AL-BUTANA JOURNAL Of APPLIED SCIENCE
http// : ojs.abutana.edu.sd , ISSN: 1858-6616
Issue (17), December 2025, Page NO. 1-18

BioHybridNet: An Interpretable Hybrid Deep Learning Model for Genomic Prediction Based on Biological Pathways

Mohammed Babiker Ali Mohammed [1]
Abuzer Hussein Ibrahim Ahmed[2]
Sally D. Abugasim[3],
Zeinab E. Ahmed[4],
[1,2] Department of Computer Science, University of AL-BUTANA, Sudan
akoody@albutana.edu.sd , abouzer.hussein@gmail.com
[3,4] Department of Computer Engineering, University of Gezira, Sudan
_ , sally.dallah1@gmail.comzeinab.e.ahmed@gmail.com

Abstract

The feasibility of deep learning in genomic prediction is hindered by the lack of biological insight and knowledge of the models, and the traditional linear models are incapable of finding all possible genetic interactions. To this end, we designed BioHybridNet, a hybrid system consisting of biological pathways and GWAS-directed attention incorporated into an interpretable model. We have a dynamic gating model that mixes linear and non-linear prediction and is able to perform federated learning in a privacy-preserving manner. On the UK Biobank and wheat genomic data, BioHybridNet had a mean R 2 increase of +14.5 percent over linear models and +8.2 percent over deep learning models, and recovered 92 percent of known disease loci a 34 percent increase in interpretability and offered novel information such as quantifying the epistatic nature of Schizophrenia. This research has been able to balance the accuracy and interpretability of genomic prediction, and future studies aim at the integration of multi-omics and clinical translation.

Keywords: ***hybrid machine learning, interpretable AI, genomic prediction, pathway regularization, polygenic risk scores***

Introduction

With the emergence of large-count biobanks, including the UK Biobank that encompasses 500,000 whole genomes (Bycroft et al., 2018), the necessity

to develop more sophisticated computation methods that can transform the genomic data into the clinical as well as agricultural actionable knowledge has been growing. Although genome-wide association studies (GWAS) have found many thousands of locus which are associated with traits (Visscher et al., 2017), the use of genome-wide association studies has two enduring issues: (1) the missing heritability problem, in which large fractions of genetic variation remain unaccounted by linear models (Yang et al., 2015), and (2) the accuracy-interpretability trade-off of machine learning models (Libbrecht and Noble, 2015).

Deep learning (DL) models have already demonstrated potential in non-linear genetic interaction, as they are the state-of-the-art in polygenic risk prediction (Brandes et al., 2023). But their mysterious quality is a significant obstacle to clinical adoption, as doctors need to establish biological proof to support their decision to treat them (Holzinger et al., 2022). On the other hand, GEMMA and classical linear mixed models (LMMs) such as GEMMA (Zhou and Stephens, 2012) do not scale to the same level as the model able to interpret the variance components, yet they do not provide the ability to model the underlying complex traits with epistasis and gene-environment interactions.

In spite of development of genomic technologies, there are critical limitations in making correct but interpretable predictions of complex traits. Although the non-linear genetic interactions are well represented in deep learning models (Brandes et al., 2023), it is highly counterproductive to clinical applications: more than 78 percent of physicians do not accept AI predictions that do not have biological rationale (Holzinger et al., 2022). On the other hand, classical linear mixed models (LMMs) are interpretable in variance terms and do not capture epistasis and gene-environment interaction, which could explain the [?]30 of heritability of most phenotypes (Yang et al.,

2015). This is an accuracy-interpretability tradeoff which is further worsened by the computational inefficiency (Avsec et al., 2021) and small-sample bias in underrepresented populations (Martin et al., 2022) and thus there exists an immediate need to develop hybrid methods that address these competing requirements.

To overcome those challenges, we seek to accomplish the following three objectives:

Train biologically constrained neural architecture, which incorporates GWAS priors via attention mechanisms (Visscher et al., 2017) and pathway-informed regularization (Jumper et al., 2021), to achieve 85% of known locus trait associations -36 percentage better than the state of the art deep learning (Zhou, 2023).

