2020

Prediction Strategies for Leveraging Information of Associated Traits under Single- and Multi-Trait Approaches in Soybeans

Reyna Persa
University of Nebraska-Lincoln, reynapersa@gmail.com

Arthur Bernardeli
Universidade Federal de Viçosa, arthurbernardeli@gmail.com

Diego Jarquin
University of Nebraska-Lincoln, jhernandezjarquin2@unl.edu

agriculture

MDPI

Article

Prediction Strategies for Leveraging Information of Associated Traits under Single- and Multi-Trait Approaches in Soybeans

Reyna Persa [1], Arthur Bernardeli [2] and Diego Jarquin [1,*]

[1] Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, NE 68583, USA; reynapersa@gmail.com
[2] Department of Agronomy, Universidade Federal de Viçosa, Viçosa, MG 36570900, Brazil; arthurbernardeli@gmail.com
* Correspondence: diego.jarquin@gmail.com

Received: 31 May 2020; Accepted: 19 July 2020; Published: 22 July 2020

Abstract: The availability of molecular markers has revolutionized conventional ways to improve genotypes in plant and animal breeding through genome-based predictions. Several models and methods have been developed to leverage the genomic information in the prediction context to allow more efficient ways to screen and select superior genotypes. In plant breeding, usually, grain yield (yield) is the main trait to drive the selection of superior genotypes; however, in many cases, the information of associated traits is also routinely collected and it can potentially be used to enhance the selection. In this research, we considered different prediction strategies to leverage the information of the associated traits ([AT]; full: all traits observed for the same genotype; and partial: some traits observed for the same genotype) under an alternative single-trait model and the multi-trait approach. The alternative single-trait model included the information of the AT for yield prediction via the phenotypic covariances while the multi-trait model jointly analyzed all the traits. The performance of these strategies was assessed using the marker and phenotypic information from the Soybean Nested Association Mapping (SoyNAM) project observed in Nebraska in 2012. The results showed that the alternative single-trait strategy, which combines the marker and the information of the AT, outperforms the multi-trait model by around 12% and the conventional single-trait strategy (baseline) by 25%. When no information on the AT was available for those genotypes in the testing sets, the multi-trait model reduced the baseline results by around 6%. For the cases where genotypes were partially observed (i.e., some traits observed but not others for the same genotype), the multi-trait strategy showed improvements of around 6% for yield and between 2% to 9% for the other traits. Hence, when yield drives the selection of superior genotypes, the single-trait and multi-trait genomic prediction will achieve significant improvements when some genotypes have been fully or partially tested, with the alternative single-trait model delivering the best results. These results provide empirical evidence of the usefulness of the AT for improving the predictive ability of prediction models for breeding applications.

Keywords: associated traits; genomic prediction; multi-trait; single-trait; soybeans; SoyNAM

1. Introduction

Genomic prediction (GP) [1] has substantially changed conventional animal and plant breeding [2,3] in the last 20 years. It uses massive information of genome-wide dense markers to describe the genomic similarities between pairs of genotypes [4,5], allowing the borrowing of information between tested and untested genotypes. The ultimate goal in GP is to predict the expected

performance of untested genotypes (testing set) via the already observed genotypes (training set). This methodology allows the increase of the screening capacity, selection intensity, and consequently the genetic gains [6].

Several authors [5,7–9] have described and compared the performance of different single-trait GP models. Most of the studies have concluded there is not a superior model outperforming the others for all cases. The most common and convenient model is the GBLUP model because it is easy to implement and computationally more convenient. Besides model performance, numerous factors affect the accuracy of GP, such as the training set size [10,11], the genetic structure of the populations [12,13], genomic and phenotypic relationships between the training/calibration and testing sets [14], marker technology or marker density [15,16], and trait architecture [17], among others [18].

Other authors [19–21] have studied/proposed more elaborated models that allow the borrowing of information between traits under the multi-trait fashion. In general, multi-trait studies have shown that the predictive ability benefits from the sharing of information between traits [21,22] that are correlated due to pleiotropy or linkage disequilibrium [23,24]. In this context, novel phenotyping platforms capturing information of secondary traits in an efficient way [25–28] can be used to leverage the predictive ability of partially tested genotypes. Additionally, information from associated or secondary traits is routinely recorded in breeding programs. When the main objective is to predict yield performance only, two different approaches can be considered if phenotypic information of the secondary or associated traits is available: (1) Combine the genomic and the phenotypic information of these traits into the single-trait fashion via co-variance structures [29,30], or (2) implement the multi-trait prediction methods that are enhanced by using genetic correlations between traits [21,28].

In this research, we aimed to study to what extent the yield predictive ability and the predictive ability of the associated traits can be improved considering prediction strategies based on multi-trait models and single-trait models by combining marker and phenotypic information of the associated traits. In order to compare the model performance of the different prediction strategies, we used phenotypic data and molecular marker information from the SoyNAM project tested in Nebraska in 2012.

2. Materials and Methods

2.1. Phenotypic Data

Phenotypic information from the SoyNAM project was analyzed in this study. It was composed of 40 bi-parental families sharing a common parent (IA3023). A detailed description of the SoyNAM set can be found in [25,31], and the access to resources at https://www.soybase.org/SoyNAM. Information on 7 traits (grain yield mh^{-1}, protein content %, oil content %, plant height, lodging, seed size, and fiber) was collected in Nebraska in 2012 (40.575256, −98.137824). A total of 2560 genotypes with phenotypic information for all 7 traits were considered in this study. The number of genotypes per family varied between 46 and 76, with a mean of 64 and a standard deviation (SD) of 6. For an easier identification, the set of traits (6) other than yield (grain yield) are referred to as associated traits (AT).

2.2. Genomic Data

Genotypes were sequenced with a 6k array [25,31]. After discarding those molecular markers with more than 50% of missing values and a minor allele frequency smaller than 3%, a total of 4250 Single Nucleotide Polymorphisms (SNPs) remained for analysis.

2.3. Models

Two different methods were considered for the prediction of missing phenotypes. Single-trait (S), where each trait was analyzed independently, and the phenotypic information used for model calibration corresponds to the same trait; and multi-trait (M), where all the traits were jointly analyzed,

and thus, the borrowing of information between traits was leveraged. Details of these two methods are provided below.

