Reichman University

Efi Arazi School of Computer Science
Machine Learning and Data Science M.Sc. Program

Predicting DNA Methylation at individual CpG sites from diluted gene expression data and site context

Dana Keydar, Adi Gotliber

Mentors: The project mentored and supervised by Prof. Zohar Yakhini and Alona Levy-Jurgenson, a PhD student at the Yakhini Research Group.

Abstract - DNA methylation plays a pivotal role in gene expression regulation. In this study, we replicate and extend Levy-Jurgenson et al.'s deep learning model for predicting DNA methylation at specific CpG positions based on a sample's gene expression profile and the surrounding sequence. We meticulously reconstructed the data preparation process and implemented their model in the PyTorch environment. Our analysis successfully reproduces the original model's results, affirming its validity and accuracy while improving accessibility and reproducibility. Furthermore, we examined the model's predictions on diluted gene expression data, exploring various retention levels of gene expression information. Our findings demonstrate that even in scenarios with significantly diluted gene expression data, the model can extract meaningful insights, highlighting its robustness and practical potential for single-cell analysis, and spatial sample utilization.

I **Introduction:**

DNA methylation is a chemical process that modifies DNA in living organisms and can significantly affect gene expression, mostly through the inhibition transcription. In humans, DNA methylation refers to the presence of a methyl group at a defined position of a cytosine and occurs mostly in CpG dinucleotides. It has been particularly shown to affect gene expression in gene promoter regions with relatively dense CpGs, known as CpG islands (CGI). When a large number of proximal CpGs are methylated, the transcription of nearby downstream genes may be inhibited.

DNA methylation has been extensively linked to alterations in gene expression, playing a key role in the manifestation of multiple diseases, such as cancer [2],[3],[4]. Hence, the sequence determinants of methylation and the relationship between methylation and expression are of great interest from a biological perspective.

In the research paper titled "Predicting Methylation from Sequence and Gene Expression Using Deep Learning with Attention" by Levy-Jurgenson et al. [1], the relationship between DNA methylation and gene expression was explored. The authors harnessed the power of deep learning, integrating an attention mechanism, to create a versatile model capable of predicting DNA

methylation patterns at individual CpG positions solely based on gene expression profiles and the contextual sequence surrounding the CpG sites. This approach yielded impressive Spearman correlation and Mean Absolute Error (MAE) results on a diverse set of CpG positions and subjects, and it also unearthed potential associations between methylation activity and specific motifs and genes, such as Nodal and Hand1. Moreover, the integration of attention mechanisms offered a fresh perspective on deriving insights from gene expression data, especially when combined with sequence information.

Although several models have been suggested to support the prediction of methylation status [5],[6],[7],[8], [9], these models were never tested on spatial data or single cell data. The greater vision that forms the context of this project is the extension of these methods to be applicable to such challenging samples. In this report we present methods and results related to a first step in this greater vision. Namely, the application of prediction methods on diluted gene expression data.

The primary objective of this project is to reproduce and extend the findings presented in the research paper titled "Predicting Methylation from Sequence and Gene Expression Using Deep Learning with Attention" by Levy-Jurgenson et al. As single cell and spatial samples are in effect diluted versions of the actual gene expression profile, we focus, in this work, on studying diluted input. We describe the investigation of the performance of MNIST Fashion classifier, on diluted images and then continue to predict DNA methylation at individual CpG sites from diluted gene expression data site context.

An additional important contribution of this project is a PyTorch version of the training and date preparation environment of Levy-Jurgenson et al, which was originally implemented in TensorFlow. The related code is available at this git repository.

II MNIST Fashion classification on diluted images

We implemented dilution tests to assess the potential effectiveness of our methylation prediction approach using diluted biological data. To elucidate the concept of dilution test validity, we initiated the process by evaluating our method on the well-established MNIST FASHION open dataset. Our initial approach involved training a basic classifier using the original images. Subsequently, we ran the inference and evaluated the classifier's performance at 10 different dilution working points, ranging from 10% to 100% retention level.

Each dilution level we randomly altered the test images so that the corresponding fraction of the pixels were turned black. Fig 1 describes the visual transformation of the test set images after the dilution process at each working point, where the retention level is the percentage of the preserved pixels. Fig 1 depicts an example of ankle boot's classification that remained correct up to a 60% retention level.

