


Setting the Bar for DNA Language Model Finetuning


Exploring modern transfer learning techniques.

Mei-Ling Fang  , **Rahul Shrestha** , **Ibrahim Ebrar Yurt** , and **Arina Zemchyk** 

TUM School of Computation, Information and Technology, Technical University of Munich

 julie.fang@tum.de

 rahul.shrestha@tum.de

 ibrahim.yurt@tum.de

 arina.zemchyk@tum.de

July 16th, 2024

1 Introduction

Recent advancements in genomics and machine learning have paved the way for innovative approaches to understanding the complexity of genetic regulation across diverse species. Among these, species-aware DNA language models (LMs) have emerged as powerful tools for capturing the intricacies of regulatory elements and their evolution[15, 9]. These models, trained on vast genomic datasets spanning hundreds of species and millions of years of evolution, offer unprecedented insights into the functional landscape of DNA. By leveraging the evolutionary context inherent in multi-species comparisons, species-aware DNA LMs not only enhance our ability to predict gene expression and identify regulatory motifs, but also provide a framework for exploring the conversation and divergence of genetic elements across evolutionary time.

Building upon the success of species-aware DNA language models, exploring efficient fine-tuning methods for these models represents a crucial next step in advancing genomics research and applications. While the pre-trained species-aware DNA LMs have demonstrated remarkable capabilities in capturing regulatory elements and their evolution across diverse species, fine-tuning these models for specific downstream tasks can potentially unlock even greater performance and biological insights, especially when used for important downstream tasks such as gene expression prediction, motif discovery, or mRNA half-life prediction. Efficient fine-tuning methods are important because they allow researchers to adapt these powerful models to downstream tasks without the need for extensive computational resources or large labeled datasets. In addition, fine-tuning these models could potentially help researchers expand the understanding of rare or understudied species. By optimizing the fine-tuning process, DNA language models are expected to be more accessible and applicable to a wider range of bi-

ological questions and practical applications in fields such as personalized medicine, conversation biology, and agricultural genomics.

The primary goals of our research project are:

1. Explore the performance of different fine-tuning techniques with DNA Language Models, by applying it to downstream tasks.
2. Explore sensitivity of the techniques to different hyperparameters (rank, batch size, learning rate).

To answer this, we delve into the exploration of state-of-the-art fine-tuning methods for species-aware DNA language models, with a particular focus on Low-Rank Adaptation (LoRA)[8], Weight-Decomposed Low-Rank Adaptation (DoRA)[12] and Infused Adapter by Inhibiting and Amplifying Inner Activations (IA3)[11]. These advanced methods are rigorously compared against our baseline approach, which involves extracting embeddings from the last layers of the DNA LM and training of a linear classifier. Our investigation extends beyond mere performance comparisons, as we also examine the transferability of the encoded knowledge within these species-aware DNA LMs. To this end, we leverage the downstream task datasets of the Nucleotide Transformer available on Huggingface[4], applying our fine-tuned models to a diverse set of genomic prediction tasks. Furthermore, we conduct a comprehensive analysis of the impact of hyperparameter adjustments on the performance of DNA LMs, seeking to identify optimal configurations. The analysis conducted allows us to access the efficiency and effectiveness of various fine-tuning methods, as well as the robustness and generalizability of the species-aware DNA LMs across different genomic contexts.

The remainder of this report is structured as follows: we first describe the fine-tuning methods for species-aware DNA LMs, followed by a presentation of the datasets and experimental results. We then discuss our

findings, their implications, and conclude the report with the future research directions.

2 Fine-tuning Methods

The adaptation of large language models (LLMs), including DNA LMs, for specific tasks presents significant challenges due to their immense size and complexity. Traditional transfer learning techniques in natural language processing (NLP), such as feature-based transfer and full model fine-tuning, often prove impractical or inefficient when applied to these models. For instance, state-of-the-art LLMs like GPT-4 (1.76 trillion parameters)[1], Google’s T5 Flan XXL (11 billion parameters)[3], and Meta’s Llama (65 billion parameters)[13] require substantial computational resources and extensive datasets for effective fine-tuning. Moreover, conventional fine-tuning methods can lead to catastrophic forgetting[5], where the model loses previously acquired knowledge.

To address these challenges, parameter-efficient fine-tuning (PEFT) techniques have emerged as promising solutions. These methods, including adapter modules[7], prefix-tuning[10], and LoRA, enable the fine-tuning of LLMs with minimal parameter updates. They are classified into three categories[12]: adapter-based, prompt-based, and LoRA and its variants, with the latter being the most promising and the main focus of recent studies in the field of domain adaptation. PEFT approaches offer a significant reduction in computational demands while preserving model performance, thereby enabling the adaptation of large-scale models using consumer-grade hardware and modest-sized datasets. Given their efficiency and effectiveness, these PEFT methods form the core focus of our research report, as we explore their application to species-aware DNA language models.

2.1 Baseline: lightweight fine-tuning

Historically, the utilization of large pre-trained language models for downstream tasks in genomics has followed a similar trajectory to that in NLP. The initial approach involved freezing the majority of the pre-trained models’ layers and only training a small number of task-specific layers appended to the model. This method has been applied to various genomic tasks, including gene expression prediction and regulatory element identification[14], as the rich representations learned from vast genomic datasets spans multiple

species, potentially capturing complex evolutionary patterns and regulatory information[9].

For our baseline model, we adopt this approach by freezing all layers of Species-LM. In addition, we append a single linear layer on top of the species-aware DNA LM and train it with the task-specific dataset.