2.3.1. Single-Trait

M1: Single-trait based on marker data G.

Consider that the phenotypic measure $\left(y_{ij}\right)$ of the ith ($i = 1, 2, \ldots, n$) genotype for the jth ($j = 1, 2, \ldots, t$) trait can be represented as the sum of a common effect $\left(\mu_j\right)$ plus a linear combination between p marker SNPs $\left(X_{ijk};\ k = 1,\ 2,\ \ldots,\ p\right)$ and their corresponding effects, plus an error term, such that:

$$y_{ij} = \mu_j + \sum_{k=1}^{p} X_{ijk} b_{jk} + e_{ij}.$$

In addition, consider further assumptions on the marker effects (IID: independent and identically distributed outcomes) following normal densities, such that $b_{jk} \sim \mathrm{N}\left(0, \sigma^2_{b_j}\right)$ and with $\sigma^2_{b_j}$ acting as the corresponding variance component. From the results from the multivariate normal density, we find that the vector of genetic effects $\mathbf{g_j} = \left\{\mathrm{g}_{ij}\right\}$, $\mathrm{g}_{ij} = \sum_{k=1}^{p} X_{ijk} b_{jk}$, for the jth trait is distributed as $\mathbf{g_j} \sim \mathrm{N}\left(\mathbf{0}, \mathbf{G}\sigma^2_{\mathbf{g}_j}\right)$, with $\mathbf{G} = \frac{\mathbf{X}\mathbf{X}'}{p}$ [4], $\sigma^2_{\mathbf{g}_j} = p \times \sigma^2_{b_j}$ and $\mathbf{X}$ representing the centered (by columns) matrix of molecular markers. After collecting the aforementioned results, **M1** becomes:

$$y_{ij} = \mu_j + \mathrm{g}_{ij} + e_{ij}, \tag{1}$$

where e_{ij}, the error term, follows IID normal densities, such that $e_{ij} \sim \mathrm{N}\left(0, \sigma^2_j\right)$ and σ^2_j represents the residual variance for the jth trait.

M2: Single-trait based on AT only.

This model is similar to M1 and is exclusively used for yield prediction; however, instead of using maker SNPs as covariates, it incorporates phenotypic information from the AT. Consider that the yield performance y_{i1} of the ith genotype can be represented as the sum of a common effect across genotypes (μ_1) plus a linear combination between the AT $\left(y_j,\ j = 2, 3, \ldots, t\right)$ and their corresponding effects (or weights), such that $\mathrm{c_i} = \sum_{j=2}^{t} y_{ij} w_j$, where $\mathrm{c_i}$ represents the phenotypic effect of the ith genotype, with $w_t \sim \mathrm{N}\left(0, \sigma^2_t\right)$ and σ^2_t as the corresponding variance component. This model can be written as follows:

$$y_{i1} = \mu_1 + \mathrm{c_i} + e_{i1}, \tag{2}$$

and from properties of the multivariate normal density, we find that the vector of phenotypic effects $\mathbf{c} = \{\mathrm{c}_i\}$ follows a multivariate density, such that $\mathbf{c} \sim \mathrm{N}\left(0, \mathbf{T}\sigma^2_T\right)$, where $\mathbf{T} = \frac{\mathbf{Y}_{\{-1\}}\mathbf{Y}'_{\{-1\}}}{t-1}$, $\mathbf{Y}_{\{-1\}}$ is the centered (by columns) matrix of AT and $\sigma^2_T = (t-1)\sigma^2_t$ is the corresponding variance component. $\mathbf{T}$ is the matrix of phenotypic relationships, and its entries describe the phenotypic similarities (based on the AT) between pairs of genotypes. This model requires perfect information of the AT for all genotypes (i.e., no missing values are allowed on these).

M3: Single-trait combining maker and AT data (G + T).

This model combines marker SNPs and the information of the AT by coupling models M1 and M2 into a single linear predictor as follows:

$$y_{i1} = \mu_j + \mathrm{g_{i1}} + \mathrm{c_i} + e_{ij}, \tag{3}$$

with all the model terms as previously defined.

2.3.2. Multi-Trait

M4: Multi-trait model based on marker data G.

Consider $\boldsymbol{y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_t\}$ as the vector of phenotypic responses for t (7) traits and it is of order $n \times t$ such that the linear predictor for all traits can be written in matrix formulation as:

$$\boldsymbol{y} = \boldsymbol{u} + \mathbf{g} + \mathbf{e}. \tag{4}$$

$\boldsymbol{u} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_t]$, with $\boldsymbol{u}_j$ as the vector of the common effect for the jth trait across genotypes; $\mathbf{g} = [\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_t]$ with $\mathbf{g}_j$ as the vector of genomic effects for the jth trait, such that $\mathbf{g} \sim \mathrm{N}(\mathbf{0}, \mathbf{G} \otimes \boldsymbol{\Sigma})$ and $\mathbf{e} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_n \otimes \mathbf{R})$, where $\boldsymbol{\Sigma}$ and $\mathbf{R}$ are the corresponding genetic and residual co-variance matrices between pairs of traits and these are of order $t \times t$ (7×7).

2.4. Prediction Schemes

To evaluate the performance of the described models for predicting grain yield and the AT as well, four different cross-validation (CV) scenarios were considered. Figure 1 provides a graphical representation of the different scenarios for predicting untested (CV00) and partially tested (CV2, CV2H, CVY) genotypes. These correspond to different prediction objectives (yield and AT), different levels of overlapping (50% and 75%) across genotypes, and different training (50% and 75%) and testing set sizes (50% and 25%), respectively.

