Plant Breeding

WILEY

Plant Breeding

ORIGINAL ARTICLE OPEN ACCESS

Machine Learning for Prediction of Resistance Scores in Wheat (*Triticum aestivum* L.)

Philipp Georg Heilmann[1] | Yohannes Fekadu Difabachew[1] | Matthias Frisch[1] | Anna Luise Moritz[2] | Andreas Stahl[3] | Benjamin Wittkop[2] | Rod J. Snowdon[2] | Michael Koch[4] | Martin Kirchhoff[5] | László Cselényi[6] | Markus Wolf[7,8] | Jutta Förster[8] | Carola Zenke-Philippi[1]

[1]Institute of Agronomy and Plant Breeding II, Justus Liebig University, Gießen, Germany | [2]Institute of Agronomy and Plant Breeding I, Justus Liebig University, Gießen, Germany | [3]Institute for Resistance Research and Stress Tolerance, Julius Kühn Institute, Quedlinburg, Germany | [4]Deutsche Saatveredelung AG, Lippstadt, Germany | [5]Nordsaat Saatzucht GmbH, Langenstein, Germany | [6]W. von Borries-Eckendorf GmbH & Co. KG, Leopoldshöhe, Germany | [7]German Seed Alliance GmbH, Holtsee, Germany | [8]Saaten-Union Biotec GmbH, Leopoldshöhe, Germany

Correspondence: Carola Zenke-Philippi (biometry.popgen@uni-giessen.de)

Received: 7 February 2024 | **Revised:** 12 July 2024 | **Accepted:** 10 October 2024

Funding: This research was supported by the German Federal Ministry of Food and Agriculture, Grant/Award number: FKZ 2818403A18.

Keywords: cross-validation | genomic prediction | machine learning | wheat

ABSTRACT

Machine learning methods were shown to improve the prediction accuracies of genomic prediction of resistance scores compared to methods like RR-BLUP, which were originally designed for metric rather than ordinal response values. We conducted a cross-validation study with 361 wheat genotypes evaluated for five fungal diseases. Our objective was to compare the prediction accuracy and the ability to identify the most resistant genotypes of 19 genomic prediction approaches. Each approach consisted of a different combination of prediction method (RR-BLUP, an alternative method with heterogeneous marker variances, Bayesian generalized linear regression with an ordinal response, support vector machine, gradient boosting machine and random forest), predictor (single SNP markers, LD-based haplotype blocks, 250 variables generated with an autoencoder and SNPs identified with incremental feature selection) and response value (untransformed and logit-transformed resistance scores). In our dataset, RR-BLUP was consistently among the methods with the largest prediction accuracies and the best abilities to identify resistant genotypes in four of five investigated traits. However, in *P. triticina*, using gradient boosting machine and random forest instead of RR-BLUP increased the prediction accuracy from 0.64 to 0.71, indicating that machine learning methods may have an advantage over linear models in genomic prediction. We also found that even though there was a positive correlation between the prediction accuracy and Cohen's κ, a measure to judge how well the most resistant genotypes can be identified, the correlation is not perfect and a large value for the prediction accuracy does not necessarily translate into an equally large κ value.

1 | Introduction

In the last two decades, genomic prediction (Meuwissen, Hayes, and Goddard 2001), which aims at predicting the phenotypic value of an individual from its genotypic data, has increasingly replaced phenotypic selection. The advantage is that only a part of all the genotypes in the breeding population have to be phenotyped or, even better, that phenotypic data that are already

Abbreviations: BGLR, Bayesian generalized linear regression; GBLUP, genomic best linear unbiased prediction; GBM, gradient boosting machine; GWAS, genome-wide association study; LD, linkage disequilibrium; RF, random forest; RMLA, estimation of the error and genetic variance components with restricted maximum likelihood and partitioning according to ANOVA variance components; RMSE, square root of the mean square error; RR-BLUP, ridge-regression best linear unbiased prediction; SNP, single nucleotide polymorphism; SVM, support vector machine; SVR, support vector machine regression.

Philipp Georg Heilmann and Yohannes Fekadu Difabachew contributed equally to this work.

available can be used for the predictions. For the remaining genotypes, only marker data are needed. This is especially beneficial for the evaluation of resistance traits, which is time consuming and expensive. The genomic prediction approach has three components: (1) the form of the genotypic data that are used as predictors, (2) the type of the response values (metric values, percentages and values on an ordinal or nominal scale) and (3) the statistical model that links predictor and response. All of these components have an influence on the prediction accurary, defined as Pearson's correlation between the observed and predicted phenotypic values.

The last component, the statistical model, is the one that receives the most attention in studies on genomic prediction. Ridge regression best linear unbiased prediction (RR-BLUP) and Bayesian methods are among the standard methods for genomic prediction (Wang et al. 2018). While RR-BLUP assumes homogeneous marker variances, most Bayesian methods (Meuwissen, Hayes, and Goddard 2001) as well as another method called "estimation of the error and genetic variance components with restricted maximum likelihood and partitioning according to ANOVA variance components" (RMLA) (Hofheinz and Frisch 2014) allow for heterogeneous marker variances. This might be a better fit for the oligogenic nature of resistance traits because the effects of some markers may be large while those of most others may be close to zero (Hofheinz and Frisch 2014). RR-BLUP, methods from the Bayesian alphabet, and RMLA were developed for the prediction of metric response values with single SNP markers. A possible alternative for the prediction of ordinal response values is Bayesian generalized linear regression as implemented in the R package BGLR (Pérez and de los Campos 2014). More recently, machine learning methods such as support vector machine (SVM), gradient boosting machine (GBM) and random forest (RF) have been used for genomic prediction of resistance scores (Azodi et al. 2019; John et al. 2022; Jones et al. 2023; Ornella et al. 2012; Ornella et al. 2014; Tomar et al. 2021). Machine learning methods are non-parametric and can be applied to metric or ordinal response values without any assumption on the underlying distribution. They also allow to reframe the prediction problem as a classification in which not the observed or predicted resistance scores of a genotype are used as response values but rather its assignment to the "top" or "flop" class (González-Camacho et al. 2018).

Other attempts to improve the prediction accuracy of genomic prediction address the predictors of the model by using haplotype blocks (Difabachew et al. 2023; Weber et al. 2023) or autoencoder features (Islam et al. 2023) as predictors instead of single SNPs or by using subsets of SNPs determined with feature selection (Heinrich et al. 2023; Li et al. 2018). Haplotype blocks group adjacent SNPs on the chromosomes together based on different criteria such as linkage disequilibrium (LD), a fixed number of markers, a fixed physical or genetic distance on the chromosome, or algorithms that aim to create haplotype block libraries that are as representative of the whole set of markers as possible (Pook et al. 2019). When haplotype blocks are used as input variables in RR-BLUP, they are able to capture local epistatic effects (Jiang, Schmidt, and Reif 2018). Autoencoder features are extracted from the encoding layer of an autoencoder. Autoencoders are unsupervised neural networks, in which the input variables are also the targets of the model output (Goodfellow, Bengio, and Courville 2016). In between the input and the output layers is at least one hidden layer with fewer nodes than input variables. These layers function as a bottleneck where the input variables are mapped to a lower dimensional representation (encoding). The number of dimensions can be selected by the user. The model then reconstructs the original input variables from this representation (decoding) in the output layer. To minimize the reconstruction loss, the model learns to preserve as much information of the original variables in the hidden layer as possible (Kramer 1991). Once the "optimal" encoding model is found, the encoded data are used as input variables in a genomic prediction model. In a study in rice, this reduction in dimensionality preserved most of the prediction accuracy while it reduced the computation time considerably (Islam et al. 2023). For feature selection, a genome-wide association study is performed to identify markers that are associated with the trait. The optimum number of markers to be used in the prediction is then determined by cross-validation and the final model is fit accordingly. The results on whether feature selection increases the prediction accuracy compared to the full set of SNPs are contradictory (Heinrich et al. 2023; Li et al. 2018).

Apart from ignoring that the response values are not normally distributed and using RR-BLUP or other methods for metric data anyway, researchers have the option to transform the response so that it better fits the normality assumption. The goal here is to avoid potentially biased results when methods that were originally intended for use with normally distributed data are applied to data on an ordinal scale (Montesinos López et al. 2015). Additionally, when marker effects are estimated with methods like RR-BLUP with an additive model, the additivity of effects can lead to genomic estimates of the genotypic value (GEGVs) that are outside of the original scale, that is, smaller than 0 or larger than 9 on a 0–9 scale. The GEGVs then have no direct translation into meaningful resistance scores. The logit transformation addresses both of these issues. It is intended to achieve a normal distribution of the data (Lesaffre, Rizopoulos, and Tsonaka 2007) and shrinks the score values at both ends of the scale so that GEGVs below or above the limits of the scale are avoided.

We designed this study in order to evaluate the potential of machine learning methods for genomic prediction not only for single SNP markers but also for alternative input features, precisely haplotype blocks and autoencoder features and for subsets of SNP markers determined with feature selection. In order to compare these newer methods with established approaches, we also included Bayesian generalized linear regression and the use of logit-transformed response values. In particular, our objectives were to compare (1) the prediction accuracy of different prediction approaches, including machine-learning methods, and (2) the ability of these approaches to identify the genotypes with the smallest resistance scores with a reference scenario (RR-BLUP with single SNP markers) for the prediction of resistance to five different fungal diseases in a panel of 361 German elite winter wheat lines.