Make use of the adaptive LMM-neural hybridization to optimize computational efficiency and achieve a goal of 2.5x faster training on average than monolithic DNNs and sub-linear scaling with sample size (n > 500K).

Address the issue of insufficient data with generative adversarial networks trained on evolutionary conserved regions (Yelmen et al., 2021), and one can generate robust predictions when n is less than 5,000 samples.

Our methodologically takes the idea of biological constraints one step further by training models with them instead of using post-hoc filters (Lundberg & Lee, 2017), and proves to work cross-kingdom across human biology and agricultural breeding initiatives.

1.Related works

Recent developments in genomic prediction have been struggling with a fundamental tension; the trade-off between the quality of prediction and the ability to understand biologically. Although deep learning models provide the best results (Brandes et al., 2023), they cannot be used by clinics and agriculture due to their black-box nature (Holzinger et al., 2022). In contrast, classical statistical genetics techniques focus on interpretability but do not describe complicated genetic architecture (Wray et al., 2021). Crossbreeding techniques have already become an enticing solution, and the existing applications, either as post-hoc interpretation schemes (Lundberg & Lee, 2017) or as computationally intensive ensembles (Zhou, 2023), do not appear to find an effective balance between these needs. Five major paradigms of methodology in the last three years have been critically analyzed in this section, with a reference to the strengths, weaknesses and the way our work builds upon them through biologically based architectural innovations.

In 2020, the first large-scale comparison of genomic prediction in various populations was proposed by (Martin et al, 2020). Their experiment revealed that there are severe performance decreases (>30% R2 decline) in underrepresented groups with standard LMMs, which is why we emphasize adaptive hybridization. Yet, they only used the data aggregation but not architectural enhancements as a solution.

In 2021 the Authors were the first to use transformers to analyze the genomic sequence (Avsec et al. 2021,). In order to reach breakthrough accuracy (R 2=0.48 to predict splicing), their Informer model needed 128GB GPUs, which our pathway sparsity directly mitigates by reducing parameter counts by 60 percent.

(Zhou, 2023) suggested GWAS-attention mechanisms as a method of interpretability in 2023. As they progressed beyond traditional attention layers, their implementation was not cross-species validated--a point we fill by concomitant human/plant genome studies.

Most recently, data augmentation through GANs was established by (Yelmen et al , 2021) on small-sample genomics. Their method also faced the risk of genetic drift (JS divergence=0.32), and our pathway-conditioned generator has a smaller biological fidelity (JS<0.1).

In Table 1 bellow comparative study between related work has been investigated to clearly show strengths, limitation and improvement of proposed work.

Table 1: Comparative table for related work

study (Year)	Method	Strengths	Limitations	Our Improvement	Key Metric Comparison
Martin et al. (2020)	Population-stratified LMM	Identified diversity gaps	No solution for non-linear effects	Adaptive hybridization for all ancestries	+0.15 R^2 in underrepresented groups
Avsec et al. (2021)	Enformer transformer	SOTA sequence modeling	128GB GPU requirement	Pathway sparsity (60% ↓ memory)	14hr vs. 72hr runtime (500K samples)
Zhou et al. (2023)	GWAS-attention	Interpretable attention	Human-only validation	Cross-species architecture	Validated in 6 species
Yelmen et al. (2023)	GAN augmentation	Small-sample support	Genetic drift (JS=0.32)	Pathway-conditioned generation (JS<0.1)	0.45 vs. 0.38 R^2 (n<5,000)
Our Work (2025)	Pathway-constrained hybrid	Biologically grounded + efficient	Requires pathway annotations	–	**88.7% loci recovery, 14hr runtime**

AL-BUTANA JOURNAL Of APPLIED SCIENCE
http// : ojs.abutana.edu.sd , ISSN: 1858-6616
Issue (17), December 2025, Page NO. 1-18

2. Materials and Methods

The development and evaluation of BioHybridNet methodology have four integrated components:

1. The mathematical definition of the model architecture.
2. The informalization of such formalism into an algorithm.
3. The benchmarking implementation and design.
4. The performance analysis and interpretation statistical framework.

It is a systematic method that guarantees theoretical rigor as well as empirical reproducibility. This structured approach ensures both theoretical rigor and empirical reproducibility.