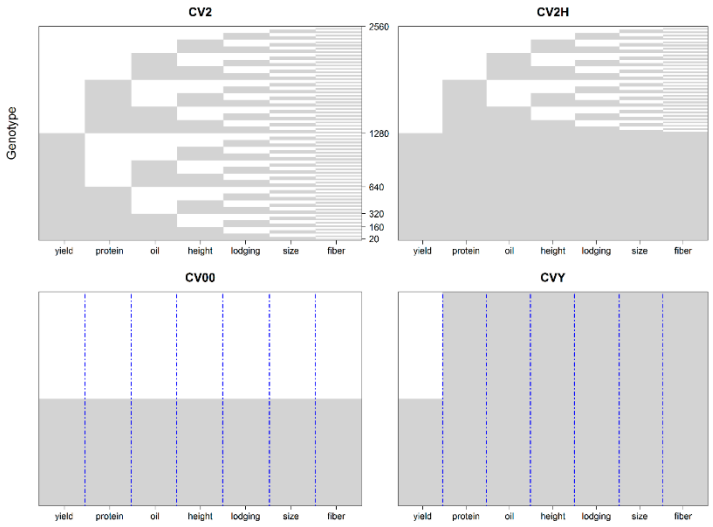

Figure 1. Graphical representation of four different cross-validation designs; horizontal gray blocks of lines (20 per block) correspond to those genotypes assigned to the training set while the horizontal blocks of white lines (20 per block) conform the testing set. CV2: for each trait, half of the phenotypic information was assigned to the training set while the remaining half to the testing set, such that this design allows different degrees of overlapping across traits for genotypes varying between no phenotypic information (20 genotypes) and information on all the 7 traits (20 genotypes). CV2H: similar to the previous scheme; however, there is phenotypic information available on all the traits for those genotypes with available yield records (training set for yield). CV00: there is phenotypic information available across traits only for those genotypes with yield records. CVY: There is phenotypic information on grain yield for half of the genotypes while full information for the remaining traits (6).

CV2 (top left in Figure 1) considers the partial information of genotypes (i.e., for a particular genotype, some traits were observed but not others). In this case for each trait, 50% (1280 genotypes) of the genotypes were assigned to training and testing sets, allowing some degree of overlapping across traits. There are 20 genotypes with no information for any of the 7 traits; also, there are 20 genotypes with information on all traits, and 140 genotypes with information for only one trait (not the same trait, 20 genotypes per trait). The full list of the number of genotypes by the number of observed traits is as follows: 20 (0), 140 (1), 420 (2), 700 (3), 700 (4), 420 (5), 140 (6), and 20 (7). In all scenarios, the rows

of the different cross-validation schemes (the same for all schemes) from Figure 1 were randomly assigned to training and testing sets, such that 50% of the genotypes comprised the training set and the remaining 50% the testing set. In this case, it is expected half of the genotypes in the previous list were assigned to each one of the sets.

CV2H (top right) also corresponds to partially tested genotypes but ensuring full information of all AT for those genotypes with yield records (training set) and partial information of the AT for the remaining genotypes (testing set). As such, 50% of the yield observations were assigned to the training sets, and consequently 75% of AT were assigned to the training set, and the remaining were assigned to the testing set (50% for yield and 25% for the AT). The full list of the number of genotypes by the number of observed traits is as follows: 20 (0), 120 (1), 300 (2), 400 (3), 300 (4), 120 (5), 20 (6), and 1280 (7). This prediction scheme allows (*i*) assessment of the effects in predictive ability when increasing the information of the AT for yield prediction, and (*ii*) assessment of the effects of increasing the training set size when predicting the AT. Under this scenario, the single-trait and the multi-trait methods can be applied. In this case, the training and the testing sets were comprised of the same rows as in the previous scheme. Similar to the previous scheme, 50% of the genotypes in this listing were expected to conform the training set while the remaining 50% composed the testing set.

CV00 (bottom left) corresponds to the case of predicting the trait performance of fully untested genotypes. No information for any of the traits is available for those genotypes in the testing set. Here, 50% of the genotypes were assigned to the training set, and the remaining 50% to the testing set.

CVY (bottom right) is exclusively used for yield prediction, and it requires the full information of the AT for all genotypes for its implementation. The objective of this scheme was to assess the levels of yield predictive ability that can be reached when considering perfect information of the AT under the single-trait and the multi-trait approaches.

Since different cross-validation schemes were intended for different prediction objectives with different models to avoid confusion, all the studied cases (strategies) are provided in Table 1. These strategies (model-cv scheme) were implemented for yield prediction only and for predicting the AT as well using the information of marker SNPs of AT and combined. With model M1 (baseline model), each trait is analyzed separately, and thus no borrowing of information between traits is possible. M2 and M3 models were exclusively used for yield prediction, and these allowed the incorporation of the AT, and of the AT with marker data in the analyses, respectively. The model M4 was used to jointly perform predictions on all traits for (*i*) partially tested (CV2, CV2H), (*ii*) untested (CV00) genotypes, (*iii*) and for yield prediction only (CVY) using information from AT.

Table 1. Strategies (model-cv scheme) for predicting yield performance and associated traits (AT) for different approaches (S, single-trait; M multi-trait) and different training/testing sizes (further details in Figure 1).

	S			**M**
	M1	**M2**	**M3**	**M4**
CV2	X			X
CV2H	X			X
CV00	X			X
CVY	X	X	X	X

2.5. Assessing Predictive Ability

Despite the adopted strategy and the approach (single-trait/multi-trait) used for performing the predictions, the predictive ability was assessed as the Pearson correlation between predicted and observed values within traits. The assignation of the training/testing sets was repeated 20 times and the mean correlation and corresponding standard deviation were computed across replicates for each strategy.

2.6. Software

Single-trait analyses (M1, M2, and M3) were performed using BGLR R-package [32] while the multi-trait analysis was implemented using the MTM R-package [20]. The fitted models were developed under the Bayesian approach considering normal priors for the main effects while Scaled-inverse Chi-Squared priors were considered for the variance components. A total of 50,000 samples were considered for the Gibbs sampler, 20,000 of these were used for burn-in and a thinning of 5. Thus, the final chain consisted of 6000 samples.

3. Results

3.1. Genomic Heritability

Table 2 presents the percentage of phenotypic variability explained for the different model terms (genomic and residual) of the baseline model M1, and the genomic heritability $\left(H^2 = \frac{\sigma^2_{g_j}}{\sigma^2_{g_j}+\sigma^2_j}\right)$ for each trait (jth). The variance components were obtained by conducting a full data set analysis (i.e., no missing values were allowed in the vector of responses) separately for each trait. Concerning the yield, around half of the phenotypic variability (48.56%) was explained by the genomic component (G). For the other traits, the heritability ranged between 0.4 (lodging) and 0.68 (protein and oil).

