Prediction of DNA Methylation With Long-Range State-Space Models

Hao Feng, Sakshar Chakravarty, and Stefano Lonardi, *Fellow, IEEE*

***Abstract*—The prediction of DNA methylation from the primary DNA sequence allows one to impute the methylation status of cytosines with insufficient sequencing coverage. Various deep learning models have been proposed in the literature, including transformer-based models and convolutional neural networks. In this study, we investigate the performance of long-range state-space models based on the Hyena architecture on the task of DNA methylation prediction on six plant species. First, we train the HyenaDNA framework to obtain a genome-wide foundation model for each species. Then, we fine-tune these foundation models using the sequence data surrounding the methylated or unmethylated cytosines. Extensive experimental results show that our model predicts DNA methylation with higher accuracy than state-of-the-art methods in the literature.**

***Index Terms*—State-space model, HyenaDNA, deep learning, DNA methylation, epigenetics.**

I. Introduction

DNA methylation is an epigenetic marker which involves the addition of a methyl group to cytosine bases. It plays a critical role in regulating gene expression and other important cellular molecular processes [1]. In vertebrates, DNA methylation is mostly found at cytosines that are followed by a guanine, which is referred to as the CG context. In plants instead, methylation can occur at cytosines in any context, i.e., CG, CHG, and CHH, where H={A,T,C} [2], [3], [4]. Methylation patterns influence critical cellular processes such as stress responses, phenotypic plasticity, and trait inheritance [5], [6]. The traditional method of identifying cytosine methylation involves sequencing DNA after being treated with sodium bisulfite (BS-Seq). BS-Seq can be costly for large eukaryotic genomes because the sequencing coverage has to be sufficiently deep to enable one to determine the methylation levels with statistical confidence. Typically, 5-10 reads covering a cytosine are necessary to reliably determine the methylation levels. Due to the non-uniformity of Illumina sequence coverage, even if the average sequencing depth is higher than 10×, a small but significant fraction of the cytosines will not have sufficient read coverage. Thus, there is a need to develop computational methods to predict the methylation levels for some of the cytosines either from the surrounding DNA sequence, or the methylation levels of nearby cytosines, or other local epigenetic signals. The prediction of methylation from the primary DNA sequence is the most convenient because it does not require any additional data or wet lab experiment, but it is also the most computationally challenging due to complex sequence-context dependencies and long-range interactions across the genome.

To address the need to impute the methylation levels for the cytosines with low sequencing coverage, several deep learning methods have been proposed in the literature to predict the methylation status from the sequence or the methylation levels of nearby cytosines [7], [8], [9], [10]. These methods include (i) CpGenie, which is a classifier based on a convolutional neural network (CNN) [11], (ii) DeepCpG, which uses a CNN and recurrent neural network [12], (iii) AMPS, which is another CNN-based classifier oriented to predict methylation in plants [13], (iv) scGPT, which is a model based on transformers and trained on single-cell data [14], (v) iDNA-ABF, which uses a bidirectional transformer (BERT) [15], [16], [17], and (vi) BERT-5mC, which also uses a BERT architecture [18]. Transformer-based methods have also been employed for predicting RNA methylation from sequence data [19]. We should note that, with the exception of AMPS, all of these DNA methylation prediction methods have focused on CG methylation in humans or other vertebrates. To the best of our knowledge, only AMPS can predict methylation in all CG, CHG, and CHH contexts, for the benefit of non-vertebrate species like plants. Based on the results in these studies, transformer-based models appear to produce the most accurate predictions from the DNA sequence, which we attribute to their powerful self-attention mechanisms. However, the attention mechanism relies on computing the relationship between every pair of tokens in a sequence; thus it incurs quadratic time complexity with respect to sequence length during both training and evaluation, making it inefficient for long DNA sequences [20], [21]. Since methylation potentially involves interactions across thousands of base pairs, there is a need for models that can efficiently capture long-range dependencies without excessive computational overhead [22], [23]. State-space models (SSMs), traditionally used for time series modeling, have recently been explored as an alternative to transformers for sequence-based tasks [24]. Unlike transformers, which use attention to explicitly model interactions between all sequence positions, SSMs use state representations to propagate information through structured recurrence. As shown by the Structured State Space sequence model and Hungry Hungry Hippos, it is possible to achieve competitive performance compared to transformers while demanding lower computation and memory costs [25], [26]. HyenaDNA adapts the SSM-based Hyena architecture to achieve subquadratic complexity in genomic sequence modeling [27], [28]. HyenaDNA employs long convolutions and data-controlled gating mechanisms to implicitly capture long-range dependencies as a replacement for the more computationally expensive attention mechanism of transformers. This architecture is particularly beneficial for genomic tasks, where regulatory elements can influence gene expression over thousands of base pairs. By leveraging memory-efficient computation over long context lengths, HyenaDNA has shown strong performance in long-range predictive tasks, e.g., chromatin profile prediction, species classification, among others.

In this work, we focus our attention on plant genomes due to their more diverse type of methylation contexts compared to mammalian genomes [29], [30]. Since plants have significantly more non-CG methylation than mammals, predictive methods trained on plants are more general compared to methods trained only on mammalian genomes. For the same reason, we used AMPS [13] as our performance benchmark because it has demonstrated strong predictive performance across all CG, CHG, and CHH contexts.

