GENETICS

Genome-wide methylome modeling via generative AI incorporating long- and short-range interactions

Fengyao Yan[1], Aristeidis G. Telonis[2], Qin Yang[1,9], Limin Jiang[1], Francine E. Garrett-Bakelman[3,4,5], Mikkael A. Sekeres[1,6], Valeria Santini[7], Michele Ceccarelli[1,8], Neha Goel[1,8], Liliana Garcia-Martinez[1,9], Lluis Morey[1,9], Maria E. Figueroa[1,2,6]*, Yan Guo[1,10]*

Using millions of methylation segments, we developed DiffuCpG, a generative artificial intelligence (AI) diffusion model designed to solve the critical challenge of missing data in high-throughput methylation technologies. DiffuCpG goes beyond conventional methods by leveraging both short-range interactions including nearby CpGs from both latitude and longitude of the dataset, local DNA sequences, and long-range interactions, including three-dimensional genome architecture and long-distance correlations, to comprehensively model the methylome. Compared to previous methods, through extensive independent validations across different tissue types, cancers, and technologies (whole-genome bisulfite sequencing, enhanced reduced representation bisulfite sequencing, single-cell bisulfite sequencing, and methylation arrays), DiffuCpG has demonstrated superior performance in accuracy, scalability, and versatility. On average, bisulfite sequencing dataset, DiffuCpG can extend the original dataset by millions of additional CpGs. As an alternative application of generative AI, DiffuCpG addresses a key bottleneck in epigenetic research and will substantially benefit studies relying on high-throughput methylation data.

INTRODUCTION

Methylation constitutes a fundamental biochemical process entailing the addition of a methyl group to macromolecules, such as DNA, RNA, or proteins. On DNA, this chemical modification frequently occurs at CpG sites, characterized by a cytosine (C) nucleotide followed by a guanine (G) nucleotide, and linked by a phosphate group (p). Within a CpG site, cytosines may undergo methylation through the addition of a methyl group to the carbon atom on position 5 of the cytosine ring, thereby yielding 5-methylcytosine (5mC). Methylation is the subject of extensive biomedical inquiry due to its pivotal role as an epigenetic mechanism governing gene expression (*1*, *2*), cellular differentiation (*3*), genomic stability (*4*), and heritability (*5*). Dysregulation of methylation patterns has been implicated in a spectrum of diseases, including cancer (*6*, *7*), neurological disorders (*7*), aging (*8*), and developmental anomalies (*9*).

Numerous high-throughput methodologies are available for the quantification of DNA methylation, among which microarray-based techniques have historically been favored. This approach uses bisulfite-converted DNA and uses hybridization to interrogate sequences representing CpG sites across the genome. However, a notable drawback of microarray-based methodologies is their inherent limitation in accommodating the vast number of CpG sites, with one of the most widely used platforms, the Infinium HumanMethylation450 BeadChip (450K), targeting less than 2% of CpG sites (*10*). Presently, bisulfite sequencing stands as the prevailing high-throughput method for methylation quantification. Bisulfite treatment facilitates the conversion of unmethylated cytosines to uracil, whereas methylated cytosines remain unaltered. Following polymerase chain reaction amplification and sequencing, the methylation status of individual cytosines can be discerned. Bisulfite sequencing can be divided into two types, whole-genome bisulfite sequencing (WGBS) and reduced representation bisulfite sequencing (RRBS). As implied by its nomenclature, WGBS endeavors to capture the methylation status of the entire genome or methylome. Conversely, RRBS selectively enriches CpG-rich genomic regions by using restriction enzymes. Although RRBS is comparatively more cost-effective than WGBS, it carries a trade-off, sacrificing coverage of genomic regions characterized by lower CpG content. The enhanced reduced representation bisulfite sequencing (ERRBS) method represents a modified version of traditional RRBS that augments genomic coverage into more distal regions (*11*). Although bisulfite sequencing methods are most commonly applied to bulk cells and return a methylation percentage, when applied at a single-cell level, the methylation state can only be binary.

Many factors influence the DNA methylation status of individual CpGs. Studies have revealed that within *cis*-CpG interactions, the methylation level of a specific CpG site is notably influenced by neighboring CpG sites, with the strength of this effect diminishing as the genomic distance increases (*12*). Furthermore, it has been demonstrated that the DNA sequence itself, through genetic variations, can influence local methylation levels (*13*) such as in methylation quantitative trait loci (meQTLs). In *trans*-CpG interactions, the influence of distant genes, such as *ACD* and *SENP7*, on methylation patterns has been elucidated through the effects of *trans*-meQTLs (*14*). Other features of the human genome may also play pivotal roles in methylation regulation. For instance, three-dimensional (3D) genomic structural distances and topologically associating domains (TADs) are emerging as critical determinants influencing methylation

[1]Sylvester Comprehensive Cancer Center, University of Miami Miller School of Medicine, Miami, FL 33136, USA. [2]Department of Biochemistry and Molecular Biology, University of Miami Miller School of Medicine, Miami, FL 33136, USA. [3]Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA 22908, USA. [4]Department of Medicine, University of Virginia, Charlottesville, VA 22908, USA. [5]Comprehensive Cancer Center, University of Virginia, Charlottesville, VA 22908, USA. [6]Division of Hematology, Department of Medicine, University of Miami Miller School of Medicine, Miami, FL 33136, USA. [7]MDS Unit, DMSC, University of Florence, AOU Careggi, Florence 50134, Italy. [8]Department of Surgery, University of Miami Miller School of Medicine, Miami, FL 33136, USA. [9]Department of Human Genetics, University of Miami Miller School of Medicine, Miami, FL 33136, USA. [10]Department of Public Health and Sciences, University of Miami, Miami, FL 33136, USA.

*Corresponding author. Email: mefigueroa@miami.edu (M.E.F.); Yxg835@med.miami.edu (Y.G.)

patterns (*15*). Understanding the interplay between these structural features and methylation dynamics holds promise for advancing our comprehension of epigenetic regulation and facilitating methylation imputation strategies.

Similar to single-nucleotide polymorphism (SNP) data, methylation data are susceptible to the issue of missing values. However, in contrast to SNP data, the imputation of missing methylation data is less effective due to the absence of haplotype information. Previous efforts have been made to address methylation imputation (*12*, *16*), with notable attempts including DeepCpG (*13*), an imputation tool that integrates convolutional neural network (CNN) and bidirectional gated recurrent unit (GRU) architectures, yielding commendable outcomes. In this study, we propose an innovative methylation imputation methodology founded upon a generative artificial intelligence (AI) approach. Generative models are adept at producing new data that closely resemble a given dataset. For instance, the Bing Image Creator leverages DALL-E 3, a formidable image generation model developed by OpenAI, to generate images based on textual descriptions. DALL-E 3 is a diffusion-based generative AI model extensively used in image generation tasks. Its methodology involves initiating the process by introducing noise to the image, known as the forward process. Subsequently, during the reverse process, the model learns and reconstructs the original image without noise. Upon completion of training, the model gains the capability to generate new images by initiating with random noise and iteratively denoising the input.

In this study, we adapted generative diffusion techniques for the imputation of methylation level, thereby broadening its utility beyond image and text generation. We used WGBS data on samples from 26 patients with acute myeloid leukemia (AML) as the training set. We validated DiffuCpG's performance on ERRBS data from 93 myelodysplastic syndrome samples. Additional independent test sets included 450K and single-cell reduced representation bisulfite sequencing (scRRBS) on 26 HepG2 cells.