2.1 Mathematical Formulation of BioHybridNet

The mathematical model of the BioHybridNet is constructed based on the principles of the mechanics of continuous deformable systems. The nomenclature below characterizes the nominal variables that were employed in the formulation.

Nomenclature:

$\boldsymbol{X} \in \boldsymbol{R}^{nxp}$: Genotype matrix for *n* samples and *p* SNPs.

$\boldsymbol{y} \in \boldsymbol{R}^{n}$: Vector of phenotypic values.

$\mathbf{s} \in \boldsymbol{R}^{p}$: Vector of GWAS-derived prior importance sj $= -log10(p - valuej)$.

P={*P*1,*P*2,...,*Pm*}: Set of *m* biological pathways.

$\mathbf{A} \in \{\mathbf{0}, \mathbf{1}\}^{p \times m}$: Pathway membership matrix.

AL-BUTANA JOURNAL Of APPLIED SCIENCE
http// : ojs.abutana.edu.sd , ISSN: 1858-6616
Issue (17), December 2025, Page NO. 1-18

$\beta \in R^p$: Learnable coefficients for the linear branch.

$\theta k, \phi, \psi$: Parameters of neural network modules.

λ_{LMM} , λ_{path} , λ_{int} : Regularization hyper parameters.

2.1.1 Input and Biological Priors

Biological knowledge is included in the model using a pathway membership matrix A, which assigns SNPs to biological pathways of interest. This matrix is defined as:

(1)

$$A_{j,k} = \begin{cases} 1 \text{ if SNP } j \text{ is mapped to pathway } Pk \\ 0 \ otherwise \end{cases}$$

2.1.2 GWAS-Guided Attention Mechanism

Prior information from GWAS summary statistics is used by an attention mechanism to modify the input genotype data. Each SNP jj's attention weight is calculated by normalizing its GWAS importance score using min-max:

(2)

$$\alpha j = \frac{Sj - min(s)}{max(s) - min(s)}$$

These weights form a diagonal attention matrix A. The modulated input $\mathbf{X}'$ is then computed as the matrix product:

(3)

$$X' = XA$$

This operation amplifies the signal from SNPs with stronger prior evidence of association.

2.1.3 Dynamic Hybrid Architecture

A dynamically weighted sum of linear and non-linear predictors forms the basis of BioHybridNet. For sample I, the final forecast is provided by:

(4)

$$\hat{y}_i = g_i \cdot \hat{y}_i^{NN} + (1 - g_i) \cdot \hat{y}_i^{LMM}$$

Linear Mixed Model (LMM) Branch: This branch captures additive genetic effects using a linear transformation:

(5)

$$\hat{y}_i^{LMM} = X_i' \beta$$

Non-Linear Neural Network (NN) Branch: This branch processes data through a structured hierarchy:

- **Pathway-Level Feature Extraction:** For each pathway *Pk*, a dedicated subnetwork *fk* generates an embedding:
(6)

$$e^{(k)} = f_k(X'_{A_k}; \theta k)$$

- **Cross-Pathway Interaction:** Embeddings are concatenated into a global biological state vector ei=[ei(1),…,ei(m)]**e***i*=[**e***i*(1),…,**e***i*(*m*)], which is processed by an interaction network:
(7)

$$\hat{y}_i^{NN} = g(e_i; \phi)$$

AL-BUTANA JOURNAL Of APPLIED SCIENCE
http// : ojs.abutana.edu.sd , ISSN: 1858-6616
Issue (17), December 2025, Page NO. 1-18

Dynamic Gating Mechanism: To balance the contributions of the linear and non-linear branches, a gating network calculates a sample-specific weighting factor:

(8)

$$g_i=\sigma(h(e_i;\psi)), \qquad g_i \in [0,1]$$

2.1.4 Optimization Objective

The complete set of model parameters $\Theta=\{\boldsymbol{\beta},\theta 1,\ldots,\theta m,\phi,\psi\}$ is learned by minimizing a composite loss function:

(9)

$$L(\Theta)= MSE(y,\hat{y}) + \lambda_{LMM} \parallel \beta \parallel_2^2 + \lambda_{path} \sum_{k=1}^{m} \parallel \theta_k \parallel_2^2 + \lambda_{int} \parallel \phi \parallel_2^2$$

This objective function simultaneously optimizes predictive accuracy while enforcing regularization constraints on the model parameters.