2 | Materials and Methods

2.1 | Phenotypic Data

We evaluated the resistances against *Puccinia triticina* (brown rust), *Fusarium graminearum*, *Septoria tritici*, *Blumeria graminis* (mildew) and *Puccinia striiformis* (yellow rust) of 378

elite wheat lines at three locations in Germany (Böhnshausen, Sachsen-Anhalt; Hovedissen, Nordrhein-Westfalen; Leutewitz, Sachsen) in 2020. Resistances were scored on a 1–9 scale in observation plots in one replication at one (*S. tritici*), two (*F. graminearum*, *P. triticina*), or three locations (*B. graminis*, *P. striiformis*). In case there was more than one location, the arithmetic mean of the two or three observations was used as the resistance score. In order to improve the readability of the manuscript, we use only the name of the disease instead of the full term for the trait, for example, "*S. tritici*" instead of "*S. tritici* resistance score."

2.2 | Genotypic Data

All wheat lines were genotyped with the 25k Illumina iSelect SNP array (SGS TraitGenetics, Gatersleben, Germany). All SNP markers with more than two recorded alleles, more than 10% missing values and an expected heterozygosity of $< 5\%$ as well as all individuals with more than 10% missing marker information were excluded from the analysis. As a result, 16,667 SNP markers and 361 genotypes remained for further analysis. Missing marker data were imputed with BEAGLE (Browning, Zhou, and Browning 2018). We used this dataset for all further calculations. There was no population structure in the dataset (Figure S1).

2.3 | Genomic Prediction Methods

We used genomic prediction based on linear models and machine learning algorithms to evaluate genomic prediction accuracy and efficiency for resistance traits. We used RR-BLUP (Meuwissen, Hayes, and Goddard 2001), RMLA (Hofheinz and Frisch 2014) and Bayesian generalized linear regression (BGLR) with an ordinal response (Pérez and de los Campos 2014). RR-BLUP was technically implemented using a transformation to an animal model (Shen et al. 2013). In order to obtain more robust results in case singular design matrices occur during the cross-validations, we used method 2 of Nazarian and Gezan 2016. The method is available in our software package SelectionTools: https://www.uni-giessen.de/de/fbz/fb09/institute/pflbz2/population-genetics/software. RR-BLUP is considered a standard genomic prediction method in plant and animal breeding programs as it provides stable prediction results (Clark and van der Werf 2013; VanRaden 2008) and is therefore, together with single SNP markers as predictors and resistance scores as response values, treated as the reference in this study.

We also used three supervised machine learning algorithms: support vector regression (SVR)/SVM, GBM and RF. Hyperparameter optimization was performed for all algorithms. SVR is a special case of SVM that is used for metric response values (Drucker et al. 1996). We used a radial basis function as the kernel and tuned the `cost`, the error margin (`margin`) and the influence reach of the individual data points (`sigma`). GBM and RF are both based on ensembles of decision trees (Breiman 2001; Friedman 2001). For GBMs, decision trees are trained in a consecutive order, each tree based on the previous one. For RFs, multiple trees are trained in parallel, each based on a different subset of the training data. The final prediction of the RF model is the average of the predictions of all trees. For both algorithms, we tuned the number of trees used by the model (`ntrees`), the random column sampling rate (`mtry`) and the minimum data points required for a split (`min_n`). We manually set a learning rate of 0.001 for GBM. Default settings were used for all other hyperparameters.

As an alternative, we treated the prediction of resistance scores as a classification task. We used SVM and GBM to predict whether a line was resistant, that is, had a resistance score y smaller than or equal to the 10% quantile Q_{10}, or not. For classification, we used a linear kernel for the SVM and only tuned the `cost` and `margin`. Learning rate for GBM was increased to 0.01 and `min_n` was manually set to 1.

We used a two-step procedure to optimize the hyperparameters for SVR, RF and GBM. The procedure was the same for all algorithms, only the hyperparameters changed (Table 1). We used a 5-fold cross-validation based on the training set to evaluate the hyperparameters. The metric used for evaluation was the square root of the mean square error (RMSE). First, we trained 10 models with hyperparameter combinations based on a maximum entropy grid (Kuhn and Frick 2024; Shewry and Wynn 1987). The essential idea of the maximum entropy grid is to sample points (i.e., combinations of hyperparameters) that cover the hyperparameter space as well as possible, which ensures that the grid search explores a broad range of hyperparameter combinations. Since the points are sampled, they vary between replications. The range of the hyperparameters is shown in Table 1. We used the results of the grid search to initialise an iterative Bayesian optimization, training 10 more models (Snoek, Larochelle, and Adams 2012). Based on the error distribution of the initial maximum entropy grid points, a Bayesian optimization approach can sample and test new combinations from the most promising

TABLE 1 | Overview of hyperparameter ranges considered during tuning.

Hyperparameter	Regression	Classification
RF		
ntrees	(200, 1000)	—
mtry	(0.01, 0.33)	—
min_n	(1, 20)	—
GBM		
ntrees	(50, 500)	(500, 2000)
mtry	(0.01, 0.2)	(0.01, 0.8)
min_n	2, 40)	—
SVR/SVM		
cost	(−10, 5)	(−10, 5)
margin	(0, 0.2)	(0, 0.2)
sigma	(−10, 0)	—

Note: Names of the listed hyperparameters correspond to the argument names used in the software.

regions of the hyperparameter space more quickly. The hyperparameter combination of the model with the smallest RMSE was used to train the final model. The optimization of the hyperparameters for classification was performed analogously, except that some of the parameter ranges in the grid were changed and Cohen's κ was used as the evaluation metric.

2.4 | Feature Engineering

In addition to the complete set of SNP markers, we used three alternative sets of predictors. For the first set, we constructed haplotype blocks based on linkage disequilibrium (LD), which can be measured by r^2 (Zhao et al. 2005). Pairwise LD values were calculated for all SNP markers on each chromosome. SNP markers were added to the left or to the right of a haplotype block as long as the average r^2 between all pairs of SNPs within a block was greater than $t = 0.7$. In order to be able to apply RR-BLUP and RMLA to multi-allelic haplotype block data, the design matrix $\mathbf{Z}$ was re-parameterized (Difabachew et al. 2023).

For the second alternative set of predictors, we extracted the outputs of the encoding layer of an autoencoder. Our autoencoder consisted of five fully connected hidden layers. The layers consisted of [4000,1000, 250,1000, 4000] nodes. The input and output layers consisted of as many nodes as there were predictor variables. The output of the centre layer, consisting of 250 nodes, was treated as the encoding and extracted after model training. We used a rectified linear unit activation function in the hidden layers and applied batch normalization to the outputs of all hidden layers except for the encoding layer. Our data consisted only of homozygous inbred lines with no heterozygous markers present after filtering. Therefore, the markers could be encoded in a binary format, represented by 0 and 1. This allowed for the use of a sigmoid activation function in the output layer. We used binary crossentropy as the loss function and Adam as the optimizer (Kingma and Ba 2015) and trained the autoencoder for 100 epochs.

The third alternative set of predictors was determined by feature selection with a RF model based on GWAS (Heinrich et al. 2023). Analogous to the grid search in the hyperparameter optimization, we conducted a 5-fold cross-valiation on the training set. First, a GWAS was conducted and the markers were ranked according to their p values. Next, RF models were trained, starting with only the most important markers and then incrementing the number of markers in a iterative procedure in steps of 50 from 100 to 1000, of 100 from 1001 to 5000 and of 1000 beyond 5000 markers. The number of markers that resulted in the largest prediction accuracy was determined as the optimum number and the marker set was then used to train another RF model on the complete training set in order to predict the phenotypic values in the validation set. Default settings were used in all RF models. The distribution of the number of SNPs selected by the feature selection procedure is shown in Figure S2.

2.5 | Response Values

For the regression approaches, we used either the resistance scores y or the logit-transformed resistance scores $y^* = \text{logit}\left(\frac{y}{10}\right) = \ln\left(\frac{1}{1-\frac{y}{10}}\right)$ (Lesaffre, Rizopoulos, and Tsonaka 2007) as response values. The division by 10 was necessary because the logit transformation can only be applied to values in the interval (0,1). For the classification methods, the observations y were transformed into two classes: Individuals in the "top" class had a y below or equal to the 10% quantile Q_{10}, and individuals in the "flop" class had a y above the 10% quantile Q_{10}.

2.6 | Prediction Approaches

We define a prediction "approach" as the combination of prediction method, predictor and response values. The name of each approach consists of three elements, divided by a hyphen. The first element is the prediction method: ridge regression BLUP (`RR-BLUP-...`), estimation of the error and genetic variance components with restricted maximum likelihood and partitioning according to ANOVA variance components (`RMLA-...`), Bayesian generalized linear regression with an ordinal response (`BGLR-...`), support vector regression (`SVR-...`), support vector machine (`SVM-...`), gradient boosting machine (`GBM-...`) and random forest (`RF-...`). The predictors can either be SNPs, indicated by `...-SNP-...` as the second element of the approaches, haplotype blocks, indicated by `...-HAP-...`, the autoencoder output, indicated by `...-AEN-...`, or a set of SNP markers determined by feature selection (`...-FS-...`). The last element of each approach is the type of the response value: The use of untransformed values y is indicated by `...-...-0` in the name of the approach, the use of logit-transformed values y^* is indicated by `...-...-1`. Classified values are denoted by `...-...-c`. For example, the approach with the name `SVR-AEN-0` means that a support vector regression was applied on the autoencoder data with the untransformed resistance scores as the response values.

