# What's your Immune Age? Uncovering Aging Pathways with Autoencoders

Luca Voros
High School

September 15, 2024

**Abstract**

*In this study, we use single-cell RNA sequencing data (scRNA-seq) from peripheral blood mononuclear cells (PBMCs) of 166 donors to predict cell-type specific, and overall biological age. We categorize each of the donor's cells into nine groupings, based on existing annotations. We developed and benchmarked four distinct neural networks: Pseudo-Bulked, Cell-Level, Pseudo-Bulked-AE, and Cell-Level-AE. The models either employed total normalized gene expression data or compressed latent representations generated by an autoencoder. Our findings show that our models that utilize autoencoder-based latent features (Cell-Level-AE and Pseudo-Bulked-AE) demonstrate superior performance than those that rely on normalized gene expression data. Additionally, we applied Random Forest Regression to integrate our predictions across cell types, which yields an estimate of the donor's age. We added SHAP (SHapley Additive exPlanations) analysis to interpret the outputs of a single model, identifying key genes that influence the aging process across individuals. The research highlights the application of single-cell data and machine learning in predicting biological age, and developing interventions aimed at potentially reversing the aging process.*

# 1 Introduction

Aging is a complex process that is marked by a decline in physiological function and increased susceptibility to disease. This process is significantly influenced by a decline of function in the immune system, referred to as immunosenescence. Chronological age as a measure of the aging process is insufficient, as it does not take note of the biological mechanisms that drive aging, many of which are tissue and cell specific. Research has been increasingly focused on molecular data as a metric for biological age, which may more accurately represent an individual's health status. Discordance between chronological and biological age indicates whether an individual's health is better or worse than expected. Understanding the driving factors of this discordance enables interventions at both the individual and population level. Identifying and shifting these biological pathways associated with specific aging patterns may improve overall health.

This study uses single-cell RNA sequencing (scRNA-seq) data from peripheral blood mononuclear cells (PBMCs) from 166 donors to develop AI models that can predict the biological age of various cell types, and overall biological age. Further analyses, such as SHAP (SHapley Additive exPlanations), are used to decipher the genes that most influence age predictions, providing biomarkers for biological aging.

# 2 Methods

## 2.1 Data collection and preprocessing

The dataset comprises gene expression profiles of approximately 36,601 genes across nine cell types: CD8+ T cells, CD4+ T cells, NK (Natural Killer) cells, B cells, gd T cells, MAIT cells, Myeloid cells, Progenitor cells, and DN T cells. In addition, the normalized single-cell gene data was accompanied by metadata with various clinical variables, including Age and Sex. To filter our analysis, we excluded genes that were not variably expressed (had near constant expression), and those that were present in fewer than three cells, resulting in a filtered dataset of approximately 1,390 genes. For model training and validation, the data was split. The models were trained on 80% of the donors, and the models were tested on the remaining 20%.

