Final Presentation

Merck - Healthy Aging

Team members: Dan Cox, Yaxin Lei, Eleonora Shantsila , Aaron Jacobson

Problem Description

We are studying one of the most inevitable natural processes — **aging**. In the health industry, this has almost always been associated with diseases.

At the outset of this project we defined the following goals.

1. To identify databases that are relevant to aging;
2. To identify features that are good biomarkers of healthy aging;
3. To build age-predictive models based on the biomarkers we identify.
4. To compare the aging process between healthy and unhealthy cohorts, as well as biomarkers from different tissues.

Break our topic up to HEALTHY + AGING

Three step breakdown of our goal

01 Identifying Datasets

- ✔ What datasets are relevant to age prediction?
- ✔ Databases: PPMI (Parkinson), EWAS (Methylation)
- ✔ Different dataset types: MRI, Blood test, Methylation

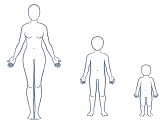
EWAS Data Hub

02 Age Predictive Modeling

- ✔ Feature selection
- ✔ Produce an accurate model for age prediction within the healthy cohort.
- ✔ Model refinement and model selection.

03 Comparative Model Analysis

- ✔ Do different age cohorts and unhealthy/unhealthy cohorts behave differently, and if so how?
- ✔ Comparing results sampled from different tissues.



Identifying Datasets

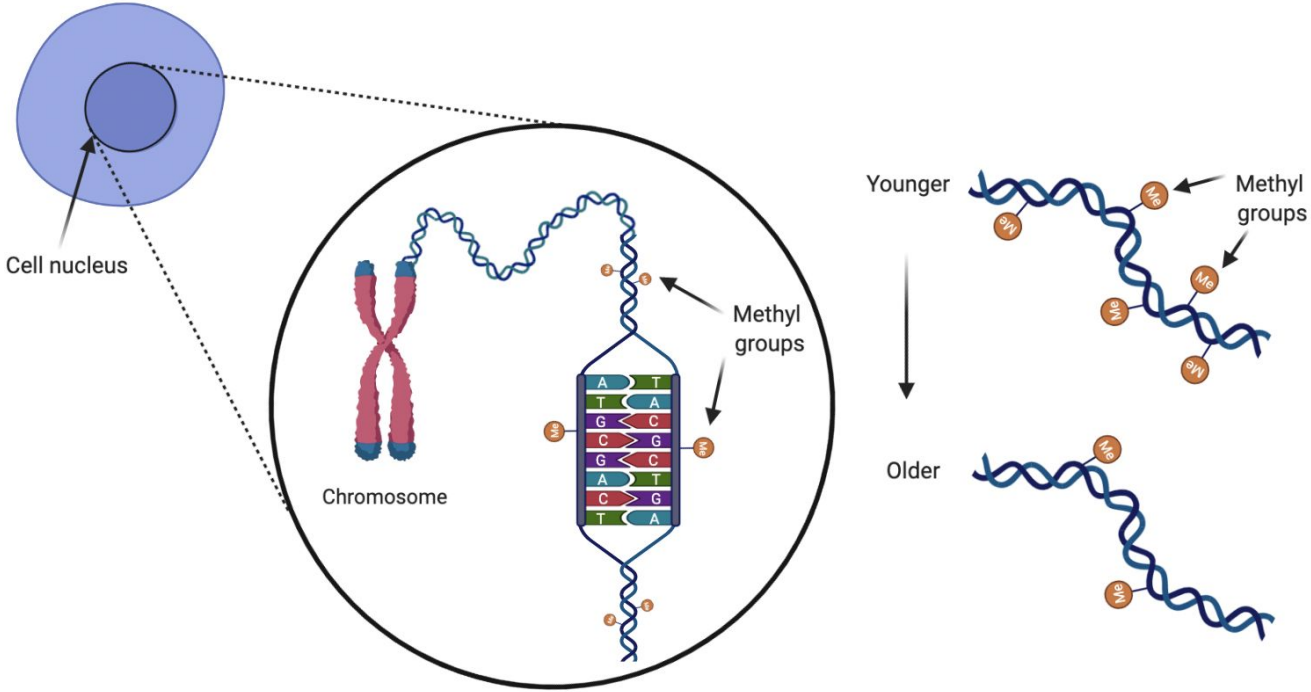
1. Dataset exploration;
2. Literature review.

1

Literature review

1	MRI	→ Available data would involve very high volumes of processing
2	Blood chemistry	→ Conducted an EDA using data from PPMI database → Results showed little association between blood chemistry and age
3	DNA methylation	→ Also conducted an EDA using data from PPMI database → Results were promising

DNA Methylation



Database overview

EWAS Data Hub

A data hub of DNA methylation array data and metadata

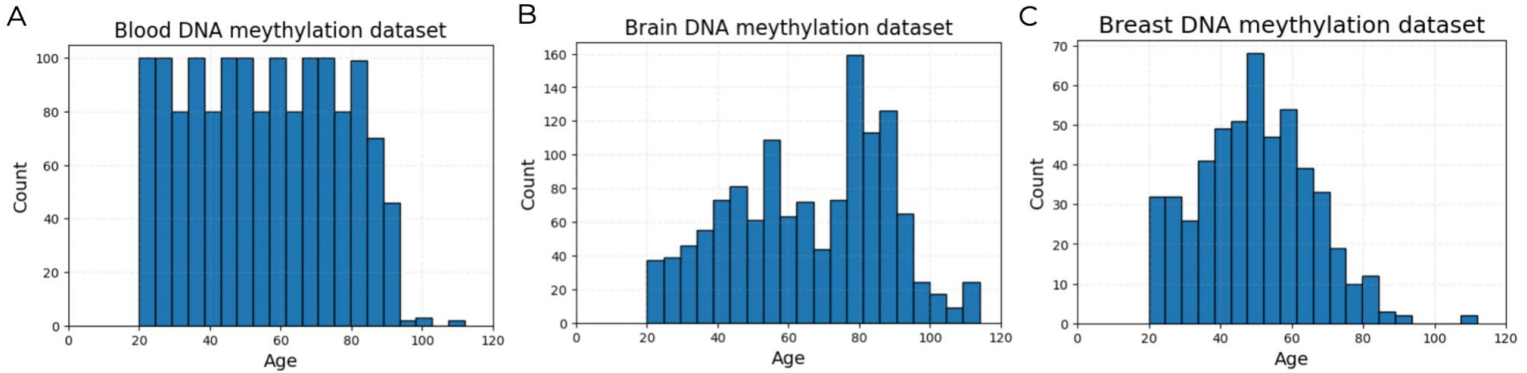
95,783
Samples

626
Tissues/cells

431
Diseases

	Healthy Control	Alzheimer's	Parkinson's	Huntington's
Whole Blood	1802	111	222	N/A
Brain	1064	811	N/A	270
Breast	520	N/A	N/A	N/A

Database overview



sample_id	tissue	age	cg02494853	cg03706273	...	cg04023335
GSM2334366	whole blood	94	0.078	0.205	...	0.139
GSM989863	whole blood	101	0.013	0.008	...	0.117
:	:	:	:	:		:
GSM1871289	whole blood	48	0.019	0.024	...	0.166