2.7 | Evaluation of the Prediction Approaches

Each prediction approach was evaluated in 200 cross-validation runs. In each of the 200 runs, the dataset was randomly divided into a training set with 289 genotypes (80%) and a validation set with 72 genotypes (20%). The same splits into training and validation set were used for all sets of predictors and algorithms. When predicting ordinal values, the prediction accuracy $r(y, \hat{y})$ was calculated as the correlation between the actual phenotypic values y and the predicted phenotypic values $\hat{y}$ in the validation set. The predicted logit-transformed resistance scores $\hat{y}^*$ were transformed back to $\hat{y}$ and the prediction accuracy was then calculated as $r(y, \hat{y})$.

Cohen's κ (Cohen 1960; Fielding and Bell 1997) as a measure for the agreement between observed and predicted class can be calculated from the confusion matrix for the class assignment (Table 2) as $\kappa = \frac{p_o - p_e}{1 - p_e}$ with $P_0 = \frac{\text{tp}+\text{tn}}{n}$ (the proportion of agreement between the observed and predicted values) and $P_e = \frac{\text{tp}+\text{fn}}{n} \times \frac{\text{tp}+\text{fp}}{n} + \frac{\text{fp}+\text{tn}}{n} \times \frac{\text{fn}+\text{tn}}{n}$ (the expected agreement by random chance) (Montesinos López, Montesinos López, and Crossa 2022). The values for κ range from -1 to 1 where $\kappa = 1$ for perfect agreement and $\kappa \leq 0$ for agreement only by random

chance (González-Camacho et al. 2018). The assignment of the observed values y to the "top" or "flop" class was based on the 10% quantile Q_{10}. Individuals in the "top" class had a y smaller than or equal to Q_{10}, and individuals in the "flop" class had a y greater than Q_{10}. This assignment led to different numbers of individuals in the "top" and "flop" classes for the different diseases (Table 3). To account for the different numbers n_{top}, an individual was assigned to the "top" class of the predictions $\hat{y}$ if its predicted value $\hat{y}$ was among the n_{top} individuals with the smallest $\hat{y}$ values for this disease and to the "flop" class otherwise. For the classification approaches `SVM-SNP-c` and `GBM-SNP-c`, the observed values in the confusion matrix resulted from the assignment of the genotypes to the "top" and "flop" classes by the algorithm. A "good" prediction can mean that (a) the prediction accuracy is high and (b) a prediction approach is able to correctly identify the genotypes with extreme resistance scores, that is, the ones that are most interesting for selection decisions, which would be reflected in a κ value of at least 0.3 to 0.5 (Kuhn and Johnson 2013).

The efficiency of an algorithm was evaluated as the mean of the computation time required for one cross-validation run.

2.8 | Software and Hardware

We used R 4.2.2 (R Core Team 2022) for all calculations except the autoencoders, which where calculated using Python 3.10 (Van Rossum and Drake 2009). The adjusted entry means of the genotypes were estimated using "ASReml-R 4.1.0.110" (Butler et al. 2017). Haplotype blocks were built and RR-BLUP and RMLA were calculated using the R package "SelectionTools 22.1." BGLR was calculated using "BGLR" version 1.1.0 (Pérez and de los Campos 2014). SVR was calculated using the package "kernlab 0.9-30" (Karatzoglou, Smola, and Hornik 2022). For RF, we used "ranger 0.16.0" (Wright and Ziegler 2017). GBMs were trained using "lightgbm 3.3.5" (Shi et al. 2023). Maximum entropy grids were constructed using "dials 1.2.0" (Kuhn and Frick 2024) and Bayesian optimization was based on "tune 1.2.1" (Kuhn 2024). We used "parsnip 1.2.1" (Kuhn and Vaughan 2024) and "tidymodels 1.2.0" (Kuhn and Wickham 2020) as wrapper packages to access all the machine learning-related packages. Autoencoders were built using "tensorflow 2.10.0" (Abadi et al. 2015). Missing marker data were imputed with "BEAGLE 5.4" (Browning, Zhou, and Browning 2018). "plink 1.90b6.12" (Chang et al. 2015; Purcell and Chang 2018) was used for recoding the data into VCF format and conducting the GWAS for incremental feature selection.

All calculations were performed on four Intel Xeon Platinum processors 8276 (28 × 2.20 GHz) with 1 TB DDR4 RAM each and 112 kernels in total. For the ML methods, a maximum of 50 kernels was used at the same time. Due to technical limitations on the package side, it was not possible to run one iteration of SVR or SVM on multiple threads. To keep comparability between machine learning algorithms, we ran 50 instances of SVR at the same time and divided the runtime by 50. This way, 50 cores could be used for training.

TABLE 2 | Confusion matrix for a classification problem with two classes.

		Predicted values		
		Top	**Flop**	**Σ**
Observed values	Top	`tp`	`fn`	`tp + fn`
	Flop	`fp`	`tn`	`fp + tn`
	Σ	`tp + fp`	`fn + tn`	n

Abbreviations: `fn`, number of false negatives; `fp`, number of false positives; n, total number of individuals; `tn`, number of true negatives; `tp`, number of true positives.

3 | Results

3.1 | Phenotypic Values

The observed resistance scores covered only a part of the available range from 1 to 9. *F. graminearum* had the smallest range with observed scores between 3 and 7. The proportion of individuals assigned to the "top" class ranged from 12% in *S. tritici* to 49% in *P. striiformis* (Table 3). An illustration of the distribution of the phenotypic data in one particular validation set can be found in the Supporting Information (Figure S3).

3.2 | Prediction Accuracy of Different Prediction Approaches

All results presented in this section are shown in Figure 1. The overall level of the prediction accuracy was determined by the trait. The reference prediction approach `RR-BLUP-SNP-0`,

TABLE 3 | Summary statistics for the five resistance scores.

Trait	Min	Q_{10}	$Z = Q_{50}$	Q_{90}	Max	n_{top} $(y \leq Q_{10})$	n_{flop} $(y > Q_{10})$
S. tritici	1.00	1.00	2.00	3.00	6.00	42	319
B. graminis	1.00	1.50	2.00	3.50	5.50	119	242
P. triticina	1.00	1.00	2.00	3.75	7.50	66	295
P. striiformis	1.00	1.00	1.33	3.50	6.50	177	184
F. graminearum	3.00	4.00	4.50	5.50	7.00	86	275

Note: The last two columns show how many of the $n = 361$ individuals have a phenotypic value y below/equal to or above the 10% quantile.

