Improving Single-Cell Transcriptomic Aging Clocks: Enhancing Accuracy and Biological Interpretability

1. Introduction

- Transcriptomic aging aims to estimate biological age based on gene expression profiles [2].
- Single-cell RNA sequencing (scRNA-seq) allows aging signatures to be studied at a cellular resolution.
- The study by Zakar-Polyák et al. [1] showed that ElasticNet can model immune aging signatures.
- A limitation of their study is the lack of feature importance or interpretability analysis, leaving uncertainty about which specific genes drive the age predictions.

2. 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?

3. Methods

Data preprocessing and model training

- Used AIDA dataset for training (~1m cells from 508 healthy human donors aged 19–75).
- Filtered genes: top 5000 most variable (variance-based per cell type) for linear models and top 3000 for nonlinear.
- Total-count normalization (10,000 counts/cell) and applied log-transformation and Z-score standardization per gene.
- Trained separate models per cell type: ElasticNet for linear (with the purpose of improving paper results) and LightGBM for nonlinear (to capture complex interactions in high-dimensional data [3]), both tuned using nested CV.

Model Application

- Applied the trained models to 4 external datasets.
- Imputed missing gene values using reference averages.
- Predicted age for each cell using all 5 models.
- Averaged predictions per cell (for the external datasets).

Evaluation

- Metrics: MAE, Pearson's r, Spearman's ρ, R^2.
- Performed feature importance via SHAP analysis.
- Applied functional enrichment analysis of top SHAP genes.

4. Results

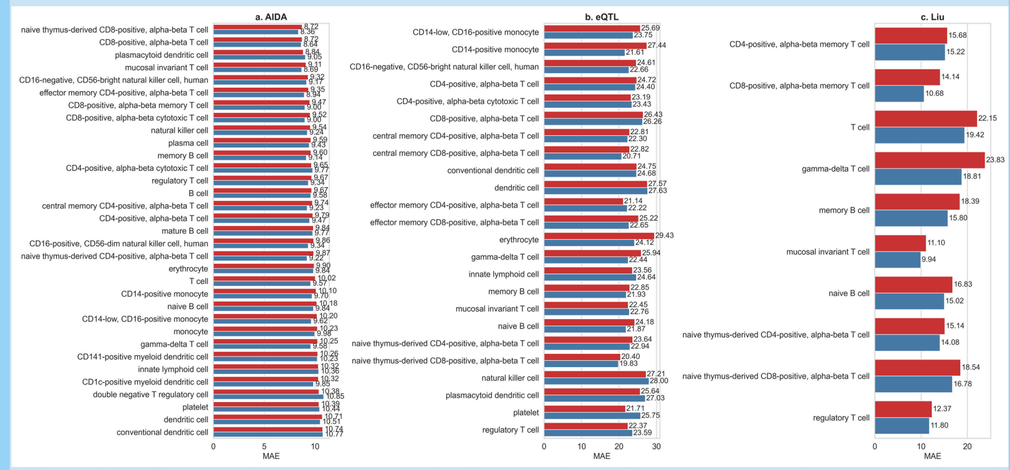

Figure 1: MAE comparison between replication models (red) and enhanced linear models (blue).

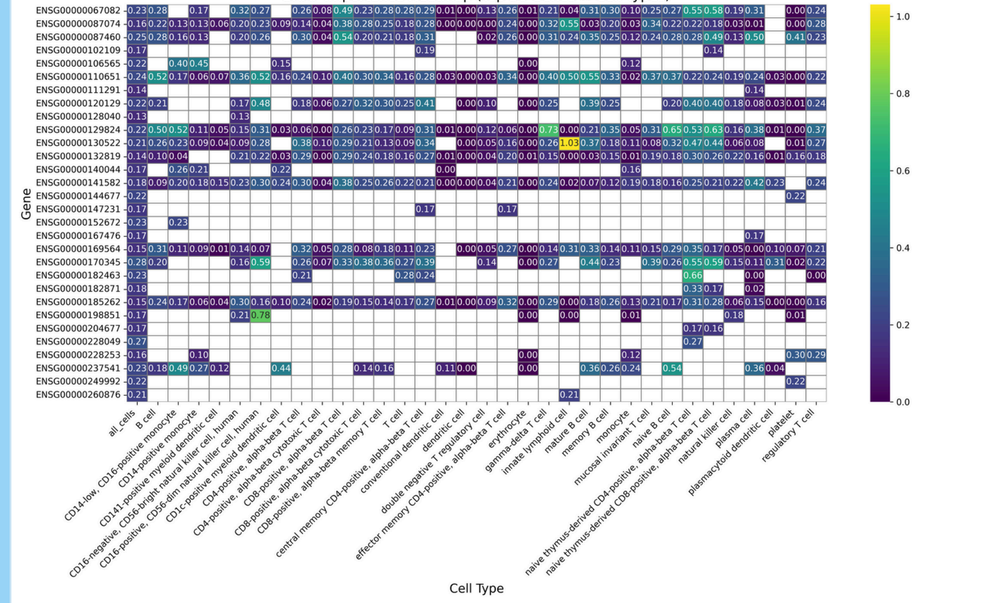

Figure 2: SHAP importance heatmap for the top 30 genes (ranked by global mean SHAP value across all cell types).

- Figure 1 showcases that enhanced models consistently outperformed replication models with median MAE reductions ranging between 0.5 and 1 year on the AIDA training dataset, and even higher improvements for the external datasets, with average MAE reductions between 2-3 years.
- Figure 2 reveals that some genes consistently show high importance across many immune cell types, while others are more cell-type-specific, with the first column highlighting a global summary of feature relevance.

5. Conclusions

- Enhanced linear models outperformed replication baselines, with MAE reductions of 1–2 years, and up to 5–6 years in some cell types (CD14-positive monocyte for the eQTL dataset), highlighting the impact of improved preprocessing and tuning.
- SHAP analysis revealed robust global and cell-type-specific aging genes, with key genes like **JUND**, **FOS**, **KLF6**, **GNAS**[1] consistently important across models.
- Enrichment analysis confirmed links between top genes and immune system processes, validating the biological relevance of learned features.

Future work

- Perform deeper hyperparameter tuning using log-scaled continuous search spaces.
- Focus on developing hybrid and domain-adaptive models to improve generalization.
- Explore datasets that are more similar in structure and composition to training data.

[1] Gene symbols in bold represent common gene names transformed for readability from Ensembl gene IDs (the ones displayed on the left of Figure 2).

References

[1] M. Zakar-Polyák, D. S. W. Lee, M. S. Arneson, et al., "Profiling the transcriptomic age of single-cells in humans", Nature Communications, vol. 15, no. 1, 2024.

[2] C. Muralidharan, E. Zakar-Polyák, A. Adami, A. A. Abbas, Y. Sharma, et al., "Human brain cell-type-specific aging clocks based on single-nuclei transcriptomics," bioRxiv, 2025.

[3] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, et al., "Light gbm: A highly efficient gradient boosting decision tree," Advances in Neural Informa tion Processing Systems, vol. 30, 2017.

Author: Vlad Alexan - V.Alexan@student.tudelft.nl

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