Ageing Prediction Models from Blood Gene Expression



Marta Duarte
Master in Data Science
Department of Computer Science
Faculty of Sciences of the University of Porto
2026

Ageing Prediction Models from Blood Gene Expression

Marta Duarte

Dissertation carried out as part of the Master in Data Science
Department of Computer Science
2026

Supervisor
Pedro G. Ferreira, Full Professor, Faculty of Sciences of the University of Porto (FCUP) and Institute for Systems and Computer Engineering, Technology and Science (INESC TEC)

Co-supervisor
Marta Moreno, Researcher, Institute for Research and Innovation in Health (i3s)

Acknowledgements

This thesis would not have been possible without the support of many people who accompanied me throughout this journey.

First and foremost, I would like to express my sincere gratitude to my supervisors, Professor Pedro Ferreira and Marta Moreno, for their invaluable guidance and dedication. They were always available whenever I needed, responded promptly, and had a remarkable ability to keep me motivated. Their support made all the difference.

To my family, the people I love most in this world. To my grandparents, whose pride and unconditional love have always warmed my heart and whose wish for my bright future has never wavered. To Marta Ribeiro, who taught me that family is not only the one we are born into but also the one we are lucky enough to find. To João Amaral, I would not be here if it were not for a conversation with you, and I am so glad I listened. Thank you for being there every step of the way. To my parents, who are simply the best, I will never find the words, nor the means, to return everything you have given me. You are by my side no matter what, you encourage me to do what I love, and you remind me every single day that I am loved. Everything I am, I owe to you. And to my sister, who annoys me endlessly and whom I love more than anything. You look up to me, but the truth is I do not know what I would do without you either.

To my friends, the ones who chose me, and let me choose them back. Sofia Minnemann, who has known me longer than I have known myself, you are not just a friend, you are family, and that is forever. Gonçalo Costa and Miguel Gamito, who both made me feel like a best friend from the very first moment, and have been putting up with me ever since. Thank you for always staying. Diogo Domingues, who made university special, we went through all of it together, the highs and the lows, and I am so glad it was you beside me. Cláudia Oliveira, who came with this master's degree and stayed as one of the best people I know, I share this diploma with you, truly, because I could not have done it without you. And Inês Yaksetig, some people go a whole lifetime without ever finding a friendship like this. I was just lucky enough to find mine in you.

Resumo

Os relógios de envelhecimento são modelos de aprendizagem automática supervisionada que preveem a idade cronológica a partir de dados biológicos de elevada dimensionalidade. A expressão génica no sangue representa um substrato particularmente acessível para esta tarefa, mas a área carece de *pipelines* reprodutíveis que combinem um pré-processamento rigoroso com uma comparação sistemática de modelos e validação interpretável das características utilizadas. Esta tese desenvolve e avalia um *pipeline* de aprendizagem automática para previsão da idade a partir de dados de *RNA-seq* de sangue total, utilizando a coorte *GTEx* v11. O *pipeline* integra um pré-filtro de expressão génica diferencial, seis métodos complementares de seleção de características avaliados em múltiplos tamanhos de conjuntos de genes, e uma estrutura comparativa de regressão que abrange modelos lineares penalizados, *ensembles* baseados em árvores, e redes neuronais. A prevenção de fuga de dados (*data leakage*) é assegurada em todas as etapas, com todo o pré-processamento ajustado exclusivamente nos dados de treino e o conjunto de teste reservado até à avaliação final. O *XGBoost* foi selecionado como o modelo preditivo final, alcançando um *MAE* de teste de 7.35 anos e R^2 de 0,506 num conjunto de teste de 152 dadores. Um *ensemble* ponderado de *Elastic Net, Random Forest* e *XGBoost* foi também avaliado, obtendo apenas uma melhoria marginal (*MAE* = 7.29 anos), insuficiente para justificar o aumento de complexidade e a perda de interpretabilidade associados. A rede neuronal não conseguiu generalizar, provavelmente devido ao tamanho insuficiente da amostra. Além disso, a sobreamostragem sintética através de *SMOGN* foi avaliada e verificou-se que não melhorava o desempenho face à distribuição de idades desequilibrada no conjunto de treino. A aceleração da idade (Δ*Age*), a diferença entre a idade prevista e a idade cronológica utilizado ao longo desta tese como *proxy* para o envelhecimento biológico, revelou um enviesamento claramente dependente da idade, com os dadores mais jovens a serem superestimados e os mais velhos subestimados, refletindo a influência residual deste desequilíbrio mesmo após a avaliação estratificada. A assinatura genética final, composta por 200 genes, foi interpretada utilizando os coeficientes do *Elastic Net* para determinar a direção da associação com a idade, e valores de*acrshortshap* calculados sobre o modelo *XGBoost* final para avaliar a importância de cada gene, sendo comparada com duas assinaturas de referência publicadas e analisada através de redes de interação proteína-proteína, confirmando a coerência biológica da assinatura para além do seu desempenho preditivo. Estes resultados demonstram

que é possível obter previsão competitiva da idade transcriptómica e assinaturas genéticas interpretáveis a partir de uma coorte modesta através de um desenho metodológico cuidadoso, fornecendo uma base reprodutível para futura validação entre coortes. Todo o código utilizado para gerar os resultados apresentados nesta tese está disponível em `https://github.com/martacduarte/blood-gene-expression-ageing-prediction`.

Palavras-chave: Relógio de envelhecimento transcriptómico, expressão génica no sangue, *RNA-seq*, previsão de idade, seleção de características, expressão génica diferencial, aprendizagem automática, *GTEx*.

Abstract

Ageing clocks are supervised machine learning models that predict chronological age from high-dimensional biological data. Blood gene expression represents a particularly accessible substrate for this task, yet the field lacks reproducible pipelines that combine rigorous preprocessing with systematic model comparison and interpretable feature validation. This thesis develops and evaluates a machine learning pipeline for age prediction from whole blood RNA-seq data using the GTEx v11 cohort. The pipeline integrates differential gene expression pre-filtering, six complementary feature selection methods evaluated at multiple gene set sizes, and a comparative regression framework spanning penalised linear models, tree-based ensembles, and neural networks. Strict data leakage prevention is enforced throughout, with all preprocessing fitted exclusively on training data and the test set held out until final evaluation. XGBoost was selected as the final predictive model, achieving a test MAE of 7.35 years and R^2 of 0.506 on a held-out set of 152 donors. A weighted ensemble of Elastic Net, Random Forest, and XGBoost was also evaluated, achieving only a marginal improvement (MAE = 7.29 years), insufficient to justify an increase in complexity and reduced interpretability. The neural network failed to generalise, likely due to insufficient sample size. Moreover, synthetic oversampling via SMOGN was assessed and found to not improve performance on the age-imbalanced training distribution. Age acceleration (ΔAge), the residual between predicted and chronological age used throughout this thesis as a proxy for biological ageing, showed a clear age-dependent bias, with younger donors overestimated and older donors underestimated, reflecting the residual influence of this imbalance even after stratified evaluation. The final 200-gene signature was interpreted using Elastic Net coefficients for direction of age association and SHAP values computed on the final XGBoost model for importance, compared against two published reference signatures and analysed through protein-protein interaction networks, confirming biological coherence alongside predictive performance. These results demonstrate that competitive transcriptomic age prediction and interpretable gene signatures can be obtained from a modest cohort through careful methodological design, providing a reproducible foundation for future cross-cohort validation. All code used to generate the results presented in this thesis is made available at `https://github.com/martacduarte/blood-gene-expression-ageing-prediction`.

Keywords: Transcriptomic ageing clock, blood gene expression, RNA-seq, age prediction, feature selection, differential gene expression, machine learning, GTEx.

Table of Contents

List of Figures viii
List of Abbreviations ix

1. Introduction 1
   1.1. Research Questions 2
   1.2. Thesis Structure 3
2. State of Art 5
   2.1. Ageing Biomarkers 5
      2.1.1. Definition 5
      2.1.2. Classes of Biomarkers 5
   2.2. Transcriptomics and Gene Expression 6
      2.2.1. Measurement Technologies 8
      2.2.2. Relevance of Ageing Biomarkers 9
      2.2.3. Challenges 10
   2.3. Transcriptomic Ageing Clocks 10
      2.3.1. Definition 11
      2.3.2. Early Transcriptomic Ageing Models 12
      2.3.3. Modern Blood-Based Transcriptomic Ageing Clocks 12
      2.3.4. Tissue-specific and Multi-tissue Transcriptomic Clocks 13
      2.3.5. Comparison with other Ageing Clocks 13
   2.4. Machine-Learning Methods for Age Prediction 14
      2.4.1. Data Preprocessing and Feature Engineering 14
      2.4.2. Classical ML Models 15
      2.4.3. Deep and Knowledge-Primed Neural Networks 16
      2.4.4. Evaluation and Calibration 17
   2.5. Multi-Omics and Multimodal Ageing Clocks 17
   2.6. Data Resources and Cohorts 18
      2.6.1. GTEx and Multi-modal Omics Resources 19
      2.6.2. Large-Scale Cohort Studies for Blood transcriptomic ageing 19
      2.6.3. Public Repositories 21
   2.7. Limitations 21
      2.7.1. Noise, Variability, and Data Imbalance in Blood Gene Expression 21
      2.7.2. Tissue Specificity and Limited Transferability 22
      2.7.3. Sample Size Limitations and Absence of Longitudinal Data 22
      2.7.4. Challenges in Causal Interpretation 23

2.7.5. Reproducibility and Absence of Standardised Pipelines 23
2.8. Positioning and Contributions of This Thesis 23
3. Methods 25
3.1. Data Acquisition and Preprocessing 26
3.1.1. Dataset 26
3.1.2. Expression Data Transformation and Quality Filtering 28
3.1.3. Exploratory Data Analysis (EDA) 28
3.2. Batch Effect Correction and Quality Control 29
3.2.1. Batch Correction 29
3.2.2. Quality Control 30
3.3. Train/Test Split 31
3.4. Differential Gene Expression (DGE) 32
3.4.1. Motivation 32
3.4.2. Implementation 33
3.4.3. Results 35
3.5. Feature selection 37
3.5.1. Preprocessing for Feature Selection 37
3.5.2. Feature Selection Methods 38
3.5.3. Results and Configuration Selection 40
3.6. Regression Modelling 42
3.6.1. Hyperparameter Optimisation 43
3.6.2. Evaluation Metrics 43
3.6.3. Models 44
3.6.4. Weighted Ensemble 49
3.7. SMOGN 49
4. Results and Discussion 51
4.1. Results 51
4.2. Model Performance 54
4.3. Interpreting Gene Expression Signal 54
4.3.1. Model Choice for Biological Interpretation 54
4.3.2. Genes Overlapping with State-of-the-Art Transcriptomic Clocks 58
4.3.3. Top Ranked Unique Genes 60
4.3.4. Overall Signature Profile 61
4.4. Protein-Protein Interaction Network of the Gene Signature 63
4.5. Technical Strengths 65
5. Conclusions 67
5.1. Answers to the Research Questions 68
5.2. Limitations 69

5.3. Future Work 70

Bibliography 70
A. Appendix 87

List of Figures

2.1. Overview of the main omics layers and their biological scope. 7
2.2. Transcriptomic ageing clock pipeline. 11

3.1. Overview of the transcriptomic age prediction pipeline developed. 25
3.2. RIN distribution. 27
3.3. Sex distribution. 27
3.4. SMCENTER distribution. 27
3.5. Age distribution. Left: histogram with mean (red) and median (green) indicated. Right: boxplot with summary statistics. 28
3.6. PCA analysis on top 2000 genes (colored by SMCENTER). 29
3.7. PCA after Combat - top 2000 genes (colored by SMCENTER). 30
3.8. Age distribution: train (left) vs test (right). 32
3.9. Voom plot: mean-variance trend. 35
3.10 Volcano plot: age-associated genes (FDR < 0.05). 36
3.11 R^2 across feature selection methods and feature set sizes. 41
3.12 Test MAE across feature selection methods and feature set sizes. 41

4.1. XGBoost predicted vs true age scatter plot (right). 52
4.2. Age acceleration by age group. 53
4.3. Top 20 genes by Elastic Net absolute coefficient magnitude on the held-out test set. Red bars indicate genes whose expression increases with age; blue bars indicate genes whose expression decreases with age. 55
4.4. Top 20 genes ranked by mean absolute SHAP value. 56
4.5. SHAP summary plot showing feature value impact and direction. 57
4.6. Overlap between the 200-gene Ridge signature and the top 200 genes from *Peters et al.* [7]. The seven shared genes are listed in the intersection. 58
4.7. STRING protein-protein interaction network of the 200-gene Ridge signature. 63

A.1. Poster presented at the Department of Computer Science Master's Day. 88

List of Abbreviations

AFAR American Federation for Aging Research 5

ARD Automatic Relevance Determination 39, 40

BiT Binarized Transcriptomic 12, 14, 23, 54

C Cytosine 9

cDNA complementary DNA 8

CIBERSORT Cell-type Identification By Estimating Relative Subsets Of RNA Transcripts 15, 70

CPM Counts Per Million 7

CV Cross-Validation 40, 48, 49

dbGaP Database of Genotypes and Phenotypes 20, 70

DESeq2 Differential Expression analysis for Sequencing data 9, 14

DGE Differential Gene Expression 32, 33, 36, 37, 40–42, 65

DNA Deoxyribonucleic Acid 1, 5–8, 10, 12–14, 17–19, 60

ENA European Nucleotide Archive 21

FastQC Fast Quality Control 9

FDR False Discovery Rate 36, 42, 65, 67

FHS Framingham Heart Study 20

FPKM Fragments Per Kilobase per Million 7

FWER Family-Wise Error Rate 35

G Guanine 9

GCT Gene Cluster Text 26, 33

GDC Genomic Data Commons 21

GEO Gene Expression Omnibus 13, 20, 21, 70

GSVA Gene Set Variation Analysis 15

GTEx Genotype-Tissue Expression project ii–iv, 2, 3, 7, 13, 16, 18, 19, 22, 23, 25–31, 33, 49, 54, 67–70

HISAT2 Hierarchical Indexing for Spliced Alignment of Transcripts 8

IQR Interquartile Range 53

k-NN k-Nearest Neighbors 40

KS Kolmogorov-Smirnov 32

LASSO Least Absolute Shrinkage and Selection Operator 11, 12, 15, 38–40, 42, 44, 65

LightGBM Light Gradient-Boosting Machine 15

MAE Mean Absolute Error ii, iv, viii, 11, 13, 14, 17, 40–51, 54, 67, 68

MDI Mean Decrease in Impurity 39

MI Mutual Information 38, 40

ML Machine Learning v, 15, 65

MRI Magnetic Resonance Imaging 6

MuTHER Multiple Tissue Human Expression Resource 20

OMICmAge Omics-based Methylation Age 18

PC Principal Component 30

PCA Principal Component Analysis 15, 23, 28, 30, 31, 68

PCR Polymerase Chain Reaction 8

ReLu Rectified Linear Unit 48

RF Random Forest 39, 40

RIN RNA Integrity Number 9, 10, 26, 30, 31, 34, 42, 67

RMSE Root Mean Squared Error 44, 50, 51

RNA Ribonucleic Acid 1, 2, 5–10, 21, 26, 30, 31, 67

RNA-seq RNA Sequencing ii–iv, 3, 7, 8, 11–14, 18–21, 23, 25, 26, 28, 29, 31, 33–35, 37, 67, 70

RPKM Reads Per Kilobase per Million 7

RUVSeq Remove Unwanted Variation from RNA-Seq Data 15, 29

SAMPID Sample Identifier 26

SHAP SHapley Additive exPlanations iv, 54–57, 60, 61, 64, 65, 67

SHIP Study of Health in Pomerania 20, 70

SMCENTER Sample Center 26, 28, 29

SMGEBTCH Sample Genotype/Expression Batch 26

SMGEBTCHT Sample Genotype/Expression Batch Type 26

SMOGN Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise ii, iv, 49, 50, 68

SMRIN Sample RNA Integrity Number 26

SMTS Sample Tissue Source 26, 27

SMTSD Sample Tissue Source Detail 26, 27

SRA Sequence Read Archive 21

STAR Spliced Transcripts Alignment to a Reference 8, 9

STRING Search Tool for the Retrieval of Interacting Genes/Proteins viii, 63, 67

SUBJID Subject Identifier 26

SVA Surrogate Variable Analysis 15, 29

Tanh Hyperbolic Tangent 48

TCGA The Cancer Genome Atlas 21

TMM Trimmed Mean of M-values 9, 34, 35

TPM Transcripts Per Million 7, 9, 26, 28, 33, 35, 36, 67

TRIzol Total RNA Isolation 8

UK United Kingdom 20, 21, 70

VST Variance Stabilizing Transformation 14

XGBoost eXtreme Gradient Boosting ii, iv, viii, 15, 42, 46–49, 51, 52, 54–58, 60–62, 64, 65, 67, 68

1. Introduction

Ageing represents the gradual, time-dependent decline in physiological function, accompanied by an increasing risk of chronic disease and mortality [1]. As global life expectancy continues to rise, the burden of age-related conditions, including cardiovascular disease, neurodegeneration, metabolic disorders, and cancer, places growing pressure on healthcare systems worldwide [2]. Understanding the molecular mechanisms that drive ageing, and developing tools to quantify its rate in individual patients has therefore become one of the central challenges at the intersection of data science and modern biology.

A fundamental distinction in this field is between chronological age, the time elapsed since birth, and biological age, which reflects the cumulative impact of genetic, environmental, and lifestyle factors on the organism [1, 2]. Chronological age increases at the same rate for everyone; biological age does not. Two individuals of the same chronological age can differ substantially in their physiological and molecular state, and it is this divergence that motivates the search for quantitative biomarkers capable of capturing biological rather than chronological age. Such biomarkers would not only improve our understanding of why some individuals age faster than others, but could also serve as targets in clinical trials of anti-ageing interventions, as risk stratification tools in preventive medicine, and as the basis for personalised health monitoring [3].

Age-associated changes have been described at multiple molecular and physiological levels, including DNA methylation, gene expression, protein and metabolite abundances, telomere length, cellular senescence, and organ structure and function [1]. These changes can be measured at scale through the field of omics, a family of high-throughput approaches that profile the complete set of a specific molecular type within a biological system. Each layer targets a different class of molecules: genomics characterises the full DNA sequence and its variants; epigenomics maps chemical modifications that regulate gene activity; transcriptomics measures the RNA molecules actively produced from genes, reflecting which programmes are currently running in a cell; proteomics quantifies the resulting proteins; and metabolomics profiles the small molecule end products of cellular metabolism.

Advances in high-throughput technologies have enabled genome-wide profiling, powering statistical and machine learning models that map these molecular features to age, often

referred to as ageing clocks [4]. Among these, DNA methylation-based clocks stand out for their accuracy and robustness [5, 6]. Clocks have also been trained on transcriptomics, proteomics, metabolomics, imaging, and clinical data [4, 7], and different data types have been combined to build multimodal predictors [4, 8, 9] with the aim of improving prediction performance while better capturing the biological complexity of ageing. While multimodal approaches have shown promise, their practical deployment is limited by data availability, the requirement for matched multi-omic measurements across large cohorts, and substantial analytical complexity, motivating the development of robust and reproducible single-modality pipelines.

Among the molecular layers available for biomarker development, transcriptomics is particularly informative. The transcriptome, the complete set of RNA molecules actively produced by a cell at a given time, reflects the dynamic functional state of cells in response to ageing, disease, and the environment. Unlike the genome, which remains largely static throughout life, transcriptomic profiles change continuously, capturing which biological pathways are active, which are stressed, and which are dysregulated. Large-scale studies across human cohorts and multiple tissues have established robust transcriptomic signatures of ageing (Section 2.2.2), patterns that are broadly conserved across species and tissues [10, 11].

Blood is a particularly attractive source for transcriptomic ageing biomarkers, as it is obtainable from a routine venipuncture rather than invasive tissue biopsy and is especially sensitive to immune ageing (Section 2.2.2). These properties make whole blood RNA sequencing a practical choice for developing ageing clocks intended for clinical use, and form the basis of the work presented in this thesis.

1.1. Research Questions

Despite substantial progress in the field of transcriptomics age prediction, several challenges remain unresolved. Blood gene expression data are high-dimensional, noisy, and subject to technical batch effects arising from differences in sample processing across laboratories and sequencing platforms. Real-world cohorts are typically age-imbalanced, with post-mortem resources such as GTEx heavily concentrated in middle and older ages, limiting model sensitivity at the extremes of the age range. Published transcriptomic clocks vary widely in their preprocessing strategies, feature selection approaches, and modelling choices, and few provide the methodological transparency or code availability

needed for independent replication. Furthermore, most studies prioritise predictive accuracy over biological interpretability, producing black-box models whose gene signatures cannot be meaningfully cross-validated against the broader ageing literature. This thesis addresses these challenges through the development and evaluation of a reproducible machine learning pipeline for age prediction from whole blood RNA-seq data. We aim to address four key research questions:

1. Can robust and accurate models be developed to predict biological age from blood gene expression data?
2. Can we overcome technical challenges posed by incomplete and imbalanced distributions of the response variable (age)?
3. Can we build models that generalize well across independent blood gene expression datasets?
4. Is it possible to identify a minimal age signature for accurate age prediction? Could such a signature serve as the basis for developing a practical age prediction device?

Throughout this thesis, biological age is operationalised as age difference (ΔAge), the residual between predicted and chronological age (Section 2.3.1). Since no ground truth biological age label exists in the GTEx cohort, all models in this work are trained and evaluated against chronological age, the only directly measurable target, with ΔAge subsequently interpreted as a proxy signal for divergence from expected biological ageing.

These questions span the full breadth from technical methodology to biological discovery and clinical translation, reflecting the dual objectives of the thesis: to produce a well-calibrated predictive model and to extract biologically meaningful insights from the gene signature it learns.

1.2. Thesis Structure

The remainder of this thesis is organised as follows. Chapter 2 reviews the state of the art in ageing biomarkers, transcriptomics, machine learning methods for age prediction, and the data resources that have shaped the field, identifying the key open challenges that motivate this work. Chapter 3 describes the full pipeline from data acquisition and preprocessing through differential gene expression analysis, feature selection and regression modelling. Chapter 4 presents the results obtained on the held-out test set and discusses

them in biological and methodological context, including model performance and diagnostic metrics, gene signature interpretation, protein-protein interaction network analysis, and comparison with published transcriptomic ageing clocks. Chapter 5 presents conclusions, addresses each research question, discusses the limitations of the work, and outlines the most important directions for future work.

2. State of Art

2.1. Ageing Biomarkers

2.1.1 Definition

Biomarkers of ageing are measurable biological characteristics that reflect age-related functional decline across organ, tissue, cellular, and molecular scales [1]. They fulfil three essential roles: (i) quantifying differences in ageing rates, (ii) tracking physiological responses to lifestyle or pharmacological interventions, and (iii) anticipating the transition from healthy ageing to age-related disease or mortality [1, 9]. The American Federation for Aging Research (AFAR) established key criteria for valid ageing biomarkers to ensure that they capture the underlying ageing process independently of chronological age [12]. The guidelines propose that ageing biomarkers must:

- Predict the rate of ageing better than chronological age, anticipating future physiological, cognitive, and physical decline across a range of age-sensitive domains.
- Monitor a basic process underlying ageing itself, rather than reflecting the effects of disease.
- Be testable repeatedly without harm to the subject, and valid in both humans and experimental animals to facilitate preclinical validation before human translation.

Given ethical and logistical barriers to perform longitudinal studies in humans, these biomarkers act as surrogates, accelerating research and enabling intervention development [9].

2.1.2 Classes of Biomarkers

Ageing biomarkers span multiple biological scales, each with distinct precision profiles, strengths, limitations, and applications [12]:

- **Targeted Molecular/Cellular** biomarkers measure individual, well characterised ageing processes (e.g. telomere length, senescence markers) [13].
- **Omics-based** biomarkers simultaneously profile thousands of molecules genome-wide (DNA, RNA, proteins) [12].

- **Imaging** biomarkers estimate tissue age from scans (brain MRI, retinal images) [14].
- **Clinical/phenotypic** biomarkers use routine measures like grip strength or blood pressure [15].

Table 2.1: Biological age estimation classes.

Class	Precision	Advantages	Disadvantages	Key Applications
Targeted Molecular/Cellular	r=0.4–0.6	Direct cause of ageing known; cheap	Too simplistic	Telomere length predicts mortality risk
Omics-based	r=0.85–0.95	Rich feature space; systemic view	Costly arrays	DNA methylation clocks predict higher risk of dying from any cause
Imaging	r=0.7–0.9	Direct tissue view	Expensive scanners	Brain age gap predicts dementia
Clinical	r=0.5–0.7	Easy clinic data	Late disease signals	Grip strength predicts 4-year mortality

r being the correlation coefficient.

Transcriptomic clocks (gene expression) are particularly well suited to blood-based ageing prediction, offering competitive predictive accuracy alongside sensitivity to dynamic regulatory signals, making them suitable for machine learning pipelines [7].

2.2. Transcriptomics and Gene Expression

The transcriptome represents the complete set of RNA molecules actively produced by a cell, tissue, or organism at a given time, acting as a dynamic indicator of gene activity that integrates genetic programming, epigenetic control, and environmental influences [16]. Figure 2.1 illustrates how transcriptomics relates to other omics layers that collectively characterise biological systems.

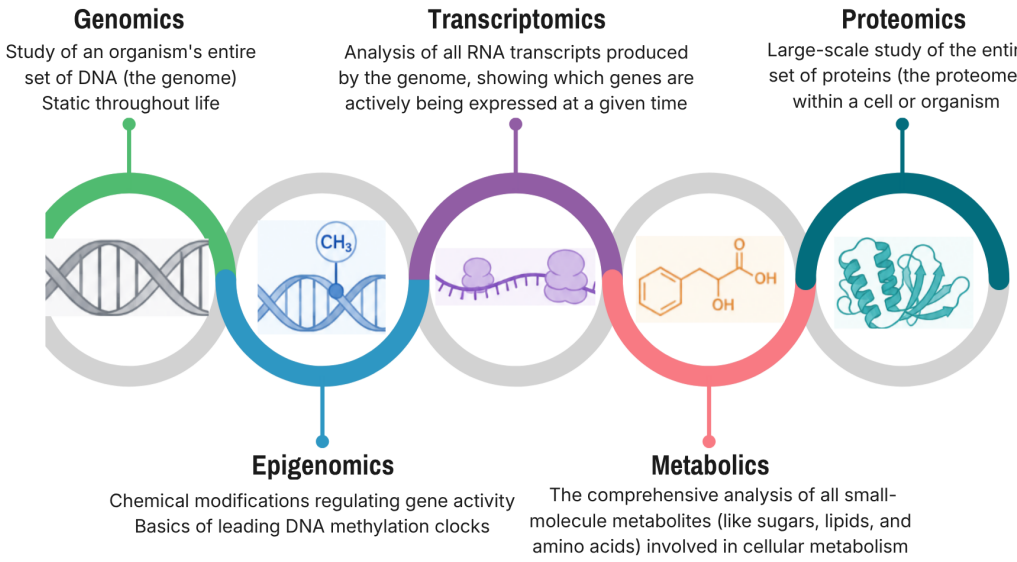


Figure 2.1: Overview of the main omics layers and their biological scope.

