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

Submitted by:

Neha Singh

Advisors:

Professor Nabil Atallah

Professor Oyeronke Ayansola

Master of Science in Bioinformatics

College of Science

Northeastern University

November 15, 2025

TABLE OF CONTENTS

ABSTRACT

Biological age, measured through DNA methylation-based epigenetic clocks, provides superior health outcome prediction compared to chronological age. However, existing models face a critical trade-off between predictive accuracy and biological interpretability. This study developed an accurate and interpretable blood-based epigenetic clock by integrating Elastic Net regression with SHAP (SHapley Additive exPlanations) analysis and pathway enrichment. Using the GSE87571 dataset (729 blood samples, ages 14-94, Illumina 450K array), rigorous feature selection identified 500 age-associated CpG sites from 485,512 candidates. Elastic Net regression with 10-fold cross-validation selected 153 CpG sites achieving mean absolute error of 2.55 years, outperforming Horvath (3.6 years) and Hannum (4.9 years) clocks. The model explained 97.56% of age variance with consistent performance across all ages. SHAP analysis validated interpretability, showing strong correlation ($r = 0.84$) between coefficients and SHAP importance. The top site cg16867657 in *ELOVL2* exhibited perfect linearity ($R^2 = 1.0$) in SHAP dependence analysis, confirming it as primary age predictor biomarker. The pathway enrichment mapped 153 CpG sites to 108 genes, revealing associations with central nervous system development ($p = 2.6\times10^{-5}$), cellular senescence ($p = 0.010$), and Hippo signaling ($p = 0.002$), connecting predictions to fundamental drivers of aging. The balanced hypermethylation (52.9%) versus hypomethylation (47.1%) suggests coordinated epigenetic remodeling. This study demonstrates that combining penalized regression with explainable SHAP-based model interpretability framework addresses the interpretability-accuracy balancing relationship. The reproducible framework integrating machine learning, SHAP validation, and pathway analysis advances methodological rigor and mechanistic understanding. Interpretability plays a critical role for achieving epigenetic clocks transition to

clinical biomarkers. The SHAP-validated approach enables identification of therapeutic targets while maintaining transparency necessary for clinical translation.

INTRODUCTION

Aging is the most basic and foundational biological process characterized by gradual decline in physiological functioning and enhanced vulnerability to diseases followed by mortality. The chronological age is a parameter for measuring time elapsed since birth; however, biological age indicates the true state of physiological and functional ability of living organism (López-Otín et al., 2023). This difference has become immensely relevant in the present-day ethos of modern medicine as biological age is a prime predictor of disease susceptibility, health outcomes and calculation of remaining life span of a living organism. The ability of biological age to immaculately quantify aging has far fetching implications in the modern-day personalized medications, enabling early detection of accelerated aging, assessing therapeutic interventions and development of plans for enhancing health as well as life span (An et al., 2025).

In connection to the foregoing discussion, it is imperative to note that epigenetic alterations have emerged as a hallmark in biomarkers. Epigenetic alterations or modifications include reversible changes to chromatin structure and DNA, which is responsible for regulating gene expression without causing any change in the genetic sequences. It is further pertinent to note that DNA and chromatin structure undergo significant alterations with age across all cell types and tissues. These changes include histone modifications, DNA methylation and chromatin remodeling, which is collectively responsible for functional decline in an organism with its increasing age (An et al., 2025). As opposed to genetic mutations, epigenetic modifications are mostly reversible making them immensely suitable for developing anti-aging interventions. Recent studies indicate that

epigenetic drift, the gradual accumulation of stochastic epigenetic modifications over time, constitutes a focal regulator of age-related issues like cancer, cardiovascular diseases and neurodegeneration. Additionally, it also regulates organismal aging (Borrego-Ruiz & Borrego, 2024).

In the arena of aging research, DNA methylation which involves addition of methyl groups to cytosine bases predominantly at cytosine-guanine dinucleotides (CpG sites) is the most comprehensive and intricately studied epigenetic modification. This modification plays a significant role in comprehending differentiations at cellular level, gene expression and the development of an organism throughout its lifespan (Mc Auley & Morgan, 2025). However, it cannot be ignored that Patterns under DNA methylation undergoes many age- related changes that can act as biomarkers of biological age. Furthermore, genome-wide methylations research has recognized thousands of CpG sites showcasing systematic alterations with age. Some of them show hypermethylation while others show hypomethylations (Seale et al., 2024). Thereby, DNA methylation has emerged as a substructure for estimating epigenetic age amid the ongoing dynamic patterns combined with easy access of DNA methylation measurements from less invasive blood samples and constant stability. Furthermore, it is pertinent to note that the distribution of age-related methylation is not uniform across the genome but is enriched in specific regulatory regions including developmental genes, polycomb group protein targets, and bivalent chromatin domains, indicating functional additions to the aging process itself.

In the year 2013, Horvath developed a multi tissue predictor using 353 CpG sites which revolutionized the concept of epidenetic clock that is a machine learning model aimed at estimating biological age with the help of DNA methylation patterns. This landmark development led to a deduction that DNA methylation modifications take place predictably enough for accurate

computational age modeling (Horvath, 2013a). Subsequently, many such clocks with varied and complex applications have been developed. Initially, the first-generation clocks such as Horvath's pan-tissue clock or Hannum's blood-specific clock aimed at predicting the chronological age while establishing correlation with age-related phenotypes. Furthermore, second gen clocks like PhenoAge and GrimAge were particularly designed and trained to estimate health outcomes rather than calculating chronological age alone, thereby offering enhanced predictability in matters of health span, mortality and morbidity (McCrory et al., 2021). Lately, third generation clocks such as DunedinPACE are used to comprehend the pace of aging rather than the age itself. It seeks to provide insight as to the speed of aging in individuals and organisms (Belsky et al., 2022). These advancements have placed epigenetic clocks at a potentially powerful stature with respect to clinical translation, forensic science and basic aging research.

Development of an epigenetic clock involves a foundational machine learning challenge which is the identification of informative subsets from a wide range of high-dimensional methylation data. It further creates challenge in the development of an accurate age predictor. Initially clocks primarily employed elastic net regression which is a penalized linear method combining L1 and L2 penalties for simultaneous prediction and feature selection. Explicit CpG contribution, automatic feature selection and efficient handling of correlated features are the most significant benefits of elastic net regression (Shireby et al., 2020a). However, as compared to the conventional elastic net models, the recent advancements suggest greater predictive performance in understanding feature interaction and capturing nonlinear relationships. These include advanced algorithms such as gradient boosting (XGBoost, LightGBM) and deep neural networks. As compared to linear baselines, deep learning models have exemplified enhanced accuracy and improved biological interpretations. However, they are often coupled with many interpretabilities

challenge (de Lima Camillo et al., 2022; Prosz et al., 2024). This performance-interpretability trade-off represents a critical challenge in the field.

Though advanced machine learning models enhance predictive accuracy, the resultant ‘black box’ model obstructs comprehension of CpG site interaction and what CpG sites cause predictions. Conventional approaches to interpretability aimed at examining coefficient magnitudes in linear models, though providing global importance, but failed to reveal feature interaction. This gap in interpretability is a limitation to biological insights as it reduces clinical trust and obstructs recognition of therapeutic targets.

A feature-based interpretability method, SHAP (SHapley Additive exPlanations) have developed to address these limitations in particular (Ponce-Bobadilla et al., 2024). It provides model-agnostic feature importance quantification based on cooperative game theory, enabling consistent interpretation across model architectures. It calculates contribution of every feature to singular prediction by taking into account all possible feature combinations and aims at ensuring accuracy as well as consistency. Applications to epigenetic clocks have revealed interactive CpG sites with nonlinear age relationships, mapped important sites to enhancers and CTCF binding regions, and connected predictions to aging pathways involving DNA damage, senescence, and metabolic dysfunction. SHAP's visualizations like beeswarm plots, dependence plots, and waterfall diagrams, together facilitate communication to both computational and biological audiences.

