Vourlaki, I.-T., Luigi-Sierra, M., Ayres, L., Ramayo-Caldas, Y., Calus, M., Karlskov-Mortensen, P., Piles, M. (2026) Integrating multi-omics in deep learning prediction of pig production traits, *Proceedings 13th World Congress on Genetics Applied to Livestock Production* 23672: https://doi.org/10.31274/wcgalp.23672 ©2026 Vourlaki, et al. CC BY-NC-ND

Integrating multi-omics in deep learning prediction of pig production traits

I.-T. Vourlaki[1], M.G. Luigi-Sierra[2], L. Ayres[3], Y. Ramayo-Caldas[1], M.P.L. Calus[3], P. Karlskov-Mortensen[2], and M. Piles[1]

[1]Animal Breeding and Genetics Program, IRTA, Caldes de Montbui, Spain; [2]Department of Veterinary and Animal Sciences, University of Copenhagen, Copenhagen, Denmark; [3]Animal Breeding and Genomics, Wageningen University & Research, the Netherlands;
[*]Corresponding author: miriam.piles@irta.es

Introduction

Genomic selection has transformed animal breeding by using dense genome-wide markers to estimate breeding values and accelerate genetic gain. By prioritising breeding value estimation over total phenotypic prediction traditional linear models (e.g., Meuwissen *et al*., 2021) mainly focus on additive effects, with non-additive components such as epistasis and dominance often not explicitly modelled. Deep learning (DL) can model these higher-order interactions in high-dimensional datasets without strict parametric assumptions (Montesinos-López *et al*., 2021). Advances in multi-omics technologies such as transcriptomics and epigenomics provide complementary regulatory insights, and integrating these layers with genotypes can improve trait prediction (Jové-Juncà *et al*., 2025; Calle-García *et al*., 2023). However, challenges like high dimensionality, heterogeneity, and noise limit predictive power. This study investigates and proposes multi-omics integration strategies within a deep-learning framework, combining SNP genotypes, gene expression profiles, and DNA methylation data to predict three key pig production traits: average daily gain (ADG), backfat thickness (BFT), and residual feed intake (RFI). The primary objective was to evaluate the efficacy of dimensionality-reduction and feature-selection methods, such as partial least squares (PLS), and multi-omics factor analysis (MOFA+, Argelaguet *et al*., 2020).

Materials & Methods

The study utilised data from 443 pigs representing three commercial breeds: Yorkshire (YY), Duroc (DD), and Landrace (LL). Three traits were analysed: ADG, BFT, and RFI, each adjusted for sex, parity, and housing. The molecular data comprised 11,900 biallelic SNPs common across breeds, transcriptomic data for 14,293 genes generated by bulk RNA-seq, and DNA methylation data obtained via reduced-representation bisulfite sequencing (RRBS), encompassing 1,095,309 CpG sites. Both transcriptomic and methylation data were derived from the *longissimus dorsi* muscle. An intermediate integration was deployed, where each omics dataset is processed through a dedicated Multilayer Perceptron (MLP) subnetwork with a bottleneck layer, then the resulting latent representations are merged into a hidden layer, which is followed by an additional hidden layer that halves the number of nodes from the previous layer. The output layer generates predictions of the phenotype. L2 regularization, dropout, batch normalization, and early stopping were implemented to further mitigate overfitting and underfitting. Predictions and loss were computed via forward propagation, followed by backpropagation to update parameters and minimize validation loss. To manage the high dimensionality of the data, several Feature Selection (FS) strategies were implemented. These included variance-based filtering to define subsets of CpG sites, and supervised selection via PLS Variable Importance in Projection (VIP) scores for all omics. Additionally, MOFA+ was employed as an unsupervised probabilistic framework to identify latent factors capturing variability across the heterogeneous layers. For FS based on

MOFA and PLS, the inputs comprised 11,900 SNPs, 14,293 transcripts, and 14,492 CpG sites, with the latter filtered using variance thresholds of 0.01 and 0.05. Model performance was assessed through two scenarios: a Leave-One-Breed-Out (LOBO) validation to test model transferability and a stratified 5-fold cross-validation considering the combined data across the three breeds to evaluate generalisation within a mixed population. FS using PLS and MOFA was restricted to the training set to ensure unbiased results. Each model, and consequently hyperparameter optimisation was conducted separately for each trait using the Keras Tuner 1.4.7 with Bayesian Optimisation to identify the most efficient configurations for activation functions, learning rates, and dropout. Moreover, each model was trained in triplicate using distinct seeds to account for initialisation variability. Prediction accuracy was assessed using the Pearson correlation coefficient between observed and predicted values.