Table 2. Percentage of phenotypic variability captured with the genomic (G) and residual terms (R) of model M1, and the broad sense genomic heritability (H^2). The variance components were computed for 7 traits from a Nested Association Mapping population (SOYNAM) comprising information of 2560 genotypes belonging to 40 bi-parental families tested in Nebraska in 2012.

	vG	vR	H^2
Yield	48.56	51.44	0.49
Protein	68.30	31.70	0.68
Oil	67.57	32.43	0.68
Height	49.68	50.32	0.50
Lodging	40.36	59.64	0.40
Size	64.11	35.89	0.64
Fiber	54.99	45.01	0.55

3.2. Phenotypic and Genomic Correlations

Table 3 presents the phenotypic (lower triangular) and genetic (upper triangular) correlations among pairs of traits. The phenotypic correlations were computed as the Pearson correlation between pairs of traits. The matrix of genetic correlations was obtained by performing the full data set analysis (no missing data) under model M4. As expected, yield and protein showed negative phenotypic (−0.310) and genetic (−0.776) correlations while yield and oil showed positive trends (0.259 and 0.800, respectively). Interestingly, while yield and plant height showed a strong negative genetic correlation (−0.720), the phenotypic correlation was close to zero (−0.094). Yield and lodging showed strong negative phenotypic and genetic correlations (−0474 and −0.921, respectively). The phenotypic and genetic correlations between yield and seed size were positive (0.149 and 0.582), the same as between yield and fiber content (0.162 and 0.566).

Table 3. Phenotypic (lower diagonal) and genetic (upper diagonal) correlation among 7 traits from the Nested Association Mapping (SoyNAM) project comprising information on 2560 genotypes belonging to 40 bi-parental families tested in Nebraska in 2012.

	Yield	Protein	Oil	Height	Lodging	Size	Fiber
yield	1	−0.776	0.800	−0.720	−0.921	0.582	0.566
protein	−0.310	1	−0.918	0.696	0.773	−0.395	−0.877
oil	0.259	−0.693	1	−0.715	−0.802	0.548	0.627
height	−0.094	0.193	−0.168	1	0.842	−0.540	−0.508
lodging	−0.474	0.280	−0.245	0.300	1	−0.631	−0.555
size	0.149	−0.057	0.214	−0.115	−0.119	1	0.105
fiber	0.162	−0.681	0.031	−0.071	−0.122	−0.134	1

3.3. Heritability

Table 4 shows for the 7 traits, the squared root of the genomic heritability (H) (the theoretical upper threshold for predictive ability) and the goodness of fit derived from the full data analysis (FD; i.e., the correlation between the predicted and observed values without considering missing values) under the single-trait (M1_FD) and the multi-trait (M4_FD) models, and the average prediction accuracy for different prediction strategies (model-CV scheme) with different levels of connectivity across traits.

Table 4. The square root of the heritability (H), the goodness of fit using single-trait (M1_FD) and multi-trait (M4_FD) models, and the average predictive ability (20 replicates) of unobserved records for 7 traits using different prediction strategies: M1_CV00, single-trait genomic prediction with 50% of missing values per trait; M1_CV2H, single-trait genomic prediction with 75% of missing values for traits other than yield (50%); M4_CV00, multi-trait genomic prediction model with 50% of missing values across traits; M4_CV2, multi-trait genomic prediction model with 50% of the genotypes partially observed (i.e., some traits but not others for the same genotype); M4_CV2H, multi-trait genomic prediction with 75% of missing values for traits other than yield (50%).

Trait	M1_H	M1_FD	M4_FD	M1_CV00		M4_CV00		M4_CV2		M1_CV2H		M4_CV2H	
				Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Yield	0.72	0.63	0.74	0.547	0.011	0.513	0.019	0.581	0.014	0.547	0.011	0.580	0.015
Protein	0.56	0.73	0.89	0.645	0.008	0.568	0.026	0.730	0.009	0.675	0.017	0.733	0.016
Oil	0.57	0.74	0.87	0.636	0.012	0.597	0.022	0.708	0.010	0.680	0.022	0.722	0.018
Height	0.71	0.60	0.71	0.469	0.016	0.487	0.014	0.534	0.015	0.548	0.023	0.569	0.022
Lodging	0.77	0.58	0.71	0.512	0.014	0.478	0.021	0.547	0.013	0.514	0.026	0.545	0.020
Size	0.60	0.70	0.80	0.616	0.012	0.600	0.014	0.620	0.014	0.638	0.014	0.648	0.016
Fiber	0.67	0.67	0.80	0.583	0.012	0.547	0.022	0.638	0.012	0.610	0.025	0.652	0.019

For grain yield, the H was 0.72 while for the other traits it varied between 0.56 (protein) and 0.77 (lodging). The squared root of the genomic heritability (H) can be considered as the highest correlation between the predicted and observed values that can be achieved for each trait when predicting the missing values using molecular marker information only. The S_FD for yield was 0.63 and it varied between 0.58 (lodging) and 0.74 (oil) for the other traits. Under the multi-trait model, the FD for yield was 0.74 and it varied between 0.71 (plant height and lodging) and 0.89 (protein).

3.4. Predictive Ability

Regarding the predictive ability under the different prediction scenarios, the single-trait approach (M1_CV00) returned an average predictive ability of 0.547 (SD = 0.011) for yield while for the other traits, it varied between 0.469 (height, SD = 0.016) and 0.645 (protein, SD = 0.008). For grain yield, the results from this strategy were used as the baseline to compare the performance of the other strategies. In addition, the results of the strategies M1_CV00, M1_CV2, M1_CV2H, and CVY were

equivalent for the yield predictive ability, and thus only those from the M1_CV00 strategy were used to contrast with the remaining strategies (M4_CV00, M4_CV2, M4_CV2H).

