Accurate aging clocks based on accumulating stochastic variation

David H. Meyer[1,2,*] and Björn Schumacher[1,2,*]

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

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

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

Aging clocks have provided one of the most significant recent breakthroughs in the biology of aging. Such clocks allow the determination of chronological and increasingly also biological age, which is prerequisite for assessing the effectiveness of interventions in the aging process and preventive treatments of age-related diseases. The most advanced aging clocks are based on age-dependent changes in DNA methylation pattern. The reproducibility of such changes over the life course has reinvigorated the debate whether a programmed process underlies aging. A programmed aging process, however, is incompatibly with the evolutionary theory of aging. Aging occurs as a consequence of a vanishing force of selective pressure post-reproduction as no fitness benefit is provided by immortality of the soma. In fact, stochastic events have been observed to increasingly occur during the aging process. Here, we test whether aging clocks could be built with entirely stochastic variation. We find that accumulating stochastic variation is sufficient to accurately predict chronological and biological age. Moreover, current aging clocks are entirely compatible with random alterations in the methylation or transcriptomic patterns. Our analysis unifies the clock measure of aging with the evolutionary theory of aging and predicts that any set of data that have a ground state at the age zero with accumulating stochastic variation could be used for building accurate aging clocks.

Introduction

Weismann [1] proposed in 1882 that aging might be programmed to benefit a population of species by freeing up resources taken by older individuals. The hypothesis of a programmed aging process, however, was later largely rejected [2–5], for a range of reasons such as the circularity of the argument and the underlying assumption of group selection. Instead, evolutionary theories of aging realized the vanishing force of natural selection post-reproductively, i.e. once the progeny carries on the genetic information, further maintenance of the parental individuals would have little fitness contribution. Most clearly, this has been stated in the disposable soma theory of aging but also the mutation accumulation and the antagonistic pleiotropy theories recognized the lack of a fitness contribution of old individuals leading to the accumulation of gene variants that can have detrimental effects at older ages as those variants are not counterselected against [2,6]. These theories are in line with the insufficiency of maintenance and repair processes that leads to stochastic damage accumulation with aging thus causing the functional decline of the organism [7]. Recent progress on aging clocks, however, has revived the idea of a potential program involved in aging [8,9]. Currently, it is controversially discussed whether aging is purely a stochastic entropy-driven event in line with the evolutionary theory of aging or whether the existence of aging clocks could show a causal relationship to aging [10,11]. Epigenetic drift has been observed during aging and was assigned to imperfect maintenance of epigenetic marks [12]. Such a drift might reduce methylation differences between genomic regions that are defined during development over time [13]. It has been proposed that age-coupled stochastic methylation changes are highly genome context specific [14], and that an information-theoretic view of DNA methylation pattern explains the observed stochasticity in line with context-specific maintenance energy consumption [15]. Indeed, early analyses with differential equations showed that CpG methylation sites can be modelled based on maintenance rates, which define the CpG site-specific equilibria [16,17]. This stochastic epigenetic drift has been shown to be conserved across species and attenuated upon caloric restriction [18].

To deepen the mechanistic understanding of epigenetic aging clocks, Levine et al. deconstructed epigenetic clock CpG sites from 12 different clocks into distinct modules and showed that while some modules might be driven by entropic alterations and regress to a methylation state of 0.5, most might be distinct and changing systematically with time [19]. Recently, Tarkhov et al. showed that aging single-cell DNA methylation changes are predominantly affected by a stochastic component [20], and that a single stochastic variable (thermodynamic biological age) can track entropic aging [21].

Here, we show that in principle any dataset that can be normalized to values between 0 and 1, containing accumulating stochastic variation can be used to build a predictor suggesting that any set of biological measurements could be used to build accurate aging clocks. We determine that the pace of aging is primarily set by the degree of stochastic variation, where increased stochasticity accelerates while reduced stochastic variation decelerates the predicted age. We further determine that current epigenetic aging clocks measure how much stochastic variation accumulated. We show that the predictive results of a model trained on simulated data with accumulating stochastic variation correlates significantly with the chronological age of human DNA methylation samples. We validate our findings in a transcriptomic dataset of *C. elegans,* and show that predictions of the most accurate transcriptomic aging clock correlate significantly with the number of times stochastic variation was added to the simulated data. Finally, we show that the predictive results of simulated transcriptomic data with accumulating stochastic variation significantly correlates with the biological age. Taken together, we establish that aging clocks could be based on any biological parameter and that precise aging clocks are compatible with the evolutionary theory of aging. No deterministic process is required but instead stochastic age-related alterations allow the precise measurement of aging.

Results

Data-type independent predictions

To investigate whether a stochastic process is sufficient to build an age predictor of any dataset, we used purely simulated random data. We set the simulated age range between 0 and 100, and defined a random ground state, i.e. a starting point as a state at the beginning of an individual's life. The ground state is motivated by reports showing a tight global regulation after the beginning of life starting with the zygote stage on the transcriptome [22], proteome [23], and epigenome [24] level, in addition to the recently proposed ground zero of organismal life and aging [25]. For our first simulations, we simulated 2000 random data points (features) uniformly distributed between 0 and 1 that we defined as the ground state. Features in prediction models can be any quantifiable data type such as for example methylation data. To model variation in the ground state, we varied the features slightly for each simulated sample by adding stochastic variation to each feature that was drawn from a normal distribution with a mean of 0 and standard deviation σ of 0.01, i.e. $N(\mu = 0, \sigma^2 = 0.01^2)$. With these parameters ~99.7 % of noise values are within the interval [-0.03, 0.03]. To test whether accumulating normal-distributed stochastic variation over time would allow building a predictor of the simulated age, we added normal-distributed stochastic variation independently to all features in the ground state 1 to 100 times, i.e. while each sample started from the same ground state (including slight variations as described above), the normal-distributed stochastic variation is added independently (**Figure 1A**, see methods for details). For example, for a sample with simulated age 2, stochastic variation would be added twice to the ground state. The stochastic variation addition was performed independently from all other samples, i.e. ground state + 1x stochastic variation sampled from the normal distribution, + 1x stochastic variation sampled from the normal distribution. A sample with simulated age 10 is simulated by taking the ground state and adding, independently sampled, normal-distributed stochastic variation 10 times (**Figure 1A**). The amount of stochastic variation added to each feature was drawn from a normal distribution with a mean of 0 and standard deviation σ of 0.05, i.e. $N(\mu = 0, \sigma^2 = 0.05^2)$. With these parameters ~99.7 % of noise values are within the interval [-0.15, 0.15]. Since the stochastic variation distribution is the same for all samples and time-steps, older samples are noisier, i.e. more divergent from the ground state. We simulated 6 sets of samples ranging from stochastic variation applied once to applied 100 times, reflecting a potential lifespan range. Note that the range from 1-100 was chosen arbitrarily. We used 3 sets of 100 samples (one sample per simulated age) to train an Elastic net regression that predicts the simulated age, i.e. the number of times stochastic variation was added. To validate the model, we used the 3 independent validation samples, i.e. samples that started with the same ground state but that added independent normal-distributed stochastic variation from the same distribution $N(\mu = 0, \sigma^2 = 0.05^2)$ (**Figure 1B**). As expected, accumulating normal-distributed stochastic variation without limits does not allow for any predictor to be build and the validation samples do not show any trend in the data (**Figure 1C**). The stochastic variation application in each time-step is independent from all other and therefore contains negative and positive values equally likely, which will lead to a cancellation of stochastic variation on average and subsequent neither a trend nor a prediction. We, therefore, next used random feature values within a range limitation between 0 and 1 because, in principle, all types of quantifiable biological parameters could be normalized to values between 0 and 1. Using the same approach as above but limiting the values between 0 and 1 after adding the stochastic variation, i.e. not allowing features to go below 0 or above 1, surprisingly allowed for an almost perfect prediction with a Pearson correlation of the independent validation data of 0.99 (p-value 2.42e-241) (**Figure 1D**). Of note, the ground state in Figure 1C and D is comparable and only containing random values between 0 and 1 for each feature. Thus, the model found pattern in the simulated data that allowed the prediction of how often stochastic variation was added to the ground state (the simulated age) even in data not used during the training process. Importantly, this will potentially work for any dataset, since our simulated starting point (ground state) consists of uniformly random data between 0 and 1, and the stochastic variation added at each time-step is randomly chosen from a normal distribution, i.e. does not require any regulation or program.

The prediction accuracy of the independent validation data was robust to the distribution from which stochastic variation was sampled for the training and validation samples (**Figure 1E**). Even predictions in which the age-related stochastic variation per time-step was smaller than the stochastic variation with which we varied the ground state for each sample ($N(\mu = 0, \sigma^2 = 0.01^2)$), showed high accuracy, e.g. the model trained on stochastic variation 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 the independent validation data (**Figure 1E**). This indicates, that even a small amount of accumulating stochastic variation per time-step is enough for an accurate prediction, e.g. stochastic variation sampled from $N(\mu = 0, \sigma^2 = 0.005^2)$ contains 99.7% of all sampled values in the interval [-0.015, 0.015] with a mean of 0.

During training, Elastic net regression assigns a coefficient to each of the 2000 features that then can be used to predict novel independent samples. The Elastic net regression coefficients for the 2000 features in our simulation in Figure 1D are highly reproducible in between independent runs, as long as the ground state (including slight variations as described above) is the same (**Figure 1F**), indicating that even random stochastic variation pattern allow for robust predictions. The prediction is possible due to a regression to the mean, which is to be expected from a stochastic process with a data range limit (**Figure 1G**). Features starting close to 0 or 1, cannot go below, respective beyond, their respective limits, resulting on average after adding random stochastic variation to a regression towards 0.5. Features starting close to 0, therefore, tend to increase after stochastic variation addition resulting in a positive Elastic net coefficient, while features close to 1 tend to decrease resulting in a negative coefficient. Features starting around 0.5 in the ground state are more noise sensitive since the added stochastic variation is equally likely to move in either direction leading on average to a cancellation of noise (**Figure 1G**). Thus, features with a ground state close to 0.5 won't allow a robust trend in the data and therefore won't add any information to the prediction model.