## 2.2 Model architectures

### 2.2.1 Autoencoder

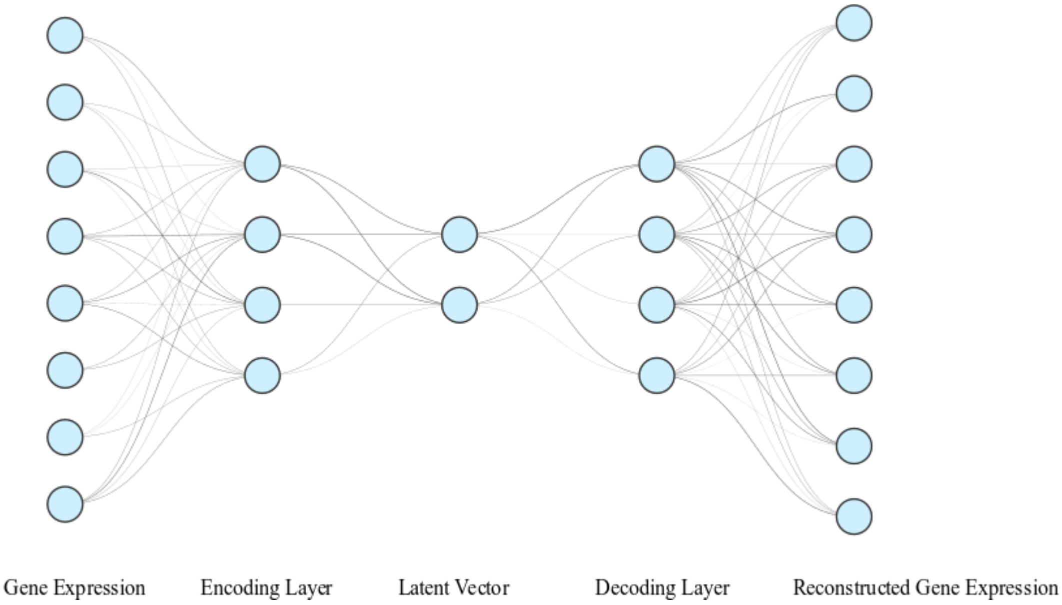

**Figure 1.** A simplified representation of our architecture. We used two encoding layers of size 286 and 222, with the ReLU activation function. The latent layer is of size 128. The decoding layers are of size 222 and 286, activation=ReLU. We chose these sizes to balance accuracy and efficiency. The output layer is the same size as the input. This model was trained on 1390 genes, and 800,000 cells for 15 epochs. Mean squared error was used as the loss function.

Autoencoders, a form of neural network, are used to compress and reconstruct high-dimensional data, such as gene expression profiles. By learning patterns and features in the large RNA-seq datasets, the autoencoders can reduce the data's complexity. These compressed representations, or latent representations, generated by autoencoders can be used for various downstream analyses, including predicting cell function, diseases, and, in this study, age. The autoencoder's ability to reconstruct high-dimensional data from its compressed form allows it to format the latent vectors optimally, preserving maximal information from the original data (Figure 1). These representations can be visualized in a two-dimensional space (Figure 2).

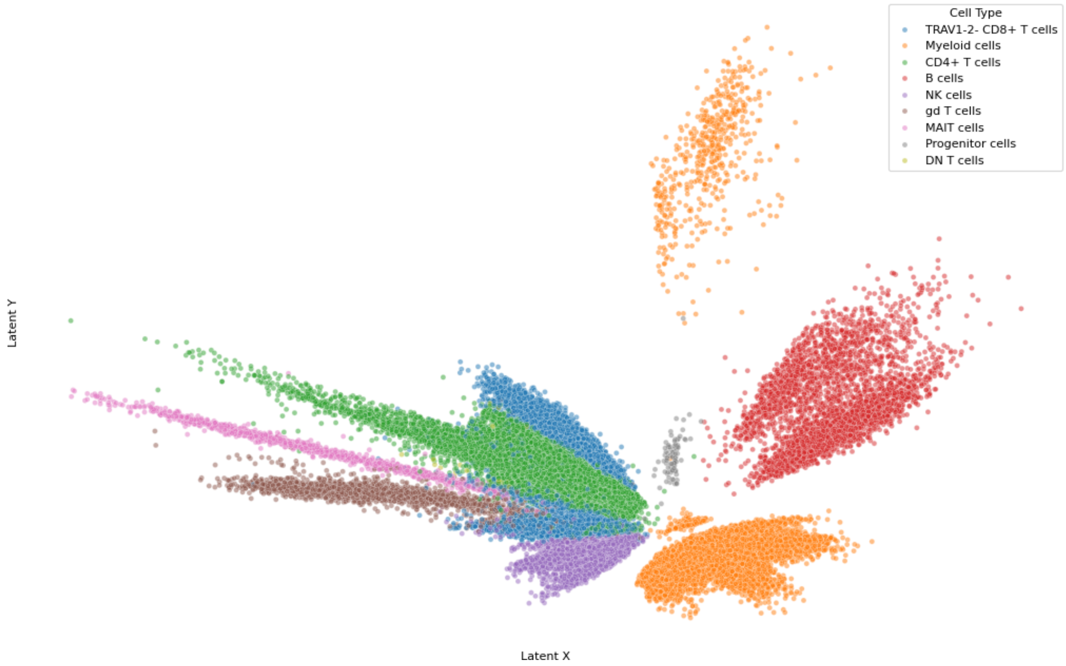

**Figure 2.** A visualization of cells in the latent space of a two-dimensional autoencoder. A separate autoencoder was trained to achieve this, the latent layer of which was then plotted and colored by cell type.

