Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Yang Shen Date

Methylation Imputation from HM450K Array to EPIC Array with Autoencoder and Nonnegative Matrix Factorization

By

Yang Shen

Master of Public Health

Department of Biostatistics and Bioinformatics

Zhaohui (Steve) Qin, PhD
(Thesis Advisor)

Anke Huels, PhD
(Reader)

Methylation Imputation from HM450K Array to EPIC Array with Autoencoder and Nonnegative Matrix Factorization

By

Yang Shen

B. S., Shandong University, 2020

Thesis Committee Chair: Zhaohui (Steve) Qin, PhD

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Public Health

in Biostatistics

2023

Abstract

Methylation Imputation from HM450K Array to EPIC Array
with Autoencoder and Nonnegative Matrix Factorization

By Yang Shen

DNA methylation is an essential epigenetic modification that plays a crucial role in gene expression regulation and cellular differentiation. DNA methylation profiling has been widely used in research to determine the development of various human diseases, including cancer, cardiovascular disease, and neurological disorders. The HumanMethylation450K (HM450K) arrays and the Enhanced DNA Methylation Profiling (EPIC) arrays are two commonly used high-throughput technologies that enable genome-wide DNA methylation profiling. The HM450K array covers approximately 450,000 CpG sites, while the EPIC array covers more than 850,000 CpG sites, and there's an overlap of around 440,000 CpG sites between the two arrays. In this study, our goal is to impute methylation levels from the HM450K array to the EPIC array to circumvent the need for expensive re-measurement using the EPIC array when HM450K array data is already available. Convolutional autoencoders and nonnegative matrix factorization (NMF) are both machine-learning techniques that are commonly used in the analysis of large-scale genomic data. Our approach involved using a convolutional autoencoder and an NMF model to capture the latent structure in the DNA methylation data and generate imputed values for all CpG sites in the EPIC arrays. We mainly focused on chromosome 18 to simplify our model. The overall RMSE was 0.0196, which was better than 0.04 from a simple linear regression model with nearby CpG sites. Our model was highly adaptable to other chromosomes and could easily adjust the dimensions of the results obtained from autoencoders to accommodate different chromosome sizes.

Methylation Imputation from HM450K Array to EPIC Array with Autoencoder and Nonnegative Matrix Factorization

By

Yang Shen

B. S., Shandong University, 2020

Thesis Committee Chair: Zhaohui (Steve) Qin, PhD

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Public Health

in Biostatistics

2023

Table of contents

1. Introduction

DNA methylation is an essential epigenetic modification that plays a crucial role in gene expression regulation and cellular differentiation. In recent years, advances in DNA methylation profiling technologies have led to significant progress in understanding of the epigenetic basis of human diseases. DNA methylation profiling has been widely used in research to determine the development of various human diseases, including cancer[1], cardiovascular disease[2], and neurological disorders[3]. These technologies have enabled researchers to identify specific DNA methylation patterns that are associated with disease onset and progression, thereby providing valuable insights into disease mechanisms and potential targets for therapeutic intervention[4]. Additionally, DNA methylation profiling can also be used for disease diagnosis and prognosis, and monitoring treatment response. As such, it has become an important tool in the field of personalized medicine[5,6].

The HumanMethylation450K (HM450K) arrays[7] and the Enhanced DNA Methylation Profiling (EPIC) arrays[8] are two commonly used high-throughput technologies that enable genome-wide DNA methylation profiling. These technologies provide a comprehensive view of DNA methylation patterns across the entire genome, allowing for the identification of differentially methylated regions. The HM450K array covers approximately 450,000 CpG sites, while the EPIC array covers more than 850,000 CpG sites, providing a larger coverage and resolution of the epigenome, and there's an overlap of around 440,000 CpG sites between the two arrays. With the increasing availability of these high-throughput technologies, there is a growing need for efficient methods to analyze and integrate the data obtained from different platforms to gain a deeper understanding of the epigenetic mechanisms underlying disease.

Our goal is to impute methylation levels from the HM450K array to the EPIC array to circumvent

the need for expensive re-measurement using the EPIC array when HM450K array data is already available. There have been several studies focused on this. The penalized functional regression (PFR) method[9] was developed for imputation from HM27K to the HM450K array. The CpG impUtation Ensemble (CUE)[10] used both statistical and machine learning methods, to impute from the HM450K to the EPIC array. However, it can only impute 85.4% EPIC only CpG sites in its case study. These methods only focused on the EPIC only CpG sites. Our approach involved using a convolutional autoencoder[11] and an nonnegative matrix factorization[12] to capture the latent structure in the DNA methylation data and generate imputed values for all CpG sites in the EPIC arrays.