RESULTS

CpG coverage

The latest CRCh38 human reference genome features 28 million CpG sites. In Fig. 1, we present a comparison of CpG coverage among four major technologies: WGBS (Fig. 1A), ERRBS (Fig. 1B), scRRBS (Fig. 1C), and 450K (Fig. 1D). Using a typical genomic span of 100,000 base pairs (bp) on chromosome 1 (200,549,495 to 200,649,494) and 20 randomly selected samples from each technology, we illustrate the disparities in CpG coverage. On average, WGBS covers 79% of CpG sites across the genome, 80% within protein-coding regions, defined as the length of protein-coding genes extending 1500 bp upstream and downstream, and 79% across all gene regions. ERRBS, on the other hand, covers 17% of CpG sites genome-wide, 20% within protein-coding regions, and 19% within all gene regions. 450K covers 0.6% of CpG sites across the genome, 0.9% within protein-coding regions, and 0.8% within all gene regions. In comparison, scRRBS covers the sparsest CpG sites (Fig. 1E). However, all technologies are susceptible to missing data issues, defined as a CpG that is detected in at least one sample in the dataset but missing in one or more samples in the same dataset. We compared four datasets to demonstrate the missing percentages by sample. For the WGBS dataset, the median missing percentage was 15% (high: 98%, low: 2%); for the ERRBS dataset, the median missing percentage detected was 79% (high: 99%, low: 59%); for the 450K dataset, the median missing percentage was 2% (high: 100%, low: 0.1%); and last, the scRRBS dataset had the highest median missing percentage at 92% (high: 98%, low: 70%) (Fig. 1F).

Performance and comparison

The DiffuCpG developed from a diffusion model was trained on 2.6 million 1000-bp-long genome segments from 26 WGBS data of 26 AML samples. It is designed to operate on multichannel 1D data guided by a custom inpainting algorithm (Fig. 2A), imputing missing methylation values within fixed-size genomic windows. Within each window, the missing methylation values can be imputed using the remaining methylation values and data from other channels such as DNA sequence. The model's design relies on two critical parameters: window size and the number of missing CpG sites within the window. We trained DiffuCpG models using WGBS data for various window sizes ranging from 500 to 10,000 bp. For each window size, a portion of methylation values was randomly removed and then imputed. Performance was assessed using the average of accuracy, F1 score, and correlation. Analysis indicated that imputation performance is positively associated with CpG density. Through extensive evaluation, a 1000-bp window size with at least 10 measured CpG sites provided the best performance, achieving a median imputation performance exceeding 80% (Fig. 2, B and C). Different feature combinations were tested when developing DiffuCpG. The combination including DNA sequence data, Hi-C data, and cross-sample confidence interval (CI) data achieved the best overall performance score (Fig. 2D). We also performed a comprehensive survey of the number of CpG sites that can be accurately imputed for different technologies (Fig. 2E). On average, the numbers of imputable CpGs are 2,885,129, 3,231,864, 1,455,000, and 238,300 for WGBS, ERRBS, scRRBS, and 450K, respectively.

The DiffuCpG's superior performance was demonstrated by comparing it to four other methylation imputation tools: MissForest (*17*), LightCpG (*12*), MethyLImp2 (*18*), and DeepCpG (*13*). The major differences between these four tools plus DiffuCpG are summarized in Table 1. MissForest and LightCpG are tree-based methods using random forest and lightGBM, respectively. MethyLImp2 is a linear regression–based method. DeepCpG used both a CNN and a bidirectional GRU. DiffuCpG uses a diffusion model and uses the greatest number of features, including short-range interaction features, neighboring CpGs, sequence, and cross-sample information, and long-range interaction features, Hi-C, and correlation. These features were prebuilt into the DiffuCpG model; thus, they do not increase the input burden for future applications. Among all the tools compared, MethyLImp2 and MissForest do not support single-sample imputation as their algorithms depend on cross-sample information.

The DiffuCpG's performance was thoroughly compared to other tools in three independent test datasets (450K, scRRBS, and ERRBS). DiffuCpG showed the best performance with 86% of imputation results less than 0.25 from the true values (Fig. 2F) when imputing from ERRBS. DiffuCpG demonstrated the highest number of imputable CpGs (Fig. 2G) in various datasets. In another performance evaluation, for the 450K test cases, DiffuCpG achieved the highest balanced accuracy (0.88) and correlation (0.93) and the lowest Root Mean Squared Error (RMSE) (0.1); for scRRBS test cases, DiffuCpG ranked first in F1 (0.91); and for ERRBS test cases, DiffuCpG attained the highest balanced accuracy (0.91) and correlation (0.88), as well as the lowest RMSE (0.19). Combining all test cases, DiffuCpG outperformed all other tools compared in all performance

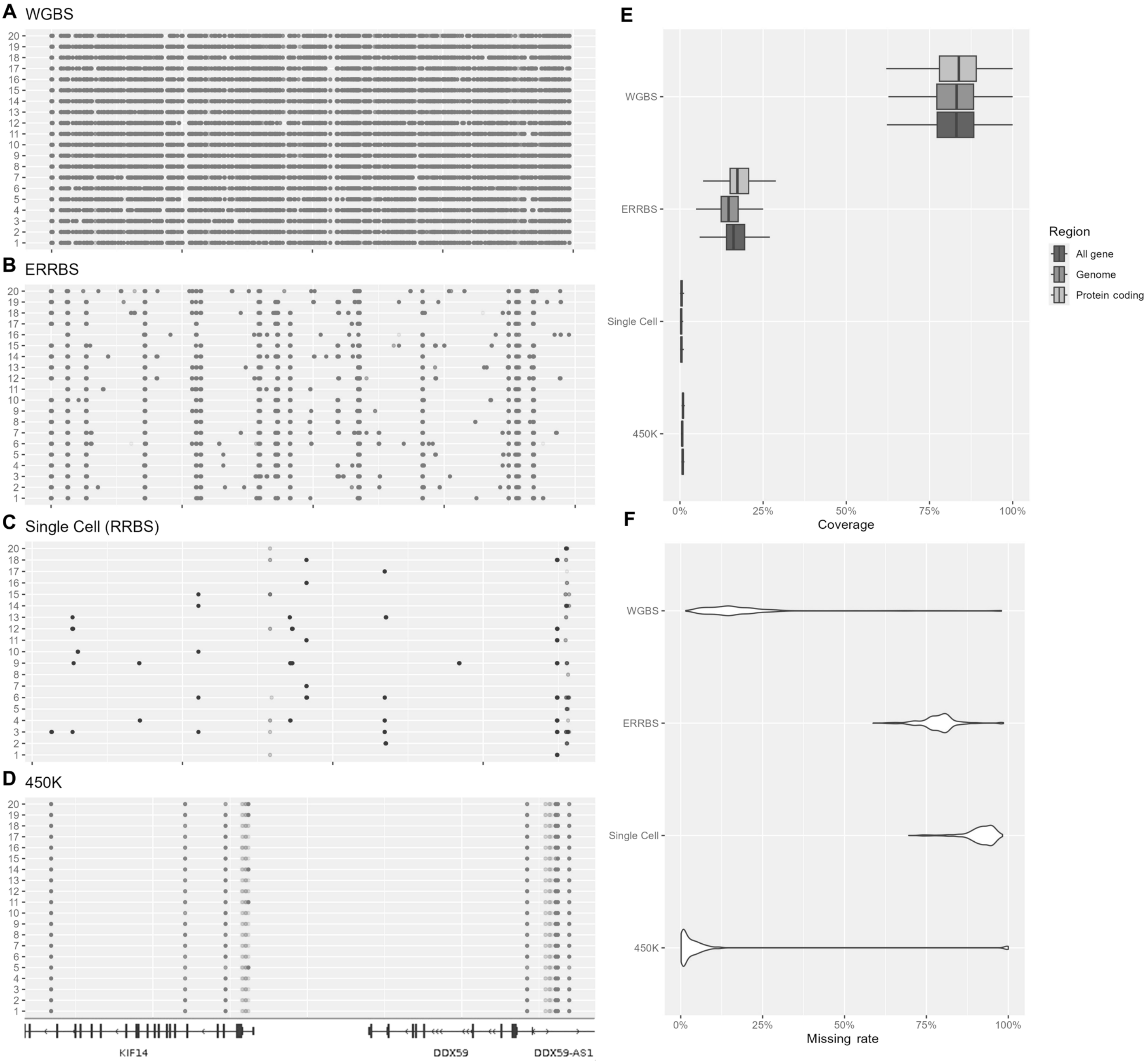

Fig. 1. Methylation data density in different platform. (**A** to **D**) CpG density covered by different methylation platforms. WGBS covers the most CpGs, whereas scRRBS covers the least. The genomic region displayed is 100,000 bp on chromosome 1 (positions 200,549,495 to 200,649,494) with 20 randomly selected samples from each technology. (**E**) Boxplots showing the percentage of covered CpGs at three different scales: genome-wide, in genes (both protein-coding and noncoding), and in protein-coding genes. (**F**) Violin plot showing the missing rate for each technology.

metrics (Fig. 2H). To verify that the sequence information is captured by our model, we shuffled the position of CpGs randomly within the 1000-bp window. The results are in purple designated as DiffuCpG on shuffled CpGs. Overall, we saw a 34% decrease in correlation, a 17% decrease in balanced accuracy, a 30% increase in RMSE, and a 32% decrease in F1 score after shuffling (Fig. 2H).