This study uses comprehensive sources for researching age-related DNA methylation in human blood i.e., GSE87571 dataset. This dataset comprises of 732 blood samples from persons between 14 to 94 years of age, profiled using the Illumina HumanMethylation450 BeadChip interrogating over 485,000 CpG sites genome-wide (Johansson et al., 2013). This dataset includes wide range of CpG islands, intergenic regions, gene promoters and bodies, making it suitable for a vigorous

clock development. Furthermore, the utility of GSE87571 dataset has been validated by much subsequent research such as Vershinina et al. (2021) and Yusipov et al. (2020). The former researchers analyzed 341 male and 388 female samples to differentiate nonlinear, deterministic and stochastic features of age-related methylation (Vershinina et al., 2021); and the latter aimed at understanding sex-specific methylation patterns with certaing CpG sites indicating varied trajectories in males and females (Yusipov et al., 2020). Investigating across such a vast range of age groups enables development of accurate clocks for different age groups.

Despite all the advancements in recent years, there still exist significant gaps in epigenetic clock research. A comprehensive review highlighted critical challenges including the complexity-interpretability trade-off, lack of standardized preprocessing protocols, and insufficient biological integration (Teschendorff & Horvath, 2025a). Several research have been conducted to assess the accuracy and interpretability of multiple algorithms using standardized metrics. Reproducibility is another major challenge as much research lack transparency in preprocessing and validation requiring a well formulated containerized pipeline ensuring continuous analysis. Additionally, though computational methods assist in identifying relevant CpG sites, a very few of them systematically map genes and perform pathway enrichment to connect predictions to mechanisms. Although recent research aims at distinguishing casual modifications from correlational changes, wholesome mechanistic comprehension is still very limited. Technical challenges persist in handling batch effects, tissue heterogeneity, and computational demands of high-dimensional data, requiring sophisticated statistical approaches and efficient workflows.

The present research seeks to address these gaps by providing a comprehensive model amalgamating explainable results, integrated machine learning and biological pathway analysis. The foremost aim is to develop a vigorous epigenetic clock using elastic net regression combined

with SHAP-based validation. This study seeks to attain specific aims which include implementing elastic net regression with standardized preprocessing and cross-validation to build an accurate age prediction model while performing automatic feature selection, enhancing model interpretability via comprehensive SHAP analysis, generating visualizations including various plots to identify key age-associated CpG sites and quantify their individual contributions to age predictions. It also aims at ensuring reproducibility through systematic documentation of all preprocessing steps, normalization procedures, and model parameters. This will bridge statistical predictions to biological mechanisms by mapping CpG sites to genes using Illumina 450K annotation databases and performing pathway enrichment analysis using established resources including KEGG and Gene Ontology databases. By combining Elastic Net's feature selection capabilities with SHAP's model-agnostic interpretability and comprehensive pathway analysis, this work advances both methodological rigor and biological insight in epigenetic clock development.

This research makes significant contributions to the existing gamut of research on epigenetic clock through several relevant innovations. Operationally, it aims to systematically filling the gap in current literature by providing a comprehensive comparison of machine learning models with SHAP-based models which will ultimately assist in evaluating predictive accuracy and biological interpretability. This integration of SHAP analysis with pathway enrichment bridges assists in identifying age-associated CpG sites for experimental validations. It helps attain statistical modeling and mechanistic understanding. By integrating advanced machine learning models, methods of interpretability and reproducible workflows, this research aims at enhancing both methodological severity as well as biological insights needed in epigenetic clock development,

additionally contributing to the broader goal of increasing human health span by comprehending and potentially intervening with the aging process of humans.

METHODS

3.1 Data Acquisition and Preprocessing

The present research uses GSE87571 dataset which is publicly accessible from the Gene Expression Omnibus repository. Johansson et al. (2013) was the first one to publish and examine this dataset on the lines of continuous DNA methylation modifications taking place across human lifespan (Johansson et al., 2013). Blood samples of 729 people have been used in this project ranging from 14 to 94 years of age. Out of the total number of samples, 341 samples are that of males and the remaining 388 are that of females. These samples are profiled using the Illumina HumanMethylation450 BeadChip array which cross examines 485,512 CpG sites over CpG islands, gene promoters, intergenic regions and gene bodies. The entire computational analysis (data cluster preprocessing and feature selection) was conducted with the aid of Northeastern University's Discovery High Performance Computing (HPC) pursued by visualization and creating statistical model on a local workstation.

Unprocessed data files are downloaded using python 3.13.5 scripts from GEO executed through Slurm batch scheduling. Three primary files were retrieved: a soft file containing sample metadata, and two compressed beta value matrices. Both matrices accommodate identical CpG sites but supportive sample subsets with duplicate subset market by .1 suffix. Horizontal merging is done to remove duplicate columns from both matrices to develop a complete dataset. Furthermore, to

avoid computational memory limitations, a memory efficient chunking strategy is used to process 50,000 CpG sites per chunk.

SOFT file in assistance with GEOparse Python library was used to extract sample metadata including sex, age and GEO sample identifiers ID (GSM IDs). Originally marked with non-descriptive names of placeholders, sample identifiers in the beta matrices are systematically arranged to their corresponding GSM ID by cross-referencing sequential sample order contained in the SOFT file. This data was then amalgamated with metadata through inner join on sample identifiers. This process helped to attain a final dataset of 729 blood samples with absolute phenotypic information and methylation profiles.

3.2 Feature Selection and CpG Site Prioritization

Due to the high-dimensional nature of DNA methylation data, CpG sited associated with chronological age was selected via feature selection. Initially, 35,230 CpG sites with missing values were excluded from total 485,512 CpG sites, making only 450,282 sites available for correlation analysis. All 729 samples were subject to calculating Pearson correlation coefficients between every CpG site methylation levels and chronological age. An uncorrected p-value threshold of $p < 1\times10^{-5}$, identifying 190,240 CpG sites (42.2% of complete sites) is used to assess statistical importance showcasing relevant age associations. This approach is very suitable for development of epigenetic clock methods that prioritize effect size much higher than the genome-wide statistical relevance for predictive modeling (Hannum et al., 2013; Horvath, 2013b). Therefore, a total of 500 top CpG sites were ranked based on absolute correlation coefficient for further machine learning analysis.

3.3 Elastic Net Regression Model Development

Elastic Net regression was used to conduct age prediction modeling. It is a penalized linear method combining L1 (Lasso) and L2 (Ridge) regularization penalties (Zou & Hastie, 2005). The method was particularly selected due to its ability to perform the function of automatic feature selection while also maintaining model interpretability via exclusive and explicit coefficient values. It is to be noted that these properties are particularly significant and advantageous for creating an epigenetic clock (Shireby et al., 2020b). The entire analysis is conducted in R version 4.4.1 with aid of glmnet package. To enhance reproducibility, the data set was conveniently divided into training sets (80%, n=583) and test sets (20%, n=146) using a fixed random seed method.

To identify the optimal regularization strength (lambda), hyperparameter optimization is conducted through 10-fold cross-validation on the training set. It helped reduce mean squared error. A mixed parameter alpha was fixed to 0.5 indicating a balanced Elastic Net that weighs L1 and L2 penalties equally. Optimal lambda value of 0.1498 was identified through cross-validation which is further used to train a model on the complete training set. Finally, a set of 153 CpG sites are selected out of the total 500 features having non-zero coefficients (69.4% feature reduction). Model performance was evaluated on the independent test set using mean absolute error (MAE), root mean squared error (RMSE), coefficient of determination (R^2), and Pearson correlation coefficient between predicted and actual ages.

3.4 Model Interpretability Using SHAP Analysis

For validating feature relevance rankings and enhancing model interpretability, Shapley Additive explanations (SHAP) analysis is carried out on previously trained Elastic Net model. SHAP is based on cooperative game theory ensuring a uniform procedure for quantifying contribution of

individual feature on predictions (Lundberg & Lee, 2017). For linear models such as Elastic Net, SHAP values can be analytically computed as the product of each feature's coefficient and its deviation from the training set mean: SHAP = coefficient × (feature_value - mean_feature_value). This formula helps assess both global feature significance as well as explanation of individual predictions.

All 146 test set samples across the 153 selected CpG sites are subject to calculations of SHAP values. CpG importance is ranked based on the Mean absolute SHAP values. It provides alternative feature significance metric. To validate consistency between approaches, correlation between absolute coefficient values and mean absolute SHAP values are assessed. Several visualizations are generated including summary beeswarm plots showcasing SHAP value distributions for the top 20 CpG sites colored by methylation level, dependence plots are generated to examine and assess the interlinking between methylation and SHAP contribution and force plots are generated to understand for different age groups showing how individual CpG sites contribute in specific age-related predictions.

