Harnessing DNA Foundation Models for Cross-Species Transcription Factor Binding Site Prediction in Plant Genomes

Maryam Haghani[1], Krishna Vamsi Dhulipalla[1], Song Li[1,2,3*]

[1]Department of Computer Science, Virginia Tech, Blacksburg, VA, USA

[2]Genetics, Bioinformatics, and Computational Biology, Virginia Tech, Blacksburg, VA, USA

[3]School of Plant and Environmental Sciences, Virginia Tech, Blacksburg, VA, USA

[*]Email: songli@vt.edu

Abstract

Accurate prediction of transcription factor binding sites (TFBSs) is crucial for understanding gene regulation. While experimental methods like ChIP-seq and DAP-seq are informative, they are labor-intensive and species-specific. Recent advancements in large-scale pretrained DNA foundation models have shown promise in overcoming these limitations. This study evaluates the performance of three such models—DNABERT-2, AgroNT, and HyenaDNA—in predicting TFBSs in plants. Using *Arabidopsis thaliana* and *Sisymbrium irio* DAP-seq data, we benchmark their accuracy against specialized methods like DeepBind and BERT-TFBS. Our results demonstrate that foundation models, particularly HyenaDNA, offer superior predictive accuracy and computational efficiency, highlighting their potential for scalable, genome-wide TFBS prediction in plants.

1 Introduction

Transcription factors (TFs) are key regulatory proteins that control gene expression by binding to specific DNA sequences called transcription factor binding sites (TFBSs). Accurate identification of these binding sites is crucial for deciphering gene regulatory networks and understanding biological processes. In plants, experimental methods such as chromatin immunoprecipitation sequencing (ChIP-seq) [1] and DNA affinity purification sequencing (DAP-seq) [2] are commonly used to map TFBSs. While powerful, these assays are labor-intensive, costly, often limited to a few species, and lack the scalability needed for comprehensive, genome-wide analyses

The advent of machine learning and deep learning approaches has revolutionized TFBS prediction in animal and human systems by learning sequence features directly from high-throughput binding data [3]. However, their application in plant genomics has lagged behind, despite the availability of extensive DAP-seq and ChIP-seq data. A true paradigm shift has emerged with the rise of foundation models—large-scale pretrained models originally developed in natural language processing and now adapted to DNA sequences. These foundation models, including DNABERT-2, a transformer-based model pretrained on genomes from 135 species, demonstrating superior performance in various DNA sequence classification tasks [4]; the Agronomic Nucleotide Transformer (AgroNT) [5], a RoBERTa-style encoder pretrained on genomes from 48 plant species with state-of-the-art performance in regulatory feature prediction across plant genomes; and HyenaDNA [6], a decoder-only model utilizing Hyena operators for long-range genomic sequence modeling at single nucleotide resolution, offering efficient training and inference, have revolutionized genomics by capturing rich, complex representations of DNA sequences that generalize across species and tasks.

However, despite extensive datasets in plants, no prior study has systematically benchmarked multiple large pretrained DNA foundation models on predicting plant transcription factor binding site data from DAP-seq experiments, or evaluated their ability to generalize across species.

In this study, we leverage the power of these pretrained DNA foundation models by fine-tuning them on DAP-seq data from *Arabidopsis thaliana* (*A. thaliana*) to capture complex sequence patterns. We further evaluate model generalization through cross-species testing on the closely related *Sisymbrium irio* (*S. irio*). We selected these two species because they both belong to the Brassicaceae family, possess relatively small genomes, and exhibit highly similar—but not identical—transcription factor binding site (TFBS) landscapes for orthologous genes. Specifically, we focus on the ABF family of transcription factors, key mediators of abscisic acid (ABA)-regulated gene expression in plants [7]. ABA is a central plant hormone involved in mediating various stress responses. Understanding how ABFs bind across the genome can help identify more effective targets for enhancing plant tolerance to abiotic stress conditions, with the potential to improve crop yields during periods of environmental adversity such as drought. Due to the biological significance of ABF genes, they were the first among thousands of plant transcription factors for which cross-species binding sites were identified using DAP-seq experiments [7]. This provides a rich, biologically validated dataset for training and evaluating the predictive power of the our foundation models. By appending a lightweight classification head and training end-to-end, our fine-tuned models learn nuanced sequence dependencies far beyond the reach of traditional TFBS predictors. We use a unified training and evaluation framework based solely on DNA sequences and benchmark their performance and speed against other specialized TFBS prediction models, including (1) DeepBind [8], a convolutional neural network (CNN)-based model for DNA/RNA binding site prediction, and (2) BERT-TFBS [9], an integration of DNABERT-2 with CNN modules and a convolutional block attention module, designed to capture both local and global sequence features for TFBS prediction. Our results demonstrate that fine-tuned foundation models substantially outperform specialized predictors (DeepBind and BERT-TFBS), and that HyenaDNA in particular achieves comparable predictive accuracy while reducing training time by over an order of magnitude. Finally, we conclude by discussing the limitations of our current methodology and outlining promising directions for future work.

