Transforming Genomic Interpretability: A DNABERT Case Study

Authors: Micaela E. Consens (1, 2), Nicolas Papernot (1, 2), Bo Wang (1, 2), Alan Moses (1)

(1) Department of Computer Science, University of Toronto, Toronto, Canada

(2) Vector Institute, Toronto, Canada

Keywords: Machine Learning, ICML

Abstract

Transformer models trained on genomic sequence data achieve strong predictive performance, but interpreting their learned representations remains difficult. We analyze DNABERT using Layer-wise Relevance Propagation (LRP), attention-based scores, and gradient-based methods on enhancer detection and non-TATA promoter classification tasks. Attribution quality is assessed via mutagenesis experiments that perturb positions ranked as important by each method. We find that the impact of targeted mutations varies substantially across tasks. In enhancer detection, mutating high-scoring positions leads to marked performance degradation, indicating reliance on localized sequence features. In contrast, for promoter classification against structured genomic background regions, model performance degrades more slowly, consistent with more distributed representations. These findings suggest that attribution-based evaluations are sensitive to dataset construction and task semantics, and that conclusions about interpretability methods should be conditioned on the structure of the underlying prediction problem.

1. Introduction

The multi-modal and complex nature of functional genomic datasets, as well as their expanding scale, are well suited for the application of deep learning tools (Zou et al., 2019). Deep neural networks, specifically convolutional neural networks (CNNs) (Zhou & Troyanskaya, 2015; Alipanahi et al., 2015; Kelley et al., 2015; Kelley et al., 2018; Quang & Xie, 2015; Wang et al., 2018; Zhou et al., 2018; Jia et al., 2021) and transformers (Ji et al., 2021; Benegas et al., 2022; Avsec et al., 2021), have been effective in predicting regulatory genomic annotations from biological sequences. Recently, transformer models like DNABERT have demonstrated superior prediction accuracy compared to their CNN counterparts (Ji et al., 2021; Avsec et al., 2021).

Despite their predictive power, the interpretability of these models is equally crucial, especially in clinical settings where understanding model failure points is essential (Eraslan et al., 2019; Novakovsky et al., 2022; Molnar et al., 2020). Although attention scores of transformers have been proposed as an interpretability solution (Clauwaert et al., 2021), limitations exist due to the reduction of model information captured and the varied relevance of different attention heads in each layer (Serrano & Smith, 2019; Chefer et al., 2020).

This paper addresses this interpretability challenge by applying Layer-wise Relevance Propagation (LRP) (Chefer et al., 2020), a technique based on the Deep Taylor Decomposition principle (Montavon et al., 2017), to DNABERT (Ji et al., 2021). To our knowledge, this is the first application of an interpretability mechanism beyond simple attention maps to transformers for biological sequences. We fine-tune DNABERT for two classification tasks (non-TATA promoters and identifying human enhancers), apply several LRP-based methods for transformer interpretability, and compare these methods to the attention-score-based method proposed by the DNABERT authors along with other interpretability approaches such as GradCAM and rollout.

2. Related Works

In recent years, deep learning methods proposed for predicting regulatory annotations in genomics from sequence have shifted from primarily CNN-based to transformer-based models (Benegas et al., 2022; Ji et al., 2021; Avsec et al., 2021). While previous papers have explored the role of transformers' interpretability in genomics (Clauwaert et al., 2021), they have not compared methods beyond attention scores or acknowledged the limitations of this method (Serrano & Smith, 2019).

2.1 Layer-Wise Relevance

Layer-wise Relevance Propagation (LRP) is a method which has been utilized to understand and interpret the decisions made by deep learning models. LRP works by attributing the contribution of each input feature to the final decision of the network. This is achieved by propagating the output prediction back to the input layer, thereby providing an indication of feature importance.

LRP is calculated as (Letzgus et al., 2021):

$$R_{p_0} = \sum_{j=0}^{Q} \frac{a_{p_0} w_{p_0 q_j}}{\sum_{i=0}^{P} a_{p_i} w_{p_i q_j}} R_{q_j}$$

Where p_i and q_j are the neurons of consecutive layers, a is the activation of the respective neuron, and w is the weight between two neurons. R_{p_0} is then

the ‘relevance’ received by neuron p_0, which is interpreted as the contribution of that neuron in its layer to the output prediction $f(x)$.

