Transformer encoder–based DNA-methylation epigenetic clock

Alexander Sharipov, 20190793, Yeskendir Assankul, 20200808, Nargiz Askarbekkyzy 20190761

https://github.com/Alex-T-Sharipov/Epigenetic_Clocks

Abstract

DNA methylation is one of the main hallmarks of aging and is believed to influence the development of age-related diseases. Here, we improve the accuracy of age inference from DNA methylation by utilizing multilayer transformer encoder with learnable down sampling. Our model achieves median absolute error below 0.5 years and outperforms the penalized linear regression and multilayer perceptron in predicting the age based on the DNA methylation values. Further, we visualize the self-attention scores calculated by the transformer architecture, revealing potential interactions among the CpG methylation sites.

1. Introduction

1.1 Motivation

Next-generation sequencing has revolutionized medicine by generating a vast amount of genomic information, including DNA methylation data. DNA methylation is a crucial epigenetic biomarker of aging[1], prompting researchers to develop "epigenetic clocks," which predict age based on DNA methylation data. Quantitative age prediction models are invaluable in the field of aging research because they allow us to evaluate the effect of medical treatments on the aging process. In addition, these models can guide experimental treatments: those DNA methylation sites that contributed the most to the model prediction are promising targets for epigenetic interventions that aim to alleviate age-related disorders.

1.2 Problem setting

Existing epigenetic clocks were traditionally based on linear regression [Horvath, 3], with the most recent models utilizing multilayer perceptron architecture [Camillo et al., 2]. However, neither of these techniques is currently considered state-of-the-art in machine learning. This project aims to improve age prediction with a multilayer transformer encoder [Vaswani et al., 4] deep learning model. Transformer encoder architecture is exceptionally well suited for the chosen dataset because it computes self-attention on the input. It can learn to relate any two CpG sites. Our dataset consists of the methylation percentage (i.e., the beta value) at each of the 18294 CpG genomic locations. Each of these locations can potentially interact with others in complex regulatory networks. Self-attention matrix allows us to look "under the hood" of the model and interpret the transformer-perceived mutual relevance of genomic locations. We visualize the self-attention matrix, which acts as a predictor of the pairwise interactions between the CpG sites.

2. Methods

2.1. Dataset collection and normalization.

The complete datasets used by preceding Horvath and Camillo et al. were unavailable at the time of this project. Therefore, we created a new dataset consisting of 48 Gene Expression Omnibus (GEO) samples used by Horvath and 58 samples used by Camillo et. al.

GEO is a database repository of genomic and epigenomic data. The records in the database are highly heterogenous since they come from different research groups and diverse experimental settings. Most of the collected samples had records for different CpG locations (i.e., different sets of columns). To combine these records into one dataset, we utilized R language to find the union of all possible CpG locations (i.e., all possible columns). Then we filled in the missing values with the mean of the column.

As suggested by Horvath, we discarded the DNA methylation values of human embryonic stem cells and natal tissue since those tend to be "outliers" that interfere with the learning process. This yielded a data set with 5560 human records, each with the DNA methylation values at 18294 genomic locations. Finally, we applied BMIQ normalization, as was done by Horvath and Camillo et al. BMIQ, or beta-mixture quantile normalization, was developed for correcting the hardware design bias in Illumina Infinium 450k DNA sequencing platform. Next, we created several additional datasets with the number of columns reduced by the Principal Component Analysis to avoid overfitting. Namely, we created datasets with 128, 256, 512, and 1024 columns. All the collected datasets were split into training, validation, and test subsets using an 80:10:10 ratio.

2.2. Penalized Linear Regression.

First, we set out to reproduce the results from Horvath by training Elastic Net, a penalized linear regression implemented in the glmnet package in R. The optimization objective of Elastic Net is as follows:

$$\min_\beta \frac{1}{N} \sum_{i=1}^{N} w_i l(\beta^T x_i) + \lambda \left[(1-\alpha)\|\beta\|_2^2/2 + \alpha\|\beta\|_1\right]$$

The negative log-likelihood "l" is minimized over the grid of multiple possible regularization weights "lambda." The penalty varies between lasso or L1 regularization (α = 1) and ridge or L2 regularization (α = 0). We used the α value 0.5, which is the middle ground between these two regularization methods. Elastic Net, therefore, learns to select only the most relevant CpG sites (Figure 2) and applies the weights β to infer the age.

