CpG Transformer for imputation of single-cell methylomes

Gaetan De Waele
Department of Data Analysis
and Mathematical Modelling
Ghent University
gaetan.dewaele@ugent.be

Jim Clauwaert
Department of Data Analysis and
Mathematical Modelling
Ghent University
jim.clauwaert@ugent.be

Gerben Menschaert
Department of Data Analysis and
Mathematical Modelling
Ghent University
gerben.menschaert@ugent.be

Willem Waegeman
Department of Data Analysis
and Mathematical Modelling
Ghent University
willem.waegeman@ugent.be

Abstract

Motivation: The adoption of current single-cell DNA methylation sequencing protocols is hindered by incomplete coverage, outlining the need for effective imputation techniques. The task of imputing single-cell (methylation) data requires models to build an understanding of underlying biological processes. Current approaches compress intercellular methylation dependencies in some way and, hence, do not provide a general-purpose way of learning interactions between neighboring CpG sites both within- and between cells.

Results: We adapt the transformer neural network architecture to operate on methylation matrices through the introduction of a novel 2D sliding window self-attention. The obtained CpG Transformer displays state-of-the-art performances on a wide range of scBS-seq and scRRBS-seq datasets. Furthermore, we demonstrate the interpretability of CpG Transformer and illustrate its rapid transfer learning properties, allowing practitioners to train models on new datasets with a limited computational and time budget.

Availability and Implementation: CpG Transformer is freely available at https://github.com/gdewael/cpg-transformer

Keywords Deep learning · Transformers · DNA methylation · Imputation

1 Introduction

DNA methylation is the addition of a methyl group to the DNA. The best-known type is CpG methylation, where the methyl group is added to the C-5 position of CG dinucleotides. Its association with a broad range of biological processes, such as gene expression regulation, is well-established [1]. CpG methylation is also known as a driving factor in developmental biology and carcinogenesis, motivating the need to study this phenomenon on the cellular level [2].

The last decade, several protocols that measure DNA methylation at single-cell resolution have been developed. These methods make use of bisulfite conversion of DNA followed by sequencing [3], both on genome-wide scale (scBS-seq) [4] and using reduced-representation protocols (scRRBS-seq) [5]. These methods have uncovered the heterogeneity and dynamics of epigenetic patterns between cells and have made it possible to describe epigenomic networks on an unprecedented scale and resolution [6].

Due to the smaller amount of genetic material available per cell, profiling single cells comes with certain challenges not encountered in bulk sequencing experiments. In practice, the genome-wide coverage of CpG sites per cell is low,

ranging from 1% for high-throughput studies [7] to 20% for low-throughput ones [4]. Furthermore, profiled sites are covered by a smaller number of reads, resulting in noisy measurements of DNA methylation. Effective imputation and denoising techniques are therefore crucial in unlocking the full potential of single-cell methylome analyses.

Prediction of methylation states in tissue samples is a well-established problem in bioinformatics, often tackled by leveraging dependencies between CpG sites. For example, variational autoencoders have been successfully applied for dimensionality reduction of methylation data [8]. Other methods focus on imputation of single CpG sites in tissue samples using, among others, linear regression [9], Random Forests [10], autoencoders [11], gradient boosting [12] or mixture models [13]. In addition to using intrasample dependencies between neighboring CpG sites, some of these methods adopt the idea of leveraging information from multiple (tissue) samples for prediction [12, 13].

Most recent work on single-cell DNA methylation imputation has built upon this idea of leveraging both intra- and intercellular correlations between methylation states. Melissa [14] first defines specific regions of interest in the genome (such as a specific promoter region), then performs generalized linear model regression on CpG sites in that region. The model leverages information from other cells through a shared prior distribution determined by a Bayesian mixture model, effectively clustering cells. DeepCpG [15] proposes a recurrent neural network (RNN) to process differences in local CpG profiles across cells. For every cell, the local CpG profile consists of a vector containing the methylation states and distances of the 25 nearest observed CpG sites up- and downstream from the target site. Along with this RNN, a convolutional neural network (CNN) processes relevant information in the DNA sequence surrounding the target site. The two streams of information are combined near the end of the network. Finally, the output head returns predictions for every cell at a single CpG site. Using similar design principles, LightCpG uses gradient boosting to obtain faster training times at the cost of a lower performance [16]. CaMelia [17], also relying on gradient boosting models, restricts its imputation to CpG sites that are also recorded in at least one other cell. It additionally discards CpG sites whose local methylation profiles are too dissimilar of the profiles in all other cells. Uniquely, CaMelia introduces the notion of using bulk tissue samples to improve performance compared to DeepCpG and trains a separate model for every cell. It remains unclear, however, whether these performance gains can be attributed to the employed methods or to the aforementioned sample selection. Additionally, it is important to note that all of the aforementioned methods compress or summarize intercellular methylation dependencies in some way, whether through a prior, the outputs at the ends of an RNN, or a single feature. No framework has been proposed yet for learning dependencies between methylation sites in a general-purpose manner.

In this work, inspiration is drawn from recent developments in self-supervised learning of natural language. In particular, the language model BERT is trained by randomly replacing words in a sentence by a unique `[MASK]` token and attempting to predict the masked word given the newly-formed sentence (called masked language modeling) [18]. In essence, this objective trains a model to fill in the gaps in a sentence. The similarity with imputation, where gaps in a matrix need to be filled in, is compelling but unexplored. In language modeling, transformer neural networks are used because of their capability of learning interactions between all input words, akin to the flow of information in a complete digraph [19].

