

Bioinformatics aspects of aging and rejuvenation

Zoltán Szarvas¹

lecturer: Csaba Kerepesi²

¹Faculty of Computer Science, Eötvös Loránd University

²HUN-REN Institute for Computer Science and Control

May 2024

Assignment 1

Train epigenetic clock for microarray-based methylation dataset[2]

```

CCGGGTCCGGCGGGGAAGAGCGGCTCAACGGCAGGGGCCCATCGCGA
GAGGCCACCGCCCCCGCGGTCCAGCCCCAGGCCCGCGCCCTCGCCCTG
GGCTGCTCCTCGGGCCCTCGACCGCCTCTGCTACTTGTGGAACGCTTC
CTCACCGCCCTTCTCCACCCCGCGCGCCGAGCCTCCCGCGCGCACGGTGGG
ATTCGGCCAAATAAGGAGAAAGGGCGCGGCCGTACGCGCGCCAGGTGC
GTGGGCGAGACCAGCTCACGCCCTCTCTCCAGCGCCAAGGCCCGGCC
ACAGCTGCTTGGCTGCAGTCGAAGCGTAGCCCGAGACAAGGAAGGGCGC
CTTGACCTCGCACTTTTGTTCGGTTCGAACGTTCTGCTCAGTGGTGGTGG
AATGACGACCGCTCTAAATCGATGCGCTAGGAGTCCATGAATAACG
GTACAGGCTTCGCGCGCGATACCGACGCTCGCCCTCGCGCCCT
CGGGGATGCCCCACCCCTCGTGGCGTCCCGCCCGTCCCGCGAGCGCG
CGCTCGGGTGCCTGGCTCTTCGACGCGGCCATGCGCGACTCGAGC
TGACGTGGTTGAGCAGCGATCGCAGCTTCCCGCACTTCCCAACCCCA
GGCGTGTGATTAGGTGACCGCACAGGCGCGCCCTTGCGCGCCCGACCT
CGGGGCTAGCGATGGGACCGCGGTGGCCCGCGACCTTCGGGGCGCGGGG
GGGAACCCCTCGCTTTCGGGGGGGCTGCTCTCTCGCGCCCTCGCGCC
CGTACACAGCCTGCTCTTGGGTCCAGGGACACTCGCTCGCTCTGAAGG
ACCCCGCTCTTCGCGCGCGCATCGCCTCTGCGCGACACCTGAAG
GCGACCAACGGGGCGCATCGACTACATCGAGGCGAGTGCCAGTGCC
CGCATCTAGGGCGCTTCGCGCTTGGCGCGCGAGGGCAGCACTGGGGC
TCTGCGGCTCTGCTTGGGGGAGGGCCTTGGGGTGTCTCAGGGGGCGCG
GGACGGCGCGCTGCTTGGGTGGCGCGGGAAGGTTGTGAGATTGAGCCC
CGAGGGCGCGCGCGCTGCTGCGCGCTGCTGCTGCTGCTGCTGCTGCT
AGCCAGGACAGCGCTGACGAGTTTCGCGGTCACTGGTCTCCCTGGAG
TGCCCAAGCTGAATCCAGAGGGGCCAGCTGCTGCTGCTTCTGCTCTCT
CGAGCTGGATTGAGCGCCTGCCACGAGCCAGGCTTCCCTGGTGAAGA
TCACGGAATGCCCAACCCAGGGAAGGGAGGCGCTGAGGCCCTCGGGAGGC
CCAAGAGGTGGCCAGAGGAGAACAGAGTGTTCCTGGCGCTTTCCTCTC
CTAGGGTGTGACAGCCACTTCCCTGGACACTGCTGAGGAAAGCGAGC
CTTGTGGAGCCACAGACTGCGAGAGCTTCTCTCACCTTCTCTCCAC
GAAGCCCTTCTGACCTTCTGCCAGCGCGGGCAGGGTTTCCCTGAGCGT
CCCCAACCATCACAGCTCAGGCCACCTCGAGAGACTCCCTTTTAGACA
GAAGCCTGGTGACAGAGCTGCTTTGAGAGTAAGCTGAGGCTGTGAGGT
TCTTACCGCCAGTTACAGATGGGCTGCTCAGCTCAGAGAGAGGGGTGG
TGACTCCCTTAGGAACACAGACTAGAGTGTGCTTCTTAAAGACAGAC
CGAGGTCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCT
ATTCTTAAATGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG
TACCCTTGGCACCGCCAAACCTGGCCCTGGGGATTCCCATGCTGCGC
AGTCACTCTGCTGCTTACCCTGACAGGCTGAGTCTCGAGGCTTCTCT
TTTGGCCCTTCCCTGGCCAGGAGCTTGGACTGGGCTCGTGTCTATCGC
AAGCGGGGAAGCTGCGAGGCCCACTCTGTGGGCTTCTATTCCTTGG
AGTACGGAAGTTAGAGGCTGGGTGGCAGAGGAAGGCGAGGCGAGC
GGCACCTGGGCACTTCCCGCACCTTCCAGCGGTGTC
AAGTGAGCTGCTGTGGTTACAGTGGCTTGGGAGCTCAGAGAGGTGAG
ACATAGGCTGGGCTCACACAGCCAGGTAAACGCAAGGTGGGGTTGGAGTC
AGGGTCTAGGGTGGCAGCTGCCAAGCTGTGCAACAAAGCTGTTTCTCGC
GGAGGCTGAGGACACACACCACTTCCCACTCAGGCTGAGCTGGAGATT
CAGAAAGAAGCCTGGAGCCAGGACAGAGGTGGTGGCGTGGATGATCT
GCTGGCAGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCT
AGGCTCTGGTGGAGTGGGACAGGACCTCGCTGTGTGACATGGGATCG
CTTACTGTTGTCCAGAGGTGCTGCTGCTGGCCAGGCGACACCTTCTCTC
CCATGCTTCTCCCTCCCAACCCAGGGGCTGGCTGGAGCACTGCTCT
CTGACGCCAGCCAACTGGGGACCTCACTTCCATCCCAAGGAACCAT
GAAAGCTGCTGTGAGCTGCTGGGCGCCTGCGAGGCTGAGGTCTGGAGT
CGGTGAGGCTGCTGAGCTGAGCTGCTGCTGCTGCTGCTGCTGCTGCTG
CTGTGAGGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCT
AGCCAACTCTCCAGCTGGATCCAGGGAAATATCAGCTTGGGCACT
GCAGTGACAGGGGCGCGCTGCCACAGGGAACACATCTTCTGCTGG
GGTTACGCGCTTCTGGGGCTGGAAGTGCCAAAGCTGGGGCAAGCT
GTGTTTACGCCACTGAACCAATTACACACAGCGGAGAAAGCAGTAA
ACAGCTTTCACC

```

Dataset

Source

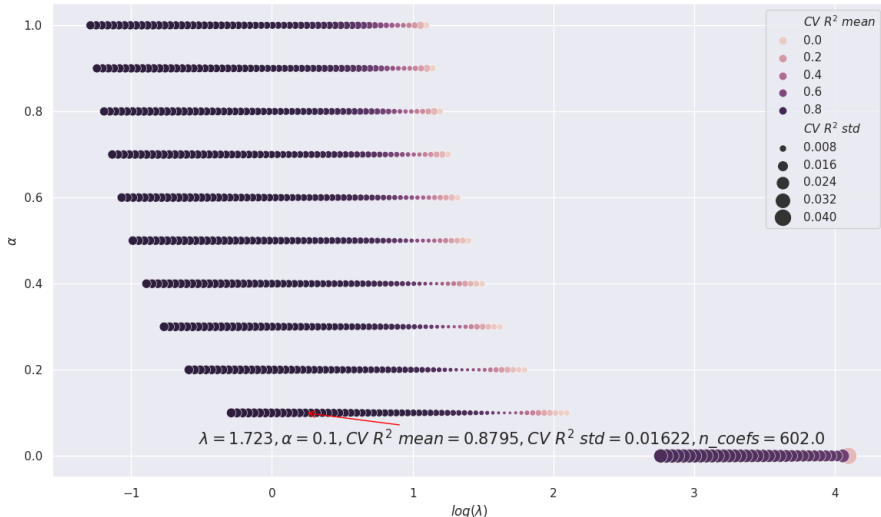
- Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates (Gene Expression Omnibus) [4]
- approximately 450k CpGs from **human whole blood**
- Illumina Infinium 450k (microarray)

Details

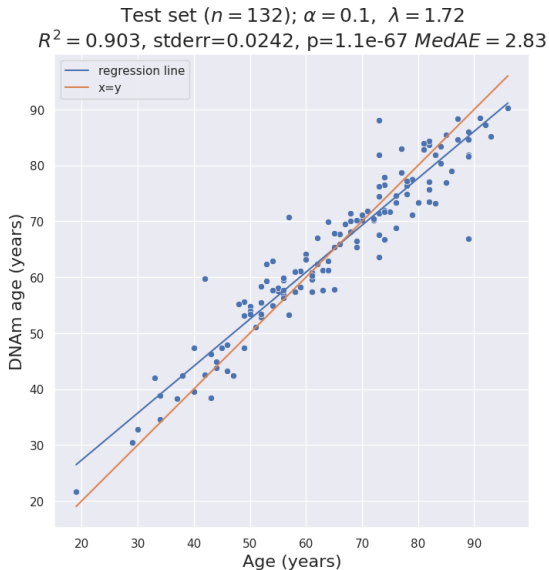
- sample N=656
- features=473 034

Hyperparameter Optimization (ElasticNet)

Hyperparameter optimization result on training data
best is selected based on $(CV R^2 \text{ mean}) - 0.5 \cdot (CV R^2 \text{ std})$



Results on Test Set



Assignment 2

Train epigenetic clock for RRBS-based methylation dataset (with missing data)[3]

```
CCCGGGTCCGGCGGGGAAGAGCGGCTCAACGCGAGGGGCCATCGCGA
GAGGCCACCGCCCCCGCGCTCCAGCCCCAGGCCCGCGCCCTCGCCCTG
GGCTGCTCCTCGCGGCCCTCGACCGCCTCTGCTACTTGTGGAACGCTTC
CTCACCGCCCTTCTCCACCCCGCGCGCCGAGCCTCCCGCGCGCACGGTGGG
ATTCGCGCAATAAAGGAGAAAGGGCGCGGCCGTAACGCGCGCCAGGTGC
GTGGGCGAGACCAGCTCACGCCCTCTCTCCAGCGCAAAGGCCCGCGCCC
ACAGCTGCCTGGCTGCAGTCAAGAGCGTAGCCCGAGACAAGGAAGGCGCG
CTTGACCTCGCACTTTTGTTCGGTTTCGAACGTTTCTGCTCAGTGGTGC
AATGACGCGCTCTTAAATCGATGCGCCTAGAGGTCCATGAATAACG
GTACAGGCTTCGCGCGCGGATCGACCGCGCTCGCCCTCGCGCCCT
CGGGGATGCCCCACCCCTCGTGCCTCCCGCGCTCCCGCGAGCGCG
CGCTCGGGTGCCTGGCTCTTCGACGCGGCCATGCGCGACTCGAGC
TGACGTGGTTGAGCAGCGGATCGCAGCTTCCCGCACTTCCCAACCCCA
GGCGTGGTATTAGGTGACCGCACAGGCGCGCCCTCGGGCGCCCGACCT
CGGGGCTTGCATGGAGCGCGCTGGCCCGCGACCTCGGGCGCGCGGG
GGGAACCCCTCGCTTTCGGGGGCGCTGCTCTCGCGCCCTCGCGCCCT
CGTACACAGCCTGCTTCTGGGTCCAGGACACTCGCTCGCTTGAAGG
ACCCCGCTCTTCGCGCGCGCATCGCCTCTGCGCGACACTGAAG
GCGACCCAAGGGGCGCATCGACTACATCGAGGCGAGTGCCAGTGCC
CGCATCTAGGGCGCTTCGCGCTTGCAGCGCGCGAGGGCAGCACTGGGG
TCTGCGGCTCTGCTTGGGGGAGGGCCTTGGGGTGTCTCAGGGGGCGCG
GGAAGCGCGCGCTGCTTGGGTGCGCGGGAAGGTTGTGAGATTGAGCCC
CGAGGGCGCGCGAGCTTCGCGGCTCTCGCGGCTCTCGCGGCTCTCG
AGCCAGGACAGCGTGAACGAGTTTCGCGGTCAAGTTGCTTCTCGGAG
TGCCAAGCTGAATCCACAGGGGCCAGCTGCTTGTCTTCTTCTCTCT
CGAGCTGGATTAGAGCGCTGCCACGAGCCAGGCGCTTCCCTGGTGAAGA
TCAAGGAATGCCCAAGGGAAGGGAGGGCCTGGAGGCGCTCGGGAGAGC
CCAAGAGGTGGCCAGGGGAGAACAGAGTGTCTTGGCGCTTTCCTCTC
CTAGGGTGTGACAGCCACTTCCCTGGACACTTCCCTGAGGAAGCGAG
CTTGTGGAGCAGACACTGCGAGAGCTTCTCTACCTCTCTCAGC
GAAGCCCTCTGACCTTCTGCGAGCGCGGGCAGGGTTTCCCTGAGCGT
CCCCAACCATACAGCTCAGGCCACTCGAGAGACTCCCTTTTAGACA
GAAGCCCTGGTGCAGAGCTGCTTGAAGTAAGCTGAGGCGCTGACAGT
TCTACAGCCAGTTACAGATGGGTGCTCAGCTCAGAGAGAGGGGTGG
TGACTCCCTTAGGAACACAGCACTAGAGTGTGCTCCCTTAAAGACAGAC
CGAGTCTGTGACTGTGACTTGCAGACAGCTCGGGTGGTGTGGTGGT
ATTCTTAAATGTCATGTCACTGGCTTTTACGCTGGGAGCCACACAGG
TACCCTTGGCACCGCCAAACCTTGGCCCTTGGGGATTCCCATGTGCG
AGTCACTCTGTGCTTACCTTACAGGCGCTAGACTCGAGGCTTCTCT
TTTGGCCCTTCCCTGGCCAGGAGCTTGGACTGGGCTCGTGCTCATCG
AAGCGGGGAAGCTGCGAGGCCCACTCTGTGGGCTCTCTATTCTCTGG
AGTACGGAAGTTAGAGGCTGGGTGGCAGAGGAAGGCGAGGCGAGCG
GGCACTGGGCACTTCCCGCACTCTCAAGCGGTGTC
AAGTGAGCTGTGCTGGTTACAGTGGCTTGGGAGCTCAGAGAGGTTGAG
ACATAGGCTGGGCTCACACAGCCAGGTAACAGCAAGGTGGGGTTGGAGTC
AGGGTCTAGGGTGGCAGCTGCCAAGCTGTGCAACAAAGCTGTTTCTCG
GGAGGCTGAGGACCAACACCACTTCCCACTCAGGCTGAGCTGGAGATT
CAGAAAGAAGCCTGGAGCCAGGACAGAGGTGGTGGCTGAGTGAITCT
GCTGGCACTGTGAGTGAAGGCTTCAAGGCTTCAAGGCTGAGGCT
AGGCTCTGGTGGAGTGGACAGGACTCGCTGTGTGACATGGGATCGAG
CTTACTGTTGTCCAGAGGTGCTGTGTGGCCAGGCGACACTTCTCTCTC
CCATGCTTCTCCCTCCCAACCCAGGGGCTGGCCTGGAGCACTGCTCT
CTGACGCCAGGCAACTGGGGACCTCACTTCCATCCCAAGGAACCAT
GAACTGCTGCTGTGAGCTGCTGGGCGCTGCTGAGGCTGAGGTCTGGAAGT
CGGTGAGGCTGTGTGAGCTGAGCTCGCTTAAAGGAGGAGAGCTGGCA
CTGTACCTTCTCTCTCTCTCTGCTGAGTATGATGACAGAGGCTCC
AGCCAACTATCCAGCTGGATCCAGGGAAATATCAGCTTGGGCACT
GCAGTGACAGGGGCGCGCTGCCACAGGGAACACATCTTCTGTG
GGTTACGCGCTTCTGGGGCTGGAAGTGCCAAAGCTGGGGCAAGCT
GTGTTTACGCCACTGAACCAATTACACACAGCGGAGAAAGCAGTAA
ACAGCTTTCACC
```

Dataset

Source

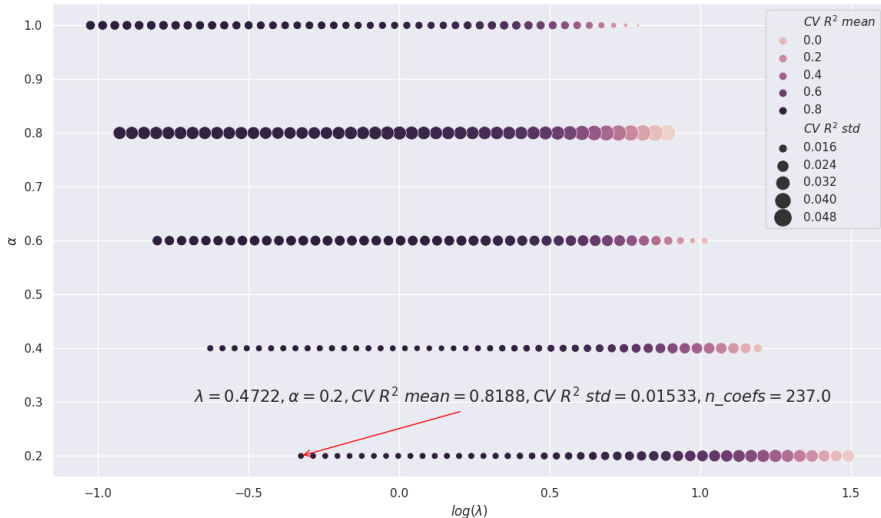
- Data from "Depression and suicide risk prediction models using blood-derived multi-omics data" [1]
- **Human blood**
- Reduced Representation Bisulfite Sequencing (RRBS)

Details

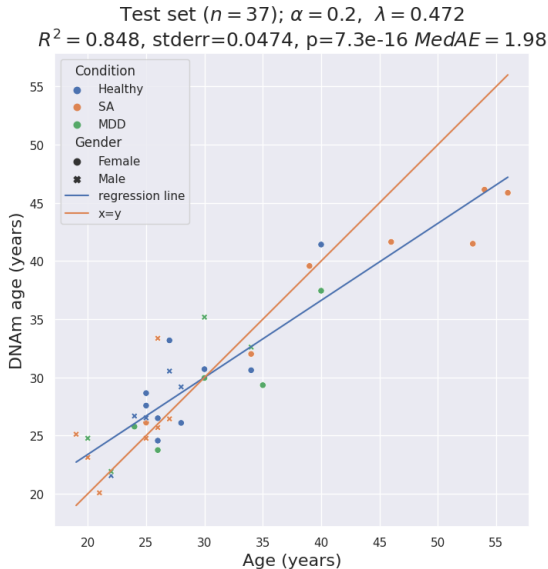
- sample N=182
- conditions:
 - Healthy/Control=87
 - Major Depressive Disorder (MDD)=39
 - Suicide Attempters (SA)=56
- features=8 722 096
- data for missing CpG sites was filled with **mean**

Hyperparameter Optimization (ElasticNet)

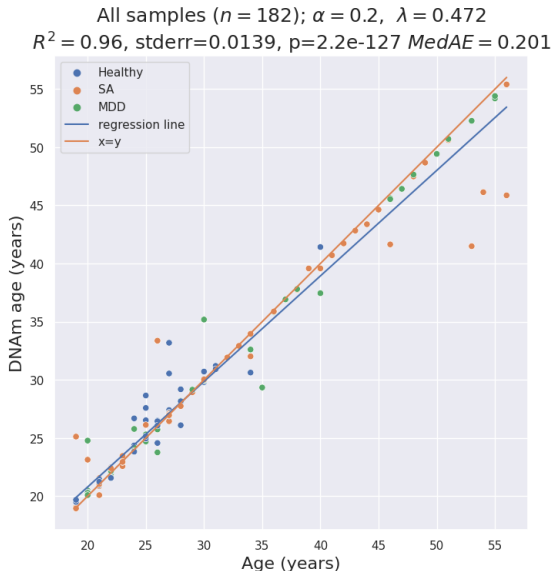
Hyperparameter optimization result on training data
best is selected based on $(CV R^2 \text{ mean}) - 0.5 \cdot (CV R^2 \text{ std})$



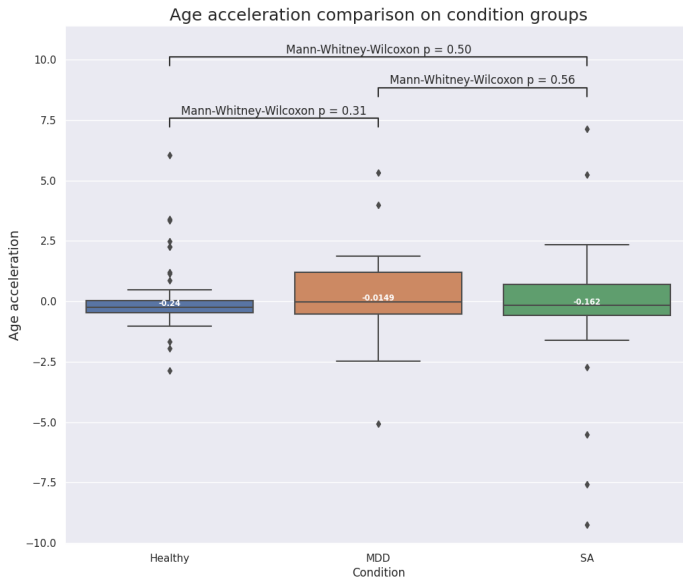
Results on Test Set



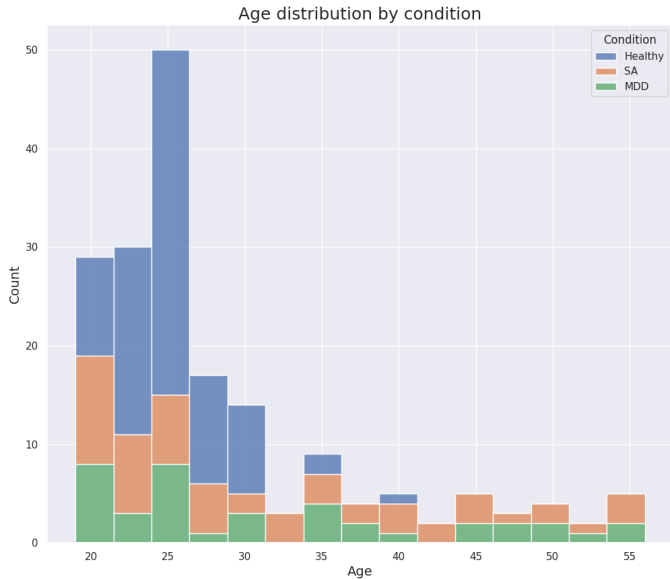
Results for All Instances



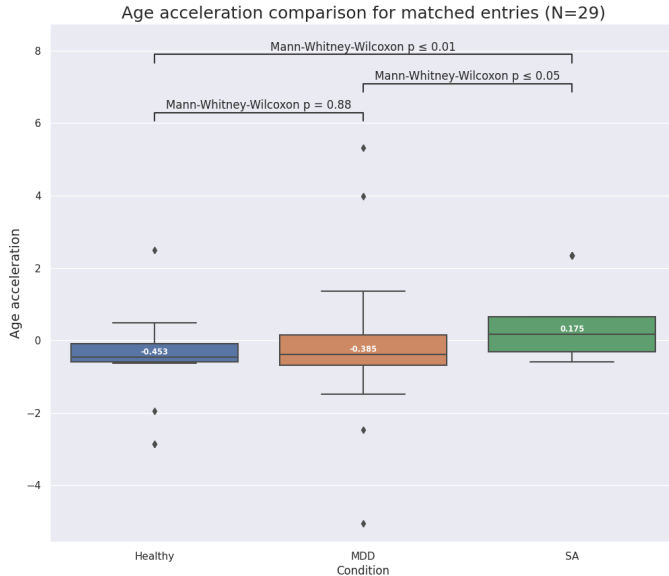
Age acceleration comparison by condition



Age distribution by condition



Age acceleration comparison by condition (controlled)



References

- [1] Youngjune Bhak et al. “Depression and suicide risk prediction models using blood-derived multi-omics data”. en. In: *Transl. Psychiatry* 9.1 (Oct. 2019), p. 262.
- [2] *Epigenetic clock assignment 1*.
https://github.com/szazo/epigenetic-clock/blob/main/assignment1_microarray.ipynb. 2024.
- [3] *Epigenetic clock assignment 2*.
https://github.com/szazo/epigenetic-clock/blob/main/assignment2_rrbs.ipynb. 2024.
- [4] *Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates*. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE40279>. [Accessed 13-05-2024]. 2012.