Developing a Large-Scale Foundation Model on the DNA Methylome Using Transformers

Kejun Ying[1], Jinyeop Song[2], Basheer Becerra[3]
[1]Harvard School of Public Health
[2]Department of Physics, MIT
[3]Bioinformatics and Integrative Genomics PhD Program, Harvard Medical School

Abstract

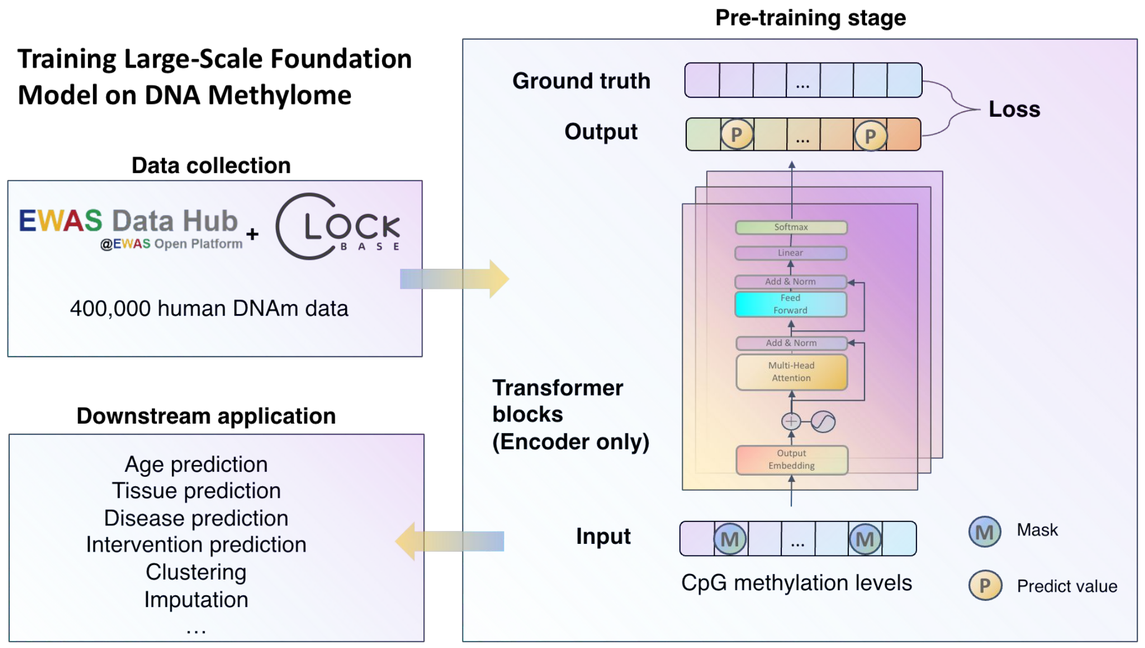

Figure 1. Schematic representation of the model architecture

- Epigenetic phenomena like **DNA methylation** play a crucial role in gene expression and are pivotal in the study of **aging and disease**.
- We introduce two transformer-based models, reminiscent of **BERT** and **GPT** architectures, tailored to interpret patterns in DNA methylation data from 13,747 patients.
- Through our model, **patient embedding** of DNA methylation pattern can be generated, providing a robust tool for personalized medicne and future research in the biology of aging and disease prediction

Dataset

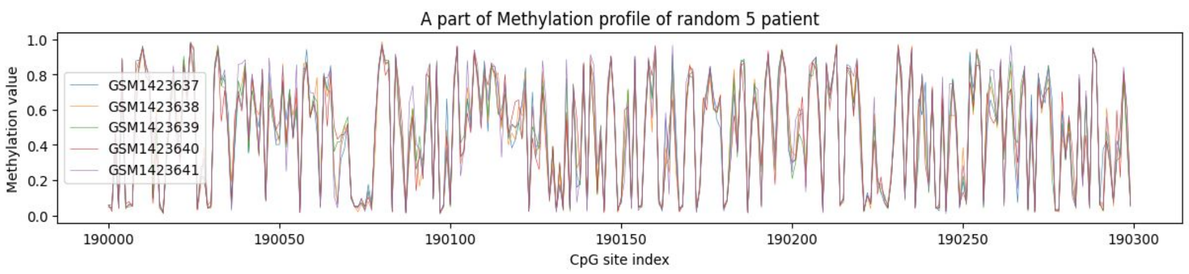

Figure 2. Representative methylation profile of random 5 patients

- Our dataset comprises DNA methylation (DNAm) profiles from 13,747 patients, curated from 2,000 distinct sites.
- DNAm profile consists of normalized methylation values over ~400000 CpG sites (Figure 2)
- Most CpG sites have constant methylation value across patients (Figure 2). So we restrain our scope into "**highly variable" CpG sites(HV CpG sites).** (Figure 3)

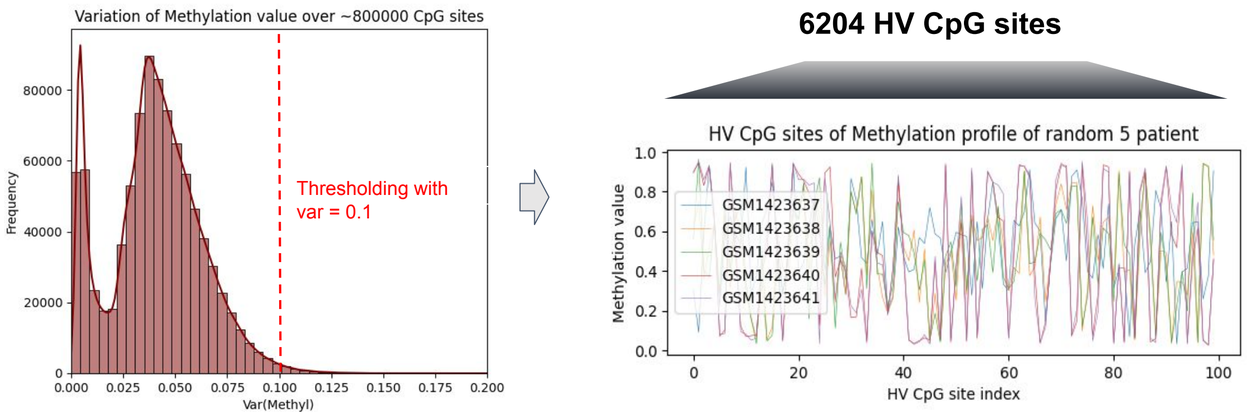

Figure 3. 6204 HV CpG sites are acquired through variation thresholding

Methods

We employed two training strategies : BERT-style and GPT-style

	Transformer Layer #	Emb dim	Layer dim	Input token size	Training Loss
BERT-style	3	64	48	1024	Masked Value Prediction
GPT-style					Masked Value Prediction + Autoregression by Cell Embedding

Table. Specifications for model and parameters

Results - Training curve & Masked Value Prediction

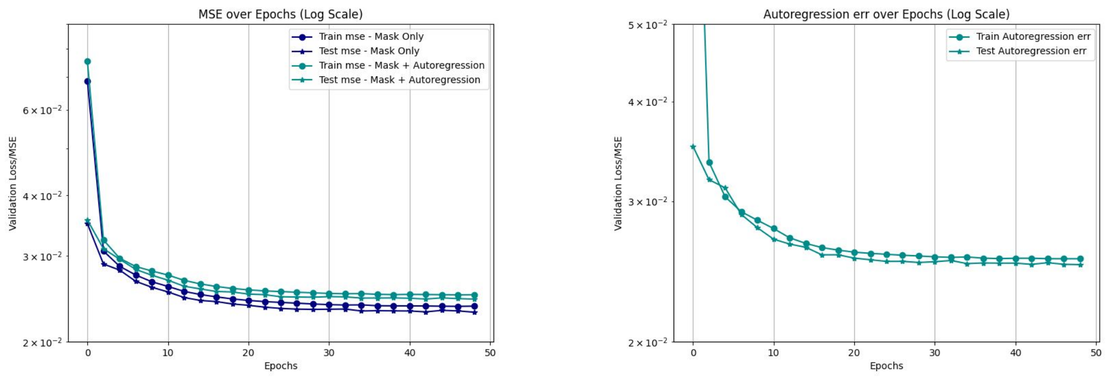

Figure 4. MSE loss curve and Autoregression curve for BERT-style and GPT-style

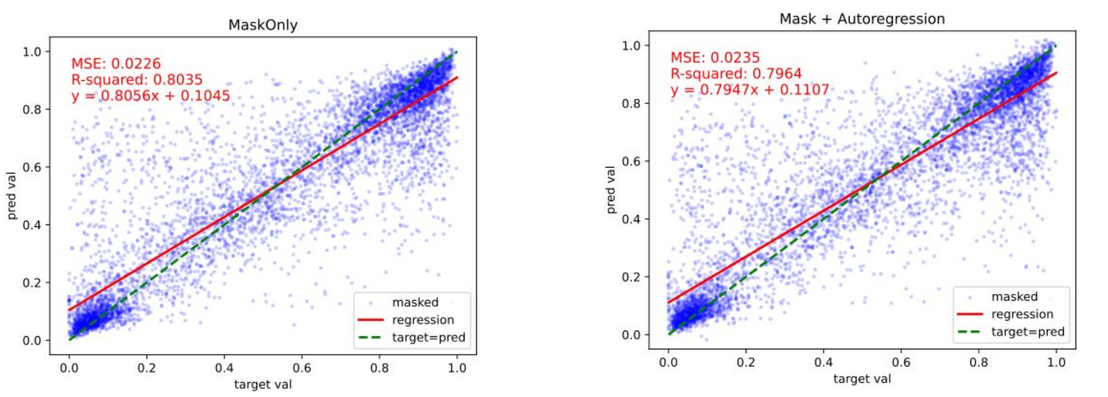

Figure 5. Masked Value Prediction performance for BERT-style and GPT-style

Results - Patient Embedding

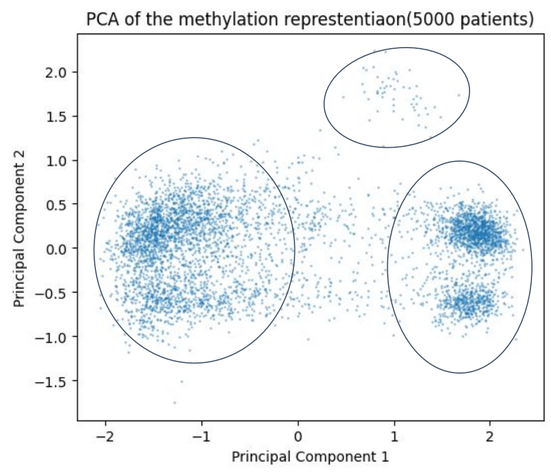

Figure 6. PCA of methylation representation

- We identified few clusters based on the first two axes of PCA. (Figure 6)

Conclusions & Future work

- We introduced an transformer-based model to process the DNAm profile.
- TODO 1- Sizing up the model
- TODO 2 - Predict the Metadata of patient, such as ages and diseases, with the patient embedding from our model

Acknowledgements

We would like to express gratitude to everyone in NLP 6.861!

References

[1] Peter A Jones. 2012. Functional epigenomics: the key to understanding the complexity of phenotypic traits. Annual review of genetics, 46:75–92.
[2] Jumper et al. 2021. Highly accurate protein structure prediction with alphafold. Nature
[3] Cui, Haotian, et al. "scgpt: Towards building a foundation model for single-cell multi-omics using generative ai." bioRxiv (2023): 2023-04.