# DMRU: Generative Deep-Learning to unravel condition specific cytosine methylation in plants

Sagar Gupta[1,2], Anchit Kumar[1], Veerbhan Kesarwani[1,2], Ravi Shankar*[1,2]

[1] Studio of Computational Biology & Bioinformatics,
The Himalayan Centre for High-throughput Computational Biology,
(HiCHiCoB, A BIC supported by DBT, India), Biotechnology Division,
CSIR-Institute of Himalayan Bioresource Technology (CSIR-IHBT),
Palampur (HP), 176061, India.

[2] Academy of Scientific and Innovative Research (AcSIR),
Ghaziabad, Uttar Pradesh- 201002

Authors' email addresses:
SG: sagargupta47321@gmail.com
AK: ranaanchit34@gmail.com
VK: veerbhan8586@gmail.com
RS: ravish@ihbt.res.in

*To whom correspondence should be addressed- Email-ravish@ihbt.res.in;

# Abstract

Epigenetic regulation of gene expression is a complex and multifaceted process that plays a crucial role in plant development and adaptation. DNA methylation at cytosines has been shown to influence gene expression by regulating chromatin structure and accessibility. Some approaches have been developed to identify DNA methylation but none of them are capable to tell the condition specific DNA methylation, making them hardly of any use. Here, we report a first of its kind an explainable Deep Encoders-Decoders generative system, DMRU, which learns the relationship between transcritpome status and DNA methylation states at any given time. It was also found that instead of looking for homology to relate condition specificity of DNA methylation marks across species, GC similarity is more relevant to the specificity of DNA methylation patterns. The developed approach, DMRU, is capable to discover the methylated cytosines in total independence from data volume and species specific models, and works with same performance across different species in universal manner. In a comprehensive benchmarking study across a huge volume of experimental data covering different conditions and species, it has consistently achieved >90% accuracy. With this all, DMRU brings a completely new chapter in methylated cytosine discovery through generative AI, giving a strong alternative to costly bisulfite sequencing experiments, saving enormous amount of time and money. DMRU may have significant impact on the development of novel strategies for crop improvement and biotechnology applications, while accelerating the research in the field of epigenomics and gene regulation. DMRU is freely available as a web-server at https://scbb.ihbt.res.in/DMRU/index.php.

**Keywords:** DNA methylation, Epigenetic regulation, Deep Learning, DenseNet, Transformers, Encoder-Decoder

# Introduction

DNA methylation is a crucial epigenetic mechanism that plays a key role in regulating gene expression and ensuring genomic stability across various biological contexts **[1]**. It involves the addition of methyl groups to specific DNA nucleotides (adenine and cytosine), primarily targeting cytosine residues in plants. The most common form of this modification is the conversion of cytosine to 5-methylcytosine (5mC) **[2].** This modification is not static; it dynamically responds to a multitude of factors, including various environmental and developmental influences. These factors can lead to changes in the methylation landscape, affecting how genes are expressed. In addition to 5mC, there are other important derivatives of DNA methylation in plants. For instance, 5-hydroxymethylcytosine (5hmC) is formed from the oxidation of 5mC and is often associated with active gene transcription **[3]**. This intermediate plays a crucial role in active DNA demethylation, where it can be further oxidized to form 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC), ultimately leading to the replacement of the methylated cytosine with an unmethylated one **[4-6]**. In addition to cytosine methylation, N6-methyladenine (6mA) **[7-8]** has been identified as an important modification in both prokaryotes and some eukaryotes **[9-12]**.

DNA methylation can be investigated through advanced techniques like next-generation sequencing (NGS). Among these, whole-genome bisulfite sequencing (WGBS) **[13]** and reduced-representation bisulfite sequencing (RRBS) **[14]** stand out as powerful methods. WGBS is a comprehensive method that provides a complete overview of DNA methylation across the entire genome. In this technique, genomic DNA is treated with sodium bisulfite, which converts unmethylated cytosines to uracils while leaving methylated cytosines unaffected **(Figure 1a)**. This conversion allows researchers to sequence the treated DNA and directly infer the methylation status at individual nucleotide positions. RRBS offers a more targeted approach, focusing on specific regions of the genome, especially CpG-rich areas where methylation is most prevalent. This technique begins with the digestion of the genomic DNA using restriction enzymes, followed by bisulfite treatment and sequencing. While it may not cover the entire genome, RRBS provides high-resolution data for specific loci, making it a cost-effective alternative to WGBS. The ability to focus on regions of interest is beneficial for studying particular genes or regulatory elements associated with traits or diseases. Both WGBS and RRBS generate data are invaluable for comparative amd regulatory studies, allowing researchers to investigate variations in DNA methylation across different conditions, developmental stages, or species.

While significant progress has been made in understanding DNA methylation in animal models, particularly in humans, there is a striking disparity in plant research. The plant kingdom comprises a vast number of species, each with unique genetic and epigenetic landscapes. However, only a limited number of plant species have been sequenced for their DNA methylation states, primarily *Arabidopsis thaliana* and *Oryza sativa* **(Figure 1b)**. In fact, fewer than 10 species have been comprehensively studied, highlighting a considerable gap in our knowledge of plant methylation dynamics. This represents a stark contrast to the extensive research conducted in human. The gap in experimental data for plant DNA methylation is also mirrored in the development of computational tools and software resources **(Supplementary Table S1)**. This lack of resources hampers the ability to properly explore the DNA methylations and their impact in various conditions in the plant kingdom. As research continues to evolve, there is a pressing need to bridge this gap, fostering a more comprehensive understanding of DNA methylation in plants and its implications for biology and agriculture.

The detection of DNA methylation has traditionally relied on experimental techniques that can be both expensive and time-consuming **[15]**. WGBS costs can vary significantly depending on the service provider, sample volume, and specific project requirements and can range between approx

$2,500 to $3,300 per sample. Additionally, methylation patterns varies significantly across different organisms, which presents further challenges in understanding these epigenetic modifications. Above all, very much like the transcriptome profiling, DNA methylome profile has unlimited possible states depending upon the condition which makes it quite unpragmatic to reveal all of them experimentally. Therefore, it has become an essentiality to develop some efficient computational approach to address these issues, a tool which could accurately profile DNA methylation profile for any given condition. To address these issues, researchers are increasingly turning to computational approaches, particularly those based on machine learning (ML). These methods claim good accuracy. They usually classify DNA methylation by framing the identification of methylation sites as a binary classification problem. In this context, ML models are trained to differentiate between true methylation sites and non-methylation sites based on genomic sequences.

While the goal remains the same for these tools, the approaches taken differ significantly in terms of the features selected and the architectures employed. ML methods rely heavily on manually engineered features, which require a certain level of expertise and prior knowledge from researchers. This reliance on manual features can make these methods less adaptable across different species or contexts, as the effectiveness of features can vary significantly depending on the biological background. Input features often include: sequence encoding, a common approach where the DNA sequences are converted into numerical formats suitable for algorithmic processing. Additional features might encompass various biochemical characteristics of the DNA, such as nucleotide composition or methylation patterns **[16-17]**. Some studies employ DNA molecular graph representations to capture structural information about the DNA, allowing for more nuanced analysis **[18]**. The choice of model architecture also plays a crucial role in determining the success of these classification tasks. A variety of deep-learning (DL) models have also been used. Approaches like Convolutional Neural Networks (CNNs) are well-suited for analyzing spatial hierarchies in data as CNNs can capture local patterns in DNA sequences. Implementation of bi-modal CNN and Long short-term memory (LSTM) could also achieve better classification results by learning intricate DNA structural and sequence features **[19]**. In addition to this, Zheng et. al., 2023 **[20]** used attention mechanism with CNN and LSTM for better classification performance **[20]**. Transformers **[21]**, which has gained prominence in various domains, offers advantages in capturing long-range dependencies in sequential data **[22-24]**. Bidirectional Encoder Representations from Transformers (BERT) **[25]** and its adaptations leverage the transformer model to enhance context understanding in sequence data **[25]**. The specific choices made in feature engineering and model selection are pivotal. They not only affect the performance of the models but also highlight the need for a robust framework that can integrate various inputs and architectures.

Despite the development of numerous ML/DL based approaches for the discovery of DNA methylation sites, none of the aforementioned approaches are capable to propose condition specific DNA methylation but report mainly the possibility of any given cytosine getting methylated in any possible condition. Process of DNA methylation depends upon several genes in the background whose dynamics may decide the location specific differential methylation status **[26-31]**. Such dynamics can be learned by incorporating the transcriptome expression profiling data along with methylation profile. All of the existing tools have ignored this. The absence of gene expression data limits the precise identification of methylated cytosines under specific conditions, as DNA methylation is a dynamic process. Current software tools are heavily reliant on the provided sequences information mainly and fail to consider gene expression, rendering them largely ineffective. To address this gap, it is crucial to develop integrated approaches that learns from DNA sequence information and transcriptome expression profiles in interdependent manner.

Based on the above concept, we have developed an innovative approach called **D**NA **M**ethylation **R**ecognition **U**nit **(DMRU)**. This system integrates a concurrent bi-modal architecture that combines DenseNet and Transformer encoder-decoder system within a generative deep learning framework. It has utilized 258 WGBS-seq and 326 RNA sequencing (RNA-Seq) data specifically from the plant species *Arabidopsis thaliana* and *Oryza sativa*. The primary objective of DMRU is to understand and learn from the interdependent variability between the methylation states present in the potential promoter regions of the genes (2kb upstream) and their corresponding RNA expression levels under specific environmental conditions. This enables DMRU to identify the most promising cytosine methylation points in the given sequence, independent of any species-specific models. In essence, DMRU's design allows it to generalize across different plant species, providing a versatile tool for researchers in plant epigenetics.

The results achieved with DMRU is groundbreaking. In a rigorous benchmarking study, the system has consistently performed with an accuracy exceeding 90%, and in cross-species universal manner. This characteristic is particularly significant as it demonstrates the system's capability to generalize findings across diverse plant species and for any sequence region, thereby expanding its applicability in various research contexts. One of the distinguishing features of DMRU is its commitment to explainability in machine learning. By employing Gradient Weighted Class Activation Mapping (Grad-CAM) **[32]**, DMRU not only identify the cytosine methylation points but also provides insights into the underlying most important factors responseible for the observed methylation pattern for any given specific region. In contrast to existing software solutions, DMRU represents a pioneering effort in the realm of deep learning for plant epigenetics. It is the first auto-generative deep learning system specifically designed to decode precise cytosine methylation points for any gene of interest completely independent of any species specific model. With this all, DMRU brings a new dawn in the area of plant epigenetics research where one can accurately identify the condition specific cytosine methylation in the promoters of any plant gene. Also, DMRU emerges as a promising alternative to costly and time consuming bisulfite sequencing experiments.

# Materials and Methods

## Dataset retrieval, processing and construction

For the development of universal and generalized model for the identification of methylated cytosines, we retrieved RNA-seq and WGBS-seq data for *Arabidopsis thaliana and Oryza sativa* from NCBI Sequence Read Archive (SRA). A total of 326 RNA-seq fastq data were collected from SRA databases which comprises 196 and 130 different read files for *A. thaliana* and *O. sativa*, respectively (**Supplementary Table S2 Sheet 1-2**). A total of 45 and 40 different experimental conditions were retrieved for *A. thaliana* and *O. sativa*, respectively. Genomic sequences, annotations and reference DNA sequences were downloaded from Ensembl Plants. Trimmomatic v0.39 **[33]** and in house developed reads processing tool, filteR **[34]**, were used to filter out poor quality reads, read trimming, and for adapter removal. Filtered reads were mapped back to the genome using Hisat2 **[35]**. After mapping, read counting was done by using Rsubread **[36]**. Furthermore, the read counts were normalized to fragments per kilo base per million mapped reads (FPKM) for paired end reads and reads per kilo base per million mapped reads (RPKM) for single end reads, using in-house developed shell scripts for every single gene across different conditions.

In a manner similar to the RNA-seq, we collected a total of 258 WGBS-seq fastq data from SRA databases having 140 and 118 different read files for *A. thaliana* and *O. sativa*, respectively (**Supplementary Table S2 Sheet 3-4**). It has to be noted that, only those WGBS-seq fastq files were downloaded whose corresponding RNA-seq fastq files were available with the same experimental condition. Trimmomatic v0.39 **[33]** and filteR **[34]** were used to filter out poor quality reads, read trimming, and for adapter removal. The filtered reads were subsequently aligned to the reference genome using Bismark v0.22.2 **[37]**, retaining only uniquely aligned reads and discarding ambiguous mappings. For methylation analysis, we utilized Samtools v0.1.9 **[38]** to sort reads by genomic coordinates and remove PCR duplicates from the Bismark output, and ultimately converting this processed SAM file to BAM output. Afterwards, the output BAM files underwent methylation extraction using the Bismark methylation extractor function, with reads exhibiting a conversion rate below 90% or fewer than three non-converted cytosines in non-CpG contexts excluded from analysis. Furthermore, we employed the CX report file generated by Bismark, which provide details of the three types of methylation sites (CG, CHH, CHG). We extracted the two kilobases (kb) upstream promoter sequences for each gene of both the species from the downloaded gene transfer file (GTF) file. Additionally, using this CX report file, we marked all methylated cytosines within the promoter sequences with a distinct character **"M"** to differentiate the methylated cytosines with unmethylated cytosines. For every promoter single sequences, corresponding gene expression profiles were also considered in the development of this dataset. The dataset having sequences from *A. thaliana* has been named as Dataset "A" and the dataset having sequences from *O. sativa* has been named as Dataset "B" in the current study. Both these datasets were split in a 70:30 ratio for training and testing purposes, including for 10-folds random train:test trials. All details on data processing are illustrated in **Figure 2**. Complete list of various sources are available in **Supplementary Table S2 Sheet 1-4**.