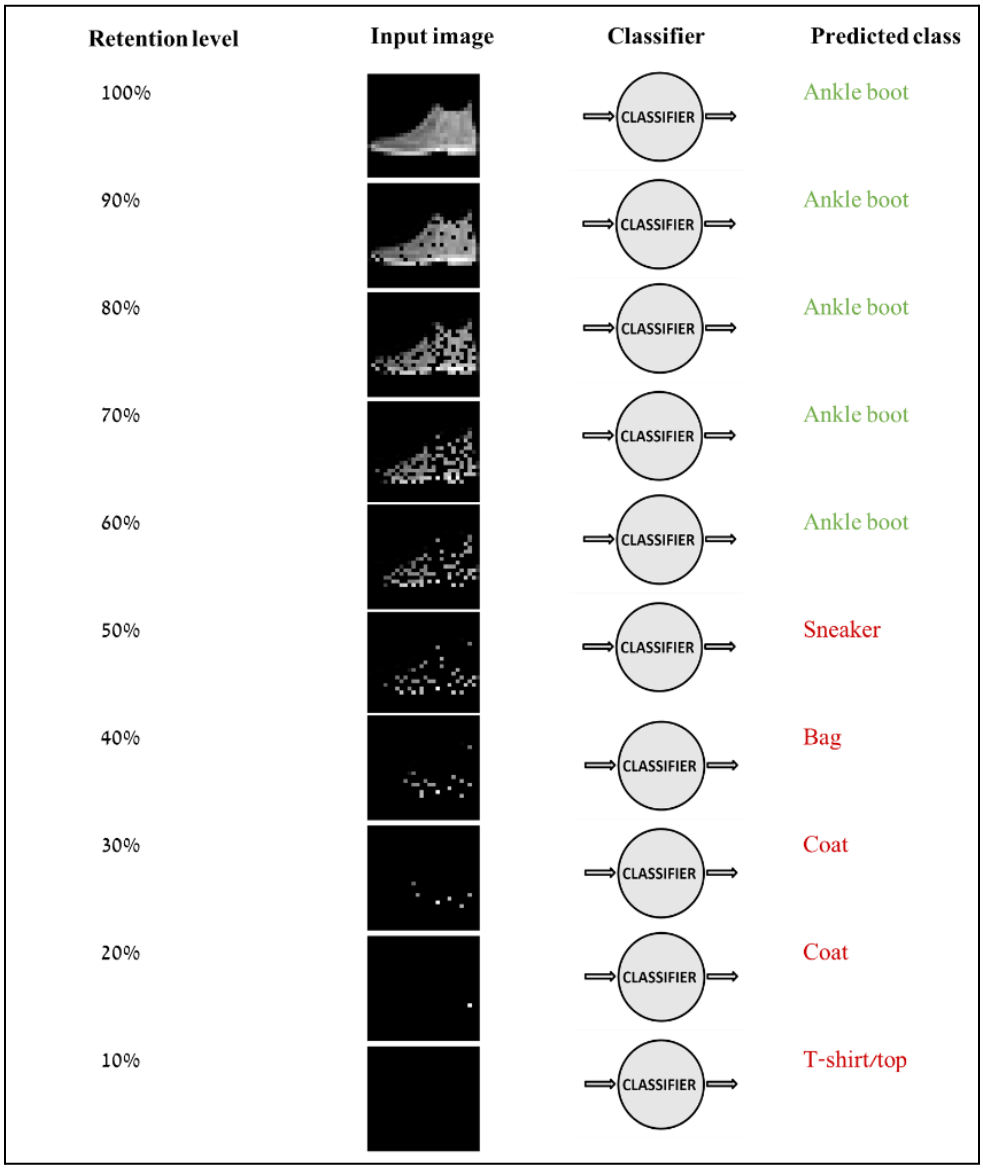

Figure 1: Example of Mnist fashion classification evaluation after the dilution process.

To evaluate the model's performance at each dilution level, we employed the classification accuracy metric, calculated as the ratio between the correct classifications and the total predictions. The accuracy trends across various dilution levels are depicted in Figure 2 and Figure 3. As anticipated, the classification accuracy diminishes with increasing dilution levels. Notably, at a dilution level of 90%, where only 10% of the pixels are retained, the classification accuracy drops to 10%, similar to random classification as the MNIST-Fashion data addresses 10 different classes.

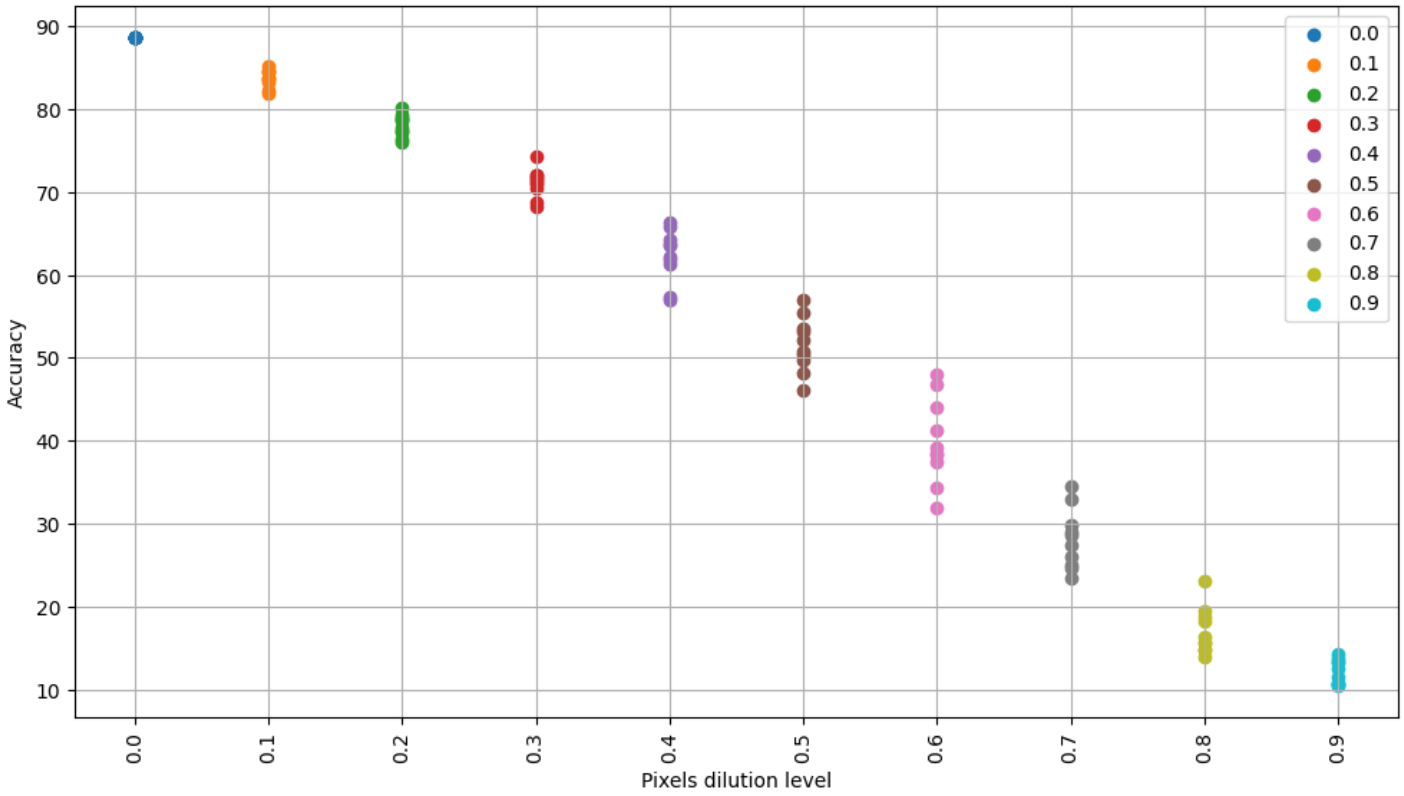

Figure 2: Scatter plot comparing the model accuracy in different pixels retention levels, where for each retention level we sampled the test data and measured the accuracy 10 times. The accuracy is calculated as the number of correct class predictions out of the total number of tested examples.

