Multi-Trait Machine and Deep Learning Models for Genomic Selection using Spectral Information in a Wheat Breeding Program

Karansher S. Sandhu[1], Shruti S. Patil[2], Michael O. Pumphrey[1], Arron H. Carter[1*]

[1] Department of Crop and Soil Sciences, Washington State University, Pullman, WA, USA, 99164

[2] School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA, USA, 99164

[*] **Correspondence:**

Dr. Arron H. Carter

ahcarter@wsu.edu

Abbreviations

ARI: anthocyanin reflectance index

CNN: convolutional neural network

GBLUP: genomic best linear unbiased predictor

GEBVs: genomic estimated breeding values

GNDVI: green normalized difference vegetation index

GS: genomic selection

MLP: multilayer perceptron

MT: multi-trait

NCPI: normalized chlorophyll pigment ratio index

NDVI: normalized difference vegetation index

NWI: normalized water index

PRI: photochemical reflectance index

RF: random forest

SVM: support vector machine

UT: uni-trait

Abstract

Prediction of breeding values and phenotypes is central to plant breeding and has been revolutionized by the adoption of genomic selection (GS). Use of machine and deep learning algorithms applied to complex traits in plants can improve prediction accuracies in the context of GS. Spectral reflectance indices further provide information about various physiological parameters previously undetectable in plants. This research explores the potential of multi-trait (MT) machine and deep learning models for predicting grain yield and grain protein content in wheat using spectral information in GS models. This study compares the performance of four machine and deep learning -based uni-trait (UT) and MT models with traditional GBLUP and Bayesian models. The dataset consisted of 650 recombinant inbred lines from a spring wheat breeding program, grown for three years (2014-2016), and spectral data were collected at heading and grain filling stages. MT-GS models performed 0-28.5% and -0.04-15% superior to the UT-GS models for predicting grain yield and grain protein content. Random forest and multilayer perceptron were the best performing machine and deep learning models to predict both traits. These two models performed similarly under UT and MT-GS models. Four explored Bayesian models gave similar accuracies, which were less than machine and deep learning-based models, and required increased computational time. Green normalized difference vegetation index best predicted grain protein content in seven out of the nine MT-GS models. Overall, this study concluded that machine and deep learning-based MT-GS models increased prediction accuracy and should be employed in large-scale breeding programs.

Core Ideas

1. Potential for combining high throughput phenotyping, machine and deep learning in breeding.
2. Multi-trait models exploit information from secondary correlated traits efficiently.
3. Spectral information improves genomic selection models.
4. Deep learning can aid plant breeders owing to increased data generated in breeding programs

Introduction

Quantitative genetics theory was proposed by Sir Ronald Fisher a century ago and established the infinitesimal model (Fisher, 1918). This theory was developed without the direct use of genotypic data and persisted for decades. With the advent of sufficient genome-wide markers paired with an infinitesimal model of quantitative genetics theory, Meuwissen et al. (2001) were the first to propose the term genomic selection (GS) for predicting breeding values in animals and plants. GS uses estimates of marker effects based on model development from a related population genotyped and phenotyped for the trait of interest. The genome-wide marker effects are then used for predicting the genomic estimated breeding values (GEBVs) of a new population, which is only genotyped (Heffner et al., 2010). GS has been extensively researched in plant breeding, focusing on optimizing marker density, training population size, family relatedness, heritability of the trait, and GS model utilized (Shengqiang et al., 2009). Most of the GS models evaluated in plant breeding have been uni-trait (UT), where single traits are predicted (Lozada & Carter, 2019; Sun et al., 2019). Recently, breeders have moved to adopt multi-trait (MT) GS models because predictions are required for multiple traits simultaneously and their combination may improve prediction accuracies (Jia & Jannink, 2012).

MT-GS models leverage shared genetic information between correlated traits (Jia & Jannink, 2012) to predict various traits simultaneously by utilizing the same set of predictors, assuming the presence of some structure in the captured output (Bhatta et al., 2020). Improved prediction accuracy of MT-GS models over UT-GS is attributed to the correlation among the training population and between the traits. Furthermore, MT-GS models are more interpretable and require less computational time than a series of UT-GS models (Montesinos-López et al., 2018a). MT models are intensively employed in other fields such as data mining, forest management, energy forecasting, and ecological modelling (Voyant et al., 2017). Jia & Jannink (2012) showed that prediction accuracy improved for primary traits with low heritability in barley (*Hordeum vulgare* L.) when a secondary correlated trait is used in MT-GS models. Sun et al. (2019) demonstrated the increase in GS prediction accuracy for wheat (*Triticum aestivum* L.) grain yield when secondary traits such as normalized difference vegetation index (NDVI) and canopy temperature were included in MT-GS models. Bhatta et al. (2020) compared UT- and MT-GS models for predicting end-use quality traits in barley and concluded that MT-GS models have better performance under both within and across environment predictions. Mixed model approaches utilizing Bayesian and genomic best linear unbiased predictor (GBLUP) are most commonly used in plant breeding programs (Endelman, 2011; Pérez & Campos, 2014).

GBLUP is a frequently used MT-GS models in plant breeding, which uses marker-based relationship matrix for predicting the performances of genotypes (Endelman, 2011; VanRaden, 2008). Several Bayesian MT-GS models (Bayes A, Bayes Lasso, Bayes B, and Bayes Cpi) are also available, which assume a prior distribution during the training process, and hence separate models are required to optimize different traits (Pérez & Campos, 2014). Bayes Lasso follows the double exponential prior distribution for performing the continuous shrinkage and variable

selection (Tishbirani, 1996). Additionally, Bayes Lasso applies a long tail student t distribution to the marker effects. Bayes A, Bayes B, and Bayes Cpi use the scaled-t, gaussian mixture (point mass at zero with gaussian distribution), and scaled-t mixture (point mass at zero with scaled-t distribution) prior distribution, respectively, during the model training (Pérez & Campos, 2014). These models assume that all markers do not contribute to the total genetic variance. Bayesian models are also computationally intensive due to Monte Carlo Markov Chain utilization for estimating the marker effects. Bayesian models are known as parametric due to the assumption of the prior relationship among features and predictors, which do not model gene by gene and higher-order interactions during the estimation of marker effects (Gianola et al., 2006; Montesinos-López et al., 2019). Hence, recently developed machine and deep learning tools provide an opportunity for the selection of the GS model.

The increasing adoption of high throughput genotyping and phenotyping tools by plant breeders has increased data generation tremendously, which requires the adoption of analytical methods used in other disciplines for complex datasets. Machine and deep learning models have been explored in previous studies for prediction in UT-GS models and have demonstrated mixed results (Bellot et al., 2018; Ma et al., 2017). Random forest (RF) and support vector machine (SVM) are commonly used ensemble machine learning models for the GS context. RF is an ensemble machine learning tool used for predicting the output by averaging the results of an extensive collection of identically distributed decision trees applied to bootstrapped samples of the training data. RF is better than the other tree-based methods like decision tree regression and bagging, as it tries to reduce the correlation between the subsets of tree by averaging results from those trees for final predictions. The reduction of correlation among the independent trees and averaging performance of the trees aids in reducing overfitting and increasing the prediction

accuracy. The important hyperparameters for RF model training include the number of trees, the number of features sampled for each iteration, the importance of each feature, and the depth of the trees (Hastie et al., 2009). SVM provides flexibility for fitting the model as it fits the best regression line by allowing a specific acceptable error in the model. Optimization of SVM models requires finding a regression line that deviates from the real line by no greater than a value called maximum error (ϵ), and at the same time the line should be flat as possible (Smola & Scholkopf, 2004).

Deep learning is the branch of machine learning that uses an artificial neural network as a prediction tool, and needs to be explored in GS owing to the plethora of data accumulated in breeding programs (Lecun et al., 2015; Samuel, 2001). Deep learning models explore the relationship between input and output variables using a combination of neurons and hidden layers to form a network similar to the biological network of neurons in the human brain. Deep learning models use different non-linear activation functions with a large number of layers, and data is transformed along with each layer to obtain the best fit for different genetic architectures (Angermueller et al., 2016). Often used deep learning models in plant breeding are multilayer perceptron (MLP) and convolutional neural network (CNN) (Pérez-Enciso & Zingaretti, 2019). Optimization of hyperparameters is required to achieve the best deep learning model performance, which is the most computationally intensive step (McKay, 1992). The most essential hyperparameters include the type of activation function, activation rate, regularization parameter, number of epochs, number of hidden layers, dropout, and stopping criteria (Pérez-Enciso & Zingaretti, 2019). Hyperparameters can be selected by using one of the four methods, namely grid search, latin hypercube sampling, random search, and optimization (McKay, 1992).

MLP is considered a feed-forward neural network, and consists of three input, hidden, and output layers. The first layer is the input layer, which receives the DNA marker information. Each neuron of the hidden layers has its characteristic weight and transforms the previous layer's data using various linear and non-linear activation functions (Lecun et al., 2015). The weight parameters and other hyperparameters are optimized using the training data using either backpropagation or stochastic gradient (Cho & Hegde, 2019). The number of the output layer is equal to the total number of response variables, depending upon the tested model. The output of a layer also depends upon the weighted average transformation of neurons from the previous layer with associated bias. CNN is a special type of neural network used for input features having a specific pattern, such as linkage disequilibrium among markers distributed along a linear chromosome. The hidden layer in MLP is replaced with multiple layers in CNN, such as convolutional, pooling, fully connected, and dense layers (Lecun et al., 2015). As opposed to neurons, CNN uses the kernels or filters in convolution layers for capturing the hidden information. A filter consists of predefined marker interval windows having the same weights. This filter is moved continuously across the input data for obtaining the weight for each window for computing the locally weighted sum. The pooling layer follows the first convolutional layer and is used for dimensionality reduction (Pérez-Enciso & Zingaretti, 2019). This layer merges the output of filters from the convolutional layer using either mean, minimum, or maximum to smoothen the results. Dropout and activation functions are employed after the convolutional and pooling layer (Pook et al., 2020).

High-throughput phenotyping applications include spectral reflectance values obtained from plants to provide information about various physiological processes and have been used in wheat (Babar et al., 2006), rice (*Oryza sativa* L.; Zheng et al., 2018) maize (*Zea mays* L.; Aguate

et al., 2017), barley (Barmeier et al., 2017), and sorghum (*Sorghum bicolor* L.; Habyarimana et al., 2020). Different vegetation indices or spectral reflectance indices (SRI) can be extracted by measuring reflection from plants. These SRI aid in the indirect selection of primary traits (grain yield or grain protein content) in wheat due to their moderate to high genetic correlation and high heritability compared to primary traits (Crain et al., 2018). Commonly utilized indices are NDVI, photochemical reflectance index (PRI), normalized water index (NWI), and green-NDVI (GNDVI), which provide information about plant biomass, photochemical pigments, plant water stress, and nitrogen status (Gitelson et al., 1996; Peñuelas et al., 1994). These indices have been used in covariate and MT-GS models to predict grain yield in wheat and demonstrate improvement in prediction accuracy (Lozada & Carter, 2019; Sun et al., 2019). To the best of our knowledge, deep learning models have not been explored for MT-GS in wheat for predicting grain yield and grain protein content using spectral information as secondary traits.