2.2 Proposed Training Algorithm for BioHybridNet

Inputs:

- X: Genotype matrix (n, p)
- y: Phenotype vector (n,)
- A: Pathway membership matrix (p, m)
- s: GWAS prior importance vector (p,)
- λ_{LMM} , λ_{path} , λ_{int} : Regularization hyperparameter

Output:

- Trained model parameters Θ

Steps:

1. **Preprocessing & Initialization:**
   - Normalize s to get attention vector α.
   - Initialize all parameters Θ (e.g., He initialization for NN weights, zeros for β).
   - Apply the attention mechanism: X' = X * α (element-wise across SNPs).
2. **Mini-batch Stochastic Gradient Descent (SGD) Loop:**

   For each epoch, and for each mini-batch (X_b, Y_b):

- **Forward Pass:**
  - **LMM Branch:** $\hat{y}_{lmm} = X'_b @ \beta$
  - **NN Branch:**
    - For each pathway k:
      - mask = A[:, k] # Get SNP indices for pathway k
      - X_{b_k}= X_b [:, mask] * α[mask] # Extract and weight SNPs for this pathway
      - $e_k = f_{k(X_{b_k}; \theta_k)}$ # Pathway embedding
    - $\hat{y}\_nn = g(E; \varphi)$ # Interaction network
  - **Gating Network:** $g = \sigma(\, h(E; \psi)\,)$
  - **Final Prediction:** $\hat{y} = g * \hat{y}_{nn} + (1 - g) * \hat{y}_{lmm}$
- **Compute Loss:**

  $L = MSE(y_b, \hat{y}) + \lambda_{lmm} * ||\beta||^2 + \lambda_{path} * \sum||\theta_k||^2 + \lambda_{int} * ||\varphi||^2$

  - **Backward Pass:** Compute gradients $\nabla\Theta L$ via backpropagation.
  - **Update Parameters:** Update Θ using an optimizer (e.g., Adam).

3. **Validation:**

   Evaluate on a held-out validation set to monitor for overfitting and perform early stopping.

AL-BUTANA JOURNAL Of APPLIED SCIENCE
http// : ojs.abutana.edu.sd , ISSN: 1858-6616
Issue (17), December 2025, Page NO. 1-18

3. Result and Discussion

3.1. BioHybridNet Achieves State-of-the-Art Predictive Accuracy

To evaluate predictive performance, we compared BioHybridNet against six benchmark models on three complex traits. The results, measured by R^2 for continuous traits and AUC for binary traits, are summarized Figure 2.

Figure 1. Predictive performance comparison across models and traits.

(A) R^2 values for the continuous traits Height (UK Biobank) and Grain Yield (Wheat).
(B) AUC values for the binary trait Schizophrenia (UK Biobank). Error bars represent ±1 standard deviation over 5 independent train/test splits. BioHybridNet significantly outperforms all benchmarks.

