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 unique function compared to CpG methylation, as it is particularly enriched in the brain and 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, 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, smoker, 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).

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 lead to a reduce 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 sets 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).

Tab-Seq which is a method that uses bisulfite conversion and Tet proteins to study 5hmC[1]. We will use TAB-seq datasets from the NIH Roadmap Epigenome Consortium (Kundaje et al. 2015) could be used 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.

[1] "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

We will train our models on H1 cells and test their performances on NPC and vice-versa.

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 j^{th} individual is:

$$\text{MortalityScore}_j = \text{CDF}(120, x_j) = 1 - \exp(-e^{\frac{xbj(exp(120*\gamma)-1)}{\gamma}})$$

where xb is the linear combination of biomarkers from the fitted model (Table 1)

Next, the phenotypic age score is then computed as:

$$\text{PhenotypicAge}_j = 141.50225 + \frac{ln(-0.00553 * ln(1 - MoralityScorej)))}{0.090165}$$

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:

$$\text{PhenotypicAge}_j = \alpha_0 + \alpha_1 \text{ x feature}_1 + \alpha_2 \text{ x feature}_2 + \ldots + \alpha_n \text{ x feature}_n$$

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, 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 (WHI0, 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).

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

Variable		Units	Weight
Albumin	Liver	g/L	-0.0336
Creatinine	Kidney	umol/L	0.0095
Glucose, serum	Metabolic	mmol/L	0.1953
C-reactive protein (log)	Inflammation	mg/dL	0.0954
Lymphocyte percent	Immune	%	-0.0120
Mean cell volume	Immune	fL	0.0268
Red cell distribution width	Immune	%	0.3306
Alkaline phosphatase	Liver	U/L	0.0019
White blood cell count	Immune	1000 cells/uL	0.0554
Age		Years	0.0804

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

Alu_repeat
Bp_to_CGI
CG_sat_50bp
CpG_sat_50bp
CpG_to_CGI
DNase
G_sat_50bp
H2AK5ac
H3K27ac
H3K27me3
H3K36me3
H3K4me1
H3K4me3
H3K79me1
H3K9ac
H3K9me3
Histone_states
Repeats
BS-seq_CCR
CpG_Island

Supplementary Tables

Citations	Samples	Model	Features	Response variable	Performance metric
(Feltus,F.A., et al. 2003)	Restriction Landmark Genome Scanning for control fibroblast and DNMT1-overexpressed fibroblast cell lines	LDA	k-mer and consensus motifs in CGI	Methylation prone CGIs vs Methylation resistant CGIs for DNMT1 overexpression among unmethylated CGIs in controls	ACC: 0.82
(Bhasin,M., et al. 2005)	MethDB (curated database of ~5,000 experimentally determined methylation of DNA fragments in species from plants to humans)[1]	SVM (best), ANN, NB, LR, k-NN, decision tree	Genomic features (binary sparse encoding of sequence)	Methylation status of DNA fragments of 39bp	SVM (polynomial kernel degree 6) metrics: ACC: 0.7506, MCC: 0.504, AUC: 0.82
(Feltus,F.A., et al. 2006)	Restriction Landmark Genome Scanning for control fibroblast and DNMT1-overexpressed fibroblast cell lines	LDA	Discriminative motifs in CGI obtained using MAST	Methylation prone CGIs vs Methylation resistant CGIs for DNMT1 overexpression among unmethylated CGIs in controls	ACC: 0.84
(Bock,C., et al. 2006)	Methylation status of CGI in the non-repetitive parts of human Chromosome 21 (HpaII-McrBC PCR method)- 149 CGIs[2]	SVM linear kernel (best), RBF SVM, Decision tree, AdaBoost	k-mer and nucleotide content, predicted DNA structure, repeat regions, TFBS, evolutionary conservation, SNP frequency	CGI methylation status for whole CGI	Linear SVM metrics: CC:0.74, ACC:0.915
(Das,R., et al. 2006)	Human brain data[3] with methylation status of ~5,500 genomic domains	SVM RBF kernel (best), K-means, LDA, LR	k-mer content and repeat regions	Methylation status of 800bp regions	RBF SVM metrics: ACC: Overall: 0.86, CGIs: 0.965, non-CGIs: 0.84
(Fang,F., et al. 2006)	Human brain data[3] with methylation status of ~5,500 genomic domains	SVM (linear kernel)	Nucleotide and dinucleotide content, Alu element, TFBSs	Methylation status of CpG-rich 200-500bp regions (CGI fragments)	ACC: 0.8303-0.8499, CC: 0.567-0.686
(Kim,S., et al. 2008)	Bisulfite treated tumor and normal human samples followed by targeted 454 sequencing of 25 gene-related CGIs	NB (best), SVM (SMO), ANN, kNN (k=3)	30bp flanking sequence of each CpG site	Methylation status of randomly selected 41 CpG sites from sequenced dataset (methylation level ≥ 0.5 or ≤ 0.01)	NB metrics: ACC:>0.75
(Bock,C., et al. 2007)	Methylation status of CGI in the non-repetitive parts of human Chromosome 21 (HpaII-McrBC PCR method)[2]	SVM (linear kernel)	DNA sequence patterns, repeat distribution, predicted DNA helix structure, predicted TFBS, genetic variation, and CGI attributes	Methylation status of CGI	CC: 0.698, ACC: 0.868
(Fan,S., et al. 2008)	Human Epigenome Project[4] data for chromosomes 6, 20, and 22, using methylation status in human CD4+ T lymphocytes	SVM (linear kernel)	Nucleotide content, Alu annotation, TFBS, and histone methylation (H3K4me1, H3K4me2, H3K4me3, and H3K9me1)	CGI methylation status	ACC: 0.8994
(Carson,M.B., et al. 2008)	Methylation status of CGI in the non-repetitive parts of human Chromosome 21 (HpaII-McrBC PCR method)[2]	Alternative decision tree (best), decision tree, AdaBoost, SVM	4-mer frequencies in CGI	Methylation status of CGIs on chromosome 21	Alternating decision tree metrics: ACC: 0.9063, AUC: 0.8906, MCC: 0.742
(Bock,C., et al. 2009)	Various vertebrate epigenomic datasets[5]	AdaStump, Decision Tree, RF, NB, LR, SVM (linear, RBF kernels)	DNA sequence content, predicted DNA structure, evolutionary history and population variation, annotation of repeats, genes, regulatory regions, chromosomal bands and isochores, histone modification	Prediction of various epigenetic features (including DNA methylation)	AdaStump metrics: for all epigenome predictions: CC: 0.498, ACC: 0.749
(Previti,C., et al. 2009)	Human Epigenome Project[4] data for chromosomes 6, 20, and 22, using methylation status for all samples, and Epigraph datasets[3]	Decision tree (best), SVM	Nucleotide content, evolutionary conservation, DNA structure prediction	CGI methylation status (2-way: methylated/unmethylated, or 4-way: methylation patterns across tissues)	Decision tree metrics: 2-way: CC:0.775, ACC: 0.9167; 4-way: CC: 0.707, ACC: 0.8939
(Lu,L., et al. 2010)	Human Epigenome Project[4] data for chromosomes 6, 20, and 22, using methylation status in human CD4+ T lymphocytes	k-NN	5-mer frequency in 499bp upstream and downstream of CpG site	Methylation status of CpG sites	ACC: 0.7745
(Fan,S., et al. 2010)	Human Epigenome Project[4] data for chromosomes 6, 20, and 22 across 1.9 million CpG sites, using methylation status in human CD4+ T lymphocytes	SVM (linear kernel)	DNA sequence derived features: GC content, GC observed/expected ratio, Alu repeats, and repeat masker. 214 TFBS and 38 histone marks.	CGI methylation status in chromosomes 6, 20, and 22	ACC: 0.94, CC: 0.81
(Zhang,W., et al. 2011)	Human Epigenome Project[4] data for chromosomes 6, 20, and 22, using methylation status in human CD4+ T lymphocytes	SVM	Sequence length, nucleotide and dinucleotide content, promoter and TFBS annotation, nucleosome positioning	Methylation status of CGI in chromosome 22	ACC: 0.9059, CC: 0.65
(Zhou,X., et al. 2012)	MethDB (curated database of ~5,000 experimentally determined methylation of DNA fragments in species from plants to humans)[1]	SVM (RBF kernel)	3-mer composition of DNA fragments	Methylation status and level for 400 human DNA fragments in MethDB	Methylation status prediction: ACC: 0.8207, MCC: 0.6411 Methylation level prediction: R: 0.8223, RMSE: 0.2042
(Zheng,H., et al. 2013)	Human Epigenome Project[4] data for chromosomes 6, 20, and 22, using methylation status in several human tissue or cell types	SVM	Gardiner-Garden criteria, 4-mer composition, conserved TFBSs and conserved elements, predicted DNA structure, functional annotation of proximal genes, nucleosome positioning, histone methylation and acetylation	Methylation status of CGI	Metric in human CD4+ lymphocyte: ACC: 0.9313, CC: 0.8302
(Gaidatzis,D., et al. 2014)	BS-seq for H1 and IMR90 cell lines	Linear regression	Dinucleotide sequence derived features created using the sequence environment of 78bp. Each nucleotide interpreted as a categorical variable with 16 states.	DNA methylation levels at CpG nucleotides within partially methylated domains	R=0.86 (for the sequence context of 140bp)
(Ma,B., et al. 2014)	Methylation array data of multiple human tissues	Support vector regression (RBF kernel) (best), linear regression	Methylation beta values in surrogate tissue	Methylation beta values for different tissues	Methylation level prediction: For probes in beta-value range 0.2 to 0.8: R^2: 0.89-0.98
(Yan,H., et al. 2015)	BS-seq for H1, NPC, IMR90 cell lines	RF (best), SVM (RBF kernel), LR, Decision Tree, NB	Nucleotide composition, 16 histone marks, RNA-seq	Methylation status of genomic segments (based on CpG_MPs tool)	RF metrics: H1: AUC: 0.99, NPC: AUC: 0.99, IMR90: AUC: 0.92
(Zhang,W., et al. 2015)	100 blood samples for 450K arrays	RF	Sequence composition, evolutionary rate, copy number variation, haplotype score, recombination rate, SNP presence, annotation of gene body, promoters, CGIs, repeats, DNase, Pol2 and TF ChIP-seq, histone marks, neighboring CpG site methylation level and distance, chromatin states	Methylation status and levels at single CpG sites	Classification: CGI: ACC: 0.98, Whole genome: ACC: 0.92, Regression: R=0.9, RMSE=0.19
(Wang,Y., et al. 2016)	GM12878 and K562 cell lines (RRBS-seq)	Deep Nets (ANN) and SVM	Genomic features, neighboring CpG sites, and Hi-C	Methylation status at CpG dinucleotides across 1kb windows	ACC: 0.721-0.897
(Fan,S., et al. 2016)	BS-seq and methylation arrays for H1 and H9 cell lines	RF (best), LR, SVM	Nucleotide, dinucleotide frequencies and NpN ratios for 500bp flanks, methylation data for 1000bp flanks, histone marks, chromosome organization, chromatin structure, evolutionary features, repeats, TFBS	Methylation status and levels at CpG sites	Metrics for RF: Classification: ACC: 0.93, MCC: 0.86, Regression: Spearman correlation coefficient: 0.7602

Supplementary Table T1: Literature survey of methylation prediction (Methods: NB: Naive Bayes, LR: Logistic Regression, k-NN: k Nearest Neighbor, RF: Random Forest, SVM Support Vector Machine, LDA: Linear Discriminant Analysis, ANN: Artificial Neural Network) (Metrics: ACC: Accuracy, MCC: Matthews Correlation Coefficient, CC: Correlation Coefficient, R: Regression Coefficient, RMSE: Root Mean Square Error) [1](Amoreira,C., et al. 2003) [2](Yamada,Y., et al. 2004) [3](Rollins,R.A., et al. 2006) [4](Eckhardt,F., et al. 2006) [5](Bock,C., et al. 2009)