2.2 Low-Rank Adaptation (LoRA)

LoRA[8] is an adaptation technique based on the hypothesis that pre-trained language models (PLMs) have a low intrinsic dimension, which can be exploited by training rank decomposition matrices inserted into each layer of the PLM. The main advantage is that LoRA performs on par with or better than full fine-tuning. Unlike adapter methods, LoRA avoids introducing the inference latency, which refers to the time it takes for a model to generate a prediction or output after receiving an input, while maintaining the original input sequence length. By adding trainable pairs of rank decomposition matrices in parallel to existing weight matrices, LoRA allows for efficient adaptation of the model to specific tasks or domains.

2.3 Weight-Decomposed Low-Rank Adaptation (DoRA)

DoRA[12] builds upon the principles of LoRA while addressing some of its limitations. The key idea behind DoRA is to decompose the weight updates into two components: magnitude (the model’s weights change) and direction (in which direction the model’s weights change). This decomposition allows for more flexible and efficient adaptation of pre-trained models, potentially leading to better performance on specific tasks. When applying on species-aware DNA language models, it can capture the subtle variations in regulatory elements across diverse species. Like LoRA, DoRA maintains the benefits of not introducing inference latency and not reducing input sequence length.

2.4 Infused Adapter by Inhibiting and Amplifying Inner Activations (IA3)

IA3[11] introduces small, trainable vectors that modulate the activations within the model’s existing architecture. These vectors act by either inhibiting or amplifying the inner activations of the model, effectively steering its behavior towards the target task without altering the original model weights. This method is particularly efficient, as it requires training only a fraction

of the parameters compared to full fine-tuning, typically less than 0.1% of the model’s total parameters

3 Experimental Results

In this section, we briefly introduce the architectural design of the species-aware DNA language model and the experimental setup. Then, we present the results of our experiments for each downstream task. In total, we explored 3 classification tasks and 1 regression task. For each task, we provide a brief description of its objective, followed by a detailed report of our experimental results.

3.1 Species-aware DNA Language Model

The species-aware DNA language model, SpeciesLM[9] is used for all experiments in this project. SpeciesLM is trained on a diverse set of highly evolutionarily diverged eukaryotes, and the specific downstream DNA LM we used in our study is trained on the 3’ regions. The architecture of SpeciesLM is based of the DNABERT model, which is a transformer-based model adapted for DNA sequences. The input of SpeciesLM consists of DNA sequences with an additional species token, which allows the model to learn species-specific patterns while still capturing conserved elements across species. The model uses k-mer tokenization, a common approach in DNA language models where sequences are split into overlapping k-mers. In our set up, k is set to 6 and the species token "candida_glabrata" is given in every experiment. The species token is concatenated with the k-merized sequence and then it is tokenized before being processed by the transformer layers.

During training, the model employs a masked language modeling objective. Random nucleotides in the input sequences are masked, and the model is trained to reconstruct these masked nucleotides from the surrounding context. The core of the model consists of multiple transformer layers, which use self-attention mechanisms to process the input sequences and capture long-range dependencies. The final layer of the model predicts the probability distribution over possible nucleotides for each masked position.

In our experimental configuration, only the encoder of the pre-trained DNA language model is extracted. The layers of the encoder are initially frozen, maintaining their pre-trained weights. A linear layer is then appended to the end of the encoder. This linear layer

serves as the task-specific output layer, mapping the encoder’s high-dimensional representations to the required output format for our particular task. In the baseline model, only the linear layer is trained. In the fine-tuning models, dependent on the fine-tuning strategies, the fraction of the encoder parameters along with the linear layer are trained.

3.2 Enhancer sequence prediction

3.2.1 Objective and Datasets

Enhancers are special regions in DNA that can boost the activity of genes, which can be located near or far from the genes they regulate, sometimes even thousands of base pairs away. Understanding these enhancers is crucial for predicting how genes might be turned on or off in different situations, which could be helpful for understanding the biological processes of diseases.

The dataset consists of a non-synthetic dataset[6] from human reference genome data (742 strong enhancers, 742 weak enhancers and 1484 non-enhancers) and a synthetic dataset (6000 synthetic enhancers and 6000 synthetic non-enhancers) produced through a generative model[4]. The classification task aims to predict the strength or effectiveness of enhancers in gene regulation using DNA sequence alone. It involves distinguishing between two types of enhancers: enhancers (positive labels) and non-enhancers (negative labels). We refer the detailed creation of the dataset to the paper[4].

3.2.2 Experimental Results

The dataset was split into 12,722 training sequences, 2,246 validation sequences and 400 testing sequences. First, we report the best accuracies (see figure 1) achieved by different fine-tuning methods for the species-aware DNA-LM, with a batch size of 128, a learning rate of 1e-3 and 5 epochs. On average, each epoch took around 6 minutes to train. The DoRA (r=2) method achieved the highest accuracy at 75.25%, followed by the baseline and IA3 both at 72.50%, and LoRA (r=2) at 72%.

Figure 2 illustrates the training losses over 5 epochs for different fine-tuning methods applied to the species-aware DNA-LM model. The Baseline model shows fluctuating losses with a notable dip at epoch 3. IA3 maintains relatively low losses throughout, while both LoRA and DoRA exhibit higher initial losses, peaking at epoch 2 before decreasing significantly.

