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

Shunichiro Tomura [1,2,*], Mark Cooper [1,2], 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: Shunichiro Tomura, s.tomura@uq.edu.au.

Abstract

The refinement of prediction accuracy in genomic prediction is a key factor in accelerating genetic gain for crop breeding. The mainstream strategy for prediction performance improvement has been developing an individual prediction model outperforming others across diverse prediction scenarios. However, this approach has limitations in situations when there is inconsistency in the superiority of individual models, attributed to the existence of complex nonlinear interactions among genetic markers. This phenomenon is expected given the No Free Lunch Theorem, which states that the average performance of an individual prediction model is expected to be equivalent to the others across all scenarios. Hence, we investigate the potential to leverage the concept of a stacked ensemble as an alternative method. We consider two traits, days to anthesis (DTA) and tiller number (TILN), measured on a Nested Association Mapping study, referred to herein as TeoNAM; a public maize (*Zea mays*) inbred W22 was crossed to five inbred Teosinte lines. The TeoNAM data set and the two traits were selected as the example of choice based on prior evidence that the traits were under the control of networks of genes and high levels of segregation diversity for the nodes of the genetic networks. Our analysis of both traits for the TeoNAM demonstrated an improvement in prediction performance, measured as the Pearson correlation, for the ensemble approach across all the proposed scenarios, for at least more than 95% of cases, compared to the six individual prediction models that contributed to the ensemble; rrBLUP, BayesB, RKHS, RF, SVR and GAT. The observed result indicates that there is a potential for ensemble approaches to enhance the performance of genomic prediction for crop breeding.

Key message: Ensemble approach can improve genomic prediction performance by combining information from individual models.

Keywords: ensemble, genomic prediction, machine learning, explainability, No Free Lunch Theorem, crop breeding

1 Introduction

Genomic selection has accelerated rates of genetic gain in plant breeding programs (Voss-Fels et al., 2019) by using gene-phenotype (G2P) 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). 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 (Prasanna et al., 2021; Dreisigacker et al., 2021).

The selection accuracy of genomic selection is dependent on the choice of G2P model. Therefore, research evaluating alternative G2P models has received considerable attention (Poudel et al., 2023), revealing that no G2P model has consistently outperformed other models. This has been attributed to the difficulty in capturing intermediate biological interactions via linear G2P models between genetic markers and complex trait phenotypes. Mackay (2014) demonstrated that including both additive (linear) and epistatic (non-linear) effects in G2P models can capture the genetic effects more precisely. Doebley et al. (1995) showed that the branch length of maize is determined by both a main effect of gene teosinte branched1 as well as interactions with multiple other genes. Such complex interactions can even be further amplified by gene-by-environment (GxE) interactions (Cooper et al., 2005; Lasky et al., 2022).

Outside of plant breeding, the absence of a best prediction model has been postulated by the No Free Lunch Theorem (Wolpert and Macready, 1997), which suggests that the average prediction performance of individual models becomes equal over all prediction problem scenarios. If we accept that this theorem applies generally, seeking superior individual G2P models for genomic selection in plant breeding programs is unlikely to be successful. A potential alternative approach could be to generate ensembles of diverse G2P models.

Ensemble approaches combine predictions from multiple models to predict the target (Farooq et al., 2022). Several ensemble approaches have been proposed, including; Bootstrap aggregating (bagging) (Breiman, 1996) and AdaBoost (Freund and Schapire, 1995), and those ideas have been adapted to random forest (Breiman, 2001) and ensemble neural network approaches (Zhou et al., 2002; Li et al., 2004; Liu and Li, 2008).

Ensemble approaches have been leveraged in agricultural science but applications are lacking in plant breeding. For crop yield prediction, Wallach et al. (2018) and Wallach et al. (2019) demonstrated higher prediction accuracy of ensemble approaches for forecasting climate change scenarios. Within the plant breeding community, a few examples have shown the superiority of ensemble approaches compared to

approaches using individual G2P models (McCormick et al., 2021; Huang and Wei, 2022; Heilmann et al., 2023; Fradgley et al., 2023). However, the investigation of factors leading to performance increases of ensemble approaches for genomic selection has not been well discussed.

In this paper, we examine the performance of ensemble approaches for genomic selection in comparison to that of the individual G2P models and dissect the factors leading to performance increases. We demonstrate the potential of the ensemble approaches using the TeoNAM population (Chen et al., 2019) as an example. The TeoNAM consists of five diverse maize-teosinte crosses designed as nested association mapping (NAM) populations. The objectives of this paper are to (1) investigate the prediction performance of six individual G2P models for a set of the chosen traits that are expected to be under a genetic architecture involving both linear and complex nonlinear interactions that can be detected using genome wide SNPs, (2) consider the implications of the No Free Lunch (NFL) Theorem and investigate an ensemble approach to assess its effectiveness in comparison to the individual genomic prediction model outcomes, and (3) outline a framework to investigate the contribution of individual prediction models to an ensemble and identify situations where prediction performance improvements can be expected from an ensemble approach.

2 Materials & Methods

2.1 Overview

This section describes the TeoNAM dataset (Chen et al., 2019), methods used for data quality assessment and cleaning, the six individual models leveraged in the experiments and the algorithms for extracting estimated SNP effect values from each of the proposed models. The six individual models selected were, rrBLUP, BayesB, Reproducing Kernel Hilbert Space (RKHS), random forest (RF), support vector regression (SVR) and graph attention network (GAT). For demonstration purposes, a naive-average model was applied as the ensemble approach, where each of the six prediction predictions was allowed to contribute equally to the ensemble.

2.2 Dataset

The TeoNAM dataset was developed from a series of five recombinant inbred line populations derived from five BC1 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 BC1 cross, each population was advanced four times by controlled self-pollination to produce the recombinant inbred lines (RILs) that were used for the analyses.

Each population dataset consisted of seven agronomic traits and fifteen domestication traits. From the trait list, day to anthesis (DTA) and tiller number (TILN) were chosen as target traits for the ensemble investigations. 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; Poudel et al., 2023).

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 grown 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 marker (SNP) and RIL numbers per cross are described in Table 1 (the detail of data preprocessing is explained in 2.3), indicating W22TIL03 possesses the highest SNP numbers, followed by W22TIL11, W22TIL01, W22TIL25 and W22TIL14. As per the number of lines, W22TIL25 has the largest number at 308, followed by W22TIL03, W22TIL14, W22TIL01 and W22TIL11. Each population contains at least 200 lines with more than 10,000 SNPs.

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

	SNPs			RIL
	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

2.3 Data cleaning & preprocessing

The original datasets were evaluated and quality control, data cleaning and preprocessing were undertaken for the following reasons. Missing marker data was one noticeable problem affecting the quality of the datasets. Imputation was undertaken when possible and when considered inappropriate, SNPs with a high number of missing values were removed. Another issue is the high proportion of the marker numbers relative to the RIL numbers in each cross. The number of markers in all the crosses is significantly higher than the number of RILs (Table 1). The RIL numbers may not be sufficient to capture complex patterns among markers and phenotypes, resulting in the curse of dimensionality (Bellman, 1957; Ramstein et al., 2019). Therefore, we undertook marker dimensionality reduction by eliminating less informative markers based on their linkage disequilibrium relationships.

The former was solved by imputing missing values with the most frequent allele from each marker after removing the markers with missing rates higher than 10%. Concurrently, RILs without target trait phenotypes were excluded from the experiment. For the latter, linkage disequilibrium (LD) filtering was leveraged by Plink (Chang et al., 2015). When the squared correlation was higher than the threshold of 0.8, the compared SNP was removed from the datasets. A higher correlation indicates that two SNPs provide similar information for prediction and thus removing either of the SNPs can be beneficial due to the reduction of the total number of SNPs rather than losing critical information for predictions.

