Legume Genomics and Genetics 2025, Vol.16, No.2, 91-99
http://cropscipublisher.com/index.php/lgg

Feature Review Open Access

Genomic Prediction of Yield and Protein Traits in Soybean Using Machine Learning Models

Xingde Wang, Tianxia Guo
Institute of Life Sciences, Jiyang College, Zhejiang A&F University, Zhuji, 311800, Zhejiang, China
Corresponding email: tianxia.guo@jicat.org

Legume Genomics and Genetics, 2025 Vol.16, No.2 doi: 10.5376/lgg.2025.16.0010
Received: 20 Feb., 2025
Accepted: 06 Apr., 2025
Published: 27 Apr., 2025

Preferred citation for this article:
Wang X.D., and Guo T.X., 2025, Genomic prediction of yield and protein traits in soybean using machine learning models, Legume Genomics and Genetics, 16(2): 91-99 (doi: 10.5376/lgg.2025.16.0010)

Abstract As a globally significant food and plant protein crop, the yield and protein content of soybeans are the core target traits in breeding. However, due to the influence of the interaction between the genetic background and environment of complex quantitative traits, the efficiency of traditional phenotypic selection and genetic improvement is limited. To enhance breeding efficiency and prediction accuracy, this study explored the applicability and effectiveness of multiple machine learning algorithms in the genomic prediction of soybean yield and protein traits. Based on the genotype (SNP) and phenotypic data of multiple soybean breeding populations in this study, machine learning models such as RR-BLUP, Support vector Machine (SVM), Random Forest (RF), Gradient enhancer (GBM), and Deep neural Network (DNN) were respectively constructed. Combined with feature selection methods such as principal Component Analysis (PCA), LASSO and Boruta, the prediction accuracy and stability of the model are systematically evaluated. The results show that nonlinear models (such as RF and GBM) have better generalization ability for complex traits under multiple environmental conditions. The multi-trait joint prediction strategy further enhanced the model's performance in composite indicators such as protein yield. This study demonstrates the potential of machine learning techniques in the genomic prediction of complex quantitative traits, providing an efficient means for auxiliary selection in soybean breeding and laying the foundation for the construction of intelligent and high-throughput breeding decision-making systems.

Keywords Soybeans; Genomic prediction; Machine learning; Yield traits; Protein content

1 Introduction

Soybean (*Glycine max*) has always been the main force in the global food system, especially in providing protein and oil. It is indispensable whether for human consumption or as feed. At present, the demand for plant protein is increasing, and the added value of soybeans has begun to gradually expand in food, industry and even global food security (Van Der Laan et al., 2024). Strong adaptability, high protein content, and the ability to grow in different climates make it the first choice for filling nutritional gaps in the face of climate change and dietary structure changes (Gill et al., 2022).

However, traditional breeding does not always keep up with the pace of such demands. Breeding high-yield and high-protein varieties is no easy task - these traits usually involve multiple genes and often interact with the environment. Just the phenotypic identification stage alone is time-consuming and laborious. Moreover, the long breeding cycle and high cost have further hindered the acceleration point (Ray et al., 2022). For this reason, an increasing number of studies are beginning to shift towards genomic prediction (GP) and machine learning (ML) technologies. With the support of high-throughput phenotypic technologies and genomic data, these new tools can not only enhance the efficiency of selection but also demonstrate stronger predictive power than traditional statistical models in the context of multi-data fusion. Especially for methods like random forests and deep learning, when combined with genomic, phenotypic and environmental variables, the prediction effect is often more stable (Yoosefzadeh-Najafabadi et al., 2021). The emergence of such models can also be regarded as a "paradigm shift" in the field of breeding.

This study explored the application of genomic prediction and machine learning models in predicting soybean yield and protein traits, compared the predictive performance of various genomic and machine learning models,

and evaluated the effect of integrating phenotypic and genomic data for trait prediction. And evaluate its practical significance for soybean breeding projects. This research will promote the modernization and precision of soybean breeding. By enhancing the efficiency and accuracy of trait prediction, machine learning methods are expected to accelerate the breeding process of high-yield and high-protein soybean varieties, thereby contributing to the development of sustainable agriculture and the realization of global food security goals.