The M4_CV00 strategy based on the multi-trait model returned a mean correlation for grain yield of 0.513 (SD = 0.019) while for the other traits, these values ranged between 0.478 (lodging) and 0.600 (seed size). Thus, when no information for any of the traits for those genotypes in the testing set was available, their predictive ability significantly decreased, except for plant height (0.469 and 0.487). When the genotypes were partially observed (50%) for some traits but not others, M4_CV2 returned an average yield predictive ability of 0.581 (SD = 0.014). This represents an improvement of 6% and of 13% with respect to the M1_CV00 and M4_CV00 strategies, respectively. For the remaining traits, the M4_CV2 strategy always returned the best results compared with strategies M1_CV00 (~1–14%) and M4_CV00 (~3–29%). Seed size and protein content were the less and more benefited traits of the borrowing of information between traits, respectively.

When the percentage of partially observed genotypes (i.e., not all the traits were observed) increased from 50% to 75% for the AT (CV2_H), benefits were not observed in yield prediction by increasing the levels of connectivity across traits. M4_CV2H returned a mean correlation of 0.580 (SD = 0.15). Under this strategy, yield is the only trait that can be compared with the other strategies because in all cases, the testing set is the same (size and genotypes). For the remaining traits, the training and testing set sizes changed with respect to the previous strategies, thus their corresponding predictive ability can be compared only between strategies M1_CV2H and M4_CV2H. Slight improvements (2%~9%) in predictive ability were observed with M4 with respect to M1; under the CV2H scheme, the most benefited trait of the borrowing of information between traits was protein.

Models M2 and M3 were used to leverage the information of the AT for yield prediction under the single-trait approach. These models require complete information of the AT for all genotypes (i.e., no missing values are allowed on these).

A fair comparison of the performance of all the different prediction strategies can be done with yield prediction only (training-testing sets).

Figure 2 presents the square root of the heritability (H) obtained with the single-trait model M1, the goodness of fit derived from the single-trait (M1_FD) and the multi-trait (M4_FD) models, and the average yield predictive ability (and SDs) for the different prediction strategies (model-CV scheme). In all the cases, the testing set comprised 50% of the genotypes while the remaining 50% were used for model calibration. M3_CVY, the single-trait strategy that combines marker SNPs and information of the AT, returned the best results (mean = 0.681 and SD = 0.010). M4_CVY, the multi-trait strategy that considers the same amount of information than the previous one, was the second-best strategy (mean = 0.607, SD = 0.014). The improvement of M3_CVY and M4_CVY with respect to the baseline strategy (M1_CVY; mean = 0.547, SD = 0.011) was 25% and 11%, respectively. Additionally, the improvement of the single-trait strategy M3_CVY with respect to the multi-trait strategy M4_CVY was 12%.

The multi-trait strategies M4_CV2 (0.581, SD = 0.014) and M4_CVH (0.580, SD = 0.015) practically performed the same and their results were slightly below M4_CVY (0.607). When only phenotypic information of the AT was used, the M2_CVY strategy returned a mean correlation of 0.516 (SD = 0.011). The less efficient strategy was M4_CV00 (0.513, SD = 0.018). Thus, under the multi-trait model, when no information of the AT is available for those genotypes in the testing set, the predictive ability is negatively affected.

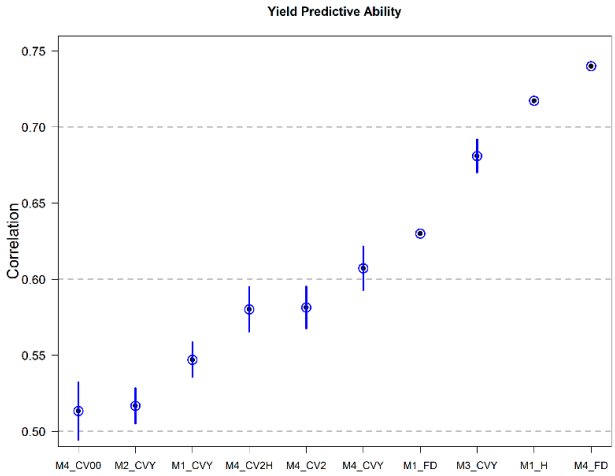

Figure 2. Square root of the heritability (H; M1_H), model goodness of fit under single-trait (M1_FD) and multi-trait (M4_FD) models, and mean predictive ability (20 replicates) and uncertainty interval of the mean plus minus the standard deviation for 7 different strategies for grain yield. M1_CVY, single-trait model using marker SNPs; M2_CVY, single-trait model using information of the AT as covariates; M3_CVY, single-trait model combining maker SNPs and phenotypic data of the AT; M4_CV00, multi-trait model for predicting fully untested genotypes; M4 _CV2; multi-trait model for predicting partially tested (50%) genotypes; M4_CV2H; multi-trait model for predicting partially tested (75%) genotypes; M4_CVY, multi-trait model for predicting partially tested genotypes for traits other than yield (100%).

4. Discussion

In this study, we compared different prediction strategies for addressing different prediction problems. Only the results for grain yield can be compared across all these prediction strategies. For the remaining traits (AT), their results are comparable only when the same training and testing sets were used for different prediction scenarios (CV00, CV2, and CV2H). The heritability varied across traits, with protein and oil being the traits with the highest values (0.68). These results suggest the predominance of the pleiotropic effect of genes [33]. On the other hand, grain yield was the trait with the lowest heritability (0.49). Thus, an improvement in predictive ability can be expected when the information of traits with high heritability in the multi-trait context is included [28,33,34].

Regarding the phenotypic and the genetic correlations, yield and protein exhibited a moderated negative phenotypic correlation (−0.310) while the genetic correlation was stronger (−0.766). On the other hand, as we expected, yield and oil showed positive phenotypic (0.259) and a genetic correlation (0.800). Oil and protein are two of the main constituents of the seed composition and these were strongly negatively correlated at the phenotypic (−0.693) and the genetic (−0.918) levels as addressed by [35,36]. The strong correlation among traits (between AT, and between AT and yield) might indicate that the inclusion of the AT can be beneficial for yield prediction and for predicting the AT as well as under the multi-trait approach [28,33].

