Can We Use Physical Characteristics of Genes to Predict Age-Related Changes in Expression?

A Classifier-Based Exploration of Predictive Gene Properties

Lovro Mlikotić [1]
Supervisors: Marcel Reinders, Inez den Hond, Bram Pronk, Gerard Bouland
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Lovro Mlikotić
Final project course: CSE3000 Research Project
Thesis committee: Kaitai Liang, Marcel Reinders, Inez den Hond, Bram Pronk, Gerard Bouland

Abstract

The aim of this research is to investigate whether physical gene characteristics can predict age-related changes in gene expression. Specifically, we analyze gene length, GC content, distance to the ends of the chromosome, and similar features to determine their connection with differential expression between young and old individuals. Among these features, gene length consistently shows a strong correlation with age-related expression patterns. However, when combined, the selected features do not provide sufficient predictive power to train a classifier capable of exceeding a modest 66% accuracy. These findings highlight the limitations of the current feature set and point toward the need for more complex feature preprocessing steps or biologically relevant features in future predictive models.

1 Introduction

No one escapes aging - not mice, not men. Over time, bones weaken, synapses falter, and cells gradually lose their ability to function, until eventually, the body as a whole begins to shut down. Aging is an inevitable but deeply complex biological process, propelled at the molecular level by the gradual accumulation of mutations and modifications in DNA and RNA. Understanding the mechanisms behind aging is not just a scientific pursuit: it holds real potential to improve quality of life. By gaining clearer picture into how and why gene expression changes with age, we can inform medical strategies aimed at mitigating the effects of aging, support the development of new treatments, and even explore the possibility of slowing or reversing aspects of age-related decline.

To understand this paper, it is imperative to understand that DNA is a linear sequence of four nucleotide bases, adenine (A), thymine (T), guanine (G), and cytosine (C), which encodes the genetic instructions for building proteins. Genes are specific regions of this sequence, and the full set of genes makes up the genome. Gene expression involves transcription of DNA into messenger RNA (mRNA), followed by translation of mRNA into proteins. Through mechanisms like alternative splicing, a single gene can give rise to multiple transcripts, and therefore, multiple protein variants. This entire process constitutes gene expression, and if a fault occurs anywhere in the process, the resulting gene will not be transcribed and translated, and therefore, not expressed.

This paper investigates the relationship between gene expression and aging, with a focus on identifying physical gene-level characteristics that may influence whether a gene is differentially expressed between young and old individuals. Specifically, the goal is to evaluate the predictive power of such features in determining age-related expression changes. If successful, this approach might help identify early signs of age-related decline and be of service to future medical and scientific efforts.

There was prior work conducted into the role of gene characteristics in aging, namely the findings from Ibañez-Solé et al. [1] and Soheili-Nezhad et al. [2], which reveal a consistent occurrence in aging biology: the preferential downregulation of long genes as individuals age. These studies together introduce and support the concept of gene length-dependent transcriptional decline (GLTD), i.e. the fact that the longer a gene is, the more susceptible it is to age-associated underexpression.

Ibañez-Solé et al. conducted a comprehensive analysis using single-cell RNA sequencing data across multiple tissues and species (mouse and human) and shown that aging is associated with a decrease in the expression of longer genes [1]. This transcriptional decline was observed across both sexes and all tissues. Additionally, they showed that environ-

mental genotoxic stressors, such as UV radiation and tobacco smoke, can induce a similar gene-length-dependent suppression in young individuals, suggesting that GLTD can be accelerated by external damage.

Similarly, Soheili-Nezhad et al. puts GLTD within the broader context of aging biology. They proposed that GLTD is not just correlated with aging, but is potentially causing it [2]. In their paper, it is suggested that GLTD happens because damage to DNA causes the cellular machinery responsible for reading genes to stall, which is an issue that affects longer genes more severely, since they take more time to be read. The authors further identified GLTD in various model organisms (mice and humans, among others) and aging-related diseases, including Alzheimer's. They highlighted how genes involved in brain cell communication found at chromosomal fragile sites are particularly vulnerable, linking GLTD to the failure of neurological systems in aging. Interestingly, interventions known to slow aging, such as caloric restriction and vitamin D supplementation, were shown to partially mitigate GLTD, indicating its modifiability.

Together, these studies describe a previously underappreciated aspect of the aging genome: that gene length significantly shapes patterns of gene expression decline. This insight emphasizes the importance of considering physical gene characteristics, not just their biological functions, in age-related gene expression research.

The observation that gene length influences expression patterns with age raises the possibility that other intrinsic gene characteristics may also contribute to age-related transcriptional changes. The aim of this paper is to identify a set of such features and assess whether they collectively have sufficient predictive power to determine if a gene is more likely to be overexpressed in young or old individuals.

The major opportunity here, which was made available by recent technological advancements, lies in the public availability of large-scale biological measurement datasets, which can now be analyzed using machine learning and statistical models.

In this study, we investigate whether physical gene-level characteristics extrapolated from such datasets can predict age-related expression changes. We begin by confirming previous findings linking gene length to differential expression with age, and extend the analysis to GC content and distance from the end of the chromosome. These features were chosen for their biological relevance and because they represent distinct properties, which are hard for a classifier to infer from one another. After evaluating their individual associations with expression patterns through statistical analysis, we combined these features (along with a few additional, biologically related ones added to increase feature diversity and model complexity) to train classifiers aimed at predicting whether a gene is more likely to be overexpressed in young or old individuals. This approach helps reveal how multiple gene features together may contribute to age-related transcriptional changes.

The main finding of this paper is that the feature set used to train and test the classifier, outlined in detail later in the paper, is not informative enough to support high-accuracy prediction of age-related differential gene expression. However, through this process, we found that gene length consistently emerged as a reliable predictor across both statistical analyses and classifier-based approaches. While it is not sufficient on its own to accurately classify gene expression changes with age, it appears to capture one of the most meaningful signals in the current feature set and is likely to remain an important component in any future, more effective models.

2 Experimental Setup

This study is conducted by applying statistical and machine learning methods to the *Tabula Muris Senis* (TMS) dataset, which is a large-scale single-cell transcriptomic atlas of aging tissues in *Mus musculus* (house mouse). Specifically, the "*All - A single-cell transcriptomic atlas characterizes ageing tissues in the mouse - 10x*" data object, which was obtained from the CZ CELLxGENE portal [3], was utilized. The dataset comprises gene expression profiles from 245,389 individual cells sampled from 23 distinct tissues, with measurements across various age groups and both sexes. The resulting data matrix contains 17,984 genes, with over 460 million non-zero entries, stored efficiently as a sparse matrix in Compressed Sparse Row (CSR) format.

The data is stored in an `AnnData` object, where the expression matrix is organized such that rows correspond to individual cells, and columns correspond to genes. Each entry in this matrix contains the raw UMI (Unique Molecular Identifier) count for a particular gene in a given cell. These counts serve as proxies for gene expression levels, and each UMI count reflects the number of observed transcriptions of a gene in a given cell, i.e. the number of observed mRNA molecules transcribed from a single gene in that cell.

A UMI value of 0 indicates that no transcripts for that gene were detected in that cell, suggesting no detectable expression. While UMI counts are inherently integers, some values may appear as floats; but these float values should still be interpreted approximately (for example, 2.5105 roughly means 2 to 3 transcripts). All things considered, UMI counts in the `AnnData` object provide a reliable, per-cell measure of gene expression.

2.0.1 Dataset Preprocessing

Before the TMS could be used for analysis, it was preprocessed. Redundant or non-informative fields from the cell metadata (such as ontology term IDs and various free-text annotations) were removed. The remaining essential metadata fields were: cell identifiers, donor identifiers, number of genes detected per cell, total UMI counts, age, sex, cell type, and tissue of origin.

Age annotations, originally present as discrete values (in months), were divided into two categories: young and old. These were cast as categorical variables with an explicit ordering (young $<$ old) to preserve their chronological ordering in later analysis tasks. Each possible age value was categorized as follows:

- 1m, 3m $\rightarrow$ young
- 18m, 21m, 24m, 30m $\rightarrow$ old

Subsequently, the gene expression data was normalized and transformed. Total-count normalization was performed, scaling each cell's total UMI count to 10,000, in order to correct for varying sequencing depths. The data was then log-transformed using the `log1p` function to stabilize variance and compress large values.

2.1 Gene Annotations

The TMS dataset alone did not contain enough data to perform the intended analysis. Therefore, two additional datasets were utilized in order to add on the absent gene annotations.

2.1.1 Gene Feature Dataset

Gene-level characteristics were obtained through processing of transcript-level data, which was procured via Ensembl BioMart [4] using the Ensembl Genes database (release 114, GRCm39) for *Mus musculus*.

To derive gene-level features from transcript-level annotations, transcript-level information was grouped by gene. For attributes that do not vary across transcripts (such as gene name, gene type, chromosome, strand, GC content, and genomic start/end positions) the shared value was simply retained for each gene. Transcript-level metrics, including transcript length, start, and end positions, were summarized per gene using statistical aggregations: mean, median, minimum, and maximum values.

In order to obtain additional relevant features, the distance (in base-pairs) of each gene from the start and end of its chromosome was calculated. Using known chromosome lengths for Mus musculus from the NCBI Genome Reference Consortium [5], the distance from the gene start to the beginning of the chromosome and from the gene end to the end of the chromosome were calculated.

2.1.2 Gene Sequence Dataset

To analyze the gene-level nucleotide sequence, normalized k-mer frequencies (where k = 3) were calculated for each gene in the mouse genome. By definition, k-mer sequences are short, fixed-length substrings of DNA of length k, commonly used to quantify local sequence patterns within a genome. For example, 3-mers such as "AAA" or "ATG" represent all possible combinations of three consecutive nucleotides.

This gene sene sequence dataset was obtained from the GRCm39 cDNA FASTA file for *Mus musculus*, downloaded from Ensembl [6].

The following process was applied to each gene's nucleotide sequence. Using a sliding window approach, all present 3-mers were extracted, and their occurrences were counted. Relative frequencies were calculated by dividing each 3-mer count by the total number of 3-mers in the sequence (L - k + 1, where L is the sequence length). This normalization accounts for gene length variability, and ensures a fair comparison across different genes. The final output of the process is a tabular dataset where each row represents a gene and each column corresponds to the normalized frequency of a possible 3-mer (e.g., "AAA", "AAC", ..., "TTT"), resulting in 64 features per gene. These sequence-derived features were subsequently used as input for statistical and machine learning analyses.

2.1.3 Gene Annotations Final Form

After both were processed, the engineered gene feature dataset and the 3-mer frequency dataset were merged using the gene ID. This combination resulted in a single gene-level dataset (containing information about 78,258 genes) that combined structural, positional, and sequence-based information. The full list of gene annotations in this final dataset can be found in Figure 5, in the Appendix.

2.2 DEG Labeling

The genes were labeled (so a classifier can be trained on them) in the following way. First, the dataset was filtered to include only genes that showed a statistically significant change in expression between young and old mice, based on a p-value threshold of 0.05. Genes

not meeting this criteria were pruned to ensure that the following steps focused only on biologically meaningful differences.

From the remaining DEGs, the top 20% most upregulated in young mice were selected and labeled as class 0, and the top 20% most upregulated in old mice as class 1. Genes with intermediate log fold changes were excluded to create a clearer separation between the two classes.

The 20% threshold was chosen after testing several options (10%, 20%, and 30%) and evaluating the resulting classifier performance. Although the 10% selection yielded similar accuracy, the 20% threshold was chosen in the end as it provided a better balance between maximizing the number of data points and maintaining model performance. In contrast, increasing the further threshold beyond 20% led to a decline in classification accuracy, most likely due to the inclusion of less strongly differentially expressed genes.

3 Responsible Research

3.1 Reflection on Replicability and Reproducibility

The inclusion of a dedicated section on replicability and reproducibility aligns with TU Delft's core values of Integrity, Engagement, and Trust. By ensuring that not only research results, but also the research process is accessible and verifiable, TU Delft actively contributes to raising the standards of academic research and promotes trust and cooperation between researchers. While planning out the research, particular attention was given to ensuring the research is highly replicable and reproducible.

3.1.1 Clear Dataset Referencing

All datasets used throughout this research are mentioned in this paper, and are properly cited. This is in line with recent initiatives from TU Delft, which aim to develop a systematic approach to proper dataset citation and organization. This structured methodology, which advocates for standardized dataset documentation in scientific publications, is consistent with citation standards used for referencing other research papers. Given the current widespread availability of large datasets, and their increased use in present-day research, it is remarkable that standardized referencing has not yet became universal practice. This research paper contributes to setting this new standard.

3.1.2 Open Source Code and Transparent Methodology

All code written for this research will be made accessible once the research is completed. It can be accessed via this GitHub repository: https://github.com/lovromlikotic/gene-expression-aging.

All of the work was conducted in Jupyter Notebook, which, in addition to being excellent interactive computing software, allows the research to be showcased through a linear, logical, and easy-to-follow structure. Although the code changed and evolved significantly throughout the course of the research, it has since been cleaned and organized. The notebooks enhance readability though structure, comprehensive explanations, and extensive code comments. The final notebooks have been refined to a level where someone with a bachelor's degree background in computer science can easily understand the process and successfully reproduce the figures and results presented in the paper.

By adhering to these principles, this work aims for a high standard for reproducibility. It is our hope that this level of transparency and openness becomes the norm across similar research efforts.

4 Results

First, the statistical analysis of the main features with regard to age-related changes in expression is presented, then the results of the classifier. Finally, the feature importance ranking produced as a result of the classifier is analyzed.

4.1 Characteristics

To begin, we examined the set of gene-level features available in our dataset to assess their potential relationship with age-associated changes in gene expression. Our goal was to identify whether individual features showed any correlation with differential expression between young and old mice, prior to using these features as input for classification models. Importantly, we focused on selecting non-overlapping, independent features to ensure that each classifier input contributed distinct information and that no feature could be trivially inferred from another.

Gene length, a well-studied characteristic known to influence age-related expression patterns, has already been discussed in the context of previous studies cited in the Introduction. Nevertheless, we performed a full analysis of gene length within our dataset for completeness. The corresponding visualization and statistical test results are provided in the Appendix.

4.1.1 CG Content

The second gene characteristic taken into account is CG content, which is simply taken to mean the proportion of the cytosine and guanine bases relative to the total gene sequence length.

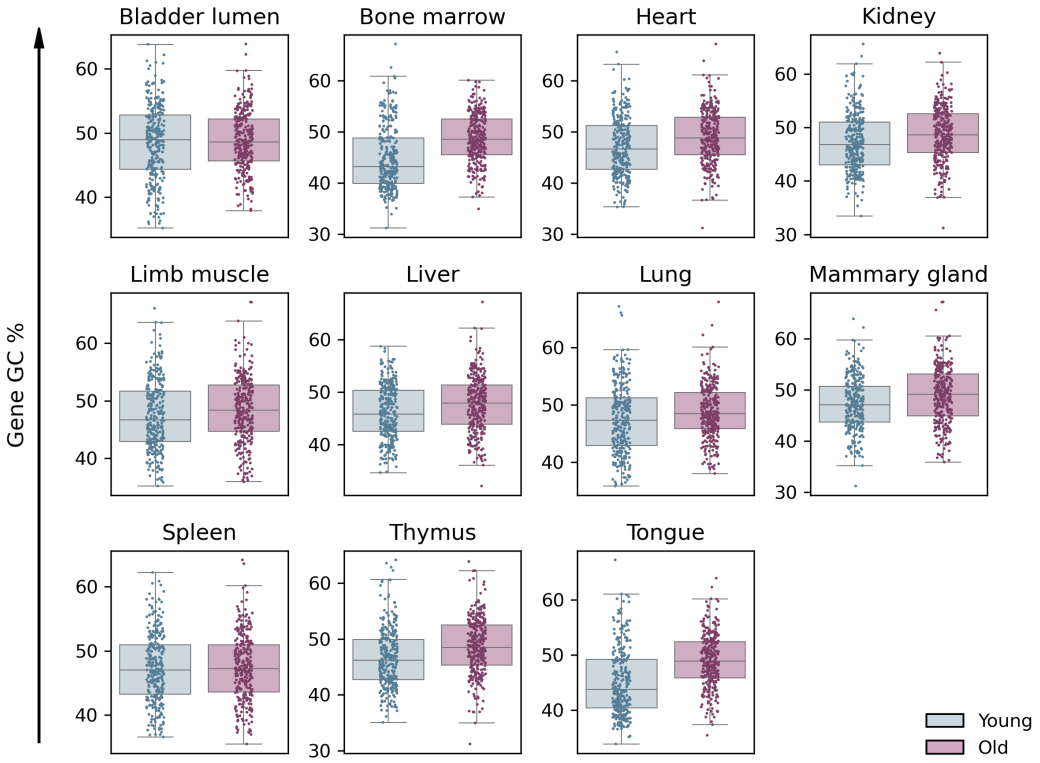


Figure 1: Each panel shows the GC content distribution of the 300 most differentially expressed genes overexpressed in young (blue) and old (purple) samples for a given tissue. Boxplots display summary statistics, while the overlaid dots correspond to individual genes.

As can be inferred from Figure 1, there is a consistent trend across most tissues indicating that genes overexpressed in old mice tend to have higher CG content compared to those overexpressed in young mice.

This observation is supported by testing using the Mann-Whitney U test, which shows statistically significant differences in GC content between the two groups in the majority of tissues. The strongest effects were seen in bone marrow, tongue, and thymus ($p < 10^{-8}$), with genes overexpressed in old showing a higher GC content. Milder but significant differences were also observed in heart, kidney, liver, lung, mammary gland, and limb muscle. Only bladder lumen and spleen did not show statistically significant differences ($p > 0.05$). The full results are provided in Table 3 in the Appendix. Overall, these results point to a broad association between higher GC content and increased gene expression in older tissues.

4.1.2 Chromosome Edge Proximity

The next gene characteristic that is considered is the distance of a gene from the ends of it's chromosome. This distance is defined as the smaller of two values: the distance from the gene's start position to the beginning of the chromosome, and the distance from the gene's end position to the end of the chromosome, measured in base pairs.

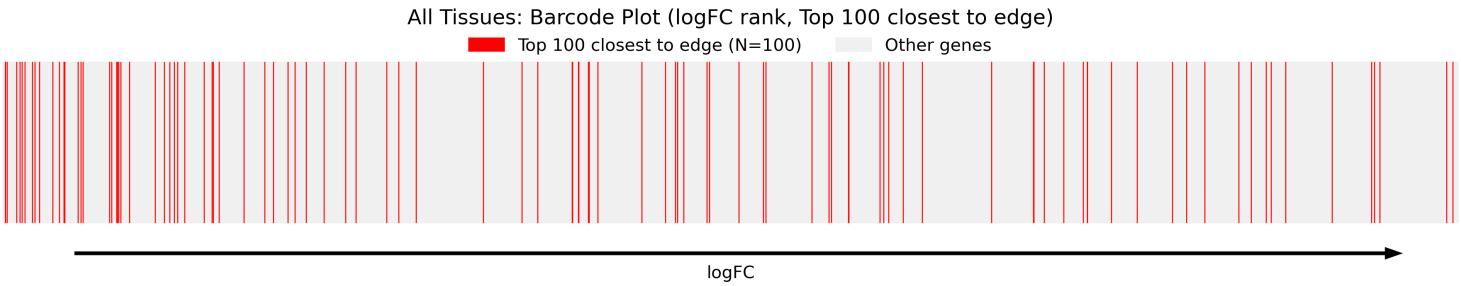


Figure 2: Each vertical line represents a gene. The genes are ranked by their logFC, with the lowest negative logFC being on the left, and the highest positive logFC being on the right. The vertical lines colored in red represent the 100 genes closest to the edge of their chromosome.

Figure 2 shows a noticeable concentration of red lines, which represent genes closest to the ends of chromosomes, towards the left side, where genes with the most negative log fold changes are located. This suggests that many of these edge-adjacent genes are more highly expressed in young mice compared to old mice. When we increase the number of genes colored red beyond the top 100 closest to the chromosome ends, the crowding effect on the left side becomes less apparent, suggesting that the strongest bias is specific to the very closest genes. These observations point to a potential relationship between a gene's chromosomal position and how its expression changes with age.

It is important to note that this visualization includes only genes that are significantly differentially expressed, as all genes with a p-value greater than 0.05 were pruned before the figure was plotted. By focusing solely on genes that pass this significance threshold, the observed patterns, such as the clustering of edge-adjacent genes, can be interpreted with greater confidence, as they are unlikely to be due to random variation.

4.2 Classifier

The preprocessing steps and the labeling strategy for each gene are detailed in the Experimental Setup section, under the DEG Labeling subsection.

To evaluate the data, we selected four classifiers, each representing different learning paradigms and levels of complexity. This allowed us to approach the problem from multiple angles and better assess the predictive power of the features.

The final feature set used as input for the classifiers can be found in Figure 5 in the Appendix. Most features in the dataset can be broadly grouped into three categories: gene/transcript length, distance to chromosome edges, and nucleotide composition (like GC content and 3-mer frequencies). A few additional features (that were readily available from the data) were also included to provide categorical context.

4.2.1 Classifier Performance

Many classifiers with varying levels of complexity were tried out, along with multiple strategies for splitting and labeling the data. Even regression-based approaches were explored but failed to achieve interpretable results. In the interest of space, these approaches are not described here, as none outperformed the two best-performing models presented below. Across

all experiments, classification accuracy peaked at a modest 66%, and even with parameter tuning, the accuracy saw no more than a 1% increase.

Classifier	Class	Precision	Recall	F1-score	Support
XGBoost	0	0.6598	0.6605	0.6601	2044
	1	0.6598	0.6592	0.6595	2042
	Avg	**0.6598**	**0.6598**	**0.6598**	**4086**
Random Forest	0	0.6630	0.6409	0.6517	2044
	1	0.6521	0.6738	0.6628	2042
	Avg	**0.6575**	**0.6574**	**0.6573**	**4086**

Table 1: Performance metrics of the two classifiers. The metrics are shown per binary class and as an average of the two classes.

Considering that the best performing XGBoost and Random Forest classifiers achieved nearly identical accuracy scores (the confusion matrix of Random Forest is shown in Figure 10) and produced extremely similar feature importance rankings and SHAP beeswarm plots, the the following analysis will focus on the results of the Random Forest classifier. For the purposes of completeness, the corresponding plots for the XGBoost classifier are included in the Appendix, beginning with Figure 7.

4.2.2 Inferred Feature Importance

To interpret the classifier outputs, we used SHAP (SHapley Additive exPlanations), a model-agnostic method based on cooperative game theory [7]. SHAP is well-suited for explaining complex models, offering consistent and locally accurate estimates of feature contributions. It provides both global and instance-level insights, allowing us to identify which gene features most strongly influenced the classification results.

As shown in Figure 11 in the Appendix, gene length, and related features such as transcript length, clearly stand out as the most prominent contributor to the classifier's performance. This suggests that gene length is the most influential feature driving the model's predictions. Figure 12, the SHAP beeswarm plot, provides further insight by illustrating how specific feature values influence the classification. Unsurprisingly, lower values of gene length-related features tend to push the prediction toward the overexpressed-in-old class. Another noteworthy observation from the beeswarm plot is that most of the 3-mer features cluster near the center, indicating their relatively inconsistent influence on the model's decisions.

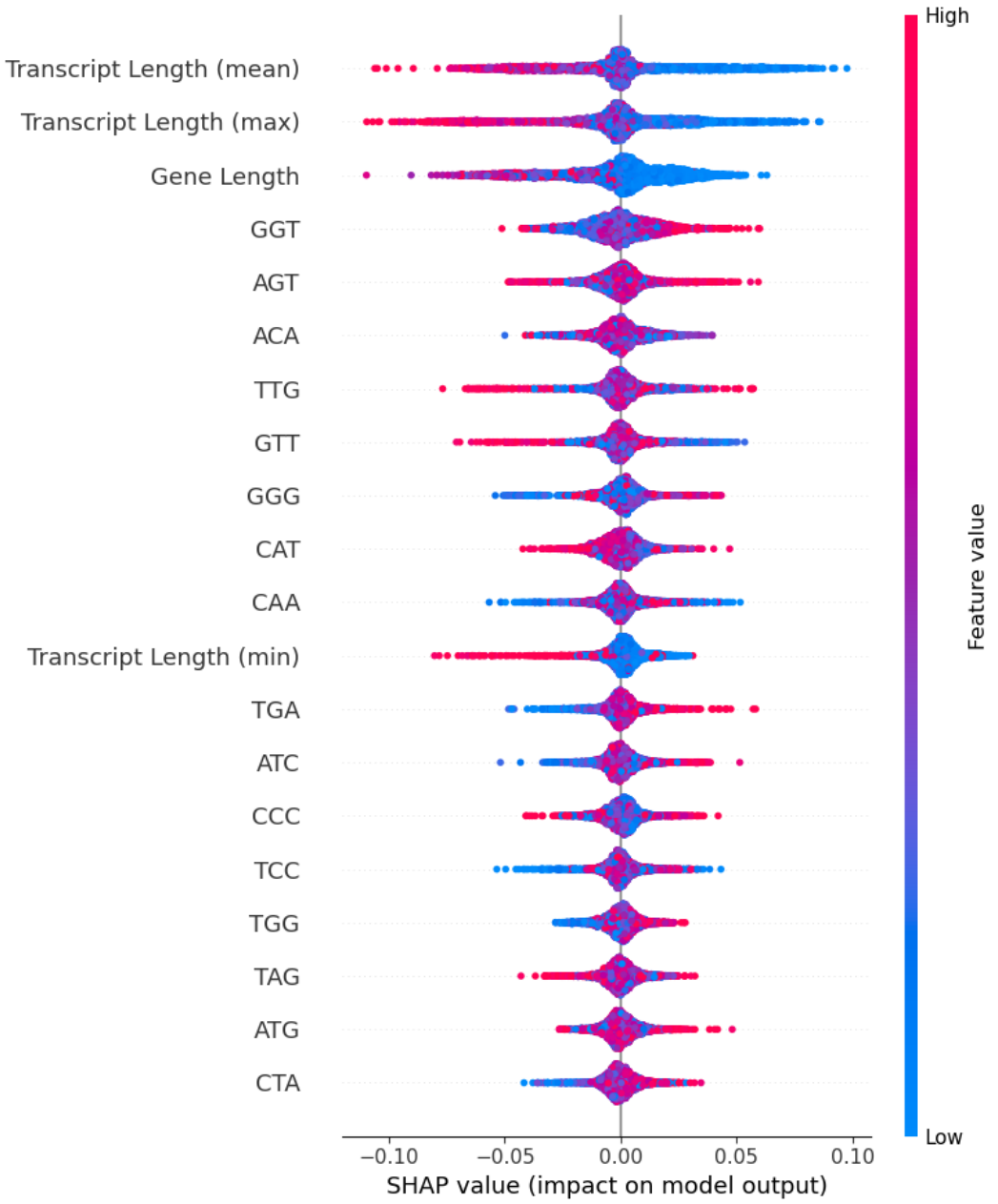


Figure 3: Random Forest. All Tissues. Combined SHAP Beeswarm. Each point represents an observation (gene). The color indicates the magnitude of the feature value for that observation. The horizontal position of the dot indicates the SHAP value. If it is to the right, it contributes toward class 1, if it is to the left, it contributes toward class 0. The magnitude of the horizontal position signifies contribution strength.

4.2.3 Feature Ranking Analysis

Considering the moderate performance of the best-performing classifier that was produced, and in order to better understand and validate the top features identified by the classifier, we needed a point of comparison. Therefore, we performed a point-biserial correlation analysis. The point-biserial correlation is a statistical method used to measure the relationship between a continuous variable and a binary categorical variable. In this case, the various gene features are the continuous variables and the binary DEG label (upregulated-in-young vs. upregulated-in-old) is the categorical variable. This analysis provides a straightforward way to assess how strongly each individual feature correlates with the class DEG label, and it does so independently of any machine learning model.

The full results of the point-biserial analysis are provided in the appendix, starting with Figure 13. This includes a plot showing the feature importance ranking of all features, as well as two additional plots: one displaying only the 3-mer frequencies, and another showing all features excluding the 3-mers.

By comparing the feature importance rankings produced by the point-biserial (PB) analysis with those derived through SHAP values from the classifier, we can assess which feature-label relationships are consistently identified by both approaches.

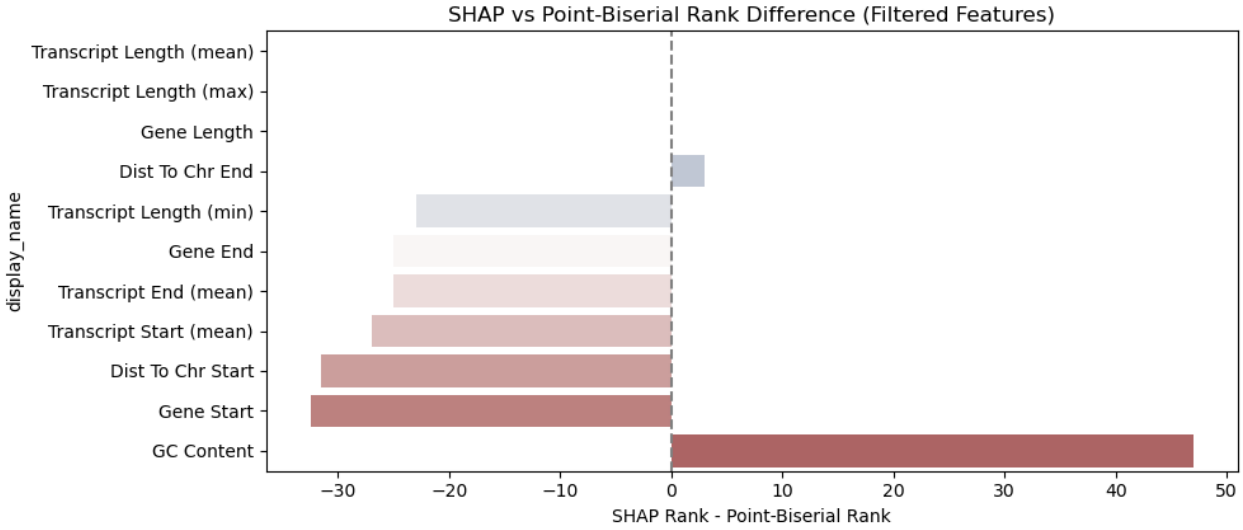


Figure 4: The plot shows the difference in feature importance rankings between SHAP and PB. The features on the plot are sorted by the absolute value of this difference. Bars extending right indicate features ranked higher by PB than by SHAP. The bars extending left indicate the opposite. The blue-to-red color gradient is solely a visual aid, and only signifies the absolute difference between the SHAP and PB ranks.

As is visible in Figure 4, there is no bar extending left or right for the top three features. This indicates a clear overlap between of the top features in both point-biserial and SHAP rankings. In other words, mean transcript length, maximum transcript length, and gene length are identified as the top three features, and in the same order, by both methods. In contrast, some features show notable differences; for example, GC content is ranked

significantly higher by the point-biserial method than by SHAP. A detailed interpretation of these differences is beyond the scope of this section and will be addressed in the Discussion.

Another important point to note is that the directionality of the features shown in Figure 4 is consistent between the point-biserial analysis and the SHAP results. In other words, both methods agree on how the value of each feature influences the class label. If a higher value of a feature favors one class and a lower value favors the other, this pattern is the same in both analyses.

4.2.4 A note on 3-mers

On the other hand, the 3-mer frequency features show virtually no agreement between the point-biserial and SHAP rankings. This is supported by Kendall's Tau correlation coefficient between the two rankings, which is -0.0893 ($p = 0.297$), and indicates no meaningful correlation. Moreover, the directionality agreement, i.e. the percentage of 3-mers for which both methods agree on whether higher or lower values are associated with a particular class, is only 48.44% (31 out of 64). These results suggest that the 3-mers lack consistent or interpretable predictive contribution when examined individually. A detailed comparison of the rankings can be found in Figure 16 of the Appendix.

However, including the 3-mer frequencies as features in the classifier still results in a 2-3% increase in overall accuracy. This unexpected improvement might mean that while the individual contribution of 3-mers may be difficult to interpret, they could still capture subtle, non-linear patterns that enhance model performance.

5 Discussion

As observed in Figure 1, genes overexpressed in older individuals tend to have a higher CG content, and in most tissues, this difference is found to be statistically significant. However, the classifier results suggest that CG content is not a major contributing factor in predicting whether a gene is overexpressed in young or old individuals. However, it's important to note that CG content ranks much lower in the classifier's feature importance ranking than it does in the one produced by point-biserial analysis. In fact, among the features shown in Figure 4, CG content exhibits the largest discrepancy in rank between the two methods. This difference likely a result of the nature of the analyses. The point-biserial correlation assesses the relationship between a single feature and the class label in isolation, while classifiers evaluate features in the context of all other features. In this case, the presence of 64 3-mer frequency features may render CG content redundant, as the classifier can potentially infer CG composition from the patterns present in those k-mer features. Nevertheless, when the 3-mers are excluded, CG content ranks among the most important features, just below the various gene and transcript length measures, as shown in Figure 13. Still, this may reflect the relatively low predictive value of the remaining features rather than the high utility of CG content itself.

As for the performance of the classifiers, there seems to be a clear pattern. Despite testing a variety of models with different levels of complexity, and even after rounds of parameter optimization, none were able to breach the accuracy threshold of approximately 66%. This plateau suggests that the limitation likely lies not with the classifiers themselves, but with the data being fed into them. In other words, the current feature set appears to lack sufficient predictive signal that would allow for a higher prediction accuracy. Supporting this interpretation is the overlap between the SHAP-based feature importance rankings

and those produced by the independent point-biserial analysis. Both methods highlight a roughly small difference between their top features (as evident in Figure 4), indicating that the classifier is likely learning the right things, but simply cannot extract more from what is available. This consistency reinforces the idea that the ceiling of what these features can offer was already reached. Therefore, improving performance must likely include a change to the preprocessing of the data before it is fed into the classifiers, or introducing entirely new types of gene-level characteristics as additional features. Specific directions for addressing this limitation are discussed in the Future Work section.

Finally, it can be said with a degree of confidence that gene length stands out as the most consistently predictive feature when differentiating whether a gene is overexpressed in young or old individuals. As can be deducted from Figures 12 and 13, longer genes tend to be overexpressed in young mice, while shorter genes are more commonly overexpressed in old mice. This pattern is not only supported by both the Random Forest and XGBoost classifiers (where gene length and related features such as the transcript lengths consistently rank at the very top), but also aligns with the independent point-biserial analysis. Furthermore, this correlation was discussed in prior research reflected upon in the introduction, reinforcing its biological plausibility. However, while gene length shows clear predictive value, it does not offer enough information on its own to classify gene expression patterns with high accuracy (Figures 10 and 7). Still, among all features tested, it appears to be the most informative, and it's reasonable to expect that any future, more accurate classifiers will include gene length as a key feature.

6 Conclusions and Future Work

This research set out to explore whether intrinsic physical characteristics of genes could be used to predict age-related changes in gene expression. Starting from the well-established observation that longer genes tend to be underexpressed with age, we aimed to identify additional features, such as GC content and proximity to chromosome ends, that might also contribute to this pattern. Using a combination of statistical tests and machine learning classifiers, we evaluated the combined predictive power of these features. While initial analyses confirmed previous findings regarding gene length and revealed moderate associations for GC content and chromosome position, the overall predictive performance of the combined feature set was limited. Classifiers plateaued at around 66% accuracy, suggesting that while the features capture some meaningful biological signal, they are not sufficient for high-accuracy prediction. Nonetheless, gene length consistently emerged as the most informative feature across both SHAP and statistical analyses, reinforcing its biological relevance. Ultimately, although the feature set tested here does not enable highly accurate classification, the study provides a possible framework for evaluating gene-level characteristics in the context of aging.

6.1 Limitations and Future Work

The future work suggestions outlined here stem largely from the main limitation of this study: the strict 10-week time constraint. As the research progressed, more and more questions emerged and several promising alternative methods and approaches came to mind.

One key area for further investigation is the role of 3-mer frequencies in classifier performance. As discussed in the Results section, these features show almost no correlation

with rankings from the statistical test, yet their inclusion improves classifier accuracy somewhat. Understanding why this happens, could shed light on hidden patterns not captured by traditional statistical methods. Additionally, a different approach to these features via different preprocessing of the gene sequences might prove fruitful.

Another important step would be to repeat this analysis using transcript-level data instead of gene-level data. Working at the transcript level would provide a larger dataset, since each gene can have multiple transcripts, allowing for more training examples while still testing the same set of features.

A logical next thing would be to formally test whether the current feature set has reached its predictive limit. While the evidence suggests a performance ceiling, this should be validated using dimensionality reduction techniques such as PCA, UMAP, or t-SNE. Showing explicitly that the data points based on these features are not cleanly separable would provide stronger justification for finding additional features.

Finally, a more exploratory type of future work could involve repeating the study with either the same data but differently processed features, or (preferably) by identifying new biologically meaningful gene-level characteristics. Extracting potentially richer and more specific features may unlock greater predictive potential and lead to a high-performing classifier.

References

[1] Olga Ibañez-Solé, Irantzu Barrio, and Ander Izeta. Age- or lifestyle-induced accumulation of genotoxicity is associated with a length-dependent decrease in gene expression. *iScience*, 26(4):106368, 2023.

[2] Sourena Soheili-Nezhad, Olga Ibañez-Solé, Ander Izeta, Jan H. J. Hoeijmakers, and Thomas Stoeger. Time is ticking faster for long genes in aging. *Trends in Genetics*, 40(4):299–307, 2024.

[3] Nicholas Schaum and Tabula Muris Consortium. All – a single-cell transcriptomic atlas characterizes ageing tissues in the mouse – 10x, 2020. Accessed: 2025-06-10.

[4] Ensembl. Biomart portal, 2024. Accessed: 2025-06-21.

[5] Genome Reference Consortium. Grcm39 genome assembly for *Mus musculus*, 2024. Accessed: 2025-05-25.

[6] Ensembl. Grcm39 cdna fasta file for *Mus musculus*, 2023. Downloaded from Ensembl, GRCm39 assembly.

[7] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. Explainable ai for trees: From local explanations to global understanding, 2019.

A Gene Annotations Final Form

Figure 5: List of the gene features on which the classifier will be trained and tested.

- `gene_stable_id`
- `gene_name`
- `gene_type`
- `chromosome`
- `strand`
- `gc_content`
- `gene_start`
- `gene_end`
- `transcript_length_mean`
- `transcript_length_median`
- `transcript_length_min`
- `transcript_length_max`
- `transcript_start_mean`
- `transcript_end_mean`
- `dist_to_chr_start`
- `dist_to_chr_end`
- all 64 possible 3-mer frequencies, including: `AAA`, `AAC`, `AAG`, `AAT`, `ACA`, ..., `TTT`

B Gene Characteristics Supplementary Figures and Tables

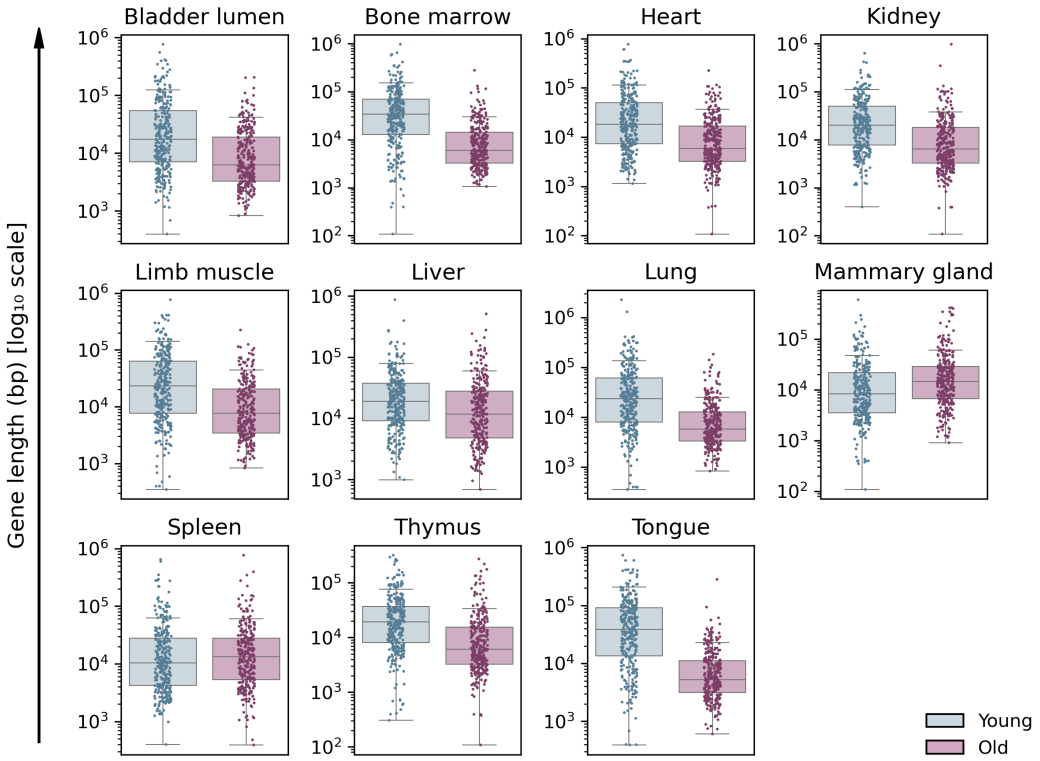


Figure 6: Each panel shows the gene length distribution of the 300 most differentially expressed genes overexpressed in young (blue) and old (purple) samples for a given tissue. Boxplots display summary statistics, while the overlaid dots correspond to individual genes.

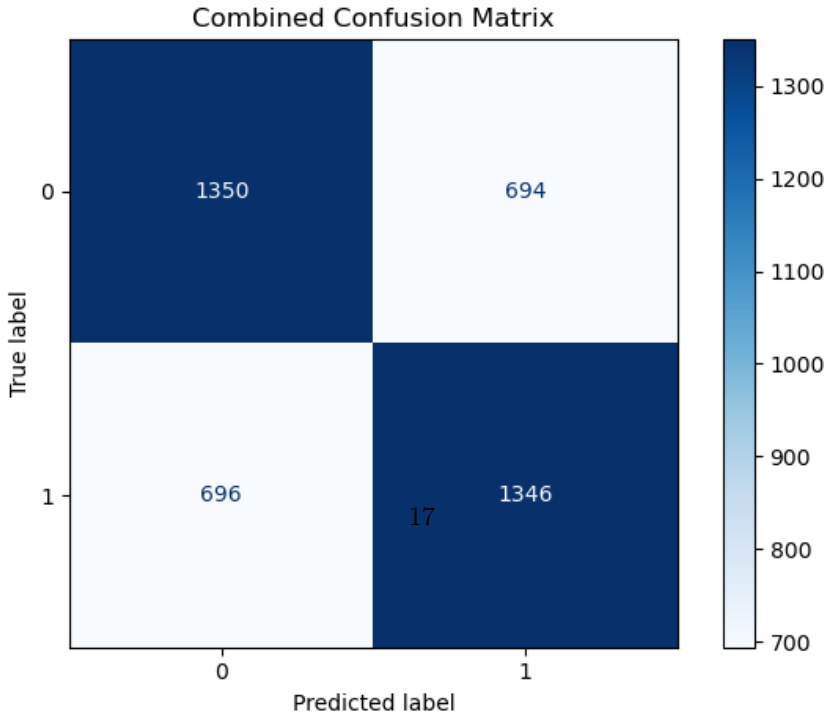
Table 2: Mann-Whitney U test results comparing gene length of top 300 genes overexpressed in old vs. top 300 genes overexpressed in young mice across tissues.

Tissue	U statistic	p-value
Bladder lumen	61391.5	1.16×10^{-14}
Bone marrow	70629.5	1.49×10^{-33}
Heart	63491.0	3.06×10^{-18}
Kidney	64075.0	2.60×10^{-19}
Limb muscle	62416.0	2.35×10^{-16}
Liver	54586.5	6.33×10^{-6}
Lung	68143.0	1.15×10^{-27}
Mammary gland	34558.5	8.75×10^{-7}
Spleen	42096.5	1.72×10^{-1}
Thymus	64473.0	4.65×10^{-20}
Tongue	75107.0	1.21×10^{-45}

Table 3: Mann-Whitney U test results comparing CG content of top 300 genes overexpressed in old vs. top 300 genes overexpressed in young mice across tissues.

Tissue	U statistic	p-value
Bladder lumen	44670.0	0.8767
Bone marrow	26362.5	1.66×10^{-18}
Heart	35550.5	8.56×10^{-6}
Kidney	36963.0	1.54×10^{-4}
Limb muscle	39765.5	1.37×10^{-2}
Liver	37049.0	1.81×10^{-4}
Lung	38732.5	3.16×10^{-3}
Mammary gland	36966.5	1.55×10^{-4}
Spleen	44596.0	0.8493
Thymus	33108.0	2.13×10^{-8}
Tongue	26319.0	1.38×10^{-18}

C XGBoost Plots



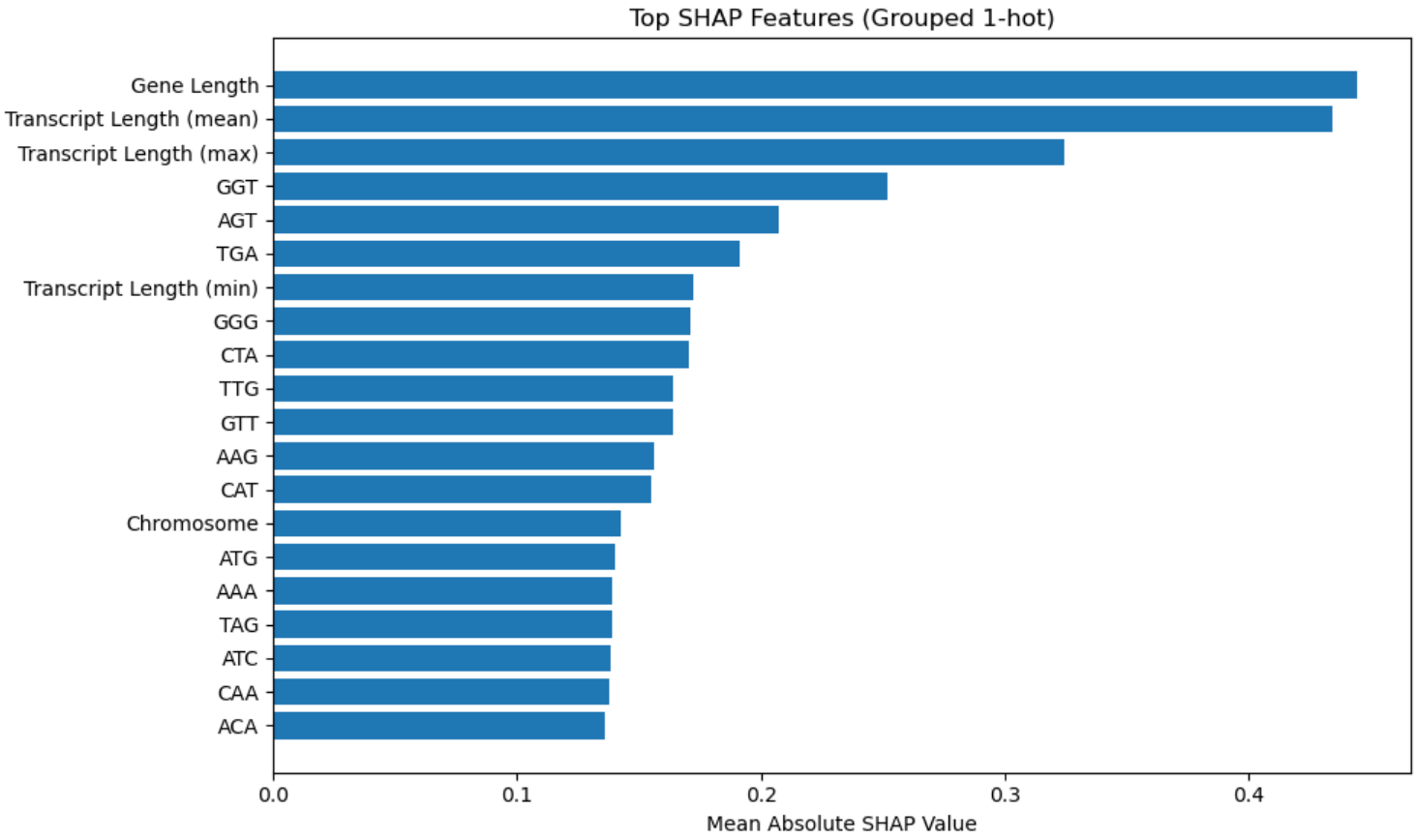


Figure 8: XGBoost. All Tissues. Combined Feature Importance Ranking.

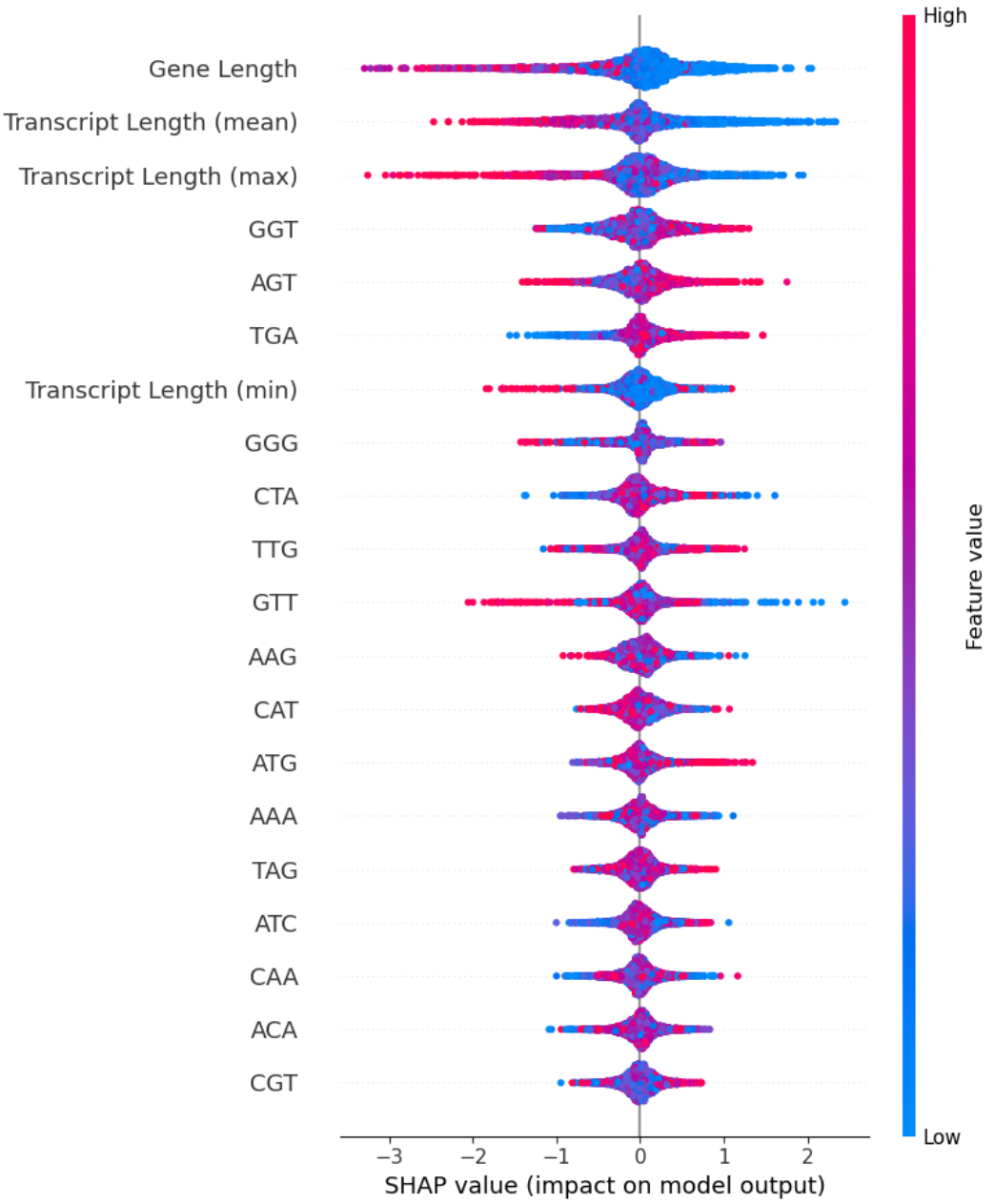


Figure 9: XGBoost. All Tissues. Combined SHAP Beeswarm.

D Random Forest Plots

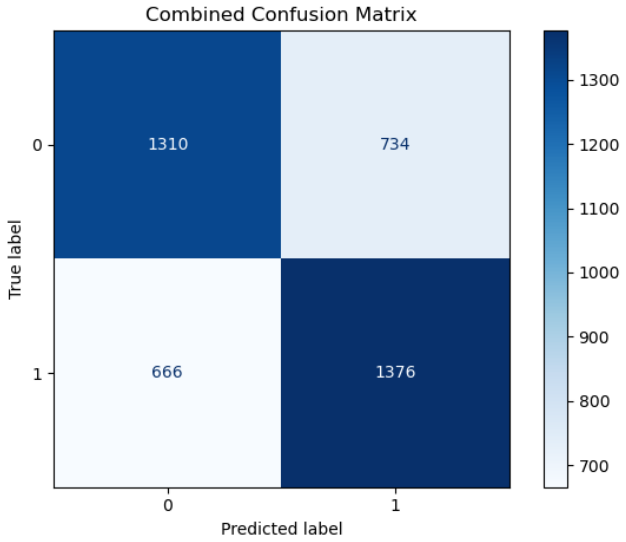


Figure 10: Random Forest. Confusion Matrix. All Tissues Combined

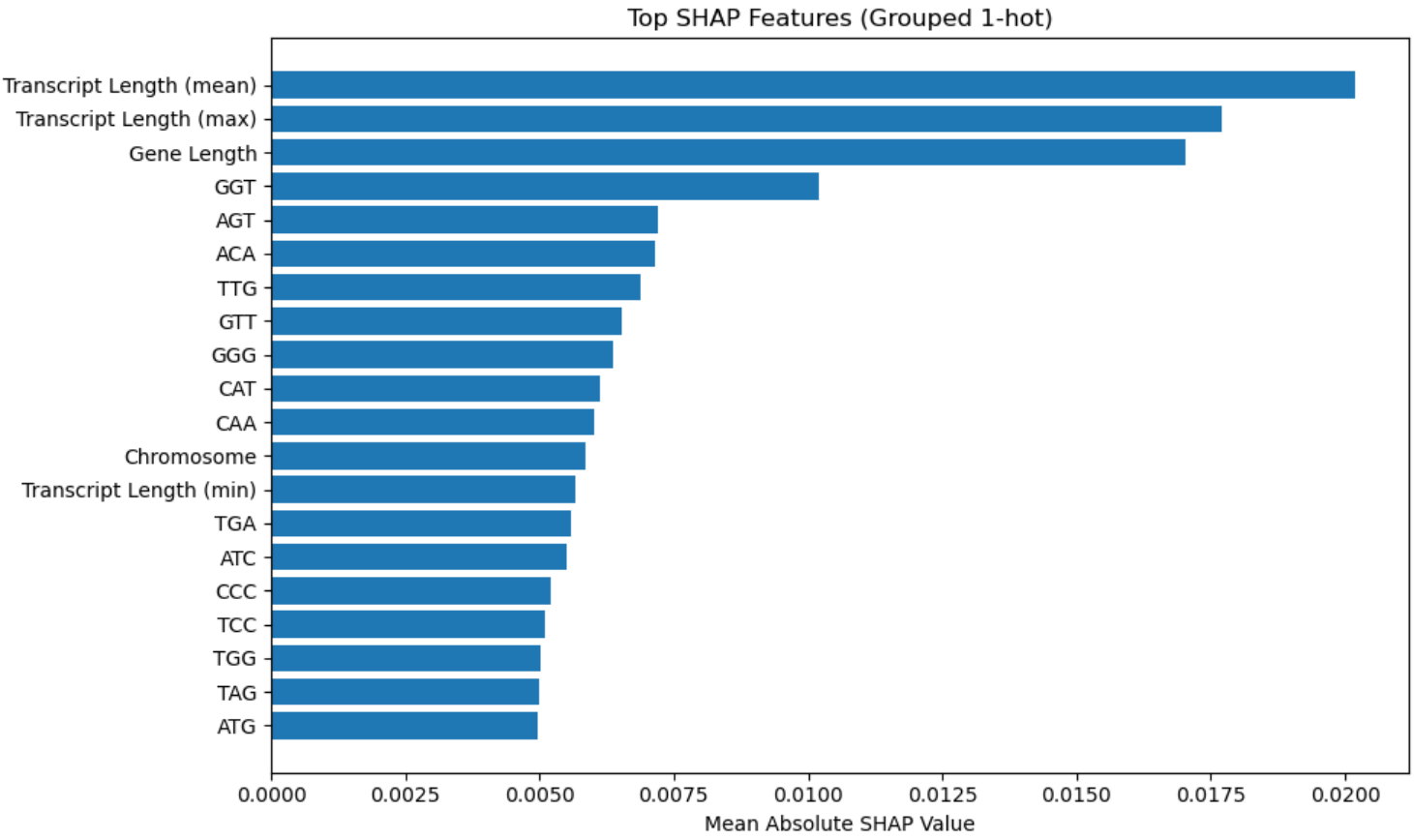


Figure 11: Random Forest. Feature Importance Ranking. All Tissues Combined.

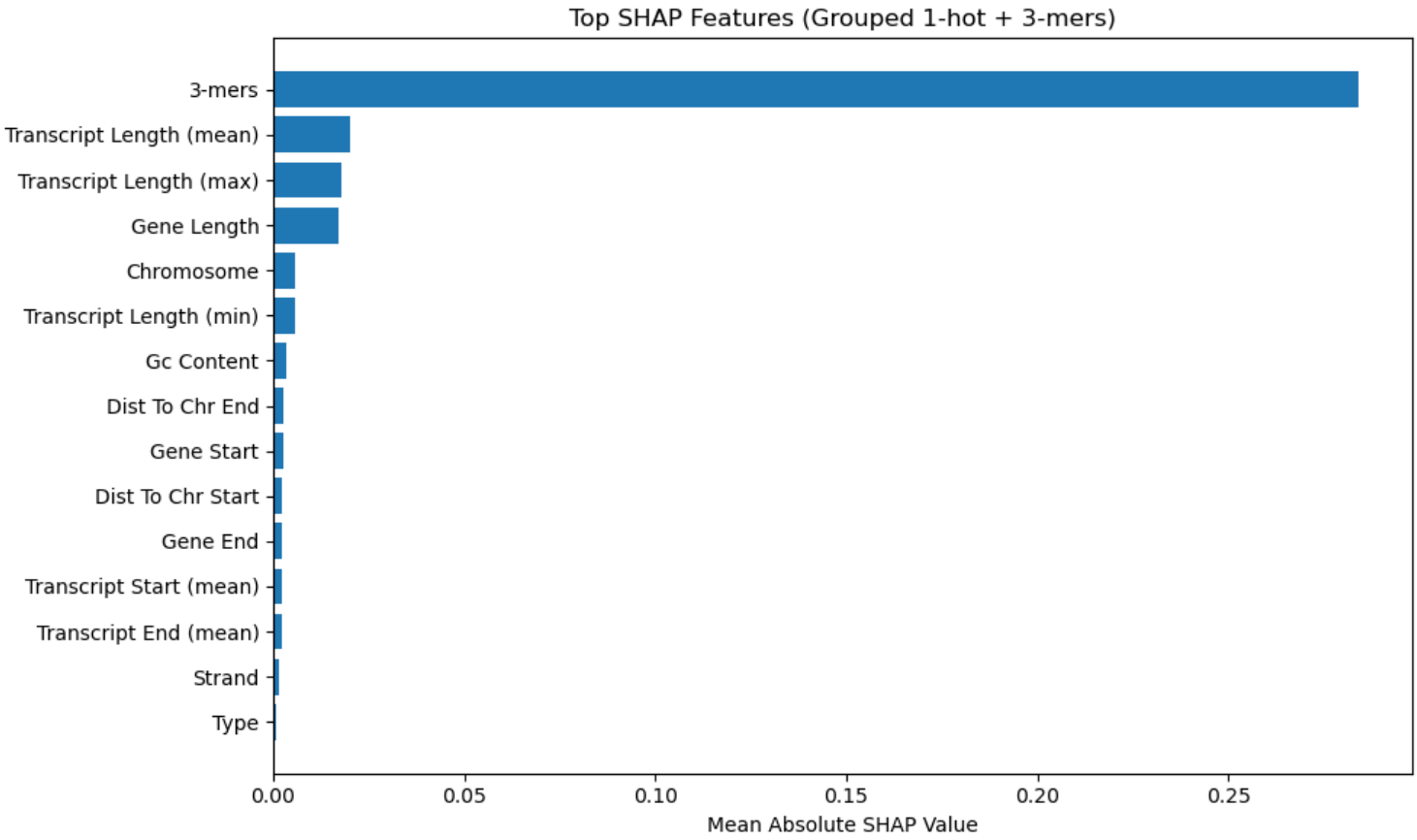


Figure 12: Random Forest. All Tissues. Combined Feature Importance Ranking. 3-mers grouped.

E Point-Biserial Analysis Plots

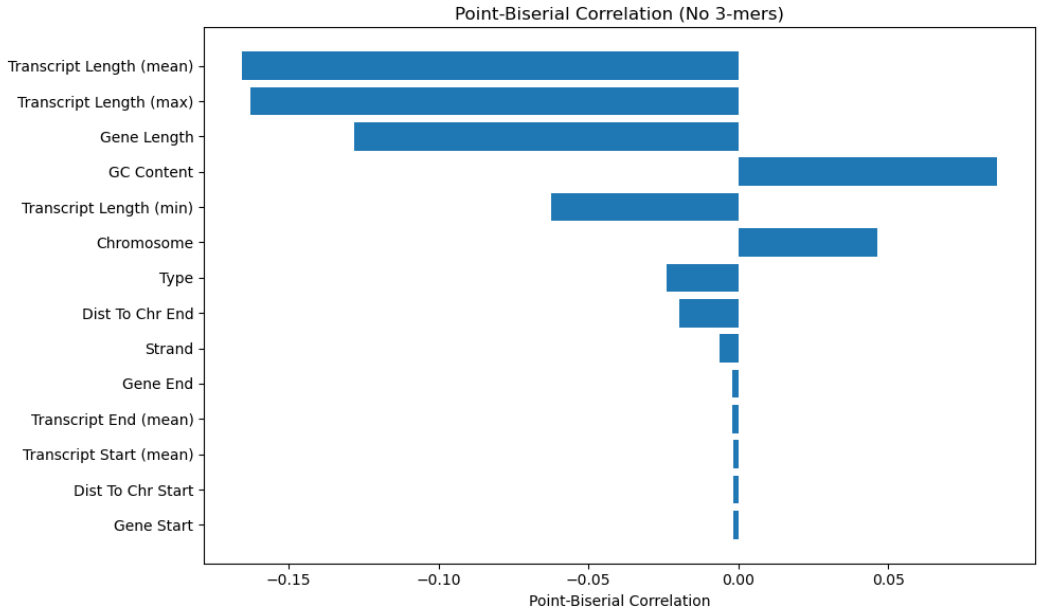


Figure 13: Feature Importance Ranking. 3-mers excluded.

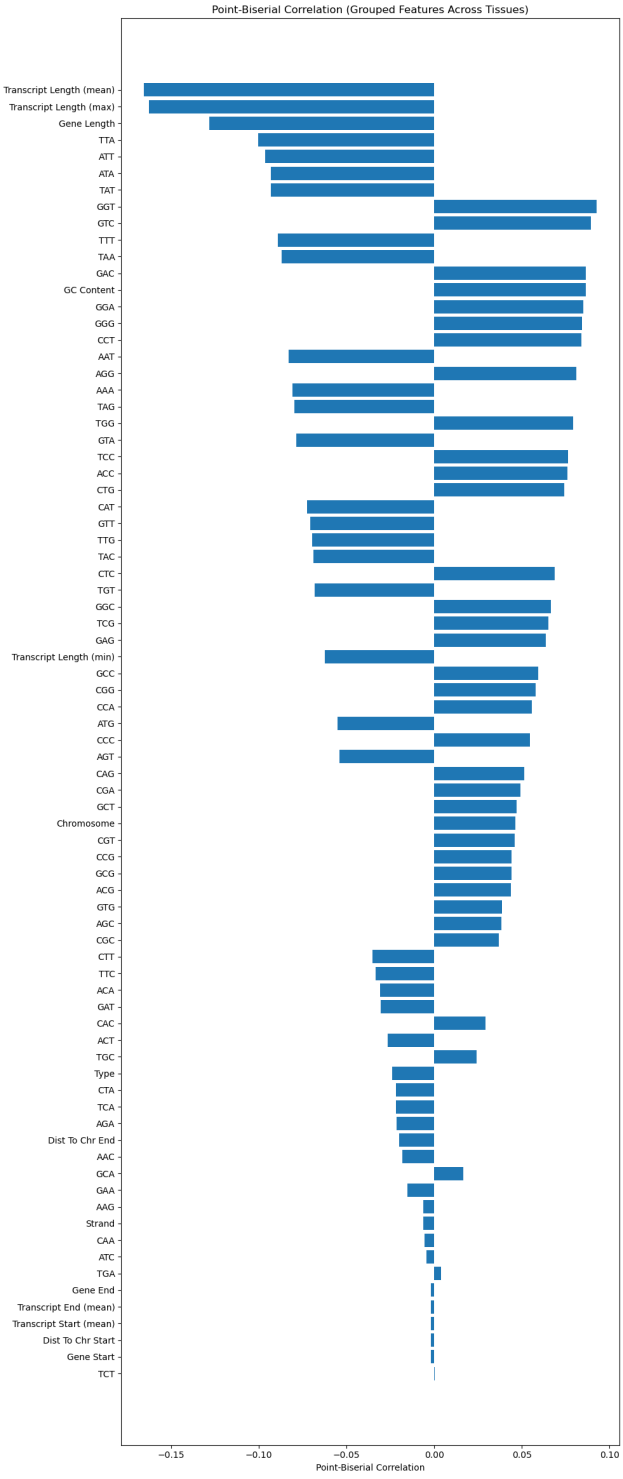


Figure 14: Feature Importance Ranking. All Featres.

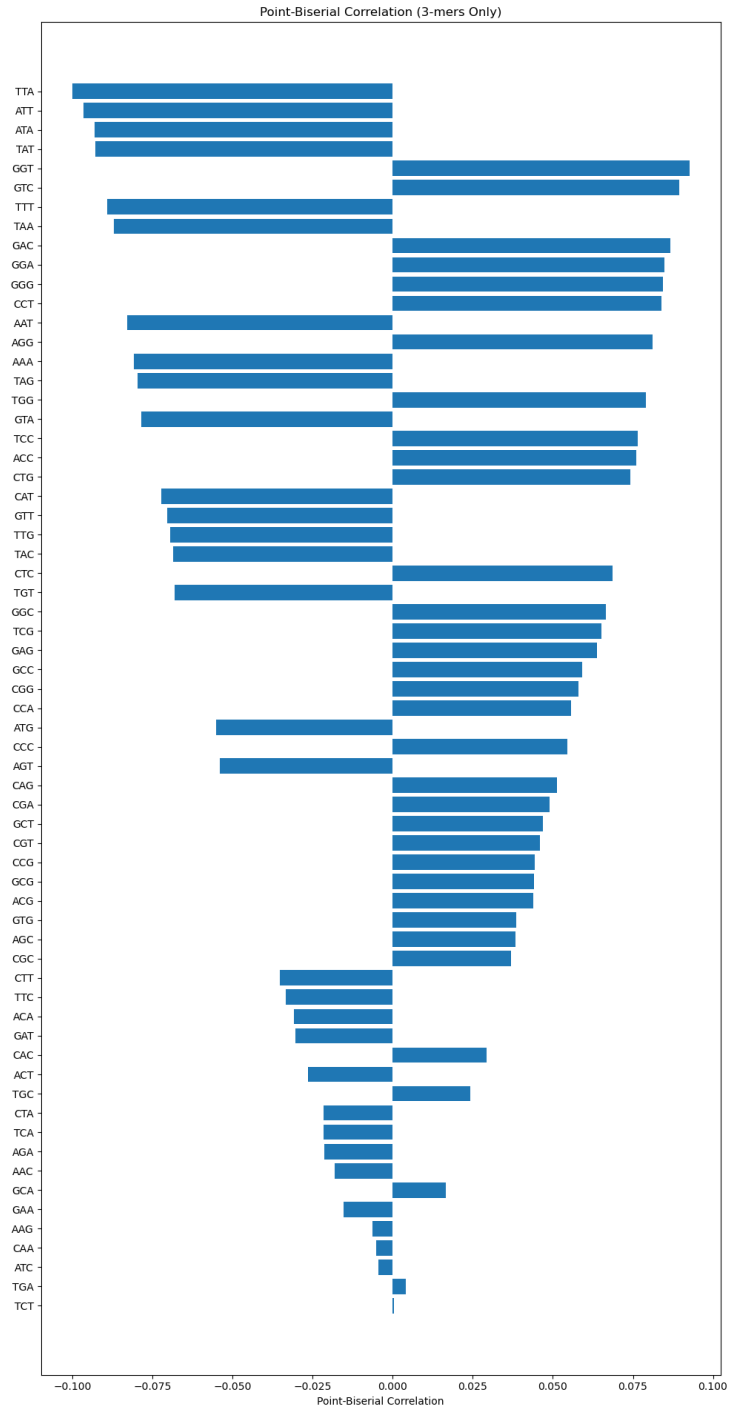


Figure 15: Feature Importance Ranking. Only 3-mers.

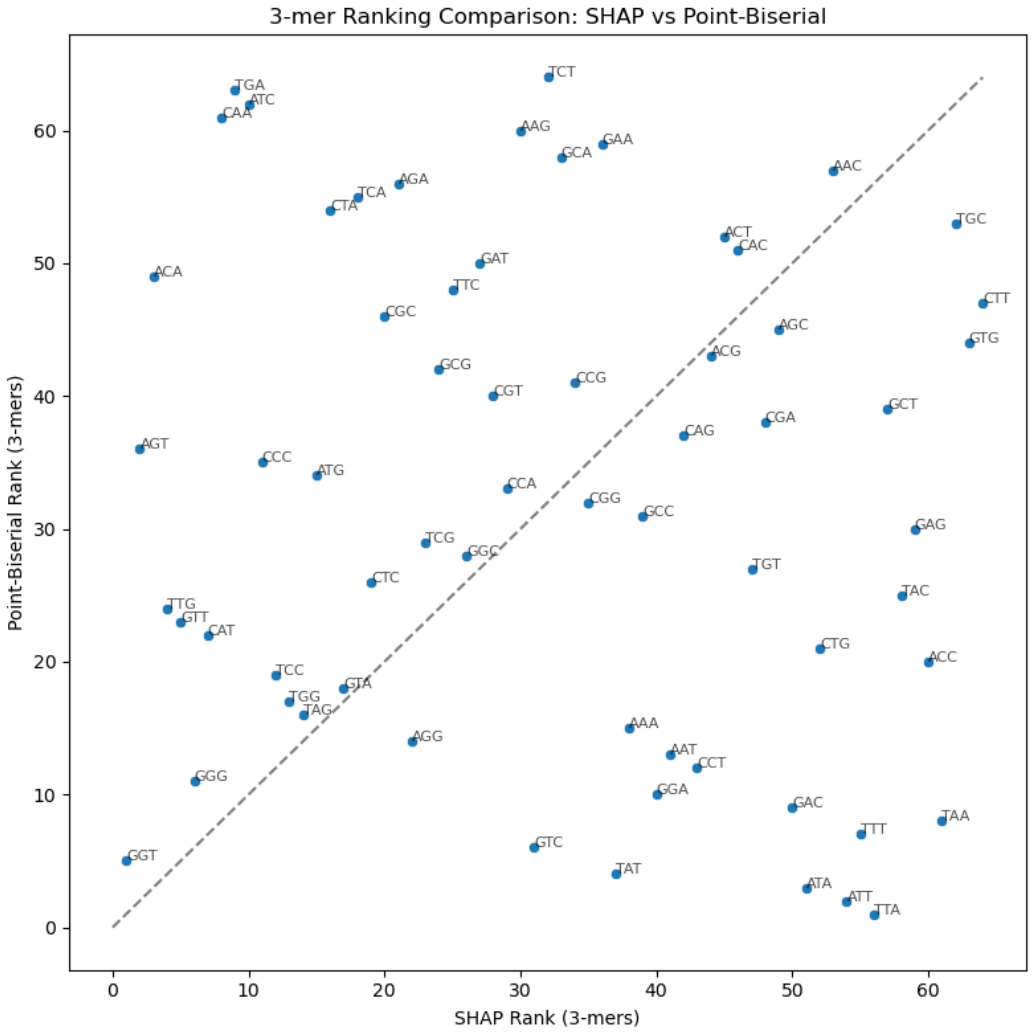


Figure 16: Each dot is 3-mer. The plot shows the correlation between the SHAP ranking of the 3-mer and the PB ranking. The line signifies a perfect match between the rankings.