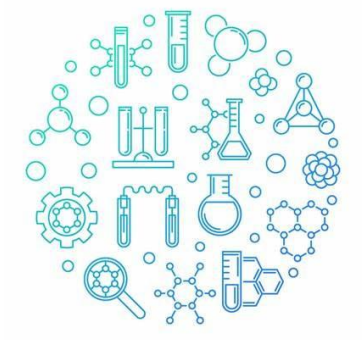
Comparison of Epigenetic Frameworks on Canine Buccal Samples

Jared Paul Guevara
University of California, Los Angeles
December 2nd, 2022

About Me



- 4th year Biochemistry student
- Joined the Pellegrini Lab in June 2022
- Born in San Diego, raised in Glendale, currently live in SGV
- Have 2 pet huskies at home named Lady and Starky
- Interested in going into Data/Software Engineering after undergrad

Figure 0. (left to right) Lady, Starky, and Foxy

Canine Epigenetic Study

- Study adapted from Trapp et al. (2021)
- Aim to analyze canine buccal swab data using the single cell-Age (scAge) framework to predict cellular aging by way of DNA methylation

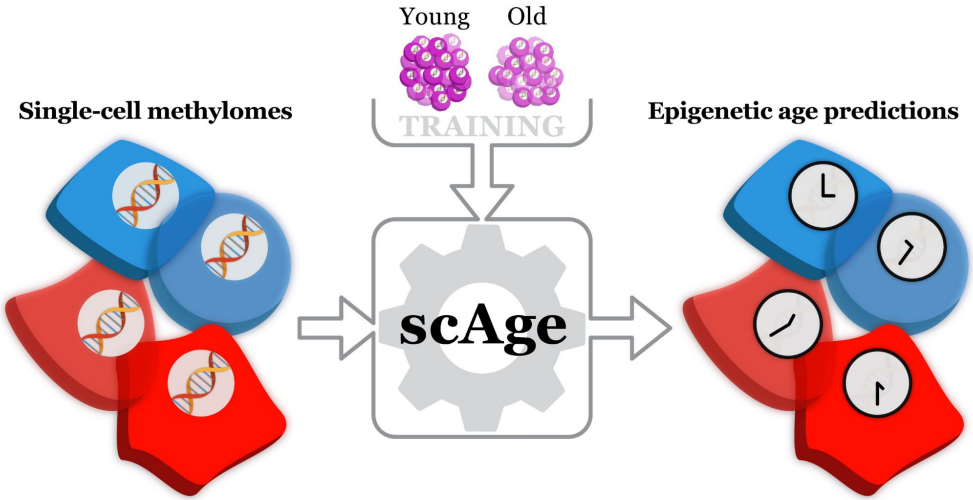
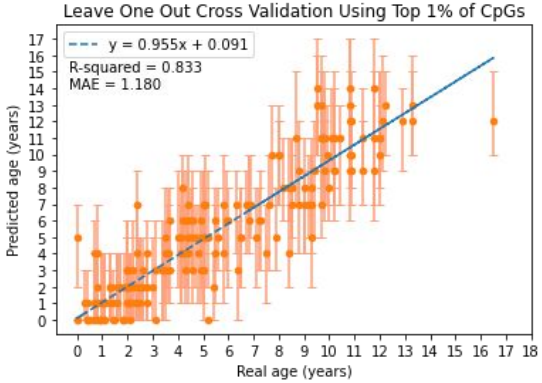


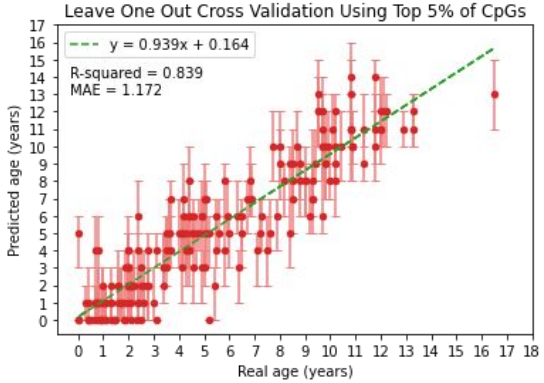
Figure 1. Simplified visual of scAge framework (Trapp et al. 2021)

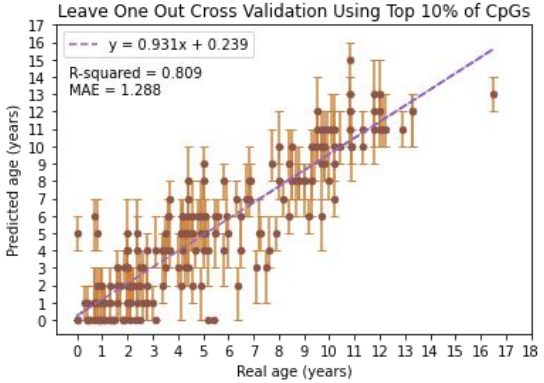
Materials and Methods (pt. 1)

