Improving Single-Cell Transcriptomic Aging Clocks

Enhancing Accuracy and Biological Interpretability

Vlad Alexan
Supervisors: Dr. Marcel Reinders, Bram Pronk, Inez den Hond, Gerard Bouland

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student:	Vlad Alexan
Final project course:	CSE3000 Research Project
Thesis committee:	Dr. Marcel Reinders, Bram Pronk, Inez den Hond, Gerard Bouland, Dr. Kaitai Liang

Abstract

Biological aging clocks estimate age from molecular data and provide insights into age-related functional decline. While aging clocks based on bulk transcriptomic data are well-studied, their single-cell counterparts remain limited and underexplored. In this study, we replicate and enhance a recent single-cell RNA-seq aging clock for human immune cells using ElasticNet, improving its performance through refined preprocessing, feature selection, and regularization. We also explore LightGBM to assess nonlinear modeling potential. Our enhanced models reduce prediction error, generalize better across external datasets, and identify biologically relevant genes through SHAP analysis. These findings support the development of accurate, interpretable, cell-type-specific aging clocks using single-cell data.

1 Introduction

Aging is a complex biological process characterized by progressive functional decline at the cellular and molecular level, increasing vulnerability to disease and mortality. One key goal of aging research is to quantify this process using estimates of "biological age", which may better reflect functional status than chronological age. Aging clocks, which are machine learning models trained on molecular data, are a promising approach for estimating biological age. They can be trained on diverse data types, such as DNA methylation levels (epigenetic clocks [1, 2]) or gene expression profiles (transcriptomic clocks[3], which are more recent and the focus of our research), to capture systemic and stochastic aspects of aging.

Historically, most transcriptomic aging clocks have relied on bulk RNA sequencing data [4, 5], which measures the average gene expression across large, mixed populations of cells. This averaging masks cell-type-specific aging signatures (distinct patterns of gene activity that differ between cell types) and limits the interpretability of model predictions. To address this limitation, single-cell RNA sequencing (scRNA-seq) enables the development of clocks that operate at single-cell resolution. A notable study by Zakar-Polyák et al. [6] demonstrated that ElasticNet[1]-based clocks trained on over one million peripheral blood mononuclear cells (PBMCs) from 508 human donors can predict donor age at the cell-type level and capture meaningful biological variation.

While promising, these models were built using linear techniques (ElasticNet regression), which may fail to capture non-linear dependencies and interactions present in high-dimensional gene expression data. Additionally, a major limitation of the reference study is the lack of in-depth feature importance or interpretability analysis, leaving uncertainty about which specific genes drive the age predictions. Moreover, their generalizability across cell types and tissues remains an open question.

This thesis addresses the following core question: *Can we improve on current models that predict biological age using single-cell gene expression data, and can we determine which specific genes are most important for making accurate predictions?* Building on the framework established by Zakar-Polyák et al. [6], we replicate and validate their baseline ElasticNet model using the approach outlined in their work, despite the lack of publicly available code. We then enhance this model through improved data preprocessing, feature selection, regularization, and hyperparameter tuning. Our work extends prior research by addressing a key limitation of the original study, namely the lack of feature importance analysis, through the use of SHAP[2]-based interpretability methods. To assess model generalizability, we eval-

[1] `https://en.wikipedia.org/wiki/Elastic_net_regularization`
[2] `https://shap.readthedocs.io/en/latest/`

uate both the replicated and enhanced linear models on four external scRNA-seq datasets, analyzing differences in predictive accuracy across cell types. Additionally, we investigate non-linear modeling approaches, specifically LightGBM[3], to examine whether they outperform ElasticNet in predicting biological age. Finally, we apply interpretability techniques, including SHAP values, and enrichment analysis, to identify genes that are most consistently associated with transcriptomic aging and assess their potential biological relevance.

2 Methodology

In order to address the research question of this thesis, we build on the aging clock framework proposed by Zakar-Polyák et al. [6], who trained ElasticNet regression models on scRNA-seq data from the AIDA dataset to predict donor age from PBMCs. Their approach enabled cell-type-specific age prediction via standardized gene expression and 5-fold cross-validation.

To assess and improve upon this work, we pursued three complementary directions. First, we replicated their ElasticNet models per cell type using donor age as target, while following the steps described the reference paper [6], since no code was provided for training the models. Second, we enhanced the replicated models by refining variance-based gene selection, enforcing donor-wise splits, and tuning ElasticNet hyperparameters (α and l_1-ratio) via nested cross-validation. We will refer to these models as "enhanced linear models" for the rest of the paper. Finally, we explored gradient boosted decision trees models (specifically LightGBM) to evaluate non-linear modeling potential. These models were optimized using nested cross-validation and hyperparameter tuning via Optuna, with the goal of assessing whether non-linear models could capture additional predictive signal beyond linear ElasticNet models.

Our methodology is motivated by prior work showing that ElasticNet provides sparse, interpretable models well-suited for biological insight [1], while LightGBM, a gradient boosted decision tree (GBDT) framework, can model complex nonlinear relationships. The reason why we decided to use GBDTs for nonlinear models is because they are known for strong performance on high-dimensional, sparse (which is the case for gene expression, single cell data), structured biological datasets and often outperform deep neural networks in tabular settings while remaining more interpretable [7]. We selected LightGBM over other GBDT implementations (such as XGBoost) for its computational efficiency, scalability, and built-in regularization [8], making it a practical choice for a limited resource, large-scale single-cell RNA-seq analysis.

2.1 Datasets

Single-cell RNA sequencing data (which captures the gene expression profile of individual cells) was used in this research, in the form of high-dimensional and sparse gene-by-cell matrices. Each dataset can be formalized as a matrix $X \in \mathbb{R}^{n \times p}$, where n is the number of cells and p the number of genes. Each row vector x_i contains the expression levels of p genes in a single cell i, and is associated with metadata such as cell type, donor ID, and donor age.

The primary training dataset was the AIDA dataset, consisting of approximately 1 million transcriptomes from PBMCs, collected from 508 healthy donors aged 19–75. Following

[3] `https://en.wikipedia.org/wiki/LightGBM`

the approach of Zakar-Polyák et al., we trained separate predictive models per cell type, as different immune cells may exhibit distinct aging signatures.

To evaluate generalizability, we tested our models on four independent external scRNA-seq datasets: Yoshida, which captures single-cell multi-omics data from pediatric and adult individuals; Liu, which includes PBMCs from individuals with continuous age labels across a wide age range; eQTL, a PBMC dataset with continuous donor ages containing the largest amount of samples ≈ 1.2 million; and Stephenson, which provides immune profiles of both healthy and COVID-19 patients at different stages, from which we retained only the healthy donors for analysis.

2.2 Preprocessing

For both model types, expression matrices were normalized using total-count scaling (10,000 counts per cell), followed by log-transformation to stabilize variance and reduce skewness due to highly expressed genes. To ensure comparability across genes and datasets, expression values were standardized per gene, using z-score normalization based on the training cells in each cross-validation fold. Finally, to focus the models on informative features, we selected the top 5,000 most variable genes per cell type for the linear models. This variance was computed across all cells of a given cell type, prior to cross-validation. Moreover, many of the $\sim$36,000 measured genes were not detectably expressed in PBMCs, and including them would introduce additional sparsity and noise. For the non-linear models, we selected a reduced set of the top 3,000 most variable genes. This smaller feature set was chosen based on empirical performance. Tree-based models are often less sensitive to small feature effects and can suffer from overfitting or inefficiency when trained on very high-dimensional data [9]. Additionally, in preliminary experiments on the non-linear models (when training models using default parameters) we noticed that the smaller input size substantially reduced memory usage, which was a crucial consideration given our limited computational resources.

2.3 Models

For each cell type, we trained both linear and non-linear models to predict donor age. Linear models consisted of separate ElasticNet regressors per cell type, trained using standardized gene expression as input and donor age as target. We employed nested cross-validation with donor-wise splitting: an outer 5-fold cross-validation loop to estimate generalization performance, and an inner 5-fold cross-validation using ElasticNetCV to tune α and l_1-ratio hyperparameters.

Subsequently, we trained LightGBM models for each cell type to explore non-linear modeling capacity. LightGBM models were optimized using nested cross-validation and hyperparameter tuning with Optuna[4], employing a similar outer 5-fold split. For each outer fold, hyperparameters such as learning rate, number of leaves, and regularization terms were tuned on an inner 3-fold cross-validation loop, compared to the 5-fold used by ElasticNet. This choice reflects a trade-off between computational efficiency and robustness, since tree-based models with such large hyperparameter spaces and high sample training dataset would reach considerably high training times, as observed empirically. Finally, models were then trained on the full training fold using the selected hyperparameters and evaluated on the held-out outer fold, similarly to the linear models. Additionally, to support

[4] `https://optuna.readthedocs.io/en/stable/`

reproducibility, all random seeds were fixed, donor splits were performed deterministically, and model configurations were version-controlled.

2.4 Evaluation

Model performance was evaluated using mean absolute error (MAE) for datasets with continuous age labels, Pearson correlation (r) between true and predicted age, and Spearman correlation (ρ) for datasets with categorical age labels (Yoshida and Stephenson). External datasets were mapped to AIDA's gene and cell-type space, and missing genes were imputed using the average log-normalized expression of that gene across all cells of the corresponding cell type from the AIDA training set.

2.5 Feature Importance

We computed feature importance for both linear and non-linear models to interpret the learned aging signatures. For this, we applied SHAP analysis to compute per-fold mean absolute SHAP values, which were visualized using heatmaps (gene-cell type matrices, including top genes overall) and frequency counts (genes appearing in top-20 lists across cell types), facilitating biological interpretation of the learned models.

