Improving and Interpreting Epigenetic Age Predictors

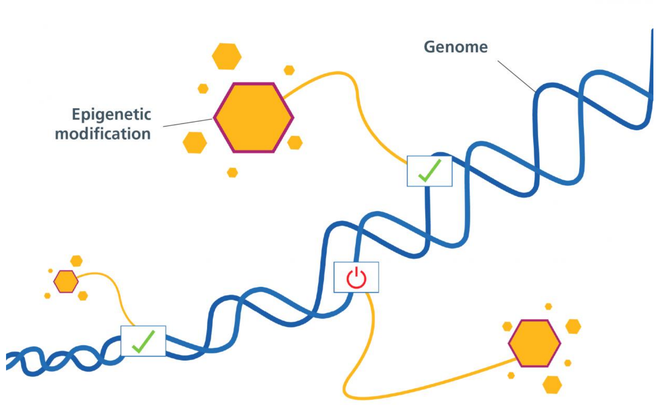
A Machine Learning Approach to Improving Epigenetic Age Predictors and Understanding How DNA Methylation Affects Aging

Author: Elena Langens, **email**: elangens@tudelft.nl

Supervisors: Bram Pronk, Inez den Hond, Gerard Bouland

Responsible Professor: Marcel Reinders

Biological Background



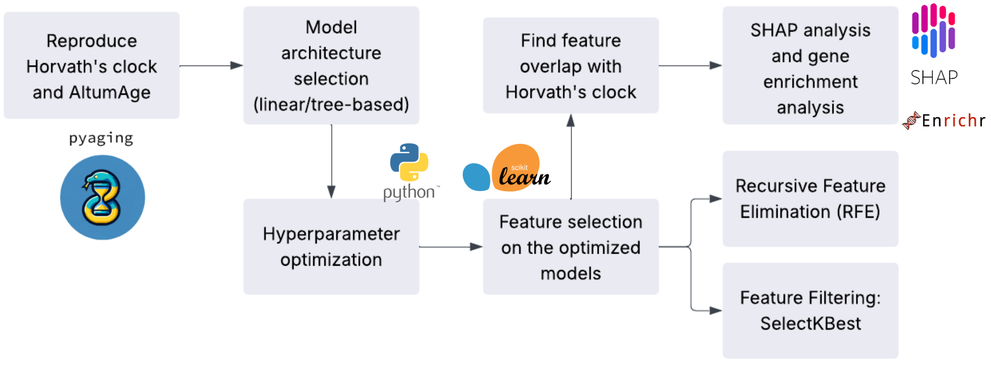
- **Epigenetics** describes how other molecules within a cell interact with DNA
- **Epigenetic modifications** can change gene expression
- Genes can be **'switched on' or 'switched off'**

- **DNA methylation** is a type of epigenetic modification
- DNA methylation involves adding a methyl group (-CH3) to a **CpG site**
- Methylation at CpG sites can **repress gene activation**
- **Aging Clocks** predict the age of a cell based on methylation levels at CpG sites
- Epigenetic aging clocks based on DNA methylation data (Horvath, AltumAge)

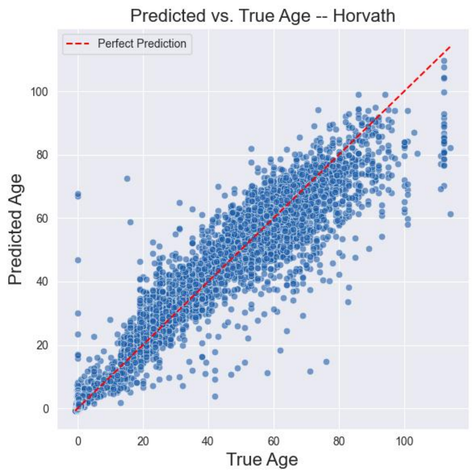
Research Question

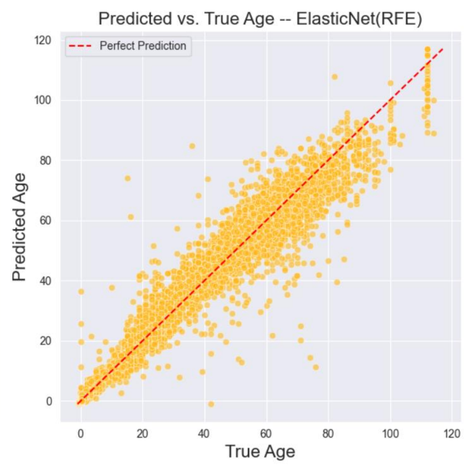
Can we reproduce or improve upon current age predictors based on epigenetic modification data and interpret the most important features for prediction? Can we use this knowledge to find biomarkers for aging?

Methodology



Results





ElasticNet hyperparameter optimization

$$\text{Obj} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \alpha\left[\lambda\sum_{j=1}^{p}|\beta_j| + (1-\lambda)\sum_{j=1}^{p}\beta_j^2\right]$$

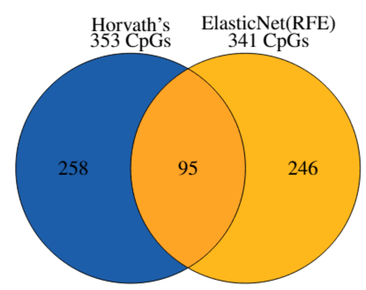
- **5-fold** cross validation for hyperparameter tuning
- Alpha = 0.01, lambda = 0.2
- L1 regularization strength of **0.02**, L2 regularization strength of **0.08**
- L1 regularization promotes **sparsity** and reduces coefficients to 0
- L2 regularization prevents **overfitting** by penalizing large coefficients but does not reduce them to 0

Improvements on Horvath's clock

Model	CpGs	MedAE	MSE	R
Horvath's clock	353	3.530	71.030	0.951
ElasticNet(RFE)	341	**2.820**	**44.085**	**0.970**

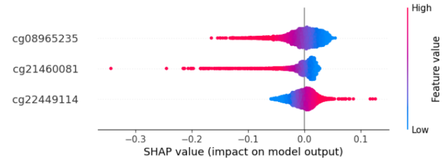
- **ElasticNet(RFE) has a lower Median Absolute Error**
- **ElasticNet(RFE) uses less CpG sites to predict age and is therefore more interpretable**

Overlapping features



SHAP and gene enrichment analysis

- 3 of the 95 overlapping genes were involved in the process **Positive Regulation of Stem Cell Differentiation**, which was statistically enriched
- Genes are mapped to CpG sites and are analyzed with SHAP



The first two CpG sites **correlate negatively** with age prediction

- The 95 (of 341) overlapping genes contributed **35%** to the total SHAP value
- **80%** of SHAP values were explained by **195** features
- The highest SHAP value was **0.087**

Conclusions

- **ElasticNet(RFE) outperforms Horvath's clock with less features.**
- **Linear models benefit from some feature reduction, but not from a very small subset of CpG sites. Aging seems to be influenced by a varied set of epigenetic changes.**
- **RFE is a better feature selection technique than feature filtering due to interaction between CpG sites.**
- **SHAP analysis indicated that age is not predicted by a small subset of features.**
- **Stem cells are biomarkers for aging.**
- **Deep learning models capture interactions between CpG sites better and serve as more accurate age predictors.**