SRI derived from wheat has been reported to correlate to grain yield, biomass, and drought tolerance in spring wheat (Gizaw et al., 2018). Grain yield and grain protein content are important selection traits in spring wheat breeding programs and are complicated by the negative correlation between them. However, GS and SRI provide an alternative for selecting these two traits simultaneously. Our previous study observed that inclusion of secondary correlated traits results in improved prediction accuracy for grain yield and grain protein content by using the rrBLUP GS model. Grain protein content and grain yield were accurately predicted when spectral data was collected at heading and grain filling stages, respectively (Sandhu et al., 2021b). Similarly, we observed that deep learning based GS models improve prediction accuracy by 3-5% in different agronomic traits in wheat (Sandhu et al., 2021a). The main objectives of this study were to 1) Optimize different MT machine and deep models for predicting grain yield

and grain protein content in wheat, 2) Compare the performance of MT-GS and UT-GS models, and 3) Compare the performances of MT-GS mixed models with machine and deep models under cross-validation and independent validation scenarios.

Materials and Methods

Field data and plant material: The data set used in this study consisted of 650 recombinant inbred lines from a nested association mapping population of spring wheat (Blake et al., 2019). The field trial was planted for three years (2014-2016) at Spillman Agronomy Farm, Pullman, WA. For detailed information about the population, field trials, and traits evaluated see Sandhu et al. (2021a,b). In brief, field trials were planted in a modified augmented design with 15-20% of the plots assigned to three replicated check cultivars. Grain protein content (%) and grain yield (t/ha) was collected using a Perten DA 700 NIR analyzer (Perkin Elmer, Sweden) and Wintersteiger Nursery Master combine (Ried im Innkreis, Austria).

Spectral reflectance at 16 different bands between the 430 and 980 nm wavelengths was collected with a handheld CROPSCAN multi-spectral radiometer at heading (Feekes growth stages 10.1) and grain filling stages (Feekes growth stages 11.1) (Large, 1954). Data from the CROPSCAN was processed with CROPSCAN MSR software, and six different SRI were derived. These indices were NDVI, PRI, NWI, anthocyanin reflectance index (ARI), normalized chlorophyll pigment ratio index (NCPI), and GNDVI (Peñuelas et al., 1994; Prasad et al., 2007; Rouse et al., 1972). Detailed information about these SRI and the physiological traits they explain is provided in **Supplementary Table 1**.

Genotyping: The population was genotyped using genotyping by sequencing and Illumina 90K SNP chip array (Poland & Rife, 2012). Initial genotyping data consisted of 73,345 high-quality polymorphic markers anchored to the Chinese Spring RefSeqv1 reference map (Jordan et al., 2018). Detailed procedure about genotyping, SNP calling, map construction, and filtration is described in previous publications (Sandhu et al., 2021a). Quality filtering involved removing monomorphic markers and markers missing more than 20% of the genotyping data. RILs missing phenotyping data in one environment and 10% genotyping data were discarded. Finally, a minor allele of < 0.05 was used, resulting in 635 RILs having 40,038 polymorphic markers.

Statistical analysis: Adjusted means for grain yield, grain protein content, and SRI were obtained using residuals derived separately for each environment using the 'lme4' function implemented in the R program using the model:

$$Y_{ij} = Block_i + Check_j + residuals_{ij}$$

Where Y_{ij} is the phenotypic values of the trait, $Check_j$ is the fixed effect of jth replicated check cultivars, and $Block_i$ is the fixed effect of the ith block. Residuals were used for obtaining the adjusted means for all the evaluated phenotypic traits (Bates et al., 2015).

Broad sense heritability was extracted using augmented randomized complete block design implemented in R (Aravind et al., 2020) with the model:

$$Y_{ij} = \mu + Block_i + Check_j + Gen_{j(i)} + e_{ij}$$

Where Y_{ij}, $Block_i$, and $Check_j$ are defined above, $Gen_{j(i)}$ is the random effect of the unreplicated genotype nested within the ith block and follows the distribution $Gen_j \sim N(0, \sigma^2_g)$, and e_{ij} is the standard normal error distributed as $e_{ij} \sim N(0, \sigma^2_e)$.

Genetic correlation between primary traits (grain yield or grain protein content) and secondary traits (individual SRI) are calculated using multivariate models, represented as

$$\begin{bmatrix} y_A \\ y_B \end{bmatrix} = \begin{bmatrix} X_A & 0 \\ 0 & X_B \end{bmatrix} \begin{bmatrix} b_A \\ b_B \end{bmatrix} + \begin{bmatrix} Z_A & 0 \\ 0 & Z_B \end{bmatrix} \begin{bmatrix} g_A \\ g_B \end{bmatrix} + \begin{bmatrix} \epsilon_A \\ \epsilon_B \end{bmatrix}$$

Where y_A and y_B are the BLUPs for the primary (A) and secondary (B) traits, X and Z denote the design matrix for the fixed and random effect, and b is the fixed effects, g is the random genetic effects, and e is the residuals for each trait. Variance components were calculated assuming $\begin{bmatrix} g_A \\ g_B \end{bmatrix} \sim N(0, H \otimes G)$, where G is the genomic relationship matrix, H is the genetic variance-covariance matrix, and $\begin{bmatrix} \epsilon_A \\ \epsilon_B \end{bmatrix} \sim N(0, I \otimes R)$, where R is the residual variance-covariance matrix, and I is the identity matrix. The genetic correlation was obtained as

$$r_G = \frac{cov(A, B)}{\sqrt{var(A) \cdot var(B)}}$$

Where $Var(A)$, and $Var(B)$ represents the genetic variance of the primary and secondary traits individually and $cov(A, B)$ is the covariance between primary and secondary traits. The complete multivariate analysis was performed using a multivariate approach in JMP genomics (SAS Insitute Inc. 2011).

Genomic prediction models:

Genomic best linear unbiased predictor (GBLUP): The UT-GS GBLUP is defined as

$$y = \mu + Zg + e$$

where y is the phenotype of interest (grain yield or grain protein content), μ is the overall mean, Z is the design matrix linking genotypes to the breeding values, g is the vector of genomic breeding values, and e is the random residuals. It is assumed that $g \sim N(0, G\sigma^2_g)$, where G is genomic relationship matrix, σ^2_g is the additive genetic variation, and $e \sim N(0, I\sigma^2_e)$ with σ^2_e as residual variance and I is the identity matrix. The MT-GS is represented as

$$\begin{bmatrix} y_A \\ y_B \end{bmatrix} = \begin{bmatrix} X_A & 0 \\ 0 & X_B \end{bmatrix} \begin{bmatrix} b_A \\ b_B \end{bmatrix} + \begin{bmatrix} Z_A & 0 \\ 0 & Z_B \end{bmatrix} \begin{bmatrix} g_A \\ g_B \end{bmatrix} + \begin{bmatrix} \epsilon_A \\ \epsilon_B \end{bmatrix}$$

Where y_A and y_B represent the primary and secondary traits, X and Z are the design matrix associating the fixed and random effects, b is the vector of means for primary and secondary traits, g and e are the vector for random genetic and residual effects. It is assumed as $\begin{bmatrix} g_A \\ g_B \end{bmatrix} \sim N(0, H \otimes G)$, where G is the genomic relationship matrix, H is the variance-covariance matrix for the two traits, and $\begin{bmatrix} \epsilon_A \\ \epsilon_B \end{bmatrix} \sim N(0, I \otimes R)$, R is the residual variance-covariance matrix between two traits (Endelman, 2011; VanRaden, 2008). In all the MT-GS models, individual SRI were included as secondary correlated traits.

Bayesian models: As GBLUP uses the relationship matrix for estimating genotypes effects, in this study we also explored the Bayes Lasso, Bayes A, Bayes B, and Bayes Cpi in UT- and MT-GS models, which consider different prior distributions. The UT-GS model is represented as

$$y_i = \mu + \sum_{j=1}^{j=p} x_{ij}\beta_j + \epsilon_i$$

where y_i is the phenotype of interest (grain yield or grain protein content), μ is the overall mean, β_j is the jth marker effect, x_{ij} is the value of the jth SNP in the ith individual, and ϵ_i is residual

error. The conditional prior distribution was separate for each of the Bayesian models employed in this study. For Bayes A, the prior distribution is $\epsilon_i \sim N(0, \sigma^2)$ with $\sigma^2 \sim \chi^{-2}(\sigma^2|df, S)$ for residuals, and $\beta_j \sim \chi^{-2}(df_\beta, S_\beta)$ for genotypic values. Initial values for the degrees of freedom for the t distribution was set to four and S was calculated as $S = var(y) * 0.4$ as suggested by Pérez & Campos (2014). Analysis was performed using BGLR and MTM packages (Campos & Grüneberg, 2016) with 20,000 Monte Carlo Markov Chain iterations, and 5,000 burn in iterations.

The MTM package was used for fitting MT Bayesian models estimating unstructured variance-covariance between traits. The model is represented as

$$y = \mu + Zu + \epsilon$$

Where y is the vector of primary and secondary phenotypic traits, μ is the mean vector for all traits, u is the predicted genotypic values for all traits with distribution as $u \sim N(0, H \otimes G)$, where G is relationship matrix, H is the variance-covariance matrix, and ϵ is vector of residuals and distributed as $\epsilon \sim N(0, I \otimes R)$, where I is the identity matrix and R is the variance-covariance matrix for the residuals (Montesinos-López et al., 2016).

