scHiMe: Predicting single-cell DNA methylation levels based on single-cell Hi-C data

Hao Zhu[1], Tong Liu[1], and Zheng Wang[1,*]

[1]Department of Computer Science, University of Miami, 1365 Memorial Drive, Coral Gables, FL 33124, USA.

[*]Corresponding author: zheng.wang@miami.edu

Abstract

Motivation: Recently a biochemistry experiment was developed to simultaneously capture the chromosomal conformations and DNA methylation levels on single cells. A computational tool to predict single-cell methylation levels based on single-cell Hi-C data becomes necessary due to the availability of this experiment. Moreover, if a promising prediction accuracy is achieved, it can prove the existence of an influential relationship between three-dimensional (3D) genome structures and DNA methylation for single cells.

Results: A computational tool named scHiMe was developed to predict the base-pair-specific methylation levels in the promoter regions genome-wide based on the single-cell Hi-C data and DNA nucleotide sequences using the graph transformer algorithm. Our evaluation results showed a high consistency between the predicted and the true methylation levels. The predicted DNA methylation levels also resulted in accurate classifications of cells into different cell types, which indicated that our algorithm captured the cell-to-cell variability in the single-cell Hi-C data and further indicated the existence of an interrelated relationship between the 3D genome and DNA methylations for single cells.

Availability: Our tool scHiMe is freely available at http://dna.cs.miami.edu/scHiMe/. A user can submit single-cell Hi-C data and the server will predict single-cell methylation levels on all of the promoter regions genome-wide.

Supplementary information: Supplementary data are available at Bioinformatics online.

1. Introduction

DNA methylation in mammals is a biological process, in which a methyl group is attached almost exclusively to the cytosine. It regulates gene expression by blocking the binding of transcription factors to DNA or by recruiting the proteins involved in gene repression (Moore, et al., 2013). Many cellular processes including embryonic development, X chromosome inactivation, and chromosome stability maintenance are related to DNA methylation (Robertson, 2005).

The chromosome conformation capture (3C) technology was developed to reveal the spatial proximities between a single pair of genomic loci in a cell population (Dekker, et al., 2002). After that, 3C-On-Chip (4C) (Simonis, et al., 2006), 3C-Carbon Copy (5C) (Dostie, et al., 2006), and Hi-C techniques (Lieberman-Aiden, et al., 2009) have been developed.

A new experiment protocol named single-nucleus methyl-3C sequencing was recently developed that could simultaneously detect both single-cell DNA methylations and the single-cell spatial proximities of the genomes (Lee, et al., 2019). Li *et al.* reported a molecular assay called Methyl-HiC that simultaneously captured the chromosomal conformations and DNA methylations in individual mouse cells (Li, et al., 2019).

2. Materials and methods

2.1 Overview

Fig. 1 shows the overview of the methodologies used in our scHiMe. The single-cell Hi-C data of the target cell was used to generate a promoter-promoter Hi-C contact matrix for the target cell. The idea of meta-cell was then applied, with which the promoter-promoter Hi-C contact matrices of the neighboring cells were aggregated to generate the aggregated promoter-promoter Hi-C contact matrix for the target cell. Based on the aggregated promoter-promoter Hi-C contact matrix, a promoter-promoter spatial interaction network was constructed. The true base-pair-specific DNA methylation values or target values for the 1000 base pairs in the target promoter were generated also based on meta-cell. Node features and edge features were generated and input into the graph transformer network, which contained five blocks of graph transformer. Details of each step will be presented in later sections.

2.2 Data sets

The first dataset used in this research contains the simultaneously-captured single-cell Hi-C data and DNA methylation data from the brain cells of a 29-year-old male caucasian (Lee, et al., 2019). The cells in this data set are from 14 different cell types, and we only kept and used nine cell types (Astro, ODC, Ndnf, Sst, Vip, Endo, L23, MG, and OPC) that contained at least 30 cells. We used the cells in the Astro, ODC, and Ndnf cell types to build the training data set, cells in the Sst and Vip cell types for the validation data set, and the cells in Endo, L23,

MG, and OPC cell types for the blind-test data set. Details about our quality controls can be found in the supplementary document. This first dataset was used for training, validation, and blind test and was also used to benchmark the performance of classifying cells into different cell types based on the predicted single-cell methylation levels of promoters.

The second dataset is similar to the first but from a 21-year-old male caucasian also from the research (Lee, et al., 2019). We picked the same four cell types: Endo, L23, MG, and OPC, and applied the previously-mentioned quality-control criteria. This dataset was only used for blind testing and clustering cells into cell types based on the predicted single-cell methylation levels of promoters.

We used another human single-cell Hi-C data as our third dataset that was from (Ramani, et al., 2017). We used three cell types: K562, HAP1, and HeLa, and selected 100 cells from each cell type that had the highest number of single-cell Hi-C contacts. Since single-cell methylation data were not included in this data set (the authors only performed single-cell Hi-C experiments), we only used it for benchmarking the performance of classifying cells into cell types based on the predicted methylation levels of promoters.

The transcription start site (TSS) definition used in this research was downloaded from the GENCODE human annotation version 19 (Frankish, et al., 2019).