Results

Comparing models trained on all features (AF) against those using variance-filtered (VF) CpG sites revealed that reducing the dimension of the epigenomic dataset did not compromise predictive accuracy. For instance, using only 1.95% of the original CpG sites (VF0.05 threshold) yielded results comparable to those from the full dataset containing over one million sites (Fig. 1). Unsupervised analysis via MOFA+ identified six latent factors capturing substantial variance across layers (e.g., 29.2% in SNPs, 45.9% in transcripts and 42.9% in CpGs).

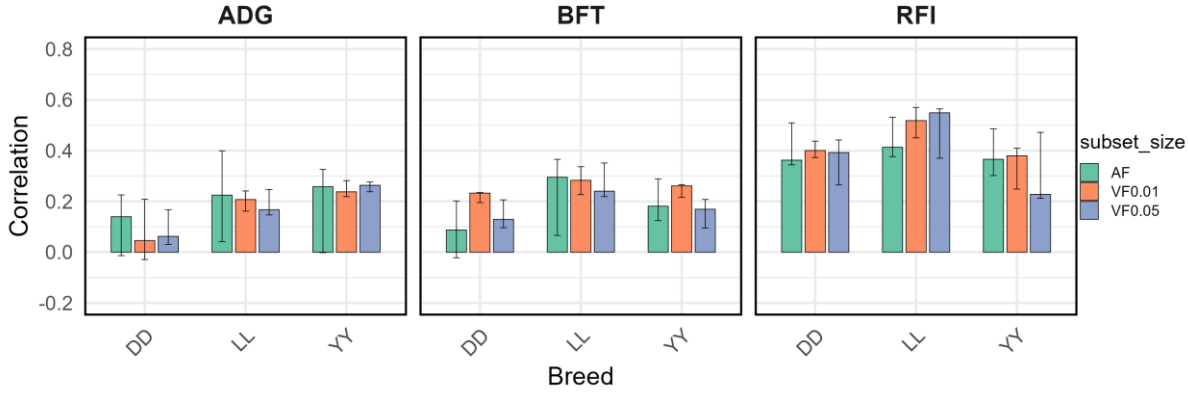

Figure 1. Prediction performance in the LOBO scenario for three input variants: AF (all features), VF0.01 (variance-filtered CpGs at 0.01; 23.1% retained), and VF0.05 (variance-filtered CpGs at 0.05; 1.95% retained)

Feature selection based on MOFA+ Factor 1 improved prediction ability compared to earlier unsupervised methods, with RFI reaching a median correlation of 0.63 in the LL breed when 5,000 features were used (Fig. 2). However, PLS emerged as the superior feature-selection approach, providing the highest correlations across all breeds and traits. In the LOBO analysis, PLS-selected features achieved the highest median correlations for RFI (0.71 in DD, 3,000 features), BFT (0.51 in LL, 1,000 features), and ADG (0.46 in LL, 1,000 features), with similar performance observed using fewer features (RFI 0.68, 1,000 features; BFT 0.50, 200 features; ADG 0.42, 200 features) (Fig. 3). Five-fold cross-validation further confirmed the robustness of PLS-based models (Fig. 4). In particular, for RFI, selecting 200 features yielded the highest median correlation (0.72). BFT was best predicted using 200 features (0.51), and for ADG, selecting 100 features also resulted in the highest median prediction accuracy (0.48).