The number of features of the ground state has an effect on the model accuracy: The prediction accuracy of the amount of normal-distributed stochastic variation plateaus after ~1000 features at an R^2 value around 0.97, showing that even models with a limited number of features are highly accurate in predicting how often normal-distributed stochastic variation was added to the ground state of independent validation samples (**Figure 1H**). Of note, Elastic net regression shrinks coefficients of some features to 0 and thereby further reduces the number of features. These results show that reproducible predictions are possible with less than 1000 features as long as there is accumulating stochastic variation and the data can be normalized between 0 and 1, i.e. predictions are not limited to DNA methylation or transcriptomic data.

We next wondered how a model trained on stochastic variation sampled from $N(\mu = 0, \sigma^2 = 0.05^2)$ would predict samples with different stochastic variation distributions. Choosing a standard deviation twice as large (σ=0.1), also doubles the interval to [-0.3,0.3] from which ~99.7 % of stochastic variation values are sampled, which increases the amount of stochastic variation added in each time step. Testing the model on data simulated with more stochastic variation per time step resulted in a faster increase and plateau of the prediction, while a reduced stochastic variation level decreased the slope of the prediction (**Figure 1I**). Samples with more stochastic variation per time step reach their maximum simulated age earlier. This analysis suggests that an increase in stochastic variation accelerates, while a decrease in stochastic variation decelerates the predicted aging process.

The simulations in Figure 1 did not restrain the stochastic variation accumulation aside from keeping the features between 0 and 1. Stochastic variation was sampled from the same normal distribution, regardless of the starting values in the ground state. Adding normal distributed stochastic variation once in this set-up does not change the simulated sample much from the ground state (**Supplement Figure 1A**), while adding stochastic variation 100 times leads to a uniform distribution of features (**Supplement Figure 1B**). Biological data on the other hand will contain features, e.g. CpG methylation levels or gene expression levels, that are under higher maintenance and less noisy. Comparing biological DNA methylation data of young and old subjects shows that features, i.e. methylation sites, starting close to the extremes (0 or 1) tend to be stronger maintained and show therefore less variance, i.e. potential stochastic variation (**Supplement Figure 1C**).

Empirically estimated stochastic variation in bulk DNA methylation

Next, we wondered whether our simulations would be applicable to data with accumulating stochastic variation that is sampled from empirically estimated stochastic variation distributions. To empirically estimate biological stochastic variation levels, we first estimated the stochastic variation distribution of data points between DNA methylation samples of 2 young subjects one year apart (16 and 17 years). We used DNA methylation data because the most widely used aging clocks are based on CpG methylation data. As shown in Supplement Figure 1C the DNA methylation sites close to 0, or 1, have low variance between young and old samples, while most of the differences in DNA methylation lie in the middle of the data distribution. Estimating the variance between subjects over all data points therefore would over-estimate the variance close to 0 and 1. Instead, to empirically estimate the variance/ stochastic variation distributions we divided the data into quantiles based on their DNA methylation level in the youngest sample (**Supplement Figure 1D**). For each of these quantiles, we estimated the distribution of the differences between the samples of the 2 young subjects that are one year apart (see methods for details) (**Supplement Figure 1E,F**). We chose samples from subjects one year apart from each other to estimate the amount of stochastic variation accumulating within a year. Note that this will overestimate the amount of stochastic variation since age-independent interindividual DNA methylation differences are not excluded. Using a similar approach as in Figure 1, we simulated 6 datasets with samples of simulated age 1-100, with the difference that we started the ground state with the youngest DNA methylation sample (GSM1007467) in the public dataset GSE41037 [26] instead of uniformly random data. The accumulating stochastic variation was not sampled from a normal distribution, but from empirically estimated distributions based on quantiles of the DNA methylation data distribution of the youngest sample (**Supplementary Figure 1D-F**, see methods for details). All number of quantiles tested allowed for accurate age predictions of the independent validation data (**Supplement Figure 1G,H**). The number of quantiles, i.e. the number of stochastic variation distributions, did not have a strong effect on either the predictions (**Supplement Figure 1G**) or the distribution of features after adding noise 100x (**Supplement Figure 1I**). These results show that empirically estimated stochastic variation distributions still allow for a prediction, however, this approach still over-estimates the overall amount of stochastic variation in the data, as can be seen in the comparison of the data distributions in Supplement Figure 1C and I.

Single-cell simulations

To improve the simulations of stochastic variation for DNA methylation data, we simulated instead of bulk data between 0 and 1, “single-cell” data for which each feature is binary, i.e. either methylated (1) or unmethylated (0) **(Figure 2A)**. Pfeifer et al. showed that the methylation pattern at single CpG sites can be modelled with differential equations containing a methylation maintenance efficiency (E_m) (the probability that a methylated site stays methylated), and a *de novo* methylation efficiency (E_d) (the probability that an unmethylated site gets methylated; $1 - E_d$ is the maintenance efficiency of the unmethylated state (E_u)) [16]. These maintenance efficiencies describe the rate by which a CpG site does not alter per time-step. Using E_m and E_d we simulated single-cell DNA methylation changes over time as depicted in Figure 2A. We started with bulk DNA methylation data from a young sample (2000 randomly chosen CpG sites from GSM1007467 [26] as the ground state) and generated 1000 cells per CpG site that average up to the bulk DNA methylation rate, i.e. a bulk DNA methylation value of 0.13 would be defined as 130 cells being methylated (1), and the remaining 870 cells being unmethylated (0). Next, we randomly alter the state of every single-cell CpG site based on the respective E_m and E_d values for each time step., i.e. for each time-step we flip a coin with the probabilities E_m (to stay methylated) and E_d (to *de novo* methylate) for each CpG site in each cell. For training and validating a predictor, we again computed the average bulk methylation levels for each site and time-point. The training and validation process of the Elastic net regression is the same as described in Figure 1B.

First, we tested how a universal maintenance efficiency rate, i.e. the same rate for all features, for 500 features would affect the accuracy of the model (**Figure 2B**). A high maintenance efficiency for both methylated and unmethylated states for all simulated CpG sites (E_m=99.9%, E_d=0.01%, i.e. E_u=99.9%) allowed for almost perfect simulated age predictions with an R^2 of 0.999 of the independent validation data (**Figure 2B, C**). This is surprising since the sample of simulated age 100 shows almost no deviation from the ground state (**Supplement Figure 2A**). Even maintenance rates of up to 99.995% (for which we expect only 25 of the simulated 500*1000=500.000 cells to change state in each time-step) resulted in a prediction with an R^2 of 0.78 in the independent validation data (**Figure 2B**). The predictor is robust in the number of features, i.e. DNA methylation sites, allowing for highly accurate age predictions with small feature sizes, whose accuracy cap after around 32 features (**Figure 2D**). Similar to Figure 1I, training the model on a maintenance rate of 99.9 % per site and testing it on data simulated with lower, respectively higher maintenance rates, showed that less maintenance leads to a quicker increase in the age prediction, i.e. accelerates the aging clock, while higher maintenance reduced the predicted age (**Figure 2E**). These results indicate that even a high maintenance rate for all simulated features allows for highly accurate age predictions, and that an increase in maintenance decelerates biological aging, while a decrease in maintenance would accelerate the aging process.

Since everything except a 100 % maintenance rate would lead to a regression to the maintenance rate-specific equilibrium for each site [16], we wondered whether a prediction would be possible if we set the starting values for each feature in the ground state to the maintenance rate-specific equilibrium before applying stochastic changes to the data. A maintenance rate of 99.9 % for methylated as well as unmethylated sites leads to a regression to the mean, i.e. 0.5 is the equilibrium state, irrespective of the starting value of each feature in the ground state. Unsurprisingly, starting the simulation with 0.5 for each feature and a 99.9% maintenance rate for each feature, i.e. CpG site, did not allow for a prediction of the simulated age, since no regression to the equilibrium state is possible (**Supplementary figure 2B**). However, just a slight deviation to 0.51 for all starting values in the ground state led to an accurate simulated age prediction via a regression to the equilibrium state, i.e. the mean (**Supplementary Figure 2C**).

Each CpG site might have a different maintenance efficiency and likely factors as accessibility, DNA sequence context, histone modifications, and protein binding affect E_m and E_d[14,17,27]. Therefore, we empirically estimated the maintenance efficiency of each CpG site from data of an old subject. According to Pfeifer et al.[16] the equilibrium of the methylation state M_{eq} is reached at

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

. Pfeifer showed that one can then estimate E_m from equation [1] for a given equilibrium state. DNA methylation trends towards the site-specific equilibrium over time [16,17]. We, therefore, estimated that the data of the sample from the oldest subject in the dataset are closest to the site-specific equilibria. While the measurement of maintenance efficiencies is not straight forward, several groups have estimated the biological range of site-specific E_m and E_d values. Pfeifer et al. estimated E_m to be ~99.9 % and E_d to be ~ 5 %, i.e. the maintenance of unmethylated regions E_u is ~95 % [16], Riggs et al. estimated the average E_m to be 95 % and for many sites bigger than 99 % [17], and Laird et al. estimated E_m to be between 95-98 % and E_d to be maximally 23 % [28]. We set maintenance levels based on these publications to E_m > 95 % and E_d < 23 %. To estimate the site-specific E_m and E_d values for the DNA methylation dataset, we set the site-specific DNA methylation equilibrium to be the value of the oldest sample in the dataset (GSM1007832 [26]) and estimated E_m and E_d within the limits defined above. This is only a rough approximation of the site-specific equilibria, but the closest available at the moment. Similar to the universal maintenance model in Figure 2B-D, highly accurate simulated age predictions are possible if E_m and E_d are empirically estimated from data (**Figure 2F**). The predictions most likely cap off earlier than in Figure 2C since most empirically estimated maintenance rates are smaller than

99.9 %, leading to a quicker convergence to the site-specific equilibria. Once the equilibria for all features is reached on average the prediction will stay stable (see also **Supplementary Figure 2B**).

