



OPEN

## Biologically informed deep learning for explainable epigenetic clocks

Aurel Prosz<sup>1</sup>, Orsolya Pipek<sup>2</sup>, Judit Börcsök<sup>1,3</sup>, Gergely Palla<sup>4,5</sup>, Zoltan Szallasi<sup>1</sup>, Sandor Spisak<sup>6</sup>✉ & István Csabai<sup>2</sup>

Ageing is often characterised by progressive accumulation of damage, and it is one of the most important risk factors for chronic disease development. Epigenetic mechanisms including DNA methylation could functionally contribute to organismal aging, however the key functions and biological processes may govern ageing are still not understood. Although age predictors called epigenetic clocks can accurately estimate the biological age of an individual based on cellular DNA methylation, their models have limited ability to explain the prediction algorithm behind and underlying key biological processes controlling ageing. Here we present XAI-AGE, a biologically informed, explainable deep neural network model for accurate biological age prediction across multiple tissue types. We show that XAI-AGE outperforms the first-generation age predictors and achieves similar results to deep learning-based models, while opening up the possibility to infer biologically meaningful insights of the activity of pathways and other abstract biological processes directly from the model.

Aging, defined as some form of functional decline over time, has always attracted a considerable interest among humankind, and has been in the focus of intense research from a wide range of perspectives<sup>1</sup>. According to the related studies, certain biomarkers can rather precisely predict the functional capability of tissues, organs and even patients<sup>2,3</sup>. Furthermore, age-related biomarkers enable the introduction of the concept of biological age<sup>4,5</sup>, which can bring additional information in the risk assessments for age-related conditions on top of chronological age.

One of the most promising age-predictive biomarkers are the ones based on DNA-methylation<sup>6–8</sup>, which can be used for basically any source of DNA from sorted cells through tissues to organs.

Age-related changes in DNA methylomes are generally occurring processes, during which up to 2–14% of all cytosine-guanine dinucleotide (CpG) sites display consistent changes in their methylation levels throughout ageing<sup>9–18</sup>.

The combination of multiple CpGs or even individual CpG sites are often used to estimate the chronological age of cells, tissues, or individuals based on their DNA methylation levels, and are generally referred to as epigenetic age estimators or epigenetic clocks. The obtained estimated age is often referred to as DNAm age, or epigenetic age<sup>8</sup>, which is highly correlated with chronological age, but also affected by other biological factors<sup>13,19–22</sup> such as health status.

Typically developed using supervised machine learning methods, DNA methylation-based age estimators often employ penalized regression models. These models are designed to autonomously identify the CpGs that are most informative for estimating age<sup>8,23</sup>. However, the construction of a multi-tissue DNA methylation based age estimator is non-trivial, due to the significant differences between different tissues<sup>19,20</sup> and the distinct biological processes that drive the observed age-related hypermethylation and hypomethylation. The first multi-tissue DNA methylation-based age estimator became widely known as Horvath's clock<sup>6</sup> (proposed by Steve Horvath), which relied on elastic net regression that selected altogether 353 CpGs from the overall 27k CpG dinucleotides in the data it was trained on, corresponding to about 8,000 microarray samples collected from patients of all ages between children and elderly. Aside some limitations<sup>24,25</sup>, Horvath's clock proved to be a remarkably accurate age estimator in a variety of studies, yielding precise results for diverse DNA sources spanning the whole human lifespan<sup>8</sup>, e.g., together with other similar DNA-methylation-based clocks<sup>26–28</sup>, Horvath's clock was used to quantify the effectiveness of a program designed to regenerate the thymus, where the mean epigenetic age was 1.5 years younger than baseline after one year of treatment<sup>29</sup>. Possible relations between epigenetic aging and

<sup>1</sup>Danish Cancer Institute, Copenhagen, Denmark. <sup>2</sup>Department of Physics of Complex Systems, ELTE Eötvös Loránd University, Budapest, Hungary. <sup>3</sup>Biotech Research & Innovation Centre (BRIC), University of Copenhagen, Copenhagen, Denmark. <sup>4</sup>Department of Biological Physics, ELTE Eötvös Loránd University, Budapest, Hungary. <sup>5</sup>Health Services Management Training Centre, Semmelweis University, Budapest, Hungary. <sup>6</sup>Institute of Enzymology, HUN-REN Research Centre for Natural Sciences, Budapest, Hungary. ✉email: spisak.sandor@ttk.hu

the previously identified aging hallmarks are in the focus of on going research, and very recent related results have shown that although epigenetic aging is distinct from genomic instability, cellular senescence and telomere attrition, it is associated with nutrient sensing, mitochondrial activity and stem cell composition<sup>30</sup>.

With the advent of the overwhelming success of neural network-based techniques and deep learning methods in pattern recognition problems in general, it became another natural alternative to use these approaches for the estimation of biological age<sup>31–34</sup>. However, in spite of their high accuracy, the way neural networks make predictions about the age of input samples is difficult to interpret, and their operation is somewhat analogous to a “black box” method, where we have no explanation regarding why some methylation profiles are estimated to be older or younger compared to others. The need for interpretable neural network-based methods has risen also in the broader field of computational biology, and a very promising advancement in this direction was achieved by Elmarakeby et al.<sup>35</sup> by the introduction of a biologically informed deep learning tool for predicting the state of prostate cancer and evaluating molecular drivers of treatment resistance for therapeutic targeting. The suggested model used a huge collection of curated biological pathways to construct a pathway-aware multi-layered hierarchical deep learning network, thereby incorporating previously acquired biologically established hierarchical knowledge in a neural network language.

