Chronological age is an imperfect measure of biological aging, it varies among individuals depending on their genetic inheritance but also on their environmental risk exposures to smoking, alcohol, diet, socio-economic factors or stress. Functional declines happen in every organ and tissue, long before one is diagnosed with a disease. There exists a tipping point separating aging gracefully and the onset of age-related diseases, which can be measured with valid and robust estimates of “biological age”. These epigenetic clocks can help to identify at-risk populations and provide clinical checkpoints for assessing therapeutic and behavioral interventions extending human lifespan. Similar to methylation, DNA hydroxymethylation of cells occurs at the C5 position in cytosine when it is immediately followed by a guanine. CpG islands are regions with a high frequency of CpG sites, and a strong correlation exist between age and DNA methylation or hydroxymethylation level in CpG sites. However, CpG hydroxymethylation has an unique function compared to CpG methylation, as it is particularly enriched in the brain and it is altered in AD.

**AIM**: Develop a DNA, CpG hydroxymethylation (5-hmC) based, epigenetic biological clock of the brain (DNAh) which could be used to monitor older people and help to detect neurological diseases related to aging.

5-Methylcytosine (5-mC) and 5-Hydroxymethylcytosine (5-hmC) are major modifications to the cytosine base in the DNA, known to be correlated with gene expression. 5-hmC is an oxidative derivative of 5-mC generated in a Ten-Eleven Translocation (TET) oxidase family mediated reaction. The role of 5-mC in transcriptional regulation is well understood, while the function of 5-hmC remains under investigation. 5-hmC is the intermediate step leading to demethylation of the cytosine.

We will construct the epigenetic clock estimating the age of the brain in two-steps: 1) we will identify surrogate biomarkers of physiological risk factors and stress factors associated with aging in a cohort of individuals and construct a phenotypic age score and 2) we will combine CpG sites of genomic data to predict the phenotypic age score.

**Research Model and Plan**

Phenotypic Features

We will identify a cohort of individuals with a specific mean age (65 years old for example), including both male and female, smokers, and various demographic and ethnicity characteristic for a genome-wide DNA methylation study. For the phenotypic features, we can collect blood samples and regress time-to-death with plasma proteins or measure cardio-vascular characteristics like hypertension, coronary heart disease, BMI, cholesterol, blood cell counts, leucocyte telomer length and regress time-to-death with them. We will compute for each identified covariate, its correlation coefficient and p-value, we will reject the ones with a p-value greater than a p-value determined using a control group (see table 1). We will then use a profiling array covering CpG islands, genes and enhancers to extract genomic features from this population[[1]](#footnote-1).

Genomic Features

Initial feature set will be the one used in ML models for DNA hydroxymethylation predictions from the research literature when applicable (see feature list). Machine learning models with more parameters tend to overfit, which leads to a decreased prediction power on unseen data points. One technique to address this issue is to perform recursive feature selection using a beam search algorithm: performing k-fold cross-validation followed by a selection of the top features with highest evaluation metric scores. DNA hydroxymethylation studies have reported that the most distinguishing features for CpG hydroxymethylation, are the active enhancer histones modifications H3K4me1 and H3K27ac, DNase, genomic derived features including CpG content, and Alu repeats (see feature table).

Now genome-wide studies represent a significant cost in time and financially, and Tab-Seq is a method that uses bisulfite conversion and Tet proteins to study 5hmC[[2]](#footnote-2). For these reasons, we may not recruit people for the study but use TAB-seq datasets from the NIH Roadmap Epigenome Consortium (Kundaje et al. 2015 for training and testing our DNA methylation predictive models. This dataset contains H1 human embryonic stem cell (GEO GSE36173) and H1-derived NPC (GEO GSM882245, GSM1463129) neural progenitor cells. TAB-Seq allows to estimate a C-to-U conversion rate (CCR) or methylation level for each cytosine in the genome – an estimator of degree of methylation (which will also be one of the features used for our models). 5-hmC is an intermediate molecular state in the demethylation pathway, and in TAB-seq the majority of CpG sites exhibit a unimodal distribution of CCRs peaking at 0.18. We could eliminate GpG sites with a low CCR (CCR is a number varying from 0: non-hydroxymethylated to 1: hydroxymethylated), i.e., less than 0.001.

