Adapting Nucleotide Transformers for Genomic Disease Prediction: LoRA Fine-Tuning vs. Linear Probing on a KEGG Benchmark

Prabashwara D.G.H.
Department of Computer Science Engineering
Faculty of Engineering, University of Moratuwa
Colombo, Sri Lanka
hansana.21@cse.mrt.ac.lk

***Abstract*—Foundation models trained on DNA sequences have shown promise in genomics, but their effectiveness in complex reasoning tasks remains limited. The Nucleotide Transformers (NT) has been applied in hybrid frameworks, which integrate large language models (LLMs) for biological reasoning. However, NT alone performs suboptimally in downstream predicting tasks. In this work, we explore incremental strategies to improve NT performance on a KEGG-derived dataset. We evaluate two approaches: (i) fine-tuning NT-500M with Low-Rank Adaptation (LoRA) on variant and reference sequences, and (ii) leveraging frozen NT-500M embeddings with a multi-layer perceptron (MLP) classifier. Results show that while LoRA finetuning provides modest gains, using frozen embeddings with MLP achieves higher accuracy (91.78%) and F1-score (78.68%). These findings highlight the potential of embedding-based approaches for improving single DNA foundation models without relying on heavy LLM integration.**

***Index Terms*—Nucleotide Transformer, LoRA, KEGG**

I. Introduction

Large pretrained models for biological sequences are rapidly improving the state of computational genomics by learning transferable sequence embeddings that can be adapted to downstream tasks [1] . The Nucleotide Transformer (NT) [2] family provides transformer-based DNA models that capture nucleotide dependencies and achieve strong performance on many tasks when fine-tuned or used as feature extractors.

Despite these advances, using a DNA foundational model for biological reasoning (for example, answering disease–variant relation questions) can be challenging. Recent works like BioReason [3] showed significant gains by combining DNA FMs with large language models (LLMs) and projecting DNA embeddings into the LLM space to support reasoning, but such hybrids are computationally expensive and complex to reproduce.

Parallel developments notably Evo2 [4] push DNA FMs to massive scale and longer context windows, confirming that larger and more diverse pretraining datasets can encode more signal. However, reliance on large models and hybrid systems makes lean deployments difficult.

Parameter-efficient tuning methods such as LoRA[5] enable adaptation of large models without full fine-tuning and have shown strong empirical performance in NLP and other domains. We evaluate whether LoRA can effectively adapt NT-500M to KEGG-derived DNA dataset for disease prediction task, and compare it to a simpler approach: extract frozen embeddings from NT and train a small MLP classifier. Our dataset choice, KEGG derived reasoning benchmark, was curated in the previous research work [3].

II. Related Work

Foundation models (FMs) for DNA have rapidly become central to computational genomics, as they allow machine learning systems to leverage pretraining on massive nucleotide corpora in a similar way that large language models exploit natural language. These models are typically based on transformer architectures or their derivatives, which can learn rich contextual embeddings of DNA sequences. Once trained, these embeddings transfer to a wide range of downstream tasks such as variant effect prediction, promoter identification, enhancer classification, gene regulatory modeling, and chromatin accessibility estimation.

One of the initial works includes DNABERT [6], which adapts the BERT model to genomics by tokenizing DNA into fixed-length k-mers (typically 3–6 base pairs). DNABERT demonstrated that pretraining with masked language modeling on large-scale genomic data yields embeddings that generalize well across species and tasks. Because it is based on the efficient BERT architecture, DNABERT has been widely adopted and extended, spawning variants such as DNABERT-2 [7], which improve training data diversity and architectural refinements. These models remain popular for labs with limited compute resources, as they strike a balance between performance and feasibility.

The NT (Nucleotide Transformer) family [2] is among the earliest and most influential examples. Pretrained on hundreds of billions of nucleotides, NT models demonstrated that transformers can capture genomic syntax and semantics in ways that generalize across diverse benchmarks. The NT-500M model, a mid-sized variant that has been pretrained on human DNA sequences, is widely adapted for downstream tasks given its size [3]. NT models provide embeddings that can be finetuned or directly applied to supervised learning tasks, showing clear improvements over traditional sequence-based methods.

