Open access • Posted Content • DOI:10.1101/2021.06.01.446559

Improving epigenetic clock performance and interpretation with deep learning — Source link

Lucas Paulo de Lima Camillo, Louis R. Lapierre, Ritambhara Singh

Institutions: Brown University

Published on: 31 Aug 2021 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: DNA methylation and CpG site

AltumAge: A Pan-Tissue DNA-Methylation Epigenetic Clock Based on Deep Learning

Lucas Paulo de Lima Camillo^{1,*}**, Louis R Lapierre**^{2}**, and Ritambhara Singh**^{3,4,*}

^{1}Department of Chemistry, Brown University, USA

^{2}Department of Molecular Biology, Cell Biology and Biochemistry, Brown University, USA

^{3}Department of Computer Science, Brown University, USA

^{4}Center for Computational Molecular Biology, Brown University, USA

*Corresponding authors

Several age predictors based on DNA methylation, dubbed epigenetic clocks, have been created in recent years. Their accuracy and potential for generalization vary widely based on the training data. Here, we gathered 143 publicly available data sets from several human tissues to develop AltumAge, a highly accurate and precise age predictor based on deep learning. Compared to Horvath's 2013 model, AltumAge performs better across both normal and malignant tissues and is more generalizable to new data sets. Interestingly, it can predict gestational week from placental tissue with low error. Lastly, we used deep learning interpretation methods to learn which methylation sites contributed to the final model predictions. We observed that while most important CpG sites are linearly related to age, some highly-interacting CpG sites can influence the relevance of such relationships. We studied the associated genes of these CpG sites and found literary evidence of their involvement in age-related gene regulation. Using chromatin annotations, we observed that the CpG sites with the highest contribution to the model predictions were related to heterochromatin and gene regulatory regions in the genome. We also found age-related KEGG pathways for genes containing these CpG sites. In general, neural networks are better predictors due to their ability to capture complex feature interactions compared to the typically used regularized linear regression. Altogether, our neural network approach provides significant improvement and flexibility to current epigenetic clocks without sacrificing model interpretability.

One of the leading challenges in the field of aging research is measuring age accurately. Accompanying healthy individuals for decades to assess whether an intervention affects the aging process is prohibitive in terms of time and funding. The creation of the 'epigenetic clocks', age predictors that use DNA methylation data, has given researchers a tool to quantitatively measure the aging process. Moreover, recent works [1] have demonstrated precise epigenetic editing based on CRISPR with targeted DNA methylation or demethylation. Consequently, epigenetic clocks have the potential of not only measuring aging but guiding epigenetic interventions.

Notably, two of the most well-known predictors are the ones developed by Hannum *et al.* [2] and Horvath [3] in 2013. Hannum *et al.* [2] developed a blood-based epigenetic clock using 71 CpG sites. Then Horvath [3] showed epigenetic clocks could also accurately predict age across tissues, developing a predictor with 353 CpG sites. Both of these works used simple regularized linear regression (ElasticNet) for feature selection and prediction [4]. More recent epigenetic clocks that predict mortality also use a linear combination of features [5, 6]. ElasticNet has been widely used to develop epigenetic clocks [2, 3, 5–9]. Nevertheless, simple linear regression typically displays high bias and fails to capture non-linear feature-feature interactions in the data.

Interactions among variables can be taken into account by expanding the feature space with feature multiplication. However, incorporating pairwise CpG site interactions is unfeasible given the high dimensionality of DNA methylation data. For his model, Horvath [3] selected 353 CpG sites out of 21,368; to account for all pairwise interactions. If such a model used the entire data, then it would have over 228 million features. The large feature space is especially challenging given the relatively low number of publicly available DNA methylation samples. Given the complexity of the epigenetic regulatory network, it is likely that important interactions among CpG sites are not captured in the current epigenetic clocks developed thus far.

Recently, Galkin *et al.* [10] showed that a deep neural network model, DeepMAge, was slightly superior to Horvath's model in blood samples. However, the authors compared Horvath's pan-tissue predictor to a model trained only in blood DNA methylation data. Moreover, there was no in-depth exploration of why their deep learning model outperformed the ElasticNet model. Similarly, Levy *et al.* [11] developed a deep learning framework to work with DNA methylation data that encodes the CpG sites into latent features for downstream analysis. They showed encouraging results for age prediction using a multi-layer perceptron; however, they investigated only one data set obtained from white blood cells. Therefore, currently, our understanding of the advantages of neural networks for this task in a pan-tissue setting is limited.

We introduce AltumAge, a deep neural network that uses beta values from 21368 CpG sites for pan-tissue age prediction (summarized in Figure 1 (a)). We hypothesized that a neural network using all available CpG sites would be better suited to predict pan-tissue age using DNA methylation data due to their ability to (1) capture higher-order feature interactions and (2) leverage important information contained in the thousands of CpG sites not selected by ElasticNet models. AltumAge uses multi-layer perceptron layers (similar to [5, 10]) that account for non-linear interactions by combining multiple features into each node of the network. We trained AltumAge on samples from 143 different experiments, which, to our knowledge, is the largest compilation of DNA methylation data sets for human age prediction. The publicly available data were obtained from multiple studies that used Illumina 27k and Illumina 450k arrays.

We show that AltumAge has a far lower error and can better generalize to new data sets than ElasticNet models. It also performs substantially better than Horvath's model for age prediction across different normal and cancer tissues. AltumAge is particularly accurate early in life when it can even measure gestational week with a low error. Finally, we apply Shapley-value based interpretation method, called SHAP [12], on AltumAge to determine the contributions of different features towards age prediction (summarized in Figure 1 (b)). We confirm that the most important CpG sites have complex interactions involved when predicting age.

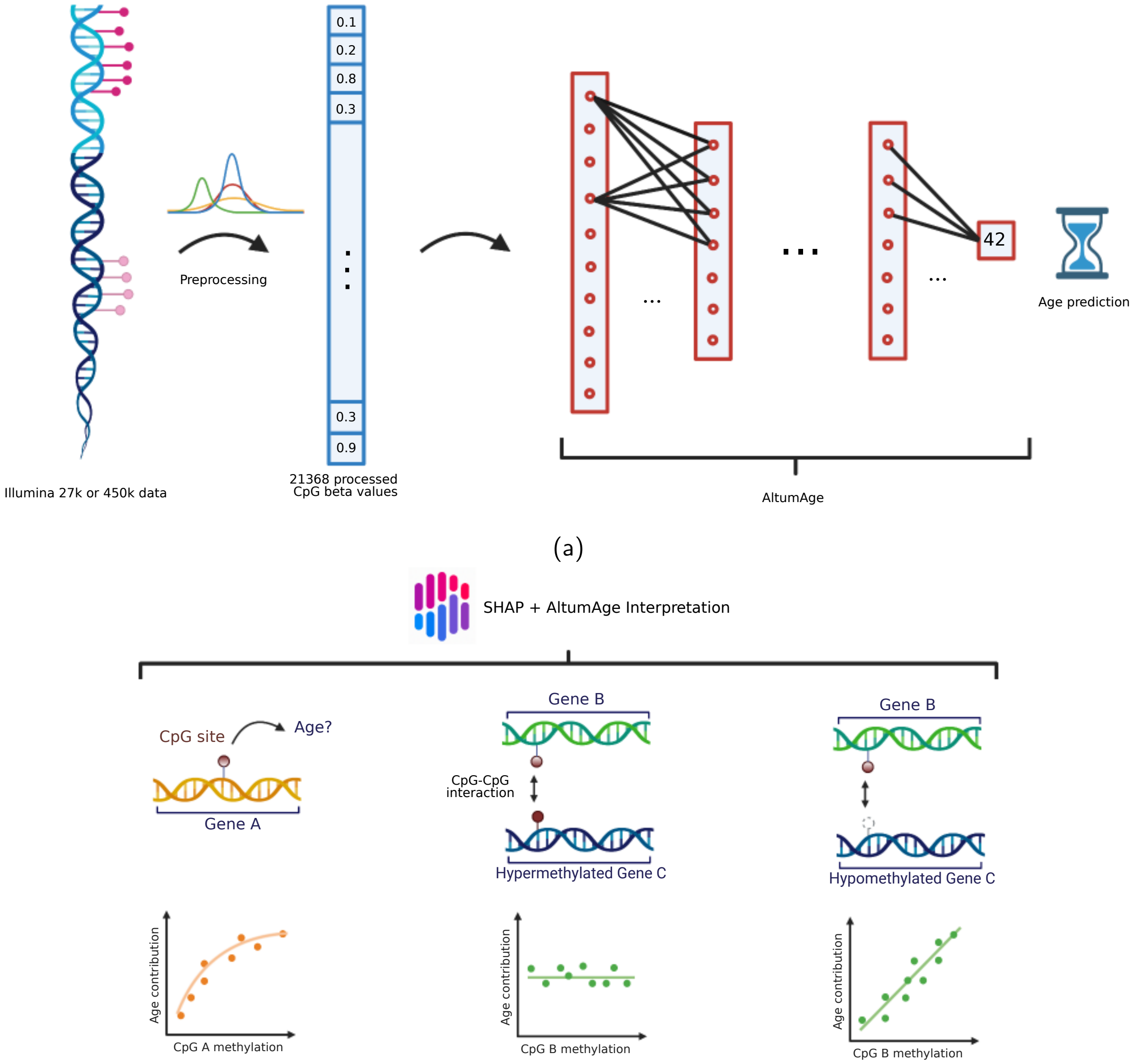
Results

Given that neural networks can capture complex variable interactions, model different data structures, and generally perform better than other machine learning models, we hypothesized that the same would be true for age prediction with DNA methylation data. For model selection, several machine learning models were trained and validated. The hyperparameters of the neural networks were tuned, and the best performer based on both the median absolute error (MAE) and mean squared error (MSE) was dubbed AltumAge. We ran some traditional machine learning methods, including random forest and support vector regression with different hyperparameters. The best performing models were chosen for comparison with AltumAge (see Methods and Supplementary Table S1).

Performance

AltumAge is a better age predictor than linear models

Differences in performance among epigenetic clocks can generally be explained by the data, the model, and the input CpG sites. We used the same training and test sets for each model to control for the data, as our large and diverse DNA methylation data might improve performance compared to other epigenetic clocks. Therefore, we compared the impact of the model and the number of CpG sites used for the input. We trained AltumAge and



(a)

(b)

Figure 1: AltumAge model and interpretation. (a) DNA methylation data from Illumina 27k and 450k arrays are normalized with BMIQ and centered at mean zero and variance one. Then 21368 CpG sites are selected as the input of the model. The information is processed through a first hidden layer with 256 nodes and the remaining seven with 64 nodes. The values of the last hidden layer nodes are combined into a single node as the age output in years. (b) For interpretation, a Shapley-values-based method, called SHAP [12], is used to determine how the methylation status of a specific CpG site affects the age output of AltumAge. Relevant CpG sites generally present a primarily linear relationship (left) with the predicted age. However, interacting CpG sites can change such relationships. In some instances, we find that when a secondary CpG site is hypermethylated (middle), the methylation status of the first CpG is irrelevant for age prediction; when it is hypomethylated (right), then the methylation status becomes essential.

a linear model using three different sets of CpG sites - (1) 353 Horvath's CpG sites, (2) 799 ElasticNet-selected CpG sites, and (3) All the 21,368 CpG sites. The results are summarized in Table 1.

Using the same set of CpG sites as features makes it easier to compare the performance of the two models

Table 1: Evaluation metrics of AltumAge and different linear models in the test set. The median absolute error (MAE) and the median error are in units of year, while the mean squared error (MSE) is in units of year-squared.

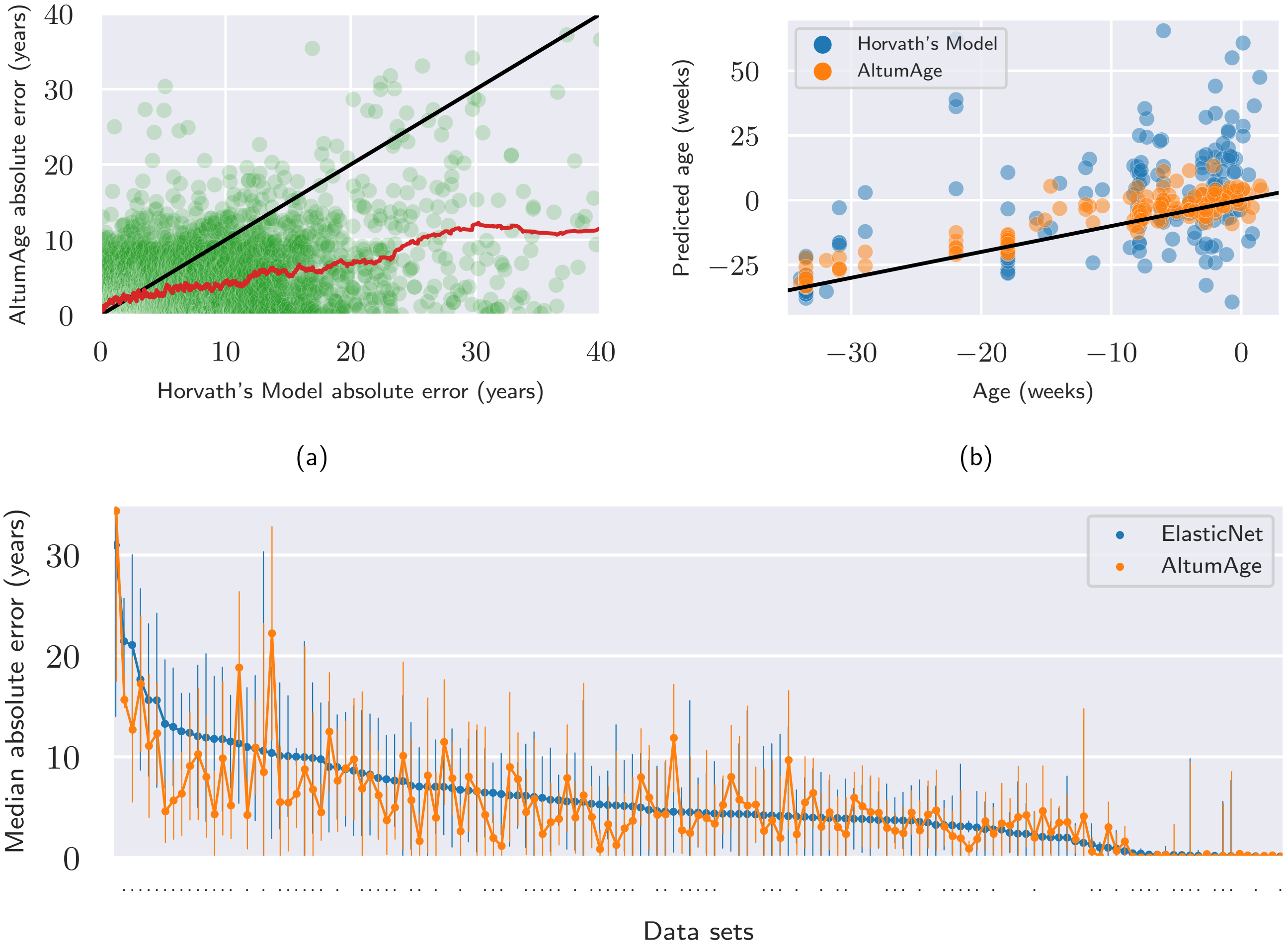
Model	CpGs	MAE	MSE	R	Median Error
AltumAge	21368	**1.926**	**29.614**	**0.980**	-0.024
AltumAge with ElasticNet CpGs	799	2.194	33.638	0.977	-0.031
AltumAge with Horvath's CpGs	353	2.638	39.724	0.973	**0.011**
ElasticNet	799	2.911	44.211	0.970	0.033
Linear Regression with Horvath's CpGs	353	3.230	52.680	0.964	0.026

directly. AltumAge outperformed the respective linear model with Horvath's CpG sites (MAE = 2.638 vs. 3.230, MSE = 39.724 vs. 52.680), ElasticNet-selected CpG sites (MAE = 2.194 vs. 2.911, MSE = 33.638 vs. 44.211), and all 21,368 CpG sites (MAE = 1.926 vs. 87.000, MSE = 29.614 vs.1.639e+20). Overall, the neural network approach outperformed the linear models in all instances. Moreover, for AltumAge, we observed that incorporating more CpG sites reduced the error. This result suggests that the expanded feature set helped improve the performance because relevant information in the epigenome is likely not considered in the couple hundred CpG sites selected by an ElasticNet model. Lastly, it is possible to compare the impact of using larger, more varied training data on the performance of an epigenetic clock. A linear regression using Horvath's 353 CpG sites trained in our data from 143 datasets outperformed Horvath's model trained on 39 datasets (MAE = 3.230 vs. 3.672; MSE = 52.680 vs. 76.023). These results suggest that even though more data lowers the prediction error, AltumAge's performance improvement is far superior to that effect.

AltumAge is a better age predictor than state-of-the-art epigenetic clocks

Horvath's model has been widely used as it is seen as the state-of-the-art pan-tissue epigenetic clock for humans [13–16]. Therefore, it is essential to contrast it with AltumAge. We applied AltumAge and Horvath's model to our data set obtained from 143 experminet. As shown in Figure 2a, AltumAge performs considerably better overall, with a 47.4% lower MAE and 60.6% lower MSE (MAE = 1.926 vs. 3.672, MSE = 29.614 vs. 76.023).

AltumAge is also more robust than Horvath's model across tissue types, with fewer tissues having high MAE. In his 2013 paper, Horvath noticed poor calibration of his model in breast, uterine endometrium, dermal fibroblasts, skeletal muscle, and heart [3]. In our test data, a similarly poor predictive power was found for these tissue types with Horvath's model (breast MAE = 9.462; uterus MAE = 5.804; fibroblast MAE = 10.804; muscle MAE = 9.470; heart not included). AltumAge, on the other hand, had much lower errors (MAE = 4.014, 2.887, 4.621, 2.480 respectively). Furthermore, Horvath's model had an MAE of over 10 years in 45 tissue types in the test data. AltumAge, on the other hand, had MAE > 10 in only three tissue types. AltumAge



(a)

(b)

(c)

Figure 2: Plots showing the improved performance of AltumAge in comparison to Horvath's model and ElasticNet. Top left (A): scatter plot of the absolute error per test sample with AltumAge and Horvath's model. The black line separates the region in the graph in which AltumAge performs better (bottom right) versus where Horvath is superior (top left). The red line is a 100-sample rolling average. AltumAge outperforms Horvath's model, particularly in difficult-to-predict data. Top right (B): scatter plot of the predicted age of each model versus the real age for data sets that had a gestational week available. Zero age is equivalent to gestational week 40. The black line represents the location where the predicted age equals the real age. As shown, AltumAge's predictions are considerably closer to the black line. Bottom (C): point plot showing the LOOCV median absolute error in each of the 143 data sets with AltumAge and an ElasticNet model. Bars represent the standard deviation of the absolute error. A dot below a bar represents data sets in which AltumAge had a lower error than the linear model. For 61.5% of data sets, AltumAge is the better performer.

is also a better age predictor in cancer samples (Supplementary Table S2). Even though the MAE is only slightly improved (MAE = 6.574 vs. 7.429), the MSE is much lower (MSE = 162.961 vs. 289.819).

Supplementary Figure S1 in particular shows how AltumAge, in contrast to Horvath's model, performs well in older ages. Better performance in older age is fundamental in defining biomarkers of age-related diseases of which age is the biggest risk factor. Horvath's model systematically underestimates such population, partly

Table 2: Evaluation metrics for blood-based data sets in DeepMAge and AltumAge, including the sample size (N) used in each test set for each model. The numbers for DeepMAge are reported from the paper [10]. The median absolute error (MAE) and the median error are in units of year, while the mean squared error (MSE) is in units of year-squared.

Data set	DeepMAge N	AltumAge N	DeepMAge MAE	AltumAge MAE	DeepMAge R	AltumAge R
GSE34639	48	20	1.92	**0.08**	0.89	**0.986**
GSE99624	16	19	2.72	**2.134**	**0.93**	0.901
GSE99624	99	40	3.74	**3.084**	**0.81**	0.681
All test blood	1293	2805	2.77	**2.283**	0.97	**0.975**

due to CpG saturation (beta value approaching 0 or 1 in certain genomic loci) [17]. Another reason might be the paper's assumption that age-related CpG changes are linearly correlated with age after 20 years of age. AltumAge resolves these two problems by incorporating an expanded feature set and not using any age transformation function that creates a bias in the data processing.

