Beyond DNA methylation: chromatin age of human tissues

A THESIS PRESENTED
BY
MUHAMMAD HAIDER ASIF
TO
THE DEPARTMENT OF COMPUTER SCIENCE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
BACHELORS OF SCIENCE
IN THE SUBJECT OF
COMPUTER SCIENCE

BROWN UNIVERSITY
PROVIDENCE, RHODE ISLAND
MAY 2022

Thesis advisor: Professor Ritambhara Singh Muhammad Haider Asif

Beyond DNA methylation: chromatin age of human tissues

Abstract

DNA methylation-based methods, especially ones based on linear regression, have proven very accurate at predicting age. More recently, neural networks have outperformed them and provided valuable insights through interpretation methods such as SHAP (SHapley Additive exPlanations) and DeepPINK. These methods, however, are limited by the assay to probe DNA methylation, which only covers part of the human genome. Histone ChIP-Seq data, on the other hand, covers the entire genome, which opens up a wider array of possible insights that we can derive using interpretation methods. Here, I gather histone ChIP-Seq data from ENCODE and Gene Expression Omnibus (GEO) for the six most commonly probed histone marks. Then, I validate the viability of histone ChIP-Seq data by using it to predict age accurately using existing methods such as ElasticNet. As seen with DNA methylation, optimized neural networks for the six histone marks out-perform ElasticNet on unseen data, and the latest deep interpretation methods such as SHAP help to identify age-related pathways and complex genomic region-region interactions. Neural networks can perform better than ElasticNet due to their ability to capture complex region-region interactions in genomic regions, which contain genes that are known to impact human aging.

Contents

Acknowledgments

I would like to thank Lucas Paulo de Lima Camillo and Professor Ritambhara Singh for introducing me to the field of aging research and for their constant guidance, help, and support. I would also like to thank Professor Louis Lapierre for his valuable feedback on my work.

0
Introduction

Aging is naturally accompanied by an increased risk of developing a range of diseases and leads to deteriorating physical and mental health. Aging research can allow us to detect the consequences of aging early on, along with the factors that might cause them. Such insights allow scientists to design interventions to alleviate the impact of aging. However, aging research is challenging since accompanying individuals for years and monitoring their health is too time and resource-intensive. Over the last decade, there have been concerted efforts to construct an accurate and precise biomarker of human aging, typically from DNA methylation data. Understanding which factors impact age predictions can allow scientists to design targeted interventions and discover novel pathways that can open the way for data-driven drug discovery that targets aging itself, the single largest risk factor for death worldwide [1].

In recent years, several DNA methylation-based, linear age predictors that use epigenetic data, dubbed epigenetic clocks, have been created. Two of the most well-known predictors are ones developed by Hannum and Horvath in 2013. [2], [3] Both these predictors rely on ElasticNet to predict age with low mean squared and median absolute errors (MSE and MAE). Recently, however, there has been an attempt to improve the accuracy of linear models by using deep learning methods. One such paper by the Singh Lab at Brown University, AltumAge [4], improved the existing best accuracy on such linear models by harnessing the flexibility and potential of neural networks. However, despite good performance, the interpretability of DNA methylation epigenetic clocks is limited by the assay to probe DNA methylation, which covers only part of the human genome. Most of the features consequential to aging are elusive and may not be real drivers of aging when tested in the laboratory.

More recently, Professor David Sinclair from Harvard Medical School has proposed the "information theory of aging", which postulates that the cause of human aging is the loss of epigenetic information analogous to the scratching of a CD [5]. It has been shown that such information can be recovered for rejuvenation [6]. Histone marks are a fundamental pillar of epigenetic information and several authors have proposed histone codes [7] that explain how such modifications to histone marks can regulate the attributes of living beings. The examples of age-related histone mark changes are plentiful. One histone mark, H3K4me3, showed a unique pattern of change during aging in C. elegans [7]. Around 30 percent of the H3K4me3 enriched regions exhibited significant and reproducible changes with age. A study by Sun et al. reported that H3K4me3 and H3K27me3 show that aged Hematopoietic stem cells exhibit an increase in the number of H3K4me3 peaks while also producing broader H3K4me3 and H3K27me3 peaks across genes related to the self-identity and renewal of hematopoietic stem cells [7]. Overall, it has been established that several histone marks change predictably with age. Consequently, I have developed deep learning networks for the six most commonly probed histone marks. These networks predict age much more accurately on unseen data than can ElasticNet, and the latest deep interpretation methods can help to identify age-related pathways due to the nature of the histone mark data. Specifically, I use SHAP [8] to identify genomic regions important to age prediction, along with relationships between the most important genomic regions. Analyzing the genes present in these genomic areas can validate our networks' results if the genes are known to impact age. Moreover, complex genomic region-region interactions can capture the possible combined role of genomic regions in predicting age.

1

Methodology

1.1 Data Pre-Processing and Description

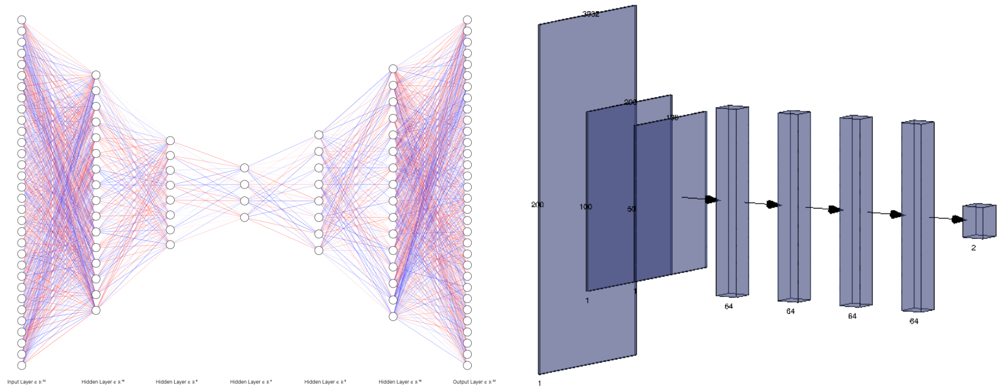
The pan-tissue histone ChIP-seq data for each of the six histone marks, H3K4me3, H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K9me3, was downloaded from the ENCODE consortium and followed their data processing guidelines. [9] Gene Expression Omnibus (GEO) data was also gathered from publicly available data sets to use for model inference and testing purposes. The data consists of 67 different tissue types including spleen, aorta, brain, and stomach. The age range represented ages from -0.542 (which represented approximate fetal age, considering a 40-week pregnancy as standard, age at birth = 0) to high 89.5. The mean age of the dataset was 44.06. The assemblies used were GRCh38 and hg19, with histone ChIP-Seq Assays.