## The Transformer-DenseNet system

We extracted gene expression data from RNA sequencing (RNA-seq) for various conditions. This dataset encompasses the expression levels of each gene across different experimental contexts, providing a rich source of information to analyze how expressions of the target gene as well as other genes also correlate with methylation states. To process this data, we developed a DenseNet

**[39]** model architecture known for its efficiency in handling vanishing gradient and retention of learnings from previous layers. We established a bi-modal architecture with DenseNet and Transformer encoder which effectively combines expression profiles and DNA methylation in the target sequence region information. This architecture enables the model to take sequences and gene expression data as input, allowing it to capture both local and global patterns in gene expression that may be linked to methylation. The expression input to the DenseNet is a vector with a size of 38,912 elements corresponding to the positions of genes holding their respective expression values for the given condition. The DenseNet architecture we developed for this study was implemented using PyTorch.

## Constructing the DenseNet architecture

The DenseNet **[39]** model developed here processes a tensor of 38,912 elements representing gene expression profiles 38,912 genes (each element corresponds to the positions of genes, each associated with its respective expression level). This architecture consisted a convolution layer with 32 convolution filters (kernel size = 3), batch normalization, and 2D max-pooling (stride = 2), and incorporates four dense blocks and three transition layers, totaling 121 layers. Dense blocks, fostering feature reuse through dense connectivity, play a crucial role. In a dense block, each layer "***l***" receives direct input from all preceding layers, producing output through operations like convolution and normalization. Transition layers effectively reduce the dimensions before the next dense block.

Expressed mathematically, let "$H_l$" represent the output feature maps of layer "*l*". The output of layer "*l*" is computed as:

$$H_l = H_{l-1} \oplus f\left(H_{l-1}, W_l\right) \tag{1}$$

Here, “ $\oplus$ ” signifies concatenation, “$W_l$” denotes the layer's weights, and “$f$” incorporates batch normalization (BN), rectified linear units (ReLU), and convolution operations.

The growth rate, denoted by the hyperparameter “$k$” in DenseNet, emerges as a pivotal determinant in the architecture's extraordinary performance. DenseNet's innovative design treats feature maps as a global network state, enabling impressive results even with a smaller growth rate. At each layer, “$k$” feature maps contribute to this global state, where the total number of input feature maps “$F_m$” at the “$l^{th}$” layer is calculated as:

$$F_m = k_c + k * (l - 1) \tag{2}$$

The term “$k_c$” denotes the total number of channels in the input layer. DenseNet consists of dense blocks, where several convolutional layers are stacked together, followed by transition layers that reduce the dimensionality and down-sample the feature maps. These transition layers incorporate batch normalization, ReLU activation, and 1x1 convolutions, making the network both efficient and manageable. To enhance computational efficiency, a 1x1 convolution layer precedes each 3x3 convolution layer. This design, known as a bottleneck layer, reduces the number of input feature maps, typically exceeding the output feature maps “$k$”, by generating 4k feature maps. This architecture's output is intertwined through a bi-modal architecture in a feed-forward layer within the transformers encoder to capture RNA expression relationship with input target sequence data to raise a bi-modal learning model to identify methylated cytosines in global and cross-species manner.

## Word representations of sequence data for Transformer

Each input sequence can be viewed as a collection of independent words or tokens. The DNA alphabet consists of four nucleotides: A, T, G, and C. By pairing these nucleotides, we can create a total of 16 unique dimers. An alphabet of the four bases with an additional character **“M”** for the methylated cytosines allows creating 3,125 unique 5-mer words, 15,625 unique 6-mer words, and 78,125 unique 7-mer words. This exponential growth in words allows for a rich representation of genetic information. In processing these sequences, overlapping windows of pentamers, hexamers, and heptamers could capture critical information about structural and functional regions within the DNA **[23-24, 40-44]**. Transformers takes two inputs i.e., source and target input. The source input consists of promoter sequences without any information of methylated cytosines with a maximum length of 5,985 words and 2,000 bases, providing a detailed representation of the DNA being analyzed. The target input typically involves promoter sequences, which include methylation information. This input has a maximum length of 1,994 words and also spans 2,000 bases. To process these sequences, each unique word is assigned a distinct integer token. This tokenization step is crucial as it converts the raw sequences into a structured format suitable for Transformers models. Once tokenized, these sequences are transformed into numeric vectors and matrices through a process known as embedding. The Transformer’s encoder processes the source sequences while the decoder generates the target sequences based on the encoded information. By feeding these embedded sequences into the model, the Transformer can learn to translate sequences effectively. This setup allows for efficient handling of large datasets and complex sequence transformations.

## Implementation of the Transformers Encoders-Decoders

The encoder component processes the input sequence by transforming it into a series of continuous representations. It does this through several layers, each consisting of multiple attention heads and feed-forward neural networks. The decoder, on the other hand, generates the output sequence based on the encoded information. It uses the context derived from the encoder's outputs and incorporates its own attention mechanisms to focus on relevant parts of the input sequence while generating each token of the output.

## The Encoder

The encoder utilizes a multi-headed attention mechanism that features self-attention layers by employing a dot-product method. These layers enable the encoder to focus on various words in the input, consider their relationships and how they relate to one another, when making decisions. This design allows the transformer model to effectively capture long-range dependencies and contextual connections within the input data. In the transformer architecture, the sequences derived from the tokenized representation of the previous step are input into the system's input layer. Each word is transformed into a word vector of a specified embedding size through the embedding process. This results in an output matrix that is structured by samples, sequence length, and embedding size, facilitating the model's ability to understand and manipulate the data.

The input sequence of word tokens is denoted as *"Z"*, with a length of *"I"*. The embedding layer projects these discrete input tokens into continuous vector embeddings. If "***v***" is the size of the vocabulary and "***d***" is the dimensionality of the word embeddings, the embedding layer is represented as a matrix $E_{embed} \in \mathbb{R}^{v \times d}$ .

$$E_{embed} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \dots \\ x_v \end{bmatrix} \begin{bmatrix} W_{11}, W_{12}, \dots, W_{1d} \\ W_{21}, W_{22}, \dots, W_{2d} \\ W_{31}, W_{32}, \dots, W_{3d} \\ \dots\dots\dots\dots\dots \\ W_{v1}, W_{v2}, \dots, W_{vd} \end{bmatrix} \tag{3}$$

Each word in matrix "$E_{embed}$" combines with its positional embedding "$P$", where "$P$" shares dimension "$d$" with the word embedding vector. The resulting matrix, "$E'_{embed}$", is obtained through " $E'_{embed} = E_{embed} + P$ ", where "$P$" is calculated using sinusoidal positional embedding equations:

$$P_{i,2j} = \sin\left(\frac{i}{10000^{2j/d_{model}}}\right) \tag{4}$$

$$P_{i,2j+1}=\cos\left(\frac{i}{10000^{2j/d_{model}}}\right) \quad (5)$$

Here, “$i$” represents the position of the token, and “$d_{model}$” is the embedding dimension. “$E'_{embed}$” matrix enters the transformer encoder block, which processes it through a multi-head attention layer. This module involves multiple heads, each dividing its query, key, and value parameters N-ways and independently processing the splits. The output from each head produces an attention score derived from the key, query, and value computations.

Initially, the model generates individual attention vectors, which are essential for capturing relationships within the input data. This step involves applying self-attention to produce contextual embeddings for each token in the sequence. The attention vectors are then concatenated and subsequently forwarded to a feed-forward network block within the encoder. To combat the risk of overfitting, dropout layers were employed. Afterwards, a normalization layer implemented as Layer Normalization follows. This normalization helps stabilize the learning process by normalizing the outputs, ensuring that the mean and variance remain consistent throughout the training. The output post-normalization is then directed to fully connected feed-forward layers and a dropout layer is introduced to prevent overfitting, followed by another normalization layer. At this stage, the output is integrated with the DenseNet output, encapsulating vital gene expression information from the RNA-seq data. The final output from this enhanced encoder structure feeds into the decoder block of the transformer.

## Decoder block of the transformer

The decoder block of the transformer architecture plays a crucial role in the overall functioning of the model, particularly in tasks involving sequence generation. Each decoder block consists of several key components: masked multi-head self-attention, multi-head attention over the encoder’s output, and position-wise feed-forward networks, all combined with residual connections and layer normalization. The second attention layer in the decoder block attends to the encoder's output, enabling the decoder to incorporate information from the input sequence. This interaction allows the model to generate contextually relevant output by aligning the generated tokens with the input representations.

**Masked Multi-Head Attention of Decoders**

The primary purpose of masking in a decoder is to prevent the model from accessing future tokens when predicting the next token in a sequence. For each distinctive head "$i$", symbolizing $i=1, 2, \ldots, h$, the calculation of three pivotal matrices ensues: "$Q_i$" (Query), "$K_i$" (Key), and "$V_i$" (Value). This is accomplished through the matrix multiplication of "$E$" with the corresponding weight matrices "$W_{qi}$", "$W_{ki}$", and "$W_{vi}$". The formulation of attention scores "$Z_i$" transpires through the application of the softmax function to the scaled dot-product attention mechanism, adhering to the expression:

$$Z_i = softmax\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) \tag{6}$$

To preserve causality during training, a masking mechanism is applied to the attention scores, specifically designed to impede the model from attending to future positions, encapsulating the essence of masked self-attention. The ensuing phase involves the computation of the weighted sum, where each head "$i$" contributes to the formation of "$Head_i$" via the formula:

$$Head_i = Z_i V_i \tag{7}$$

A concatenation of the outputs from all heads happens, culminating in a subsequent multiplication with the output weight matrix "$W_o$", yielding the definitive output of the Masked Multi-Head Attention mechanism in the decoder:

$$MaskedMultiHead(E) = \left(head_1 \oplus head_{2\ldots\ldots} \oplus head_h\right) W_o \tag{8}$$

This exposition details the steps involved in using masked self-attention across multiple heads within the decoder of a Transformer model. During the decoding phase, the process begins with the Masked Multi-Head Attention mechanism, which analyzes the input sequence while ensuring that future tokens remain concealed. Once this initial processing is complete, the output from the encoder is integrated as a secondary input for the decoder. Next, the Decoder Multi-Head Attention mechanism operates similarly to the encoder's attention, allowing the model to selectively focus on specific elements of both the input sequence and the encoder's output. Afterwards, the output is refined through two Feed-forward and normalization layer. Following these steps, the final output is evaluated to estimate conditional probabilities, which is a key part of the decoding process. Building on the normalized output, the model calculates a conditional probability distribution over the vocabulary for the next token in the sequence. This is accomplished using the softmax function, which converts the raw scores into a probability distribution. The resulting probabilities guide the selection of the next token, making this step essential for generating outputs that are coherent and contextually relevant. The conditional probability $P\langle y_t | y_1, \ldots, y_{t-1}, x\rangle$ for the next token "$y_t$" is computed using the following equation:

$$P\langle y_t | y_1, \ldots, y_{t-1}, x\rangle = softmax\left(w_o h_t\right) \tag{9}$$

where, "$w_o$" is the weight matrix, and "$h_t$" is the hidden state at position "$t$". This process repeats iteratively for each time step until the entire sequence is decoded. The model uses the generated tokens as input for subsequent time steps, allowing for autoregressive decoding.

Developed by Papineni et al, in 2002 **[45]**, Bilingual Evaluation Understudy (BLEU) has become a standard benchmark for assessing the quality of translations by comparing them to reference translations. The final BLEU score is computed as the weighted geometric mean of the precision scores of 5-grams levels in the generated sequence compared to the reference sequence, adjusted by the brevity penalty as given below:

$$BLEU\ Score = BP \times \exp\left(\sum_{n=1}^{N} w_n \times \log\left(P_n\right)\right) \tag{10}$$

where, “*BP*” is the Brevity Penalty, accounting for the length of the generated sequence relative to the reference. “*N*” is the maximum order of 5-grams considered. “$w_n$” is the weight assigned to each n-gram precision. “$P_n$” is the precision of 5-grams in the predicted sequence. This formula quantifies the precision of the model's output against reference sequences, considering 5-gram orders and assigning appropriate weights.

## Evaluation of the generated output from the Transformer-Decoder

The decoder segment of the transformer is designed to generate multiple words as part of its output sequence comprising both the methylated and unmethylated cytosines bases for each provided sequence. The crucial step post-decoding involves matching these generated words with target sequences. To take the final decision on the full length promoter output sequence generated from the raised model, sequences were mapped back to the original target sequence and per base cytosines analysis was carried to compare the agreement with actual experimental data for methylated states of cytosines for the given experimental condition for every generated sequence. In this way, the accuracy of the raised model was calculated for the complete promoter output sequence. This entire part of the study has used 70% of the Dataset “A” as the training set to build the model and its performance on the remaining 30% totally unseen test dataset. **Figure 3** depicts the operation of the implemented deep-learning encoder-decoder system. Further to this, the evaluation was also done through 10-fold random independent train-test trial runs while building the model from the scratch every time maintaining the same 70:30 ratio of train:test datasets to evaluate the consistency of the developed method across variable learning and testing datasets. Model overfitting level was also tested using MAE values. Full methods details are provided in the **Supplementary File Materials and Methods** section.

# Results and Discussion

## Data collection and formation of datasets

DNA methylation has been reported to be a function of sequence properties and several genes **[26-31, 46-52]**. Some of well identified such genes in plants are: DNA methyltransferase MET1, DDM1, VIM1–3, chromomethylase CMT1-3, histone methyltransferases kryptonite (KYP), Su(var)3-9 homologue 5 (SUVH5), IDN2, IDN2, DRM2, and SUVH6 **[26-31]**. However, these are usually fundamental genes which are not condition specific but usually expressed uniform way. While the fact is that in plants the regulatory DNA methylations are highly spatio-temporal, suggesting involvement of several other molecular factors in making the epigenetic regulation highly specific and time controlled. The present study started with this view where it has successfully attempted to locate such molecular factors which could answer the differential methylation pattern with a substantial accuracy which would be helpful in detecting them economically and completely free from the existing bottlenecks faced by experimental approaches.