Since each cross dataset consists of two sub-datasets from two distinctive environments, the sub-datasets can be concatenated into one data set with a dummy variable representing different environments to increase the available training data size. The datasets after undertaking all the preprocessing steps are used as input for the six G2P prediction models. The final SNPs and RIL numbers are summarised in Table 1. The final number of SNPs varied depending on the combination of RILs in the training set for each sample.

2.4 Individual models

2.4.1 Parametric models

Parametric models have been widely leveraged for genomic prediction. They do not only capture linearity but can also capture nonlinearity by adding terms representing the interaction between SNPs. 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 (Rothenberg, 1971).

One of the most well-known parametric models in genomic prediction is 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 below (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 genomic estimated breeding values (GEBVs) 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 marker effect 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 (VanRaden, 2008), 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 $\boldsymbol{\beta}_j \sim N(0, \mathbf{I}, \sigma^2_{\boldsymbol{\beta}_j})$ where $\mathbf{I}$ is an identity matrix and $\sigma^2_{\boldsymbol{\beta}_j}$ is the variance of 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 with rrBLUP (Habier et al., 2007) and thus both models share the same mechanisms and assumptions (Clark and van der Werf, 2013). The rrBLUP model implementation was chosen as one of the G2P 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 marker effect $\boldsymbol{\beta}_j$ to have different distributions (Wang et al., 2018). This variation in the distribution is expected to capture more realistic marker effects whenever it is unrealistic for all the marker effects to be from the same distribution with small effects, as rrBLUP assumed. Another noticeable characteristic of the Bayesian models is the application of prior distributions, constructed from prior knowledge, assumptions and statistics (Kruschke, 2011), 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 for our experiment because it has been widely used as one of the standard models in genomic prediction (John et al., 2022; Farooq et al., 2022; Abdollahi-Arpanahi et al., 2020; Meher et al., 2022; Plavšin et al., 2022). For BayesB some marker effects are expected not to be influential on the target phenotypes and their 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 the experiment.

2.4.2 Semiparametric models

Semiparametric models possess both parametric and non-parametric properties, indicating that the models can capture both linear and non-linear relationships from two different approaches. One of the semi-parametric 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 non-linear relationships among the data points. Any relationship that may not be captured in a lower 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 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. Those equations imply that the RKHS model consists of both additive and non-additive components, aimed at capturing 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.

2.4.3 Nonparametric models

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. One large class of such models are the examples of machine learning (ML) methods. From the various ML methods, we selected random forest (Breiman, 2001), support vector regression (Drucker et al., 1996) and graph attention network (Veličković et al., 2018) for our investigation of ensemble prediction.

Random forest (RF) contains a collection of decision trees (Belson, 1959) trained by sub-train datasets sampled from the original training dataset (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 generated from 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 datasets 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 machine-learning models in genomic prediction (González-Camacho et al., 2018; Sandhu et al., 2021; John et al., 2022; Farooq et al., 2022).

Support vector regression (SVR) (Drucker et al., 1996) is another machine learning approach that draws a hyperplane between data points labelled with continuous 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):

$$U(\sum_{i=1}^{N} \xi_i^* + \sum_{i=1}^{N} \xi_i) + \frac{1}{2}(\mathbf{w}^t\mathbf{w})$$

$$\begin{aligned} s.t. \quad y_i - (\mathbf{w}^t\mathbf{v}_i) - b &\leq \epsilon + \xi_i \\ (\mathbf{w}^t\mathbf{v}_i) + b - y_i &\leq \epsilon + \xi_i^* \\ \xi_i^* &\leq 0 \\ \xi_i &\leq 0 \end{aligned} \tag{4}$$

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 also be equipped with a kernel, mapping data points to another dimension for capturing non-linear interactions. SVR has also been leveraged for various experiments in genomic prediction as a standard method (John et al., 2022; Li et al., 2023; Yu et al., 2021; An et al., 2021), selected as one prediction model for our experiment.

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. Markers and phenotypes can be represented as nodes, and edges from the marker nodes to the phenotype nodes represent markers affecting the observed phenotype. Hence, edges are directed. In this experiment, we do not allow connections between marker nodes because the allowing the edges

between them did not improve the prediction performance and resulted in the exponential increase of computational time. GAT was chosen due to its transductive nature, indicating that a trained GAT can be utilised for a graph whose structure changed from the training phase given that attribute types in both nodes and edges are the same as the attributes used in training (Rossi et al., 2018). This characteristic can enhance the applicability of graph neural networks as graph structure can often be modified from the training phase considering its practical use.

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., 2018):

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

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 GEBV from the model.

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 there was more than one node type, phenotype nodes and marker nodes 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 twenty channels and a dropout rate of 0.9 was applied to every layer. Leaky Rectified Exponential Linear Unit (LeakyReLU) function was used for the activation functions. The total number of heads

were set as 8. The model is trained with 150 epochs by mini-batched graphs with a batch size of eight. AdamW was selected for the optimiser with a learning rate of 0.005 and weight decay of 0.00005. Due to computational limitations, the evaluation of model performance under various hyperparameter values was done heuristically rather than systematically and thus there is a potential that another combination of hyperparameter values can exceed the performance of our chosen model.

2.5 Naive ensemble-average models

A naive ensemble-average model combines the prediction results from more than one model, implemented concurrently beforehand, to achieve a higher prediction performance rather than leveraging models individually (Sagi and Rokach, 2018). Firstly each individual model outputs GEBVs from the given genetic markers. Those individual GEBVs are assembled as a GEBV matrix. Each row contains GEBVs from each individual model for a specific target genotype. This matrix is used as input data for the ensemble model, resulting in outputting prediction values as the final GEBVs. This 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 GEBVs, each $\boldsymbol{\eta}$ represents GEBVs from each individual model and $f(\cdot)$ is the ensemble function.

We used the results from all six different individual models (rrBLUP, BayesB, RKHS, RF, SVR and GAT) for the ensemble approach to satisfy one of the assumptions that individual models need to be diverse enough for ensemble approaches to achieve higher prediction accuracy (Wallach et al., 2018). Hence, six diverse individual GEBVs were generated per target genotype and were assembled to develop a GEBV matrix for the ensemble model.

As per the actual ensemble function f(·), a naive ensemble-average model was used in this study for demonstration purposes, by taking the arithmetic average of the GEBVs from all the individual 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 six for this study.

2.6 Marker effect estimation

The extraction of estimated marker effect values from each model can suggest how each model estimates the marker effect of respective SNPs for predicting target phenotypes, allowing a comparison of the G2P prediction models at their information-capturing level. For rrBLUP and BayesB, the marker effect was extracted using $\boldsymbol{\beta}$ in absolute values as all the genetic markers are represented in the range of 0 to 2 in a discrete way.

For RKHS and SVR, Shapley value (Shapley et al., 1953) was employed. The Shapley value was originally aimed at measuring the fair distribution of resources and rewards to players based on their contribution level under cooperative scenarios in game theory (Winter, 2002). In this study, each SNP is regarded as a player and the Shapley value allocates the marker effect to the SNPs based on the changes in the prediction values achieved following including and excluding the specific SNPs in combination with the inclusion of effects from other SNPs. A larger Shapley value is allocated to SNPs causing larger changes in prediction 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. Hence, the Shapley value calculates the average gap of prediction values between the case where the target feature is included and the case without the target feature. For RKHS, the Shapley value was implemented using iml (v0.11.3) (Molnar, 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 for each SNP is calculated by averaging the Shapley values for the target SNP from all the genotypes in a test dataset.

For RF, the marker effect was 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 sum of the squared value of 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 feature importance was extracted using Sklearn (v1.2.2) in Python.