### 2.2.2 Neural networks

**Pseudo-Bulked Model** We aggregate the gene expression data for each cell type by averaging across all cells in a single donor. The averaged expression profiles are then used to train a neural network to predict the donor's age. (Figure 3)

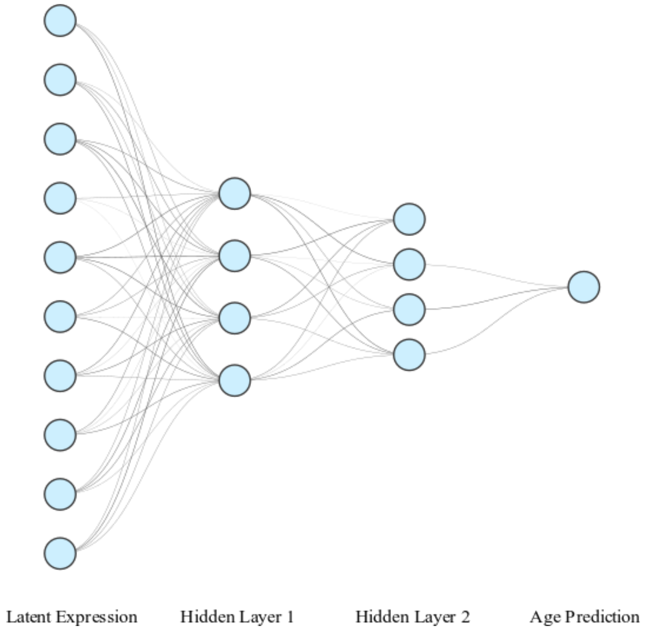

**Figure 3.** A simplified form of the model's architecture. All models share similar architecture. They have two fully connected hidden layers of size 20, with ReLU activation and L2 regularization to prevent overfitting. Each hidden layer is followed by a Dropout layer, with a dropout rate of 25% to prevent overfitting. There is one output node, for age. The Cell-Level models were trained for 10 epochs, while the Pseudo-Bulked models were trained for 40 epochs. Mean Squared Error was used as the loss function.

**Cell-Level Model** This model was trained on individual cell data, and predicts the donor's age based on each individual cell's gene expression profile. After predicting on our test set, the age predictions of an individual's cells in a cell type are then averaged to determine the cell-type's age. (Figure 3)

**Pseudo-Bulked-AE Model** Similar to the Pseudo-Bulked model but using features derived from the autoencoder, which compresses the gene expression data into 128-dimensional latent space prior to training. (Figure 3)

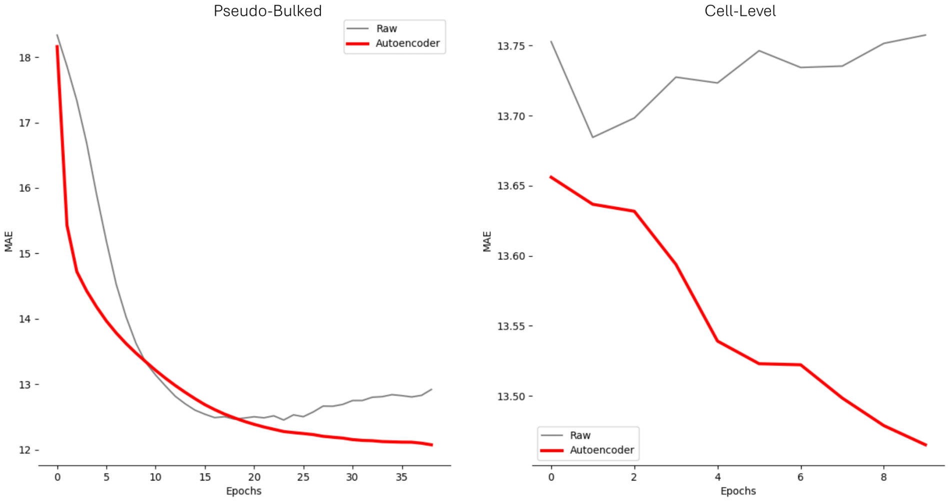

**Figure 4.** Model performance (measured by validation MAE) of the four models plotted over the training epochs. Overfitting can be seen when the validation MAE starts to rise, and the model gets worse at predicting on new data.