2 Genetic Basis of Soybean Yield and Protein Traits

2.1 Major and minor QTLs associated with yield and quality in soybean

The two types of traits, yield and protein, are not determined by a single gene. Behind them, there is usually a group of QTLS working together - some large and some small (Figure 1). Years of mapping and GWAS studies have actually identified many key loci. Some QTLS have obvious effects, and many are micro-effect QTLS with long-term effects (Diers et al., 2018). For instance, major QTLS related to protein content are often located on chromosomes 20 and 15. Those QTLS that regulate yield and quality tend to cluster in the same area. This overlap is not accidental and is likely related to pleiotropy or close linkage (Tayade et al., 2023). The *POWR1* gene is a case in point - it is precisely the "protagonist" in a QTL on chromosome 20, regulating proteins while influencing oil and yield (Goettel et al., 2022). Similar candidate genes are still being discovered, adding many new "parts" to the breeding toolbox (Doszhanova et al., 2024; Dong et al., 2025).

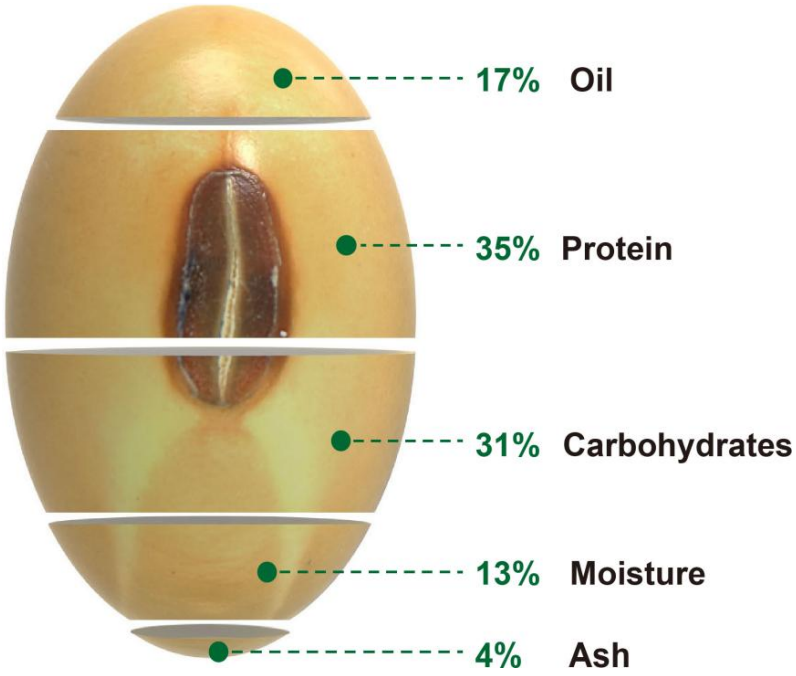

Figure 1 Composition of stored mature soybean seeds. The percentage value indicates the relative weight of the corresponding component in a seed (Liu, 1997) (Adopted from Duan et al., 2023)

2.2 Genotype-phenotype interactions and their impact on complex traits

Not every genotype behaves the same in every environment. Soybean yield and protein traits are precisely the "typical representatives" under the influence of this G×E interaction. Some QTLS are highly active in a certain environment but their performance weakens in a different location. Even the "temperament" of alleles can change when the genetic background is altered (Gao et al., 2024). So, it is not enough to just look at the QTL itself; one also needs to consider which genotype it falls on and what environment it encounters. Vymyslicky et al. (2025) emphasized the significance of introducing G×E analysis into the breeding process. The results of online analysis are also quite interesting. Many genes do not act alone but rather have complementary or superior effects on each other. Some genes even have multiple traits (Fang et al., 2017). These intricate interaction relationships make it more difficult to predict traits and further highlight the necessity of multi-environment tests and higher-order models.

2.3 Genetic control of protein content and amino acid composition

Protein is not only a matter of "how much", but also of "what kind". The control mechanisms of content and amino acid composition are actually quite different. Alleles of varieties like Danbaikong have high-protein QTL on chromosome 20. In some contexts, protein can be increased without significantly reducing yield, although there is often a certain negative correlation between protein and oil and yield (Patil et al., 2017). Studies have found that the genes influencing the total protein content and the mechanisms regulating the amino acid profile may be separate, but the two often run in the same network (Duan et al., 2023; Hu et al., 2023). This is also why the properties of proteins seem particularly "tangled". However, there have been breakthroughs in genomics, transcriptomics and proteomics now. Guo et al. (2022) and Liu et al. (2023) have identified many candidate genes and regulatory modules, laying the foundation for the subsequent breeding of high-protein and nutritionally fortified soybean varieties.