A detailed example of imputing an ERRBS sequenced Myelodysplastic Syndromes (MDS) sample is illustrated in Fig. 3. We selected a random genomic region on chromosome 1, spanning from 1,003,000 to 1,006,000 and covering 215 CpGs. The methylation levels of 179 CpGs detected by ERRBS are displayed on the top track, whereas the remaining 36 CpGs had no coverage. After randomly removing the methylation levels of 50% (89) of these CpGs, the remaining 90 CpGs were used for imputation. The subsequent tracks show the imputation results from each tool. All tools successfully imputed methylation levels for the removed 89 CpGs. Because MethyLImp2 and MissForest do not support single-sample imputation, additional samples were used for these two tools. For the 36 CpGs not

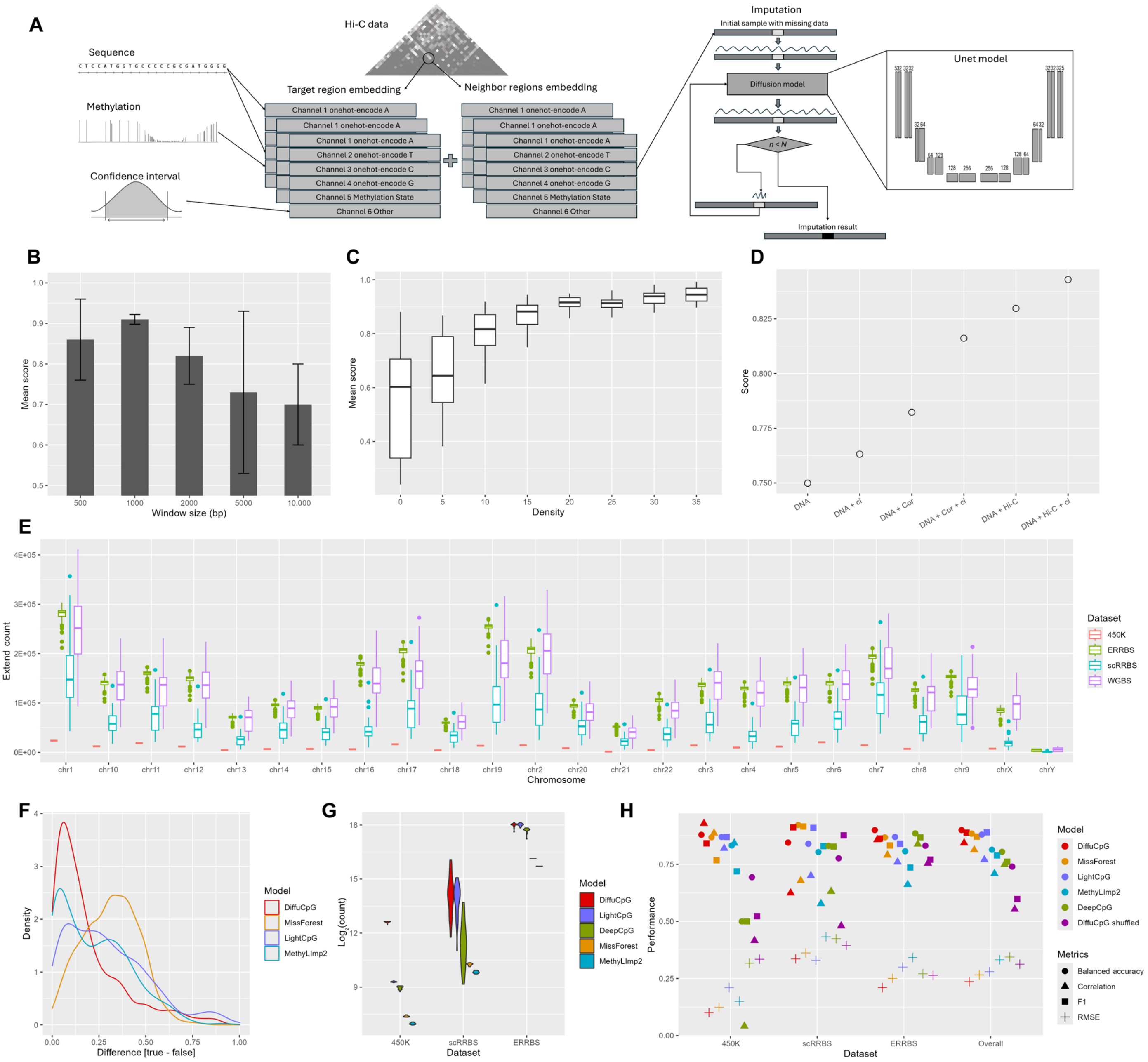

Fig. 2. Model performance comparison. (**A**) Depiction of the diffusion model used in our study. (**B**) Bar plots showing the results from our test for the ideal window size (1000 bp) for DiffuCpG. (**C**) Boxplots showing the density of CpGs can affect DiffuCpG performance. The density is defined as the number of available CpGs sites (not missing methylation level) in a 1-kb region. (**D**) Performance of DiffuCpG with different feature combinations. (**E**) Extendable (imputable) CpG sites using DiffuCpG by each chromosome. (**F**) Performance comparison of methylation tools for nonbinary imputation. DiffuCpG performed the best. The *x* axis denotes the difference between the true value (from ERRBS) and the imputed value. The *y* axis denotes the density of the difference. DeepCpG was excluded from this comparison because it only generates binary outcome. (**G**) Comparison of the number of imputable CpGs by different tools. Under the same conditions, for each technology, DiffuCpG can impute the greatest number of CpGs. (**H**) Performance metrics (balanced accuracy, correlation, F1 score, and RMSE) compared for each technology. Under the same conditions, DiffuCpG performed best in most scenarios.

covered by ERRBS, all are within 1-kb windows containing at least another 10 measured CpGs. Consequently, DiffuCpG could impute them with high accuracy based on previous density testing.

Cross-tissue, cross-disease, and cross-technology robustness

TADs influence DNA methylation by maintaining consistent epigenetic states within their domains and insulating methylation changes at their boundaries. Given that TADs are largely conserved across different tissue types (*19*, *20*), we hypothesized that most methylated CpGs are similarly conserved across tissues. Analyses show that DiffuCpG performed consistently across different tissues, diseases, and technologies. Similar to gene expression, methylation exhibits tissue specificity.

To estimate the average proportion of tissue-specific methylated CpGs (TS-CpGs) in normal tissues, we analyzed 450K methylation

Table 1. Methylation imputation tool comparison. Blank space indicates no association.

	DiffuCpG	DeepCpG	LightCpG	MethyLImp2	MissForest
Methods	Diffusion	CNN, GRU	lightGBM	Linear regression	Random forest
Short-range feature					
Sequence	✓	✓	✓		
Neighbor CpGs	✓	✓	✓	✓	✓
Cross-sample CpG	✓			✓	✓
Long-range feature					
Hi-C	✓				
Correlation	✓				
Functionality					
Binary imputation	✓	✓	✓	✓	✓
Nonbinary imputation	✓		✓	✓	✓
Single-sample imputation	✓	✓	✓		
Multisample imputation	✓	✓	✓	✓	✓
Multi-CpG imputation	✓			✓	✓
Evaluation					
Independent validation	✓	✓	✓	✓	✓
Cross-tissue	✓	✓		✓	✓
Cross-disease	✓				✓
Cross-technology	✓				✓

data from five tissue types: lung, breast, liver, prostate, and skin. Using Limma (*21*) to identify differentially methylated CpGs, we found that the lung, breast, liver, and prostate exhibited less than 4% TS-CpGs, with the liver showing the highest proportion at 3.2% (Fig. 4A). In the corresponding tumor types, the liver and prostate had the highest proportions of TS-CpGs (Fig. 4B). A differential methylation analysis between tumors and matched normal tissues revealed that the liver had the most differentially methylated CpGs (3.2%), whereas other tumor-normal comparisons exhibited less than 3% (Fig. 4C). These findings suggest that tissue and disease specificity may have a minor influence on the cross-tissue and cross-disease applicability of the DiffuCpG model.

We further validated the tissue-mixed DiffuCpG model trained on WGBS data across various methylation platforms. The model demonstrated strong performance, achieving an average score of 0.89 on ERRBS, 0.86 on 450K, and 0.83 on scRRBS (Fig. 4D). The relatively lower performance on 450K and scRRBS was attributed to the lower density of CpG coverage in these datasets. This limitation can be alleviated by training models directly on 450K data, which relies more on cross-sample consistency than neighboring CpGs. In cross-tissue validation, we applied tissue-specific models to different tissue types, achieving strong results (Fig. 4E). For the Glioblastoma Multiforme (GBM) specific model, performance was notably better when tested on AML data (0.98) compared to GBM data (0.91). The key reason for the better performance on AML data is that AML data are sequenced with WGBS, providing a higher CpG coverage. As shown in Fig. 2C, the density of available methylation sites during imputation positively affects model performance. We also classified CpGs into TS-CpGs and nontissue-specific CpGs (NTS-CpGs). Both tissue-specific and tissue-mixed models performed well on TS-CpGs and NTS-CpGs in normal (Fig. 4F) and tumor tissues (Fig. 4G), with no notable performance differences between these categories. These results highlight the robustness of DiffuCpG for methylation imputation across diverse contexts. Despite the tissue and disease specificity of methylation patterns, DiffuCpG consistently performs well, demonstrating its broad applicability and reliability.

DISCUSSION

Missing data are a critical issue in biomedical research as it can affect the statistical power, validity, reliability, and generalizability of study findings. Addressing and mitigating the effects of missing data require careful consideration and appropriate methods. Imputation is a commonly used tool for handling missing data, with traditional statistical approaches including mean, median, and mode imputation, as well as regression models. Recent advances in AI have introduced a series of deep learning models capable of handling complex and large-scale biological data (*22*, *23*). Among these, diffusion models represent one of the latest innovations, offering sophisticated methods for data imputation and enhancing the robustness of biomedical research.

A diffusion machine learning model (*24*) is a type of model inspired by the concept of diffusion processes, which describe how substances spread or propagate over time. In the context of machine learning, diffusion models have been particularly influential in generative modeling and have found applications in various domains, including image synthesis (*25*), signal processing (*26*), and natural language processing (*27*). These models simulate the process of gradual transformation from a simple initial state to a complex final state, often leveraging probabilistic frameworks. In this study, we repurposed the diffusion model for imputing methylation values.

In imputation tasks, the diffusion model starts with incomplete data and iteratively refines it, gradually filling in the missing values

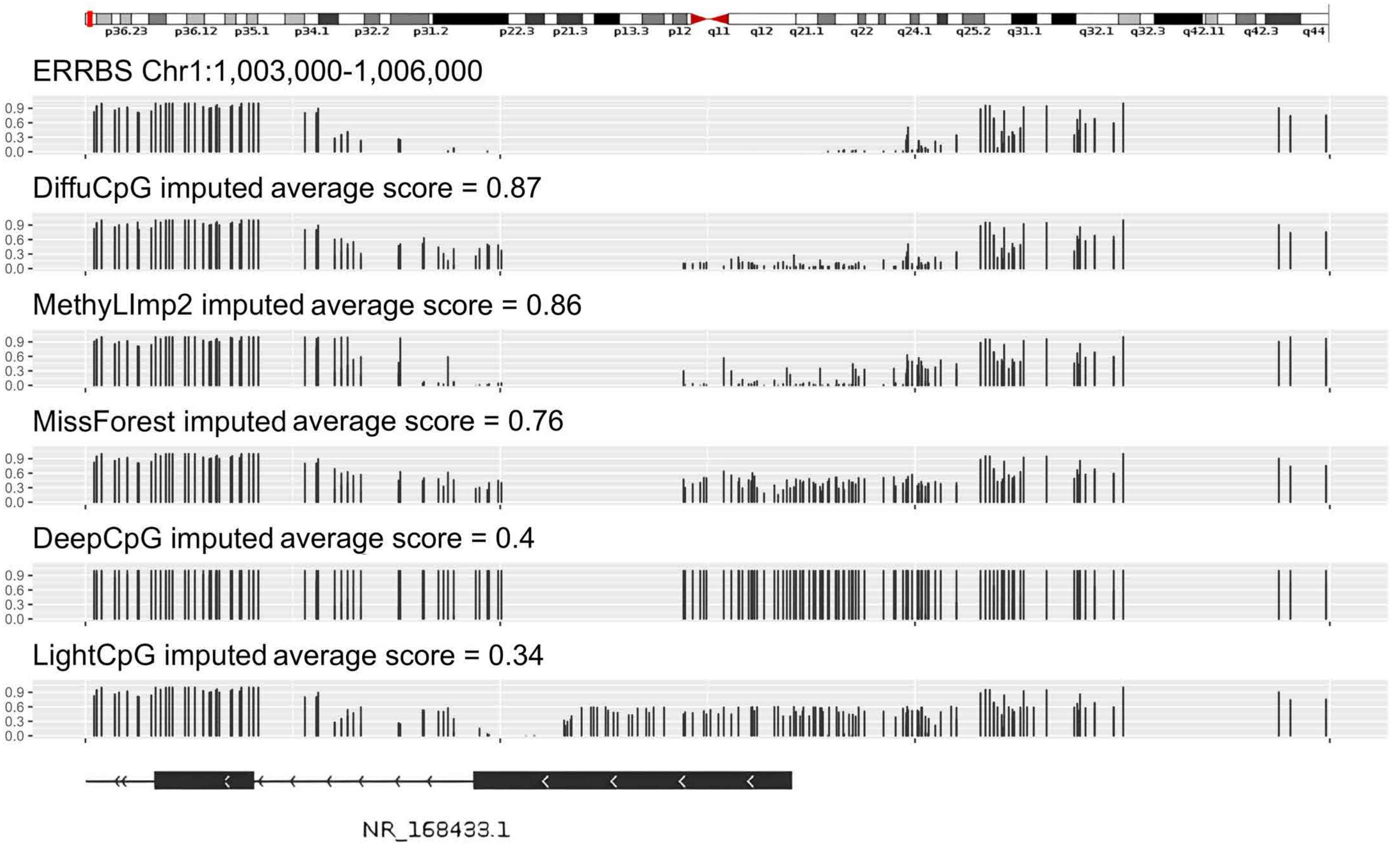

Fig. 3. Track on a CpG dense region on chromosome 1 from position 1,003,000 to 1,006,000, covering 179 CpGs by ERRBS. The ERRBS methylation levels with 50% randomly removed are shown in the first track. The subsequent tracks display the imputed results from each imputation tool after randomly removing 50% of methylation levels, where greater similarity to the first track indicates better performance. The score is the average of correlation, F1, and balanced accuracy. DiffuCpG achieved the highest overall score of 0.8. DeepCpG, which only supports binary imputation, shows uniform line heights.

with plausible content that aligns with the overall data structure. This ensures that the imputed data are consistent with the observed portions while maintaining the natural variability and complexity of the original dataset. At the core of the diffusion model is a U-Net, originally designed for biomedical image segmentation. The U-Net uses a symmetric encoder-decoder structure with skip connections, facilitating the precise reconstruction of images from their latent representations. This architecture excels in tasks requiring detailed localization and high-quality image segmentation. In the diffusion framework, the U-Net serves as the denoising model, leveraging its architectural strengths to achieve enhanced image denoising and generation. The combination of diffusion model and U-Net is particularly promising for applications like free-form inpainting.

