applied system innovation

MDPI

Article

Leveraging Explainable Artificial Intelligence for Genotype-to-Phenotype Prediction: A Case Study in *Arabidopsis thaliana*

Pierfrancesco Novielli [1,2,*], Nelson Nazzicari [3], Stefano Pavan [1,*], Chiara Delvento [1], Domenico Diacono [2], Claudia Zoani [4], Roberto Bellotti [2,5] and Sabina Tangaro [1,2,*]

[1] Dipartimento di Scienze del Suolo, della Pianta e degli Alimenti, Università degli Studi di Bari Aldo Moro, 70125 Bari, Italy; chiara.delvento@uniba.it
[2] Istituto Nazionale di Fisica Nucleare, Sezione di Bari, 70125 Bari, Italy; domenico.diacono@ba.infn.it (D.D.); roberto.bellotti@uniba.it (R.B.)
[3] CREA—Council for Agricultural Research and Analysis of Agricultural Economics, Research Centre for Animal Production and Aquaculture, 26900 Lodi, Italy; nelson.nazzicari@crea.gov.it
[4] Dipartimento Sostenibilità, Circolarità e Adattamento al Cambiamento Climatico dei Sistemi Produttivi e Territoriali, Divisione Biotecnologie e Agroindustria—ENEA, C.R. Casaccia, 00123 Roma, Italy; claudia.zoani@enea.it
[5] Dipartimento Interateneo di Fisica "M. Merlin", Università degli Studi di Bari Aldo Moro, 70125 Bari, Italy
* Correspondence: pierfrancesco.novielli@uniba.it (P.N.); stefano.pavan@uniba.it (S.P.); sabina.tangaro@uniba.it (S.T.)

Abstract

Predicting phenotypes from genomic data can significantly advance agriculture. Genomic selection, which uses genome-wide DNA markers to identify individuals with high genetic value, enhances the accuracy of breeding programs. While linear models are routinely used for genomic selection (GS), machine learning (ML) models offer complementary potential. In this study, robust ML-based models were developed to predict five phenotypic traits—three related to flowering time and two to leaf number—in *Arabidopsis thaliana*, a model plant with a fully sequenced genome. Using explainable artificial intelligence (XAI), specifically SHapley Additive exPlanations (SHAP) values, we identified SNPs that contributed most to trait prediction. Many of these SNPs were located in or near genes known to regulate flowering and stem elongation, such as DOG1 and VIN3, supporting the biological plausibility of the model. SHAP also enabled local interpretability at the single-plant level, revealing the genotypic basis of individual predictions. Our results indicate that integrating ML with XAI improves model interpretability and provides predictive performance comparable to traditional methods. This approach confirms known genotype–phenotype relationships and highlights new candidate loci, paving the way for functional validation. The proposed methodology offers promising applications in precision breeding and translation of insights from *Arabidopsis* to crop species.

Keywords: genotype-to-phenotype prediction; explainable artificial intelligence; machine learning; SHAP; *Arabidopsis thaliana*; regression analysis

Academic Editor: Charles Tijus

Received: 13 August 2025
Revised: 1 October 2025
Accepted: 14 October 2025
Published: 27 October 2025

Citation: Novielli, P.; Nazzicari, N.; Pavan, S.; Delvento, C.; Diacono, D.; Zoani, C.; Bellotti, R.; Tangaro, S. Leveraging Explainable Artificial Intelligence for Genotype-to-Phenotype Prediction: A Case Study in *Arabidopsis thaliana*. *Appl. Syst. Innov.* **2025**, *8*, 164. https://doi.org/10.3390/asi8060164

1. Introduction

Predicting phenotypes from genomic information is a pivotal challenge in modern breeding, which needs to enhance cultivars for complex traits, including adaptation to climate changes, while enhancing resource use efficiency. Genomic selection (GS) has

emerged as a promising approach, enabling breeders to select plants with desired traits directly from genotypic data. GS uses genome-wide markers to forecast phenotypes or breeding values, as initially proposed by Meuwissen et al. [1]. State-of-the-art methods such as the Best Linear Unbiased Predictor (BLUP) and its variants, including genomic BLUP (GBLUP) and Ridge Regression BLUP (RR-BLUP), have become central to predicting additive genetic effects in plant breeding across diverse species and traits [2,3]. However, as breeding challenges become more complex, particularly with non-additive genetic effects, machine learning (ML) techniques have emerged to complement these traditional methods [4–7]. Recently, transformer-based architectures and multi-task learning have also shown promise for crop genomic prediction. In particular, models such as Cropformer and MtCro have reported improved ability to capture complex genomic patterns and to enhance prediction across multiple traits [8–10].

Ensemble-based methods like Random Forests (RFs) and gradient boosting techniques such as XGBoost (XGB) and CatBoost have gained traction for their ability to capture complex genetic architectures and improve prediction accuracy [11–13]. Studies indicate that ML algorithms can enhance accuracy in genomic selection [14–16], making it possible to utilize intricate and complex data.

Moreover, to make these sophisticated ML models more interpretable and biologically meaningful, combining ML with explainable artificial intelligence (XAI) has gained traction, offering the potential to illuminate the genetic mechanisms underlying trait prediction [17,18]. XAI techniques aim to augment the interpretability of ML models, providing transparency and insights into the biological bases underlying genotype-to-phenotype predictions [19].

XAI comprises methods that make complex models transparent by attributing predictions to input features at both the global (model-level) and local (sample-level) scales. Among these, SHAP (SHapley Additive exPlanations) provides theoretically grounded, additive attributions for individual predictions and has become a widely adopted tool for interpreting black-box models [20]. XAI has been increasingly applied across the life sciences to bridge predictive accuracy with biological insight: from clinical decision support and healthcare model design [21,22] to omics-driven discovery where interpretable attributions help prioritize variants, genes, and pathways for hypothesis generation [23]. Building on this literature, our study leverages SHAP to connect predictive performance with mechanistic clues in genomic prediction, enabling trait- and plant-specific interpretations.

In the context of crop improvement, GS has revolutionized the association of Single-Nucleotide Polymorphisms (SNPs) with breeding values, expediting breeding cycles and enhancing genetic gains in crops [3,24]. Despite ongoing research, the biological mechanisms driving genotype-to-phenotype predictions remain elusive. While integrating associated SNPs through supervised feature selection techniques into prediction models has proven promising, outcomes have been varied. Consequently, there is a growing need to explore non-linear models and leverage data from feature selection algorithms to refine predictions [25,26]. Identifying critical loci is crucial not only for model construction but also for data reduction through feature selection in subsequent iterations.

In this study, feature selection was applied not only to optimize model accuracy but also to reduce data dimensionality. This is particularly important given the high number of SNPs relative to the number of samples, as it reduces computational time and mitigates the risk of overfitting, all while maintaining model interpretability within the context of XAI.

Exploring a variety of ML algorithms is advisable when predicting genotype-to-phenotype relationships, as each algorithm operates based on unique assumptions and biases. No single algorithm universally delivers optimal performance across all traits [27]. In this study, robust prediction models were developed for five phenotypic traits in *Ara-*

bidopsis thaliana, a model plant for genomic studies for which high-quality genomes and phenotypes of more than 1000 genotypes are publicly available.

By employing explainable artificial intelligence, specifically SHapley Additive exPlanations (SHAP) values, we enhance interpretability in genomic prediction models. SHAP values provide insights into the contribution of each feature to the model's predictions, elucidating the underlying mechanisms driving genotype-to-phenotype predictions. By offering local explanations, SHAP enables precise, plant-specific insights, identifying the most relevant SNPs for each plant. This capability is particularly valuable for precision breeding, where decisions must be tailored to individual plants.

While previous studies have explored the application of machine learning models for genotype-to-phenotype prediction, our work contributes to the field by systematically integrating SHAP-based explainability with ensemble methods across multiple traits in *Arabidopsis thaliana*. This integrated approach, applied to a large-scale dataset from the 1001 Genomes Project, enables trait-specific interpretation at the individual plant level, which is essential for precision breeding.

This comprehensive background sets the stage for exploring genomic prediction methodologies in *Arabidopsis thaliana*, highlighting the intersection of machine learning and traditional breeding practices in crop improvement endeavors.

2. Materials and Methods

In Figure 1, the workflow of the analysis is depicted. We utilized publicly available SNP genotype data, integrating it with various phenotypic variables to predict these variables based on genetic factors. The initial step involved preprocessing the SNP data. The processed data were then fed into regression algorithms to obtain phenotype predictions. The results obtained from mixed linear models were compared with those derived from an XAI-based approach. Specifically, for the XAI approach, additional steps were taken to interpret the results obtained.