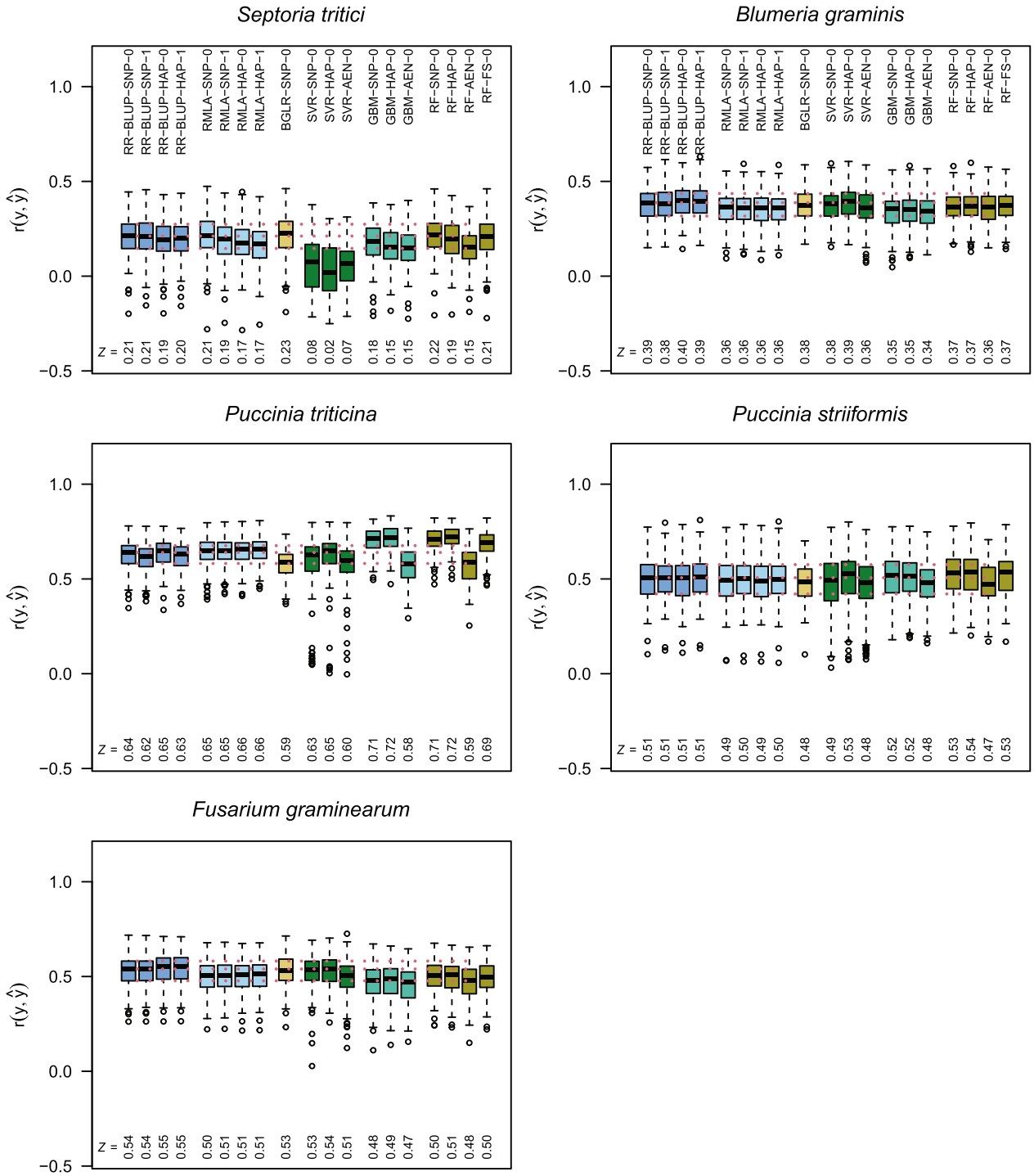

FIGURE 1 | Prediction accuracies for genomic prediction of resistance scores for *S. tritici*, *B. graminis*, *P. triticina*, *P. striiformis* and *F. graminearum* with different prediction approaches. The boxplots show the correlations $r(y,\hat{y})$ between the observed phenotypic values y and the predicted phenotypic values $\hat{y}$ in the validation set for 200 cross-validation runs. Predictions were made with methods ridge regression BLUP (`RR-BLUP-…`), estimation of the error and genetic variance components with restricted maximum likelihood and partitioning according to ANOVA variance components (`RMLA-…`), Bayesian generalized linear regression (`BGLR-…`), support vector regression (`SVR-…`), gradient boosting machine (`GBM-…`) and random forest (`RF-…`). Predictors were either the full set of 16,667 SNP markers (`…-SNP-…`), haplotype blocks based on linkage disquilibrium (`…-HAP-…`), 250 autoencoder features (`…-AEN-…`), or SNP markers identified by feature selection (`…-FS…-`). The response values were either the untransformed resistance scores (`…-...-0`) or the logit-transformed resistance scores (`…-...-1`). Red dotted lines: quartiles from RR-BLUP with 16,667 SNPs (reference). *Z*: median of the correlations $r(y,\hat{y})$ in the 200 cross-validation runs.

RR-BLUP with SNP markers as predictors and the untransformed resistance scores as the response, resulted in medians of $r(y,\hat{y})$ from 0.21 in *S. tritici* to 0.64 in *P. triticina*.

In *S. tritici*, medians of the prediction accuracy ranged from 0.19 to 0.21 in the RR-BLUP approaches and from 0.17 to 0.21 in the RMLA approaches, with the smaller values in the approaches

that used haplotype blocks as predictors. BGLR-SNP-0 had a median of 0.23, the largest value that was observed in this trait. Medians for SVR-SNP-0 and SVR-HAP-0 were 0.08 and 0.02, respectively, while the SVR approach with autoencoder features as predictors, SVR-AEN-0, had a median of 0.07. Medians for the GBM approaches ranged from 0.15 for GBM-HAP-0 to 0.18 for GBM-SNP-0. Medians for the random forest approaches were between 0.15 when autoencoder features were used as predictors (RF-AEN-0) and 0.22 when single SNPs were used instead (RF-SNP-0).

In *B. graminis*, all medians of the correlations $r(y, \hat{y})$ were between 0.34 and 0.40. The largest median, 0.40, was observed with approach RR-BLUP-HAP-0, and the smallest values of 0.34 and 0.35 with the GBM approaches. The medians of the other approaches were in between.

The largest prediction accuracies of all traits were observed in *P. triticina*. The reference approach RR-BLUP-SNP-0 had a median of 0.64, with medians of the other RR-BLUP approaches ranging from 0.62 to 0.65. Medians of the RMLA approaches were 0.65 with untransformed and 0.66 with logit-transformed response values. The median of BGLR-SNP-0 was 0.59. The medians of the SVR approaches ranged from 0.60 for autoencoder features as predictors (SVR-AEN-0) to 0.65 for haplotype blocks (SVR-HAP-0). Medians of the GBM and RF approaches were similar: 0.71 for SNPs as predictors (GBM-SNP-0 and RF-SNP-0), 0.72 for haplotype blocks (GBM-HAP-0 and RF-HAP-0) and 0.58 and 0.59 for autoencoder features (GBM-AEN-0 and RF-AEN-0, respectively). The random forest approach with incremental feature selection (RF-FS-0) was in between with a median of 0.69.

All medians of the prediction accuracies in *P. striiformis* were in the range between 0.47 (for appraoch RF-AEN-0) and 0.53 (approaches SVR-HAP-0, RF-SNP-0 and RF-FS-0). The median of the reference, RR-BLUP-SNP-0, was 0.51 in this case.

In *F. graminearum*, the reference approach RR-BLUP-SNP-0 resulted in a median of the prediction accuracies of 0.54, as did the corresponding approach with haplotype blocks. When logit-transformed response values were used instead, the medians of the prediction accuracies increased to 0.55. RMLA approaches resulted in medians of 0.50 with single SNPs and untransformed response values (RMLA-SNP-0) and 0.51 otherwise. The median of approach BGRL-SNP-0 was 0.53. Among the machine learning methods, the SVR approaches had the largest medians with 0.54 for SVR-HAP-0 and 0.53 for SVR-SNP-0. The smallest medians were observed in the GBM and RF approaches with values of 0.48 for GBM-SNP-0 and RF-AEN-0 and 0.47 for GBM-AEN-0. The remaining RF approaches resulted in medians of 0.50 or 0.51.

3.3 | Identification of the Most Resistant Genotypes

Figure 2 visualizes the results presented in this section. When Cohen's κ was used to evaluate the approaches for how well they were able to identify the most resistant genotypes, the overall level of the κ values was again dependent on the trait.

In *S. tritici*, the reference approach RR-BLUP-SNP-0 had a median of 0.11, as did the corresponding approach with haplotype blocks, RR-BLUP-HAP-0. Using logit-transformed response values led to medians of 0.13 in the RR-BLUP approaches. A similar pattern could be observed in the RMLA approaches, with medians of 0.07 for RMLA-SNP-0 and 0.08 for RMLA-HAP-0 and 0.10 and 0.11 for RMLA-SNP-1 and RMLA-HAP-1, respectively. BGLR-SNP-0 had a median of 0.10. The medians were 0.01 for SVR-SNP-0 and 0.13 and SVR-HAP-0. The use of autoencoder features led to a median of 0.04 in SVR-AEN-0 and the classification approach SVM-SNP-c resulted in a median of 0.02. In the GBM approaches based on regression, the medians ranged between 0.09 for GBM-HAP-0 and 0.13 for GBM-AEN-0. The classification approach GBM-SNP-c had a median of -0.03. The medians of the RF approaches were 0.11 for RF-SNP-0, RF-HAP-0 and RF-FS-0 and 0.10 for RF-AEN-0.

In *B. graminis*, all medians were between 0.21 and 0.23, with the exception of the classification approaches SVM-SNP-c and GBM-SNP-c with median of 0.05 and 0.11, respectively.

In *P. triticina*, the reference RR-BLUP-SNP-0 had a median of 0.16. The medians of the remaining RR-BLUP approaches ranged from 0.15 to 0.18 and showed more variation than the reference. The RMLA approaches resulted in medians of 0.22 for RMLA-SNP-1 and 0.20 for the others. BGRL-SNP-0 had a median of 0.16. The medians of the SVR approaches were either 0.15 or 0.16, with a smaller median of 0.04 for the classification approach SVM-SNP-c. GBM-SNP-0 and GBM-HAP-0 resulted in medians of 0.25. Smaller medians of 0.16 and 0.11 were observed for GBM-AEN-0 and GBM-SNP-c. The pattern for the random forest approaches was similar, with medians of 0.25, 0.28 and 0.15 for approaches RF-SNP-0, RF-HAP-0 and RF-AEN-0, respectively. Approach RF-FS-0 was in between with a median of 0.20.

In *P. striiformis*, the range of the κ values was smaller than for the other traits. The RR-BLUP approaches resulted in medians of 0.30 (for RR-BLUP-SNP-0) to 0.33 (for RR-BLUP-HAP-1). All RMLA approaches had medians of 0.28. The median of BGLR-SNP-0 was 0.31. SVR-HAP-0 had a median of 0.31, compared to medians of 0.28 and 0.25 in the other SVR/SVM approaches. The medians of the GBM approaches were between 0.25 and 0.30. The medians of RF-SNP-0 and RF-HAP-0 were 0.28 and 0.30, respectively, compared to medians of 0.25 for RF-AEN-0 and 0.28 for RF-FS-0.