**Cell-Level-AE Model** Similar to the Cell-Level model but using features derived from the autoencoder, which compresses the gene expression data into 128-dimensional latent space prior to training. (Figure 3)

**Random Forest** In this study, we employed Random Forest Regression, configured with 1000 estimators, to identify the relationship between predicted cell type ages, and the overall age of the donor. The input feature was 18-dimensional, consisting of 9 features representing cell-type ages, and 9 features consisting of the number of that cell type, in that donor. Random Forests operate by constructing a multitude of decision trees, each processing the data independently and coming to its own conclusion. The final decision is made through an aggregation of the individual predictions. Random Forest Regressors were chosen due to their superior performance compared to alternative methods, such as Linear Regression.

**Evaluating performance of models** Model performance is evaluated by Mean Absolute Error - which is the average magnitude of errors in a set of predictions, agnostic of their direction. In this context, MAE quantifies the average difference between the predicted ages and the actual ages. Therefore, lower MAE indicates better overall model performance.

# 3 Results

## 3.1 Neural network performance

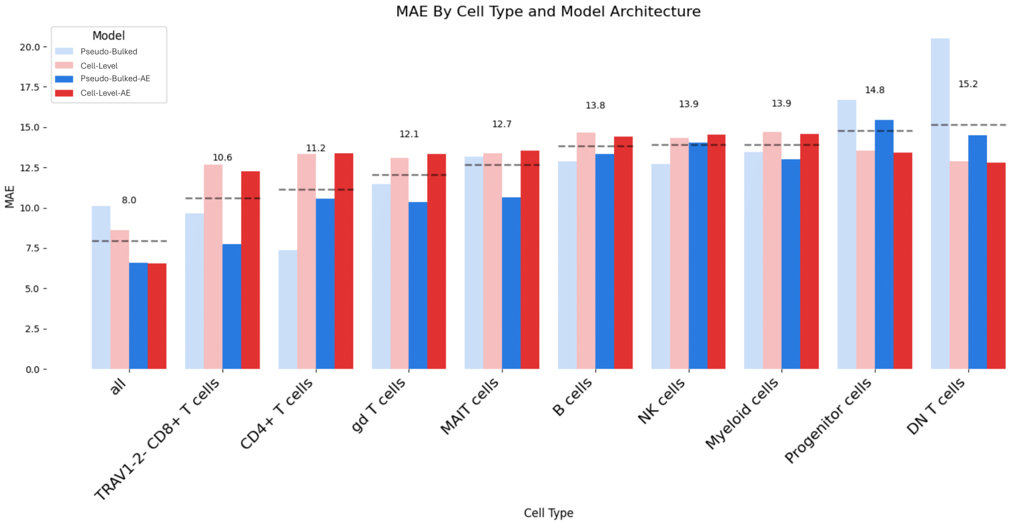

**Figure 5.** A figure summarizing the performance of the four models across cell types, and the random forest model (marked "all").

The Cell-Level models perform better in the random forest cell-type aggregate, and the Autoencoder models also perform much better than the raw models in the random forest. Interestingly, across cell types, the Pseudo-Bulked models perform better than the Cell-Level models. The Autoencoder models had much more consistent results, as well as often better results. Additionally, the cell types are ranked by their performance - so the cell types that best indicate age are: CD8+ T cells, CD4+ T cells, and gd T cells in that order. This seems to agree with the agglomerate literature.

As illustrated in Figure 4, the Autoencoder models performed significantly better than the non-Autoencoder models when it came to predicting overall individual age. We get a better idea of this when we look at the training history of the four models (Figure 5).

Overfitting is a common problem in machine learning, particularly in situations where the dataset size is fairly small. Overfitting is where the model memorizes the training data without actually understanding the true patterns. Evidence for this can be observed happening in Figure 5, where for both the Pseudo-Bulked and Cell-Level “raw” models, the validation MAE rises at the tail end of training. Additionally, the Cell-Level “raw” model performs far worse than its autoencoder-based counterpart during training, not making any progress after the first epoch.

