Reproducing “Quantifying the Stochastic Component of Epigenetic Aging”

Tong, H., Dwaraka, V.B., Chen, Q. et al. 2024

https://doi.org/10.1038/s43587-024-00600-8

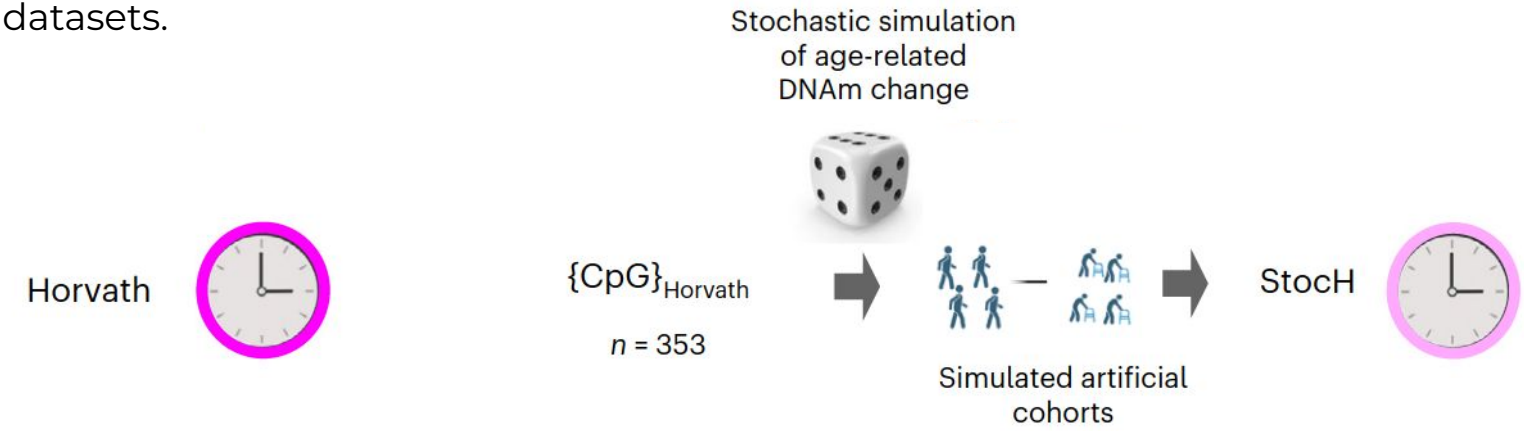
Team: Oksana Kotovskaya

Main hypothesis:

According to the paper, approximately 66–75% of the accuracy underpinning Horvath's clock could be driven by a stochastic process.

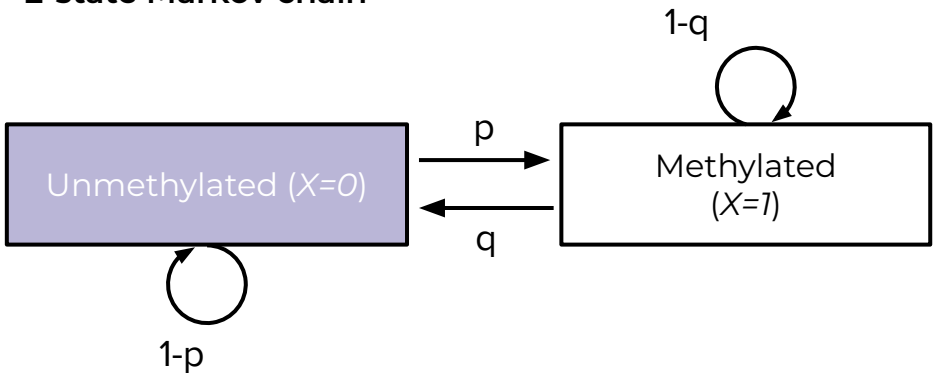
Objectives:

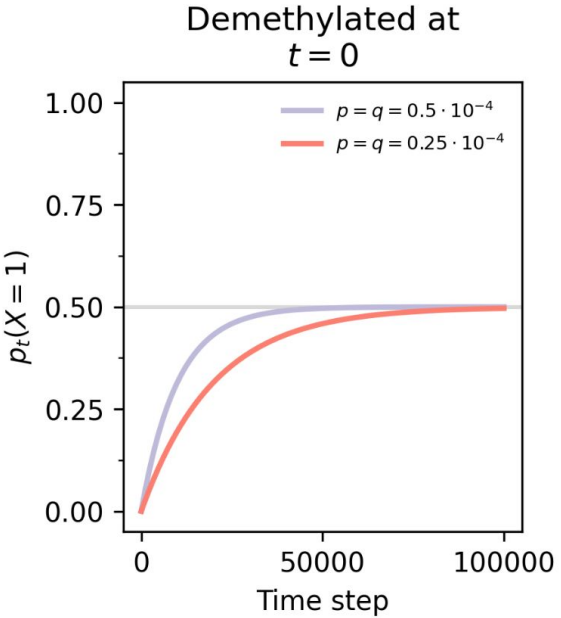
1. Reproduce the stochastic simulation for Horvath's clock;
2. Construct a stochastic analog (StocH clock) using DNAm dataset;
3. Quantify the stochastic component of Horvath's clock in one of the sorted immune cell datasets.