The goodness of fit was significantly higher under the multi-trait model compared with the single-trait approach. Additionally, these values were higher than the squared root of the heritability obtained with the single-trait model. The square root of the heritability can be considered as the upper threshold of the predictive ability under the single-trait model. These results provide insights about the levels of prediction accuracy that can be reached when the phenotypic information from the AT is combined with the marker data. We should recall that H was computed using maker and yield phenotypic information from the same trait (yield), thus an increase in predictive ability is expected with the addition of the information from the AT. In our case, the multi-trait strategies improved the results from the single-trait strategy, which only uses marker data, when at least 50% of the genotypes were partially observed.

The different cross-validation schemes attempt to mimic some realistic scenarios that breeders might face in their breeding programs. When the main interest is the yield prediction only, sometimes there might be available information of the AT for those genotypes in the testing set. Thus, this information can be used for improving the yield predictive ability to some degree with respect to the single-trait baseline strategy. In our case, the M4_CV2 strategy, which includes information of the AT for some of the genotypes (50%) in the testing set, improved the results of the baseline model (CV1_CV00) in about 6%. Increasing the information of the AT up to 75% (M4_CV2H), ensuring complete information for all genotypes in the training set, did not benefit the yield predictive ability in comparison with the previous case.

On the other hand, when no information for the AT was available for any of the genotypes in the testing set, the yield predictive ability decreased to 0.513. This result contrasts with other studies [21,28,33], where the predictive ability practically remained unchanged with respect to the baseline model (single-trait model). When the information of all the AT was available for all genotypes (M4_CVY = 0.607), slight improvements were observed with respect to all the previous strategies (single-trait and multi-trait) and these ranged between 5% and 18%.

When the single-trait strategy was implemented considering the same information as in the previous strategy (i.e., perfect information for AT and marker data) but leveraged in a different way, the predictive ability was significantly improved with respect to all previous strategies (single-trait and multi-trait). The M3_CVY strategy returned a mean correlation 0.681 and this represents a relative improvement between 12% and 33% with respect to all the previous strategies. In this case, the information from the AT was included using a variance-covariance matrix between traits in addition to the molecular marker information. Other authors have shown similar improvements predicting yield when combining information from secondary traits with maker information in comparison with the baseline strategy. For example, [30] showed an improvement of 22% when combining soybean canopy coverage recorded in the early stages of plant growth (days 14–33 after planting) with marker data. Additionally, [29] showed an improvement of 13% when combining the soybean canopy data with the marker information via a hybrid matrix where the importance of both components was sequentially weighted. Other studies have shown similar improvements when combining both sources of information (marker and AT) for yield prediction in maize [37] and wheat [38].

The strategy that only considers the information of the AT for yield prediction (M2_CVY) decreased the predictive ability of the baseline strategy (M1_CVY) by around 6%. Additionally, these results were slightly better than those from the multi-trait strategy when the information from the AT was not available for those genotypes in the testing set (M4_CV00). Other authors also showed a reduction in the yield predictive ability when using the information from other traits only. For example, [29] showed a significant reduction in the soybean predictive ability of around 48% with respect to the M1_CVY strategy; [30] also showed a decrease of around 25% in two out of three predicted soybean trials.

With respect to the other traits, we observed that the predictive ability was improved (6–14%) only when the genotypes in the testing set were partially observed. Additionally, small improvements were observed when the training sample size was increased under the multi-trait approach (M4_CV2H) and these improvements were higher than those from the single-trait (M1_CV2H) for the same partitions. In this case, these results provide evidence of the benefits of the borrowing of information between correlated traits. On the other hand, when the information of the AT was absent for all genotypes, the predictive ability was reduced (except for plant height) under the multi-trait strategy (M4_CV00) with respect to the baseline strategy (M1_CV00).

In cases of full data availability, M3 showed to be more appropriate. However, perfect information of the AT is not always available in the routine of a soybean breeding program. For these cases, it is feasible to apply multi-trait models in scenarios where there is correspondence up to a certain degree between genotypes in training and testing sets across the correlated traits (i.e., M4_CV2). Additionally, when the testing sets are reduced in terms of the number of genotypes for the secondary traits, the yield predictive ability remained unchanged with respect to the previous strategy; however, the M4_CV2H

predictive ability of the remaining traits was slightly improved. Further studies should consider a better selection of the AT to be included for yield prediction and the prediction of these ATs as well. For example, questions like up to what extent can the predictive ability be improved by adding only those ATs with the strongest phenotypic/genomic correlation with yield, or what is the lowest phenotypic/genetic correlation threshold that should be considered to ensure significantly higher correlations than those derived from the baseline model should be studied.

In general, the results from this study showed that improvements in the yield predictive ability can be accomplished only when the information from the AT is available for those genotypes in testing sets when performing the single-trait and the multi-trait strategies. In both approaches, the predictive ability was significantly improved with respect to the baseline strategy (M1_CVY). We also showed that the same information (i.e., markers and AT) can be leveraged differently depending on the adopted strategy. The single-trait strategy (M3_CVY) more efficiently combined the maker and phenotypic information from the AT than the multi-trait strategy (M4_CVY). A practical implication of adopting one approach or the other is that the single-trait approach is much easier to implement and is computationally less demanding. Another important implication of these results is that the use of AT is also convenient because in many cases, the measurement of these traits does not require an evaluation of extensive field trials (plots with or without replicates) to be conducted. In these cases, the information of the AT could significantly improve the yield predictive ability, thus saving resources (water, land, seed availability, time, money, etc.).

5. Conclusions

In this study, we compared different methods for effectively including the information of associated traits for yield prediction and for predicting the AT as well under different cross-validation schemes using the single-trait and multi-trait approaches. The multi-trait model was shown to improve the performance of the conventional single-trait model but only when the information from the AT was available for those genotypes in the testing set. However, the predictive ability was reduced in comparison with the single-trait baseline model when predicting completely untested genotypes. As was expected, under the multi-trait fashion, the best results were obtained when there was available complete or partial information of the AT; however, we observed that a more efficient use of this information could be accomplished under the single-trait parameterization by combining the complete information of the AT with marker SNPs via covariance structures.