The SNP effect of GAT is 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. By 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 an feature, $x^{'}$ is a baseline value of x and m is the total number of interpolation steps of integral. Similar to the Shapley value, Integrated Gradients also returns element-wise marker effects, and thus the final SNP marker effect was estimated by averaging the effect of the target SNP across the given genotypes as a test dataset. Pyg (v2.4.0) was used for the implementation.

3 Assessment criteria

3.1 Experimental flow

The experimental framework for evaluation of the individual G2P models and the ensemble is shown in Figure 1. The TeoNAM dataset is randomly split into training and test datasets. After genetic markers are selected, based on the LD-filtering using genetic marker information from the training dataset, the training dataset is leveraged to train the six individual G2P models, followed by the performance evaluation process which generates GEBVs for the corresponding test dataset. A GEBV matrix is developed by assembling the test dataset GEBVs from the six individual models to estimate the corresponding ensemble models. As for the six individual models, the prediction performance of the ensemble model is evaluated by the prediction performance for the same test dataset. Lastly, the marker effects for each of the six individual models are calculated by applying the approaches explained in Section 2.6.

To ensure the generalisability of the observed result for the TeoNAM ensemble analysis (Figure 1), we repeatedly implemented the models for several different settings. Three different training-test ratios (0.8-0.2, 0.65-0.35 and 0.5-0.5, respectively) were leveraged with a random sampling of 500 times. Hence, each population-trait combination of the process depicted in Figure 1 was evaluated over 1,500 prediction results (3 ratios*500 samples). The influence of the three training-test ratios is also analysed by analysing prediction performance metrics for the three different ratios

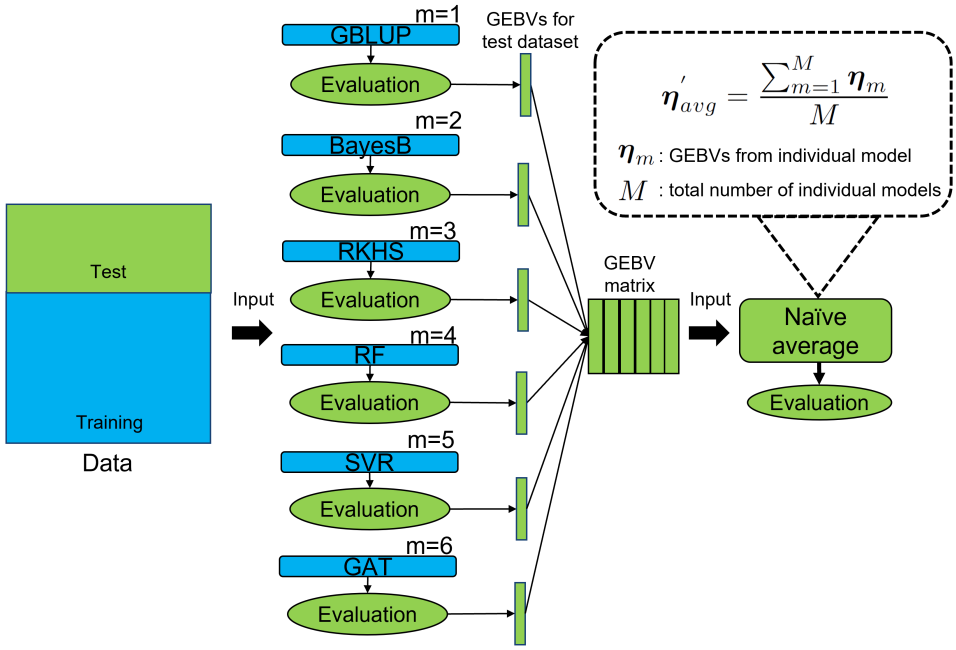

Fig. 1 Experimental flow. Blue and green colours represent components related to a training and test dataset respectively. Each individual model is trained by genetic markers in a training dataset and is evaluated by a test dataset. GEBVs for a test dataset are extracted to develop a GEBV matrix used as input data for the ensemble model.

3.2 Metrics

The Pearson correlation between the predicted and observed trait value in the test set is used as one metric used of the performance of the individual and ensemble 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 a predicted trait value, y_i is the true trait value observed for RILs in the test set and $\bar{x}$ and $\bar{y}$ are the mean values of x and y, respectively. The value ranges from 1 to -1. The value 1 indicates that all predicted values are the same as the actual trait value and as the correlation value decreases, the gap between the predicted and true trait values becomes larger. Thus, the correlation value of a model with higher prediction performance should be close to 1. The second metric used to assess the individual and ensemble prediction models is the mean squared error (MSE) between the predicted and observed trait values. The MSE measures the variance of the gap between the predicted and true trait values. 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 true trait value and $\hat{y}_i$ is the predicted value. A model with MSE close to 0 is regarded as a high-performance model because it suggests that on average the gap between the predicted and true

values is small. The value is squared to assess the cumulative effect of positive and negative deviation values.

4 Results

4.1 Ensemble of models outperformed the best individual genomic prediction models

The overall prediction result of each individual prediction model and the ensemble is shown in Figure 2. The ensemble-average model demonstrated higher prediction accuracies compared to individual models

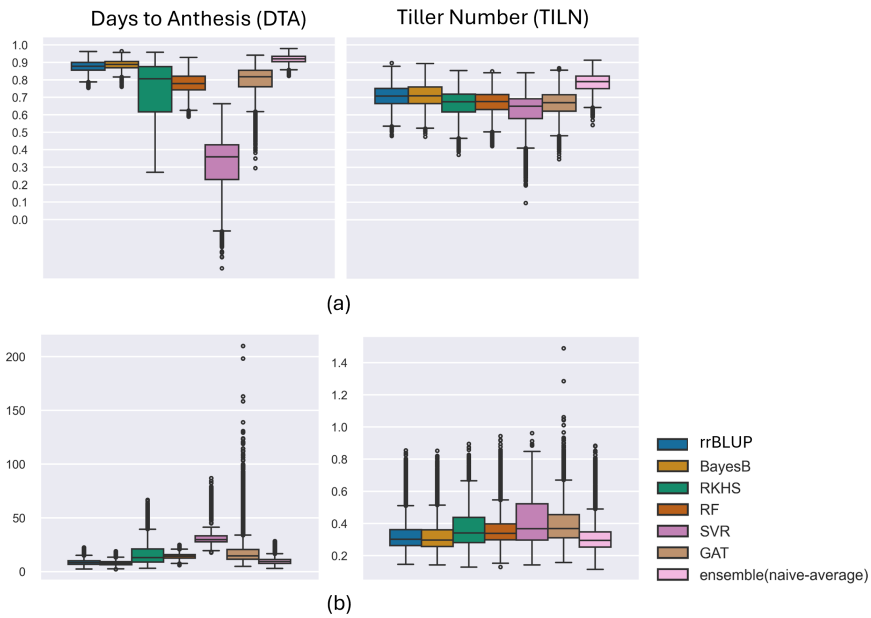

Fig. 2 A comparison of genomic prediction performance of the ensemble-average model versus each of the individual models. The performance of genomic prediction models was measured with (a) the Pearson correlation and (b) mean squared error (MSE). Horizontal lines within the boxes represent the median value and individual points represent outliers.

of the ensemble across multiple traits. While the model with the lowest prediction errors depended on the trait analysed, BayesB was the individual model with the highest selection accuracies and lowest prediction accuracies when averaged across different training-test sets. However, there was no consistent "best" individual model across all the training test sets.

For the DTA trait, the highest prediction accuracies were observed with the ensemble-average model, but the lowest prediction errors were observed with an individual model, BayesB. The median prediction accuracy of the best individual model, BayesB, was 0.887 while the median prediction accuracy of the

ensemble-average model was 0.919. The median mean squared error of the best individual model and best model overall, BayesB, was 8.048 while the median mean squared error of the ensemble-average model was 10.076.