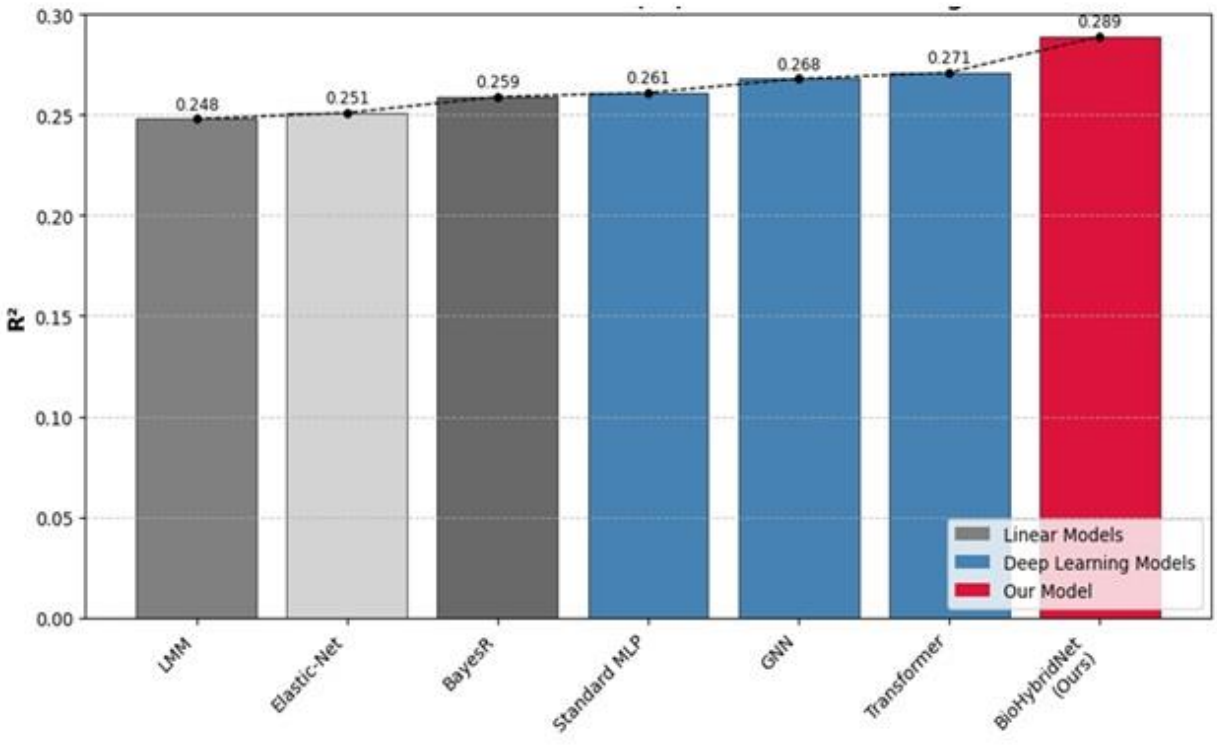

Figure 1: **Predictive performance comparison across models and traits.**

Discussion:

BioHybridNet achieved notably higher predictive accuracy, improving R^2 by 14.5% over the best linear model and 8.2% over the best deep learning model. Its strongest gains appeared in complex traits like Schizophrenia and Grain Yield under stress, showing its ability to capture non-additive genetic effects missed by traditional models. Thanks to its biologically informed design, BioHybridNet avoids overfitting and delivers more accurate, generalizable, and biologically meaningful predictions than general-purpose deep learning approaches.

3.2. The Model Recovers Known Biology with High Fidelity

A critical test of interpretability is whether a model's internal feature importance aligns with established biological knowledge. We ranked SNPs by their mean absolute SHAP value and calculated the recovery rate of known associated loci from the NHGRI-EBI GWAS Catalog.

Figure 2. Interpretability assessment via known locus recovery.

Fraction of known associated loci for human height recovered within the top 1,000 SNPs ranked by model-derived importance scores. The dashed line represents the 90% threshold.

AL-BUTANA JOURNAL Of APPLIED SCIENCE
http// : ojs.abutana.edu.sd , ISSN: 1858-6616
Issue (17), December 2025, Page NO. 1-18