The overall level of the κ values was highest in *F. graminearum*. All RR-BLUP and RMLA approaches as well as BGLR-SNP-0 had medians of 0.36. The largest median for this trait, 0.38, was observed for approach SVR-SNP-0. The medians of SVR-HAP-0, SVR-AEN-0 and SVM-SNP-c were 0.36, 0.31 and 0.30, respectively. The medians of the GBM regression approaches ranged from 0.26 to 0.31 and the median of the classification approach GBM-SNP-c was 0.24. The medians of the RF approaches ranged from 0.28 to 0.31.

3.4 | Correlation Between r and κ

Figure 3 visualizes the relationship between the prediction accuracy $r(y, \hat{y})$ and Cohen's κ. The means of both measures

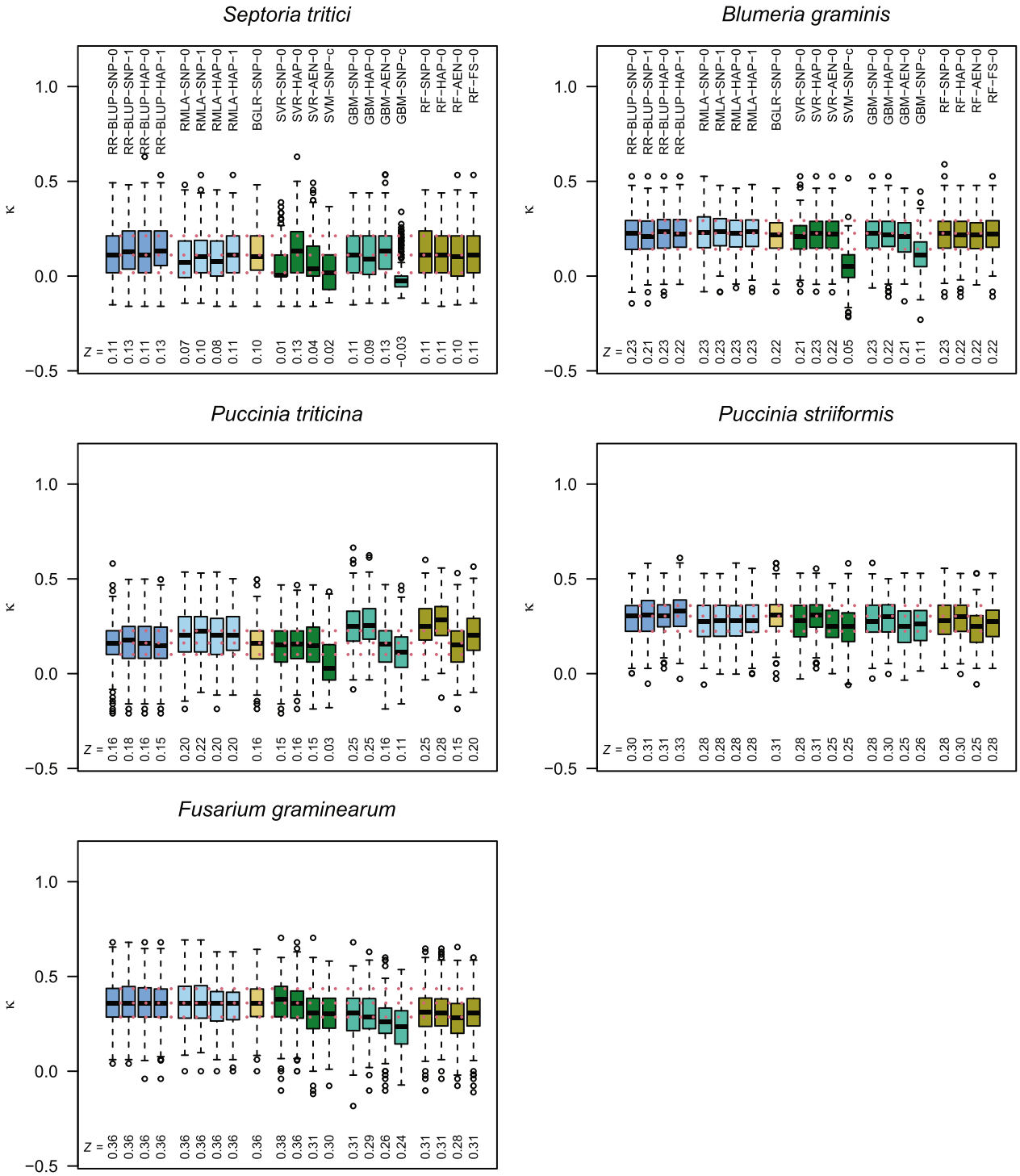

FIGURE 2 | Cohen's κ for genomic prediction of resistance scores for *S. tritici*, *B. graminis*, *P. triticina*, *P. striiformis*, and *F. graminearum* with different prediction approaches. The boxplots show the κ values for the agreement between the assignment to the "top" class (y or $\hat{y}$ equal to or below the 10% quantile Q_{10}) and the "flop" class (y or $\hat{y}$ greater than the 10% quantile Q_{10}) in the validation set for 200 cross-validation runs. Predictions were made with methods ridge regression BLUP (`RR-BLUP-…`), estimation of the error and genetic variance components with restricted maximum likelihood and partitioning according to ANOVA variance components (`RMLA-…`), Bayesian generalized linear regression (`BGLR-…`), support vector regression (`SVR-…`), support vector machine (`SVM-…`), gradient boosting machine (`GBM-…`), and random forest (`RF-…`). Predictors were either the full set of 16,667 SNP markers (`…-SNP-…`), haplotype blocks based on linkage disquilibrium (`…-HAP-…`), 250 autoencoder features (`…-AEN-…`), or SNP markers identified by feature selection (`…-FS…-`). The response values were either the untransformed resistance scores (`…-…-0`), the logit-transformed resistance scores (`…-…-1`), or classifications based on the 10% quantile Q_{10} (`…-…-c`). Red dotted lines: quartiles from RR-BLUP with 16,667 SNPs (reference). Z: median of the κ values in the 200 cross-validation runs.

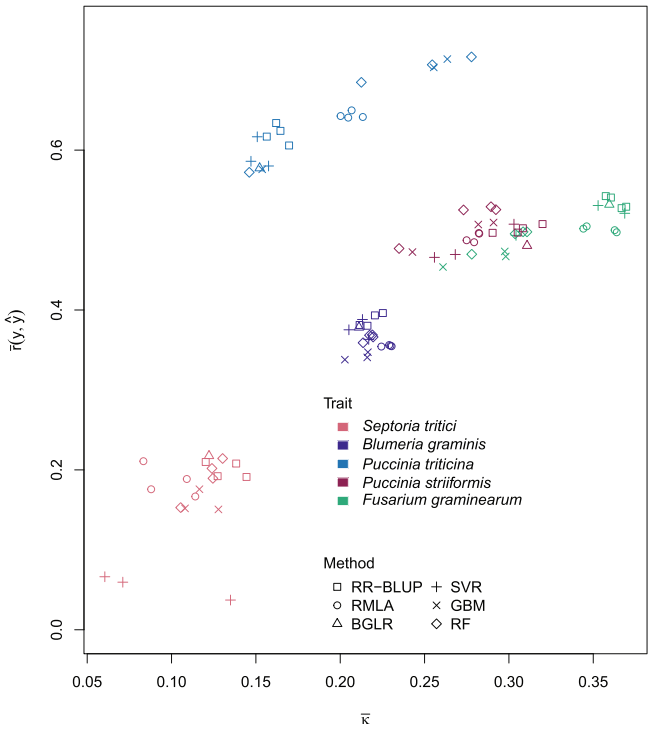

FIGURE 3 | Mean values of correlations $r(y,\hat{y})$ between the observed phenotypic values y and the predicted phenotypic values $\hat{y}$ in the validation set and of Cohen's κ for 200 cross-validation runs. Displayed are the values for the resistance scores for *S. tritici*, *B. graminis*, *P. triticina*, *P. striiformis* and *F. graminearum* for different prediction approaches. Predictions were made with methods ridge regression BLUP (`RR-BLUP`), estimation of the error and genetic variance components with restricted maximum likelihood and partitioning according to ANOVA variance components (`RMLA`), Bayesian generalized linear regression (`BGLR`), support vector regression (`SVR`), gradient boosting machine (`GBM`) and random forest (`RF`). Predictors were either the full set of 16,667 SNP markers, haplotype blocks based on linkage disquilibrium, 250 autoencoder features, or SNP markers identified by feature selection. The phenotypic values used as response values were either the untransformed resistance scores or the logit-transformed resistance scores. Different predictors and response values are not visualized.

showed a correlation across the traits. *P. triticina*, which had the largest mean prediction accuracies of around 0.60, had mean κ values of 0.12 to 0.28. Larger mean κ values of between 0.23 and 0.37 were observed in *P. striiformis* and *F. graminearum* together with smaller mean prediction accuracies of around 0.50. Within the traits, a linear relationship between $r(y,\hat{y})$ and κ could be observed in *P. triticina* and, to a smaller extent, in *P. striiformis* and *F. graminearum*, but not in *S. tritici* and *B. graminis*.