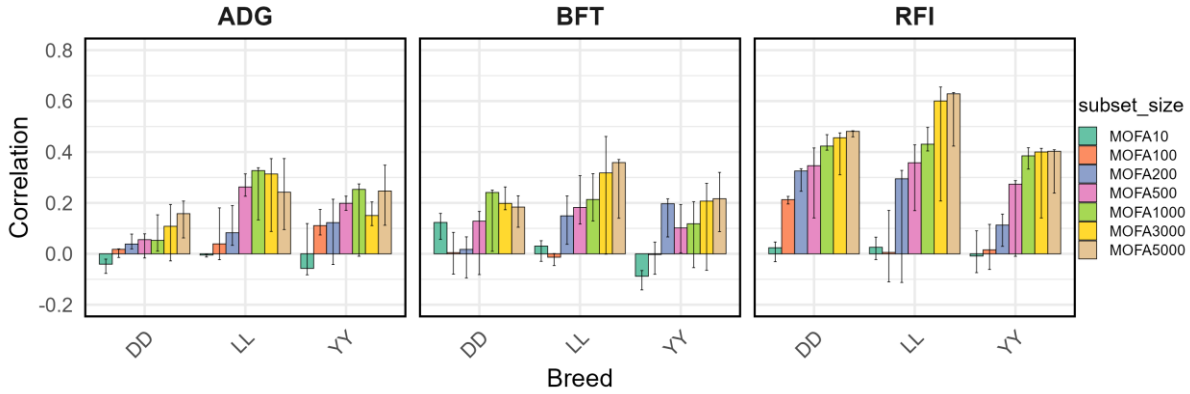

Figure 2. Prediction performance using MOFA in the LOBO prediction scenario for each omics layer. Bars are color-coded to indicate the size of the selected feature subsets.

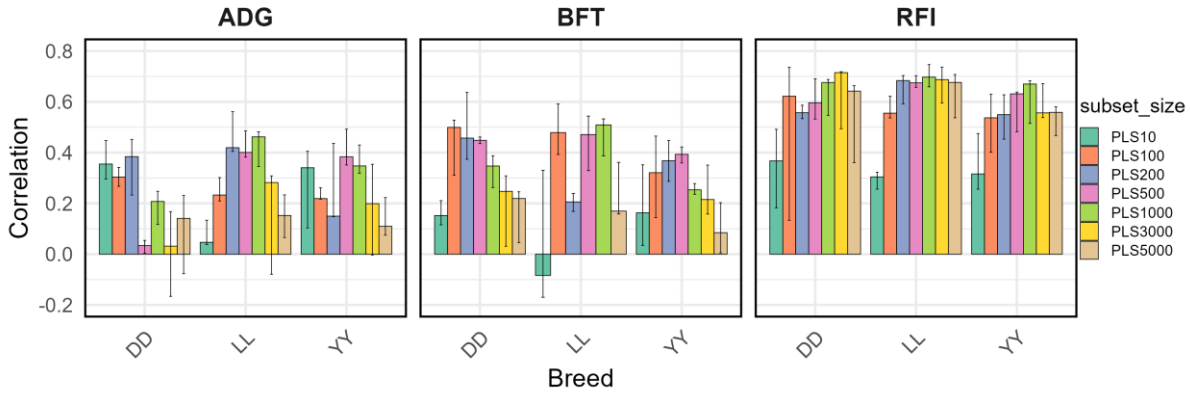

Figure 3. Prediction performance using PLS in the LOBO prediction scenario for each omics layer. Bars are color-coded to indicate the size of the selected feature subsets.