3 Overview of Genomic Prediction and Machine Learning Methods

3.1 Common machine learning algorithms (RR-BLUP, SVM, RF, GBM, DNN, etc.)

When it comes to genomic prediction, the most commonly used models at the beginning were actually some rather "old-fashioned" linear models, such as RR-BLUP or GBLUP. These models are fast in operation and straightforward in logic, especially when traits are mainly controlled by additive effects (Rao et al., 2025). However, when it comes to non-additive effects or complex interactions, these linear methods become a bit challenging. At this time, SVM and SVR, which can handle nonlinearity, come in handy (Wang et al., 2022). Later on, ensemble methods such as Random Forest (RF) and GBM began to gain popularity because they are better at capturing complex relationships. In recent years, deep learning has also joined the fray. Technologies like MLP, CNN, and even DNN have been employed to capture higher-order patterns from a vast array of labels, especially when non-additive structures are quite evident (Montesino-Lopez et al., 2021). Of course, they are not always "invincible". Many times, it still depends on the volume of data and the type of trait.

3.2 Feature selection and dimensionality reduction techniques (PCA, LASSO, Boruta, etc.)

High-dimensional data is much more troublesome to process. An obvious problem is that when there are too many features, the model is prone to deviation. PCA is often used to compress the dimensions first, trying to retain most of the variations while eliminating the redundanities (Monaco et al., 2021; Conard et al., 2023). Regularization methods like LASSO are also commonly used. They automatically compress variables with small weights to 0, thereby filtering out irrelevant features (Lourenco et al., 2022). Packaging methods like Boruta are even more "meticulous". They repeatedly compare the true features with the scrambled "false features" and pick out the most contributing part from them (Lopez et al., 2023). These techniques are particularly crucial for deep learning because such models can easily overfit if they have too many features and insufficient samples.

3.3 Model training, cross-validation, and generalization performance evaluation

Whether a model is well-trained or not cannot be judged merely by the performance of the training set. Cross-validation is a basic skill. Whether it is the K-fold or the retention method, the purpose is to test whether its performance on new data can still be stable (Lopez et al., 2023). Parameter tuning is also indispensable, especially for nonlinear models. For instance, grid search or automatic optimization using AutoML has basically become a standard feature. In terms of evaluation, indicators such as predictive correlation, RMSE or AUC are generally considered (Abdollahi-Arpanahi et al., 2020). Nowadays, many people still use ensemble prediction, combining the results of multiple models. Although it is more complex, the results are usually more stable (Azodi et al., 2019). As for the interpretability of models, there are now ways to open up "black box" models like deep learning. Tools like SHAP can tell you which variable contributes the most in the prediction (Watson, 2021).

4 Model Construction and Prediction Performance Comparison

4.1 Input variable design: SNP encoding and integration of environmental factors

Not all input data can be directly used for prediction. For complex traits such as soybean yield and protein, the variables need to be designed before modeling. SNP information often needs to be encoded first - sometimes

additive, and sometimes dominant or dominant. How to combine them depends on the target and data situation. Environmental information should not be overlooked either. Factors such as the planting location, year, and even management methods, although diverse, make the model more stable when added. Especially when considering the interaction between genotype and environment, these variables can provide additional information support and are very helpful for improving the adaptability of prediction (Norberg et al., 2019).

4.2 Accuracy comparison and error analysis across ML models

Everyone wants to find the "strongest model", but the reality is often that different algorithms perform quite differently when facing different datasets. Integrated methods like Random Forest (RF) and Gradient Elevator (GBM) do a relatively solid job in capturing nonlinearity and complex relationships. Although SVM is not the most complex model, it has the advantages of fast calculation and acceptable accuracy, especially when the amount of data is large (Chakraborty et al., 2020). Deep learning is also often mentioned. For example, the multi-layer perceptron (MLP) can indeed achieve high accuracy, but it also has high requirements for sample size and parameter tuning. If not tuned properly, overfitting is easy (Chandra and Goyal, 2021; Nguyen et al., 2021). Which model is good after all? It is difficult to explain in one sentence. Usually, it is necessary to compare metrics such as R^2, RMSE, and MAE to see the actual error. Incidentally, the performance of the training set and the test set also needs to be compared to determine whether there is overfitting (Robinson et al., 2017; Zhou et al., 2021). When choosing a model in the end, people usually do not rely on just one. Instead, they try several types and then select the one with stable performance through cross-validation.

4.3 Optimization of multi-trait joint prediction strategies

If there is more than one goal, for instance, one wants to predict both yield and protein content at the same time, then a different approach is needed. Multi-task learning frameworks, or model stacking, are commonly used practices at present. They can share genetic and environmental information to improve overall performance. Adding some feature selection methods or hybrid models can usually boost the accuracy and explanatory power (Chakraborty et al., 2020). However, how these strategies are combined still depends on specific requirements, such as whether cross-environment prediction is needed or whether they are related to multiple traits. It cannot be generalized.