3.5 | Computation Times

The computation times of all approaches can be found in Table 4. RR-BLUP with all 16,667 single SNP markers as predictors was the fastest method with a computation time of 0.68 second per individual cross-validation run on average. However, SVR with autoencoder features was faster when considering the averaged runtime of the parallelization. RMLA and BGLR had a computation time that was about twice as long for the same set of predictors. The computation times of the machine learning methods with single SNP markers were longer with 1.65 min for SVR (averaged), 5.02 min for GBM regression and 3.78 min for RF. When haplotype blocks instead of single SNPs were used, computation times of both RR-BLUP and RMLA increased compared to single SNP markers. An increase of the computation time was also observed for GBM regression, RF and SVR. The use of autoencoder features as predictors in the machine learning methods reduced their computation times to around 1 min or less. The computation time of SVM (averaged) was about 2.87 min and longer compared to that of SVR with single SNP markers. In contrast, the GBM classification took more than twice as long per run as the corresponding regression approach. Computation time was much longer for SVR and SVM compared to other approaches when considering the runtime of the individual cross-validation runs. SVR took 82 min with SNPs and 132 min with haplotype blocks. The autoencoder-based approach (`SVR-AEN-0`) was closer to the other machine learning approaches with 6.36 minutes of individual computation time. SVM took slightly longer than the SNP-based approach `SVR-SNP-0`.

4 | Discussion

4.1 | Prediction Accuracy of Different Prediction Approaches

4.1.1 | Trait

The overall level of the prediction accuracy was determined by the trait (Figure 1). Prediction accuracies between the different approaches varied less for *B. graminis*, *P. striiformis* and *F. graminearum* and more for *S. tritici*, the trait with the smallest overall prediction accuracies with medians around 0.20, and *P. triticina*, the trait with the largest overall prediction accuracies with medians around 0.60 or greater (Figure 1).

The wheat lines in this study are either registered elite varieties or genotypes that are already close to registration. They have therefore been bred for resistance against a variety of pathogens which is reflected in the distribution of the phenotypic values: The observations only cover part of the available scale from 1 to 9 and the larger values, indicating less resistance, are relatively rare (Table 3 and Figure S3). Small prediction accuracies could therefore be at least partially due to the low variation in the response values. In order to obtain reliable results for the genomic predictions, other authors suggest a training set of diverse lines which is continually updated with new breeding material and which can be phenotyped once per season (Juliana et al. 2017).

4.1.2 | Prediction Method

Predictions made with RMLA resulted in similar prediction accuracies as predictions made with RR-BLUP in most cases (Figure 1), even though the genetic architecture of resistance traits is made up of major and minor genes and should, in theory, be captured better by a prediction model like RMLA that allows for heterogeneous marker variances (Hofheinz and Frisch 2014).

TABLE 4 | Computation times in minutes for the different prediction approaches.

Prediction method	Computation time (in minutes)
RR-BLUP-SNP-0	0.68
RR-BLUP-SNP-1	0.68
RR-BLUP-HAP-0	4.97
RR-BLUP-HAP-1	4.74
RMLA-SNP-0	1.45
RMLA-SNP-1	1.45
RMLA-HAP-0	2.10
RMLA-HAP-1	2.10
BGLR-SNP-0	1.33
SVR-SNP-0	1.65 (82.51)
SVR-HAP-0	2.65 (132.53)
SVR-AEN-0	0.13 (6.36)
SVM-SNP-c	2.87 (143.44)
GBM-SNP-0	5.02
GBM-HAP-0	7.2
GBM-AEN-0	0.95
GBM-SNP-c	12.8
RF-SNP-0	3.78
RF-HAP-0	4.89
RF-AEN-0	0.75
RF-FS-0	2.14

Note: The table contains the average values of 200 cross-validation runs for all five traits. For SVR/SVM, due to parallelization, we provide the averaged time per run for 200 cross-validation runs and the time required for a single run in brackets (). Predictions were made with methods ridge regression BLUP (RR-BLUP-…), estimation of the error and genetic variance components with restricted maximum likelihood and partitioning according to ANOVA variance components (RMLA-…), Bayesian generalized linear regression (BGLR-…), support vector regression (SVR-…), support vector machine (SVM-…), gradient boosting machine (GBM-…) and random forest (RF-…). Predictors were either the full set of 16,667 SNP markers (…-SNP-…), haplotype blocks based on linkage disquilibrium (…-HAP-…), 250 autoencoder features (…-AEN-…), or SNP markers identified by feature selection (…-FS…-). The response values were either the untransformed resistance scores (…-...-0), the logit-transformed resistance scores (…-...-1), or classifications based on the 10% quantile Q_{10} (…-...-c).

Other studies on genomic prediction of rust in wheat found that Bayesian methods, which also allow for heterogeneous marker variances, are not necessarily superior to RR-BLUP or genomic BLUP (GBLUP) for the prediction of resistance scores in empirical datasets (Tehseen et al. 2021; Mahmood et al. 2022) even though simulation studies predict that they should be (Meher, Rustgi, and Kumar 2022). A study on both empirical and simulated datasets found the same discrepancy between properties of the methods that should result in better prediction accuracies in theory—and do in simulated datasets—and the actual performance in real-life data (John et al. 2022).

Bayesian generalized linear regression with ordinal response values (approach BGLR-SNP-0) also led to correlations $r(y,\hat{y})$ that were mostly similar to those of RR-BLUP-SNP-0, except for *P. triticina*, in which the values were smaller (Figure 1). This was true regardless of the distribution of the phenotypic values in the validation set. The use of a method specifically designed for ordinal response values therefore did not result in greater prediction accuracies than the use of methods designed for metric response values.

For the machine learning approaches, we did not observe larger prediction accuracies than for the reference approach except for GBM-SNP-0, GBM-HAP-0, RF-SNP-0 and RF-HAP-0 in *P. triticina*. Since we showed in another study that haplotype blocks also led to larger prediction accuracies in this trait compared to single SNPs (Difabachew et al. 2023), we hypothesize that local epistatic effects that can be incorporated by haplotype blocks and machine learning methods, but not by RR-BLUP with single SNPs, may play a role here (Jiang, Schmidt, and Reif 2018; Momen et al. 2018). The prediction accuracies for SVR with single SNP markers (SVR-SNP-0) were generally in the range of those for the corresponding RR-BLUP approach (RR-BLUP-SNP-0), with a difference in the medians of 0.02 at most, except for *S. tritici*. Predictions made with method RF mostly had medians that were 0.01 to 0.04 points greater than those for the corresponding GBM approaches (Figure 1). Only for *P. triticina*, the medians were similar for GBM and RF. Our results partially confirm and partially contradict the results of others. For example, RF resulted in larger prediction accuracies compared to RR-BLUP in the prediction of *P. striiformis* (Tomar et al. 2021) and *F. graminearum* (Rutkoski et al. 2012). In a recent simulation study on genomic prediction with machine learning methods, SVM, RF and GBM showed larger prediction accuracies in a dataset with clear population structure but not in a dataset in which population structure was absent (Jones et al. 2023). The latter corresponds to our dataset (Figure S1), possibly explaining the equal performance of linear and machine learning genomic prediction approaches in four of the five traits in our study. In an extensive study spanning six crops with mostly quantitative traits that compared the prediction accuracy of RR-BLUP, Bayes A and B, Bayesian LASSO, Baysian ridge regression, SVR with linear and nonlinear kernels, gradient tree boosting, artificial neural networks and convolutional neural networks, the results were similar to ours: No single genomic prediction method performed best in all crop/trait combinations, and RR-BLUP was close to the method with the largest prediction accuracy in most cases (Azodi et al. 2019). The same result was found in another study on a simulated animal dataset and three real-life datasets for maize (Lourenço et al. 2024). Our study confirms these findings for resistance traits in wheat.

4.1.3 | Predictor

Replacing single SNP markers with haplotype blocks led to mostly similar prediction accuracies for the corresponding methods, with only small decreases or increases (Figure 1). It has to be noted that there are other possibilities for defining haplotype blocks. In this study, haplotype blocks were built

based on an LD threshold of $r^2 > 0.7$. Other thresholds as well as other methods like building blocks based on a fixed number of markers, fixed window sizes in cM or kilobases on the chromosome, or haplotype block libraries created with the R package HaploBlocker (Pook et al. 2019) are alternative options which have already been investigated in greater detail for this dataset (Difabachew et al. 2023) and others (Weber et al. 2023) and have been shown to increase prediction accuracy in some but not in all cases.

Using autoencoder features as predictors in the machine learning methods resulted in medians of the prediction accuracies that were either similar to or smaller than those of the other approaches, regardless of the method they were used in (Figure 1). Their use led to a reduction in the computation time compared to other predictors for the machine learning methods (Table 4). However, since the computation of the autoencoder features also needs time and the prediction accuracy is generally decreased compared to other predictors, their use as inputs for the machine learning methods was not advantageous in our dataset. More complex studies (Islam et al. 2023) demonstrate the feasibility of preserving prediction accuracy with a reduced set of autoencoder features. We found larger prediction accuracies for GBM and RF than for RR-BLUP with single SNP markers in *P. triticina*, albeit with longer computation times (Table 4). Further research is required to find an easily applicable way to use the autoencoder while maintaining the prediction accuracy and thus save a lot of computation time.

When sets of markers determined by feature selection were used as predictors in a random forest prediction approach (`RF-FS-0`), prediction accuracies were similar to those obtained with the full set of SNPs in nearly all cases (Figure 1), even though the distributions of the numbers of selected SNPs were different between the traits (Figure S2). The findings from other authors in this respect are contradictory: Some found substantial increases with incremental feature selection compared to using the full set of SNPs (Heinrich et al. 2023) while the results of others are similar to ours (Li et al. 2018). We conclude that while feature selection can be beneficial in some cases, further research is needed to determine under which circumstances exactly it can improve the prediction accuracy.