Received 30 May 2025; revised 8 October 2025; accepted 10 November 2025. Date of publication 19 November 2025; date of current version 10 December 2025. This work was supported in part by NSF IIS under Grant 2444456 to SL. *(Corresponding author: Stefano Lonardi.)*

The authors are with the Department of Computer Science, Engineering University of California, Riverside, Riverside, CA 92521 USA (e-mail: hfeng031@ucr.edu; schak026@ucr.edu; stelo@ucr.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/TCBBIO.2025.3634738, provided by the authors.

Digital Object Identifier 10.1109/TCBBIO.2025.3634738

In this study, we investigate for the first time the performance of HyenaDNA in predicting DNA methylation from the primary DNA sequence, taking advantage of its ability to model long-range genomic interactions while maintaining computational scalability. We train the Hyena architecture using the DNA sequence surrounding methylated and unmethylated cytosines in six plant species. As it was done in AMPS, we predict cytosine methylation in all contexts, i.e., CG, CHG, and CHH. However, our architecture uses the DNA sequence exclusively, while AMPS also takes advantage of genomic annotations. We investigate the performance of three model sizes for HyenaDNA by pretraining the HyenaDNA framework on the genomes of the six plant species. Finally, we compare the predictive accuracy of HyenaDNA against other published methods, and we also examine the performance when predicting methylation across contexts and across species. The ability of HyenaDNA to "transfer" predictive knowledge from one species to another would be very valuable, because obtaining high-quality methylation data for a new species through bisulfite sequencing is laborious and expensive.

II. Problem Formulation

Let a genome be represented as a sequence G over the alphabet $\{A, C, G, T\}$. Given a cytosine site $c \in G$, we denote by $S(c)$ the subsequence of length L centered at c. We define $M(c) \in \{0, 1\}$ as the methylation state of cytosine c, where $M(c) = 1$ if c is methylated, and $M(c) = 0$ otherwise. The problem can thus be expressed as predicting the mapping $f : S(c) \rightarrow M(c)$. In other words, the objective is to design a system which predicts the methylation state of a cytosine given the DNA sequence in the window containing the selected cytosine at its center.

III. Methods

Our approach follows a two-stage framework of pre-training and fine-tuning using HyenaDNA. The pre-training input consists of genomic DNA sequences. During pre-training, we train a HyenaDNA backbone of two Hyena layers using a masked language modeling objective. In the fine-tuning stage, we pair the genomic DNA sequences surrounding cytosines with binary methylation labels for CG, CHG, or CHH contexts as inputs and use a decoder layer connected to the backbone to perform the binary classification of the methylation status. We use balanced sampling of methylated and unmethylated sites and an 80:10:10 train/validation/test split.

Since the original HyenaDNA foundation model was trained on the human genome, we retrained our instances of HyenaDNA from scratch, and then fine-tuned them for the methylation prediction task. Our HyenaDNA models were trained on the genomes of tomato (*Solanum lycopersicum*), thale cress (*Arabidopsis thaliana*), rice (*Oryza sativa*), cowpea (*Vigna unguiculata*), cucumber (*Cucumis sativus*), and common liverwort (*Marchantia polymorpha*).

For each species, we tested three model sizes for HyenaDNA, namely 128 k, 436 k, and 940 k parameters. These model sizes correspond to two Hyena layers with a width of 64, 128, and 192, respectively (as shown in Table I). The original HyenaDNA study describes experiments using up to eight Hyena layers, but we chose a smaller depth for our models due to the relatively small context range used for our tasks.

For the training dataset, we fragmented the genomes into sequences of 6,400 bp, twice the length that would be used for the downstream classification task, to ensure that the models are provided adequate contextual information. We used a single nucleotide tokenizer (where each nucleotide A, T, C, G is treated as a discrete token, along with special tokens for sequence boundaries and padding) as this approach performed best on the classification tasks in [27], [28].

TABLE I
Hyperparameters of HyenaDNA

	HyenaDNA-128k	HyenaDNA-436k	HyenaDNA-940k
Model size	128k	436k	940k
Model width	64	128	192
Decoder head size	130	258	386
Architecture	Two Hyena layers + MLP decoder layer		
Optimizer	AdamW		
Batch size	128		
Tokenization	single nucleotide		

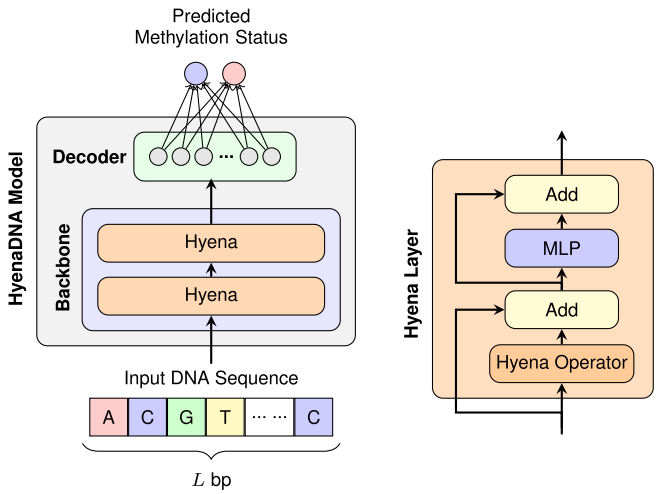

Fig. 1. The HyenaDNA model structure; two Hyena layers constitute the architecture backbone (pre-training); a downstream decoder layer is used for classification (fine-tuning); input sequences are processed using a single-nucleotide tokenizer; the right side shows the components of the Hyena layer: the Hyena operator combines long convolutions with element-wise gating, and the long convolutions are implicitly parameterized by an MLP [27].

For the downstream task, we fine-tuned the base species-specific models to predict the methylation status in the contexts of CG, CHG, and CHH. We first extracted genomic sequences of length 3,200 bp around each methylated/unmethylated cytosine for the selected context (1,600 bp upstream and 1,600 bp downstream of the cytosine site). The window size of 3,200 bp was chosen because it was shown to be optimal in [13]. A linear classifier downstream of the HyenaDNA model was trained via supervised learning on the sequence dataset and their associated methylation labels. The architecture of our classifier is illustrated in Fig. 1. We applied global average pooling over all sequence positions before passing the output to the decoder classification layer, as opposed to using the last token embedding. In both the pre-training and fine-tuning steps, we applied dropout with a probability of 0.1 to help mitigate overfitting and increase robustness against unseen sequences. In total, we produced 54 fine-tuned HyenaDNA models (6 species $\times$ 3 model sizes $\times$ 3 contexts).

Each dataset was curated to maintain an equal 50:50 ratio of methylated to unmethylated cytosines, with a total size of 500,000 samples for each dataset. The size of the dataset was chosen to be consistent with [13]. We extracted 250,000 DNA windows (3,200 bp long) centered at a methylated cytosine in the genome and 250,000 DNA windows (3,200 bp long) centered at an unmethylated cytosine in the genome. If the number of methylated cytosines or the number of unmethylated cytosines was lower than 250,000, we took an equal number of samples from both classes. The data was split into an 80:10:10 ratio for the purposes of training, validation, and testing. We trained our models with a batch size of 128. Detailed statistics for each

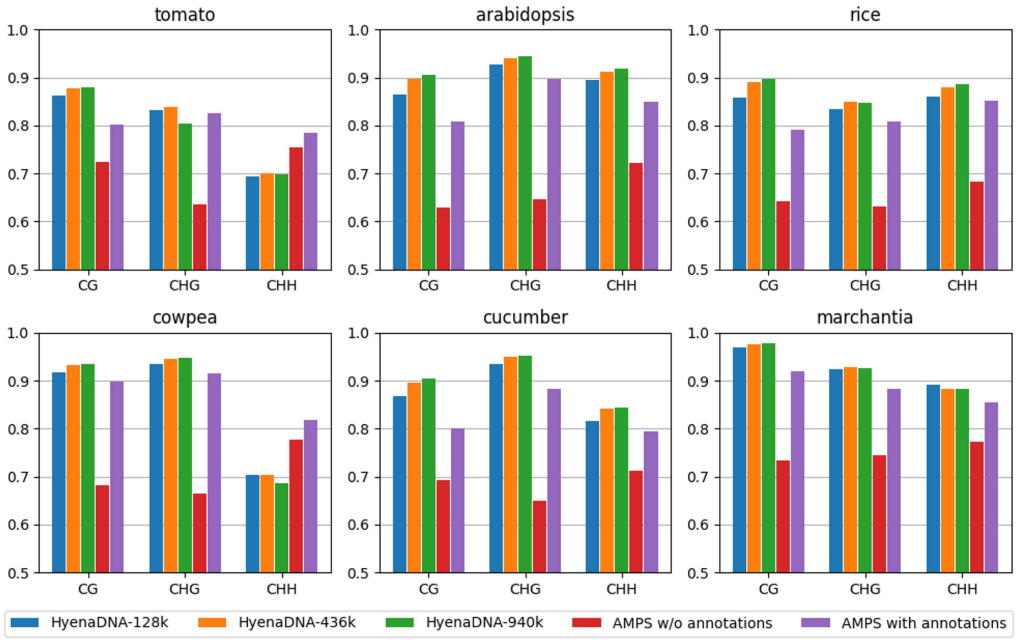

Fig. 2. Context and species-specific accuracy comparisons between HyenaDNA and AMPS (sequence-only and sequence+annotation). In 15 contexts/species combinations out of 18 (83.4%), HyenaDNA models (irrespective of their sizes) outperformed AMPS with and without annotations.

TABLE II
PLANT GENOMES USED IN THIS STUDY

Common name	Accession number
Tomato (*Solanum lycopersicum*)	GCA_000188115.4 (*SL3.1*)
Thale cress (*Arabidopsis thaliana*)	GCA_000001735.2 (*TAIR10.1*)
Rice (*Oryza sativa*)	GCA_001433935.1 (*IRGSP-1.0*)
Cowpea (*Vigna unguiculata*)	GCA_004118075.2 (*ASM411807v2*)
Cucumber (*Cucumis sativus*)	GCA_000004075.3 (*Cucumber_9930_V3*)
Liverwort (*Marchantia polymorpha*)	GCA_003032435.1 (*Marchanta_polymorpha_v1*)

dataset are shown in Supplemental Table I. We trained using an adaptive learning rate schedule, with performance monitored on cross-entropy validation loss for early stopping with a patience of three epochs. We set a maximum limit of 100 epochs for both the pre-training and fine-tuning steps. For the largest models (HyenaDNA-940 k) we used gradient checkpointing during training to reduce memory usage.

IV. EXPERIMENTAL RESULTS

We collected experimental results on three classification problems, namely (i) species and context-specific prediction, where the test data matches the species and context of the trained model, (ii) cross-context prediction, where the test data matches the species of, but differs in context from the trained model, (iii) cross-species prediction, where the test data is of a different species, but from the same context as the trained model. The experiments were performed for the six species listed in Table II, and for the CG, CHG, and CHH contexts. We compared our experimental results with [13]. We did not compare any other published method, because in [13] it was shown that AMPS outperformed iDNA-ABF [17], MRCNN [31], CpGenie [11], and SMEP [32]. We compared our HyenaDNA architecture against AMPS without annotations (i.e., sequence only) and against AMPS with annotations.

A. Species and Context-Specific Performance

We first evaluated each of our fine-tuned HyenaDNA models on the test data set matching both the species and context on which the models were trained. Fig. 2 shows that our HyenaDNA models (irrespective of their sizes) outperformed both sequence-only AMPS and sequence+annotation AMPS in 15 contexts/species combinations out of 18 (83.4%). The highest improvement over AMPS with annotations was achieved by HyenaDNA-940 k on the rice genome for the CG context, with an accuracy increase of 10.5%. The highest accuracy was obtained on the marchantia data set (CG context) by the HyenaDNA-940 k model with an accuracy of 97.7%. We observed that, in general, HyenaDNA models with a higher number of parameters yielded higher accuracies in most cases. In 17 contexts/species combinations out of 18, HyenaDNA-436 k performed better than HyenaDNA-128 k. In 12 contexts/species combinations out of 18, HyenaDNA-940 k performed better than HyenaDNA-436 k. There were a few cases in which the HyenaDNA models yielded a lower accuracy compared to AMPS, namely tomato (CHG and CHH context), and cowpea (CHH context). In the CHG context for tomato, HyenaDNA-436 k performed similarly to AMPS with annotations, while HyenaDNA-128 k performed marginally better and HyenaDNA-940 k performed worse. In the context of CHH in tomato and CHH in cowpea, both AMPS models outperformed HyenaDNA by a significant margin. In terms of precision and recall, we observed a similar performance by HyenaDNA in these contexts as shown in Supplemental Figs. 2 and 3. We do not have a good explanation for the lower predictive performance of HyenaDNA in these three specific cases.

B. Cross-Context Performance

For the cross-context problem, we tested our fine-tuned HyenaDNA models on data from the same species but in different contexts. In total,

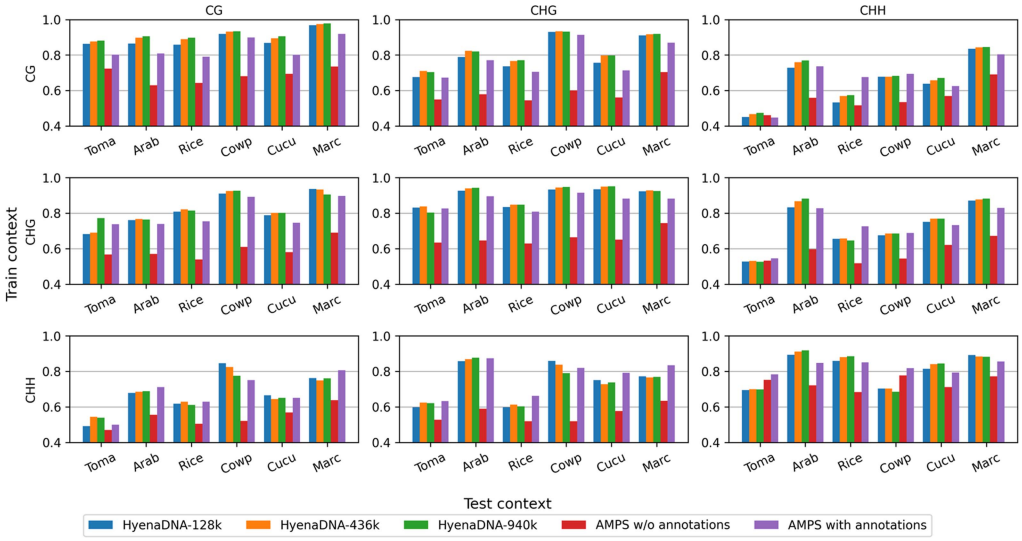

Fig. 3. Cross-context testing accuracy comparisons between HyenaDNA and AMPS. HyenaDNA shows consistent improvement in accuracy scores over AMPS, especially in cross-context cases between CG and CHG, with the 940 k model outperforming both classes of AMPS in 12 of 12 such cases (100%).