It was suggested that if age-related DNA methylation changes were due to entropic alterations, it would lead to a bias against DNA methylation aging clock sites that start around 0.5 since these sites won't regress towards the mean [19]. Hence, if clock sites started around the mean and regressed away from it, it would argue for either a regulated mechanism, clonal expansion, or cellular selection[19]. This would be in line with our results in Supplement Figure 2B, i.e. if the equilibrium of all sites is exactly 0.5, starting at 0.5 will not allow a prediction of the simulated age. However, since the maintenance rate and equilibrium of each DNA methylation site is site specific, not all sites will regress towards the mean but might even regress away from it. Such a regression away from the mean is still in line with stochasticity and entropic alterations. In our simulations each time-step is purely random stochastic variation based on the maintenance rates. So, while the site-specific maintenance rates give a framework in which each feature, i.e. CpG site, will change, the change itself is purely stochastic. Site-specific E_m and E_d values indeed allowed for an accurate simulated age prediction even if all features start at 0.5 in the ground state (**Supplementary Figure 2D**). The effect of the stochastic variation after 100 times-steps in comparison with the ground state, shows that in our simulations features starting close to 0, respective 1, have less variation as features starting close to 0.5 in the ground state (**Supplement Figure 2E**), which resembles the comparison of a young and an old human DNA methylation dataset (**Supplement Figure 1C**). Without site-specific stochastic variation the prediction was driven by the regression to the mean, indicated by a negative slope between the ground state and the coefficients of the Elastic net regression (**Figure 1G**). Site-specific stochastic variation on the other hand does not show a correlation between the ground state and the coefficients of the Elastic net regression; features close to 0 for example can have negative coefficients indicating that these sites do not regress towards 0.5 (**Supplement Figure 2F**). This suggests that even a regression away from the mean could be explained via a stochastic process.

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

Public aging clocks

Next, we were wondering whether published DNA methylation aging clocks might also mainly measure stochastic variation. For this, we generated again samples based on single-cell simulations with empirically estimated site-specific E_m and E_d values (and as mentioned above the limits E_m > 95 % and E_d < 23 %) and tested them with several published clocks. Surprisingly, Horvath's pan-tissue DNA methylation clock [29] predicts a linear increase of the amount of stochastic variation generated based on empirically estimated E_m and E_d values until it caps off at an predicted age around ~60 years (**Supplementary Figure 3A**). The time-steps in our simulations are arbitrary and not directly comparable to the predicted age, since our simulated age tracks how often we added stochastic variation, and the predicted age is epigenetic age in years. We were, therefore, wondering whether the fast cap-off of the predicted age after the linear increase might be due to the amount of stochastic variation we apply in each time-step. The amount of stochastic variation is affected by the site-specific E_m and E_d values, which are empirically estimated from the oldest sample in the used dataset. Due to the nature of the estimation either E_m or E_d are fixed, allowing the other to be estimated from data. Note that multiple E_m and E_d values will regress to the same equilibrium over time (compare equation 1). To stay within biologically meaningful regions, we set the limits which in the E_m and E_d values have to lie to be $95\% < E_m \leq 100\%$ and $0\% \leq E_d < 23\%$ as explained above. The lower the limit for E_m, respective the higher the limit for E_d, the higher the stochastic variation per time-step on average, since each site (feature) is potentially less well maintained, leading to a quicker regression to the equilibrium (the perfect maintenance would be E_d=0, and E_m=1). For example, CpG sites with E_m =99% and E_d=1% will regress towards 0.5 slower than CpG sites with E_m=90% and E_d=10%. The real limits for E_m and E_d are currently not known and we, therefore, estimated them as explained above. We wondered whether we could estimate the limits for E_m and E_d such that the epigenetic age prediction of our simulated data would be as accurate as possible regarding the simulated age. We tested multiple combinations of limits for E_m and E_d and calculated the R^2 as a measure of accuracy between the predicted and the simulated age (**Figure 3A**). Horvath's epigenetic clock has the highest accuracy in predicting the simulated age with the limits $97\% < E_m \leq 100\%$ and $0\% \leq E_d < 5\%$, which is indeed a narrower range for E_m and E_d as we previously assumed (**Figure 3A**). Indeed, the prediction with Horvath's epigenetic clock caps-off later with these new limits (**Figure 3B**, compare **Supplement Figure 3A**). Even more surprisingly, the same is true if all CpG sites were simulated with a universal maintenance efficiency of 99%, with Pearson correlations going as high as 0.97 (the R^2, however, is with ~0.5 lower than for the empirically estimated maintenance efficiencies in Figure 3A,B) for a maintenance efficiency of 99 % for all CpG sites (**Figure 3C**). The Pearson correlations are robust to the universal methylation maintenance efficiency, but peak at 99% (**Supplement Figure 3B**). Setting a low maintenance efficiency of 90 % leads to a reduced Pearson correlation (**Supplement Figure 3B**) since the features reach the equilibrium faster and therefore cap off quicker (compare **Figure 2B**). Setting a high maintenance efficiency of 99.95 % reduces the Pearson correlation as well, due to the reduced speed of convergence (**Supplement Figure 3B**). Notably, Horvath's clock predicts an old age of 69.4 years for a dataset with DNA methylation levels of 0.5 for all CpG sites.

Even the second generation aging clock PhenoAge [30], which is built on a proxy for biological age based on clinical biomarkers, showed the same behavior (**Figure 3D-F, Supplement Figure 3C,D**). The previously assumed limits for E_m and E_d led to a similar linear increase, and early cap-off (**Supplement Figure 3C**), which could be improved upon estimating better limits (**Figure 3D,E**), which coincidentally are the same as estimated with Horvath's clock, i.e. $97\% < E_m \leq 100\%$ and $0\% \leq E_d < 5\%$. PhenoAge as well significantly correlates with the simulated age of samples simulated with a universal maintenance efficiency of 99% (**Figure 3F**), which as well was robust to the maintenance efficiency chosen (**Supplement Figure 3D**).

Interestingly, the effect of the age of the sample used for the ground state is in line with the expectation that we had if epigenetic clocks would indeed measure the amount of stochastic variation in the data. Starting the ground state with a sample from a 16-year-old and simulating the addition of up to 100 stochastic variations shows the previously shown linear increase in predicted age with a cap-off (**Supplementary Figure 3E**). Starting the ground state with a sample from a mid-aged 37-year-old, starts the prediction higher, shows a smaller linear increase in the predicted age, and leads to a quicker arrival and longer time at the cap (**Supplement Figure 3F**). Starting the ground state with a sample of an old 81-year-old, does not show a difference in the prediction upon stochastic variation, indicating that the ground state is already containing as much stochastic variation as we would expect at the cap-off (**Supplement Figure 3G**). All tested first generation aging clocks showed the same pattern as Horvath's and the PhenoAge clock. Vidal-Bralo's blood aging clock [31], Lin's clock [32], and even Weidner's aging clock that is based on just 3 CpG sites [33] significantly correlated with the simulated age, independent on whether empirically estimated or universal maintenance efficiencies were applied (**Supplementary Figure 4A-F**).

In conclusion, we show here that first as well as second generation aging clocks significantly correlate with the amount of stochastic variation added to a young biological starting point and this irrespective of whether empirically estimated or universal maintenance rates were assumed. This indicates that chronological as well as biological age correlate with stochastic changes in DNA methylation data.

Stochastic data-based aging clock

Motivated by these results, we wondered whether a clock built on simulated data could predict the chronological age of biological samples with known chronological age. To this end, we generated data in the same manner as above, starting from the youngest sample of the biological dataset (GSM1007467 [26]), empirically estimated E_m and E_d values (within the limits found above, i.e.

$97\% < E_m \leq 100\%$ and $0\% \leq E_d < 5\%$), trained an Elastic net regression model on simulated samples, and predicted the chronological age of biological samples. As before, starting from the same ground state, the data is randomly changed based on the site-specific E_m and E_d values in each time-step (**Figure 2A**), i.e. a sample of simulated age 1 underwent one single-cell stochastic variation application, a sample of age 100 underwent the same stochastic variation application independently 100 times; older simulated samples accumulated therefore more stochastic variation on average. Note, that the scale and units of the simulated age are arbitrary, and different from the chronological age of biological samples. The Elastic net regression is trained to predict the simulated age, i.e. how often stochastic variation was added to the starting values of all features in the ground state, as described in Figure 1B. Similar to Horvath's epigenetic clock, we found it to be beneficial if we applied a scaling of the simulated age before training the model (see methods for details). Note, that since the simulated age has arbitrary time-steps this won't interfere with any conclusion drawn from the predictions of a model trained with scaled simulated age. The trained age predictor was then applied to the independent blood DNA methylation dataset, excluding the youngest and oldest sample as they were used to define the ground state and as basis for estimating the methylation maintenance efficiencies, respectively. Surprisingly, a small simulated training dataset with one simulated sample per stochastic variation generation starting with the CpG sites from Horvath's epigenetic clock led to a significant Pearson correlation of 0.87 (p-value=1.66e-119) of chronological age and the predicted simulated age, i.e. how often stochastic variation was applied to the data (**Figure 3G**). While the scale, units, and axes of the prediction is different as mentioned above, this indicates that accumulating stochastic changes correlate linearly with chronological age. To verify that this is reproducible not only with Horvath's epigenetic clock sites, but also randomly chosen CpG sites, we used the same set-up as explained above, but with different randomly chosen ground states with different features sizes, i.e. amount of CpG sites used for training. Interestingly, the significant correlation between the chronological age and the stochastic data-based model predictions seems to be largely robust to the amount of CpG sites used for training (**Supplement Figure 5A**). In order to validate that chronological age is the relevant factor for the prediction, we used the same setup as above but permutated the chronological age of all samples randomly. The results show that no significant correlation can be made (**Supplement Figure 5B)**. This indicates that the results of a predictor that is trained to predict how often independent stochastic variation was applied to a common ground state, indeed correlate with the chronological age of biological samples.