2 Methods

2.1 Benchmark Datasets

We compiled DAP-seq binding sites for ABA-Responsive Element Binding Factors (AREB/ABF) from two public resources: `Malley2016` (Plant Cistrome Database) [2], consisting of binding regions for multiple TFs in *A. thaliana*, including AREB/ABF2, and `Sun2022` [7], comprising of AREB/ABFs1-4 binding sites across four Brassicaceae species, notably *Arabidopsis thaliana* (*A. thaliana*) and *Sisymbrium irio* (*S. irio*). We focused on *A. thaliana*, a model organism in plant biology and human disease research [10]. Using these data, we defined three evaluation protocols:

1. **Cross-chromosome** (`Sun2022`): Leave-one-chromosome-out on *A. thaliana* AREB/ABFs across chromosomes 1–5 (14,374 samples in total). For each experiment, one chromosome serves as the test set and the remaining four as train/validation (see Supplementary Table S1). Because `Sun2022` sequences varied in length, we standardized them to maximum length of 264bp by padding shorter sequences.

2. **Cross-dataset** (`Malley2016 - Sun2022`): Models were trained on AREB/ABF2 binding sites from `Malley2016` (*A. thaliana*) and evaluated on the corresponding AREB/ABF2 sites

in `Sun2022`. Since `Malley2016` includes only AREB/ABF2, we extracted the same TF's sites from `Sun2022`, which—being released later—provides an independent test set. After removing duplicate sequences across datasets, 14,569 unique samples remained (see Supplementary Table S2). All sequences were then standardized to 201bp (the `Malley2016` length) by trimming longer and padding shorter `Sun2022` sequences.

3. **Cross-species** (`Sun2022`): Train on *A. thaliana* and test on *Si* (and vice versa) for AREB/ABF 1–4. *A. thaliana* contributes 14,374 binding site samples; *Si* contributes 10,558; total = 24,932 across both species. Sequences were standardized as above (see Supplementary Table S3). As in the cross-chromosome protocol, `Sun2022` sequences varied in length, so we standardized them to the maximum observed length of 264bp by padding any shorter sequences.

For each evaluation protocol, we generated an equal number of negative samples for both the training and test sets to maintain balanced class proportions. Negative samples were created by randomly permuting the nucleotides of each positive sequence, preserving sequence length and base composition while disrupting biologically meaningful motifs, a standard approach in TFBS prediction [9]. To enable early stopping and reduce overfitting during training, we performed stratified 5-fold cross-validation on the training set, ensuring balanced class proportions in each fold. In each round of cross-validation, one fold was reserved for validation, while the remaining four folds were used for model training.

