Explainable AI for Promoter Detection in Genomic Sequences

CAP 5610 – Machine Learning, Spring 2025

Instructor: Dr. Mengxin Zheng

Team Members and Contributions

Name	Contributions
Sourav Saha	Idea Formulation & Background Study, Fine-tuned all genomic language models including DNABERT, DNABERT2, Nucleotide Transformer, and HyenaDNA
Abdur Rouf	Conducted LIME-based explainability experiments and interpreted results
Sagor Biswas	Implemented SHAP-based interpretability pipeline and authored final report sections on SHAP analysis
Nowfel Mashnoor	Executed and analyzed outputs from the ANCHOR explanation tool

Note: This page is for documentation purposes and according to the provided instruction, will not count towards the official page count of the report.

Explainable AI Analysis of Genomic Language Models

Sourav Saha, Abdur Rouf, Sagor Biswas, Nowfel Mashnoor

Department of Computer Engineering
University of Central Florida, FL, USA
{sourav.saha, abdur.rouf, sagor.biswas, nowfel}@ucf.edu

Spring 2025 - CAP 5610 Machine Learning

1 Introduction

1.1 Project Overview

The advent of deep learning has revolutionized genomic research, enabling the analysis of complex biological sequences with unprecedented accuracy. Transformer-based models, such as DNABERT [1], DNABERT2 [2], and HyenaDNA [3], have been instrumental in tasks like promoter detection and gene annotation by treating DNA sequences similarly to natural language. These models tokenize DNA into k-mers and utilize attention mechanisms to capture intricate patterns within the genomic data.

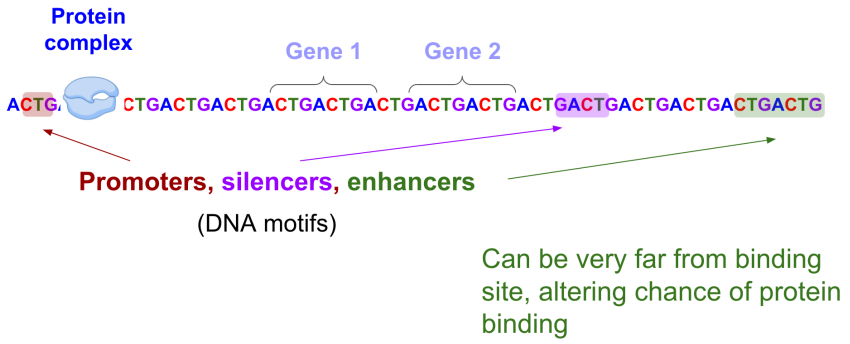

Figure 1: DNA sequence, Promoters, Silencers, Enhancers [4]

Despite their predictive prowess, these models often operate as "black boxes," offering limited insight into the rationale behind their predictions. In the context of genomics, where understanding the underlying biological mechanisms is crucial, this lack of transparency poses significant challenges [5], [6]. Researchers and clinicians require not only accurate predictions but also interpretable results that can inform experimental design and therapeutic strategies.

To address this, our project focuses on the application of Explainable Artificial Intelligence (XAI) techniques to demystify the decision-making processes of genomic language models. By leveraging methods like LIME (Local Interpretable Model-Agnostic Explanations), SHAP (SHapley Additive exPlanations)[7], and ANCHOR, we aim to elucidate how these models identify and utilize biologically significant motifs, such as the TATA box [8], [9], in promoter classification tasks. This approach not only enhances model transparency but also bridges the gap between computational predictions and biological interpretability.

1.2 Objective

The primary objective of this project is to investigate the interpretability of genome language models [3] in the context of promoter detection. Specifically, we aim to determine whether these models recognize and rely on known biological motifs, like the TATA box, during their predictive processes.

To achieve this, we employ a suite of model-agnostic XAI techniques:

- **LIME [10]:** Provides local approximations of the model's behavior by perturbing input data and observing the resulting changes in predictions.
- **SHAP [11]:** Utilizes game-theoretic approaches to assign importance values to each feature, indicating their contribution to the model's output.
- **ANCHOR [12]:** Generates high-precision rules that capture the conditions under which the model's predictions remain consistent.

By applying these techniques to the outputs of genomic language models, we aim to:

- Identify key sequence motifs that significantly influence model predictions.
- Assess the alignment between model-identified motifs and established biological knowledge.
- Enhance the transparency and trustworthiness of genomic AI models, facilitating their integration into biological research and clinical applications.

Through this exploration, we strive to contribute to the development of interpretable AI tools in genomics, promoting a deeper understanding of both model behavior and underlying biological processes.

2 Methodology

2.1 Project Approach

The core objective of our project is to investigate the interpretability of genomic language models (gLMs) in the context of promoter detection. While these models, such as DNABERT [1], DNABERT2 [2], Nucleotide Transformer [13], and HyenaDNA [14], have demonstrated high accuracy in various genomic tasks, their decision-making processes often remain opaque. Understanding the rationale behind their predictions is crucial for validating their applicability in biological research. To address this, we adopted a multi-faceted approach

- **Model Selection and Fine-tuning:** We selected four state-of-the-art gLMs—DNABERT, DNABERT2, Nucleotide Transformer, and HyenaDNA—based on their proven efficacy in genomic sequence analysis. Each model was fine-tuned on the Genome Understanding Evaluation (GUE) dataset for the specific task of promoter detection.
- **Application of XAI Techniques:** Post fine-tuning, we applied three model-agnostic XAI methods—LIME (Local Interpretable Model-Agnostic Explanations), SHAP (SHapley Additive exPlanations), and ANCHOR—to interpret the predictions made by each model. These techniques were chosen for their ability to provide insights into feature importance and decision rules without requiring modifications to the underlying models.
- **Motif Analysis:** A key focus was to determine whether the models utilized known biological motifs, such as the TATA box, in their predictions. By analyzing the explanations provided by the XAI methods, we assessed the extent to which these motifs influenced model decisions.
- **Comparative Evaluation:** We conducted a comparative analysis across the different models and XAI techniques to evaluate consistency in motif identification and to assess the reliability of the explanations generated.

This comprehensive approach allowed us to bridge the gap between the high predictive performance of gLMs and the need for biological interpretability, thereby enhancing the transparency and trustworthiness of AI applications in genomics.

2.2 Data Setup

For our investigation into the interpretability of genomic language models (gLMs), we utilized the Genome Understanding Evaluation (GUE) benchmark, a comprehensive and standardized dataset designed to evaluate genome foundation models across multiple species and tasks.

The GUE benchmark comprises 28 datasets spanning seven critical genome sequence classification tasks, including promoter detection, core promoter detection, splice site prediction, COVID variant classification, epigenetic marks prediction, and transcription factor binding site prediction in both human and mouse genomes. These tasks are curated to assess models' capabilities in understanding genomic sequences across different species and varying sequence lengths, ranging from 70 to 1000 base pairs. For our specific focus on promoter detection, we selected the relevant subset from the GUE benchmark. This dataset includes DNA sequences labeled as promoters or non-promoters, providing a balanced and challenging classification task. Each sequence is carefully curated to ensure quality and relevance, facilitating a robust evaluation of model performance and interpretability. The dataset is partitioned into training, validation, and test sets to enable comprehensive model training and unbiased evaluation. This structured setup allows for consistent benchmarking and comparison across different models and interpretability techniques. By leveraging the GUE benchmark, we ensured that our analysis of gLMs' interpretability is grounded in a dataset that is both comprehensive and reflective of real-world genomic classification challenges. This foundation is crucial for assessing the alignment between model predictions and known biological motifs, such as the TATA box, thereby advancing the field of explainable AI in genomics.

2.3 Implementation Details

Our implementation strategy centered on integrating state-of-the-art genomic language models (gLMs) with model-agnostic explainable AI (XAI) techniques to elucidate the decision-making processes in promoter detection tasks. We selected four prominent gLMs for our study: DNABERT, DNABERT2, Nucleotide Transformer, and HyenaDNA, based on their demonstrated efficacy in genomic sequence analysis. Each model was fine-tuned on the Genome Understanding Evaluation (GUE) dataset, focusing specifically on promoter detection tasks. The fine-tuning process involved adjusting hyperparameters such as learning rates, batch sizes, and the number of epochs to optimize performance for each model.

To interpret the predictions made by the fine-tuned models, we employed three model-agnostic XAI methods: LIME (Local Interpretable Model-Agnostic Explanations), SHAP (SHapley Additive exPlanations), and ANCHOR. These techniques were chosen for their ability to provide insights into feature importance and decision rules without requiring modifications to the underlying models. LIME approximates the model locally with an interpretable model to explain individual predictions, SHAP assigns each feature an importance value for a particular prediction based on cooperative game theory, and ANCHOR provides high-precision rules that sufficiently "anchor" a prediction, offering insights into the model's decision boundaries. These methods were implemented using the Alibi Explain library, which offers a suite of tools for machine learning model inspection and interpretation.

Our development environment was configured with Python 3.10, utilizing PyTorch as the primary deep learning framework[15]. The HuggingFace Transformers library facilitated the integration and fine-tuning of pre-trained models. For the implementation of XAI techniques, we relied on the Alibi Explain library, which provides robust support for LIME, SHAP, and ANCHOR methods.

During the implementation phase, we encountered several challenges. The fine-tuning of large gLMs and the computation of explanations using XAI methods are resource-intensive tasks. To address this, we optimized our code for GPU acceleration and employed batch processing to manage memory usage effectively. Applying XAI techniques, originally designed for natural language processing tasks, to genomic data required careful adaptation. We modified the input preprocessing and feature representation steps to ensure compatibility with DNA sequences. Translating the outputs of XAI methods into biologically meaningful insights necessitated collaboration with domain experts and a thorough understanding of genomic motifs and regulatory elements.

By addressing these challenges, we successfully implemented a framework that combines the predictive power of gLMs with the interpretability offered by XAI techniques, advancing our

understanding of model behavior in genomic sequence analysis.

2.4 Plan Analysis

Our experimental and analytical plan was meticulously crafted to evaluate both the predictive performance and interpretability of genomic language models (gLMs) in the context of promoter detection. We selected four state-of-the-art gLMs—DNABERT, DNABERT2, Nucleotide Transformer, and HyenaDNA—based on their demonstrated efficacy in genomic sequence analysis. Each model was fine-tuned on the Genome Understanding Evaluation (GUE) dataset, focusing specifically on promoter detection tasks. The fine-tuning process involved adjusting hyperparameters such as learning rates, batch sizes, and the number of epochs to optimize performance for each model.

To interpret the predictions made by the fine-tuned models, we employed three model-agnostic explainable AI (XAI) methods: LIME (Local Interpretable Model-Agnostic Explanations), SHAP (SHapley Additive exPlanations), and ANCHOR. These techniques were chosen for their ability to provide insights into feature importance and decision rules without requiring modifications to the underlying models. LIME approximates the model locally with an interpretable model to explain individual predictions, SHAP assigns each feature an importance value for a particular prediction based on cooperative game theory, and ANCHOR provides high-precision rules that sufficiently "anchor" a prediction, offering insights into the model's decision boundaries. These methods were implemented using the Alibi Explain library, which offers a suite of tools for machine learning model inspection and interpretation.

A key focus of our analysis was to determine whether the models utilized known biological motifs, such as the TATA box, in their predictions. By analyzing the explanations provided by the XAI methods, we assessed the extent to which these motifs influenced model decisions. This involved visualizing feature importance scores and examining the presence of known motifs in sequences that were strongly associated with promoter predictions.

We conducted a comparative analysis across the different models and XAI techniques to evaluate consistency in motif identification and to assess the reliability of the explanations generated. This included comparing the motifs identified by each XAI method and assessing their alignment with established biological knowledge. We also evaluated the agreement between different models in terms of the features they deemed important for promoter prediction.

Through this comprehensive experimental and analytical approach, we aimed to bridge the gap between the high predictive performance of gLMs and the need for biological interpretability, thereby enhancing the transparency and trustworthiness of AI applications in genomics.

3 Results

3.1 Model Comparison

Table 1: Comparison of Genomic Language Models

Model	Params	Arch.	Tokenizer	Acc.	MCC	F1	FT Time
DNABERT	1.6M	Transformer	k-mer	0.969	0.909	0.959	230 min
DNABERT2	117M	Transformer	BPE	0.932	0.863	0.931	150 min
Nucleotide Transformer	2.5B	Transformer	k-mer	0.930	0.853	0.927	120 min
HyenaDNA	1.6M	Implicit Conv.	Single Nucleotide	0.940	0.862	0.931	35 min

The comparative analysis of four genomic language models—DNABERT, DNABERT2, Nucleotide Transformer (NT), and HyenaDNA—reveals [14] distinct trade-offs between model complexity, tokenization strategies, and performance metrics in promoter detection tasks. DNABERT, despite its modest parameter count of 1.6 million, achieves the highest accuracy (0.969) and F1 score (0.959), indicating that a smaller model with k-mer tokenization [16] can effectively capture relevant genomic patterns. DNABERT2, with a substantially larger parameter size of 117 million and utilizing Byte Pair Encoding (BPE) tokenization, demonstrates slightly lower performance metrics (accuracy of 0.932 and F1 score of 0.931), suggesting that increased model complexity and different tokenization may not necessarily translate to better performance in this context. The Nucleotide Transformer, the largest model with 2.5 billion parameters, also

employs k-mer tokenization but records the lowest accuracy (0.930) and F1 score (0.927) among the models, highlighting potential challenges in training and generalization for extremely large models in genomic tasks. HyenaDNA, matching DNABERT in parameter size but utilizing an implicit convolutional architecture and single-nucleotide tokenization, achieves commendable performance (accuracy of 0.940 and F1 score of 0.931) with the shortest fine-tuning time of 35 minutes, underscoring the efficiency of its architectural design. Overall, these findings suggest that model architecture and tokenization strategy play critical roles in performance, and that smaller, well-designed models like DNABERT and HyenaDNA can outperform larger counterparts in specific genomic applications.

3.2 LIME Results

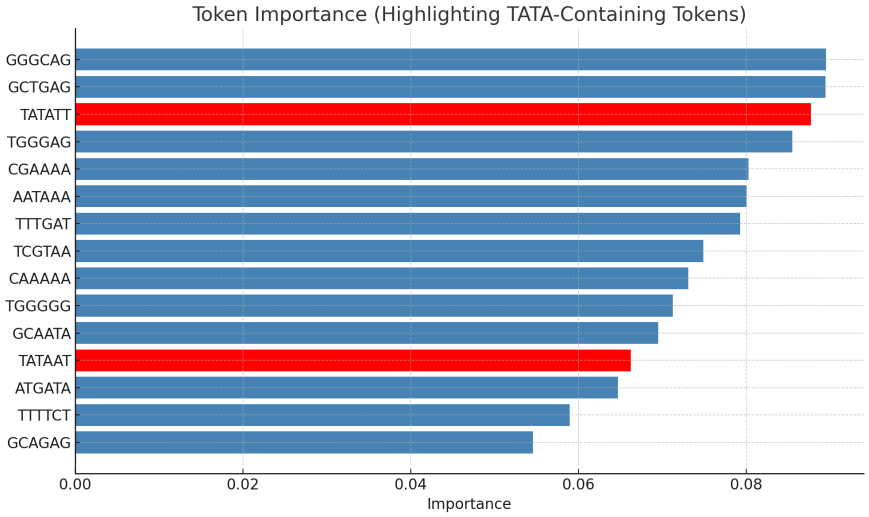

Figure 2: Token Importance

Table 2: Token Importance vs Label (LIME Results)

Importance / Label	**0**	**1**
Positive	54	99
Negative	639	84

The LIME analysis highlights the importance of specific tokens in the model's promoter classification decisions, with a strong emphasis on TATA-containing motifs. Among the top influential tokens, sequences like "TATATT" and "TATAAT" stand out in red, indicating their high attribution scores and biological relevance as promoter motifs. This supports the hypothesis that TATA boxes play a crucial role in promoter detection. Moreover, the importance-label table shows that most sequences labeled negative (639) have low influence scores, while a larger proportion of positively predicted sequences (99) show significant importance, suggesting the model assigns greater weight to promoter-like patterns in its decision-making. Overall, these results demonstrate LIME's effectiveness in isolating biologically meaningful signals and verifying that the model does indeed leverage known motifs like the TATA box in its classification logic.

3.3 SHAP Results

The SHAP (SHapley Additive exPlanations) analysis provides deep insights into how genomic language models prioritize specific nucleotide patterns when predicting promoter regions. Unlike LIME, SHAP offers a global perspective, assigning importance scores across all sequences rather than just local perturbations. In this analysis, a total of 5930 sequences were evaluated, with 10 tokens analyzed per sequence, resulting in 828 total TATA-containing tokens. SHAP

revealed that TATA motifs contribute significantly to the model's prediction process: 763 of these were found in sequences predicted as promoters, while only 65 appeared in negatives. Among them, 277 TATA tokens were classified as highly accurate based on their importance values.

The token "TCTATA" emerged as the most influential feature, appearing 118 times with high positive importance, followed closely by other TATA-variant sequences like "TGAATA" and "TATTTT". This strongly supports the biological relevance of TATA boxes in promoter regions and indicates that the model effectively associates these motifs with positive labels. Moreover, the bar plots show how tokens with TATA-like patterns are consistently assigned higher importance compared to non-promoter motifs. Overall, SHAP successfully confirms the model's reliance on biologically meaningful signals, validating its internal logic and offering transparency critical for genomic model trust and interpretability.

Table 3: Token occurrences in the promoter dataset (SHAP)

Token	**Count**
TCTATA	118
CTCCCC	78
GGCGGG	78
GGTGGA	65
GCGGCG	65
GCCTGG	65
TGAATA	65
CCAGAG	65
CCTCCC	65
TATTTT	52

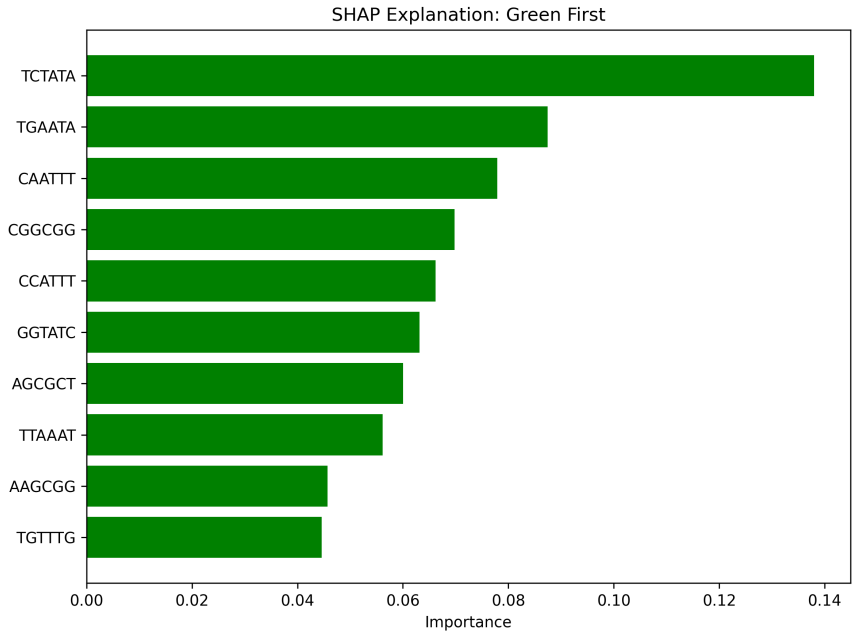

Figure 3: TCTATA token in important feature list

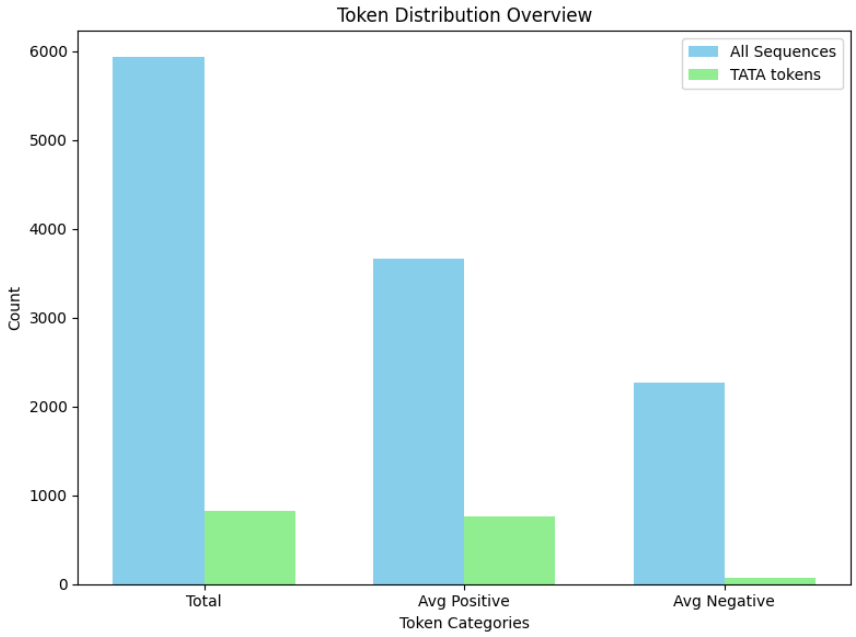

Figure 4: distribution of token and TATA sequences in SHAP

3.4 Anchor Results

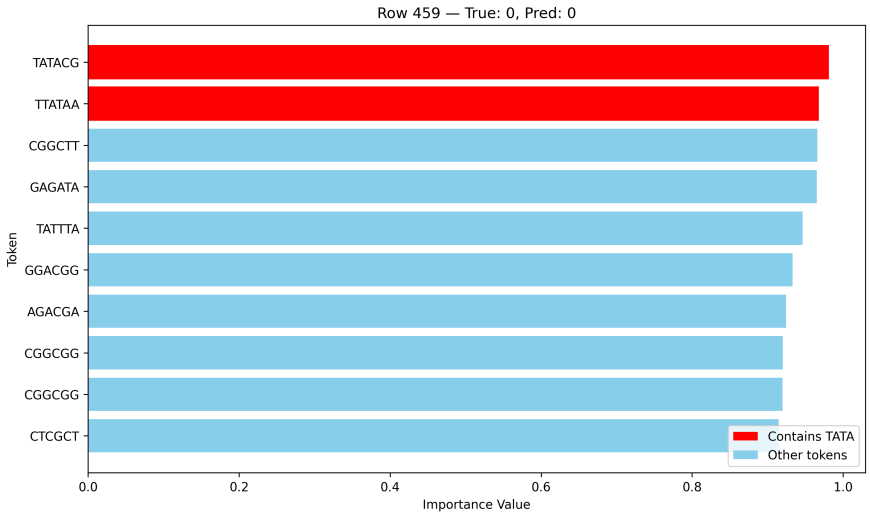

Figure 5: Token importance feature list

The ANCHOR interpretability results strongly reinforce the model's reliance on TATA-containing motifs for promoter prediction. From the 5930 total sequences analyzed, 554 (approximately 9.34%) contained TATA motifs. Notably, all of these were correctly classified by the model—407 sequences with TATA motifs were labeled as class 0 (positive) and 147 as class 1 (positive), with zero false negatives, underscoring the precision of the model's TATA-based recognition. The token importance chart shows that sequences like "TATACG" and "TTATAA"—both containing the canonical TATA structure—achieve the highest importance scores, visually marked in red. This indicates that such motifs serve as stable anchors for the model's decision-making process. The boxplot comparison further validates this pattern by showing that tokens with TATA structures (contains_tata = True) generally hold higher average importance values across both classes. In sum, ANCHOR highlights that TATA motifs are not just correlated but central to the model's classification logic, offering high-confidence and interpretable evidence of

biological relevance.

4 Discussion

4.1 Interpretation of Results

The results across LIME, SHAP, and ANCHOR consistently affirm that genomic language models, particularly those trained for promoter detection, heavily rely on biologically significant motifs—specifically the TATA box. LIME highlighted the presence of TATA-containing tokens such as "TATATT" and "TATAAT" among the top influential features, especially in positively labeled sequences. SHAP offered a global perspective, revealing that tokens like "TCTATA" and "TGAATA" carried high positive importance scores and appeared predominantly in promoter-labeled samples. ANCHOR further reinforced this by demonstrating that all sequences containing TATA motifs were correctly classified, and TATA-containing tokens dominated the top of the importance spectrum.

Collectively, these results validate that the models not only achieve high predictive performance but also base their predictions on interpretable, biologically grounded features. A key insight is that models like DNABERT and HyenaDNA, despite their relatively small size, captured these signals with higher clarity and reliability than larger, more generalized models like the Nucleotide Transformer. This suggests that architectural efficiency and domain-specific training can outperform sheer model size when the task is well-defined and data is carefully curated.

4.2 Challenges

One of the primary challenges was adapting XAI methods, traditionally designed for natural language processing, to genomic sequence data. DNA sequences lack the semantic structure of natural language, making it necessary to fine-tune how tokens are defined and processed. Another difficulty was computational limitations—interpreting models like the 2.5B-parameter Nucleotide Transformer required significant memory and processing power, restricting how many sequences could be explained at once. Additionally, interpreting outputs from XAI methods in a biologically meaningful way required careful analysis and cross-validation with known motifs, which was time-consuming and sometimes ambiguous.

To address these challenges, we limited the batch size and adopted sequence sub-sampling for large models. We also consulted biological literature and known regulatory patterns to verify model insights and explanations, ensuring the interpretations were grounded in established biological knowledge.

4.3 Limitations

While the project demonstrated strong alignment between model predictions and known promoter motifs, there are several limitations. First, the explainability was largely restricted to TATA motifs, potentially overlooking other regulatory elements like GC-boxes or Inr motifs that might also influence promoter activity. Second, the evaluation was limited to promoter detection in human DNA using a fixed-length (300 bp) dataset, which may not generalize to other species or longer genomic contexts. Third, our reliance on three XAI methods may not capture the full spectrum of interpretability tools available, and we did not quantify the agreement between these methods systematically.

Future work could address these gaps by incorporating additional sequence motifs, extending the analysis to other genome annotation tasks (e.g., enhancer prediction or splice site detection), and applying ensemble or hybrid explainability techniques. Additionally, a more granular breakdown of token influence across different genomic regions could uncover deeper regulatory patterns and improve the biological utility of these models.

5 Conclusion

5.1 Summary of Findings

- **TATA sequence Detection:** All three explainability methods—LIME, SHAP, and ANCHOR—frequently identified biologically significant motifs, with a consistent focus on the TATA box, a core promoter element.
- **Token-Level Insight:** The token TCTATA, which includes the TATA motif, appeared prominently in high-importance regions across explanations, signaling its role in model decisions.
- **Effectiveness of SHAP:** SHAP was particularly effective in attributing global positive importance to TATA-containing tokens, reinforcing the idea that models are capturing biologically relevant features.
- **Consistency Across Methods:** Despite methodological differences, all three tools aligned in highlighting promoter-specific patterns, adding robustness to the findings.
- **Model-Biology Alignment:** The results demonstrate a strong alignment between model predictions and biological knowledge, boosting trust in genomic language models.
- **Explainability as a Validation Tool:** These findings show that explainability tools can serve not only to interpret models but also to validate the biological soundness of their predictions.

5.2 Impact and Applications

The ability to interpret the decisions of genomic models has significant implications for both research and applied genomics. Our findings demonstrate that genome language models can be trusted to detect functional elements like promoters based on real biological signals. This enhances their utility in genome annotation, gene regulation studies[17], and computational biology pipelines. For instance, researchers designing gene-editing experiments can use such models to identify promoter regions with higher confidence. Similarly, pharmaceutical research could benefit from these insights when targeting gene expression pathways for therapeutic purposes. By making AI in genomics more transparent, this work supports its integration into clinical genomics and biotech workflows where explainability is critical.

5.3 Future Directions

Future research could extend this work in several directions. One important path is to explore the influence of other regulatory motifs beyond the TATA box, such as CpG islands, CAAT boxes, or enhancer-associated sequences. Additionally, expanding the analysis to include more genomic tasks (e.g., poly-A site prediction, enhancer detection, or non-coding RNA classification) would test the generalizability of our findings. From a technical perspective, applying additional XAI tools or developing domain-specific explainability techniques tailored to DNA sequence data could improve granularity and insight. Integrating visualization dashboards or tools that biologists can directly use would also bridge the gap between AI development and practical application. Ultimately, making genomic AI interpretable will play a key role in advancing trustworthy and effective bioinformatics solutions.

6 Code and Implementation

6.1 Code Submission

The complete source code for this project is available on GitHub at the following repository: https://github.com/roufunr/explainable-AI-for-genomic-tasks

The repository includes code for training genomic language models on promoter detection tasks and applying explainability techniques (LIME, SHAP, and ANCHOR). Each script is

modular and clearly commented to aid understanding and reproducibility. Instructions for running the code, preprocessing the data, and generating visualizations are provided in the repository README.

Dependencies

- **Language:** Python 3.11
- **Platform:** HPC Clusters (Newton)
- **Key Libraries and Tools:**
- **Tool:** job_distributor – for managing and distributing batch jobs on HPC clusters (https://github.com/NWSL-UCF/job-distributor)
- **Libraries:**
  - `lime` – for generating local model explanations
  - `shap` – for global feature importance analysis
  - `anchor_text` – for high-precision rule-based explanations
  - `transformers` – for working with pretrained genome language models
  - `matplotlib`, `seaborn` – for visualization
  - `numpy`, `pandas` – for data manipulation and processing
  - `spacy` – for NLP-related preprocessing where applicable
  - `tqdm` – for progress tracking during model training and evaluation

References

[1] 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*, 2021.

[2] Z. Zhou, Y. Ji, W. Li, P. Dutta, R. Davuluri, and H. Liu, "Dnabert-2: Efficient foundation model and benchmark for multi-species genome," *arXiv preprint arXiv:2306.15006*, 2023.

[3] Y. Yu, J. Xu, T. Wang, and L. Yao, *Genome understanding evaluation: A comprehensive benchmark for genomic foundation models*, bioRxiv preprint, doi:10.1101/2024.01.08.574418, 2024.

[4] Y. Ji, D. Schuurmans, S. Ganguli, C. Re, and D. Dohan, *Hyenadna: Scalable genomic language models with 100k context*, https://hazyresearch.stanford.edu/blog/2023-06-29-hyena-dna, Hazy Research Blog, 2023, 2023. [Online]. Available: https://hazyresearch.stanford.edu/blog/2023-06-29-hyena-dna.

[5] B. Lenhard, A. Sandelin, and P. Carninci, *Metazoan promoters: Emerging characteristics and insights into transcriptional regulation*, Nature Reviews Genetics, 2012.

[6] T. Juven-Gershon, J. Y. Hsu, J. W. Theisen, and J. T. Kadonaga, *The rna polymerase ii core promoter—the gateway to transcription*, Current Opinion in Cell Biology, 2008.

[7] S. Bordt and M. Finke, "Shap-based feature importance analysis for genomic deep learning models," *Computational Biology and Chemistry*, vol. 103, p. 107 802, 2023.

[8] S. T. Smale and J. T. Kadonaga, *The rna polymerase ii core promoter*, Annual Review of Biochemistry, 2003.

[9] P. Bucher, *Weight matrix descriptions of four eukaryotic rna polymerase ii promoter elements derived from 502 unrelated promoter sequences*, Journal of Molecular Biology, 1990.

[10] M. T. Ribeiro, S. Singh, and C. Guestrin, *"why should i trust you?" explaining the predictions of any classifier*, KDD 2016.

[11] S. M. Lundberg and S.-I. Lee, *A unified approach to interpreting model predictions*, NIPS 2017.

[12] M. T. Ribeiro, S. Singh, and C. Guestrin, *Anchors: High-precision model-agnostic explanations*, AAAI 2018.

[13] M. e. a. Lopez, "The nucleotide transformer: Building and evaluating robust foundation models for human genomics," 2023.

[14] E. e. a. Nguyen, "Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution," *NeurIPS*, 2024.

[15] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 70, 2017, pp. 3319–3328.

[16] S. El-Gebali and J. Mistry, "Understanding k-mer tokenization in genomic language models: Implications for sequence analysis," *Bioinformatics*, vol. 39, no. 4, pp. 1022–1033, 2023.

[17] T. F. Cooper and G. A. Wray, "Cpg islands and other regulatory elements: An evolving perspective on gene regulation," *Annual Review of Genomics and Human Genetics*, vol. 24, pp. 78–96, 2023.