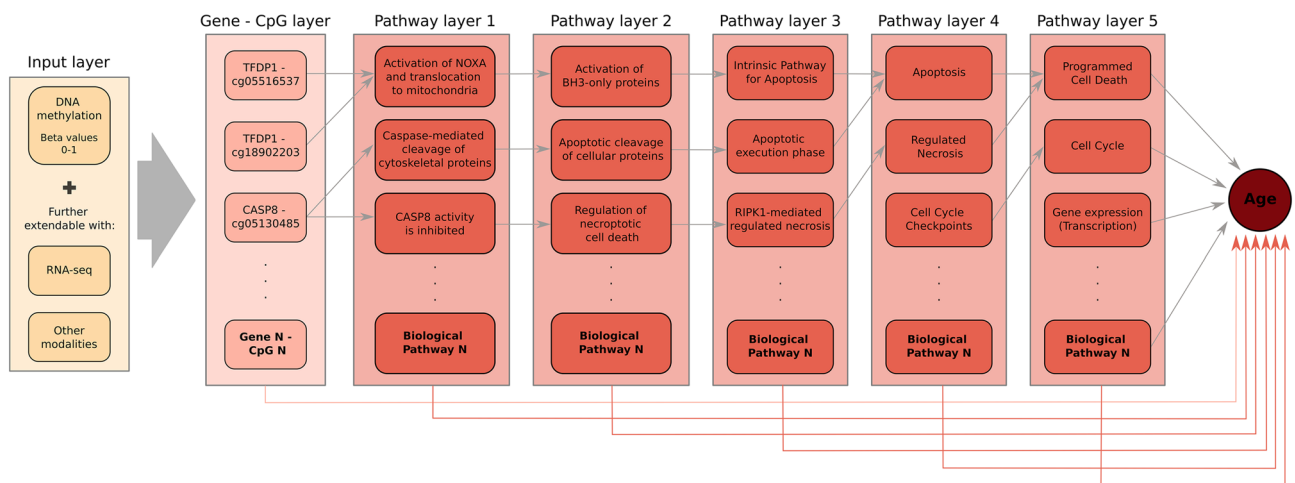
Inspired by this, here we propose a similar, biologically informed, explainable deep learning model for predicting the chronological age across multiple tissue types based on their methylation profiles. The structure of the neural network follows the hierarchy dictated by the biological pathways, in complete analogy with the tool presented by Elmarakeby et al.<sup>35</sup>. We compare the performance of the obtained method to that of elastic net regression in different use cases, including e.g., the data set by Gill et al.<sup>36</sup> related to the rejuvenation of fibroblast cells. According to these studies, beside a slight gain in the prediction precision, the most important benefit of our approach is given by the versatile possibilities for comparing the importance of different CpGs, genes, biological pathways or entire pathway branches and layers in predicting the age across the human lifespan.

## Results

### Explainable deep-learning age prediction model

We created a deep learning prediction model named XAI-AGE (XAI stands for Explainable AI) that integrates previously identified biologically hierarchical information in a neural network model for predicting the biological age based on DNA methylation data. The training of the model relied on the available chronological age of the patients in the training set. The construction of this pathway-aware multilayered hierarchical network was based on 3007 manually curated biological pathways parsed from the Reactome Pathway Knowledgebase<sup>37</sup>. The individual’s molecular profile as DNA methylation beta values was entered into the XAI-AGE model as input and spread across a layer of nodes representing a set of genes through weighted links. This input layer can be extended in a modular way to incorporate multiple data modalities, such as gene expression, gene mutation status or other measurable features representable on the gene level.

Subsequent layers of the network encode a collection of routes with increasing degrees of abstraction, representing complicated biological activities. The layers closer to the input layer correspond to finer biological pathways and deeper layers represent the higher levels of the hierarchy in the Reactome Pathway Knowledgebase as illustrated in Fig. 1. The connections between various layers are bound to follow known descendant-ascendant



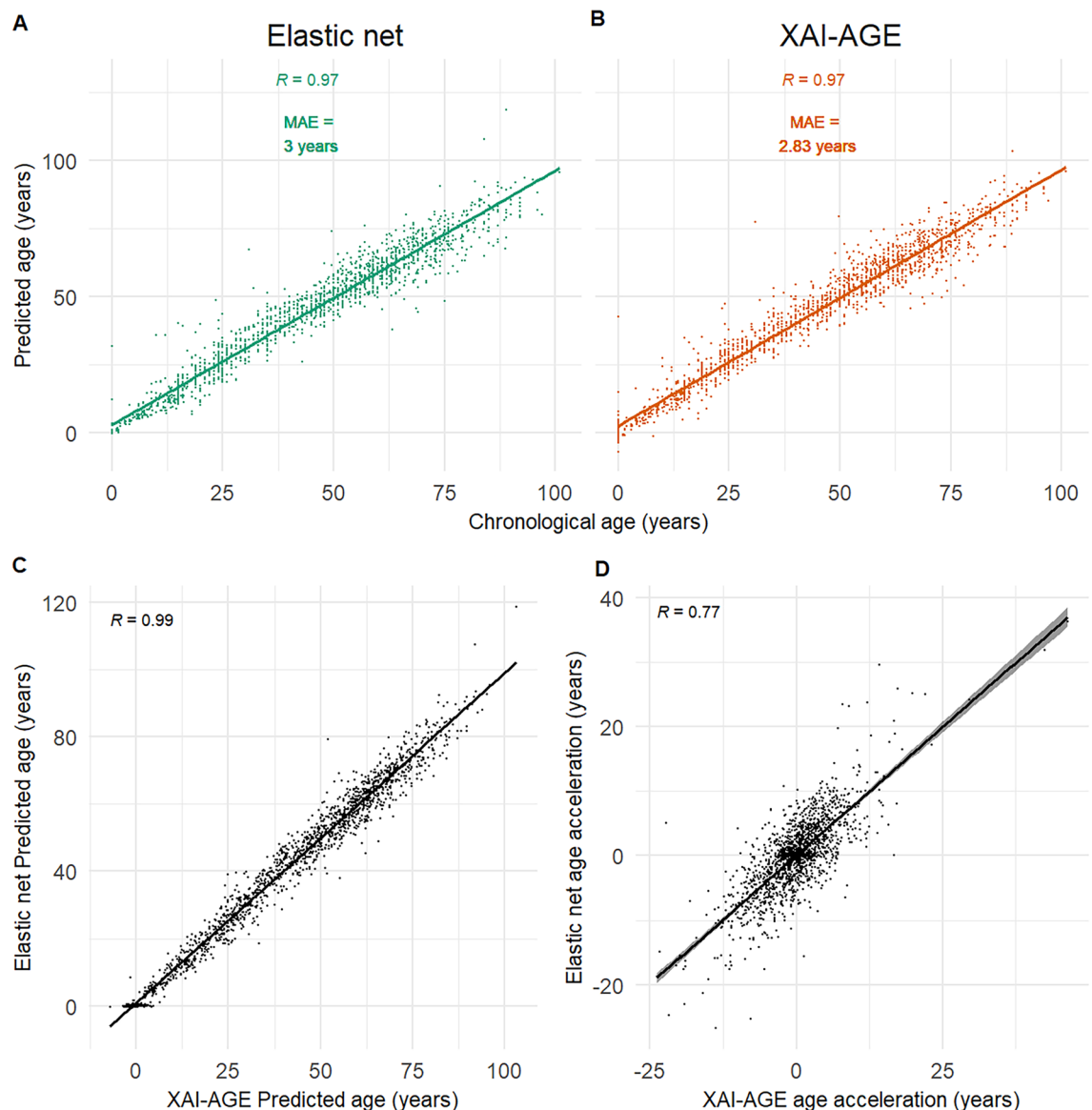
**Figure 1.** Schematic representation of the XAI-AGE model. The gene-CpG layer receives input data in the form of CpG methylation beta values. From here on, the information propagates in a restricted manner where nodes in the later layers are connected only if they are annotated jointly with the given node in the current layer according to the ReactomeDB. From left to right, the model’s abstraction becomes progressively more complex, as later levels in the neural network correspond to higher levels in the hierarchy defined by the ReactomeDB. The prediction of the chronological age for any given sample is given by the arithmetic mean of the outputs obtained for each individual layer. As indicated, the input layer can be easily extended by any new modalities which can be represented on the level of genes<sup>35</sup>.

relations among encoded properties, genes, and pathways, making the network interpretable by design. The architecture of the model is shown in more details in Supplementary Table S2 in the Supplementary Material.