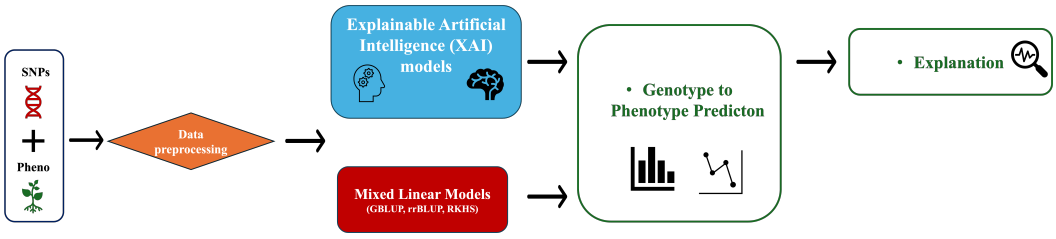

Figure 1. Workflow of the analysis. The process involves data preprocessing, regression using mixed linear models and XAI models, genotype-to-phenotype prediction, and explanation of the predictions.

2.1. Materials

The SNP genotypic data used in our study pertains to the model plant species *Arabidopsis thaliana*, for which extensive genomic and phenotypic information is publicly available [28]. Specifically, we used the SNP matrix file of the 1001 Genomes Consortium [29,30], which contains about 10 million variants and is filtered for heterozygous loci that might introduce bias due to duplicated regions [31]. This dataset encompasses genetic information from 1135 *A. thaliana* accessions selected from a globally representative hierarchical collection [32]. The genetic data consisted of approximately 10 million biallelic SNPs that had undergone quality control filtering.

Regarding the phenotypic variables targeted for prediction, five traits were considered (Table 1), namely Days Until Emergence of Visible Flowering Buds in the Center of the Rosette from Time of Sowing (DTF1), Days Until the Inflorescence Stem Elongated to 1 cm (DTF2), Days Until First Open Flower (DTF3), Rosette Leaf Number (RL), and Cauline Leaf

Number (CL). Table 1 also provides the number of accessions available for each trait. Plants were grown under controlled conditions, as specified in Table 1. For the phenotypic traits analyzed, not all 1135 accessions were available. As shown in Table 1, each phenotype was measured on a specific subset: 936 for DTF1, 931 for DTF2, 923 for DTF3, 850 for RL, and 904 for CL. The machine learning models were trained and validated on these subsets, using only accessions with complete phenotype data.

Table 1. Overview of the phenotypic traits selected for this study. Growth conditions: Seeds from 1135 *Arabidopsis thaliana* accessions (1001 Genomes Consortium, 2016) were surface-sterilized in 95% ethanol for 5 min and air-dried. After 6 days of stratification at 4 °C in 0.1% agarose, seeds were distributed across 4800 pots as four replicates in a randomized block design. Plants were grown in controlled chambers (16 h light/8 h dark, 16 °C constant temperature, 65% humidity). Trays were rotated every other day to minimize position effects.

ID	Phenotypic Trait	Abbreviation	Number of Accessions	URL
1	Days until emergence of visible flowering buds in the center of the rosette from time of sowing	DTF1	936	https://arapheno.1001genomes.org/phenotype/703/ (accessed on 4 April 2024)
2	Days until the inflorescence stem elongated to 1 cm	DTF2	931	https://arapheno.1001genomes.org/phenotype/701/ (accessed on 4 April 2024)
3	Days until first open flower	DTF3	923	https://arapheno.1001genomes.org/phenotype/702/ (accessed on 4 April 2024)
4	Rosette leaf number	RL	850	https://arapheno.1001genomes.org/phenotype/704/ (accessed on 4 April 2024)
5	Cauline leaf number	CL	904	https://arapheno.1001genomes.org/phenotype/705/ (accessed on 4 April 2024)

2.2. Methods

2.2.1. Preprocessing

SNP data underwent preprocessing to enhance quality and informativeness, ensuring suitability for accurate and robust downstream genotype–phenotype prediction analyses (Figure 2). In more detail, the input Variant Call Format (VCF) file containing 10,709,949 SNPs was initially filtered using the software PLINK 1.9 [33]. Variants with a minor allele frequency (MAF) of less than 0.05 were removed from the dataset to filter out rare variants and reduce their impact on subsequent analyses [34]. Subsequently, SNPs were subjected to linkage disequilibrium (LD) pruning using PLINK 1.9, a tool for whole-genome association analysis. At each step of LD pruning, pairs of variants within the current window were assessed for LD, with variants having an R^2 greater than the threshold of 0.8 being identified. This threshold was chosen to minimize collinearity between SNPs while retaining informative markers. The greedy pruning approach employed by PLINK 1.9 ensures that we systematically reduce LD within a window size, further optimizing the feature set for model training [35,36].

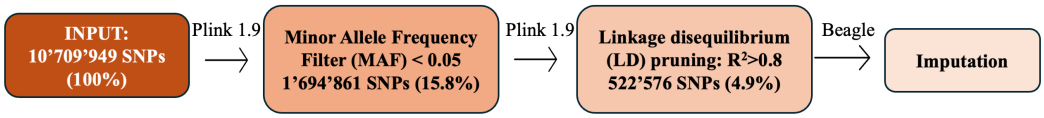

Figure 2. SNP data preprocessing workflow. The figure illustrates the preprocessing steps applied to the SNP dataset. The input consisted of 10.7 M SNPs. After applying a minor allele frequency (MAF) filter (<0.05), 1.69 M SNPs remained (15.8% of the initial dataset). Linkage disequilibrium (LD) pruning with $R^2 > 0.8$ further reduced the set to 522 K SNPs (4.9% of the initial dataset). The final step involves imputation to handle missing genotype data.

Lastly, for handling missing data, genotype imputation was conducted using Beagle 5.1, a state-of-the-art imputation tool. Beagle uses phased haplotypes to accurately impute missing genotypes by leveraging population-level linkage patterns. This ensures the completeness of the genetic data without biasing model training due to incomplete genotypes [37,38].

After the completion of the genotype data preprocessing steps, the following encoding of genomic data was conducted for subsequent analyses:

- "0/0", representing genotypes homozygous for the reference allele (REF), were encoded as "0";
- "1/1", indicating genotypes homozygous for the alternate allele (ALT), were encoded as "1".

2.2.2. Regression Analysis

In this study, we compared the results of various regression models to predict each phenotypic trait from genotypic data. We evaluated both state-of-the-art genomic prediction models, represented by mixed linear regressors, and machine learning-based models. The validation procedure, depicted in Figure 3, was designed to measure the performance of the models accurately.

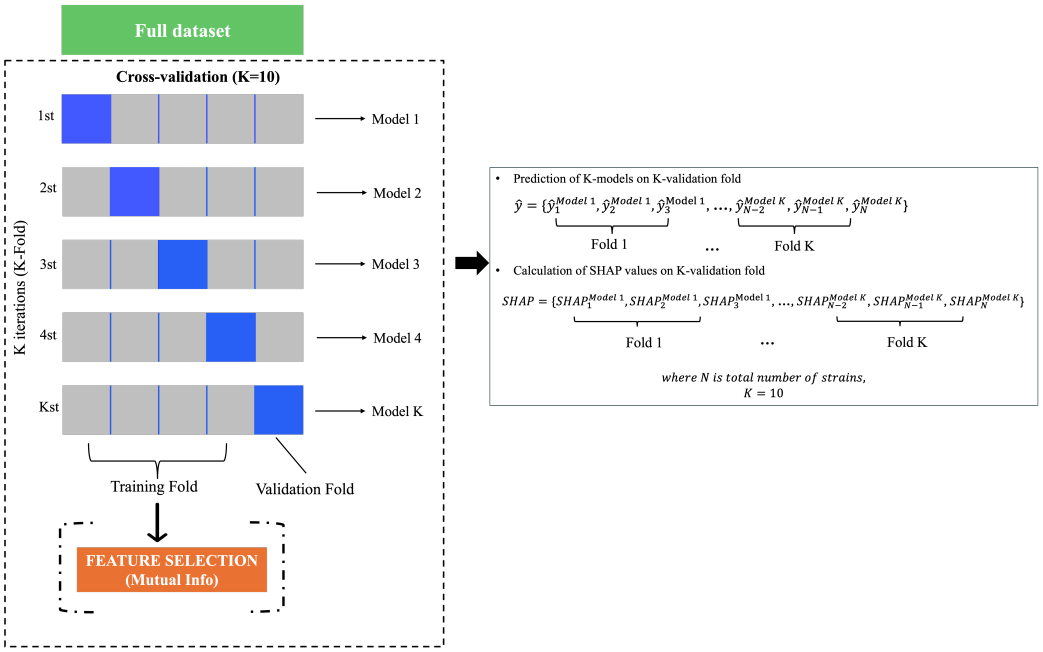

Figure 3. Workflow of the validation procedure. The full dataset was divided into 10 folds for cross-validation. In each iteration, 9 folds were used for training and 1 fold was used for validation. Feature selection was performed within each training fold to prevent data leakage. The SHAP values were then calculated for each model across the validation folds, providing insights into the contribution of each feature to the model's predictions. This comprehensive validation approach ensured the reliability and generalizability of the predictive models.