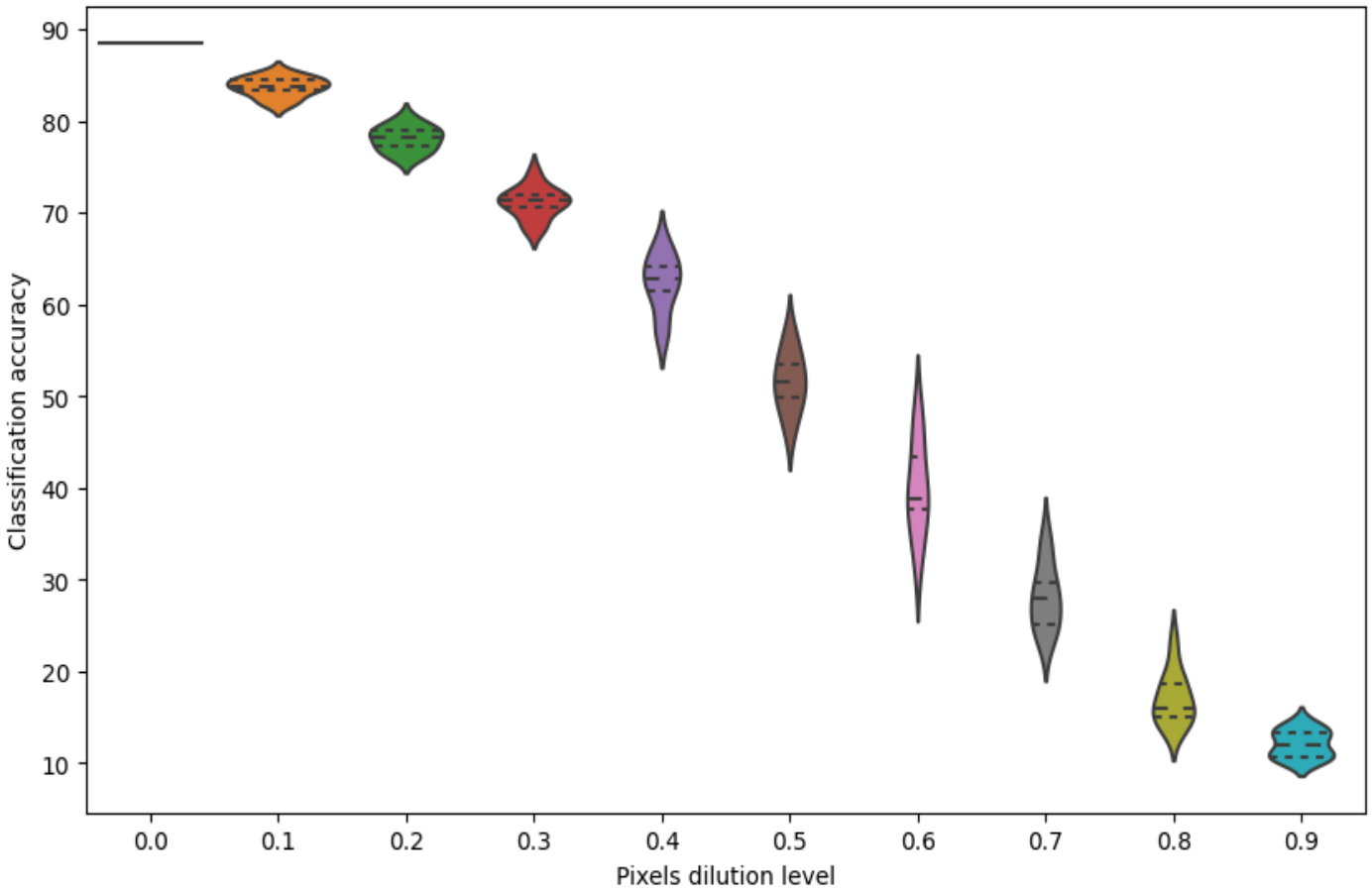

Figure 3: Violin plot comparing the model accuracy distributions in different pixels retention levels, where for each retention level we sampled the test data and measured the accuracy 10 times. The accuracy is calculated as the number of correct class predictions out of the total number of tested examples.

As evident from the distributions displayed in Figure 3, we observe that at the extreme ends of the retention spectrum–specifically, for the lower dilution levels of 0.1, 0.2, as well as the high dilution level of 0.9–there is relatively less variation between different samples belonging to the same dilution category. In contrast, for the middle dilution levels, we notice a higher degree of variability, as the randomly preserved pixels chosen have a more pronounced impact on the model's performance.

For a detailed implementation of the MNIST FASHION classifier, and the implementation of the dilution tests and analyses, please refer to this notebook.

Ⅲ Methylation:

In this main part of the project, we aimed to replicate and build upon the research conducted by Levy-Jurgenson et al [1]. To facilitate the reproducibility and retraining of their model, we meticulously recreated the data preparation process and further implemented the model using PyTorch. We validated the original model's results using the re-prepared data and then assessed the results derived from the PyTorch implementation. Finally, we experimented with the model performance using diluted test data, while the model trained using the non-diluted data.

a. Data preparation

We reconstructed the datasets required for training and validating our model, adhering closely to the methodology outlined in the original paper [1]. The core datasets utilized include gene expression and methylation level data from patients with BRCA and LUAD conditions, coupled with essential human genome sequence information, CpG locations, and gene locations data. Leveraging this raw information, we generated four data files, their structure is demonstrated in the image below:

1. Sequences Centered around Each CpG Site
2. Distances between CpG Sites and Genes
3. Gene Expression Per Sample (Subject)
4. Methylation Level Data Per Sample and CpG Site (which are the labels for our learning task)