**Step 1**: Using a phenotypic age estimator

We will use the “phenotypic age” estimator developed in “epigenetic biomarker of aging for lifespan and health span”.

The model was a Cox penalized regression model where the hazard of aging-related mortality (mortality from diseases of the heart, malignant neoplasms, chronic lower respiratory disease, cerebrovascular disease, Alzheimer’s disease, Diabetes mellitus, nephritis, nephrotic syndrome, and nephrosis) was regressed on forty-two clinical markers and chronological age. 10 variables (including chronological age) were selected for the phenotypic age predictor.

These nine biomarkers and chronological age were then included in a parametric proportional hazards model based on the Gompertz distribution. The Gompertz regression is a parametrized proportional hazards model which has been extensively used for modeling mortality data. Based on this model, the 10-year (120 months) estimation of mortality risk of jth individual is:

MortalityScorej = CDF(120, xj) = 1 – exp(-

where xb is the linear combination of biomarkers from the fitted model (the phenotypic features set we will have identified)

Next, the phenotypic age score is then computed as:

PhenotypicAgej = 141.50225 +

**Step 2**: develop an epigenetic aging clock of the brain by regressing the phenotypic age (from step 1) on TET-assisted TAB-seq protocol.

We will use various machine learning algorithm to predict the phenotypic age with genomic features of 5-hmC in selected CpG sites using a CCR threshold. We will start with a simple regression model, in which we will model phenotypic age score (obtained above) as a linear combination of the genetic features mentioned in the genomic features section:

PhenotypicAgej = α0 +α1 x feature1 + α2 x feature2 + … + αn x featuren

n: number of top critical features selected by the beam search algorithm

To assess the performances of each algorithm, we will compute various scores such as: regression coefficient, regression coefficient R^2, p-values, Mean Square Error, Mean Absolute Error, Mean Relative Error and plot these various metrics.