Interestingly, AltumAge is also accurate in predicting age in early life (Figure 2b, Supplementary Table S3). The MAE of 0.058 years, or 21.2 days, was achievable through a fine-grained encoding of age based on the gestational week in the 12 data sets where it was available. In the US in 2013, the average birth occurred at an estimated 38.5 weeks [18]. This number has changed slightly over time, and since preterm deliveries skew the average more than late-term births, we considered gestational week 40 as age 0 in such data sets. The resultant error is markedly lower than the 0.302 year MAE of Horvath's model. Overall, AltumAge outperforms Horvath's Model for young and old ages, for which the study of age-related factors can be beneficial.

Additionally, we report AltumAge results in comparison to DeepMAge, as it is a recent deep-learning model with an architecture similar to AltumAge [10]. The model code for DeepMAge is not publicly available, nor the description reproducible. Therefore, we were only able to contrast the reported results in the paper for our overlapping test data sets (Table 2), as DeepMAge is a blood-based epigenetic clock. We observe that AltumAge gives lower MAE for all the selected datasets and higher correlation for two out of four datasets. Note that the performance is not directly comparable due to different training and test sizes. However, we hypothesize that this improvement is likely due to the pan-tissue training data.

AltumAge is more generalizable than ElasticNet models

Leave-one-data-set-out cross-validation (LOOCV) provides a way to understand the generalization of a model to new data sets. In this case, one out of the 143 data sets in the training set was left out of model fitting to predict the age for the test set of left out data set. To find the performance of a specific model type across all data sets, 143 different models were consequently fitted for each model type (Figure 2c, Table 3). LOOCV

Table 3: Leave-one-data-set-out cross validation evaluation metrics for AltumAge (with different number of CpG sites), ElasticNet, and the average of AltumAge and ElasticNet. The median absolute error (MAE) and the median error are in units of year, while the mean squared error (MSE) is in units of year-squared.

Model	MAE	MSE	R	Median Error
AltumAge and ElasticNet Mean	**3.336**	**66.949**	**0.955**	0.060
AltumAge	3.620	76.364	0.948	**0.016**
AltumAge with ElasticNet CpG sites	3.856	78.524	0.946	-0.054
ElasticNet	3.878	77.339	0.947	0.162

tests how the model performs for unseen data sets.

Since AltumAge uses 21368 CpG sites, it is expected to be more prone to noise and overfitting than a model with low variances, such as ElasticNet regression, with only a subset of CpG sites. Nevertheless, its MSE is almost identical (MSE = 76.364 vs. 77.339), with AltumAge slightly outperforming its MAE (MAE = 3.620 vs. 3.878).

Given similar results, AltumAge may be simply learning the information contained in the ElasticNet model. One way to determine how similar the predictions of both models are is to look at their correlation of predictions. However, as both AltumAge and ElasticNet are correlated with age, they are inevitably highly correlated (Pearson's correlation coefficient (r) = 0.999). The residuals of each model, in contrast, are not correlated with age by definition. Analyzing the correlation between the residuals of each model can show how similar the predictions are. The residuals of each model are only moderately correlated (r = 0.739). AltumAge, when trained with only the selected features from the ElasticNet regression, performed similarly (MAE = 3.856, MSE = 78.524), and the residuals were more correlated with the ElasticNet residuals (r = 0.806). Interestingly, by averaging the ElasticNet and AltumAge predictions, the performance is further improved, with MAE and MSE 14.0% and 13.4% lower (MAE = 3.336, MSE = 66.949).

As the results of the LOOCV weigh more heavily larger data sets, which are typically blood samples, it is also worth looking at the median of the evaluation metrics for the 143 data sets. A model might be performing extremely well in those large data sets but might have a high error for smaller data sets, skewing the overall MAE and MSE. The median data set MAE (MMAE) and median data set MSE (MMSE) are useful metrics for this evaluation. MMAE and MMSE can be more informative in regards to generalization to new data sets. AltumAge with the whole set of CpGs or only a subset performed similarly (MMAE = 4.216 vs. 4.187), while the ElasticNet had the highest (MMAE = 4.484). AltumAge had a lower MAE in 61.5% of data sets. A similar result is observed with MMSE, with AltumAge outperforming the ElasticNet model (MMSE = 40.367 vs. 58.904).

There does not seem to be specific tissue types in which AltumAge, because of the high number of parameters, performs notably worse than the ElasticNet model (Supplementary Figure S9). AltumAge had at least a 50% worse MAE than ElasticNet in data sets spanning 14 tissue types, while ElasticNet had at least a 50% worse MAE in 46. The overlap consisted of 7 tissue types. These results suggest that AltumAge can better generalize to new tissue types and data sets than ElasticNet models.

Inference

Neural networks, particularly in the context of deep learning, used to be seen as "black-box" methods, as their interpretability was difficult. On the other hand, regardless of the predictive power of ElasticNet models, they are easily understandable. Recently, various methods have been proposed to extract the contribution of features towards a prediction in neural networks. They include interpretation based on model gradients [19–21], attention [22], among others. One such inference method is SHAP [12], which uses a game-theoretic approach to aid in the explanation of machine learning methods. It can measure how one feature contributes to the output of deep neural networks. For our case, the SHAP value can be conceived as how much the value of one CpG site affects the age output of the model in years. Through the architecture of neural networks, it can also determine which CpG sites most highly interact with each other.

To support the results obtained by SHAP, we also applied another method of determining feature importance called DeepPINK [23]. It works by comparing the original features with fake features. The knockoff features can be generated in many different ways, as long as they simulate the original data structure but are not related to the output. DeepPINK contrasts the relevance of the fake features against the regular input features to determine which ones are truly related to the output. It can also be used for feature selection with a controllable false discovery rate (FDR). It is worth highlighting the difficulty in feature selection in DNA methylation data. Most experiments have a couple dozen or a couple hundred samples. Depending on the type of platform used, the number of beta values for the CpG sites analyzed can vary from around 27 thousand to around 850 thousand. DeepPINK, even with a high FDR of 0.5, only selects 78 features. The fact that other sets of CpG sites unrelated to Horvath's 353 also perform similarly well emphasizes the difficulty in finding the "true" age-related CpG sites.

We present results for model inference using SHAP that assist with understanding AltumAge. These results are supported by the importance scores obtained from DeepPINK.

AltumAge captures important CpG-CpG interactions

As epigenetic modifications can significantly influence gene expression, they can also impact genes that affect other epigenetic changes. Some CpG sites interact with others through the gene expression network and can work in tandem. AltumAge, through SHAP, can measure how hyper- or hypomethylation of secondary CpG sites affect the relationship of a CpG of interest and age.

Figure 3 shows scatter plots of the nine most important CpG sites based on SHAP-based importance values. These nine CpG sites are representative of other similarly important sites and account for 0.60% of the total model importance according to SHAP (or 9.78% for DeepPINK, see Supplementary Figure S2). These dependence plots show both the relationship of a CpG site with the predicted age and how that relationship can be affected by the value of a second CpG site for a DNA methylation sample. This secondary CpG site has the highest interaction with the first CpG site, as determined by SHAP. As observed, most of the CpG sites have a mostly linear relationship to the output. This observation explains how ElasticNet, even typically displaying high bias, can perform well in age prediction with DNA methylation data. However, some relationships are not completely linear. For instance, the first and fourth most important CpG sites (cg22736354 and cg06493994, Figures 3a and 3b) have a slight curvature for the lower standardized beta values, even though the plot is mostly linear. Moreover, some of the CpG sites are interacting with others to determine how relevant they are for age. For example, in the third and sixth most important CpG sites (cg10523019 and cg13460409, Figures 3c and 3f), the value of another CpG site (cg26394940) determines how important they are for age prediction. The standardized beta value for cg26394940 affects the final output by changing the slope of the relationship, with a higher value of cg26394940 decreasing the influence of cg10523019 and cg13460409 on the age output. Overall, SHAP shows that the non-linear interaction between CpG sites may partly explain the improvement in the performance of neural networks compared to linear models.