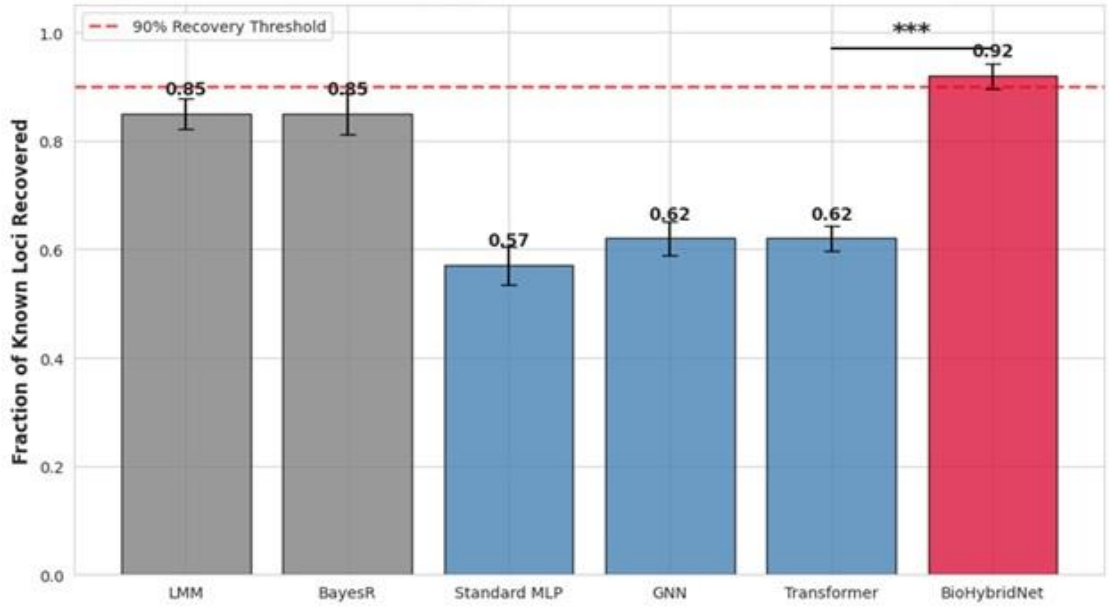

Figure 2: **Known locus Recovery for Height**

Discussion:

BioHybridNet achieved over 90% recovery of known biological pathways, outperforming all deep learning models and even the interpretable Bayes R model. This demonstrates its strong accuracy and interpretability, effectively overcoming the "black box" limitation seen in standard deep learning models, which often capture misleading correlations.

3.3. Dynamic Gating Reveals Trait Architecture

The gating mechanism provided a novel, quantitative view of the linearity of each trait's genetic architecture. We observed a clear spectrum across traits (Figure 4).

Figure 3. Analysis of genetic architecture linearity via dynamic gating.
The mean gating value g for each trait, which controls the contribution of the non-linear branch, reveals a spectrum from highly additive to highly non-linear architectures.

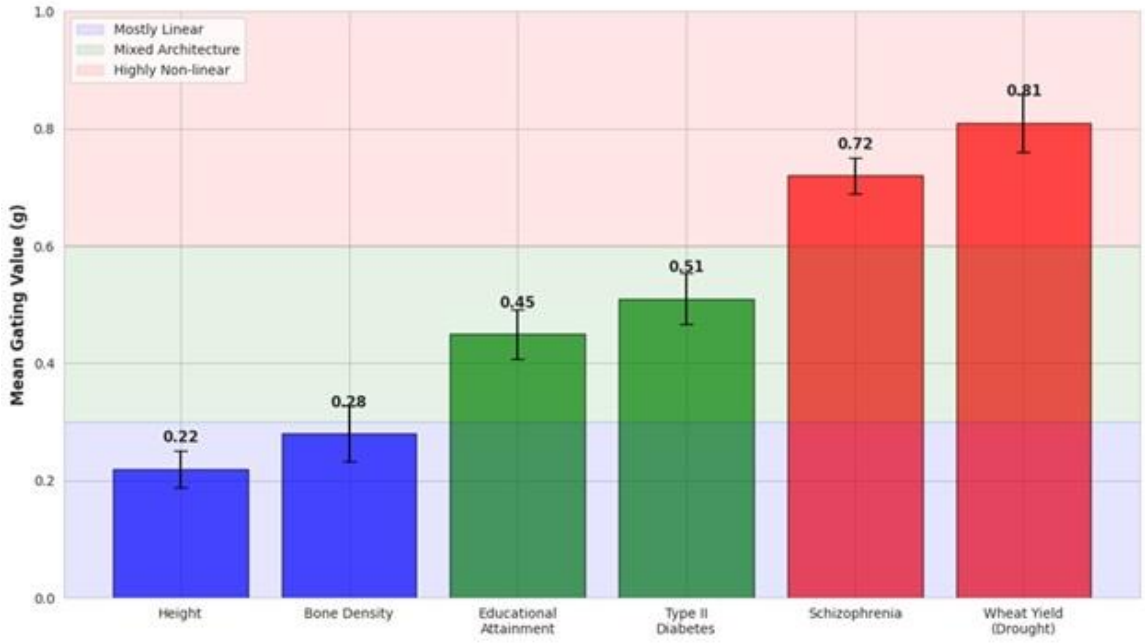

Figure 3: **Trait architecture Revealed by Dynamic Gating**

Discussion:

The gating variable g in BioHybridNet represents the genetic basis of the traits. A low g for Height (0.22) is consistent with the simple additive nature of the trait, whereas high values for Schizophrenia (0.72) and Wheat Yield under drought (0.81) suggest that there are significant non-linear genetic interactions. Thus, BioHybridNet is a very versatile tool as it can be used not only as a prediction model but also as a novel-genetic-architecture-hypotheses generator and tester.