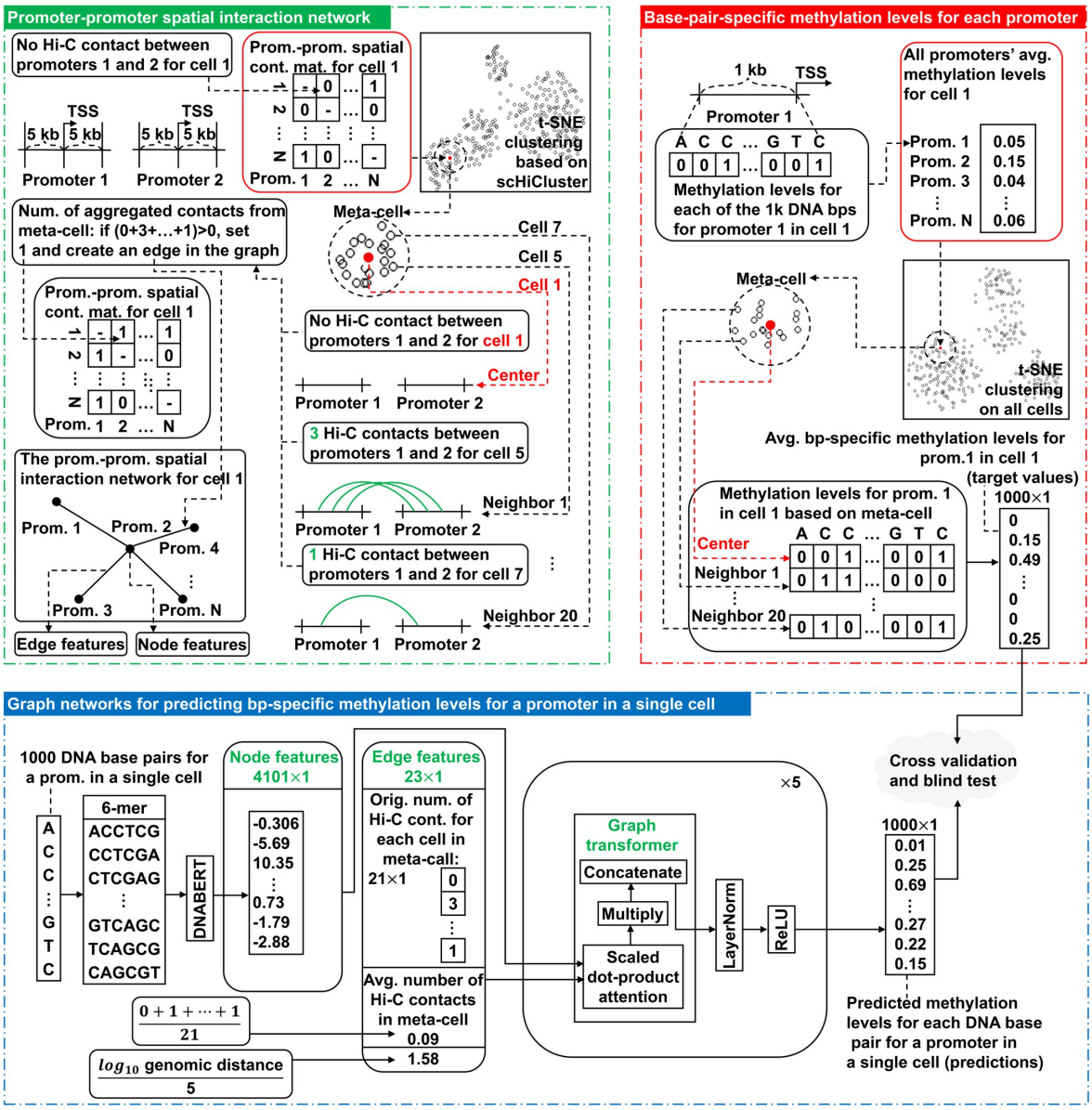


Figure 1. The overview of the architectures of scHiMe. The top left area illustrates the process of building promoter-promoter spatial interaction networks. The top right area shows how the true DNA methylation values, or the target values for the machine learning algorithm, are defined based on meta-cells. The bottom area shows the architecture of the graph transformer including the illustrations of node and edge features.

2.3 Building meta-cells for single-cell Hi-C data

Due to the sparsity of the single-cell Hi-C data of a target cell, we combined the target cell and other cells, which share similar Hi-C patterns

with the target cell, into a meta-cell and used the single-cell Hi-C contacts of the meta-cell as the Hi-C contacts for the target cell. For the cells used for training and validation, their cell types were considered known information. Therefore, the similar or neighboring cells used to build each meta-cell were selected from the same cell type as the target or central cell. Details can be found in the supplementary file.

2.4 Building meta-cells for single-cell DNA methylation data

Due to the sparsity of the single-cell methylation data, we combined the target cell with other neighboring cells, which share similar methylation patterns with the target cell, into meta-cell and used the aggregated single-cell methylation levels of the meta-cell as the methylation levels for the target cell. The details can be found in the supplementary document.

2.5 Promoter-promoter spatial interaction network

In the promoter-promoter spatial interaction network, each node represents a promoter. The aggregated Hi-C contacts from meta-cells were used to define edges in the promoter-promoter spatial interaction networks. An edge was created between two promoters if at least one aggregated Hi-C contact exists between the two promoter regions. The promoter region used here was defined as the genomic region that contains 5 kb upstream and downstream of the TSS plus the TSS of a gene.

2.6 Node and edge features

The node features were the encodings of DNA nucleotide sequence for each 1-kb promoter/node, which was generated by DNABERT (Ji, et al., 2021). Specifically, we generated the 6-mer representation for each base pair in the DNA sequence, and input the representations of all base pairs into DNABERT, which outputted 4101 real numbers as the codings of the promoter. For the promoters existing in the negative strand, we used their complementary sequences in the positive strand in the order of 5' to 3' (this 5' to 3' is in terms of the positive strand) as the sequences that were input to DNABERT.

The first edge feature contains 21 integers, which are the numbers of Hi-C contacts, between the two promoters that the edge connects, in the target cell and 20 neighboring cells in the meta-cell. The second edge feature is the average number of occurrences of nonzero Hi-C contacts in the meta cell. An example is given in the supplementary document. In total, the edge feature for each edge consists of 23 values.

2.7 Graph transformer

The node features that are input to the o-th block of the graph transformer are $P^{(o)} = \{p_1^{(o)}, p_2^{(o)}, \dots, p_n^{(o)}\}$, where p stands for a promoter, and n is the number of promoters. Based on the node features, a query vector for the i-th node feature is defined as: $q_{h,i}^{(o)} = W_{h,q}^{(o)} p_i^{(o)} + b_{h,q}^{(o)}$, where h indicates that it is the h-th head in the block, and $W_{h,q}^{(o)}$ and $b_{h,q}^{(o)}$ are the trainable parameters for the query vector.

Similarly, a key vector for the j-th node feature is defined as: $k_{h,j}^{(o)} = W_{h,k}^{(o)} p_j^{(o)} + b_{h,k}^{(o)}$.