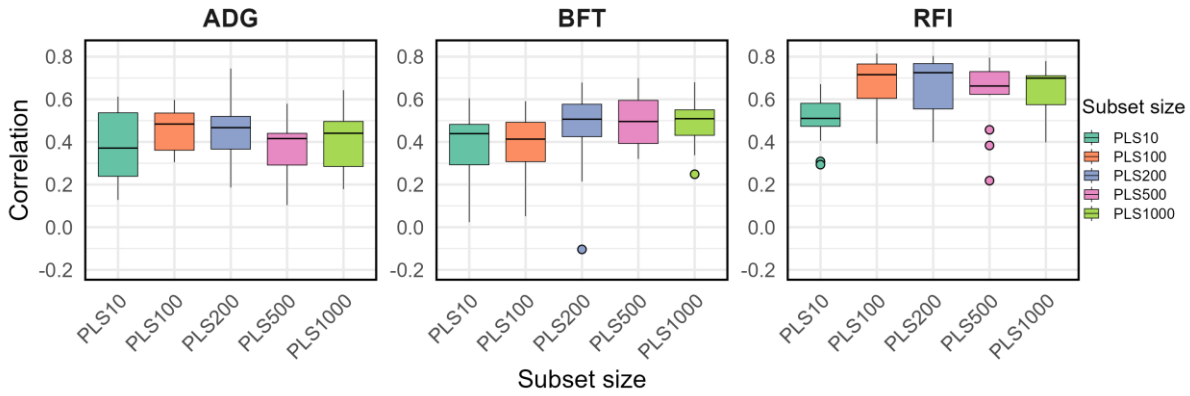

Figure 4. Prediction performance for 5-fold cross-validation for the combined dataset using PLS-selected features for each omic, with colour-coded bars indicating subset sizes. PLS inputs included 11,900 SNPs, 14,293 transcripts, and 14,492 CpGs.

Discussion

Retaining high-variance CpG sites using univariate variance-based filtering resulted in equal or improved prediction accuracy compared to using 1M sites. Removing low-variance sites reduces noise without compromising model performance. Moreover, multivariate filtering methods like PLS and MOFA further increased prediction accuracy, as they account for correlations and interactions among predictors. As a supervised multivariate method, PLS

successfully identifies latent components that both account for feature correlations and maximise the covariance between molecular profiles and target phenotypes, resulting in the selection of more informative subsets. Notably, small omics feature subsets (100–200 features) frequently matched or outperformed models using thousands of predictors. This suggests that reducing redundancy and noise by selecting the most informative features improves predictive performance while mitigating overfitting and the curse of dimensionality. Model performance was consistent across LOBO and five-fold cross-validation schemes, indicating that the integrated multi-omics signals capture shared biological relationships and maintain predictive accuracy across breeds and genetic backgrounds. RFI exhibited particularly strong predictive performance, with the highest median correlation reaching 0.72, likely because RFI integrates multiple feed-efficiency processes, such as metabolism and energy regulation, that are well captured by genomic, transcriptomic, and epigenomic data. In comparison, previous studies based on SNP-only feature selection combined with machine-learning methods, such as support vector machines and gradient boosting, reported substantially lower prediction accuracies for phenotypic RFI (0.27–0.28; Piles *et al*., 2021). The high performance observed here, achieved through intermediate multi-omics integration and feature selection, underscores the added value of incorporating transcriptomic and epigenomic information into breeding programmes. This approach is particularly valuable for complex traits that are difficult or costly to measure, such as RFI, and can serve as a complementary tool to support phenotyping and decision-making particularly as the acquisition of targeted omics subsets becomes increasingly economically viable. Although deep learning is computationally demanding, PLS-based feature selection reduces dimensionality, accelerates training, and improves scalability.

Conclusions

Integrating multiple omics layers through a deep learning framework with intermediate integration and PLS-based feature selection enables accurate prediction of complex pig traits using small, carefully selected subsets of SNP, gene expression, and CpG methylation features. This approach achieves high predictive performance without relying on the full omics set, highlighting the value of data-driven feature selection. It is generalizable across traits and breeds, stable across initializations, and captures trait-informative signals. By improving prediction of challenging traits such as RFI, it provides practical support for breeding programs aiming to enhance productivity and efficiency.

Acknowledgements

This study was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101000236 (GEroNIMO). We gratefully acknowledge LF, Breeding & Genetics, Danish Agriculture & Food Council F.m.b.A., for providing the data used in this study.

References

Argelaguet R., Arnol D., Bredikhin D., *et al*. (2020) Genome Biol. 21:111. https://doi:10.1186/s13059-020-02015-1

Calle-García J., Ramayo-Caldas Y., Zingaretti L.M., *et al*. (2023) Genet. Sel. Evol. 55:29. https://doi:10.1186/s12711-023-00803-4

Jové-Juncà T., Haas V.P., Calus M.P.L., *et al*. (2025) Animal 20:101742. https://doi:10.1016/j.animal.2025.101742

Meuwissen T.H.E., Hayes B.J., and Goddard M.E. (2001) Genetics 157:1819–1829. https://doi.org/10.1093/genetics/157.4.1819

Montesinos-López O.A., Montesinos-López A., Pérez-Rodríguez P., *et al*. (2021) BMC Genomics 22:19. https://doi.org/10.1186/s12864-020-07316-2

Piles M., Bergsma R., Gianola D., *et al*. (2021) Front. Genet. 12:611506. https://doi:10.3389/fgene.2021.611506