For the TILN trait, the highest prediction accuracies and lowest prediction errors were observed with the ensemble-average model. The median prediction accuracy of the best individual model, BayesB, was 0.708 while the median prediction accuracy of the ensemble-average model was 0.790. Unlike the DTA trait, the ensemble-average model had the lowest median mean squared error of all models. The median mean squared error of the best individual model, BayesB, was 0.315 while the median mean squared error of the ensemble-average model was 0.270.

4.2 No consistent winner amongst individual genomic prediction models

There was no consistent winner amongst individual models across traits and samples of training-test sets. The individual model with the highest prediction accuracies varied across training-test set sizes, while the individual model with the lowest prediction errors varied across traits. Table 2 presents the percentage of sample training-test sets in which each model achieved the highest prediction accuracy or lowest mean squared error.

Table 2 Percentage of superior performance achieved by each model across 500 samples achieved by training-test ratio (0.8, 0.65, 0.5), trait (DTA, TILN) and metric (Pearson correlation and MSE)

		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**

The highest prediction accuracies were observed with BayesB between 0.44% and 2.08% and 0.24% and 0.40% of training-test set samples across training-test sizes for DTA and TILN respectively while the highest prediction accuracies were observed with rrBLUP between 0.04% and 1.24% and 0.00% of training-test set samples across training-test ratios for DTA and TILN respectively. The highest prediction accuracies were observed with the other individual models (RKHS, RF, SVR and GAT) in less than 1.00% of training-test set samples.

The individual model with the highest percentage of lowest prediction errors across training-test set samples was training-test, set size and trait dependent. For DTA, the lowest prediction errors were observed with rrBLUP for larger training set sizes, but the lowest prediction errors were observed with BayesB for smaller training set sizes. The lowest prediction errors were observed with rrBLUP in 39.32% of training-test sets with a ratio of training-test set size of 0.8 while the lowest prediction errors were

observed with BayesB in 72.00% and 94.12% of training-test sets with ratios of training-test set sizes of 0.65 and 0.5, respectively. For TILN, the lowest prediction errors among the individual models were observed with BayesB across all ratios of training-test set sizes.

4.3 Large variation in the total genetic values and SNP effects predicted by individual genomic prediction models

Large magnitudes of variation were observed in the total genetic values and SNP effects across classical and machine learning genomic prediction models. Positive associations were observed in pairwise comparisons between predicted total genetic values of individual genomic prediction models. Generally, there was no associations were observed in pairwise comparisons between predicted SNP effects of individual genomic prediction models. The exception to this trend was a strong positive association of predicted SNP effects between rrBLUP and BayesB models.

Figure 3 presents pairwise comparisons of predicted total genetic values and SNP effects between the individual prediction models. The upper off-diagonal scatter diagrams compare predicted total genetic values. The lower off-diagonal scatter diagrams compare the predicted SNP effects.

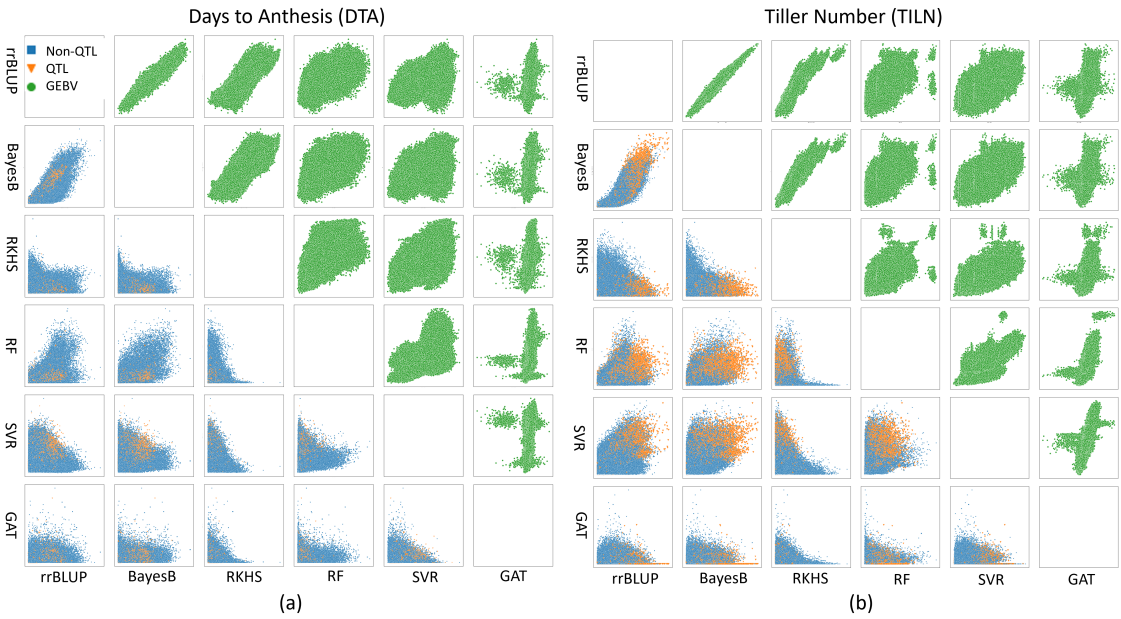

Fig. 3 Pairwise comparison of individual models at GEBV (top right triangle) and SNP effect (the bottom left) levels for both traits. The green circles, blue squares and orange triangles represent GEBVs, Non-QTL effect and QTL effect introduced in Chen et al. (2019) respectively for the days to anthesis (DTA) trait (a) and the tiller number (TLN) trait (b)

For predictions of total genetic values, the strength of the associations varied among the individual genomic prediction models. Strong positive associations were observed among the classical genomic

prediction models. High Pearson correlations were observed between predicted total genetic values of rrBLUP and BayesB (0.9997), rrBLUP and RKHS (0.9981) and BayesB and RKHS (0.9977). Weaker, but still highly positive, associations were observed among the predicted total genetic values of classical and machine learning genomic prediction models with the exception of comparisons with the GAT model. As an example, Pearson correlations were observed between predicted total genetic values of rrBLUP and RF (0.9972), rrBLUP and SVR (0.9954) and rrBLUP and GAT (0.9964). The weakest positive associations were observed among the predicted total genetic values of machine learning models. Pearson correlations were observed between predicted total genetic values of RF and SVR (0.9965), RF and GAT (0.9968) and SVR and GAT (0.9966).

For predictions of SNP effects, a lack of association was consistently observed among the individual genomic prediction models with the exception of rrBLUP and BayesB. Among the classical genomic prediction models, the Pearson correlations between predicted SNP effects of rrBLUP and BayesB, rrBLUP and RKHS, and BayesB and RKHS were 0.7688, 0.0158 and 0.0199 respectively. Among the classical and machine learning genomic prediction models, the Pearson correlations between predicted SNP effects of rrBLUP and RF, rrBLUP and SVR and rrBLUP and GAT were 0.1386, 0.2844 and 0.3530 respectively. The Pearson correlations between predicted SNP effects of BayesB and RF, BayesB and SVR and BayesB and GAT were 0.3828, 0.3021 and 0.3607 respectively. The Pearson correlations between predicted SNP effects of RKHS and RF, RKHS and SVR and RKHS and GAT were 0.0274, -0.0110 and 0.0120 respectively. Among the machine learning genomic prediction models, the Pearson correlations between predicted SNP effects of RF and SVR, RF and GAT and SVR and GAT were 0.2505, 0.1494 and 0.3102 respectively.

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

The ensemble-average model demonstrated higher prediction accuracies and lower prediction errors compared to the individual, constituent models of the ensemble across multiple traits (Figure 4).

