Investigating the Relationship between DNA Methylation in Whole Blood and Chemical Exposure for Predictive Modeling

Hugo Nørholm

Thesis submitted for the degree of
Master in Programming and system architecture
60 credits

Department of Informatics
The Faculty of Mathematics and Natural Sciences

UNIVERSITY OF OSLO

Spring 2023

Investigating the Relationship between DNA Methylation in Whole Blood and Chemical Exposure for Predictive Modeling

Hugo Nørholm

Chemical Exposure for Predictive Modeling

Investigating the Relationship between DNA Methylation in Whole Blood and

Printed: Reprosentralen, University of Oslo

Abstract

This study investigates the potential impacts of Polybrominated Biphenyl (PBB) exposure, smoking habits, and benzene exposure on methylation in whole blood samples. These exposures have proven to affect the general health of people exposed to them, and both smoking and PBB have been proven to affect DNA methylation.

The main objective of this project is to attempt to use a similar methodology used for designing epigenetic clocks to create a predictive model for said exposures. Along the way we will also be using existing epigenetic clocks to look at the age acceleration of the subjects to see if its affected by the exposures as well. Additionally we will look if blood cell composition were affected.

I generated a variety of models for PBB and smoking using Elastic Net neural network, linear regression and Naive Bayes. Differences between groups of different exposure were analysed by applying linear regression with residual and intrinsic age acceleration as the dependent variables. The datasets were used to calculate cell populations to see if there was an observable effect on cell population caused by exposure. I also analyzed some of the best predictors used in the prediction models to see if they could be used as biomarkers for their given exposure.

I trained ten prediction models for PBB exposure and ten additional models for smoking habits. For PBB my models performance ranged from an RMSE of 1.33-3.56 and for smoking the models had an accuracy in the range of 0.84-0.65. I also managed to identify several CpG sites that had strong statistical relation to exposure. As well as an observable effect on both CD4T and CD8T cells in the blood.

This work has found that Elastic Net is a well-suited algorithm for predictive models for exposure, as has also been shown in previously published work. It also indicates that neither exposure PBB or cigarette smoking have a significant impact on age acceleration. I did however find that PBB and smoking and benzene seemed to have an impact on CD4T and CD8T cell population in blood. Additionally I was able to identify CpGs that might be usable as biological markers for exposure of PBB, smoking and benzene exposure.

Acknowledgements

I would like to express my thanks to my main supervisor, Marcin W. Wojewodzic, for their guidance and support throughout this project and for being a better supervisor than I deserved. I also want to acknowledge my internal supervisor Torgjørn Rognes for their assistance.

Lastly, my deepest thanks goes to my mother, who has been a constant source of encouragement during my studies. This achievement would not have been possible without her support.

Contents

List of Figures

List of Tables

Part I
Background

1 Introduction

Study to explore existing computational methods and develop new methods for estimating omics clocks in the context of chemical exposure using machine learning to analyze epigenetic signatures for detection the effects of exposure to chemical mixtures, while integrating with public databases.

2 Epigenetics

The basis for this thesis is going to be epigenetics which is the study of how your behaviour and environment can cause changes that affect your genes without altering the DNA sequence. The term epigenetics was first coined by the developmental biologist Conrad H. Waddington (1905–1975) to summarize a new branch of biology which focuses on the links between gene and protein expression [59]. However the meaning of the term has changed since then to now become the study of how your behaviors and environment can cause changes that affect the way your genes work. This includes changes that can affect your overall health and genetic age which is what this work is going to look further into. Epigenetic modifications can occur through a variety of mechanisms, including DNA methylation, histone modification, and RNA-mediated regulation.

2.1 DNA methylation

DNA methylation is a common epigenetic modification that plays a key role in the regulation of gene expression in eukaryotic cells. It involves the addition of a methyl group to the cytosine base of DNA molecules, usually at CpG dinucleotides [56].

DNA methylation can occur in different regions of the genome, including promoter regions, enhancers, and gene bodies. Methylation of promoter regions is generally associated with transcriptional repression, while methylation of gene bodies is associated with transcriptional elongation and alternative splicing.

The addition of methyl groups to DNA molecules can affect gene expression by altering the accessibility of DNA to transcription factors and other proteins that regulate gene expression. Specifically, DNA methylation can block the binding of transcription factors to promoter regions, prevent the binding of RNA polymerase to the transcriptional start site, and recruit proteins that repress gene expression.

Furthermore, DNA methylation is a heritable modification, meaning that it can be passed down from one generation to the next [24]. In this way, DNA methylation can play a key role in epigenetic inheritance and the regulation of development.

Aberrant DNA methylation patterns have been associated with a variety of diseases, including cancer, cardiovascular disease, and neurological disorders. As a result, DNA methylation has emerged as a potential target for therapeutic interventions and a promising biomarker for disease diagnosis and prognosis.

2.2 MicroRNAs

MicroRNAs (miRNAs) are a class of small non-coding RNAs that play a key role in post-transcriptional gene regulation in eukaryotic cells [6]. miRNAs can regulate gene expression by binding to messenger RNA (mRNA) molecules and either degrading them or preventing their translation into proteins.

In epigenetics, miRNAs are considered an important part of the regulatory machinery that controls gene expression by affecting chromatin structure and function. miRNAs can target specific chromatin-modifying enzymes or transcription factors, thereby altering the epigenetic state of genes and regulating their expression [22].

Furthermore, miRNAs have been shown to play an important role in the regulation of cell differentiation, development, and disease. Dysregulation of miRNA expression has been implicated in a variety of human diseases, including cancer, cardiovascular disease, and neurodegenerative disorders [39].

Overall, miRNAs are an important component of the epigenetic machinery that helps to regulate gene expression in response to environmental cues and developmental signals. Their potential as therapeutic targets or biomarkers for disease underscores the importance of continued research in this field.

2.3 Histone modification

Histone modification is a key mechanism in epigenetics, which refers to changes in gene expression that are not caused by alterations in the DNA sequence itself [5]. Epigenetic modifications, including histone modifications, can affect how genes are expressed or silenced without changing the underlying genetic code.

In particular, histone modifications involve the addition or removal of chemical groups, such as acetyl, methyl, or phosphate groups, to the histone proteins that make up the nucleosomes [43]. These modifications can alter the structure of chromatin and affect the accessibility of DNA to transcription factors and other proteins that regulate gene expression. For example, the addition of acetyl groups to histone proteins is associated with open chromatin and active gene expression, while the addition of methyl groups can either promote or repress gene expression, depending on the location of the modification and the context of other modifications.

Histone modifications are reversible and can be dynamically regulated in re-

sponse to environmental cues or developmental signals. They are known to play critical roles in a wide range of biological processes, including embryonic development, differentiation, and disease [69]. Understanding the mechanisms and effects of histone modifications is essential for elucidating the epigenetic regulation of gene expression and identifying potential targets for therapeutic interventions.

3 Genetics

Genetics is the study of genes, heredity, and genetic variation in living organisms. It encompasses the mechanisms by which genetic information is passed from one generation to the next and how this information is expressed and regulated within cells.

3.1 Genome

The genome is the complete set of genetic instructions encoded within an organism's DNA. It contains all the genes that are necessary for an organism to develop and function. The genome is organized into chromosomes, which are long strands of DNA that are coiled and packaged within the nucleus of a cell. The DNA is made up of four different nucleotide bases adenine, cytosine, guanine and thymine, A, C, G and T for short and the DNA code is based on the order they appear in.

3.2 Gene

Genes are segments of DNA that contain the instructions for making specific proteins or RNA molecules. These instructions are encoded in the sequence of nucleotide bases that make up the DNA molecule. Genes are the fundamental units of heredity and determine many of an organism's traits.

One example of a gene is the BRCA1 gene, which is associated with an increased risk of breast and ovarian cancer. Mutations in this gene can lead to a disruption of its normal function and increase the likelihood of developing cancer [41].

Another example is the CFTR gene, which is associated with cystic fibrosis. Mutations in this gene can cause a defect in the transport of chloride ions across cell membranes, leading to the buildup of mucus in the lungs and other organs [65].

4 CpGs and CpG islands

CpG is a short term for cytosine-phosphate-guanine dinucleotide, which is a sequence of nucleotides found in DNA. CpG dinucleotides are usually underrep-

resented in the genome, but they are frequently clustered in regions known as CpG islands. CpG sites should not be confused with GpC sites which is when the order of the guanine and cytosine are the other way around. CpG islands are regions with a high amount of CpG sites. To qualify as an island it has to be a region of at least 200 base pairs with a GC percentage greater than 50 percent and an expected CpG ratio of at least 60 percent. The expected CpG ratio by is calculated by (number of C * number of G)/amount of base pairs.

Figure 1: CpG sequence of one DNA strand versus C-G base pair on complementary strands [33]

The reason CpG islands are important in epigenetics is because of how they are related to promoters. A promoter is a sequence of DNA to which proteins bind to initiate transcription of a single RNA transcript from the DNA downstream of the promoter. In humans 70 percent of promoters that are near the transcription start site of a gene contain a CpG island.The promoter can be hindered in its function when methylation of CpG islands leads to silencing of the genes.

4.1 Methylation in CpG sites

DNA methylation is a well-researched epigenetic mechanism that regulates gene expression by adding or removing a methyl group. In CpG sites this occurs by methyl groups binding to the cytosines forming 5-methylcytosines. When multiple CpG sites are methylated in CpG islands of promoters it leads to silencing of the gene. This is not the only cause of a gene being silenced, but when a gene is silenced the CpG sites in the associated promoter CpG island usually methylates leading to stable silencing of the gene [8].

5 Extrinsic and intrinsic factors of epigenetics

Steve Horvath and Kenneth Raj [35] categorizes the factors that affect methylation as either intrinsic or extrinsic. Intrinsic factors are the factors that are the same no matter do not come from the environment that one lives is such as cell differences and age. Extrinsic factors are environmental factors that affect epigenetic markers. This can include diet, smoking habits or chemical exposure such

as through contaminated water or unsafe work environments. All these factors cause different changes in methylation and must be evaluated differently.

5.1 Celltypes

Different cell types methylate differently, therefore it is important to consider what cell type we are testing when looking at epigenetics. This is even more prevalent when comparing methylation changes in cells from different types of tissue. [73]

5.2 Age

There is a well-documented relationship between age and DNA methylation. DNA methylation is known to change with age, and these changes can have important implications for health and disease.

Several studies have shown that DNA methylation patterns change over time, with some CpG sites becoming more methylated and others becoming less methylated with increasing age. For example, a study by Hannum et al. [31] found that DNA methylation levels at 353 CpG sites were strongly associated with age across a wide range of tissues, and that these sites could be used to accurately predict an individual's age.

Other studies have shown that age-related changes in DNA methylation can be tissue-specific. In a study by Rakyan et al. [62] found that DNA methylation patterns in blood cells were strongly associated with age, but that the patterns of methylation in other tissues, such as brain and muscle, were less strongly associated with age.

Age-related changes in DNA methylation have been linked to a variety of health outcomes. For example, changes in DNA methylation patterns have been implicated in the development of age-related diseases, such as cancer, cardiovascular disease, and Alzheimer's disease [35].

5.3 Smoking habits

Cigarette smoking is a well known for being a major causal risk factor for various diseases including cancers, cardiovascular disease, chronic obstructive pulmonary disease, and osteoporosis [38]. When comparing current smokers against people who have never smoked using the Illumina BeadChip 450K array on blood derived DNA samples this article [38] found that the samples were statistically significantly differentially methylated. The genes associated with the CpGs sites that were found to be methylated in this study have also been associated with severe smoking related traits.

There are many harmful chemicals found in cigarettes, and they can cause a

range of health problems, including cancer, heart disease, and lung disease. Here are some of the most dangerous chemicals found in cigarettes [20, 58]:

1. Nicotine: Nicotine is the addictive substance in cigarettes. It increases heart rate, blood pressure, and constricts blood vessels [21].

2. Tar: Tar is the sticky substance that collects in the lungs of smokers. It contains many harmful chemicals, including polycyclic aromatic hydrocarbons (PAHs), which are known to cause cancer [63].

3. Carbon monoxide: Carbon monoxide is a poisonous gas that is produced when cigarettes are burned. It interferes with the ability of the body to transport oxygen to vital organs, including the brain and heart [32].

4. Formaldehyde: Formaldehyde is a colorless gas with a strong odor. It is used in the production of many products, including cigarettes. Formaldehyde is a carcinogen that has been linked to several types of cancer [37].

5. Benzene: Benzene is a colorless chemical that is used in the production of many products, including cigarettes. It is a carcinogen that has been linked to several types of cancer [80].

6. Acetone: Acetone is a colorless chemical that is used in the production of many products, including cigarettes. It is a toxic substance that can cause damage to the central nervous system, kidneys, and liver [79].

5.4 Chemical exposure

Chemical exposure is something we are constantly being warned about both in the food and drink we consume as well as from the environment that we live in and these chemicals are often linked to health issues and shortened lifespan.
Arsenic contaminated drinking water has of example been associated with a variety of adverse health effects and shortened lifespan are consumed by an estimated 200 million people worldwide [55]. There are also studies that show a clear differentiation in methylation between people who have been exposed to arsenic and those who have not [19].
Air pollution exposure is estimated to contribute to approximately seven million early deaths every year worldwide. And emerging data indicates that air pollution exposure changes the epigenetic mark, DNA methylation [64].

5.4.1 Polybrominated biphenyl (PBB)

Polybrominated biphenyls (PBBs) are a class of persistent organic pollutants that were once widely used as flame retardants in a variety of consumer products, including electronics, textiles, and plastics. PBBs are similar in structure and toxicity to polychlorinated biphenyls (PCBs), which have been banned in many countries due to their health and environmental impacts. PBBs have also been linked to a range of adverse health effects in humans and animals.

Here are some of the health effects associated with exposure to PBBs:

1. Endocrine disruption: PBBs have been shown to disrupt the endocrine system by mimicking the effects of hormones such as estrogen. This can lead to a range of adverse health effects, including reproductive problems, developmental delays, and thyroid dysfunction [17].

2. Cancer: Animal studies have shown that PBBs can cause cancer in various organs, including the liver, thyroid, and mammary glands [57].

3. Neurotoxicity: PBBs have been shown to cause neurotoxicity in animals, with effects including behavioral changes, learning and memory deficits, and motor dysfunction [42].

4. Immune system effects: PBBs have been shown to affect the immune system, with effects including decreased antibody production, altered immune cell function, and increased susceptibility to infectious diseases [53].

5. Developmental effects: PBBs can cross the placenta and affect fetal development. Animal studies have shown that PBB exposure during pregnancy can lead to developmental delays, altered behavior, and decreased survival rates in offspring [66].

Overall, exposure to PBBs has been associated with a range of adverse health effects, and their use has been banned in many countries. However, PBBs are persistent organic pollutants and can remain in the environment for a long time, continuing to pose a potential risk to human health and the environment.

5.4.2 Benzene

Benzene as mentioned earlier chemical is found in cigarettes but it is used in the production of many other chemicals and is present in crude oil ans gasoline. Benzene exposure can have serious health effects, including:

1. Cancer: Benzene is a known carcinogen and can cause leukemia, a cancer of the blood-forming organs, and other cancers such as non-Hodgkin's lymphoma [80].

2. Blood disorders: Benzene exposure can cause a decrease in red blood cells, leading to anemia, and a decrease in white blood cells, which can weaken the immune system and make the body more susceptible to infections [80].

3. Reproductive effects: Long-term exposure to benzene can affect the reproductive system and cause menstrual disorders and infertility in women and decrease sperm count in men [61].

4. Neurological effects: Benzene exposure can cause dizziness, headaches, tremors, and loss of consciousness [80].

5. Skin disorders: Benzene exposure can cause skin irritation, rashes, and burns.

6. Respiratory effects: Benzene exposure can cause respiratory distress, including coughing, wheezing, and shortness of breath [80].

6 Measuring methylation

Illumina inc. is the leading company when it comes to tools used for measuring methylation in DNA form human samples. Their first array, HumanMethylation27k BeadChip could read methylation values of 27,000 CpGs. This tool was improved upon with Illumina HumanMethylation450k BeadChip which read 450,000 CpGs and has been commonly used to investigate DNA methylation in human tissues [23]. This has very recently been replaced by Illumina HumanMethylationEPIC BeadChip (EPIC) covering over 850,000 CpGs.
All three versions of this tool work in the same way, they use bisulfite conversion to observe the methylation status of the CpGs, This method involves exposing to cytosine's to bisulfite and observing what happens to them. The methylated cytosine's will remain unchanged while the unmethylated ones will turn into uracil [45].

7 Epigenetic clocks

An epigenetic clock is a method to measure age based on biochemical data. The method is based on DNA methylation levels and measuring the build up of methyl groups in someones DNA. The main motivation for developing these kinds of clocks are to aid in biological research as age is a very fundamental characteristic for most living organisms. An accurate measure of biological age could be useful for

- testing the validity of various theories of biological aging,
- diagnosing various age related diseases and for defining cancer sub types,
- predicting the onset of various diseases,
- serving as data point when evaluating therapeutic methods including rejuvenation approaches,
- studying developmental biology and cell differentiation,
- forensic applications, for example to estimate the age of a suspect based on blood left on a crime scene.

The main interest for this paper is how an epigenetic clock could be used to measure the impact of chemical exposure on the overall health of a population.

7.1 Age acceleration and reported causes

Age acceleration is calculated by looking at the methylation value of age ralated CpGs and comparing them to the average methylation value of other people of the same age. This tells us if a persons epigenetic age is different from their chronological age. If the difference to the average is a positive number that means that their epigenetic age is older than their chronological one and if the difference is negative that means that their epigenetic age is younger.

7.2 The evolution epigenetic clocks

The relationship between age and DNA methylation has been known since the late 1960s, however the history of epigenetic clocks really kicks off in 2011 when an article was published by a UCLA team that demonstrated that DNA methylation levels in saliva could generate age predictors with an average accuracy of 5.2 years meaning that they predicted age with an error of $\pm$ 5.2 years.

7.3 Evaluation of existing clocks

In order to evaluate the functionality of existing clocks I have compiled a table which gives a basic overview of their functionality and accuracy.

Clock	Correlation	Error	Array	Sample type	Age	Sample size	Algorithm	Clock type
Bocklandt et al. [10]	0.73	5.2 (Avg.Error)	27k	Saliva	21-55	68	Regression	CA-based
Hannum et al. [31]	0.91	4.9 (RMSE)	450k	Blood	19-101	656	Elastic Net	CA-based
Horvath Skin & Blood [36]	0.9-0.95	2.5-3 (MAD)	450k & EPIC	Skin & blood	0-94	905	Elastic Net	CA-based
Horvath Pan-tissue [34]	0.96	3.6 (MAD)	27k & 450k	51 types	0-100	3931	Elastic Net	CA-based
Zhang [89]	0.99	2.04 (RMSE)	EPIC	Blood & saliva	2-104	13566	Elastic Net	CA-based
Alsaleh [3]	0.97	2.6 (MAD)	EPIC	Blood	0-88	527	Elastic Net	CA-based
Alsaleh minimal [3]	0.9	4.6 (MAD)	EPIC	Blood	0-88	527	Elastic Net	CA-based
ABEC [44]	0.95	1.13 (MAD)	EPIC	Blood	19-59	1592	Elastic Net	CA-based
eABEC [44]	0.97	1.25 (MAD)	EPIC	Blood	18-88	2227	Elastic Net	CA-based
cABEC [44]	0.97	1.30 (MAD)	450k & EPIC	Blood	18-88	2227	Elastic Net	CA-based
DeepMAge [27]	0.97	2.77 (MAD)	27k & 450k	Blood		4930	Deep learn	CA-based
Levine PhenoAge [46]			27k, 450k & EPIC	Blood	21-100	9926/456	Elastic Net	PT-based
GrimAge [50]			450k & EPIC	Blood	66 (mean)	1731	Elastic Net	PT-based
AltumAge [48]	0.98	2.071(MAD)	27k & 450k	20 types		8050	Deep learn	CA-based

Table 1: Table comparing different epigenetic clocks

7.4 Data challenges

As with most data sets, the one I will be using is likely to have imperfections and thus present some data challenges.

7.4.1 Normalization

Normalization in machine learning refers to the technique of adjusting a dataset to fit within a pre-specified range, typically between zero and one. While normalization is not a requisite step for every dataset applied in machine learning algorithms, it becomes particularly beneficial when the dataset features vary significantly in their ranges. For instance, suppose we have two features, x and y, where x ranges from 1 to 10 and y spans from 1 to 10,000. In this case, due to its larger numerical values, feature y will exert a greater influence on the predicted outcomes, regardless of its actual predictive power. By normalizing both features to the same range, we can circumvent this issue, ensuring that each feature contributes proportionately to the prediction, thereby facilitating more balanced and accurate model outcomes.

7.4.2 Batch effects

Batch effects refers to non-biological factors that cause changes in the data from when the data is gathered. These factors could for example be differences in capturing times, handling personnel, equipment, laboratory conditions, and technology platforms. [81]

7.4.3 Missing values

Missing values in a dataset are instances where an observation lacks one or more of its attributes. A basic, often-used strategy to handle this is to simply discard the incomplete observations and continue with the data analysis. However, if the dataset contains a substantial number of missing values, discarding these data points might not be a feasible solution due to the potential loss of significant information. Thus, it becomes necessary to explore more sophisticated techniques to address missing data, such as data imputation or predictive modeling based on other available data points. One effective strategy to manage missing values involves normalization, a process that scales numeric attributes to a standard range. This method can substantially reduce the impact of missing values on subsequent analyses by minimizing the variability between different attribute scales, thereby ensuring that each attribute contributes equally to the overall analysis.

7.5 Cell deconvolution

The complex composition of different cell types within a tissue can be estimated by deconvolution [67]. This is important as we already established that cell type

is important to methylation. We can therefore use deconvolution to enhance our cell data to improve the predictions.

7.6 Elastic net

From table 1 we can observe that elastic net is the most commonly used algorithm for developing epigenetic clocks. Elastic Net is a technique for linear regression that blends the L1 (Lasso) and L2 (Ridge) regularization approaches. It seeks to overcome certain limitations inherent in these two methods by incorporating a penalty term that represents a linear mix of the L1 and L2 penalties. This dual approach simultaneously achieves two objectives: it compresses the coefficients towards zero (similar to Ridge regression) and carries out feature selection (akin to Lasso regression). Elastic Net is especially valuable when working with high-dimensional data sets containing a significant number of predictors that may be interrelated.

$$L = \Sigma(\hat{Y}i - Yi)^2 + \lambda\Sigma|\beta|$$

Figure 2: Lasso Formula: Sum of squared errors + Sum of the absolute value of coefficients.

Here, $\Sigma(\hat{Y}i - Yi)^2$ represents the sum of the squared differences between the predicted ($\hat{Y}i$) and actual (Yi) outputs, which is the loss we want to minimize. $\lambda\Sigma|\beta|$ represents the sum of the absolute values of the coefficients, which acts as a penalty term, discouraging large coefficients and thereby controlling overfitting. λ is the regularization parameter which balances the trade-off between the loss and the penalty term.

$$L = \Sigma(\hat{Y}i - Yi)^2 + \lambda\Sigma\beta^2$$

Figure 3: Ridge Formula: Sum of squared errors + Sum of the squares of coefficients.

Like the Lasso formula, $\Sigma(\hat{Y}i - Yi)^2$ represents the sum of the squared differences between the predicted and actual outputs. The penalty term $\lambda\Sigma\beta^2$ is the sum of the squares of the coefficients. This serves to control overfitting by discouraging large coefficients. However, unlike Lasso which tends to select one variable from a group of highly correlated features, Ridge regression will consider all of them. This means it tends to distribute the coefficient values among correlated predictors, instead of assigning a zero coefficient to some of them, as Lasso does. This results in a model that may be better at handling scenarios where predictors have strong correlations with each other.

$$L = \Sigma(\hat{Y}i - Yi)^2 + \lambda_1 \Sigma\beta^2 + \lambda_2 \Sigma|\beta|$$

Figure 4: Elastic Net Formula: Ridge + Lasso.

This formula combines the loss and penalty terms from both Ridge and Lasso regression. This means it not only minimizes the sum of squared differences between predicted and actual outputs, but also imposes both the Lasso and Ridge penalties on the coefficients. The λ_1 and λ_2 parameters control the balance between the Ridge and Lasso penalty terms. This hybrid approach helps handle situations where there are correlations between predictors or where there are more predictors than observations.

In all the above formulas, L is the loss function to be minimized, β represents the coefficients or weights for each predictor in the model, and λ_1, λ_2, and λ are the regularization parameters that control the impact of the penalty terms in the loss function. These formulas provide a mathematical basis for understanding how Elastic Net, Ridge, and Lasso regression control overfitting and make predictions based on the given predictors.

7.7 Deep learning

The other major algorithm in table 1 is the deep learn algorithm. Deep learn is not exactly on specific algorithm but a subset of algorithms which have been inspired by the structure and function of the human brain. The most important aspect of the deep learn algorithm is the layers of nodes which the inputs are fed through each node applying its own weight. The weights of each node is then updated based on the result and this process is repeated to train the model.

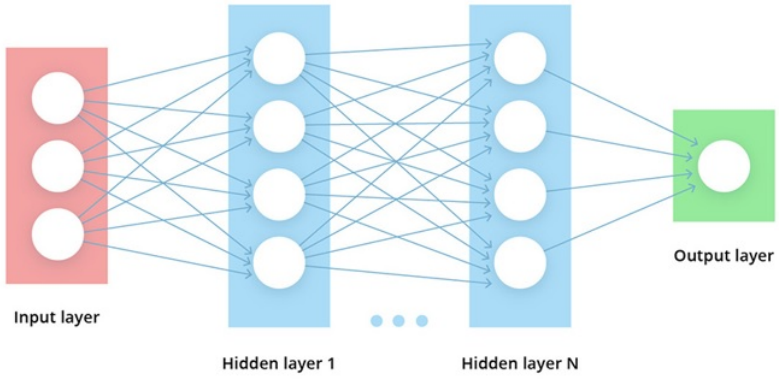


Figure 5: Visualization of layers in a deep learn algorithm [18]

8 Technologies

8.1 Python vs R

Initially when considering this project I assumed that python was going to be the most suitable environment for the job. However learning more about the epigenetics I learnt that R might be a better fit. The reason I assumed that python would be best is because of my previous history with using it for machine learning and other projects. Python is a very good general purpose language is highly versatile this can make exploratory data analysis quite smooth. Python also sports many libraries that help carry out data science and machine learning functions. R also has those same benefits but it seems to have better tools for data processing, plotting and graphing. This means that R is probably better for this project which is very focused on processing and visualising data. In addition an earlier project titled "Building Epigenetic Clocks for Estimating Ageing in Life After" [75] by Kristin Aurora Sydhagen that I am attempting to build on used R to build her which pushes me further towards using R as I won't have to recreate many of the solutions that she has already implemented.

Julia is a high-level, dynamic programming language. It is a more general purpose language when compared to R and python and it is also designed to give users the speed of C/C++. This means that it is a good choice for writing all kinds of applications. It also boasts to be as easy to use as Python. Looking at this comparison we can see some evidence for this speed especially when compared to the worst case scenario for python, the exception to this speed is when writing R-like vectorized code.

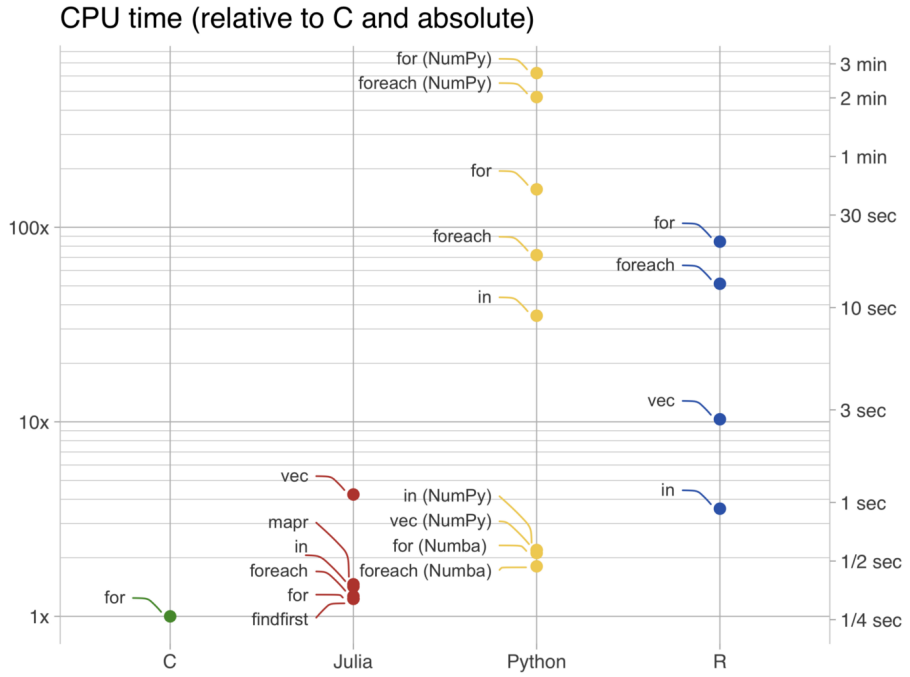


Figure 6: Comparisons of Python, R and Julia [16]

8.2 Data type

The data set that I am going to use are in the IDAT file format. This is a proprietary format of Illumina inc that is used by their scanners [70]. There are two IDAT files for each sample, one for the green channel and one for the red channel. The red channels tracks which locations are methylated and the green locations the unmethylated ones.The main issue with this file format is that due to it's proprietary nature there is a limited amount of tools that can be used to read and process the data. R has the minfi package to process these files, this package has been recommend to me and in my own limited testing it seems to work nicely. Python and Julia also has packages for reading these files however if these packages are not as good that might be a bottleneck that forces me into using the R programming language.

9 Goals

The aim of this project was to analyse existing methods for using methylation data to predict biological age through epigenetic clocks and see how these methods could be used to detect or measure chemical exposure. I wanted to explore how exposure affected the results of Age acceleration from these existing clocks as well as try to use the method that these clocks used to calculate age

to develop my own model for predicting the exposure of subjects. In addition I wanted to explore which CpG's were affected by each chemical to then identify which sections of the genome had been changed by the genome and how that might affect health.

9.1 Aim 1 Analyse how qualitatively and quantitatively exposure has an effect on age acceleration

There is substantial evidence that exposure to harmful chemicals reduce a persons life expectancy through for example. [47]lung cancer related to smoking. I am therefore going to see how these exposures affect the biological age of the subjects and testing the following hypotheses.

H_a : Exposure to chemicals will increase age acceleration in the subject and subjects that are more exposed will show a bigger impact on age acceleration

H_a : Exposure to chemicals will increase age acceleration in the subject equally

H_a : Exposure to chemicals will have little to no impact on age acceleration To test these statements I am going to calculate age acceleration using existing clocks and then plot it against the chemical exposure or in the cases were I only have exposed and not exposed I will create a density plot of the age acceleration of exposed and not exposed.

9.2 Aim 2 Identify CpGs related to exposure of different chemicals

During the development of the genetic clocks researchers found that some CpGs were much more related to aging and that making the clock using only those sites was beneficial for the creation of the clocks [7]. These are the hypotheses I am going to test to investigate whether a similar selection of CpGs can be done when looking at their relation to chemical exposure.

H_a : Some CpG's are clearly more or less relevant to exposure of each chemical

H_a : All CpG's more or less equally relevant to exposure of each chemical

H_a : None of the CpG's are relevant to exposure of each chemical To test the CpGs I will run linear models for each of the CpGs relation to the subjects exposure status individually I will then use the R2 scores of these models to evaluate how much each CpG is impacted by exposure of the chemical. In addition I will look at which CpGs were selected by the elastic net model since this method is also supposed to identify the most important features in the set.

9.3 Aim 3 Develop a model to predict exposure of subject based on methylation signatures

I will test the method of using the same method of using an elastic net model to predict biological age to try and predict the subject chemical exposure. I will also be testing the newer method of using neural networks as well as a linear model for regression models and a Bayesian classifier for classification models as a simple control to see if these more advanced methods provide a significant benefit.

H_a : The model is able to accurately predict chemical exposure

H_a : The model is able to predict chemical exposure but not with a degree of accuracy that is required for such a tool

H_a : The model is not be able to predict chemical exposure at all After training and tuning a variety of models I will measure the accuracy of each model as well as their rate of false positives and negatives where applicable. I will then compare the results to see how all the models performed overall and in relation to each other.

Part II
Method

10 Finding and evaluating additional data sets

Initially, I had a dataset obtained from an article that described the association between exposure to polybrominated biphenyl (PBB) and genome-wide DNA methylation differences in peripheral blood. This dataset formed the foundation for my search for additional datasets to explore, with specific requirements that needed to be met, which were as follows:

- The set should be in the Illumine 850k EPIC array format.
- The set should provide access to raw date in the form of IDAT files.
- The set needs to measure some form of chemical exposure either as how much the subject was exposed or simply as exposed and not exposed.
- The set should provide at least 400 samples to provide a good sample size and make it convenient to split the set into a training set and a validation set.

10.1 Prowling articles on Pubmed and google scholar

To find additional datasets that met these requirements, I started by searching for relevant articles on PubMed and Google Scholar. My aim was to identify datasets that could be helpful in furthering my investigation.

10.2 Available sets on GEO

Another approach I took was to search for available datasets on the Gene Expression Omnibus (GEO), which is a public repository of microarray and next-generation sequencing data. I searched for datasets that were in the Illumina 850k EPIC array format and that measured some form of chemical exposure. After many deliberations I ended up including 450k data as well as sets that did not consist of raw idat files.

10.3 Other sources of data

Lastly, I also considered other sources for potential datasets, such as collaborations with other researchers and contacting authors of relevant articles to request access to their datasets.

Set ID	Tissue type	N	Chemical exposure	Covariates	Array type	Existing age clock
GSE116339 [13]	Whole blood	674	pbb	sex, age	EPIC	Aurora clock [75]
GSE147430 [84]	Whole blood	132	smoking	none	450k	methylclock
GSE50660 [87]	Whole blood	464	smoking	sex, age	450k	methylclock
GSE85210 [85]	Whole blood	253	smoking	none	450k	methylclock
GSE54690 [9]	Whole blood	27	smoking	sex, age, cigarettes per day	450k	methylclock
GSE106648 [49]	Whole blood	279	smoking	sex, age, disease	450k	methylclock
GSE50967 [1]	Whole blood	12	benzene	none	450k	methylclock

Table 2: Data set table

11 Downloading and parsing data

Once I have identified potential datasets, I will download and parse the data using various tools and packages, such as the Python methylprep package [71], the R Biocoductor package, or by downloading zip files directly. I will then preprocess the data using methods such as Noob or Swan to normalize the data and calculate beta values.

11.1 Python methylprep package

The Python Methyprep package is a useful tool for preprocessing and analyzing methylation data. This package provides a variety of functions for quality control, data normalization, and differential methylation analysis. With Methyprep, one can load methylation data from file formats such as CSV and BED, and easily filter out low-quality data points or batch effects. Additionally, Methyprep includes methods for imputing missing data and performing normalization using algorithms like quantile normalization and ComBat. Methyprep also provides functions for identifying differentially methylated regions (DMRs) and genes (DMGs) between different samples or groups, and can output results in an easy-to-interpret format. For me the main function of the methylprep package was to download the IDAT files and generate sample sheets from the GEO. I did consider using this package for further processing of the data but I found to be lacking when handling the large datasets I was tackling.

11.2 R biocoductor package

The R Bioconductor package is a useful tool for downloading data from the Gene Expression Omnibus (GEO), a public repository of functional genomics datasets. Bioconductor provides functions for accessing and downloading data from GEO, allowing researchers to easily access and analyze a wealth of genomic data. With Bioconductor, users can search for specific datasets using keywords or GEO accession numbers, and can download and preprocess data directly within R. Additionally, Bioconductor provides functions for quality control, normalization, and differential expression analysis of GEO data, allowing for comprehensive analysis of gene expression patterns across different experimental conditions.

11.3 Downloading zip files

Downloading and extracting zip files containing methylation data using 7-Zip is a straightforward process. First, navigate to the website or source from which you wish to download the zip file. Once you have located the file, download it to your computer. Once the download is complete, locate the file on your computer and right-click on it. From the drop-down menu that appears, select "7-Zip" and then select "Extract Here" or "Extract to [filename]". This will extract the contents of the zip file to a folder with the same name as the zip file.

Once the extraction is complete, you can access the methylation data within the folder and begin analyzing it.

12 Prepossessing

The R minfi package is a powerful tool for analyzing methylation data. This package is specifically designed to work with data from the Illumina Infinium MethylationEPIC BeadChip and the 450K array [25]. Minfi provides functions for quality control, normalization, and differential methylation analysis [4]. With Minfi, you can preprocess and filter methylation data to remove low-quality probes and batch effects. Additionally, Minfi includes methods for identifying differentially methylated regions (DMRs) and genes (DMGs) between different samples or groups. One of the strengths of Minfi is its ability to work with large datasets, allowing for high-throughput analysis of methylation data.

12.1 Noob

The Noob method is a popular method for normalizing methylation data that is generated by the Illumina Infinium platform [82]. The method stands for "normal-exponential out-of-band," and it involves normalizing the data in several steps. First, the raw signal intensities are corrected for background noise using the normal-exponential model. Next, any non-specific binding effects are corrected using the out-of-band model. Finally, the normalized intensities are adjusted for dye bias, and the resulting values are transformed into beta values. The Noob method is particularly effective at correcting for batch effects and other technical variations that can arise in Illumina methylation data.

12.2 SWAN

The SWAN (Subset-quantile Within Array Normalization) method is a popular method for normalizing methylation data generated by the Illumina Infinium platform. The SWAN method is a type of quantile normalization that aims to remove technical variations between samples and probes by normalizing the methylation values within a given probe type (either type I or type II probes) and between the two different probe types [52]. The SWAN method is effective at correcting for technical variations that arise from probe design and hybridization differences, and is particularly useful for analyzing methylation data that has been generated from different Infinium platforms or different array versions.

12.3 NoobSwan

The NoobSWAN method is a normalization technique that combines the Noob and SWAN methods for normalizing Illumina Infinium methylation data. This method aims to correct for both technical and biological variations in the data. The NoobSWAN method first applies the Noob normalization method to correct

for background noise, non-specific binding effects, and dye bias. Then, the SWAN method is used to correct for technical variations arising from probe design and hybridization differences. The combination of these two methods allows for the removal of technical variations while preserving the biological variation in the data. The NoobSWAN method has been shown to produce more accurate and reproducible results than either Noob or SWAN alone and is becoming increasingly popular for analyzing large-scale methylation datasets.

12.4 Calculating Beta values

The minfi package in R provides functions for analyzing DNA methylation data generated by the Illumina Infinium platform, including the calculation of beta values. The package uses the raw intensity data from the IDAT files generated by the Illumina platform to calculate beta values for each CpG site. The first step in the process is to perform background correction and color balance on the raw data. Next, the raw intensities for each probe type (methylated and unmethylated) are extracted and normalized a method specified by the user in my case the NoobSWAN method to correct for technical variation. Finally, beta values are calculated as the ratio of the methylated intensity to the sum of the methylated and unmethylated intensities, with an adjustment for background intensity. The resulting beta values range from 0 to 1, representing the proportion of methylation at each CpG site.

12.5 PCA analysis

Principal Component Analysis (PCA) is a widely used method for reducing the dimensionality of high-dimensional datasets in various fields, including genomics and bioinformatics. In R, PCA analysis can be performed using the prcomp() function, which computes the principal components of a dataset.

The prcomp() function takes the data matrix as input, and optional arguments can be used to control the centering and scaling of the data. The function returns the principal components of the data as a matrix, where each column represents a principal component and each row corresponds to a sample in the dataset.

To visualize the results of PCA analysis, the ggplot2 package can be used to create a biplot, which displays the scores and loadings of each principal component. The scores correspond to the position of each sample in the reduced-dimensional space, while the loadings correspond to the contribution of each variable (or feature) to each principal component.

PCA analysis can be useful for identifying patterns or clusters in the data, as well as for identifying outliers or sources of variability in the data. It can also be used as a preprocessing step for downstream analysis such as clustering, classification, or differential expression analysis.

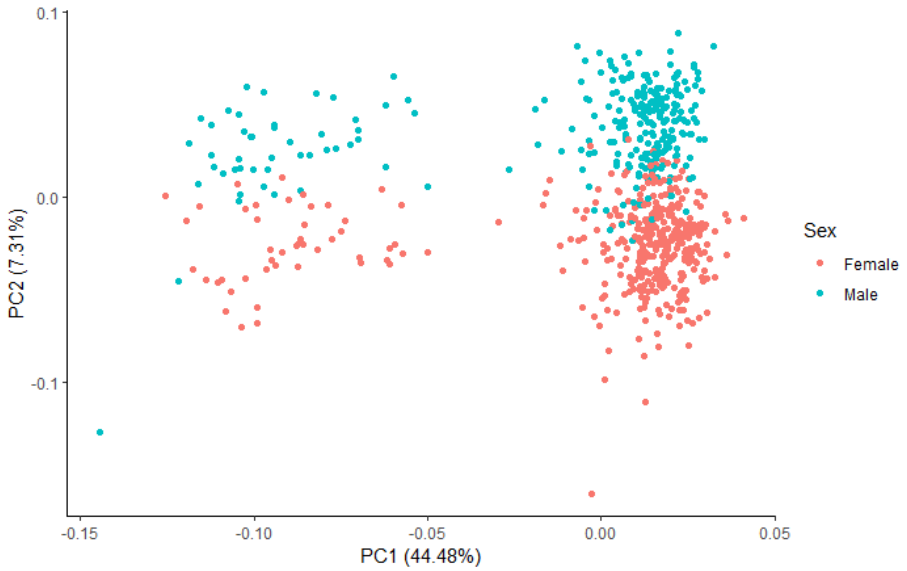


Figure 7: PCA plot for PBB colored for sex

This figure shows a PCA plot of the beta-values from GSE116339, normalized with Noob and SWAN. The colors represents different sex.

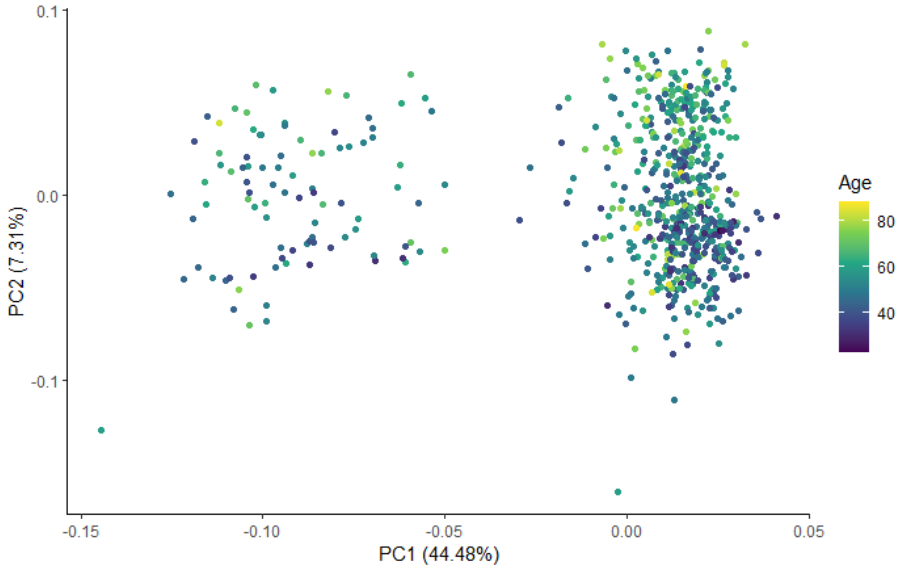


Figure 8: PCA plot for PBB colored for age

This figure shows a PCA plot of the beta-values from GSE116339, normalized with Noob and SWAN. The colors represents different age values.

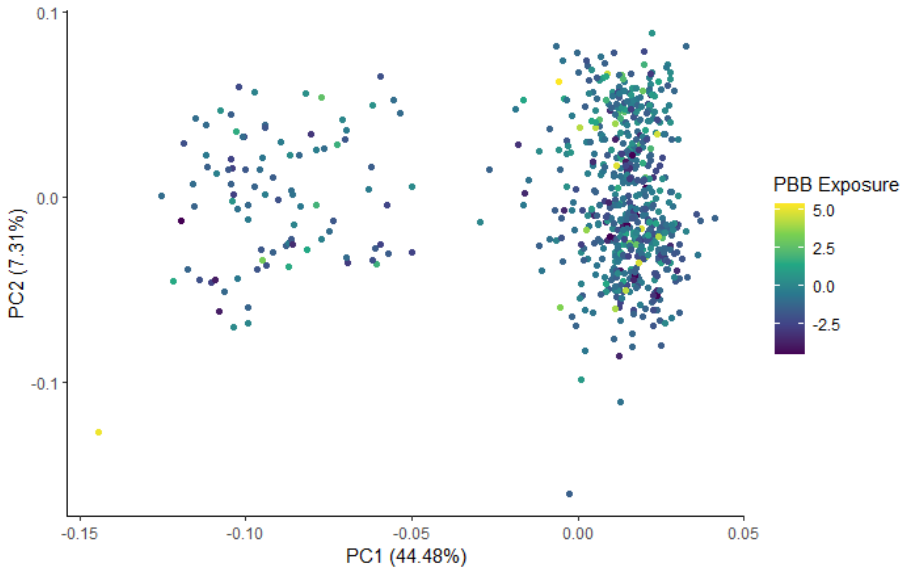


Figure 9: PCA plot for PBB colored for PBB exposure

This figure shows a PCA plot of the beta-values from GSE116339, normalized with Noob and SWAN. The colors represents different values of ln(total pbb).

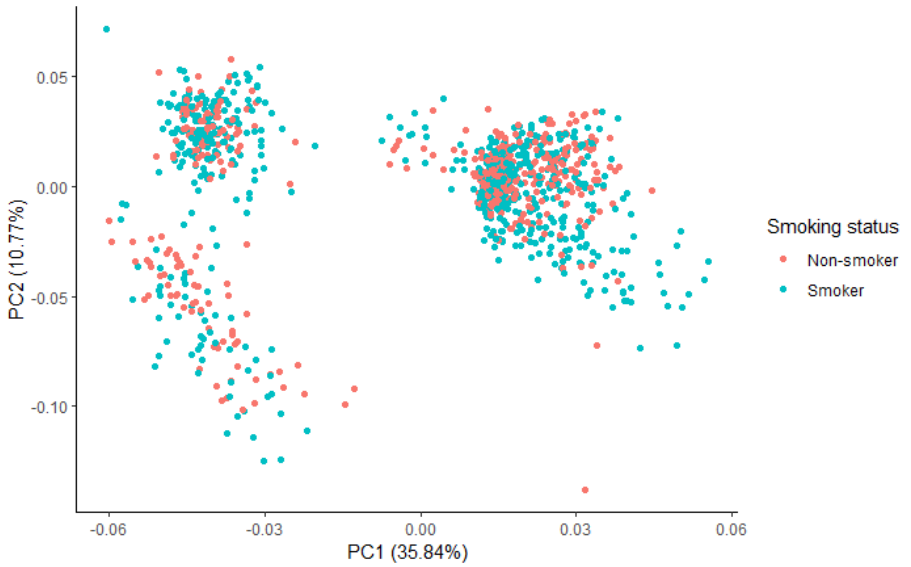


Figure 10: PCA plot for Smoking colored for smoking status

This figure shows a PCA plot of the beta-values from GSE147430, GSE85210, GSE50660, GSE54690 and GSE106648. The colors represents wether the subject is a smoker or not.

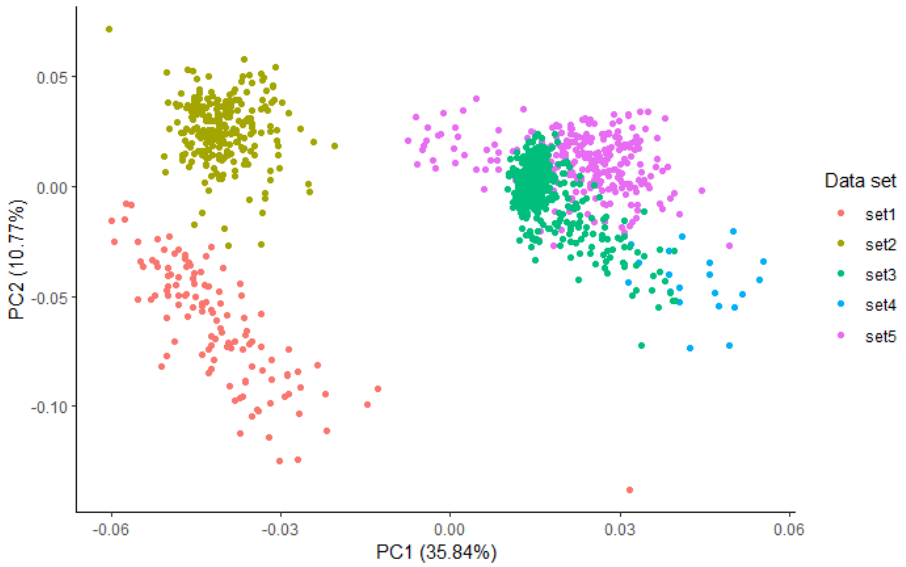


Figure 11: PCA plot for smoking colored for set, set 1: GSE147430, set 2: GSE85210, set 3: GSE50660, set 4: GSE54690 and set 5: GSE106648

This figure shows a PCA plot of the beta-values from GSE147430, GSE85210, GSE50660, GSE54690 and GSE106648. The colors represents what set the subject is a part of.

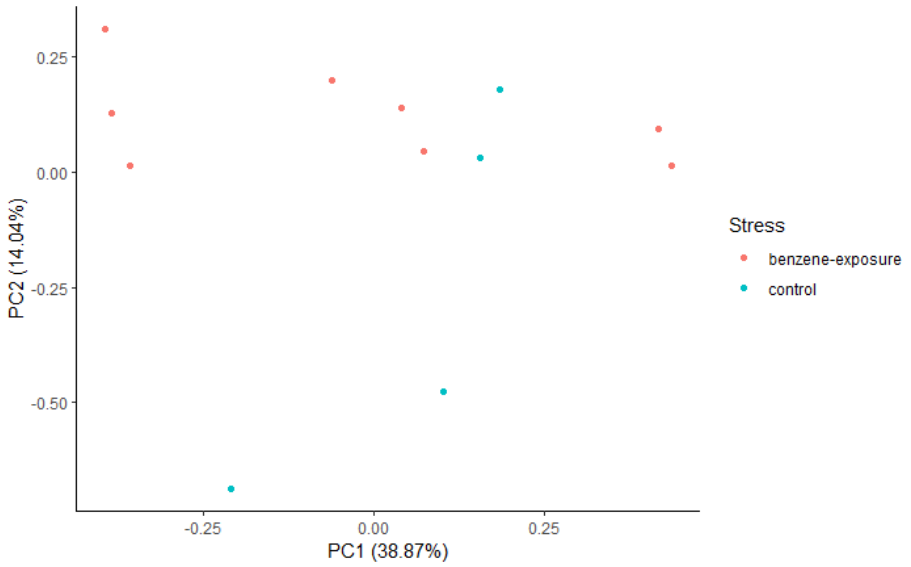


Figure 12: PCA plot for Benzene colored for exposure

This figure shows a PCA plot of the beta-values from GSE50967. The colors represents wether the subject has ben exposed to benzene.

13 Age acceleration

Epigenetic age acceleration is when the biological (epigenetic) age is higher than the chronological age. In other words, based on the DNA methylation patterns, the body appears to be aging faster biologically than would be expected based on the number of years since birth.

Intrinsic and extrinsic epigenetic age acceleration are two specific types of age acceleration:

- **Intrinsic Epigenetic Age Acceleration (IEAA)**: This is a measure of age acceleration independent of changes in blood cell composition, which are known to change with age. It's calculated by adjusting the epigenetic age for measures of blood cell counts. The concept here is to identify the aging speed of the body's cells themselves, separate from the shifts in blood cell populations that occur with age.
- **Extrinsic Epigenetic Age Acceleration (EEAA)**: This measure takes into account changes in blood cell composition, which are known to be reflective of immune system aging. It is calculated by first adjusting the DNA methylation age measure for blood cell counts and then forming an epigenetic measure of immune system age. In simple terms, it's a measure of immune system aging.

So, the key difference between intrinsic and extrinsic age acceleration is whether changes in blood cell composition (reflecting immune system aging) are considered. IEAA focuses on cellular aging regardless of changes in blood cell composition, while EEAA incorporates these changes as part of the aging process.

13.1 Cell deconvolution

In order to calculate extrinsic acceleration we need some measure of blood cell counts. to do this we will use cell deconvolution which is a computational technique used in the analysis of biological data. This process helps in the extraction of signals from complex data sets, such as blood samples, by separating it into its constituent cell types. By using reference data for each cell type, the deconvolution algorithm estimates the proportion of each cell type present in the mixed sample. In the context of extrinsic acceleration, this means we can obtain a more precise measure of blood cell counts, which we can then correlate to the individual's physical condition.

13.1.1 Minfi

For our EPIC data we ill use the minfi package and its estimateCellCounts function. This function is used to estimate the proportion of different types of cells in a mixed cell population using methylation data.

The function works on the basis of reference datasets that contain known cell type methylation profiles. When presented with a mixed sample, it uses these reference datasets to estimate the proportions of each cell type within the sample. This is accomplished by comparing the methylation patterns in the mixed sample to those in the reference datasets.

13.1.2 Meffil

For my 450k data I only have post processed bata values so I could not use the minfi package. Instead I used the mefill package [72] since it has a function to estimate cell counts based on methylation beta values.This package, meffil, has been specifically designed to handle methylation beta values and provides a robust solution for my dataset. It's function, meffil.estimate.cell.counts.from.betas(), uses reference-based cell deconvolution methodology similar to minfi, but designed for beta values, allowing me to estimate cell proportions in my mixed cell population. This alternative approach allowed me to leverage the available post-processed beta values and avoid the need for raw data. By using this function, I was able to gain insights into the cell composition of my samples.

13.2 PBB set

For the PBB set I will be using the Aurora clock [75] to calculate biological age. I will then use minfi as mentioned to calculate cell counts in order to get age acceleration by getting the residuals from linear models of biological age age and biological age age + cell populations.

13.3 Smoking set

For the smoking set I will use the methylclock package to calculate age using its implementation og Zhang et als [89] EN clock. This package in addition to calculating biological age implements the meffil package [72] to calculate cell counts so that it can calculate both intrinsic and extrinsic age acceleration.

14 Model building

To build my models in R I used a package called caret. The caret (Classification And REgression Training) package in R provides a unified framework for building and evaluating predictive models. This model includes all the functions I needed to separate my training data, train my models using K-fold cross-validation and all the different algorithms that I wanted to test out

In order to separate my training data and test data I used the createDataPartition() function from the caret package. this function allowed me to my to create a 70%/30% split between my training set and my test set while maintaining the distribution of exposure of the subjects

14.1 K-fold cross-validation

To tune my models instead of doing this manually i implemented a method called k-fold cross-validation. This technique used in machine learning to evaluate the performance of a model on a limited amount of data.

The basic idea of k-fold cross-validation is to divide the dataset into k subsets of roughly equal size, where k is a positive integer. The model is then trained on k-1 of these subsets, called the training set, and evaluated on the remaining subset, called the validation set. This process is repeated k times, each time using a different subset as the validation set and the remaining subsets as the training set.

At the end of each iteration, the performance of the model is measured by calculating a metric such as accuracy or mean squared error on the validation set. The k performance metrics are then averaged to obtain a single performance estimate for the model.

The advantage of using k-fold cross-validation is that it provides a more reliable estimate of the model's performance than a single train-test split, especially when the dataset is small or imbalanced. It also allows for a more thorough evaluation of the model's ability to generalize to new data.

To go even further I am going to use repeated cross validation which means that the process of splitting the data into K subsets and evaluating the model

is repeated multiple times. In other words, it involves performing K-fold cross-validation multiple times, with different random splits of the data into training and testing sets each time.

The goal of repeated cross-validation is to obtain a more stable estimate of the model's performance by averaging the results of multiple cross-validation runs. This can help to reduce the variance of the estimate and produce a more reliable estimate of the model's true performance.

I ended up using 5 folds repeated 5 times which means that my final result would be based on an estimate of the model's performance would be the average of the results from the 25 cross-validation runs.

14.2 Tuning grid

A tuning grid is a collection of hyperparameter values that are systematically searched to find the best combination for a given model. Hyperparameters are parameters that are not learned during model training, but rather are set by the user before training starts. The performance of a model can be highly dependent on the choice of hyperparameters, so it is essential to optimize them to achieve the best possible results.

In the context of the caret package, a tuning grid is specified using the expand.grid() function, which creates a data frame containing all possible combinations of hyperparameter values. When using the train() function in caret, the model is trained on each combination of hyperparameters in the tuning grid, and the best combination is chosen based on the performance metric specified (e.g., accuracy, mean squared error).

14.3 Different models

14.3.1 Elastic net

For the elastic net model I specifically used the glmnet model. In caret, the glmnet model has two main hyperparameters: α and λ. These hyperparameters control the type and amount of regularization applied in the model. The glmnet model is an elastic net model that combines Lasso (L1) and Ridge (L2) regularization techniques [76].

1. **Alpha (α)**: The alpha hyperparameter controls the mixing of Lasso (L1) and Ridge (L2) regularization in the model. The alpha value ranges from 0 to 1.

    - When $\alpha = 0$, the model becomes a Ridge regression model. In Ridge regression, L2 regularization is applied, which adds the squared val-

ues of the coefficients multiplied by the regularization parameter (lambda) to the loss function. This encourages the model to have smaller coefficients, reducing the risk of overfitting and improving generalization.

- When $\alpha = 1$, the model becomes a Lasso regression model. In Lasso regression, L1 regularization is applied, which adds the absolute values of the coefficients multiplied by the regularization parameter (lambda) to the loss function. This encourages the model to have some coefficients exactly equal to zero, leading to sparse models that can be more interpretable.
- When $0 < \alpha < 1$, the model is an Elastic Net regression model, which combines both L1 and L2 regularization. This can lead to a balance between the sparsity of Lasso regression and the smoothness of Ridge regression, making it useful for datasets with correlated features.

2. **Lambda (λ)**: The lambda hyperparameter determines the amount of overall regularization applied in the model. A larger lambda value results in stronger regularization, which can help prevent overfitting by reducing the complexity of the model. However, setting lambda too high may cause underfitting, as the model becomes too simple to capture the underlying patterns in the data.

For my tuning grid for the alpha values I used a range from 0 to 1 with increments of 0.1. For the lambda value I fit a model using the training set and with alpha = 0.5 and lambda = to 0 to see the minimum value where all coefficients equalled 0 and created a range from 1e-2 to this value across 100 increments.

14.3.2 Neural network

For my neural network I used the nnet model, the nnet model is a single-hidden-layer feedforward neural network implemented using the nnet package [83]. There are two main hyperparameters when using nnet in caret: size and decay.

1. **Size**: The size hyperparameter controls the number of hidden units in the single hidden layer of the neural network. The hidden units are responsible for learning and representing the complex patterns in the input data. Increasing the number of hidden units can enhance the model's capacity to learn complex patterns, which may lead to better performance. However, having too many hidden units can cause overfitting, as the model becomes too complex and starts to capture noise in the data. It is essential to find an appropriate balance by trying different values for the size hyperparameter.

2. **Decay**: The decay hyperparameter controls the amount of weight decay (L2 regularization) applied to the neural network's weights. Weight decay

is a regularization technique that adds a penalty term to the loss function, which is proportional to the sum of the squared values of the weights multiplied by the decay parameter. This encourages the neural network to have smaller weights, which can help prevent overfitting by reducing the model's complexity. Setting the decay parameter too high may lead to underfitting, as the model becomes overly simplified and unable to capture the underlying patterns in the data.

For building my models I went with a tuning grid with size values 1, 3, and 5, and decay values 0, 0.1, and 0.2.

14.3.3 Linear regression

The lm model refers to linear regression, which is a simple linear approach for modeling the relationship between a dependent variable and one or more independent variables. Linear regression does not have any hyperparameters to tune, unlike other more complex models such as neural networks or elastic net.

14.3.4 Bayesian

The naive_bayes model refers to the Naive Bayes classifier, which is a simple probabilistic classification algorithm based on Bayes' theorem with the assumption of independence among the features. When using the Naive Bayes classification in R, there is one primary hyperparameter: laplace.

The laplace hyperparameter controls the Laplace smoothing (also known as additive smoothing) applied to the probabilities. Laplace smoothing is used to avoid zero probabilities for features that have not been observed in the training data for a particular class. By adding a small constant (specified by the laplace parameter) to the observed counts, the probabilities are smoothed, and the model becomes more robust to unseen features. The default value for the laplace parameter in the Naive Bayes classifier is 0, which means no smoothing is applied. However, it is common to use a small positive value (e.g., 1) to avoid zero probabilities. For my tuning grid I used the values 0, 0.5, and 1

15 Model evaluation

Evaluating data models is an important step in the data modeling process, as it helps to determine the effectiveness and accuracy of the model in making predictions or classifications. Depending on the type of model, different evaluation metrics may be used, such as mean squared error, accuracy, precision, or recall. The choice of evaluation metric should be based on the specific requirements of the problem being addressed, as well as the nature of the data and the model. It is also important to consider the potential limitations and assumptions of the model, as well as the quality and representatives of the data used to train and test the model. In this project I had to distinct type of models which is a

15.1 Regression model

- Root Mean Squared Error (RMSE): This is the square root of the MSE. It is a popular metric as it gives more weight to large errors.
- Mean Absolute Error (MAE): This is the average absolute difference between the predicted values and the actual values. It measures the average magnitude of the errors in the predictions.
- R-squared (R^2): This measures the proportion of the variance in the dependent variable that is explained by the independent variables. It ranges from 0 to 1, with higher values indicating a better fit of the model.

15.2 Classification model

When analyzing the classification model it is very useful to create a confusion matrix. A confusion matrix is a table that is often used to evaluate the performance of a classification model. It compares the actual labels of a data set to the predicted labels generated by the model.

The table is usually a square matrix, where each row represents the instances in a predicted class, while each column represents the instances in an actual class. Therefore, the diagonal elements of the matrix represent the instances that were correctly classified by the model, while the off-diagonal elements represent the instances that were misclassified.

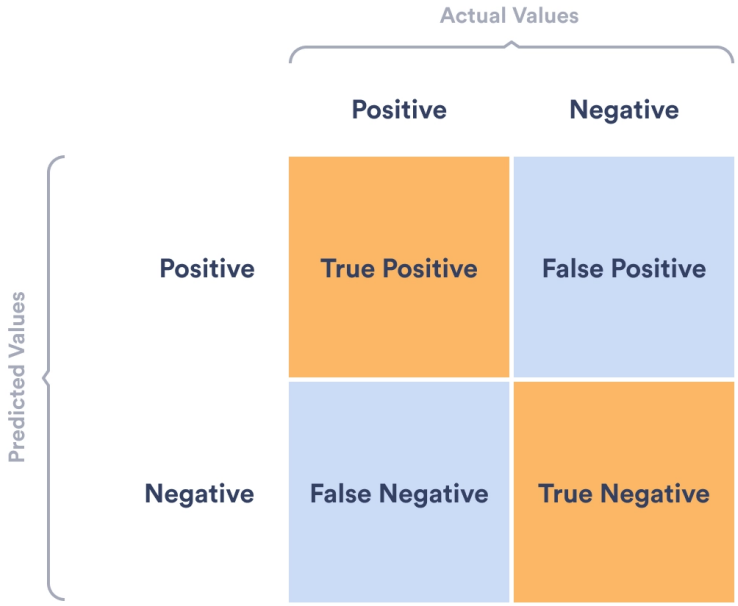


Figure 13: Binary confusion matrix [2]

In this matrix you measure how many of the following classification results occur:

- TP (True Positive) is the number of instances that were actually positive and were correctly predicted as positive by the model.
- TN (True Negative) is the number of instances that were actually negative and were correctly predicted as negative by the model.
- FP (False Positive) is the number of instances that were actually negative but were predicted as positive by the model.
- FN (False Negative) is the number of instances that were actually positive but were predicted as negative by the model.

By analyzing the confusion matrix, you can compute various performance metrics such as accuracy, precision, recall, F1-score, and others, which can help you to understand how well your model is performing on the given data set.

- Accuracy: This is the proportion of correct predictions made by the model.
- Precision: This is the proportion of true positives (correctly predicted positives) out of all positive predictions made by the model.

- Recall: This is the proportion of true positives out of all actual positives in the data set.
- F1 score: This is the harmonic mean of precision and recall, and is a balanced measure that takes into account both precision and recall.

Part III
Results

16 Cell deconvolution

16.1 Cell composition of PBB subjects

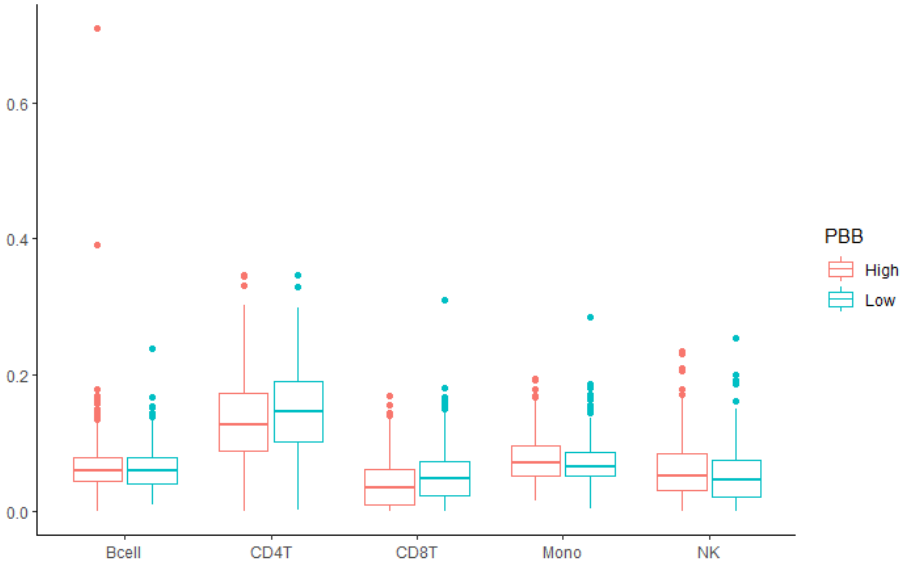


Figure 14: Box Plot of Cell Composition by PBB exposure high and low (higher or lower than the median value)

Upon inspecting the box plot in figure 14 representing the cell composition of subjects exposed to PBB, it appears that both CD4T and CD8T are lower in subjects with higher ln(PBB) values determined by wether the subjects ln(PBB) value weas higher or lower than the median. This observation for CD4T is confirmed by a statistical model depicting the relationship between CD4T and ln(PBB) 15, which shows an estimate of -0.004281. With a p-value of 0.00651, this model suggests a high probability that the observed difference can be attributed to PBB exposure. Similarly, the model 16 for CD8T yields consistent results, with an estimate of -0.0039498 and an even more significant p-value of 3.58e-05, further substantiating the likelihood of PBB causing these differences.

```
Call:
glm(formula = CD4T ~ ln.totalpbb., data = targets)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.143737  -0.046383  -0.003897   0.042794   0.211861

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.136613   0.002680  50.979  < 2e-16 ***
ln.totalpbb. -0.004281   0.001568  -2.729  0.00651 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.003979655)

    Null deviance: 2.7000  on 672  degrees of freedom
Residual deviance: 2.6703  on 671  degrees of freedom
AIC: -1805.5
```

Figure 15: Summary of generalized linear model of CD4T ∼ PBB

```
Call:
glm(formula = CD8T ~ ln.totalpbb., data = targets)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.057553  -0.031218  -0.005655   0.022212   0.261568

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.0433568  0.0016220  26.730  < 2e-16 ***
ln.totalpbb. -0.0039498  0.0009493  -4.161 3.58e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.001457962)

    Null deviance: 1.00353  on 672  degrees of freedom
Residual deviance: 0.97829  on 671  degrees of freedom
AIC: -2481.3
```

Figure 16: Summary of generalized linear model of CD8T ∼ PBB

16.2 Cell composition of smoking subjects

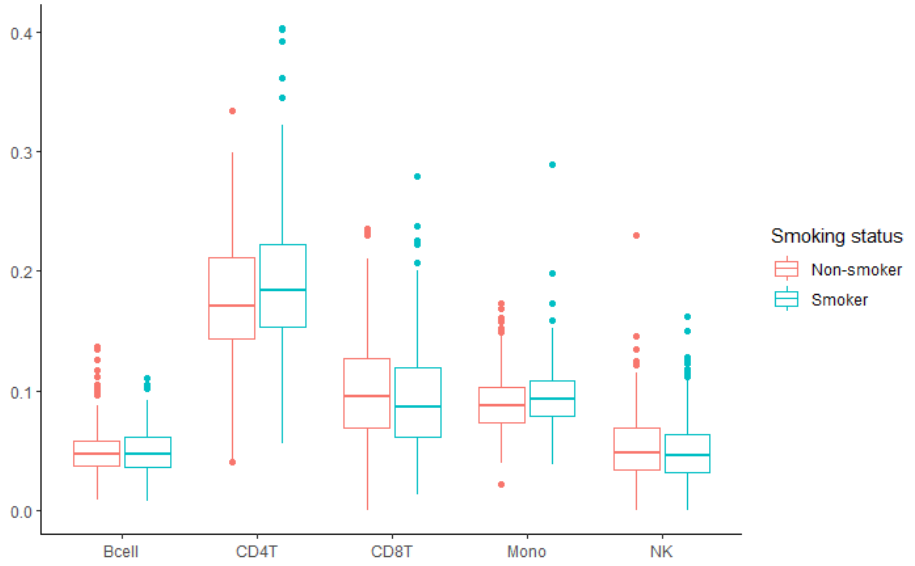


Figure 17: Box Plot of Cell Composition by Smoking

The box plot in figure 17, which represents cell composition for smokers and non-smokers, suggests that the CD4T level is elevated in smokers while CD8T appears to be slightly reduced. The statistical model examining the relationship between CD4T and smoking, shown in figure 18, corroborates this observation. The estimate of 0.013244 suggests that smokers tend to have higher CD4T values, and given the model's p-value of 0.000758, there is a high probability that this difference is not merely coincidental.

In a similar vein, the statistical model analyzing the relationship between CD8T and smoking, displayed in figure 19, shows an estimate of -0.005594. This suggests that smokers have higher CD8T values. Although the p-value of 0.0829 is not as convincing as that of the CD4T model, it still indicates a substantial correlation between CD8T levels and smoking status.

```
Call:
glm(formula = CD4T ~ smoking_status, data = cell)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.135172  -0.035011  -0.005289   0.034670   0.213580

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.175930   0.002984  58.950  < 2e-16 ***
smoking_statusSmoker 0.013244   0.003917   3.381 0.000758 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.002850138)

    Null deviance: 2.2015  on 762  degrees of freedom
Residual deviance: 2.1690  on 761  degrees of freedom
AIC: -2302.2
```

Figure 18: Summary of generalized linear model of CD4T ∼ smoking

```
Call:
glm(formula = CD8T ~ smoking_status, data = cell)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.09941  -0.03139  -0.00622   0.02688   0.18551

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.099413   0.002455  40.494   <2e-16 ***
smoking_statusSmoker -0.005594   0.003222  -1.736   0.0829 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.001928719)

    Null deviance: 1.4736  on 762  degrees of freedom
Residual deviance: 1.4678  on 761  degrees of freedom
AIC: -2600.1
```

Figure 19: Summary of generalized linear model of CD8T ∼ smoking

16.3 Cell composition of benzene subjects

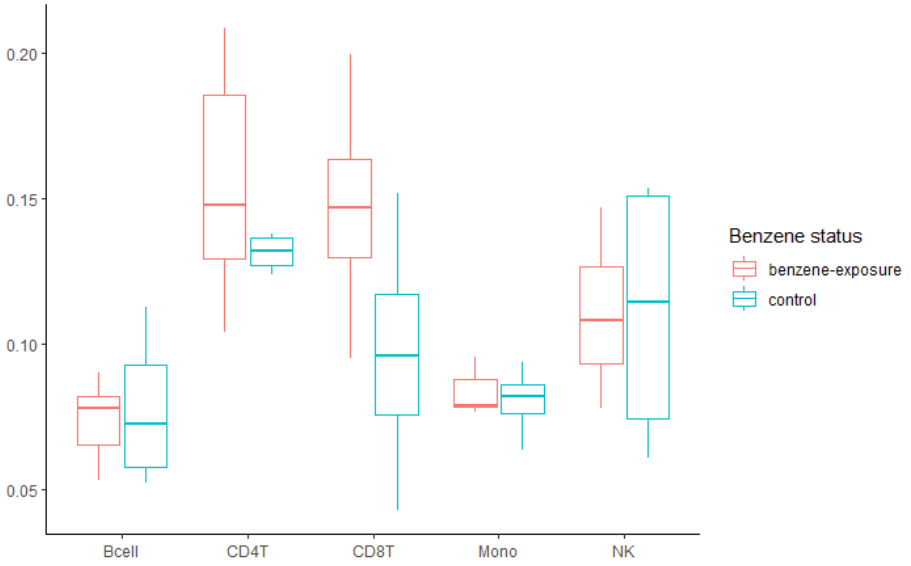


Figure 20: Box Plot of Cell Composition by Benzene exposed subjects and non exposed controls

The plot shown in figure 20, which represents the cell composition of subjects exposed to benzene compared to controls, suggests that both CD4T and CD8T levels are considerably higher in subjects exposed to benzene. This observation is substantiated by the statistical model depicted in figure 21, which analyzes the relationship between CD4T and benzene exposure. Here, an estimate of -0.02352 indicates that the control group tends to have lower CD4T values. However, given the model's p-value of 0.257, it implies that the difference may not be statistically significant.

In the case of the statistical model examining the relationship between CD8T and benzene exposure (figure 22), an estimate of -0.04774 suggests that the control group has lower CD8T values. The model's p-value of 0.0708 indicates a suggestive, though not highly significant, correlation between CD8T levels and benzene exposure.

```
Call:
glm(formula = CD4T ~ benzene_status, data = cell)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.050642  -0.012812  -0.003851   0.011463   0.053470

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.15491    0.01130  13.705  8.3e-08 ***
benzene_statuscontrol -0.02352    0.01958  -1.201    0.257
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.00102215)

    Null deviance: 0.011697  on 11  degrees of freedom
Residual deviance: 0.010222  on 10  degrees of freedom
AIC: -44.763
```

Figure 21: Summary of generalized linear model of CD4T ~ benzene exposure

```
Call:
glm(formula = CD8T ~ benzene_status, data = cell)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.053718  -0.019759   0.002185   0.018998   0.055042

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.14444    0.01364  10.593 9.35e-07 ***
benzene_statuscontrol -0.04774    0.02362  -2.021   0.0708 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.001487315)

    Null deviance: 0.020950  on 11  degrees of freedom
Residual deviance: 0.014873  on 10  degrees of freedom
AIC: -40.263
```

Figure 22: Summary of generalized linear model of CD8T ~ benzene exposure

17 Age acceleration analysis

When analysing age acceleration for the different subjects, we first have to calculate the DNAm age or predicted age for all the subjects and then compare this to the age of the subject. This has to be done a bit differently for each

set as the data sets are in different formats EPIC and 450k. This can then be plotted against the exposure data for each subject. The type of plot used will depend on how exposure is recorded in the data sets.

17.1 PBB age acceleration analysis

For this set since it the data was in EPIC the biological age was calculated using Auroras' clock [75]. Then since our data for PBB was recorded using a numeric value (on ln scale) to represent how much PBB each samples subject had in their body it was possible to run a scatter plot of age acceleration vs PBB exposure.

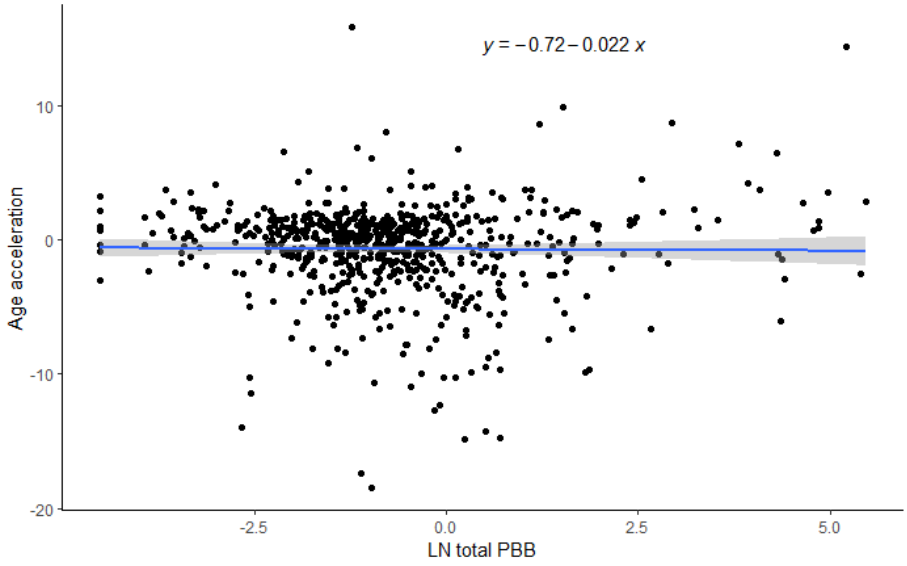


Figure 23: Scatter plot of PBB exposure and age acceleration of subjects

```
Call:
lm(formula = ageAcceleration ~ ln.totalpbb., data = targets)

Residuals:
     Min       1Q   Median       3Q      Max
-17.7433  -1.2489   0.5981   1.8279  16.4862

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.71742    0.14577  -4.922 1.08e-06 ***
ln.totalpbb. -0.02240    0.08531  -0.263    0.793
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.432 on 671 degrees of freedom
Multiple R-squared:  0.0001028, Adjusted R-squared:  -0.001387
F-statistic: 0.06896 on 1 and 671 DF,  p-value: 0.7929
```

Figure 24: Summary of linear model of age acceleration ~ PBB

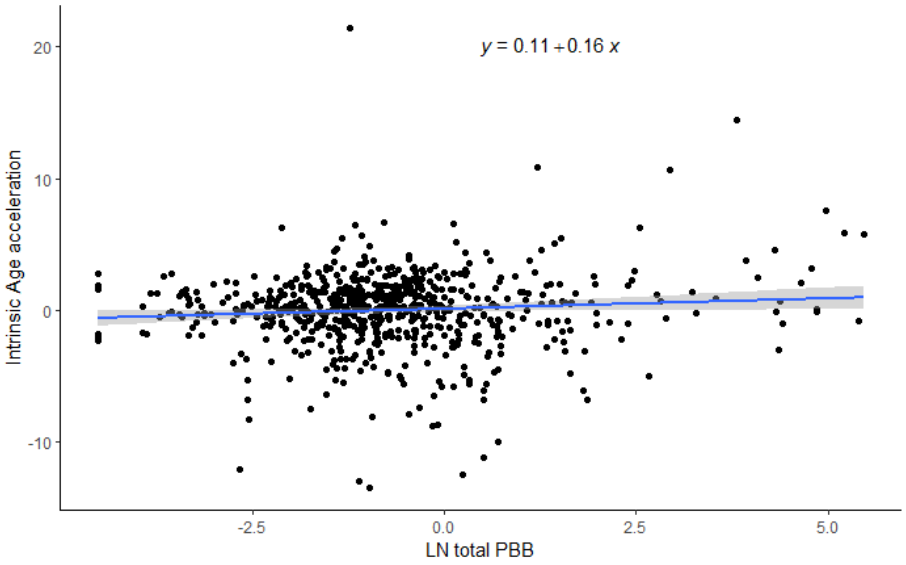


Figure 25: Scatter plot of PBB exposure and Intrinsic age acceleration of subjects

```
Call:
lm(formula = intAgeAcceleration ~ ln.totalpbb., data = targets)

Residuals:
     Min       1Q   Median       3Q      Max
-13.4102  -1.2241   0.2729   1.5116  21.4649

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.11242    0.12288   0.915   0.3606
ln.totalpbb.  0.15658    0.07192   2.177   0.0298 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.893 on 671 degrees of freedom
Multiple R-squared:  0.007015,  Adjusted R-squared:  0.005535
F-statistic:  4.74 on 1 and 671 DF,  p-value: 0.02981
```

Figure 26: Summary of linear model of intrinsic age acceleration ~ PBB

To use a scatter plot to find correlation between PBB and age acceleration, we first need to plot the data points. If the points are tightly clustered around a line that is sloping upwards from left to right, then we can say that the two variables have a positive correlation. This means that as the values of one variable increase, so do the values of the other variable.

Conversely, if the points are tightly clustered around a line that is sloping downwards from left to right, then we can say that the two variables have a negative correlation. This means that as the values of one variable increase, the values of the other variable decrease.

However when looking at both our plots 23,24 we can observe that the points are scattered randomly with no apparent pattern, when fitting a linear model we get a p-value of 0.7929 24 which indicates that PBB does not have an effect on acceleration. The same is not true of our linear model for intrinsic age acceleration 26 which yields a p-value of 0.0298.

17.2 Smoking age acceleration analysis

When looking at the smoking set a different clock had to be used because the data sets analyzed were in the 450k format. Therfore Aurora's clock [75] would not be good way of calculating the methylation age. The clock used here to calculate biological age is the elastic net clock included in the methylclock package [15]. Then since the smoking subjects are either smokers or non-smokers this was plotted using density plots one for smoker and one for non-smokers. Since our populations are similar in size the the resulting difference in plots should provide meaningful insight.

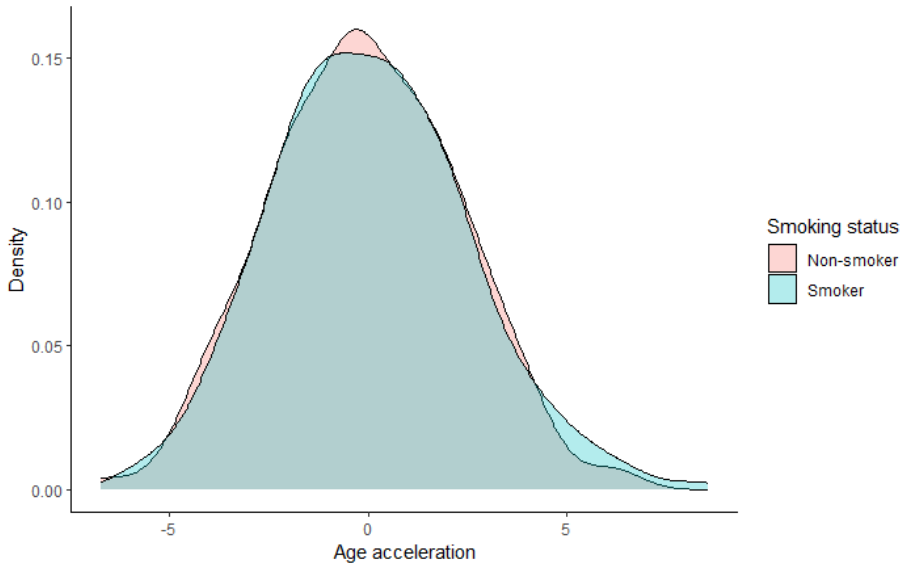


Figure 27: Density graph of age acceleration for smoking (green) and non-smoking (red) subjects

```
Call:
lm(formula = AgeAcceleration ~ as.factor(smoking_status), data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-6.6531 -1.7168 -0.0374  1.6615  8.5454

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   -0.07257    0.13578  -0.534    0.593
as.factor(smoking_status)Smoker  0.12498    0.17819   0.701    0.483

Residual standard error: 2.429 on 761 degrees of freedom
Multiple R-squared:  0.000646,  Adjusted R-squared:  -0.0006672
F-statistic: 0.4919 on 1 and 761 DF,  p-value: 0.4833
```

Figure 28: Summary of linear model of age acceleration $\sim$ smoking

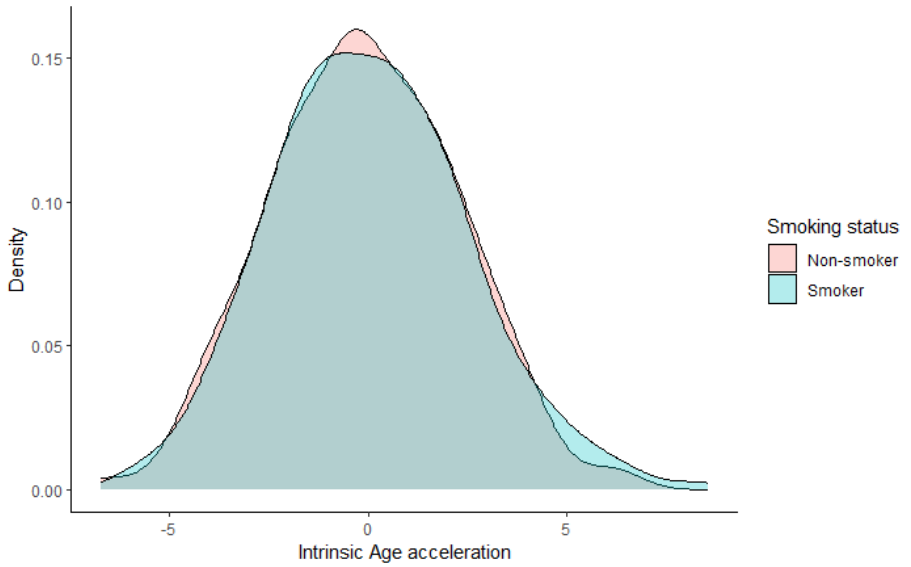


Figure 29: Density graph of intrinsic age acceleration for smoking and non smoking subjects

```
Call:
lm(formula = IntAgeAcceleration ~ as.factor(smoking_status),
    data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-6.8330 -1.7047 -0.1182  1.6409  9.1352

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   -0.09786    0.13292  -0.736    0.462
as.factor(smoking_status)Smoker  0.16855    0.17444   0.966    0.334

Residual standard error: 2.378 on 761 degrees of freedom
Multiple R-squared:  0.001225,  Adjusted R-squared:  -8.71e-05
F-statistic: 0.9336 on 1 and 761 DF,  p-value: 0.3342
```

Figure 30: Summary of linear model of intrinsic age acceleration $\sim$ smoking

A density plot is a graphical representation of the distribution of a continuous variable. In the case of age acceleration, a density plot can show the distribution of age acceleration values for smokers and non-smokers separately.

If there is a difference in the shape of the two density plots27,29, it may suggest that smoking has an impact on age acceleration. For example, if the density

plot for smokers is shifted to the right (i.e., has higher age acceleration values) compared to the density plot for non-smokers, it suggests that smoking may be associated with higher age acceleration.

However, when looking at plot 27,29, we observe no drastic differences in the shape between the density plots there. Additionally when looking at statistical tests 28,30 we observe p-values of 0.4833 and 0.3342 respectively.

18 Relevant CpGs

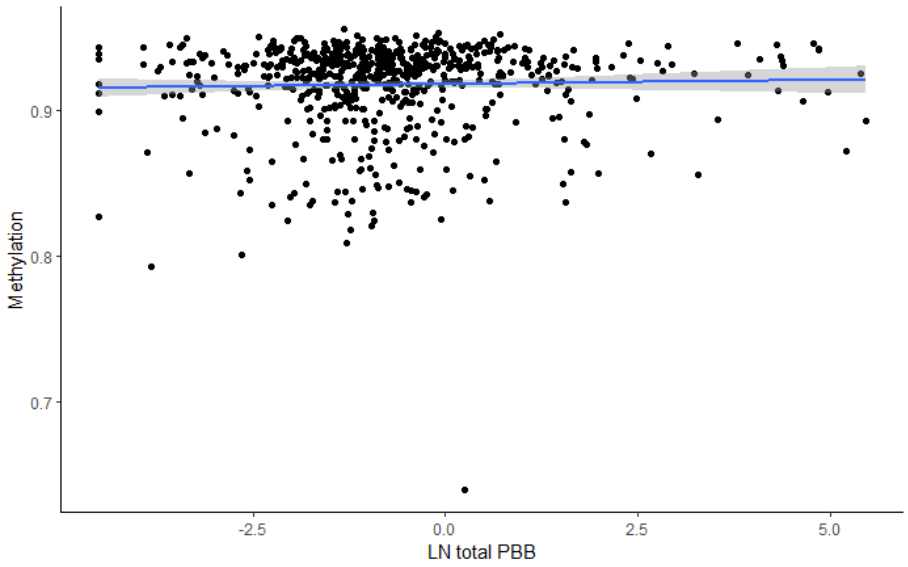
Initially, the objective was to compare the relevant CpGs identified by the Elastic Net model and my method to see if both techniques produced similar outcomes. To achieve this, I retrieved the CpGs from the Elastic Net model that were assigned non-zero weights. Since my method maintains all CpGs in the ranking process without exclusion, I specifically chose to examine the top CpGs, rounded up to the nearest ten based on their R2 scores, to match the quantity selected by the Elastic Net model. This approach was designed to see if there were any overlaps between the CpGs chosen by the Elastic Net and those selected by my method.

Moreover, I decided to scrutinize the top two CpGs chosen by each method for every stressor, with the aim to evaluate their correlation with the target variable. When selecting the best two CpGs from my method, I referred to those with the highest R2 scores, which is consistent with the ranking principle used for the Elastic Net model. When it comes to the CpGs chosen by the Elastic Net model, we specifically examined the predictor with the most substantial absolute weight, as the weights can be either negative or positive. Therefore, it is more suitable to evaluate them based on their distance from zero rather than purely on their magnitude.

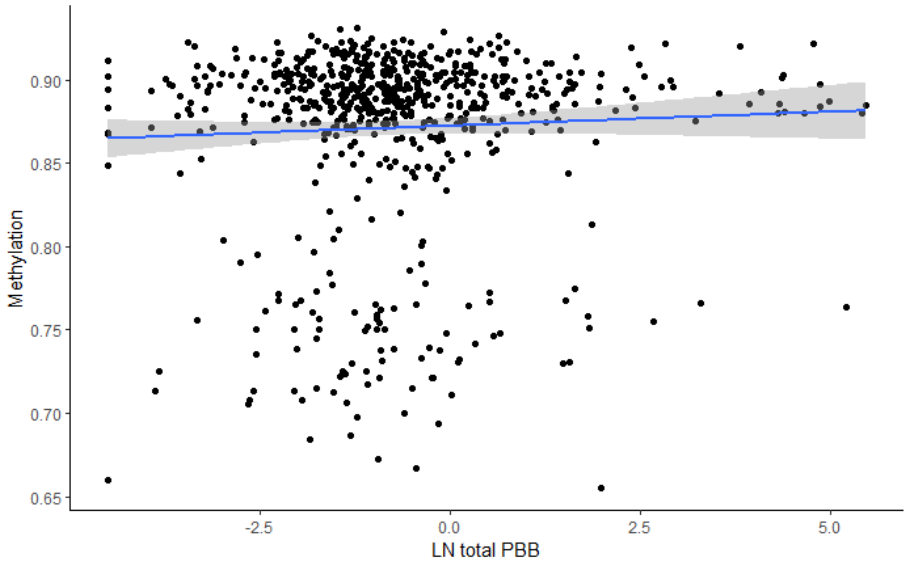
18.1 Relevant CpGs for PBB

Upon examining the top CpGs chosen by both my method and the Elastic Net for predicting PBB exposure, it's evident that there are only eighteen CpGs common to both selections, as depicted in figure 31. When two distinct models select different predictors for a data model, it suggests that they have recognized different variables as being most crucial or influential in forecasting the model's outcome.

Figure 31: Venn diagram showing overlap of CpGs selected for predicting PBB exposure



(a) Scatter plot of methylation of CpG cg19859270 and PBB exposure

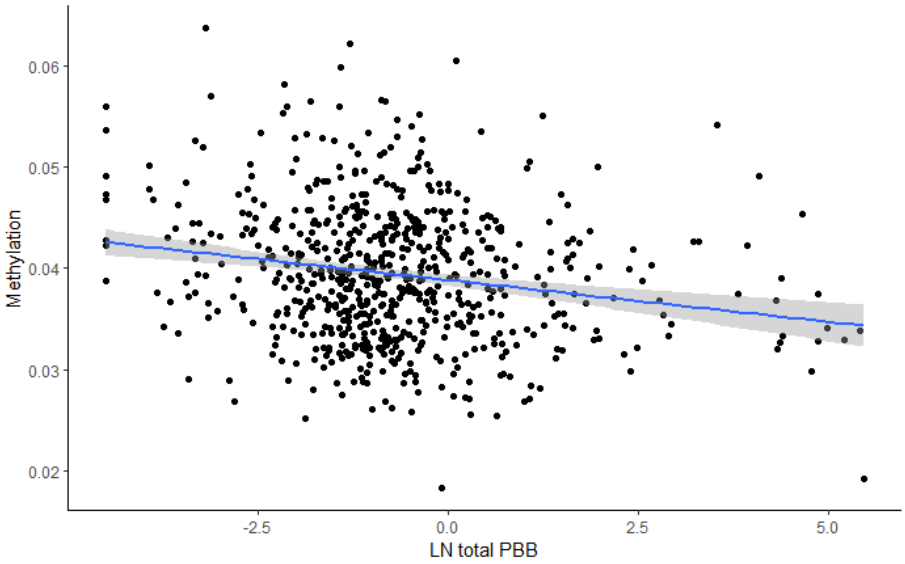


(b) Scatter plot for methylation of CpG cg0265716 and PBB exposure

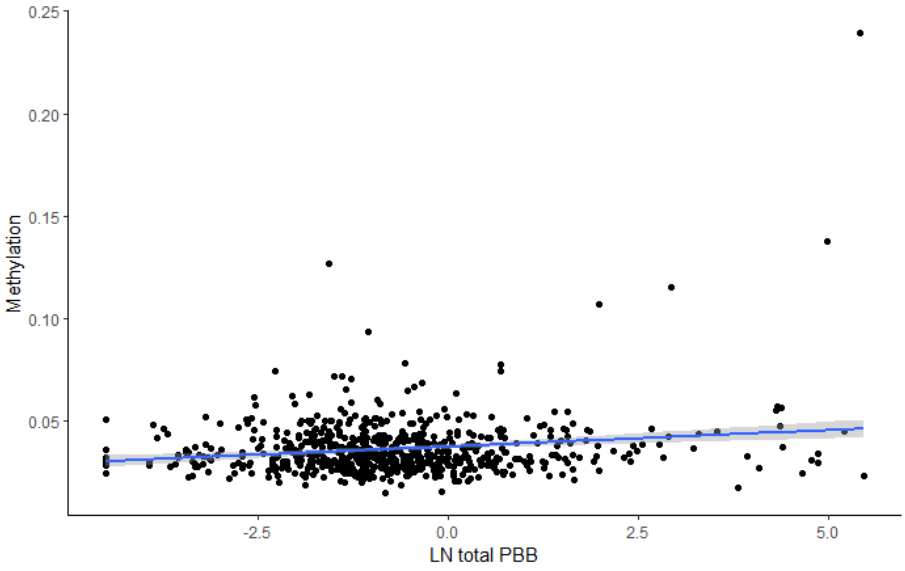
Figure 32: Scatter plots for PBB exposure and methylation for the top two CpGs chosen by my method

These figures show the plots of total PBB (transformed by natural logarithm) against the methylation (percentage) for the top two CpGs as selected by my method for identifying relevant CpGs. Figure 32a shows cg19859270 specifically

and figure 32b shows cg0265716.



(a) Scatter plot for methylation of CpG cg04158069 and PBB exposure



(b) Scatter plot for methylation of CpG cg18108008 and PBB exposure

Figure 33: Scatter plots for PBB exposure and methylation for the top two CpGs chosen by elastic net

CpG	R2	Slope
cg19859270	0.393	-0.036
cg02657160	0.285	-0.043

(a) Table with metrics for the top two CpGs chosen by my method for PBB

CpG	EN weight
cg04158069	-6.590
cg18108008	3.836

(b) Table with metrics for the top two CpGs chosen by elastic net for PBB

Table 3: Tables with metrics for the top two CpGs chosen by elastic net and my method for PBB

The table 3 indicates:

- **CpG site cg19859270:**
  - *R2*: The coefficient of determination (R2) for this site is 0.393, suggesting that about 39.3% of the variability in methylation in this CpG site can be explained by PBB.
  - *Slope*: The slope is -0.036, indicating a negative correlation between PBB and methylation meaning that higher PBB values would lead to lower methylation values.
- **CpG site cg02657160:**
  - *R2*: The R2 value for this site is 0.285, implying that roughly 28.5% of the variability in methylation in this CpG site can be explained by PBB exposure.
  - *Slope*: The slope of -0.043 suggests a negative correlation between PBB an methylation at this CpG site.

For the CpGs chosen by the elastic net for PBB, the weights are as follows:

- **cg04158069**: The EN weight for this CpG site is -6.590. The negative weight suggests that an increase in the PBB exposure would result in a decrease methylation level at this CpG site.
- **cg18108008**: The elastic net (EN) weight for this site is 3.836. The positive weight implies that an increase in the PBB exposure would lead to a a increase in methylation at this CpG site.

18.2 Relevant CpGs for smoking

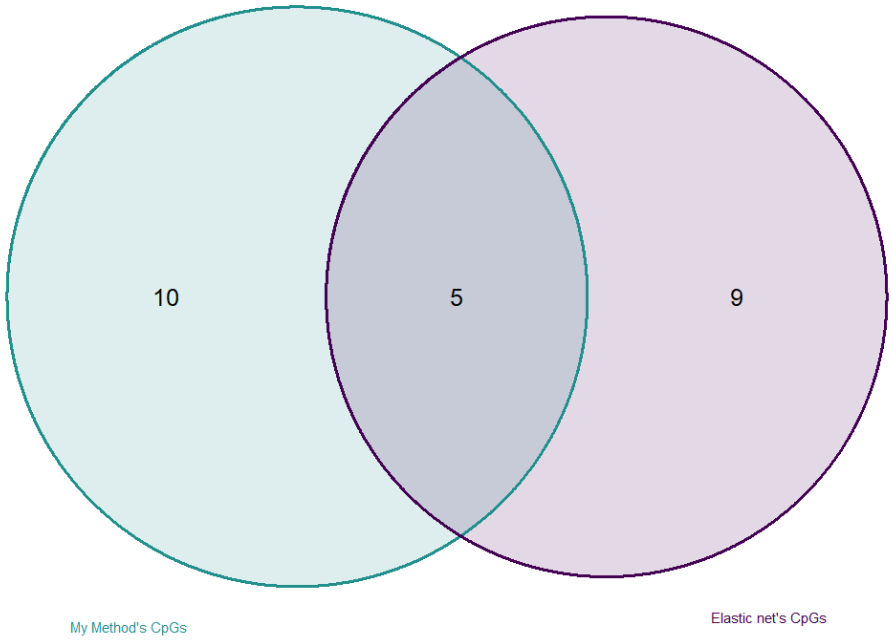
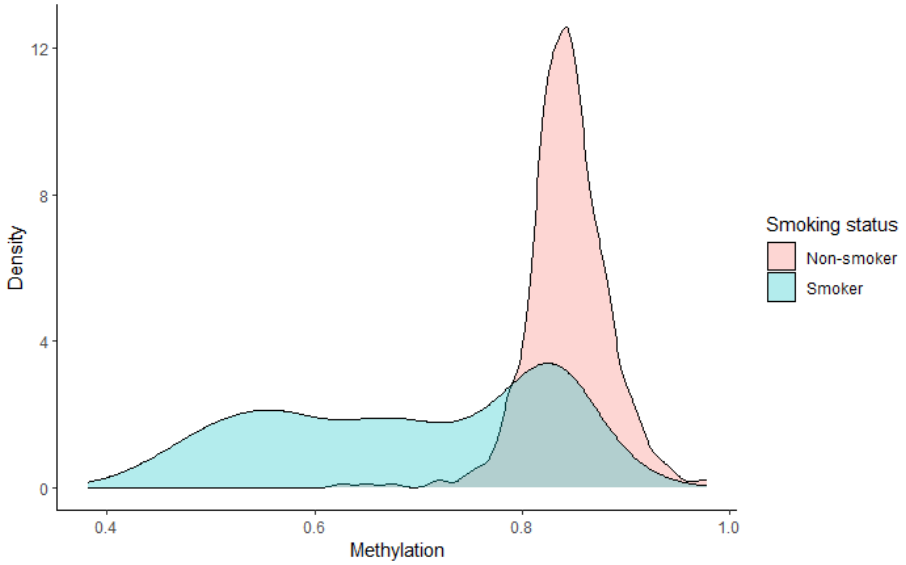
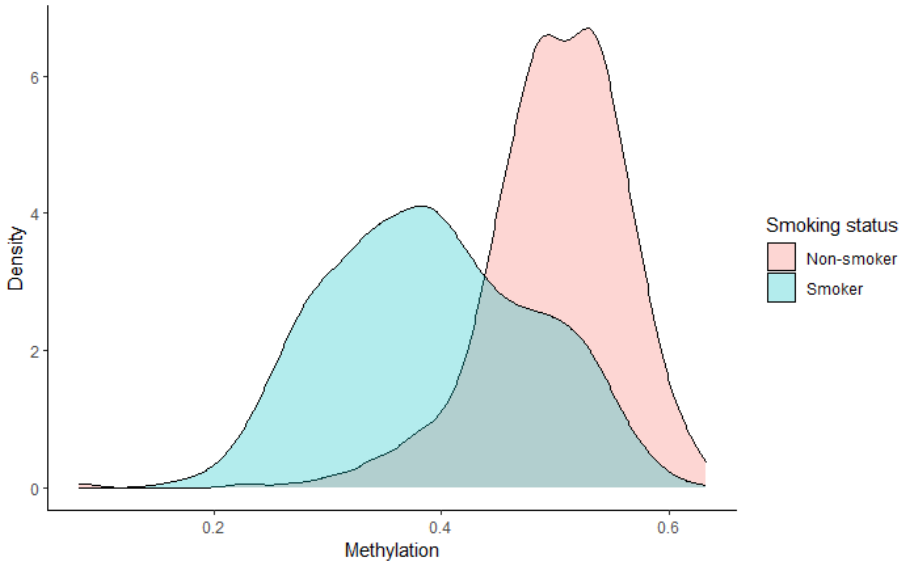


Figure 34: Venn diagram showing overlap of CpGs selected for predicting smoking status



(a) Density plot for methylation of CpG ccg05575921 and smoking status



(b) Density plot for methylation of CpG cg21566642 and smoking status

Figure 35: Density plots for smoking status for the top two CpGs chosen by both elastic net and my method

CpG	R2	Slope	EN weight
cg05575921	0.33	-0.15	-6.67
cg21566642	0.31	-0.11	-3.93

Table 4: Table with metrics for the top two CpGs chosen by both elastic net and my method for smoking

The table 4 presents the metrics for the top two CpG sites chosen by both elastic net regression and my alternative method for their association with smoking. The metrics provided include R2, slope, and elastic net (EN) weight. Let's analyze each CpG site's metrics:

- **CpG site cg05575921:**
  - *R2*: 0.33 indicates that 33% of the variance in methylation beta values can be explained by the model. This suggests a moderate level of association between this CpG site's methylation status and smoking.
  - *Slope*: -0.15 implies that as smoking exposure increases, the methylation beta value at this site tends to decrease. The negative slope indicates an inverse relationship between smoking and methylation at this specific site.
  - *EN weight*: -6.67, a negative value, highlights that this CpG site has a strong association with smoking according to the elastic net regression model. The more negative the value, the stronger the association.
- **CpG site cg21566642:**
  - *R2*: 0.31 shows that 31% of the variance in methylation beta values can be accounted for by the model, suggesting a moderately strong association between methylation at this CpG site and smoking.
  - *Slope*: -0.11, similar to the first CpG site, demonstrates an inverse relationship between smoking and methylation at this site. The decrease in methylation is less pronounced compared to cg05575921, but still noteworthy.
  - *EN weight*: -3.93, although less negative than cg05575921, still indicates a significant association with smoking according to the elastic net regression model.

In summary, both CpG sites (cg05575921 and cg21566642) show a moderate association with smoking, as evidenced by their R2 values. Both sites also exhibit an inverse relationship between methylation and smoking, with cg05575921 having a slightly stronger effect. Elastic net regression weights further confirm the association of these CpG sites with smoking, with cg05575921 having a stronger association than cg21566642.

18.3 Relevant CpGs for Benzene

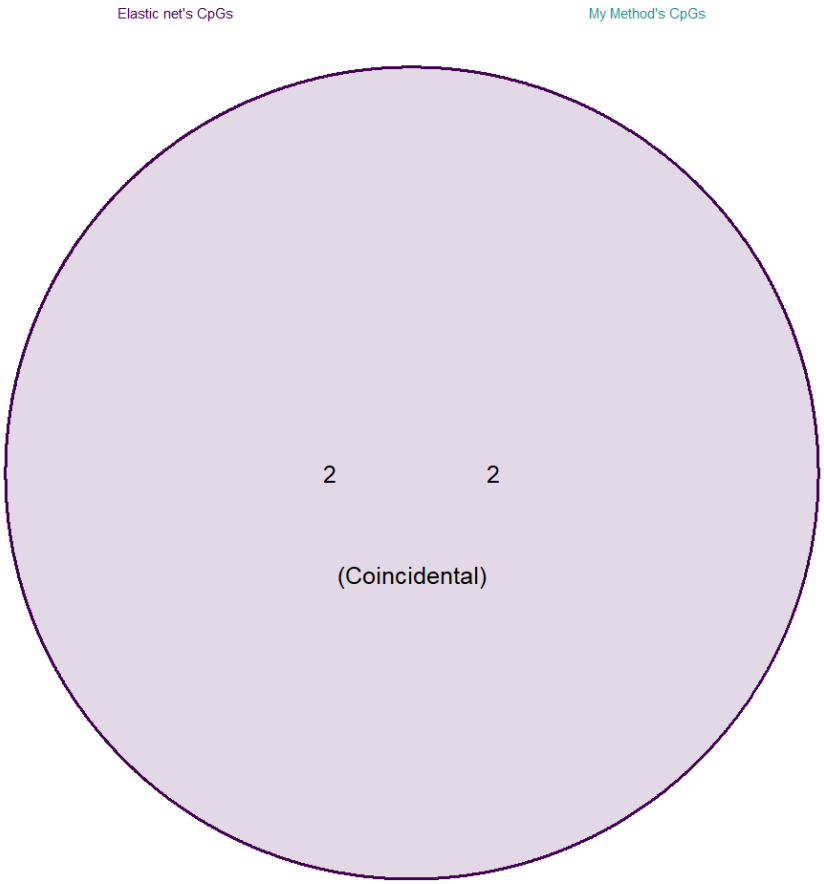
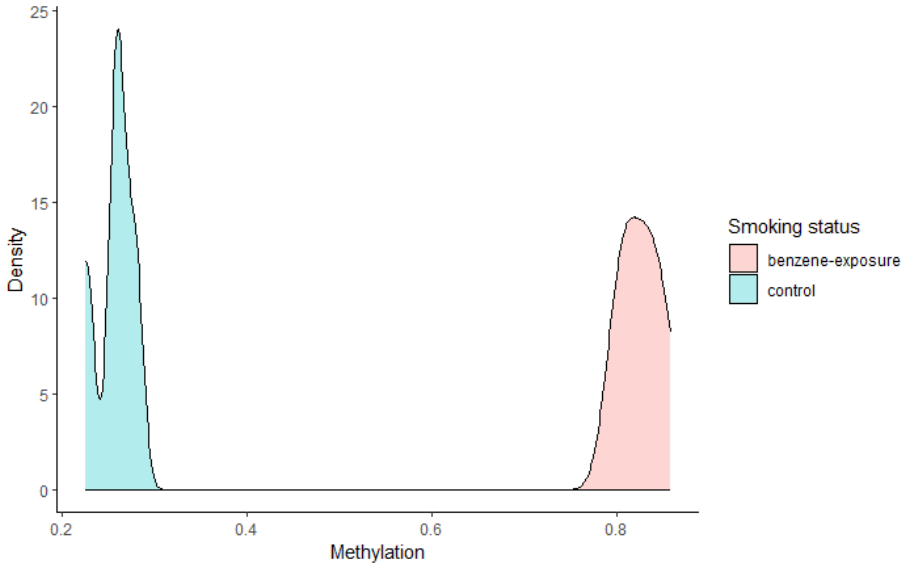
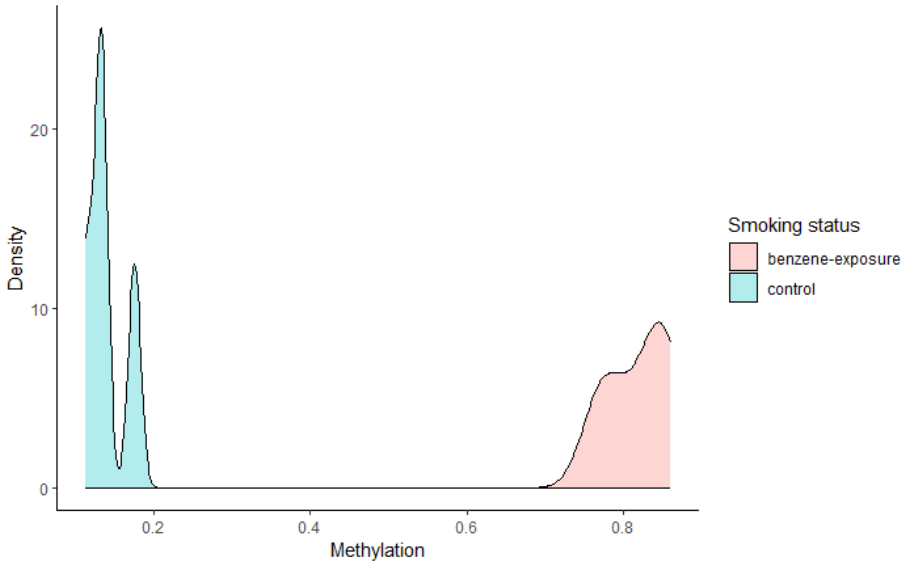


Figure 36: Venn diagram showing overlap of CpGs selected for predicting Benzene exposure



(a) Density plot for methylation of CpG cg07156839 and benzene exposure status



(b) Density plot for methylation of CpG cg20139683 and benzene exposure status

Figure 37: Density plots for benzene exposure status for the top two CpGs chosen by both elastic net and my method

CpG	R2	Slope	EN weight
cg07156839	0.994	-0.569	-7.822
cg20139683	0.990	-0.679	-3.709

Table 5: Table with metrics for the top two CpGs chosen by both elastic net and my method

The table 5 presents the metrics for the top two CpG sites chosen by both elastic net regression and an alternative method for their association with benzene exposure. The metrics provided include R2, slope, and elastic net (EN) weight. Let's analyze each CpG site's metrics:

- **CpG site cg07156839:**
  - *R2*: 0.995 indicates that 99.5% of the variance in methylation beta values can be explained by the model. This suggests a very strong level of association between this CpG site's methylation status and benzene exposure.
  - *Slope*: -0.569 implies that as benzene exposure increases, the methylation beta value at this site tends to decrease. The negative slope indicates a direct relationship between benzene exposure and methylation at this specific site.
  - *EN weight*: -7.822, a negative value, highlights that this CpG site has a strong association with benzene exposure according to the elastic net regression model. The more negative the value, the stronger the association.
- **CpG site cg20139683:**
  - *R2*: 0.997 shows that 99.7% of the variance in methylation beta values can be accounted for by the model, suggesting an extremely strong association between methylation at this CpG site and benzene exposure.
  - *Slope*: -0.679, like the first CpG site, demonstrates an inverse relationship between benzene exposure and methylation at this site. The decrease in methylation is more pronounced compared to cg07156839.
  - *EN weight*: -3.709, although less negative than cg07156839, still indicates a significant association with benzene exposure according to the elastic net regression model.

In summary, both CpG sites (cg07156839 and cg20139683) show an extremely strong association with benzene exposure, as evidenced by their R2 values. The cg07156839 site exhibits a direct relationship between methylation and benzene exposure, while the cg20139683 site demonstrates an inverse relationship. Elastic net regression weights further confirm the association of these CpG sites with benzene exposure, with cg07156839 having a stronger association than cg20139683.

Model	RMSE	Rsquared	MAE
MEN	1.43	0.19	1.02
MEN1k	2.33	0.057	1.99
MEN100	1.33	0.30	0.99
MEN10	1.38	0.26	1.02
MNN1k	1.72	0.088	1.36
MNN100	1.73	0.17	1.36
MNN10	1.73	0.18	1.36
MLR1k	3.56	0.01	2.48
MLR100	1.55	0.16	1.18
MLR10	1.37	0.25	1.01

Table 6: The error metrics of the different PBB prediction models
The model column refers to which algorithms and predictors were used, MEN: Elastic net all CpGs, MEN1k: Elastic net top 1000 CpGs, MEN100: Elastic net top 100 CpGs, MEN10: Elastic net top 10 CpGs, MNN1k: neural net top 1000 CpGs, MNN100: neural net top 100 CpGs, MNN10: neural net top 10 CpGs, MLR1k: logistic regression top 1000 CpGs, MLR100: logistic regression top 100 CpGs and MLR10: logistic regression top 10 CpGs.

19 PBB model evaluation

1. **RMSE (Root Mean Square Error):** This measures the average squared difference between the predicted and actual values. Lower values indicate a better model fit. In this case, the MEN100 model has the lowest RMSE value (1.33), indicating that it has the best fit among the tested models.

2. **R-squared (R^2):** This represents the proportion of the variance in the dependent variable that is predictable from the independent variable(s). R^2 values range from 0 to 1, with higher values indicating a better model fit. The MEN100 model has the highest R-squared value (0.30), meaning it explains 30% of the variance in the data, which makes it the best model among the tested models based on this metric as well.

3. **MAE (Mean Absolute Error):** This measures the average absolute difference between the predicted and actual values. Like RMSE, lower values indicate a better model fit. The MEN100 model has the lowest MAE value (0.99), suggesting that it has the best fit among the tested models based on this metric too.

Overall, based on the given error metrics, the MEN100 model appears to be the best performing model for predicting PBB exposure using DNA methylation data. However, it's important to note that the highest R-squared value (0.30) suggests that there is still a significant portion of the variance in the data unexplained by the model. Further refinement or exploration of other modeling techniques might be necessary to improve the prediction accuracy.

20 Smoking model evaluation

In order to evaluate the performance of various smoking prediction models, several key metrics were utilized: Balanced Accuracy, Precision, Recall (Sensitivity), F1 Score, Specificity, and Negative Predictive Value (NPV). These metrics provide a comprehensive understanding of each model's ability to accurately predict both positive and negative cases, while minimizing false positives and false negatives. The values for each of these metrics, for every model under consideration, are detailed in the table below. Please note that higher values for each metric indicate superior performance. The models were constructed using different algorithms and varying numbers of CpGs. The detailed results are as follows:

1. **Balanced Accuracy:** This metric considers both sensitivity and specificity and is particularly useful when dealing with imbalanced datasets. Higher values indicate better performance. The MEN1k model has the highest balanced accuracy (0.8448), making it the best model in terms of this metric.

2. **Precision:** This measures the proportion of true positive predictions out of all positive predictions made. Higher values indicate better performance. The MNN100 model has the highest precision (0.7692), suggesting it has the best ability to correctly identify positive cases while minimizing false positives.

3. **Recall (Sensitivity):** This measures the proportion of true positive predictions out of all actual positive cases. Higher values indicate better performance. The MNB10 model has the highest recall (0.9071), suggesting it has the best ability to identify positive cases in the dataset.

4. **F1 Score:** This is the harmonic mean of precision and recall, providing a single metric that considers both false positives and false negatives. Higher values indicate better performance. The MEN1k model has the highest F1 score (0.8173), making it the best model based on this metric.

5. **Specificity:** This measures the proportion of true negative predictions out of all actual negative cases. Higher values indicate better performance. The MNN1k model has the highest specificity (0.8209), suggesting it has the best ability to identify negative cases in the dataset.

6. **Negative Predictive Value (NPV):** This measures the proportion of true negative predictions out of all negative predictions made. Higher values indicate better performance. The MNB10 model has the highest NPV (0.9133), indicating the best ability to correctly identify negative cases while minimizing false negatives.

Model	Balanced Accuracy	Precision	Recall	F1	Specificity	Neg Pred Value
MEN	0.8155	0.7301	0.8500	0.7855	0.7811	0.8820
MEN1k	0.8448	0.7640	0.8786	0.8173	0.8109	0.9056
MEN100	0.8173	0.7484	0.8286	0.7864	0.8060	0.8710
MEN10	0.8180	0.7346	0.8500	0.7881	0.7861	0.8827
MNN1k	0.7854	0.7447	0.7500	0.7473	0.8209	0.8250
MNN100	0.8390	0.7692	0.8571	0.8108	0.8209	0.8919
MNN10	0.8302	0.7469	0.8643	0.8013	0.7960	0.8939
MNB1k	0.6552	0.5621	0.6786	0.6149	0.6318	0.7384
MNB100	0.7340	0.6391	0.7714	0.6990	0.6965	0.8140
MNB10	0.7944	0.6649	0.9071	0.7674	0.6816	0.9133

The model column refers to which algorithms and predictors were used, MEN: Elastic net all CpGs, MEN1k: Elastic net top 1000 CpGs, MEN100: Elastic net top 100 CpGs, MEN10: Elastic net top 10 CpGs, MNN1k: neural net top 1000 CpGs, MNN100: neural net top 100 CpGs, MNN10: neural net top 10 CpGs, MNB1k: naive bayes top 1000 CpGs, MNB100: naive bayes top 100 CpGs and MNB10: naive bayes top 10 CpGs.

Table 7: The error metrics of the different smoking prediction models

Part IV
Discussion

21 Development and accuracy of prediction models

I trained ten prediction models for PBB exposure and ten additional moddels for smoking habits. When looking at the precision metrics of RMSE for PBB my model ranged from an RMSE of 1.33-3.56 this range is a bit skewed by my two worst performing models with an RMSE of 3.56 and 2.33 and if we disregard these outliers the range of RMSE becomes 1.33-1.73

For smoking habits effect we look at the models balanced accuracy for performance and the models I trained ended up with an accuracy within the range of 0.8448-0.6552.

21.1 Exposure to PBB

Looking at the RMSE, which measures the average squared difference between predicted and actual values, we find that the MEN100 model (Elastic Net with the top 100 CpGs) has the lowest value of 1.33. This suggests that, on average, the MEN100 model's predictions deviate less from the actual values, indicating a better fit than the other models.

The R-squared value represents the proportion of variance in the dependent variable that can be predicted from the independent variables. In this case, the MEN100 model also has the highest R-squared value of 0.30. While this is the best among the tested models, it is important to note that it's relatively low in absolute terms, indicating that the model accounts for only 30% of the variance in the data. This suggests that there are other factors influencing PBB exposure that are not captured by the top 100 CpGs.

The Mean Absolute Error (MAE) measures the average absolute difference between predicted and actual values, with lower values indicating a better fit. Once again, the MEN100 model performs the best, with an MAE of 0.99.

Overall, the MEN100 model appears to have the best performance based on the provided metrics. However, despite its relative success, further work is needed to improve these models. The relatively low R-squared value for even the best model when comparing to the existing prediction models for predicting genetic age as seen in table 1 suggests that a significant portion of the variance in PBB exposure remains unexplained. This could involve using more or different CpGs, employing different algorithms, or incorporating additional types of data beyond DNA methylation.

21.2 Smoking

The MEN1k model, which uses the Elastic Net algorithm and the top 1000 CpGs, performed exceptionally well according to two of the metrics: Balanced

Accuracy and F1 Score. The high Balanced Accuracy of 0.8448 suggests that this model performs well overall, with a good balance between sensitivity and specificity. This is particularly important in imbalanced data sets This is good because in my smoking set of 1137 samples 467 of them are smokers which is significantly less than half. The high F1 Score of 0.8173 indicates that the model has a strong balance between Precision and Recall, suggesting it can reliably identify both positive and negative cases and minimize errors.

However, the MEN1k model did not outperform all other models in every metric. The MNN100 model, which uses a Neural Net algorithm with the top 100 CpGs, had the highest Precision of 0.7692. This suggests that, while it may not have the highest overall accuracy or balance between positive and negative predictions, it is particularly good at minimizing false positives.

The MNB10 model, which uses the Naive Bayes algorithm and only the top 10 CpGs, showed the highest Recall and Negative Predictive Value. With a Recall of 0.9071, this model is the most successful at identifying positive cases, but it might come with a higher rate of false positives. Its high Negative Predictive Value of 0.9133 shows that it is also strong at correctly identifying negative cases and minimizing false negatives.

Interestingly, the MNN1k model, which uses the Neural Net algorithm and the top 1000 CpGs, showed the highest Specificity of 0.8209. This indicates its strength in correctly identifying negative cases, which is crucial in certain contexts.

22 Effect of stressors on age acceleration

Aurora's clock [75] and methylclock [15] implementation of Zhang [89] were used to analyse age acceleration for the different stressors Aurora's clock for PBB and Zhang for smoking. All hypotheses were tested with both extrinsic and intrinsic measures of Age Acceleration for which I examined the variations in age acceleration.

22.1 PBB exposure

The scatter plots shown in figures 23 and 24 provide a visual representation of the relationship between PBB exposure and age acceleration. Ideally, if there were a strong positive or negative correlation between the two, the data points would be closely grouped along a line sloping upwards (for a positive correlation) or downwards (for a negative correlation). This would indicate that as one variable increases, the other does too (or conversely, decreases).

In this case, however, the data points appear to be scattered with no visible pattern, suggesting a lack of correlation between PBB exposure and age accel-

eration. This is further substantiated by the p-value of 0.7929 obtained from the linear model 24. In statistical terms, a p-value this high provides weak evidence against the null hypothesis that there is no relationship between the two variables. As such, based on this analysis, it would appear that PBB exposure does not significantly influence extrinsic age acceleration.

However, a different story emerges when examining intrinsic age acceleration. The p-value for this model 26 is 0.0298, which is less than the conventional threshold of 0.05 for statistical significance. This suggests that PBB exposure might be associated with intrinsic age acceleration, indicating a need for further investigation.

These findings provide preliminary evidence suggesting that exposure to PBB might influence the biological aging process, at least as measured by intrinsic age acceleration. I found that PBB exposed had lower values of both CD4T and CD8T and these are both taken into account when calculating intrinsic age acceleration which might cause the improved result.

22.2 Smoking habits

When inspecting of the density plots referred to as figures 27 and 29, it becomes apparent that there are no significant differences in the shapes of these distributions between smokers and non-smokers. If smoking had a considerable impact on age acceleration, we might anticipate a noticeable shift in the distribution of one group compared to the other. For instance, if the distribution of age acceleration values for smokers were shifted to the right (i.e., depicting higher values), this would suggest that smoking could be associated with increased age acceleration.

However, the observed similarity in the distributions suggests that the age acceleration does not differ dramatically between smokers and non-smokers, at least not in a way that could be discerned from these density plots. This visual interpretation is further supported by the p-values of 0.4833 and 0.3342 from the statistical models 28 and 30 respectively. In a statistical context, these p-values are well above the conventional significance level of 0.05, indicating that we fail to reject the null hypothesis that there is no difference in either intrinsic or extrinsic age acceleration between smokers and non-smokers.

This suggests that, according to the data and methods employed in this study, smoking may not have a pronounced effect on age acceleration. However, this does not negate the vast body of evidence that indicates smoking has numerous other detrimental health effects [20, 21, 32, 58, 63]. Furthermore, this observation is specific to the measure of age acceleration used in this thesis which uses predicted age calculated through Zhang clock [89] and does not rule out potential effects of smoking detected by other methods or measures of biological aging.

23 Cell deconvolution

Since the tissue cell composition of the samples was already computed for use in intrinsic age acceleration I was additionally able to investigate the difference ins in cell type proportion for subjects exposed to PBB, smoking and benzene compared non exposed subjects. PBB exposed to a lower amount of PBB seemingly had a lower amount of both CD4T and CD8T cells than the subjects with low exposure. For smoking I observed higher CD4T values in smokers and lower CD8T values when comparing to non non smokers. Benzene exposure observed significantly higher Values of both CD4T and CD8T in comparison to the control subjects.

23.1 PBB exposure

The analysis of the cell composition in subjects exposed to PBB, as presented in figure 14, reveals some interesting patterns. Both CD4T and CD8T cell levels seem to be lower in subjects with higher ln(PBB) values. This finding is intriguing because CD4T and CD8T cells are integral components of our immune response. This is supported by the article [14] that the set is from which surmised that PBB affected CpGs related to immune function and autoimmune disease. Additionally I found an article [51] that found that PBB negatively affected the immune system in rats.

Statistical models (15 and 16) provide further support for these observations. The CD4T model shows an estimate of -0.004281 with a p-value of 0.00651, implying a strong correlation between higher PBB exposure and reduced CD4T levels. Similarly, the CD8T model presents an estimate of -0.0039498 with a very significant p-value of 3.58e-05, suggesting a similar trend.

In terms of health implications, it's essential to understand the distinct roles that CD4T and CD8T cells play in our immune system. CD4T cells, commonly referred to as helper T cells, are instrumental in managing the immune response. They do not directly destroy infected cells, but rather signal other immune cells to do so. Therefore, they are a key component in a coordinated, effective immune response. [90]

On the other hand, CD8T cells, also known as cytotoxic T cells, are an integral part of our immune system. They actively seek out and destroy infected cells, as well as cancer cells. They are our body's primary defense against internal threats such as viral infections and tumor growth [68].

Consequently, lower levels of these two critical cell types could potentially impair our immune response [40]. A diminished helper T cell population could lead to a disorganized or less effective immune response, while a reduction in cytotoxic T cells could weaken our body's ability to directly combat infections and malignancies. This implies that individuals with reduced CD4T and CD8T

cell levels could potentially be more susceptible to infections and, potentially, cancer development. This is also supported by findings that show that PBB has been shown to be carcinogenic in animals [57].

23.2 Smoking habits

The data suggests that smoking has a significant impact on the cellular composition of blood, specifically on CD4T and CD8T cell levels.

Firstly, the CD4T cell level appears to be higher in smokers than in non-smokers. This is supported by the statistical model 18 with an estimate of 0.013244, suggesting an elevation in CD4T cell levels in smokers. The model's p-value of 0.000758, a value well below the threshold of 0.05 for statistical significance, reinforces the likelihood that this observation isn't merely due to chance but rather points to a genuine effect of smoking on CD4T cell levels.

Conversely, the CD8T cell levels in smokers seem to be slightly reduced. The statistical model 19 provides an estimate of -0.005594, hinting at a decrease in CD8T cell levels in smokers. However, the model's p-value of 0.0829, while not below the commonly accepted significance threshold of 0.05, still suggests a potential relationship between CD8T cell levels and smoking, albeit less convincing than the correlation with CD4T cell levels.

In terms of health implications, these changes in CD4T and CD8T cell levels might affect the immune response. CD4T cells, or helper T cells, play a crucial role in orchestrating the immune response [90], and their increase in smokers may indicate an effort by the immune system to counteract the harmful effects of smoking. On the other hand, CD8T cells, or cytotoxic T cells, are responsible for killing infected cells and cancer cells [68], and their decrease in smokers might impair the immune system's ability to combat infections or malignancies effectively.

23.3 Benzene exposure

The analysis of cell composition in subjects exposed to benzene, as depicted in figure 20, reveals noticeable differences in the levels of CD4T and CD8T cells compared to non-exposed controls. Both CD4T and CD8T levels appear to be substantially higher in benzene-exposed subjects.

Upon examining the relationship between CD4T levels and benzene exposure through a statistical model (figure 21), we find an estimate of -0.02352. This value suggests that CD4T levels tend to be lower in the control group. However, with a p-value of 0.257, the statistical significance of this relationship is questionable. In other words, the observed variation in CD4T levels might be due to factors other than benzene exposure,

Similarly, the statistical model for CD8T levels (figure 22) indicates an estimate of -0.04774, signifying lower CD8T levels in the control group. The p-value of this model is 0.0708, which, while not achieving conventional thresholds of statistical significance, suggests a possible correlation between CD8T levels and benzene exposure.

From a health perspective, it's important to note that CD4T and CD8T cells are fundamental elements of our immune system. The observed increase in these cell types in benzene-exposed individuals may be indicative of the body's attempt to initiate a robust immune response to counteract the potential harm caused by benzene exposure.

However, one must consider the potential adverse consequences of such an immune response. While an initial increase in CD4T and CD8T cells may be beneficial, a prolonged state of heightened immune activity could result in immune system dysregulation [26]. This could lead to persistent inflammation, a condition that has been associated with various health complications [54].

Chronic inflammation, as a result of continuous immune activation, is a known risk factor for several serious health conditions, including cardiovascular diseases and cancer. It is therefore crucial that we consider the potential long-term health implications of benzene exposure on the immune system. These findings are also corroborated by this paper on benzene-associated immunosuppression and chronic inflammation in humans [30].

24 Specific CpGs for investigated stressors

24.1 PBB exposure

The two CpGs chosen by elastic net were also found by the article [14] that published the dataset that I used for my analysis which heavily supports that these sites are affected by PBB.

24.1.1 cg19859270

The CpG site cg19859270 has an R-squared value of 0.393, suggesting that approximately 39.3% of the variability in its methylation can be attributed to PBB exposure. Furthermore, the slope of -0.036 indicates an inverse relationship between PBB exposure and methylation at this site. As such, increases in PBB exposure are associated with decreases in methylation values at this CpG site. However the magnitude of the slope is rather small as methylation values are between 1 and 0 a slope value of -0.036 Indicates that this change in methylation is likely not significant

24.1.2 cg04158069

The CpG site cg02657160 has an R-squared value of 0.285, indicating that about 28.5% of the variability in its methylation can be accounted for by PBB exposure. The slope of -0.043 further corroborates this relationship, showing a negative correlation between PBB exposure and methylation at this site. This suggests that as PBB exposure increases, there is a corresponding decrease in methylation at this CpG site, implying a potential influence of PBB on the methylation but again the slope value of -0.043 is very small and might suggest that the changes in methylation due to PBB exposure are not significant. It's important to take into account that while these patterns of methylation change in response to PBB exposure are noted, the actual biological impact of such small changes remains unclear.

24.1.3 cg04158069

The CpG site cg04158069 shows an elastic net weight of -6.590, indicating an inverse relationship between PBB exposure and methylation at this site. As PBB exposure increases, a decrease in the methylation level at this site is expected.

24.1.4 cg18108008

The CpG site cg18108008 has an elastic net weight of 3.836. This positive weight implies a direct relationship with PBB exposure. That is, an increase in PBB exposure is predicted to result in an increase in the methylation level at this CpG site. This relationship can be valuable for the prediction of PBB exposure levels based on observed methylation data at this particular site.

24.2 Smoking habits

For smoking I found to cg05575921 and cg21566642 to be the two most significant CpGs this is backed up the article Tobacco Smoking Leads to Extensive Genome-Wide Changes in DNA Methylation [38, 88]. This article includes both these CpG sites and found a similar relationship where smokers where found to have a significant lower amount of methylation at given sites. This is further enforced in another paper by Epigenetic Signatures of Cigarette Smoking [38].

24.2.1 cg05575921

The CpG site cg05575921 exhibits notable characteristics in relation to smoking. Its R-squared value of 0.33 indicates that 33% of the variance in methylation beta values can be accounted for by the model, suggesting a moderate association between the methylation status of this particular CpG site and smoking behavior. The slope of -0.15 implies an inverse relationship; as smoking exposure increases, methylation at this site tends to decrease. Further supporting this association is the Elastic Net (EN) weight of -6.67. This negative value denotes a strong relationship with smoking according to the Elastic Net regression

model, with more negative values indicating stronger associations. Together, these findings suggest that the methylation status of cg05575921 could serve as a meaningful biological indicator of smoking exposure.

This CpG has been used by others to predict lung cancer risk [60]. This is very intruiging as lung cancer is generally considered to be the biggest risk of cigarette smoking.

24.2.2 cg21566642

The CpG site cg21566642 also displays significant associations with smoking. An R-squared value of 0.31 suggests that 31% of the variance in methylation beta values can be explained by the model, indicating a moderately strong relationship between the methylation status at this CpG site and smoking. The slope of -0.11 further supports this association, revealing an inverse relationship between smoking exposure and methylation at this site. Although the decrease in methylation is less pronounced compared to cg05575921, it remains significant. The Elastic Net (EN) weight of -3.93, despite being less negative than that of cg05575921, still underscores a considerable association with smoking as per the elastic net regression model. Thus, the methylation status of cg21566642 potentially serve as a marker for smoking exposure, albeit potentially less sensitive than cg05575921.

Interestingly one publication [12] found that cg21566642 was also related to coffee consumption, it was in fact the only CpG that met their criteria for genome wide significance of P=3.7×10^{-10}.However the article did consider that most smokers also drink coffee and when adjusted for smoking their statistical significance was reduced to P=3.7×5.4^{-4}.

24.3 Benzene exposure

For smoking I found to cg07156839 and cg20139683 to be the two most significant CpGs. I could not however find any articles that supported this find. One interesting thing is that even though benzene is considered to be one of the harmful chemicals in cigarette smoking the articles [38, 88] I found that detailed CpGs related to smoking. I did however find that findings that reported that cg07156839 were related to male infertility [74], however this study was done on methylation in sperm cells. This is mostly interesting as benzene exposure has been show to have an impact on male fertility [61].

24.3.1 cg07156839

For the CpG site cg07156839, the R-squared value of 0.995 indicates a very strong association between this site's methylation status and benzene exposure, accounting for 99.5% of the variance in methylation beta values. The slope of -0.569 further corroborates this relationship, illustrating that an increase in

benzene exposure corresponds to a decrease in the methylation beta value at this site. This negative slope suggests a direct inverse relationship between benzene exposure and methylation at cg07156839. In the context of the elastic net regression model, the EN weight of -7.822 for this site underscores a strong association with benzene exposure. The more negative this value, the stronger the association, making cg07156839 a notable indicator in the context of benzene exposure.

24.3.2 cg20139683

The CpG site cg20139683 exhibits an extremely strong association with benzene exposure, as indicated by an R-squared value of 0.997, which means 99.7% of the variance in methylation beta values can be accounted for by the model. The slope of -0.679 illustrates a strong inverse relationship; as benzene exposure increases, methylation at this site significantly decreases. This decrease in methylation is more pronounced compared to cg07156839, another CpG site. The Elastic Net (EN) weight of -3.709, while less negative than that of cg07156839, still highlights a significant association with benzene exposure according to the elastic net regression model. Taken together, these results suggest that methylation status of cg20139683 could serve as a highly sensitive biological indicator of benzene exposure.

The lack of articles outlining external effects on these two CpG combined with my findings that methylation in these sites are not only strongly related to benzene exposure but that the change in methylation is significant indicates that these sites could be specific to benzene exposure. This specificity could potentially serve as a useful biomarker for benzene exposure in future research and environmental monitoring efforts.

25 Data availability issues and quality of data

One of the biggest problems I encountered during the thesis was lack of available data sets. For PBB I had a single set with raw EPIC data which is What I wanted but It would have been great to have one or two more sets that measured PBB to make sure that the effects observed on the sets were caused by PBB and not some other common variable in the cohort.

For smoking I only had 450k data available which means that there were roughly 400 000 CpGs that could potentially have a relationship with smoking that I am unable to test. Additionally three of the smoking related sets GSE50660, GSE54690 and GSE106648 did not include RAW idat files only a csv file containing the calculated beta values. This posed a significant constraint as I could not perform a comprehensive analysis on the raw data, limiting my ability to detect more subtle relationships or patterns. The absence of RAW idat files meant that I couldn't reprocess the data or perform any quality control checks,

thus leaving potential issues like batch effects, poor probe performance, or outlier detection unaddressed. These conditions could potentially skew the results and lead to inaccurate conclusions.

My benzene set like the smoking set was in 450k format and post processed beta values. And also like for PBB I could only find one set that contained metadata on benzene exposure. A unique issue with this set was that it only contained twelve subjects which means that this set was unfit for creating predictive models. Additionally this small sample size makes it challenging to draw definitive conclusions and significantly reduces the statistical power of the study. The inherent variability among the subjects, along with the potential for outliers, might disproportionately influence the results. Furthermore, this limited number of subjects might not accurately represent the larger population, thus reducing the external validity of my findings related to benzene.

I also had an issue were many of my sets namely GSE147430, GSE85210 and GSE50967 did not contain any data on covariates aside from the main stressor in the study. This meant that I could not look at the age acceleration of subjects in these sets which limited my observations significantly. Additionally this also meant that I was unable to verify whether there were any other factors that impacted my findings on specific CpGs related to stressors.

26 Tissue type

The principle that methylation patterns are cell-specific is a fundamental tenet of epigenetics [29, 91]. This is partly attributable to the necessity of cell differentiation, which requires distinct active genes for various tissue types. Additionally, the impact of external factors will differ among tissues [11]. For instance, smoking has been shown to disproportionately alter methylation patterns in lung tissue compared to blood [28, 77, 86].

This means that an examination of methylation changes without taking into account the tissue type could lead to a partial understanding of the situation. In this project, blood was the sole tissue type used for methylation analysis. Therefore, it's important to recognize that the results predominantly reflect changes in blood cells and may not fully capture the impacts in other tissues. This does not mean that there is guaranteed something to find by looking at methylation in other types of tissue but it is definitely worth considering, especially tissue types that are known to be affected by the stressors such as lung tissue for smoking [78] and sperm cells for benzene [61].

27 Main discoveries and future work

All of my stressors indicated some level of effect on blood cell composition, namely the levels of CD4T and CD8T. This suggests a significant immune response to these stressors, potentially altering the cellular landscape of the blood. Notably, these shifts in CD4T and CD8T levels might influence the overall immune function and health status of an individual.The main takeaway from these observations is that it supports previous articles on the health impacts of these stressors [30, 38, 53, 80].

My prediction model results were not as good as I had hoped when setting out on this project. For PBB my best model was MEN100 which was an elastic net model using my top 100 CpGs and for smoking it was the MEN 100 model which also was an elastic net model using my top 100 CpGs for smoking. While these models both show that the pursuit of creating predictive models for these exposures is not hopeless, I was hoping for results as good as those found when trying to predict epigenetic ageing. However I do believe that improvements can be made, particularly by refining the model features and perhaps incorporating more advanced machine learning techniques. Additionally, the data itself was a limiting factor. The data available for smoking and PBB is not as readily available as methylation data with peoples age. It would also be incredibly interesting If it would be possible to analyse methylation in the same subject before and after exposure just as you have been able to track someones methylation as they age and observe the differences.

For PBB and smoking the CpGs I idetified as relevant were also the ones found in similar prior research [14, 38, 88]. This provides further evidence that these CpG sites are reliably associated with these exposures, reinforcing their potential value in related health and environmental studies.

The discovery that I am most excited for is the potential of cg07156839 and cg2013968 as specific biomarkers for detecting benzene exposure. What I am hoping for here is that these CpG sites could serve as reliable indicators for benzene exposure in future studies. If these associations are proven consistent across larger and more diverse population samples, it could significantly enhance our ability to quickly and accurately assess benzene exposure levels in individuals. This could have significant implications for environmental health research and public health interventions, potentially enabling targeted preventative measures for those at high risk of exposure. Furthermore, understanding the specific methylation patterns associated with benzene exposure could provide valuable insights into the underlying biological mechanisms of benzene-related health effects. It's an exciting prospect to consider, and I look forward to seeing how this line of research progresses.

A GitHub

A GitHub has been created that contains all the code I used during this project. It does not include my result data or the raw The repository is private. GitHub accounts with access to the repository has therefore been created. Credentials are provided in table 8, and a link to the repository is also below:

https://github.com/HugoNorholm/Master

Username	Password
HugoMasterObserver1	HugoMaster1
HugoMasterObserver2	HugoMaster2

Table 8: GitHub account table

References

[1] ai gao ai. *GSE50967*. 2016. URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE50967.

[2] Plat AI. *Confusion Matrix in Machine Learning*. 2021. URL: https://plat.ai/blog/confusion-matrix-in-machine-learning/ (visited on 05/01/2023).

[3] Hussain Alsaleh and Penelope R. Haddrill. "Identifying blood-specific age-related DNA methylation markers on the Illumina MethylationEPIC® BeadChip." In: *Forensic Science International* 303 (Sept. 2019), p. 109944. DOI: 10.1016/j.forsciint.2019.109944.

[4] Martin J. Aryee et al. "Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA Methylation microarrays." In: *Bioinformatics* 30.10 (2014), pp. 1363–1369. DOI: 10.1093/bioinformatics/btu049.

[5] Andrew J Bannister and Tony Kouzarides. "Regulation of chromatin by histone modifications." In: *Cell Research* 21.3 (2011), pp. 381–395. DOI: 10.1038/cr.2011.22.

[6] David P Bartel. "MicroRNAs." In: *Cell* 116.2 (Jan. 2004), pp. 281–297. DOI: 10.1016/s0092-8674(04)00045-5.

[7] Christopher G. Bell et al. "DNA methylation aging clocks." In: *Genome Biology* 20.1 (2019). DOI: 10.1186/s13059-019-1824-y.

[8] Adrian Bird. "DNA methylation patterns and epigenetic memory." In: *Genes amp; Development* 16.1 (2002), pp. 6–21. DOI: 10.1101/gad.947102.

[9] Celine Boby. *GSE54690*. 2014. URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE54690.

[10] Sven Bocklandt et al. "Epigenetic predictor of age." In: *PLoS ONE* 6.6 (2011). DOI: 10.1371/journal.pone.0014821.

[11] B. C. Christensen et al. "Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context." In: *PLoS Genet* 5.8 (2009), e1000602. ISSN: 1553-7390. DOI: 10.1371/journal.pgen.1000602.

[12] Yu-Hsuan Chuang et al. "Coffee consumption is associated with DNA methylation levels of human blood." In: *European Journal of Human Genetics* 25.5 (2017), pp. 608–616. DOI: 10.1038/ejhg.2016.175.

[13] Sarah Curtis. *GSE116339*. 2018. URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE116339.

[14] Sarah W. Curtis et al. "Exposure to polybrominated biphenyl (PBB) associates with genome-wide DNA methylation differences in peripheral blood." In: *Epigenetics* 14.1 (2019), pp. 52–66. DOI: 10.1080/15592294.2019.1565590.

[15] Pelegri-Siso D et al. "methylclock: a Bioconductor package to estimate DNA methylation age." In: *Bioinformatics* 37.12 (Sept. 2020), pp. 1759–1760. ISSN: 1367-4803. URL: https://doi.org/10.1093/bioinformatics/btaa825.

[16] Daniel Moura. *R vs. Python vs. Julia*. [Online; accessed May 31, 2022]. 2021. URL: https://miro.medium.com/max/1400/1*OH_n58xfBC7HSP2U8ZC1GQ.png.

[17] PO Darnerud. "Toxic effects of brominated flame retardants in man and in wildlife." In: *Environment international* 29.6 (2003), pp. 841–853.

[18] Databricks. *Deep Learning*. [Online; accessed May 31, 2022]. 2022. URL: https://databricks.com/wp-content/uploads/2019/01/deep-learning.jpg.

[19] Kathryn Demanelis et al. "Association of arsenic exposure with whole blood DNA methylation: An epigenome-wide study of Bangladeshi adults." In: *Environmental Health Perspectives* 127.5 (May 2019), p. 057011. DOI: 10.1289/ehp3849.

[20] Centers for Disease Control and Prevention. "Harms of Cigarette Smoking and Health Benefits of Quitting." In: *CDC fact sheet* (2021). URL: https://www.cdc.gov/tobacco/data_statistics/fact_sheets/index.htm.

[21] National Institute on Drug Abuse. "How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis for Smoking-Attributable Disease: A Report of the Surgeon General." In: *NIH Publication No. 10-5647* (2010). URL: https://www.ncbi.nlm.nih.gov/books/NBK53017/.

[22] Ana Eulalio, Eric Huntzinger, and Elisa Izaurralde. "Getting to the root of MIRNA-mediated gene silencing." In: *Cell* 132.1 (Jan. 2008), pp. 9–14. DOI: 10.1016/j.cell.2007.12.024.

[23] Nora Fernandez-Jimenez et al. "Comparison of illumina 450k and epic arrays in placental DNA methylation." In: *Epigenetics* 14.12 (2019), pp. 1177–1182. DOI: 10.1080/15592294.2019.1634975.

[24] Maximilian H. Fitz-James and Giacomo Cavalli. "Molecular mechanisms of transgenerational epigenetic inheritance." In: *Nature Reviews Genetics* 23.6 (2022), pp. 325–341. DOI: 10.1038/s41576-021-00438-5.

[25] Jean-Philippe Fortin, Timothy J. Triche, and Kasper D. Hansen. "Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi." In: *Bioinformatics* 33.4 (2017). DOI: 10.1093/bioinformatics/btw691.

[26] C. Franceschi and J. Campisi. "Chronic inflammation (inflammaging) and its potential contribution to age-associated diseases." In: *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 69.Suppl 1 (2014). DOI: 10.1093/gerona/glu057.

[27] Fedor Galkin et al. "DeepMAge: A methylation aging clock developed with deep learning." In: *Aging and disease* 12.5 (2021), p. 1252. DOI: 10.14336/ad.2020.1202.

[28] X. Gao et al. "Relationship of tobacco smoking and smoking-related DNA methylation with epigenetic age acceleration." In: *Oncotarget* 7.30 (2016), pp. 46878–46889. ISSN: 1949-2553. DOI: 10.18632/oncotarget.9795.

[29] E. R. Gibney and C. M. Nolan. "Epigenetics and gene expression." In: *Heredity* 105.1 (2010), pp. 4–13. ISSN: 1365-2540. DOI: 10.1038/hdy.2010.54. URL: https://doi.org/10.1038/hdy.2010.54.

[30] Helen Guo, Stacy Ahn, and Luoping Zhang. "Benzene-associated immunosuppression and chronic inflammation in humans: A systematic review." In: *Occupational and Environmental Medicine* 78.5 (2020), pp. 377–384. DOI: 10.1136/oemed-2020-106517.

[31] Gregory Hannum et al. "Genome-wide methylation profiles reveal quantitative views of human aging rates." In: *Molecular Cell* 49.2 (2013), pp. 359–367. DOI: 10.1016/j.molcel.2012.10.016.

[32] U.S. Department of Health and Human Services. "The Health Consequences of Smoking - 50 Years of Progress: A Report of the Surgeon General." In: *Centers for Disease Control and Prevention (US)* (2014). URL: https://www.ncbi.nlm.nih.gov/books/NBK179276/.

[33] Helixitta. *CpG sequence of one DNA strand versus C-G base pair on complementary strands*. Jan. 2016. URL: https://en.wikipedia.org/wiki/CpG_site.

[34] Steve Horvath. "DNA methylation age of human tissues and cell types." In: *Genome Biology* 14.10 (2013). DOI: 10.1186/gb-2013-14-10-r115.

[35] Steve Horvath and Kenneth Raj. "DNA methylation-based biomarkers and the epigenetic clock theory of ageing." In: *Nature Reviews Genetics* 19.6 (Apr. 2018), pp. 371–384. DOI: 10.1038/s41576-018-0004-3.

[36] Steve Horvath et al. "Epigenetic clock for skin and blood cells applied to Hutchinson Gilford progeria syndrome and ex vivo studies." In: *Aging* 10.7 (2018), pp. 1758–1775. DOI: 10.18632/aging.101508.

[37] National Cancer Institute. "Formaldehyde and Cancer Risk." In: *National Institutes of Health* (2011). URL: https://www.cancer.gov/about-cancer/causes-prevention/risk/substances/formaldehyde/formaldehyde-fact-sheet.

[38] Roby Joehanes et al. "Epigenetic signatures of Cigarette Smoking." In: *Circulation: Cardiovascular Genetics* 9.5 (Sept. 2016), pp. 436–447. DOI: 10.1161/circgenetics.116.001506.

[39] Peter A. Jones and Stephen B. Baylin. "The epigenomics of cancer." In: *Cell* 128.4 (Feb. 2007), pp. 683–692. DOI: 10.1016/j.cell.2007.01.029.

[40] Yoav Keynan et al. "The role of regulatory T cells in chronic and acute viral infections." In: *Clinical Infectious Diseases* 46.7 (2008), pp. 1046–1052. DOI: 10.1086/529379.

[41] Mary-Claire King, Jeffrey H Marks, and Jessica B Mandell. "Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2." In: *Science* 302.5645 (2003), pp. 643–646.

[42] PR Kodavanti. "Neurotoxicity of persistent organic pollutants: possible mode(s) of action and further considerations." In: *Dose-Response* 3.2 (2005), pp. 273–305.

[43] Tony Kouzarides. "Chromatin modifications and their function." In: *Cell* 128.4 (2007), pp. 693–705. DOI: 10.1016/j.cell.2007.02.005.

[44] Yunsung Lee et al. "Blood-based epigenetic estimators of chronological age in human adults using DNA methylation data from the Illumina MethylationEPIC array." In: *BMC Genomics* 21.1 (2020). DOI: 10.1186/s12864-020-07168-8.

[45] Fatjon Leti et al. "Methods for CPG methylation array profiling via bisulfite conversion." In: *Methods in Molecular Biology* (2018), pp. 233–254. DOI: 10.1007/978-1-4939-7471-9_13.

[46] Morgan E. Levine et al. "An epigenetic biomarker of aging for lifespan and healthspan." In: *Aging* 10.4 (2018), pp. 573–591. DOI: 10.18632/aging.101414.

[47] Yanping Li et al. "Healthy lifestyle and life expectancy free of cancer, cardiovascular disease, and type 2 diabetes: Prospective cohort study." In: *BMJ* (2020), p. l6669. DOI: 10.1136/bmj.l6669.

[48] Lucas Paulo de Lima Camillo, Louis R Lapierre, and Ritambhara Singh. "A pan-tissue DNA-methylation epigenetic clock based on deep learning." In: *npj Aging* (2022). DOI: 10.1038/s41514-022-00085-y. eprint: https://www.nature.com/articles/s41514-022-00085-y.pdf. URL: https://doi.org/10.1038/s41514-022-00085-y.

[49] Yun Liu. *GSE106648*. 2018. URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE54690.

[50] Ake T. Lu et al. "DNA methylation GrimAge strongly predicts lifespan and healthspan." In: *Aging* 11.2 (2019), pp. 303–327. DOI: 10.18632/aging.101684.

[51] M I Luster, R E Faith, and J A Moore. "Effects of polybrominated biphenyls (PBB) on immune response in rodents." In: *Environmental Health Perspectives* 23 (1978), pp. 227–232. DOI: 10.1289/ehp.7823227.

[52] Jovana Maksimovic, Lavinia Gordon, and Alicia Oshlack. "SWAN: Subset quantile Within-Array Normalization for Illumina Infinium HumanMethylation450 BeadChips." In: *Genome Biology* 13.6 (2012), R44. DOI: 10.1186/gb-2012-13-6-r44.

[53] Espen Mariussen. "Polybrominated diphenyl ethers (PBDEs) and polybrominated biphenyls (PBBs) in human blood samples from Norway: a study on levels and temporal trends." In: *Environment international* 39 (2012), pp. 210–219.

[54] Ruslan Medzhitov. "Origin and physiological roles of inflammation." In: *Nature* 454.7203 (2008), pp. 428–435. DOI: 10.1038/nature07201.

[55] Richard R Meehan et al. "DNA methylation as a genomic marker of exposure to chemical and environmental agents." In: *Current Opinion in Chemical Biology* 45 (Mar. 2018), pp. 48–56. DOI: 10.1016/j.cbpa.2018.02.006.

[56] Lisa D Moore, Thuc Le, and Guoping Fan. "DNA methylation and its basic function." In: *Neuropsychopharmacology* 38.1 (2012), pp. 23–38. DOI: 10.1038/npp.2012.112.

[57] NTP. *Toxicology and carcinogenesis studies of polybrominated biphenyls (CAS no. 59536-65-1) in F344/N rats and B6C3F1 mice (feed studies).* Tech. rep. Technical Report Series, No. 298, 1986.

[58] World Health Organization. "Tobacco." In: *WHO fact sheet* (2021). URL: https://www.who.int/news-room/fact-sheets/detail/tobacco.

[59] Paul Peixoto et al. "From 1957 to nowadays: A brief history of epigenetics." In: *International Journal of Molecular Sciences* 21.20 (2020), p. 7571. DOI: 10.3390/ijms21207571.

[60] Rob Philibert et al. "Using CG05575921 methylation to predict lung cancer risk: A potentially bias-free precision epigenetics approach." In: *Epigenetics* 17.13 (2022), pp. 2096–2108. DOI: 10.1080/15592294.2022.2108082.

[61] Diana Poli et al. "Sex difference and benzene exposure: Does it matter?" In: *International Journal of Environmental Research and Public Health* 19.4 (2022), p. 2339. DOI: 10.3390/ijerph19042339.

[62] Vardhman K. Rakyan et al. "Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains." In: *Genome Research* 20.4 (2010), pp. 434–439. DOI: 10.1101/gr.103101.109.

[63] International Agency for Research on Cancer. "IARC Monographs on the Evaluation of Carcinogenic Risks to Humans. Volume 83. Tobacco Smoke and Involuntary Smoking." In: *World Health Organization, International Agency for Research on Cancer* (2004). URL: https://www.ncbi.nlm.nih.gov/books/NBK53017/.

[64] Christopher F. Rider and Chris Carlsten. "Air pollution and DNA methylation: Effects of exposure in humans." In: *Clinical Epigenetics* 11.1 (2019). DOI: 10.1186/s13148-019-0713-2.

[65] John R Riordan et al. "Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA." In: *Science* 245.4922 (1989), pp. 1066–1073.

[66] SL Schantz, JJ Widholm, and DC Rice. "Effects of PCB exposure on neuropsychological function in children." In: *Environmental health perspectives* 111.3 (2003), pp. 357–376.

[67] Marco Schmidt et al. "Deconvolution of cellular subsets in human tissue based on targeted DNA methylation analysis at individual CPG sites." In: *BMC Biology* 18.1 (2020). DOI: 10.1186/s12915-020-00910-4.

[68] Megan E. Schmidt and Steven M. Varga. "The CD8 T cell response to respiratory virus infections." In: *Frontiers in Immunology* 9 (2018). DOI: 10.3389/fimmu.2018.00678.

[69] Mona D. Shahbazian and Michael Grunstein. "Functions of site-specific histone acetylation and deacetylation." In: *Annual Review of Biochemistry* 76.1 (2007), pp. 75–100. DOI: 10.1146/annurev.biochem.76.052705.162114.

[70] Mike L Smith et al. "Illuminaio: An open source IDAT parsing tool for Illumina microarrays." In: *F1000Research* 2 (2013), p. 264. DOI: 10.12688/f1000research.2-264.v1.

[71] Steven Smith et al. *Methylprep: A Python package for DNA methylation array preprocessing.* https://github.com/AlexsLemonade/Methylprep. Version 1.6.0. 2021.

[72] Matthew Suderman, Gibran Hemani, and Josine Min. *meffil: Efficient algorithms for DNA methylation.* R package version 1.3.5. 2023. URL: https://github.com/perishky/meffil.

[73] Mònica Suelves et al. "DNA methylation dynamics in cellular commitment and differentiation." In: *Briefings in Functional Genomics* (2016). DOI: 10.1093/bfgp/elw017.

[74] Kumar Mohanty Sujit et al. "Genome-wide differential methylation analyses identifies methylation signatures of male infertility." In: *Human Reproduction* 33.12 (2018), pp. 2256–2267. DOI: 10.1093/humrep/dey319.

[75] Kristin Aurora Sydhagen. "Building Epigenetic Clocks for Estimating Ageing in Life After Cancer." In: (2022).

[76] J. Kenneth Tay, Balasubramanian Narasimhan, and Trevor Hastie. "Elastic Net Regularization Paths for All Generalized Linear Models." In: *Journal of Statistical Software* 106.1 (2023), pp. 1–31. DOI: 10.18637/jss.v106.i01.

[77] Andrew E. Teschendorff et al. "Correlation of Smoking-Associated DNA Methylation Changes in Buccal Cells With DNA Methylation Changes in Epithelial Cancer." In: *JAMA Oncology* 1.4 (July 2015), pp. 476–485. ISSN: 2374-2437. DOI: 10.1001/jamaoncol.2015.1053. eprint: https://jamanetwork.com/journals/jamaoncology/articlepdf/2293216/coi150026.pdf. URL: https://doi.org/10.1001/jamaoncol.2015.1053.

[78] Michael J. Thun et al. "50-year trends in smoking-related mortality in the United States." In: *The New England Journal of Medicine* 368.4 (2013), pp. 351–364. DOI: 10.1056/NEJMsa1211127.

[79] Agency for Toxic Substances and Disease Registry. "Toxicological Profile for Acetone." In: *U.S. Department of Health and Human Services, Public Health Service* (1994). URL: https://www.atsdr.cdc.gov/toxprofiles/tp21.pdf.

[80] Agency for Toxic Substances and Disease Registry. "Toxicological Profile for Benzene." In: *U.S. Department of Health and Human Services, Public Health Service* (2007). URL: https://www.atsdr.cdc.gov/toxprofiles/tp3.pdf.

[81] Hoa Thi Tran et al. "A benchmark of batch-effect correction methods for single-cell RNA sequencing data." In: *Genome Biology* 21.1 (2020). DOI: 10.1186/s13059-019-1850-9.

[82] Timothy J. Triche et al. "Low-level processing of Illumina Infinium DNA Methylation BeadArrays." In: *Nucleic Acids Research* 41.7 (2013), e90. DOI: 10.1093/nar/gkt090.

[83] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer, 2002. URL: https://www.stats.ox.ac.uk/pub/MASS4/.

[84] Xuting Wang. *GSE147430*. 2021. URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE147430.

[85] Xuting Wang. *GSE85210*. 2017. URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE85210.

[86] Xiaohui Wu et al. "Effect of tobacco smoking on the epigenetic age of human respiratory organs." In: *Clinical Epigenetics* 11.1 (2019), p. 183. ISSN: 1868-7083. DOI: 10.1186/s13148-019-0777-z. URL: https://doi.org/10.1186/s13148-019-0777-z.

[87] Tsun-Po Yang. *GSE50660*. 2014. URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE50660.

[88] Sonja Zeilinger et al. "Tobacco smoking leads to extensive genome-wide changes in DNA methylation." In: *PLoS ONE* 8.5 (2013). DOI: 10.1371/journal.pone.0063812.

[89] Qian Zhang et al. "Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing." In: *Genome Medicine* 11.1 (2019). DOI: 10.1186/s13073-019-0667-1.

[90] Jinfang Zhu, Hidehiro Yamane, and William E. Paul. "Differentiation of effector CD4 T cell populations." In: *Annual Review of Immunology* 28.1 (2010), pp. 445–489. DOI: 10.1146/annurev-immunol-030409-101212.

[91] D. Zilberman and S. Henikoff. "Genome-wide analysis of DNA methylation patterns." In: *Development* 134.22 (2007), pp. 3959–3965. ISSN: 0950-1991. DOI: 10.1242/dev.001131.