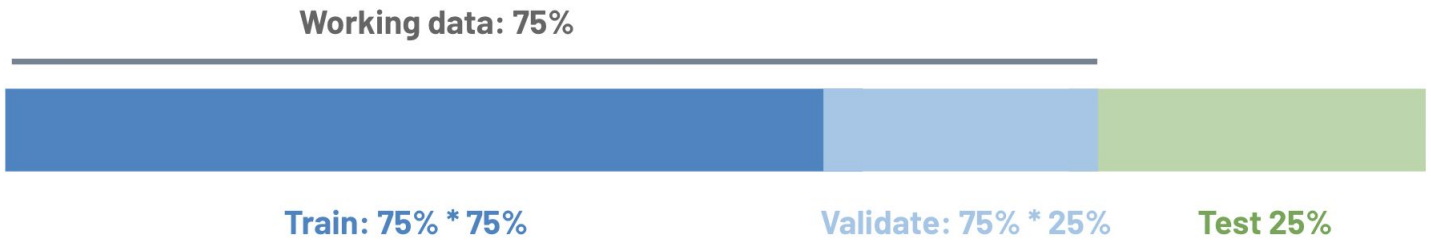
1066 x 375,603

Age Prediction Modeling

1. Standard procedures
2. Feature selection
3. Healthy cohort age predictive modeling
4. Model and feature transferability

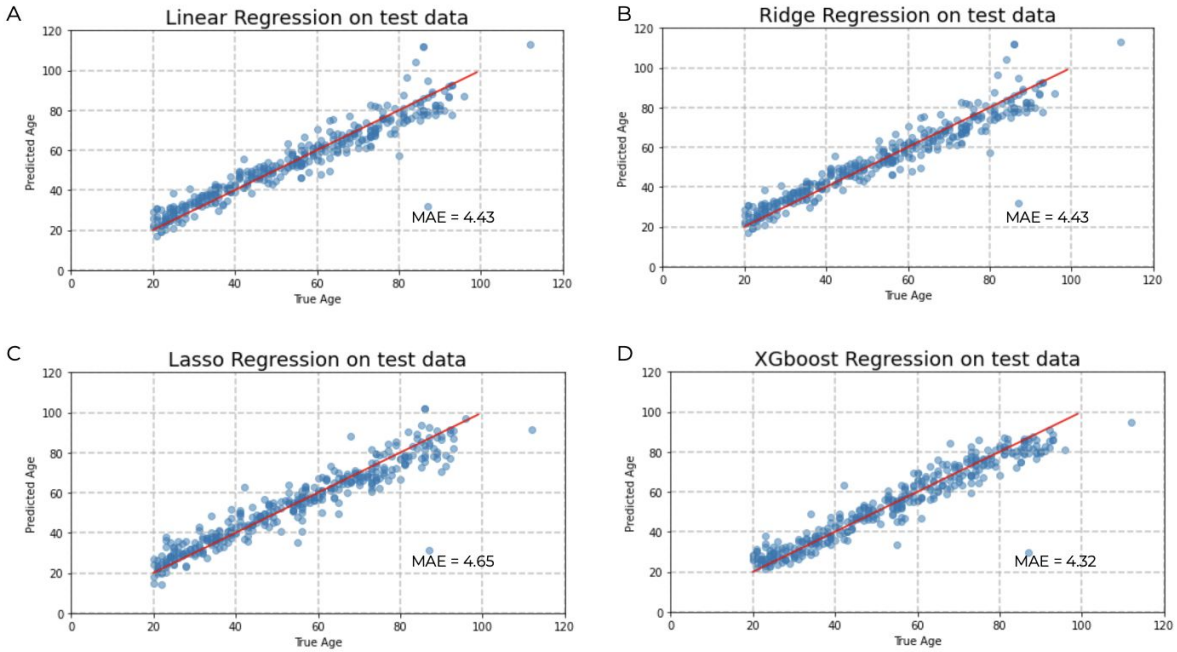
2

Standard Procedure



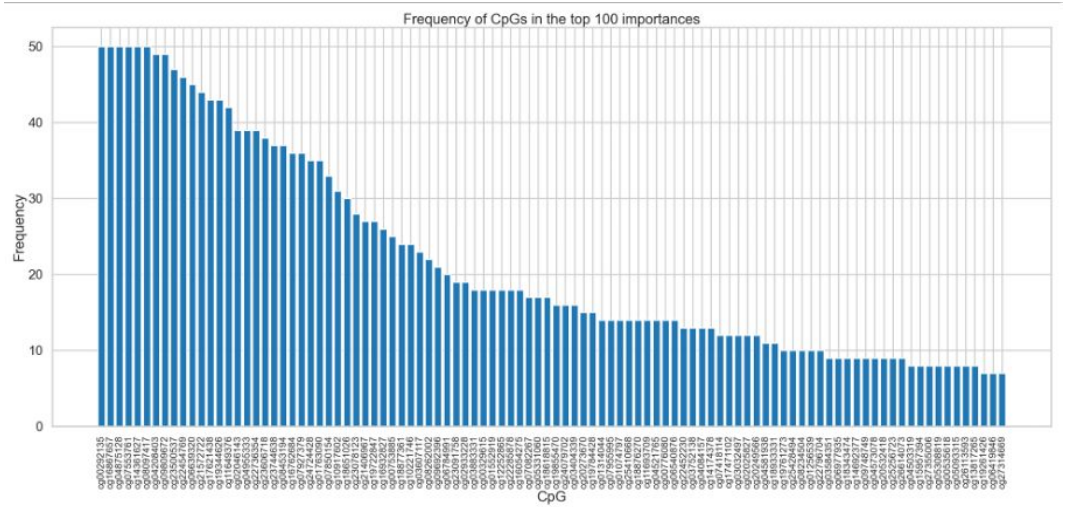
- Dropped all columns with greater than 10% of NAs
- Removed young individuals who are under 20 years of age
- Standardized NA imputation by computing the column mean
- Standardized train, validate, test splits