Modelling single-cell methylation dynamics of one CpG

2-state Markov chain

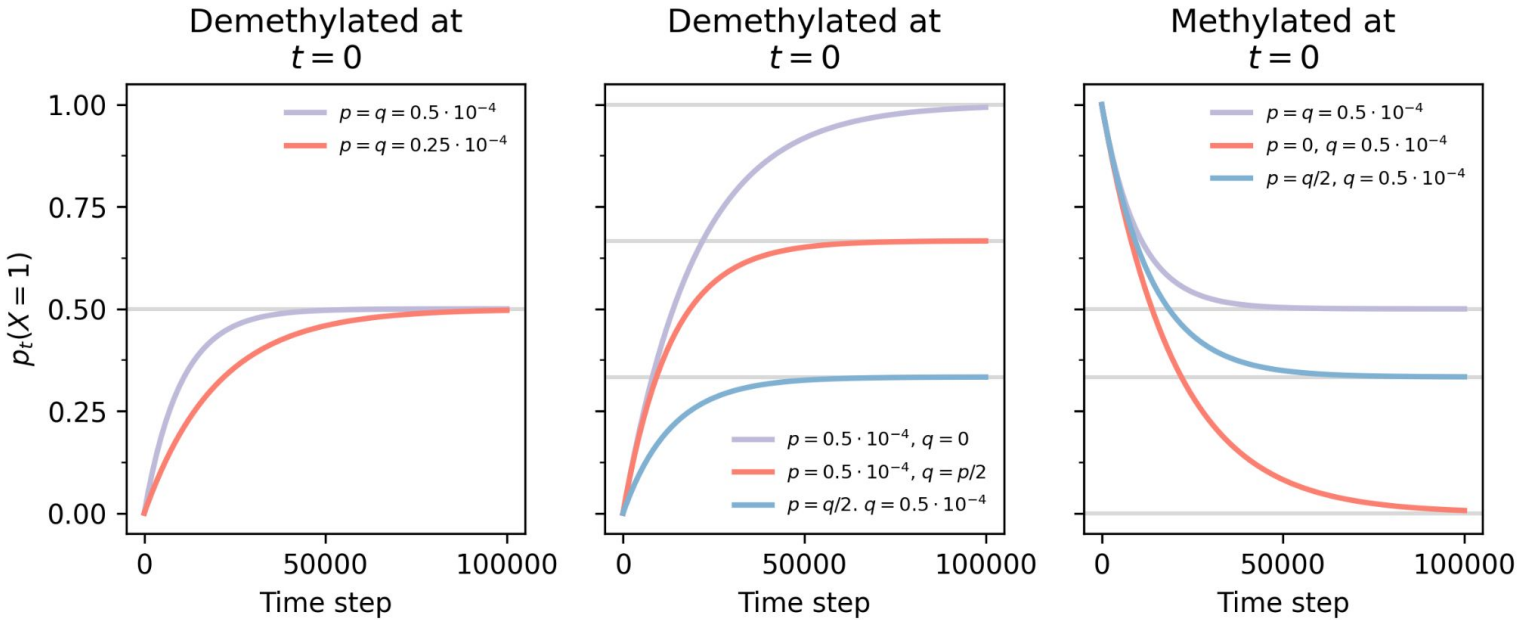




Markov process is set by two vectors:

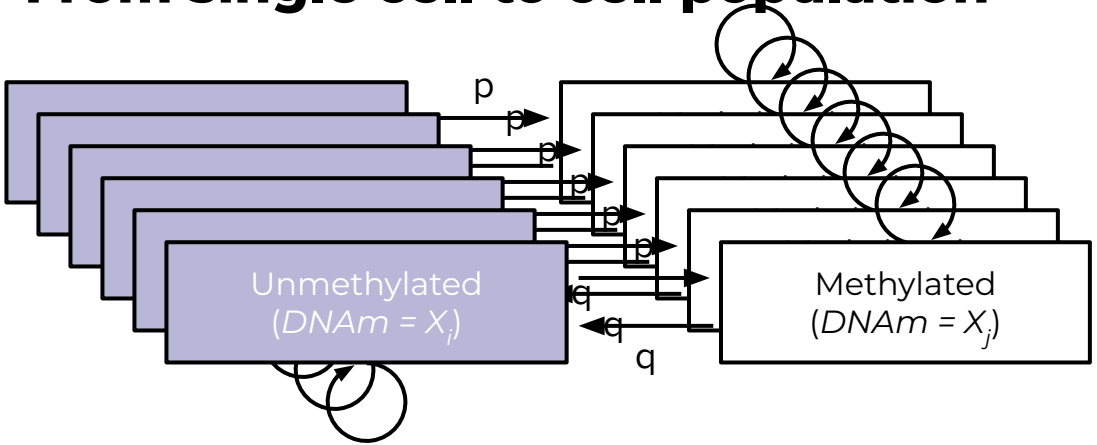
1. Vector of initial states;
2. Vector of transition probabilities.

Methylation dynamics in single cell-single site simulation



Markov processes become stationary over time and do not remember the initial state

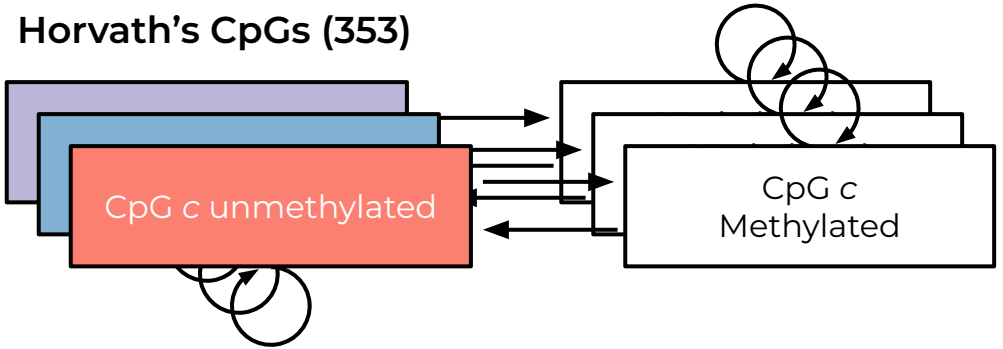
From single cell to cell population



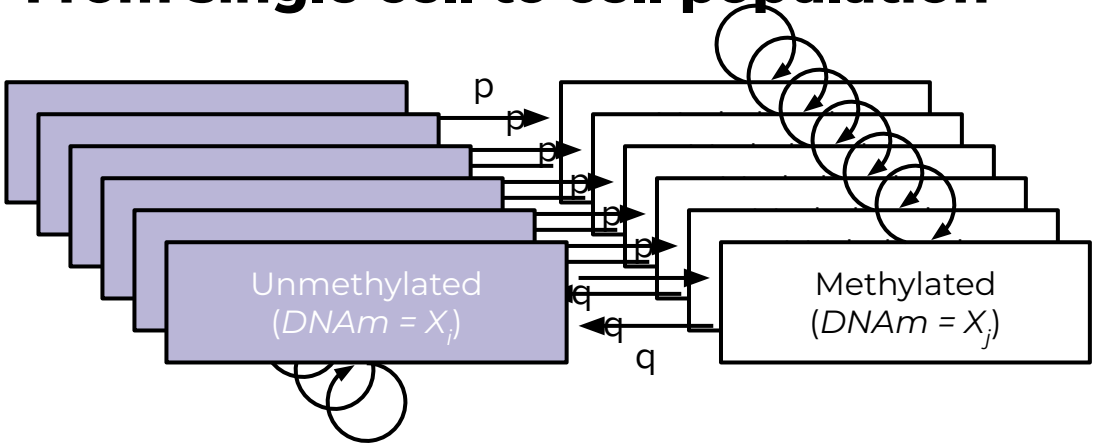
DNAm (fraction of methylated sites) as a proxy of each particular CpG methylation status

From single site to set of sites

Changes in all sites are modelled independently



From single cell to cell population

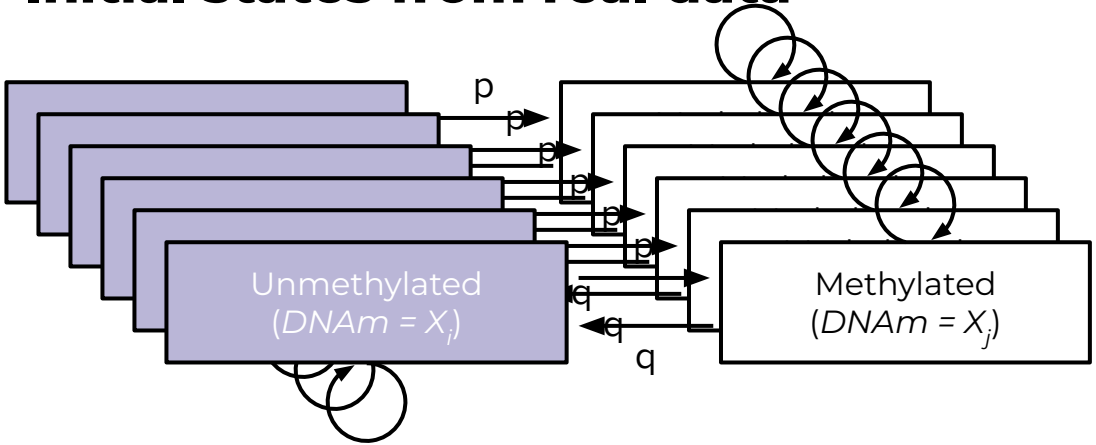


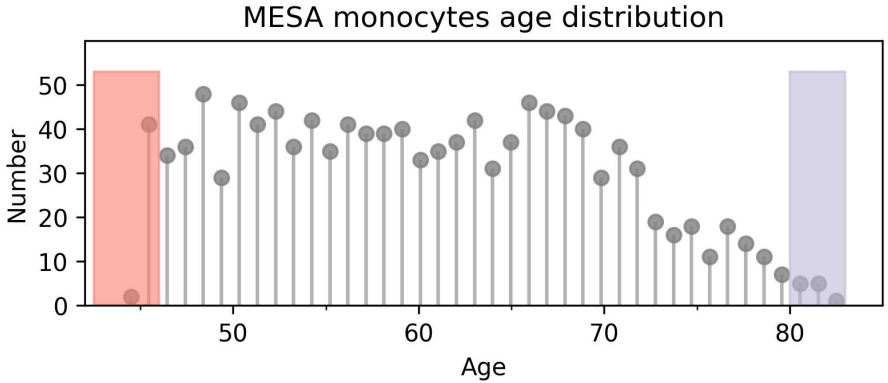
DNAm (fraction of methylated sites) as a proxy of each particular CpG methylation status

Markov process is set by two vectors:

1. Vector of initial states;
2. Vector of transition probabilities.

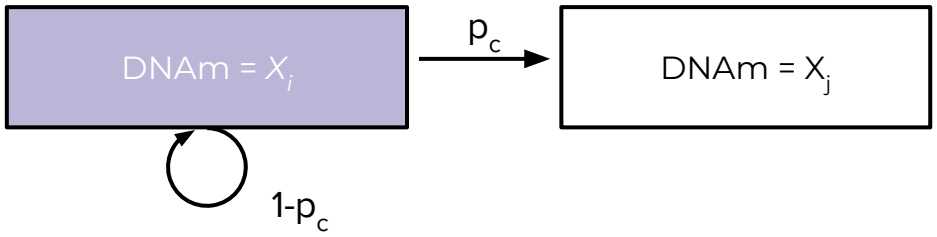
Initial states from real data

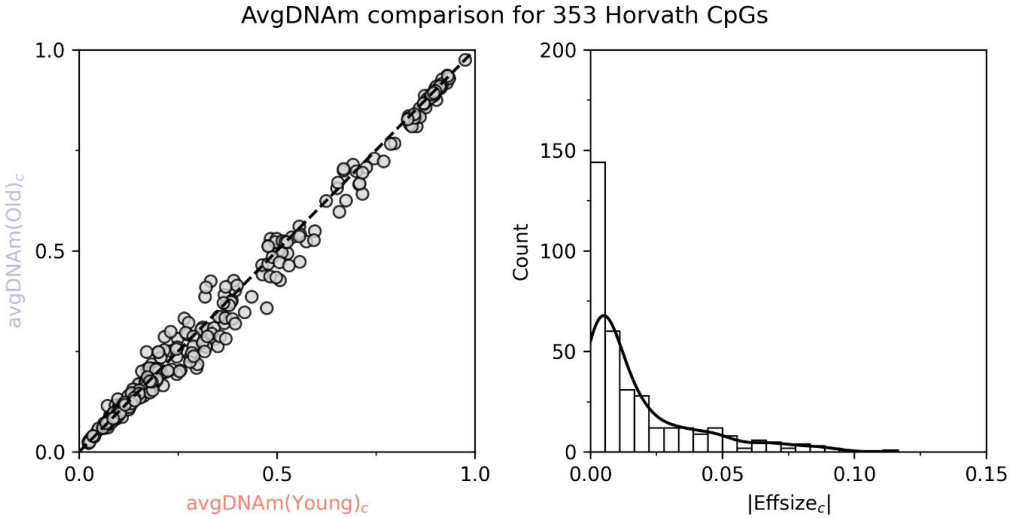




Initial DNAm from real data: average DNAm for young samples (avgDNAm$_c$(Young))

Transition probabilities from real data





At each timestep **status of methylation change** at CpG *c* can happen with probability p_c

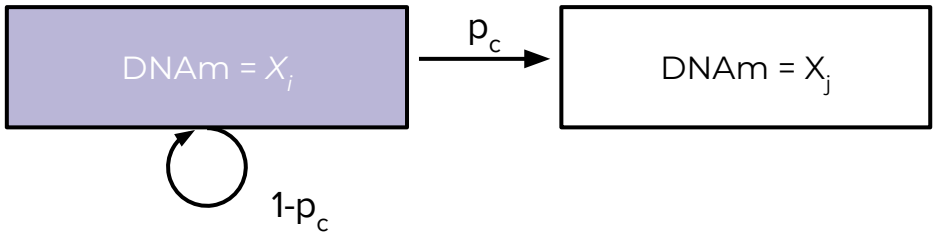
$$p_c = 1 - e^{-\gamma|\mathrm{EffSize}_c|}$$

where *γ* is parameter, characterising **global methylation change rate** associated with age.

Effect size (EffSize) is also derived from data:

$$\mathrm{EffSize}_c = \mathrm{avgDNAm}_c(\mathrm{Old}) - \mathrm{avgDNAm}_c(\mathrm{Young})$$

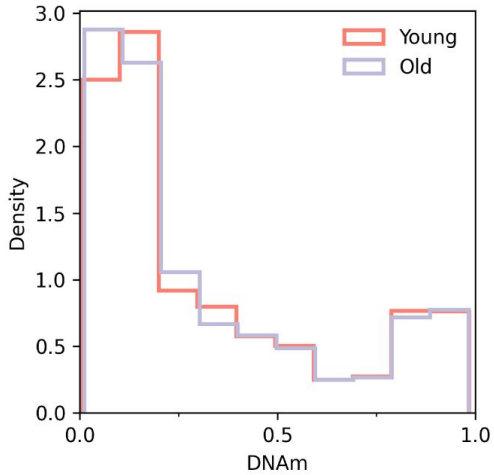
Methylation change at each step ($X_i \rightarrow X_j$)



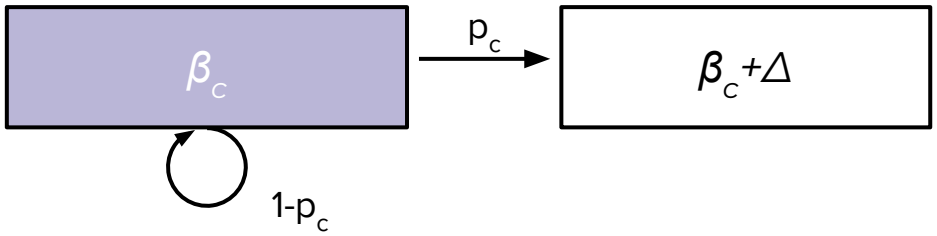
- Since we have a population, the change in methylation status of CpG *c* will change of DNAm from $\beta_c^{(t)}$ at time step *t* to $\beta_c^{(t+1)}$.
- **The size of DNAm change is defined as a random deviation r_c** from truncated normal distribution:

$$r_c \sim \mathcal{N}_+(0, \sigma)$$

DNAm are beta-distributed



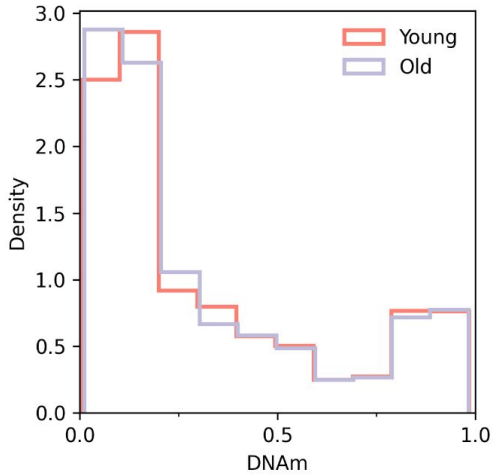
Methylation change at each step ($X_i \rightarrow X_j$)



- Since we have a population, the change in methylation status of CpG *c* will change of DNAm from $\beta_c^{(t)}$ at time step *t* to $\beta_c^{(t+1)}$.
- **The size of DNAm change is defined as a random deviation r_c** from truncated normal distribution:

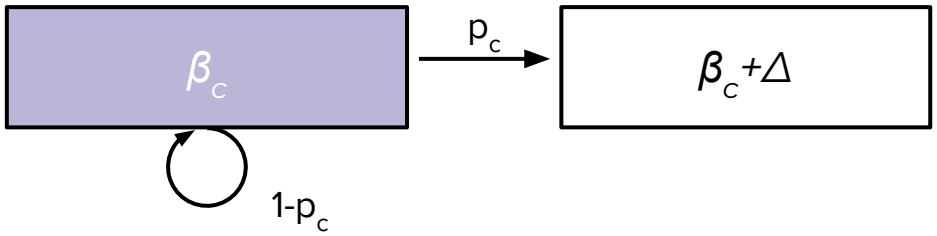
$$r_c \sim \mathcal{N}_+(0, \sigma)$$

DNAm are beta-distributed



- **Sign of change is same as sign of effect size**: $\text{sign}(\text{EffSize}_c)$

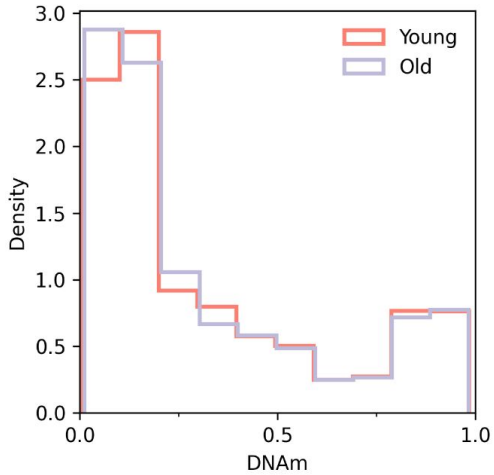
Methylation change at each step ($X_i \rightarrow X_j$)



- Since we have a population, the change in methylation status of CpG *c* will change of DNAm from $\beta_c^{(t)}$ at time step *t* to $\beta_c^{(t+1)}$.
- **The size of DNAm change is defined as a random deviation r_c** from truncated normal distribution:

$$r_c \sim \mathcal{N}_+(0, \sigma)$$

DNAm are beta-distributed

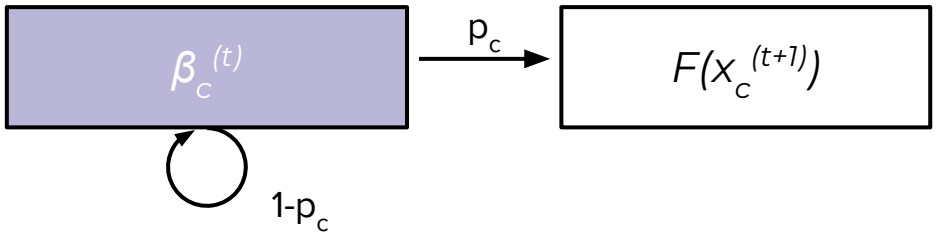


- **Sign of change is same as sign of effect size**: $\mathrm{sign}(\mathrm{EffSize}_c)$
- Since the DNAms $\beta_c^{(t)}$ are beta-distributed and the *r* belongs to truncated normal distribution we need to **transform the DNAm to normal distribution**, change the value, and then convert it back to the beta-values. Thus,

$$\beta_c^{(t)} \quad \rightarrow \quad x_c^{(t)} = iF(\beta_c^{(t)}),$$

where *iF* is inverse of the normal cumulative distribution function.

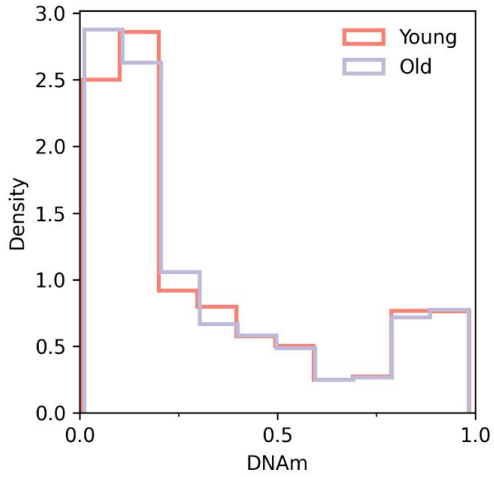
Methylation change at each step ($X_i \rightarrow X_j$)



- Since we have a population, the change in methylation status of CpG *c* will change of DNAm from $\beta_c^{(t)}$ at time step *t* to $\beta_c^{(t+1)}$.
- **The size of DNAm change is defined as a random deviation r_c** from truncated normal distribution:

$$r_c \sim \mathcal{N}_+(0, \sigma)$$

DNAm are beta-distributed



- **Sign of change is same as sign of effect size**: $\text{sign}(\text{EffSize}_c)$
- Since the DNAms $\beta_c^{(t)}$ are beta-distributed and the *r* belongs to truncated normal distribution we need to **transform the DNAm to normal distribution**, change the value, and then convert it back to the beta-values. Thus,

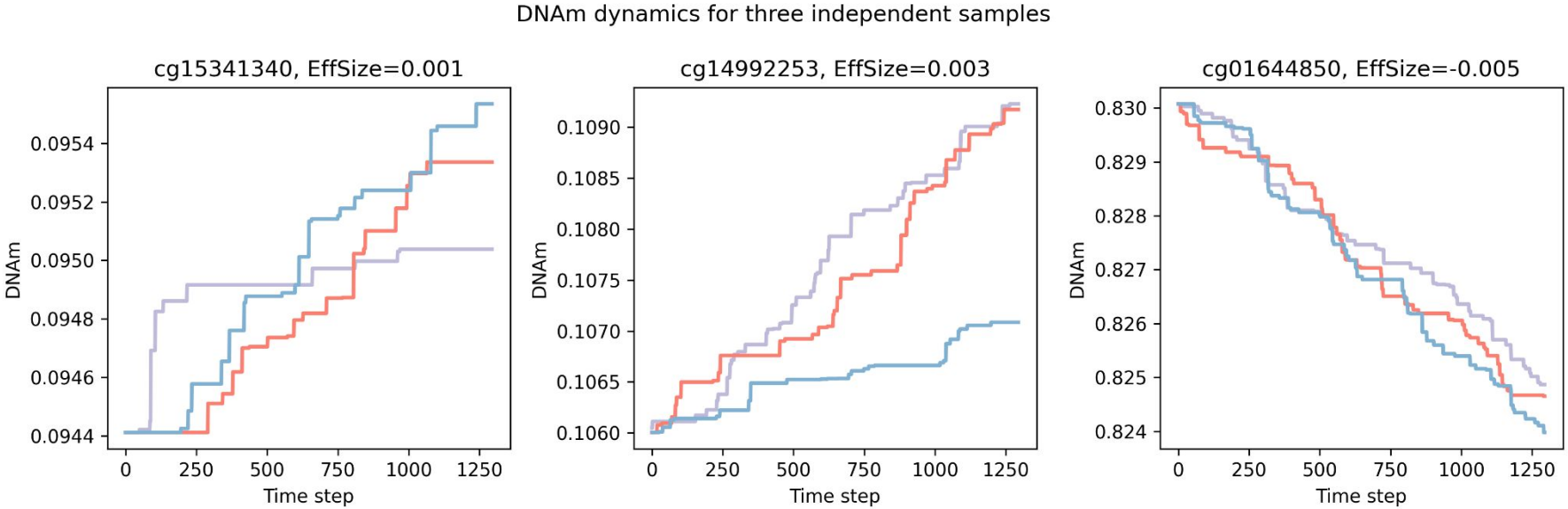
$$\beta_c^{(t)} \quad \rightarrow \quad x_c^{(t)} = iF(\beta_c^{(t)}),$$

where *iF* is inverse of the normal cumulative distribution function. Next,

$$\boxed{x_c^{(t+1)} = x_c^{(t)} + \text{sign}(\text{EffSize}_c) \cdot r_c,} \quad r_c \sim \mathcal{N}_+(0, \sigma)$$

and, finally, $\beta_c^{(t+1)} = F(x_c^{(t+1)})$

Simulated DNAm dynamics (for fixed γ, σ)



DNAm changes with time **monotonously**, as expected.

One can see that the **increase of EffSize increases chance of DNAm change**, while its' **sign defines direction** (increase or decrease) of DNAm change.

Optimized parameters γ, σ

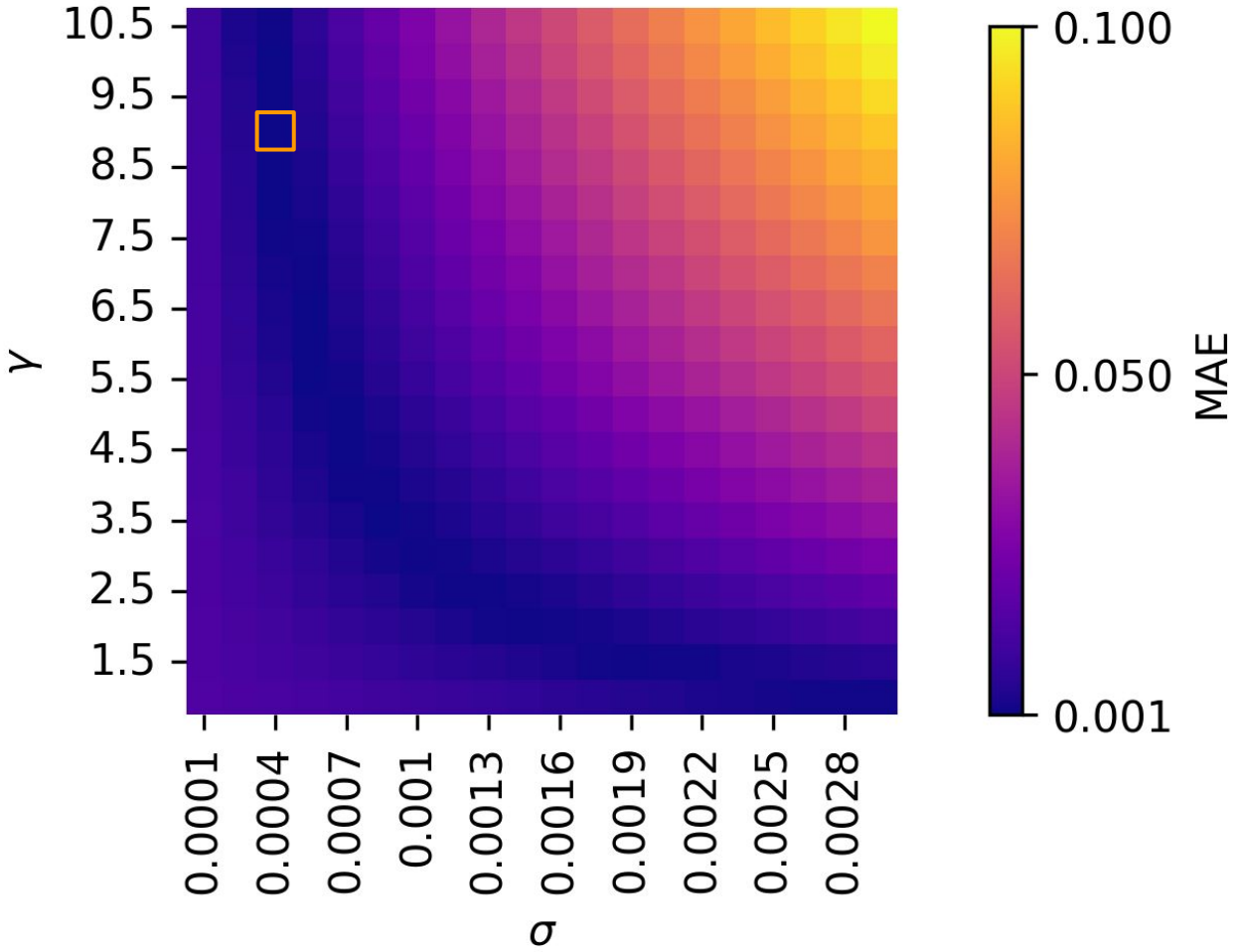
I ran simulation for each pair of parameters on grid from the right.

Optimal parameters obtained:

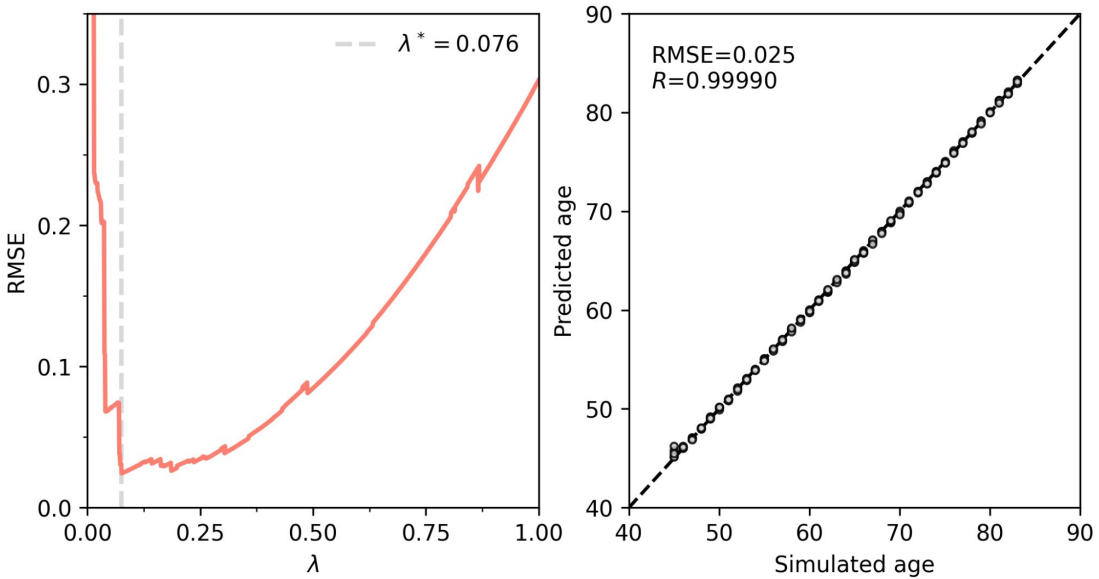
γ = 9.0,

σ = 0.0004

with mean absolute error (MAE) between observed and simulated effect size: 0.003.



Construction of the StocH clocks



Dataset:
3 artificial cohorts of 5 samples per age from 45 to 83 years – train, test, simulation.

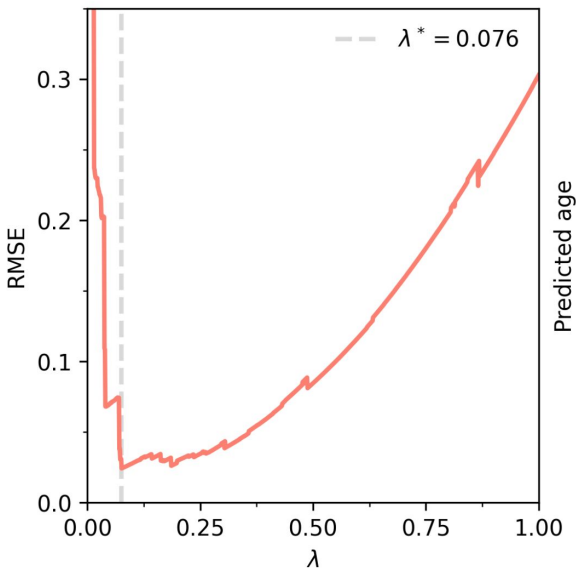
Train linear model with regularization for a set of penalties λ (from 0 for 1 with step 0.001) on **train cohort**.

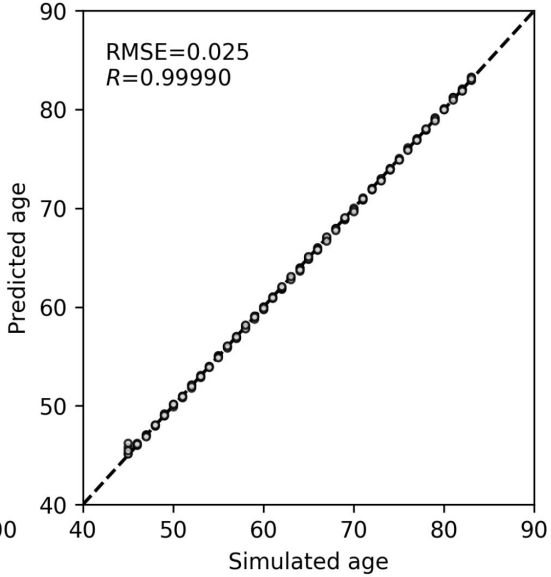
Test models on test cohort to find best λ^* (left figure).

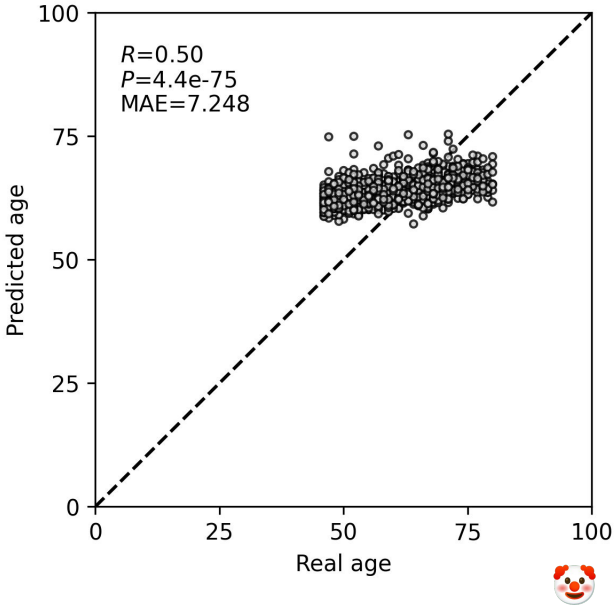
Validate results on validation cohort (right figure).

Best ElasticNet model is our **StocH**.

Age predicted with StocH is far from chronological one



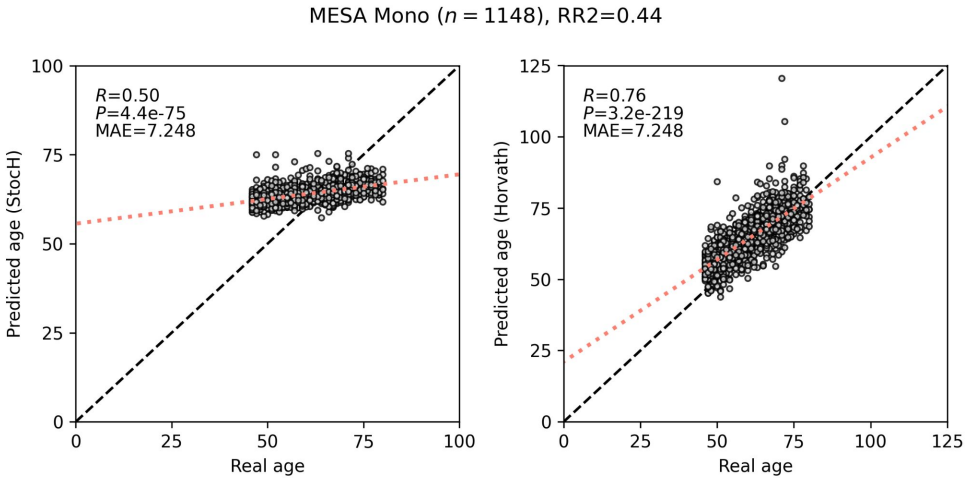


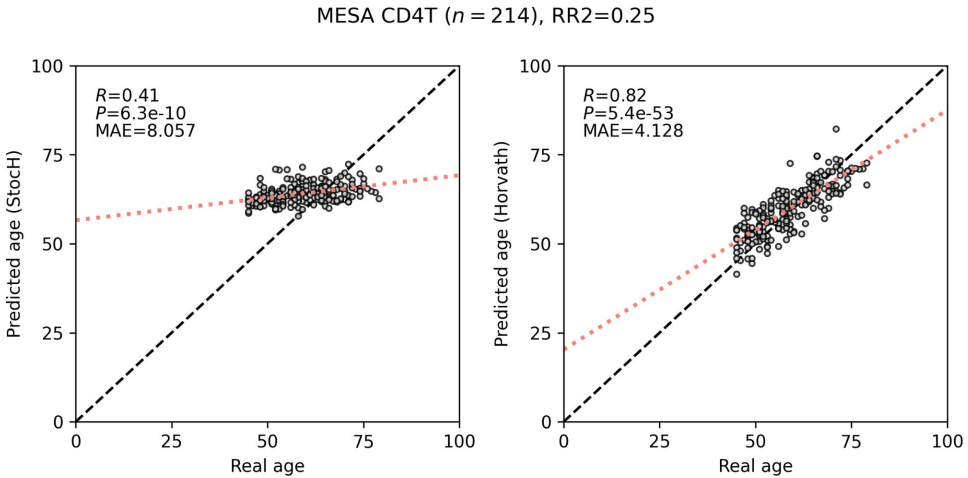


R – Pearson correlation.

Results has not been reproduced.

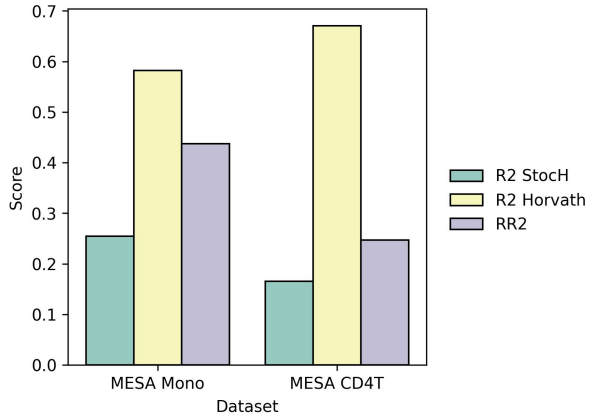
Quantification of stochastic component of Horvath's clock





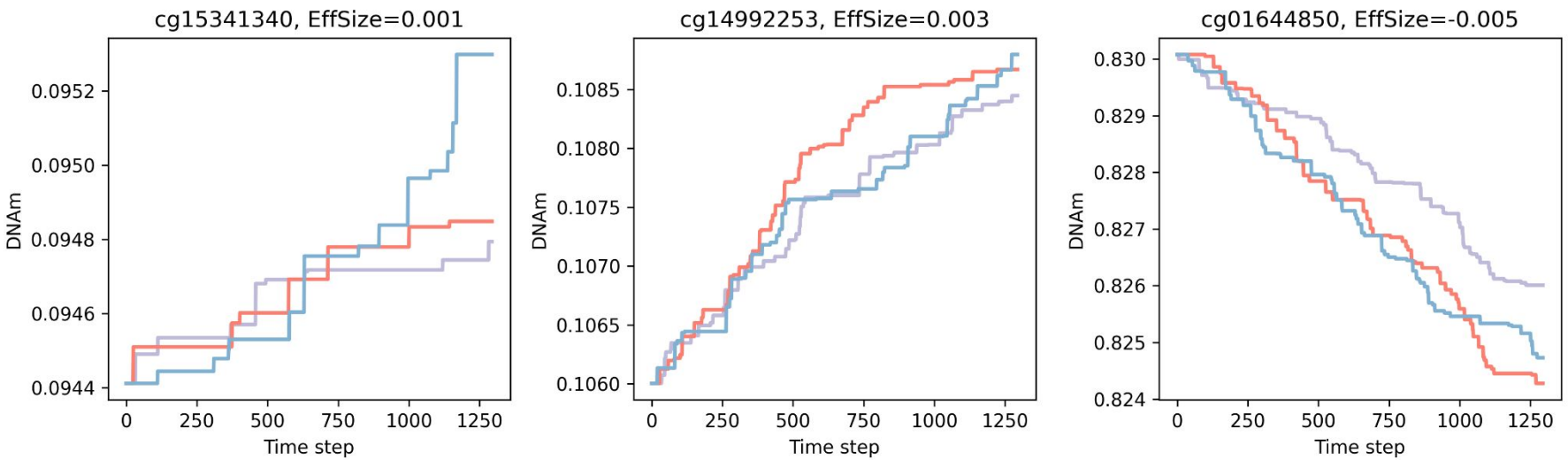
(left) Scatter plots of predicted age versus chronological age for the StocH clock and Horvath's clock.

(bottom) Quantified stochastic component (RR2 0.44 for monocytes ad 0.25 for T-cells). The obtained results are less than proposed in the paper (0.66–0.75).



Questions

- Can resulted simulated **DNAm exceed 1?** There is no restricting conditions.
- Which datasets authors used to build clocks? Did they used **any validation** data?
- There is no information about definition in paper. Did authors use determination scores or just **squared Pearson correlation?** The R2 for StocH-clocks can be negative (I have an evidence).
- Could such a **difference in results** be explained by difference in optimal in parameters? Not really:



Conclusions

- As a result, the project achieved all of its objectives.
- Not all the values obtained during the following of the paper methods were reproduced. Particularly, the RR2 values obtained from analysis are lower for two used datasets (MESA Mono and CD4T).
- The analysis is available as Python code for review and evaluation.

Suggestions

- Can we consider non-homogeneous Markov process by define time-dependent global methylation rate (gamma)?
- Can we use stochastic clocks as random matrix in mixed linear models?

Follow QR-code to see meme in project repository

Evidence for my honest excitement of the project results: