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 is a tipping point separating aging gracefully and the onset of age-related diseases, by developing valid and robust estimates of "biological age". By identifying at-risk populations, we will be able to 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, it is particularly enriched in the brain and is altered in AD.

Biological question: Develop a DNA based on CpG hydroxymethylation (5-hmC) aging clock.

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.

Data
Tab-Seq is a method that uses bisulfite conversion and Tet proteins to study 5hmC[1]. TAB-seq datasets from the NIH Roadmap Epigenome consortium (Kundaje et al. 2015) will be used for training and testing our predictive models. This dataset contains H1 human embryonic stem cell (GEO GSE36173) and H1-derived NPC (GEO GSM882245, GSM1463129) neural progenitor cells. . AB-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 feature to train our model). We will eliminate GpG sites with a low CCR (CCR is a number varying from 0, non-hydroxymethylated to 1, hydroxymethylated)

Feature selection
Initial feature sets will be the ones used in previous ML models when available[2]. 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 to perform recursive feature selection using a beam search algorithm which using k-fold cross-validations will select the top n features sets with highest evaluation metric scores. It has been reported that the most distinguishing feature in similar studies, has been the active enhancer histones modifications H3K4me1 and H3K27ac, DNase, genomic derived features including CpG content, and Alu repeats (see table).

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:

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

where xb = the linear combination of biomarkers from the fitted model (Table 1)
Next, the phenotypicAge is then computed as:

[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 thymidines 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
[2] "Linking the Epigenome to the Genome: Correlation of Different Features to DNA Methylation of CpG Islands": paragraph: Feature

$$\text{PhenotypicAge}_j = 141.50225 + {\ln(-0.00553 * \ln(1 - \text{MoralityScore}j))}/{0.090165}$$

Step 2: develop an epigenetic biomarker of phenotypic age by regressing phenotypic age (from step 1) on TET-assisted BS-seq (TAB-seq) protocol.
We will use various machine learning algorithm to predict the phenotypic age with feature sets for 5-hmC related to CpG sites. We will start with a simple regression model, followed by Random Forrest Regression, XGBoost, neural network and compare their performance on validation and test datasets.

DeepH&M: Estimating single-CpG hydroxymethylation and methylation levels from enrichment and restriction enzyme sequencing methods https://www-ncbi-nlm-nih-gov.proxy1.library.jhu.edu/pmc/articles/PMC7458459/
https://www.ks.uiuc.edu/Research/methylation/

Measuring biological aging in humans: A quest https://onlinelibrary.wiley.com/doi/10.1111/acel.13080

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
Constant			-19.9067
Gamma			0.0077

Table 1

- 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)

(B) Random Forest OFS for NPC 5-hmC status prediction	
	Features
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	