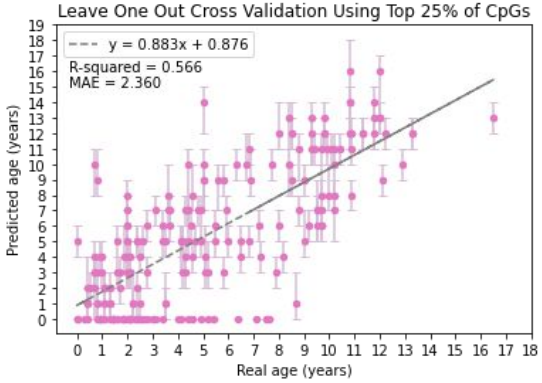
- Canine metadata from preexisting collections of canine buccal swabs
  - Samples collected in November 2019 through February 2020
- Data cleaning and analysis in Python
  - Removal of invalid samples, from 218 → 204 samples
  - BiSulfite Bolt (BSBolt) used to generate matrix of methylation data from all 204 canine samples
  - scAge used to generate single-cell methylomes, process existing samples, and train age predictions
- Primary method of analysis: Leave One Out Cross Validation (LOOCV)
  - Comparing between different thresholds of CpG site percentages: 1%, 5%, 10%, 25%, 50%, 100%
  - Maximum age tested was 20 years, since the max age of all samples within the metadata was around 16.5 years old
  - Age step used was 1 year

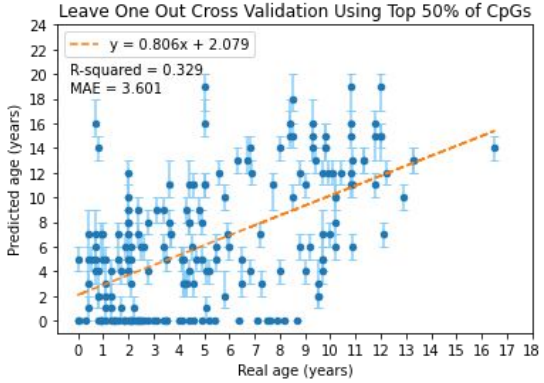
Results: Predicted vs. Real Age Using scAge