In this study, we have investigated the relationship between DNA methylation profiles and transcriptome expression profiles through two important plant species, *Arabidopsis thaliana* and *Oryza sativa*. Our investigation was centered on the 2-kilobase (kb) upstream regions as the potential promoter regions where DNA methylations at cytosines have been reported critical in regulating the downstream gene's expression. The methylation profiles for all genes in both *A. thaliana* and *O. sativa* were recorded from various published studies. The methodology is thoroughly detailed in the methods section, with **Figure 2b** providing a visual representation of how these datasets were curated and utilized in this study. From the SRA database, a total of 260 RNA-seq fastq files and 232 WGBS-seq fastq files were populated for *A. thaliana* and *O. sativa*. This data encompassed a wide array of experimental conditions, 45 for *A. thaliana* and 40 for *O. sativa*. To ensure the integrity of the analysis, only those WGBS-seq data were included which had corresponding RNA-seq data for the same experimental conditions, allowing us to directly associate the gene expression levels with methylation status as well as build scope to consider the influence the of other genes too. For each gene, the 2kb upstream potential promoter sequences were extracted. These sequences were annotated for methylated cytosines, clearly demarcating methylated cytosines from the unmethylated ones. Each promoter sequence was represented multiple times, because of varying methylation states of cytosines across different conditions. In terms of scale, the dataset for *O. sativa* comprised 54,646,596 promoter sequences, while *A. thaliana* contributed 51,425,584 promoter sequences. These datasets have been named as Dataset “A” (*A. thaliana*) and Dataset “B” (*O. sativa*) in the current study, also are illustrated in **Figure 2**.

## Conjoint deep-learning on sequence properties and conditional transcriptome profiles raises an efficient model to detect cytosine methylation

The implementation of the DMRU model represents a significant advancement, particularly in the study of gene regulation through the interplay of DNA methylation, methylation region’s sequence information, and surrounding transcriptome’s profile including the expression of the downstream gene. As already mentioned above, the fundamental hypothesis was consideration of the fact that in plants DNA methylation is differential, where many other genes could be involved in the selection of condition specific cytosine methylation. Thus, a relationship between such expression pattern of causal genes, methylation spot’s local sequence information, and downstream gene’ expression may become a guide to evolve an efficient model where using the RNA-seq profile one may decode the DNA methylation pattern for the selected gene’s promotoer region.

Central to this implementation is a composite deep co-learning system that leverages both Transformers and DenseNet architectures, which are adept at capturing complex patterns. This model was designed to understand the interdependent variability between methylation states in gene promoters and their corresponding RNA expression levels under varying environmental conditions. To rigorously evaluate the model's efficacy, the training and testing phases utilized Dataset "A," having 70:30 split ratio for training and testing. It was ensured in this entire study that there was no overlap and redundancies existed between the training and testing sets, eliminating possibilities of any skews. The initial training phase involved 1,200 distinct promoter sequences paired with RNA-seq data across different experimental conditions. Complete implementation details of the deep-learning model is already covered in the methods section.

Once the dataset was processed through the integrated Transformers-DenseNet system, the model was initially evaluated on an unseen test set consisting of 200 promoter sequences, which produced an impressive accuracy of 90.31% **(Figure 4a)**. This level of accuracy indicates that the model was successful in identifying methylated cytosines within the sequences, showcasing its potential for application in genomic studies to detect DNA methylation just using RNA-seq data. This system was further improved to obtain better results by conducting hyper-parameter optimization steps which is reported in the next section.

A very interesting now question arises that which features were most contributory in the rise of such accurate model to decode cytosine methylation. In this direction, ablation analysis was performed which evaluated one after another considered properties, and performance in their presence and in their absence. In the present section, we discuss the sequence based features' ablation analysis. A total of six distinct sequence-derived word representations were utilized: dimers, trimers, tetramers, pentamers, hexamers, and heptamers. Each representation corresponded to the number of consecutive nucleotides or characters considered within the sequences, thus providing varying lengths of contextual information. The underlying hypothesis was that integrating these representations would lead to additive effects that enhance the model's translational performance. The results indicated a clear trend: while individual word representations demonstrated limited capability most notably, the integration of these representations markedly improved performance due to enhanced sharing of contextual information **(Figure 4b)**. As more complex representations were introduced, significant increase in accuracy was observed: dimeric representations returned 26.02% accuracy, trimeric representation achieved 39.8%, tetrameric reached 57.13%, and pentameric representation attained 72.77%. This progression continued with hexameric and heptameric representations, which returned accuracy values of of 74% and 78%, respectively. It is evident from these findings that the contribution of higher-order words improved the performance mainly because of increase in the vocabulary due to higher number of representational words. But that came at the cost of computational complexity and memory requirements. For example, the GPU memory required for dinucleotides for the training set was 12 GB, which took 238 seconds of execuation time for single epoch. This shot up by manifolds when heptamers were considered where the required memory was 46 GB and took 1,456 seconds to complete the single epoch training process for the same set of training instances. Yet, it was very clear that the true potential of these representations was realized when they were combined. For instance, the pairing of pentameric and hexameric representations (totaling 3,991 words) resulted in a significant leap in the accuracy to 85.85%. Following this, the combination of hexamers with heptamers (3,989 words) further enhanced performance, yielding an accuracy of 89.32%. The best instance of the additive benefits of combining representations was illustrated when incorporating pentameric, hexameric, and heptameric representations, resulting in an impressive accuracy of 91.48% with a total of 5,985 representative words. These results emphasize the necessity of information sharing among different representations, as they collectively harnessed their unique attributes to raise a superior model. The findings are illustrated in performance plots **(Figure 4a)**, which clearly show the improvements in accuracy associated with various combinations of sequence representations.

## Hyper-paramter optimization further improves the model

Above the model was developed with prefixed hyperparameters, though not completely random but values understood from our previous works on other problems **[23-24, 40, 53]**. Initially, the main task was to determine if the identified sequence and expression properties were forming the right features or not. Additionally, the most suitable deep-learning architecture too had to be finalized. Once these were done and a substantial accuracy was obtained, hyperparamter optimization was the next essential to improve the performance of the developed deep-learning system.

An exploration into the performance of varying numbers of encoders and decoders for the Transformers-DenseNet system revealed a notable trend in the system's efficiency as layers were incrementally added. Specifically, the most pronounced increase in performance was observed at the addition of the eighth encoder-decoder pair, marking a significant point in the model's ability to handle the complexities of the task at hand. This enhancement in performance remained stable through the inclusion of subsequent layers up to the eighth, after which there was a noticeable decline in the effectiveness **(Figure 4c)**. The output from the multi-head attention layer was subjected to a dropout layer with a dropout fraction of 0.12, effectively regularizing the model to mitigate overfitting by randomly deactivating a fraction of the neurons during training. This output was then normalized to ensure that subsequent layers received data that was statistically stable and appropriately scaled. The architecture continued with a Feed-Forward network comprising 48 nodes, followed by another dropout layer with a slightly higher dropout fraction of 0.14. This strategic layering approach allowed the model to not only learn complex representations but also to maintain a robust training regimen that promotes generalizability. Next, a second feed-forward layer was integrated, this time featuring 64 nodes, further enriching the model's capacity for feature extraction and transformation. This layer was followed by another normalization stage, ensuring that the subsequent computations maintained a uniform scale and distribution. The output from the DenseNet component was then channeled through a feed-forward layer within the transformers' encoder, which played a crucial role in capturing the variability present in RNA expression profiles under specific conditions, along with their corresponding methylated sequence data. This architecture was designed to facilitate this interaction enabled the model to leverage rich biological data effectively, enhancing its performance capabilities. In terms of model training, the output from the encoder was transferred to the decoder's components through a masked multi-head attention mechanism, a design choice that crucially prevented information leakage during the training phase and ensured that the model adhered to the sequence generation rules intrinsic to natural language processing. This attention to detail culminated into the final processing via the decoder's feed-forward layer, responsible for generating coherent outputs in response to the processed inputs. To evaluate performance quantitatively, the model utilized the cross-entropy loss function for loss calculation, employing the Adam optimizer to enhance training efficiency and convergence speed. The learning rate for this optimizer was finally set at 0.001. The model underwent training across 100 epochs, with a batch size of 4 to optimize computational resources while maintaining the integrity of gradient updates. The hyperparameters defining the model's output layer were fine-tuned, resulting in a final configuration with softmax activation function, the cross-entropy loss function to guide learning, and the Adam optimizer. The “pt” format provided the model graph definition and weights to the pytorch structure while saving the model for translation. **Figure 3** illustrates the finally implemented deep-learning architecture. Upon evaluation against the test set, the optimized model achieved an impressive accuracy of approximately 93.17%, as illustrated in **Figure 4a**, a significant improvement by ~3% in accuracy exhibited by the optimized model after the hyperparameter tuning **(Figure 4a)**.

## Remarkably consistent performance observed across wide range of experimental data

The investigation into the performance of the co-learning system when applied to the unseen promoter sequences yielded significant insights into its capabilities and performance. By evaluating 10,000 promoter sequences from Dataset “A” that were not included in the training and testing phases, we aimed to ascertain the model's generalizability beyond its initial training data. Remarkably, the model achieved an impressive accuracy of 92.82% **(as illustrated in Figure 4d)**, underscoring its effectiveness in capturing the intricate co-variability between RNA expression levels and the specific methylation preferences associated with these sequences in condition specific manner. This high level of accuracy suggests that the model learned well from the training data and internalized the underlying biological relationships, allowing it to make accurate discovery even when confronted with previously unseen data.

Following this promising result, we developed two co-learning model, Model “A” from Dataset “A” and Model “B” from Dataset “B”. We further tested these models on the testset of Dataset “A” and Dataset “B”, each comprising a total of 30,000 unique promoter sequences corresponding to *A. thaliana* and *O. sativa*, respectively. Both the model's performance remained consistent, with an average accuracy of 92.74% across the datasets. More specifically, the accuracy for *A. thaliana* was recorded at 92.6% from Model “A”, while for *O. sativa* from Model “B”, it slightly improved to 92.88% **(Figure 4d)**. This consistency in accuracy across datasets reinforces the model can reliably discover methylation patterns in diverse genomic and experimental conditions. Afterwards, we assessed contribution from the integration of the RNA expression profile handled through DenseNet, the DL system was raised again without it. This analysis marked decrease in the model's accuracy, plummeting to 67.23% (p-value >> 0.01). Such a significant decline advocates for the necessity of jointly analyzing sequence data and RNA expression profiles which ultimately enriches the model's performance. 10-fold random trials performance, raising models form randomly created training and testing data every time, concurred with the above observed performance level and scored in the same range consistently for Dataset “A”. The MAE for training was found to be 0.0685, while for testing it was 0.0714, resulting in a very small difference of only 0.0029. A t-test comparing the MAE values for the training and test sets yielded a highly insignificant result of ~15%, much above the significance threshold of 5% or lower, further confirming the absence of any possibility of any significant over-fitting (**Supplementary Table S2 Sheet 5)**. To conduct an unbiased performance testing without any potential recollection of data instances, it was ensured that no overlap and redundancy existed across any of the used data.

## Accurate cross-species identification of DNA methylation where GC similarity is the key, not homology

Detecting DNA methylation computationally itself is a big challenge, whose next level comes with condition specific detection of DNA methylation. It it becomes top challenge when one looks for developing a tool which could work cross-species in total universal manner also. Universality is the requirement as any such tool should be capable to address any plant genome. Unfortunately, most of these challenges have remained unaddressed, which DMRU has successfully solved now.

The species specific models discussed above can tell condition specific methylation for any condition for the given species. But could they work with same efficiency on other species? In the context of cross-species application and universality of DMRU algorithm, the raised model was subjected to rigorous assessment to evaluate its effectiveness in identifying cross-species DNA methylation patterns. For this evaluation, we utilized two distinct datasets, referred to as Dataset "A" and Dataset "B," comprising sequences from two model plant, *A. thaliana* and *O. sativa*. The approach involved training a model using sequences from *A. thaliana* and subsequently testing it on sequences from *O. sativa*, while conversely, a model trained on *O. sativa* was tested on *A. thaliana* sequences. The model raised from *O. sativa* achieved a maximum accuracy of 83.9% (average accuracy 76.98%) on Dataset “A” covering *A. thaliana* **(Figure 4e)**. Similarly, the model raised from *A. thaliana* achieved a maximum accuracy of 79.89% (average accuracy 73.63%) on Dataset “B” **(Figure 4e)**. Both results clearly highlights that the raised models left enough scope for further improvement as they were far below than the accuracy levels observed for within species runs.

To develop a universal model capable of identifying DNA methylation promoter sequences from gene expression data across various species, our initial approach centered on leveraging homology between the sequences where the query sequences/genes were looked for their corresponding homologous ones from *A. thaliana.* This approach aimed to harness evolutionary relationships as has been the practice to expect similar functionality for homologous genes. However, despite rigorous testing across species, the model underperformed, achieving an accuracy of only around 83%, as depicted in **Figure 5a-b**. This outcome prompted us to rethink our approach as well as highlighted how wrong one can go while traditionally relying on the homology, especially for epigenetic controls. In epigenetic control it is GC isochore which should be looked for instead of homology. Therefore, the GC content (GC%) of the 2 Kb upstream regions for the genes were evaluated in both *A. thaliana and O. sativa*. By grouping genes with similar GC% value ranges, we created a new model based on these clusters. This innovative approach proved fruitful, when we applied the GC%-based model to the datasets from both species, the results were significantly improved. The model derived from *A. thaliana* demonstrated an impressive accuracy exceeding 90% on *O. sativa* (Dataset “B”), while the model based on *O. sativa* also achieved above 90% accuracy on *A. thaliana* (Dataset “A”) **(Figure 5a-b)**. These outcomes underscore the critical role which GC content plays in the context of gene regulation and methylation identification.