Building on this line, the Evo family [4], [8], [9] of models has pushed the scale of DNA FMs dramatically. Evo2 [4] trained on trillions of nucleotides with context lengths far beyond standard transformers, exemplifies the gains from ultra-large-scale pretraining. Evo models are specifically designed to capture long-range dependencies in DNA, such as gene regulation mechanisms spanning thousands of base pairs. This enables better performance on tasks like enhancer–promoter interaction prediction. However, the enormous compute and memory requirements of Evo2 highlight the challenges of deploying very large FMs in practical settings. Other variants in the Evo family explore trade-offs between model size, training data volume, and context length, offering researchers different options depending on resource availability.

In parallel, researchers have proposed specialized architectures that adapt transformers to genomic data. DNAHyena [10], for example, introduces a hyena-based operator that allows efficient modeling of very long DNA sequences with reduced memory overhead compared to vanilla transformers. DNAHyena retains the representational capacity of attention while scaling more favorably to genomic contexts that can reach hundreds of thousands of tokens. This makes it especially promising for tasks requiring long-range reasoning without incurring the prohibitive costs of Evo-scale transformers. Beyond these families, additional foundation models have emerged. For example, Enformer [11] extends transformers with convolutional layers to explicitly capture both local motifs and long-range dependencies, achieving state-of-the-art results in gene expression prediction. Similarly, GenSLMs [12] adapt large-scale sequence-to-sequence models to genomic prediction tasks.

At the same time, another emerging direction has focused on combining DNA FMs with language-based reasoning systems, reflecting the broader shift toward hybrid AI approaches. For instance, BioReason [3] integrates DNA FMs and large language models (LLMs), showing significant improvements on KEGG-derived reasoning benchmarks. By jointly processing sequence embeddings and textual prompts inside an LLM, BioReason demonstrates how biological sequence understanding can be enriched with textual reasoning. However, this integration increases system complexity and inference cost, motivating exploration of simpler and cheaper alternatives for specific downstream applications.

In parallel, researchers have placed considerable attention on fine-tuning strategies that make large foundation models more adaptable to practical tasks. Instead of retraining entire models from scratch, which is often computationally infeasible, modern approaches like LoRA [5] and LoRA+ [13] focus on adjusting only a small portion of the model's parameters or layering lightweight components on top of pretrained embeddings. These strategies allow large models to be tailored to specific downstream applications without incurring prohibitive computational costs.

A related and widely adopted practice is linear probing, which involves using frozen embeddings from pretrained models as fixed feature representations and training simple classifiers on top. In this setup, the foundation model serves purely as a feature extractor, while lightweight models such as multilayer perceptrons or logistic regression are optimized to perform the downstream task. Despite their simplicity, linear probing approaches often achieve competitive performance, demonstrating that pretrained embeddings already capture rich and meaningful patterns. This highlights the effectiveness of pretraining in encoding information that can be leveraged across diverse genomic tasks without the need for full model fine-tuning

III. Methodology

This section describes the dataset used, the experimental setup, and the two approaches we employed to improve the performance of the NT-500M model: parameter efficient fine-tuning using Low-Rank Adaptation (LoRA) and embedding-based downstream classification. We also detail the preprocessing steps, model configurations, and evaluation Metrics.

A. Dataset

We use a KEGG-derived biological reasoning dataset, originally introduced in the BioReason study [3]. The Kyoto Encyclopedia of Genes and Genomes (KEGG) [14] is a biological database that links genomic sequences with biological functions and disease associations. The curated dataset consists of question–answer pairs, where each question relates to a genetic variant or a biological pathway, and the answer corresponds to the associated disease. Each instance includes both a reference sequence (representing the canonical DNA sequence) and a variant sequence (containing single nucleotide variants or non-SNV mutations). The dataset comprises 1,159 training instances, 144 validation instances, and 144 test instances, spanning a total of 37 classes.

B. Model Setup

Our experiments focus exclusively on the nucleotide-transformer-500m-human-ref (NT-500M), a 500-million parameter transformer pre-trained on large-scale human genomic datasets. The NT-500M model outputs dense nucleotide embeddings that capture both local k-mer information and long-range dependencies. To evaluate its downstream performance, we compare two distinct strategies: (i) fine-tuning NT-500M with LoRA, and (ii) freezing NT-500M and training a lightweight classifier on its embeddings.

C. LoRA Fine-Tuning

We first explore parameter-efficient fine-tuning using Low-Rank Adaptation (LoRA). LoRA modifies the attention layers of a transformer by introducing low-rank decomposition matrices into the query and value projections. Instead of updating the full parameter set of NT-500M, LoRA trains only these additional low-rank matrices, significantly reducing memory and compute requirements. For our experiments, LoRA is applied to NT-500M under three input settings:

- Variant fine-tuning: The model is fine-tuned using only the variant DNA sequences.
- Reference fine-tuning: The model is fine-tuned using only the canonical reference sequences.

Fine-tuning is performed with cross-entropy loss, optimized using AdamW with a learning rate of 5e-5 and weight decay of 0.01. Early stopping is employed to prevent overfitting. This setup evaluates whether LoRA can adapt NT-500M embeddings to better discriminate disease classes directly from raw nucleotide sequences.

D. Embed ding-Based Classification

While LoRA adapts the transformer parameters, our second approach investigates whether frozen embeddings from NT-500M already contain sufficient information for classification. In this setup, NT-500M is kept frozen, and we extract embeddings for reference, variant, and concatenated sequences. The extracted embeddings are then fed into a multi-layer perceptron (MLP) classifier. The MLP consists of two fully connected layers with ReLU activations, batch normalization, and dropout for regularization. The final output layer applies a softmax function to predict disease classes. We evaluate three input configurations:

- Reference embeddings only
- Variant embeddings only
- Concatenated reference + variant embeddings

This design allows us to assess whether integrating both reference and variant contexts provides complementary signals for disease prediction. Training is conducted with a learning rate of 1e-4 and batch size of 32, using cross-entropy loss.

E. Evaluation Metrics

We adopt four widely used metrics to evaluate performance: accuracy, precision, recall, and F1-score. Accuracy provides an overall measure of correct predictions, while precision and recall evaluates the trade-off between false positives and false negatives. F1-score, as the harmonic mean of precision and recall, offers a balanced assessment of classification quality. To ensure reproducibility, each experiment is run three times with different random seeds, and we report the average scores.

IV. Experiments

For downstream disease classification, we compared two adaptation strategies of NT-500M: parameter-efficient fine-tuning using LoRA and linear probing with shallow classifiers. In the LoRA setup, NT-500M was wrapped with 'PeftModelForSequenceClassification', inserting low-rank adapters into the attention layers. Training was performed using the AdamW optimizer with weight decay 0.01, up to 10 epochs with early stopping based on validation F1-score.

In the linear probing setup, NT-500M was kept frozen, and embeddings were extracted and fed into a lightweight MLP classifier consisting of a single hidden layer with ReLU activation and dropout. The classifier was optimized with Adam (learning rate 1e-3) for up to 30 epochs, with dropout and batch normalization for regularization. This setup corresponds to linear probing, where the FM serves purely as a feature extractor while the shallow classifier adapts to the target task.

TABLE I
Comparison of LoRA Fine-Tuning and Embedding-Based Classification Using NT-500M on KEGG-Derived Dataset

Approach	**Accuracy**	**Precision**	**Recall**	**F1-score**
LoRA, variant sequence	0.867	0.668	0.670	0.665
LoRA, reference sequence	0.859	0.576	0.631	0.589
LoRA, reference + variant (1000)	0.867	0.626	0.652	0.635
Embeddings + MLP (ref + var, 1000)	**0.918**	**0.794**	**0.807**	**0.787**
Embeddings + MLP (ref only, 1000)	0.890	0.729	0.760	0.734
Embeddings + MLP (var only, 1000)	0.904	0.756	0.770	0.750

The baselines included LoRA fine-tuning applied to two different input configurations: variant sequence and reference sequence. These were compared against three embedding based setups, in which embeddings were extracted from NT-500M and fed into a multi-layer perceptron classifier. The embedding-based setups consisted of reference-only embeddings, variant-only embeddings, and concatenated reference + variant embeddings.

The results across all configurations are summarized in Table 1. LoRA fine-tuning on variant sequences achieved an accuracy of 86.72%, precision of 66.80%, recall of 66.98%, and F1-score of 66.51%. Reference-only LoRA fine-tuning achieved slightly lower performance with an accuracy of 85.94% and an F1-score of 58.88%. In contrast, the embedding-based MLP classifiers substantially outperformed all LoRA configurations. When concatenating reference and variant embeddings, the classifier achieved the highest accuracy of 91.78% and an F1-score of 78.68%, while reference-only and variant-only embeddings yielded 89.04% and 90.41% accuracy with F1-scores of 73.41% and 74.99%, respectively.