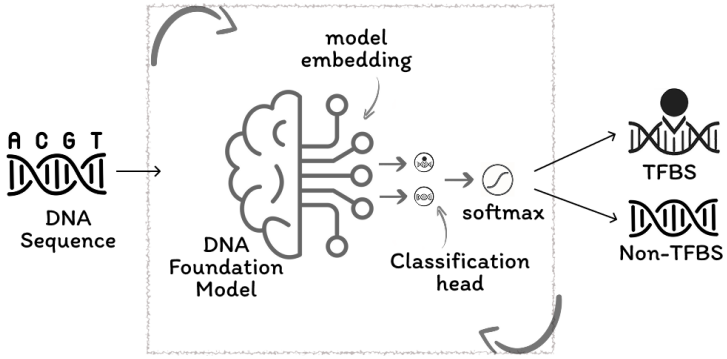

Figure 1: Overview of the fine-tuning framework for transcription factor binding sites (TFBS) prediction using a DNA foundation model: raw DNA sequences are input into a pre-trained DNA foundation model, which encodes them into high-dimensional embeddings. These embeddings are passed through a classification head that predicts whether the sequence corresponds to a TFBS or a non-TFBS. The loop arrows indicate that both the foundation model and the classification head are jointly updating the entire model end-to-end for the TFBS prediction task.

2.2 Model Architecture

To predict TFBS, we attach a two-neuron classification head directly atop the final token representations of each pre-trained DNA foundational model and fine-tune all model parameters—including the new head—on our labeled TFBS examples (Figure 1). These foundation models were initially trained via self-supervised learning on large-scale genomic corpora. We applied this approach to three different models:

Agronomic Nucleotide Transformer: Building upon Nucleotide Transformer [11], Agronomic Nucleotide Transformer (AgroNT) is a 1B-parameter RoBERTa-style [12] encoder pre-trained on genomes from 48 plant species. It uses a 6-mer tokenizer and 40 transformer layers that generate 1,500-dimensional embeddings per token. The model was evaluated on a range of prediction tasks,

including regulatory feature detection, RNA processing, and gene expression, and demonstrated state-of-the-art performance [5]

DNABERT-2: A modern foundation model, with 117 million parameters, designed for modeling genomic sequences in multiple species. It uses Byte Pair Encoding (BPE) [13] tokenization over multi-species genomic sequences (32.5B bases) of 135 species [4] and a 12 layer BERT encoder producing 768 dimensional embeddings [4]. Following sequence encoding, we apply mean pooling over token embeddings and feed the result into our two-class output head.

HyenaDNA: HyenaDNA is a decoder-only model that uses Hyena operators [14]—implicit long-range convolutional filters—to process up to 1M nucleotide tokens at single-base resolution. Its architecture forgoes k-mer tokenization in favor of single-nucleotide tokens and is trained via next-token prediction [6]. Its ability to model long-range interactions while maintaining single-nucleotide precision enables it to generalize effectively across tasks and species, establishing HyenaDNA as a robust and efficient foundation model for genomics [6].

2.3 Baseline Models

We have used recent architectures, specifically designed for binding-site prediction task to benchmark the performance and efficiency of our fine-tuning approach against them. We used two different models:

DeepBind: DeepBind employs a four-stage convolutional neural network to predict DNA- and RNA-binding specificity without manual feature engineering. First, raw nucleotide sequences are transformed into one-hot encoded matrices, which convolutional layers then scan using learned motif detectors to extract informative patterns. Second, these feature maps are passed through a rectification layer—applying an absolute value operation—to emphasize signal, then through a pooling layer that both enhances feature robustness and reduces overfitting. Finally, the condensed features feed into a fully connected layer that calculates a binding probability score. For our classification, we apply a softmax to this score to yield TFBS versus non-TFBS probabilities

BERT-TFBS: BERT-TFBS [9] integrates three modules: (1) a pretrained DNABERT-2 [4] encoder to capture long-range dependencies, (2) a CNN stack to extract high-order local features from the sequence, and (3) a Convolutional Block Attention Module (CBAM) [15] to enhances local features by the spatial and channel attention mechanisms on the convolutional outputs. The output module integrates all the sequence features from the three modules and passes them through a multi-layer perceptron output head for binary TFBS classification.

3 Results

We evaluated our fine-tuned models alongside classic DeepBind and the hybrid BERT-TFBS model (Section 2.2) across three distinct experimental scenarios to benchmark transcription factor binding site (TFBS) prediction performance: 1. cross-chromosome evaluation, 2. cross dataset evaluation, and 3. cross species evaluation.