5 Multi-Environment Trials and Environmental Interaction Modeling

5.1 Impact of G×E interactions on prediction accuracy

The performance of soybeans in different environments is actually not that easy to predict. Even if the genotypes are the same, changes in climate or planting and management methods may still cause deviations in traits. This so-called interaction between genotype and environment (G×E) often leads people to underestimate or overestimate the true potential of a certain strain (Burgueno et al., 2011; Li et al., 2024). Especially for traits like yield or protein that are influenced by multiple factors, if G×E is not taken into account, the prediction results are often unreliable, and the selected materials are also hard to guarantee stability. Multi-environment tests (MET) and more complex statistical modeling come into play at this point - not all models can capture this interaction, but as long as the models keep up, they can get closer to the real performance.

5.2 Model expansion with environmental covariates (e.g., E-GP, reaction norm models)

The poor performance of many predictive models may not lie in "genes", but rather in the fact that they have overlooked the variable of "environment". Once data such as soil, climate and management measures are incorporated into the model, things become different. New methods such as the reaction norm model, E-GP, and the hybrid model of factor analysis are precisely aimed at the G×E problem. They can make the model both more accurate and easier to interpret (Piepho and Williams, 2024). Moreover, their advantages lie not only in the better processing of current data, but also in their ability to predict genotype expression in untested environments, and even help to develop management plans that are suitable for the environment (Mumford et al., 2023). This type of model provides good tool support for breeding varieties that are resistant to climate fluctuations.

5.3 Case studies on climate adaptability and stable high-protein trait prediction

Although it may only be a few percentage points of improvement, a 7% increase in prediction accuracy can influence the selection direction of a breeding season. Some studies that combined environmental data with genomics and phenotypes did indeed reach this level (Fernandes et al., 2024). Compared with those models that completely ignore environmental changes, response norm or factor analysis methods perform more stably and can identify candidate genotypes for stable production and high protein under variable and even extreme conditions (Burgueno et al., 2011; Li et al., 2024). For breeders, the practicality of these methods does not lie in the three words "accurate prediction", but in their ability to select materials that take into account both adaptability and performance with greater confidence.

6 Case Studies: Application of ML-Based Prediction in Breeding Populations

6.1 Genomic prediction in the U.S. soybean core collection

At first, no one expected machine learning to perform so steadily in such complex materials, but it did perform well in breeding big data scenarios like the US Soybean Core Germplasm Bank. Common models like RF (Random Forest), SVM (Support Vector Machine), and MLP (Multi-Layer Perceptron) have been frequently used in high-throughput phenotypic and genotyping data, capable of predicting agronomic traits such as yield. In particular, some studies have attempted to run RF using hyperspectral reflection data from different environments, and the accuracy rate can reach 84%. If another layer of model combination is added, it can even jump to 0.93 (Figure 2) (Yoosefzadeh-Najafabadi et al., 2021). The greatest significance of this method does not lie in how "smart" the algorithm is, but in enabling breeders to select potential materials from the vast germplasm resource bank earlier and more accurately, and thus the entire breeding process becomes faster.