The LoRA fine-tuning experiments highlight that adapting NT-500M parameters directly provides only modest gains, with precision and recall remaining relatively low. Variant-only fine-tuning slightly outperforms reference-only fine-tuning, suggesting that variant information carries more discriminative signals for disease prediction. By contrast, embedding-based classification demonstrates the advantage of leveraging pre-trained NT-500M representations directly. The frozen embeddings already encode rich nucleotide features, allowing a small MLP classifier to effectively differentiate between disease classes. Concatenating reference and variant embeddings provides complementary information: the reference sequence captures the canonical genetic context, while the variant highlights the alterations. The improved performance with concatenation underscores the value of integrating both sequence types for variant effect prediction.

Analysis of misclassified examples reveals that LoRA fine-tuning often struggles with cases involving multiple variants associated with similar phenotypes, leading to ambiguous predictions. The embedding-based MLP handles such cases more effectively, likely because it can directly compare the reference and variant embeddings in a joint space. Furthermore, LoRA models tended to overfit to more frequent disease classes, whereas the embedding classifier maintained more balanced precision and recall across both common and rare classes, demonstrating its robustness to uneven label distributions. Overall, these experiments confirm that frozen embeddings with a lightweight MLP classifier outperform LoRA finetuning in both accuracy and F1-score. They also reveal that the concatenation of reference and variant embeddings provides the most discriminative representation, emphasizing the importance of combining canonical and altered sequence information for reasoning-style genomic tasks. These findings suggest that embedding-based downstream training offers a scalable and effective strategy for adapting DNA foundation models like NT-500M to variant–disease prediction tasks without incurring the computational cost of fine-tuning large transformer weights.

V. Discussion

The experimental results provide several insights into the capabilities of the NT-500M model for reasoning-based genomic tasks. Overall, the experiments show that parameter-efficient fine-tuning with LoRA yielded only modest improvements, whereas embedding-based downstream classification achieved substantially better performance. This contrast highlights a key property of large DNA foundation models: they encode rich, general-purpose representations that can be effectively leveraged without extensively modifying internal parameters. The relatively limited gains from LoRA fine-tuning are partly attributable to the input constraints of the transformer architecture. Since NT-500M accepts only a single DNA sequence at a time, it was not possible to jointly encode both reference and variant sequences during fine-tuning. As a result, the model could not directly exploit the complex interactions between canonical genetic context and sequence alterations. In contrast, the embedding-based approach allowed flexible downstream architectures to process both sequence representations together, enabling better capture of discriminative features. Embedding-based classification results demonstrate that the pretrained representations of NT-500M already contain biologically meaningful features capable of supporting high-accuracy predictions. By freezing the model and training a relatively lightweight MLP on top of the embeddings, the model is able to specialize for the disease classification task without the computational overhead or risk of overfitting associated with fine-tuning the full transformer. Concatenating reference and variant embeddings leads to the highest performance, which indicates that these two sources of information provide complementary perspectives. The reference sequence contextualizes the canonical structure and organization of the gene or genomic region, while the variant sequence highlights deviations that may directly contribute to disease phenotypes. The combination allows the classifier to reason over both the baseline genetic context and the perturbations introduced by variants.

Analysis of misclassified cases provides additional insights. Fine-tuned LoRA models tend to struggle in instances where multiple variants lead to similar phenotypic outcomes, suggesting that they lack the granularity to differentiate subtle effects. In contrast, the embedding-based MLP handles such cases more effectively, likely because it can simultaneously compare reference and variant embeddings in a shared representation space, allowing more nuanced distinctions. Furthermore, the LoRA fine-tuned models exhibit bias towards frequently occurring classes, whereas embedding-based classification maintains more balanced performance across both common and rare diseases, indicating better generalization to diverse genomic contexts.

These observations have broader implications for the use of DNA foundation models in reasoning and variant effect prediction. Firstly, they suggest that leveraging pretrained embeddings directly may often be preferable to fine-tuning, especially when computational resources are limited or when the dataset is relatively small. Secondly, the strong performance of concatenated embeddings highlights the value of integrating multiple sources of sequence information to capture complementary biological signals. Finally, the success of the embedding-based MLP approach points to a scalable strategy for adapting large DNA models to specialized tasks, enabling researchers to exploit the rich information encoded in foundation models without incurring the significant computational costs of full model adaptation.