For the DTA trait, the average of the median prediction accuracy of the individual models (rrBLUP, BayesB, RKHS, RF, SVR, GAT) was 0.719 while the median prediction accuracy of the ensemble-average model was 0.919. The average of the median mean squared error of the individual models was 16.838 while the median mean squared error of the ensemble-average model was 10.167.

For the TILN trait, the average of the median prediction accuracy of the individual models (rrBLUP, BayesB, RKHS, RF, SVR, GAT) was 0.671 while the median prediction accuracy of the ensemble-average

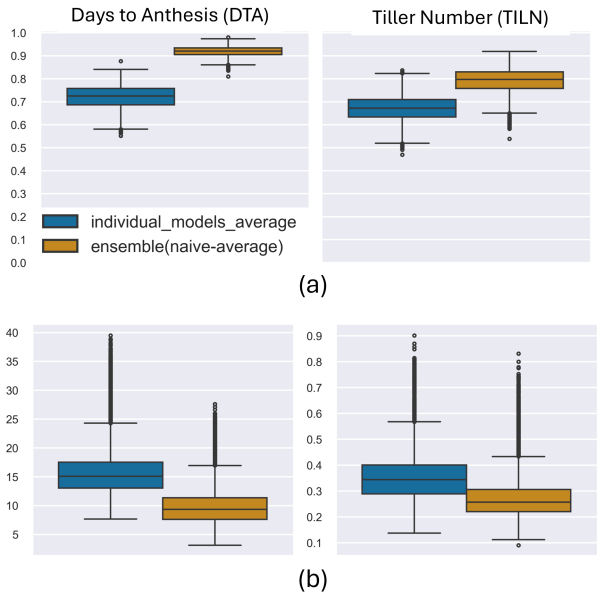

Fig. 4 A comparison of genomic prediction performance of the ensemble-average model versus the average of individual models. The performance of genomic prediction models was measured with (a) the Pearson correlation and (b) mean squared error (MSE). Horizontal lines within the boxes represent the median value and individual points represent outliers.

model was 0.790. The average of the median mean squared error of the individual models was 0.356 while the median mean squared error of the ensemble-average model was 0.277.

4.5 Classical genomic prediction models outperform machine learning models

Figure 2 indicates that classical genomic prediction models (rrBLUP, BayesB and RKHS) demonstrated higher prediction accuracies and lower prediction errors than machine learning models (RF, SVR and GAT), although the magnitude of differences was trait-dependent.

For the DTA trait, SVR had considerably lower Pearson correlations (<0.4) and higher mean squared errors (>20) between predicted and observed values than other individual models. RF and GAT demonstrated comparable prediction accuracies and prediction errors to RKHS with average Pearson correlations (0.78, 0.79 and 0.74) and mean squared errors (19.36, 35.62 and 17.50), respectively. rrBLUP and BayesB demonstrated the highest prediction accuracies (0.88 and 0.89) and mean squared errors (8.74 and 8.06) of all the individual models.

For the TILN trait, the differences in prediction accuracies and prediction errors between the classical genomic prediction models and machine learning models were smaller. SVR demonstrated the lowest average Pearson correlations (0.31) and highest mean squared errors (14.28) between predicted and observed values than other individual models. RF and GAT demonstrated comparable prediction

accuracies and prediction errors to RKHS with average Pearson correlations (0.67, 0.67 and 0.66) and mean squared errors (0.36, 0.40 and 0.37), respectively. rrBLUP and BayesB demonstrated the highest prediction accuracies (0.71 and 0.71) and mean squared errors (0.33 and 0.32) of all the individual models.

5 Discussion

The results of this study highlight three points for discussion:

i There was no consistent winner among individual genomic prediction model

ii Ensembles of models provided improved prediction outcomes compared to individual models

iii The improvement in prediction outcomes of the ensemble over individual models was trait-dependent

5.1 No consistent winner amongst individual genomic prediction models

For TILN in four of the crosses, prediction performance for the individual models was similar. The exception was W22TIL14. However, the ranking for performance prediction varied among the G2P models. For example, for TILN, RKHS outperformed the other individual models in cross W22TIL11. However, its prediction performance dropped in cross W22TIL14, while BayesB outperformed the others. SVR performed similarly to rrBLUP, BayesB and RKHS in cross W22TIL01, while its prediction performance decreased in W22TIL03. This lack of consistency in the relative prediction performance demonstrates the limitations of individual G2P models for general prediction applications. No all-rounder G2P model was identified from among the six individual models considered. The relative performance of the individual G2P models for prediction was affected by the combinations of traits and crosses. The phenomenon of the lack of a clear prediction winner for complex problems is well-known as the No Free Lunch Theorem (Wolpert and Macready, 1997). The theorem states that the averaged performance of all models over a problem state space is equivalent to that for the others, indicating that no model can outperform others for all scenarios. The applicability of the theorem has been investigated for a wide range of prediction optimisation problems, the structure of which is relevant for applications of genomic prediction for quantitative traits that are targeted in plant breeding (Fernández-Delgado et al., 2014; Gómez and Rojas, 2016; Plavšin et al., 2022; Merrick and Carter, 2021).

The implications of the theorem can also be examined in Figure 3 at the SNP effect and GEBV levels whenever differences in individual SNP effects sum to similar GEBV predictions. Therefore, we argue that it is not strategic to focus on developing only an individual G2P model for a diverse range of genomic prediction tasks to reach a higher level of prediction performance. The ensemble approach can be one solution to overcome the prediction limitation indicated by the No Free Lunch Theorem, as observed from this investigation.

5.2 Ensembles of models provided improved prediction outcomes compared to individual models

For the DTA and TILN traits within the TeoNAM experiment, the consensus is the performance superiority of the ensemble model in comparison to the individual G2P models (Figure 2, Table 2). Although the ensemble model did not achieve the lowest MSE level for DTA, it was almost equivalent to the individual models that achieved the highest performance.

One possible explanation for the success of the ensemble of models can be the differences in captured information among the individual G2P models. Individual models can capture only a fragment of total G2P information based on the properties of any individual type of algorithm to predict phenotypes. Given the high G2P dimensionality of quantitative traits, such information may not be sufficient to achieve constant high prediction performance, resulting in similar performance on average. This phenomenon is called the Rashomon effect (Breiman, 2001), which states that there is a set of models (Rashomon sets) sharing similar prediction performance while they capture different feature effects due to distinct properties of the prediction algorithms. Therefore, by combining alternative information pieces from the different G2P models through the ensemble approaches, a more comprehensive representation of trait architecture information can be achieved by the ensemble, which can be utilised to conduct more accurate (higher Pearson correlation) and precise (lower MSE) trait predictions. This concept was theoretically demonstrated by Wallach et al. (2018), concluding that if each component model in an ensemble model is randomly chosen to diversify their algorithms, the ensemble model achieves higher prediction performance than the individual performance of the component models. Although the selection of the six component models in our experiment is one of many possibilities, each of the prediction algorithms is distinctively diverse, resulting in satisfying the diversity assumption. For example, the parametric models (rrBLUP, BayesB) represent models that focus on SNP main effects first followed by consideration of interactions. In contrast, the semi-parametric model RKHS provides a mechanism for explicit consideration of SNP interaction terms. Further, the non-parametric machine learning models focus on SNP interactions.

Another potential explanation for the superior performance of the ensemble of models can be the avoidance of stagnation in local optima. Each of the six G2P modelling algorithms may achieve different local optima within their possible prediction range, the size of which depends on the given data and algorithms. Algorithms that are stuck on local optima can inhibit opportunities to explore the problem state space for alternative solutions closer to the global optima, where more precise predictions may be achieved. The application of another algorithm as the next layer of a set of individual models can increase the likelihood of shifting the prediction values from each individual model focused on exploiting local optima, towards the global optima, which have not been able to be reached individually (Sagi and Rokach,

2018). This interpretation proposes that the ensemble approaches enhance the ability of the collective ensemble of search algorithms to reach problem solutions outside the reach of individual models. If the superior prediction models do not plot within the hypothetical prediction space of each model, neither of them can correctly predict the value, resulting in lower prediction performance. However, as Dietterich (2000) mentioned, the prediction of values outside the hypothetical spaces of individual models can be feasible by creating a new hypothetical space through the ensemble of individual models. In short, the compensation of the individual weakness of each prediction model enables the construction of solutions for complex tasks with higher performance (Dong et al., 2020).