LRP has been widely used in interpreting CNN-based models, providing valuable insights into their decision-making process (Bach et al., 2015). LRP has also been applied to transformers, particularly to demonstrate multi-headed attention results in "redundant" heads (Voita et al., 2019). A recent paper has further adapted the LRP method for explaining transformer classification decisions (Chefer et al., 2020).

2.2 Layer-Wise Relevance for Transformers

LRP has two important features. Firstly, LRP satisfies the conservation rule (Montavon et al., 2017), meaning relevancies are preserved as they move through each layer of the model such that the total relevance at the output layer is equal to the total relevance of the input layer. Secondly, LRP assumes ReLU activations, i.e., that there are only non-negative feature maps (Bach et al., 2015). However, in transformers the conservation rule is challenged by skip connections and matrix multiplications (Chefer et al., 2020), and transformers use the GELU non-linearity (Hendrycks & Gimpel, 2016), which outputs both positive and negative values.

Throughout this paper we apply versions of the modified LRP method proposed by Chefer et al. (2020) which accounts for the conservation rule in skip connections and matrix multiplications as well as GELU non-linearities.

3. Datasets

The datasets for both tasks, identifying non-TATA promoters as well as identifying human enhancers were taken from a recent paper (Martinek et al., 2022), with code available here: https://github.com/ML-Bioinfo-CEITEC/genomic_benchmarks. The test datasets for each task consist of roughly 30% of all data points and are relatively balanced. We collected the data from HuggingFace (Wolf et al., 2019) (https://huggingface.co/katarinagresova) and formatted the sequences including kmer-izing for DNABERT.

3.1 Human Non-TATA Promoters

Promoters are a region of sequence of DNA that binds a protein initiating the gene transcription. They are usually located close (from -200 to 50bp) to the transcription splice site (TSS). The dataset taken from HuggingFace was adapted from this paper (Umarov & Solovyev, 2017).

3.2 Human Enhancers

This dataset originates from the Ensembl database (Howe et al., 2021), release 100, itself taken from the VISTA Enhancer Browser project (Visel et al., 2007). As this dataset had variable length sequences, and 'N' encoded nucleotides, part of the processing for using this dataset was removing all sequences less than 200bp in length, and removing sequences with 'N' nucleotide codes, before kmer-izing.

4. Methods

4.1 Fine-tuning DNABERT

We employed the pretrained DNABERT model for k-mer (k = 6) linked on the DNABERT Github. We fine-tuned the pretrained model for both tasks, identifying non-TATA promoters and identifying human enhancers, by using a maximum sequence length of 200. The hidden dropout probability was set to 0.1 to prevent overfitting, and the learning rate was set to $2e^{-4}$. Weight decay was set to 0.01. The models were fine-tuned for a total of 5 epochs. We employed a batch size of 48 for fine-tuning, leveraging the per-GPU setting to optimize memory usage and computational efficiency.

The warmup phase of the training, the period during which the learning rate gradually increases to its maximum value, constituted 10% of the total training duration. We employed four NVIDIA Tesla T4 GPUs for this task, and used an Intel(R) Xeon(R) Silver 4110 CPU operating at a base frequency of 2.10GHz for efficient handling of the non-GPU computations involved in the fine-tuning process.

4.2 Scoring Position Contributions

We computed interpretability scores of each k-mer (k=6) for 500 samples of each class from the test dataset for both tasks. These scores were calculated using several methods, including LRP, Gradient-based methods, and attention scores from the Transformer model. Each score represents the importance of each k-mer (and is then converted to positions in the original sequence for downstream analysis) according to the specific interpretability method.

The attention score is computed as $softmax(QK^T/ \sqrt{d_k})$ where Q, K are the query and key matrices, respectively, and d_k is the dimension of the query and key vectors. The attention score reported for each sequence is calculated as the sum of attention scores from the start to end tokens.

The LRP score computes the LRP equation for each layer in the model following Equation 1. The LRP score function computes LRP for each layer, and returns a 'rollout' of these relevance scores. The LRP_last score computes LRP only for the last layer, while the full_LRP score computes the full LRP for the model, where relevance scores are then summed, providing a single relevance score for each token in the input.

The rollout score computes the average of attention matrices from the start of the model to the end.

