Methylencoder: A transformer encoder architecture trained on long-read DNA methylation data achieves significant performance gains by encoding each of the 30M CpGs in the genome and incorporating long-range fragment-level attention

Alexander P. Fields, Erle M. Holgersen, Alex Campbell, Sara Fakhretaha-Aval, Pedro Mendez, Rigel J. Kishton, Siddhartha Bagaria, Samuel S. Gross, Justin K. Valley, Alexander M. Aravanis, Arash Jamshidi
Moonwalk Biosciences, South San Francisco, CA

Abstract

Cytosine methylation is a key epigenetic modification and biomarker that regulates gene expression, controls development and cell fate, and responds dynamically to environmental exposures. Transformer architectures are a promising tool to model the complexity of DNA methylation patterns. Here, we introduce Methylencoder, a 1B-parameter transformer-based encoder model for DNA methylation compatible with long-read methylation data. Because Methylencoder represents reads by their methylation patterns, it is able to process long reads (10+ kbp, each derived from a single cell) within a context size of 128 CpGs (or tokens). This distinguishes Methylencoder both from microarray-based models, which consider only a population-average signal at <2% of CpGs, and from models that represent full nucleotide sequences, which would require an impractically large context size to encode long-read data without truncation. Additionally, many genomic sequence models do not consider epigenetics at all.

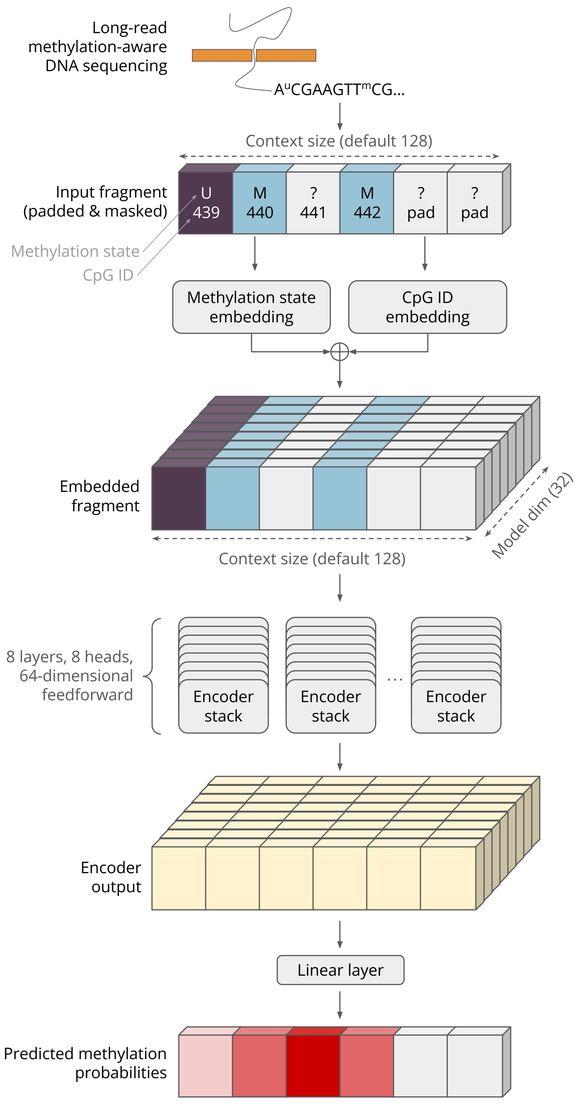
We train and apply Methylencoder on two datasets: an in-house collection of 172 human long-read nanopore sequencing samples from a variety of source tissues (2B total usable fragments, mean 64 CpGs/fragment, 128B total tokens), and a literature atlas of 207 human purified-cell short-read samples (≤300 bp/fragment, mean 2.5 CpGs/fragment). For long-read data, Methylencoder outperforms simpler models such as the multivariate independent Bernoulli model and Bernoulli mixture models at tasks including methylation-state imputation (≥20% relative reduction of cross-entropy loss) and inference of source sample (4.4% relative reduction of cross-entropy loss). For short-read data, Methylencoder outperforms only for the imputation task (12% relative loss reduction).

To quantify the utility of the additional methylation states in long-read data, we measured Methylencoder's performance as a function of context size. Relative to a context size of 128 CpGs, restricting to 16, 8, or 4 CpG context increases cross-entropy loss by 4, 8, or 18%, respectively. Short-read data, for which 68% of fragments contain ≤2 CpGs and 0.3% contain >16 CpGs, are largely excluded from this benefit.

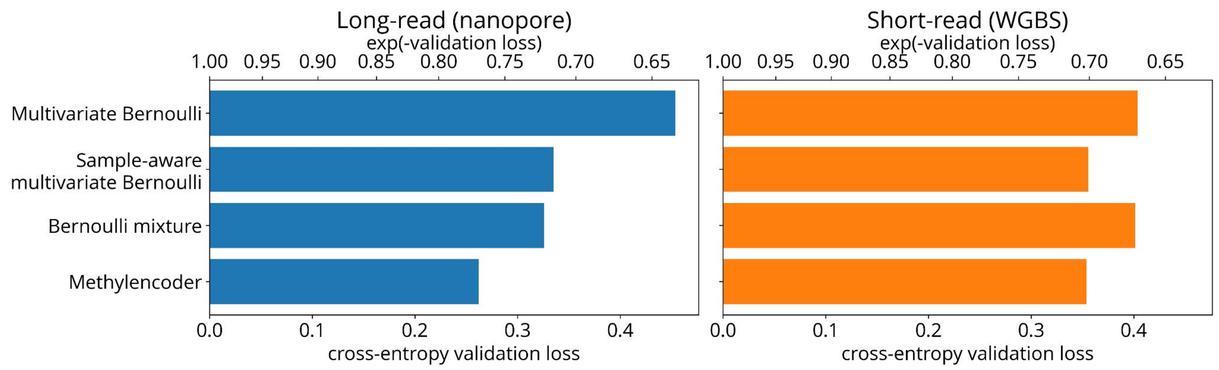
Because Methylencoder only tokenizes the 30M human CpG positions, it can store an embedding for each, a strategy infeasible if attempting to represent all 3B human nucleotide positions. To investigate whether Methylencoder takes advantage of the richness of this embedding space, we measured its performance after forcing genomically proximal CpGs to share an embedding. Grouping CpGs into intervals of 64, 256, or 1024 bp increased cross-entropy loss relatively by 3, 12, or 22%, respectively, indicating that the model learns distinguishing features of even closely separated CpGs.

By leveraging a transformer-based encoder architecture, a CpG-level tokenization that preserves the information contained in long-read data, and a rich embedding that encodes every CpG, Methylencoder achieves state-of-the-art epigenetic modeling with the potential for future applications such as source deconvolution or enhancer mapping.

Methylencoder architecture



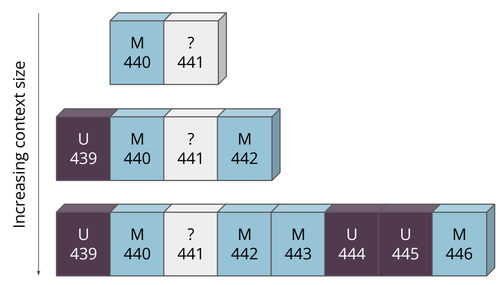
Methylencoder outperforms other models at methylation imputation



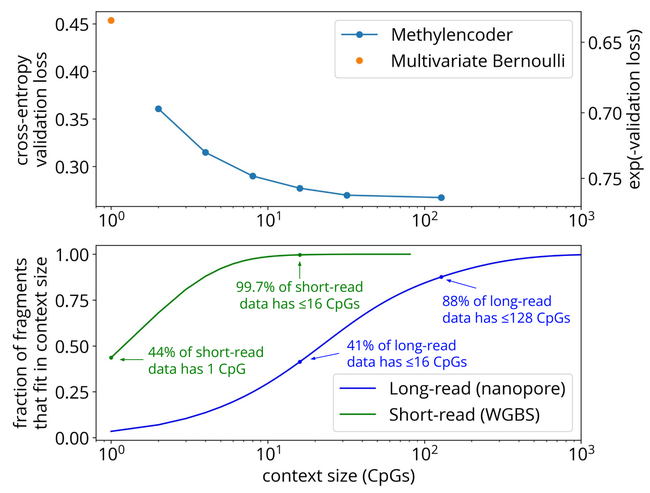
For both long-read (left) and short-read (right) data, Methylencoder delivers lower validation cross-entropy imputation loss than other models. Methylencoder's performance benefit is statistically significant relative to all other models except for the short-read sample-aware multivariate Bernoulli model (Wilcoxon p=0.056, comparing paired performance for each sample within the dataset). Note that the sample-aware multivariate Bernoulli model has the benefit of knowledge of which sample each fragment derived from; this information is not provided to Methylencoder or the other models.

Model	Description	# parameters	Training method
Multivariate Bernoulli	For each CpG, predict average methylation seen in training set	# CpGs	Single-step calculation of average methylation
Sample-aware multivariate Bernoulli	For each CpG, predict average methylation seen in the sample of origin's training set	# samples × # CpGs	Single-step calculation of average methylation
Bernoulli mixture	32-component mixture model, each component with its own mixture fraction and methylation probabilities	32 × (# CpGs + 1)	Stochastic minibatch optimization with Adam Early stopping when validation loss increases
Methylencoder	Described in "Methylencoder architecture"	32 × # CpGs + O(32^2)	Stochastic minibatch optimization with Adam Terminate after fixed duration

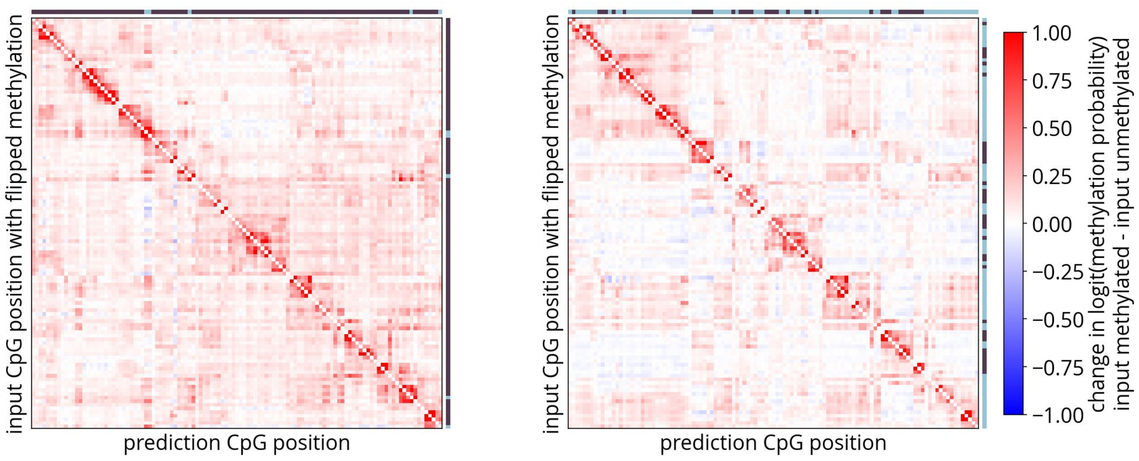
Methylation imputation benefits from larger context than short-read data can provide



As context size increases, more methylation states are provided to the model near each masked position, giving the model more information with which to make predictions. By default, the long-read Methylencoder model provides a context size of 128 CpGs, enough to encode 88% of long-read fragments without truncation. Reduction of context size diminishes performance, particularly for values ≤16 CpGs. The low number of CpGs per fragment in short-read data precludes the model from learning long-range methylation patterns, regardless of context size.

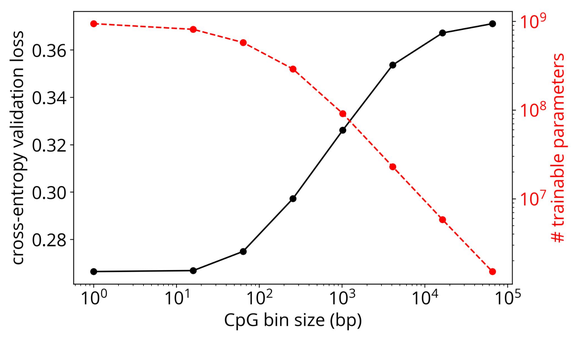


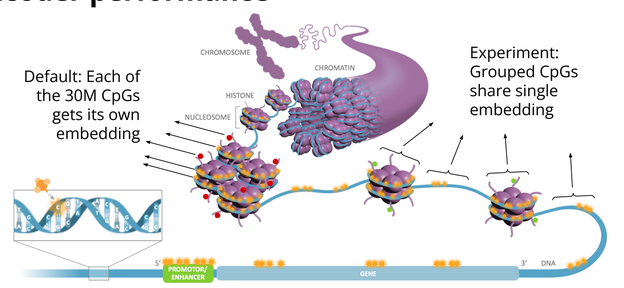
Within-fragment attention allows long-range propagation of predicted methylation



Methylencoder's predicted methylation probabilities at each position along a fragment respond variably to changes in methylation elsewhere in the fragment. In each plot, the color at position (x, y) shows the difference in predicted logit-transformed methylation likelihood at CpG x when the input methylation at CpG y is methylated vs unmethylated, for an input fragment overlapping the promoter of the SKI gene (chr1:2229609-2233059). The original methylation state at each CpG is plotted at the top and right of each plot, with light blue indicating a methylated position. For both fragments, each perturbation tends to have a stronger, localized effect, as well as a weaker, long-range effect.

Richer CpG embeddings improve Methylencoder performance

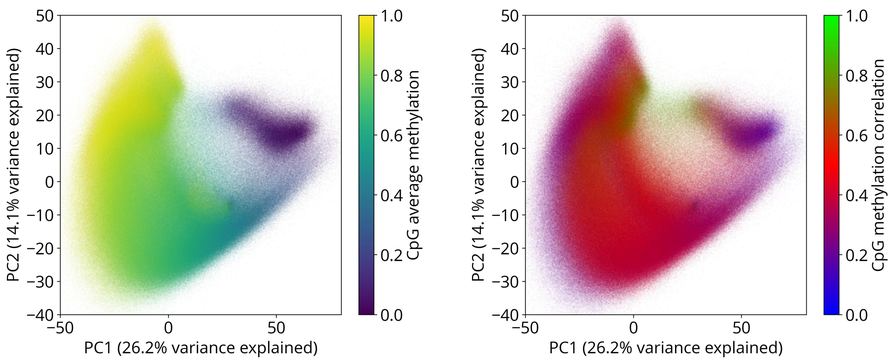




By default, Methylencoder assigns every genomic CpGs its own embedding. Grouping nearby CpGs to share embedding values causes imputation performance to degrade, especially for groupings ≥64 bp.

Learned CpG embeddings reflect biology of methylation

Methylencoder learns a 32-dimensional embedding vector for each of the 30M CpGs in the genome. In these plots, these CpG embeddings are projected along their lowest two principal components and colored according to average methylation or within-fragment Pearson correlation of methylation with adjacent CpGs. A distinct cluster of low-methylation CpGs (many of which are part of CpG islands) is readily visible. More subtly, a set of CpGs whose methylation is high correlated with their neighbors is visible near the top center of the plot.



Methylencoder improves inference of sample origin for long-read data

Methylencoder can be adapted to infer each fragment's sample of origin by prepending each input fragment with a "class" token, the encoder output of which is sent to a fully connected multilayer perceptron (MLP) model (ReLU activation, hidden dimensions 512-512). For comparison, an ordinary MLP model can be trained for the same task by directly connecting the one-hot encoded methylation state at each CpG to a fully connected network (ReLU activation, hidden dimensions 32-512-512, where the value of 32 matches Methylencoder's model dimension). Alternatively, the sample-aware multivariate Bernoulli model can be adapted to sample inference by calculating the posterior likelihood of a fragment being generated by each sample's multivariate Bernoulli model (mathematically equivalent to a Naive Bayes classifier).