Random forests (RF): In RF bootstrap sampling, a subset of features was selected randomly as predictors for splitting the tree nodes. Each tree is chosen for lowering the loss function in the final prediction (Smith et al., 2013). The RF model can be represented as

$$\hat{y}_i = \frac{1}{B}\sum_{b=1}^{B} T_b(x_i)$$

289 Where $\hat{y}_i$ is the predicted value of the individual with genotype x_i, B is the number of bootstrap
290 samples, and T is the total number of trees. RF is computationally less intensive, as each tree is
291 independent of each other and can be computed on different units or nodes (Waldmann, 2016).
292 The working of the RF can be grouped into four main steps:

293 1. Bootstrap sampling is used to select the individual plant i (y_i, x_i) with replacement. The
294 sampled individual can appear several times or not, mainly bootstrap sampled $b = (1, \ldots,$
295 $B)$.
296 2. Selection is performed for the number of features (max features) or input variables at
297 random $(SNP_j, j = (1, \ldots, J)$, and the best set of features are selected that minimize the
298 loss function obtained as MSE.
299 3. Splitting is performed at each node into two new subsets (child nodes) for the genotype
300 of SNP_j.
301 4. Steps 2 and 3 are repeated for each node until a minimum node or the specified max
302 depth is reached. The final predicted value of an individual of genotype x_i is the average
303 of the values predicted by the decision trees in the forest.

304 The important hyperparameters for RF model training include the number of trees, the
305 number of features sampled for each iteration, the importance of each feature, and the depth of
306 the trees (Hastie et al., 2009). We used randomized and grid search cross-validation for selecting
307 the best hyperparameter's combination. The combination used for grid search cross-validation
308 after the randomized search was the number of trees (200, 300, 500, 1000), max features (auto,
309 sqrt), and max depth (40, 60, 80, 100). The random forest regressor and Scikit learn libraries
310 were used for analysis in Python 3.7 (Gulli and Pal, 2017).
311

Support vector machine (SVM): SVM uses kernel functions for mapping the input space into a high dimensional feature space using the relationship between phenotypes and marker genotype (Smola & Scholkopf, 2004). The relationship between the phenotype and genotype is given as

$$f(x) = wx + b$$

Where *w* is the unknown weight and *b* is the constant, reflecting the maximum allowed bias. The learning of function $f(x)$ is performed by minimizing the loss function

$$C\sum_{i=1}^{n} L(e_i) + \frac{1}{2}\|w\|^2$$

Where e_i = y - $f(x)$ is the associated error with the i^{th} training data point, $\|w\|^2$ represents model complexity, *C* is a positive regularization parameter controlling the tradeoff between training error and model complexity, and *L* is the loss function (Vapnik, 2013). Herein, we selected the ϵ insensitive loss function for *L* and is represented as

$$L_\epsilon(e) = \begin{cases} 0 \; if |e| < \epsilon \\ |e| - \epsilon \; otherwise \end{cases}$$

L (loss function) is zero if the absolute error is less than predefined ϵ, while if absolute error is greater than ϵ, *L* is the difference between absolute error and ϵ. Usually, ϵ insensitive loss function is represented in term of slack variables ($\xi\xi^*$). The resulting optimization equation can be represented as

$$min_{\,w,\,b,\,\xi,\,\xi^*}\left(C\sum_{i=1}^{n} L(\xi\xi^*) + \frac{1}{2}\|w\|^2\right)$$

The solution to this minimization problem is of the form

$$f(x) = \sum_{i=1}^{n} \alpha_i x_i x_j + b$$

Where $x_i x_j$ is the inner product of linear function. It was replaced with a non-linear kernel function, namely the gaussian radial basis function. The parameters *C*, ξ, and kernel are optimized using cross-validation in the Scikit library (Pedregosa et al., 2011).

Multilayer perceptron (MLP): MLP is a feed-forward neural network consisting of three main layers, namely input, hidden, and output layers. A detailed representation of the MLP used is represented in **Figure 1A**, where the output of a layer depends upon the weighted average transformation of neurons from the previous layer with associated bias. The output of a hidden layer is represented as

$$Z_i = b_{(i-1)} + W_i f_{(i-1)}(x)$$

Where Z_i is the output from the i^{th} hidden layer, W_i is the weight associated with the neurons, $f_{(i-1)}$ represents the activation function linking the associated weights and bias from the previous layer, and this process is repeated until the output layer. In the case of UT-GS models, the output layer is a vector of GEBVs, and in MT-GS, it contains two vectors having GEBVs and spectral information.

Hyperparameters were optimized for the MLP models using inner grid search cross-validation and the Keras function's internal capabilities. The grid search cross-validation used 80% of the training data, where 80% of this dataset is used for optimizing the hyperparameters and the remaining 20% for validation using Keras independent split validation functions (Cho & Hegde, 2019). The hyperparameters that provided the least MSE on the validation set were selected, and later used on the testing set. Detailed information about the hyperparameter optimization is referred to in a previous publication (Sandhu et al., 2021a). All the MLP

algorithms were implemented in Python 3.7 using Keras and Scikit learn libraries (Gulli and Pal, 2017).

Convolutional neural network (CNN): CNN is a special case of neural network used for input features having a specific pattern. The complete layout for CNN is provided in **Figure 1B**. Initial values for the hyperparameters were based on previous findings on UT-GS models (Sandhu et al., 2021a). The CNN model used here consists of one input layer, two convolutional layers, two pooling layers, a dense layer, a flatten layer, two dropouts, and an output layer. Grid search cross-validation was used for selecting hyperparameters, namely, filters (16, 32, 64, 128), learning rate (0.01, 0.05, 0.1), activation function (logistic, linear, tanh, relu), batch size (64, 128), epochs (150, 200), and solver (adam, sgd, lbfgs). These hyperparameters were selected based on previous findings and other studies (Waldmann et al., 2020). Early stopping, dropout, and regularization techniques were applied to control model overfitting. Early stopping involves stopping the training process as validation error reaches a minimum, using Keras-provided API (Callbacks) (Pedregosa et al., 2011). Dropout involves assigning a fixed set of training neurons with a weight to zero for controlling the overfitting and reducing complexity. We used a dropout rate of 0.2 during hyperparameter optimization in MLP and CNN based on Srivastava et al. (2014).

Cross-validation and independent prediction: The performances of all nine UT- and MT-GS models were evaluated using five-fold cross-validation. During five-fold cross-validation, 80% of the data was used for model training, and the remaining 20% for model testing within each environment. Two hundred replications were used to assess the model's performance, and the

mean was reported as prediction accuracy. Each replication consisted of five model iterations, where the testing set was rotated for each iteration. Prediction accuracy was obtained as the Pearson correlation coefficient between actual (observed) phenotypic value and the calculated GEBVs. Instant accuracy was reported, which involves the average correlation coefficient of iterations. Comparisons were made between UT- and MT-GS models where a single SRI was included in the MT-GS model. Similarly, machine and deep learning-based MT-GS models were compared with their Bayesian and GBLUP counterparts. MT-GS models used six SRI individually in the model, and the best performing SRI was identified for each trait with different models.

Independent predictions were performed by training models on previous year(s) data and predicting the phenotype in future years. We tested the performance of both UT- and MT-GS with the inclusion of spectral information. In brief, the GS model trained on 2014 and 2015 data was used to predict the 2016 and similarly the model trained on 2014 data to predict the 2015 environment. The GS analysis was computationally intensive, and this problem was resolved by working on Washington State Universities high speed computing cluster (https://hpc.wsu.edu/).

Results

Phenotypic summary and heritability: Average phenotypic values and heritabilities are provided for grain yield and grain protein content under three environments (**Table 1**). Grain yield and grain protein content had low and moderate heritability. The six SRI used in this study had moderate to high heritability (**Table 2**), and 2015 had the lowest heritability for phenotypic and spectral traits (**Table 1 and 2**). Phenotypic and genetic correlation between phenotypic traits and SRI was obtained at both heading and grain filling stages (**Table 3, Supplementary Table 2**

and 3). Grain protein content and grain yield had a high and significant correlation with SRI collected at heading and grain filling stage (**Table 3 and Supplementary Table 2**).

Genomic selection using uni- and multi-trait models: We evaluated nine different models (GBLUP, Bayes A, Bayes B, Bayes Cpi, Bayes Lasso, RF, SVM, MLP, and CNN) for predicting grain protein content and grain yield using five-fold cross-validation (**Figure 2 and 3**). A single SRI was included in each model for predicting both traits under MT-GS models and their average results are depicted for comparison with UT-GS models (**Figure 2 and 3**). In the case of grain yield, MT-GS models either gave an equal or higher prediction accuracy than UT-GS models (**Supplementary Table 4**). Furthermore, machine and deep learning models performed better than the traditional GBLUP and Bayesian models under UT and MT models. The improvement in prediction accuracy with MT-GS models for grain yield varied from 0 to 28.5%, with maximum improvement observed in the 2014 environment, and the lowest increase was observed for the 2015 environment (**Figure 2**). RF and MLP performed best for predicting grain yield under all the environments, closely followed by GBLUP (**Supplementary Table 4 and Figure 2**). Both models performed superior for UT- and MT-GS models compared to other machine and deep learning models. Four bayesian models, namely, Bayes A, Bayes B, Bayes Cpi, and Bayes Lasso, produced almost the same prediction accuracy for grain yield under the UT- and MT-GS models (**Supplementary Table 4 and Figure 2**). SVM resulted in the lowest prediction accuracy for grain yield and CNN observed the lowest increase under the MT-GS models (**Figure 2 and Supplementary Table 4**).

Similarly, MT-GS models increased prediction accuracy for grain protein content compared to UT-GS counterparts (**Figure 3 and Supplementary Table 4**). MLP and RF

418 performed superior to other models and gave similar accuracy under different scenarios (**Figure**
419 **3 and Supplementary Table 4**). The performance of MT-GS models varied from -0.04 to 15%
420 compared to the UT-GS models. Similar to grain yield, SVM performed poorest for UT- and
421 MT-GS models to predict grain protein content. We observed similar trends in improvement in
422 prediction accuracy with MT-GS models for all environments compared to grain yield, with
423 2014 observing the largest increase (**Figure 2 and 3**). The maximum prediction accuracy for
424 grain protein content was 0.62 in 2014 with RF, 0.56 in 2015 with MLP, and 0.61 in 2016 with
425 RF MT-GS models (**Figure 3**). We observed that similar to grain yield, MT-GS model for CNN
426 resulted in the lowest increase in prediction accuracy.

427

428 **Performances of multi-trait genomic selection models using individual SRI:** We evaluated
429 each SRI's relationship with grain protein content and grain yield across all nine MT-GS models
430 assessed in this study (**Figure 4 and 5**). Inclusion of any SRI in the MT-GS model resulted in
431 higher prediction accuracy than the UT-GS model for grain protein content and grain yield.
432 **Figure 4** shows the prediction accuracies for grain protein content for the three environments
433 with six individual SRIs in the MT-GS models. GNDVI was the best performing index for seven
434 out of the nine models in each environment (**Figure 4**). RF and MLP resulted in the greatest
435 improvement in prediction accuracy by including GNDVI in the MT-GS model. SVM was the
436 only model where GNDVI performed worse than other indices for most environments (**Figure**
437 **4**). The maximum prediction accuracy for grain protein content in 2014, 2015, and 2016 was
438 0.67, 0.61, and 0.68, respectively, with RF by inclusion of GNDVI in the MT-GS model.