The GEO data was downloaded as fastq files for both the control and the histone mark accession codes, before being fed into ENCODE's Chip-seq pipeline [9], which processed this data and converted it into bigWig files, which is a much more compressed and processable form of data.

The data downloaded directly from ENCODE was in the form of bigWig files and consisted of both fold chain over control, and signal p-value files. For my experiments, I only use fold chain over control files for each accession code. The human genome has about 3 billion base pairs, which makes using the raw data in a model almost impossible due to the sheer size of its' feature space. This is why I took the mean of the fold-change value of a 10Kb region to construct 10Kb bins, which is an ideal size since the average human gene size is about 5-7 Kb. Therefore, this bin size can allow us to reveal important genes by using interpretation methods such as SHAP values on the models, while at the same time predicting age accurately. The data was then converted into mean pickle files for each histone mark for faster data retrieval for model training purposes. The final data contained approximately 277 samples for H3K4me3, 269 samples for H3K27ac, 215 samples for H3K27me3, 225 samples for H3K36me3, 232 samples for H3K4me1, and 216 samples for H3K9me3, each with 30321 features due to the 10Kb binning. The number of samples to features ratio is not ideal, but increasing the size of the bins further leads to the loss of important data and features, for example, a 100Kb bin would have too many different genes and hence might not lead to valuable results post-processing. This led me to use an auto-encoder to reduce the dimensionality of the feature space to improve the samples to features space ratio. This auto-encoder introduced Gaussian noise into the data before training, which allowed it to learn the underlying representation of the data better and increase the robustness and generalizability of the data. For most of the models, the latent size or the reduced feature space was of size 50, with one model having a latent size of 300. These latent sizes along with the other selected hyper-parameters were selected because they resulted in the lowest reconstruction loss for the auto-encoder and the best age predictions. The samples were finally divided by a 0.8-0.2 train-test split to ensure the performance of the models on unseen data, which left training samples for each histone mark at approximately 200. This split was made while making sure that biological replicates for a certain sample stayed in the same set, so as to not leak information into the testing set. In comparison to already existing age predictors, this is a surprisingly low number of samples due to the low availability of histone ChIP-seq data. However, this does not keep the models from predicting age with a low mean squared error and median absolute error (MSE, MAE), while at the same time providing deeper insight into which genomic regions specifically go into predicting the age using interpretation techniques.

1.2 Final Model Structure

The final optimized neural network models for each of the histone marks included a denoising auto-encoder followed by a feed-forward neural network. The auto-encoder served as a means to both reduce the dimensionality of the feature space, by having a latent size



(a) Auto-encoder setup

(b) Final Model Structure

Figure 1.1: 1.1a is not made to scale as the real input size was 30321, and respective layer sizes for the auto-encoder were 200 -> 100 -> 50 -> 100 -> 200 -> 30321. The middle layer in 1.1a represents the latent feature space, which is used as input to the feed-forward network after being reduced from input size in 1.1b. The feed-forward network passes the latent features through four Dense layers, before outputting a mean value and it's standard deviation in a Dense layer of size 2.

much smaller than the input size, and also to learn a better underlying distribution of the data by applying Gaussian noise to the inputs. The auto-encoder consisted of an encoder and a decoder, both of which had 3 Dense Layers. To avoid over-fitting, every Dense layer was followed by a Dropout layer barring the last one. Before passing through the Dropout layer, the outputs of each Dense layer were regularized using L1 and L2 activity regularizers and passed through a SELU activation function. The reason for using a SELU activation function was that it doesn't remove negative inputs like RELU, or allow only a certain level of negative values like LeakyRELU, rather it scales the negative values which in our case helps the model learn relationships in the feature space better. Furthermore, an activity regularizer puts a heavier penalty on the output based on the coefficient provided (a hyperparameter).

L1 and L2 regularization penalties are computed as:

$$L1 = l_1 \cdot \sum_{i=0}^{n} |y_{true} - y_{predicted}| \tag{1.1}$$

$$L2 = l_2 \cdot \sum_{i=0}^{n} (y_{true} - y_{predicted})^2 \tag{1.2}$$

where l_1 and l_2 represent the coefficients, y_{true} represents the ground truth label or true age value, and $y_{prediction}$ represents the model's prediction. By penalizing wrong predictions more, the regularizer encourages the model to learn sparse features better, which in turn reduces over-fitting.

The feed-forward neural network started with a Batch Normalization layer which normalizes that batch around its mean value, allowing the model training to be more stable because it reduces the effect of randomness in the parameter initialization and inputs. [10] This is followed by a variable number of hidden layers with both activity and kernel l1, l2 regularization. All the hidden layers are each followed by a SELU activation function, another Batch Normalization, and a dropout. I found that using Batch Normalization again after the activation function improves performance, which is why it was included in the final models. Finally, there is one Dense layer without any regularizer or dropout, but with SELU activation. Naturally, this would be followed by a final Dense layer with size 1, representing the predicted age, but in my case, I outputted a Normal distribution rather than a single value to more accurately capture the prediction, the error associated with it, and the standard deviation of the model. The mean of this normal distribution represented the predicted age, and the standard deviation provided an additional mean to see the variance of the prediction from its actual value. Both the model and the auto-encoder used an Adam optimizer.

1.3 Experimental Setup and Model Development

Before arriving at a final model, I began my experiments with a simple feed-forward neural network with 3 Dense Layers and a RELU activation function with 1 output layer. The initial experiments were meant to make sure that data was viable to predict age accurately. Moreover, before making the feed-forward network more complicated I tried two types of Recurrent Neural Networks, LSTMs, and GRUs, hoping to exploit the sequential nature of the human genome. However, both these methods came with a big complexity overhead which would make the problem of over-fitting hard to deal with, while at the same time taking too long to run, which was not viable given the time constraints of the project. Furthermore, I experimented with different kinds of layers (Batch Normalization, Dense, Dropout), regularizers (l1, l2, both), optimizers, and activation functions, with different hyper-parameters and at different positions before coming up with a final model structure. However, initially, the final model structure only had the feed-forward part of the model. This performed well on both training and validation sets, but the ratio between the number of samples and the number of features would lead to high variance and hence I ended up using an auto-encoder to reduce the dimensionality of the data. The auto-encoder initially did not perform as well, which is why I introduced Gaussian noise into the model which helped it learn the underlying distribution of the data better.