In the next step, the above analysis was extended forward for the model's final assessment to determine how well it performed to identify cross-species DNA methylation on other unseen species. Based on experimental data availability for validation, this was done for *Arabidopsis lyrata* and *Solanum tuberosum*, for which WGBS-seq and RNA-seq data were available. Here also, the DMRU model developed from Dataset “A” achieved an average accuracy >90% for both the species, affirming its efficacy in handling cross-species conditions. **Figure 5c** provides the comparison details for both species by DMRU for cross-species performance. All these series of validation studies reinforced that DMRU performed consistently with high accuracy and emerged as a promising first of its kind tool for cross-species methylation identification without any dependency on predefined models while implementing generative and explainable AI innovative way. It can be employed successfully to even determine the methylated cytosines in newly sequenced plant genomes besides capturing the condition specific methylation profile with high accuracy and consistency.

## DMRU outstands in comparative benchmarking studies

To evaluate the performance of our software tool in this study, we compared DMRU with six different tools: iDNA-MS **[54]**, iDNA-ABT **[55]**, iDNA-ABF **[56],** MaskDNA-PGD **[20]**, MuLan-Methyl **[57]**, and DeepPGD **[19]**. The three out of the compared six tools viz. iDNA-MS, DeepPGD, and MuLan-Methyl have bugs to run. Initially two different datasets based on *Arabidopsis thaliana* and *Oryza sativa*, were used to evaluate the performance of the developed model of DMRU with respect to three different tools, representing most recent and different approaches to detect DNA methylation: iDNA-ABT (BERT+ transductive information maximization (TIM)), MaskDNA-PGD (CNN, Bi-LSTM and attention mechanism), and iDNA-ABF (BERT based approach). The performance measure on the test set of these dataset which consists 12 different conditions gave an idea how the compared algorithms in their existing form perform. All these three tools were trained and tested across these datasets where DMRU outperformed all of them, for all the performance metrics considered. All the three tools benchmarked here are not capable to identify condition specific DNA methylation and give a uniform cytosine methylation profile for all conditions, making them hardly of any use in real world applications. As highlighted in **Figure 6 a-c**, all these tools output exactly the same result and fail to capture the dynamic condition specific methylation patterns in *O. sativa*. On the other hand **(Figure 6d)**, DMRU captured the conditions specific methlyations in these regions. **Supplementary Figure S1** also presents dynamic condition specific methylation patterns in *A. thaliana*. **Supplementary Figure S2-3** show that the compared tools did not breach even 80% accuracy mark when the accuracy was calculated for every sequence under different conditions. With this all, DMRU stands at present the only approach which could capably suggest differential methylation patterns across different conditions, emerging also as a good alternative for costly bisulfite sequencing experiments.

## Explainable deep learning gives insight into the most critical condition specific factors functioning for DNA methylation

Significant efforts are underway to enhance the transparency and interpretability of deep learning. DLs perform exceptionally good but are hard to explain the causality. Explainability of DL models has been the frontier of DL research at the present. It is crucial to make DL models more interpretative. Selvaraju et al., 2017 **[32]**, introduced the Grad-CAM technique or Gradient-weighted Class Activation Mapping, which is a powerful technique used in the field of DL offering a visual explanation of deep learning models.

DMRU has implemented Grad-CAM scoring scheme to pull out the genes from the transcriptome profile which appear to be critical for condition specific methylation for any given sequence stretch. This approach involves computing the gradient of a target class, denoted as "c," alongside the global average of the activation feature map "F" derived from the first convolutional layer "l." The gradient provides insight into how the model's performances change with respect to small variations in the input, allowing identification of the features critical for distinguishing between different methylation states. By synthesizing these gradients with the activation maps, DMRU generates class-discriminative profiles which highlight the importance of specific features, expressing genes from the transcriptome profile in the present condition. To refine these profiles further, DMRU applies a weighted combination of the activation maps, effectively emphasizing the contribution of each feature based on its gradient score. Subsequently, a ReLU function is applied which ensures that only features with a positive influence are retained. This methodological framework allows us to visually represent and interpret which genes are responsible for the DNA sequence methylation states. The mathematical formulation and operational principles of Grad-CAM are described as follows:

$$GradCAM_l^c = ReLU\left(\sum_l \alpha_l^c F^l\right) \qquad (12)$$

$$\alpha_l^c = \frac{1}{L}\sum_i \sum_j \left(\frac{\partial y^c}{\partial A_{ij}^l}\right) \qquad (13)$$

Where, $\alpha_l^c$ denotes the neuron importance weights, highlighting the most relevant features that contribute to the models accuracy. ***'L'*** is the vector length in the feature map, and $A_{ij}^l$ represents the activation at ***"ij<sup>th</sup>"*** position in the input array. Full details of Grad-CAM implementation is covered in **Supplementary File Materials and Methods** section.

To illustrate the effectiveness of the implemented approach, case of gene AT1G13450 from *Arabidopsis thaliana* has been considered as example here. This gene is pivotal due to its role in binding to one of the cis-acting elements, BoxII, which resides within the upstream promoter region of light-responsive genes, regulating plant growth. In this context, we utilized Grad-CAM to visualize the influence of gene expression on DNA methylation. To extract relevant features, we employed the first convolutional layer of our Transformer-DenseNet model, effectively enabling Grad-CAM to elucidate crucial gene features linked to DNA methylation. This method proved to be a valuable tool for understanding the underlying mechanisms of DNA methylation. The findings of this analysis is illustrated by **Figure 7a** where the top 100 genes ranked by their Grad-CAM scores across multiple conditions were clustered and presented. To further contextualize our findings within biological frameworks, we conducted Gene Ontology (GO) analysis on the genes identified through Grad-CAM. GO analysis of these high Grad-CAM scoring top 100 genes unveiled significant insights that a large proportion of these genes were intricately linked to essential plant developmental processes **(Figure 7b-d)**.

Notably, they played vital roles in maintaining RNA binding, facilitating floral whorl developmental mechanisms, and driving stamen development, which is crucial for reproduction in flowering plants **(Figure 7c)**. This emphasizes the foundational importance of these genes in not just growth, but also in the adaptations of *A. thaliana*. In *Arabidopsis*, floral organs develop in concentric whorls, and AT1G13450 acts as a regulator of genes that determine the identity of these organs. Specifically, it affect the expression of genes responsible for establishing the correct patterning of the second and third whorls. It also plays role in controlling floral symmetry, the proper arrangement of floral organs and it also helps in establishing floral symmetry, an essential feature of floral development **[58-59]**. AT1G13450 is involved in interactions with different TFs such as LEAFY (LFY) and APETALA1 (AP1), which are known to regulate floral organ identity

**[59]**. The interaction of these transcription factors can influence the development of floral organs in *Arabidopsis* too. AT1G13450 contributes to the integration of various signals needed for the correct positioning and formation of floral organs. Moreover, our analysis revealed that these genes exhibit enrichment in various molecular functions. Key activities identified include extra-cellular transport, floral whorl and organ development, response to abscisic acid, which are essential for signaling pathways that regulate growth and development, indicative of their roles in biochemical reactions crucial for metabolism and synthesis **(Figure 7c-d)**. The examination of KEGG pathways further corroborated our findings, illustrating the involvement of these genes in these important metabolic pathways, as depicted in **Figure 7b-d**.

This application demonstration while considering gene AT1G13450 clearly displayed the level of accuracy and importance DMRU is capable to attain, while churning out relevant information also. The Grad-CAM part of DMRU very aptly proposed the relevant genes in the system which could be involved with this gene where DNA methylation appears to have significant stake in its regulation.

## Web-server implementation

The DMRU server hosts the software in a user friendly manner for the identification of methylated cytosines within plant DNA sequences. To identify methylated cytosine, users are required to upload their DNA sequences in FASTA format and its associated experimental condition transcriptome expression profile in “tsv” format, followed by clicking the submit button. Once the user submits the data, the back end processing begins with the input data which is then passed through Transformer-DenseNet engine in the back-end which identifies the methylated cytosine for the given promoter sequences. DMRU server can handle upto 40 sequences in one submission. Subsequently, a results page featuring relevant information in an interactive format is presented. This result page consists of two tabs, **(i)** in the first tab, identified methyated cytosines were presented where the user can visualize the methylated cytosines within the given input sequences and also users get the capability to download the results in a tabular format, facilitating further downstream analysis and integration with other tools. **(ii)** The results page also provides a separate tab for explainability scoring line plot in the form of importance score while highlighting the genes important for the methylation in the sequence. This line plot is implemented through plotly.js. Users can download results in a tabular format. The server implementation has been done using PHP, JavaScript, HTML5, and the Apache-Linux platform. The majority of codes were developed in Shell and Python scripts for jobs related to data curation and back-end processing. The project was executed within the Ubuntu Linux platform's Open Source OS environment. **Figure 8** provides an overview of DMRU webserver implementation. The DMRU system is freely accessible at https://scbb.ihbt.res.in/DMRU/.

## Conclusion

DNA methylation at cytosines regulate the downstream gene to attain spatio-temporal gene expression pattern,. This is very crucial for developement, survival, and adaptations mechanism of plants. Currently, only though Bisulfite based sequencing experiments one can find the methylation of DNA and its differential patterns, making its detection a costly affair. No existing software tools are capable to tell the differential methylation patterns. Here, we have develpoed a deep-learning based encoder:decoder system in generative manner to suggest the condition specific cytosine methylation with consistently high accuracy. It was tested across a large number of experimental data, covering several conditions. It also worked with same level of accuracy in cross-species manner. DMRU now makes it possible for anyone to find out the methylation pattern of their target genes promoter for any given condition. They just need to load their RNA-seq data and the target sequences. DMRU goes further and also tells about the most influential genes associated with the observed methylation patterns. With this all, DMRU emerges as a first of its kind software tool, and a much cheaper, faster, and accurate alternative to costly and time consuming experiments like Bisulfite-seq.

## Acknowledgements

The work was carried out under the aegis of The Himalayan Centre for High-throughput Computational Biology (HiCHiCoB), a BIC supported by DBT, Govt. of India. SG, AK, and VK are thankful to DBT, India for financial support as project associateship. SG and VK are also thankful to Academy of Scientific and Innovative Research (AcSIR) for their Ph.D. enrollment. All authors are thankful to the Director, CSIR-IHBT, for his kind support for this study. This MS has CSIR-IHBT MSID ##.

## Author’s contributions

SG carried out the major parts of this study. AK and SG developed the web-server of DMRU. VK assisted in the study. RS conceptualized, designed, analyzed, and supervised the entire study. SG and RS wrote the MS.

## Declaration of competing interest

The authors declare that they have no competing interests.

## Software and Data availability

All the secondary data used in the present study were publicly available and their due references and sources have been provided in **Supplementary Table S1-2**. The software has also been made available as a webserver at https://scbb.ihbt.res.in/DMRU.

## Figure Legend

**Figure 1: a)** A DNA double strand with methylated (violet) and unmethylated (green) CpG sites is shown before and after bisulfite treatment and polymerase chain reaction (PCR). Methylated cytosine remains unaffected by bisulfite treatment, while unmethylated cytosine is converted to uracil, which is subsequently changed to thymine during PCR on the original strand. In the complementary strand, unmethylated cytosine is converted to adenine. **b)** Distribution of RRBS and MeDIP-sequencing based data in public domain. Barring a few plant species, majority plant species are still not sequenced. For some species the data is much more than the rest.

**Figure 2: a-b)** Distribution of WGBS and RNA-sequencing studies retrieved for the study collected from different projects from NCBI. **c)** Flowchart representation of dataset formation. The Dataset “A” and "B" were created for two plant species, *Arabidopsis thaliana* and *Oryza sativa*. Both the datasets contained the instances originating from RNA-seq and WGBS-seq, respectively.

**Figure 3: Implementation of the DMRU Deep Co-learning system using Transformers and DenseNet to annotate DNA methylation.** The encoder processes the promoter sequences, without methylation representations, in penta, hexa, and heptamer words representations, while the decoder receives the methylated sequences representations for the corresponding sequence. Eight layers of encoder:decoder pairs were incorporated. In the parallel, a DenseNet consisting of 121 layers accepts input in the form of the associated RNA expression profile. The learned representations from the encoder and DenseNet are merged and passed to the encoder-decoder multi-head attention layer, which also receives input from decoder layers. Subsequently, the resulting output undergoes conditional probability estimation, playing a pivotal role in the decoding process. Following layer normalization, the model proceeds to calculate the conditional probability distribution over the

vocabulary for the next token. Finally, the resultant tokens were converted back into words, representing the methylated states for the given sequence.

**Figure 4: Comprehensive performance assessment of the raised deep co-learning models. a)** The model was evaluated on an unseen test set consisting of randomly selected 200 promoter sequences from Dataset "A" before and after hyper-parameter optimization. **(b)** Ablation analysis for six main properties was conducted to evaluate their impact on model performance. These words representations appeared highly additive and complementary to each other as the performance increased substantially as they were combined together. **(c)** Hyper-parameters optimization effect on the performance due to varying numbers of encoders and decoders for the Transformers-DenseNet system. At the addition of the eighth encoder-decoder pair, a significant change in the model's ability to handle the complexities was observed. **d-e)** Self and cross-species performance of DMRU across plant species *A. thaliana* and *O. sativa*. For this evaluation, we utilized two distinct datasets, referred to as Dataset "A" and Dataset "B," comprising sequences from two model plant species, *A. thaliana* and *O. sativa*. The approach involved training a model using sequences from *A. thaliana* and subsequently testing it on the sequences from *A. thaliana and O. sativa,* and vice-versa. This self and cross-species assessment were critical to ascertain the accuracy and adaptability of the models across different genomic contexts.

**Figure 5: Performance comparisons for the GC% based universal model of DMRU across species.** Performance assessment for the model raised using **a)** *Arabidopsis thaliana* data, and **b)** *Oryza sativa* data, with self genes (Blue coded), homology based cross-species (orange coded), and GC% similarity based cross-species testing. **c)** Cross-species performance of DMRU based on GC % clusters. Here, GC% cluster based models from *A. thaliana* was tested across *A. lyrata* and *S. tuberosum* genes. Consistently >90% accuracy was observed for every GC% clusters. The GC% of

the 2 Kb upstream regions for the genes were evaluated in both *A. thaliana and O. sativa*. By grouping genes with similar GC% value ranges, we created new models based on these clusters, the results were significantly improved. These results underscore the critical role which GC content plays in the context of gene regulation and methylation.

**Figure 6: Performance benchmarking of the universal model of DMRU based on GC% for *Oryza sativa*.** The performance over the 12 different experimental conditions for the compared software tools were benchmarked for the upstream region (have 500 bases shown) of gene Os01g0110100, for the tools **(a)** iDNA-ABF, **(iii)** iDNA-ABT, **c)** MaskDNA-PGD, and **d)** DMRU. Here also, comparative benchmarking returned outstanding results, reinforcing DMRU as the most reliable software too to detect differential methylation. As can be seen, other tools keep giving same methylation patterns in all experimental conditions, while DMRU very accurately captured the condition specific methylation patterns.

**Figure 7: Explainability of DMRU system.** It helps to understand the most impactful genes features possibly responsible for the observed methylation patterns. **(a)** The top 100 genes ranked by their Grad-CAM scores for their methylation impact across multiple conditions were clustered for the target gene AT1G13450, **(b-d)** Gene Ontology (GO) analysis on the genes identified through Grad-CAM. This application demonstration while considering this target gene aptly proposed the relevant genes which concur very well with respect to their common functions. Conditions specific differential methylation appears to control specific functionality.

**Figure 8: DMRU webserver implementation. a)** Input data box where the user can either paste or load the input files, **b)** Inputs are DNA sequence in FASTA format and gene expression profile in “tsv” format output from RNA-seq data analysis. **c)** The methylated cytosines are represented in an

interactive form. Download option for the result in the tabular format. **d)** Also, important feature scoring is implemented and is represented in the form of interactive line plot depicting the distribution of scoring distribution for the most influential genes associated with the observed DNA methylation pattern for the user provided target sequence.

## Supplementary Information

**Supplementary Table S1: Brief list of some of the published tools for DNA methylation identification.**

**Supplementary Table S2: Complete details of various sources of WGBS-seq & RNA-seq and information of ten fold random independent trials.**

**Supplementary Figure S1: Performance and benchmarking of the universal model of DMRU based on GC% for *Arabidopsis thaliana*.** The performance over the 12 different experimental conditions for the compared software tools were benchmarked for the upstream region (have 500 bases shown) of gene Os01g0110100, for the tools **(a)** iDNA-ABF, **(iii)** iDNA-ABT, **c)** MaskDNA-PGD, and **d)** DMRU.

**Supplementary Figure S2: Performance accuracy of the benchmarked software tools (DMRU and iDNA-ABT) across different conditions.** iDNA-ABT did not breach even 80% accuracy mark while DMRU easily crosses 90% when the accuracy was calculated for every sequence under different conditions. Rendering DMRU effective across different experimental conditions.

**Supplementary Figure S3: Performance accuracy of the software tools across different conditions. (a)** iDNA-ABF and **b)** MaskDNA-PGD. These software tools did not breach even 80% accuracy mark when the accuracy was calculated for every sequence under different conditions.

# References

1. Niederhuth,C.E. and Schmitz,R.J. (2017) Putting DNA methylation in context: from genomes to gene expression in plants. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms, 1860, 149–156.
2. Ehrlich,M. and Wang,R.Y.-H. (1981) 5-Methylcytosine in Eukaryotic DNA. Science, 212, 1350–1357.
3. Kriaucionis,S. and Heintz,N. (2009) The Nuclear DNA Base 5-Hydroxymethylcytosine Is Present in Purkinje Neurons and the Brain. Science, 324, 929–930.
4. He,Y.-F., Li,B.-Z., Li,Z., Liu,P., Wang,Y., Tang,Q., Ding,J., Jia,Y., Chen,Z., Li,L., et al. (2011) Tet-Mediated Formation of 5-Carboxylcytosine and Its Excision by TDG in Mammalian DNA. Science, 333, 1303–1307.
5. Ito,S., Shen,L., Dai,Q., Wu,S.C., Collins,L.B., Swenberg,J.A., He,C. and Zhang,Y. (2011) Tet Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine. Science, 333, 1300–1303.
6. Pfaffeneder,T., Hackner,B., Truß,M., Münzel,M., Müller,M., Deiml,C.A., Hagemeier,C. and Carell,T. (2011) The Discovery of 5-Formylcytosine in Embryonic Stem Cell DNA. Angewandte Chemie International Edition, 50, 7008–7012.
7. Osorio-Concepción,M., Lax,C., Navarro,E., Nicolás,F.E. and Garre,V. (2021) DNA Methylation on N6-Adenine Regulates the Hyphal Development during Dimorphism in the Early-Diverging Fungus Mucor lusitanicus. Journal of Fungi, 7, 738.

8. Luo,G.-Z., Blanco,M.A., Greer,E.L., He,C. and Shi,Y. (2015) DNA N6-methyladenine: a new epigenetic mark in eukaryotes? Nat Rev Mol Cell Biol, 16, 705–710.

9. Wu,K.-J. (2020) The epigenetic roles of DNA N6-Methyladenine (6mA) modification in eukaryotes. Cancer Letters, 494, 40–46.

10. Zhang,G., Huang,H., Liu,D., Cheng,Y., Liu,X., Zhang,W., Yin,R., Zhang,D., Zhang,P., Liu,J., et al. (2015) N6-Methyladenine DNA Modification in Drosophila. Cell, 161, 893–906.

11. Mondo,S.J., Dannebaum,R.O., Kuo,R.C., Louie,K.B., Bewick,A.J., LaButti,K., Haridas,S., Kuo,A., Salamov,A., Ahrendt,S.R., et al. (2017) Widespread adenine N6-methylation of active genes in fungi. Nat Genet, 49, 964–968.

12. Fu,Y., Luo,G.-Z., Chen,K., Deng,X., Yu,M., Han,D., Hao,Z., Liu,J., Lu,X., Doré,L.C., et al. (2015) N6-Methyldeoxyadenosine Marks Active Transcription Start Sites in Chlamydomonas. Cell, 161, 879–892.

13. Lister,R., Pelizzola,M., Dowen,R.H., Hawkins,R.D., Hon,G., Tonti-Filippini,J., Nery,J.R., Lee,L., Ye,Z., Ngo,Q.-M., et al. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. Nature, 462, 315–322.

14. Meissner,A., Gnirke,A., Bell,G.W., Ramsahoye,B., Lander,E.S. and Jaenisch,R. (2005) Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Research, 33, 5868–5877.

15. Flusberg,B.A., Webster,D.R., Lee,J.H., Travers,K.J., Olivares,E.C., Clark,T.A., Korlach,J. and Turner,S.W. (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods, 7, 461–465.

16. Xu,H., Jia,P. and Zhao,Z. (2021) Deep4mC: systematic assessment and computational prediction for DNA N4-methylcytosine sites by deep learning. Briefings in Bioinformatics, 22, bbaa099.

17. Liu,Q., Chen,J., Wang,Y., Li,S., Jia,C., Song,J. and Li,F. (2021) DeepTorrent: a deep learning-based approach for predicting DNA N4-methylcytosine sites. Briefings in Bioinformatics, 22, bbaa124.

18. Liu,M., Sun,Z.-L., Zeng,Z. and Lam,K.-M. (2022) MGF6mARice: prediction of DNA N6-methyladenine sites in rice by exploiting molecular graph feature and residual block. Briefings in Bioinformatics, 23, bbac082.

19. Teragawa,S., Wang,L. and Liu,Y. (2024) DeepPGD: A Deep Learning Model for DNA Methylation Prediction Using Temporal Convolution, BiLSTM, and Attention Mechanism. International Journal of Molecular Sciences, 25, 8146.

20. Zheng,Z., Le,N.Q.K. and Chua,M.C.H. (2023) MaskDNA-PGD: An innovative deep learning model for detecting DNA methylation by integrating mask sequences and adversarial PGD training as a data augmentation method. Chemometrics and Intelligent Laboratory Systems, 232, 104715.

21. Vaswani,A., Shazeer,N., Parmar,N., Uszkoreit,J., Jones,L., Gomez,A.N., Kaiser,Ł. and Polosukhin,I. (2017) Attention is All you Need. In Advances in Neural Information Processing Systems. Curran Associates, Inc., Vol. 30.

22. Tang,X., Zheng,P., Li,X., Wu,H., Wei,D.-Q., Liu,Y. and Huang,G. (2022) Deep6mAPred: A CNN and Bi-LSTM-based deep learning method for predicting DNA N6-methyladenosine sites across plant species. Methods, 204, 142–150.

23. Gupta,S., Kesarwani,V., Bhati,U., Jyoti and Shankar,R. (2024) PTFSpot: deep co-learning on transcription factors and their binding regions attains impeccable universality in plants. Briefings in Bioinformatics, 25, bbae324.

24. Gupta,S. and Shankar,R. (2023) miWords: transformer-based composite deep learning for highly accurate discovery of pre-miRNA regions across plant genomes. Briefings in Bioinformatics, 24, bbad088.

25. Devlin,J., Chang,M.-W., Lee,K. and Toutanova,K. (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 10.48550/arXiv.1810.04805.

26. Lindroth,A.M., Cao,X., Jackson,J.P., Zilberman,D., McCallum,C.M., Henikoff,S. and Jacobsen,S.E. (2001) Requirement of CHROMOMETHYLASE3 for Maintenance of CpXpG Methylation. Science, 292, 2077–2080.

27. Jackson,J.P., Lindroth,A.M., Cao,X. and Jacobsen,S.E. (2002) Control of CpNpG DNA methylation by the KRYPTONITE histone H3 methyltransferase. Nature, 416, 556–560.

28. Ebbs,M.L., Bartee,L. and Bender,J. (2005) H3 Lysine 9 Methylation Is Maintained on a Transcribed Inverted Repeat by Combined Action of SUVH6 and SUVH4 Methyltransferases. Molecular and Cellular Biology, 25, 10507–10515.

29. Ebbs,M.L. and Bender,J. (2006) Locus-Specific Control of DNA Methylation by the Arabidopsis SUVH5 Histone Methyltransferase. The Plant Cell, 18, 1166–1176.

30. Ausin,I., Greenberg,M.V.C., Simanshu,D.K., Hale,C.J., Vashisht,A.A., Simon,S.A., Lee,T., Feng,S., Española,S.D., Meyers,B.C., et al. (2012) INVOLVED IN DE NOVO 2-containing complex involved in RNA-directed DNA methylation in Arabidopsis. Proc Natl Acad Sci U S A, 109, 8374–8381.

31. Ausin,I., Mockler,T.C., Chory,J. and Jacobsen,S.E. (2009) IDN1 and IDN2: two proteins required for de novo DNA methylation in Arabidopsis thaliana. Nat Struct Mol Biol, 16, 1325–1327.

32. Selvaraju,R.R., Cogswell,M., Das,A., Vedantam,R., Parikh,D. and Batra,D. (2017) Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In.pp. 618–626.

33. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30, 2114–2120.

34. Gahlan,P., Singh,H.R., Shankar,R., Sharma,N., Kumari,A., Chawla,V., Ahuja,P.S. and Kumar,S. (2012) De novo sequencing and characterization of Picrorhiza kurrooa transcriptome at two temperatures showed major transcriptome adjustments. BMC Genomics, 13, 126.

35. Kim,D., Paggi,J.M., Park,C., Bennett,C. and Salzberg,S.L. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol, 37, 907–915.

36. Liao,Y., Smyth,G.K. and Shi,W. (2019) The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic Acids Research, 47, e47.

37. Krueger,F. and Andrews,S.R. (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics, 27, 1571–1572.

38. Danecek,P., Bonfield,J.K., Liddle,J., Marshall,J., Ohan,V., Pollard,M.O., Whitwham,A., Keane,T., McCarthy,S.A., Davies,R.M., et al. (2021) Twelve years of SAMtools and BCFtools. GigaScience, 10, giab008.

39. Huang,G., Liu,Z., van der Maaten,L. and Weinberger,K.Q. (2017) Densely Connected Convolutional Networks. In.pp. 4700–4708.

40. Gupta,S., Jyoti, Bhati,U., Kesarwani,V., Sharma,A. and Shankar,R. (2024) PTF-Vāc: Ab-initio discovery of plant transcription factors binding sites using explainable and generative deep co-learning encoders-decoders. 10.1101/2024.01.28.577608.

41. Zhou,T., Yang,L., Lu,Y., Dror,I., Dantas Machado,A.C., Ghane,T., Di Felice,R. and Rohs,R. (2013) DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. Nucleic Acids Research, 41, W56–W62.

42. Sharma,N.K., Gupta,S., Kumar,A., Kumar,P., Pradhan,U.K. and Shankar,R. (2021) RBPSpot: Learning on appropriate contextual information for RBP binding sites discovery. iScience, 24, 103381.

43. Hombach,D., Schwarz,J.M., Robinson,P.N., Schuelke,M. and Seelow,D. (2016) A systematic, large-scale comparison of transcription factor binding site models. BMC Genomics, 17, 388.

44. Jyoti, Ritu, Gupta,S. and Shankar,R. (2024) Comprehensive analysis of computational approaches in plant transcription factors binding regions discovery. Heliyon, 10, e39140.

45. Papineni,K., Roukos,S., Ward,T. and Zhu,W.-J. (2002) BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL ’02. Association for Computational Linguistics, USA, pp. 311–318.

46. Niederhuth,C.E. and Schmitz,R.J. (2017) Putting DNA methylation in context: from genomes to gene expression in plants. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms, 1860, 149–156.

47. Marinus,M.G. and Løbner-Olesen,A. (2014) DNA Methylation. EcoSal Plus, 6, 10.1128/ecosalplus.ESP-0003–2013.

48. Bock,C., Paulsen,M., Tierling,S., Mikeska,T., Lengauer,T. and Walter,J. (2006) CpG Island Methylation in Human Lymphocytes Is Highly Correlated with DNA Sequence, Repeats, and Predicted DNA Structure. PLOS Genetics, 2, e26.

49. Liu,G., Zhang,F., Jiang,Y., Hu,Y., Gong,Z., Liu,S., Chen,X., Jiang,Q. and Hao,J. (2017) Integrating genome-wide association studies and gene expression data highlights dysregulated multiple sclerosis risk pathways. Mult Scler, 23, 205–212.

50. Pavlovic,M., Ray,P., Pavlovic,K., Kotamarti,A., Chen,M. and Zhang,M.Q. (2017) DIRECTION: a machine learning framework for predicting and characterizing DNA methylation and hydroxymethylation in mammalian genomes. Bioinformatics, 33, 2986–2994.

51. Cokus,S.J., Feng,S., Zhang,X., Chen,Z., Merriman,B., Haudenschild,C.D., Pradhan,S., Nelson,S.F., Pellegrini,M. and Jacobsen,S.E. (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature, 452, 215–219.

52. Baubec,T. and Schübeler,D. (2014) Genomic patterns and context specific interpretation of DNA methylation. Current Opinion in Genetics & Development, 25, 85–92.

53. Ritu, Gupta,S., Sharma,N.K. and Shankar,R. (2022) DeepPlnc: Bi-modal deep learning for highly accurate plant lncRNA discovery. Genomics, 114, 110443.

54. Lv,H., Dao,F.-Y., Zhang,D., Guan,Z.-X., Yang,H., Su,W., Liu,M.-L., Ding,H., Chen,W. and Lin,H. (2020) iDNA-MS: An Integrated Computational Tool for Detecting DNA Modification Sites in Multiple Genomes. iScience, 23.

55. Yu,Y., He,W., Jin,J., Xiao,G., Cui,L., Zeng,R. and Wei,L. (2021) iDNA-ABT: advanced deep learning model for detecting DNA methylation with adaptive features and transductive information maximization. Bioinformatics, 37, 4603–4610.

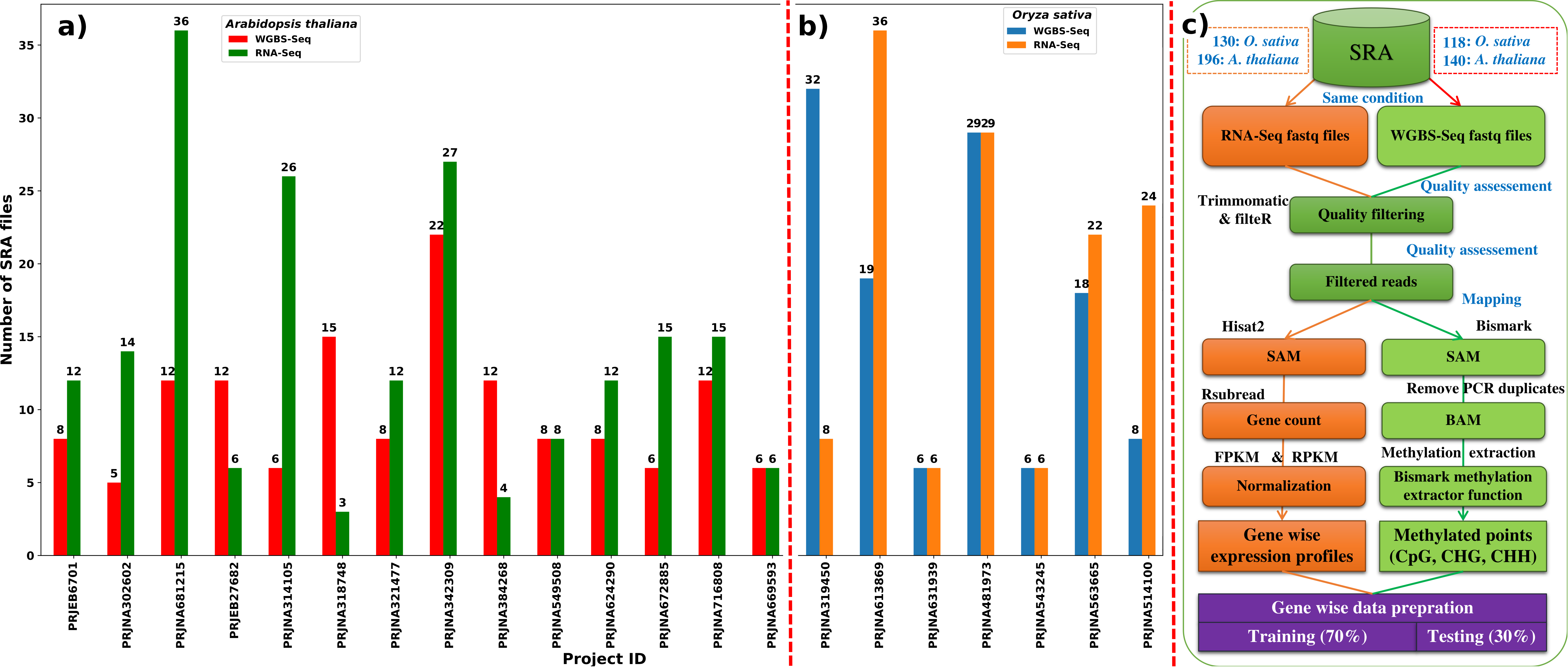
56. Jin,J., Yu,Y., Wang,R., Zeng,X., Pang,C., Jiang,Y., Li,Z., Dai,Y., Su,R., Zou,Q., et al. (2022) iDNA-ABF: multi-scale deep biological language learning model for the interpretable prediction of DNA methylations. Genome Biol, 23, 219.

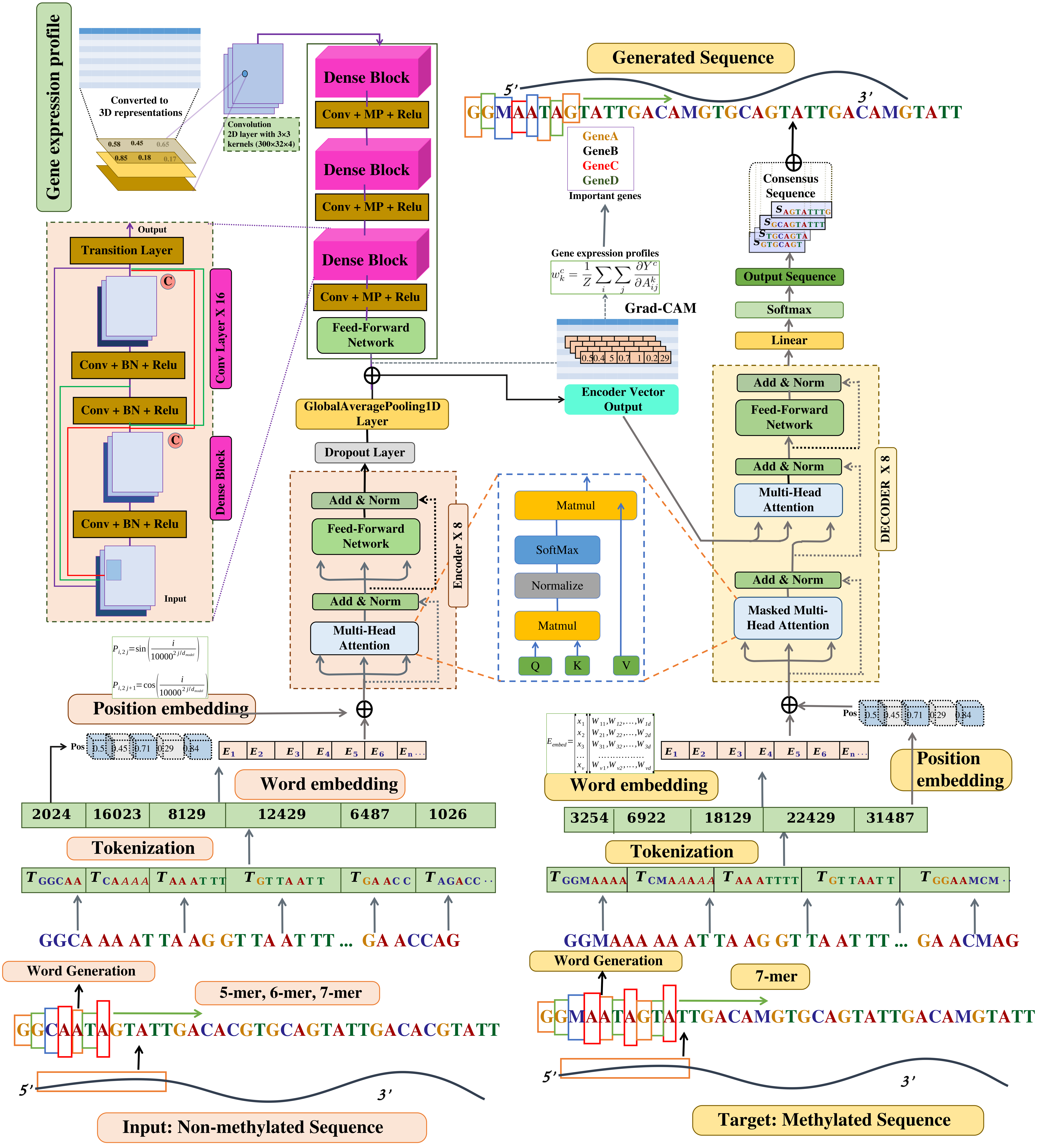
57. Zeng,W., Gautam,A. and Huson,D.H. (2023) MuLan-Methyl—multiple transformer-based language models for accurate DNA methylation prediction. GigaScience, 12, giad054.

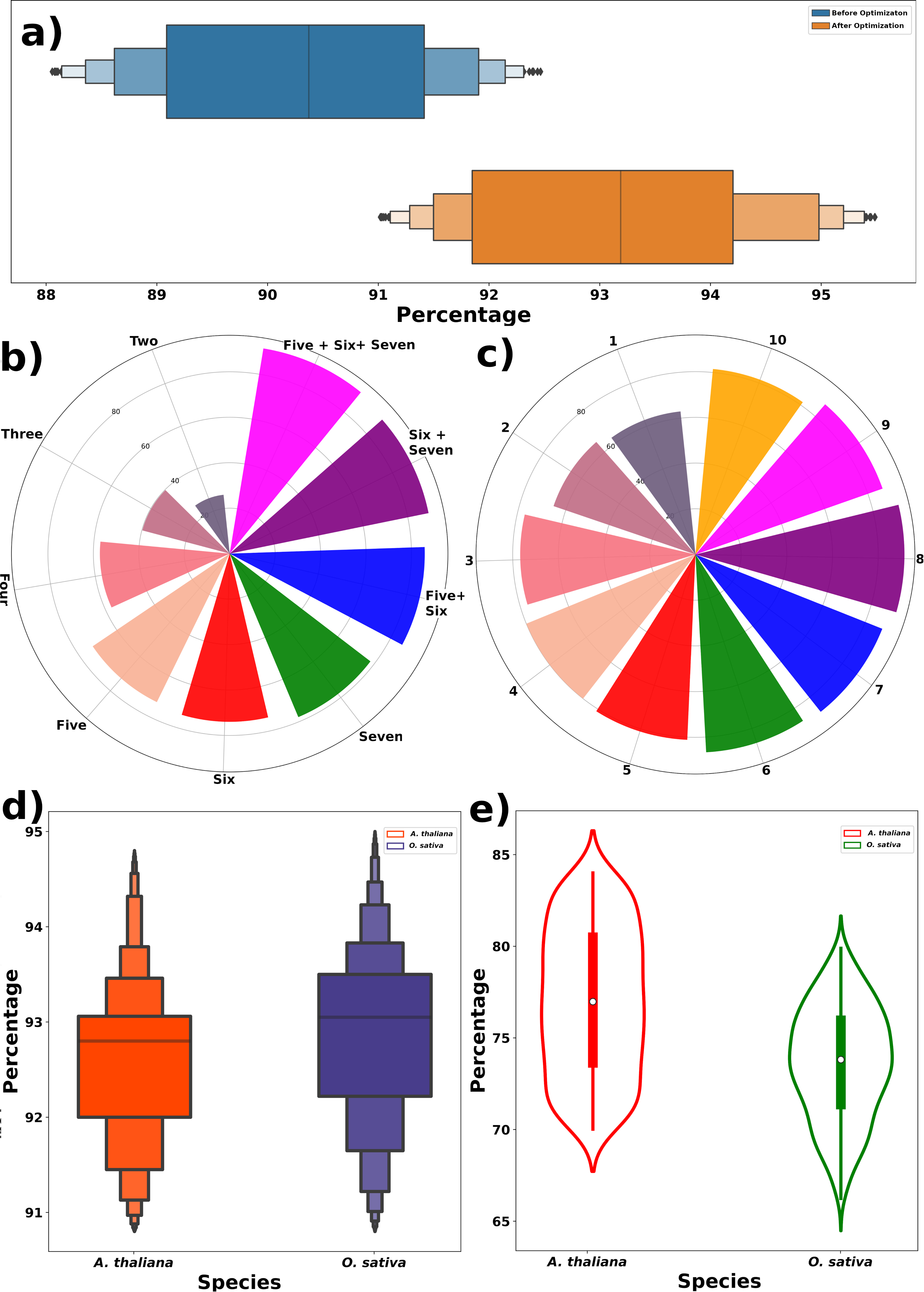
58. Brewer,P.B., Howles,P.A., Dorian,K., Griffith,M.E., Ishida,T., Kaplan-Levy,R.N., Kilinc,A. and Smyth,D.R. (2004) PETAL LOSS, a trihelix transcription factor gene, regulates perianth architecture in the Arabidopsis flower. Development, 131, 4035–4045.

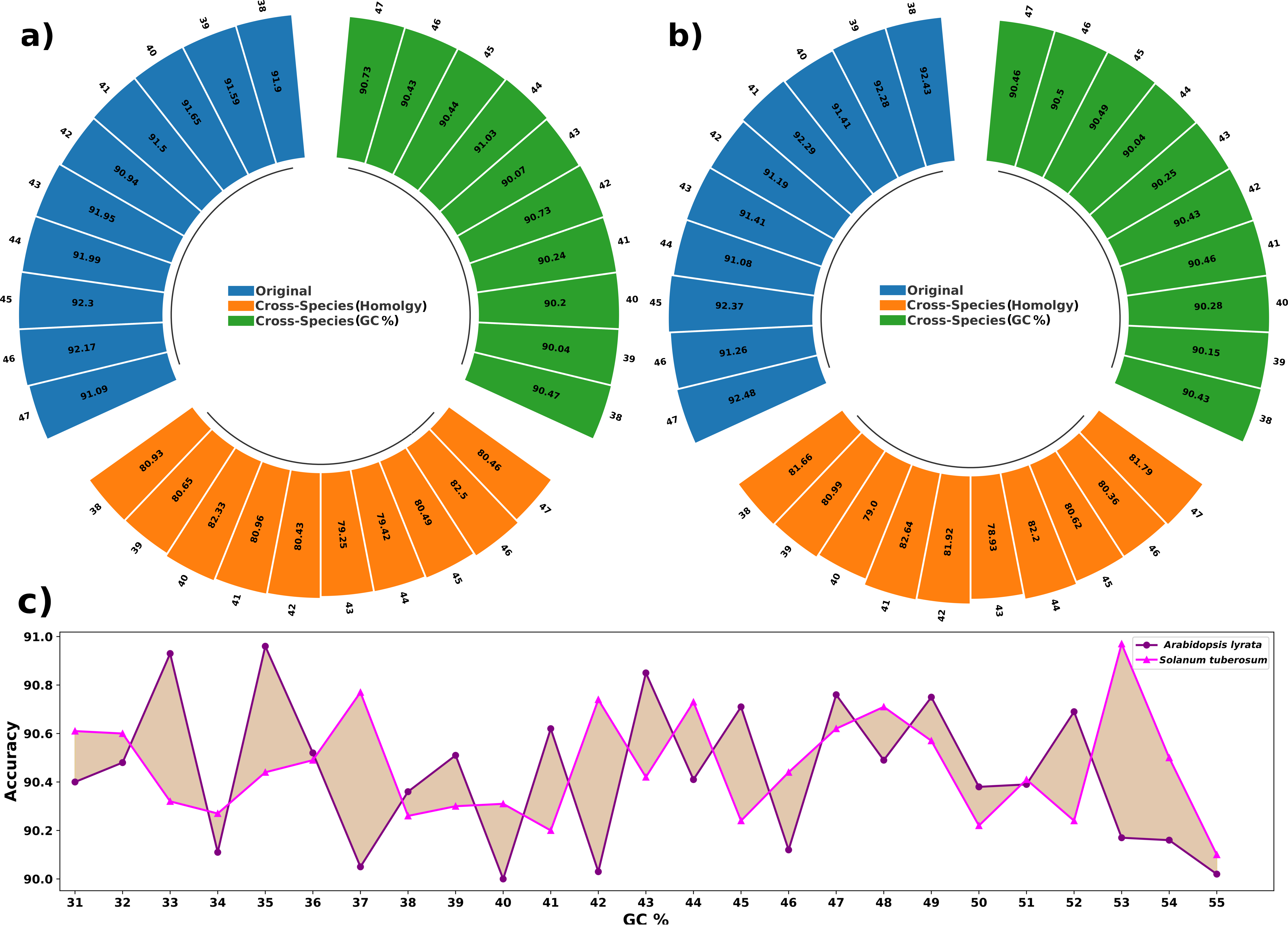
59. Huala,E. and Sussex,I.M. (1992) LEAFY Interacts with Floral Homeotic Genes to Regulate Arabidopsis Floral Development. The Plant Cell, 4, 901–913.

a)
CH3
ATAACGTCGATAT-3`
TATTGCAGCTATA-5`
CH3
Bisulfite
Treatment
ATAACGTCGATAT-3`
TATTGCAGCTATA-5`
CH3
PCR
ATAACGTCGATAT-3`
TATTGCAGCTATA-5`
ATAACGTCGATAT-3`
TATTGCAGCTATA-5`
NH2
HSO3-
OH-
cytosine
cytosine sulphonate
SO3-
H2O
NH4+
uracil sulphonate
OH-
HSO3-
uracil
5-methylcytosine
CH3
b)
MeDIP-Seq
RRBS-Seq
Number of SRA files
150
125
100
75
50
25
Species
Arabidopsis lyrata
Arabidopsis thaliana
Arabis alpina
Arachis hypogaea
Brassica rapa
Glycine max
Gossypium hirsutum
Oryza sativa
Phaseolus vulgaris
Populus deltoides
Populus nigra
Populus trichocarpa
Solanum lycopersicum
Solanum pimpinellifolium
Sorghum bicolor
Triticum aestivum
Vitis vinifera
Zea mays
Betula papyrifera
Cicer arietinum
Hordeum vulgare
Physcomitrella patens
Picea abies
Quercus lobata
Scopelophila cataractae
Taraxacum officinale
Vachellia drepanolobium


a)
Arabidopsis thaliana
WGBS-Seq
RNA-Seq
Number of SRA files
0
5
10
15
20
25
30
35
8
12
5
14
12
36
12
6
6
26
15
3
8
12
22
27
12
4
8
8
8
12
6
15
12
15
6
6
PRJEB6701
PRJNA302602
PRJNA681215
PRJEB27682
PRJNA314105
PRJNA318748
PRJNA321477
PRJNA342309
PRJNA384268
PRJNA549508
PRJNA624290
PRJNA672885
PRJNA716808
PRJNA669593
Project ID
b)
Oryza sativa
WGBS-Seq
RNA-Seq
32
8
19
36
6
6
29
29
6
6
18
22
8
24
PRJNA319450
PRJNA613869
PRJNA631939
PRJNA481973
PRJNA543245
PRJNA563665
PRJNA514100
c)
130: O. sativa
196: A. thaliana
SRA
118: O. sativa
140: A. thaliana
Same condition
RNA-Seq fastq files
WGBS-Seq fastq files
Quality assessement
Trimmomatic & filteR
Quality filtering
Quality assessement
Filtered reads
Mapping
Hisat2
Bismark
SAM
SAM
Rsubread
Remove PCR duplicates
Gene count
BAM
FPKM & RPKM
Methylation extraction
Normalization
Bismark methylation extractor function
Gene wise expression profiles
Methylated points (CpG, CHG, CHH)
Gene wise data prepration
Training (70%)
Testing (30%)


Gene expression profile
Converted to 3D representations
0.58 0.45 0.65
0.85 0.18 0.17
Convolution 2D layer with 3×3 kernels (300×32×4)
Dense Block
Conv + MP + Relu
Dense Block
Conv + MP + Relu
Dense Block
Conv + MP + Relu
Feed-Forward Network
Output
Transition Layer
C
Conv + BN + Relu
Conv + BN + Relu
C
Conv + BN + Relu
Input
Conv Layer X 16
Dense Block
GlobalAveragePooling1D Layer
Dropout Layer
Add & Norm
Feed-Forward Network
Add & Norm
Multi-Head Attention
Encoder X 8
Matmul
SoftMax
Normalize
Matmul
Q
K
V
Generated Sequence
5′
3′
GGMAATAGTATTGACAMGTGCAGTATTGACAMGTATT
GeneA
GeneB
GeneC
GeneD
Important genes
Gene expression profiles
$w_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial Y^c}{\partial A_{ij}^k}$
Grad-CAM
0.5 0.4 5 0.7 1 0.2 29
Encoder Vector Output
Consensus Sequence
$S_{AGTATTTG}$
$S_{GCAGTATTT}$
$S_{TGCAGTA}$
$S_{GTGCAGT}$
Output Sequence
Softmax
Linear
Add & Norm
Feed-Forward Network
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
DECODER X 8
$P_{i,2j} = \sin\left(\frac{i}{10000^{2j/d_{model}}}\right)$
$P_{i,2j+1} = \cos\left(\frac{i}{10000^{2j/d_{model}}}\right)$
Position embedding
Pos 0.5 0.45 0.71 0.29 0.84
$E_1$ $E_2$ $E_3$ $E_4$ $E_5$ $E_6$ $E_n$...
Word embedding
2024 16023 8129 12429 6487 1026
Tokenization
$T_{GGCAA}$ $T_{CAAAA}$ $T_{AAATTT}$ $T_{GTTAATT}$ $T_{GAACC}$ $T_{AGACC}$...
GGCA AA AT TA AG GT TA AT TT ... GA ACCAG
Word Generation
5-mer, 6-mer, 7-mer
GGCAATAGTATTGACACGTGCAGTATTGACACGTATT
5′
3′
Input: Non-methylated Sequence
$E_{embed} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_v \end{bmatrix} \begin{bmatrix} W_{11}, W_{12}, \ldots, W_{1d} \\ W_{21}, W_{22}, \ldots, W_{2d} \\ W_{31}, W_{32}, \ldots, W_{3d} \\ \ldots \\ W_{v1}, W_{v2}, \ldots, W_{vd} \end{bmatrix}$
Word embedding
$E_1$ $E_2$ $E_3$ $E_4$ $E_5$ $E_6$ $E_n$...
Pos 0.5 0.45 0.71 0.29 0.84
Position embedding
3254 6922 18129 22429 31487
Tokenization
$T_{GGMAAAA}$ $T_{CMAAAAA}$ $T_{AAATTTT}$ $T_{GTTAATT}$ $T_{GGAAMCM}$...
GGMAAA AA AT TA AG GT TA AT TT ... GA ACMAG
Word Generation
7-mer
GGMAATAGTATTGACAMGTGCAGTATTGACAMGTATT
5′
3′
Target: Methylated Sequence


a)
Before Optimizaton
After Optimization
88
89
90
91
92
93
94
95
Percentage
b)
Two
Five + Six+ Seven
Three
Six +
Seven
Four
Five+
Six
Five
Six
Seven
20
40
60
80
c)
1
10
2
9
3
8
4
7
5
6
20
40
60
80
d)
A. thaliana
O. sativa
95
94
93
92
91
Percentage
A. thaliana
O. sativa
Species
e)
A. thaliana
O. sativa
85
80
75
70
65
Percentage
A. thaliana
O. sativa
Species


a)
Original
Cross-Species (Homolgy)
Cross-Species (GC %)
38 91.9
39 91.59
40 91.65
41 91.5
42 90.94
43 91.95
44 91.99
45 92.3
46 92.17
47 91.09
47 90.73
46 90.43
45 90.44
44 91.03
43 90.07
42 90.73
41 90.24
40 90.2
39 90.04
38 90.47
38 80.93
39 80.65
40 82.33
41 80.96
42 80.43
43 79.25
44 79.42
45 80.49
46 82.5
47 80.46
b)
Original
Cross-Species (Homolgy)
Cross-Species (GC %)
38 92.43
39 92.28
40 91.41
41 92.29
42 91.19
43 91.41
44 91.08
45 92.37
46 91.26
47 92.48
47 90.46
46 90.5
45 90.49
44 90.04
43 90.25
42 90.43
41 90.46
40 90.28
39 90.15
38 90.43
38 81.66
39 80.99
40 79.0
41 82.64
42 81.92
43 78.93
44 82.2
45 80.62
46 80.36
47 81.79
c)
*Arabidopsis lyrata*
*Solanum tuberosum*
Accuracy
91.0
90.8
90.6
90.4
90.2
90.0
GC %
31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55

a)
b)
c)
d)
Condition No
12
11
10
9
8
7
6
5
4
3
2
1
Upstream in bases
0
100
200
300
400
500
methylation state
C
M

a)
Number of Times
40
38
36
34
32
30
AT1G04367
AT1G06727
AT1G12820
AT1G19394
AT1G19490
AT1G28290
AT1G32240
AT1G43670
AT1G47280
AT1G56750
AT1G60400
AT1G61040
AT1G64100
AT1G67250
AT1G74600
AT1G80820
AT2G01540
AT2G03870
AT2G03880
AT2G05705
AT2G05812
AT2G07055
AT2G08570
AT2G08900
AT2G12170
AT2G18938
AT2G18940
AT2G19810
AT2G21660
AT2G21670
AT2G22430
AT2G28350
AT2G32050
AT2G33830
AT2G40280
AT2G45470
AT2G46050
AT3G07395
AT3G18720
AT3G28920
AT3G51440
AT3G58340
AT3G62980
AT4G08965
AT4G17483
AT4G17770
AT4G37870
AT4G39260
AT5G01060
AT5G01385
AT5G04640
AT5G05125
AT5G07570
AT5G09565
AT5G11090
AT5G26570
AT5G27330
AT5G39700
AT5G48905
AT5G51590
AT5G58510
AT5G62000
AT5G66100
Genename
b)
16
12
8
4
0
0
5
10
AT1G13450 target gene
GERMAN MIR393B MicroRNA Targets
GERMAN MIR393B target gene
MYB101 COL A Transcription Factor Targets
GO:0048438 floral whorl development
PPD2 TARGET GENES Transcription Factor Targets
LAFOS AUXIN-PERCEPTION-SIGNALING PCG H3K27ME3 DIFF
R-ATH-5608118.1 Auxin signalling
AG-MUTANT BOLTING DNLiterature Reference gene sets
AT3G25990 COLAMP A Transcription Factor Targets
AT3G25990 target gene
GO:0035198
GT1 COL A Transcription Factor Targets
GUSTAFSON MIR393 MicroRNA Targets
LU MIR393A MicroRNA Targets
GERMAN AT1G13450 MIR393B target
GO:0048438 MYB101 floral development
Auxin AUXIN-PERCEPTION-SIGNALING GENES DIFF
AT3G25990 AG-MUTANT BOLTING COLAMP
GTI GUSTAFSON LU MIR393
number of genes
10
15
20
25
color
0.04
0.03
0.02
0.01
c)
8
6
4
2
0
0
3
6
Regulatory RNA binding
Extracellular transport
Auxin binding
MicroRNA binding
SiRNA binding
Response to virus
MiRNA transport
Extracelluar Regulatory RNA transport
Auxin MicroRNA binding
SiRNA binding
Response to virus
MiRNA transport
number of genes
5
10
15
20
25
color
0.04
0.03
0.02
d)
7.5
5.0
2.5
0.0
0
5
Response to alcohol
Extracellular tranport
Floral organ development
MiRNA transport
Response to lipid
Response to abscisic acid
Floral whorl development
Response to virus
Extracellular Response transport alcohol
Floral organ development
MiRNA transport
Response abscisic lipid acid
Floral whorl virus development
number of genes
10
15
20
25
30
color
0.04
0.03

a
DMRU
DNA Methylation Regulation Unit
The Computing DNA of Himalayas
HIMCoB
Home
Implementation
About Authors
About Us
DNA Methylation Regulation Unit
Enter your promoter sequences ( Minimum length of Sequence 2,000 base )
Paste your sequence here
Use Example
Clear Sequences
Download Example files
OR Upload Files
Promoter Sequence File
Gene Expression File
Transcriptome Sequence File
Submit

About DMRU...

Epigenetic regulation of gene expression is a complex and multifaceted process that plays a crucial role in plant development and adaptation. DNA methylation at cytosines has been shown to influence gene expression by regulating chromatin structure and accessibility. Some approaches have been developed to identify DNA methylation but none of them are capable to tell the condition specific DNA methylation, making them hardly of any use. Here, we report a first of its kind an explainable Deep Encoders-Decoders generative system, DMRU, which learns the relationship between transcritpome status and DNA methylation states at any given time. It was also found that instead of looking for homology to relate condition specificity of DNA methylation marks across the species, looking for GC similarity is more related to the specificity of DNA methylation patterns. The developed approach, DMRU, is capable to discover the methylated cytosines in total independence from data volume and species specific models, and works with same performance across different species in universal manner. In a comprehensive benchmarking study across a huge volume of experimental data covering different conditions and species, it has consistently achieved >90% accuracy. With this all, DMRU brings a completely new chapter in methylated cytosine discovery through generative AI, giving a strong alternative to costly bisulfite sequencing experiments, saving enormous amount of time and money.DMRU may have significant impact on the development of novel strategies for crop improvement and biotechnology applications, while accelerating the research in the field of epigenomics and gene regulation.

Citation:

**DMRU: Generative Deep-Learning to unravel condition specific cytosine methylation in plants**
Sagar Gupta, Anchit Kumar, Veerbhan Kesarwani, Ravi Shankar* **bioRxiv, 2025** Read research article here

b
Methylation Regulation Unit
Enter your promoter sequences ( Minimum length of Sequence 2,000 base )

```
>Test_Seq1
ATGGTATGTTTTTTTATTGAAAGTGATAAAGACTTACCAGACAATACCATAAGCACCTCGGCC
GATGGGTTCAATAGGAGGGATGTATTTTGAAGATAGCTCGAAAATGTTACCGAGTACATTGTA
CATAACATATCTACCATCATATGTCAAGATCCCTCCATCTTCCCTTTTCTCCATCTCTTCTTTG
GAAGAAGAACTCGGATCAAGACAGGTTTTAAGATTAAAAAAGGAAGCAAGTTTTAAAAGACT
GTTTCTCAAGTAGCTGTTGTTGCCGTTAAAAAGGCAAAGCAGTGAGTAAGAGGAGAGAAAAT
GGTGGCACCTTTTGCTTTTGGTGGGCCAAATGAAGACAAACTTTTTTATTTGATCCGTTGAACT
TCGAAATTTGAATGAGAATTATTATTACAAGGTTAAAACTTGACTTCTGCATCATTAAAGAAAA
ATTCAAAAGAAACAAAAGAGAAGTATACTATAGAAAAAGCTATCTCATTGTGGATCAATAGGT
CCCTAGGAACCAATAGTAAGTAAAAAGCAAAGTGGTAAACATACAAAAGCGAGATTATTCTC
AAGTCTCATTTCAATGTCAAAACTCTCTCTTGTGTGCTAAGTAAACAGAACCTGTGTGAAAGT
AAAACACCTCTGAACTCATTTTCACTATCTCACAAACCTGCATTATCGATCCAATTCTCCCAT
CTTATATCTTCTTACATATTAAGGATCAGAGAAGGTGATCAAACCTGATAATTATGTAGATGTA
GCTGAAGCTAAGATATAAGGCTCAGGAACCACAAAAACATATCACTGAATATTTCAAATTAG
CTTCAGCATTTTATGAGTCAAAATACTAAGAAACAATGCAAATTCTCTCATGAAAAGGCTCTC
```

Use Example
Clear Sequences
Download Example files
OR Upload Files
Promoter Sequence File
Gene Expression File
Transcriptome Sequence File
Submit
DMRU
Gene expression profile
Converted to 3D representations
0.58 0.45 0.65
0.85 0.18 0.17
Convolution 2D layer with 3x3 kernels (300×32×4)
Dense Block
Conv + MP + Relu
Dense Block
Conv + MP + Relu
Dense Block
Conv + MP + Relu
Feed-Forward Network
Output
Transition Layer
C
Conv + BN + Relu
Conv + BN + Relu
C
Conv + BN + Relu
Input
Conv Layer X 16
Dense Block
GlobalAveragePooling1D Layer
Dropout Layer
Add & Norm
Feed-Forward Network
Add & Norm
Multi-Head Attention
Encoder X 8
Matmul
SoftMax
Normalize
Matmul
Q
K
V
Generated Sequence
5'
3'
GGMAATAGTATTGACAMGTGCAGTATTGACAMGTATT
GeneA
GeneB
GeneC
GeneD
Important genes
Gene expression profiles
$w_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial Y^c}{\partial A_{ij}^k}$
Grad-CAM
0.5 0.4 5 0.7 1 0.2 29
Encoder Vector Output
Consensus Sequence
AGTATTTG
GCAGTATTT
TGCAGTA
GTGCAGT
Output Sequence
Softmax
Linear
Add & Norm
Feed-Forward Network
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
DECODER X 8
$P_{i,2j} = \sin\left(\frac{i}{10000^{2j/d_{model}}}\right)$
$P_{i,2j+1} = \cos\left(\frac{i}{10000^{2j/d_{model}}}\right)$
Position embedding
Pos 0.5 0.45 0.71 0.29 0.84
$E_1$ $E_2$ $E_3$ $E_4$ $E_5$ $E_6$ $E_n$...
Word embedding
2024 16023 8129 12429 6487 1026
Tokenization
$T_{GGCAA}$ $T_{CAAAA}$ $T_{AAATTT}$ $T_{GTTAATT}$ $T_{GAACC}$ $T_{AGACC}$
GGCA AA AT TA AG GT TA AT TT ... GA ACCAG
Word Generation
5-mer, 6-mer, 7-mer
GGCAATAGTATTGACACGTGCAGTATTGACACGTATT
5'
3'
Input: Non-methylated Sequence
$E_{embed} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \dots \\ x_v \end{bmatrix} \begin{bmatrix} W_{11}, W_{12}, \dots, W_{1d} \\ W_{21}, W_{22}, \dots, W_{2d} \\ W_{31}, W_{32}, \dots, W_{3d} \\ \dots \\ W_{v1}, W_{v2}, \dots, W_{vd} \end{bmatrix}$
$E_1$ $E_2$ $E_3$ $E_4$ $E_5$ $E_6$ $E_n$...
Pos 0.5 0.45 0.71 0.29 0.84
Word embedding
Position embedding
3254 6922 18129 22429 31487
Tokenization
$T_{GGMAAAA}$ $T_{CMAAAAA}$ $T_{AAATTTT}$ $T_{GTTAATT}$ $T_{GGAAMCM}$
GGMAAA AA AT TA AG GT TA AT TT ... GA ACMAG
Word Generation
7-mer
GGMAATAGTATTGACAMGTGCAGTATTGACAMGTATT
5'
3'
Target: Methylated Sequence
c
Results
Identified Methylated Cytosines In The Given Sequences
Importance Score Of Various Genes Responsible For Methylation In The Given Sequence
Download
Seq IDs
Sequences

```
>Test_Seq1  ATGGTATGTTTTTTTATTGAAAGTGATAAAGACTTACCAGACAATACCATAAGCACCTCGGCCGATGGGTTCAATAGGAGGGATGTATTTTGAAGATAGCTCGAAAATGTTACCGAGTACATTGTACATAACATATCTACCATCATATGTCAAGATCCCTCCATCTTCCCTTTTCTCCATCT
            ATGGTATGTTTTTTTATTGAAAGTGATAAAGACTTACCAGAMAATACMATAAGCACCTCGGCCGATGGGTTMAATAGGAGGGATGTATTTTGAAGATAGMTCGAAAATGTTAMMGAGTACATTGTACATAACATATCTACMATMATATGTMAAGATCCCTMCATCTTCMMTTTTCTMMATCT

>Test_Seq2  ATGGTATGTTTTTTTATTGAAAGTGATAAAGACTTACCAGACAATACCATAAGCACCTCGGCCGATGGGTTCAATAGGAGGGATGTATTTTGAAGATAGCTCGAAAATGTTACCGAGTACATTGTACATAACATATCTACCATCATATGTCAAGATCCCTCCATCTTCCCTTTTCTCCATCT
            ATGGTATGTTTTTTTATTGAAAGTGATAAAGAMTTAMCAGAMAATACCATAAGCAMCTMGGMCGATGGGTTCAATAGGAGGGATGTATTTTGAAGATAGCTCGAAAATGTTACMGAGTACATTGTACATAACATATCTACCATCATATGTCAAGATMCMTCMATMTTMMCTTTTMTMMATCT

>Test_Seq3  ATGGTATGTTTTTTTATTGAAAGTGATAAAGACTTACCAGACAATACCATAAGCACCTCGGCCGATGGGTTCAATAGGAGGGATGTATTTTGAAGATAGCTCGAAAATGTTACCGAGTACATTGTACATAACATATCTACCATCATATGTCAAGATCCCTCCATCTTCCCTTTTCTCCATCT
            ATGGTATGTTTTTTTATTGAAAGTGATAAAGACTTAMCAGACAATAMCATAAGCACCTCGGCCGATGGGTTCAATAGGAGGGATGTATTTTGAAGATAGCTCGAAAATGTTACCGAGTACATTGTACATAACATATMTACCATCATATGTCAAGATMCCTCMATCTTMCCTTTTMTMCATMT

>Test_Seq4  ATGGTATGTTTTTTTATTGAAAGTGATAAAGACTTACCAGACAATACCATAAGCACCTCGGCCGATGGGTTCAATAGGAGGGATGTATTTTGAAGATAGCTCGAAAATGTTACCGAGTACATTGTACATAACATATCTACCATCATATGTCAAGATCCCTCCATCTTCCCTTTTCTCCATCT
            ATGGTATGTTTTTTTATTGAAAGTGATAAAGAMTTAMCAGACAATAMCATAAGCACCTCGGCCGATGGGTTMAATAGGAGGGATGTATTTTGAAGATAGCTMGAAAATGTTACCGAGTACATTGTAMATAAMATATMTAMMATCATATGTCAAGATCMCTCCATCTTCCCTTTTCTMCATCT
```

d
Results
Identified Methylated Cytosines In The Given Sequences
Importance Score Of Various Genes Responsible For Methylation In The Given Sequence
Download
Test_Seq1 ▼
Test_Seq1
Importance Score
0
1
2
3
4
5
6
Gene
AT4G05690
AT1G09143
AT3G41768
AT5G22580
AT4G25050
AT4G04225
AT5G48490
AT4G40030
AT1G76080
AT3G09630
AT1G43170
AT3G27690
AT2G30860
AT5G26000
AT5G40950
AT1G55490
AT5G09810
AT1G37130
AT1G07940
AT1G33120
AT5G14740
AT3G04400
AT5G47700
AT2G10940
AT4G00100
AT2G32690
AT5G54770
AT4G27090
AT5G26742
AT3G62870
AT1G07890
AT4G35100
AT5G20290
AT1G52220
AT3G47070
AT2G16600
AT3G08580
AT3G44310
AT4G18730
AT3G15360
AT4G13940
AT1G04800
AT4G27520
AT1G78370
AT4G13770
AT4G34670