Predicting epimutation rates based on DNA sequence and chromatin information

Vorhersage von Epimutationsraten auf Grundlage von DNA-Sequenz und Chromatinzuständen

Wissenschaftliche Arbeit zur Erlangung des Grades
B.Sc.
an der TUM School of Life Sciences der Technischen Universität München.

Betreut von Prof. Frank Johannes
Lehrstuhl für Pflanzenepigenomik

Eingereicht von Constantin Goeldel
Vimystraße 1c
85354 Freising
constantin.goeldel@tum.de

Eingereicht am 10.01.2024 in Freising

Erklärung

Ich versichere hiermit, dass ich die von mir eingereichte Abschlussarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Constantin Goeldel

Freising, 10.01.2024

Bachelor's Thesis
Predicting Epimutation Rates from Sequence and Chromatin Information

Constantin Goeldel
Department for plant epigenomics, supervised by Prof. Dr. Frank Johannes
Technical University of Munich, School of Life Sciences, Emil-Ramann-Str. 4 85354 Freising, Germany
(Dated: January 10, 2024)

Die Mechanismen, die der Etablierung und Aufrechterhaltung der DNA-Methylierung zugrundeliegen, sind noch nicht vollständig aufgeklärt. Dabei ist diese epigenetische Modifikation ein wichiger Bestandteil der Regulation der Genexpression und kann ortsabhängig Gene silencen oder deren Transkriptionsrate erhöhen. In Pflanzen kommt 5'-Cytosin-Methylierung in den drei Kontexten CG, CHH und CHG (H = A, T, C) vor. Aus charakteristischen Kombinationen dieser Kontexte lassen sich Genklassen definieren, die sich hinsichtlich Expressionsverhalten, Methylierungslevel und Epimutationsraten unterscheiden. In dieser Bachelorarbeit verbinde ich diese Eigenschaften mit divergierenden Anreicherungen von Chromatinzuständen entlang der Gene und zeige, dass Entstehung und Verlust von CG-Methylierung mit nur wenigen Prediktoren akkurat vorhergesagt werden kann. Zudem präsentiere ich eine Datenstruktur und Modellarchitektur, mit der die Prognose von Epimutationsraten auf das gesamte Genom ausgeweitet werden kann.

I. INTRODUCTION

Epigenomics is the study of stably heritable phenotypes resulting from changes in a chromosome without alterations in the DNA sequence[1]. Developing an understanding of the processes modulating the inheritance of epigenomic information carriers has long been a goal of the field due to its central role in embryonic development, aging and disease development[2–4]. Among the epigenetic modifications influencing the phenotype are cytosine methylation, histone modifications & variants, nucleosomal positioning as well as the three-dimensional configuration of the genome inside the nucleus, which regulate the transcriptional activity of nearby genes[5]. While covalent modifications are usually preserved throughout cell divisions, marks can be spontaneously lost or established *de novo*[6]. In the case of symmetric CG or CHG methylation contexts (where H = A, T or C), this may occur through stochastic errors in maintenance methylation when restoring the original cytosine methylation onto the newly synthesised strand after DNA replication[7]. However, the rate of gain (α) and loss (β) of cytosine methylation is not constant but rather varies by sequence and chromatin context[8]. In this work, I aim to contribute to understanding the factors influencing the variability of the epimutation rate.

Recent progress has been aided by improvements in measurement technologies such as *Whole-Genome-Bisulfite-Sequencing* (WGBS), *Chromatin ImmunoPrecipitation and DNA-Sequencing* (ChIP-Seq) and the decreasing cost of Next-Generation Sequencing (NGS)[9]. This allows for large scale quantitative analyses of the epigenome at site-level resolution. To estimate the mutation rate of cytosine methylation, *Mutation Accumulation Lines* (MA-Lines) are used, in which a species of interest is sequenced at different points in time in a clonally derived pedigree (Fig. 1). In a pairwise comparison, the methylation divergence between any two samples can be correlated with the generational time difference, from which epimutation rates can be deduced[10]. However, the cultivation and sequencing of the MA-Lines take particular time and resources, especially as generational times of the studied organisms increase. Experimental adjustment can partially overcome this shortcoming, for example by treating branches in trees as analogous to offspring in the pedigree, which enables age-estimations and phylogenetic explorations[11]. A general method of predicting epimutation rates and stable state methylation levels nevertheless remains desirable for reducing experimental iteration time while hopefully unveiling biological patterns itself.

II. METHODS

A. Estimating epimutation rates from MA-Lines

The generational time difference between any two samples in the pedigree is dependent on the generation of the latest common ancestor t_{ij} (Fig. 1). The divergence between two compared samples is the absolute difference between the methylation states of the equivalent diploid loci in both samples where

$$s(k) = \begin{cases} 1 \text{ if Methylated (m/m)} \\ \frac{1}{2} \text{ if Differentially Methylated (m/u)} \\ 0 \text{ if Unmethylated (u/u)} \end{cases}$$

averaged over all sites.

$$D_{ij} = \sum_{k=1}^{N} |s_{ik} - s_{jk}| \, N^{-1}$$

Applied for all pairs of samples, a divergence matrix is obtained, which is then used to fit a model of 5mC divergence where ϵ_{ij} is the normally distributed residual error, c is the intercept and $D^{\bullet}_{ij}(M_{\Theta})$ is the expected divergence between samples i and j as a function of an

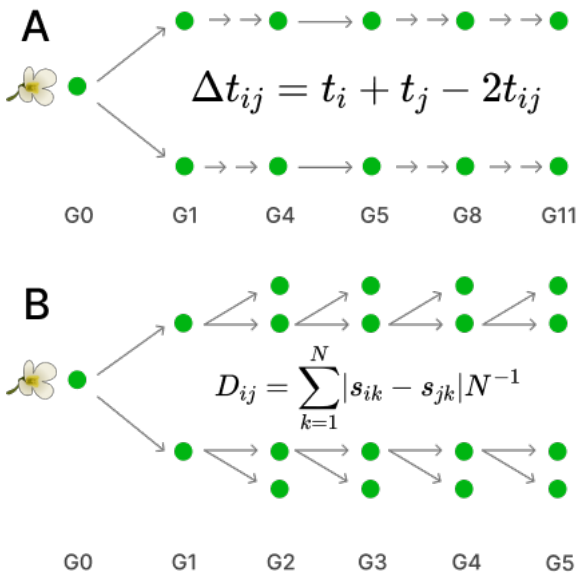

FIG. 1: Pedigrees of the Col-0 *A. Thaliana* MA-Lines used to estimate the epimutation rates. Upon sequencing the genomic DNA after bisulfite conversion, and alignment to the TAIR10 reference genome, methylation states were called using Methimpute[12]. **A:** (GSE153055)[10] **B:** (GSE204837)[13].

underlying epimutation model M with parameter vector Θ[10].

$$D_{ij} = c + D^{\bullet}_{ij}\left(M_\Theta\right) + \epsilon_{ij}$$

Without repeating the exact algorithm for determining the state transition model M, which is outlined in the original *AlphaBeta* paper[10], it is crucial to note that optimising $D^{\bullet}_{ij}\left(M_\Theta\right)$ using

$$\nabla \sum_{q=1}^{\binom{n}{2}} \left(D_q - D^{\bullet}_{q}\left(M_\Theta\right) - c\right)^2 = \mathbf{0}$$

where n is the total number of samples in the pedigree scales in $O(n^2)$. It is further slowed by the repeated matrix multiplication necessary for each iteration of the *Nelder-Mead minimisation* algorithm. Using the original implementation of the *AlphaBeta* software therefore effectively prohibited the estimation of the epimutation rates for sufficiently small slices of the *A. Thaliana* genome. Significant effort has been spent to increase the performance of the implementation, e.g. by porting it from the interpreted R version to the compiled language Rust, using a highly optimized matrix-multiplication library, running the calculations for each pair in parallel and switching to a column-oriented data storage engine which allows efficient vectorization of the divergence calculation. Combined, these improvements speed up the estimation by more than 200 times for the datasets used in this work.

This speedup moves the bottleneck in estimating epimutation rates for every cytosine in the genome from calculation to data availability: As a rule-of-thumb, there should be enough methylatable sites included in the divergence estimation that at least one mutation can be expected per generation of the pedigree. With genome-wide gain rates (α) and loss rates (β) at around 10^{-4} per CG site per haploid genome per generation[7], and two (1A) or four (1B) samples per generation in the MA-datasets, I divided the genome into overlapping slices of 512 CG-Dinucleotides each offset by 64 nucleotides. For each of the resulting 87.000 slices, AlphaBeta was run to estimate the epimutation rates.

B. Metaprofiles: Assessing variability of epigenomic traits grouped by arbitrary features

Slicing the genome into windows along the whole genome is one method to overcome the lack of data for single site estimation. Another, which we first applied in *Stochasticity in gene body methylation*[14], is to group sites by a common annotation, for example membership in certain types of genes, reasoning that sites with common features also share similar epimutation characteristics. The set of sites that belong to said feature can then be further divided into subsets, e.g. windows along the length of the feature, to capture variances that occur within it.

A pseudocode algorithm to implement this, with an arbitrary number of windows, also including the up- and downstream regions of the feature, might look like this:

```
SELECT Count(*) as num_sites,
       AlphaBeta(*) as (alpha, beta),
       SteadyStateObs(*),
       SteadyStatePred(alpha, beta) ,
       NUM_WINDOWS *
            Abs(s.location - (f.strand = "+" ?
                f.start : f.end))
            / (f.end - f.start) as percentile
FROM sites s
JOIN features f ON s.chromosome = f.chromosome and
    s.strand = f.strand
WHERE s.location + OFFSET >= f.start and s.location
    - OFFSET <= f.end
GROUP BY percentile, f.feature
```

The functions used in the code are defined as follows: The observed methylation level for any set of sites **S** is given by the methylation state s_k, averaged over all sites in the set.

$$M(S) = \frac{\sum_k^N s_k}{N}$$

In a selfing system, the methylation level can also be predicted from the epimutation rates $\alpha_S = P(\mathrm{CG} \to \mathrm{mCG})$ and $\beta_S = P(\mathrm{mCG} \to \mathrm{CG})$ using a model developed by *van der Graaf et. al*[15]:

$$\widehat{M}\left(S, t_\infty\right) = s(\widehat{\pi})$$

where the probability of being in any of the possible methylation states is given by:

$$\begin{aligned}
\widehat{\pi}\left(c^u c^u\right) &= \frac{\beta\left((1-\beta)^2 - (1-\alpha)^2 - 1\right)}{(\alpha+\beta)\left((\alpha+\beta-1)^2 - 2\right)} \\
\widehat{\pi}\left(c^u c^m\right) &= \frac{4\alpha\beta(\alpha+\beta-2)}{(\alpha+\beta)\left((\alpha+\beta-1)^2 - 2\right)} \\
\widehat{\pi}\left(c^m c^m\right) &= \frac{\alpha\left((1-\alpha)^2 - (1-\beta)^2 - 1\right)}{(\alpha+\beta)\left((\alpha+\beta-1)^2 - 2\right)}.
\end{aligned}$$

Previously, we applied this method of combining chromatin information and epimutation rates to certain subsets of the genome. Specifically, we analyzed gene body methylated (gbM) genes, which are extensively methylated on CG dinucleotides, but not on cytosines in sequence contexts CHG and CHH[16]. They also have a higher epimutation rate than the average of the whole genome[14]. Dividing them further into subsets of sites grouped by the relative position of a dinucleotide within its respective gene, we created metaprofiles of CG density, H2A.Z occurrence, epimutation rates and steady state methylation. Curiously, while gbM gene methylation levels are stable when averaged over the whole gene, they are highly dynamic at individual sites. This is driven by an increase of the methylation gain rate (α) towards the center of the gene[8,14].

Creating metaprofiles is not constrained to just DNA bases, rather being generally applicable to all annotations that can be stably located on a continuous chromosome. A weak constraint is that the assorted feature should be smaller than the annotation by which it is grouped, otherwise the placement in a percentile window becomes arbitrary. However, this can be overcome by reducing the length of the feature, for example by using its midpoint as the reference location. Currently, the metaprofile software library supports computational biology formats like BED, GFF, BigWig and several formats of histone modification and heterogeneity score files. Additionally, it supports arbitrary, user-defined data formats through its Python API and Polars dataframe backend.

In this work, I extend the metaprofiles to UM and teM genes and analyze the relative frequency of additional histone modifications as well as all chromatin states from *Zhang et al*[17]. For this reduced dataset, I build a multiple regression model to predict epimutation rates from the feature vector of each window along the metaprofile. The features that provide the most information towards predicting the epimutation rates were selected by a Sequential Feature Selector, using both the forward and backward pass options.

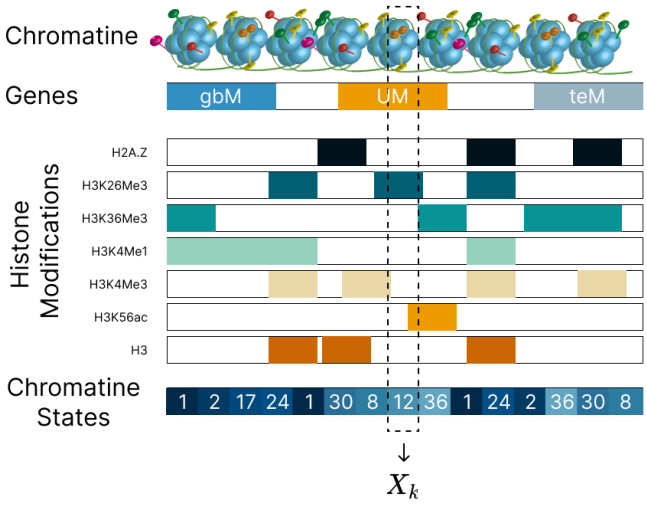

FIG. 2: Composition of the feature vector X_k describing a genetic locus k. DNA bases are encoded using a dictionary mapping Base Letter → Int8. Gene types are treated accordingly. The presence of a histone modification is encoded as a binary flag and chromatin state are represented by their state number. Nucleosome chain image copied from the Michael Goldstein Lab of the Washington University School of Medicine in St. Louis.

C. Describing the epigenome by a feature vector

We previously[14] postulated defining a feature vector X based on DNA sequence and/or chromatin information which can uniquely define all n mutational subsets S_n. Such a subset is here defined as the set of all sites in the genome that share the same epimutation rates. Each subset S_n has its own corresponding gain rate α_{S_n} and loss rate β_{S_n}. To answer questions about the characteristics of this feature vector, such as the choice and minimum number of features necessary, a set partition function which can accurately map the feature vector of a specific site to its mutational subset S_n is required. In this work, I propose a particular formulation of the feature vector $\mathbf{X}$ and evaluate its eligibility for predicting $\mathbf{X} \rightarrow (\alpha, \beta)$.

What are the fundamental upper and lower bounds on the complexity of such a feature vector? The most generic possible formulation might be just the whole DNA sequence of the organism in question, as with infinite computation resources, a physically accurate simulation of the full organism should be able to reproduce the original completely[18]. At the other extreme, providing the state of the entire organism, down to the position of each molecule is certainly sufficient feature information but equally unachievable.

I used a feature vector X_k that is uniquely defined for all k DNA bases in the genome (although due to data-availability, my analysis is limited to CG-Dinucleotides), combining sequence, genetic, histone modifications and

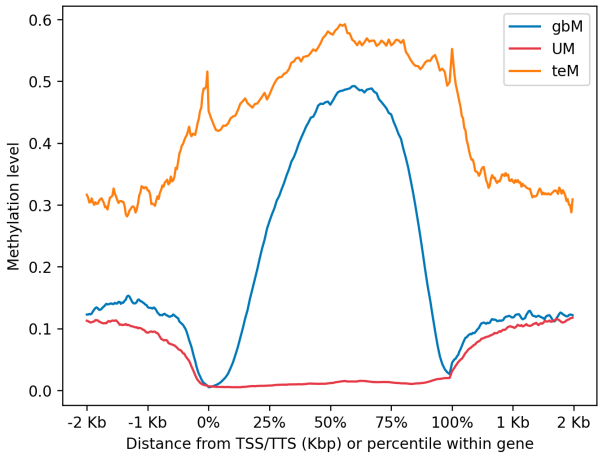

FIG. 3: Metaplot of CG steady state methylation in different gene types. Metaprofile windows span 1% of the gene length and include sites up to 2Kbp up- and downstream of the gene.

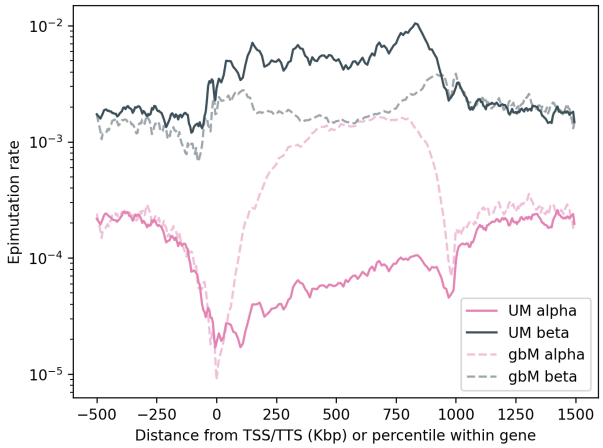

FIG. 4: Metaprofile of methylation gain rates α and β in sliding windows along UM and gbM genes. Using a logarithmic scale, the change in both rates becomes visible.

variants as well as chromatin state information:

$$X_k = \begin{pmatrix} \text{DNA Base} \\ \text{Gene Type} \\ \text{Histone Modification}_1 \\ .. \\ \text{Histone Modification}_n \\ \text{Chromatin State} \end{pmatrix}$$

where DNA Bases are encoded as $G \to 1, C \to 2$, etc., gene type is encoded as gbM $\to 1$, UM $\to 2$, teM $\to 3$, chromatin states as their respective state number and histone modifications - as they are the only used epigenetic markers that overlap at single sites - as binary according to their presence at site k (Fig. 2). After experimenting with several different feature representations, I chose this set of features mostly due to data availability, but also due to it's highly regular structure and ease of computability. Noticeably absent are features containing information about a site's chromatin neighborhood, for example nucleosomal positioning or contacts in topologically associated domains, which would probably be beneficial to add in future iterations.

III. RESULTS

A. Intragenic methylation levels vary between gene types

The metaplots of observed steady state mCG methylation levels in the three different analysed gene types show a deviation from background methylation levels. UM genes contain almost no steady state methylation while the gbM genes are characterized by low methylation towards the 5' and 3' end and high methylation levels in the center. The methylation level of teM genes (which are methylated in CG and non-CG contexts CHH and CHG) also rise throughout the gene body but - unlike CG - never drop below the level of their surrounding dinucleotides. This teM background methylation level is also higher than that of gbM and UM genes. The difference in methylation reflects the role of mCG as a regulator of gene expression. Usually, it acts as a repressor of CpG-dense promotors[19]. This is in accordance with the increased mCG level in teM genes, which have very low transcriptional activity[20] However, increased steady state methylation in gbM genes, which are moderately expressed[21], has the opposite effect, with a small but significant gain in expression levels[22]. The most common of the analyzed gene types, UM genes, possess no intragenic methylation. While gbM genes are mostly constitutively expressed, UM gene activity is more dependent on tissue or cell type[23].

B. Change in steady state methylation is driven by change in 5mCG gain rate (α)

When comparing each gene types' steady state methylation levels with their corresponding α and β metaprofiles (Fig. 4), it appears that in all gene types, mCG correlates more strongly with the gain rate rather than the loss rate. The effect is strongest in gbM, were towards the center of the gene, α rises to eight times its background level, matching the several fold increase in mCG while β rises only moderately. This is reflected by the Pearson correlation coefficient (PCC), which is 0.96 for α but only -0.14 for β. The absence of mCG towards the 5' and 3' end of the gene meanwhile correlates with a combined fall in α and increase in β in these regions.

The rise in methylation towards the center of teM genes similarly correlates with an increase in α (PCC:

0.73) around the center while β remains constant (PCC: -0.20) (Fig. 5).

UM genes differ from the previous two gene types in that their absence of intragenic CG methylation is the result of **both** an increased loss rate (PCC: 0.95) as well as a decreased gain rate (PCC: -0.77) (Fig. 4) . Together, they drive an almost complete absence of mCG in UM genes. Outside of the gene body, both rates track their counterpart in gbM metaprofiles closely. The decline in gain rates at the transcription start site between UM and gbM are similar, hinting at a shared pathway.

C. Distribution of epigenetic marks along gene metaprofiles

Combining the metaprofiles of steady state methylation, epimutation rates, CG density and H2A.Z from *Stochasticity* with additional histone modifications and 36 chromatin states from the *Plant chromatin State Database*[24], each in gbM, UM and teM gene contexts yields figures 5 and 6. The occurrence of epigenetic marks along the windows of the gene metaprofiles is highly variable, with almost every chromatin feature having a unique distribution.

Despite their uniqueness, the frequency distributions of different features can be grouped into different classes: First, some features are generally not present in and around the analyzed genes. This class includes chromatin states 30 to 36 which broadly correspond to heterochromatin[8,25,26] and are found in intergenic regions, in transposable elements as well as close to the centromere[24]. This explains their absence in gene bodies. The noisy pattern of these states in teM genes is a consequence of their rare presence in these already less common genes, leading to small sample sizes, often just one or two occurrences per window. For this reason, I will omit teM from further analysis.

A second class of features shows a sustained decrease of intragenic occurrence levels from background presence in all gene contexts. This includes S14, S15 and S21 with an especially strong effect in S16, S20, and S29 (all significant at $p < 0.01$ using Welch's t-test, rejecting the null hypothesis that enrichment levels have the same mean in intragenic and up- and downstream regions). Previously, these states have been broadly classified as occurring in and around promotors[8,24]. This clustering is compatible with their increased presence just upstream of the transcription start site but does not account for their almost equally high enrichment downstream of the transcription termination site. A possible explanation might be genes that are read on both strands, like the antisense transcript of FLOWERING LOCUS C in the *A. thaliana* vernalisation pathway[27]. These can arise from independent promoters, bidirectional promoters of divergent transcription units or cryptic promoters[28,29], which would likely contain the familiar chromatin state combinations. Additionally, promotor-like chromatin states downstream of a gene might be part of the promotor of the next gene if they are located within two kilobases of each other which is the offset chosen for this analysis. The alternative explanation, namely that these states are not as closely constrained to promotors as previously assumed, must also be considered: In the case of S16 and S29, the *Plant chromatin state database* already includes intergenic regions in the referential locations of these states. This addition is worth considering for the remaining states in this class.

Several other chromatine states previously clustered as promotor-related fit this description more accurately: States 17, 18, 19 and 22 - 26 show a procession of spikes wandering from the promotor towards the TSS and into the gene, stopping at around 10% of the length of the gene. They are sharpest just around the TSS, an effect that is strongest in gbM genes. Histone modifications H2A.Z, H3K4me3 and K3K56ac also peak close to the TSS. It is a look inside the delicate machinery regulating gene expression. All these chromatine states have previously been clustered as promotor states with some crossing into 5'UTRs and coding sequences[8,24], a decision based upon their accessible DNA and enrichment of histone acetylation. The metaprofile reveals the opportunity to further divide these states into finer-grained subcategories according to their relative position and role in controlling transcription.

Other chromatin features are also concentrated around specific locations: S1 and S2 are clustered around the TTS while CG density, S20, S21 and S29 peak at both ends of the gene. Others, like S4, S7 or S28 show a wider rise inside the gene body.

There are several chromatin states that are durably enriched inside the gene body. Marks which rise in both gbM and UM contexts in a similar pattern include H3K4me1, H3K36me3, S3 and S9 (S3 only significant at $p < 0.01$ in gbM but significant in UM at $p < 0.05$; others significant at $p < 0.01$ in both contexts.) . In H3K36me3, the enrichment peaks towards the 5' end, S3 and S9 peak towards the 3' end while H3K4me1 has a long plateau of enrichment. The clustering of S3 and S9 with H3K4me1 is expected as this histone modification is the preferential epigenetic mark of these states. Interestingly, although the occurrence patterns of S3 and S9 look very similar, they regulate for different gene regions: S3 is enriched in coding sequences while S9 is most often present in introns[24]. In both cases, the classification of each state as gbM over non-gbM in *Epimutation Hotspots*[8] seems premature: Although gbM enrichment levels surpass those of non-gbM genes, the difference between them in these two states is much smaller than in other states, e.g. S5, S6 (Test statistics of the Wilcoxon signed-rank test differ by order of magnitude for S3/S9 compared to S5/S6; differences between gbM and UM are significant at $p < 0.01$ in all states). The role of H3K4me1

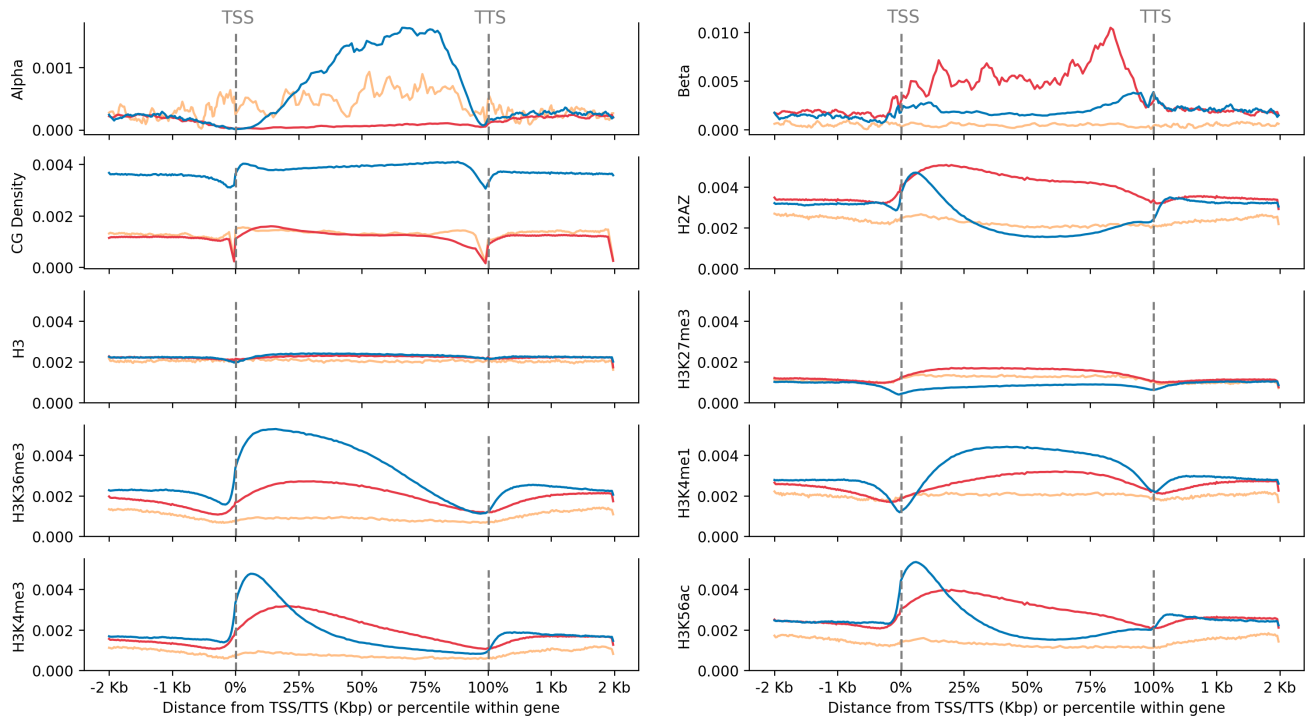

FIG. 5: Metaprofiles of epimutation rates, CG density and Histone Modifications along the length of approximately 5000 gbM genes, 12000 unmethylated (UM) genes and 1000 transposable element (TE)-like genes from Zhang et al.[17]. Genes were separated into 1000 sliding windows, each spanning 1% of the length of the gene, including 2kb upstream and downstream of the gene. The α and β rates each window were estimated using AlphaBeta. Feature levels of different genes were normalized by the total number of genes of this type and their average length.

in enhancers[30–32] cannot be captured by these metaprofiles but appears to be complicated by its increased presence in the body of gbM and UM genes. H3K36me3 on the other hand is found in actively transcribed euchromatin and is associated with exons through nucleosome positioning[33].

S4, S5, S6, S7 and to a lesser extent S28 constitute marks that rise in the body of gbM genes but not in UM genes ($p < 0.01$) in S4 and S6 in UM while significant in gbM, others significant in both gene types). S5 and S6 are symmetrically enriched around the center of the gene while S4 peaks towards the 3' end and S7 and S28 lean towards the 5' end. Although S5 and S6 share their most influential histone modifications - H3K4me1 and H3K36me3[24] - with S3 and S9, and also preferentially occur at the same locations - coding sequences and introns, respectively -, they differ strongly in their enrichment metaprofiles: They don't tilt towards the 3' end and are not present in UM contexts.

To investigate where these differences were coming from, I reanalyzed the emission parameters for each chromatin feature used for chromatin state calling[24] (Fig. 7). To explain the symmetry and lack of enrichment in UM genes in states 5 & 6 compared to 3 & 9, I searched for features that were similar within the pairs but dissimilar between the pairs. This only yielded the two H2A.Z measurements in the dataset (red box). The shape of the metaprofile distributions of states 3 and 9 suggests a negative correlation of state enrichment with H2A.Z level, which would also account for the tilt towards the 3' end. Both of this is reminiscent of the correlation between H2A.Z and α that we first discussed in *Stochasticity*[14].

Similarly, to explain the assortment into coding regions (S3 & S5) or introns (S6 & S9), the pair dissimilarity search resulted in two candidates across 8 data sources, namely the histone variants H3.3 and H3.1 (purple box). I unfortunately did not have access to histone variant data to further investigate this finding.

The last group of chromatin states is present at a higher level in UM than in gbM. Chief among them are S8, S10-S12, which are enriched in UM genes but rarely present in gbM genes (all $p << 0.01$). A similar - if weaker - pattern can be observed in H3K27me3 ($p << 0.01$), which, along with H2A.Z and H3K4me2 is among the preferential epigenetic marks of these states[24]. The antagonal enrichment of these features might be the key to the diverging epimutation rates and methylation levels between UM and gbM genes.

Overall, the observed levels of chromatin feature enrichment are in good concordance[34] with the existing literature. For some states, especially those clustered

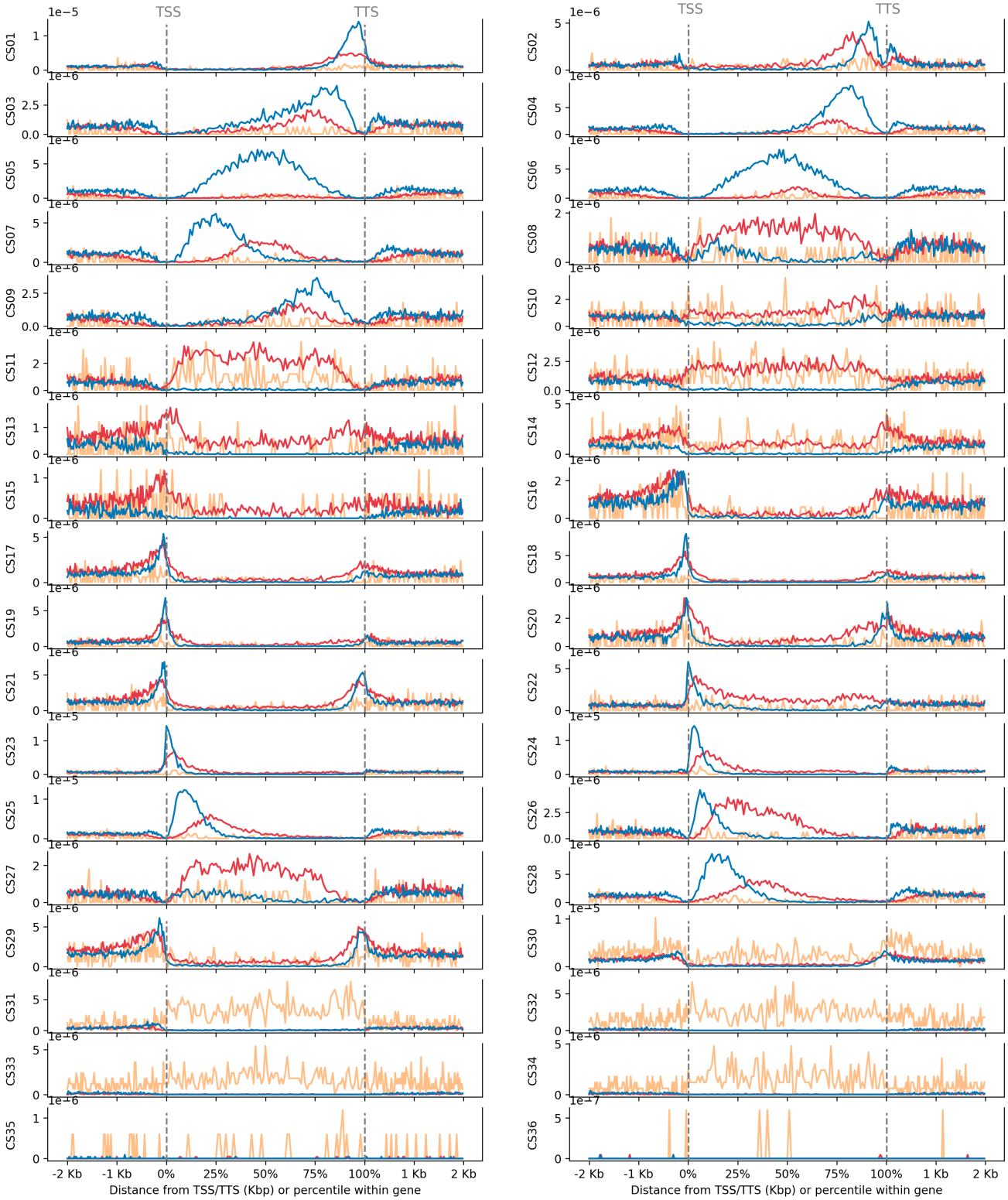

FIG. 6: Metaprofiles of chromatine states along the length of approximately 5000 gbM, 12000 UM and 1000 teM genes from Zhang et al.[17]. Genes were separated into 1000 sliding windows, each spanning 1% of the length of the gene, including 2kb upstream and downstream of the gene. Feature levels of different genes were normalized by the total number of genes of this type and their average length.

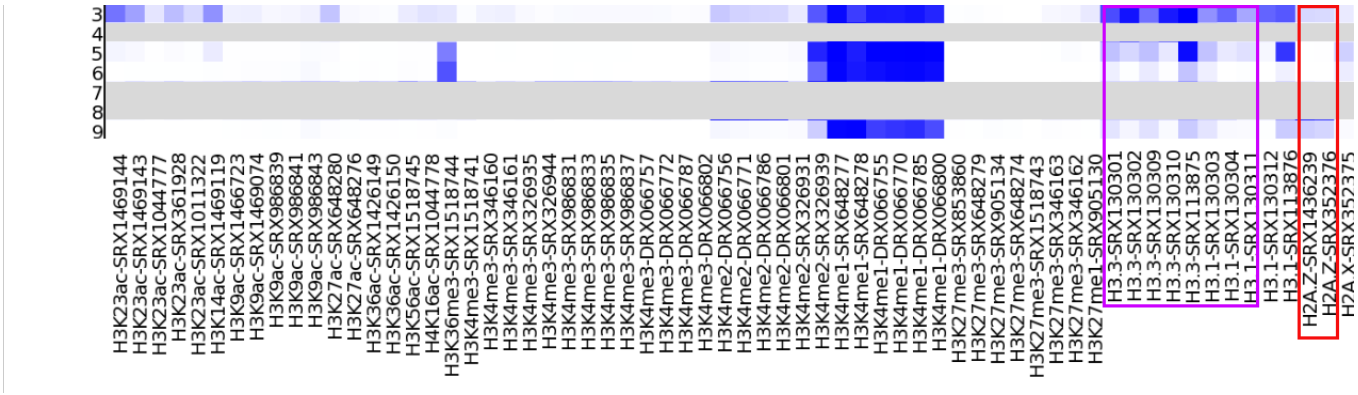

FIG. 7: Emission parameters of histone modifications for chromatin state calling. Adapted from the *Plant Chromatin State Database*[24] https://systemsbiology.cau.edu.cn/chromstates/images/At_emissions.png. Cropped from 36 chromatin states and 216 features to focus on relevant states and chromatin features. To explain the difference between two pairs of chromatine state metaplots, Boxes indicate features that were similar within pairs of chromatine states but different between the pairs.

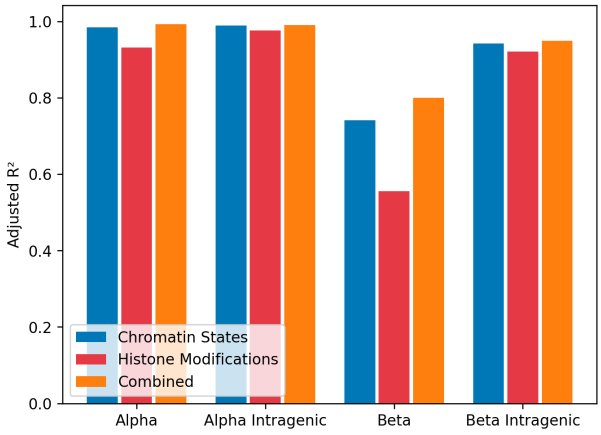

FIG. 8: Adjusted coefficients of determination of different epimutation rate models in gbM genes. The multilinear regression models use either histone modifications (7 features), chromatine states (36 features) or the combination of both (43 features). Results are grouped by the dependent variable.

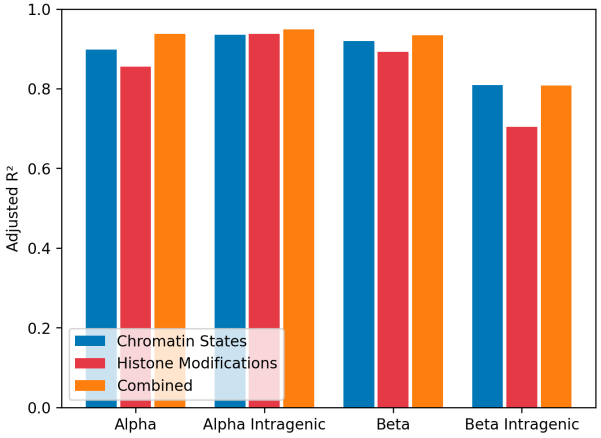

FIG. 9: Adjusted coefficients of determination of different epimutation rate models in UM genes.

around the promotor, current rough descriptions of location and purpose can be refined by using the above metaprofiles. For that, the generality of these findings based on aggregated data must first be validated, though.

D. Modelling intragenic epimutation rates with chromatin information

Initially, I built models that predicted epimutation rates in a constrained context, e.g. just α in gbM genes only, which demonstrate high accuracy (often $R^2 > 0.95$). This allowed me to evaluate the relative prognostic powers of different groups of predictors. Figs (8) and (9) plot the adjusted coefficient of determination R^2 of multilinear regressions for different sets of predictors depending on the chosen epimutation rate, gene type and gene region. The results show that both chromatin states and histone modifications allow for accurate prediction individually with some synergistic effects when used in conjunction. Multilinear regression mostly results in a better fit within the gene body than if up- and downstream regions are included, except for UM genes, where β decreases if constrained to intragenic regions . Prediction fidelity is generally higher for α than for β.

For teM genes, the model is incapable of correctly fitting to the epimutation rates, leading to abysmal adjusted coefficients of determination (Fig. 10) with a maximum of 0.552 for α when using histone modifications as the predictors. For several models, adjusted R^2 values even turn negative as the additional degrees of freedom introduced by the multitude of predictors outweigh the

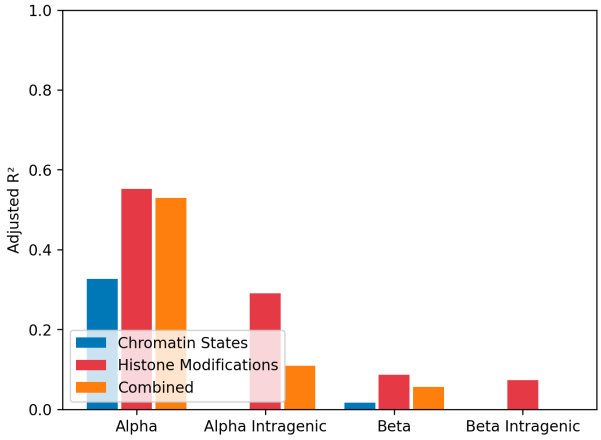

FIG. 10: Adjusted coefficients of determination of different epimutation rate models in teM genes.

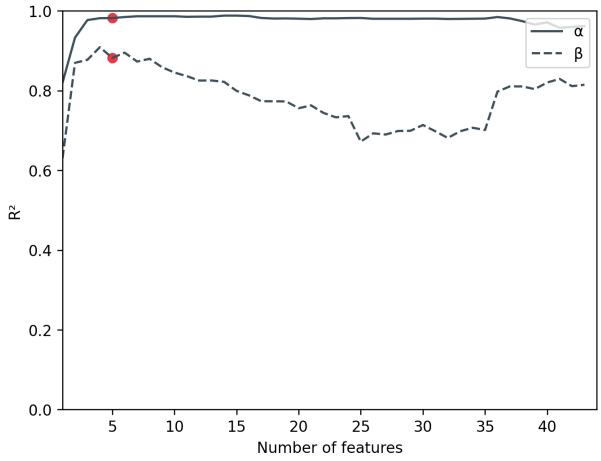

FIG. 11: R^2 values depending on the number of features selected for use in decision tree models of gbM gain and loss rate. The R^2 value shown is the higher of the forward or backward pass of the sequential feature selector for any given number of features.

gains in accuracy[35].

1. *Feature selection*

The main problem with this analysis is the high degree of interdependence between the features, as chromatin states are derived from histone modification enrichment and histone modifications themselves promote or inhibit the modification of surrounding nucleosomes[36]. Multicollinearity problems are also indicated by the small eigenvalues of the regression coefficient matrix, often below 10^{-10}, and large condition numbers (above 10^7).

To address this issue, the features with the most predictive power were selected using sequential feature selection, both backwards and forwards, reducing the total number of features. For these and later regressions, I used a decision tree regressor as these consistently yielded better results than multilinear regression. This might be due to their ability to represent conditional logic, which more closely resembles biological processes than purely linear dependencies[37]. Figures 11 and 12 show the coefficient of determination in relation to the number of features a given model uses. It can be seen that in the beginning the prediction fidelity rises with the number of features but soon plateaus or even decreases as more features are added. The decrease is an indicator of the model overfitting to the training data and therefore performing worse on the validation data (50/50 split). A local optimum is often reached within the first five and at most ten features.

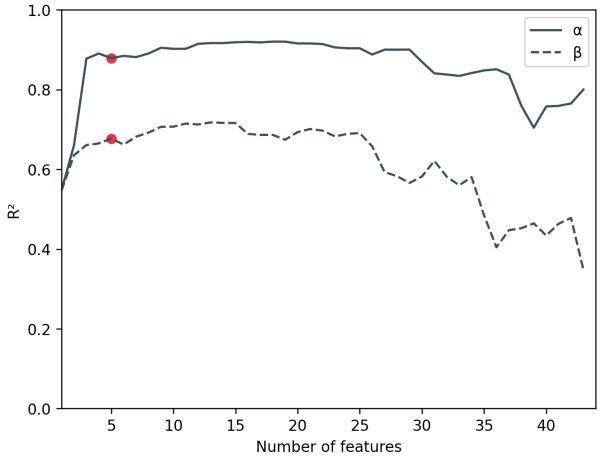

FIG. 12: R^2 values depending on the number of features selected for use in decision tree models of gbM gain and loss rate. $R^2 = \max(\text{forward}, \text{backward})$ using a bidirectional sequential feature selector.

The selected features differ between the models but share some overlap. I generally observed that the particular selection of features was variable between runs, being highly sensitive to model parameters and the test/-train split of the data. This probably occurs due to the similarities between many chromatin states, which makes them more interchangeable. However, the pattern of R^2 quickly peaking before plateauing and finally decreasing was stable between runs.

For gbM genes, the five most relevant features (red dot in figures 11 & 12) for predicting the stochastic gain rate were H2AZ, H3K27me3, H3K56ac, CG Density and S11. For β, it was S06, S08, S18, S27 and S30. In UM genes, S1, S15, S25, S26 and H3K4me1 were selected for α and S02, S05, S18, S31 and H3K56ac were selected for β. In previous iterations which were performed with multilinear regression, often the same features were selected for both α and β. In these cases, their corresponding regres-

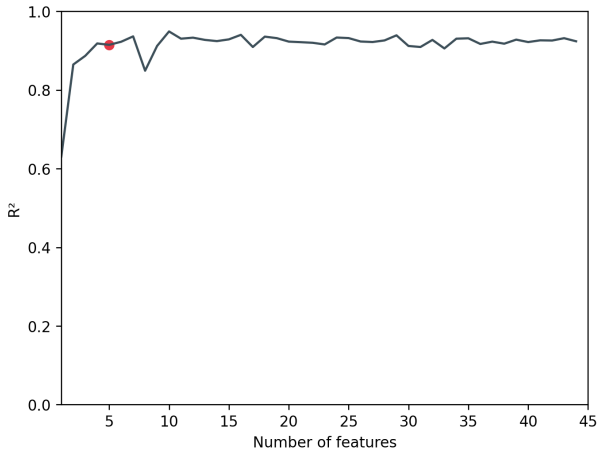

FIG. 13: $R^2 = \max(\text{forward}, \text{backward})$ values depending on the number of features selected by a bidirectional sequential feature selector. The decision tree model combines metaprofiles from gbM, UM and teM genes, including up- and downstream regions to predict both α and β.

sion coefficients had opposite signs (Appendix, Table I), which might be interpreted as gain and loss being modulated in the same biological pathway. Alas, the instability of feature selection prohibits further interpretation in this direction. Stronger results must instead rely on mutant MA lines, connecting feature knockout with affected epimutation rates[14].

2. *Combined model*

Lastly, I built a model encompassing the dependent variables of all contexts and the whole metaprofile, not just intragenic windows. Using a decision tree regressor, the model predicts epimutation gain and loss rate in gbM, UM and teM genes to a R^2 value of 0.932. Feature selection yielded features S6, S13, H2A.Z, H3K27me3 and CG density. Interestingly, although I provided the model with information on the current gene type, the feature selector chose not to use it. Apparently, the current selection of chromatin features is sufficient for a decision tree to derive this information independently. Against our earlier prediction, that "CG density may not even be necessary in a predictive model once H2A.Z enrichment is accounted for"[14], both features are highly useful for predicting epimutation rates across different rates and contexts.

The predictive capability of this combined model is an encouraging result for even more general approaches to forecasting spontaneous epimutation rates.

E. Expanding epimutation rate prediction to the entire genome

In addition to modelling epimutation rates for the windows of metaprofiles along different gene types, my core goal for this thesis was to build a model capable of forecasting rates at single-site resolution. This was mostly unsuccessful.

I built a machine learning model using the transformer architecture[38], which is the foundation of recent progress in deep learning and natural language processing (NLP). At its core lies an attention mechanism by which the model can selectively focus on a subset of the input tokens. This allows it to discard uninteresting information and pick up on complex patterns of tokens, even if they are not located close to each other. Recent developments have addressed shortcomings of the transformer architecture for use in genomics, most importantly the limitations in context length arising from quadratic scaling of attention[39].

Inspired by recent large language model (LLM) developments and *HyenaDNA*[39], I initially attempted to use unsupervised learning. I provided the model with with unlabeled data, including feature vectors for every site (Fig. 2) (encoder) and the entire methylomes of every sample in the MA-lines (decoder). Doing so required rearchitecting the usual transformer to handle two-dimensional (upstream/downstream in the genome and forward/backward in time), instead of one-dimensional (forward/backward in a sentence) sequence data. Ideally, the model would then learn to predict the inheritance of a single cytosine based on the surrounding chromatin features, its methylation state in the previous generation (to prevent it looking ahead in time, a causal mask is applied), and the methylation status of surrounding CG dinucleotides.

Although the model did show some learning, it failed to pick up any meaningful patterns and improvement quickly stalled. I hypothesise that this is due to the inherently stochastic nature of epimutations and their infrequency. An analogy might be a human trying to predict the outcome of a strongly biased coin flip with a 1/1000 probability of landing on "heads". With enough context, you might be able to correctly deduce its probability distribution, but, by rationally always choosing "tails" when predicting the next independent throw, will inadvertently miss the few times that the outcome is different.

Therefore, I subsequently built a supervised model that predicts said probability distribution, namely the epimutation rates α and β. This also had the benefit of reducing the amount of data to be fed into the now encoder-only transformer to just sequence and chromatin feature data. The rates were estimated by applying AlphaBeta to overlapping slices of the genome with the final value for each site being the average of all windows in which it is present. Unfortunately, I was unable to get usable rate estimations using this process

because the results where highly noisy. Repeating the process with increasingly larger window sizes to address a lack of data yielded similarly noisy predictions. Applying previous techniques like multilinear and decision tree regressions to the prepared dataset yielded regression coefficients close to zero and the neural network fared no better.

Previous works have applied similar methods to successfully calculate divergence rates for the whole genome[8], which makes me optimistic that these problems can be overcome with more development time and data. Other works have successfully built models predicting the presence of DNA methylation from chromatin features[40,41], demonstrating that the problem can be tackled in practise.

IV. DISCUSSION

The goal of accurately predicting DNA methylation dynamics from a limited set of DNA and chromatin information is not just useful in itself but will hopefully lead to novel insights into the molecular processes involved. In this thesis, advances in that direction include insights into the precise positioning of promotor-related chromatine states, the cooperative role of H2A.Z and CG density in influencing epimutation rates and further evidence that variations in steady state methylation are primarily driven by changes in α.

For sites grouped by certain genome annotations, I show that high fidelity prognoses are possible with as few as five chromatin features which is advantageous for simplifying experiments. However, the best choice of features is still unclear and more experimental work is required to separate those features that have a causal influence on inheritance mutagenesis from those that are simply correlated.

Further insights into the stochasticity of epimutations in different contexts might come from metaprofiles of histone modifications, epimutation rates and CG density in every chromatin state or an extension of the current work to mutant MA lines and other species. Additionally, if the difficulties in estimating epimutation rates in progressively smaller slices of the genome can be overcome, the outlined genome-wide epimutation neural network might be a strong candidate for a unified epimutation model.

V. ACKNOWLEDGEMENTS

I thank Zhilin for finding and preparing most datasets used in this study as well as Patrick and Ming for being patient beta testers of AlphaBeta-rs, sitting through many bugs and error messages. Your feedback made the end product much better. I also want to thank my parents for last minute reviews (Corinna as well!), providing an excellent writers retreat and not giving up on trying to understand what it is I actually do in these papers. Teresa, it was a pleasure exchanging our respective setbacks during the course of our bachelors theses, thank you for keeping me sane and being a wonderful part of my life. Last but not least, thank you Frank for having given me the opportunity of joining your lab this early on. Your trust has enabled me to make my first steps in actual research which is the most exciting part of my studies!

VI. DATA AVAILABILITY

This work relies on a reanalysis of publicly available data. To compute the epimutation rates, mutation accumulation lines data available as GSE204837[13], GSE64463 (MA1 & MA2)[15] and GSE153055 (MA3)[10] was used. To compute the chromatin information, I used histone modification reads from GSE128434[42], chromatin state assignments from the *Plant chromatin State Database*[24] and gene type annotations[17] , each aligned with the *A. Thaliana* TAIR10.1 reference genome.

The code written to create metaprofiles and run AlphaBeta on them is publicly available at https://github.com/constantingoeldel/alphabeta-rs/ under the GPLv3 license.

Code written for data preparation, modelling and analysis, as well a figure generation is publicly available at https://github.com/constantingoeldel/epimutation_model under the GPLv3 license. It contains interactive notebooks which source intermediate data from this public spreadsheet: https://docs.google.com/spreadsheets/d/1eIdupjcWapVGPNMNTeKbnaPIroOXP6taFhrDMn5z3D4/edit?usp=sharing. Running the analysis.ipynb notebook should reproduce the figures and statistical models used in this work without requiring further input.

Appendix A: Shared features possess opposite regression coefficients

Many of the feature selection runs resulted in predictors that shared features between α and β. In those cases, their coefficients had the opposite sign. An example is given below:

Feature	Alpha	Beta
H2AZ	-1.87e-8	3.34e-8
S32	-5.98e-6	1.04e-5
H3K4Me1	-1.54e-8	3.28e-8
CG Density	2.16e-1	
H3K27Me3	9.83e-8	
H3K36Me3		-1.16e-8
S28		4.09e-5

TABLE I: Regression coefficients of the five most important features for the prediction of alpha and beta, which were found by feature selection. H2AZ, chromatin State 32 and H3K4Me1 were selected for both epimutation rates.

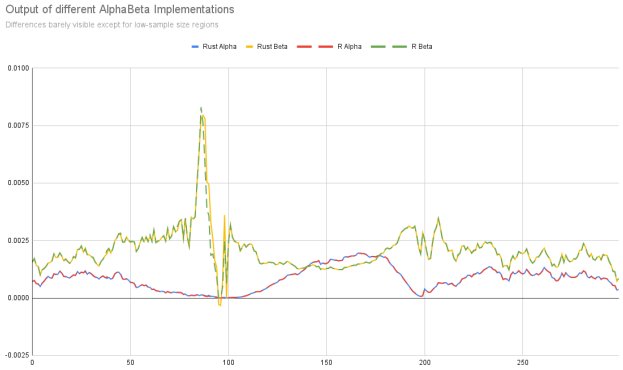

FIG. 14: Assessment of the differences between two AlphaBeta implementations written in R and Rust, respectively, for gbM metaplots[14].

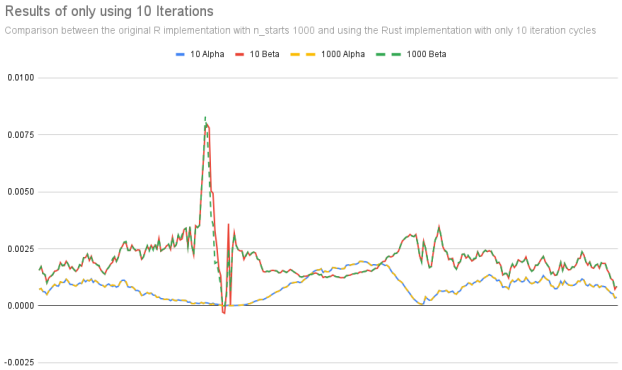

FIG. 15: Assessment of the impact of using a different number of iterations for AlphaBeta implementations written in R and Rust, respectively, for gbM metaplots[14].

Appendix B: Evaluating the alternative AlphaBeta implementation

To assess the correctness of the Rust implementation, previous analyses performed with the original implementation were repeated, allowing for a comparison between their respective outputs (Fig. 14). The resulting epimutation rates had an average difference of 4 / 1 000 000 for α and 4 / 100 000 for β. Differences were higher in regions with low sample sizes.

The original authors were concerned that convergence on epimutation rates using the Nelder Mead Algorithm is not always stable[10], which is addressed by including an additional optimization term controlling for steady state methylation and repeating it for a number of iterations with a default of 1000. To evaluate the need of using this many iterations, I compared the results from the original implementation using 1000 iterations with the Rust implementation using 10 iterations (Fig. 15). Again, the differences were sufficiently small, of the same order of magnitude as the comparison between the implementations at the same number of iterations. A further reduction of iteration yields diminishing returns on performance, as - in accordance with Amdahl's law which states that the overall performance improvement gained by optimizing a single part of the system is limited by the fraction of time that the improved part is actually used - the program execution time is then dominated by the divergence calculation.

[1] S. L. Berger, T. Kouzarides, R. Shiekhattar, and A. Shilatifard, Genes & development **23**, 781 (2009).
[2] J. C. Kiefer, Developmental dynamics: an official publication of the American Association of Anatomists **236**, 1144 (2007).
[3] C. López-Otín, M. A. Blasco, L. Partridge, M. Serrano, and G. Kroemer, Cell (2023).
[4] S. Virani, J. A. Colacino, J. H. Kim, and L. S. Rozek, ILAR Journal **53**, 359 (2012), https://academic.oup.com/ilarjournal/article-pdf/53/3-4/359/1851606/ilar-53-359.pdf.
[5] M. Curradi, A. Izzo, G. Badaracco, and N. Landsberger, Molecular and cellular biology **22**, 3157 (2002).
[6] C. Becker, J. Hagmann, J. Müller, D. Koenig, O. Stegle, K. Borgwardt, and D. Weigel, Nature **480**, 245 (2011).
[7] F. Johannes and R. J. Schmitz, New Phytologist **221**, 1253 (2019).
[8] R. R. Hazarika, M. Serra, Z. Zhang, Y. Zhang, R. J. Schmitz, and F. Johannes, Nature plants **8**, 146 (2022).
[9] M. Suzuki, W. Liao, F. Wos, A. D. Johnston, J. DeGrazia, J. Ishii, T. Bloom, M. C. Zody, S. Germer, and J. M. Greally, Genome research **28**, 1364 (2018).
[10] Y. Shahryary, A. Symeonidi, R. R. Hazarika, J. Denkena, T. Mubeen, B. Hofmeister, T. Van Gurp, M. Colomé-Tatché, K. J. Verhoeven, G. Tuskan, *et al.*, Genome biology **21**, 1 (2020).
[11] N. Yao, Z. Zhang, L. Yu, R. Hazarika, C. Yu, H. Jang, L. M. Smith, J. Ton, L. Liu, J. Stachowicz, *et al.*, bioRxiv

, 2023 (2023).

[12] A. Taudt, D. Roquis, A. Vidalis, R. Wardenaar, F. Johannes, and M. Colomé-Tatché, BMC genomics **19**, 1 (2018).

[13] A. Briffa, E. Hollwey, Z. Shahzad, J. D. Moore, D. B. Lyons, M. Howard, and D. Zilberman, Cell Systems **14**, 953 (2023).

[14] C. Goeldel and F. Johannes, Current Opinion in Plant Biology **75**, 102436 (2023).

[15] A. Van Der Graaf, R. Wardenaar, D. A. Neumann, A. Taudt, R. G. Shaw, R. C. Jansen, R. J. Schmitz, M. Colomé-Tatché, and F. Johannes, Proceedings of the National Academy of Sciences **112**, 6676 (2015).

[16] R. K. Tran, J. G. Henikoff, D. Zilberman, R. F. Ditt, S. E. Jacobsen, and S. Henikoff, Current Biology **15**, 154 (2005).

[17] Y. Zhang, J. M. Wendte, L. Ji, and R. J. Schmitz, Proceedings of the National Academy of Sciences **117**, 4874 (2020).

[18] Or is the DNA not enough? After all, extracting DNA from a cell and putting it into a test tube has not yet let to the synthetic reemergence of life. Is the entire history of life, going back to its origin, required?

[19] S. Feng, S. J. Cokus, X. Zhang, P.-Y. Chen, M. Bostick, M. G. Goll, J. Hetzel, J. Jain, S. H. Strauss, M. E. Halpern, *et al.*, Proceedings of the National Academy of Sciences **107**, 8689 (2010).

[20] A. J. Bewick and R. J. Schmitz, Current Opinion in Plant Biology **36**, 103 (2017), 36 Genome studies and molecular genetics.

[21] D. Zilberman, Genome biology **18**, 1 (2017).

[22] A. J. Bewick, L. Ji, C. E. Niederhuth, E.-M. Willing, B. T. Hofmeister, X. Shi, L. Wang, Z. Lu, N. A. Rohr, B. Hartwig, *et al.*, Proceedings of the National Academy of Sciences **113**, 9111 (2016).

[23] R. Horvath, B. Laenen, S. Takuno, and T. Slotte, Heredity **123**, 81 (2019).

[24] Y. Liu, T. Tian, K. Zhang, Q. You, H. Yan, N. Zhao, X. Yi, W. Xu, and Z. Su, Nucleic Acids Research **46**, D1157 (2018).

[25] C. Wang, C. Liu, D. Roqueiro, D. Grimm, R. Schwab, C. Becker, C. Lanz, and D. Weigel, Genome research **25**, 246 (2015).

[26] F. Roudier, I. Ahmed, C. Bérard, A. Sarazin, T. Mary-Huard, S. Cortijo, D. Bouyer, E. Caillieux, E. Duvernois-Berthet, L. Al-Shikhley, *et al.*, The EMBO journal **30**, 1928 (2011).

[27] R. Ietswaart, Z. Wu, and C. Dean, Trends in Genetics **28**, 445 (2012).

[28] V. Pelechano and L. M. Steinmetz, Nature Reviews Genetics **14**, 880 (2013).

[29] N. D. Trinklein, S. F. Aldred, S. J. Hartman, D. I. Schroeder, R. P. Otillar, and R. M. Myers, Genome research **14**, 62 (2004).

[30] A. Rada-Iglesias, Nature genetics **50**, 4 (2018).

[31] A. Pekowska, T. Benoukraf, J. Zacarias-Cabeza, M. Belhocine, F. Koch, H. Holota, J. Imbert, J.-C. Andrau, P. Ferrier, and S. Spicuglia, The EMBO journal **30**, 4198 (2011).

[32] A. Sharifi-Zarchi, D. Gerovska, K. Adachi, M. Totonchi, H. Pezeshk, R. J. Taft, H. R. Schöler, H. Chitsaz, M. Sadeghi, H. Baharvand, *et al.*, BMC genomics **18**, 1 (2017).

[33] E. J. Wagner and P. B. Carpenter, Nature reviews Molecular cell biology **13**, 115 (2012).

[34] Though I am confused about H3, which as one of the core histones[43] should be present in every nucleosome. Yet, its level is below that of several modifications present on H3, such as H3K36me3. Is H3 in this case just the canonical H3, omitting its variants? Is it included as a validation of the experimental and analytical methods? In any case, it is remarkably stable across gene types and windows. This indicates that nucleosomal spacing has no outsize effect on the distribution of features within these genes.

[35] Adjusted R^2 is given by

$$\bar{R}^2 = 1 - \frac{SS_{\text{res}}/\text{df}_{\text{res}}}{SS_{\text{tot}}/\text{df}_{\text{tot}}}$$

where df_{res} is the degrees of freedom of the estimate of the population variance around the model, and df_{tot} is the degrees of freedom of the estimate of the population variance around the mean. (https://en.wikipedia.org/wiki/Coefficient_of_determination#Adjusted_R2).

[36] M. G. Goll and T. H. Bestor, Genes & development **16**, 1739 (2002).

[37] P. Geurts, A. Irrthum, and L. Wehenkel, Molecular Biosystems **5**, 1593 (2009).

[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, Advances in neural information processing systems **30** (2017).

[39] E. Nguyen, M. Poli, M. Faizi, A. Thomas, C. Birch-Sykes, M. Wornow, A. Patel, C. Rabideau, S. Massaroli, Y. Bengio, *et al.*, arXiv preprint arXiv:2306.15794 (2023).

[40] L. Zhuo, R. Wang, X. Fu, and X. Yao, BMC genomics **24**, 742 (2023).

[41] S. Tsukiyama, M. M. Hasan, H.-W. Deng, and H. Kurata, Briefings in Bioinformatics **23**, bbac053 (2022).

[42] Z. Lu, A. P. Marand, W. A. Ricci, C. L. Ethridge, X. Zhang, and R. J. Schmitz, Nature Plants **5**, 1250 (2019).

[43] E. J. Draizen, A. K. Shaytan, L. Mariño-Ramírez, P. B. Talbert, D. Landsman, and A. R. Panchenko, Database **2016**, baw014 (2016).