2. Method

2.1 Data Pre-processing

We used a total of 172 patients who had both HM450K and EPIC data. In order to avoid potential biases, we removed duplicate observations from the same patient and randomly assigned the remaining 172 observations into a training dataset consisting of 155 patients and a testing dataset consisting of 17 patients. To simplify our models and reduce the cost of training models, all experiments were based on chromosome 18. We also removed CpG sites with partially missing methylation levels, resulting in 5,594 CpG sites on chromosome 18 from the HM450K dataset and 13,173 CpG sites from the EPIC dataset, and there's an overlap of 4,933 CpG sites between the two arrays. To facilitate analysis, we standardized the methylation levels of each CpG site by subtracting the minimum value and dividing by the range, transforming the beta values to the whole standardized scale ranging from 0 to 1. To capture local information, we re-sorted all CpG

sites according to their locations on chromosome 18.

2.2 Dimension reduction: Convolutional Autoencoder

Autoencoder is a type of neural network architecture that can be used for unsupervised learning. An autoencoder consists of an encoder and a decoder. The encoder first maps the input data into the embedding space. The decoder returns the embedding representation back to the input data by reconstruction. By minimizing the difference between the input data and the reconstructed output, the autoencoder method effectively learns the underlying structure and patterns of the input data, and creates a more compressed representation that captures the most important information. This compressed representation can then be used as a feature set for downstream tasks.

Convolutional autoencoder is a variant of autoencoder. Similar to the traditional autoencoder, it consists of an encoder network and a decoder network. However, the encoder and decoder networks of a convolutional autoencoder are composed of convolutional layers, which are specifically designed to identify spatial patterns in the data.

In this study, we utilized convolutional autoencoders to learn compressed representations of DNA methylation data from the HM450K and EPIC arrays. Convolution layers in the autoencoder models were used to capture the spatial relationships between adjacent CpG sites and extract the important features of the input data. By reducing the dimensionality of the original data, the autoencoder models were able to compress the input information into a more compact representation, which also helped to mitigate the effects of noise in the input data. The use of separate autoencoder models for each array allowed us to capture the unique features of the methylation data in each array. The architecture of our autoencoder model was designed to achieve a better performance for our DNA methylation data. Specifically, the encoder part of the model contained a single convolutional layer followed by a maxpooling layer, which allowed for the

efficient extraction of features from the input data. This was then followed by four fully connected layers, with the embedding layer being the final layer of the encoder. The decoder part of the model consisted of three fully connected layers, one maxunpooling layer, and a convolutional layer. The activation function used in most layers was ReLU, except for the identity function used in the embedding layer and the Sigmoid function used in the convolutional layer of the decoder. These functions are defined as:

$$\mathrm{ReLU}(x) = \max\{0, x\} = \begin{cases} x, & \text{if } x > 0 \\ 0, & \text{if } x \leq 0 \end{cases},$$

$$\mathrm{Identity}(x) = x,$$

$$\mathrm{Sigmoid}(x) = \frac{1}{1 + \exp(-x)}$$

As shown in Figure 1, this architecture allowed us to effectively reduce the dimensionality of the data while preserving important features.

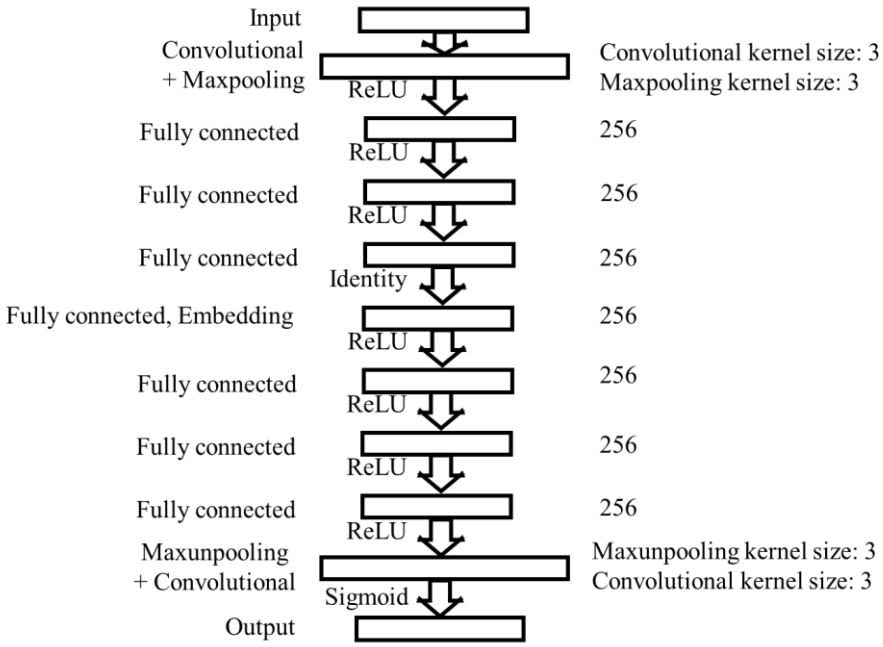

Figure 1: Architecture of our convolutional autoencoder models. The convolutional autoencoder models for both arrays had very similar structures, with the only difference being the dimensions of the input and output layers.

2.3 Nonnegative Matrix Factorization

After dimension reduction, the next goal was to generate the EPIC data from HM450K data, so that we could obtain the original data of EPIC through the decoder part of the autoencoder for EPIC array. Nonnegative matrix factorization (NMF) was widely used as a linear, non-negative approximate data representation. The NMF algorithm factorizes the input data matrix into two non-negative matrices, one representing the basis and the other the coefficients, such that their product approximates the input data matrix. NMF can be defined as:

$$\mathbf{V} \approx \mathbf{H} \cdot \mathbf{W}$$

where **V** is the data matrix needed to be factorized, **H** is the matrix describing the information from each observation, and **W** is the matrix containing the basis of the data matrix or the latent structure of the data. Specifically, each column of **V** represents an observation, each row of **H** represents the contribution of each basis component to the observation, and each column of **W** represents the weight of each basis component across all observations.

To infer the latent structure in and between two reduced-dimensional DNA methylation datasets, we employed NMF as a model. In training dataset, we concatenated the HM450K data, as matrix $\mathbf{V}_1$, and the EPIC data, as matrix $\mathbf{V}_2$, along the columns to obtain a larger matrix as **V** for factorization during training. After NMF, the **H** matrix of the relevant observation information and the **W** matrix of the latent structure and the information of methylation levels in HM450K data, as matrix $\mathbf{W}_1$, and the EPIC data, as matrix $\mathbf{W}_2$, were generated. During the training process, we used the mean square error (MSE) as the loss function to minimize the reconstruction error between **V** and the product of **H** and **W**. In the testing dataset, only HM450K data was used as matrix $\mathbf{V}_{01}$ to be factorized into $\mathbf{H}_0 \cdot \mathbf{W}_1$ where matrix $\mathbf{W}_1$ was the matrix from training process and was fixed during the testing step. Subsequently, the new EPIC data $\mathbf{V}_{02}$ was generated as the product of

matrix $\mathbf{H}_0$ and matrix $\mathbf{W}_2$. During the testing step, we computed the loss as the MSE between the original EPIC data and the newly generated EPIC data $\mathbf{V}_{02}$. Figure 2 showed the architecture of our NMF model.

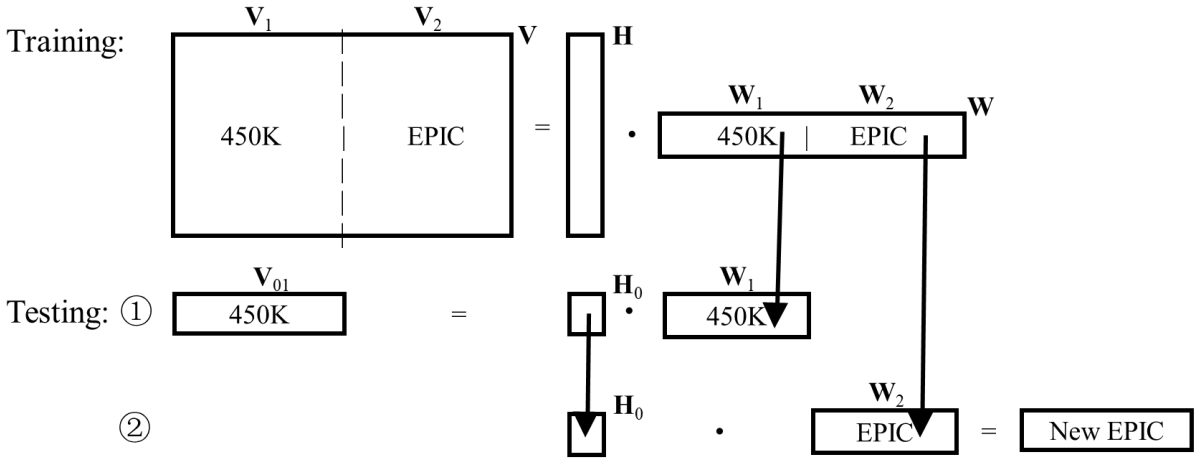

Figure 2: Architecture of our NMF model

The MSE loss function is defined as:

$$\mathrm{MSE}(\vec{x}, \vec{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2$$

3. Results

After data cleaning, there were 5,594 CpG sites in the HM450K dataset and 13,173 CpG sites in the EPIC data (on chromosome 18). Our main goal in the autoencoder step was to achieve accurate reconstruction from the output with a low MSE loss. As depicted in Figure 1, we chose uniform kernel sizes of 3 for the convolutional layer, maxpooling layer, and maxunpooling layer. We set the dimensions of the fully connected layers at 256, as having too many parameters would require more storage and reduce the speed. To avoid the overfitting during the training process, we split the dataset into the training and testing as specified before. During the training process, we used the MSE as the loss function to minimize the reconstruction error between input and output. We

set the maximum epoch to 50. Eventually, we achieved the best MSE loss of 0.0066 for the testing dataset of the HM450K array and 0.0084 for the testing dataset of the EPIC array.

From the embedding layer we obtained the reduced-dimensional DNA methylation datasets. Then we put them into the NMF model with MSE as the loss function. We experimented with different ranks in the NMF with a maximum epoch to 100. The rank parameter in NMF determines the number of components in the factorization. A higher rank can result in better accuracy, but it can also lead to overfitting. We observed the MSE loss for the reduced-dimensional data of the testing dataset between the HM450K array and the EPIC array. Ultimately, we found that a rank of 25 yielded the best results, with an initial MSE loss of 0.1611 being reduced to 0.0018.

Furthermore, once we generated the new testing reduced-dimensional data of the EPIC array, we could feed it back to the embedding layer in the convolutional autoencoder of EPIC array and, subsequently, got the results from the output layer. These predicted results represent the imputed methylation levels for the EPIC array based on the original data of the HM450K array. Further, we evaluated the performance of our method by comparing the predicted results with the original EPIC data. The MSE loss between these predicted results and the original EPIC data was calculated and found to be 0.0086, which was slightly higher than the reconstruction loss of EPIC array in the testing dataset, which was 0.0084.

To calculate the final MSE loss, we restored all the standardized data back to their respective original ranges. The final MSE loss was 0.0004, indicating a very good performance of our method. To evaluate the performance of our method, we generated scatter plots in Figure 3, which clearly showed the predicted values to be very close to the original values in the testing dataset of the EPIC array.

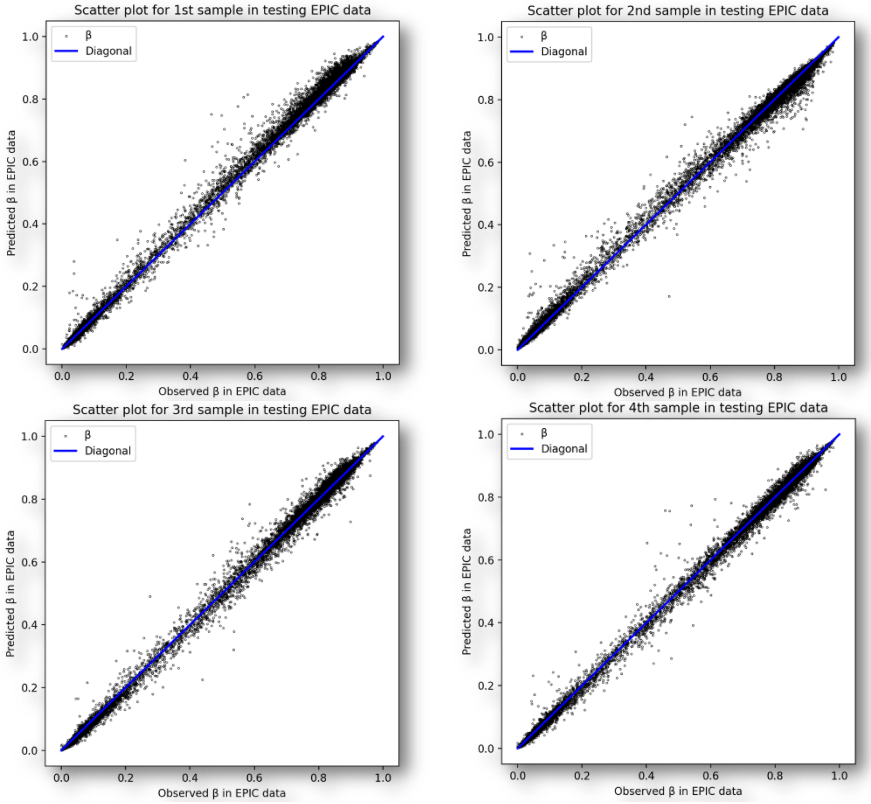

Figure 3. Comparisons of observed and predicted beta values of DNA methylation levels of first 4 samples in the testing dataset of EPIC arrays

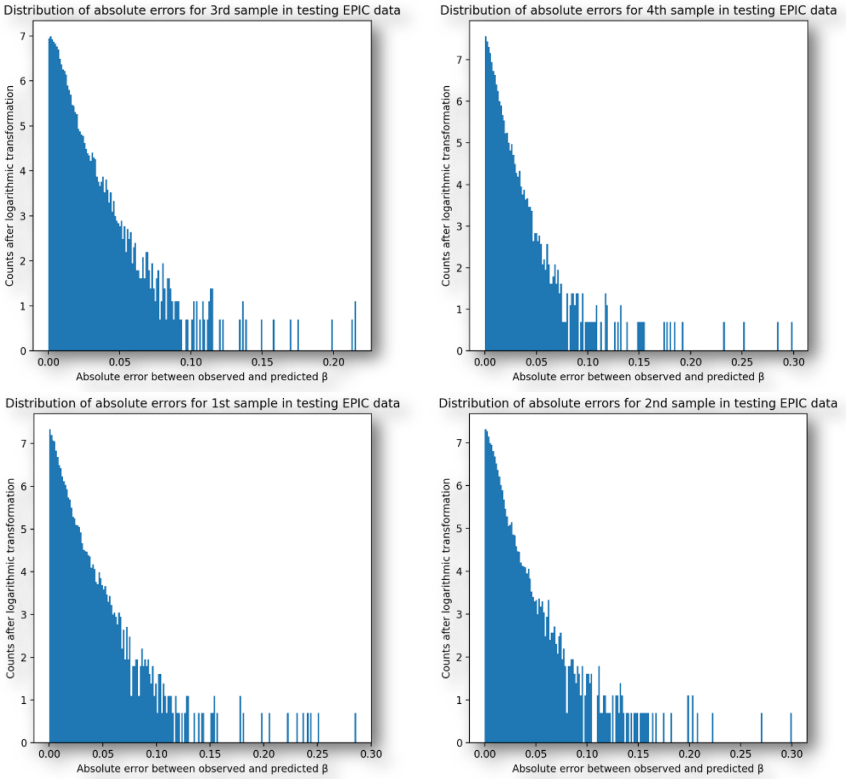

Figure 4. Frequency distribution of Absolute Error between predicted and observed beta values of DNA methylation levels of first 4 samples in the testing dataset of EPIC arrays

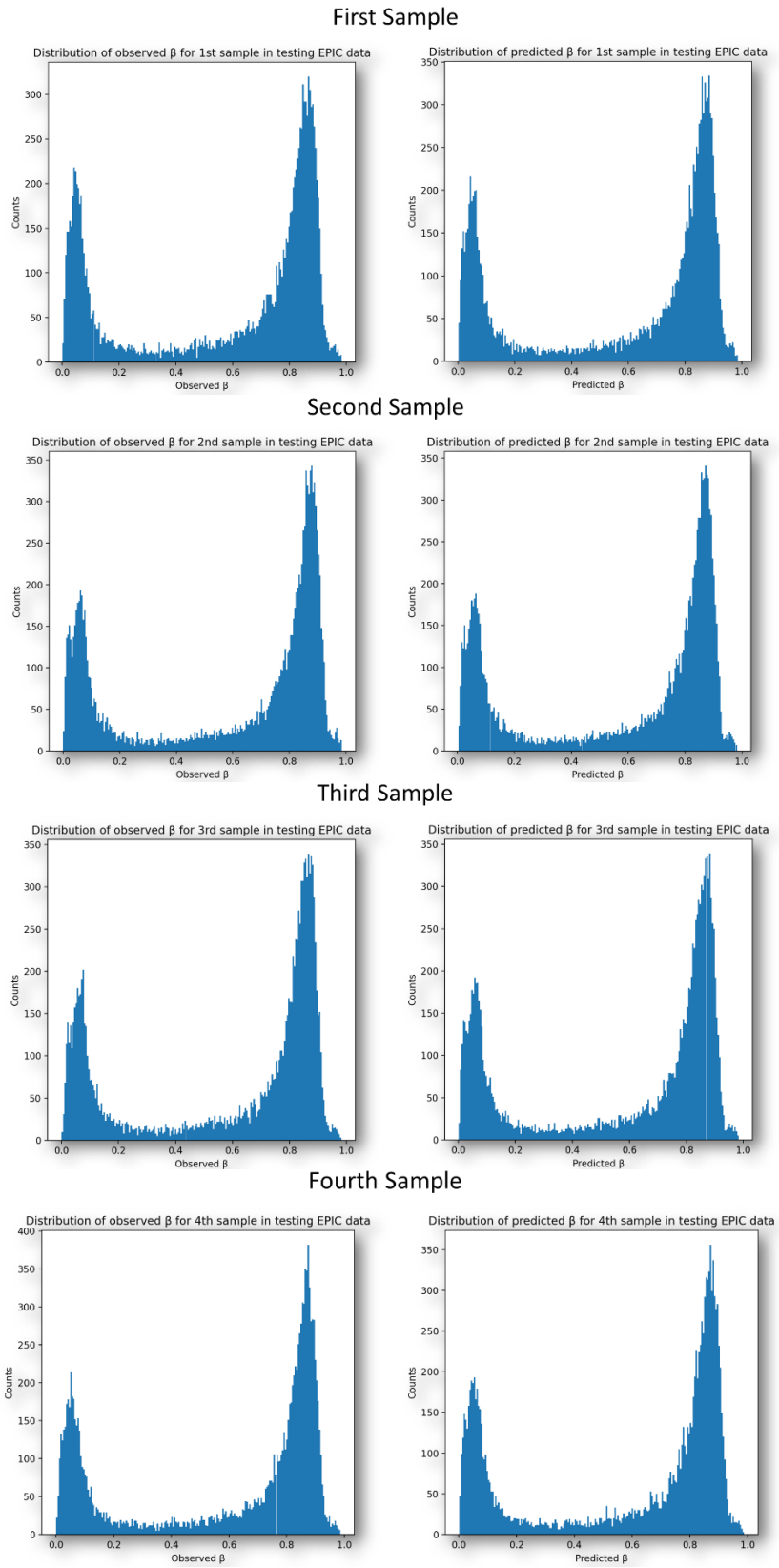

Figure 5. Comparison of observed and predicted beta value distributions of DNA methylation levels of first 4 samples in the testing dataset of EPIC arrays (Left: observed distribution, Right: predicted distribution)

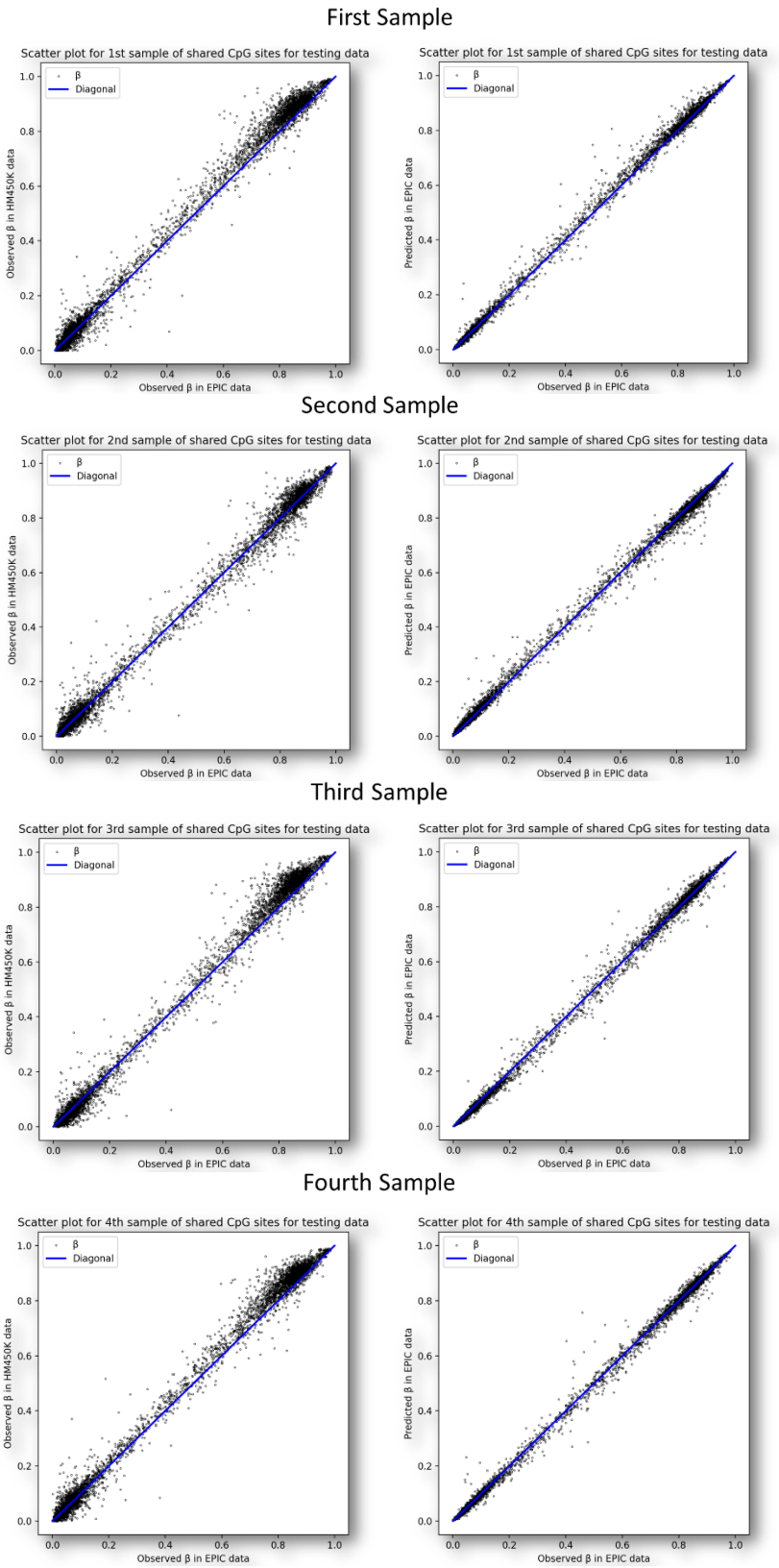

Figure 6. Comparisons of observed and predicted beta values of shared CpG sites in both arrays of first 4 samples in the testing dataset (Left: comparisons of observed betas in HM450K array and observed betas in EPIC array, Right: comparisons of observed and predicted betas in EPIC array)

As shown in Figure 4, most of the differences between the predicted beta values and the original beta values in the testing dataset of the EPIC array are almost negligible, which indicates that our model's predictions are very accurate. Furthermore, Figure 5 shows that our model captures the original distribution of beta values in a simple sample and does not predict all beta values as fixed for all samples. This further confirms the robustness of our method and its ability to capture the true biological variations in the data. Previous studies have assumed that CpG sites common to both the HM450K and EPIC arrays would exhibit same or very similar methylation levels, but the results depicted in Figure 6 suggest that this is not necessarily the case. By using our model to predict the methylation levels of these shared CpG sites, we can achieve more accurate and reliable results across different platforms.

4. Discussions

Autoencoder and nonnegative matrix factorization are powerful techniques that can handle the high-dimensional nature of DNA methylation data. Our model successfully captured the complex relationships between CpG sites in the two arrays, and produced accurate imputations of beta values in the EPIC array. The overall RMSE (root mean square error) was 0.0196, which was better than 0.04 from simple linear regression model with nearby CpG sites. The improvement in RMSE compared to a simple linear regression model suggests that our approach was effective at capturing non-linear relationships between CpG sites. The overall MAE (mean absolute error) was 0.0122. The low MAE indicated that our model had high accuracy in predicting methylation levels.

Our model had the capability to impute methylation levels of all CpG sites in the EPIC array, including those that are also present in the HM450K array. Our model would be highly adaptable to other chromosomes and could easily adjust the dimensions of the results obtained from

autoencoders to accommodate different chromosome sizes. Our model can be expanded to facilitate imputation from EPIC version 1 to the most recent version, EPIC version 2. By including more comprehensive coverage from both arrays, we anticipate that the imputation quality will improve. A real data application of our method can enhance the power of meta-analyses that merge data from both the 450K and EPIC arrays. This would allow for the integration of a larger number of samples and CpG sites, potentially increasing the statistical power of the analyses and enabling the identification of more robust associations between DNA methylation and various traits or diseases.

Despite the promising results, our study had some limitations. The relatively small sample size may have limited the generalizability of our findings to larger populations. Additionally, the requirement for a long running time and large storage space for the data and the results may limit the practical application of our model. Further research is needed to address these limitations and to refine our approach.

References

1. Michael Klutstein, Deborah Nejman, Razi Greenfield, Howard Cedar; DNA Methylation in Cancer and Aging. *Cancer Res* 15 June 2016; 76 (12): 3446–3450. https://doi.org/10.1158/0008-5472.CAN-15-3278
2. Palou-Márquez, G., Subirana, I., Nonell, L. et al. DNA methylation and gene expression integration in cardiovascular disease. *Clin Epigenet* 13, 75 (2021). https://doi.org/10.1186/s13148-021-01064-y
3. Younesian S, Yousefi A-M, Momeny M, Ghaffari SH, Bashash D. The DNA Methylation in Neurological Diseases. *Cells*. 2022; 11(21):3439. https://doi.org/10.3390/cells11213439
4. Mahmood, Niaz, and Shafaat A. Rabbani. DNA Methylation Readers and Cancer: Mechanistic and Therapeutic Applications. *Frontiers in Oncology* 9, (2019). https://doi.org/10.3389/fonc.2019.00489.
5. Mohamad Hesam Shahrajabian, Wenli Sun, and Qi Cheng. DNA methylation as the most important content of epigenetics in traditional Chinese herbal medicine, *Journal of Medicinal Plants Research* 13, no. 16 (2019): 357-369. https://doi.org/10.5897/JMPR2019.6803
6. Tang, J., Xiong, Y., Zhou, H. H. and Chen, X. P., DNA methylation and personalized medicine. *J Clin Pharm Ther*, 39: 621-627 (2014). https://doi.org/10.1111/jcpt.12206
7. Naeem, H., Wong, N.C., Chatterton, Z. et al. Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the HumanMethylation450 array. *BMC Genomics* 15, 51 (2014). https://doi.org/10.1186/1471-2164-15-51

8. Mansell, G., Gorrie-Stone, T.J., Bao, Y. et al. Guidance for DNA methylation studies: statistical insights from the Illumina EPIC array. *BMC Genomics* 20, 366 (2019). https://doi.org/10.1186/s12864-019-5761-7
9. Li, G., Zhang, G., Li, Y. (2022). DNA Methylation Imputation Across Platforms. In: Guan, W. (eds) Epigenome-Wide Association Studies. *Methods in Molecular Biology*, vol 2432. Humana, New York, NY. https://doi.org/10.1007/978-1-0716-1994-0_11
10. Gang Li, Laura Raffield, Mark Logue, Mark W. Miller, Hudson P. Santos Jr, T. Michael O'Shea, Rebecca C. Fry & Yun Li (2021) CUE: CpG impUtation ensemble for DNA methylation levels across the human methylation450 (HM450) and EPIC (HM850) BeadChip platforms, *Epigenetics*, 16:8, 851-861, DOI: 10.1080/15592294.2020.1827716
11. Z. Cheng, H. Sun, M. Takeuchi and J. Katto, Deep Convolutional AutoEncoder-based Lossy Image Compression, *2018 Picture Coding Symposium (PCS)*, San Francisco, CA, USA, 2018, pp. 253-257, doi: 10.1109/PCS.2018.8456308.
12. Y. -X. Wang and Y. -J. Zhang, Nonnegative Matrix Factorization: A Comprehensive Review, in *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 6, pp. 1336-1353, June 2013, doi: 10.1109/TKDE.2012.51.