The GradCAM score computes the gradient of the output with respect to feature maps and then performs a weighted combination of these maps. If y_c denotes the output for class c, A^k denotes the k-th feature map (or attention map in this case), and dy_c/dA^k denotes the gradient of y_c with respect to A^k, the operation can be denoted as: $GradCAM = ReLU(sum((dy_c/dA^k) * A^k))$.

4.3 Mutagenesis Experiments

We ran mutagenesis experiments to determine how well the interpretability scores explained the predictions of the fine-tuned DNABERT models. The mutagenesis experiment targeted the most relevant base pairs identified by each score and mutagenized them. In this case, a steep decrease in the model's accuracy indicates the mutagenized positions are important to the classification task.

To run the mutagenesis experiments, we took the positions identified by each interpretability score and mutated them in the original sequence, taking a single nucleotide at a time (either 'A', 'T', 'G', or 'C') and returning a different nucleotide chosen randomly from the remaining three. We then kmer-ized the mutated sequence again, and sent it to the model for classification to evaluate its performance after mutagenesis. Note that a single base pair mutation in the original sequence can affect up to 6 k-mers in the k-merized sequence sent to the model.

5. Results

5.1 Model Performance

On the full test dataset, the model achieved 86% accuracy on the human enhancer task and 93% accuracy on the non-TATA promoter task after fine-tuning.
For each task, we sample 500 sequences per class and measure the drop in classification accuracy when mutating either the top 10% most important nucleotides ($n = 20$) or all nucleotides ($n = 200$), as identified by attribution methods.

Results in Table 1 show that the fine-tuned DNABERT model exhibits different learning behaviors across tasks. For the non-TATA promoter task, the model more strongly learned to identify negative sequences, random fragments of human genes located downstream of first exons, than non-TATA promoters themselves. This is evidenced by a large drop in accuracy for non-promoter classification when all nucleotides are mutated ($n = 200$), compared to minimal degradation at $n = 0$. In contrast, the accuracy drop for non-TATA promoter identification is more moderate, indicating weaker reliance on promoter-specific signals.

For the human enhancer task, the model learned features associated with the enhancer class. The accuracy drops substantially for enhancer classification as more nucleotides are mutated, while non-enhancer classification degrades less sharply. Negative examples in this task consist of randomly sampled, length-matched genomic regions from GRCh38 that do not overlap annotated enhancers.

Together, these results suggest two distinct learning strategies. When separating enhancers from random genomic backgrounds, the model primarily learns features of the positive class. When separating non-TATA promoters from structured gene-sampled regions, which may present a more structured negative pattern, the model instead relies more heavily on identifying negative-class structure.

Table 1: Accuracy on DNABERT fine-tuned models in classifying specific classes with 500 samples.

(Note: LRP indicates the results reported are the best of either LRP or LRP_last.)*

Task	**n = 0**	**n = 20 (LRP*)**	**n = 20 (Attention)**	**n = 200**
Non-TATA	0.878	0.784	0.788	0.678
Non-promoter	0.968	0.814	0.854	0.306
Enhancer	0.876	0.49	0.512	0.386
Non-enhancer	0.87	0.838	0.852	0.62

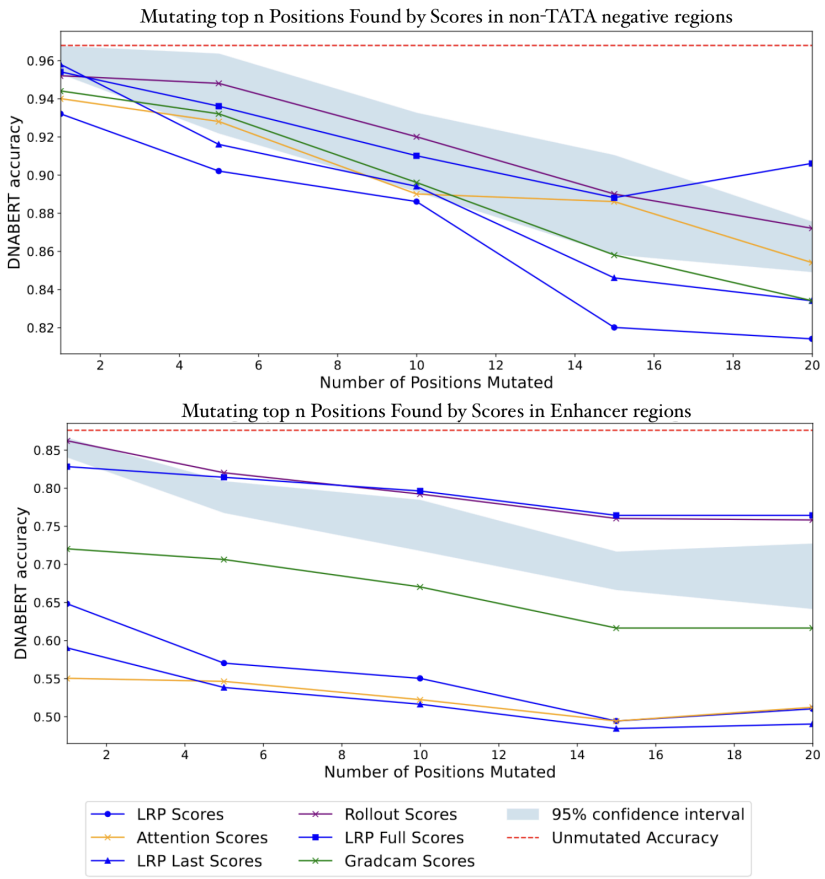
5.2 Mutagenesis Experiments

Across both tasks, mutagenesis experiments reveal that the impact of perturbing attribution-ranked positions depends strongly on the structure of the classification problem rather than on any single attribution method. For the enhancer task, mutating high-scoring positions identified by multiple methods leads to a noticeable reduction in model accuracy at n=20n = 20n=20, indicating that several attribution approaches capture enhancer-associated signal that is relevant to the model's predictions. However,

no method produces a sharp or exclusive degradation in performance under sparse mutation.

Figure 1 shows mutagenesis results for n=1,5,10,15,20n = 1, 5, 10, 15, 20n=1,5,10,15,20 across both tasks. In the non-TATA promoter task, where negative examples consist of structured gene-adjacent regions, model accuracy decreases gradually as more positions are mutated. For most attribution methods, performance remains within the 95% confidence interval of random mutagenesis over a wide range of nnn, suggesting that the learned representation is distributed across many input positions. Under these conditions, sparse mutagenesis provides limited separation between attribution methods.

Taken together, these results indicate that mutagenesis-based faithfulness tests are sensitive to task semantics and dataset construction. When predictive signal is localized, as in enhancer detection against random genomic background, targeted mutation can expose influential positions. In contrast, when models rely on distributed sequence features, attribution-guided sparse perturbations—regardless of the scoring method—produce only modest changes in performance. These findings motivate a more detailed analysis of how different interpretability signals, including attention, reflect learned representations under varying task conditions.



[Figure 1: Top graph shows the drop in accuracy of the non-TATA-DNABERT model pictured after mutating the k most important positions identified by each interpretability score. Bottom graph depicts the drop in accuracy of enhancer-DNABERT model pictured after mutating the k most important positions.]

6. Conclusion

To the best of our knowledge, this work represents an early comparison of interpretability methods for transformer models beyond simple attention mechanisms in the context of genomics. Through mutagenesis-based evaluation, we show that attribution faithfulness depends strongly on task design and dataset structure. In particular, sparse mutagenesis does not consistently recover random-performance baselines at small mutation budgets, suggesting that predictive information in genomic sequences may be distributed across many positions rather than concentrated in a small set of motifs.

These findings indicate that conclusions drawn from attribution-based evaluations should be conditioned on the underlying prediction task and the nature of the negative

class, rather than interpreted as evidence for or against a specific interpretability method. Further work on more diverse tasks, larger datasets, longer sequence contexts, and broader mutational regimes will be necessary to better understand when and how different interpretability signals reflect learned genomic representations.

We leave as future work a more systematic analysis of the biological relevance of motifs recovered by different attribution methods, as well as the extension of relevance propagation techniques to regression-based genomic models such as Enformer(Avsec et al., 2021). Recent work has proposed adaptations of LRP for regression settings, suggesting a path toward applying similar interpretability analyses to continuous genomic prediction tasks (Letzgus et al., 2021).

References

- Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8):831–838, 2015.
- Avsec, Ž., Agarwal, V., Visentin, D., et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18:1196–1203, 2021.
- Bach, S., Binder, A., Montavon, G., et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One*, 10(7):e0130140, 2015.
- Benegas, G., Batra, S. S., and Song, