Explaining Biological Age Discrepancies through DNA Methylation

Klára Hirmanová

Responsible Professor: Prof.dr.ir. Marcel Reinders

Supervisors: Bram Pronk, Inez den Hond, Gerard Bouland

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 21, 2025

Name of the student: Klára Hirmanová
Final project course: CSE3000 Research Project
Thesis committee: Prof.dr.ir. Marcel Reinders, Bram Pronk, Inez den Hond, Gerard Bouland, Dr. Kaitai Liang

Abstract

Biological age, as estimated by epigenetic clocks, can differ significantly from an individual's chronological age. These discrepancies, referred to as positive and negative age acceleration, may reflect underlying biological variation in the healthy human aging processes. This thesis investigates which DNA methylation features drive such differences in healthy individuals.

To explore this, two established aging clocks were replicated, and the residuals between predicted and chronological age were used to label samples as positively or negatively accelerated. A classification model was developed to distinguish between these groups, with XGBoost achieving the best performance. Model interpretation using SHAP values identified key CpG sites associated with acceleration. Gene Ontology analysis of these features revealed enrichment in biological processes such as immune response, DNA repair, neurodevelopment, and oxidative stress. Feature importance was also analyzed across age groups, revealing age-sensitive patterns in methylation influence.

This work provides a reproducible pipeline for interpreting biological age discrepancies and contributes to a deeper understanding of the biological features associated with different aging trajectories.

1 Introduction

Aging is a biological process characterized by gradual functional deterioration that ultimately compromises an organism's survival. It is a universal phenomenon observable throughout the lifetime of every organism. Considerable research has been devoted to understanding how aging might be slowed, halted, or even reversed. This has raised the question of how organisms evolved such intricate biological systems, yet appear to fail at the seemingly simpler task of self-maintenance. Extensive studies have focused on age-related diseases, such as dementia, which express across different tissues and often appear unrelated. While some individuals may develop one condition and others another, it is generally expected that certain tissues in each individual deteriorate more rapidly than others. Although these age-related pathologies may not seem directly connected, the underlying biological mechanisms at the cellular level may provide a unifying explanation [1].

Several key hallmarks of aging have been identified, such as genome instability, cellular senescence, and epigenetic alterations, which are grouped into three functional categories: primary hallmarks, which are the initiating causes of cellular damage; antagonistic hallmarks, which are responses to this damage and can be beneficial in moderation but harmful when dysregulated; and integrative hallmarks, which emerge when the damage exceeds the cell's compensatory capacity and lead to functional decline [2]. Among these, epigenetic alterations are classified as a primary hallmark of aging. This term reffers to widespread and often reversible modifications that affect gene expression without altering the underlying DNA sequence. These include changes in DNA methylation, histone modifications, chromatin remodeling, and non-coding RNA regulation. In particular, DNA methylation, which involves the addition of a methyl group to cytosine residues in CpG dinucleotides, plays a crucial role in regulating gene expression, maintaining genomic stability, and determining cell identity. DNA methylation levels at specific CpG sites are quantified using *beta values*, which range from 0 (completely unmethylated) to 1 (fully methylated). A beta value represents the proportion of methylated signal over the total signal, providing a normalised and interpretable measure of methylation intensity at each CpG site. Age-associated shifts in methylation patterns can disrupt gene regulation and contribute to cellular dysfunction and aging phenotypes [3].

A concept in the field of aging is the distinction between chronological and biological age. Chronological age simply states the number of years an organism has lived, whereas biological age can be calculated from various mathematical models based on the underlying biology of the organism. Biological age is often a better predictor of overall health, functional status, and susceptibility to disease [4]. A discrepancy between an individual's predicted biological age and their chronological age is referred to as *age acceleration*. Positive age acceleration, where biological age is larger than chronological age, is often associated with an increased risk age-related diseases or cancer [5], while negative age acceleration may indicate healthier aging pathways, typical trend within (super)centenarians [6].

One prominent approach to estimating biological age involves the use of aging clocks, which are statistical or machine learning models trained on age-related biomarkers to predict an individual's biological age. Among the most widely studied are epigenetic aging clocks, which are trained on DNA methylation data to make these predictions. Two influential models have been developed. The Horvath2013 clock uses 353 CpG sites and relies on Elastic Net regression, enabling age prediction from multiple tissues [5]. On the other hand, AltumAge, a recently developed deep learning model, uses over 21,000 CpGs and captures complex non-linear interactions [7].

While epigenetic clocks like Horvath2013 and AltumAge achieve high predictive accuracy, relatively little is known about what drives their errors and why two individuals of the same chronological age might receive different biological age estimates. Existing studies have largely focused on enhancing predictive performance or applying these models to disease contexts such as cancer. In contrast, the mechanisms underlying positive and negative biological age acceleration in healthy individuals remain underexplored. This project addresses that gap by first replicating the predictions of these clocks to validate the models performance. We then focus on the residuals, the differences between predicted and chronological age, and form a binary classification task of positive and negative age acceleration. Using feature selection and importance techniques, we identify the CpG sites most predictive of these residuals.

This Research Project aims to reproduce state-of-the-art epigenetic aging clocks and based on their predictions under-

stand features, which drive the residual differences between chronological and biological age. To ensure that the results reflect regular aging processes rather than disease-related effects, this project focuses on healthy tissue samples. By identifying the CpG features associated with residuals in biological age in the absence of a disease or other pathology, this work contributes new insight into the variability of aging within the healthy population.

To achieve the primary objective, this work will reproduce performance metrics of published epigenetic age prediction models. Given these predictions, a classification task is defined to differentiate between positive and negative residuals. Feature selection techniques are applied to reduce dimensionality and to help prevent overfitting of the classifiers. Multiple classification models are evaluated using nested cross-validation. On the best-performing classifier, SHAP analysis is used to identify the most influential features contributing the most to the positively and negatively accelerated predictions. Finally, we examine how the importance of specific features changes through the lifespan by comparing results across distinct chronological age groups.

2 Methodology and Experimental Setup

To analyse age discrepancies through DNA methylation a pipeline was developed. Starting with the acquisition of DNA methylation data, we replicate selected epigenetic clocks, calculate residuals to classify positive or negative age acceleration, and train machine learning models to be able to classify these patterns. Finally, we use feature selection and importance methods to identify CpG sites that contribute most to these biological age discrepancies and compare between chronological age groups. An overview of the complete pipeline is shown in Figure 1.

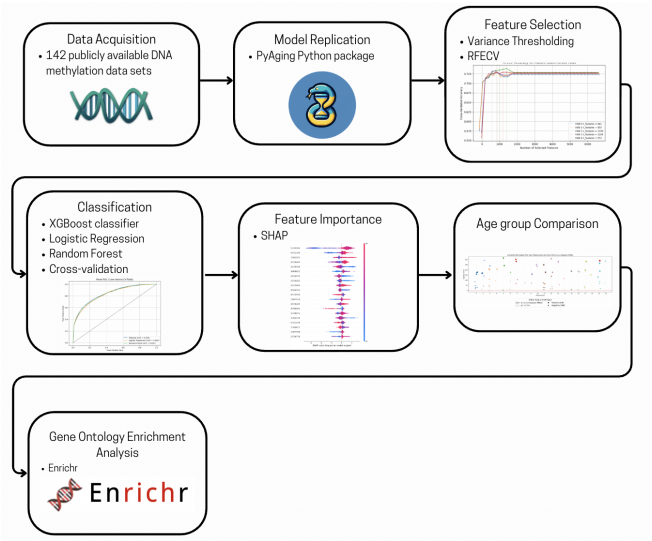

Figure 1: High-level overview of the pipeline used to identify and interpret age-associated CpG features from DNA methylation data.

2.1 Experimental Setup and Data

This study uses human DNA methylation data from healthy tissue samples, originally sourced from 142 publicly available datasets across the Gene Expression Omnibus (GEO), ArrayExpress, and The Cancer Genome Atlas (TCGA). These datasets were compiled and preprocessed by the authors of the AltumAge model [7], who applied preprocessing steps and imputed missing values using KNN imputation. Only samples with complete chronological age and valid beta values were kept.

Each sample contains methylation beta values at 21,368 CpG sites, derived from either the Illumina HumanMethylation27 or HumanMethylation450 platforms. Tumor samples present in the original dataset were excluded from this study to focus on normal aging processes. No missing values are present in the final dataset used, due to preprocessing by the AltumAge authors.

The analysis pipeline was implemented in Python 3.9. Major packages used include `pandas` (v2.2.3), `numpy` (v2.0.2), `scikit-learn` (v1.6.1), `xgboost` (v2.1.4), `goatools` (v1.4.12) and `pyaging` (v0.1.22) for clock replication.

2.2 Model Selection and Replication

In this project we replicated two epigenetic aging clocks. Horvath2013, a linear model trained on 353 CpG sites using Elastic Net regression, and AltumAge, a deep learning model based on over 21,000 CpGs. Both models were implemented using the `pyaging` Python package.

For Horvath2013, only the 353 CpGs used in the original model were extracted from the full dataset and matched by the CpG ID. For AltumAge, all 21,368 CpGs were retained. Both models were applied to the preprocessed beta values provided by the AltumAge authors. No additional preprocessing was performed, as the dataset was already fully processed.

To evaluate how well each model was replicated, we compared the published performance metrics to those obtained from our implementation on the same datasets. For each of the four metrics—mean absolute error (MAE), mean squared error (MSE), Pearson correlation coefficient (R), and median error—we calculated the absolute difference between the published value and the replicated value for each dataset. We then averaged these absolute differences across datasets and normalized them by the mean of the published values. This yields a replication accuracy score defined as:

$$\text{Replication Accuracy} = 1 - \frac{\text{MAE}_{\text{datasets}}(|\text{Replicated} - \text{Published}|)}{\text{Mean}_{\text{datasets}}(|\text{Published}|)} \quad (1)$$

where a score of 1.00 indicates perfect replication.

2.3 Classification

Building on the predictions produced by the replicated Horvath2013 aging clock, we defined a classification task to distinguish between biologically positively and negatively accelerated aging profiles. Residuals were calculated as the difference between predicted biological age and chronological age, and samples were labeled based on whether this residual was positive or negative. To focus on more biologically meaningful cases of positively and negatively accelerated aging, samples with residuals smaller than one year in absolute value were excluded.

Three classification models were trained and evaluated to ensure the best performing model was selected for further analysis. Logistic Regression was selected as a baseline model. Random Forest was included as a more flexible, nonlinear alternative. Finally, XGBoost was selected based on its improved performance on methylation data compared to Random Forest [8], and its ability to handle high-dimensional inputs, model complex nonlinear interactions, and remain robust to class imbalance and multicollinearity [9].

2.4 Feature Selection

The DNA methylation dataset used in this study includes beta values for 21,368 CpG sites per sample. Given this high dimensionality, feature selection techniques were applied as part of the classification pipeline.

As a first step, a variance threshold filter was applied to eliminate features with near-constant values across all samples. This unsupervised method is computationally inexpensive and helps remove non-informative features that are unlikely to contribute to the classification task. A threshold of 0.01 was chosen, meaning that any feature with variance below this value was discarded.

On the filtered subset of features, a supervised feature selection technique, Recursive Feature Elimination with Cross-Validation (RFECV), was used. RFECV iteratively evaluates subsets of features, removing the least important ones based on the performance of a gradient-boosted decision tree classifier (XGBoost). Feature importance was evaluated using the average gain across decision trees, and model performance was assessed using accuracy. To ensure balanced representation of positively and negatively accelerated classes, a stratified 3-fold cross-validation was used during feature selection. To make the process computationally feasible while preserving sufficient resolution, 200 features were removed per iteration.

2.5 Nested Cross-Validation and Hyperparameter Tuning

To ensure that classification results did not depend on a particular train/test split, a nested cross-validation procedure was implemented. The outer loop consisted of a 5-fold stratified cross-validation to evaluate performance of the best model from the training stage. Within each outer fold, feature selection and hyperparameter tuning were performed on the training portion of the data using an inner 3-fold stratified cross-validation. Different random seeds and independent shuffling were used to ensure that splits for RFECV and hyperparameter search were distinct, avoiding any potential data leakage. This structure ensures that the outer test data remains completely unseen throughout model development. A full schematic of the procedure is shown in Figure A.1 in the appendix.

The hyperparameters for each learning algorithm were tuned using randomized search, following the search space for XGBoost and Random Forest defined in [10] and [8]. Table 1 summarizes the complete set of hyperparameter ranges explored.

Model performance was assessed using the area under the receiver operating characteristic curve (ROC AUC). ROC curves were generated from the predicted class probabilities of each model and averaged across folds.

Learning Algorithm	**Hyperparameter Range**
Logistic Regression	`C`: continuous values in $[10^{-5}, 10^{2}]$; `penalty`: ℓ_1 or ℓ_2 regularization; `solver`: `liblinear`.
Random Forest	`bootstrap`: `[True, False]`; `max_depth`: integers from 10 to 110; `min_samples_leaf`: integers from 1 to 4; `min_samples_split`: integers from 2 to 10; `n_estimators`: integers from 100 to 300.
XGBoost	`min_child_weight`: integers from 1 to 10; `gamma`: continuous values in [0.0, 0.5]; `max_depth`: integers from 10 to 110; `learning_rate`: continuous values in [0.01, 0.11]; `subsample`: continuous values in [0.0, 1.0]; `colsample_bytree`: continuous values in [0.0, 1.0]; `n_estimators`: integers from 100 to 300.

Table 1: Hyperparameter search ranges used during model tuning for each machine learning algorithm. Each range defines the set of values explored in the random search using 3-fold cross validation.

2.6 Age Group Division

To analyze patterns of biological age acceleration across the lifespan, samples were grouped into three age categories based on key inflection points in DNA methylation dynamics and to ensure sufficient sample sizes within each group [11]:

- **Young**: below 20 years,
- **Middle-aged**: 20 to 55 years,
- **Older adults**: 56 years and above.

2.7 Final Feature Selection and SHAP

To interpret model predictions and identify biologically meaningful CpG sites, SHAP (SHapley Additive exPlanations) interaction values were computed using `TreeExplainer` within the nested cross-validation framework. For each outer fold, SHAP interaction values were calculated for the best-performing model, and main effects were extracted for interpretation.

To ensure robustness in feature interpretation, a two-step filtering strategy was applied when aggregating SHAP values across the outer-loop cross-validation folds. First, stability selection was performed by keeping only those CpG sites selected in at least 60% of the folds. Second, SHAP thresholding was applied by keeping CpG sites whose cumulative absolute main effect SHAP values accounted for at least 60%

of total importance. The intersection of these criteria defined the final feature set used for feature importance analysis.

2.8 Feature Importance

To assess how feature contributions varied across the lifespan, mean main effect SHAP values were computed separately for each age group (young, middle-aged, older adults). The final set of CpG features was then categorised into four groups based on their age-related SHAP patterns and subsequently evaluated for biological relevance.

- **Top Positively Accelerating:** Features with a z-score of their mean SHAP main effect greater than or equal to 1, indicating a strong positive contribution to predicted biological age acceleration.
- **Top Negatively Accelerating:** Features with a z-score of their mean SHAP main effect less than or equal to –1, indicating a strong negative contribution to predicted biological age acceleration.
- **Sign-Flipping (Positive Shift):** Features whose contributions were negative in the young age group and positive in older adults.
- **Sign-Flipping (Negative Shift):** Features whose contributions were positive in the young age group and negative in older adults.

To evaluate whether sign-flipping CpG sites showed differences in methylation across age groups, a one-way ANOVA was performed on the methylation levels. P-values were adjusted using the Benjamini-Hochberg procedure, with statistical significance defined at $p_{FDR} < 0.05$. For each significant CpG, effect size was computed using eta squared (η^2):

$$\eta^2 = \frac{\text{SS}_{\text{between}}}{\text{SS}_{\text{total}}} \tag{2}$$

where $\text{SS}_{\text{between}}$ is the between-group sum of squares and SS_{total} is the total variance. CpG sites with $\eta^2 \geq 0.14$ were considered to have a large effect size.

2.9 Gene Set Ontology Enrichment Analysis

To investigate possible functional relevance of selected CpG features, we annotated them using the Illumina HumanMethylation450 manifest file[1]. When a CpG was mapped to multiple genes, all associated gene names were retained.

Gene set enrichment analysis was performed using Enrichr[2]. Separate gene lists were generated for each CpG category: top positively accelerating, top negatively accelerating, and sign-flipping features with positive or negative shifts (the latter two filtered based on statistically significant ANOVA results). For each enrichment analysis, the background set comprised all genes annotated in the Illumina HumanMethylation450 manifest. The analysis was restricted to Gene Ontology (GO) biological process terms, with statistical significance defined as an FDR-adjusted p-value below 0.05.

To improve interpretability of enriched GO terms, we performed semantic similarity-based clustering. Pairwise Resnik semantic similarities was computed between all enriched terms based on the GO hierarchical structure and term annotation frequencies [12]. This approach inspired by [13] allowed us to identify biologically coherent clusters of processes, rather than interpreting terms individually.

3 Responsible Research

The research presented in this Research Project follows the principles of responsible research and ethical integrity. The analysis is based on real human methylation data that has been fully anonymised, thus no personally identifiable information is accessible at any stage of the project. All biological samples were originally collected with informed consent under appropriate ethical oversight in public repositories Gene Expression Omnibus (GEO) and ArrayExpress.

It is important to acknowledge the limitations of the data used. Although it comes from real human samples, the data may not be fully representative of the broader population in terms of ethnicity, lifestyle, or other demographic and environmental factors. Therefore, the findings of this project may not generalize to all population groups and should be interpreted with this caution in mind.

The dataset used in this study was provided by the authors of the AltumAge model via a publicly accessible Github repository [3]. Although accessed through this secondary source, the underlying data originate from publicly available repositories.

This project is designed to be fully reproducible. The datasets used are publicly available, and all methodological steps necessary for replication are described in Section 2. The complete source code is accessible via a public GitHub repository[4]. The findings of this research aim to enhance understanding of the biological mechanisms underlying healthy aging and will not be used for any discriminatory or commercial purposes.

4 Results

We began by asking whether selected aging are fully reproducible. To answer this, we created a replication pipeline of the Horvath2013 and AltumAge models. As shown in Table 2, Horvath2013 achieved perfect replication (1.000) across all evaluation metrics, reflecting its fully deterministic nature. In contrast, AltumAge includes stochastic components, leading to minor deviations in our replication. Its normalized accuracy scores ranges from 0.632 to 0.993, with the lowest agreement observed for the median error metric.

After excluding samples with an absolute residual value below 1 year, a total of 4,119 samples were kept for further development, while 1,125 samples were filtered out.

From the initial variance thresholding step, 14,888 low-variance features were removed, resulting in a reduced feature set of 6,481 CpG sites.

[1] https://emea.support.illumina.com/downloads/infinium_humanmethylation450_product_files.html

[2] https://maayanlab.cloud/Enrichr/

[3] https://github.com/rsinghlab/AltumAge

[4] https://github.com/KlaraHirm/CSE3000

Performance Metric	Horvath2013	AltumAge
Mean Absolute Error	1.000	0.906
Mean Squared Error	1.000	0.927
Pearson Correlation (R)	1.000	0.993
Median Error	1.000	0.632

Table 2: Replication accuracy scores for the Horvath2013 and AltumAge epigenetic clocks across four evaluation metrics: mean absolute error (MAE), mean squared error (MSE), Pearson correlation coefficient (R), and median error. Scores were normalized such that a value of 1.000 indicates perfect agreement between our replicated results and the originally published values. Horvath2013 achieved perfect replication across all metrics, while AltumAge showed slightly lower agreement, particularly for median error.

Feature selection using Recursive Feature Elimination with Cross-Validation (RFECV) led to a significant reduction in dimensionality, as shown in Figure 2. While the optimal number of features varied across folds, from 451 up to 1,172, the cross-validated accuracy curves showed consistent trends. Performance improved sharply with the first few hundred features and then stabilized. This consistency across folds supports the stability of the model and suggests that a compact, informative subset of CpG sites drives prediction performance.

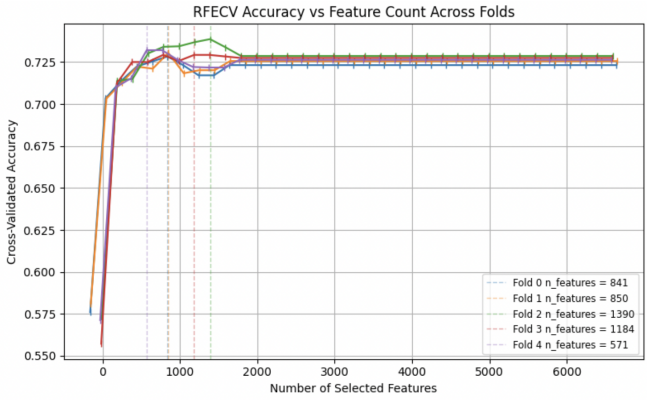

Figure 2: Recursive Feature Elimination with Cross-Validation (RFECV) accuracy across five cross-validation folds, plotted against the number of selected features. Each colored line represents the accuracy curve for one fold. While the optimal number of features varied across folds—from 571 to 1,390—the accuracy curves consistently show rapid initial improvement followed by a plateau. Vertical dashed lines indicate the fold-specific feature count that achieved peak accuracy.

To benchmark classification performance, we compared XGBoost to a baseline linear Logistic Regression model as well as non-linear Random Forest model. As shown in Figure 3, XGBoost achieved a slightly higher mean AUC score than both baseline models. The results suggest that XGBoost offers a modest advantage in distinguishing biologically positively and negatively accelerated profiles. This improvement is likely due to its ability to capture complex non-linear relationships and interactions between features, while also benefiting from more effective regularization and gradient-based optimization compared to Random Forest.

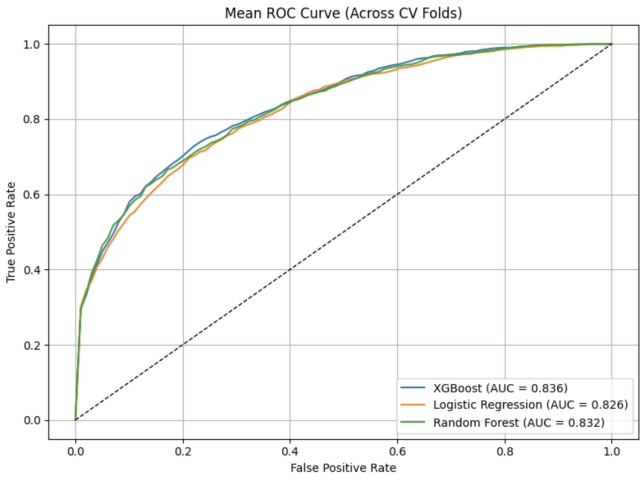

Figure 3: Mean ROC AUC curves from nested 5-fold cross-validation. XGBoost achieved an average AUC of 0.836, compared to 0.826 for Logistic Regression and 0.832 for Random Forest.

Classification performance metrics are summarized in Table 3. XGBoost slightly outperformed both Logistic Regression and Random Forest, achieving the highest accuracy, F1 score, and AUC. While Random Forest showed slightly lower log loss.

Performance Metric	XGBoost	Logistic Regression	Random Forest
Accuracy	**0.7492**	0.7373	0.7395
F1 Score	**0.7529**	0.7424	0.7406
AUC	**0.8371**	0.8266	0.8331
LogLoss	0.5043	0.5057	**0.502**1

Table 3: Classification performance across four evaluation metrics. XGBoost achieved the highest accuracy, F1 score, and AUC, indicating slightly better overall performance. Random Forest obtained the lowest log loss.

Across the nested cross-validation folds as defined in Section 2.5, a total of 391 features were selected in at least 60% of outer folds, satisfying the stability selection criterion. Separately, 551 features accounted for 60% of the total absolute SHAP contribution. The intersection of these two sets resulted in a final feature set of 312 CpGs, used for further analysis and biological interpretation.

Z-score filtering identified 11 CpG sites with strongest positive SHAP contributions, corresponding to 12 different genes, and 7 CpG sites with stronglest negative contributions, each linked to a distinct gene. These features had the most pronounced impact on the XGBoost classifier's predictions. Positive SHAP values reflect features that contribute to increasing acceleration, while negative values indicate those associated with a negatively accelerating effect. Their average SHAP values across samples are shown in Figure 4, highlighting the top positively and negatively contributing CpGs.

To explore how feature importance varies with age, SHAP values were averaged within each age group for the final set of selected features. Features were selected based on the groups defined in Section 2. These patterns were visualized in a heatmap (Figure A.2), where red indicates a positive effect on predicted biological age acceleration and blue indicates

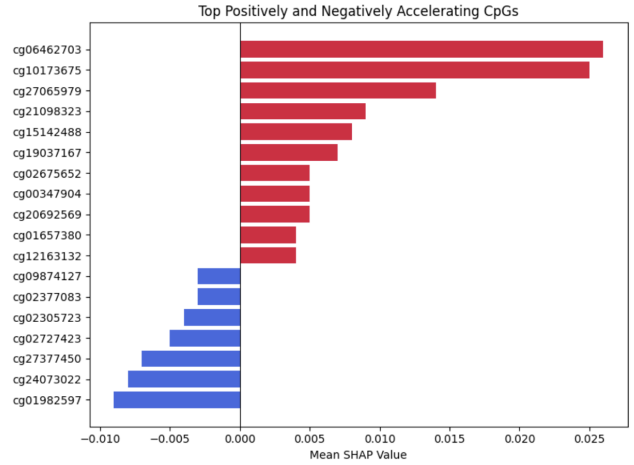

Figure 4: Horizontal bar plot of mean SHAP values for top accelerating CpG sites. Positively accelerating features (red) exhibit high positive SHAP values, while negatively accelerating features (blue) show negative contributions. Bar lengths represent average SHAP magnitude across all samples.

a negative effect. This visualization highlights how the influence of individual CpG sites on model predictions shifts across different stages of the lifespan. Based on these patterns, a negative shift was defined as a feature with a positive effect in the young group and a negative effect in the old group, while a positive shift was defined as a feature with a negative effect in the young group and a positive effect in the middle group.

One-way ANOVA revealed a subset of sign-flipping CpGs as defined in Section 2.5 with statistically significant differences in methylation across age groups (FDR-adjusted $p < 0.05$). A Manhattan-style plot (Figure 5) was used to visualize their chromosomal distribution. CpGs were grouped by shift direction, positive or negative, and those with a large effect size ($\eta^2 \geq 0.14$) within each group were selected separately for GO enrichment analysis. The positively shifting group mapped to 4 significant genes, while the negatively shifting group mapped to 34 significant genes.

GO enrichment analysis was carried out separately for the positively and negatively accelerating CpG sites. In total, 86 GO terms were significantly enriched for the positively accelerating features (FDR-adjusted $p < 0.05$), while 35 terms were identified for the negatively accelerating set. Figures A.3 and A.4 present the top 30 enriched terms from each analysis. The enriched terms for the positively accelerating CpGs were primarily related to immune signaling and neurodevelopmentn. For the negatively accelerating CpGs, the terms were associated with DNA repair, neuronal development, and immune defense. Semantic similarity between GO terms was assessed using the Resnik method.

In addition, a separate enrichment analysis was performed on CpG sites that showed a positive shift with age, defined as having negative acceleration in the young group and positive acceleration in the middle-aged group. Although only four genes were mapped from this set after ANOVA analysis, they were associated with 24 significantly enriched GO

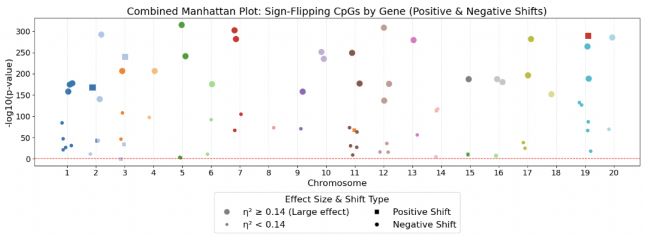

Figure 5: Combined Manhattan-style plot of ANOVA results for sign-flipping CpG sites. Each point represents gene associated with a sign-flipping CpG, with chromosomal position on the x-axis and $-\log_{10}(p\text{-value})$ on the y-axis. Marker shape indicates the shift direction: squares for positive shifts and circles for negative shifts in SHAP contribution across age groups. Dot size reflects effect size (η^2), with larger dots denoting biologically relevant effects ($\eta^2 \geq 0.14$). The red dashed line marks the significance threshold at $p = 0.05$.

terms (FDR-adjusted $p < 0.05$). These terms are visualized in Figure A.5 and highlight biological processes that become increasingly active with age, including oxidative stress response, protein processing, and amyloid-beta metabolism. In contrast, 34 genes were mapped from CpG sites showing a negative shift with age, but no significant GO terms were identified in the enrichment analysis for this group.

5 Discussion

5.1 Reproducibility

Our replication experiments showed perfect agreement with the Horvath2013 clock across all evaluation metrics, demonstrating the reproducibility of linear models. In comparison, replication accuracy for the AltumAge model was slightly lower, particularly with respect to median error. This difference likely reflects the complex nonlinear relationships captured by the deep learning architecture, which can introduce some degree of randomness. Nonetheless, the accuracy in correlation and mean absolute errors for AltumAge support its overall reliability and replicability.

5.2 Classifier Evaluation

Across all evaluation settings, XGBoost consistently outperformed both Logistic Regression and Random Forest classifiers. In nested cross-validation, XGBoost achieved an average AUC score of 83.7%, compared to 82.7% for Logistic Regression and 83.3% for Random Forest. The best-performing XGBoost fold reached a AUC score of 84.8% using the following hyperparameters: `learning_rate` ≈ 0.065, `max_depth = 24`, `n_estimators = 293`, `subsample` ≈ 0.99, `gamma` ≈ 0.086, `colsample_bytree` ≈ 0.98, and `min_child_weight = 8`. The lowest-performing XGBoost fold scored 82.6% with `learning_rate` ≈ 0.075, `max_depth = 12`, `n_estimators = 217`, `subsample` ≈ 0.53, `gamma` ≈ 0.049, `colsample_bytree` ≈ 0.35, and `min_child_weight = 7`.

Logistic Regression showed comparatively stable performance, with its best fold reaching AUC score of 84.9% using `solver = 'liblinear'`, `penalty = 'l2'`, and `C` $\approx$

0.905. Its lowest-performing fold scored 81.2% with the same solver and penalty but a higher regularization strength of $C \approx 6.716$.

Random Forest achieved AUC scores ranging from 82.1% to 83.9% across folds. The best-performing fold used `n_estimators = 157`, `max_depth = 69`, `min_samples_leaf = 1`, `min_samples_split = 2`, and `bootstrap = False`. The lowest-performing fold used `n_estimators = 221`, `max_depth = 70`, `min_samples_leaf = 1`, `min_samples_split = 8`, and `bootstrap = False`.

These results show that XGBoost, although slightly, is the best-performing classifier in the methylation setting, consistent with findings from [8], albeit in a different classification task.

5.3 Gene Set Ontology Enrichment Analysis

Top Positively Accelerating Features

Positively accelerating features, those driving positively accelerating classifier prediction, are enriched in immune-related processes, neurodevelopmental regulation, and cellular response pathways.

One cluster among the enriched terms is the involvement of genes linked to chronic and dysregulated immune responses. For example, *Regulation of Interleukin-18 Production* (GO:0032661), *Toll-Like Receptor 2 Signaling Pathway* (GO:0034134), and *I-kappaB Phosphorylation* (GO:0007252) cluster together semantically and highlight inflammatory signaling pathways that are central to the immune system. These findings relate to the concept of "inflammaging", where persistent, low-grade immune activation contributes to functional decline in aging [14] [15].

Complementing this, the cluster with lowest common ancestor *response to molecule of bacterial origin (GO:0002237)* with keyterms *Response, Lipoteichoic* includes terms such as *Response to Lipoteichoic Acid* (GO:0070391) and *Cellular Response to Bacterial Lipopeptide* (GO:0071221). These terms indicate that responses to bacterial components may be associated with increased biological age. Additionally, the enrichment of *Negative Regulation of Mitochondrial Membrane Permeability* (GO:0035795), although not part of the same cluster, suggests a link between immune activity and mitochondrial stress, an interaction recognised as hallmark of aging [2]. Together, these findings suggest relationship between immune-related pathways and accelerated epigenetic aging.

Among the enriched terms are *Negative Regulation of Synapse Assembly* (GO:0051964) and *Negative Regulation of Synapse Organization* (GO:1905809), both of which relate to processes that suppress the formation of synaptic connections in the brain. This may indicate impaired synaptic plasticity, process essential for learning, memory, and cognitive development [16].

Another coherent group, labeled *developmental process (Development, Ganglion)*, contains terms such as *Sympathetic Nervous System Development* (GO:0048485), *Ganglion Development* (GO:0061548), and *Embryonic Heart Tube Development* (GO:0035050). Since the dataset contains prenatal samples, the enrichment of these terms may partly reflect signals originating from fetal tissue. At the same time, their presence may suggest that some features capture aspects of early biological development, potentially indicating meaningful variation in prenatal maturation.

These clustered enrichments highlight that positively accelerating CpG features are not randomly associated with biological functions but are concentrated in processes already linked to aging, including immune response, neural development, and prenatal growth.

Top Negatively Accelerating Features

Negatively accelerating features are enriched in pathways related to DNA repair, neuronal developmenta and immune defense. These enrichments may reflect mechanisms that support cellular maintenance and resilience against aging.

A top cluster with lowest common ancestor *DNA damage response*, labeled with keywords *Repair, Excision*, includes terms such as *Double-Strand Break Repair via Nonhomologous End Joining* (GO:0006303), *Base-Excision Repair* (GO:0006284), and *DNA Damage Response* (GO:0006974). Since DNA damage accumulates with age and contributes to genomic instability, a known hallmark of aging [2], these results point to efficient repair processes as potential markers of slower biological aging.

The cluster with keywords *Generation, Neurons* features terms like *Neuroblast Proliferation* (GO:0007405), *Generation of Neurons* (GO:0048699), and *Neurogenesis* (GO:0022008), suggesting preserved neurogenic activity. Since neurogenesis typically declines with age, this enrichment could indicate delayed neural aging or greater cognitive robustness. This complements the enrichment of synapse-related repression terms among positively accelerating features, which instead signal reduced plasticity and cognitive decline associated with accelerated biological aging.

Immune-related pathway *Regulation of Defense Response to Virus by Host* (GO:0050691) may reflect effective immune monitoring and stress responses that protect against age-related dysfunction.

Lastly, clusters labeled *regulation of DNA metabolic process (Regulation, Double)* and *regulation of primary metabolic process (Regulation, Transcription)* include terms such as *Positive Regulation of Transcription by RNA Polymerase II* (GO:0045944). This appears to contradict findings that show increased transcriptional elongation by RNA Polymerase II with age [17], potentially highlighting the need to differentiate between regulated transcriptional processes and those arising from age-related dysregulation

Together, these findings highlight biological processes that may slow biological aging, supporting genomic stability, neural function and immune readiness.

Age-Associated Positive Shift Features

Features showing a positive shift with age (negative in younger individuals and increasingly positive in middle group) point to biological processes that are mostly inactive or even suppressed early in life, but become more active with age. This may reflect how the body shifts from efficient internal regulation in youth to activating more stress or repair pathways as damage builds up over time.

One example is the cluster *response to chemical (Response, Cellular)*, which includes *Cellular Response to Reactive Oxygen Species* (GO:0034614) and *Response to Hydrogen Peroxide* (GO:0042542). These are involved in reacting to oxidative stress, a type of damage caused by unstable molecules like free radicals. Their negative association in young individuals likely reflects low stress levels and efficient energy production by mitochondria. This is consistent with the idea that young organisms have higher capacity to manage oxidative stress, but their ability to do so weakens with age [18]. The increase with age may reflect the body compensating for accumulating oxidative damage [2].

Another important cluster involves pathways related to amyloid-beta, a protein linked to Alzheimer's disease. These pathways, such as *Amyloid-Beta Formation* (GO:0034205) and *Amyloid Precursor Protein Catabolic Process* (GO:0042987), show a negative association with biological age in younger individuals. This likely reflects control of these processes early in life, helping to prevent harmful protein build-up in the brain. Research has shown that amyloid-beta plaques begin forming many years before any noticeable symptoms of Alzheimer's appear [19]. While all samples in this dataset came from healthy tissues, the increase in these features in older groups may reflect normal age-related changes in amyloid-beta activity, which can occur even without disease, or possibly early-stage shifts that developed before the disease clinically appeared.

Overall, these shifts suggest that some pathways are kept in check early in life and only start to become active when the body begins to experience more stress or loses control over certain functions. The transition from negative to positive values may reflect a broader change from stability to compensation or damage control.

5.4 Limitations

While this research provides insights into the features driving predicted biological age acceleration in healthy samples, it is necessary to address limitations of this work.

First, due to the time constraint, the analysis relied on preprocessed methylation data obtained from the authors of the AltumAge model. Although this ensured consistency and reproducibility of the aging clocks, it limited control over preprocessing steps.

Second, age acceleration was modeled as a binary classification task, based on residuals from the Horvath2013 clock. Samples with almost perfect prediction, thus near-zero residuals, were excluded to focus on biologically accelerating cases. However, this simplification may have discarded informative samples. While this work tried to find meaning behind those residuals and we looked at overall feature importances, certain residuals might have still been just an error of the model and skew the results.

Lastly, the Gene Ontology (GO) enrichment analysis while sourced and semantically clustered, the interpretation still remains partially speculative, as no experimental validation was used to confirm functional links between the identified CpG features and age acceleration.

5.5 Future Work

Future work should aim to address limitations stated in previous subsection and further explore the mechanisms underlying predicted biological age acceleration in aging clocks.

To reduce reliance on externally preprocessed data, future analyses could begin from raw methylation arrays and either replicate the preprocessing pipeline defined by the AltumAge authors or develop a new pipeline specific to the objective. The work should also ensure that the underlying data is ethically sourced and represents diverse population groups, including variation in ethnicity, age, and environmental background, to ensure generalisability, and fairness of interpreted results.

Incorporating the results of multiple epigenetic clocks, such as AltumAge, could provide a more comprehensive view on biological age acceleration. These clocks differ in CpG coverage, training, and underlying learning algorithm, and their residuals may capture different aspects of aging. Comparing and integrating these outputs could help differentiate between clock-specific noise/errors and shared biological signals.

Rather than treating residuals as a binary outcome, future models could treat age acceleration as a continuous variable or explore multi-class approach. This would preserve more information and allow for more detailed biological insights.

Finally, to increase the power of gene-level analyses, future studies could explore different filtering or grouping strategies, such as statistical test or clustering-based selection, to find a more functionally coherent set of genes. On top of that, longitudinal datasets tracking individuals over time are needed to validate whether interpreted biological pathways results in accelerated aging or disease risk.

6 Conclusion

This Research project investigated the mechanisms underlying discrepancies between biological and chronological age, as predicted by epigenetic aging clocks. While existing models like Horvath2013 and AltumAge offer accurate predictions of biological age based on DNA methylation data, relatively little is known about what drives the residuals, cases where the predicted biological age differs from an actual chronological age. This work focused on understanding which DNA methylation features are associated with these residuals in healthy human tissue samples, with the goal of identifying biological pathways linked to positively or negatively accelerated aging.

The study successfully replicated two widely used epigenetic clocks and defined a classification task to distinguish between positively and negatively accelerated individuals. Using a nested cross-validation framework and rigorous feature selection, XGBoost was identified as the best-performing model for classifying age acceleration status. SHAP-based explanation of the classifier was used to identify features with most predictive value. Gene Ontology enrichment analysis of these features highlighted biologically meaningful processes, including immune signaling, DNA repair, neuronal development, and oxidative stress response that have been previously identified in the aging process.

One of the contributions of this thesis is the grouping of CpG feature importance across age groups, revealing "sign-flipping" features whose influence on predicted biological age shifts with chronological age. This approach provides an age-related view for interpreting changes in DNA methylation and their potential role in healthy aging.

The study has several limitations which have to be addressed, including reliance on preprocessed data, exclusive use of residuals from a single aging clock, and limited biological interpretability. These constraints form future research opportunities, such as using raw data, integrating multiple clocks, modeling acceleration as a continuous variable, and validating predictions against longitudinal methylation studies.

In conclusion, this work deepens our understanding of biological aging by identifying methylation features that distinguish between positively and negatively accelerated aging in healthy individuals. It provides a reproducible pipeline for residual-based classification and interpretation, and it raises important questions about the biological meaning of predicted age discrepancies. As aging clocks become the state-of-the-art tools for age prediction and may be in the future used in practical setting, interpreting their residuals will be essential for understanding why we age the way we do.

Acknowledgements

I would like to acknowledge the use of generative AI tools for stylistic and grammatical support during the writing of this thesis. All of the content, analysis, and interpretation are my own.

References

[1] Robin Holliday. Causes of aging. *Annals of the New York Academy of Sciences*, 854(1):61–71, nov 1998.

[2] Carlos López-Otín, Maria A Blasco, Linda Partridge, Manuel Serrano, and Guido Kroemer. Hallmarks of aging: An expanding universe. *Cell*, 186(2):243–278, January 2023.

[3] Lisa D Moore, Thuc Le, and Guoping Fan. Dna methylation and its basic function. *Neuropsychopharmacology*, 38(1):23–38, July 2012.

[4] Linpei Jia, Weiguang Zhang, and Xiangmei Chen. Common methods of biological age estimation. *Clin. Interv. Aging*, 12:759–772, May 2017.

[5] S. Horvath. DNA methylation age of human tissues and cell types. *Genome Biol.*, 14(10), December 2013.

[6] A. Daunay et al. Centenarians consistently present a younger epigenetic age than their chronological age with four epigenetic clocks based on a small number of cpg sites. *Aging*, 14(19):7718–7733, Oct 2022.

[7] L. R. Lapierre L. P. de Lima Camillo and R. Singh. A pan-tissue DNA-methylation epigenetic clock based on deep learning. *NPJ Aging*, 8(1), April 2022.

[8] Baoshan Ma, Bingjie Chai, Heng Dong, Jishuang Qi, Pengcheng Wang, Tong Xiong, Yi Gong, Di Li, Shuxin Liu, and Fengju Song. Diagnostic classification of cancers using dna methylation of paracancerous tissues. *Scientific Reports*, 12(1), 2022.

[9] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery.

[10] Nguyen Quoc Khanh Le, Duyen Thi Do, Fang-Ying Chiu, Edward Kien Yee Yapp, Hui-Yuan Yeh, and Cheng-Yu Chen. Xgboost improves classification of mgmt promoter methylation status in idh1 wild-type glioblastoma. *Journal of Personalized Medicine*, 10(3):128, 2020.

[11] Shui-Ying Tsang, Tanveer Ahmad, Flora W. K. Mat, Cunyou Zhao, Shifu Xiao, Kun Xia, and Hong Xue. Variation of global dna methylation levels with age and in autistic children. *Human Genomics*, 10(1), 2016.

[12] Xiaoli Xue, Wei Zhang, and Anjing Fan. Comparative analysis of gene ontology-based semantic similarity measurements for the application of identifying essential proteins. *PLOS ONE*, 18(4):e0284274, 2023.

[13] Guangchuang Yu. *enrichplot: Visualization of Functional Enrichment Result*, 2022. R package version 1.16.2.

[14] Claudio Franceschi, Paolo Garagnani, Paolo Parini, Cristina Giuliani, and Aurelia Santoro. Inflammaging: a new immune–metabolic viewpoint for age-related diseases. *Nature Reviews Endocrinology*, 14(10):576–590, 2018.

[15] Antero Salminen and Kai Kaarniranta. Nf-b signaling in the aging process. *Journal of Clinical Immunology*, 29(4):397–405, May 2009.

[16] Sara N. Burke and Carol A. Barnes. Neural plasticity in the ageing brain. *Nature Reviews Neuroscience*, 7(1):30–40, 2006.

[17] Cédric Debès, Antonios Papadakis, Sebastian Grönke, Özlem Karalay, Luke S. Tain, Athanasia Mizi, Shuhei Nakamura, Oliver Hahn, Carina Weigelt, Natasa Josipovic, Anne Zirkel, Isabell Brusius, Konstantinos Sofiadis, Mantha Lamprousi, Yu-Xuan Lu, Wenming Huang, Reza Esmaillie, Torsten Kubacki, Martin R. Späth, Bernhard Schermer, Thomas Benzing, Roman-Ulrich Müller, Adam Antebi, Linda Partridge, Argyris Papantonis, and Andreas Beyer. Ageing-associated changes in transcriptional elongation influence longevity. *Nature*, 616(7958):814–821, 2023.

[18] Jiao Meng, Zhenyu Lv, Xinhua Qiao, Xiaopeng Li, Yazi Li, Yuying Zhang, and Chang Chen. The decay of redox-stress response capacity is a substantive characteristic of aging: Revising the redox theory of aging. *Redox Biology*, 11:365–374, 2017.

[19] George S. Bloom. Amyloid- and tau. *JAMA Neurology*, 71(4):505, 2014.

A Appendix A

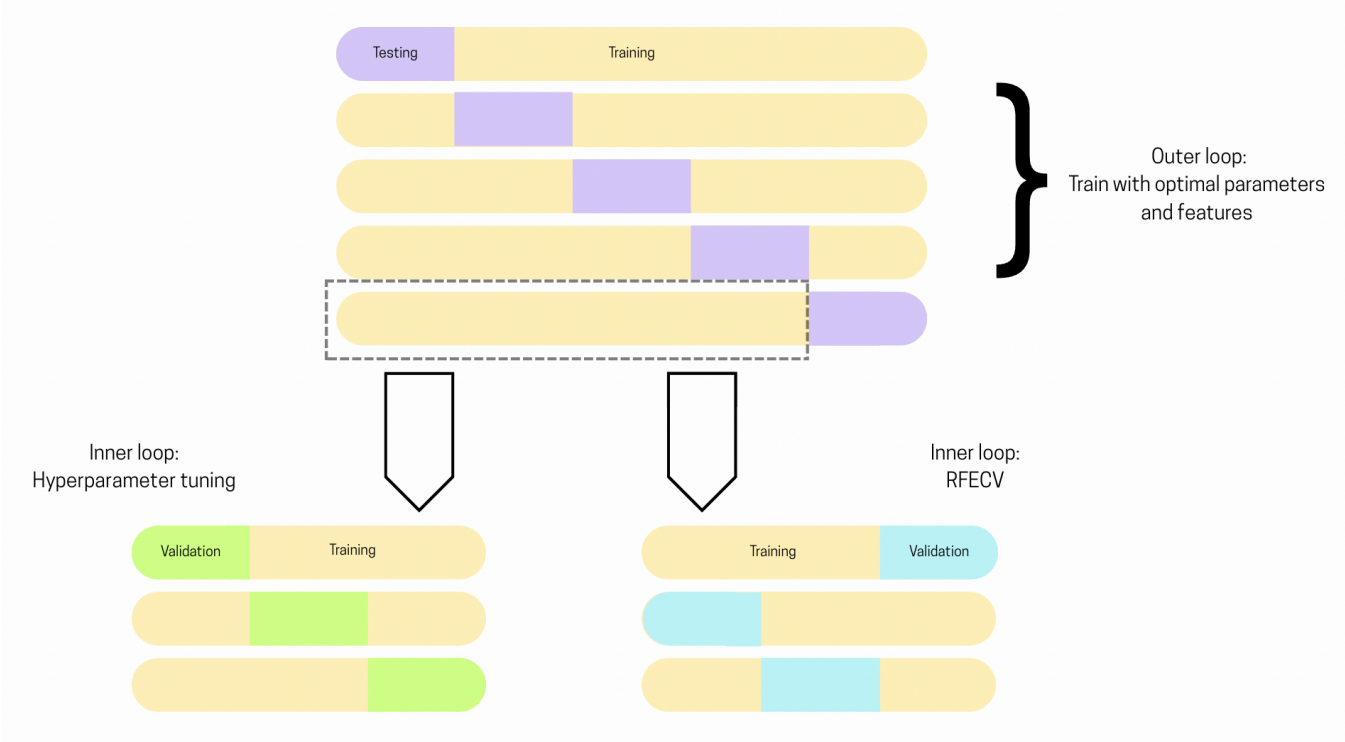

Figure A.1: Nested cross-validation framework used for model development. The outer loop partitions the data into training and testing folds to evaluate model generalization. Within each outer training fold, two inner loops are executed: one for hyperparameter tuning using cross-validation (left), and one for recursive feature elimination with cross-validation (RFECV, right). The final model is then evaluated on the corresponding outer test fold.

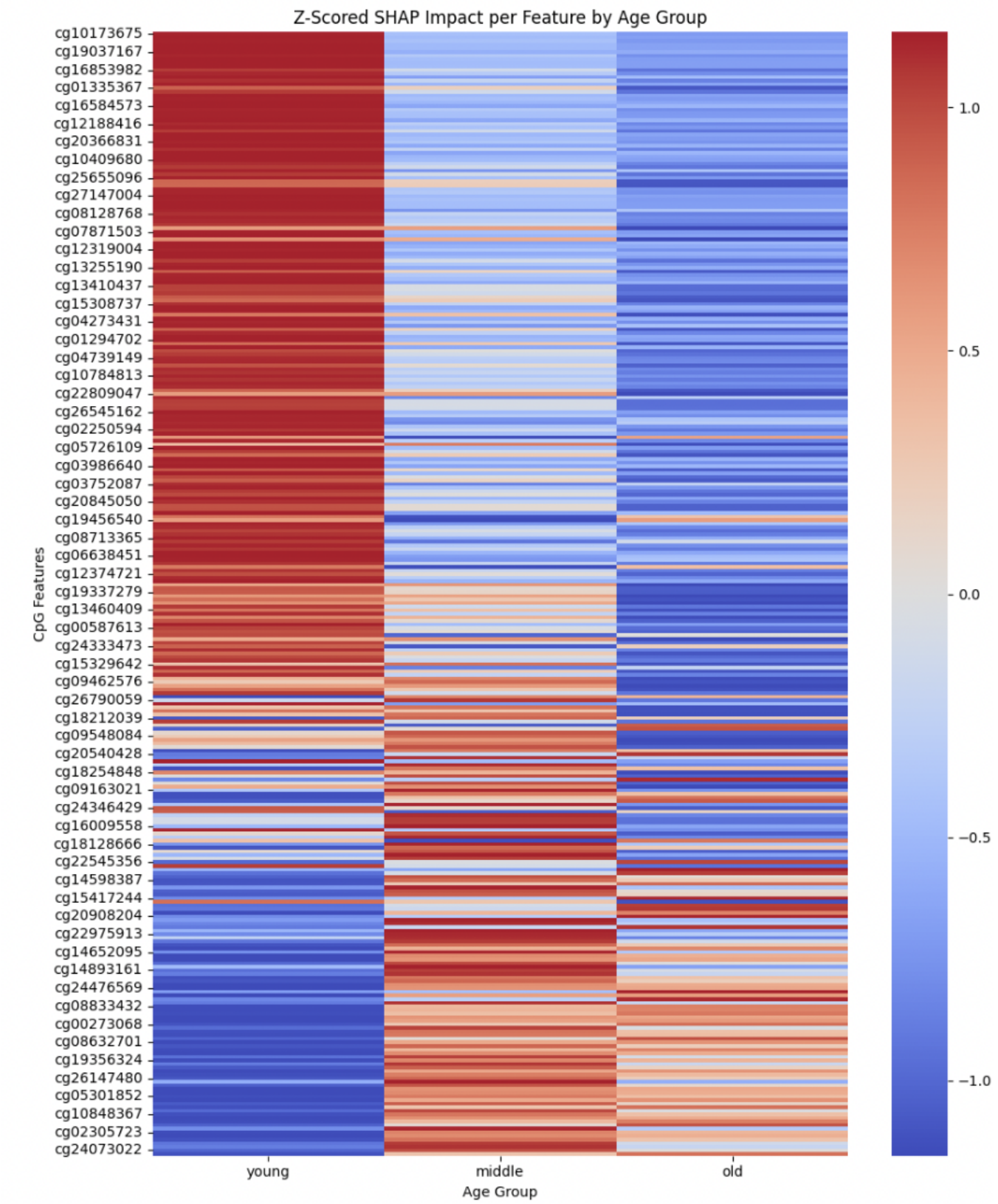

Figure A.2: Z-scored mean SHAP values of selected CpG features stratified by age group. Each row corresponds to a CpG site, and columns represent the young, middle-aged, and older groups. Red indicates greater contribution towards predicted biological age positive acceleration, while blue indicates contribution toward negative acceleration. Features are ordered by their SHAP impact in the young group.

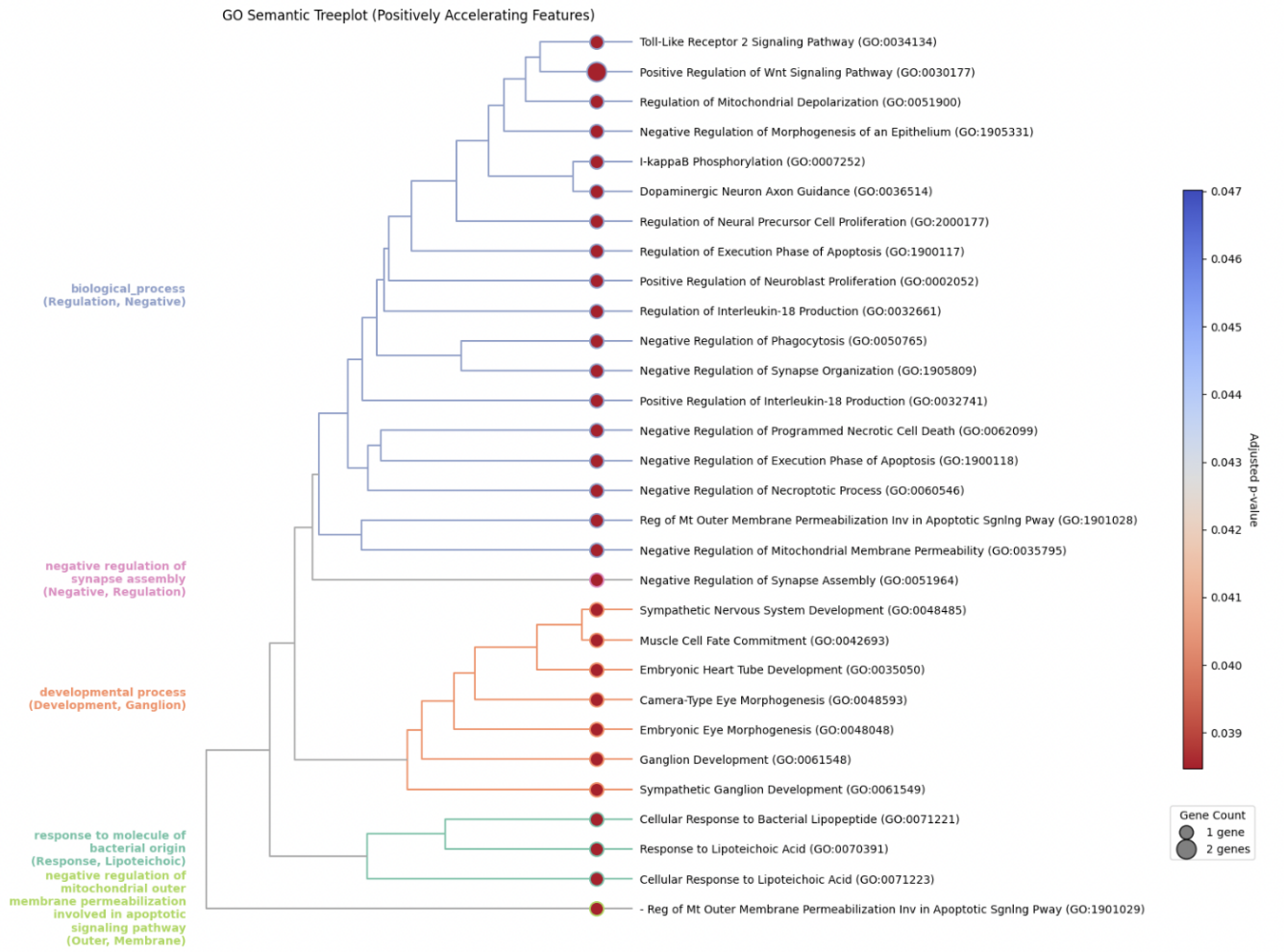

Figure A.3: Hierarchical clustering of the top 30 enriched Gene Ontology (GO) biological processes associated with genes linked to CpG sites identified as positively accelerating. Each node represents a GO term, with bubble size indicating the number of associated genes and color denoting the adjusted p-value (FDR ≤ 0.05). Semantic similarity between terms was calculated using the Resnik method, and branch colors indicate clusters of functionally related terms. Cluster labels are generated by combining the most specific common GO ancestor of the terms within each cluster with the most frequent keywords extracted from the corresponding GO term names. The visualization highlights biological pathways reflecting processes associated with features that exhibit a positive influence on predicted acceleration.

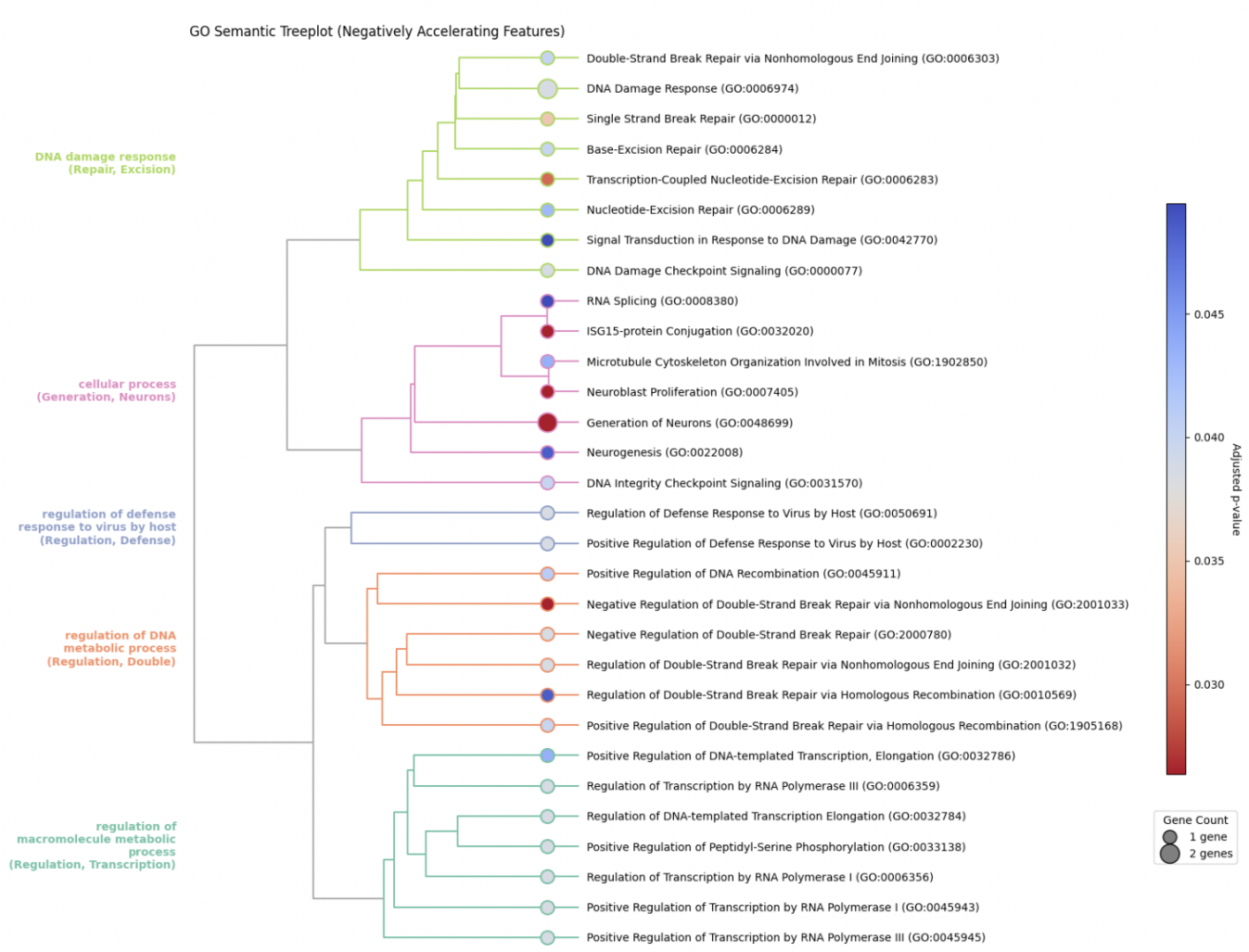

Figure A.4: Hierarchical clustering of the top 30 enriched Gene Ontology (GO) biological processes associated with genes linked to CpG sites identified as negatively accelerating. Each node represents a GO term, with bubble size indicating the number of associated genes and color denoting the adjusted p-value (FDR ≤ 0.05). Semantic similarity between terms was calculated using the Resnik method, and branch colors indicate clusters of functionally related terms, with representative cluster labels shown in matching text. The visualization highlights biological pathways reflecting processes associated with features that exhibit negative influence on predicted acceleration.

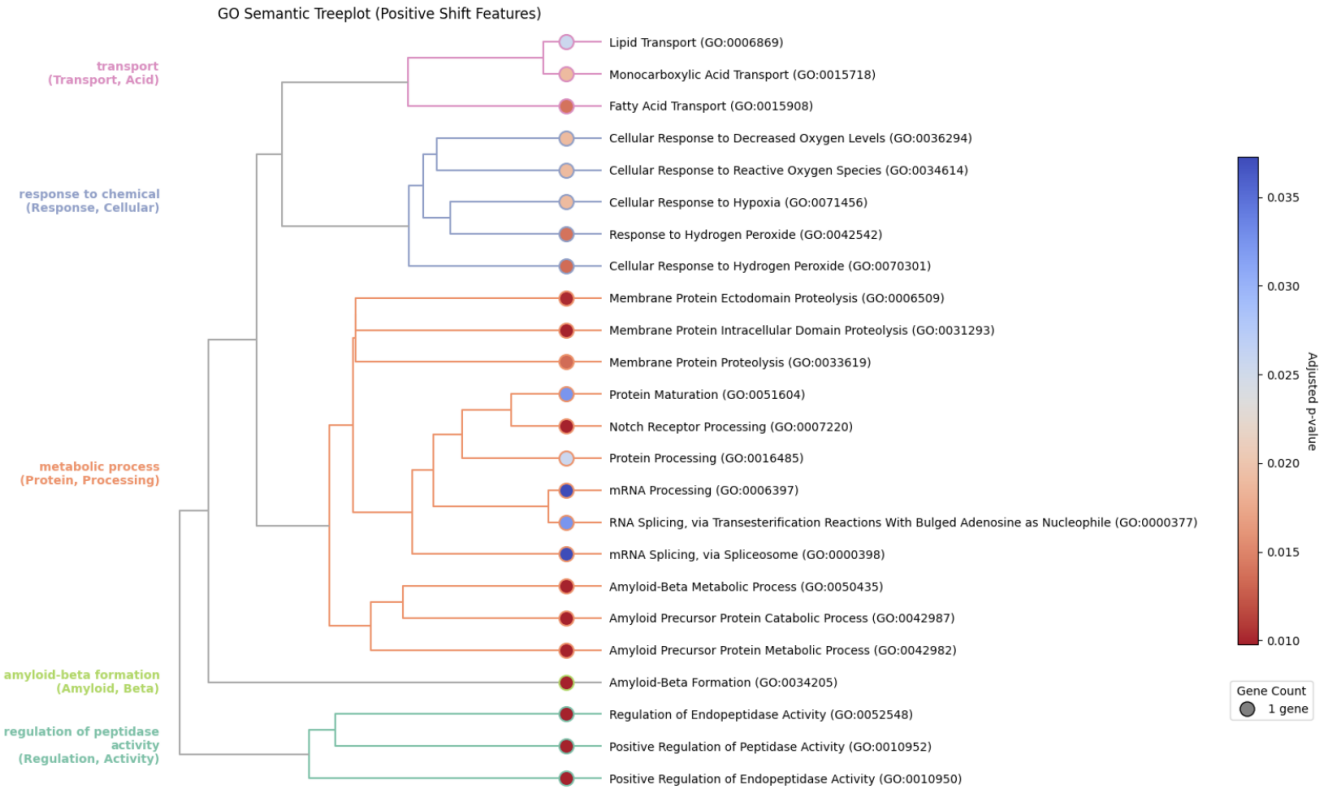

Figure A.5: Hierarchical clustering of the top 30 enriched Gene Ontology (GO) biological processes associated with genes linked to CpG sites showing a positive shift with age defined as negative acceleration in the young group and positive acceleration in the middle-aged group. Each node represents a GO term, with bubble size indicating the number of associated genes and color denoting the adjusted p-value (FDR ≤ 0.05). Semantic similarity between terms was calculated using the Resnik method, and branch colors indicate clusters of functionally related terms, with representative cluster labels shown in matching text. The visualization highlights biological pathways increasingly activated with age, particularly among features that transition from negatively accelerating to positively accelerating influence on predicted biological age.