3.5 Biological Annotation and Pathway Enrichment Analysis

A comprehensive genomic annotation is conducted with the aid of IlluminaHumanMethylation450kanno.ilmn12.hg19 R package to draw relation of predictive CpG sites with the biological mechanism of aging. This process provides uniform and standard annotations based on the hg19 human genome assembly. Each of 153 CpG sites was evaluated to its closest gene(s), CpG island context (island, shore, shelf, or open sea), genomic coordinates and gene region (promoter, gene body, or intergenic). Finally, it recognized 108 unique genes in relation to age predictive CpG sites.

To comprehend if these genes are preferentially involved in biological process of aging, a pathway enrichment analysis is conducted. Gene lists are submitted to a web-based tool named Enrichr for gene set enrichment analysis (Xie et al., 2021). Enrichment is assessed for KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways and Gene Ontology (GO) Biological Process terms. Furthermore, Enrichr employs a Fisher's exact test coupled with Benjamini-Hochberg false discovery rate correction to understand and evaluate statistical importance. These results are exported and visualized in R with help of bar plots representing $\log_{10}$(p-value) to recognize biological processes most nearly connected with age related predictive methylation modifications.

RESULTS

4.1 Feature Selection and Correlation Analysis

Correlation analysis uncovered widespread age-associated DNA methylation across the blood samples. 190,240 CpG sites making 42.2% of the 450,282 CpG sites demonstrated significant age correlations ($p < 1\times10^{-5}$). The top 500 age-associated CpG sites exhibited exceptionally strong linear relationships with chronological age, with correlation coefficients ranging from $|r| = 0.71$ to 0.95 (all $p < 1\times10^{-300}$). The most age-correlated CpG site was cg16867657 ($r = 0.9464$, $p < 1\times10^{-300}$) shown in figure1. This figure marks strong positive correlation between cg16867657 methylation and chronological age, with no substantial sex differences observed between females ($n = 388$) and males ($n = 341$). This CpG site is located on ELOVL2 gene which is discovered as a premier aging biomarker (C.Zapico & Ubelaker, 2013; Garagnani et al., 2012).

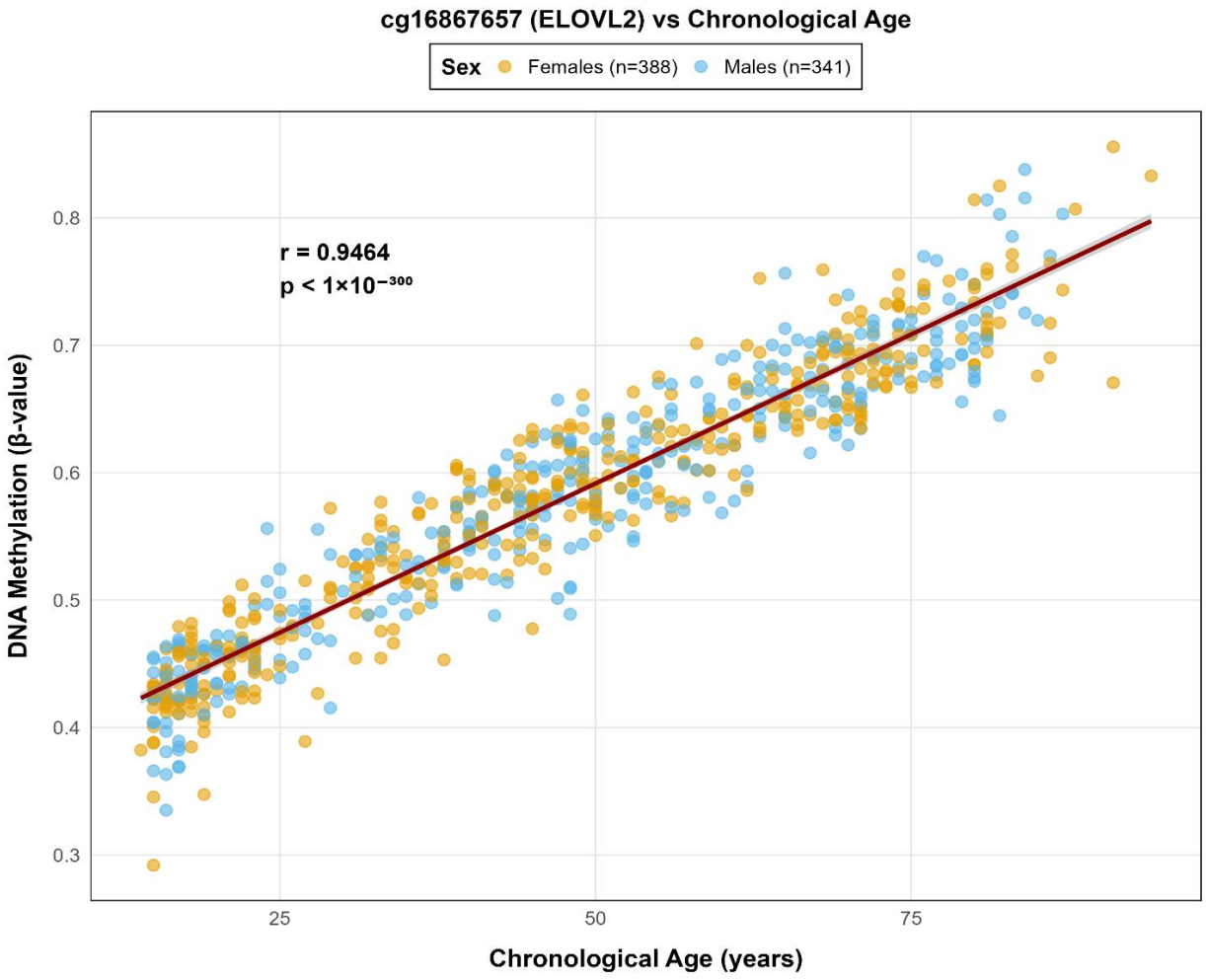

Figure 1. *cg16867657 showing methylation versus chronological age with females (orange) and males (blue). Irrespective of sex, similar age-related hypermethylation (r = 0.9464) is seen across ages.*

As shown in Figure 2, different, coordinated methylation patterns were found by hierarchically grouping the top 20 age-associated CpG sites. Twelve of these locations exhibit age-related hypomethylation, whereas eight clearly indicate a tendency of progressive hypermethylation with increasing age. The correlation coefficients show high connections in both directions, ranging from r = –0.8559 for cg11649376 to r = +0.9464 for cg16867657. Consistent age-related gradients from younger samples on the left to older samples on the right are further highlighted by the heatmap.

Interestingly, the annotation bars' low sex-based clustering indicates that these methylation changes are mostly unrelated to biological sex.