The goal of the auto-encoder was to reconstruct the data with minimal reconstruction error while learning the underlying distribution well enough to fit the information in (latent-size) number of features rather than 30321 features. At the same time, the auto encoder's learning could not be too specific to the training data since that combined with the possible over-fitting of the feed-forward network would produce poor results on unseen data. The goal of the model, on the other hand, was to use the data representation outputted by the auto-encoder to learn features of the data that would allow it to perform well on both the training and testing data.

For this purpose, I developed a 4-fold Cross-validation strategy which allowed me to see the performance of a certain model on unseen data, as it trained the model on 3/4 of the training data, and validated it on the remaining 1/4. The strategy alternated the validation set each time to make sure it covered the entire data. However, even with this strategy, it was too difficult to pick hyper-parameters that would produce the best results on the validation data. For this purpose, after initial tests to narrow down hyper-parameter ranges, I ran a grid-search on all the hyper-parameters involved in both the auto-encoder and the feed-forward neural network for all six histone marks. This included two options for the batch size, 2 options for the number of hidden layers in the feed-forward network, 3 options for the learning rate for both the auto-encoder and the feed-forward network, 4 options for the dropout rate, 3 options for the coefficients for the regularizers in both the auto-encoder and the feed-forward network, 4 options for the latent size for the auto-encoder and 3 options for the Gaussian noise introduced into the inputs of the auto-encoder. With the different number of options, a total of **1728** combinations were tried for each histone mark, and the top 10 models that produced the lowest validation Mean Squared and Median Absolute Error (MSE, MAE) for each histone mark were selected. From this list, I selected 3-5 models that produced the lowest reconstruction error for the auto-encoder along with the lowest MSE for the feed-forward network on the training and validation sets. After selecting the top models, I further fine-tuned their hyper-parameters to extract better performance on the validation set. The final models for each of the histone marks performed significantly better than optimized ElasticNet, which was optimized using the sci-kit learn built-in Cross-validation function. The final CV results are tabulated below.

Histone Marks	NN MSE	ElasticNet MSE	NN MAE	ElasticNet MAE
H3K4me3	**74.55**	141.52	3.5	5.90
H3K27ac	**60.72**	148.35	2.8	6.78
H3K27me3	**68.61**	168.35	2.3	5.77
H3K36me3	**62.83**	144.06	2.5	5.35
H3K4me1	**72.87**	141.75	3.4	4.47
H3K9me3	**83.43**	235.75	3.6	7.74

Table 1.1: Mean 4-fold Cross-validation metrics for the optimized Neural Networks vs. optimized ElasticNet. Bold values in each row represent the lowest MSE. We can see that the optimized NN does better than the optimized ElasticNet for all histone marks.

2

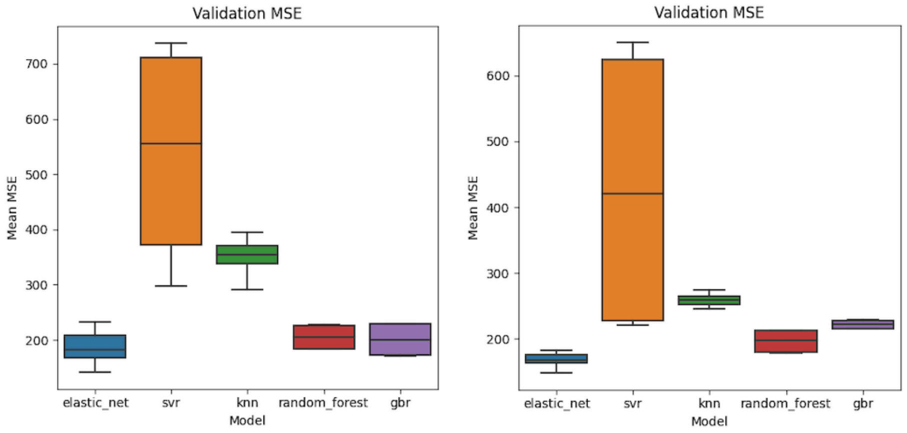
Classical ML Methods predict age accurately using ChIP-Seq data

Since histone mark ChIP-seq data has not been used for age prediction before, if classical machine learning methods can predict age decently using this data, we will know that the data has the potential to predict age accurately. Some of the classical methods that I experimented with for this paper were Support Vector Regression (SVR), K-nearest Neighbors (KNN), Random Forest (RF), Gradient Boosting Regression (GBR), and ElasticNet. Linear regression is known to predict age accurately using DNA methylation data, and since ElasticNet has linear regression at its core, I hypothesized that it should predict age accurately from Histone ChIP-Seq data too. I tabulate 4-fold cross-validation metrics below. These results are achieved by the best hyper-parameters for each of the methods, along with the best input scalars (None, Standard, Robust, Quantile), and the best age-transformers (None, Log-linear (appendix)).

Histone Marks	ElasticNet MSE	SVR MSE	GBR MSE	KNN MSE	RF MSE
H3K4me3	**141.52**	298.23	171.12	290.51	184.33
H3K27ac	**148.35**	220.54	215.77	246.04	179.20
H3K27me3	**168.35**	268.72	256.63	228.66	257.51
H3K36me3	**144.06**	244.7	183.16	209.1	163.67
H3K4me1	**141.75**	224.23	261.86	254.04	248.15
H3K9me3	**235.75**	310.57	293.06	286.06	303.42

Table 2.1: Mean 4-fold Cross-validation metrics for the Best models vs. optimized ElasticNet. Bold values in a row indicate the lowest MSE value for a histone mark. We can see that ElasticNet outperforms all other classical methods.

Figure 2.1 illustrates the combined errors across several different types of scalars and age transformations across 4-folds for each model and histone mark. Some methods and scalar/age combinations do better than others, but overall we see a low variance for all methods but SVR for all histone marks. This low variance indicates that the methods perform consistently across different scalars/age transformations and the 4-folds. Overall, the classical methods perform decently with the best MSE values near the 200 range for almost all histone marks.