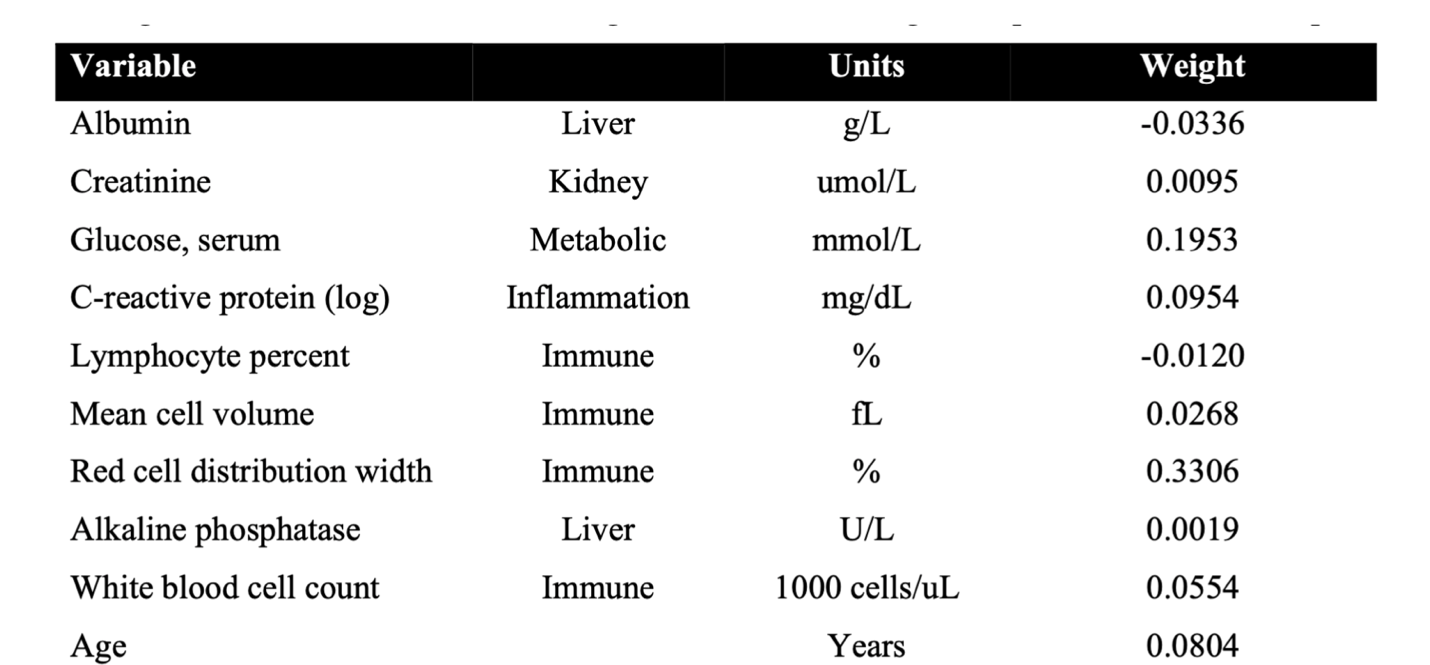
**Validation of the DNAh score**

We are planning to compute a brain biological age, DNAh, for five independent large-scale samples: 1) two samples from Women’s Health Initiative (WHI), the Framingham Heart Study (FHS), the Normative Aging Study (NAS) and the Jackson Heart Study (JHS). We will select a study if it can provide all the genomic features we identified during the training of our models. We will evaluate how our DNAh correlates with chronological age and we will compare our score with other approaches (see table T1).

**Conclusion**

There are been numerous DNA methylation clocks created in recent years, and as indicated in table T1, many models have been designed to predict DNA hydroxymethylation in the cells. By combining research performed in such biological clocks and existing models, this study will allow to create a novel brain epigenetic clock which could be a critical tool in assessing individuals at risk of neuronal aged related diseases (Alzheimer’s disease, Parkinson’s disease).

1. Measuring biological aging in humans: A quest <https://onlinelibrary.wiley.com/doi/10.1111/acel.13080>
2. Albert T. Higgins-Chen et al., “Aging biomarkers and the brain”, Seminars in Cell and Developmental Biology 116 (2021) 180–193
3. Morgan E. Levine et al., “An epigenetic biomarker of aging for lifespan and healthspan”, AGING 2018, Vol. 10, No. 4
4. Milos Pavlovi et al., “DIRECTION: a machine learning framework for predicting and characterizing DNA methylation and hydroxymethylation in mammalian genomes”, Bioinformatics, 33(19), 2017, 2986–2994, doi:10.1093/bioinformatics/btx316
5. Dan Jian et al., “DNA hydroxymethylation combined with carotid plaques as a novel biomarker for coronary atherosclerosis”, AGING 2019, Vol. 11, No. 10
6. Ake T. Lu et al., “DNA methylation GrimAge strongly predicts lifespan and healthspan”, AGING 2019, Vol. 11, No. 2
7. Christopher G. Bell et al., “DNA methylation aging clocks: challenges and recommendations”, Bell et al. Genome Biology, <https://doi.org/10.1186/s13059-019-1824-y>
8. Clemens Wrzode et al., “Linking the Epigenome to the Genome: Correlation of Different Features to DNA Methylation of CpG Islands”, doi:10.1371/journal.pone.0035327
9. Satyanarayan Rao et al., Systematic prediction of DNA shape changes due to CpG methylation explains epigenetic effects on protein–DNA binding, https://doi.org/10.1186/s13072-018-0174-4