To evaluate the performance of the models, a 10-fold cross-validation (K = 10) was adopted. Ten different models were trained, and predictions were calculated for each cycle of the cross-validation on the remaining validation fold. The machine learning-based approach also involved feature selection algorithms due to the necessity of reducing the number of features, as an excessive number of variables compared to the number of accessions could lead to overfitting [39]. The primary goal of feature selection was not solely to enhance predictive performance but also to reduce the dimensionality of the dataset, improving computational efficiency and minimizing the risk of overfitting. Given the large number of SNPs compared to samples, feature selection is critical for ensuring interpretability, particularly when using SHAP for explainable artificial intelligence (XAI). To ensure integrity and independence of the test set, feature selection was nested within the cross-validation to prevent data leakage. Performing feature selection on all of the data and then cross-validating could bias the performance analysis since the test data in each fold of the CV procedure would also be used to choose the features [40,41]. Thus, feature selection was applied to each split of the CV, followed by training an ML regression model.

For the computation of SHAP values aimed at providing an explanation of results, the SHAP value for each strain was calculated by combining the SHAP values across the ten models trained in cross-validation. This comprehensive validation approach ensured the reliability and generalizability of the predictive models developed in this study.

2.2.3. Mixed Linear Models

Ridge Regression BLUP (rrBLUP) assumes a linear mixed additive model where each marker is assigned an effect as a solution of the equation

$$y = \mu + Gu + \epsilon \tag{1}$$

where y is the array of observed phenotypes with mean value μ, where G is the genotype matrix with values in [0, 1, 2] (for diploid SNP markers), $u \sim N(0, I\sigma_u^2)$ contains the marker effects to be estimated, and ϵ contains the residuals. The standard ridge regression solution becomes

$$\hat{u} = G'(GG' + \lambda I)^{-1}(y - \mu) \tag{2}$$

with $\lambda = \sigma_e^2/\sigma_u^2$ being the ridge parameter, i.e., the ratio between residual and marker variance. We obtained the estimated allele effects by solving the above equations on the training set. We then used $\hat{u}$ to predict $\hat{y}$ for the samples in the validation set. Cross-validations, model solutions and predictions, and performance metrics were computed using the R package GROAN [6], which internally leveraged the rrBLUP package [42].

2.2.4. XAI Models

Feature Selection

Feature selection plays a pivotal role in data preprocessing as it determines the essential features for analysis. Not only does it eliminate irrelevant features, but it also identifies the most significant ones, thereby enhancing the model's performance. In our work, we employed a feature selection technique based on Mutual Information (MI) Gain [43–45]. This method, a univariate filtering approach, offers improved accuracy to the model [46]. MI estimates the mutual information for continuous target variables in regression problems (phenotypic traits), leveraging the entropy of the variables (SNPs). It quantifies the dependency between variables, with higher values indicating stronger dependency. Essentially, MI measures the information one variable provides about another. Like other feature selection techniques, its goal is to reduce the size of the input feature set. This reduction

in features can simplify the problem, reduce computational time, and potentially enhance model performance.

For each studied phenotypic trait, we analyzed its Manhattan plot. We then selected only SNPs with mutual information above the 0.90 quantile. To test the robustness of this choice, we further evaluated alternative thresholds (0.80, 0.85, 0.95, 0.975, and 0.99), and the results are reported in Supplementary Table S2.

Machine Learning Regressors

Ensemble learning has emerged as a highly popular option for regression tasks in recent years. Among the various choices available, Random Forest (RF) has garnered widespread adoption [47]. Random Forests are built by bootstrapping the data sample and growing multiple regression trees, each utilizing a different bootstrap sample from the original dataset. Notably, RF introduces an additional layer of randomness compared to bagging strategies, as each tree is grown with a different set of predictors randomly selected at each node split. One of the primary advantages of RF over traditional regression strategies lies in its robustness against overfitting. This characteristic makes it particularly valuable in practical applications where generalization is crucial. Additionally, RF offers the advantage of providing insights into feature importance, thus facilitating the interpretation of model results. This capability is especially valuable in domains such as clinical applications, where understanding the relationship between independent and dependent variables is essential. For this study, RF regression was implemented using the default parameters of the scikit-learn package.

Gradient boosting is a class of ensemble machine learning algorithms widely used for regression predictive modeling tasks. These ensembles are built from decision tree models, with trees added sequentially to the ensemble and trained to correct prediction errors made by prior models. This approach, known as boosting, enhances the predictive power of the ensemble. XGBoost, short for Extreme Gradient Boosting, represents an efficient open-source implementation of the gradient boosting algorithm. It stands out for its computational efficiency and effectiveness, often surpassing other open-source implementations in terms of performance. XGBoost is designed to optimize both execution speed and model accuracy, making it a preferred choice for many regression tasks [48]. In this study, XGBoost regression was employed using the default parameters of the xgboost Python package, ensuring consistency and ease of implementation in the analysis pipeline.

CatBoost, a recently introduced open-source machine learning tool, represents a robust algorithm within the realm of gradient boosted decision trees. It has demonstrated promising outcomes across various standard machine learning tasks [49]. One of the distinguishing features of the CatBoost model is its utilization of ordered boosting and its adept handling of categorical features during the training process, marking a significant advancement compared to conventional gradient boosted decision trees. Notably, CatBoost has showcased superior performance when dealing with non-numeric or categorical features, outperforming other gradient boosting algorithms [50]. For this study, we employed the default parameters of the CatBoost Python package.

Explainability

To conduct the explainable artificial intelligence analysis and ensure interpretation and transparency of the results, we computed SHAP (SHapley Additive exPlanations) values. The calculation of SHAP facilitates the interpretation of complex machine learning models by providing insights into the contribution of individual features to model predictions. SHAP values offer explanations for model predictions by utilizing classic equations derived from cooperative game theory [51]. These values quantify the degree of interaction between

features and the predicted values, providing insights into the impact of each feature on the overall predictions. The computation of SHAP values involves evaluating the difference in model output predictions with and without specific features, considering all possible feature subsets. Consequently, the model necessitates retraining on all subsets F of the complete set S of features ($F \subseteq S$). The SHAP value for the $j-th$ feature of the instance x is determined by aggregating it across all possible subsets:

$$SHAP_j(x) = \sum_{F \subseteq S-\{j\}} \frac{|F|!(|S|-|F|-1)!}{|S|!}[f_x(F \cup j) - f_x(F)] \quad (3)$$

where $|F|!$ represents the permutations of features in the subset F, $(|S|-|F|-1)!$ represents the permutations of features in the subset $S-(F \cup \{j\})$, $|S|!$ is the total number of feature permutations, and $f_x(F \cup j)$ and $f_x(F)$ denote, respectively, the regression score obtained by including and not including the *j-th* feature.

To prevent data leakage, feature selection was nested within the cross-validation process. For the SHAP-based interpretability analysis, we computed SHAP values using the subset of features that were consistently selected across all folds. This strategy ensures that the interpretation reflects robust and recurrent patterns rather than fold-specific artifacts, while also improving computational efficiency by reducing the dimensionality of the explanation space.

2.2.5. Evaluation Metrics

We assessed the performance of the regression models by computing the following metrics for each algorithm:

- Pearson correlation:

$$corr = \frac{\sum_{i=1}^{n}(y_i - m_y)(\hat{y}_i - m_{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - m_y)^2 \sum_{i=1}^{n}(\hat{y}_i - m_{\hat{y}})^2}} \quad (4)$$

- Coefficient of determination:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - m_y)^2} \quad (5)$$

- Root mean squared error:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (6)$$

where $\hat{y}_i$ is the predicted values, $m_{\hat{y}}$ denotes their average, y_i is the observed values of phenotypic traits, and m_y denotes their average. These metrics were computed for each algorithm to allow a fair comparison across models. All metrics were averaged across the 10 folds of cross-validation to ensure robust and unbiased estimates of model performance.

All machine learning models (Random Forest, XGBoost, and CatBoost) were implemented in Python (v.3.12) using the scikit-learn, xgboost, and catboost libraries, respectively. In addition, rrBLUP analyses were run in R using the GROAN/rrBLUP packages (see Section 2.2.3). Model explainability was performed with the `shap` library (TreeExplainer) to compute SHAP values and generate summary plots.

