

Development and Interpretation of an Epigenetic Clock Using DNA Methylation Data From GSE87571

Presented by: Neha Singh , MSc

Advisors: Dr. Nabil Atallah, & Dr. Oyeronke Ayansola (BINF7700: Bioinfo Research Directions)

Abstract

Purpose: Biological age predicts health outcomes better than chronological age, but existing epigenetic clocks face an accuracy-interpretability trade-off.

Methods: We developed a blood-based epigenetic clock using GSE87571 dataset. Elastic Net regression with 10-fold cross-validation selected 153 age-predictive CpG sites from 485,512 candidates. SHAP analysis validated model interpretability.

Results: Our model achieved MAE of 2.55 years ($R^2 = 0.976$). The top predictor, cg16867657 (ELOVL2), showed perfect SHAP linearity ($R^2 = 1.0$). Pathway enrichment revealed associations with CNS development, cellular senescence, and Hippo signaling.

Conclusion: Combining Elastic Net with SHAP validation achieves both high accuracy and biological interpretability, advancing epigenetic clock development for clinical translation.

Introduction

Why Biological Age Matters:

Aging drives chronic diseases, yet individuals age at different rates. While chronological age counts years lived, biological age reveals true physiological state and better predicts disease risk and mortality.

Epigenetic Clocks:

DNA methylation refers to addition of methyl groups at CpG sites which regulates gene expression and changes systematically with age. Since Horvath's 2013 clock, these methylation patterns have been used to estimate biological age from blood samples.

The Challenge:

Current epigenetic clocks face a critical trade-off: advanced models achieve high accuracy, but obscure which CpG sites drive predictions and why they matter biologically. This interpretability gap limits therapeutic target identification and clinical translation.

Our Approach:

We integrated Elastic Net regression (feature selection), SHAP analysis (CpG contribution quantification), and pathway enrichment (mechanistic discovery) to achieve both accuracy and interpretability.