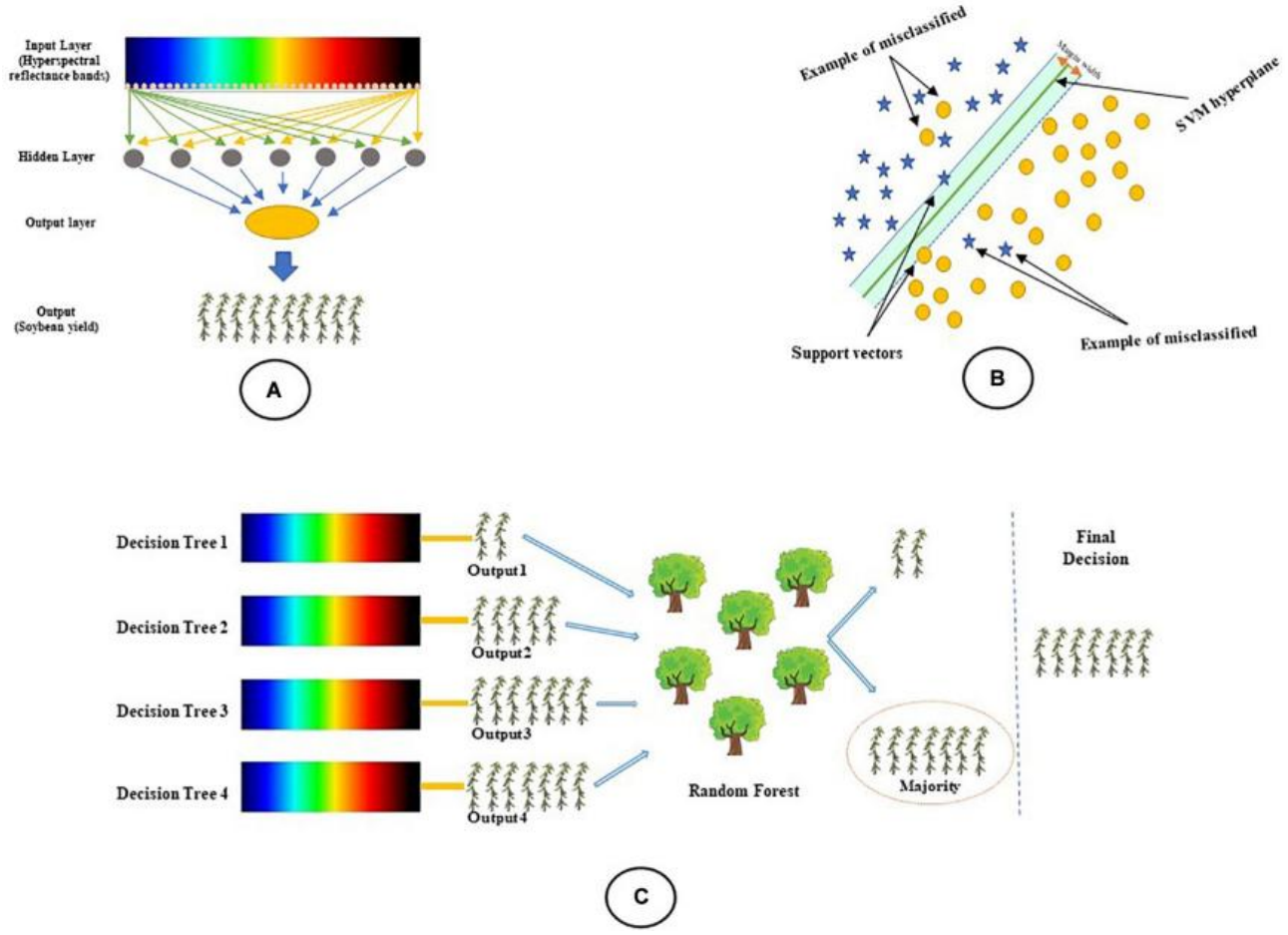

Figure 2 A schematic representation of the machine learning algorithms used in this study to classify the soybean yield using reflectance bands: (A) Multilayer perceptron, (B) Support vector machine, and (C) Random forest (Adopted from Yoosefzadeh-Najafabadi et al., 2021)

6.2 Trait prediction of breeding materials in China's Huang-Huai Region

The situation of breeding work in the Huanghuai region is not so ideal. The climate fluctuates greatly and management methods vary greatly. Relying solely on traditional methods is indeed a struggle. Against this backdrop, machine learning models are highly regarded because they can take into account complex genotype-environmental interactions. Especially when combined with feature selection and integrated models, some customized predictions can better meet the actual needs of the region. Whether it is a variety with strong adaptability or materials with relatively stable traits, the model can screen them out in advance, which has gained a lot of initiative for local breeding (Parmley et al., 2019).

6.3 Prediction evaluation and selection efficiency in commercial high-protein/high-yield varieties

In real commercial scenarios, goal-oriented traits such as high protein and high yield were mainly achieved through linear models in the past, but now machine learning is gradually becoming the main force. It's not that traditional models are ineffective; rather, in situations where trait inheritance is complex and nonlinear interaction is obvious, the advantages of ML begin to emerge. Especially the integrated model or deep learning architecture combined with feature optimization can not only improve the prediction accuracy, but also shorten the computing time (Yoosefzadeh-Najafabadi et al., 2021). From the perspective of breeding enterprises, the efficiency of screening out high-quality varieties has indeed improved. That is to say, it is no longer a luxury to turn potential materials into market products more quickly and accurately.

7 Conclusions and Future Perspectives

Machine learning has indeed performed outstandingly in genomic prediction in recent years, but it is not without flaws. Some regularized regressions, ensemble models, and even deep learning architectures can indeed better handle the complex "genotype-phenotype" relationships brought about by high-dimensional data, especially when the genetic structure of traits is not simple to begin with. However, to be fair, no matter how high the accuracy rate is, the model is still picky about the data. The scale, quality and target traits of the data will all directly affect its performance. Sometimes, when deep models are used, the training time is several times longer, but the result is only slightly better than that of traditional linear models. Moreover, many ML methods are inherently "black box" and cannot explain why predictions can succeed - this is actually quite fatal for breeding decisions (especially in the context of biological research).