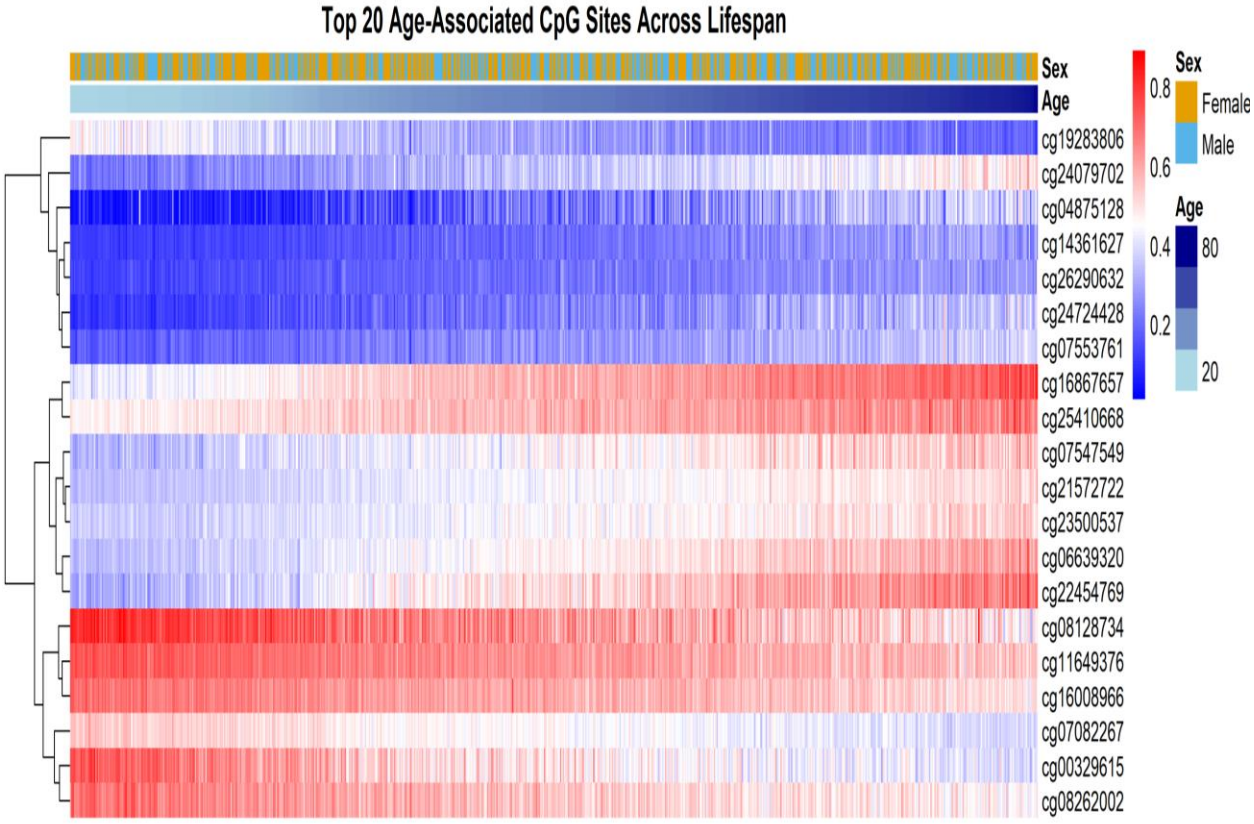

Figure 2. *Heatmap of top 20 CpG sites showing highest correlation with age across 729 samples.*

(hypomethylation in blue & hypermethylation in red)

4.2 Elastic Net Model Development and Performance

A total of 153 CpG sites were selected from the top 500 CpGs using Elastic Net regression, representing a 69.4% reduction in features. Optimal hyperparameters were determined through cross-validation, with $\alpha = 0.5$ and $\lambda = 0.1498$, achieving a root mean squared error (RMSE) of 3.49 years. The resulting model demonstrated excellent predictive performance on the held-out test set ($n = 146$). The mean absolute error of 2.55 years substantially outperformed published epigenetic

clocks: Horvath (MAE = 3.6 years) (Horvath, 2013b) and Hannum (MAE = 4.9 years) (Hannum et al., 2013). The model explained 97.56% of age variance (R^2 = 0.9756), with a Pearson correlation of r = 0.9877 ($p < 2.2\times10^{-16}$) between predicted and actual age. Figure 3 illustrates strong consistency between predicted and actual age across the age range. The age predictions are clustered tightly around the line of perfect prediction, showing model performed consistently well across all age groups.

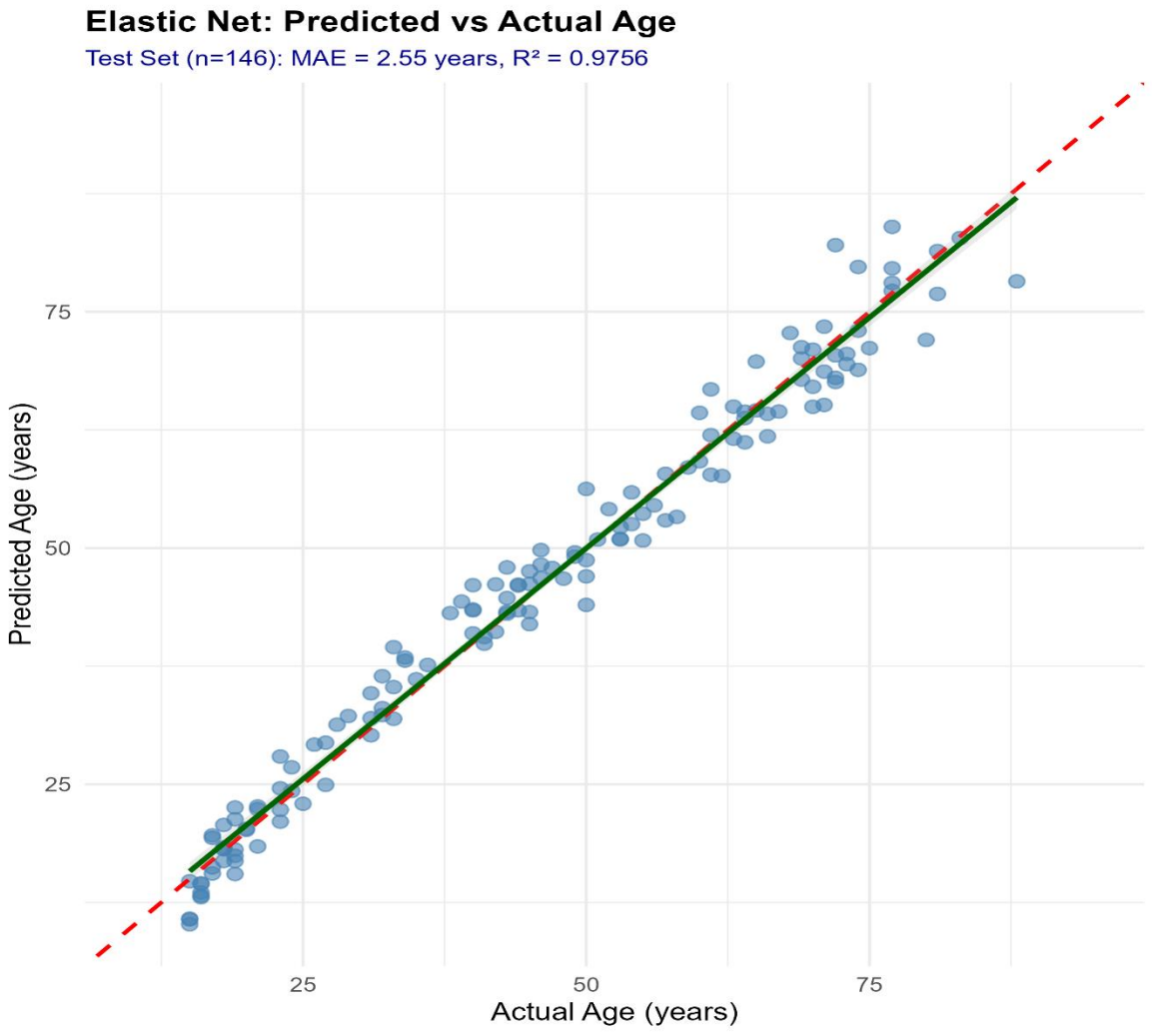

Figure 3. *Predicted age versus actual age in test set (n = 146). Blue points depict individual predictions; red dashed line is for perfect prediction. (R^2 = 0.9756, MAE = 2.55 years)*

4.3 Feature Importance and SHAP Validation

The Elastic Net model retained 153 CpG sites with coefficients ranging from -15.98 to +28.15. Out of 153, 81 sites are hypermethylated and 72 CpGs exhibit hypomethylation, demonstrating balanced methylation patterns. The top 20 CpG sites are sorted by absolute coefficient magnitude in Figure 4. The strongest age predictor was cg16867657 (β = +28.15), where an increase of 2.8 years in expected age is associated with a 0.1-unit increase in methylation. Cg03607117 (β = +22.27), cg00292135 (β = +18.01), and cg14361627 (β = +17.97) were additional noteworthy hypermethylation sites. The most notable hypomethylation sites were cg02395812 (β = –15.98), cg12580096 (β = –15.80), and cg19722847 (β = –14.94).

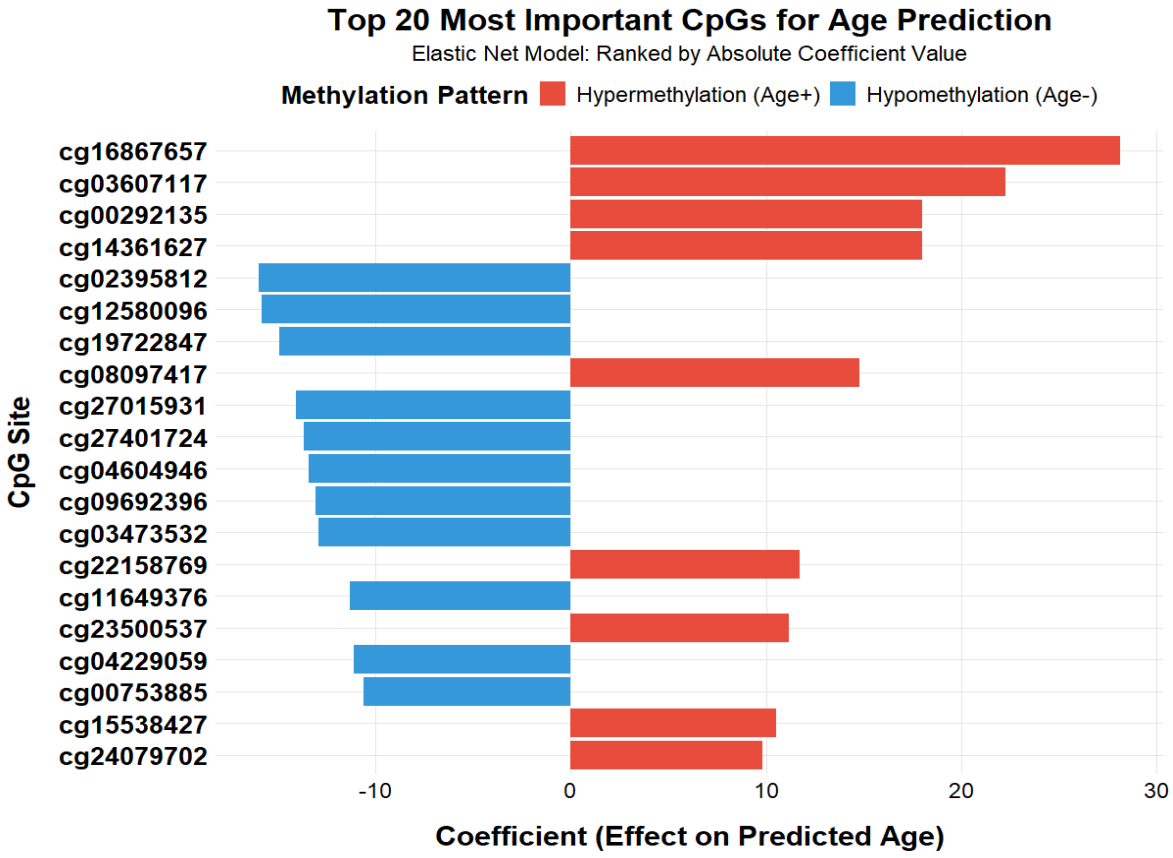

Figure 4. *Top 20 CpG sites displayed with absolute coefficient. Red stands for hypermethylation and blue marks hypomethylation.*

These feature importance rankings were independently confirmed by SHAP analysis, and there was a substantial correlation ($r = 0.84$, $p < 2.2\times10^{-1}$) between absolute coefficients and mean absolute SHAP values. The most significant predictor, according to both methods, was cg16867657 (ELOVL2) (mean |SHAP| = 2.49 years). Age prediction and cg16867657 methylation showed perfect linearity (R2 = 1.0, slope = 28.15 years) according to SHAP dependency analysis (Figure 5 in appendix). The slope precisely matched the Elastic Net coefficient ($\beta = +28.15$), confirming that the link was biological rather than overfitting and supporting the linearity assumption of the model.

4.4 Gene Annotation and Mapping

The final model had 153 CpG sites, of which 116 (75.8%) matched to 108 distinct genes and 37 (24.2%) were found in intergenic regions. Multiple age-associated CpGs were found in seven genes, including KLF14 (3 CpGs) and ABCC4, ASPA, FHL2, LRRC23, ZAR1, and ZYG11A (2 CpGs each). Preferential methylation alterations in regulatory regions were suggested by the analysis of genomic context, which revealed that 63 sites (41.2%) were in CpG islands, 40 (26.1%) in CpG island coasts, 12 (7.8%) in shelves, and 38 (24.8%) in open sea regions (Figure 6 in appendix). The distribution of chromosomes was wide, with chromosomes 1, 7, and 2 having the largest representation. In terms of gene regions, effects on both promoters and gene bodies were shown by the presence of 97 CpGs in gene bodies, 46 in TSS1500, 25 in 5'UTR, and 19 in TSS200 areas. Important genes included well-known aging indicators like FHL2, KLF14, and ELOVL2. Intergenic CpGs imply that epigenetic aging may be influenced by other regulatory factors.

4.5 Pathway Enrichment Analysis

The 108 previously discovered genes were examined using Enrichr for Gene Ontology (GO) Biological Process and KEGG pathway enrichment (adjusted $p < 0.05$) in order to investigate the biological functions and pathways linked to age-related DNA methylation alterations. Significant enrichment for physiological and developmental processes was found by GO analysis. The most enriched phases was "central nervous system development," which was followed by "brain development," "enteroendocrine cell differentiation," and "cellular hypotonic response" as demonstrated in figure 7. The findings of GO confirms that alterations in methylation patterns in the neurodevelopmental genes plays a role in epigenetic aging.

Five enriched pathways were found using KEGG pathway analysis (Figure 7). The strongest enrichment was found in hippo signaling, which regulates organ size and cell proliferation. This was followed by cellular aging, fluid shear stress and atherosclerosis, maturity-onset diabetes of the young, and calcium signaling. Notably, pro-inflammatory signaling and cell cycle arrest are two important signs of aging that are highlighted by the enrichment of the cellular senescence pathway. A number of established aging indicators, such as ELOVL2, KLF14, FOXG1, TP73, and HDAC4, as well as genes linked to neuronal, metabolic, and oxidative stress-related aging, were confirmed by cross-referencing the 108-gene set with aging literature. Together, these findings validate our feature selection and demonstrate that the model captures biologically meaningful age-associated epigenetic changes rather than random correlations.