The incorporation of autoencoders into these models seems to improve this issue, with both of the autoencoder models not overfitting, as well as converging into a lower MAE than the “raw” models- meaning that they trained much more effectively. Overfitting usually occurs when the model spots a pattern in the data, that occurs purely by chance. This is more likely to happen when you have less data, and more features (because there is a higher chance that a feature is falsely correlated with the labels). By reducing the complexity of the input from 1390 genes into 128 latent features, the models are better able to understand the real patterns behind the data - and successfully generalize to new data. This is known as the curse of dimensionality. (Kuo et all, 2005)

## 3.2 Analysis of feature impact with SHAP

SHAP (SHapley Additive exPlanations), is an approach that explains the output of machine learning models by assigning each feature (genes, in this case) values based on their contribution to the output of a model. In the context of predicting age with gene expression, SHAP can tell us which of the genes inputted to our neural networks are driving the age prediction the most. The plot indicates direction of effect, so for example - down-regulation of CEBPD is associated with prediction of increased age, while the opposite is true for CCL5 where up-regulation predicts increased age.

Figure B-E shows the predicted age of four individuals, compared to their actual age. These four individuals in particular had high discordances between their actual and predicted biological ages. This indicates that their gene-expression differed in a crucial way from the gene expression of those in their age group. The genes colored in blue are pulling the age prediction down (meaning that they are predictive of being younger), while the genes

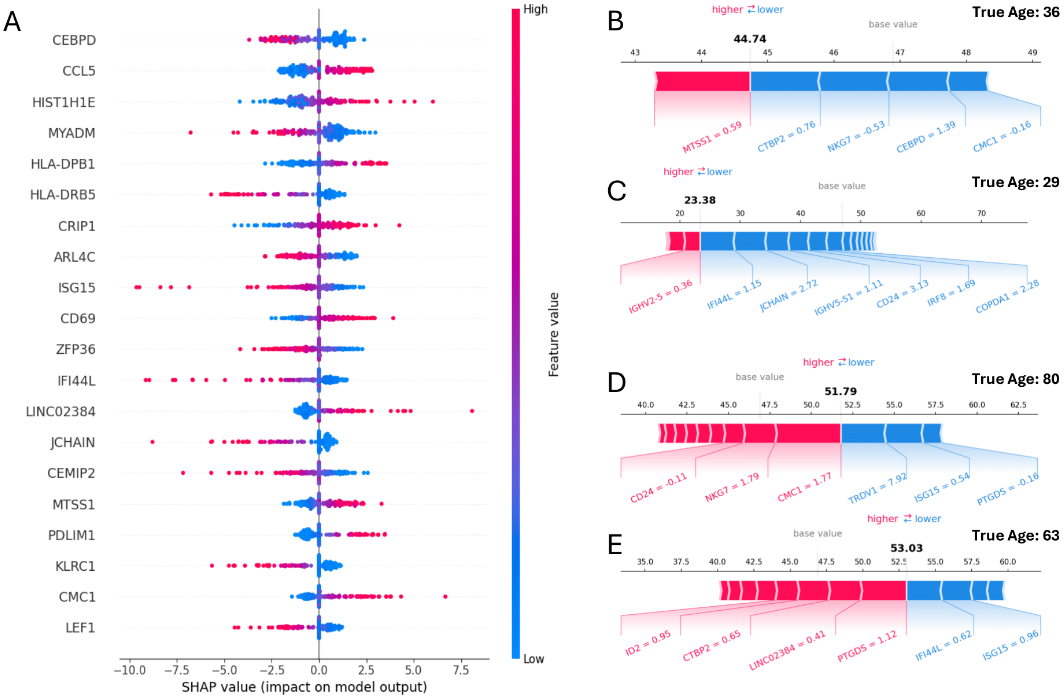

**Figure 6.** A) summary of the 20 most impactful genes on model prediction. B-E) Force plots of individuals who had large discordances between predicted and actual age. We used the Pseudo-Bulked model with SHAP analysis to identify genes that impact the prediction of age. Using an autoencoder model would be tedious with SHAP (due to its additional state as a black box), as well as greatly increase the complexity of the problem.

colored in red are pushing the age prediction up (meaning that they are predictive of being older).

