Unveiling Optimal Models for Phenotype Prediction in Soybean Branching: An In-depth Examination of 11 Non-linear Regression Models, Highlighting SVR and SHAP Importance

Wei Zhou

wei.zhou@famu.edu

Florida Agricultural and Mechanical University

Zhengxiao Yan

Florida Agricultural and Mechanical University

Liting Zhang

Florida Agricultural and Mechanical University

Article

Keywords: Phenotype Prediction, Non-linear Regression, Artificial intelligence, feature importance

Posted Date: September 7th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3232751/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on March 11th, 2024. See the published version at https://doi.org/10.1038/s41598-024-55243-x.

Abstract

Plant breeding is gaining importance as a sustainable tool to address the challenges posed by a growing global population and enhance food security. Advanced high-throughput omics technologies are utilized to accelerate crop improvement and develop resilient varieties with higher yield performance. These technologies generate vast genetic data, which can be exploited to manipulate key plant characteristics for crop improvement. The integration of big data and AI in plant breeding has the potential to revolutionize the field and increase food security.

By using branching data (phenotype) of 1918 soybean accessions and 42k SNP polymorphic data (genotype), this study systematically compared 11 non-linear regression AI models, including four deep learning models (DBN regression, ANN regression, Autoencoders regression, and MLP regression) and seven machine learning models (e.g., SVR, XGBoost regression, Random Forest regression, LightGBM regression, GPS regression, Decision Tree regression, and Polynomial regression). After being evaluated by four valuation metrics: R2 (R-squared), MAE (Mean Absolute Error), MSE (Mean Squared Error), and MAPE (Mean Absolute Percentage Error), it was found that the SVR, ANN, and Autoencoder outperformed other models and could obtain a better prediction accuracy if they were used for phenotype prediction.

To support the evaluation of deep learning methods, feature importance and GO enrichment analyses were conducted. After comprehensively comparing four feature importance algorithms, there was no significant difference among the feature importance ranking score among these four algorithms, but the SHAP value could provide rich information on genes with negative contributions, and SHAP importance was chosen for feature selection. The genes identified by the SVR model plus SHAP importance combination clearly grouped into three clusters on the soybean whole genome. Our GO enrichment results also confirmed the prediction accuracy of this methods combination.

The results of this study offer valuable insights for AI-mediated plant breeding, addressing challenges faced by traditional breeding programs. The method developed has broad applicability in phenotype prediction, minor QTL mining, and plant smart-breeding systems, contributing significantly to the advancement of AI-based breeding practices and transitioning from experience-based to data-based breeding.

Introduction

Now, with the global population projected to reach nearly 10 billion by 2050 [1,2,3]and climate change causing significant alterations in growing conditions [4,5], there is an urgent need for breeders to employ new technologies such as molecular markers, genomic selection, and artificial intelligence to enhance breeding efficiency and progress in order to feed the world in a rapidly changing environment. Classical breeding methods have limitations in providing substantial insights into the genome structure of plant species due to the lack of correlation between genotype and phenotype [6,7,8]. However, with the advent of the genomics revolution in the early 2000s [9], plant breeders found themselves overwhelmed by genomic data that exceeded the capacity of traditional statistical techniques [10]. Concurrently, the Next Generation Sequencing (NGS) technology has significantly accelerated the availability of complete genome sequences for desired plant species, resulting in large datasets [11,12]. High-coverage and high-quality reference genome sequences offer comprehensive information on genes, genome composition, and serve as a foundation for understanding genome variation, enabling "omics" investigations of target species [13,14]. The utilization of machine- and deep-learning algorithms in the context of complex traits in plants holds the potential to enhance prediction accuracies. With the significant increase in data collected from breeding programs and the slow rate of genetic gain improvement, it has become imperative to explore the capabilities of artificial intelligence for data analysis. Traditional plant breeding methods, ill-suited for handling vast amounts of data and making precise decisions, have prompted the rapid development of machine learning, which has now found widespread use in various scientific domains, including plant genotyping and phenotyping [15,16,17,18,19]. Scientists have been actively investigating the application of artificial intelligence in plant breeding, aiming to intelligently and efficiently extract valuable insights from breeding datasets using relevant models and algorithms [20,21,22].

A comparative analysis of multilayer perceptron (MLP), Support Vector Machine (SVM), and Bayesian Threshold Genomic Best Linear Unbiased Prediction (TGBLUP) for predicting ordinal traits in plant breeding revealed no statistical difference among different numbers of layers in the MLP model [23]. This finding suggests that a conventional neuronal network model with a single layer is sufficient for accurate predictions. In the context of deep learning for genomic prediction of complex traits in polyploid species such as strawberry and blueberry, it was observed that interactions between hyperparameter combinations and the number of convolutional filters and regularization in the initial layers significantly influenced model performance [24]. To estimate genomic breeding values (GEBVs) for a cattle population, the efficiency of three machine learning methods, namely Random Forests (RF), Gradient Boosting Machine (GBM), and XgBoost, was evaluated using subsets of single nucleotide polymorphisms (SNPs) to construct genomic relationship matrices (GRMs) [25]. The study found that RF and particularly GBM were effective in identifying a subset of SNPs directly linked to candidate genes influencing the growth trait. Moreover, the subsets of SNPs derived from RF and GBM outperformed evenly spaced subsets across the cattle genome, demonstrating significantly improved genomic prediction accuracy [25]. In comparison, RF and GBM consistently outperformed XgBoost in genomic prediction accuracy. Through a comparative analysis of machine- and deep-learning-based unitrait and multitrait models, as well as a traditional genomic best linear unbiased predictor (GBLUP) and Bayesian models, in a study involving 650 recombinant inbred lines (RILs) of spring wheat, it was determined that the multitrait genomic selection (MT-GS) models exhibited superior

performance compared to the unitrait genomic selection (UT-GS) models [26]. Among the MT-GS models, random forest and multilayer perceptron emerged as the top-performing models for predicting both traits.

In plant breeding, multi-trait and multi-environment interactions are prevalent, necessitating the use of more robust models and algorithms to leverage correlations between traits and environmental factors in order to enhance prediction accuracy in breeding programs. In a study comparing the prediction performance of multi-trait deep learning (MTDL) models to the Bayesian multi-trait and multi-environment (BMTME) model proposed by Montesinos-López et al., it was observed that the BMTME model provided the best predictions in two out of three datasets [27]. In scenarios without genotype-by-environment (GxE) interaction, the MTDL model exhibited superior performance, while among models accounting for GxE interaction, the BMTME model outperformed the others. These findings indicate that the MTDL model demonstrates competitiveness in making predictions within the context of genomic selection. In the context of genotype-by-environment (GXE) prediction, effective variable selection plays a crucial role in many applications of predictive modeling. Okser et al. analyzed the potential challenges associated with regularized machine learning models, such as model overfitting to the training data and identifiability of predictive variants, using examples from human disease classification and quantitative trait prediction[28]. These findings have implications not only in human research but also in animal and plant breeding.

Recent advancements in field phenomics have significantly contributed to the study of stress response traits in soybean (Glycine max). The performance of complex traits is influenced by both genetic and environmental factors, which can be challenging to dissect due to the need for multiple replications of numerous genotypes under diverse environmental conditions. To address these challenges, deep learning techniques have been employed to integrate genotype and weather variables for predicting soybean yield. Shook et al. reported that the Long Short-Term Memory (LSTM) - Recurrent Neural Networks model effectively isolates key weather events and genetic interactions that impact yield, seed oil, seed protein, and maturity[29]. This enables the prediction of genotypic responses in previously unseen environments. To provide interpretability of important time windows during the growing season, a temporal attention mechanism has been developed for the LSTM model. The output of this interpretable model offers valuable insights for plant breeders [30]. In a study by Yoosefzadeh-Najafabadi et al. (2021), the robustness of three common machine learning (ML) algorithms, namely multilayer perceptron (MLP), support vector machine (SVM), and random forest (RF), was assessed to predict soybean (Glycine max) seed yield using hyperspectral reflectance[31]. The findings revealed that soybean breeders could effectively utilize an ensemble-stacking algorithm, employing either the full or selected spectra reflectance, to identify high-yielding soybean genotypes at early growth stages among a large number of genotypes.

Despite these advancements, challenges persist in linking phenotypic data to genotypes for the identification of genotypes with higher genetic gain. The application of artificial intelligence in plant breeding involves mining the deep associations between genotype and phenotype. To harness the potential of AI in plant breeding, it is crucial to intelligently and efficiently analyze breeding datasets using appropriate models and algorithms.

With the advancement of modern biotechnology, DNA molecular marker technology has become increasingly prevalent in crop breeding research [32,33]. Among these technologies, single nucleotide polymorphism (SNP) stands out due to its high density across the genome, abundance of polymorphic information, strong specificity, ease of detection, and genetic stability. SNPs have bridged the gaps left by first-generation markers (e.g., restriction fragment length polymorphism, RFLP) and second-generation markers (e.g., microsatellite DNA polymorphism) to a significant extent[34]. Consequently, SNPs are recognized as the third-generation genetic marker method and are widely employed in genetic research. SNPs reflect DNA genetic variations at the single base level and can be located in close proximity to genes, offering substantial potential for detecting correlations between alleles and phenotypes of individual genes [34]. The integration of next-generation sequencing (NGS) technology with precise phenotyping has increased the prospects of discovering candidate genes and their allelic variants that control traits of interest[32]. Therefore, future research focusing on next-gen AI is crucial for bridging the gap between phenotypes and genotypes and facilitating crop improvement programs. The development of NGS technology further expands opportunities for utilizing SNPs as phenotypic clues. However, the field of smart breeding still requires more robust statistical models for analyzing ordered phenotypes and improving the accuracy of candidate genotype selection. There is an urgent need to develop methods that can elucidate the complex biological perspectives underlying many current artificial intelligence approaches. By opening these "black boxes" and providing meaningful explanations, breeders and researchers can enhance breeding efficiency through informed AI decisions and outputs. Therefore, future research in next-generation AI is an essential prerequisite for bridging the phenotype-genotype gap and facilitating crop improvement programs. In this project, we propose to combine genotype data generated from next-generation sequencing with artificial intelligence techniques to advance our understanding of genome structure, develop genomic selection models for legume and other specific crop species, establish connections between phenomes and genomes, and provide a technical reference for future efficient artificial intelligence breeding.