By the way, relying on a single data source is indeed not enough. Nowadays, more and more people are beginning to try to incorporate phenotypic, multi-omics (such as genomic, transcriptomic, proteomic) and even remote sensing data, making the models more "diverse" and comprehensive. Deep learning is quite popular in this regard and has a decent ability to handle heterogeneous and multimodal data. Once the model can capture the "signals" from such complex combinations, it will be more confident in predicting the traits that are influenced by both environmental and genetic factors. Moreover, using public data and commercial data together, supplemented by interpretable modeling, sounds like a truly feasible approach for the breeding scenario (rather than just a laboratory demo).

Ultimately, the future development of ML breeding is more like making up for deficiencies in several directions: one is speed, no one wants to train a model for several days; One aspect is interpretability. It's not just about "calculating accurately", but also about "explaining clearly". Another point is how to connect multi-source data. Don't let them be scattered here and there. Automated and interpretive tools like AutoML and iML are already on the way. Their goal is simple - to enable breeders to "get started", rather than being deterred by a bunch of complex models. As data accumulates more and more and species are studied more deeply, these models will eventually have to support the core pillar of smart agriculture, especially when developing soybean varieties that are high-yielding, stress-resistant and have a relatively high protein content.

Acknowledgments

We would like to express our gratitude to the reviewers for their valuable feedback, which helped improve the manuscript.

Conflict of Interest Disclosure

The authors affirm that this research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Abdollahi-Arpanahi R., Gianola D., and Peñagaricano F., 2020, Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes, Genetics, Selection, Evolution, 52: 12.
https://doi.org/10.1186/s12711-020-00531-z

Azodi C., Bolger E., Mccarren A., Roantree M., De Los Campos G., and Shiu S., 2019, Benchmarking parametric and machine learning models for genomic prediction of complex traits, G3: Genes|Genomes|Genetics, 9: 3691-3702.
https://doi.org/10.1534/g3.119.400498

Burgueño J., Crossa J., Cotes J., Vicente F., and Das B., 2011, Prediction assessment of linear mixed models for multienvironment trials, Crop Science, 51: 944-954.
https://doi.org/10.2135/CROPSCI2010.07.0403

Chakraborty D., Elhegazy H., Elzarka H., and Gutierrez L., 2020, A novel construction cost prediction model using hybrid natural and light gradient boosting, Advanced Engineering Informatics, 46: 101201.
https://doi.org/10.1016/j.aei.2020.101201

Chandra R., and Goyal S., 2021, Evaluation of deep learning models for multi-step ahead time series prediction, IEEE Access, 9: 83105-83123.
https://doi.org/10.1109/ACCESS.2021.3085085

Conard A., DenAdel A., and Crawford L., 2023, A spectrum of explainable and interpretable machine learning approaches for genomic studies, Wiley Interdisciplinary Reviews: Computational Statistics, 15(5): e1617.
https://doi.org/10.1002/wics.1617

Diers B., Specht J., Rainey K., Cregan P., Song Q., Ramasubramanian V., Graef G., Nelson R., Schapaugh W., Wang D., Shannon G., McHale L., Kantartzi S., Xavier A., Mian R., Stupar R., Michno J., An Y., Goettel W., Ward R., Fox C., Lipka A., Hyten D., Cary T., and Beavis W., 2018, Genetic architecture of soybean yield and agronomic traits, G3: Genes|Genomes|Genetics, 8: 3367-3375.
https://doi.org/10.1534/g3.118.200332

Dong Q., Cheng Y., Li Y., Tong Y., Liu D., Yu J., Zhao N., Liu B., Ding X., and Xu C., 2025, Genome-wide association study and genomic prediction of essential agronomic traits in diversity panel of soybean varieties, Agronomy, 15(5): 1181.
https://doi.org/10.3390/agronomy15051181

Doszhanova B., Zatybekov A., Didorenko S., Fang C., Abugalieva S., and Turuspekov Y., 2024, Genome-wide association study of seed quality and yield traits in a soybean collection from Southeast Kazakhstan, Agronomy, 14(11): 2746.
https://doi.org/10.3390/agronomy14112746

Duan Z., Li Q., Wang H., He X., and Zhang M., 2023, Genetic regulatory networks of soybean seed size, oil and protein contents, Frontiers in Plant Science, 14: 1160418.
https://doi.org/10.3389/fpls.2023.1160418