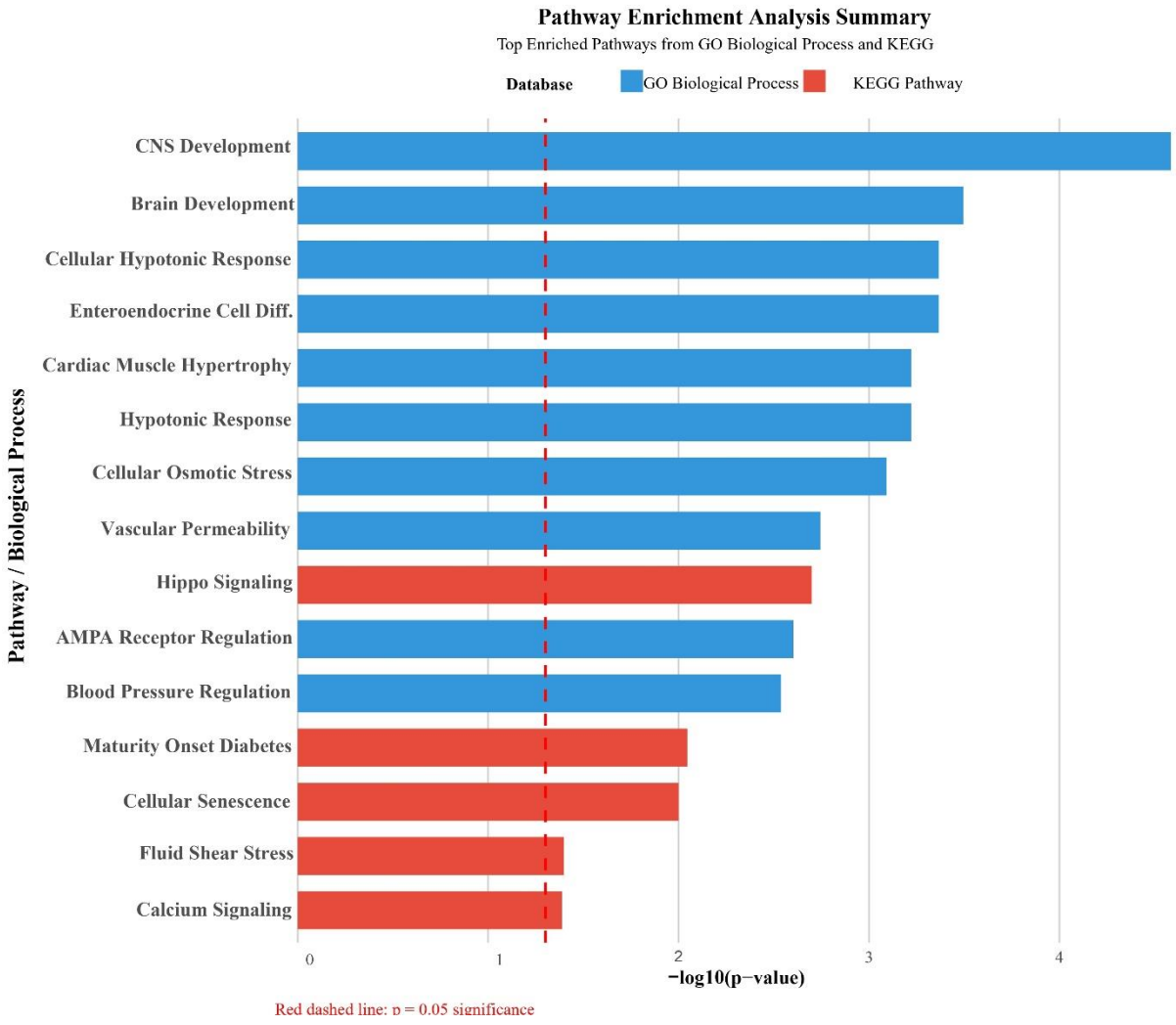

Figure 7. *Pathway enrichment summary for 108 age-associated genes. Blue bars represent GO Biological Process; red bars represent KEGG pathways. Red dashed line is the p (0.05) significance threshold.*

DISCUSSION

This study analysed blood-based DNA methylation data (729 samples) to develop an accurate and transparent epigenetic clock by using elastic net combined with SHAP. The machine learning model achieved MAE of 2.55 years, showing substantial improvement in comparison to Horvath (MAE = 3.6 years) and Hannum (MAE = 4.9 years) clocks. These findings demonstrate that careful

methylation feature selection in elastic net with proper SHAP validation, produces both highly accurate and biologically interpretable age predictors.

This degree of accuracy is from three methodological factors. First, considerable sample size of 729 covering wide age range (14-94 years) provided robust training data for capturing age-related methylation patterns (Teschendorff & Horvath, 2025b). Second, feature selection favored locations with greatest correlations, enriching for functionally important CpGs. The top-ranked site, cg16867657 in ELOVL2, has been regularly verified as a prime aging biomarker across multiple populations and tissues (El-Shishtawy et al., 2024). According to recent research, ELOVL2 methylation may have significant functions in aging processes rather than only being correlative, especially when it comes to controlling the synthesis of extremely long-chain polyunsaturated fatty acids(El-Shishtawy et al., 2024). Third, a sparse model of 153 CpGs avoided overfitting while preserving accuracy was produced by Elastic Net's balanced L1/L2 regularization.

SHAP analysis validated strong correlation ($r = 0.84$) between its mean absolute values and Elastic Net coefficients. This confirmed consistency between traditional and game-theory-based features, emphasizing model interpretability and biological plausibility. The cg16867657 linear relationship similarity in SHAP dependence plots ($R^2 = 1.0$) and in elastic net ($R^2 = 0.9464$) validates that the model captured genuine biological signals rather than overfitting to noise. This revealed that for blood-based age prediction using properly selected characteristics, linear connections dominate, indicating Elastic Net's applicability for this application. Therefore, SHAP fills a major gap in epigenetic clock research, where more complicated models compromise interpretability for minor accuracy gains.

Significant correlations with cellular senescence ($p = 0.010$), Hippo signaling ($p = 0.002$), and central nervous system development ($p = 2.6\times10^{-5}$) were revealed by pathway enrichment analysis. The developmental programming theory of aging, which postulates that age-related methylation patterns reflect the ongoing influence of early-life gene regulatory programs, is supported by the prevalence of neurodevelopmental pathways (Teschendorff & Horvath, 2025b). The enrichment of cellular senescence pathways is especially noteworthy, as senescence is a fundamental aging hallmark including permanent cell cycle arrest and pro-inflammatory secretions (Ajoolabady et al., 2025). Increased senescent cell load is linked to tissue deterioration, chronic inflammation, and age-related diseases such neurological and cardiovascular conditions, according to mounting data. Senescence-associated genes are included in predicted CpGs, indicating that biological aging processes are captured by our approach. Moreover, enrichment of the Hippo signaling pathway, which regulates organ growth and tissue homeostasis, raises the possibility that age-related methylation changes are a result of a decline in regeneration potential and an increase in the accumulation of senescent cells. Rather than random drift, planned epigenetic remodeling is implied by the almost equal proportions of hypermethylated and hypomethylated CpGs.

Given individual variability in aging, our model approaches the theoretical limits for blood-based clocks in comparison to earlier research. When training data is sufficient and features are appropriately selected, Elastic Net performance approximates that of more sophisticated models, while giving direct quantification of CpG contributions. These findings underscore the usefulness of interpretable models in translational and mechanistic studies.

Despite these positives, certain restrictions warrant examination. Although significant, the sample size is still less than that of mega-cohort studies (>10,000 samples), which may improve generalizability and accuracy (Teschendorff & Horvath, 2025b). Only blood tissue was analyzed;

tissue-specific clocks could enhance forecasts for target organs. Since only a portion of age-associated CpGs may have a direct impact on aging, the cross-sectional design prevents causal conclusion (Ying et al., 2024). Additionally, the age range may not effectively represent pediatric or centenarian groups, which could benefit from specialized clocks. The necessity for rigorous experimental design and normalization is highlighted by technical factors, such as the 450K array's low coverage (~2% of CpGs), age-related changes in blood cell composition, and possible batch effects (Teschendorff & Horvath, 2025b).

The designed clock has various potential applications. Age acceleration has been connected to higher mortality, cardiovascular disease, dementia, and cancer (Teschendorff & Horvath, 2025b). Its 2.55-year precision allows for the detection of accelerated or decelerated aging in longitudinal research. The 153 CpGs found represent possible biomarkers for monitoring biological aging and responsiveness to treatments. SHAP-based interpretability helps mechanistic research by quantifying individual CpG contributions, directing experimental validation and potential therapeutic targeting.

Future research should address critical directions. Generalizability will be ensured by validation across separate cohorts from various racial and geographic origins. Second-generation clocks with improved predictive and mechanistic capacity may be produced by integrating multi-omics data, such as transcriptomics and proteomics (Teschendorff & Horvath, 2025b). Longitudinal investigations are needed to directly evaluate aging rates, and single-cell methylation methods can resolve cellular heterogeneity, differentiating intrinsic aging from compositional changes. Targeted epigenetic editing could be used to experimentally validate potential CpGs and determine their causal roles in aging symptoms (Ying et al., 2024).

In nutshell, development of accurate and interpretable epigenetic clocks can be achieved by combining linear regression model with SHAP validation. The integration of machine learning models with biological pathway analysis and interpretable models enhances both the methodological strength of the study and our insight into the underlying biological mechanisms and biomarker studies.

CONCLUSION

This study developed an accurate and interpretable epigenetic clock achieving MAE = 2.55 years, outperforming Horvath (3.6 years) and Hannum (4.9 years) clocks. Elastic Net regression identified 153 age-predictive CpG sites with cg16867657 in *ELOVL2* demonstrating perfect SHAP linearity ($R^2 = 1.0$). Pathway enrichment revealed associations with nervous system development, cellular senescence, and Hippo signaling, connecting predictions to fundamental aging mechanisms.