439 Similarly, the inclusion of individual SRI in the MT-GS models increased prediction accuracy
440 compared to UT-GS models for grain yield (**Figure 5**). Unlike grain protein content, there was

441 no individual SRI that performed best across all environments and models. The maximum
442 prediction accuracy for 2014 was 0.72 with inclusion of NWI, 2015 was 0.40 with inclusion of
443 NDVI, and 2016 was 0.64, with inclusion of NDVI. Overall, we observed that NWI, NDVI, and
444 GNDVI provide the necessary information for increasing prediction accuracy for grain yield
445 (**Figure 5**).

446

447 **Independent predictions with uni- and multi-trait genomic selection models:** Independent
448 predictions involve prediction across environments where models were trained on previous year
449 datasets, and predictions were made for upcoming years. We used four UT- and MT-GS models,
450 namely, GBLUP, RF, MLP, and CNN, for the independent predictions as they performed
451 consistently better with cross-validation for both traits and under all the environments. SVM was
452 excluded because of its poor performance for both traits. All the Bayesian models performed
453 similarly, but had less accuracy, so were excluded from independent predictions because of huge
454 computational time limitations. The machine and deep learning models were approximately four
455 times faster than the Bayesian models. **Figure 6** shows the results for independent prediction for
456 both traits using UT- and MT-GS models.

457 There was an increase of 17% and 11% in prediction accuracies for grain yield and grain protein
458 content with MT-GS over the UT-GS models (**Supplementary Table 5**). RF and MLP
459 performed consistently better under UT- and MT-GS models to predict grain yield and grain
460 protein content, closely followed by CNN. The highest average prediction for grain yield and
461 grain protein content was 0.29 and 0.41 with MLP and RF, compared to 0.20 and 0.34 with the
462 traditional UT GBLUP model (**Supplementary Table 5**). There was a varied increment in

463 prediction accuracies with different MT-GS models for grain yield and grain protein content, as
464 depicted in **Figure 6**.

465

466 **Discussion:**

467 Grain yield and grain protein content are highly important target traits in wheat breeding
468 programs and for other cereal grains. The generally negative correlation between them, along
469 with lower heritability, creates a problem in efficiently selecting both the traits simultaneously.
470 GS and high throughput phenotyping have the potential for reducing the challenges associated
471 with selection for these two traits. GS may increase selection efficiency, reduce the generation
472 advancement time, and increase selection intensity (Heffner et al., 2010). Similarly, spectral
473 information from phenomics data allows for indirect selection by using SRI as a proxy indicator.
474 The increased use of machine and deep learning models in other disciplines has prompted its use
475 in plant breeding. Several MT selection studies have conducted and demonstrated the potential
476 for increasing prediction accuracy (Bhatta et al., 2020; Jia & Jannink, 2012). This study
477 evaluated the potential of MT machine and deep learning models for predicting grain yield and
478 grain protein content in wheat using spectral information as a secondary trait. The spectral
479 information acts as a proxy indicator for selection, is correlated with the primary trait of interest,
480 and has higher heritability. We observed that machine and deep learning models, namely RF and
481 MLP, resulted in an increased prediction accuracy for both traits when spectral information was
482 included in MT-GS models.

483

484 **Trait characterization and association**

The greatest advantage of MT-GS models is borrowing the information from secondary traits to increase prediction accuracies for the primary trait (Calus & Veerkamp, 2011). Understanding the genetic architecture of each traits is the first consideration when developing MT models. The primary traits of interest in this study were grain yield and grain protein content, which were evaluated for three years and heritabilities were low to moderate (**Table 1**), suggesting a considerable influence of non-genetic effects. Lower heritability for grain yield and grain protein content was expected, as these traits are controlled by a large number of small effect QTLs and are genetically complex. Similar results were obtained in previous studies for grain yield and grain protein content (Sun et al., 2017). Secondary traits (SRI used in this study) have high heritability and correlate positively with the primary traits (**Table 2**). Furthermore, the collection of SRI is easy and could be performed using high-throughput techniques (Sankaran et al., 2015). This suggested that their inclusion in MT-GS models may improve prediction accuracy, increase selection intensity, and lead to faster breeding cycles (Crain et al., 2018).

GNDVI was the best performing index for seven out of nine models evaluated in this study, and could be a useful proxy index for selecting grain protein content in breeding programs. Association of GNDVI with nitrogen status and translocation is related to the increase in prediction accuracies for grain protein content in the MT-GS models (Gitelson et al., 1996). The high correlation of primary traits with SRI indicates a direct connection between them. Grain protein content has a lower heritability value than GNDVI, and hence a lesser amount of variation is accounted by GS models for grain protein content under the UT-GS model. The higher accuracy observed in MT-GS models can be attributed to capturing more genetic variation, as GNDVI is genetically correlated with grain protein content. Grain protein content had the highest correlation with GNDVI, which measures reflection from the green region (550

nm) of the plant vegetation spectrum and provides information about the plant's nitrogen status (Gitelson et al., 1996). Another advantage of using GNDVI is that it allows the prediction of grain protein content earlier in the selection pipeline, saving the time, cost, and effort to harvest and collect data from a large number of field plots. Using GNDVI also aids in improving MT-GS models, which can be used early in the selection pipeline to select improved lines for advancement in the breeding program.

Grain yield is a complex trait resulting from a myriad of interactions, including nutrient status, water availability, biotic and abiotic stress. Three SRI, namely NWI, NDVI, and GNDVI, resulted in the highest prediction accuracy for grain yield under the MT-GS models. These indices each measure a part of NIR (900-970 nm), and this spectrum determines the water status and biomass of the plants, suggesting NIR is useful for predicting grain yield, especially in the Pacific Northwest US, where wheat is grown under dryland conditions. Identifying multiple SRI that increase prediction accuracy provided insight that model analysis should not rely on a single SRI in each year. Expression of several physiological processes in plants is dependent upon the plants genetic makeup, factors like light, temperature, humidity, day length, etc, and the growth stage when the plant experiences stress. Different SRI are able to capture these various genotype by environment interactions, along with environmental variation that may exist from year to year, which contribute to final grain yield estimations. The inclusion of multiple SRI in the model helps explain the unknown variance component ignored in UT-GS models. We were able to identify the three SRI that are more influential for predicting grain yield out of the six SRI explored. Using these three SRI will reduce computation time and cost for data management to make selections.

Potential of the machine and deep learning models in a breeding program

This study aimed to explore machine and deep learning potential in wheat breeding programs using MT-GS models. Deep learning is a new machine learning branch using a dense network of neurons to explore the dataset's complicated hidden relationship. We concluded that MT random forest and multilayer perceptron resulted in an improvement of 23-31% prediction accuracy for both traits under cross-validation and independent prediction compared to the GBLUP and Bayesian models. Similar results were obtained by Ma et al. (2018) for predicting grain yield, plant height, and grain length in wheat using UT-deep learning models and rrBLUP. Their study demonstrated potential for the utilization of deep learning models in plant breeding. Deep learning based GS models gave 0-5% higher prediction accuracies for various agronomic traits in wheat in our previous work (Sandhu et al., 2021a). Montesinos-López et al. (2018b) concluded that deep learning models were superior for six out of nine traits evaluated in maize and wheat over the traditional GBLUP. Additionally, Montesinos-López et al. (2018a) showed that MT deep learning models performed superior to the MT Bayesian models when genotype by environment interaction was not included for predicting grain yield in wheat. These results open up a new path for improved breeding selection, that could translate into higher rates of genetic gain.

Machine and deep learning models, unlike Bayesian models, are highly flexible for mapping complex interactions present between predictors and responses, and thereby interpreting the trend of the current dataset (Liu et al., 2019). Bayesian models often include selection of a specific subset of markers which explain the most variation in the response, in contrast to machine and deep learning models, that explore the whole data space during model training (Pérez & Campos, 2014). As grain yield and grain protein content are polygenic, relying

554 on Bayesian models might not be a good strategy, as it is often unknown how many QTLs are
555 present in a particular population and are expressing under the specific environment. Hence,
556 machine and deep learning models are suitable for complex traits as they explore all possible
557 relationships between markers and traits. Machine and deep learning models also account for
558 interactions among predictors and remove redundant information using filters, nodes, or neurons
559 (Crossa et al., 2019), and modelling of this interaction is especially important for MT models
560 when primary and secondary traits are correlated . Our results confirm this, obtaining the highest
561 prediction accuracy with a MT-MLP model, compared to a UT-MLP model. The neuron weight
562 updates in the hidden layers leads to a combination of attributes for capturing the most suitable
563 hierarchical representation. Similarly, random forest development uses independent tree
564 branches, and the conclusion is based on the forest's average with individual tree branches being
565 uncorrelated, instead of any one individual branch (Waldmann, 2016). In this way, both random
566 forest and MLP models are an improvement for mapping complex interactions and resulted in
567 the highest prediction accuracy for both grain protein content and grain yield in this study.

568 The computation time increased linearly with inclusion of more variables such as SRI in
569 this study for MT GBLUP and Bayesian models; however, MT machine and deep learning
570 models are well developed for parallelism in computation (Lecun et al., 2015). We observed that
571 MT machine and deep learning models were four times faster than the MT Bayesian models, due
572 to their capacity to parallelize the operations. With the continuous increase in secondary traits
573 owing to utilization of high throughput phenotyping tools in the breeding program, breeders need
574 to shift from traditional Bayesian models to more computationally efficient models, and machine
575 and deep learning models provide a new avenue in this regard. Grain yield and grain protein
576 content had different heritability in each environment, with 2015 being lowest. The MT machine

577 and deep learning models resulted in the maximum increase in prediction accuracy for both traits
578 in the 2015 environment, showing that these models are also superior for lower heritability traits
579 or environments, which is often seen in plants (González-Camacho et al., 2018).
580
581 It is a misconception that machine and deep learning models should be used only on large
582 datasets having thousands of individuals, which is difficult for traits like grain yield or grain
583 protein content which are evaluated at mid to late stages in breeding programs (Angermueller et
584 al., 2016). However, our results and other related work indicate that machine and deep learning
585 models have similar or higher performance compared to traditional GBLUP and Bayesian
586 models when using data from hundreds of individuals for model training (Zingaretti et al., 2020).
587 Furthermore, a large dataset with 100k individuals was used for predictions with deep learning
588 models in the GS context and did not observe any superiority over the traditional linear mixed
589 models (Bellot et al., 2018). Zingaretti et al. (2020) and Montesinos-López et al. (2018a)
590 demonstrated deep learning models have higher prediction accuracy than GBLUP and other
591 mixed models when using 1233 strawberry and 250 wheat genotypes. These results, combined
592 with ours, suggest that training population size plays a minor role compared to the genetic
593 architecture of the trait and secondary traits utilized in the models. However, adequate training
594 population sizes remain important in GS.