3. Results

The aim of this study was to implement regression models predicting phenotypic traits exclusively from genotypic data. Our focus lies on the outcomes of the artificial intelligence workflow, as we examined the regression performance of various machine learning algorithms alongside traditional algorithms and evaluated the biological significance of the genomic position of SNP loci exerting the most significant influence on predictions.

Following the procedures outlined in the Methods Section, after preprocessing the SNPs and performing feature selection to reduce the variables from approximately 5×10^5 to 5×10^4, we applied different regression models for each of the following phenotypic traits: DTF1, DTF2, DTF3, CL, and RL (Table 1). The results of the predictive models are presented in Table 2, in which the models are compared based on the R^2, Pearson coefficient, and RMSE. The values reported in the tables represent the mean and standard deviation obtained through cross-validation.

Table 2. Evaluation metrics for five phenotypic traits: Days to Flowering Time 1 (DTF1), Days to Flowering Time 2 (DTF2), Days to Flowering Time 3 (DTF3), Cauline Leaf Number (CL), and Rosette Leaf Number (RL). The table compares the performance of different regression models—CatBoost, Random Forest (RF), and XGBoost—with and without feature selection (Feat sel) and Ridge Regression BLUP (rrBLUP) without feature selection. Performance metrics include Pearson Correlation, Coefficient of Determination (R^2), and Root Mean Squared Error (RMSE), with standard deviations calculated through cross-validation. Bold values indicate the best-performing model for each metric within each phenotypic trait.

Trait	Regressor	Pearson Correlation	R^2	RMSE
DTF1	CatBoost	0.830 ± 0.034	0.682 ± 0.056	14.60 ± 1.05
	CatBoost + Feat sel	$\mathbf{0.833 \pm 0.034}$	$\mathbf{0.688 \pm 0.057}$	$\mathbf{14.45 \pm 1.11}$
	RF	0.798 ± 0.037	0.618 ± 0.055	16.05 ± 0.96
	RF + Feat sel	0.802 ± 0.038	0.627 ± 0.057	15.84 ± 1.07
	XGBoost	0.780 ± 0.036	0.605 ± 0.056	16.30 ± 0.91
	XGBoost + Feat sel	0.778 ± 0.042	0.602 ± 0.065	16.37 ± 1.08
	rrBLUP	0.830 ± 0.038	0.678 ± 0.061	14.69 ± 1.16
DTF2	CatBoost	0.824 ± 0.037	0.672 ± 0.063	14.79 ± 1.38
	CatBoost + Feat sel	$\mathbf{0.831 \pm 0.032}$	$\mathbf{0.683 \pm 0.056}$	$\mathbf{14.55 \pm 1.32}$
	RF	0.788 ± 0.036	0.603 ± 0.054	16.33 ± 1.16
	RF + Feat sel	0.794 ± 0.036	0.617 ± 0.056	16.04 ± 1.19
	XGBoost	0.756 ± 0.047	0.562 ± 0.073	17.12 ± 1.24
	XGBoost + Feat sel	0.772 ± 0.026	0.588 ± 0.042	16.69 ± 1.15
	rrBLUP	0.826 ± 0.036	0.675 ± 0.061	14.74 ± 1.28
DTF3	**CatBoost**	$\mathbf{0.824 \pm 0.020}$	$\mathbf{0.672 \pm 0.031}$	14.80 ± 0.75
	CatBoost + Feat sel	0.823 ± 0.024	0.672 ± 0.038	$\mathbf{14.79 \pm 0.77}$
	RF	0.786 ± 0.025	0.598 ± 0.034	16.41 ± 0.91
	RF + Feat sel	0.796 ± 0.023	0.615 ± 0.032	16.05 ± 0.86
	XGBoost	0.755 ± 0.017	0.568 ± 0.027	17.02 ± 0.80
	XGBoost + Feat sel	0.745 ± 0.019	0.550 ± 0.030	17.35 ± 0.60
	rrBLUP	0.823 ± 0.028	0.671 ± 0.040	14.81 ± 0.92
CL	CatBoost	0.658 ± 0.047	0.423 ± 0.070	3.42 ± 0.38
	CatBoost + Feat sel	0.660 ± 0.049	0.425 ± 0.075	3.41 ± 0.39
	RF	0.654 ± 0.051	0.406 ± 0.068	3.47 ± 0.41
	RF + Feat sel	0.653 ± 0.050	0.409 ± 0.070	3.46 ± 0.40
	XGBoost	0.613 ± 0.076	0.354 ± 0.126	3.61 ± 0.44
	XGBoost + Feat sel	0.590 ± 0.060	0.315 ± 0.101	3.71 ± 0.35
	rrBLUP	$\mathbf{0.680 \pm 0.049}$	$\mathbf{0.449 \pm 0.072}$	$\mathbf{3.34 \pm 0.41}$

Table 2. *Cont.*

Trait	Regressor	Pearson Correlation	R^2	RMSE
RL	**CatBoost**	**0.743 ± 0.046**	**0.541 ± 0.071**	**11.23 ± 0.79**
	CatBoost + Feat sel	0.736 ± 0.046	0.532 ± 0.070	11.34 ± 0.68
	RF	0.714 ± 0.050	0.480 ± 0.068	11.97 ± 0.74
	RF + Feat sel	0.715 ± 0.047	0.487 ± 0.066	11.88 ± 0.75
	XGBoost	0.678 ± 0.045	0.443 ± 0.080	12.40 ± 1.03
	XGBoost + Feat sel	0.666 ± 0.076	0.425 ± 0.125	12.52 ± 0.89
	rrBLUP	0.743 ± 0.056	0.540 ± 0.085	11.25 ± 1.02

Among the ML models, CatBoost outperformed both Random Forest and XGBoost across all the traits under investigation. The CatBoost model, whether used alone or with preceding feature selection, showed performance that was generally comparable to standard parametric linear models across all traits considered. Therefore, CatBoost presents itself as a viable alternative to rrBLUP. Additionally, the integration of XAI with ML models enhances the interpretability of prediction results.

Feature selection resulted in modest improvements for certain phenotypic traits, such as DTF1, DTF2, and CL. More importantly, despite the reduction in features, the performance of the models remained stable across all traits, demonstrating that feature selection contributed to computational efficiency without compromising accuracy.

We further assessed the robustness of feature selection by varying the mutual information percentile threshold (0.80, 0.85, 0.90, 0.95, 0.975, and 0.99) within a CatBoost pipeline. As reported in Supplementary Table S2, performance remained stable across thresholds for all traits (Pearson, R^2, RMSE within overlapping standard deviations), supporting the choice of the 0.90 quantile as a parsimonious operating point that balances predictive performance with computational efficiency by limiting the number of features entering subsequent analyses.

A closer inspection of the results by trait confirms this trend (Table 2). For DTF1, CatBoost with feature selection achieved the best values across all metrics (Pearson 0.833 ± 0.034, $R^2 = 0.688 \pm 0.057$, and RMSE 14.45 ± 1.11), with rrBLUP performing very closely. For DTF2, CatBoost with feature selection again led for all metrics (Pearson 0.831 ± 0.032, $R^2 = 0.683 \pm 0.056$, and RMSE 14.55 ± 1.32). For DTF3, CatBoost (with and without feature selection) provided the top Pearson and R^2 values (≈ 0.824 and ≈ 0.672), with the lowest RMSE obtained by CatBoost with feature selection (14.79 ± 0.77). For CL, rrBLUP yielded the best performance on all three metrics (Pearson 0.680 ± 0.049, $R^2 = 0.449 \pm 0.072$, and RMSE 3.34 ± 0.41). Finally, for RL, CatBoost reached the top values (Pearson 0.743 ± 0.046, $R^2 = 0.541 \pm 0.071$, and RMSE 11.23 ± 0.79).

Additionally, to provide further insights into the performance and interpretability of the CatBoost regressor, we present visualizations in Figures 4 and 5. Figures 4a–c and 5a,b depict scatter plots of the analyzed phenotypic traits, with actual values on the x-axis (y_i) and predicted values on the y-axis ($\hat{y}_i$). Notably, a close adherence of predicted values to the actual ones is observed, especially for the days to flowering traits. Except for RL, where Catboost without feature selection performed better, for all other traits, the combination of Catboost preceded by feature selection obtained better results. Therefore, in Figure 4a–c and in Figure 5c, the results related to the Catboost model preceded by feature selection are represented, whereas in Figure 5d, the scatter plot of the Catboost model without feature selection is represented.

The scatter plots in Figures 4 and 5 show tight alignment of predicted vs. measured values for the three flowering time traits, consistent with their higher heritability, and a wider spread for leaf number traits.