This work demonstrates that combining Elastic Net with SHAP-based explainable AI addresses the interpretability-accuracy trade-off in clock development. The perfect SHAP validation confirms high accuracy without sacrificing transparency, essential for clinical translation. Integration of pathway enrichment bridges statistical modeling and mechanistic understanding. The reproducible framework advances both technical rigor and biological insight in epigenetic aging research. Interpretability plays a critical role in epigenetic clocks transition to clinical biomarkers. The SHAP-validated approach enables identification of therapeutic targets while maintaining explainability. The future improvements suggest requirement of longitudinal measurements, multi-omics integration, or cell-type-specific analyses to break 2.55-year accuracy theoretical limits.

Future research should pursue independent validation across diverse populations, longitudinal tracking of aging rate, multi-omics integration, experimental validation through epigenetic editing, and single-cell clock development. By maintaining emphasis on interpretability and biological validation, the field can develop clocks serving as both accurate biomarkers and tools for understanding fundamental mechanisms of human aging.

REFERENCES

Ajoolabady, A., Pratico, D., Bahijri, S., Eldakhakhny, B., Tuomilehto, J., Wu, F., & Ren, J. (2025). Hallmarks and mechanisms of cellular senescence in aging and disease. In *Cell Death Discovery* (Vol. 11). Springer Nature. https://doi.org/10.1038/s41420-025-02655-x

An, Y., Wang, Q., Gao, K., Zhang, C., Ouyang, Y., Li, R., Ma, Z., Wu, T., Zhou, L., Xie, Z., Zhang, R., & Wu, G. (2025). Epigenetic Regulation of Aging and its Rejuvenation. In *MedComm* (Vol. 6). John Wiley and Sons Inc. https://doi.org/10.1002/mco2.70369

Belsky, D. W., Caspi, A., Corcoran, D. L., Sugden, K., Poulton, R., Arseneault, L., Baccarelli, A., Chamarti, K., Gao, X., Hannon, E., Harrington, H. L., Houts, R., Kothari, M., Kwon, D., Mill, J., Schwartz, J., Vokonas, P., Wang, C., Williams, B. S., & Moffitt, T. E. (2022). DunedinPACE, a DNA methylation biomarker of the pace of aging. *ELife*, *11*. https://doi.org/10.7554/eLife.73420

Borrego-Ruiz, A., & Borrego, J. J. (2024). Epigenetic Mechanisms in Aging: Extrinsic Factors and Gut Microbiome. In *Genes* (Vol. 15). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/genes15121599

C.Zapico, S., & Ubelaker, D. H. (2013). Applications of physiological bases of ageing to forensic sciences. Estimation of age-at-death. In *Ageing Research Reviews* (Vol. 12, pp. 605–617). https://doi.org/10.1016/j.arr.2013.02.002

de Lima Camillo, L. P., Lapierre, L. R., & Singh, R. (2022). A pan-tissue DNA-methylation epigenetic clock based on deep learning. *Npj Aging*, *8*. https://doi.org/10.1038/s41514-022-00085-y

El-Shishtawy, N. M., El Marzouky, F. M., & El-Hagrasy, H. A. (2024). DNA methylation of ELOVL2 gene as an epigenetic marker of age among Egyptian population. *Egyptian Journal of Medical Human Genetics*, *25*. https://doi.org/10.1186/s43042-024-00477-7

Garagnani, P., Bacalini, M. G., Pirazzini, C., Gori, D., Giuliani, C., Mari, D., Di Blasio, A. M., Gentilini, D., Vitale, G., Collino, S., Rezzi, S., Castellani, G., Capri, M., Salvioli, S., & Franceschi, C. (2012). Methylation of ELOVL2 gene as a new epigenetic marker of age. *Aging Cell*, *11*, 1132–1134. https://doi.org/10.1111/acel.12005

Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S. V., Klotzle, B., Bibikova, M., Fan, J. B., Gao, Y., Deconde, R., Chen, M., Rajapakse, I., Friend, S., Ideker, T., & Zhang, K. (2013). Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Molecular Cell*, *49*, 359–367. https://doi.org/10.1016/j.molcel.2012.10.016

Horvath, S. (2013a). DNA methylation age of human tissues and cell types. *Genome Biology*, *14*. https://doi.org/10.1186/gb-2013-14-10-r115

Horvath, S. (2013b). DNA methylation age of human tissues and cell types. *Genome Biology*, *14*. https://doi.org/10.1186/gb-2013-14-10-r115

Johansson, Å., Enroth, S., & Gyllensten, U. (2013). Continuous Aging of the Human DNA Methylome Throughout the Human Lifespan. *PLoS ONE*, *8*. https://doi.org/10.1371/journal.pone.0067378

López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M., & Kroemer, G. (2023). Hallmarks of aging: An expanding universe. In *Cell* (Vol. 186, pp. 243–278). Elsevier B.V. https://doi.org/10.1016/j.cell.2022.11.001

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *2017-December*, 4766–4775.

Mc Auley, M. T., & Morgan, A. E. (2025). Investigating Aging and DNA Methylation: A Path to Improving Health Span? In *Yale Journal of Biology and Medicine* (Vol. 98, pp. 237–244). Yale Journal of Biology and Medicine Inc. https://doi.org/10.59249/BYOI5042

McCrory, C., Fiorito, G., Hernandez, B., Polidoro, S., O'Halloran, A. M., Hever, A., Ni Cheallaigh, C., Lu, A. T., Horvath, S., Vineis, P., & Kenny, R. A. (2021). GrimAge Outperforms Other Epigenetic Clocks in the Prediction of Age-Related Clinical Phenotypes and All-Cause Mortality. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences*, *76*, 741–749. https://doi.org/10.1093/gerona/glaa286

Ponce-Bobadilla, A. V., Schmitt, V., Maier, C. S., Mensing, S., & Stodtmann, S. (2024). Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. *Clinical and Translational Science*, *17*. https://doi.org/10.1111/cts.70056

Prosz, A., Pipek, O., Börcsök, J., Palla, G., Szallasi, Z., Spisak, S., & Csabai, I. (2024). Biologically informed deep learning for explainable epigenetic clocks. *Scientific Reports*, *14*. https://doi.org/10.1038/s41598-023-50495-5

Seale, K., Teschendorff, A., Reiner, A. P., Voisin, S., & Eynon, N. (2024). A comprehensive map of the aging blood methylome in humans. *Genome Biology*, *25*. https://doi.org/10.1186/s13059-024-03381-w

Shireby, G. L., Davies, J. P., Francis, P. T., Burrage, J., Walker, E. M., Neilson, G. W. A., Dahir, A., Thomas, A. J., Love, S., Smith, R. G., Lunnon, K., Kumari, M., Schalkwyk, L. C., Morgan, K., Brookes, K., Hannon, E., & Mill, J. (2020a). Recalibrating the epigenetic clock: Implications for assessing biological age in the human cortex. *Brain*, *143*, 3763–3775. https://doi.org/10.1093/brain/awaa334

Shireby, G. L., Davies, J. P., Francis, P. T., Burrage, J., Walker, E. M., Neilson, G. W. A., Dahir, A., Thomas, A. J., Love, S., Smith, R. G., Lunnon, K., Kumari, M., Schalkwyk, L. C., Morgan, K., Brookes, K., Hannon, E., & Mill, J. (2020b). Recalibrating the epigenetic clock: implications for assessing biological age in the human cortex. *Brain*, *143*(12), 3763–3775. https://doi.org/10.1093/brain/awaa334

Teschendorff, A. E., & Horvath, S. (2025a). Epigenetic ageing clocks: statistical methods and emerging computational challenges. In *Nature Reviews Genetics* (Vol. 26, pp. 350–368). Nature Research. https://doi.org/10.1038/s41576-024-00807-w

Teschendorff, A. E., & Horvath, S. (2025b). Epigenetic ageing clocks: statistical methods and emerging computational challenges. *Nature Reviews Genetics*, *26*(5), 350–368. https://doi.org/10.1038/s41576-024-00807-w

Vershinina, O., Bacalini, M. G., Zaikin, A., Franceschi, C., & Ivanchenko, M. (2021). Disentangling age-dependent DNA methylation: deterministic, stochastic, and nonlinear. *Scientific Reports*, *11*. https://doi.org/10.1038/s41598-021-88504-0