5.3 Improvement in prediction outcomes of ensemble model was trait dependent

Both individual and ensemble G2P models demonstrated higher prediction performance for DTA compared to TILN with consistency. This contrast suggests greater difficulty for TILN prediction, which may be a consequence of the existence of nonlinear complex marker interactions involved in the genetic architecture of the branching network determining TILN in the diverse crosses (Doebley et al., 1995; Bertheloot et al., 2020; Powell et al., 2022). While networks of genes are involved in the control of DTA (Dong et al., 2012), following the results of Buckler et al. (2009), approximate models based on contributions of additive effects appear to have been sufficient to account for the phenotypic diversity for DTA in the TeoNAM crosses. The observed performance pattern of RF and SVR indicates the presence of complex genetic marker interactions. They often achieve higher prediction performance by capturing the effects of non-additivity rather than additive relationships. The lower performance of RF, SVR and GAT in DTA may be a consequence of their lower ability to construct the simpler models mainly aimed at detecting additive effects. Hence, the rrBLUP and BayesB focused primarily on additive models may have been able to outperform RF and SVR. RKHS gave more variable prediction results for DTA suggesting important contributions of interactions between SNPs. The GAT machine learning model, on the other hand, is a more complex prediction model exploiting graph structures present in data, which may have contributed to the ability of the method to capture both additive and nonadditive effects to some extent.

6 Further Opportunitiess

Several limitations might have affected the observed results from this analysis of the data sets available for the TeoNAM experiment.

Firstly, the hyperparameters of the G2P models have been adjusted heuristically rather than through systematic investigation such as cross-validation or Bayesian methods, primarily due to computational expensiveness. In addition, a number of the parameters are set as the recommended default. Hence, the optimisation of (hyper) parameters may result in further performance improvement for both individual and ensemble models.

The refinement of the data preprocessing can be another factor contributing to improving model performance. The current method adopted a simple imputation method by replacing missing values with the most frequent allele in each SNP. More complex imputation methods, through the development of a reference genome as an example, may enhance the imputation accuracy. The identification of the trait contribution level for each SNP was based on an RF model. Other approaches possessing the ability to calculate the contribution level should also be considered as a further research opportunity. Thirdly, the addition of other G2P prediction methods, in addition to the six considered here, may increase the prediction accuracy. Considering the diversity of the available G2P prediction models, as discussed above, is a promising area for further investigation.

The observed superior performance in the ensemble model motivates the further exploration of ensemble methods, which can be oriented by different components. One approach can be the optimisation of weights entailed with prediction results from each individual model. In this study, we applied a na¨ıve ensemble where each of the six individual prediction models contributed equally in the ensemble averaging step. This is one of many possible weighting approaches. Hence, the investigation of weight optimisation can be one critical research point as Wallach et al. (2016) suggested. In some applications, higher prediction performance was achieved by tuning weights accompanying each individual model (Yu et al., 2021; Liang et al., 2021; McCormick et al., 2021; Wang et al., 2023). A grid-search approach can be considered as one option, but the computational time for large ensembles can be prohibitive as all the possible weight combinations between the models need to be tested. Another approach can be a contribution-based approach, which is based on the SNP effect level in each individual model as explained in Section 2.6. A third method can be the determination of the optimum weights through training by ML-based approaches. A fully connected neural network consists of a number of weights including the edges between the input and a hidden layer. The more important connection is weight during the training phase, and by extracting such weight information, the weight of each individual model can be optimally adjusted.

The development of other ensemble models can also be a significant component of the extension. While the simple ensemble method such as average and median improved prediction performance in this experiment and that of Wallach et al. (2018), the merit of ensembles of information can be amplified by

developing a complex nonlinear model capturing detailed interactions between prediction results from the individual models. The application of statistical models such as regressions can be one approach, but such simple models may not be able to significantly improve prediction performance and the performance gap between averaging used in this experiment can be small. On the other hand, ML models can be developed to capture nonlinear interactions, expected to achieve higher levels of prediction performance. One drawback of this approach is the requirement for larger data sets for training and thus may not be suitable if the total number of individual models is relatively large compared to the total number of records. Another approach can be information selection (Wallach et al., 2016). Although Wallach et al. (2018) demonstrated that an increase in the total number of included individual models led to performance improvement, research in ensemble approaches (Zhou et al., 2002; Li et al., 2004; Huang and Wei, 2022) has shown the effectiveness of model selection methods. This improvement has been attributed to reducing noise or redundant information to enhance the performance of an ensemble model, which can strongly affect the performance of an ensemble model that is complex or susceptible to noise. These early results widen the research target in developing efficient and effective ensemble approaches and several selection methods have been proposed, including pairwise-based, correlation-based, cluster-based, and importance-level-based ensemble identification.

7 Conclusion

We investigated the prediction performance of a stacked ensemble model over six individual G2P models for two traits for the TeoNAM dataset. Prior evidence indicates that both traits are under the control of networks of genes. Thus, we hypothesised for the TeoNAM experiment an ensemble approach could have advantages over individual prediction models. Our results demonstrated an advantage for predictions based on the ensemble of models. Therefore, there is a potential that ensemble approaches can be a promising tool for genomic prediction. We interpret the demonstrated advantage of the ensemble of models for the TeoNAM study to be a consequence of the diversity of the G2P information captured by the individual G2P models and the incorporation of more comprehensive G2P information by the ensembles. Opportunities were identified for further research to investigate the effectiveness of and improve ensemble approaches for purposes of generalising the method for application to other datasets.

Declarations

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)

Authors' contributions MC OP developed the research ideas. ST, OP and MC developed the experimental design and analyses ST undertook data analysis, developed the ensemble model and wrote the initial draft. OP refined and contributed to writing later versions of the manuscript. MC helped refine the manuscript.

Data availability All the datasets leveraged in this study are collected by Chen et al. (2019) and full genotype and phenotype data are publicly available at: https://datacommons.cyverse.org/browse/iplant/home/shared/panzea/genotypes/GBS/TeosinteNAM and https://gsajournals.figshare.com/articles/dataset/Supplemental_Material_for_Chen_et_al_2019/9250682.

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

References

Voss-Fels, K.P., Cooper, M., Hayes, B.J.: Accelerating crop genetic gains with genomic selection. Theoretical and Applied Genetics **132**, 669–686 (2019)

Meuwissen, T.H., Hayes, B.J., Goddard, M.E.: Prediction of total genetic value using genome-wide dense marker maps. Genetics **157**(4), 1819–1829 (2001)

Heffner, E.L., Sorrells, M.E., Jannink, J.-L.: Genomic selection for crop improvement. Crop Science **49**(1), 1–12 (2009)

Gaynor, R.C., Gorjanc, G., Bentley, A.R., Ober, E.S., Howell, P., Jackson, R., Mackay, I.J., Hickey, J.M.: A two-part strategy for using genomic selection to develop inbred lines. Crop Science **57**(5), 2372–2386 (2017)

Powell, O., Gaynor, R.C., Gorjanc, G., Werner, C.R., Hickey, J.M.: A two-part strategy using genomic selection in hybrid crop breeding programs. BioRxiv, 2020–05 (2020)

Cooper, M., Gho, C., Leafgren, R., Tang, T., Messina, C.: Breeding drought-tolerant maize hybrids for the us corn-belt: discovery to product. Journal of Experimental Botany **65**(21), 6191–6204 (2014)