3 Responsible Research

3.1 Ethical Considerations

This research utilizes publicly available single-cell transcriptomic datasets, such as the AIDA dataset and external validation sets from previous studies, all of which were collected and anonymized by their original authors. No personally identifiable information is included, and all datasets are used strictly for academic purposes.

The focus of our study is the development of machine learning models to estimate biological age at the single-cell level. These models are not intended for clinical or diagnostic use without rigorous validation. We are aware that the interpretation of "biological age" can carry implications for individual health or longevity assessments, and we emphasize that our findings should not be misused to make deterministic claims about health status at the cellular or organismal level.

3.2 Reproducibility

To promote reproducibility, we designed our modeling pipeline to follow transparent and standard practices. The datasets used are publicly accessible, and all preprocessing steps (including normalization, filtering, and gene selection) are described in detail. We perform 5-fold cross-validation with donor-wise splitting to prevent data leakage, and we report multiple evaluation metrics (MAE, Pearson, and Spearman correlation) to ensure robust assessment.

We replicated baseline models from Zakar-Polyák et al. [6] despite the absence of publicly available code, by interpreting their described methodology. Enhancements to these models, such as improved regularization and feature selection, as well as the non-linear models we trained, are reproducible through the detailed description we provided. Upon successful completion of the research, we intend to publish our code on a public repository to support

further research and verification. The code will be made available at: `https://github.com/Vlad13503/Aging-Clock`.

3.3 Use of Generative AI

Generative AI tools were used in this project mainly to help reformulate and refine phrases and ideas, clarify technical descriptions, optimize plot annotations, and, in some cases, interpret or provide insight into the biological meaning of some results. These tools assisted in improving the clarity and consistency of the presentation but did not generate scientific content (code pipelines), results or influence the design or decisions taken throughout the project. All scientific contributions are the result of our original work and all outputs from generative tools were critically reviewed and adjusted. Example prompts that were used are: "How could we group the attached images into a panel in Overleaf?" or "How could we rephrase the following paragraph to use more scientific terms, without changing its meaning: [...]?".

4 Results

We systematically evaluated two modeling approaches (enhanced linear models and non-linear LightGBM models) for predicting biological age from single-cell transcriptomic data, with the intention to see whether our enhancements can improve both accuracy and generalizability. All models were trained on the AIDA dataset and assessed on four independent external datasets. We first present an overview of the datasets used and their structure, followed by an analysis of model performance and interpretability.

4.1 Data Overview

We analyzed five transcriptomic datasets: AIDA (used to train the models) and four external datasets (Yoshida, Liu, eQTL, and Stephenson) used for model validation. Figure 6 summarizes key characteristics of these datasets, including the number of cells, cell types, donors, and age ranges.

The AIDA dataset contains over 1 million immune cells from 508 donors across 33 cell types, covering an age range from 19 to 75 years. The external datasets vary considerably in size and complexity. The eQTL dataset includes nearly 1.25 million cells and 981 donors, making it the largest in terms of both number of samples and donor diversity. Yoshida and Stephenson offer broader age ranges (up to 92 years), but noticeably fewer donors and more heterogeneous age labeling, since they contain both continuous and categorical age groups. Liu is smaller in scope, with only 14 cell types and 46 donors, ranking last on both accounts. Together, these datasets span diverse populations and conditions, supporting robust generalization analysis of our aging models.

4.2 Performance of Enhanced Linear Models

We begin by comparing our enhanced linear models to the published results reported by Zakar-Polyák et al. [6]. Since the authors did not publish their code used for training the linear models, we could not directly apply their exact models. Instead, we used their metrics and evaluation results for reference and compared them to the performance of our enhanced models.

Figure 7 shows a scatter plot of mean absolute error (MAE) per cell type, comparing our enhanced linear models to the reported performance from the original paper. Most points (25 out of the 33) lie above the diagonal, indicating a lower error for our models. Similarly, Figure 8 compares the Pearson correlation between true and predicted donor ages for each cell type, again showing higher correlation for our models in many cases (21 out of the 33). Together, these comparisons confirm that our enhanced models match or exceed the reported performance in the original study.

We next evaluate the performance of our enhanced linear models trained on the AIDA dataset by comparing them to a reimplementation of the ElasticNet aging clocks proposed by Zakar-Polyák et al. [6]. These reimplementation models, referred to as "replication models" (shown in red in Figure 1), were trained according to the process described in the original publication (so using the same gene selection and preprocessing steps), without our enhancements.

Figure 1 summarizes the results across three datasets. On the AIDA dataset (panel **a**), the enhanced models consistently outperformed replication models across almost all cell types, with median MAE reductions ranging between 0.5 and 1 year and relative improvements reaching up to 10% in some subsets. This indicates that careful enhancements to preprocessing and even light model tuning can visibly improve the accuracy of transcriptomic aging clocks.

To assess the generalizability of our models, the same enhanced models that were trained on the AIDA dataset were applied (without retraining) to four external scRNA-seq datasets, two of which are displayed in figure 1: the eQTL (panel **b**) and Liu (panel **c**) datasets. For each cell type, the model trained on AIDA was applied to matching cell types in the eQTL and Liu datasets. Expression profiles from the external datasets were normalized and log-transformed in the same way as for AIDA, and missing genes were imputed using average expression values from the AIDA training data, following the pipeline described in the reference paper [6] to facilitate comparison of results. The enhanced models yielded substantial MAE improvements over replication models on eQTL (average reduction of 2 years) and more variable but still favorable results on Liu (average reduction of 2–3 years). These findings demonstrate that relatively simple enhancements to model training and preprocessing not only improve performance on the training dataset, but also lead to even greater accuracy and generalizability on external datasets.

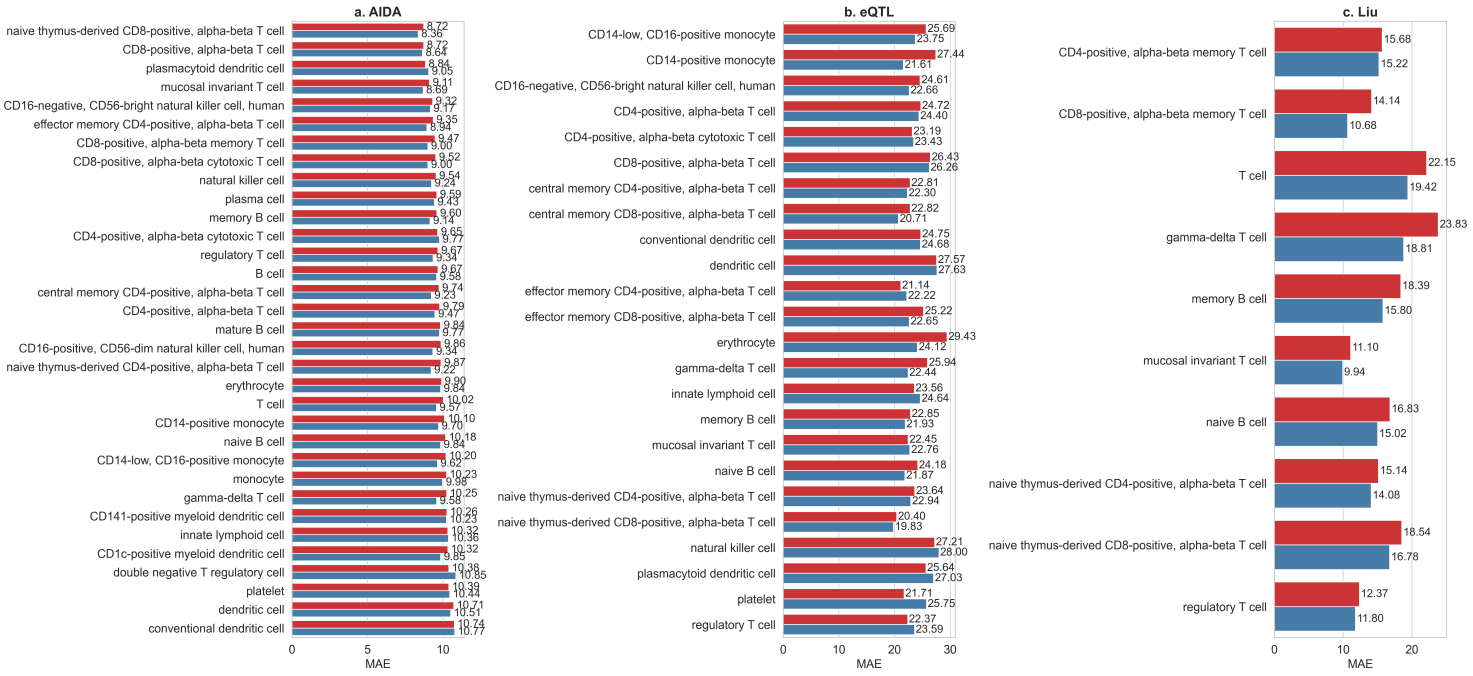


Figure 1: Mean absolute error (MAE) comparison between replication models (red) and enhanced linear models (blue) across three datasets. **(a)** MAE on AIDA (training set) using nested cross-validation. **(b)** External evaluation on the eQTL dataset. **(c)** External evaluation on the Liu dataset. Each pair of blue and red bars represents a distinct immune cell type with lower values indicating better performance.

To interpret the biological relevance of our enhanced models, we analyzed feature importance using SHAP values. Figure 2 displays a heatmap of the top 30 genes based on their SHAP importance across cell types. We chose the top 30 genes to provide sufficient granularity for identifying both shared and cell-type-specific aging markers while keeping the heatmap readable. The figure reveals that some genes show consistently high importance across many immune cell types, while others are more cell-type-specific. Additionally, the *all cells* column represents the mean SHAP importance of each gene averaged across all individual cell types, providing a global summary of feature relevance. This suggests that transcriptomic aging signatures comprise both shared and cell-type-specific components.

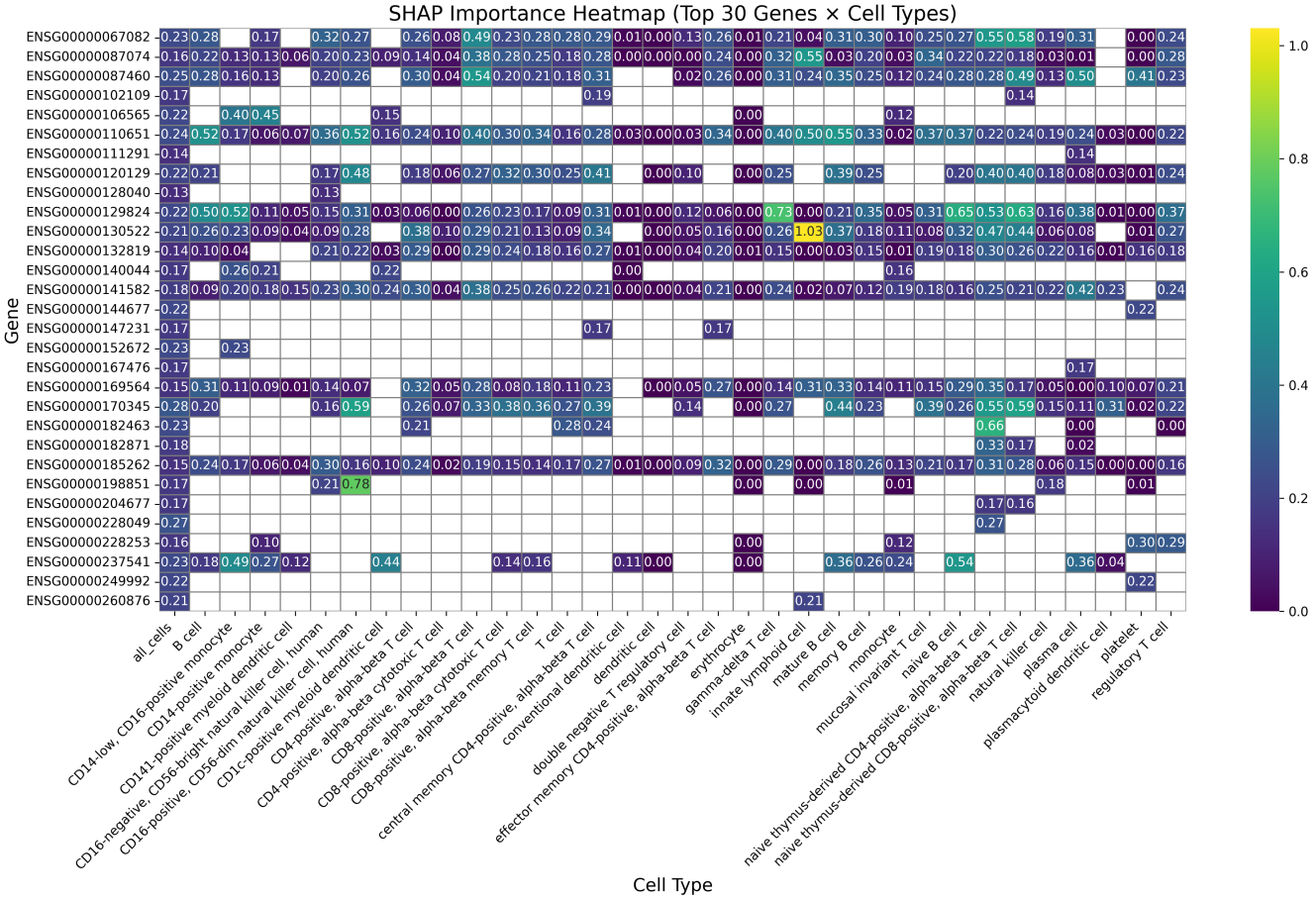


Figure 2: SHAP importance heatmap for the top 30 genes (ranked by global mean SHAP value across all cell types). Each cell shows the mean absolute SHAP value of a gene for a specific cell type, with brighter colors indicating higher importance. Empty (white) cells indicate that the gene was not selected during model training for that specific cell type, and thus no SHAP value is available. The *all cells* column represents the mean SHAP value of each gene across all cell types, serving as a global summary of gene relevance.

Finally, Figure 9 quantifies the prevalence of SHAP-important genes across cell types. For each cell type, we identified the top 20 genes based on the highest mean absolute SHAP values. We selected the top 20 genes per cell type to balance interpretability and focus, since we noticed (after also trying other amounts) that 20 was the choice that captured a relevant amount of both high and low frequency genes. We then ranked genes by how often they appeared in these cell-type-specific top-20 lists. Genes with high appearance frequency (such as ENSG00000110651[5]–**CD81**[6], ENSG00000067082–**KLF6**, ENSG00000087460–**GNAS**, and ENSG00000141582–**CBX4**) are consistently important across diverse immune contexts (they were present in the top-20 lists for more than 20 of the 33 cell types), suggesting they may be robust, cell-type-agnostic biomarkers of aging. In contrast, genes with low frequency (such as ENSG00000173020–**GRK2**, ENSG00000142327–**RNPEPL1**, and ENSG00000213145–**CRIP1**) may reflect specialized aging processes within specific immune compartments, since they were present in less than 7 out of the 33 cell types.

To further explore the biological relevance of the top-ranked genes, we performed functional enrichment analysis using Enrichr[7] on the top 30 SHAP genes (namely those shown in Figure 2), using the top 5,000 genes based on variance selected during model training as background (Figure 5 a,b). The KEGG 2021 Human analysis (Figure 5 a) showed that the top SHAP-selected genes were strongly associated with key immune and signaling pathways, including those involved in T cell differentiation (Th1/Th2 and Th17), IL-17 signaling, and serotonin-related processes. Some pathways also pointed to links with immune-related diseases, such as Chagas disease and T-cell leukemia. Similarly, the GO Biological Process 2025 results (Figure 5 b) highlighted processes like T cell activation, immune cell growth, and antibody production, along with broader functions like gene regulation and skeletal development. Together, these findings indicate that the genes identified by the model are biologically meaningful and reflect immune system changes commonly seen with aging.

4.3 Performance of Nonlinear Models

To complement our analysis of linear models, we continue with the LightGBM nonlinear models we trained and evaluated on the same data splits and cell types as the enhanced linear models, using the same plots (allowing for direct comparison), for which figure 3 compares the predictive performance of the enhanced linear models (green) and nonlinear models (orange) across three datasets. On AIDA (panel **a**), the nonlinear models manage to achieve a slightly lower MAE values on most cell types (improved with 0.3–0.5 years on average), suggesting potential gains from modeling interactions and nonlinear effects within the training domain.

However, when applied to the external datasets eQTL (panel **b**) and Liu (panel **c**), the nonlinear models exhibited higher MAE in most cases, which outweighed the few cell types in which we noticed some improvements (such as regulatory T cell, platelet, innate lymphoid cell etc. in the eQTL dataset). This reduction in generalization performance may be due to overfitting to AIDA-specific distributions, or more likely due to domain shift effects that nonlinear models are more sensitive to, since the four external datasets were significantly different in structure and distribution from AIDA. These findings suggest that while nonlinear models offer flexible modeling capacity, their utility in cross-dataset

[5]ENSG identifiers refer to Ensembl gene IDs, which are unique identifiers assigned to genomic features (in this case to genes).

[6]Gene symbols in bold represent their common gene names (standardized abbreviations).

[7]`https://maayanlab.cloud/Enrichr/`

generalization remains limited without further regularization, tuning or domain adaptation strategies.

For completeness, we also include in the Appendix some additional plots we obtained (Figure 11 and Figure 12) summarizing cell-type-specific correlation metrics (between predicted and actual age) for both linear and nonlinear models across all four external datasets. Compared to their nonlinear counterparts, the enhanced linear models demonstrated more stable and broadly positive correlation values across external datasets (Figure 11). While the nonlinear models (Figure 12) occasionally achieved higher peak correlations in specific cell types (such as monocytes in the Stephenson dataset), their performance was more variable in less abundant or noisier data populations. These observations reinforce the trade-off between flexibility and generalizability in aging clock model design.