4.1.4 | Response Values

When logit-transformed resistance scores (approaches `RR-BLUP-SNP-1`, `RR-BLUP-HAP-1`, `RMLA-SNP-1` and `RMLA-HAP-1`) were used as response variables instead of the untransformed resistance scores (approaches `RR-BLUP-SNP-0`, `RR-BLUP-HAP-0`, `RMLA-SNP-0` and `RMLA-HAP-0`), differences between the prediction accuracies were small with a maximum of 0.02 points in the medians of the prediction accuracies of the corresponding approaches (Figure 1). We conclude that the logit transformation could successfully address the problem of GEGVs outside the interpretable range and yields predictions with a similar accuracy to those obtained with untransformed data in our dataset. However, it did not improve the predictions by a change in the distribution of the response values. These findings are supported by a study on *P. striiformis* infection in wheat in which the use of logarithmic, boxcox and square root transformations on the observed data did not result in consistent increases in the prediction accuracies obtained with RR-BLUP (Merrick et al. 2022).

4.2 | Identification of the Most Resistant Genotypes

Overall, κ should have a value between 0.3 and 0.5 for acceptable agreement between the classes (Kuhn and Johnson 2013), indicating that an approach is able to identify the most resistant genotypes. We found values in this range only for *F. graminearum*. In the other traits, the κ values were usually smaller.

The patterns for the comparisons between the κ values of the regression approaches in terms of the prediction methods, predictors and response values were the same as for the prediction accuracy (Figure 2). The use of alternative prediction methods, predictors and logit-transformed response values led to medians of the κ values that were either smaller than or similar to the reference approach `RR-BLUP-SNP-0`. The only exception was *P. triticina*, with an increase for GBM and RF from a median of the κ values of 0.16 for `RR-BLUP-SNP-0` to 0.25 for `GBM-SNP-0`, `GBM-HAP-0` and `RF-SNP-0` and 0.28 for `RF-HAP-0`. Autoencoder features as predictors led to smaller κ values in most cases in comparison to `RR-BLUP-SNP-0` (Figure 2). We could not confirm the superiority of SVM for the identification of superior genotypes that was found in 16 wheat datasets (Ornella et al. 2014).

In most studies on genomic prediction, only the prediction accuracy $r(y,\hat{y})$ is reported. However, while a large value for the prediction accuracy indicates that the predictions are accurate on average, this is different from the correct identification of the most resistant genotypes, which are the ones that are interesting for selection. Ideally, a prediction approach would yield both large κ values as well as have a large prediction accuracy. We found a positive correlation between the means of the prediction accuracy $r(y,\hat{y})$ and the means of κ across the traits (Figure 3). Apart from the smaller range of the κ values, these findings are mostly similar to those for rust resistance in wheat (Ornella et al. 2014; González-Camacho et al. 2018). However, both measures must be considered together when the suitability of a method identify superior genotypes is evaluated: In *P. triticina*, the prediction accuracies were largest for all traits, with mean values around 0.6, while the means of the κ values were between 0.12 and 0.28. In contrast, the mean prediction accuracies in *F. graminearum* were around 0.5, but the means of the κ values were all greater than 0.25 (Figure 3). Our findings show that even if κ and $r(y,\hat{y})$ are positively correlated, a large prediction accuracy does not automatically translate into a κ value that is sufficient for the selection of superior genotypes.

4.3 | Summary

A good genomic prediction model is supposed to extract the relevant information from the genotypic data while simultaneously dealing with the noise which comes from other factors. Linear models like RR-BLUP make simplifying assumptions in this situation, particularly when they include only additive effects, like

in our study. The questions then become if there are additional patterns in the genotypic data that cannot be captured by linear models and if machine learning methods are able to find these patterns. In our dataset, RR-BLUP was consistently among the methods with the largest prediction accuracies and the best abilities to identify resistant genotypes in four of the five investigated traits. Compared to machine learning methods, RR-BLUP is implemented in most genomic prediction software. It is easy to apply without the need for hyperparameter tuning and consequently very fast. Additionally, the resulting marker effects are easy to interpret and understand. However, we found substantial increases in the prediction accuracies and κ values compared to the reference approach `RR-BLUP-SNP-0` in *P. triticina*, indicating that investing the additional effort to fine-tune such a method may be worth it. We also found that even though there was a positive correlation between the prediction accuracy and Cohen's κ, a measure to judge how well the most resistant genotypes can be identified, the correlation is not perfect and a large value for the prediction accuracy does not necessarily translate into an equally large κ value. This shows that the prediction accuracy should not be the only measure that is used to select a "good" genomic prediction method.

Author Contributions

Matthias Frisch, Rod Snowdon and Andreas Stahl conceived the study. Michael Koch, Martin Kirchhoff, László Cselényi, Markus Wolf and Jutta Förster collected the field data and genotypic data. Anna Moritz, Andreas Stahl, Benjamin Wittkop and Matthias Frisch evaluated the field data. Yohannes Difabachew carried out the genomic predictions with RR-BLUP, RMLA and BGLR. Philipp Heilmann carried out the genomic predictions with SVR/SVM, GBM and RF. Philipp Heilmann, Yohannes Difabachew and Carola Zenke-Philippi wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgments

The project was funded by the Federal Ministry of Food and Agriculture (BMEL) based on a decision of the Parliament of the Federal Republic of Germany via the Federal Office for Agriculture and Food (BLE) under the innovation support programme (FKZ 2818403A18). Open Access funding enabled and organized by Projekt DEAL.

Conflicts of Interest

Michael Koch is employed by Deutsche Saatveredelung AG. Martin Kirchhoff was employed by Nordsaat Saatzucht GmbH and is employed by Nordzucker AG. László Cselényi is employed by W. von Borries-Eckendorf GmbH & Co. KG. Markus Wolf is employed by German Seed Alliance GmbH. Jutta Förster is employed by Saaten-Union Biotec GmbH. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Data Availability Statement

The genotypic and phenotypic data as well as the scripts used for this study can be downloaded from https://github.com/czp-jlu/resistance.

References

Abadi, M., A. Agarwal, P. Barham, et al. 2015. "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems." https://www.tensorflow.org/. Software available from tensorflow.org.

Azodi, C. B., E. Bolger, A. McCarren, M. Roantree, G. de los Campos, and S.-H. Shiu. 2019. "Benchmarking Parametric and Machine Learning Models for Genomic Prediction of Complex Traits." *G3: Genes, Genomes, Genetics* 9, no. 11: 3691–3702.

Breiman, L. 2001. "Random forests." *Machine Learning* 45: 5–32.

Browning, B. L., Y. Zhou, and S. R. Browning. 2018. "A One-Penny Imputed Genome From Next Generation Reference Panels." *American Journal of Human Genetics* 103: 338–348.

Butler, D. G., B. R. Cullis, A. R. Gilmour, B. G. Gogel, and R. Thompson. 2017. *ASReml-R Reference Manual Version 4*. Hemel Hempstead, HP1 1ES, UK: VSN International Ltd. https://asreml.kb.vsni.co.uk/knowledge-base/asreml_r_documentation/.

Chang, C. C., C. C. Chow, LCAM Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee. 2015. "Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets." *GigaScience* 4, no. 1: s13742–015.

Clark, S. A., and J. van der Werf. 2013. "Genomic Best Linear Unbiased Prediction (gBLUP) for the Estimation of Genomic Breeding Values." edited by C. Gondro, J. van der Werf, and B. Hayes, *Genome-Wide Association Studies and Genomic Prediction*. Totowa, NJ: Humana Press, pp. 321–330.

Cohen, J. 1960. "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement* 20, no. 1: 37–46.

Difabachew, Y. F., M. Frisch, A. L. Langstroff, et al. 2023. "Genomic Prediction With Haplotype Blocks in Wheat." *Frontiers in Plant Science* 14: 1168547.

Drucker, H., C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik. 1996. "Support Vector Regression Machines." edited by M. C. Mozer, M. Jordan, and T. Petsche, *Advances in Neural Information Processing Systems*, Vol. 9. MIT Press, pp. 155–161. https://proceedings.neurips.cc/paper_files/paper/1996/file/d38901788c533e8286cb6400b40b386d-Paper.pdf.

Fielding, A. H., and J. F. Bell. 1997. "A Review of Methods for the Assessment of Prediction Errors in Conservation Presence/Absence Models." *Environmental Conservation* 24, no. 1: 38–49.

Friedman, J. H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics* 29: 1189–1232.

González-Camacho, J. M., L. Ornella, P. Pérez-Rodríguez, D. Gianola, S. Dreisigacker, and J. Crossa. 2018. "Applications of Machine Learning Methods to Genomic Selection in Breeding Wheat for Rust Resistance." *Plant Genome* 11, no. 2: 170104.

Goodfellow, I., Y. Bengio, and A. Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Heinrich, F., T. M. Lange, M. Kircher, F. Ramzan, A. O. Schmitt, and M. Gültas. 2023. "Exploring the Potential of Incremental Feature Selection to Improve Genomic Prediction Accuracy." *Genetics Selection Evolution* 55, no. 1: 78.

