

Development and Application of Aging Clocks

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln



vorgelegt von

David Helmut Meyer

aus Cloppenburg, Deutschland

Köln, 2024

The work described in this dissertation was conducted from September 2018 to May 2024 under the supervision of Prof. Dr. Björn Schumacher at the Institute for Genome Stability in Ageing and Disease, CECAD research center, University of Cologne.

Chapter 2 and 3 of this thesis have been published:

1. Meyer, D. H. & Schumacher, B. BiT age: A transcriptome-based aging clock near the theoretical limit of accuracy. *Aging Cell* 20, e13320 (2021).
2. Meyer, D. H. & Schumacher, B. Aging clocks based on accumulating stochastic variation. *Nature Aging* (2024)

Chapter 4 is a manuscript in preparation with the working title “Neuron-type specific aging-rate reveals age decelerating interventions preventing neurodegeneration” and is currently submitted.

Contents

Summary	1
1 Introduction.....	3
1.1 Aging Theories.....	3
1.1.1 Programmed Aging Theories.....	3
1.1.2 Evolutionary Aging Theories.....	3
1.1.3 Damage Accumulation Theories of Aging	4
1.2 Biological Age Prediction.....	5
1.2.1 Hochschild’s Method.....	6
1.2.2 Principal Component Analysis (PCA).....	7
1.2.3 Klemra and Doubal’s Method (KDM)	7
1.2.4 PhenoAge	9
1.2.5 Homeostatic Dysregulation	9
1.2.6 DNA Methylation Aging Clocks.....	10
1.2.7 Transcriptomic Aging Clocks.....	12
1.2.8 Proteomic Aging Clocks	13
1.2.9 Other Aging Clocks	13
1.3 Aims of this Thesis.....	14
2 BiT age: A transcriptome-based aging clock near the theoretical limit of accuracy.....	15
3 Aging clocks based on accumulating stochastic variation.....	16
4 Neuron-type specific aging-rate reveals age decelerating interventions preventing neurodegeneration	17
5 Discussion.....	18

5.1	Aging Clocks.....	18
5.1.1	Constructing and Validating Accurate Biological Aging Clocks	18
5.1.2	Understanding Their Underlying Mechanisms.....	21
5.1.3	Utilization in Identifying and Evaluating Longevity Interventions	22
5.2	Stochastic Biological Variation	23
5.2.1	Stochastic Epigenetic Variation	23
5.2.2	Stochastic Transcriptomic Variation.....	27
5.3	Cross-Species Transcriptome Comparisons.....	32
6	Conclusion	34
7	References.....	35
8	Acknowledgements	60
9	Additional Contributions	62

Summary

Aging clocks have emerged as powerful tools in the field of aging biology. These clocks utilize various biomarkers to estimate the biological age, overall health status, and pace of aging of an organism. Unlike chronological age, which measures linearly the time elapsed since birth, biological age considers factors such as genetics, lifestyle, and environmental exposure that affect the aging process and lead to inter-individual differences. By providing an assessment of an organisms' health status, aging clocks can aid personalized healthcare, and accelerate aging research by giving surrogate endpoints in clinical trials for the identification and evaluation of geroprotective interventions. Currently, the field primarily focuses on three key areas of research: 1.) the search and validation of accurate biological aging clocks is still ongoing, with various clocks being built for different species and data modalities. 2.) The underlying mechanisms and interpretation of aging clocks is under debate with no clear consensus on what aging clocks are measuring. 3.) And lastly, the use of aging clocks in the identification and evaluation of geroprotective interventions. This is currently largely constrained to their use as surrogate endpoints of clinical trials, limiting their applicability.

In this thesis, we first developed an accurate transcriptomic aging clock based on the novel concept of binarization (BitAge). Transcriptomic aging clocks faced limitations due to the inherent variability and age-dependent increase in variation of transcriptomic data, leading to their relative underperformance compared to epigenetic aging clocks. Here, I show that binarizing transcriptomic data enables the usage of transcriptomic data for training aging clocks that rival epigenetic aging clocks. Leveraging existing lifespan data for the nematode *Caenorhabditis elegans* for temporal rescaling, moreover, allowed highly accurate predictions, not only of the chronological, but especially the biological age.

In the second part of this thesis, we investigated the underlying mechanisms of aging clocks and what they ultimately might be measuring. We show that accumulating stochastic variation is sufficient to build aging clocks that accurately predict the chronological and biological age. All tested epigenetic aging clocks, including the most recent pan-mammalian clock, and our own transcriptomic aging clock, correlate with the amount of artificially added stochastic variation to a biological ground state. Surprisingly, we found that an aging clock can be built using just one biological sample and artificially induced stochastic variation accumulation. Even clocks trained with only one biological sample enabled highly correlated predictions with the chronological age and revealed significant differences among samples subjected to lifespan interventions.

In the last part of this thesis, we applied our aging clocks to a pseudo-bulk dataset of neuronal cell classes of the nematode *Caenorhabditis elegans*. We identified almost two-fold aging rate differences

between the youngest and oldest predicted neuron classes, and showed that these biological age differences are associated with neurodegeneration *in vivo*. We then used the predicted age of all neuronal cell classes to identify transcriptomic trajectories over the predicted age (NeuronAge). We show that enriched pathways of genes that are correlated with NeuronAge are conserved in human and mice, thereby bringing the field of cross-species aging transcriptome comparisons to species as far evolutionary apart as *Caenorhabditis elegans* and humans. Lastly, we performed an *in silico* drug screen and identified known and novel neuroprotective small molecule compounds that could be validated *in vivo*, demonstrating that our identified hits do decelerate the age-related neurodegeneration in *Caenorhabditis elegans*.

1 Introduction

1.1 Aging Theories

Understanding the biology of aging, a key factor to the loss of physiological integrity, organ decline, and diseases, as well as defining the scope of what constitutes aging, remains one of the biggest challenges¹⁻³. Numerous aging theories have emerged to explain the underlying causes and consequences, with over 300 theories reviewed as early as 1990 by Medvedev⁴. These theories can be broadly classified into three categories: programmed aging⁵⁻¹¹, the evolutionary theory of aging¹²⁻¹⁵, and the damage accumulation theory of aging¹⁶⁻¹⁸.

1.1.1 Programmed Aging Theories

Programmed aging¹¹, also termed phenoptosis¹⁹, was first proposed by Weismann in 1882 as a genetic process evolved to benefit future populations by clearing frail individuals and freeing up resources¹⁰. This theory therefore implies that aging is favored by natural selection, adaptive, and determined by specific genes⁵. Indeed, there are examples in nature where a rapid programmed degeneration, especially following reproduction, can be found. Semelparity, a reproductive strategy characterized by a single reproductive event followed by death, exemplifies this²⁰. Salmon species die after first reproduction^{21,22}, which can be abrogated by castration²³. And there is evidence that the nematode *Caenorhabditis elegans* degrades its own intestine to feed yolk to the next generation²⁴. It has been suggested that senescence and death is beneficial for evolutionary adaptability²⁵. And the adaptive senescence theory proposes that aging has several beneficial roles in preventing overpopulation, accelerating evolution by faster generation turnover, and to prevent penalties by pathogen exposure^{19,26,27}.

1.1.2 Evolutionary Aging Theories

The non-programmed evolutionary theories of aging started with the mutation accumulation theory from Medawar, who proposed that aging is caused by various harmful mutations that are deleterious late in life¹². The antagonistic pleiotropy hypothesis suggests the existence of pleiotropic genes with beneficial effects early in life but deleterious effects later in life¹³. As the organism ages, selective pressure decreases, especially once it reaches reproductive maturity. Beneficial early-life traits play a significant role in driving reproductive success, while detrimental late-life effects have less impact, potentially leading to the selection for antagonistic pleiotropic genes¹³. The disposable soma theory offers another perspective, suggesting that evolutionary resources are divided between reproduction and somatic maintenance²⁸. According to this theory, organisms must balance investment in reproduction against investment in maintaining somatic tissues, which might affect longevity and

health-span²⁸. All three theories have in common that they only consider individual selection, in contrast to programmed aging theories, which rely on group selection, a highly controversial concept and mostly disregarded by evolutionary biologists^{29–31}. For some genes pleiotropic effects have been suggested, while there is currently no evidence for any gene that evolved to induce aging^{32,33}. The existence of antagonistic pleiotropic genes suggests that aging may be adjustable, and supports program-like features, i.e. not an evolved program, but a side effect of the main genes' function³⁴.

1.1.3 Damage Accumulation Theories of Aging

Due to the inherent imperfections in nature's physical properties, biochemical processes and ultimately all biological processes are prone to errors¹⁸. Additionally, stochastic DNA damage can arise from both endogenous sources, such as metabolic byproducts and oxidative stress, as well as exogenous sources, including exposure to environmental factors like radiation and chemical toxins¹⁶. To maintain proper cell function and genomic integrity, repair mechanisms are essential³⁵. Imperfect DNA repair of these damages leads to mutations³⁶, or unrepaired lesions¹⁶. Unrepaired transcription blocking lesions have been proposed to shape the aging transcriptome and explain why long genes are downregulated with progressing age^{37–40}. Somatic mutations are detrimental⁴¹ and have been shown to scale with lifespan⁴². As the ability of the cell to function not only depends on its intact genomic information, but also on the interplay of the proteome, accumulation of translation errors has been proposed to play a role in the aging process as well⁴³. The accumulation of these errors might be one driver of the observed stoichiometry loss of protein complexes⁴⁴. These small errors, while often not immediately detrimental to cell survival, gradually diminish the stability of the regulatory network over time⁴⁵. In postmitotic cells, the accumulation of these errors can disrupt crucial cellular functions, leading to dysfunction, senescence, or cell death³⁴. Cell division can effectively dilute damage, making it potentially sufficient to handle most forms of damage³⁴. However, mutations and epimutations are exceptions as they can persist despite cell division^{46,47}. The accumulation of these deleterious age-related changes on all levels has been termed the *deleterome* and proposed to drive the aging process³⁴.

1.2 Biological Age Prediction

While chronological age serves as the universal measure of time elapsed since birth, biological age refers to an individual's health and functional status relative to their chronological age, considering factors such as genetics, lifestyle, and environmental exposures that influence the aging process⁴⁸. The difference between biological and chronological age was already explored in Benjamin's 1947 study⁴⁹. He used subjective measurements to quantify the biological age of a person, without, however, validating against mortality or other measures of functional age⁴⁹. If biological age can be measured, it will be an important tool for informed clinical decisions and could potentially accelerate aging research and geroprotective drug discovery⁵⁰. The hypothesis that ionizing radiation induces a general age-acceleration process⁵¹ motivated studies aiming to quantify biological age acceleration in survivors of the Hiroshima atomic bomb compared to age-matched control groups⁵²⁻⁵⁴. First studies of irradiated and nonirradiated subjects in skin⁵³ and erythrocytes⁵⁴ showed no evidence for age acceleration, but led to the first study trying to compute the biological age (physiological or functional age) with a multiple regression model⁵², i.e.:

$$B_i = b_0 + \sum_{j=1}^m b_j * x_{ji} ,$$

where B_i is the biological age of the i -th subject, b_0 is the intercept, b_j is the coefficient for the j -th biomarker and x_{ji} is the j -th biomarker value for the i -th subject. The coefficients are computed via the method of least squares⁵². It was proposed that the biomarkers x_j for the regression should be chosen such that they strongly correlate with chronological age to get insights into the "clock" and accelerate experimental gerontology⁵⁵. Validation of multiple regression models based on clinical variables showed that subjective health⁵⁶, and hypertension⁵⁷ can be distinguished by biological age differences from the chronological age, i.e. age acceleration. These early successes, however, were surrounded by controversy and skepticism, particularly concerning the feasibility of accurately predicting biological age^{58,59}. It was argued that there is neither a single underlying aging process, nor a general (linear) aging rate making biological age predictions unfeasible, and that differences between the predicted and chronological age are rather due to individual differences, diseases, measurement errors, or regression-to-the-mean^{59,60}. Additionally, it was emphasized that the strength of association between the chronological age and a biomarker is not a good selection criterion for a "clock", as biomarkers with perfect correlations would not allow for any measurement of aging rate differences^{59,61}, or might not be causally related to aging, e.g. the degree of baldness⁶². Nevertheless, amidst the debate, proponents of biological aging measurements argued for continued research in the field^{61,63}.

1.2.1 Hochschild's Method

Hochschild proposed a new method for biological age prediction aiming to alleviate some of the criticism, especially:

- 1) The use of the strength of association with the chronological age as a selection criterium.
- 2) The regression-to-the-mean of multiple linear regression⁶², and
- 3) The lack of biological age validation⁶⁴.

He proposed to:

- 1.) Avoid multiple linear regression by instead using multiple simple regressions, each involving one biomarker, and
- 2.) To reverse the regression (i.e. predict the biomarker level by the chronological age, instead of predicting chronological age by biomarker level) to then convert the resulting regression coefficients for a measure of biological age⁶².

For each non-multicollinear biomarker x_j he calculates the intercept $b_{j,2}$ and the slope $b_{j,3}$ on chronological age (C):

$$x_j = b_{j,2} + b_{j,3} * C .$$

The predicted age P_j according to the j-th biomarker x_j is:

$$P_j = b_{j,0} + b_{j,1} * x_j .$$

To compute the needed intercept $b_{j,0}$ and slope $b_{j,1}$ he algebraically converts the coefficients:

$$b_{j,0} = -\frac{b_{j,2}}{b_{j,3}}$$

$$b_{j,1} = \frac{1}{b_{j,3}}$$

This biological age measure was compared with mortality risk factors, revealing small to moderate effects⁶⁴. Subsequently, the Hochschild method was adapted by including chronological age in an orthogonal regression, motivated by the fact that chronological age is the strongest indicator of biological age, further improving, albeit artificially, the correlation between the predicted and the chronological age⁶⁵.

1.2.2 Principal Component Analysis (PCA)

Nakamura et al. advocated to first calculate a principal component analysis (PCA) to then use multiple linear regression not on chronological age, but on the first principal component of the biomarker data and subsequently transforming the predictions to years. This largely alleviated the regression-to-the-mean problem and predicted significant older ages in diabetic or hypertensive subjects⁶⁶, and Down's syndrome patients⁶⁷. Alternatively, the principal components of the biomarker data can be used to calculate a multiple linear regression on chronological age⁶⁸. The latter has the advantage that PCA transforms the data into a new set of orthogonal and uncorrelated variables (principal components).

1.2.3 Klemra and Doubal's Method (KDM)

Klemra and Doubal recognized the need for a more mathematically formalized definition of biological age and a method (denoted as KDM) that does not rely on multiple linear regression⁶⁹. Their method is similar to Hochschild's method in the use of single-biomarker regression, but proposes a mathematically optimal way to compute biological age given six assumptions⁶⁹:

- 1.) The pace of the natural aging process is varying across species and to some degree within species.
- 2.) The average biological age (B) of a population is the chronological age (C) and differences correspond to differences in the pace of aging: $B = C + R_B(0; s_B^2)$, where R_B is a random variable with zero mean and the variance (s_B^2).
- 3.) The individual pace of aging affects all biomarkers similarly.
- 4.) Biological age biomarker can be affected by independent random effects.
- 5.) All biomarkers are functionally uncorrelated.
- 6.) The biomarkers are approximately linear with respect to age.

The biomarker x can then be expressed as:

$$x = F_x(B) + R_x(0; s_x^2),$$

with F_x being the biomarker-specific function over the biological age (B), and $R_x(0; s_x^2)$ the random variable depicting the age-independent random effects. KDM then estimates the biological age by minimizing the distance between the m regression lines (for the m biomarkers with slopes k and intercepts q) and the m biomarker points for an individual in an m -dimensional space with the optimal estimate B_E given at⁶⁹:

$$B_E = \frac{\sum_{j=1}^m (x_j - q_j) \frac{k_j}{s_j^2}}{\sum_{j=1}^m \left(\frac{k_j}{s_j}\right)^2}$$

Where s_j is the standard deviation of the random variable indicating the linear regression residual of the j -th biomarker x_j :

$$R_j(0; s_j^2) = x_j - (k_j B + q_j)$$

As the biological age is not known, Klemra and Doubal give an estimation of s_j with the assumption that the correlation coefficients for all biomarker are the same. The accuracy of B_E is highly dependent on the number of biomarkers and complex to compute. To alleviate this problem KDM was extended to include chronological age (C) scaled by the variance (s_B^2) of the biological age ⁶⁹:

$$B_{EC} = \frac{\sum_{j=1}^m (x_j - q_j) \frac{k_j}{s_j^2} + \frac{C}{s_B^2}}{\sum_{j=1}^m \left(\frac{k_j}{s_j}\right)^2 + \frac{1}{s_B^2}}$$

With the caveat that (s_B^2) is not known and needs to be estimated⁶⁹. Despite several assumptions, estimation of parameters and complex computations, KDM (B_{EC}) outperformed multiple linear regression and the Hochschild method in simulation studies⁶⁹. However, especially assumption 5.), the requirement of functionally uncorrelated biomarkers, hinders its application. Cho et al., therefore, extended the method (denoted as KDM2) by first calculating a PCA to use the uncorrelated principal components of the biomarkers as the variables x_j and further reduced the complexity of KDM (B_{EC}) by substituting s_j with the mean squared residuals of the regression on chronological age, instead of biological age⁶⁸. KDM2 is easier to compute, more broadly applicable, and was shown to perform as well as the original KDM (B_{EC}) when comparing its biological age predictions with the work ability index health questionnaire⁶⁸, and mortality⁷⁰. Conversely, Mitnitski et al. showed that KDM is predicting mortality to a lower accuracy than chronological age on its own, and this independent on whether chronological age was included into the biological age prediction (B_{EC}) or not (B_E)⁷¹.

1.2.4 PhenoAge

To overcome limitations in the accuracy of mortality prediction, Levine et al. replaced the regression on chronological age with a Cox penalized regression on mortality⁷². The mortality score (defined as the 120-month mortality risk) is defined as:

$$MortalityScore = 1 - e^{-\frac{(e^{\gamma*120}-1)*e^{xb}}{\gamma}}$$

Where γ is a parameter that needs to be estimated, and xb is the linear combination of the m biomarkers (including chronological age), i.e.:

$$xb = b_0 + \sum_{j=1}^m b_j x_j$$

The mortality score is then transformed into a biological age (B) estimate in years⁷²:

$$B \approx 141.5 + \frac{\ln(-0.0055 * \ln(1 - MortalityScore))}{0.09}$$

The resulting biological age estimator (PhenoAge) is mostly used as a surrogate measure of phenotypic age that is then subsequently predicted with DNA methylation data⁷², as outlined in the DNA methylation clock section below.

1.2.5 Homeostatic Dysregulation

The homeostatic dysregulation method tries to measure multi-system physiological dysregulation by calculating the Mahalanobis distance (MHBD)⁷³ of multivariate joint distributions of biomarkers⁷⁴. The MHBD is a statistical distance measuring how rare a specific combination of biomarkers is relative to a reference population, i.e. an individual with higher MHBD would be more distant from the population mean, be potentially more physiologically dysregulated, and have a higher mortality risk^{74,75}. The method assumes multivariate normality, which is a conservative assumption for aging biomarker and might therefore lead to underperformance of the method.

1.2.6 DNA Methylation Aging Clocks

Epigenetic modifications encompass chemical alterations to DNA, such as methylation of cytosine residues, and modifications to chromatin structure, such as histone acetylation or methylation. These modifications result in changes in gene expression⁷⁶ and can impact the accessibility of the maintenance machinery, such as DNA repair mechanisms, thereby influencing genome stability⁷⁷. The modifications are known to change with age and are hypothesized to causally contribute to the aging process^{78–80}. In 1973 Vanyushin et al. demonstrated that global DNA methylation levels decrease with age in multiple tissues of rats⁸¹, which was supported in 1983 by Wilson & Jones demonstrating that DNA methylation decreases in aging fibroblasts, with mouse cells exhibiting a quicker decline than human cells, while immortal cell lines displayed a more stable DNA methylation pattern⁸². Subsequent studies showed distinct methylation patterns in specific CpG islands with aging. For instance, the oestrogen receptor 5' CpG island, which are hypermethylated with aging⁸³, while others like transposable element CpG islands are hypomethylated with aging⁸⁴. Hypermethylation was found to be especially associated with bivalent chromatin domain promoters⁸⁵, loci within CpG islands⁸⁶, and at Polycomb-target genes⁸⁷. The significance of age-dependent epigenetic alterations was further highlighted in a comparative study involving young and old monozygotic twins⁸⁸, and a longitudinal twin study during childhood⁸⁹. These studies revealed that younger twins were epigenetically more similar than older twins, suggesting that the genetic background alone cannot account for the observed differences. Noteworthy though, already early longitudinal studies suggested that methylation maintenance is partly under genetic control, shedding light on the interplay between genetic factors and epigenetic changes during aging⁹⁰.

These findings of age-related epigenetic changes motivated the first epigenetic predictor of age (aging clock) based on saliva samples of 34 pairs of identical twins between 21 and 55 years of age⁹¹. While this initial study lacked validation in independent data, Hannum et al. subsequently demonstrated that an aging clock could be built using whole blood DNA methylation data from a mixed population of 656 individuals between 19 to 101 years, which was validated in an independent cohort⁹². Although many epigenetic age-related alterations are tissue-specific⁹³, Koch & Wagner identified 5 CpG sites facilitating age predictions across different tissue types⁹⁴. Later, Horvath significantly improved the chronological age predictions and demonstrated that a human pan-tissue aging clock can be built and validated in a broader range of tissues⁹⁵. Subsequently, numerous DNA methylation aging clocks have emerged, with first generation aging clocks focusing on predicting the chronological age. Single tissue first generation epigenetic aging clocks have been developed and fine-tuned for whole blood^{96–99}, breast tissue¹⁰⁰, saliva¹⁰¹, and the human cortex¹⁰² to improve the chronological age prediction accuracy. Similarly, pan-tissue first generation aging clocks have been improved by using deep

learning^{103,104}, or fine-tuning predictions to a subset of tissues¹⁰⁵, and cancer types¹⁰⁶, and have been developed for a variety of species such as mice^{107,108}, rats¹⁰⁹, and naked mole rats¹¹⁰. Moreover, a first generation DNA methylation aging clock tailored for single-cell data has been developed¹¹¹. Although first-generation aging clocks are trained to predict chronological age, studies suggested that the difference between the predicted age and the chronological age, i.e. delta age, can be predictive of all-cause mortality^{112,113}, is associated with frailty¹¹⁴, and diseases such as Down's syndrome¹¹⁵, or neuropathological measurements¹¹⁶. This association with mortality is attenuated with higher chronological age prediction accuracy, limiting the usage of first-generation aging clocks for biological age prediction¹¹⁷.

Second-generation aging clocks aim to improve this mortality and disease-risk association by using variables indicative of health. Conceptionally, first generation aging clocks may exclude CpG sites that are relevant for the biological age, but don't show a strong age-dependent trajectory. Recognizing this limitation, PhenoAge⁷² and GrimAge¹¹⁸ are not trained on chronological age, but a surrogate measure via a two-step method. PhenoAge first used a Cox penalized regression model to regress the hazard of mortality based on clinical markers and chronological age. The epigenetic clock was then trained to predict this phenotypic age, which led to significant improvements in mortality and health-span prediction⁷². Similarly, GrimAge defined surrogate biomarkers with smoking pack-years and mortality-associated plasma proteins, which are then used to predict time-to-death. Finally, the prediction is transformed into a biological age estimate¹¹⁸. DunedinPoAm diverges from the conventional chronological or biological age predictors by quantifying the rate of aging in a two-step approach¹¹⁹. First, they define the reference rate of aging in a longitudinal dataset of young adults between 26 and 38 years based on 18 blood-chemistry biomarkers¹²⁰, and then use whole-genome methylation data from the same individuals at age 38 to predict the pace of aging, which is significantly associated with mortality and physical function¹¹⁹. Subsequently, DunedinPACE included further data and improved the reliability of the aging pace predictions¹²¹.

Noticing that some CpG sites in technical replicates are highly variable¹²², and that most sites measured with different methylation arrays are not correlated (median correlation of 0.15 between Illumina 450K and EPIC BeadChips in blood samples)¹²³, Higgins-Chen et al. investigated the reliability of epigenetic clocks¹²⁴. Specifically, they tested both first- and second-generation aging clocks and found up to 8.6 years deviation between predictions of technical replicates. To address this issue, they proposed using principal components to minimize the effect of technical variation by extracting age-related covariance, and retrained the first- and second-generation aging clocks to improve reliability¹²⁴. Similarly, Kriukov et al. recognized uncertainty in the prediction outcomes, specifically out-of-distribution uncertainty due to a covariate shift in the dataset, and proposed an uncertainty-aware

clock with a Gaussian Process Regressor¹²⁵. Moqri et al. instead used a biologically-informed set of CpG sites that are highly bound by Polycomb repressive complex 2 (PRC2) and are low-methylated in young organisms to define a biomarker that is assay-agnostic, robust to site-specific technical variability and conserved across species¹²⁶.

The conservation of PRC2-bound aging clock CpG sites across species was also found in the first third-generation aging clock that demonstrated that a pan-mammalian clock can be built and predict the relative age of 185 species to a high accuracy¹²⁷. Subsequently, pan-mammalian maximum lifespan, average gestation time, and age at sexual maturity predictors have been built¹²⁸. Notably, all first-, second-, and third-generation aging clocks mentioned so far are correlative, and not causative, which explains why many aging clocks can be found throughout the epigenome and only limited information can be gained by analyzing clock sites¹²⁹. Recognizing this short-fall, Mendelian-randomization was used to identify CpG sites that are causal to aging-related traits and defined a causality-enriched clock, which might allow for more causal biomarkers of aging¹³⁰. However, despite these advancements in identifying causal CpG sites¹³⁰ and recent more mechanism-based models¹³¹, the biological interpretation of epigenetic aging clocks remains limited.

1.2.7 Transcriptomic Aging Clocks

In contrast, transcriptomic aging clocks offer potentially easier interpretability by directly measuring gene expression levels¹³². In addition, age-related changes in the abundance of genes potentially integrate age-related changes in DNA methylation^{133,134}, histone modifications^{135,136}, the 3D genome organization¹³⁷, and RNA polymerase stalling due to potential transcription blocking lesions³⁷. Even before the advent of epigenetic aging clocks, first transcriptomic clocks in the nematode *Caenorhabditis elegans* based on microarray data of single worms demonstrated that chronological age predictions are possible^{138,139}. Subsequently, it was shown that a transcriptomic clock for *Caenorhabditis elegans* can predict biological age¹⁴⁰, and microarray-based transcriptomic aging clocks for human blood samples¹⁴¹, muscle¹⁴², and brain tissues¹⁴³ were built. To overcome limitations of microarray platforms, e.g. detection of only a subset of the transcriptome, RNA-seq data from the Genotype-Tissue Expression (GTEx) project¹⁴⁴ were used to build tissue age predictors^{145,146}. A clock based on a human fibroblast dataset derived from cell culture of healthy donors enabled chronological age predictions and showed accelerated aging in samples from patients with Hutchinson-Gilford progeria syndrome¹⁴⁷. To improve accuracy and interpretability biologically-informed deep neural networks have been applied¹⁴⁸, and a hypothesis-driven clock based on repetitive element expression has shown accurate age predictions in human samples and *Caenorhabditis elegans*¹⁴⁹. And recently, single-cell RNA-seq data has been used to build cell-type specific aging clocks in mouse¹⁵⁰ and humans¹⁵¹. Although the accuracy and interpretability of transcriptomic clocks were steadily

improving, they were limited by training and predicting chronological age, thereby limiting their usability. Recently, a neuronal single-cell clock for mice used the proliferative fraction of cells as a biological age score, thus improving the predictions beyond chronological age¹⁵².

1.2.8 Proteomic Aging Clocks

Proteomic aging clocks face greater limitations compared to DNA methylation or transcriptomic clocks. This is primarily due to the scarcity of available data and the absence of a standardized approach in generating proteomics data. With different techniques detecting diverse subsets of proteins, ensuring consistency becomes challenging¹⁵³. Nevertheless, proteomic aging clocks have been developed, with the first clock being based on human blood plasma proteins, achieving a prediction accuracy for chronological age with an r^2 of 0.88¹⁵⁴, which was later improved to an r^2 of 0.94¹⁵⁵. A longitudinal dataset of human blood proteins was used to predict the chronological age with a Pearson correlation of 0.88, and identified multiple common health conditions that increased the predicted age¹⁵⁶. The combination of diverse human proteomic datasets, spanning various tissues and proteomic techniques, identified 85 common age-associated proteins, leading to the development of an aging clock with a Spearman correlation of 0.88¹⁵³. This was later improved by including more datasets to a chronological age predictor with a Pearson correlation of 0.96¹⁵⁷, which was subsequently replicated in a novel dataset¹⁵⁸. A longitudinal blood immunome proteomics dataset measured systemic inflammation and its predictions associated with multimorbidity, frailty, and cardiovascular aging¹⁵⁹. A recent preprint predicted all-cause mortality with a penalized Cox regression to improve biological age predictions¹⁶⁰, while another recent preprint argued that proteomic aging clocks can be trained on chronological age and remain strong predictors of mortality, multimorbidity, and frailty¹⁶¹. Following up on the latter, it was demonstrated that blood plasma proteome data could even be used to not only build a blood aging clock, but multiple clocks for different organs. Importantly, this study revealed that these organ-specific clocks facilitated organ-specific risk assessment¹⁶².

1.2.9 Other Aging Clocks

Over the years a manifold of data aside DNA methylation, transcriptomic, and proteomic data has been used to build aging clocks: Chromatin accessibility based on ATAC-seq data¹⁶³; cell-free DNA nucleosome distance, and cell-free DNA fragment size¹⁶⁴; microbiome data^{165,166}; physical activity¹⁶⁷; facial images^{168–170}; lipidomics^{171,172}; glycans^{173–175}; metabolomics^{176–178}; histone modification ChIP-seq data¹⁷⁹; or a combination of multiple OMICs data^{180,181}.

1.3 Aims of this Thesis

As outlined above, a variety of aging clocks based on various biomarkers have been described. While DNA methylation aging clocks have been widely used due to their accuracy and applicability, transcriptional aging clocks have lagged behind. In Aim 1, the goal was to improve the accuracy of transcriptomic aging clocks to match that of established DNA methylation aging clocks and develop a second-generation aging clock predicting biological age.

While transcriptomic aging clocks are easier to interpret than DNA methylation aging clocks, understanding the causal relationships underlying the aging process remains challenging. Moreover, since most aging clocks are trained on chronological age, highly accurate clocks might lose any insights into the causal factors of aging. In Aim 2, we aimed to investigate current aging clocks and identify the underlying mechanism enabling accurate age predictions irrespective of the dataset used.

Aging clocks serve as valuable endpoints in intervention studies, facilitating the discovery of anti-aging compounds and treatments. In Aim 3, we explored the potential of the clock developed in Aim 1 not as an endpoint, but as an *in silico* screening tool for identifying novel compounds that counteract age-related neurodegeneration.

The studies for Aim 1 and 2 have been published, as indicated in the chapters below. The study for Aim 3 has been submitted for publication.

2 BiT age: A transcriptome-based aging clock near the theoretical limit of accuracy

David H. Meyer¹, Björn Schumacher¹

¹Correspondence

Published in *Aging Cell* **20**, 1–17 (2021). DOI: 10.1111/accel.13320

Author contributions:

- D.H.M. conceived and designed the study and performed all bioinformatics analysis;
- B.S. coordinated the project and together with D.H.M. designed the study.
- All authors wrote the paper.



BiT age: A transcriptome-based aging clock near the theoretical limit of accuracy

David H. Meyer^{1,2} | Björn Schumacher^{1,2}

¹Institute for Genome Stability in Ageing and Disease, Medical Faculty, University of Cologne, Cologne, Germany

²Cologne Excellence Cluster for Cellular Stress Responses in Ageing-Associated Diseases (CECAD), Center for Molecular Medicine Cologne (CMMC), University of Cologne, Cologne, Germany

Correspondence

David H. Meyer, Institute for Genome Stability in Ageing and Disease, Medical Faculty, University of Cologne, Joseph-Stelzmann-Str. 26, 50931 Cologne, Germany.

Email: david.meyer@uni-koeln.de

Björn Schumacher, Institute for Genome Stability in Ageing and Disease, Medical Faculty, University of Cologne, Joseph-Stelzmann-Str. 26, 50931 Cologne, Germany.

Email: bjoern.schumacher@uni-koeln.de

Funding information

Deutsche Krebshilfe, Grant/Award Number: 70112899; Deutsche Forschungsgemeinschaft, Grant/Award Number: EXC 2030 - 390661388, GRK2407, KFO 286, KFO 329, SCHU 2494/10-1, SCHU 2494/11-1, SCHU 2494/3-1, SCHU 2494/7-1 and SFB 829; H2020-MSCA-ITN-2018, Grant/Award Number: Healthage and ADDRESS ITNs

Abstract

Aging clocks dissociate biological from chronological age. The estimation of biological age is important for identifying gerontogenes and assessing environmental, nutritional, or therapeutic impacts on the aging process. Recently, methylation markers were shown to allow estimation of biological age based on age-dependent somatic epigenetic alterations. However, DNA methylation is absent in some species such as *Caenorhabditis elegans* and it remains unclear whether and how the epigenetic clocks affect gene expression. Aging clocks based on transcriptomes have suffered from considerable variation in the data and relatively low accuracy. Here, we devised an approach that uses temporal scaling and binarization of *C. elegans* transcriptomes to define a gene set that predicts biological age with an accuracy that is close to the theoretical limit. Our model accurately predicts the longevity effects of diverse strains, treatments, and conditions. The involved genes support a role of specific transcription factors as well as innate immunity and neuronal signaling in the regulation of the aging process. We show that this binarized transcriptomic aging (BiT age) clock can also be applied to human age prediction with high accuracy. The BiT age clock could therefore find wide application in genetic, nutritional, environmental, and therapeutic interventions in the aging process.

KEYWORDS

aging, aging clock, biological aging, biomarkers, *Caenorhabditis elegans*, RNA sequencing, transcriptome

1 | INTRODUCTION

Aging is the driving factor for several diseases, the declining organ function, and overall progressive loss of physiological integrity. Aging biomarkers that predict the biological age of an organism are important for identifying genetic and environmental factors that influence the aging process and for accelerating studies examining potential rejuvenating treatments. Diverse studies tried to identify biomarkers and predict the age of individuals, ranging from proteomics,

transcriptomics, the microbiome, frailty index assessments to neuroimaging, and DNA methylation (Galkin et al., 2020). Currently, the most common predictors are based on DNA methylation. The DNA methylation marks themselves might influence the transcriptional response, but aging also affects the transcriptional network by altering the histone abundance, histone modifications, and the 3D organization of chromatin. The difference in RNA molecule abundance, thereby, integrates a variety of regulation and influences resulting in a notable gene expression change during the lifespan of an organism

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. *Aging Cell* published by Anatomical Society and John Wiley & Sons Ltd.



(Lai et al., 2019). These changes sparked interest in the identification of transcriptomic aging biomarkers, an RNA expression signature for age classification, and the development of transcriptomic aging clocks.

Peters et al. extended previous classification approaches to a regression, which allows the computation of the predicted age and developed a transcriptional aging clock based on whole-blood microarray samples for half of the human genome and reported an r^2 of up to 0.6, an average difference of 7.8 years, and an association of the predicted age to blood pressure as well as smoking status (Peters et al., 2015). Similarly, Mamoshina et al. build a transcriptomic aging clock of human muscle tissue. A deep feature selection model performed best with an r^2 of 0.83 and a mean absolute error of 6.24 years (Mamoshina et al., 2018). However, microarray data have the drawbacks of a limited range of detection, high background levels, and the detection of just a subset of the transcriptome. Instead, by applying an ensemble of linear discriminant analysis classifiers on RNA-seq data, a model with an r^2 of 0.81, a mean absolute error of 7.7 years, and a median absolute error of 4.0 years were obtained in a dataset derived from cell culture of healthy donors (Fleischer et al., 2018). The same model also predicted an accelerated age in 10 patients with the premature aging disease Hutchinson-Gilford progeria syndrome (HGPS).

While a large variety of data, techniques, and analyses have been used to identify aging biomarkers and aging clocks in humans, issues remain with regard to pronounced variability and difficulties in replicability. Indeed, a recent analysis of gene expression, plasma protein, blood metabolite, blood cytokine, microbiome, and clinical marker data showed that individual age slopes diverged among the participants over the longitudinal measurement time and subsequently that individuals have different molecular aging pattern, called ageotypes (Ahadi et al., 2020). These interindividual differences show that it is still difficult to pinpoint biomarkers for aging in humans.

Model organisms, instead, can give a more controllable view on the aging process and biomarker discovery. *Caenorhabditis elegans* has revolutionized the aging field and has vast advantages as a model organism. Even isogenic nematodes in precisely controlled homogeneous environments have surprisingly diverse lifespans; however, the underlying causes are still incompletely understood. Several predictive biomarkers of *C. elegans* aging have been described, and a first transcriptomic clock of *C. elegans* aging using microarray data of 104 single wild-type worms predicted the chronological age with 71% accuracy (Golden et al., 2008). When the prediction was based on modular genetic subnetworks inferred from microarray data with support vector regression, the age of sterile *fer-15* mutants at 4 timepoints was predicted with an r^2 of 0.91. The same approach on the 104 individual N2 wild-type worms yielded an r^2 of 0.77 indicating that for microarray data subnetworks of genes result in better prediction compared with single gene predictors, likely due to the noisiness of the data type (Fortney et al., 2010). Although the accuracy of this model is reasonable, it is limited by the fact that no lifespan-affecting genotypes or treatments were tested and that the validation dataset, although tested on single worms, resulted in an

increased prediction error. Recently, an initial age prediction based on microarray data predicted 60 RNA-seq samples with a Pearson correlation of 0.54 and was improved to an r of 0.86 when the chronological age was rescaled by the median lifespan of the corresponding sample (Tarkhov et al., 2019). Even though this model instead of chronological age predicted the biological age of a variety of *C. elegans* genotypes, it is limited by the accuracy of the prediction. Moreover, the biological age is not reported in days, but as a variable with values between 0 and ~2.5, which makes it harder to interpret.

To date, no aging clock for *C. elegans* has been built solely on RNA-seq data and been shown to predict the biological age of diverse strains, treatments, and conditions to a high accuracy. In this study, we build such a transcriptomic aging clock that predicts the biological age of *C. elegans* based on high-throughput gene expression data to an unprecedented accuracy. We combine a temporal rescaling approach, to make samples of diverse lifespans comparable, with a novel binarization approach, which overcomes current limitations in the prediction of the biological age. Moreover, we show that the model accurately predicts the effects of several lifespan-affecting factors such as insulin-like signaling, a dysregulated miRNA regulation, the effect of an epigenetic mark, translational efficiency, dietary restriction, heat stress, pathogen exposure, the diet-, and dosage-dependent effects of drugs. This combination of rescaling and binarization of gene expression data therefore allows for the first time to build an accurate aging clock that predicts the biological age regardless of the genotype or treatment. Lastly, we show how our binarized transcriptomic aging (BIT age) clock model has the potential to improve the prediction of the transcriptomic age of humans and might therefore be universally applicable to assess biological age.

2 | RESULTS

2.1 | Temporal scaling and transcriptome data binarization allow precise biological clock predictions

We downloaded and processed 1,020 publicly available RNA-seq samples for adult *C. elegans* out of which for 972 samples corresponding lifespan data were available (Table S1). 900 samples were used for the training and testing of the model, the remainder for validation purposes (Figure 1). Out of the 900 samples most (409) were wild-type N2 worm populations. A significant portion of 171 samples contained reads of temperature-sensitive sterile strains such as *glp-1* or *fem-1* or double mutants thereof. 59 samples contained a mutation in the insulin-like growth factor 1 receptor *daf-2* and 45 a mutation in the dietary-restriction mimic strain *eat-2* either as a single or as a combination with a different mutation. 216 samples did not cluster in one of the mentioned groups and contain a variety of different strains. 112 of the samples span 14 different RNAis in 51 samples and 61 empty vector controls. Slightly more than half of the samples (486) were sequenced from a population that was undergoing a treatment (excluding RNAi or empty vector) that is different

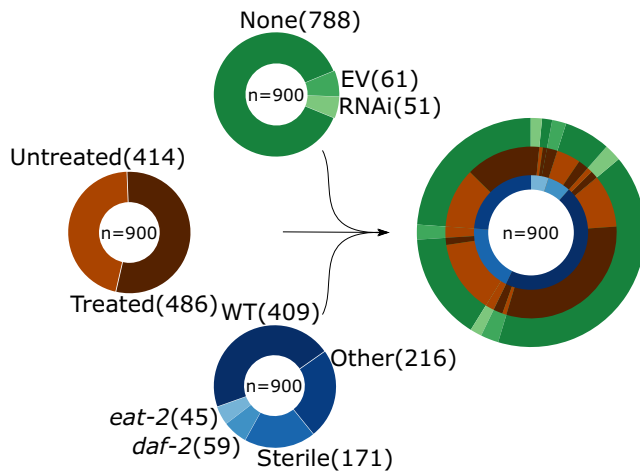


FIGURE 1 Data overview. Overview of the processed published data utilized in the training of the model. Pie charts show the distribution of different genotypes (blue), treatments (brown), and RNAis (green). The convoluted pie chart on the right shows the overlap of the three classes. The partition “Sterile” contains multiple different genotypes that cannot give rise to progeny and *daf-2*, as well as *eat-2*, might contain additional mutations. For a more detailed view, see Table S1

from the standard treatment of an *Escherichia coli* OP50 diet at 20°C. The convoluted circle plot on the right side of Figure 1 shows the overlap of the different possible combinations of strains, RNAi, and treatments in our training samples.

We only downloaded and processed data for which the corresponding publication reported a median lifespan. The lifespan data are required to make strains with vastly different lifespans comparable. Without rescaling, an RNA-seq sample of a long-lived nematode beyond the normal lifespan of a wild-type worm would not be comparable to a wild-type sample, since no sample would be able to be generated. Lifespan-altering manipulations, for example, a temperature shift, a *daf-2* mutation, or oxidative damage, were shown to just shift the lifespan curve by stretching or shrinking it (Stroustrup et al., 2016). One interpretation would be that all lifespan-affecting interventions converge on similar pathways, which affect the risk of death in a similar pattern, just at different velocities. Moreover, there have been descriptions of a transcriptional drift during *C. elegans* aging (Hastings et al., 2019; Tarkhov et al., 2019), which might be due to a (dys-)regulation of single transcription factors (Mann et al., 2016) and the suppression of this transcriptional drift might slow down the aging process (Rangaraju et al., 2015). Notably, age prediction could be improved by rescaling the chronological age by the median lifespan (Tarkhov et al., 2019).

We, therefore, employed a strategy similar to Tarkhov et al. and rescaled the lifespan by the corresponding median lifespan of the sample. We set the median lifespan of a standard wild-type N2 worm to $\mu = 15.5$ days of adulthood. Using this standard lifespan, we calculated a correction factor to determine the biological age of a sample. For example, the correction factor of a strain with a measured median lifespan of 31 days would be $\mu/31 = 0.5$ and thereby assuming

a uniform aging rate reduction of 50%. This correction factor would be applied to each RNA-seq sample of the same strain and experiment. A sample sequenced, for example, at day 10 of adulthood, would be corrected to $10 \cdot 0.5 = 5$ days of biological age. Applying the individual correction factors for each RNA-seq sample allows us to build a classifier of the biological, instead of the chronological age. Importantly, by defining a standard lifespan of 15.5 days we are able to predict the biological age in days instead of a variable between 0 and 2.5 as reported by Tarkhov et al.

Owing to the fact that the public data were generated in multiple laboratories with different protocols and sequencers (see Table S1 for details), we expected noisy data with a strong batch effect. Indeed, the results of an elastic net regression (see Methods for details) on the raw counts-per-million (CPM) reads resulted in a mediocre model with an r^2 of 0.78, a Pearson correlation of 0.89 ($p = 2.82e-304$), a Spearman correlation of 0.86 ($p = 9.97e-258$), a mean absolute error (MAE) of 1.02 days, a median absolute deviation (MAD) of 0.71 days, and a root-mean-square-error (RMSE) of 1.51 days. Figure S1a shows the comparison of the rescaled biological age of the strains on the x-axis and the age predicted by the elastic net regression on the y-axis. Interestingly, the overall absolute error and the variance in the absolute error of the prediction increase strongly after ~5 days (Figure S2).

In order to mitigate this increase in variance, we developed a novel approach and binarized the transcriptome data by setting the value of each gene to 1, if the CPM is bigger than the median CPM of the corresponding sample and 0 otherwise (see Methods for details), thereby reducing the noise, but retaining the information whether a gene is strongly transcribed or not. After this binarization, we trained an elastic net regression model with nested cross-validation to obtain the best parameter setting and optimal set of genes (see Methods for details) that predict the biological age remarkably well with an r^2 of 0.96, a Pearson correlation of 0.98 ($p < 1e-304$), a Spearman correlation of 0.96 ($p < 1e-304$), a mean absolute error of 0.46 days, a median absolute error of 0.33 days, and a RMSE of 0.66 days (Figure S1b).

Interestingly, especially the increased variance in older samples, as seen in our initial analysis in Figure S1a, diminished and showed a strong improvement in overall accuracy. Comparison of the absolute error terms of the raw CPM and the binarized data prediction shows that the absolute error of the binarized prediction is lower than the prediction based on the raw CPMs regardless of the biological age of the worms. Furthermore, while the initial prediction on the raw data starts to get especially inaccurate starting from day 5, the increase in the binarized data is far less pronounced (Figure S2a). Interestingly, also the variance of the absolute error terms stays more stable in the binarized data than the raw data and thereby demonstrating a more robust prediction regardless of the true age of the worms (Figure S2b).

These results show that the binarization approach strongly improves the prediction, especially in older samples, which have been shown to contain a noisier transcriptome. Indeed, this age-dependent noisiness so far hindered the identification of proper aging biomarkers.



The binarization therefore might facilitate the identification by reducing the noise, while retaining the important information. To verify our prediction further, eight independent datasets, not used in the nested cross-validation for optimization of the parameter and gene set, were predicted with an r^2 of 0.91, a Pearson correlation of 0.97 ($p = 2.43e-58$), a Spearman correlation of 0.91 ($p = 6.58e-38$), a mean error of 0.92 d, a median error of 0.53 d, and a RMSE of 1.40 d (Figure S1c).

The results show that the overall prediction is highly accurate; however, although lower than the increase in deviation in the raw data, the binarized data as well show a decrease in accuracy in samples with an older biological age (see also Figure S2). This might be due to the lower sample size of older animals, but might also be influenced by the nature of bulk RNA sequencing itself. Figure S3a shows a standard lifespan curve of *C. elegans*. Until ~day 8, 100% of non-censored worms are alive. Starting from day 8, the first worms die, until the median lifespan is reached at ~15.5 days and the maximum at ~24 days. We can assume that the biological age of worms at the same chronological age follows a normal distribution (Figure S3b). In other words, in a plate of synchronized worms at day 8 we would expect to see that most worms are also at a biological age of 8 days. However, some worms will be healthier while others are already close to death and will therefore be the worms that start dying early. While the peak of this bell curve will therefore be the chronological age of the worm population, some worms will be biologically younger and some older (Figure S3b). Starting from the next day, the first part of the worm population will die (Figure S3c). Assuming the normal distribution of the biological age of the worms and a hypothetical maximum biological age as shown with the dotted line in Figure S3d, we can hypothesize that the biologically older worms will die off first and thereby truncate the biological age distribution on the right side of the curve (Figure S3d). This truncation will shift the true median biological age toward the left side, as indicated by the green line. This becomes more noticeable at the median lifespan of 15.5 days, where by definition 50% of the population is dead (Figure S3e). Following the same reasoning from above, we see that the right half of the biologically older worms died, while the younger half of the population stayed alive. However, this clearly skews the distribution, since the oldest 50% of the population is dead and therefore will not contribute to the average biological age anymore. Indeed, the median biological age will be the median of the remaining, alive worms, that is, the left part of the curve. This will result in a shift of biological age, especially for chronologically older populations (Figure S3f). In consideration of this biological age shift, an RNA-seq sample sequenced at 15.5 days will have a younger true population-median biological age, which will introduce a bias into the regression model. The bias will be not as pronounced in younger samples, since most of the population will still be alive (Figure S3b).

To alleviate this bias, we calculated a second correction term that takes into consideration the hypothetical biological age distribution of the sequenced population (methods for details). Applying this correction before the optimization of the regression resulted in an improved prediction model, especially for the independent

dataset. The new model utilizes 576 genes (Table S2) and predicts the full dataset slightly better, with an r^2 of 0.96, a Pearson correlation of 0.98 ($p < 1e-304$), a Spearman correlation of 0.96 ($p < 1e-304$), a mean error of 0.45 d (-1.63% compared with pre-correction model), a median error of 0.32 d (-2.15%), and a RMSE of 0.64 d (-3.47%) (Figure 2a). The independent dataset is now predicted with an r^2 of 0.94, a Pearson correlation of 0.98 ($p = 1.13e-62$), a Spearman correlation of 0.92 ($p = 6.24e-38$), a mean error of 0.76 d (-17.45%), a median error of 0.53 d, and a RMSE of 1.01 (-28.28%) (Figure 2b). These data indicate that it might be worthwhile including a correction for the survival bias of worms in older populations. The comparison to the prediction on the unbinarized validation data after applying the second correction term showed a strong improvement in accuracy upon binarization with a 48.27% reduction in the mean error (Figure S4a, Table S3).

To confirm that not every gene set of 576 genes results in a similar prediction, we randomly sampled 576 genes and recorded the resulting absolute errors and r^2 values. The boxplot in Figure 2c shows the distribution of r^2 values centering around the mean of 0.488 with a standard deviation of 0.117. The blue dot shows the result of our predicted gene set as a clear outlier at 0.96. The MAE and MAD are centered around 1.27 d and 0.911 d with a standard deviation of 0.066 and 0.063, respectively (Figure S4b).

To assess the precision of the age prediction, we next probed how close this model approaches the theoretical limit of a biological clock. The datasets are annotated in whole days alive from adulthood and thereby including a variance of ± 12 h to the actual chronological age. Random sampling of this error alone gives a mean error of 0.236 (± 0.006) d, a median error of 0.187 (± 0.006) d, and a r^2 of 0.986 (± 0.002). However, since lifespan assays, even done under the same conditions in the same laboratory, will vary, we can assume that the reported median lifespan, used for the temporal rescaling, will also be including an inherent experimental error. Indeed, it has been shown that lifespan assays are heavily affected by the number of animals and less, but substantially, by the scoring frequency, thereby indicating that many lifespan studies are underpowered and often driven by stochastic variation (Petrascheck & Miller, 2017). Computing the mean and SD of lifespan assays for one genotype with the same treatment for several publications shows that the variation is indeed on average ~7% for one standard deviation from the mean with a range between 5.44% and 8.83% (Table S3). An assumption of a moderate 5% deviation between assays increases the mean error to 0.302 (± 0.007) d, the median error to 0.244 (± 0.008) d, and reduces the r^2 to 0.98 (± 0.002). These theoretical optima, shown as dotted lines in the boxplots in Figure 2c and Figure S4b, clearly display the quality of our prediction. We conclude that the prediction based on the set of 576 genes is close to the theoretical optimum.

Next, we compared our model to a previous model (Tarkhov et al., 2019) that described three sets of aging-associated genes. The first set, consisting of 327 genes, was generated by a meta-analysis of publicly available microarray data, the second consists of 902 age-associated genes generated by the analysis of 60

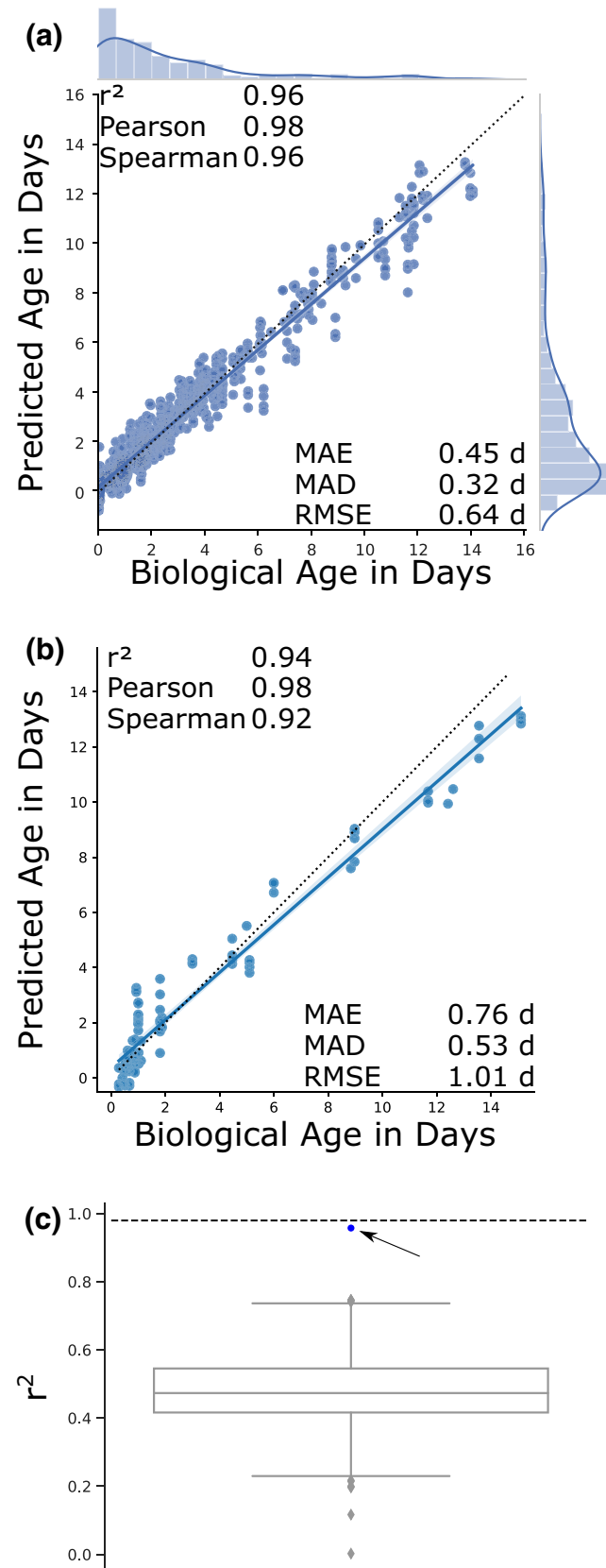


FIGURE 2 Biological age prediction. (a) Results of the biological age prediction computed by cross-validation. The x-axis shows the rescaled biological age in days starting from adulthood additionally corrected by the second rescaling approach. The y-axis shows the predicted age computed by the elastic net regression after the second rescaling approach on binarized gene expression data. Every blue dot displays one RNA-seq sample. The regression line with the 95% confidence interval is shown in blue, and the dotted line shows the perfect linear correlation. The distribution of the data is shown on the side of the plot. r^2 = coefficient of determination, Pearson = Pearson correlation, Spearman = Spearman correlation, MAE = mean absolute error in days, MAD = median absolute deviation in days, RMSE = root-mean-square-error in days. (b) Prediction of the model on eight independent datasets consisting of 94 samples at different time points. The x-axis shows the rescaled biological age in days starting from adulthood additionally corrected by the second rescaling approach. The y-axis shows the predicted age computed by the elastic net regression after the second rescaling approach on binarized gene expression data. For more details on the data, see Table S1. (c) The y-axis shows the r^2 of a given prediction. The box plot displays 1,000 random models with 576 genes. The prediction by our final model with an r^2 of 0.96 is shown as a blue dot and indicated by the arrow. The dotted line shows the theoretical limit of prediction given by the limit of accuracy in the chronological age annotation as well as variance in the lifespan data used for rescaling

prediction of the 900 RNA-seq samples with an r^2 of 0.52 and a mean error of 1.33 d (195.18% increase compared with our final model). The gene set of 902 genes performed similarly, with an r^2 of 0.57 and a mean error of 1.40 d (210.37% increase). Finally, the sparse predictor provided an r^2 of 0.57 and a mean error of 1.36 d (202.07% increase) (Figure S5a–c; for further quality measurements, see Table S3). Remarkably, binarization improves the prediction of these three gene sets as well to an r^2 of 0.74, 0.78, and 0.62, respectively (Figure S5d,e, Table S3). Although the r^2 of the sparse predictor increased to 0.62, the MAE and MAD increased and thereby also show that a single quality assessment is not enough to give a good evaluation (Figure S5f).

Next, we also evaluated the prediction of the independent datasets from Figure 2b with the three previously published gene sets. The gene set of 71 genes performed worst with an r^2 of 0.35 and a MAE of 1.95 d (+156.07% compared with our final model). The gene set derived from microarray data and the gene set with 902 genes performed better with an r^2 of 0.44 and a MAE of 2.20 d (+188.11%), respectively, an r^2 of 0.43 and a MAE of 2.31 d (+203.24%) (Figure S6a–c; for further quality measurements, see Table S3). Remarkably, the binarization could also improve the prediction in this case to an r^2 of 0.87 for the gene set derived from microarray data, 0.85 for the gene set of 902 genes, and 0.72 for the sparse predictor (Figure S6d–f; for further quality measurements, see Table S3).

These comparisons indicate that binarization is improving the quality of regression models overall and that our new model consisting of 576 binarized genes predicts the biological age of *C. elegans* to a high accuracy and superior to previously existing models.

RNA-seq samples, and finally, a sparse subset with only 71 genes that Tarkhov et al. used for their biological age prediction. The gene set derived from microarray data performed worst on the



2.2 | Transcriptomic clock correctly predicts multiple lifespan-affecting factors

Since our model is able to predict the biological age to a high accuracy, we next tested the capability of the model to predict the effect of multiple lifespan-affecting factors. We used the previously determined 576 predictor genes and trained an elastic net regression on the 900 RNA-seq samples, excluding the data for the respective publication. This is thereby a different cross-validation approach where we excluded a whole experimental dataset at a time.

First, we tested the well-known effect of insulin-like signaling on the biological age and saw that a *daf-2* mutation reduces the predicted biological age compared with the WT strain of the same experiment by 41.3% in 4-day adult *C. elegans* (Figure 3a). The even longer-lived *daf-2*; *rsk-1* double mutant is accordingly predicted

to be even younger with a significant reduction of 56.8% in 4-day adults (Figure 3b).

To determine whether short-lived mutants can also be predicted correctly, we next tested *mir-71*, which has been shown to regulate the global miRNA abundance during aging and to directly influence lifespan (Inukai et al., 2018). Compared to WT, *mir-71* mutants are predicted to be 56% older in 5-day adults (Figure 3c). In addition, samples of a gain-of-function *skn-1* mutation, that is, detrimental for lifespan, are predicted to be 77.2% older than wild-type worms at day 2 (Figure 3d). Interestingly, this adverse effect can be rescued by a loss-of-function mutation in *wdr-5* and the subsequent abolishment of the epigenetic mark H3K4me3 (Nhan et al., 2019), which is remarkably also reflected in our prediction. Loss of protein homeostasis decreases overall fitness and is a hallmark of aging. In *C. elegans*, the loss of uridine U34 2-thiolation in *tut-1*; *elpc-1* double

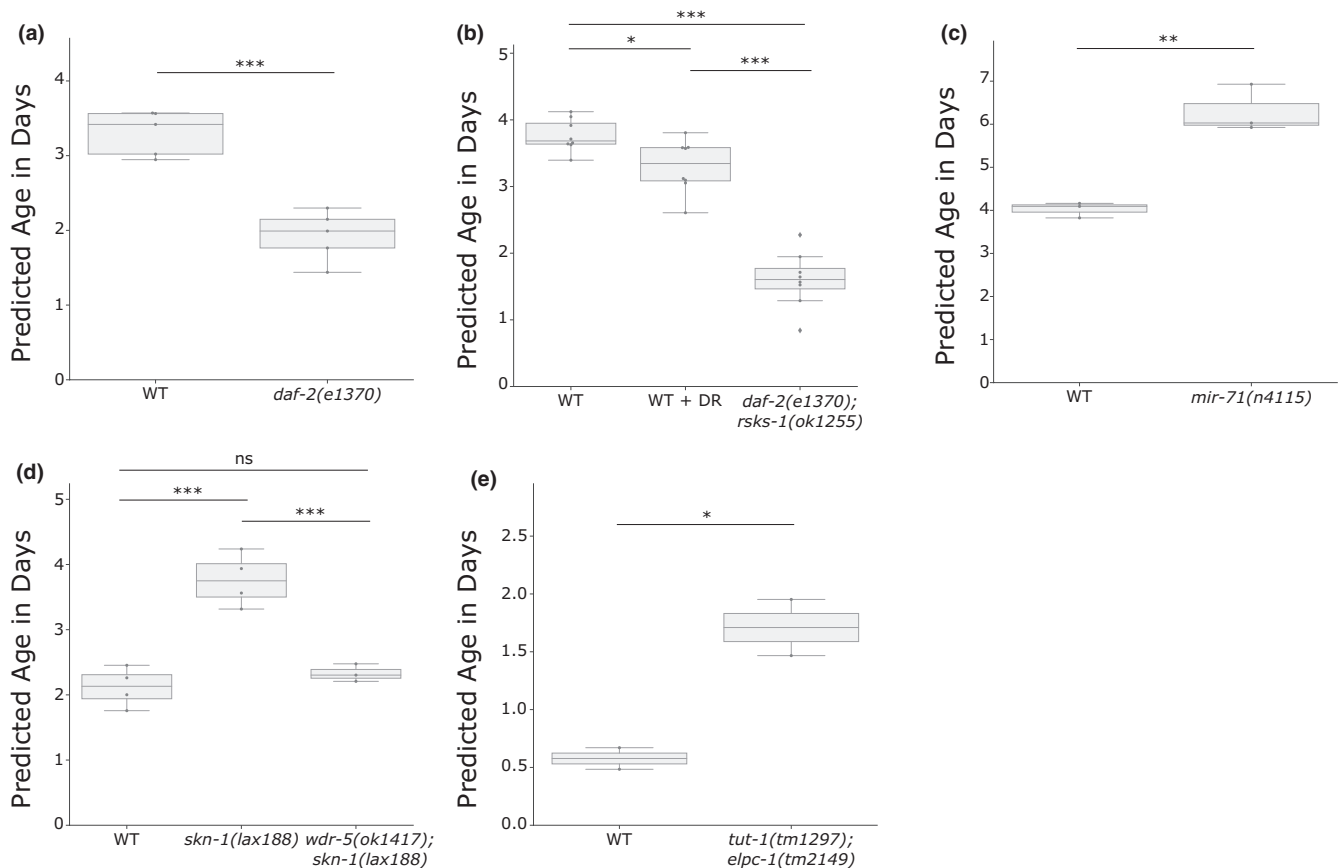


FIGURE 3 Biological age prediction of short- and long-lived mutants. The box plots show the predicted biological age in days on the y-axis. Assuming the properties of a uniform temporal rescaling, a lower predicted age will equal a longer lifespan. The corresponding whole dataset was set aside for the training of the final model for the corresponding plot. Blue dots display single RNA-seq samples. (a) The lifespan-extending *daf-2(e1370)* strain is predicted to be biologically younger than WT samples of the same chronological age (4.5 days). Note that the WT strain in this publication had a longer lifespan (19.4 days) than the standard 15.5 days and is thereby also predicted to be biologically younger than its chronological age. Data from GSE36041. (b) Dietary restriction (DR) and the long-lived double mutant *daf-2(e1370)*; *rsk-1(ok1255)* are predicted to be significantly younger than WT samples of the same chronological age (4 days). Data from GSE119485. (c) The lifespan-shortening *mir-71(n4115)* mutation significantly increased the predicted biological age compared to samples of the same chronological age (5 days). Data from GSE72232. (d) The gain-of-function mutant *skn-1(lax188)* significantly increased the biological age, while an additional mutation in the epigenetic regulator *wdr-5* rescues the biological age back to WT levels (2 days). Data from GSE123531. (e) The double mutant *tut-1(tm1297)*; *elpc-1(tm2149)* significantly increases the biological age (chronological age of 1 day). Data from GSE67387. * $p < 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$, independent two-sided *t* tests were used for comparisons in (a), (c), and (e). One-way ANOVA with a post hoc Tukey test was used in (b) and (d). Table S3 contains more detailed statistics

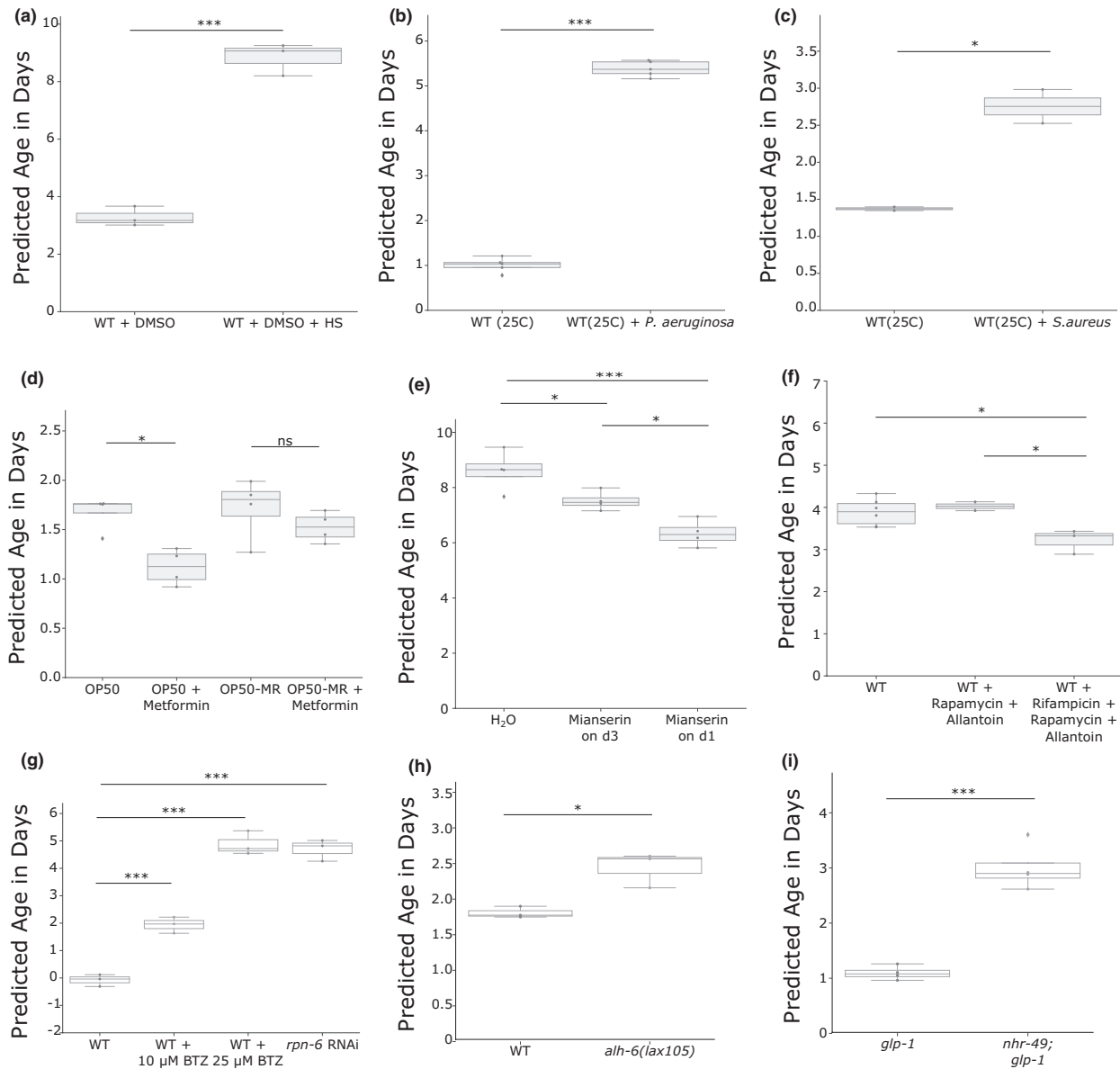


FIGURE 4 Biological age prediction of a variety of treatments and stressors. The box plots show the predicted biological age in days on the y-axis. Assuming the properties of a uniform temporal rescaling, a lower predicted age will equal a longer lifespan. The corresponding whole dataset was set aside for the training of the final model for the corresponding plot. Blue dots display single RNA-seq samples. (a) Heat shock induces a strong increase in the predicted biological age at a chronological age of 3 days in WT. Data from PRJNA523315. (b) Pathogen infection by *Pseudomonas aeruginosa* at 25°C at a chronological age of day 1 increases significantly the predicted age. Data from GSE122544. (c) Pathogen infection by *S. aureus* at 25°C at a chronological age of day 1 increases significantly the predicted age. Data from GSE57739. (d) The bacterial strain-dependent effect of metformin is resembled in the prediction. The box plots show wild-type worm populations at a chronological age of day 2 with either a standard OP50 *E. coli* diet or a Metformin-resistant OP50 (OP50-MR) strain with or without 50 mM Metformin. A two-way ANOVA showed a significant treatment effect ($p = 0.004$). Data from E-MTAB-7272. (e) The dosage-dependent effect of Mianserin is resembled in the prediction. The box plots show wild-type worm populations at a chronological age of day 10 either treated with water or 50 μ M Mianserin on day 3 or day 1. A one-way ANOVA showed significance ($p = 0.0008$). Data from GSE63528. (f) The effect of drug combinations at the chronological age of 6 days is resembled in the prediction. A one-way ANOVA showed significance ($p = 0.02$). Data from GSE108263. (g) An independent dataset without a reported lifespan sequenced at the chronological age of day 1. Wild-type worms were treated with either 10 μ M or 20 μ M of the proteasome inhibitor Bortezomib (BTZ), or RNAi against the proteasomal subunit *rpn-6*. Data from GSE124178. (h) An independent dataset without a reported lifespan sequenced at the chronological age of day 3. Data from GSE121920. The predicted median lifespan reduction of 35.7% is similar to the reported lifespan reduction of 33.5% (Pang & Curran, 2014). (i) An independent dataset without a reported lifespan sequenced at the chronological age of day 2. Data from GSE158729. The predicted median lifespan reduction of 63.96% is similar to the reported lifespan reduction of 50%–60.69% (Ratnappan et al., 2014). * $p < 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$, independent two-sided t tests were used for comparisons in (a), (b), (c), (h), and (i). One-way ANOVA with a post hoc Tukey test was used in (e), (f), and (g). Two-way ANOVA with a post hoc Tukey test was used in (d). Table S3 contains more detailed statistics



mutants has been shown to have a negative impact on the efficiency of translation and to promote protein aggregation (Nedialkova & Leidel, 2015). Strikingly, this effect on translational efficiency is also reflected in the transcriptomic aging clock for day 1 adults, which are predicted to be 196% older than their wild-type counterpart (Figure 3e).

These data show that the BiT age clock can effectively predict the biological age of a variety of mutants and pathways, ranging from the insulin pathway, miRNAs, and the epigenetic mark H3K4me3 to translational efficiency.

Since both, long-lived and short-lived strains, are predicted with the correct pattern, we next asked whether we could predict the effect of dietary restriction (DR) on the biological age. Although the effect was slight, the dietary-restricted worms are predicted to be 12.9% younger than their normal-fed counterpart at day 4 of adulthood (Figure 3b). DR-induced longevity was shown to depend on the PMK-1/p38 signaling-regulated innate immune response. In *C. elegans*, *sek-1* is part of the PMK-1/p38 signaling cascade and required for longevity in dietary-restricted worms (Wu et al., 2019). Noticeably, the same trend can be observed in our prediction for day 6 adults (Figure S7a). A two-way ANOVA showed a significant interaction between the effects of the strain and dietary restriction ($p = 0.004$), which indicates that the effect of DR is dependent on *sek-1* activation. Although in this dataset, the adjusted p -value of the effect of DR in WT worms is not significant ($p = 0.057$), it is interesting to note that the dietary-restricted worms are on average 32% younger than the *ad libitum* fed WT worms. This biological age reduction is thereby showing a stronger effect than the 12.9% reduction in Figure 3b. This could be due to strain differences in the different laboratories or suggest that positive effects of DR add up over time.

Next, we decided to test whether different lifespan-shortening stressors can be predicted correctly. Both heat stress (Figure 4a) and pathogen exposure to either *P. aeruginosa* or *S. aureus* (Figure 4b,c) showed a strong increase in the predicted biological age. Heat stress increased the prediction by 169.3% in day 3 adults. *Pseudomonas aeruginosa* increased the predicted age by 421.4%. And *S. aureus* increased the biological age prediction by 101%, in day 1 adults.

While heat or pathogen exposure can lead to a quick demise of the animals, we wondered whether more subtle changes in lifespan by different diets and subsequent nutrient metabolism could also be detected. It was shown that an *E. coli* K12 variant's indole secretion extends fecundity and overall healthspan and lifespan in *C. elegans*, while an isogenic *E. coli* strain (K12tnaA) with a deletion in the indole-converting gene does not have these benefits. This effect on healthspan was reported to be not yet visible in worms on day 8, but showed a significant difference only at the next tested timepoint on day 15 (Sonowal et al., 2017). Intriguingly, the same pattern can be observed in RNA-seq samples of day 3 and day 12 (Figure S7b). A two-way ANOVA showed a significant treatment effect ($p = 0.034$) indicating the sensitivity of the approach. Moreover, in accordance with the published results, a subsequent post hoc Tukey test showed no difference between the diets on day 3 (adjusted $p = 0.9$), while day

12 showed a 15.3% increased biological age in the K12tnaA diet (adjusted $p = 0.0506$). Consistent with the link between diet-dependent changes in nutrient metabolism and lifespan, it has been shown that the lifespan-extending effect of Metformin is, at least partially, regulated by a bacterial nutrient pathway (Pryor et al., 2019). A two-way ANOVA of the predicted biological age of day-2 adults, grown on either *E. coli* OP50 or a Metformin-resistant OP50 strain, with or without Metformin showed as well a significant bacteria effect ($p = 0.045$) as a significant drug effect ($p = 0.004$). A subsequent post hoc Tukey test showed a significant reduction in the biological age of Metformin-treated wild-type worms grown on OP50 (−34.5%), but no significant effect in worms grown on Metformin-resistant OP50 (Figure 4d).

Next, we asked whether the effect of the duration time of a drug might be reflected on the transcriptomic age. The antidepressant Mianserin has been shown to extend the lifespan of *C. elegans* by inhibiting serotonergic signals, which is lessening the age-dependent transcriptional drift. This effect is more pronounced in animals that were treated starting from day 1, compared to starting the treatment from day 3 (Rangaraju et al., 2015). Our prediction of day 10 adults resembles this conclusion; a one-way ANOVA showed a significant difference ($p = 0.0008$) and an ensuing post hoc Tukey test revealed statistical significance between all three cases, with the biggest effect in worms treated from day 1 (Figure 4e).