595 The main issue with machine and deep learning models is the lack of biological
596 significance as different hyperparameters in the models handle different data parts. These models
597 might not be useful in providing genetic insight for the primary and secondary traits employed,
598 and hence genome-wide association studies are an important complement. Furthermore,
599 compared to GBLUP, machine and deep learning models were computationally intensive

because of the extra step of hyperparameter optimization. Hyperparameter optimization is required separately for each trait and could significantly discourage users when results are needed quickly in breeding programs for making selections (Cho & Hegde, 2019). Plant breeders are often interested in each predictor's significance in the models, which is not possible in deep learning because of their black-box nature due to many hidden layers, neurons, and filters. Finally, machine and deep learning model implementation requires a sufficient background in computer science, mathematics, and machine learning, which will require additional efforts by plant breeders, or accomplished through efficient collaborations. Although there are some potential hindrances in implementing machine and deep learning models, their utilization can result in the improvement of prediction accuracy for complex traits of interest in breeding programs. Overall, this study presents the benefits to utilizing MT machine and deep learning models for predicting complex traits under selection while efficiently using spectral information.

Conclusion: In this study, we evaluated the potential of MT machine and deep learning models for predicting grain yield and grain protein content using spectral information in a wheat breeding program. The model's performances were compared with traditional UT and MT GBLUP and Bayesian models under cross-validation and independent predictions. Our results showed the vast potential for MT machine and deep learning models with spectral information as a proxy phenotype. Random forest and multilayer perceptron were the best performing models for both traits under all the evaluated scenarios, closely followed by convolutional neural network and GBLUP. Green normalized difference vegetation index was the best SRI for predicting grain protein content for most MT models under cross-validation and independent predictions. Furthermore, machine and deep learning models were competitive with Bayesian

623 models as they were less computationally intensive than Bayesian models. This and previous
624 studies on deep learning for predicting complex traits shows enormous potential for the
625 utilization of these models in plant breeding programs to enhance genetic gain for quantitative
626 traits.

627

628 **Conflict of Interest**

629 The authors declare no conflict of interest.

630

631 **Authors contributions**

632 KS: conceptualized the idea, analyzed data, and drafted the manuscript; SP: assisted in data
633 analysis and edited the manuscript; MP: conducted field trials, edited the manuscript and
634 obtained the funding for the project; AC: supervised the study, conducted field trials, edited the
635 manuscript and obtained the funding for the project.

636

637 **Funding**

638 This project was supported by the Agriculture and Food Research Initiative Competitive Grant
639 2017-67007-25939 (WheatCAP) and 2016-68004-24770 from the USDA National Institute of
640 Food and Agriculture and Hatch project 1014919.

641

642 **References**

643 Abdollahi-Arpanahi, R., Gianola, D., & Peñagaricano, F. (2020). Deep learning versus

parametric and ensemble methods for genomic prediction of complex phenotypes. *Genetics Selection Evolution*, *52*(1), 12. https://doi.org/10.1186/s12711-020-00531-z

Aguate, F. M., Trachsel, S., González Pérez, L., Burgueño, J., Crossa, J., Balzarini, M., … de los Campos, G. (2017). Use of hyperspectral image data outperforms vegetation indices in prediction of maize yield. *Crop Science*, *57*(5), 2517–2524. https://doi.org/10.2135/cropsci2017.01.0007

Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2016). Deep learning for computational biology. *Molecular Systems Biology*, *12*(7), 878. https://doi.org/10.15252/msb.20156651

Gulli, A., and Pal, S. (2017). Deep learning with keras. Birmingham: Packt Publishing Ltd.

Aravind, J., Sankar, S. M., Wankhede, D. P., Kaur, V. (2020). augmentedRCBD: analysis of augmented randomised complete block designs. R package version 0.1.3. https://doi.org/10.2307/2527837

Babar, M. A., Reynolds, M. P., Van Ginkel, M., Klatt, A. R., Raun, W. R., & Stone, M. L. (2006). Spectral reflectance to estimate genetic variation for in-season biomass, leaf chlorophyll, and canopy temperature in wheat. *Crop Science*, *46*(3), 1046–1057. https://doi.org/10.2135/cropsci2005.0211

Barmeier, G., Hofer, K., & Schmidhalter, U. (2017). Mid-season prediction of grain yield and protein content of spring barley cultivars using high-throughput spectral sensing. *European Journal of Agronomy*, *90*, 108–116. https://doi.org/10.1016/j.eja.2017.07.005

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

Bellot, P., de los Campos, G., & Pérez-Enciso, M. (2018). Can deep learning improve genomic prediction of complex human traits? *Genetics*, *210*(3), 809–819. https://doi.org/10.1534/genetics.118.301298

Bhatta, M., Gutierrez, L., Cammarota, L., Cardozo, F., Germán, S., Gómez-Guerrero, B., … Castro, A. J. (2020). Multi-trait genomic prediction model increased the predictive ability for agronomic and malting quality traits in barley (Hordeum vulgare L.). *G3: Genes, Genomes, Genetics*, *10*(3), 1113–1124. https://doi.org/10.1534/g3.119.400968

Blake, N. K., Pumphrey, M., Glover, K., Chao, S., Jordan, K., Jannick, J. L., … Talbert, L. E. (2019). Registration of the triticeae-cap spring wheat nested association mapping population. *Journal of Plant Registrations*, *13*(2), 294–297. https://doi.org/10.3198/jpr2018.07.0052crmp

Calus, M. P., & Veerkamp, R. F. (2011). Accuracy of multi-trait genomic selection using different methods. *Genetics Selection Evolution*, *43*(1), 26. https://doi.org/10.1186/1297-9686-43-26

Cho, M., & Hegde, C. (2019). "Reducing the search space for hyperparameter optimization using group sparsity" in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* (Vol. 2019-May, pp. 3627–3631). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ICASSP.2019.8682434

684 Crain, J., Mondal, S., Rutkoski, J., Singh, R. P., & Poland, J. (2018). Combining high-throughput
685 phenotyping and genomic information to increase prediction and selection accuracy in
686 wheat breeding. *The Plant Genome*, *11*(1), 170043.
687 https://doi.org/10.3835/plantgenome2017.05.0043

688 Crossa, J., Martini, J. W., Gianola, D., Pérez-Rodríguez, P., Jarquin, D., Juliana, P., … Cuevas, J.
689 (2019). Deep kernel and deep learning for genome-based prediction of single yraits in
690 multienvironment breeding trials. *Frontiers in Genetics*, *10*, 1168.
691 https://doi.org/10.3389/fgene.2019.01168

692 de los Campos, G., and Grüneberg, A. (2016). MTM (multi-trait model) package,
693 http://quantgen.github.io/MTM/vignette.html (accesed 10.20.2020).

694 Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R
695 package rrBLUP. *The Plant Genome*, *4*, 250–255.
696 https://doi.org/10.3835/plantgenome2011.08.0024

697 Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian
698 Inheritance. *Royal Sociey of Edinburgh*, *52*, 399–433.

699 Gianola, D., Fernando, R. L., & Stella, A. (2006). Genomic-assisted prediction of genetic value
700 with semiparametric procedures. *Genetics*, *173*(3), 1761–1776.
701 https://doi.org/10.1534/genetics.105.049510

702 Gitelson, A. A., Kaufman, Y. J., & Merzlyak, M. N. (1996). Use of a green channel in remote
703 sensing of global vegetation from EOS- MODIS. *Remote Sensing of Environment*, *58*(3),
704 289–298. https://doi.org/10.1016/S0034-4257(96)00072-7

705 Gizaw, S. A., Godoy, J. G. V., Garland-Campbell, K., & Carter, A. H. (2018). Using spectral
706 reflectance indices as proxy phenotypes for genome-wide association studies of yield and
707 yield stability in pacific northwest winter wheat. *Crop Science*, *58*(3), 1232–1241.
708 https://doi.org/10.2135/cropsci2017.11.0710

709 González-Camacho, J. M., Ornella, L., Pérez-Rodríguez, P., Gianola, D., Dreisigacker, S., &
710 Crossa, J. (2018). Applications of machine learning methods to genomic selection in
711 breeding wheat for rust resistance. *The Plant Genome*, *11*(2), 170104.
712 https://doi.org/10.3835/plantgenome2017.11.0104

713 Habyarimana, E., Franceschi, P. De, Ercisli, S., & Baloch, F. S. (2020). Genome-wide
714 association study for biomass related traits in a panel of sorghum bicolor and S . bicolor × S
715 . halepense populations. *Frontiers in Plant Science*, *11*, 551305.
716 https://doi.org/10.3389/fpls.2020.551305

717 Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data*
718 *mining, inference, and prediction.* Springer Science & Buisness Media.

719 Heffner, E. L., Lorenz, A. J., Jannink, J. L., & Sorrells, M. E. (2010). Plant breeding with
720 genomic selection: gain per unit time and cost. *Crop Science*, *50*(5), 1681–1690.
721 https://doi.org/10.2135/cropsci2009.11.0662

722 Jia, Y., & Jannink, J. L. (2012). Multiple-trait genomic selection methods increase genetic value
723 prediction accuracy. *Genetics*, *192*(4), 1513–1522.

https://doi.org/10.1534/genetics.112.144246

Jordan, K. W., Wang, S., He, F., Chao, S., Lun, Y., Paux, E., Sourdille, P., … Akhunov, E. D. (2018). The genetic architecture of genome-wide recombination rate variation in allopolyploid wheat revealed by nested association mapping. *The Plant Journal*, *95*(6), 1039–1054. https://doi.org/10.1111/tpj.14009

Koch, P., Wujek, B., Golovidov, O., & Gardner, S. (2017). "Automated hyperparameter tuning for effective machine learning" in *Proceedings of the SAS Global Forum 2017 Conference*. Carry, NC, 1–23.