Age Prediction – Whole Blood Tissue:
Modeling with all features, **375,603**

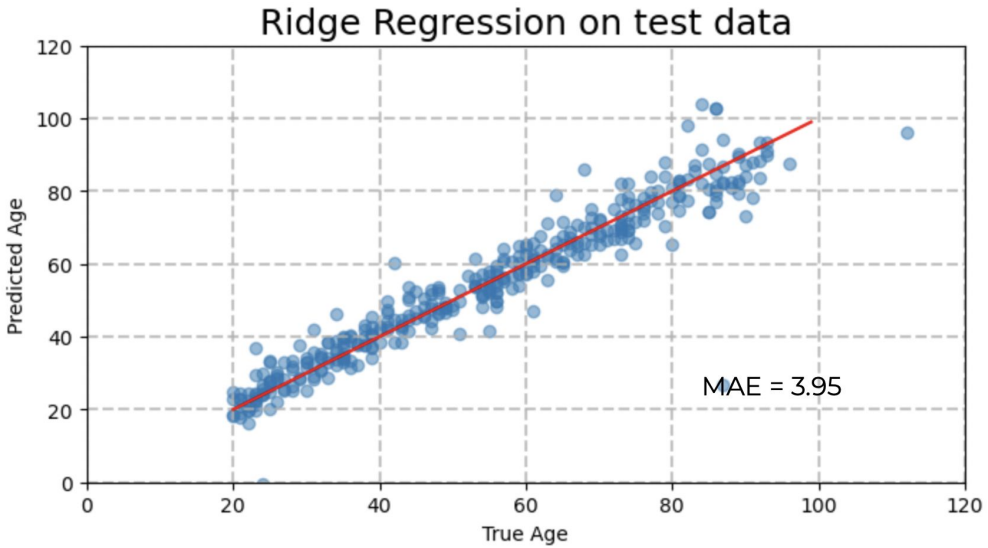


Feature selection:
Using XGboost importance scores

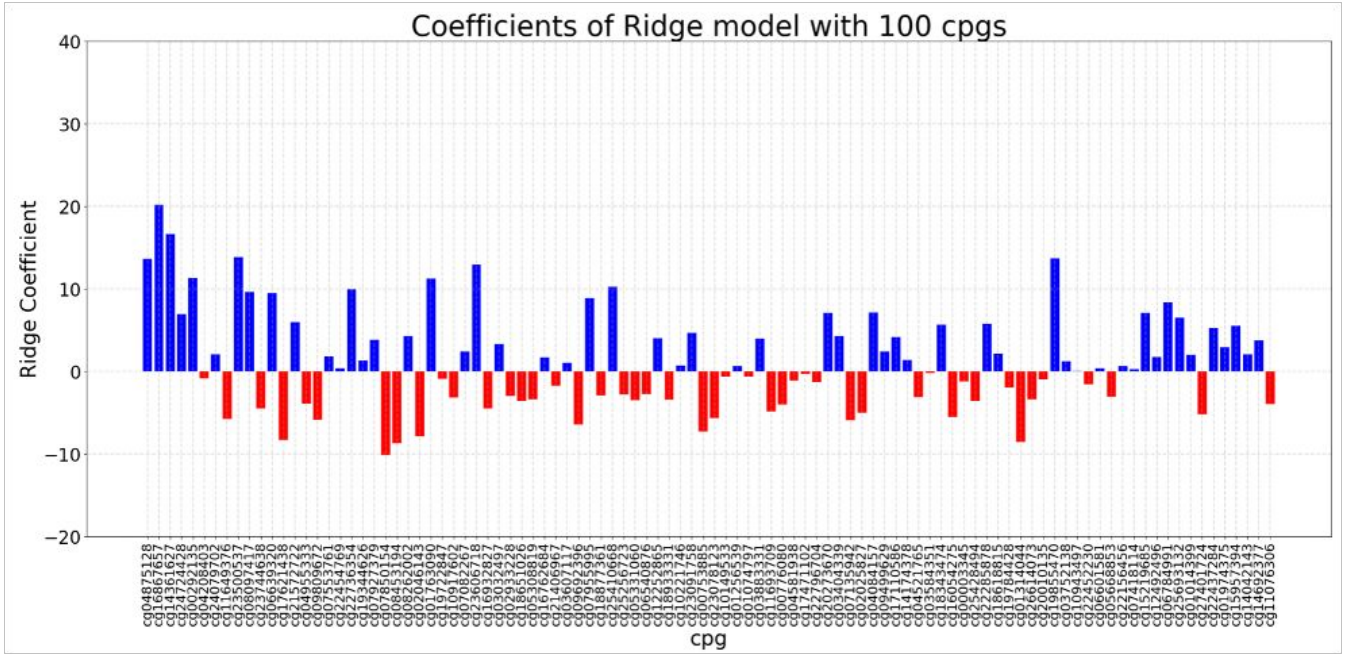
- Optimize XGboost model with all features
- Cycle for 50 cycles
  - Randomly split training data 70/30
  - Fit data with an XGboost model
  - Record importance scores
- Determine which features most often occur in the top 100 importance scores
- Select the features that appear most often



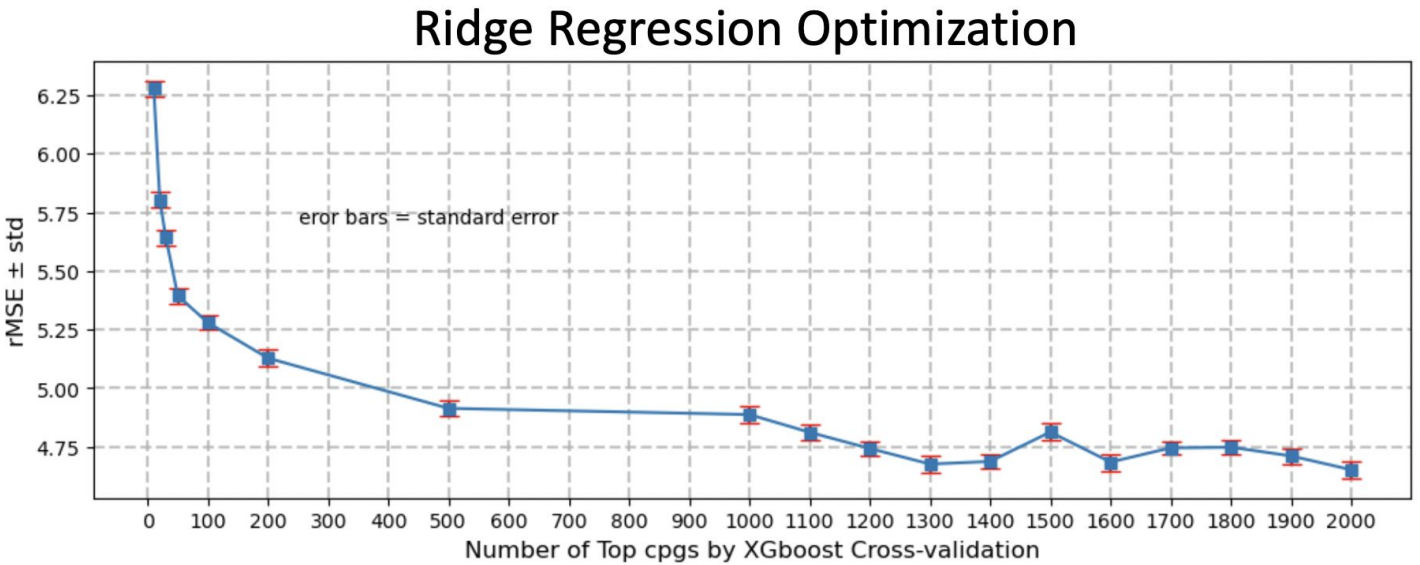
Feature Selection:
Blood Model Result: Top 100 cpgs



Feature Selection:
Blood Model Result: Top 100 cpgs

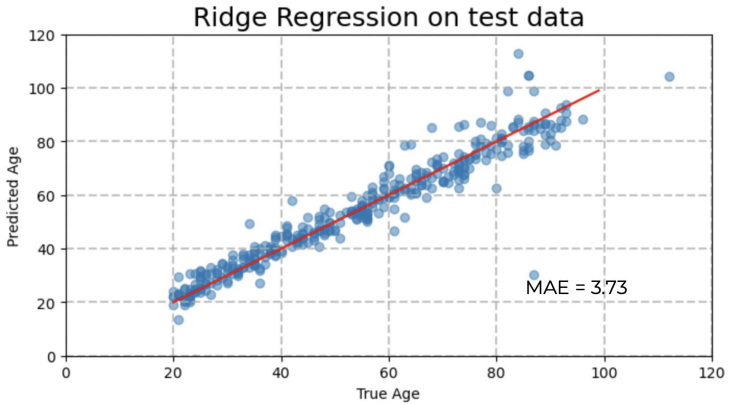


Feature Selection:
Blood model, optimizing the number of features

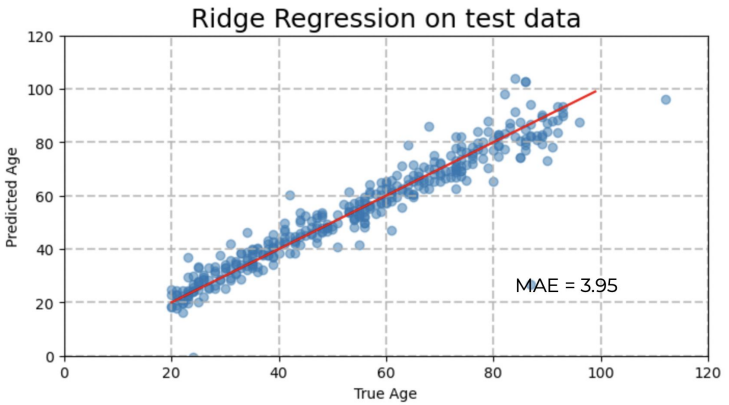


Feature Selection:
Blood Model Result: Top 1000 cpgs

1000 CpGs

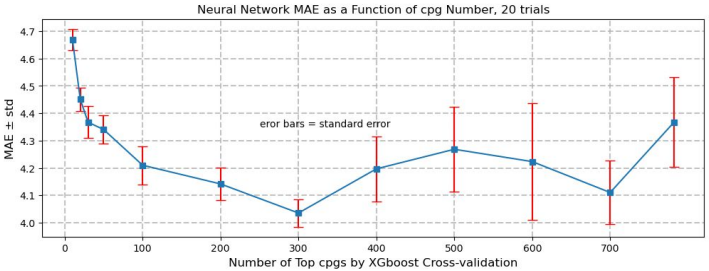


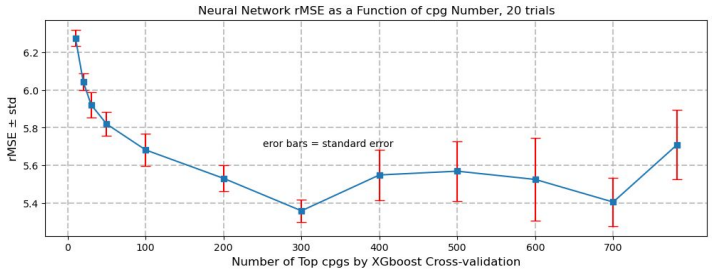
100 CpGs



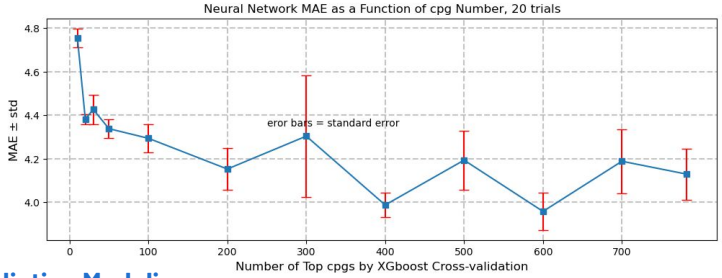
Feature Selection:
Blood Model Result - Overall

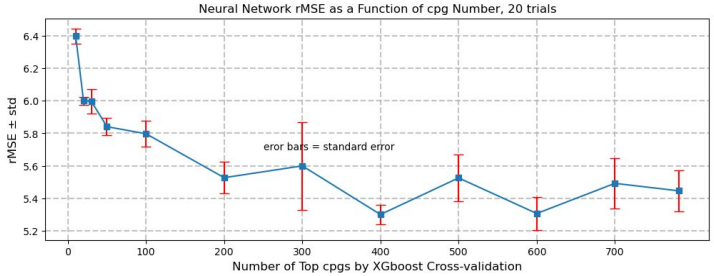
Whole Blood Neural network: 3 layers (hidden layer node number 128->56>28)



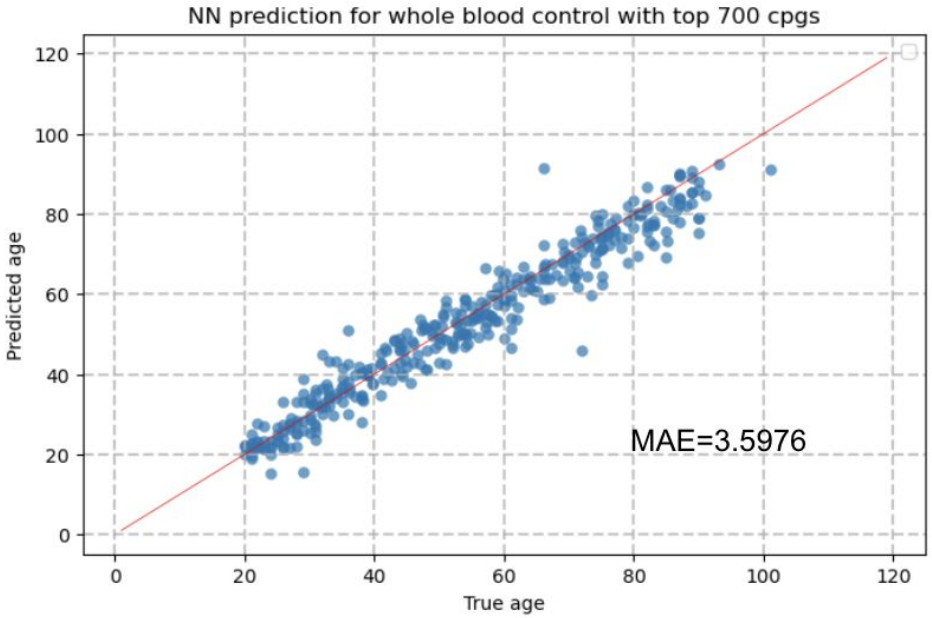


Whole Blood Neural network: 2 layers (hidden layer node number 128->56).





Feature Selection:
Blood Model Result - Overall



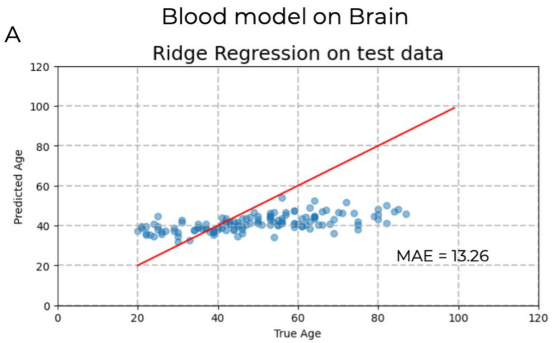
MAE of 3.597 years
2 hidden layer neural networks (hidden layer node number 128->64)

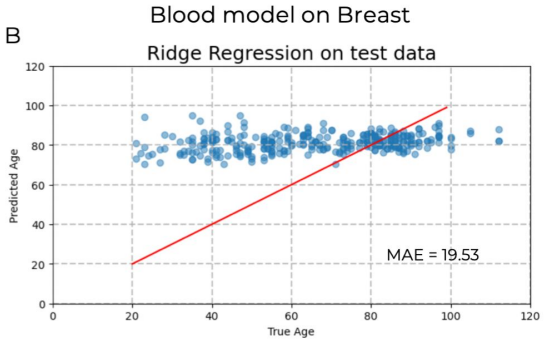
Feature Selection:
Blood Model Result - Overall

Model	MSE	rMSE	MAE	r^2	Corr
1000 cpgs					
Linear	88.680	9.417	6.669	0.803	0.912
Ridge	36.610	6.051	3.733	0.918	0.959
Lasso	36.830	6.609	3.866	0.918	0.958
XGboost	34.580	5.880	4.109	0.923	0.961
100 cpgs					
Linear	41.635	6.453	4.224	0.907	0.953
Ridge	37.580	6.130	3.950	0.916	0.957
Lasso	37.510	6.125	3.881	0.916	0.957
Xgboost	35.380	5.948	4.126	0.921	0.960
700 cpgs					
Neural Net	23.470	4.841	3.597	—	—

Are models transferable across different tissues?