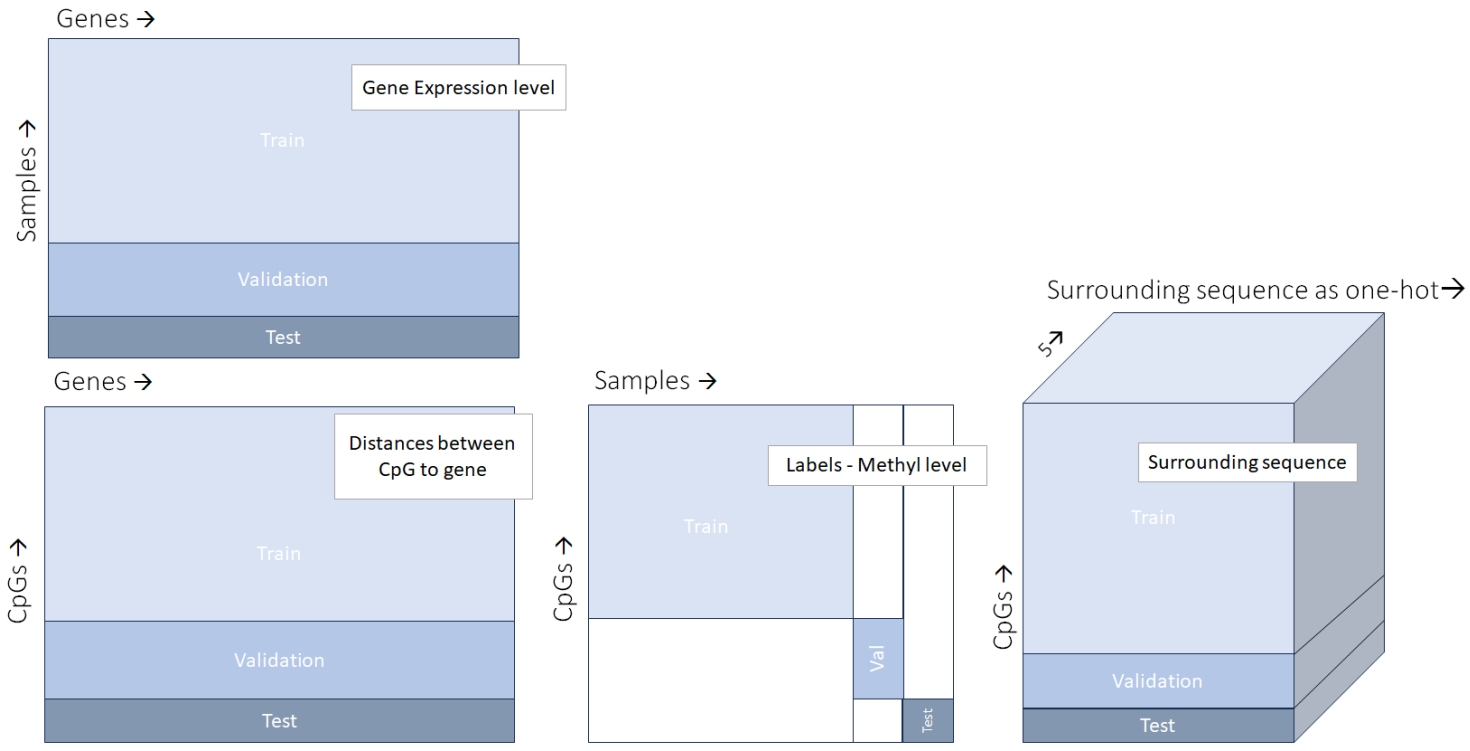

Figure 4: Data structure of the four data parts, including the train validation and test separation illustration. The test set consists of subjects/samples and CpGs that have not been included in the train or validation set. The association of each subject and CpG to either one of the train, validation or test sets was random.

We defined a single training example,to represent one sample (subject) and one CpG. The data has the same structure as in the original paper [1]. It contains the following components:

1. The subject's gene expression vector e of all available genes, where each entry, ei, represents the expression level of a gene gi.
2. The sequence surrounding the CpG of interest, represented as a one-hot matrix.
3. A vector d, for each CpG of interest, where di is computed based on the distance, in base-pairs, between gi and the CpG of interest. Specifically, a gene residing within the first 2,000 base pairs received a value of 1, the next 2,000 a value of 0.5 and so on until the last bucket of 2,000 was given a value of 0.5^9. Beyond this point di was set to 0. For genes residing on a different chromosome this value was also set to 0.

The full data preparation process and the corresponding code, is available in the CH3 Data Preparation Documentation.pdf

b. Methods

Our project involves implementing the model in PyTorch and reproducing the results achieved in the paper [1]. To confirm the validity of each step, we first replicated the original model's performance using our dataset and the original TensorFlow implementation. Then, we proceeded to train and test our PyTorch model using our dataset, successfully reproducing the results at this stage as well. To access the model code and instructions with the training and validation environment visit the Model Code Repository.

For the dilution data experiments, to simulate gene expression data that is typical for spatial samples, we introduced a random dilution process to the gene expression test data, wherein the strength of gene expression per sample, serves as the probability of a gene's inclusion in the sample's diluted data. The selection process is as follows: The retention level required, acts as the number of times we conduct the random selection of a gene to be included in the gene expression list for a specific sample, the selection function can choose a certain gene more then once, as we sample with replacement. The probability for a gene to be selected is the gene expression fraction out of the gene expression sum of the sample. We conducted this procedure for multiple retention levels: 20,000, 15,000, 10,000, 5000, 1000, and 100 genes. In Figure 5, we visually depict the persistence of gene expression information across different retention levels. The figure clearly shows that even with a reduction to 10,000 or 5,000 selected genes, the essential information and prominent patterns remain discernible, although with somewhat reduced intensity.