An interesting and challenging question is whether the combination of different lifespan-extending drugs might have a synergistic effect. Admasu et al. reported that not all combinations of drugs have an additive effect. While the combination of Rapamycin with Allantoin had no effect on the lifespan of wild-type worms, the triple combination with Rifampicin surprisingly had the biggest effect (Admasu et al., 2018). Interestingly, while the administration of rifampicin, rapamycin, and allantoin significantly reduced the predicted age by 17.7% (Figure 4f), the double combination of rapamycin and allantoin did not change the predicted lifespan, which is in accordance with the published lifespan results.

Lastly, we decided to check the biological age prediction of independent validation data and downloaded three datasets for which no direct lifespan data (i.e., in the same publication) were published and which contained treatments and strains that were not included in any of the analyses and nested cross-validations above. We first tested the effect of proteotoxic stress on the transcriptional age with samples of two different dosages of the proteasome inhibitor bortezomib (BTZ) and the knockdown by RNAi of the proteasomal subunit RPN-6.1 and saw a significant increase in the biological age of all three samples (Figure 4g). Notably, the effect of BTZ shows a dose dependency. *rpn-6.1* RNAi has been shown to strongly reduce the lifespan of WT worms (Vilchez et al., 2012), and BTZ supposedly mimics the effects by directly blocking the proteasome and has been shown to dramatically reduce the lifespan of starved worms (Webster et al., 2017). Moreover, although no direct lifespan data are available for normal-fed worms, 10 μ M BTZ leads to an early death starting from day 3 (Finger et al., 2019), while 25 μ M even increased mortality (Fabian Finger, personal communication). Next, we



tested samples with a mutation in *alh-6* (Yen et al., 2020), which resulted in a 35.7% reduction in the predicted lifespan (Figure 4h). This is remarkably close to the previously reported 33.5% lifespan reduction in *alh-6(lax105)* (Pang & Curran, 2014). Lastly, we tested *glp-1* and *nhr-49*; *glp-1* samples for which no direct lifespan measurement was available. A mutation in *nhr-49* was previously reported to decrease the lifespan in a *glp-1* background by 50–60.69% (Ratnapan et al., 2014), which is in line with the predicted mean 63.96% decrease (Figure 4i).

These results demonstrate that the nested cross-validation was sufficient to prevent overfitting, that our model extends beyond the data described here and that even lifespan-affecting stressors unknown to the model, for example, proteasomal stress, are correctly predicted.

We next wondered how well the aging clock that is measured at one specific timepoint could predict the median lifespan. The prediction of the median lifespan from the biological age assumes a uniform lifespan shift. In other words, if the biological age ratio of two strains or treatments stays constant, we are able to compute the predicted median lifespan. For example, if a sample is twice as long lived as its control, we assume a uniform 50% reduction in the biological age compared with the control, regardless of the timepoint of sequencing; that is, the biological age will be half regardless of the chronological age. The aforementioned intrinsic biases in the chronological age and lifespan assays, however, limit the precision of the predicted median lifespan, especially in chronologically younger samples as here the intrinsic experimental error of ± 12 h has a greater influence (Figure S8). Nonetheless, the predicted median lifespan is within the theoretical error bounds in most of the tested samples, indicating that not only biological age but also median lifespan could be predicted by the transcriptomic clock (Table S4).

Nonetheless, the aforementioned 41.3% biological age reduction in *daf-2* in 4-day adults corresponds to a 1.71-fold lifespan extension. This *daf-2* strain is reported to be 2.6-fold longer-lived than its control; however, even with the theoretically optimal prediction, the predicted lifespan effect will vary due to the aforementioned intrinsic biases to around 2.6 ± 0.5 -fold. Since the WT sample of this dataset (Zarse et al., 2012) was already longer lived than our standard 15.5 days, we also computed the comparison against 15.5 days which resulted in a 2.31-fold increase in lifespan for *daf-2*.

In addition, it cannot be excluded *per se* that some mutations or treatments might affect the lifespan non-uniformly over time, which would result in an additional bias in the model (Table S4). Indeed, our analysis of the 2 DR datasets (Figure 3b and Figure S7a) might indicate such a bias (even though all values are within the lifespan error bounds). The 12.9% reduction in biological age at day 4 (Figure 3b) corresponds to a 1.15-fold lifespan extension (in comparison with the theoretical 1.36 ± 0.26 -fold extension). The samples on two additional days of DR (Figure S7a), however, are predicted to be 1.47 times longer lived (theoretical 1.61 ± 0.22 -fold extension).

In conclusion, we demonstrated that the BiT age clock of *C. elegans* is highly accurate and versatile usable. We showed that it correctly predicts the effects of insulin-like signaling, a modified miRNA

regulation, the effect of an aberrant active transcription factor, and the reversal of this effect by an epigenetic mark, translational efficiency, dietary restriction, and the requirement of the intact innate immune system on its lifespan-extending effect, heat stress as well as pathogen exposure, and the effects of diet-depending metabolites. Lastly, we also showed that the predictor is able to correctly identify the effect of Metformin through the host's microbiota, the dosage-dependent effect of drugs, and the counterintuitive fact that the combination of lifespan-extending drugs might not be necessarily synergistic. Strikingly, our model extends beyond the data used for the nested cross-validation and is able to correctly predict the biological age of worms, for which no direct lifespan data were available. The BiT age clock could thus facilitate the assessment of pro- and anti-aging effects of genetic, metabolic, environmental, or pharmacological interventions as it determines the biological age and predicts median lifespan.

2.3 | The predictor genes are enriched in age-related processes, the innate immune response, and neuronal signaling

For the final model, we calculated the regression coefficients of the 576 genes based on all the 900 training samples for which lifespan data were available (Figure 1, Table S1). The final regression model utilizes 576 genes, out of which 294 have a negative coefficient and thereby are mostly expressed in young worms, while 282 genes have a positive coefficient and thereby increase the predicted age if active (the genes with the corresponding regression coefficients can be found in Table S2). Intriguingly, the protein-coding genes with a negative coefficient were enriched on the X-chromosome and are significantly less expressed from chromosomes I and II (Figure S9a). Protein-coding genes with a positive coefficient show a opposite trend and are significantly enriched on chromosomes I and II, while depleted from chromosome IV (Figure S9b,c). Interestingly, a gene set enrichment analysis of the genes with a negative coefficient, so those that are associated with younger samples, is enriched in age-related categories that are downregulated with aging (Figure 5a). Moreover, the 294 genes are enriched in the *pmk-1*, *elt-2*, *pqm-1*, and *daf-16* transcription factor target category (Figure 5b). A motif search at the promoter regions of the genes with a negative coefficient corroborates this finding and shows a significant enrichment in the GATA transcription factors PQM-1 and ELT-3 (Figure S10a). Although the gene set enrichment analysis with WormExp did not show a significant enrichment of transcription factors in the gene set with a positive coefficient, the motif search also identified the GATA motif enriched at the promoter regions (Figure S10b). Notably, the GATA transcription factor *elt-6* is within the top 30% of genes with a positive coefficient in our gene set and thereby correlated with older worms and has been shown to increase during normal aging and to increase the lifespan upon knock down by RNAi (Budovskaya et al., 2008). Interestingly, genes associated with younger worms are also enriched in genes that are upregulated in germline-ablated animals (Figure 5c), which in general exhibit an increased lifespan.



Genes with a positive coefficient on the other hand are enriched in categories that show an increase with age (Figure 5d).

A subsequent functional enrichment analysis (s. methods) revealed a strong enrichment of signal peptides (i.e. proteins that are targeted to the secretory pathway by their signal sequence), transporter activity, and neuropeptides, which suggest that especially systemic responses influence the aging process (Figure 5e). Neurotransmitters, although not directly enriched in the GO-term analysis, might as well play an important role: *hlc-1* is one of the genes with the strongest increase in predicted age of our gene set. It has been previously shown to be present at the presynaptic terminal of cholinergic neurons and to regulate the normal secretion of acetylcholine neurotransmitter and Wnt vesicles (Tikiyani et al., 2018). In the same manner, the dopamine receptor *dop-4* is in the top 25% of genes with a negative coefficient and has been shown to promote healthy proteostasis and the innate immunity as well as detoxification genes (Joshi et al., 2016). Interestingly, the innate immune response and cytochrome P450 enrichment in our gene set might indicate a role of a general stress response, detoxification, and drug metabolism during the aging process. Consistent with a general stress response, we also find *csa-1* in the list of genes with a positive coefficient, which might indicate an increased DNA damage load in older worms.

To conclude, these results further validate the genes used for the age prediction and indicate that the aging process might be driven

by the dysregulation of single transcription factors (Figure 5b) and a systemic signal transmitted by secreted peptides (Figure 5e).

2.4 | Improved Human age prediction by the BiT age clock

To demonstrate that our novel approach is also usable for other organisms, we employed a recent human dermal fibroblast RNA-seq dataset generated from cell culture of 133 healthy individuals with ages between 1 and 94, and 10 patients with Hutchinson-Gilford progeria syndrome (HGPS) with ages between 2 and 9 (Fleischer et al., 2018). Fleischer et al. showed that an LDA ensemble approach can predict the age of the 133 healthy patients with a r^2 of 0.81, a mean error of 7.7 years, and a median error of 4.0 years. Moreover, they find a statistical increase in the predicted biological age of HGPS patients, as would be expected from a premature aging disease. However, as they mention, the ensemble method has some limitations, that is, the discretization of age, the computational cost, and the difficult interpretation of the influence of gene expression changes on the predicted age.

Our regression-based method is fast to compute, does not require the discretization of age, and directly allows the effect interpretation of the activity of single genes on the predicted age.

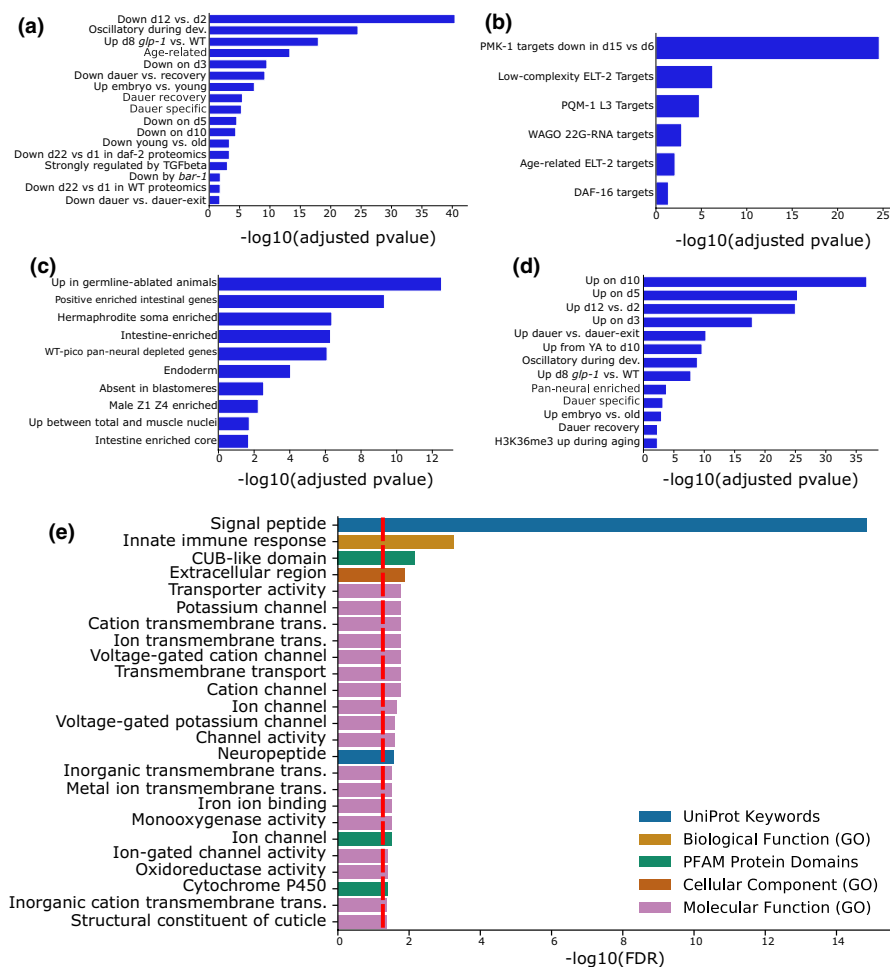


FIGURE 5 Functional analysis of the predictor genes. (a–d) WormExp gene set enrichment analysis for the 576 predictor genes. The x-axis displays the $-\log_{10}$ of the adjusted p -value. Only statistically significant (adjusted $p < 0.05$) enrichments are shown. (a–c) Gene set enrichment analyses for the genes with a coefficient ≤ 0 for the Development/Dauer/Aging category (a), the TF Targets category (b), and the Tissue category (c). (d) Gene set enrichment analyses for the genes with a coefficient > 0 for the Development/Dauer/Aging category. (e) Functional enrichment analysis for the 576 predictor genes by String and geneSCF. The x-axis displays the $-\log_{10}$ of the FDR. The red line displays an FDR of 0.05. Different enrichment categories are color-coded

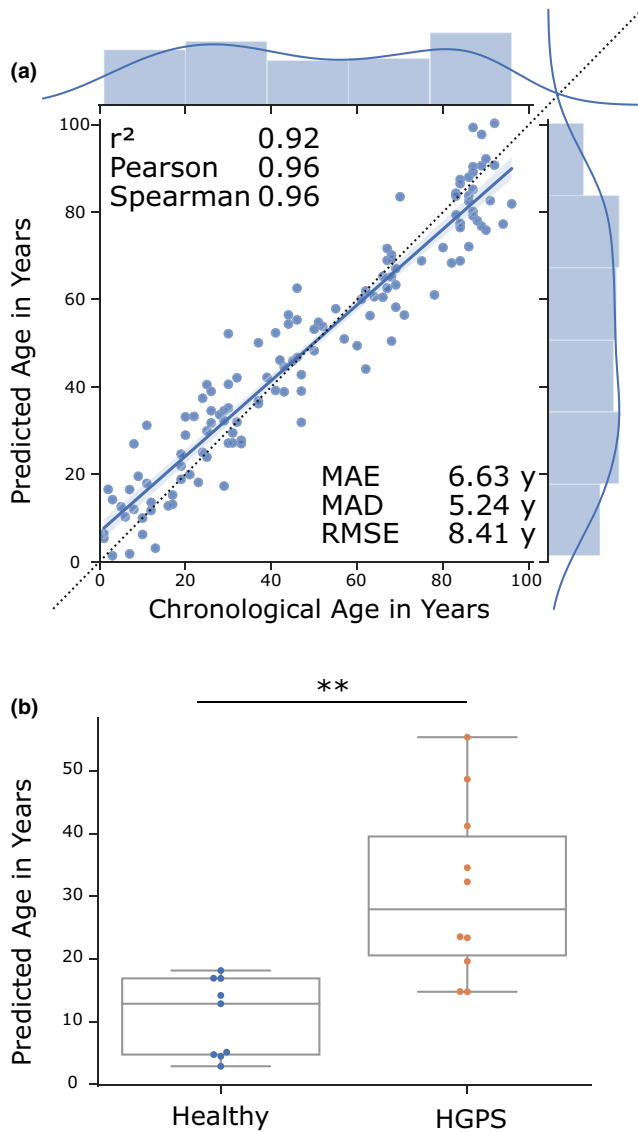


FIGURE 6 Transcriptomic human aging clock. (a) Results of the age prediction computed by cross-validation on human fibroblast gene expression data. The x-axis shows the chronological age in years. The y-axis shows the predicted age computed by an elastic net regression on binarized gene expression data. Every blue dot displays one RNA-seq sample. The regression line with the 95% confidence interval is shown in blue, and the dotted line shows the perfect linear correlation. The distribution of the data is shown on the side of the plot. r^2 = coefficient of determination, Pearson = Pearson correlation, Spearman = Spearman correlation, MAE = mean absolute error in years, MAD = median absolute deviation in years, RMSE = root-mean-square-error in years. Data from GSE113957. (b) Box plots of age predictions of samples from Hutchinson-Gilford progeria syndrome patients (red) and predictions of age-matched healthy controls (blue) by the elastic net regression of binarized gene expression data. Progeria samples are predicted to be significantly older than age-matched healthy controls. Data from GSE113957. ** $p \leq 0.01$, calculated by an independent two-sided t test. Table S3 contains more detailed statistics

Using the elastic net regression on the unbinarized data resulted in a model of 132 predictor genes and in a similar prediction quality as the elastic net regression by Fleischer et al. (Figure S11a), and

similarly, the HGPS samples are not predicted to be biologically older (Figure S11b). However, binarization of the data before calculating the elastic net regression improved the results dramatically to an r^2 of 0.92, a Pearson correlation of 0.96 ($p = 7.87e-73$), a Spearman correlation of 0.96 ($p = 9.31e-73$), a MAE of 6.63 years, a MAD of 5.24 years, and a RMSE of 8.41 years (Figure 6a). Moreover, our model predicts the HGPS patients to be significantly older (Figure 6b). This new model contains 141 predictor genes (Table S5), out of which 25 are significantly enriched in the biological process regulation of cell death. Interestingly, among the predictor genes the forkhead transcription factor FOXO1—a regulator of the aging process in *C. elegans* and mammals—is positively correlated with age thus further supporting the evolutionary conservation of transcriptionally regulated longevity mechanisms (Martins et al., 2016).

To summarize, these data indicate that elastic net regression on binarized gene expression data is not only usable in the nematode *C. elegans*, but also in more complex organisms like humans.

3 | DISCUSSION

The molecular understanding of aging on the genetic, epigenetic, transcriptomic, proteomic, and metabolomic level has made steady progress over the recent years. Since the initial discovery of genetic mechanisms that determine longevity, *C. elegans* has remained an important model system not only for the genetics of aging but also for devising molecular intervention strategies. However, up to date no single model could predict the biological age of any organism to a high accuracy in diverse strains, treatments, and conditions. In our study, we show that the binarization of gene expression data allows a biological age prediction of *C. elegans* to an unprecedented accuracy and for the first time the prediction of a variety of lifespan-affecting factors. Additionally, we show that the binarization approach, even without the biological rescaling, might be applicable to and improving the predictions in other organisms. This is in contrast to the currently most widely used epigenetic clocks, which are limited to organisms with DNA methylation marks. Moreover, our results suggest that especially the innate immune system and neuronal signaling are important for an accurate prediction and therefore also might play an essential role in the aging process.

Binarization of the gene expression data hugely improved the predictability of the biological age. Interestingly, the biggest deviation from the true biological age is in the samples treated with heat shock or in *mir-71*, *eat-2*, and *skn-1* (*gof*) mutants. Heat-shock treatment and an *eat-2* mutation have been shown to exhibit a different aging trajectory and to diverge from the temporal scaling approach proposed by Stroustrup (Stroustrup et al., 2016). Similarly, *skn-1* (*gof*) and *mir-71* display a sharp drop in lifespan (Inukai et al., 2018; Nhan et al., 2019) that cannot totally be accounted for with our median lifespan-rescaling approach. Incorporating the whole lifespan curve could therefore improve the prediction



even further. In this regard, it is also noteworthy that the utilized bulk-sequencing data introduce several biases that might not be reflected in a simple rescaling approach. We tried to alleviate some of the potential biases with our second rescaling approach, which should reduce the error that is introduced by the fact that especially the biologically older part of a population dies off first. However, it has been published that *C. elegans* dies of at least two different types of death (Zhao et al., 2017): either an early death with a swollen pharynx, induced by an increased bacterial content, or a later death with an atrophied pharynx. This might introduce a different bias, since the initial transcriptional response close to an early death might be different from the response to a later death. Nevertheless, even with these limitations our model predicts the biological age of worms remarkably well.

The increasing error and increase in variance of the age predictor in older worms is especially visible in the unbinarized model. This might be due to the known age-dependent increase in transcriptional variety that limits the ability of the regression model to pick an accurate subset of genes. Different hypotheses have been proposed that try to explain this transcriptional noise. In *C. elegans*, it might be partially regulated by a microRNA feedback loop that is dependent on *mir-71* (Inukai et al., 2018), serotonergic signals (Rangaraju et al., 2015), and the decline of the GATA transcription factor *ELT-2* during aging (Mann et al., 2016). One interesting possibility is the idea that the increasing noise is driven by accumulating somatic mutations over the course of aging. Indeed, Enge et al. demonstrated an increase in the transcriptional noise as well as an age-dependent accumulation of somatic mutations in single human pancreatic cells; however, they did not find any support for a causal relationship between exonic mutations and transcriptional dysregulation (Enge et al., 2017).

3.1 | Transcription factors

Similar to Tarkhov et al., we find an enrichment in targets of DAF-16, the GATA transcription factors *PQM-1* and *ELT-2*, and *PMK-1* in our predictor gene set. DAF-16 is known to be involved in a variety of stress responses and longevity pathways (Sun et al., 2017). GATA transcription factors have been found to be relevant for a variety of tissue-specific stress responses and to have a functional role in the aging process (Budovskaya et al., 2008). Moreover, deactivation of *elt-2* has been described as a major driver of normal *C. elegans* aging (Mann et al., 2016) and *pqm-1* has been shown to decline with age and to be involved in *daf-2*-mediated longevity (Tepper et al., 2013). The p38 MAPK family member *pmk-1* is an important gene in the nematode's pathogen defense system and innate immunity.

3.2 | Innate immune response

The innate immune system of *C. elegans* has been linked to several lifespan-affecting pathways (Ermolaeva & Schumacher, 2014).

Schmeisser et al. (2013), for example, showed that dietary restriction (DR)-dependent lifespan extension requires a limited neuronal ROS signaling via a reduced mitochondrial complex 1 activity that activates *PMK-1/p38*. Furthermore, it has been shown that the intestinally produced and secreted innate immunity-related protein *IRG-7* can lead to the activation of the *p38-ATF-7* pathway and is required for the longevity in germlineless nematodes (Yunger et al., 2017). Apart from long-lived mutants, *PMK-1* expression was also observed to decline with normal age, leading to an innate immunosenescence in *C. elegans* that has been proposed to be a driving factor of the aging process (Youngman et al., 2011). This immunosenescence and the overall involvement of the innate immune system in aging has also been shown in other model organisms and might demonstrate an evolutionary conservation. Our work falls in line with these reports and supports an important role of the innate immune response in *C. elegans* aging.

3.3 | Neuronal signaling

Our model also shows an enrichment in neuropeptide signaling. Neuronal communication is important for the organism's homeostasis when responding to different stressors and a changing environment and has been implicated in the aging process. It has also recently been shown that the suppression of excitatory neurotransmitter and neuropeptide signaling is partially required for the longevity of *daf-2* mutants (Zullo et al., 2019) and similarly a glia-derived neuropeptide signaling pathway that affects the aging rate and healthspan of worms has been described and shows the potential for neuropeptide involvement in the aging process (Yin et al., 2017). In line with this, we find *hic-1* and *dop-4* in our predictor gene set. *hic-1* is important for the regulation of acetylcholine neurotransmitter (Tikiyani et al., 2018) and might therefore indicate a role of *hic-1* in the locomotion defect that occurs with aging (Glenn et al., 2004). Besides the role of *dop-4* in the innate immune response (Joshi et al., 2016), it has also been implicated in the slowing down of habituation (Ardiel et al., 2016). Older worms have been shown to exhibit a greater habituation and a slower recovery from it (Beck & Rankin, 1993). The fact that *dop-4* has a negative coefficient in our age prediction suggests that it is less transcribed in older worm populations, thereby making it an interesting target for the cause of increasing habituation with age.

3.4 | Human data

Lastly, we demonstrated that binarized gene expression data also allow building an accurate human age prediction. Currently, the analysis is limited by the data amount and future studies should include more high-quality data from different cohorts with different environments and populations. Optimally, the data would be generated with biopsies from different tissues of living donors without the need of cell culture. Nevertheless, we demonstrated that binarization improves the level of prediction beyond the current standard and that it also allows for a



prediction by an elastic net regression, which results in an easy interpretable gene set. Interestingly, we found a significant enrichment in the biological process regulation of cell death, including FOXO1, which indicates that certain age-related pathways, such as insulin signaling, are indeed relevant for multiple species and evolutionarily conserved.

4 | CONCLUSIONS

The binarized expression of our 576 genes is sufficient to predict the biological age of *C. elegans* independent of the underlying genetics or environment with an accuracy near the theoretical limit. Our analysis suggests that the innate immune response, neuronal signaling, and single transcription factors are major regulators of the aging process independent of the strain and treatment. Although the temporal rescaling approaches will not be applicable in humans, we have also shown how the binarization approach improves the chronological age prediction of a recent human dataset. Our work establishes that an accurate aging predictor can be built on binarized transcriptomic data that extends beyond the training data, predicts lifespan effects across diverse genetic, environmental, or therapeutic interventions, is employable in distinct species, and might thus serve as a universally applicable aging clock.

5 | MATERIALS AND METHODS

5.1 | Data processing

The quality of the data was checked with FastQC, and the data were preprocessed with Fastp with the following parameters: -g to trim polyG read tails caused by sequencing artifacts, -x to trim polyX, -q 30 for base quality filtering, and -e 30 to filter for an average quality score. Paired-end samples were processed together. After preprocessing, the samples were mapped with STAR-2.7.1a with the following parameters: --outFilterType BySJout --outFilterMultimapNmax 20 --alignSJoverhangMin 8 --alignSJDBoverhangMin 1 --outFilterMismatchNmax 999 --outFilterMismatchNoverReadLmax 0.04 --alignIntronMin 20 --alignIntronMax 1000000 --alignMatesGapMax 1000000 --quantMode GeneCounts.

The genome directories were generated with the ce11 genome, WBcel235.96 without rRNA and the parameter -genomeSAindexNbases 12 for *C. elegans* and the hg38 genome, GRCh38.97 without rRNA, and the parameter -genomeSAindexNbases 14 for human data. The parameter -sjdbOverhang was set to the read length of the sample -1.

The validation samples with the IDs GSE106079, GSE127917, GSE138129, and GSE141041 were mapped with Salmon-1.1 with a k-mer length of 31 and the following parameters: -l A -validateMappings -gcBias -seqBias.

The raw counts for the validation samples with the IDs GSE93826 and GSE138035 were directly downloaded from the gene expression omnibus.

The counts for unstranded RNA-seq were merged into one csv file, and edgeR was used to generate count per millions (CPM).

Functional enrichment analysis was done with String v.11 and geneSCF, and the gene set enrichment analysis with WormExp.

5.2 | Binarization

To binarize the data first zero CPMs were masked by NaN. For the remaining data, the median for each sample was calculated and genes bigger the median were set to 1, while genes smaller or equal to the median were set to 0, finally genes masked by NaN were set to 0 as well.

5.3 | Temporal rescaling

For the temporal rescaling, we set the median lifespan of a standard worm to 15.5 days of adulthood. We calculated a correction factor for every sample by dividing this standard lifespan by the median lifespan reported by the publication of the corresponding sample. We restricted the training data to this subset of samples for which a lifespan was reported in the associated publication, because even a wild-type worm under standard conditions can show dramatically different median lifespans in between different laboratories. For example, the median lifespan of N2 wild-type worms at the same standard conditions in the datasets we used ranges from 15 days in GSE112753 to 24 days in PRJNA508378, which increases to a range from 14 days (GSE65765) to 30.55 days (GSE92902) just by including FUDR-treated worms. Without requiring the lifespan data from the same publication and just setting the lifespan to the standard 15.5 days, we would introduce a twofold bias in the rescaled biological age, which would reduce the prediction of the model accordingly. The chronological age of each sample is multiplied with this correction factor to result in the approximated biological age of the sample. The chronological age, correction factor, and biological age for every sample can be seen in Table S1.

The datasets GSE106079 and GSE93826 were not associated with any publication and thereby no lifespan data were available. However, both datasets consist of a time course of *C. elegans* aging and would therefore be valuable validation data. Since the strains used in both datasets should not show strong deviations in the median lifespan from wild-type worms, we assumed that the lifespan is 15.5 days in both cases. Since this lifespan is approximated and should therefore include a bias as shown above, we would expect the prediction error to be higher than usual.

5.4 | 2nd rescaling approach

For the 2nd rescaling of the biological age, we set the maximum biological age of the worm to 15.5 days. Assuming a normal distribution of biological age around the chronological age of a worm population



and further assuming that, on average, worms will die according to their biological age, we can assume that the maximum biological age of a worm is the median lifespan of 15.5 days. Worms living longer than the median lifespan were biologically younger and therefore did not cross the line of 15.5 days (see Figure S3). Since the first wild-type worms under standard conditions start dying at around 9 days of adulthood, the oldest worms at day 8 should be biologically around 15.5 days old. Therefore, we approximated the standard deviation to be 8/3. Centering a normal distribution at 8 days with a SD of 8/3 will contain 99.73% of the area under the curve within day 0 to day 16.

Next, we approximated that the biological age distribution is not changing over time and that the SD over 8/3 stays stable. To calculate the median of the data after trimming the data at the maximum age of 15.5 days, we first need to calculate how much data are trimmed. We approximate this by utilizing the error function:

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt$$

implemented in the SciPy library.

The approximation of the percentage p of data that is remaining on the left side from the maximum lifespan of 15.5 days on the biological age x is as follows:

$$p = \frac{1}{2} \operatorname{erf} \left(\frac{15.5 - x}{\frac{8}{3} \sqrt{2}} \right) + 0.5$$

Here, $\frac{15.5 - x}{8/3}$ calculates how many SDs the biological age is apart from the maximum age of 15.5 days. And $\operatorname{erf} \left(\frac{15.5 - x}{\frac{8}{3} \sqrt{2}} \right)$ calculates the percentage of the area under the bell curve for the calculated number of SDs. If the biological age would be one SD away from the maximum age of 15.5 days, that is, 8/3 days, the area under the curve would be ~68.2%. However, this value corresponds to the area on the left and the right of the median. Since we are only interested in one side, we have to divide the area by 2 and add 50%, that is, 0.5, for the opposite side. With this, p will approximate the area under the curve that is remaining after trimming the right side from the maximum lifespan of 15.5 days.

To get the approximation of the new median percentage for the trimmed bell curve, we can divide p by 2. This new median percentage can be used to calculate the median in days by reverting the calculation. First, we subtract the new median percentage from 0.5 to get the deviation from the original median percentage, that is, 0.5, and use the inverse error function to approximate s , the number of standard deviations that the new median is shifted to the left of the old median:

$$s = \operatorname{erf}^{-1} \left(0.5 - \frac{p}{2} \right) \sqrt{2} * 2$$

The new median m , in other words the new rescaled biological age, can then be calculated by the following:

$$m = x - s * \frac{8}{3}$$

where 8/3 is the standard deviation that we set in the beginning and x the biological age, that is, the original median.

5.5 | Model fitting—Parameter search

The age prediction models use an elastic net regression as implemented by Python's sklearn. The random_state was set to 0, the max_iter to 1,000, and positive=False. The best parameter settings for alpha and the L1/L2 ratio were selected using a parameter grid search with a nested cross-validation approach. To avoid overfitting during the training, we split the data into multiple partitions. Every sample of the same genetic background, with the same treatment, and RNA interference of the same rounded biological age to days was considered to be one partition. This makes sure that samples with a similar transcriptome are taken out together during the process. The elastic net regression is trained on the remaining data, and the partition that got taken out will be predicted. To get an overview of the accuracy of the model, this process is repeated for the partitions in the dataset. In the end, every sample will be predicted exactly once, which allows the comparison of the predicted with the true biological age.

A simple cross-validation like this gives an overview of the accuracy of the model; however, to select the best parameter setting, a nested cross-validation is required, since otherwise information may leak into the model and introduce another type of overfitting. Therefore, after splitting the data into the test and the train partitions (the outer loop), the latter will be split again into an inner test and train partition (the inner loop). This inner cross-validation will be computed for every parameter set to compute the average of the absolute error for each parameter setting.

This will be done for every partitioning in the outer loop to select the most stable parameter set. The parameters selected by this approach for the binarized data are alpha = 0.075 and l1_ratio = 0.3.

5.6 | Model fitting—Optimal gene set

To obtain the optimal gene set without overfitting, a similar approach was taken. Instead of looping over different parameter settings, the cross-validation for the gene set loops over a list of the genes with the highest absolute coefficients. First, for every training partition in the outer loop the full model with alpha = 0.075 and l1_ratio = 0.3 is computed. This will result in a model, where every gene is annotated with a coefficient. In the binarized model, the sum of the coefficients for all genes that are 1 in the sample added to the intercept equals the predicted age. Therefore, a negative coefficient will result in a younger predicted age, while a positive coefficient will increase the predicted age. Next, we loop over different subsets of the top genes to identify the approximately optimal and smallest gene set for the given partition. For every gene set, the inner cross-validation loop is computed and the gene set with the smallest average absolute error is saved. This



will be done again for every partition in the outer loop to gain multiple gene sets. Similar to the parameter search, the most stable gene set is taken by retaining only those genes that were used by every partition. This stable gene set selected by this approach for the binarized data after the second rescaling are the 576 genes described in Table S2. This final model starts at an intercept of 103.55 hrs (4.31 days).

5.7 | Using the clock

To predict the biological age of new data, one has to start with binarizing the transcriptome as described above. The elastic net coefficients (column 2 in Table S2) are added up for all of the 576 genes with a value of 1 after binarization. Finally, the intercept of 103.55 hr has to be added to get the final prediction of the biological age in hours. The code is included in <https://github.com/Meyer-DH/AgingClock/>

5.8 | Motif search

The set of genes with a coefficient >0 , respective ≤ 0 , was used as input for the findMotifs function of Homer-4.9.1-6 with the parameters -len 8,10 -start -300 -end 100. To make sure that the maximum number of genes got recognized by Homer, we first converted the Wormbase IDs to the sequence name with WormBase's SimpleMine and added "CELE_" in front of it. These identifiers were then searched in the "worm.description" file of Homer to gain the corresponding RefSeq IDs that are recognized by the program. The p-values were calculated with a hypergeometric test.

5.9 | Median lifespan fold change prediction

The median lifespan fold change can be predicted by the biological age of the strain of interest and its control, assuming a uniform age effect. The median lifespan of each strain can be computed by dividing the chronological age by the biological age and multiplying it by 15.5 days. To compute the fold change, the median lifespan of interest is divided by the control lifespan, or easier, the biological age of the strain of interest can be divided by the biological age of the control, if the chronological age is the same.

The theoretical range of lifespan fold change predictions in Figure S8 was calculated with the Python package Uncertainties. The chronological age bias was set to 0.5 days and the lifespan assay bias to 5%. The code is included in <https://github.com/Meyer-DH/AgingClock/>

5.10 | Figure details

All plots were done with Seaborn-0.9.0. Boxplots: The center line represents the median; the box limits the bottom, and top quartiles of the data and the whiskers show the 1.5x interquartile range.

5.11 | Statistics

ANOVA and *t* tests were computed with Python's pingouin library v.0.3.3. post hoc Tukey test were computed with Python's Statsmodels library v.0.10.1.

5.12 | Citations of the age predictors from the literature

Because currently no general consensus of quality assessment exists and different measurements are being reported, we state the measurements as reported in the cited paper in the introduction. Some of the most common used assessments are as follows:

1. Mean absolute error (MAE): the mean of the absolute difference in predicted and true age.
2. Root-mean-square-error (RMSE): the square root of the average squared differences. Larger errors have a larger effect on the RMSE than on MAE.
3. Median absolute deviation (MAD): the median absolute difference in predicted and true age.
4. Pearson correlation (*r*): measurement of how the predicted and true age changes together. Evaluates linear relationships.
5. Spearman correlation (*r*): similar to Pearson correlation, but evaluates the monotonic relationship. Other than Pearson correlation, the variables do not need to change at a linear rate.
6. Coefficient of determination (r^2): the fraction of the variance that is predictable with the model. Often the r^2 is the square of the correlation coefficient; however, this is not true in the general case. The value can get negative if the model fits worse than a horizontal line.

ACKNOWLEDGMENTS

We thank the Regional Computing Center of the University of Cologne (RRZK) for providing computing time on the DFG-funded High Performance Computing (HPC) system CHEOPS as well as support. D.M. is member of the Cologne Graduate School of Ageing Research. B.S. acknowledges funding from the Deutsche Forschungsgemeinschaft (SCHU 2494/3-1, SCHU 2494/7-1, SCHU 2494/10-1, SCHU 2494/11-1, CECAD EXC 2030 - 390661388, SFB 829, KFO 286, KFO 329, and GRK2407), the Deutsche Krebshilfe (70112899), and the H2020-MSCA-ITN-2018 (Healthage and ADDRESS ITNs). Open access funding enabled and organized by ProjektDEAL.

CONFLICT OF INTEREST

The authors declare no competing interests.

AUTHOR CONTRIBUTIONS

D.M. conceived and designed the study and performed all bioinformatics analysis; B.S. coordinated the project and together with D.M. designed the study. All authors wrote the paper.



DATA AVAILABILITY STATEMENT

We used publicly available datasets. The details can be found in Table S1.

CODE AVAILABILITY STATEMENT

Code for the binarization, age-correction, and the prediction of new samples can be found on <https://github.com/Meyer-DH/AgingClock/>.

ORCID

David H. Meyer  <https://orcid.org/0000-0002-5667-4720>

Björn Schumacher  <https://orcid.org/0000-0001-6097-5238>

REFERENCES

- Admasu, T. D., Chaithanya Batchu, K., Barardo, D., Ng, L. F., Lam, V. Y. M., Xiao, L., Cazenave-Gassiot, A., Wenk, M. R., Tolwinski, N. S., & Gruber, J. (2018). Drug synergy slows aging and improves healthspan through IGF and SREBP lipid signaling. *Developmental Cell*, 47(1), 67–79.e5. <https://doi.org/10.1016/j.devcel.2018.09.001>
- Ahadi, S., Zhou, W., Rose, S. M. S., Sailani, M. R., Contrepolis, K., Avina, M., Ashland, M., Brunet, A., & Snyder, M. (2020). Personal aging markers and ageotypes revealed by deep longitudinal profiling. *Nature Medicine*, 26, 83–90.
- Ardiel, E. L., Giles, A. C., Yu, A. J., Lindsay, T. H., Lockery, S. R., & Rankin, C. H. (2016). Dopamine receptor DOP-4 modulates habituation to repetitive photoactivation of a *C. elegans* polymodal nociceptor. *Learning & Memory*, 23, 495–503.
- Beck, C. D. O., & Rankin, C. H. (1993). Effects of aging on habituation in the nematode *Caenorhabditis elegans*. *Behavioural Processes*, 28, 145–163.
- Budovskaya, Y. V., Wu, K., Southworth, L. K., Jiang, M., Tedesco, P., Johnson, T. E., & Kim, S. K. (2008). An elt-3/elt-5/elt-6 GATA transcription circuit guides aging in *C. elegans*. *Cell*, 134, 291–303. Retrieved from <https://linkinghub.elsevier.com/retrieve/pii/S0092867408007071>
- Enge, M., Arda, H. E., Mignardi, M., Beausang, J., Bottino, R., Kim, S. K., & Quake, S. R. (2017). Single-cell analysis of human pancreas reveals transcriptional signatures of aging and somatic mutation patterns. *Cell*, 171(2), 321–330.e14. <https://doi.org/10.1016/j.cell.2017.09.004>
- Ermolaeva, M. A., & Schumacher, B. (2014). Insights from the worm: The *C. elegans* model for innate immunity. *Seminars in Immunology*, 26(4), 303–309. <https://doi.org/10.1016/j.smim.2014.04.005>
- Finger, F., Ottens, F., Springhorn, A., Drexel, T., Proksch, L., Metz, S., Cochella, L., & Hoppe, T. (2019). Olfaction regulates organismal proteostasis and longevity via microRNA-dependent signalling. *Nature Metabolism*, 1(3), 350–359. <https://doi.org/10.1038/s42255-019-0033-z>
- Fleischer, J. G., Schulte, R., Tsai, H. H., Tyagi, S., Ibarra, A., Shokhirev, M. N., Huang, L., Hetzer, M. W., & Navlakha, S. (2018). Predicting age from the transcriptome of human dermal fibroblasts. *Genome Biology*, 19, 1–8.
- Fortney, K., Kotlyar, M., & Jurisica, I. (2010). Inferring the functions of longevity genes with modular subnetwork biomarkers of *Caenorhabditis elegans* aging. *Genome Biology*, 11, R13–<https://doi.org/10.1186/gb-2010-11-2-r13>
- Galkin, F., Mamoshina, P., Aliper, A., de Magalhães, J. P., Gladyshev, V. N., & Zhavoronkov, A. (2020). Biohorology and biomarkers of aging: Current state-of-the-art, challenges and opportunities. *Ageing Research Reviews*, 60, 101050. Retrieved from <https://linkinghub.elsevier.com/retrieve/pii/S1568163719302582>
- Glenn, C. F., Chow, D. K., David, L., Cooke, C. A., Gami, M. S., Iser, W. B., Hanselman, K. B., Goldberg, I. G., & Wolkow, C. A. (2004). Behavioral deficits during early stages of aging in *Caenorhabditis elegans* result from locomotory deficits possibly linked to muscle frailty. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 59, 1251–1260. Retrieved from <https://academic.oup.com/biomedgerontology/article-lookup>. <https://doi.org/10.1093/gerona/59.12.1251>
- Golden, T. R., Hubbard, A., Dando, C., Herren, M. A., & Melov, S. (2008). Age-related behaviors have distinct transcriptional profiles in *Caenorhabditis elegans*. *Aging Cell*, 7, 850–865. <https://doi.org/10.1111/j.1474-9726.2008.00433.x>
- Hastings, J., Mains, A., Virk, B., Rodriguez, N., Murdoch, S., Pearce, J., Bergmann, S., Le Novère, N., & Casanueva, O. (2019). Multi-omics and genome-scale modeling reveal a metabolic shift during *C. elegans* aging. *Frontiers in Molecular Biosciences*, 6, 1–18.
- Inukai, S., Pincus, Z., de Lencastre, A., & Slack, F. J. (2018). A microRNA feedback loop regulates global microRNA abundance during aging. *RNA*, 24, 159–172. Retrieved from <http://rnajournal.cshlp.org/lookup>. <https://doi.org/10.1261/rna.062190.117>
- Joshi, K. K., Matlack, T. L., & Rongo, C. (2016). Dopamine signaling promotes the xenobiotic stress response and protein homeostasis. *EMBO Journal*, 35, 1885–1901. Retrieved from <https://onlinelibrary.wiley.com/doi/abs/10.15252/embj.201592524>
- Lai, R. W., Lu, R., Danthi, P. S., Bravo, J. I., Goumba, A., Sampathkumar, N. K., & Benayoun, B. A. (2019). Multi-level remodeling of transcriptional landscapes in aging and longevity. *BMB Reports*, 52, 86–108.
- Mamoshina, P., Volosnikova, M., Ozerov, I. V., Putin, E., Skibina, E., Cortese, F., & Zhavoronkov, A. (2018). Machine learning on human muscle transcriptomic data for biomarker discovery and tissue-specific drug target identification. *Frontiers in Genetics*, 9, 1–10. Retrieved from <https://www.frontiersin.org/article/10.3389/fgene.2018.00242/full>
- Mann, F. G., Van Nostrand, E. L., Friedland, A. E., Liu, X., & Kim, S. K. (2016). Deactivation of the GATA transcription factor ELT-2 is a major driver of normal aging in *C. elegans*. *PLOS Genetics*, 12(4), e1005956–<https://doi.org/10.1371/journal.pgen.1005956>
- Martins, R., Lithgow, G. J., & Link, W. (2016). Long live FOXO: Unraveling the role of FOXO proteins in aging and longevity. *Aging Cell*, 15, 196–207.
- Nedialkova, D. D., & Leidel, S. A. (2015). Optimization of codon translation rates via tRNA modifications maintains proteome integrity. *Cell*, 161(7), 1606–1618. <https://doi.org/10.1016/j.cell.2015.05.022>
- Nhan, J. D., Turner, C. D., Anderson, S. M., Yen, C., Dalton, H. M., Cheesman, H. K., Ruter, D. L., Uma Naresh, N., Haynes, C. M., Soukas, A. A., Pukhila-Worley, R., & Curran, S. P. (2019). Redirection of SKN-1 abates the negative metabolic outcomes of a perceived pathogen infection. *Proceedings of the National Academy of Sciences*, 116, 22322–22330. Retrieved from <http://www.pnas.org/lookup/doi/10.1073/pnas.1909666116>
- Pang, S., & Curran, S. P. (2014). Adaptive capacity to bacterial diet modulates aging in *C. elegans*. *Cell Metabolism* 19, 221–231. <https://doi.org/10.1016/j.cmet.2013.12.005>
- Peters, M. J., Joehanes, R., Pilling, L. C., Schurmann, C., Conneely, K. N., Powell, J., Reinmaa, E., Sutphin, G. L., Zhernakova, A., Schramm, K., Wilson, Y. A., Kobes, S., Tukiainen, T., Ramos, Y. F., Göring, H. H. H., Fornage, M., Liu, Y., Gharib, S. A., Stranger, B. E., ... Johnson, A. D. (2015). The transcriptional landscape of age in human peripheral blood. *Nature Communications*, 6, 8570. Retrieved from <http://www.nature.com/articles/ncomms9570>
- Petrasccheck, M., & Miller, D. L. (2017). Computational analysis of lifespan experiment reproducibility. *Frontiers in Genetics*, 8, 1–11.
- Pryor, R., Norvaisas, P., Marinos, G., Best, L., Thingholm, L. B., Quintaneiro, L. M., De Haes, W., Esser, D., Waschina, S., Lujan, C., Smith, R. L., Scott, T. A., Martinez-Martinez, D., Woodward, O., Bryson, K., Laudes, M., Lieb, W., Houtkooper, R. H., Franke, A., ...



- Cabreiro, F. (2019). Host-microbe-drug-nutrient screen identifies bacterial effectors of metformin therapy. *Cell*, 178, 1299–1312.e29.
- Rangaraju, S., Solis, G. M., Thompson, R. C., Gomez-Amaro, R. L., Kurian, L., Encalada, S. E., Niculescu, A. B., Salomon, D. R., & Petrascheck, M. (2015). Suppression of transcriptional drift extends *C. elegans* lifespan by postponing the onset of mortality. *Elife*, 4, e08833. Retrieved from <https://elifesciences.org/articles/08833>
- Ratnappan, R., Amrit, F. R. G., Chen, S.-W., Gill, H., Holden, K., Ward, J., Yamamoto, K. R., Olsen, C. P., & Ghazi, A. (2014). Germline signals deploy NHR-49 to modulate fatty-acid β -oxidation and desaturation in somatic tissues of *C. elegans*. *PLoS Genetics*, 10(12), e1004829–<https://doi.org/10.1371/journal.pgen.1004829>
- Schmeisser, S., Priebe, S., Groth, M., Monajembashi, S., Hemmerich, P., Guthke, R., Platzer, M., & Ristow, M. (2013). Neuronal ROS signaling rather than AMPK/sirtuin-mediated energy sensing links dietary restriction to lifespan extension. *Molecular Metabolism*, 2, 92–102. <https://doi.org/10.1016/j.molmet.2013.02.002>
- Sonowal, R., Swimm, A., Sahoo, A., Luo, L., Matsunaga, Y., Wu, Z., Bhingarde, J. A., Ejzak, E. A., Ranawade, A., Qadota, H., Powell, D. N., Capaldo, C. T., Flacker, J. M., Jones, R. M., Benian, G. M., & Kalman, D. (2017). Indoles from commensal bacteria extend healthspan. *Proceedings of the National Academy of Sciences*, 114, E7506–E7515.
- Stroustrup, N., Anthony, W. E., Nash, Z. M., Gowda, V., Gomez, A., López-Moyado, I. F., Apfeld, J., & Fontana, W. (2016). The temporal scaling of *Caenorhabditis elegans* ageing. *Nature*, 530, 103–107.
- Sun, X., Chen, W.-D., & Wang, Y.-D. (2017). DAF-16/FOXO transcription factor in aging and longevity. *Frontiers in Pharmacology*, 8, 1–8. <https://doi.org/10.3389/fphar.2017.00548>
- Tarkhov, A. E., Alla, R., Ayyadevara, S., Pyatnitskiy, M., Menshikov, L. I., Shmookler Reis, R. J., & Fedichev, P. O. (2019). A universal transcriptomic signature of age reveals the temporal scaling of *Caenorhabditis elegans* aging trajectories. *Scientific Reports*, 9(1), <https://doi.org/10.1038/s41598-019-43075-z>
- Tepper, R. G., Ashraf, J., Kaletsky, R., Kleemann, G., Murphy, C. T., & Bussemaker, H. J. (2013). PQM-1 complements DAF-16 as a key transcriptional regulator of DAF-2-mediated development and longevity. *Cell*, 154, 676–690. Retrieved from <https://linkinghub.elsevier.com/retrieve/pii/S0022202X15370834>
- Tikiyani, V., Li, L., Sharma, P., Liu, H., Hu, Z., & Babu, K. (2018). Wnt secretion is regulated by the tetraspan protein HIC-1 through its interaction with neurabin/NAB-1. *Cell Reports*, 25(7), 1856–1871.e6. <https://doi.org/10.1016/j.celrep.2018.10.053>
- Vilchez, D., Morantte, I., Liu, Z., Douglas, P. M., Merkwirth, C., Rodrigues, A. P. C., Manning, G., & Dillin, A. (2012). RPN-6 determines *C. elegans* longevity under proteotoxic stress conditions. *Nature*, 489, 263–268.
- Webster, C. M., Pino, E. C., Carr, C. E., Wu, L., Zhou, B., Cedillo, L., Kacergis, M. C., Curran, S. P., & Soukas, A. A. (2017). Genome-wide RNAi screen for fat regulatory genes in *C. elegans* identifies a proteostasis-AMPK axis critical for starvation survival. *Cell Reports*, 20(3), 627–640. <https://doi.org/10.1016/j.celrep.2017.06.068>
- Wu, Z., Isik, M., Moroz, N., Steinbaugh, M. J., Zhang, P., & Blackwell, T. K. (2019). Dietary restriction extends lifespan through metabolic regulation of innate immunity. *Cell Metabolism*, 29, 1–23.
- Yen, C.-A., Ruter, D. L., Turner, C. D., Pang, S., & Curran, S. P. (2020). Loss of flavin adenine dinucleotide (FAD) impairs sperm function and male reproductive advantage in *C. elegans*. *Elife*, 9, 1–22. Retrieved from <https://elifesciences.org/articles/52899>
- Yin, J. A., Gao, G., Liu, X. J., Hao, Z. Q., Li, K., Kang, X. L., Li, H., Shan, Y. H., Hu, W. L., Li, H. P., & Cai, S. Q. (2017). Genetic variation in glia-neuron signalling modulates ageing rate. *Nature*, 551(7679), 198–203. <https://doi.org/10.1038/nature24463>
- Youngman, M. J., Rogers, Z. N., & Kim, D. H. (2011). A decline in p38 MAPK signaling underlies immunosenescence in *Caenorhabditis elegans*. S. K. Kim, ed. *PLoS Genetics*, 7, e1002082. Retrieved from <https://dx.plos.org/10.1371/journal.pgen.1002082>
- Yunger, E., Safra, M., Levi-Ferber, M., Haviv-Chesner, A., & Henis-Korenblit, S. (2017). Innate immunity mediated longevity and longevity induced by germ cell removal converge on the C-type lectin domain protein IRG-7 M.-W. Tan, ed. *PLoS Genetics*, 13, e1006577. Retrieved from <https://dx.plos.org/10.1371/journal.pgen.1006577>
- Zarse, K., Schmeisser, S., Groth, M., Priebe, S., Beuster, G., Kuhlow, D., Guthke, R., Platzer, M., Kahn, C. R., & Ristow, M. (2012). Impaired insulin/IGF1 signaling extends life span by promoting mitochondrial L-proline catabolism to induce a transient ROS signal. *Cell Metabolism*, 15(4), 451–465. <https://doi.org/10.1016/j.cmet.2012.02.013>
- Zhao, Y., Gilliat, A. F., Ziehm, M., Turmaine, M., Wang, H., Ezcurra, M., Yang, C., Phillips, G., McBay, D., Zhang, W. B., Partridge, L., Pincus, Z., & Gems, D. (2017). Two forms of death in ageing *Caenorhabditis elegans*. *Nature Communications*, 8(1), 1–8. <https://doi.org/10.1038/ncomms15458>
- Zullo, J. M., Drake, D., Aron, L., O'Hern, P., Dhamne, S. C., Davidsohn, N., Mao, C.-A., Klein, W. H., Rotenberg, A., Bennett, D. A., Church, G. M., Colaiacovo, M. P., & Yankner, B. A. (2019). Regulation of lifespan by neural excitation and REST. *Nature*, 574, 359–364.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Meyer DH, Schumacher B. BiT age: A transcriptome-based aging clock near the theoretical limit of accuracy. *Aging Cell*. 2021;20:e13320. <https://doi.org/10.1111/ace1.13320>

Supplementary materials and methods

Programs and methods citations

The following programs and methods have been used in this study:

FastQC (Andrews et al. 2010), Fastp (Chen et al. 2018), STAR-2.7.1a (Dobin et al. 2013), Salmon-1.1 (Patro et al. 2017), edgeR (Robinson et al. 2009), String v.11 (Szklarczyk et al. 2019), geneSCF (Subhash & Kanduri 2016), WormBase's SimpleMine (Harris et al. 2020), Homer-4.9.1-6 (Heinz et al. 2010), WormExp (Yang et al. 2016).

The following Python libraries have been used:

pingouin v.0.3.3 (Vallat 2018), Statsmodels v.0.10.1 (Seabold & Perktold 2010), Scipy-1.5.1 (Virtanen et al. 2020), sklearn-0.23.1 (Varoquaux et al. 2011), Uncertainties-3.1.1 (Lebigot n.d.), seaborn-0.9.0 (Waskom et al. 2018)

Datasets have been downloaded from the gene expression omnibus (Edgar et al. 2002).

Supplementary References

Andrews S, Krueger F, Segonds-Pichon, Anne Biggins L, Krueger C & Wingett S (2010) FastQC. Available at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>.

Chen S, Zhou Y, Chen Y & Gu J (2018) Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, i884–i890.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M & Gingeras TR (2013) STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21.

Edgar R, Domrachev M & Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30, 207–210. Available at: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/30.1.207>.

Harris TW, Arnaboldi V, Cain S, Chan J, Chen WJ, Cho J, Davis P, Gao S, Grove CA, Kishore R, Lee RYN, Muller HM, Nakamura C, Nuin P, Paulini M, Raciti D, Rodgers FH, Russell M, Schindelman G, Auken K V., Wang Q, Williams G, Wright AJ, Yook K, Howe KL, Schedl T, Stein L & Sternberg PW (2020) WormBase: a modern Model Organism Information Resource. *Nucleic Acids Res.* 48, D762–D767.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H & Glass CK (2010) Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* 38, 576–589. Available at: <http://dx.doi.org/10.1016/j.molcel.2010.05.004>.

Lebigot EO Uncertainties: a Python package for calculations with uncertainties. Available at: <https://pythonhosted.org/uncertainties/>.

Patro R, Duggal G, Love MI, Irizarry RA & Kingsford C (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* 14, 417–419. Available at: <http://www.nature.com/articles/nmeth.4197>.

Robinson MD, McCarthy DJ & Smyth GK (2009) edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140.

Seabold S & Perktold J (2010) Statsmodels: Econometric and Statistical Modeling with Python. *PROC.*

9th PYTHON Sci. CONF, 57. Available at: <http://statsmodels.sourceforge.net/>.

Subhash S & Kanduri C (2016) GeneSCF: A real-time based functional enrichment tool with support for multiple organisms. *BMC Bioinformatics* 17, 1–10. Available at: <http://dx.doi.org/10.1186/s12859-016-1250-z>.

Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ & Von Mering C (2019) STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, D607–D613.

Vallat R (2018) Pingouin: statistics in Python. *J. Open Source Softw.* 3, 1026.

Varoquaux G, Buitinck L, Louppe G, Grisel O, Pedregosa F & Mueller A (2011) Scikit-learn: Machine Learning in Python Fabian. *J. Mach. Learn. Res.*

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P, Vijaykumar A, Bardelli A Pietro, Rothberg A, Hilboll A, Kloeckner A, Scopatz A, Lee A, Rokem A, Woods CN, Fulton C, Masson C, Häggström C, Fitzgerald C, Nicholson DA, Hagen DR, Pasechnik D V., Olivetti E, Martin E, Wieser E, Silva F, Lenders F, Wilhelm F, Young G, Price GA, Ingold GL, Allen GE, Lee GR, Audren H, Probst I, Dietrich JP, Silterra J, Webber JT, Slavič J, Nothman J, Buchner J, Kulick J, Schönberger JL, de Miranda Cardoso JV, Reimer J, Harrington J, Rodríguez JLC, Nunez-Iglesias J, Kuczynski J, Tritz K, Thoma M, Newville M, Kümmerer M, Bolingbroke M, Tartre M, Pak M, Smith NJ, Nowaczyk N, Shebanov N, Pavlyk O, Brodtkorb PA, Lee P, McGibbon RT, Feldbauer R, Lewis S, Tygier S, Sievert S, Vigna S, Peterson S, More S, Pudlik T, Oshima T, Pingel TJ, Robitaille TP, Spura T, Jones TR, Cera T, Leslie T, Zito T, Krauss T, Upadhyay U, Halchenko YO & Vázquez-Baeza Y (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* 17, 261–272.

Waskom M, Botvinnik O, O’Kane D, Hobson P, Ostblom J, Lukauskas S, Gemperline DC, Augspurger T, Halchenko Y, Cole JB, Warmenhoven J, de Ruiter J, Pye C, Hoyer S, Vanderplas J, Villalba S, Kunter G, Quintero E, Bachant P, Martin M, Meyer K, Miles A, Ram Y, Brunner T, Yarkoni T, Williams ML, Evans C, Fitzgerald C, Brian & Qalieh A (2018) mwaskom/seaborn: v0.9.0 (July 2018). Available at: <https://doi.org/10.5281/zenodo.1313201>.

Yang W, Dierking K & Schulenburg H (2016) WormExp: A web-based application for a *Caenorhabditis elegans*-specific gene expression enrichment analysis. *Bioinformatics* 32, 943–945.

Supporting Information Legends

Figure S1. Alternative models

(A) Results of the biological age prediction computed by cross-validation. The x-axis shows the rescaled biological age in days starting from adulthood. The y-axis shows the predicted age computed by an elastic net regression on unbinarized CPMs. Every blue dot displays one RNA-seq sample. The regression line with the 95 % confidence interval is shown in blue and the dotted line shows the perfect linear correlation. The distribution of the data is shown on the side of the plot. r^2 = coefficient of determination, Pearson= Pearson correlation, Spearman= Spearman correlation, MAE= mean absolute error in days, MAD= median absolute deviation in days, RMSE= root-mean-square-error in days.

(B) Results of the biological age prediction computed by cross-validation. The x-axis shows the rescaled biological age in days starting from adulthood. The y-axis shows the predicted age computed by an elastic net regression on binarized gene expression data. Every blue dot displays one RNA-seq sample. The regression line with the 95 % confidence interval is shown in blue and the dotted line shows the perfect linear correlation. The distribution of the data is shown on the side of the plot. r^2 = coefficient of determination, Pearson= Pearson correlation, Spearman= Spearman correlation, MAE= mean absolute error in days, MAD= median absolute deviation in days, RMSE= root-mean-square-error in days.

(C) Prediction of the model on 8 independent datasets consisting of 94 samples at different time points. The x-axis shows the biological age in days starting from adulthood before the second rescaling approach. The y-axis shows the predicted age computed by an elastic net regression on binarized gene expression data. For more details on the datasets see the Table S1.

Figure S2. Comparison of the binarized and unbinarized model error

(A) The absolute error distribution between the predicted and true biological age is plotted for either the unbinarized (red) or binarized (blue) data. The x-axis shows the true biological age in days. The y-axis the absolute error in days. While the unbinarized model strongly increases the absolute prediction error with age, the increase is less pronounced with the binarized model.

(B) The bar plots show the standard deviation of the absolute prediction errors in days. The x-axis shows the true biological age in days. While the binarized model stays relatively stable over age, the unbinarized model increases the variance in the prediction error.

Figure S3. Explanation of the 2nd rescaling

(A, C, E) Standard lifespan curves of *C. elegans* with a median lifespan of 15.5 days. The X mark the chronological age for which we show the hypothetical age distributions in (B, D, F) respectively. (B, D, F) show the biological age distribution around the chronological age marked by the X. The biggest portion of the age-synchronized worm population will be as old as the chronological age. However, assuming a normal distribution of the biological age, we can assume that a part of the population is biologically younger, respective older. The green lines indicate the median biological age of the living worm population. The dotted line displays the maximum lifespan.

(A, B) All non-censored worms are still alive in the population, i.e. no worm crossed the maximum lifespan line. The population age median is equal to the peak of the distribution.

(C, D) The first (biologically older) worms died, leading to a truncation of the alive distribution of biological age in the population. This has the consequence that the true median of the alive fraction of the worms will be shifted to the left, away from the peak of the distribution.

(E, F) At the median lifespan, 50 % of the population has died. Assuming a uniform shift of the biological age distribution results in the truncation of the right half of the distribution. The true population median is therefore even further shifted to the left.

Figure S4. Comparison of the model with unbinarized data, random genes and the theoretical limit

(A) Prediction of the 8 independent datasets consisting of 94 samples at different time points. The x-axis shows the rescaled biological age in days starting from adulthood additionally corrected by the second rescaling approach. The y-axis shows the predicted age computed by an elastic net regression on unbinarized CPMs. For more details on the data see the Table S1. (B) The y-axis shows the mean absolute error (MAE), respective the median absolute deviation (MAD) of a given prediction in days. The box plots display the results of 1000 random models with 576 (binarized) genes. The prediction by our final model with a MAE of 0.45 and a MAD of 0.32 is shown as the blue dots and indicated by

arrows. The dotted lines show the theoretical limit of prediction given by the limit of accuracy in the chronological age annotation as well as variance in the lifespan data used for rescaling.

Figure S5. Comparison of our gene set to published gene sets

Results of the biological age prediction computed by cross-validation based on different gene sets predicted by Tarkhov et al. (Tarkhov et al. 2019). The x-axis shows the rescaled biological age in days starting from adulthood additionally corrected by the second rescaling approach. The y-axis shows the predicted age computed by an elastic net regression on unbinarized (A, B, C) or binarized (D, E, F) gene expression data. Every blue dot displays one RNA-seq sample. The regression lines with the 95 % confidence intervals are shown in blue and the dotted lines show the perfect linear correlation. The distribution of the data is shown on the side of the plot. r^2 = coefficient of determination, Pearson= Pearson correlation, Spearman= Spearman correlation, MAE= mean absolute error in days, MAD= median absolute deviation in days, RMSE= root-mean-square-error in days.

(A) Prediction based on the unbinarized CPMs of 327 genes generated by a meta-analysis of publicly available microarray data.

(B) Prediction based on the unbinarized CPMs of 902 age-associated genes generated by an RNA-seq experiment.

(C) Prediction based on the unbinarized CPMs of a sparse subset with 71 genes.

(D) Prediction based on the binarized CPMs of the 327 genes generated by a meta-analysis of publicly available microarray data shown in (A).

(E) Prediction based on the binarized CPMs of the 902 age-associated genes generated by an RNA-seq experiment shown in (B).

(F) Prediction based on the binarized CPMs of the sparse subset with 71 genes shown in (C).

Figure S6. Comparison of our gene set to published gene sets on the validation data

Prediction of the 8 independent datasets consisting of 94 samples at different time points based on different gene sets predicted by Tarkhov et al. (Tarkhov et al. 2019). The x-axis shows the rescaled biological age in days starting from adulthood additionally corrected by the second rescaling approach. The y-axis shows the predicted age computed by an elastic net regression on unbinarized

(A, B, C) or binarized (D, E, F) gene expression data. Every blue dot displays one RNA-seq sample. The regression lines with the 95 % confidence intervals are shown in blue and the dotted lines show the perfect linear correlation. The distribution of the data is shown on the side of the plot. r^2 = coefficient of determination, Pearson= Pearson correlation, Spearman= Spearman correlation, MAE= mean absolute error in days, MAD= median absolute deviation in days, RMSE= root-mean-square-error in days.

(A) Prediction based on the unbinarized CPMs of 327 genes generated by a meta-analysis of publicly available microarray data.

(B) Prediction based on the unbinarized CPMs of 902 age-associated genes generated by an RNA-seq experiment.

(C) Prediction based on the unbinarized CPMs of a sparse subset with 71 genes.

(D) Prediction based on the binarized CPMs of the 327 genes generated by a meta-analysis of publicly available microarray data shown in (A).

(E) Prediction based on the binarized CPMs of the 902 age-associated genes generated by an RNA-seq experiment shown in (B).

(F) Prediction based on the binarized CPMs of the sparse subset with 71 genes shown in (C).

Figure S7. Biological age prediction of additional samples

(A) The genotype-dependent effect of dietary restriction (DR) is resembled in the prediction of chronologically 6-day adults. A two-way ANOVA shows a significant interaction effect ($p=0.004$) between the genotype and the diet. AL = *ad libitum* fed. Data from GSE92909.

(B) The change in diet from K12 to K12 Δ *tnaA* *E. coli* shows an increasing trend, especially in chronologically older population, as indicated by the different colors. A two-way ANOVA shows a significant diet effect ($p=0.03$) and almost significant interaction effect ($p=0.067$). Data from GSE101910.

Figure S8. Theoretical error in the prediction of the median lifespan from the biological age

This plot visualizes the intrinsic random error that propagates from the biological age calculation to the fold-change. The x-axis shows the chronological age in days starting from adulthood. The y-axis shows the calculated fold-change between 2 lifespan curves. 3 lifespan comparisons are shown (color-coded). The control median lifespan is always set to 15.5 days, while the second lifespan is variable at 8 days (blue), 15.5 days (orange), and 31 days (green). The same intrinsic biases as in Fig. 2c and Fig. S4b are considered, i.e. a chronological age reporting error of +/- 12 h and a moderate 5 % lifespan variation. For each chronological age point the biological age was calculated with error propagation. The 2 biological age points were then used to approximate the lifespan fold-change for the 3 examples shown. The lines show the average fold-change, e.g. if both lifespans were at 15.5 days (orange), the expected fold-change is at 1.0, i.e. no change. The random error especially introduces a potential bias in the prediction based on chronologically younger samples, i.e. the shadow around the lines.

Figure S9. Chromosome enrichment

(A) Chromosome distribution of the 286 protein-coding predictor genes with a coefficient ≤ 0 in blue and the number of protein-coding genes that would be expected if the genes were randomly distributed among the chromosomes in red.

(B) Chromosome distribution of the 260 protein-coding predictor genes with a coefficient > 0 in blue and the number of protein-coding genes that would be expected if the genes were randomly distributed among the chromosomes in red.

(C) Differences of the observed to the expected numbers in percent for the protein-coding genes with a coefficient > 0 in blue and with a coefficient ≤ 0 in red.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, Hypergeometric tests were performed and the resulting p-values were corrected with the Benjamini-Hochberg procedure. Table S3 contains more detailed statistics.

Figure S10. Motif enrichment nearby the TSS

Results of a motif enrichment analysis for the region -300 bp to +100 bp from the transcription start site of the genes with a coefficient ≤ 0 (A) and genes with a coefficient > 0 (B). The columns show the name of the transcription factor in the first column with the known motif in the second column. Column 3 and 4 show the percentage of target genes, respective background genes, containing the

motif in the described region. Column 5 shows the fold change enrichment, column 6 the corresponding Hypergeometric p-value and the last column the Benjamini-Hochberg adjusted q-value.

Figure S11. Unbinarized human data

(A) Results of the age prediction computed by cross-validation on human fibroblast gene expression data. The x-axis shows the chronological age in years. The y-axis shows the predicted age computed by an elastic net regression on unbinarized gene expression data. Every blue dot displays one RNA-seq sample. The regression line with the 95 % confidence interval is shown in blue and the dotted line shows the perfect linear correlation. The distribution of the data is shown on the side of the plot. r^2 = coefficient of determination, Pearson= Pearson correlation, Spearman= Spearman correlation, MAE= mean absolute error in years, MAD= median absolute deviation in years, RMSE= root-mean-square-error in years. Data from GSE113957.

(B) Box plots of age predictions of samples from Hutchinson–Gilford progeria syndrome patients (red) and predictions of age-matched healthy controls (blue) by the elastic net regression of unbinarized gene expression data. Progeria samples show no significant increase in the predicted age compared to age-matched healthy controls. Data from GSE113957.

The p-value was calculated by an independent two-sided t-test. Table S3 contains more detailed statistics.

Supplementary Tables

The following supplementary tables can be found at:

<https://onlinelibrary.wiley.com/doi/full/10.1111/accel.13320>

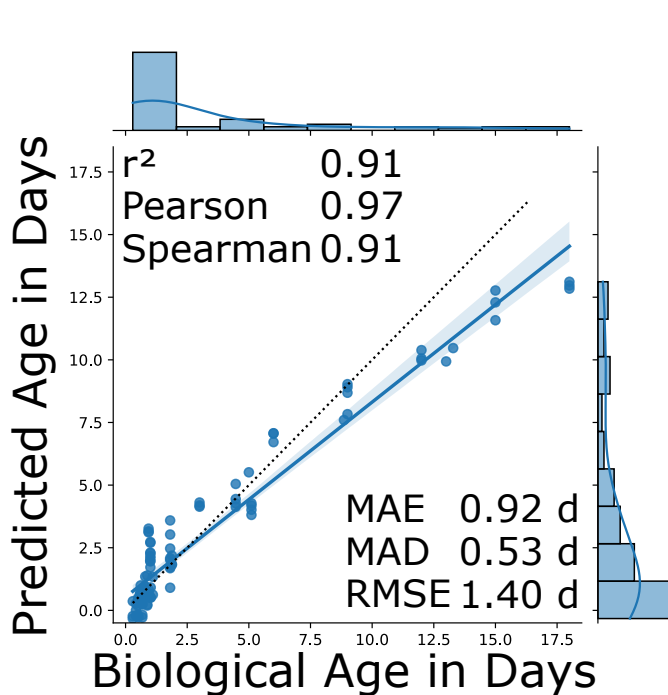
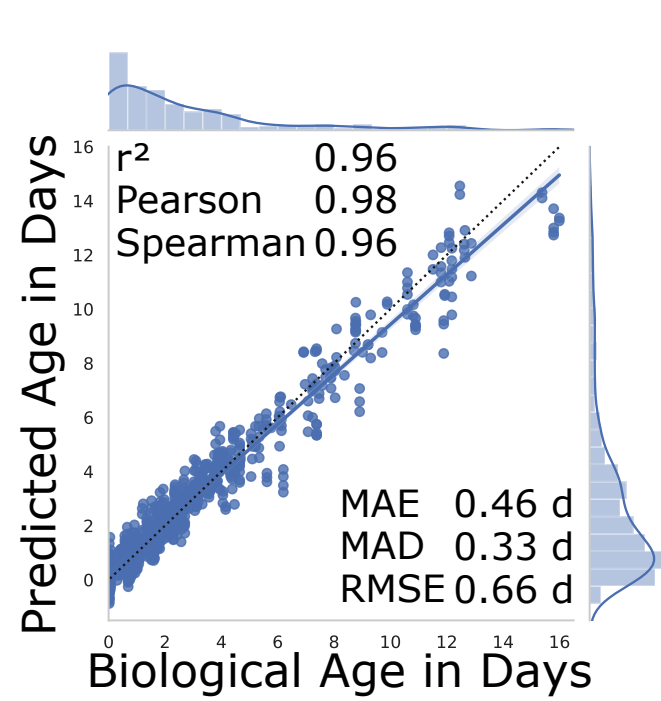
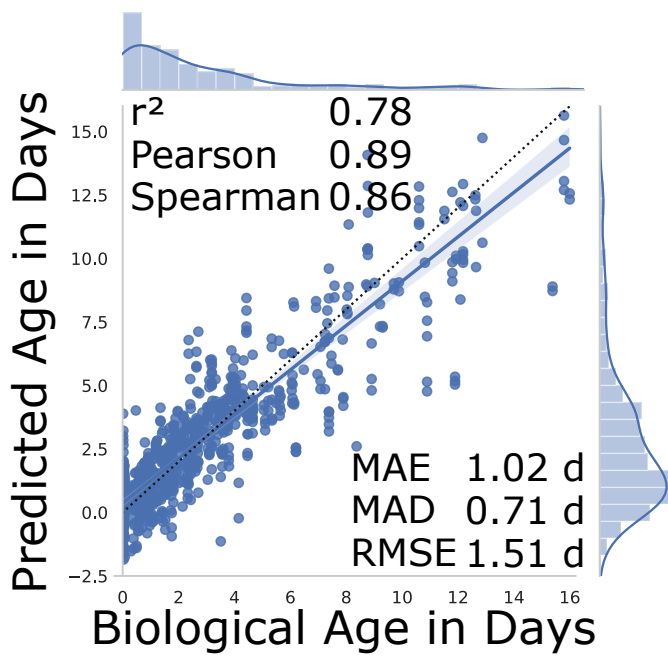
Table S1. Data overview

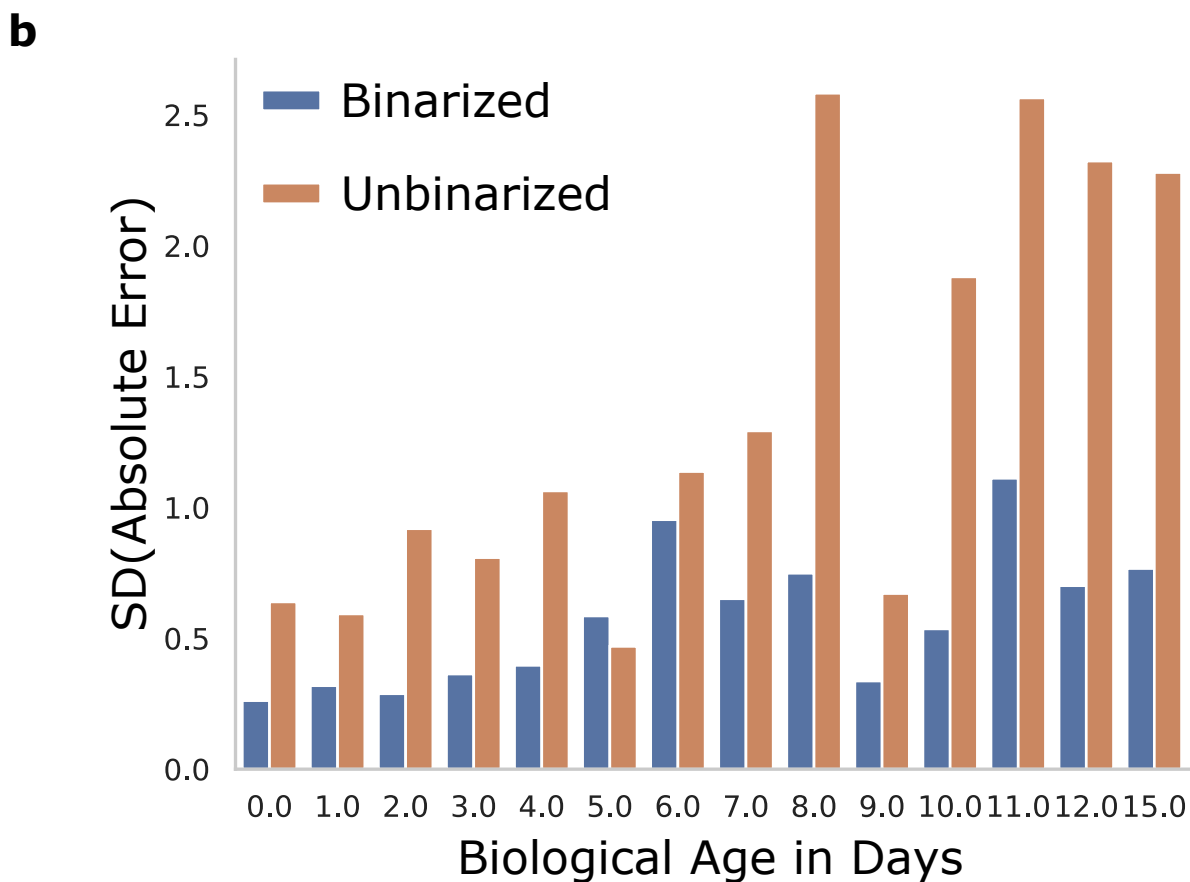
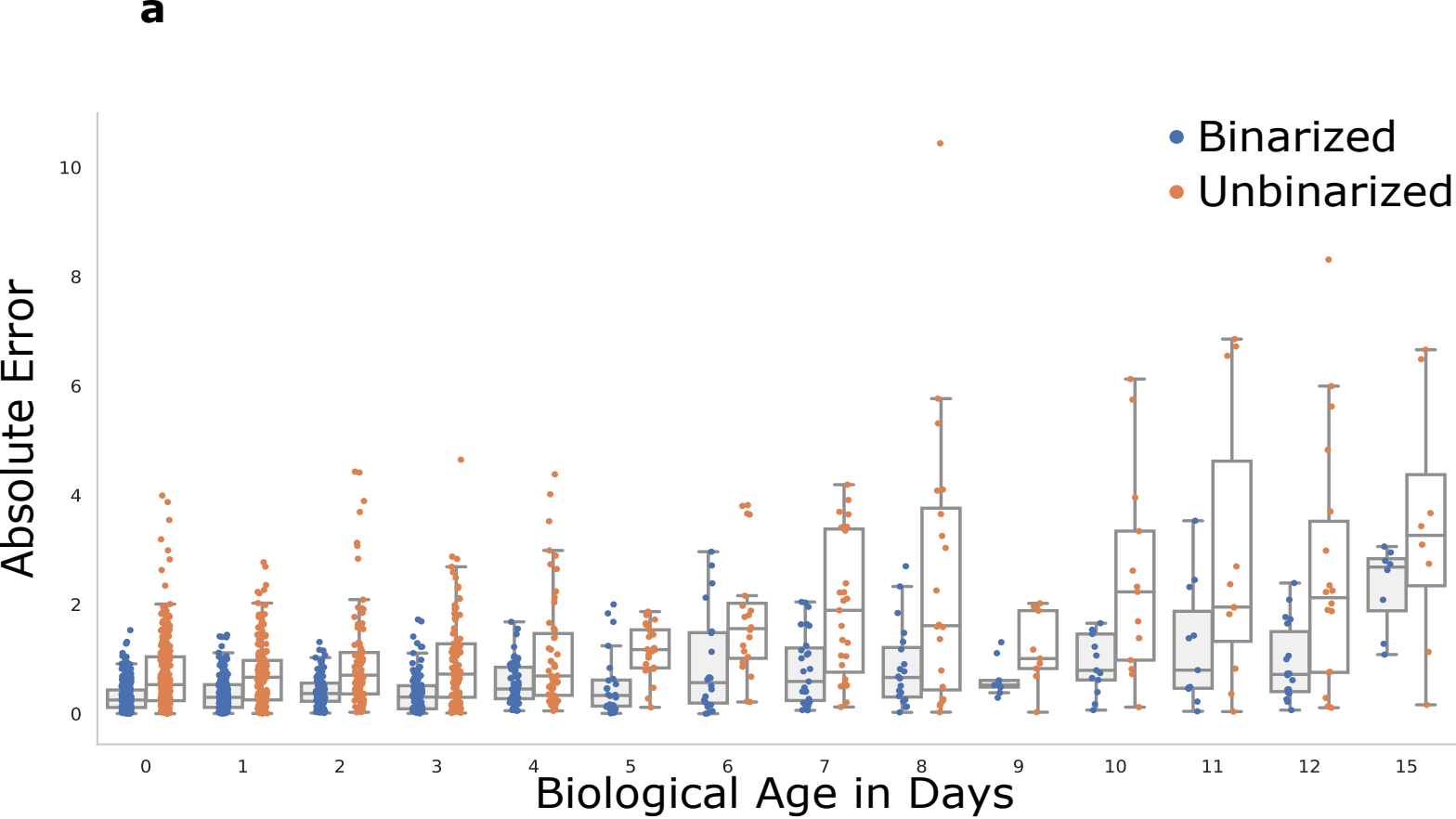
Table S2. *C. elegans* age prediction gene set

Table S3. Statistics

Table S4. Lifespan Prediction

Table S5. Human age prediction gene set





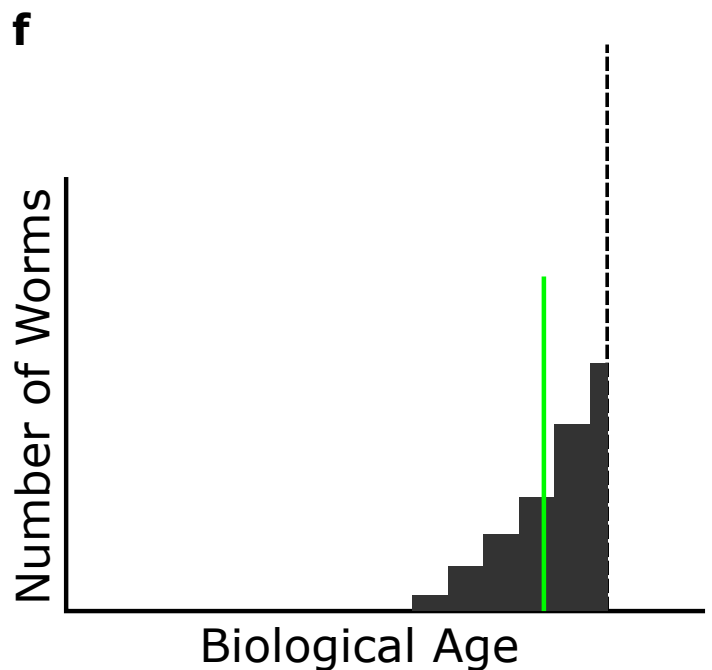
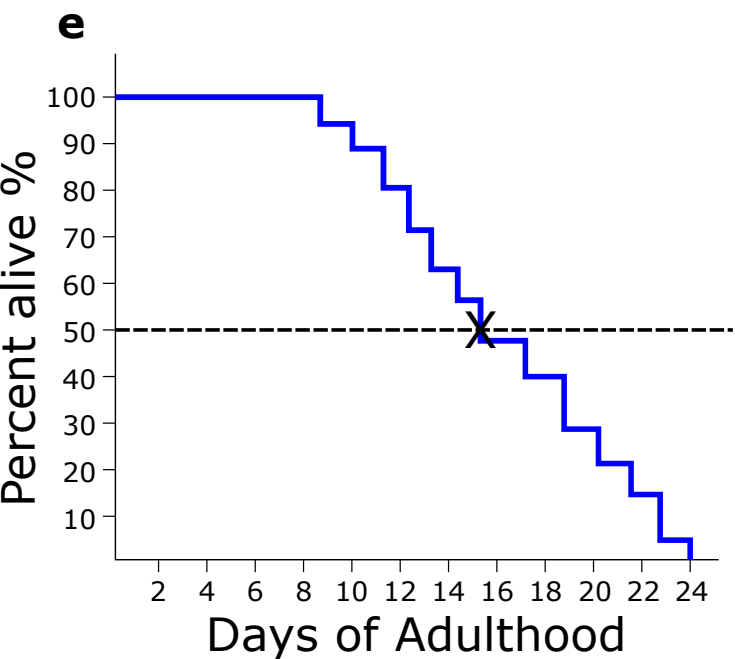
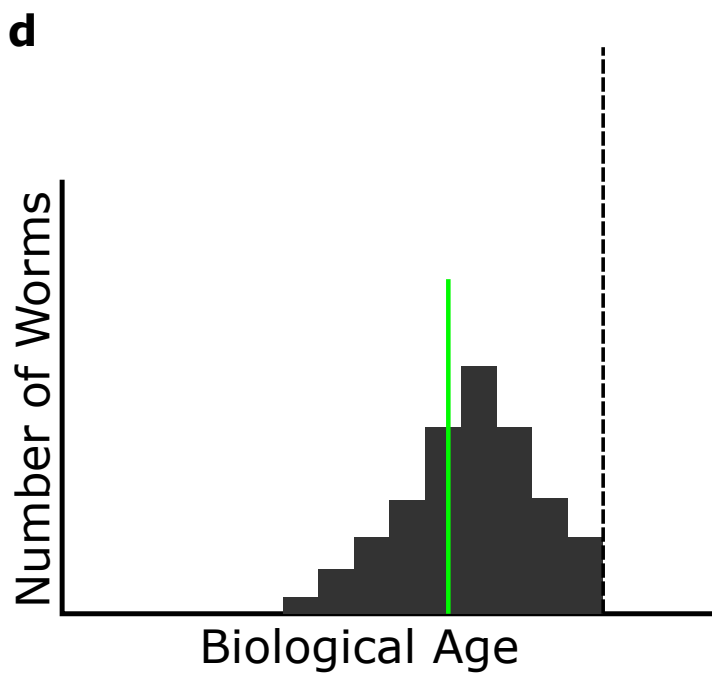
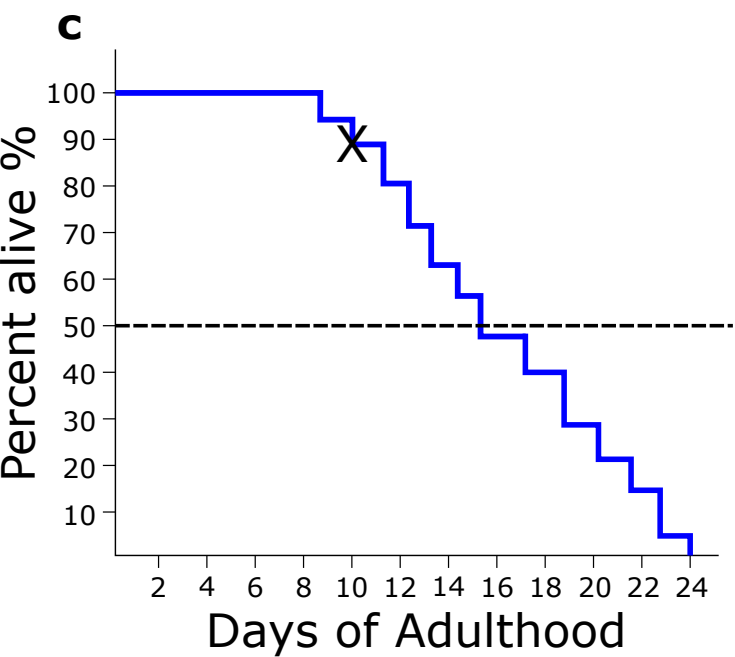
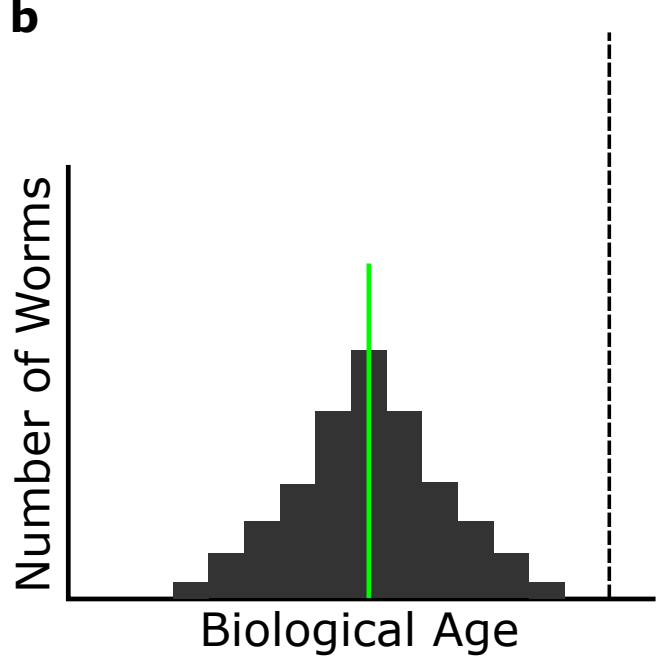
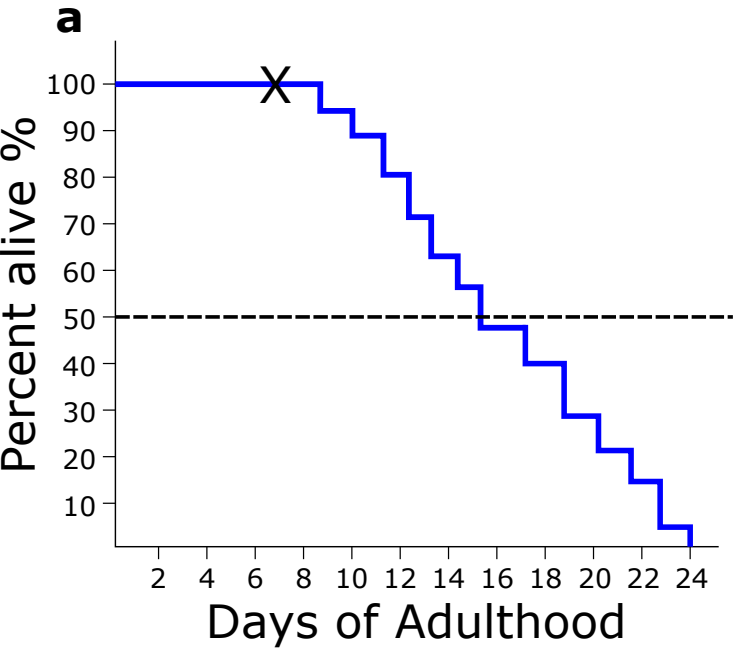
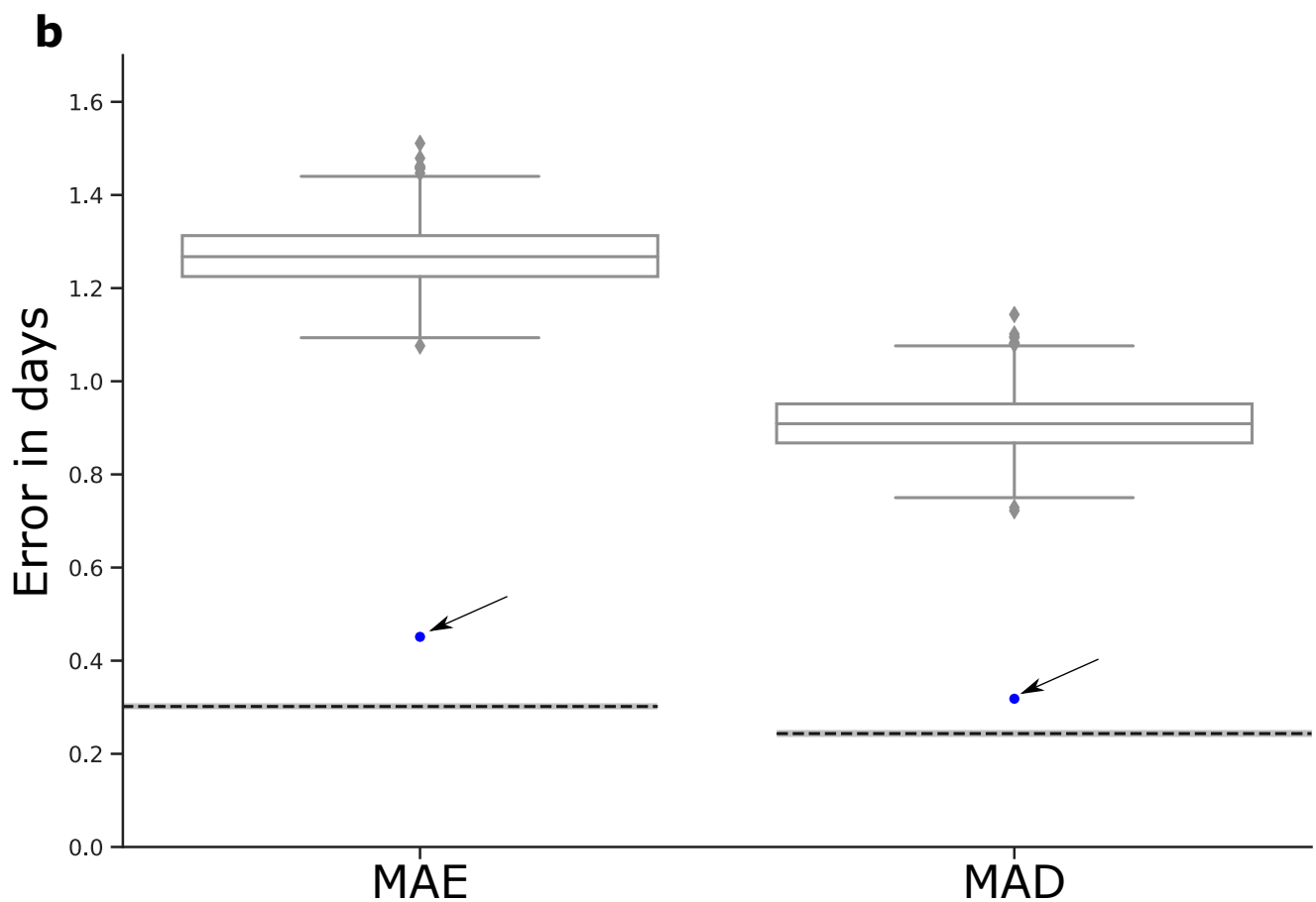
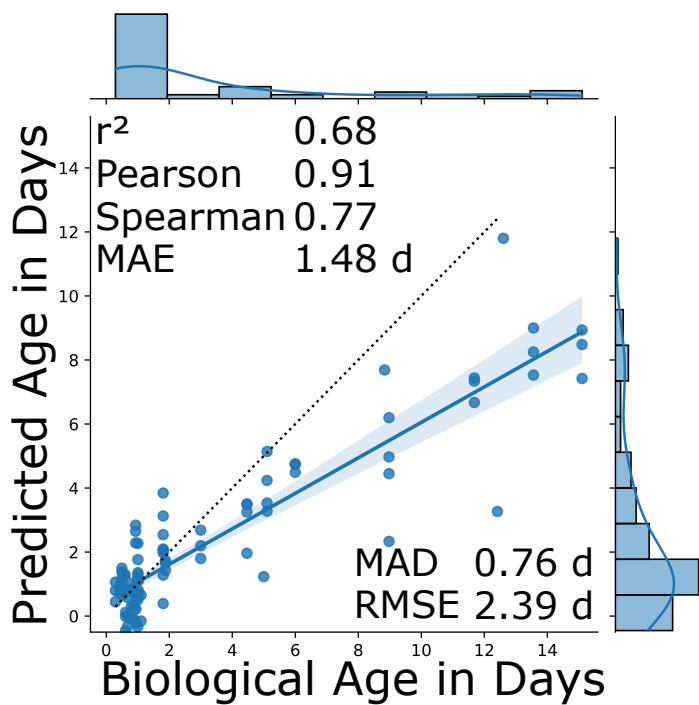


Fig. S3



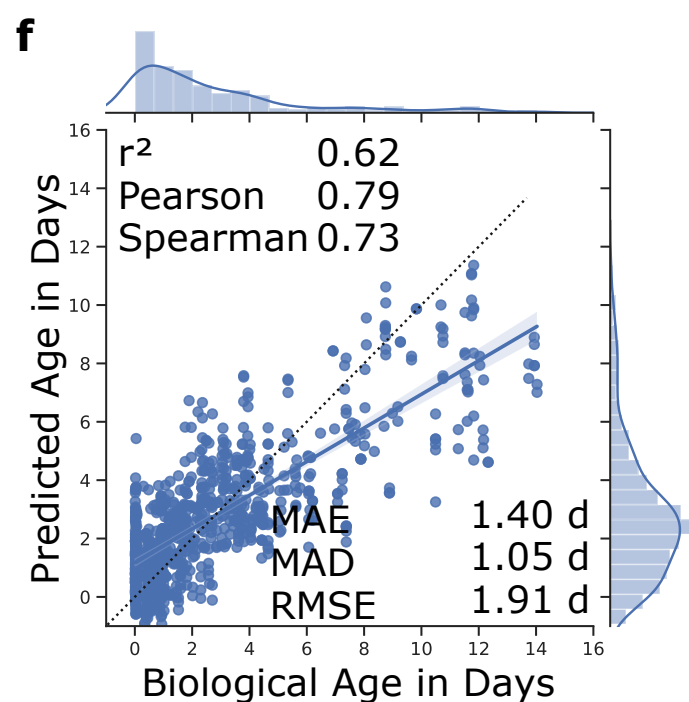
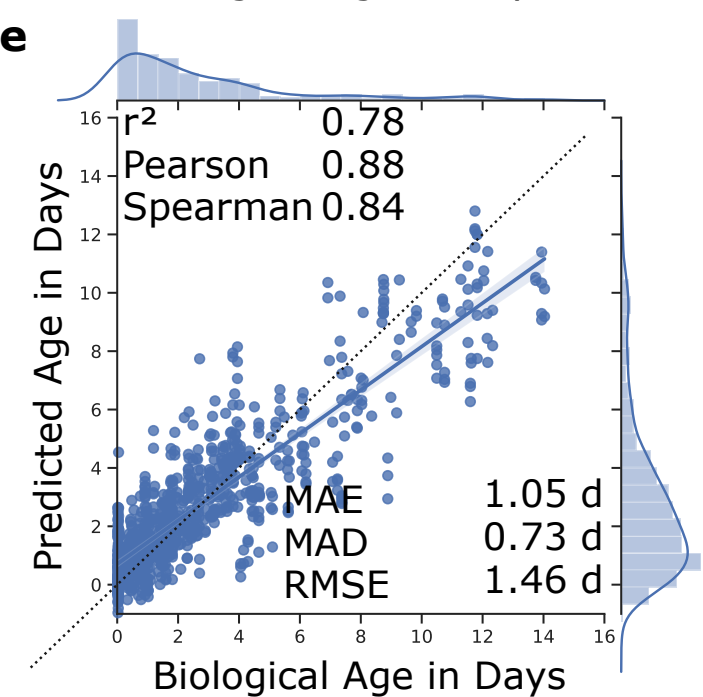
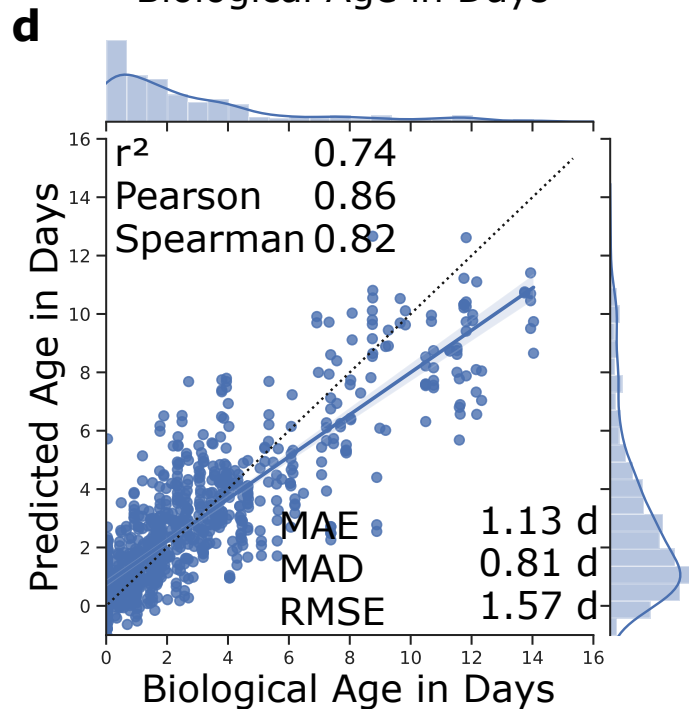
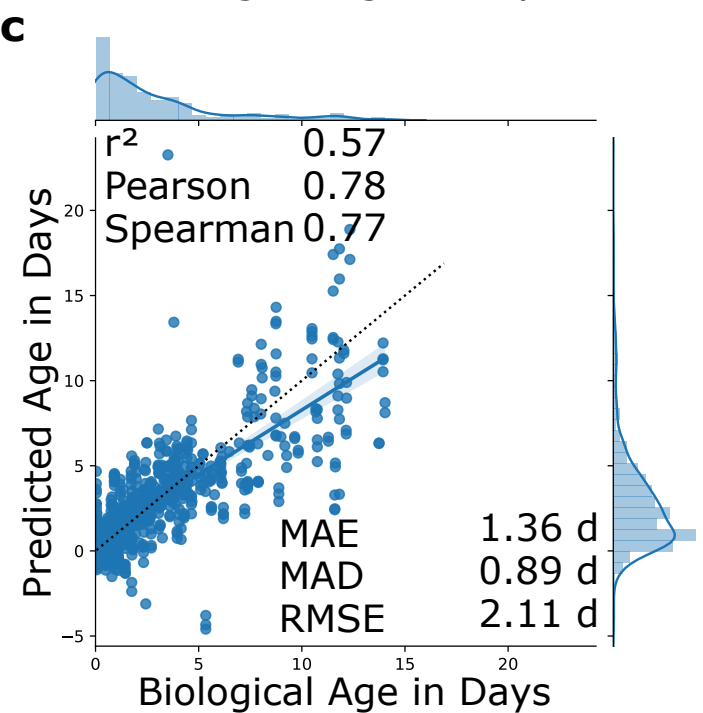
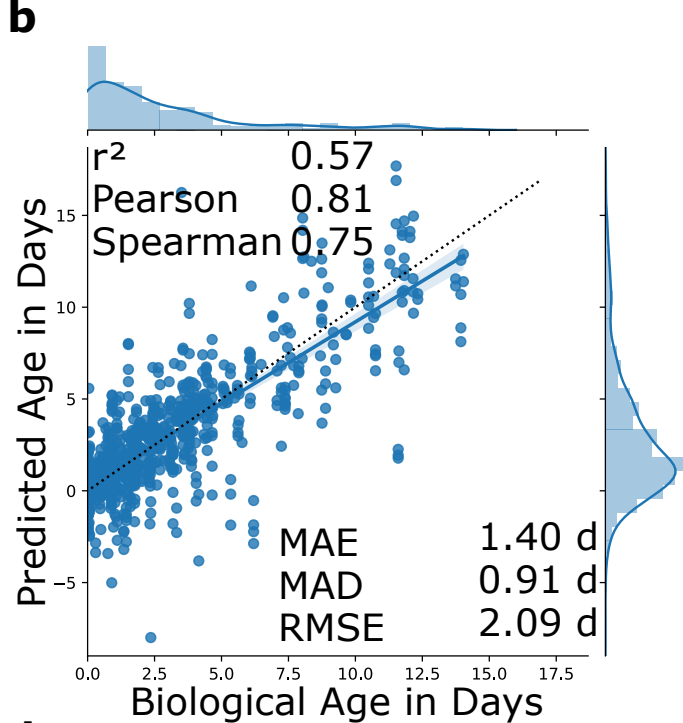
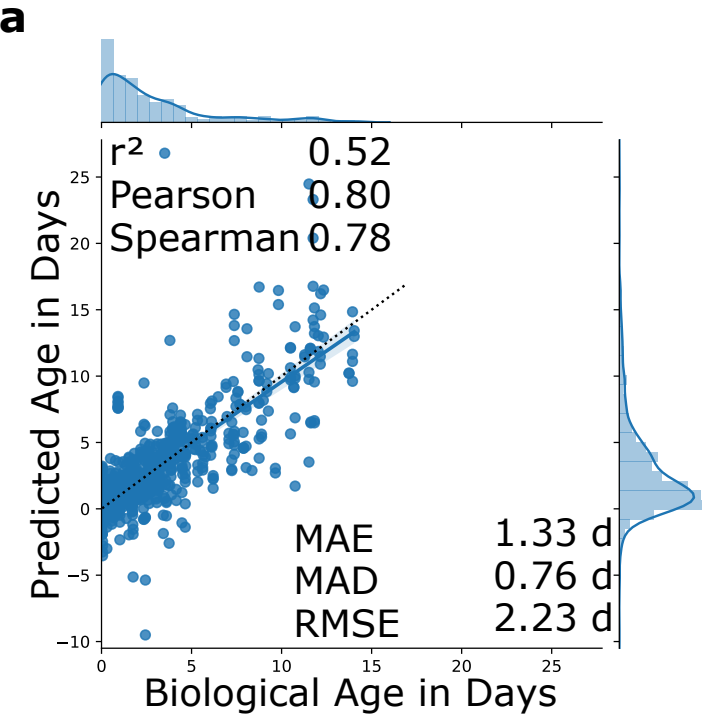


Fig. S5

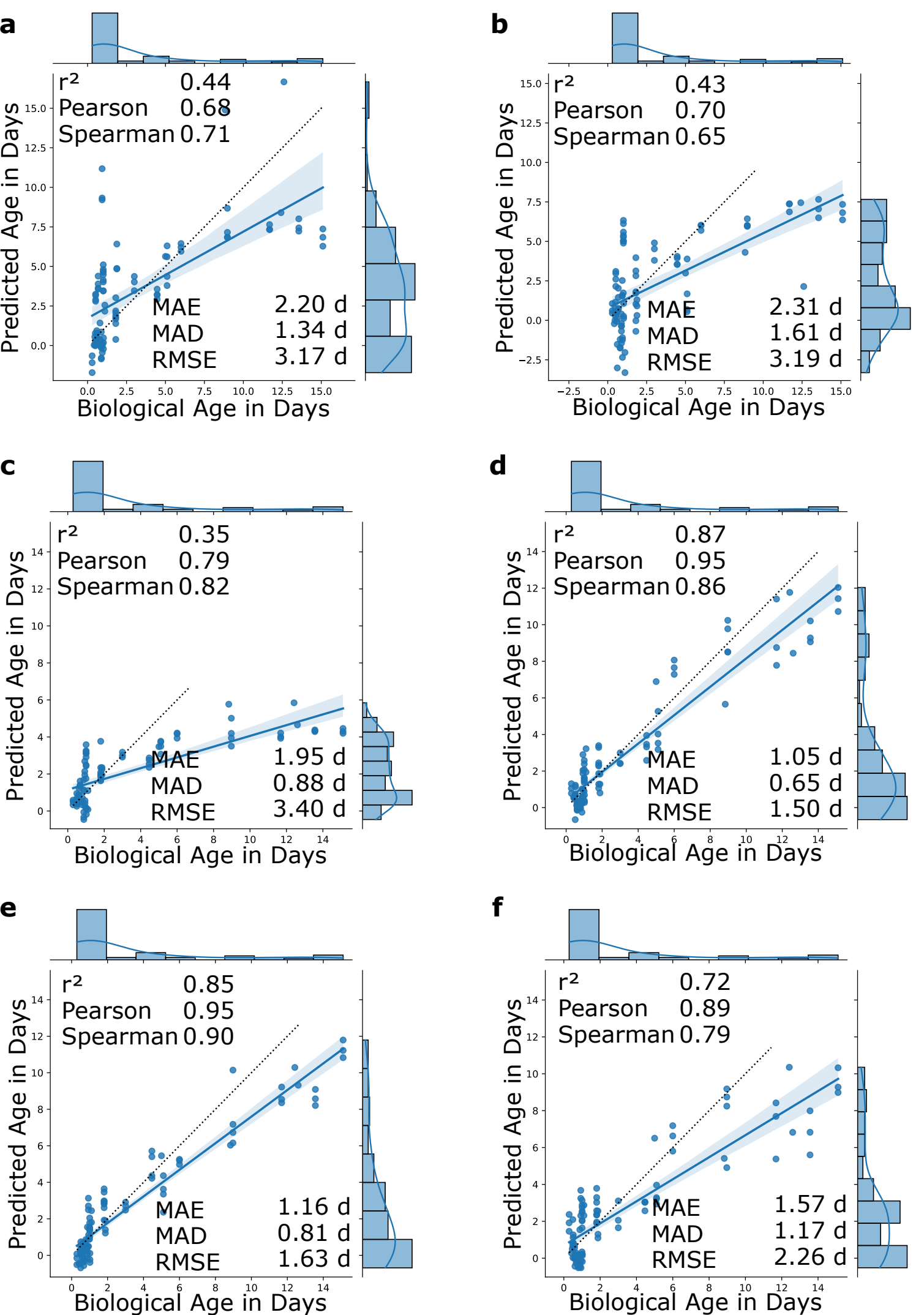
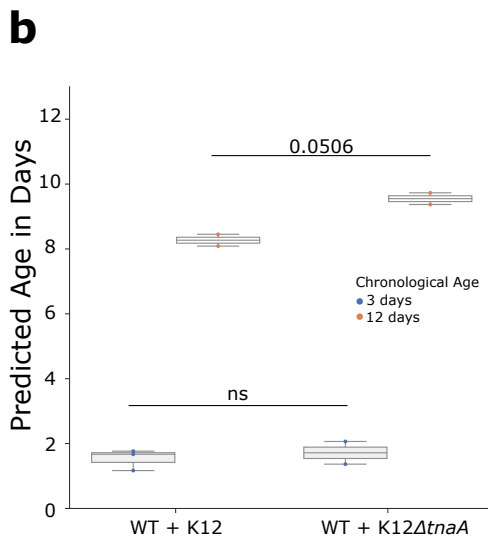
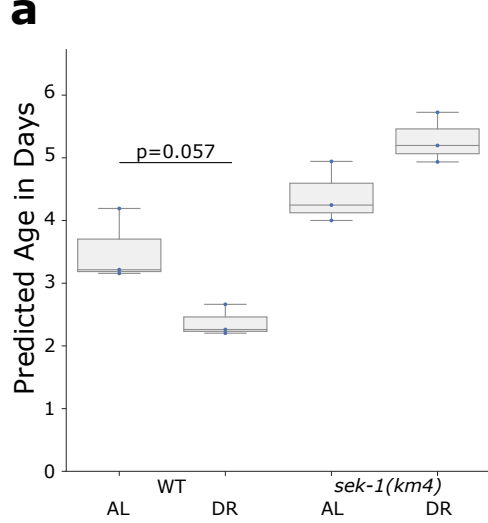
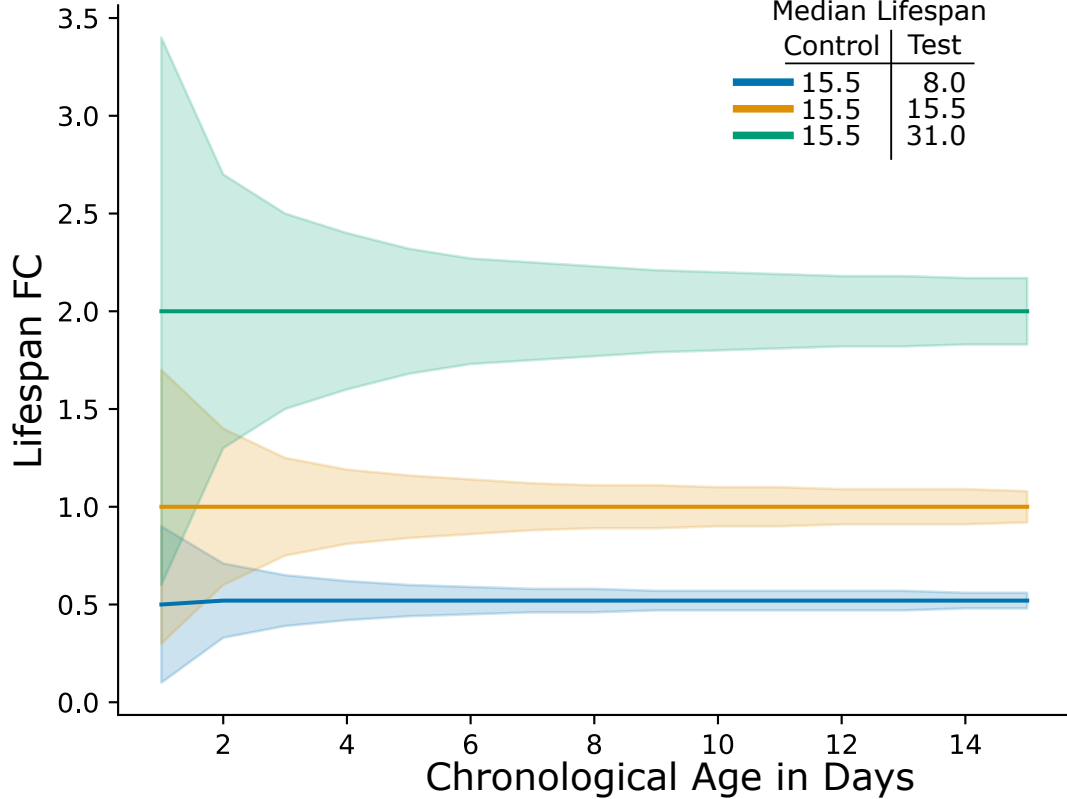
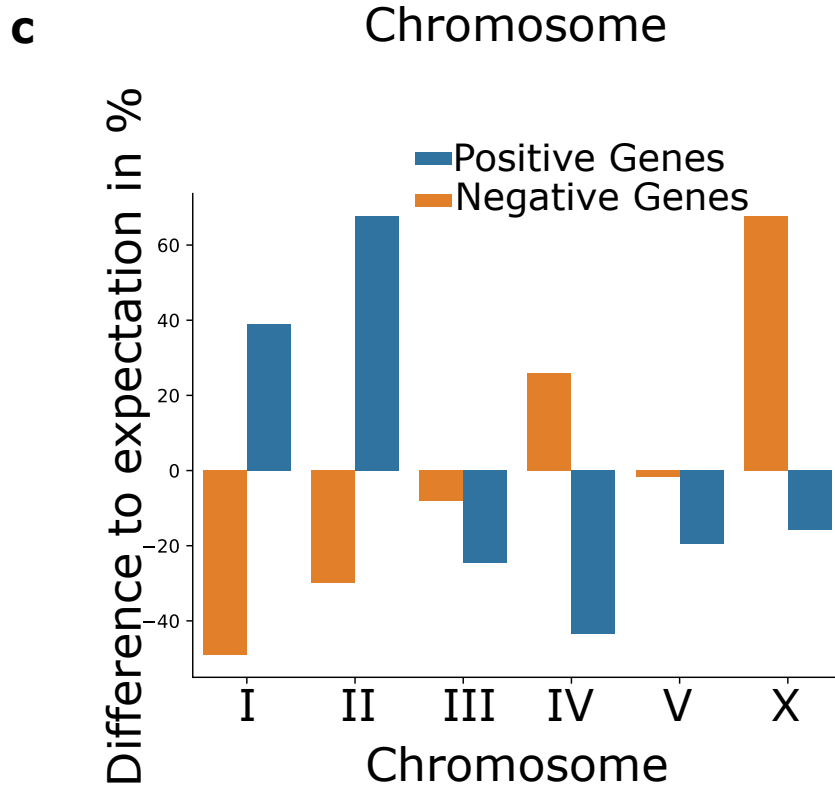
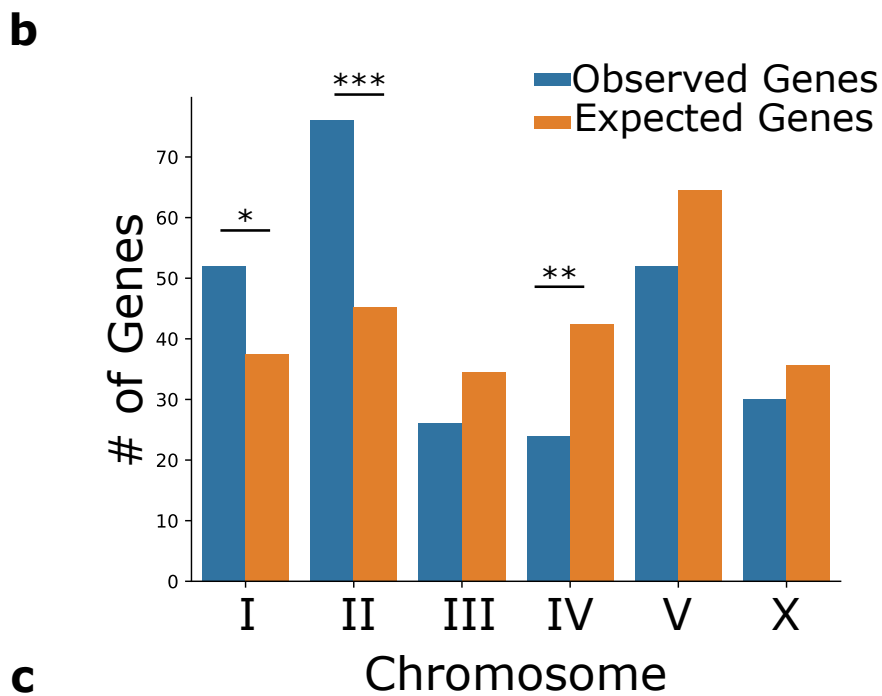
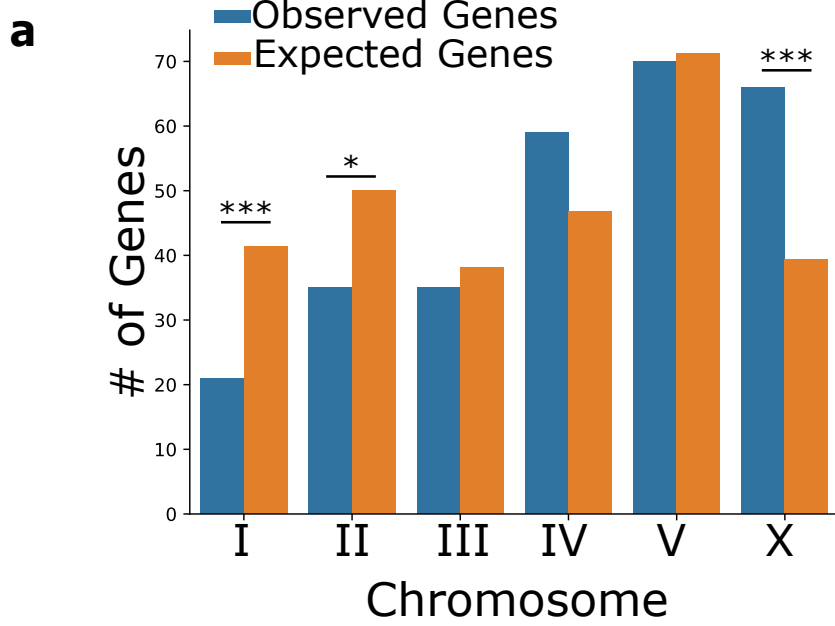




Fig. S6







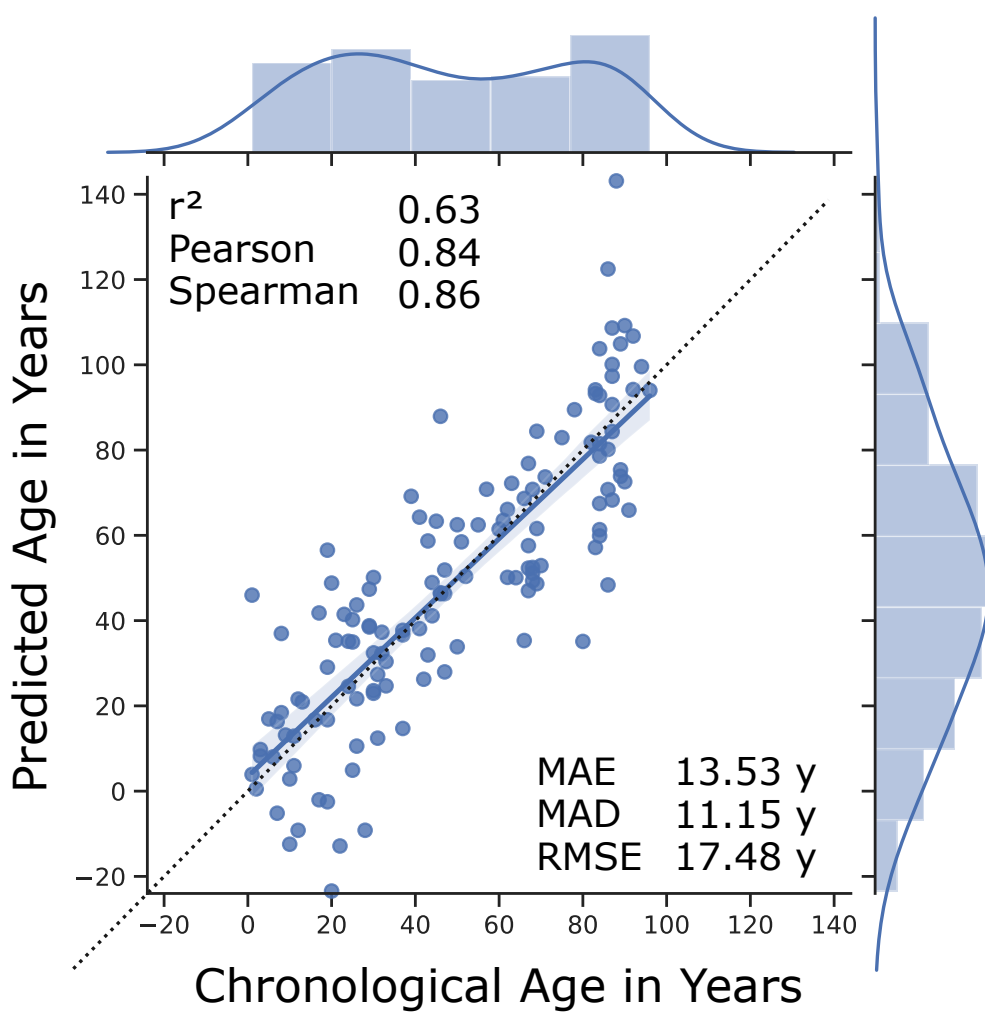
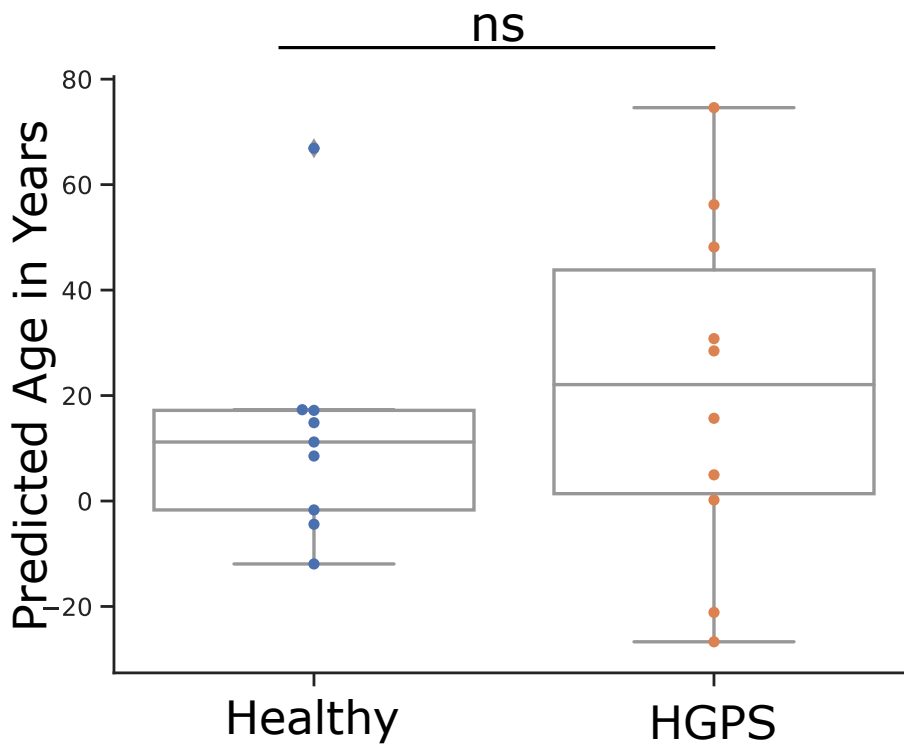


a

Name	Motif	% of Genes	% of BG	FC	p-value	q-value
PQM-1(Gata)		29.93	17.10	1.75	1.09E-07	1.2E-06
ELT-3(Gata)		31.02	20.59	1.51	3.00E-05	1.65E-04

b

Name	Motif	% of Genes	% of BG	FC	p-value	q-value
ELT-3(Gata)		30.04	19.91	1.51	7.74E-05	8.51E-04
PQM-1(Gata)		23.72	16.52	1.44	1.97E-03	0.01

a**b**

3 Aging clocks based on accumulating stochastic variation

David H. Meyer¹, Björn Schumacher¹

¹Correspondence

Published in *Nature Aging* (2024). DOI: 10.1038/s43587-024-00619-x

Author contributions:

- D.H.M did all data analysis.
- D.H.M and B.S. conceived and designed the study, and wrote the manuscript.

1 Aging clocks based on accumulating stochastic variation

2 David H. Meyer^{1,2,*} and Björn Schumacher^{1,2,*}

3 ¹Institute for Genome Stability in Aging and Disease, Medical Faculty, University Hospital and University of
4 Cologne, Joseph-Stelzmann-Str. 26, 50931 Cologne, Germany

5 ²Cologne Excellence Cluster for Cellular Stress Responses in Aging-Associated Diseases (CECAD), Center for
6 Molecular Medicine Cologne (CMMC), University of Cologne, Joseph-Stelzmann-Str. 26, 50931 Cologne, Germany

7 *To whom correspondence should be addressed. E-mail: david.meyer@uni-koeln.de, bjoern.schumacher@uni-
8 koeln.de

9 **Aging clocks have provided one of the most important recent breakthroughs in the biology of aging**
10 **and may provide indicators for the effectiveness of interventions in the aging process and preventive**
11 **treatments of age-related diseases. The reproducibility of accurate aging clocks has reinvigorated**
12 **the debate on whether a programmed process underlies aging. Here, we show that accumulating**
13 **stochastic variation in purely simulated data is sufficient to build aging clocks, and that first and**
14 **second-generation aging clocks are compatible with the accumulation of stochastic variation in DNA**
15 **methylation or transcriptomic data. We find that accumulating stochastic variation is sufficient to**
16 **predict chronological and biological age, indicated by significant prediction differences in smoking,**
17 **calorie restriction, heterochronic parabiosis and partial reprogramming. Moreover, we demonstrate**
18 **that the pan-mammalian clock can be recapitulated through stochastic variation. While our**
19 **simulations may not explicitly rule out a programmed aging process, our results suggest that**
20 **stochastically accumulating changes in any set of data that have a ground state at age zero are**
21 **sufficient for generating aging clocks.**

22 Introduction

23 Weismann's 1881 proposition suggested an aging program to benefit species by freeing up resources
24 from older individuals¹. This hypothesis was later largely rejected²⁻⁵, for a range of reasons such as the
25 circularity of the argument and the assumption of group selection. Evolutionary theories of aging
26 realized the vanishing force of natural selection post-reproductively, notably stated in the disposable
27 soma, the mutation accumulation, and the antagonistic pleiotropy theories of aging^{2,6}. Mutations that
28 abruptly limit post-reproductive life are observed in semelparous species, while iteroparous species
29 typically show a gradual functional decline due to insufficient maintenance and repair mechanisms,
30 leading to stochastic damage accumulation with aging⁷. Progress on aging clocks has revived the idea
31 of a potential aging program⁸, questioning whether aging is primarily a stochastic entropy-driven
32 event, whether aging clocks could show a causal relationship^{9,10}, and whether it involves programmatic
33 aspects¹¹⁻¹⁶. Intrinsic flaws in a software code of life¹⁷, an adaptive pathogen control program^{11,18}, or

34 developmental processes^{13,15} were suggested to cause aging. Age-dependent selective mortality may
35 depend not only on remaining fertility, but also on intergenerational resource transfer, explaining a
36 quantity-quality tradeoff, and potentially allowing a programmed process to affect aging¹⁹.

37 Epigenetic drift, observed during aging, was assigned to imperfect maintenance of epigenetic marks²⁰,
38 reducing methylation differences between genomic regions that are defined during development over
39 time²¹. It has been proposed that age-coupled stochastic methylation changes are highly genome
40 context specific²², and that an information-theoretic view of DNA methylation pattern explains the
41 observed stochasticity in line with context-specific maintenance energy consumption²³. Differential
42 equations showed that CpG methylation sites can be modelled based on maintenance rates, defining
43 CpG site-specific equilibria^{24,25}. Horvath's epigenetic clock was suggested to result from an imperfect
44 epigenetic maintenance system²⁶ and increased DNA methylation entropy was observed in older
45 individuals²⁷. This stochastic epigenetic drift is conserved across species and attenuated upon caloric
46 restriction²⁸. Age-related variably methylated positions are reproducible, not driven by cell-type
47 composition, linked to developmental and DNA damage response genes, enriched at polycomb
48 repressed regions, and associated with expression of polycomb repressive complex 2²⁹. Moreover,
49 ~30% of the mouse genome might be affected by age-related epigenetic disorder, which are enriched
50 in the Petkovich clock³⁰, and a clock using these biological disorder measurements could be built³¹.

51 To deepen the mechanistic understanding of epigenetic aging clocks, CpG sites from 12 clocks were
52 deconstructed into distinct modules some of which might be driven by entropic alterations that regress
53 to a methylation state of 0.5, while most modules change systematically with time³². Recently, it was
54 demonstrated that initializing CpG values at either 0 % or 100 % could accurately predict the simulated
55 age in single-cell simulations, irrespective of stochastic, co-regulated, or a combined simulation.
56 Starting every CpG site at 0 % or 100 %, they could either remain unchanged or regress towards 0.5³³,
57 suggesting that a single stochastic variable could track entropic aging³⁴.

58 Here, we show that datasets that contain accumulating stochastic variation, and are normalized
59 between 0 and 1, can be used to build an age predictor suggesting that any set of biological
60 measurements could be used to build accurate aging clocks. The pace of predicted aging is primarily
61 set by the degree of stochastic variation, where increased stochasticity accelerates while reduced
62 stochastic variation decelerates the predicted age. We validated our findings in transcriptomic datasets
63 of *C. elegans*, and determined that predictions of the transcriptomic aging clock and the amount of
64 added stochastic variation correlate significantly. The predictive results of simulated transcriptomic
65 data with accumulating stochastic variation significantly correlates with the chronological age.
66 Epigenetic aging clocks measure how much stochastic variation accumulated, and the predictive
67 results of a model trained on simulated data with accumulating stochastic variation correlates

68 significantly with the chronological age of human DNA methylation samples. We validated and
69 replicated our results on data from the Mammalian Methylation Consortium³⁵, showing that a variety
70 of mammalian species and interventions can be correctly predicted. We establish that accumulation
71 of stochastic variation is enabling the construction of pan-mammalian clocks, which are capable of
72 detecting biological age deceleration and acceleration¹⁵, and the rejuvenation trajectory over a
73 reprogramming time-course in human cells. Our analyses suggest that aging clocks could be based on
74 any biological parameter with stochastic age-related alterations for precise measurements of aging,
75 without the need for a deterministic process.

76 Results

77 Data-type independent predictions

78 To investigate whether a stochastic process is sufficient to build an age predictor of any dataset, we
79 simulated random data with an age range between 0-100. We used 2000 random data points (features)
80 uniformly distributed between 0 and 1 as the ground state. The ground state is motivated by the
81 proposed ground zero of organismal aging³⁶. Features in prediction models can be any quantifiable
82 data type normalized to values between 0 and 1. To test if accumulating normal-distributed stochastic
83 variation over time enables building an age predictor, we independently added such variation to all
84 features in the ground state 1 to 100 times (**Extended Data Figure 1A**, see methods for details). We
85 simulated 6 sets of samples, applying stochastic variation from once to 100 times, reflecting a potential
86 lifespan range. Note that the range from 1-100 was chosen arbitrarily. Using 3 sets of 100 samples we
87 trained an Elastic net regression that predicts the simulated age, i.e. the number of times stochastic
88 variation was added. To validate the model, we used the 300 independent validation samples, starting
89 with the same ground state but that adding independent stochastic variation from the same
90 distribution (**Extended Data Figure 1B**). Although the stochastic variation application makes the data
91 noisier in each time-step and appears to be countable, no predictor can be built as the validation
92 samples lack any trend in the data (**Extended Data Figure 1C**, Pearson correlation: -0.05). Stochastic
93 variation contains negative and positive values equally likely thus on average canceling out the
94 variation precluding a trend or a prediction. When, however, we used the same approach as above but
95 constrained the values between 0 and 1 after adding the stochastic variation, we observed an almost
96 perfect prediction with a Pearson correlation of the independent validation data of 0.99 (p-value<1e-
97 16, full statistics of all analyses can be found in Source Data) (**Extended Data Figure 1D**). Thus, the
98 model found pattern in the simulated data allowing the prediction of how often stochastic variation
99 was added to the ground state (simulated age) even in independent validation data. Importantly, this
100 will potentially work for any dataset, since our simulated starting point (ground state) consists of

101 uniformly random data between 0 and 1, and the stochastic variation added at each time-step is
102 randomly chosen from a normal distribution, i.e. does not require any regulation or program.

103 To account for the non-normal distribution of values that are bounded by 0 and 1, we transformed the
104 values before adding stochastic variation using the logit transform and transformed the data back via
105 the expit (inverse-logit) transformation (**Figure 1A**). A predictor built on these transformed data
106 replicates the model in Extended Data Figure 1D, further establishing the validity of accumulating
107 stochastic variation in predicting age independent from whether a data transformation was used or
108 not (**Figure 1B**, Pearson correlation: 0.95).

109 The prediction accuracy of the independent validation data was robust to the distribution from which
110 stochastic variation was sampled for the training and validation samples (**Figure 1C, Extended Data
111 Figure 1E**). The logit transformed data require a slightly higher data range from which the stochastic
112 variation is sampled (**Figure 1C**). Even predictions in which the age-related stochastic variation per
113 time-step was smaller than the stochastic variation with which we varied the ground state for each
114 sample ($N(\mu = 0, \sigma^2 = 0.01^2)$), showed high accuracy, e.g. the model trained on stochastic variation
115 sampled from $N(\mu = 0, \sigma^2 = 0.005^2)$ per time-step still had a median R^2 of 0.79 for the prediction of
116 the independent validation data (**Extended Data Figure 1E**). This indicates, that even a small amount
117 of accumulating stochastic variation per time-step is enough for an accurate prediction.

118 During training, Elastic net regression assigns a coefficient to each of the 2000 features that then can
119 be used to predict novel independent samples. The Elastic net regression coefficients for the 2000
120 features in our simulation in Figure 1B and Extended Data Figure 1D are reproducible in between
121 independent runs with the same ground state (**Figure 1D, Extended Data Figure 1F**), indicating that
122 even random stochastic variation pattern allow for robust predictions. The prediction is possible due
123 to a regression to the mean, which is to be expected from a stochastic process with a data range limit
124 (**Figure 1E, Extended Data Figure 1G**). Features starting close to 0 tend to increase after stochastic
125 variation addition resulting in a positive Elastic net coefficient, while features close to 1 tend to
126 decrease resulting in a negative coefficient. Features starting around 0.5 in the ground state are more
127 noise sensitive since the added stochastic variation is equally likely to move in either direction leading
128 on average to a cancellation of noise (**Figure 1E, Extended Data Figure 1G**).

129 The prediction accuracy of the amount of normal-distributed stochastic variation plateaus after ~2000
130 features at an R^2 value around 0.97, showing that even models with a limited number of features are
131 highly accurate (**Figure 1F, Extended Data Figure 1H**). Of note, Elastic net regression shrinks
132 coefficients of some features to 0 and thereby further reduces the number of features. These results
133 show that reproducible predictions are possible with less than 2000 features, i.e. much less than is
134 usually available in biological datasets involving any omics approaches, as long as there is accumulating

135 stochastic variation and the data can be normalized between 0 and 1, i.e. predictions are not limited
136 to DNA methylation or transcriptomic data.

137 We next wondered how a model trained on stochastic variation sampled from $N(\mu = 0, \sigma^2 = 0.2^2)$
138 would predict samples with different stochastic variation distributions. Choosing a standard deviation
139 twice as large ($\sigma=0.4$), also doubles the interval from which $\sim 99.7\%$ of stochastic variation values are
140 sampled, which increases the amount of stochastic variation added in each time step. Testing the
141 model on data simulated with more stochastic variation per time step resulted in a faster increase and
142 plateau of the prediction, while a reduced stochastic variation level decreased the slope of the
143 prediction (**Figure 1G, Extended Data Figure 1I**). Samples with more stochastic variation per time step
144 reach their maximum simulated age earlier. This analysis suggests that an increase in stochastic
145 variation accelerates, while a decrease in stochastic variation decelerates the predicted aging process.

146 Transcriptomic biological age prediction

147 We next wondered whether an age predictor based on gene expression data applied to data with
148 accumulation of stochastic variation would show a comparable correlation result. We have recently
149 developed a highly accurate biological age predictor of *C. elegans* with the Binarized Transcriptome
150 Aging (BitAge) clock³⁷. We defined the ground state as the biologically youngest adult RNA-seq sample
151 (GSM2916344³⁸) in our dataset and simulated stochastic variation similarly as explained in Extended
152 Data Figure 1A, i.e. with (non-empirically estimated) normal distributed variation. In accordance with
153 our results in Figure 1B and Extended Data Figure 1D, BitAge predictions as well correlate linearly with
154 the amount of stochastic variation in the data (**Figure 2A**, Pearson correlation: 0.81). The correlation
155 is robust to the amount of stochastic variation added in each time-step, with a peak in Pearson
156 correlation of 0.81 at stochastic variation sampled from a normal distribution with a standard deviation
157 of 0.01 (**Extended Data Figure 2A**). This indicates that the predicted transcriptomic age of *C. elegans*
158 correlates with age-dependent stochastic variation in the data.

159 Next, we wondered whether a stochastic data-based clock could predict the biological age of biological
160 samples. The stochastic data-based clock predictions significantly correlated (Pearson correlation:
161 0.72) with the biological age of 993 independent *C. elegans* RNA-seq samples from 61 independent
162 public datasets for which the biological age could be calculated (**Figure 2B, Supplement Table 1**, see
163 methods for details). This prediction is robust to the number of features, i.e. genes, used in the
164 simulation (**Extended Data Figure 2B**). A permutation of the biological age does not correlate with the
165 predicted simulated age (**Extended Data Figure 2C**).

166 To test whether a stochastic age predictor could identify age acceleration and deceleration across a
167 wide spectrum of aging interventions, we divided the 993 transcriptome samples into long-lived (>20d

168 median lifespan), normal-lived, and short-lived (<8d median lifespan). Plotting the predictions against
169 the chronological age shows small but significant differences. A multivariate linear regression with the
170 chronological age, the median lifespan, and its interaction term, shows a significant median lifespan
171 effect with a negative coefficient, i.e. a longer lifespan leads to a lower prediction based on the
172 stochastic data-based clock ($p=0.015$) (**Figure 2C**). This indicates that accumulating stochastic variation
173 scales mostly with the chronological age, but also shows a significant lifespan effect, i.e. biological age
174 prediction. A lifespan extending treatment that was shown to reduce transcriptional drift (a measure
175 of transcriptomic variance) is the anticonvulsant Mianserin³⁹. Consistent with limiting gene expression
176 variation, we found that Mianserin dose-dependently decreases the predicted age with the stochastic
177 data-based clock in independent data (**Figure 2D**, one-way ANOVA p-value: 0.006, post hoc Tukey test
178 50 μ M Mianserin vs. Control p-value: 0.03). 50 μ M Mianserin shows a (non-significant) lower slope as
179 well as generally lower predicted values over a time-course (p -value=7.3e-04) compared to control
180 samples (**Figure 2E**). These results indicate that the stochastic transcriptomic data-based clock
181 predictions of *C. elegans* can predict the chronological age and the biological age deceleration of a
182 pharmacological intervention affecting transcription drift.

183 Single-cell DNA methylation simulations

184 The most well-established aging clocks in mammals including humans are based on age-related
185 changes of epigenetic CpG sites. We assessed whether simulations based on accumulating stochastic
186 variation might be applicable to epigenetic data. Adding normally distributed stochastic variation once
187 in the simulation in Figure 1 did not change the simulated sample much from the ground state
188 (**Extended Data Figure 3A**), while adding stochastic variation 100 times lead to a uniform distribution
189 of features (**Extended Data Figure 3B**). However, CpG methylation sites are typically under higher
190 maintenance and less noisy. Comparing biological DNA methylation data of young and old subjects
191 shows that the methylation sites, starting close to the extremes (0 or 1) show indeed less variance
192 (**Extended Data Figure 3C**).

193 We next simulated instead of bulk data between 0 and 1, “single-cell” data for which each feature is
194 binary, i.e. either methylated (1) or unmethylated (0) (**Figure 3A**). Note that this is a simplification for
195 diploid organisms, however, this should not affect the results, as in theory the different alleles could
196 be represented as different features in the simulations. It has been shown that bulk methylation
197 pattern at single CpG sites can be modelled with differential equations containing a methylation
198 maintenance efficiency (E_m) (the probability that a methylated site stays methylated), and a *de novo*
199 methylation efficiency (E_d) (the probability that an unmethylated site gets methylated; $1 - E_d$ is the
200 maintenance efficiency of the unmethylated state (E_u))²⁴. These maintenance efficiencies describe the
201 rate by which a CpG site does not alter per time-step. We simulated single-cell DNA methylation

202 changes in a stochastic system over time as depicted in Figure 3A using a variety of maintenance
203 efficiencies, i.e. site-specific efficiencies that are either estimated from data, randomly chosen, or
204 universal efficiencies that are fixed to one value for all CpG sites.

205 First, we tested how a universal maintenance efficiency rate, i.e. the same rate for all 500 features
206 would affect the accuracy of the model (**Figure 3B**). A high maintenance ($E_m=99.9\%$, $E_d=0.01\%$, i.e.
207 $E_u=99.9\%$) yielded almost perfect simulated age predictions ($R^2=0.999$) on the independent validation
208 data (**Figure 3B, C**). A simulated age of 100 shows minimal deviation from the ground state,
209 demonstrating high accuracy with small effect sizes (**Extended Data Figure 3D**). Even maintenance
210 rates of up to 99.995% resulted in a prediction with an R^2 of 0.78 (**Figure 3B**). The predictor is robust
211 in the number of features allowing for highly accurate age predictions with small feature sizes, whose
212 accuracy cap after around 32 features (**Figure 3D**). Training the model on $E_m=99.9\%$ and testing it on
213 data simulated with lower, respectively higher E_m , showed that less maintenance accelerates, while
214 higher maintenance decelerates the aging clock (**Figure 3E**). These results indicate that even a high
215 maintenance rate yields accurate age predictions, and that an increased maintenance decelerates,
216 while a decrease in maintenance accelerates the predicted age.

217 A maintenance rate of 99.9% for methylated as well as unmethylated sites leads to a regression to the
218 equilibrium (0.5). Starting the simulation at the equilibrium and $E_m=99.9\%$ did not allow for a
219 prediction of the simulated age, since no regression to the equilibrium state is possible (**Extended Data**
220 **Figure 3E**, Pearson correlation: 0.05). However, a slight deviation to 0.51 for all starting values in the
221 ground state led to an accurate simulated age prediction via a regression to the equilibrium state
222 (**Extended Data Figure 3F**, Pearson correlation: 0.95).

223 Similar to the universal maintenance model (**Figure 3B-D**), accurate simulated age predictions are
224 possible if E_m and E_d are empirically estimated from data (see Methods) (**Figure 3F**, Pearson
225 correlation: 0.81). The predictions cap off earlier than in Figure 3C due to lower maintenance rates,
226 leading to a quicker convergence to the site-specific equilibria (see also **Extended Data Figure 3E**).

227 Site-specific E_m and E_d values allow accurate simulated age prediction even when starting at 0.5
228 (**Extended Data Figure 3G**, Pearson correlation: 0.99). Such a site-specific regression away from the
229 mean is still in line with stochasticity and entropic alterations. While the site-specific maintenance
230 rates give a framework in which each feature will change, the change itself is purely stochastic.
231 Stochastic variation after 100 times-steps shows less variation in features starting close to 0 or 1 than
232 those starting close to 0.5 (**Extended Data Figure 3H**), resembling the comparison of young and old
233 human DNA methylation datasets (**Extended Data Figure 3C**). Without site-specific stochastic variation
234 predictions were driven by the regression to the mean (**Figure 1E**, **Extended Data Figure 1G**), while
235 site-specific stochastic variation showed no correlation (**Extended Data Figure 3I**), suggesting a

236 regression away from the mean could be explained via a stochastic process, arguing against a recent
237 report that suggested clock sites starting around 0.5 couldn't be entropic³².

238 In conclusion, accurate age predictors can be built by simulating DNA methylation changes purely with
239 stochastic variation based on the maintenance efficiency rates of methylated and unmethylated sites.
240 In addition, DNA methylation sites can have equilibria unequal to 0.5, allowing for a stochastic
241 regression away from the mean, and even sites close to the site-specific equilibria can confer
242 information for the aging clock.

243 Public aging clocks

244 Next, we were wondering whether published DNA methylation aging clocks might also mainly measure
245 stochastic variation. Horvath's pan-tissue DNA methylation clock²⁶ predicts a linear increase of the
246 amount of stochastic variation generated based on empirically estimated E_m and E_d values until it
247 caps off at an predicted age around ~60 years (**Extended Data Figure 4A**, Pearson correlation: 0.91).
248 The time-steps in our simulations are arbitrary and not directly comparable to the predicted age, since
249 our simulated age tracks how often we added stochastic variation, and the predicted age is epigenetic
250 age in years. We wondered whether we could estimate the range-limits of the site-specific E_m and E_d
251 such that the epigenetic age prediction of our simulated data would be as accurate as possible
252 regarding the simulated age. We tested multiple combinations of limits for E_m and E_d and calculated
253 the R^2 as a measure of accuracy between the predicted and the simulated age (**Figure 4A**). Horvath's
254 epigenetic clock has the highest accuracy in predicting the simulated age with the limits $97\% < E_m \leq$
255 100% and $0\% \leq E_d < 5\%$, suggesting higher site-specific maintenance with a narrower range for E_m
256 and E_d than previously assumed (**Figure 4A**). Indeed, the prediction with Horvath's epigenetic clock
257 caps-off later with these new limits (**Figure 4B**, Pearson correlation: 0.91, compare **Extended Data**
258 **Figure 4A**). These results suggest that the site-specific maintenance rates are sufficient to explain the
259 predictability of Horvath's aging clock.

260 Randomly choosing E_m and E_d within the limits $97\% < E_m \leq 100\%$ and $0\% \leq E_d < 5\%$ allowed
261 simulations with highly significant Pearson correlations as well (Median Pearson correlation: 0.89,
262 **Extended Data Figure 4B**). The same is even true if instead of site-specific maintenance rates all CpG
263 sites were simulated with a universal maintenance efficiency of 99 % that was not inferred from a
264 biological sample and could therefore not be confounded (**Figure 4C**, Pearson correlation: 0.97). The
265 Pearson correlations are robust to the universal methylation maintenance efficiency, but peak at 99%
266 (**Extended Data Figure 4C**). A low maintenance efficiency of 90 % reduces the Pearson correlation
267 (**Extended Data Figure 4C**) since the features reach the equilibrium faster and therefore cap off quicker
268 (compare **Figure 3B**). A high maintenance efficiency of 99.95 % reduces the Pearson correlation due to

269 the reduced speed of convergence (**Extended Data Figure 4C**). Notably, Horvath's clock predicts an old
270 age of 69.4 years for a dataset with DNA methylation levels of 0.5 for all CpG sites. These results suggest
271 that no biologically inferred maintenance rate is required but instead indicates that stochastic variation
272 is sufficient for age prediction.

273 Next, we tested the second generation aging clock PhenoAge⁴⁰ (**Figure 4D-F, Extended Data Figure 4D-**
274 **F**). The previously assumed limits for E_m and E_d led to a similar linear increase, and early cap-off of
275 the predicted PhenoAge (**Extended Data Figure 4D**, Pearson correlation: 0.89). Improved limits (**Figure**
276 **4D,E**), coincide with those estimated for Horvath's clock. PhenoAge significantly correlates with the
277 simulated age of samples simulated with random E_m and E_d within the limits (Median Pearson
278 correlation: 0.84, **Extended Data Figure 4E**), or a universal maintenance efficiency of 99% (**Figure 4F**,
279 Pearson correlation: 0.94), which as well was robust to the maintenance efficiency chosen (**Extended**
280 **Data Figure 4F**).

281 We next tested how ground states defined at different ages might affect the age simulations. Starting
282 the ground state with a sample from a 16-year-old and simulating the addition of up to 100 stochastic
283 variations results in the linear increase in predicted age (**Extended Data Figure 4G**, Pearson correlation:
284 0.89). Starting from a 37-year-old, starts the prediction higher, shows a smaller linear increase in the
285 predicted age, and leads to a quicker arrival and longer time at the cap (**Extended Data Figure 4H**).
286 Starting from an 81-year-old, does not show a difference in the prediction upon stochastic variation,
287 indicating that the ground state is already containing as much stochastic variation as we would expect
288 at the cap-off (**Extended Data Figure 4I**, Pearson correlation: 0.09). These results affirm that our
289 simulations are robust to the choice of the ground state and that the predictions are scaled accordingly.

290 All tested first generation aging clocks⁴¹⁻⁴³ and the second generation aging clock GrimAge⁴⁴,
291 significantly correlated with the simulated age irrespective of whether empirically estimated, random,
292 or universal maintenance rates were assumed (**Extended Data Figure 5A-H**).

293 Employing the Gillespie algorithm⁴⁵ for event-based simulations, where time-steps are not uniform but
294 the time until the next event is calculated, recapitulates our results (**Extended Data Figure 5I**, Pearson
295 correlation: 0.98), indicating that our simulations are robust to the method used.

296 Stochastic data-based aging clock

297 We next aimed to address whether a clock built on simulated DNA methylation data (see Methods)
298 could predict the chronological age of mammalian biological samples. A simulated training dataset
299 with the CpG sites from Horvath's epigenetic clock led to a significant Pearson correlation of 0.87 (p-
300 value <1e-16) of chronological age and the predicted simulated age (**Extended Data Figure 6A**). This
301 linear correlation holds for randomly chosen CpG sites, and is robust across different feature sizes

302 **(Extended Data Figure 6B)**, while randomly permuting the chronological age of samples leads to non-
303 significant correlations **(Extended Data Figure 6C)**.

304 To exclude any potentially confounding effects of cell-type heterogeneity⁴⁶, we estimated the cell-
305 type composition to subsequently correct the biological samples to obtain cell-type heterogeneity-
306 adjusted CpG beta-values. Using cell-type corrected data did not affect the performance of the
307 stochastic data-based clock **(Figure 5A)**, Pearson correlation 0.87, $p < 1e-16$), and an additional cell-type
308 correction of the simulated samples still showed a Pearson correlation of 0.81 ($p < 1e-16$) indicating
309 highly correlated predictions of the biological samples **(Extended Data Figure 6D)**. Additionally, we
310 used a multivariate linear regression of the form

311 $Age \sim PredictedAge + CellTypeFractions$.

312 This multivariate linear regression approach also showed a significant Predicted Age variable ($p < 1e-16$,
313 Source Data) for the predictions of the stochastic data-based clock. These results indicate that cell-
314 type heterogeneity does not have a major role in the predictive power of stochastic variation
315 accumulation.

316 We further probed for potential confounding effects by expanding the analysis to 11,146 independent
317 whole blood or peripheral blood leukocyte samples from 15 different datasets. Stochastic data-based
318 prediction of those samples still resulted in a Pearson correlation of 0.57 ($p < 1e-16$) **(Extended Data**
319 **Figure 6E)**.

320 When instead of an adolescent ground state, we initiated the stochastic data-based clock with a fetal
321 sample the Pearson correlation improved to 0.72 **(Figure 5B)**, with 9 out of 15 datasets reaching
322 correlations ≥ 0.8 **(Extended Data Figure 7)**. By comparison, Horvath's original clock predicts the same
323 samples with a Pearson correlation of 0.85, and 10 out of 15 datasets with a correlation ≥ 0.8
324 **(Extended Data Figure 8)**.

325 In conclusion, our analysis shows that simulating epigenetic stochastic data starting from one young
326 biological sample with site-specific maintenance rates, allows significantly correlated predictions with
327 the chronological age of independent biological samples.

328 Biological age prediction

329 Recently, a pan-mammalian clock suggested that instead of stochastic damage accumulation, aging
330 might be a consequence of a developmental process as the clock sites were associated with genes
331 implicated in developmental gene regulation¹⁵. To assess whether stochastic variation accumulation
332 might also allow a prediction of the biological age, we next investigated the predictive power of a
333 stochastic data-based clock on the data from the Mammalian Methylation Consortium^{15,35,47}.

334 We used 4 stochastic clocks starting from the youngest blood sample from *Tursiops truncatus* with
335 different maintenance rates (see Methods). All 4 clocks are on average highly significantly correlated
336 with independent data, even from different species (**Figure 5C, Extended Data Figure 9A**),
337 demonstrating that even one biological sample alone with simulated stochastic variation accumulation
338 is sufficient to build aging clocks that are strongly correlated with the relative age of a variety of
339 mammalian species.

340 Lu et al. further validated their clock on interventions that are known to slow biological age¹⁵. Applying
341 our stochastic data-based clocks (Clock 1-4) on independent intervention data predicts significant age
342 deceleration for growth hormone receptor knock-out (GHRKO), mutant Tet3, or calorie restricted (CR)
343 mice after multiple test correction (**Figure 5D, Extended Data Figure 9B-D**). Each intervention group
344 showed on average strong effect sizes for all 4 clocks (see **Source Data** for full statistics). GHRKO liver
345 samples have a Cohen's d of 1.96 for Clock 1 (**Extended Data Figure 9B**), Tet3 mutant Cerebral Cortex
346 samples have a Cohen's d of 3.7 for Clock 1 (**Extended Data Figure 9C**), and calorie restricted liver
347 samples have a Cohen's d of 1.65 (**Extended Data Figure 9D**). In a dataset of human smokers, previous
348 smokers, and never smokers our stochastic clocks predict a significant age acceleration trajectory in
349 the smokers over the study course as calculated by a multivariate regression analysis (**Figure 5D,**
350 **Extended Data Figure 9E**). We further validated our 4 clocks on an independent dataset on parabiosis
351 in young and old mice⁴⁸. A multivariate regression analysis showed that the predictions of Clock 1-4
352 are all highly significantly correlated with the chronological age (**Figure 5E** p-value: 7.8e-18, **Extended**
353 **Data Figure 9F-H** p-values: 6.1e-12, 5.6-09, 1.3e-06 respectively). Clock 1 and 2 additionally showed a
354 significant interaction term, indicating that heterochronic parabiosis in old mice leads to a younger
355 predicted age compared to isochronic parabiosis, while there is no difference in young mice. These
356 results further validate the chronological age prediction in independent datasets and corroborate that
357 biological age is robustly predictable with accumulating stochastic variation.

358 To assess the effect of the ground state on predictions we build clocks for 12 different species orders,
359 resulting on average in highly significant correlations with values ranging from 0.6 for Clock 1 starting
360 from a Monotremata sample to 0.85 for Clock 1 starting from a Artiodactyla sample (**Figure 6A,**
361 **Extended Data Figure 10A-B**). Clock 2-4 show similar results (**Extended Data Figure 10C-E**). A clock
362 built from the ground state of one order does not improve the prediction accuracy of species within
363 the same order on average (**Figure 6A**).

364 To assess whether 'age-reversal' could be measured by a stochastic data-based clock, we applied it to
365 an independent reprogramming time-course of human dermal fibroblasts⁴⁹. Despite differences in
366 species, tissue-type and platform, a rejuvenation trajectory became evident, with a decreasing
367 predicted age starting from 11 days of intermediate reprogramming and reaching the final lowest

368 predicted age at 28 days (**Figure 6B**, one-way ANOVA p-value: 8.4e-09). These results show that the
369 stochastic data-based clock could identify study-/tissue- and platform-independent signatures of age
370 and captures biological aging as shown by the gradual decrease of the predicted age over the
371 reprogramming time-course, as well as correctly predicted biological age-differences in interventions.

372 Discussion

373 During aging a range of biomolecular parameters show increased 'noise' such as stochastic DNA
374 methylation drifts, degrading transcriptional networks in mouse muscle stem cells⁵⁰, and increased
375 cell-to-cell gene expression variation⁵¹. Transcriptomic variation can result from intrinsic (biochemical
376 fluctuations and transcriptional bursting)⁵² and extrinsic noise like stochastic DNA damage⁵³.
377 Predominantly affecting long genes⁵⁴, transcription-blocking DNA lesions might explain the age-
378 associated systemic transcript-length imbalance^{55,56}. The role of stochasticity in transcription remains
379 subject to debate as a recent study reported a lack of evidence for increased transcriptional single-cell
380 noise in aged tissues⁵⁷.

381 Stochastic changes occur during DNA methylation site copying or maintenance, like DNA repair and
382 subsequent Dnmt1 recruitment⁵⁸, or DNA replication⁵⁹ as replication timing during S-phase itself has
383 been shown to affect methylation maintenance levels⁶⁰. The information-theoretic view of the
384 epigenome²³ suggests that higher maintenance, and therefore lower information loss, consumes more
385 energy and is focused on more crucial regions of the genome.

386 The increased entropy with aging has been associated with higher hemi-methylation²³, is correlated
387 with chronological age, and longer-lived mice showed a lower entropy at age-related CpGs⁶¹, which
388 are enriched in transcription factors and regulators of development and growth⁶². The epigenetic
389 maintenance system (EMS) theory²⁶ postulates that age-related epigenetic changes are the footprint
390 of an imperfect maintenance system, leading to an increase in errors over time. CpG maintenance in
391 genomic regions that are important for development might become less relevant during aging, leading
392 to faster stochastic variation accumulation. It was suggested that only 10 % of CpG sites are driven by
393 biological stochastic variation⁶³. Our single-cell simulation results, in contrast, are in line with a recent
394 report that showed a majority of CpG sites change stochastically³³ even though only ~500 CpG sites
395 could be analyzed due the low coverage of single-cell data⁶⁴.

396 The most trivial model of a stochastic process that can potentially be used for an age prediction, is a
397 process that starts at a ground state of all 0's and has a certain low probability to switch to 1. Such a
398 system will inevitably accrue changes, i.e. 1's, over time. If the probability to switch from 0 to 1 is high
399 enough for an accumulation over the timeframe of a lifespan, the sum of 1's can be used as the
400 simplest predictor of age. The accumulation of DNA mutations could be seen of one example of this

401 simplest case. Similarly, simulated stochastic changes in single-cell DNA methylation using an
402 exponential decay approach starting with either 0 or 1 for all sites before applying stochastic changes,
403 allowed for accurate predictions of the simulated age, in line with the regression-to-the-mean model,
404 since each site starts at the extreme and can only diverge from it³³.

405 In contrast to a multiplicative model, which models a gradual slowdown of methylation change over
406 time³³, we modeled the stochastic variation accumulation in an additive manner, i.e. without a
407 dependency of the random variation on the state of the system. We show that stochastic data-based
408 clocks also predict chronological age and lifespan effects in transcriptome data of *C. elegans* and could
409 measure the age deceleration resulting from reduced transcription drift through Mianserin
410 treatment³⁹.

411 First as well as second generation DNA methylation aging clocks significantly correlate with the amount
412 of stochastic variation in the data, suggesting that chronological and biological aging clocks are
413 measuring stochastic variation. The prediction of all tested clocks caps off after a certain amount of
414 stochastic variation, possibly indicating an approach to site-specific equilibria. Cell-type composition
415 was shown to change with age and to affect clock predictions^{65,66}. While this is an important aspect for
416 the interpretation of clocks and the analysis of differentially methylated regions, correcting for cell-
417 type composition did not change our results, and our DNA methylation simulations incorporating fixed
418 or random maintenance rates cannot be confounded by a composition change over age. In line with
419 this, age-related variably methylated positions are suggested to not be driven by variations in cell type
420 composition^{29,67}. Publicly available clock predictions significantly correlate with the simulated age even
421 if the same constant maintenance rate for all CpGs, or even random maintenance rates, are used. A
422 cell-type corrected stochastic data-based clock maintains accurate predictions of independent cell-
423 type corrected biological samples, underscoring that cell-type composition is not critical for the
424 predictive power of stochastic variation accumulation. While estimating E_m and E_d values is imperfect
425 and likely cell-type dependent, our stochastic simulations are robust regardless of whether
426 maintenance rates are estimated, randomly chosen, or fixed to a universal value.

427 We replicated our results on data from the Mammalian Methylation Consortium³⁵. Contrary to
428 previous proposals that age-related CpG sites were not stochastic marks accrued with age¹³⁻¹⁵, our
429 results show that a stochastic process and one single biological sample as ground state are sufficient
430 to (1) build predictors significantly correlated with the relative age in various mammalian species, and
431 (2) predict the age-accelerating or decelerating effects of interventions such as growth hormone
432 receptor knockout, calorie-restriction, or smoking.

433 OSKM reprogramming has been suggested to revert cellular aging by resetting the DNA methylation
434 landscape via de-differentiation⁶⁸. The predictions with a stochastic data-based clock of a

435 reprogramming time-course indeed follows the expected rejuvenation trajectory. Our work suggests
436 that interventions (potentially even rejuvenation) could reduce and perhaps reverse stochastic
437 variation.

438 The fact that aging clocks strongly correlate with the amount of stochastic variation cautions with
439 regards to the identification of causal effects. CpG sites that show faster stochastic variation
440 accumulation are likely less efficiently maintained and less important for cell survival or homeostasis,
441 making aging clock CpG sites unsuitable for the development of novel geroprotectors¹⁰. Indeed, many
442 chronological aging clocks can be built from DNA methylation data and clock CpG sites might have
443 limited value for understanding biology or anti-aging interventions⁶⁹.

444 Stochastic data-based aging clocks demonstrate the compatibility of precise measures of the pace of
445 aging with entropy-driven stochastic variations in biological processes such as age-associated damage
446 accumulation. These results emphasize that a precise aging pace measure does not require a
447 programmed process, but is consistent with a stochastic nature of the molecular alterations. While we
448 show that accumulation of stochastic variation is sufficient to build aging clocks, the limitation of our
449 study is that a deterministic aging trajectory could also be measured by a programmed clock. Thus, our
450 results do not completely rule out the existence of deterministic processes. In certain species
451 deterministic processes regulate the aging process, as seen in the monarch butterfly's aging rate
452 variation with migration routes⁷⁰. Maintenance and repair mechanisms were selected during evolution
453 for early but not indefinite somatic maintenance, as for instance the limitation of somatic DNA repair
454 capacities by the DREAM complex in *C. elegans*⁷¹. Somatic proteostasis declines rapidly in nematodes,
455 as the heat shock response is repressed during reproduction onset via programmed *jmjd-3.1* reduction,
456 which can be alleviated by removing the germline consistently with the disposable soma theory⁷². The
457 genetically programmed limitations of such maintenance and repair capacities could then result in the
458 age-dependent accumulation of stochastic damage.

459 Stochastic errors might start accumulating from conception, in line with the suggestion that aging
460 starts from mid-embryonic development⁷³. This might start a vicious spiral, since every additional error
461 could disturb the intricate regulatory networks including maintenance systems thus allowing for more
462 errors to be made⁷⁴. It will be interesting to explore in how far a tightening of regulatory mechanisms
463 could slow the aging process, consistently with the epigenetic maintenance system (EMS) theory²⁶.

464 We propose that in addition to methylation clocks, any set of biological measures, whether molecular
465 or physiological, could in principle be used for building aging clocks, as long as the data have a range
466 limit and experience accumulating stochastic variation. The sufficiency of stochasticity for building
467 aging clocks unifies the exact determination of age and the reduced maintenance of homeostatic
468 processes driving the aging process. Indeed, our analysis predicts that the level of such stochasticity

469 sets the pace of aging. Reinstating regulatory tightness could therefore provide opportunities for aging
470 decelerating therapies.

471

472 Methods

473 Bulk Simulations

474 A ground state was generated with 2000 (or indicated otherwise) random features between 0 and 1.
475 From this ground state 6 independent sets of 100 samples each (one sample per age from 1-100) were
476 generated. Each of these 600 samples started from the same ground state with slight deviations, i.e.
477 each sample started with stochastic variation generated from $N(\mu = 0, \sigma^2 = 0.01^2)$ added to the
478 ground state to simulate biological variation. To model age-dependent stochastic variation
479 accumulation, random noise was generated from a normal distribution $N(\mu = 0, \sigma^2)$ with
480 `random.randn()` from Numpy v.1.18.5⁷⁵. The standard deviation σ used for generation of stochastic
481 variation that is applied at each time-step is indicated in the figure legends. The simulated age of each
482 sample defined how often stochastic variation generated from $N(\mu = 0, \sigma^2)$ was independently added
483 to the ground state. For example, for a sample with simulated age 2, stochastic variation would be
484 added twice to the ground state. The stochastic variation addition was performed independently from
485 all other samples, i.e. ground state + 2x stochastic variation independently sampled from the normal
486 distribution. A sample with simulated age 10 is simulated by taking the ground state and adding,
487 independently sampled, normal-distributed stochastic variation 10 times (**Extended Data Figure 1A**).
488 After stochastic variation addition values were kept between 0 and 1, by setting values bigger 1 to 1
489 and values smaller 0 to 0 (except for the results in Extended Data Figure 1C, where no limits were
490 applied). To train a predictor of the simulated age we used 3 sets of 100 independent samples for
491 training of an Elastic net regression model with `ElasticNetCV` from `sklearn v.0.23.1`⁷⁶ with the following
492 parameter: `l1_ratio=[0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9]`. The remaining 3 sets of 100 independent
493 samples were used as a hold-out validation dataset.

494 Logit transform

495 Analysis done with the logit transform were processed the following way. The ground state was first
496 transformed with `logit()` from Scipy⁷⁷. Stochastic variation was generated and applied as described
497 above and added to the logit-transformed ground state. After stochastic variation addition values were
498 transformed back with the inverse-logit transform `expit()` from Scipy⁷⁷.

499 Human single-cell simulations

500 The ground state of single-cell simulations consists of 2000 (or indicated otherwise) randomly chosen
501 CpG sites of the youngest sample in GSE41037 ⁷⁸ (GSM1007467). For the clock starting from a fetal
502 sample, a umbilical cord blood sample in GSE154915 (GSM4682890) was chosen. Each of the features
503 (CpG sites) is a number between 0 and 100 % and used to generate 1000 cells with binary values for
504 each feature. A ground state value of 0.13, i.e. 13 % methylated, generates 1000 cells for which 130
505 are 1 (methylated), and 870 are 0 (unmethylated). One sample therefore consists of 2000 (or indicated
506 otherwise) features with each 1000 simulated cells with binary values of either 1 or 0. Note that our
507 ground state is derived from bulk sequencing and not single-cell data, since single-cell omics come with
508 large technical problems and drawbacks including the sparsity of sequencing coverage, which make it
509 unfavorable as a starting point for our simulations⁶⁴. Next, for each feature a methylation maintenance
510 efficiency E_m and *de novo* methylation efficiency E_d was generated. As indicated in the figure legends,
511 we either simulated data with a universal maintenance efficiency for all features, random efficiencies,
512 or we estimated E_m and E_d from empirical data. For the empirical maintenance estimation, we set the
513 site-specific DNA methylation equilibrium to be the value of the oldest sample in the dataset
514 (GSM1007832 ⁷⁸), as DNA methylation trends towards the equilibrium over time ^{24,25} and estimated
515 E_m and E_d from the equation given by Pfeifer et al. ²⁴:

$$M_{eq} = \frac{E_d}{1 + E_d - E_m} \quad [1]$$

516
517 , where M_{eq} is the equilibrium of the methylation state. Several groups suggested a biological range
518 for E_m and E_d values, with E_m to be on average ~99.9 % and E_d to be ~ 5 % ²⁴, E_m to be ~95 % and
519 for many sites bigger than 99 % ²⁵, or E_m to be between 95-98 % and E_d to be maximally 23 % ⁷⁹. These
520 limits guide our simulations, ensuring both E_m and E_d are within biologically meaningful regions (
521 $95\% < E_m \leq 100\%$ and $0\% \leq E_d < 23\%$). Note that the values inferred by those 3 publications
522 only serve as an estimation of the biologically meaningful regions ($95\% < E_m \leq 100\%$ and $0\% \leq$
523 $E_d < 23\%$), but not for the estimation of the site-specific values itself. Due to the nature of this
524 empirical estimation either E_m or E_d are fixed, allowing the other to be estimated from data. Note,
525 that it is unlikely that all sites will have reached their equilibria with old age. This is therefore only a
526 rough approximation of the site-specific equilibria, and that multiple E_m and E_d values will regress to
527 the same equilibrium over time (compare equation 1). The lower the limit for E_m , respective the higher
528 the limit for E_d , the higher the stochastic variation per time-step on average, since each site (feature)
529 is potentially less well maintained, leading to a quicker regression to the equilibrium (the perfect
530 maintenance would be $E_d=0\%$, and $E_m=100\%$). For example, CpG sites with $E_m=99\%$ and $E_d=1\%$ will
531 regress towards 0.5 slower than CpG sites with $E_m=90\%$ and $E_d=10\%$. Next, we randomly altered the
532 state of every single-cell CpG site based on the respective E_m and E_d values for each time step., i.e.

533 for each time-step we flip a coin with the probabilities E_m (to stay methylated) and E_d (to *de novo*
534 methylate) for each CpG site in each cell. 100 (or indicated otherwise) age steps, i.e. stochastic
535 variation applications, from 0 to 99 (or indicated otherwise) were simulated. The simulations for
536 GrimAge needed 450k Human Methylation Beadchip data and started from the youngest human blood
537 sample in GSE40279 (GSM990528)⁸⁰. The maintenance rates were estimated from the oldest sample
538 (GSM989863). For training and validating a predictor, we again computed the average bulk
539 methylation levels for each site and time-point. The training and validation process of the Elastic net
540 regression is the same as described in Extended Data Figure 1B.

541 **Cell-type correction**

542 The cell-type composition was first estimated with EpiDISH⁸¹ with the parameter
543 `ref.m=centDHSbloodDMC.m` and `method='RPC'` in R-4.3. The estimated cell-type composition was
544 subsequently used in a regression-based correction approach⁸². Briefly, a linear model is fit for every
545 CpG site using the cell-type composition values via $\text{lm}(x \sim B + NK + CD4T + CD8T + Mono + Neutro + Eosino)$ to
546 estimate the variance in the data that is predicted by the blood cell-type proportions. The remaining
547 residuals depict the variance that is cell-type independent and can be added to the mean methylation
548 value for each site to obtain the adjusted beta values⁸². Additionally, we calculated a multivariate linear
549 regression model of the form

550 $\text{Age} \sim \text{PredictedAge} + \text{CellTypeFractions}$

551 which gives p-values for each of the variables, i.e. also whether the predicted age is significantly
552 associated with the chronological age when also correcting for cell-type fractions.

553 **Public aging clocks**

554 We downloaded the Elastic net regression coefficients for Horvaths pan-tissue clock²⁶, Vidal-Bralo's
555 blood aging clock⁴¹, Lin's 99-CpG clock⁴², Weidner's 3-CpG clock⁴³, and Levine's PhenoAge⁴⁰ clock and
556 applied them to simulated data. The data were simulated as defined above, with the difference that
557 we only used the clock-specific CpG sites as the features in the ground state, and we started the
558 arbitrary simulated age at 16, i.e. the age of the subject of the ground state sample. Stochastic variation
559 was simulated either with a universal maintenance efficiency for all CpG sites, or with empirically
560 estimated maintenance rates as defined above. For GrimAge⁴⁴ predictions we uploaded the simulated
561 datasets to the webpage: <https://dnamage.genetics.ucla.edu/>.

562 **Human stochastic data-based clock**

563 The stochastic data-based clock was computed based on simulations described above. The scale and
564 units of the simulated age are arbitrary since we do not know when or in which time-steps the noise

565 increases, and are therefore different from the chronological age of biological samples. We found that
566 a rescaling of the simulated age before training and testing the model is beneficial. First, we rescaled
567 via min-max scaling the simulated age to be within 0 and 1, multiplied it by 400 and subtracted 120.
568 Note that this transformation on the arbitrary time-steps will not interfere with the correlation
569 analyses. For the correlation analyses, we excluded the youngest (GSM1007467, or GSM4682890; from
570 which the ground state was sampled), and the oldest (GSM1007832; from which the maintenance
571 efficiencies were estimated as described above) to not confound the correlation between the
572 chronological age of samples in GSE41037⁷⁸, and the predicted age. To train a predictor of the
573 simulated age we used 1 set of 1 independent sample per age step from 1 to 73 for training of an
574 Elastic net regression model with ElasticNetCV from sklearn v.0.23.1⁷⁶ with the following parameter:
575 $l1_ratio=[0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9]$, $alphas=[1]$. The clock was validated on 11,146
576 independent whole blood or peripheral blood leukocyte samples from the Illumina Infinium 450k
577 Human Methylation Beadchip and the Illumina Infinium MethylationEPIC Beadchip (GSE84727,
578 GSE87571, GSE80417, GSE40279, GSE87648, GSE42861, GSE50660, GSE106648, GSE179325,
579 GSE210254, GSE210255, GSE72680, GSE147740, GSE55763, GSE117860).

580 **Pan-mammalian clocks**

581 The pan-mammalian stochastic data-based clocks (Clock 1-4) are built on the youngest blood sample
582 from *Tursiops truncatus* as the ground state (or stated otherwise) from the Illumina
583 HorvathMammalianMethylChip40 BeadChip platform. Clock 1 used empirically estimated
584 maintenance efficiency rates from the oldest sample of the same tissue and species as the ground
585 state for all CpG sites of Lu's pan-mammalian relative age-clock. Clock 2 uses the same CpG sites, but
586 non-empirically estimated 99% maintenance rate for all sites (or stated otherwise). Clock 3 is the same
587 as Clock1 but utilizes all 37554 CpG sites. Clock 4 is the same as Clock2 but utilizes all 37554 CpG sites.
588 To train a predictor of the simulated age we used 1 set of 1 independent sample per age step from 1
589 to 67 for training of an Elastic net regression model with ElasticNetCV from sklearn v.0.23.1⁷⁶ with the
590 following parameter: $l1_ratio=[0.01, 0.001]$, $alphas=[1]$. The predictor was trained to predict $-\log(-$
591 $\log(\text{SimulatedAge}/\text{MaxAge}))$ as described by Lu et al.¹⁵, where MaxAge is the number of age steps
592 simulated, i.e. 67. To get the relative age back, the predictions are transformed back via $\exp(-\exp(-$
593 $\text{PredictedAge}))$. Lu et al. employed leave-one-fraction-out and leave-one-species-out cross-validation
594 to get an unbiased estimate of the clock's accuracy¹⁵. Since the stochastic data-based clock only needs
595 one biological sample as a ground state we directly applied the clock to all samples, thereby further
596 reducing the risk of an accuracy bias. To calculate the Pearson correlation of the predicted and relative
597 age of species, only species with at least 5 samples (or stated otherwise) were taken. Note that the
598 species have distinct age ranges, which is affecting the Pearson correlation values. For the validation
599 of our stochastic data-based clocks on interventions with known lifespan effects for growth hormone

600 receptor-knockout, *Tet3* knockout, or calorie restricted mice, we calculated the adjusted FDR and used
601 the t-value of a two-sided t-test for the color gradient (Control vs. experimental mice; a positive value
602 indicating a younger predicted age in the experimental mice).

603 The statistics for the liver samples of the parabiosis dataset (GSE224361) and the slope difference of
604 smoking individuals (GSE50660) were calculated with Python's
605 statsmodels.regression.linear_model.OLS and the following regression models:

606 Parabiosis (GSE224361):

$$607 \quad \text{PredictedAge} \sim \text{ChronologicalAge} + \text{HeterochronicParabiosis} + \text{ChronologicalAge} \\ 608 \quad \quad \quad * \text{HeterochronicParabiosis}$$

609 Where HeterochronicParabiosis is a binary variable indicating whether the parabiosis was
610 heterochronic or isochronic.

611 Smoking (GSE50660):

$$612 \quad \text{PredictedAge} \sim \text{ChronologicalAge} + \text{ExSmoker} + \text{CurrentSmoker} + \text{ChronologicalAge} \\ 613 \quad \quad \quad * \text{ExSmoker} + \text{ChronologicalAge} * \text{CurrentSmoker}$$

614 Where ExSmoker and CurrentSmoker are binary variables indicating the smoking status of the
615 sequenced individuals. The significant interaction term *ChronologicalAge * CurrentSmoker*
616 indicates a steeper slope, i.e. faster aging trajectory, and is shown as negative values in Figure 5D. The
617 smoking dataset and the reprogramming time-course dataset of human dermal fibroblasts (GSE54848)
618 ⁴⁹ were generated with the Illumina Infinium HumanMethylation450 BeadChip array and was
619 converted by the Array Converter Algorithm of the Mammalian Methylation Consortium before
620 predicting the samples ¹⁵.

621 Gillespie algorithm

622 For the simulations we adapted the code from⁸³. We modelled each CpG site with 2 different
623 equations, one for the methylation, one for the demethylation. The probability of switching the state
624 from one to the other was set to 0.1 for both equations. tmax was set to 5 and nrmax to 8000. The
625 arbitrary time-steps (of 0-5) were scaled to be within the same range of the predicted age. Note that
626 this does not affect the Pearson correlation results.

627 Public RNA-seq processing

628 All 994 public RNA-seq samples were downloaded and processed the same. First, we preprocessed
629 samples with Fastp v0.20.0 ⁸⁴ with the following parameters -g -x -q 30 -e 30. After preprocessing, the
630 samples were mapped with Salmon v1.1 ⁸⁵ and the parameters -validateMappings -seqBias and for

631 paired-end samples additionally `-gcBias`. The decoy-aware index for Salmon was generated with the
632 WS281 transcriptome build from Wormbase⁸⁶. The results of Salmon were combined to the gene-level
633 with `tximport v1.14.2`⁸⁷. Raw counts were log10-transformed after the addition of one pseudo-count,
634 each sample was min-max normalized to bring each sample within the data range 0-1, and genes 0 in
635 all 994 samples were filtered out. To binarize the data zeroes were masked by NaN, the median was
636 calculated, and genes bigger than the median were set to 1, and all other genes to 0³⁷.

637 **Transcriptomic stochastic variation simulation**

638 The ground state consists of all (or indicated otherwise) gene counts (normalized as described above)
639 of the biologically youngest sample (GSM2916344³⁸). From this ground state 10 independent samples
640 for each time-step (from 1 to 16) were generated (based on the distribution that resulted in the best
641 correlation with BitAge (**Extended Data Figure 2A**)) and used to train an Elastic net regression as
642 described above (see Bulk simulations). Note that the simulated age range is arbitrary, and the scale
643 and unit not directly comparable to the biological age. Similar to the epigenetic stochastic-data based
644 clock we found a rescaling of the arbitrary simulated time-steps by 2 to be beneficial, i.e. we multiplied
645 the simulated age by 2 before training and testing the data. The Elastic net regression model was then
646 used to predict the biological age of the 993 remaining (excluding the youngest sample which was used
647 for the ground state) *C. elegans* samples. The biological age is calculated by temporal rescaling of the
648 chronological age by the median lifespan. Briefly, we set a reference lifespan of a standard worm
649 population to 15.5 days of adulthood and calculate a rescaling factor for every sample by dividing this
650 reference lifespan by the median lifespan reported by the publication of the corresponding sample.
651 This rescaling factor is multiplied with the chronological age of the sample³⁷.

652 **Statistics & Reproducibility**

653 All indicated public data were used for validation, except for samples used as the ground state or to
654 estimate maintenance rates as indicated. No statistical method was used to predetermine sample size.
655 Stochastic variation accumulation simulations were done at least N=3 times as indicated in the figure
656 legends and can be reproduced with the public code. Data analyses were not performed blinded.
657 Statistical tests used, are indicated in the figure legends. Full statistics can be found in the **Source Data**
658 **File**. All data plots were done with Seaborn-0.11.0⁸⁸ and Matplotlib-3.3.0⁸⁹. Boxplots are shown with
659 the center line depicting the median, the box limits the bottom, respective top quartiles, and the
660 whiskers the 1.5x interquartile range. Scatterplots showing a linear regression model fit are shown
661 with a 95% confidence interval. Pearson correlations were computed with Scipy-1.5.1's `stats.pearsonr`
662 function⁷⁷ and two-sided tests. Effect sizes (Cohen's d and Hedges g) for pair-wise comparisons were
663 computed with Pingouin-0.3.6's `compute_effsize` function⁹⁰.

664

665 Data availability statement

666 The human DNA methylation data is available at NCBI GEO (accession code GSE84727, GSE87571,
667 GSE80417, GSE40279, GSE87648, GSE42861, GSE50660, GSE106648, GSE179325, GSE210254,
668 GSE210255, GSE72680, GSE147740, GSE55763, GSE117860, GSE41037, GSE54848, GSE223748, and
669 GSE224361). The accession codes for all 994 public *C. elegans* RNA-seq samples can be found in
670 Supplementary Table 1. The WS281 transcriptome version of *C. elegans* was downloaded from
671 Wormbase⁸⁶.

672 Code availability statement

673 The code for the simulations can be found in a supplementary file and at [https://github.com/Meyer-](https://github.com/Meyer-DH/StochasticAgingClock)
674 [DH/StochasticAgingClock](https://github.com/Meyer-DH/StochasticAgingClock). The BiT age clock code can be found at [https://github.com/Meyer-](https://github.com/Meyer-DH/AgingClock)
675 [DH/AgingClock](https://github.com/Meyer-DH/AgingClock). The Gillespie algorithm can be found at [https://github.com/karinsasaki/gillespie-](https://github.com/karinsasaki/gillespie-algorithm-python/blob/master/build_your_own_gillespie_solutions.ipynb)
676 [algorithm-python/blob/master/build_your_own_gillespie_solutions.ipynb](https://github.com/karinsasaki/gillespie-algorithm-python/blob/master/build_your_own_gillespie_solutions.ipynb). The
677 ArrayConverterAlgorithm can be found at
678 [https://github.com/shorvath/MammalianMethylationConsortium/tree/main/UniversalPanMammali-](https://github.com/shorvath/MammalianMethylationConsortium/tree/main/UniversalPanMammalianClock/R_code/ArrayConverterAlgorithm)
679 [anClock/R_code/ArrayConverterAlgorithm](https://github.com/shorvath/MammalianMethylationConsortium/tree/main/UniversalPanMammalianClock/R_code/ArrayConverterAlgorithm)

680 Acknowledgements

681 We thank Khrystyna Totska, Robert Bayersdorf, and Arturo Bujarrabal-Dueso for comments on the
682 manuscript and the Regional Computing Center of the University of Cologne (RRZK) for providing
683 computing time and support on the DFG-funded High Performance Computing (HPC) system CHEOPS.
684 D.M. was supported by the Cologne Graduate School of Ageing Research. B.S. acknowledges funding
685 from the Deutsche Forschungsgemeinschaft (Reinhart Koselleck-Project 524088035, FOR 5504 project
686 496650118, SCHU 2494/3-1, SCHU 2494/7-1, SCHU 2494/10-1, SCHU 2494/11-1, SCHU 2494/15-1,
687 CECAD EXC 2030 – 390661388, SFB 829, KFO 286, KFO 329, and GRK 2407), the Deutsche Krebshilfe
688 (70114555), the H2020-MSCA-ITN-2018 (Healthage and ADDRESS ITNs) and the John Templeton
689 Foundation Grant (61734).

690 Author Contributions Statement

691 D.H.M and B.S. conceived and designed the study, and wrote the manuscript. D.H.M did all data
692 analysis.

693 Competing Interests Statement

694 The authors declare no competing interests

695 Tables

696 Supplementary Table 1 – List of the 994 RNA-seq samples used

697 Figure Legends

698 **Figure 1 – Normal-distributed stochastic variation accumulation simulations enable aging clock**
699 **construction for simulated data.** (A) Sample generation explanation with logit transform. (B)
700 Accumulating stochastic variation in logit transformed data enables accurate simulated age
701 predictions. The x-axis shows the number of times stochastic variation was added to the ground state.
702 The y-axis shows the prediction of the independent validation data (n=300). (C) The predictions of the
703 independent validation data are robust to the stochastic variation distribution. The x-axis shows the
704 standard deviation of the normal distribution from which the stochastic variation was sampled. The y-
705 axis shows the R² value of the independent validation data predictions (N=3 independent repeats; each
706 with n=300 independent samples). (D) Coefficients of independent models are highly correlated if
707 trained on samples starting from the same ground. Shown are the coefficients of N=2000 features. (E)
708 The prediction in (B) is possible due to a regression to the mean. The x-axis shows the starting values
709 of the 2000 features of the simulated ground state, the y-axis the Elastic net regression coefficients for
710 the model in (B) (trained on n=300). (F) The accuracy of predictions caps off after ~2000 features in the
711 ground state. The x-axis shows how many features were randomly sampled for the ground state. The
712 y-axis shows the R² as a measure of model accuracy. (N=10 independent repeats for Features
713 Sizes<1000, N=3 independent repeats otherwise; each with n=300, 3 independent samples per time
714 point). (G) The amount of stochastic variation sets the pace of aging. The Elastic net regression model
715 was trained with stochastic variation sampled from $N(\mu = 0, \sigma^2 = 0.2^2)$ and tested on independent
716 samples generated from the same ground state, but with varying degrees of stochastic variation (color-
717 coded, as indicated in the panel). All simulated datasets consist of n=300 independent samples.
718 Boxplots in Figure 1 C,F are shown with the center line depicting the median, the box limits the bottom,
719 respective top quartiles, and the whiskers the 1.5x interquartile range.

720 **Figure 2 - Normal-distributed stochastic variation accumulation simulations enable aging clock**
721 **construction for transcriptomic data.** (A) The simulated age (x-axis), and BitAge³⁷ predictions (y-axis)
722 significantly correlate (Pearson correlation of 0.81, p-value 5.99e-41, two-sided test). n=160, 10
723 independent samples per time point. Variation was sampled with a SD of 0.01. (B) The predictions of
724 a transcriptomic stochastic data-based clock (y-axis) correlates significantly (Pearson correlation 0.72,
725 p-value 5.7e-150, two-sided test) with the biological age (x-axis) of the n=993 independent RNA-seq
726 from 61 independent public datasets (**Supplement Table 1**). (C) There is a significant association
727 between the median lifespan and the predicted age of the clock used in (B) (median lifespan coefficient
728 p-value= 0.015, full statistics in **Source Data**). The regression model fit with a 95% confidence interval

729 (shadowed area) is shown for Long-lived (>20d median lifespan in blue), Short-lived (<8d median
730 lifespan in green) and Normal-lived (orange). (D) Mianserin shows a dose-dependent decrease of the
731 predicted age of the clock used in (B). ANOVA (p-value: 0.006) with a two-sided Tukey post-hoc test
732 was used (50 μ M Mianserin vs. Control adjusted p-value 0.026, full statistics in **Source Data**). Boxplots
733 are shown with the center line depicting the median, the box limits the bottom, respective top
734 quartiles, and the whiskers the 1.5x interquartile range. (E) 50 μ M Mianserin shows a lower predicted
735 age over the whole time-course (2-way ANOVA treatment p-value: 7.3e-04, full statistics in **Source**
736 **Data**). The regression model fit with a 95% confidence interval (shadowed area) is shown for worms
737 receiving 50 μ M Mianserin (orange) and Control worms (blue).

738 **Figure 3 – Single-cell DNA methylation stochastic variation accumulation simulations enable aging**
739 **clock construction for simulated data.** (A) Explanation of single-cell simulations. (B) The accuracy of
740 the model is dependent on the methylation maintenance efficiency rate. A stochastic data-based clock
741 was trained with 500 features and universal maintenance efficiencies E_m and E_d and used to predict
742 the simulated age of 300 independent validation samples. The x-axis shows the methylation
743 maintenance efficiency E_m in %. The y-axis shows the R^2 . N=3 independent experiments with different
744 ground states are shown for each maintenance efficiency. Boxplots are shown with the center line
745 depicting the median, the box limits the bottom, respective top quartiles, and the whiskers the 1.5x
746 interquartile range. (C) Single-cell simulation of DNA methylation sites with E_m and E_u of 99.9 % allows
747 to build a clock with highly accurate predictions ($R^2=0.999$) of independent validation data (n=300).
748 The x-axis shows the true simulated age. (D) The accuracy of predictions with a universal maintenance
749 efficiency rate of 99.9 % caps off after ~32 features with an R^2 of 0.99. The x-axis shows the amount of
750 features of the stochastic data-based clock. The y-axis shows the R^2 . N=10 independent repeats for
751 Features Sizes<1000, N=3 independent repeats otherwise; each with n=300, 3 samples per time point.
752 Boxplots are shown with the center line depicting the median, the box limits the bottom, respective
753 top quartiles, and the whiskers the 1.5x interquartile range. (E) The maintenance efficiency rate sets
754 the pace of aging. The stochastic data-based clock was trained with a maintenance efficiency of
755 $E_m=E_u=99.9\%$, and tested on independent samples generated from the same ground state, but with
756 varying maintenance efficiencies (color-coded, as indicated in the panel). All simulated datasets consist
757 of n=300 independent samples. (F) Biologically estimated maintenance rates allow for highly accurate
758 predictions. Site-specific E_m and E_u values were estimated from data (see methods for details). The
759 simulations were done the same as in C) but with site-specific maintenance rates. (n=300).

760 **Figure 4 – Epigenetic aging clock predictions correlate significantly with the amount of stochastic**
761 **variation.** (A) The methylation maintenance efficiency limits affect the simulation and subsequent
762 prediction with Horvath's epigenetic clock²⁶. The x-axis shows the limit of E_m . Color-coded is the limit

763 of E_d . The y-axis shows the R^2 between the predicted epigenetic age by Horvath's epigenetic clock²⁶
764 and the simulated age. N=3 independent repeats, each consisting of n=73 independent samples.
765 Boxplots are shown with the center line depicting the median, the box limits the bottom, respective
766 top quartiles, and the whiskers the 1.5x interquartile range. (B) Horvath's epigenetic age prediction²⁶
767 of samples simulated based on biologically estimated maintenance rates with the limits $E_m > 97\%$
768 and $E_d < 5\%$, correlates significantly with the simulated age. N=73 independent samples. (C) Horvath's
769 epigenetic age prediction²⁶ of samples simulated based on a universal maintenance efficiency rate of
770 99% for all features, correlates significantly with the simulated age. N=73 independent samples. (D)
771 The methylation maintenance efficiency limits affect the simulation and subsequent prediction with
772 PhenoAge⁴⁰. The x-axis shows the limit of E_m . Color-coded is the limit of E_d . The y-axis shows the R^2
773 between the predicted epigenetic age by PhenoAge⁴⁰ and the simulated age. N=3 independent
774 repeats, each consisting of n=73 independent samples. Boxplots are shown with the center line
775 depicting the median, the box limits the bottom, respective top quartiles, and the whiskers the 1.5x
776 interquartile range. (E) Biological age prediction with PhenoAge⁴⁰ of samples simulated based on
777 biologically estimated maintenance rates with the limits $E_m > 97\%$ and $E_d < 5\%$, correlates
778 significantly with the simulated age. N=73 independent samples. (F) Biological age prediction with
779 PhenoAge⁴⁰ of samples simulated based on a universal maintenance rate of 99% for all features,
780 correlates significantly with the simulated age. N=73 independent samples.

781 **Figure 5 - Single-cell DNA methylation stochastic variation accumulation simulations enable aging**
782 **clock construction for pan-mammalian chronological and biological age predictions.** (A) The
783 predictions of a stochastic data-based clock, correlates significantly (Pearson correlation 0.87, p-
784 value<1e-16, two-sided test) with the chronological age of the cell-type corrected independent healthy
785 biological validation samples (GSE41037, n=392)⁷⁸. (B) The validation of the stochastic data-based
786 clock starting from a fetal sample (GSM4682890) on 11,146 independent samples from 15
787 independent datasets (GSE84727, GSE87571, GSE80417, GSE40279, GSE87648, GSE42861, GSE50660,
788 GSE106648, GSE179325, GSE210254, GSE210255, GSE72680, GSE147740, GSE55763, GSE117860)
789 shows a significant correlation (Pearson correlation 0.72, p-value<1e-16, two-sided test). (C) A circle
790 plot showing the Pearson correlation between the relative age of blood samples of the corresponding
791 species and the predictions of Clock 1 as a green line around the circle. Species are shown for which at
792 least 5 blood samples were available in the dataset GSE223748. The colors within the circle show the
793 taxonomic order of the corresponding species, as listed on the left side. (D) Validation of Clock 1-4 on
794 interventions with known lifespan effects in mouse and humans. Age-matched growth hormone
795 receptor-knockout (GHRKO) with 30 normal (12 liver, 12 kidney, 6 cerebral cortex) and 29 GHRKO (11
796 liver, 12 kidney, 6 cerebral cortex) mice¹⁵. *Tet3* knockout mice with 28 normal (14 striatum, 14 cerebral
797 cortex) and 16 *Tet3* (8 striatum, 8 cerebral cortex) mice¹⁵. 36 calorie restricted (CR) mice with 59

798 normal mice¹⁵. The effect of smoking on human aging⁹¹. The color gradient for mice is based on the
799 sign of the t-test, the color of the human data is based on the interaction coefficient. The annotated
800 values show the adjusted FDR, full statistics in **Source Data**. (E) Independent validation of Clock 1 on
801 parabiosis in young and old mice (GSE224361). Liver samples of mice that received either isochronic
802 (orange) or heterochronic (blue) parabiosis are shown. A multivariate regression shows a significant
803 age variable ($p < 1 \times 10^{-16}$), and interaction variable ($p = 1.22 \times 10^{-3}$), full statistics in **Source Data**. The
804 regression model fit with a 95% confidence interval (shadowed area) is shown.

805 **Figure 6 - Single-cell DNA methylation stochastic variation accumulation simulations enable**
806 **predictions for various species and reprogramming.** (A) Heatmap showing median Pearson
807 correlations of species within the same taxonomic order between the predicted age of Clock 1 trained
808 on the youngest blood sample from species of the corresponding taxonomic order in the columns
809 (Artiodactyla: *Tursiops truncatus*, Carnivora: *Odobenus rosmarus divergens*, Lagomorpha: *Oryctolagus*
810 *cuniculus*, Monotremata: *Tachyglossus aculeatus*, Perissodactyla: *Equus caballus*, Pilosa: *Choloepus*
811 *hoffmanni*, Proboscidea: *Loxodonta africana*, Rodentia: *Marmota flaviventris*, Sirenia: *Trichechus*
812 *manatus*, Suidae: *Sus scrofa*, Tubulidentata: *Orycteropus afer*) and the relative age for all species in the
813 rows. Values are shown for tissues and species for which at least 5 samples were available. (B) The
814 stochastic data-based clock in Figure 5C was used on an independent reprogramming time-course
815 dataset of human dermal fibroblasts (GSE54848)⁴⁹. The x-axis shows the time in days of
816 reprogramming, the y-axis shows the predicted simulated age. 1-way ANOVA $p = 8.36 \times 10^{-9}$ (Statistics in
817 **Source Data**). The lineplot shows the mean values with a 95% confidence interval (shadowed area).

818 References

- 819 1. Weismann, A. *Ueber die Dauer des Lebens; ein Vortrag*. (G. Fischer, 1882).
820 doi:10.5962/bhl.title.21312.
- 821 2. Kirkwood, T. B. & Cremer, T. Cyto gerontology since 1881: a reappraisal of August Weismann
822 and a review of modern progress. *Hum. Genet.* **60**, 101–21 (1982).
- 823 3. Vijg, J. & Kennedy, B. K. The Essence of Aging. *Gerontology* **62**, 381–5 (2016).
- 824 4. Kowald, A. & Kirkwood, T. B. L. Can aging be programmed? A critical literature review. *Aging*
825 *Cell* **15**, 986–998 (2016).
- 826 5. Medawar, P. B. An unsolved problem of biology. *Med. J. Aust.* (1952) doi:10.5694/j.1326-
827 5377.1953.tb84985.x.
- 828 6. Williams, G. C. Pleiotropy, natural selection, and the evolution of senescence. *Evolution (N. Y.)*
829 **11**, 398–411 (1957).
- 830 7. Schumacher, B., Pothof, J., Vijg, J. & Hoeijmakers, J. H. J. The central role of DNA damage in
831 the ageing process. *Nature* **592**, 695–703 (2021).
- 832 8. Mitteldorf, J. An epigenetic clock controls aging. *Biogerontology* **17**, 257–265 (2016).
- 833 9. Wagner, W. The Link Between Epigenetic Clocks for Aging and Senescence. *Front. Genet.* **10**,

- 834 1–6 (2019).
- 835 10. Schork, N. J., Beaulieu-Jones, B., Liang, W., Smalley, S. & Goetz, L. H. Does Modulation of an
836 Epigenetic Clock Define a Geroprotector? *Adv. Geriatr. Med. Res.* **4**, 1–11 (2022).
- 837 11. Lidsky, P. V, Yuan, J., Rulison, J. M. & Andino-Pavlovsky, R. Is Aging an Inevitable Characteristic
838 of Organic Life or an Evolutionary Adaptation? *Biochem.* **87**, 1413–1445 (2022).
- 839 12. de Magalhães, J. P. & Church, G. M. Genomes Optimize Reproduction: Aging as a
840 Consequence of the Developmental Program. *Physiology* **20**, 252–259 (2005).
- 841 13. Magalhães, J. P. Programmatic features of aging originating in development: aging
842 mechanisms beyond molecular damage? *FASEB J.* **26**, 4821–4826 (2012).
- 843 14. Gems, D. The hyperfunction theory: An emerging paradigm for the biology of aging. *Ageing
844 Res. Rev.* **74**, 101557 (2022).
- 845 15. Lu, A. T. *et al.* Universal DNA methylation age across mammalian tissues. *Nat. Aging* **5**, 410–
846 410 (2023).
- 847 16. Gems, D., Singh Virk, R., de Magalhães, J. P., Virk, R. S. & Magalhães, J. P. de. Epigenetic Clocks
848 and Programmatic Aging. *Preprints* (2023) doi:10.20944/preprints202312.1892.v1.
- 849 17. Magalhães, J. P. De. Ageing as a software design flaw. *Genome Biol.* 1–20 (2023)
850 doi:https://doi.org/10.1186/s13059-023-02888-y.
- 851 18. Lidsky, P. V & Andino, R. Could aging evolve as a pathogen control strategy? *Trends Ecol. Evol.*
852 **37**, 1046–1057 (2022).
- 853 19. Lee, R. D. Rethinking the evolutionary theory of aging: transfers, not births, shape senescence
854 in social species. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 9637–42 (2003).
- 855 20. Issa, J. Aging and epigenetic drift: a vicious cycle. *J. Clin. Invest.* **124**, 24–9 (2014).
- 856 21. Min, B., Jeon, K., Park, J. S. & Kang, Y. Demethylation and derepression of genomic
857 retroelements in the skeletal muscles of aged mice. *Aging Cell* **18**, 1–13 (2019).
- 858 22. Shipony, Z. *et al.* Dynamic and static maintenance of epigenetic memory in pluripotent and
859 somatic cells. *Nature* **513**, 115–119 (2014).
- 860 23. Jenkinson, G., Pujadas, E., Goutsias, J. & Feinberg, A. P. Potential energy landscapes identify
861 the information-theoretic nature of the epigenome. *Nat. Genet.* **49**, 719–729 (2017).
- 862 24. Pfeifer, G. P., Steigerwald, S. D., Hansen, R. S., Gartler, S. M. & Riggs, A. D. Polymerase chain
863 reaction-aided genomic sequencing of an X chromosome-linked CpG island: Methylation
864 patterns suggest clonal inheritance, CpG site autonomy, and an explanation of activity state
865 stability. *Proc. Natl. Acad. Sci. U. S. A.* **87**, 8252–8256 (1990).
- 866 25. Riggs, A. D. & Xiong, Z. Methylation and epigenetic fidelity. *Proc. Natl. Acad. Sci.* **101**, 4–5
867 (2004).
- 868 26. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **16**, 96 (2013).
- 869 27. Seale, K., Horvath, S., Teschendorff, A., Eynon, N. & Voisin, S. Making sense of the ageing
870 methylome. *Nat. Rev. Genet.* **0123456789**, (2022).
- 871 28. Maegawa, S. *et al.* Caloric restriction delays age-related methylation drift. *Nat. Commun.* **8**,
872 539 (2017).
- 873 29. Sliker, R. C. *et al.* Age-related accrual of methylomic variability is linked to fundamental
874 ageing mechanisms. *Genome Biol.* **17**, 191 (2016).

- 875 30. Petkovich, D. A. *et al.* Using DNA Methylation Profiling to Evaluate Biological Age and
876 Longevity Interventions. *Cell Metab.* **25**, 954-960.e6 (2017).
- 877 31. Bertucci-richter, E. M., Shealy, E. P. & Parrott, B. B. Age related disorder in DNA methylation
878 patterns underlies epigenetic clock signals, but displays distinct responses to epigenetic
879 rejuvenation events. *bioRxiv* (2022).
- 880 32. Levine, M. E., Higgins-chen, A., Thrush, K., Minter, C. & Niimi, P. Clock Work: Deconstructing
881 the Epigenetic Clock Signals in Aging, Disease, and Reprogramming. *bioRxiv [Preprint]* (2022)
882 doi:<https://doi.org/10.1101/2022.02.13.480245>.
- 883 33. Tarkhov, A. E. *et al.* Nature of epigenetic aging from a single-cell perspective. *bioRxiv*
884 *[Preprint]* (2022) doi:10.1101/2022.09.26.509592.
- 885 34. Tarkhov, A. E., Denisov, K. A. & Fedichev, P. O. Aging clocks , entropy , and the limits of age-
886 reversal. *bioRxiv [Preprint]* 1–12 (2022) doi:10.1101/2022.02.06.479300.
- 887 35. Haghani, A. *et al.* DNA methylation networks underlying mammalian traits. *Science* **381**,
888 eabq5693 (2023).
- 889 36. Gladyshev, V. N. The Ground Zero of Organismal Life and Aging. *Trends Mol. Med.* **27**, 11–19
890 (2021).
- 891 37. Meyer, D. H. & Schumacher, B. BiT age: A transcriptome-based aging clock near the
892 theoretical limit of accuracy. *Aging Cell* **20**, 1–17 (2021).
- 893 38. Senchuk, M. M. *et al.* Activation of DAF-16/FOXO by reactive oxygen species contributes to
894 longevity in long-lived mitochondrial mutants in *Caenorhabditis elegans*. *PLOS Genet.* **14**,
895 e1007268 (2018).
- 896 39. Rangaraju, S. *et al.* Suppression of transcriptional drift extends *C. elegans* lifespan by
897 postponing the onset of mortality. *Elife* **4**, e08833 (2015).
- 898 40. Levine, M. E. *et al.* An epigenetic biomarker of aging for lifespan and healthspan. *Aging*
899 (*Albany. NY*). **10**, 573–591 (2018).
- 900 41. Vidal-Bralo, L., Lopez-Golan, Y. & Gonzalez, A. Simplified Assay for Epigenetic Age Estimation
901 in Whole Blood of Adults. *Front. Genet.* **7**, 1–7 (2016).
- 902 42. Lin, Q. *et al.* DNA methylation levels at individual age-associated CpG sites can be indicative
903 for life expectancy. *Aging (Albany. NY)*. **8**, 394–401 (2016).
- 904 43. Weidner, C. I. *et al.* Aging of blood can be tracked by DNA methylation changes at just three
905 CpG sites. *Genome Biol.* **15**, R24 (2014).
- 906 44. Lu, A. T. *et al.* DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging*
907 (*Albany. NY*). **11**, 303–327 (2019).
- 908 45. Gillespie, D. T. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81**,
909 2340–2361 (1977).
- 910 46. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture
911 distribution. *BMC Bioinformatics* **13**, 86 (2012).
- 912 47. Arneson, A. *et al.* A mammalian methylation array for profiling methylation levels at
913 conserved sequences. *Nat. Commun.* **13**, 783 (2022).
- 914 48. Poganik, J. R. *et al.* Biological age is increased by stress and restored upon recovery. *Cell*
915 *Metab.* **35**, 807-820.e5 (2023).

- 916 49. Ohnuki, M. *et al.* Dynamic regulation of human endogenous retroviruses mediates factor-
917 induced reprogramming and differentiation potential. *Proc. Natl. Acad. Sci.* **111**, 12426–12431
918 (2014).
- 919 50. Hernando-Herraez, I. *et al.* Ageing affects DNA methylation drift and transcriptional cell-to-cell
920 variability in mouse muscle stem cells. *Nat. Commun.* **10**, 4361 (2019).
- 921 51. Bahar, R. *et al.* Increased cell-to-cell variation in gene expression in ageing mouse heart.
922 *Nature* **441**, 1011–1014 (2006).
- 923 52. Eldar, A. & Elowitz, M. B. Functional roles for noise in genetic circuits. *Nature* **467**, 167–73
924 (2010).
- 925 53. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic Gene Expression in a Single
926 Cell. *Science (80-.)*. **297**, 1183–1186 (2002).
- 927 54. Gyenis, A. *et al.* Genome-wide RNA polymerase stalling shapes the transcriptome during
928 aging. *Nat. Genet.* **55**, 268–279 (2023).
- 929 55. Stoeger, T. *et al.* Aging is associated with a systemic length-associated transcriptome
930 imbalance. *Nat. Aging* **2**, 1191–1206 (2022).
- 931 56. Ibañez-Solé, O., Barrio, I. & Izeta, A. Age or lifestyle-induced accumulation of genotoxicity is
932 associated with a length-dependent decrease in gene expression. *iScience* **26**, 106368 (2023).
- 933 57. Ibañez-Solé, O., Ascensión, A. M., Araúzo-Bravo, M. J. & Izeta, A. Lack of evidence for
934 increased transcriptional noise in aged tissues. *Elife* **11**, 1–33 (2022).
- 935 58. Mortusewicz, O., Schermelleh, L., Walter, J., Cardoso, M. C. & Leonhardt, H. Recruitment of
936 DNA methyltransferase I to DNA repair sites. *Proc. Natl. Acad. Sci.* **102**, 8905–8909 (2005).
- 937 59. Petryk, N., Bultmann, S., Bartke, T. & Defossez, P. Staying true to yourself: mechanisms of
938 DNA methylation maintenance in mammals. *Nucleic Acids Res.* **49**, 3020–3032 (2021).
- 939 60. Aran, D., Toperoff, G., Rosenberg, M. & Hellman, A. Replication timing-related and gene body-
940 specific methylation of active human genes. *Hum. Mol. Genet.* **20**, 670–680 (2011).
- 941 61. Mzhui, K. *et al.* Genetic loci and metabolic states associated with murine epigenetic aging.
942 *Elife* **11**, 1–30 (2022).
- 943 62. Horvath, S. & Raj, K. DNA methylation-based biomarkers and the epigenetic clock theory of
944 ageing. *Nat. Rev. Genet.* **19**, 371–384 (2018).
- 945 63. Vershinina, O., Bacalini, M. G., Zaikin, A., Franceschi, C. & Ivanchenko, M. Disentangling age -
946 dependent DNA methylation : deterministic , stochastic , and nonlinear. *Sci. Rep.* 1–12 (2021)
947 doi:10.1038/s41598-021-88504-0.
- 948 64. Cuomo, A. S. E., Nathan, A., Raychaudhuri, S., MacArthur, D. G. & Powell, J. E. Single-cell
949 genomics meets human genetics. *Nat. Rev. Genet.* (2023) doi:10.1038/s41576-023-00599-5.
- 950 65. Zhang, Q. *et al.* Improved precision of epigenetic clock estimates across tissues and its
951 implication for biological ageing. *Genome Med.* **11**, 54 (2019).
- 952 66. Tomusiak, A. *et al.* Development of a novel epigenetic clock resistant to changes in immune
953 cell composition. *bioRxiv [Preprint]* (2023) doi:https://doi.org/10.1101/2023.03.01.530561.
- 954 67. Dabrowski, J. K. *et al.* Probabilistic inference of epigenetic age acceleration from cellular
955 dynamics. *bioRxiv [Preprint]* 2023.03.01.530570 (2023).
- 956 68. Simpson, D. J., Olova, N. N. & Chandra, T. Cellular reprogramming and epigenetic

- 957 rejuvenation. *Clin. Epigenetics* **13**, 170 (2021).
- 958 69. Porter, H. L. *et al.* Many chronological aging clocks can be found throughout the epigenome:
959 Implications for quantifying biological aging. *Aging Cell* **20**, 1–13 (2021).
- 960 70. Herman, W. S. & Tatar, M. Juvenile hormone regulation of longevity in the migratory monarch
961 butterfly. *Proc. R. Soc. London. Ser. B Biol. Sci.* **268**, 2509–2514 (2001).
- 962 71. Bujarrabal-Dueso, A. *et al.* The DREAM complex functions as conserved master regulator of
963 somatic DNA-repair capacities. *Nat. Struct. Mol. Biol.* **30**, 475–488 (2023).
- 964 72. Labbadia, J. & Morimoto, R. I. Repression of the Heat Shock Response Is a Programmed Event
965 at the Onset of Reproduction. *Mol. Cell* **59**, 639–50 (2015).
- 966 73. Kerepesi, C., Zhang, B., Lee, S.-G., Trapp, A. & Gladyshev, V. N. Epigenetic clocks reveal a
967 rejuvenation event during embryogenesis followed by aging. *Sci. Adv.* **7**, 1–12 (2021).
- 968 74. Belikov, A. V. Age-related diseases as vicious cycles. *Ageing Res. Rev.* **49**, 11–26 (2019).
- 969 75. Harris, C. R. *et al.* Array programming with {NumPy}. *Nature* **585**, 357–362 (2020).
- 970 76. Varoquaux, G. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* (2011)
971 doi:10.1145/2786984.2786995.
- 972 77. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat.*
973 *Methods* **17**, 261–272 (2020).
- 974 78. Horvath, S. *et al.* Aging effects on DNA methylation modules in human brain and blood tissue.
975 *Genome Biol.* **13**, R97 (2012).
- 976 79. Laird, C. D. *et al.* Hairpin-bisulfite PCR: Assessing epigenetic methylation patterns on
977 complementary strands of individual DNA molecules. *Proc. Natl. Acad. Sci.* **101**, 204–209
978 (2004).
- 979 80. Hannum, G. *et al.* Genome-wide Methylation Profiles Reveal Quantitative Views of Human
980 Aging Rates. *Mol. Cell* **49**, 359–367 (2013).
- 981 81. Teschendorff, A. E., Breeze, C. E., Zheng, S. C. & Beck, S. A comparison of reference-based
982 algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies.
983 *BMC Bioinformatics* **18**, 105 (2017).
- 984 82. Jones, M. J., Islam, S. A., Edgar, R. D. & Kobor, M. S. Adjusting for Cell Type Composition in
985 DNA Methylation Data Using a Regression-Based Approach BT - Population Epigenetics:
986 Methods and Protocols. in (eds. Haggarty, P. & Harrison, K.) 99–106 (Springer New York,
987 2017). doi:10.1007/7651_2015_262.
- 988 83. Gillespie algorithm. [https://github.com/karinsasaki/gillespie-algorithm-
989 python/blob/master/build_your_own_gillespie_solutions.ipynb](https://github.com/karinsasaki/gillespie-algorithm-python/blob/master/build_your_own_gillespie_solutions.ipynb).
- 990 84. Chen, S., Zhou, Y., Chen, Y. & Gu, J. Fastp: An ultra-fast all-in-one FASTQ preprocessor.
991 *Bioinformatics* **34**, i884–i890 (2018).
- 992 85. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-
993 aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
- 994 86. Davis, P. *et al.* WormBase in 2022—data, processes, and tools for analyzing *Caenorhabditis*
995 *elegans*. *Genetics* **220**, (2022).
- 996 87. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: Transcript-level
997 estimates improve gene-level inferences [version 2; referees: 2 approved]. *F1000Research* **4**,

998 1–22 (2016).

999 88. Waskom, M. L. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).

1000 89. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. \& Eng.* **9**, 90–95 (2007).

1001 90. Vallat, R. Pingouin: statistics in Python. *J. Open Source Softw.* **3**, 1026 (2018).

1002 91. Tsaprouni, L. G. *et al.* Cigarette smoking reduces DNA methylation levels at multiple genomic
1003 loci but the effect is partially reversible upon cessation. *Epigenetics* **9**, 1382–1396 (2014).

1004

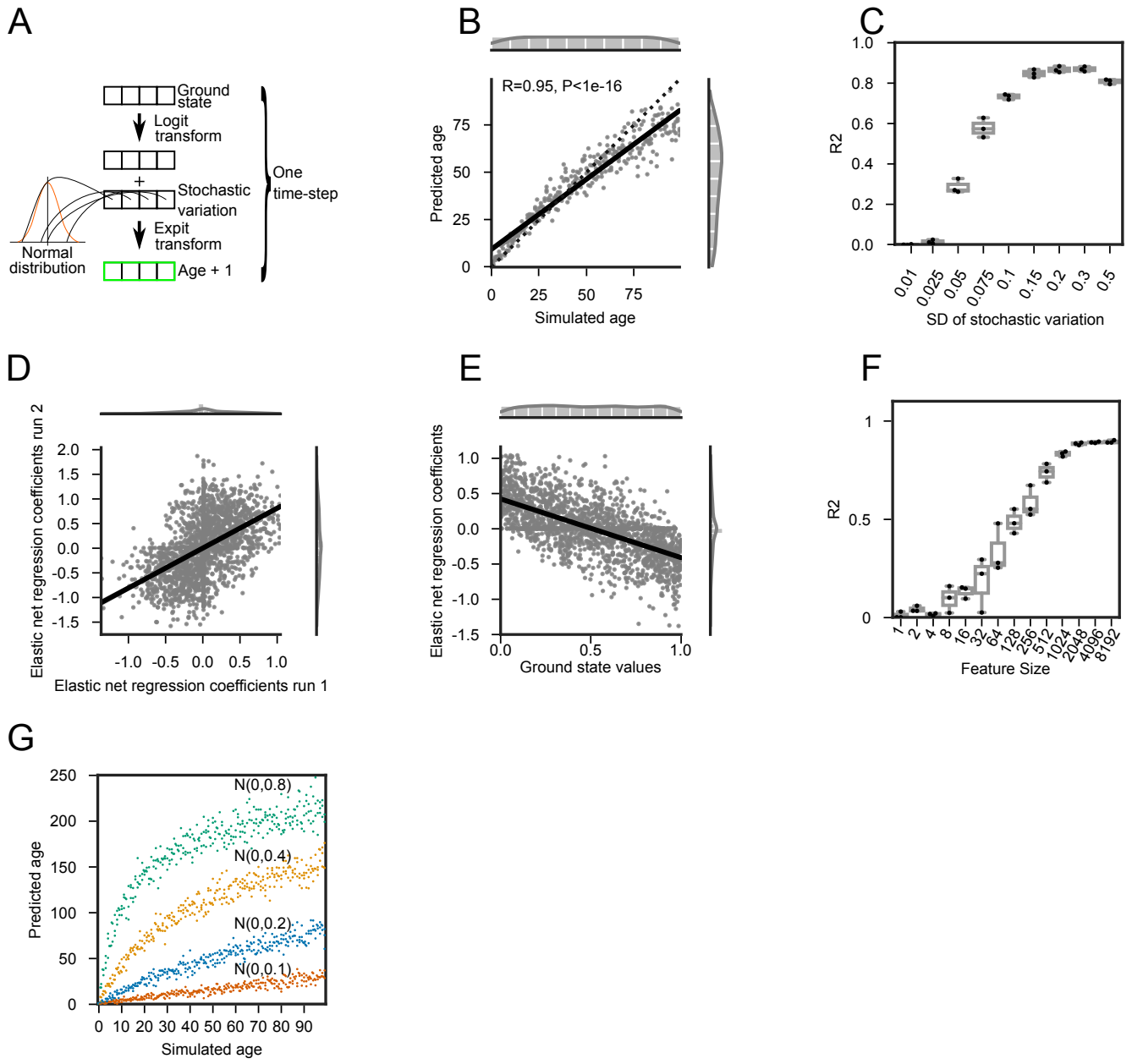


Figure 1

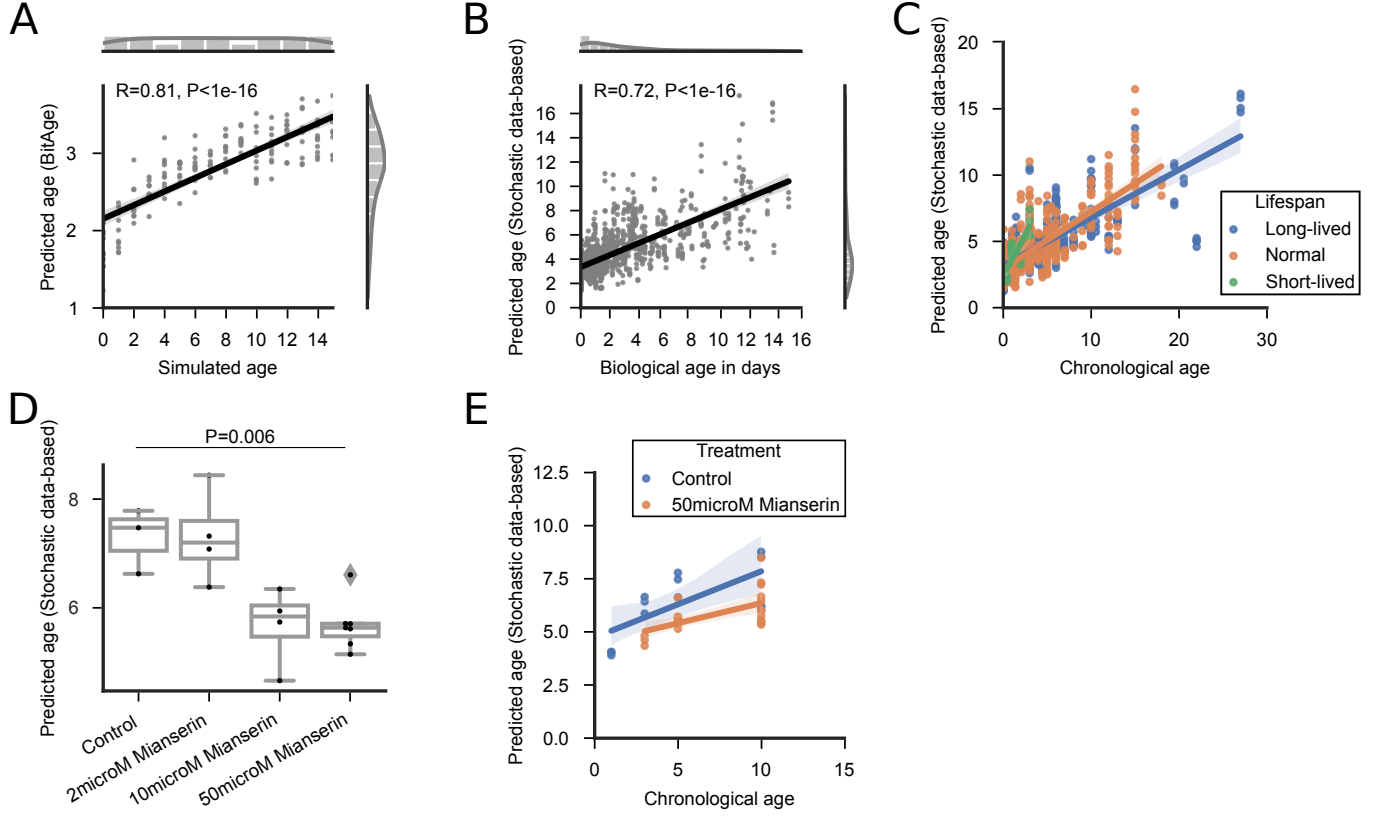


Figure 2

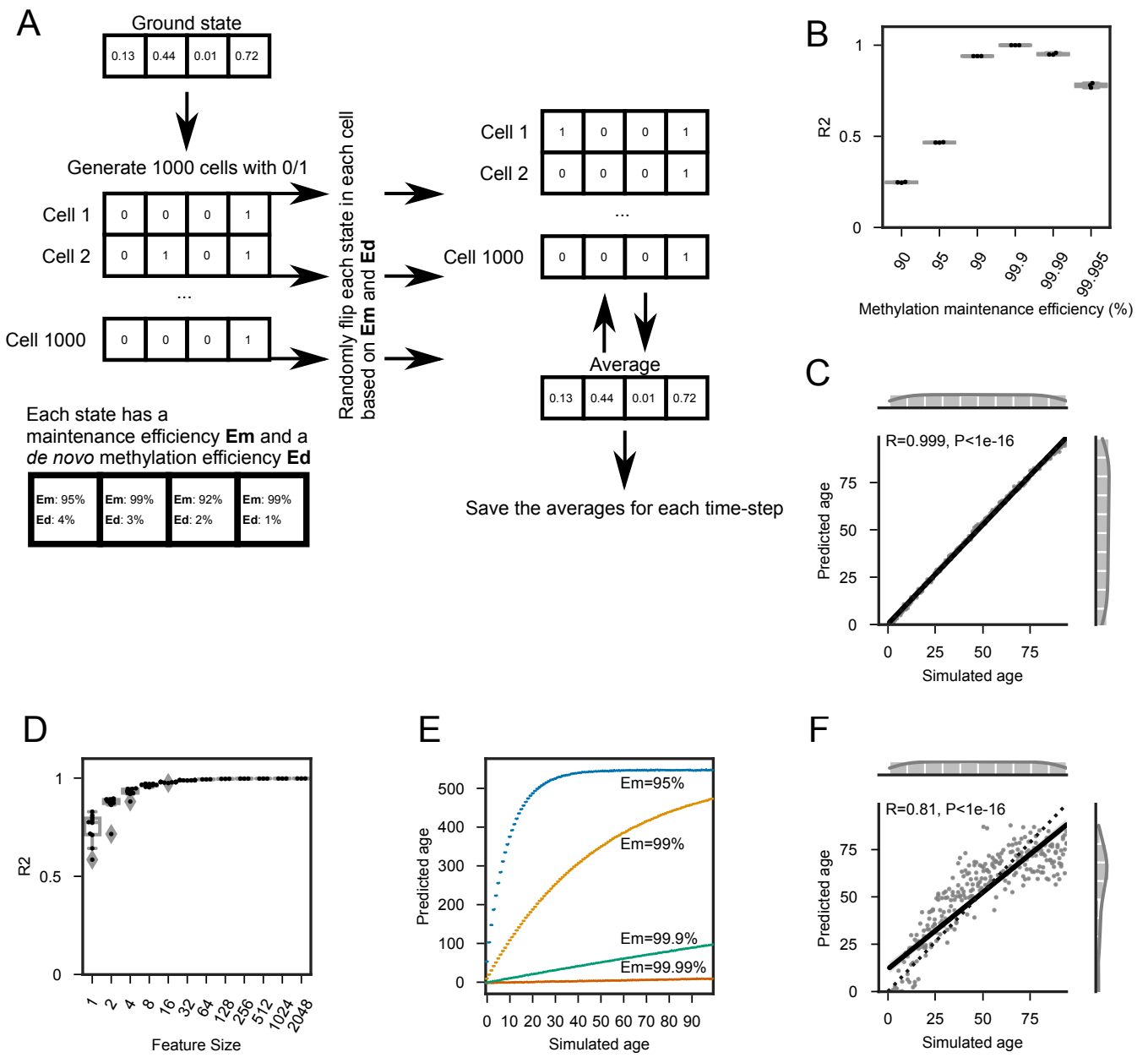


Figure 3

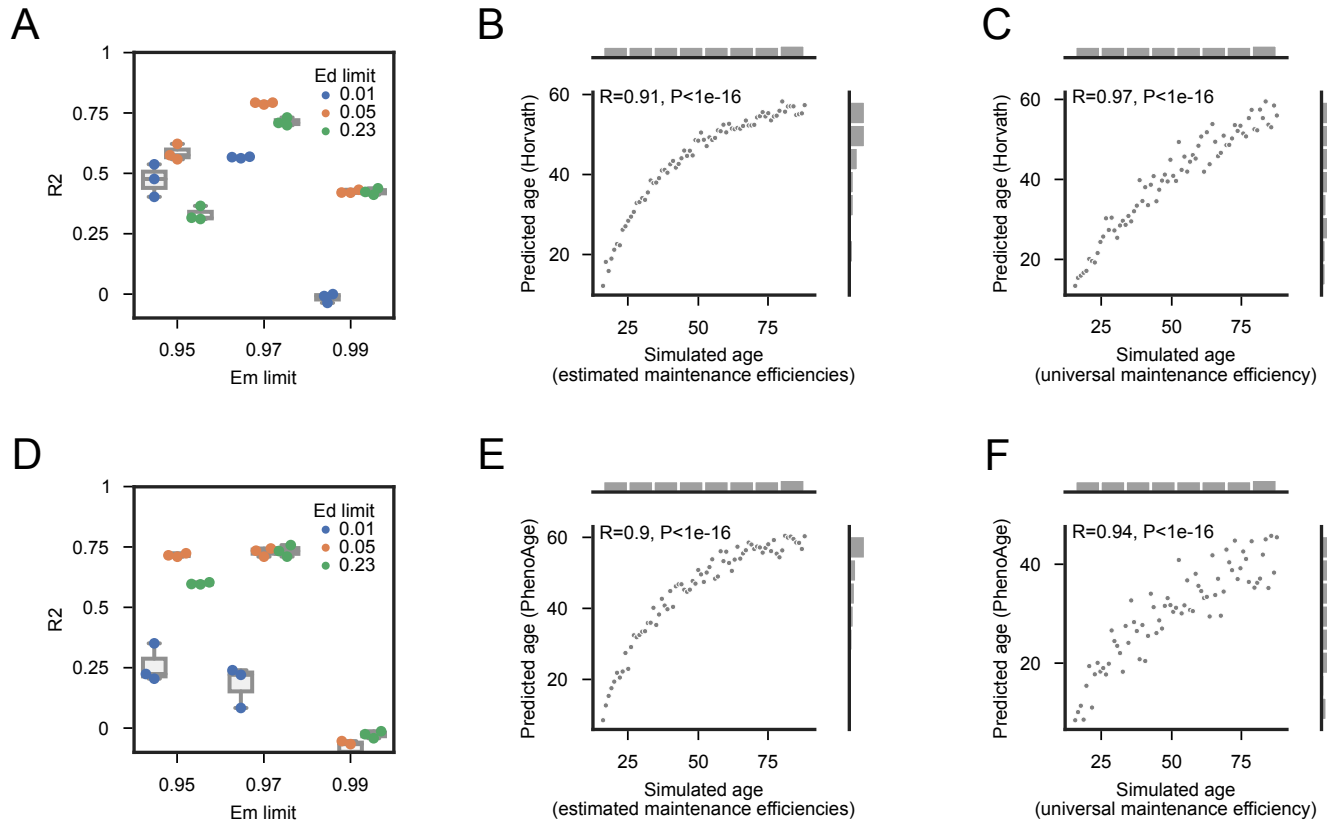


Figure 4

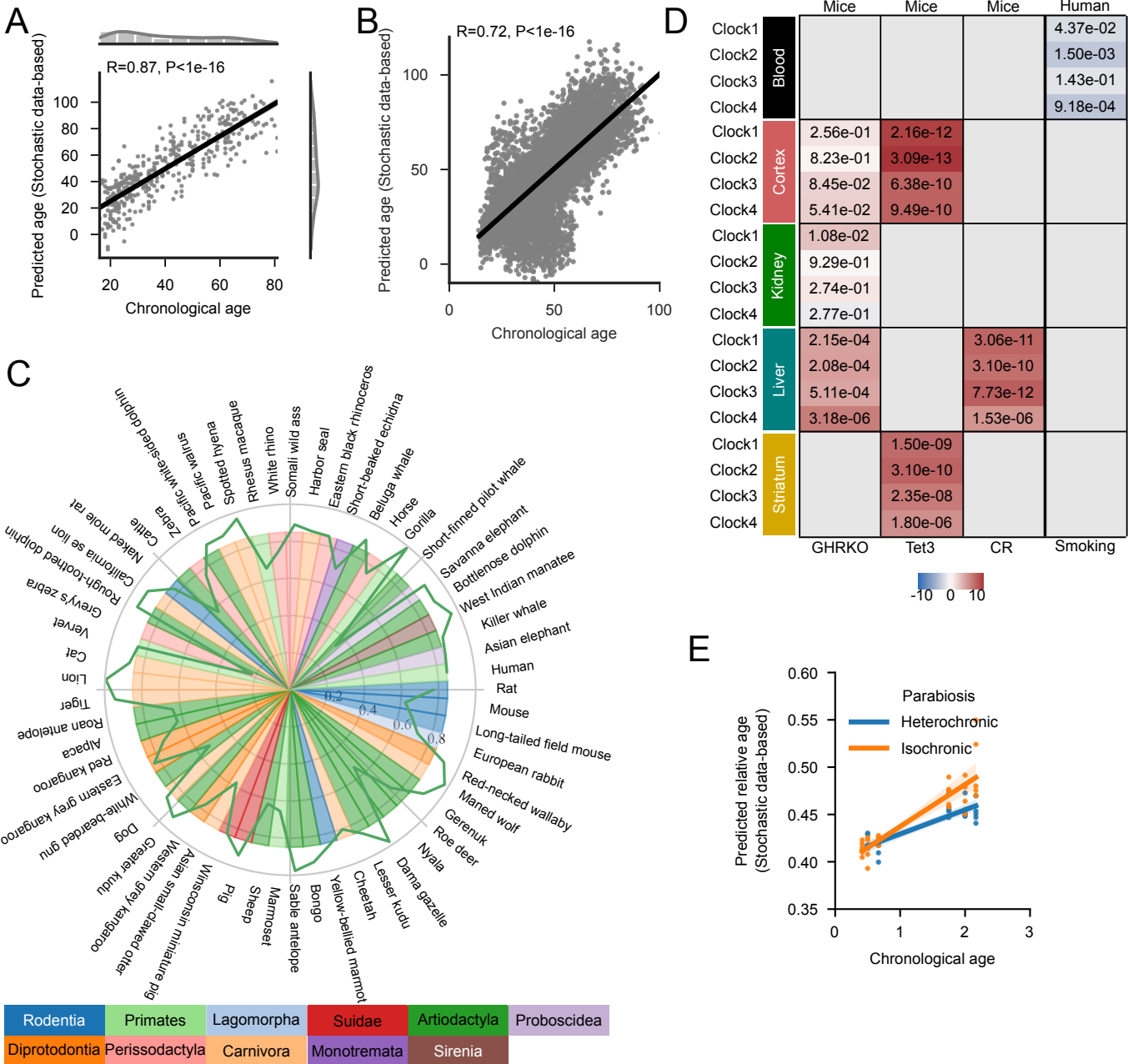


Figure 5

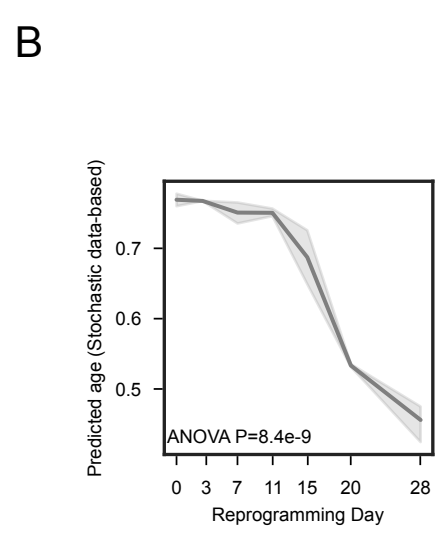
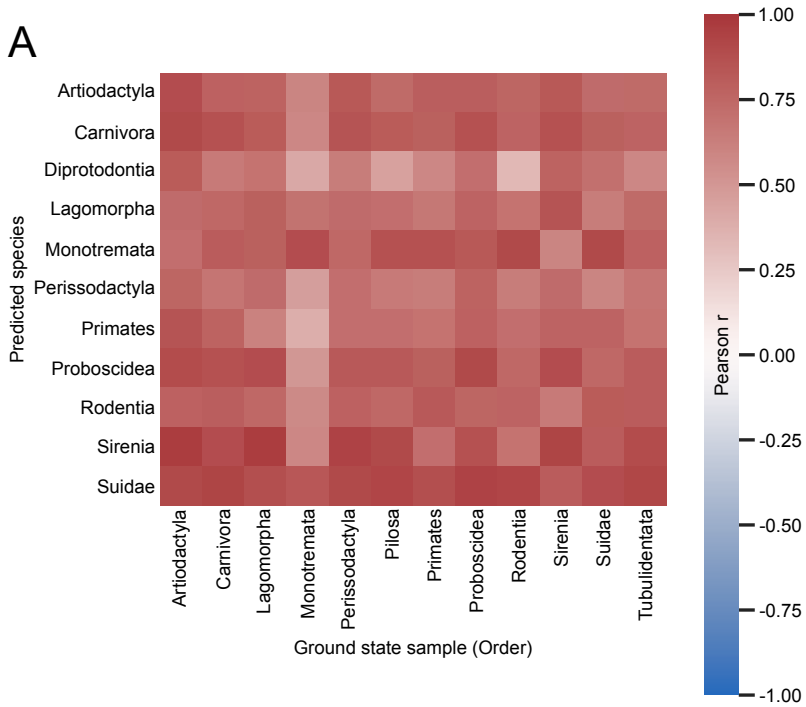