2.3. Multilayer perceptron.

Next, we reproduced the results from Camillo et al. by training a multilayer perceptron. Each of the five hidden layers had 32 neurons, followed by the Gaussian dropout layer, batch normalization layer, and SeLU (scaled exponential linear unit) activation. The perceptron was trained for 500 epochs using the Adam optimizer with a 0.002 learning rate. However, no improvement was observed after 300 epochs (Figure 1).

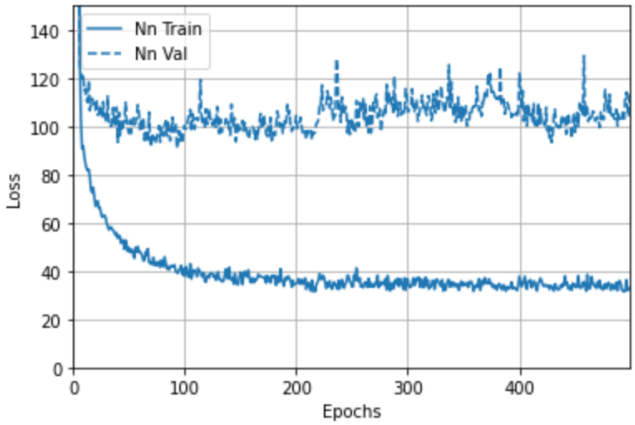

Figure 1: *Training and validation loss of multilayer perceptron*

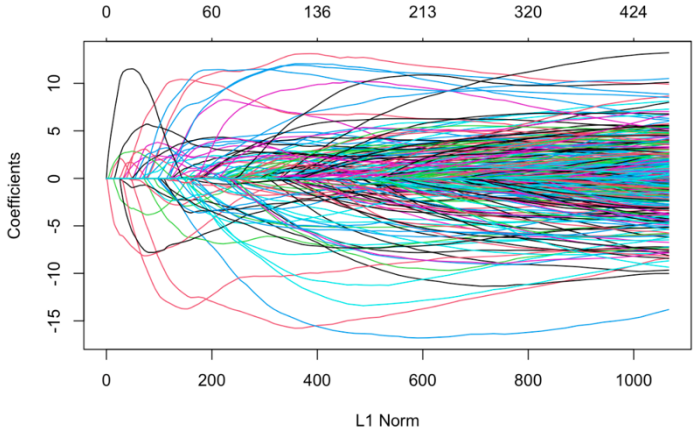

Figure 2: *The coefficients of the linear regression and the L1 norm with varying lambda*

2.4. Transformer encoder

The main obstacle to applying a transformer encoder to our data is the extremely high number of CpG sites (18294). The RAM requirements of conventional systems cannot accommodate the transformer model when attempting to process the input tensor. To resolve this issue, we came up with two main strategies. The first strategy involved applying the Principal Component Analysis to obtain the dataset with fewer features. The second strategy entailed a "learnable down sampling," which was implemented with a linear layer that projects the 18294 – dimensional input into a lower, 256 - dimensional space.

We then designed two transformer encoder architectures with different implementations of self-attention. Overall, both architectures follow a similar overarching design which involves stacking N transformer encoder layers, as is shown in Figures 2 and 3. From now on, we refer to these architectures as transformer encoder A and transformer encoder B.

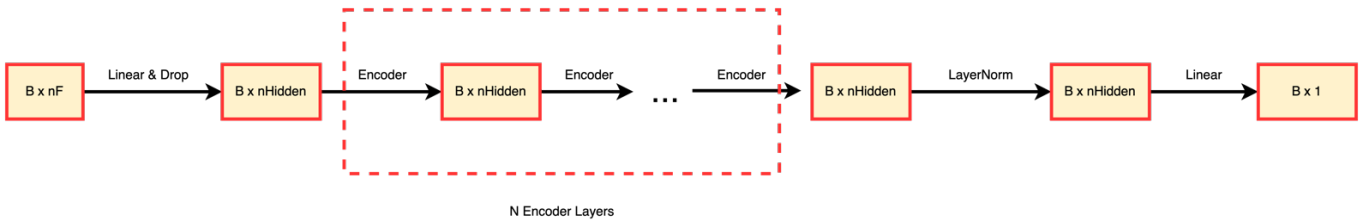

Figure 2: *The overview of the implemented multilayer transformer encoder. First, a linear layer may be applied to reduce the number of feature dimensions. This is followed by the dropout layer, and the result is fed into the first out of*

the N encoder layers. Finally, the result of the encoder layers is normalized, and the final linear layer projects it into just one dimension: the predicted age of the human subject.