The nature of missing data in bisulfite sequencing experiments is often random, making its imputation analogous to regenerating masked regions in an image inpainting process, specifically free-form inpainting. Free-form inpainting involves adding new content to an image in regions specified by an arbitrary binary mask. Diffusion models have shown notable advantages in free-form inpainting compared to other generative models, such as generative adversarial networks. Given the similarity between methylation imputation and free-form image inpainting, we hypothesize that the diffusion model can achieve superior imputation performance for methylation data.

A thorough comparison with four other tools [MissForest (*17*), LightCpG (*12*), MethyLImp2 (*18*), and DeepCpG (*13*)] demonstrated the superior performance of DiffuCpG in terms of the number of imputable CpGs, balanced accuracy, correlation, F1 score, and RMSE. This improved performance can be attributed to the model's superiority and additional features. The diffusion model offers several advantages over traditional and other deep learning models in biomedical research, including a better capability to capture nonlinear and nonnumerical patterns, greater robustness in dealing with noise, improved generalization and scalability across datasets, and superior representation of underlying data distribution. These characteristics make the diffusion model a more realistic and reliable data generation and imputation tool. Another reason for the better performance of DiffuCpG can be attributed to the additional features. Previous attempts are methylation imputation focused. Additional independent validations were conducted: cross-tissue (lung, breast, liver, prostate, and skin), cross-disease (lung, breast, liver, prostate, and skin cancers), and cross-technologies (ERRBS, scRRBS, and 450K). These results show that DiffuCpG is applicable on a much broader scale.

This study underscores the intricate regulatory dynamics of DNA methylation, influenced by key genomic features such as CpG interactions, genetic variations, and 3D genomic structures. The development of DiffuCpG signifies a leap forward in methylation analysis, demonstrating superior performance across diverse genomic

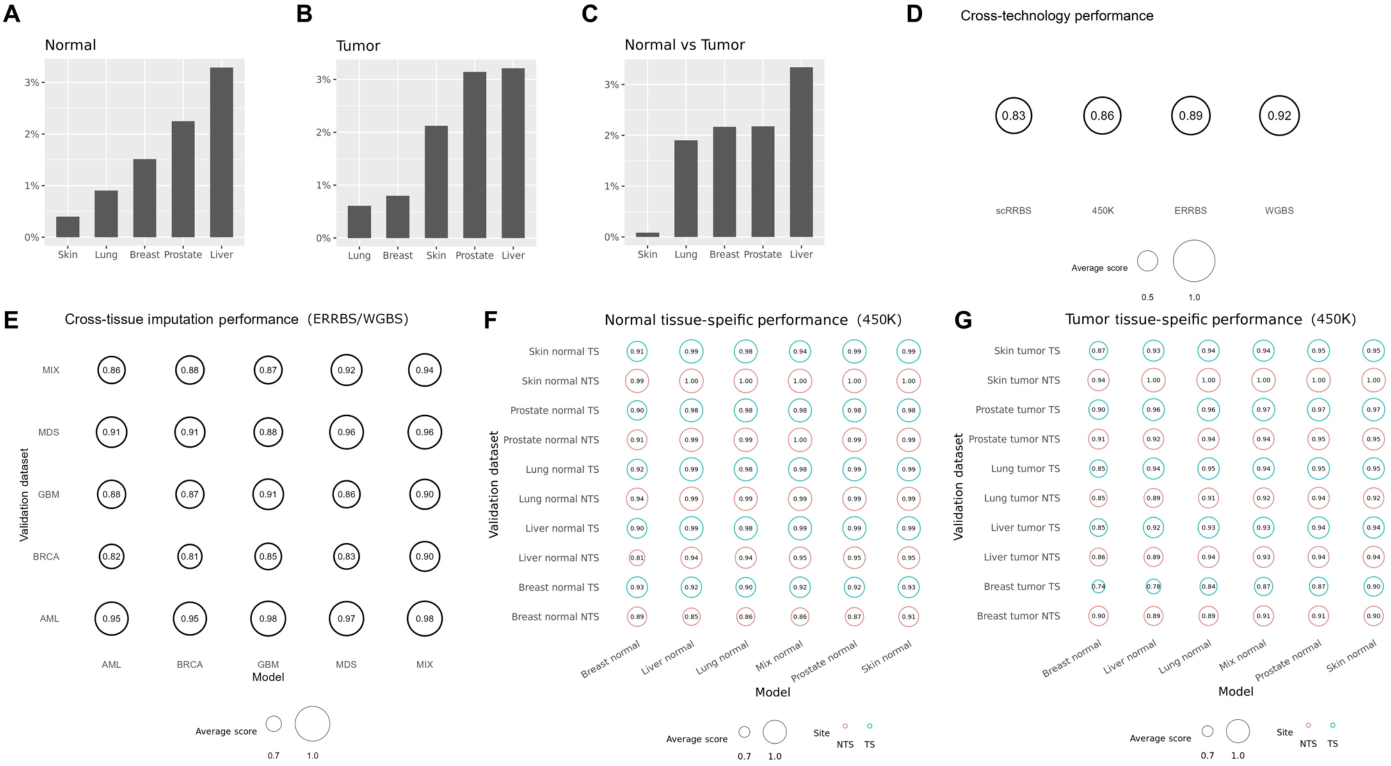

Fig. 4. Model performance and tissue specificity. (**A**) Bar plots showing the percentage of CpGs that are tissue specific. (**B**) Bar plots showing the percentage of CpGs that are cancer specific. (**C**) Bar plots showing the percentage of CpGs that are differentially methylated between tumor and normal tissues within matching organs. (**D**) DiffuCpG performed well across different technologies. The model trained on WGBS was used as a training model. (**E**) Cross-tissue performance of DiffuCpG in both tissue-specific and tissue-mixed settings. (**F**) DiffuCpG performed well for both TS-CpGs and NTS-CpGs for normal tissues. (**G**) DiffuCpG performed well for both TS-CpGs and NTS-CpGs for tumor tissues.

contexts and datasets. Its success underscores its potential to advance methylation research and broaden applications in biomedical and clinical settings. In summary, DiffuCpG is a cutting-edge methylation imputation tool that harnesses generative AI to tackle challenges associated with missing methylation data with exceptional accuracy and versatility.

METHODS

Datasets

We used 14 datasets for this study, including 5 original datasets (three bisulfite sequencing and two Hi-C) and 9 public datasets [five 450K (*28*–*32*), one RRBS (*33*), one Hi-C (*34*), one scRRBS (*35*), and one ERRBS (*11*, *36*)]. The methylation datasets cover five tissue types, eight diseases, and five types of methylation technology. The exact sample size, accession numbers, and usage (training or testing) are available in table S1.

AML and MDS patient samples

Clinical bone marrow or peripheral blood specimens were collected from deidentified AML specimens and deidentified MDS specimens. In all cases, mononuclear cells (MNCs) had been isolated through Ficoll density centrifugation at the time of diagnosis and frozen for later use. Genomic DNA was isolated from MNCs for AML cases using the AllPrep DNA/RNA kit from Qiagen (Valencia, CA) according to the manufacturer's instructions. For MDS cases, CD34+ cells were isolated from human bone marrow samples using the Miltenyi MACS magnetic bead purification system and then used for DNA extraction using the AllPrep kit. Institutional Review Board approval was obtained at the University of Miami Miller School of Medicine. Written informed consent was obtained from all patients at the time of collection.

WGBS experiment

DNA was submitted for WGBS. The Methyl-Seq Swift Kit was used for library preparation using 75 ng of input DNA. Libraries were sequenced on a NovaSeq Illumina's system. Adapters were trimmed from fastq files with cutadapt (v. 1.18) with parameters -U10 and -m25. Reads were then mapped on GRCh37 with bismark and bismark_methylation_extractor (v. 0.22.1). Mapped data were destranded and were filtered for coverage to keep CpGs with coverage between 10 and 400 reads.

ERRBS experiment

The ERRBS was performed according to standard procedures (*37*). Before adapter ligation, the methylated adapters were diluted to 150 nM. The gel size selection step targeted DNA fragments within the 150- to 450-bp range. The prepared libraries were then sequenced using an Illumina HiSeq 3000 platform. Sequencing reads were trimmed with Trim Galore (version 0.6.2) and aligned to a bisulfite-converted human genome (hg19) using Bowtie2 and Bismark (versions 2.4.1 and 0.23.1, respectively). The aligned reads were then collapsed by strand and filtered based on coverage, keeping only loci with read counts between 10 and 400.

Hi-C experiment

Hi-C experiments were conducted in duplicate using 1 million freshly thawed CD34+ cells (*2*). The Arima-HiC Kit (Arima Genomics, A510008) was used according to the manufacturer's instructions for low-input cross-linking and library preparation, using the Accel-NGS 2S Plus DNA Library Kit (Swift Biosciences, 21024). Subsequently, the libraries were subjected to paired-end sequencing on a NovaSeq 6000 platform. Hi-C contacts were called with Juicer (version 1.22.01). Additional information regarding the Hi-C experiment can be found in our previous publication (*2*).

U-Net

The U-Net (*24*) architecture used in our experiments closely resembles the original U-Net model introduced within the diffusion framework. However, we have made specific modifications to ensure optimal performance on our datasets, which predominantly consist of multichannel 1D samples, such as DNA sequences and methylation arrays. Our customized model includes several key components: a time stamp Multilayer Perceptron (MLP) module for encoding temporal information, a convolution module for processing the initial input, and a downsampling module featuring ResNet submodules. In addition, the architecture incorporates a middle block with two ResNet submodules and an attention layer, enhancing the model's ability to capture complex patterns. This is complemented by an upsampling module that mirrors the downsampling module and a final convolution module to refine the output. All ResNet and convolution modules are modified to accept the multichannel, 1D input. Our U-Net model accepts input with five channels, as depicted in Fig. 2A. The first four channels are one-hot encodings of the sequence information, whereas the fifth channel represents the methylation array. Furthermore, additional information can be incorporated by adding more channels, such as cross-sample statistics. We will delve into more details of methylation data preparation in later sections.

Denoising diffusion probabilistic models

The diffusion model operates through a dual-process mechanism comprising a forward diffusion process and a reverse denoising process. The U-Net model is integral to the denoising phase, performing the critical task of noise reduction. Collectively, these processes define the latent space of the diffusion model. The underlying concept is that the diffusion model takes an input and progressively adds Gaussian noise to it over a series of T steps during the forward diffusion process. The model then learns to denoise the resulting noisy data by reversing these T steps in the denoising process. Training focuses on the U-Net model, which, over N epochs (where $N > T$), learns to effectively remove the noise and reconstruct the original input from the random noise. At the culmination of the forward diffusion process, the output is essentially random Gaussian noise. Through extensive training, the U-Net model acquires the capability to denoise this output, thus reproducing the input from seemingly random noise.

The forward diffusion process can be expressed as the following equation

$$q\left(X_t|X_{t-1}\right)=N\left(X_t;\mu_t=\sqrt{1-\beta_t}X_{t-1},\ \sum t=\beta_t I\right) \tag{1}$$

If we reparametrize using the following

$$a_t = 1-\beta_t, \bar{a}_t = \prod_{s}^{t} a_s \tag{2}$$

Then, the reparametrized form of forward diffusion process is defined as

$$X_t \sim q\left(X_t|X_0\right)=N[X_t;\sqrt{\bar{a}_t}X_0,\left(1-\bar{a}_t\right)I] \tag{3}$$

In the equation above, β_t is a variance scheduler and can be computed at each diffusion step t, I is the identity matrix, X_0 is the initial input, and X_t is the input at step t of the diffusion process. This equation allows us to calculate the noised input at step t directly without iterative calculate all the steps during the forward diffusion process. N stands for Gaussian distribution.

The reverse diffusion process can be defined as

$$p_\theta\left(X_{t-1}|X_t\right)=N[X_{t-1};\mu_\theta\left(X_t,t\right),\ \sum\theta\left(X_t,t\right)] \tag{4}$$

The U-Net model will be trained to simulate the Gaussian distribution by predicting μ_θ and $\sum\theta$, which represents the mean and variance in a Gaussian distribution, respectively. Further investigation revealed that $\sum\theta$ variance can be kept the same throughout the reverse diffusion processes and only μ_θ mean needs to be predicted at each step t.

Instead of training our U-Net to predict the mean μ_θ, which requires us to optimize a variance lower bound, a reparametrized mean μ_θ allowed us to train the U-Net to predict the noise at each step t, and then the U-Net model can be trained using an MSE loss function; the final objective function can be written as follows

$$\|\epsilon-\epsilon_\theta\left(X_t,t\right)\|^2=\left\|\epsilon-\epsilon_\theta\left[\sqrt{\bar{a}_t}X_0+\sqrt{\left(1-\bar{a}_t\right)}\epsilon,t\right]\right\|^2 \tag{5}$$

In the equation above, ϵ represents the U-Net trained, at each step t; the U-Net can now be optimized using MSE loss function, and the U-Net will predict the noise instead of the Gaussian distribution. Through a denoising algorithm, we can denoise the input at each step t for a total of T steps to regenerate the original input. After the U-Net ϵ is trained, the mean μ_θ can be computed as follows

$$\tilde{\mu}_\theta\left(X_t,t\right)=\frac{1}{\sqrt{a_t}}\left[X_t-\frac{\beta_t}{\sqrt{1-\bar{a}_t}}\epsilon_\theta\left(X_t,t\right)\right] \tag{6}$$

Training and inpainting

The training process of a diffusion model is both systematic and intuitive. In Algorithm 1, it begins with the random selection of a timestep t from the range $\{1\ldots, T\}$. At this selected step t, Gaussian noise is added to the input data according to predefined equations (Eqs. 1 and 3). Subsequently, a U-Net architecture is used to predict the added noise, and the mean squared error (MSE) between the predicted noise and the actual noise is calculated. This process is iteratively repeated, with the model parameters being updated until the loss converges to a satisfactory level. Our training process closely resembles the original training algorithm, except that our input is five-channel, 1D data, as opposed to the original three-channel, 2D image data. Training is conducted using complete methylation arrays only.

```
repeat
    $X_0 \sim q(X_0)$
    $t \sim Uniform(\{1, \ldots, T\})$
    $\epsilon \sim N(0, I)$
    $gradient\_descent(unet, \nabla_\theta \|\epsilon - \epsilon_\theta(\sqrt{\bar{a}_t}X_0 + \sqrt{(1 - \bar{a}_t}\epsilon, t)\|^2)$
until converged
```

Algorithm 1. Diffusion training.

After training the diffusion model, we can leverage the U-Net architecture to perform various tasks such as generation, where the input is random noise, or inpainting, where the input is partially complete data. Specifically, our goal is to generate missing data points in a methylation array. This task closely resembles free-form inpainting in image generation. Drawing inspiration from the inpainting algorithm (*38*) using a U-Net model, we have adapted and modified the algorithm to suit our specific task of methylation imputation. Below is a tabular representation of our customized algorithm for this purpose. The input $X_{missing}$, $M_{missing}$ stands for methylation array with missing data, missing data mask, and inversed missing data mask.

In Algorithm 2, as also depicted in Fig. 2A, the inpainting (imputation) process unfolds through a series of structured steps. Initially, the input data with missing values are subjected to Gaussian noise according to Eq. 3. Subsequently, within a loop of T steps, the model iteratively denoises the input using Eq. 6. During each iteration, the partially denoised missing portion of input is recombined with existing values to ensure the model infers the missing values from the patterns present in the existing data. This iterative denoising and restructuring compel the model to generate predictions that are consistent with the surrounding data. Last, after T steps, the input with missing values is returned as the output, with the previously missing values now filled in with predicted values that harmonize with the existing data patterns.

Our primary contribution in Algorithm 2 is that, after every step of denoising using Eq. 4, we combined the partially denoised missing data with the original data before starting the next step. This approach encourages the model to generate missing data that are consistent with the surrounding patterns, thereby enhancing the accuracy and coherence of the imputed values. Through this iterative refinement, our method effectively achieves the goal to accurately impute missing data points, leveraging the inherent structure and relationships within the data.

In our training samples, we use five channels as depicted in Fig. 2A, with the first four channels containing sequence information and the fifth channel representing methylation data. During the training of our U-Net model, we observed that the model also attempts to impute the sequence data in the first four channels. However, our primary focus is on the imputation of the methylation data encoded in the fifth channel. To address this, we modified the loss function (Eq. 7) to assign greater weight to the fifth channel. This adjustment ensures that the model prioritizes the accurate imputation of methylation data over the sequence data, aligning the training process with our specific objective of enhancing methylation data imputation. The new loss function is as follows, where M_{seq} stands for the sequence channels mask, and M_{methy} stands for the methylation channel mask

$$loss = 0.9 \times \left[\|\epsilon - \epsilon_\theta(X_t, t)\|^2 \times M_{seq}\right] + 0.1 \times \left[\|\epsilon - \epsilon_\theta(X_t, t)\|^2 \times M_{methy}\right] \tag{7}$$

The training process follows the original training procedure (*24*), except that the diffusion steps T is set to 2000 to have the best outcome according to our tests; the number of training epochs is also set to 2000 to adequately accommodate for large diffusion steps T.

```
Input: $X_{missing}$, $M_{missing}$
$\sim M_{missing} = !\, M_{missing}$
$z \sim N(0, I)$
$t \sim T$
$X_t \sim q(X_{missing})$
for $t = T, \ldots, 1$ do
    $z \sim N(0, I)$ if $t > 1$, else $z = 0$
    $X_{t-1} = \frac{1}{\sqrt{a_t}}\left(X_t - \frac{1 - a_t}{\sqrt{1 - \bar{a}_t}}\epsilon_\theta(X_t, t)\right) + \sigma_t z$
    $X_{t-1} = X_{t-1} * M_{missing} + X_{missing} * \sim M_{missing}$
return $X_0$
```

Algorithm 2. Diffusion inpainting (imputation).

Channels and features

As previously mentioned, an input sample for the U-Net model used by diffusion consists of five-channel, 1D data. The first four channels contain sequence data encoded using one-hot encoding, whereas the fifth channel represents the methylation array. Each position in the methylation array corresponds to a base pair. If a position is not a CpG site—meaning it cannot be methylated—the value at that position is set to −1. Similarly, if the methylation state of a CpG site is missing, the value is also set to −1. To ensure the integrity of the training process and the accuracy of the imputation, only complete samples (those with no missing positions) are used. This approach guarantees that the diffusion model can generate a value for each CpG location during the imputation process. The length for all channels is set to 1000 bp (1 kb); this is also tested extensively to have the best imputation performance.

Additional features are added to channels to enhance imputation performance, specifically long-range interaction matrix, Hi-C interaction matrix, and cross-sample CIs. As depicted in Fig. 2A, both long-range interaction and Hi-C data can be incorporated into an input sample by identifying a pair of interacting base pair regions and concatenating one after another to form a five-channel, 2-kb sample. We also calculated the 95% CI across samples for each CpG site and incorporated them into the input samples. Because, for each CpG site, CI generated three measurements, in total, three additional channels are added to the sample bring the total number of channels to 8. For each CI channel, if a location is not a CpG site, the value is set to −1. To keep the training consistent, we moved the methylation array to the last channel. The final channel configuration looks like the following: (1) nucleotide A one-hot encoding, (2) nucleotide T one-hot encoding, (3) nucleotide C one-hot encoding, (4) nucleotide G one-hot encoding, (5) 95% CI lower bound, (6) 95% CI upper bound, (7) 95% CI, and (8) methylation states.

Because we are not interested in imputing first seven channels, they are mainly used as supporting information; we assigned lower loss weights to the first seven channels. We are more interested in imputation of the methylation states and assigned higher loss weights to the last channel. This is done through Eq. 7.

Model validation

A series of independent cross-tissue, cross-disease, and cross-technology validations was conducted. For each validation, 50% of the CpGs' methylation value was removed and then imputed back. Three metrics were used to measured performance: balanced accuracy, F1 score, and correlation. Balanced accuracy is a metric commonly used to evaluate classification models, particularly in datasets with imbalanced class distributions. It is defined as the average of recall (sensitivity or true-positive rate) calculated for each class, ensuring that the performance of all classes is weighted equally, regardless of their prevalence in the dataset.

In the context of classifying methylation states, there are two categories: methylated (represented by 1) and unmethylated (represented by 0). Our statistical analysis revealed that the methylated state (1) is significantly more prevalent than the unmethylated state (0), with a ratio of ~8:1 in both the AML and MDS datasets. This imbalance highlights the importance of using balanced accuracy as a more suitable metric for evaluating the classification performance of our model as it accounts for disparities in class distribution. The general formula for balanced accuracy can be expressed as follows

$$\frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i + FN_i}$$

where C is the number of classes, TP is the calculation of the true-positive number, and FN is the calculation of the false-negative number. We have included this information in our latest revision.

The F1 score is a measure used to evaluate the accuracy of a binary classification model. It combines two important metrics: precision and recall. The F1 score is the harmonic mean of precision and recall, providing a single metric that balances both. Correlation is Pearson's correlation coefficient between measured methylation values (bisulfite sequencing or methylation array) and DiffuCpG imputed methylation values. In the study, we use the average score, which is the average of balanced accuracy, F1 score, and correlation to represent the overall performance of a model.

Supplementary Materials

This PDF file includes:

Table S1

REFERENCES AND NOTES

1. Z. Siegfried, I. Simon, DNA methylation and gene expression. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **2**, 362–371 (2010).
2. A. G. Telonis, Q. Yang, H. T. Huang, M. E. Figueroa, MIR retrotransposons link the epigenome and the transcriptome of coding genes in acute myeloid leukemia. *Nat. Commun.* **13**, 6524 (2022).
3. D. A. Khavari, G. L. Sen, J. L. Rinn, DNA methylation and epigenetic control of cellular differentiation. *Cell Cycle* **9**, 3880–3883 (2010).
4. V. Colot, L. Maloisel, J. L. Rossignol, DNA repeats and homologous recombination: A probable role for DNA methylation in genome stability of eukaryotic cells. *J. Soc. Biol.* **193**, 29–34 (1999).
5. E. L. Greer, Y. Shi, Histone methylation: A dynamic mark in health, disease and inheritance. *Nat. Rev. Genet.* **13**, 343–357 (2012).
6. M. Kulis, M. Esteller, DNA methylation and cancer. *Adv. Genet.* **70**, 27–56 (2010).
7. A. G. Telonis, D. A. Rodriguez, P. M. Spanheimer, M. E. Figueroa, N. Goel, Genetic ancestry-specific molecular and survival differences in admixed patients with breast cancer. *Ann. Surg.* **279**, 866–873 (2024).
8. S. Younesian, A. M. Yousefi, M. Momeny, S. H. Ghaffari, D. Bashash, The DNA methylation in neurological diseases. *Cells* **11**, 3439 (2022).
9. E. R. Adelman, H. T. Huang, A. Roisman, A. Olsson, A. Colaprico, T. Qin, R. C. Lindsley, R. Bejar, N. Salomonis, H. L. Grimes, M. E. Figueroa, Aging human hematopoietic stem cells manifest profound epigenetic reprogramming of enhancers that may predispose to leukemia. *Cancer Discov.* **9**, 1080–1101 (2019).
10. S. Fan, C. Li, R. Ai, M. Wang, G. S. Firestein, W. Wang, Computationally expanding infinium HumanMethylation450 BeadChip array data to reveal distinct DNA methylation patterns of rheumatoid arthritis. *Bioinformatics* **32**, 1773–1778 (2016).
11. F. E. Garrett-Bakelman, C. K. Sheridan, T. J. Kacmarczyk, J. Ishii, D. Betel, A. Alonso, C. E. Mason, M. E. Figueroa, A. M. Melnick, Enhanced reduced representation bisulfite sequencing for assessment of DNA methylation at base pair resolution. *J. Vis. Exp.* **96**, e52246 (2015).
12. L. Jiang, C. Wang, J. Tang, F. Guo, LightCpG: A multi-view CpG sites detection on single-cell whole genome sequence data. *BMC Genomics* **20**, 306 (2019).
13. C. Angermueller, H. J. Lee, W. Reik, O. Stegle, DeepCpG: Accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol.* **18**, 67 (2017).
14. S. Villicana, J. Castillo-Fernandez, E. Hannon, C. Christiansen, P. C. Tsai, J. Maddock, D. Kuh, M. Suderman, C. Power, C. Relton, G. Ploubidis, A. Wong, R. Hardy, A. Goodman, K. K. Ong, J. T. Bell, Genetic impacts on DNA methylation help elucidate regulatory genomic processes. *Genome Biol.* **24**, 176 (2023).
15. A. Monteagudo-Sanchez, D. Noordermeer, M. V. C. Greenberg, The impact of DNA methylation on CTCF-mediated 3D genome organization. *Nat. Struct. Mol. Biol.* **31**, 404–412 (2024).
16. P. Di Lena, C. Sala, A. Prodi, C. Nardini, Methylation data imputation performances under different representations and missingness patterns. *BMC Bioinformatics* **21**, 268 (2020).
17. P. Di Lena, C. Sala, A. Prodi, C. Nardini, Missing value estimation methods for DNA methylation data. *Bioinformatics* **35**, 3786–3793 (2019).
18. A. Plaksienko, P. Di Lena, C. Nardini, C. Angelini, methyLImp2: Faster missing value estimation for DNA methylation data. *Bioinformatics* **40**, btae001 (2024).
19. J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, B. Ren, Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
20. S. S. P. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, E. L. Aiden, A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
21. M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, G. K. Smyth, limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
22. X. Wu, W. Li, H. Tu, Big data and artificial intelligence in cancer research. *Trends Cancer* **10**, 147–160 (2024).
23. L. Messeri, M. J. Crockett, Artificial intelligence and illusions of understanding in scientific research. *Nature* **627**, 49–58 (2024).
24. J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **33**, 6840–6851 (2020).
25. B. Xia, Y. L. Zhang, S. Y. Wang, Y. T. Wang, X. L. Wu, Y. P. Tian, W. M. Yang, L. Van Gool, DiffIR: Efficient diffusion model for image restoration, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (IEEE, 2023), pp. 13049–13059.
26. R. Durall, A. Ghanim, M. R. Fernandez, N. Ettrich, J. Keuper, Deep diffusion models for seismic processing. *Comput. Geosci.* **177**, 105377 (2023).
27. M. Y. Zhang, Z. G. Cai, L. Pan, F. Z. Hong, X. Y. Guo, L. Yang, Z. W. Liu, MotionDiffuse: Text-driven human motion generation with diffusion model. *IEEE Trans. Pattern Anal. Mach. Intell.* **46**, 4115–4128 (2024).
28. Cancer Genome Atlas Network, Genomic classification of cutaneous melanoma. *Cell* **161**, 1681–1696 (2015).
29. Cancer Genome Atlas Network, Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
30. Cancer Genome Atlas Research Network, Comprehensive and integrative genomic characterization of hepatocellular carcinoma. *Cell* **169**, 1327–1341.e23 (2017).
31. Cancer Genome Atlas Research Network, The molecular taxonomy of primary prostate cancer. *Cell* **163**, 1011–1025 (2015).
32. Cancer Genome Atlas Research Network, Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
33. T. M. R. Noviello, A. M. Di Giacomo, F. P. Caruso, A. Covre, R. Mortarini, G. Scala, M. C. Costa, S. Coral, W. H. Fridman, C. Sautes-Fridman, S. Brich, G. Pruneri, E. Simonetti, M. F. Lofiego, R. Tufano, D. Bedognetti, A. Anichini, M. Maio, M. Ceccarelli, Guadecitabine plus ipilimumab in unresectable melanoma: Five-year follow-up and integrated multi-omic analysis in the phase 1b NIBIT-M4 trial. *Nat. Commun.* **14**, 5914 (2023).
34. E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, J. Dekker, Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
35. Y. Hou, H. Guo, C. Cao, X. Li, B. Hu, P. Zhu, X. Wu, L. Wen, F. Tang, Y. Huang, J. Peng, Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Res.* **26**, 304–319 (2016).

36. S. Li, F. E. Garrett-Bakelman, S. S. Chung, M. A. Sanders, T. Hricik, F. Rapaport, J. Patel, R. Dillon, P. Vijay, A. L. Brown, A. E. Perl, J. Cannon, L. Bullinger, S. Luger, M. Becker, I. D. Lewis, L. B. To, R. Delwel, B. Lowenberg, H. Dohner, K. Dohner, M. L. Guzman, D. C. Hassane, G. J. Roboz, D. Grimwade, P. J. Valk, R. J. D'Andrea, M. Carroll, C. Y. Park, D. Neuberg, R. Levine, A. M. Melnick, C. E. Mason, Distinct evolution and dynamics of epigenetic and genetic heterogeneity in acute myeloid leukemia. *Nat. Med.* **22**, 792–799 (2016).

37. A. Akalin, F. E. Garrett-Bakelman, M. Kormaksson, J. Busuttil, L. Zhang, I. Khrebtukova, T. A. Milne, Y. Huang, D. Biswas, J. L. Hess, Base-pair resolution DNA methylation sequencing reveals profoundly divergent epigenetic landscapes in acute myeloid leukemia. *PLoS Genet.* **8**, e1002781 (2012).

38. A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, L. Van Gool, RePaint: Inpainting using denoising diffusion probabilistic models, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2022), pp. 11461–11471.

Acknowledgments: This study was supported by the Biostatistics and Bioinformatics Shared Resource (SCR022890) and the Onco-Genomics Shred Resource of Sylvester (SCR022502) of Sylvester Comprehensive Cancer Center, University of Miami. We acknowledge the Research Computing at The University of Virginia for providing computational resources (https://rc.virginia.edu). **Funding:** This study was supported by funding R01ES030993 and P30CA240139 from the National Cancer Institute, USA. M.E.F. was partially supported by a Leukemia & Lymphoma Society Specialized Center of Research award (7028-22). L.M. was supported by BCRF and R01GM141349 and R01GM146409 from the National Institute of General Medical Sciences, USA. V.S. was supported by Associazione Italiana Ricerca sul Cancro (AIRC) IG-26537-2021. F.G.-B. was partially supported by the UVA Cancer Center through the NCI Cancer Center Support Grant P30CA44579. **Author contributions:** Conceptualization: Y.G., F.Y., M.A.S., M.E.F., N.G., and L.J. Investigation: V.S., L.G.-M., A.T., F.Y., and F.G.-B. Methodology: M.C., Y.G., F.Y., M.E.F., L.G.-M., and L.J. Resources: V.S., M.C., Q.Y., Y.G., M.E.F., L.M., and N.G. Data curation: V.S., M.C., Q.Y., Y.G., A.T., F.Y., M.E.F., N.G., and F.G.-B. Formal analysis: M.C., A.T., and F.Y. Software: Q.Y., A.T., and F.Y. Validation: V.S., Q.Y., Y.G., L.G.-M., F.Y., M.E.F., and F.G.-B. Visualization: Q.Y., Y.G., F.Y., and L.J. Supervision: V.S., Y.G., M.E.F., and N.G. Project administration: Y.G. Funding acquisition: V.S. and Y.G. Writing—original draft: Y.G., F.Y., M.A.S., and M.E.F. Writing—review and editing: V.S., Q.Y., Y.G., A.T., F.Y., M.A.S., M.E.F., N.G., and F.G.-B. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. The DiffuCpG's tissue-specific models and tissue-mixed models in Docker can be found on zenodo.org with record ID 14622375 (DOI: 10.5281/zenodo.14622375). The DiffuCpG source code is also available to download from the same website or GitHub (https://github.com/yay135/DiffuCpG) and can be used to train additional models from other tissues. Other datasets and their accession numbers are available in table S1.

Submitted 27 September 2024
Accepted 5 March 2025
Published 11 April 2025
10.1126/sciadv.adt4152