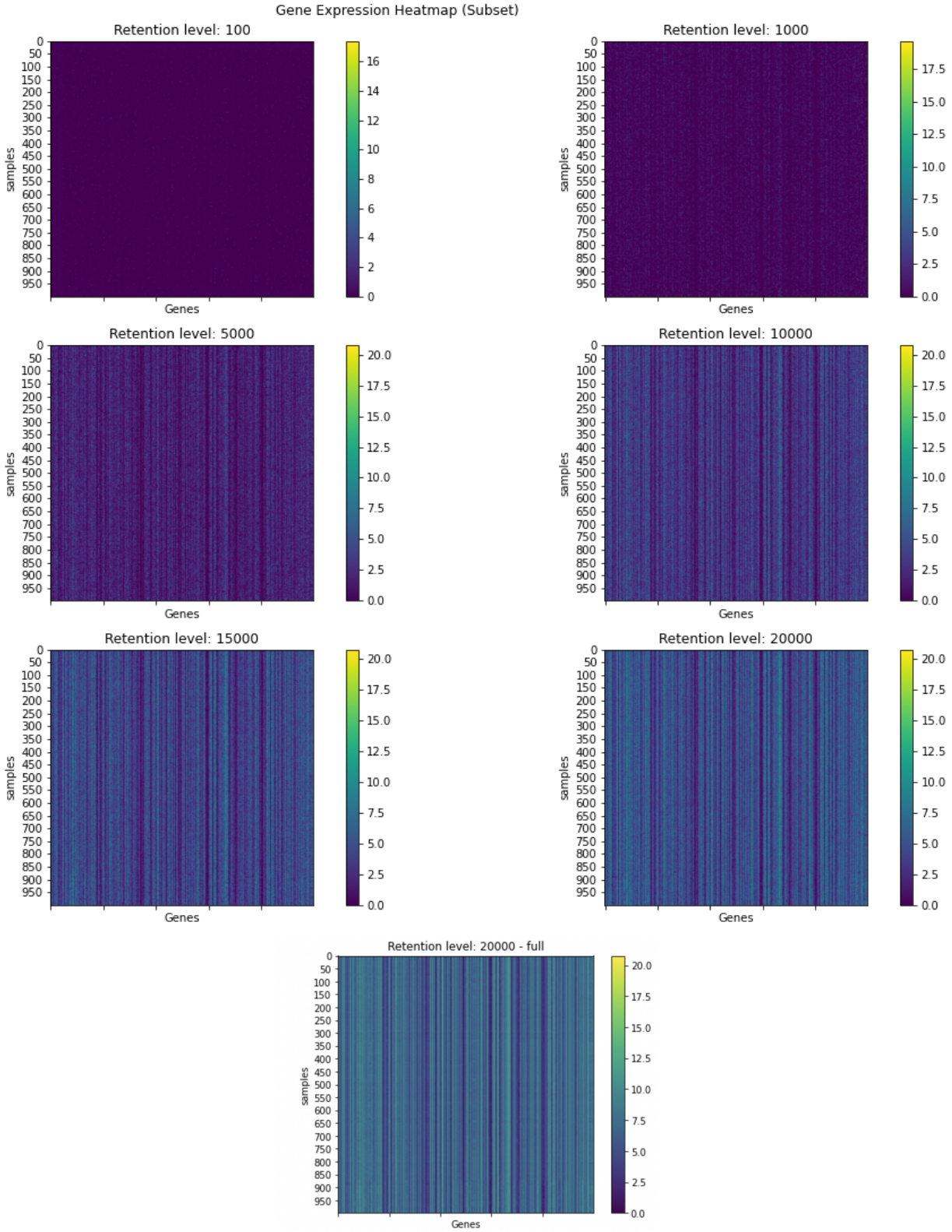

Fig 5: Visualization of a gene expression subset data, in different retention levels. The columns represent an arbitrary selection of samples, and the rows represent an arbitrary selection of genes. Each graph depicts the gene expression data at the indicated retention level. The colors represent the gene expression levels after the dilution process.

c. Results

Predicting Methylation Levels

We evaluated both models on held-out test sets in which both CpGs and Samples are disjoint to the ones included in the training. The reported results in [1] are: MAE of 0.14 and 0.8 Spearman correlation. In our reproduction experiments (with the new data). While utilizing the original Tensorflow model, we achieved MAE of 0.173 and 0.740 Spearman correlation, and with our new PyTorch model MAE of 0.167 and 0.742 Spearman correlation.

TABLE I

Methylation level prediction accuracy

	MAE- Mean Absolute Error	Spearman Correlation
Levy-Jurgenson et al- TF Model1 results as reported in the paper.	0.14	0.8
Levy-Jurgenson et al- TF model1 results with this project's reconstructed data.	0.173	0.740
PyTorch model1 with this project reconstructed data	0.167	0.742

Training Graphs

The following graphs demonstrate the training and validation performance during the training process. The experiments were monitored via clear ml, and the full details can be reviewed at the project dashboard. The training metrics can be observed in Figure 6 and Figure 7.