To determine the relative importance of particular genes, pathways and biological processes contributing to the model prediction, we examined each layer and used the DeepLIFT<sup>38</sup> attribution approach to get the overall importance score of the neurons. Since the architecture is constrained by the underlying genes and biological processes, we can assume that the obtained importance scores can be used to test biological hypotheses across different subsets of the data. We note that the importance score is also a signed quantity, making it possible to infer trends in the dataset, however, the exact meaning of the direction is still not well understood. Hence, we included both positive and negative trends found during the analysis.

### Analysis of a pan-tissue data set

The XAI-AGE model was trained and first tested on a pan-tissue data set (details are given in the Methods), and for comparison, an elastic net regression model similar to Horvath's original regressor was also trained and evaluated on the same dataset. The performance of the two models was measured using the Pearson correlation coefficient and the median absolute error (MAE)<sup>39</sup>. As indicated in Fig. 2, we obtained 3 years MAE for the elastic net (Fig. 2A), and 2.83 years MAE for XAI-AGE (Fig. 2B) on the test set of the pan-tissue dataset, whereas the



**Figure 2.** Predicted age as a function of the chronological age. We show the scatter plot of the age estimated based on the methylation data according to elastic net regression (A) and according to XAI-AGE (B). The predicted age and the age acceleration (defined by subtracting the chronological age from the predicted age) for the elastic net as a function of the same quantity according to the XAI-AGE model is also shown in (C,D). The Pearson correlation coefficients and the median absolute errors are indicated beside the plots, the number of samples were  $n = 1619$ .

Pearson's correlation coefficient was 0.97 for both models. Furthermore, the two models showed high correlation with each other as well when considering either the predicted age (Fig. 2C) or the age acceleration (Fig. 2D), defined as the difference between the predicted age and the chronological age. To further validate the XAI-AGE model's performance, the results were replicated in a 5-fold cross-validation setting, where an artificial neural network, where all the neurons are connected between the layers (fully connected dense network) were trained as well and compared to XAI-AGE and the elastic net models (Supplementary Fig. S1). According to these tests, the MAE was significantly lower for XAI-AGE when compared to the elastic net model (Mann-Whitney U test,  $p$ -value = 0.028), while the dense fully connected neural network outperformed both models. However, it is important to note that the dense network contained more than 200 times more parameters. The neural network architecture for the fully connected dense model is shown on Supplementary Table S3.

The performance of the model was also examined by considering the various tissue types as displayed in Supplementary Fig. S2 in the Supplementary Material. By taking into consideration the varying amount of observations for the different tissue types, the results indicate that XAI-AGE provided the most accurate results for whole blood and blood PBMC tissue types, but performed poorly for blood cord, bone marrow, and esophagus.

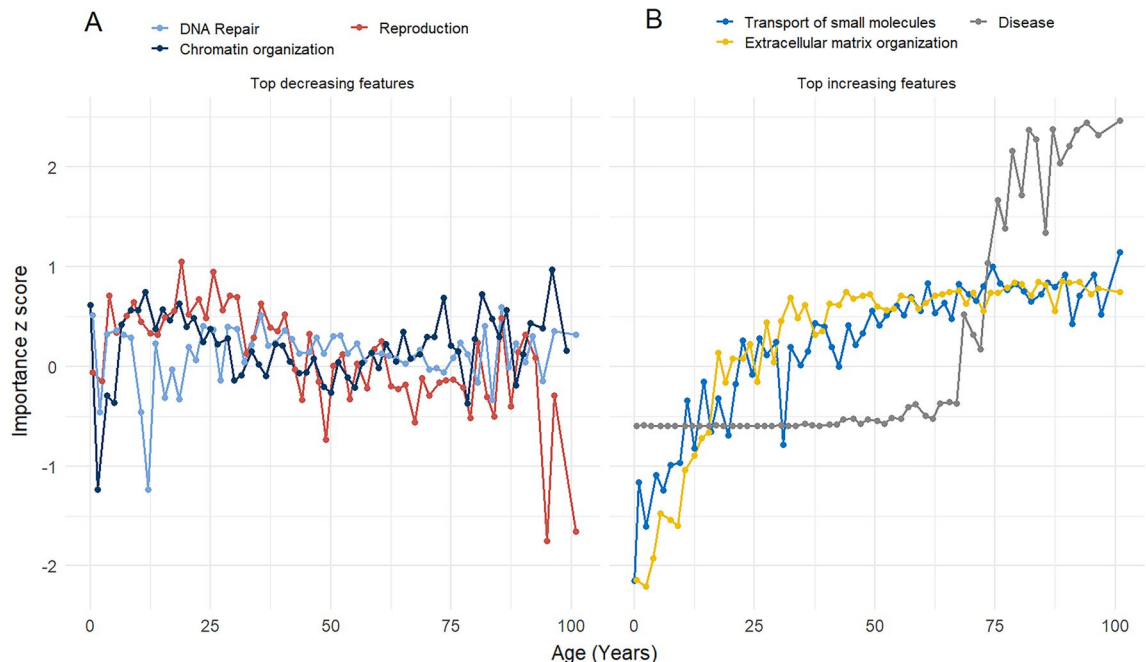
Next, we investigated the explainable representations that XAI-AGE learnt from the pan-tissue cohort. Using the DeepLIFT attribution approach<sup>38</sup>, the feature importance scores were retrieved from each layer and neuron in the model. The top six characteristics that exhibited the greatest change between the beginning and the end of the timeline (from chronological age zero to the maximum of the cohort) were further classified based on whether they caused a positive or negative trend.

In Fig. 3, we display the results for the last layer (corresponding to the top level in the hierarchy of the ReactomeDB), whereas similar plots for the other layers are presented in the Supplementary Material (Supplementary Figs. S3–S7).

From the features with a decreasing  $z$ -score over time, the top three features included the DNA Repair (R-HSA-73894), Chromatin organization (R-HSA-4839726) and the Reproduction (R-HSA-1474165) pathways. The top features where an increasing trend was observed in the  $z$ -score consisted of the Transport of small molecules (R-HSA-382551), Extracellular matrix organization (R-HSA-1474244) and a general pathway category called Disease (R-HSA-1643685). Interestingly, the latter exhibits a particular dynamics during the aging process, it remains constant until approximately the age of 70 then switches to a rapidly increasing tendency.

To demonstrate the advantages of XAI-AGE even further, a Plotly Dash graphical interface was built<sup>40</sup>, that renders Sankey plots similar to the one presented in Ref.<sup>35</sup>. This enables interactive navigation between the different layers of the network (each corresponding to a given level in the hierarchy of biological pathways according to the ReactomeDB), highlighting the features that contribute the most to the predictions (accessible at: <https://k8plex-krft.vo.elte.hu/notebook/report/xgrp0j-sankeymethy/>).

Since the links in this network indicate that the given pair of nodes are annotated to be related according to the ReactomeDB, one can track the flow of information between the layers, and infer the relevant sources that



**Figure 3.** Standardized importance score as a function of chronological age. We show the  $z$ -score based on the distribution of the importance scores for the neurons (features) in the last layer of the network (corresponding to the top level in the hierarchy of the ReactomeDB) with the largest change in the  $z$ -score across the chronological age. The top 3 features where the trend in the  $z$ -score is negative are displayed in (A), whereas the top 3 features where the  $z$ -score is increasing with age are shown in (B).

contributed to the prediction. As an illustration, in Fig. 4, we show the layer-wise standardized and ordered importance score for the samples in the pan-tissue dataset.

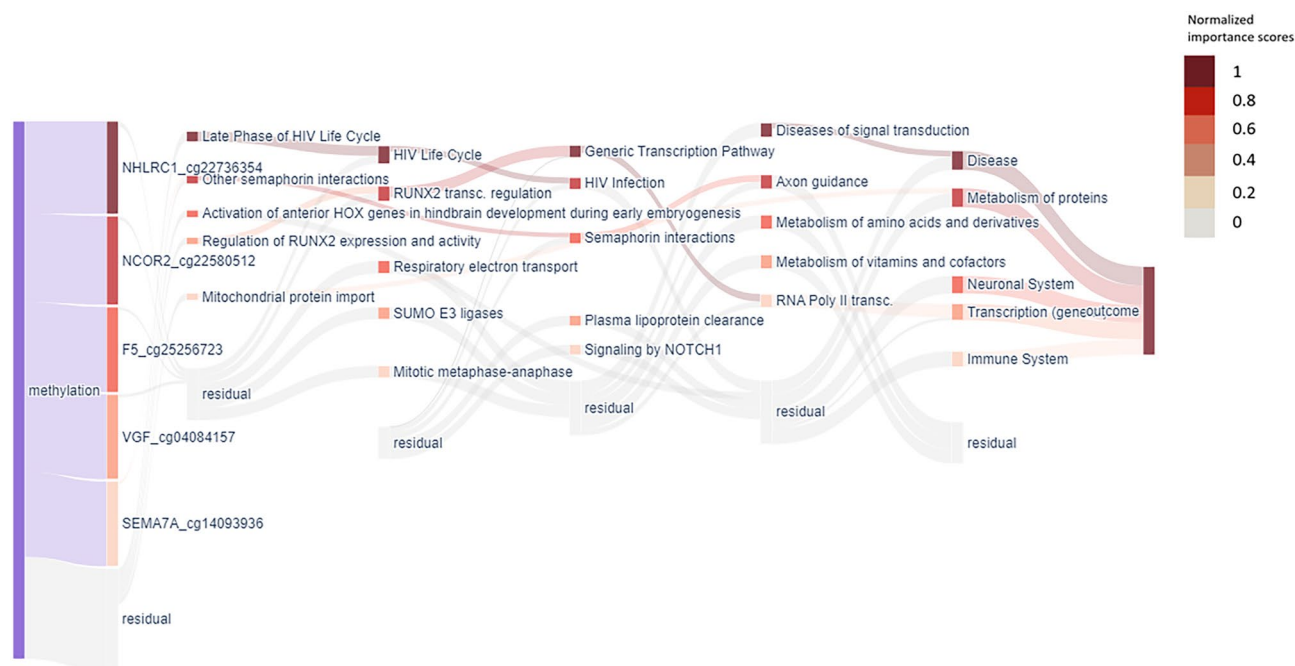
### Measuring the biological age during fibroblast reprogramming

We also applied XAI-AGE to estimate the biological age of dermal fibroblast cells derived from middle age donors used in the reprogramming study by Gill et al.<sup>36</sup>. In this study, the cells were harvested for DNA methylation and RNA-sequencing in different time points during the reprogramming process. In the present study we used the methylation data for calculating the biological age predicted by XAI-AGE for both the treated and the (non-treated) control cells. According to Fig. 5, our age estimation framework gave results similar to that obtained by the Horvath clock-like elastic net. Both epigenetic clocks precisely predicted the biological age of the cells in the negative control and failed to reprogram group, as well as a significant drop for the transiently reprogrammed cells. However, according to the original study by Gill et al.<sup>36</sup>, the methylation levels go significantly down across all gene groups for these cells, which could provide a simple explanation for this effect. As anticipated, the predicted biological age of the iPSC cells was close to zero. Interestingly, the predicted biological age of the negative control cells shows a positive trend in time, consistent with the recent findings by Levine et al.<sup>41</sup>.

Similarly to the analysis of the pan-tissue cohort, we calculated the importance scores for both the individual neurons and all the layers. This allowed the study of the importance in the age prediction of the different features and biological pathways during the reprogramming process. The results for the last layer in the neural network (highest level in the biological pathway hierarchy) are displayed in Fig. 6, showing the top six features according to the magnitude of the change over time for the negative controls and unsuccessfully reprogrammed cells (Fig. 6A), as well as for the transiently reprogrammed cells (Fig. 6B). The time-dependent dynamics of the importance scores shows interesting differences between the negative controls and the reprogrammed cells, e.g., the Metabolism of proteins (R-HSA-392499) and Muscle contraction (R-HSA-397014) significantly changed in the negative direction in the transiently reprogramming group, while the Chromatin organization (R-HSA-4839726) and Circadian clock (R-HSA-400253) increased. The extracted importance scores from the other layers of the model can be seen on Supplementary Figs. S8–S12.

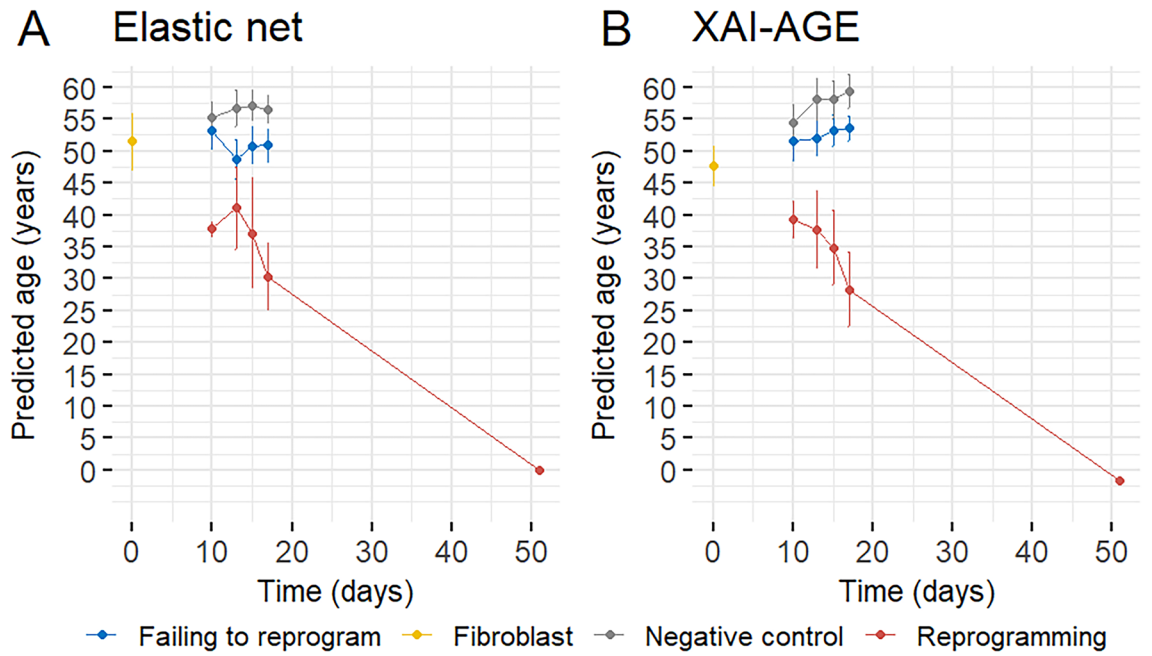
### Biological age in umbilical cord plasma transfusion

As a further application of XAI-AGE, we also analysed a recently published dataset by Clement et al.<sup>42</sup>, related to umbilical cord plasma transfusion. Heterochronic parabiosis studies have shown favorable benefits in aged animals getting youthful blood across a variety of tissues<sup>43</sup>. The study presented in Ref.<sup>42</sup>, examined whether infusion of plasma or plasma-derived factors from young donors could be used to mitigate human age-related conditions by administering human umbilical cord plasma concentrate to elderly patients (n = 18, mean age = 74) and monitoring epigenetic age-related measures for a period of 10 weeks. The authors have shown that the

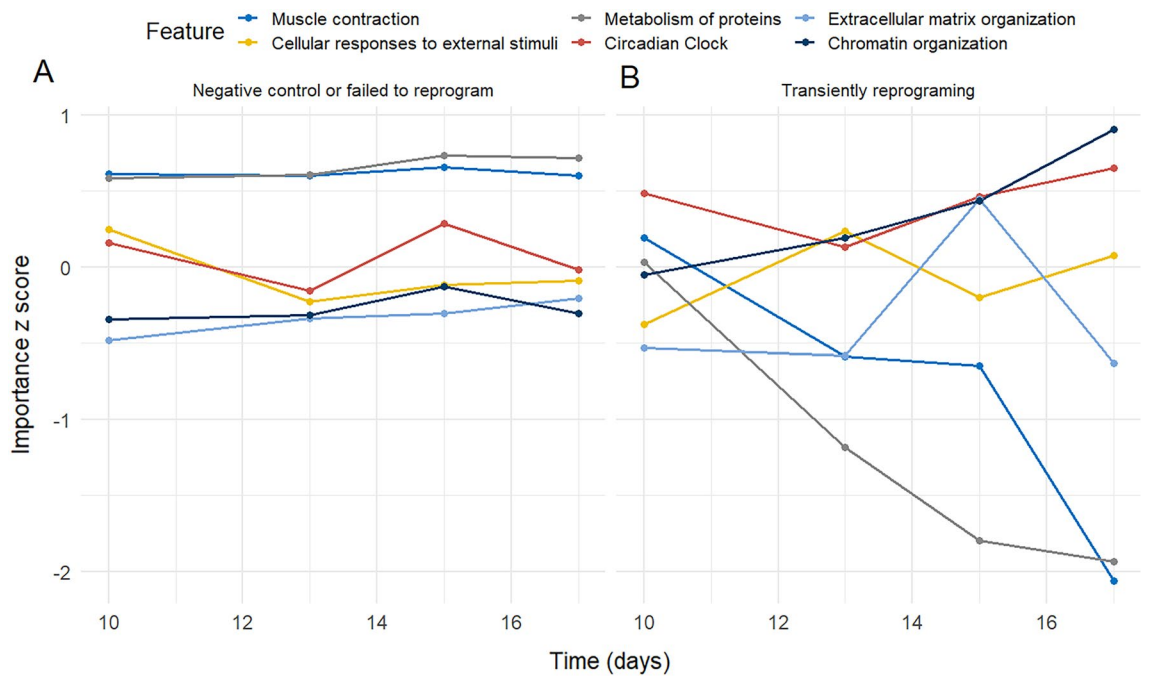


**Figure 4.** Sankey diagram of the top features calculated across all the samples in the pan-tissue dataset. The visualization of XAI-AGE layer structure shows the normalized relative importance score difference between the old (>65 years) and young samples for various nodes inside each layer. Darker colours indicate a larger difference between the importance scores for the two age groups at the given node. Only the top 5 nodes in each layer are displayed here, where the other nodes are indicated by the single semi-transparent nodes (named residual) at the bottom for each layer.





**Figure 5.** Comparing the biological age during the time-dependent differentiation process of human fibroblast cells from three donors. At time zero, one can observe undifferentiated fibroblast cells, and as time progresses, the biological age as determined by (A) elastic net and (B) XAI-AGE models are calculated. The cells can be further subdivided into categories that include fibroblasts utilized as negative controls, cells that failed to be transiently reprogrammed, and cells that were successfully transiently reprogrammed. The differentiation endpoint of induced pluripotent stem cells is also displayed on the final day of the measurement. The error bars represent the standard deviation across the three donors. A small positive trend can be also observed in the negative control group for the predicted age with slightly better correlation for XAI-AGE ( $R=0.38$ ,  $p=0.0038$ ) than for elastic net ( $R = 0.2$ ,  $p = 0.23$ ).



**Figure 6.** Time dependence of the standardised importance score values in the human fibroblast experiment. We show the six scores from the last layer of the network where the largest deviation can be observed between the negative control or the failed to reprogram group (A) and the transiently reprogrammed group (B).

treatment lowered DNA methylation-based GrimAge measure by an average of 0.82 years, indicating a decrease in the risk of morbidity and mortality. However, other epigenetic clocks that estimate chronological age did not detect a significant age-reversal effect.

In the present work, using this data, we first estimated the chronological age of the individuals using XAI-AGE. The comparison between the predicted age and the chronological age stratified by the pre-treatment and post-treatment samples is shown in the Supplementary Material in Supplementary Fig. S13, indicating a high correlation between the two variables. Next, we compared the age acceleration (corresponding to the difference between the estimated value and the actual chronological age) predicted by XAI-AGE between the two groups of samples derived from the same individuals, similarly as was described in<sup>42</sup>. A paired t-test was performed and reported no significant changes.