In conclusion, our analysis shows that simulating stochastic data starting from a young biological sample with site-specific maintenance rates, allows building an Elastic net regression model whose predictions are significantly correlated with the chronological age of biological samples. The only biological data in the regression model is the ground state as well as the estimated maintenance rate efficiencies, indicating that our simulated stochastic changes to the ground state capture a process that is resembling the aging process.

Transcriptomic biological age prediction

To verify, that not only human DNA methylation data can be used, and that stochastic variation not only enables the prediction of the chronological, but also biological age, we next sought to validate our approach in transcriptomic data from *C. elegans*. *C. elegans* has the unique advantage that the effect on lifespan of genetic, environmental, and pharmacological interventions is known and corresponding transcriptome data are available thus allowing the determination of relative biological age. Moreover, we can directly estimate the biological age via temporal rescaling and we have recently shown that binarization of transcriptomic data allows for an highly accurate biological age prediction of *C. elegans* with the BitAge clock [34]. First, we tested whether BitAge would be able to predict a linear trend in the simulated data, similar to the epigenetic clocks in Figure 3 and S4. For this we defined the ground state as the biologically youngest adult sample (GSM2916344 [35]) in our dataset and simulated stochastic variation similar as explained in Figure 1A. We rescaled the log-transformed RNA-seq counts to be within 0 and 1, and simulated samples by adding stochastic variation for an age range of 1-16 (the average *C. elegans* lifespan after temporal rescaling is ~16 days [34]). Note that the age range is arbitrary, and the scale and unit not directly comparable to the biological age. The stochastic variation was sampled and added as explained in Figure 1A. We simulated 10 samples per age step and binarized them as described previously [34]. These simulated binarized samples were used as input for the BitAge prediction. In accordance with the epigenetic clock results, BitAge predictions as well correlate linearly with the amount of stochastic variation in the data (**Figure 3H**). The correlation is robust to the amount of stochastic variation added in each time-step, with a peak in Pearson correlation of 0.81 at stochastic variation sampled from a normal distribution with a standard deviation of 0.01 (**Supplement Figure 5C**). This indicates that not only the predicted human epigenetic age, but also the predicted transcriptomic age of *C. elegans* correlates with age-dependent stochastic variation in the data.

In Figure 3G and Supplement Figure 5A,B we have shown that a stochastic-data based clock is linearly correlated with the chronological age of humans. Next, we wondered whether we could not only build a predictor that is significantly correlated with the chronological, but also the biological age of an organism. For this we simulated data based on the distribution that resulted in the best correlation with BitAge, i.e. a normal distribution with a standard deviation of 0.01 (**Supplement Figure 5C**), and trained an Elastic net regression model using the same approach as defined above. We started with the scaled counts of the youngest adult RNA-seq sample, applied stochastic variation 1-16 times sampled from a normal distribution with mean of 0 and standard deviation of 0.01, and trained an Elastic net regression model to predict the simulated age, i.e. how often stochastic variation was added to the ground state. Similar to our results above, we found it to be beneficial if we applied a scaling of the simulated age before training the model (see methods for details). Note, that since the simulated age has arbitrary time-steps this won't interfere with any conclusion drawn from the predictions of a model trained with scaled simulated age. Once trained we applied the model to predict *C. elegans* RNA-seq samples with known biological age. For this we processed 994 RNA-seq samples for which the biological age could be calculated (Supplementary Table 1, see methods for details). The Elastic net regression model based on a transcriptomic ground state and stochastic variation is significantly correlated with the biological age of public RNA-seq samples (excluding the youngest sample which was used for the ground state) with a Pearson correlation of up to 0.7 between the predicted simulated age and the biological age of *C. elegans* samples (**Figure 3I**). This prediction is again robust to the number of features, i.e. genes, used in the simulation (**Supplement Figure 5D**). And similar to the epigenetic results in Figure 3G and Supplement Figure 5A,B, a permutation of the biological age does not correlate with the predicted simulated age, indicating that the correlation is driven by the biological age (**Supplement Figure 5E**). The only biological data incorporated in this model is the biological ground state, while the stochastic variation is sampled from a normal distribution centered at 0 with a fixed standard deviation. These results indicate that not only human epigenetic clocks, but also *C. elegans* transcriptomic clocks correlate linearly with stochastic variation in the data. Moreover, a transcriptomic stochastic-data based clock can be built, whose predictions are significantly correlated with the biological age of samples.

Discussion

Biological systems are known to become noisier with progressing age. Aging has been shown to affect stochastic DNA methylation drifts and subsequent degradation of transcriptional networks in mouse muscle stem cells [36], increased cell-to-cell gene expression variation has been observed with aging [37], and the stability of transcriptional networks has been closely linked to aging and stress resistance [38]. These stochastic methylation changes could appear every time a DNA methylation site has to be copied or maintained. Stochastic DNA damage, a central factor in the aging process [7], leads to DNA repair and Dnmt1 recruitment to maintain DNA methylation pattern during the repair process [39]. DNA replication involves the copying and maintenance of all CpG methylation sites [40]. Interestingly, replication timing during S-phase itself has been shown to affect methylation maintenance levels [41]. This is in line with Jenkinson et al.'s information-theoretic approach to the epigenome [15], since higher maintenance, and therefore lower information loss, consumes more energy and is, therefore, focused on more crucial regions of the genome. Such regions also tend to be replicated earlier, e.g. constitutively active housekeeping genes always replicate early [42].

It was suggested that 90 % of CpG sites are driven by non-stochastic genetic and environmental factors, while only 10 % are driven by biological stochastic variation [43]. Our single-cell simulation results, in contrast, are in line with a recent publication by Tarkhov et al. [20] that showed increased single-cell DNA methylation level heterogeneity with age. Tarkhov et al. also simulated stochastic changes in single-cell DNA methylation based on a different approach via exponential decay and that when starting a simulation with either 0 or 1 for all sites before applying stochastic changes, a prediction with high accuracy is possible. This is in line with the regression-to-the-mean model, since each site starts at the extreme and can only diverge from it [20].

Here, we extended on these results, by showing that not only simulated datasets starting at the extremes (0 or 1), but also uniformly distributed or biological starting points allow for robust predictions with normal-distributed as well as with empirically estimated noise. Our extended simulations also do not only apply to DNA methylation data, but essentially all data types which have range limits or can be normalized to be within 0 and 1. Therefore, in addition to methylation clocks, any set of biological measures, whether molecular or physiological, could be used for building aging clocks.

We find that first as well as second generation DNA methylation aging clocks significantly correlate with the amount of stochastic variation in the data suggesting that chronological as well as biological aging clocks are measuring stochastic variation. Interestingly, the prediction of all tested clocks caps off after a certain amount of stochastic variation in the data and stays stable afterwards. This could indicate that the simulated data approached the site-specific equilibria and stochastic deviations from it are not interfering with the predicted age anymore. In line with this, it has been shown that epigenetic clocks tend to underestimate the age of older subjects [44], which could be due to the DNA methylation sites reaching their equilibrium states.

We have validated the predictive power of stochastic data to predict biological age by using our recently developed transcriptomic BitAge aging clock [34] with data simulated from a young *C. elegans* transcriptome. BitAge was designed to diminish the amount of aging-irrelevant variability in the data by binarization, which allowed a highly accurate biological age prediction. Interestingly, even after binarization with BitAge the amount of simulated stochastic variation significantly correlates with the predicted age by BitAge. As the lifespan data of each of those *C. elegans* datasets is known, this approach allows determining how well biological aging could be predicted. The significant predictive power of a clock build on stochastic data thus further emphasizes the underlying effect of stochasticity in driving biological aging.

The fact that aging clocks strongly correlate with the amount of stochastic variation also cautions with regards to the identification of causal effects. CpG sites that show increased random variation faster are likely to be less efficiently maintained and are therefore not that important for cell survival or homeostasis. Targeting aging clock CpG sites might thus be unsuitable for the development of novel geroprotectors [11]. This is also in line with a recent report showing that many chronological aging clocks can be built from DNA methylation data and that clock CpG sites have limited value for the understanding of biology or potential anti-aging interventions [45].

The validity of entirely stochastic data-based aging clocks demonstrates the compatibility of precise measures of the pace of aging with the evolutionary theory of aging and entropy-driven stochastic variations in biological processes such as damage accumulation as the main drivers of aging. These results emphasize that a precise measure of the pace of aging does not require a programmed process underlying aging and, in contrast, underline the stochastic nature of the molecular alterations during aging. While maintenance levels will be highest in younger years, stochastic errors will start accumulating from conception. This will start a vicious spiral, since every additional error might disturb the intricate regulatory networks of the cell, leading to stronger and stronger defects in maintenance and thereby allowing for more errors to be made. Our data also suggest that the main accelerator and decelerator of aging is the degree of random variation. Reducing such random variation by tightening regulatory mechanisms such as methylation maintenance and gene expression would be predicted to slow the aging process and could provide targetable approaches.

Lastly, the stochasticity underlying aging clocks unifies the concept of exact determination of age and the evolutionary theory of aging. Aging occurs as a consequence of reduced maintenance of homeostatic processes. Maintenance and repair genes have been selected for maximizing the transfer of genes through the germline [46]. After reproduction the selective force for somatic maintenance and repair vanishes and, therefore, the tightness of regulatory processes is loosening and consequently stochasticity in the physiological processes increases. Indeed, our analysis predicts that the level of such stochasticity sets the pace of aging. Reinstating regulatory tightness could therefore provide opportunities for aging decelerating therapies.

Acknowledgements

We thank Khrystyna Totska for comments on the manuscript and the Regional Computing Center of the University of Cologne (RRZK) for providing computing time and support on the DFG-funded High Performance Computing (HPC) system CHEOPS. D.M. was supported by 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, SCHU 2494/15-1, CECAD EXC 2030 – 390661388, SFB 829, KFO 286, KFO 329, and GRK 2407), the Deutsche Krebshilfe (70114555), the

H2020-MSCA-ITN-2018 (Healthage and ADDRESS ITNs) and the John Templeton Foundation Grant (61734).

B.S. is co-founder of Agevio Therapeutics, Inc.

Methods

Bulk Simulations

A ground state was generated with 2000 (or indicated otherwise) random features between 0 and 1. From this ground state 6 independent sets of 100 samples each (one sample per age from 1-100) were generated. Each of these 600 samples started from the same ground state with slight deviations, i.e. each sample started with stochastic variation generated from $N(\mu = 0, \sigma^2 = 0.01^2)$ added to the ground state to simulate biological variation. To model age-dependent stochastic variation accumulation, random noise was generated from a normal distribution $N(\mu = 0, \sigma^2)$ with random.randn() from Numpy v.1.18.5 [47]. The standard deviation σ used for generation of stochastic variation that is applied at each time-step is indicated in the figure legends. The simulated age of each sample defined how often stochastic variation generated from $N(\mu = 0, \sigma^2)$ was independently added to the ground state. After stochastic variation addition values were kept between 0 and 1, by setting values bigger 1 to 1 and values smaller 0 to 0 (except for the results in Figure 1C, where no limits where applied). To train a predictor of the simulated age we used 3 sets of 100 independent samples for training of an Elastic net regression model with ElasticNetCV from sklearn v.0.23.1 [48] with the following 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 samples were used as a hold-out validation dataset.

Stochastic variation quantile distribution estimation from bulk data

To estimate stochastic variation from bulk DNA methylation data we compared 2 public blood DNA methylation samples from a healthy 16-year-old male subject (GSM1007467) and a healthy 17-year-old male subject (GSM1007336) from GSE41037 [26]. We chose samples from subjects one-year apart to model the amount of stochastic variation that accumulates within a year. Since DNA methylation sites close to 0 or 1 have a smaller stochastic variation distribution than sites in the middle of the distribution, we divided the data into 5, 10, 15, or 20 quantiles based on the DNA methylation levels of the youngest subject (**Supplementary Figure 1D**) and estimated the stochastic variation for each of the quantiles. To estimate the distribution of each quantile, we first computed the DNA methylation site specific differences between GSM1007336 and GSM1007467 (**Supplementary Figure 1E**), and used Python's Fitter() function from fitter v.1.4.0 [49] with the parameter distributions=['lognorm'] to estimate the best fitting lognorm distribution for each quantile (**Supplementary Figure 1F**).

Noise quantile distribution application

The ground state consists of randomly sampled 2000 (or indicated otherwise) CpG sites of the youngest sample in GSE41037 [26] (GSM1007467). Each CpG site was sorted into the corresponding quantile and stochastic variation was generated from the fitted distribution with Scipy's [50] stats.lognorm.rvs() function with the parameters scale, loc, and s estimated as described above with the Fitter() function. The simulated age of each sample defined how often stochastic variation generated from the fitted function was independently added to the ground state. After stochastic variation addition values were kept between 0 and 1, by setting values bigger 1 to 1 and values smaller 0 to 0. To train a predictor of the simulated age we used 3 sets of 100 independent samples for training of an Elastic net regression model with ElasticNetCV from sklearn v.0.23.1 [48] with the following 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 samples were used as a hold-out validation dataset.

Single-cell simulations

The ground state of single-cell simulations consists of 2000 (or indicated otherwise) randomly chosen CpG sites of the youngest sample in GSE41037 [26] (GSM1007467). Each of the features (CpG sites) is a number between 0 and 100 % and used to generate 1000 cells with binary values for each feature. A ground state value of 0.13, i.e. 13 % methylated, generates 1000 cells for which 130 are 1 (methylated), and 870 are 0 (unmethylated). One sample therefore consists of 2000 (or indicated otherwise) features with each 1000 simulated cells with binary values of either 1 or 0. Next, for each feature a methylation maintenance efficiency E_m and *de novo* methylation efficiency E_d was generated. As indicated in the figure legends, we either simulated data with a universal maintenance efficiency for all features, or we estimated E_m and E_d from empirical data. For the empirical maintenance estimation, we set the site-specific DNA methylation equilibrium to be the value of the oldest sample in the dataset (GSM1007832 [26]) and estimated E_m and E_d from the equation given by Pfeifer et al. [16]:

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

, where M_{eq} is the equilibrium of the methylation state, while retaining E_m and E_d within the limits E_m > 95 % and E_d < 23 % (or indicated otherwise), as defined in previous publications[16,17,28].

Public aging clocks

We downloaded the Elastic net regression coefficients for Horvaths pan-tissue clock[29], Vidal-Bralo's blood aging clock [31], Lin's 99-CpG clock [32], Weidner's 3-CpG clock[33], and Levine's PhenoAge[30] clock and applied them to simulated data. The data were simulated as defined above, with the difference that we only used the clock-specific CpG sites as the features in the ground state, and we started the arbitrary simulated age at 16, i.e. the age of the subject of the ground state sample. Stochastic variation was simulated either with a universal maintenance efficiency for all CpG sites, or with empirically estimated maintenance rates as defined above.

Stochastic data-based clock

The stochastic data-based clock was computed based on simulations described above and used for predictions on all samples in the dataset GSE41037 [26]. We found that a rescaling of the simulated age before training and testing the model is beneficial. First, we rescaled via min-max scaling the simulated age to be within 0 and 1, multiplied it by 400 and subtracted 120. Note that this transformation on the arbitrary time-steps will not interfere significantly with the correlation analyses. For the correlation analyses, we excluded the youngest (GSM1007467; from which the ground state was sampled), and the oldest (GSM1007832; from which the maintenance efficiencies were estimated as described above) to not confound the correlation between the chronological age of samples in GSE41037 [26], and the predicted age. To train a predictor of the simulated age we used 1 set of 100 independent samples for training of an Elastic net regression model with ElasticNetCV from sklearn v.0.23.1 [48] with the following parameter: l1_ratio=[0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9].

Public RNA-seq processing

All 994 public RNA-seq samples were downloaded and processed the same. First, we preprocessed samples with Fastp v0.20.0 [51] with the following parameters -g -x -q 30 -e 30. After preprocessing, the samples were mapped with Salmon v1.1 [52] and the parameters –validateMappings –seqBias and for paired-end samples additionally –gcBias. The decoy-aware index for Salmon was generated with the WS281 transcriptome build from Wormbase [53]. The results of Salmon were combined to the gene-level with tximport v1.14.2 [54]. Raw counts were log10-transformed after the addition of one pseudo-count, each sample was min-max normalized to bring each sample within the data range 0-1, and genes 0 in all 994 samples were filtered out. Binarization and prediction with BitAge were applied as described previously [34].

Transcriptomic stochastic variation simulation

The ground state consists of all (or indicated otherwise) gene counts (normalized as described above) of the biologically youngest sample (GSM2916344 [35]). From this ground state 10 independent samples for each time-step (from 1 to 16) were generated and used to train an Elastic net regression as described above (see Bulk simulations). Similar to the epigenetic stochastic-data based clock we found a rescaling of the arbitrary simulated time-steps by 2 to be beneficial, i.e. we multiplied the simulated age by 2 before training and testing the data. The Elastic net regression model was then used to predict the biological age of the 993 remaining *C. elegans* samples.

Code availability statement

Code for all simulations will be made public at https://github.com/Meyer-DH/ .

Figure Legends

Figure 1

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 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 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 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. 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. 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.

F) Independent Elastic net regression models are highly correlated if trained on samples starting from the same ground state. 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.

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). 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). 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.

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 $N(\mu = 0, \sigma^2 = 0.05^2)$. 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.

Figure 2

A) Explanation of single-cell simulations. Briefly, values from a bulk DNA methylation sample are used to generate 1000 binary cells for each feature (CpG site). Each feature is flipped randomly based on the site-specific methylation maintenance efficiencies E_m and E_d for each time-step. After each time-step, i.e. stochastic variation application, the average of all 1000 cells per feature is calculated and used as a sample for subsequent analyses.

B) The accuracy of the model is dependent on the methylation maintenance efficiency rate. An Elastic net regression model was trained on n=300 samples (3 samples per time point) starting from the same ground state with 500 features and universal maintenance efficiencies E_m and E_d (E_u) and used to predict the simulated age of 300 independent validation samples. The x-axis shows the methylation maintenance efficiency E_m in %, E_d was set to 100-E_m, i.e. E_u=E_m). The y-axis shows the R² as a measure of model accuracy of the independent validation data. All samples within the training and validation dataset started from the same ground state of 500 randomly sampled features from the youngest healthy sample (GSM1007467) in GSE41037 [26]. 3 independent experiments with different ground states are shown for each maintenance efficiency.