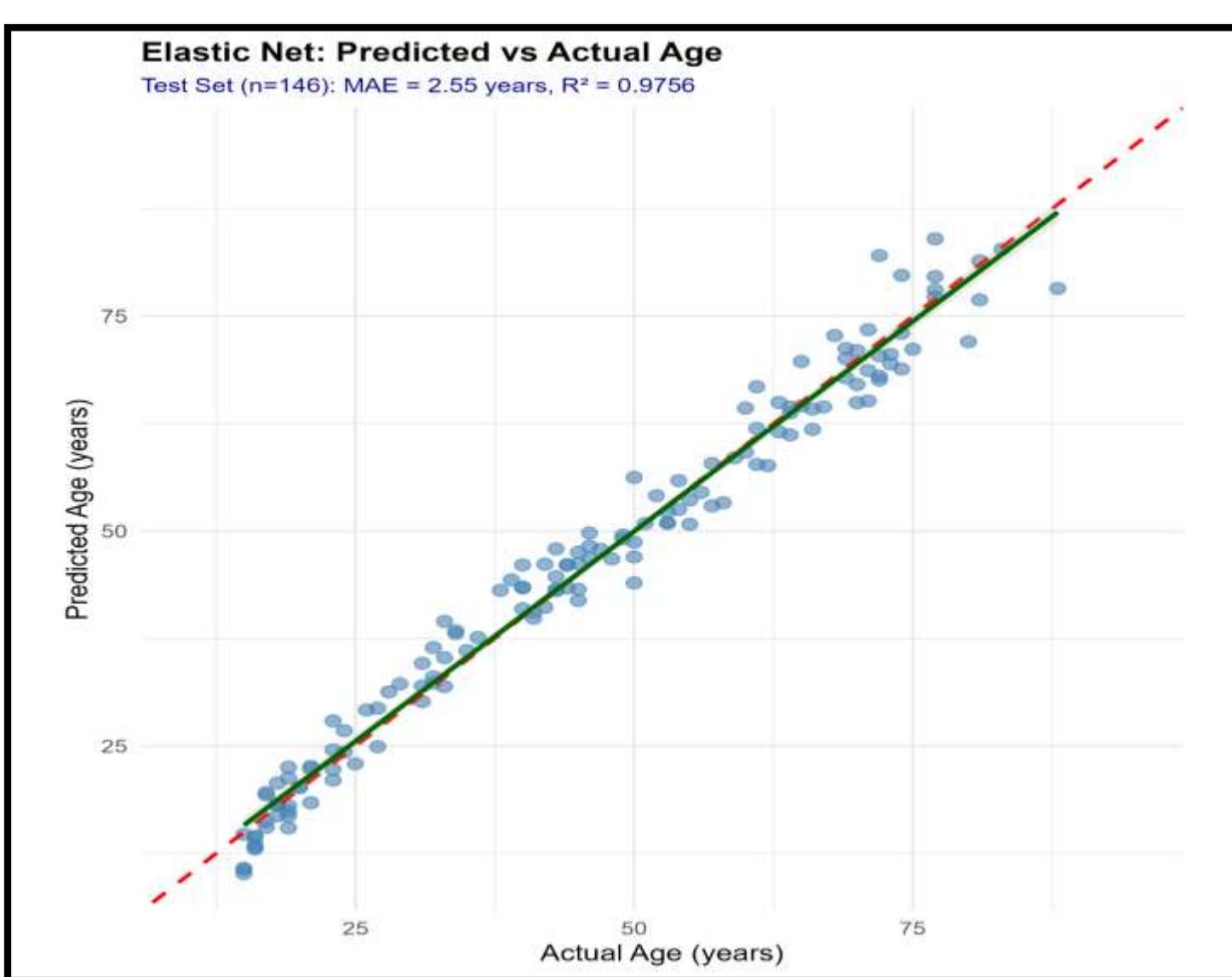


Figure 1. Predicted versus actual age

Model	R2	MAE (years)	RMSE (years)
Elastic Net	0.976	2.55	3.21

Table 1. performance metrics of Elastic Net

Methods and Materials

Dataset & Feature Selection:

- GSE87571: 729 blood samples, ages 14-94 (Illumina 450K array)
- Top 500 age-correlated CpGs selected ($p < 1 \times 10^{-5}$)

Modeling:

- Elastic Net regression with 10-fold cross-validation
- 80/20 train/test split → 153 CpG sites selected

Interpretability & Validation:

- SHAP analysis quantified CpG contributions
- Validated coefficient-SHAP correlation ($r = 0.84$)

Biological Annotation:

- 153 CpGs mapped to 108 genes
- Pathway enrichment: KEGG & GO Biological Process

Evaluation:

- Performance metrics: MAE, R^2 , comparison to Horvath/Hannum clocks

Results

Model Performance:

Elastic Net achieved MAE = 2.55 years ($R^2 = 0.976$), outperforming Horvath (3.6 years) and Hannum (4.9 years) clocks. Strong correlation between predicted and actual age ($r = 0.9756$) across ages 14-94.

Feature Importance:

Model selected 153 CpGs with balanced methylation: 81 hypermethylated, 72 hypomethylated. Top predictor cg16867657 (ELOVL2) showed strongest association with age.

SHAP Validation:

SHAP analysis confirmed feature rankings. cg16867657 showed perfect linearity ($R^2 = 1.0$), validating biological signal over overfitting.

Biological Pathways:

153 CpGs mapped to 108 genes. Enrichment revealed associations with CNS development ($p = 2.6 \times 10^{-5}$), cellular senescence ($p = 0.010$), and Hippo signaling ($p = 0.002$), connecting predictions to fundamental aging mechanisms.