Xie, Z., Bailey, A., Kuleshov, M. V., Clarke, D. J. B., Evangelista, J. E., Jenkins, S. L., Lachmann, A., Wojciechowicz, M. L., Kropiwnicki, E., Jagodnik, K. M., Jeon, M., & Ma'ayan, A. (2021). Gene Set Knowledge Discovery with Enrichr. *Current Protocols*, *1*. https://doi.org/10.1002/cpz1.90

Ying, K., Liu, H., Tarkhov, A. E., Sadler, M. C., Lu, A. T., Moqri, M., Horvath, S., Kutalik, Z., Shen, X., & Gladyshev, V. N. (2024). Causality-enriched epigenetic age uncouples damage and adaptation. *Nature Aging*, *4*, 231–246. https://doi.org/10.1038/s43587-023-00557-0

Yusipov, I., Bacalini, M. G., Kalyakulina, A., Krivonosov, M., Pirazzini, C., Gensous, N., Ravaioli, F., Milazzo, M., Giuliani, C., Vedunova, M., Fiorito, G., Gagliardi, A., Polidoro, S., Garagnani, P., Ivanchenko, M., & Franceschi, C. (2020). Age-related DNA methylation changes are sex-specific: a comprehensive assessment. *Aging*, *12*, 24057–24080. https://doi.org/10.18632/aging.202251

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *67*, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

APPENDIX

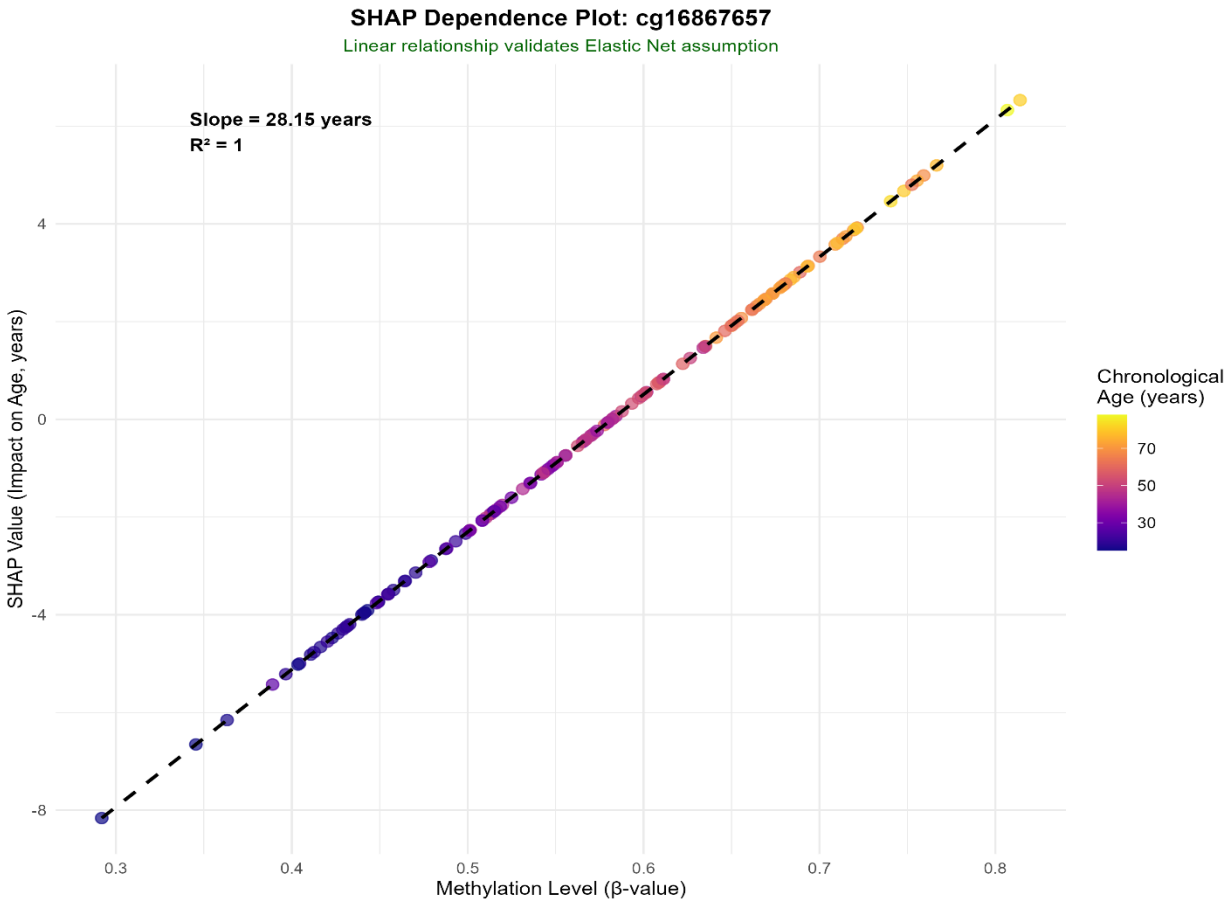

Figure 5. *SHAP dependence plot for cg16867657 showing perfect linear relationship (R^2 = 1.0, slope = 28.15 years).*

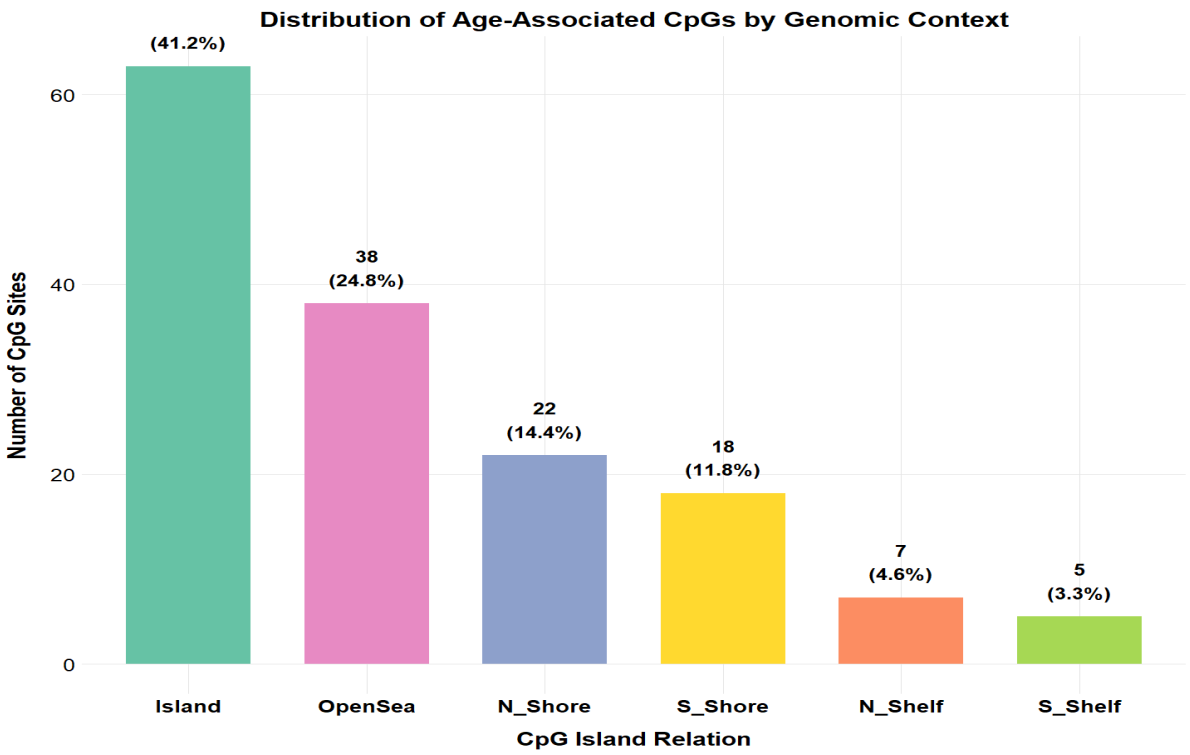

Figure 6. *Bar graph showing 153 CpGs distribution across the genome.*