Results

Result 1. "Evaluation of the prediction accuracy of 11 AI models

Given the complex relationship between phenotype and genotype, genes can contribute positively or negatively to traits. Some genes have a significant impact, while others have a minor influence. Moreover, the contribution and role of genes can also be influenced by the environment. In essence, the relationship between phenotype and genotype can be described as follows: the dependent variable (phenotype) cannot be predicted or estimated based on the values of the independent variables (genotype). So, we applied the non-linear regression algorithm in this study.

Considering the specific problem of predicting phenotypes and the characteristics of the SNP genotype dataset, we conducted a comparison using the most suitable non-linear regression machine learning and deep learning models. Seven machine learning models and four deep learning models were selected for this purpose. The seven machine learning models include SVR (Support Vector Regression), XGBoost (Extreme Gradient Boosting) regression, Random Forest regression, LightGBM regression, Gaussian Processes (GPS) regression, Decision Tree regression and Polynomial regression; The four deep learning models include Deep Belief Network (DBN) regression, Artificial Neural Networks (ANN) regression, Autoencoders regression, and Multi-Layer Perceptron (MLP) regression.

In this study, we collected phenotype data from 2220 soybean accessions and applied the corresponding SNP genotype data in our research. To address the large dataset size and redundancy of genotype data, we employed two steps. First, we used the one-hot encoder to convert the genotype data (ATCG nucleotide code) into an array. Then, Principal Component Analysis (PCA) was used to reduce the dimensionality of the data. Finally, the chosen models were applied using the respective algorithms.

To determine the optimal parameters, we employed the Gridsearch method to fine-tune the hyperparameters and identify the best model for phenotype prediction. In order to evaluate the performance of the regression models, we utilized four evaluation metrics: R2 (R-squared), MAE (Mean Absolute Error), MSE (Mean Squared Error), and MAPE (Mean Absolute Percentage Error). These metrics were used to assess the prediction accuracy of each model.

Our results showed that, among seven machine learning models and four deep learning models, Polynomial regression exhibited the highest training performance, with an R2 value of 1.000, a MAE value of 0.00, an MSE value of 0.000, and an MAPE value of 0.000 during training, indicating a very close match between predicted and actual values. (Table 1). Furthermore, during testing, the Polynomial regression model achieved the lowest MAPE value of 0.080, indicating its good performance on unseen data. However, when comparing with other evaluation indicators, among all the models evaluated, the Autoencoder model had the highest R2 value for the test set, reaching an impressive 0.969. Additionally, the Autoencoder model obtained the lowest MAE value of 0.055 and the lowest MSE value of 0.008 during testing, indicating an excellent fit. (Table 1). It is worth noting that, the Autoencoder model exhibited a higher MAPE value during the training phase, with a value of 1.831. Furthermore, for the test phase, the MAPE value was reported as "NAN" (Not a Number) in Table 1, suggesting that there might have been issues with data or model performance during the calculation.

Table 1
Comparison of machine learning models

Valuation Metrics	SVR	XGBoost	randomforest	LightGBM	GPS	Decision tree	Polynomial regression	Autoencoders	ANN	DBN	MLP
R2 for train	0.926	0.908	0.897	0.967	0.798	0.303	1.000	0.973	0.995	0.984	0.905
R2 for test	0.637	0.589	0.565	0.585	0.599	0.096	0.614	0.969	0.605	0.544	0.303
R2 Relative difference	0.369	0.426	0.453	0.492	0.285	1.035	0.479	0.003	0.487	0.576	0.997
MAE for train	0.110	0.123	0.116	0.068	0.176	0.282	0.000	0.052	0.011	0.051	0.085
MAE for test	0.237	0.247	0.249	0.249	0.253	0.337	0.216	0.055	0.238	0.254	0.319
MAE Relative difference	0.734	0.665	0.726	1.146	0.357	0.178	2.000	0.064	1.826	1.334	1.161
MSE for train	0.019	0.024	0.027	0.009	0.053	0.183	0.000	0.007	0.001	0.004	0.025
MSE for test	0.096	0.109	0.115	0.110	0.106	0.239	0.102	0.008	0.105	0.126	0.185
MSE Relative difference	1.327	1.273	1.239	1.706	0.669	0.269	2.000	0.117	1.950	1.875	1.526
MAPE for train	0.040	0.046	0.044	0.025	6.634	0.106	0.000	1.831	0.004	1.819	0.032
MAPE for test	0.086	0.091	0.092	0.092	9.247	0.125	0.080	NaN	0.089	8.949	0.116
MAPE Relative difference	0.720	0.651	0.717	1.130	0.329	0.156	2.000	NAN	1.836	1.324	1.134

Among the seven machine learning models evaluated, lightGBM demonstrated the highest R2 value for the training set, achieving an impressive score of 0.967. Following closely is SVR, which obtained an R2 value of 0.926 for the training phase. Moving on to the test set, the SVR model performed the best, achieving the highest R2 value of 0.637, closely followed by GPS (Gaussian Processes) with an R2 value of 0.599. Regarding the Mean Absolute Error (MAE) metric, the lightGBM model exhibited the lowest value for the training set, registering a MAE of 0.068. Conversely, for the test set, the SVR model showcased the lowest MAE value of 0.237. Considering the Mean Squared Error (MSE) metric, the lightGBM model demonstrated the lowest value for the training set, yielding an MSE of 0.009. On the other hand, for the test set, the SVR model achieved the lowest MSE value, which amounted to 0.096. Focusing on the Mean Absolute Percentage Error (MAPE), the lightGBM model displayed the lowest value for the training set, obtaining an MAPE of 0.025. In contrast, the SVR model secured the lowest MAPE for the test set, recording a value of 0.086.

Among the four deep learning models, in training phase, ANN model got the highest R2 for train value of 0.995; and the lowest MAE for train value of 0.011, lowest MSE for train value of 0.001 and Lowest MAPE for train value of 0.004 (Table 1). when comparing with evaluation metrics R2, MAE and MSE in testing phase, the Autoencoder model got the best performance as mentioned above.

In order to further evaluate these 11 models, we conducted prediction accuracy evaluation based on Mean Absolute Error (MAE) (Fig. 1), as well as overfitting evaluation based on Mean Squared Error (MSE) (Fig. 2).

The probability plots of standardized residuals for each regression model provide a clear visual representation. The true values and predictions of the autoencoder model align well along the 45-degree line, with MAE of 0.05 for the training set and 0.06 for the test set. This demonstrates that the model's predictions adhere to the normality assumption. Similarly, the SVR model (MAE train = 0.11 and MAE test = 0.24), XGBoost model (MAE train = 0.12 and MAE test = 0.25), and Gaussian Process model (MAE train = 0.18 and MAE test = 0.25) also show good alignment between true values and predictions. On the other hand, the Multiplayer Perception model, Decision Tree model, and Deep Belief Network model exhibit a looser aggregation of true values and predictions, with data points scattered more loosely along the 45-degree line (Fig. 1).

To further assess the performance of each model and gain a deeper understanding of the relative disparities between testing and training results, considering the magnitudes of the values being compared, we employed Relative Difference Analysis (RDA) on all four evaluation metrics (Table 1). Our findings revealed that among the 11 models analyzed, Autoencoders demonstrated the most favorable performance, with a relative difference value of 0.003 for R2, 0.064 for MAE, and 0.117 for MSE. Additionally, the Decision Tree model achieved the lowest relative difference value for MAPE, with a value of 0.156. However, it also exhibited the highest R2 relative difference value at 1.035. On the other hand, the Polynomial Regression model displayed the highest relative difference values, with 2.000 for MAE, 2.000 for MSE, and 2.00 for MAPE (Table 1). In summary, when considering only the three metrics, R2, MAE, and MSE, the autoencoder outperformed all other models in the analysis of relative differences.

We conducted a thorough evaluation of these 11 models by analyzing overfitting based on Mean Squared Error (MSE) values. The results of our analysis indicate that SVR, lightGBM, Autoencoder, and ANN models fit both the training and test data exceptionally well, demonstrating a stable performance (Fig. 2). While the testing loss of the MLP model shows significant fluctuations when the hidden layer size is below 400, it exhibits a robust fit for the training and test data when the hidden layer size exceeds 400.

On the contrary, the Decision tree and DBN models demonstrate relatively poorer fits. As evident from the figures, the Decision tree model displays the least disparity between training and testing losses when the maximum depth (MAX Depth) is set to 5.0. Yet, when the depth is either below 5.0 or above 5.0, the gap between training and testing losses tends to widen. Regarding the DBN model, a relatively stable gap between training and testing losses is maintained for hidden layer sizes below 100. However, when the hidden layer size exceeds 100, the gap gradually increases. Similarly, the Polynomial regression model performs well when the polynomial degree is below 7. However, when the degree surpasses 9, there is a sharp increase in the gap between the training and testing losses (Fig. 2). Both the Random forest and Gaussian process models exhibit a growing gap between training and testing losses with an increase in the maximum depth or the degree of freedom (degree of freedom) (Fig. 2).

In summary, based on our comprehensive analysis, it is evident that Autoencoder, SVR, and ANN outperform the other models in relative terms. These models are suitable for genotype to phenotype prediction and minor QTL mapping. It could be the powerful tools in AI assisted breeding practice.

Result 2. Assessing the Performance of Four Feature Selection Methods for SVR

Our objective is to discover the most effective artificial intelligence model and utilize feature selection techniques to pinpoint genes responsible for specific physiological activities in plants. These identified genes will aid in precise phenotype prediction and gene function mining. To ensure the model's reliability, efficiency, low computational requirements, versatility, and openness, this study employs the Support Vector Regression (SVR) model as an illustrative example. We assess four distinct feature selection algorithms: Variable Ranking, Permutation, SHAP, and Correlation Matrix. Apart from the feature importance data, the Correlation Matrix method also provides valuable insights. A heatmap is employed to visualize the strength of correlations. In Fig. 3, we present the heatmap showcasing the top 100 features identified through the Correlation Matrix analysis based on the SVR model (Fig. 3). Additionally, the SHAP output plot offers a concise representation of the distribution and variability of SHAP values for each feature. Figure 4 illustrates the summary beeswarm plot of the top 20 features derived from our SHAP importance analysis based on the SVR model. This plot effectively captures the relative effect of all the features in the entire dataset (Fig. 4).

We ranked all SNPs based on the absolute values of feature importance obtained from four feature selection methods respectively (see supplementary 1). Considering that the ranking results do not follow a normal distribution and the assumptions of equal variances, we conducted a significance analysis of the differences in these rankings using the Wilcoxon signed-rank test, instead of the paired t-test.

Our results showed that the difference between Variable ranking and Permutation ranking is significant at P-value 0.05 level. The difference between Variable ranking and ranking of Correlation Matrix or SHAP were not significant. The difference between Permutation ranking and ranking of Correlation Matrix or SHAP were not significant. The difference between Correlation Matrix ranking and SHAP ranking was not significant also (Table 2.).

Table 2
significant difference analysis of feature importance results of four methods by Wilcoxon signed-rank test

Comparison	Test Statistic	P-value	significance
variable_ranking vs permutation_ranking	871839.0	0.046	YES
variable_ranking vs heatmap_ranking	914741.0	0.823	NO
variable_ranking vs SHAP_ranking	919254.5	0.970	NO
heatmap_ranking vs permutation_ranking	904463.5	0.518	NO
permutation_ranking vs SHAP_ranking	918051.0	0.931	NO
SHAP_ranking vs heatmap_ranking	919427.5	0.976	NO

Compare to the importance results of other three methods, SHAP importance provide very rich information of negative contribution genes (Supplementary 1.). Understanding the positive and negative contributions is vital for studying the gene's function and its role in plant physiological activities. Consequently, in the subsequent biological analysis, we made use of the SHAP importance results from our research.

Result 3. The soybean branching related gene analysis

According to our PCA analysis findings, 780 features account for 95% of the variance associated with soybean branching biological activities. Furthermore, 1202 genes explain 99% of the variance contributing to these activities, while 1752 genes account for 99.99% of the variance. In this section, we focused on the top 1202 genes (representing 99% of the variance contribution) based on their SHAP feature importance rankings.

By utilizing BLAST, we compared the sequences corresponding to these 1202 SNPs with the annotated genes in the soybase database (https://www.soybase.org/). Out of these, 253 SNPs exhibited a complete match with their respective genes (Supplementary 2.). Out of these 253 genes, 9 pairs of SNPs correspond to the same gene, resulting in 244 unique and non-overlapping genes. We conducted a Gene Ontology (GO) analysis on these 244 genes and mapped their respective positions to the soybean chromosomes. Our mapping results showed that these 244 genes were clearly grouped into three clusters on chromosome I and II (Fig. 5.).

Our statistical findings indicate the presence of 177 genes on chromosome 1, comprising 106 genes with a negative contribution and 71 genes with a positive contribution. Additionally, chromosome 2 contains 33 genes, consisting of 25 genes with a negative contribution and 8 genes with a positive contribution (refer to Table 3). Remarkably, a substantial portion, 86.07% (210 out of 244), of the identified genes associated with branching are located on chromosomes 1 and 2. Unraveling the functions of these genes and understanding their reactivation mechanisms will be instrumental in uncovering the precise control mechanisms governing soybean branching.

Table 3
the Statistics of Gene Distribution

chromosome	negative gene number	positive gene number	Total genes
01G	111	74	185
02G	25	8	33
03G	3	1	4
04G	0	1	1
05G	0	1	1
06G	2	0	2
07G	1	0	1
08G	0	0	0
09G	1	2	3
10G	1	0	1
11G	6	2	8
12G	1	0	1
13G	0	0	0
14G	0	1	1
15G	1	0	1
16G	2	1	3
17G	0	0	0
18G	0	2	2
19G	1	2	3
20G	1	2	3
Total	156	97	253

We conducted GO enrichment analysis on these 244 genes from three aspects: molecular function, cellular components, and biological process. Among these 244 genes, 25 of them were not found in any GO Ontologies, accounting for 10.25% of the total number. Our analysis results revealed that the GO terms related to Biological Processes could be clustered into 15 categories, with a total occurrence of 45 genes. The most prominent category was "signal transduction" (11 out of 45), followed by "cell differentiation" and "lipid metabolic process," each accounting for 4 out of 45 genes. Regarding Molecular Function, the GO terms could be grouped into 20 categories, with a total of 193 gene occurrences. The most prevalent category was "protein binding" (49 out of 193), followed by "sequence-specific DNA binding transcription factor activity" (22 out of 193), and "DNA binding" (20 out of 193). Concerning Cellular Components, the GO terms could also be classified into 20 categories, with a total of 404 gene occurrences. The most significant category was the "nucleus" (113 out of 404), followed by "cytoplasm" (55 out of 404), and "plasma membrane" (51 out of 404). For detailed results, please refer to Fig. 6 and Supplement 3.

Furthermore, we performed Gene Ontology enrichment analysis using the agriGO database. The outcomes revealed the functional distribution of 244 genes associated with biological processes (Fig. 7). Notably, these processes exhibited a significant level (level 19) of overall metabolic activities. We observed a negative regulation between multicellular organismal processes and cell recognition. Additionally, a complex interplay of negative and positive regulations among reproduction-related processes, including reproductive process, pollination, pollen-pistil interaction, and recognition of pollen were detected (Fig. 7).

Both of the chromosome location pattern of 244 genes and the negative regulation among physical activity revealed by GO enrichment analysis verified that the branching related genes identified in this study were accurate and reliable.**Discussion**:

1. Linear Regression Models or Nonlinear Artificial Intelligence Models

In the realm of animal and plant breeding, many contemporary breeding approaches continue to rely on linear regression models or manually constructed linear features to capture the interactions between genotypes and phenotypes. However, the intricate diversity and complexity of gene interactions, such as complementary effects, additive effects, duplicate effects, epistatic dominance, epistatic recessiveness, and inhibiting

effects, pose challenges for linear regression models to accurately represent the genotype-phenotype relationships. Nonlinear models, on the other hand, excel at handling high-dimensional data and capturing intricate patterns, making them particularly advantageous when dealing with large and heterogeneous datasets commonly encountered in plant breeding. A recent review by Montesinos-López et al. compared 23 independent studies on linear and nonlinear prediction performance, indicating that nonlinear models outperformed linear ones in 47% of the studies when considering gene-environment interactions (G × E) and in 56% when ignoring G × E interactions[35]. Another study by Gabur et al. using real plant breeding program data demonstrated that Machine Learning methods have the potential to outperform current approaches, increasing prediction accuracies, drastically reducing computing time, and improving the detection of important alleles involved in qualitative or quantitative traits[36]. Traditional univariate and multivariate statistics have limited efficiency in analyzing data affected by the complex interactions between genotypes and environments (G × E). To handle the large-scale and non-deterministic nature of such data, nonlinear nonparametric machine learning techniques are more effective than classical statistical models [37].

Modern biological data, encompassing genomic sequence analysis, SNP chip arrays, and hyperspectral phenomics, often involves high dimensionality, necessitating effective tools to understand the underlying genetic mechanisms and identify patterns associated with specific traits [31]. A profound comprehension of the biological interactions between the genetic makeup of a genotype and its environmental conditions is vital in understanding rare diversity [36]. Machine learning models have proven highly valuable, particularly in dealing with large heterogeneous datasets frequently encountered in plant breeding populations [38]. While current literature does not clearly establish deep learning's superiority over conventional genome-based prediction models in terms of prediction power, deep learning algorithms are more adept at capturing nonlinear patterns. Their ability to integrate data from various sources, a requirement in genomic selection-assisted breeding, leads to improved prediction accuracy for large plant breeding datasets. Applying deep learning to extensive training-testing datasets is crucial for maximizing its benefits [24]. An advantage of deep learning for genomic prediction over standard linear model methods lies in its potential to consider all genetic interactions, including dominance and epistasis, which are crucial in most polyploids. Nonlinear system models, such as deep neural networks, have the capability to analyze and account for complex non-additive effects, setting them apart from linear regression models. Optimizing deep learning algorithms significantly enhances the predictive capacity of whole-genome selection, as they can handle the complexities of gene interactions.

Although there is an ongoing debate, certain studies [39] have shown that no single linear or nonlinear algorithm outperforms across all species and trait combinations. Ensemble predictions, which combine results from multiple algorithms, consistently performed well. Nonlinear algorithms performed best for a similar number of traits, but their performance varied more between traits. Artificial neural networks, while not being the best performer for any trait, identified strategies, such as feature selection and seeded starting weights, that boosted their performance to nearly match other algorithms. These results highlight the significance of carefully selecting the appropriate algorithm for trait value prediction.

In this study, we focused on exploring 11 different nonlinear regression artificial intelligence models. After comparing prediction accuracy and overfitting, we selected Support Vector Regression (SVR) as the non-linear regression model to explore branching-related genes. SVR, a supervised machine learning algorithm used for regression tasks, aims to find a function that best fits the training data while minimizing margin violations. Unlike traditional regression algorithms, SVR focuses on finding a hyperplane with the maximum margin and fewest training samples falling within that margin. SVR offers advantages over traditional regression methods, including robustness against outliers and the ability to handle non-linear relationships between features and targets. However, it can be computationally expensive, especially for large datasets. We utilized high-quality soybean branching phenotype data from 1918 soybean accessions along with the corresponding 42k SNP genotype data in our research.

In conclusion, our research indicates that nonlinear artificial intelligence models exhibit good performance in predicting from genotype to phenotype. However, comparing the strengths and weaknesses of linear and nonlinear models requires rigorous comparisons and extensive testing beyond a small number of tests. The optimal choice of model should be carefully considered based on specific data and research objectives.

1. Feature Dimensionality and Algorithmic Advances

Genomic selection relies on advancements in the availability of genotyping data to predict agriculturally relevant phenotypic traits [40]. Various factors affecting the predictability of phenotype prediction have been extensively discussed [41,42]. The success of phenotype prediction is contingent on the quality of the training models and the characteristics of the datasets used for training and testing. Consequently, the number of observations and markers utilized is expected to play a crucial role in determining predictability [40]. With the development of sequencing technology, a large amount of genotype data will be generated, such as the up to 42K SNP markers used in our study. In the context of genomic selection, there arises a scenario where the number of loci assayed (p) exceeds the number of accessions (samples) (n), commonly known as the "curse of dimensionality" ($n \ll p$)[43]. Excessively high dimensions increase data noise and processing of high-dimensional data requires large computational power, while it may overfit a model and generate poor prediction. The concept of feature dimension refers to the number of features relative to the sample size. When both the dimension and sample size are substantial, predictability issues are generally not significant. However, problems arise when the sample size is limited, and the samples are described with an excessive number of dimensions, particularly in

the presence of noise. In such cases, a sufficient sample size is required to effectively discover patterns and enhance prediction accuracy. Conversely, when the sample size is relatively small compared to the feature dimension, leading to an excess of feature dimensions, overfitting becomes a concern. Therefore, reducing the feature dimension reasonably becomes essential to improve prediction accuracy. Dimensionality reduction algorithms are commonly employed to achieve this goal, involving the selection of the most representative features in the dataset to reduce their quantity. Two main algorithms for dimensionality reduction are Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA). The primary difference lies in how they identify new features. LDA uses class information to search for new features that maximize separability between classes, while PCA achieves this by considering the variance of each feature. In this study, we mainly utilize the PCA dimensionality reduction algorithm to reduce the feature dimensions. During data preprocessing, we employ Principal Component Analysis (PCA) to reduce noise in the samples. Subsequently, through feature importance analysis (feature selection), we identify genes or molecular markers closely related to the target trait (branching) to enhance the model's prediction accuracy. Simulation studies have indicated that a larger number of markers leads to improved phenotype prediction accuracy [44]. Similarly, increasing the number of samples in the training population can also yield better phenotype prediction accuracy [45]. In a comparison of six linear and six non-linear algorithms, Azodi et al. found that hyperparameter selection was necessary for all non-linear algorithms, and feature selection before model training was crucial for artificial neural networks when the number of markers greatly exceeded the number of training lines[39].

Bommert et al. investigated 16 high-dimensional data sets and compared 22 feature selection methods in terms of accuracy and computing time. Their findings suggested that there is no group of feature selection methods that outperformed all others; however, some filter methods performed much better on a specific data set[46]. Similar results were observed also in our study (Table 2.).

One of the simplest techniques to improve machine learning models is to select a better machine learning algorithm. In general, unsupervised feature selection methods are less prone to overfitting [47], and have the ability to improve predictions, while discarding redundant data points. Model-agnostic local explanation methods, such as Shapley additive explanations (SHAP) and Local interpretable model-agnostic explanations (LIME), have the potential to overcome interpretability issues by consistently and transparently quantifying the input's effect on prediction across most model types[48]. However, concerns remain about potential bias, engineered explanations, and the risk of false conclusions made by inexperienced users. Explainability for genotype to phenotype prediction is an emerging area in genomic prediction studies. The interest in explainability and interpretability suggests that new deep learning algorithms may offer enhanced interpretability [35]. Additionally, interpretable machine learning extracts associations based on the correlation between features and outcomes, but interpretations often aim to suggest possible causal relationships from features, necessitating further investigation to establish causality. Factors like untested variables not included in the genotype data, such as epigenetic or environmental features, may cause both the genomic region and the predicted phenotype to be associated with each other without the genomic region being the causal feature. Therefore, the relationship between the training and testing populations significantly impacts prediction [40]. The ultimate validation of phenotype prediction models lies in their performance on breeding populations.

Conclusion

In this study, we focused on phenotype prediction as a crucial step in Smart-breeding and breeding specialist systems. We conducted a thorough comparison of 11 non-linear regression artificial intelligence models and found that SVR, ANN, and Autoencoder exhibited superior performance in phenotype prediction when compared to other models. These three models are well-suited for use in genomic selection and accurate phenotype prediction.

Moreover, we evaluated four feature importance algorithms and discovered that there were no significant differences among them. However, the SHAP method stood out as it could differentiate between genes with positive and negative contributions, providing valuable information on genes with negative effects. Consequently, SHAP proved to be more suitable than other feature importance methods for genetic research.

To validate our feature importance results, we performed GO analysis, and the genes identified using the combination of SVR and SHAP approach were effectively grouped into three clusters in the entire soybean genome view. The GO enrichment results further supported the accuracy of our predictions.

The method developed in this research has broad applicability in phenotype prediction, minor QTL mining, and plant smart-breeding systems. It marks a significant contribution to the advancement of AI-breeding and fosters the transition from experience-based breeding to data-based breeding.

Dataset and methods

1. Dataset

The original genotypic data is from soybase data bank: https://soybase.org/snps/. The SoySNP50K iSelect BeadChip has been used to genotype the USDA Soybean Germplasm Collection [49]. The complete data set for 20,087 G. max accessions genotyped with 42,509 SNPs is available.

Soybean accessions and phenotypic data used in this study were obtained from the USDA Soybean Germplasm Collection (http://www.ars-grin.gov/npgs/). Branching phenotype data was extracted and used for analysis. Total 1918 accessions are available. Missing data and SNPs with minor allele frequencies below 0.1 were excluded, leaving 42,291 SNPs for analysis.

2. Data preprocessing

The genotype data was loaded into a pandas DataFrame, where the columns represented the SampleID and Genotype information (Reference SNP cluster ID, i.e. RS number). The Genotype column contained the genetic data for each sample, represented using the letters 'A', 'T', 'C', 'G', and '-' for absent data.

To prepare the data for machine learning algorithms, we performed one-hot encoding, a process that converted the categorical genotype data into a numerical format. Each unique genotype category was transformed into a binary vector, with 1 marking the category and 0 marking all other categories. The OneHotEncoder from scikit-learn was employed to perform this encoding. The result was a new DataFrame with columns for each unique genotype category. We then reintegrated the SampleID column back into the one-hot encoded DataFrame, preserving each sample's identifier alongside its corresponding one-hot encoded genotype information.

To enhance data analysis and reduce noise, we employed Principal Component Analysis (PCA) from scikit-learn. PCA is a dimensionality reduction technique that transformed the high-dimensional one-hot encoded genotype data into a lower-dimensional representation. By finding new orthogonal axes (principal components), PCA captured the maximum variance in the data. We saved the results of the PCA analysis in a new DataFrame while retaining the SampleID column to identify each sample.

Subsequently, both the processed genotype data and phenotype data were sorted by the SampleID, ensuring their alignment for use in training machine learning models.

3. Model training and evaluatiion

In our research, we employed a total of 11 non-linear regression models, comprising both machine learning and deep learning models. Specifically, we utilized seven machine learning models and four deep learning models for our investigation. The machine learning models included Support Vector Regression (SVR), Extreme Gradient Boosting (XGBoost) regression, Random Forest regression, LightGBM regression, Gaussian Processes (GPS) regression, Decision Tree regression, and Polynomial regression. On the other hand, the deep learning models encompassed Deep Belief Network (DBN) regression, Artificial Neural Networks (ANN) regression, Autoencoders regression, and Multi-Layer Perceptron (MLP) regression.

To assess each model's performance, we loaded the sorted dataset, which contained the features (X) and the target variable (y), into every model. To enhance model performance, we partitioned the data into training and testing sets using the 'train_test_split' function from the scikit-learn library with a test size of 0.2.

For hyperparameter tuning, we defined a hyperparameter grid using the 'param_grid' dictionary and conducted a grid search to find the best hyperparameters for each model. We used the GridSearchCV from scikit-learn, which performed a systematic exploration of different hyperparameter combinations and selected the best combination based on the Mean Absolute Error (MAE) scoring metric. To ensure robust evaluation, we employed 5-fold cross-validation (cv = 5) to evaluate the performance of each hyperparameter combination.

After the grid search, we obtained the best model for each regression technique, including the best hyperparameters determined during the cross-validation process. These best models were then used to make predictions on the test set (X_test), with the 'predict' method providing the predicted target values (y_pred) based on the test features.

To evaluate the performance of each model, we considered four key metrics: R2 Score, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Mean Absolute Percentage Error (MAPE). R2 Score, ranging from 0 to 1, quantifies the proportion of variance in the target variable explained by the model, with 1 indicating a perfect fit. MAE calculates the average of the absolute differences between the predicted and true target values, representing the magnitude of errors. MSE computes the average of the squared differences between the predicted and true target values, penalizing larger errors more than MAE. MAPE measures the percentage difference between the predicted and true target values, providing a relative error metric.

SVR (Support Vector Regression) is a powerful algorithm for non-linear regression tasks. It finds a hyperplane that best fits the data while minimizing error. SVR uses kernel functions to map data into a higher-dimensional space, allowing complex decision boundaries. Tuning hyperparameters like the kernel type and regularization parameter is essential for optimal performance. SVR is versatile and accurate, suitable for handling complex datasets.

Random forest regression is used to detect nonlinear combinations of variables and complex interactions among them. We determined variable importance using the "h2o.randomForest" function from H2O.ai (version 3.32.0.4) with "ntrees = 1,000" as the hyper-tuned parameter.

XGBoost (Extreme Gradient Boosting) is an ensemble learning algorithm that combines the predictions of multiple weak models (decision trees) to create a strong predictive model. It handles complex datasets, provides feature importance analysis, and efficiently handles missing values. Appropriate hyperparameter tuning and feature engineering are crucial for optimal performance.

Deep Belief Network (DBN) is a generative model with multiple layers of hidden units, usually used for unsupervised learning. It can be adapted for regression tasks to model the relationship between input features and a continuous target variable.

Artificial Neural Networks (ANNs) can be used for regression tasks, training on gradient data to predict continuous output values based on input features. Proper architecture selection, data preprocessing, and hyperparameter tuning are essential for good performance.

Autoencoder is a type of neural network used for feature learning and data compression. It can be adapted for regression tasks using denoising techniques or traditional autoencoder architectures.

Polynomial regression models nonlinear relationships between a dependent variable and independent variables. The model takes the form: $y = \beta + \beta x + \beta x^2 + ... + \beta x + \varepsilon$. Careful model selection and evaluation are needed to avoid overfitting.

Multilayer Perceptron (MLP) Regression is a powerful supervised learning algorithm based on artificial neural networks. It consists of multiple layers of interconnected neurons and can learn complex non-linear relationships between input features and target values.

Decision tree regression builds a tree based on the training data to predict continuous numerical values. The model splits the data based on different features and thresholds to minimize the overall variance or mean squared error of the target variable.

Gaussian Processes (GPS) Regression is a probabilistic machine learning technique that models the underlying function mapping inputs to outputs using Gaussian processes. It provides a distribution over possible functions and handles uncertainty in regression tasks. GPS regression is powerful but can be computationally demanding for large datasets.

Gradient boosting machines are an ensemble learning method used for prediction. They combine weak prediction models to improve predictive performance and are known for their fast-computational learning. The model construction is done stage-wise and is generalized using a differentiable loss function. LightGBM is an efficient gradient boosting framework widely used for regression tasks. It is known for its speed and scalability, making it suitable for handling large-scale datasets.

4. Feature importance analysis

The main objective of the feature importance analysis is to identify the most impactful features that influence predictions in a testing model. In our research, we explored four feature importance algorithms based on the SVR model: Variable Ranking, Correlation Matrix, Permutation Importance, and SHAP (SHapley Additive exPlanations).

To initiate the analysis, we loaded the sorted dataset containing the features (X) and the target variable (y). We then divided the data into training and testing sets using the "train_test_split" function from scikit-learn, with a test size of 0.2. Subsequently, an SVR model was created and trained with the use of optimal hyperparameters obtained from GridsearchCV. We employed this trained SVR model in various feature importance algorithms to compute feature importance scores.

Variable Ranking, a widely-used approach in feature selection analysis, was utilized to rank features based on their importance concerning the target variable. In the case of linear regression or linear kernel SVR, the feature importance is directly determined by the coefficients (or weights) obtained from the fitted model. Features with higher absolute coefficient values are considered more influential in making predictions.

The Correlation Matrix method was employed to investigate relationships between features and the target variable, as well as among different features. Features with strong correlations to the target variable were deemed more important, while lower correlations among features were preferred to avoid multicollinearity, which could lead to instability in the model's estimates. A heatmap was used to visualize the strength of the correlations.

Permutation Importance, a model-agnostic method, was applied to evaluate the significance of each feature through post-training analyses. This involved randomly permuting the values of a single feature while keeping other features unchanged and measuring the impact on the model's performance metric (e.g., accuracy, R2 score, or Mean Absolute Error). If shuffling a feature resulted in a significant decline in model performance, it indicated the importance of that specific feature. Permutation Importance provided a model-agnostic means to assess the contribution of individual features to the model's overall performance.

Finally, we employed SHAP (SHapley Additive exPlanations), an advanced and model-agnostic feature importance method based on game-theoretically optimal Shapley values. SHAP values were computed at the individual prediction level, considering all possible feature interactions. They explained how a feature's value contributed to the prediction for a specific data point compared to a baseline reference. SHAP values provided insights into both the direction and magnitude of feature influence, following cooperative game theory principles. SHAP was especially useful for interpreting complex models and explaining individual predictions in a model-agnostic manner [50, 51]. In our research, we utilized the "shap.utils.permutation_importance" function to determine the permutation SHAP feature importance scores, which complemented our analysis.

5. Gene Ontology analysis

SNPs identified by feature importance analysis were searched in SoyBase data site (https://soybase.org/snps/) by RS number. And the flank sequence of corresponding SNP was used to BLAST in Glycine max Genome DB database (http://www.plantgdb.org/GmGDB/) for confirmation. The gene names which SNPs hit to the same location (including CDS, UTR and intron) were collected for GO (gene ontology) analysis. All the genes identified by BLAST were analyzed by GO term enrichment tool at SoyBase website (https://soybase.org/goslimgraphic_v2/dashboard.php). The GO distribution analysis, related charts and gene location map were generated by GO term enrichment tool at SoyBase website (https://www.soybase.org/goslimgraphic_v2/dashboard.php). Gene ontology enrichment analysis (Fig. 7)was carried out by Analysis tools Glycine max Singular Enrichment Analysis (SEA) at agriGO database (http://bioinfo.cau.edu.cn/agriGO/) [52,53].

6. Wilcoxon Signed Rank Test

The Wilcoxon Signed Rank Test is a non-parametric statistical test used to compare the mean ranks of two related or paired samples. It is a suitable alternative when data does not meet the normality assumptions required by parametric tests like the paired t-test. This test is applicable when the data is measured on at least an ordinal scale. The Wilcoxon Signed Rank Test is commonly used in the following scenarios: Paired Data: When you have two sets of related or paired data points, such as before and after measurements on the same subjects. Non-Normal Data: When the data is not normally distributed, making parametric tests inappropriate.

The test works by calculating the differences between the paired observations, ranking these differences based on their absolute values (disregarding direction), and summing the positive and negative ranks separately. The test statistic is the smaller of the two sums. By comparing the test statistic to critical values from the Wilcoxon Signed Rank Table, the significance level can be determined.By default, the Wilcoxon Signed Rank Test is a one-tailed test, assessing if the median of the differences between paired samples is significantly different from zero. It can be modified for a two-tailed test to check if the median is significantly different from a specific value.The Wilcoxon Signed Rank Test is valuable for working with non-parametric data and paired samples, providing an alternative to parametric tests like the paired t-test for hypothesis testing. If the test statistic is statistically significant (falling in the critical region), it indicates a significant difference between the paired observations. If not statistically significant, there is insufficient evidence to conclude a significant difference.

Declarations

Acknowledgements

This project received funding from the Dean's Fund of College of Agriculture and Food Science, Florida Agricultural and Mechanical University, Tallahassee, Florida.

Funding

Open access funding provided by College of Agriculture and Food Science, Florida Agricultural and Mechanical University.

Author information

Authors and Affiliations

Wei Zhou, College of Agriculture and Food Science, Florida Agricultural and Mechanical University, Tallahassee, Florida, 32307, USA

Zhengxiao Yan, FAMU-FSU College of Engineering, Florida State University, Tallahassee, Florida, 32306, USA

Liting Zhang, Department of Computer Science, Florida State University, Tallahassee, FL, 32306, USA

Contributions

Conceptualization: W. Z., Original draft: W.Z. Writing and editing: W. Z., Z. Y., L. Z, Supervision: W. Z., All authors reviewed and approved final version.

Corresponding author

Correspondence to Wei Zhou, wei.zhlou@famu.edu

Ethics declarations

Competing interests

The authors declare no competing interests.

Additional information

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary Information

Supplementary Information.

Supplementary 1, 2 and 3

Data Availability

The datasets generated and/or analyzed during the current study are available in the ScienceDB repository, https://www.scidb.cn/en/anonymous/MkFWdklm.

References

1. Bongaarts J. Human population growth and the demographic transition. Philosophical Transactions of the Royal Society B: Biological Sciences. 2009;364(1532):2985–90.
2. Lutz W, KC S. Dimensions of global population projections: what do we know about future population trends and structures? Philosophical Transactions of the Royal Society B: Biological Sciences. 2010;365(1554):2779–91.
3. Searchinger T, Waite R, Hanson C, Ranganathan J, Dumas P, Matthews E, Klirs C. Creating a sustainable food future: A menu of solutions to feed nearly 10 billion people by 2050. 2019, Final report.
4. Thornton PK, Lipper L. How does climate change alter agricultural strategies to support food security? Intl Food Policy Res Inst; 2014 Apr 11.
5. Oliver TH, Morecroft MD. Interactions between climate change and land use change on biodiversity: attribution problems, risks, and opportunities. Wiley Interdisciplinary Reviews: Climate Change. 2014;5(3):317–35.
6. Tester M, Langridge P. Breeding technologies to increase crop production in a changing world. Science. 2010;327(5967):818–22.
7. Chowdhury S, Kumar S. Okra breeding: recent approaches and constraints. Annals of Biology. 2019;35(1):55–60.
8. Kulwal PL, Mir RR, Varshney RK. Efficient breeding of crop plants. InFundamentals of Field Crop Breeding 2022 (pp. 745–777). Springer, Singapore.
9. Hilgartner S. Reordering life: knowledge and control in the genomics revolution. MIT press; 2017 May 19.
10. Bhat JA, Ali S, Salgotra RK, Mir ZA, Dutta S, Jadon V, Tyagi A, Mushtaq M, Jain N, Singh PK, Singh GP. Genomic selection in the era of next generation sequencing for complex traits in plant breeding. Frontiers in genetics. 2016;7:221.
11. Joyce AR, Palsson BØ. The model organism as a system: integrating'omics' data sets. Nature reviews Molecular cell biology. 2006;7(3):198–210.
12. Mallick H, Rahnavard A, McIver LJ, Ma S, Zhang Y, Nguyen LH, Tickle TL, Weingart G, Ren B, Schwager EH, Chatterjee S. Multivariable association discovery in population-scale meta-omics studies. PLoS computational biology. 2021;17(11):e1009442.
13. Feuillet C, Leach JE, Rogers J, Schnable PS, Eversole K. Crop genome sequencing: lessons and rationales. Trends in plant science. 2011;16(2):77–88.
14. Wei L, Xiao M, Hayward A, Fu D. Applications and challenges of next-generation sequencing in Brassica species. Planta. 2013;238(6):1005–24.
15. Shakoor N, Lee S, Mockler TC. High throughput phenotyping to accelerate crop breeding and monitoring of diseases in the field. Current opinion in plant biology. 2017; 38:184–92.
16. Wang X, Xu Y, Hu Z, Xu C. Genomic selection methods for crop improvement: Current status and prospects. The Crop Journal. 2018;6(4):330–40.
17. Mochida K, Koda S, Inoue K, Hirayama T, Tanaka S, Nishii R, Melgani F. Computer vision-based phenotyping for improvement of plant productivity: a machine learning perspective. GigaScience. 2019;8(1):giy153.

18. Grinberg NF, Orhobor OI, King RD. An evaluation of machine-learning for predicting phenotype: studies in yeast, rice, and wheat. Machine Learning. 2020;109(2):251–77.
19. Feng X, Zhan Y, Wang Q, Yang X, Yu C, Wang H, Tang Z, Jiang D, Peng C, He Y. Hyperspectral imaging combined with machine learning as a tool to obtain high-throughput plant salt-stress phenotyping. The Plant Journal. 2020;101(6):1448–61.
20. Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, De Los Campos G, Burgueño J, González-Camacho JM, Pérez-Elizalde S, Beyene Y, Dreisigacker S. Genomic selection in plant breeding: methods, models, and perspectives. Trends in plant science. 2017;22(11):961–75.
21. Pérez-Enciso M, Zingaretti LM. A guide on deep learning for complex trait genomic prediction. Genes. 2019;10(7):553.
22. van Dijk AD, Kootstra G, Kruijer W, de Ridder D. Machine learning in plant science and plant breeding. Iscience. 2021;24(1):101890.
23. Montesinos-López OA, Martín-Vallejo J, Crossa J, Gianola D, Hernández-Suárez CM, Montesinos-López A, Juliana P, Singh R. A benchmarking between deep learning, support vector machine and Bayesian threshold best linear unbiased prediction for predicting ordinal traits in plant breeding. G3: Genes, Genomes, Genetics. 2019;9(2):601 – 18.
24. Zingaretti LM, Gezan SA, Ferrão LF, Osorio LF, Monfort A, Muñoz PR, Whitaker VM, Pérez-Enciso M. Exploring deep learning for complex trait genomic prediction in polyploid outcrossing species. Frontiers in plant science. 2020;11:25.
25. Li B, Zhang N, Wang YG, George AW, Reverter A, Li Y. Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. Frontiers in genetics. 2018;9:237.
26. Sandhu K, Patil SS, Pumphrey M, Carter A. Multitrait machine-and deep-learning models for genomic selection using spectral information in a wheat breeding program. The Plant Genome. 2021;14(3):e20119.
27. Montesinos-López OA, Montesinos-López A, Crossa J, Gianola D, Hernández-Suárez CM, Martín-Vallejo J. Multi-trait, multi-environment deep learning modeling for genomic-enabled prediction of plant traits. G3: Genes, genomes, genetics. 2018;8(12):3829–40.
28. Okser S, Pahikkala T, Airola A, Salakoski T, Ripatti S, Aittokallio T. Regularized machine learning in the genetic prediction of complex traits. PLoS genetics. 2014;10(11): e1004754.
29. Shook J, Wu L, Gangopadhyay T, Ganapathysubramanian B, Sarkar S, Singh AK. Integrating genotype and weather variables for soybean yield prediction using deep learning. bioRxiv. 2018 May 25:331561.
30. Shook J, Gangopadhyay T, Wu L, Ganapathysubramanian B, Sarkar S, Singh AK. Crop yield prediction integrating genotype and weather variables using deep learning. Plos one. 2021;16(6):e0252402.
31. Yoosefzadeh-Najafabadi M, Earl HJ, Tulpan D, Sulik J, Eskandari M. Application of Machine Learning Algorithms in Plant Breeding: Predicting Yield From Hyperspectral Reflectance in Soybean. Front Plant Sci. 2021;11:624273. doi: 10.3389/fpls.2020.624273. PMID: 33510761; PMCID: PMC7835636.
32. Poland JA, Rife TW. Genotyping-by-sequencing for plant breeding and genetics. The plant genome. 2012;5(3).
33. Berkman PJ, Lai K, Lorenc MT, Edwards D. Next-generation sequencing applications for wheat crop improvement. American journal of botany. 2012;99(2):365–71.
34. Kumar S, Banks TW, Cloutier S. SNP discovery through next-generation sequencing and its applications. International journal of plant genomics. 2012;2012.
35. Montesinos-López O. A., Montesinos-López A., Pérez-Rodríguez P., Barrón-López J. A., Martini J. W. R., Fajardo-Flores S. B., et al. (2021). A Review of Deep Learning Applications for Genomic Selection. BMC Genomics 22, 19. 10.1186/s12864-020-07319-x.
36. Gabur I, Simioniuc DP, Snowdon RJ, Cristea D. Machine Learning Applied to the Search for Nonlinear Features in Breeding Populations. Front Artif Intell. 2022;5:876578. doi: 10.3389/frai.2022.876578. PMID: 35669178; PMCID: PMC9164111.
37. Niazian, M.; Niedbała, G. Machine Learning for Plant Breeding and Biotechnology. Agriculture 2020, 10, 436. https://doi.org/10.3390/agriculture10100436
38. Collins, A., and Yao, Y. (2018). “Machine learning approaches: data integration for disease prediction and prognosis,” in Applied Computational Genomics. Translational Bioinformatics, Vol 13, ed Y. Yao (Singapore: Springer). doi: 10.1007/978-981-13-1071-3_10
39. Azodi CB, Bolger E, McCarren A, Roantree M, de Los Campos G, Shiu SH. Benchmarking parametric and machine learning models for genomic prediction of complex traits. G3: Genes, Genomes, Genetics. 2019;9(11):3691 – 702.
40. Tong H, Nikoloski Z. Machine learning approaches for crop improvement: Leveraging phenotypic and genotypic big data. J Plant Physiol. 2021;257:153354. doi: 10.1016/j.jplph.2020.153354. Epub 2020 Dec 29. PMID: 33385619.
41. Nakaya A, Isobe SN. Will genomic selection be a practical method for plant breeding?. Annals of botany. 2012;110(6):1303–16.
42. Danilevicz MF, Gill M, Anderson R, Batley J, Bennamoun M, Bayer PE, Edwards D. Plant Genotype to Phenotype Prediction Using Machine Learning. Front Genet. 2022;13:822173. doi: 10.3389/fgene.2022.822173. PMID: 35664329; PMCID: PMC9159391.
43. Ramstein GP, Jensen SE, Buckler ES. Breaking the curse of dimensionality to identify causal variants in Breeding 4. Theoretical and Applied Genetics. 2019;132(3):559–67.

44. Solberg TR, Sonesson AK, Woolliams JA, Meuwissen TH. Genomic selection using different marker types and densities. Journal of animal science. 2008;86(10):2447–54.
45. Heffner EL, Jannink JL, Sorrells ME. Genomic selection accuracy using multifamily prediction models in a wheat breeding program. The Plant Genome. 2011;4(1).
46. Bommert A, Sun X, Bischl B, Rahnenführer J, Lang M. Benchmark for filter methods for feature selection in high-dimensional classification data. Computational Statistics & Data Analysis. 2020;143:106839.
47. Guyon I, Elisseeff A. An introduction to variable and feature selection. Journal of machine learning research. 2003;3(Mar):1157–82.
48. Slack D., Hilgard S., Jia E., Singh S., Lakkaraju H. (2020). "Fooling LIME and SHAP," in Proceedings of the AAAI/ACM Conference on AI. New York, NY, USA. Ethics, and SocietyACM, 180–186. 10.1145/3375627.3375830.
49. Song Q, Hyten DL, Jia G, Quigley CV, Fickus EW, Nelson RL, Cregan PB. Fingerprinting soybean germplasm and its utility in genomic research. G3: Genes, genomes, genetics. 2015;5(10):1999–2006.
50. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017;30.
51. Lundberg SM, Erion GG, Lee SI. Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888. 2018 Feb 12.
52. Tian T, Liu Y, Yan H, You Q, Yi X, Du Z, Xu W, Su Z. agriGO v2. 0: a GO analysis toolkit for the agricultural community, 2017 update. Nucleic acids research. 2017;45(W1):W122-9.
53. Du Z, Zhou X, Ling Y, Zhang Z, Su Z. agriGO: a GO analysis toolkit for the agricultural community. Nucleic acids research. 2010;38(suppl_2):W64-70.

Figures

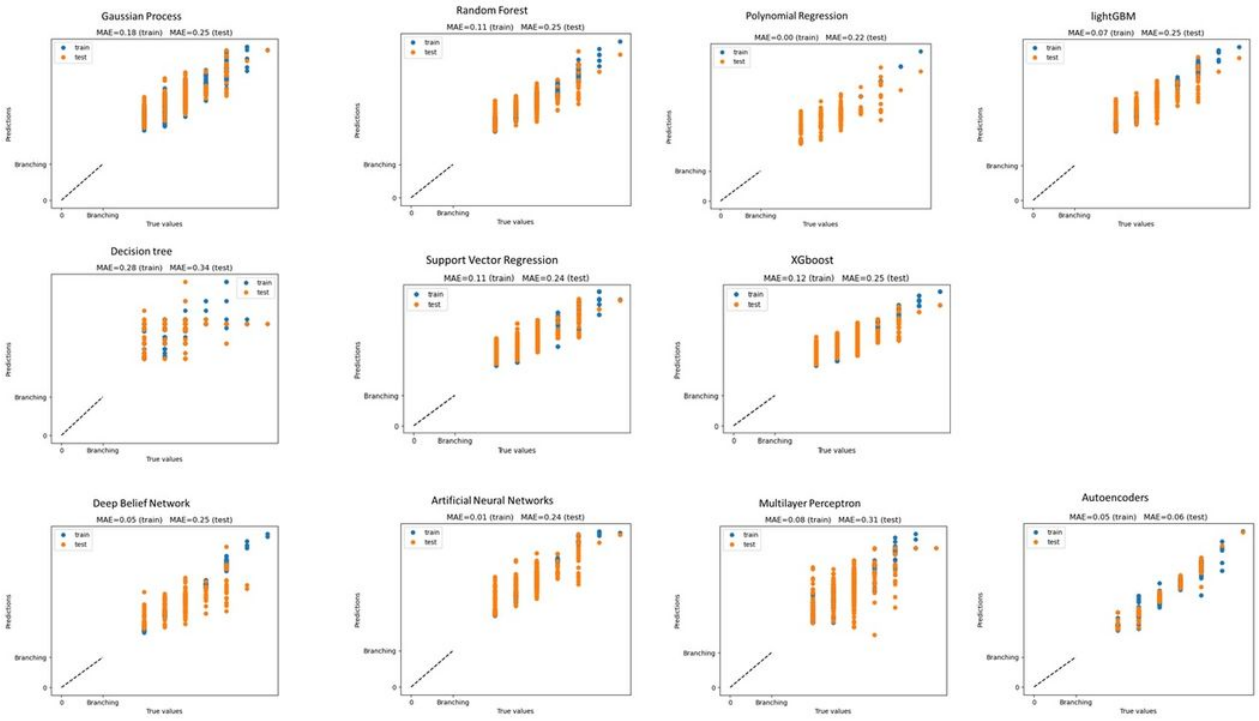

Figure 1

Histogram of Prediction Accuracy Evaluation of 11 Models by MAE Value

In Figure 1, the histogram displays accuracy scores in the model evaluation using Mean Absolute Error (MAE). The blue dots represent the target values of the training data (y_train_pred), while the orange dots correspond to the target values of the testing data (y_test_pred). The X-axis represents the true values, and the Y-axis represents the prediction values.

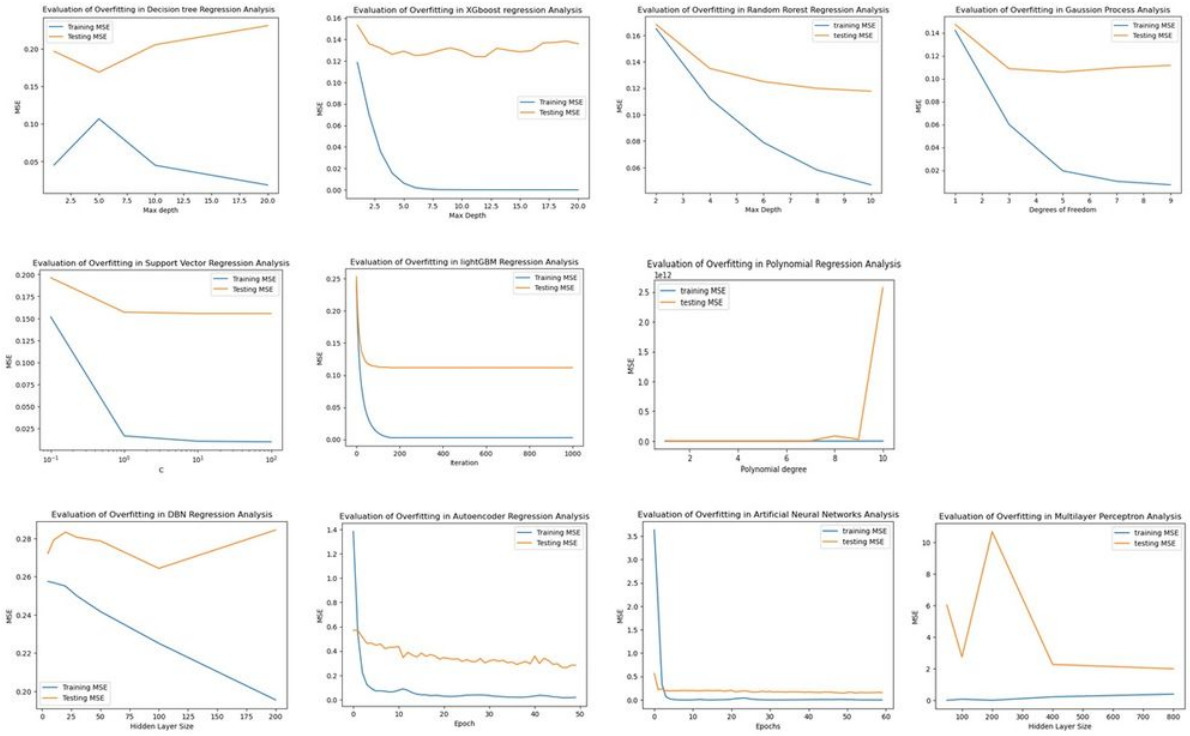

Figure 2

Overfitting Evaluation of 11 Models Based on MSE Value

In Figure 2, the histogram illustrates the evaluation of overfitting for each model. The blue line represents the Mean Squared Error (MSE) of the training data, while the orange line represents the MSE of the testing data. The Y-axis indicates the values of MSE, and the X-axis corresponds to different parameters for each model:

For Decision tree, XGboost, and Random forest, the X-axis represents the max depth.

For Gaussian process, the X-axis represents degrees of freedom.

For SVR (Support Vector Regression), the X-axis represents the c-value.

For lightGBM, the X-axis represents the number of iterations.

For polynomial regression, the X-axis represents the polynomial degree.

For DBN (Deep Belief Network) regression and Multilayer perception, the X-axis represents the hidden layer size.

For Autoencoder and ANN (Artificial Neural Network) models, the X-axis represents the number of epochs.

Each model's performance and overfitting tendencies can be observed and compared using these representations.

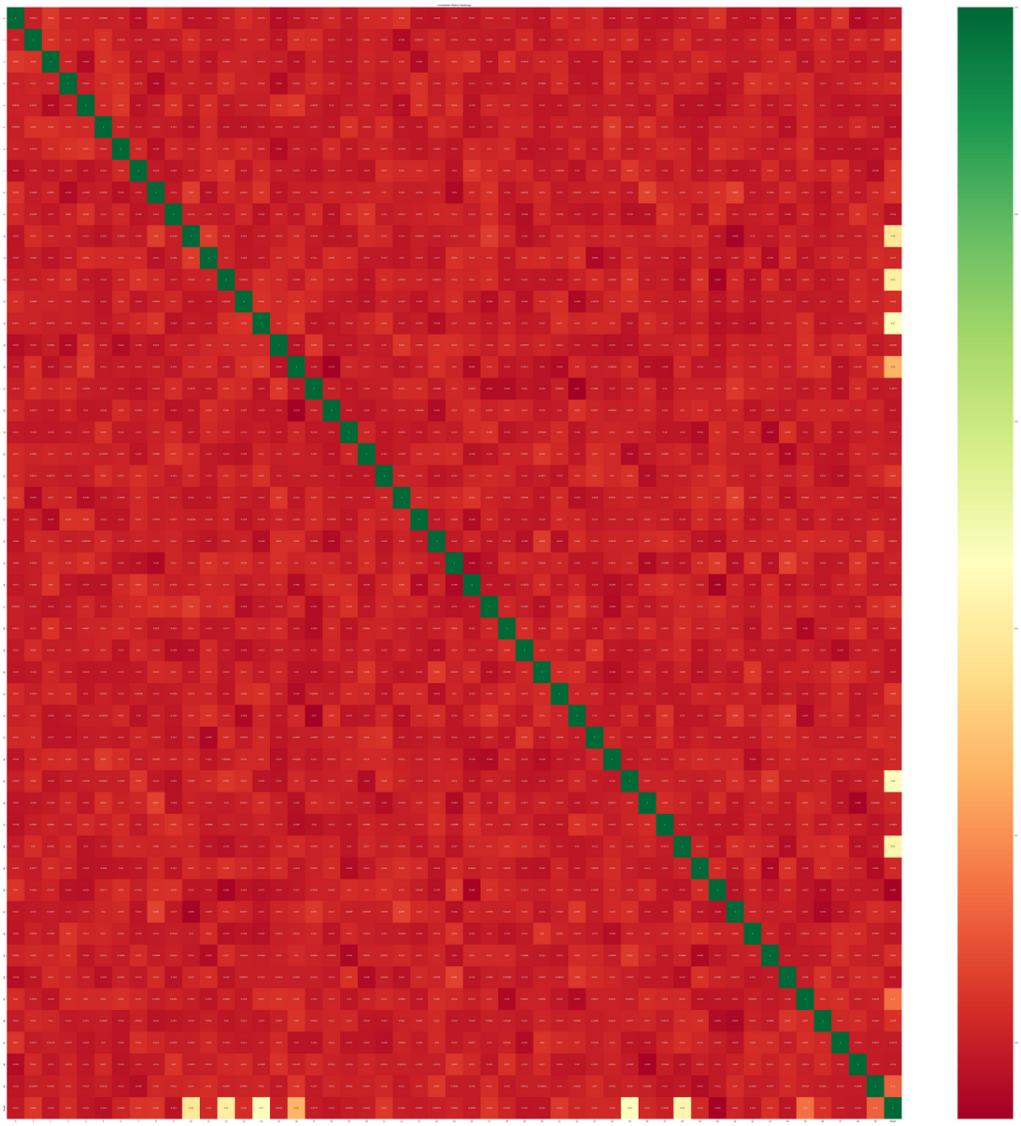

Figure 3

Correlogram of Top 100 Features (SNP) Identified in SVR Correlation Analysis.

The figure 3 displays a heatmap representing the correlations between the top 100 features (Single Nucleotide Polymorphisms - SNP) identified in the SVR (Support Vector Regression) correlation analysis. The heatmap uses varying shades of the color gray, with higher values indicating stronger correlations between the variables. This visualization allows for a clear and visual assessment of the interrelationships among the features, providing valuable insights into their associations and potential implications in the study.

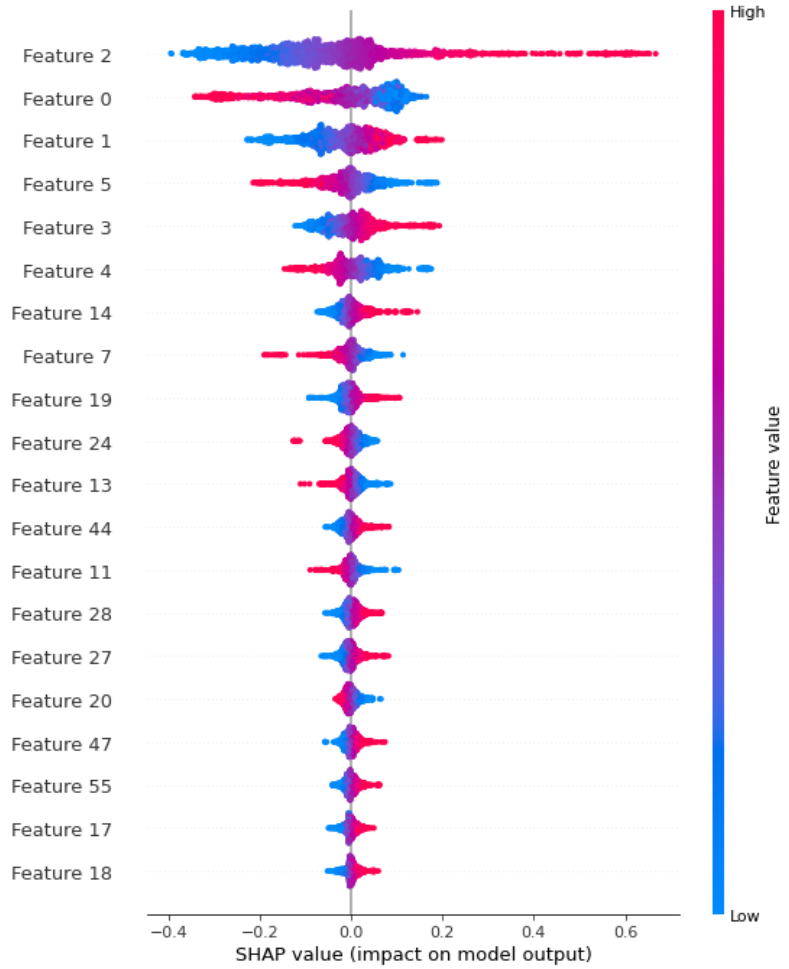

Figure 4

Summary Beeswarm Plot of Top 20 Features from SHAP Importance Analysis based on SVR Model.

This figure presents a beeswarm plot summarizing the top 20 features derived from our SHAP (SHapley Additive exPlanations) importance analysis using the SVR (Support Vector Regression) model. The plot visually captures the relative effect of each feature across the entire dataset, allowing for a comprehensive understanding of their respective influences. The beeswarm plot provides an intuitive representation of the feature importances, aiding in the identification of key contributors to the model's predictions and facilitating insightful data-driven decisions.

