EPISTEMIC UNCERTAINTY CHALLENGES AGING CLOCK RELIABILITY IN PREDICTING REJUVENATION EFFECTS

Dmitrii Kriukov[1, *], **Ekaterina Kuzmina**[1], **Evgeniy Efimov**[1], **Dmitry V. Dylov**[1, 2], and **Ekaterina E. Khrameeva**[1]

[1]Skolkovo Institute of Science and Technology, Moscow, Russia
[2]Artificial Intelligence Research Institute, Moscow, Russia
[*]Corresponding author: dmitrii.kriukov@skoltech.ru

December 1, 2023

ABSTRACT

Epigenetic aging clocks have been widely used to validate rejuvenation effects during cellular reprogramming. However, these predictions are unfalsifiable, since the true biological age of reprogrammed cells remains inaccessible. We present a multifaceted analytical framework to consider rejuvenation predictions from the uncertainty perspective. We discover that DNA methylation profiles of reprogramming are not represented in the aging data used for clock training, which introduces high epistemic uncertainty in aging predictions. Moreover, predictions of different published clocks are poorly consistent with each other and suggest even zero or negative rejuvenation. We show that the high prediction uncertainty challenges the reliability of rejuvenation effects observed during in vitro reprogramming prior to pluripotency and throughout embryogenesis. Conversely, our method also reveals a significant age increase after in vivo reprogramming. We propose to include uncertainty estimation in future aging clocks to avoid the risk of misinterpreting the results of biological age prediction.

Keywords Rejuvenation · cell reprogramming · epistemic uncertainty · epigenetic aging clocks · dataset shift · DNA methylation

Introduction

Reprogramming aged somatic cells into pluripotency or other progenitor states was repeatedly shown to ameliorate various aging-associated features, either by applying different transcription factors (TFs) or by introducing small molecules [1, 2, 3, 4, 5]. To quantify the effect of age reversal, researchers employ various methods, including the so-called "epigenetic aging clock" models built from DNA methylation (DNAm) data using diverse machine learning (ML) approaches. These approaches are employed most widely to compare the "biological age" of reprogrammed cells to that of control cells [6] (Fig. **1**a,b). These clocks are easy to use and can assess aging from the organismal to cellular levels, which is especially helpful in cases when large-scale parameters of organismal aging cannot be measured, such as *in vitro* cellular reprogramming. A similar line of thought has recently been used to demonstrate rejuvenation in the course of embryonic development [7, 8, 9].

The primary assumption of aging clocks is that the deviation Δ of predicted age from the chronological age C represents an accelerated or decelerated aging, that is, an increase or decrease in the biological age B [10, 11]. One can express it as $B = C + \Delta$. Since biological age cannot be measured directly (*i.e.*, it has no ground truth), the epigenetic age estimated by the clocks is therefore considered a proxy measure of the biological age [12]. Because of that, and also because the DNAm patterns are some of the best indicators of past influences and future health outcomes, the epigenetic age is proposed to serve as a biomarker for measuring the effects of pro-longevity interventions in clinical trials [13, 14, 10].

The obstacle is that before the aging clocks could be adopted by clinicians, these models should provide an estimate of uncertainty for their own predictions. Uncertainty manifests itself in three ways [15, 16]: (i) model choice uncertainty (part of uncertainty from a broader category, called epistemic uncertainty) reflects how well a proposed model (its architecture, parameters, metrics, *etc.*) reflects the real underlying process (Fig. **1**c); (ii) out-of-distribution (OOD) uncertainty (another type of epistemic uncertainty) arises when the testing data do not represent the training data distribution leading to a high risk of model prediction failure (Fig. **1**d); (iii) aleatoric uncertainty originates from data variations that cannot be reduced to zero by the model (*e.g.*, when the same DNA methylation level corresponds to different ages) (Fig. **1**e).

Dataset shift [17] is a term for describing the case of OOD sampling, where the testing population is under-represented in the training distribution. Dataset shift could be decomposed into a covariate shift (*e.g.*, differently distributed DNAm values) and a response shift (*e.g.*, different age ranges). The batch effect is one notorious example of dataset shift in the field of omics data analysis [18].

From the clinical standpoint, epistemic uncertainty must be estimated to make safe conclusions about whether to trust a model or not [15]. Specifically, epistemic uncertainty resulting from a dataset shift should be scrutinized, given the batch effects are so common in biological data [18, 19]. However, most popular DNAm aging clocks cannot satisfy this criterion (Fig. **1**f), because they are typically built using algorithms from the penalized multivariate linear regression (MLR) family [20] (*e.g.*, ElasticNet). Such algorithms do not yield information on any of the uncertainties, except for the error between the chronological and the predicted ages in the training data (*e.g.*, mean or median absolute errors, MAE or MedAE).

In this work, we question the applicability of existing aging clock methodology to measuring rejuvenation by taking a closer look at prediction uncertainty (Fig. **1**). We reanalyzed published data of putative rejuvenation in the extreme cases of anticipated dataset shift, such as cellular reprogramming and embryonic development.

Because biological age measurements cannot be verified explicitly, we introduce four different indirect approaches to this problem: (i) Is there a covariate shift in DNAm values between the datasets of aging and rejuvenation? (ii) Do different aging clocks agree with each other in predicting rejuvenation? (iii) Can the rejuvenation datasets be employed reciprocally to predict normal aging? And, (iv) Given an aging clock capable of estimating its own uncertainty (Fig. **1**g), would it demonstrate a significant age reversal in the putative rejuvenation experiments?

We propose a framework for answering these questions. By leveraging this framework, we expect to elucidate the most critical drawbacks of applying aging clock models to rejuvenation studies, which should be solved in order to drive a wider adoption of these models among the longevity community.

Results

Covariate shift can lead to biologically meaningless predictions

To introduce the concept of covariate shift in the field of aging clocks, we began by exploring a simple, low-dimensional example. We used two parameters (biomarkers) to construct an elementary aging clock for predicting chronological age in humans: weight and height (Fig. **2**a; see Methods). These two biomarkers strongly correlate with age during the first twenty years of human life [21], therefore they can legitimately be used for age prediction. We analyzed the data of body measurements performed for male humans ranging from 1 to 25 years old (an approximate end of body growth), among which there were both healthy controls [21] and individuals with the achondroplasia disorder [22, 23] typically characterized by a shorter length of arms and legs [24].

Following the common framework of aging clocks construction established in earlier works [25, 26, 27], we trained a multivariate linear regression (MLR) model using body measurements of a cohort of healthy individuals to predict their chronological age, which yielded good performance on the training data ($MAE = 2.3$ years, $R2 = 0.84$). For the achondroplasia cohort, these clocks predicted consistently lower ages (Fig. **2**b), which would be viewed as decelerated aging in the context of other aging clocks. However, this interpretation has no biological support because the average lifespan of people with achondroplasia is around 10 years shorter than that of control individuals due to early-life mortality [24]. We assume that this underestimation of ages in the achondroplasia cohort is caused by a covariate shift in the analyzed data, leading to a huge OOD uncertainty, which is not taken into account by the model. Indeed, the distributions of covariates (weight and height) differ between the training and testing data (Fig. **2**a, Kolmogorov-Smirnov (KS) test for distribution equality yields P-value < 0.0003). In general, any significant differences between the observed distributions in training and testing samples should caution us against applying ML models uncritically.

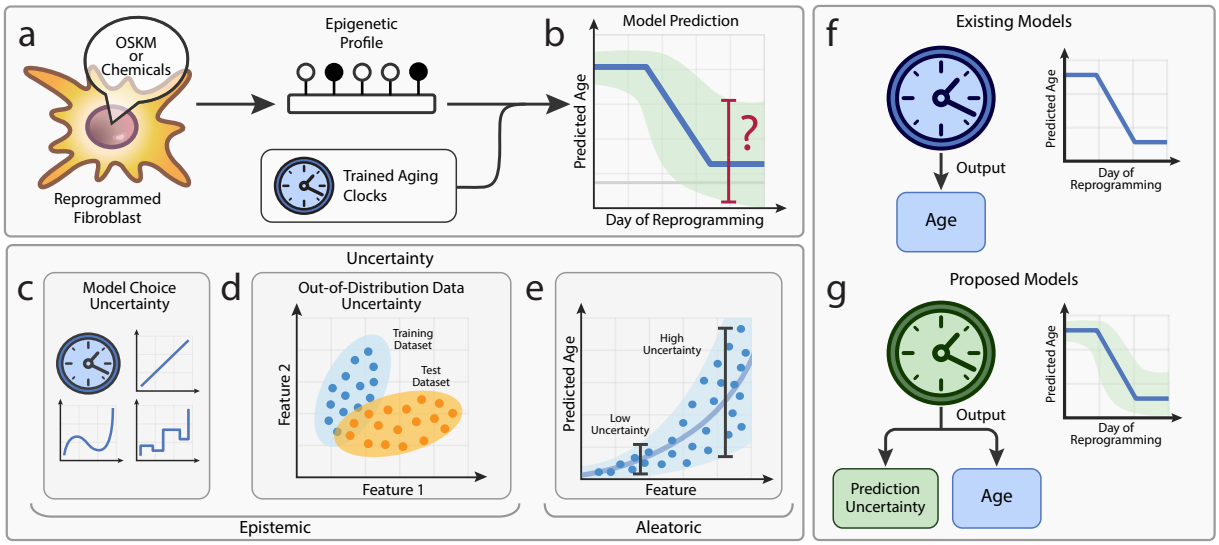

Figure 1: **Prediction uncertainty is an essential component for clinically relevant aging clocks.** **a**, A common pipeline for testing rejuvenation effect using epigenetic aging clocks. **b**, Current aging clocks estimate biological age during the reprogramming process; however, they lack epistemic uncertainty quantification. **c**, By selecting a model to make predictions from data, a researcher implicitly introduces model uncertainty. **d**, Out-of-distribution uncertainty arises when the testing data samples are not represented in the training distribution. **e**, Aleatoric uncertainty comes from the intrinsic variability in data, *e.g.* when the same level of a feature corresponds to different ages. **f**, None of the existing aging clocks estimate epistemic uncertainty. **g**, We propose to use aging clocks capable of predicting uncertainty, which could mitigate the potentially erroneous effects of clock predictions on clinical decision-making.

Moving beyond this simplistic case, we proceed with a deeper exploration of possible covariate shifts in the context of epigenetic aging clocks.

Datasets of reprogramming and embryogenesis exhibit significant covariate shifts relative to aging datasets

DNA methylation in mammals is generally observed in a CpG context (*i.e.*, at the cytosines followed by a guanine nucleotide) [28]. As most popular DNAm profiling methods such as bead microarrays and reduced-representation bisulfite sequencing (RRBS) also deliver information regarding mainly CpG methylation [29], we will further refer to DNAm sites as CpG sites (CpGs).

Covariate shift can arise from various intrinsic and technical factors, including differences in sampling sources and locations, tissue cell content, sample handling techniques, instrumental effects, *etc.* [18]. Given that aging clock models are built to include only a handful of CpGs, it is reasonable to focus on these specific sites for covariate shift detection (see Methods). Clearly, different datasets may harbor different subsets of sites, and aging clocks have been shown to perform at least slightly better when being trained and tested on the same tissue types [30], so, for instance, comparing CpGs from an aging skin dataset with CpGs profiled during fibroblast reprogramming is advised. In addition, the success of multi-tissue epigenetic clock predictions in different conditions was already evaluated elsewhere [31], so we focused exclusively on the single-tissue predictions.

To estimate the extent of covariate shift in DNAm studies, we analyze four representative scenarios in the order of increasing expected difference between the distributions of DNAm patterns: (i) One aging dataset split into two subsets; (ii) Two independent aging datasets; (iii) Aging *vs.* cellular reprogramming *in vitro* and *in vivo*; and (iv) Aging *vs.* early embryogenesis, for which epigenetic rejuvenation was also demonstrated [7]. For every scenario, we performed the principal component analysis (PCA) to demonstrate if there are separate clusters between the datasets, and the Kolmogorov-Smirnov (KS) test to explicitly calculate the percentage of CpGs that have statistically similar distributions in both datasets in question.

First, we examine a DNAm dataset of aging human skin [32] split randomly into the training and testing subsets. As anticipated, we detect no discernible covariate shift: the subsets are indistinguishable by the PCA (Fig. **2**c), DNAm value distributions of at least top-4 age-correlated CpGs from both subsets perfectly overlay each other (Fig. **2**d), and the KS test for the distribution similarity further confirms the lack of covariate shift (Fig. **2**e).

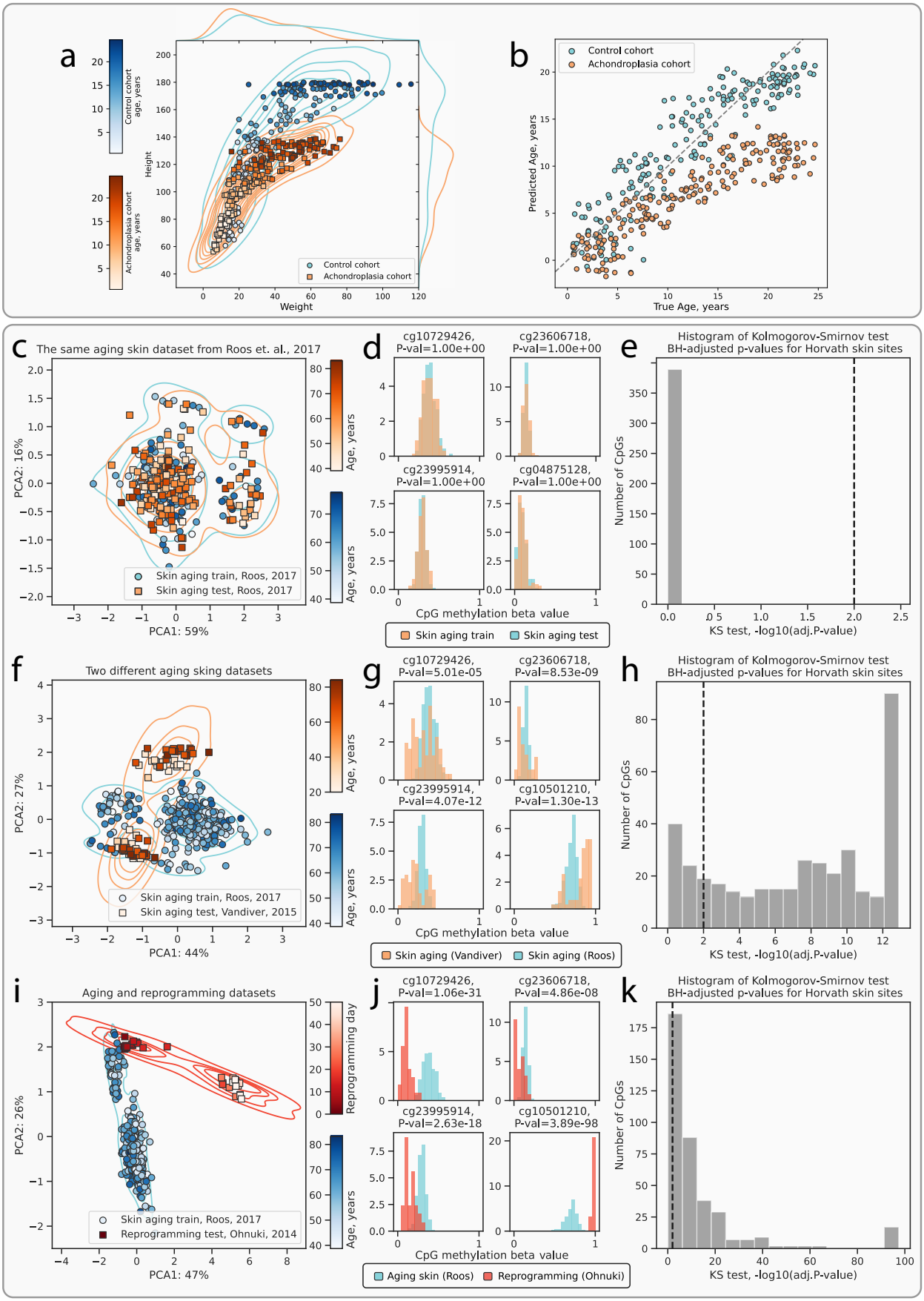
a
Control cohort age, years
Achondroplasia cohort age, years
Height
Weight
Control cohort
Achondroplasia cohort
b
Control cohort
Achondroplasia cohort
Predicted Age, years
True Age, years
c
The same aging skin dataset from Roos et. al., 2017
PCA2: 16%
PCA1: 59%
Age, years
Skin aging train, Roos, 2017
Skin aging test, Roos, 2017
d
cg10729426, P-val=1.00e+00
cg23606718, P-val=1.00e+00
cg23995914, P-val=1.00e+00
cg04875128, P-val=1.00e+00
CpG methylation beta value
Skin aging train
Skin aging test
e
Histogram of Kolmogorov-Smirnov test BH-adjusted p-values for Horvath skin sites
Number of CpGs
KS test, -log10(adj.P-value)
f
Two different aging sking datasets
PCA2: 27%
PCA1: 44%
Age, years
Skin aging train, Roos, 2017
Skin aging test, Vandiver, 2015
g
cg10729426, P-val=5.01e-05
cg23606718, P-val=8.53e-09
cg23995914, P-val=4.07e-12
cg10501210, P-val=1.30e-13
CpG methylation beta value
Skin aging (Vandiver)
Skin aging (Roos)
h
Histogram of Kolmogorov-Smirnov test BH-adjusted p-values for Horvath skin sites
Number of CpGs
KS test, -log10(adj.P-value)
i
Aging and reprogramming datasets
PCA2: 26%
PCA1: 47%
Reprogramming day
Age, years
Skin aging train, Roos, 2017
Reprogramming test, Ohnuki, 2014
j
cg10729426, P-val=1.06e-31
cg23606718, P-val=4.86e-08
cg23995914, P-val=2.63e-18
cg10501210, P-val=3.89e-98
CpG methylation beta value
Aging skin (Roos)
Reprogramming (Ohnuki)
k
Histogram of Kolmogorov-Smirnov test BH-adjusted p-values for Horvath skin sites
Number of CpGs
KS test, -log10(adj.P-value)

Figure 2: **Identification of covariate shift and its impact on aging clock models. a**, An example of weight and height covariate shift between the control and achondroplasia cohorts. **b**, Ages predicted by a model trained on the weight and height measurements of the control cohort. The predictions are significantly biased for the achondroplasia cohort, which is a purely technical phenomenon caused by shifted covariates. **c,f,i**, Principal component analysis (PCA) of DNAm samples shows no covariate shift between the training and testing splits of the same aging skin dataset [32] **(c)**, a moderate covariate shift between the different aging skin datasets [32, 33] **(f)**, and a strong covariate shift between the aging skin dataset [32] and the *in vitro* fibroblast reprogramming dataset [36] **(i)**. **d,g,j**, Histograms of beta values for individual DNAm sites demonstrating no shifts between the two subsets of data **(d)**, moderate shifts between aging skin datasets from different studies **(g)**, and strong shifts between the aging skin and reprogramming datasets **(j)**. **e,h,i**, Histograms of $-log_{10}(adj.P-values)$ demonstrating no DNAm sites rejected by the KS two-sample test at the 0.01 significance level (see Methods) confirming the absence of covariate shift **(e)**, 81% of DNAm sites rejected by the KS test confirming the presence of moderate covariate shift **(h)**, 86% of DNAm sites rejected by the KS test confirming the presence of strong covariate shift **(k)**. Percentages on axes of **c**, **f**, **i** demonstrate the amount of variance explained by the corresponding principal components. Representative sites for histograms **d**, **g**, **j** were chosen from the top-four sites ordered by their correlation with chronological age.

Second, for two independent datasets of aging human skin [32, 33], moderate covariate shift is evident from similar analysis (Fig. **2**f,g), with the KS test indicating substantial differences in individual distributions (Fig. **2**h): 81% of sites are rejected by the test (i.e., have different distributions). On the other hand, a joint analysis of two aging mouse liver datasets [34, 35] displays minimal covariate shift (1% of rejected CpGs,Extended Data Fig. **1**d-f).

Third, a comparison of the aging human skin dataset [32] with two datasets of *in vitro* human fibroblast reprogramming [36, 4] reveals strong covariate shifts: at early stages, fibroblasts closely resemble aging skin samples in their principal component (PC) coordinates (Fig. **2**i; Extended Data Fig. **1**a), but, as the reprogramming progresses through the maturation phase, a notable departure from the skin samples can be observed (86% and 69% of rejected CpGs, respectively; Fig. **2**j,k; Extended Data Fig. **1**b,c). Such behavior of samples during the *in vitro* reprogramming might suggest that reprogrammed cells acquire some phenotype unobservable *in vivo*.

To shed more light on this hypothesis, we further compare a dataset of *in vivo* reprogramming in mouse liver [3] with merged aging mouse liver samples from [34, 35], as these two studies demonstrated no significant difference in the previous analysis. As a result, we detect a moderate covariate shift according to PCA and the KS test (32% of rejected CpGs, Extended Data Fig. **1**g-i), which might imply that the *in vivo* conditions are better at preserving the normal phenotypic characteristics.

Fourth, to address the "ground zero" hypothesis of epigenetic rejuvenation during embryogenesis [37], we compare mouse embryos [38] with blood aging samples [34], as both of them were used previously to demonstrate this phenomenon [7]. The first stages of embryogenesis strongly diverge from the aging samples, while later stages align closer to the aging cluster on PCA, with the KS test detecting moderate covariate shift (15% of rejected CpGs, Extended Data Fig. **1**j-l).

These results collectively suggest that the DNAm covariates can be significantly shifted relative to each other dataset, thereby implicitly increasing the risk of failed clock predictions. Given these risks, we advocate for the routine checks of covariate shifts between datasets using the aforementioned methods or other techniques reviewed elsewhere [17] before applying aging clocks.

Aging clocks are inconsistent in their predictions for reprogramming-induced rejuvenation

When choosing a specific machine learning model to construct an aging clock, a researcher inevitably introduces model uncertainty (Fig. **3**a). The chosen model family (*e.g.*, Linear Regression) is an assumption about how the true process of epigenetic aging, which is not accessible *a priori*, works. Therefore, by training different aging clocks on different data or using different models, one can expect to obtain poorly consistent predictions [39].

To demonstrate how model uncertainty manifests itself, we leverage nine aging clocks trained on different CpG sets and tissue types [26, 27, 40, 41, 42, 43, 44, 45] and apply them to two *in vitro* reprogramming datasets (see Methods). As expected, all clocks vary greatly in their dynamics and amplitudes of predicted ages across the reprogramming timeline (Fig. **3**b, Extended Data Fig. **2**a). We further focused on the period of the first three weeks of reprogramming (from initiation to maturation), as the end of the third week (approximately day 20) marks the loss of somatic identity and increasing risk of teratoma formation [1]. A comparison of absolute differences of ages estimated by the models at the beginning and at the last available time point before the end of this period (day 15 for Ohnuki et al. and day 17 for Gill et al.) exhibits evident inconsistencies for both datasets, ranging from the Horvath clock [26] predicting age reversal by 40 years to the Hannum clock [27] predicting age increase by 13 years (Fig. **3**c, Extended Data Fig. **2**a,b).

To test the hypothesis that these discrepancies might have arisen from the differences in training datasets rather than from the clock models themselves (most of which are based on ElasticNet regression), we trained different ML models on the same dataset of aging human skin [32] and discovered that their predictions of rejuvenation demonstrate a considerable instability (Fig. **3**d, Extended Data Fig. **2**c). Despite these inconsistencies, most models indicated positive rejuvenation. Therefore, the clocks could potentially serve as qualitative (or binary) predictors of rejuvenation for the *in vivo* studies.

To explore this possibility, we developed our aging clocks anew by fitting a Lasso regression over the combined dataset of aging mouse liver [34, 35] (see Methods) and applied these clocks to the *in vivo* reprogramming dataset [3]. We had to develop clocks *de novo*, since this dataset contained almost no CpGs from the existing clock models mentioned before. Although these new clocks performed quite robustly for the testing subset (Fig. **3**e), they failed to register any rejuvenation between the old control mice and old mice treated with the OSKM factors (two-sided Mann-Whitney-Wilcoxon (MWW) test, P-value=0.11, Fig. **3**f). Notably, all models consistently predicted higher ages for all control liver samples, suggesting a response shift (*i.e.*, $P_{train}(Y) \neq P_{test}(Y)$) between the training and testing datasets [17], which might result from the differences in DNAm patterns between the training and testing mouse strains (inbred C57BL/6 and transgenic i4F-B, respectively).

Taken together, our findings highlight a considerable instability of predictions with regard both to the different datasets utilized for training and to the choice of a model family employed for prediction. This lack of agreement between aging clocks could supposedly result from the covariate and response shift discovered previously in this research, as well as from the manifestation of model uncertainty, leading to divergent rejuvenation dynamics.

Reprogramming data cannot be used to predict normal aging

We further examine the accuracy of aging clocks applied to predict reprogramming by employing a mathematically rigorous approach. When the regression models are trained, they need to satisfy a certain degree of correlation, expressed, for example, as the R^2 score or mean absolute error (MAE). Supposedly, in an ideal case, when a clock predicts age with absolute accuracy ($R^2 = 1.0$, $MAE = 0$), its predicted ages can be used interchangeably with the true chronological ages (because they are equal) to train another model on the testing data, and to predict ages in the original training dataset with the same absolute accuracy. In reality, this ideal case is never observed due to the technical and biological variations in the training and testing samples and sampling techniques, and due to model under- or overfitting, but we hypothesize that if these effects are small (*i.e.*, if there is little epistemic uncertainty), then the reciprocal prediction of training data by the ages predicted for testing data should still be possible, albeit with some degree of error.

To evaluate this mutual interchangeability (*i.e.*, commutativity) of datasets from the perspective of model training, we developed an Inverse Train-Test Procedure (ITTP, see Methods for details, Fig. **4**a,b). Considering the availability of ground truth measurements, we divided the ITTP use cases into two categories. In case 1, true ages are available for the testing dataset X_{te}, Y_{te} (Fig. **4**a), which corresponds to comparing two datasets of aging. First, we train model 1 (*e.g.*, linear regression) on the training set X_{tr}, Y_{tr} to predict the ages of test samples $\hat{Y}_{te}$, where the hat symbol ˆ denotes predicted values. Second, model 2 is trained using the testing features X_{te} and the ages predicted by the model 1 $\hat{Y}_{te}$. If the datasets are indeed interchangeable, similarly good performance metrics are expected for the predictions made by model 2 for the original training samples (Fig. **4**a, Case 1 panel).

In case 2, we do not know the true ages for the testing dataset, which corresponds to leveraging an aging DNAm dataset $\{X_{tr}, Y_{tr}\}$ and a reprogramming dataset $\{X_{rep}\}$ (Fig. **4**b), since we do not expect the biological age of reprogrammed cells to stay approximately the same as their chronological age, as would be the case for a dataset of normal aging. As before, we train model 1 on the training data and predict the ages of reprogramming samples $\hat{Y}_{rep}$. In this scenario, we cannot validate our predictions due to the lack of ground truth values of biological age for reprogrammed cells, so we can only assume that the predictions are correct and use them to train model 2. If, as a result, we observe good performance metrics between model 2 predictions and the ages of the original training dataset, then we can still assume interchangeability, regardless of the intermediary predictions of model 1. On the other hand, if we obtain poor prediction accuracy for model 2, then we infer that these datasets are not interchangeable, and that such prediction failure was supposedly caused by a large epistemic uncertainty, presumably caused by a substantial dataset shift. In summary, the ITTP approach contributes to our multifaceted estimation of uncertainty of predictions in aging and reprogramming.

We further apply this procedure for the pairs of datasets described previously (Fig. **2**; Extended Data Fig. **1**) and use the Lasso regression model family as models 1 and 2 (see Methods). Models performe well for the aging human skin dataset [32] (Case 1, Fig. **2**c,d), both when applying model 1 to predict testing data ages ($r = 0.934$, $MAE = 2.8$ years) and when applying model 2 to predict ages of the original training data ($r = 0.831$, $MAE = 4.4$ years) using $\hat{Y}_{te}$ instead of Y_{te} to fit model 2. Therefore, we infer that the training and testing datasets are interchangeable and

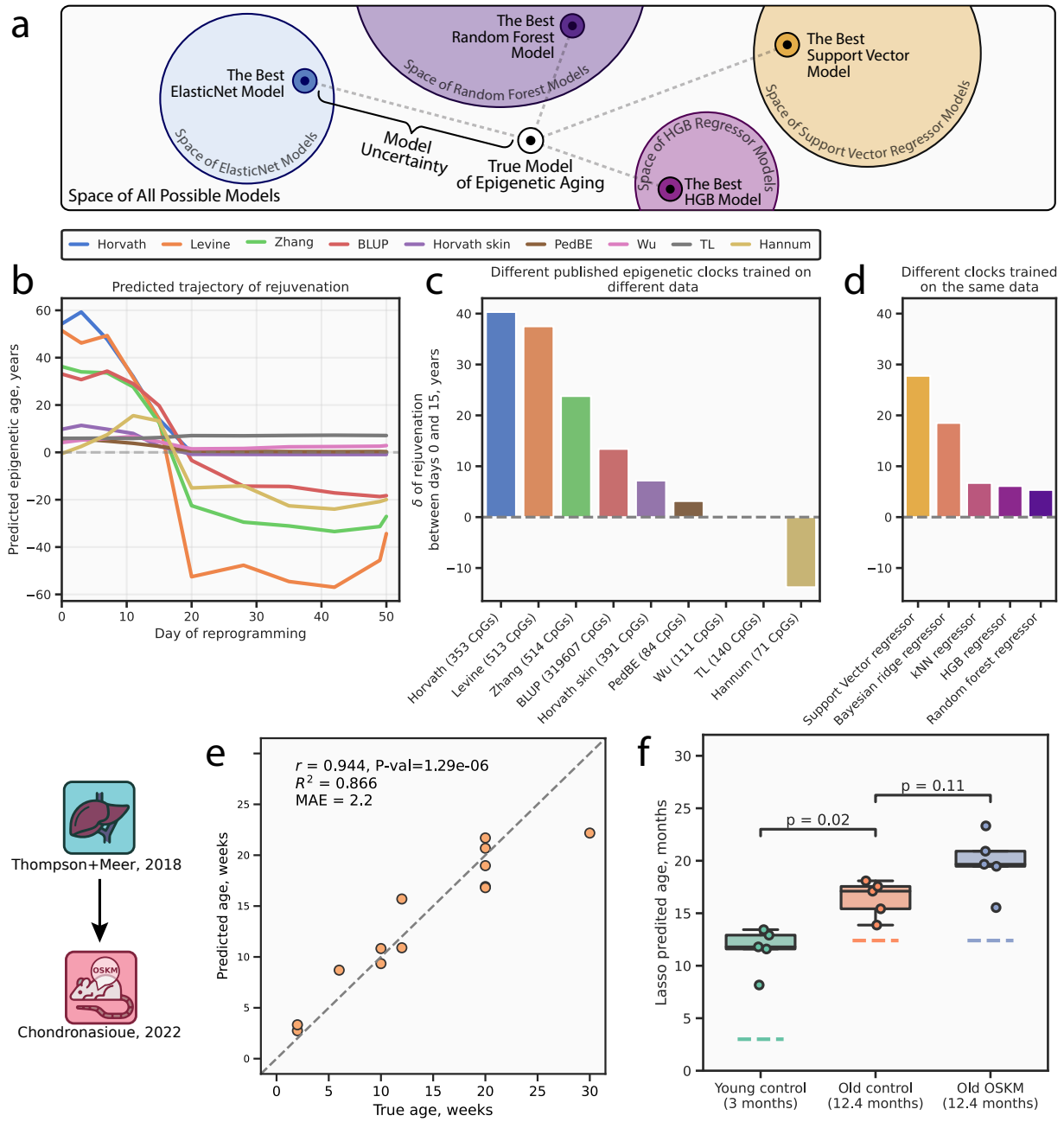

Figure 3: **Inconsistency between the estimations of reprogramming-induced rejuvenation provided by the aging clocks.** **a**, A schematic representation of model uncertainty stemming from the choice of a particular model. **b**, Published aging clocks (most of which are ElasticNet-based) trained on different DNAm datasets predict different trajectories of rejuvenation, as cellular reprogramming progresses. Dashed line represents zero epigenetic age. **c**, Aging clocks showing differences in the rejuvenation effect accumulated between reprogramming days 0 and 15, bar colors match line colors from **a** and represent different published clocks. The dashed line represents lack of changes in epigenetic age. **d**, Inconsistency of predictions holds for the clocks built using different ML model types and trained on the same dataset. Bar colors represent different models. The dashed line represents lack of changes in epigenetic age. **e**, Performance scatter plot of the *de novo* Lasso clock model on the testing subset (see Methods). Pearson's correlation coefficient (r), the associated P-value (*Pval*), R^2 score (R^2), and mean absolute error (*MAE*) are displayed. The dashed line corresponds to the points of equality between predicted and chronological ages. Dots represent individual samples. **f**, Epigenetic ages predicted by the *de novo* Lasso clock for the *in vivo* liver reprogramming dataset [3] containing samples from the young (n=5) and old (n=5) control mice, and old reprogrammed mice (n=5). Samples from transiently reprogrammed mice exhibit insignificantly increased (P-value = 0.11, two-sided MW test) epigenetic age in comparison to the control group. Dashed lines represent chronological ages for the respective cohorts. Blue icons indicate the training dataset and orange icons indicate the testing dataset. A detailed description of the datasets is presented in Supplementary Tables 1 and 2.

can be used reciprocally to predict each other, which is expected for a dataset of aging split randomly into two parts. Similarly, we obtain good reciprocal predictions for the mouse blood samples subsets [34] (Extended Data Fig. **3**e)

In accordance to the already demonstrated evidence that the datasets of aging mouse liver [34, 35] exhibit no significant covariate shift (Extended Data Fig. **1**d), we observe that they both pass the ITTP relatively well (Extended Data Fig. **3**b,c) for the initial model training. A slightly lower accuracy of predicting Meer ages with clocks trained on Thompson data ($r = 0.737$, $MAE = 6.14$ months) might be explained by the presence of a number of outliers in the Meer dataset (Extended Data Fig. **1**d). At the second step, however, we obtained high performance metrics ($r = 0.975$, $MAE = 1.49$ months), which is another evidence of absence of a significant covariate shift between these datasets.

For the case 2 applications of ITTP, the results were more diverse. Both the Ohnuki et al. (**4**e) and the Gill et al. (Extended Data Fig. **3**a) datasets of *in vitro* fibroblast reprogramming yielded bad predictions for the aging human skin data ($r = 0.235$, $MAE = 52.7$ years and $r = 0.239$, $MAE = 34.3$ years for the second-step models trained on the Ohnuki et al. and Gill et al. datasets, respectively), from which we concluded that the reprogramming datasets cannot be used to predict normal aging skin data, provided that we consider age predictions for the reprogramming data to be accurate.

The dataset of *in vivo* reprogramming in mouse liver [3], on the other hand, passed the second step of ITTP with remarkable success, demonstrating performance metrics of $r = 0.968$ and $MAE = 2.21$ months (Extended Data Fig. **??**d). It is worth mentioning that the predictions at step 1 were shifted upward, while the aging trend was still captured well ($r = 0.826$). Finally, by applying ITTP to the datasets of aging mouse blood [34] and mouse embryogenesis [38], we obtained prediction failure similar to that of *in vitro* reprogramming (Extended Data Fig. **3**f).

To summarize, the ITTP method highlights the concerns of applying aging clocks to the reprogramming data (or to any other OOD scenario). As an empirical test to discover dataset shift, it helps assess the risk of prediction failure. The results presented above unambiguously showed that normal aging cannot be predicted using the reprogramming data, which immediately prompts to inquire, whether, in return, the ages in reprogramming can be correctly predicted using data on normal aging.

Uncertainty-aware clocks reveal insignificance of age reversal

In clinical settings, where decision-making often relies on the level of uncertainty, an ML model is required to estimate not only the desired outcome, but also the uncertainty of its predictions [15]. A robust model should ideally warn of extreme uncertainty when making predictions on shifted datasets. Most aging clock papers, following Horvath [26] and Hannum et al. [27], adopted the ElasticNet model that lacks inherent uncertainty estimation. Consequently, to address this gap, we trained a Gaussian Process Regressor (GPR) model [46, 47], a variant of which was recently employed in aging clocks as well [48].

To exemplify the principle of GPR, we trained and tested it on a single CpG site (Fig. **5**a,b). A Gaussian process (GP) can be viewed as a probability distribution over all possible functions that can be fitted over the training observations [46]. Therefore, for every input methylation value, a fitted GPR model provides an estimation of the most probable age, as well as the probability distribution around this estimation with a finite variance that represents a credible interval for the prediction. Credible interval is grounded in Bayesian statistics and is calculated for every individual prediction relying on the training data and prior model assumptions, so it should not be confused with a confidence interval, which describes only the distribution of multiple predictions.

When the GPR fits the training data well, it computes a finite error produced by the variations of age points related to the same methylation value (*i.e.*, aleatoric uncertainty). The further methylation values depart outside the training distribution, the higher epistemic uncertainty is assigned by the model, and, hence, the wider a credible interval becomes, reflecting that the model is unfamiliar with this kind of data (Fig. **5**b).

GPR thus provides an estimation of total prediction uncertainty, comprising both its aleatoric and epistemic components, and presenting it as a credible interval of several standard deviations for every individual prediction. It is important to acknowledge that GPR predictions are significantly influenced by the selected prior distributions over functions, so the results may slightly vary depending on the choice of these assumptions.

In accordance with the previous sections, we employed GPR models trained on the aging datasets to predict rejuvenation trajectories in the respective reprogramming scenarios, and to determine the corresponding credible intervals (see Methods). For the Ohnuki et al. dataset of *in vitro* reprogramming, a skin-trained GPR (Extended Data Fig. **4**a) predicted a notable age decline from day 11 through day 28 (Fig. **5**c), which aligned well with our previous observations obtained with ElasticNet (Fig. **3**a). However, the accompanying credible interval of two standard deviations revealed how extremely uncertain the model was about these predictions, especially in the case of the late reprogramming phase.

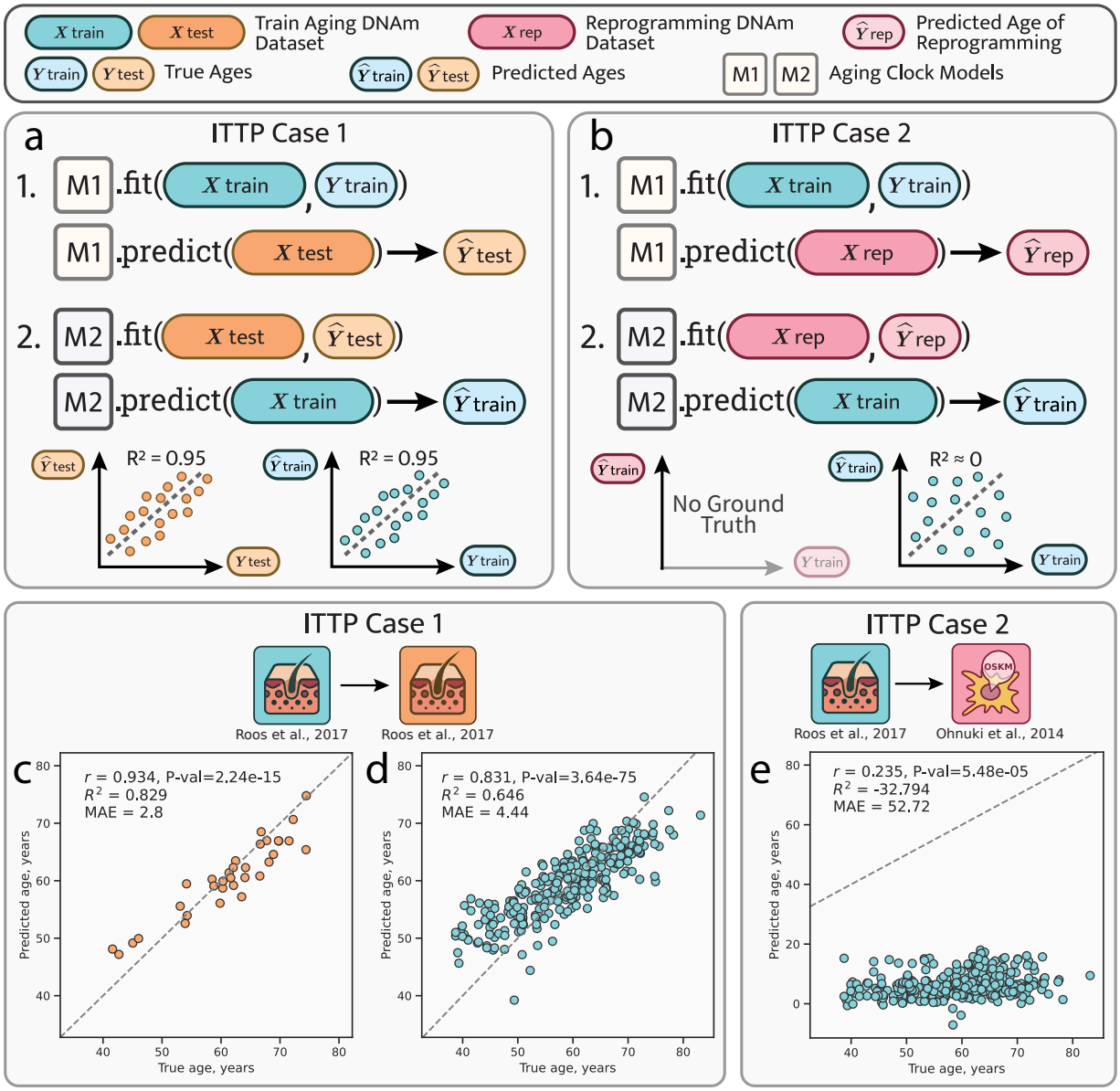

Figure 4: **Inverse Train-Test Procedure (ITTP) demonstrates the impossibility of predicting aging data with reprogramming data.** **a**, ITTP case 1 demonstrates the procedure with both datasets having true values of age (falsifiable case). Two datasets are interchangeable if similar prediction performance metrics are obtained on both steps of the procedure. **b**, ITTP case 2 demonstrates the procedure with only one dataset having true values of age and another doesn't (unfalsifiable case). In this case, only the second-step performance metric is computable. A hypothetical reprogramming dataset fails to pass the ITTP if the second-step performance metrics are naught. **c,d**, Application of ITTP to human aging skin dataset [32] split into training and testing subsets as 90% to 10% correspondingly (see Methods). The performance metrics computed both at step 1 **c**, and step 2 **d** are high, thus, the pair of datasets are interchangeable. **e**, Application of ITTP to human aging skin dataset [32] for training and human reprogramming fibroblasts dataset [36] for test. Poor performance metrics at the second step qualify datasets as interchangeable — the reprogramming dataset can not be used for the prediction of aging data. Blue icons indicate the training dataset and orange icons indicate the testing dataset. A red icon indicates the reprogramming dataset. A detailed description of the datasets is presented in Supplementary Tables 1 and 2.

This uncertainty cast doubt on the significance of any rejuvenation effect until the 20-th day of reprogramming, where complete erasure of somatic identity was observed in the respective paper [36].

Assessing the Gill et al. [4] *in vitro* dataset with GPR yielded similar results, with two notable differences (Fig. **5**d). Firstly, the credible interval at reprogramming day 0 was slightly narrower, suggesting that these samples were more similar to the training set. Secondly, the model indicated a significant rejuvenation effect between days 0 and 17 (with the rejuvenation coefficient P-value of 0.014), hinting that rejuvenation might indeed occur towards the end of the maturation phase. However, the sizeable credible interval at day 17 (spanning from 0 to approximately 70 years) precludes making confident statements.

We revisited the *in vivo* reprogramming dataset using a mouse liver-trained GPR (Extended Data Fig. **4**b) and observed a significant negative rejuvenation effect between the control and OSKM-treated old mice, with a P-value of 0.018 (Fig. **5**e). This outcome underscores the importance of incorporating individual prediction uncertainties to resolve differences between groups that are otherwise undistinguishable by simpler methods such as Lasso regression. On the other hand, as in the case of Lasso-based predictions, we found that the GPR had a systematic bias in its estimations. This bias did not seem to affect credible intervals as strongly though, which might suggest that GPR underestimates credible intervals in the event of dataset shift, resulting, in our case, from the comparison of two different mouse strains.

As predicted by the mouse blood-trained GPR (Extended Data Fig. **4**c), the dynamics of epigenetic age during mouse embryogenesis [38] displayed a local minimum on embryonic day 8.5 (Fig. **5**f), in alignment with the findings reported by Kerepesi et al. [7]. However, the large credible intervals accompanying these predictions prevent us from designating day 8.5 as the "ground zero" of epigenetic age, as indicated by the P-value of 0.54 for the age decline between days 3.5 and 8.5, and the P-value of 0.37 for the subsequent age increase between days 8.5 and 10.5. It is worth noting, that the credible intervals narrowed in the course of embryonic development, supporting our earlier observation of a larger shift between the early days of embryogenesis and aging (Extended Data Fig. **1**j).

Our findings indicate that a GPR model capable of assessing its prediction uncertainty can effectively detect covariate shift in a test dataset, assigning elevated uncertainty to samples not represented in the training data. Given these insights, we propose that the future aging clock models should incorporate the capability to quantify their prediction uncertainty. Such feature would dramatically enhance the reliability of these models, allowing for improved control and mitigation of potential prediction failures.

Discussion

Epigenetic clocks have been widely used to demonstrate age acceleration and deceleration in a variety of research contexts [12]. However, there is still a reluctance to include clock measurements as endpoints in the clinical longevity intervention trials [49]. In addition to the fact that the existing clocks fail to capture some aging-associated conditions [20], we also point out that they cannot be easily validated and relied upon, as they lack the ability to estimate the degree of uncertainty in their predictions.

In this study, we present a computational framework for validating the rejuvenation effects predicted by epigenetic aging clocks. Epigenetic age reversal is unfalsifiable, because we cannot access the ground truth values of biological age. Hence, we have to rely on the indirect evidence. For that, we included four approaches in our framework: covariate shift estimation, comparison of different clock models, the Inverse Train-Test Procedure (ITTP), and the prediction uncertainty estimation using a Gaussian Process Regression (GPR) model.

In a simplified case of covariate shift, clocks trained on the weight and height of normal individuals predicted lower ages for the achondroplasia individuals, in contradiction to the demographic data on their lifespans. Thus, we demonstrate how the presence of covariate shift can distort model performance. Next, we show that the DNA methylation (DNAm) data of reprogramming-induced rejuvenation (RIR) exhibits strong covariate shift, which might also lead to a systematic error in its age prediction.

By applying nine different published aging clocks [26, 27, 40, 41, 42, 43, 44, 45] to the *in vitro* reprogramming data, we illustrate that the magnitude of rejuvenation effect achieved by the end of the maturation phase of reprogramming highly depends on the clock model type and data we use for training and prediction. For different models, the age reversal effect can span two orders of magnitude, including null and even negative rejuvenation.

We developed the ITTP approach to assess the interchangeability of the training and testing datasets. This procedure revealed that the *in vitro* reprogramming datasets cannot predict normal aging, which challenged the premise that normal aging can accurately predict reprogramming.

A GPR-based aging clock can estimate uncertainty of its own age predictions in the form of standard deviations (that are used for credible interval calculation), even for the samples out of its training distribution. This clock demonstrates

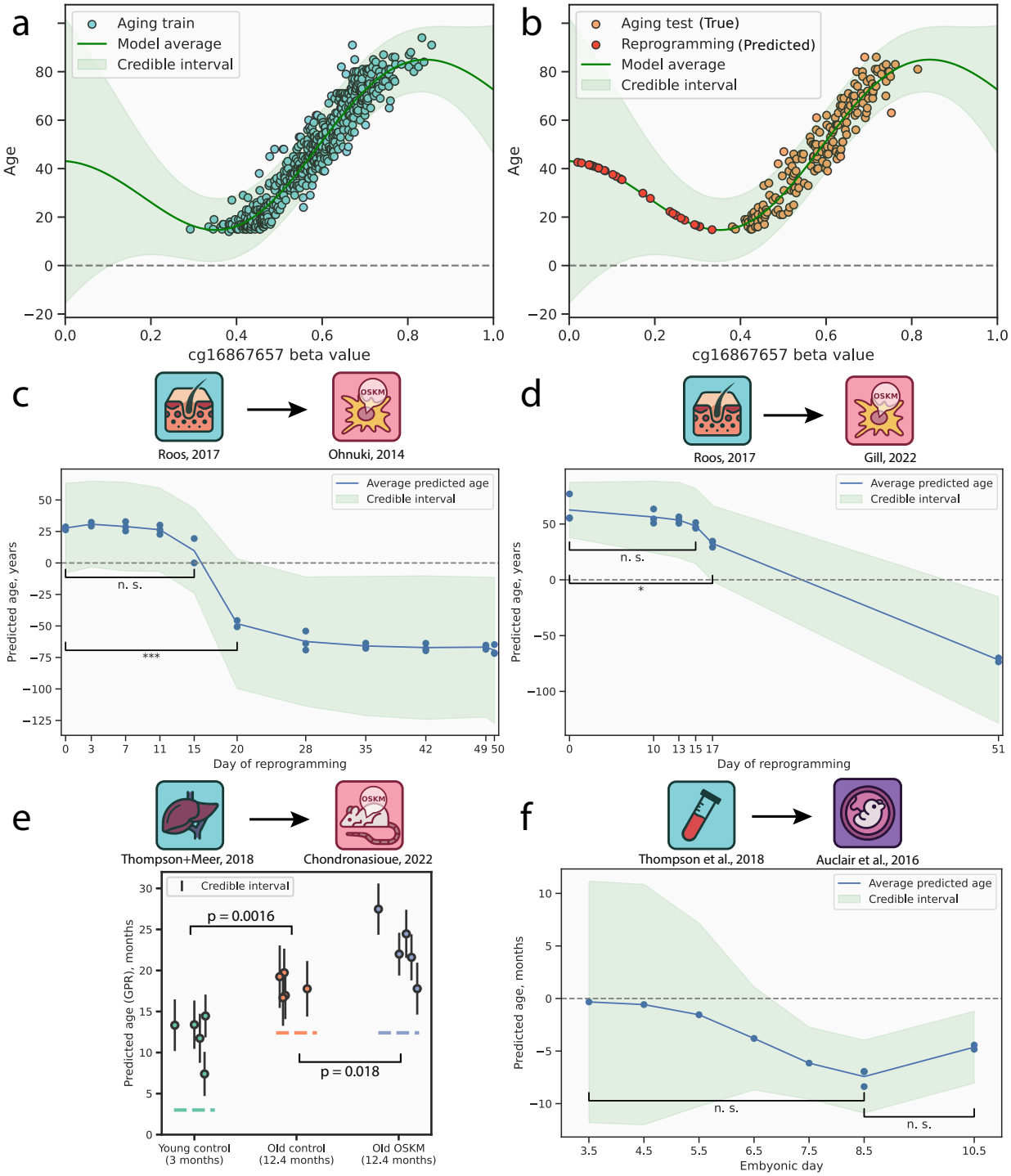
a
Aging train
Model average
Credible interval
Age
cg16867657 beta value
b
Aging test (True)
Reprogramming (Predicted)
Model average
Credible interval
Age
cg16867657 beta value
c
Roos, 2017
Ohnuki, 2014
OSKM
Average predicted age
Credible interval
Predicted age, years
n. s.
***
Day of reprogramming
d
Roos, 2017
Gill, 2022
OSKM
Average predicted age
Credible interval
Predicted age, years
n. s.
*
Day of reprogramming
e
Thompson+Meer, 2018
Chondronasioue, 2022
OSKM
Credible interval
p = 0.0016
p = 0.018
Predicted age (GPR), months
Young control (3 months)
Old control (12.4 months)
Old OSKM (12.4 months)
f
Thompson et al., 2018
Auclair et al., 2016
Average predicted age
Credible interval
Predicted age, months
n. s.
n. s.
Embyonic day

Figure 5: **Estimation of epistemic uncertainty for rejuvenation datasets with Gaussian Process regression (GPR) model. a**, A simplified case of a GPR model trained to predict chronological age from a single methylation site (see Methods). **b**, Demonstration of increasing of prediction uncertainty (expressed as two standard deviations credible interval) as points move away from the training distribution. **c**, Predicted rejuvenation trajectory and prediction uncertainty (expressed as two standard deviations credible interval) for human fibroblasts *in vitro* reprogramming dataset [36]. No significant rejuvenation event is detected before day 20 of reprogramming. **d**, Predicted rejuvenation trajectory and prediction uncertainty (expressed as two standard deviations credible interval) for human fibroblasts *in vitro* reprogramming dataset [**?**]. No significant rejuvenation event is detected before day 17 of reprogramming. **e**, Predicted epigenetic age and individual prediction uncertainties (expressed as two standard deviations credible interval) for murine *in vivo* reprogramming dataset [3]. **f**, Predicted epigenetic age trajectory and prediction uncertainty (expressed as two standard deviations credible interval) for murine embryogenesis dataset [38].
no statistically significant difference between days 0 and 15 of *in vitro* reprogramming, and some significance between days 0 and 17 in another dataset. Nevertheless, the accompanying credible interval at day 17 spans around 70 years, which prevents us from drawing definitive conclusions of observed rejuvenation.

The *in vivo* liver reprogramming data exhibit a covariate distribution closer to that of normal aging. Moreover, it manages to pass the ITTP test, suggesting that the *in vivo* reprogramming might preserve the organismal states better than the *in vitro* procedure. However, we show that both the Lasso and the GPR clock models surprisingly predict old reprogrammed samples to be either of the same age or even significantly older than the old controls. This finding prompts further inquiries into the nature of processes accompanying the *in vivo* reprogramming, especially in light of a recent study describing impaired liver function and premature death of mice with continued OSKM induction [50].

To exemplify our approach to the "ground zero" of epigenetic age in embryonic development [37], we leveraged a dataset spanning days 3.5 through 10.5 post-conception [38]. The mid-embryogenesis stages cluster well with the aging data on PCA, and their GPR-predicted credible intervals are narrower than those of the earlier stages. However, it appears insufficient, because the dataset as a whole shows significant covariate shift, fails at the ITTP, and features too wide credible intervals to demonstrate any significance between the stages of highest (day 3.5) and lowest (day 8.5) predicted age.

We hypothesize that the GPR model assigns such large credible intervals both to the *in vitro* reprogramming and embryogenesis, because the totipotent and pluripotent states are too unfamiliar to a model trained purely on differentiated somatic cells. Thus, we have shown that an aging clock model performing well within aging datasets will likely fail to reliably predict rejuvenation events not represented in the training dataset. Including progenitor cells in the training samples could be beneficial to decrease this uncertainty.

We did not aim to comprehensively cover all available datasets of putative rejuvenation, be it reprogramming, embryogenesis, or other interventions. Importantly, our work should also not be viewed as an attempt to prove or disprove whether rejuvenation actually occurs. Limiting ourselves to the most vivid examples, we illustrate that the existing aging clocks cannot serve as reliable biomarkers of any rejuvenation. Moreover, aging clocks that are trained to predict chronological age, so-called first generation clocks [41], bear other drawbacks as well. They often rely on overly optimistic assumptions [11] and may be subject to the biomarker paradox, formulated as: "A hypothetical biomarker that approaches perfect correlation with chronological age could be replaced by chronological age and would be insensitive to differences in aging among individuals." [51]. Notably, the most accurate epigenetic aging clocks, while precise in age prediction, fail to predict mortality or onset of age-related diseases [45]. Although semi-supervised [51] and unsupervised [52] models are not subject to this paradox, they still face the challenge of outperforming chronological age in mortality prediction, and it is not obvious which model assumptions should be used to satisfy this criterion. The more promising direction is the second-generation aging clocks [41, 53] that are trained to predict mortality rather than chronological age. However, they require accumulating both the biomarker data and individual death times post-collection, posing significant practical and ethical challenges, and they are also hardly applicable to the *in vitro* experiments and embryonic development.

Nevertheless, the aforementioned issues do not refute the necessity of developing a reliable surrogate health measure [54] that is still crucial for evaluating longevity drugs and other interventions in clinical trials of geroprotectors [55]. There are already a handful of criteria for the potential biomarkers of aging: the association with mortality, the responsiveness to longevity interventions, and the minimally invasive procedure to obtain data [56]. We propose that in order to become approved by a wider longevity community, single-point age clock predictions should at least be supplemented with the uncertainty estimation aimed at identifying the out-of-distribution samples [15], acknowledging that this estimation could itself be flawed. Additionally, we recommend assessing potential covariate shift between the datasets before applying any aging clock models, relying on the methods discussed in this work.

We believe that a clinically relevant and reliable aging clock should have a well-defined training target, such as mortality, and the capability to estimate prediction uncertainties, alerting researchers to possible misinterpretations of their trial results. These new criteria may complicate the development of aging clocks, but they should also advance the field and bring clocks closer to becoming the true health estimator.

Methods

Data and code availability

All datasets of DNA methylation (DNAm) in this study were obtained from the Gene Expression Omnibus (GEO) under the corresponding accession numbers (GSE IDs). Dataset details are available in Supplementary Table 1. We processed these datasets and made them accessible via our repository at `https://github.com/ComputationalAgingLab/reprogramming_ood`, which also includes the code for replicating our original analysis and the detailed installation instructions.

Height and weight datasets

We sourced data for our toy example of dataset shift from the WHO [21] for the control cohort and from Hoover et al. [23] for the achondroplasia cohort. The datasets included height and weight means and standard deviations across various ages: 60–228 months for the control and 0–240+ months for the achondroplasia cohort. We interpolated the control dataset using the methodology by Andres et al. [57] to align age ranges (0–240+ months). Assuming a normal joint distribution for height and weight [22], we sampled points with corresponding means and covariance matrices. Age was uniformly sampled from 0 to 276 months. 1000 samples were generated for each cohort.

Using 1000 samples from the control cohort, we constructed a bivariate linear regression model with the Python *scikit-learn* package [58]. Post-training, this model was utilized for age prediction in the achondroplasia cohort.

Principles of CpG selection

The selection of CpG sites for aging clocks is a nuanced challenge, with various authors suggesting distinct, minimally overlapping subsets [39]. To tackle this, we utilized CpG sites from established clocks most relevant to each dataset pair. For instance, in analyzing covariate shift between the aging skin and fibroblast reprogramming datasets, Horvath's skin clocks [40] were employed. While any accurate age-predicting CpG subset could suffice, we predominantly used known subsets for methodological simplicity where possible. The detailed description of dataset pairs, clock models, and the amount of clock sites observed in the datasets is specified in Supplementary Table 2.

Principal component analysis for covariate shift visualization.

Principal component analysis (PCA), depicted in Fig. **2**c,f,i and Extended Data Fig. **1**a,d,g,j, was conducted on merged datasets to select CpG sites for the covariate shift analysis (refer to previous sections). This analysis used the Python *scikit-learn* library [58].

Kolmogorov-Smirnov test for a shift of individual covariates.

We employed a two-tailed Kolmogorov-Smirnov (KS) test on selected CpG sites (refer to site selection principles above) to detect covariate shifts between dataset pairs. This test assessed the null hypothesis that beta value distributions for a specific site are identical across pairs of datasets. We applied the Benjamini-Hochberg correction to the computed P values, considering a corrected P value below 0.01 as indicative of a significant distributional shift. The KS statistics and P values were calculated using *scipy* [59], and the multiple testing correction was performed with *statsmodels* [60] in Python.

Testing *in vitro* reprogramming datasets with different published epigenetic aging clocks

We evaluated the consistency of predictions across nine aging clock models using the *methylclock* package in R, which predicts epigenetic age from input matrices of methylation site beta values. The predictions generated by *methylclock* are available in our code repository.

Testing *in vitro* reprogramming datasets with different models of aging clocks trained on the same dataset

To evaluate prediction consistency across different aging clock model families, we utilized five machine learning models from the *scikit-learn* package: k-neighbors regressor, random forest regressor, support vector regressor, Bayesian ridge regressor, and histogram-based gradient boosting regressor. These models were trained on the aging skin blood dataset [32] using 5-fold cross-validation and hyperparameter optimization via grid-search, assessing the performance with mean squared error metric. Performance metrics for both training and test subsets, including those for optimally tuned models, are presented in Table 1.

Model	r **Train**	r **Test**	R^2 **Train**	R^2 **Test**	MAE **Train**	MAE **Test**
KNeighborsRegressor	0.844	0.654	0.699	0.402	3.915	6.037
RandomForestRegressor	0.980	0.847	0.934	0.650	1.793	4.585
SVR	0.963	0.933	0.921	0.854	1.603	2.923
BayesianRidge	0.986	0.932	0.971	0.863	1.235	2.750
HistGradientBoostingRegressor	0.999	0.883	0.998	0.752	0.237	3.729

Table 1: Comparison of Model Performance

Lasso fit for *in vivo* reprogramming testing

We observed that the *in vivo* reprogramming dataset [3] exhibited limited overlap with three established mouse aging clocks (Thompson [34] - 7/582 sites, Meer [35] - 3/90 sites, Petkovich [61] - 16/436 sites), potentially impairing clock predictions. Consequently, we developed a new clock using a Lasso penalized regression model, trained on combined liver samples from Thompson [34] and Meer [35]. Utilizing the *LassoCV* class from *scikit-learn* [58], we identified the optimal regularization hyperparameter α through 5-fold cross-validation. The final model, selecting 22 of 16849 CpG sites, exhibited strong test performance ($MAE = 2.2$ months, $R^2 = 0.866$), as detailed in Fig. **3**E. This model was then applied to predict epigenetic age in the *in vivo* reprogramming dataset [3].

Inverse Train-Test procedure with Lasso regression model

ITTP procedure can applied in principle to any pair of datasets to test their interchangeability. However, in practice, the outcome of the procedure will depend on the generalizing abilities of the chosen model. Thus, it is crucial to use models from the same family (e.g. linear model) at step 1 and step 2 of the ITTP. We decided to choose a linear regression model with Lasso penalization as a base model (i.e. model 1 and model 2 according to the scheme in Fig. **4**A) for ITTP because it has generalization properties equivalent to ElasticNet (most often used for aging clocks), but is simpler in terms of the training process (optimizing one hyperparameter instead of two). Below we provide a detailed algorithm for the ITTP procedure for both cases discussed in this study presuming that the Lasso model is used as the base model.

ITTP case 1

ITTP case 1 (also referred to as the falsifiable case) considers a pair of datasets having ground truth values: train dataset $\{X_{tr}, Y_{tr}\}$ and test dataset $\{X_{te}, Y_{te}\}$. Let two different initializations of the Lasso regression model be m_1^0 and m_2^0, where superscript 0 denotes the model state before training and superscript $*$ will denote the model after training. Then, ITTP case 1 can performed by the following algorithm:

Step 1

1. Train model m_1^0 on $\{X_{tr}, Y_{tr}\}$. Select the optimal regularization parameter, α, for Lasso regression employing cross-validation to evaluate the model performance across a range of alpha values $[\alpha_{min}, \alpha_{max}]$. For that split the dataset $\{X_{tr}, Y_{tr}\}$ into multiple training and validation sets (we used 5-fold splitting), train the model on each, and assess performance using a mean squared error metric. The α value yielding the best average performance across all folds is chosen as the optimal.
2. Apply the trained model m_1^* to test dataset X_{te} predicting $\hat{Y}_{te}$.
3. Compute performance metrics for the model m_1^* predictions. We propose to compute widely used regression metrics as R^2, mean absolute error (MAE), and Pearson correlation coefficient (r), i.e. metrics should be computed for the pair $\hat{Y}_{te}$ and Y_{te}.

Step 2

1. Train model m_2^0 on $\{X_{te}, \hat{Y}_{te}\}$. Select the optimal regularization parameter, α, as in Step 1.
2. Apply the trained model m_2^* to train dataset X_{tr} predicting $\hat{Y}_{tr}$.
3. Compute performance metrics for the model m_2^* predictions (R^2, MAE, r), i.e. metrics should be computed for the pair $\hat{Y}_{tr}$ and Y_{tr}.

Compare the obtained performance metrics on steps 1 and 2 of ITTP. If metrics are satisfactory and comparable then the datasets are interchangeable.

Within the study, we used a threshold $R^2 > 0.25$ to state that the model is performing satisfactorily. If satisfactory metrics are obtained only in step 2 then we state that datasets are not fully interchangeable, but the test dataset still contains the information for train dataset prediction.

ITTP case 2

ITTP case 2 (also referred to as the unfalsifiable case) considers a pair of datasets where only ground truth values are accessible for only train dataset: train dataset $\{X_{tr}, Y_{tr}\}$ and test dataset $\{X_{rep}\}$ (we denoted test dataset X_{rep} emphasizing that reprogramming datasets do not have ground truth values of biological age). The algorithm for this case is similar to ITTP case 1 with distinctions in step 1:

Step 1

1. Train model m_1^0 on $\{X_{tr}, Y_{tr}\}$. Select the optimal regularization parameter, α, as was described in step 1 of ITTP case 1.
2. Apply the trained model m_1^* to test dataset X_{rep} predicting $\hat{Y}_{rep}$.

Step 2

1. Train model m_2^0 on $\{X_{rep}, \hat{Y}_{rep}\}$. Select the optimal regularization parameter, α, as was described in step 1 of ITTP case 1.
2. Apply the trained model m_2^* to train dataset X_{tr} predicting $\hat{Y}_{tr}$.
3. Compute performance metrics for the model m_2^* predictions (R^2, MAE, r), i.e. metrics should be computed for the pair $\hat{Y}_{tr}$ and Y_{tr}.

In the second case of ITTP, we rely only on the performance metrics computed in step 2. If metrics are satisfactory the datasets are "probably" interchangeable. Otherwise, datasets cannot be used for prediction of each other according to the chosen linear model assumption.

For the training Lasso models during ITTP steps, we used *LassoCV* class from *scikit-learn* library [58] which conducts a simultaneous search for the optimal Lasso regularization hyperparameter α with cross-validation over the training subset.

Inference prediction uncertainty with Gaussian Process regression model

A Gaussian Process regression (GPR) model was developed to predict the age of samples given their methylome. GPR is a flexible non-parametric Bayesian approach for regression. In our model, the used inputs are the same CpG sites used for covariate shift analysis from published and Lasso newly constructed (see the section about Lasso training) aging clocks (Supplementary Table 2), and the outputs are the ages of sample donors. The model was trained using the python *scikit-learn* package with the default hyper-parameters adjustments.

Since aging clocks based on GPR have been described in detail elsewhere [**?**] (we use an equivalent methodology), here we will focus only on the part of the prediction uncertainty derivation that is essential for our study. A Gaussian Process (GP) is a probability distribution over possible functions that fit a set of points. Formally, it is a collection of random variables, any finite number of which have a joint Gaussian distribution [46]. Given train samples set $X = \{x_1, \ldots, x_n\} \in \mathbb{R}^d$, a mean function $m : \mathbb{R}^d \to \mathbb{R}$ and a covariance function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, a GP can be written as $f(x) \sim GP(m(x), k(x, x'))$ if the outputs $f = (f(x_1), \ldots, f(x_n))^T$ have a Gaussian distribution described by $f \sim \mathbb{N}(\mu, \Sigma)$, where $\mu = m(x_1, \ldots, x_n)$ and $\Sigma_{i,j} = k(x_i, x_j)$. The mean function is usually assumed to be the zero function, and the covariance function is a kernel function chosen based on assumptions about the function to be modeled. We tried different kernel functions and found the sum of the Radial Basis Function (RBF) and the white noise kernels the best in terms of prediction performance metrics. Interestingly, this kernel was also used by authors of the

previous GPR aging clocks [**?**]. The RBF kernel is defined as:

$$k(x_i, x_j) = s^2 exp\left(-\frac{||x_i - x_j||^2}{2l^2}\right) \tag{1}$$

where s^2 is the variance hyper-parameter, and l is the length-scale hyper-parameter (iteratively optimizing during training procedure) which controls the smoothness of the modeled function, or how fast it can vary. Since GP assumes the output variable includes additive Gaussian noise part $\varepsilon \sim \mathbb{N}(0, \sigma^2)$, i.e. $y_i = f(x_i) + \varepsilon$, the vector of outputs are viewed as $y \sim \mathbb{N}(0, \Sigma + \sigma^2 I)$. The term $\sigma^2 I$ reflects the white noise kernel added to the model to account noise of observations, which corresponds to the aleatoric part of prediction uncertainty.

Given a test point x^*, its output distribution is defined by $f^*|x^*, X, y$ which is a conditional Gaussian distribution having the following form:

$$f^*|x^*, X, y \sim \mathbb{N}(\mu^*, (\sigma^*)^2) = \mathbb{N}\left((k^*)^T(\Sigma + \sigma^2 I)^{-1}y, k(x^*, x^*) - (k^*)^T(\Sigma + \sigma^2 I)^{-1}(k^*)\right) \tag{2}$$

where $k^* = (k(x_1, x^*), \ldots, k(x_n, x^*))^T$.

Thus, given the training data, the distribution of predictions of a new point is given by a closed analytical form of Gaussian distribution. In our model, the inputs are DNAm methylation vectors, and the outputs are the ages of donors. The mean of the distribution μ^* can be used as the final prediction of the regression model (corresponds to the prediction of ElasticNet or other models). At the same time, the variance of the distribution $(\sigma^*)^2$ expresses the level of total prediction uncertainty — one of the most important aspects of our research. One can see that the magnitude of the uncertainty depends on the vector of covariates k^* of the new sample x^* with the training set samples X. Since the RBF kernel relies on the quadratic distance between samples, the total prediction uncertainty for an OOD sample will increase, as the sample moves away from the training distribution, until it reaches the limiting value $(\sigma^*)^2 \approx k(x^*, x^*) = s^2$ determining the upper bound of prediction uncertainty the model can estimate for OOD sample.

Testing rejuvenation effect with meta-regression approach

The Gaussian process model, yielding uncertainty levels in individual sample predictions as Gaussian distribution variances, enables statistical comparison of two predictions via, for example, the z-test (if two variances are assumed to be equal). For comparing prediction groups, each with unique variances, we employed advanced meta-analysis techniques. Utilizing *meta_regression* function from *pymare* library for Python, we accounted for individual age prediction variances in two *in vitro* reprogramming groups (*e.g.*, days 0 and 15). This function, which incorporates average ages, variances, and a binary group indicator in the design matrix, uses a restricted maximum likelihood approach to optimize meta-regression coefficients, providing coefficient estimates and their P values.

Acknowledgments

We would like to thank Leonid Peshkin for substantive discussions in the early stages of preparation of the work.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

[1] Nelly Olova, Daniel J Simpson, Riccardo E Marioni, and Tamir Chandra. Partial reprogramming induces a steady decline in epigenetic age before loss of somatic identity. *Aging cell*, 18(1):e12877, 2019.

[2] Kristen C Browder, Pradeep Reddy, Mako Yamamoto, Amin Haghani, Isabel Guillen Guillen, Sanjeeb Sahu, Chao Wang, Yosu Luque, Javier Prieto, Lei Shi, et al. In vivo partial reprogramming alters age-associated molecular changes during physiological aging in mice. *Nature Aging*, 2(3):243–253, 2022.

[3] Dafni Chondronasiou, Diljeet Gill, Lluc Mosteiro, Rocio G Urdinguio, Antonio Berenguer-Llergo, Mònica Aguilera, Sylvere Durand, Fanny Aprahamian, Nitharsshini Nirmalathasan, Maria Abad, et al. Multi-omic rejuvenation of naturally aged tissues by a single cycle of transient reprogramming. *Aging Cell*, 21(3):e13578, 2022.

[4] Diljeet Gill, Aled Parry, Fátima Santos, Hanneke Okkenhaug, Christopher D Todd, Irene Hernando-Herraez, Thomas M Stubbs, Inês Milagre, and Wolf Reik. Multi-omic rejuvenation of human cells by maturation phase transient reprogramming. *Elife*, 11:e71624, 2022.

[5] Jae-Hyun Yang, Christopher A Petty, Thomas Dixon-McDougall, Maria Vina Lopez, Alexander Tyshkovskiy, Sun Maybury-Lewis, Xiao Tian, Nabilah Ibrahim, Zhili Chen, Patrick T Griffin, et al. Chemically induced reprogramming to reverse cellular aging. *Aging (Albany NY)*, 15(13):5966, 2023.

[6] Daniel J Simpson, Nelly N Olova, and Tamir Chandra. Cellular reprogramming and epigenetic rejuvenation. *Clinical Epigenetics*, 13(1):1–10, 2021.

[7] Csaba Kerepesi, Bohan Zhang, Sang-Goo Lee, Alexandre Trapp, and Vadim N Gladyshev. Epigenetic clocks reveal a rejuvenation event during embryogenesis followed by aging. *Science advances*, 7(26):eabg6082, 2021.

[8] Alexandre Trapp, Csaba Kerepesi, and Vadim N Gladyshev. Profiling epigenetic age in single cells. *Nature Aging*, 1(12):1189–1201, 2021.

[9] Csaba Kerepesi and Vadim N Gladyshev. Intersection clock reveals a rejuvenation event during human embryogenesis. *Aging Cell*, page e13922, 2023.

[10] Jarod Rutledge, Hamilton Oh, and Tony Wyss-Coray. Measuring biological age using omics data. *Nature Reviews Genetics*, 23(12):715–727, 2022.

[11] Marije H Sluiskes, Jelle J Goeman, Marian Beekman, P Eline Slagboom, Hein Putter, and Mar Rodriguez-Girondo. Clarifying the biological and statistical assumptions of cross-sectional biological age predictors. *bioRxiv*, pages 2023–01, 2023.

[12] Paul D Yousefi, Matthew Suderman, Ryan Langdon, Oliver Whitehurst, George Davey Smith, and Caroline L Relton. Dna methylation-based predictors of health: applications and statistical considerations. *Nature Reviews Genetics*, 23(6):369–383, 2022.

[13] Alex Zhavoronkov, Polina Mamoshina, Quentin Vanhaelen, Morten Scheibye-Knudsen, Alexey Moskalev, and Alex Aliper. Artificial intelligence for aging and longevity research: Recent advances and perspectives. *Ageing research reviews*, 49:49–66, 2019.

[14] Josh Mitteldorf. A clinical trial using methylation age to evaluate current antiaging practices. *Rejuvenation Research*, 22(3):201–209, 2019.

[15] Michelle Chua, Doyun Kim, Jongmun Choi, Nahyoung G Lee, Vikram Deshpande, Joseph Schwab, Michael H Lev, Ramon G Gonzalez, Michael S Gee, and Synho Do. Tackling prediction uncertainty in machine learning for healthcare. *Nature Biomedical Engineering*, 7(6):711–718, 2023.

[16] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110:457–506, 2021.

[17] Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2008.

[18] Tristan Zindler, Helge Frieling, Alexandra Neyazi, Stefan Bleich, and Eva Friedel. Simulating combat: how batch correction can lead to the systematic introduction of false positive results in dna methylation microarray studies. *BMC bioinformatics*, 21:1–15, 2020.

[19] Michael F Adamer, Sarah C Brüningk, Alejandro Tejada-Arranz, Fabienne Estermann, Marek Basler, and Karsten Borgwardt. recombat: batch-effect removal in large-scale multi-source gene-expression data integration. *Bioinformatics Advances*, 2(1):vbac071, 2022.

[20] Xiaoyue Mei, Joshua Blanchard, Connor Luellen, Michael J Conboy, and Irina M Conboy. Fail-tests of dna methylation clocks, and development of a noise barometer for measuring epigenetic pressure of aging and disease. *Aging (Albany NY)*, 15(17):8552, 2023.

[21] World Health Organization et al. *WHO child growth standards: length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: methods and development*. World Health Organization, 2006.

[22] Alasdair GW Hunter, Jacqueline T Hecht, and Charles I Scott Jr. Standard weight for height curves in achondroplasia. *American journal of medical genetics*, 62(3):255–261, 1996.

[23] Julie E Hoover-Fong, Kerry J Schulze, Adekemi Y Alade, Michael B Bober, Ethan Gough, S Shahrukh Hashmi, Jacqueline T Hecht, Janet M Legare, Mary Ellen Little, Peggy Modaff, et al. Growth in achondroplasia including stature, weight, weight-for-height and head circumference from clarity: achondroplasia natural history study—a multi-center retrospective cohort study of achondroplasia in the us. *Orphanet journal of rare diseases*, 16(1):1–19, 2021.

[24] Richard M Pauli. Achondroplasia: a comprehensive clinical review. *Orphanet journal of rare diseases*, 14(1):1–49, 2019.

[25] James William Hollingsworth, Asaji Hashizume, and Seymour Jablon. Correlations between tests of aging in hiroshima subjects–an attempt to define" physiologic age". *The Yale journal of biology and medicine*, 38(1):11, 1965.

[26] Steve Horvath. Dna methylation age of human tissues and cell types. *Genome biology*, 14(10):1–20, 2013.

[27] Gregory Hannum, Justin Guinney, Ling Zhao, LI Zhang, Guy Hughes, SriniVas Sadda, Brandy Klotzle, Marina Bibikova, Jian-Bing Fan, Yuan Gao, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular cell*, 49(2):359–367, 2013.

[28] Kirsten Seale, Steve Horvath, Andrew Teschendorff, Nir Eynon, and Sarah Voisin. Making sense of the ageing methylome. *Nature Reviews Genetics*, 23(10):585–605, 2022.

[29] Shizhao Li and Trygve O Tollefsbol. Dna methylation methods: Global dna methylation and methylomic analyses. *Methods*, 187:28–43, 2021.

[30] AT Lu, Zhe Fei, A Haghani, TR Robeck, JA Zoller, CZ Li, R Lowe, Q Yan, J Zhang, H Vu, et al. Universal dna methylation age across mammalian tissues. *Nature aging*, 3(9):1144–1166, 2023.

[31] Daigo Okada. Application of a mathematical model to clarify the statistical characteristics of a pan-tissue dna methylation clock. *GeroScience*, pages 1–15, 2023.

[32] Leonie Roos, Johanna K Sandling, Christopher G Bell, Daniel Glass, Massimo Mangino, Tim D Spector, Panos Deloukas, Veronique Bataille, and Jordana T Bell. Higher nevus count exhibits a distinct dna methylation signature in healthy human skin: implications for melanoma. *Journal of Investigative Dermatology*, 137(4):910–920, 2017.

[33] Amy R Vandiver, Rafael A Irizarry, Kasper D Hansen, Luis A Garza, Arni Runarsson, Xin Li, Anna L Chien, Timothy S Wang, Sherry G Leung, Sewon Kang, et al. Age and sun exposure-related widespread genomic blocks of hypomethylation in nonmalignant skin. *Genome biology*, 16(1):1–15, 2015.

[34] Michael J Thompson, Karolina Chwiałkowska, Liudmilla Rubbi, Aldons J Lusis, Richard C Davis, Anuj Srivastava, Ron Korstanje, Gary A Churchill, Steve Horvath, and Matteo Pellegrini. A multi-tissue full lifespan epigenetic clock for mice. *Aging (Albany NY)*, 10(10):2832, 2018.

[35] Margarita V Meer, Dmitriy I Podolskiy, Alexander Tyshkovskiy, and Vadim N Gladyshev. A whole lifespan mouse multi-tissue dna methylation clock. *Elife*, 7:e40675, 2018.

[36] Mari Ohnuki, Koji Tanabe, Kenta Sutou, Ito Teramoto, Yuka Sawamura, Megumi Narita, Michiko Nakamura, Yumie Tokunaga, Masahiro Nakamura, Akira Watanabe, et al. Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential. *Proceedings of the National Academy of Sciences*, 111(34):12426–12431, 2014.

[37] Vadim N Gladyshev. The ground zero of organismal life and aging. *Trends in molecular medicine*, 27(1):11–19, 2021.

[38] Ghislain Auclair, Sylvain Guibert, Ambre Bender, and Michael Weber. Ontogeny of cpg island methylation and specificity of dnmt3 methyltransferases during embryonic development in the mouse. *Genome biology*, 15(12):1–16, 2014.

[39] Fedor Galkin, Polina Mamoshina, Alex Aliper, João Pedro de Magalhães, Vadim N Gladyshev, and Alex Zhavoronkov. Biohorology and biomarkers of aging: Current state-of-the-art, challenges and opportunities. *Ageing research reviews*, 60:101050, 2020.

[40] Steve Horvath, Junko Oshima, George M Martin, Ake T Lu, Austin Quach, Howard Cohen, Sarah Felton, Mieko Matsuyama, Donna Lowe, Sylwia Kabacik, et al. Epigenetic clock for skin and blood cells applied to hutchinson gilford progeria syndrome and ex vivo studies. *Aging (Albany NY)*, 10(7):1758, 2018.

[41] Morgan E Levine, Ake T Lu, Austin Quach, Brian H Chen, Themistocles L Assimes, Stefania Bandinelli, Lifang Hou, Andrea A Baccarelli, James D Stewart, Yun Li, et al. An epigenetic biomarker of aging for lifespan and healthspan. *Aging (albany NY)*, 10(4):573, 2018.

[42] Lisa M McEwen, Kieran J O'Donnell, Megan G McGill, Rachel D Edgar, Meaghan J Jones, Julia L MacIsaac, David Tse Shen Lin, Katia Ramadori, Alexander Morin, Nicole Gladish, et al. The pedbe clock accurately estimates dna methylation age in pediatric buccal cells. *Proceedings of the National Academy of Sciences*, 117(38):23329–23335, 2020.

[43] Ake T Lu, Anne Seeboth, Pei-Chien Tsai, Dianjianyi Sun, Austin Quach, Alex P Reiner, Charles Kooperberg, Luigi Ferrucci, Lifang Hou, Andrea A Baccarelli, et al. Dna methylation-based estimator of telomere length. *Aging (Albany NY)*, 11(16):5895, 2019.

[44] Xiaohui Wu, Weidan Chen, Fangqin Lin, Qingsheng Huang, Jiayong Zhong, Huan Gao, Yanyan Song, and Huiying Liang. Dna methylation profile is a quantitative measure of biological aging in children. *Aging (Albany NY)*, 11(22):10031, 2019.

[45] Qian Zhang, Costanza L Vallerga, Rosie M Walker, Tian Lin, Anjali K Henders, Grant W Montgomery, Ji He, Dongsheng Fan, Javed Fowdar, Martin Kennedy, et al. Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing. *Genome medicine*, 11:1–11, 2019.

[46] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

[47] Jie Wang. An intuitive tutorial to gaussian processes regression. *arXiv preprint arXiv:2009.10862*, 2020.

[48] Miri Varshavsky, Gil Harari, Benjamin Glaser, Yuval Dor, Ruth Shemer, and Tommy Kaplan. Accurate age prediction from blood using of small set of dna methylation sites and a cohort-based machine learning algorithm. *bioRxiv*, pages 2023–01, 2023.

[49] Jamie N Justice, Luigi Ferrucci, Anne B Newman, Vanita R Aroda, Judy L Bahnson, Jasmin Divers, Mark A Espeland, Santica Marcovina, Michael N Pollak, Stephen B Kritchevsky, et al. A framework for selection of blood-based biomarkers for geroscience-guided clinical trials: report from the tame biomarkers workgroup. *Geroscience*, 40:419–436, 2018.

[50] Alberto Parras, Alba Vílchez-Acosta, Gabriela Desdín-Micó, Sara Picó, Calida Mrabti, Elena Montenegro-Borbolla, Céline Yacoub Maroun, Amin Haghani, Robert Brooke, María del Carmen Maza, et al. In vivo reprogramming leads to premature death linked to hepatic and intestinal failure. *Nature Aging*, pages 1–12, 2023.

[51] Petr Klemera and Stanislav Doubal. A new approach to the concept and computation of biological age. *Mechanisms of ageing and development*, 127(3):240–248, 2006.

[52] Emma Pierson, Pang Wei Koh, Tatsunori Hashimoto, Daphne Koller, Jure Leskovec, Nick Eriksson, and Percy Liang. Inferring multidimensional rates of aging from cross-sectional data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 97–107. PMLR, 2019.

[53] Ake T Lu, Austin Quach, James G Wilson, Alex P Reiner, Abraham Aviv, Kenneth Raj, Lifang Hou, Andrea A Baccarelli, Yun Li, James D Stewart, et al. Dna methylation grimage strongly predicts lifespan and healthspan. *Aging (albany NY)*, 11(2):303, 2019.

[54] Nicholas J Schork, Brett Beaulieu-Jones, Winnie Liang, Susan Smalley, and Laura H Goetz. Does modulation of an epigenetic clock define a geroprotector? *Advances in geriatric medicine and research*, 4(1), 2022.

[55] Lisa Melton. Scientists hone tools to measure aging and rejuvenation interventions. *Nature Biotechnology*, 2023.

[56] Mahdi Moqri, Chiara Herzog, Jesse R Poganik, Jamie Justice, Daniel W Belsky, Albert Higgins-Chen, Alexey Moskalev, Georg Fuellen, Alan A Cohen, Ivan Bautmans, et al. Biomarkers of aging for the identification and evaluation of longevity interventions. *Cell*, 186(18):3758–3775, 2023.

[57] Aline Andres, Holly R Hull, Kartik Shankar, Patrick H Casey, Mario A Cleves, and Thomas M Badger. Longitudinal body composition of children born to mothers with normal weight, overweight, and obesity. *Obesity*, 23(6):1252–1258, 2015.

[58] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[59] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[60] Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

[61] Daniel A Petkovich, Dmitriy I Podolskiy, Alexei V Lobanov, Sang-Goo Lee, Richard A Miller, and Vadim N Gladyshev. Using dna methylation profiling to evaluate biological age and longevity interventions. *Cell metabolism*, 25(4):954–960, 2017.