Prasanna, B.M., Cairns, J.E., Zaidi, P., Beyene, Y., Makumbi, D., Gowda, M., Magorokosho, C., Zaman-Allah, M., Olsen, M., Das, A., *et al.*: Beat the stress: breeding for climate resilience in maize for the tropical rainfed environments. Theoretical and Applied Genetics **134**(6), 1729–1752 (2021)

Dreisigacker, S., Crossa, J., Pérez-Rodríguez, P., Montesinos-López, O.A., Rosyara, U., Juliana, P., Mondal, S., Crespo-Herrera, L., Govindan, V., Singh, R.P., Braun, H.-J.: Implementation of genomic selection in the cimmyt global wheat program, findings from the past 10 years. Crop Breeding, Genetics and Genomics **3**(2) (2021)

Poudel, P., Naidenov, B., Chen, C., Alderman, P.D., Welch, S.M.: Integrating genomic prediction and genotype specific parameter estimation in ecophysiological models: overview and perspectives. in silico Plants, 007 (2023)

Mackay, T.F.: Epistasis and quantitative traits: using model organisms to study gene–gene interactions. Nature Reviews Genetics **15**(1), 22–33 (2014)

Doebley, J., Stec, A., Gustus, C.: teosinte branched1 and the origin of maize: Evidence for epistasis and the evolution of dominance. Genetics **141**(1), 333–346 (1995)

Cooper, M., Podlich, D.W., Smith, O.S.: Gene-to-phenotype models and complex trait genetics. Australian Journal of Agricultural Research **56**, 895–918 (2005)

Lasky, J.R., Josephs, E.B., Morris, G.P.: Genotype–environment associations to reveal the molecular basis of environmental adaptation. The Plant Cell (2022)

Wolpert, D.H., Macready, W.G.: No free lunch theorems for optimization. IEEE Transactions on Evolutionary Computation **1**(1) (1997)

Farooq, M., Dijk, A.D.J., Nijveen, H., Mansoor, S., Ridder, D.: Genomic prediction in plants: opportunities for ensemble machine learning based approaches. F1000 Research (2022)

Breiman, L.: Bagging predictors. Machine learning **24**, 123–140 (1996)

Freund, Y., Schapire, R.E.: A desicion-theoretic generalization of on-line learning and an application to boosting. In: European Conference on Computational Learning Theory, pp. 23–37 (1995). Springer

Breiman, L.: Random forest. Machine Learning **45**, 5–32 (2001)

Zhou, Z.-H., Wu, J., Tang, W.: Ensembling neural networks: Many could be better than all. Artificial

Intelligence **137**(1-2), 239–263 (2002)

Li, K., Huang, H., Ye, X., Cui, L.: A selective approach to neural network ensemble based on clustering technology. International Conference on Machine Learning and Cybernetics (2004)

Liu, T.-Y., Li, G.-Z.: Fault diagnosis by using selective ensemble learning based on mutual information. The Second International Symposium on Optimization and Systems Biology, 191–197 (2008)

Wallach, D., Martre, P., Liu, B., Asseng, S., Ewert, F., Thorburn, P.J., Ittersum, M., Aggarwal, P.K., Ahmed, M., Basso, B., Biernath, C., Cammarano, D., Challinor, A.J., Sanctis, G.D., Dumont, B., Rezaei, E.E., Fereres, E., Fitzgerald, G.J., Gao, Y., Garcia-Vila, M., Gayler, S., Girousse, C., Hoogenboom, G., Horan, H., Izaurralde, R.C., Jones, C.D., Kassie, B.T., Kersebaum, K.C., Klein, C., Koehler, A., Maiorano, A., Minoli, S., Müller, C., Kumar, S.N., Nendel, C., O'Leary, G.J., Palosuo, T., Priesack, E., Ripoche, D., Rötter, R.P., Semenov, M.A., Stöckle, C., Stratonovitch, P., Streck, T., Supit, I., Tao, F., Wolf, J., Zhang, Z.: Multimodel ensembles improve predictions of crop–environment–management interactions. Global Change Biology **24**(11) (2018)

Wallach, D., Makowski, D., Jones, J.W., Brun, F.: Multimodel ensembles. Working with Dynamic Crop Models (Third Edition), 425–443 (2019)

McCormick, R.F., Truong, S.K., Rotundo, J., Gaspar, A.P., Kyle, D., Van Eeuwijk, F., Messina, C.D.: Intercontinental prediction of soybean phenology via hybrid ensemble of knowledge-based and data-driven models. in silico Plants **3**(1), 004 (2021)

Huang, W.-H., Wei, Y.-C.: A split-and-merge deep learning approach for phenotype prediction. Frontiers in Bioscience-Landmark **27**(3), 78 (2022)

Heilmann, P.G., Frisch, M., Abbadi, A., Kox, T., Herzog, E.: Stacked ensembles on basis of parentage information can predict hybrid performance with an accuracy comparable to marker-based gblup. Frontiers (2023)

Fradgley, N., Gardner, K.A., Bentley, A.R., Howell, P., Mackay, I.J., Scott, M.F., Mott, R., Cockram, J.: 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**(1), 002 (2023)

Chen, Q., Yang, C.J., York, A.M., Xue, W., Daskalska, L.L., DeValk, C.A., Krueger, K.W., Lawton, S.B., Spiegelberg, B.G., Schnell, J.M., Neumeyer, M.A., Perry, J.S., Peterson, A.C., Kim, B., Bergstrom,

L., Yang, L., Barber, I.C., Tian, F., Doebley, J.F.: Teonam: A nested association mapping population for domestication and agronomic trait analysis in maize. Genetics **213**(3), 1065–1078 (2019)

Dong, Z., Danilevskaya, O., Abadie, T., Messina, C., Coles, N., Cooper, M.: A gene regulatory network model for floral transition of the shoot apex in maize and its dynamic modeling (2012)

DeWitt, N., Guedira, M., Lauer, E., Murphy, J.P., Marshall, D., Mergoum, M., Johnson, J., Holland, J.B., Brown-Guedira, G.: Characterizing the oligogenic architecture of plant growth phenotypes informs genomic selection approaches in a common wheat population. BMC genomics **22**(1), 1–18 (2021)

Powell, O.M., Barbier, F., Voss-Fels, K.P., Beveridge, C., Cooper, M.: 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), 1–9 (2022)

Bellman, R.: Dynamic programming princeton university press princeton. New Jersey Google Scholar, 24–73 (1957)

Ramstein, G.P., Jensen, S.E., Buckler, E.S.: Breaking the curse of dimensionality to identify causal variants in breeding. Theoretical and Applied Genetics **32**, 559–567 (2019)

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., Lee, J.J.: Second-generation plink: rising to the challenge of larger and richer datasets. GigaScience **4**(1) (2015)

Rothenberg, T.J.: Identification in parametric models. Econometrica: Journal of the Econometric Society, 577–591 (1971)

Ray, S., Jarquin, D., Howard, R.: Comparing artificial-intelligence techniques with state-of-the-art parametric prediction models for predicting soybean traits. The Plant Genome (2022)

Pérez, P., Los Campos, G.: Genome-wide regression and prediction with the bglr statistical package. Genetics **198**(2), 483–495 (2014)

VanRaden, P.M.: Efficient methods to compute genomic predictions. Journal of Dairy Science **91**(11), 4414–4423 (2008)

Bernardo, R., Yu, J.: Prospects for genome-wide selection for quantitative traits in maize. Crop Science **47**, 1082–1090 (2007)

Endelman, J.B.: Ridge regression and other kernels for genomic selection with r package rrblup. The

Plant Genome **4**(3), 250–255 (2011)