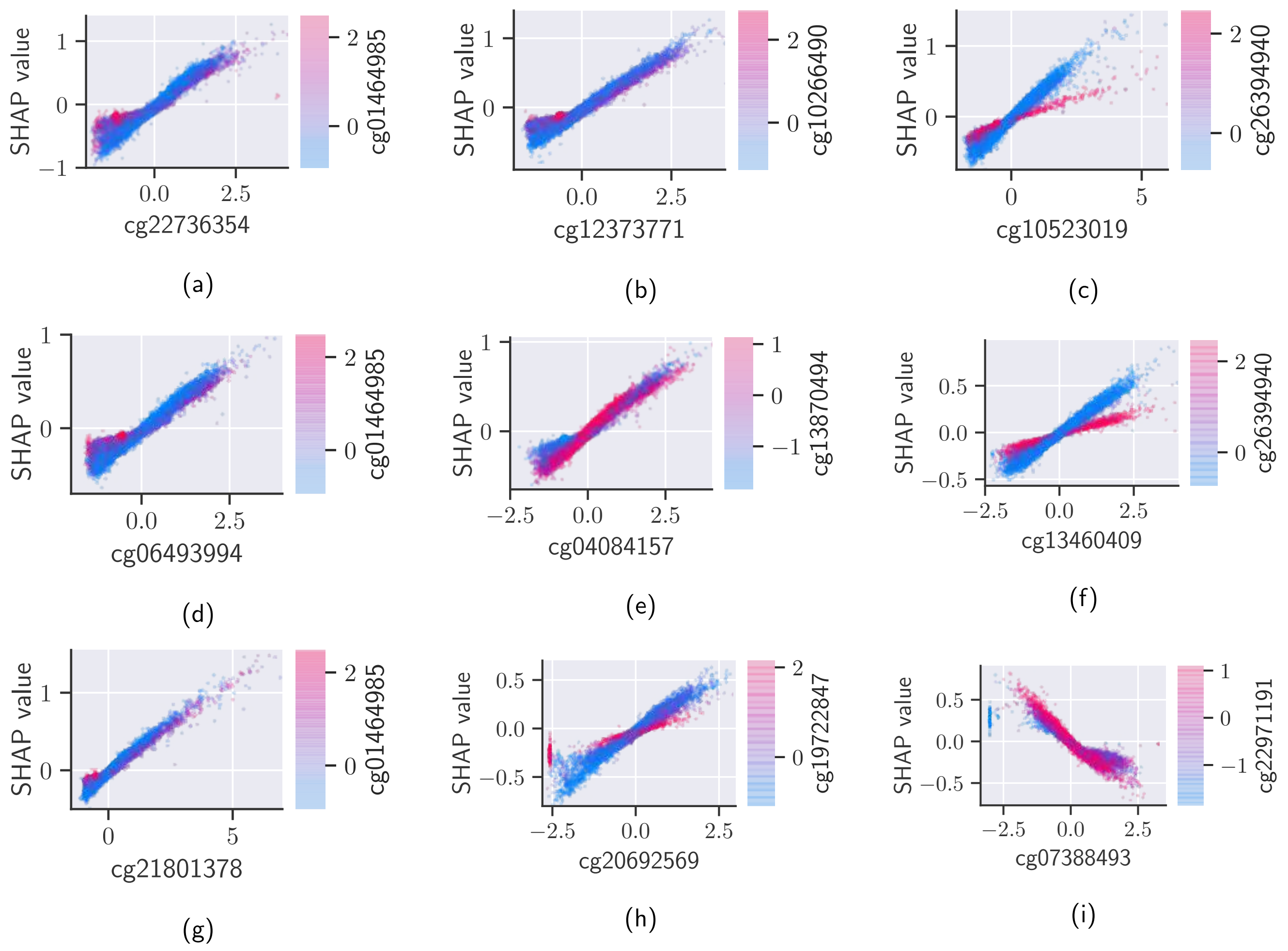
Note that despite their important effects on the predicted age, some of the CpG sites that interact with the most important CpG sites are themselves not particularly relevant for the output. For instance, the cg26394940 mentioned above ranks 385 and 1113 according to SHAP and DeepPINK, respectively, out of 21368. cg01464985, the CpG site with the highest interaction with three out of the top nine CpGs, ranks 2741 and 11582 according to SHAP and DeepPINK. Therefore, an ElasticNet model would likely not select CpG sites whose beta values themselves are not directly relevant but are critical in their influence on other important CpG sites. Supplementary Figure S3 displays their dependence plots, showing how little their SHAP values directly affect age. These results suggest that DNA loci that regulate other loci in aging are relevant for age prediction and may be missed by linear models.

Lastly, it is possible to understand better the function of particular genes based on their SHAP relations. The aforementioned cg01464985 is located in the gene ZNF512, a zinc finger nuclease that likely regulates gene transcription. Consequently, the methylation status of ZNF512, while not directly important to age, may regulate how crucial other genes are to aging. An even clearer picture can be deduced from the cg26394940 located inside the genes PRR34/PRR34-AS1, which code for long noncoding RNAs (lncRNAs). From the SHAP dependence plots only, it is possible to hypothesize that PRR34/PRR34-AS1 regulates the genes in which cg10523019 (Figure 3(c)) and cg13460409 (Figure 3(f)) are located (RHBDD1 and RIPPLY3), and when PRR34/PRR34-AS1 is hypermethylated, its expression is lowered, and the methylation status of cg10523019 and cg13460409 is not that relevant anymore for aging. While not much is known about PRR34/PRR34-AS1, PRR34-AS1 seems to increase expression of the longevity-related transcription factor FOXO3 through inhibition of miR-498 [24]. Furthermore, RHBDD1 is a direct target of FOXO3 in humans [25]. This fact may explain why when cg26394940 (PRR34-AS1) is hypermethylated, the methylation status of cg10523019 (RHBDD1) does not contribute as much to age, likely due to downregulation of FOXO3. In any case, laboratory experiments would have to be performed to more thoroughly characterize these relationships; however, it is possible to obtain data-driven hypotheses from these dependence plots.

Characterization of CpG sites by model interpretation

CCCTC-Binding factor (CTCF) is a transcription factor involved in the negative regulation of several cellular processes. It also contributes to long-range DNA interactions by affecting chromatin architecture. Important CpG sites are overwhelmingly closer to CTCF binding sites (Supplementary Figure S4). This suggests that epigenetic alterations proximal to such loci may alter chromatin packing by affecting CTCF binding, as chromatin structure modifications have been associated with aging [26].

Because of the close relationship between chromatin and aging, we hypothesized that different chromatin states would influence the importance of each CpG site. ChromHMM is a Hidden Markov Model used for the characterization of chromatin states [27]. Annotations for several cell lines and tissue types are widely available online. Since AltumAge is a pan-tissue epigenetic clock, we used the mode of the 18-state annotation from 41 different tissues obtained from ENCODE for each CpG location [28] (Supplementary Figure S5, Supplementary Table S4). The Kruskal-Wallis H-test (see Methods) confirms our hypothesis with both SHAP and DeepPINK importance values ($H = 406.0$, $p = 9.91e\text{-}76$; $H = 52.9$, $p = 1.49e\text{-}5$). The chromatin state with the highest DeepPINK normalized median importance was heterochromatin (DeepPINK importance = 2.04e-14%, top 64th percentile of all CpG sites). DeepPINK, because of the L1 regularization in the algorithm, tends to reduce non-



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 3: Dependence plots of the nine most important CpG sites (a-i) in AltumAge based on SHAP values. They are ordered from top left to bottom right in terms of importance. The x-axis shows the standardized beta values for each specific CpG site; the y-axis, its SHAP value, and the coloring scheme, the standardized beta values for the CpG site with the highest interaction. The effect of a specific CpG site on the predict age can vary drastically based on a second CpG site.

relevant feature importance towards zero, and there were only 29 CpG sites characterized as heterochromatic. Despite these limitations, this result emphasizes the importance of chromatin packing with aging, as it is related to genome stability and maintenance. The chromatin state with the highest SHAP normalized median importance was the 5' flanking region (SHAP importance = 4.58e-3%, top 62nd percentile of all CpG sites). This region contains promoters and sometimes enhancers and is, thus, typically involved in gene regulation.

The importance of each CpG site to age prediction does not seem related to chromosome number for both SHAP and DeepPINK importance values according to the Kruskal-Wallis H-test (H = 23.2, p = 0.33; H = 25.2, p = 0.24).

Importance values were also divided by gene type as some genes, e.g. transposable elements, are associated with aging (Supplementary Figure S6, Supplementary Table S5). Several categories, such as scaRNA, have very few instances since only a couple of the 21368 CpG sites analyzed were contained within scaRNA genes, making the results difficult to interpret. Nevertheless, some observations should be noted. The gene types with the highest DeepPINK normalized mean importance with over 100 CpG sites are protein-coding genes, lncRNAs, unprocessed and processed transcribed pseudogenes. It is expected that protein-coding genes would constitute the bulk of important age-related CpG sites, but it is interesting that lncRNAs, many known to be implicated in the aging process, are also highly important [29].

Aging-related pathways

One of the main interpretation advantages of AltumAge, compared to other ElasticNet models, is that it uses 21368 CpG sites. CpG sites in aging-related genes are often not selected within the couple dozen or couple hundred features of an ElasticNet model, thus making analyses of these CpG sites of interest impossible. AltumAge allows a closer look at the relationship of CpG sites in aging-related pathways even when these CpG sites are not particularly important for the final age prediction.

SIRT, mTOR, and AMPK are some of the most well-known pathways that affect aging [30–32]. Out of Horvath's 353 CpG sites, only one from these pathways was selected (cg11299964, located in MAPKAP1). Nevertheless, it is worth analyzing the relative importance of the other CpG sites in the aging-related pathway.

Unexpectedly, all of the CpG sites in SIRT genes do not appear very relevant, at least directly, for age prediction using AltumAge. Located in SIRT2, cg27442349, accounting for 0.01302% of the total SHAP importance and ranked 954, has the highest SIRT SHAP importance value (Supplementary Figure S7). Located in SIRT7, cg21770145, accounting for 7.89e-12% of total DeepPINK importance and ranked 1426, has the highest SIRT DeepPINK importance value (Supplementary Figure S7).

Out of the 67 proteins participating in the mTOR signaling pathway according to the PID Pathways data set [33], cg11299964, located in MAPKAP1, has the highest SHAP importance of 0.023%, ranking 149. It was the only CpG site from the three main age-related pathways used in Horvath's model. cg05546044, located in MAPK1, has the highest DeepPINK importance of 0.029%, ranking 233. Surprisingly, mTOR was not particularly relevant, with its most important CpG site being cg07029998 (SHAP importance = 0.00811%, rank 3149; DeepPINK importance = 1.12e-12%, rank 2610) (Supplementary Figure S7).

In terms of the AMPK pathway, out of the proteins that directly activate or inhibit AMPK from the KEGG database [34], cg22461835, located in ADRA1A, has the highest SHAP importance of 0.019%, ranking 257. All AMPK-related CpG sites had low (less than 10e-13%) DeepPINK importance values.

Overall, out of all the CpG sites located in SIRT genes, none was significant. In the mTOR and AMPK pathways, some genes were relatively important, ranking in the top 300. We performed KEGG pathway analysis on the genes related to the top-ranking nine CpG sites using KEGGMapper [35]. We found the following genes associated with three of them - NHLRC1 involved in proteolysis; NDUFS5, involved in metabolic pathways, including oxidative phosphorylation and thermogenesis; and FZD9, involved in a range of age-related diseases, including cancer and neurodegeneration. Note that DNA methylation affects gene expression depending on its position. A methylated CpG site in an enhancer, promoter, or gene body may impact gene regulation differently. These findings may shine a light on how methylation in specific loci in aging-related pathways can contribute to age prediction, an insight that is not possible to obtain using regular ElasticNet models.

Discussion

The creation of new quantitative aging measurements has been rapidly expanding with the burgeoning field of the biology of aging. Epigenetic clocks are a tool that can aid researchers to understand better and to measure the aging process. In 2013, Horvath showed it was possible to use just a couple of CpG sites to predict a person's age based on DNA methylation accurately. It was a giant leap in the field. However, his 2013 ElasticNet model or other versions that rely on linear models are still widespread despite recent advances in machine learning. The accuracy of such linear models was so good that it was difficult to imagine a model significantly outperforming it [36]. Other deep learning methods, which slightly outperform ElasticNet models, have focused thus far only in a single tissue type [10] [11].

