9012W

GEM: A deep generative framework for synthetic generation of plasma cfDNA methylation profiles

Harbinger Health

Sarah Baalbaki[1], Shiva Farashahi[1], Dorna Kashef[1], Kieran I Chacko[1]*.
[1]Harbinger Health, Cambridge, MA *Corresponding author: kchacko@harbinger-health.com

BACKGROUND

Methylation analysis of cell-free DNA (cfDNA) is a powerful approach for liquid biopsy, providing insights into health and disease from a simple blood draw. However, machine learning applications in this space face challenges: restricted training data, class imbalances, subtle signals of interest, technical noise, and biological noise from cfDNA's heterogeneous cell-type origins that can obscure condition-specific patterns. Synthetic data generation addresses these challenges by producing realistic read-level cfDNA methylation data to expand datasets, model background variation, and improve signal detection. We present GEM (Generative Epigenomic Modeling), a novel diffusion-based framework built on Denoising Diffusion Implicit Models (DDIMs). GEM learns detailed methylation signatures from cfDNA reads and generates high-fidelity synthetic methylation data closely resembling real cfDNA data, providing a new foundation for creating biologically realistic synthetic datasets.

CONCLUSIONS

- We developed GEM to model methylation patterns in non-cancer cfDNA, generating reads across 100 genomic regions.
- Our results show that GEM produces biologically realistic synthetic datasets closely matching real data.
- Ongoing work includes conditioning on technical and biological factors and will be further extended to generate cancer fragments and model disease-specific variation.
- GEM will enhance our large-scale simulation platform [4] to enable generation of more generalizable synthetic samples.
- Our approach offers broad impact, supporting general applications such as data augmentation, rare condition modeling, and the simulation of controlled signal-to-noise datasets.

METHODS

Sample Collection and Dataset: Blood samples (NCT05435066) were collected from cancer free individuals across the ages of 30-79. A subset of these samples (N=150) were selected for model development, including 100 samples for model training and 50 held-out non-cancer samples for generation. Extracted cfDNA was analyzed using a custom targeted bisulfite sequencing hybrid capture assay (HHx; 18.6 Mb). From this panel, 100 regions with highly heterogenous methylation patterns (>20 CpG sites) were selected to train the model and evaluate the model's ability to generate diverse methylation patterns.

Input representation: Each cfDNA read was standardized to 190 bp and encoded as a binary sequence representing CpG methylation states, restricted to CpG sites (channel 1). A second binary channel (channel 2) distinguished true versus padded CpGs, ensuring consistent sequence lengths across regions. The relative starting CpG position was added as a conditioning variable to preserve positional context **(Fig 1)** [3].

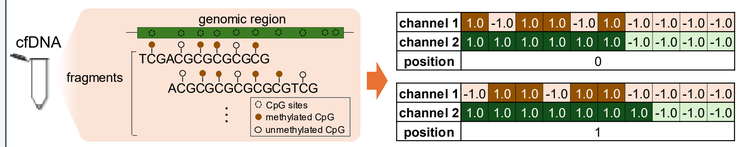

Figure 1. Encoding of cfDNA reads.

Model framework: We adapted a U-Net based DDIM framework [1,2] to learn methylation signatures from cfDNA fragments and generate synthetic read-level data. This framework is designed to capture local CpG dependencies across neighboring sites, while preserving broader region-level context with attention mechanisms. For each genomic region, a separate continuous DDIM model was trained on cfDNA fragments from the 100 non-cancer individuals selected for training, with outputs binarized using 0-thresholding.

Generation: Synthetic reads were generated at different start positions using the DDIM reverse process, beginning from Gaussian noise and iteratively denoising into realistic sequences. At the sample level, 50 synthetic non-cancer samples were generated with coverage matched to the 50 real held-out non-cancer samples, using read counts at each start position to preserve coverage profiles. Synthetic BAM files were reconstructed using these positions as references, incorporating model-derived methylation states and read lengths, and were subsequently used to compute methylation-level metrics (defined to reflect region-wide methylation levels) **(Fig 2).**

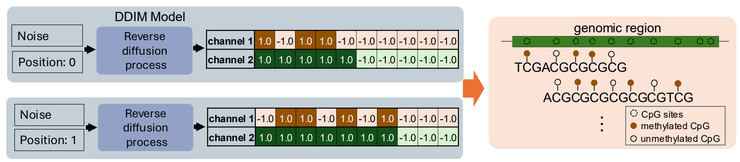

Figure 2. Example cfDNA reads are generated by conditioning on start positions (shown for positions 0 and 1) and denoising from gaussian noise.

RESULTS

Generated reads match real reads in CpG count distributions

Real and generated cfDNA reads showed similar CpG count distributions (t-test: $p > 0.07$), indicating no statistically significant differences between the two groups and suggesting GEM successfully reproduces true read length characteristics. **(Fig 3)**.

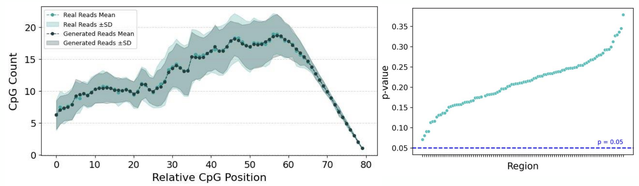

Figure 3. Average CpG count across relative CpG positions within a read for a representative region using 50 real and 50 generated samples (left), and p-values across all regions comparing the CpG counts of the same real and generated samples (right). Higher p-values indicate stronger agreement.