Endelman, J.B., Jannink, J.-L.: Shrinkage estimation of the realized relationship matrix. G3 Genes—Genomes—Genetics **2**(11), 1405–1413 (2012)

Lipka, A.E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P.J., Gore, M.A., Buckler, E.S., Zhang, Z.: Gapit: genome association and prediction integrated tool. Bioinformatics **28**(18), 2397–2399 (2012)

Wang, X., Yang, Z., Xu, C.: A comparison of genomic selection methods for breeding value prediction. Life Medical Sciences **60**(10), 925–935 (2015)

Habier, D., Fernando, R.L., Dekkers, J.C.M.: The impact of genetic relationship information on genome-assisted breeding values. Genetics **177**(4), 2389–2397 (2007)

Clark, S.A., Werf, J.: Genomic best linear unbiased prediction (gblup) for the estimation of genomic breeding values. Genome-Wide Association Studies and Genomic Prediction, 321–330 (2013)

Gianola, D., Campos, G., Hill, W.G., Manfredi, E., Fernando, R.: Additive genetic variability and the bayesian alphabet. Genetics **1**, 347–363 (2009)

Wang, X., Xua, Y., Huc, Z., Xu, C.: Genomic selection methods for crop improvement: Current status and prospects. The Crop Journal **6** (2018)

Kruschke, J.K.: Tutorial: Doing bayesian data analysis with r and bugs. the Annual Meeting of the Cognitive Science Society **33** (2011)

Kruschke, J.K.: What to believe: Bayesian methods for data analysis. Trends in Cognitive Sciences **14**(7), 293–300 (2010)

John, M., Haselbeck, F., Dass, R., Malisi, C., Ricca, P., Dreischer, C., Schultheiss, S.J., Grimm, D.G.: A comparison of classical and machine learning-based phenotype prediction methods on simulated data and three plant species. Frontiers in Plant Science (2022)

Abdollahi-Arpanahi, R., Gianola, D., Peñagaricano, F.: Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. Genetics Selection Evolution **52**(12) (2020)

Meher, P.K., Rustgi, S., Kumar, A.: Performance of bayesian and blup alphabets for genomic prediction: analysis, comparison and results. the Genetic Society **128**, 519–530 (2022)

Plavšin, I., Gunjača, J., Galić, V., Novoselović, D.: Evaluation of genomic selection methods for wheat

quality traits in biparental populations indicates inclination towards parsimonious solutions. Agronomy **12**(5) (2022)

Gianola, D., Kaam, J.B.C.H.M.: Reproducing kernel hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics **178**(4), 2289–2303 (2008)

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

Howard, R., Carriquiry, A.L., Beavis, W.D.: Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. G3 Genes—Genomes—Genetics **4**(6), 1027–1046 (2014)

Drucker, H., Burges, C.J., Kaufman, L., Smola, A., Vapnik, V.: Support vector regression machines. Advances in neural information processing systems **9** (1996)

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph attention networks. arXiv:1710.10903 (2018)

Belson, W.A.: Matching and prediction on the principle of biological classification. Journal of the Royal Statistical Society. Series C (Applied Statistics) **8**(2), 65–75 (1959)

Liu, Y., Wang, Y., Zhang, J.: New machine learning algorithm: Random forest. Information Computing and Applications, 246–252 (2012)

Rokach, L., Maimon, O.: Decision trees. Data Mining and Knowledge Discovery Handbook, 165–192 (2005)

Ville, B.: Decision trees. WIREs Computational Statistics **5**(6), 448–455 (2013)

Qi, Y.: Random forest for bioinformatics. Ensemble Machine Learning, 307–323 (2012)

González-Camacho, J.M., Ornella, L., Pérez-Rodríguez, P., Gianola, D., Dreisigacker, S., Crossa, J.: Applications of machine learning methods to genomic selection in breeding wheat for rust resistance. The Plant Genome **11**(2) (2018)

Sandhu, K., Patil, S.S., Pumphrey, M., Carter, A.: Multitrait machine- and deep-learning models for genomic selection using spectral information in a wheat breeding program. The Plant Genome (2021)

Li, T., Jiang, S., Fu, R., Wang, X., Cheng, Q., Jiang, S.: Ip4gs: Bringing genomic selection analysis to

breeders. Frontiers in Plant Science **14**, 1131493 (2023)

Yu, T., Zhang, W., Han, J., Li, F., Wang, Z., Cao, C.: An ensemble learning approach for predicting phenotypes from genotypes. In: 2021 20th International Conference on Ubiquitous Computing and Communications (IUCC/CIT/DSCI/SmartCNS), pp. 382–389 (2021). IEEE

An, B., Liang, M., Chang, T., Duan, X., Du, L., Xu, L., Zhang, L., Gao, X., Li, J., Gao, H.: Kcrr: a nonlinear machine learning with a modified genomic similarity matrix improved the genomic prediction efficiency. Briefings in Bioinformatics **22**(6), 132 (2021)

Rossi, A., Tiezzi, M., Dimitri, G.M., Bianchini, M., Maggini, M., Scarselli, F.: Inductive–transductive learning with graph neural networks. In: Artificial Neural Networks in Pattern Recognition: 8th IAPR TC3 Workshop, ANNPR 2018, Siena, Italy, September 19–21, 2018, Proceedings 8, pp. 201–212 (2018). Springer

Brody, S., Alon, U., Yahav, E.: How attentive are graph attention networks? arXiv preprint arXiv:2105.14491 (2021)

Sagi, O., Rokach, L.: Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **8**(4), 1249 (2018)

Shapley, L.S., et al.: A value for n-person games (1953)

Winter, E.: The shapley value. Handbook of game theory with economic applications **3**, 2025–2054 (2002)

Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. Advances in neural information processing systems **30** (2017)

Molnar, C.: Interpretable machine learning (2020)

Ishwaran, H.: The effect of splitting on random forests. Machine learning **99**, 75–118 (2015)

Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International Conference on Machine Learning, pp. 3319–3328 (2017). PMLR

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

Gómez, D., Rojas, A.: An empirical overview of the no free lunch theorem and its effect on real-world machine learning classification. Neural Computation **28**(1), 216–228 (2016)

Merrick, L.F., Carter, A.H.: Comparison of genomic selection models for exploring predictive ability of complex traits in breeding programs. The Plant Genome (2021)

Breiman, L.: Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical science **16**(3), 199–231 (2001)

Dietterich, T.G.: Ensemble methods in machine learning. International Workshop on Multiple Classifier Systems, 1–15 (2000)

Dong, X., Yu, Z., Cao, W., Shi, Y., Ma, Q.: A survey on ensemble learning. Frontiers of Computer Science **14**, 241–258 (2020)

Bertheloot, J., Barbier, F., Boudon, F., Perez-Garcia, M.D., Peron, T., Citerne, S., Dun, E., Beveridge, C., Godin, C., Sakr, S.: Sugar availability suppresses the auxin-induced strigolactone pathway to promote bud outgrowth. New Phytologist **225**(2), 866–879 (2020)

Buckler, E.S., Holland, J.B., Bradbury, P.J., Acharya, C.B., Brown, P.J., Browne, C., Ersoz, E., Flint-Garcia, S., Garcia, A., Glaubitz, J.C., *et al.*: The genetic architecture of maize flowering time. Science **325**(5941), 714–718 (2009)

Wallach, D., Mearns, L.O., Ruane, A.C., Rötter, R.P., Asseng, S.: Lessons from climate modeling on the design and use of ensembles for crop modeling. Climatic Change **139**, 551–564 (2016)

Liang, M., Miao, J., Wang, X., Chang, T., An, B., Duan, X., Xu, L., Gao, X., Zhang, L., Li, J., *et al.*: Application of ensemble learning to genomic selection in chinese simmental beef cattle. Journal of Animal Breeding and Genetics **138**(3), 291–299 (2021)

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