We show that AltumAge is overall a better age predictor than the original 2103 pan-tissue epigenetic clock. There are several reasons, including (1) the more comprehensive and larger data for training the model; (2) the capability of neural networks to detect complex CpG-CpG interactions; and (3) the expanded feature set with 21368 CpG sites instead of 353. The improved performance of AltumAge LOOCV against an ElasticNet model was not as substantial as in the test set. This is likely because of the difficulty in generalizing to new datasets. There are several data preprocessing and experimental effects that differentiate the DNA methylation among studies. ElasticNet models, which have low variance, are better able to accommodate such differences. Nevertheless, many studies, especially for specific species, create entirely new epigenetic clocks. In those cases, neural networks are vastly superior to simple linear models.

Deep learning models have shown promise in several biological tasks, given their good performance on unstructured data. They have been for many years seen as "black-box" models, but new tools have made it possible to get insights as profound, if not more detailed, than simple ElasticNet models. AltumAge provides a detailed relationship between each one of 21368 CpG sites and age, showing that while most CpG sites are mostly linearly related with age, some important ones are not. Given recent advances in epigenetic editing [1], finding the sweet spot for DNA methylation to delay or reverse aging may be necessary for future interventions to tackle the disease. AltumAge allied with other deep learning inference methods can provide information on highly interacting CpG sites. The primary locus of an epigenetic editing intervention, given its place in the genome, may be difficult to target because of the chromatin structure. Consequently, knowing secondary CpG sites that affect how the CpG of interest interacts with age can also guide such interventions. We show that one can obtain biological hypotheses for the same from the data using AltumAge. For example, we observe that cg26394940 located inside the genes PRR34/PRR34-AS1 could regulate genes with sites cg10523019 (RHBDD1) cg13460409 (RIPPLY3). Analysis of ChromHMM annotations shows that the top-ranking CpG sites are associated with heterochromatin and gene regulatory regions. Finally, we also highlight the age-related KEGG pathways obtained for genes with these CpG sites, indicating that the model is learning valuable biological information from the data.

In future work, it would be interesting to create a deep learning model with Illumina's EPIC array with the roughly 850 thousand CpG sites to understand more deeply how genomic location can affect influence in aging. By having several CpG sites in a single gene, it is also possible to better understand how methylation in different positions may affect the contribution of a particular gene to the aging process. Currently, however, there are only a few EPIC array publicly available data sets.

Overall, we have shown that deep learning represents an improvement in performance over current approaches for epigenetic clocks while at the same time providing new, relevant biological insights about the aging process.

Methods

DNA methylation data sets

In total, we gathered 143 publicly available data sets from the Gene Expression Omnibus, Array Express, and The Cancer Genome Atlas, totaling 15090 normal and 1057 cancer samples. DNA methylation data from the Illumina Infinium HumanMethylation27 BeadChip and the Illumina Infinium HumanMethylation450 BeadChip platforms were used. 21368 CpG sites from both platforms were selected from each array, similarly to Horvath's 2013 paper [3]. The data was normalized using the beta mixture quantile normalization (BMIQ) with the optimized code from Horvath, called BMIQCalibration [3, 37]. Then, each data set was split 60% for training ($n = 8999$), of which one third was used for model validation, and 40% for testing ($n = 6091$). To validate and train the final neural networks, the beta value of each CpG site was scaled so that the mean would equal zero and variance, one. The full list of data sets used is available in the paper's GitHub repository (https://github.com/rsinghlab/AltumAge).

For twelve data sets in which gestational week was available, the encoding for age is the following:

$$y = 7 * \frac{w - 40}{365} \tag{1}$$

where w is the gestational week, and y is the age in years. A gestational week below 40 would have negative age; for instance, 30 weeks would be encoded as $7 * (30 - 40)/365 = -0.192$.

CpG site annotation

For the annotation of CpG sites, GENCODE and Zhou et al's annotations were used [38, 39]. 41 data sets from ENCODE with the 18-state ChromHMM information were gathered [28]. Since AltumAge is a pan-tissue clock, the mode of each state was chosen for each CpG site. This is the list of accession codes: ENCFF717HFZ, ENCFF718AGZ, ENCFF371WNR, ENCFF318XQO, ENCFF340OUL, ENCFF893CAJ, ENCFF151PZS, ENCFF098CED, ENCFF273PJW, ENCFF377YFI, ENCFF773VYR, ENCFF928QES, ENCFF786HDE, ENCFF827FZN, ENCFF364PIY, ENCFF802QCI, ENCFF021NNN, ENCFF510ZEI, ENCFF175NGE, ENCFF670DBL, ENCFF825ZCZ, ENCFF912ILE, ENCFF725WBV, ENCFF829SZB, ENCFF483NRC, ENCFF717RYX, ENCFF249ZBG, ENCFF205OTD, ENCFF765OKG, ENCFF820YPQ, ENCFF685BMF, ENCFF545ZMG, ENCFF294UQS, ENCFF104ZSA, ENCFF370EGY, ENCFF860FWW, ENCFF177TTP, ENCFF151ZGD, ENCFF743GHZ, ENCFF990YHL, and ENCFF036WIO.

Model selection

Since virtually only papers using ElasticNet for epigenetic clocks have been published, multiple different machine learning models were tested in the validation set. The evaluation metrics were median absolute error (MAE), mean squared error (MSE), Pearson's correlation coefficient (R), and median error.

To select the best performing model, we tried some traditional machine learning methods, including random forest and support vector regression, alongside neural networks with different hyperparameters. All code was written in Python 3.8.6 with packages `numpy` version 1.20.3 and `pandas` version 1.2.4 and ran with the arm64 Mac M1 processor.

The non-neural network models were trained with package `scikit-learn` version 0.24.1. They were: support vector regression with all features; random forest with all features; ElasticNet with hyperparameter λ selected with cross validation with 20 values; ElasticNet with λ so that the number of features selected was 353; linear regression with Horvath's 353 CpG sites.

All the neural networks were trained with `tensorflow` version 2.4.0. They were trained using the Adam optimizer (learning rate = 0.0001) for 1000 epochs, with an early stopping if the validation loss did not improve after 400 epochs.

Holding constant the learning rate, the maximum number of epochs, and the activation function (ReLU), the number of fully-connected hidden layers was varied from two to eleven and the number of nodes per layer from 64 to 512. The neural networks converged at around 400 epochs and did not overfit if trained for longer. The performance of the best architectures was similar, so the one with 256 nodes for the first hidden layer and 64 nodes for the other seven hidden layers was chosen to balance performance and ease of training. Then, the ReLU activation function was compared with SeLU, with the latter improving all evaluation metrics. Finally, as batch normalization typically assists with training for deep neural network, we tried to add it between hidden layers. However, the performance decreased. Therefore, we dubbed the deep neural network with 256 nodes for the first hidden layer and 64 nodes for the following seven layers with SeLU activation as AltumAge.

Another handful of models were also validated: AltumAge using only Horvath's 353 CpG sites; AltumAge using only the selected CpG sites from the cross-validated ElasticNet; and AltumAge using the 78 CpG sites selected by DeepPINK with a false discovery rate (FDR) of 0.5. Lastly, Horvath's model was validated based on the instructions from Horvath's paper [3].

The validation metrics with the full list of models is in Supplementary Table S1. Support vector regression was by far the worst performer (MAE = 7.07, MSE = 186.631), being the only model with Pearson's correlation coefficient below 0.9. Random forest and Horvath's model were the next poorest predictors (MAE = 4.366, MSE = 78.494 vs MAE = 3.637, MSE = 74.581). Next, AltumAge with 78 CpG sites selected by DeepPINK was a slight improvement over Horvath's model, even using 78% fewer features. The worst performing neural network with only two hidden layers and 64 nodes in each had an MSE less than half of Horvath's model (MSE = 33.648 vs MSE = 74.581), and a much lower MAE (MAE = 2.279 vs MAE = 3.637). The model with both the lowest MAE and MSE was AltumAge using all 21368 CpG sites (MAE = 2.071, MSE = 30.075). Based on these results from the validation set, AltumAge with all features performed the best.

The final models used in the test set from Table 1 were identically as in model validation, with the exception that the neural networks did not have early stopping.

SHAP and DeepPINK

To obtain the SHAP values for AltumAge, the python package `shap` version 0.35.0 was used. With the entire training data set, the function `GradientExplainer` resulted in all SHAP values. For the DeepPINK importance values and feature selection, the standard architecture and number of epochs was used [23]. To create the knockoff features for DeepPINK, the function `knockoff.filter` from the R 4.0.2 package `knockoff` version 0.3.3 was used with the importance statistic based on the square-root lasso.

Both SHAP and DeepPINK importance values were normalized so that their sum would equal to 100. Therefore, each importance value represents a percent contribution of a certain feature.

Equations

ElasticNet models are trained by minimizing the following loss function:

$$\mathcal{L}(\hat{\beta}) = \frac{\sum_{i=1}^{n}(y_i - \hat{\beta}^{\mathsf{T}}\mathbf{x}_i)^2}{2n} + \lambda\left(\frac{1-\alpha}{2}\sum_{j=1}^{m}\hat{\beta}_j^2 + \alpha\sum_{j=1}^{m}|\hat{\beta}_j|\right), \tag{2}$$

where n is the number of samples, m is the number of independent variables (CpG sites), y is the dependent variable (age), $\mathbf{x}$ is the vector of independent variables (beta values for each CpG site), $\hat{\beta}$ is the vector of estimated coefficients in the linear regression, α is a parameter for the proportion of L1 to L2 penalty, and λ is a hyperparameter. As observed in the left side of Equation 2, only the linear combination of the model coefficients with the CpG sites, $\hat{\beta}^{\mathsf{T}}\mathbf{x}$, is minimized, without considering feature interactions.

The number of combinations can be calculated as:

$$\binom{m}{k} = \frac{m!}{k!(m-k)!}, \tag{3}$$

where m is the number of features and $k = 2$ for pairwise interactions only.

Shannon's information entropy has the formula:

$$S = -\sum_{i}(\beta * \log(\beta)) \tag{4}$$

where S is the entropy, β is each CpG beta value, and $i = 21368$.