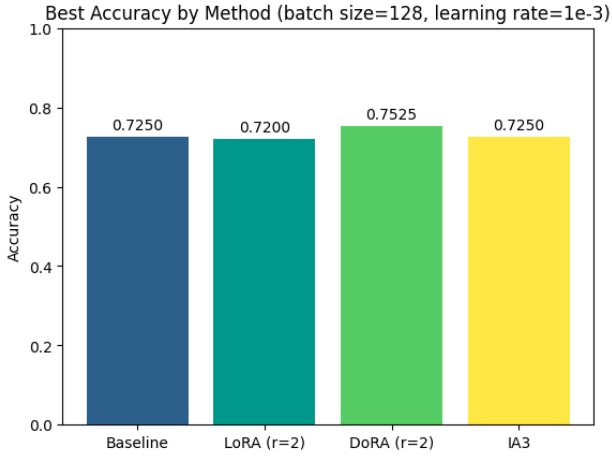


Figure 1 Best accuracy by methods

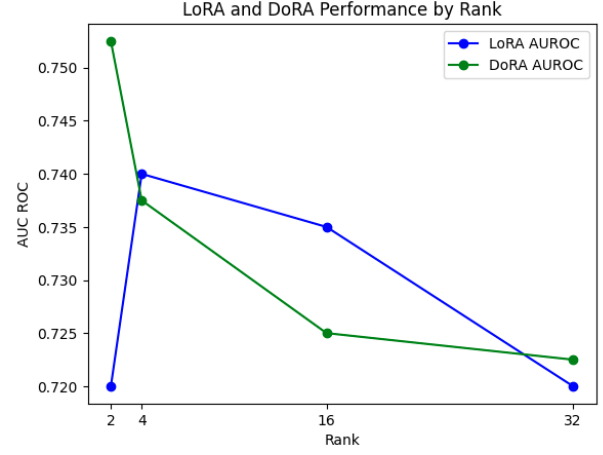


Figure 3 LoRA and DoRA Performance by Rank

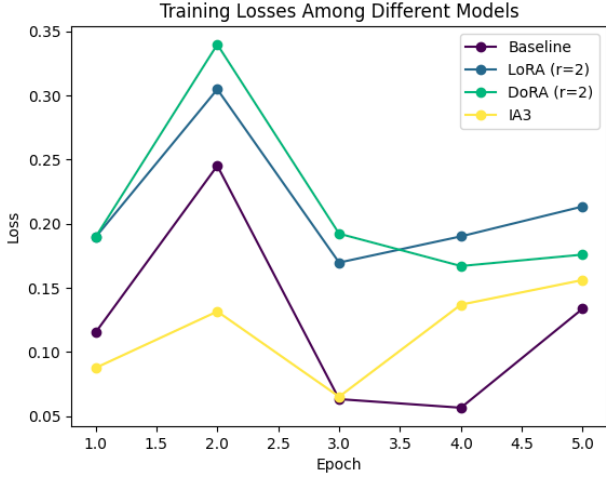


Figure 2 Training losses among different fine-tuning models

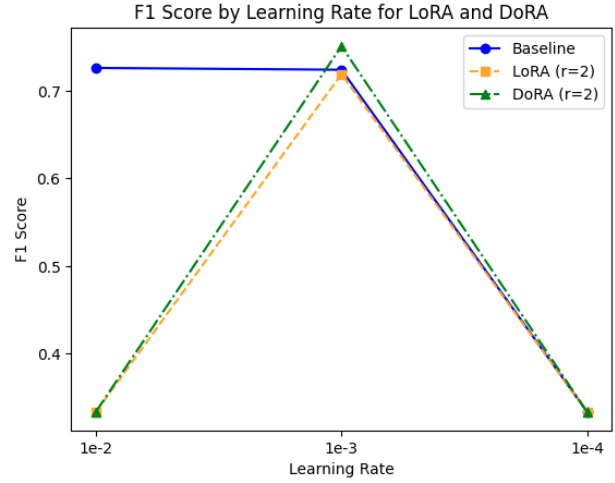


Figure 4 F1 Score by Learning Rate for LoRA and DoRA

Subsequently, we compared the AUROC performance of LoRA and DoRA methods across different ranks, see figure 3. The DoRA method achieves the highest AUROC at rank 2, while the LoRA method shows a more gradual decrease in performance as the rank increases.

F1 scores across different learning rates among various models are also investigated, see figure 4. The Baseline model (blue) maintains a relatively stable F1 score at higher learning rates, but drops significantly at 1e-4. Both LoRA and DoRA models achieve their highest F1 scores at a learning rate of 1e-3, with DoRA outperforming LoRA. However, their performance declines at learning rates of 1e-2 and 1e-4.

3.3 Promoter sequence prediction

3.3.1 Objective and Datasets

Promoters are regulatory regions of DNA located near the start of a gene. They control when and how much a gene is transcribed into RNA. Promoters typically contain specific DNA sequences that act as binding sites for proteins involved in initiating transcription.

This promoter detection task is to identify DNA sequences as TATA-box[2] promoters or non-TATA promoters. The TATA box, also called the Goldberg-Hogness box, is a specific type of promoter element found in many eukaryotic genes. The name "TATA" comes from the DNA sequence of this element, which typically contains a TATAAA consensus sequence. These TATA boxes are usually located about 25-35 base pairs upstream of the transcription start site. They serve as a binding site for the TATA-binding protein (TBP), which is part of the transcription factor TFIID. This binding helps position the RNA

polymerase II enzyme at the correct start site for transcription. Not all promoters contain TATA boxes, which are non-TATA promoters. The presence of a TATA box can influence the regulation and expression pattern of a gene.

The dataset consists of 47767 train sequences and 5299 test sequences. Each sequence is 300 base pairs (bp) long, spanning 249 bp upstream and 50 bp downstream of transcription start sites. We refer the detailed creation of the dataset to the paper[4].

3.3.2 Experimental Results

The dataset was split into 40,601 training sequences, 7,166 validation sequences and 5,299 testing sequences. The results, as shown in Figure 5, indicate that both LoRA and DoRA models significantly outperform the baseline model, achieving nearly identical high accuracies. The IA3 also demonstrates improved performance over the baseline.