Overall, the findings indicate that embedding-driven strategies not only improve predictive performance but also provide a more robust and interpretable framework for reasoning over genomic sequences. By maintaining the integrity of the pretrained representations and combining multiple sequence perspectives, researchers can develop models that better reflect the complex interplay between reference and variant DNA sequences, thereby advancing the capabilities of computational genomics and supporting applications in disease prediction and biological reasoning.

VI. Future Works

While this study demonstrates the effectiveness of embedding-based classification using NT-500M, several directions remain to further improve performance and applicability. One promising avenue is integrating multiple DNA foundation models, which could provide complementary embeddings and enhance robustness, particularly for complex variants or rare diseases. Exploring more sophisticated downstream architectures, such as attention-based classifiers or graph neural networks, may also better capture relationships between reference and variant sequences.

Incorporating multi-modal biological information, such as gene expression, proteomics, or epigenetic data, could further improve predictive accuracy and functional understanding of variants. Enhancing interpretability remains important; future work could focus on attention visualization or embedding space analysis to reveal which sequence features drive predictions. Finally, scaling these methods to larger datasets and clinical applications will require optimizing computational efficiency, enabling genome-wide or population-scale analyses.

VII. Conclusion

We performed an extensive comparison of two practical adaptation strategies for NT-500M on a KEGG-derived biological reasoning dataset: parameter-efficient LoRA fine-tuning versus frozen embeddings with an MLP classifier. Across metrics and ablations, the linear probing approach achieved better performance (best accuracy 91.78%, F1 78.68%) and greater training stability. These results suggest that, for classification tasks with moderate labeled data, frozen DNA FM embeddings are a highly effective and resource-efficient choice. The findings also provide practical guidance for deploying DNA foundation models in settings where compute is limited.

References

[1] R. Schmirler, M. Heinzinger, and B. Rost, “Fine-tuning protein language models boosts predictions across diverse tasks,” Nat. Commun., vol. 15, no. 1, p. 7407, Aug. 2024, doi: 10.1038/s41467-024-51844-2.

[2] 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, doi: 10.1038/s41592-024-02523-z.

[3] A. Fallahpour et al., “BioReason: Incentivizing Multimodal Biological Reasoning within a DNA-LLM Model,” May 29, 2025, arXiv: arXiv:2505.23579. doi: 10.48550/arXiv.2505.23579.

[4] G. Brixi et al., “Genome modeling and design across all domains of life with Evo 2,” Feb. 21, 2025, Genomics. doi: 10.1101/2025.02.18.638918.

[5] E. J. Hu et al., “LoRA: Low-Rank Adaptation of Large Language Models,” Oct. 16, 2021, arXiv: arXiv:2106.09685. doi: 10.48550/arXiv.2106.09685.

[6] Y. Ji, Z. Zhou, H. Liu, and R. V. Davuluri, “DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome,” Bioinformatics, vol. 37, no. 15, pp. 2112–2120, Aug. 2021, doi: 10.1093/bioinformatics/btab083.

[7] Z. Zhou, Y. Ji, W. Li, P. Dutta, R. Davuluri, and H. Liu, “DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome,” Mar. 18, 2024, arXiv: arXiv:2306.15006. doi: 10.48550/arXiv.2306.15006.

[8] A. T. Merchant, S. H. King, E. Nguyen, and B. L. Hie, “Semantic mining of functional de novo genes from a genomic language model,” Dec. 18, 2024, Synthetic Biology. doi: 10.1101/2024.12.17.628962

[9] E. Nguyen et al., “Sequence modeling and design from molecular to genome scale with Evo”.

[10] E. Nguyen, M. Poli, and M. Faizi, “HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution”.

[11] Ž. Avsec et al., “Effective gene expression prediction from sequence by integrating long-range interactions,” Nat. Methods, vol. 18, no. 10, pp. 1196–1203, Oct. 2021, doi: 10.1038/s41592-021-01252-x.

[12] M. Zvyagin et al., “GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics,” 2022

[13] S. Hayou, N. Ghosh, and B. Yu, “LoRA+: Efficient Low Rank Adaptation of Large Models,” July 04, 2024, arXiv: arXiv:2402.12354. doi: 10.48550/arXiv.2402.12354

[14]] M. Kanehisa and S. Goto, “KEGG: Kyoto Encyclopedia of Genes and Genomes”.