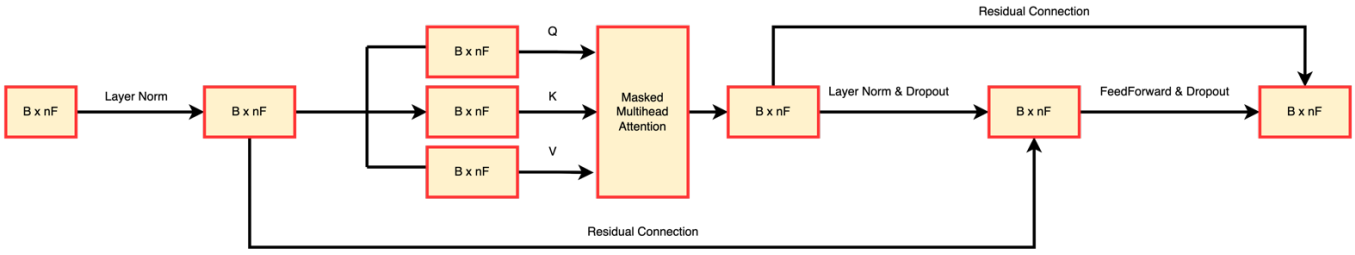

Figure 3: *The architecture of the encoder layer. The input is normalized and then projected into the Query Q, Key K, and Value V, which are then fed to the multi-head attention. Next, the result is normalized and passed through the dropout before adding the residual connection. Finally, the sum is passed through the feedforward network and the dropout layer before adding another residual.*

2.4.1 Transformer Encoder A.

The first implementation of the transformer encoder involves calculating self-attention over all CpG sites of the input. (Figure 4). This architecture was tested with 1 head and 128 hidden dimensions corresponding to the PCA-processed 128-dimensional dataset. In addition, we evaluated this architecture with 4 heads and 4*128 hidden dimensions corresponding to up sampled data with 128 columns (Figure 5).

Finally, we trained this architecture with 1 attention head, 256 hidden dimensions, and learnable feature down-sampling performed by the fully connected layer. In addition, we applied the “all CpG attention” architecture to the original 18294-dimensional dataset after adding a fully connected layer that projects the data points down to the 256-dimensional space.

2.4.1 Transformer Encoder B.

The second implementation involves distributing the CpG sites into 16 unique partitions, one for each attention head (Figure 6). The self-attention is then calculated only between the pairs of CpG sites in the same partition. We evaluated this architecture with 128, 256, and 512 hidden dimensions corresponding to the number of columns in datasets with reduced dimensions. The direct correspondence of the hidden dimension to the number of columns was chosen to increase the interpretability of the attention.

Further, we applied the "partitioned attention" architecture to the original 18294-dimensional dataset after adding a fully connected layer that projects the data points down to the 256-dimensional space.

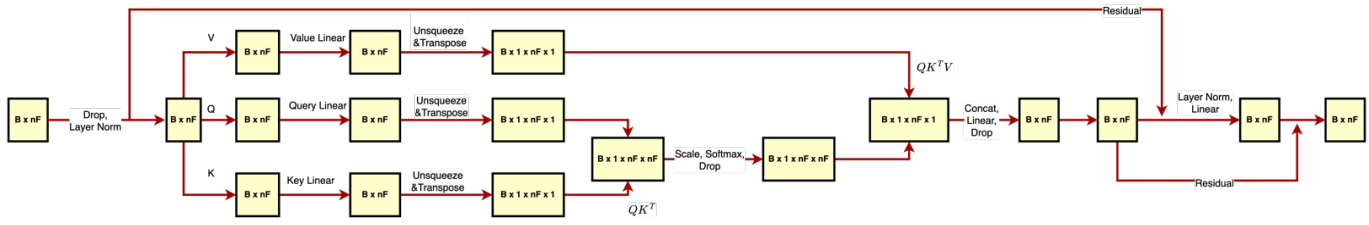

Figure 4: *1-headed self-attention across all CpG sites (Transformer A). The self-attention matrix QxK^T has 1 x nF x nF dimensions, meaning that it has the attention score for each CpG pair.*