Furthermore, the importance score for each feature in each layer was extracted and compared between the pre-treatment and post-treatment groups. In Supplementary Figs. S14–S19, the six top features from the last layer according to the magnitude of the difference between the two groups are shown, of which three correspond to the top features where this difference is positive, and the other three are the top features where the difference is negative. Our results indicate that the Cell-cycle (R-HSA-1640170), Cell-Cell communication (R-HSA-1500931) and the Reproduction (R-HSA-1474165) pathways were more important in the post-treatment samples, while the Circadian clock (R-HSA-400253), Mitophagy (R-HSA-5205647) and the Vesicle-mediated transport (R-HSA-5653656) pathways were more important in the pre-treatment group. Overall, the XAI-AGE results are less informative for this data set that may indicate that either the input data is not robust enough or may indicate weak points of XAI-AGE.

More extended data analyses of the results and the comparison of the importance scores from the other layers of the network are described in the Supplementary Materials.

## Discussion

In this paper, we present an accurate and explainable neural network architecture allowing not only the estimation of age based on DNA methylation data with high precision but also the easy interpretation of results that are comparable across tissues, age groups, and differentiation processes in the case of cell lines. The resulting model can be used to generate hypotheses and visualize the underlying mechanisms connected to aging. We have demonstrated this feature of the model by examining the importance scores of the individual neurons in predicting the age when the neural network was trained on different datasets. In this aspect, probably the most noteworthy result was obtained for the pan-tissue dataset, where the standardised importance score for the Disease pathway (corresponding to a neuron in the last layer of the neural network) displayed a particular behaviour when plotted as a function of age, showing a roughly constant flat curve that is replaced by a rapidly increasing function at the age of 70.

The second important observation is related to the DNA Repair pathway, which demonstrated a decreasing tendency in the pan-tissue cohort when the importance *z*-score was visualized as a function of age (Fig. 3A). The DNA repair pathway is part of the DNA damage response system that is responsible for the maintenance of genome integrity. Living organisms are constantly exposed to exogenous and endogenous DNA damage. Unrepaired or faulty repair of DNA damage leads to the accumulation of somatic mutations as an organism ages, making genome instability a hallmark of aging<sup>1</sup>. The importance of DNA repair mechanisms to counteract the time- and exposure-dependent accumulation of DNA damage is highlighted by the fact that inherited mutations in genes that are involved in these pathways underlie several segmental premature ageing-like syndromes in humans<sup>44</sup>. Our result is in agreement with accumulating evidence suggesting that the integrity and maintenance of the genome are strongly associated with aging<sup>45,46</sup>. The Chromatin organization pathway was also selected as one of the top decreasing features in the last layer of the network based on the change in the importance *z*-score across the chronological age of the individuals (Fig. 3A). The Chromatin organization pathway includes chromatin modifying enzymes involved in processes that result in the specification, formation or maintenance of the physical structure of eukaryotic chromatin. The identification of this pathway as one of the top features in the XAI-AGE network is coherent with the well-established fact that epigenetic changes affecting DNA methylation patterns, histone modifications and chromatin remodeling are the hallmarks of ageing<sup>1</sup>.

Biological pathways that demonstrated the largest difference between the importance scores of the old (> 65 years) and young samples are shown in Fig. 4. In the third layer, one of the top 5 nodes is the Mitotic metaphase and anaphase pathway that regulate the proper segregation of chromosomes into daughter cells. Recently, several epigenetic mitotic clocks were developed, such as epiTOC<sup>47</sup>, epiTOC2<sup>48</sup> and solo-WCGW<sup>49</sup>. epiTOC and epiTOC2 rely on CpG sites in CpG-rich regions that are marked by the polycomb repressive complex 2 (PRC2), which are generally unmethylated across numerous different fetal tissue types, to calculate the rate of stem cell division<sup>47,48</sup>. On the other hand, solo-WCGW focuses on DNA methylation loss at partially methylated domains (PMDs) that showed increased hypomethylation with age and appeared to track the accumulation of cell divisions<sup>49</sup>. It seems that the identification of the Mitotic metaphase and anaphase pathway as significantly different between old and young individuals by the XAI-AGE model captures a different association between mitotic processes and ageing than the previously described epigenetic mitotic clocks since we did not identify overlapping genes between the Mitotic metaphase and anaphase pathway and the described epigenetic mitotic models. However, a detailed analysis of these interesting findings is to be explored in further studies.

We calculated the standardised importance scores from the last layer of the XAI-AGE model using the data from a fibroblast rejuvenation experiment<sup>36</sup>. The largest difference between the negative control or failed to reprogram group and the transiently reprogrammed group were observed in six biological pathways (Fig. 6). Among these pathways are the Extracellular matrix organization and the Muscle contraction pathways that likely to reflect the observations made by Gill and colleagues that the reprogrammed fibroblasts produced youthful levels of

collagen proteins, and showed partial functional rejuvenation of their migration speed<sup>36</sup>. Interestingly, the Circadian clock pathway and several known associated pathways, such as the Cellular response to external stimuli, Chromatin organization and Metabolism of proteins, were also identified as important by the XAI-AGE model in the fibroblast reprogramming process during which the DNA methylation age measured by the multi-tissue epigenetic clock was significantly decreased<sup>36</sup>. The circadian clock is an endogenous, biological timing mechanism that responds to several external stimuli to maintain the synchronization of internal biological processes among themselves and with exogenous environmental cycles<sup>50</sup>. The core clock genes, including *CLOCK1*, *BMAL1*, *PER* and *CRY* genes, are rhythmically expressed and form a negative feedback loop that drives circadian oscillations.

The underlying transcription-translation feedback system of the circadian clock regulates the expression of clock-controlled genes that are involved in various processes, e.g., metabolism and chromatin remodelling<sup>51</sup>. A growing body of evidence suggests a link between the disruption of the circadian rhythms and ageing. Studies have shown that disturbances in the circadian clock and sleep homeostasis are linked to increased incidence of a variety of age-related health problems, such as neurodegenerative diseases, metabolic disorders, cardiovascular disease, obesity and cancer<sup>52–55</sup>. Furthermore, the transcription factor *BMAL1*, which is the co-activator of the circadian clock, exhibited decreased regulatory activity with age independently from cell-type and tissue-type<sup>56</sup>.

According to the chrono-epigenetic theory, circadian oscillations of cytosine modification at specific CpG sites are robust in young individuals but diminish with age, potentially as a result of changed activity of ten-eleven translocation (TET) and DNA methyltransferase (DNMT) maintenance enzymes. Age-related changes in amplitudes of the oscillations precede linear DNA methylation changes and might predict age-dependent linear outcomes<sup>57</sup>. Our results suggest that the synchronization of oscillatory rhythms of internal biological processes is associated not only with ageing but also with rejuvenation of human cells by maturation phase transient reprogramming.

Additional advantages of the model include the modular construction of the underlying neural network: the input layers can be modified to incorporate additional modalities, allowing the integration of multiomics data, as demonstrated in an analogous example by Gill et al.<sup>36</sup>. In the case of age prediction, the logical next step would be to include RNA-seq data alongside DNA methylation values in the model. This can be easily accomplished by vertically increasing the input layer in the model and making the new data modalities representable at the level of genes. This modularity applies to the deeper layers in the model (corresponding to higher levels in the pathway hierarchy according to the ReactomeDB) as well. The core structure in the Reactome Pathway can be freely altered, enlarged, or replaced by another database. Along this line, the incorporation of the so-called Hallmarks of Aging<sup>1</sup> into the interaction network to make it more aging-specific is an intriguing study topic for the future.

There are also some limitations to our analysis, e.g., compared to other deep learning based models like DeepMAge optimised solely for prediction accuracy, XAI-AGE performs worse by around half a year MAE<sup>34,58</sup>. Regarding our training data, there are substantial class imbalances of tissues and age groups, and batch effects from the various data sources that may be included can potentially bias the results. A curation bias can also alter the results of the Reactome Pathway Database, which is another issue. For instance, the HIV pathway was over-represented in our data, which may play little role in predicting the biological age because the same critical age-related genes are present in several pathways and the neural network amplifies the value of these neurons for better prediction. Pre-training the neural network on the CpG data level (for instance, by changing the architecture to an Autoencoder) and then fine-tuning it to predict the biological age is a potential solution to this issue.

Clarifying the causal relationship between the many CpG-s, genes, and biological processes associated with aging would be a future goal for biologically informed deep learning approaches. The primary advantage of XAI-AGE over other epigenetic clocks is the direct comparison and inference of relationships between more abstract data layers than using raw input data alone. Further supplementation of the model with additional biological data modalities, such as incorporating RNA-seq at the gene level or evaluating the data as a time series, as we demonstrated with the fibroblast reprogramming dataset, could facilitate the future discovery of causal relationships. Using XAI-AGE could assist by analyzing computationally its interpretable network, or by domain experts using the Sankey diagram interactive visualization.

## Methods and data

### Applied data sources in this study

All data used in the study is publicly available. The complete list of the data sources are shown in Supplementary Table S1. All methods were carried out in accordance with relevant guidelines and regulations.

### Training the model on the pan-tissue data set

We trained and tested XAI-AGE with a set of 6547 patient samples across 54 cohorts and multiple tissues (Supplementary Table S1), divided into 75% training, 25% testing, to predict the chronological age based on the DNA methylome of the individuals. This estimation was later used to also infer the biological age, defined by the chronological age prediction of the model.

### Fibroblast cell reprogramming data

In the study by Gill et al.<sup>36</sup> the cells were harvested for DNA methylation and RNA-sequencing in different time points during the reprogramming process. Altogether 96 cells were analysed during the study from three different individuals which can be further subdivided into the categories of cells that were measured prior to the reprogramming phase (fibroblasts), negative controls that received mock treatment, cells that failed to reprogram, and cells that transiently reprogrammed successfully. Cells that had been fully reprogrammed (iPSC) and were sampled on the final day were also measured.



## Umbilical cord plasma transfusion data

The dataset contains 36 whole-blood samples collected at the beginning and at the end of the 10-week experiment period. In the present study, we used the already trained XAI-AGE model to estimate the biological age and biological age acceleration for each sample, and the results were compared between the pre-treatment and post-treatment groups.

## Data and materials availability

All data used in the study were downloaded from the Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo/>) and from the The Cancer Genome Atlas data portal (<https://portal.gdc.cancer.gov/>) databases. The corresponding dataset ID-s are listed in the supplementary information file. Any data not public can be requested.

## Code availability

The code for running inferences with the XAI-AGE model can be accessed at: <https://github.com/Paureel/XAI-AGE>.

Received: 17 January 2023; Accepted: 20 December 2023

Published online: 15 January 2024

## References

- López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The hallmarks of aging. *Cell* **153**(6), 1194–1217 (2013).
- Baker, G. & Sprott, R. Biomarkers of aging. *Exp. Gerontol.* **23**, 223–239 (1988).
- Warner, H. R. The future of aging interventions. *J. Gerontol. A* **59**, B692–B696 (2004).
- Jylhävä, J., Pedersen, N. L. & Hägg, S. Biological age predictors. *EBioMedicine* **21**, 29–36 (2017).
- Field, A. E., Wang, T., Havas, A., Ideker, T. & Adams, P. D. Dna methylation clocks in aging: Categories, causes, and consequences. *Mol. Cell* **71**, 882–895 (2018).
- Horvath, S. Dna methylation age of human tissues and cell types. *Genome Biol.* **14**, R115 (2013).
- Lee, H. Y., Lee, S. D. & Shin, K.-J. Forensic DNA methylation profiling from evidence material for investigative leads. *BMB Rep.* **49**, 359–369 (2016).
- Horvath, S. & Raj, K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat. Rev. Genet.* **19**, 371–384 (2018).
- Berdyshev, G., Korotaev, G., Boiarskikh, G. & Vaniushin, B. Nucleotide composition of DNA and RNA from somatic tissues of humpback and its changes during spawning. *Biokhimiia* **31**, 988–993 (1967).
- Ahuja, N., Li, Q., Mohan, A. L., Baylin, S. B. & Issa, J. P. Aging and DNA methylation in colorectal mucosa and cancer. *Cancer Res.* **58**, 5489–5494 (1998).
- Fraga, M. F. & Esteller, M. Epigenetics and aging: The targets and the marks. *Trends Genet.* **23**(8), 413–418 (2007).
- Bollati, V. *et al.* Decline in genomic DNA methylation through aging in a cohort of elderly subjects. *Mech. Ageing Dev.* **130**(4), 234–239 (2009).
- Christensen, B. C. *et al.* Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CPG island context. *PLoS Genet.* **5**, e1000602 (2009).
- Rodríguez-Rodero, S., Fernández-Morera, J., Fernandez, A., Menéndez-Torre, E. & Fraga, M. Epigenetic regulation of aging. *Discov. Med.* **10**, 225–233 (2010).
- Mugatroyd, C., Yonghe, W., Bockmühl, Y. & Spengler, D. The Janus face of DNA methylation in aging. *Ageing* **2**(2), 107–110 (2010).
- ...Teschendorff, A. E. *et al.* Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res.* **20**(4), 440–446 (2010).
- Bell, J. T. *et al.* Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PLoS Genet.* **8**, e1002629 (2012).
- Zheng, S. C., Widschwendter, M. & Teschendorff, A. E. Epigenetic drift, epigenetic clocks and cancer risk. *Epigenomics* **8**(5), 705–719 (2016).
- Bernstein, B. E. *et al.* The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.* **28**, 1045–1048 (2010).
- ...Li, Y. *et al.* The DNA methylome of human peripheral blood mononuclear cells. *PLoS Biol.* **8**(11), 1–9 (2010).
- Thompson, R. F. *et al.* Tissue-specific dysregulation of DNA methylation in aging. *Ageing Cell* **9**(4), 506–518 (2010).
- Baubec, T. & Schübeler, D. Genomic patterns and context specific interpretation of DNA methylation. *Curr. Opin. Genet. Dev.* **25**, 85–92 (2014).
- Palla, G. *et al.* Hierarchy and control of ageing-related methylation networks. *PLoS Comput. Biol.* **17**(9), e1009327. <https://doi.org/10.1371/journal.pcbi.1009327> (2021).
- ...Horvath, S. *et al.* Epigenetic clock for skin and blood cells applied to Hutchinson Gilford progeria syndrome and ex vivo studies. *Ageing* **10**, 1758–1775 (2018).
- ...Zhang, Q. *et al.* Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing. *Genome Med.* **11**, 54 (2019).
- Hannum, G. *et al.* Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol. Cell* **49**(2), 359–367 (2013).
- Levine, M. E. *et al.* An epigenetic biomarker of aging for lifespan and healthspan. *Ageing* **10**, 573–591 (2018).
- Lu, A. T. *et al.* DNA methylation Grimace strongly predicts lifespan and Healthspan. *Ageing* **11**, 303–327 (2019).
- Fahy, G. M. *et al.* Reversal of epigenetic aging and immunosenescent trends in humans. *Ageing Cell* **18**(6), e13028 (2019).
- Kabacik, S. *et al.* The relationship between epigenetic age and the hallmarks of aging in human cells. *Nat. Aging* **2**, 484–493 (2002).
- Aliferi, A. *et al.* DNA methylation-based age prediction using massively parallel sequencing data and multiple machine learning models. *Forensic Sci. Int.* **37**, 215–226 (2018).
- Galkin, F. *et al.* Human gut microbiome aging clock based on taxonomic profiling and deep learning. *iScience* **23**(6), 101199 (2020).
- Levy, J. J. *et al.* Methylnet: An automated and modular deep learning approach for DNA methylation analysis. *BMC Bioinform.* **21**, 108 (2020).
- Galkin, F., Mamoshina, P., Kochetov, K., Sidorenko, D. & Zhavoronkov, A. Deepmage: A methylation aging clock developed with deep learning. *Ageing Dis.* **12**, 1252–1262 (2021).
- Elmarakeby, H. *et al.* Biologically informed deep neural network for prostate cancer discovery. *Nature* **598**, 1–5 (2021).
- Gill, D. *et al.* Multi-omic rejuvenation of human cells by maturation phase transient reprogramming. *Elife* **11**, e71624 (2021).
- Fabregat, A. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res.* **46**, 11 (2017).
- Shrikumar, A., Greenside, P. & Kundaje, A. *Learning Important Features Through Propagating Activation Differences* (2017).

39. McEwen, L. *et al.* The pedbe clock accurately estimates dna methylation age in pediatric buccal cells. *Proc. Natl. Acad. Sci. USA* **117**, 201820843 (2019).
40. Hossain, S. *Visualization of Bioinformatics Data with Dash Bio* 126–133 (2019).
41. Minteer, C. *et al.* Revisiting the bad luck hypothesis: Cancer risk and aging are linked to replication-driven changes to the epigenome. *bioRxiv* <https://doi.org/10.1101/2022.09.14.507975> (2022).
42. Clement, J. *et al.* Umbilical cord plasma concentrate has beneficial effects on DNA methylation grimage and human clinical biomarkers. *Aging Cell* **09**, e13696 (2022).
43. Conboy, I. *et al.* Rejuvenation of aged progenitor cells by exposure to a young systemic environment. *Nature* **433**, 760–764 (2005).
44. Hoeijmakers, J. H. J. DNA damage, aging, and cancer. *N. Engl. J. Med.* **361**(15), 1475–1485 (2009).
45. Schumacher, B., Pothof, J., Vijg, J. & Hoeijmakers, J. H. J. The central role of DNA damage in the ageing process. *Nature* **592**(7856), 695–703 (2021).
46. Melzer, D., Pilling, L. C. & Ferrucci, L. The genetics of human ageing. *Nat. Rev. Genet.* **21**(2), 88–101 (2020).
47. Yang, Z. *et al.* Correlation of an epigenetic mitotic clock with cancer risk. *Genome Biol.* **17**, 205 (2016).
48. Teschendorff, A. E. A comparison of epigenetic mitotic-like clocks for cancer risk prediction. *Genome Med.* **12**, 56 (2020).
49. Zhou, W. *et al.* Dna methylation loss in late-replicating domains is linked to mitotic cell division. *Nat. Genet.* **50**, 591–602 (2018).
50. Benitah, S. A. & Welz, P. S. Circadian regulation of adult stem cell homeostasis and aging. *Cell Stem Cell* **26**, 817–831 (2020).
51. Takahashi, J. S. Transcriptional architecture of the mammalian circadian clock. *Nat. Rev. Genet.* **18**, 164–179 (2017).
52. Masri, S. & Sassone-Corsi, P. The emerging link between cancer, metabolism, and circadian rhythms. *Nat. Med.* **24**, 1795–1803 (2018).
53. Reinke, H. & Asher, G. Crosstalk between metabolism and circadian clocks. *Nat. Rev. Mol. Cell Biol.* **20**, 227–241 (2019).
54. Patke, A., Young, M. W. & Axelrod, S. Molecular mechanisms and physiological importance of circadian rhythms. *Nat. Rev. Mol. Cell Biol.* **21**, 67–84 (2020).
55. Nassan, M. & Videnovic, A. Circadian rhythms in neurodegenerative disorders. *Nat. Rev. Neurol.* **18**, 7–24 (2022).
56. Maity, A. K., Hu, X., Zhu, T. & Teschendorff, A. E. Inference of age-associated transcription factor regulatory activity changes in single cells. *Nat. Aging* **2**, 548–561 (2022).
57. Oh, E. S. & Petronis, A. Origins of human disease: The chrono-epigenetic perspective. *Nat. Rev. Genet.* **22**, 533–546 (2021).
58. de Lima Camillo, L. P., Lapiere, L. R. & Singh, R. A pan-tissue dna-methylation epigenetic clock based on deep learning. *npj Aging* **8**(1), 4 (2022).

## Acknowledgements

Supported by the the European Union project RRF-2.3.1-21-2022-00004 within the framework of the MILAB Artificial Intelligence National Laboratory. G.P. received funding partly from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 101021607 and from the National Research, Development and Innovation Office under grant no. K128780. S.S. received funding from National Research Development and Innovation Office Hungary, under grant no. FK142835.

## Author contributions

I.C., G.P., S.S. and A.P. devised the project, the main conceptual ideas and proof outline. O.P. performed the data collection. A.P. implemented the model, processed the experimental data, performed the analysis and designed the figures. A.P. and G.P. drafted the manuscript. J.B., Z.S. and S.S. aided in interpreting the results and worked on the manuscript. All authors provided critical feedback and helped shape the research, the analysis, and the manuscript. All authors reviewed the manuscript.

## Funding

Open access funding provided by HUN-REN Research Centre for Natural Sciences.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-023-50495-5>.

**Correspondence** and requests for materials should be addressed to S.S.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024