Furthermore, Figures 4d–f and 5c,d present the results of feature importance based on SHAP values calculated by Equation (3). SHAP values provide insights into how specific features influence predictions, with positive and negative values indicating contributions towards high or low values of the phenotypic variable under regression, respectively. Each data point in each row of the summary plot represents the SHAP value of that particular SNP for a specific strain. Higher absolute SHAP values denote greater feature relevance in the prediction. The top 20 SNPs are represented in descending order of importance.

Moreover, from the analysis of SHAP values, further insight into the direction of effects can be deduced. Certain SNPs with ALT alleles (highlighted in red points) are associated with higher phenotypic values, while those with REF alleles (highlighted in blue points) are linked to lower phenotypic values. Conversely, some SNPs exhibit the opposite pattern, suggesting a correlation between alternative alleles and lower phenotypic values.

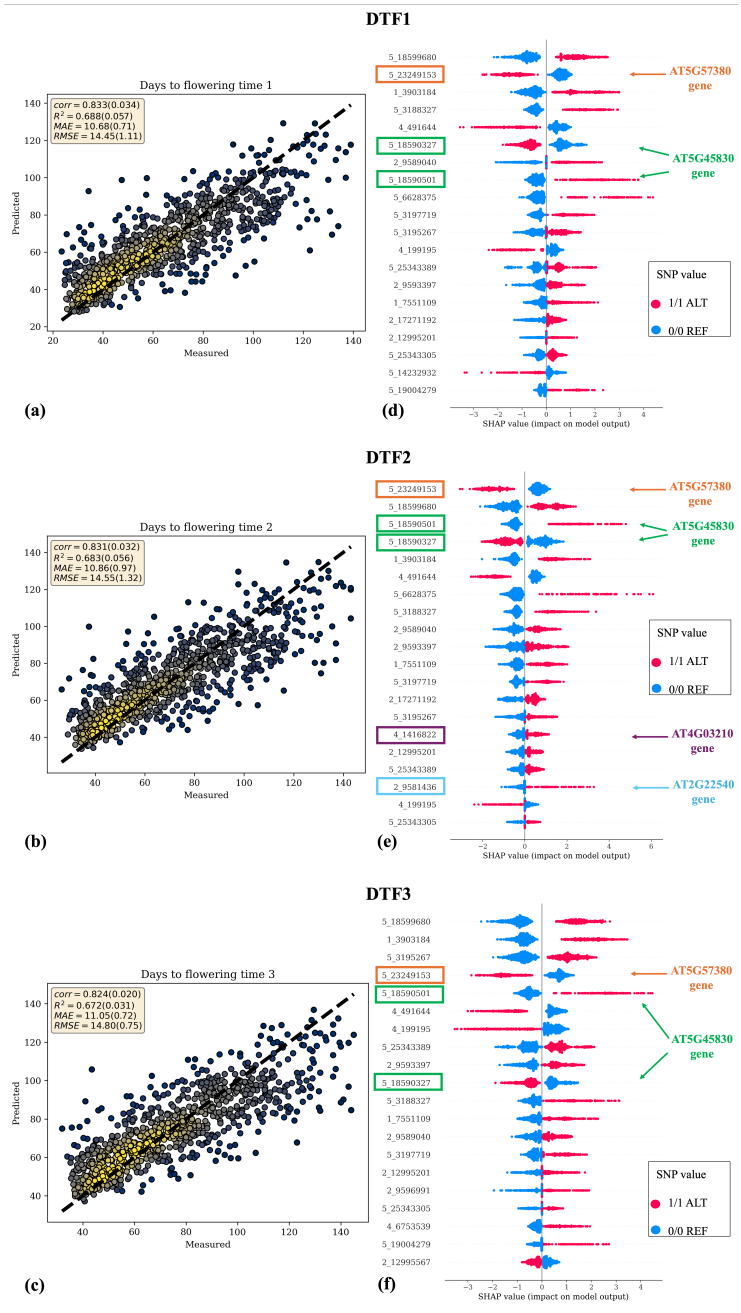

Figure 4. Scatter plots (**a**–**c**) and SHAP summary plots (**d**–**f**) for days to flowering time traits. Scatter plots represent actual phenotypic values on the x-axis and predicted values $\hat{y}_i$ on the y-axis, along

with evaluation metrics of regression models. The yellow color in the scatter plots represents the density of data points, with higher densities shown in more intense yellow regions. SHAP summary plots depict the feature importance of SNPs based on SHAP values. The 20 most important SNPs are ranked from most to least important based on the absolute values of their SHAP scores. Additionally, the color code provides information on the direction of effects, with red indicating the ALT allele and blue indicating the REF allele. Each point in every row represents the SHAP value of a specific SNP for each strain. The SNPs within the rectangles are those for which a biological meaning has been found, as they fall within genes of interest, which are described in more detail in the Discussion Section.

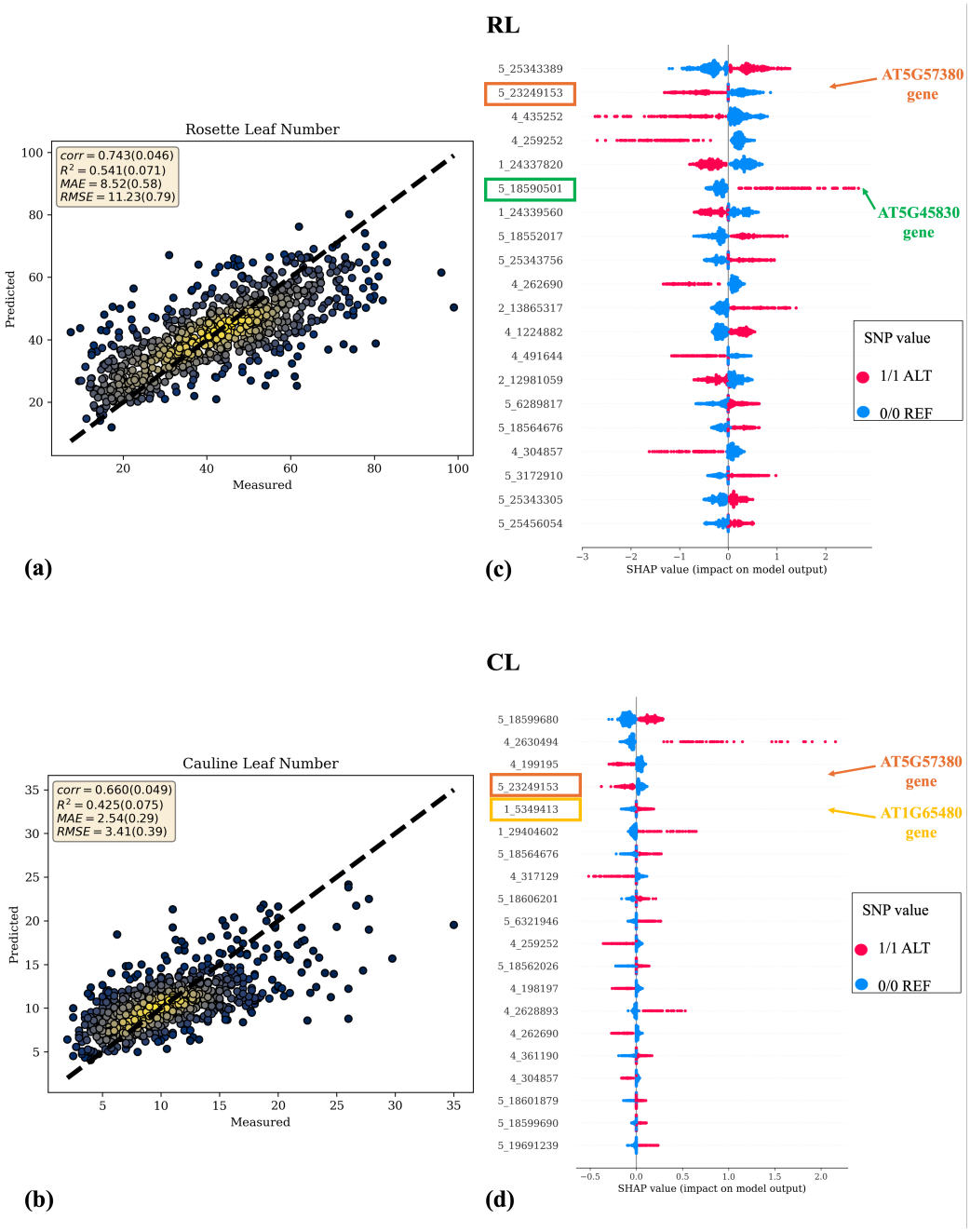

Figure 5. Scatter plots (**a**,**b**) and SHAP summary plots (**c**,**d**) for leaf number traits. Scatter plots represent actual phenotypic values on the x-axis and predicted values $\hat{y}_i$ on the y-axis, along with evaluation metrics of regression models. The yellow color in the scatter plots represents the density of data points, with more intense yellow regions indicating higher data point densities. SHAP summary plots depict the feature importance of SNPs based on SHAP values. The 20 most important SNPs are ranked from most to least important based on the absolute values of their SHAP scores. Additionally,

the color code provides information on the direction of effects, with red indicating the ALT allele and blue indicating the REF allele. Each point in every row represents the SHAP value of a specific SNP for each strain. The SNPs within the rectangles are those for which a biological meaning has been found, as they fall within genes of interest, which are described in more detail in the Discussion Section.