The edge features input to the transformer block are represented as $E^{(o)} = \{e_1^{(o)}, e_2^{(o)}, \dots, e_n^{(o)}\}$. Specifically, the edge feature for the edge between node i and node j after the h head-attentions is encoded as: $e_{h,ij}^{(o)} = W_{h,e}^{(o)} e_{ij}^{(o)} + b_{h,e}^{(o)}$, where $e_{ij}^{(o)}$ is the original edge features between node i and node j, and $W_{h,e}^{(o)}$ and $b_{h,e}^{(o)}$ are trainable parameters.

The encoded edge features will be added to the key vector for calculating the multi-head attention: $\alpha_{h,ij}^{(o)} = \frac{\langle q_{h,i}^{(o)}, k_{h,j}^{(o)} + e_{h,ij}^{(o)} \rangle}{\sum_{u \in \mathcal{N}(i)} \langle q_{h,i}^{(o)}, k_{h,j}^{(o)} + e_{h,ij}^{(o)} \rangle}$

, where $\langle q, k + e \rangle$ is the exponential scale dot-product: $\langle q, k \rangle = \exp\left(\frac{q^T k}{\sqrt{d}}\right)$ with d is the hidden size of each head attention.

After the above h-th head attention is completed, a value vector is defined as: $v_{h,j}^{(o)} = W_{h,v}^{(o)} p_j^{(o)} + b_{h,v}^{(o)}$, where $W_{h,v}^{(o)}$ and $b_{h,v}^{(o)}$ are trainable parameters.

Message aggregation is performed to integrate the value vectors of all the radius-one neighboring nodes, each of which is labeled as j, to node i as: $\hat{p}_i^{(o+1)} = \|_1^h [\sum_{j \in \mathcal{N}(i)} \alpha_{h,ij}^{(o)} (v_{h,j}^{(o)} + e_{h,ij}^{(o)})]$, where $\|_1^h$ is the concatenation operation that is performed on all h head attentions.

Different from the paper (Shi, et al., 2020), we directly applied the normalization layer and ReLU layer to the multi-head attention as follows, which, based on our evaluation, resulted in better performance (results of comparisons not shown): $p_i^{(o+1)} = ReLU(LayerNorm(\hat{p}_i^{(o+1)}))$.

Also different from the paper (Shi, et al., 2020), which only aggregated head attention for the node features, we also aggregated the head attention for the edge features as follows, which resulted in better performance (results of comparisons not shown): $\hat{e}_{ij}^{(o+1)} = \|_1^h (\alpha_{h,ij}^{(o)}, e_{ij}^{(o)})$.

A normalization and ReLU layer are added to get the edge features that will be input to the next block: $e_{ij}^{(o+1)} = ReLU(LayerNorm(\hat{e}_{ij}^{(o+1)}))$.

For the last block, the ReLU layer and the normalization layer are not applied. Instead, the concatenation operation on all h head attentions is replaced by the average on multi-head output as: $p_i^{(o+1)} = \frac{1}{\mathrm{H}} \sum_{h=1}^{H} [\sum_{j \in \mathcal{N}(i)} \alpha_{h,ij}^{(o)} (v_{h,j}^{(o)} + e_{h,ij}^{(o)})]$.

2.8 Training, validation, and blind test

Three cell types in the first data set were used as the training data, which were "Astro", "ODC", and "Ndnf" cell types. The total number of graphs used for training is 21,745. The graph transformer that we used was modified based on the transformer_conv in the torch_geometric.nn.conv pack (Paszke, et al., 2019), which was published by (Shi, et al., 2020). We tested different values of the number of transformer blocks including 3, 5, 7, and 10, and different values of head attentions including 1, 2, and 5, and found the number of blocks of 5 and the number of head attentions of 1 resulted in the best performance, which was used in our final model. Other details can be found in the supplementary document.

2.9 Evaluation

2.9.1 Naive predictors

Since there is no other published tool to compare performance, we built two naive predictors that make predictions directly based on the methylation levels of the neighboring nodes in the promoter-promoter spatial interaction networks. The details about these native predictors can be found in the supplementary file.

2.9.2 Performance metrics

To benchmark the performance of scHiMe, we calculated the Pearson's correlation (PCC), Matthews correlation (MCC), average precision (AP), and area under the curve (AUC) of the receiver operating characteristic curve (ROC) between the predictions and ground truths. Our evaluations were performed only on the base pairs that had true methylation levels available.

We parsed the predicted methylation levels on all the cytosines and guanines in CpGs because guanine in the positive strand indicates cytosine in the negative strand.

2.9.3 Cell classification and its evaluation

The predicted methylation levels on all the CpGs of all promoters were input into the K-means clustering algorithm with the number of clusters set to four. To evaluate the clustering results, or evaluate whether the groupings of cells in the clustering results match the true cell types, we applied the adjusted rand index (ARI) to measure the similarities between the clustering results and the true cell types. We named this type of ARI score ARI-all-data. A close to zero ARI score indicates random labeling, and the ARI score of one indicates that two clusterings are identical.

To visualize the predicted methylation levels on all the CpGs of all promoters, we input the data into the t-distributed stochastic neighbor embedding (t-SNE) algorithm, and the top two components were obtained, which were treated as the X and Y coordinates for plotting the cells in a two-dimensional (2D) space. Another type of ARI score was calculated based on the clustering results of these top two components, which was referred to as ARI-two-components in this paper. Details can be found in the supplementary document.

2.9.4 Importance of individual CpGs in promoters for clustering cells

We used the random forest algorithm from the sklearn package (Pedregosa, et al., 2011) to measure the importance of individual CpGs in promoters for classifying cells. Random forest was used to classify a cell into one of the four cell types: Endo, L23, MG, and OPC. We used the predicted methylation levels on all of the CpGs, including both cytosines and guanines, from all of the promoters genome-wide as the features. More details can be found in the supplementary document.

2.9.5 Cell clustering based on housekeeping genes

Housekeeping genes are typically constitutive genes that are required for the maintenance of basic cellular functions. We downloaded the human housekeeping genes from the HRT Atlas database (Hounkpe, et al., 2021). Among all the housekeeping genes that we downloaded, 1967 of them matched the genes that were used to define the promoters in our research and were used in this analysis. We randomly selected another 1967 promoter from all of the promoters that we had in the research and used them as the baseline analysis.

2.9.6 Collective Influence algorithm

We applied the collective influence (CI) algorithm (Morone and Makse, 2015) to the promoter-promoter spatial interaction networks built on data set 2. A list of influencers ranked from the most significant to the least significant influencers was outputted from the CI algorithm for each cell. To find the relationship between the ranking position of each promoter in the influencer list and the ARI score of using the promoter to classify cells, we defined two ways of calculating the average ranking position in the influencer list for each promoter, which were named avg_inf_rank1 and avg_inf_rank2. Details can be found in the supplementary document.

3. Results

3.1 Blind test results on data set 1 and data set 2

Table 1 shows the evaluation results for the predictions from scHiMe based on five evaluation metrics, which were conducted on 328 and 1053 cells in the blind-test data sets generated from data sets 1 and 2, respectively. We can find that the overall absolute difference between the scHiMe-predicted and true methylation levels is less than 0.1 on average, and the AUC of the ROC for scHiMe are above 0.92 on both

data sets. The performances of scHiMe are much better than the performances of the two naive predictors, which indicates the contributions of the graph transformer in terms of capturing graph topology and data patterns in the networks and improving performances.

Table1. The evaluation results for scHiMe and two naive predictors on the blind-test cells on data sets 1 and 2.

Predicted method	Data set	PCC	Average precision	MCC	AUC	Average abs. error
scHiMe	1	0.723	0.768	0.701	0.934	0.097
scHiMe	2	0.716	0.767	0.694	0.937	0.099
Naive 1	1	0.059	0.189	0.058	0.533	0.286
Naive 1	2	0.097	0.203	0.095	0.550	0.274
Naive 2	1	0.485	0.377	0.479	0.813	0.154
Naive 2	2	0.545	0.441	0.530	0.836	0.148

To further analyze the performances of scHiMe, we plotted the absolute differences between the scHiMe-predicted and true methylation levels for different ranges of the true methylation levels, see **Fig. 2 A** and **B** for data sets 1 and 2, respectively. We can see that when the true methylation levels are 0 and 1, the absolute differences reach the minimum, which are 0.057 and 0.208 for data set 1, and 0.060 and 0.234 for data set 2, respectively. The largest absolute differences were found when the true methylation levels are in the range of (0.5, 0.6], which are 0.412 and 0.413 for data sets 1 and 2, respectively. However, **Fig. 2 C** and **D** show that almost all of the true methylation levels are at 0 and 1, which are 77.9% and 18.5% for data set 1, and 78.9% and 19.3% for data set 2, respectively. The ROC of the scHiMe predictions are shown in **Fig. 2 E** and **F** for data sets 1 and 2, respectively.

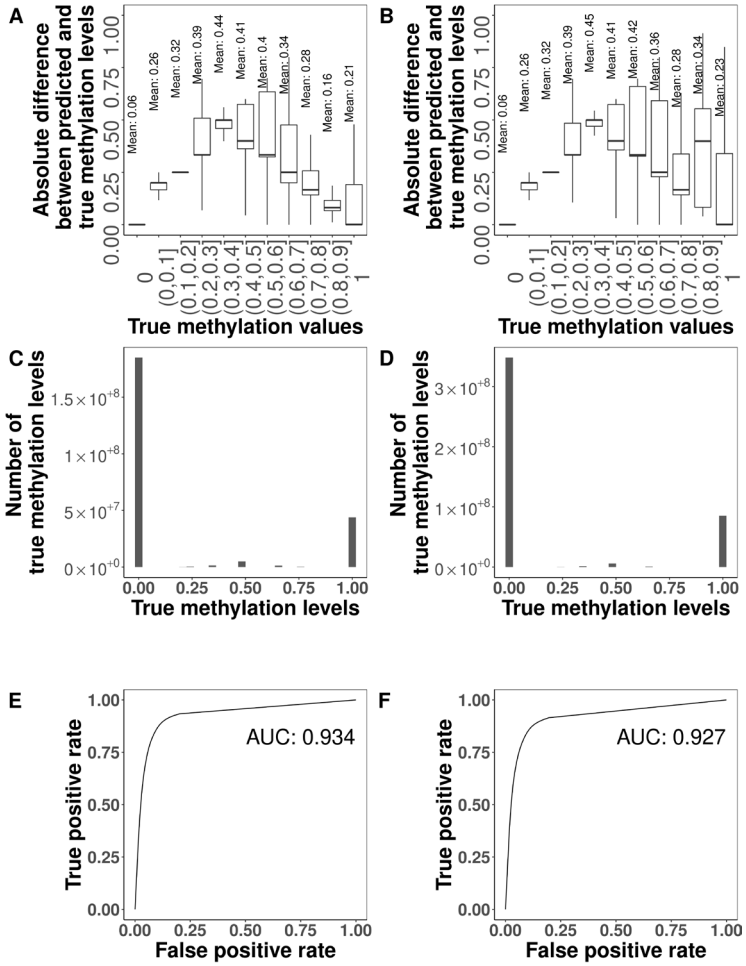


Figure 2. A-B. Absolute differences between the scHiMe-predicted and true methylation levels on the blind-test cells in data sets 1 and 2, respectively. The numbers shown in A and B are the mean values for different ranges of true methylation levels. **C-D.** Distributions of the true methylation levels on data sets 1 and 2, respectively. **E-F.** ROC curves and the AUC scores for the predictions of scHiMe on the blind-test cells in data sets 1 and 2, respectively.

3.2 Cell-type classification based on the methylation levels of all promoters

Fig. 3 A and **B** illustrate the first two components from t-SNE based on the true methylation levels on data sets 1 and 2, respectively. For the cytosines and guanines of all CpGs of all promoters, if there are true methylation levels available, the true methylation levels were used for

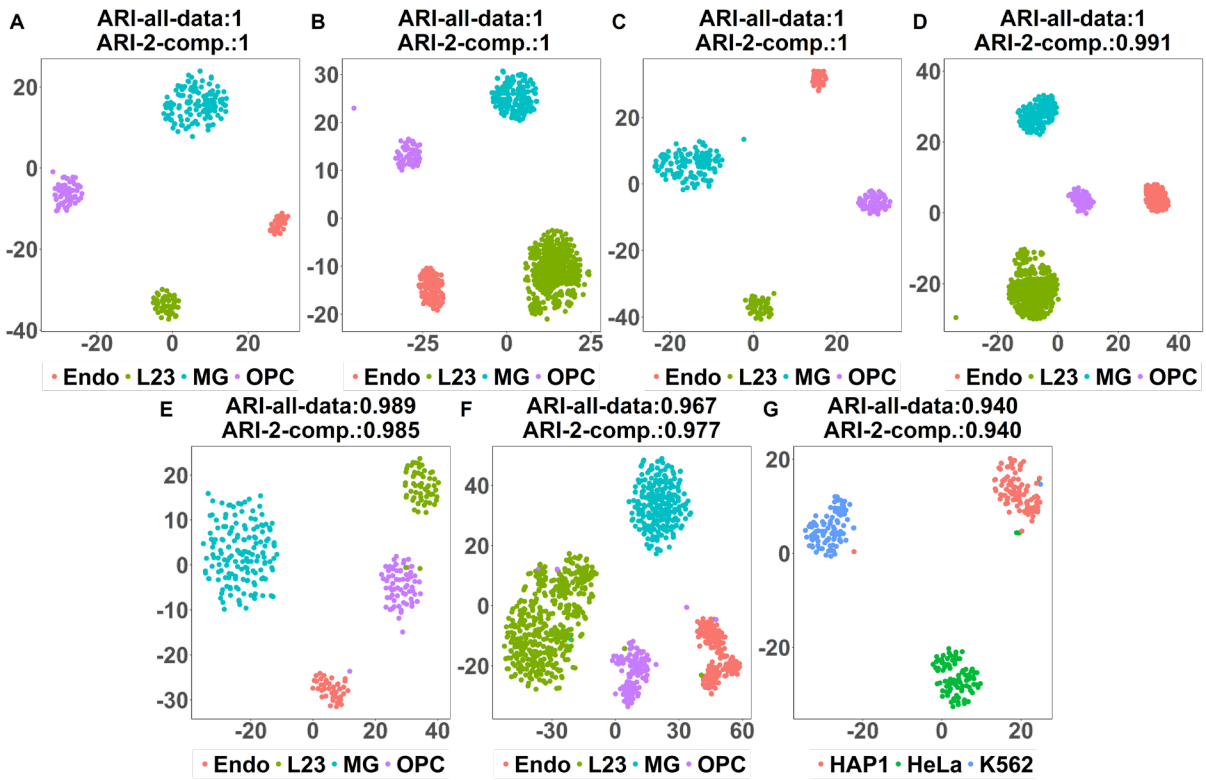
the cytosines and guanines; if not available, -1 was used. It can be found that the cells were clustered almost perfectly to their real cell types, which matched the results of a similar experiment in (Lee, et al., 2019).

Fig. 3 C and **D** are similar to **Fig. 3 A** and **B**, but the scHiMe-predicted methylation levels were used for the cytosines and guanines of CpGs that have true methylation levels available. For the cytosines and guanines of CpGs that did not have true methylation levels available, -1 was still used. It can be found that almost all of the cells were still correctly classified. This demonstrates that the predictions from scHiMe maintain the unique patterns of single-cell methylation data for each cell type.

For **Fig. 3 E** and **F**, we used the scHiMe-predicted methylation levels for all of the cytosines and guanines of CpGs, no matter whether the true methylation levels were available or not. This experiment tested whether scHiMe-predicted methylation levels resulted in the correct classification of cells even for the cytosines and guanines of CpGs that did not have true methylation data available from the wet-lab experiments. **Fig. 3 E** and **F** demonstrate that almost all of the cells were still correctly classified into partitions, which matched their true cell types.

Fig. 3 G was similar to **Fig. 3 E** and **F**, but was on data set 3, which was another independent data set. It can be found that scHiMe achieved a similar performance on this data set.

Figure 3. A. Visualization of the top two components from t-SNE based on true methylation levels on data set 1. For the Cs and Gs of CpGs that did not have methylation



data available, -1 was used. **B**. Similar to A, but on data set 2. **C**. Visualization of the top two components from t-SNE based on scHiMe-predicted methylation levels on data set 1. For the Cs and Gs of CpGs that did not have methylation data available, -1 was used. **D**. Similar to C, but on data set 2. **E**. Visualization of the top two components from t-SNE based on scHiMe-predicted methylation levels on all Cs and Gs of CpGs no matter whether true methylation levels are available or not. **F**. Similar to E, but on data set 2. **G**. Similar to E, but on data set 3.

3.3 Cell-type classification based on the methylation levels of individual promoters and individual cytosines and guanines in CpGs

Fig. 4 A shows the ARI-all-data scores of clustering cells based on the predicted methylation levels of all cytosines and guanines of the individual or selected promoters on data set 2. We used the scHiMe-predicted methylation levels for each promoter to cluster cells and then ranked the promoters based on the ARI-all-data scores. It can be found that using only the first-ranked promoter resulted in an ARI score of 0.664 and using the top 32 promoters brought the ARI score to 0.898. However, using more than top 50 promoters did not further improve the ARI scores. The ARI scores and the visualizations of clustering results on the top two components for the No. 1-200 and top 1-200 promoters can be found in supplementary data.

We further tested the performance of clustering cells using the scHiMe-predicted methylation level of each cytosine or guanine of a CpG on data set 2. Supplementary **Fig. S1** shows the distribution of the number of cytosines and guanines having various values of feature importance. We ranked all of the cytosines and guanines based on their feature importance and clustered cells using the top n cytosines or guanines. The ARI-all-data scores for the top 1-1000 were plotted in **Fig. 4 B**. It can be found that when the top 28 cytosines or guanines were used, the ARI score reached 0.572 and remained there even when the number of cytosines or guanines went up to 1000. The ARI-all-data scores and the visualizations of the top two t-SNE components of the top 1-1000 cytosines or guanines can be found in the supplementary data.

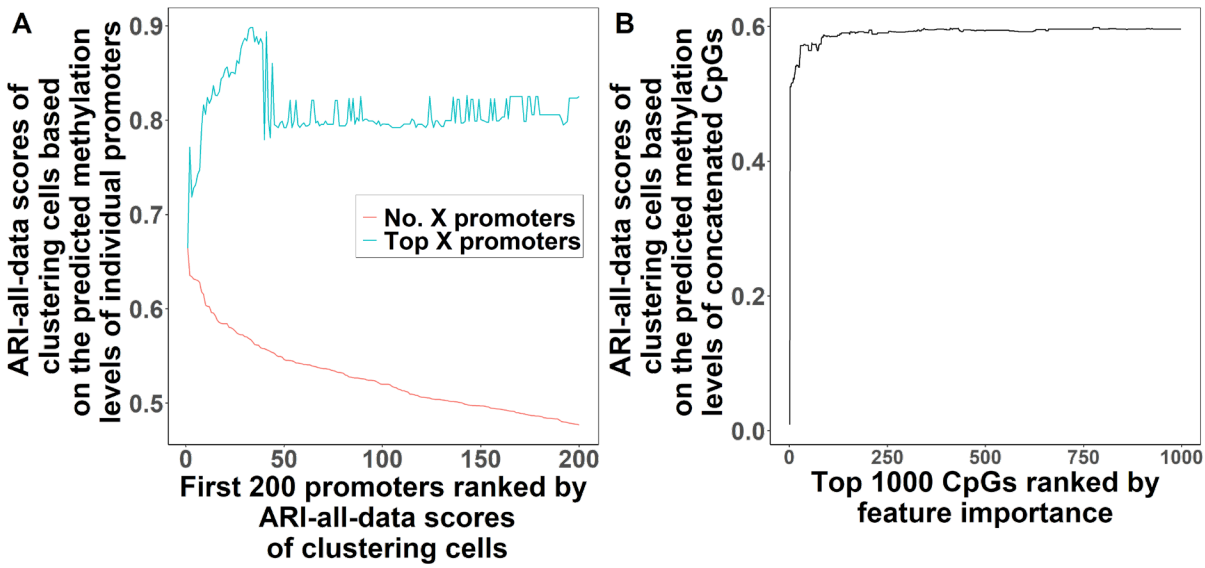


Figure 4. A. ARI-all-data scores of clustering on the first 200 promoters ranked by ARI scores. The first-ranked promoter was labeled as No. 1, the second-ranked No. 2, and so on. The red line indicated the ARI-all-data scores when only No. n ($1 \leq n \leq 200$) promoter was used, and the values of n were shown as the X-axis. The blue line indicated the performance when top n promoters were used, which meant when all of the first-ranked, second-ranked, …, and the n-th-ranked promoters were used. **B.** The ARI-all-data scores of the top 1000 CpGs that were ranked by feature importance.

3.4 Performance from the perspective of network degree

The degree of a promoter in a promoter-promoter spatial interaction network was defined as the number of edges that connected to the promoter. **Fig. 5 A** shows the evaluation metrics for different degree values from 1 to 40 on data set 2. It can be found that the performances are stable when the degree is from 1 to 20, and the performances slightly increase and then vary when the degree is larger than 20. The best AUC 0.965 is with degree 34, and the best Pearson correlation, Matthews correlation coefficient, average precision, and average absolute difference, which are 0.864, 0.853, 0.886, and 0.045, respectively, are with degree 36.

Fig. 5 B illustrates the average degree distribution for all of the promoters on data set 2. We gathered the degree values of each promoter in the promoter-promoter spatial interaction networks of different cells and then calculated an average degree value for each promoter. **Fig. 5 B** shows that most of the promoters have average degrees of 2 and 3.

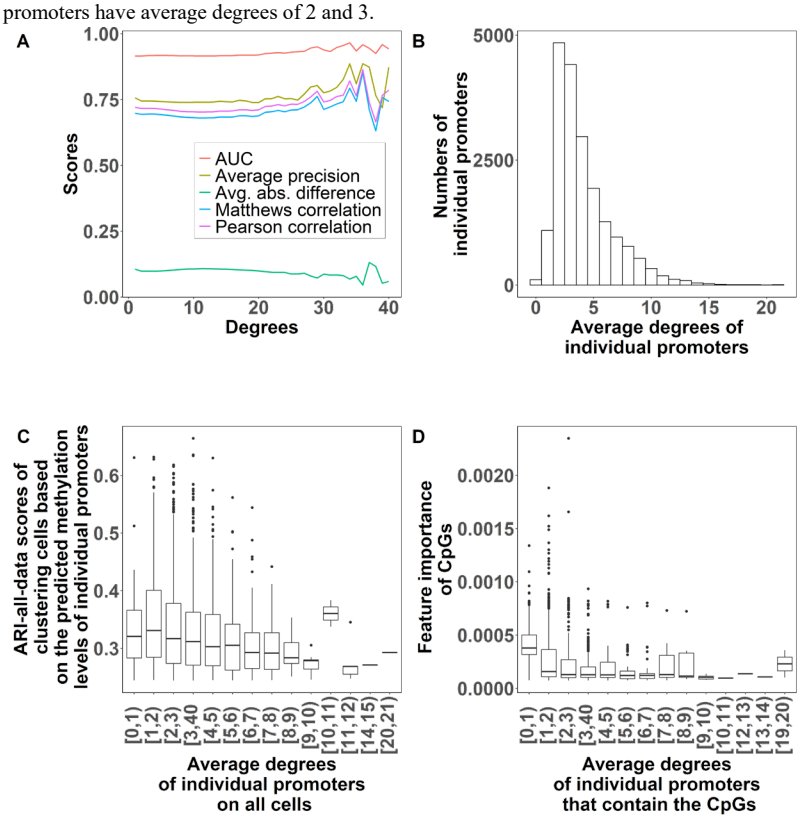


Figure 5. A. Evaluation matrices for different degree values on data set 2. **B.** Distribution of average degree for all of the promoters on data set 2. **C.** ARI-all-data scores of clustering cells based on the predicted methylation levels of the individual promoters of different average degrees. **D.** Feature importance for all cytosines and guanines of CpGs from the individual promoters of different average degrees.

We further tested the performances when using the individual promoters having various average degrees to cluster cells. **Fig. 5 C** shows the ARI-all-data scores when using the individual promoters of different average degrees to cluster cells. The promoters used in **Fig. 5 C** were the top 3000 promoters ranked by the ARI-all-data scores when each of them was used to cluster cells. It can be found that the promoters with average degrees between 10 and 11 are associated with the highest ARI scores although it needs to be noticed that few promoters fall into the degree range of [10, 11). Other than that, the promoters in the degree range of [1, 2) resulted in slightly higher ARI scores when used to cluster cells into different types. In general, the increase in the average degree values is associated with the decrease in ARI scores. Supplementary **Fig. S2** shows the same plot but generated on all promoters, which shows that the promoters with smaller degrees are associated with higher ARI scores.

We gathered the feature importance generated by random forest for all cytosines and guanines of CpGs, and then visualized the feature importance concerning the degrees of the promoters that contained the cytosines and guanines, see **Fig. 5 D**. Only the top 3000 cytosines or guanines that have the highest feature importance values were used for the analysis in **Fig. 5 D**. Interestingly, the higher the average degree value is, the lower mean feature importance the cytosines and guanines have.

3.5 Performance from the perspective of network influencer

We also benchmarked the performance of scHiMe from the perspective of network influencers. Supplementary **Fig. S3** shows the distribution of the number of influencers found for all the cells in data set 2. It can be found that 85% of the cells have from 14007 to 17822 influencers detected from their promoter-promoter spatial interaction networks. We studied the degrees of the top 10 influencers for all cells in data set 2, see the distribution in **Fig. 6 A**. We did find higher mean degree values for the first-ranked and second-ranked influencers, which are 7.374 and 6.964, respectively. For the influencers ranked from the third to the tenth, they have similar mean degree values.

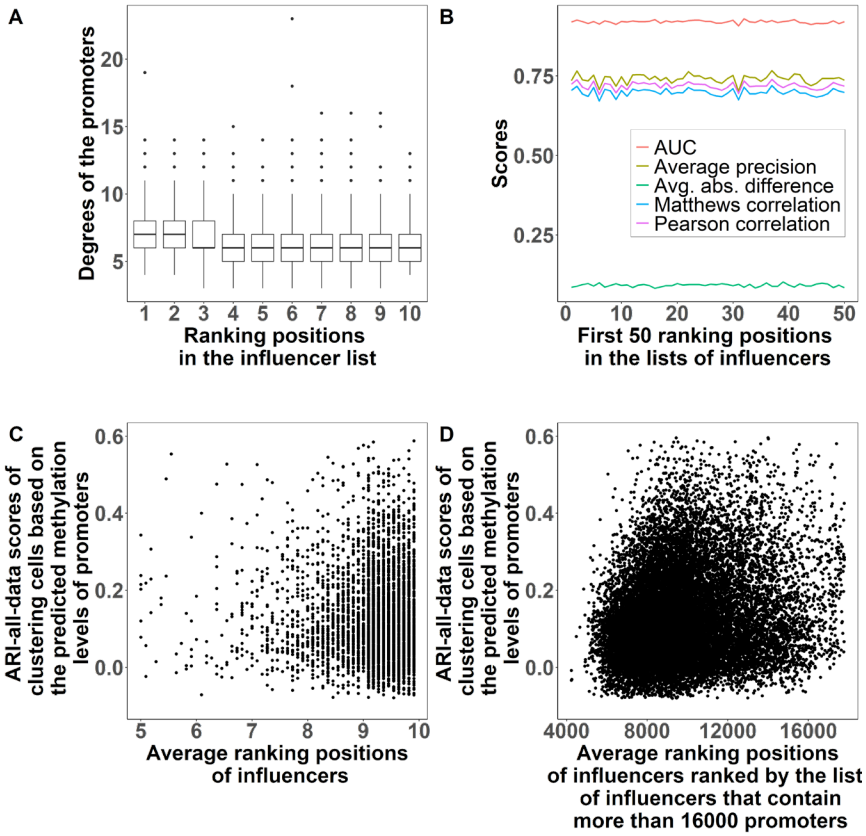


Figure 6. A. The distribution of the degrees of the top 10 influencers for all cells in data set 2. **B.** Evaluation matrices for top 50 influencers on data set 2. **C.** ARI-all-data scores of clustering cells based on the predicted methylation levels on all cells based on the top 10 most significant influencers (promoters). **D.** Similar to **C**, but on the cells that have at least 16000 influencers (promoters) in the list.

Fig. 6 B shows the evaluation metrics for the influencers ranked from the first to the fourteenth on data set 2. We did not find a trend of increasing or decreasing for all of the evaluation metrics with the increase of the influencer ranking positions.

We further investigated the relationship between influencer significance and the feature importance of the cytosines and guanines in CpGs on data set 2, see supplementary **Fig. S4** and **S5**. Similar to the five evaluation metrics, we did not find an obvious relationship between influencers and feature importance.

We studied whether the most significant influencers resulted in higher ARI-all-data scores when the predicted methylation levels on the influencers were used to cluster cells into different cell types. **Fig. 6 C** shows our benchmark results on all of the cells but only considers the top 10 most significant influencers. **Fig. 6 D** illustrates the results on the cells that have at least 16000 influencers detected, which has a Pearson correlation of 0.098 between influencer ranking positions and ARI-all-data scores. In summary, the results from **Fig. 6 C** and **D** do not show a strong correlation or relationship between the influencer ranking positions and ARI scores. The average ranking positions in **Fig. 6 C** and **D** were calculated using avg_inf_rank1. Supplementary **Fig. S6** shows the distributions of the number of influencers with various

avg_inf_rank1 values associated with **Fig. 6 C** and **D**, respectively. Supplementary **Fig. S7** shows the results using avg_inf_rank2 on data set 2, which indicates the same conclusions.

Last but not least, we picked up the influencers that were first-ranked in the influencer lists of all cells and counted the number of occurrences of each promoter as the first-ranked influencer among all cells. The promoters that have the same number of occurrences were used to cluster cells, and the ARI-all-data scores were calculated. Besides the first-ranked influencers, we also conducted the same experiments for the second-ranked, third-ranked, fourth-ranked, and fifth-ranked influencers, see Supplementary **Fig. S8 A-E**. No significant Pearson correlation or relationship was found between the ARI-all-data scores and the number of occurrences. We also picked up the top five most occurred influencers for influencer ranking positions 1 to 93 (the shortest influencer list contained 93 influencers) and used the top five influencers to clustering cells, see Supplementary **Fig. S8 F**, from which the same conclusion was drawn.

3.6 Cell clustering and feature importance for housekeeping genes

Housekeeping genes are the genes that are required for maintaining basic cellular functions, which are essential for the existence of cells. Housekeeping genes usually are considered to be expressed in all cells with stable expression levels. Therefore, if the scHiMe-predictions are accurate, the predicted methylation levels on the promoters of housekeeping genes should be less effective in distinguishing cell types compared to the promoters of baseline genes. **Fig. 7 A** shows the ARI-all-data scores when housekeeping genes and randomly-selected genes were used to classify cells. The mean of ARI scores for using housekeeping genes is 0.128, which is lower than the ARI score for using randomly-selected genes, which is 0.146, with a P-value of 0.0028. **Fig. 7 B** shows the feature importance of cytosines and guanines of CpGs when housekeeping genes and randomly-selected genes are used, which have the means of 1.542×10^{-5} and 1.601×10^{-5}, respectively, with a P-value of 0.0059. These results further proved that the predictions from scHiMe matched biological meanings.

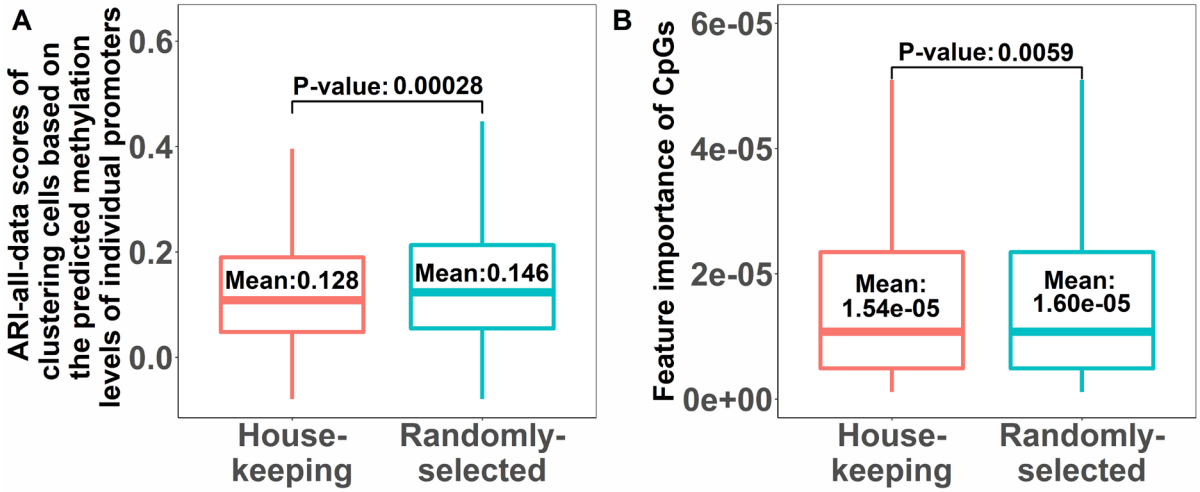


Figure 7. A. ARI scores of using scHiMe-predicted housekeeping genes and randomly-selected genes to classify cells. The P-value between the ARI scores in the two groups is shown in the figure. **B.** Feature importance of the cytosines and guanines of the promoter CpGs for housekeeping genes and randomly-selected genes when their scHiMe-predictions were used to classify cells. The P-value between the ARI scores in the two groups is shown in the figure.

4. Discussion and conclusion

We benchmarked the predictions of scHiMe from multiple perspectives and found that the predictions were accurate and maintained cell-to-cell variability. The DNA sequence used for all of the cells was the same as we used the reference genome to generate related features. Therefore, if we only used DNA sequences as machine learning features, the predictions on all the cells would be the same, which meant the cells would not be correctly classified into cell types. However, our predictions resulted in almost perfect cell-type classifications, which indicated that the single-cell Hi-C data that we used indeed provided useful information for cell-specific characteristics, and the graph transformer that we applied successfully captured the patterns in the graph-structured data.

More importantly, the successful performance of our deep learning algorithm indicated that the single-cell 3D genome structures did have a relationship with single-cell DNA methylations although from this single study we were not certain about which determined which or how one of them influenced the other.

Acknowledgment

Funding: This research was partially supported by the National Institutes of Health grant [1R35GM137974 to Z.W.].

Conflicts of interest: There is no conflict of interest declared.

References

Dekker, J., *et al.* Capturing chromosome conformation. *Science* 2002;295(5558):1306-1311.

Dostie, J., *et al.* Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res* 2006;16(10):1299-1309.

Frankish, A., *et al.* GENCODE reference annotation for the human and mouse genomes. 2019;47(D1):D766-D773.
Hounkpe, B.W., *et al.* HRT Atlas v1.0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets. *Nucleic Acids Res* 2021;49(D1):D947-D955.
Ji, Y., *et al.* DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* 2021.
Lee, D.S., *et al.* Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. *Nat Methods* 2019;16(10):999-1006.
Li, G., *et al.* Joint profiling of DNA methylation and chromatin architecture in single cells. *Nat Methods* 2019;16(10):991-993.
Lieberman-Aiden, E., *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 2009;326(5950):289-293.
Moore, L.D., Le, T. and Fan, G. DNA methylation and its basic function. *Neuropsychopharmacology* 2013;38(1):23-38.
Morone, F. and Makse, H.A. Influence maximization in complex networks through optimal percolation. *Nature* 2015;524(7563):65-68.
Paszke, A., *et al.* Pytorch: An imperative style, high-performance deep learning library. 2019;32:8026-8037.
Pedregosa, F., *et al.* Scikit-learn: Machine learning in Python. 2011;12:2825-2830.
Ramani, V., *et al.* Massively multiplex single-cell Hi-C. *Nat Methods* 2017;14(3):263-266.
Robertson, K.D. DNA methylation and human disease. *Nat Rev Genet* 2005;6(8):597-610.
Shi, Y., *et al.* Masked label prediction: Unified message passing model for semi-supervised classification. 2020.
Simonis, M., *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet* 2006;38(11):1348-1354.