Figure 6

Extended Data Figure Legends

Extended Data Figure 1 – Normal-distributed stochastic variation accumulation simulations with value limits enable aging clock construction for simulated data

- A) Sample generation explanation. One time-step is defined as the addition of one-time stochastic variation, i.e. random noise, to each feature of the ground state that is sampled from a normal distribution centered at 0 (Top). Samples with different simulated ages are generated starting from the same ground state, but independently from each other (Bottom). A sample of age 1 adds normal-distributed stochastic variation once to the ground state, a sample of age 2 twice independently, and so on.
- B) Model training and validation explanation. For training and validation 3 sets of independent samples are generated from the same ground state as explained in Extended Data Figure 1A. 3 sets comprising the whole age-range, e.g. 1-100, are used as an input for an Elastic net regression to train a predictor that predicts the simulated age of a sample, i.e. how often stochastic variation was added to the ground state. The 3 independent datasets are used to validate the model and assess the accuracy.
- C) Unlimited stochastic variation does not allow for any prediction. All samples within the training and validation dataset started from the same ground state of 2000 uniformly randomly sampled features between 0 and 1. For every whole simulated age step from 1 to 100, normal-distributed stochastic variation sampled from $N(\mu = 0, \sigma^2 = 0.05^2)$ was added. $n=300$ samples (3 independent samples per age step) were used for training of the Elastic net regression model to predict the simulated age, and $n=300$ independent samples were used for validation. The x-axis shows the true simulated age, i.e. the number of times random stochastic variation was added to the ground state. The y-axis shows the prediction of the Elastic net regression model of the independent validation data ($n=300$, 3 independent samples per time point). The sides show the distribution of the samples.
- D) Same as C), but after addition of stochastic variation the values were kept within the range of 0-1, e.g. values bigger to 1 were set to 1 ($n=300$, 3 independent samples per time point). Limiting the values after stochastic variation application allows to build highly accurate predictors of the simulated age.
- E) The predictions of the independent validation data are robust to the stochastic variation distribution. The samples were simulated the same as in D) with different stochastic variation distributions ($n=300$, 3 independent samples per time point). The x-axis shows the standard deviation of the normal distribution from which the stochastic variation was sampled, i.e.

$N(\mu = 0, \sigma^2 = 0.005^2)$ has a narrow noise distribution with 99.7 % of the sampled data within the range $[-0.015, 0.015]$, while $N(\mu = 0, \sigma^2 = 0.01^2)$ has a wide distribution with 99.7 % of the sampled data within the range $[-0.3, 0.3]$. The y-axis shows the R^2 value between the simulated age and the predicted age of the independent validation data (N=3 independent repeats; each with n=300, 3 samples per time point). Boxplots are shown with the center line depicting the median, the box limits the bottom, respective top quartiles, and the whiskers the 1.5x interquartile range.

- F) Independent Elastic net regression models are highly correlated if trained on samples starting from the same ground state (consisting of N=2000 uniformly randomly sampled features between 0 and 1). The x-axis shows the coefficients of the Elastic net regression of D), and the y-axis shows the coefficients of an independent Elastic net regression on samples that started with the same ground state, but with independent stochastic variation application (trained on n=300, 3 samples per time point).
- G) The prediction in D) is possible due to a regression to the mean. The x-axis shows the starting values of the 2000 features of the simulated ground state, the y-axis the Elastic net regression coefficients for the model in D) (trained on n=300, 3 samples per time point). Features starting close to 0 have a positive coefficient, indicating an increase over the simulated time period, while features close to 1 have a negative coefficient, indicating a decrease. Features close to 0.5 are more sensitive to random changes and are closer to 0.
- H) The accuracy of predictions caps off after ~1000 features in the ground state. The x-axis shows how many uniformly randomly features were sampled for the ground state that was used to build and validate an Elastic net regression model the same as in D) (trained on n=300, 3 samples per time point). The y-axis shows the R^2 as a measure of model accuracy. Of note, the Elastic net regression will shrink coefficients of features to 0 and thereby reduce the features relevant for the prediction further. (N=10 independent repeats for Features Sizes<1000, N=3 independent repeats otherwise; each with n=300, 3 samples per time point). Boxplots are shown with the center line depicting the median, the box limits the bottom, respective top quartiles, and the whiskers the 1.5x interquartile range.
- I) The amount of stochastic variation sets the pace of aging. The Elastic net regression model was trained the same as in D) with stochastic variation sampled from $N(\mu = 0, \sigma^2 = 0.05^2)$ (n=300, 3 samples per time point). Color-coded are different independent validation samples, generated from the same ground state, but with stochastic variation from different normal distributions. Samples with stochastic variation from a distribution with a narrower standard deviation ($N(\mu = 0, \sigma^2 = 0.025^2)$) accumulate less noise and are predicted to age slower, i.e. the slope of the prediction is lower. Samples with stochastic variation from a distribution with

a wider standard deviation ($N(\mu = 0, \sigma^2 = 0.1^2)$, $N(\mu = 0, \sigma^2 = 0.2^2)$) accumulate noise faster, have a steeper slope of prediction, and reach the maximum age faster. The x-axis shows the true simulated age, i.e. the number of times stochastic variation was added to the ground state. The y-axis shows the prediction of the Elastic net regression model of the independent validation data. All 4 simulated datasets consist of $n=300$, 3 samples per time point.

Extended Data Figure 2 – The effect of the feature size and the amount of stochastic variation on transcriptomic stochastic variation accumulation simulations

- A) The BitAge predictions in Figure 2A are robust to the distribution from which the stochastic variation is sampled. The x-axis shows the standard deviation of the normal distribution (centered at 0) from which stochastic variation for the simulations is sampled. The y-axis shows the Pearson correlation between the BitAge prediction of the simulated samples and the number of stochastic variation additions of the samples. Stochastic variation sampled from a normal distribution centered at 0 and a standard variation of 0.01 shows the highest Pearson correlation. $N=5$ independent experiments are shown. Boxplots are shown with the center line depicting the median, the box limits the bottom, respective top quartiles, and the whiskers the 1.5x interquartile range.
- B) The feature size is largely irrelevant for the model in Figure 2B). Predictions of Elastic net regression models trained on more than 100 features are significantly correlated with the biological age of *C. elegans* samples. The x-axis shows the number of randomly selected features, i.e. genes, for the ground state, which were subsequently used to generate data based on stochastic variations (see methods for details). These simulated samples were used to train the Elastic net regression. The y-axis shows the Pearson correlation between the biological age of the 993 independent samples (excluding the sample from which the ground state was sampled) and the prediction of the independent stochastic-data based model. $N=10$ independent experiments are shown. Boxplots are shown with the center line depicting the median, the box limits the bottom, respective top quartiles, and the whiskers the 1.5x interquartile range.
- C) Verification of Extended Data Figure 2B). Using the same approach as in Extended Data Figure 2B, but with randomly shuffled biological ages of the *C. elegans* samples shows no significant correlation, indicating that biological age, and not a confounding variable is correlated with the predictions of the model based on simulated data. The x-axis shows the number of randomly selected features, i.e. genes, for the ground state, which were subsequently used to generate data based on stochastic variations (see methods for details). These simulated samples were used to train the Elastic net regression. The y-axis

shows the Pearson correlation between the biological age of the 993 independent samples (excluding the sample from which the ground state was sampled) and the prediction of the stochastic-data based model. N=10 independent experiments are shown. Boxplots are shown with the center line depicting the median, the box limits the bottom, respective top quartiles, and the whiskers the 1.5x interquartile range.

Extended Data Figure 3 – DNA methylation stochastic variation accumulation simulations

- A) Comparison between the ground state on the x-axis, and the ground state (N=2000 uniformly randomly sampled features between 0 and 1) after applying stochastic variation from $N(\mu = 0, \sigma^2 = 0.05^2)$, i.e. Gaussian noise, once on the y-axis.
- B) Comparison between the ground state on the x-axis, and the ground state (N=2000 uniformly randomly sampled features between 0 and 1) after applying stochastic variation from $N(\mu = 0, \sigma^2 = 0.05^2)$, i.e. Gaussian noise, 100 times on the y-axis.
- C) Comparison of human blood DNA methylation data of the youngest (x-axis= GSM1007467) and oldest (y-axis= GSM1007832) subjects in the public dataset GSE41037⁷⁸. Every dot depicts a DNA methylation site (n=21389). Values close to 0 and 1 show less variation than values closer to 0.5.
- D) Comparison of the ground state on the x-axis (2000 randomly sampled features from the youngest healthy sample (GSM1007467⁷⁸)) and the ground state after applying 100x single cell stochastic variation steps with a universal maintenance efficiency rate of 99.9 %, i.e. the maintenance efficiency rate is fixed to be the same for all features (y-axis).
- E) Starting single-cell simulations with a ground state consisting of 2000 features at 0.5 with a universal maintenance of 99 % allows no prediction. An Elastic net regression model was trained on n=300 samples (3 samples per time point) starting from the same ground state in which all features were set to 0.5, and universal maintenance efficiencies E_m and E_u of 99 %. The x-axis shows the true simulated age, i.e. the number of times stochastic variation was added to the ground state. The y-axis shows the prediction of the Elastic net regression model of the independent validation data (n=300, 3 samples per time point). The sides show the distribution of the samples.
- F) Starting single-cell simulations with a ground state consisting of 2000 features at 0.51 with a universal maintenance of 99 % allows for an accurate age prediction. The training and validation were done the same as in B) with the difference that all features in the ground state started at 0.51. (n=300, 3 samples per time point).
- G) Starting single-cell simulations with a ground state consisting of 2000 features at 0.5 with biologically estimated maintenance rates allows for an accurate prediction. The training and

validation were done the same as in B) with the difference that E_m and E_u values were estimated from biological data (see methods for details). (n=300, 3 samples per time point).

- H) Comparison of the ground state on the x-axis (2000 randomly sampled features from the youngest healthy sample (GSM1007467⁷⁸)) and the ground state after applying 100x single cell stochastic variation steps (y-axis) with empirically estimated maintenance efficiency rates with the limits $E_m > 95\%$ and $E_d < 23\%$.
- I) The prediction in Figure 3F) is not due to a regression to the mean, different to Figure 1. The x-axis shows the starting values of the 2000 randomly sampled features from the youngest healthy sample (GSM1007467⁷⁸) as the ground state, the y-axis the Elastic net regression coefficients for the model in Figure 3F) (n=300, 3 samples per time point). All ground state features can have positive as well as negative coefficients, indicating that the prediction is not based on a regression to the mean.

Extended Data Figure 4 – Epigenetic aging clock predictions correlate robustly with the amount of stochastic variation

- A) Horvath's epigenetic age prediction²⁶ of samples simulated based on biologically estimated maintenance rates with the limits $E_m > 95\%$ and $E_d < 23\%$ starting from biological data from a young human blood sample (GSM1007467)⁷⁸, correlates significantly with the simulated age, i.e. how often stochastic variation was applied to the ground state. N=73 independent samples, one per age step from 16 to 88 are shown.
- B) Horvath's epigenetic age prediction²⁶ of samples simulated based on random maintenance rates within the limits $97\% < E_m \leq 100\%$ and $0\% \leq E_d < 5\%$ starting from biological data from a young human blood sample (GSM1007467)⁷⁸, correlates significantly with the simulated age, i.e. how often stochastic variation was applied to the ground state. The y-axis shows the Pearson correlation between the simulated age and Horvath's age prediction. N=30 independent experiments with each n=73 independent samples. Boxplots are shown with the center line depicting the median, the box limits the bottom, respective top quartiles, and the whiskers the 1.5x interquartile range.
- C) Pearson correlation of Horvath's epigenetic age prediction²⁶ of simulated data and the true simulated age for different universal methylation maintenance efficiencies. 5 independent experiments (each containing n=73 independent samples, one per age step from 16 to 88) with different ground states are shown for each maintenance efficiency. Boxplots are shown with the center line depicting the median, the box limits the bottom, respective top quartiles, and the whiskers the 1.5x interquartile range.
- D) Biological age prediction with PhenoAge⁴⁰ of samples simulated based on biologically estimated maintenance rates with the limits $E_m > 95\%$ and $E_d < 23\%$ starting from biological

data from a young human blood sample (GSM1007467)⁷⁸, correlates significantly with the simulated age, i.e. how often stochastic variation was applied to the ground state. N=73 independent samples, one per age step from 16 to 88 are shown.

- E) Biological age prediction with PhenoAge⁴⁰ of samples simulated based on random maintenance rates within the limits $97\% < E_m \leq 100\%$ and $0\% \leq E_d < 5\%$ starting from biological data from a young human blood sample (GSM1007467)⁷⁸, correlates significantly with the simulated age, i.e. how often stochastic variation was applied to the ground state. The y-axis shows the Pearson correlation between the simulated age and PhenoAge's age prediction. N=30 independent experiments with each n=73 independent samples. The boxplot is shown with the center line depicting the median, the box limits the bottom, respective top quartiles, and the whiskers the 1.5x interquartile range.
- F) Pearson correlation of biological age predictions with PhenoAge⁴⁰ of simulated data and the true simulated age for different universal methylation maintenance efficiencies. 5 independent experiments (each containing n=73 independent samples, one per age step from 16 to 88) with different ground states are shown for each maintenance efficiency. Boxplots are shown with the center line depicting the median, the box limits the bottom, respective top quartiles, and the whiskers the 1.5x interquartile range.
- G) Horvath's epigenetic age prediction²⁶ of samples simulated based on biologically estimated maintenance rates with the limits $E_m > 97\%$ and $E_d < 5\%$ starting from biological data from a young human blood sample age 16 (GSM1007467)⁷⁸, correlates significantly with the simulated age, i.e. how often stochastic variation was applied to the ground state. The simulation is the same as in Extended Data Figure 4A, but with a simulated age range from 0-99 for an easier comparison with Extended Data Figure 4H,I. N=100 independent samples, one per age step from 0 to 99 are shown.
- H) Horvath's epigenetic age prediction²⁶ of samples simulated based on biologically estimated maintenance rates with the limits $E_m > 97\%$ and $E_d < 5\%$ starting from biological data from a middle-aged human blood sample age 37 (GSM1007384)⁷⁸, still correlates significantly with the simulated age, i.e. how often stochastic variation was applied to the ground state. The predicted age starts at a later time-point than the predictions in Extended Data Figure 4G, and reaches the cap-off earlier. N=100 independent samples, one per age step from 0 to 99 are shown.
- I) Horvath's epigenetic age prediction²⁶ of samples simulated based on biologically estimated maintenance rates with the limits $E_m > 97\%$ and $E_d < 5\%$ starting from biological data from an old human blood sample age 81 (GSM1007791)⁷⁸, does not correlate significantly with the simulated age, i.e. how often stochastic variation was applied to the ground state. Starting the

ground state at an old age does not allow for a correlation between the predicted epigenetic age and the amount of stochastic variation in the data, since the prediction already starts in the cap-off. N=100 independent samples, one per age step from 0 to 99 are shown.

Extended Data Figure 5 – All tested epigenetic clock predictions correlate significantly with the amount of stochastic variation

- A) Vidal-Bralo's epigenetic age prediction⁴¹ of samples simulated based on biologically estimated maintenance rates with the limits $E_m > 97\%$ and $E_d < 5\%$ starting from biological data from a young human blood sample (GSM1007467)⁷⁸, correlates significantly with the simulated age, i.e. how often stochastic variation was applied to the ground state. N=73 independent samples, one per age step from 16 to 88 are shown.
- B) Vidal-Bralo's epigenetic age prediction⁴¹ of samples simulated based on a universal maintenance rate of 99% for all features (CpG sites) starting from biological data from a young human blood sample (GSM1007467)⁷⁸, correlates significantly with the simulated age, i.e. how often stochastic variation was applied to the ground state. N=73 independent samples, one per age step from 16 to 88 are shown.
- C) Lin's epigenetic age prediction⁴² of samples simulated based on biologically estimated maintenance rates with the limits $E_m > 97\%$ and $E_d < 5\%$ starting from biological data from a young human blood sample (GSM1007467)⁷⁸, correlates significantly with the simulated age, i.e. how often stochastic variation was applied to the ground state. N=73 independent samples, one per age step from 16 to 88 are shown.
- D) Lin's epigenetic age prediction⁴² of samples simulated based on a universal maintenance rate of 99% for all sites starting from biological data from a young human blood sample (GSM1007467)⁷⁸, correlates significantly with the simulated age, i.e. how often stochastic variation was applied to the ground state. N=73 independent samples, one per age step from 16 to 88 are shown.
- E) Weidner's epigenetic age prediction⁴³ of samples simulated based on biologically estimated maintenance rates with the limits $E_m > 97\%$ and $E_d < 5\%$ starting from biological data from a young human blood sample (GSM1007467)⁷⁸, correlates significantly with the simulated age, i.e. how often stochastic variation was applied to the ground state. N=73 independent samples, one per age step from 16 to 88 are shown.
- F) Weidner's epigenetic age prediction⁴³ of samples simulated based on a universal maintenance rate of 99% for all sites starting from biological data from a young human blood sample (GSM1007467)⁷⁸, correlates significantly with the simulated age, i.e. how often stochastic variation was applied to the ground state. N=73 independent samples, one per age step from 16 to 88 are shown.

- G) GrimAge's epigenetic age prediction⁴⁴ of samples simulated based on biologically estimated maintenance rates with the limits $E_m > 97\%$ and $E_d < 5\%$ starting from biological data from a young human blood sample generated with the 450k Human Methylation Beadchip (GSM990528)⁸⁰, correlates significantly with the simulated age, i.e. how often stochastic variation was applied to the ground state. N=20 independent samples are shown.
- H) GrimAge's epigenetic age prediction⁴⁴ of samples simulated based on a universal maintenance rate of 99% for all sites starting from biological data from a young human blood sample generated with the 450k Human Methylation Beadchip (GSM990528)⁸⁰, correlates significantly with the simulated age, i.e. how often stochastic variation was applied to the ground state. N=20 independent samples are shown.
- I) Horvath's epigenetic age prediction²⁶ of samples simulated with Gillespie's algorithm with a universal maintenance efficiency rate of 90% for all features (CpG sites) starting from biological data from a young human blood sample (GSM1007467)⁷⁸, correlates significantly with the simulated age, i.e. how often stochastic variation was applied to the ground state. Since the ground state was starting from a sample of a 16-year-old human, we set the starting point of the simulated age to 16. The time-steps in Gillespie's algorithm are not fixed, in total N=15999 simulations were computed.

Extended Data Figure 6 – Human stochastic data-based clock predictions correlate significantly with the chronological age

- A) The predictions of an Elastic net regression model based on simulated data, correlates significantly (Pearson correlation 0.87, p-value<1e-16, two-sided test) with the chronological age of the independent healthy biological validation samples (GSE41037, n=392)⁷⁸. The simulated data is based on biologically estimated maintenance rates starting with Horvath's epigenetic clock CpG sites from biological data from a young human blood sample. The x-axis shows the chronological age of the subjects from which blood DNA methylation data was processed. The y-axis shows the predicted simulated age, i.e. the prediction how often stochastic variation was added to the ground state and is therefore on a different scale and unit than the x-axis.
- B) The feature size is largely irrelevant for stochastic data-based models in Extended Data Figure 6A. Predictions of Elastic net regression models trained on more than 500 random CpG sites (features) are significantly correlated with the chronological age. The x-axis shows the number of randomly selected features, i.e. CpG sites, for the ground state, which were subsequently used to generate data based on stochastic variations (see methods for details). These simulated samples were used to train the Elastic net regression. The y-axis shows the Pearson correlation between the chronological age of the

n=392 healthy samples in GSE41037⁷⁸ (excluding the sample from which the ground state was sampled, and the oldest sample from which maintenance efficiencies were estimated) and the prediction of the independent stochastic-data based model. N=5 independent experiments are shown. Boxplots are shown with the center line depicting the median, the box limits the bottom, respective top quartiles, and the whiskers the 1.5x interquartile range.

- C) Verification of Extended Data Figure 6B). Using the same approach as in Extended Data Figure 6A, but with randomly shuffled chronological ages shows no significant correlation, indicating that chronological age, and not a confounding variable is correlated with the predictions of the model based on simulated data. The x-axis shows the number of randomly selected features, i.e. CpG sites, for the ground state, which were subsequently used to generate data based on stochastic variations (see methods for details). These simulated samples were used to train the Elastic net regression. The y-axis shows the Pearson correlation between the permuted chronological age of healthy samples in GSE41037⁷⁸ (excluding the sample from which the ground state was sampled, and the oldest sample from which maintenance efficiencies were estimated) and the prediction of the stochastic-data based model. N=3 independent experiments are shown. Boxplots are shown with the center line depicting the median, the box limits the bottom, respective top quartiles, and the whiskers the 1.5x interquartile range.
- D) The same analysis as in Figure 5A, but the simulated stochastic data were additionally cell-type corrected and then used to train the clock (Pearson correlation 0.81, $p < 1e-16$, two-sided test).
- E) The validation of the stochastic data-based clock in Figure 5A on 11,146 independent samples from 15 independent datasets (GSE84727, GSE87571, GSE80417, GSE40279, GSE87648, GSE42861, GSE50660, GSE106648, GSE179325, GSE210254, GSE210255, GSE72680, GSE147740, GSE55763, GSE117860) shows a highly significant correlation (Pearson correlation 0.57, $p\text{-value} < 1e-16$).

Extended Data Figure 7 – Human stochastic data-based clock predictions correlate significantly with the chronological age of independent validation data

The validation of the stochastic data-based clock starting from a fetal sample (GSM4682890) on 11,146 independent samples from 15 independent datasets A) GSE106648, B) GSE84727, C) GSE87571, D) GSE80417, E) GSE40279, F) GSE87648, G) GSE179325, H) GSE50660, I) GSE42861, J)

GSE210254, K) GSE210255, L) GSE72680, M) GSE147740, N) GSE55763, O) GSE117860. See Figure 5B for a combined plot. The Pearson correlation and its p-value, calculated with a two-sided test, are shown in the figure panels.

Extended Data Figure 8 – Horvath’s epigenetic age prediction results for the same 15 datasets

Horvath’s epigenetic age prediction on the same 11,146 samples from 15 independent datasets used in Extended Data Figure 7. A) GSE106648, B) GSE84727, C) GSE87571, D) GSE80417, E) GSE40279, F) GSE87648, G) GSE179325, H) GSE50660, I) GSE42861, J) GSE210254, K) GSE210255, L) GSE72680, M) GSE147740, N) GSE55763, O) GSE117860. Note that GSE40279 and GSE42861 were used during test and training in Horvath’s original publication. Similar to Extended Data Figure 7 GSE87648 and GSE147740 do not show any correlation between the predicted and the chronological age. The Pearson correlation and its p-value, calculated with a two-sided test, are shown in the figure panels.

Extended Data Figure 9 – Stochastic data-based clock predictions correlate significantly with the chronological and biological age of pan-mammalian data

- A) The same circle plot as in Figure 5C, but for Clock 2-4. The Pearson correlation of the relative age of all blood samples of a given species and their predicted age of the stochastic data-based clocks are shown as lines around the circle. Species are shown for which at least 5 blood samples were available. The species are clock-wise sorted by maximum lifespan, starting with *Rattus norvegicus* (3.8 years) in the center right, and ending with *Homo sapiens* (122.5 years). The colors within the circle show the taxonomic order of the corresponding species, as listed on the right side. Clock 2 (99% maintenance rate for all CpG sites used in Lu’s pan-mammalian relative age clock¹⁵), Clock 3 (CpG site-specific empirically estimated maintenance rates from the oldest sample of *Tursiops truncatus* for all 37554 CpG sites), and Clock 4 (99% maintenance rate for all 37554 CpG sites) correlate on average highly significantly.
- B) Example comparison for Figure 5D. Predictions of Clock 1 for GHRKO (n=11 biologically independent samples) vs. WT (n=12 biologically independent samples) liver samples show significantly lower values for GHRKO samples (two-sided adjusted p-value 2.15e-04, full statistics in **Source Data 1**). Boxplots are shown with the center line depicting the median, the box limits the bottom, respective top quartiles, and the whiskers the 1.5x interquartile range.
- C) Example comparison for Figure 5D. Predictions of Clock 1 for *Tet3* (n=8 biologically independent samples) vs. WT (n=44 biologically independent samples) cerebral cortex

samples show significantly lower values for *Tet3* samples (two-sided adjusted p-value 2.16e-12, full statistics in **Source Data 1**). Boxplots are shown with the center line depicting the median, the box limits the bottom, respective top quartiles, and the whiskers the 1.5x interquartile range.

- D) Example comparison for Figure 5D. Predictions of Clock 1 for calorie restricted (CR) (n=59 biologically independent samples) vs. normal fed (n=36 biologically independent samples) liver samples show significantly lower values for CR samples (two-sided adjusted p-value 3.06e-11, full statistics in **Source Data 1**). Boxplots are shown with the center line depicting the median, the box limits the bottom, respective top quartiles, and the whiskers the 1.5x interquartile range.
- E) Example comparison for Figure 5D. Current-smoker vs. ex-smoker vs. never-smoker aging trajectories are color-coded. The lines show the linear regression model fit of Seaborn's `lmplot` function⁸⁸, and the shadow around the lines the 95% confidence interval. Current-smoker show a steeper aging trajectory (slope) compared to never- or ex-smoker.
- F) The same as Figure 5E, but for Clock 2. A multivariate regression of chronological age, the parabiosis treatment, and the interaction shows a significant age variable ($p=6.11e-12$), and interaction variable ($p=1.13e-02$). The regression model fit with a 95% confidence interval (shadowed area) is shown.
- G) The same as Figure 5E, but for Clock 3. A multivariate regression of chronological age, the parabiosis treatment, and the interaction shows a significant age variable ($p=5.6e-09$). The regression model fit with a 95% confidence interval (shadowed area) is shown.
- H) The same as Figure 5E, but for Clock4. A multivariate regression of chronological age, the parabiosis treatment, and the interaction shows a significant age variable ($p=1.29e-06$). The regression model fit with a 95% confidence interval (shadowed area) is shown.

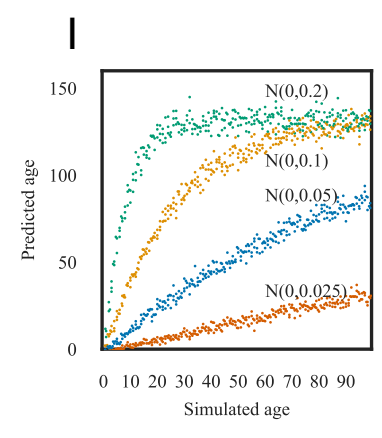
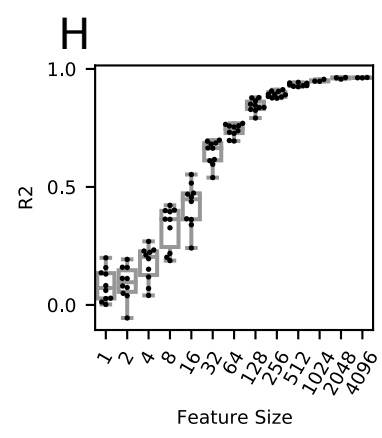
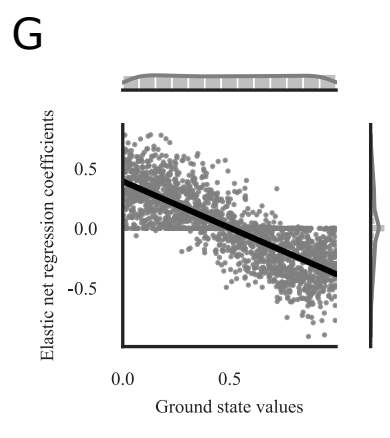
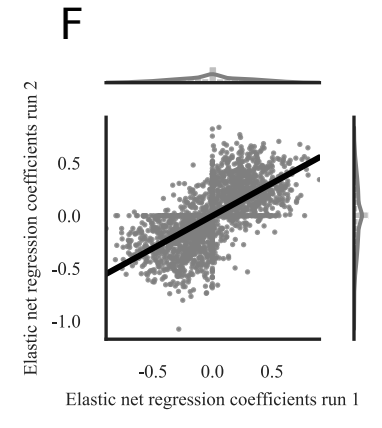
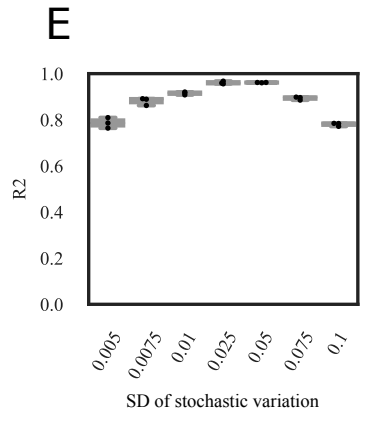
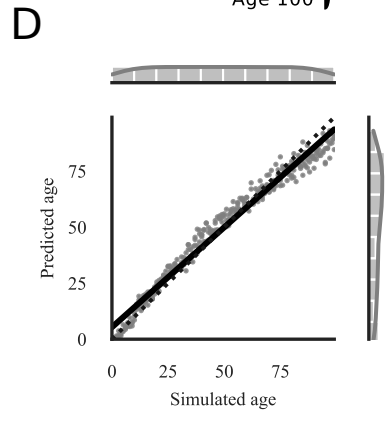
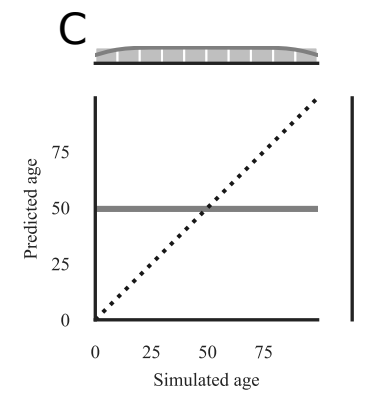
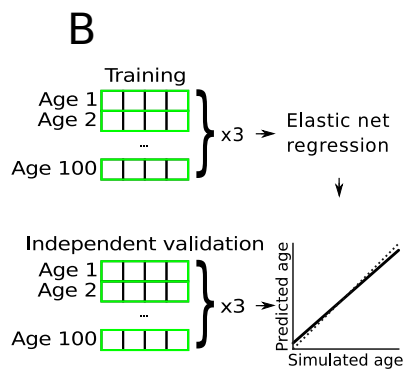
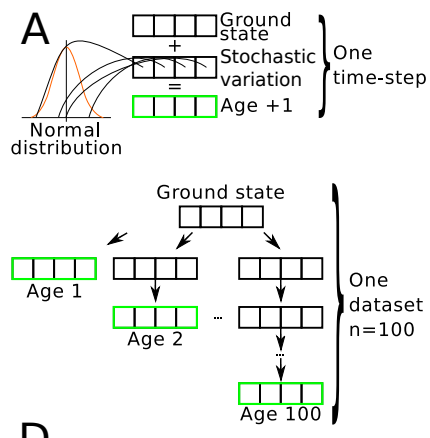
Full statistics can be found in **Source Data 1**.

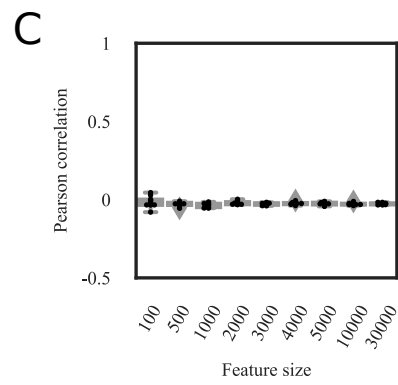
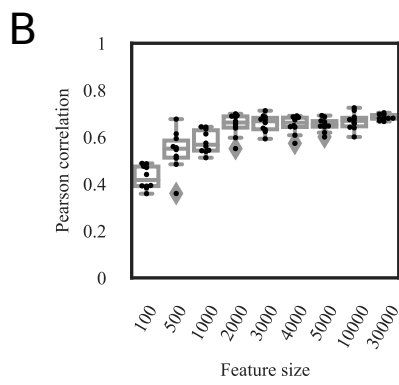
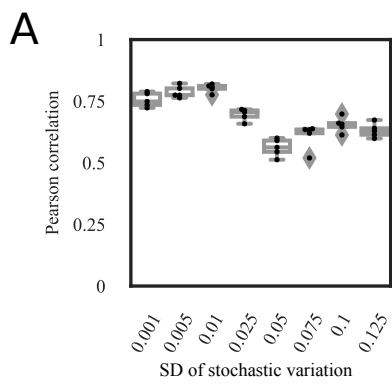
Extended Data Figure 10 – Stochastic data-based clock predictions for pan-mammalian data are robust to the choice of the ground state species

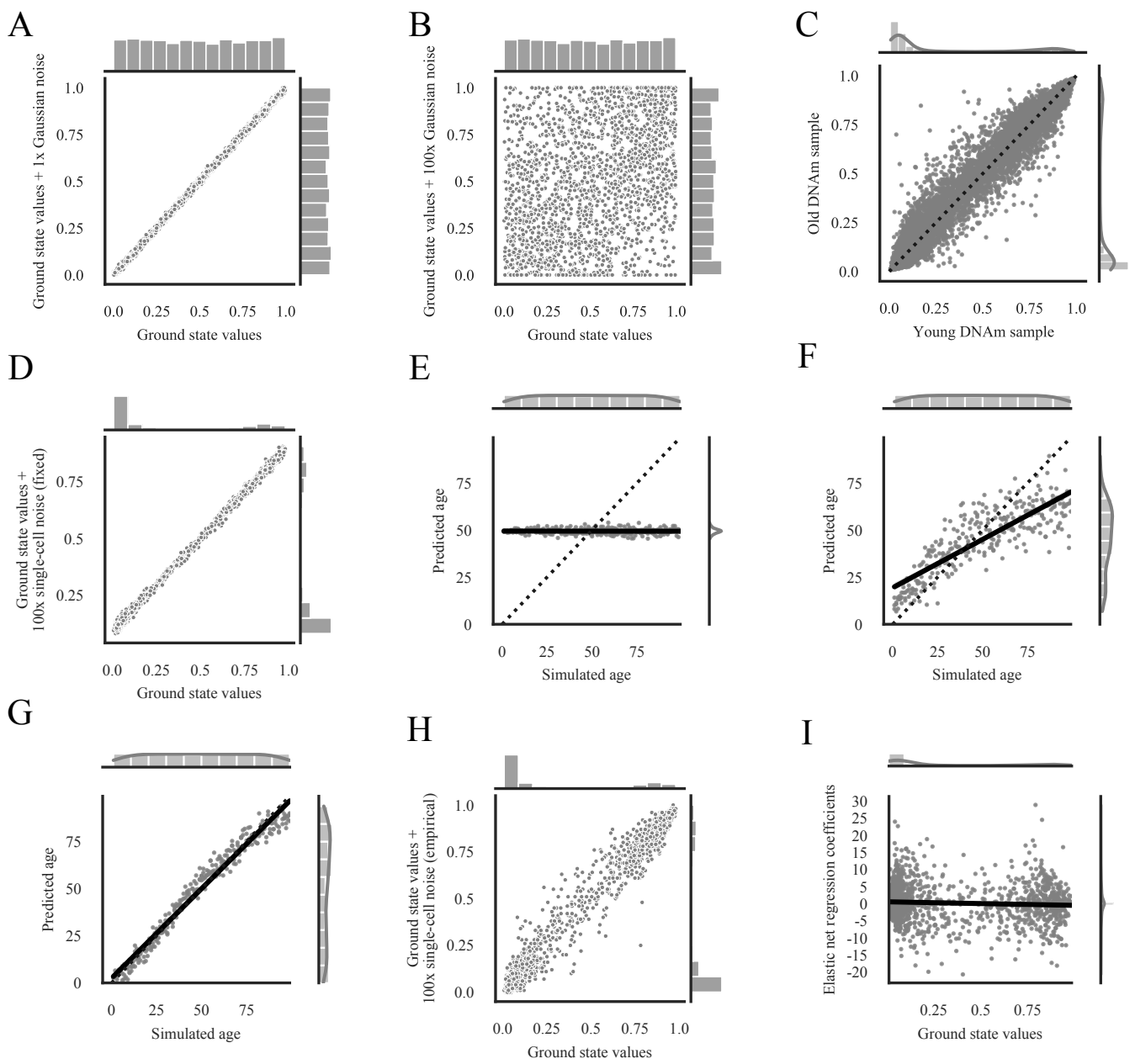
- A) Heatmap showing Pearson correlations between the predicted age of Clock 1 trained on the youngest blood sample from species of the corresponding taxonomic order in the columns (Artiodactyla: *Tursiops truncatus*, Carnivora: *Odobenus rosmarus divergens*, Lagomorpha: *Oryctolagus cuniculus*, Monotremata: *Tachyglossus aculeatus*, Perissodactyla: *Equus caballus*, Pilosa: *Choloepus hoffmanni*, Proboscidea: *Loxodonta africana*, Rodentia: *Marmota flaviventris*, Sirenia: *Trichechus manatus*, Suidae: *Sus scrofa*, Tubulidentata: *Orycteropus afer*) and the relative age for all species in the rows. The

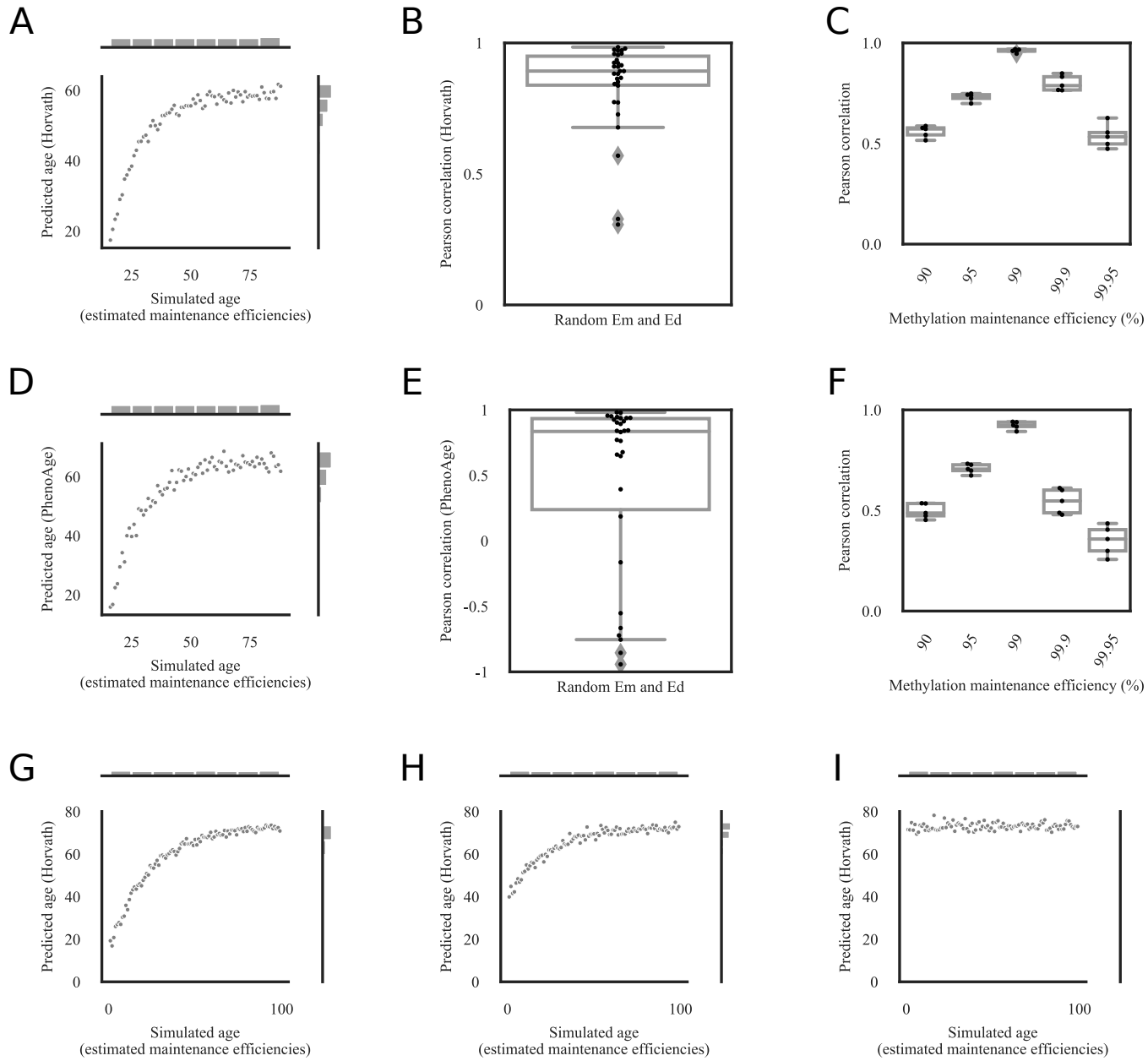
Artiodactyla column corresponds to Figure 5C. Values are shown for tissues and species for which at least 5 samples were available.

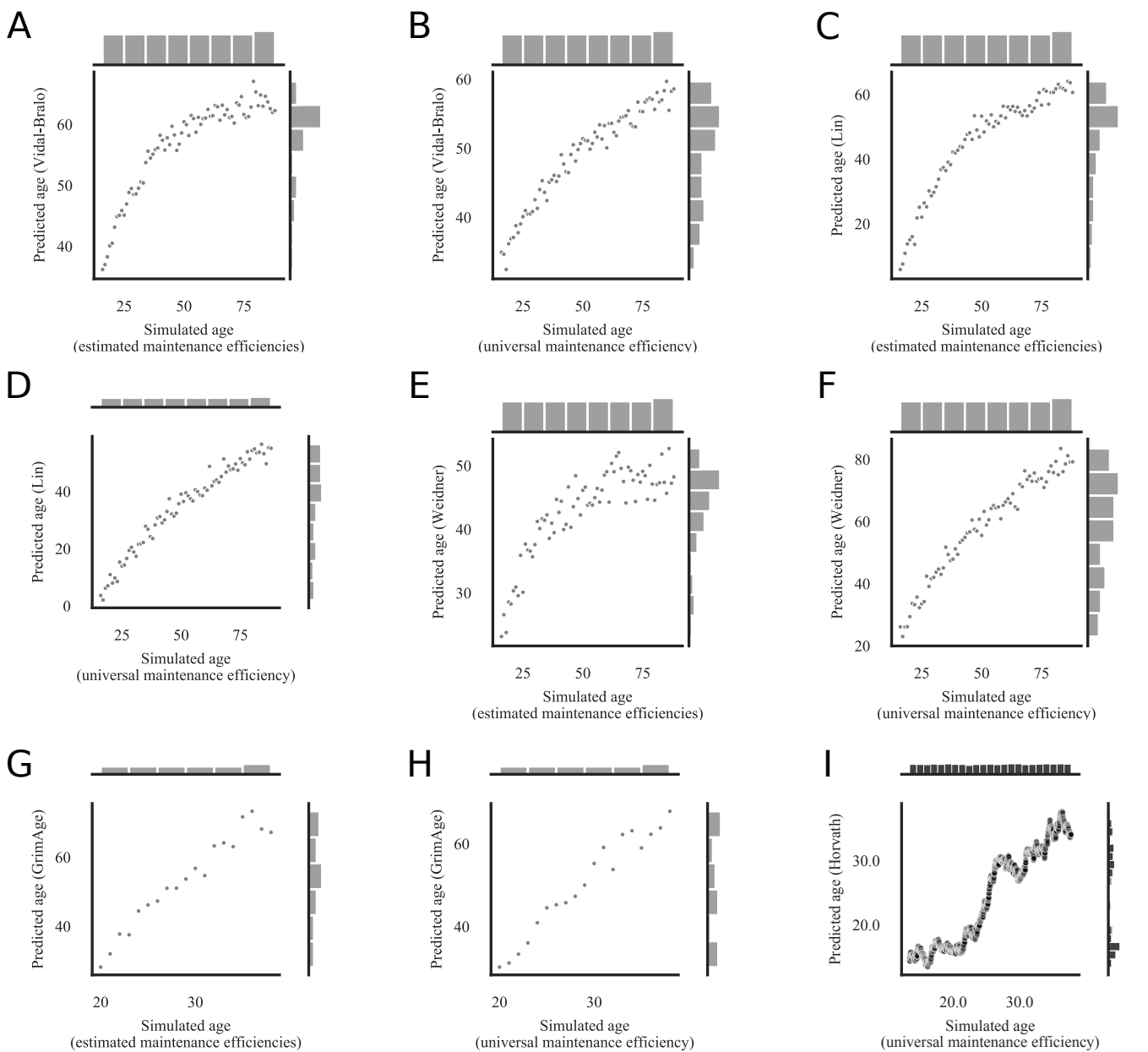
- B) The box-plots show the distribution of Pearson correlation values of Extended Data Figure 10A. Clock 1 trained on samples starting from a Monotremata ground state with accumulating variation show on average a lower accuracy. For each of the 12 clocks (based on a different ground state as shown on the x-axis) the n=57 biologically independent species orders (as indicated in Extended Data Figure 10A) are shown as dots. Boxplots are shown with the center line depicting the median, the box limits the bottom, respective top quartiles, and the whiskers the 1.5x interquartile range.
- C) The same as Extended Data Figure 10B but for Clock 2 trained with 99.99% maintenance rate for all sites of Lu's pan-mammalian relative age-clock. For each of the 12 clocks (based on a different ground state as shown on the x-axis) the n=57 biologically independent species orders (as indicated in Extended Data Figure 10A) are shown as dots. Boxplots are shown with the center line depicting the median, the box limits the bottom, respective top quartiles, and the whiskers the 1.5x interquartile range.
- D) The same as Extended Data Figure 10B but for Clock 3 trained on empirically estimated maintenance rates from the species specified in Extended Data Figure 10A for all 37443 CpG sites. For each of the 12 clocks (based on a different ground state as shown on the x-axis) the n=57 biologically independent species orders (as indicated in Extended Data Figure 10A) are shown as dots. Boxplots are shown with the center line depicting the median, the box limits the bottom, respective top quartiles, and the whiskers the 1.5x interquartile range.
- E) The same as Extended Data Figure 10B but for Clock 4 train with 99.99% maintenance rate for all 37443 CpG sites. For each of the 12 clocks (based on a different ground state as shown on the x-axis) the n=57 biologically independent species orders (as indicated in Extended Data Figure 10A) are shown as dots. Boxplots are shown with the center line depicting the median, the box limits the bottom, respective top quartiles, and the whiskers the 1.5x interquartile range.



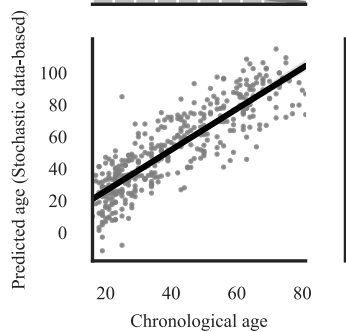
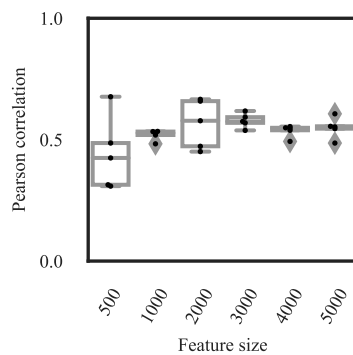
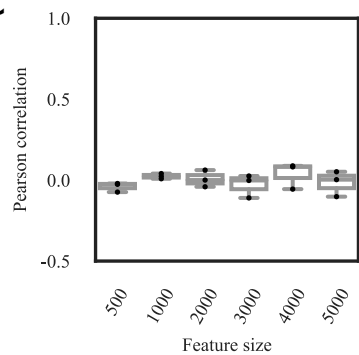
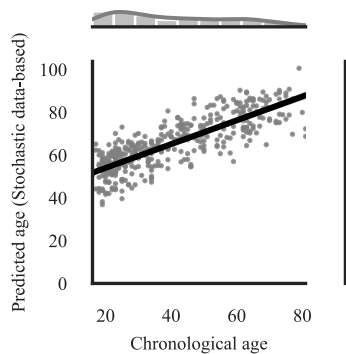
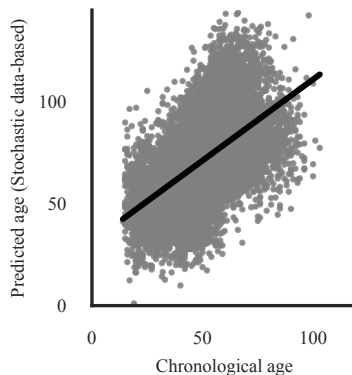


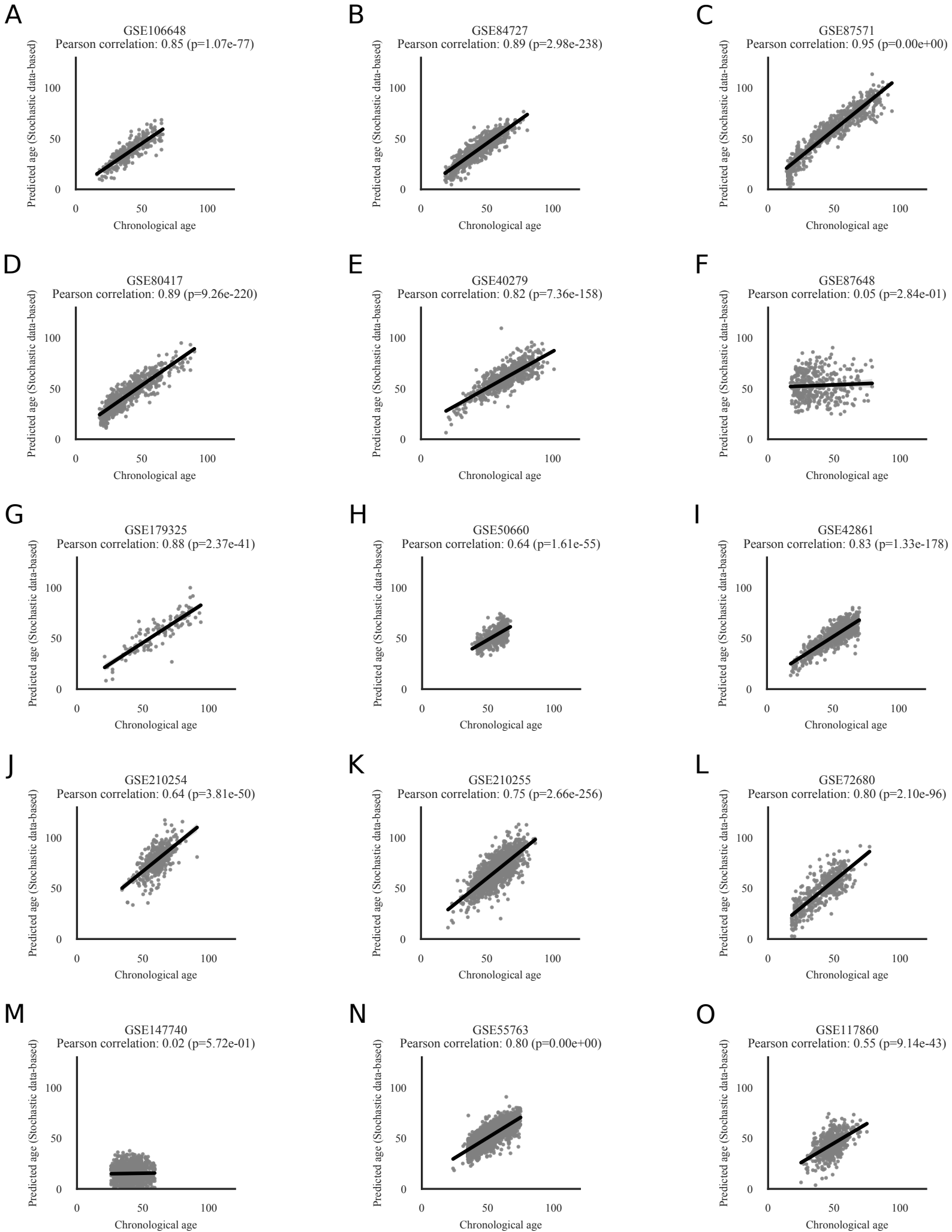


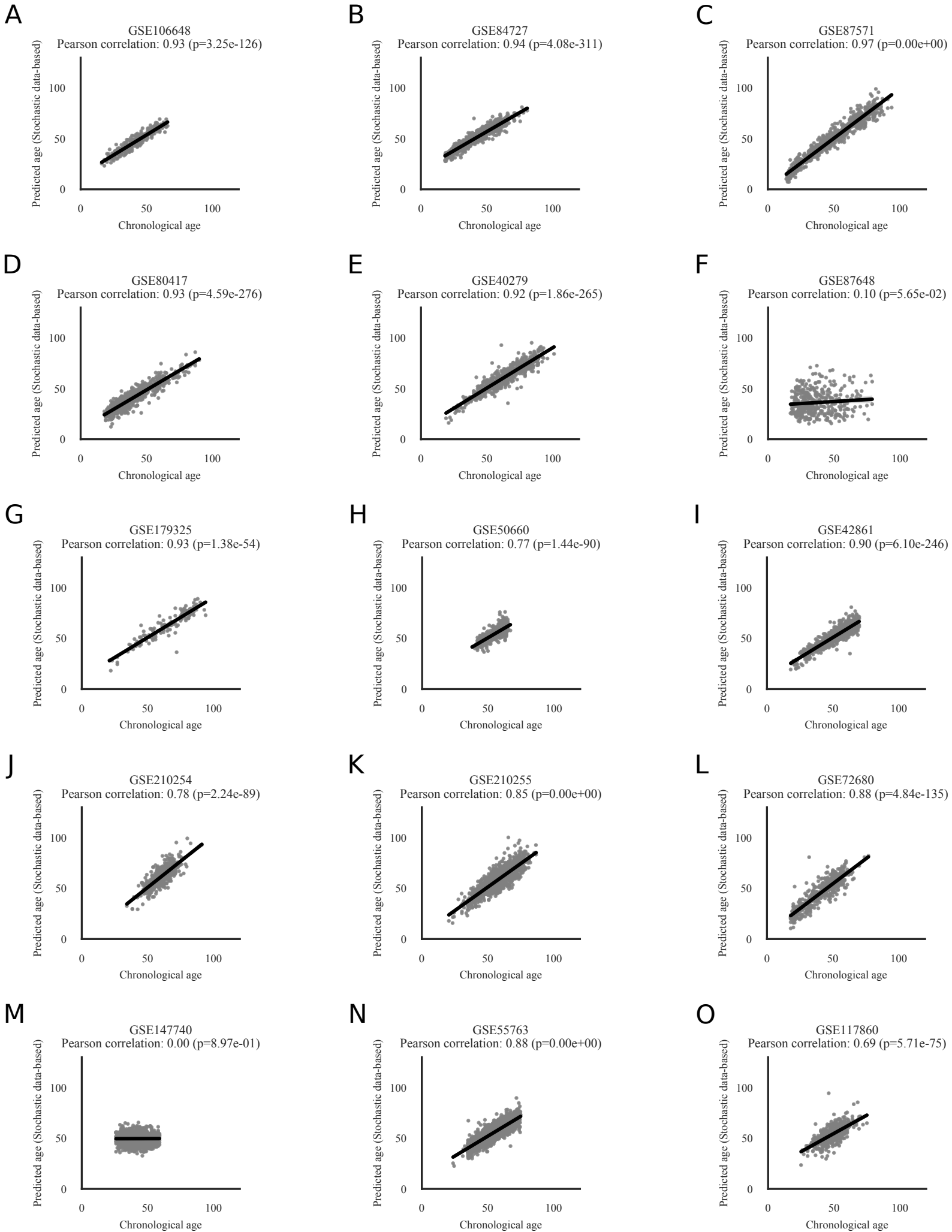


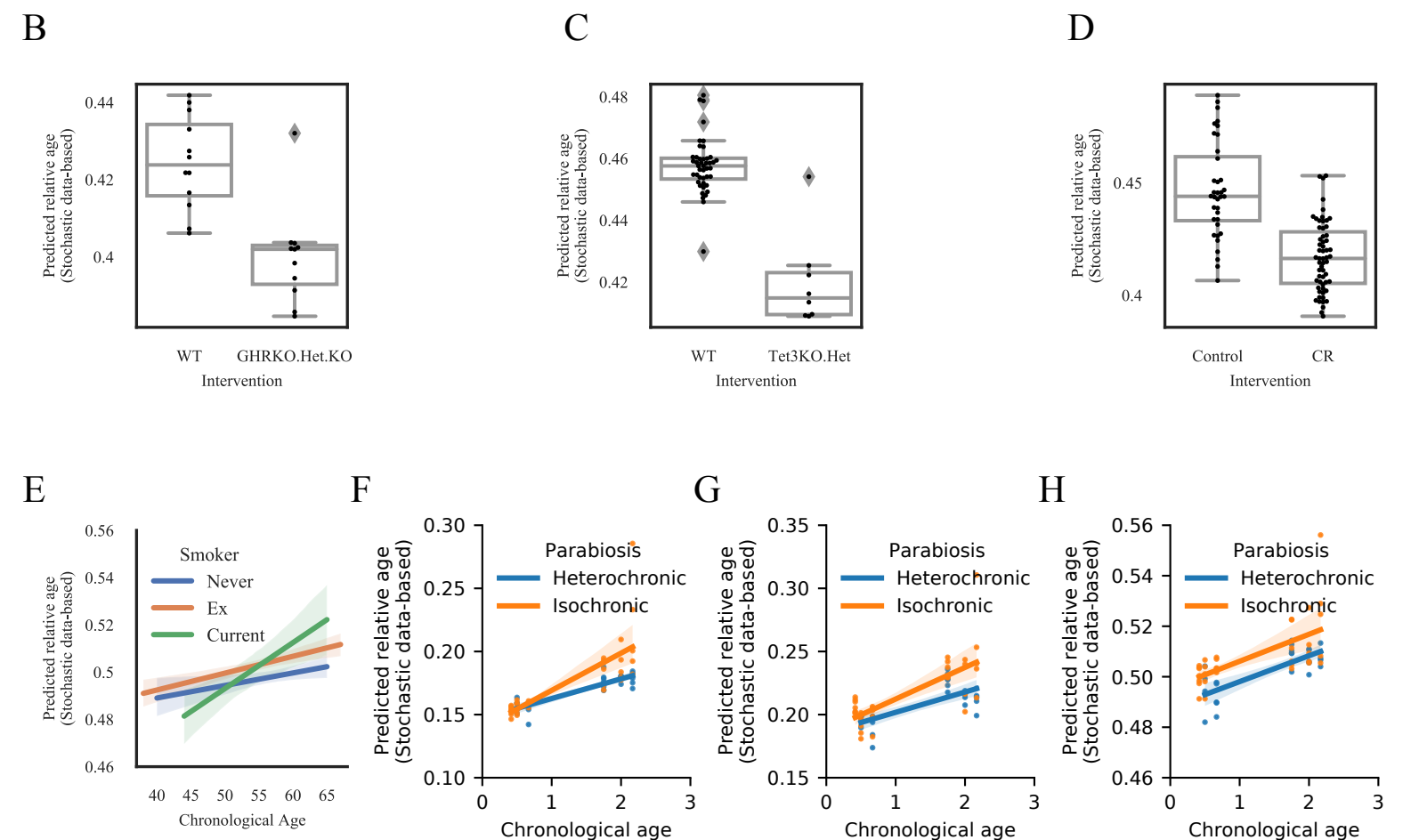
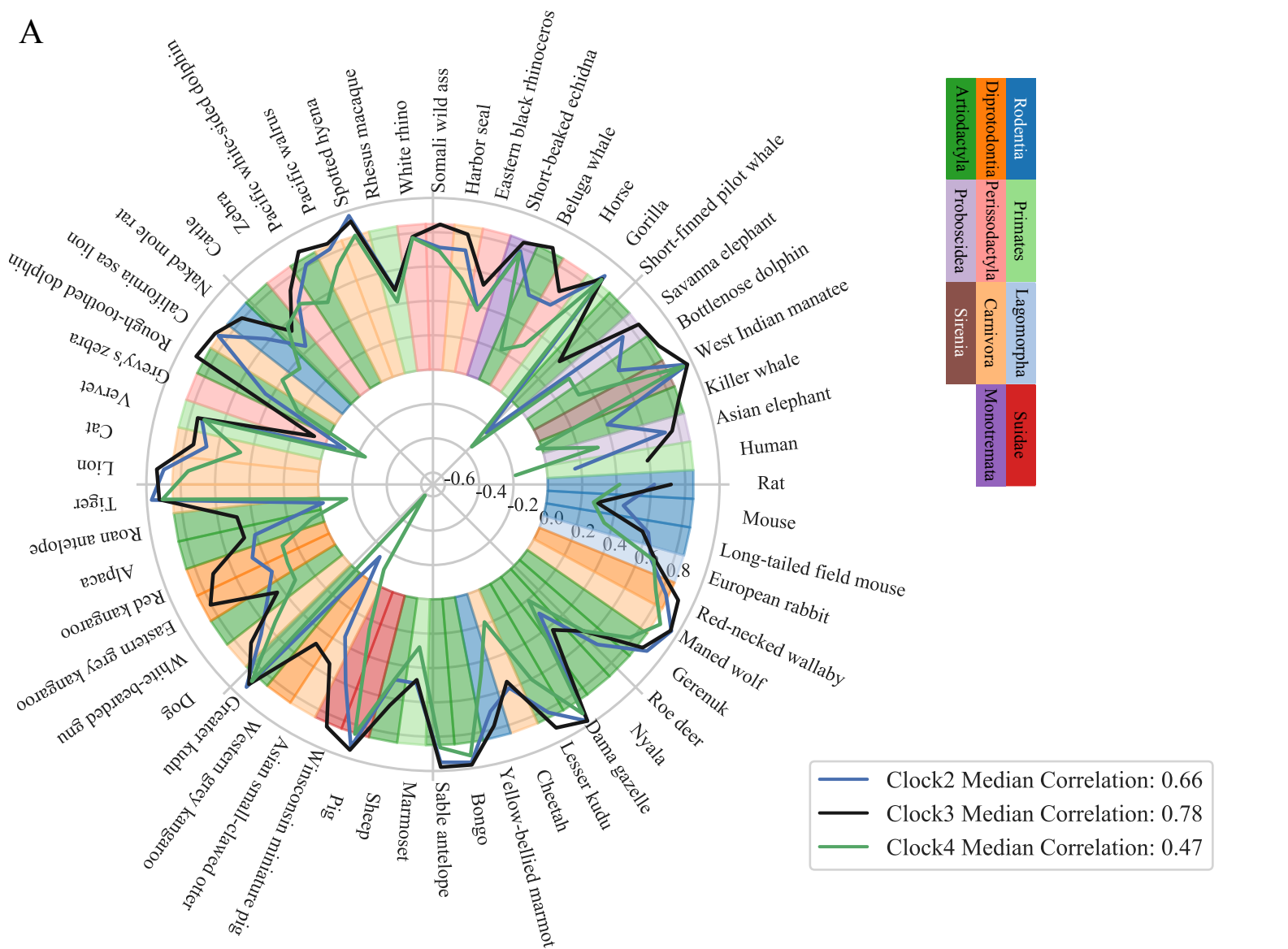


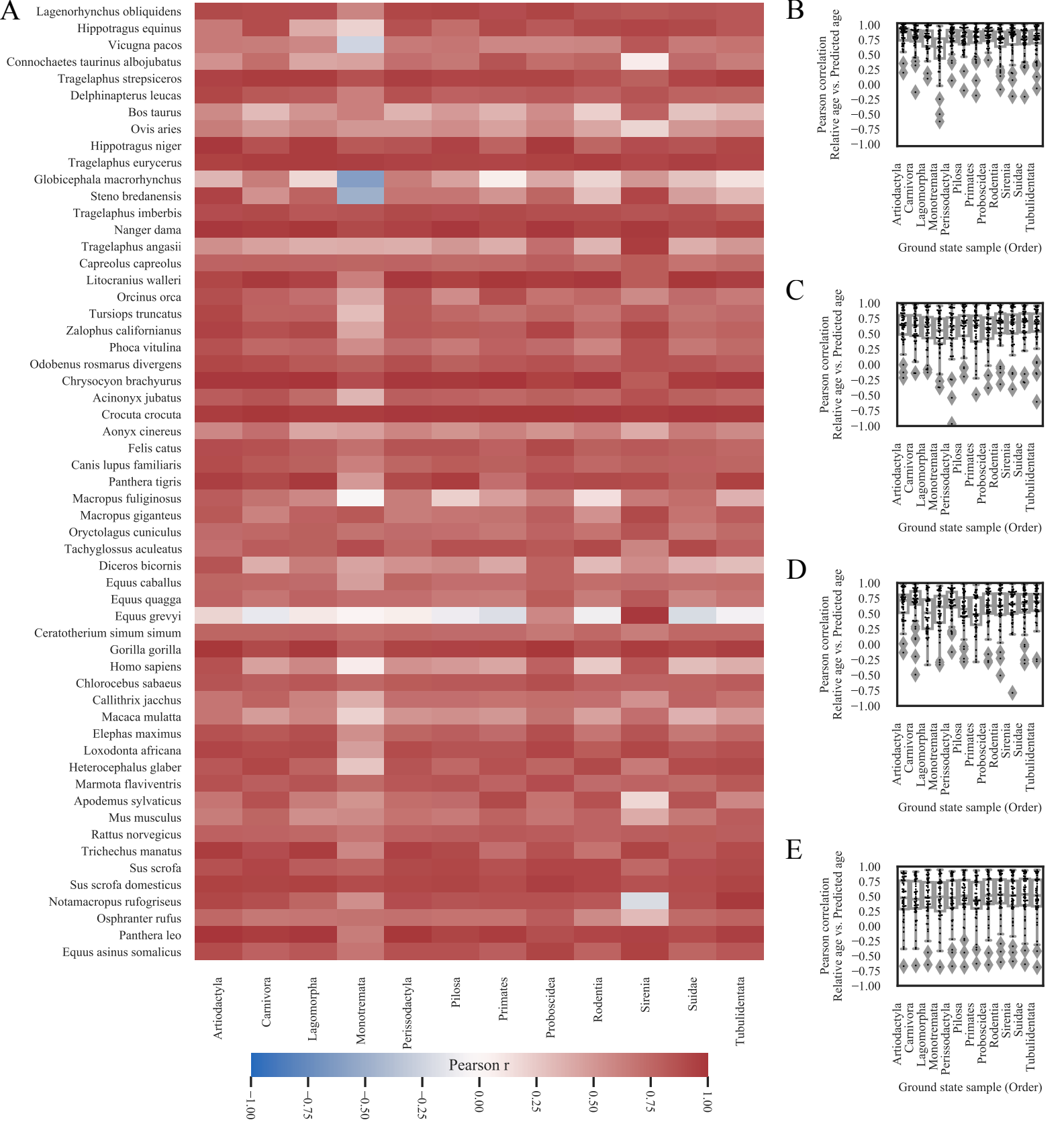
Extended Data Figure 5

A**B****C****D****E**









4 Neuron-type specific aging-rate reveals age decelerating interventions preventing neurodegeneration

David H. Meyer¹, Christian Gallrein¹, Björn Schumacher

¹these authors contributed equally

Manuscript submitted.