The systematic analysis of the transcriptomic landscape is made possible through the use of high-throughput technologies that quickly measure the abundance of thousands of RNA transcripts per sample. Computational processing of the raw sequencing data converts each gene's reads into a single numerical expression value. Several units have been used for this purpose. Counts Per Million (CPM) corrects for sequencing depth and serves primarily as an intermediate value in further calculations. Reads Per Kilobase per Million (RPKM) and its paired-end equivalent, Fragments Per Kilobase per Million (FPKM), additionally normalise for gene length and were the standard in early RNA-seq studies. However, because their denominators depend on the composition of the sample's own transcriptome, they are not strictly comparable across samples [17]. Transcripts Per Million (TPM) addresses this limitation by normalising for gene length first and then for sequencing depth, ensuring that values sum to the same total in every sample and making cross-sample comparisons more reliable [17, 18]. TPM is now the preferred unit in bulk RNA-seq studies and is the unit in which the Genotype-Tissue Expression project (GTEx) data used in this thesis are provided. The result is a high-dimensional dataset where each sample is represented as a vector of expression values, one per gene. These profiles reveal the functional state of cells by showing which biological pathways are active (e.g., inflammation, DNA repair, metabolism), which pathways are stressed (e.g., unfolded protein response, oxidative stress), and which processes are dysregulated (e.g.,

cancer signaling, neurodegeneration) [19]. Unlike genomic sequences (a static "instruction manual" that remains nearly identical throughout life), transcriptomic profiles capture real-time execution, revealing which genes are actively transcribed, to what extent they are expressed, and how their activity responds to internal signals and external conditions. This dynamic nature makes transcriptomics uniquely suited to ageing research, where the gradual accumulation of pathway dysregulation, stress responses, and metabolic reprogramming defines the biology of ageing [16].

2.2.1 Measurement Technologies

Transcriptomic profiling evolved from older microarray technology, which measured only a list of known genes with limited sensitivity to extreme expression levels, to RNA sequencing (RNA-seq), the current gold standard [17]. RNA-seq measures all genes without prior knowledge, distinguishes protein variants (isoforms) encoded by the same gene, and accurately quantifies expression across a wide range of levels. A typical bulk RNA-seq pipeline includes:

- **RNA extraction from cells:** Total RNA is isolated from whole blood or tissue using commercial kits (e.g., *RNeasy, TRIzol*) that remove contaminants and stabilize fragile RNA molecules. This proves particularly challenging for blood, which contains heterogeneous cell types (leukocytes, erythrocytes) with differing RNA content and stability [17].
- **cDNA library preparation:** RNA is reverse-transcribed into stable complementary DNA (cDNA) fragments, tagged with adapters for sequencing compatibility, and amplified via Polymerase Chain Reaction (PCR) to create sufficient material. Ribosomal RNA (80-90% of total RNA) is removed to enrich messenger RNA [17].
- **High-throughput sequencing:** Illumina *NovaSeq/NextSeq* platforms generate 50-150 base pairs paired-end reads at 20-100 million depth per blood sample, ensuring statistical power to detect low expressed genes (e.g., < 1 read per million) while capturing highly abundant transcripts [17].
- **Read Alignment:** Sequencing reads are mapped to the human reference genome (e.g., *GRCh38.p14*) using splice-aware aligners (*STAR, HISAT2*) that accommodate intron-exon boundaries. Multi-mapping reads and contaminants (bacteria, mouse DNA) are filtered, yielding 85-95% uniquely aligned reads for quantification [17].

- **Gene-level quantification:** Converts aligned reads into gene-level counts. Tools such as *Salmon* and *Kallisto* use pseudoalignment, a computationally efficient approximation that maps reads to a transcript index without producing a full genome alignment. Alternatively, STAR, beyond its role as a splice-aware aligner, returns true genome alignments and can also be paired with *featureCounts* or *STARsolo* for quantification. In high-quality blood samples, these methods typically detect tens of thousands of expressed genes [17].

- **Normalization:** Library size differences and RNA composition biases are corrected using methods like Transcripts Per Million (TPM), *DESeq2* median-of-ratios [20], or *edgeR* Trimmed Mean of M-values (TMM), ensuring cross-sample comparability [17].

- **Quality control:** *FastQC* [21] evaluates sequencing read quality, finds leftover adapter sequences for posterior removal (trimming) and shows per sequence G/ C content for one sample; *MultiQC* provides similar metrics and visualizations across multiple samples; Degraded RNA (samples with RIN <6.5), >50% duplicate reads, or extreme library sizes are flagged/removed to prevent technical artifacts from confounding age signals [17].

When applied to whole blood, this pipeline begins with samples obtained from a simple blood draw (venipuncture) rather than more invasive tissue biopsies, while capturing systemic immune ageing (”inflammaging”) through leukocyte transcriptomes [22].

2.2.2 Relevance of Ageing Biomarkers

Large-scale meta-analyses across human cohorts, model organisms, and diverse tissues have established robust transcriptomic signatures of ageing: increased activity (upregulation) of genes involved in inflammation, innate/adaptive immunity (interferon signaling), and connective tissue repair (extracellular matrix remodelling), alongside decreased activity (downregulation) of genes regulating mitochondrial energy production, ribosome function and protein synthesis pathways [10, 11]. In blood specifically, immune and inflammatory pathways dominate the ageing signature, reflecting the progressive remodelling of the immune system with age, while metabolic and mitochondrial decline is more pronounced in tissues such as brain and muscle [7, 23]. Because blood transcriptomes excel at capturing immune system ageing (immunosenescence), they are ideal candidates for systemic ageing clocks.

2.2.3 Challenges

Transcriptomic ageing clocks face three principal challenges requiring sophisticated computational mitigation:

- **Biological heterogeneity:** Gene expression in blood varies substantially due to factors independent of ageing, including circadian rhythms [24]; subclinical infections, which transiently activate interferon-stimulated gene expression programmes in circulating leukocytes [25]; and medication use, with commonly prescribed drugs such as statins broadly altering cholesterol and immune pathway gene expression in blood [26].
- **Technical artifacts:** Batch effects (systematic differences in gene expression measurements arising from samples being processed at different times, locations, or with different reagents or instruments) represent one of the most critical confounders in high-throughput transcriptomic studies and can account for a substantial proportion of total expression variance [27]; RNA quality, quantified using the RNA Integrity Number (RIN), also degrades progressively after blood draw, making rapid processing and rigorous quality control essential to prevent technical artifacts from confounding biological signals [28].
- **High dimensionality ($p \gg n$ problem):** thousands of genes vs 100-1000 samples risks overfitting curse of dimensionality [29].

These challenges demand preprocessing pipelines that address or mitigate their effects. In this work, we have the following preprocessing steps: RNA quality filter, batch correction, differential gene expression and feature selection.

2.3. Transcriptomic Ageing Clocks

Transcriptomics ageing clocks use gene expression data and supervised learning to estimate chronological or biological age, typically from whole-blood RNA profiles. In contrast to DNA methylation clocks that rely on relatively stable epigenetic markers, transcriptomic clocks operate at the gene expression level and are therefore thought to be more sensitive to short-term physiological and environmental changes such as infection, inflammation, or drug exposure [7, 30].

2.3.1 Definition

Transcriptomic ageing clocks are supervised machine learning models that predict age from high-dimensional gene expression profiles generated by microarrays or RNA-seq [7, 30]. Each sample's transcriptome, represented as thousands of normalized gene expression values for each transcript, is treated as an input feature vector, and regression algorithms such as elastic net, LASSO, random forests, or ensemble methods are trained to map these features to chronological age [5, 7].

The difference between predicted age ($\hat{y}$) and chronological age (y), often termed age acceleration (ΔAge = $\hat{y} - y$), represents deviations that are possibly associated with biological ageing, where positive values indicate accelerated and negative values decelerated ageing [6]. Figure 2.2 illustrates the full transcriptomic ageing clock pipeline, from blood collection to biological age estimation.

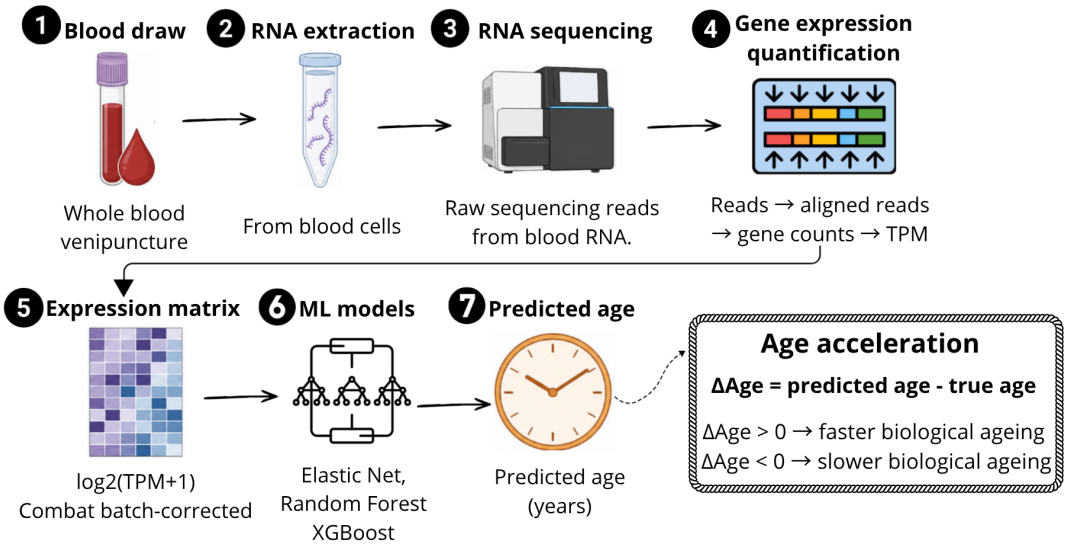


Figure 2.2: Transcriptomic ageing clock pipeline.

Model performance is typically measured using the coefficient of determination (R^2), the correlation between predicted and chronological age, and the Mean Absolute Error (MAE) [31]. Robust clocks aim to retain performance across independent cohorts, sequencing platforms, and transcript quantification pipelines, while maintaining some biological interpretability through the analysis of age-associated pathways and sets of related genes as opposed to individual genes [5, 7].

2.3.2 Early Transcriptomic Ageing Models

Early transcriptomic ageing models, developed roughly between 2010 and 2018, demonstrated that gene expression profiles could predict chronological age with moderate accuracy but lower performance than contemporary DNA methylation clocks [5, 32]. In a landmark meta-analysis of 14,983 peripheral blood samples, *Peters et al.* identified 1,497 age-associated genes and built a transcriptomic age predictor with correlations ranging between 0.6 and 0.7 and mean absolute errors of 7 to 8 years across cohorts [7]. Subsequent work in human dermal fibroblasts showed that RNA-seq profiles from 133 donors could be used to train ensemble machine learning models that predicted donor age with a median error of about 4 years. Furthermore, the models detected accelerated ageing in progeria patients, supporting the biological relevance of transcriptomic age acceleration [30].

Methodologically, these early clocks typically relied on differential expression analysis (for example, using limma) to preselect several hundred to a few thousand age-associated genes, followed by penalized regression such as LASSO or elastic net with internal cross-validation restricted to a single cohort [33, 34]. However, performance remained modest (correlations generally 0.6–-0.8, MAE 4–-10 years), and models often showed limited generalization to independent datasets due to batch effects, platform differences, and the high dimensionality of gene expression data relative to sample size [7, 27, 35]. These limitations motivated the development of larger bulk RNA-seq datasets processed with improved normalisation and batch-correction pipelines, as well as more advanced modelling strategies, leading to the second generation of transcriptomic clocks described in Section 2.3.3 [7, 36].

2.3.3 Modern Blood-Based Transcriptomic Ageing Clocks

Modern blood-based transcriptomic ageing clocks represent a second generation of models that leverage large-scale RNA-seq datasets, advanced preprocessing pipelines, and more sophisticated feature engineering to substantially improve upon the performance of early microarray-based predictors (Section 2.3.2). These advances have closed the gap between transcriptomic clocks and epigenetic models while maintaining sensitivity to dynamic physiological states [36, 37].

A prominent example is the *BiT*-age clock [36], which introduced a binarisation strategy that converts thousands of continuous gene expression values into high/low binary states.

When applied to human blood RNA-seq data from GTEx, elastic net regression achieved correlations around 0.9 and median absolute errors below 3 years on held-out test samples and showed strong cross-platform performance when evaluated on independent microarray cohorts [36]. Related work [38] has constructed clocks from repetitive element expression in blood RNA-seq, aggregating transcript levels of age-associated transposable elements to achieve correlations of 0.85–-0.90 and MAE of 4–-5 years across multiple Gene Expression Omnibus (GEO) cohorts, illustrating how pathway or element level aggregation can reduce dimensionality while retaining signal.

Other approaches combine dimensionality reduction with clinical features. Clinical ageing clocks based on principal component analysis, for example, compress multi-omic or clinical blood profiles into a small number of components and then train age prediction models on these lower dimensionality representations, obtaining correlations around 0.9 with good generalisability across health conditions [37].

2.3.4 Tissue-specific and Multi-tissue Transcriptomic Clocks

Transcriptomic ageing signatures exhibit both conserved and tissue-specific components [10, 39]. Large-scale meta analyses across human cohorts, model organisms, and multiple tissues consistently report the same conserved transcriptomic ageing pattern described in Section 2.2.2 [10, 39]. However, the magnitude and specific genes involved vary substantially across tissues: inflammatory signatures dominate in blood (reflecting immunosenescence), while metabolic decline is more pronounced in brain, muscle, and adipose tissue [7]. Clocks trained on a single tissue type are optimised for that tissue's specific expression patterns, and their generalisability to other tissues remains limited [4, 7]. Multi tissue or organism-wide transcriptomic clocks, by contrast, aim to capture systemic ageing signals at the cost of some tissue specificity [31]. Recent advances in single-cell and spatial transcriptomics further reveal that ageing rates can differ even between neighbouring cell types within the same organ, underscoring the value of context-specific clock development [40].

2.3.5 Comparison with other Ageing Clocks

Transcriptomic clocks generally exhibit slightly lower raw predictive accuracy than leading DNA-methylation clocks but greater responsiveness to acute physiological changes (Table 2.2). In blood, second-generation epigenetic clocks (e.g., *Horvath, Hannum, GrimAge*) routinely achieve correlations above 0.90–0.96 and MAE of 2–-3.5 years on external

test sets. Well-designed transcriptomic clocks, including *BiT*-age and repetitive-element models, report correlations of 0.85–-0.95 and MAE of 2.5–-6 years, approaching epigenetic performance in favourable settings while remaining more sensitive to short-term immune modulation. The most recent multi-species transcriptomic clock achieved human tissue correlations of 0.91–-0.94 [11], approaching the performance of leading epigenetic clocks. Clinical and phenotypic clocks, based on routine blood tests (e.g., creatinine, albumin) or physical measures (e.g., gait speed, grip strength), typically yield more modest correlations of 0.60–-0.80 and higher MAE, but excel in accessibility and cost [7, 37].

Table 2.2: Biological age clock types.

Clock type	**r (external test)**	**MAE (years)**	**Strengths**	**Limitations**	**Examples**
DNA methylation	0.90–0.96	2–3.5	High accuracy, stability	Less dynamic	*Horvath* [5], *Hannum* [32]
Transcriptomic	0.85–0.95	2.5–6	Intervention sensitivity, immune insight	Batch effects, heterogeneity	*Peters* [7], *BiT*-age [36], *Sood* [41], *Tyshkovskiy* [11]
Clinical/phenotypic	0.60–0.80	6–10	Cheap, scalable	Late signals	*PhenoAge* [6]

2.4. Machine-Learning Methods for Age Prediction

Machine learning models for age prediction map biological information, such as transcriptomic profiles described in Section 2.2, to estimates of chronological or biological age, usually through supervised regression methods. These approaches span classical statistical models, tree based ensembles, and deep neural architectures, with ongoing advances in feature engineering and calibration to address the challenges of batch effects, heterogeneity, and overfitting, as highlighted in Sections 2.2.3 and 2.3 [42].

2.4.1 Data Preprocessing and Feature Engineering

Preprocessing of gene expression data is critical for transcriptomic age prediction, since it converts raw RNA-seq counts into stable, comparable features that mitigate technical noise while preserving biological signal. Standard pipelines include log-transformation or variance-stabilising normalisation (e.g., *DESeq2 VST*) [20] steps, followed by batch

correction using *Combat* [35], Surrogate Variable Analysis (SVA) [43] or removal of unwanted variation (*RUVSeq*) [44]. Feature engineering then reduces the thousands of genes space through strategies such as:

- **Dimensionality reduction:** Principal Component Analysis (PCA) or autoencoders compress profiles into 10–200 components capturing >95% variance [45].
- **Sparsity-inducing penalisation:** LASSO/elastic net select hundreds of age-associated genes during model fitting [34].
- **Non-linear transformations:** Binarisation (high/low expression thresholds) or gene-set aggregation (e.g., inflammatory pathway scores) [36].
- **Biological priors:** Cell composition deconvolution (e.g. *CIBERSORTx*) or pathway scoring (e.g. *GSVA*) incorporate cell-type or functional context [46].

2.4.2 Classical ML Models

Classical machine learning models remain standard tools for age prediction from transcriptomic data because they handle high-dimensional features relatively well and can be tuned to different signal-to-noise ratios. In blood-based ageing clocks, they are typically applied on top of the preprocessing and feature-engineering steps described in Section 2.4.1, using cross-validation to select regularisation strength and other hyperparameters [42].

Penalised linear models such as LASSO, Ridge, and Elastic Net are particularly effective when sample sizes are modest (hundreds to a few thousand individuals) and only a subset of genes carries most of the age-related signal. Their regularisation terms control overfitting in the $p \gg n$ setting and can produce sparse or shrunk coefficient vectors that are easier to interpret biologically, making them natural baselines for transcriptomic ageing clocks [34, 47, 48].

Tree-based ensemble methods, including Random Forest, Gradient Boosting, XGBoost, and LightGBM, are advantageous when age–gene relationships are non-linear or involve higher order interactions, for example threshold behaviour in inflammatory or stress response pathways. These models can exploit complex structure in high-dimensional spaces and often achieve strong predictive performance, provided there are sufficient samples to tune hyperparameters such as tree depth, learning rate, and number of trees without overfitting [49–52].

Feed-forward neural networks become competitive as cohort sizes increase and when feature engineering (e.g. principal components or pathway scores) has already reduced dimensionality, allowing the network to model residual non-linear patterns in gene expression with manageable risk of overfitting [53].

In this thesis, these complementary properties motivate a comparative evaluation of penalised linear models, tree-based ensembles, and neural networks for blood-based transcriptomic age prediction on the GTEx v11 whole blood cohort, with a focus on robustness, generalisation, and biological interpretability.

2.4.3 Deep and Knowledge-Primed Neural Networks

Deep learning methods use multi-layer neural networks to learn complex, non-linear representations of gene-expression data. In transcriptomics, fully connected neural networks, convolutional architectures, and autoencoders have been applied to predict age and other phenotypes from bulk and single-cell expression matrices, often starting from dimension-reduced features or gene-set scores to mitigate the $p \gg n$ problem [53]. When sufficiently large cohorts are available, these architectures can match or exceed the accuracy of tree-based ensembles, but they are sensitive to hyperparameter choices and require careful regularisation (dropout, weight decay, early stopping) to avoid overfitting [53].

Knowledge-primed neural networks incorporate prior biological structure into the network design or input representation, aiming to improve interpretability and robustness [54, 55]. Examples include architectures where nodes or edges correspond to known pathways or gene regulatory networks, graph neural networks built on co-expression or protein–protein interaction graphs, and models that operate on pathway or cell-type scores rather than individual genes. Such approaches can highlight biologically meaningful subnetworks associated with ageing while still leveraging the expressive power of deep learning. However, they require more data compared with penalised linear models, and their advantages are most pronounced in settings with large multi-omics or single-cell datasets and strong prior knowledge of pathway structure, conditions that are still emerging for blood-based ageing clocks [56].

2.4.4 Evaluation and Calibration

To rigorously evaluate ageing clocks, all metrics must be reported on unseen external test sets in addition to training/validation splits, thus ensuring realistic performance estimates. Core performance metrics include [57]:

- **Pearson correlation (r):** Measures linear association between predicted age ($\hat{y}$) and chronological age (y); $r = 0.94$ means predictions preserve 94% of age ranking.
- **Coefficient of determination (R^2):** Quantifies variance explained by the model; ($R^2 = 0.85$) indicates 85% of age variance is captured, leaving 15% irreducible error (biological + technical noise). Superior to r for comparing model capacity.
- **Mean Absolute Error (MAE):** Reports average prediction error in biological years; MAE=3.2 years means predicted age deviates 3.2 years from true age on average. Most intuitive for clinical interpretation (”off by 3 years”).
- **Age acceleration ($\Delta Age = \hat{y} - y$)** distributions should have mean close to zero and approximate normality across age groups so that positive values can be interpreted as accelerated ageing and negative values as decelerated ageing.

In addition, performance should be evaluated on independent external datasets using the same pipeline. Well-fitted pipelines that are robust to batch effects and population differences maintain high R^2 and low MAE even in this scenario.

2.5. Multi-Omics and Multimodal Ageing Clocks

The single-omic clocks described in the preceding sections capture only one dimension of the ageing process (Figure 2.1). Several recent studies have therefore proposed multi-omics and multimodal ageing clocks that combine information from DNA methylation, transcriptomics, proteomics, metabolomics, clinical biomarkers, and imaging, aiming to improve both predictive accuracy and biological interpretability [4, 58].

Different omic layers capture complementary aspects of ageing: epigenetic clocks reflect cumulative regulatory change; transcriptomic clocks capture dynamic pathway activity and immune status; proteomic and metabolomic profiles report on phenotypes closer to organ function; and clinical or imaging markers summarise overt physiological decline [4]. Comparative analyses across fifteen omics clocks spanning glycans, proteins, metabolites, lipids, DNA methylation, and clinical measures have shown that while approximately

95% of chronological age variance is shared across clock types, each modality still contributes independent biological signal not captured by the others [59]. This finding simultaneously motivates multi-omics integration and raises the question of whether the marginal gain in predictive accuracy justifies the substantial increase in data requirements and analytical complexity. Recent large-scale efforts have attempted to resolve this tension by distilling multi-omics information into a single deployable readout. OMICmAge, for example, incorporates proteomic, metabolomic, and clinical data through epigenetic biomarker proxies, yielding a clock that is quantifiable from DNA methylation alone while retaining the biological breadth of multiple data layers; applied across discovery and validation cohorts totalling over 45,000 participants, it demonstrated stronger associations with incident chronic disease and mortality than established single-omic epigenetic clocks [8].

Despite their promise, multi-omics clocks face substantial practical limitations. Cohorts with matched data across all modalities are rare, and the effective sample size is restricted to the intersection of individuals with complete measurements, compounding the high-dimensionality problem already present in single-omic settings (Section 2.2.3). Each data type carries distinct normalisation requirements and batch structure (methods appropriate for RNA-seq cannot be naively applied to methylation arrays or metabolomic profiles) and misaligned scaling can cause one modality to dominate the combined predictor [60]. Empirical evidence from GTEx-based multi-modal age prediction studies illustrates these constraints directly. When integrating transcriptomics, DNA methylation, telomere length, and histological imaging from the same cohort, combining modalities yielded only marginal performance improvements over the best single modality alone, with the limited intersection of samples across data types identified as the primary limiting factor [61, 62]. Multi-omics assays are also expensive and logistically demanding, limiting clinical feasibility relative to single-omic readouts, and multimodal architectures tend to be harder to interpret, complicating attribution of age acceleration to specific biological mechanisms.

2.6. Data Resources and Cohorts

Developing and evaluating transcriptomic ageing clocks requires access to large, well-characterised datasets that provide high-quality gene expression measurements alongside phenotypic metadata. This section summarises the main data resources and large-scale cohorts that have shaped the field, including those used in the key studies reviewed in this chapter.

2.6.1 GTEx and Multi-modal Omics Resources

The Genotype-Tissue Expression project (GTEx) has become the central reference resource for transcriptomic ageing studies, providing RNA-seq profiles from 54 non-diseased post-mortem tissue sites collected from nearly 1,000 human donors, totalling more than 17,000 samples profiling over 58,000 genes [63]. For several tissues, additional data modalities are available alongside gene expression, including histological images and, for selected tissues, DNA methylation arrays generated by companion initiatives. GTEx uses standardised collection, processing, and quality control protocols across all sites, and donor metadata including age, sex, body mass index, and cause of death are publicly available.

2.6.2 Large-Scale Cohort Studies for Blood transcriptomic ageing

Beyond GTEx, transcriptomic ageing research draws on a set of large, well-established population cohorts and public data repositories that have collectively generated the microarray and RNA-seq datasets used to discover, train, and validate blood-based ageing signatures. Table 2.3 summarises the most important of these resources.

Table 2.3: Transcriptomic datasets.

Dataset/-Cohort	Country	N (blood)	Platform	Age Range	Key Contributions	Access
Framingham Heart Study (FHS) [64]	USA	∼5,600	Microarray	40–80	Largest single contributor to *Peters et al. 2015* meta-analysis; multi-generational design	dbGaP
Rotterdam Study [65]	Netherlands	∼881	Microarray	45+	Major European replication cohort; frequently used for external validation	dbGaP / GEO
InChianti [66]	Italy	∼698	Microarray	21–100	Broad age range including very elderly; contributed to *Peters et al. 2015*	GEO
Kora [67]	Germany	∼993	Microarray	25–74	Population-based Bavarian cohort; contributed to *Peters et al. 2015*	GEO
SHIP-TREND [68]	Germany	∼970	Microarray	20–80	Large northern German cohort; contributed to *Peters et al. 2015*	GEO
MuTHER [69]	UK	∼856 (female twins)	Microarray	Adults	Multi-tissue including blood-derived immune cell lines; key for tissue-specific ageing comparisons	GEO
UK Biobank [70]	UK	500,000+ (ongoing)	RNA-seq	40–69	Largest population cohort; RNA-seq progressively released; future reference dataset	Controlled access
GEO / *ArrayExpress* [71, 72]	International	Varies	Microarray + RNA-seq	Varies	Primary public repositories for depositing and accessing blood transcriptomic studies	Open controlled

While UK Biobank data has not yet been explored specifically for developing transcriptomic ageing clocks, their work represents the next generation of resources in this field, with its scale expected to enable substantially more robust clock development than any previously available cohort. GEO and *ArrayExpress* together serve as the central archives through which these and hundreds of smaller studies are made accessible to the research community, enabling meta-analyses, cross-platform validation, and replication of published signatures.

2.6.3 Public Repositories

Beyond cohort-specific datasets, several public repositories provide access to the raw and processed data used in molecular ageing studies. The Gene Expression Omnibus (GEO) [71] and *ArrayExpress* host the majority of published microarray and RNA-seq ageing datasets and are the primary sources for external validation cohorts in transcriptomic clock studies. The Sequence Read Archive (SRA) [73] and European Nucleotide Archive (ENA) [74] provide raw sequencing reads enabling reprocessing with standardized pipelines. Disease focused repositories such as The Cancer Genome Atlas (TCGA) [75] and the Genomic Data Commons (GDC) [76], while primarily oncology oriented, have contributed bioinformatics tools and practices now widely applied in ageing research. For future transcriptomic clock development, the UK Biobank, which is progressively releasing blood RNA-seq data from its 500,000 participant cohort, represents the most significant emerging resource, offering age ranges and sample sizes that surpass currently available datasets.

2.7. Limitations

Despite substantial methodological progress, several fundamental challenges continue to limit the accuracy, generalisability, and clinical utility of blood-based transcriptomic ageing clocks [4].

2.7.1 Noise, Variability, and Data Imbalance in Blood Gene Expression

Blood gene expression is highly variable, affected by technical factors including RNA quality, sequencing depth, batch effects, and library preparation protocol, as well as biological factors such as subclinical infections, medication use, circadian rhythms, and lifestyle. Even after normalisation and batch correction, residual noise can confound age related signals and reduce the stability of selected gene sets, yielding clocks that perform well

within a single dataset but degrade when applied to data generated under different conditions or platforms [31, 36]. An additional challenge, relevant to the GTEx blood dataset used in this thesis, is the imbalance of age distributions in real-world cohorts. The GTEx blood cohort used in this work is heavily concentrated in the 50–70 year range, with very few donors below 30 years of age (the full age distribution is shown in Figure 3.5, Section 3.1.1). This pattern reflects both the post-mortem nature of GTEx sample collection and the general difficulty of obtaining younger donors in population-based biobanks.

The practical consequence for age prediction is systematic bias: models trained on such distributions tend to overestimate age in younger individuals and underestimate it in older ones, because the loss function is dominated by the age range where most training samples lie [61, 62]. Addressing this imbalance requires explicit resampling or re-weighting strategies to ensure the model learns from the full age range rather than fitting only the majority [77].

2.7.2 Tissue Specificity and Limited Transferability

Ageing processes differ across organs and cell types, and transcriptomic signatures are strongly tissue-specific. Blood clocks mostly capture immune and systemic aspects of ageing and may not reflect changes predominant in other tissues [31]. Moreover, models trained in one cohort or platform often show reduced accuracy when transferred to other populations or technologies, limiting the idea of a single universal transcriptomic clock [4, 7, 11].

2.7.3 Sample Size Limitations and Absence of Longitudinal Data

Beyond age distribution imbalance, sample sizes in transcriptomic ageing studies remain small relative to the number of measured genes, increasing the risk of overfitting and unstable feature selection. Like many transcriptomic datasets, the GTEx blood dataset comprises approximately 800 samples profiled over 58,000 genes, and thus suffers from the curse of high-dimensionality. Longitudinal data with repeated blood sampling across the decades spanning the lifespan of a large group of donors remain rare, making it difficult to distinguish model errors from true divergence between biological and chronological age [4].

2.7.4 Challenges in Causal Interpretation

Current transcriptomic clocks are optimised for prediction rather than causal inference. Strong contributions of specific genes or pathways to age prediction do not imply that those genes drive ageing, or that modifying their expression would alter biological age. Instead, clocks aim to capture the downstream consequences of disease processes, environmental exposures, or shifts in blood cell type composition [4]. Approaches that prioritise interpretability, such as sparse feature selection methods that produce small, named gene signatures, partially address this limitation by enabling cross-validation against published literature, even if they cannot establish causality.

2.7.5 Reproducibility and Absence of Standardised Pipelines

There is no standard workflow for building blood transcriptomic ageing clocks. Studies differ in preprocessing, normalization, covariate handling, feature selection, model choice and evaluation, and often provide limited external validation and code availability [31]. This limits the reproducibility and fair comparison of methods, slowing progress towards clinically robust and widely accepted blood-based aging clocks [9].

2.8. Positioning and Contributions of This Thesis

The work presented in this thesis addresses a gap in the transcriptomic ageing clock literature: the absence of reproducible, fully scripted pipelines that combine biologically motivated preprocessing with systematic model comparison and interpretable gene signature validation. While studies such as *Peters et al.* [7] established the feasibility of blood transcriptomic age prediction and BiT-age [36] demonstrated the performance gains achievable with large-scale RNA-seq data, neither provides a fully reproducible pipeline applicable to a new cohort, nor a systematic evaluation of the methodological choices that most influence prediction quality.

Building on prior multi-modal ageing studies from the same group [61, 62], this thesis develops such a pipeline using GTEx v11 whole blood RNA-seq data. By selecting Ridge-based gene ranking over dimensionality reduction approaches such as PCA, the pipeline preserves individual gene identities throughout, enabling cross-validation of the final 200-gene signature against published ageing signatures [7, 41] and protein-protein interaction network analysis. All models are trained and evaluated against chronological age, the only ground-truth label available in GTEx; biological age is subsequently operationalised

as age acceleration (ΔAge = $\hat{y} - y$, Section 2.3.1), positioning the work at the intersection of predictive modelling and biological discovery rather than treating accuracy as the sole objective.

The thesis directly addresses four research questions stated in Section 1.1, with particular focus on whether a compact, biologically interpretable gene signature consistent with established transcriptomic hallmarks of blood ageing can be extracted from a modest, age-imbalanced cohort. These contributions provide a methodologically rigorous and biologically grounded foundation for future cross-cohort validation and targeted panel design.

3. Methods

This chapter describes the methodology used to develop and evaluate a transcriptomic ageing clock from GTEx whole-blood RNA-seq data. The pipeline comprises data acquisition, preprocessing and quality control, train/test partitioning, differential gene expression analysis, feature selection and machine-learning model development. An overview of the complete workflow is presented in Figure 3.1.

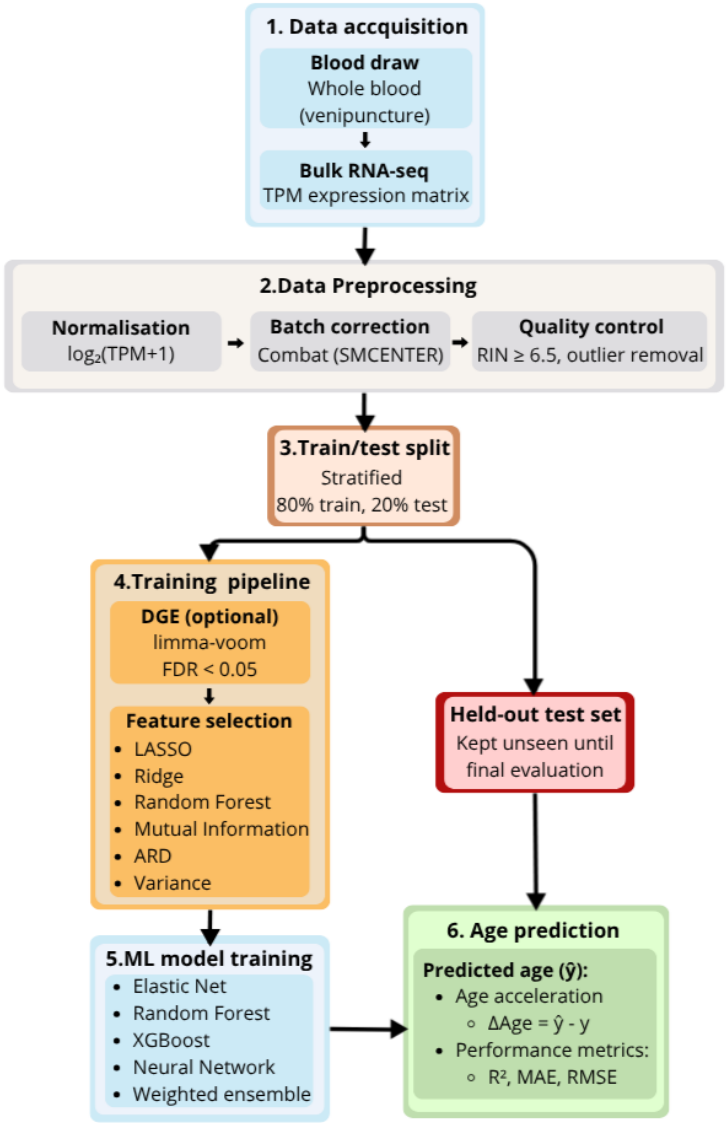


Figure 3.1: Overview of the transcriptomic age prediction pipeline developed.

3.1. Data Acquisition and Preprocessing

3.1.1 Dataset

Gene expression data were obtained from the Genotype-Tissue Expression project (GTEx) [63] project version 11, a large-scale multi-tissue resource providing bulk RNA-seq profiles across hundreds of human donors. The raw expression matrix was provided in Gene Cluster Text (GCT) format containing Transcripts Per Million (TPM) values for 74,628 genes. The GCT file was parsed directly in Python by skipping the two line format header and reading the tab delimited expression matrix, with gene identifiers retained as row indices and sample identifiers (SAMPID) as column headers. The resulting matrix was transposed to samples × genes orientation, with each row representing a sample and each column a gene, as required for downstream processing. Sample-level metadata and subject-level phenotypic annotations (including age) were retrieved from the accompanying GTEx v11 annotation files [78, 79]. Alignment across these three sources was performed using sample identifiers (SAMPID) and subject identifiers (SUBJID), with subject identifiers derived from sample identifiers by extracting the first two hyphen delimited fields (e.g., *GTEx-16NFA-0005-SM-9JGFN* $\rightarrow$ *GTEx-16NFA*). For whole blood, each donor contributes at most one sample, ensuring a one-to-one correspondence between sample and individual within this tissue. Only samples present in all three sources and carrying a valid age annotation were retained.

A metadata table was assembled containing sample and subject identifiers and available covariates, including sex (SEX), tissue descriptors (SMTS, SMTSD), RNA integrity number (SMRIN), sequencing centre (SMCENTER), and batch variables (SMGEBTCH, SMGEBTCHT). Sex annotations were harmonised across sources, as GTEx encodes sex numerically in some files (1 = Male, 2 = Female) and as strings in others. All values were normalised to a consistent ”Male”/”Female” encoding.

The distributions of key metadata covariates are shown in Figures 3.2, 3.3 and 3.4. The cohort is male-dominated (541 male, 262 female), reflecting GTEx donor demographics. Sequencing centre B1 accounts for the majority of samples (n = 568), followed by C1 (n = 229) and D1 (n = 6), a distribution that motivates the batch correction step described in Section 3.2. RNA integrity numbers are generally high (median RIN $\sim$8.3), with a small tail of lower quality samples, confirming the overall technical quality of the dataset.

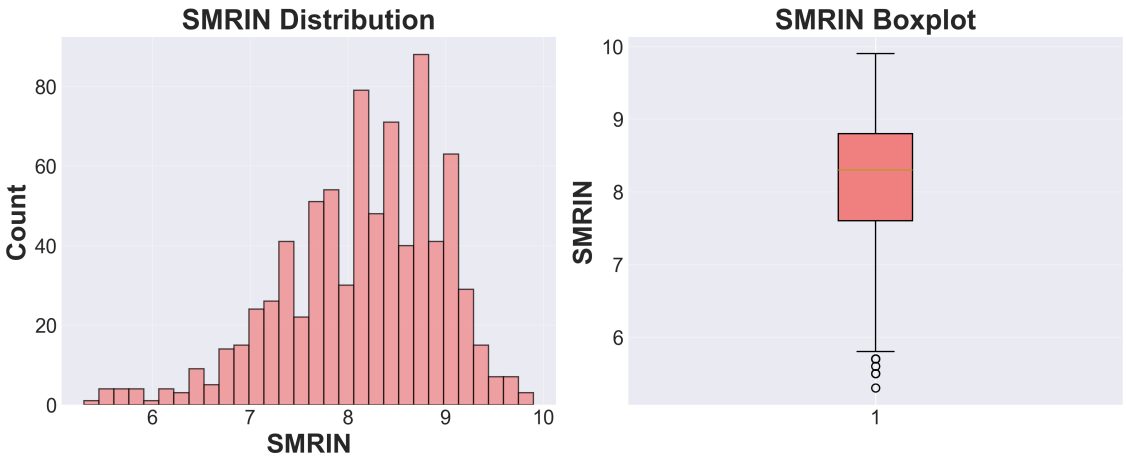


Figure 3.2: RIN distribution.

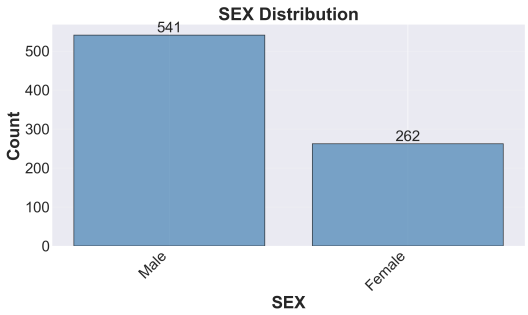


Figure 3.3: Sex distribution.

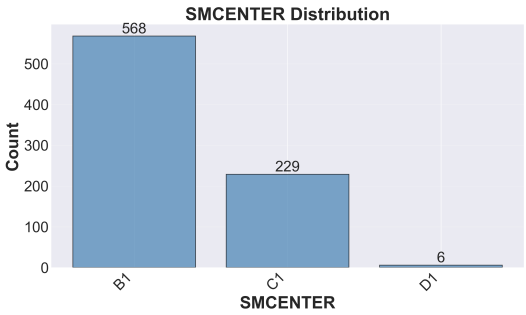


Figure 3.4: SMCENTER distribution.

This study focused exclusively on whole blood, selected for its clinical accessibility and its established use in transcriptomic ageing clock development. Tissue filtering was applied to the sample attributes file using the SMTS/ SMTSD tissue annotation fields, retaining samples labelled as ”Whole Blood”. This yielded 4,369 samples at the metadata level. Following alignment with the expression matrix and removal of samples lacking age annotations, the final cohort comprised 802 samples across 74,628 genes. The loss of 81.62% of samples relative to the metadata count reflects the fact that the GTEx expression matrix spans all tissues, and whole blood represents only a fraction of the total collected samples. This loss is therefore expected and does not indicate data quality issues.

The final cohort of 802 samples with ages ranging from 20 to 70 years (mean 52.8 $\pm$ 13.0, median 55.0, Q1 = 47.0, Q3 = 63.0) is illustrated in Figure 3.5. The age distribution is right-skewed towards middle and older age, reflecting the GTEx recruitment strategy, with a notable underrepresentation of donors below 40 years; a property that was addressed during model training via resampling (Section 3.7).

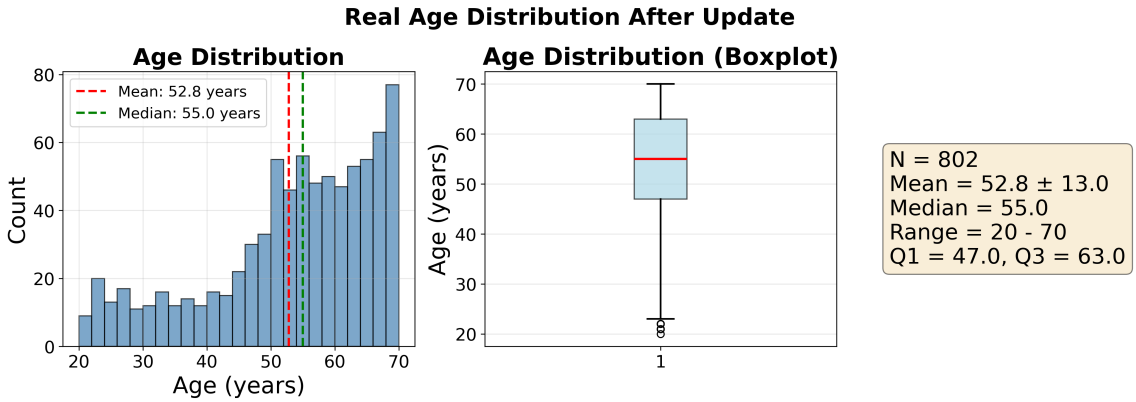


Figure 3.5: Age distribution. Left: histogram with mean (red) and median (green) indicated. Right: boxplot with summary statistics.

3.1.2 Expression Data Transformation and Quality Filtering

Raw TPM values were $\log_2$-transformed using the standard variance stabilising shift $\log_2(\text{TPM} + 1)$, which compresses the dynamic range of highly expressed genes and renders the distribution more suitable for linear modelling. The transformed values ranged from 0.00 to 19.61. The $\log_2$-transformed matrix was carried forward for all downstream steps; the raw TPM matrix was retained for reference only. The expression data exhibited high sparsity, with a median of 83.4% zero values per gene across samples, consistent with the known properties of bulk RNA-seq data. Metadata quality filtering was applied by discarding any sample for which more than 10% of metadata fields were missing. No samples were removed at this stage, confirming the completeness of the GTEx annotation files for the retained cohort. Genes with more than 10% missing values across samples were similarly flagged for removal, though no such genes were identified in the final matrix.

3.1.3 Exploratory Data Analysis (EDA)

Principal Component Analysis (PCA) was performed on the top 2,000 most variable genes, after standardisation to zero mean and unit variance, as an unsupervised quality check. The first two principal components explained 28.29% and 16.09% of variance respectively, with 10 components collectively capturing approximately 70% of total variance. As shown in Figure 3.6, samples coloured by sequencing centre (SMCENTER) reveal partial separation between centres B1 and C1, which motivates the batch correction step described in Section 3.2.

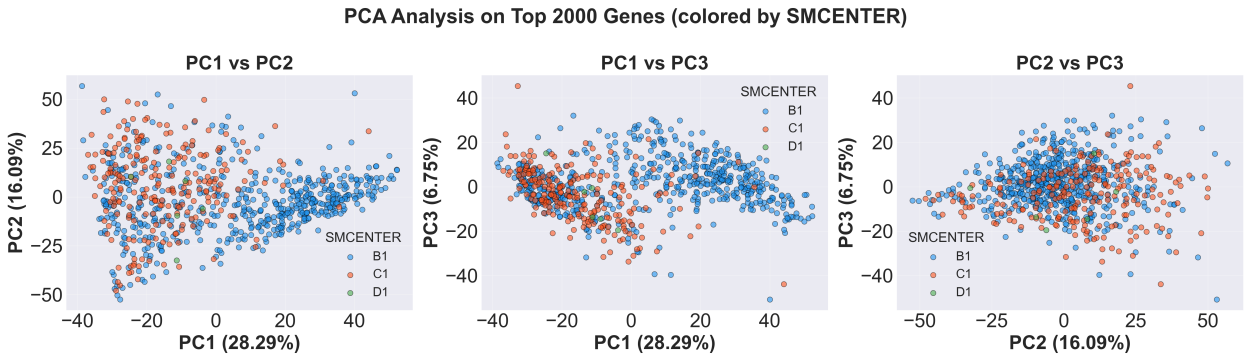


Figure 3.6: PCA analysis on top 2000 genes (colored by SMCENTER).

3.2. Batch Effect Correction and Quality Control

3.2.1 Batch Correction

Technical batch effects are a well documented source of systematic variation in large-scale RNA-seq datasets such as GTEx, where samples are processed across multiple sequencing centres [27]. Failure to account for these effects can confound downstream analyses and introduce centre-specific biases into age predictions. To address this, *ComBat* batch correction was applied to the $\log_2$-transformed expression matrix, using the sequencing centre identifier (SMCENTER) as the batch variable [35] . *ComBat* was selected because batch labels (sequencing centre, SMCENTER) were known a priori, making supervised correction both appropriate and more precise than unsupervised alternatives such as SVA and *RUVSeq* [43, 44], which infer hidden sources of variation directly from the data and can inadvertently remove biological signal when technical and biological variation are correlated. For each gene, *ComBat* estimates an additive shift (location) and a multiplicative scaling factor per batch, capturing how the mean and variance of that gene's expression differ across sequencing centres due to technical processing differences. These parameters are estimated using empirical Bayes shrinkage, which pools information across all genes to stabilize estimates, particularly important for the D1 centre, which contained only 6 samples [63].

Prior to correction, the expression matrix was temporarily transposed back to genes x samples orientation as required by *pycombat* [80], the Python implementation of the original R *ComBat* method. The matrix was inspected for missing and infinite values; none were detected. Zero-variance genes were identified and removed prior to correction, as *ComBat* cannot estimate batch parameters for genes with no variation across samples. A total of 3,810 such genes were removed, reducing the feature space from 74,628 to 70,818 genes. The three sequencing centres represented in the cohort were B1 (n = 567), C1

(n = 229), and D1 (n = 6), all satisfying *ComBat's* minimum batch size requirement of at least two samples per batch. A random seed (seed = 42) was fixed prior to all stochastic operations, and sample-label alignment was explicitly verified after each filtering step to ensure correspondence between expression matrices, metadata, and age labels was maintained throughout the pipeline. The corrected data ranged from −2.31 to 19.64. The presence of negative values is an expected artefact of *ComBat's* additive location-shift model, which subtracts batch-specific offsets without imposing a non-negativity constraint on the output. These values are biologically interpretable as genes with expression near zero whose batch offset subtraction results in a small negative residual, and they are routinely accepted in *ComBat* corrected datasets in the literature.

The effectiveness of batch correction was assessed by PCA on the top 2,000 most variable genes before and after correction. Prior to *ComBat*, as shown in Figure 3.6, centres B1 and C1 exhibited partial separation. Following correction and subsequent quality control filtering, as shown in Figure 3.7, samples from both centres showed substantially greater intermixing along PC1 and PC2, PC2 and PC3. However, residual structure remained after correction. This incomplete separation is consistent with the known confounding between technical and biological variation in GTEx, where sequencing centres may differ not only in processing protocols but also in donor recruitment demographics [63]. Residual centre-associated variation is therefore likely to reflect a combination of both sources and does not indicate a failure of the correction procedure.

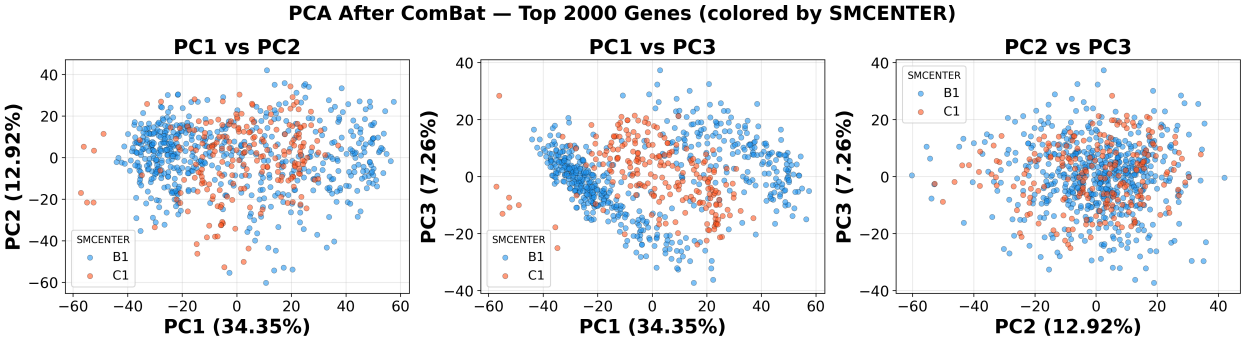


Figure 3.7: PCA after Combat - top 2000 genes (colored by SMCENTER).

3.2.2 Quality Control

Following batch correction, two sequential quality control filters were applied. First, samples with an RNA Integrity Number (RIN) below 6.5 were removed. RIN is a standardised measure of RNA sample quality scored from 1 to 10, where higher values indicate more intact RNA [28]. A threshold of RIN $\geq$ 6.5 was applied, slightly more conservative than the

minimum of 6.0 established by the GTEx project for postmortem tissue collection [81], and consistent with general recommendations for reliable bulk RNA-seq quantification [28]. Samples with low RIN values contain degraded RNA, which distorts gene expression profiles by reducing measured counts for longer transcripts and introducing systematic noise that is unrelated to biological age. This threshold eliminated 25 samples. Second, outlier samples were identified by computing the median expression value per sample across all genes and flagging those deviating by more than two standard deviations from the cohort mean; 20 samples were removed by this criterion. In total, 45 samples were excluded during quality control, yielding a final cohort of 757 samples across 70,818 genes. This matrix was carried forward to all downstream analyses. PCA after ComBat shown in Figure 3.7 was generated after quality control filtering and therefore reflects the final 757 sample cohort. Consequently, D1 sequencing centre does not appear in Figure 3.7, all six of its samples were removed during RIN and outlier filtering, leaving only B1 and C1 visible in the plot.

3.3. Train/Test Split

The 757 sample cohort was partitioned into a training set and a held-out test set using a stratified random split via *scikit-learn* [82] with a fixed random seed (seed=42) to ensure reproducibility. The test set comprised 20% of the data (n=152), with the remaining 80% reserved for training (n=605). In a stratified split, the data are divided such that the proportion of samples from each group is maintained in both partitions, unlike a simple random split which does not guarantee this balance. Since age is a continuous variable, stratification requires it to be discretised into groups first. Five quantile-based age bins were used, meaning the age range was divided into five intervals each containing approximately the same number of samples, regardless of where the age boundaries fell. This approach is preferable to equal-width intervals because the GTEx age distribution is skewed towards older donors, which would result in very few samples in the youngest equal-width bins and therefore unreliable stratification. The five bins spanned the age intervals [20.0, 42.2[, [42.2, 52.0[, [52.0, 58.0[, [58.0, 65.0[and [65.0, 70.0] years, containing 152, 160, 144, 172, and 129 samples respectively, confirming approximately uniform stratification across the age range.

The equivalence of the resulting partitions was assessed using two complementary statistical tests, chosen to capture different aspects of distributional similarity. A two-sample t-test was applied to compare the mean ages of the training and test sets. This test asks

whether the average age in the two sets is statistically different, producing a p-value under the assumption that the data are approximately normally distributed. A p-value of 0.7198 indicated no statistically significant difference in mean age between sets. A Kolmogorov-Smirnov (KS) test was additionally applied as a non-parametric complement, since, unlike the t-test, it makes no assumption about the shape of the distribution and instead compares the full empirical distributions of the two sets, capturing potential differences in spread, skewness, and tails that the t-test would not detect. A KS statistic of 0.046 with p = 0.9467 further confirmed that the train and test sets are drawn from statistically indistinguishable age distributions (train: 52.7 $\pm$ 13.0 years, range 20–70; test: 52.3 $\pm$ 13.1 years, range 21–70), as shown in Figure 3.8

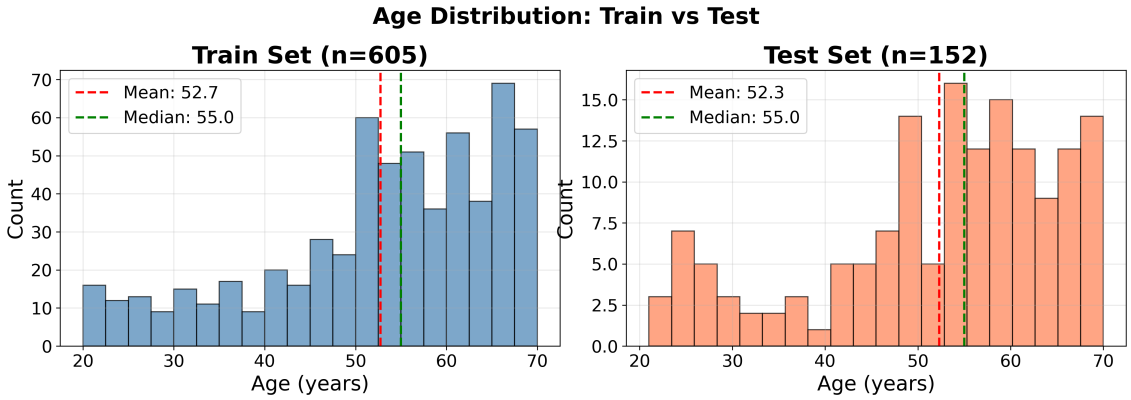


Figure 3.8: Age distribution: train (left) vs test (right).

It is important to note that all subsequent feature selection and model training steps were performed exclusively on the training set, with the test set remaining unseen until final evaluation. This strict separation is essential to prevent data leakage, which would otherwise produce artificially optimistic performance estimates.

3.4. Differential Gene Expression (DGE)

3.4.1 Motivation

Differential Gene Expression (DGE) analysis is a statistical framework used to identify genes whose expression levels vary significantly in association with a variable of interest, in this case, chronological age [83]. Rather than treating all 70,818 genes as equally informative, DGE pre-filtering was evaluated as a biologically motivated dimensionality reduction step for two reasons. First, by restricting the input to genes with a statistically significant age association, downstream feature selection and modelling would be applied

on a cleaner, more relevant feature space, reducing noise and potentially improving prediction accuracy. Second, genes that pass a statistical significance threshold are known to associate with age and are therefore more likely to produce biologically interpretable results.

3.4.2 Implementation

DGE analysis was implemented in R using the limma package from the Bioconductor ecosystem [83]. Limma represents the standard for linear model based differential expression analysis in transcriptomics and has no equivalent of comparable maturity in Python. All remaining pipeline steps were implemented in Python.

At its core, limma fits a separate linear model to each gene, testing whether the association between that gene's expression and age is statistically significant after accounting for the other covariates in the design matrix. With tens of thousands of genes being tested simultaneously, a naive per-gene approach would produce unreliable variance estimates because each estimate is based on a limited number of samples. Limma addresses this through empirical Bayes moderation: rather than estimating the variance of each gene in complete isolation, it borrows information across all genes to compute a stabilised variance estimate for each one. Genes with similar expression levels share variance information, making the resulting test statistics, known as moderated t-statistics, more reliable and better powered than those from an ordinary linear model. The output for each gene is an age coefficient representing how much expression changes per year of age, a moderated t-statistic, a raw p-value, and a Benjamini-Hochberg adjusted p-value used for significance calling.

Count-Based DGE Limma was applied using raw read counts as input via the limma-voom pipeline, which is the methodologically correct approach for RNA-seq data. Unlike other steps in the pipeline, DGE used the GTEx raw count matrix as input, rather than the $\log_2$(TPM+1) matrix. DGE analysis requires count data for correct variance modelling, while the downstream machine learning pipeline learns better from the linearized $\log_2$(TPM+1) representation.

Raw counts were loaded from the GTEx counts GCT file and the 605 training samples selected. Low-expression genes were removed by retaining only genes with at least 10 counts in at least 20% of samples (121 samples), reducing the gene space from 74,628 to 21,013 genes. This filtering step is standard practice in RNA-seq analysis and serves

to remove genes that are either not expressed or expressed at levels too low to provide reliable statistical estimates [17].

The count matrix, age labels, and metadata were aligned to their common 605 sample identifiers, sorted alphabetically to ensure consistent ordering across all three sources. Alignment was explicitly verified using two assertion checks: that the column names of the count matrix were identical in content and order to the sample identifiers in the age label vector, and identically so for the metadata table. The pipeline halts with an error if either condition is not satisfied.

Trimmed Mean of M-values (TMM) normalisation was then applied using *edgeR* [84]. Different samples in an RNA-seq experiment produce different total numbers of sequenced reads, known as library size. Without correction, a gene could appear more expressed in one sample simply because more reads were sequenced overall, rather than due to any true biological difference. TMM addresses an additional related problem: a small number of highly expressed genes can dominate the count distribution of a sample and distort comparisons across samples. TMM normalisation corrects for both library size and composition differences by computing scaling factors for each sample based on a trimmed subset of genes, excluding the most extreme observations in both directions before computing the normalisation factor. The normalisation factors in this analysis ranged from 0.402 to 1.901, indicating that uncorrected samples differed by up to 4.7-fold in library composition, a systematic difference large enough to substantially distort gene expression comparisons across samples if left uncorrected.

The design matrix was then constructed from the TMM-normalised, aligned and verified data. Each row corresponds to a training sample and each column to a modelled variable. Age was included as the primary variable of interest, and sex and RIN were included as covariates to account for their known confounding influence on gene expression profiles. Without these covariates, the age coefficient estimated by limma would partially reflect variation due to sex differences and sample quality rather than true age associated expression changes. The resulting design matrix comprised 605 samples and 4 columns: intercept, age, sex, and RIN.

The voom transformation was subsequently applied to convert the normalised count data into a form suitable for limma's linear modelling framework [85]. It is important to recognise a property of count data known as heteroscedasticity: the variance of a gene's counts is not constant but depends on its mean expression level. Lowly expressed genes tend

to have high relative variability, while highly expressed genes tend to be more stable. Standard linear models assume constant variance across observations and will produce incorrect inference if this assumption is violated. Voom addresses this by estimating the mean-variance relationship of the log-counts across all genes and using this relationship to assign precision weights to each observation. These weights are then incorporated into the linear model fitting, allowing limma to account for the heteroscedasticity inherent in count data and produce well calibrated statistical inference. The voom mean-variance trend plot (Figure 3.9 confirmed a well behaved decreasing trend, as expected for properly normalised RNA-seq data, validating the appropriateness of the transformation.

With TMM normalisation applied, the design matrix constructed and precision weights computed by voom, limma's empirical Bayes moderation was then applied on the calibrated data. The combination of voom weighting and limma moderation is what makes limma-voom the appropriate and recommended framework for differential expression analysis of RNA-seq count data, as opposed to applying limma directly to normalised continuous values such as TPM.

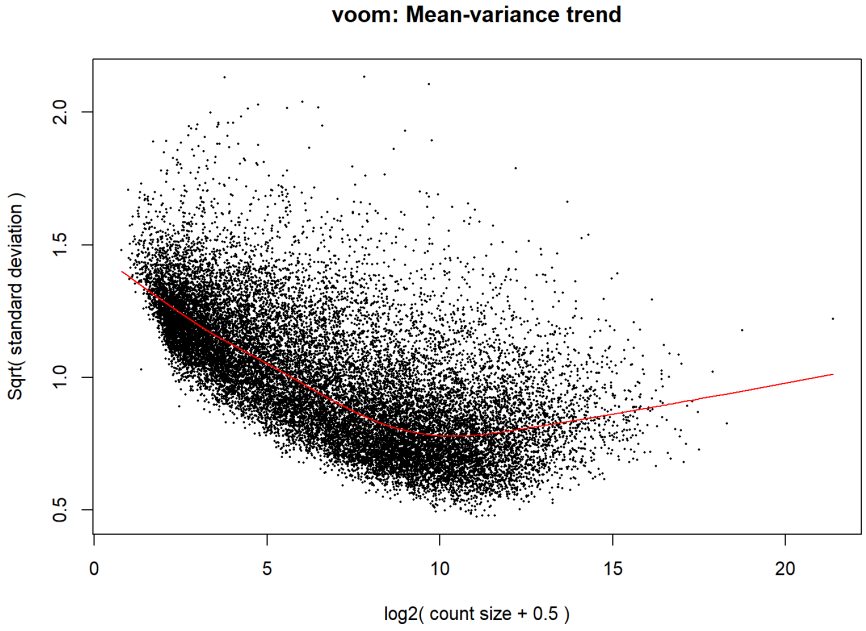


Figure 3.9: Voom plot: mean-variance trend.

3.4.3 Results

Two broad approaches exist for correcting multiple testing errors. The first is the Family-Wise Error Rate (FWER) [29], which controls the probability of making even a single false positive among all tests. Methods such as Bonferroni correction implement this by

dividing the significance threshold by the number of tests, in this case, 21,013, yielding a per-gene threshold of approximately 0.05/21,013 = 0.0000024. While this provides very strong protection against false positives, it is excessively conservative in a genomic context: when thousands of genes are genuinely associated with age, the vast majority of true signals would be discarded to avoid even one false discovery. The second approach is the FDR, which instead controls the expected proportion of false positives among all genes declared significant. An FDR threshold of 0.05 means that at most 5% of the genes called significant are expected to be false discoveries, a much more permissive and practically appropriate criterion for large-scale genomic studies where some degree of false discovery is acceptable in exchange for retaining genuine signal. When testing 21,013 genes simultaneously, applying an uncorrected p-value threshold of 0.05 would declare approximately 1,050 genes significant purely by chance; the Benjamini-Hochberg procedure [86] corrects for this by adjusting each gene's p-value to reflect the rank-ordered multiple testing burden, ensuring that the overall false discovery rate remains controlled at the specified threshold. Genes with a Benjamini-Hochberg adjusted p-value below 0.05 were retained as age-associated. Of the 21,013 genes entering the analysis after low-expression filtering, 15,217 (72.4%) passed this threshold. After matching against the TPM matrix by Ensembl gene identifier, 15,217 genes were carried forward to feature selection using DGE-based pre-filtering. The volcano plot (Figure 3.10) illustrates the distribution of age coefficients and significance levels across all tested genes.

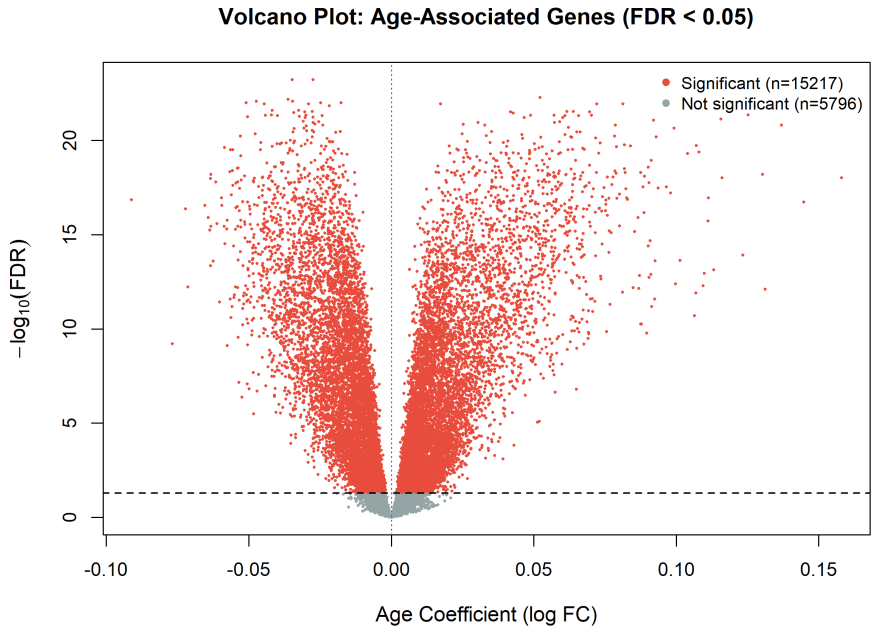


Figure 3.10: Volcano plot: age-associated genes (FDR < 0.05).

3.5. Feature selection

Feature selection was a central step in the analysis pipeline because the expression matrix, even after DGE pre-filtering, contained a large number of genes relative to the number of training samples. In this high dimensional setting, reducing dimensionality was necessary not only to improve computational efficiency but also to limit overfitting and to identify a compact subset of genes with the strongest predictive value for chronological age. Feature selection was therefore designed as a systematic comparison of multiple methods, each capturing a different notion of gene relevance, evaluated at four target gene set sizes across two pipeline configurations: the no- DGE configuration and the count-based DGE configuration.

3.5.1 Preprocessing for Feature Selection

Prior to feature selection, two preprocessing steps were applied to the training data. First, a basic expression filter was applied to remove lowly expressed genes, retaining only those with a mean $\log_2$ expression of at least 0.1 and a non-zero fraction of at least 0.05 across training samples. The filter criteria were derived exclusively from the training set to prevent data leakage, and the same gene list was then applied to the test set to ensure both partitions share an identical feature space, a requirement for applying the trained model to unseen data. In the no- DGE configuration this reduced the feature space from 70,818 to 27,322 genes, corresponding to the removal of 61.4% of the original variables. In the count-based DGE configuration, the input was already restricted to 15,217 age-associated genes, of which 15,008 remained after the expression filter, corresponding to a reduction of only 1.4%. This much smaller reduction reflects the fact that the DGE pre-filtering had already removed most lowly expressed genes at the count level during the low-expression filtering step described in Section 3.4.2.

Second, a quantile transformation was applied to each gene's expression values, mapping them to a normal distribution estimated from the training data. The same transformation fitted on the training set was then applied to the test set. This approach was selected over standard scaling because RNA-seq expression values often contain extreme observations and heavy tailed distributions that can disproportionately influence both regression-based and tree-based models. By mapping values to a Gaussian-like distribution, the transformation reduces the influence of outliers while preserving the relative ordering of samples within each gene. In preliminary experiments, this transformation

consistently produced better downstream prediction performance across all feature selection methods compared to standard scaling.

3.5.2 Feature Selection Methods

Six feature selection methods were evaluated at four gene set sizes (50, 100, 150, and 200), yielding 24 configurations per pipeline. For each method, gene ranking and selection were performed exclusively on the training data to prevent data leakage. The resulting gene list was then used to subset both the training and test sets to the same feature space, ensuring the trained model can be applied to unseen data. The methods and the rationale for their usage in this work are described below.

Variance-based selection ranks genes by their variance across training samples and selects the top n most variable genes. This is a simple, model-free approach that does not use age labels; therefore it serves as a baseline approach rather than a biologically-informed strategy.

Mutual Information (MI) regression estimates the statistical dependency between each gene's expression and age. Unlike correlation, which only captures linear relationships, mutual information detects any form of statistical association. This association is estimated using a *k*-nearest neighbours approach, which measures the local density of samples in the joint expression-age space without assuming any particular distribution. Genes are ranked by their mutual information score and the top n are retained.

LASSO regularisation fits a linear model with an *L1* penalty that reduces a majority of coefficients down to zero, thus producing a sparse coefficient vector. Regularisation strengths were tested across 100 values on a logarithmic scale between 10^{-4} and 10^{2}, and the smallest value that yielded no more than the target number of genes (50, 100, 150, or 200) as non-zero coefficients was selected. This approach allowed the gene set size to be controlled precisely while still benefiting from the sparsity-inducing properties of LASSO. The resulting coefficients are directly interpretable as linear age associations, making LASSO particularly suited to contexts where biological interpretability is important.

Ridge regularisation also adds a penalty to the linear model, but unlike LASSO it shrinks all coefficients towards zero without ever making them exactly zero. Because Ridge never eliminates genes through null coefficients, selection is instead performed by ranking genes by absolute coefficient magnitude and retaining the top n. Importantly, while LASSO tends to arbitrarily select one gene from a group of correlated genes and zero out

the rest, even if they are equally informative, Ridge spreads coefficient weight across correlated genes, so the ranking better reflects which groups of genes carry age-associated signals. This is particularly relevant in gene expression data where co-regulated genes share biological information. Ridge was fitted using a regularisation strength of α = 0.0001, selected by 5-fold cross-validation across a range of 10^{-4} to 10^{2}, and is computationally very fast, completing all four target sizes in under one minute. The convergence to the minimum available alpha suggests that Ridge required minimal regularisation to produce stable coefficient rankings on this dataset.

Random Forest (RF) feature importance can be measured in several ways, namely Mean Decrease in Impurity (MDI). MDI ranks genes based on how much each gene reduces prediction error across all trees. For feature selection purposes, a single Random Forest with fixed parameters (100 trees, maximum depth of 10, minimum 5 samples per leaf, seed = 42) was fitted on the training data. As Random Forest here serves purely as a gene ranker rather than a predictive model, fixed parameters were used rather than hyperparameter optimisation; a standard approach for importance-based feature selection [49]. This method captures non-linear associations and gene interactions that linear methods cannot detect.

Automatic Relevance Determination (ARD) Bayesian regression [87] is a probabilistic extension of linear regression in which each gene is assigned not only a coefficient but also a precision parameter, a measure of how confidently the model believes that gene is relevant. For uninformative genes, the model assigns very high precision, which effectively drives the gene's coefficient to near zero. For informative genes, the model retains a meaningful coefficient. This mechanism is similar in spirit to LASSO but operates through a probabilistic framework rather than a fixed penalty. Genes are ranked by the absolute magnitude of their posterior coefficients and the top n are retained.

Table 3.1 summarises the six feature selection methods evaluated in this thesis, comparing them across stringency, interpretability, computational runtime, and the key parameters used.

Table 3.1: Feature selection methods comparison.

Method	**Stringency**	**Inter-pretability**	**Runtime DGE (all 4 sizes)**	**Key Parameters Used**
Variance	None	Low	∼0.5s	-
MI	Low	Medium	∼2.9min	k-NN estimator
LASSO	High	High	∼154min	α swept 10^{-4}–10^{2}
Ridge	Medium	Medium	∼2.0min	α = 0.0001, selected by 5-fold CV
RF	Low	Low	∼2.3min	100 trees, max depth 10, min samples leaf 5, seed 42
ARD	High	Medium	∼19.6min	Default *scikit-learn* priors

Each resulting gene set was evaluated using an Elastic Net probe model to obtain preliminary estimates of test MAE and R^2 across all methods and gene set sizes, enabling a fair and consistent comparison. These results are reported in Section 3.5.3, where the final configuration is selected.

3.5.3 Results and Configuration Selection

Two pipeline configurations were systematically compared. Both pipelines started from the full expression matrix of 70,818 genes. After basic expression filtering, the no- DGE pipeline retained 27,322 genes, which were then passed directly to feature selection after quantile transformation. The DGE pipeline instead restricted the input to 15,217 age-associated genes identified by limma-voom (Section 3.4), of which 15,008 remained after the expression filter, before feature selection was applied. The best performing configuration from each pipeline is summarised in Table 3.2.

Table 3.2: Best configuration results.

Configuration	**Best Method**	**N genes**	**Test MAE (years)**	R^2	**Gap (years)**	**Wall Time (hours)**
No- DGE	LASSO	200	7.53	0.499	4.5	∼6
DGE	Ridge	200	7.81	0.469	3.62	∼3

Figures 3.11 and 3.12 show test MAE and explained variance (R^2) as a function of the number of selected genes for all six feature selection methods in the DGE configuration. Ridge selection consistently achieved the lowest MAE across all gene set sizes, with *Ridge_200* identified as the best overall configuration (MAE = 7.81 years, R^2 = 0.469). Variance-based selection performed poorly regardless of gene set size, confirming that expression variance alone is not a reliable proxy for age association.

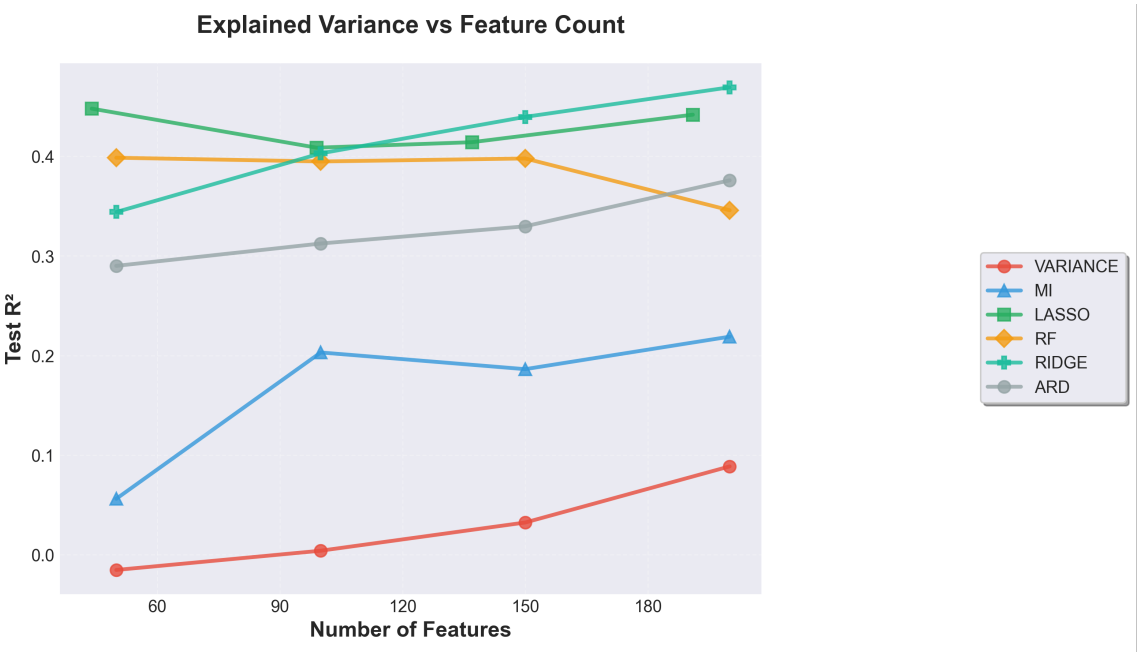


Figure 3.11: R^2 across feature selection methods and feature set sizes.

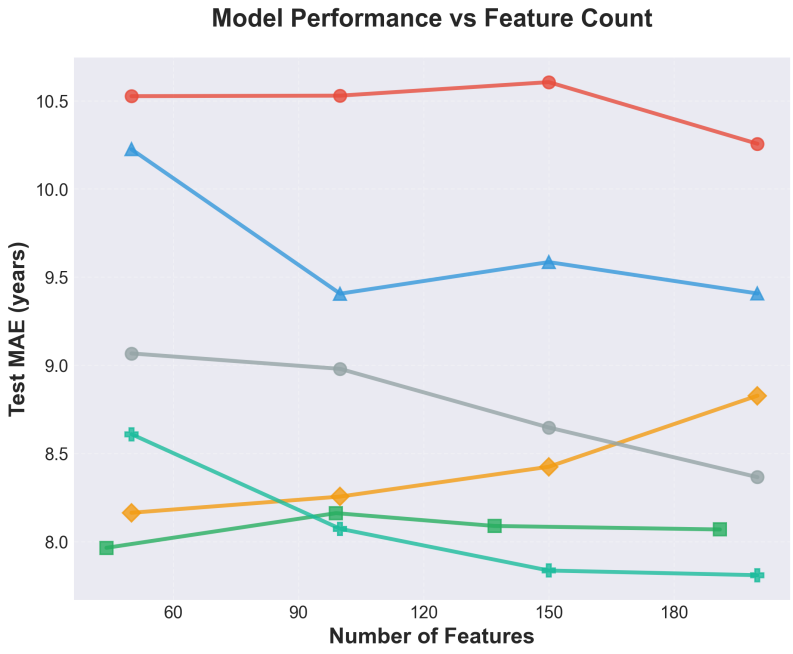


Figure 3.12: Test MAE across feature selection methods and feature set sizes.

The DGE pipeline with Ridge selection for 200 genes was chosen as the final configuration for the following reasons. First, DGE pre-filtering restricts the feature space to genes with a statistically demonstrated association with age at FDR < 0.05, estimated using a linear model that accounts for biological and technical covariates, namely sex and RIN. Differentially expressed genes have stronger biological evidence of age association. Second, the DGE-filtered input space is substantially cleaner, improving the signal-to-noise ratio for downstream feature selection methods. Third, genes selected from a DGE-filtered space are not only predictive but also statistically age-associated, making them more meaningful for biological interpretation and cross-validation against published ageing signatures [7, 41]. Fourth, this approach mirrors the standard workflow in transcriptomic ageing studies [7], making results more comparable to published literature. Fifth, the performance difference between the best no- DGE configuration (*LASSO_200*, MAE 7.53) and the selected DGE configuration (*Ridge_200*, MAE 7.81) is only 0.28 years, well within the margin of variability expected from a single train/test split [29], and is therefore not a meaningful justification for sacrificing biological signal. Finally, the approximately 50% reduction in feature selection wall time (from $\sim$6 to $\sim$3 hours) leads to more efficient model training, while represents a practically significant efficiency gain, reducing energy consumption [88]. Final model performance is reported in Section 3.6.

3.6. Regression Modelling

The goal of regression modelling was to predict continuous age from the selected gene expression features. Four regression models were trained and evaluated, chosen to encompass several model families: a penalised linear model (Elastic Net), two decision tree ensemble methods with different approaches for improving model accuracy (Random Forest and Gradient Boosting Trees) and a neural network. Different model families make different assumptions about the relationship between gene expression and age, and comparing them allows us to empirically identify the most appropriate model for our data. All models were implemented using scikit-learn [82] , with the exception of XGBoost for which its dedicated Python library was used [50]. The training set comprised 605 samples across the final 200 genes previously selected by 1) the count-based DGE pipeline described in Section 3.4 and 2) the Ridge feature selection described in Section 3.5. The held-out test set of 152 samples was never used during training, hyperparameter optimisation, or ensemble construction.

3.6.1 Hyperparameter Optimisation

Hyperparameters are settings that must be set before training begins and control how the model learns. For example, hyperparameters handle how strongly to regularise, how deep to grow trees, or what learning rate to use. They cannot be learned from the training data directly and must be selected by testing different configurations.

Hyperparameter optimization was performed for each model independently using *RandomizedSearchCV* with 5-fold cross-validation, scoring on negative mean absolute error. Rather than exhaustively testing all combinations, *RandomizedSearchCV* randomly samples a fixed number of hyperparameter combinations from a statistical distribution, which is more computationally feasible for large search spaces compared to trying out millions of model fits.

In 5-fold cross-validation, the 605 training samples are randomly shuffled and divided into five equal groups. Each group takes a turn as the validation set while the model is trained on the remaining four groups. The cross-validated MAE for a given hyperparameter setting is the average MAE and standard deviation across all five validation folds, providing an estimate of generalisation performance without using the test set. A fixed random seed (seed = 42) was used for both the fold assignment and all model initialisations to ensure full reproducibility.

To prevent data leakage, the *StandardScaler* for models requiring scaling was fitted only on the training folds and applied to each validation fold. Fitting the scaler on all data including the validation fold would constitute leakage and produce over-optimistic performance estimates.

3.6.2 Evaluation Metrics

Model performance was assessed on the held out test set of 152 samples using three primary metrics:

- **Mean Absolute Error (MAE)**, equation 3.1, measures the average absolute difference between predicted and true age in years and is the primary metric used throughout this work, as it is directly interpretable and robust to outliers.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{3.1}$$

- **Coefficient of determination (R^2)**, equation 3.2, measures the proportion of variance in chronological age explained by the model. An R^2 of 1.0 indicates perfect prediction while R^2 of 0 indicates no better than predicting the mean age.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.2}$$

- **Root Mean Squared Error (RMSE)**, equation 3.3, penalises large errors more heavily than MAE because individual errors are squared before averaging, meaning predictions that deviate substantially from the true age contribute disproportionately to the final score. This makes RMSE more sensitive to outliers.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{3.3}$$

where $\hat{y}_i$ is the predicted age, y_i is the true chronological age, $\bar{y}$ is the mean true age, and n is the number of test samples.

3.6.3 Models

Elastic Net is a linear regression model that combines *L1* (LASSO) and *L2* (Ridge) regularisation penalties, discussed in Section 3.5.2. The *L1* component encourages sparsity by driving some coefficients to zero, while the *L2* component stabilises the solution by shrinking correlated coefficients together. The balance between these two penalties is controlled by the *l1_ratio* hyperparameter (0 = pure Ridge, 1 = pure LASSO), and the overall regularisation strength is controlled by alpha (higher values produce simpler models with smaller coefficients). Elastic Net requires standardised input features and was therefore trained within a *scikit-learn* Pipeline containing *StandardScaler*. It was included because it is a highly interpretable model, producing gene coefficients that are directly interpreted as linear age associations.

The alpha search range was set from 10^{-4} to 10^{2} on a logarithmic scale, covering the full spectrum from almost no regularisation to very strong shrinkage. The *l1_ratio* was searched from 0 to 1 in 21 steps, (0, 0.05, 0.10, ..., 0.95, 1.0), covering the full spectrum from pure Ridge to pure LASSO behaviour. Elastic Net was tuned by sampling 40 hyperparameter settings. The best-performing combination of hyperparameters was alpha = 0.202 and *l1_ratio* = 0.15, indicating a predominantly Ridge-like solution with only mild sparsity. This result is biologically informative: the low *l1_ratio* suggests that the age signal is distributed across many correlated genes rather than concentrated in a few

highly predictive ones, which is consistent with the known polygenic and systemic nature of transcriptomic ageing. Cross-validated MAE was 5.83 $\pm$ 0.39 years, with stable performance across folds (range 5.30 to 6.22 years), indicating consistent generalisation. On the full training set the model achieved R^2 = 0.815 and MAE = 4.38 years, with test set performance of R^2 = 0.471 and MAE = 7.75 years.

Random Forest is an ensemble of decision trees trained on bootstrap samples of the training data (random subsets of samples drawn with replacement). Each tree is also trained on a random subset of genes at each split, further increasing diversity. Predictions are made by averaging across all trees, which reduces variance compared to a single tree. Random Forest does not require feature scaling and can capture non-linear relationships between gene expression and age. The following hyperparameters were considered and the selected values are indicated in bold:

- ***max_depth*** (maximum depth of each tree, controlling model complexity): restricted to finite values (5, 10, 15, 20), explicitly excluding unlimited depth, as fully grown trees memorise the training data completely and with only 605 samples would lead to severe overfitting (selected: **10**).
- ***n_estimators*** (number of trees in the forest): values of 100, 200, and 500 were tested, covering a range from computationally efficient to highly stable ensembles (selected: **500**).
- ***min_samples_split*** (minimum number of training samples required to split a node): values of 2, 5, and 10 were tested. Higher values prevent the model from learning patterns from very small groups of samples (selected: **2**).
- ***min_samples_leaf*** (minimum number of samples required in a leaf node): values of 1, 2, and 4 were tested. Higher values produce smoother predictions by requiring more evidence before finalising a prediction (selected: **1**).
- ***max_features*** (fraction of genes considered at each split): values of sqrt, log2, and 0.3 were tested. The value 0.3 introduces sufficient diversity between trees and reduces correlation among them, particularly important when many genes carry overlapping age-associated signals (selected: **0.3**).

Random Forest was tuned using 50 combinations of hyperparameters. The selection of *max_depth* = 10 and *max_features* = 0.3 confirms that moderate depth and high feature

subsampling were beneficial, consistent with a dataset where age signal is broadly distributed. Cross-validated MAE was 7.80 $\pm$ 0.59 years with notable variability across folds (range 7.23 to 8.73 years). On the full training set the model achieved R^2 = 0.915 and MAE = 3.05 years, while test performance was R^2 = 0.406 and MAE = 7.96 years. The large train-test gap indicates overfitting, suggesting that the chosen random forest model struggled to generalise to the test set data.

XGBoost is a gradient boosting framework. In gradient boosting, trees are built sequentially, with each new tree fitted to the residuals of the previous ensemble. Unlike Random Forest, which builds all trees independently, XGBoost builds each tree to improve the results of the previous one, thereby focusing on difficult to predict samples. Regularisation is applied through *L1* and *L2* penalties on leaf weights, and additional hyperparameters control tree depth, subsampling rate, and learning rate. XGBoost does not require feature scaling and was included because gradient boosting methods consistently achieve state-of-the-art performance on tabular data [50]. The following hyperparameters were considered and the selected values are indicated in bold:

- ***max_depth*** (maximum depth of each tree): searched over 3, 4, 5, and 6, as deeper trees produced extreme overfitting without improving test performance in preliminary experiments (selected: **3**).
- ***n_estimators*** (number of trees in the ensemble): searched over 100, 200, 300, and 500 (selected: **300**).
- ***learning_rate*** (shrinkage factor applied to each tree's contribution; lower values require more trees but often generalise better): searched over 0.01, 0.05, and 0.1 (selected: **0.1**).
- ***min_child_weight*** (minimum sum of sample weights required in a leaf node; higher values prevent the model from learning patterns from rare subgroups of donors): searched over 1, 3, 5, and 10 (selected: **3**).
- ***subsample*** (fraction of training samples randomly selected to fit each tree, introducing stochasticity to reduce overfitting): searched over 0.6, 0.7, 0.8, and 0.9, with 0.6 included as a lower bound to allow more aggressive stochasticity (selected: **0.7**).
- ***colsample_bytree*** (fraction of genes randomly selected for each tree, analogous to max_features in Random Forest): searched over 0.6, 0.7, 0.8, and 0.9 (selected: **0.9**).

- ***reg_alpha*** (*L1* regularisation on leaf weights, which pushes some leaf values towards zero): searched over 0, 0.1, 1.0, and 10.0 (selected: **10.0**).

- ***reg_lambda*** (*L2* regularisation on leaf weights, which shrinks all leaf values proportionally): searched over 0.1, 1.0, and 10.0 (selected: **1.0**).

XGBoost was tuned using 50 hyperparameter combinations. The convergence to the maximum available *L1* regularisation (*reg_alpha* = 10.0) and very shallow trees (*max_depth* = 3) is notable. Shallow trees limit the complexity of each individual learner, and strong *L1* regularisation pushes many leaf weights towards zero, both of which constrain the model's capacity to memorise training data. That cross-validation selected these values suggests the data benefited from aggressive regularisation; the model generalised better by being kept simple than by being allowed to fit complex patterns. Cross-validated MAE was 6.92 $\pm$ 0.69 years with moderate fold variability (range 6.17 to 7.70 years). On the full training set the model exhibited near-perfect fit (R^2 = 0.996, MAE = 0.68 years), while test performance was R^2 = 0.506 and MAE = 7.35 years.

Neural Network (MLPRegressor) is a multi-layer perceptron with fully connected hidden layers. Each neuron applies a non-linear activation function to a weighted sum of its inputs, allowing the network to learn complex non-linear patterns. The network was trained using gradient descent with early stopping; if validation loss did not improve for 10 consecutive iterations, training was halted to prevent overfitting. Neural Networks require standardized inputs and are sensitive to learning rate and architecture choices. This model was included to assess whether complex non-linear patterns in the gene expression data could provide additional predictive power. MLPRegressor was selected over more complex architectures such as convolutional or recurrent networks because the input is a fixed length vector of gene expression values with no spatial or temporal structure, making feedforward networks the appropriate choice for this task. More complex deep learning architectures were explicitly ruled out on the grounds of insufficient sample size. The following hyperparameters were considered, with the selected values indicated in bold:

- ***hidden_layer_sizes*** (number of layers and neurons per layer; for example (200, 100) means two hidden layers with 200 neurons in the first and 100 in the second): restricted to shallow two-layer architectures (100,), (100, 50), (200, 100), and (100, 100, 50) as deeper networks would be unable to train reliably on 605 samples (selected: **(200, 100)**).

- ***activation*** (non-linear function applied after each neuron's weighted sum): searched over *ReLu*, which computes max(0, x), and *Tanh*, which maps values to the range −1 to 1 (selected: ***Tanh***).
- ***alpha*** (*L2* weight regularisation strength, penalising large network weights to prevent overfitting): searched over 0.0001, 0.001, and 0.01 (selected: **0.001**).
- ***learning_rate_init*** (initial step size for gradient descent weight updates; too large causes instability, too small causes slow convergence): searched over 0.0001, 0.001, and 0.01 (selected: **0.01**).

The Neural Network was tuned using 40 random candidates. Cross-validated performance was poor and highly unstable (CV MAE = 8.72 ± 0.84 years, CV R^2 = −0.038 $\pm$ 0.338), with negative R^2 values in three out of five folds, indicating that the model performed worse than simply predicting the mean age in most folds. Test set performance was also no better than random chance (R^2 = −0.096, MAE = 9.47 years). The instability across folds suggests that there is insufficient data to properly tune the model's parameter-dense architecture. With 605 training samples and a (200, 100) neuron architecture comprising over 20,000 trainable parameters, the network was severely underdetermined. The Neural Network was therefore excluded from the ensemble constructed in subsequent steps.

Table 3.3 summarises all four models evaluated. Models whose names appear in bold were included in the weighted ensemble described in Section 3.6.4. Train and test MAE are reported on the full training set and held-out test set respectively.

Table 3.3: Hyperparameter search results per model.

Model	**Number of Hyperparameters**	**Number of Sampled Hyperparameter Combinations**	**Train MAE (years)**	**Test MAE (years)**	**Wall Time**
Elastic Net	2	40	4.38	7.75	∼7s
Random Forest	5	50	3.05	7.96	∼30s
XGBoost	8	50	0.68	7.35	∼227s
Neural Network	4	40	2.72	9.47	∼166s

3.6.4 Weighted Ensemble

A weighted ensemble was constructed by combining the predictions of the three models capable of generalization, Elastic Net, XGBoost, and Random Forest, ranked by their cross-validated MAE computed entirely from the training set. Ranking by CV MAE rather than test MAE ensures that no information from held out data influences the ensemble construction. Each model was assigned a weight inversely proportional to its cross-validated MAE, normalised to sum to one: Elastic Net 0.386, XGBoost 0.325, Random Forest 0.289. Although XGBoost achieved the best test-set performance (Section 4.1), Elastic Net obtained the lowest cross-validated MAE during training, and therefore received the highest ensemble weight; weighting was deliberately based on CV performance rather than test performance to avoid any influence from the held-out set. The ensemble prediction for each test sample was computed as the weighted average of the three models' predictions.

Models were grouped into an ensemble because each one makes different errors on different samples according to their underlying algorithm. Elastic Net captures linear age associations through regularised coefficients, XGBoost focuses on difficult samples through sequential tree correction, and Random Forest provides robust variance reduction through bootstrap averaging of hundreds of decision trees. Their predictions are combined using weights reflecting their CV performance on the training set, with the aim of reducing overall test MAE compared to any individual model's performance.

3.7. SMOGN

The GTEx whole blood training set has a pronounced age imbalance, with donors aged 50–-70 representing 65.6% of training samples while donors under 40 comprised only 16.9% (Figure 3.5). To address this issue, Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise (SMOGN) was evaluated as a pre-processing step prior to regression modelling [77]. SMOGN uses a relevance function to identify underrepresented target values and generates synthetic training samples by interpolating between real donors in the feature space, while undersampling overrepresented regions. A custom mean-based relevance function was implemented to identify rare age groups without requiring manual parameter tuning. SMOGN simultaneously generated 276 synthetic donors concentrated in the under 50 age range and undersampled overrepresented age groups (50–-70 years), yielding a rebalanced training set of 764 samples. The test set was not modified at any point during this process.

The pipeline using SMOGN produced a test ensemble MAE of 7.53 years (R^2 = 0.466, RMSE = 9.55), worse on all three metrics compared to the ensemble without rebalancing (MAE = 7.29 years, R^2 = 0.515, RMSE = 9.10). One possible explanation is that synthetic interpolations in a 200-dimensional gene expression space do not correspond to biologically plausible expression profiles, introducing artificial signal that does not generalise to the real test set. SMOGN was therefore not adopted in the final pipeline. The age imbalance remains a limitation of this work and is discussed further in Section 5.2.

4. Results and Discussion

4.1. Results

The five regression models described in Section 3.6 were evaluated on the held out test set of 152 samples. Full test set performance is summarised in Table 4.1.

Among individual models, XGBoost achieved the best performance MAE (7.35 years, R^2 = 0.506), followed by Elastic Net (7.75 years, R^2 = 0.471) and Random Forest (7.96 years, R^2 = 0.406). The Neural Network failed to generalise (MAE = 9.47 years, R^2 = −0.096) and was excluded from the ensemble, as detailed in Section 3.6.3.

The weighted ensemble of Elastic Net, XGBoost, and Random Forest (Section 3.6.4) achieved a test MAE of 7.29 years and R^2 = 0.515, only marginally better than XGBoost alone (MAE = 7.35 years, R^2 = 0.506). Given this negligible gain, and the added computational cost and loss of single-model interpretability that the ensemble introduces, XGBoost was selected as the final predictive model.

Table 4.1: Per-model test set performance.

Model	R^2	**MAE**	**RMSE**
Elastic Net	0.471	7.75	9.51
Random Forest	0.406	7.96	10.07
XGBoost	**0.506**	**7.35**	**9.18**
Neural Network	-0.096	9.47	13.68
Ensemble	*0.515*	*7.29*	*9.10*

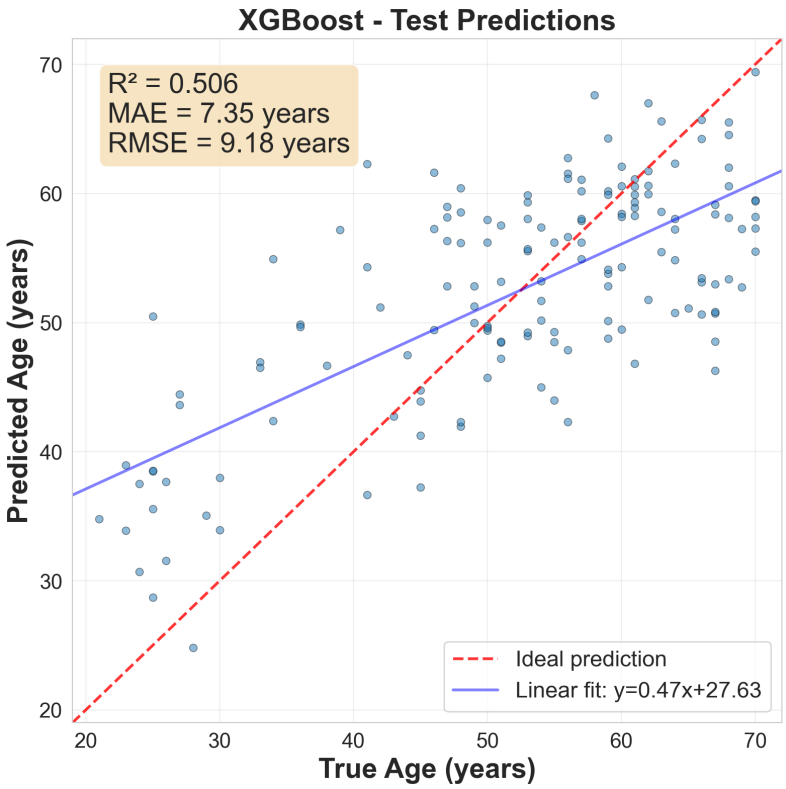


Figure 4.1: XGBoost predicted vs true age scatter plot (right).

As shown in Figure 4.1, XGBoost predictions map to the true age reasonably well across the full age range. However, the linear fit ($y = 0.47x + 27.63$) deviates from the ideal prediction line, indicating regression to the mean: younger donors tend to be overestimated and older donors underestimated. This pattern is a known shortcoming of age prediction models trained on imbalanced age distributions, where the loss function is dominated by the majority age group, in this case donors aged 50–70, reducing the model's sensitivity at the extremes of the age range [29].

To quantify this pattern directly, age acceleration (ΔAge = $\hat{y} - y$, Section 2.3.1) was computed for each test-set donor using the chosen XGBoost model. Across the full test set, the median ΔAge was −0.43 years, indicating no systematic bias at the aggregate level; the mean (+0.12 years) was consistent with this. However, this aggregate figure conceals a strong age-dependent pattern, visualised as a boxplot of ΔAge by age group in Figure 4.2 and summarised by median in Table 4.2.

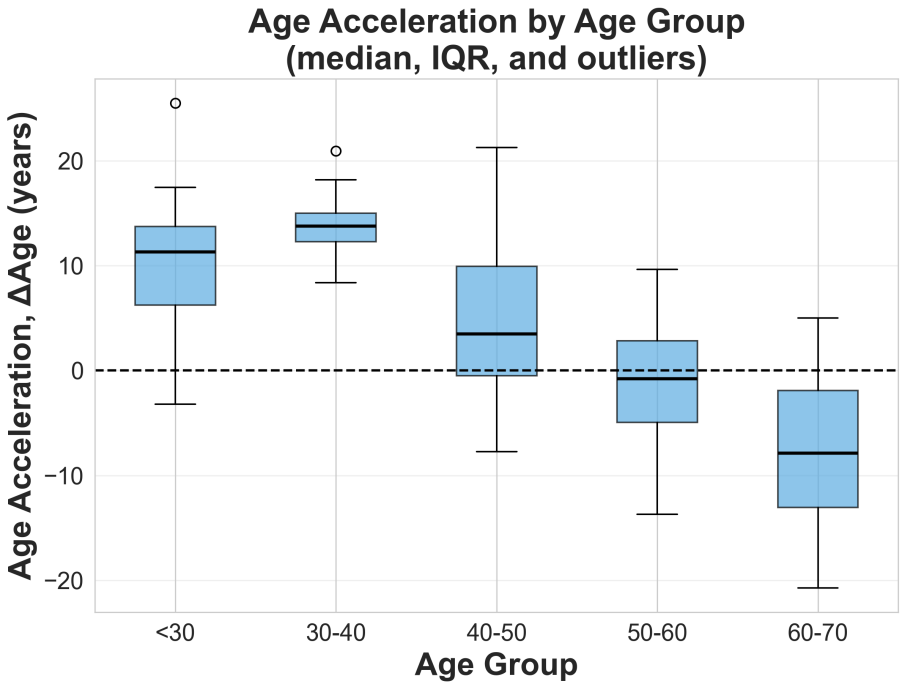


Figure 4.2: Age acceleration by age group.

Table 4.2: Median age acceleration and IQR per age group.

Age group	**n**	**Median ΔAge (years)**	**IQR (years)**
<30	18	+11.30	6.24 to 13.74
30–40	8	+13.77	12.30 to 15.01
40–50	31	+3.49	-0.50 to 9.93
50–60	48	-0.77	-4.96 to 2.83
60–70	47	-7.88	-13.05 to -1.91

Donors under 40 are predicted, by median, 11 to 14 years older than their true chronological age, while donors aged 60–70 are predicted to be nearly 8 years younger. The interquartile ranges in Table 4.2, together with the boxplot in Figure 4.2, show that while the direction of bias is consistent and monotonic across age groups, the spread within each bin is substantial relative to the differences between adjacent groups, particularly 40–50 and 50–60, whose Interquartile Range (IQR)s overlap. The smallest age groups (<30, n=18; 30–40, n=8) also show the most extreme individual outliers (Figure 4.2), and their bias estimates should be interpreted with appropriate caution given the limited sample size. This pattern clearly shows the regression-to-mean behaviour visible in the parity plot (Figure 4.1, and confirms that the apparent absence of bias at the aggregate level

results from these opposing directional errors cancelling out across age groups, rather than from the model being equally accurate across the age range. Because ΔAge is the operational proxy for biological age used throughout this thesis (Section 2.3.1), this finding has a direct implication for interpretation: ΔAge values for donors at the extremes of the training age range should be treated with caution, both because of the systematic bias demonstrated here and because of the small sample sizes underlying the youngest group estimates.

4.2. Model Performance

The chosen model, XGBoost, achieved a test MAE of 7.35 years and R^2 of 0.506, consistent with the performance range reported for blood transcriptomic clocks trained on datasets of comparable size. *Peters et al.* [7], the primary reference for this type of work, reported correlations of 0.6-0.7 and MAE of 7–8 years across independent cohorts using nearly 15,000 samples, a range the present model matches with approximately twenty times fewer training samples. More recent clocks such as BiT-age [36] achieve substantially lower MAE ($\sim$3 years), within the 2.5-6 year range reported across modern transcriptomic clocks (Table 2.2). This range reflects models trained on far larger and more diverse cohorts than the present whole-blood GTEx sample; BiT-age in particular was trained on all GTEx tissues rather than only whole blood, limiting direct comparison. The age-dependent bias in ΔAge (Section 4.1) is consistent with findings from comparable GTEx-based transcriptomic clock studies [61, 62], confirming this is a dataset-level limitation rather than a modelling failure, and will require larger, more balanced cohorts to resolve.

4.3. Interpreting Gene Expression Signal

4.3.1 Model Choice for Biological Interpretation

Biological interpretation of the gene signature requires understanding both the direction of each gene's association with age and the importance of each gene to the final model's predictions. These two questions are addressed using two complementary methods: Elastic Net coefficients for direction, and SHapley Additive exPlanations (SHAP) values computed on the final XGBoost model for importance.

Elastic Net is used for the directional analysis because, as a linear model, it assigns one coefficient per gene with an unambiguous sign: positive values indicate genes whose

expression increases with age, negative values indicate genes whose expression decreases. This single-number summary is directly comparable to the up/down-regulation language used throughout the transcriptomic ageing literature (Section 2.2.2) and to specific genes reported in prior studies such as *Peters et al.* [7]. XGBoost, by contrast, captures non-linear and interaction effects that have no single coefficient or fixed sign, making it unsuitable for this kind of feature direction analysis despite being the more accurate predictive model.

Of the 200 selected genes, 93 have positive Elastic Net coefficients and 107 have negative coefficients, with eight genes at the tail of the ranking driven to zero by the *L1* penalty, indicating negligible individual predictive contribution after accounting for other genes in the signature. Figure 4.3 shows the top 20 genes ranked by absolute coefficient magnitude, coloured by direction of association.

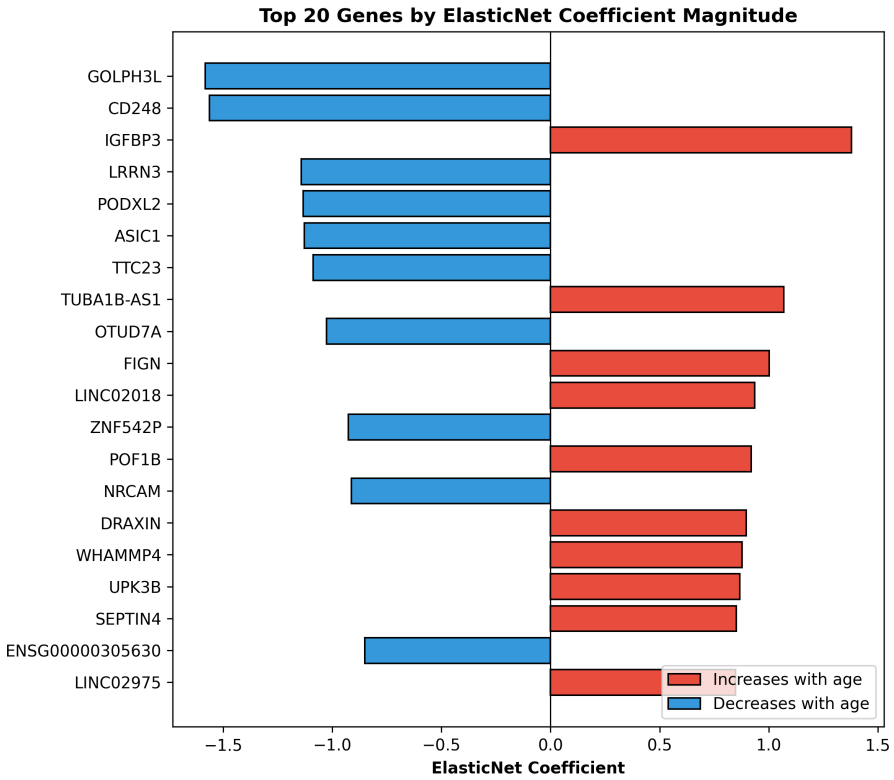


Figure 4.3: Top 20 genes by Elastic Net absolute coefficient magnitude on the held-out test set. Red bars indicate genes whose expression increases with age; blue bars indicate genes whose expression decreases with age.

To verify that Elastic Net's direction is not caused by using a different, less accurate model than the one actually deployed, SHAP [89] values were computed using the XGBoost model to obtain a model-agnostic, non-linear measure of gene importance. This comparison confirms whether the genes Elastic Net identifies as important also significantly

contribute to the XGBoost model.

Figure 4.4 shows the top 20 genes ranked by mean absolute SHAP value for XGBoost , computed on the held-out test set (n=152). Figure 4.5 shows the full distribution of SHAP values for the same genes across all test samples, including the direction of each gene's effect. Each point represents one test sample; horizontal position indicates the SHAP value (impact on predicted age) and colour indicates the gene's expression level (red = high, blue = low).

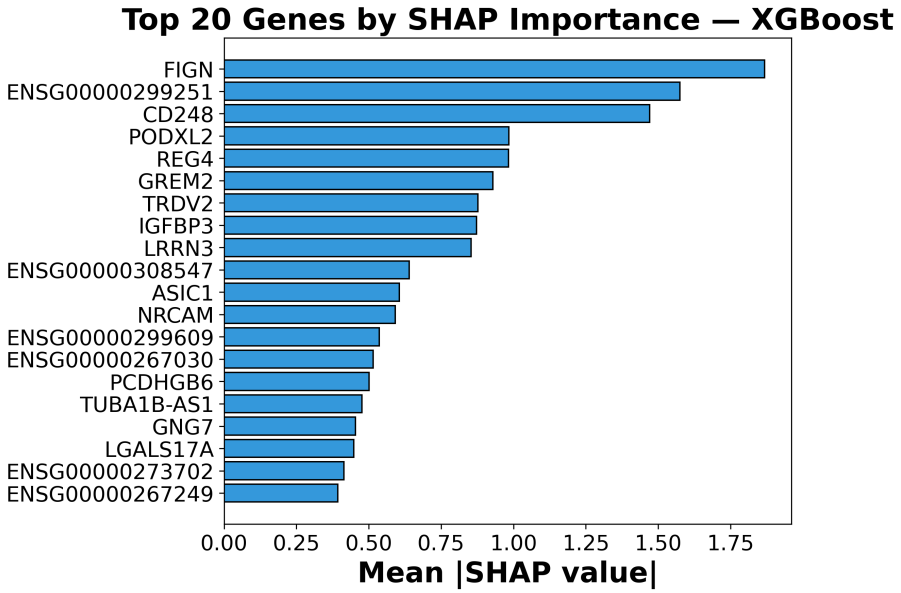


Figure 4.4: Top 20 genes ranked by mean absolute SHAP value.

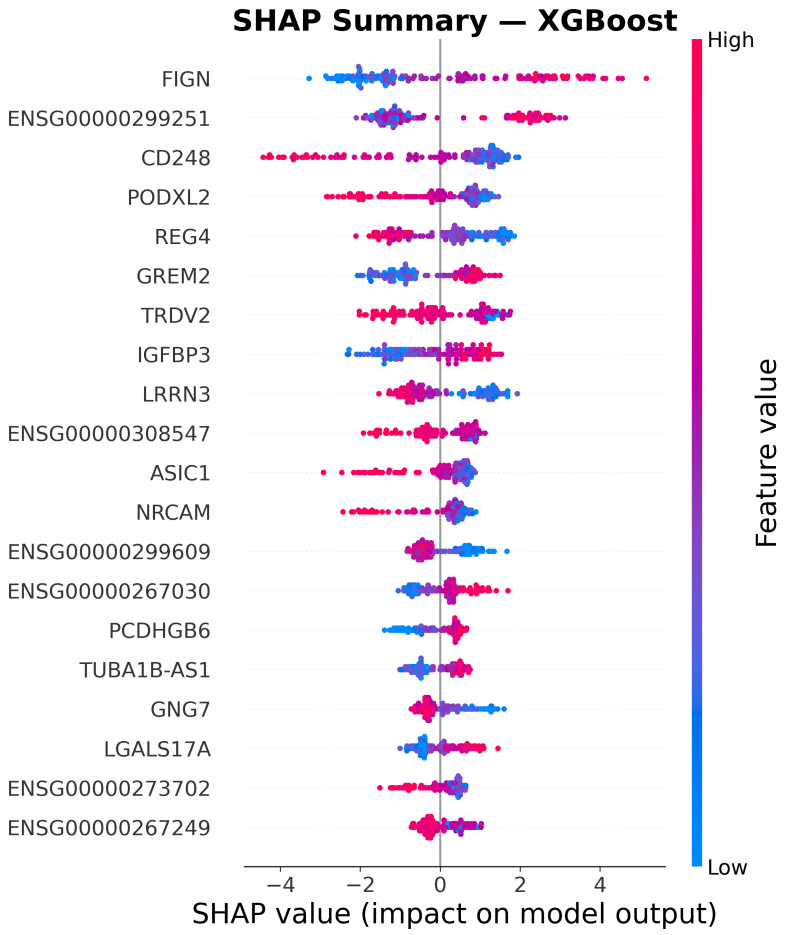


Figure 4.5: SHAP summary plot showing feature value impact and direction.

The top 20 genes ranked by mean absolute SHAP value in XGBoost overlap with the top 20 genes ranked by absolute Elastic Net coefficient in 8 of 20 cases (40%). Eight genes are consistently identified as important by both a linear and the non-linear, gradient boosted decision tree-based model: *FIGN, CD248, PODXL2, IGFBP3, LRRN3, ASIC1, NRCAM*, and *TUBA1B-AS1*. Genes identified as important by both models constitute the strongest candidates within the signature, since their importance is corroborated by two models and does not depend on an assumption of linearity. Notably, *CD248* and *LRRN3* are both established blood-specific ageing markers from *Peters et al.* [7] and their consistent ranking as top features in both models lends additional biological credibility to the gene signature.

The remaining genes identified by only one method are examined further in Section 4.3.4, where specific cases of agreement and disagreement between the two models are discussed.

4.3.2 Genes Overlapping with State-of-the-Art Transcriptomic Clocks

As established in Section 4.3.1, Elastic Net coefficients are used throughout this section to characterise the direction and relative strength of each gene's association with age, since XGBoost (the chosen predictive model) does not produce interpretable coefficients per gene. All genes discussed below were part of the fixed 200 gene set selected by Ridge regression (Section 3.5) and used identically across all four regression models; only the coefficient values, used here for direction, come specifically from the Elastic Net model.

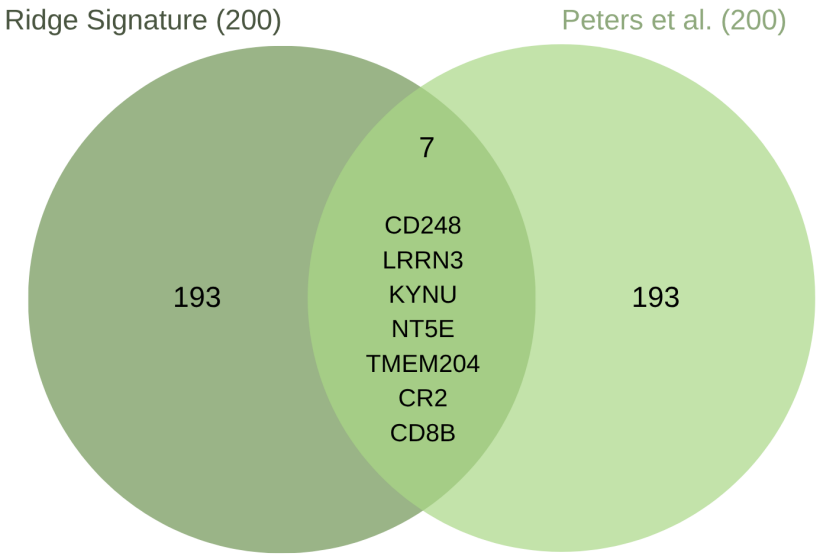


Figure 4.6: Overlap between the 200-gene Ridge signature and the top 200 genes from *Peters et al.* [7]. The seven shared genes are listed in the intersection.

Seven genes were shared with the top 200 genes from *Peters et al.* [7] as shown in Figure 4.6. Given the entirely different platforms, cohorts, and analytical methods between studies, this overlap provides meaningful cross-study validation, as genes identified under two different conditions are less likely to reflect technical artefacts and instead more likely to represent genuine transcriptomic markers of ageing in blood.

The two most strongly weighted shared genes are *CD248* and *LRRN3*. CD248 (endosialin, coefficient -1.563) is a protein expressed in the cells that form and maintain blood vessel walls. Its consistent downregulation with age across two independent signatures suggests a progressive loss of vascular cell activity with age, supporting *CD248* as a robust transcriptomic marker of vascular ageing in blood [1]. LRRN3 (coefficient −1.142) is a protein involved in cell adhesion in the nervous system, whose expression is also detected in circulating immune cells. Its downregulation with age in blood reflects

the progressive decline of immune cell populations and their signalling capacity, a process known as immunosenescence, consistently reported in blood transcriptomic ageing studies [7, 23].

KYNU (kynureninase, coefficient +0.626) is an enzyme involved in the breakdown of tryptophan, an essential amino acid. This metabolic pathway is relevant to ageing because it is the main route through which cells produce NAD+, a molecule essential for energy metabolism. The levels of this metabolic process decline with age, thereby contributing to cellular dysfunction [10]. The upregulation of *KYNU* with age is consistent with increased activity of this pathway in aged blood, and *Peters et al.* [7] also identified it as one of the most consistently supported metabolic markers of blood ageing.

NT5E (or CD73, coefficient −0.222) is a protein on the surface of immune cells that produces adenosine, a molecule that acts as a natural brake on inflammatory responses, helping the immune system avoid excessive inflammation. Its downregulation with age decreases adenosine production in aged immune cells. This reduction in their ability to control inflammation contributes to the chronic low-grade inflammatory state observed in older individuals [22]. CR2 (coefficient −0.109) is a receptor on B lymphocytes, immune cells responsible for producing antibodies, whose decline follows the well-documented reduction in B cell number and function with age [22]. CD8B (coefficient +0.0007) is a protein on cytotoxic T lymphocytes, the immune cells that destroy virus-infected and cancerous cells. Its presence in the signature, even with minimal predictive weight, is consistent with the progressive decline of cytotoxic T cell function that is one of the best-characterised features of immune ageing [22]. *TMEM204* (coefficient −0.183) is a transmembrane protein involved in cell-cell interactions whose age-associated change is less well characterised in the literature.

In regard to the transcriptomic clock developed by *Sood et al.* [41], two of the genes identified by our methodology were shared with their 150-gene signature: *POF1B* and *MAPK8IP2*. No gene was common to all three gene sets simultaneously, which is expected given the very different methodologies employed by each transcriptomic clock.

POF1B (coefficient +0.920) is involved in organising the actin cytoskeleton, the internal network of protein filaments that gives cells their shape and allows them to move and divide. Its upregulation with age in blood, replicated in *Sood et al.* [41], is consistent with age-associated changes in the properties of circulating immune cells that accompany immunosenescence [1]. MAPK8IP2 (coefficient +0.409) is a protein that organises the

cellular processes responsible for responding to stress, such as DNA damage, oxidative stress, and inflammation, all of which increase in frequency with age. Its upregulation in both signatures points to progressively elevated stress signalling in aged blood cells, consistent with the inflammaging framework [10].

4.3.3 Top Ranked Unique Genes

This section examines genes ranked highly by the interpretive methods established in Section 4.3.1, focusing on two complementary questions: which genes are ranked highly by both the linear Elastic Net model and the non-linear XGBoost model, and which genes are ranked highly by only one. Genes important to both models represent the strongest candidates for genuine age-associated signal, as their importance does not depend on an assumption of linearity. Genes important to only one method are informative in distinct ways: a gene ranked highly only by Elastic Net may reflect a linear association not deemed important by the gradient boosting model, while a gene ranked highly only by XGBoost may capture a non-linear or interaction effect invisible to a linear model.

Cross-referencing the top 30 genes by absolute Elastic Net coefficient against the top 30 genes by mean absolute XGBoost SHAP value identifies ten genes that rank highly in both: *CD248*, and *LRRN3*, already discussed in Section 4.3.2 for their overlap with *Peters et al.* [7], and *IGFBP3, PODXL2, ASIC1, TUBA1B-AS1, FIGN*, and *NRCAM*, along with two genes lacking an annotated symbol in the current *GENCODE* release. The latter group is discussed below. Two notable disagreements between the methods are also discussed: *GOLPH3L*, the single most strongly weighted gene by Elastic Net coefficient, which is absent from XGBoost's top 30; and *REG4* and *GREM2*, which rank highly in XGBoost but do not appear among Elastic Net's top 30.

FIGN (fidgetin; Elastic Net rank 10, coefficient +1.003; XGBoost SHAP rank 1) shows the strongest agreement of any gene in the signature, ranking first in non-linear importance despite a comparatively modest linear coefficient. FIGN controls the dynamics of microtubules, the internal protein rods that form the structural skeleton of cells and are essential for cell division and the transport of molecules inside cells. Its upregulation with age suggests progressive disruption of this internal structural system, consistent with the cytoskeletal dysfunction and impaired cell division fidelity observed in aged cells [1].

PODXL2 (Elastic Net rank 5, coefficient −1.134; XGBoost SHAP rank 4) and *ASIC1* (Elastic Net rank 6, coefficient −1.128; XGBoost SHAP rank 11) are both consistently ranked

highly by both models. ASIC1 is an ion channel, a protein that acts as a gate in the cell membrane, controlling the flow of charged particles into cells in response to acidic conditions, which typically signal inflammation or tissue damage. Its downregulation with age may reflect a declining capacity of aged immune cells to sense and respond to inflammatory signals. PODXL2 encodes a cell-surface glycoprotein implicated in cell adhesion and is less well characterised in the ageing literature, though its consistent downregulation across both models makes it a candidate for further investigation.

IGFBP3 (Elastic Net rank 3, coefficient +1.380; XGBoost SHAP rank 8) is the most strongly upregulated gene by linear coefficient and also within the top 10 of the non-linear ranking, corroborating its importance across both models. It is a protein that circulates in the blood and controls the availability of insulin-like growth factor 1 (IGF-1), a hormone that promotes cell growth and division. With age, the balance of this growth signalling system shifts: IGF-1 levels decline and IGFBP3 increasingly promotes signals that inhibit growth and trigger cellular senescence. The IGF-1/IGFBP3 axis is one of the most evolutionarily conserved regulators of lifespan, implicated in longevity across species from worms to humans [10].

Meanwhile, *GOLPH3L* (Elastic Net rank 1, coefficient −1.582) is the most strongly downregulated gene by linear coefficient, yet does not appear in the XGBoost top 30, suggesting its apparent importance may be specific to the linear model. GOLPH3L is involved in regulating protein transport inside cells, coordinating the movement of proteins from the Golgi apparatus to their final destinations; its proposed link to proteostasis failure, a recognised hallmark of cellular ageing [1], should therefore be treated as a hypothesis warranting independent validation rather than a confirmed marker.

Conversely, *REG4* and *GREM2* rank highly in XGBoost (SHAP ranks 5 and 6 respectively) but fall outside the Elastic Net top 30, illustrating that the non-linear model captures additional age-associated signal not well represented by a linear combination of gene expression values.

4.3.4 Overall Signature Profile

Taken together, the genes examined in Sections 4.3.2 and 4.3.3 reveal a directional profile consistent with established transcriptomic hallmarks of blood ageing. Upregulated genes are enriched in growth factor signalling (*IGFBP3*) and cytoskeletal remodelling (*FIGN, POF1B*), while downregulated genes include markers of vascular integrity (*CD248,*

PODXL2), neuronal-immune cross-talk (*LRRN3, ASIC1*), and lymphocyte function (*CR2*). This profile mirrors the findings of *Peters et al.* [7] and the universal transcriptomic hallmarks of mammalian ageing reported by *Tyshkovskiy et al.* [11], supporting the conclusion that the pipeline has captured genuine transcriptomic ageing signal.

This conclusion is strengthened by the cross-model agreement established in Section 4.3.3: genes such as *CD248, IGFBP3, LRRN3, PODXL2, ASIC1*, and *FIGN* are consistently ranked as important by both the linear Elastic Net model and the non-linear XGBoost model, indicating that their contribution to age prediction does not depend on the specific modelling assumptions of either method. At the same time, the disagreements identified between the two models, most notably *GOLPH3L*'s strong linear coefficient unsupported by XGBoost's importance ranking, and the converse pattern for *REG4* and *GREM2*, are a reminder that not every gene in the signature carries equally robust evidence, and that interpretability findings from a single model should be treated with appropriate caution.

4.4. Protein-Protein Interaction Network of the Gene Signature

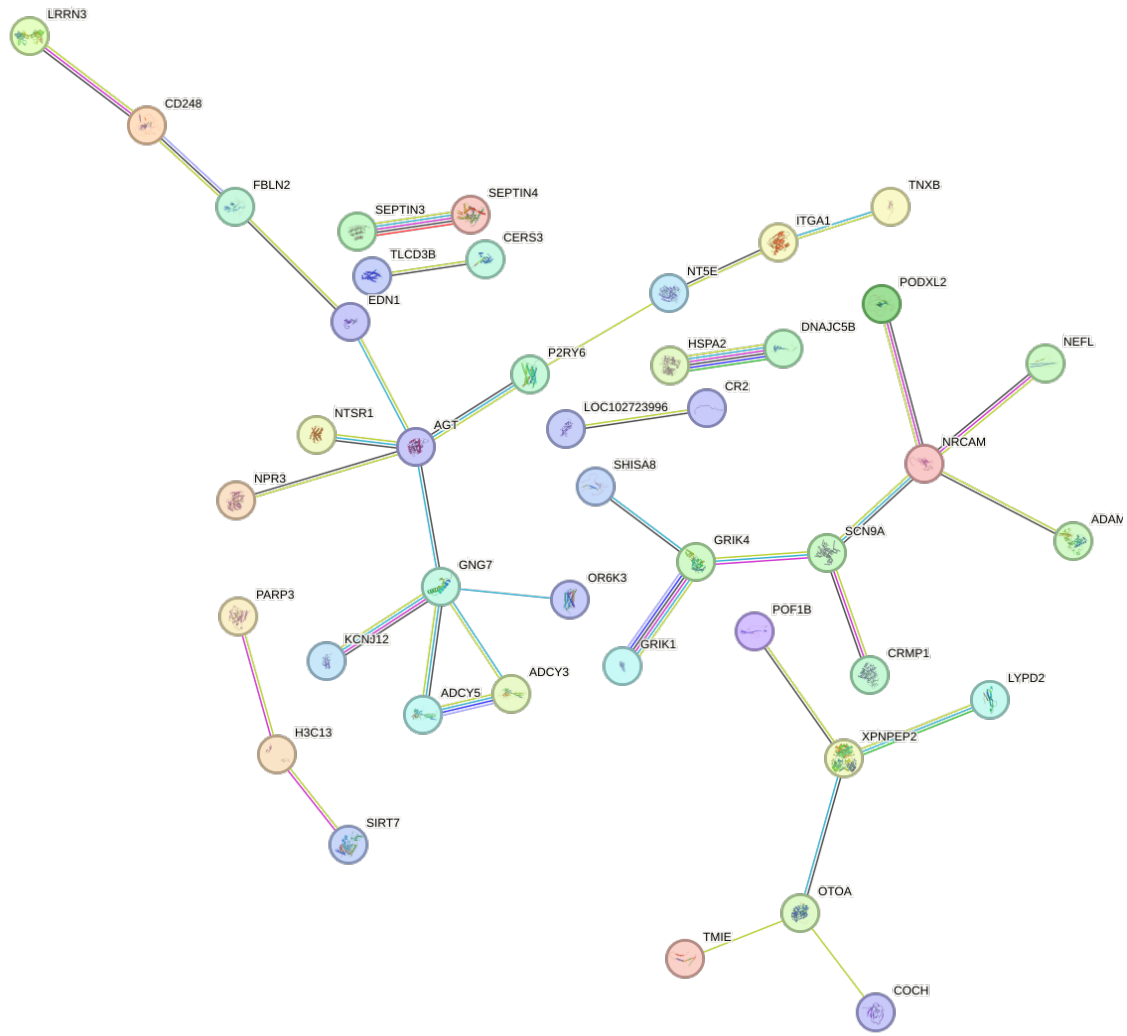


Figure 4.7: STRING protein-protein interaction network of the 200-gene Ridge signature.

Proteins rarely act in isolation; rather, they interact with other proteins to carry out biological functions, forming networks of connected molecules that work together in the same pathways or complexes. Submission of the 200 Ridge selected genes to STRING [90], a database of known and predicted protein-protein interactions, yielded a significant enrichment result ($p = 0.0403$), indicating that the proteins encoded by the signature have significantly more connections between them than would be expected for a random set of 200 proteins from the human genome. This suggests that the signature encodes a functionally connected biological network rather than a collection of individually predictive but

biologically unrelated genes, consistent with the use of Ridge-based ranking, which preserves co-expressed gene groups and therefore naturally retains pathway-level structure in the selected set. Each node in the resulting network represents a protein encoded by a signature gene, and each edge represents a known or predicted interaction between two proteins; nodes with no detected interactions are omitted. Figure 4.7 shows the resulting network, with two clusters of particular interest discussed below.

The largest cluster centres on AGT (angiotensin precursor) and directly connects two of the signature's strongest individual predictors, *CD248* and *LRRN3* (Section 4.3.2), via the intermediate genes *FBLN2* and *EDN1*. AGT itself carries a near-zero Elastic Net coefficient and ranks poorly in XGBoost's SHAP importance (rank 182 of 200), indicating it contributes little to either model's predictions individually; its hub position therefore reflects biological centrality rather than independent predictive value, consistent with redundancy among co-regulated genes in this region of the network. AGT is the starting point of the renin-angiotensin system, a hormonal pathway that regulates blood pressure and fluid balance and becomes progressively dysregulated with age, contributing to vascular stiffening, chronic inflammation, and cardiovascular disease [1]. The cluster also includes *NPR3* (Elastic Net coefficient −0.735 rank 34, XGBoost SHAP rank 30) and *P2RY6* (Elastic Net coefficient −0.386 rank 106, XGBoost SHAP rank 49), both moderately important receptors involved in blood pressure regulation and vascular signalling, and the *GNG7/ADCY3/ADCY5* signalling chain (a G-protein/cAMP signalling pathway), downregulated with age (coefficients −0.339, −0.571, −0.296 respectively) and ranked within the top 60 of 200 genes by XGBoost importance, indicating a coordinated reduction in cellular responsiveness to hormonal signals with age [10]. Taken together, this cluster links the strongest individually predictive genes in the signature, *CD248* and *LRRN3*, to a broader, biologically coherent network of vascular and hormonal regulation in blood.

The second cluster of interest connects *NRCAM* (Elastic Net coefficient −0.913 rank 14, XGBoost SHAP rank 12), *SCN9A* (Elastic Net coefficient −0.424 rank 93, XGBoost SHAP rank 118), and *PODXL2* (Elastic Net coefficient −1.134 rank 5, XGBoost SHAP rank 4), three genes involved in cell adhesion and electrical signalling that are also relevant to immune cell biology. NRCAM and PODXL2 are both strongly supported by both models, while SCN9A's importance is more specific to the linear model. Their co-downregulation and network connectivity reinforce the neuronal-immune cross-talk theme observed in the gene-level analysis (Section 4.3.4), suggesting that the progressive loss of this signalling capacity in blood cells is a coordinated, pathway-level feature of ageing rather than an

isolated single-gene effect.

Several smaller clusters were also identified in the network, involving genes of lower individual predictive importance (e.g. *NT5E, GRIK4, XPNPEP2* and their respective connections); these are not discussed further here but are visible in Figure 4.7.

4.5. Technical Strengths

In regards to the ML methodology described in this thesis, the DGE pre-filtering step via limma-voom on raw count data ensures that every gene entering feature selection carries a statistically demonstrated age association at FDR < 0.05 from a covariate-adjusted model, providing biological grounding absent from purely data-driven pipelines. This mirrors the standard workflow in the field [7] and makes the selected genes more meaningful for biological interpretation and cross-study comparison. Furthermore, the strict prevention of data leakage throughout (test set held out from all processing decisions and scalers fitted inside cross-validation folds) ensures the reported evaluation metrics reflect genuine generalisation to unseen data. Data leakage is a common and often underreported issue in machine learning pipelines that produces optimistic performance estimates that do not hold in practice [29].

For gene signature interpretation, Ridge based gene ranking preserves co-regulated gene groups rather than arbitrarily eliminating correlated partners as LASSO would. This directly contributed to the biologically coherent protein-protein interaction network structure observed in the final signature. The low optimal Elastic Net *l1_ratio* (0.15) confirmed that the age signal is distributed across many correlated genes, consistent with the polygenic nature of transcriptomic ageing, making Ridge the more appropriate ranking tool for this dataset.

Interpretability was further strengthened by validating Elastic Net's directional findings against SHAP values computed directly on XGBoost, (Section 4.3.1). Rather than relying on a single model's interpretation, this dual approach identified a core set of genes whose importance is robust to the choice of model, and explicitly flagged genes where the two methods disagree (Section 4.3.4), providing a more honest and defensible account of the gene signature than either method could offer alone.

Finally, the pipeline is fully scripted and publicly available on GitHub at `https://github.com/martacduarte/blood-gene-expression-ageing-prediction`. The use of fixed random seeds throughout and data validation at each filtering step to verify that expression

matrix identifiers, age labels, and metadata retain the expected order ensures full reproducibility and prevents silent misalignment errors that are a common source of data leakage in transcriptomic pipelines. The provided pipeline code directly addresses the absence of standardised pipelines identified as a key limitation in the transcriptomic clock literature [4].

5. Conclusions

This thesis developed and evaluated a reproducible machine learning pipeline for predicting chronological age from whole blood RNA-seq gene expression data, using the resulting age predictions to derive age acceleration (ΔAge) as a proxy for biological ageing, with GTEx v11 as the primary data resource. Starting from a raw expression matrix of 74,628 genes across 803 samples, the pipeline applied $\log_2$(TPM+1) transformation, ComBat batch correction using sequencing centre as the batch variable, RNA quality filtering (RIN $\geq$ 6.5), and outlier removal, yielding a final cohort of 757 samples. A stratified 80/20 train-test split (Section 3.3) then separated 605 training samples from a held-out test set of 152 donors, with all subsequent steps performed exclusively on the training set to prevent data leakage. Differential gene expression analysis via limma-voom on the training samples identified 15,217 age-associated genes at FDR < 0.05 from a covariate-adjusted linear model, providing a biologically justified input space for feature selection. Ridge based ranking then identified the 200 most informative genes, which served as input to a multi-model regression framework comprising Elastic Net, Random Forest, XGBoost, and a Neural Network.

XGBoost was chosen over the weighted ensemble (Section 4.1), achieving a test MAE of 7.35 years and R^2 of 0.506 on the held-out set of 152 donors, consistent with the state of the art for blood transcriptomic clocks trained on comparable data scales [7]. Predictions showed an age-dependent bias consistent with the imbalanced training distribution, with younger donors overestimated and older donors underestimated (Section 4.1). Biological validation of the gene signature, interpreted using Elastic Net coefficients for direction and SHAP values computed on XGBoost for importance, confirmed cross-study replication with both *Peters et al.* [7] (7 shared genes) and *Sood et al.* [41] (2 shared genes), with functionally relevant markers of immunosenescence, tryptophan metabolism, vascular ageing, and intracellular trafficking. STRING protein-protein interaction network analysis confirmed significant interaction enrichment among the 200 selected genes (p = 0.0403), with biologically interpretable hub clusters pointing to the renin-angiotensin axis, G-protein/cAMP signalling, and neuronal-immune cross-talk as coordinated pathway-level features of the blood ageing signature.

5.1. Answers to the Research Questions

Research question 1: Can robust and accurate models be developed to predict biological age from blood gene expression data?

Yes, within the constraints of the available dataset. The final model, XGBoost, achieves MAE of 7.35 years and R^2 of 0.506 on a held-out test set, matching the performance range reported by *Peters et al.* [7] across independent cohorts despite using substantially fewer training samples. Biological age is operationalised in this thesis as age acceleration (ΔAge = $\hat{y} - y$, Section 2.3.1); while the mean ΔAge across the test set is close to zero (+0.12 years), per-age-group analysis (Section 4.1) reveals systematic over-prediction in donors under 40 and under-prediction in donors over 60, a direct consequence of the age-imbalanced training distribution. Performance and calibration are therefore constrained primarily by the small and age-imbalanced GTEx whole blood training set and would be expected to improve with larger, more balanced cohorts.

Research question 2: Can we overcome technical challenges posed by incomplete and imbalanced distributions of the response variable?

Partially. Stratified splitting with quantile-based age bins ensured statistically indistinguishable train and test age distributions (Section 3.3), addressing the evaluation side of the imbalance problem. However, the imbalance in the training distribution itself was not successfully corrected: SMOGN synthetic oversampling failed to improve prediction accuracy and degraded performance on all three metrics (Section 3.7), and per-age-group ΔAge analysis (Section 4.1) confirms the underlying imbalance still limits the model's sensitivity at the extremes of the age range, with median ΔAge ranging from +13.8 years in the 30–40 group to −7.9 years in the 60–70 group. Full resolution of this challenge requires larger training datasets with broader age coverage rather than synthetic augmentation.

Research question 3: Can we build models that generalise well across independent blood gene expression datasets?

This question remains open. The pipeline was evaluated exclusively on a held-out subset of the GTEx cohort, which shares the same collection protocol, sequencing centre distribution, and donor demographic profile as the training data, so cross-cohort generalisation was not feasible to demonstrate within the scope of this work. However, the pipeline was deliberately designed to support this validation step: Ridge based gene ranking was selected over dimensionality-reduction approaches such as PCA specifically to preserve

individual gene identities (Section 2.8), so the final 200-gene signature can in principle be applied directly to independent cohorts without re-deriving the feature space. Demonstrating this generalisation remains the most important outstanding validation step before clinical utility can be claimed.

Research question 4: Is it possible to identify a minimal age signature for accurate age prediction, and could such a signature serve as the basis for a practical age prediction device?

A compact 200-gene signature was identified that achieves competitive prediction performance, with cross-study replication in two independent reference signatures and a statistically significant protein-protein network confirming functional connectivity. Whether 200 genes constitutes a minimal signature, however, is not something this work can answer definitively. For context, established clinically deployed gene expression panels vary considerably in size, from 21 genes in the *Oncotype DX* breast cancer recurrence assay [91] to 70 genes in the *MammaPrint* assay [92], suggesting 200 genes is not inherently impractical for this class of diagnostic tool. However, determining whether this specific signature could be meaningfully reduced, and whether a smaller panel would maintain comparable predictive performance, would require dedicated feature reduction experiments and validation on independent cohorts, which fall outside the scope of this thesis and are identified as a priority for future work (Section 5.3). Similarly, whether the present pipeline constitutes a practical ”device” depends on how that term is defined: as a data-driven bioinformatic pipeline (blood sample $\rightarrow$ sequencing $\rightarrow$ computational prediction), the work presented here is already functional in that sense, but translation to a point-of-care physical assay would require the panel reduction and validation steps just described.

5.2. Limitations

The small and age-imbalanced training set is the primary limitation, producing the systematic ΔAge bias reported in Section 4.1 and limiting the model's exposure to young blood transcriptomes (Sections 3.7, 4.2). Sample size also constrained model complexity, preventing effective use of deep learning architectures.

The pipeline was trained and evaluated within a single cohort, meaning cross-cohort generalisation remains undemonstrated. Both training and test sets share the same GTEx collection protocol, sequencing centre distribution, and post-mortem sampling strategy,

and drops in performance when evaluated on independent datasets collected under different conditions is a known challenge for transcriptomic clocks [7, 27]. *ComBat* correction reduced but did not fully eliminate between-centre batch effects, and residual technical variation may inflate apparent age associations in a subset of genes.

Finally, the models are optimised for prediction rather than causal inference. Strong gene coefficients indicate correlation with age, not causation. Therefore, the signature may partially capture the downstream consequences of age-associated processes, such as immune cell compositional shifts, medication effects and subclinical disease, rather than intrinsic molecular ageing mechanisms.

5.3. Future Work

The most impactful near-term addition would be to perform external validation on independent, publicly available cohorts such as those used in the Peters et al. meta-analysis [7] (Framingham Heart Study, Rotterdam Study, *InChianti, KORA, SHIP-TREND*), accessible through GEO and dbGaP. This would establish whether the clock maintains its performance outside the GTEx context and if the uncovered gene signatures should be considered for experimental validation. Addressing the age imbalance requires larger datasets with broader and more uniform age coverage. The UK Biobank blood RNA-seq database [70], which is progressively releasing expression data from its 500,000-participant cohort, represents the most significant near-term resource for this purpose and would be expected to substantially improve prediction accuracy, particularly at the extremes of the age range. Integration of cell type deconvolution approaches such as CIBERSORTx, a computational tool that estimates the proportions of different immune cell types present in a blood sample from gene expression data alone, would allow disentanglement of expression changes driven by immune cell compositional shifts from those reflecting intrinsic cell-level ageing, moving the signature towards more robust biological interpretability. Finally, the identified 200-gene signature represents a candidate panel of blood-based ageing biomarkers, comparable in scale to existing clinically deployed gene expression panels (Section 5.1). Reduction to a smaller targeted panel compatible with platforms such as *NanoString* [93], a targeted gene expression measurement technology suitable for clinical settings that can measure hundreds of specific genes from a small blood sample, followed by prospective validation in clinical cohorts with longitudinal follow-up, would establish whether predicted age acceleration associates with health outcomes and mortality.

References

[1] C. López-Otín, M. A. Blasco, L. Partridge, M. Serrano, and G. Kroemer, “The hallmarks of aging,” *Cell*, vol. 153, no. 6, p. 1194–1217, Jun. 2013. [Online]. Available: http://dx.doi.org/10.1016/j.cell.2013.05.039 [Cited on pages 1, 5, 58, 59, 60, 61, and 64.]

[2] L. Partridge, J. Deelen, and P. E. Slagboom, “Facing up to the global challenges of ageing,” *Nature*, vol. 561, no. 7721, p. 45–56, Sep. 2018. [Online]. Available: http://dx.doi.org/10.1038/s41586-018-0457-8 [Cited on page 1.]

[3] B. K. Kennedy, S. L. Berger, A. Brunet, J. Campisi, A. M. Cuervo, E. S. Epel, C. Franceschi, G. J. Lithgow, R. I. Morimoto, J. E. Pessin, T. A. Rando, A. Richardson, E. E. Schadt, T. Wyss-Coray, and F. Sierra, “Geroscience: Linking aging to chronic disease,” *Cell*, vol. 159, no. 4, p. 709–713, Nov. 2014. [Online]. Available: http://dx.doi.org/10.1016/j.cell.2014.10.039 [Cited on page 1.]

[4] J. Rutledge, H. Oh, and T. Wyss-Coray, “Measuring biological age using omics data,” *Nature Reviews Genetics*, vol. 23, no. 12, p. 715–727, Jun. 2022. [Online]. Available: http://dx.doi.org/10.1038/s41576-022-00511-7 [Cited on pages 2, 13, 17, 21, 22, 23, and 66.]

[5] S. Horvath, “Dna methylation age of human tissues and cell types,” *Genome Biology*, vol. 14, no. 10, Dec. 2013. [Online]. Available: http://dx.doi.org/10.1186/gb-2013-14-10-r115 [Cited on pages 2, 11, 12, and 14.]

[6] M. E. Levine, A. T. Lu, A. Quach, B. H. Chen, T. L. Assimes, S. Bandinelli, L. Hou, A. A. Baccarelli, J. D. Stewart, Y. Li, E. A. Whitsel, J. G. Wilson, A. P. Reiner, A. Aviv, K. Lohman, Y. Liu, L. Ferrucci, and S. Horvath, “An epigenetic biomarker of aging for lifespan and healthspan,” *Aging*, vol. 10, no. 4, p. 573–591, Apr. 2018. [Online]. Available: http://dx.doi.org/10.18632/aging.101414 [Cited on pages 2, 11, and 14.]

[7] M. J. Peters, R. Joehanes, L. C. Pilling, C. Schurmann, K. N. Conneely, J. Powell, E. Reinmaa, G. L. Sutphin, A. Zhernakova, K. Schramm, Y. A. Wilson, S. Kobes, T. Tukiainen, M. A. Nalls, D. G. Hernandez, M. R. Cookson, R. J. Gibbs, J. Hardy, A. Ramasamy, A. B. Zonderman, A. Dillman, B. Traynor, C. Smith, D. L. Longo,

D. Trabzuni, J. Troncoso, M. van der Brug, M. E. Weale, R. O'Brien, R. Johnson, R. Walker, R. H. Zielke, S. Arepalli, M. Ryten, A. B. Singleton, Y. F. Ramos, H. H. H. Göring, M. Fornage, Y. Liu, S. A. Gharib, B. E. Stranger, P. L. De Jager, A. Aviv, D. Levy, J. M. Murabito, P. J. Munson, T. Huan, A. Hofman, A. G. Uitterlinden, F. Rivadeneira, J. van Rooij, L. Stolk, L. Broer, M. M. P. J. Verbiest, M. Jhamai, P. Arp, A. Metspalu, L. Tserel, L. Milani, N. J. Samani, P. Peterson, S. Kasela, V. Codd, A. Peters, C. K. Ward-Caviness, C. Herder, M. Waldenberger, M. Roden, P. Singmann, S. Zeilinger, T. Illig, G. Homuth, H.-J. Grabe, H. Völzke, L. Steil, T. Kocher, A. Murray, D. Melzer, H. Yaghootkar, S. Bandinelli, E. K. Moses, J. W. Kent, J. E. Curran, M. P. Johnson, S. Williams-Blangero, H.-J. Westra, A. F. McRae, J. A. Smith, S. L. R. Kardia, I. Hovatta, M. Perola, S. Ripatti, V. Salomaa, A. K. Henders, N. G. Martin, A. K. Smith, D. Mehta, E. B. Binder, K. M. Nylocks, E. M. Kennedy, T. Klengel, J. Ding, A. M. Suchy-Dicey, D. A. Enquobahrie, J. Brody, J. I. Rotter, Y.-D. I. Chen, J. Houwing-Duistermaat, M. Kloppenburg, P. E. Slagboom, Q. Helmer, W. den Hollander, S. Bean, T. Raj, N. Bakhshi, Q. P. Wang, L. J. Oyston, B. M. Psaty, R. P. Tracy, G. W. Montgomery, S. T. Turner, J. Blangero, I. Meulenbelt, K. J. Ressler, J. Yang, L. Franke, J. Kettunen, P. M. Visscher, G. G. Neely, R. Korstanje, R. L. Hanson, H. Prokisch, L. Ferrucci, T. Esko, A. Teumer, J. B. J. van Meurs, and A. D. Johnson, "The transcriptional landscape of age in human peripheral blood," *Nature Communications*, vol. 6, no. 1, Oct. 2015. [Online]. Available: http://dx.doi.org/10.1038/ncomms9570 [Cited on pages viii, 2, 6, 9, 10, 11, 12, 13, 14, 22, 23, 42, 54, 55, 57, 58, 59, 60, 62, 65, 67, 68, and 70.]

[8] Q. Chen, V. B. Dwaraka, N. Carreras-Gallo, J. F. Armstrong, R. Sehgal, M. A. Argentieri, A. Richmond, A. Aparicio, K. Mendez, Y. Chen, S. Begum, P. Kachroo, N. Prince, T. Guo, H. Went, T. Mendez, A. Lin, L. Turner, M. Moqri, S. H. Chu, R. S. Kelly, S. T. Weiss, N. J. W. Rattray, V. N. Gladyshev, E. Karlson, C. E. Wheelock, E. A. Mathé, A. Dahlin, M. J. McGeachie, R. E. Marioni, A. T. Higgins-Chen, R. Smith, and J. Lasky-Su, "Omicmage quantifies biological age by integrating multi-omics with electronic medical records," *Nature Aging*, vol. 6, no. 3, p. 722–737, Feb. 2026. [Online]. Available: http://dx.doi.org/10.1038/s43587-026-01073-7 [Cited on pages 2 and 18.]

[9] M. Moqri, C. Herzog, J. R. Poganik, J. Justice, D. W. Belsky, A. Higgins-Chen, A. Moskalev, G. Fuellen, A. A. Cohen, I. Bautmans, M. Widschwendter, J. Ding, A. Fleming, J. Mannick, J.-D. J. Han, A. Zhavoronkov, N. Barzilai, M. Kaeberlein,

S. Cummings, B. K. Kennedy, L. Ferrucci, S. Horvath, E. Verdin, A. B. Maier, M. P. Snyder, V. Sebastiano, and V. N. Gladyshev, “Biomarkers of aging for the identification and evaluation of longevity interventions,” *Cell*, vol. 186, no. 18, p. 3758–3775, Aug. 2023. [Online]. Available: http://dx.doi.org/10.1016/j.cell.2023.08.003 [Cited on pages 2, 5, and 23.]

[10] J. P. de Magalhães, J. Curado, and G. M. Church, “Meta-analysis of age-related gene expression profiles identifies common signatures of aging,” *Bioinformatics*, vol. 25, no. 7, p. 875–881, Feb. 2009. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btp073 [Cited on pages 2, 9, 13, 59, 60, 61, and 64.]

[11] A. Tyshkovskiy, D. Kholdina, M. Davitadze, A. Molière, A. Moldakozhayev, Y. Tongu, T. Kasahara, D. Glubokov, A. Eames, L. M. Kats, A. Vladimirova, K. Ying, H. Liu, B. Zhang, U. Khasanova, M. Moqri, J. M. Van Raamsdonk, D. E. Harrison, R. Strong, T. Abe, S. E. Dmitriev, and V. N. Gladyshev, “Universal transcriptomic hallmarks of mammalian ageing and mortality,” *Nature*, vol. 654, no. 8117, p. 173–188, May 2026. [Online]. Available: http://dx.doi.org/10.1038/s41586-026-10542-3 [Cited on pages 2, 9, 14, 22, and 62.]

[12] R. Chen, Y. Wang, S. Zhang, G. Bulloch, J. Zhang, H. Liao, X. Shang, M. Clark, Q. Peng, Z. Ge, C. Cheng, Y. Gao, M. He, and Z. Zhu, “Biomarkers of ageing: Current state□of□art, challenges, and opportunities,” *MedComm – Future Medicine*, vol. 2, no. 2, 2023. [Online]. Available: http://dx.doi.org/10.1002/mef2.50 [Cited on page 5.]

[13] E. H. Blackburn, E. S. Epel, and J. Lin, “Human telomere biology: A contributory and interactive factor in aging, disease risks, and protection,” *Science*, vol. 350, no. 6265, p. 1193–1198, Dec. 2015. [Online]. Available: http://dx.doi.org/10.1126/science.aab3389 [Cited on page 5.]

[14] J. H. Cole, S. J. Ritchie, M. E. Bastin, M. C. Valdés Hernández, S. Muñoz Maniega, N. Royle, J. Corley, A. Pattie, S. E. Harris, Q. Zhang, N. R. Wray, P. Redmond, R. E. Marioni, J. M. Starr, S. R. Cox, J. M. Wardlaw, D. J. Sharp, and I. J. Deary, “Brain age predicts mortality,” *Molecular Psychiatry*, vol. 23, no. 5, p. 1385–1392, Apr. 2017. [Online]. Available: http://dx.doi.org/10.1038/mp.2017.62 [Cited on page 6.]

[15] D. P. Leong, K. K. Teo, S. Rangarajan, P. Lopez-Jaramillo, A. Avezum, A. Orlandini, P. Seron, S. H. Ahmed, A. Rosengren, R. Kelishadi, O. Rahman, S. Swaminathan,

R. Iqbal, R. Gupta, S. A. Lear, A. Oguz, K. Yusoff, K. Zatonska, J. Chifamba, E. Igumbor, V. Mohan, R. M. Anjana, H. Gu, W. Li, and S. Yusuf, "Prognostic value of grip strength: findings from the prospective urban rural epidemiology (pure) study," *The Lancet*, vol. 386, no. 9990, p. 266–273, Jul. 2015. [Online]. Available: http://dx.doi.org/10.1016/S0140-6736(14)62000-6 [Cited on page 6.]

[16] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by rna-seq," *Nature Methods*, vol. 5, no. 7, p. 621–628, May 2008. [Online]. Available: http://dx.doi.org/10.1038/nmeth.1226 [Cited on pages 6 and 8.]

[17] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, and A. Mortazavi, "A survey of best practices for rna-seq data analysis," *Genome Biology*, vol. 17, no. 1, Jan. 2016. [Online]. Available: http://dx.doi.org/10.1186/s13059-016-0881-8 [Cited on pages 7, 8, 9, and 34.]

[18] G. P. Wagner, K. Kin, and V. J. Lynch, "Measurement of mrna abundance using rna-seq data: Rpkm measure is inconsistent among samples," *Theory in Biosciences*, vol. 131, no. 4, p. 281–285, Aug. 2012. [Online]. Available: http://dx.doi.org/10.1007/s12064-012-0162-3 [Cited on page 7.]

[19] Z. Wang, M. Gerstein, and M. Snyder, "Rna-seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, no. 1, p. 57–63, Jan. 2009. [Online]. Available: http://dx.doi.org/10.1038/nrg2484 [Cited on page 8.]

[20] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for rna-seq data with deseq2," *Genome Biology*, vol. 15, no. 12, Dec. 2014. [Online]. Available: http://dx.doi.org/10.1186/s13059-014-0550-8 [Cited on pages 9 and 14.]

[21] Michigan State University Research Technology Support Facility, "Fastqc tutorial & faq," https://rtsf.natsci.msu.edu/genomics/technical-documents/fastqc-tutorial-and-faq.aspx, accessed on October 2025. [Cited on page 9.]

[22] L. Ferrucci and E. Fabbri, "Inflammageing: chronic inflammation in ageing, cardiovascular disease, and frailty," *Nature Reviews Cardiology*, vol. 15, no. 9, p. 505–522, 2018. [Online]. Available: http://dx.doi.org/10.1038/s41569-018-0064-2 [Cited on pages 9 and 59.]

[23] E. B. van den Akker, W. M. Passtoors, R. Jansen, E. W. van Zwet, J. J. Goeman, M. Hulsman, V. Emilsson, M. Perola, G. Willemsen, B. W. Penninx, B. T. Heijmans, A. B. Maier, D. I. Boomsma, J. N. Kok, P. E. Slagboom, M. J. Reinders, and M. Beekman, "Meta□analysis on blood transcriptomic studies identifies consistently coexpressed protein–protein interaction modules as robust markers of human aging," *Aging Cell*, vol. 13, no. 2, p. 216–225, Nov. 2013. [Online]. Available: http://dx.doi.org/10.1111/acel.12160 [Cited on pages 9 and 59.]

[24] C. S. Möller-Levet, S. N. Archer, G. Bucca, E. E. Laing, A. Slak, R. Kabiljo, J. C. Y. Lo, N. Santhi, M. von Schantz, C. P. Smith, and D.-J. Dijk, "Effects of insufficient sleep on circadian rhythmicity and expression amplitude of the human blood transcriptome," *Proceedings of the National Academy of Sciences*, vol. 110, no. 12, Feb. 2013. [Online]. Available: http://dx.doi.org/10.1073/pnas.1217154110 [Cited on page 10.]

[25] E. L. Tsalik, R. Henao, M. Nichols, T. Burke, E. R. Ko, M. T. McClain, L. L. Hudson, A. Mazur, D. H. Freeman, T. Veldman, R. J. Langley, E. B. Quackenbush, S. W. Glickman, C. B. Cairns, A. K. Jaehne, E. P. Rivers, R. M. Otero, A. K. Zaas, S. F. Kingsmore, J. Lucas, V. G. Fowler, L. Carin, G. S. Ginsburg, and C. W. Woods, "Host gene expression classifiers diagnose acute respiratory illness etiology," *Science Translational Medicine*, vol. 8, no. 322, Jan. 2016. [Online]. Available: http://dx.doi.org/10.1126/scitranslmed.aad6873 [Cited on page 10.]

[26] M. Obeidat, N. Fishbane, Y. Nie, V. Chen, Z. Hollander, S. J. Tebbutt, Y. Bossé, R. T. Ng, B. E. Miller, B. McManus, S. Rennard, P. D. Paré, and D. D. Sin, "The effect of statins on blood gene expression in copd," *PLOS ONE*, vol. 10, no. 10, p. e0140022, Oct. 2015. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0140022 [Cited on page 10.]

[27] J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry, "Tackling the widespread and critical impact of batch effects in high-throughput data," *Nature Reviews Genetics*, vol. 11, no. 10, p. 733–739, Sep. 2010. [Online]. Available: http://dx.doi.org/10.1038/nrg2825 [Cited on pages 10, 12, 29, and 70.]

[28] A. Schroeder, O. Mueller, S. Stocker, R. Salowsky, M. Leiber, M. Gassmann, S. Lightfoot, W. Menzel, M. Granzow, and T. Ragg, "The rin: an rna integrity number

for assigning integrity values to rna measurements," *BMC Molecular Biology*, vol. 7, no. 1, Jan. 2006. [Online]. Available: http://dx.doi.org/10.1186/1471-2199-7-3 [Cited on pages 10, 30, and 31.]

[29] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer New York, 2009. [Online]. Available: http://dx.doi.org/10.1007/978-0-387-84858-7 [Cited on pages 10, 35, 42, 52, and 65.]

[30] J. G. Fleischer, R. Schulte, H. H. Tsai, S. Tyagi, A. Ibarra, M. N. Shokhirev, L. Huang, M. W. Hetzer, and S. Navlakha, "Predicting age from the transcriptome of human dermal fibroblasts," *Genome Biology*, vol. 19, no. 1, Dec. 2018. [Online]. Available: http://dx.doi.org/10.1186/s13059-018-1599-6 [Cited on pages 10, 11, and 12.]

[31] M. N. Shokhirev and A. A. Johnson, "Modeling the human aging transcriptome across tissues, health status, and sex," *Aging Cell*, vol. 20, no. 1, Dec. 2020. [Online]. Available: http://dx.doi.org/10.1111/acel.13280 [Cited on pages 11, 13, 22, and 23.]

[32] G. Hannum, J. Guinney, L. Zhao, L. Zhang, G. Hughes, S. Sadda, B. Klotzle, M. Bibikova, J.-B. Fan, Y. Gao, R. Deconde, M. Chen, I. Rajapakse, S. Friend, T. Ideker, and K. Zhang, "Genome-wide methylation profiles reveal quantitative views of human aging rates," *Molecular Cell*, vol. 49, no. 2, p. 359–367, Jan. 2013. [Online]. Available: http://dx.doi.org/10.1016/j.molcel.2012.10.016 [Cited on pages 12 and 14.]

[33] G. K. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, no. 1, p. 1–25, Jan. 2004. [Online]. Available: http://dx.doi.org/10.2202/1544-6115.1027 [Cited on page 12.]

[34] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 58, no. 1, p. 267–288, Jan. 1996. [Online]. Available: http://dx.doi.org/10.1111/j.2517-6161.1996.tb02080.x [Cited on pages 12 and 15.]

[35] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical bayes methods," *Biostatistics*, vol. 8, no. 1,

p. 118–127, Apr. 2006. [Online]. Available: http://dx.doi.org/10.1093/biostatistics/kxj037 [Cited on pages 12, 15, and 29.]

[36] D. H. Meyer and B. Schumacher, "Bit age: A transcriptome□based aging clock near the theoretical limit of accuracy," *Aging Cell*, vol. 20, no. 3, Mar. 2021. [Online]. Available: http://dx.doi.org/10.1111/acel.13320 [Cited on pages 12, 13, 14, 15, 22, 23, and 54.]

[37] S. Fong, K. Pabis, D. Latumalea, N. Dugersuren, M. Unfried, N. Tolwinski, B. Kennedy, and J. Gruber, "Principal component-based clinical aging clocks identify signatures of healthy aging and targets for clinical intervention," *Nature Aging*, vol. 4, no. 8, p. 1137–1152, Jun. 2024. [Online]. Available: http://dx.doi.org/10.1038/s43587-024-00646-8 [Cited on pages 12, 13, and 14.]

[38] T. J. LaRocca, A. N. Cavalier, and D. Wahl, "Repetitive elements as a transcriptomic marker of aging: Evidence in multiple datasets and models," *Aging Cell*, vol. 19, no. 7, Jun. 2020. [Online]. Available: http://dx.doi.org/10.1111/acel.13167 [Cited on page 13.]

[39] S. A. McCarroll, C. T. Murphy, S. Zou, S. D. Pletcher, C.-S. Chin, Y. N. Jan, C. Kenyon, C. I. Bargmann, and H. Li, "Comparing genomic expression patterns across species identifies shared transcriptional profile in aging," *Nature Genetics*, vol. 36, no. 2, p. 197–204, Jan. 2004. [Online]. Available: http://dx.doi.org/10.1038/ng1291 [Cited on page 13.]

[40] E. D. Sun, O. Y. Zhou, M. Hauptschein, N. Rappoport, L. Xu, P. Navarro Negredo, L. Liu, T. A. Rando, J. Zou, and A. Brunet, "Spatial transcriptomic clocks reveal cell proximity effects in brain ageing," *Nature*, vol. 638, no. 8049, p. 160–171, Dec. 2024. [Online]. Available: http://dx.doi.org/10.1038/s41586-024-08334-8 [Cited on page 13.]

[41] S. Sood, I. J. Gallagher, K. Lunnon, E. Rullman, A. Keohane, H. Crossland, B. E. Phillips, T. Cederholm, T. Jensen, L. J. van Loon, L. Lannfelt, W. E. Kraus, P. J. Atherton, R. Howard, T. Gustafsson, A. Hodges, and J. A. Timmons, "A novel multi-tissue rna diagnostic of healthy ageing relates to cognitive health status," *Genome Biology*, vol. 16, no. 1, Sep. 2015. [Online]. Available: http://dx.doi.org/10.1186/s13059-015-0750-x [Cited on pages 14, 23, 42, 59, and 67.]

[42] A. J. Clark and J. W. Lillard, “A comprehensive review of bioinformatics tools for genomic biomarker discovery driving precision oncology,” *Genes*, vol. 15, no. 8, p. 1036, Aug. 2024. [Online]. Available: http://dx.doi.org/10.3390/genes15081036 [Cited on pages 14 and 15.]

[43] J. T. Leek and J. D. Storey, “Capturing heterogeneity in gene expression studies by surrogate variable analysis,” *PLoS Genetics*, vol. 3, no. 9, p. e161, Sep. 2007. [Online]. Available: http://dx.doi.org/10.1371/journal.pgen.0030161 [Cited on pages 15 and 29.]

[44] D. Risso, J. Ngai, T. P. Speed, and S. Dudoit, “Normalization of rna-seq data using factor analysis of control genes or samples,” *Nature Biotechnology*, vol. 32, no. 9, p. 896–902, Aug. 2014. [Online]. Available: http://dx.doi.org/10.1038/nbt.2931 [Cited on pages 15 and 29.]

[45] Springer-Verlag, 2002. [Online]. Available: http://dx.doi.org/10.1007/b98835 [Cited on page 15.]

[46] C. B. Steen, C. L. Liu, A. A. Alizadeh, and A. M. Newman, *Profiling Cell Type Abundance and Expression in Bulk Tissues with CIBERSORTx*. Springer US, 2020, p. 135–157. [Online]. Available: http://dx.doi.org/10.1007/978-1-0716-0301-7_7 [Cited on page 15.]

[47] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, no. 2, p. 301–320, Mar. 2005. [Online]. Available: http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x [Cited on page 15.]

[48] A. E. Hoerl and R. W. Kennard, “Ridge regression: Biased estimation for nonorthogonal problems,” *Technometrics*, vol. 12, no. 1, p. 55–67, Feb. 1970. [Online]. Available: http://dx.doi.org/10.1080/00401706.1970.10488634 [Cited on page 15.]

[49] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, p. 5–32, Oct. 2001. [Online]. Available: http://dx.doi.org/10.1023/A:1010933404324 [Cited on pages 15 and 39.]

[50] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge*

Discovery and Data Mining, ser. KDD '16. ACM, Aug. 2016, p. 785–794. [Online]. Available: http://dx.doi.org/10.1145/2939672.2939785 [Cited on pages 42 and 46.]

[51] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[52] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001. [Online]. Available: http://www.jstor.org/stable/2699986 [Cited on page 15.]

[53] C. Angermueller, T. Pärnamaa, L. Parts, and O. Stegle, "Deep learning for computational biology," *Molecular Systems Biology*, vol. 12, no. 7, Jul. 2016. [Online]. Available: http://dx.doi.org/10.15252/msb.20156651 [Cited on page 16.]

[54] J. Ma, M. K. Yu, S. Fong, K. Ono, E. Sage, B. Demchak, R. Sharan, and T. Ideker, "Using deep learning to model the hierarchical structure and function of a cell," *Nature Methods*, vol. 15, no. 4, p. 290–298, Mar. 2018. [Online]. Available: http://dx.doi.org/10.1038/nmeth.4627 [Cited on page 16.]

[55] N. Fortelny and C. Bock, "Knowledge-primed neural networks enable biologically interpretable deep learning on single-cell sequencing data," *Genome Biology*, vol. 21, no. 1, Aug. 2020. [Online]. Available: http://dx.doi.org/10.1186/s13059-020-02100-5 [Cited on page 16.]

[56] Y. Zhao, J. Shao, and Y. W. Asmann, "Assessment and optimization of explainable machine learning models applied to transcriptomic data," *Genomics, Proteomics amp; Bioinformatics*, vol. 20, no. 5, p. 899–911, Aug. 2022. [Online]. Available: http://dx.doi.org/10.1016/j.gpb.2022.07.003 [Cited on page 16.]

[57] C. Miller, T. Portlock, D. M. Nyaga, and J. M. O'Sullivan, "A review of model evaluation metrics for machine learning in genetics and genomics," *Frontiers in Bioinformatics*, vol. 4, Sep. 2024. [Online]. Available: http://dx.doi.org/10.3389/fbinf.2024.1457619 [Cited on page 17.]

[58] N. Rivero-Segura, O. Bello-Chavolla, O. Barrera-Vázquez, L. Gutierrez-Robledo, and J. Gomez-Verjan, "Promising biomarkers of human aging: In search of

a multi-omics panel to understand the aging process from a multidimensional perspective," *Ageing Research Reviews*, vol. 64, p. 101164, Dec. 2020. [Online]. Available: http://dx.doi.org/10.1016/j.arr.2020.101164 [Cited on page 17.]

[59] E. Macdonald-Dunlop, N. Taba, L. Klarić, A. Frkatović, R. Walker, C. Hayward, T. Esko, C. Haley, K. Fischer, J. F. Wilson, and P. K. Joshi, "A catalogue of omics biological ageing clocks reveals substantial commonality and associations with disease risk," *Aging*, vol. 14, no. 2, p. 623–659, Jan. 2022. [Online]. Available: http://dx.doi.org/10.18632/aging.203847 [Cited on page 18.]

[60] R. Argelaguet, B. Velten, D. Arnol, S. Dietrich, T. Zenz, J. C. Marioni, F. Buettner, W. Huber, and O. Stegle, "Multi□omics factor analysis—a framework for unsupervised integration of multi□omics data sets," *Molecular Systems Biology*, vol. 14, no. 6, Jun. 2018. [Online]. Available: http://dx.doi.org/10.15252/msb.20178124 [Cited on page 18.]

[61] A. Moraes, M. Moreno, R. Ribeiro, and P. G. Ferreira, *Predicting Age from Human Lung Tissue Through Multi-modal Data Integration*. Springer Nature Switzerland, 2023, p. 644–658. [Online]. Available: http://dx.doi.org/10.1007/978-3-031-45275-8_43 [Cited on pages 18, 22, 23, and 54.]

[62] R. Ribeiro, A. Moraes, M. Moreno, and P. G. Ferreira, "Integration of multi-modal datasets to estimate human aging," *Machine Learning*, vol. 113, no. 10, p. 7293–7317, Jul. 2024. [Online]. Available: http://dx.doi.org/10.1007/s10994-024-06588-x [Cited on pages 18, 22, 23, and 54.]

[63] F. Aguet, S. Anand, K. G. Ardlie, S. Gabriel, G. A. Getz, A. Graubert, K. Hadley, R. E. Handsaker, K. H. Huang, S. Kashin, X. Li, D. G. MacArthur, S. R. Meier, J. L. Nedzel, D. T. Nguyen, A. V. Segrè, E. Todres, B. Balliu, A. N. Barbeira, A. Battle, R. Bonazzola, A. Brown, C. D. Brown, S. E. Castel, D. F. Conrad, D. J. Cotter, N. Cox, S. Das, O. M. de Goede, E. T. Dermitzakis, J. Einson, B. E. Engelhardt, E. Eskin, T. Y. Eulalio, N. M. Ferraro, E. D. Flynn, L. Fresard, E. R. Gamazon, D. Garrido-Martín, N. R. Gay, M. J. Gloudemans, R. Guigó, A. R. Hame, Y. He, P. J. Hoffman, F. Hormozdiari, L. Hou, H. K. Im, B. Jo, S. Kasela, M. Kellis, S. Kim-Hellmuth, A. Kwong, T. Lappalainen, X. Li, Y. Liang, S. Mangul, P. Mohammadi, S. B. Montgomery, M. Muñoz-Aguirre, D. C. Nachun, A. B. Nobel, M. Oliva, Y. Park, Y. Park, P. Parsana, A. S. Rao, F. Reverter,

J. M. Rouhana, C. Sabatti, A. Saha, M. Stephens, B. E. Stranger, B. J. Strober, N. A. Teran, A. Viñuela, G. Wang, X. Wen, F. Wright, V. Wucher, Y. Zou, P. G. Ferreira, G. Li, M. Melé, E. Yeger-Lotem, M. E. Barcus, D. Bradbury, T. Krubit, J. A. McLean, L. Qi, K. Robinson, N. V. Roche, A. M. Smith, L. Sobin, D. E. Tabor, A. Undale, J. Bridge, L. E. Brigham, B. A. Foster, B. M. Gillard, R. Hasz, M. Hunter, C. Johns, M. Johnson, E. Karasik, G. Kopen, W. F. Leinweber, A. McDonald, M. T. Moser, K. Myer, K. D. Ramsey, B. Roe, S. Shad, J. A. Thomas, G. Walters, M. Washington, J. Wheeler, S. D. Jewell, D. C. Rohrer, D. R. Valley, D. A. Davis, D. C. Mash, P. A. Branton, L. K. Barker, H. M. Gardiner, M. Mosavel, L. A. Siminoff, P. Flicek, M. Haeussler, T. Juettemann, W. J. Kent, C. M. Lee, C. C. Powell, K. R. Rosenbloom, M. Ruffier, D. Sheppard, K. Taylor, S. J. Trevanion, D. R. Zerbino, N. S. Abell, J. Akey, L. Chen, K. Demanelis, J. A. Doherty, A. P. Feinberg, K. D. Hansen, P. F. Hickey, F. Jasmine, L. Jiang, R. Kaul, M. G. Kibriya, J. B. Li, Q. Li, S. Lin, S. E. Linder, B. L. Pierce, L. F. Rizzardi, A. D. Skol, K. S. Smith, M. Snyder, J. Stamatoyannopoulos, H. Tang, M. Wang, L. J. Carithers, P. Guan, S. E. Koester, A. R. Little, H. M. Moore, C. R. Nierras, A. K. Rao, J. B. Vaught, and S. Volpi, “The gtex consortium atlas of genetic regulatory effects across human tissues,” *Science*, vol. 369, no. 6509, p. 1318–1330, Sep. 2020. [Online]. Available: http://dx.doi.org/10.1126/science.aaz1776 [Cited on pages 19, 26, 29, and 30.]

[64] T. R. Dawber, G. F. Meadors, and F. E. Moore, “Epidemiological approaches to heart disease: The framingham study,” *American Journal of Public Health and the Nations Health*, vol. 41, no. 3, p. 279–286, Mar. 1951. [Online]. Available: http://dx.doi.org/10.2105/ajph.41.3.279 [Cited on page 20.]

[65] A. Hofman, S. D. Murad, C. M. van Duijn, O. H. Franco, A. Goedegebure, M. Arfan Ikram, C. C. W. Klaver, T. E. C. Nijsten, R. P. Peeters, B. H. C. Stricker, H. W. Tiemeier, A. G. Uitterlinden, and M. W. Vernooij, “The rotterdam study: 2014 objectives and design update,” *European Journal of Epidemiology*, vol. 28, no. 11, p. 889–926, Nov. 2013. [Online]. Available: http://dx.doi.org/10.1007/s10654-013-9866-z [Cited on page 20.]

[66] L. Ferrucci, S. Bandinelli, E. Benvenuti, A. Di Iorio, C. Macchi, T. B. Harris, and J. M. Guralnik, “Subsystems contributing to the decline in ability to walk: Bridging the gap between epidemiology and geriatric practice in the inchianti study,” *Journal of the American Geriatrics Society*, vol. 48, no. 12, p. 1618–1625, Dec. 2000.

[Online]. Available: http://dx.doi.org/10.1111/j.1532-5415.2000.tb03873.x [Cited on page 20.]

[67] R. Holle, M. Happich, H. Löwel, and H. Wichmann, “Kora - a research platform for population based health research,” *Das Gesundheitswesen*, vol. 67, no. S 01, p. 19–25, 2005. [Online]. Available: http://dx.doi.org/10.1055/s-2005-858235 [Cited on page 20.]

[68] H. Volzke, D. Alte, C. O. Schmidt, D. Radke, R. Lorbeer, N. Friedrich, N. Aumann, K. Lau, M. Piontek, G. Born, C. Havemann, T. Ittermann, S. Schipf, R. Haring, S. E. Baumeister, H. Wallaschofski, M. Nauck, S. Frick, A. Arnold, M. Junger, J. Mayerle, M. Kraft, M. M. Lerch, M. Dorr, T. Reffelmann, K. Empen, S. B. Felix, A. Obst, B. Koch, S. Glaser, R. Ewert, I. Fietze, T. Penzel, M. Doren, W. Rathmann, J. Haerting, M. Hannemann, J. Ropcke, U. Schminke, C. Jurgens, F. Tost, R. Rettig, J. A. Kors, S. Ungerer, K. Hegenscheid, J.-P. Kuhn, J. Kuhn, N. Hosten, R. Puls, J. Henke, O. Gloger, A. Teumer, G. Homuth, U. Volker, C. Schwahn, B. Holtfreter, I. Polzer, T. Kohlmann, H. J. Grabe, D. Rosskopf, H. K. Kroemer, T. Kocher, R. Biffar, U. John, and W. Hoffmann, “Cohort profile: The study of health in pomerania,” *International Journal of Epidemiology*, vol. 40, no. 2, p. 294–307, Feb. 2010. [Online]. Available: http://dx.doi.org/10.1093/ije/dyp394 [Cited on page 20.]

[69] D. Glass, A. Viñuela, M. N. Davies, A. Ramasamy, L. Parts, D. Knowles, A. A. Brown, K. Hedman, K. S. Small, A. Buil, E. Grundberg, A. C. Nica, P. Di Meglio, F. O. Nestle, M. Ryten, R. Durbin, M. I. McCarthy, P. Deloukas, E. T. Dermitzakis, M. E. Weale, V. Bataille, and T. D. Spector, “Gene expression changes with age in skin, adipose tissue, blood and brain,” *Genome Biology*, vol. 14, no. 7, Jul. 2013. [Online]. Available: http://dx.doi.org/10.1186/gb-2013-14-7-r75 [Cited on page 20.]

[70] C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O’Connell, A. Cortes, S. Welsh, A. Young, M. Effingham, G. McVean, S. Leslie, N. Allen, P. Donnelly, and J. Marchini, “The uk biobank resource with deep phenotyping and genomic data,” *Nature*, vol. 562, no. 7726, p. 203–209, Oct. 2018. [Online]. Available: http://dx.doi.org/10.1038/s41586-018-0579-z [Cited on pages 20 and 70.]

[71] R. Edgar, “Gene expression omnibus: Ncbi gene expression and hybridization array data repository,” *Nucleic Acids Research*, vol. 30, no. 1, p. 207–210, Jan.

2002. [Online]. Available: http://dx.doi.org/10.1093/nar/30.1.207 [Cited on pages 20 and 21.]

[72] A. Brazma, “Arrayexpress–a public repository for microarray gene expression data at the ebi,” *Nucleic Acids Research*, vol. 31, no. 1, p. 68–71, Jan. 2003. [Online]. Available: http://dx.doi.org/10.1093/nar/gkg091 [Cited on page 20.]

[73] R. Leinonen, H. Sugawara, and M. a. Shumway, “The sequence read archive,” *Nucleic Acids Research*, vol. 39, no. Database, p. D19–D21, Nov. 2010. [Online]. Available: http://dx.doi.org/10.1093/nar/gkq1019 [Cited on page 21.]

[74] R. Leinonen, R. Akhtar, E. Birney, L. Bower, A. Cerdeno-Tarraga, Y. Cheng, I. Cleland, N. Faruque, N. Goodgame, R. Gibson, G. Hoad, M. Jang, N. Pakseresht, S. Plaister, R. Radhakrishnan, K. Reddy, S. Sobhany, P. Ten Hoopen, R. Vaughan, V. Zalunin, and G. Cochrane, “The european nucleotide archive,” *Nucleic Acids Research*, vol. 39, no. Database, p. D28–D31, Oct. 2010. [Online]. Available: http://dx.doi.org/10.1093/nar/gkq967 [Cited on page 21.]

[75] J. N. and Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart, “The cancer genome atlas pan-cancer analysis project,” *Nature Genetics*, vol. 45, no. 10, p. 1113–1120, 2013. [Online]. Available: http://dx.doi.org/10.1038/ng.2764 [Cited on page 21.]

[76] R. L. Grossman, A. P. Heath, V. Ferretti, H. E. Varmus, D. R. Lowy, W. A. Kibbe, and L. M. Staudt, “Toward a shared vision for cancer genomic data,” *New England Journal of Medicine*, vol. 375, no. 12, p. 1109–1112, 2016. [Online]. Available: http://dx.doi.org/10.1056/NEJMp1607591 [Cited on page 21.]

[77] P. Branco, L. Torgo, and R. P. Ribeiro, “SMOGN: a pre-processing approach for imbalanced regression,” in *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, ser. Proceedings of Machine Learning Research, P. B. Luís Torgo and N. Moniz, Eds., vol. 74. PMLR, 22 Sep 2017, pp. 36–50. [Online]. Available: https://proceedings.mlr.press/v74/branco17a.html [Cited on pages 22 and 49.]

[78] GTEx Portal, “GTEx v11 Sample Metadata,” https://gtexportal.org/home/downloads/adult-gtex/metadata, accessed on September 2025. [Cited on page 26.]

[79] ——, “GTEx v11 Bulk Tissue Expression,” https://gtexportal.org/home/downloads/adult-gtex/bulk_tissue_expression, accessed on September 2025. [Cited on page 26.]

[80] A. Behdenna, M. Colange, J. Haziza, A. Gema, G. Appé, C.-A. Azencott, and A. Nordor, “pycombat, a python tool for batch effects correction in high-throughput molecular data using empirical bayes methods,” *BMC Bioinformatics*, vol. 24, no. 1, Dec. 2023. [Online]. Available: http://dx.doi.org/10.1186/s12859-023-05578-5 [Cited on page 29.]

[81] L. J. Carithers, K. Ardlie, M. Barcus, P. A. Branton, A. Britton, S. A. Buia, C. C. Compton, D. S. DeLuca, J. Peter-Demchok, E. T. Gelfand, P. Guan, G. E. Korzeniewski, N. C. Lockhart, C. A. Rabiner, A. K. Rao, K. L. Robinson, N. V. Roche, S. J. Sawyer, A. V. Segrè, C. E. Shive, A. M. Smith, L. H. Sobin, A. H. Undale, K. M. Valentino, J. Vaught, T. R. Young, and H. M. a. Moore, “A novel approach to high-quality postmortem tissue procurement: The gtex project,” *Biopreservation and Biobanking*, vol. 13, no. 5, p. 311–319, Oct. 2015. [Online]. Available: http://dx.doi.org/10.1089/bio.2015.0032 [Cited on page 31.]

[82] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, “Scikit-learn: Machine learning in python,” *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011. [Cited on pages 31 and 42.]

[83] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, “limma powers differential expression analyses for rna-sequencing and microarray studies,” *Nucleic Acids Research*, vol. 43, no. 7, p. e47–e47, Jan. 2015. [Online]. Available: http://dx.doi.org/10.1093/nar/gkv007 [Cited on pages 32 and 33.]

[84] M. D. Robinson and A. Oshlack, “A scaling normalization method for differential expression analysis of rna-seq data,” *Genome Biology*, vol. 11, no. 3, Mar. 2010. [Online]. Available: http://dx.doi.org/10.1186/gb-2010-11-3-r25 [Cited on page 34.]

[85] C. W. Law, Y. Chen, W. Shi, and G. K. Smyth, “voom: precision weights unlock linear model analysis tools for rna-seq read counts,” *Genome Biology*, vol. 15, no. 2, Feb. 2014. [Online]. Available: http://dx.doi.org/10.1186/gb-2014-15-2-r29 [Cited on page 34.]

[86] Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate: A practical and powerful approach to multiple testing,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 57, no. 1, p. 289–300, Jan. 1995. [Online]. Available: http://dx.doi.org/10.1111/j.2517-6161.1995.tb02031.x [Cited on page 36.]

[87] M. E. Tipping, “Sparse bayesian learning and the relevance vector machine,” *Journal of machine learning research*, vol. 1, no. Jun, pp. 211–244, 2001. [Cited on page 39.]

[88] E. Strubell, A. Ganesh, and A. McCallum, “Energy and policy considerations for deep learning in nlp,” in *Proceedings of the 57th annual meeting of the association for computational linguistics*, 2019, pp. 3645–3650. [Cited on page 42.]

[89] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf [Cited on page 55.]

[90] D. Szklarczyk, R. Kirsch, M. Koutrouli, K. Nastou, F. Mehryary, R. Hachilif, A. L. Gable, T. Fang, N. T. Doncheva, S. Pyysalo, P. Bork, L. J. Jensen, and C. von Mering, “The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest,” *Nucleic Acids Research*, vol. 51, no. D1, p. D638–D646, Nov. 2022. [Online]. Available: http://dx.doi.org/10.1093/nar/gkac1000 [Cited on page 63.]

[91] S. Paik, S. Shak, G. Tang, C. Kim, J. Baker, M. Cronin, F. L. Baehner, M. G. Walker, D. Watson, T. Park, W. Hiller, E. R. Fisher, D. L. Wickerham, J. Bryant, and N. Wolmark, “A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer,” *New England Journal of Medicine*, vol. 351, no. 27, p. 2817–2826, Dec. 2004. [Online]. Available: http://dx.doi.org/10.1056/NEJMoa041588 [Cited on page 69.]

[92] L. J. van ’t Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, “Gene expression profiling predicts clinical outcome of breast cancer,” *Nature*, vol. 415, no.

6871, p. 530–536, Jan. 2002. [Online]. Available: http://dx.doi.org/10.1038/415530a [Cited on page 69.]

[93] G. K. Geiss, R. E. Bumgarner, B. Birditt, T. Dahl, N. Dowidar, D. L. Dunaway, H. P. Fell, S. Ferree, R. D. George, T. Grogan, J. J. James, M. Maysuria, J. D. Mitton, P. Oliveri, J. L. Osborn, T. Peng, A. L. Ratcliffe, P. J. Webster, E. H. Davidson, L. Hood, and K. Dimitrov, "Direct multiplexed measurement of gene expression with color-coded probe pairs," *Nature Biotechnology*, vol. 26, no. 3, p. 317–325, Feb. 2008. [Online]. Available: http://dx.doi.org/10.1038/nbt1385 [Cited on page 70.]

A. Appendix

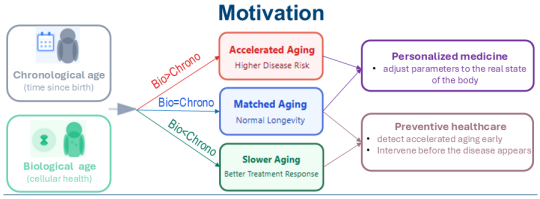
The following poster was presented at the Department of Computer Science Master's Day, Faculty of Sciences of the University of Porto, 2025/26, summarising the motivation, methods, and preliminary results of the work presented in this thesis. Note that the poster reflects preliminary results from an earlier stage of the project; the final model, results, and conclusions are presented in the main body of this thesis.

2025/26

Aging Prediction Models From Blood Gene Expression

Marta Duarte[1] | Pedro Ferreira[1] and Marta Moreno
[1]Faculdade de Ciências da Universidade do Porto

Motivation



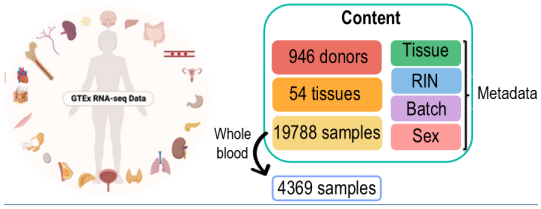
Scientific challenges

Problem
- Noisy signal
- Batch effects
- Age imbalance

Solution
- Preprocessing
- Feature engineering

Output
- Transcriptomic ageing clock

Data

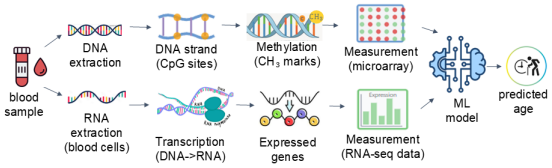


Objective(s)

1. Built robust and accurate transcriptomic age prediction models
2. Handle imbalance and incomplete age distributions
3. Ensure cross-dataset model generalization
4. Identify a minimal predictive gene signature

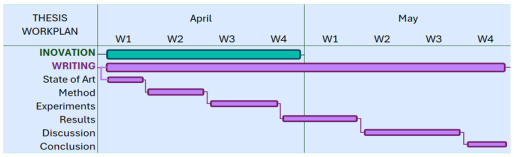
State-of-the-Art

DNA methylation clock

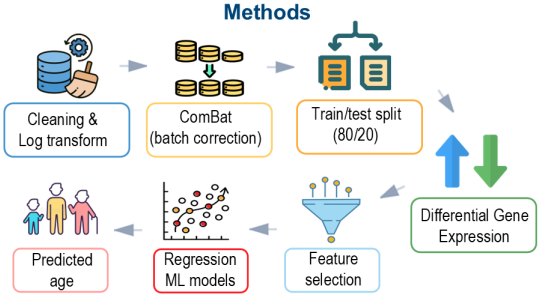


Transcriptomic clock

Workplan



Methods



Preliminary results

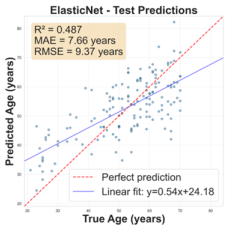


Figure 1 - Each point represents one donor. The slope below 1 (0.54) reflects regression-to-the-mean, with younger ages overestimated and older ages underestimated.

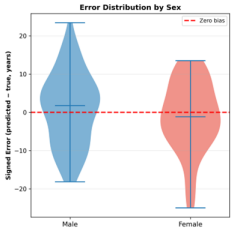
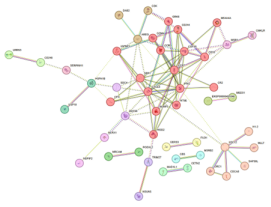


Figure 2 - Distribution of signed prediction errors (predicted − true age) by sex on the test set. Males show a positive median error (ages overestimated) while females show a negative median error (ages underestimated)



105 nodes | 15 clusters
94 edges | 1.79 av. deg.

Cluster 1: 17 nodes, enriched for regulation of interleukin-1 beta production, consistent with inflammaging

Figure 3 - Protein-protein interaction network of selected genes

Conclusions

By combining machine learning with blood gene expression, this work opens a path towards personalized health monitoring: identifying individuals whose tissues age faster than expected may enable early intervention and lifestyle adaptation to slow biological aging.

References

[1] López-Otín, Carlos, et al. "The hallmarks of aging." *Cell* 153.6 (2013): 1194-1217.
[2] Levine, Morgan E., et al. "An epigenetic biomarker of aging for lifespan and healthspan." *Aging (albany NY)* 10.4 (2018): 573.
[3] Moraes, Athos, et al. "Predicting age from human lung tissue through multi-modal data integration." *International Conference on Discovery Science*. Cham: Springer Nature Switzerland, 2023.
[4] Smyth, Gordon K. "Linear models and empirical bayes methods for assessing differential expression in microarray experiments." *Statistical applications in genetics and molecular biology* vol. 3 (2004): Article3. doi:10.2202/1544-6115.1027
[5] Peters, Marjolein J., et al. "The transcriptional landscape of age in human peripheral blood." *Nature communications* 6.1 (2015): 8570.

U.PORTO FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO | [dcc] DEPARTAMENTO DE CIÊNCIA DE COMPUTADORES FACULDADE DE CIÊNCIAS DA UNIVERSIDADE DO PORTO | M:DS

Figure A.1: Poster presented at the Department of Computer Science Master's Day.