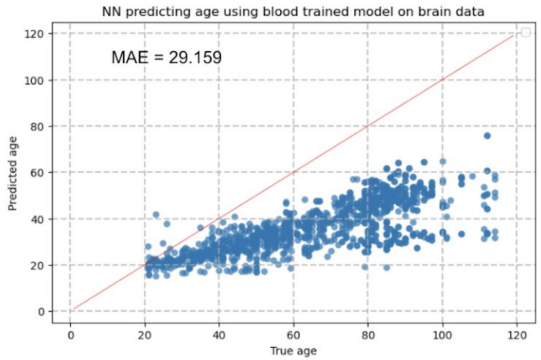
Ridge Regression

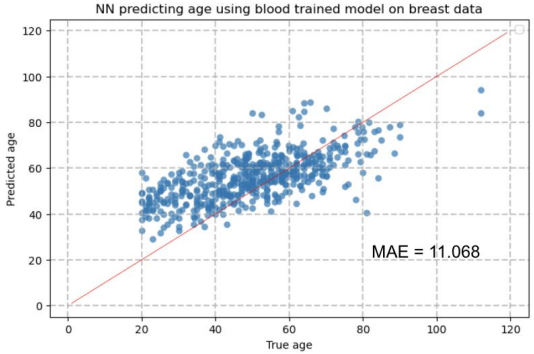




We observe a general **underprediction** when the blood-fitted ridge regression and neural network models are applied to methylation data from brain tissue.

Neural Network

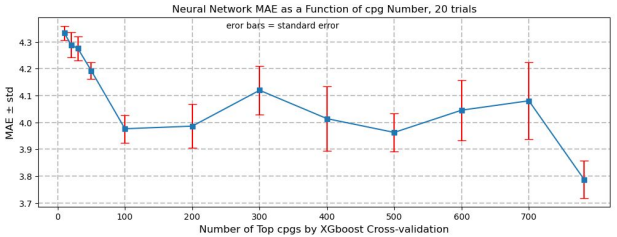


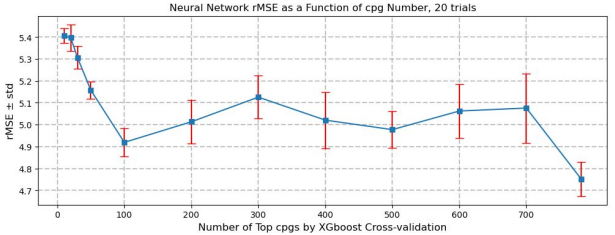


And an **overprediction** when they are applied to breast data although not as severe for the neural network compared to that of the ridge regression.

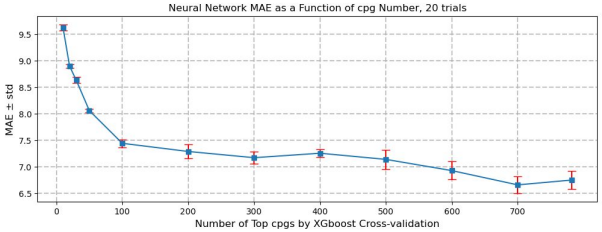
Are features transferable across different tissues?

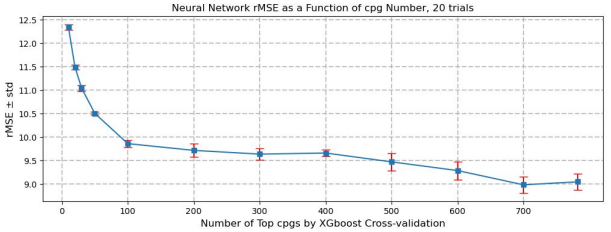
While transferring the blood-based models without modification to other tissues did not work well, we **investigated whether the top-ranked features generated by XGBoost cross validation using methylation data from blood are valuable at all for predicting age with methylation data from other tissues**



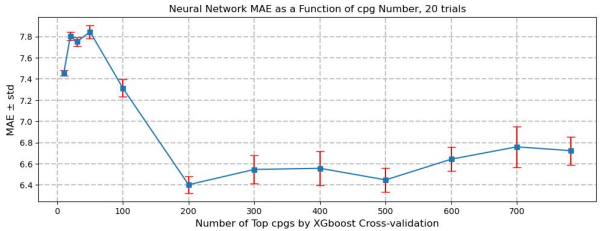


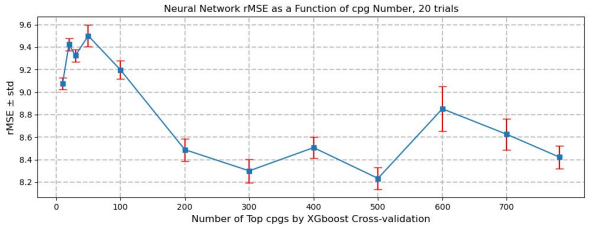
Leukocyte





Brain

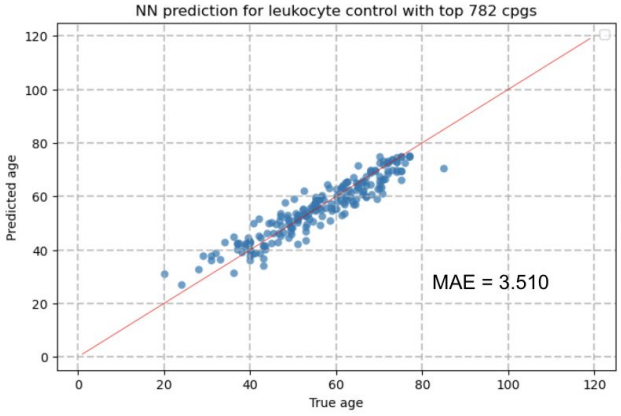




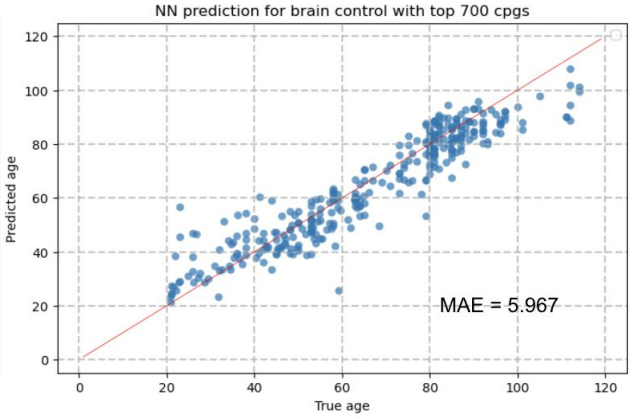
Breast

Are features transferable across different tissues?