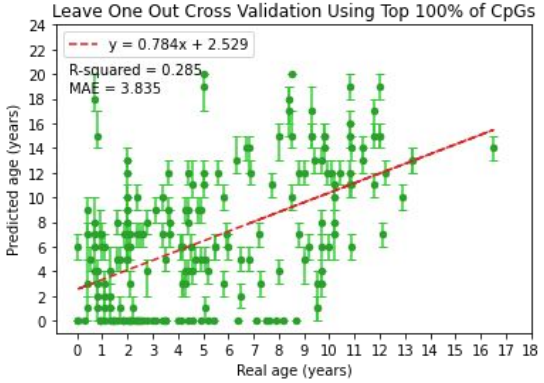




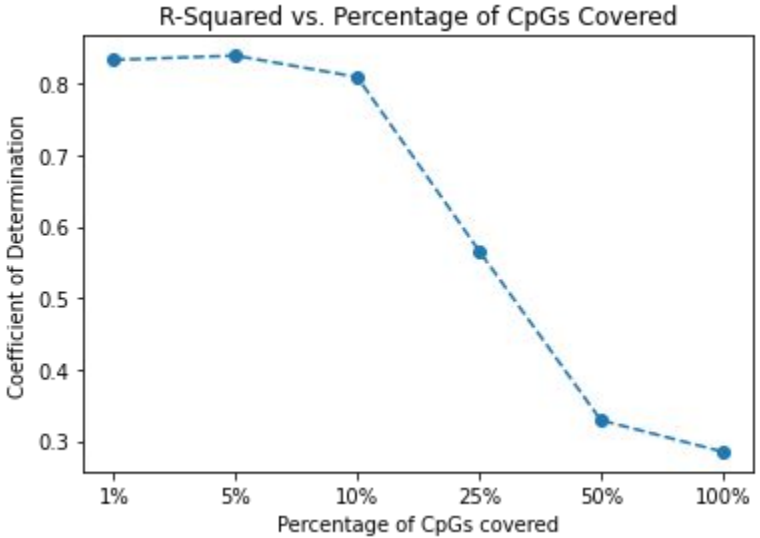


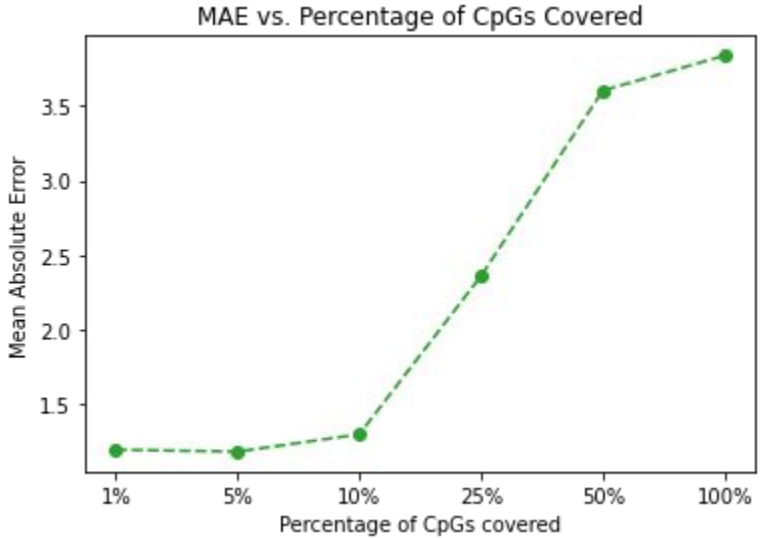




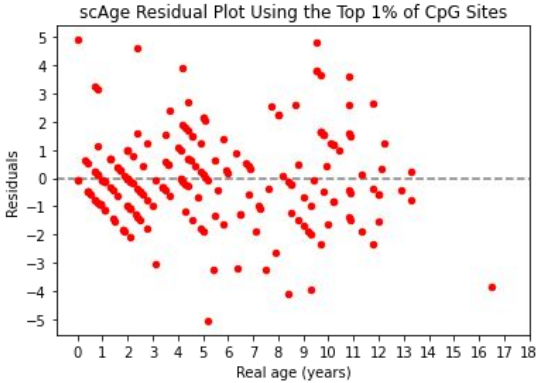


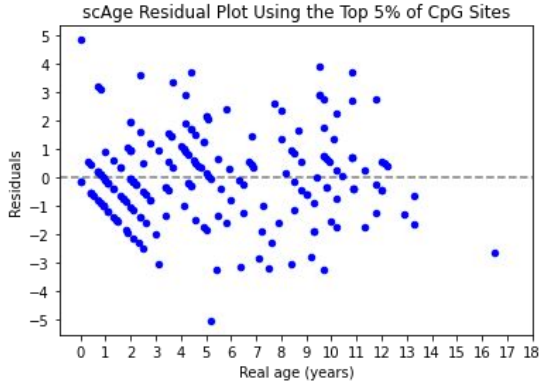
Results: R-squared & MAE vs. Percentage of CpG Sites