Author contributions:

- Conceptualization: C.G., D.H.M., B.S.;
- Experimentation: C.G.;
- Bioinformatics: D.H.M.;
- Data analysis: C.G., D.H.M.;
- Manuscript writing: C.G., D.H.M, B.S.;
- Visualization: C.G., D.H.M.;
- Supervision: B.S.;
- Funding acquisition: B.S.
- The analyses for Figure 1A,B; Figure 3A-D; Figure 4A,B; Figure 5; Figure 6A,B; Supplement Figure 1A-C; Supplement Figure 2A-L and N-W; and Supplement Figure 3 were done by D.H.M.
- The experiments for Figure 2A-C; Figure 4C; Figure 6C; Supplement Figure 2M; and Supplement Figure 4 were done by C.G.

1 **Neuron-type specific aging-rate reveals age decelerating interventions**
2 **preventing neurodegeneration**

3
4 Christian Gallrein^{1,2,3,◇}, David H. Meyer^{1,2,◇}, and Björn Schumacher^{1,2,*}

5 ¹Institute for Genome Stability in Aging and Disease, Medical Faculty, University and University Hospital of Cologne, Joseph-
6 Stelzmann-Str. 26, 50931 Cologne, Germany

7 ²Cologne Excellence Cluster for Cellular Stress Responses in Aging-Associated Diseases (CECAD), Center for Molecular
8 Medicine Cologne (CMCC), University of Cologne, Joseph-Stelzmann-Str. 26, 50931 Cologne, Germany

9 ³Leibniz Institute on Aging – Fritz Lipmann Institute (FLI), Beutenbergstraße 11, 07745 Jena, Germany

10 * correspondence: bjoern.schumacher@uni-koeln.de ◇ these authors contributed equally

11
12 **Keywords**

13 aging, transcriptomic clock, neurodegeneration, neuronal aging intervention

14
15 **Different types of neurons show distinct susceptibility to age-dependent functional decline and**
16 **degeneration and are linked to different types of neurodegenerative disorders. The underlying**
17 **reasons for different aging trajectories of distinct neuron types are poorly understood. Here, we**
18 **employ aging clocks to assess whether distinct neurons differ in their biological age trajectory in**
19 ***C. elegans* where the identity of each single neuron is known. We find that specifically ciliated**
20 **sensory neurons with high neuropeptide expression show the most rapidly progressing biological**
21 **age. The more rapidly aging neurons show a reduction of mitochondrial respiration and elevated**
22 **protein translation gene expression. Reducing protein translation with cycloheximide effectively**
23 **protected fast aging neurons. We show that the *C. elegans* neuronal aging pattern are highly**
24 **correlated with human brain aging and contrasted by geroprotective interventions. We performed**
25 **an *in silico* drug screen and identified known and novel neuroprotective small molecule compounds.**
26 **We show that the natural occurring plant metabolite syringic acid and the piperazine derivative**
27 **vanoxerine delay neuronal degeneration and propose that they could serve as neuroprotective**
28 **interventions. We also identify neurotoxins that accelerate neurodegeneration indicating that this**
29 **approach could reveal interventions as well as risk factors for neuronal aging.**

30

31

32

33 Introduction

34 Age-related diseases are the leading causes of death in high-income countries with the highest
35 prevalence of ischemic heart disease, Alzheimer's Disease (AD) and other dementias, cancer of the
36 respiratory system, intestinal cancer, kidney failure, and diabetes mellitus¹. The distinctive types of
37 chronic diseases of aging indicate that different organs can become the weak-point of an organism that
38 will cause its death and it was reported that nearly 20% of the population show accelerated aging in
39 single organs compared to the rest of their body². In mice, molecular changes across multiple tissues
40 occurring with aging have been described and unique molecular aging trajectories were identified³⁻⁵.
41 Indeed, different organs age at different and individual paces^{2,6,7}. How distinct cell types are differing
42 in their susceptibility to age-related functional decline is largely unexplored.

43 Aging is also the highest risk factor for the development of neurodegenerative diseases like AD or
44 Parkinson's Disease (PD). The distinct neurodegenerative diseases are triggered by the functional
45 decline of distinct neuron types. For instance, in AD, the first tissues degenerating are the
46 parahippocampal gyrus and the olfactory bulb, followed by the degeneration of the hippocampus,
47 leading to the characteristic clinical dementia symptoms⁸⁻¹¹. PD in contrast, affects mostly the
48 dopaminergic innervation of the mid brain and, according to more recent studies, also the cerebellum,
49 resulting in the perturbation of motility and the induction of tremor syndromes^{12,13}. Why and how
50 distinct brain regions are differently affected in neurodegenerative diseases and whether different
51 neuron classes would show intrinsically different aging trajectories is poorly understood. In order to
52 identify neuron-specific aging trajectories, we employed the nematode *Caenorhabditis elegans* as
53 experimental model with a well characterized neuronal system of 302 neurons in total, and ≥ 128
54 different neuron classes¹⁴. Importantly, the nematode system uniquely allows the assessment of the
55 integrity of individual neurons during the normal aging process *in vivo*.

56 Aging trajectories are being increasingly well determined by aging clocks, which are predictive models
57 of molecular signatures that estimate an individual's chronological, and recently also biological age.
58 Aging clocks based on DNA methylation or transcriptome data could predict biological age differences,
59 resulting from genetic age acceleration or deceleration as well as lifespan-affecting treatments^{2,15,16}.
60 Cellular heterogeneity and tissue composition might affect the prediction accuracy and interpretation
61 of bulk aging clocks¹⁷, and recent single-cell transcriptomic clocks highlight cell-type specific aging
62 pattern¹⁸. Still, aging clocks relying on bulk data provide a solid reference because they capture broader
63 aging trajectories, summarizing molecular profiles from diverse cell types, and averaging-out outlier
64 profiles, thus providing general insights into aging patterns and dynamics. While aging clocks have been
65 used to identify biological aging differences across individuals, among distinct treatments, and recently
66 also distinct organs, they have not been used to identify differences among distinct cell types.

67 Here, we asked whether distinct aging trajectories could be identified among the different and well
68 characterized neuron types in *C. elegans*. We show that individual neuron types in young adult animals
69 show a distinct biological age prediction that we find to correspond to the pace of their degeneration
70 in vivo. Neuron types with a higher predicted age show the earliest neurodegeneration during
71 adulthood while predicted young neuron types remain intact. The transcriptomic differences between
72 fast and slow aging neuron types suggest protein translation as a crucial driver of the biological aging
73 process. Slowing down translation with the established translation inhibitor cycloheximide (CHX)
74 reduced the neuronal degeneration of fast aging neurons. We determined that the transcriptional
75 pattern of biological age difference among these neuronal cell types are correlated with transcriptional
76 pattern of human and mouse brain aging and anti-correlated with anti-aging interventions supporting
77 the general validity of our approach. Using a transcriptomic resource incorporating thousands of
78 different compounds in human cell lines (CMAP)¹⁹, we identified pharmacological compounds that we
79 show to prolong the integrity of neurons *in vivo*. We demonstrate that the natural occurring plant
80 metabolite syringic acid and the piperazine derivative vanoxerine delay neuronal degeneration and
81 propose that they could serve as neuroprotective interventions. In reverse, our approach can also
82 identify neurotoxins that accelerate neurodegeneration thus serving as risk assessment. Our data
83 suggest that the mechanisms underlying neuron-type specific aging rates allow the identification of
84 therapeutic interventions that could slow neuronal aging and prevent neurodegeneration.

85

86 **Results**

87 To investigate whether different neuron classes age differently within an organism we applied our
88 previously developed binarized transcriptomic biological age predictor (BitAge)¹⁶ on pseudo-bulk data
89 from the *C. elegans* Neuronal Gene Expression Map & Network (CeNGEN) dataset²⁰, which comprises
90 128 distinct neuron classes from young adult worms. The BitAge clock is trained on the biological age
91 of isogenic whole worm populations and highly accurately predicts not only the chronological age but
92 especially the biological age. The 128 neuron age predictions range from \approx 98 h in FLP neurons to
93 \approx 177 h in ADL neurons (**Figure 1 A**) suggesting that different neurons might indeed show an almost
94 two-fold biological age difference in young adult worms. We independently confirmed the different age
95 predictions of the neuron types in the neuronal pseudo-bulk data from young adult worms in a recent
96 cell atlas of *C. elegans* aging (Calico)²¹ which contains 67 of the 128 neuron classes from the CeNGEN
97 dataset. Also, the Calico dataset of day 1 adult worms exhibits the same predicted age distribution
98 (**Supplement Figure 1A**), and the BitAge predictions on both datasets are significantly correlated to
99 each other (**Supplement Figure 1B**, Pearson correlation 0.64, p-value = 5.92e-09). The youngest and
100 oldest 10% of the 67 neuron classes (as predicted in the CeNGEN dataset) show a stronger correlation

101 (Pearson correlation 0.75, p-value = 2e-03), while the middle group has a higher prediction variance.
102 These results indicate that BitAge is able to predict age differences across different datasets robustly
103 and that neurons indeed show biological age differences on the transcriptome level in young adult
104 worms.

105 Next, we sought to replicate the neuronal age predictions with a method that is not based on the
106 assumption of directed transcriptome changes but stochastic variation accumulation instead. Recently,
107 we showed that all current aging clocks might be driven by accumulating stochastic variation, and that
108 for biological age predictions almost no biological data is required ²². Instead, simulating the aging
109 process by adding stochastic variation to a biological ground state is sufficient to build a predictor that
110 is enabling chronological, as well as biological age predictions. We have shown that for transcriptomic
111 data of *C. elegans* one biological sample as the starting point and stochastic variation drawn from a
112 normal distribution is sufficient for accurate age predictions. This stochastic clock is thereby trained on
113 a different concept than BitAge, which is trained to find biological age pattern in biological samples.
114 Also, the predictions with this stochastic clock are significantly correlated with the BitAge predictions
115 on the CeNGEN dataset (**Supplement Figure 1C**, Pearson correlation 0.65, p-value = 5.5e-17). Similar
116 to Supplement Figure 1B, the 10% youngest and oldest 10% of the 128 CeNGEN neurons show a
117 stronger correlation (Pearson correlation 0.87, p-value = 1.15e-08), while the middle group has a higher
118 prediction variance. These results corroborate the biological age prediction differences, especially in
119 the youngest and oldest neuron groups.

120 The aging transcriptome in species ranging from *C. elegans* to mice was recently shown to exhibit a
121 gene length dependent transcriptional decline (GLTD), where long genes are downregulated with age,
122 while short genes are upregulated ²³⁻²⁵. The reduced expression of long genes likely results from the
123 heightened susceptibility of long genes to accrue transcription-blocking DNA damage. In line with this
124 feature of the aging transcriptome, the marker genes of the 10 % oldest predicted neurons are
125 significantly shorter than expected by chance (adjusted p-value = 0.013), while the marker genes of the
126 10 % youngest neurons are significantly longer than expected (adjusted p-value = 0.037) (**Figure 1B**).
127 These results indicate that even in chronologically young, but biologically old neurons a gene-length
128 dependent transcriptome imbalance can be observed.

129 Taken together, biological and stochastic aging clock measurements and GLTD all suggest specific
130 neuron types to age more rapidly than others.

131

132 *Neuron specific age predictions are associated with differential degeneration*

133 To assess whether the predicted neuron-specific age differences are associated with different degrees
134 of neuron-specific cellular degeneration, we next chose three young (I2, OLL, PHC) and three old (ASI,

135 ASJ, ASK) predicted neurons (**Figure 2A**) and scored their degeneration over the chronological age
136 (**Figure 2B**). Green fluorescent protein (GFP) was expressed under promoters specific for those neurons
137 (ASI, ASJ, ASK, and OLL) or, alternatively, under promoters specifically expressed in a subset of neurons
138 to make identification and segmentation of the selected neurons easier (I2 and PHC) (**Figure 2A**).
139 Macroscopic aberrations on the neurites were counted and subsequently, neurons were classified as
140 healthy, damaged, or severely damaged. In accordance with our predictions, the three young predicted
141 neurons show less degeneration than the old predicted neurons at all analyzed timepoints (**Figure 2C**).
142 On the first day of adulthood, that is closest to the age of the nematodes used for transcriptomics and
143 subsequently for our prediction, I2, OLL, and PHC exhibit a minimal degeneration offset between 10 –
144 20% of all nematodes analyzed. Upon aging, there is a slight, non-significant increases of the fraction
145 of nematodes displaying degeneration (up to $\approx 35\%$) for the young predicted neurons. This fraction of
146 degeneration is consistent with our previous observation of approximately 35% degeneration at day 7
147 of adulthood in URY neurons²⁶, which we here predict to belong to the top 10 youngest neurons (**Figure**
148 **1A**).

149 The ASI, ASJ, and ASK neurons, which are predicted to be biologically older, exhibit >45% damaged
150 neurons already on the first day of adulthood. Interestingly, there is a significant increase of
151 neurodegeneration in ASI (p-value=0.027, Cohen's $h=-1.02$, i.e. a large effect) and ASK (p-value=0.049,
152 Cohen's $h=-0.6$, i.e. a medium effect) neurons upon aging which is in stark contrast to the young
153 predicted neurons where no such elevated degeneration levels could be observed. These results
154 suggest that the predicted biological age differences of the youngest, respective oldest neurons are
155 biologically meaningful and can serve as a predictor of neuronal degeneration already at day 1 of
156 adulthood.

157

158 *Environmental exposition could be a discriminator for premature aging in neurons*

159 Next, we aimed to understand potential commonalities among the youngest, as well as among the
160 oldest neurons. For this, we adapted a hierarchical whole-animal connectome for *C. elegans*²⁷ with a
161 rough anatomical correspondence on the x-axis and directional flow of neuronal signaling on the y-axis
162 and color-coded it with the predicted biological age (**Figure 3A**). The predicted oldest neurons cluster
163 in the top middle part of the network and consist mostly of sensory neurons, while the youngest
164 neurons cluster further to the right. 6 out of the 10 oldest neurons are amphid neurons (ADL, ASJ, ASK,
165 ASG, ADF, ASI), the primary chemosensory organ of *C. elegans*, which is mostly ciliated²⁸. Indeed,
166 comparing the 14 amphid neurons of the CeNGEN dataset with the remaining 114 neurons shows a
167 significant increased biological age (**Figure 3B**, p-value=1.8e-07), which can be replicated in the Calico
168 dataset (**Supplement Figure 2A**, p-value=4.2e-08), and with the stochastic data-based clock predictions

169 **(Supplement Figure 2B, p-value=7.2e-22)**. Amphid neurons, as part of the sensory system, express a
170 variety of neuropeptides, neurotransmitters, receptors, and innexins to transmit the sensed cues. The
171 number of expressed neuropeptides and receptors are significantly higher in amphid neurons
172 **(Supplement Figure 2C,D)**, while the number of neurotransmitter or innexins is not significantly
173 changed **(Supplement Figure 2E,F)**²⁹. Moreover, the number of neuropeptides and the number of
174 receptors per neuron are significantly positively correlated with the predicted biological age in the
175 CeNGEN dataset **(Supplement Figure 2G,H)**, the Calico dataset **(Supplement Figure 2I,J)**, and the
176 predictions with the stochastic data-based clock **(Supplement Figure 2K,L)**. Depletion of *unc-31* leads
177 to reduced neuropeptide release and exhibits a mild, but significant reduction of degeneration in ASI
178 neurons **(Supplement Figure 2M, p-value=0.03, Cohen's h=0.36)**. The number of innexins and number
179 of total synapses per neuron do not show a significantly positive, and potentially rather a negative
180 correlation with the predicted ages **(Supplement Figure 2N-S)**.

181 Amphid neurons are part of the ciliated neuron classes; comparing all 28 ciliated neurons with the
182 remaining 100 neurons also shows a significant increased biological age **(Figure 3C, p-value=0.0006)**,
183 which can be replicated in the Calico dataset **(Supplement Figure 2T, p-value=9e-05)**, and with the
184 stochastic data-based clock predictions **(Supplement Figure 2U, p-value=2.3e-14)**. The ciliated neurons
185 can be further divided into five distinct classes dependent on where its cilia terminate²⁸. Neurons with
186 cilia exposed to the environment show the highest predicted biological age **(Figure 3D, 1-way ANOVA:**
187 **8.8e-06)**, while the other ciliated neuron classes are not significantly different from not-ciliated
188 neurons. A similar effect can be observed in the Calico dataset **(Supplement Figure 2V, 1-way ANOVA:**
189 **1.8e-08)**, and the predictions with the stochastic data-based clock **(Supplement Figure 2W, 1-way**
190 **ANOVA: 6.6e-18)**. These results indicate that the oldest neurons are functionally related and mostly
191 consist of ciliated sensory neurons; that contact to the environment; and the production, potentially
192 the translational load, of neuropeptides are associated with more rapid biological age progression.

193

194 *Transcriptional Clustering identified reduced translation efficacy as potential driver of neuronal aging*

195 We next sought to identify the age-related transcriptional patterns and signatures underlying the
196 biological age distinctions across the 128 neuron classes. To mitigate data variance and extract
197 overarching trends, we initially categorized the neurons into five distinct groups based on their
198 predicted age, followed by a fuzzy clustering analysis. We identified 4 transcriptional clusters **(Figure**
199 **4A, Supplement Figure 3A)**: Cluster 1 shows a general increase over the predicted age and is enriched
200 for stress-induced pathways including DNA repair, response to DNA damage stimulus, transcription-
201 related pathways, and synthesis of ribosomal components, while oxidative phosphorylation is under-
202 represented **(Figure 4B)**. Cluster 2 shows the highest expression in the youngest age group, generally

203 declines over the predicted age time-course, and is enriched in mRNA processing, active translation,
204 and oxidative phosphorylation. Cluster 3 is showing an increase until the last age group, in which it
205 sharply declines, and is enriched in ribosomal genes and proteasome core complex genes. Cluster 4 is
206 especially increased in the oldest age group and is enriched in cilia, immune response, and
207 neuropeptide genes. Conversely, translation-related pathways and oxidative phosphorylation are
208 under-represented. As shown above, the oldest age group is enriched in amphid neurons (ADL, ASJ,
209 ASK, ASG, ADF, ASI, AWA, ASEL), which are mostly ciliated²⁸ and exposed to the environment (**Figure**
210 **3**). The strong enrichment of translation-related pathways (Cluster 1) in the highly expressed genes in
211 the most rapidly aging neurons and the lower translation-related pathways (Cluster 4) in the most
212 slowly aging neurons is consistent with recent studies on transcriptional changes with chronological
213 age in the brain of different organisms^{30–32}.

214

215 *Inhibition of translation alleviates neurodegeneration in fast aging neurons*

216 Based on the enrichment of active protein biosynthesis processes in the accelerated aging neurons, we
217 aimed to test whether translational activity contributes to neurodegeneration. Hence, we treated
218 nematodes for 24 h with the translation inhibitor cycloheximide (CHX) and scored neurite degeneration
219 in ASK and ASJ neurons (from the group of old predicted neurons), as well as in I2 and OLL neurons (as
220 representatives of the young predicted neurons). In the young predicted neurons, no effect of CHX or
221 a DMSO-control treatment was observed (**Figure 4C**). In contrast, old predicted neurons exhibited
222 significantly less neurite deterioration upon CHX-treatment, with a Cohen's h of 1.4, i.e. large effect
223 size, for the ASI, and a Cohen's h of 0.79, i.e. medium to large effect size for the ASK neuron. These
224 results indicate that translational activity in the old predicted neurons is responsible for the premature
225 neurodegeneration that was observed.

226

227 *Comparison with mammalian brain aging*

228 In order to see whether the biological age-related transcriptional patterns of chronologically young
229 adult *C. elegans* neurons (NeuronAge) might be conserved to higher organisms, we next compared the
230 conserved KEGG pathway enrichments of NeuronAge with mouse and human datasets. We computed
231 age-correlations of z-score normalized gene counts for all human brain regions in the GTEx dataset³³,
232 the Tabula Muris Senis (TMS) dataset⁵, and an additional mouse Hypothalamus aging cohort
233 (GSE157025). Similarly, we calculated the enriched pathways for several “anti-aging” treatments like
234 young serum injections³⁴, the platelet factor PF4³⁵, sport in humans³⁶, and krilloil in *C. elegans*³⁷. An
235 unbiased clustering analysis revealed that the aging-trajectories of *C. elegans*, mouse, and humans
236 cluster together. We validated the clustering of NeuronAge by including the conserved pathway

237 enrichments for NeuronAge on the Calico dataset. The trajectories of the anti-aging interventions
238 formed a separate cluster that negatively correlates with the brain aging, irrespective of the organism
239 (**Figure 5**). These results indicate that neuronal transcriptomic aging trajectories are conserved from
240 nematodes to mice and humans and that known anti-aging treatments largely anti-correlate with the
241 aging datasets supporting their geroprotective effectiveness.

242

243 *Identification of drugs preserving neuronal function*

244 As the NeuronAge trajectories cluster together with human neuronal aging trajectories, we sought to
245 use transcriptome data to identify small molecule compounds that could delay neuronal aging. We
246 used the transcriptome resource CMAP consisting of 470k transcriptomes of 19,841 different
247 pharmacological compound treatments in human cell lines¹⁹. We focused on the 3,566 samples for the
248 terminally differentiated neuronal cell line NEU, consisting of 2,467 different molecules (**Figure 6A,B**).
249 Based on the NeuronAge prediction, this analysis identified both negatively correlated compounds
250 (potentially neuro-protective / anti-aging active) and positively correlated compounds (potentially
251 neuro-toxic / pro-aging active). Pathways enriched upon CHX treatment compared to control are
252 significantly negatively correlated with pathways enriched in NeuronAge (Pearson Correlation -0.22),
253 indicating that the beneficial effect of CHX that we saw is mirrored in the transcriptome. To identify
254 neuro-protective small molecule compounds, we ranked the enriched pathways for all 170 molecules
255 that remained after filtering and before using an absolute correlation threshold of 0.25 (see Source
256 Data). The top anti-NeuronAge compound hits contain several (9 out of 16) for which a protective effect
257 for neurons has been previously documented, thus validating our approach (**Figure 6B**). The glycogen
258 synthase kinase 3 (GSK3) inhibitor *AR-A014418* was shown to inhibit beta-amyloid induced
259 neurodegeneration³⁸; the selective serotonin reuptake inhibitor *fluoxetine* protects against
260 neurotoxicity and neurodegeneration³⁹⁻⁴¹; the PPAR-alpha activator *gemfibrozil* exhibits
261 neuroprotective effects via upregulating pro-survival factors and suppressing inflammation⁴²; the
262 kinase inhibitor *sorafenib* protects against neurodegeneration in *C. elegans*⁴³; the selective aryl
263 hydrocarbon receptor modulator *3,3'-diindolylmethane* (DIM) is neuroprotective and promotes brain-
264 derived neurotrophic factor (BDNF)^{44,45}; the insulin-sensitizing agent *rosiglitazone* exhibits
265 neuroprotective effects in the eye and the brain⁴⁶⁻⁴⁸; the p38 MAPK inhibitor *SB202190* was shown to
266 reduce hippocampal apoptosis and rescue spatial learning as well as memory deficits in rats⁴⁹;
267 *dibutyryl-cAMP-Na* (dBcAMP) elevates cAMP levels and protects against neurodegeneration in stab
268 wound or kainic acid injuries⁵⁰⁻⁵²; and the catecholamine-O-methyltransferase inhibitor *tolcapone* was
269 shown to improve cognitive function⁵³.

270 2 out of 15 “pro-NeuronAge” compounds were shown to be detrimental, while 2 are potentially
271 protective. *BAY-K8644* is known to be neurotoxic⁵⁴; and *amiodarone* induces neuronal apoptosis⁵⁵ and
272 is known to induce adverse neurological effects⁵⁶. *Tacedinaline/CI-994* is a class I histone deacetylase
273 inhibitor correlates positively with NeuronAge, was shown to promote functional recovery following
274 spinal cord injury⁵⁷, and to enhance synaptic and structural neuroplasticity⁵⁸. Of note, this effect might
275 be due to a hormetic response^{59–61}. Likewise, *resveratrol* is potentially neuroprotective⁶² due to a
276 hormetic response⁶³. More than half of the top hits have, however, not been tested in neurons. It is
277 conceivable that a short-term “pro-NeuronAge” effect might be hormetic and anti-aging after more
278 time has passed, potentially explaining the effect of *resveratrol* and *tacedinaline*.

279 In summary, 11 out of 31 compounds have neuroprotective evidence, out of which 9 are predicted to
280 revert NeuronAge, i.e. “anti-NeuronAge”, with our *in-silico* approach, while 2 out of 31 compounds are
281 known to be neurotoxic, both of which are predicted correctly to be “pro-aging”, giving weight to the
282 potential that an *in-silico* screen has to identify novel compounds.

283

284 *Identification of neuro-protective molecule compounds*

285 Next, we aimed to determine whether compounds that we predicted to be anti-NeuronAge, i.e.
286 neuroprotective, could indeed prevent the age-related functional decline of aging neurons. We chose
287 two compounds, that were among the most strongly anti-correlated with NeuronAge patterns, BRD-
288 K13195996 and vanoxerine (**Figure 6B**). The chemical identity of the phenolic compound BRD-
289 K13195996 is 3-Hydroxy-4,5-dimethoxybenzoic acid, which is related to 4-Hydroxy-3,5-
290 dimethoxybenzoic acid that is known as syringic acid. Syringic acid is a naturally occurring secondary
291 compound derived from edible plants and fruits, among those olives, walnuts, and grapes – and
292 furthermore red wine and honey⁶⁴. A correlation between the anti-oxidative properties of syringic acid
293 and reduced neurotoxicity following bisphenol A insult has recently been shown⁶⁵, yet no clear
294 mechanism is reported so far⁶⁶. Given the dietary availability of syringic acid, we chose to test its effect
295 on rapidly aging neurons in *C. elegans*. Vanoxerine is a potent dopamine uptake inhibitor and has been
296 developed as cocaine-abuse medication⁶⁷, and, moreover, vanoxerine was observed to impede
297 colorectal cancer stem cell functions by repressing G9a expression⁶⁸. Vanoxerine was so far not
298 reported to exhibit neuroprotection or anti-aging effects.

299 We applied either compounds to nematodes for a 24 h short-term treatment. We assessed neurite
300 degeneration in ASJ and ASK neurons (exemplarily for the old predicted neurons) and observed a
301 significantly reduced deterioration for both compounds, with Cohen’s h ranging from 0.8 to 0.97, i.e.
302 large effect sizes (**Figure 6C**). Applying either of the compounds to nematodes showed no significant
303 adverse effects on OLL neurons (**Supplement Figure 4A**). This indicates that both compounds interfere

304 with the physiological degeneration process of the old predicted neurons and are able to restore a
305 healthy neuron state.

306 Next, we assessed whether our NeuronAge compound predictions could also identify neurotoxic
307 compounds and hence serve for compound risk assessment. We tested the 5-HT_{1A} serotonin receptor
308 antagonist WAY-100635, for which so far no adverse effects on neuron health have been reported, in
309 nematode I2 and OLL neurons (as representatives of healthy young neurons). We observed that WAY-
310 100635 induced significant neurite deterioration in I2 (p-value=0.02, Cohen's h=-0.76) but not in OLL
311 (p-value=0.08, Cohen's h=-0.34) neurons (**Figure 6C**). This indicates that WAY-100635 does not have an
312 indiscriminate effect on all neurons but is more selective, potentially depending on surface receptor
313 expression, presentation, or specific neuronal metabolism patterns.

314 Taken together, we could validate the anti-NeuronAge compound prediction method by identifying
315 known neuroprotectors as well as discovering previously unknown neuroprotective molecules. In
316 reverse, a positively correlated NeuronAge prediction could identify neurotoxic compound properties.

317

318 **Discussion**

319 Why distinct neuron types exhibit different susceptibility to age-dependent degeneration and the
320 associated neurodegenerative diseases has remained largely unclear. While differences in inter-
321 individual aging are commonly known, differential aging of organs within the same organism, and aging
322 variance between cells of the same tissue have recently been observed²⁻⁵. Here, we aimed to
323 understand the aging pace of distinct neuron types to elucidate whether and why specific neurons age
324 faster than others. We employed an aging clock approach to predict the age of distinct neuron types.
325 Aging clocks are increasingly useful for measuring biological age and hold the promise to assessing an
326 individual's biological age. Employing the 'BiT age' and the 'stochastic aging clock' on the single neuron
327 transcriptomics dataset (CeNGEN) we predicted the biological age of the 128 distinct neuron types in
328 *C. elegans*. We observed that the youngest predicted neurons' biological ages were roughly the
329 chronologic age of the nematodes, while the oldest predicted neurons' biological ages were about 1.5-
330 fold as old. The age dependent decline of long gene expression confirmed the distinct biological age of
331 certain cell types at the same chronological age of the animal. The nematode model provides the
332 distinct advantage that the integrity of single neurons could be followed during aging in live animals.
333 We indeed observed that the age-dependent degeneration of specific neurons corresponds to the
334 prediction of the aging clocks evidencing their reliability in identifying neuron-type specific aging
335 trajectories.

336 In the nematode the identity and connectivity of each individual neuron is known, thus allowing to
337 address how their distinct biological age is linked to their biological function. We found that particularly

338 ciliated sensory neurons show an accelerated age trajectory. This might be linked to their exposure to
339 the environment but could also indicate their functional requirement during larval development where
340 the sensing of environmental condition, for instance for deciding to enter dauer stage amid food
341 scarcity or overcrowding, is pivotal for survival. Similar to the exposed ciliated neurons in the
342 nematode, olfactory neurons in the nose of higher organisms are constantly exposed to the
343 environment. Indeed, in humans olfactory capacity is known to decline with age and olfactory
344 dysfunction is an early sign of neurodegenerative diseases⁶⁹. Our results shed light on the possibility
345 that this exposure to the environment might lead to a faster pace of aging and subsequent
346 neurodegeneration.

347 We show that the neuron-type specific biological age differences can be used to determine
348 neuroprotective transcriptome compositions. We identify distinct transcriptional patterns over the
349 neuronal biological aging course and predicted that reduction of the translational load might reduce
350 the pace of neuronal aging. Age-related changes in the translational machinery and the translation rate
351 can be observed across various species and downregulation of translation has been shown to extend
352 lifespan and healthspan parameters as evidenced by dietary restriction studies, downregulation of
353 mTOR, or CHX treatment⁷⁰. Recently, it has been shown that stoichiometric changes of the ribosome
354 are prominent in the aging brain of *Nothobranchius furzeri*, leading to enrichment in protein aggregates
355 in old brains³⁰, which is in line with the over-enrichment of ribosomal proteins in the insolublome of
356 old *C. elegans*⁷¹. Consistent with prediction and literature, we observed that transient inhibition of
357 translation by CHX treatment is sufficient to reduce degeneration of the fast-aging neuronal cell types.

358 As neuron function is highly conserved from nematode to human, we tested whether age-dependent
359 transcriptome changes might be similar. Indeed, we found that the transcriptional patterns of
360 nematode neuronal biological age differences are significantly correlated with mouse and human brain
361 aging trajectories, and anti-correlated with known anti-aging interventions such as young plasma
362 treatment or sport. This conservation shows the potential of identifying conserved mechanisms that
363 underlie the aging trajectories and might determine the susceptibility of specific neuron types to
364 undergo degeneration and potentially contribute to specific neurodegenerative diseases.

365 We used the conservation of transcriptome age-trajectories to *in silico* screen for novel compounds
366 using the human CMAP dataset¹⁹. Such approaches are highly valuable as recent studies employed a
367 transcriptome-based approach for drug screening using the CMAP resource to successfully identify
368 geroprotective compounds that either induce a ‘youthful’ state as predicted through an age-
369 classification approach leveraging the GTEx³³ transcriptomic dataset⁷², by mimicking longevity FOXO3
370 overexpression⁷³, or a “youthful” matreotype⁷⁴. Here, we extended this approach further and used
371 transcriptomic data from *C. elegans* to compare to the CMAP resource to identify neuro-protective

372 small molecule compounds. This approach successfully picked up nine known geroprotective
373 interventions and also identified molecule compounds whose effect on neuroprotection was previously
374 unexplored. By testing the top scoring compounds, we indeed found that they extend the integrity of
375 fast aging neurons indicating a biological age-deceleration. Conversely, the pro-aging prediction
376 revealed neurotoxic effects of compounds and could thus be highly valuable in risk assessment.

377 As proof of concept, we determined that syringic acid and vanoxerine effectively preserved the integrity
378 of fast aging neurons. Syringic acid indeed has been suggested to exert neuroprotective effects through
379 its antioxidant properties⁶⁶. Both vanoxerine and WAY-100635 have been developed to treat cocaine
380 addiction but our data indicate starkly contrasting effects on neuronal aging. The high-affinity
381 dopamine reuptake inhibitor vanoxerine was initially developed as antidepressant and has entered
382 clinical trials for treatment of cocaine addiction⁷⁵ and, based on its property as ion channel blocker, for
383 atrial fibrillation or atrial flutter⁷⁶. The 5-HT1A serotonin receptor antagonist WAY-100635 has been
384 tested preclinically for cocaine addiction⁷⁷ and treatment of depressive disorders⁷⁸. We propose that
385 NeuronAging clocks and the aging-associated gene expression responses we determined here could be
386 useful in risk assessment given the strong similarities we show between the *C. elegans* neuronal aging
387 trajectories and human brain aging

388 Taken together, we here define the biological basis for the distinct susceptibility of neurons to undergo
389 age-dependent degeneration. We establish the utility of employing aging clocks to identify neuron-
390 type specific aging rates and based on their transcriptome profiles reveal conserved aging pattern also
391 present in human brain aging. We show that this approach is suitable for identifying neuroprotective
392 molecules and propose that they could be useful in delaying neuronal aging and protect from age-
393 associated degeneration.

394

395

396 **Methods**

397 *C. elegans culture.*

398 Nematodes were cultured on nematode growth medium (NGM) agar plates at 20 °C under standard
399 conditions unless stated otherwise. All age statements given in in this publication consider the first
400 day of adulthood as day 1. A complete strain list can be found in the supplement.

401

402 *Neurite imaging*

403 Nematodes were synchronized by L4 picking and grown on standard NGM for one day, four days, or
404 seven days. For imaging nematodes were placed in a drop of 250 mM NaN₃ on a 2% agarose pad.
405 Imaging was performed on a Zeiss Imager.M2 at 400fold magnification. Z-Stacks of nematode heads /
406 tails were acquired employing 2 μm step width. Acquisition time was set between 300 ms to 3 s per
407 plane to achieve a good signal to noise ratio and was strongly depending on the imaged expression
408 strain.

409

410 *Scoring of neurite degeneration*

411 Recorded Z-stack images of neurons were analyzed by hand, counting blebs, large spherical
412 outgrowths, branching, breaks, and necrosis on the dendrites of the analyzed neurons. Images were
413 classified according to the degree of aberration: necrotic neurons, broken or truncated neurons,
414 neurons with ≥ 10 blebs, or ≥ 3 outgrowths were scored as 'severely damaged'; neurons with 5 – 9
415 blebs, or 2 outgrowths were classified as 'mildly damaged'; and neurons with less < 5 blebs or < 2
416 outgrowths were classified as 'healthy'. See Figure 2B for exemplary images.

417

418 *Compound treatment*

419 Standard NGM plates seeded with OP50 were coated with compounds by dropwise adding compounds,
420 dissolved in 300 – 1000 μl medium, directly to the plate's surface. Plates were dried for at least 1 h
421 before transferring L4 stage nematodes onto them. Nematodes were incubated with the compounds
422 for 24 h and then used for neurite imaging. Final concentration of compounds was: Cycloheximide –
423 2 mM; Syringic acid – 2.5 mM; Vanoxerine – 10 nM; WAY-100635 – 25 nM. Control nematodes were
424 incubated on appropriate solvent control coated plates (either water or ≤ 5 % DMSO).

425

426 *BitAge prediction*

427 The BitAge clock¹⁶ was used as described previously. Briefly, each sample was binarized, i.e. genes
428 higher than the median expression value within each sample after removing genes with zero counts,
429 were set to 1, and the remaining genes to 0. The BitAge coefficients for the 576 clock genes are added
430 up for all genes in a given sample that is 1 after binarization. After adding the BitAge intercept the
431 results show the predicted biological age.

432

433 *Stochastic-data based clock*

434 The stochastic-data based clock was used as described previously²². Briefly, each sample was log10-
435 transformed after the addition of one pseudo-count. Subsequently, the samples were min-max
436 normalized to bring each sample within the range [0,1], and then binarized as described above. The
437 normalized counts were then added up for all 1010 stochastic-data based clock genes (see Source
438 Data). Note that the stochastic clock might result in slightly different genes every-time a clock is trained.

439

440 *Gene length analysis*

441 First, we downloaded the differential expressed genes for each neuron and all other cells in the
442 CeNGEN²⁰ dataset (<https://cengen.shinyapps.io/CengenApp/>) with the statistical test “Wilcoxon on
443 single cells”. This gives a list of genes with log fold changes, and percentage expression in the specific
444 neuron and all other cells. We further filtered this list of significant genes by requiring that the gene is
445 expressed in at least 90% of cells of the specific neuron and at most 10% in all other cell types. 39
446 neurons had no genes with these requirements, i.e. 89 neurons were used for further analysis. Next,
447 we used the marker genes of the 10% oldest (31 genes), respective 10% youngest neurons (33 genes)
448 and calculated the density distribution of the log10-normalized gene lengths. To calculate a two-sided
449 permutation test, we calculated the median log10 gene length of the genes and compared it to the
450 100.000 permutation median. The permutation for the old marker genes used 31 genes, the
451 permutation for the young marker genes 33 genes out of all marker genes of the 89 neurons (303
452 genes).

453

454 *Neuronal connectome mapping*

455 We downloaded and adapted the Cytoscape file from Cook et al.²⁷ by adding head neurons, deleting
456 non-neuronal cell-types, and color-coding neurons by their predicted Age with BitAge on the CeNGEN
457 dataset.

458 Median total number of synapses calculation

459 The NeuroType.xlsx file was downloaded from <https://www.wormatlas.org/neuronalwiring.html>. For
460 each of the 128 neuron classes we summed up the median total number of synapses in the head, tail,
461 and mid-body.

462

463 *Fuzzy clustering*

464 To cluster general trajectories over the predicted age, we first summarized the gene expression into 5
465 bins: 1) [97-110], 2) (110,120], 3) (120,130], 4) (130,140], 5) (140, 180]. Within each bin, we computed
466 the median expression level for each gene. To make the genes comparable and bring them onto the
467 same scale we calculated the z-score across the 5 bins for each of the 9950 genes with non-zero
468 standard deviation. Next, we used fuzzy clustering with Mfuzz v2.58.0⁷⁹ to identify trajectories across
469 the predicted age bins. The elbow method was computed with the Dmin function of Mfuzz and
470 indicated an optimal number of 4 clusters. The genes belonging to each cluster were subsequently used
471 for a pathway enrichment analysis with clusterprofiler v4.9.2.002⁸⁰, with maxGSSize=500 and the list
472 of all 9950 genes as the background gene list.

473

474 *Heatmap*

475 We processed several public datasets for the heatmap:

- 476 1. TPM normalized gene expression values for human brain tissues from the GTEx v8 data⁸¹
477 release (i.e. Amygdala, Anterior Cingulate Cortex, Caudate Basal Ganglia, Cerebellar
478 Hemisphere, Cerebellum, Cortex, Frontal Cortex, Hippocampus, Hypothalamus, Nuclear
479 Accumbens Basal, Putamen Basal Ganglia, Substantia Nigra) were correlated with the
480 chronological age (the midpoints of the publicly available age bins).
- 481 2. The mouse aging time-course for Hypothalamus data from GSE157025 was downloaded and
482 a gene-wise correlation with the chronological age calculated.
- 483 3. Whole brain data from the Tabula Muris Senis (GSE132040) was downloaded and edgeR
484 v3.40.2⁸² to calculate normalized expression values. The normalized expression values were
485 correlated to the chronological age.
- 486 4. Differentially expressed genes for mouse Hippocampus data treated with the platelet factor
487 PF4 or saline control were downloaded from GSE173254. We multiplied the logFCs by -1 to
488 always compare treatment vs. control, instead of control vs. treatment.

- 489 5. Differentially expressed genes for mouse Hippocampus data treated with young serum or
490 sham were downloaded from GSE234667.
- 491 6. *C. elegans* whole worm data treated with Krill oil from GSE207152. We used edgeR v3.40.2⁸²
492 to calculate normalized expression values and calculated z-scores for each gene over all
493 samples. The z-score normalized expression values were used for a regression model:
494 $expression = \beta_0 + \beta_1 * Age + \beta_2 * Krilloil + \beta_3 * (Age * Krilloil)$, where β_0 is the
495 intercept term, β_1 is the coefficient for the Age variable, β_2 is the coefficient for the Krilloil-
496 treatment variable, and β_3 is the coefficient for the interaction between Age and Krilloil, i.e.
497 the difference in the slope over age.
- 498 7. Differential expressed genes upon physical activity were downloaded from the supplementary
499 data from PMID: **30927700**.

500 The Pearson correlation values of all genes of 1.) - 3.), the logFC of all genes for 4.)-5.), the β_3
501 coefficients for 6.) were used to calculate enriched pathways analysis with fgseaMultilevel from the
502 fgsea R package⁸³ with nPermSimple=1000 for all conserved KEGG pathways between *C. elegans*,
503 mouse, and humans. For 7.) the “anti-aging/AD” genes were split into genes that are up-, respective
504 down-regulated upon exercise. Both gene sets were used for an enrichment analysis with enricher from
505 enrichplot v1.18.0 with maxGSSize=500, minGSSize=5, and the all genes quantified in the
506 supplementary data from PMID: **30927700** as the background gene list. For both enrichment analyses
507 the enrichment fold change, i.e. number of observed genes per pathway divided by the number of
508 expected genes per pathway, was calculated. Finally, the fold changes were combined, i.e. pathways
509 with a bigger fold change enrichment in the downregulated genes were multiplied by -1.

510 The clustering was done on the normalized enrichment scores for 1.)-6.) and the fold change
511 enrichment score for 7.) with the Ward method and a correlation distance matrix.

512

513 CMAP

514 The CMAP resource uses the L1000 array, which measures 978 landmark transcripts, which can be used
515 to infer most of the remaining transcriptome with high accuracy¹⁹. Here we used all available L1000
516 datasets for a human differentiated neuronal cell line. Despite only measuring a subset of landmark
517 genes and inferring the rest, the CMAP resource has shown to be highly valuable for drug screens. We
518 downloaded the aggregated level 5 L1000 Connectivity Map¹⁹ data from GSE92742 and did a pathway
519 enrichment analysis with fgseaMultilevel from the fgsea R package⁸³ with nPermSimple=1000 and the
520 conserved KEGG pathways between human and *C. elegans* for each of the samples in the level 5
521 dataset. To compare it to the NeuronAge trajectory, we first calculated the z-score of each gene that

522 has at least some gene counts across the 128 neurons of the CeNGEN dataset. We correlated these z-
523 score normalized genes with the biological age prediction from BitAge. The resulting Pearson
524 correlation values for all genes were used for a pathway enrichment analysis with fgseaMultilevel and
525 the conserved KEGG pathways between human and *C. elegans*. To identify whether any compound
526 might revert the NeuronAge gene expression trajectory on the pathway level, we correlated the
527 normalized enrichment scores (NES) of the NeuronAge KEGG pathway enrichment analysis with all NES
528 that we calculated from the CMAP dataset. Next, we filtered for only compounds that were tested in
529 the neuronal cell line (“NEU”). Compounds had to be tested at least twice, with all measurements
530 resulting in correlations in the same direction. Additionally, we filtered for compounds that were
531 measured at 24h and 6h and took only those compounds that showed a stronger correlation into the
532 same direction at the 24h timepoint compared to the 6h timepoint. Lastly, we filtered out those
533 compounds that had no information available at PubCHEM and used a correlation threshold of 0.25,
534 respective -0.25.

535

536 Statistics

537 All data are presented as mean \pm SD. Number of cohorts (N), individuals (n), and technical replicates
538 (N) is stated in the figures and their respective figure legends. The applied statistical tests are
539 mentioned in the figure legends and the respective p-values are directly reported in the diagrams. All
540 statistics were done two-sided if not stated otherwise. Independent t-tests were calculated with
541 Python’s Scipy⁸⁴ v1.5.1 stats.ttest_ind function. Kruskal-Wallis tests were calculated with Python’s
542 Scipy v1.5.1 stats.kruskall() function. One-way ANOVA’s were calculated with Python’s pingouin
543⁸⁵v0.3.6 anova function and the parameter ss_type=2. Cohen’s h⁸⁶ as a measure of effect size was
544 calculated by hand with Python’s numpy⁸⁷ v1.18.5. Plots were generated with Python’s seaborn⁸⁸
545 v0.11.0, matplotlib⁸⁹ v3.3.0, or GraphPad Prism 9. Boxplots are shown with the center line depicting
546 the median, the box limits the bottom, respective top quartiles, and the whiskers the 1.5x interquartile
547 range. Scatterplots showing a linear regression model fit are shown with a 95% confidence interval.

548

549 *Data availability*

550 The unfiltered TPM counts and the Cell Marker list was downloaded from the CENGEN dataset were
551 assessed at <https://cengen.shinyapps.io/CengenApp/> .

552 The Calico dataset was downloaded from <https://c.elegans.aging.atlas.research.calicolabs.com/data> .

553 The neuron-specific information was assessed at <https://www.wormatlas.org/> .

554 The gene length information was downloaded from <https://wormbase.org/>.

555 The CMAP data were downloaded from GSE92742.

556 Data for the heatmap were downloaded either from the GEO database: GSE157025, GSE132040,
557 GSE173254, GSE234667, GSE207152. From the Supplementary data from PMID: **30927700**. Or the
558 GTEx v8 database: https://gtexportal.org/home/downloads/adult-gtex/bulk_tissue_expression

559

560

561 **Acknowledgement**

562 We thank the Regional Computing Center of the University of Cologne (RRZK) for providing computing
563 time on the DFG-funded High Performance Computing (HPC) system CHEOPS as well as support. Worm
564 strains were provided by the National Bioresource Project (supported by The Ministry of Education,
565 Culture, Sports, Science and Technology, Japan), the Caenorhabditis Genetics Center (funded by the
566 NIH National Center for Research Resources, US), and the C. elegans Gene Knockout Project at the
567 Oklahoma Medical Research Foundation (part of the International C. elegans Gene Knockout
568 Consortium). We express our gratitude to Piali Sengupta for sharing PY6457 strain.

569 C.G. was supported by the Peter and Traudl Engelhorn Foundation. D.H.M. is member of the Cologne
570 Graduate School of Ageing Research. B.S. acknowledges funding from the Deutsche
571 Forschungsgemeinschaft (Reinhart Koselleck-Project 524088035, FOR 5504 project 496650118, SCHU
572 2494/3-1, SCHU 2494/7-1, SCHU 2494/10-1, SCHU 2494/11-1, SCHU 2494/15-1, CECAD EXC 2030 –
573 390661388, and GRK2407), the Deutsche Krebshilfe (70114555), Deutsche José Carreras Leukämie-
574 Stiftung (DJCLS 04 R/2023), and the John Templeton Foundation Grant (61734).

575 **Author contributions**

576 *Conceptualization:* C.G., D.H.M., B.S.; *Experimentation:* C.G.; *Bioinformatics:* D.H.M; *Data analysis:*
577 C.G., D.H.M.; *Manuscript writing:* C.G., D.H.M, B.S.; *Visualization:* C.G., D.H.M; *Supervision:* B.S.;
578 *Funding acquisition:* B.S.

579 **Declaration of Interests**

580 All authors declare no competing interests.

581

582

583 **Figure Legends**

584 *Figure 1*

585 A) Distribution of transcriptomic age predictions. The 128 neurons of the CeNGEN dataset were
586 predicted with BitAge and sorted by their predicted age. The x-axis shows the rank of the
587 prediction in ascending order, the y-axis the predicted age. The ten youngest neurons and their
588 respective age predictions are outlined in blue; the ten oldest neurons and their respective age
589 predictions are displayed in orange.

590 B) The marker genes of neurons in the CeNGEN dataset show a gene length dependent
591 transcriptional decline. The log₁₀ gene length (x-axis) of the marker genes of the 10% youngest
592 neurons (blue), and of the 10% oldest neurons (orange) are compared to random permutation
593 of all marker genes with the same number of genes. The two-sided permutation test compared
594 the median log₁₀ gene length. The y-axis shows the probability density of the values on the x-
595 axis.

596

597 *Figure 2 – Predicted neuron age and degeneration onset and progression correlate*

598 A) Representative fluorescence images of the analysed neurons grouped by prediction age – I2,
599 OLL, and PHC in blue; ASI, ASJ, and ASK in orange. Small schematic images taken from
600 WormAtlas.

601 B) Representative fluorescence images (Z-stack maximum projections) of nematodes expressing
602 neuronal volume markers, classified according to the severity of observed degeneration.
603 Orange arrows indicate blebs, red arrow heads indicate spheric outgrowths. Nematode heads
604 are outlined by a dashed line. Scale bar represents 50 μm.

605 C) Fraction-plots displaying the fraction of nematodes expressing neuronal volume markers in
606 different neurons categorized as ‘healthy’, ‘mildly damaged’, and ‘severely damaged’. Three to
607 four cohorts were analysed, comprised of 10 – 30 individual nematodes, for every timepoint
608 indicated. Kruskal-Wallis-test was employed to test for significant differences upon aging within
609 the neuron classes.

610

611 *Figure 3 – Environment-exposed ciliated neurons are old predicted*

612 A) We adapted a previously published connectome of *C. elegans*²⁷. Only neuronal cells are shown
613 in a largely directional information flow on the vertical axis, with sensory neurons (triangles)
614 on top, interneurons (hexagons) in the middle, and motor neurons (circles) on the bottom²⁷.
615 The horizontal axis roughly shows the anatomical orientation with the head region on the left,

616 and posterior neurons on the right. Chemical synapses and gap junctions are indicated as faded
617 grey lines. The size of the neurons indicates the number of cells within this neuron class. The
618 predicted age (BitAge based on the CeNGEN dataset) is color coded from blue (young) to
619 orange (old). The oldest neurons cluster in the middle top part and are largely sensory neurons.

620 B) Dot/Box-plot showing BitAge predictions grouped by Amphid neurons / non-amphid neurons.
621 Two-sided t-test was performed to test for significant age differences.

622 C) Dot/Box-plot showing BitAge predictions grouped by ciliated neurons / non-ciliated neurons.
623 Two-sided t-test was performed to test for significant age differences.

624 D) Dot/Box-plot showing ciliated neurons' age predictions divided into 5 classes depending on
625 where its cilia terminate. ANOVA + Tukey post hoc test was performed to test for significant
626 differences.

627

628 *Figure 4 – Fuzzy clustering reveals translation dynamics as potential driver of neuron aging*

629 A) Fuzzy clustering on z-score normalized genes over the predicted aging course of the 128
630 CeNGEN neurons identified four clusters. The 128 neurons were merged into five age-
631 prediction bins (1) 97-110h, (2) 110-120h, (3) 120-130h, (4) 130-140h, (5)140-180h.

632 B) Pathway analysis of the four clusters shows age-related pathways. The log fold change (logFC)
633 of the pathway enrichment is color-coded (from blue = under-represented to red = over-
634 represented). Circle size displays the $-\log_{10}$ false discovery rate ($-\log_{10}\text{FDR}$), for ease of
635 interpretation the circle size legend displays the values as the FDR. The number of genes (n) in
636 each cluster is annotated.

637 C) Fraction-plots displaying the fraction of nematodes expressing neuronal volume markers in
638 different neurons categorized as 'healthy', 'mildly damaged', and 'severely damaged' that were
639 treated with 2 mM cycloheximid (CHX) for 24 h. Three to four cohorts were analysed, consisting
640 of 10 – 30 individual nematodes. Mann-Whitney-test was employed to test for significant
641 differences.

642

643 *Figure 5 - Neuronal aging trajectories are conserved across C. elegans, mice, and humans*

644 A) The normalized enrichment scores of the conserved KEGG pathways for the indicated aging
645 trajectories or treatment-effects were used for an unbiased clustering analysis. The matrix is
646 color-coded according to the Pearson correlation between the indicated comparisons, non-
647 significant correlations are colored white. The colors on the side indicate the species and
648 whether it was an aging trajectory or an anti-aging treatment.

649

650 *Figure 6 – Compound prediction algorithm identifying neuro-protective / neurotoxic compounds*

- 651 A) Flowchart explaining the *in silico* drug screening. We computed and correlated the conserved
652 KEGG pathway enrichments for NeuronAge and all compounds from the CMAP dataset that
653 are measured on the neuronal cell line NEU. To obtain a manageable list of compounds we
654 filtered for compounds that were measured at least twice, show consistent correlations in all
655 measurements, have a stronger correlation at the 24h timepoint compared to the 6h
656 timepoint, have information in PubCHEM, and at least an absolute correlation value of 0.25.
- 657 B) The top anti-NeuronAge and pro-NeuronAge compounds after the filtering steps ranked
658 according to their Pearson correlation. Previously published neuro-protective (blue) or
659 neurotoxic (orange) compounds are indicated. Three, in regards to their effect on neuronal
660 health, uncharacterized compounds are highlighted by black arrows and their structural
661 formula is given.
- 662 C) Fraction-plots displaying the fraction of nematodes expressing neuronal volume markers in
663 different neurons categorized as ‘healthy’, ‘mildly damaged’, and ‘severely damaged’ that were
664 treated with 2.5 mM syringic acid (SA), 10 nM vanoxerine (VX), or 25 nM WAY-100635 (WAY)
665 for 24 h. Three to four cohorts were analysed, consisting of 10 – 25 individual nematodes.
666 Mann-Whitney-test was employed to test for significant differences.

667

668

669 **References**

- 670 1. Organization, W. H. *World health statistics 2019: Monitoring health for the SDGs, sustainable*
671 *development goals*. (World Health Organization, 2019, 2019).
- 672 2. Oh, H. S.-H. *et al.* Organ aging signatures in the plasma proteome track health and disease.
673 *Nature* **624**, 164–172 (2023).
- 674 3. Schaum, N. *et al.* Ageing hallmarks exhibit organ-specific temporal signatures. *Nature* **583**,
675 596–602 (2020).
- 676 4. Pálovics, R. *et al.* Molecular hallmarks of heterochronic parabiosis at single-cell resolution.
677 *Nature* **603**, 309–314 (2022).
- 678 5. Tabula Muris Consortium. A single-cell transcriptomic atlas characterizes ageing tissues in the
679 mouse. *Nature* **583**, 590–595 (2020).
- 680 6. Tian, Y. E. *et al.* Heterogeneous aging across multiple organ systems and prediction of chronic
681 disease and mortality. *Nat. Med.* **29**, 1221–1231 (2023).
- 682 7. Ahadi, S. *et al.* Personal aging markers and ageotypes revealed by deep longitudinal profiling.
683 *Nat. Med.* **26**, 83–90 (2020).
- 684 8. Van Hoesen, G. W., Augustinack, J. C., Dierking, J., Redman, S. J. & Thangavel, R. The
685 parahippocampal gyrus in Alzheimer’s disease. Clinical and preclinical neuroanatomical
686 correlates. *Ann. N. Y. Acad. Sci.* **911**, 254–74 (2000).

- 687 9. Echávarri, C. *et al.* Atrophy in the parahippocampal gyrus as an early biomarker of Alzheimer's
688 disease. *Brain Struct. Funct.* **215**, 265–71 (2011).
- 689 10. Braak, H. & Braak, E. Neurofibrillary changes confined to the entorhinal region and an
690 abundance of cortical amyloid in cases of presenile and senile dementia. *Acta Neuropathol.*
691 **80**, 479–86 (1990).
- 692 11. Murphy, C. Olfactory and other sensory impairments in Alzheimer disease. *Nat. Rev. Neurol.*
693 **15**, 11–24 (2019).
- 694 12. Li, T., Le, W. & Jankovic, J. Linking the cerebellum to Parkinson disease: an update. *Nat. Rev.*
695 *Neurol.* **19**, 645–654 (2023).
- 696 13. Surmeier, D. J. Determinants of dopaminergic neuron loss in Parkinson's disease. *FEBS J.* **285**,
697 3657–3668 (2018).
- 698 14. White, J. G., Southgate, E., Thomson, J. N. & Brenner, S. The structure of the nervous system
699 of the nematode *Caenorhabditis elegans*. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **314**, 1–340
700 (1986).
- 701 15. Lu, A. T. *et al.* Universal DNA methylation age across mammalian tissues. *Nat. aging* **3**, 1144–
702 1166 (2023).
- 703 16. Meyer, D. H. & Schumacher, B. BiT age: A transcriptome-based aging clock near the
704 theoretical limit of accuracy. *Aging Cell* **20**, e13320 (2021).
- 705 17. Tomusiak, A. *et al.* Development of a novel epigenetic clock resistant to changes in immune
706 cell composition. *bioRxiv [Preprint]* (2023) doi:<https://doi.org/10.1101/2023.03.01.530561>.
- 707 18. Buckley, M. T. *et al.* Cell-type-specific aging clocks to quantify aging and rejuvenation in
708 neurogenic regions of the brain. *Nat. aging* **3**, 121–137 (2023).
- 709 19. Subramanian, A. *et al.* A Next Generation Connectivity Map: L1000 Platform and the First
710 1,000,000 Profiles. *Cell* **171**, 1437–1452.e17 (2017).
- 711 20. Taylor, S. R. *et al.* Molecular topography of an entire nervous system. *Cell* **184**, 4329–4347.e23
712 (2021).
- 713 21. Roux, A. E. *et al.* Individual cell types in *C. elegans* age differently and activate distinct cell-
714 protective responses. *Cell Rep.* **42**, 112902 (2023).
- 715 22. Schumacher, B., Meyer, D. & Meyer, D. H. *Accurate aging clocks based on accumulating*
716 *stochastic variation*. *Research Square* (2023).
- 717 23. Stoeger, T. *et al.* Aging is associated with a systemic length-associated transcriptome
718 imbalance. *Nat. aging* **2**, 1191–1206 (2022).
- 719 24. Ibañez-Solé, O., Barrio, I. & Izeta, A. Age or lifestyle-induced accumulation of genotoxicity is
720 associated with a length-dependent decrease in gene expression. *iScience* **26**, 106368 (2023).
- 721 25. Gyenis, A. *et al.* Genome-wide RNA polymerase stalling shapes the transcriptome during
722 aging. *Nat. Genet.* **55**, 268–279 (2023).
- 723 26. Gallrein, C. *et al.* Novel amyloid-beta pathology *C. elegans* model reveals distinct neurons as
724 seeds of pathogenicity. *Prog. Neurobiol.* **198**, 101907 (2021).
- 725 27. Cook, S. J. *et al.* Whole-animal connectomes of both *Caenorhabditis elegans* sexes. *Nature*
726 **571**, 63–71 (2019).
- 727 28. Inglis, P. N., Ou, G., Leroux, M. R. & Scholey, J. M. The sensory cilia of *Caenorhabditis elegans*.

- 728 *WormBook* **8**, 1–22 (2007).
- 729 29. Altun, Z. F. Neurotransmitter Receptors in *C. elegans*. *WormAtlas* (2011)
730 doi:10.3908/wormatlas.5.202.
- 731 30. Kelmer Sacramento, E. *et al.* Reduced proteasome activity in the aging brain results in
732 ribosome stoichiometry loss and aggregation. *Mol. Syst. Biol.* **16**, e9596 (2020).
- 733 31. Ximerakis, M. *et al.* Single-cell transcriptomic profiling of the aging mouse brain. *Nat.*
734 *Neurosci.* **22**, 1696–1708 (2019).
- 735 32. Buchwalter, A. & Hetzer, M. W. Nucleolar expansion and elevated protein translation in
736 premature aging. *Nat. Commun.* **8**, 328 (2017).
- 737 33. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis:
738 multitissue gene regulation in humans. *Science* **348**, 648–60 (2015).
- 739 34. Fitz, N. F. *et al.* Extracellular Vesicles in Young Serum Contribute to the Restoration of Age-
740 Related Brain Transcriptomes and Cognition in Old Mice. *Int. J. Mol. Sci.* **24**, 12550 (2023).
- 741 35. Schroer, A. B. *et al.* Platelet factors attenuate inflammation and rescue cognition in ageing.
742 *Nature* **620**, 1071–1079 (2023).
- 743 36. Berchtold, N. C. *et al.* Hippocampal gene expression patterns linked to late-life physical
744 activity oppose age and AD-related transcriptional decline. *Neurobiol. Aging* **78**, 142–154
745 (2019).
- 746 37. SenGupta, T. *et al.* Krill oil protects dopaminergic neurons from age-related degeneration
747 through temporal transcriptome rewiring and suppression of several hallmarks of aging. *Aging*
748 (*Albany, NY*). **14**, 8661–8687 (2022).
- 749 38. Bhat, R. *et al.* Structural insights and biological effects of glycogen synthase kinase 3-specific
750 inhibitor AR-A014418. *J. Biol. Chem.* **278**, 45937–45 (2003).
- 751 39. Zhang, F. *et al.* Fluoxetine protects neurons against microglial activation-mediated
752 neurotoxicity. *Parkinsonism Relat. Disord.* **18 Suppl 1**, S213-7 (2012).
- 753 40. Chung, E. S. *et al.* Fluoxetine prevents LPS-induced degeneration of nigral dopaminergic
754 neurons by inhibiting microglia-mediated oxidative stress. *Brain Res.* **1363**, 143–50 (2010).
- 755 41. Li, I.-H. *et al.* Study on the neuroprotective effect of fluoxetine against MDMA-induced
756 neurotoxicity on the serotonin transporter in rat brain using micro-PET. *Neuroimage* **49**,
757 1259–70 (2010).
- 758 42. Ivraghi, M. S. *et al.* Neuroprotective effects of gemfibrozil in neurological disorders: Focus on
759 inflammation and molecular mechanisms. *CNS Neurosci. Ther.* **30**, e14473 (2024).
- 760 43. Liu, Z. *et al.* Inhibitors of LRRK2 kinase attenuate neurodegeneration and Parkinson-like
761 phenotypes in *Caenorhabditis elegans* and *Drosophila* Parkinson's disease models. *Hum. Mol.*
762 *Genet.* **20**, 3933–42 (2011).
- 763 44. Rzemieniec, J., Wnuk, A., Lason, W., Bilecki, W. & Kajta, M. The neuroprotective action of 3,3'-
764 diindolylmethane against ischemia involves an inhibition of apoptosis and autophagy that
765 depends on HDAC and AhR/CYP1A1 but not ER α /CYP19A1 signaling. *Apoptosis* **24**, 435–452
766 (2019).
- 767 45. Lee, B. D. *et al.* 3,3'-Diindolylmethane Promotes BDNF and Antioxidant Enzyme Formation via
768 TrkB/Akt Pathway Activation for Neuroprotection against Oxidative Stress-Induced Apoptosis
769 in Hippocampal Neuronal Cells. *Antioxidants (Basel, Switzerland)* **9**, (2019).

- 770 46. Normando, E. M. *et al.* The retina as an early biomarker of neurodegeneration in a rotenone-
771 induced model of Parkinson's disease: evidence for a neuroprotective effect of rosiglitazone in
772 the eye and brain. *Acta Neuropathol. Commun.* **4**, 86 (2016).
- 773 47. Zhao, Z. *et al.* Rosiglitazone Exerts an Anti-depressive Effect in Unpredictable Chronic Mild-
774 Stress-Induced Depressive Mice by Maintaining Essential Neuron Autophagy and Inhibiting
775 Excessive Astrocytic Apoptosis. *Front. Mol. Neurosci.* **10**, 293 (2017).
- 776 48. Mishra, J., Chaudhary, T. & Kumar, A. Rosiglitazone synergizes the neuroprotective effects of
777 valproic acid against quinolinic acid-induced neurotoxicity in rats: targeting PPAR γ and HDAC
778 pathways. *Neurotox. Res.* **26**, 130–51 (2014).
- 779 49. Yang, S. *et al.* Protective effects of p38 MAPK inhibitor SB202190 against hippocampal
780 apoptosis and spatial learning and memory deficits in a rat model of vascular dementia.
781 *Biomed Res. Int.* **2013**, 215798 (2013).
- 782 50. Rao, M. S. & Abd-El-Basset, E. M. dBcAMP Rescues the Neurons From Degeneration in Kainic
783 Acid-Injured Hippocampus, Enhances Neurogenesis, Learning, and Memory. *Front. Behav.*
784 *Neurosci.* **14**, 18 (2020).
- 785 51. Abd-El-Basset, E. M. & Rao, M. S. Dibutyl Cyclic Adenosine Monophosphate Rescues the
786 Neurons From Degeneration in Stab Wound and Excitotoxic Injury Models. *Front. Neurosci.*
787 **12**, 546 (2018).
- 788 52. Shohami, E. *et al.* The Ras inhibitor S-trans, trans-farnesylthiosalicylic acid exerts long-lasting
789 neuroprotection in a mouse closed head injury model. *J. Cereb. Blood Flow Metab.* **23**, 728–38
790 (2003).
- 791 53. Apud, J. A. *et al.* Tolcapone improves cognition and cortical information processing in normal
792 human subjects. *Neuropsychopharmacology* **32**, 1011–20 (2007).
- 793 54. Tran, N. K. C. *et al.* Ginsenoside Re blocks Bay k-8644-induced neurotoxicity via attenuating
794 mitochondrial dysfunction and PKC δ activation in the hippocampus of mice: Involvement of
795 antioxidant potential. *Food Chem. Toxicol.* **178**, 113869 (2023).
- 796 55. Liao, R. *et al.* Amiodarone-Induced Retinal Neuronal Cell Apoptosis Attenuated by IGF-1 via
797 Counter Regulation of the PI3k/Akt/FoxO3a Pathway. *Mol. Neurobiol.* **54**, 6931–6943 (2017).
- 798 56. Algharably, E. A. el-H., Di Consiglio, E., Testai, E., Kreutz, R. & Gundert-Remy, U. Prediction of
799 the dose range for adverse neurological effects of amiodarone in patients from an in vitro
800 toxicity test by in vitro-in vivo extrapolation. *Arch. Toxicol.* **95**, 1433–1442 (2021).
- 801 57. Zhang, S., Fujita, Y., Matsuzaki, R. & Yamashita, T. Class I histone deacetylase (HDAC) inhibitor
802 CI-994 promotes functional recovery following spinal cord injury. *Cell Death Dis.* **9**, 460 (2018).
- 803 58. Gräff, J. *et al.* Epigenetic priming of memory updating during reconsolidation to attenuate
804 remote fear memories. *Cell* **156**, 261–76 (2014).
- 805 59. Shao, L.-W. *et al.* Histone deacetylase HDA-1 modulates mitochondrial stress response and
806 longevity. *Nat. Commun.* **11**, 4639 (2020).
- 807 60. McIntyre, R. L., Daniels, E. G., Molenaars, M., Houtkooper, R. H. & Janssens, G. E. From
808 molecular promise to preclinical results: HDAC inhibitors in the race for healthy aging drugs.
809 *EMBO Mol. Med.* **11**, e9854 (2019).
- 810 61. Zeyn, Y. *et al.* Histone deacetylase inhibitors modulate hormesis in leukemic cells with mutant
811 FMS-like tyrosine kinase-3. *Leukemia* **37**, 2319–2323 (2023).
- 812 62. Griñán-Ferré, C. *et al.* The pleiotropic neuroprotective effects of resveratrol in cognitive

- 813 decline and Alzheimer's disease pathology: From antioxidant to epigenetic therapy. *Ageing*
814 *Res. Rev.* **67**, 101271 (2021).
- 815 63. Calabrese, E. J., Mattson, M. P. & Calabrese, V. Resveratrol commonly displays hormesis:
816 occurrence and biomedical significance. *Hum. Exp. Toxicol.* **29**, 980–1015 (2010).
- 817 64. Srinivasulu, C., Ramgopal, M., Ramanjaneyulu, G., Anuradha, C. M. & Suresh Kumar, C.
818 Syringic acid (SA) – A Review of Its Occurrence, Biosynthesis, Pharmacological and Industrial
819 Importance. *Biomed. Pharmacother.* **108**, 547–557 (2018).
- 820 65. Helli, B. *et al.* The Protective Effects of Syringic Acid on Bisphenol A-Induced Neurotoxicity
821 Possibly Through AMPK/PGC-1 α /Fndc5 and CREB/BDNF Signaling Pathways. *Mol. Neurobiol.*
822 (2024) doi:10.1007/s12035-024-04048-0.
- 823 66. Ogut, E., Armagan, K. & Gül, Z. The role of syringic acid as a neuroprotective agent for
824 neurodegenerative disorders and future expectations. *Metab. Brain Dis.* **37**, 859–880 (2022).
- 825 67. Rothman, R. B., Baumann, M. H., Priszczano, T. E. & Newman, A. H. Dopamine transport
826 inhibitors based on GBR12909 and bupropion as potential medications to treat cocaine
827 addiction. *Biochem. Pharmacol.* **75**, 2–16 (2008).
- 828 68. Bergin, C. J. *et al.* The dopamine transporter antagonist vanoxerine inhibits G9a and
829 suppresses cancer stem cell functions in colon tumors. *Nat. cancer* **5**, 463–480 (2024).
- 830 69. Fatuzzo, I. *et al.* Neurons, Nose, and Neurodegenerative Diseases: Olfactory Function and
831 Cognitive Impairment. *Int. J. Mol. Sci.* **24**, (2023).
- 832 70. Takauji, Y. *et al.* Restriction of protein synthesis abolishes senescence features at cellular and
833 organismal levels. *Sci. Rep.* **6**, 18722 (2016).
- 834 71. Xie, X. *et al.* Quantification of Insoluble Protein Aggregation in *Caenorhabditis elegans* during
835 Aging with a Novel Data-Independent Acquisition Workflow. *J. Vis. Exp.* **46**, 248–256 (2020).
- 836 72. Janssens, G. E. *et al.* Transcriptomics-Based Screening Identifies Pharmacological Inhibition of
837 Hsp90 as a Means to Defer Aging. *Cell Rep.* **27**, 467-480.e6 (2019).
- 838 73. McIntyre, R. L. *et al.* Inhibition of the neuromuscular acetylcholine receptor with atracurium
839 activates FOXO/DAF-16-induced longevity. *Aging Cell* **20**, e13381 (2021).
- 840 74. Statzer, C. *et al.* Youthful and age-related matreotypes predict drugs promoting longevity.
841 *Aging Cell* **20**, e13441 (2021).
- 842 75. Gorelick, D. A., Gardner, E. L. & Xi, Z. X. Agents in development for the management of
843 cocaine abuse. *Drugs* **64**, 1547–1573 (2004).
- 844 76. Piccini, J. P. *et al.* Randomized, double-blind, placebo-controlled study to evaluate the safety
845 and efficacy of a single oral dose of vanoxerine for the conversion of subjects with recent
846 onset atrial fibrillation or flutter to normal sinus rhythm: RESTORE SR. *Hear. Rhythm* **13**, 1777–
847 1783 (2016).
- 848 77. Kohtz, A. S., Zhao, J. & Aston-Jones, G. Serotonin signaling in hippocampus during initial
849 cocaine abstinence drives persistent drug seeking. *J. Neurosci.* **44**, e1505212024 (2024).
- 850 78. Kosari-Nasab, M. *et al.* Serotonin 5-HT_{1A} receptors modulate depression-related symptoms
851 following mild traumatic brain injury in male adult mice. *Metab. Brain Dis.* **34**, 575–582
852 (2019).
- 853 79. Kumar, L. & E Futschik, M. Mfuzz: a software package for soft clustering of microarray data.
854 *Bioinformatics* **2**, 5–7 (2007).

855 80. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innov.*
856 *(Cambridge 2, 100141 (2021).*

857 81. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–5
858 (2013).

859 82. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential
860 expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–40 (2010).

861 83. Korotkevich, G. *et al.* Fast gene set enrichment analysis. *bioRxiv* (2021) doi:10.1101/060012.

862 84. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat.*
863 *Methods* **17**, 261–272 (2020).