Large, E. C. (1954). Growth stages in cereals illustratin of the feeks scale. *Plant Pathology*, *3*(4), 128–129. https://doi.org/10.1111/j.1365-3059.1954.tb00716.x

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436-444. https://doi.org/10.1038/nature14539

Liu, Y., Wang, D., He, F., Wang, J., Joshi, T., & Xu, D. (2019). Phenotype prediction and genome-wide association study using deep convolutional neural network of soybean. *Frontiers in Genetics*, *10*, 1091. https://doi.org/10.3389/fgene.2019.01091

Lozada, D. N., & Carter, A. H. (2019). Accuracy of single and multi-trait genomic prediction models for grain yield in US Pacific Northwest winter wheat. *Crop Breeding, Genetics and Genomics*, *1*(1), e190012. https://doi.org/10.20900/cbgg20190012

Ma, W., Qiu, Z., Song, J., Li, J., Cheng, Q., Zhai, J., & Ma, C. (2018). A deep convolutional neural network approach for predicting phenotypes from genotypes. *Planta*, *248*(5), 1307–1318. https://doi.org/10.1007/s00425-018-2976-9

McKay, M.D. (1992). "Latin hypercube sampling as a tool in uncertainty analysis of computer models" in *Proceedings of the 24th Conference on Winter Simulation;* December 1992; 557–564.

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

Montesinos-López, A., Montesinos-López, O. A., Gianola, D., Crossa, J., & Hernández-Suárez, C. M. (2018a). Multi-environment genomic prediction of plant traits using deep learners with dense architecture. *G3: Genes, Genomes, Genetics*, *8*(12), 3813–3828. https://doi.org/10.1534/g3.118.200740

Montesinos-López, O. A., Martín-Vallejo, J., Crossa, J., Gianola, D., Hernández-Suárez, C. M., Montesinos-López, A., … Singh, R. (2019a). A benchmarking between deep learning, support vector machine and Bayesian threshold best linear unbiased prediction for predicting ordinal traits in plant breeding. *G3: Genes, Genomes, Genetics*, *9*(2), 601–618. https://doi.org/10.1534/g3.118.200998

Montesinos-López, O. A., Martín-Vallejo, J., Crossa, J., Gianola, D., Hernández-Suárez, C. M., Montesinos-López, A., … Singh, R. (2019b). New deep learning genomic-based prediction model for multiple traits with binary, ordinal, and continuous phenotypes. *G3: Genes, Genomes, Genetics*, *9*(5), 1545-1556. https://doi.org/10.1534/g3.119.300585

763 Montesinos-López, O. A., Montesinos-López, A., Crossa, J., Gianola, D., Hernández-Suárez, C.
764 M., & Martín-Vallejo, J. (2018b). Multi-trait, multi-environment deep learning modeling for
765 genomic-enabled prediction of plant traits. *G3: Genes, Genomes, Genetics*, *8*(12), 3829–
766 3840. https://doi.org/10.1534/g3.118.200728

767 Montesinos-López, O. A., Montesinos-López, A., Crossa, J., Toledo, F. H., Pérez-Hernández, O.,
768 Eskridge, K. M., & Rutkoski, J. (2016). A genomic bayesian multi-trait and multi-
769 environment model. *G3: Genes, Genomes, Genetics*, *6*(9), 2725–2774.
770 https://doi.org/10.1534/g3.116.032359

771 Pedregosa, F., Michel, V., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., … Duchesnay, E.
772 (2011). Scikit-learn: machine learning in python. *Journal of Machine Learning Research*,
773 *12*, 2825-2830.

774 Peñuelas, J., Gamon, J. A., Fredeen, A. L., Merino, J., & Field, C. B. (1994). Reflectance indices
775 associated with physiological changes in nitrogen- and water-limited sunflower leaves.
776 *Remote Sensing of Environment*, *48*(2), 135–146. https://doi.org/10.1016/0034-
777 4257(94)90136-8

778 Pérez-Enciso, M., & Zingaretti, L. M. (2019). A guide for using deep learning for complex trait
779 genomic prediction. *Genes*, *10*(7), 553. https://doi.org/10.3390/genes10070553

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

782 Poland, J. A., & Rife, T. W. (2012). Genotyping-by-sequencing for plant breeding and genetics.
783 *The Plant Genome*, *5*(3), 92–102. https://doi.org/10.3835/plantgenome2012.05.0005

784 Pook, T., Freudenthal, J., Korte, A., & Simianer, H. (2020). Using local convolutional neural
785 networks for genomic prediction. *Frontiers in Genetics*, *11*, 561497.
786 https://doi.org/10.3389/fgene.2020.561497

787 Prasad, B., Carver, B. F., Stone, M. L., Babar, M. A., Raun, W. R., & Klatt, A. R. (2007).
788 Genetic analysis of indirect selection for winter wheat grain yield using spectral reflectance
789 indices. *Crop Science*, *47*(4), 1416–1425. https://doi.org/10.2135/cropsci2006.08.0546

790 R Core Team. (2017). A language and environment for statistical computing. Vienna, Austria: R
791 foundation for statistical computing. Retrieved from https://www.R-project.org/.

792 Rouse Jr, J.W., Haas R. H., Deering, D. W. & Schell, J. A. (1972). Monitoring vegetation
793 systems in the Great Plains with ERTS. In: S.C. Freden, et al., editors, Third Earth
794 Resources Technology Satellite-I Symposium. Washington, DC.

795 Samuel A.L. (1959). Some studies in machine learning. *IBM Journal of Research and*
796 *Development, 44*(1.2). doi: 10.1147/rd.441.0206

797 Sandhu, K. S., Lozada, D. N., Zhang, Z., Pumphrey, M. O., & Carter, A. H. (2021a). Deep
798 learning for predicting complex traits in spring wheat breeding program. *Frontiers in Plant*
799 *Science*, *11*, 613325. https://doi.org/10.3389/fpls.2020.613325

800 Sandhu, K. S., Mihalyov, P. D., Lewien, M. J., Pumphrey, M. O., & Carter, A. H. (2021b).
801 Combining genomic and phenomic information for predicting grain protein content and

grain yield in spring wheat. *Frontiers in Plant Science*, *12*, 170. https://doi.org/10.3389/fpls.2021.613300

Sankaran, S., Khot, L. R., & Carter, A. H. (2015). Field-based crop phenotyping□: multispectral aerial imaging for evaluation of winter wheat emergence and spring stand. *Computers and Electronics in Agriculture*, *118*, 372–379. https://doi.org/10.1016/j.compag.2015.09.001

SAS Institute Inc 2011. MP Genomics. Release 6.0. SAS Inst., Cary, NC.

Shengqiang, Z., Dekkers, J. C. M., Fernando, R. L., & Jannink, J. L. (2009). Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: A barley case study. *Genetics*, *182*(1), 355–364. https://doi.org/10.1534/genetics.108.098277

Smith, P. F., Ganesh, S., & Liu, P. (2013). A comparison of random forest regression and multiple linear regression for prediction in neuroscience. *Journal of Neuroscience Methods*, *220*(1), 85–91. https://doi.org/10.1016/j.jneumeth.2013.08.024

Smola, A., & Scholkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*, 199–222.

Srivastava, N., Hinton, G., Krizhevsky, A., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(56), 1929-1958. https://doi.org/10.5555/2627435.2670313

Sun, J., Poland, J. A., Mondal, S., Crossa, J., Juliana, P., Singh, R. P., … Sorrells, M. E. (2019). High-throughput phenotyping platforms enhance genomic selection for wheat grain yield across populations and cycles in early stage. *Theoretical and Applied Genetics*, *132*(6), 1705–1720. https://doi.org/10.1007/s00122-019-03309-0

Sun, J., Rutkoski, J. E., Poland, J. A., Crossa, J., Jannink, J., & Sorrells, M. E. (2017). Multitrait, random regression , or simple repeatability model in high-throughput phenotyping data improve genomic prediction for wheat grain yield. *The Plant Genome*, *10*(2), 1-12. https://doi.org/10.3835/plantgenome2016.11.0111

Tishbirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*(11), 4414–4423. https://doi.org/10.3168/jds.2007-0980

Vapnik, V. (2013). The nature of statistical learning theory. Springer science & buisness media.

Voyant, C., Notton, G., Kalogirou, S., Nivet, M. L., Paoli, C., Motte, F., & Fouilloy, A. (2017). Machine learning methods for solar radiation forecasting: A review. *Renewable Energy*, *105*, 569-582. https://doi.org/10.1016/j.renene.2016.12.095

Waldmann, P. (2016). Genome-wide prediction using Bayesian additive regression trees. *Genetics Selection Evolution*, *48*(1), 42. https://doi.org/10.1186/s12711-016-0219-8

Waldmann, P., Pfeiffer, C., & Mészáros, G. (2020). Sparse convolutional neural networks for genome-wide prediction. *Frontiers in Genetics*, *11*, 25.

https://doi.org/10.3389/fgene.2020.00025

Zhang, A., Wang, H., Beyene, Y., Semagn, K., Liu, Y., Cao, S., … Zhang, X. (2017). Effect of trait heritability, training population size and marker density on genomic prediction accuracy estimation in 22 bi-parental tropical maize populations. *Frontiers in Plant Science*, *8*, 1916. https://doi.org/10.3389/fpls.2017.01916

Zheng, H., Cheng, T., Li, D., Zhou, X., Yao, X., Tian, Y., … Zhu, Y. (2018). Evaluation of RGB, color-infrared and multispectral images acquired from unmanned aerial systems for the estimation of nitrogen accumulation in rice. *Remote Sensing*, *10*(6), 824. https://doi.org/10.3390/rs10060824

Zingaretti, L. M., Gezan, S. A., Ferrão, L. F. V., Osorio, L. F., Monfort, A., Muñoz, P. R., … Pérez-Enciso, M. (2020). Exploring deep learning for complex trait genomic prediction in polyploid outcrossing species. *Frontiers in Plant Science*, *11*, 25. https://doi.org/10.3389/fpls.2020.00025

Table 1. Average phenotypic values and broad sense heritability of grain yield and grain protein content for the three environments (2014-2016).

Environment	Grain yield		Grain protein content	
	Phenotype (t/ha)	Heritability	Phenotype (%)	Heritability
2014	2.0	0.38	14.4	0.57
2015	1.7	0.24	12.2	0.35
2016	2.4	0.40	12.6	0.63

Table 2. Broad sense heritability of six different spectral reflectance indices obtained for each environment for utilization in multi-trait genomic selection models.

Environment	NDVI[a]	PRI[b]	NWI[c]	ARI[d]	NCPI[e]	GNDVI[f]
2014	0.75	0.93	0.60	0.74	0.64	0.69
2015	0.66	0.81	0.60	0.56	0.55	0.64
2016	0.80	0.60	0.74	0.76	0.82	0.72

[a] NDVI, Normalized difference vegetation index; [b] PRI, Photochemical reflectance index; [c] NWI, Normalized water index; [d] ARI, Anthocyanin reflectance index; [e] NCPI, Normalized chlorophyll pigment ratio index; [f] GNDVI, Green normalized difference vegetation index

Table 3. Phenotypic correlation between six different spectral reflectance indices collected at grain filling stage with grain yield and grain protein content for the three environments (2014-2016).