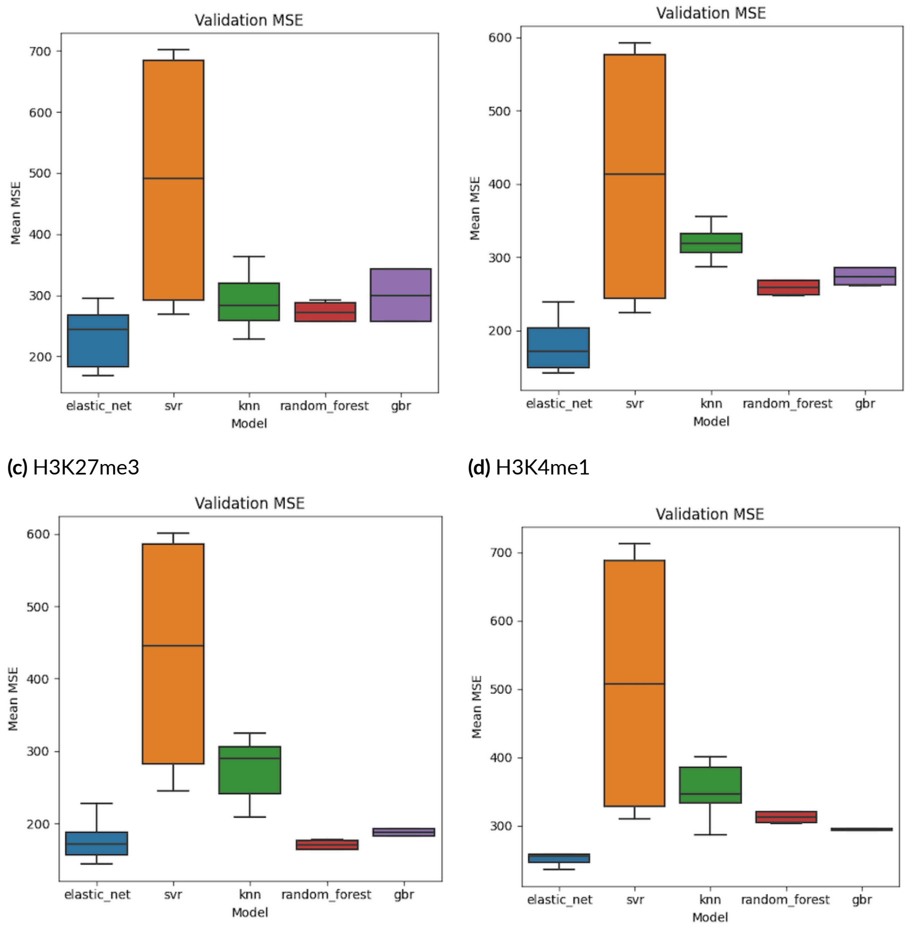


(a) H3K4me3

(b) H3K27ac

Figure 2.1: 4-Cross Validation Results for each Classical ML Method for each histone mark. Each figure shows ElasticNet, SVR, KNN, RF, and GBR on the x-axis in the same order. The y-axis represents the Mean MSE values for each method. A box-plot for each method shows its results with and without a Standard/Robust/Quantile input Scalar, and with or without a Log-linear age transform. The combined results for all four folds are shown for each histone mark.

ElasticNet, as predicted due to its use in Horvath's and Hannum's models [3], [2], performs the best out of the lot for all histone marks, confirming its ability to predict age with high accuracy. The low MSE values for most classical methods confirm that data is viable, and can be used with more complex models such as deep learning networks.



(c) H3K27me3

(d) H3K4me1

(e) H3K36me3

(f) H3K9me3

Figure 2.1: 4-Cross Validation Results for each Classical ML Method for each histone mark. Each figure shows ElasticNet, SVR, KNN, RF, and GBR on the x-axis in the same order. The y-axis represents the Mean MSE values for each method. A box-plot for each method shows its results with and without a Standard/Robust/Quantile input Scalar, and with or without a Log-linear age transform. The combined results for all four folds are shown for each histone mark.

3

Deep Learning Methods improve upon the accuracy of age prediction

3.1 Neural Networks perform better on unseen data than ElasticNet

One of the ways to measure a model's performance on unseen data is by using a train-test split. I used a train-test split of 0.8-0.2, meaning that 80% of the data for each histone mark was used for training and 20% of the data for each histone mark was used for testing, as a way to measure how well the best models for each histone mark perform on data they have never seen before. Tabulated below in Table 3.1 we see the Mean Squared Errors (MSE), the Median Absolute Errors (MAE), and the Pearson Correlation Coefficients (R) of the best neural networks and the most optimal ElasticNet models for each histone mark.

Histone Marks	NN MSE	ElasticNet MSE	NN MAE	ElasticNet MSE	NN R	ElasticNet R
H3K4me3	91.88	189.39	**4.28**	7.04	0.894	0.774
H3K27ac	180.11	170.12	**7.64**	9.64	0.766	0.772
H3K27me3	117.24	143.85	**3.64**	7.37	0.93	0.924
H3K36me3	76.52	161.82	**3.46**	7.47	0.901	0.788
H3K4me1	178.51	195.89	**3.83**	7.00	0.827	0.802
H3K9me3	61.87	277.1	**4.36**	7.42	0.95	0.77

Table 3.1: Testing metrics for the optimized Neural Networks vs. optimized ElasticNet. Test data was unseen throughout training. In bold we see the lowest MAE for each histone mark.

We can see that the best neural networks for each histone mark perform better than ElasticNet on unseen data. We see a lower MAE value on the testing set for each of the histone marks with the biggest difference in H3K36me3, with NN MAE at 3.46 and ElasticNet MAE at a 115.9% higher value of 7.47. Figure 3.1 illustrates the differences between the test MAE values of optimized neural network models vs. an optimized ElasticNet for all histone marks.