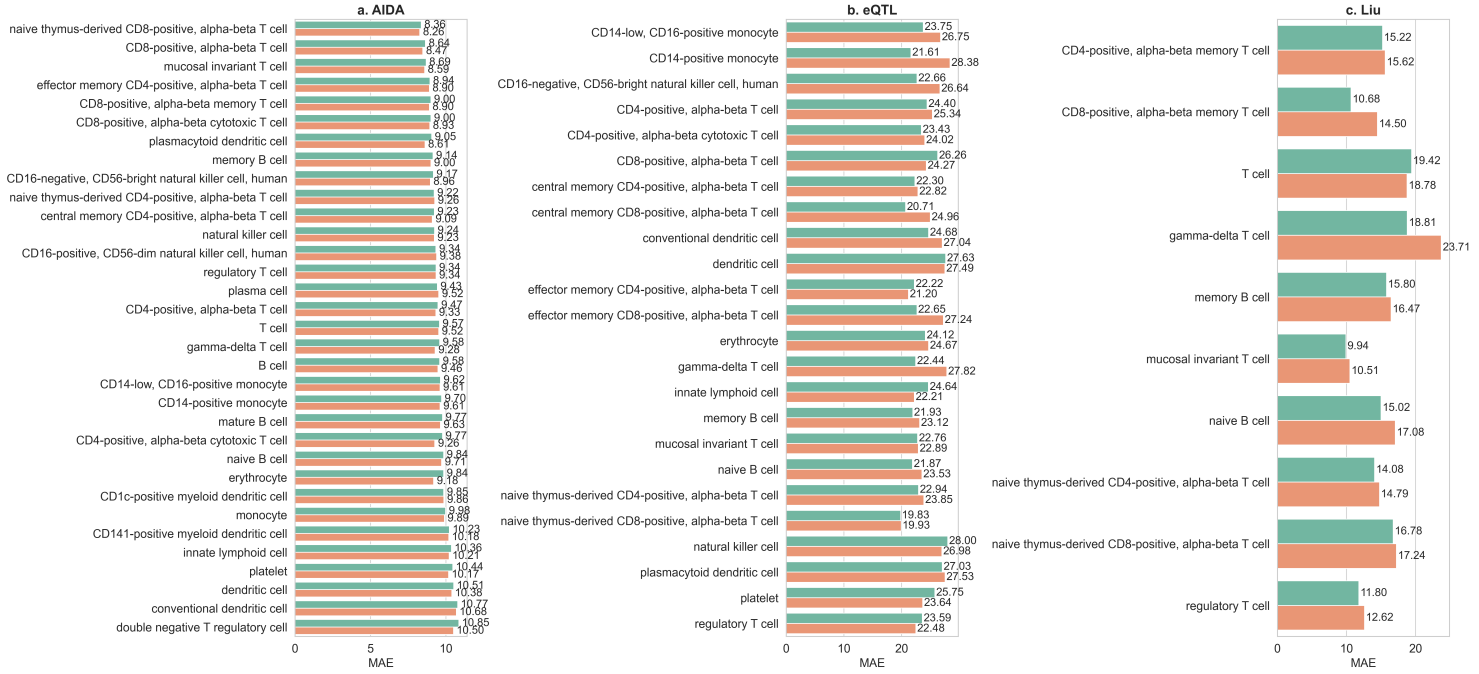


Figure 3: Mean absolute error (MAE) comparison between enhanced linear models (green) and nonlinear LightGBM models (orange) across three datasets. **(a)** MAE on AIDA (training set) using nested cross-validation. **(b)** External evaluation on the eQTL dataset. **(c)** External evaluation on the Liu dataset. Each pair of green and orange bars represents a distinct immune cell type with lower values indicating better performance.

We next examined feature importance in the nonlinear models using SHAP, which is well-suited for interpreting tree-based models [10]. Figure 4 presents a SHAP heatmap of the top 30 genes across all cell types. The top genes were selected based on their global mean SHAP values across cell types.

In this plot, brighter colors represent higher SHAP importance, while empty (white) cells indicate that the gene was not selected by the model for that specific cell type. The *all cells* column shows the average SHAP value across all individual cell types, serving as a global summary. Compared to the linear model heatmap, the nonlinear version highlights more

extreme differences in gene importance across cell types (with noticeably more green and yellow cells, so importance scores between 1 and 2), consistent with the higher flexibility and sparsity of LightGBM models.

Notably, several genes were consistently important across both model types. In the nonlinear models, top-ranked genes included ENSG00000130522 (**JUND**), ENSG00000170345 (**FOS**), and ENSG00000067082 (**KLF6**), which also appeared prominently in the linear models. This overlap highlights a robust set of aging-associated genes that are recoverable across different model types.

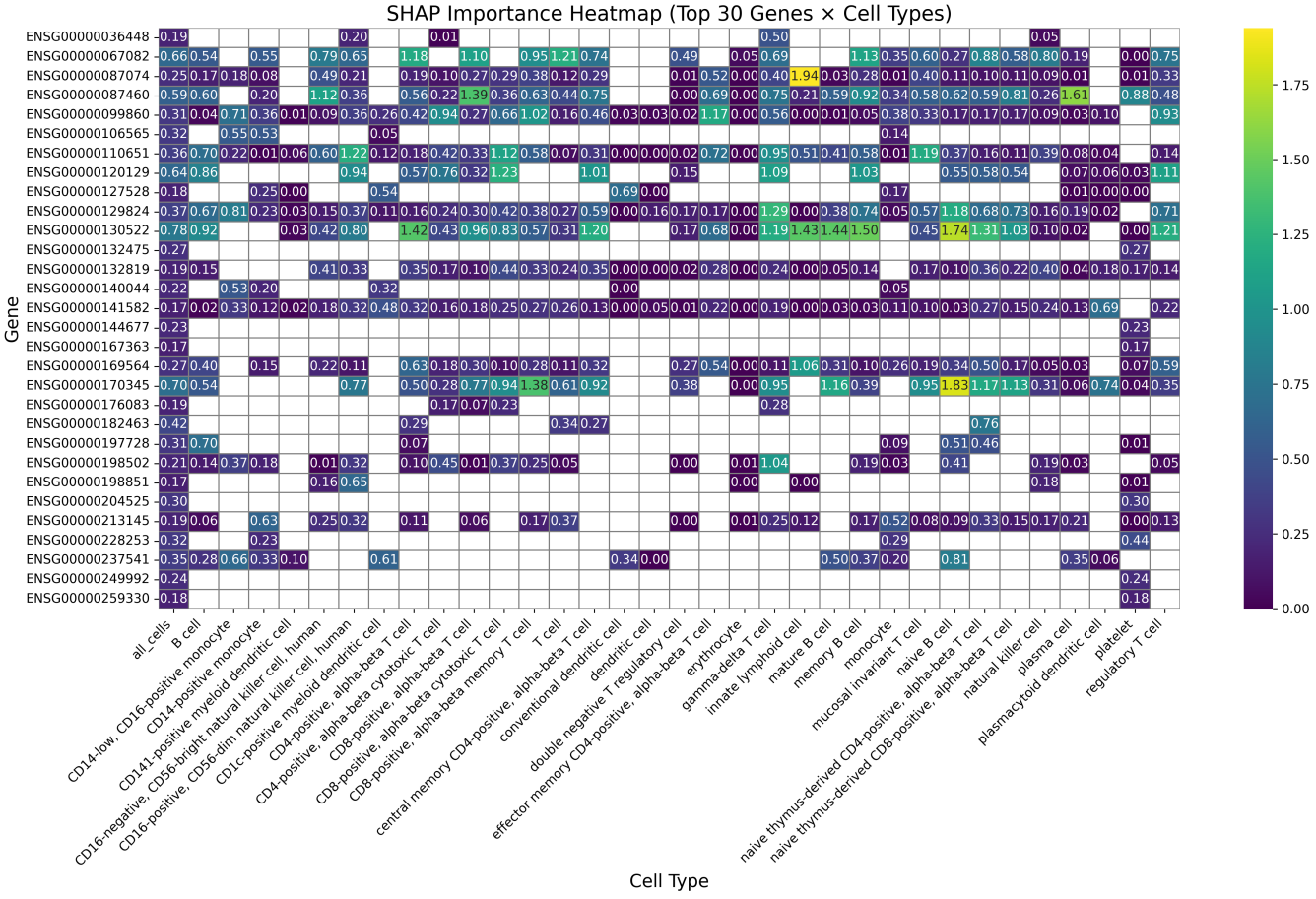


Figure 4: SHAP importance heatmap for the top 30 genes (ranked by global mean SHAP value across all cell types) in the nonlinear LightGBM models. Each cell shows the mean absolute SHAP value of a gene for a specific cell type, with brighter colors indicating higher importance. Empty (white) cells indicate that the gene was not selected during model training for that specific cell type, and thus no SHAP value is available. The *all cells* column represents the mean SHAP value of each gene across all cell types, serving as a global summary of gene relevance.

Finally, figure 10 presents the frequency of SHAP-important genes across cell types. Similar to the linear models, for each cell type we identified the top 20 genes with the highest

mean SHAP values. We then ranked genes by how many times they appeared in these per-cell-type top-20 lists. Many of the top-ranked genes—such as ENSG00000087460 (**GNAS**), ENSG00000067082 (**KLF6**), ENSG00000170345 (**FOS**) and ENSG00000129824 (**RPS4Y1**) again overlapped with the linear model's top features, underscoring their robustness and consistency of top contributing features across models.

We applied the same enrichment strategy (as we did for the linear models) to the top 30 SHAP genes from the nonlinear models, using the top 3,000 genes based on variance selected during model training as background (Figure 5 c,d). Notably, the enriched KEGG pathways included strong immune signatures such as allograft rejection, autoimmune thyroid disease, T-cell leukemia, and Th1/Th2 and Th17 differentiation, the last two overlapping with the results for the linear models. The GO results again highlighted transcriptional regulation and immune activation terms, with substantial thematic overlap with the linear model's enrichments. This convergence across modeling approaches suggests that shared functional programs underlie the most robust transcriptomic aging markers.

KEGG 2021 Human | Bar Graph | Table | Clustergram | Appyter

Hover each row to see the overlapping genes.

10 entries per page | Search:

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	Parathyroid hormone synthesis, secretion and action	0.0003108	0.02372	28.40	229.39
2	Chagas disease	0.0004313	0.02372	25.22	195.43
3	Th1 and Th2 cell differentiation	0.0008748	0.03208	19.52	137.42
4	Th17 cell differentiation	0.001597	0.04391	15.68	100.95
5	Serotonergic synapse	0.003008	0.05478	29.40	170.70
6	Circadian entrainment	0.003243	0.05478	28.22	161.72
7	Amphetamine addiction	0.003486	0.05478	27.13	153.51
8	IL-17 signaling pathway	0.005116	0.05833	22.01	116.13
9	Human T-cell leukemia virus 1 infection	0.005397	0.05833	10.00	52.19
10	MAPK signaling pathway	0.005397	0.05833	10.00	52.19

(a) KEGG 2021 Human enrichment results (linear models)

GO Biological Process 2025 | Bar Graph | Table | Clustergram | Appyter

Hover each row to see the overlapping genes.

10 entries per page | Search:

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	Regulation of Transcription by RNA Polymerase II (GO:0006357)	0.0004182	0.08210	6.67	51.90
2	Positive Regulation of T Cell Activation (GO:0050870)	0.0009175	0.08210	19.18	134.15
3	Skeletal System Development (GO:0001501)	0.002154	0.08210	35.31	216.79
4	Positive Regulation of Transcription by RNA Polymerase II (GO:0045944)	0.004909	0.08210	5.71	30.35
5	Positive Regulation of Lymphocyte Proliferation (GO:0050671)	0.005416	0.08210	21.34	111.37
6	Activation of Adenylate Cyclase Activity (GO:0007190)	0.006417	0.08210	330.80	1670.13
7	Adenylate Cyclase-Activating Adrenergic Receptor Signaling Pathway (GO:0071880)	0.006417	0.08210	330.80	1670.13
8	Humoral Immune Response Mediated by Circulating Immunoglobulin (GO:0002455)	0.006417	0.08210	330.80	1670.13
9	Sensory Perception of Smell (GO:0007608)	0.006417	0.08210	330.80	1670.13
10	Positive Regulation of Immune Response (GO:0050778)	0.008844	0.08210	16.35	77.28

(b) GO Biological Process 2025 enrichment results (linear models)

KEGG 2021 Human | Bar Graph | Table | Clustergram | Appyter

Hover each row to see the overlapping genes.

10 entries per page | Search:

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	Allograft rejection	0.00002189	0.0009720	82.39	884.00
2	Autoimmune thyroid disease	0.00002189	0.0009720	82.39	884.00
3	Type I diabetes mellitus	0.00003600	0.0009720	67.36	689.27
4	Graft-versus-host disease	0.00003600	0.0009720	67.36	689.27
5	Epstein-Barr virus infection	0.00003738	0.0009720	18.33	186.85
6	Human T-cell leukemia virus 1 infection	0.00004690	0.001016	17.42	173.65
7	Th1 and Th2 cell differentiation	0.00006449	0.001198	24.79	239.24
8	Th17 cell differentiation	0.0001485	0.002147	19.67	173.39
9	Viral myocarditis	0.0001487	0.002147	38.89	342.81
10	Systemic lupus erythematosus	0.0001942	0.002525	35.17	300.56

(c) KEGG 2021 Human enrichment results (nonlinear models)

GO Biological Process 2025 | Bar Graph | Table | Clustergram | Appyter

Hover each row to see the overlapping genes.

10 entries per page | Search:

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	Regulation of Transcription by RNA Polymerase II (GO:0006357)	0.000003670	0.0008846	12.73	159.29
2	Positive Regulation of T Cell Activation (GO:0050870)	0.00005420	0.006531	26.02	255.61
3	Regulation of DNA-templated Transcription (GO:0006355)	0.0001914	0.01538	7.92	67.84
4	Positive Regulation of Immune Response (GO:0050778)	0.0003105	0.01871	29.50	238.28
5	Regulation of Cell Cycle (GO:0051726)	0.0004255	0.02051	14.66	113.81
6	Positive Regulation of Transcription by RNA Polymerase II (GO:0045944)	0.0007573	0.03042	7.63	54.84
7	Negative Regulation of Transcription by RNA Polymerase II (GO:0000122)	0.001143	0.03935	8.35	56.59
8	Positive Regulation of Immune System Process (GO:0002684)	0.002065	0.05530	37.99	234.86
9	Antigen Processing and Presentation of Exogenous Peptide Antigen (GO:0002478)	0.002065	0.05530	37.99	234.86
10	Positive Regulation of DNA-templated Transcription (GO:0045893)	0.003358	0.07579	5.55	31.62

(d) GO Biological Process 2025 enrichment results (nonlinear models)

Figure 5: Enrichment analysis results for the top 30 SHAP genes derived from linear (panels a and b) and nonlinear models (panels c and d). Shown are the top 10 enriched terms (by adjusted p-value) in KEGG and GO Biological Process categories for each model type.

5 Discussion

Our enhanced linear models consistently outperformed the replicated models across the AIDA dataset and external validation sets. The observed reductions in MAE (such as up to ∼0.7 years for naive thymus-derived CD4+ T cells in AIDA and ∼5–6 years for certain cell types, such as erythrocytes, in the eQTL and Liu datasets, as shown in Figure 1) demonstrate that careful feature selection, standardized preprocessing, and parameter tuning can significantly enhance predictive accuracy. Importantly, these results support prior claims that linear models (ElasticNet) are well-suited for scRNA-seq aging clocks [6], while showing that performance remains sensitive to pipeline design and implementation choices, such as hyperparameter tuning.

When comparing our enhanced models with those reported in the reference study (Figures 7 and 8), we observe that our models replicate the core findings, achieving lower MAE (typically improved by ≈0.5 year) and Pearson correlations between true and predicted age (r values around 0.2–0.3 for the top-performing cell types). This confirms the robustness of single-cell aging signatures and supports the reproducibility of transcriptomic clocks under standardized reimplementation.

Regarding model generalizability, our results are aligned with the original study's observations: the clocks generalized well to external PBMC datasets such as Yoshida and Liu, but less so to the eQTL dataset (although we still saw noticeable improvements), where reduced gene expression variability contributed to lower performance [6]. This highlights a known limitation of current single-cell clocks, namely that they remain sensitive to dataset-specific biases, also noted by Hongming Zhu et al. [11] and others.

We further explored nonlinear models using LightGBM with hyperparameter optimization and nested cross-validation. As shown in Figure 3, these models slightly outperformed the enhanced linear models (and implicitly the replication models) on the AIDA training data, suggesting some benefit from capturing nonlinear interactions. However, their performance declined when evaluated on external datasets, likely due to increased sensitivity to domain shift, which is understandable due to the high discrepancy between the external datasets' distribution6. This performance gap underscores a key tradeoff: while nonlinear models are more expressive, they may generalize less effectively unless regularized, tuned properly or adapted to mitigate dataset-specific biases.

In terms of interpretability, SHAP analysis revealed that both linear and nonlinear models prioritize a mixture of globally predictive and cell-type-specific genes (Figures 2 and 4). Interestingly, several genes were consistently ranked highly in both model types, including **JUND** (ENSG00000130522), **FOS** (ENSG00000170345), **GNAS** (ENSG00000087460), and **KLF6** (ENSG00000067082), suggesting that these may represent robust, cell-type-agnostic biomarkers of immune aging.

To complement our SHAP-based interpretation, we conducted enrichment analysis on the top 30 SHAP-ranked genes for both linear and nonlinear models (Figure 5). In the linear models, the strongest KEGG enrichments included parathyroid hormone signaling and Th cell differentiation, which were linked to key immune and regulatory genes like **FOS**, **GNAS**, **CD3E**, and **HLA-DQA2**. GO terms also highlighted transcriptional control (notably through genes like **JUND**, **KLF6**, and **CBX4**) and T cell activation, reinforcing the relevance of immune and regulatory mechanisms in aging. In the nonlinear models, enrichment results were even more pronounced for immune-related KEGG pathways including allograft rejection, autoimmune thyroid disease, and Epstein-Barr virus infection, many involving HLA genes and **CD3E**. Likewise, transcriptional regulation and immune activation

remained top GO enrichments, with genes like **ZNF683**, **FOS**, and **JDP2** consistently contributing. These findings provide evidence that the most influential genes identified by our models are consistently associated with known aging-related and immune-regulatory functions. They suggest that immune remodeling and transcriptional shifts are central themes in transcriptomic aging and that the models, despite their small architectural differences, converge on biologically meaningful signals.

In a broader context, our work contributes to the growing field of single-cell aging clocks by demonstrating that enhancements in preprocessing and model tuning can yield meaningful performance gains even within the widely used ElasticNet framework. This supports the call from recent reviews [12, 2] for increased focus on model interpretability and reproducibility in aging biomarker research.

Together, these results suggest that transcriptomic aging signatures are at least partially robust to modeling assumptions, with key aging genes recoverable across both linear and nonlinear paradigms. Nevertheless, cross-cohort generalization remains a challenge, particularly when training and test populations differ in donor composition, size, sequencing depth, or preprocessing protocols. Prior work has shown that single-cell RNA-seq data is often plagued by batch effects and similar cohort-specific biases that can impair model transferability unless corrected [13].