The implications of these results are that (i) the single-trait approach is much easier to implement than the multi-trait model and since it returned the best results, we recommend its use for yield prediction; and (ii) the convenience of using the information of correlated traits is evident when some of these traits can be measured easier than grain yield or do not require extensive field trials.

Finally, one of the disadvantages of using the information of correlated traits for effectively improving yield predictive ability of genotypes is that these must be partially tested in order to measure the AT. This fact poses challenges to fully implement the alternative single-trait or the conventional multi-trait strategies to leverage the availability of the information on the AT for those genotypes in the training set. Perhaps, a solution would consider predicting the missing values of the AT of those genotypes in the testing set using the single-trait approach first, and then combine the predicted and the observed values to perform predictions of yield records under the alternative single-trait or multi-trait approach as a second step. This is still an area of development, and we expect that future research will consider/address these challenges.

Author Contributions: R.P. conducted integrated the data sets, performed the data analysis and the results, and wrote the first draft of the manuscript. A.B. contributed to all sections, commented on the applications of the different prediction strategies and their implications in breeding programs, and wrote the first draft of the document. D.J. conceptualized the study, supervised the data analysis, results, and contributed to the first draft. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Agriculture and Food Research Initiative Grant number NEB-21-176 from the USDA National Institute of Food and Agriculture, Plant Health and Production and Plant Products: Plant Breeding for Agricultural Production, A1211, Accession No.1015252. And the APC was funded by NEB-21-176.

Acknowledgments: We are thankful to two anonymous reviewers and the Assigned Editor for their valuable comments, suggestions, and positive criticisms to a previous version of the manuscript.

Conflicts of Interest: The 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.

References

1. Meuwissen, T.H.E.; Hayes, B.J.; Goddard, M.E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **2001**, *157*, 1819–1829.
2. Jannink, J.L.; Lorenz, A.J.; Iwata, H. Genomic selection in plant breeding: From theory to practice. *Brief. Funct. Genom.* **2010**, *9*, 166–177. [CrossRef] [PubMed]
3. Crossa, J.; Pérez-Rodriguez, P.; Cuevas, J.; Montesinos-López, O.; Jarquín, D.; de los Campos, G.; Burgueno, J.; Gonzalez-Camacho, J.M.; Perez-Elizalde, S.; Beyene, Y.; et al. Genomic selection in plant breeding: Methods, models, and perspectives. *Trends Plant Sci.* **2017**, *22*, 961–975. [CrossRef] [PubMed]
4. Van Raden, P.M. Efficient methods to compute genomic predictions. *J. Dairy Sci.* **2008**, *91*, 4414–4423. [CrossRef] [PubMed]
5. De los Campos, G.; Hickey, J.M.; Pong-Wong, R.; Daetwyler, H.D.; Calus, M.L. Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* **2013**, *193*, 327–345. [CrossRef] [PubMed]
6. Xu, Y.; Li, P.; Zou, C.; Lu, Y.; Xie, C.; Zhang, X.; Prasana, B.M.; Olsen, M.S. Enhancing genetic gain in the era of molecular breeding. *J. Exp. Bot.* **2017**, *68*, 2641–2666. [CrossRef]
7. Heslot, N.; Yang, H.P.; Sorrells, M.E.; Jannink, J.L. Genomic selection in plant breeding: A comparison of models. *Crop Sci.* **2012**, *52*, 146–160. [CrossRef]
8. Rutkoski, J.; Benson, J.; Jia, Y.; Brown-Guedira, G.; Jannink, J.L.; Sorrels, M. Evaluation of genomic prediction methods for fusarium head blight resistance in wheat. *Plant Genome* **2012**, *5*, 51–61. [CrossRef]
9. Howard, R.; Carriquiry, A.L.; Beavis, W.D. Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. *G3 Genes Genomes Genet.* **2014**, *4*, 1027–1046. [CrossRef]
10. Crossa, J.; Beyene, F.; Kassa, S.; Pérez, P.; Hickey, J.M.; Chen, C.; de los Campos, G.; Burgueno, J.; Windhausen, V.S.; Bucler, E.; et al. Genomic prediction in maize breeding populations with genotyping-bysequencing. *G3 Genes Genomes Genet.* **2013**, *3*, 1903–1926. [CrossRef]
11. Zhang, J.; Song, Q.; Cregan, P.B.; Jiang, G.L. Genome-wide association study, genomic prediction and marker-assisted selection for seed weight in soybean (Glycine max). *Theor. Appl. Genet.* **2016**, *129*, 117–130. [CrossRef] [PubMed]
12. Endelman, J.B. Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* **2011**, *4*, 250–255. [CrossRef]
13. Heslot, N.; Jannink, J.L.; Sorrells, M.E. Perspectives for genomic selection applications and research in plants. *Crop Sci.* **2015**, *55*, 1–12. [CrossRef]
14. Crossa, J.; Pérez, P.; Hickey, J.; Burgueño, J.; Ornella, J.; Ceron-Rojas, J.; Zhang, X.; Dreisigacker, S.; Babu, R.; Li, Y.; et al. Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity* **2014**, *112*, 48–60. [CrossRef] [PubMed]
15. Poland, J.; Rutkoski, J. Advances and challenges in genomic selection for disease resistance. *Annu. Rev. Phytopathol.* **2016**, *54*, 79–98. [CrossRef]
16. Liu, C.; Sikumaran, S.; Jarquin, D.; Crossa, J.; Dreisigacker, S.; Sansaloni, C.; Reynolds, M. Comparison of array-and sequencing-based makers for genome-wide association mapping and genomic prediction in spring wheat. *Cropscience* **2020**. [CrossRef]
17. Jarquin, D.; Kocak, K.; Posadas, L.; Hyma, K.; Jedlicka, J. Genotyping by Sequencing for Genomic Prediction in a Soybean Breeding Population. *BMC Genom.* **2014**, *15*, 740. [CrossRef] [PubMed]
18. Jarquín, D.; Howard, R.; Graef, G.; Lorenz, A. Response surface analysis of genomic prediction accuracy values using quality control covariates in soybean. *Evol. Bioinform.* **2019**, *15*, 1–7. [CrossRef]

19. Guo, G.; Zhao, F.; Wang, Y.; Zhang, Y.; Du, L.; Su, G. Comparison of single-trait and multiple-trait genomic prediction models. *BMC Genet.* **2014**, *15*, 30. [CrossRef]
20. De los Campos, G.; Grüneberg, A. MTM Package. 2016. Available online: http://quantgen.github.io/MTM/vignette.html (accessed on 30 May 2020).
21. Bhatta, M.; Gutierrez, L.; Cammarota, L.; Cardozo, F.; Germán, S.; Gómez-Guerrero, B.; Pardo, M.F.; Lanaro, V.; Sayas, M.; Castro, A.J. Multi-trait Genomic Prediction Model Increased the Predictive Ability for Agronomic and Malting Quality Traits in Barley (*Hordeum vulgare* L.). *G3 Genes Genomes Genet.* **2020**, *10*, 1113–1124. [CrossRef]
22. Velazco, J.G.; Jordan, D.R.; Mace, E.S.; Hunt, C.H.; Malosetti, M.; van Eeuwijk, F.A. Genomic Prediction of Grain Yield and Drought-Adaptation Capacity in Sorghum Is Enhanced by Multi-Trait Analysis. *Front. Plant Sci.* **2019**, *10*, 997. [CrossRef] [PubMed]
23. Lynch, M.; Walsh, B. *Genetics and Analysis of Quantitative Traits*; Sinauer: Sunderland, MA, USA, 1998; Volume 1.
24. Neyhart, J.L.; Lorenz, A.J.; Smith, K.P. Multi-trait Improvement by Predicting Genetic Correlations in Breeding Crosses. *G3 Genes Genomes Genet.* **2019**, *9*, 3153–3165. [CrossRef]
25. Xavier, A.; Hall, B.; Hearst, A.A.; Cherkauer, K.A.; Rainey, K.M. Genetic Architecture of Phenomic-Enabled Canopy Coverage in Glycine max. *Genetics* **2019**, *206*, 1081–1089. [CrossRef]
26. Sun, J.; Rutkoski, J.E.; Poland, J.A.; Crossa, J.; Jannink, J.L.; Sorrels, M.E. Multitrait, Random Regression, or Simple Repeatability Model in High-Throughput Phenotyping Data Improve Genomic Prediction for Wheat Grain Yield. *Plant Genome* **2017**, *10*, 1–12. [CrossRef]
27. Moreira, F.F.; Hearst, A.A.; Cherkauer, K.A.; Rainey, K.M. Improving the efficiency of soybean breeding with high-throughput canopy phenotyping. *Plant Methods* **2019**, *15*, 139. [CrossRef] [PubMed]
28. Rutkoski, J.; Poland, J.; Mondal, S.; Autrique, E.; Párez, L.G.; Crossa, J.; Reynolds, M.; Singh, R. Canopy temperature and vegetation indices from high-throughput phenotyping improve accuracy of pedigree and genomic selection for grain yield in wheat. *G3 Genes Genomes Genet.* **2016**, *6*, 2799–2808. [CrossRef]
29. Howard, R.; Jarquin, D. Genomic prediction using canopy coverage image and genotypic information in soybean via a hybrid model. *Evol. Bioinform.* **2019**, *15*, 117693431984002. [CrossRef] [PubMed]
30. Jarquin, D.; Howard, R.; Xavier, A.; Das Choudhury, S. Increasing predictive ability by modeling interactions between environments, genotype and canopy coverage image data for soybeans. *Agronomy* **2018**, *8*, 51. [CrossRef]
31. Diers, B.W.; Specht, J.; Rainey, K.M.; Cregan, P.; Song, Q.; Ramasubramanian, V.; Graef, G.; Nelson, R.; Schapaugh, W.; Wang, D.; et al. Genetic Architecture of Soybean Yield and Agronomic Traits. *G3 Genes Genomes Genet.* **2018**, *8*, 3367–3375. [CrossRef] [PubMed]
32. Pérez, P.; de los Campos, G. Genome-wide regression and prediction with the BGLR statistical package. *Genetics* **2014**, *198*, 483–495. [CrossRef] [PubMed]
33. Li, Y.; Reif, J.C.; Hong, H.; Li, H.; Liu, Z.; Ma, Y.; Li, J.; Tian, Y.; Li, Y.; Li, W.; et al. Genome-wide association mapping of QTL underlying seed oil and protein contents of a diverse panel of soybean accessions. *Plant Sci.* **2018**, *266*, 95–101. [CrossRef] [PubMed]
34. Jia, Y.; Jannink, J.L. Multiple-trait genomic selection methods increase genetic value prediction accuracy. *Genetics* **2012**, *192*, 1513–1522. [CrossRef] [PubMed]
35. Kwon, S.H.; Torrie, J.H. Heritability of and Interrelationships among Traits of Two Soybean Populations. *Crop Sci.* **1964**, *4*, 196–198. [CrossRef]
36. Patil, G.; Mian, R.; Vuong, T.; Pantalone, V.; Song, Q.; Chen, P.; Shannon, G.J.; Carter, T.C.; Nguyen, H.T. Molecular mapping and genomics of soybean seed protein: A review and perspective for the future. *Theor. Appl. Genet.* **2017**, *130*, 1975–1991. [CrossRef] [PubMed]
37. Aguate, F.M.; Trachsel, S.; González-Pérez, L.; Burgueño, J.; Crossa, J.; Balzarini, M.; Gouache, D.; Bogard, M.; de los Campos, G. Use of hyperspectral image data outperforms vegetation indices in prediction of maize yield. *Crop Sci.* **2017**, *57*, 2517–2524. [CrossRef]
38. Montesinos-López, O.A.; Montesinos-López, A.; Crossa, J.; de los Campos, G.; Alvarado, G.; Suchismita, M.; Rutkoski, J.; Gonzalez-Perez, L.; Burgeno, J. Predicting grain yield using canopy hyperspectral reflectance in wheat breeding data. *Plant Methods* **2017**, *13*, 4. [CrossRef] [PubMed]