Cross-Species Transferability of Genomic Language Models for DNA Methylation Prediction

Literature Review for Research Lab

Mohammad Rezaei
Student ID: 5547733
m.rezaeib-barzani@tu-braunschweig.de
Program: AIMS

Supervisor:	Corinna Thoben
Primary Advisor:	Prof. Dr. Tim Kacprowski

Technische Universität Braunschweig
Peter L. Reichertz Institute for Medical Informatics
Division Data Science in Biomedicine
Rebenring 56
38106 Braunschweig

Semester	WS25/26 SS26
Planned Start	01.01.2026
Planned End	31.06.2026

Braunschweig, December 10, 2025

Contents

1 Abstract

DNA methylation is a central epigenetic mechanism that shapes gene regulation and phenotypic variation in plants. While whole-genome bisulfite sequencing (WGBS) and nanopore-based assays provide base-resolution methylation maps, generating such data across many species and conditions remains costly. Recent "genomic language models", including transformer-based architectures pretrained on large-scale DNA sequence corpora, offer a promising route to predict methylation states directly from sequence and to transfer information across species.

This project investigates how three representative models. Nucleotide Transformer based on transformer architecture and pretrained on genomic sequences is considered our base benchmark. DNABERT-2 is a genome foundation model that uses an efficient Byte-Pair Encoding (BPE) tokenizer instead of traditional k-mers, resulting in improved training efficiency. UniMethylNet is a convolution neural network combined with Long Short-Term Memory and self-attention blocks. We first establish standardized, quality-controlled methylation labels and genomic splits, and evaluate all models in zero-shot and fine-tuned settings for both within-species and cross-species prediction. We then perform targeted experiments on architectural and tokenization choices (context window length, number of fine-tuned layers, k-mer configurations), and a fine-tuning strategy using Low-Rank Adaptation (LoRA) to probe which design decisions most strongly affect transferability.

2 Introduction and Motivation

DNA methylation is a fundamental epigenetic modification influencing gene regulation, development, and environmental responses in plants. In most plant genomes, cytosine methylation (5mC) occurs in three sequence contexts—CG, CHG, and CHH (H $\in \{A, C, T\}$)—and shapes chromatin state, transposable element (TE) silencing, and gene expression. Accurate base-resolution methylation maps therefore provide essential "ground-truth" information on regulatory architecture, but generating such maps at scale remains costly and technically challenging, especially across diverse species and environmental conditions.

Recent progress in high-throughput sequencing has produced large numbers of whole-genome bisulfite sequencing (WGBS) and nanopore methylation datasets. Nevertheless, many biological contexts still lack dense, high-quality coverage, and existing datasets are often biased towards a few model species such as *Arabidopsis thaliana*. Methods such as PlantDeepMeth and DeepPlant already leverage WGBS or nanopore signal data to predict plant methylation states from local sequence and coverage features, but they remain limited by the availability of labeled data and species-specific training. [1, 2, 3] This motivates leveraging AI models, especially those trained on vast amounts of unlabeled DNA sequence, to impute or predict methylation states where direct measurement is missing or sparse. Researchers commonly combine sequence context with genomic features such as CpG island proximity, transcription factor binding sites, histone modification marks, and chromatin accessibility to improve prediction accuracy. [1, 4, 5]

Recent advances in large "genomic language models"—neural networks pretrained on vast amounts of DNA sequence—offer a new paradigm for methylation prediction. Models such as DNABERT-2 and the Nucleotide Transformer series use transformer architectures to learn rich, context-dependent sequence representations by self-supervised training on billions of nucleotides. [6, 7, 8, 9] These foundation models can exploit abundant unlabeled sequence data to capture genomic patterns across species and can be adapted to downstream tasks (e.g., methylation prediction) via fine-tuning or shallow classifier heads trained on relatively small labeled datasets.

In plant epigenomics, the cross-species transferability of such models is of particular interest: can a model trained on one plant species accurately predict methylation patterns in another species, possibly with only limited fine-tuning? Existing methods like PlantDeepMeth, UniMethylNet, MuLan-Methyl, and DeepPlant suggest that cross-species prediction is feasible but strongly dependent on model architecture, training data diversity, and methylation context. [1, 5, 10, 2] This project focuses on 5mC methylation in plants and investigates how genomic language models, particularly transformer-based architectures, can (i) predict methylation states based on DNA sequence alone, (ii) generalize across species, and (iii) support the identification of phenotype-linked genomic regions.