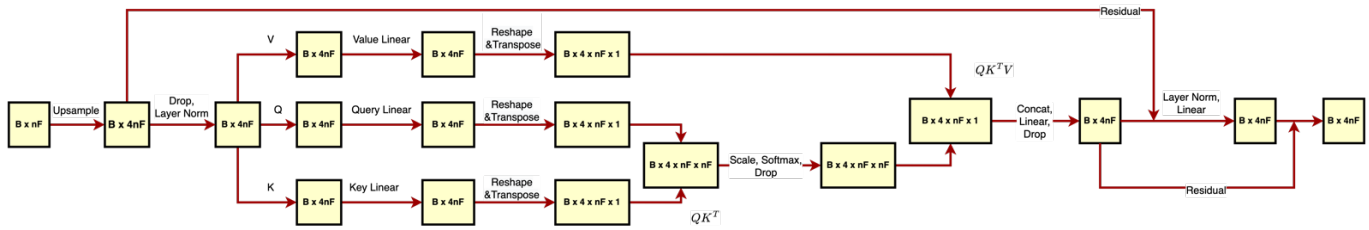

Figure 5: *4-headed self-attention across all CpG sites (Transformer A). The self-attention matrix QxK^T has 4 x nF x nF dimensions, meaning that it has 4 attention scores for each CpG pair*

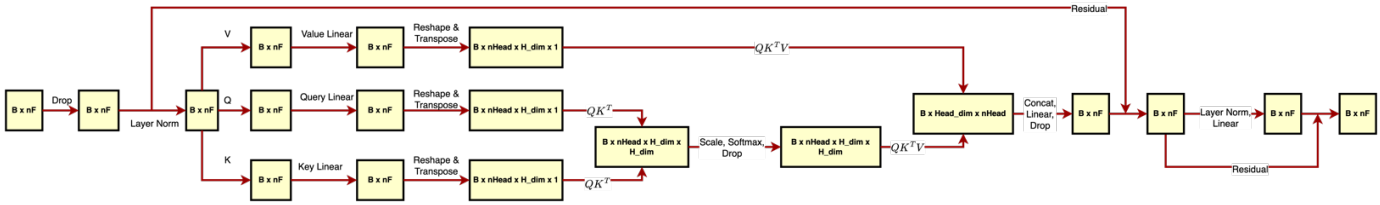

Figure 6: *partitioned attention (Transformer B). CpG sites are distributed into nHead unique partitions. Self-attention is calculated only between the pairs of CpG sites in the same partition.*

Transformer architectures were trained for 500 or 1000 epochs using Adam optimizer with a 0.0001 learning rate.

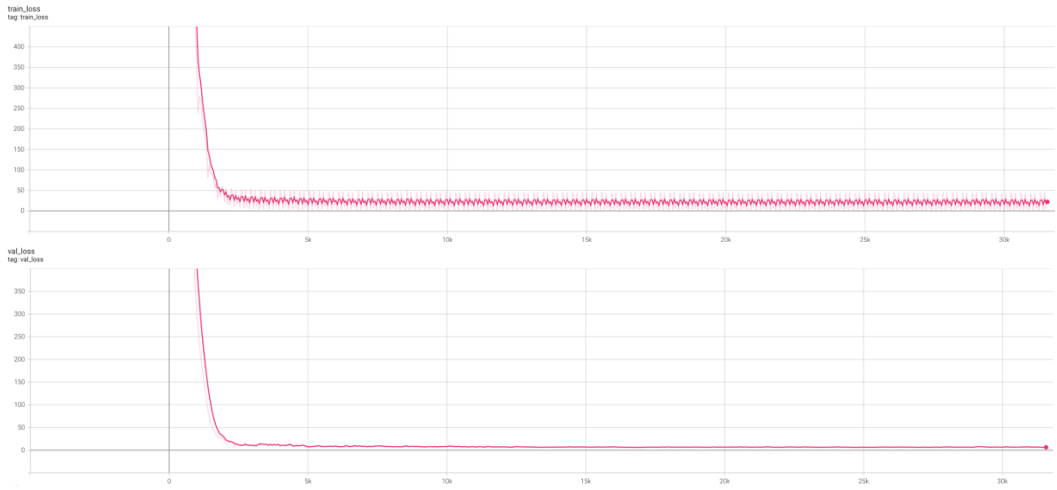

Figure 7: *The training and validation loss of Transformer A with 128 hidden dimensions, 1 attention head, and learnable down sampling.*

3. Results.

Following training, the median absolute error of each model was calculated on the test dataset.