Biological systems can be elegantly represented by graphs [20]. For example, the interactions of genes form distinct pathways in a regulatory network. Consequently, models should ideally reason over graphs or mimic graph structure. Most of the current deep learning practices in bioinformatics do not reflect this reality. For example, fully-connected layers learn a set of fixed weights for all inputs and are hence unable to reason over how correlations between inputs differ when their contents change. Transformers mimic graph structure using a self-attention mechanism to explicitly reason over how every input is influenced by the others [21]. Because of this, transformers scale quadratically in computational- and memory cost with the number of inputs. They have previously been shown to outperform other neural architectures in DNA sequence annotation tasks [22] and protein representation learning [23]. Recently, the use of transformers in biology has gone beyond 1D sequences. For example, MSA Transformer [24] adapts axial attention [25] to MSAs for unsupervised protein structure learning. AlphaFold2 [26] also speculatively uses two-dimensional attention to process MSAs. By processing 2D inputs, full self-attention learns $\mathcal{O}(n^2m^2)$ pairwise interactions for a $\mathbb{R}^{n\times m}$ matrix, making vanilla transformers impossible to apply on high-dimensional methylation data.

We introduce CpG Transformer, an adaptation of the transformer neural network architecture to operate on partially-observed methylation matrices using a novel 2D sliding window self-attention, thereby obtaining state-of-the-art imputation performances on a wide range of datasets. The inputs to CpG Transformer consist of the CpG matrix along with their respective positions on the genome and the DNA sequences surrounding them. Cell identity is communicated to the model through learned cell embeddings. The model recombines signals from CpG sites in a graph-like manner. Because of this, the architecture excels at transfer learning, a prospect of great interest to practitioners with different datasets, downstream objectives, and limited computational resources. In addition, ablation studies and model interpretation demonstrate the contributing factors to single-cell DNA methylation.

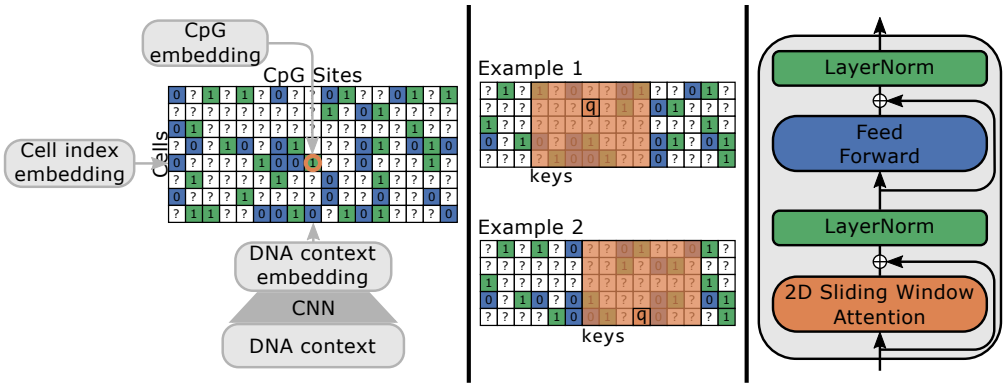

Figure 1: **Left:** Inputs to CpG Transformer. Cell, DNA, and CpG embeddings are applied row-, column-, and element-wise, respectively. **Middle:** 2D sliding window attention, with full self-attention over rows and sliding windows over columns. Every query attends to all keys within a window as shown in the figure. **Right:** A single CpG Transformer layer.

2 Methods

Here, CpG Transformer is described for the imputation of DNA methylation data. Our architectural contributions are twofold. Firstly, CpG Transformer draws inspiration from collaborative filtering approaches to formulate its inputs to the transformer layers [27]. Transformer layers model the interactions between matrix entries. In this sense, CpG Transformer can be regarded as *contextualized collaborative filtering*. Second, we extend sliding window self-attention [28] to 2D inputs, where full attention is applied in the row dimension and sliding window attention is applied over columns.

Model Inputs The input to CpG Transformer is a three-dimensional tensor $\mathbf{H} \in \mathbb{R}^{n \times m \times d_{\text{model}}}$, where $\mathbf{H}_{i,j} \in \mathbb{R}^{d_{\text{model}}}$ represents the input representation at cell i (rows) and methylation site j (column) of the methylation matrix. Every representation $\mathbf{H}_{i,j}$ is the result of linear combination of a concatenation of three embeddings: $\mathbf{H}_{i,j} = \boldsymbol{W} \cdot [\boldsymbol{h}_{i,j}^{\text{CpG}}, \boldsymbol{h}_i^{\text{cell}}, \boldsymbol{h}_j^{\text{DNA}}]$ (Figure 1). All three embeddings consist of 32 hidden dimensions, and are combined to $d_{\text{model}} = 64$ dimensions by $\boldsymbol{W}$. The CpG embedding $\boldsymbol{h}_{i,j}^{\text{CpG}}$ is obtained by embedding the methylation state (unknown `?`, unmethylated `0` or methylated `1`) of CpG site j in cell i. Similarly, row-wise cell embeddings $\boldsymbol{h}_i^{cell}$ encode a hidden representation for cell indices. Finally, DNA sequence information is included in the model by taking 1001 nucleotide windows centered around the methylation sites and processing them with a CNN to obtain column-wise DNA embeddings $\boldsymbol{h}_j^{\text{DNA}}$. In all experiments, the CNN architecture is adapted from DeepCpG, consisting of 2 convolutional layers, each followed by a max-pooling layer [15]. The exact parameters of the CNN backbone are elaborated in Supplementary Section 1.