Generated samples preserve methylation signal

Methylation metrics from 50 real and 50 generated non-cancer samples exhibited similar variability across mean methylation, confirming GEM preserves biologically realistic methylation signals **(Fig 4)**.

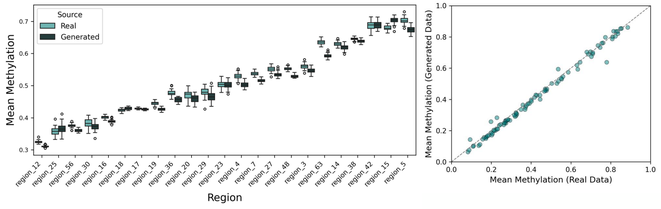

Figure 4. Mean methylation boxplot across a random subset of regions showing similar distributions (left) and scatter plot showing strong correspondence between the means across real and generated samples per region (right).

Generated samples resemble real samples

A UMAP fit on methylation metrics from real non-cancer samples showed synthetic samples clustering among the real held-out non-cancer data, supporting sample-level dataset fidelity **(Fig 5).**

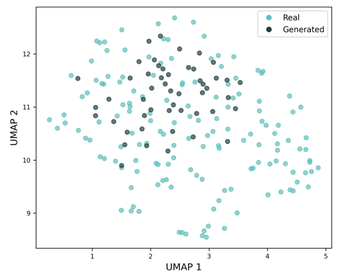

Figure 5. UMAP for real and generated samples.

Generated reads capture methylation patterns within regions

To assess position-specific learning, we computed 5-CpG hepitype probabilities, defined as the fraction of reads spanning 5 consecutive CpGs with each binary methylation pattern, ranging from fully unmethylated (00000) to fully methylated (11111). Generated and real distributions closely matched (MSE= 5.5e-5±5.1e-5), confirming GEM captures detailed CpG methylation dependencies **(Fig 6).**

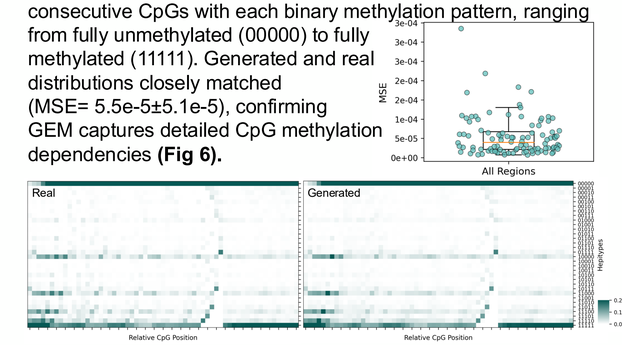

Figure 6. MSE distribution between real and generated hepitype probabilities across all regions (top) and hepitype heatmap for an example region showing close agreement between real and generated hepitype probabilities (bottom).

Age-conditioned generation preserves age-associated methylation signals

As a proof of concept, GEM was conditioned on age as an additional variable and evaluated in an age-confounded region, where methylation levels correlate with age in real samples. GEM replicated age-associated methylation patterns in age-conditioned generation, preserving biologically relevant patterns **(Fig 7).**

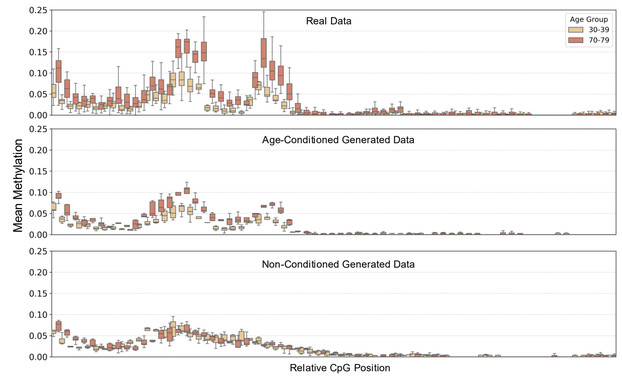

Figure 7. Mean methylation for younger and older age groups in real data (top) and age-conditioned data (center) show similar methylation patterns, compared to non-conditioned data (bottom).

REFERENCES

1. J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv*:2010.02502, 2020.
2. T. Chen, R. Zhang, and G. Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. *arXiv preprint arXiv*:2208.04202, 2022.
3. S. Farashahi, Y. Wu, E. Massaad, F. Hantash, H. Ashrafian, D. Kashef, & K. Chacko. Denoising Models Enhance Detection of Tumor-Derived cfDNA fragments and Cancer Tissue signal in Liquid Biopsy. Clinical Cancer Research, 31, 2025.
4. K. Pettie, S. Farashahi, J. Killian, D. Kashef, J. Hubbell, F. Hantash, J. Charlton, & K. Chacko. FabricaTM: A large-scale data simulation platform isolates tumor signal from cell-free DNA and improves tissue of origin prediction accuracy. *Clinical Cancer Research*, 30, 2024.

ACKNOWLEDGEMENTS

We would like to thank Yifan Wu for their assistance with the implementation and Jocelyn Charlton for their feedback on the poster.
We are thankful to the lab team at Harbinger Health for their dedication and expertise in managing the sample preparation.
We gratefully acknowledge all participants for their contributions, without whom this research would not have been possible.

www.harbinger-health.com