4. Discussion

In this study, we aimed to develop and evaluate an artificial intelligence workflow for genotype-to-phenotype prediction in *Arabidopsis thaliana*, integrating both machine learning and mixed linear models with explainable AI techniques. The workflow was applied to DNA polymorphism data from the *Arabidopsis* 1001 Genomes Project [29] to predict five phenotypic traits, including days to flowering and leaf number. Combining linear and machine learning models may thus be a valuable strategy, as the latter can capture complex, non-additive genetic effects that traditional linear approaches may not fully capture [52–54]. By combining predictive modeling with SHAP-based interpretability, our objective was not only to achieve accurate trait prediction but also to gain biologically meaningful insights into the genetic variants contributing to these traits.

A key innovation of our workflow lies in the use of SHAP values within the XAI framework. This integration not only identifies the variants contributing most to overall predictions, but also provides individual-level explanations, pinpointing specific SNPs that drive each plant's predicted phenotype. While functional validation will be necessary to confirm the causal role of these variants, such local interpretability offers a promising avenue for future precision breeding applications by helping to prioritize candidate SNPs for further investigation.

Although the application of a data-driven feature selection procedure did not substantially improve model accuracy for all traits, it significantly reduced the number of input variables, which is critical in high-dimensional genomic datasets. This step contributed to minimizing computational costs and enhancing SHAP-based model interpretability by limiting the number of features involved in the explanation space.

In our study, three traits described the number of days from sowing to specific flowering phenological stages (DTF1–3), while two traits (RL and CL) captured leaf number. Flowering time is a major breeding target, as this trait strongly influences adaptation to environmental stress and yield [55,56]. Leaf number also impacts crop productivity [57,58] and, in *Arabidopsis*, rosette and cauline leaf numbers are established morphological proxies of flowering time [59]. Our results confirm that traits related to days to flowering display higher genetic heritability than those related to leaf number, in agreement with previous reports [60,61].

Notably, several high-SHAP (Figures 4 and 5) SNPs mapped within genes with well-established roles in flowering and developmental timing (Supplementary Table S1). Specifically, SNPs 5_18590327 and 5_18590501, associated with DTF1, DTF2, and DTF3, reside in the *DELAY OF GERMINATION 1* (DOG1, AT5G45830) gene, which regulates flowering time [62]. Similarly, SNP 5_23249153 is located in the *VERNALIZATION INSENSITIVE 3* (VIN3, AT5G57380) gene, which represses *FLOWERING LOCUS C* (FLC) during vernalization [63]. Other SNPs of interest for DTF2 include 4_1416822 in *XYLOGLUCAN ENDOTRANSGLUCOSYLASE/HYDROLASE 9* (XTH9, AT4G03210), involved in inflorescence stem elongation [64], and 2_9581436 in *AGAMOUS-LIKE 22* (AGL22, AT2G22540), a regulator of flowering [65]. For DTF3, SNP 2_9596991 resides in the gene *ATKH6* (AT2G22600), a nucleic acid-binding protein with a K-homologous (KH) domain linked to floral development [66]. Interestingly, several of these SNPs also contributed to leaf number traits, highlighting a shared genetic basis. Consistently, the above-mentioned *VIN3* SNP 5_23249153 also showed high SHAP values for RL and CL, and the *DOG1* SNP 5_18590501 was also strongly associated with RL (Supplementary Table S1). Moreover, SNP 1_24337820,

located near *FLOWERING LOCUS T* (FT, AT1G65480), displayed high SHAP values for RL, consistent with the role of FT as a key flowering integrator [67]. These findings reinforce the physiological and genetic connection between flowering time and leaf number and support the use of leaf number as a morphological indicator of flowering earliness in *Arabidopsis*.

One of the strengths of our study is the use of a comprehensive public dataset encompassing over a thousand genotypes. Nonetheless, some limitations should be acknowledged. First, the analysis was restricted to five phenotypic traits. Second, the workflow was trained and evaluated exclusively on the Arabidopsis 1001 Genomes dataset. This reliance on a single source may limit the generalizability of the findings. While the internal cross-validation design provides a robust estimate of predictive performance within this dataset, future validation on additional datasets or other species will be essential to fully assess the broader applicability of the workflow.

An important consideration emerging from our results is the trade-off between computational efficiency and interpretability. While rrBLUP offers computational simplicity and fast execution suitable for large-scale breeding programs, CatBoost combined with SHAP provides deeper biological insights at the cost of increased computational complexity. For routine genomic selection where speed is prioritized, rrBLUP remains highly valuable. However, when biological understanding and precision breeding decisions are paramount, the additional computational investment in ensemble methods with XAI techniques like SHAP proves worthwhile, as demonstrated by our identification of functionally relevant SNPs in key flowering genes.

In summary, our results demonstrate the potential of integrating AI with XAI techniques to advance genomic selection methodologies. SHAP, in particular, provides both global and local interpretability, enabling the identification of SNPs that drive trait prediction at the level of individual plants. This dual perspective combines predictive accuracy with biological insight, making the approach a valuable resource for precision breeding.

5. Conclusions

In this study, we implemented and benchmarked machine learning regressors against established genomic regression methods for predicting phenotypic traits from genotype data. Across the five traits considered, machine learning achieved performance comparable to, and at times exceeding, state-of-the-art genomic regressors, indicating its capacity to model complex genetic structures and offering a credible alternative within genomic prediction workflows.

We further integrated explainable AI to ground these predictions biologically. SHAP attributions quantified the contribution of individual SNPs to trait predictions, providing complementary global and plant-level explanations. This framework clarifies why specific variants are prioritized by the models and strengthens the interpretability of genotype–phenotype links.

Our analysis targeted three flowering time traits and two leaf number traits, both of direct relevance to breeding under changing climates. The accuracy obtained suggests practical utility for tasks such as yield estimation and planning for shifts in growing seasons.

The present analysis is limited to five traits and to a single public cohort of A. thaliana accessions. Performance estimates and feature attributions are derived from internal validation and may not extend beyond this population. External evaluation on independent cohorts and additional species will be required to assess generalizability and robustness across distinct genetic backgrounds and environments.

Overall, the results support the use of machine learning for genotype–phenotype prediction and demonstrate the utility of coupling these models with SHAP-based interpre-

tation. The combination of competitive accuracy and traceable, variant-level attributions can inform biologically grounded, model-assisted breeding decisions.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/asi8060164/s1.

Author Contributions: Conceptualization, P.N. and S.T.; methodology, P.N., D.D., R.B., and S.T.; software, P.N. and N.N.; validation, S.P., C.D., and C.Z.; formal analysis, P.N.; investigation, P.N.; resources, D.D.; data curation, P.N., C.D., and N.N.; writing—original draft preparation, P.N., S.P., N.N., and C.D.; writing—review and editing, P.N., N.N., S.P., C.D., D.D., C.Z., R.B., and S.T.; visualization, P.N.; supervision, S.T.; project administration, S.T.; funding acquisition, S.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by The National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.4—Call for tender No. 3138 of 16 December 2021 of the Italian Ministry of University and Research funded by the European Union—NextGenerationEU. Award Number: Project code: CN00000022, Concession Decree No. 1031 of 17 February 2022 adopted by the Italian Ministry of University and Research, CUP H93C22000440007, Project title: "National Research Centre for Agricultural Technologies (Agritech)"; the METROFOOD-IT project has received funding from the European Union—NextGenerationEU, PNRR—Mission 4 "Education and Research" Component 2: from research to business, Investment 3.1: Fund for the realization of an integrated system of research and innovation infrastructures—IR0000033 (D.M. Prot. n.120 del 21/06/2022).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are either publicly available in the databases cited in the bibliography or available from the corresponding author on request.