C) Single-cell simulation of DNA methylation sites based on E_m and E_u allows to build highly accurate predictions. An Elastic net regression model was trained on n=300 samples (3 samples per time point) starting from the same ground state and universal maintenance efficiencies E_m and E_u of 99.9 %. 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. The R² is 0.999. All samples within the training and validation dataset started from the same ground state of 500 randomly sampled features from the youngest healthy sample (GSM1007467) in GSE41037 [26]

D) The accuracy of predictions with a universal maintenance efficiency rate of 99.9 % caps off after ~32 features in the ground state with an R² of 0.99. The x-axis shows how many features were sampled for the ground state that was used to build and validate an Elastic net regression model the same as in B) and C). The y-axis shows the R² 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.

E) The maintenance efficiency rate sets the pace of aging. The Elastic net regression model was trained the same as in B) and C) with a maintenance efficiency of E_m=E_u=99.9 %. Color-coded are different independent validation samples, sampled from the same ground state, but with different maintenance efficiency rates. Samples with higher methylation efficiencies (99.99%)

accumulate less stochastic variation and are predicted to age slower, i.e. the slope of the prediction is lower. Samples with lower maintenance efficiencies (95%, 99%) accumulate stochastic variation 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.

F) Biologically estimated maintenance rates allow for highly accurate predictions. Site-specific E_m and E_u values were estimated from data (see methods for details). The simulations were done the same as in C) but with site-specific maintenance rates.

Figure 3

A) The methylation maintenance efficiency limits affect the simulation and subsequent prediction with Horvath's epigenetic clock [29]. The x-axis shows the limit of E_m, i.e. E_m has to be bigger than the depicted limit. Color-coded is the limit of E_d, i.e. E_d has to be smaller than the depicted limit. The site-specific maintenance efficiencies are estimated as described in the methods based on the oldest sample in the dataset (GSM1007832 [26]) and to be within the specified limits. The y-axis shows the R² as a measure of accuracy between the predicted epigenetic age by Horvath's epigenetic clock [29] and the simulated age, i.e. how often stochastic variation was applied to the ground state. E_m > 97 % and E_d < 5 % has the highest accuracy.

B) Horvath's epigenetic age prediction[29] 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) [26], 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.

C) Horvath's epigenetic age prediction[29] of samples simulated based on a universal maintenance efficiency rate of 99 % for all features (CpG sites) starting from biological data from a young human blood sample (GSM1007467) [26], 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.

D) The methylation maintenance efficiency limits affect the simulation and subsequent prediction with PhenoAge[30]. The x-axis shows the limit of E_m, i.e. E_m has to be bigger than the depicted limit. Color-coded is the limit of E_d, i.e. E_d has to be smaller than the depicted limit. The site-specific maintenance efficiencies are estimated as described in the methods based on the oldest sample in the dataset (GSM1007832 [26]) and to be within the specified limits. The y-axis shows the R² as a measure of accuracy between the predicted epigenetic age by

PhenoAge[30] and the simulated age, i.e. how often stochastic variation was applied to the ground state. E_m > 97 % and E_d < 5 % has the highest accuracy.

E) Biological age prediction with PhenoAge[30] 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) [26], 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.

F) Biological age prediction with PhenoAge[30] 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) [26], 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.

G) The predictions of an Elastic net regression model based on simulated data , correlates significantly (Pearson correlation 0.87, p-value=1.66e-119) with the chronological age of the biological samples (GSE41037) [26] . 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.

H) Transcriptomic age prediction with BitAge [34] of simulated data starting with the biologically youngest adult *C. elegans* sample in the used dataset (GSM2916344 [35]) as the ground state and subsequent addition of stochastic variation sampled from a normal distribution centered at 0 with a standard variation of 0.01. The x-axis shows the simulated age, i.e. how often stochastic variation was added to the ground state, the y-axis shows the results of the BitAge prediction in days after binarization of the simulated data. The number of stochastic variation additions and the BitAge prediction significantly correlate with a Pearson correlation of 0.81, p-value=8.74e-39.

I) The predictions of an Elastic net regression model based on simulated data starting with the biologically youngest adult *C. elegans* sample (GSM2916344 [35]) as the ground state and subsequent addition of stochastic variation sampled from a normal distribution centered at 0 with a standard variation of 0.01, correlates significantly (Pearson correlation 0.7, p-value=3e-147) with the biological age of the 993 RNA-seq samples (excluding the sample from which the ground state was sampled) . The x-axis shows the biological age of the 993 adult *C. elegans*

RNA-seq samples (excluding the sample from which the ground state was sampled), the y-axis shows the predicted simulated age of the Elastic net regression model based on stochastic-data, 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.

Supplementary Figure Legends

Supplementary Figure 1

A) Comparison between the ground state on the x-axis, and the ground state 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 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 GSE41037 [26]. Every dot depicts a DNA methylation site. Values close to 0 and 1 show less variation than values closer to 0.5.

D) Quantile distribution of the ground state. The histogram depicts the data distribution of the DNA methylation data of our ground state (GSM1007467 [26]). The 5 quantile ranges are depicted with the vertical black lines (one quantile contains all DNA methylation sites between two consecutive lines). The left-most line shows 0, the right-most line 1. Since most CpG sites start with a DNA methylation close to 0, more quantiles with narrower ranges are close to 0.

E) Distribution of the variance between two samples of young subjects one-year apart from each other. The x-axis shows the youngest sample (GSM1007467) in the used public dataset GSE41037 [26]. The y-axis shows the difference between GSM1007336 and GSM1007467.

F) Empirical distributions of 5 quantiles. The colored lines show the empirically data distributions of 5 quantiles. For each of the quantiles shown in Supplement Figure 1D we calculated the differences (see Supplement Figure 1E) of all CpG sites falling into the quantile and estimated the empirical distribution from these differences. For each of the quantiles a different distribution is estimated. The distribution of Quantile 1 (blue) is the narrowest around 0, indicating that CpG sites in the Quantile 1, i.e. the left-most quantile in Supplement Figure 1D, have the least amount of stochastic variation. Quantile 4 is the quantile with the biggest range in Supplement Figure 1D and broadest empirical distribution.

G) Stochastic variation sampled from empirically estimated distributions allows accurate age predictions. To empirically estimate the stochastic variation between 2 samples, the data were split into quantiles (Supplement Figure 1D-F, see methods for details). The number of quantiles has no effect on the accuracy of the model. The x-axis shows the number of quantiles into which the data was split to predict stochastic variation distributions, i.e. the number depicts the number of stochastic variation distributions from which stochastic variation is sampled. The y-axis shows the R^2 of the independent validation data.

H) Prediction results of the independent validation data for empirical stochastic variation with 5 quantiles. All samples within the training and validation dataset started from the same ground state of 2000 randomly sampled features from the youngest healthy sample (GSM1007467) in GSE41037 [26] . For every whole simulated age step from 1 to 100, stochastic variation sampled from the empirically estimated stochastic variation distributions 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 samples were used for validation. 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 predicted age 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.

I) Comparison of the ground state on the x-axis (2000 randomly sampled features from the youngest healthy sample (GSM1007467 [26])) and the ground state after 100x stochastic variation additions sampled from the 5 quantiles on the y-axis.

Supplementary Figure 2

A) Comparison of the ground state on the x-axis (2000 randomly sampled features from the youngest healthy sample (GSM1007467 [26])) 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).

B) 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.

C) 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.

D) 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).

E) Comparison of the ground state on the x-axis (2000 randomly sampled features from the youngest healthy sample (GSM1007467 [26])) and the ground state after applying 100x single cell stochastic variation steps with empirically estimated maintenance efficiency rates (y-axis).

F) The prediction in Figure 2F) 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 (GSM1007467 [26]) as the ground state, the y-axis the Elastic net regression coefficients for the model in Figure 2F). 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.

Supplementary Figure 3

A) Horvath's epigenetic age prediction[29] 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) [26], correlates significantly with the simulated age, i.e. how often stochastic variation was applied to the ground state.

B) Pearson correlation of Horvath's epigenetic age prediction[29] of simulated data and the true simulated age for different universal methylation maintenance efficiencies. 5 independent experiments with different ground states are shown for each maintenance efficiency.

C) Biological age prediction with PhenoAge[30] 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) [26], correlates significantly with the simulated age, i.e. how often stochastic variation was applied to the ground state.

D) Pearson correlation of biological age predictions with PhenoAge[30] of simulated data and the true simulated age for different universal methylation maintenance efficiencies. 5 independent experiments with different ground states are shown for each maintenance efficiency.

E) Horvath's epigenetic age prediction[29] 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) [26], 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 Supplement Figure 3A, but with a simulated age range from 0-99 for an easier comparison with Supplement Figure 3F,G.

F) Horvath's epigenetic age prediction[29] 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) [26], 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 Supplement Figure 3E, and reaches the cap-off earlier.

G) Horvath's epigenetic age prediction[29] 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) [26], 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.

Supplementary Figure 4

A) Vidal-Bralo's epigenetic age prediction[31] 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) [26], correlates significantly with the simulated age, i.e. how often stochastic variation was applied to the ground state.

B) Vidal-Bralo's epigenetic age prediction[31] 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) [26], correlates significantly with the simulated age, i.e. how often stochastic variation was applied to the ground state.

C) Lin's epigenetic age prediction[32] 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) [26], correlates significantly with the simulated age, i.e. how often stochastic variation was applied to the ground state.

D) Lin's epigenetic age prediction[32] 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) [26], correlates significantly with the simulated age, i.e. how often stochastic variation was applied to the ground state.

E) Weidner's epigenetic age prediction[33] 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) [26], correlates significantly with the simulated age, i.e. how often stochastic variation was applied to the ground state.

F) Weidner's epigenetic age prediction[33] 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) [26], correlates significantly with the simulated age, i.e. how often stochastic variation was applied to the ground state.

Supplementary Figure 5

A) The feature size is largely irrelevant for stochastic data-based models. 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 samples in GSE41037 [26] (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.

B) Verification of Supplement Figure 5A). Using the same approach as in Supplement Figure 5A, 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 samples in GSE41037 [26] (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.

C) The BitAge predictions in Figure 3H 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.

D) The feature size is largely irrelevant for the model in Figure 3I). 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 samples (excluding the sample from which the ground state was sampled) and the prediction of the stochastic-data based model.

E) Verification of Supplement Figure 5D). Using the same approach as in Supplement Figure 5D, 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 samples (excluding the sample from which the ground state was sampled) and the prediction of the stochastic-data based model.

Supplementary Table 1

List of the 994 RNA-seq samples used for calculating the biological age

References

1. Weismann, A. *Ueber die Dauer des Lebens; ein Vortrag*. (G. Fischer, 1882). doi:10.5962/bhl.title.21312.

2. Kirkwood, T. B. & Cremer, T. Cytogerontology since 1881: a reappraisal of August Weismann and a review of modern progress. *Hum. Genet.* **60**, 101–21 (1982).

3. Vijg, J. & Kennedy, B. K. The Essence of Aging. *Gerontology* **62**, 381–5 (2016).

4. Kowald, A. & Kirkwood, T. B. L. Can aging be programmed? A critical literature review. *Aging Cell* **15**, 986–998 (2016).

5. Medawar, P. B. An unsolved problem of biology. *Med. J. Aust.* (1952) doi:10.5694/j.1326-5377.1953.tb84985.x.

6. Williams, G. C. Pleiotropy, natural selection, and the evolution of senescence. *Evolution (N. Y).* **11**, 398–411 (1957).

7. 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).

8. Mitteldorf, J. An epigenetic clock controls aging. *Biogerontology* **17**, 257–265 (2016).

9. Horvath, S. & Raj, K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat. Rev. Genet.* **19**, 371–384 (2018).

10. Wagner, W. The Link Between Epigenetic Clocks for Aging and Senescence. *Front. Genet.* **10**, 1–6 (2019).

11. Schork, N. J., Beaulieu-Jones, B., Liang, W., Smalley, S. & Goetz, L. H. Does Modulation of an Epigenetic Clock Define a Geroprotector? *Adv. Geriatr. Med. Res.* **4**, 1–11 (2022).

12. Issa, J. Aging and epigenetic drift: a vicious cycle. *J. Clin. Invest.* **124**, 24–9 (2014).

13. Min, B., Jeon, K., Park, J. S. & Kang, Y. Demethylation and derepression of genomic retroelements in the skeletal muscles of aged mice. *Aging Cell* **18**, 1–13 (2019).

14. Shipony, Z. *et al.* Dynamic and static maintenance of epigenetic memory in pluripotent and somatic cells. *Nature* **513**, 115–119 (2014).

15. 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).

16. 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).

17. Riggs, A. D. & Xiong, Z. Methylation and epigenetic fidelity. *Proc. Natl. Acad. Sci.* **101**, 4–5 (2004).

18. Maegawa, S. *et al.* Caloric restriction delays age-related methylation drift. *Nat. Commun.* **8**, 539 (2017).

19. Levine, M. E. & Higgins-chen, A. Clock Work: Deconstructing the Epigenetic Clock Signals in Aging, Disease, and Reprogramming. (2022).

20. Tarkhov, A. E. *et al.* Nature of epigenetic aging from a single-cell perspective Results Bulk tissue epigenetic aging. (2022).

21. Tarkhov, A. E., Denisov, K. A. & Fedichev, P. O. Aging clocks , entropy , and the limits of age-reversal. 1–12 (2022).

22. Piras, V., Tomita, M. & Selvarajoo, K. Transcriptome-wide variability in single embryonic development cells. *Sci. Rep.* **4**, 1–9 (2014).

23. Smits, A. H. *et al.* Global absolute quantification reveals tight regulation of protein expression in single Xenopuseggs. *Nucleic Acids Res.* **42**, 9880–9891 (2014).

24. Eckersley-Maslin, M. A., Alda-Catalinas, C. & Reik, W. Dynamics of the epigenetic landscape during the maternal-to-zygotic transition. *Nat. Rev. Mol. Cell Biol.* **19**, 436–450 (2018).

25. Gladyshev, V. N. The Ground Zero of Organismal Life and Aging. *Trends Mol. Med.* **27**, 11–19 (2021).

26. Horvath, S. *et al.* Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biol.* **13**, R97 (2012).

27. Han, L., Lin, I. G. & Hsieh, C. L. Protein binding protects sites on stable episomes and in the

chromosome from de novo methylation. *Mol. Cell. Biol.* **21**, 3416–24 (2001).

28. Laird, C. D. *et al.* Hairpin-bisulfite PCR: Assessing epigenetic methylation patterns on complementary strands of individual DNA molecules. *Proc. Natl. Acad. Sci.* **101**, 204–209 (2004).

29. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **16**, 96 (2013).

30. Levine, M. E. *et al.* An epigenetic biomarker of aging for lifespan and healthspan. *Aging (Albany. NY).* **10**, 573–591 (2018).

31. 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).

32. 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).

33. 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).

34. Meyer, D. H. & Schumacher, B. BiT age: A transcriptome-based aging clock near the theoretical limit of accuracy. *Aging Cell* 1–17 (2021) doi:10.1111/acel.13320.

35. Senchuk, M. M. *et al.* Activation of DAF-16/FOXO by reactive oxygen species contributes to longevity in long-lived mitochondrial mutants in Caenorhabditis elegans. *PLOS Genet.* **14**, e1007268 (2018).

36. 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).

37. Bahar, R. *et al.* Increased cell-to-cell variation in gene expression in ageing mouse heart. *Nature* **441**, 1011–1014 (2006).

38. Kogan, V., Molodtsov, I., Menshikov, L. I., Reis, R. J. S. & Fedichev, P. Stability analysis of a model gene network links aging, stress resistance, and negligible senescence. *Sci. Rep.* **5**, 1–12 (2015).

39. Mortusewicz, O., Schermelleh, L., Walter, J., Cardoso, M. C. & Leonhardt, H. Recruitment of DNA methyltransferase I to DNA repair sites. *Proc. Natl. Acad. Sci.* **102**, 8905–8909 (2005).

40. Petryk, N., Bultmann, S., Bartke, T. & Defossez, P. Staying true to yourself: mechanisms of DNA methylation maintenance in mammals. *Nucleic Acids Res.* **49**, 3020–3032 (2021).

41. 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).

42. Holmquist, G. P. Role of replication time in the control of tissue-specific gene expression. *Am. J. Hum. Genet.* **40**, 151–73 (1987).

43. Vershinina, O., Bacalini, M. G., Zaikin, A., Franceschi, C. & Ivanchenko, M. Disentangling age - dependent DNA methylation : deterministic , stochastic , and nonlinear. *Sci. Rep.* 1–12 (2021) doi:10.1038/s41598-021-88504-0.

44. Khoury, L. Y. El *et al.* Systematic underestimation of the epigenetic clock and age acceleration in older subjects. doi:10.1186/s13059-019-1810-4.

45. 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).

46. Kirkwood, T. B. L. Understanding the odd science of aging. *Cell* **120**, 437–447 (2005).

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

48. Varoquaux, G. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* (2011) doi:10.1145/2786984.2786995.

49. Fitter v.1.4.0. https://github.com/cokelaer/fitter.

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

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

52. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).

53. Davis, P. *et al.* WormBase in 2022—data, processes, and tools for analyzing Caenorhabditis elegans. *Genetics* **220**, (2022).

54. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences [version 2; referees: 2 approved]. *F1000Research* **4**, 1–22 (2016).