Statistical Analysis

Given that the importance values grouped by ChromHMM state and chromosome do not fulfill the assumption of homoscedasticity, we used the Kruskal-Wallis H-test, which is a non-parametric version of ANOVA, for such tests. The Kruskal-Wallis H-test is more flexible despite having lower statistical power. The H statistic and p-value were computed using the function `kruskal` with standard parameters from python package `scipy` version 1.6.2.

To assess the performance of the models, we used median absolute error (MAE), mean squared error (MSE), median error, and Pearson's correlation coefficient (R). For the LOOCV, we also used two other statistics to compare the models, namely the median data set MAE (MMAE) and the median data set MSE (MMSE), which can inform on the expected performance of the model in a new data set.

Data and Code Availability

The list of all the data sets used, a summary of the results per data set, and detailed instructions to run AltumAge can be found in the paper's GitHub repository (https://github.com/rsinghlab/AltumAge). The GitHub also links to a Google Drive where our gathered DNA methylation data is publicly available.

Author Contributions

L.P.L.C conceived of the presented idea. R.S and L.P.L.C designed the methodology and the experiments. L.P.L.C conducted all the experiments. L.R.L assisted with the analysis and biological interpretation of the results. All authors discussed the results and contributed to the final manuscript.

Funding

L.R.L is funded by a R01 (AG051810) and R21(AG068922) from the National Institute on Aging.

Competing Interests Statement

The authors declare that they have no conflict of interest.

References

[1] James K. Nuñez et al. "Genome-Wide Programmable Transcriptional Memory by CRISPR-Based Epigenome Editing". English. In: *Cell* 184.9 (Apr. 2021), 2503–2519.e17. ISSN: 0092-8674, 1097-4172. DOI: `10.1016/j.cell.2021.03.025`.

[2] Gregory Hannum et al. "Genome-Wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates". In: *Molecular cell* 49.2 (Jan. 2013), pp. 359–367. ISSN: 1097-2765. DOI: `10.1016/j.molcel.2012.10.016`.

[3] Steve Horvath. "DNA Methylation Age of Human Tissues and Cell Types". In: *Genome Biology* 14.10 (Dec. 2013), p. 3156. ISSN: 1474-760X. DOI: `10.1186/gb-2013-14-10-r115`.

[4] Hui Zou and Trevor Hastie. "Regularization and Variable Selection via the Elastic Net". en. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2 (Apr. 2005), pp. 301–320. ISSN: 1369-7412, 1467-9868. DOI: `10.1111/j.1467-9868.2005.00503.x`.

[5] Morgan E. Levine et al. "An Epigenetic Biomarker of Aging for Lifespan and Healthspan". In: *Aging (Albany NY)* 10.4 (Apr. 2018), pp. 573–591. ISSN: 1945-4589. DOI: `10.18632/aging.101414`.

[6] Ake T. Lu et al. "DNA Methylation GrimAge Strongly Predicts Lifespan and Healthspan". In: *Aging (Albany NY)* 11.2 (Jan. 2019), pp. 303–327. ISSN: 1945-4589. DOI: `10.18632/aging.101684`.

[7] Michael J. Thompson et al. "A Multi-Tissue Full Lifespan Epigenetic Clock for Mice". In: *Aging (Albany NY)* 10.10 (Oct. 2018), pp. 2832–2854. ISSN: 1945-4589. DOI: `10.18632/aging.101590`.

[8] Steve Horvath et al. "Reversing Age: Dual Species Measurement of Epigenetic Age with a Single Clock". en. In: *bioRxiv* (May 2020), p. 2020.05.07.082917. DOI: `10.1101/2020.05.07.082917`.

[9] V. J. Sugrue et al. "Castration Delays Epigenetic Aging and Feminises DNA Methylation at Androgen-Regulated Loci". en. In: *bioRxiv* (Nov. 2020), p. 2020.11.16.385369. DOI: `10.1101/2020.11.16.385369`.

[10] Polina Mamoshina Fedor Galkin and Polina Mamoshina Fedor Galkin. "DeepMAge: A Methylation Aging Clock Developed with Deep Learning". en. In: *Aging and disease* (), p. 0. ISSN: 2152-5250. DOI: `10.14336/AD.2020.1202`.

[11] Joshua J. Levy et al. "MethylNet: An Automated and Modular Deep Learning Approach for DNA Methylation Analysis". In: *BMC Bioinformatics* 21.1 (Mar. 2020), p. 108. ISSN: 1471-2105. DOI: `10.1186/s12859-020-3443-8`.

[12] Scott M. Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". en. In: *Advances in Neural Information Processing Systems* 30 (2017), pp. 4765–4774.

[13] Gregory M. Fahy et al. "Reversal of Epigenetic Aging and Immunosenescent Trends in Humans". en. In: *Aging Cell* 18.6 (2019), e13028. ISSN: 1474-9726. DOI: `10.1111/acel.13028`.

[14] Li Chen et al. "Effects of Vitamin D3 Supplementation on Epigenetic Aging in Overweight and Obese African Americans With Suboptimal Vitamin D Status: A Randomized Clinical Trial". In: *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 74.1 (Jan. 2019), pp. 91–98. ISSN: 1079-5006. DOI: `10.1093/gerona/gly223`.

[15] Nelly Olova et al. "Partial Reprogramming Induces a Steady Decline in Epigenetic Age before Loss of Somatic Identity". In: *Aging Cell* 18.1 (Feb. 2019). ISSN: 1474-9718. DOI: `10.1111/acel.12877`.

[16] Kara N. Fitzgerald et al. "Potential Reversal of Epigenetic Age Using a Diet and Lifestyle Intervention: A Pilot Randomized Clinical Trial". en. In: *Aging* (Mar. 2021). ISSN: 1945-4589. DOI: `10.18632/aging.202913`.

[17] Louis Y. El Khoury et al. "Systematic Underestimation of the Epigenetic Clock and Age Acceleration in Older Subjects". In: *Genome Biology* 20.1 (Dec. 2019), p. 283. ISSN: 1474-760X. DOI: `10.1186/s13059-019-1810-4`.

[18] "National Vital Statistics Reports, Volume 64, Number 5, (06/01/2015)". en. In: (), p. 20.

[19] Ramprasaath R Selvaraju et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.

[20] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks". In: *arXiv preprint arXiv:1703.01365* (2017).

[21] Daniel Smilkov et al. "Smoothgrad: removing noise by adding noise". In: *arXiv preprint arXiv:1706.03825* (2017).

[22] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473* (2014).

[23] Yang Lu et al. "DeepPINK: Reproducible Feature Selection in Deep Neural Networks". en. In: *Advances in Neural Information Processing Systems* 31 (2018), pp. 8676–8686.

[24] Zhaoming Liu et al. "Long Noncoding RNA PRR34-AS1 Aggravates the Progression of Hepatocellular Carcinoma by Adsorbing microRNA-498 and Thereby Upregulating FOXO3". In: *Cancer Management and Research* 12 (2020), p. 10749.

[25] Ashley E Webb, Anshul Kundaje, and Anne Brunet. "Characterization of the direct targets of FOXO transcription factors throughout evolution". In: *Aging cell* 15.4 (2016), pp. 673–685.

[26] Lucas Paulo de Lima Camillo and Robert B. A. Quinlan. "A Ride through the Epigenetic Landscape: Aging Reversal by Reprogramming". en. In: *GeroScience* (Apr. 2021). ISSN: 2509-2723. DOI: `10.1007/s11357-021-00358-6`.

[27] Jason Ernst and Manolis Kellis. "ChromHMM: Automating Chromatin-State Discovery and Characterization". en. In: *Nature Methods* 9.3 (Mar. 2012), pp. 215–216. ISSN: 1548-7105. DOI: `10.1038/nmeth.1906`.

[28] Carrie A. Davis et al. "The Encyclopedia of DNA Elements (ENCODE): Data Portal Update". eng. In: *Nucleic Acids Research* 46.D1 (Jan. 2018), pp. D794–D801. ISSN: 1362-4962. DOI: `10.1093/nar/gkx1081`.

[29] Sukhleen Kour and Pramod C. Rath. "Long Noncoding RNAs in Aging and Age-Related Diseases". en. In: *Ageing Research Reviews* 26 (Mar. 2016), pp. 1–21. ISSN: 1568-1637. DOI: `10.1016/j.arr.2015.12.001`.

[30] Lijun Zhao et al. "Sirtuins and Their Biological Relevance in Aging and Age-Related Diseases". In: *Aging and Disease* 11.4 (July 2020), pp. 927–945. ISSN: 2152-5250. DOI: `10.14336/AD.2019.0820`.

[31] Thomas Weichhart. "mTOR as Regulator of Lifespan, Aging and Cellular Senescence". In: *Gerontology* 64.2 (2018), pp. 127–134. ISSN: 0304-324X. DOI: `10.1159/000484629`.

[32] Kristopher Burkewitz, Yue Zhang, and William B. Mair. "AMPK at the Nexus of Energetics and Aging". In: *Cell metabolism* 20.1 (July 2014), pp. 10–25. ISSN: 1550-4131. DOI: `10.1016/j.cmet.2014.03.002`.

[33] Andrew D. Rouillard et al. "The Harmonizome: A Collection of Processed Datasets Gathered to Serve and Mine Knowledge about Genes and Proteins". In: *Database* 2016.baw100 (Jan. 2016). ISSN: 1758-0463. DOI: `10.1093/database/baw100`.

[34] Minoru Kanehisa and Yoko Sato. "KEGG Mapper for Inferring Cellular Functions from Protein Sequences". eng. In: *Protein Science: A Publication of the Protein Society* 29.1 (Jan. 2020), pp. 28–35. ISSN: 1469-896X. DOI: `10.1002/pro.3711`.

[35] Minoru Kanehisa and Yoko Sato. "KEGG Mapper for inferring cellular functions from protein sequences". In: *Protein Science* 29.1 (2020), pp. 28–35.

[36] Steve Horvath and Kenneth Raj. "DNA Methylation-Based Biomarkers and the Epigenetic Clock Theory of Ageing". en. In: *Nature Reviews Genetics* 19.6 (June 2018), pp. 371–384. ISSN: 1471-0064. DOI: `10.1038/s41576-018-0004-3`.

[37] Andrew E. Teschendorff et al. “A Beta-Mixture Quantile Normalization Method for Correcting Probe Design Bias in Illumina Infinium 450 k DNA Methylation Data”. In: *Bioinformatics* 29.2 (Jan. 2013), pp. 189–196. ISSN: 1367-4803. DOI: `10.1093/bioinformatics/bts680`.