we evaluate 36 cross-context cases (two test contexts for each of three training contexts across six species). The test accuracy is shown in Fig. 3. We observed that the HyenaDNA-940 k models outperformed AMPS with annotations in 23 out of 36 cases (64%) and AMPS without annotations in 35 out of 36 cases (97%). The accuracy improvements of HyenaDNA were particularly consistent in cross-context cases between CG and CHG, where the HyenaDNA-940 k models outperformed both AMPS classifiers in 12 out of 12 cases (100%). The improvement of HyenaDNA models for the CHH context was less consistent. In particular, the cross-context accuracy for the CHH context in rice for all HyenaDNA models was lower than AMPS with annotations. Among the HyenaDNA models, the large models achieved the highest accuracy in 16 out of 36 cases (44%), the medium models in 13 out of 36 cases (36%), and the small models in 7 out of 36 cases (19%).

C. Cross-Species Performance

To investigate the ability of HyenaDNA to predict methylation across species, we evaluated our fine-tuned models on test datasets of the same context but in different species. We measured the accuracy across 90 cross-species cases (five test species, six training species across three contexts). Fig. 4 shows that the performance of HyenaDNA on this problem was somewhat inconsistent. For instance, HyenaDNA-128 k trained on the rice CHG context and tested on the cowpea CHG context yielded 82.0% accuracy; however, the same model trained on the marchantia CHH context and tested on the tomato CHH only achieved 48.8% accuracy. Surprisingly, some of these results were asymmetrical; for instance, HyenaDNA-128 k trained on the tomato CHH context and tested on the marchantia CHH context set yielded a much higher accuracy of 62.8%. We observed that (i) the HyenaDNA-128 k models outperformed AMPS with annotations in only 15 of 90 cases (17%), (ii) the HyenaDNA-436 k models outperformed AMPS with annotations in 11 of 90 cases (12%), and (iii) the HyenaDNA-940 k models outperformed AMPS with annotations in 12 of 90 cases (13%). Compared to AMPS without annotations, we observed an improved performance in 68 of 90 cases (76%) for the HyenaDNA-128 k models, 75 of 90 cases (83%) for the HyenaDNA-436 k models, and 75 of 90 (83%) for the HyenaDNA-940 k models.

D. Runtime Analysis

To determine the runtime advantage of Hyena with respect to a transformer-based model, we carried out an additional set of experiments. First, we collected the runtime of HyenaDNA-436 k for the inference of the methylation status on the cucumber genome (CG context) over various input sequences, ranging from 1,600 bp to 12,800 bp. Then, we ran another inference over the same inputs using a fine-tuned DNABERT-2 architecture. Both models were fine-tuned with input samples of 3,200 bp. Fig. 5 shows that HyenaDNA scales significantly better than the DNABERT-2. When the input was 1,600 bp, HyenaDNA was 7.2× faster than DNABERT-2. At 3,200 bp, HyenaDNA was 9.0× faster. At 12,800 bp, HyenaDNA was 17.5× faster. To complete this analysis, Supplemental Fig. 6 shows the predictive accuracy of DNABERT-2 and HyenaDNA on this data set. Observe that the predictive accuracy in both models peaks with inputs of 3-4 thousand base pairs, which is expected since both models were fine-tuned on 3,200 bp long sequences.

E. Hyperparameter Analysis

Here we investigated the effects of some important hyper-parameters on the performance of HyenaDNA, namely (a) residual dropout, (b) training set size, and (c) window size.

To evaluate the effects of changes to the residual dropout, we chose the HyenaDNA-436 k model pre-trained on the cucumber genome, and fine-tuned using a residual dropout of 0.1, in addition to the 0.1 embedding dropout used in the previous experiments. Supplemental Fig. 1 shows that in all cases there is a small drop in predictive accuracy when a residual dropout is incorporated.

To test the effects of changes to the training set size, we used the HyenaDNA-436 k model pre-trained on the cucumber genome, and we fine-tuned using samples of 25%, 50%, 75%, and 100% of the original training set. As expected, Supplemental Fig. 4 shows a positive correlation between HyenaDNA's accuracy and the training data set size across all contexts on the cucumber genome.

To investigate the effect of using a longer input window, we fine-tuned HyenaDNA-436 k on all contexts and species using a 6,400 bp window size (double the size of the default window used in all previous experiments). Supplemental Fig. 5 shows that the 6,400 bp models

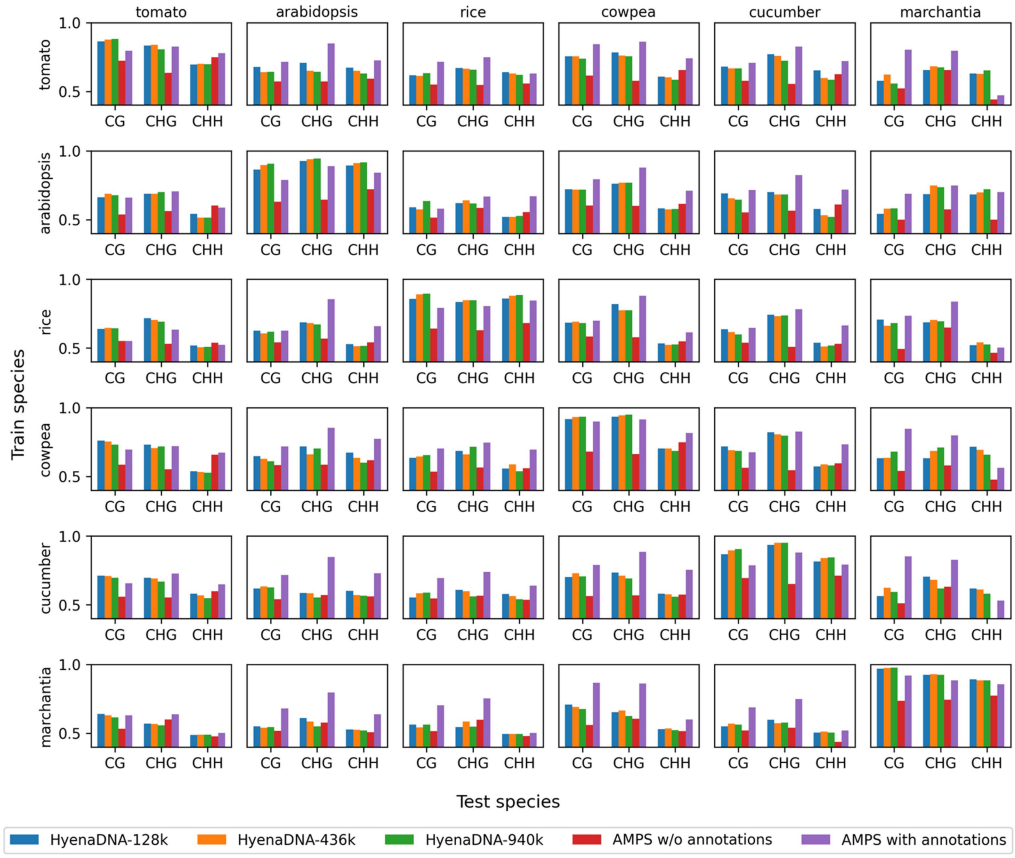

Fig. 4. Cross-species testing accuracy comparisons between HyenaDNA and AMPS. HyenaDNA-940 k outperforms AMPS without annotations in a majority of cases, but falls behind AMPS with annotations.

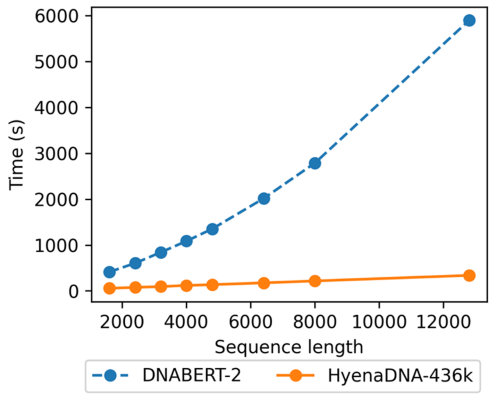

Fig. 5. Comparing DNABERT-2 and HyenaDNA-436 k's runtime for different input sequence lengths.

performed worse than the 3,200 bp HyenaDNA models in 17 out of 18 cases, with the only exception of marchantia CHH (which shows a minimal improvement).

V. Discussion

Our experimental results showed that the Hyena architecture was very accurate in predicting the methylation status in plants using exclusively the primary DNA sequence. Our Hyena-based models demonstrated significant accuracy improvements compared to the sequence-only AMPS classifiers for species/context-specific and cross-context problems. In some cases, we observed that larger Hyena-based models achieved a lower accuracy compared to smaller models, indicating possible issues with overfitting. An example of this behavior is the species/context-specific case for tomato in the CHG context, where the experimental results indicate that the optimal model size may be 128 k or even lower. The large gap in prediction accuracy between AMPS without annotations and AMPS with annotations for this particular context suggests that the methylation state of a cytosine strongly depends on its genomic localization (i.e., within a gene or a repeat), which our HyenaDNA models do not use. We have considered the possibility that smaller training set sizes in certain contexts (i.e., fewer than 500,000 samples) may have contributed to differences in the classifier's performance. However, we did not find a correlation between training set size and performance across our experiments. In cross-species situations, HyenaDNA performed on par with AMPS without annotations, but frequently underperformed compared to AMPS with annotations. This suggests that gene and repeat annotations could be key to accurately predicting methylation across different species. In cross-context situations, we found that HyenaDNA performed better than AMPS, even with annotations, in most cases. Our findings support the observation reported in [13] that cross-context predictions between CG and CHG overall yield good accuracies, while this is not the case with cross-context predictions involving CHH.

While our study demonstrates the effectiveness of HyenaDNA for DNA methylation prediction, we acknowledge our method's performance is limited in cross-species contexts. This suggests that in general, models trained on one plant species do not readily transfer to others, and new species will likely require retraining with their own methylation data. However, we hypothesize that species which are evolutionarily closely related and share a high-level of sequence similarity may

show methylation patterns similar enough to the extent that retraining would not be necessary. Additionally, several questions remain open for further exploration. A primary concern about deep learning models is their black-box nature, which makes it difficult to explain the reasons behind the predictions of these models [33], [34]. In this study, understanding the workings of the models can provide deeper insights into the regulatory mechanisms that control DNA methylation in the cell. Therefore, priority should be given to the interpretability analysis of the HyenaDNA models, focusing on understanding the features and sequence motifs that contribute most to their predictions. To interpret the weights of the CNN-based AMPS classifiers, the authors used Grad-CAM [35] to identify motifs within the input sequence that are most informative with respect to the classification of methylation status. To the best of our knowledge, there are no equivalent tools for HyenaDNA architectures. Future work should involve the analysis of the weights of the HyenaDNA models and study their layer activations. Otherwise, model-agnostic approaches to interpretability analysis, such as SHAP [36] and LIME [37], could be used for this purpose.

The authors of AMPS have shown that incorporating gene and repeat annotations significantly increased the prediction accuracy. It is possible that adding gene and repeat annotations could increase the overall performance of HyenaDNA as well. Possible improvements in performance could be achieved by carrying out additional experiments on the hyperparameters and/or the training datasets. For instance, it could be worth investigating the effect of other sequence window sizes on the performance of the HyenaDNA classifier. This could help determine whether local sequence motifs or long-range interactions are more relevant for specific methylation contexts. It may also be interesting to explore the effects of pretraining on all six plant species, to test the possibility of creating a single foundational model. Along these lines, the recent Evo foundation model, trained on millions of prokaryotic and phage genomes, demonstrated impressive performance in predictive and generative tasks [38], [39]. We believe that further studies on HyenaDNA may pave the way for similar results on plant species as well.

Acknowledgment

We thank Dr. S. Sereshki for consultation and guidance on the AMPS architecture, and for providing insight into the associated genomic datasets and experimental results.

Conflict of interest statement: The authors declare that they have no competing interests.

Software Availability

The code and its execution instructions are available from GitHub (https://github.com/hfeng95/HyenaDNA-methyl).

References

[1] B. Jin, Y. Li, and K. D. Robertson, "DNA methylation: Superior or subordinate in the epigenetic hierarchy?," *Genes Cancer*, vol. 2, no. 6, pp. 607–617, Jun. 2011.

[2] R. Sun and P. Zhu, "Advances in measuring DNA methylation," *Blood Sci.*, vol. 4, no. 1, pp. 8–15, Jan. 2022.

[3] J. A. Law and S. E. Jacobsen, "Establishing, maintaining and modifying DNA methylation patterns in plants and animals," *Nature Rev. Genet.*, vol. 11, no. 3, pp. 204–220, Mar. 2010.

[4] J. Du, L. M. Johnson, S. E. Jacobsen, and D. J. Patel, "DNA methylation pathways and their crosstalk with histone methylation," *Nature Rev. Mol. Cell Biol.*, vol. 16, no. 9, pp. 519–532, Sep. 2015.

[5] H. Zhang, Z. Lang, and J.-K. Zhu, "Dynamics and function of DNA methylation in plants," *Nature Rev. Mol. Cell Biol.*, vol. 19, no. 8, pp. 489–506, Aug. 2018, doi: 10.1038/s41580-018-0016-z.

[6] S. N. Bogan and S. V. Yi, "Potential role of DNA methylation as a driver of plastic responses to the environment across cells, organisms, and populations," *Genome Biol. Evol.*, vol. 16, no. 2, Feb. 2024, Art. no. evae022, doi: 10.1093/gbe/evae022.

[7] T. Yue et al., "Deep learning for genomics: From early neural nets to modern large language models," *Int. J. Mol. Sci.*, vol. 24, no. 21, 2023, Art. no. 15858. [Online]. Available: https://www.mdpi.com/1422-0067/24/21/15858

[8] W. S. Alharbi and M. Rashid, "A review of deep learning applications in human genomics using next-generation sequencing data," *Hum. Genomic.*, vol. 16, no. 1, Jul. 2022, Art. no. 26, doi: 10.1186/s40246-022-00396-x.

[9] H. Feng et al., "Benchmarking DNA foundation models for genomic sequence classification," Aug. 2024, doi: 10.1101/2024.08.16.608288.

[10] S. Zhang, R. Fan, Y. Liu, S. Chen, Q. Liu, and W. Zeng, "Applications of transformer-based language models in bioinformatics: A survey," *Bioinf. Adv.*, vol. 3, no. 1, Jan. 2023, Art. no. vbad001.

[11] H. Zeng and D. K. Gifford, "Predicting the impact of non-coding variants on DNA methylation," *Nucleic Acids Res.*, vol. 45, no. 11, Jun. 2017, Art. no. e99.

[12] C. Angermueller, H. J. Lee, W. Reik, and O. Stegle, "DeepCpG: Accurate prediction of single-cell DNA methylation states using deep learning," *Genome Biol.*, vol. 18, no. 1, Apr. 2017, Art. no. 67, doi: 10.1186/s13059-017-1189-z.

[13] S. Sereshki, N. Lee, M. Omirou, D. Fasoula, and S. Lonardi, "On the prediction of non-CG DNA methylation using machine learning," *NAR Genomic. Bioinf.*, vol. 5, no. 2, May 2023, Art. no. lqad045, doi: 10.1093/nargab/lqad045.

[14] H. Cui et al., "scGPT: Toward building a foundation model for single-cell multi-omics using generative AI," *Nature Methods*, vol. 21, no. 8, pp. 1470–1480, Aug. 2024.

[15] Y. Ji, Z. Zhou, H. Liu, and R. V. Davuluri, "DNABERT: Pre-trained bidirectional encoder representations from transformers model for DNA-language in Genome," *Bioinformatics*, vol. 37, no. 15, pp. 2112–2120, Aug. 2021.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423/

[17] J. Jin et al., "iDNA-ABF: Multi-scale deep biological language learning model for the interpretable prediction of DNA methylations," *Genome Biol.*, vol. 23, no. 1, Oct. 2022, Art. no. 219. [Online]. Available: https://doi.org/10.1186/s13059-022-02780-1

[18] S. Wang, Y. Liu, Y. Liu, Y. Zhang, and X. Zhu, "BERT-5mC: An interpretable model for predicting 5-methylcytosine sites of DNA based on BERT," *PeerJ*, vol. 11, Dec. 2023, Art. no. e16600.

[19] N. N. Soylu and E. Sefer, "BERT2Ome: Prediction of 2'-O-methylation modifications from RNA sequence by transformer architecture based on BERT," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 20, no. 3, pp. 2177–2189, May/Jun. 2023.

[20] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[21] F. D. Keles, P. M. Wijewardena, and C. Hegde, "On the computational complexity of self-attention," in *Proc. Int. Conf. Algorithmic Learn. Theory*, 2023, pp. 597–619.

[22] S. Kim, H. J. Park, X. Cui, and D. Zhi, "Collective effects of long-range DNA methylations predict gene expressions and estimate phenotypes in cancer," *Sci. Rep.*, vol. 10, no. 1, Mar. 2020, Art. no. 3920. [Online]. Available: https://doi.org/10.1038/s41598-020-60845-2

[23] M. Bartkuhn and R. Renkawitz, "Long range chromatin interactions involved in gene regulation," *Biochimica et Biophysica Acta Mol. Cell Res.*, vol. 1783, no. 11, pp. 2161–2166, 2008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167488908002607

[24] A. Gu et al., "Combining recurrent, convolutional, and continuous-time models with linear state-space layers," in *Proc. 35th Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 572–585.

[25] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," in *Proc. Int. Conf. Learn. Representations*, 2022. [Online]. Available: https://openreview.net/pdf?id=COZDy0WYGg

[26] D. Y. Fu, T. Dao, K. K. Saab, A. W. Thomas, A. Rudra, and C. Ré, "Hungry hungry hippos: Towards language modeling with state space models," in *Proc. Int. Conf. Learn. Representations*, 2023.

[27] E. Nguyen et al., "HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution," in *Proc. 37th Int. Conf. Neural Inf. Process. Syst.*, 2023, Art. no. 1872.

[28] M. Poli et al., "Hyena hierarchy: Towards larger convolutional language models," in *Proc. 40th Int. Conf. Mach. Learn.*, 2023, pp. 28043–28078.
[29] L. Tirot, P. E. Jullien, and M. Ingouff, "Evolution of CG methylation maintenance machinery in plants," *Epigenomes*, vol. 5, no. 3, Sep. 2021, Art. no. 19.
[30] J. Fang et al., "Mechanistic basis for maintenance of CHG DNA methylation in plants," *Nature Commun.*, vol. 13, no. 1, Jul. 2022, Art. no. 3877, doi: 10.1038/s41467-022-31627-3.
[31] Q. Tian, J. Zou, J. Tang, Y. Fang, Z. Yu, and S. Fan, "MR-CNN: A deep learning model for regression of genome-wide DNA methylation," *BMC Genomic.*, vol. 20, no. Suppl 2, Apr. 2019, Art. no. 192.
[32] Y. Wang et al., "A deep learning approach to automate whole-genome prediction of diverse epigenomic modifications in plants," *New Phytol*, vol. 232, no. 2, pp. 880–897, Aug. 2021.
[33] A. Madsen, S. Reddy, and S. Chandar, "Post-hoc interpretability for neural NLP: A survey," *ACM Comput. Surv.*, vol. 55, no. 8, Dec. 2022, Art. no. 155, doi: 10.1145/3546577.
[34] S. Chakraborty et al., "Interpretability of deep learning models: A survey of results," in *Proc. 2017 IEEE SmartWorld Ubiquitous Intell. Comput. Adv. Trusted Computed Scalable Comput. Commun. Cloud Big Data Comput. Internet People Smart City Innov.*, 2017, pp. 1–6.
[35] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Oct. 2019, doi: 10.1007/s11263-019-01228-7.
[36] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 4768–4777.
[37] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?": Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.
[38] E. Nguyen et al., "Sequence modeling and design from molecular to Genome scale with Evo," *Science*, vol. 386, no. 6723, 2024, Art. no. eado9336. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.ado9336
[39] M. Poli et al., "StripedHyena: Moving beyond transformers with hybrid signal processing models," Dec. 2023. [Online]. Available: https://github.com/togethercomputer/stripedhyena