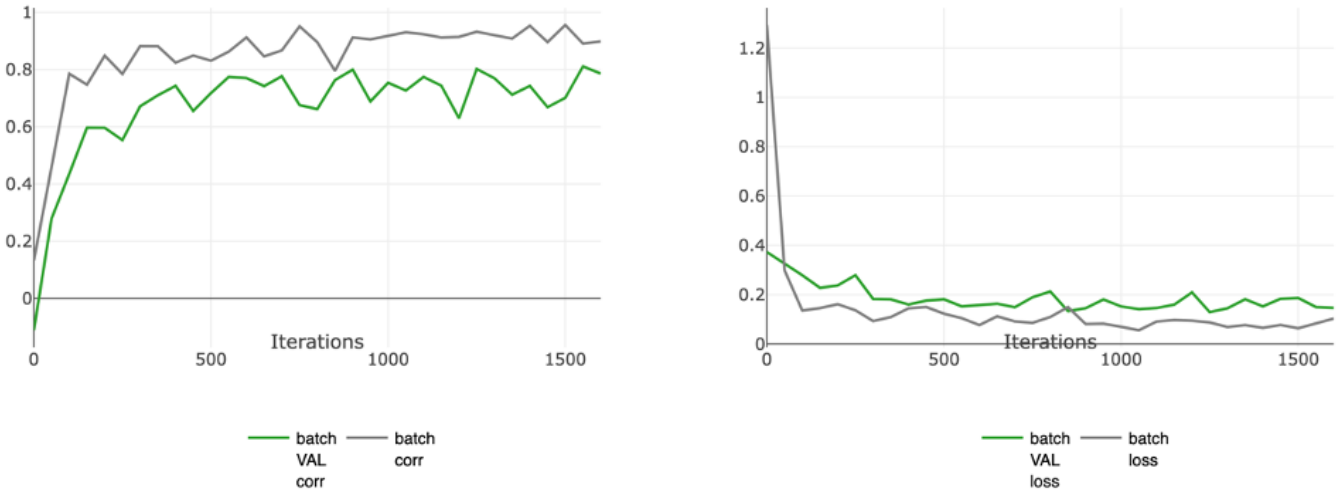

Figure 6: Training graphs of the original TensorFlow model, re-trained with the reconstructed data. The metrics presented in the graphs are the Mean Absolute Error (MAE) loss (Left) and Spearman Correlation (Right).

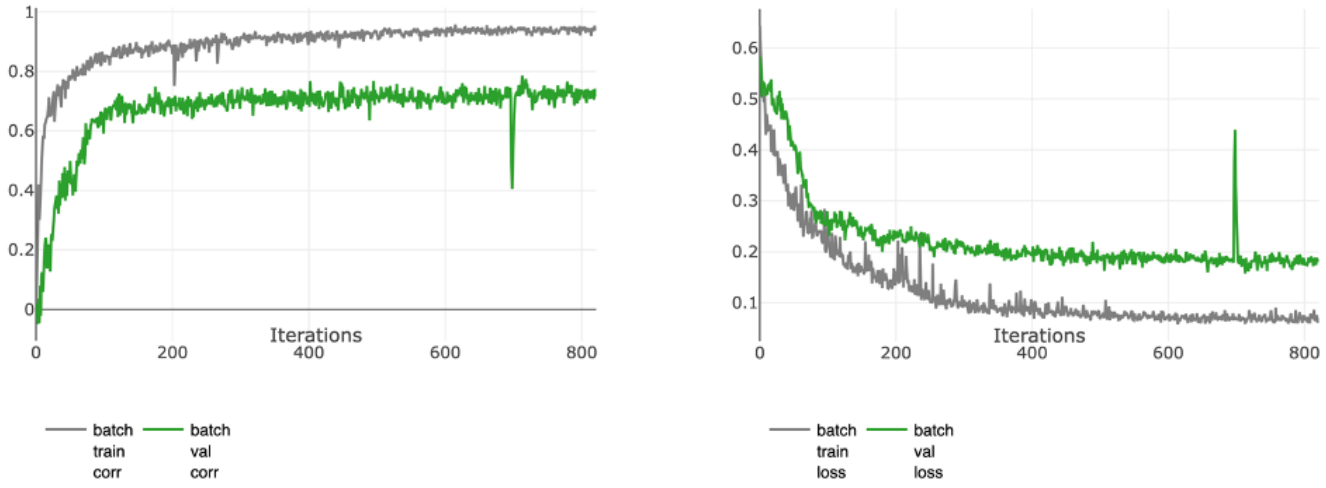

Figure 7: Training graphs of the PyTorch model, trained with the reconstructed data. The metrics presented in the graphs are the Mean Absolute Error (MAE) loss (Left) and Spearman Correlation (Right).

Predicting Methylation Levels For Diluted Data

We conducted the dilution tests at various retention levels, as indicated in the methods: 100, 1000, 5000, 10,000, 15,000, and 20,000 selected genes. Notably, the 20,000-genes working point is particularly interesting because, despite having a total of 20,157 genes available, the dilution test setting allows a random selection of genes based on their strength. As a result, fewer than 20,000 genes are actually chosen. This working point provides an interesting opportunity for comparison with the original full gene expression dataset.

At each working point we performed 10 iterations, and derived the model accuracy statistics based on these repeated experiments. As anticipated, better performance was achieved with higher retention of gene expression data and the performance decreased as we approached complete removal of gene expressions.

At the 20,000 and 15,000 gene retention levels, the model obtained identical performance to that observed with the original undiluted dataset, yielding an MAE of 0.17 and a Spearman correlation of 0.74. For the 10,000 and 5,000 gene retention levels, there was a slight reduction in data quality, although not statistically significant, resulting in MAE values of 0.73 and 0.72, respectively. At lower retention levels, specifically with 1,000 and 100 chosen genes, a more substantial reduction in data quality was evident, resulting in Spearman correlations of 0.68 and 0.66, respectively. Notably, even at the lowest retention levels of 1,000 and 100, the results are informative, and significantly far from random.

TABLE 2

Model's prediction accuracy of methylation levels, for diluted test data

Retention Level	Avg MAE	Avg Spearman Correlation
100	0.23	0.66
1000	0.22	0.68
5000	0.19	0.72
10000	0.18	0.73
15000	0.17	0.74
20000	0.17	0.74

In the figures below, we depict both the Spearman correlation metric and the MAE metric (loss) based on 10 experiments, illustrating the performance variation for each retention level, with data derived from 10 samplings.