Despite the strengths of our approach, several limitations should be acknowledged. First, the lack of publicly available training code from Zakar-Polyák et al. [6] required us to independently reimplement their ElasticNet models (to be able to improve them), potentially introducing deviations despite our best efforts to replicate the original methodology. Second, hyperparameter tuning for both our linear and nonlinear models was constrained by computational resources: the Delft Blue Supercomputer (which we used due to our local infrastructure lacking the necessary memory requirements) imposed a limit of 200 GB memory, a maximum of four concurrent jobs, and a considerable waiting time for each job, which significantly slowed the training process. As a result, our current optimization was limited in depth and scope. More exhaustive tuning may further improve model performance, especially for nonlinear architectures. Third, the external validation datasets differed substantially from the AIDA training data in terms of cell type composition, donor count, and age distributions (Figure 6), which likely contributed to reduced cross-cohort generalization. Finally, our current imputation strategy for missing genes for the external datasets may not fully capture dataset-specific gene expression differences, as it relies on average expression values from the AIDA training data.

Overall, our study contributes to the growing literature on single-cell aging clocks by demonstrating that enhanced preprocessing and hyperparameter tuning can yield substantial improvements in ElasticNet-based models, while highlighting the complementary strengths and limitations of nonlinear tree-based alternatives. These findings motivate future research into hybrid and domain-adaptive models that combine the interpretability of linear clocks with the flexibility of nonlinear architectures.

6 Conclusions

This study addressed the research question: *Can we improve on current models that predict biological age using single-cell gene expression data, and can we determine which specific genes are most important for making accurate predictions?*

We demonstrated that enhanced linear models based on ElasticNet regression can robustly predict transcriptomic age across a wide range of immune cell types. By carefully

replicating and improving the models of Zakar-Polyák et al. [6], we achieved consistent reductions in mean absolute error (MAE) on the AIDA training set as well as on external validation datasets, with typical improvements of 1–2 years and up to 5–6 years in certain cell types. These gains were made possible by systematic enhancements to preprocessing, feature selection, and cross-validation strategies.

In addition to linear models, we explored nonlinear age predictors based on LightGBM with hyperparameter optimization. These models slightly outperformed linear models on the AIDA training set, suggesting they can capture more complex patterns. However, nonlinear models showed reduced generalizability on external datasets, likely due to increased sensitivity to distributional shifts and computational constraints that limited deeper hyperparameter tuning. This illustrates the trade-off between model expressiveness and robustness when applied to new cohorts.

To interpret the biological relevance of our predictions, we used SHAP to identify both globally predictive and cell-type-specific aging genes. Notably, genes such as **JUND**, **FOS**, **KLF6**, and **GNAS** emerged as highly important across both linear and nonlinear models, consistently ranking among the most predictive genes. Enrichment analysis revealed that these genes were significantly associated with immune system regulation (such as T cell activation, Th cell differentiation), transcriptional control, and aging-related signaling pathways. These biological functions are widely implicated in immune aging and inflammaging (chronic, low-grade inflammation that typically develops as people age), further validating the relevance of our gene signatures.

Overall, our findings underscore the importance of methodological rigor in developing transcriptomic aging clocks and demonstrate that interpretable, high-performance models are attainable through careful optimization. Future work could address the computational challenges we met and build upon our findings and models in several ways. First, more extensive and parallelized hyperparameter optimization could be applied for both model types (nonlinear models are more sensitive to tuning) to obtain more precise predictions, for instance by replacing fixed categorical grids with log-scaled continuous search spaces for hyperparameters. Second, future work could focus on developing hybrid (linear and nonlinear) or domain-adaptive (improving generalization across datasets) modeling strategies to improve cross-cohort transferability. Third, training and validation should be expanded to include tissue-specific and disease-specific datasets to assess the generalizability of aging signatures across biological contexts. Finally, future work should explore external validation on larger single-cell datasets that are more similar in structure and composition to the training data (in terms of donor count, cell type representation, and donor diversity), to enable more comparable and generalizable model evaluations across cohorts. These efforts could enable more robust and generalizable single-cell aging clocks for biomedical research.

References

[1] S. Horvath, "Dna methylation age of human tissues and cell types," *Genome biology*, vol. 14, no. 10, pp. 1–20, 2013. [Online]. Available: https://doi.org/10.1186/gb-2013-14-10-r115

[2] A. T. Lu, A. Quach, J. G. Wilson, A. P. Reiner, A. Aviv, K. Raj, L. Hou, A. A. Baccarelli, Y. Li, J. D. Stewart *et al.*, "Dna methylation grimage strongly predicts lifespan and healthspan," *Aging (albany NY)*, vol. 11, no. 2, pp. 303–327, 2019. [Online]. Available: https://doi.org/10.18632/aging.101684

[3] C. Muralidharan, E. Zakar-Polyák, A. Adami, A. A. Abbas, Y. Sharma, R. Garza, J. G. Johansson, D. A. M. Atacho, É. Renner, M. Palkovits, C. Kerepesi, J. Jakobsson, and K. Pircs, "Human brain cell-type-specific aging clocks based on single-nuclei transcriptomics," *bioRxiv*, 2025. [Online]. Available: https://www.biorxiv.org/content/early/2025/03/02/2025.02.28.640749

[4] P. Mamoshina, M. Volosnikova, I. V. Ozerov, E. Putin, E. Skibina, F. Cortese, and A. Zhavoronkov, "Machine learning on human muscle transcriptomic data for biomarker discovery and tissue-specific drug target identification," *Frontiers in Genetics*, vol. 9, p. 242, 2018. [Online]. Available: https://doi.org/10.3389/fgene.2018.00242

[5] D. H. Meyer and B. Schumacher, "Bit age: A transcriptome-based aging clock near the theoretical limit of accuracy," *Aging Cell*, vol. 20, no. 8, p. e13320, 2021. [Online]. Available: https://doi.org/10.1111/acel.13320

[6] E. Zakar-Polyák, A. Csordas, R. Pálovics, and C. Kerepesi, "Profiling the transcriptomic age of single-cells in humans," *Communications Biology*, vol. 7, no. 1, p. 1397, 2024. [Online]. Available: https://doi.org/10.1038/s42003-024-07094-5

[7] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, 2022. [Online]. Available: https://doi.org/10.1016/j.inffus.2021.11.011

[8] 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," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[9] C. Han, N. Rao, D. Sorokina, and K. Subbian, "Scalable feature selection for (multitask) gradient boosted trees," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 885–894. [Online]. Available: https://doi.org/10.48550/arXiv.2109.01965

[10] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "Explainable ai for trees: From local explanations to global understanding," *arXiv preprint arXiv:1905.04610*, 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1905.04610

[11] H. Zhu, J. Chen, K. Liu, L. Gao, H. Wu, L. Ma, J. Zhou, Z. Liu, and J.-D. J. Han, "Human pbmc scrna-seq–based aging clocks reveal ribosome to inflammation balance as a single-cell aging hallmark and super longevity," *Science Advances*, vol. 9, no. 26, p. eabq7599, 2023. [Online]. Available: https://www.science.org/doi/abs/10.1126/sciadv.abq7599

[12] S. Horvath and K. Raj, "Dna methylation-based biomarkers and the epigenetic clock theory of ageing," *Nature Reviews Genetics*, vol. 19, no. 6, pp. 371–384, 2018. [Online]. Available: https://doi.org/10.1038/s41576-018-0004-3

[13] U. Shaham, K. P. Stanton, J. Zhao, H. Li, K. Raddassi, R. Montgomery, and Y. Kluger, "Removal of batch effects using distribution-matching residual networks," *Bioinformatics*, vol. 33, no. 16, pp. 2539–2546, 2017. [Online]. Available: https://doi.org/10.48550/arXiv.1610.04181