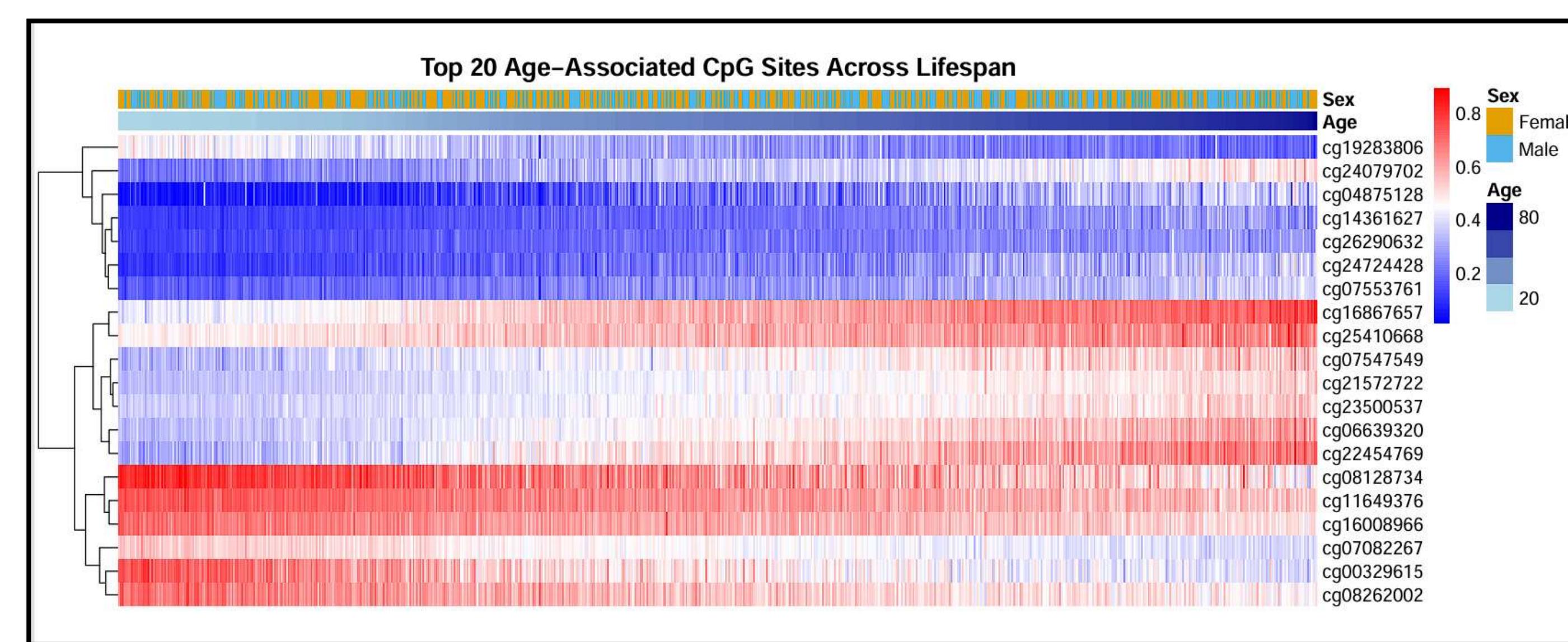


Figure 2. Top 20 CpG sites show coordinated hypermethylation and hypomethylation with age

Discussion

Model Performance & Interpretability:

Our Elastic Net-SHAP framework achieved MAE = 2.55 years, outperforming Horvath (3.6 years) and Hannum (4.9 years) clocks while maintaining interpretability. Strong correlation between coefficients and SHAP values validates consistent feature importance, addressing the accuracy-interpretability trade-off in epigenetic clock research.

ELOVL2 as Premier Aging Biomarker:

The top predictor, cg16867657 in ELOVL2 gene, showed exceptional correlation ($r = 0.9464$) with perfect SHAP linearity. ELOVL2 regulates very long-chain fatty acid synthesis and has been validated across populations as a robust aging biomarker. Its sex-independent hypermethylation pattern makes it ideal for clinical applications.

Biological Mechanisms of Aging:

Pathway enrichment connects our predictions to fundamental aging processes. CNS development supports developmental programming theory as age-related methylation reflects persistent early-life regulatory programs. Cellular senescence links predictions to senescent cell accumulation, a key aging hallmark. Hippo signaling suggests declining regenerative capacity. Moreover, Balanced hyper/hypomethylation indicates coordinated epigenetic remodeling rather than stochastic drift.