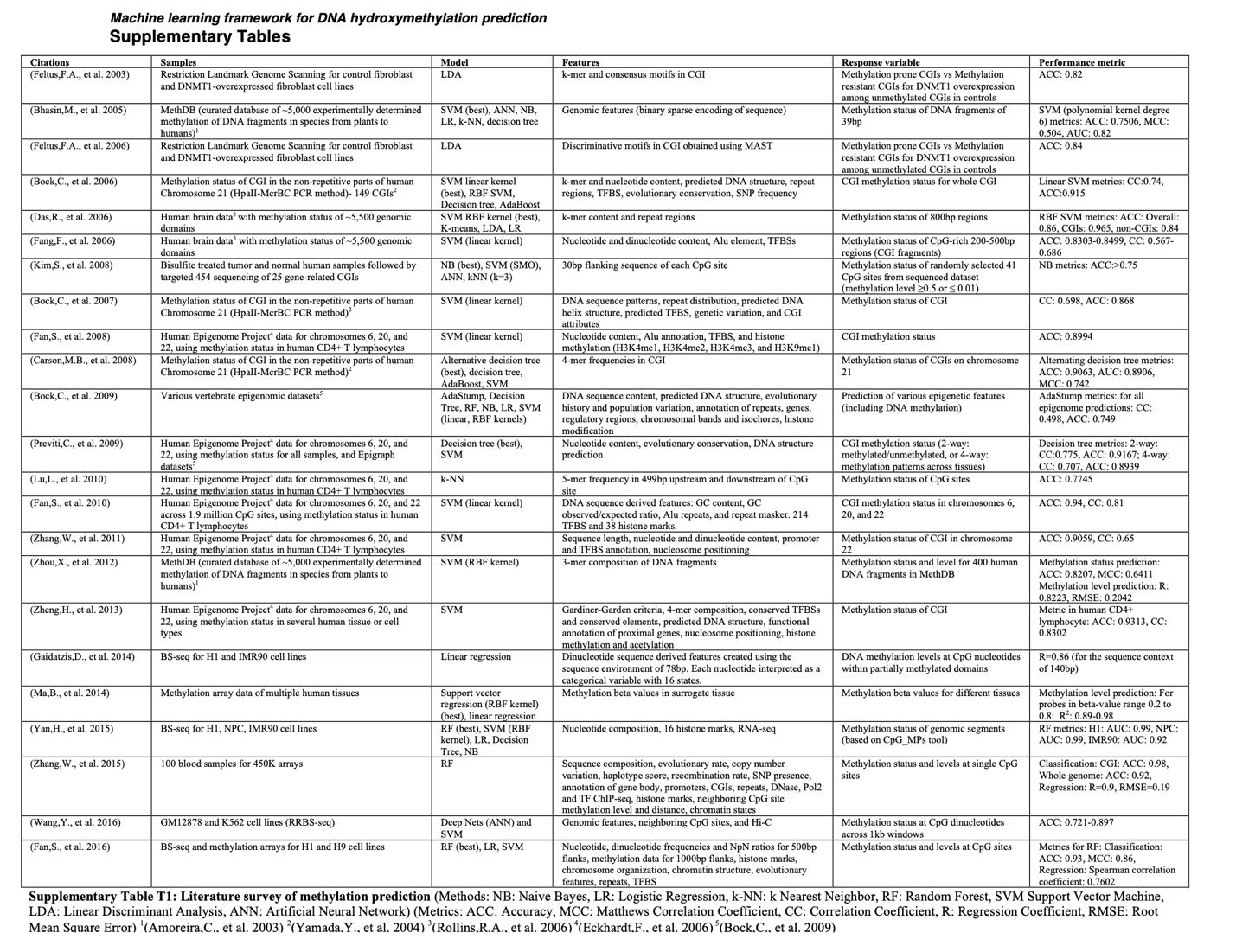
**Table 1: Phenotypic aging variables**

**5-hmc features used for model prediction**

* Distances to transcription start sites (4 features)
* CpG island-specific attributes (7 features)
* Genomic attributes (11 features)
* Repeat, Alu-Y and DNA/DNA alignment features (19 features)
* Single nucleotide polymorphism
* Periodic CpG distances (8 features)
* Closest CpGs (6 features)
* Sequence - dinucleotides (16 features)
* Sequence - tetranucleotides (257 features)
* CpG flanking sequence (4 features)
* DNA structure (43 features)
* Evolutionary conservation (4 features)
* Histone modification data (92 features)

**Top most discriminative features for hydroxymethylation CpG sites**





1. For example, the Illumina Infinium Methylation EPIC array or TAB-Seq from the same company see below. [↑](#footnote-ref-1)
2. “In Tab-Seq protocol, 5hmC is first protected with a glucose moiety that allows selective interaction and subsequent oxidation of 5mC with the Tet proteins. The oxidized genomic DNA is then treated with bisulfite, where 5hmC remains unchanged and is read as a cytosine, while 5mC and unmethylated cytosines are deaminated to uracil and read as thymidine upon sequencing. Deep sequencing of TAB-treated DNA compared with untreated DNA provides accurate representation of 5hmC localization in the genome.” From https://www.illumina.com/science/sequencing-method-explorer/kits-and-arrays [↑](#footnote-ref-2)