Supplementary Figures

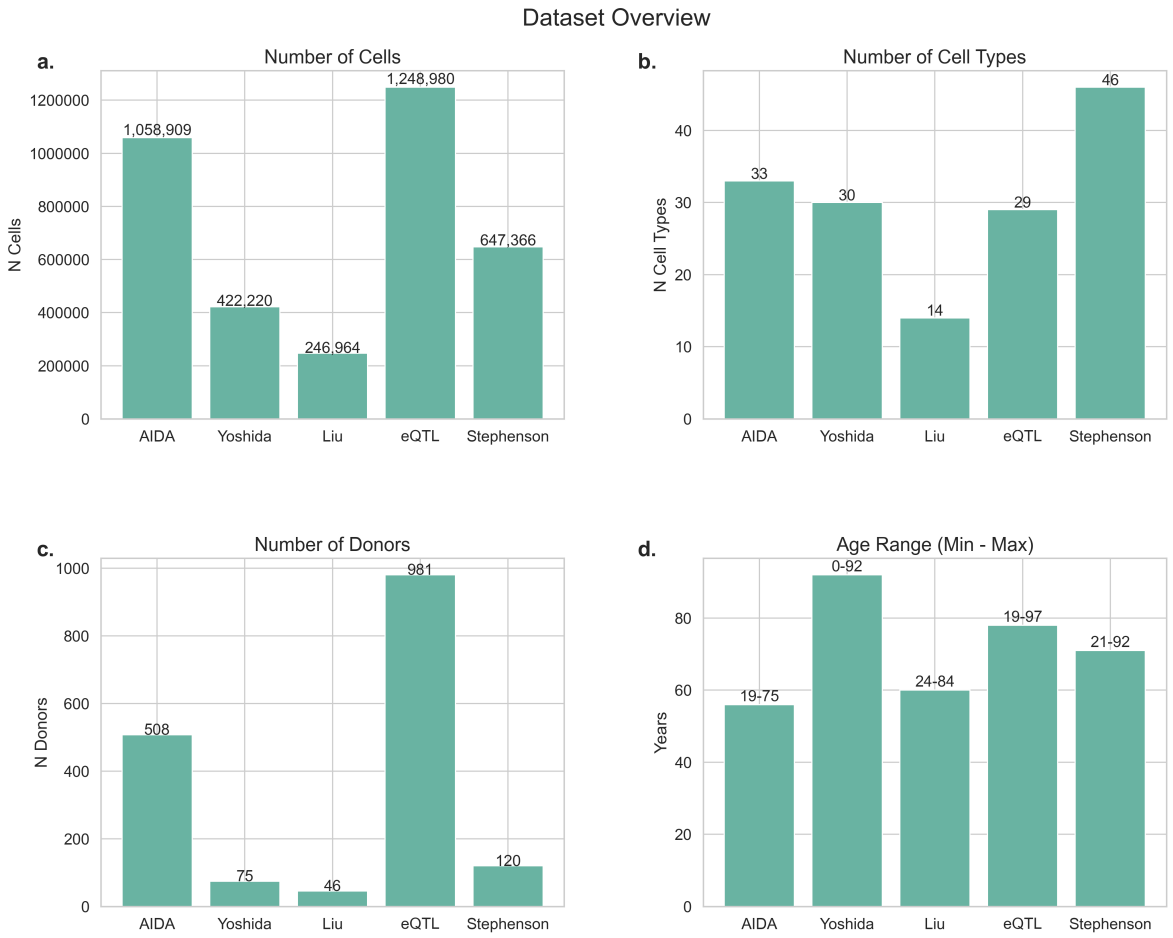


Figure 6: Overview of the five single-cell transcriptomic datasets used in this study. **(a)** Number of cells per dataset. **(b)** Number of distinct annotated immune cell types in each dataset. **(c)** Number of unique donors represented in each dataset. **(d)** Range of donor ages (in years), based on metadata from the development stage field of each dataset.

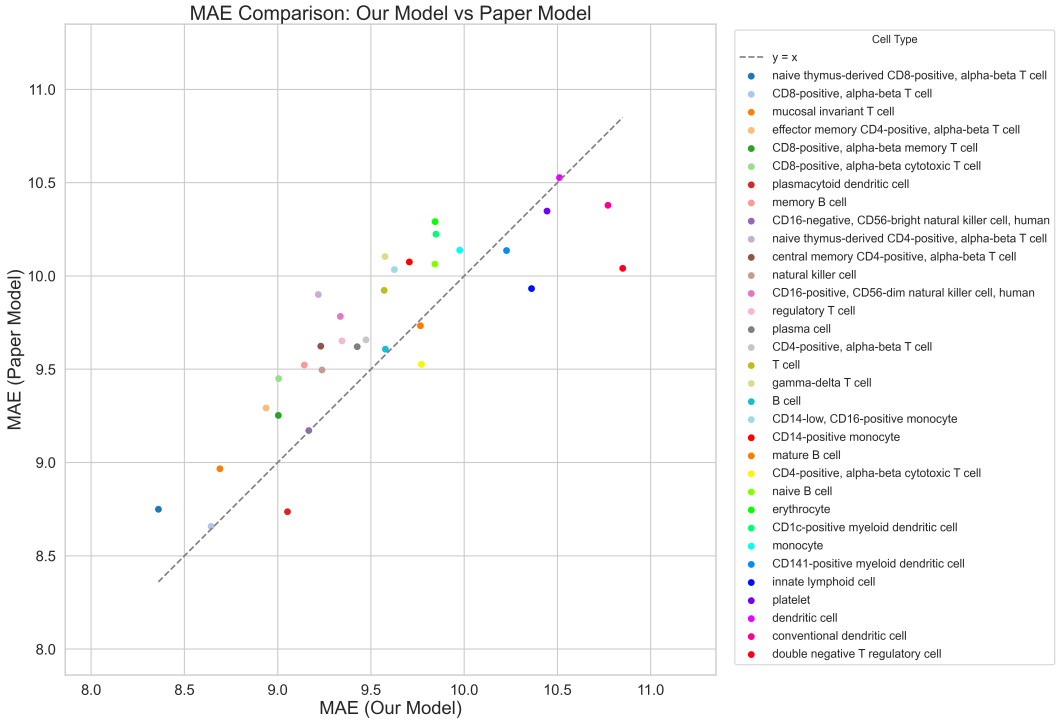


Figure 7: Scatter plot comparing the MAE of our enhanced linear models to the MAE reported in the published results from the original paper, across cell types. Each point represents a specific cell type, where the y-axis shows the MAE from the original paper and the x-axis shows the MAE of our enhanced model. Points above the diagonal line ($y = x$) indicate improved performance (lower error) in our models compared to the paper.

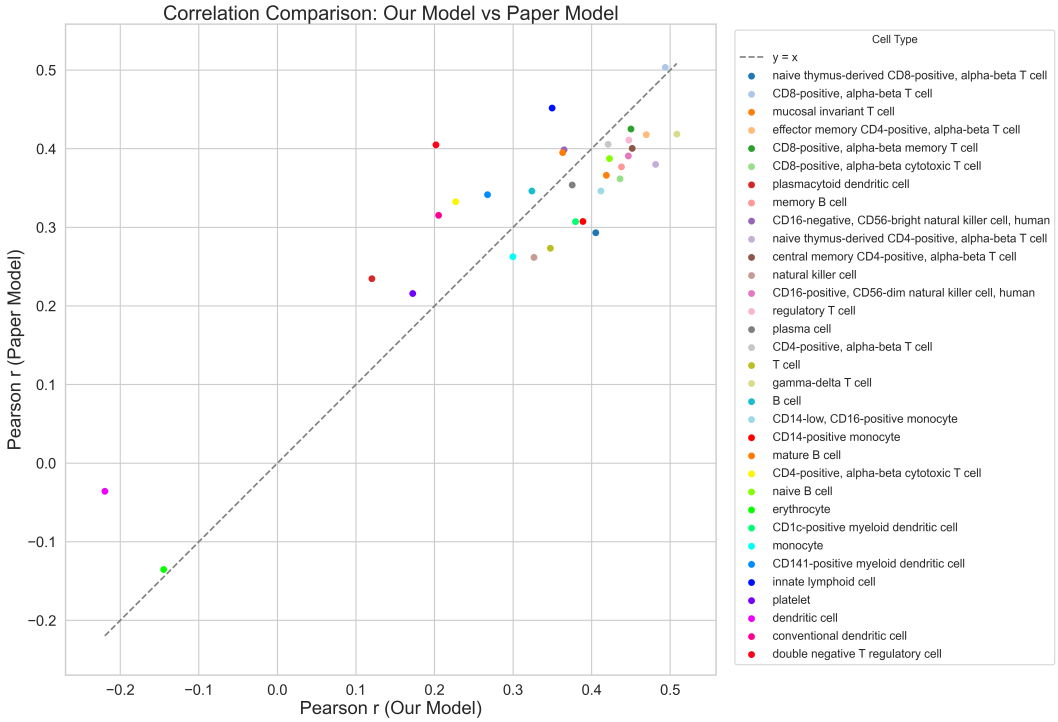


Figure 8: Scatter plot comparing the Pearson correlation between true and predicted donor ages for our enhanced linear models versus the values reported in the original paper, across cell types. Each point represents a specific cell type, with the y-axis showing the original model's correlation and the x-axis showing that of our enhanced model. Points below the diagonal line ($y = x$) reflect higher correlation (stronger age prediction accuracy) in our models.

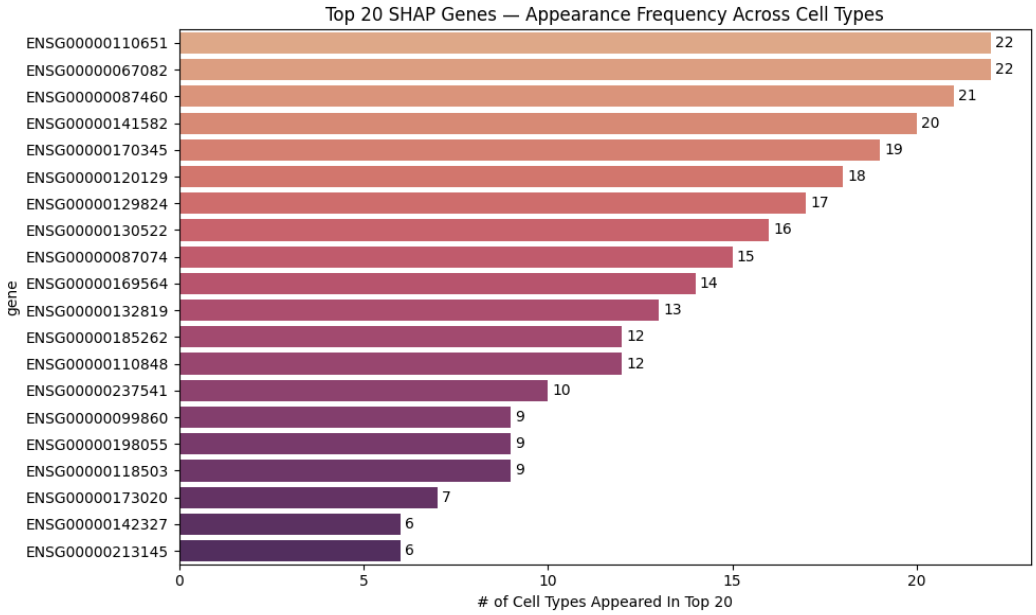


Figure 9: Frequency of appearance of genes in cell-type-specific top-20 SHAP lists for the linear models. For each cell type, we selected the 20 genes with the highest mean absolute SHAP values, then counted how many cell types each gene appeared in. The top 20 genes by appearance frequency are shown.