For all models, hyperparameters were set to each architecture's default values. Training employed binary cross-entropy loss optimized via AdamW. Early stopping was triggered after 10 epochs without validation loss improvement, with the checkpoint exhibiting the lowest validation loss retained for testing. Given the binary classification nature of TFBS prediction, performance was comprehensively measured using five metrics: accuracy (ACC), F1 score, Matthews correlation coefficient (MCC), area under the receiver operating characteristic curve (ROC-AUC), and area under the precision-recall curve (PR-AUC). This consistent fine-tuning and evaluation pipeline

facilitates a direct comparison between specialized TFBS predictors and large-scale pretrained genomic models. Implementation details, including parameter counts, are described in Supplementary Section S2. The following sections present the results for each of the three evaluation scenarios.

3.1 Cross-Chromosome Evaluation on *A. thaliana* AREB/ABF1-4 Binding Sites

To assess each model's ability to generalize to unseen genomic contexts, we evaluated them using a leave-one-chromosome-out scheme on AREB/ABF1-4 transcription factor binding sites of *A. thaliana* species from `Sun2022` dataset. In this protocol, each of the five *A. thaliana* chromosomes is held out in turn for testing, while the remaining four serve for training and validation via 5-fold cross-validation. Table 1 summarizes the macro-mean performance metrics, as well as total training and inference times, aggregated over all 25 runs. The distribution of these metrics and runtimes for each held-out chromosome is shown in Supplementary Figure S1 (see Supplementary Table S6 for the per-chromosome averages across the five folds).

Table 1: Macro-mean performance metrics for AREB/ABF transcription factor binding site regions in *A. thaliana*, using a **leave-one-chromosome-out** protocol. Each of the five chromosomes is used once as a test set, and for each setting, 5-fold cross-validation is performed. Reported metrics are macro-averaged across all 25 runs (5 chromosomes $\times$ 5 folds). The top value in each column is highlighted in bold.

Model	ACC	F1	MCC	ROC-AUC	PR-AUC	Train Time(s)	Test Time(s)
DeepBind	0.893	0.893	0.786	0.957	0.960	506	**3.2**
BERT-TFBS	0.943	0.945	0.886	0.974	0.975	4,372	30
DNABERT-2	0.956	0.955	0.911	0.988	0.988	4,307	29
AgroNT	**0.971**	**0.970**	**0.941**	**0.997**	**0.998**	30,447	194
HyenaDNA	0.952	0.951	0.907	0.987	0.988	**216**	3.8

Statistical analysis using ANOVA followed by Tukey's HSD test reveals that across all performance metrics and chromosome splits, HyenaDNA performs comparably to DNABERT-2 and BERT-TFBS. Although AgroNT is statistically superior in some metrics, it requires the greatest computational time (30,447s training, 194s testing). BERT-TFBS and DNABERT-2 offer a favorable balance, achieving ACCs of $\approx$0.94–0.96 and ROC-AUCs of $\approx$0.974–0.988, with moderate runtimes ($\approx$4,300s to train, $\approx$30s to test). DeepBind, while fast to train ($\approx$506s), shows lower overall accuracy (ACC=0.893, MCC=0.786). HyenaDNA emerges as the most efficient choice: it matches DNABERT-2's high PR-AUC (0.987), achieves solid ACC (0.952) and MCC (0.907), yet completes training in only 216s with 3.8s per-chromosome inference. This remarkable combination of near–state-of-the-art accuracy and minimal computation makes HyenaDNA particularly well-suited for large-scale, genome-wide binding-site prediction.

3.2 Cross-Dataset Evaluation on *A. thaliana* AREB/ABF2 Binding Regions

To assess cross-dataset generalization, we trained on *A. thaliana* AREB/ABF2 peaks from the `Malley2016` dataset and evaluated on the corresponding *A. thaliana* AREB/ABF2 peaks in `Sun2022` (see Section2.1, cross-dataset (`Malley2016 - Sun2022`).