[38] Adam Frankish et al. “GENCODE Reference Annotation for the Human and Mouse Genomes”. eng. In: *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D766–D773. ISSN: 1362-4962. DOI: `10.1093/nar/gky955`.

[39] Wanding Zhou, Peter W. Laird, and Hui Shen. “Comprehensive Characterization, Annotation and Innovative Use of Infinium DNA Methylation BeadChip Probes”. In: *Nucleic Acids Research* 45.4 (Feb. 2017), e22–e22. ISSN: 0305-1048. DOI: `10.1093/nar/gkw967`.

[40] David A Sinclair, Matthew D LaPlante, and Catherine Delphia. *Lifespan: Why We Age–and Why We Don’t Have To*. English. 2019. ISBN: 978-1-5011-9197-8.

Supplementary Information

Entropy

It has been shown that the DNA methylation entropy is correlated with aging rate in blood tissue [2]. Entropy sits at the corner of the Information Loss Theory of Aging [40], which purports that the aging process is caused by loss of epigenetic information.

Here, it is possible to determine the relationship between the entropy of the DNA methylation beta values of the 21368 CpG sites and age through SHAP. AltumAge was fitted again with all CpG sites plus Shannon's information entropy (Equation 4), and SHAP values were obtained. The dependence plot is shown in Figure S8. It has a similar profile as the ones for the other features, being mostly linear with the slope being determined by other CpGs. The top three interacting CpG sites are cg14244577, cg01511567, and cg26394940, located in DDX19A, SSRP1, and MIRLET7BHG and PRR34. The first is an RNA helicase. The second makes part of FACT, a chromatin transcriptional elongation factor. It interacts with histones H2A/H2B to effect nucleosome disassembly and transcription elongation. MIRLET7BHG and PRR34 are genes that code for long non coding RNAs. Surprisingly, the entropy of the 21368 CpG sites, according to SHAP values, appears to be generally negatively correlated with age. This goes contrary to Hannum et al.'s results [2]. Moreover, when cg14244577 is highly methylated, entropy had almost no relationship to AltumAge's output. These differences might arise for some reasons. Hannum et al. only used blood DNA methylation from the Illumina 450k array, whereas I used DNA methylation data from multiple tissues. Another reason is the difference in Hannum et al.'s direct correlation between age and entropy, as opposed to understanding how entropy interacts with CpG sites to determine a person's age. In AltumAge specifically, it appears that a higher entropy is negatively related with age.

Table S1: Evaluation metrics for all models in the validation set. The median absolute error (MAE) and the median error are in units of year, while the mean squared error (MSE) is in units of year-squared.

Model	CpGs	MAE	MSE	R	Median Error
AltumAge	21368	**2.071**	**30.075**	**0.98**	0.012
512-512-512-512	21368	2.08	30.906	0.979	0.021
256-64-64-64-64-64-64-64-64-64-64-64	21368	2.111	30.917	0.979	0.022
256-256-256-256	21368	2.113	31.261	0.979	0.007
256-64-64-64-64-64-64-64-64-64	21368	2.118	30.922	0.979	-0.016
256-256-256	21368	2.129	31.069	0.979	0.029
512-512	21368	2.135	31.471	0.979	-0.01
256-64-64-64-64-64-64-64	21368	2.145	31.492	0.979	-0.007
AltumAge with BatchNorm	21368	2.145	30.249	0.98	-0.536
512-512-512	21368	2.147	30.822	0.979	-0.009
128-128-128	21368	2.152	31.595	0.979	0.107
256-64-64-64	21368	2.155	32.033	0.978	-0.016
256-64-64-64-64-64	21368	2.167	31.6	0.979	-0.028
256-256	21368	2.181	31.472	0.979	0.128
64-64-64-64	21368	2.19	31.562	0.979	-0.035
64-64-64	21368	2.199	32.052	0.978	-0.1
128-128-128-128	21368	2.206	31.0	0.979	-0.065
AltumAge with ElasticNet CpGs	1504	2.227	30.899	0.979	-0.012
128-128	21368	2.228	32.341	0.978	0.04
64-64	21368	2.279	33.648	0.977	0.084
AltumAge with Horvath's CpGs	21368	2.705	41.144	0.972	-0.027
ElasticNet	1504	2.768	40.422	0.972	0.041
ElasticNet with 353 CpGs	353	3.031	59.034	0.96	0.037
Linear Regression with Horvat's CpGs	353	3.33	54.359	0.963	0.099
AltumAge with DeepPINK CpGs	78	3.422	61.028	0.959	**-0.003**
Horvath's 2013 Model [3]	353	3.637	74.581	0.949	-0.135
Random Forest	21368	4.366	78.494	0.947	-0.142
Support Vector Regression	21368	7.07	186.631	0.877	-0.209

Table S2: Evaluation metrics for AltumAge and Horvath's model in the cancer data sets. The median absolute error (MAE) and the median error are in units of year, while the mean squared error (MSE) is in units of year-squared.

Model	MAE	MSE	R	Median Error
AltumAge	**6.574**	**162.961**	**0.620**	-0.454
Horvath's 2013 Model [3]	7.429	289.819	0.522	**0.389**

Table S3: Evaluation metrics for AltumAge and Horvath's model in the test data with labeled gestational week. The median absolute error (MAE) and the median error are in units of year, while the mean squared error (MSE) is in units of year-squared. One outlier, with MAE $>$ 40 years for both models, was removed to avoid skewing the statistics.

Model	MAE	MSE	R	Median Error
AltumAge	**0.058**	**0.634**	**0.302**	**-0.049**
Horvath's 2013 Model [3]	0.302	6.754	0.206	-0.262

Figure S1: Scatter plot showing the improved performance of AltumAge in comparison to Horvath's 2013 model for older ages. The black line represents the location where the predicted age equals the real age. AltumAge's predictions are generally closer to the black line. Horvath's predictions tends to give lower performance in higher ages.

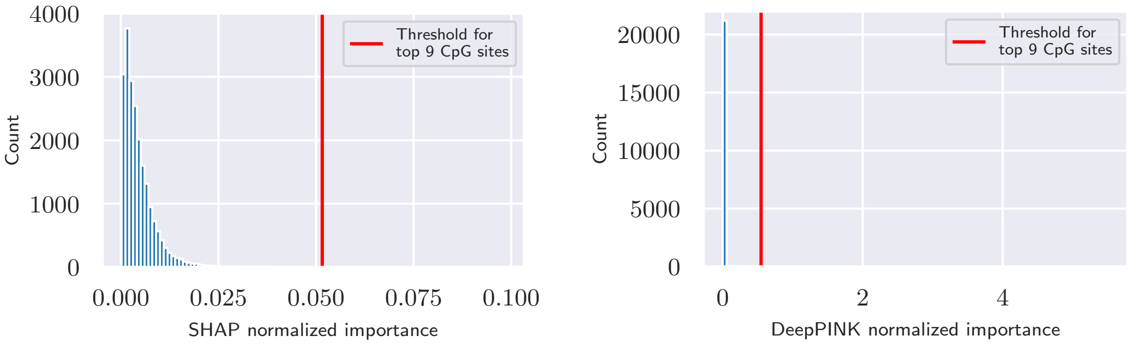


Figure S2: Histograms of the normalized importance values of all AltumAge CpG sites according to SHAP and DeepPINK. The red line represents the threshold for the top nine CpG sites. These have a much higher importance than most other CpG sites.

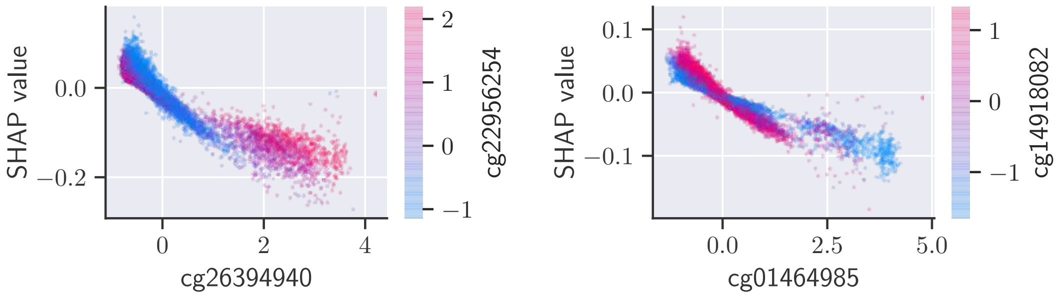


Figure S3: Dependence plots of two CpG sites that interact highly with some of the most important CpGs in AltumAge based on SHAP values. The x-axis shows the standardized beta values for each specific CpG site; the y-axis, its SHAP value, and the coloring scheme, the standardized beta values for the CpG site with the highest interaction. cg26394940, left, has the highest interaction with two of the top nine most important CpG sites; cg01464985, with three. Their overall SHAP values are low, generally less than 0.2 in magnitude.

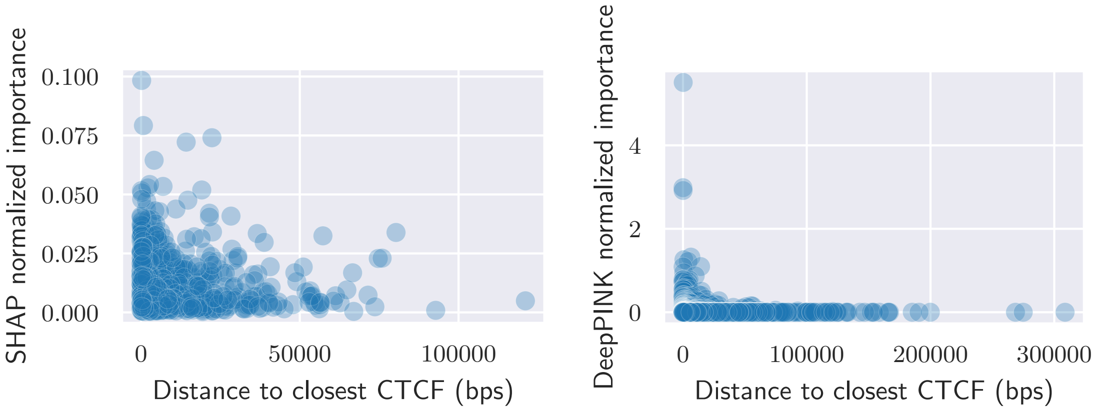


Figure S4: Scatter plots of the normalized importance values of the top 1000 most important CpG sites according to SHAP and DeepPINK by distance to CTCF binding site in basepairs. The importance of each CpG site tends to decline the farther away it is from the closest CTCF binding site.