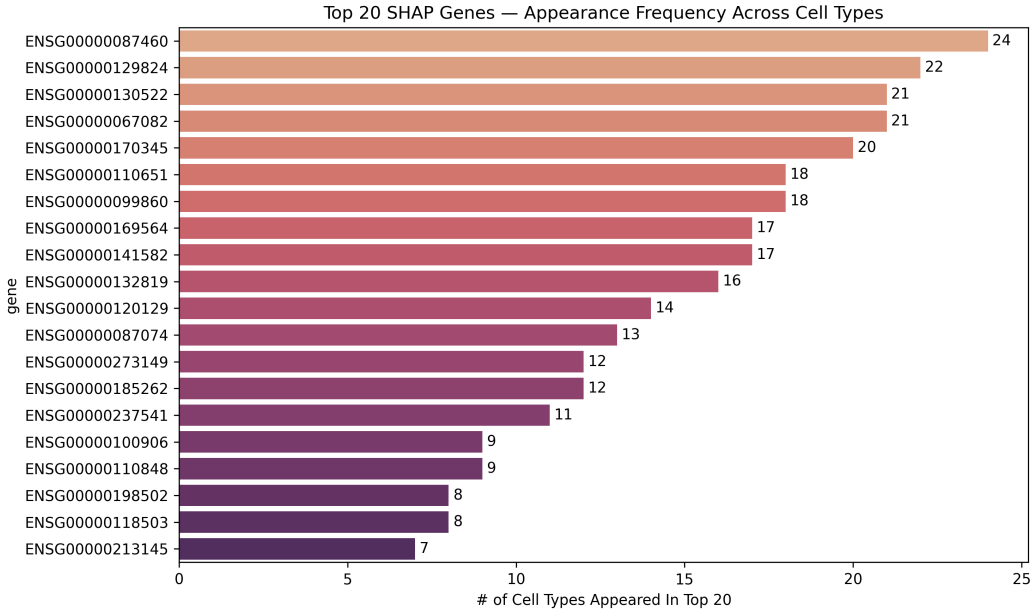
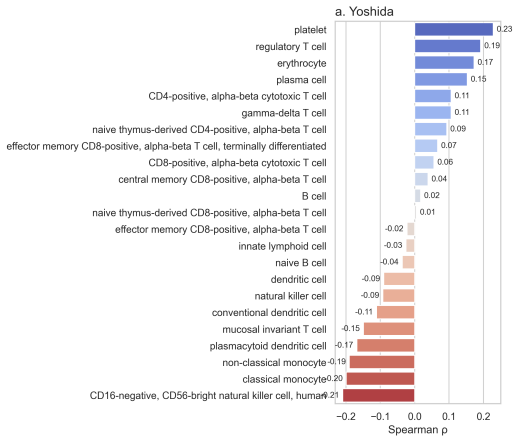
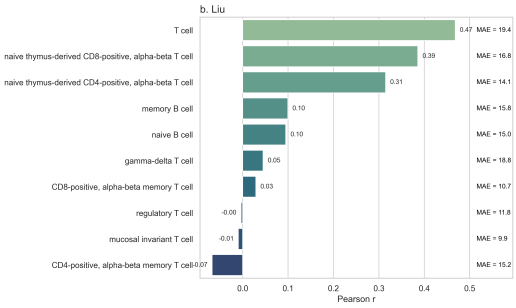


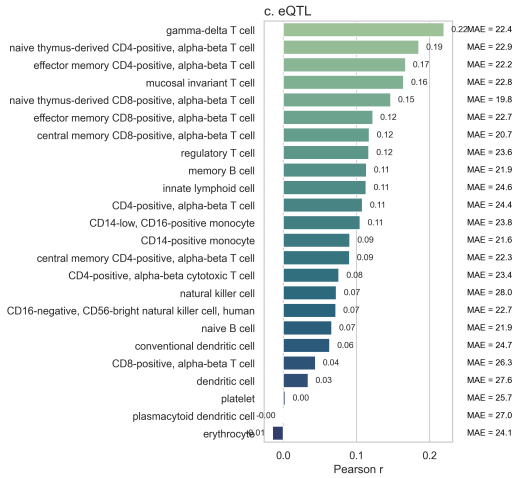
Figure 10: Frequency of appearance of genes in cell-type-specific top-20 SHAP lists for the nonlinear models. For each cell type, we selected the 20 genes with the highest mean absolute SHAP values, then counted how many cell types each gene appeared in. The top 20 genes by appearance frequency are shown.



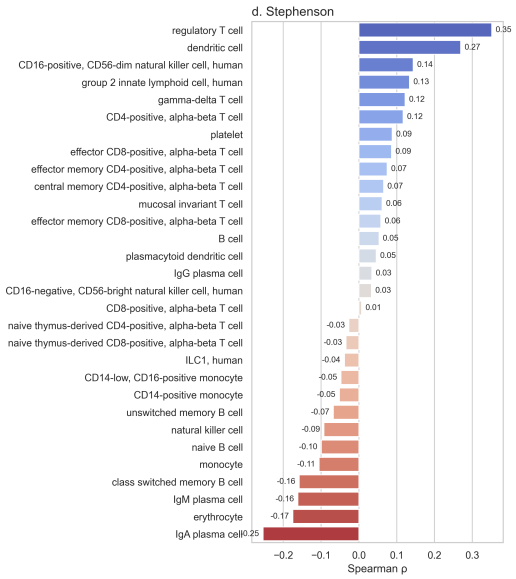
(a) Yoshida dataset (linear models)



(b) Liu dataset (linear models)

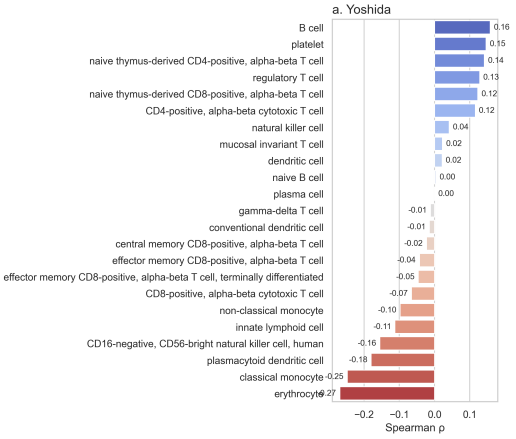


(c) eQTL dataset (linear models)

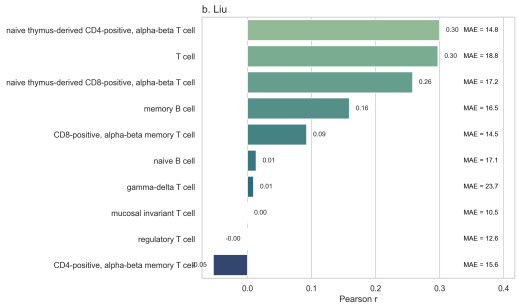


(d) Stephenson dataset (linear models)

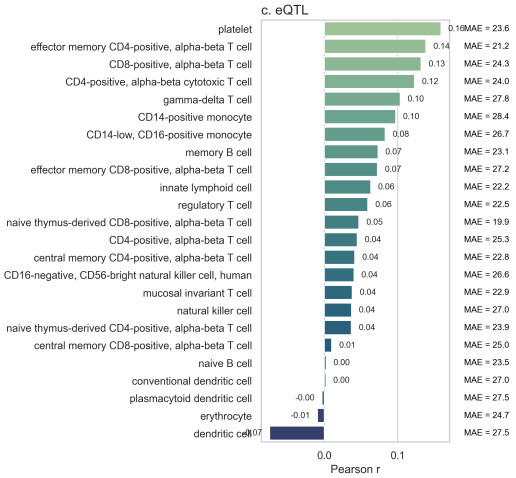
Figure 11: Cell-type-specific model performance (Spearman or Pearson correlation between predicted and actual age) across all external datasets for enhanced linear models. Higher values represent higher correlation, so stronger age prediction accuracy. Liu and eQTL (panels b and c) have the MAE displayed as an additional metric, since these two datasets contain only numerical age labels.



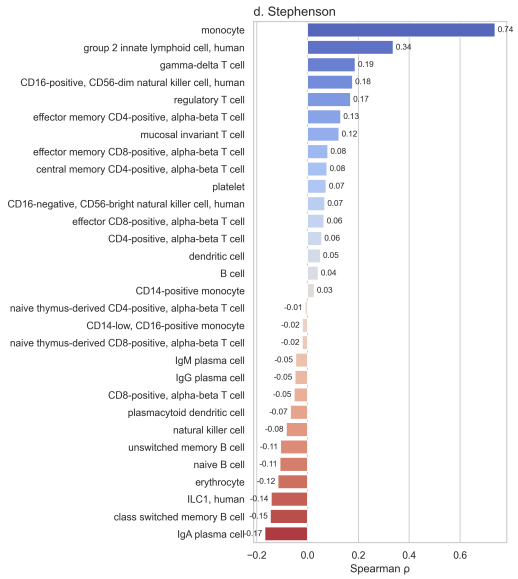
(a) Yoshida dataset (nonlinear models)



(b) Liu dataset (nonlinear models)



(c) eQTL dataset (nonlinear models)



(d) Stephenson dataset (nonlinear models)

Figure 12: Cell-type-specific model performance (Spearman or Pearson correlation between predicted and actual age) across all external datasets for nonlinear LightGBM models. Higher values represent higher correlation, so stronger age prediction accuracy. Liu and eQTL (panels b and c) have the MAE displayed as an additional metric, since these two datasets contain only numerical age labels.