Furthermore, we see all histone mark models but H3K27ac achieves a lower MSE than ElasticNet with the biggest difference in H3K9me3 with NN MSE at 61.87 and ElasticNet MSE at a 347.9% higher value of 277.1. On the other hand, ElasticNet had a better MSE than H3K27ac's network by a mere 5%. Finally, most NN's achieved over 0.8 R with 3 networks

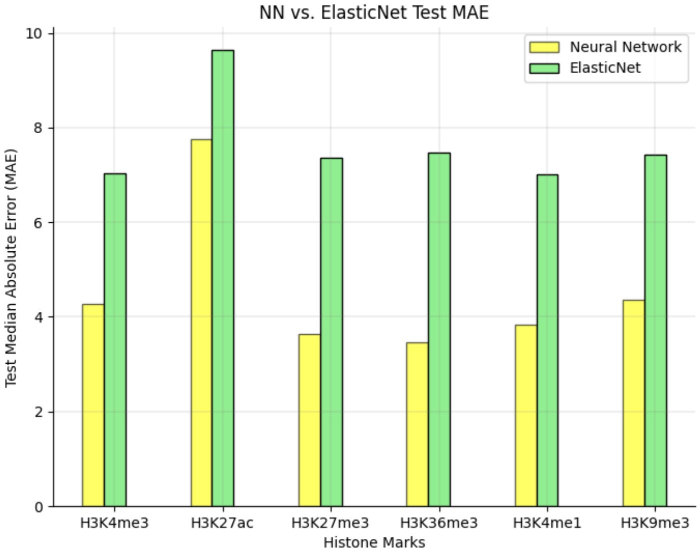


Figure 3.1: Optimized Neural Network vs. Optimized ElasticNet: MAE comparison for each histone Mark. The height of all yellow bars (NN) is less than green bars ((ElasticNet)) illustrating the better performance of neural networks.

achieving more than 0.9. ElasticNet however had only one model with more than 0.9 R and only one model with over 0.8. The biggest difference in the R-value that we saw was for H3K9me3's network which had an R-value of 0.95 as compared to ElasticNet's 0.77 which is 18% lower. Overall, we can see that neural networks do much better than ElasticNet, with lower MSEs, MAEs, and R values.

When it comes to specific age values, Figure 3.2, illustrates a comparison between ElasticNet's and the optimal Neural Network's predicted age values for the testing set. The black line represents the ground truth, or the actual age value, the purple dots represent ElasticNet's predictions, and the blue dots represent the Neural Network's predictions.

We can see in Figure 3.2 that the optimal Neural Networks perform better than ElasticNet on very small ages, which include zero and fetal ages which are encoded by the data based on

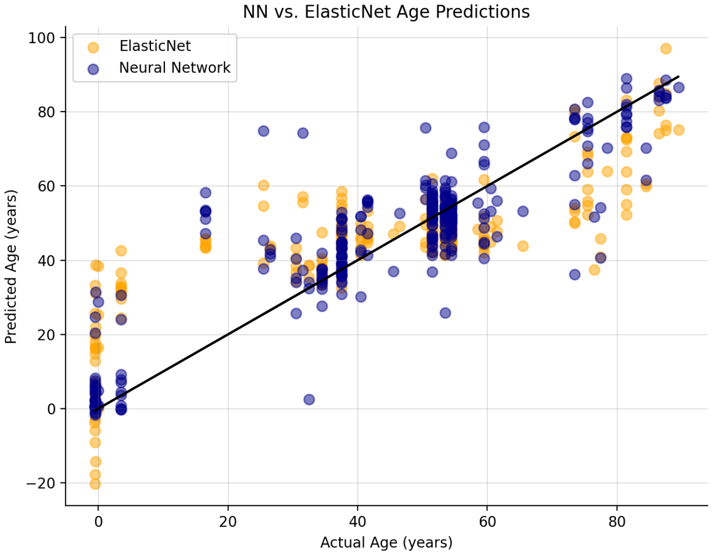


Figure 3.2: Optimized Neural Network vs. Optimized ElasticNet: This figure compares the age predictions of the two models on the test data. ElasticNet is represented by orange, and the neural network is represented by dark blue dots. The black line represents the ground truth or actual age. The very dark blue color towards the middle of the figure indicates multiple samples in the same spot. It is worth noting that ElasticNet predictions are right below the blue dots in this case and hence are not visible, but predictions look similar between 30-and 70 years.

the gestational week. Anything below the 40 weeks is encoded as a negative age, and surprisingly the optimized neural networks can predict it very accurately when ElasticNet mostly undershoots or overshoots. The optimized neural networks also perform much better than ElasticNet for very high age values, this trend is especially visible for ages > 70, where the neural networks can predict age accurately, however, ElasticNet highly undershoots. Overall, we see that the predictions are generally consistent and near the actual age, barring a few outliers.

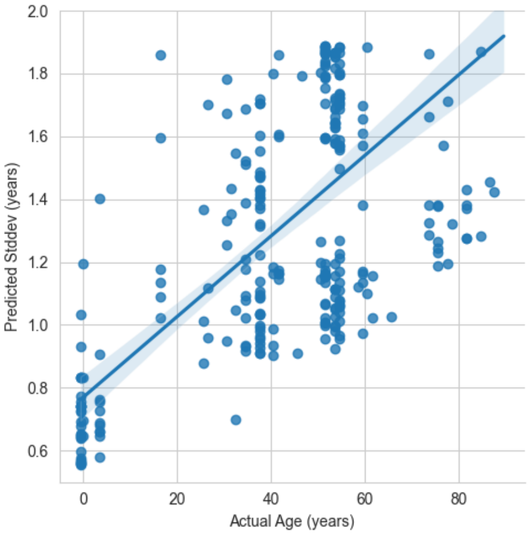


Figure 3.3: Neural Network Prediction Standard Deviation vs. Actual Age in the Test Set. Each data point represents the standard deviation in the predicted age. As age increases, the standard deviation increases, as is shown by the line of best fit. The extended color is to account for errors in calculation, but the trend stays the same.

3.2 Neural Networks Capture Epigenetic Drift

Epigenetic drift is the divergence of the epigenome as a function of age due to stochastic changes in methylation. [11] This effect states that as an individual's age increases, there is a higher chance of external factors impacting their age, which reduces our confidence in our predictions. This is exhibited by the optimized neural networks' age predictions in the test set and is illustrated in Figure 3.3. Although there is some scatter in terms of the standard deviation value, the line of best fit has a positive gradient. This indicates that for test predictions, the standard deviation is lower on lower values of age, and increases as the value of age increases. Hence, the optimized Neural Network can capture the Epigenetic drift effect, further validating our results as they align with known biological phenomena.

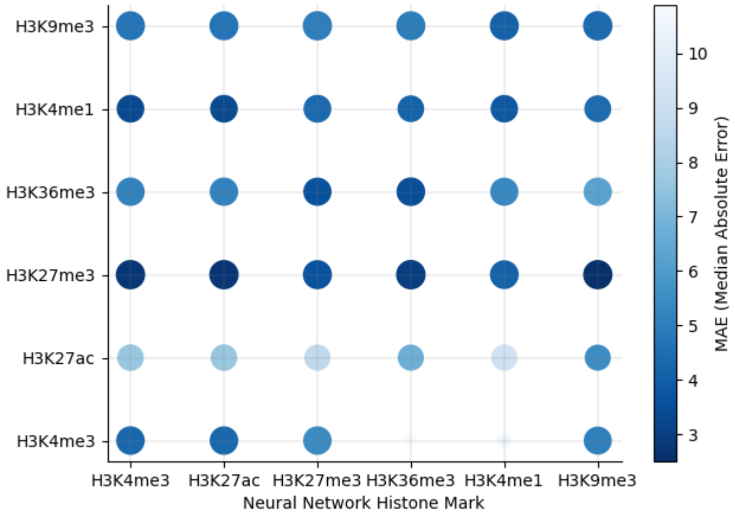


Figure 3.4: Pan-histone mark Age Predictor Comparison. The x-axis represents the histone mark whose neural network was used to train and test the rest of the histone marks' data. The size of the bubbles represents R, the bigger the R-value the larger the bubble, and the color represents the MAE value, the bigger the MAE, the lighter the bubble. For each histone mark, the bubble at HM-HM should be treated as standard, and the color and size of the rest of the bubbles should be compared with it.

3.3 Pan-Histone Mark Age Prediction

Training and testing several models for different histone marks comes at a cost. Separate grid searches need to be run for each mark to arrive at the best model for each, and just this step increases the time taken by 18 days per mark (time taken for a 1728 hyperparameter search to run on Oscar). Hence, if there exists a possibility to use a single neural network to predict age accurately for multiple histone marks, that can be a big step towards improving the scalability of neural networks to predict age using histone mark ChIP-Seq data.

Figure 3.4 shows a bubble plot showing the performance of each of the six neural networks on other histone mark data. Networks optimized for each of the histone marks on the x-axis are used to predict age on data from other histone marks. The results are shown by using bubble color and size in the plot. A darker color indicates a higher MAE value, and a smaller bubble indicates a lower R-value. Our aim, therefore, is to determine whether certain networks, can make accurate predictions for more than one histone marks' data. For this, we can look at each row, and compare the bubbles for that histone mark to the bubble at its intersection with itself. H3K27me3 for example has a low MAE at its intersection with itself, but darker bubbles at the H3K4me3, H3K27ac, and H3K9me3 indicate that predictions from those models have a lower MAE value than H3K27me3's own model. A similar trend can be seen for H3K27ac. H3K36me3 and H3K9me3's networks predict a lower MAE on H3k27ac's data than its own optimized NN. One reason for H3K9me3's model to predict age accurately for H3K27me3's data could be that they are both repressors. [12] H3K4me3 and H3K27ac's networks also predict age accurately on H3K27me3's data and one reason for that may be that they're all found in promoter regions. [12] Furthermore, H3K36me3's network predicting H3k27ac's data with a lower MAE is also possible because they are both activators. However, any of these conclusions cannot be made for certain, since there can be other reasons as to why a network would perform better on the test set, such as data distribution between the train-test set.

What we can see though, is that all networks perform consistently well for H3K36me3, H3K4me1, and H3K9me3. H3K27ac, and HK36me3. Overall, the presence of equally shaped bubbles (indicating similar R values), and similarly colored bubbles in each row, indicate that it is possible to predict ages accurately for more than one histone mark using a single model.

4

Interpretation

One key advantage of using ElasticNet over neural networks is that they are easier to understand as compared to neural networks. However, interpretation methods allow us to dive deeper into the inner workings of neural networks and examine which features impact predictions the most. One way of doing so is by using SHAP values. SHAP (SHapley Additive exPlanations) is a game-theoretic approach to explaining the output of any machine learning model. [8] Its goal is to measure the impact of a feature on the prediction of a model. It assigns SHAP values to each of the different features in a model, with positive SHAP values indicating a positive impact on the prediction (pushes age up), and a negative SHAP value indicating otherwise (pushes age down). In our case, it allows us to identify genomic regions that are important to aging, along with identifying complex relationships between different genomic regions.

In order to obtain SHAP values for the optimized neural networks, I had to alter the final model structure. First, I had to remove the auto-encoder section, because SHAP functions are unable to take in multiple models as inputs, and channel inputs through them. Second, I had to alter the final layer of the model to return a single value as the age prediction rather than a distribution, since SHAP functions do not accept models with multiple/distribution outputs.

4.1 Optimized NN's capture age-related genomic regions

I obtained the best SHAP value for each of the six neural networks optimized for each histone mark. For H3K4me3's Model, the genomic region **chr17:7300001-7400000** had the highest SHAP value. This means that the **chr17:7300001-7400000** region had the highest impact on the age-predicted by the H3K4me3 model. This region contains genes such as eIF5A, GPS2, and TNK1 that are known to impact aging in humans.

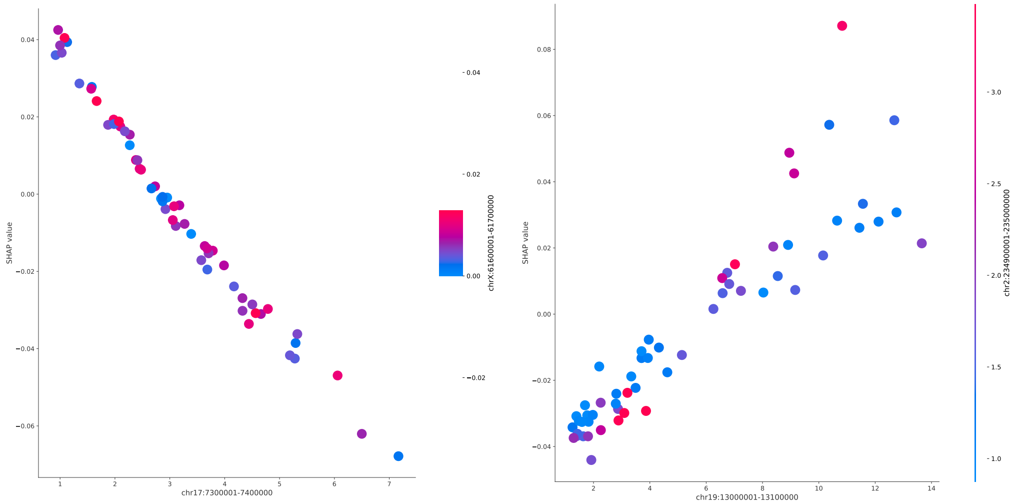
A reduction in eIF5A hypusination in several genetic regimes affects brain mitochondrial respiration resembling age-typical mitochondrial decay. [13] Furthermore, Mitochondrial function decline during brain aging is suspected to play a key role in age-induced cognitive decline and neurodegeneration. [13] This means that eIF5A is a gene that plays an important role in determining an individual's epigenetic age. GPS2, on the other hand, is known to be an independent predictor of mortality for individuals with ages > 85. [14] Moreover, a study conducted in adults with and without Alzheimer's disease with ages from 50-80+ showed that TNK1-A/A genotype frequency increased with higher age in individuals without any cognitive disabilities, and decreased with higher age in individuals. [15]