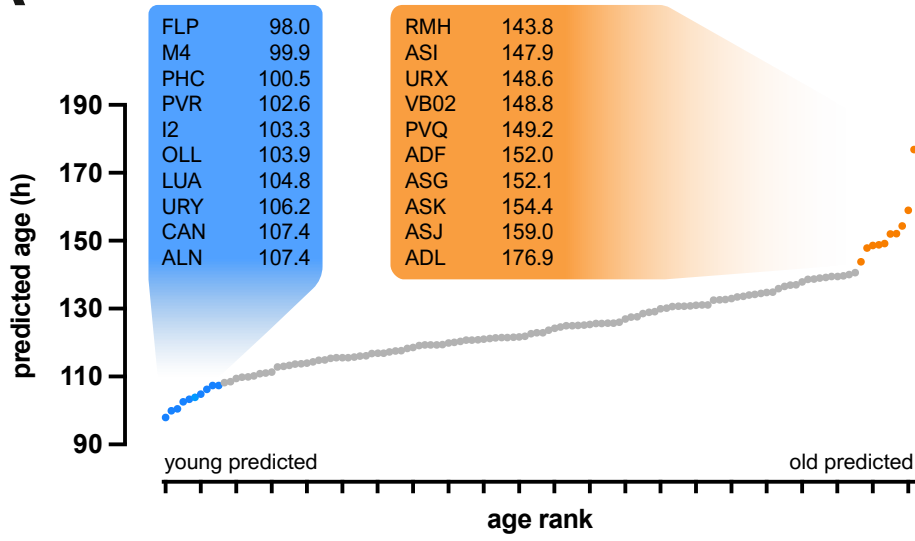
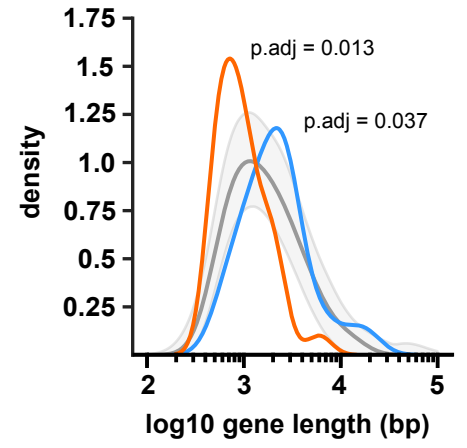
864 85. Vallat, R. Pingouin: statistics in Python. *J. Open Source Softw.* **3**, 1026 (2018).

865 86. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences.* (Elsevier, 1977).
866 doi:10.1016/C2013-0-10517-X.

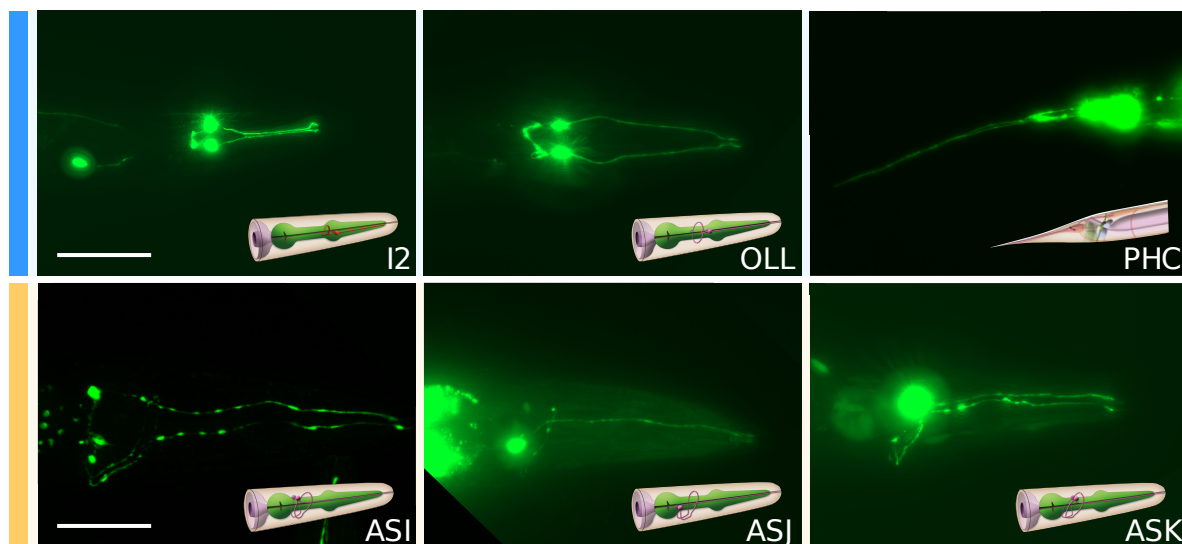
867 87. Harris, C. R. *et al.* Array programming with {NumPy}. *Nature* **585**, 357–362 (2020).

868 88. Waskom, M. L. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).

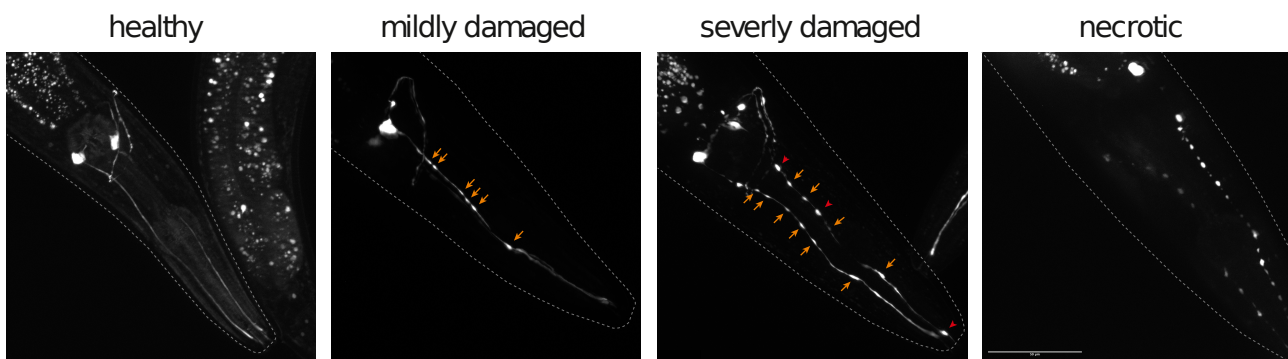
869 89. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. \& Eng.* **9**, 90–95 (2007).
870
871

A**B**

A

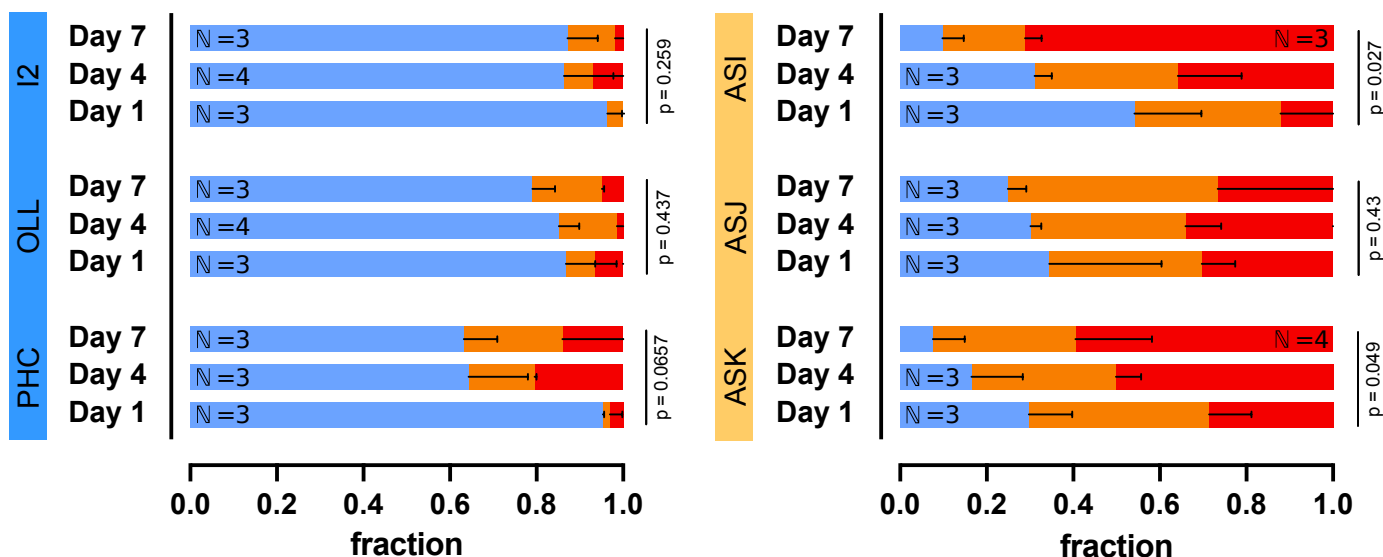


B

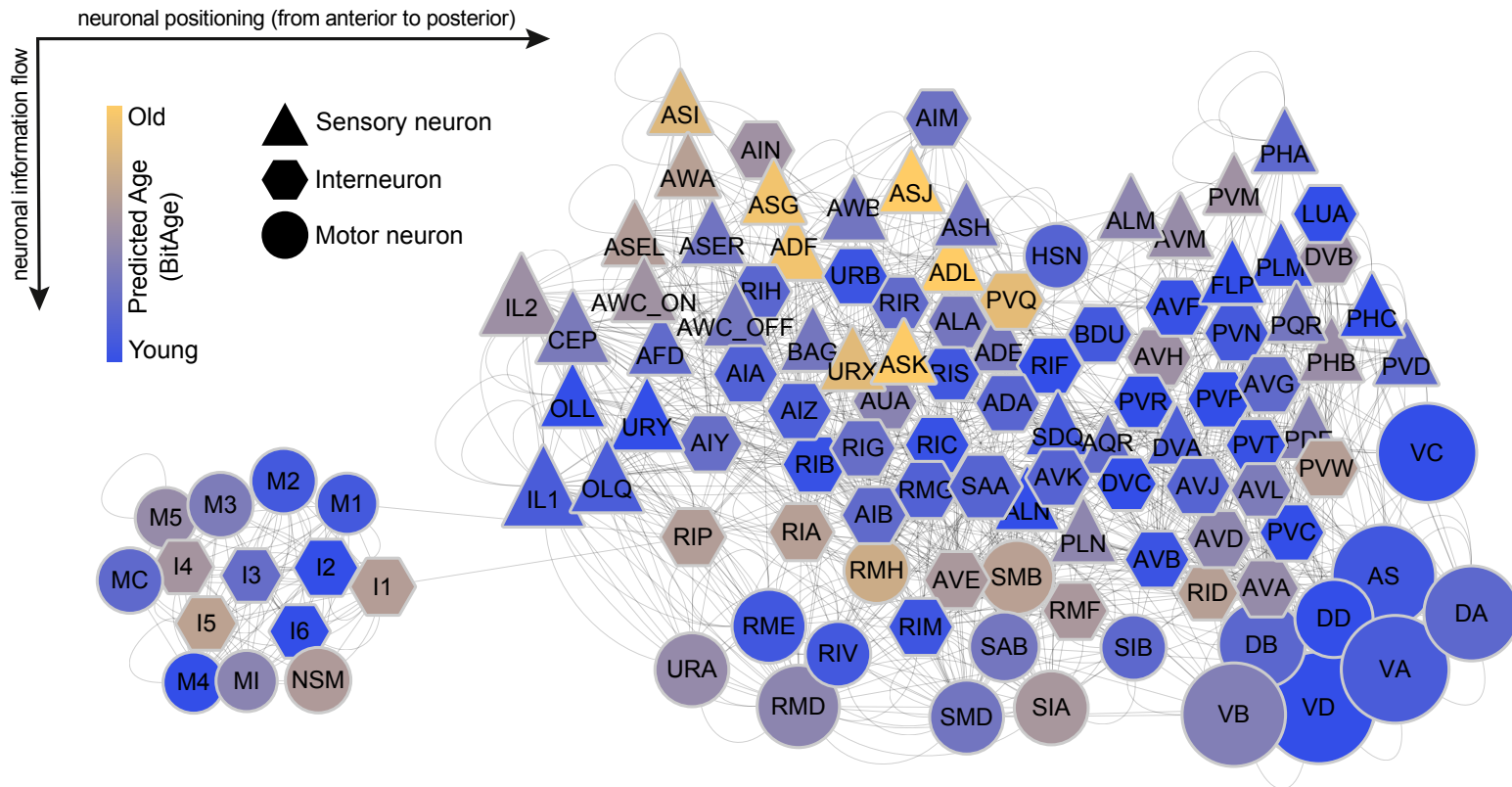


C

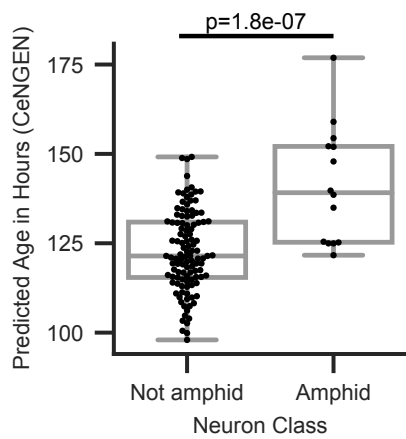
healthy mildly damaged severely damaged



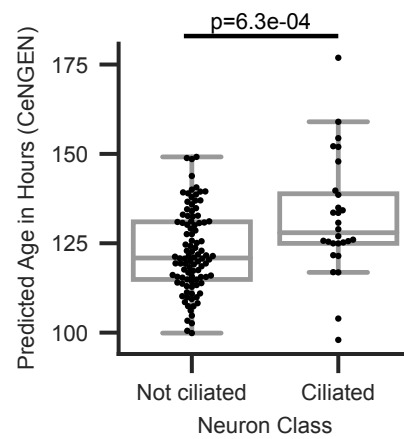
A



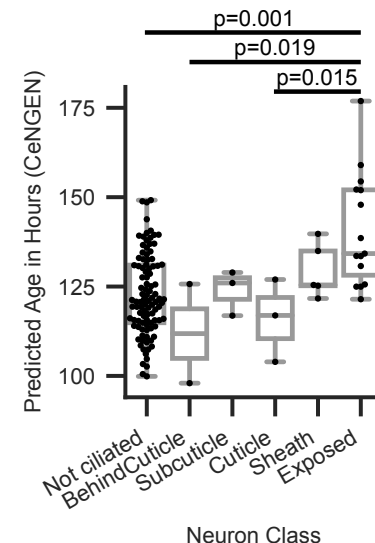
B



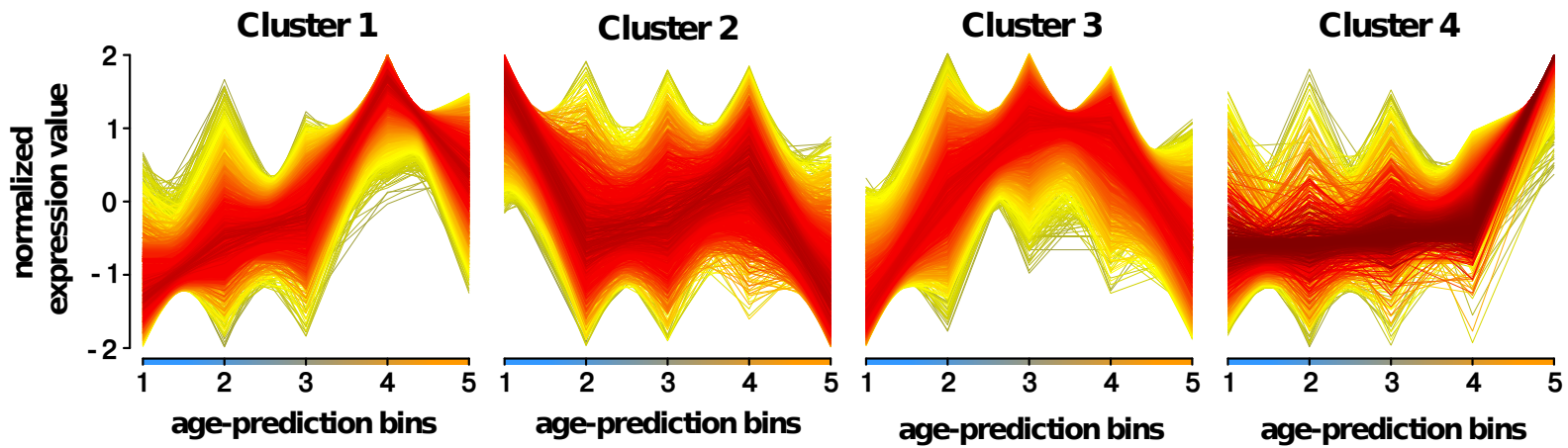
C



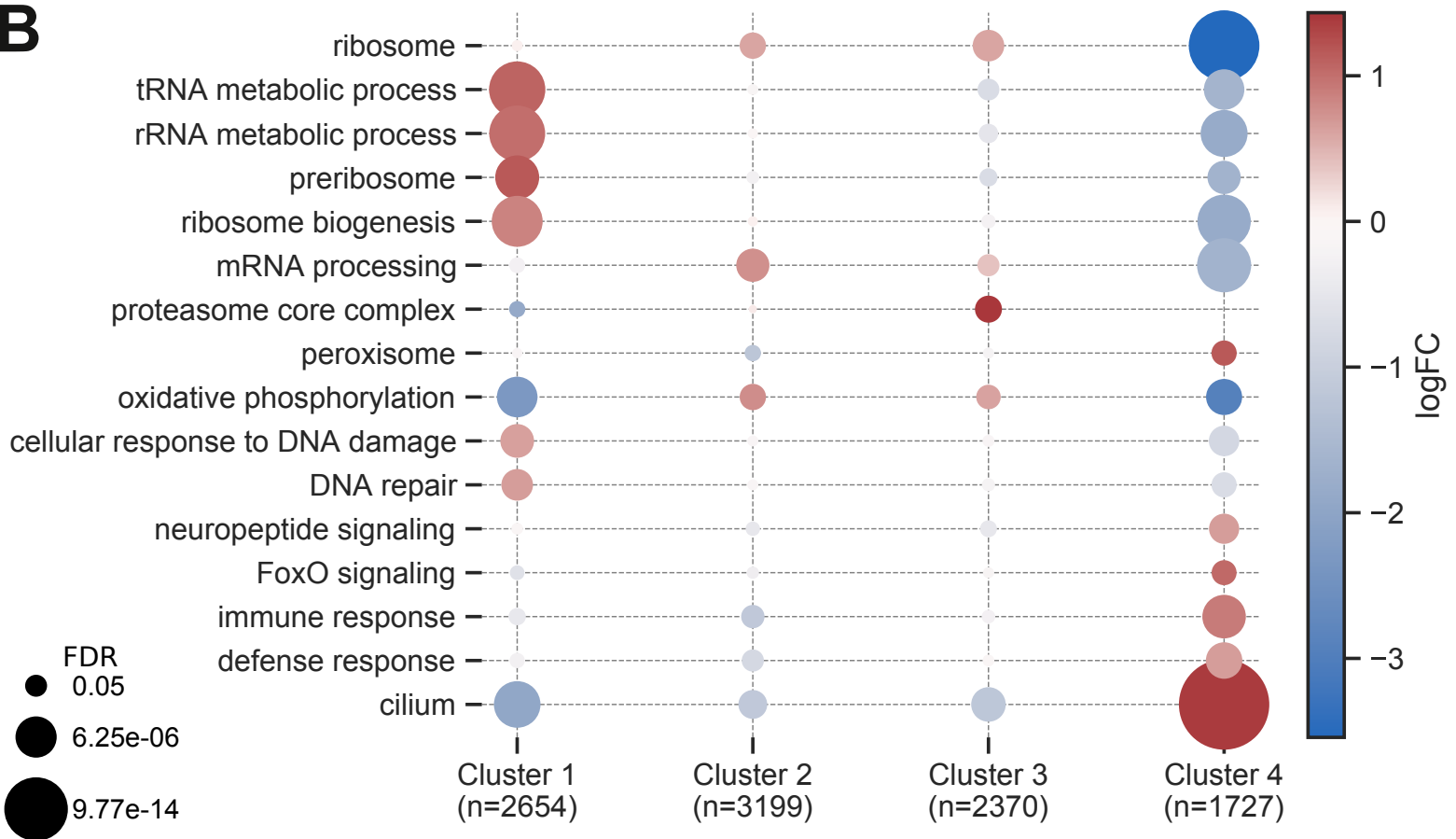
D



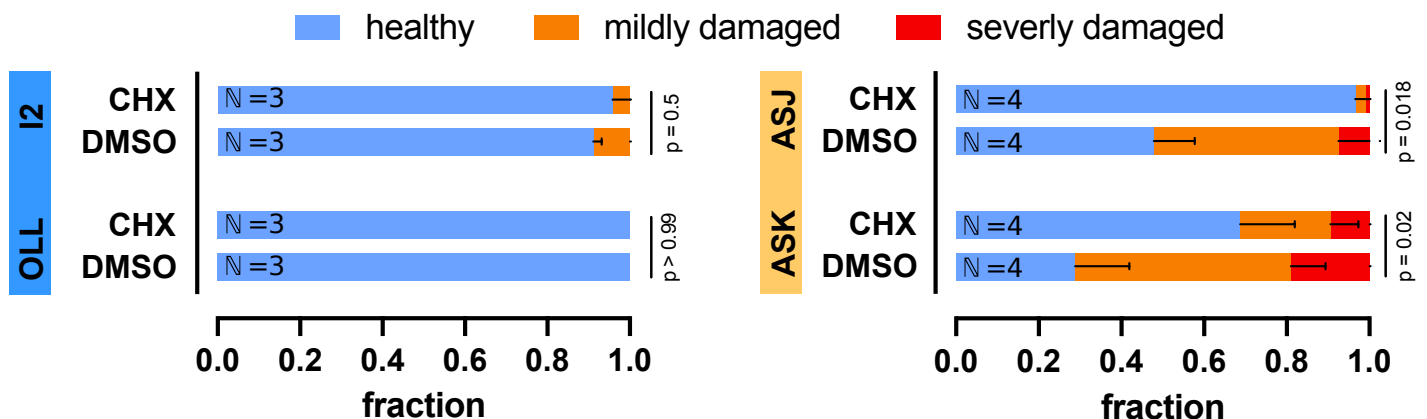
A

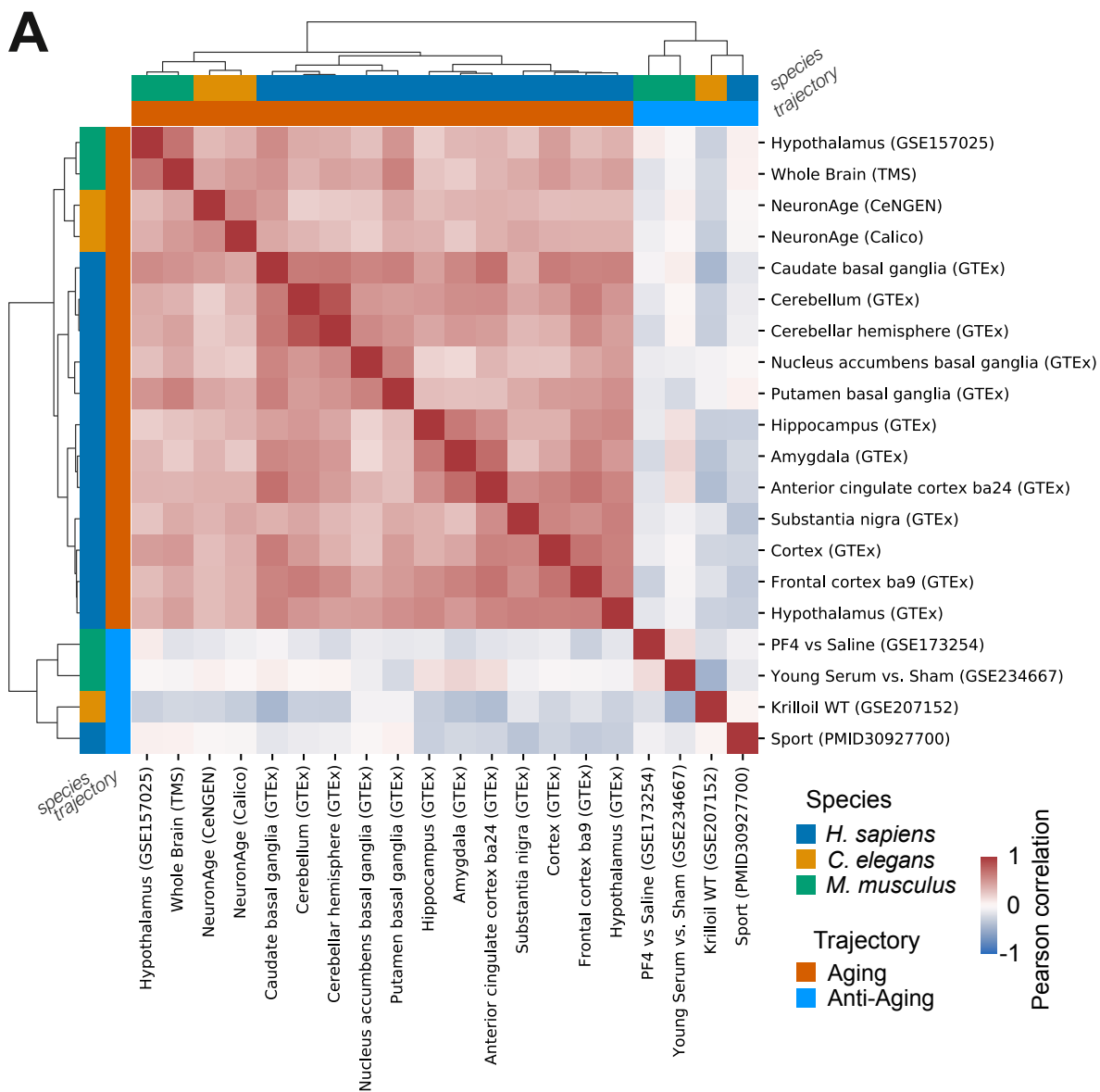


B

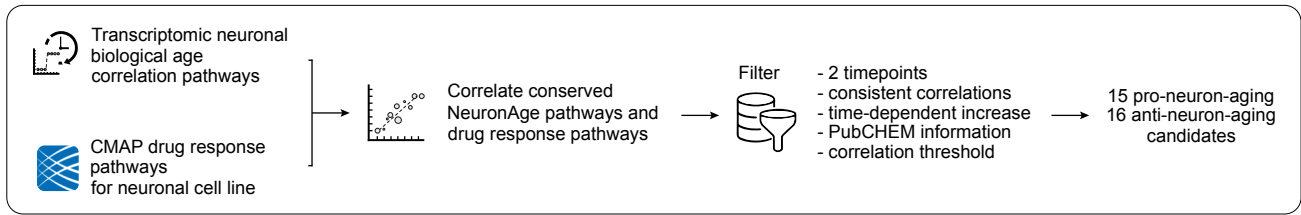


C

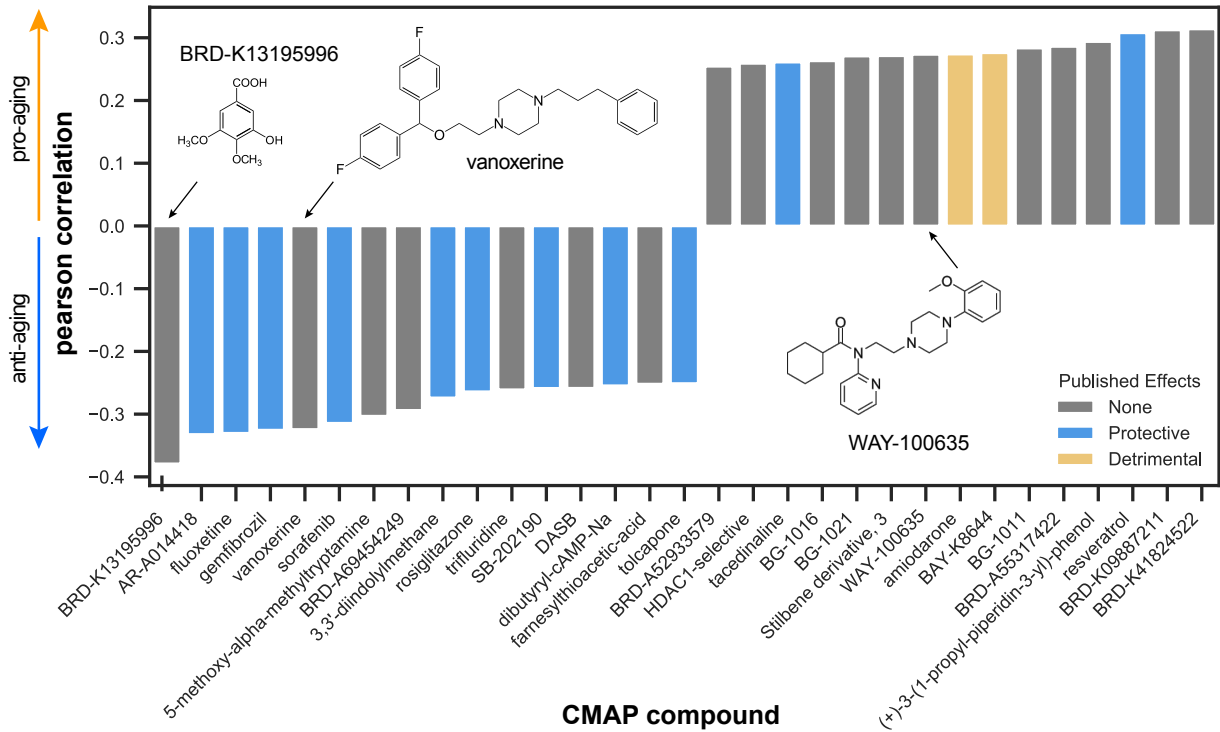




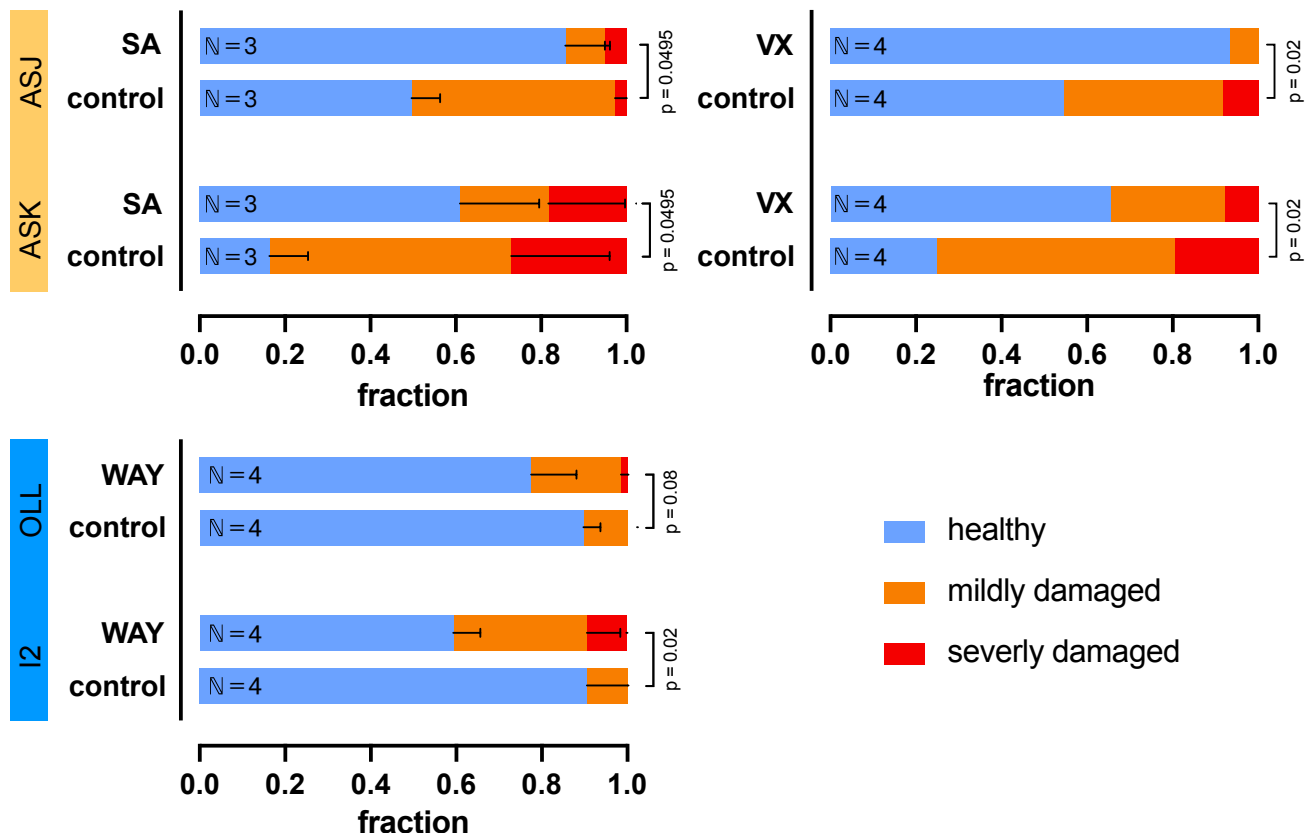
A



B



C



1 Supplementary Figure Legends

2 Supplementary Figure 1

- 3 A) Distribution of transcriptomic age predictions. The 67 neurons of the Calico day 1 dataset were
4 predicted with BitAge and sorted by their predicted age. The x-axis shows the rank of the
5 prediction in ascending order, the y-axis the predicted age.
- 6 B) BitAge predictions on the CeNGEN and the Calico day 1 dataset are highly correlated (Pearson
7 correlation 0.64, p-value 5.8e-09). The x-axis shows the BitAge predictions of the CeNGEN
8 dataset, the y-axis BitAge predictions of the Calico day 1 dataset. 67 neurons are plotted. The
9 regression model fit with a 95% confidence interval (shadowed area) is shown.
- 10 C) Predictions with BitAge and a stochastic data-based clock on the CeNGEN dataset are highly
11 correlated (Pearson correlation 0.65, p-value 5.5e-17). The x-axis shows the BitAge predictions
12 of the CeNGEN dataset, the y-axis stochastic data-based clock predictions of the CeNGEN
13 dataset. All 128 neurons are plotted. The regression model fit with a 95% confidence interval
14 (shadowed area) is shown.

15 Supplementary Figure 2

- 16 A) Amphid neurons are predicted to be significantly older than non-amphid neurons in the Calico
17 day 1 dataset. Two-sided t-test p-value: 4.2e-08.
- 18 B) Amphid neurons are predicted to be significantly older than non-amphid neurons in the
19 CeNGEN dataset with a stochastic data-based clock. Two-sided t-test p-value: 7.2e-22.
- 20 C) Amphid neurons express significantly more neuropeptides than non-amphid neurons. Two-
21 sided t-test p-value: 5.95e-12.
- 22 D) Amphid neurons express significantly more receptor genes than non-amphid neurons. Two-
23 sided t-test p-value: 8.45e-09
- 24 E) The number of neurotransmitters is not significantly different in amphid and non-amphid
25 neurons. Two-sided t-test p-value: 0.94
- 26 F) The number of innexins is not significantly different in amphid and non-amphid neurons. Two-
27 sided t-test p-value: 0.053
- 28 G) The number of expressed neuropeptides (y-axis) is significantly correlated (Pearson correlation
29 0.26, p-value 3e-03) with the predicted age by BitAge in the CeNGEN dataset. The regression
30 model fit with a 95% confidence interval (shadowed area) is shown.
- 31 H) The number of expressed receptor genes (y-axis) is significantly correlated (Pearson correlation
32 0.22, p-value 1.2e-02) with the predicted age by BitAge in the CeNGEN dataset. The regression
33 model fit with a 95% confidence interval (shadowed area) is shown.

- 34 I) The number of expressed neuropeptides (y-axis) is significantly correlated (Pearson correlation
35 0.34, p-value $4.4e-03$) with the predicted age by BitAge in the Calico day 1 dataset. The
36 regression model fit with a 95% confidence interval (shadowed area) is shown.
- 37 J) The number of expressed receptor genes (y-axis) is significantly correlated (Pearson correlation
38 0.42, p-value $4.7e-04$) with the predicted age by BitAge in the Calico day 1 dataset. The
39 regression model fit with a 95% confidence interval (shadowed area) is shown.
- 40 K) The number of expressed neuropeptides (y-axis) is significantly correlated (Pearson correlation
41 0.43, p-value $4.8e-07$) with the predicted age by a stochastic data-based clock in the CeNGEN
42 dataset. The regression model fit with a 95% confidence interval (shadowed area) is shown.
- 43 L) The number of expressed receptor genes (y-axis) is significantly correlated (Pearson correlation
44 0.39, p-value $6.9e-06$) with the predicted age by a stochastic data-based clock in the CeNGEN
45 dataset. The regression model fit with a 95% confidence interval (shadowed area) is shown.
- 46 M) Fraction-plots displaying the fraction of nematodes expressing neuronal volume markers in the
47 ASI neuron categorized as 'healthy', 'mildly damaged', and 'severely damaged'. Three to four
48 cohorts were analysed, comprised of 10 – 30 individual nematodes, for every timepoint
49 indicated. Kruskal-Wallis-test was employed to test for significant differences.
- 50 N) The number of expressed innexin genes (y-axis) is significantly anti-correlated (Pearson
51 correlation -0.19 , p-value $3.7e-02$) with the predicted age by BitAge in the CeNGEN dataset.
52 The regression model fit with a 95% confidence interval (shadowed area) is shown.
- 53 O) The number of expressed innexin genes (y-axis) is not-significantly anti-correlated (Pearson
54 correlation -0.2 , p-value $1.1e-01$) with the predicted age by BitAge in the Calico day 1 dataset.
55 The regression model fit with a 95% confidence interval (shadowed area) is shown.
- 56 P) The number of expressed innexin genes (y-axis) is not-significantly anti-correlated (Pearson
57 correlation -0.12 , p-value $1.9e-01$) with the predicted age by a stochastic data-based clock in
58 the CeNGEN dataset. The regression model fit with a 95% confidence interval (shadowed area)
59 is shown.
- 60 Q) The number of total synapses (y-axis) is not-significantly anti-correlated (Pearson correlation -
61 0.04 , p-value $6.6e-01$) with the predicted age by BitAge in the CeNGEN dataset. The regression
62 model fit with a 95% confidence interval (shadowed area) is shown.
- 63 R) The number of total synapses (y-axis) is not-significantly anti-correlated (Pearson correlation -
64 0.21 , p-value $1.3e-01$) with the predicted age by BitAge in the Calico day 1 dataset. The
65 regression model fit with a 95% confidence interval (shadowed area) is shown.
- 66 S) The number of total synapses (y-axis) is not-significantly anti-correlated (Pearson correlation -
67 0.07 , p-value $4.6e-01$) with the predicted age by a stochastic data-based clock in the CeNGEN
68 dataset. The regression model fit with a 95% confidence interval (shadowed area) is shown.

- 69 T) Ciliated neurons are predicted to be significantly older than non-ciliated neurons in the Calico
70 day 1 dataset. Two-sided t-test p-value: $9.04e-05$
- 71 U) Ciliated neurons are predicted to be significantly older than non-ciliated neurons in the
72 CeNGEN dataset with a stochastic data-based clock. Two-sided t-test p-value: $2.29e-14$
- 73 V) Ciliated neurons are divided into 5 classes depending on where its cilia terminate. Neurons
74 with exposed cilia are significantly older than non-ciliated neurons or neurons which cilia
75 terminate in the cuticle or behind the cuticle in the Calico day 1 dataset (one-way ANOVA p-
76 value: $1.79e-08$, with a post-hoc Tukey test).
- 77 W) Ciliated neurons are divided into 5 classes depending on where its cilia terminate. Neurons
78 with exposed cilia are significantly older than non-ciliated neurons or neurons which cilia
79 terminate in the cuticle or behind the cuticle in the CeNGEN dataset with a stochastic data-
80 based clock (one-way ANOVA p-value: $6.6e-18$, with a post-hoc Tukey test).

81

82 Supplementary Figure 3

- 83 A) Output of the Dmin function of the Mfuzz R package. Soft clustering for cluster numbers
84 ranging from 2-12 (x-axis) were calculated with Dmin. For each cluster number, Dmin calculates
85 the distance between the centroids of the clusters (centroid distance) and reports the
86 minimum centroid distance across 3 repetitions (y-axis). The optimal cluster number is
87 estimated from the “elbow” of the plot (indicated by a dashed line), i.e. the cluster number
88 which shows a sharp decline in the minimum centroid distance (cluster number=4).

89

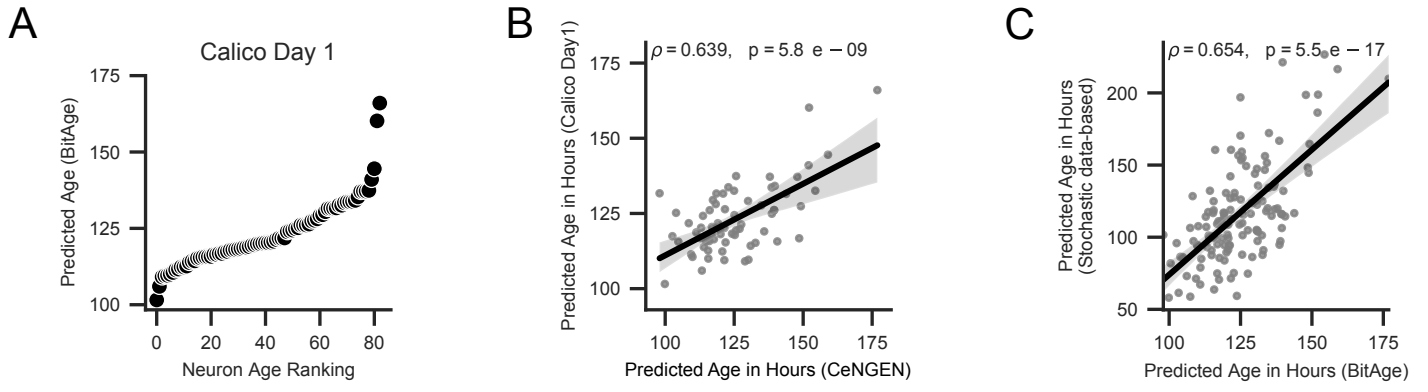
90 Supplementary Figure 4

- 91 A) Fraction-plots displaying the fraction of nematodes expressing neuronal volume markers in the
92 OLL neuron categorized as ‘healthy’, ‘mildly damaged’, and ‘severely damaged’ that were
93 treated with 2.5 mM syringic acid (SA) or 10 nM vanoxerine (VX) for 24 h. Three to four cohorts
94 were analysed, consisting of 10 – 25 individual nematodes. Kruskal-Wallis-test was employed
95 to test for significant differences.

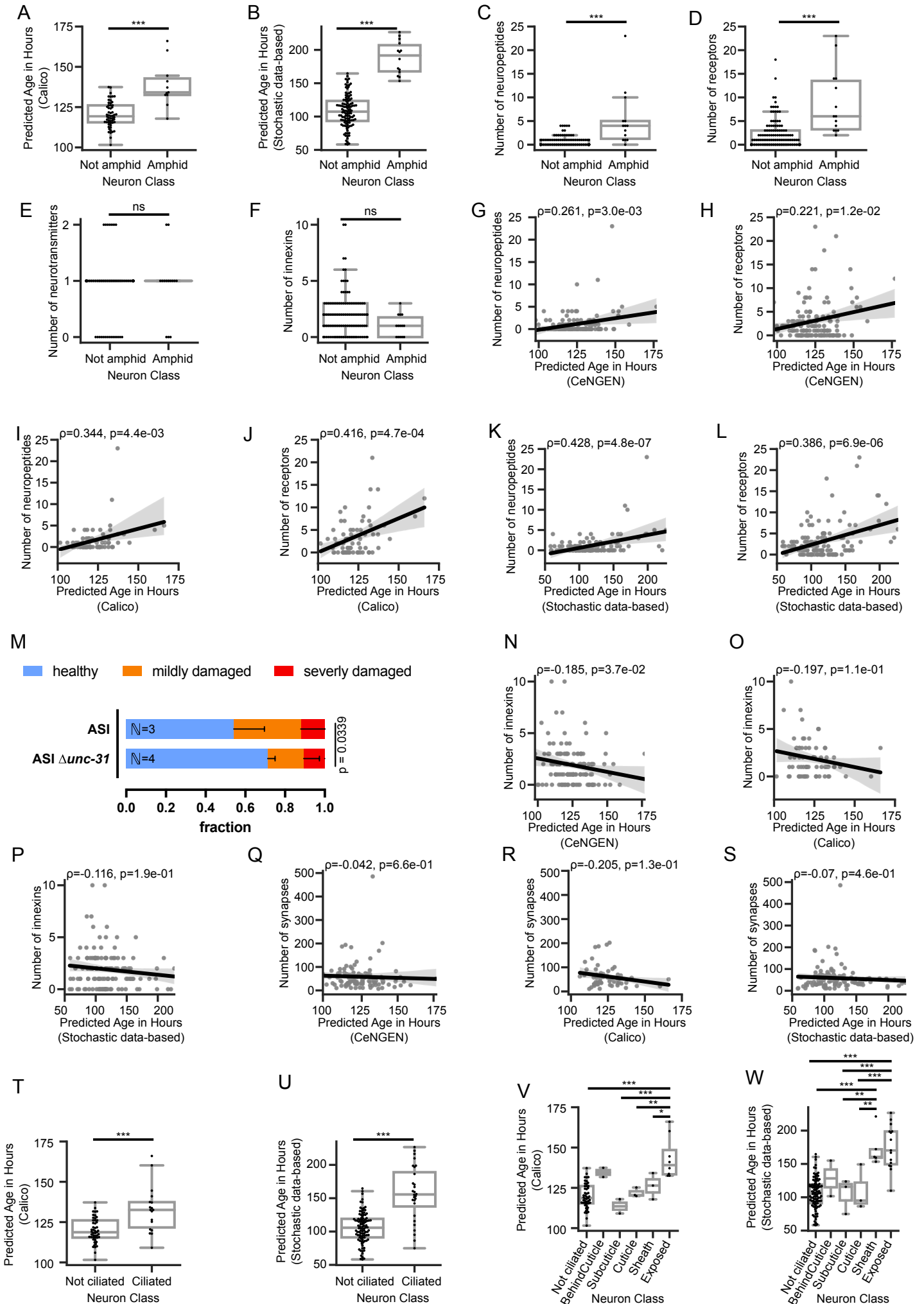
96

97

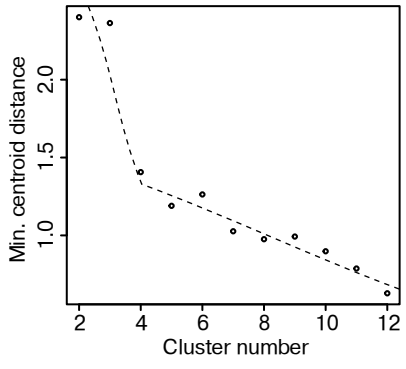
- Supplemental Figure 1 -



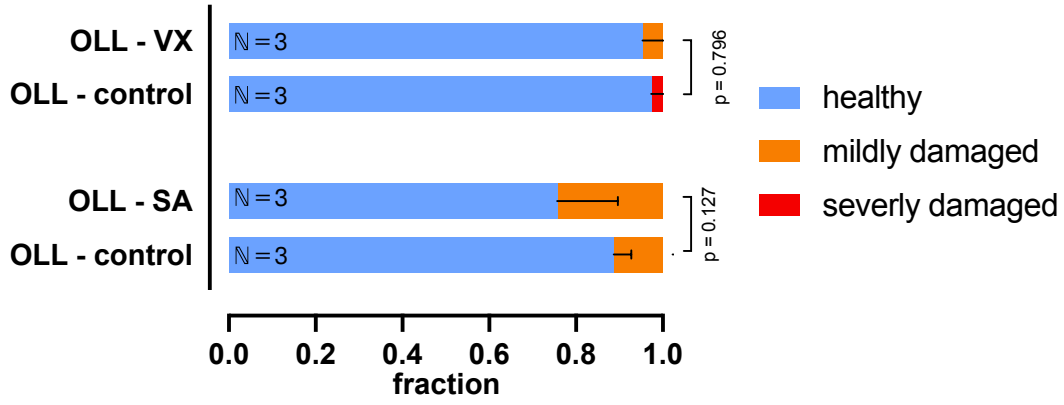
- Supplemental Figure 2 -



A



A



5 Discussion

5.1 Aging Clocks

Currently, the aging clock field faces three major ongoing challenges:

- 1.) Constructing and validating accurate biological aging clocks¹⁸²
- 2.) Understanding their underlying mechanisms^{183,184}
- 3.) Utilizing them in identifying and evaluating longevity or health-span interventions⁵⁰

5.1.1 Constructing and Validating Accurate Biological Aging Clocks

Aging clocks, and more broadly aging biomarkers, should be highly accurate in predicting biological age, robust against technical and biological variability, and ideally transfer to different populations and potentially even species¹⁸². Currently, they are mostly limited by the usage of cross-sectional data¹⁸⁵, highlighting the urgent need for more longitudinal studies¹⁸². To assess their validity and robustness they should be validated in diverse datasets¹⁸². Genomics data is known to lack diversity, with a disproportional amount of data coming from Caucasian populations¹⁸⁶, leading to potentially lower generalizability and biases against unrepresented populations^{187–189}. Similarly, sociodemographic and socioeconomic characteristics are currently mostly overlooked¹⁹⁰. In addition to the lack of validation in these groups, recent validation efforts have revealed that most epigenetic aging clocks are affected by technical variation (with variations up to 8.6 years in replicates)¹²⁴ and unwanted biological variation associated with the circadian rhythm¹⁹¹.

With BitAge¹⁹² we developed the first highly accurate biological age predictor for transcriptomic data (**Chapter 2**). We introduced the concept of binarization to remove unwanted variation in the data to subsequently improve the prediction accuracy. By leveraging existing lifespan data with temporal rescaling and survivor-bias correction for *Caenorhabditis elegans* RNA-seq data, we demonstrated that biological age prediction of an organism with transcriptomic data is highly accurately possible.

A survivor-bias correction is not only relevant for whole-worm populations of *Caenorhabditis elegans*, but also generally for aging clocks that are built on chronological age. With advancing age there is a positive selection for individuals with higher physiological capacities, as more frail subjects die earlier¹⁹³, leading to biases in first-generation aging clocks. Even second-generation aging clocks that are trained on the biological age or mortality data from a variety of datasets might still contain a similar bias. PhenoAge⁷², for example, first built a Cox penalized regression model on blood biochemistry data from NHANES III¹⁹⁴, and then used this model on data from the InCHIANTI study¹⁹⁵. The NHANES III data were collected between 1988 and 1994, while the InCHIANTI data were collected between 1998 and 2000. Subjects of the same chronological age of the InCHIANTI study were therefore up to 12 years

younger at the time of NHANES III data collection. This temporal difference suggests that subjects of the InCHIANTI study may have been exposed to beneficial factors, i.e. medical progress, for a relatively longer time period than subjects of the same chronological age in the NHANES III cohort⁴⁸. Consequently, the subjects of the NHANES III cohort might be at a greater hazard risk, because they were born earlier, compared to subjects of the same chronological age in the InCHIANTI study⁴⁸. It is not known whether this second type of bias has a strong impact on the prediction accuracy, but it shows potential accuracy limits of clocks built on various datasets potentially generated over decades apart.

Our BitAge clock has been independently used and validated in a variety of studies: BitAge was retrained and applied to a human transient reprogramming time-course dataset, indicating a rejuvenation trend¹⁹⁶. It identified positive effects of krill oil on the predicted biological age, mirroring measured health-span benefits¹⁹⁷. The human fibroblast BitAge clock failed to predict an aging trajectory in peripheral blood mononuclear cell RNA-seq data from healthy people or people with HIV¹⁹⁸, possibly due to cell-type differences. It identified an increased transcriptomic age of *Caenorhabditis elegans* worms expressing human A β and Tau as a model of Alzheimer's disease¹⁹⁹. It predicted a significantly higher transcriptional age of *hlh-30* mutants in an adult diapause model of *Caenorhabditis elegans* that could be rescued by an additional *daf-1* mutation, mirroring the detrimental effects of a loss of *hlh-30* that can be partially rescued by a *daf-1* mutation²⁰⁰. And lastly, binarization of gene expression has been adapted and demonstrated to improve single-cell age classification²⁰¹. These results corroborate that the *Caenorhabditis elegans* BitAge clock is a robust predictor of the biological age and that the concept of binarization is something to explore as a preprocessing step in machine learning approaches. Despite these corroborating results more extensive validation on newly generated data would be needed to see whether all possible aging trajectories have been faithfully covered. We trained the *Caenorhabditis elegans* clock using all available adult RNA-seq data paired with corresponding lifespan data available at the time of publication. This dataset encompassed a wide range of mutants, gene knockdowns, and treatments. However, it is important to acknowledge that unknown aging trajectories might exist that are not captured by the current model, leading to inaccurate predictions. Further validation efforts will help identify potential mis-predicted conditions, offering valuable insights into novel aging trajectories. Future research might also improve BitAge to further strengthen its generalizability:

- 1.) New RNA-seq samples with corresponding lifespan data after the submission of the original BitAge publication should be included.
- 2.) The effect of RNA-seq library size on binarization should be investigated and potentially all samples should be subsampled before binarization.

- 3.) The effect of different library preparation methods, i.e. polyA enrichment vs. riboMinus depletion of rRNA vs. total RNA, should be investigated and potentially only samples from one preparation method should be included.
- 4.) A clock version without any acute treatments, e.g. without heat shock or pathogen exposure, might be beneficial for “natural” aging prediction.
- 5.) A clock without minimizing the number of genes might reduce overfitting.
- 6.) The second correction step that we apply to correct for a possible survivor-bias is currently approximated by a complicated function and could be improved, both in clarity and potentially accuracy with a simulation approach.
- 7.) The model assumption for the second correction step, i.e. that biological age is normally distributed with the same standard deviation irrespective of the chronological age, should be tested with single worm RNA-seq and could lead to adaptations and improvements of the correction.
- 8.) The current proof-of-principle human BitAge clock needs more training data and should be extended to other cell- and tissue-types.

The stochastic data-based clocks on the other hand were initially not developed as tools for measuring the biological age of organisms or individuals, but rather as a tool to elucidate the underlying mechanism of aging clocks (**Chapter 3**)²⁰². However, we demonstrated that the idea of accumulating stochastic variation enables the construction of aging clocks across organisms and data modalities. Importantly, the concept of accumulating stochastic variation could enable the development of aging clocks even in scenarios where limited data is available, potentially requiring as little as a single biological sample.

We have validated both the transcriptomic stochastic data-based clock and BitAge on pseudo-bulk data from a neuronal *Caenorhabditis elegans* single-cell RNA-seq dataset, indicating that our clocks are robust and even single neuron biological age predictions are possible (**Chapter 4**).

Cell-type composition changes are a confounding factor in the development of aging clocks^{203,204}. If BitAge were cell-type confounded one could expect that neurons that degenerate earlier would be predicted younger, as the cell-type specific genes of the fast-degenerating neurons would potentially cease to be expressed in older worms. In **Chapter 4** we scored neurodegeneration by macroscopic aberrations on the neurites, which might not coincide with a complete transcriptional stop of these cells. Nevertheless, we saw faster degeneration of older neurons indicating that BitAge identified a more general aging trajectory that is not cell-type confounded. The stochastic data-based clock is not directly trained on biological data (except the ground state/starting point of the simulations) and is therefore less likely to be affected by cell-type composition changes. Indeed, the transcriptomic

stochastic data-based clock predictions correlated significantly with the BitAge predictions and neurodegeneration in **Chapter 4**. For the stochastic DNA methylation clock, we additionally demonstrated that correcting for cell type composition changes does not affect the resulting Pearson correlations strongly.

In conclusion, both BitAge and the stochastic data-based clock have been validated in several ways, indicating their robust predictions, which enable their usage in a variety of settings.

5.1.2 Understanding Their Underlying Mechanisms

The underlying mechanisms of aging clocks are a current focus of research. A recent study used several epigenetic clocks to investigate whether specific aging hallmarks affect the predicted age in multiple cell types¹⁸³. Interestingly, the results indicate that damage-induced senescence, and DNA double strand breaks might not affect the epigenetic age, while mitochondrial function and Rapamycin treatment did¹⁸³. Conversely, a study applying an epigenetic clock on four different progeroid mouse models (*Ercc1*, *LAKI*, *Polg*, and *Xpg*) suggested that progeroid mice with deficiencies in DNA repair (*Ercc1*, or *Xpg*), but not mice suffering from lamina defects (*LAKI*) or mitochondrial DNA mutation accumulation (*Polg*) showed an increased epigenetic age²⁰⁵. This result would suggest a connection between stochastic DNA damage and epigenetic drift. The recent pan-mammalian clock study, on the other hand, found a set of conserved CpG sites that gain methylation with age, are enriched in Polycomb repressive complex 2 (PRC2) -binding sites, and involved in the regulation of developmental gene expression¹²⁷. The authors argue that the deterministic features of these age-related changes are evidence that aging is not only a consequence of random cellular damage, but the continuation of developmental processes¹²⁷. In **Chapter 3** we showed, however, that the pan-mammalian clock results are largely reproducible with our stochastic data-based clocks and that accumulating stochastic variation is sufficient to build a clock, whose predictions are correlated with the chronological as well as biological age across species. Our results therefore support the theory that random cellular damage and variation is sufficient for aging. Our results are also in line with the epigenetic maintenance system theory, i.e. that the tightness of maintenance corresponds to the tick rate of aging clocks⁹⁵. While the tightness of the maintenance system at birth is genetically determined, we propose that the aging process is not driven by the continuation of any (developmental) process, but rather the implications of an imperfect maintenance system, i.e. a passive consequence. Species that evolved later sexual maturity must also have evolved a maintenance system that is sufficient to ensure reproduction at a later time. This improved maintenance leads to a slower accrual of damage, e.g. epimutations, less stochastic noise, thereby a slower ticking of the clock, longer-lasting regulatory tightness, and ultimately longer lifespan. In **Chapter 5.2** I will go into more detail of stochastic noise in biological systems and ways to measure it.

5.1.3 Utilization in Identifying and Evaluating Longevity Interventions

Ageing clocks hold immense promise in advancing personalized healthcare and are critical tools in speeding up the identification and evaluation of longevity interventions⁵⁰. It has been suggested that a validated ageing clock can serve as a surrogate endpoint for clinical trials, expediting the identification of potential geroprotective treatments⁵⁰. Several clinical trials are ongoing with epigenetic clocks as a surrogate endpoints⁵⁰. One limitation of using an ageing clock as the surrogate endpoint is that it is not possible to estimate the biological age of the whole organism from one datatype of one organ alone, as was already noted in 1947⁴⁹. Different datatypes (DNA methylation data, transcriptomics, proteomics etc.) might capture different aspects of the biological age and different organs will age at different rates in different individuals^{206,207}. The recent plasma proteomics clock that allowed organ-specific risk assessments is an important step for improving applicability of ageing clocks¹⁶². While it is a strength of this study to be able to assess organ-specific health risks from blood plasma, providing valuable insights, it may not fully capture ageing rate differences observed across diverse tissues and cannot directly measure the biological age of different cell types. While not being discussed in the publication, I speculate that the organ-specific ageing clocks measure the rate of organ-specific cell death. It is conceivable that an unhealthy organ might exhibit higher rates of cell death, leading to the release of more organ-specific proteins into the bloodstream²⁰⁸. A healthy individual with well-functioning organs might have fewer non-blood organ-specific proteins present in blood plasma. As the function of organs generally decreases with age, this differential protein release may enable the ageing clocks to capture organ-specific ageing signatures, similar to the reported identification of tissue-specific cell death using cell-free DNA methylation pattern in blood samples²⁰⁹.

We have shown that BitAge can predict biological age differences not only in bulk whole-worm population RNA-seq data, but also in neuronal cell-type specific pseudo-bulk data (**Chapter 4**). We applied our BitAge clock on a pseudo-bulk dataset of single neuron classes from *Caenorhabditis elegans* and identified biological age differences of almost 2-fold between the youngest and oldest predicted neurons. We validated these cell-type specific predictions *in vivo* and observed that neurons with a predicted older age at the first day of adulthood degenerate faster throughout adulthood. Importantly, we demonstrated that BitAge captured biological age differences despite the fact that it was not trained on single-cell RNA-seq data, but whole-worm populations.

We used the identified transcriptomic neuronal ageing trajectories to identify novel neuro-protective compounds in an *in silico* drug screen that we validated *in vivo*. This demonstrates that ageing clocks can be used not only as a surrogate endpoint but also as a screening strategy. This underscores the versatility and the potential of ageing clocks in accelerating the discovery and evaluation of interventions aimed at promoting health-span and longevity.

5.2 Stochastic Biological Variation

Biochemical systems, such as enzymatic reactions, are inherently stochastic in nature, driven by random movement and collisions between molecules^{210,211}. Despite this, most developmental processes are deterministic²¹², with gene regulatory networks²¹³ and epigenetic memory²¹⁴ playing a key role for the cells to ignore stochasticity and act in a deterministic fashion. Importantly, stochastic cell fate decisions, such as olfactory receptor gene expression in the mouse olfactory sensory neurons²¹⁵, exist²¹⁶. Aside these stochastic cell fate decisions that are due to overwhelming number of choices, e.g. olfactory receptor genes, generating cell-to-cell variability in isogenic cell populations is important for the ability to respond to different environmental cues²¹⁷. It was suggested that pluripotent cells will be noisier than differentiated cells to enable higher plasticity and flexibility, and that noise should be highest during cell fate transitions²¹⁷. However, this is not entirely consistent with previous reports demonstrating increased cell-to-cell variability with age²¹⁸. Indeed, it has been shown that strong chromatin regulation can lead to plasticity without noise²¹⁹, and that noise and plasticity are largely independent traits for core cellular components²²⁰.

5.2.1 Stochastic Epigenetic Variation

Age-related stochastic DNA methylation drift could even restrict plasticity and lead to phenotypes such as stem cell exhaustion⁴⁶, and it has been suggested to be a determinant of mammalian lifespan^{221,222}. Early work defined this epigenetic drift as a deficient methylation metabolism (similar to an imperfect maintenance system)²²³. Subsequent definitions included age-related changes due to environmental or stochastic causes^{88,224}. While epigenetic drift is often defined to be the stochastic component due to an imperfect maintenance system, and global decrease in stability and precision of DNA methylation with age⁴⁶, it has also been defined as the collection of changes that are not common across individuals²²⁵. Both definitions are indeed used interchangeably^{225,226}, and it has been argued that these sites affected by stochasticity are not useful for epigenetic clocks²²⁵. However, as we have shown in **Chapter 3**, stochastic changes indeed do allow for consistent pattern in the data that can be learned by an age predictor and thereby allow the prediction of the chronological and biological age.

In bulk DNA methylation samples, environmental and stochastic changes overlap, further complicated by different sources of heterogeneity within the samples²²⁷: cell-type heterogeneity or contamination and potential cell-type composition changes^{228,229}, and allele-specific methylation^{230,231}. Several methods have been defined to distill specific sources of within-sample heterogeneity²²⁷:

- 1.) The *proportion of discordant reads (PDR)* method quantifies locally disordered methylation as the number of discordant reads, i.e. reads where not all CpG sites where

either methylated or unmethylated, divided by the total amount of reads²³². It quantifies variation within a read, especially induced by stochastic effects.

- 2.) *Epipolymorphisms* are computed with the Tsallis entropy of epiallele frequency for 4-CpG windows to measure variation among reads²³³.
- 3.) Similarly, *methylation entropy* quantifies variation among reads and is computed by Shannon entropy (a specific case of Tsallis entropy)²³⁴.
- 4.) The *proportion of disordered neighbor pairs (regional disorder)* method quantifies within each read the proportion of CpG neighbors with differing methylation state and averages this across a 200bp window.²³⁵
- 5.) The *fraction of discordant read pairs (FDRP)* method is employed to quantify heterogeneity at each individual CpG site by measuring the similarity of DNA methylation patterns in pairwise read comparisons²²⁷.
- 6.) *MeConcord* uses Hamming distance and similar to *FDRP* quantifies discordant read pairs but enables the usage of higher sequencing coverage²³⁶.

Each of these methods has its own strengths and application scenarios²²⁷. The *proportion of discordant reads*, *epipolymorphisms*, *methylation entropy*, and *regional disorder* are increasing with age, suggesting stochastic processes underlying age-related DNA methylation changes^{92,237-240}. Conversely, an analysis with *MeConcord* on an aging and a replicative senescence dataset found a higher proportion of disordered reads in young, respective non-senescent, cell populations, while the fraction of uniformly/ordered regions increased with age and senescence²³⁶. Similarly, a longitudinal between-sample analysis of DNA methylation heterogeneity suggested that only 10% of CpG sites might be stochastically changing with age, while the trajectories of the remaining 90% are determined by genetics and environment²⁴¹. It is important to note for the latter analysis that CpG sites were defined as stochastic if the signal-to-noise ratio of an individual over the aging-time-course is lower than the signal-to-noise ratio for all individuals and timepoints pooled together. The signal-to-noise ratio for individuals was defined as the average over the time-course of the ratio of a regressed methylation level of a linear regression model and the absolute residuals of the same model²⁴¹. The signal-to-noise ratio of the pooled dataset was defined as the average ratio of all individuals of the average methylation value over a 10-year sliding window and the deviation from it²⁴¹. By definition the pooled signal-to-noise ratio is therefore expected to be lower, as the deviation from each individual from the average methylation value of a 10-year window is expected to be higher than the residuals computed for each year from a regression model. It is therefore not surprising that most CpG sites are found to be non-stochastic with this definition. Indeed, a recent single-cell DNA methylation analysis found that 92% of the 502 age-related CpG sites for which sufficient coverage was available, behaves stochastically, as defined by the Pearson inter-cell correlation coefficient²⁴².

The approach we took in **Chapter 3** is different from these methods in that we first used artificially accumulating stochastic variation to build an age predictor that then was applied to biological samples. We do not calculate the disorder or entropy within a region or between reads, as described above. Instead, we quantify biological age directly by using a clock that learned the artificially induced accumulation of stochastic variation pattern. Interestingly, regional disorder, calculated by the *proportion of disordered neighbor pairs* method, has recently been used to build an epigenetic aging clock for a small mouse dataset, reaching similar accuracy as standard DNA methylation clocks²³⁵, and it was shown that the rate of the age-related regional disorder increase associates negatively with maximum lifespan of species²⁴³. This result corroborates our analyses, as regional disorder is thought to be largely induced by stochastic processes²³⁵. While both approaches quantified biological age using patterns of stochasticity our method has the advantage that it works on an individual CpG-level and only requires one single biological sample as the starting point of the simulations for the training data, as it is trained to predict how often stochastic variation was added to the ground state. Our biologically-hypothesis-driven simulations are based on the imperfect maintenance system and stochastic epimutations, and are therefore also alleviating potential problems with cell-type composition changes in the training data. By directly modeling this imperfect maintenance system, our approach may provide deeper insights into the underlying mechanisms of aging clocks, and aging-related epigenetic changes, especially stochastic epimutations.

It has been suggested that stochastic epimutations reflect errors during stem cell division, that species-specific rates of methylation drift reflect stem cell turnover differences, and that inflammation increases, while caloric restriction decrease this rate²⁴⁴. Interestingly though, there is evidence that human embryonic stem cells preserve their epigenetic state not by copying epigenetic information during replication, but by balancing methylation turnover rates²⁴⁵. Somatic cells on the other hand transmit more epigenetic memory during replication, which leads to higher persistence of random epimutations and subsequently DNA methylation drift²⁴⁵. Somatic cells have lower methylation turnover rates than stem cells, but are highly context-specific with methylation loss rates being dependent on replication timing, while methylation gain rates being correlated with nucleosome occupancy and lamina-associated sites²⁴⁵. Earlier work has modeled site-specific maintenance rates, revealing that average DNA methylation levels can be modeled using these rates and that site-specific rates for methylation loss and gain exist, contributing to our understanding of epigenetic dynamics^{246,247}. Indeed, random epimutations, the balance of these site-specific maintenance rates and clonal transmission of epigenetic memory gives an explanation for age-related hypomethylation in late replicating domains²⁴⁸, and hypermethylation in Polycomb-bound CpG islands^{245,249}. Polycomb-bound regions are cell-composition-change-independently enriched in age-related variably methylated positions (aVMPs) in blood²⁵⁰. This cell-composition-independent site-specific increase in

epigenetic stochasticity is in line with an information-theoretic view outlining context-specific energy-requirements and supply for epigenetic maintenance²⁵¹.

Polycomb repressive complex 2 (PRC2) is an important maintenance enzyme that is involved in establishing and stabilizing cell fates²⁵², and has a complex relationship with DNA methylation in CpG islands^{253,254}. Actively transcribed genes are unmethylated at the promoter, but might additionally be methylated at flanking regions and the gene body to prevent repressive PRC2 binding. Conversely, inactive genes have a methylated promoter, or are bivalently repressed by PRC2-binding at unmethylated promoter regions²⁵³. This bivalency has been shown to be important for maintaining epigenetic plasticity by protecting against irreversible silencing²⁵⁵. The age-dependent increase of methylation in these bivalent CpG sites is a universal biomarker of cellular aging¹²⁶, and the average rate of methylation change in especially these bivalent promoter regions is negatively associated with maximum lifespan of a species²⁵⁶. The state of these unmethylated PRC2-bound bivalent promoters has to be copied with every cell division, i.e. the methylation status of all CpGs, PRC2-binding itself, and the bivalent chromatin state (H3K4me3 and H3K27me3) has to be maintained²⁵⁷. As no maintenance system is 100% accurate, it is therefore conceivable that with mitosis gradually errors occur that will lead to a slow but steady loss of bivalency, increase of methylation and decrease of PRC2-binding. Interestingly, PRC2 is recruited to DNA double-strand breaks which might additionally put strains on the faithful maintenance of bivalent promoter sites²⁵⁸. More generally, unwanted stochastic DNA methylation changes are potentially induced every-time the epigenome has to be maintained. During replication, adaptation to an external stimulus, or repair of DNA damage. It has also been suggested that somatic mutations lead to epimutations not only at the mutated site, but broadly within 10 kilobases around the mutated DNA base²⁵⁹. As we have shown in **Chapter 3**, these unwanted stochastic DNA methylation changes underly current epigenetic clocks and are the footprint of the potential underlying cause of aging, i.e. stochastic damage and an imperfect maintenance system. Our results suggest that the clock genes or CpG sites themselves might not be causally related to aging, and might indeed be rather unimportant, as potentially those sites most strongly affected by accumulating stochastic variation are those least well maintained and therefore potentially those that are less relevant for survival. Our results don't rule out that causal clock genes or CpG sites exist. Indeed, a recent study leveraged epigenome-wide Mendelian randomization to identify CpG sites that are potentially causal for aging-related traits and used these to inform a new epigenetic aging clock¹³⁰. Our results do, however, suggest that (passive imperfect-maintenance driven) accumulation of stochastic variation is sufficient to construct aging clocks and that the maintenance system and regulatory tightness of especially also DNA methylation are prime targets for aging decelerating therapies.

5.2.2 Stochastic Transcriptomic Variation

In addition to the role of CpG methylation in modulating transcriptional variability, transcriptomic noise can be affected by various factors²⁶⁰. Transcriptomic variation stems from both cell-extrinsic and cell-intrinsic noise^{261,262}. Cell-extrinsic noise is driven by external signaling^{263,264}, or cellular state variations in volume^{265,266}, mitochondrial content²⁶⁷, or nuclear shape²⁶⁸. This cellular context can be used to predict transcript variability, demonstrating the impact cell-extrinsic noise has on gene expression²⁶⁹. Cell-intrinsic noise arises from the inherent stochastic nature of biochemical fluctuations^{262,270}, and transcriptional bursting²⁷¹. Cell-intrinsic noise is assumed to be Poisson distributed, i.e. an increase in the mean expression reduces noise, whereas a decrease in mean expression increases noise²⁷². Constitutive expressed housekeeping genes, however, have been shown to exhibit a sub-Poissonian stochasticity due to mRNA degradation mechanics, which reduces the amount of noise further²⁷³. Additionally, nuclear retention is reducing cytoplasmic variations, by mitigating cell-intrinsic noise in mammalian cells^{269,274}. Both noise components have different consequences dependent on the context: well-regulated cell cycle genes have high extrinsic noise, i.e. different cellular contexts or external signaling should affect expression levels. However, these genes should exhibit low intrinsic noise, meaning that under the same cellular state and context, the expression levels should be stable²⁷⁵.

Similar to epigenetic stochasticity, stochastic gene expression can offer advantages to cells by providing flexibility in adaptation and balancing cell fate^{260,270}. Transcriptome-wide transcriptional noise increases during developmental stages²⁷⁶. And, interestingly, the base excision repair machinery increases cell-intrinsic transcriptional noise, without altering mean expression values, to increase cellular responsiveness to fate specification signals²⁷⁷. Conversely, cell-extrinsic stochastic variation induced by DNA damage is decidedly disadvantageous¹⁶. The potential effects of an increasing number of transcription-blocking lesions can be especially observed in long genes³⁷, leading to the observed age-associated gene-length-dependent transcription decline^{38–40}. A fuzzier and less-well-regulated nucleosome landscape, potentially induced by stochastic epigenetic changes, has been proposed as the underlying cause of a distinct type of transcriptional noise characterized by a conserved age-associated increased RNA Polymerase II speed, which subsequently might contribute to an increase of circular RNAs, erroneous splicing, and increased number of mismatches²⁷⁸. Of note, it has been suggested that an aging-induced histone depletion and subsequent less-well-regulated chromatin structure in yeast might lead to a reduction of cell-intrinsic transcriptional noise until a short catastrophe phase with increased noise right before death²⁷⁹.

Indeed the age-dependent increase in transcriptional noise remains a subject of debate due to various definitions and measurement methods^{280,281}, and difficulties in dissecting the cell-intrinsic and cell-extrinsic components²⁷⁵.

Most methods define transcriptional noise based on single-cell RNA-seq:

- 1.) The ratio between biological variation, defined as a distance from each cell to the cell cluster mean, and technical variation, defined as a spike-in based distance²⁸².
- 2.) The mean of the Euclidean distance of each cell to the cell cluster mean²⁸².
- 3.) Global coordination level as a measure of the dependency between random gene sets within a single cell²⁸³.
- 4.) An integrated Bayesian hierarchical model that filters out technical variation of single-cell RNA-seq via spike-ins²⁸⁴.
- 5.) The spearman correlations of the residuals of per-gene regression models with age²⁸⁵.
- 6.) The difference from the median (or overdispersion) method, defined as the distance between the squared coefficient of variation of normalized read counts and the median expression value²⁸⁶.
- 7.) Scallop: A membership score for each cell based on its cluster assignment consistency across multiple bootstrapped iterations²⁸¹.
- 8.) The mutual information between pairs of transcription factors and target genes, as a measure of communication efficiency within the network²⁸⁷.
- 9.) Allele-specific sequencing with previously²⁶² derived formulas for intrinsic- and extrinsic noise components dissection²⁷⁵.

Using these methods various studies have demonstrated an age-dependent increase in transcriptional noise in: mouse cardiomyocytes²¹⁸; hematopoietic stem cells²⁸³; multipotent progenitors²⁸³; lymphocytes²⁸⁸; muscle stem cells¹³³; liver cells²⁸⁹; dermal fibroblasts²⁹⁰; drosophila brain cells²⁸³; human pancreatic endocrine cells²⁸²; and senescent fibroblasts²⁹¹. Mutual information of transcription-factor : target-gene pairs is decreasing with age, in line with increasing entropy in the same muscle samples²⁸⁷. Note, however, that the expression level of transcription factors is not equivalent to its activity²⁹², thereby biasing this analysis.

In addition to these single-cell RNA-seq measures of transcriptional noise, transcriptional drift, defined as the variance of the log-fold changes for genes within each functional group or pathway in bulk RNA-seq, is increasing with age²⁹³. And a network entropy measure incorporating protein-protein interaction network information with bulk RNA-seq gene expression data from humans, showed a small but significant increase with age from 25-80 years, however, a significantly smaller network

entropy for the oldest age group (92-97 years)²⁹⁴. Interestingly, a slightly adapted version of this network entropy measure saw a similar pattern in mice muscle tissue, with entropy increasing from young (3-6 months) to old (21-24 months) mice, to then slightly decrease again in the oldest age group (27-29 months)²⁹⁵.

Several studies, however, detected no or a more nuanced age-dependent effect on transcriptional noise: early quantitative reverse transcription-polymerase chain reaction experiments in mouse hematopoietic stem cells, lymphocytes, and granulocytes did not show a significantly increased noise²⁹⁶; most, but not all single-cell RNA-seq lung cell populations (including lung-resident immune cells) showed increased transcriptional noise²⁹⁷; transcriptional variation estimated with the *overdispersion method* and the *distance of each cell to the cell cluster mean method* resulted in different outcomes dependent on the cell type in kidney, lung, and spleen cells²⁹⁸; the coefficient of variation in mouse brain cell types showed no clear uniform increase in variation²⁹⁹. Moreover a comparative analysis of five of the above mentioned methods on seven published datasets did not result in conclusive results, with some datasets showing opposite results depending on the method being used²⁸¹.

These non-conclusive results were mostly generated on the whole-transcriptome level. It was already described that certain gene groups, e.g. housekeeping genes, can suppress their cell-intrinsic transcriptional noise to a sub-Poissonian level as discussed above²⁷³. And that cell-cycle genes, for example, should generally show higher cell-extrinsic noise²⁷⁵. In line with an information-theoretic view outlining context-specific energy-requirements for maintenance²⁵¹, it is therefore conceivable, that a specific subset of genes accumulates stochastic variation at a faster rate than others. Indeed it has been shown that genes lacking CpG islands (CGI- genes) change with age and potentially even drive age-related physiological degeneration³⁰⁰. These CGI- genes are getting noisier with age due to their increased euchromatinization and subsequent increased susceptibility to transcription factor binding³⁰⁰. Interestingly, this age-related increase in transcriptional noise was especially observed in CGI- genes, while genes with CpG islands (CGI+ genes) did not change with age or potentially even showed a decrease in noise³⁰⁰. CGI+ genes in this study were defined as those with transcription start sites surrounded by both experimentally validated CpG islands and with a GC content $\geq 50\%$ for at least 200 bp³⁰⁰. CGI+ genes overlap with PRC2 target genes³⁰¹, but the link between PRC2 and CGI+, as well as the mechanism of cell-type and context-specific regulation remain subject of debate³⁰². It is puzzling why CGI- genes, i.e. genes less likely to be bound by PRC2, show an age-dependent transcriptional noise increase, while CGI+ genes, i.e. genes that are enriched in PRC2-binding, are not showing the same transcriptomic age-related stochastic variation accumulation³⁰⁰. DNA methylation changes at PRC2-bound bivalent promoter regions are a universal aging biomarker^{126,256}, and as we have proposed

in **Chapter 3**, these age-related changes are potentially due to stochastic variation accumulation due to an imperfect maintenance system. Why, therefore, are the age-related (stochastic) changes not translated to the transcriptional level?

While the answer to this is elusive, I propose two non-exclusive hypotheses: First, it could lie within the complex interplay of PRC2-binding and DNA methylation: PRC2-bound CGI+ genes are unmethylated and repressed²⁵³, stochastic accumulation of DNA methylation at these promoter regions will therefore lead to a gradual loss of bivalency, but potentially still to a repression of the regulated genes. Conversely, methylated CGI+ promoter regions that stochastically lose DNA methylation over time become more susceptible to PRC2-binding²⁵³, potentially leading to increased bivalency, but still a predominantly repressive environment. CGI- promoter regions on the other hand, which are not enriched in PRC2-binding, do not have this regulation buffer, e.g. unmethylated active genes that gain stochastic DNA methylation with age are progressively getting repressed and as the DNA methylation increase is stochastic the gene repression will be as well, leading to the increased noise and therefore potentially explaining the difference in CGI- and CGI+ genes.

However, it is important to note that while this hypothesis might provide a plausible explanation for the observed differences between CGI- and CGI+ genes, there is some data that may contradict it. While the study showing the age-dependent transcriptional noise increase of CGI- genes, did not divide PRC2-bound and PRC2-unbound CGI+ genes, it showed, however, that the PRC2-associated gene repression histone modification H3K27me3 is not changing global levels³⁰⁰. Interestingly, however, a single-cell chromatin modification profiling study revealed that especially PRC2-mediated histone modifications like H3K27me3 are showing significant increases in single-cell variability in older subjects³⁰³. This PRC2-mediated H3K27me3 age-dependent noise subsequently leads to higher transcriptional noise in older subjects³⁰³. Bringing both studies together, it seems that, while no global H3K27me3 level changes can be observed, their deposition variability is changing with age. Subsequently, especially PRC2 target genes show higher transcriptional noise with aging. This contradicts the above proposed hypothesis, but could potentially be explained by the additional requirements of H3K27me3 deposition aside from the regulation of CpG methylation.

This noise increase of PRC2 target genes might be masked by the broader CGI+ category of the first study, i.e. the subset of PRC2-bound genes that increase transcriptional variation might not be enough to lead to a significant increase of transcriptional noise of the whole CGI+ gene set.

A (non-exclusive) alternative is that transcription of genes with CpG-dense regions, like CpG islands, is less impacted by single stochastic epimutations, thereby allowing for a buffer of stochasticity and more stable gene expression. This epigenetic buffer and potentially stricter regulation may also underlie the

observed correlation between higher CpG-density at promoter regions and extended lifespan across species^{304,305}.

It is essential to consider several caveats though. First, one study was conducted in mouse kidneys³⁰⁰, while the other focused on human immune cells³⁰³. Additionally, the studies employed different methods to compute transcriptional noise, and as described earlier, the outcomes of these methods might yield contrasting results. It remains interesting to see what causes the differences and what role CpG density, DNA methylation, PRC2-binding, and histone modifications play.

Moreover, these two studies underscore the importance of the specific gene sets examined in transcriptional noise analyses, as they are highly relevant to the outcome. Potentially interesting pattern might be overlooked if studied on the whole transcriptome level. Future studies should compare several transcriptional noise methods not only on the whole transcriptome, but especially also subsets like CGI-, CGI+, or PRC2-bound genes.

In **Chapter 3** we have demonstrated a novel method for measuring biological aging in an organism using artificially noisy data. By repeatedly adding normal distributed noise to a bulk RNA-seq dataset of a young *Caenorhabditis elegans* sample, we developed a predictor that shows a significant correlation with the chronological age and is capable of distinguishing lifespan differences. This approach highlights the potential utility of artificially induced noise accumulation in biological age predictions and in assessing biological aging processes. Our method offers a distinct approach to measuring age-related transcriptional noise compared to the above-mentioned methods. Especially, the Elastic Net Regression approach that is used to train the stochastic data-based clock reduces the number of genes used for the age prediction to a smaller subset of predictor genes, i.e. it does not use the whole transcriptome in the final clock. The clock therefore only retains those predictor genes that it found to be most important for the accuracy of the predictions. This might alleviate part of the problem of the non-conclusive results mentioned above, i.e. that a gene-set-specific transcriptional noise increase might be masked by the whole transcriptome. Anecdotally, the stochastic data-based clock predictor genes used in **Chapter 4**, show especially a significant overlap with H3K27me3-bound regions in data from the ChIP-Atlas database³⁰⁶. This is in line with the results showing that especially H3K27me3 variability and subsequent transcriptional variability is observed in older humans³⁰³. As training the stochastic data-based clock will result in slight changes in the predictor genes after every training process, it will be interesting to see how robust these enrichments are in a more systematic analysis.

5.3 Cross-Species Transcriptome Comparisons

In **Chapter 4** we predicted the biological age of single neuron classes with BitAge (**Chapter 2**) and our stochastic data-based clock (**Chapter 3**). We used these predictions to identify correlations between genes and the neuronal transcriptomic aging trajectories.

The enriched pathways for these correlations are significantly correlated with human and mice brain aging trajectories and conversely anti-correlated with anti-aging treatments. These results indicated the conservation of brain aging trajectories even in species as far evolutionary apart as *Caenorhabditis elegans* and humans.

Cross-species analyses provide a powerful approach to uncover conserved biological mechanisms and pathways associated with longevity and diseases^{307,308}. Several studies have identified common aging trajectories across mammalian species and determinants of maximum lifespan, and consistently identified DNA repair, metabolism, and stress-related pathways. An early cross-species meta-analysis of microarray datasets from mice, rats, and humans identified common signatures of aging characterized by an upregulation of inflammation and lysosomal genes, and a downregulation of collagen and mitochondrial genes³⁰⁹. A comparison of 33 mammalian species identified DNA repair and stress-related pathways to be enriched with lifespan variation³¹⁰. This was replicated in a study on humans, naked mole rats, and mice, indicating a higher DNA repair gene expression in longer lived species³¹¹. Similarly, cultured fibroblast cells of 13 rodents, 2 bats, and one shrew identified expression of DNA repair to be upregulated, and proteolysis pathways to be downregulated in longer lived species³¹². The comparison of 3 long-lived whale species, and 5 additional mammals with a novel pathway-ranking method confirmed higher expression of DNA maintenance and repair, and immune response genes in longer lived species³¹³. The longevity signature of 41 mammalian species was confirmatory enriched in translation, DNA repair, and anti-correlated with oxidative phosphorylation and proteolysis³¹⁴. And a comparison of 103 mammalian species identified organ-specific maintenance-related transcription and translation fidelity pathways, as well as DNA repair to be essential for longevity³¹⁵. Bats have evolved exceptional longevity and certain species live over ten times longer than expected from their body size^{316,317}. The extended health-span of bats and their body-size-independent longevity has been linked to a unique age-related expression pattern involving DNA repair, autophagy, immunity, and tumor suppression, as observed in cross-species comparisons of bats with humans, mice, and wolves³¹⁷. It was suggested that mammalian transcriptomic aging signatures and signatures of maximum lifespan exhibited significant similarity³¹⁸. Focusing on 26 Rodentia and Eulipotyphla species, the negative correlation between maximum lifespan and inflammation and metabolism, as well as the positive correlation between DNA repair expression and maximum lifespan was again confirmed. However, while energy metabolism does show an overlap to an aging signature, most pathways were anti-correlated. Especially DNA repair and inflammation

showed opposite trends between the effect on maximum lifespan and aging, e.g. DNA repair is higher expressed in longer lived species, but gets downregulated with age³¹⁹. This was corroborated by an analysis of human, rat, and mouse aging signatures, which were shown to coincide in downregulation of DNA repair and oxidative phosphorylation, and in upregulation of inflammation, lysosomes, and ribosomes³¹⁴. A cross-vertebrae aging trajectory transcriptome comparison of humans, mice, zebrafish *Danio rerio*, and killifish *Nothobranchius furzeri* suggested that the aging transcriptome might shift away from cancer-associated signatures, i.e. downregulation of replication and DNA repair, and towards signatures of chronic degenerative diseases, i.e. upregulation of inflammation and lysosomes³²⁰. The identification of possible determinants of maximum lifespan is still in its early stages and relies on cross-species comparisons. The evidence thus far highlights the pivotal role of DNA repair and maintenance pathways, which is in line with the central role of DNA damage and genome instability in the aging process¹⁶, and the fact that somatic mutation rates scale inversely with lifespan⁴². Notably, stricter DNA repair and maintenance would potentially lead to a slower accrual of stochastic variation, thereby explaining why long-lived species accrue stochastic variation at a slower rate (**Chapter 3**). Recently, we have shown that the DREAM complex is the master regulator of somatic DNA repair³²¹. It will be interesting to see, whether long-lived species have distinct mutations in the DREAM complex, or varying levels of DREAM complex activity.

In **Chapter 4** we have extended the cross-species comparison to *Caenorhabditis elegans*, mouse, and human data, indicating that even species as different as *Caenorhabditis elegans* and humans share conserved neuronal transcriptomic aging trajectories on the pathway level. We used this conservation of aging-dependent pathways changes to identify compounds that counteract neurodegeneration in the nematode system. Our cross-species comparison not only enabled the identification of these compounds but also underscores the vast potential of inter-species analyses in uncovering therapeutic avenues.

6 Conclusion

To conclude, in this thesis I 1.) have used novel techniques to build a highly accurate transcriptomic biological age predictor for the nematode *Caenorhabditis elegans*; 2.) identified that accumulating stochastic variation is the common underlying feature of current aging clocks; 3.) demonstrated that aging clocks can be built with as little as one biological sample and still capture significant biological age differences; and 4.) used these biological, and stochastic aging clocks to identify different aging-rates in neuronal cell classes of the nematode *Caenorhabditis elegans*, which I then used to infer conserved cross-species aging pathways, and subsequent to *in silico* screen compounds that could be validated *in vivo* to delay neurodegeneration.

7 References

1. López-Otín, C., Blasco, M. a., Partridge, L., Serrano, M. & Kroemer, G. The hallmarks of aging. *Cell* **153**, 1194–217 (2013).
2. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. Hallmarks of aging: An expanding universe. *Cell* **186**, 243–278 (2023).
3. Gems, D. & de Magalhães, J. P. The hoverfly and the wasp: A critique of the hallmarks of aging as a paradigm. *Ageing Res. Rev.* **70**, 101407 (2021).
4. Medvedev, Z. A. An attempt at a rational classification of theories of ageing. *Biol. Rev. Camb. Philos. Soc.* **65**, 375–98 (1990).
5. Libertini, G. Programmed (Adaptive) Aging Theories. in *Encyclopedia of Gerontology and Population Aging* 1–5 (Springer International Publishing, 2019). doi:10.1007/978-3-319-69892-2_54-1.
6. Pamplona, R., Jové, M., Gómez, J. & Barja, G. Programmed versus non-programmed evolution of aging. What is the evidence? *Exp. Gerontol.* **175**, 112162 (2023).
7. Kowald, A. & Kirkwood, T. B. L. Can aging be programmed? A critical literature review. *Aging Cell* **15**, 986–998 (2016).
8. Goldsmith, T. C. Emerging programmed aging mechanisms and their medical implications. *Med. Hypotheses* **86**, 92–6 (2016).
9. Goldsmith, T. C. Aging as an evolved characteristic - Weismann's theory reconsidered. *Med. Hypotheses* **62**, 304–8 (2004).
10. Weismann, A. *Ueber die Dauer des Lebens; ein Vortrag.* (G. Fischer, 1882). doi:10.5962/bhl.title.21312.
11. Longo, V. D., Mitteldorf, J. & Skulachev, V. P. Programmed and altruistic ageing. *Nat. Rev. Genet.* **6**, 866–72 (2005).
12. Medawar, P. B. An unsolved problem of biology. *Med. J. Aust.* (1952) doi:10.5694/j.1326-5377.1953.tb84985.x.
13. Williams, G. C. Pleiotropy, natural selection, and the evolution of senescence. *Evolution (N. Y.)*

- 11**, 398–411 (1957).
14. Hamilton, W. D. The moulding of senescence by natural selection. *J. Theor. Biol.* **12**, 12–45 (1966).
 15. Lee, R. D. Rethinking the evolutionary theory of aging: transfers, not births, shape senescence in social species. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 9637–42 (2003).
 16. Schumacher, B., Pothof, J., Vijg, J. & Hoeijmakers, J. H. J. The central role of DNA damage in the ageing process. *Nature* **592**, 695–703 (2021).
 17. HARMAN, D. Aging: a theory based on free radical and radiation chemistry. *J. Gerontol.* **11**, 298–300 (1956).
 18. Gladyshev, V. N. The origin of aging: imperfectness-driven non-random damage defines the aging process and control of lifespan. *Trends Genet.* **29**, 506–12 (2013).
 19. Skulachev, V. P. Aging is a specific biological function rather than the result of a disorder in complex living systems: biochemical evidence in support of Weismann’s hypothesis. *Biochemistry. (Mosc).* **62**, 1191–1195 (1997).
 20. Young, T. P. A General Model of Comparative Fecundity for Semelparous and Iteroparous Life Histories. *Am. Nat.* **118**, 27–36 (1981).
 21. Dickhoff, W. W. 13 - SALMONIDS AND ANNUAL FISHES: DEATH AFTER SEX. in *Development, Maturation, and Senescence of Neuroendocrine Systems* (eds. Schreibman, M. P. & Scanes, C. G.) 253–266 (Academic Press, 1989). doi:<https://doi.org/10.1016/B978-0-12-629060-8.50017-5>.
 22. Morbey, Y. E., Brassil, C. E. & Hendry, A. P. Rapid Senescence in Pacific Salmon. *Am. Nat.* **166**, 556–568 (2005).
 23. Robertson, O. H. PROLONGATION OF THE LIFE SPAN OF KOKANEE SALMON (ONCORHYNCHUS NERKA KENNERLYI) BY CASTRATION BEFORE BEGINNING OF GONAD DEVELOPMENT. *Proc. Natl. Acad. Sci. U. S. A.* **47**, 609–621 (1961).
 24. Gems, D., Kern, C. C., Nour, J. & Ezcurrea, M. Reproductive Suicide: Similar Mechanisms of Aging in *C. elegans* and Pacific Salmon. *Front. cell Dev. Biol.* **9**, 688788 (2021).
 25. Leopold, A. C. Senescence in Plant Development: The death of plants or plant parts may be of

- positive ecological or physiological value. *Science* **134**, 1727–32 (1961).
26. Lidsky, P. V & Andino, R. Could aging evolve as a pathogen control strategy? *Trends Ecol. Evol.* **37**, 1046–1057 (2022).
 27. Lidsky, P. V, Yuan, J., Rulison, J. M. & Andino-Pavlovsky, R. Is Aging an Inevitable Characteristic of Organic Life or an Evolutionary Adaptation? *Biochemistry. (Mosc)*. **87**, 1413–1445 (2022).
 28. Kirkwood, T. B. L. Evolution of ageing. *Nature* **270**, 301–304 (1977).
 29. Okasha, S. Why Won't the Group Selection Controversy Go Away? *Br. J. Philos. Sci.* **52**, 25–50 (2001).
 30. Eldakar, O. T. & Wilson, D. S. Eight criticisms not to make about group selection. *Evolution* **65**, 1523–1526 (2011).
 31. Okasha, S. The Group Selection Controversy. *Evolution and the Levels of Selection* 0 (2006) doi:10.1093/acprof:oso/9780199267972.003.0006.
 32. Ungewitter, E. & Scoble, H. Antagonistic pleiotropy and p53. *Mech. Ageing Dev.* **130**, 10–17 (2009).
 33. Rodríguez, J. A. *et al.* Antagonistic pleiotropy and mutation accumulation influence human senescence and disease. *Nat. Ecol. Evol.* **1**, 55 (2017).
 34. Gladyshev, V. N. Aging: progressive decline in fitness due to the rising deleteriome adjusted by genetic, environmental, and stochastic processes. *Aging Cell* **15**, 594–602 (2016).
 35. Chatterjee, N. & Walker, G. C. Mechanisms of DNA damage, repair, and mutagenesis. *Environ. Mol. Mutagen.* **58**, 235–263 (2017).
 36. Vijg, J. From DNA damage to mutations: All roads lead to aging. *Ageing Res. Rev.* **68**, 101316 (2021).
 37. Gyenis, A. *et al.* Genome-wide RNA polymerase stalling shapes the transcriptome during aging. *Nat. Genet.* **55**, 268–279 (2023).
 38. Ibañez-Solé, O., Barrio, I. & Izeta, A. Age or lifestyle-induced accumulation of genotoxicity is associated with a length-dependent decrease in gene expression. *iScience* **26**, 106368 (2023).
 39. Soheili-Nezhad, S., Ibañez-Solé, O., Izeta, A., Hoeijmakers, J. H. J. & Stoeger, T. Time is ticking

- faster for long genes in aging. *Trends Genet.* **40**, 299–312 (2024).
40. Stoeger, T. *et al.* Aging is associated with a systemic length-associated transcriptome imbalance. *Nat. aging* **2**, 1191–1206 (2022).
 41. Ren, P., Zhang, J. & Vijg, J. Somatic mutations in aging and disease. *GeroScience* (2024) doi:10.1007/s11357-024-01113-3.
 42. Cagan, A. *et al.* Somatic mutation rates scale with lifespan across mammals. *Nature* **604**, 517–524 (2022).
 43. ORGEL, L. E. The maintenance of the accuracy of protein synthesis and its relevance to ageing. *Proc. Natl. Acad. Sci. U. S. A.* **49**, 517–521 (1963).
 44. Kelmer Sacramento, E. *et al.* Reduced proteasome activity in the aging brain results in ribosome stoichiometry loss and aggregation. *Mol. Syst. Biol.* **16**, e9596 (2020).
 45. Kogan, V., Molodtsov, I., Menshikov, L. I., Shmookler Reis, R. J. & Fedichev, P. Stability analysis of a model gene network links aging, stress resistance, and negligible senescence. *Sci. Rep.* **5**, 13589 (2015).
 46. Issa, J. Aging and epigenetic drift: a vicious cycle. *J. Clin. Invest.* **124**, 24–9 (2014).
 47. Ren, P., Dong, X. & Vijg, J. Age-related somatic mutation burden in human tissues. *Front. aging* **3**, 1018119 (2022).
 48. Jackson, S. H. D., Weale, M. R. & Weale, R. A. Biological age--what is it and can it be measured? *Arch. Gerontol. Geriatr.* **36**, 103–15 (2003).
 49. BENJAMIN, H. Biologic versus chronologic age. *J. Gerontol.* **2**, 217–27 (1947).
 50. Moqri, M. *et al.* Biomarkers of aging for the identification and evaluation of longevity interventions. *Cell* **186**, 3758–3775 (2023).
 51. Blair, H. A. *Data Pertaining to Shortening of Life Span by Ionizing Radiation. University of Rochester Atomic Energy Project Report No. UR-442, Rochester, New York.* (1956).
 52. Hollingsworth, J. W., Hashizume, A. & Jablon, S. Correlations between tests of aging in Hiroshima subjects--an attempt to define 'physiologic age'. *Yale J. Biol. Med.* **38**, 11–26 (1965).
 53. HOLLINGSWORTH, J. W., ISHII, G. & CONARD, R. A. Skin aging and hair graying in Hiroshima.

- Geriatrics* **16**, 27–36 (1961).
54. HOLLINGSWORTH, J. W., HAMILTON, H. B. & ISHII, G. Age-related changes in erythrocyte agglutinability in Hiroshima subjects. *J. Appl. Physiol.* **16**, 1093–6 (1961).
 55. Comfort, A. Test-battery to measure ageing-rate in man. *Lancet (London, England)* **2**, 1411–4 (1969).
 56. Webster, I. W. & Logie, A. R. A relationship between functional age and health status in female subjects. *J. Gerontol.* **31**, 546–50 (1976).
 57. Furukawa, T., Inoue, M., Kajiya, F., Inada, H. & Takasugi, S. Assessment of biological age by multiple regression analysis. *J. Gerontol.* **30**, 422–34 (1975).
 58. Wilson, D. L. Aging hypotheses, aging markers and the concept of biological age. *Exp. Gerontol.* **23**, 435–8 (1988).
 59. Costa, P. T. & McCrae, R. R. Concepts of functional or biological age: a critical view. in *Principles of Geriatric Medicine* 30–37 (1985).
 60. Costa, P. T. & McCrae, R. R. Measures and markers of biological aging: ‘a great clamoring ... of fleeting significance’. *Arch. Gerontol. Geriatr.* **7**, 211–4 (1988).
 61. Salthouse, T. A. Functional age: Examination of a concept. in *Age, health, and employment*. 78–92 (Prentice-Hall, Inc, 1986).
 62. Hochschild, R. Improving the precision of biological age determinations. Part 1: A new approach to calculating biological age. *Exp. Gerontol.* **24**, 289–300 (1989).
 63. Dean, W. & Morgan, R. F. In defense of the concept of biological aging measurement--current status. *Arch. Gerontol. Geriatr.* **7**, 191–210 (1988).
 64. Hochschild, R. Can an index of aging be constructed for evaluating treatments to retard aging rates? A 2,462-person study. *J. Gerontol.* **45**, B187-214 (1990).
 65. Krøll, J. & Saxtrup, O. On the use of regression analysis for the estimation of human biological age. *Biogerontology* **1**, 363–8 (2000).
 66. Nakamura, E., Miyao, K. & Ozeki, T. Assessment of biological age by principal component analysis. *Mech. Ageing Dev.* **46**, 1–18 (1988).

67. Nakamura, E. & Tanaka, S. Biological ages of adult men and women with Down's syndrome and its changes with aging. *Mech. Ageing Dev.* **105**, 89–103 (1998).
68. Cho, I. H., Park, K. S. & Lim, C. J. An empirical comparative study on biological age estimation algorithms with an application of Work Ability Index (WAI). *Mech. Ageing Dev.* **131**, 69–78 (2010).
69. Klemner, P. & Doubal, S. A new approach to the concept and computation of biological age. *Mech. Ageing Dev.* **127**, 240–8 (2006).
70. Levine, M. E. Modeling the rate of senescence: can estimated biological age predict mortality more accurately than chronological age? *J. Gerontol. A. Biol. Sci. Med. Sci.* **68**, 667–74 (2013).
71. Mitnitski, A., Howlett, S. E. & Rockwood, K. Heterogeneity of Human Aging and Its Assessment. *J. Gerontol. A. Biol. Sci. Med. Sci.* **72**, 877–884 (2017).
72. Levine, M. E. *et al.* An epigenetic biomarker of aging for lifespan and healthspan. *Aging (Albany, NY)*. **10**, 573–591 (2018).
73. Mahalanobis, P. C. On the Generalised Distance in Statistics. *Proc. Natl. Inst. Sci. India* **2**, 49–55 (1936).
74. Cohen, A. A. *et al.* A novel statistical approach shows evidence for multi-system physiological dysregulation during aging. *Mech. Ageing Dev.* **134**, 110–7 (2013).
75. Parker, D. C. *et al.* Association of Blood Chemistry Quantifications of Biological Aging With Disability and Mortality in Older Adults. *J. Gerontol. A. Biol. Sci. Med. Sci.* **75**, 1671–1679 (2020).
76. Jaenisch, R. & Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.* **33 Suppl**, 245–54 (2003).
77. Karakaidos, P., Karagiannis, D. & Rampias, T. Resolving DNA Damage: Epigenetic Regulation of DNA Repair. *Molecules* **25**, 1–28 (2020).
78. Sedivy, J. M., Banumathy, G. & Adams, P. D. Aging by epigenetics--a consequence of chromatin damage? *Exp. Cell Res.* **314**, 1909–17 (2008).
79. Ciccarone, F., Tagliatesta, S., Caiafa, P. & Zampieri, M. DNA methylation dynamics in aging: how far are we from understanding the mechanisms? *Mech. Ageing Dev.* **174**, 3–17 (2018).

80. Fraga, M. F. & Esteller, M. Epigenetics and aging: the targets and the marks. *Trends Genet.* **23**, 413–8 (2007).
81. Vanyushin, B. F., Nemirovsky, L. E., Klimenko, V. V, Vasiliev, V. K. & Belozersky, A. N. The 5-methylcytosine in DNA of rats. Tissue and age specificity and the changes induced by hydrocortisone and other agents. *Gerontologia* **19**, 138–52 (1973).
82. Wilson, V. L. & Jones, P. A. DNA Methylation Decreases in Aging But Not in Immortal Cells. *Science (80-.).* **220**, 1055–1057 (1983).
83. Issa, J. P. *et al.* Methylation of the oestrogen receptor CpG island links ageing and neoplasia in human colon. *Nat. Genet.* **7**, 536–40 (1994).
84. Bollati, V. *et al.* Decline in genomic DNA methylation through aging in a cohort of elderly subjects. *Mech. Ageing Dev.* **130**, 234–239 (2009).
85. Rakyan, V. K. *et al.* Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. *Genome Res.* **20**, 434–9 (2010).
86. Christensen, B. C. *et al.* Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. *PLoS Genet.* **5**, e1000602 (2009).
87. 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**, 440–6 (2010).
88. Fraga, M. F. *et al.* Epigenetic differences arise during the lifetime of monozygotic twins. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 10604–9 (2005).
89. Wong, C. C. Y. *et al.* A longitudinal study of epigenetic variation in twins. *Epigenetics* **5**, 516–26 (2010).
90. Bjornsson, H. T. *et al.* Intra-individual Change Over Time in DNA Methylation With Familial Clustering. *JAMA* **299**, 2877–2883 (2008).
91. Bocklandt, S. *et al.* Epigenetic Predictor of Age. *PLoS One* **6**, e14821 (2011).
92. Hannum, G. *et al.* Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Mol. Cell* **49**, 359–367 (2013).
93. Thompson, R. F. *et al.* Tissue-specific dysregulation of DNA methylation in aging. *Aging Cell* **9**,

- 506–518 (2010).
94. Koch, C. M. & Wagner, W. Epigenetic-aging-signature to determine age in different tissues. *Aging (Albany, NY)*. **3**, 1018–27 (2011).
 95. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **14**, R115 (2013).
 96. Weidner, C. I. *et al.* Aging of blood can be tracked by DNA methylation changes at just three CpG sites. *Genome Biol.* **15**, R24 (2014).
 97. Vidal-Bralo, L., Lopez-Golan, Y. & Gonzalez, A. Simplified Assay for Epigenetic Age Estimation in Whole Blood of Adults. *Front. Genet.* **7**, 1–7 (2016).
 98. Lin, Q. *et al.* DNA methylation levels at individual age-associated CpG sites can be indicative for life expectancy. *Aging (Albany, NY)*. **8**, 394–401 (2016).
 99. Galkin, F., Mamoshina, P., Kochetov, K., Sidorenko, D. & Zhavoronkov, A. DeepMAge: A Methylation Aging Clock Developed with Deep Learning. *Aging Dis.* **12**, 1252–1262 (2021).
 100. Castle, J. R. *et al.* Estimating breast tissue-specific DNA methylation age using next-generation sequencing data. *Clin. Epigenetics* **12**, 45 (2020).
 101. Galkin, F., Kochetov, K., Mamoshina, P. & Zhavoronkov, A. Adapting Blood DNA Methylation Aging Clocks for Use in Saliva Samples With Cell-type Deconvolution. *Front. Aging* **2**, 1–7 (2021).
 102. Shireby, G. L. *et al.* Recalibrating the epigenetic clock: implications for assessing biological age in the human cortex. *Brain* **143**, 3763–3775 (2020).
 103. Prosz, A. *et al.* Biologically informed deep learning for explainable epigenetic clocks. *Sci. Rep.* **14**, 1306 (2024).
 104. de Lima Camillo, L. P., Lapierre, L. R. & Singh, R. A pan-tissue DNA-methylation epigenetic clock based on deep learning. *npj Aging* **8**, 4 (2022).
 105. Horvath, S. *et al.* Epigenetic clock for skin and blood cells applied to Hutchinson Gilford Progeria Syndrome and ex vivo studies. *Aging (Albany, NY)*. **10**, 1758–1775 (2018).
 106. Zhu, T. *et al.* CancerClock: A DNA Methylation Age Predictor to Identify and Characterize Aging Clock in Pan-Cancer. *Front. Bioeng. Biotechnol.* **7**, 1–12 (2019).
 107. Stubbs, T. M. *et al.* Multi-tissue DNA methylation age predictor in mouse. *Genome Biol.* **18**, 68

- (2017).
108. Thompson, M. J. *et al.* A multi-tissue full lifespan epigenetic clock for mice. *Aging (Albany. NY)*. **10**, 2832–2854 (2018).
 109. Levine, M. *et al.* A rat epigenetic clock recapitulates phenotypic aging and co-localizes with heterochromatin. *Elife* **9**, 1–19 (2020).
 110. Horvath, S. *et al.* DNA methylation clocks tick in naked mole rats but queens age more slowly than nonbreeders. *Nat. Aging* **2**, 46–59 (2021).
 111. Trapp, A., Kerepesi, C. & Gladyshev, V. N. Profiling epigenetic age in single cells. *Nat. aging* **1**, 1189–1201 (2021).
 112. Marioni, R. E. *et al.* DNA methylation age of blood predicts all-cause mortality in later life. *Genome Biol.* **16**, 25 (2015).
 113. Perna, L. *et al.* Epigenetic age acceleration predicts cancer, cardiovascular, and all-cause mortality in a German case cohort. *Clin. Epigenetics* **8**, 64 (2016).
 114. Breitling, L. P. *et al.* Frailty is associated with the epigenetic clock but not with telomere length in a German cohort. *Clin. Epigenetics* **8**, 21 (2016).
 115. Horvath, S. *et al.* Accelerated epigenetic aging in Down syndrome. *Aging Cell* **14**, 491–5 (2015).
 116. Levine, M. E., Lu, A. T., Bennett, D. A. & Horvath, S. Epigenetic age of the pre-frontal cortex is associated with neuritic plaques, amyloid load, and Alzheimer’s disease related cognitive functioning. *Aging (Albany. NY)*. **7**, 1198–211 (2015).
 117. Zhang, Q. *et al.* Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing. *Genome Med.* **11**, 54 (2019).
 118. Lu, A. T. *et al.* DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging (Albany. NY)*. **11**, 303–327 (2019).
 119. Belsky, D. W. *et al.* Quantification of the pace of biological aging in humans through a blood test, the DunedinPoAm DNA methylation algorithm. *Elife* **9**, 1–25 (2020).
 120. Belsky, D. W. *et al.* Quantification of biological aging in young adults. *Proc. Natl. Acad. Sci.* **112**, E4104–E4110 (2015).

121. Belsky, D. W. *et al.* DunedinPACE, a DNA methylation biomarker of the pace of aging. *Elife* **11**, 1–26 (2022).
122. Sugden, K. *et al.* Patterns of Reliability: Assessing the Reproducibility and Integrity of DNA Methylation Measurement. *Patterns* **1**, 100014 (2020).
123. Logue, M. W. *et al.* The correlation of methylation levels measured using Illumina 450K and EPIC BeadChips in blood samples. *Epigenomics* **9**, 1363–1371 (2017).
124. Higgins-Chen, A. T. *et al.* A computational solution for bolstering reliability of epigenetic clocks: Implications for clinical trials and longitudinal tracking. *Nat. aging* **2**, 644–661 (2022).
125. Kriukov, D., Kuzmina, E., Efimov, E., Dylov, D. V & Khrameeva, E. E. Epistemic uncertainty challenges aging clock reliability in predicting rejuvenation effects. *bioRxiv* 2023.12.01.569529 (2023) doi:10.1101/2023.12.01.569529.
126. Moqri, M. *et al.* PRC2 Clock: a Universal Epigenetic Biomarker of Aging. *bioRxiv* 2022.06.03.494609 (2022) doi:10.1101/2022.06.03.494609.
127. Lu, A. T. *et al.* Universal DNA methylation age across mammalian tissues. *Nat. aging* **3**, 1144–1166 (2023).
128. Li, C. Z. *et al.* Epigenetic predictors of species maximum lifespan and other life history traits in mammals. *bioRxiv* 2023.11.02.565286 (2023) doi:10.1101/2023.11.02.565286.
129. Porter, H. L. *et al.* Many chronological aging clocks can be found throughout the epigenome: Implications for quantifying biological aging. *Aging Cell* **20**, 1–13 (2021).
130. Ying, K. *et al.* Causality-enriched epigenetic age uncouples damage and adaptation. *Nat. aging* **4**, 231–246 (2024).
131. Dabrowski, J. K. *et al.* Probabilistic inference of epigenetic age acceleration from cellular dynamics. *bioRxiv [Preprint]* 2023.03.01.530570 (2023) doi:10.1101/2023.03.01.530570.
132. Frenk, S. & Houseley, J. Gene expression hallmarks of cellular ageing. *Biogerontology* **19**, 547–566 (2018).
133. Hernando-Herraez, I. *et al.* Ageing affects DNA methylation drift and transcriptional cell-to-cell variability in mouse muscle stem cells. *Nat. Commun.* **10**, 4361 (2019).

134. Michalak, E. M., Burr, M. L., Bannister, A. J. & Dawson, M. A. The roles of DNA, RNA and histone methylation in ageing and cancer. *Nat. Rev. Mol. Cell Biol.* **20**, 573–589 (2019).
135. Wang, Y., Yuan, Q. & Xie, L. Histone Modifications in Aging: The Underlying Mechanisms and Implications. *Curr. Stem Cell Res. Ther.* **13**, 125–135 (2018).
136. Feser, J. & Tyler, J. Chromatin structure as a mediator of aging. *FEBS Lett.* **585**, 2041–2048 (2011).
137. Sun, L., Yu, R. & Dang, W. Chromatin architectural changes during cellular senescence and aging. *Genes (Basel)*. **9**, (2018).
138. Golden, T. R., Hubbard, A., Dando, C., Herren, M. A. & Melov, S. Age-related behaviors have distinct transcriptional profiles in *Caenorhabditis elegans*. *Aging Cell* **7**, 850–865 (2008).
139. Fortney, K., Kotlyar, M. & Jurisica, I. Inferring the functions of longevity genes with modular subnetwork biomarkers of *Caenorhabditis elegans* aging. *Genome Biol.* **11**, R13 (2010).
140. Tarkhov, A. E. *et al.* A universal transcriptomic signature of age reveals the temporal scaling of *Caenorhabditis elegans* aging trajectories. *Sci. Rep.* **9**, 7368 (2019).
141. Peters, M. J. *et al.* The transcriptional landscape of age in human peripheral blood. *Nat. Commun.* **6**, 8570 (2015).
142. Mamoshina, P. *et al.* Machine Learning on Human Muscle Transcriptomic Data for Biomarker Discovery and Tissue-Specific Drug Target Identification. *Front. Genet.* **9**, 1–10 (2018).
143. González-Velasco, O., Papy-García, D., Le Douaron, G., Sánchez-Santos, J. M. & De Las Rivas, J. Transcriptomic landscape, gene signatures and regulatory profile of aging in the human brain. *Biochim. Biophys. Acta - Gene Regul. Mech.* **1863**, 194491 (2020).
144. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–60 (2015).
145. Yang, J., Huang, T., Petralia, F., Long, Q. & Zhang, B. Synchronized age-related gene expression changes across multiple tissues in human and the link to complex diseases. *Sci. Rep.* 1–16 (2015) doi:10.1038/srep15145.
146. Ren, X. & Kuan, P. F. RNAAgeCalc: A multi-tissue transcriptional age calculator. *PLoS One* **15**, e0237006 (2020).

147. Fleischer, J. G. *et al.* Predicting age from the transcriptome of human dermal fibroblasts. *Genome Biol.* **19**, 1–8 (2018).
148. Holzschek, N. *et al.* Modeling transcriptomic age using knowledge-primed artificial neural networks. *npj Aging Mech. Dis.* **7**, 15 (2021).
149. LaRocca, T. J., Cavalier, A. N. & Wahl, D. Repetitive elements as a transcriptomic marker of aging: Evidence in multiple datasets and models. *Aging Cell* acel.13167 (2020) doi:10.1111/acel.13167.
150. Neumann, J. F., Leote, A. C., Liersch, M. & Beyer, A. Predicting murine age across tissues and cell types using single cell transcriptome data. *bioRxiv* 2022.10.19.512922 (2023) doi:10.1101/2022.10.19.512922.
151. Zhu, H. *et al.* Human PBMC scRNA-seq-based aging clocks reveal ribosome to inflammation balance as a single-cell aging hallmark and super longevity. *Sci. Adv.* **9**, eabq7599 (2023).
152. Buckley, M. T. *et al.* Cell-type-specific aging clocks to quantify aging and rejuvenation in neurogenic regions of the brain. *Nat. aging* **3**, 121–137 (2023).
153. Johnson, A. A., Shokhirev, M. N., Wyss-Coray, T. & Lehallier, B. Systematic review and analysis of human proteomics aging studies unveils a novel proteomic aging clock and identifies key processes that change with age. *Ageing Res. Rev.* **60**, 101070 (2020).
154. Tanaka, T. *et al.* Plasma proteomic signature of age in healthy humans. *Aging Cell* **17**, e12799 (2018).
155. Lehallier, B. *et al.* Undulating changes in human plasma proteome profiles across the lifespan. *Nat. Med.* **25**, 1843–1850 (2019).
156. Earls, J. C. *et al.* Multi-Omic Biological Age Estimation and Its Correlation With Wellness and Disease Phenotypes: A Longitudinal Study of 3,558 Individuals. *J. Gerontol. A. Biol. Sci. Med. Sci.* **74**, S52–S60 (2019).
157. Lehallier, B., Shokhirev, M. N., Wyss-Coray, T. & Johnson, A. A. Data mining of human plasma proteins generates a multitude of highly predictive aging clocks that reflect different aspects of aging. *Aging Cell* **19**, e13256 (2020).
158. Johnson, A. A., Shokhirev, M. N. & Lehallier, B. The protein inputs of an ultra-predictive aging

- clock represent viable anti-aging drug targets. *Ageing Res. Rev.* **70**, 101404 (2021).
159. Sayed, N. *et al.* An inflammatory aging clock (iAge) based on deep learning tracks multimorbidity, immunosenescence, frailty and cardiovascular aging. *Nat. aging* **1**, 598–615 (2021).
 160. Kuo, C.-L. *et al.* Proteomic aging clock (PAC) predicts age-related outcomes in middle-aged and older adults. *medRxiv* (2024) doi:10.1101/2023.12.19.23300228.
 161. Wang, Y. *et al.* Proteomic aging clock predicts mortality and risk of common age-related diseases in diverse populations. *medRxiv* **1**, 2023.09.13.23295486 (2023).
 162. Oh, H. S.-H. *et al.* Organ aging signatures in the plasma proteome track health and disease. *Nature* **624**, 164–172 (2023).
 163. Morandini, F. *et al.* ATAC-clock: An aging clock based on chromatin accessibility. *GeroScience* **46**, 1789–1806 (2024).
 164. Shtumpf, M. *et al.* Aging clock based on nucleosome reorganisation derived from cell-free DNA. *Aging Cell* e14100 (2024) doi:10.1111/accel.14100.
 165. Gopu, V. *et al.* An accurate aging clock developed from large-scale gut microbiome and human gene expression data. *iScience* **27**, 108538 (2024).
 166. Galkin, F. *et al.* Human Gut Microbiome Aging Clock Based on Taxonomic Profiling and Deep Learning. *iScience* **23**, 101199 (2020).
 167. Pyrkov, T. V. *et al.* Extracting biological age from biomedical data via deep learning: too much of a good thing? *Sci. Rep.* **8**, 5210 (2018).
 168. Qawaqneh, Z., Mallouh, A. A. & Barkana, B. D. Deep Convolutional Neural Network for Age Estimation based on VGG-Face Model. *arXiv* (2017).
 169. Bobrov, E. *et al.* PhotoAgeClock: deep learning algorithms for development of non-invasive visual biomarkers of aging. *Aging (Albany, NY)*. **10**, 3249–3259 (2018).
 170. Rothe, R., Timofte, R. & Van Gool, L. Deep Expectation of Real and Apparent Age from a Single Image Without Facial Landmarks. *Int. J. Comput. Vis.* **126**, 144–157 (2018).
 171. Unfried, M. *et al.* LipidClock: A Lipid-Based Predictor of Biological Age. *Front. aging* **3**, 828239

- (2022).
172. Latumalea, D., Unfried, M., Barardo, D., Gruber, J. & Kennedy, B. K. A lipidome Aging Clock shows Age Acceleration in individuals with Autism. *bioRxiv* 2024.02.01.578331 (2024) doi:10.1101/2024.02.01.578331.
 173. Mijakovac, A. *et al.* Heritability of the glycan clock of biological age. *Front. cell Dev. Biol.* **10**, 982609 (2022).
 174. Krištić, J. *et al.* Glycans are a novel biomarker of chronological and biological ages. *J. Gerontol. A. Biol. Sci. Med. Sci.* **69**, 779–89 (2014).
 175. Giron, L. B. *et al.* Immunoglobulin G N-glycan markers of accelerated biological aging during chronic HIV infection. *Nat. Commun.* **15**, 3035 (2024).
 176. Hwangbo, N. *et al.* A Metabolomic Aging Clock Using Human Cerebrospinal Fluid. *J. Gerontol. A. Biol. Sci. Med. Sci.* **77**, 744–754 (2022).
 177. Mutz, J., Iniesta, R. & Lewis, C. M. Metabolomic Age (MileAge) predicts health and lifespan: a comparison of multiple machine learning algorithms. *medRxiv* 2024.02.10.24302617 (2024) doi:10.1101/2024.02.10.24302617.
 178. Faquih, T. *et al.* Robust metabolomic age prediction based on a wide selection of metabolites. *medRxiv* 2023.06.03.23290933 (2023) doi:10.1101/2023.06.03.23290933.
 179. Camillo, L. P. de L., Asif, M. H., Horvath, S., Larschan, E. & Singh, R. Histone mark age of human tissues and cells. *bioRxiv* 2023.08.21.554165 (2023) doi:10.1101/2023.08.21.554165.
 180. Chen, Q. *et al.* OMICmAge: An integrative multi-omics approach to quantify biological age with electronic medical records. *bioRxiv: the preprint server for biology* (2023) doi:10.1101/2023.10.16.562114.
 181. Macdonald-Dunlop, E. *et al.* A catalogue of omics biological ageing clocks reveals substantial commonality and associations with disease risk. *Aging (Albany, NY)*. **14**, 623–659 (2022).
 182. Moqri, M. *et al.* Validation of biomarkers of aging. *Nat. Med.* **30**, 360–372 (2024).
 183. Kabacik, S. *et al.* The relationship between epigenetic age and the hallmarks of aging in human cells. *Nat. aging* **2**, 484–493 (2022).

184. Gems, D., Singh Virk, R. & de Magalhães, J. P. Epigenetic Clocks and Programmatic Aging. *Preprints* (2023) doi:10.20944/preprints202312.1892.v1.
185. Sluiskes, M. H. *et al.* Clarifying the biological and statistical assumptions of cross-sectional biological age predictors: an elaborate illustration using synthetic and real data. *BMC Med. Res. Methodol.* **24**, 58 (2024).
186. Fatumo, S. *et al.* A roadmap to increase diversity in genomic studies. *Nat. Med.* **28**, 243–250 (2022).
187. Cohen, A. A., Morissette-Thomas, V., Ferrucci, L. & Fried, L. P. Deep biomarkers of aging are population-dependent. *Aging (Albany, NY)*. **8**, 2253–2255 (2016).
188. Horvath, S. *et al.* An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease. *Genome Biol.* **17**, 171 (2016).
189. Mamoshina, P. *et al.* Population Specific Biomarkers of Human Aging: A Big Data Study Using South Korean, Canadian, and Eastern European Patient Populations. *J. Gerontol. A. Biol. Sci. Med. Sci.* **73**, 1482–1490 (2018).
190. Watkins, S. H. *et al.* Epigenetic clocks and research implications of the lack of data on whom they have been developed: a review of reported and missing sociodemographic characteristics. *Environ. epigenetics* **9**, dvad005 (2023).
191. Koncevičius, K. *et al.* Epigenetic age oscillates during the day. *Aging Cell* e14170 (2024) doi:10.1111/accel.14170.
192. Meyer, D. H. & Schumacher, B. BiT age: A transcriptome-based aging clock near the theoretical limit of accuracy. *Aging Cell* **20**, e13320 (2021).
193. Piantanelli, L. *et al.* Use of mathematical models of survivorship in the study of biomarkers of aging: the role of heterogeneity. *Mech. Ageing Dev.* **122**, 1461–75 (2001).
194. Burt, V. L. & Harris, T. The third National Health and Nutrition Examination Survey: contributing data on aging and health. *Gerontologist* **34**, 486–490 (1994).
195. Moore, A. Z. *et al.* Change in Epigenome-Wide DNA Methylation Over 9 Years and Subsequent Mortality: Results From the InCHIANTI Study. *J. Gerontol. A. Biol. Sci. Med. Sci.* **71**, 1029–1035 (2016).

196. Gill, D. *et al.* Multi-omic rejuvenation of human cells by maturation phase transient reprogramming. *Elife* **11**, 1–23 (2022).
197. SenGupta, T. *et al.* Krill oil protects dopaminergic neurons from age-related degeneration through temporal transcriptome rewiring and suppression of several hallmarks of aging. *Aging (Albany, NY)*. **14**, 8661–8687 (2022).
198. Mikaeloff, F. *et al.* Transcriptomics age acceleration in prolonged treated HIV infection. *Aging Cell* **22**, 6–11 (2023).
199. Holcom, A. *et al.* Simultaneous neuronal expression of human amyloid- β and Tau genes drives global phenotypic and multi-omic changes in *C. elegans*. *bioRxiv Prepr. Serv. Biol.* (2023).
200. Nonninger, T. J. *et al.* A primordial TFEB-TGF β signaling axis systemically regulates diapause and stem cell longevity. *bioRxiv* 2023.10.06.561181 (2023) doi:10.1101/2023.10.06.561181.
201. Yu, D. *et al.* CellBiAge: Improved single-cell age classification using data binarization. *Cell Rep.* **42**, 113500 (2023).
202. Meyer, D. H. & Schumacher, B. Aging clocks based on accumulating stochastic variation. *Nat. aging* **4**, (2024).
203. Tomusiak, A. *et al.* Development of a novel epigenetic clock resistant to changes in immune cell composition. *bioRxiv [Preprint]* (2023) doi:<https://doi.org/10.1101/2023.03.01.530561>.
204. Zhang, Z. *et al.* Deciphering the role of immune cell composition in epigenetic age acceleration: Insights from cell-type deconvolution applied to human blood epigenetic clocks. *Aging Cell* **23**, e14071 (2024).
205. Perez, K. *et al.* DNA repair-deficient premature aging models display accelerated epigenetic age. *Aging Cell* **23**, e14058 (2024).
206. Tian, Y. E. *et al.* Heterogeneous aging across multiple organ systems and prediction of chronic disease and mortality. *Nat. Med.* **29**, 1221–1231 (2023).
207. Ahadi, S. *et al.* Personal aging markers and ageotypes revealed by deep longitudinal profiling. *Nat. Med.* **26**, 83–90 (2020).
208. Eguchi, A., Wree, A. & Feldstein, A. E. Biomarkers of liver cell death. *J. Hepatol.* **60**, 1063–1074 (2014).

209. Lehmann-Werman, R. *et al.* Identification of tissue-specific cell death using methylation patterns of circulating DNA. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E1826-34 (2016).
210. Fedoroff, N. & Fontana, W. Genetic networks. Small numbers of big molecules. *Science* **297**, 1129–31 (2002).
211. McAdams, H. H. & Arkin, A. It's a noisy business! Genetic regulation at the nanomolar scale. *Trends Genet.* **15**, 65–69 (1999).
212. Menn, D. J. & Wang, X. Stochastic and Deterministic Decision in Cell Fate. in *Encyclopedia of Life Sciences* 1–7 (Wiley, 2014). doi:10.1002/9780470015902.a0025319.
213. Davidson, E. H. & Levine, M. S. Properties of developmental gene regulatory networks. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 20063–6 (2008).
214. Hemberger, M., Dean, W. & Reik, W. Epigenetic dynamics of stem cells and cell lineage commitment: digging Waddington's canal. *Nat. Rev. Mol. Cell Biol.* **10**, 526–37 (2009).
215. Magklara, A. & Lomvardas, S. Stochastic gene expression in mammals: lessons from olfaction. *Trends Cell Biol.* **23**, 449–456 (2013).
216. Losick, R. & Desplan, C. Stochasticity and cell fate. *Science* **320**, 65–8 (2008).
217. Pujadas, E. & Feinberg, A. P. Regulated noise in the epigenetic landscape of development and disease. *Cell* **148**, 1123–1131 (2012).
218. Bahar, R. *et al.* Increased cell-to-cell variation in gene expression in ageing mouse heart. *Nature* **441**, 1011–4 (2006).
219. Bajić, D. & Poyatos, J. F. Balancing noise and plasticity in eukaryotic gene expression. *BMC Genomics* **13**, 343 (2012).
220. Lehner, B. Conflict between Noise and Plasticity in Yeast. *PLoS Genet.* **6**, e1001185 (2010).
221. Mendelsohn, A. R. & Larrick, J. W. Epigenetic Drift Is a Determinant of Mammalian Lifespan. *Rejuvenation Res.* **20**, 430–436 (2017).
222. Crofts, S. J. C., Latorre-Crespo, E. & Chandra, T. DNA methylation rates scale with maximum lifespan across mammals. *Nat. aging* **4**, 27–32 (2024).
223. Cooney, C. A. Are somatic cells inherently deficient in methylation metabolism? A proposed

- mechanism for DNA methylation loss, senescence and aging. *Growth. Dev. Aging* **57**, 261–273 (1993).
224. Kaminsky, Z. A. *et al.* DNA methylation profiles in monozygotic and dizygotic twins. *Nat. Genet.* **41**, 240–5 (2009).
225. Jones, M. J., Goodman, S. J. & Kobor, M. S. DNA methylation and healthy human aging. *Aging Cell* **14**, 924–932 (2015).
226. Kochmanski, J., Montrose, L., Goodrich, J. M. & Dolinoy, D. C. Environmental deflection: The impact of toxicant exposures on the aging epigenome. *Toxicol. Sci.* **156**, 325–335 (2017).
227. Scherer, M. *et al.* Quantitative comparison of within-sample heterogeneity scores for DNA methylation data. *Nucleic Acids Res.* **48**, e46 (2020).
228. Qi, L. & Teschendorff, A. E. Cell-type heterogeneity: Why we should adjust for it in epigenome and biomarker studies. *Clin. Epigenetics* **14**, 31 (2022).
229. Loyfer, N. *et al.* A DNA methylation atlas of normal human cell types. *Nature* **613**, 355–364 (2023).
230. Chandler, L. A., Ghazi, H., Jones, P. A., Boukamp, P. & Fusenig, N. E. Allele-specific methylation of the human c-Ha-ras-1 gene. *Cell* **50**, 711–717 (1987).
231. Meaburn, E. L., Schalkwyk, L. C. & Mill, J. Allele-specific methylation in the human genome: implications for genetic studies of complex disease. *Epigenetics* **5**, 578–82 (2010).
232. Landau, D. A. *et al.* Locally Disordered Methylation Forms the Basis of Intratumor Methylome Variation in Chronic Lymphocytic Leukemia. *Cancer Cell* **26**, 813–825 (2014).
233. Landan, G. *et al.* Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. *Nat. Genet.* **44**, 1207–14 (2012).
234. Xie, H. *et al.* Genome-wide quantitative assessment of variation in DNA methylation patterns. *Nucleic Acids Res.* **39**, 4099–108 (2011).
235. Bertucci-Richter, E. M., Shealy, E. P. & Parrott, B. B. Epigenetic drift underlies epigenetic clock signals, but displays distinct responses to lifespan interventions, development, and cellular dedifferentiation. *Aging (Albany, NY)*. **16**, 1002–1020 (2024).

236. Zhang, X. & Wang, X. MeConcord: a new metric to quantitatively characterize DNA methylation heterogeneity across reads and CpG sites. *Bioinformatics* **38**, i307–i315 (2022).
237. Sziráki, A., Tyshkovskiy, A. & Gladyshev, V. N. Global remodeling of the mouse DNA methylome during aging and in response to calorie restriction. *Aging Cell* **17**, e12738 (2018).
238. Wang, T. *et al.* Epigenetic aging signatures in mice livers are slowed by dwarfism, calorie restriction and rapamycin treatment. *Genome Biol.* **18**, 57 (2017).
239. Bertucci, E. M., Mason, M. W., Rhodes, O. E. & Parrott, B. B. The aging DNA methylome reveals environment-by-aging interactions in a model teleost. *bioRxiv* 2021.03.01.433371 (2021) doi:10.1101/2021.03.01.433371.
240. Koike, Y. *et al.* Age-related demethylation of the TDP-43 autoregulatory region in the human motor cortex. *Commun. Biol.* **4**, 1107 (2021).
241. Vershinina, O., Bacalini, M. G., Zaikin, A., Franceschi, C. & Ivanchenko, M. Disentangling age-dependent DNA methylation: deterministic, stochastic, and nonlinear. *Sci. Rep.* **11**, 9201 (2021).
242. Tarkhov, A. E. *et al.* Nature of epigenetic aging from a single-cell perspective. *bioRxiv* 2022.09.26.509592 (2023) doi:10.1101/2022.09.26.509592.
243. Bertucci-Richter, E. M. & Parrott, B. B. The rate of epigenetic drift scales with maximum lifespan across mammals. *Nat. Commun.* **14**, 7731 (2023).
244. Maegawa, S. *et al.* Caloric restriction delays age-related methylation drift. *Nat. Commun.* **8**, 539 (2017).
245. Shipony, Z. *et al.* Dynamic and static maintenance of epigenetic memory in pluripotent and somatic cells. *Nature* **513**, 115–9 (2014).
246. Pfeifer, G. P., Steigerwald, S. D., Hansen, R. S., Gartler, S. M. & Riggs, A. D. Polymerase chain reaction-aided genomic sequencing of an X chromosome-linked CpG island: Methylation patterns suggest clonal inheritance, CpG site autonomy, and an explanation of activity state stability. *Proc. Natl. Acad. Sci. U. S. A.* **87**, 8252–8256 (1990).
247. Riggs, A. D. & Xiong, Z. Methylation and epigenetic fidelity. *Proc. Natl. Acad. Sci.* **101**, 4–5 (2004).
248. Aran, D., Toperoff, G., Rosenberg, M. & Hellman, A. Replication timing-related and gene body-specific methylation of active human genes. *Hum. Mol. Genet.* **20**, 670–680 (2011).

249. Gal-Yam, E. N. *et al.* Frequent switching of Polycomb repressive marks and DNA hypermethylation in the PC3 prostate cancer cell line. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 12979–84 (2008).
250. Sliker, R. C. *et al.* Age-related accrual of methylomic variability is linked to fundamental ageing mechanisms. *Genome Biol.* **17**, 191 (2016).
251. Jenkinson, G., Pujadas, E., Goutsias, J. & Feinberg, A. P. Potential energy landscapes identify the information-theoretic nature of the epigenome. *Nat. Genet.* **49**, 719–729 (2017).
252. Laugesen, A., Højfeldt, J. W. & Helin, K. Role of the Polycomb Repressive Complex 2 (PRC2) in Transcriptional Regulation and Cancer. *Cold Spring Harb. Perspect. Med.* **6**, a026575 (2016).
253. Corley, M. & Kroll, K. L. The roles and regulation of Polycomb complexes in neural development. *Cell Tissue Res.* **359**, 65–85 (2015).
254. Li, H. *et al.* Polycomb-like proteins link the PRC2 complex to CpG islands. *Nature* **549**, 287–291 (2017).
255. Kumar, D., Cinghu, S., Oldfield, A. J., Yang, P. & Jothi, R. Decoding the function of bivalent chromatin in development and cancer. *Genome Res.* **31**, 2170–2184 (2021).
256. Horvath, S., Zhang, J., Haghani, A., Lu, A. T. & Fei, Z. Fundamental equations linking methylation dynamics to maximum lifespan in mammals. *bioRxiv* 2023.05.21.541643 (2023) doi:10.1101/2023.05.21.541643.
257. Hugues, A., Jacobs, C. S. & Roudier, F. Mitotic Inheritance of PRC2-Mediated Silencing: Mechanistic Insights and Developmental Perspectives. *Front. Plant Sci.* **11**, 1–11 (2020).
258. Campbell, S., Ismail, I. H., Young, L. C., Poirier, G. G. & Hendzel, M. J. Polycomb repressive complex 2 contributes to DNA double-strand break repair. *Cell Cycle* **12**, 2675–83 (2013).
259. Koch, Z., Li, A., Evans, D. S., Cummings, S. & Ideker, T. Somatic mutation as an explanation for epigenetic aging. *bioRxiv* (2023) doi:10.1101/2023.12.08.569638.
260. Nikopoulou, C., Parekh, S. & Tessarz, P. Ageing and sources of transcriptional heterogeneity. *Biol. Chem.* **400**, 867–878 (2019).
261. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297**, 1183–6 (2002).

262. Swain, P. S., Elowitz, M. B. & Siggia, E. D. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 12795–800 (2002).
263. de Nadal, E., Ammerer, G. & Posas, F. Controlling gene expression in response to stress. *Nat. Rev. Genet.* **12**, 833–45 (2011).
264. Weake, V. M. & Workman, J. L. Inducible gene expression: diverse regulatory mechanisms. *Nat. Rev. Genet.* **11**, 426–37 (2010).
265. Kempe, H., Schwabe, A., Crémazzy, F., Verschure, P. J. & Bruggeman, F. J. The volumes and transcript counts of single cells reveal concentration homeostasis and capture biological noise. *Mol. Biol. Cell* **26**, 797–804 (2015).
266. Padovan-Merhar, O. *et al.* Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. *Mol. Cell* **58**, 339–52 (2015).
267. das Neves, R. P. *et al.* Connecting variability in global transcription rate to mitochondrial variability. *PLoS Biol.* **8**, e1000560 (2010).
268. Shim, A. R. *et al.* Dynamic Crowding Regulates Transcription. *Biophys. J.* **118**, 2117–2129 (2020).
269. Battich, N., Stoeger, T. & Pelkmans, L. Control of Transcript Variability in Single Mammalian Cells. *Cell* **163**, 1596–610 (2015).
270. Kaern, M., Elston, T. C., Blake, W. J. & Collins, J. J. Stochasticity in gene expression: from theories to phenotypes. *Nat. Rev. Genet.* **6**, 451–64 (2005).
271. Eldar, A. & Elowitz, M. B. Functional roles for noise in genetic circuits. *Nature* **467**, 167–73 (2010).
272. Thattai, M. Universal Poisson Statistics of mRNAs with Complex Decay Pathways. *Biophys. J.* **110**, 301–305 (2016).
273. Weidemann, D. E., Holehouse, J., Singh, A., Grima, R. & Hauf, S. The minimal intrinsic stochasticity of constitutively expressed eukaryotic genes is sub-Poissonian. *Sci. Adv.* **9**, eadh5138 (2023).
274. Bahar Halpern, K. *et al.* Nuclear Retention of mRNA in Mammalian Tissues. *Cell Rep.* **13**, 2653–62 (2015).

275. Sun, M. & Zhang, J. Allele-specific single-cell RNA sequencing reveals different architectures of intrinsic and extrinsic gene expression noises. *Nucleic Acids Res.* **48**, 533–547 (2020).
276. Piras, V., Tomita, M. & Selvarajoo, K. Transcriptome-wide variability in single embryonic development cells. *Sci. Rep.* **4**, 7137 (2014).
277. Desai, R. V. *et al.* A DNA repair pathway can regulate transcriptional noise to promote cell fate transitions. *Science* **373**, (2021).
278. Debès, C. *et al.* Ageing-associated changes in transcriptional elongation influence longevity. *Nature* **616**, 814–821 (2023).
279. Liu, P., Song, R., Elison, G. L., Peng, W. & Acar, M. Noise reduction as an emergent property of single-cell aging. *Nat. Commun.* **8**, 680 (2017).
280. Bartz, J., Jung, H., Wasiluk, K., Zhang, L. & Dong, X. Progress in Discovering Transcriptional Noise in Aging. *Int. J. Mol. Sci.* **24**, 1–11 (2023).
281. Ibañez-Solé, O., Ascensión, A. M., Araúzo-Bravo, M. J. & Izeta, A. Lack of evidence for increased transcriptional noise in aged tissues. *Elife* **11**, 1–33 (2022).
282. Enge, M. *et al.* Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns. *Cell* **171**, 321–330.e14 (2017).
283. Levy, O. *et al.* Age-related loss of gene-to-gene transcriptional coordination among single cells. *Nat. Metab.* **2**, 1305–1315 (2020).
284. Vallejos, C. A., Marioni, J. C. & Richardson, S. BASiCS: Bayesian Analysis of Single-Cell Sequencing Data. *PLoS Comput. Biol.* **11**, e1004333 (2015).
285. Işıldak, U., Somel, M., Thornton, J. M. & Dönertaş, H. M. Temporal changes in the gene expression heterogeneity during brain development and aging. *Sci. Rep.* **10**, 4080 (2020).
286. Kolodziejczyk, A. A. *et al.* Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation. *Cell Stem Cell* **17**, 471–85 (2015).
287. Sivakumar, S., LeFebvre, R., Menichetti, G., Mugler, A. & Ambrosio, F. The fidelity of genetic information transfer with aging segregates according to biological processes. *bioRxiv* 2022.07.18.500243 (2022).

288. Martinez-Jimenez, C. P. *et al.* Aging increases cell-to-cell transcriptional variability upon immune stimulation. *Science* **355**, 1433–1436 (2017).
289. de Jong, T. V., Moshkin, Y. M. & Guryev, V. Gene expression variability: the other dimension in transcriptome analysis. *Physiol. Genomics* **51**, 145–158 (2019).
290. Salzer, M. C. *et al.* Identity Noise and Adipogenic Traits Characterize Dermal Fibroblast Aging. *Cell* **175**, 1575–1590.e22 (2018).
291. Wiley, C. D. *et al.* Analysis of individual cells identifies cell-to-cell variability following induction of cellular senescence. *Aging Cell* **16**, 1043–1050 (2017).
292. Calkhoven, C. F. & Ab, G. Multiple steps in the regulation of transcription-factor level and activity. *Biochem. J.* **317** (Pt 2, 329–42 (1996).
293. Rangaraju, S. *et al.* Suppression of transcriptional drift extends *C. elegans* lifespan by postponing the onset of mortality. *Elife* **4**, e08833 (2015).
294. Menichetti, G., Bianconi, G., Castellani, G., Giampieri, E. & Remondini, D. Multiscale characterization of ageing and cancer progression by a novel network entropy measure. *Mol. Biosyst.* **11**, 1824–31 (2015).
295. Clemens, Z. *et al.* The biphasic and age-dependent impact of *klotho* on hallmarks of aging and skeletal muscle function. *Elife* **10**, 1–39 (2021).
296. Warren, L. A. *et al.* Transcriptional instability is not a universal attribute of aging. *Aging Cell* **6**, 775–82 (2007).
297. Angelidis, I. *et al.* An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. *Nat. Commun.* **10**, 963 (2019).
298. Kimmel, J. C. *et al.* Murine single-cell RNA-seq reveals cell-identity- and tissue-specific trajectories of aging. *Genome Res.* **29**, 2088–2103 (2019).
299. Ximerakis, M. *et al.* Single-cell transcriptomic profiling of the aging mouse brain. *Nat. Neurosci.* **22**, 1696–1708 (2019).
300. Lee, J.-Y. *et al.* Misexpression of genes lacking CpG islands drives degenerative changes during aging. *Sci. Adv.* **7**, (2021).

301. Ku, M. *et al.* Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS Genet.* **4**, e1000242 (2008).
302. Owen, B. M. & Davidovich, C. DNA binding by polycomb-group proteins: searching for the link to CpG islands. *Nucleic Acids Res.* **50**, 4813–4839 (2022).
303. Cheung, P. *et al.* Single-Cell Chromatin Modification Profiling Reveals Increased Epigenetic Variations with Aging. *Cell* **173**, 1385-1397.e14 (2018).
304. McLain, A. T. & Faulk, C. The evolution of CpG density and lifespan in conserved primate and mammalian promoters. *Aging (Albany, NY)*. **10**, 561–572 (2018).
305. Mayne, B., Berry, O., Davies, C., Farley, J. & Jarman, S. A genomic predictor of lifespan in vertebrates. *Sci. Rep.* **9**, 17866 (2019).
306. Zou, Z., Ohta, T., Miura, F. & Oki, S. CHIP-Atlas 2021 update: a data-mining suite for exploring epigenomic landscapes by fully integrating ChIP-seq, ATAC-seq and Bisulfite-seq data. *Nucleic Acids Res.* **50**, W175–W182 (2022).
307. Gorbunova, V., Seluanov, A., Zhang, Z., Gladyshev, V. N. & Vijg, J. Comparative genetics of longevity and cancer: insights from long-lived rodents. *Nat. Rev. Genet.* **15**, 531–40 (2014).
308. Ma, S. & Gladyshev, V. N. Molecular signatures of longevity: Insights from cross-species comparative studies. *Semin. Cell Dev. Biol.* **70**, 190–203 (2017).
309. de Magalhães, J. P., Curado, J. & Church, G. M. Meta-analysis of age-related gene expression profiles identifies common signatures of aging. *Bioinformatics* **25**, 875–81 (2009).
310. Fushan, A. A. *et al.* Gene expression defines natural changes in mammalian lifespan. *Aging Cell* **14**, 352–65 (2015).
311. MacRae, S. L. *et al.* DNA repair in species with extreme lifespan differences. *Aging (Albany, NY)*. **7**, 1171–84 (2015).
312. Ma, S. *et al.* Cell culture-based profiling across mammals reveals DNA repair and metabolism as determinants of species longevity. *Elife* **5**, 1–25 (2016).
313. Toren, D. *et al.* Gray whale transcriptome reveals longevity adaptations associated with DNA repair and ubiquitination. *Aging Cell* **19**, e13158 (2020).

314. Tyshkovskiy, A. *et al.* Distinct longevity mechanisms across and within species and their association with aging. *Cell* **186**, 2929-2949.e20 (2023).
315. Liu, W. *et al.* Large-scale across species transcriptomic analysis identifies genetic selection signatures associated with longevity in mammals. *EMBO J.* **42**, e112740 (2023).
316. Munshi-South, J. & Wilkinson, G. S. Bats and birds: Exceptional longevity despite high metabolic rates. *Ageing Res. Rev.* **9**, 12–19 (2010).
317. Huang, Z. *et al.* Longitudinal comparative transcriptomics reveals unique mechanisms underlying extended healthspan in bats. *Nat. Ecol. Evol.* **3**, 1110–1120 (2019).
318. Takasugi, M., Yoshida, Y., Nonaka, Y. & Ohtani, N. Gene expressions associated with longer lifespan and aging exhibit similarity in mammals. *Nucleic Acids Res.* **51**, 7205–7219 (2023).
319. Lu, J. Y. *et al.* Comparative transcriptomics reveals circadian and pluripotency networks as two pillars of longevity regulation. *Cell Metab.* **34**, 836-856.e5 (2022).
320. Aramillo Irizar, P. *et al.* Transcriptomic alterations during ageing reflect the shift from cancer to degenerative diseases in the elderly. *Nat. Commun.* **9**, 327 (2018).
321. Bujarrabal-Dueso, A. *et al.* The DREAM complex functions as conserved master regulator of somatic DNA-repair capacities. *Nat. Struct. Mol. Biol.* **30**, 475–488 (2023).

8 Acknowledgements

I would like to thank Prof. Dr. Björn Schumacher for giving me the opportunity to embark on this journey and supervising me along the ride. I am particularly grateful for the trust and encouragement to work fully independent, and his constructive criticism and guidance when I was lost in data.

I would like to thank Prof. Dr. Andreas Beyer and Prof. Dr. Peter Tessarz for providing support and guidance as members of my Thesis Advisory Committee, and for Prof. Dr. Andreas Beyer for being the second examiner of this thesis. I am also especially grateful for Prof. Dr. Barak Rotblat for being the third examiner of this thesis, and to Prof. Dr. Matthias Hammerschmidt for accepting the role of Chair of my Thesis Defense Committee.

I would like to acknowledge the Cologne Graduate School for Ageing Research for giving structure, guidance, and support especially during the first three years of my PhD. And I am extremely grateful to Daniela, Jenny, Julia, and the rest of the CGA coordination team for their continued support.

To my CGA class of 2018: Aish, Filippo, He, Nils, Edoardo, Álvaro, Sadig, and Jo?o, thank you for the fun times, lots of coffee, and support.

I would like to express my sincere gratitude to all current and former Schumacher lab members for their support, criticism, feedback, discussions, and welcoming environment, fun times, and for not forgetting me while I was working from home. Especially thanks to Christian, Yao, Simon, Robert, and Arturo for the fruitful collaborations; and to Khrysty, Olga, and Richard for the bioinformatics discussions.

To Dr. Stephanie Panier, Prof. Dr. Boris Pfander, and all members of the Panier and Pfander labs for their feedback and criticism.

To my parents, Andreas and Christiane, my sisters Sarah and Vanessa, my brother Marvin, and the rest of my family for their endless support, and belief in my abilities. I am deeply grateful for their presence in my life.

And most importantly, I would like to express my deepest gratitude to my wife Ayumi for her unwavering love, support, and understanding throughout this journey. Her encouragement, patience, and sacrifices have been the cornerstone of my success, and I am forever grateful for her presence by my side. And to my dear Tsukasa and Sora, who entered our lives during this challenging yet rewarding period, you have brought immeasurable joy and inspiration. Your presence motivates me and provides the opportunity to completely clear my mind.

I would also like to extend my thanks to all those who have supported me in ways both seen and unseen, and whom I might have missed here. Your encouragement, understanding, and support has been appreciated and valued.

Thank you!

9 Additional Contributions

During my PhD I contributed to several other publications that are not directly included in my thesis:

1. Gallrein C., Williams A., **Meyer D.H.**, Messling J., Garcia A., Schumacher B. baz-2 enhances systemic proteostasis in vivo by regulating acetylcholine metabolism. **Cell Reports** 42(12):113577 (2023)
2. Bujarrabal-Dueso, A., Sendtner, G., **Meyer, D. H.**, Chatzinikolaou, G., Stratigi, K., Garinis, George A., Schumacher, B. The DREAM complex functions as conserved master regulator of somatic DNA-repair capacities. **Nat. Struct. Mol. Biol.** 30, 475–488 (2023)
3. Selle, J., [...], **Meyer, D.H.**, [...], Alcazar, M.A. Perinatal Obesity Sensitizes for Premature Kidney Aging Signaling. **Int. J. Mol. Sci.** 24, 2508 (2023)
4. Wang, S., **Meyer, D. H.** & Schumacher, B. Inheritance of paternal DNA damage by histone-mediated repair restriction. **Nature** 613, 365–374 (2023)
5. Braun, F., [...], **Meyer, D.H.**, [...], Kurschat, C. Loss of genome maintenance accelerates podocyte damage and aging. **bioRxiv [Preprint]** (2023)
6. Uszkoreit, S., **Meyer, D. H.**, Rechavi, O., Schumacher, B. Sensory neurons safeguard from mutational inheritance by controlling the CEP-1 / p53-mediated DNA damage response in primordial germ cells. **bioRxiv [Preprint]** (2022)
7. Wang, S., **Meyer, D. H.** & Schumacher, B. H3K4me2 regulates the recovery of protein biosynthesis and homeostasis following DNA damage. **Nat. Struct. Mol. Biol.** 27, 1165–1177 (2020)