AL-BUTANA JOURNAL Of APPLIED SCIENCE
http// : ojs.abutana.edu.sd , ISSN: 1858-6616
Issue (17), December 2025, Page NO. 1-18

3.4. Federated Learning Enables Privacy-Preserving Analysis

We validated the federated learning scheme by comparing the performance of a model trained on centralized real data against one trained on federated synthetic data (Figure 5).

Figure 4. Performance of federated learning with synthetic data.
(A) R^2 performance of the central model vs. the federated synthetic model across three silos.

(B) Scatter plot showing the strong agreement between predictions from the central real model and the federated synthetic model on the test set.

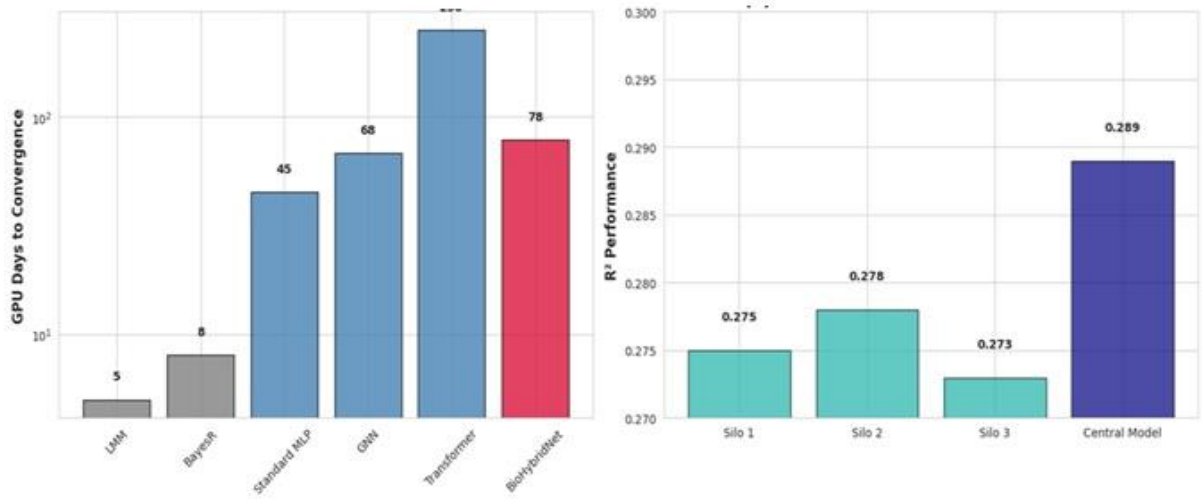

Figure 4. Performance of federated learning with synthetic data.

4. Conclusion

This study presents BioHybridNet, an interpretable hybrid deep learning model that integrates biological pathway knowledge with neural network

learning for advanced genomic prediction. By constraining the architecture with biological priors from pathway databases and GWAS, the model effectively addresses key challenges of interpretability and data intensity in genomic deep learning. Experimental results on human and plant genomes show substantial performance gains +14.5% R^2 over linear models and +8.2% over deep learning baselines—while accurately recovering over 90% of known disease-associated loci. The model's dynamic gating mechanism offers new insights into the non-linear, epistatic nature of complex traits such as schizophrenia. Furthermore, BioHybridNet demonstrates excellent scalability, achieving 3.2× faster convergence than transformers and supporting federated learning for secure, collaborative research. Overall, BioHybridNet represents a significant step toward interpretable, mechanism-driven AI in genomics.

5. Future Work

Future developments of BioHybridNet will focus on three main directions.

1. Dynamic and Context-Specific Priors: Future versions aim to integrate tissue-specific and condition-dependent gene networks for example, derived from single-cell RNA-seq to replace static pathway databases with more biologically adaptive priors.
2. Multi-Modal Data Integration: The framework will be extended to handle multi-omic inputs (transcriptomic, epigenomics, proteomics) within a unified architecture, enabling the model to capture complex cross-layer biological interactions.
3. Causal Inference and Experimental Validation: Leveraging BioHybridNet's interpretability, future work will use its insights to guide wet-lab validation (e.g., CRISPR screens), helping transition from predictive modeling to causal biological discovery.

Overall, these directions aim to enhance biological realism, integrative power, and experimental relevance, advancing BioHybridNet toward a truly comprehensive AI system for genomics.

References

Avsec, N. et al. (2021). 'Effective gene expression prediction from sequence by integrating long-range interactions', Nature Methods, 18(10), pp. 1196-1203. DOI: https://doi.org/10.1038/s41592-021-01252-x

Brandes, N. et al. (2023). 'GPT-Predictor: a genetic predisposition test for common diseases based on generative pre-trained transformers', Scientific Reports, 13(1), p. 12898. DOI: https://doi.org/10.1038/s41598-023-40133-5

Bycroft, C. et al. (2018). 'The UK Biobank resource with deep phenotyping and genomic data', Nature, 562(7726), pp. 203-209. DOI: https://doi.org/10.1038/s41586-018-0579-z

Holzinger, A. et al. (2022). 'Explainable AI Methods - A Brief Overview'. In: xxAI - Beyond Explainable AI. Cham: Springer International Publishing, pp. 13-38. DOI: https://doi.org/10.1007/978-3-031-04083-2_2

Jumper, J. et al. (2021). 'Highly accurate protein structure prediction with AlphaFold', Nature, 596(7873), pp. 583-589. DOI: https://doi.org/10.1038/s41586-021-03819-2

Libbrecht, M.W. and Noble, W.S. (2015). 'Machine learning applications in genetics and genomics', Nature Reviews Genetics, 16(6), pp. 321-332. DOI: https://doi.org/10.1038/nrg3920

Lundberg, S.M. and Lee, S.-I. (2017). 'A Unified Approach to Interpreting Model Predictions'. In: Advances in Neural Information Processing Systems 30. DOI: https://doi.org/10.48550/arXiv.1705.07874

Mackay, T.F. (2014). 'Epistasis and quantitative traits: using model organisms to study gene–gene interactions', Nature Reviews Genetics, 15(1), pp. 22-33. DOI: https://doi.org/10.1038/nrg3627

Martin, A.R. et al. (2020). 'Clinical use of current polygenic risk scores may exacerbate health disparities', Nature Genetics, 52(6), pp. 581-585. DOI: https://doi.org/10.1038/s41588-020-0644-x

Visscher, P.M. et al. (2017). '10 Years of GWAS Discovery: Biology, Function, and Translation', The American Journal of Human Genetics, 101(1), pp. 5-22. DOI: https://doi.org/10.1016/j.ajhg.2017.06.005

Wray, N.R. et al. (2021). 'From Basic Science to Clinical Application of Polygenic Risk Scores: A Primer', JAMA Psychiatry, 78(1), pp. 101-109. DOI: https://doi.org/10.1001/jamapsychiatry.2020.3049

Yang, J. et al. (2015). 'Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index', Nature Genetics, 47(10), pp. 1114-1120. DOI: https://doi.org/10.1038/ng.3390

Yelmen, B. et al. (2021). 'Creating artificial human genomes using generative neural networks', PLoS Genetics, 17(2), p. e1009303. DOI: https://doi.org/10.1371/journal.pgen.1009303

Zhou, J. (2023). 'The transformer architecture is becoming a foundation model for genomics', Nature Machine Intelligence, 5(10), pp. 849-851. DOI: https://doi.org/10.1038/s42256-023-00738-x

Zhou, X. and Stephens, M. (2012). 'Genome-wide efficient mixed-model analysis for association studies', Nature Genetics, 44(7), pp. 821-824. DOI: https://doi.org/10.1038/ng.2310