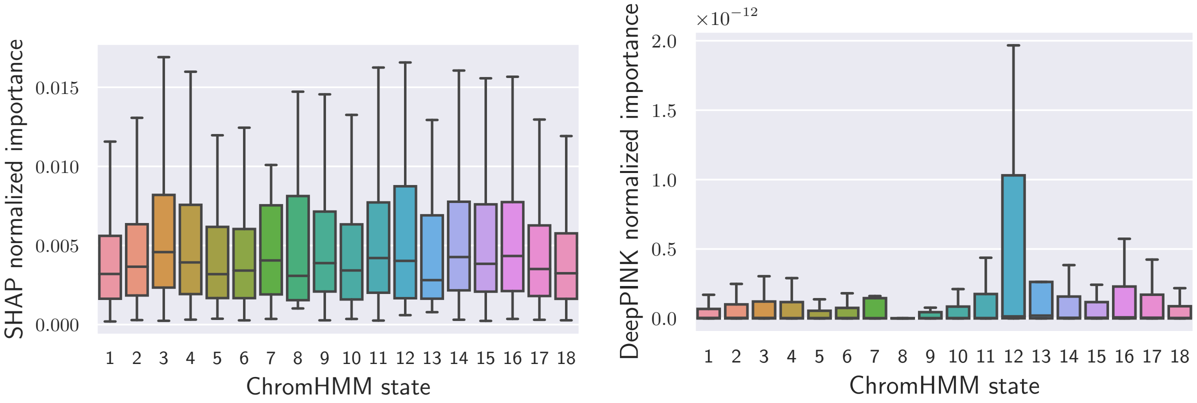


Figure S5: Box plots of SHAP and DeepPINK normalized importance values by ChromHMM state. Outliers were removed for better figure visualization. No specific ChromHMM state stands out in importance.

Table S4: List of ChromHMM states by ChromHMM state ID.

ChromHMM state ID	ChromHMM state
1	Active TSS
2	Flanking TSS
3	Flanking TSS Upstream
4	Flanking TSS Downstream
5	Strong transcription
6	Weak transcription
7	Genic enhancer1
8	Genic enhancer2
9	Active Enhancer 1
10	Active Enhancer 2
11	Weak Enhancer
12	ZNF genes and repeats
13	Heterochromatin
14	Bivalent/Poised TSS
15	Bivalent Enhancer
16	Repressed PolyComb
17	Weak Repressed PolyComb
18	Quiescent/Low

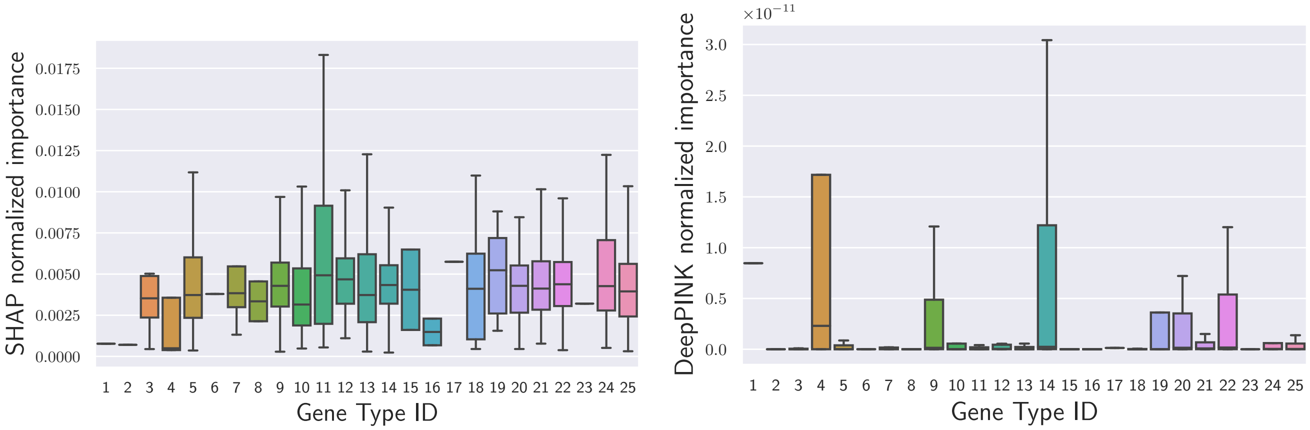


Figure S6: Box plots of SHAP and DeepPINK normalized importance values by GENCODE gene type. It is difficult to visualize any effect since very few CpG sites were located in genes of certain types, leading to high error bars.

Table S5: List of GENCODE gene types by gene ID index.

Gene type ID	Gene type
1	IG C gene
2	IG C pseudogene
3	IG V gene
4	IG V pseudogene
5	TEC
6	TR C gene
7	TR V gene
8	TR V pseudogene
9	lncRNA
10	miRNA
11	miscRNA
12	polymorphic pseudogene
13	processed pseudogene
14	protein coding
15	pseudogene
16	rRNA pseudogene
17	scaRNA
18	snRNA
19	snoRNA
20	transcribed processed pseudogene
21	transcribed unitary pseudogene
22	transcribed unprocessed pseudogene
23	translated unprocessed pseudogene
24	unitary pseudogene
25	unprocessed pseudogene

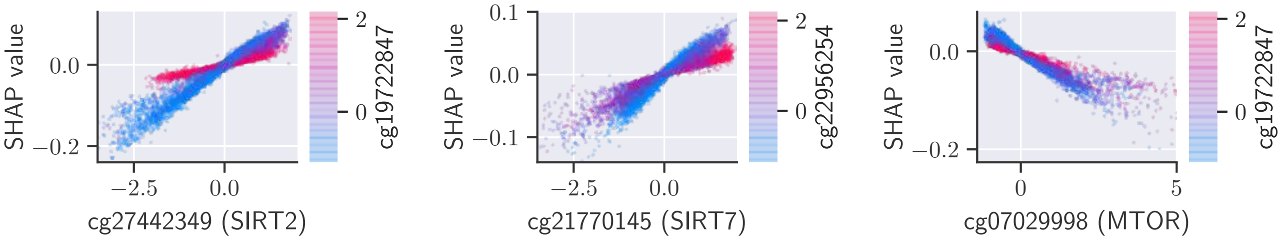


Figure S7: SHAP dependence plots of three CpG sites in SIRT2, SIRT7, and MTOR. The x-axis shows the standardized beta values for each specific CpG site; the y-axis, its SHAP value, and the coloring scheme, the standardized beta values for the CpG site with the highest interaction. These are the most important CpG sites according to SHAP for AltumAge in the SIRT and MTOR pathways.

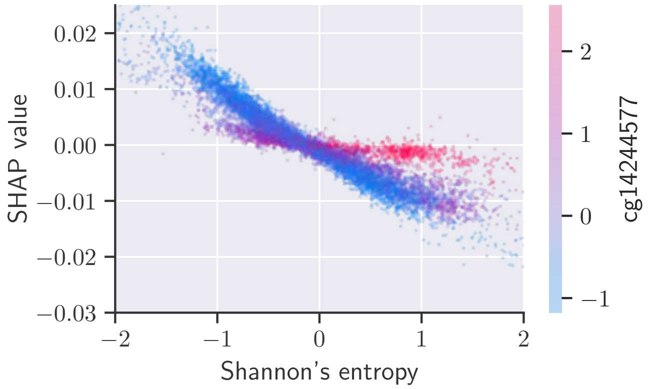


Figure S8: Dependence plot of the SHAP values for Shannon's entropy (standardized). Its impact on the final model output, regardless of the value, is minimal, below 0.03 years in magnitude.

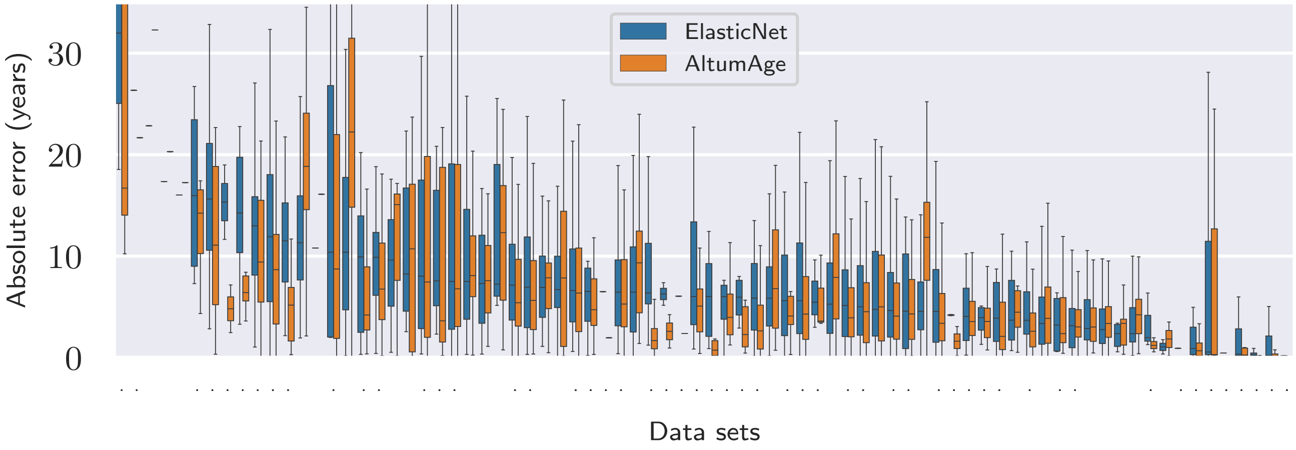


Figure S9: Box-and-whisker diagram showing the LOOCV absolute error per tissue type with AltumAge and an ElasticNet model. The box shows the top quartile, median, and bottom quartile, while the whiskers encompass a maximum of 1.5 times the inter-quartile range. For ease of visualization, outliers outside of the whiskers range were removed. A dot below a bar represents data sets in which AltumAge had a lower error than the linear model. For 53 out of 78 tissue types (67.9%), AltumAge performed better.