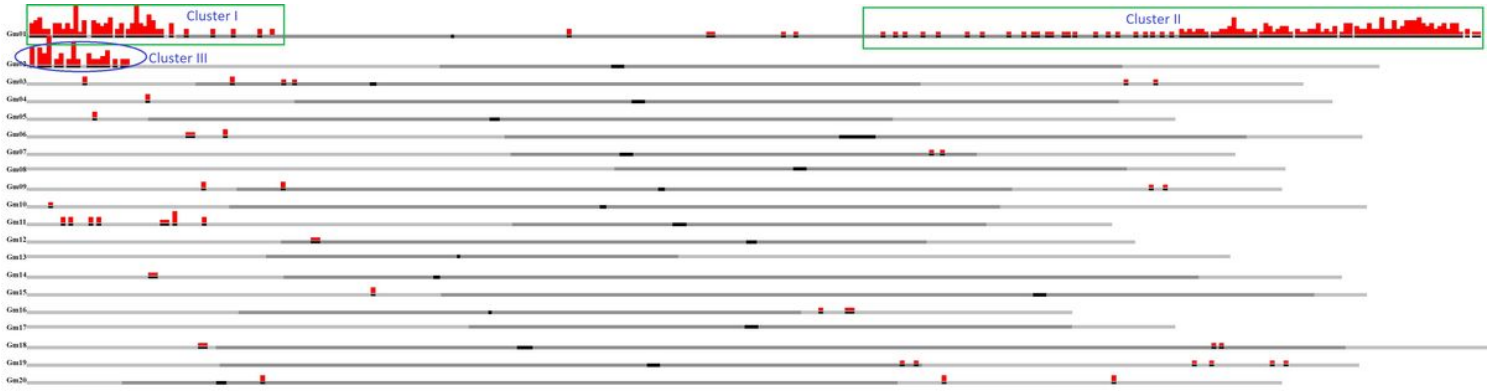

Figure 5

Whole Genome View of 244 Identified Genes

The figure 5 presents a visual representation of 244 identified genes, where each red dot represents a corresponding gene from the BLAST (Basic Local Alignment Search Tool) hit. The genes displayed in the plot are related to soybean branching. Notably, three major clusters of these genes

have been identified and are highlighted within circles on chromosomes 1 and 2. This comprehensive genome view provides valuable insights into the spatial distribution and clustering patterns of the branching-related genes, aiding in the exploration and understanding of their potential functional significance.

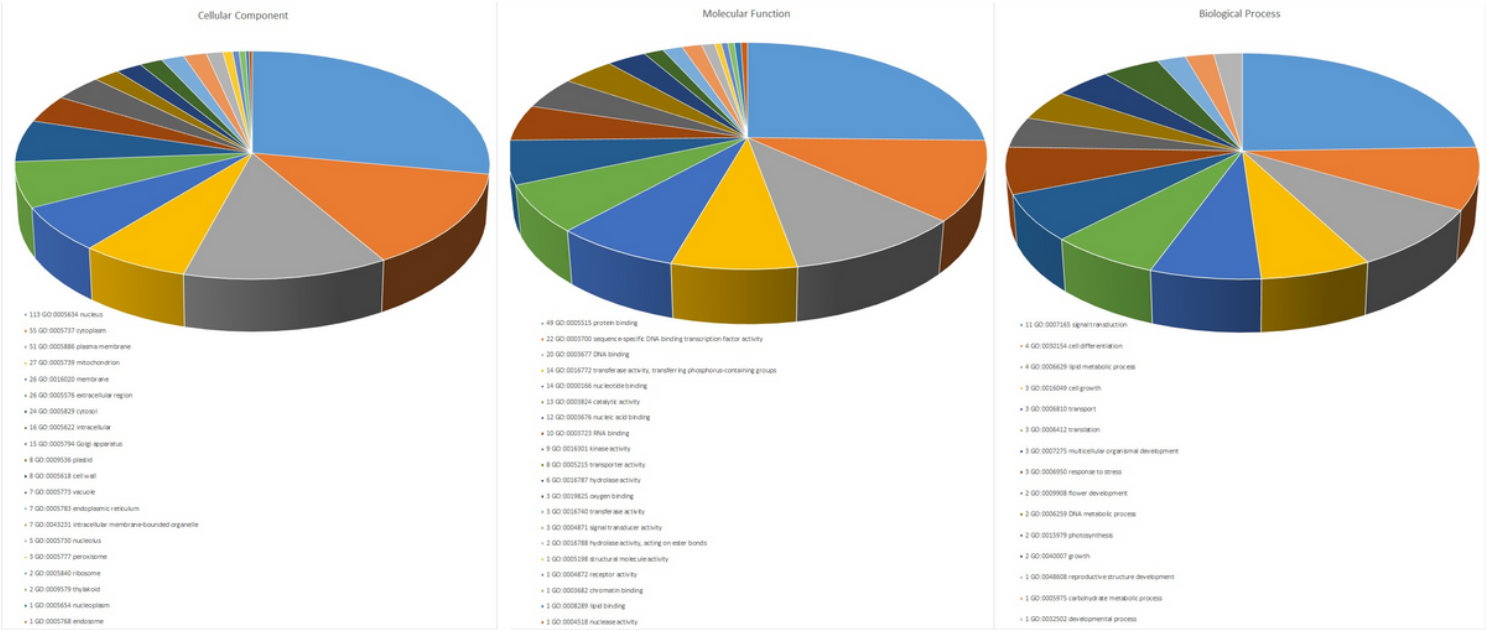

Figure 6

Analysis of GO Ontologies Distribution

The figure 6 displays three pie charts representing the distribution of three kinds of Gene Ontology (GO) ontologies, namely Cellular Component, Molecular Function, and Biological Process. Each pie chart is color-coded to distinguish different types of GO, and the size of each segment represents the proportion of that specific GO type within its respective ontology category. The accompanying number table provides the count of genes associated with each GO type, followed by the ID and category of the corresponding GO term. This analysis provides a comprehensive overview of the functional annotations of the genes in the study, highlighting their involvement in various cellular components, molecular functions, and biological processes.

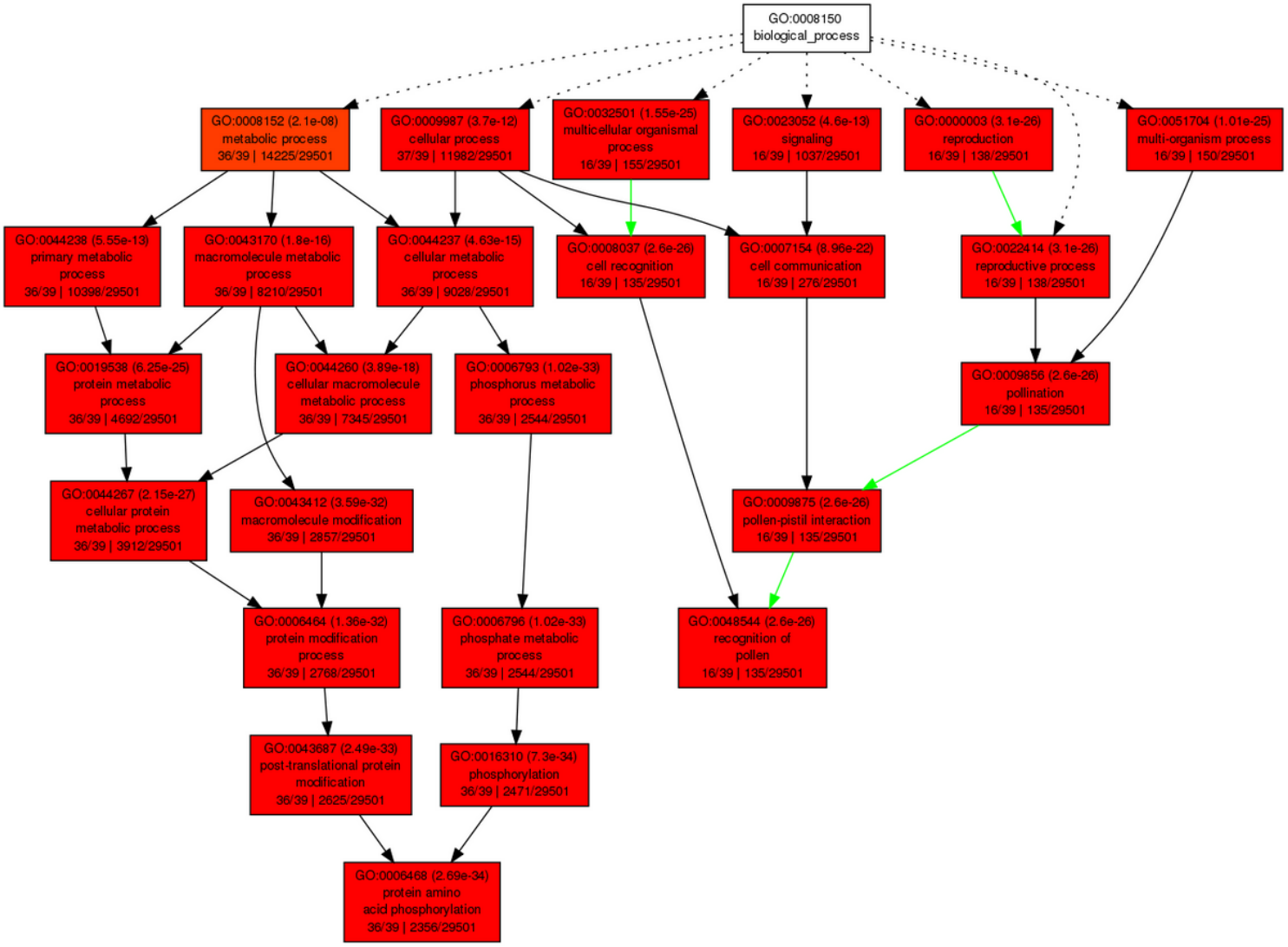

Figure 7

GO Term Enrichment Analysis of 244 Genes using AgriGo Database Corresponding to Biological Function

The figure presents the results of GO term enrichment analysis performed on the 244 genes using the AgriGo database, focusing on their biological functions. The color shading in the illustration ranges from red to yellow, representing the significance levels of the enriched GO terms, with red indicating strong significance and yellow indicating weaker significance.Furthermore, different arrow types are employed to indicate the regulation relationships between the enriched GO terms and the genes. For instance, a green arrow signifies negative regulation, while other arrow types correspond to various regulation types.This analysis provides valuable insights into the functional annotations and regulatory relationships of the studied genes, shedding light on their roles and potential biological implications in the context of the AgriGo database.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementarylegend.docx
- supplementary1.summaryoffeatureimportanceresults.xls
- supplymentary2.BLASTResultsof1023SoybeanSNPsfromPCAAnalysis.xlsx
- Supplementary3.GOanlysisof244genes.xlsx