Trait	Environment	NDVI[a]	PRI[b]	NWI[c]	ARI[d]	NCPI[e]	GNDVI[f]
Grain yield	2014	0.33***	0.32***	0.36***	-0.16***	-0.30***	0.37***
	2015	0.06	0.03	0.09*	-0.09*	-0.12*	0.04
	2016	0.20***	0.15***	0.19***	-0.23***	-0.18***	0.20***
Grain protein content	2014	0.26***	0.10	0.27***	-0.12*	-0.20***	0.29***
	2015	.08*	0.01	-0.03	0.27***	0.02	0.18***

2016	-0.19***	-0.10*	-0.22***	0.12*	0.13*	-0.15***

[a] NDVI, Normalized difference vegetation index; [b] PRI, Photochemical reflectance index; [c] NWI, Normalized water index; [d] ARI, Anthocyanin reflectance index; [e] NCPI, Normalized chlorophyll pigment ratio index; [f] GNDVI, Green normalized difference vegetation index; *** significant at P < 0.0001; ** significant at P < 0.001; * significant at P < 0.05

Figure 1. Outline for multilayer perceptron model with one input, three hidden and an output layer. The connections between layers are represented with neurons (A). The representation of convolutional neural network employed in this study is provided with multiple layers (B). Figure 1A was redrawn from code in http://www.texample.net/tikz/examples/neural-network.

Figure 2. Prediction accuracies for grain yield with nine different uni- and multi-trait genomic selection models under the three different environments (2014-16) (**A-C**) using five-fold cross-validation. The x-axis represents the nine genomic selection models with faceting separating the uni- and multi-trait models.

Figure 3. Prediction accuracies for grain protein content with nine different uni- and multi-trait genomic selection model under the three different environments (2014-16) (**A-C**) using five-fold cross-validation. The x-axis represents the nine genomic selection models with faceting separates the uni- and multi-trait models.

Figure 4. Prediction accuracies for grain protein content for the three environments (**A-C**) with inclusion of six different spectral reflectance indices in the nine multi-trait genomic selection models. The x-axis represents the individual spectral reflectance indices and multi-trait genomic selection models are separated with facets for comparing across model performances.

Figure 5. Prediction accuracies for grain yield for the three environments (**A-C**) with inclusion of six different spectral reflectance indices in the nine multi-trait genomic selection models. The x-axis represents the individual spectral reflectance indices and multi-trait genomic selection models are separated with facets for comparing across model performances.

Figure 6. Independent prediction accuracies for grain yield (**A-C**) and grain protein content (**D-F**) using four different uni- and multi-trait genomic selection models. The first digit of the year represents the testing environment, and the second year represents the training environment. The x axis represents the different models and faceting separate the uni- and multi-trait models.