Acknowledgments: The authors would like to thank the resources made available by ReCaS, a project funded by the MIUR (Italian Ministry for Education, University and Research) in the "PON Ricerca e Competitività 2007–2013-Azione I-Interventi di rafforzamento strutturale" *PONa*3_00052, Avviso 254/Ric, University of Bari.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Meuwissen, T.H.; Hayes, B.J.; Goddard, M. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **2001**, *157*, 1819–1829. [CrossRef] [PubMed]
2. Molenaar, H.; Boehm, R.; Piepho, H.P. Phenotypic selection in ornamental breeding: It's better to have the BLUPs than to have the BLUEs. *Front. Plant Sci.* **2018**, *9*, 385430. [CrossRef] [PubMed]
3. Ma, W.; Qiu, Z.; Song, J.; Li, J.; Cheng, Q.; Zhai, J.; Ma, C. A deep convolutional neural network approach for predicting phenotypes from genotypes. *Planta* **2018**, *248*, 1307–1318. [CrossRef] [PubMed]
4. Wang, K.; Abid, M.A.; Rasheed, A.; Crossa, J.; Hearne, S.; Li, H. DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants. *Mol. Plant* **2023**, *16*, 279–293. [CrossRef]
5. Guo, T.; Li, X. Machine learning for predicting phenotype from genotype and environment. *Curr. Opin. Biotechnol.* **2023**, *79*, 102853. [CrossRef]
6. Nazzicari, N.; Biscarini, F. Stacked kinship CNN vs. GBLUP for genomic predictions of additive and complex continuous phenotypes. *Sci. Rep.* **2022**, *12*, 19889. [CrossRef]
7. Gill, M.; Anderson, R.; Hu, H.; Bennamoun, M.; Petereit, J.; Valliyodan, B.; Nguyen, H.T.; Batley, J.; Bayer, P.E.; Edwards, D. Machine learning models outperform deep learning models, provide interpretation and facilitate feature selection for soybean trait prediction. *BMC Plant Biol.* **2022**, *22*, 180. [CrossRef]
8. Wang, H.; Yan, S.; Wang, W.; Chen, Y.; Hong, J.; He, Q.; Diao, X.; Lin, Y.; Chen, Y.; Cao, Y.; et al. Cropformer: An interpretable deep learning framework for crop genomic prediction. *Plant Commun.* **2025**, *6*, 101223. [CrossRef]
9. Chao, D.; Wang, H.; Wan, F.; Yan, S.; Fang, W.; Yang, Y. MtCro: Multi-task deep learning framework improves multi-trait genomic prediction of crops. *Plant Methods* **2025**, *21*, 12. [CrossRef]

10. Montesinos-Lopez, A.; Crespo-Herrera, L.; Dreisigacker, S.; Gerard, G.; Vitale, P.; Saint Pierre, C.; Govindan, V.; Tarekegn, Z.T.; Flores, M.C.; Pérez-Rodríguez, P.; et al. Deep learning methods improve genomic prediction of wheat breeding. *Front. Plant Sci.* **2024**, *15*, 1324090. [CrossRef]
11. Crossa, J.; Pérez-Rodríguez, P.; Cuevas, J.; Montesinos-López, O.; Jarquín, D.; De Los Campos, G.; Burgueño, J.; González-Camacho, J.M.; Pérez-Elizalde, S.; Beyene, Y.; et al. Genomic selection in plant breeding: Methods, models, and perspectives. *Trends Plant Sci.* **2017**, *22*, 961–975. [CrossRef]
12. Pérez-Rodríguez, P.; Gianola, D.; González-Camacho, J.M.; Crossa, J.; Manès, Y.; Dreisigacker, S. Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. *G3 Genes Genomes Genet.* **2012**, *2*, 1595–1605. [CrossRef]
13. Cuevas, J.; Crossa, J.; Soberanis, V.; Pérez-Elizalde, S.; Pérez-Rodríguez, P.; Campos, G.d.l.; Montesinos-López, O.; Burgueño, J. Genomic prediction of genotype× environment interaction kernel regression models. *The Plant Genome* **2016**, *9*, plantgenome2016.03.0024. [CrossRef] [PubMed]
14. 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]
15. Wang, X.; Shi, S.; Wang, G.; Luo, W.; Wei, X.; Qiu, A.; Luo, F.; Ding, X. Using machine learning to improve the accuracy of genomic prediction of reproduction traits in pigs. *J. Anim. Sci. Biotechnol.* **2022**, *13*, 60. [CrossRef] [PubMed]
16. Novielli, P.; Romano, D.; Pavan, S.; Losciale, P.; Stellacci, A.M.; Diacono, D.; Bellotti, R.; Tangaro, S. Explainable artificial intelligence for genotype-to-phenotype prediction in plant breeding: A case study with a dataset from an almond germplasm collection. *Front. Plant Sci.* **2024**, *15*, 1434229. [CrossRef]
17. Novielli, P.; Magarelli, M.; Romano, D.; de Trizio, L.; Di Bitonto, P.; Monaco, A.; Amoroso, N.; Stellacci, A.M.; Zoani, C.; Bellotti, R.; et al. Climate Change and Soil Health: Explainable Artificial Intelligence Reveals Microbiome Response to Warming. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 1564–1578. [CrossRef]
18. Novielli, P.; Magarelli, M.; Romano, D.; Di Bitonto, P.; Stellacci, A.M.; Monaco, A.; Amoroso, N.; Bellotti, R.; Tangaro, S. Leveraging explainable AI to predict soil respiration sensitivity and its drivers for climate change mitigation. *Sci. Rep.* **2025**, *15*, 12527. [CrossRef]
19. van Hilten, A.; Kushner, S.A.; Kayser, M.; Ikram, M.A.; Adams, H.H.; Klaver, C.C.; Niessen, W.J.; Roshchupkin, G.V. GenNet framework: Interpretable deep learning for predicting phenotypes from genetic data. *Commun. Biol.* **2021**, *4*, 1094. [CrossRef]
20. Ali, S.; Abuhmed, T.; El-Sappagh, S.; Muhammad, K.; Alonso-Moral, J.M.; Confalonieri, R.; Guidotti, R.; Del Ser, J.; Díaz-Rodríguez, N.; Herrera, F. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Inf. Fusion* **2023**, *99*, 101805. [CrossRef]
21. Hulsen, T. Explainable artificial intelligence (XAI): Concepts and challenges in healthcare. *AI* **2023**, *4*, 652–666. [CrossRef]
22. Novielli, P.; Baldi, S.; Romano, D.; Magarelli, M.; Diacono, D.; Di Bitonto, P.; Nannini, G.; Di Gloria, L.; Bellotti, R.; Amedei, A.; et al. Personalized colorectal cancer risk assessment through explainable AI and Gut microbiome profiling. *Gut Microbes* **2025**, *17*, 2543124. [CrossRef] [PubMed]
23. Toussaint, P.A.; Leiser, F.; Thiebes, S.; Schlesner, M.; Brors, B.; Sunyaev, A. Explainable artificial intelligence for omics data: A systematic mapping study. *Briefings Bioinform.* **2024**, *25*, bbad453. [CrossRef] [PubMed]
24. Voss-Fels, K.P.; Cooper, M.; Hayes, B.J. Accelerating crop genetic gains with genomic selection. *Theor. Appl. Genet.* **2019**, *132*, 669–686. [CrossRef]
25. Spindel, J.; Begum, H.; Akdemir, D.; Collard, B.; Redoña, E.; Jannink, J.; McCouch, S. Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. *Heredity* **2016**, *116*, 395–408. [CrossRef]
26. Rice, B.; Lipka, A.E. Evaluation of RR-BLUP genomic selection models that incorporate peak genome-wide association study signals in maize and sorghum. *Plant Genome* **2019**, *12*, 180052. [CrossRef]
27. Azodi, C.B.; Bolger, E.; McCarren, A.; Roantree, M.; de Los Campos, G.; Shiu, S.H. Benchmarking parametric and machine learning models for genomic prediction of complex traits. *G3 Genes Genomes Genet.* **2019**, *9*, 3691–3702. [CrossRef]
28. Sims, J.; Schlögelhofer, P.; Kurzbauer, M.T. From microscopy to nanoscopy: Defining an Arabidopsis thaliana meiotic atlas at the nanometer scale. *Front. Plant Sci.* **2021**, *12*, 672914. [CrossRef]
29. 1001 Genomes Consortium. 1001 Genomes Project, 2016. Available online: https://1001genomes.org/ (accessed on 4 April 2024).
30. 1001 Genomes Consortium. Data Center — 1001 Genomes Project (v3.1), 2016. Available online: https://1001genomes.org/data/GMI-MPI/releases/v3.1/SNP_matrix_imputed_hdf5/ (accessed on 4 April 2024).
31. Jaegle, B.; Pisupati, R.; Soto-Jiménez, L.M.; Burns, R.; Rabanal, F.A.; Nordborg, M. Extensive sequence duplication in Arabidopsis revealed by pseudo-heterozygosity. *Genome Biol.* **2023**, *24*, 44. [CrossRef]
32. Alonso-Blanco, C.; Andrade, J.; Becker, C.; Bemm, F.; Bergelson, J.; Borgwardt, K.M.; Cao, J.; Chae, E.; Dezwaan, T.M.; Ding, W.; et al. 1135 genomes reveal the global pattern of polymorphism in Arabidopsis thaliana. *Cell* **2016**, *166*, 481–491. [CrossRef]
33. Chang, C.C.; Chow, C.C.; Tellier, L.C.; Vattikuti, S.; Purcell, S.M.; Lee, J.J. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* **2015**, *4*, s13742-015-0047-8. [CrossRef]