Architecture	Median absolute error, years	Architecture	Median absolute error, years
Linear regression	1.0263	Transformer Encoder B (512 hidden dimensions, PCA)	15.1408
Transformer Encoder B (256 hidden dimensions, learnable down sampling)	3.0730	Transformer Encoder A (1 head, 128 hidden dimensions, PCA)	15.2431
Multilayer perceptron	3.4164	Transformer Encoder A (4 heads, 512 hidden dimensions, PCA)	16.0375
Transformer Encoder B (128 hidden dimensions, PCA)	15.2848	Transformer Encoder A (1 head, 256 hidden dimensions, learnable down sampling)	0.4895
Transformer Encoder B (256 hidden dimensions, PCA)	15.1529		

The best performer by far was Transformer Encoder A with 1 head and 256 hidden dimensions. It achieved a stellar 0.4895 median absolute error, beating the second best model by more than 2-fold.

The second-best model was penalized linear regression. Linear regression achieved a median error of just over one year. Linear regression identified 460 genomic locations which have high relevance to the prediction of age. Curiously, only 4 of these sites happen to coincide with the results of Horvath. This is likely due to the difference of our datasets.

The third-best performer was Transformer Encoder B with learnable down sampling, which achieved a median error of about three years and managed to narrowly outperform the linear regressor with 3.4 median error.

However, all the other transformer architectures which relied on the principal component analysis for the dimensionality reduction performed significantly worse, with a median absolute error of 15 years or more. This result demonstrates that the CpG methylation data is highly nonlinear, explaining the loss of relevant information, which results in poor performance of the models. The number of hidden dimensions did not have a very pronounced effect on the model's performance; however, the increase in the number of hidden dimensions led to a slight but consistent improvement in the performance of Transformer Encoder B.

In addition to evaluating the accuracy of the trained transformers, we extracted the scaled and "soft maxed" self-attention, or Q x K^T, from the first encoder layer of each model. Visualizing attention maps (Figures 6, 7) is essential for model interpretability since it reveals the transformer-perceived mutual relevance of the CpG sites. Furthermore, this has a solid potential to accelerate the rate of biomedical discovery by demonstrating the potential coregulation of the genes proximal to these CpG sites.

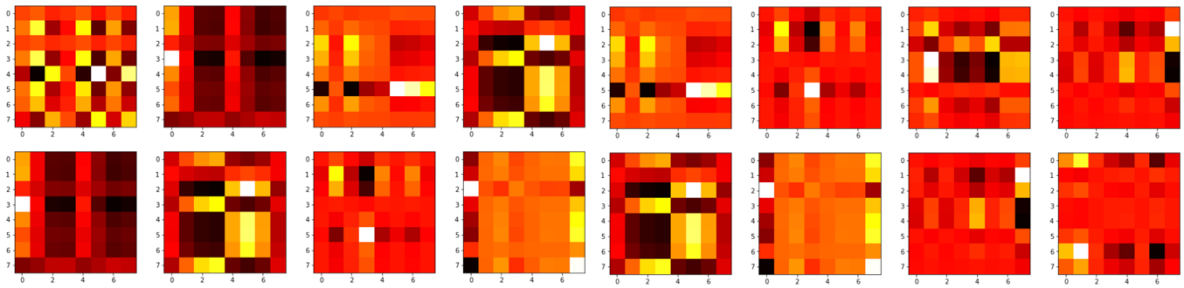

Figure 6: *Visualization of the attention map (QxK^T after applying scaling and soft max) produced by the first encoder layer of Transformer B with 128 hidden dimensions. Each of the 128 hidden dimensions is partitioned into 16 heads, each having 16 CpG site representations. The attention score is calculated for each pair of CpG sites within the partition and can be interpreted as the relevance of the two sites to each other.*

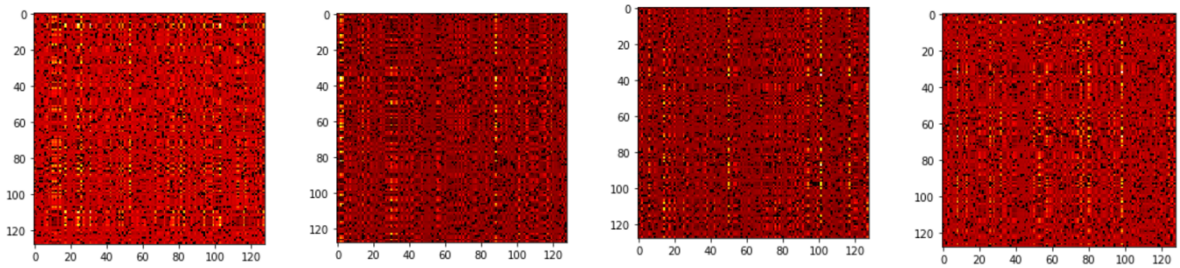

Figure 7: *Visualization of the attention map (QxK^T after applying scaling and soft max) produced by the first encoder layer of Transformer A with 4*128 hidden dimensions and 4 heads. Each head receives 128 hidden dimensions, the same as the input dimension of the data. In other words, each head can calculate the attention score for each pair of CpG site representations in the input.*

4. Discussion.

Transformer Encoder A with 1 head, 256 hidden dimensions and learnable down sampling achieved by far the lowest median error, which is almost half that of the next closest model, which is the linear regressor. This can be attributed to its central feature, the self-attention on the input. Transformer self-attention lends itself incredibly well to learning the underlying structure of the training dataset, which consists of the methylation percentage of a large set of genomic locations. These sites are not independent and are likely to interact through gene regulatory networks. Even the "linearly distant" CpG sites may be spatially quite close in the complex nonlinear chromatin structure. This means that the model should consider the interaction even between distant sites. Transformers are an excellent fit for this dataset because they don't make any assumptions about the temporal or spatial relationships across the data and can therefore model the interaction among any pair of the CpG sites. In addition, our best-performing transformer encoder architecture utilizes multi-head self-attention, which defines "saliency weights" between each element of the set and the rest of the elements. This means that the encoder's self-attention will be able to learn multiple relationships between the methylation states

of the provided genomic locations. Multi-head self-attention allows the model to learn multiple different rules for coregulation among the CpG sites, making it more expressive. These benefits of attention were confirmed by our results, which show that the transformer outperforms the "attention-less" multilayer perceptron.
Interestingly, only the transformer encoder models with learnable down sampling achieved strong performance. Meanwhile, the models that relied on the principal component analysis for the data reduction were all severely limited, with the median absolute error exceeding the 15-year threshold. This finding demonstrates that PCA is ineffective at reducing the dimensionality of the DNA methylation data, likely due to its nonlinear nature. Ultimately, the application of PCA significantly degraded the dataset quality, leading to a drastic drop in performance.

5. Contributions.

Yeskendir Assankul (1) implemented the multilayer perceptron, (2) contributed to the presentation, (3) contributed to the development of dataset loaders, (4) visualized training and validation loss, (5) trained the models, saved the model files, (6) contributed to the final report, (7) contributed to the transformer architecture development
Nargiz Askarbekkyzy (1) implemented Transformer Encoder architecture, (2) implemented the training loop and dataset loaders, (3) designed the figures used in this report, (4) contributed to the presentation, (5) visualized training and validation loss, (6) trained the models, saved the model files, (7) contributed to the final report
Alexander Sharipov (1) collected and normalized the dataset using R, (2) implemented and trained penalized linear regression, (3) suggested and implemented multi-head attention architectures in Transformer Encoders A and B, (4) extracted and visualized the attention maps, (4) contributed to debugging of transformer models, (5) visualized training and validation loss, (6) contributed to the final report, (7) contributed to the presentation, and (8) trained the models, saved the model files, and collected the results into a GitHub repository

6. Data and code availability

All of the pre-trained models can be found in the following public cloud storage link:
https://drive.google.com/drive/folders/11JlLkkm6oNcmSivNUq0Nyb25hPNxg45M?usp=share_link

The Python Notebooks with the aforementioned models along with the R markdown file with dataset preprocessing code and the linear regression can be found in the following public github repository:
https://github.com/Alex-T-Sharipov/Epigenetic_Clocks

7. References.

1. Johnson, A. A., Akman, K., Calimport, S. R., Wuttke, D., Stolzing, A., & de Magalhães, J. P. (2012). The role of DNA methylation in aging, rejuvenation, and age-related disease. Rejuvenation Research, 15(5), 483–494. https://doi.org/10.1089/rej.2012.1324

2. de Lima Camillo, L.P., Lapierre, L.R. & Singh, R. A pan-tissue DNA-methylation epigenetic clock based on deep learning. npj Aging 8, 4 (2022). https://doi.org/10.1038/s41514-022-00085-y

3. Horvath, S. DNA methylation age of human tissues and cell types. Genome Biol 14, 3156 (2013). https://doi.org/10.1186/gb-2013-14-10-r115

4. arXiv:1706.03762

5 .arXiv:1901.02860

6. https://github.com/rsinghlab/AltumAge

7. arXiv:1706.03762