893

894

895

896

897

898

899

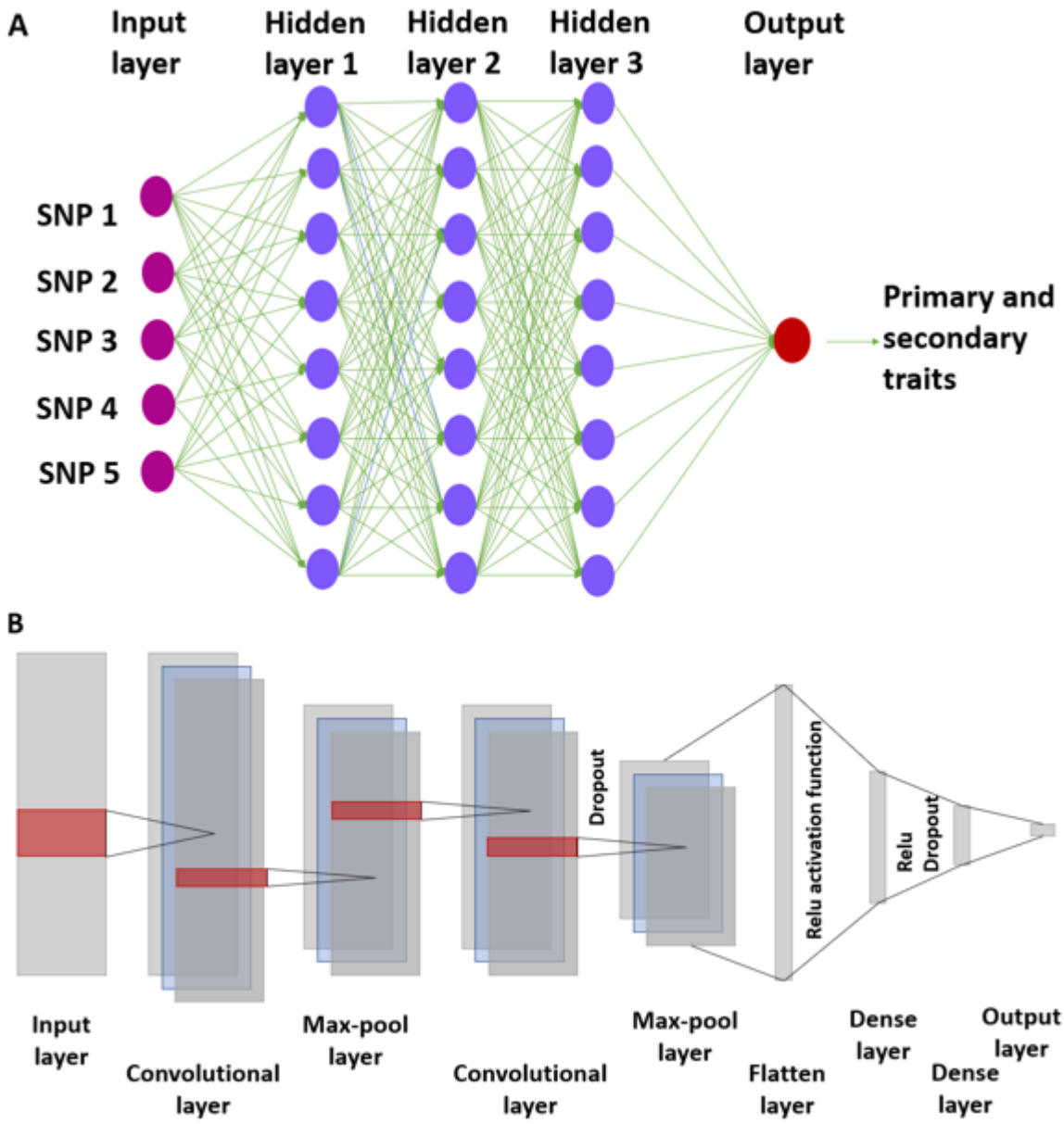
900

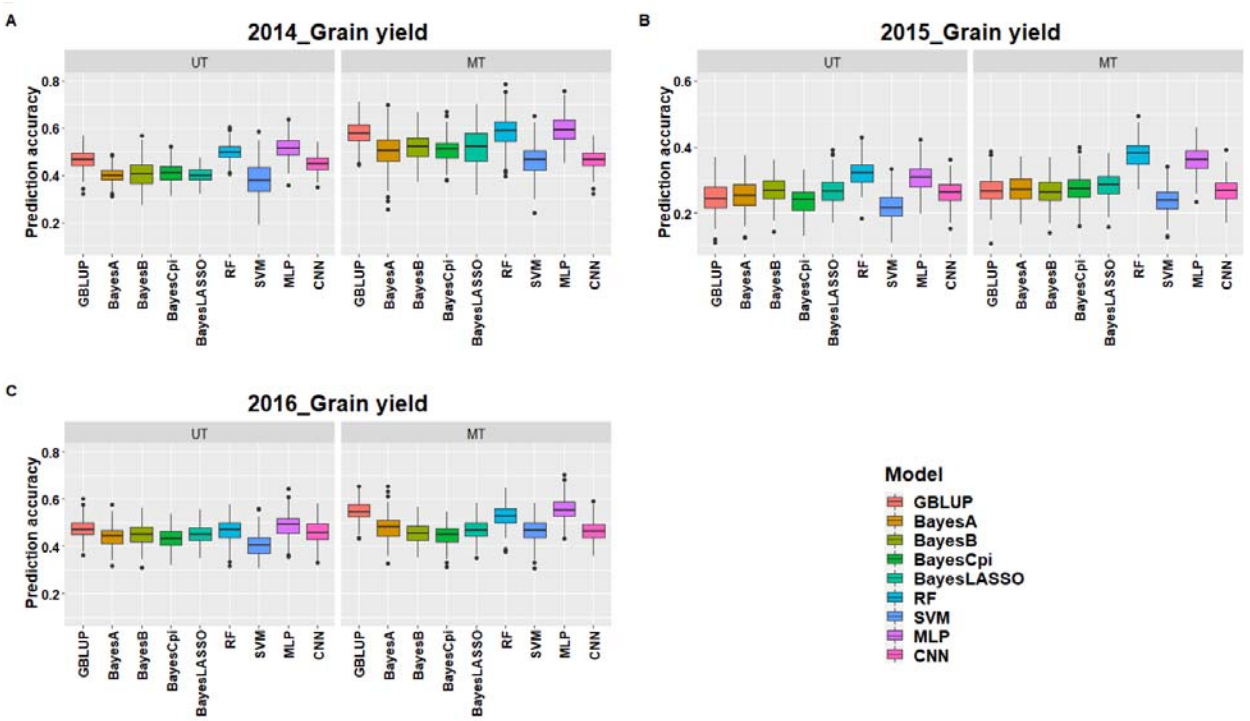
901

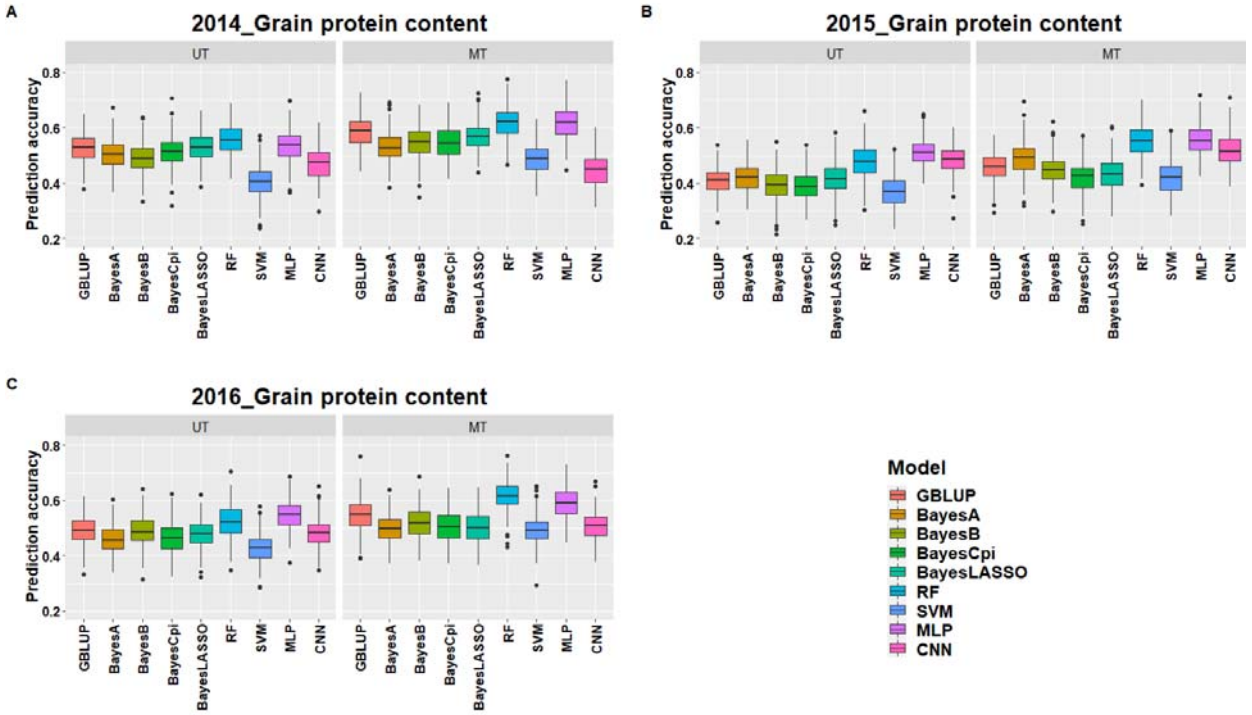
902

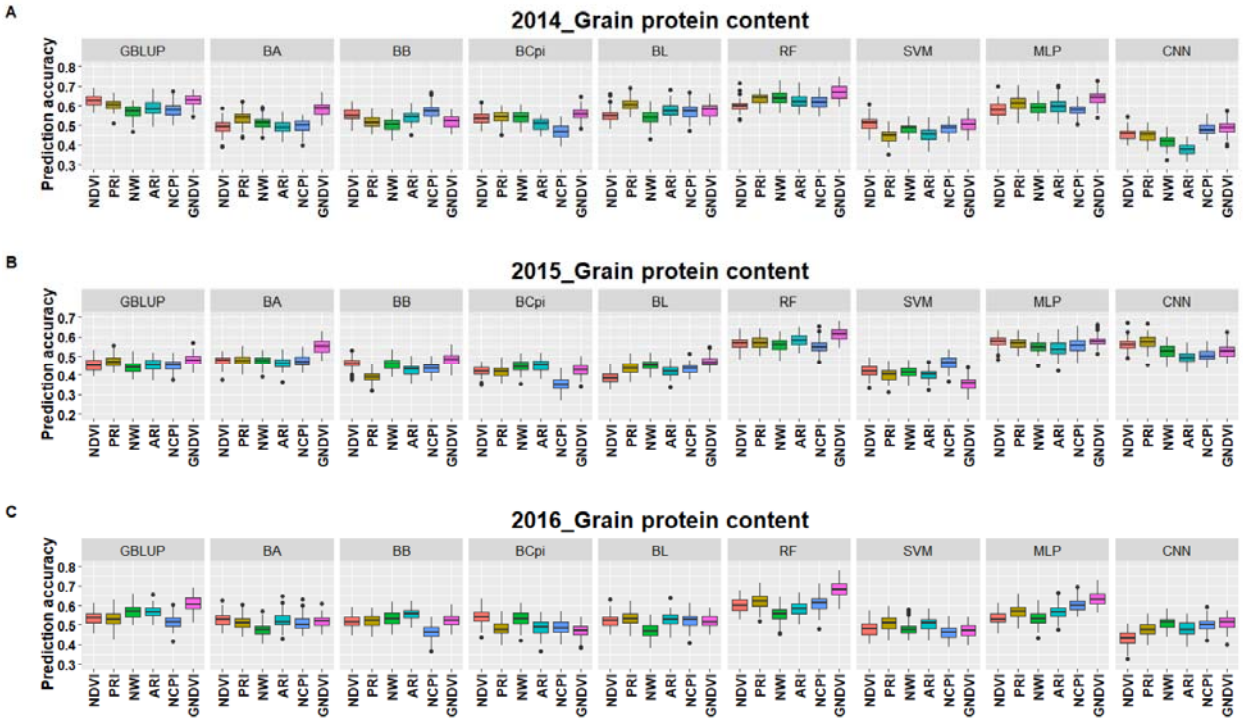
903

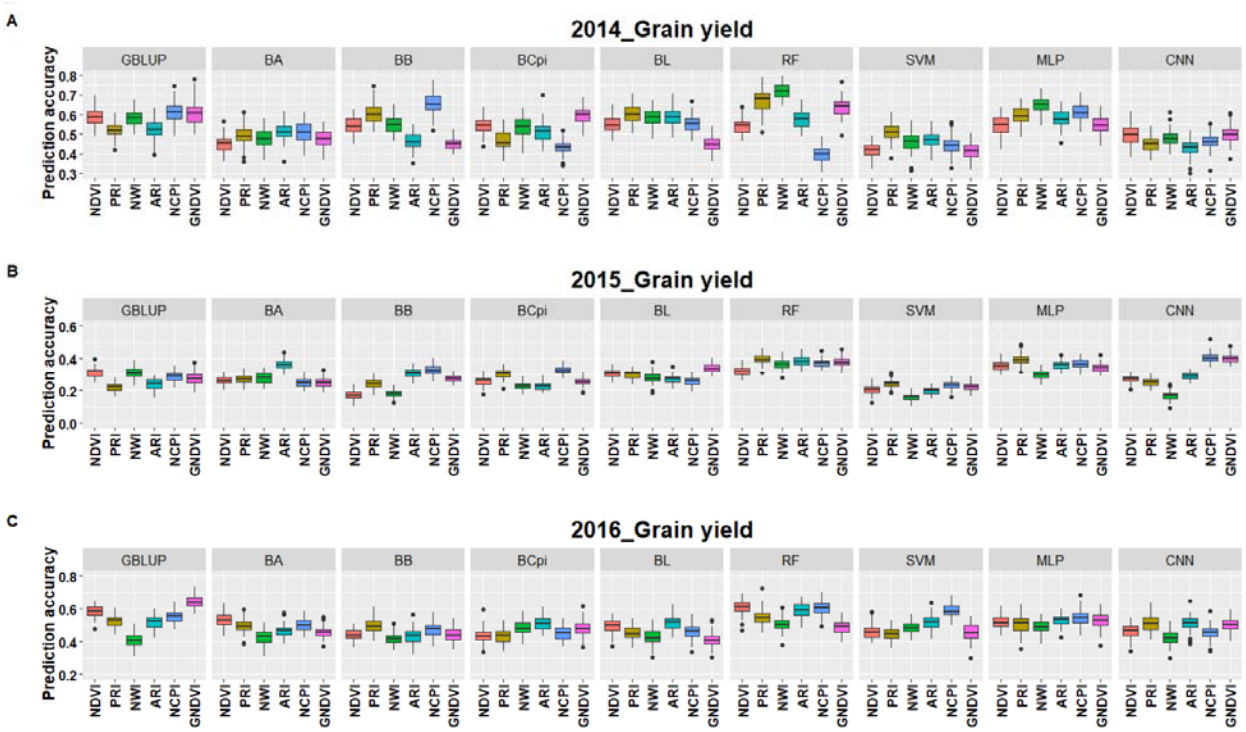
904

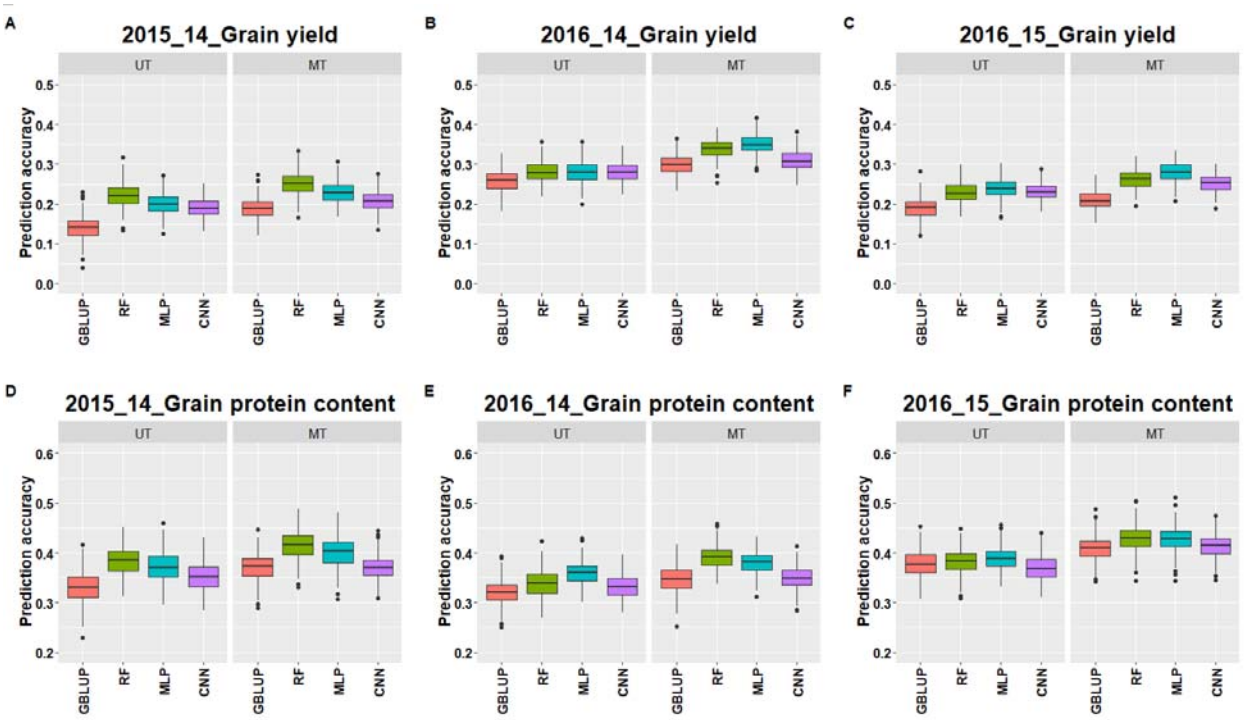












943