Integrating Multi-Omics and Phenotypic Datasets for Genomic Breeding Value Prediction

OPEN ACCESS

Volume: 13

Special Issue: 2

Month: January

Year: 2026

E-ISSN: 2582-0397

P-ISSN: 2321-788X

Citation:
Patole, Rajendra Ramesh, et al. "Integrating Multi-Omics and Phenotypic Datasets for Genomic Breeding Value Prediction." *Shanlax International Journal of Arts, Science and Humanities*, vol. 13, no. 2, 2026, pp. 97–103.

DOI:
https://doi.org/10.34293/sijash.v13iS2-i2-Jan.10528

Rajendra Ramesh Patole
Research Scholar
Vidyalankar School of Information Technology
Wadala East, Mumbai, Maharashtra, India

Dr. Abhishek Garg
Associate Professor
Mangalayatan University, Aligarh, Uttar Pradesh, India

Dr. Mandar Sohani
Associate Professor
Vidyalankar Institute of Technology, Wadala East, Mumbai, Maharashtra, India

Abstract
Genomic Breeding Value (GBV) prediction plays a crucial role in modern livestock breeding by enabling early identification of genetically superior animals. Traditional genomic selection approaches primarily rely on single-omics data, particularly single nucleotide polymorphisms (SNPs), which limits their ability to capture the biological complexity underlying economically important traits. This paper proposes an integrated framework that combines multi-omics data, including genomics, transcriptomics, epigenomics, metabolomics, microbiomics, along with phenotypic information to enhance GBV prediction. The proposed approach provides a comprehensive representation of genotype–phenotype relationships, improves predictive accuracy, and increases robustness. By integrating multiple biological layers, the framework supports biologically informed, data-driven breeding decisions and contributes to sustainable precision livestock breeding.

Keywords: Genomic Breeding Value, Multi-Omics Integration, Phenotypic Data, Livestock Breeding, Precision Agriculture

Introduction

The Genomic Breeding Value (GBV) prediction has now become a pillar to the modern livestock breeding programs; it gives the opportunity to find out the genetically superior animals at an early and precise stage. Predicting GBV at an early age of an animal aids in minimizing the generation interval, improving selection efficiency and increasing genetic gain of economically valuable traits in animals (e.g., milk yield, fertility, disease resistance and feed efficiency). Historically, the estimation of GBV has been based mainly on genomic data especially single nucleotide polymorphisms (SNPs) with phenotype records.

Although genomic selection has made a great step towards breeding, numerous multifaceted phenomena are regulated through biological events going beyond genomic variation. Dynamics of gene expression, epigenetic control, metabolic pathways, microbial interaction, and environmental condition collaborate to express the phenotype. Models which rely on one layer of biological data typically can not capture these complex interactions, which restricts prediction and biological understanding.

The recent developments in high-throughput sequencing and omics technologies have led to the production of large-scale biological data, such as transcriptomics, epigenomics, metabolomics, and microbiomics. When these multi-omics datasets are combined with the phenotypic information, it gives a more detailed representation of the genotype phenotype relationship. This integration can enhance the accuracy of prediction of GBV and provide more insight into the biological processes that drive the variation of traits.

The work is aimed at the combination of multi-omics and phenotypic data to build an effective model of predicting genomic breeding value and help to make more informed and sustainable decisions in the field of precision livestock breeding.

Problem Statement

Despite the major improvement in breeding livestock through genomic selection, the predictive power of Genomic Breeding Value (GBV) is restricted due to the use of a mono-omics data (mainly genomic markers) including single nucleotide polymorphisms (SNPs). Multifaceted economic characteristics of livestock are determined by the interaction of multifaceted layers of biology, such as gene expression, epigenetic control, metabolic activity, microbial composition, and the environment. Single-layered models that only take into account one of the biological information levels do not capture these interactions, leading to lower predictive ability and lower biological interpretability. Moreover, the genetic architecture is not the only factor influencing phenotypic variation, and molecular regulation and management conditions also play any role and should be a part of traditional prediction models. The absence of a coordinated method inhibits the use that breeding programs can give to accessible biological data. Thus, there is an urgent necessity in the development of a complete system that will combine multi-omics and phenotypic data to enhance the precision, strength, and feasibility of genomic prediction of breeding value in precision livestock breeding.

Objectives

The main aim of the research is to combine multi-omics and phenotypic data to predict the genomic breeding value (GBV) carefully.

1. To integrate genomic, transcriptomic, epigenomic, metabolomic, microbiomic and phenotypic data to form a single analysis system in estimating GBV.
2. To model the interplay of delicate biological processes between and among various layers of the cellular systems that dictate economically significant livestock characteristics.
3. To enhance the accuracy, robustness, and reliability of GBV prediction in the case of single-omics methods.
4. To enhance breeding choices grounded in evidence and biologically enlightened to enhance livestock accurately.
5. In general, the proposed study will contribute to the field of genomic selection using the integration of holistic biological data.

Literature Review

Multi-omics data combined with sophisticated predictive modelling was found to be a ground-breaking method of enhancing breeding precision in crops and livestock to overcome the shortcomings of classic uni-layer genomic prediction.

Amin, Zaman and Park (2025) proposed an inclusive model, which combines genomics, transcriptomics, proteomics, and phenotypic data to boost climate-resistant crop breeding. It has been noted by their study that the relevance of connecting the molecular-level variation to the field-level performance is of particular importance when the conditions of climate stress are encountered. The authors have shown a high level of predictive modelling and multi-omics signals by incorporating them into the study to achieve better trait

stability and adaptability, and the application of integrative data-driven breeding strategies to global climate change (Amin et al., 2025).

Cembrowska-Lech et al. (2023) suggested a multi-omics system based on artificial intelligence and high-level plant phenotyping in horticulture. They combined phenotypes based on imaging with genomic and metabolomic data, allowing them to characterise traits in more detail, both on the molecular scale and on the morphological scale. It revealed that the presence of AI-mediated heterogeneous information combination significantly enhanced the accuracy of phenotyping and efficiency of decision-making, highlighting the possibilities of an intelligent multi-omics integration in the current breeding pipelines (Cembrowska-Lech et al., 2023).

Cao et al. (2022) considered the use of multi-omics methods in molecular breeding of soybean with emphasis on genomics, transcriptomics, proteomics, and metabolomics. Their study showed that multi-layer biological data can be used to speed up the discovery of traits and functional validation. The authors emphasized that integrated omics solutions are more effective than conventional genomic methods of selection because they describe regulatory and metabolic pathways that contribute to yield and stress tolerance (Cao et al., 2022).

Chao et al. (2024) examined how various omics databases can be integrated to improve the efficiency of crop breeding. Data harmonization and interoperability was one of the key concerns of multi-omics research that their paper has emphasized. The authors showed that the systematic data integration will improve the prediction of traits and breeding decision support by linking the genomic and phenotypic repositories to the integrative bioinformatics platforms (Chao et al., 2024).

Hu et al. (2021) compared the prediction of agronomic and nutritional characteristics of oat using multi-omics in different environments and genetically diverse populations. Their findings indicated that genomics in combination with metabolomics and transcriptomics had very high prediction scores compared to the application of genomic data. The article stressed the importance of the environmental aware multi-omics modelling to facilitate consistent and predictable forecasts across populations of breeders (Hu et al., 2021).

Knoch et al. (2021) applied multi-omics based prediction to predict hybrid performance of canola. They have enhanced predictability of hybrid vigour and yield-related characteristics by using genomic, transcriptomic and metabolic profiles. Their analysis has established that multi-omics integration has the capacity of capturing the effects of heterosis in a more favorable way that cannot be well captured through the single-omics models (Knoch et al., 2021).

Mahmood et al. (2022) provided a comprehensive review of the multi-omics revolution in plant breeding, saying that the latter has been applied in enhancing the efficiency of breeding. The authors described the ways in which an integrated strategy of omics could be applied in the disaggregation of traits, accelerated selection cycles, and precision farming. Their evaluation legitimized the assumption that multi-omics combination is required in the case of intricate trait make-up of breeding schemes (Mahmood et al., 2022).

Montesinos-Lopez et al. (2024) reviewed multimodal deep learning methods that can be used to predict in plant breeding in a genomic-enabled way. The article demonstrated the significance of deep learning frameworks in successful combination of genomic, phenotypic and environmental data. They demonstrated in their study that multimodal learning models work better compared to conventional statistical such that they could model nonlinear interactions among omics layers (Montesinos-Lopez et al., 2024).

Tahir et al. (2022) investigated the efficiency of the multi-omics information as the predictor of the fertility performance of heifers in the context of genomic forecasting. The study with the combination of genomics, transcriptomics and metabolomics showed improved prediction of fertility-related traits, which are otherwise difficult to model. The authors emphasized the significance of multi-omics integration in complex reproductive features in livestock breeding (Tahir et al., 2022).

Wang et al. (2023) have designed a multi-omics-based deep neural network-based genomics prediction model, which is called DNNGP and incorporates the application of multi-omics in plants. Their model was

found to perform better as compared to the traditional genomic prediction models particularly when it comes to the characteristics that are subjected to complicated regulation mechanisms. The study has validated that deep learning has been handy in multi-omics integration of data in the breeding (Wang et al., 2023).

Multi-view BLUP is a statistical model proposed by Wu and colleagues (2025), which is applicable in post-omics integrative prediction. They also used their approach on the current BLUP models and they involved several biological data perspectives to generate a more improved prediction whose results are interpretable. This article provided the answer between the conventional breeding models and the new multi-omics integration methods (Wu et al., 2025).

Wu, Luo and Xiao (2024) were able to show that using multi-omics data in conjunction with machine learning methods enhanced genomic prediction of maize yield. Their analysis proved that a combination of transcriptomic and metabolomic characteristics and genomic markers was more effective in predicting yield, especially in fluctuating environmental settings (Wu et al., 2024).

Wang (2024) conducted a review of integrative omics technologies to enhance livestock breeding. It was highlighted in the study that genomics should be integrated with functional omics layers to maximize genetic assessment and breeding effectiveness. The author emphasized the idea that the multi-omics integration will help in the sustainable production of livestock through making more accurate and bio-informed decisions on selections (Wang, 2024).

Wu et al. (2022) explored the use of multi-omic data in the breeding of barley and reported that this approach demonstrated a substantial increase in prediction capability. Through their work, they found the combined use of metabolomic and transcriptomic data with genomics enhanced the robustness in predicting traits, especially traits related to stress (Wu et al., 2022).

Yang et al. (2021) summarized the use of multi-omics technologies in crop improvement, describing their application in the discovery of traits, functional genomics and predictive breeding. The authors pointed out that integration in multi-omics can conduct a system-level appreciation of plant biology, which is vital in enhancing complicated agronomic traits (Yang et al., 2021).

In general, the analyzed literature sources demonstrate a coherent conclusion that the combination of multi-omics and phenotypic data is far more predictive and biologically interpretable than the use of the single-omics strategies. These results clearly support the argument of integrated data structures in genomic breeding values prediction in precise agricultural practices and animal enhancement.

Methodology

The approach that will be used in this research is the integration of multi-omics and phenotypic data to enable a correct prediction of Genomic Breeding Value (GBV). The workflow will include dataset preparation, data preprocessing, feature engineering, model development and evaluation. This is a systematic format, which provides proper representation of intricate biological data, and allows good predictive modelling.

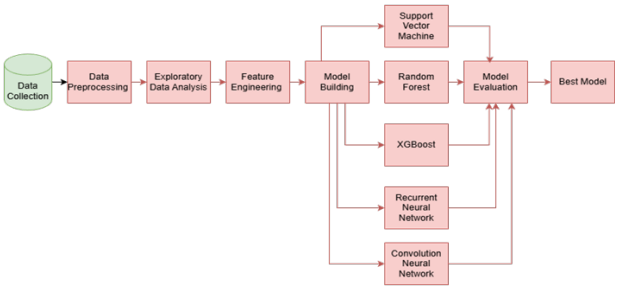

Figure 1 Methodology Diagram

Dataset Description

The dataset is also 29 input features and one output variable, Genomic Breeding Value (GBV) of various dairy farming organizations. The records are the IDs of individual cows denoted by animal_id. The input characteristics consist of the phenotypic characteristics which are milk yield, body weight, fat percentage, feed ratio, and disease resistance score. Moreover, there are genomic data in the form of single nucleotide polymorphisms (SNPs), transcriptomic gene expressions, epigenomic factors like DNA methylation scores, metabolomic features and rumen microbiota profiles. Such variables as ambient temperature and the type of nutrition are also considered as environmental and management-related. Collectively, those characteristics compose a multi-omics data set that includes genomics, transcriptomics, epigenomics, proteomics, and microbiomics, which allows analyzing the combination of factors that affect GBV.

Data Preprocessing

Preprocessing of data was necessary in order to have uniformity and reliability among the non similar data sources. The continuous variables which included milk yield, body weight and calving interval had missing values that were imputed by using mean or median values depending on the distribution of the data. In the categorical variables; breed, sex and nutrition type, missing data was substituted with the mode.

Minimum maximum normalization was done to all numerical features (continuous) to eliminate scale differences and convergence of the models, putting all data in the [0, 1] interval. Such variables as milk_yield, fat percentage, feed ratio, and disease resistance score were normalized respectively. One-hot encoding was used to convert categorical traits (breed, sex, farm location, nutrition type and management system) so that there was no imposition of ordinal correlation between them.

Since multi-omics features (especially, genotype_snp_array and miRNA_expression_profile) are dimensional, the dimensionality reduction was conducted in the form of Principal Component Analysis (PCA). PCA decreased the feature space even though it maintains over 95% of the initial variance which effectively mitigates the curse of dimensionality and increases the computational efficiency.

Feature Engineering

The feature engineering was done to improve the predictive performance and to include relationships of biological significance. Features of domain-driven interaction like the fat-to-protein ratio, which was an indicator of the milk quality, were developed. The production values were normalized with the environmental conditions like temperature and humidity index to generate adjusted milk yield which enabled better representation of environmental stress effects.

Omics aggregation was used to enhance interpretability in complex layers of omics, like gene expression level, metabolomic profile and rumen microbiota profile. Summaries at the pathway level were derived based on the Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) databases. This transformation reduced the high-dimensional molecular signals to biologically significant features that could be used to predict a model.

Model Development

The models used in the machine learning were the Random Forest (RF), the Support Vector Machine (SVM) with a Radial Basis Function kernel, and XGBoost. RF was chosen due to its strength and capacity to address nonlinear interactions and the main hyperparameters were optimized by grid search. Complex nonlinear relationships were modeled using SVM and regularization and kernel parameters were optimized using cross-validation. XGBoost was also used due to its high predictive capability and capacity to regulate, and the optimization of its parameters was done through the Bayesian optimization.

The deep learning ones were Deep Neural Network (DNN), Convolutional Neural Network (CNN), and Recurrent Neural Network (RNN). The DNN model included an input layer, three hidden layers with ReLU

the activation, and a linear output layer, which was trained with Adam optimizer and Mean Squared Error loss. Using one-dimensional convolution and one-dimensional pooling layers, CNNs were put on high-dimensional omics data in the form of matrices, e.g. reshaped SNP arrays. A time-sensitive RNN based on LSTM was used to capture long-term behavior by modeling the sequential relationship among time-sensitive omics data.

Model Evaluation

An 80:20 split was used to separate the dataset into training and testing and stratified sampling using breed was used to maintain population structure. The training was done with five-fold cross-validation to enhance robustness and minimize overfitting. Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R2) were used to test model performance as they are a complete measure of its accuracy, error level, and variance explained.

Software and Tools

All experiments were conducted using Python 3.10.

- Scikit-learn for preprocessing, PCA, Random Forest, and SVM
- XGBoost for gradient boosting models
- TensorFlow/Keras and PyTorch for DNN, CNN, and RNN implementation

Expected Outcome

The suggested multi-omics and phenotypic integration model would likely enhance the prediction of genomic breeding value (GBV) in livestock, which is based on full biological information.

- Stronger predictive power and precision of GBV than conventional single-omics techniques, in the form of the combination of complementary phenotypic and molecular data.
- Enhanced imaging of multilayered genotype phenotype interactions, such as genomics, transcriptomics, epigenomics, metabolomics, microbiomics, and phenotypic phenotype.
- Improved biological interpretability via further understanding of the relative impact and effect of various omics characteristics in estimating GBV.
- Support for data-driven and precise breeding decisions, enabling breeders to identify superior animals more reliably and promote sustainable livestock improvement strategies.

Conclusion

This paper has shown the significance of using the combination of multi-omics and phenotypic data to effectively predict Genomic Breeding Value (GBV) in livestock breeding programs. Conventional uni-omics strategies have weaknesses in terms of their capacity to capture the biological complexity of economically significant traits. This study offers a more holistic model of genotype epigenotype and phenotype linkages by treating data integration as a holistic data model. The suggested methodology will allow achieving better predictive performance, greater robustness, and biological interpretability, which will inform and make reliable breeding decisions. Combinations of various molecular layers and phenotypic reports are in line with the ideas of precision breeding and livestock sustainability. On the whole, this piece of work leads to the development of the genomic selection practices, as it highlights the importance of the multi-omics integration and provides a framework that can be utilized by the future studies and be used in the context of the contemporary livestock breeding systems.

References

1. Amin, A., Zaman, W., & Park, S. (2025). Genes, 16, 809.
2. Cembrowska-Lech, D., Krzeminska, A., Miller, T., Nowakowska, A., Adamski, C., Radaczynska, M.,

Mikiciuk, G., & Mikiciuk, M. (2023). Biology, 12, 1298.
3. Cao, P., Zhao, Y., Wu, F., Xin, D., Liu, C., Wu, X., Lv, J., Chen, Q., & Qi, Z. (2022). International Journal of Molecular Sciences, 23, 4994.
4. Chao, H., Zhang, S., Hu, Y., Ni, Q., Xin, S., Zhao, L., Ivanisenko, V. A., Orlov, Y. L., & Chen, M. (2024). Journal of Integrative Bioinformatics, 20, 20230012.
5. Hu, H., Campbell, M. T., Yeats, T. H., Zheng, X., Runcie, D. E., Covarrubias-Pazaran, G., Broeckling, C., Yao, L., Caffe-Treml, M., Gutierrez, L., & Smith, K. P. (2021). Theoretical and Applied Genetics, 134, 4043–4054.
6. Knoch, D., Werner, C. R., Meyer, R. C., Riewe, D., Abbadi, A., Lucke, S., Snowdon, R. J., & Altmann, T. (2021). Theoretical and Applied Genetics, 134, 1147–1165.
7. Mahmood, U., Li, X., Fan, Y., Chang, W., Niu, Y., Li, J., Qu, C., & Lu, K. (2022). Frontiers in Plant Science, 13, 1062952.
8. Montesinos-Lopez, O. A., Chavira-Flores, M., Kismiantini, Crespo-Herrera, L., Saint Pierre, C., Li, H., Fritsche-Neto, R., Al-Nowibet, K., Montesinos-Lopez, A., & Crossa, J. (2024). Genetics, 228, iyae161.
9. Tahir, M. S., Porto-Neto, L. R., Reverter-Gomez, T., Olasege, B. S., Sajid, M. R., Wockner, K. B., Tan, A. W., & Fortes, M. R. (2022). Journal of Animal Science, 100, skac340.
10. Wang, K., Abid, M. A., Rasheed, A., Crossa, J., Hearne, S., & Li, H. (2023). Molecular Plant, 16, 279–293.
11. Wu, B., Xiong, H., Zhuo, L., Xiao, Y., Yan, J., & Yang, W. (2025). Journal of Genetics and Genomics, 52, 839–847.
12. Wu, C., Luo, J., & Xiao, Y. (2024). Molecular Breeding, 44, 14.
13. Wang, H. (2024). Animal Molecular Breeding, 14.
14. Wu, P. Y., Stich, B., Weisweiler, M., Shrestha, A., Erban, A., Westhoff, P., & Inghelandt, D. V. (2022). BMC Genomics, 23, 200.
15. Yang, Y., Saand, M. A., Huang, L., Abdelaal, W. B., Zhang, J., Wu, Y., Li, J., Sirohi, M. H., & Wang, F. (2021). Frontiers in Plant Science, 12, 563953.