Hofheinz, N., and M. Frisch. 2014. "Heteroscedastic Ridge Regression Approaches for Genome-Wide Prediction With a Focus on Computational Efficiency and Accurate Effect Estimation." *G3: Genes, Genomes, Genetics* 4, no. 3: 539–546.

Islam, T., C. Kim, H. Iwata, H. Shimono, and A. Kimura. 2023. "DeepCGP: A Deep Learning Method to Compress Genome-Wide Polymorphisms for Predicting Phenotype of Rice." *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 20, no. 3: 2078–2088.

Jiang, Y., R. H. Schmidt, and J. C. Reif. 2018. "Haplotype-Based Genome-Wide Prediction Models Exploit Local Epistatic Interactions Among Markers." *G3: Genes, Genomes, Genetics* 8, no. 5: 1687–1699.

John, M., F. Haselbeck, R. Dass, et al. 2022. "A Comparison of Classical and Machine Learning-Based Phenotype Prediction Methods on Simulated Data and Three Plant Species." *Frontiers in Plant Science* 13: 932512.

Jones, D., R. Fornarelli, M. Derbyshire, M. Gibberd, K. Barker, and J. Hane. 2023. "The Pursuit of Genetic Gain in Agricultural Crops Through the Application of Machine-Learning to Genomic Prediction." *Frontiers in Genetics* 14: 1186782.

Juliana, P., R. P. Singh, P. K. Singh, et al. 2017. "Genomic and Pedigree-Based Prediction for Leaf, Stem, and Stripe Rust Resistance in Wheat." *Theoretical and Applied Genetics* 130: 1415–1430.

Karatzoglou, A., A. Smola, and K. Hornik. 2022. "kernlab: Kernel-Based Machine Learning Lab." https://CRAN.R-project.org/package%3Dkernlab. R package version 0.9-30.

Kingma, D. P., and J. Ba. 2015. "Adam: A Method for Stochastic Optimization." In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* edited by Y. Bengio, and Y. LeCun. https://arxiv.org/abs/1412.6980.

Kramer, M. A. 1991. "Nonlinear Principal Component Analysis Using Autoassociative Neural Networks." *AIChE Journal* 37, no. 2: 233–243.

Kuhn, M. 2024. "tune: Tidy Tuning Tools." https://CRAN.R-project.org/package%3Dtune. R package version 1.2.1.

Kuhn, M., and H. Frick. 2024. "dials: Tools for Creating Tuning Parameter Values." https://CRAN.R-project.org/package%3Ddials. R package version 1.2.1.

Kuhn, M., and K. Johnson. 2013. *Applied Predictive Modeling*. New York, NY: Springer.

Kuhn, M., and D. Vaughan. 2024. "parsnip: A Common API to Modeling and Analysis Functions." https://CRAN.R-project.org/package%3Dparsnip. R package version 1.2.1.

Kuhn, M., and H. Wickham. 2020. "tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles." https://www.tidymodels.org.

Lesaffre, E., D. Rizopoulos, and R. Tsonaka. 2007. "The Logistic Transform for Bounded Outcome Scores." *Biostatistics* 8, no. 1: 72–85.

Li, B., N. Zhang, Y.-G. Wang, A. W. George, A. Reverter, and Y. Li. 2018. "Genomic Prediction of Breeding Values Using a Subset of SNPs Identified by Three Machine Learning Methods." *Frontiers in Genetics* 9: 237.

Lourenço, V. M., J. O. Ogutu, R. A. P. Rodrigues, A. Posekany, and H.-P. Piepho. 2024. "Genomic Prediction Using Machine Learning: A Comparison of the Performance of Regularized Regression, Ensemble, Instance-Based and Deep Learning Methods on Synthetic and Empirical Data." *BMC Genomics* 25, no. 1: 152.

Mahmood, Z., M. Ali, J. I. Mirza, et al. 2022. "Genome-Wide Association and Genomic Prediction for Stripe Rust Resistance in Synthetic-Derived Wheats." *Frontiers in Plant Science* 13: 788593.

Meher, P. K., S. Rustgi, and A. Kumar. 2022. "Performance of Bayesian and BLUP Alphabets for Genomic Prediction: Analysis, Comparison and Results." *Heredity* 128, no. 6: 519–530.

Merrick, L. F., D. N. Lozada, X. Chen, and A. H. Carter. 2022. "Classification and Regression Models for Genomic Selection of Skewed Phenotypes: A Case for Disease Resistance in Winter Wheat (Triticum aestivum L.)."*Frontiers in Genetics* 13: 835781.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. "Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps." *Genetics* 157: 1819–1829.

Momen, M., A. A. Mehrgardi, A. Sheikhi, et al. 2018. "Predictive ability of Genome-Assisted Statistical Models Under Various Forms of Gene Action." *Scientific Reports* 8: 12309.

Montesinos López, O. A., A. Montesinos López, and J. Crossa. 2022. *Multivariate Statistical Machine Learning Methods for Genomic Prediction*. Cham: Springer.

Montesinos López, O. A., A. Montesinos López, P. Pérez-Rodríguez, G. de los Campos, K. Eskridge, and J. Crossa. 2015. "Threshold Models for Genome-Enabled Prediction or Ordinal Categorical Traits in Plant Breeding." *G3: Genes, Genomes, Genetics* 5: 291–300.

Nazarian, A., and S. A. Gezan. 2016. "GenoMatrix: A Software Package for Pedigree-Based and Genomic Prediction Analyses on Complex Traits." *Journal of Heredity* 107, no. 4: 372–379.

Ornella, L., P. Pérez, E. Tapia, et al. 2014. "Genomic-Enabled Prediction With Classification Algorithms." *Heredity* 112: 616–626.

Ornella, L., S. Singh, P. Perez, et al. 2012. "Genomic Prediction of Genetic Values for Resistance to Wheat Rusts." *The Plant Genome* 5: 136–148.

Pérez, P., and G. de los Campos. 2014. "Genome-Wide Regression and Prediction With the BGLR Statistical Package." *Genetics* 198, no. 2: 483–495.

Pook, T., M. Schlather, G. de Los Campos, M. Mayer, C. C. Schön, and H. Simianer. 2019. "HaploBlocker: Creation of Subgroup-Specific Haplotype Blocks and Libraries." *Genetics* 212, no. 4: 1045–1061.

Purcell, S., and C. Chang. 2018. "Plink v1.90b6.12." https://www.cog-genomics.org/plink/1.9/.

R Core Team. 2022. "R: A Language and Environment for Statistical Computing." Vienna, Austria. https://www.R-project.org.

Rutkoski, J., J. Benson, Y. Jia, G. Brown-Guedira, J.-L. Jannink, and M. Sorrells. 2012. "Evaluation of Genomic Prediction Methods for Fusarium Head Blight Resistance in Wheat." *Plant Genome* 5: 51–61.

Shen, X., M. Alam, F. Fikse, and L. Rönnegård. 2013. "A Novel Generalized Ridge Regression Method for Quantitative Genetics." *Genetics* 193, no. 4: 1255–1268.

Shewry, M. C., and H. P. Wynn. 1987. "Maximum Entropy Sampling." *Journal of Applied Statistics* 14, no. 2: 165–170.

Shi, Y., G. Ke, D. Soukhavong, et al. 2023. "lightgbm: Light Gradient Boosting Machine." https://CRAN.R-project.org/package%3Dlightgbm. R package version 3.3.5.

Snoek, J., H. Larochelle, and R. P. Adams. 2012. "Practical Bayesian Optimization of Machine Learning Algorithms." edited by F. Pereira, C. J. Burges, L. Bottou, and K. Q. Weinberger, *Advances in Neural Information Processing Systems*, Vol. 25. Curran Associates, Inc., pp. 2951–2959.

Tehseen, M. M., Z. Kehel, C. P. Sansaloni, et al. 2021. "Comparison of Genomic Prediction Methods for Yellow, Stem, and Leaf Rust Resistance in Wheat Landraces From Afghanistan." *Plants* 10: 558.

Tomar, V., G. S. Dhillon, D. Singh, et al. 2021. "Evaluations of Genomic Prediction and Identification of New Loci for Resistance to Stripe Rust Disease in Wheat (Triticum aestivum L.)."*Frontiers in Genetics* 12: 710485.

Van Rossum, G., and F. L. Drake. 2009. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace. https://api.semanticscholar.org/CorpusID:61259041.

VanRaden, P. M. 2008. "Efficient Methods to Compute Genomic Predictions." *Journal of Dairy Science* 91, no. 11: 4414–4423.

Wang, X., Y. Xu, Z. Hu, and C. Xu. 2018. "Genomic Selection Methods for Crop Improvement: Current Status and Prospects." *Crop Journal* 6, no. 4: 330–340.

Weber, S. E., M. Frisch, R. J. Snowdon, and K. P. Voss-Fels. 2023. "Haplotype Blocks for Genomic Prediction: A Comparative Evaluation in Multiple Crop Datasets." *Frontiers in Plant Science* 14: 1217589.

Wright, M. N., and A. Ziegler. 2017. "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R." *Journal of Statistical Software* 77, no. 1: 1–17.

Zhao, H., D. Nettleton, M. Soller, and J. C. M. Dekkers. 2005. "Evaluation of Linkage Disequilibrium Measures Between Multi-Allelic Markers as Predictors of Linkage Disequilibrium Between Markers and QTL." *Genetics Research* 86, no. 1: 77–87.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.