Fang C., Ma Y., Wu S., Liu Z., Wang Z., Yang R., Hu G., Zhou Z., Yu H., Zhang M., Pan Y., Zhou G., Ren H., Du W., Yan H., Wang Y., Han D., Shen Y., Liu S., Liu T., Zhang J., Qin H., Yuan J., Yuan X., Kong F., Liu B., Li J., Zhang Z., Wang G., Zhu B., and Tian Z., 2017, Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean, Genome Biology, 18: 161.
https://doi.org/10.1186/s13059-017-1289-9

Fernandes I., Vieira C., Dias K., and Fernandes S., 2024, Using machine learning to combine genetic and environmental data for maize grain yield predictions across multi-environment trials, Theoretical and Applied Genetics, 137: 189.
https://doi.org/10.1007/s00122-024-04687-w

Gao W., Ma R., Li X., Liu J., Jiang A., Tan P., Xiong G., Du C., Zhang J., Zhang X., Fang X., Yi Z., and Zhang J., 2024, Construction of genetic map and QTL mapping for seed size and quality traits in soybean (*Glycine max* L.), International Journal of Molecular Sciences, 25(5): 2857.
https://doi.org/10.3390/ijms25052857

Gill M., Anderson R., Hu H., Bennamoun M., Petereit J., Valliyodan B., Nguyen H., Batley J., Bayer P., and Edwards D., 2022, Machine learning models outperform deep learning models, provide interpretation and facilitate feature selection for soybean trait prediction, BMC Plant Biology, 22: 180.
https://doi.org/10.1186/s12870-022-03559-z

Goettel W., Zhang H., Li Y., Qiao Z., Jiang H., Hou D., Song Q., Pantalone V., Song B., Yu D., and An Y., 2022, *POWR1* is a domestication gene pleiotropically regulating seed quality and yield in soybean, Nature Communications, 13: 3051.
https://doi.org/10.1038/s41467-022-30314-7

Guo B., Sun L., Jiang S., Ren H., Sun R., Wei Z., Hong H., Luan X., Wang J., Wang X., Xu D., Li W., Guo C., and Qiu L., 2022, Soybean genetic resources contributing to sustainable protein production, Theoretical and Applied Genetics, 135: 4095-4121.
https://doi.org/10.1007/s00122-022-04222-9

Hu Y., Liu Y., Wei J., Zhang W., Chen S., and Zhang J., 2023, Regulation of seed traits in soybean, aBIOTECH, 4: 372-385.
https://doi.org/10.1007/s42994-023-00122-8

Li W., Boer M., Joosen R., Zheng C., Percival-Alwyn L., Cockram J., and Van Eeuwijk F., 2024, Modeling QTL-by-environment interactions for multi-parent populations, Frontiers in Plant Science, 15: 1410851.
https://doi.org/10.3389/fpls.2024.1410851

Liu K., 1997, Chemistry and nutritional value of soybean components, Soybeans, 2: 25-113.
https://doi.org/10.1007/978-1-4615-1763-4_2

Liu S., Liu Z., Hou X., and Li X., 2023, Genetic mapping and functional genomics of soybean seed protein, Molecular Breeding, 43: 29.
https://doi.org/10.1007/s11032-023-01373-5

López O., González B., López A., and Crossa J., 2023, Statistical machine-learning methods for genomic prediction using the SKM library, Genes, 14(5): 1003.
https://doi.org/10.3390/genes14051003

Lourenço V., Ogutu J., Rodrigues R., and Piepho H., 2022, Genomic prediction using machine learning: a comparison of the performance of regularized regression, ensemble, instance-based and deep learning methods on synthetic and empirical data, BMC Genomics, 25: 152.
https://doi.org/10.1101/2022.06.09.495423

Monaco A., Pantaleo E., Amoroso N., Lacalamita A., Lo Giudice C., Fonzino A., Fosso B., Picardi E., Tangaro S., Pesole G., and Bellotti R., 2021, A primer on machine learning techniques for genomic applications, Computational and Structural Biotechnology Journal, 19: 4345-4359.
https://doi.org/10.1016/j.csbj.2021.07.021

Montesinos-López O., Montesinos-López A., Pérez-Rodríguez P., Barrón-López J., Martini J., Fajardo-Flores S., Gaytán-Lugo L., Santana-Mancilla P., and Crossa J., 2021, A review of deep learning applications for genomic selection, BMC Genomics, 22: 19.
https://doi.org/10.1186/s12864-020-07319-x

Mumford M., Forknall C., Rodriguez D., Eyre J., and Kelly A., 2023, Incorporating environmental covariates to explore genotype×environment×management (G×E×M) interactions: a one-stage predictive model, Field Crops Research, 304: 109133.
https://doi.org/10.1016/j.fcr.2023.109133

Nguyen H., Vu T., Vo T., and Thai H., 2021, Efficient machine learning models for prediction of concrete strengths, Construction and Building Materials, 266(Part B): 120950.
https://doi.org/10.1016/j.conbuildmat.2020.120950

Norberg A., Abrego N., Blanchet F., Adler F., Anderson B., Anttila J., Araújo M., Dallas T., Dunson D., Elith J., Foster S., Fox R., Franklin J., Godsoe W., Guisan A., O'Hara, B., Hill N., Holt R., Hui F., Husby M., Kålås, J., Lehikoinen A., Luoto M., Mod H., Newell G., Renner I., Roslin T., Soininen J., Thuiller W., Vanhatalo J., Warton D., White M., Zimmermann N., Gravel D., and Ovaskainen O., 2019, A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels, Ecological Monographs, 89(3): e01370.
https://doi.org/10.1002/ECM.1370

Parmley K., Higgins R., Ganapathysubramanian B., Sarkar S., and Singh A., 2019, Machine learning approach for prescriptive plant breeding, Scientific Reports, 9: 17132.
https://doi.org/10.1038/s41598-019-53451-4

Patil G., Mian R., Vuong T., Pantalone V., Song Q., Chen P., Shannon G., Carter T., and Nguyen H., 2017, Molecular mapping and genomics of soybean seed protein: a review and perspective for the future, Theoretical and Applied Genetics, 130: 1975-1991.
https://doi.org/10.1007/s00122-017-2955-8

Piepho H., and Williams E., 2024, Factor-analytic variance-covariance structures for prediction into a target population of environments, Biometrical Journal, 66(6): e202400008.
https://doi.org/10.1002/bimj.202400008

Rao Y., Zhang L., Gao L., Wang S., and Yang L., 2025, ExAutoGP: enhancing genomic prediction stability and interpretability with automated machine learning and SHAP, Animals, 15(8): 1172.
https://doi.org/10.3390/ani15081172

Ray S., Jarquín D., and Howard R., 2022, Comparing artificial-intelligence techniques with state-of-the-art parametric prediction models for predicting soybean traits, The Plant Genome, 16(1): e20263.
https://doi.org/10.1002/tpg2.20263

Robinson R., Palczewska A., Palczewski J., and Kidley N., 2017, Comparison of the predictive performance and interpretability of random forest and linear models on benchmark data sets, Journal of Chemical Information and Modeling, 57(8): 1773-1792.
https://doi.org/10.1021/acs.jcim.6b00753

Tayade R., Imran M., Ghimire A., Khan W., Nabi R., and Kim Y., 2023, Molecular, genetic, and genomic basis of seed size and yield characteristics in soybean, Frontiers in Plant Science, 14: 1195210.
https://doi.org/10.3389/fpls.2023.1195210

Van Der Laan L., Parmley K., Saadati M., Pacin H., Panthulugiri S., Sarkar S., Ganapathysubramanian B., Lorenz A., and Singh A., 2024, Genomic and phenomic prediction for soybean seed yield, protein, and oil, The Plant Genome, 18(1): e70002.
https://doi.org/10.1002/tpg2.70002

Vymyslický T., Trněný O., Rietman H., Balko C., Đorđević V., Ranđelović P., and Dybová M., 2025, Phenotypic characterization of soybean genetic resources at multiple locations: breeding implications for enhancing environmental resilience, yield and protein content, Frontiers in Plant Science, 16: 1422162.
https://doi.org/10.3389/fpls.2025.1422162

Wang K., Abid M., Rasheed A., Crossa J., Hearne S., and Li H., 2022, DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants, Molecular Plant, 16(1): 279-293.
https://doi.org/10.1016/j.molp.2022.11.004

Watson D., 2021, Interpretable machine learning for genomics, Human Genetics, 141: 1499-1513.
https://doi.org/10.1007/s00439-021-02387-9

Yoosefzadeh-Najafabadi M., Earl H., Tulpan D., Sulik J., and Eskandari M., 2021, Application of machine learning algorithms in plant breeding: predicting yield from hyperspectral reflectance in soybean, Frontiers in Plant Science, 11: 624273.
https://doi.org/10.3389/fpls.2020.624273

Zhou Y., Liu Y., Wang D., and Liu X., 2021, Comparison of machine-learning models for predicting short-term building heating load using operational parameters, Energy and Buildings, 253: 111505.
https://doi.org/10.1016/j.enbuild.2021.111505