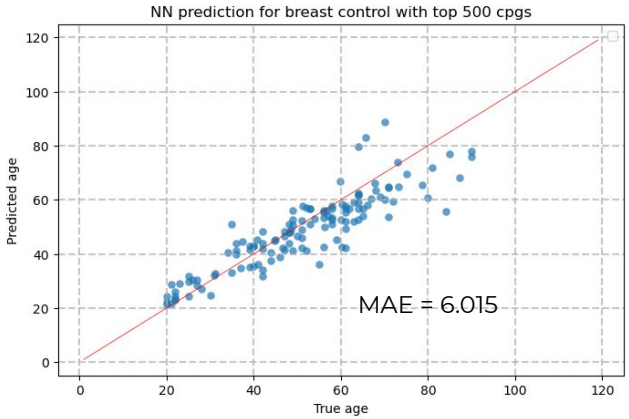
Leukocyte



Brain



Breast

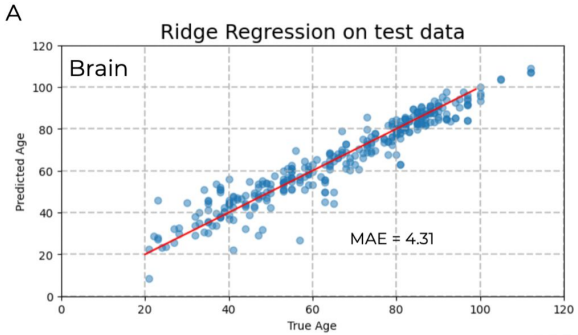


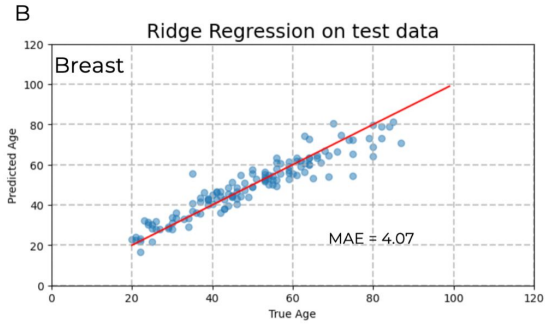
- For breast and brain model predictions, we observed reasonable prediction performance, although significantly worse than that of the blood prediction with blood top ranked cpgs.
- However, the **Leukocyte model had great performance using the blood top ranked cpgs**. This may be because most leukocytes are produced in our bone marrow from the same kind of stem cells that produce red blood cells, although leukocytes are existent in many parts of our body other than blood.

Age Prediction – Brain, Breast Tissue

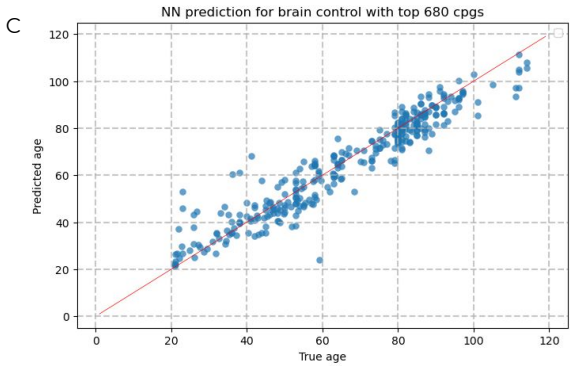
Brain | **Breast**

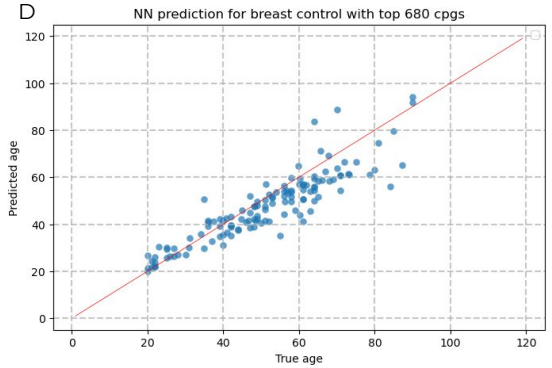
Ridge Regression





Neural Network





Comparative Modeling

1. Transferring healthy models to unhealthy cohorts
2. Transferring significant healthy CpG sites to unhealthy cohorts
3. Classification model for healthy vs unhealthy

3

Transferring healthy models to uneahlthy cohorts: brain

Healthy Control	XGBoost	Healthy Control	Compare	Unhealthy	Train	Healthy Control
~400,000 CpGs		Top 100 CpGs		Contains 55 of top 100 CpGs		Train on 55 CpGs in unhealthy data

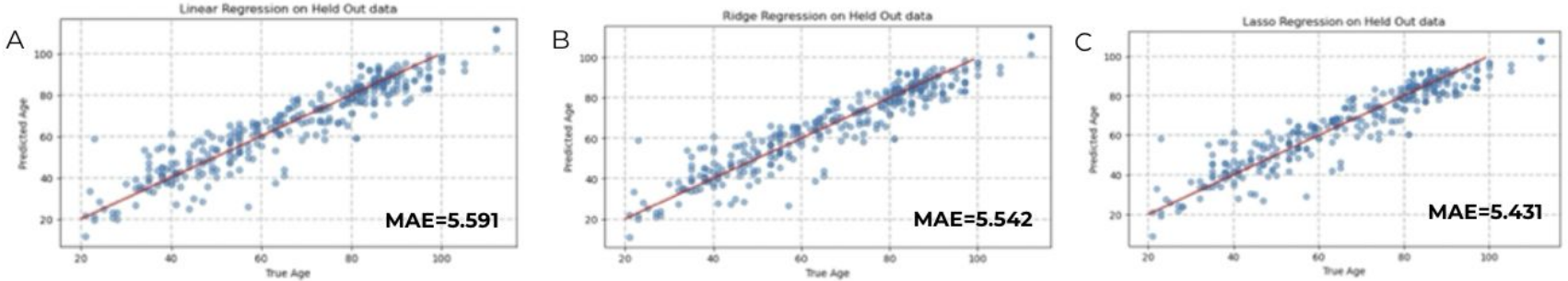
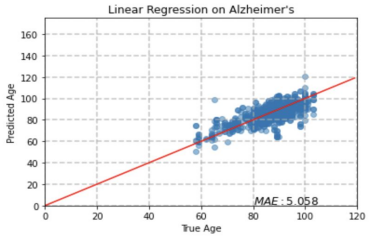
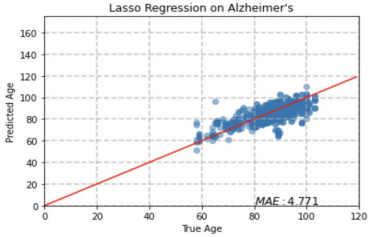
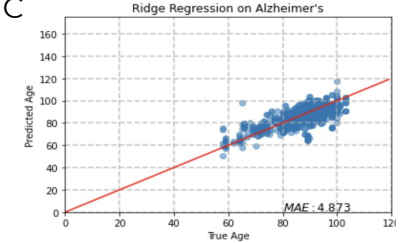
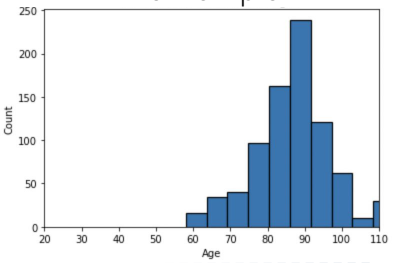
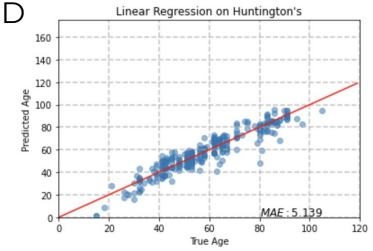
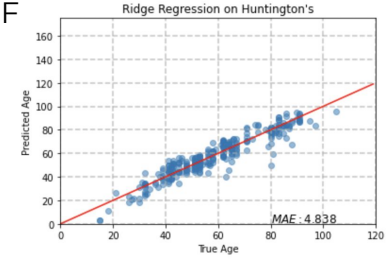
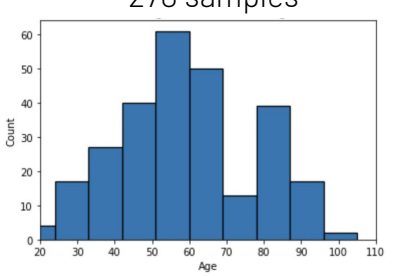
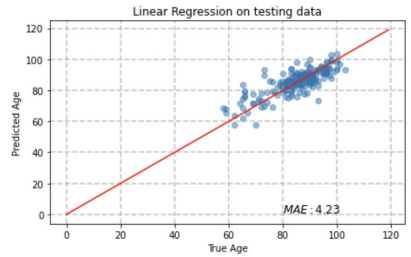
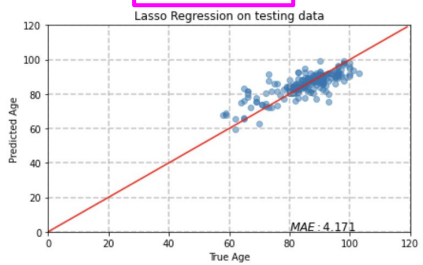
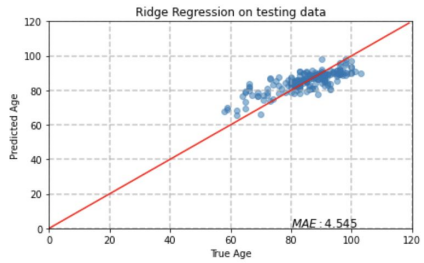
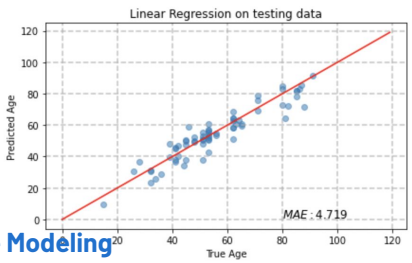
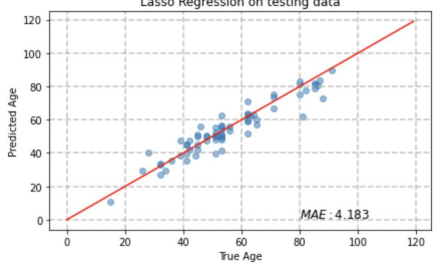


Figure: Results of healthy model with 55 CpGs on healthy cohort

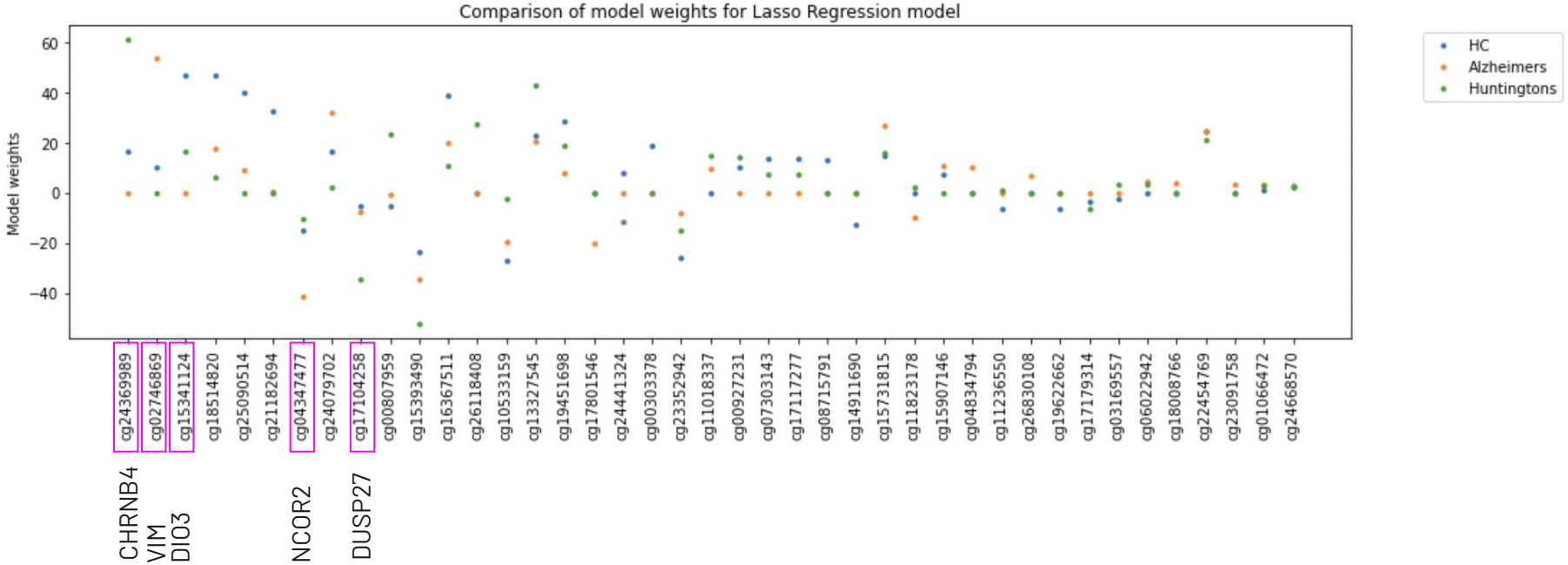
Transferring healthy models to uneahlthy cohorts: brain

	Linear	Lasso	Ridge	Cohort age distribution
Healthy	MAE=5.591	MAE=5.431	MAE=5.542	
Alzheimer's	A 	MAE=4.771 B 	C 	781 samples 
Huntington's	D 	MAE=4.471 E 	F 	270 samples 

Transferring significant healthy CpG sites to unhealthy cohorts

	Linear	Lasso	Ridge
Healthy	MAE=5.591	MAE=5.431	MAE=5.542
Alzheimer's	MAE=4.230 	MAE=4.171 	MAE=4.545 
Huntington's	MAE=4.719 	MAE=4.183 	MAE=4.282 

Comparing model weights



Classification for healthy vs unhealthy

Logistic Regression

Unhealthy: Alzheimer's patients

	Recall	Accuracy
Healthy	0.80	**0.73**
Unhealthy	0.62	

Unhealthy: Alzheimer's and Huntington's patients

	Recall	Accuracy
Healthy	0.72	**0.68**
Unhealthy	0.63	

Neural Network

Unhealthy: Alzheimer's patients

	Recall	Accuracy
Healthy	0.77	**0.74**
Unhealthy	0.70	

Biological significance

4

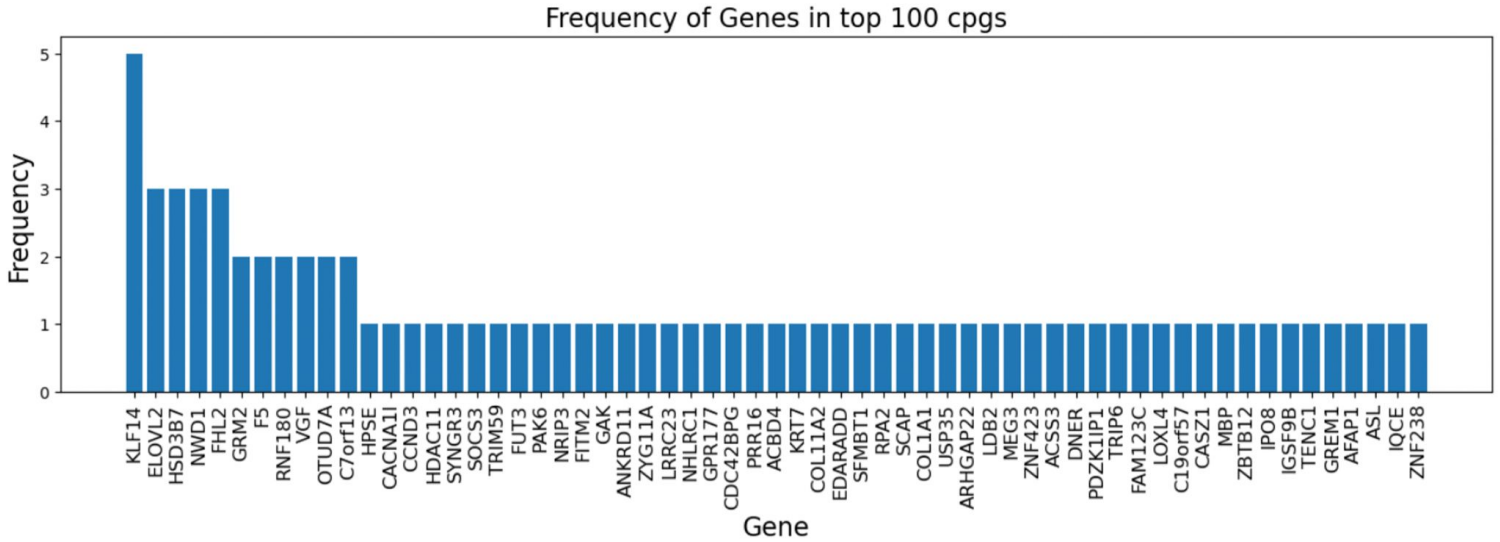
Biological Significance

Genes associated with the top 100 cpgs

cpg	gene	cpg	gene	cpg	gene	cpg	gene
cg14361627	KLF14	cg07927379	C7orf13	cg18933331		cg04084157	VGF
cg16867657	ELOVL2	cg19722847	IPO8	cg17471102	FUT3	cg10149533	
cg24724428	ELOVL2	cg10917602	HSD3B7	cg20010135	HSD3B7	cg17110586	
cg11649376	ACSS3	cg21406967	TRIP6	cg25256723	F5	cg09499629	KLF14
cg24079702	FHL2	cg09692396	LRRC23	cg06540876	ZBTB12	cg25428494	HPSE
cg04875128	OTUD7A	cg16762684	MBP	cg12580096	C19orf57	cg09748749	ASL
cg08097417	KLF14	cg01763090	OTUD7A	cg11693709	PAK6	cg04503319	ANKRD11
cg00292135	C7orf13	cg23078123	GPR177	cg19784428	NWD1	cg20249566	NWD1
cg02046143	IGSF9B	cg25410668	RPA2	cg01256539	PRR16	cg25693132	GRM2
cg07553761	TRIM59	cg07082267		cg04521765	LOXL4	cg11220950	SYNGR3
cg21572722	ELOVL2	cg02933228	CDC42BPG	cg01314044		cg00808969	USP35
cg04208403	ZNF423	cg23606718	FAM123C	cg22285878	KLF14	cg03752138	SOCS3
cg23500537		cg07955995	KLF14	cg01074797	PDZK1IP1	cg09648727	
cg08262002	LDB2	cg05331060		cg10943497	MEG3	cg15957394	AFAP1
cg04955333	IQCE	cg18651026	COL11A2	cg03032497		cg16008966	
cg09809672	EDARADD	cg10221746		cg19855470	CACNA1I	cg21186299	VGF
cg06639320	FHL2	cg05308819		cg00776080	TENC1	cg20273670	
cg17621438	RNF180	cg18618815	COL1A1	cg16054275	F5	cg18725681	FITM2
cg22736354	NHLRC1	cg18877361		cg23091758	NRIP3	cg22016779	DNER
cg22454769	FHL2	cg12252865	HDAC11	cg01552919	GAK	cg01676322	ACBD4
cg19344626	NWD1	cg16932827		cg04581938		cg21296230	GREM1
cg23744638		cg03883331		cg03404339	KRT7	cg26614073	SCAP
cg07850154	RNF180	cg07135942	ZNF238	cg00003345	CASZ1	cg01014399	
cg08453194	CCND3	cg03607117	SFMBT1	cg02025827	HSD3B7	cg06784991	ZYG11A
cg07927379	C7orf13	cg00753885		cg22796704	ARHGAP22	cg18343474	

Biological Significance

Gene frequency among top 100 cpgs



Biological Significance

Genes annotated

Rank	cpg	Gene	Function	Zinc finger	Refs related to aging
1	cg14361627	KLF14	Krüppel-Like Factor 14 (KLF14), transcription factor, master regulator of gene expression in the adipose tissue	x	16, 8, 5, 7
2	cg16867657	ELOVL2	Fatty Acid Elongase 2, involved in the synthesis of very long polyunsaturated fatty acids		21, 15, 14, 17,5, 7
3	cg24724428	ELOVL2			15, 14, 17,5, 7
4	cg11649376	ACSS3	Acyl-CoA Synthetase Short Chain Family Member 3, Ligates acetate and CoA6		1
5	cg24079702	FHL2	Four And A Half LIM Domains 2, Assembly of extracellular membranes, double zinc finger, LIM protein	x	5,17,2,
6	cg04875128	OTUD7A	OTU Deubiquitinase 7A, deubiquitinizing enzyme and possible tumor suppressor, zinc finger	x	21, 17,7
7	cg08097417	KLF14		x	21, 16, 8, 5, 7
8	cg00292135	C7orf13	Not much known		
9	cg02046143	IGSF9B	Immunoglobulin Superfamily Member 9B, cell adhesion, localized to inhibitory synapses		21, 7
10	cg07553761	TRIM59	Tripartite Motif Containing 59, E3 ubiquitin ligase, zinc finger, RING finger protein	x	15, 7
11	cg21572722	ELOVL2			15, 14, 17,5, 7
12	cg04208403	ZNF423	Zinc Finger Protein 423, Krüppel-Like Factor, zinc finger transcription factor, KO affects adipogenesis	x	16
13	cg23500537				21
14	cg08262002	LDB2	LIM Domain Binding 2, adapter molecule, binds LIM		14,15
15	cg04955333	IQCE	IQ Motif Containing E, signaling by GPCR and Hedgehog		21
16	cg09809672	EDARADD	EDAR Associated Death Domain, Ectodysplasin-A receptor-associated adapter protein		21, 16, 4, 9
17	cg06639320	FHL2		x	21, 5,17,2,
18	cg17621438	RNF180	E3 Ubiquitin-Protein Ligase RNF180, promotes protein degradation by the proteasome pathway	x	21
19	cg22736354	NHLRC1	E3 Ubiquitin-Protein Ligase NHLRC1 , promotes protein degradation by the proteasome pathway	x	9
20	cg22454769	FHL2		x	21, 5,17,2,
21	cg19344626	NWD1	NACHT And WD Repeat Domain Containing 1, modulator of androgen receptor activity		
22	cg23744638				21
23	cg07850154	RNF180		x	

Published: 04 April 2014

Ubiquitin sets the timer: impacts on aging and longevity

Éva Kevei & Thorsten Hoppe

Nature Structural & Molecular Biology 21, 290–292(2014) | Cite this article

MINI REVIEW ARTICLE

Front. Aging Neurosci., 06 December 2019 | https://doi.org/10.3389/fnagi.2019.00324

Perturbations of Ubiquitin-Proteasome-Mediated Proteolysis in Aging and Alzheimer's Disease

Ashok N. Hegde, Spencer G. Smith, Lindsey M. Duke, Allison Pourquoi and Savannah Vaz

Conclusions

- We have been able to build models to predict age with a mean error of 3.6 years across the entire adult lifespan.
- From the ~ 400,000 DNA methylation sites (CpG sites) we started with, we have identified ~700 that are optimal for age predictive modeling.
- Models are not transferable across tissues, but many CpGs are.
- Models developed with brain tissue from healthy individuals can also predict the ages of patients with neurodegenerative diseases.
- Our top ranked CpGs are often associated with genes that regulate adipose-tissue gene expression and the ubiquitin-proteasome protein degradation pathway.

Thank you

Harvard IACS

Chris Tanner

Phoebe Wong

Merck Data Scientists

Antong Chen

Gregory Bryman