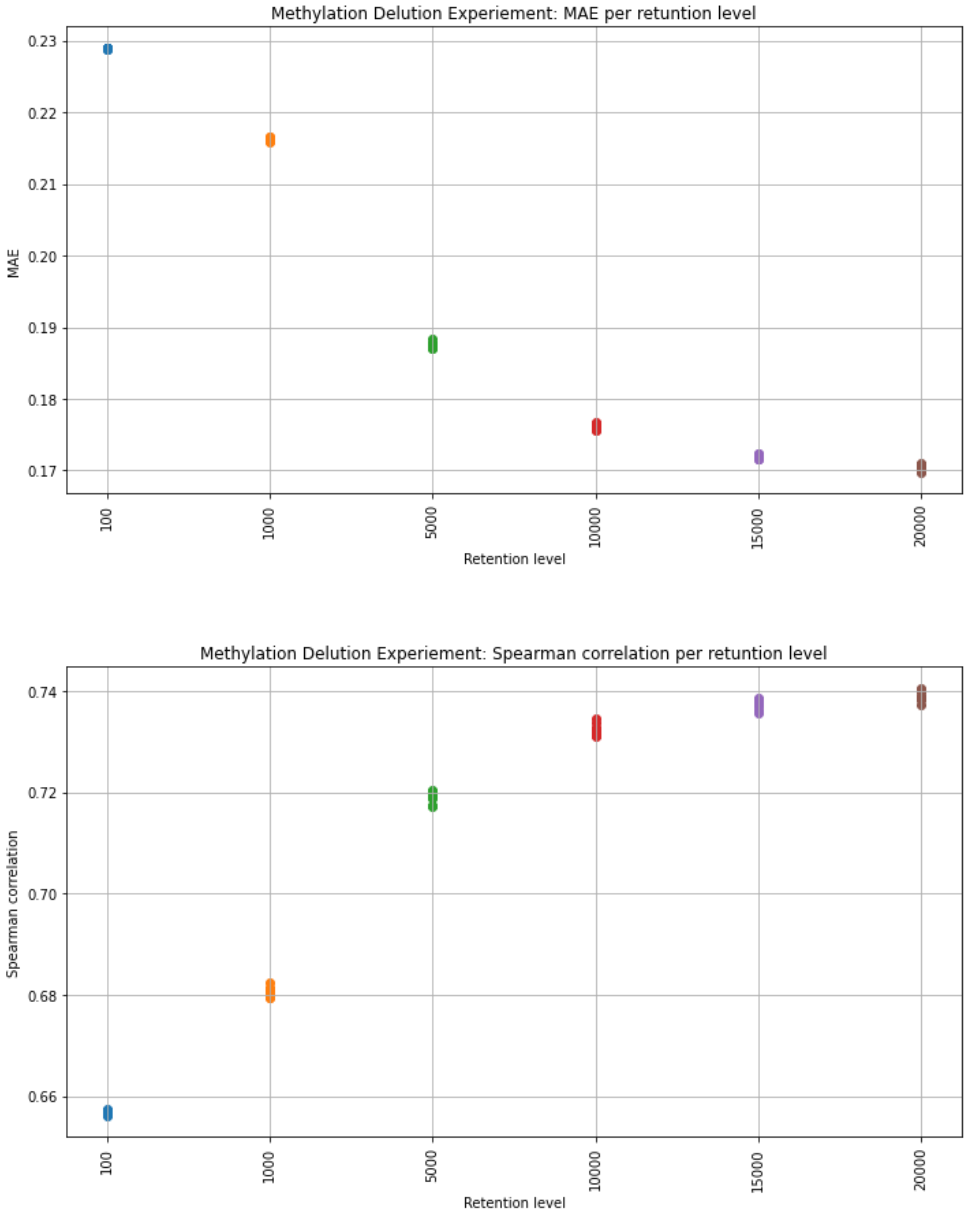

Fig 8: MAE and Spearman correlation for different retention levels. For each retention level the data was sampled 10 times.

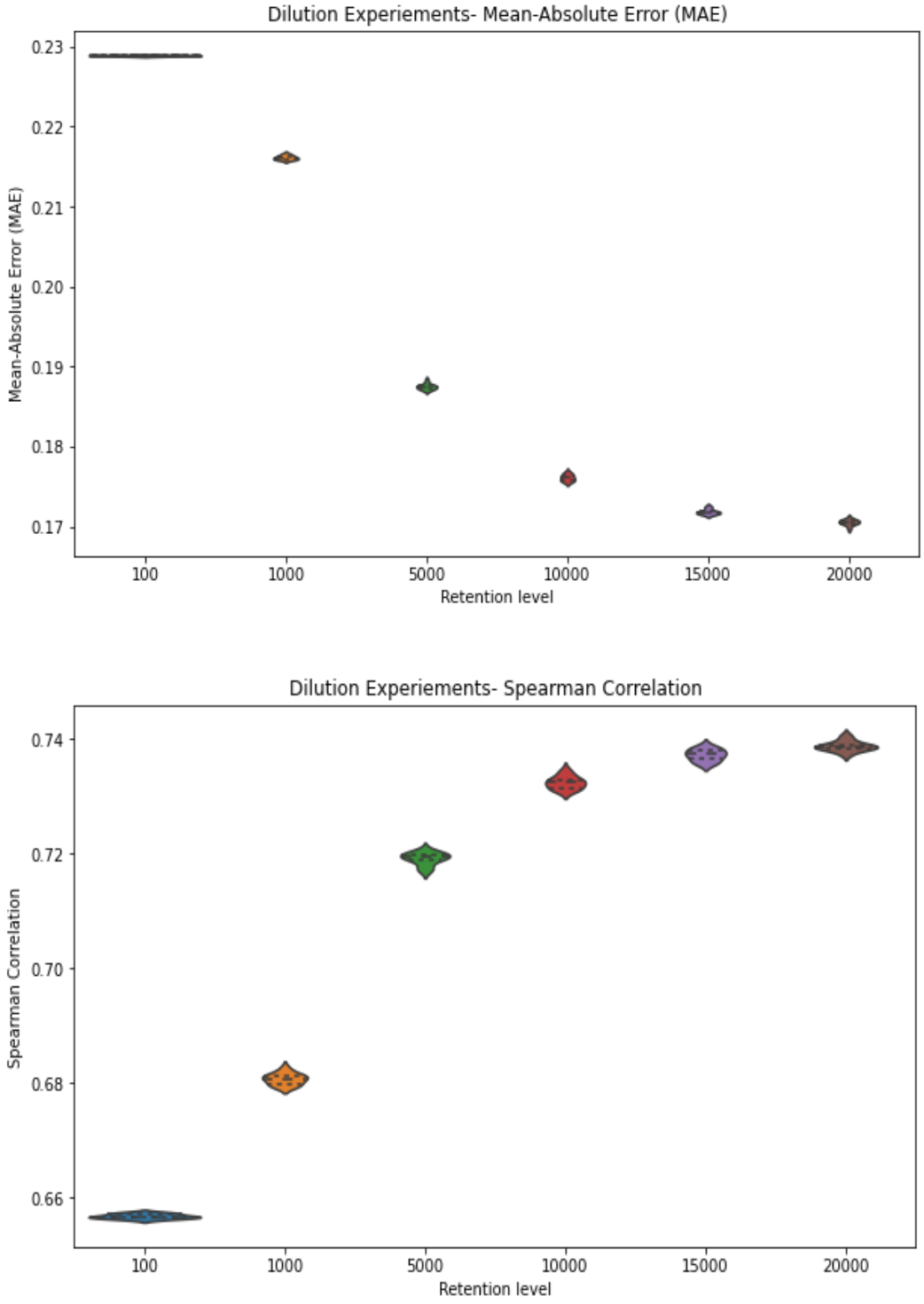

Figure 9: Violin plots of the evaluation metrics distributions, MAE at the top graph and Spearman correlation at the bottom graph.

As noted earlier in the violin plots of the MNIST Fashion dilution test, Figures 8 and 9 further illustrate that at the middle retention levels, the model's results exhibit higher variance, while for the lowest and highest retention levels, the distribution appears notably more compressed. For more insight into the dilution test code please explore the Dilution Test Code and Report.

IV Discussion

The experiments conducted using diluted gene expression data clearly illustrate the remarkable potential for predictability even when dealing with partial or diluted datasets. Specifically, when transitioning from the complete gene expression set to a subset of chosen 5,000 genes, the effect on the model's prediction capability remained relatively modest.The Spearman correlation only dipped slightly from 0.74 to 0.72. Furthermore, the mean absolute error exhibited a negligible increase of just 9%, rising from 0.17 to 0.19.

However, as we delve into lower retention levels, such as 100 and 1,000, we start to observe more pronounced performance declines. Yet, it's worth highlighting that even at these reduced retention levels, predictability remains informative and significantly deviates from randomness. For instance, at a retention level of only 100 chosen genes, the mean absolute error increases to 0.23, and the Spearman correlation decreases to 0.66. These results underscore the model's ability to extract meaningful insights from highly diluted data, reaffirming its robustness and potential for practical application.In particular, our results lay solid foundation to continued work with single cell and spatial transcriptomics data.

IV Future work

In order to advance our research and enhance the robustness of our proposed methodology, several avenues for future work emerge. First and foremost, we believe it is essential to explore the potential benefits of scaling the gene expression values. Currently, the gene expression data values range between 0 to 14, and scaling techniques such as normalization or standardization may be applied to ensure consistency and facilitate more effective feature extraction and modeling.

Furthermore, upon a thorough examination of the results involving diluted data, specifically retention level 100, it becomes evident that the model's performance remains robust. Although the mean squared error (MSE) increases to 0.23 and the Spearman correlation decreases to 0.66, the model consistently demonstrates significant outperformance of random chance. This observation underscores the significant role that sequence data plays in the model's learning process. Nonetheless, it is still clear that the model indeed acquires essential information from gene expression, exerting a substantial influence on its performance.

Therefore, for future research, we can explore the influence of each data component.Firstly, we can reduce the model's reliance on sequence data by training it exclusively using the distance

data between CpG sites and genes, along with gene expression information, without involving sequence data. Alternatively, we could experiment with training the model using a weighted loss function that balances the sequence component and the gene-expression component. Such an effort would enable a more profound understanding of the extent to which the model learns from each data type, and provide greater control over its learning process. By exploring this dimension of data learnability, we can enhance our methodology and render it adaptably to a variety of biological scenarios.

Additionally, future investigations should focus on optimizing our training dataset by exclusively utilizing gene expression information for genes that exist within the distance matrix. This approach has the potential to streamline the learning process, reduce noise, and enhance the interpretability of the model's outcomes by prioritizing biologically relevant features.

A paramount objective in our future research endeavors is to expand the scope and applicability of our model by using real spatial data. To date, our work has relied on diluted test datasets, but a significant future step should focus on using real spatial data as test data and even trying to train the model on real spatial data.

References:

1. Levy-Jurgenson, A., Tekpli, X., Kristensen, V.N., Yakhini, Z. (2019). Predicting Methylation from Sequence and Gene Expression Using Deep Learning with Attention. In: Holmes, I., Martín-Vide, C., Vega-Rodríguez, M. (eds) Algorithms for Computational Biology. AlCoB 2019. Lecture Notes in Computer Science(), vol 11488. Springer, Cham. https://doi.org/10.1007/978-3-030-18174-1_13

2. Monique GP van der Wijst, Amanda Y van Tilburg, Marcel HJ Ruiters, and Marianne G Rots. Experimental mitochondria-targeted dna methylation identifies gpc methylation, not cpg methylation, as potential regulator of mitochondrial gene expression. Scientific Reports, 7(1):177, 2017.

3. Deborah Nejman et al. Molecular rules governing de novo methylation in cancer. Cancer research, 74(5):1475–1483, 2014

4. Catherine S Grasso, Marios Giannakis, Daniel K Wells, Tsuyoshi Hamada, Xinmeng Jasmine Mu, Michael Quist, Jonathan A Nowak,Reiko Nishihara, Zhi Rong Qian, Kentaro Inamura, et al. Genetic mechanisms of immune evasion in colorectal cancer. Cancer discovery, 2018.

5. Manoj Bhasin et al. Prediction of methylated cpgs in dna sequences using a support vector machine. FEBS letters, 579(20):4302–4308, 2005.

6. Rajdeep Das et al. Computational prediction of methylation status in human genomic sequences. Proceedings of the National Academy of Sciences, 103(28):10713–10716, 2006.

7. Weiwei Zhang et al. Predicting genome-wide dna methylation using methylation marks, genomic position, and dna regulatory elements. Genome biology, 16(1):14, 2015

8. Yiheng Wang et al. Predicting dna methylation state of cpg dinucleotide using genome topological features and deep networks. Scientific reports, 6:19598, 2016.
9. Baoshan Ma et al. Predicting dna methylation level across human tissues. Nucleic acids research, 42(6):3515–3528, 2014.