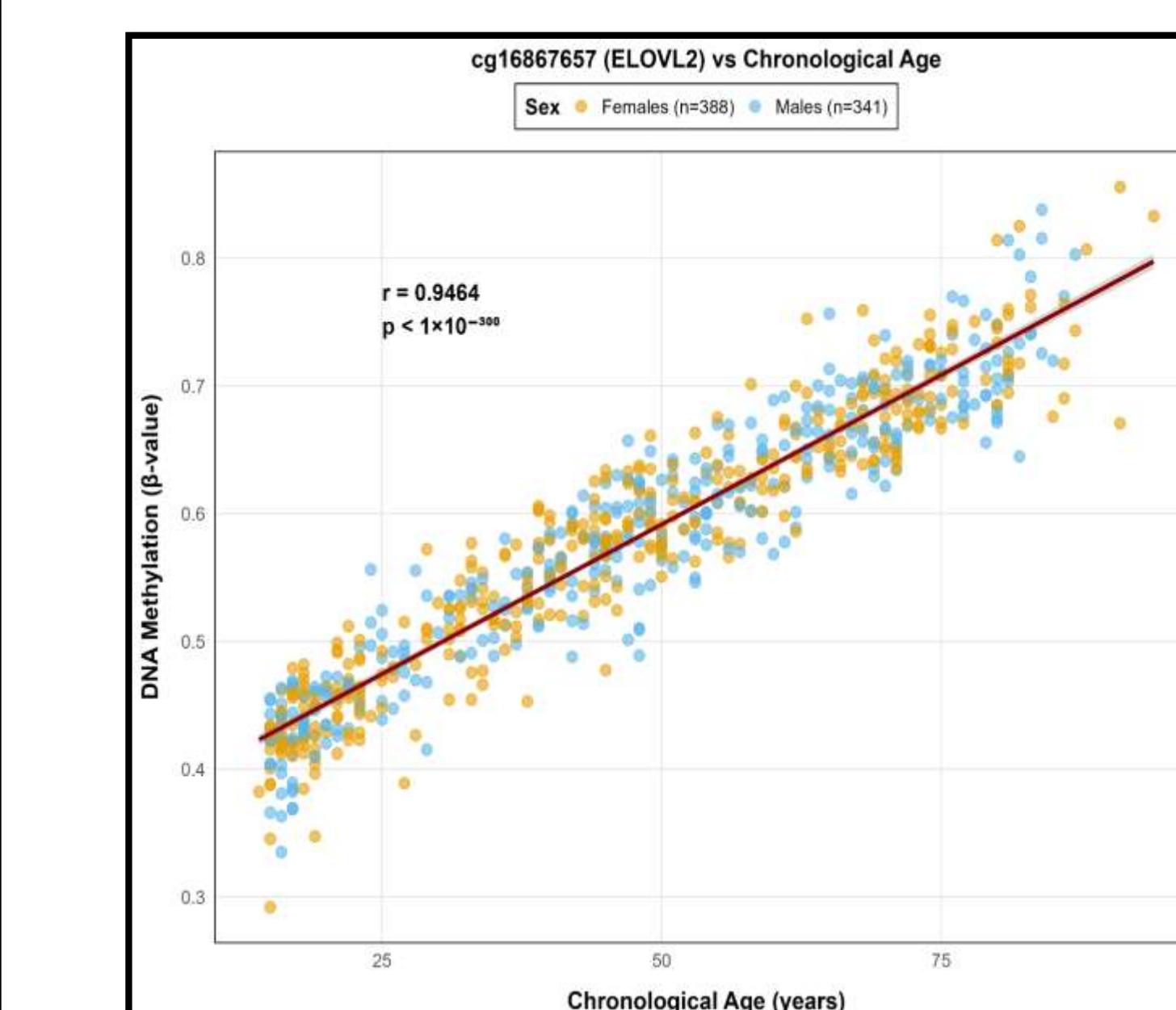


Figure 3. cg16867657 shows strongest age correlation independent of biological sex

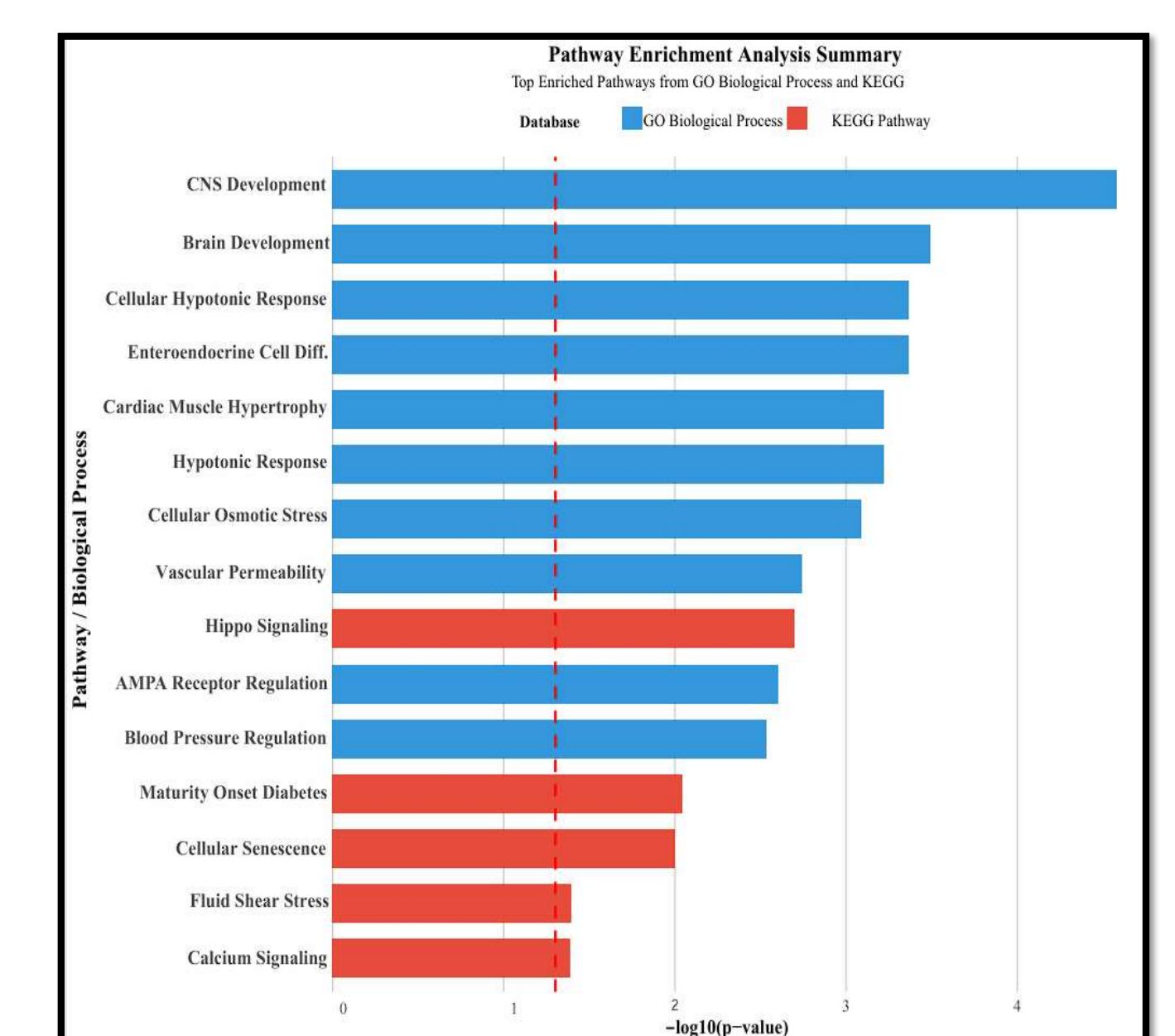


Figure 4. Pathway enrichment summary for 108 age-associated genes. (Blue bars =GO Biological Process; red bars =KEGG; Red dashed line = threshold).

Conclusions

This study demonstrates that combining Elastic Net regression with SHAP validation achieves both high accuracy and biological interpretability in epigenetic clock development. By integrating pathway enrichment analysis, we transformed the clock from a purely predictive tool into a mechanistic instrument for understanding human aging.

The 153 validated CpG sites represent potential biomarkers for monitoring biological aging and therapeutic interventions. SHAP-based interpretability enables therapeutic target identification while building clinical trust necessary for translation.

Future validation across diverse populations, longitudinal aging rate tracking, and experimental validation through epigenetic editing will advance understanding from correlation to causation, ultimately contributing to health span extension strategies.

Contact

Neha Singh
singh.neha3@northeastern.edu
College of Science
Northeastern University
375 Queen St W, Toronto, ON M5V 2A5

References

- Horvath S. (2013). DNA methylation age of human tissues and cell types. *Genome Biol.*
- Hannum G. et al. (2013). Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol Cell.*
- Johansson Å. et al. (2013). Continuous aging of the human DNA methylome throughout the human lifespan. *PLoS ONE.*
- López-Otín C. et al. (2023). Hallmarks of aging: An expanding universe. *Cell.*
- Teschendorff A.E., Horvath S. (2025). Epigenetic ageing clocks: statistical methods and emerging computational challenges. *Nat Rev Genet.*
- Lundberg S.M., Lee S.I. (2017). A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst.*
- Shireby G.L. et al. (2020). Recalibrating the epigenetic clock: Implications for assessing biological age in the human cortex. *Brain.*
- McCrory C. et al. (2021). GrimAge outperforms other epigenetic clocks in prediction of age-related clinical phenotypes. *J Gerontol.*
- Garagnani P. et al. (2012). Methylation of ELOVL2 gene as a new epigenetic marker of age. *Aging Cell.*
- Belsky D.W. et al. (2022). DunedinPACE, a DNA methylation biomarker of the pace of aging. *eLife.*