CpG Transformer Transformer layers employ self-attention to explicitly reason over how every input is influenced by the others [21]. All n entries of an input $\boldsymbol{X}$ are once encoded as a query and once as a key via learned linear layers. For this model setup, the input to the transformer layers is $\mathbf{H}$. Taking the inner product of the queries $\boldsymbol{Q}$ with the keys $\boldsymbol{K}$ results in an $n \times n$ matrix, whose values can be loosely interpreted as the importance of input j for input i, at row i and column j. These values are normalized and multiplied by a value matrix $\boldsymbol{V}$ (obtained via linear combination of the input $\boldsymbol{X}$ with learned weights) to produce outputs for every input entry in a matrix $\boldsymbol{Z}$. This process can be performed multiple times in parallel using separate weight matrices, constituting different attention heads. Corresponding outputs $\boldsymbol{Z}$ for every head can then be concatenated and linearly combined to an appropriate hidden dimension size. The scaled dot-product self-attention mechanism first described by Vaswani et al. [21] is given by the following equations, where d_k denotes the hidden dimensionality of the queries and keys:

$$\begin{aligned} \boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} &= \boldsymbol{X}\boldsymbol{W}_q^\top, \boldsymbol{X}\boldsymbol{W}_k^\top, \boldsymbol{X}\boldsymbol{W}_v^\top \\ \boldsymbol{Z} &= \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d_k}}\right)\boldsymbol{V} \end{aligned} \tag{1}$$

Intuitively, this mechanism simply learns how inputs should be recombined in order to propagate to an output at every position. As such, no structure in the input is assumed and fixed-length inputs are not required, as identical model

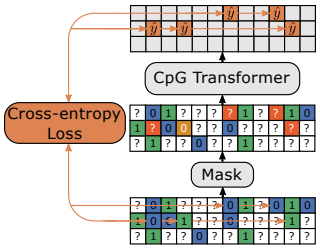

Figure 2: **Masked language modeling.** Positive and negative sites are indicated in green and blue, respectively. Sites to train on (orange) are either masked (80%) or randomized (20%). The model is optimized to infer the original methylation state given the corrupted input using the cross-entropy loss.

weights are used for every position. For an input methylation matrix with n cells (rows) and m methylation sites (columns), an $n \cdot m \times n \cdot m$ attention matrix is obtained. Because m can easily exceed millions, it is impossible to apply vanilla transformers to methylation data. To reduce the dimensionality and computational complexity of this operation, known autocorrelation between neighboring methylation sites can be leveraged. It is known that CpG sites in close proximity of each other on the genome are often correlated [29]. Hence, instead of modeling all possible pairwise interactions in the methylation matrix, we limit self-attention to interactions between neighboring CpG sites. More formally, for any input query, the keys are defined as all the entries in a window around that query, as in Figure 1. This way, the computational complexity of self-attention is reduced from $\mathcal{O}(n^2m^2)$ to $\mathcal{O}(n^2mw)$, with a window size w. Sliding windows of the methylation matrix are efficiently computed through the unfold operation [30]. Additionally, the sliding windows allow to slice the $n \times m$ methylation matrix in multiple $n \times b$ bins with minimal context fragmentation and, therefore, minimal performance losses. In order to communicate relative distances of CpG sites to the model, every self-attention operation is supplied with relative sinusoidal positional encodings [31]. A more-detailed figure illustrating 2D sliding window attention is given in Supplementary Figure S1.

CpG Transformer employs a stack of four identical layers (Figure 1). The layer structure is equal to the one defined by Vaswani et al. [21]. Each layer has two sub-layers. The first consists of the previously-described 2D sliding window attention mechanism, using a window size $w = 21$ and 8 heads with 8 hidden dimensions each. The window size is selected considering a trade-off between computational complexity and inclusion of biological information. A larger window size means that CpG Transformer recombines information from more neighboring sites at the cost of computational- and memory complexity. The input and output dimensionality of the attention layer is $d_{\text{model}} = 64$. The second sub-layer employs a position-wise fully-connected feed-forward network consisting of two linear combinations with a ReLU activation in between: $\max(0, \boldsymbol{X}\boldsymbol{W}_1 + \boldsymbol{b}_1)\boldsymbol{W}_2 + \boldsymbol{b}_2$. The dimensionality of input and output is $d_{\text{model}} = 64$ and the inner-layer has 256 hidden dimensions. A residual connection [32] followed by layer normalization [33] is employed around both sub-layers. The outputs of the last transformer layer are reduced to one hidden dimension by an output head and subjected to a sigmoid operation to obtain final predictions $\hat{\boldsymbol{Y}} \in \mathbb{R}^{n \times m}$ for all inputs.

Training Objective We adapt the masked language modeling (MLM) objective for DNA methylation imputation [18]. MLM is a type of denoising autoencoding in which the loss function acts only on the subset of inputs that are perturbed. For CpG Transformer, the inputs are corrupted by masking random observed sites to the `?` token. In addition, 20% of the tokens that would be masked are instead randomized to a random state (`0` or `1`), sampled proportionally to the distribution of methylation states in the input. Finally, the cross-entropy loss optimizes the model to return the original methylation states given the corrupted input. An overview of the procedure is given in Figure 2.

Datasets Five publicly available datasets originating from both scBS-seq [4] and scRRBS-seq [5] experiments are obtained from the Gene Expression Omnibus.

The first dataset (GSE56879) consists of 20 mouse embryonic stem cells cultured in Serum. The second dataset is obtained from the same study and is made up of 12 cells of the same type cultured in 2i medium [4]. Both datasets were profiled using scBS-seq. A third dataset (GSE65364) comprises 25 human hepatocellular carcinoma cells profiled using scRRBS-seq [34]. scRRBS-seq profiles of 30 human monoclonal B-cell lymphocytes form a fourth dataset (GSE125499; sc05) [35]. The final dataset (GSE87197) consists of 122 hematopoietic stem cells and progenitor cells profiled using scBS-seq [7]. This dataset includes 18 hematopoietic stem cells, 18 multipotent progenitors, 19 common myeloid progenitors, 24 multi-lymphoid progenitors, 22 granulocyte macrophage progenitors and 21 common lymphoid progenitors. In the remainder of this paper, these datasets are referred to as Ser, 2i, HCC, MBL, and Hemato, respectively. Corresponding reference genomes are as follows: Ser and 2i use genome build NCBIM37. GRCh38 is used by Hemato, and GRCh37 serves as reference genome for HCC and MBL.

Table 1: **Performance comparison of CpG Transformer with other methods.** Sparsity is defined as the percentage of entries in the methylation matrix that are unobserved. Best performers are indicated in bold. The reported metrics are computed for all cells together.

			ROC AUC			PR AUC		
Dataset	**# Cells**	**Sparsity (%)**	**DeepCpG**	**CaMelia**	**CpG Transformer**	**DeepCpG**	**CaMelia**	**CpG Transformer**
Ser	20	77.8	90.21	90.22	**91.49**	92.77	92.86	**93.81**
2i	12	77.9	84.80	83.02	**85.57**	71.69	68.87	**73.16**
HCC	25	88.5	96.89	97.42	**97.94**	92.58	94.10	**95.11**
MBL	30	90.8	88.22	89.17	**92.05**	87.61	87.6	**91.32**
Hemato	122	98.4	88.85	89.16	**90.37**	95.60	95.84	**96.32**

Table 2: **Ablation study on Ser dataset.** The original model is compared to four models for which one type of input is removed.

Model	**ROC AUC**
Original	91.49
Without cell emb.	84.45
Without CpG emb.	71.15
Without DNA emb.	91.07
Without positional enc.	90.46

For all datasets, binary methylation states are obtained by assigning a positive (methylated) label when $\frac{\#(\text{reads}_{\text{positive}})}{\#(\text{reads}_{\text{total}})} \geq 0.5$. We use holdout validation to test the performance of the models. For all datasets and experiments, chromosome 5 and 10 constitute the validation and test set, respectively. All other chromosomes are used in training. More instructions on how to obtain and preprocess the datasets, as well as their corresponding reference genomes, are available on the GitHub page of CpG Transformer.

Models and Training CpG Transformer is compared to two competing methods: DeepCpG [15] and CaMelia [17]. A comparison with Melissa is not considered since the method is described as complementary to whole-genome imputation methods [14]. To ensure a fair comparison, all models are trained using the same data preprocessing and splits. Due to this, performances are expected to deviate slightly from those reported in their respective manuscripts. Considering reproducibility concerns, a full list of differences in our implementations of DeepCpG and CaMelia is given in Supplementary Section 1.

All CpG Transformer models are trained on 2 V100 GPUs using Adam as optimizer [36]. A learning rate of $5 \cdot 10^{-4}$ with linear warmup over the first 1000 steps is used. The learning rate is multiplicatively decayed by a factor 0.9 after every epoch. Models are trained for a maximum of 100 epochs, with early stopping after no validation loss decrease has been observed for 10 epochs. The model arising from the epoch with the best validation loss is kept as final model. A dropout rate of 0.20 on elements of the attention matrix is employed during training. The methylation matrix is batched by binning into segments of 1024 columns (methylation sites). One such a segment makes up a batch. For every batch, the number of sites that are masked or randomized equals 10% the number of columns in the bin for the Hemato dataset and 25% for all other datasets. For the Hemato dataset, we additionally randomly subsample 32 rows (cells) every training batch to reduce complexity and increase training speed. Finally, because random masking negatively biases evaluation, test performance is measured by masking every methylation site in the dataset separately in smaller batches.

3 Results

3.1 Imputation performance

To benchmark CpG Transformer, we evaluate against one competing deep learning method, DeepCpG [15], and one traditional machine learning method, CaMelia [17]. The resulting imputation performances in terms of area under the receiver operating characteristic curve (ROC AUC) and area under the precision-recall curve (PR AUC) for all datasets are shown in Table 1. CpG Transformer outperforms existing models on all datasets. Cell-specific performance evaluation (Supplementary Figure S2). shows that CpG Transformer is, out of all cells, only outperformed by competing methods for one 2i cell and five Hemato cells.

A small ablation study (Table 2) on the Ser dataset shows the importance of the different inputs to the model. The original model is compared to four models, each trained and evaluated in a scenario where one specific input is left out: $\boldsymbol{h}_{i,j}^{\text{CpG}}$, $\boldsymbol{h}_{i}^{\text{cell}}$, $\boldsymbol{h}_{j}^{\text{DNA}}$, or the positional encodings. Without the CpG embedding, the model can only rely on cell identity and the DNA context from their own site and neighboring CpGs. The model without this embedding displays the lowest performance, illustrating the key importance of dependencies between methylation states for their prediction. Without cell embeddings, cell identity is lost and the prediction for every site is the same for all cells. As the second-most important input for the Ser dataset, this embedding highlights CpG Transformer's capability to exploit cell heterogeneity. Without positional encodings, the model has no way of knowing how far away two CpG sites are from each other. Since column-wise correlation between CpG sites decreases with distance [29], their role is to inform the effect of genomic distance on the degree of correlation in a flexible way. In practice, a minimal but noticeable effect of this encoding on

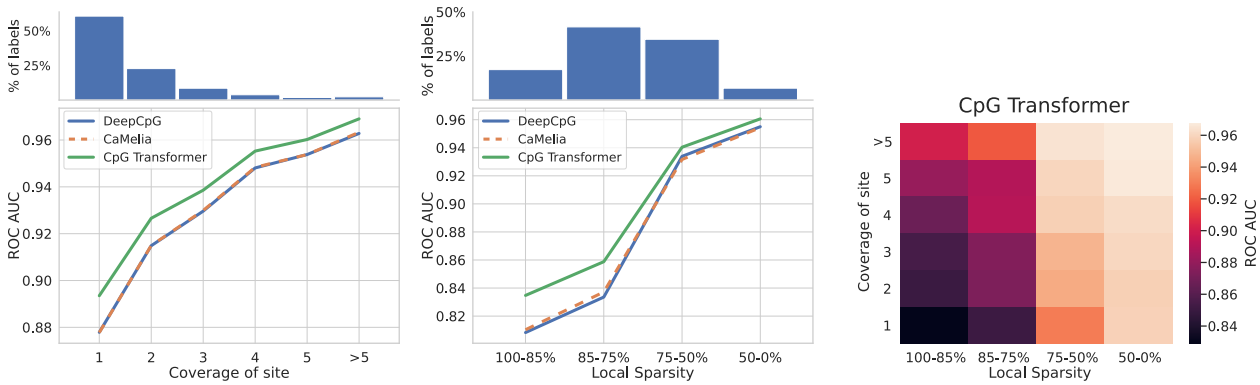

Figure 3: **Dependency of performance on sequencing depth for the Ser dataset. Left:** ROC AUC in function of coverage (# reads) of the label. On top of the plot the percentage of labels belonging to each bin is shown. **Middle:** ROC AUC in function of local sparsity (defined as the percentage of unobserved entries in the local window used for prediction). **Right:** ROC AUC in function of both factors for CpG Transformer. The biggest gradient in performance is observed for the local sparsity direction.

the Ser dataset is observed. Finally, the DNA embeddings, informing the model of DNA context surrounding CpGs, is indicated as the least important input for imputation of the Ser dataset.

The performance of all models is heavily dataset-dependent, indicating their varying quality. Since single-cell sequencing experiments suffer from low sequencing depth, we hypothesize that performance is negatively influenced by limited coverage both at the CpG site in question (noisy labels) and in its neighborhood (in terms of number of unobserved entries, termed local sparsity). To test this, the performance of the Ser dataset in function of these factors is plotted (Figure 3). Similar plots for the other datasets are shown in Supplementary Figure S3. It is observed that CpG sites covered by a smaller number of reads have a less-confident label, resulting in negatively-biased performance at evaluation. In addition, CpG sites with a higher local sparsity are harder to predict, presumably due to providing a noisier estimate of local methylation profiles. By making a heatmap of performance in function of both these factors, it is observed that local sparsity is most causal of lower predictive performance.

3.2 Model interpretation

Because CpG Transformer recombines information from CpG sites in a general way, it lends itself well to model interpretation methods. Here, we study the problem of attributing the model prediction to its input features via the use of Integrated Gradients [37]. This method computes the gradients of the prediction with respect to the input features to measure how every input contributes to prediction. Contibutions are obtained by decomposing the difference in prediction of the input sample with an all-zero baseline. For CpG Transformer, contribution scores are obtained for all inputs to the first transformer layer. Since four transformer layers with a window size of 21 are employed, the total receptive field for any prediction constitutes the 81 surrounding CpG sites for all n cells ($n \times 81$). Because Integrated Gradients returns contribution scores for all hidden dimensions, they are summed to obtain a single score for every input matrix entry. An example contribution for the Ser dataset is shown in Figure 4A.

The contributions of the individual matrix entries can be decomposed into those of their constituent embeddings $[\boldsymbol{h}_{i,j}^{\text{CpG}}, \boldsymbol{h}_i^{\text{cell}}, \boldsymbol{h}_j^{\text{DNA}}]$ by backpropagating Integrated Gradients one layer further. In doing so, contribution matrices similar to the one shown in Figure 4A are obtained for all three embeddings (Supplementary Figure S4). Performing this for 1% of the samples in the test set, it is possible to investigate how embeddings contribute to prediction depending on the sample. This way, the total contribution of the embeddings in function of the prediction error, local sparsity and cells are obtained for the Ser dataset 4B-D. The same plots for the other datasets are shown in Supplementary Figure S5. We find that CpG embeddings relatively contribute more to predictions when the model is confident (predicting an output probability close to 0 or 1). In cases where the local sparsity is low (i.e. a high number of observed sites), CpG Transformer can rely more on local methylation profiles to make a prediction, increasing the relative importance of CpG embeddings. Between different cells, relative contribution differences are negligible. Figure 4E shows the contributions of neighboring observed CpG sites in function of their distance from the prediction site. A decreasing trend is observed with distance, with one bell-shaped bump appearing at +- 160 nucleotides from the prediction site and another smaller bump at +- 400 nucleotides. This relation has been reported on the same cell types in literature by Song et al. [38], who suggested a relation between nucleosome modifications and DNA methylation.

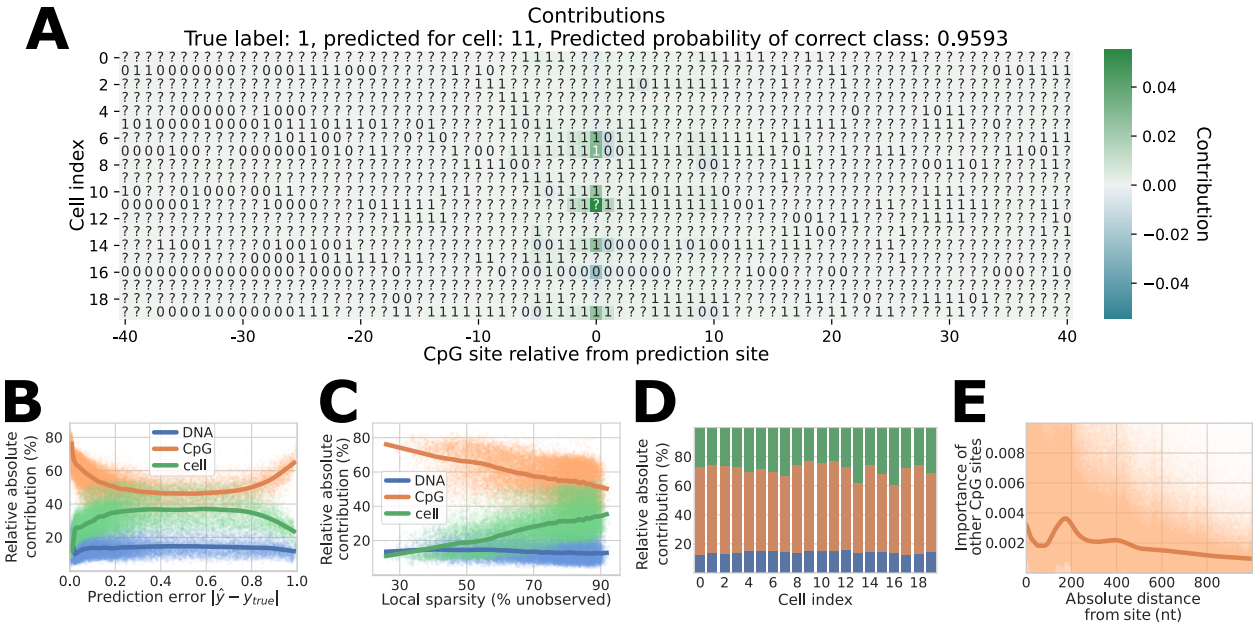

Figure 4: **Integrated Gradients interpretation for the Ser dataset. All trend lines are obtained with LOWESS. [39] A:** Example Integrated Gradients contribution plot. The contributions for the prediction of a single CpG site in cell 11 are shown. Matrix entries with a positive and negative contribution are colored in green and blue, respectively. **B:** Relative absolute contribution of the three input embeddings in function of prediction error. **C:** Relative absolute contribution in function of local sparsity. **D:** Relative absolute contribution stratified per cell. **E:** Contribution of observed CpG sites in function of their absolute distance to the prediction site.

3.3 Transfer learning

Because of the generality of CpG Transformer's self-attention mechanism, it is expected that it learns general-purpose representations of DNA methylation dynamics. In this respect, CpG Transformer is envisioned to transfer well to downstream tasks and other datasets. In this paper, transfer learning is examined in the context of improved convergence speed, a prospect of great interest for practitioners with a limited time and computational budget.

As an experiment, the training dynamics of the HCC datasets are investigated (Figure 5). Three models are trained: one with randomly initialized weights, one with initial weights transferred from the model trained on the MBL dataset and one transferred from the Ser dataset. All model weights apart from the cell embeddings are transferred. The highest achieved performance of all models is similar (97.94, 97.93, and 97.96 ROC AUC, respectively), but the models with transferred weights converge substantially faster. The model transferred from the MBL dataset reaches a ROC AUC within 0.5% of the best performance after only 50 training steps, whereas the model transferred from Ser reaches the same threshold after 250 training steps. This faster initial training is presumably due to the likeness of the HCC and MBL datasets: both containing human cell types profiled using scRRBS-seq, whereas the Ser dataset consists of mouse cells sequenced using scBS-seq. This could also explain the apparent overfitting of the MBL-transferred model after a couple epochs.

4 Discussion

Prior work on DNA methylation imputation has attempted to model intercellular methylation correlations via feature engineering or feature learning. CpG Transformer differentiates itself from these methods due to its novel 2D sliding window attention mechanism, providing a general-purpose way of learning interactions between neighboring CpG sites both within- and between cells. This approach gives rise to many advantages over competing methods. Most simple of all, state-of-the-art imputation performances are obtained over DeepCpG [15] and CaMelia [17]. Secondly, our method lends itself well to interpretation and transfer learning. Finally, because CpG Transformer's model architecture does not compress intercellular methylation relations in any way and uses learned cell embeddings to encode cell identity in a flexible way, we envision CpG Transformer to scale better to datasets containing diverse cell types.

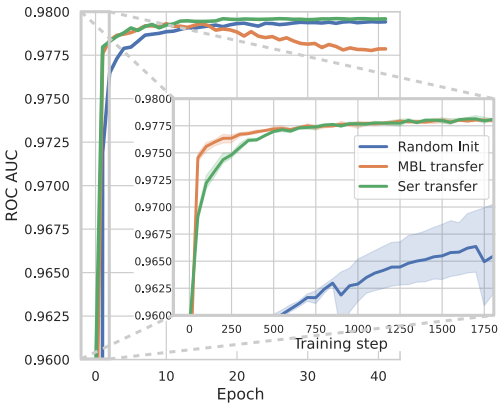

Figure 5: **Transfer learning dynamics on the HCC dataset.** The model initialized randomly is shown in blue. Models with weights initialized from the MBL and Ser dataset are shown in orange and green, respectively. Mean training curves for the first 1750 steps are additionally shown with error bands indicating the max and min performance over three runs.

CpG Transformer allows the prediction of methylation states of thousands of CpG sites in parallel. It does, however, scale quadratically with the number of cells in the dataset. Given the size of the datasets used in this study, this did not pose a problem. For datasets consisting of thousands of cells, however, the application of CpG Transformer as outlined here becomes impossible. In this case, practitioners would need to split their dataset in multiple smaller subsets in which cells are as similar as possible. Alternatively, further extensions of the 2D sliding window attention mechanism could be made in order to allow inputs with a large number of cells. Self-attention sparsity in the cell dimension could, for example, be enforced through clustered attention [40]. In doing so, interactions would only be modeled between clustered, closely-related cells, instead of between all cells. We consider such extensions to be future work.

The proposed sliding window attention attends to neighboring sites within a fixed window, irregardless of whether these neighbors have an observed label or not. A possible disadvantage of this strategy may be that, in cases with extreme sparsity, the model may not be able to properly estimate local methylation profiles. Another approach would be not to model interactions within a fixed local window, but instead to attend to the n nearest neighboring observed entries in every cell via, for example, a `gather` operation. This mechanism would attend to a fixed number of observed CpG sites independent of local sparsity. Since such a mechanism would attend to sites far away on the genome in high sparsity settings, its added value is not straightforwardly estimated. We consider a comparison between these two types of attention to be future work.

Model analysis and interpretation show that local sparsity is an obstacle for the performance of imputation models. Figure 3 surprisingly shows that lowly-covered sites (whose labels are expected to be more noisy) can be more accurately predicted in a densely-covered neighborhood. Some nuances should be made regarding genomic regions that are densely covered but only by a small number of reads for every site. CpG Transformer's masking and randomizing objective falsely assumes no structure in noise and missingness. In reality, for example, one read covering two neighboring unmethylated sites could falsely report methylated signal for both sites if bisulfite treatment failed to convert the corresponding sequence. Hence, lowly-covered sites in densely-covered neighborhoods may be collectively noisy in the same, non-random way. Most contemporary imputation methods, including CpG Transformer, have no way of coping with systematic noise and missingness. In these cases, models will most likely propagate and amplify the noise, potentially compromising biologically-relevant results.

Notwithstanding the above-mentioned considerations, given careful evaluation, CpG Transformer can greatly enhance single-cell methylation studies. A cautious practitioner may, for example, wish to only retain imputations in regions where local sparsity is low and coverage of labels is high. To aid researchers in understanding their imputation results, interpretation methods are introduced. In addition, transfer learning experiments show that CpG Transformer can be used to obtain state-of-the-art imputation performances on a limited time and computational budget.

Data Availability

The data underlying this article is freely available from the Gene Expression Omnibus following the identifiers listed in Section 2. Source code for CpG Transformer is freely available at https://github.com/gdewael/cpg-transformer.

Funding

This work was supported by Ghent University [BOFGOA2020000703 to G.D.W.]. W.W. also received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" Programme.

References

[1] Howard Cedar. Dna methylation and gene activity. *Cell*, 53(1):3–4, 1988.

[2] Adrian Bird. Dna methylation patterns and epigenetic memory. *Genes & development*, 16(1):6–21, 2002.

[3] Felix Krueger, Benjamin Kreck, Andre Franke, and Simon R Andrews. Dna methylome analysis using short bisulfite sequencing data. *Nature methods*, 9(2):145, 2012.

[4] Sébastien A Smallwood, Heather J Lee, Christof Angermueller, Felix Krueger, Heba Saadeh, Julian Peat, Simon R Andrews, Oliver Stegle, Wolf Reik, and Gavin Kelsey. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature methods*, 11(8):817–820, 2014.

[5] Hongshan Guo, Ping Zhu, Xinglong Wu, Xianlong Li, Lu Wen, and Fuchou Tang. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome research*, 23(12):2126–2135, 2013.

[6] Christof Angermueller, Stephen J Clark, Heather J Lee, Iain C Macaulay, Mabel J Teng, Tim Xiaoming Hu, Felix Krueger, Sébastien A Smallwood, Chris P Ponting, Thierry Voet, et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature methods*, 13(3):229–232, 2016.

[7] Matthias Farlik, Florian Halbritter, Fabian Müller, Fizzah A Choudry, Peter Ebert, Johanna Klughammer, Samantha Farrow, Antonella Santoro, Valerio Ciaurro, Anthony Mathur, et al. Dna methylation dynamics of human hematopoietic stem cell differentiation. *Cell stem cell*, 19(6):808–822, 2016.

[8] Joshua J Levy, Alexander J Titus, Curtis L Petersen, Youdinghuan Chen, Lucas A Salas, and Brock C Christensen. Methylnet: an automated and modular deep learning approach for dna methylation analysis. *BMC bioinformatics*, 21(1):1–15, 2020.

[9] Pietro Di Lena, Claudia Sala, Andrea Prodi, and Christine Nardini. Missing value estimation methods for dna methylation data. *Bioinformatics*, 35(19):3786–3793, 2019.

[10] Weiwei Zhang, Tim D Spector, Panos Deloukas, Jordana T Bell, and Barbara E Engelhardt. Predicting genome-wide dna methylation using methylation marks, genomic position, and dna regulatory elements. *Genome biology*, 16(1):1–20, 2015.

[11] Yeping Lina Qiu, Hong Zheng, and Olivier Gevaert. A deep learning framework for imputing missing values in genomic data. *bioRxiv*, page 406066, 2018.

[12] Luli S Zou, Michael R Erdos, D Leland Taylor, Peter S Chines, Arushi Varshney, Stephen CJ Parker, Francis S Collins, and John P Didion. Boostme accurately predicts dna methylation values in whole-genome bisulfite sequencing of multiple human tissues. *BMC genomics*, 19(1):1–15, 2018.

[13] Fangtang Yu, Chao Xu, Hong-Wen Deng, and Hui Shen. A novel computational strategy for dna methylation imputation using mixture regression model (mrm). *BMC bioinformatics*, 21(1):1–17, 2020.

[14] Chantriolnt-Andreas Kapourani and Guido Sanguinetti. Melissa: Bayesian clustering and imputation of single-cell methylomes. *Genome biology*, 20(1):1–15, 2019.

[15] Christof Angermueller, Heather J Lee, Wolf Reik, and Oliver Stegle. Deepcpg: accurate prediction of single-cell dna methylation states using deep learning. *Genome biology*, 18(1):1–13, 2017.

[16] Limin Jiang, Chongqing Wang, Jijun Tang, and Fei Guo. Lightcpg: a multi-view cpg sites detection on single-cell whole genome sequence data. *BMC genomics*, 20(1):1–17, 2019.

[17] Jianxiong Tang, Jianxiao Zou, Mei Fan, Qi Tian, Jiyang Zhang, and Shicai Fan. Camelia: imputation in single-cell methylomes based on local similarities between cells. *Bioinformatics*, 2021.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[19] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[20] Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature reviews genetics*, 5(2):101–113, 2004.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[22] Jim Clauwaert and Willem Waegeman. Novel transformer networks for improved sequence labeling in genomics. *bioRxiv*, page 836163, 2020.

[23] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Debsindhu Bhowmik, et al. Prottrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing. *arXiv preprint arXiv:2007.06225*, 2020.

[24] Roshan Rao, Jason Liu, Robert Verkuil, Joshua Meier, John F Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. *bioRxiv*, 2021.

[25] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. *arXiv preprint arXiv:1912.12180*, 2019.

[26] John Jumper, R Evans, A Pritzel, T Green, M Figurnov, K Tunyasuvunakool, O Ronneberger, R Bates, A Zidek, A Bridgland, et al. High accuracy protein structure prediction using deep learning. *Fourteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstract Book)*, 22:24, 2020.

[27] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pages 173–182, 2017.

[28] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[29] Shawn J Cokus, Suhua Feng, Xiaoyu Zhang, Zugen Chen, Barry Merriman, Christian D Haudenschild, Sriharsa Pradhan, Stanley F Nelson, Matteo Pellegrini, and Steven E Jacobsen. Shotgun bisulphite sequencing of the arabidopsis genome reveals dna methylation patterning. *Nature*, 452(7184):215–219, 2008.

[30] Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *arXiv preprint arXiv:2007.14062*, 2020.

[31] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[33] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[34] Yu Hou, Huahu Guo, Chen Cao, Xianlong Li, Boqiang Hu, Ping Zhu, Xinglong Wu, Lu Wen, Fuchou Tang, Yanyi Huang, et al. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell research*, 26(3):304–319, 2016.

[35] Helene Kretzmer, Anat Biran, Noelia Purroy, Camilla Lemvigh, Kendell Clement, Michaela Gruber, Hongcang Gu, Laura Rassenti, Arman W Mohammad, Connie Lesnick, et al. Preneoplastic alterations define cll dna methylome and persist through disease progression and therapy. *Blood cancer discovery*, 2(1):54, 2021.

[36] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[37] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017.

[38] You Song, Honglei Ren, and Jinzhi Lei. Collaborations between cpg sites in dna methylation. *International Journal of Modern Physics B*, 31(20):1750243, 2017.

[39] William S Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836, 1979.

[40] Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. Efficient content-based sparse attention with routing transformers. *Transactions of the Association for Computational Linguistics*, 9:53–68, 2021.