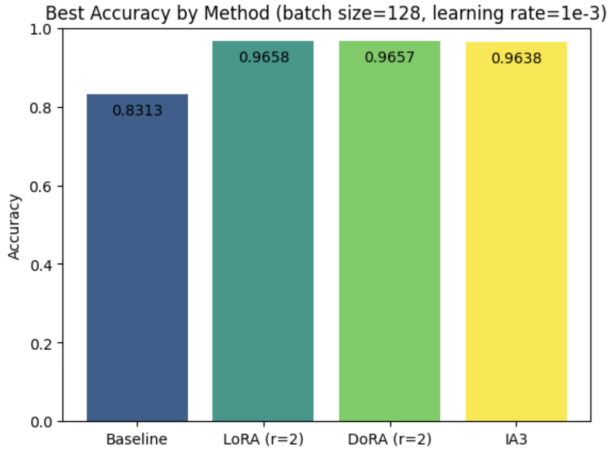


Figure 5 Best accuracy by methods

Figure 6 compares the accuracy of LoRA and DoRA models across different ranks. Both models show accuracy improvements as the rank increases, with the highest accuracy observed at rank 16 for both models. LoRA starts with a slightly lower accuracy at rank 2, but achieves marginally higher accuracy at its peak compared to DoRA.

We also analyzed the sensitivity of the models to different learning rate, Figure 7. The baseline model experiences a significant drop in accuracy when the learning rate increases from $1e-3$ to $1e-2$. Both LoRA and DoRA models with the rank set to 16 maintain high accuracy at a learning rate of $1e-3$, but show a notable decrease at $1e-2$, indicating their effectiveness at lower learning rates.

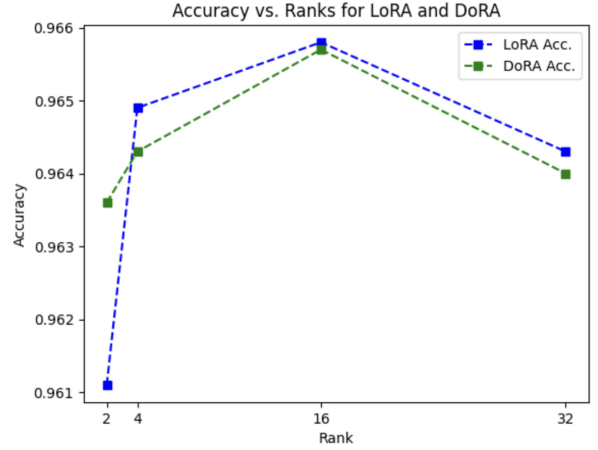


Figure 6 LoRA and DoRA Performance by Rank

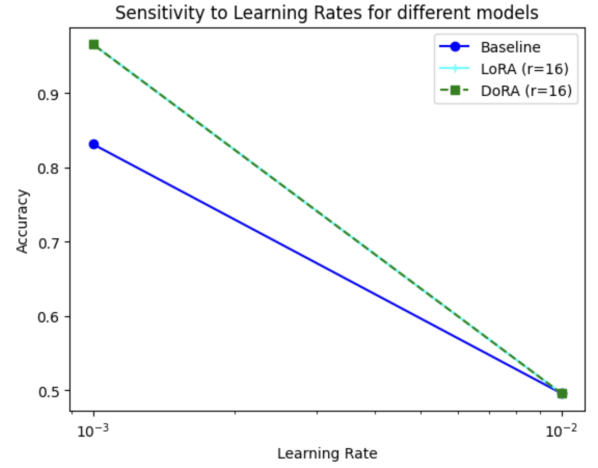


Figure 7 Accuracy in different learning rates

During training with a batch size of 512, we encountered a CUDA out-of-memory error. The average training time with batch size 128 is approximately 30 minutes. Figure 8 shows that LoRA configurations consistently exhibit shorter runtimes compared to DoRA configurations at each rank, highlighting the computational efficiency of LoRA. However, other factors such as the size of the dataset and hyperparameters might affect the runtime.

3.4 Epigenetic marks prediction

3.4.1 Objective and Datasets

Chemical modifications on histone proteins and DNA in the yeast genome, known as epigenetic marks, regulate gene expression without altering the DNA sequence. These modifications include acetylation and methylation of nucleosome occupancies, which affect chromatin structure and gene activity. Typically, acetylation results in a relaxed chromatin state and enhanced

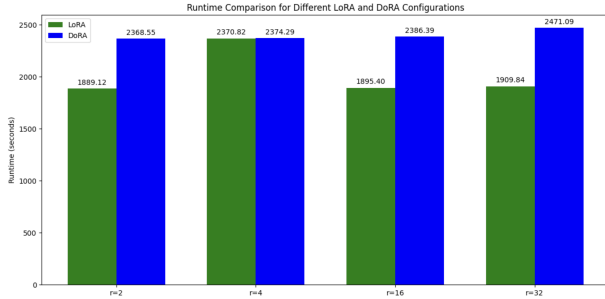


Figure 8 Runtime comparison between different configurations

gene expression, whereas methylation can either activate or repress gene expression depending on its specific context and location.

The dataset consisted of epigenetic marks identified in the yeast genome, namely acetylation and methylation nucleosome occupancies. The nucleosome occupancy values in the datasets were obtained and further processed into positive and negative observations to provide epigenetic training data for the following histone marks: H3, H4, H3K9ac, H3K14ac, H4ac, H3K4me1, H3K4me2, H3K4me3, H3K36me3 and H3K79me3.

3.4.2 Experimental Results

For our experiments, we selected a subset of the dataset for the histone mark, "H3". This was split into 11447 training sequences, 1497 validation sequences and 2021 test sequences. The labels are almost equally split into two classes: 0 (48.8%) and 1 (51.2%). Since the dataset is somewhat balanced, we've chosen accuracy as the metric to measure the model's performance.

All experiments were performed for 5 epochs with a learning rate of 0.001, batch size of 64 and weight decay of 0.001. For LoRA and DoRA, the LoRA alpha is set to 32, and the LoRA dropout is set to 0.01.

In Figure 9, we report the best-performing accuracy of different methods. The baseline, which involves, freezing all layers except the final classification layer, had the lowest performance of 0.82. All the other methods performed similarly, with IA3 scoring a bit lower at 0.87 than LoRA and DoRA.

We performed another experiment to observe how the rank affects accuracy. In Figure 10, we can see that the accuracy peaks around a rank of 16 and 32 for both LoRA and DoRA. The accuracy drops when increasing the rank to 64, although it is not significant. These results show that the rank doesn't have a major influence in the model's accuracy for this downstream task.

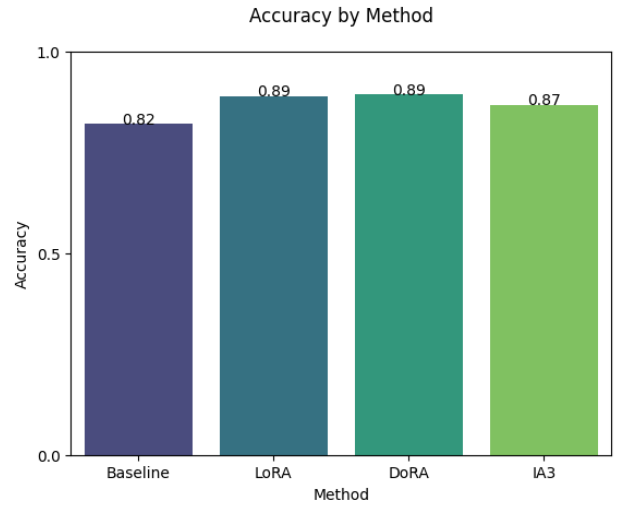


Figure 9 Comparison of accuracy for Epigenetic marks prediction

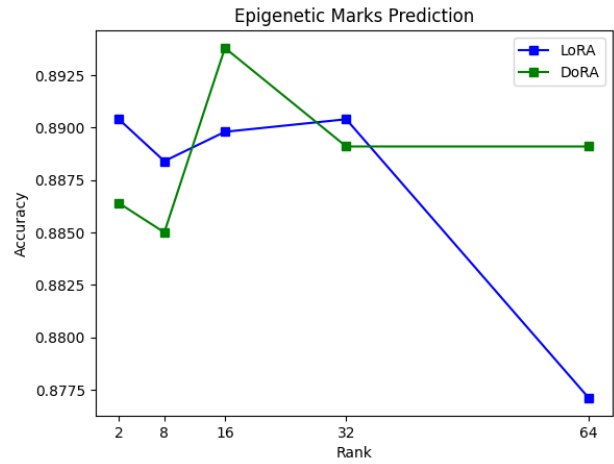


Figure 10 Accuracy by ranks for Epigenetic marks prediction

3.5 *Saccharomyces cerevisiae* expression prediction

3.5.1 Objective and Datasets

This dataset contains genome sequences from *Saccharomyces cerevisiae*, commonly known as baker's yeast, along with corresponding expression values for each sequence. *Saccharomyces cerevisiae* is a model organism extensively used in molecular and cellular biology due to its well-characterized genome, ease of genetic manipulation, and relevance to higher eukaryotes, including humans. The primary objective of analyzing this dataset is to develop predictive models capable of forecasting gene expression levels based on the underlying genome sequences, with a specific focus on the impact of different 3' sequences. Such models are essential for advancing our understanding of gene

regulation, cellular processes, and the functional genomics of *Saccharomyces cerevisiae*.

The dataset is derived from a reporter assay designed to measure the impact of different 3' sequences on gene expression. In this assay, all parts of the sequence are constant except for the 3' sequence, isolating its effect on gene expression. The 3' end genomic region encodes various regulatory processes, including mRNA stability, 3' end processing, and translation. By employing selected fine-tuning methods, we aim to identify patterns within the 3' sequence data that can predict gene expression levels. These predictions can help elucidate the regulatory networks and pathways influenced by the 3' sequences, enhancing our understanding of the organism's biology and its applications in biotechnology and research.

3.5.2 Experimental Results

The dataset was split into 11337 training sequences, 1418 validation sequences and 1417 test sequences. All the training was done for 5 epochs. For LoRA and DoRA configurations we always used `lora_alpha=32`, `lora_dropout=0.01`, and `bias="none"`. The hyperparameters we changed across experiments were batch size, learning rate and rank (for LoRA and DoRA).

First, we report the sensitivity of the models to different learning rates and batch sizes. Figure 11 and 12 shows these results for LoRA, for learning rates and batch sizes respectively. We also had similar results in our experiments with DoRA.

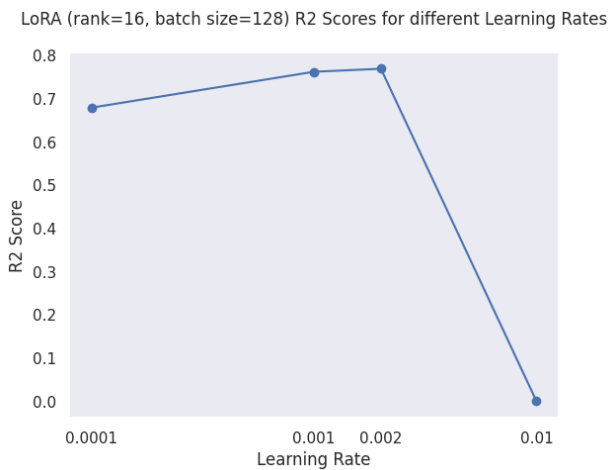


Figure 11 R2 Score by Learning Rates

Figure 13 compares the R2 Score of LoRA and DoRA models across different ranks for batch size 16, and learning rate $2e-3$. For this setup we can see that DoRA consistently performed very well across

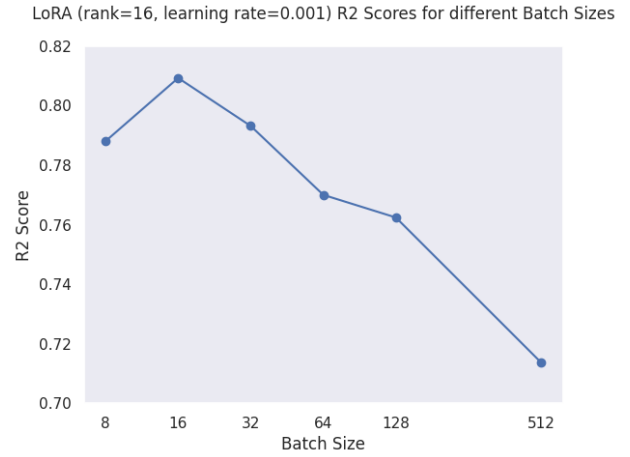


Figure 12 R2 Score by Batch Sizes

different ranks, while LoRA with rank 2 did not work. DoRA also outperformed LoRA for all different ranks. Best result in Figure 13 is with DoRA rank 32 which is also the best result throughout our experiments.

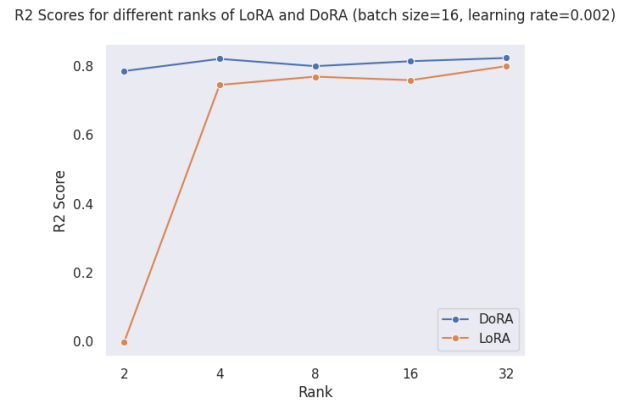


Figure 13 LoRA and DoRA performance by ranks

Next, we report the best R2 scores achieved by different fine-tuning methods overall (see Figure 14) and also for a fixed batch size, and learning rate (see Figure 15). In Figure 14 hyperparameters for each method is as follows,

- Baseline: batch size=32, learning rate=0.002
- IA3: batch size=32, learning rate=0.01
- LoRA: batch size=32, learning rate=0.002, rank=16
- DoRA: batch size=16, learning rate=0.002, rank=32

As expected, all the parameter efficient fine-tuning techniques had better scores than the baseline.

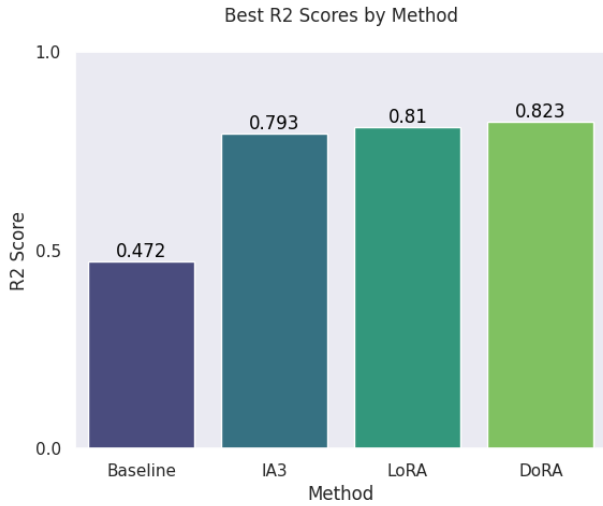


Figure 14 Best R2 Score by methods

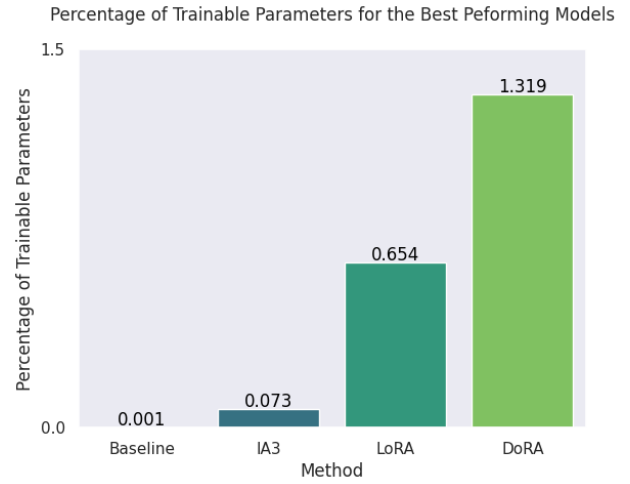


Figure 16 Percentage of trainable parameters by method

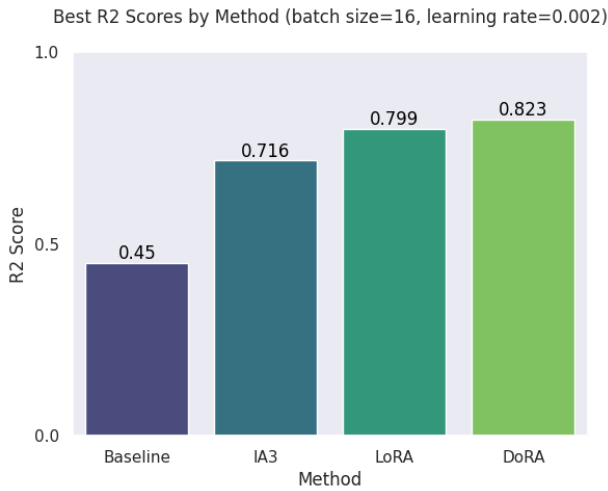


Figure 15 Best R2 Score by methods with batch size=16 and learning rate=0.002

It is important to note that even though our experiments focused mainly on LoRA and DoRA and we had less training runs for IA3 and Baseline, we can see that IA3 had comparable performance to LoRA and DoRA with significantly lower trainable parameters as shown in Figure 16

Regarding baseline and IA3 models, we have observed a surprising similar pattern when testing R2 score against different learning rates. For both models when increasing the learning rate the performance of the models drastically improves (see Figure 17 and 18). In case of IA3 the highest learning rate, which is equal to 0.01, even yields the best performance across all hyperparameter setups ($R2 = 0.79$ with batch size = 32).

Our assumption is that since for baseline and IA3 the fraction of parameters is significantly less than in

LoRA and DoRA, the models exhibit similar performance change.

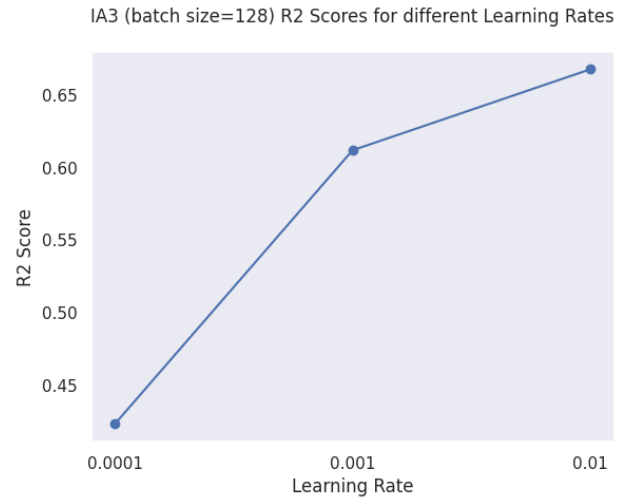


Figure 17 R2 Score by Learning Rates for IA3

Overall, for this dataset small/medium batch sizes seem to work best with 16 and 32 yielding best results. In terms of learning rate, for LoRA and DoRA 0.002 worked the best while for IA3 and baseline higher learning rates seem to perform better.

Finally, as mentioned earlier our best performing model was DoRA with rank 32, batch size 16, learning rate 0.002. On test set it had $R2 \text{ Score} = 0.839$ and $(\text{Pearson } R)^2 = 0.862$ (which was 0.704 in the paper [9]).

4 Discussion

In this section, we start by comparing the performance of both the finetuned species aware DNA

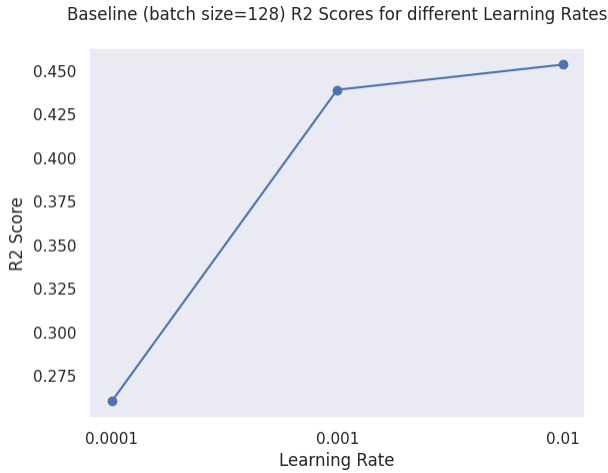


Figure 18 R2 Score by Learning Rates for baseline

language model and the finetuned Nucleotide Transformer model on downstream tasks, as it is essential to evaluate their capabilities and to identify which model is more effective and suitable for specific genomic prediction tasks. In the end, we briefly talk about results on other regression tasks from [9].

4.1 Comparative Analysis with Nucleotide Transformer

While both models are DNA language models, the species-aware DNA language model is specifically designed to leverage species information for improved performance on regulatory element prediction tasks, whereas the Nucleotide Transformer is a larger, more general-purpose model trained on a wider range of species. The authors of Nucleotide Transformer used probing through embedding extraction at various layers, as well as a parameter-efficient fine-tuning method called LoRA. They used a batch size of 8 and the Adam optimizer with a learning rate of $3e-3$ for fine-tuning.

On the enhancer prediction task, the fine-tuned Nucleotide Transformer achieved an MCC score of 0.593, while our fine-tuned species-aware DNA language model reached an accuracy of 0.7525. Both models surpassed the previous state-of-the-art LSTM-CNN baseline of 0.395. However, it's important to note that MCC and accuracy cannot be directly compared due to their different scales and interpretations. The species-aware model's higher accuracy suggests it may be particularly effective at capturing species-specific regulatory elements, but a direct performance comparison would require using the same metric for both models.

A similar case exists with the epigenetic marks prediction task, where the paper achieved a MCC score of 0.814, while the fine-tuned SpeciesLM had an accuracy of 0.89 with LoRA and DoRA. Even though it isn't comparable, the results show that SpeciesLM performed well in this downstream task.

For the promoter non-TATA prediction task, our species-aware DNA language model achieved an impressive F1 score of 0.9658, which is competitive with the larger Nucleotide Transformer's score of 0.977. This strong performance highlights the potential of our smaller species-aware model. Its ability to recognize important genomic elements, particularly promoter regions without TATA boxes, suggests good transferability across different genomic tasks. Notably, we also found that diverging the hyperparameters do not contribute to a better performance in several downstream tasks.

Overall, the species-aware DNA language model's success in fine-tuning demonstrates its adaptability and the effectiveness of our approach. Given its smaller size and competitive performance, our species-aware DNA language model could offer advantages in computational efficiency and resource requirements, making it suitable for researchers with limited resources or those needing faster inference times. Additionally, the model's species-awareness feature may provide benefits when working with diverse genomic datasets across different organisms.

4.2 Another Regression Task: RNA Half-life Prediction

We have also applied the same regression approach in Section 3.5 for two different RNA half-life Prediction tasks that are also mentioned in [9] (*S. cerevisiae* and *S. pombe*). Even though we have tried various ranks, batch sizes, learning rates and epochs, validation loss never really had a decreasing trend. Our best results had around 0.2 R2 score which is way worse than the results in the paper.

5 Conclusion

To this end, we have shown that species-aware DNA LMs, when combined with parameter-efficient fine-tuning methods such as LoRA, DoRA, and IA3, can significantly enhance the predictive capabilities of these models for downstream genomic tasks. Our findings indicate that compared to larger and more general-purpose models, fine-tuning based on smaller

scaled pre-trained language models not only achieve the competitive performance, but also maintain computational efficiency, suggesting potential advantages in computational efficiency and resource requirements for researchers with limited resources or those needing faster inference times.

Moving forward, potential future steps could include:

- Further exploration of fine-tuning methods: investigating additional parameter-efficient fine-tuning methods and comparing their performance on species-aware DNA LMs to identify the most effective approaches for specific genomic prediction tasks; probing alternative fine-tuning paradigms other than parameter-efficient fine-tuning methods, such as adapter-based fine-tuning, prompt-based fine-tuning.
- Expansion to diverse genomic contexts: extending the application of species-aware DNA LMs and fine-tuning methods to a broader range of genomic prediction tasks, including rare or understudied species, to enhance the understanding of genetic regulation across diverse organisms.
- Integration of multi-omics data: exploring the integration of multi-omics data with species-aware DNA LMs and fine-tuning methods to enhance the understanding of complex biological processes and regulatory networks.

Overall, our exploration effort has laid the groundwork for advancing genomics research and applications by leveraging state-of-the-art fine-tuning methods for species-aware DNA language models, and it has opened up exciting opportunities for future research and development in the field of computational molecular medicine.

References

- [1] OpenAI Josh Achiam et al. “GPT-4 Technical Report”. In: 2023. URL: <https://api.semanticscholar.org/CorpusID:257532815>.
- [2] J Y Chen, Min Ding, & David S. Pederson. “Binding of TFIID to the CYC1 TATA boxes in yeast occurs independently of upstream activating sequences.” In: *Proceedings of the National Academy of Sciences of the United States of America* 91 25 (1994), pp. 11909–13. URL: <https://api.semanticscholar.org/CorpusID:25068345>.
- [3] Hyung Won Chung et al. “Scaling Instruction-Finetuned Language Models”. In: *ArXiv abs/2210.11416* (2022). URL: <https://api.semanticscholar.org/CorpusID:253018554>.
- [4] Hugo Dalla-Torre et al. “The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics”. In: *bioRxiv* (2023). URL: <https://api.semanticscholar.org/CorpusID:255943445>.
- [5] Wenyu Du et al. “Unlocking Continual Learning Abilities in Language Models”. In: 2024. URL: <https://api.semanticscholar.org/CorpusID:270710911>.
- [6] Qitao Geng, Runtao Yang, & Lina Zhang. “A deep learning framework for enhancer prediction using word embedding and sequence generation.” In: *Biophysical chemistry* 286 (2022), p. 106822. URL: <https://api.semanticscholar.org/CorpusID:247006210>.
- [7] Neil Houlsby et al. “Parameter-Efficient Transfer Learning for NLP”. In: *ArXiv abs/1902.00751* (2019). URL: <https://api.semanticscholar.org/CorpusID:59599816>.
- [8] J. Edward Hu et al. “LoRA: Low-Rank Adaptation of Large Language Models”. In: *ArXiv abs/2106.09685* (2021). URL: <https://api.semanticscholar.org/CorpusID:235458009>.
- [9] Alexander Karollus et al. “Species-aware DNA language models capture regulatory elements and their evolution”. In: *Genome Biology* 25 (2023). URL: <https://api.semanticscholar.org/CorpusID:260117337>.
- [10] Xiang Lisa Li & Percy Liang. “Prefix-Tuning: Optimizing Continuous Prompts for Generation”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* abs/2101.00190 (2021). URL: <https://api.semanticscholar.org/CorpusID:230433941>.
- [11] Haokun Liu et al. “Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning”. In: *ArXiv abs/2205.05638* (2022). URL: <https://api.semanticscholar.org/CorpusID:248693283>.

- [12] Shih-yang Liu et al. “DoRA: Weight-Decomposed Low-Rank Adaptation”. In: *ArXiv* abs/2402.09353 (2024). URL: <https://api.semanticscholar.org/CorpusID:267657886>.
- [13] Hugo Touvron et al. “LLaMA: Open and Efficient Foundation Language Models”. In: *ArXiv* abs/2302.13971 (2023). URL: <https://api.semanticscholar.org/CorpusID:257219404>.
- [14] Jian Zhou et al. “Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk”. In: *Nature genetics* 50 (2018), pp. 1171–1179. URL: <https://api.semanticscholar.org/CorpusID:49868839>.
- [15] Zhihan Zhou et al. “DNABERT-S: LEARNING SPECIES-AWARE DNA EMBEDDING WITH GENOME FOUNDATION MODELS”. In: *ArXiv* (2024). URL: <https://api.semanticscholar.org/CorpusID:267658114>.