3 State of the Art

3.1 Genomic Foundation Models and Language Modeling

The concept of pretraining large neural models on DNA sequences borrows from NLP's foundation models. [8] Models such as DNABERT-2 [6] and the Nucleotide Transformer [7] illustrate this approach. DNABERT-2 replaces traditional fixed k-mer tokens with Byte-Pair Encoding (BPE) to build an efficient "DNA vocabulary," enabling compact multi-species pretraining. [6] Nucleotide Transformer models are a suite of transformer models pretrained on human and diverse animal genomes. These models integrate data from 3,202 human genomes and 850 non-human genomes, producing context-sensitive embeddings that transfer well to various genomics tasks. These foundation models demonstrate that broad pretraining can capture key genomic signals without supervision and can be fine-tuned at low cost on various downstream genomics tasks. [7] Genomic LMs excel with self-supervised pretraining, learning nucleotide "language" by masked prediction on millions of sequences. These foundation models offer a general-purpose approach to molecular prediction, leveraging vast unlabeled sequence data to overcome annotation scarcity. [8, 7]

3.2 Transfer Learning and Fine-Tuning for Methylation

Transfer learning is central to adapting foundation models to specific methylation tasks. For instance, CpGPT [11] demonstrates a large transformer pretrained on 1,500 human methylation datasets. Although CpGPT was trained on human data, its success illustrates how pretraining on large methylation data can produce generalizable features. In plants, transfer learning often means adapting animal-trained architectures to plant methylomes. Guo et al. [1] created PlantDeepMeth by modifying the DeepCpG architecture and retraining it on plant methylation data. PlantDeepMeth achieved strong performance on Brassica rapa and A. thaliana, maintaining high accuracy when the Arabidopsis-trained model was applied to B. rapa. This cross-species transfer success underscores that properly adapted deep models can generalize across plant genomes. Another example of transfer learning is Dodlapati et al., who addressed sparse single-cell methylomes by transferring knowledge between tissues. In their study, they used transfer learning with Kullback–Leibler (KL) divergence loss to train predictive models for completing methylome profiles with very low coverage. Specifically, KL divergence serves as a loss function that measures the distance between the predicted probability distribution of methylation states and the true biological distribution, allowing the model to optimize for probability accuracy rather than just magnitude error. The model architecture consists of four subunits: a neighboring DNA feature extractor, a neighboring methylation feature extractor, a fusing network that combines the two network outputs, and a classifier head that uses information from all three subunits to make a prediction for methylation patterns in target sequences. The model variants were trained on data-rich cell types, and then transfer learning was used to predict methylation profiles of less-known tissues for

the same species. [12]

In summary, transfer learning – whether from humans to plants, between tissues, or across species – has become a key strategy. Models pretrained on large genomic corpora can capture conserved methylation patterns, and fine-tuning on a new species' data often yields enhanced cross-species generalization.

3.3 Cross-Species Prediction in Plants

Multiple studies have directly tested cross-species methylation predictions. Sereshki et al. trained sequence-based classifiers on six plant species to predict cytosine methylation in CG, CHG, and CHH sequences. They tested the cross-species prediction problem, which proved to be challenging, although they demonstrated that providing gene and repeat annotations allows existing classifiers to significantly improve their prediction accuracy. [4] Contrastingly, PlantDeepMeth [1] reports strong cross-species generalization between B. rapa and A. thaliana. Similarly, iDNA-ABF outperformed existing predictors and showed robust performance across species. iDNA-ABF is a multi-scale deep biological language learning model that enables the interpretable prediction of DNA methylation based on genomic sequences only. It highlights the power of deep language learning in capturing both sequential and functional semantic information from background genomes. Integrating the interpretable analysis mechanism enables the model to explain what it learns, helping build the mapping from the discovery of important sequential determinants to the in-depth analysis of their biological functions. [13] This is particularly useful in the second part of our study, where we are interested in finding important regulatory regions in the genome. Another important novelty of their method was the utilization of adversarial training, which increased model robustness and accuracy.

UniMethylNet achieved 88 precent accuracy on 20 datasets, exhibiting superior cross-species and cross-type generalization. It recognizes different methylation types (4mC, 5hmC, and 6mA) across 12 species. [5] Chen et al. developed DeepPlant, a BiLSTM and Transformer model to detect 5mC from Oxford Nanopore signals. DeepPlant shows high whole-genome methylation correlations across species. [2] MuLan-Methyl uses an ensemble of transformer LMs pretrained on DNA and then fine-tuned on predicting the DNA methylation status of each type. MuLan-Methyl aims at identifying three types of DNA methylation sites. The self-attention mechanism of transformers produces importance scores, which can be used to identify motifs. In summary, cross-species transferability varies by approach. Traditional supervised predictors tend to be species-specific. In contrast, models pretrained on diverse data (PlantDeepMeth, iDNA-ABF, UniMethylNet, DeepPlant, MuLan) show substantially improved generalization across plant species.

3.4 Model Architectures and Innovations

A variety of model architectures have been applied to DNA methylation prediction in plants and other organisms. PlantDeepMeth builds on the earlier DeepCpG-style combination of convolutional and recurrent layers, using CNNs to extract local sequence features and bidirectional LSTMs to aggregate them over longer ranges. [1] DeepPlant instead combines bidirectional LSTM layers with Transformer blocks to integrate nanopore signal and sequence information for cross-species 5mC detection in plants. [2] MuLan-Methyl further illustrates the benefit of ensembling by aggregating predictions from five independently pretrained transformer models, leading to improved robustness and accuracy. [10]

Smaller Nucleotide Transformer models can sometimes match or even surpass larger variants when properly scaled and trained on diverse genomic corpora; in many settings, increasing sequence diversity and task alignment is more beneficial than merely increasing model size. [7, 8] Suzuki et al. showed

that model performance often depends more on input tokenization than on model size. In particular, overlap-based k-mer tokenization improves performance by preserving local sequence context and capturing motifs at multiple offsets. [9] DNABERT-2 replaces fixed k-mers with BPE-derived tokens, arguing that variable-length sub-sequences capture frequent genomic motifs more compactly and allow models to focus capacity on biologically meaningful units. [6]

In general, transformer architectures are particularly attractive for methylation prediction because multi-head self-attention can capture long-range dependencies, while layer-wise attention maps and attribution methods facilitate interpretability. Multi-scale or hierarchical models such as iDNA-ABF explicitly operate at different sequence scales and provide position-wise importance scores, offering insights into regulatory elements and motif combinations. [13] Overall, innovations in architecture (depth, attention span, ensembling) and tokenization (fixed vs. variable-length tokens, overlapping vs. non-overlapping k-mers) are central design choices for genomic language models and are expected to have a strong impact on cross-species methylation prediction.

3.5 Impact of Fine-Tuning and Model Scale on Transferability

Fine-tuning pretrained models greatly enhances cross-species performance. MuLan-Methyl applies a pretrain-and-fine-tune paradigm. Nucleotide Transformer models also outperformed baselines when fine-tuned. Dalla-Torre et al. showed that increasing parameters yields diminishing returns, [7] while Suzuki et al. found smaller transformers with optimal tokenization rivaled larger models. [9] Overall, transferability benefits more from diverse training data and smart tokenization than model size alone. Recent studies also highlight that fine tunning strategies, for example Low-Rank Adaptation (LoRA), fine-tunes only a subset of model parameters, significantly reducing computation while maintaining high accuracy.[6]

3.6 Methylation Types and Ground-Truth Data

Existing studies differ in the methylation types and experimental assays used to generate ground-truth labels. Many plant studies focus on cytosine methylation (5mC) in CG, CHG, and CHH contexts derived from WGBS data. PlantDeepMeth, for example, predicts binary methylated/unmethylated states in all three contexts using WGBS data from *Arabidopsis thaliana* and *Brassica rapa*. [1] DeepPlant uses nanopore signal data to infer 5mC states across plant genomes, illustrating how long-read sequencing can provide direct methylation measurements at single-molecule resolution. [2, 3] UniMethylNet and MuLan-Methyl extend beyond 5mC to other methylation variants such as 4mC, 5hmC, and 6mA across multiple species, highlighting the potential of unified models to handle diverse epigenetic marks. [5, 10]

In the present work, we will focus on 5mC cytosine methylation in CG, CHG, and CHH contexts, using base-resolution methylation calls from plant WGBS or nanopore-based studies as ground truth. Sites with sufficient coverage (e.g., ≥ 10 reads) will be labeled as methylated or unmethylated based on established thresholds (e.g., methylation ratio ≥ 0.5 vs. < 0.5), and low-coverage positions will be excluded or treated separately. This explicit definition of methylation type and labeling rules is essential for consistent training and evaluation across species and for meaningful comparison with existing methods.

4 Objectives

(Rewrite according to comments)

5 Planned Approach

5.1 Materials & Methods

5.1.1 Data assembly and ground truth

We focus on cytosine 5mC methylation in CG, CHG, and CHH contexts in plants. Ground-truth labels will be derived from published bisulfite or nanopore-based methylation datasets for *Arabidopsis thaliana* and *Brassica rapa*, together with their corresponding reference genomes and annotations. [Add 1000 Genome reference here] [1, 14, 15, 3, 16]

Methylation type and labels. For each cytosine, methylation ratios will be computed from aligned reads.Sites will be labeled as methylated or unmethylated based on methylation-ratio cutoffs (e.g., ≥ 0.5 vs. < 0.5), applied consistently within each species. Context (CG, CHG, CHH) will either be modeled explicitly (e.g., as an additional input feature) or separated into context-specific tasks.

Standard splits and evaluation metrics. To avoid information leakage, train/validation/test splits will be defined at the level of chromosomes or large, non-overlapping genomic blocks per species. Entire chromosomes or blocks will be reserved as unseen test data for each species. Cross-species tests will train models on one species (e.g., *A. thaliana*) and evaluate them on held-out chromosomes of the other species (e.g., *B. rapa*). Primary evaluation metrics will include AUROC, AUPRC, MCC, and F1-score, with additional stratified metrics by genomic annotation (genes vs. transposable elements) and methylation context (CG/CHG/CHH). Region-based summaries (e.g., per gene or fixed 10 kb windows) will support later LMM analyses.

5.1.2 Comparing Models Performance

We will evaluate performance of the three selected models:

Nucleotide Transformer (NT). We will use a pretrained Nucleotide Transformer model as a frozen feature extractor.[7] For each cytosine, a fixed-length window around the site will be tokenized according to the NT scheme, and contextual embeddings will be extracted. A lightweight classifier head (e.g., linear or MLP) will then be trained on top of the frozen embeddings using training data from *A. thaliana* and/or *B. rapa*.

DNABERT2. A pretrained DNABERT model will be used as another frozen feature extractor.[6] The same procedure as for NT will be followed: frozen embeddings, and a shallow classifier trained on methylation labels.

UniMethylNet (baseline model). UniMethylNet[5] will be trained directly on the methylation prediction task using the same training data and splits, serving as a non-transformer baseline against which we compare NT and DNABERT2 representations.

Within-species and cross-species evaluation. Within-species baselines will train classifier heads' on training chromosomes of *A. thaliana* or *B. rapa* and evaluate on held-out chromosomes of the same species. Cross-species baselines will train on one species (e.g., *A. thaliana*) and evaluate on test chromosomes of the other species without changing pretrained NT and DNABERT weights.

5.1.3 Comparing fine-tuning strategies

We then investigate performance fine-tuning strategies with NT and DNABERT2 on methylation labels.

Fine-tuning strategy. Starting from pretrained NT and DNABERT2 checkpoints, we will attach a binary classification head for methylated vs. unmethylated cytosines. Two main regimes will be considered:

- *Head + top-layer fine-tuning:* lower transformer layers are frozen, while the top few layers (e.g., last 2–4 blocks) and the classification head are fine-tuned.
- *Full-model fine-tuning (resource-permitting):* the entire model is fine-tuned with a small learning rate and appropriate regularization.

Class imbalance will be addressed via class-weighted or focal loss if needed, and early stopping will be applied based on validation AUROC/AUPRC.

Training data: one species or close family. *A. thaliana* will be the primary training species. *B. rapa* may optionally be included in joint training if time and compute allow, but by default *A. thaliana* will be treated as the main "source" species and *B. rapa* as the "target" species for transfer testing.

Comparison with UniMethylNet. UniMethylNet will be trained per species (without comparable large-scale pretraining). Fine-tuned NT and DNABERT2 will be compared against UniMethylNet on within-species and cross-species test sets to quantify the advantages of transformer-based pretrained representations.

5.1.4 Evaluation pipeline

All experiments will use a common evaluation pipeline.

Training and validation. Models will be trained on predefined training splits with early stopping based on validation performance (e.g., AUROC). Metrics will be tracked separately for each species, context, and architecture configuration.

Test evaluation. On held-out test chromosomes, we will compute AUROC, AUPRC, MCC, F1, and accuracy for each model and setting, stratifying results by genomic annotation (e.g., genes vs. transposable elements) and methylation context (CG/CHG/CHH).

Region-level summarization. Per-site predictions will be aggregated into per-region scores (e.g., MCC per gene or per 10 kb window).

5.2 Timeline

The planned duration is four months, from 1 December 2025 to the end of April 2026.

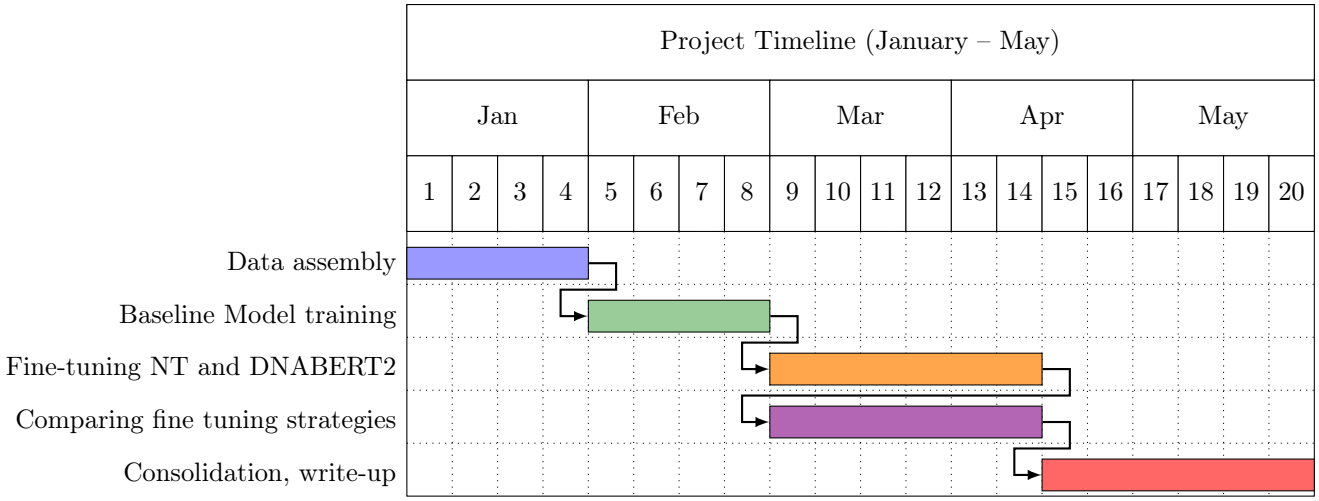

Figure 1: Weekly Gantt chart of the planned work from December 2025 to April 2026.

- **January: Data assembly**
  - Download all required sequencing and reference data for *A. thaliana* and *B. rapa*.
  - IDefine binary methylation labels and sequence windows.
  - Set up train/validation/test splits and initial evaluation scripts.
- **February: Baseline Model training**
  - Train UniMethylNet as a baseline.
  - Run within-species and cross-species experiments for both species.
  - Compute standard metrics and identify any technical issues early.
- **March - April: Fine tuning NT and DNABERT2**
  - Implement fine-tuning for NT and DNABERT2 (head-only vs. head + top layers).
  - Run main fine-tuning experiments on *A. thaliana* (with optional *B. rapa* augmentation if feasible).
- **April - May: Consolidation, optional experiments, and analysis**
  - Finalize analyses across models, species, and architectures.

- Integrate results, generate figures and tables, and write up Methods, Results, and Discussion sections.

6 Future Work

Beyond the experiments planned in this project, there are several promising directions for extending genomic language models for methylation prediction. In particular, retrieval-enhanced transformer architectures and parameter-efficient fine-tuning methods such as LoRA suggest ways to decouple global genomic context from model size and to adapt large DNA models under realistic compute constraints.

6.1 Retrieval-Enhanced Transformers for Genomic Methylation

DeepMind's Retrieval-Enhanced Transformer (RETRO) augments an autoregressive transformer with an external retrieval database of text chunks.[17] Instead of encoding all information into model parameters or a very long attention window, RETRO retrieves k nearest-neighbour chunks for each input segment from a large database and integrates them via an encoder and cross-attention into the decoder. This design decouples *capacity* (stored in the retrieval corpus) from *parametric size* (stored in the transformer), enabling smaller models to match or surpass much larger baselines when paired with a sufficiently rich retrieval index.[17]

A natural extension for plant methylation is a RETRO-style architecture where the external database consists of genomic sequence chunks (possibly across species) annotated with relevant features (e.g., methylation profiles, gene annotations, regulatory elements). For a focal cytosine, the model could:

1. encode a local sequence window around the site using a DNA language model,
2. retrieve genomically similar or functionally relevant windows from the database (possibly constrained to orthologous regions or matching regulatory annotations),
3. encode the retrieved windows with a separate encoder, and
4. integrate them through cross-attention when predicting the methylation state at the focal site.

This design offers several potential advantages:

- **Global context:** the model can condition on long-range and cross-chromosomal information without requiring extremely long self-attention windows.
- **Cross-species transfer:** retrieval can draw on homologous or functionally similar regions from other species, potentially improving predictions in sparsely annotated genomes.
- **Interpretability:** retrieved regions provide explicit, inspectable evidence the model uses for its predictions (e.g., specific transposable element families or conserved regulatory motifs).

At the same time, a RETRO-like system introduces substantial engineering complexity. Building and maintaining a dense retrieval index over millions of genomic windows requires:

- precomputing and storing embeddings for large parts of one or more genomes,

- implementing an efficient nearest-neighbour search system (e.g., FAISS) at training and inference time,
- carefully defining what constitutes "relevance" (sequence similarity, shared annotation, co-methylation patterns, or task-specific learned similarity).

For a single or a few plant genomes, this is conceptually feasible on modern hardware but likely exceeds the time and compute budget of the present 4-month project. It is therefore best framed as a medium- to long-term extension.

6.2 Parameter-Efficient Fine-Tuning with LoRA

Low-Rank Adaptation (LoRA) is a parameter-efficient fine-tuning method that freezes the pretrained transformer weights and injects small trainable low-rank matrices into existing weight matrices (typically the attention and feed-forward projections).[18] Instead of updating a full $d \times d$ matrix, LoRA learns a rank-r decomposition $BA^\top$ (with $r \ll d$), which reduces the number of trainable parameters by several orders of magnitude while keeping inference-time latency almost unchanged.[18]

For large DNA language models such as DNABERT and Nucleotide Transformer, LoRA is attractive because:

- it enables fine-tuning models with hundreds of millions (or more) parameters on commodity GPUs,
- multiple task-specific adapters (e.g., different species or phenotypes) can be stored and swapped cheaply, while sharing the same frozen backbone,
- it reduces the risk of catastrophic forgetting by constraining updates to low-rank subspaces.

In the methylation setting, LoRA could be used to learn plant- and task-specific adapters on top of general-purpose DNABERT or Nucleotide Transformer checkpoints. This would allow exploring richer models than full fine-tuning would permit under limited compute, while still adapting to methylation-specific signals (e.g., context-dependent patterns in CG/CHG/CHH).

7 AI Use Disclaimer

According to category 2 (include the prompt and other deatils)

8 References

References

[1] Z. Guo *et al.*, "Plantdeepmeth: A deep learning model for predicting DNA methylation states in plants," *Plants*, vol. 14, no. 11, p. 1724, Jun. 2025.

[2] H.-X. Chen *et al.*, "Accurate cross-species 5mc detection for oxford nanopore sequencing in plants with DeepPlant," *Nat. Commun.*, vol. 16, no. 1, p. 3227, Apr. 2025.

[3] P. Ni, N. Huang, F. Nie, J. Zhang, Z. Zhang, B. Wu, L. Bai, W. Liu, C. L. Xiao, F. Luo *et al.*, "Genome-wide detection of cytosine methylations in plants from nanopore data using deep learning," *Nature Communications*, vol. 12, p. 5976, 2021.

[4] S. Sereshki, N. Lee, M. Omirou, D. Fasoula, and S. Lonardi, "On the prediction of non-CG DNA methylation using machine learning," *NAR Genomics Bioinforma.*, vol. 5, no. 2, p. lqad045, Mar. 2023.

[5] M. Zhang *et al.*, "UniMethylNet: A universal DNA methylation site prediction network integrating a neural network and an attention mechanism," *J. Chem. Inf. Model.*, p. acs.jcim.5c02000, Oct. 2025.

[6] Z. Zhou, Y. Ji, W. Li, P. Dutta, R. Davuluri, and H. Liu, "DNABERT-2: Efficient foundation model and benchmark for multi-species genome," *ICLR 2024*, Mar. 2024, arXiv:2306.15006.

[7] H. Dalla-Torre *et al.*, "Nucleotide transformer: building and evaluating robust foundation models for human genomics," *Nat. Methods*, vol. 22, no. 2, pp. 287–297, Feb. 2025.

[8] F. Guo *et al.*, "Foundation models in bioinformatics," *Natl. Sci. Rev.*, vol. 12, no. 4, p. nwaf028, Mar. 2025.

[9] S. Suzuki, K. Horie, T. Amagasa, and N. Fukuda, "Genomic language models with k-mer tokenization strategies for plant genome annotation and regulatory element strength prediction," *Plant Mol. Biol.*, vol. 115, no. 4, p. 100, Aug. 2025.

[10] W. Zeng, A. Gautam, and D. H. Huson, "Mulan-methyl—multiple transformer-based language models for accurate DNA methylation prediction," *GigaScience*, vol. 12, p. giad054, Dec. 2022.

[11] L. P. De Lima Camillo *et al.*, "CpGPT: a foundation model for DNA methylation," *Systems Biology*, Oct. 2024.

[12] S. Dodlapati, Z. Jiang, and J. Sun, "Completing single-cell DNA methylome profiles via transfer learning together with KL-divergence," *Front. Genet.*, vol. 13, p. 910439, Jul. 2022.

[13] 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, p. 219, Oct. 2022.

[14] K. Zhang, L. Zhang, Y. Cui, Y. Yang, J. Wu, J. Liang, X. Li, X. Zhang, Y. Zhang, Z. Guo *et al.*, "The lack of negative association between TE load and subgenome dominance in synthesized *Brassica* allotetraploids," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 120, p. e2305208120, 2023.

[15] H. Chen, T. Wang, X. He, X. Cai, R. Lin, J. Liang, J. Wu, G. King, and X. Wang, "BRAD v3.0: An upgraded brassicaceae database," *Nucleic Acids Research*, vol. 50, pp. D1432–D1441, 2022.

[16] F. Cunningham, J. E. Allen, J. Allen, J. Alvarez-Jarreta, M. R. Amode, I. M. Armean, O. Austine-Orimoloye, A. G. Azov, I. Barnes, R. Bennett *et al.*, "Ensembl 2022," *Nucleic Acids Research*, vol. 50, pp. D988–D995, 2022.

[17] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. van den Driessche, J. Lespiau, B. Damoc, A. Clark, D. de Las Casas, A. Guy, J. Menick, R. Ring, T. Hennigan, S. Huang, L. Maggiore, C. Jones, A. Cassirer, A. Brock, M. Paganini, G. Irving, O. Vinyals, S. Osindero, K. Simonyan, J. W. Rae, E. Elsen, and L. Sifre, "Improving language models by retrieving from trillions of tokens," *arXiv preprint arXiv:2112.04426*, 2021.

[18] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.