chr19:13000001-13100000 was the region with the highest SHAP value for H3K27ac's network. NFIX, a gene found in the region, plays a role in regulating progenitor cell biology within the embryonic and postnatal cerebellum, and an ongoing role within multiple neuronal and glial populations within the adult cerebellum. [16] Similarly, other networks captured regions such as **chr12:53900001-54000000** with genes HOXC13 and HOXC12, [17], [18] which are known to impact aging.

Overall we can see that the neural networks pick on the regions in the human genome that we know to be related to aging, hence, validating our claim that neural networks not only predict age accurately but capture the right features to do so. This can have very promising prospects, as regions picked up by neural networks that are currently under-researched in terms of aging might also turn out to be important predictors of human age.

4.2 Optimized NN's capture complex genomic region-region interactions

Figure 4.1 shows scatter plots for each optimized networks' most important features, plotted with the corresponding region that the feature interacts with most. These plots show the relationship between a genomic region and age prediction, along with how it's relationship with another region effects its prediction. As observed, most of the regions have a linear relationship to the output. This explains how ElasticNet, even with high bias, can predict age ac-

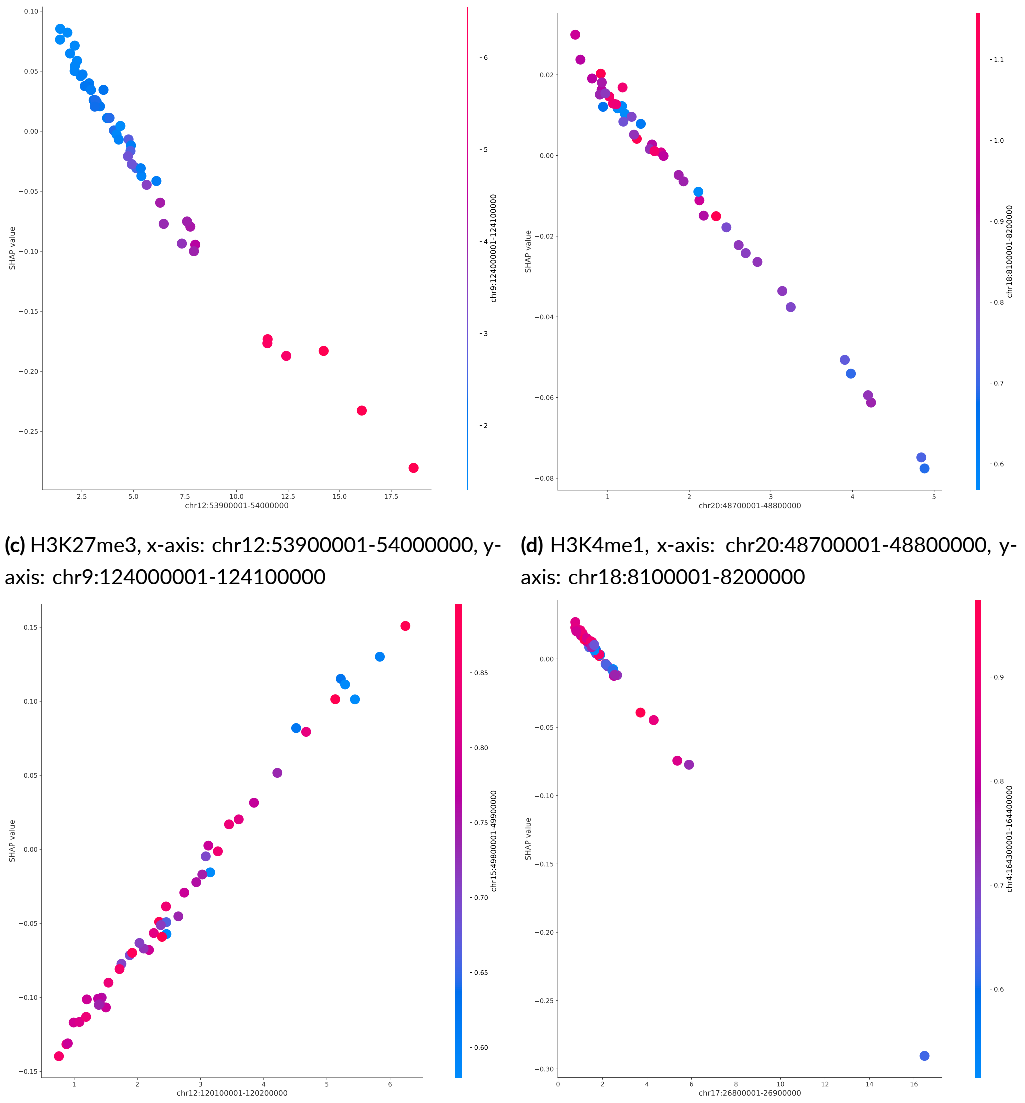


(a) H3K4me3, x-axis: chr17:7300001-7400000, y-axis: chrX:61600001-61700000

(b) H3K27ac, x-axis: chr19:13000001-13100000, y-axis: chr2:234900001-235000000

Figure 4.1: SHAP value graph showing interactions between the most important feature and the feature it interacts with the most for each histone mark. The most important feature is on the x-axis, and the feature it interacts with the most is on the y-axis. Values on the each axis represent the SHAP value for the corresponding region. The color bar shows interaction, with more pink value showing more region-region interaction

curately using the data. However, at least one histone mark's figure shows evidence of higher complexity. H3K27ac's region does not have a linear relationship with its most interacting feature. Although ElasticNet won't be able to capture such a relationship, neural networks have the ability to model such relationships, which in turn, improves their age prediction accuracy. Furthermore, we can see from figure 4.1b, that **chr2:234900001-235000000**'s SHAP values affect the curvature of the plot, which means that interactions between **chr19:-13000001-13100000**, and **chr2:234900001-235000000** impact H3K27ac's networks' predictions. Overall, these SHAP value plots show that it is possible for the neural networks to predict age more accurately than ElasticNet because of the fact that they can capture complex interactions between genomic regions better.



(c) H3K27me3, x-axis: chr12:53900001-54000000, y-axis: chr9:124000001-124100000

(d) H3K4me1, x-axis: chr20:48700001-48800000, y-axis: chr18:8100001-8200000

(e) H3K36me3, x-axis: chr12:120100001-120200000, y-axis: chr15:49800001-49900000

(f) H3K9me3, x-axis: chr17:26800001-26900000, y-axis: chr4:164300001-164400000

Figure 4.1: SHAP value graph showing interactions between the most important feature and the feature it interacts with the most for each histone mark. The most important feature is on the x-axis, and the feature it interacts with the most is on the y-axis. Values on the each axis represent the SHAP value for the corresponding region. The color bar shows interaction, with more pink value showing more region-region interaction

5

Conclusion

ElasticNet, and linear models in general, have been very successful at predicting age accurately for some time now. Recent studies have shown that neural networks can do the same while providing valuable insights into the data used. However, all the existing experiments have been conducted using DNA methylation data, which is good at predicting age, but its interpretability is limited by the assay to probe DNA methylation which only covers part of the human genome. Histone ChIP-seq data, on the other hand, covers the entire genome and opens up far more interpretation possibilities than DNA methylation data while at the same time predicting age very accurately.

This thesis set out with three major goals. First, validate Histone ChIP-Seq data's viability in an age prediction task by using existing methods such as ElasticNet to predict age accurately. It accomplishes that as shown in Chapter 2, with ElasticNet predicting age with decent accuracy using Histone ChIP-Seq data. Second, show that deep learning methods predict age with higher accuracy as compared to existing methods. It accomplishes that as in Chapter 3 we see that neural networks for each histone mark achieve a lower Median Absolute Error (MAE) than ElasticNet. Third, show that predicting age using Histone ChIP-Seq data opens up a wide array of interpretation possibilities. It achieves this by showing in Chapter 4 that neural networks captured features (genomic regions) that are known to impact aging in humans and capture complex genomic region-region interactions. The interpretation in this thesis, however, only scratches the surface of the amount of insight that is possible to be gained from the features neural networks use to predict age. Future work in regards to the thesis will include furthering the biological insights gained from different interpretation methods such as DeepPINK. Moreover, I will also use ElasticNet and the optimized neural networks to predict age from Gene Expression Omnibus data, as it is from an entirely new dataset. Performance on the GEO data can determine the generalizability of the optimized neural networks.

Overall, I have shown that epi-genetic age can be predicted with high accuracy using neural networks and Histone ChIP-Seq data while broadening the range of interpretation that can be performed.

Bibliography

[1] Lucas Paulo de Lima Camillo and Robert B A Quinlan. “A ride through the epigenetic landscape: aging reversal by reprogramming”. en. In: *GeroScience* 43.2 (Apr. 2021), pp. 463–485.

[2] Gregory Hannum et al. “Genome-wide methylation profiles reveal quantitative views of human aging rates”. en. In: *Mol. Cell* 49.2 (Jan. 2013), pp. 359–367.

[3] Steve Horvath. “DNA methylation age of human tissues and cell types”. en. In: *Genome Biology* 14.10 (2013), R115.

[4] Lucas Paulo de Lima Camillo, Louis R Lapierre, and Ritambhara Singh. “AltumAge: A pan-tissue DNA-methylation epigenetic clock based on deep learning”. en. June 2021.

[5] David A Sinclair and Matthew D Laplante. *Lifespan: The Revolutionary Science of Why We Age–and Why We Don’t Have To. First Atria Books hardcover edition*. New York: Atria Books, 2019.

[6] Zuyun Liu et al. “Underlying features of epigenetic aging clocks in vivo and in vitro”. en. In: *Aging Cell* 19.10 (Oct. 2020).

[7] Sun-Ju Yi and Kyunghwan Kim. “New insights into the role of histone changes in aging”. en. In: *Int. J. Mol. Sci.* 21.21 (Nov. 2020), p. 8241.

[8] Scott M Lundberg and Su-In Lee. “A unified approach to interpreting model predictions”. In: *Advances in Neural Information Processing Systems* 30 (2017).

[9] The ENCODE Project Consortium. “An integrated encyclopedia of DNA elements in the human genome”. en. In: *Nature* 489.7414 (Sept. 2012), pp. 57–74.

[10] Sergey Ioffe and Christian Szegedy. “Batch Normalization: Accelerating deep network training by reducing internal covariate shift”. In: (Feb. 2015). arXiv: 1502.03167 [cs.LG].

[11] Sonia Shah et al. “Genetic and environmental exposures constrain epigenetic drift over the human life course”. en. In: *Genome Res.* 24.11 (Nov. 2014), pp. 1725–1733.

[12] https://www.abcam.com/epigenetics/histone-modifications.. Accessed: 2022-4-14.

[13] Yongtian Liang et al. “eIF5A hypusination, boosted by dietary spermidine, protects from premature brain aging and mitochondrial dysfunction”. en. In: *Cell Rep.* 35.2 (Apr. 2021), p. 108941.

[14] Michela Zanetti et al. “Predictors of short- and long-term mortality among acutely admitted older patients: role of inflammation and frailty”. en. In: *Aging Clin. Exp. Res.* 34.2 (Feb. 2022), pp. 409–418.

[15] Davide Seripa et al. “Role of CLU, PICALM, and TNK1 genotypes in aging with and without Alzheimer’s disease”. en. In: *Mol. Neurobiol.* (June 2017).

[16] James Fraser et al. “Cell-type-specific expression of NFIX in the developing and adult cerebellum”. en. In: *Brain Struct. Funct.* 222.5 (July 2017), pp. 2251–2270.

[17] Kamil C Kural et al. “Pathways of aging: comparative analysis of gene signatures in replicative senescence and stress induced premature senescence”. en. In: *BMC Genomics* 17.S14 (Dec. 2016).

[18] Jasmina-Ziva Rozman et al. “DNA methylation and hydroxymethylation profile of CD34+-enriched cell products intended for autologous CD34+ cell transplantation”. en. In: *DNA Cell Biol.* 36.9 (Sept. 2017), pp. 737–746.