34. Tabangin, M.E.; Woo, J.G.; Martin, L.J. The effect of minor allele frequency on the likelihood of obtaining false positives. In *BMC Proceedings*; BioMed Central: London, UK, 2009; Volume 3, pp. 1–4.
35. Ye, S.; Gao, N.; Zheng, R.; Chen, Z.; Teng, J.; Yuan, X.; Zhang, H.; Chen, Z.; Zhang, X.; Li, J.; et al. Strategies for obtaining and pruning imputed whole-genome sequence data for genomic prediction. *Front. Genet.* **2019**, *10*, 673. [CrossRef] [PubMed]
36. Ren, D.; Teng, J.; Diao, S.; Lin, Q.; Li, J.; Zhang, Z. Impact of marker pruning strategies based on different measurements of marker distance on genomic prediction in dairy cattle. *Animals* **2021**, *11*, 1992. [CrossRef] [PubMed]
37. Pook, T.; Mayer, M.; Geibel, J.; Weigend, S.; Cavero, D.; Schoen, C.C.; Simianer, H. Improving imputation quality in BEAGLE for crop and livestock data. *G3 Genes Genomes Genet.* **2020**, *10*, 177–188. [CrossRef]
38. Nothnagel, M.; Ellinghaus, D.; Schreiber, S.; Krawczak, M.; Franke, A. A comprehensive evaluation of SNP genotype imputation. *Hum. Genet.* **2009**, *125*, 163–171. [CrossRef] [PubMed]
39. Salam, M.A.; Azar, A.T.; Elgendy, M.S.; Fouad, K.M. The effect of different dimensionality reduction techniques on machine learning overfitting problem. *Int. J. Adv. Comput. Sci. Appl.* **2021**, *12*, 641–655. [CrossRef]
40. Samala, R.K.; Chan, H.P.; Hadjiiski, L.; Koneru, S. Hazards of data leakage in machine learning: A study on classification of breast cancer using deep neural networks. In Proceedings of the Medical Imaging 2020: Computer-Aided Diagnosis, Houston, TX, USA, 15–20 February 2020; Volume 11314, pp. 279–284.
41. Balajee, J.M.; Sathish, G.; Saravanan, N. Data wrangling and data leakage in machine learning for healthcare. *Int. J. Emerg. Technol. Innov. Res.* **2018**, *5*, 553–557.
42. Endelman, J.B. Ridge regression and other kernels for genomic selection with R package rrBLUP. *The Plant Genome* **2011**, *4*, plantgenome2011.08.0024. [CrossRef]
43. Kozachenko, L.F.; Leonenko, N.N. Sample estimate of the entropy of a random vector. *Probl. Peredachi Informatsii* **1987**, *23*, 9–16.
44. Ross, B.C. Mutual information between discrete and continuous data sets. *PloS ONE* **2014**, *9*, e87357. [CrossRef]
45. Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating mutual information. *Phys. Rev. E* **2004**, *69*, 066138. [CrossRef]
46. Guo, H.; Yu, Z.; An, J.; Han, G.; Ma, Y.; Tang, R. A two-stage mutual information based Bayesian Lasso algorithm for multi-locus genome-wide association studies. *Entropy* **2020**, *22*, 329. [CrossRef] [PubMed]
47. Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, *2*, 18–22.
48. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
49. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 6725–6734.
50. Rahman, S.; Irfan, M.; Raza, M.; Moyeezullah Ghori, K.; Yaqoob, S.; Awais, M. Performance analysis of boosting classifiers in recognizing activities of daily living. *Int. J. Environ. Res. Public Health* **2020**, *17*, 1082. [CrossRef]
51. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4765–4774.
52. Danilevicz, M.F.; Gill, M.; Anderson, R.; Batley, J.; Bennamoun, M.; Bayer, P.E.; Edwards, D. Plant genotype to phenotype prediction using machine learning. *Front. Genet.* **2022**, *13*, 822173. [CrossRef]
53. Cheng, C.Y.; Li, Y.; Varala, K.; Bubert, J.; Huang, J.; Kim, G.J.; Halim, J.; Arp, J.; Shih, H.J.S.; Levinson, G.; et al. Evolutionarily informed machine learning enhances the power of predictive gene-to-phenotype relationships. *Nat. Commun.* **2021**, *12*, 5627. [CrossRef]
54. Okser, S.; Pahikkala, T.; Airola, A.; Salakoski, T.; Ripatti, S.; Aittokallio, T. Regularized machine learning in the genetic prediction of complex traits. *PLoS Genet.* **2014**, *10*, e1004754. [CrossRef]
55. Deva, C.; Dixon, L.; Urban, M.; Ramirez-Villegas, J.; Droutsas, I.; Challinor, A. A new framework for predicting and understanding flowering time for crop breeding. *Plants People Planet* **2024**, *6*, 197–209. [CrossRef]
56. Franks, S.J. The unique and multifaceted importance of the timing of flowering. *Am. J. Bot.* **2015**, *102*, 1401–1402. [CrossRef]
57. Xie, X.; Ge, Y.; Walia, H.; Yang, J.; Yu, H. Leaf-counting in monocot plants using deep regression models. *Sensors* **2023**, *23*, 1890. [CrossRef]
58. Farjon, G.; Itzhaky, Y.; Khoroshevsky, F.; Bar-Hillel, A. Leaf counting: Fusing network components for improved accuracy. *Front. Plant Sci.* **2021**, *12*, 575751. [CrossRef]
59. Pouteau, S.; Albertini, C. The significance of bolting and floral transitions as indicators of reproductive phase change in Arabidopsis. *J. Exp. Bot.* **2009**, *60*, 3367–3377. [CrossRef] [PubMed]
60. Grimm, D.G.; Roqueiro, D.; Salomé, P.A.; Kleeberger, S.; Greshake, B.; Zhu, W.; Liu, C.; Lippert, C.; Stegle, O.; Schölkopf, B.; et al. easyGWAS: A cloud-based platform for comparing the results of genome-wide association studies. *Plant Cell* **2017**, *29*, 5–19. [CrossRef] [PubMed]
61. Seymour, D.K.; Chae, E.; Grimm, D.G.; Martín Pizarro, C.; Habring-Müller, A.; Vasseur, F.; Rakitsch, B.; Borgwardt, K.M.; Koenig, D.; Weigel, D. Genetic architecture of nonadditive inheritance in Arabidopsis thaliana hybrids. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, E7317–E7326. [CrossRef]

62. Huo, H.; Wei, S.; Bradford, K.J. DELAY OF GERMINATION1 (DOG1) regulates both seed dormancy and flowering time through microRNA pathways. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, E2199–E2206. [CrossRef] [PubMed]
63. Bond, D.M.; Wilson, I.W.; Dennis, E.S.; Pogson, B.J.; Jean Finnegan, E. VERNALIZATION INSENSITIVE 3 (VIN3) is required for the response of Arabidopsis thaliana seedlings exposed to low oxygen conditions. *Plant J.* **2009**, *59*, 576–587. [CrossRef]
64. Hyodo, H.; Yamakawa, S.; Takeda, Y.; Tsuduki, M.; Yokota, A.; Nishitani, K.; Kohchi, T. Active gene expression of a xyloglucan endotransglucosylase/hydrolase gene, XTH9, in inflorescence apices is related to cell elongation in Arabidopsis thaliana. *Plant Mol. Biol.* **2003**, *52*, 473–482. [CrossRef]
65. Méndez-Vigo, B.; Martínez-Zapater, J.M.; Alonso-Blanco, C. The flowering repressor SVP underlies a novel Arabidopsis thaliana QTL interacting with the genetic background. *PLoS Genet.* **2013**, *9*, e1003289. [CrossRef]
66. Zhang, Y.; Ma, Y.; Liu, R.; Li, G. Genome-wide characterization and expression analysis of KH family genes response to ABA and SA in Arabidopsis thaliana. *Int. J. Mol. Sci.* **2022**, *23*, 511. [CrossRef]
67. Huang, X.; Ding, J.; Effgen, S.; Turck, F.; Koornneef, M. Multiple loci and genetic interactions involving flowering time genes regulate stem branching among natural variants of Arabidopsis. *New Phytol.* **2013**, *199*, 843–857. [CrossRef]