Figures

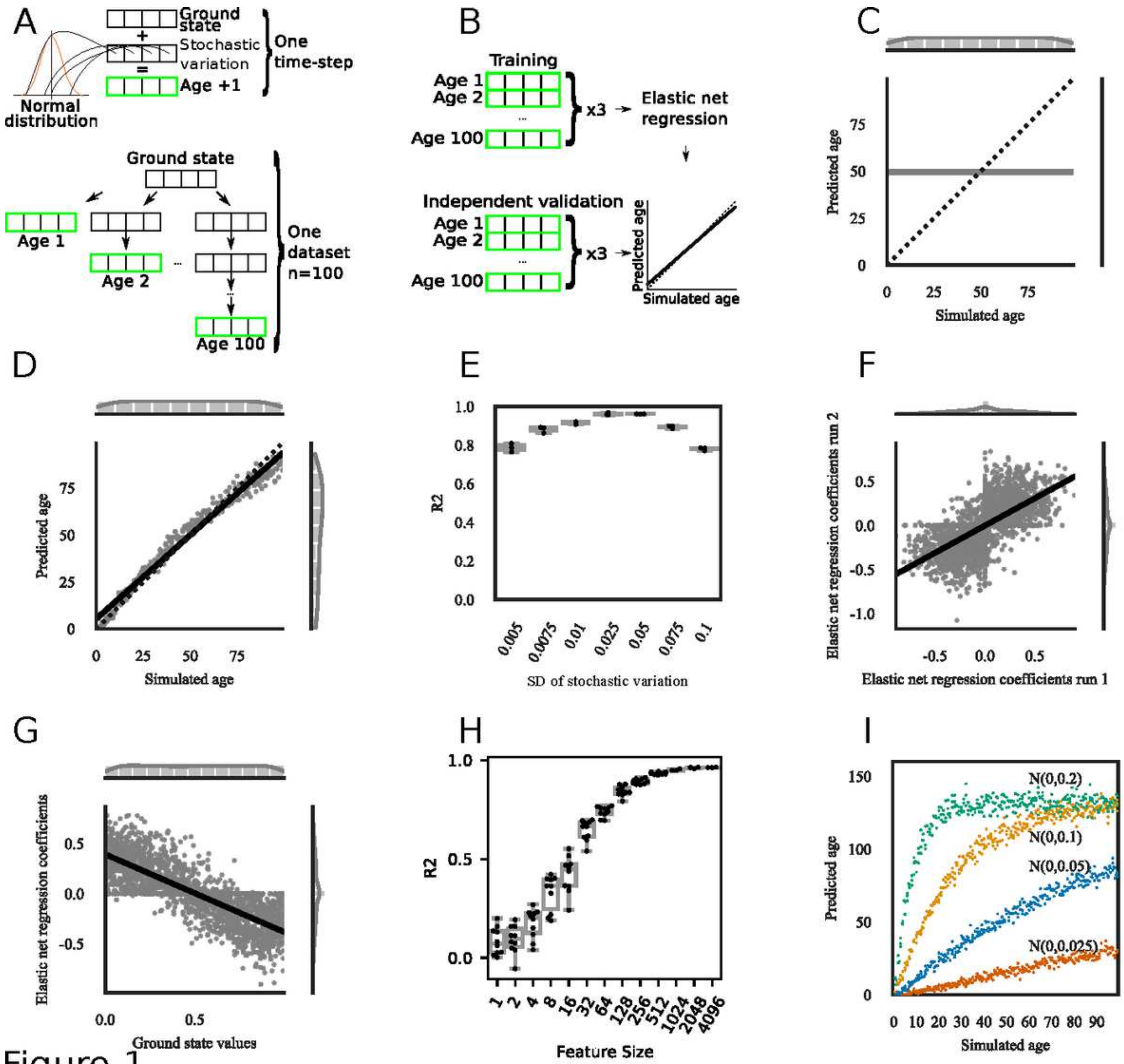

Figure 1

Figure 1

See manuscript for figure legend.

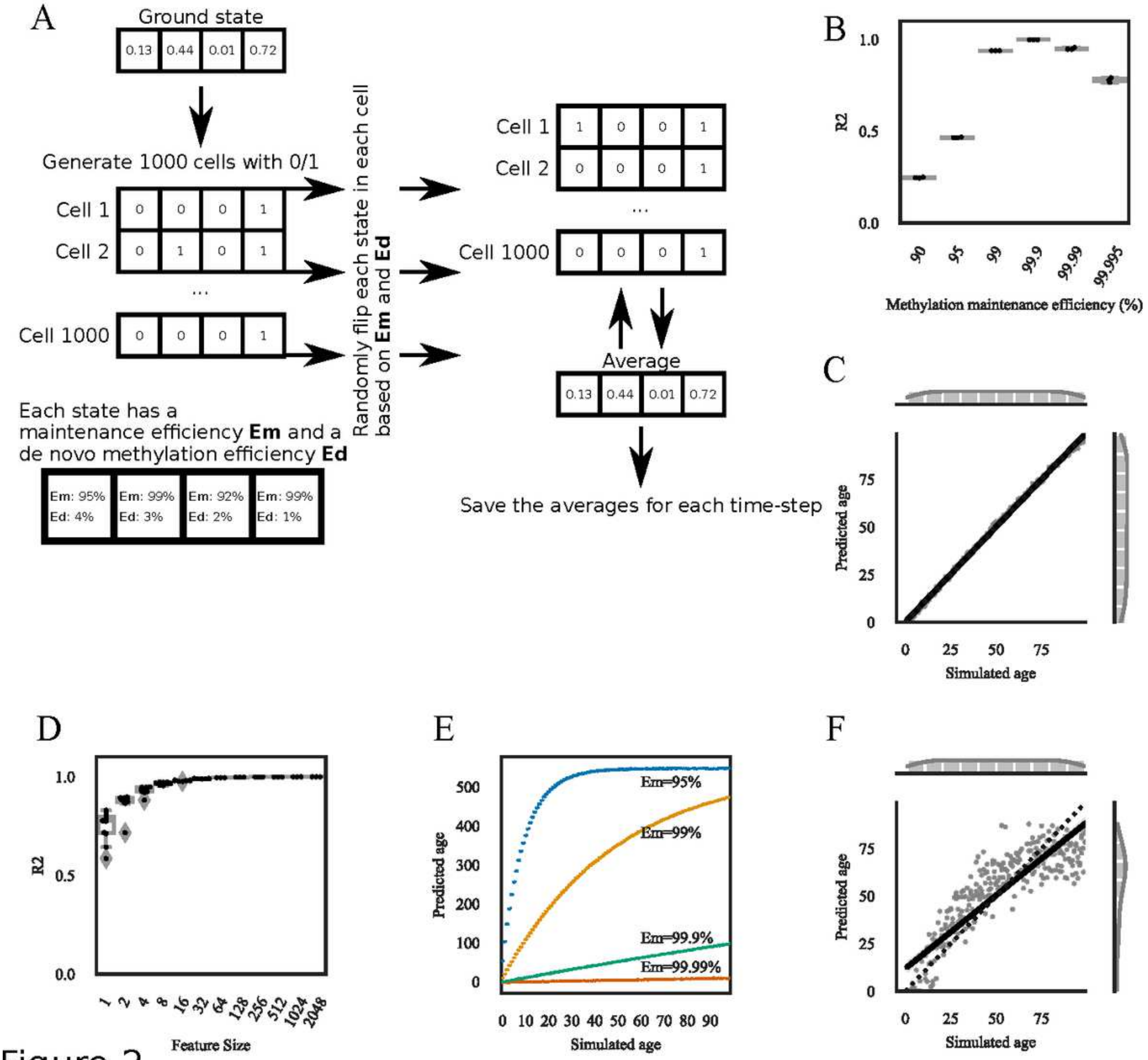

Figure 2

Figure 2

See manuscript for figure legend.

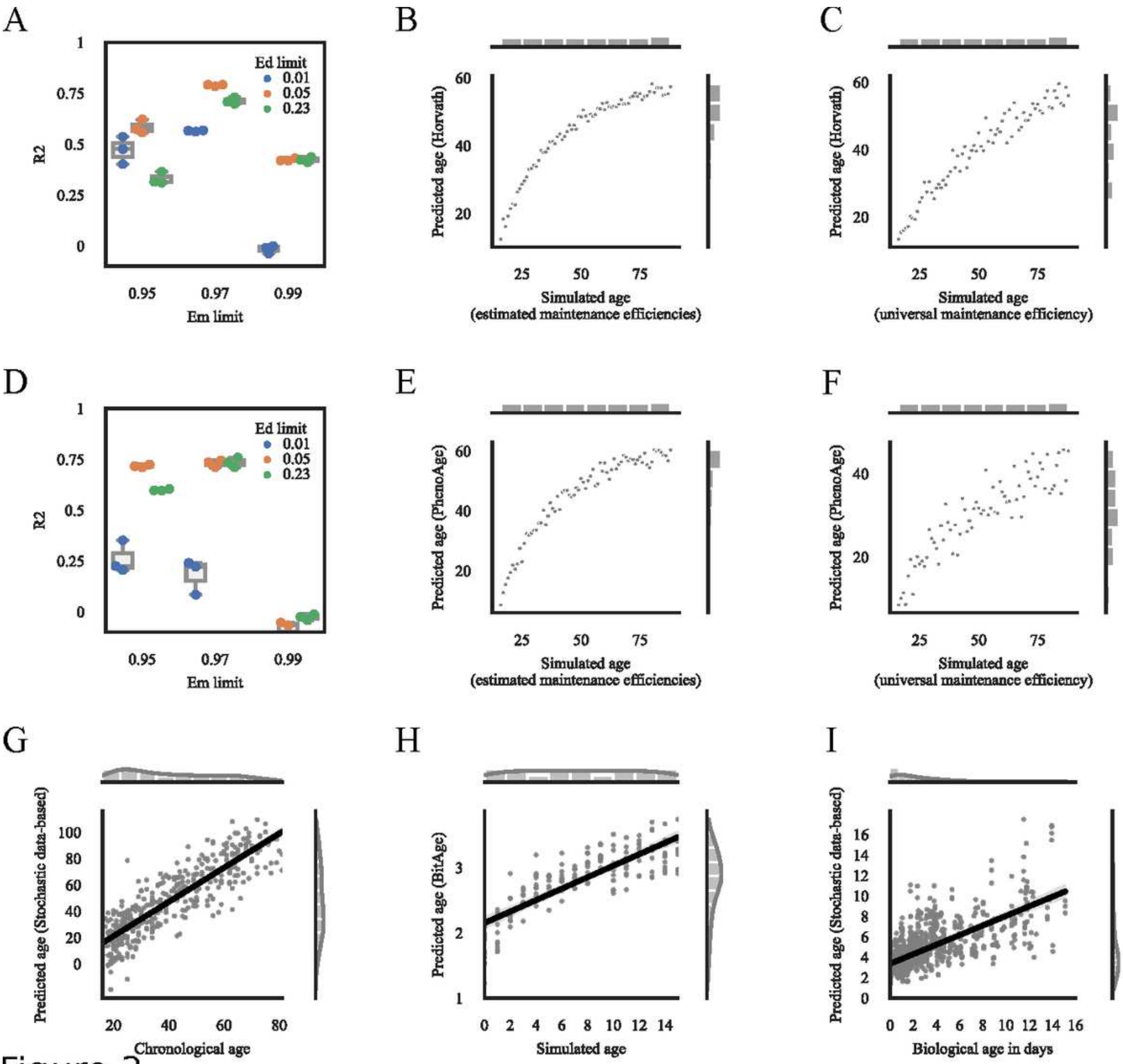

Figure 3

Figure 3

See manuscript for figure legend.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTable1.csv
- SupplementFigures06122022.pdf