Figure A demonstrates that the top genes which had the most effect on how our model predicted age were: CEBPD, CCL5, and HIST1H1E. We tested the top 100 most impactful genes using the pathway analysis tool Enrichr (https://maayanlab.cloud/Enrichr/enrich), which identifies common biological pathways among the genes that are predictive of age. Table 1 shows the top five pathways uncovered, which included pathways relating to interferon signaling and cytokine activation.

SHAP analysis shows that downregulation of genes involved in interferon signaling (including ISG15 and IFI44L) was associated with aging (Figure 6A). This suggests that interferon

| Relevant Pathways | | |
|---|---|---|
| Reactome Pathway (ID) | P-Value | Adjusted P-Value |
| Interferon Gamma Signaling (R-HSA-877300) | 0.0000055 | 0.0013 |
| Cytokine Signaling In Immune System (R-HSA-1280215) | 0.00001 | 0.0013 |
| Interferon Signaling (R-HSA-913531) | 0.000065 | 0.0047 |
| Immune System Interferon Signaling (R-HSA-913531) | 0.00007 | 0.0047 |

**Table 1.** We tested the top 100 most impactful genes using the pathway analysis tool Enrichr (https://maayanlab.cloud/Enrichr/enrich), which identifies common biological pathways among the genes that are predictive of age.

levels fall as an individual ages. ISG15, in particular, is strongly involved in inhibiting viral replication - as demonstrated by the large number of viral immune-evasion proteins that target it (Perng et al, 2018). These results highlight how aging is associated with defective interferon signaling, leading to a weaker immune system, demonstrated for example by inadequate responses to viruses (Feng et al, 2021). Particularly, interferon signaling is crucial in response to SARS-CoV-2 (Feng et al, 2021), elderly people being more prone to mortality to COVID-19.

Some individuals displayed significant differences between chronological age, and immune age. SHAP-based force plots can be used to identify the genes driving these discrepancies (Figure 6B-E). For instance, the individual shown in Figure 6D showed the greatest discordance - being predicted to be 52, when their chronological age was 80. This difference was driven primarily by TRDV1, ISG15, and PTGDS. Many of the genes that contributed to discordant predictions were involved in interferon signaling (including IFI44l, IRF8, and ISG15) (Figure 6B-E). This suggests that some individuals may develop age-related immune dysfunction, such as diminished interferon levels, earlier than others (while others might preserve their function for longer).

# 4 Conclusion

This study clearly demonstrates the potential of single-cell RNA seq data paired with AI for predicting immune age. Notably, the integration of autoencoders ensemble methods underline the benefits of complexity-reduction and the use of multi-cell type data - specifically in the context of small datasets. The results present a framework for the prediction of immune-age as a helpful indicator for health. This study additionally identifies several genes and pathways, such as those involved in interferon signaling, that play a significant role in immune-aging and the possible discordance between chronological age and biological age. The analysis of these discordant cases offer unique insights into the aging process, and can illustrate why certain individuals may age faster, or slower. These findings point to potential mechanisms that can decelerate, or possibly reverse the aging process.

# 5 Acknowledgements

The work described here was performed during a 4 week internship at the Digital Environment Research Institute, Queen Mary University of London, under the supervision of Prof Michael Barnes and Dr Jack Gisby.

# References

[1] Terekhova M, Swain A, Bohacova P, Aladyeva E, Arthur L, Laha A, Mogilenko DA, Burdess S, Sukhov V, Kleverov D, Echalar B. Single-cell atlas of healthy human blood unveils age-related loss of NKG2C$^{+}$ GZMB CD8+ memory T cells and accumulation of type 2 memory T cells. Immunity. 2023 Dec 12;56(12):2836-54.

[2] Kuo FY, Sloan IH. Lifting the curse of dimensionality. Notices of the AMS. 2005 Dec;52(11):1320-8.

[3] Perng, YC., Lenschow, D.J. ISG15 in antiviral immunity and beyond. Nat Rev Microbiol 16, 423–439 (2018). https://doi.org/10.1038/s41579-018-0020-5

[4] Feng E, Balint E, Poznanski SM, Ashkar AA, Loeb M. Aging and Interferons: Impacts on Inflammation and Viral Disease Outcomes. Cells. 2021 Mar 23;10(3):708. doi: 10.3390/cells10030708. PMID: 33806810; PMCID: PMC8004738.