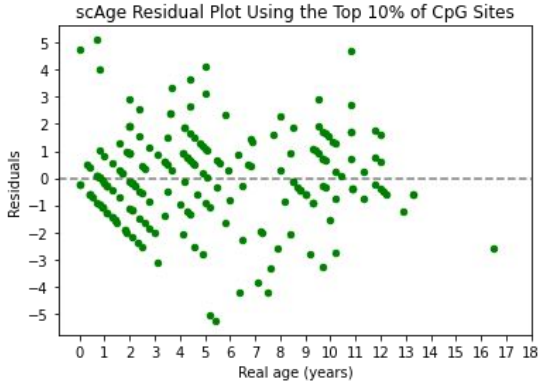


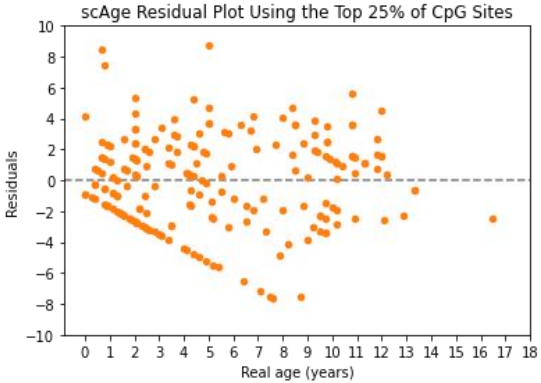


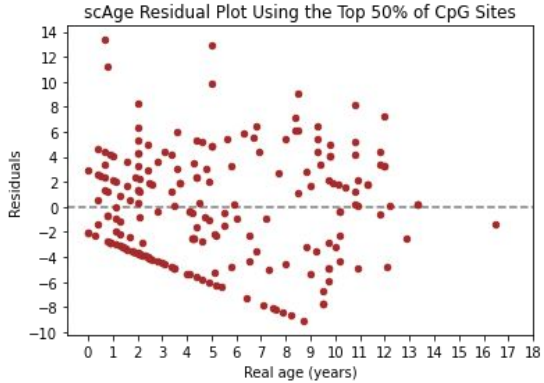
Results: Residual Plots using scAge

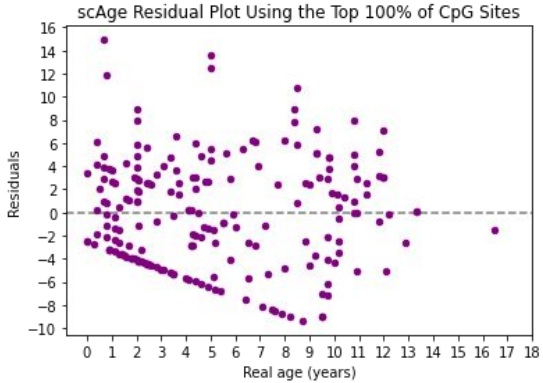




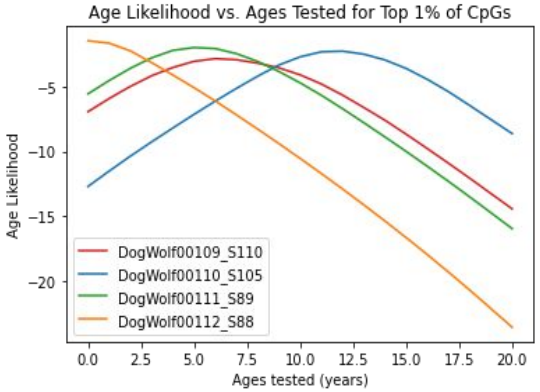


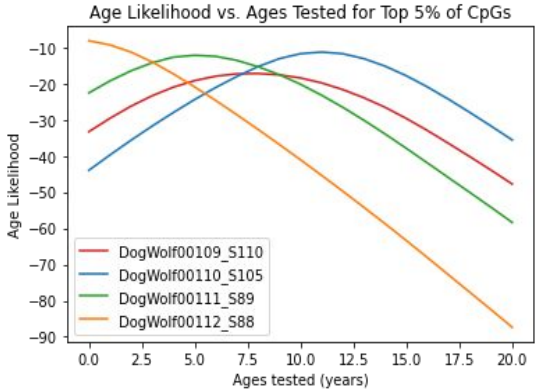


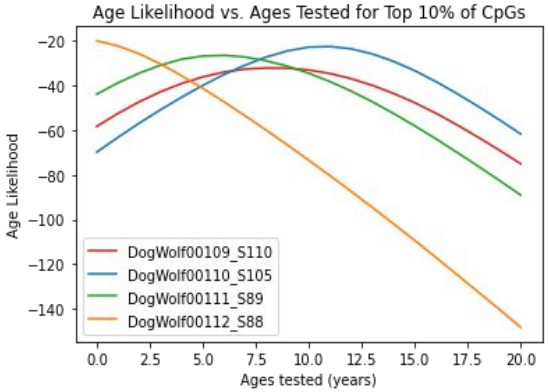


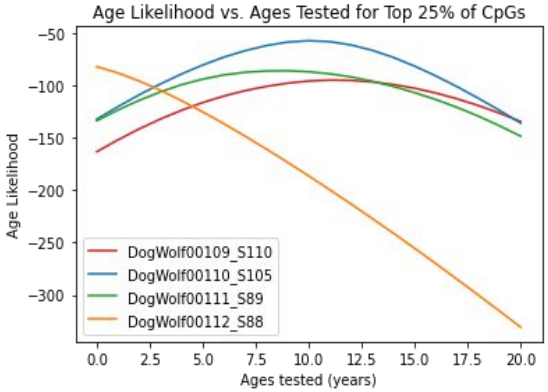


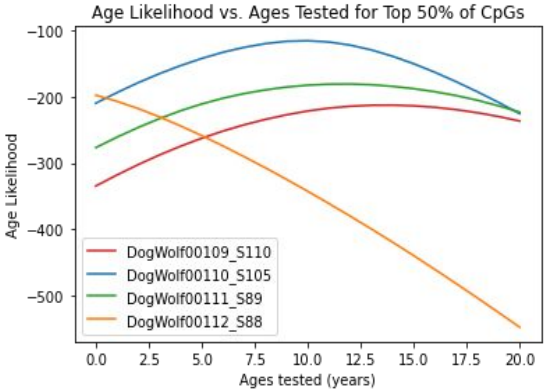
Results: Age Likelihood vs. Ages Tested of 4 Samples

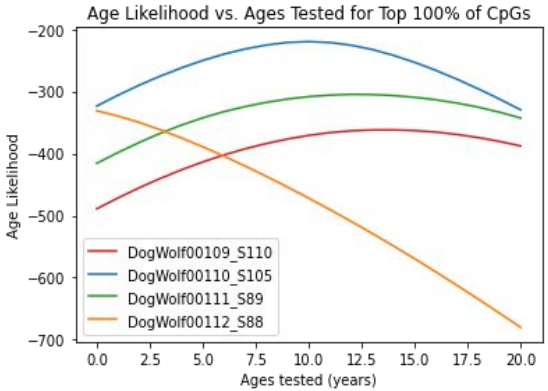




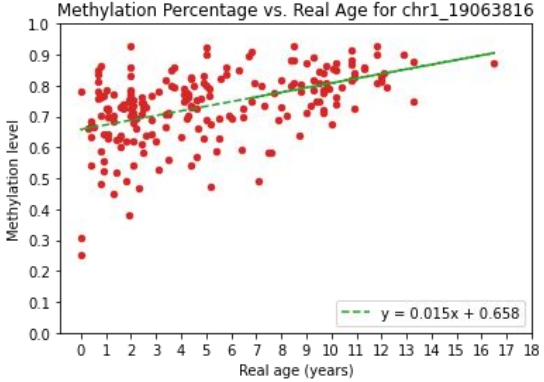


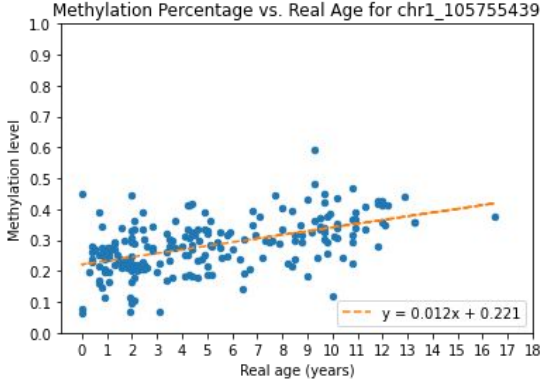


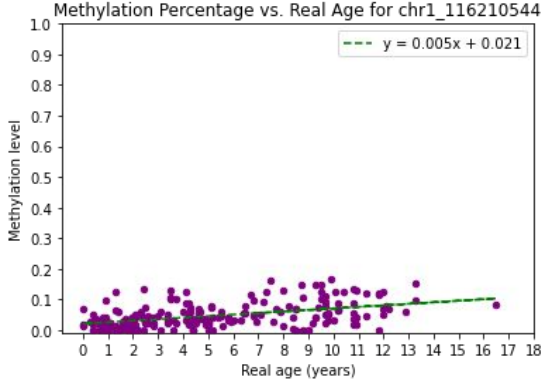


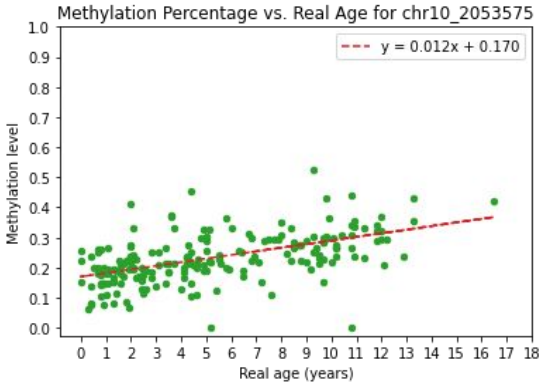


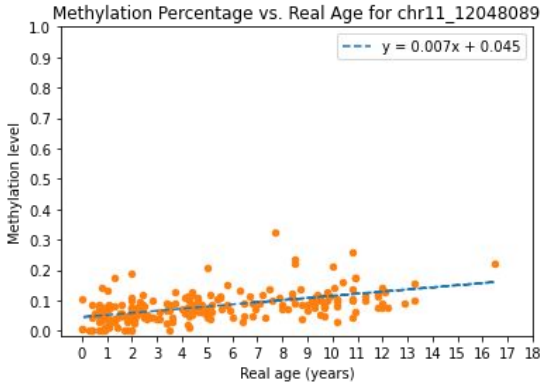
Results: Methylation Level vs. Real Age of 6 CpG Sites

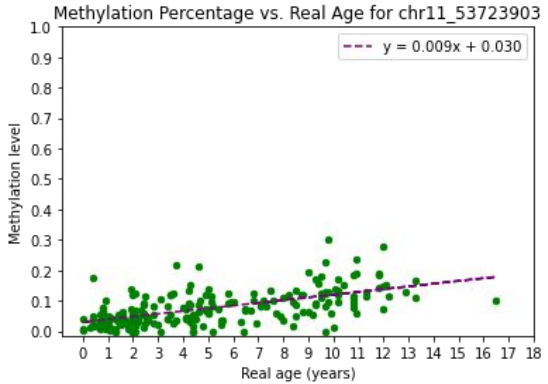












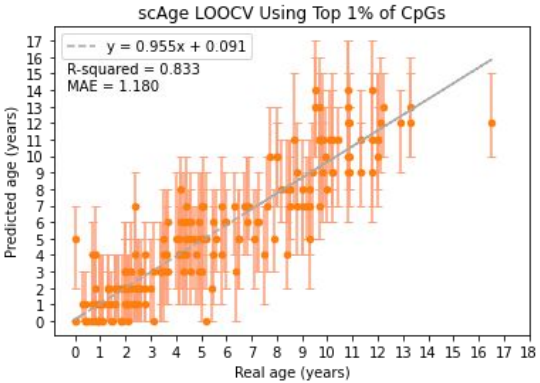
Conclusions (pt. 1)

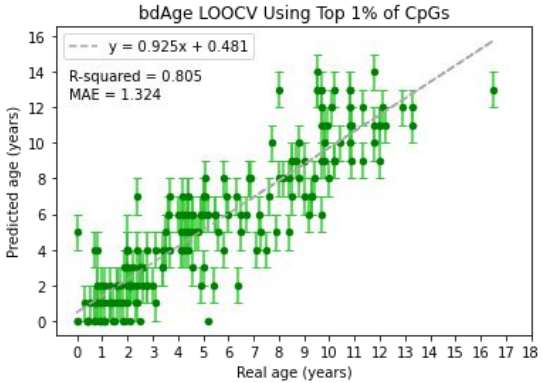
- Observed a slightly higher R-squared value and slightly lower MAE value when analyzing the top 5% of CpG sites compared to the top 1% of sites
- Besides top 5%, expected trend is present for other percentages
  - Higher percent of CpG sites covered → less accurate measurements
- Likelihood of ages becomes increasingly more negative with higher percentages of CpG sites

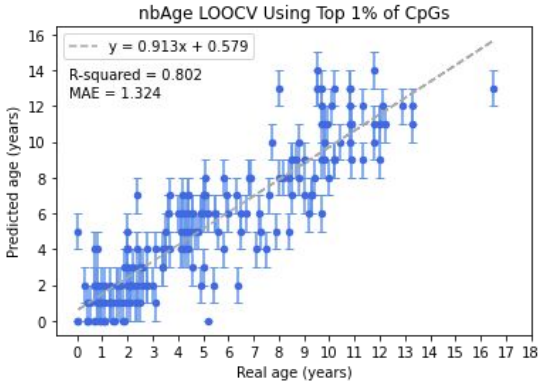
Materials and Methods (pt. 2)

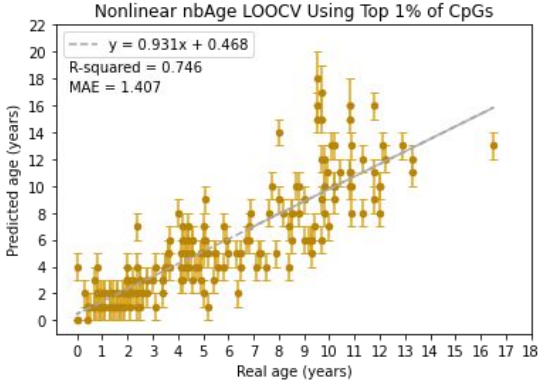
- Comparison of different epigenetic frameworks
  - scAge vs. binomial distribution-Age (bdAge) vs. non binomial distribution-Age (nbAge) vs. Nonlinear (NL) bdAge vs. NL nbAge
  - Method of analysis: LOOCV
- Analysis conducted on only the top 1% of CpG sites per framework
- Compared the predicted vs. real ages, R-squared/MAE values, and residual plot
- Top selected CpG site for the linear frameworks (scAge, bdAge, nbAge) compared vs. the top selected CpG site for the nonlinear frameworks

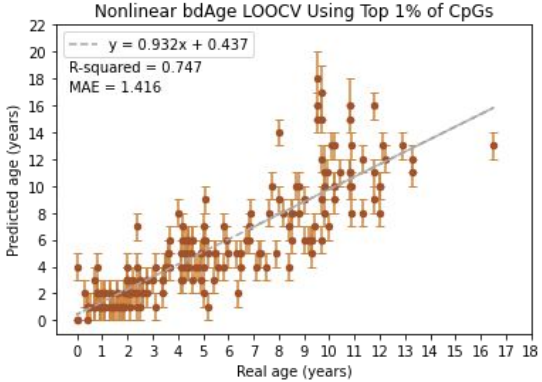
Results: Predicted vs. Real Ages of Different Frameworks



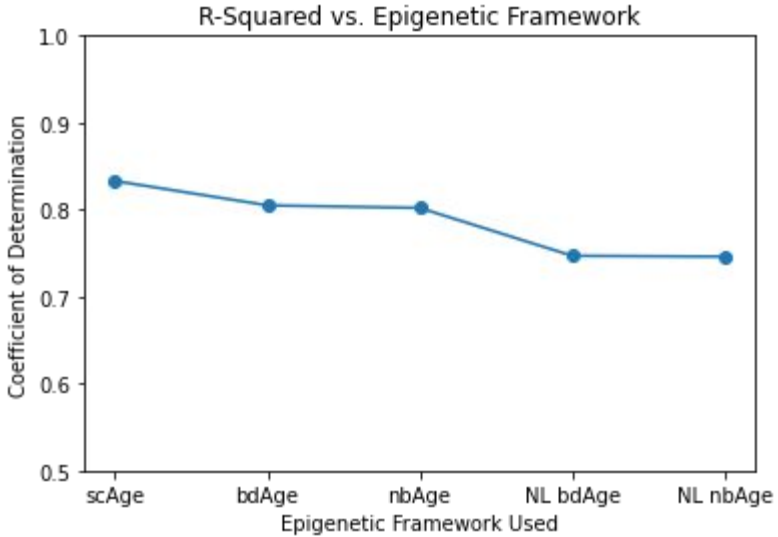


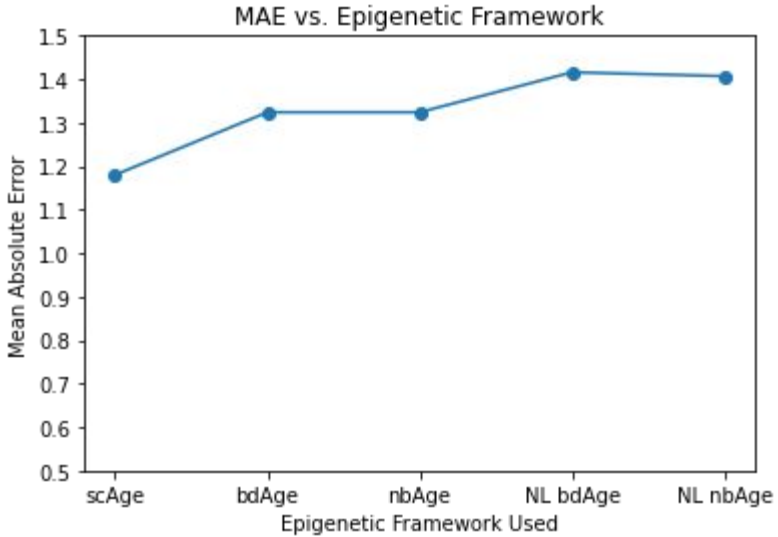




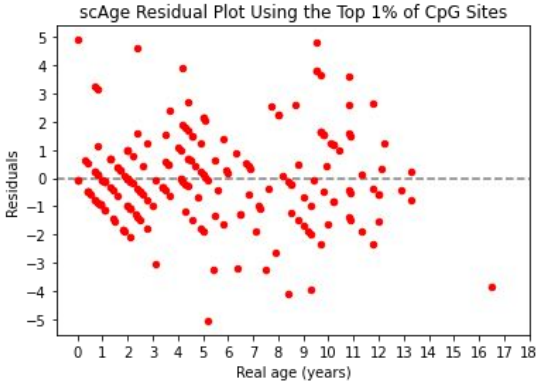


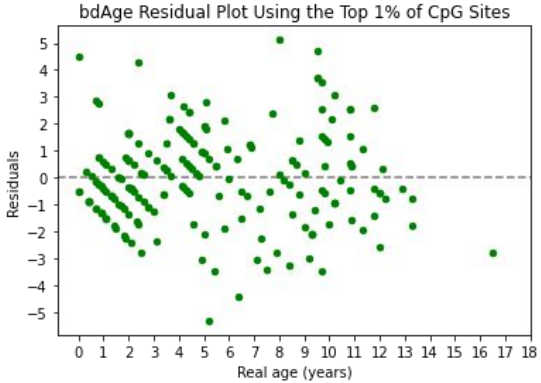
Results: R-squared and MAE vs. Different Frameworks

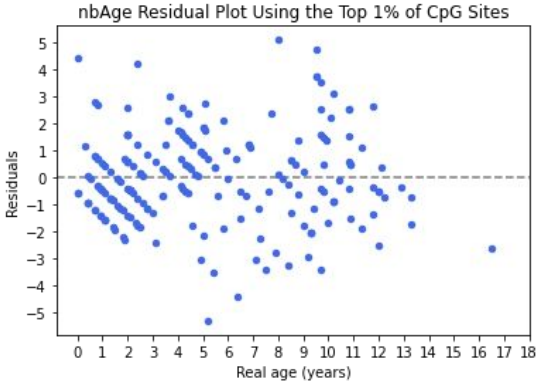


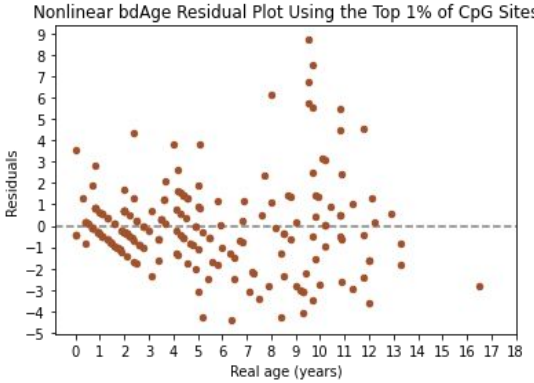


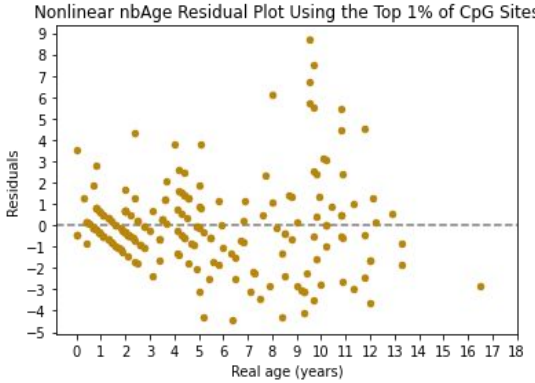
Results: Residual Plots of Different Frameworks



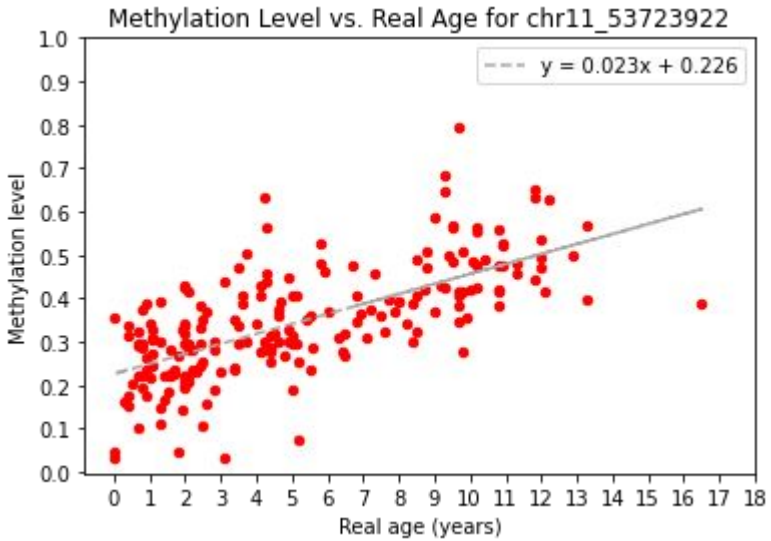




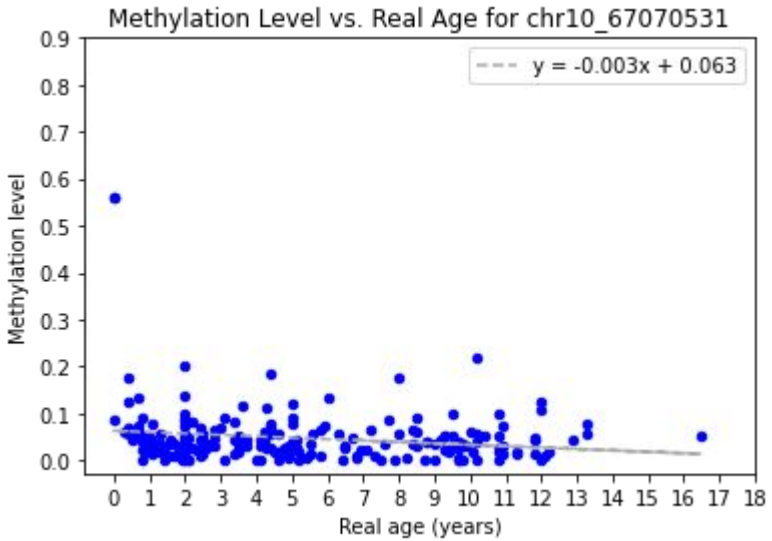




Results: Methylation Level vs. Real Age of Top Selected CpG



Top selected site for linear frameworks
(scAge, bdAge, nbAge)



Top selected site for nonlinear frameworks
(NL bdAge, NL nbAge)

Conclusions (pt. 2)

- Observed higher R-squared values for the nonlinear frameworks (around +0.1 difference for NL bdAge and NL nbAge)
- Conversely, higher MAE values towards the median years for nonlinear frameworks
    - Most outliers in the ~9-10 years old range
- As of now, scAge has the highest accuracy among all frameworks for this dataset at the top 1% of CpG sites

Discussion

- Investigate certain outlier samples (ex. Samples with predicted age of 0)
- Address negative linear trend in residual plots
- Test different thresholds of CpG site percentages across all frameworks
- Use different machine learning methods (ex. Ridge/LASSO regression)

References

- Trapp, A., Kerepesi, C. & Gladyshev, V.N. Profiling epigenetic age in single cells. *Nat Aging* 1, 1189–1201 (2021). https://doi.org/10.1038/s43587-021-00134-3