Figure 2 (a) illustrates the fold-wise distribution of five evaluation metrics (ACC, F1, MCC, ROC-AUC, and PR-AUC) for the cross-dataset scenario, with mean values reported in Supplementary Table S7. The three foundation models, HyenaDNA, DNABERT-2, and AgroNT, outperform the classic CNN DeepBind and the hybrid BERT-TFBS, each achieving average ROC-AUC and PR-AUC above 0.95 and average F1 and MCC above 0.85. In contrast, DeepBind peaks near 0.89 average ROC-AUC, 0.77 F1, and 0.64 average MCC. BERT-TFBS fails to converge in 2 out of 5 folds, leading to high variability and resulting in an average F1 score of 0.74 and MCC of 0.49,

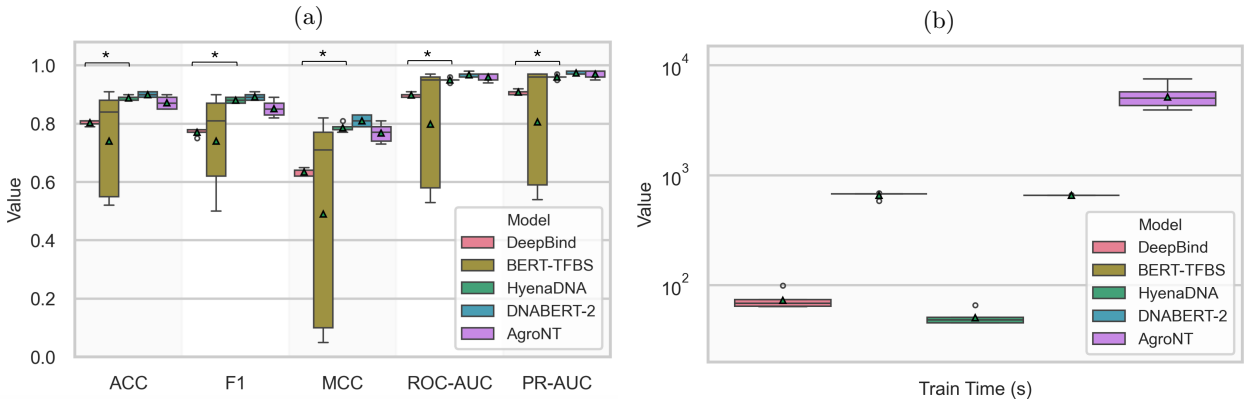

Figure 2: Distribution of **(a)** test-set performance metrics and **(b)** training time (in seconds). Results are based on 5-fold cross-validation, with each fold trained on the `Malley2016` AREB/ABF2 dataset and evaluated on the `Sun2022` *A. thaliana* AREB/ABF2 test set in a **cross-dataset setting**. Stars indicate statistically significant differences.

with other metrics remaining below 0.8. Due to this inconsistency, statistical testing was performed only between HyenaDNA and the other models using both t-test and Wilcoxon test. HyenaDNA significantly outperforms DeepBind across all primary metrics, with p-values consistently at or below 1.1×10^{-2} for Wilcoxon test and below 6.7×10^{-6} for t-test, while no other models showed statistically significant superiority over HyenaDNA (see Supplementary Section S3.2 for detailed p-values). Figure2 (b) shows the per-fold training times (see Supplementary Table S7 for each model's average time). Despite its large receptive field and single-nucleotide resolution, HyenaDNA is significantly faster than all other models, with p-values less than 0.032 for the Wilcoxon test and below 0.021 for the t-test (detailed p-values are given in Supplementary Section S3.2). It fine-tunes in just under one minute—on par with DeepBind. By comparison, DNABERT-2 and BERT-TFBS require roughly eleven minutes per fold, and the one-billion-parameter AgroNT demands over an hour and a half. These results demonstrate that HyenaDNA achieves state-of-the-art cross-dataset performance while maintaining DeepBind-level efficiency, outperforming both purely convolutional and hybrid alternatives in speed–accuracy trade-offs.

3.3 Cross-Species Transferability of AREB/ABF1-4 Binding Sites

To evaluate how well our models generalize beyond a single species, we assembled a cross-species dataset of AREB/ABF 1–4 DAP-seq peaks from two Brassicaceae: *A. thaliana* and its wild relative *S. irio* from `Sun2022`. We then carried out a leave-one-species-out evaluation—training once on *A. thaliana* and testing on *S. irio*, then vice versa.

Figure 3 presents the per-run distributions of different metrics for cross-species evaluation, with average values reported in Supplementary Table S8 from 5-fold cross-validation in each transfer direction. Despite the phylogenetic distance between these two species, all models maintain excellent performance under both transfer settings; even the simplest DeepBind baseline achieves average ROC-AUC ≥ 0.954 and average PR-AUC ≥ 0.956 for both directions. To evaluate statistical differences among models, we performed ANOVA followed by Tukey's HSD test. For ROC-AUC and PR-AUC, HyenaDNA performs comparably to AgroNT, DNABERT-2, and BERT-TFBS. For other metrics, it performs similarly to DNABERT-2 and BERT-TFBS, while requiring substantially less training time in both transfer directions, particularly compared to AgroNT. This demonstrates HyenaDNA's favorable balance of accuracy and efficiency, achieving state-of-the-art cross-species generalization without the runtime overhead associated with larger models.

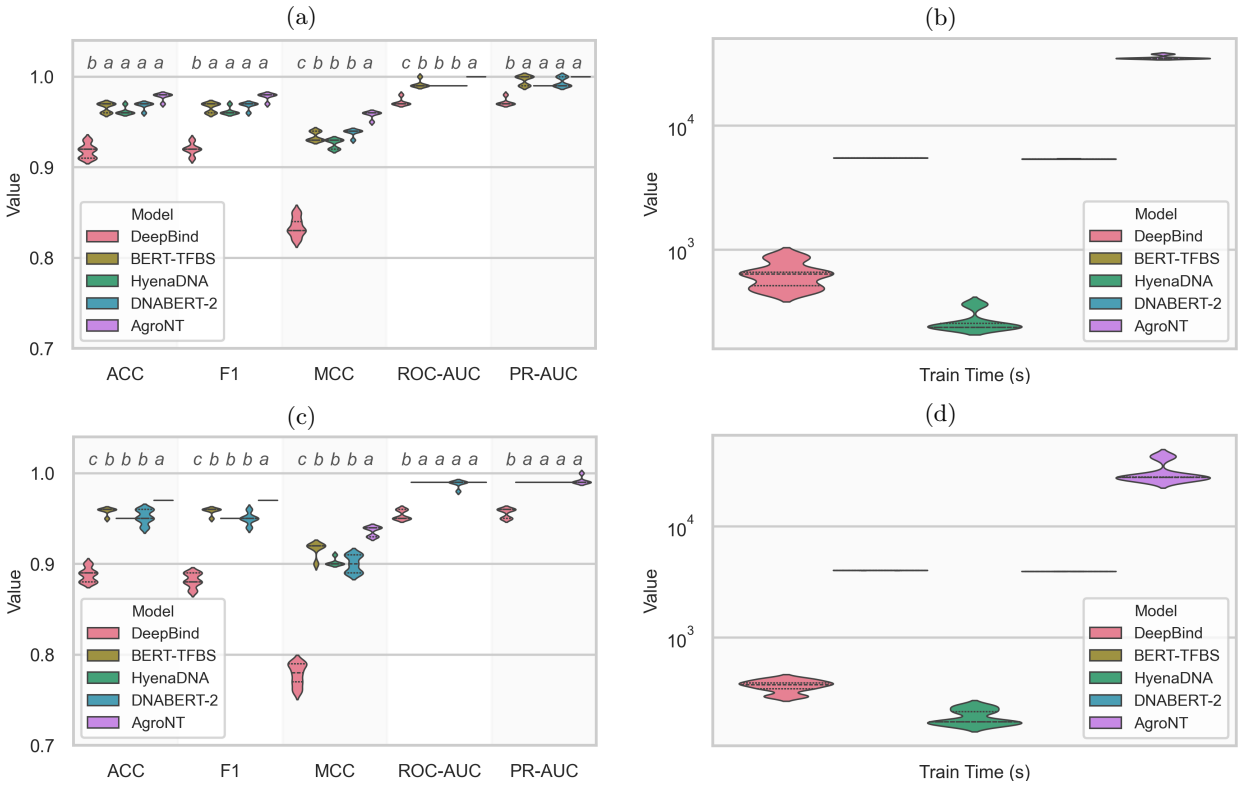

Figure 3: Distribution of test-set performance metrics and training time (in seconds) across DeepBind, BERT-TFBS, AgroNT, DNABERT-2, HyenaDNA in **cross-species evaluation**. Each distribution reflects results from 5-fold cross-validation, where models trained on each fold are evaluated on the test species. **(a–b)**: trained on *A. thaliana*, tested on *S. irio*; **(c–d)**: trained on *S. irio*, tested on *A. thaliana*. Letters on top of each box indicate statistical groupings; models sharing the same letter are not significantly different at the 0.05 significance level.

This robust cross-species performance is underpinned by the high conservation of AREB/ABF binding motifs: Sun et al. (2022) showed that the DNA-binding preferences of AREB/ABF transcription factors are essentially conserved across Brassicaceae species, including *A. thaliana* and *S. irio*, and comparison of the position weight matrices of all AREB/ABFs revealed little difference in binding site preference. Also, swap-DAP assays revealed extensive overlap, indicating that TF–DNA contacts are conserved despite species divergence [7]. This high degree of motif and binding-site conservation provides a mechanistic justification for why a model trained on one species can accurately predict AREB/ABF binding regions in the other.

3.4 Ablation Study on HyenaDNA

In this section, we conduct an ablation analysis of HyenaDNA to better understand how each component of the model contributes to its overall performance. We selected this model for its state-of-the-art predictive accuracy, exceptional training and inference efficiency, and minimal tunable parameter count among all fine-tunable models. Supplementary Table S9 lists the trainable parameters in HyenaDNA under the three freezing regimes. We can see that when the backbone is fully frozen (`Backbone`), only the 258-parameter classification head (<0.1%) is updated. Allowing embeddings and the classification head to train (`Backbone.layers`) increases the parameters to <0.6% of all parameters in the full model (represented by `None` in Table 2).

Tables 2a–2b compare these regimes across cross-chromosome, cross-species and cross-dataset

tasks. In every case, updating only the embeddings and classification head (`Backbone.layers`) delivers performance nearly identical to full fine-tuning while adjusting under 0.6% of the model. These results demonstrate HyenaDNA's parameter-efficient adaptability: by freezing over 99% of its weights, it maintains high accuracy and AUC, substantially reducing both computational load and memory usage with minimal impact on predictive power. Supplementary Figures S2-4 show the full distribution of metrics for each evaluation scenario.

Table 2: Performance metrics of HyenaDNA ablation experiments across different evaluation scenarios. Each sub-table displays the mean values of the performance metrics.

(a) Macro-mean performance for cross chromosome evaluation of AREB/ABF1-4 transcription factors of *A. thaliana*

Frozen Component	ACC	F1	MCC	ROC-AUC	PR-AUC
backbone	0.679	0.680	0.360	0.745	0.728
backbone.layers	0.892	0.892	0.785	0.957	0.961
none	0.952	0.951	0.907	0.987	0.988

(b) Mean performance for cross-dataset evaluation (trained on `Malley2016` AREB/ABF2, tested on `Sun2022` A. AREB/ABF2)

Frozen Component	ACC	F1	MCC	ROC-AUC	PR-AUC
backbone	0.592	0.554	0.190	0.644	0.612
backbone.layers	0.836	0.828	0.682	0.916	0.924
none	0.888	0.880	0.786	0.950	0.960

(c) Mean performance for cross-species evaluation for AREB/ABF1-4 transcription factors for *A. thalaina* and *S. irio*.

Split	Frozen Component	ACC	F1	MCC	ROC-AUC	PR-AUC
train:*A. thaliana* **test:***S. irio.*	backbone	0.680	0.688	0.364	0.750	0.734
	backbone.layers	0.908	0.906	0.816	0.968	0.970
	none	0.962	0.962	0.926	0.990	0.990
train:*S.irio.* **test:***A. thaliana*	backbone	0.684	0.680	0.374	0.754	0.736
	backbone.layers	0.888	0.888	0.780	0.958	0.958
	none	0.950	0.950	0.902	0.990	0.990

4 Conclusion

This study demonstrates the superior performance of large-scale pre-trained DNA foundation models, specifically DNABERT-2, AgroNT, and HyenaDNA, in predicting transcription factor binding sites (TFBSs) in plants. Our evaluations on *Arabidopsis thaliana* and *Sisymbrium irio* datasets reveal that these models, particularly HyenaDNA, outperform traditional architectures in both predictive accuracy and computational efficiency.

Looking ahead, our goal is to expand our analysis by incorporating a broader range of plant species to assess the generalizability of these models across diverse genomes. Currently, TFBS data for the same transcription factor (ABFs) is only available for a few species within the Brassicaceae family. However, upcoming multi-DAP-seq experiments [16] are expected to generate a rich dataset that will enable further testing of the computational pipeline developed in this work. Additionally, we plan to enhance model performance by integrating transcription factor-specific information, enabling the development of TF-aware models capable of transferring knowledge across different TFs. These advancements hold promise for scalable, genome-wide TFBS prediction, facilitating deeper insights into plant gene regulatory networks.

References

[1] Kerstin Kaufmann, Jose M Muino, Magne Østerås, Laurent Farinelli, Pawel Krajewski, and Gerco C Angenent. Chromatin immunoprecipitation (ChIP) of plant transcription factors followed by sequencing (ChIP-seq) or hybridization to whole genome arrays (ChIP-CHIP). Nature protocols, 5(3):457–472, 2010.

[2] Ronan C O'Malley, Shao-shan Carol Huang, Liang Song, Mathew G Lewsey, Anna Bartlett, Joseph R Nery, Mary Galli, Andrea Gallavotti, and Joseph R Ecker. Cistrome and epicistrome features shape the regulatory DNA landscape. Cell, 165(5):1280–1292, 2016.

[3] Wei Shen, Jian Pan, Guanjie Wang, and Xiaozheng Li. Deep learning-based prediction of tfbss in plants. Trends in Plant Science, 26(12):1301–1302, 2021.

[4] Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. DNABERT-2: Efficient foundation model and benchmark for multi-species genome. arXiv preprint arXiv:2306.15006, 2023.

[5] Javier Mendoza-Revilla, Evan Trop, Liam Gonzalez, Maša Roller, Hugo Dalla-Torre, Bernardo P de Almeida, Guillaume Richard, Jonathan Caton, Nicolas Lopez Carranza, Marcin Skwark, et al. A foundational large language model for edible plant genomes. Communications Biology, 7(1):835, 2024.

[6] Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, et al. HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution. Advances in neural information processing systems, 36:43177–43201, 2023.

[7] Ying Sun, Dong-Ha Oh, Lina Duan, Prashanth Ramachandran, Andrea Ramirez, Anna Bartlett, Kieu-Nga Tran, Guannan Wang, Maheshi Dassanayake, and José R Dinneny. Divergence in the ABA gene regulatory network underlies differential growth control. Nature plants, 8(5):549–560, 2022.

[8] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nature biotechnology, 33(8):831–838, 2015.

[9] Kai Wang, Xuan Zeng, Jingwen Zhou, Fei Liu, Xiaoli Luan, and Xinglong Wang. BERT-TFBS: a novel BERT-based model for predicting transcription factor binding sites by transfer learning. Briefings in Bioinformatics, 25(3):bbae195, 2024.

[10] Alan M Jones, Joanne Chory, Jeffery L Dangl, Mark Estelle, Steven E Jacobsen, Elliot M Meyerowitz, Magnus Nordborg, and Detlef Weigel. The impact of arabidopsis on human health: diversifying our portfolio. Cell, 133(6):939–943, 2008.

[11] Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P de Almeida, Hassan Sirelkhatim, et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. Nature Methods, 22(2):287–297, 2025.

[12] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692, 2019.

[13] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1715–1725, 2016.

[14] Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. Hyena hierarchy: Towards larger convolutional language models. In International Conference on Machine Learning, pages 28043–28078. PMLR, 2023.

[15] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV), pages 3–19, 2018.

[16] Leo A Baumgart, Sharon I Greenblum, Abraham Morales-Cruz, Peng Wang, Yu Zhang, Lin Yang, Cindy Chen, David J Dilworth, Alexis C Garretson, Nicolas Grosjean, Guifen He, Emily Savage, Yuko Yoshinaga, Ian K Blaby, Chris G Daum, and Ronan C O'Malley. Recruitment, rewiring, and deep conservation in flowering plant gene regulation. bioRxiv, 2025.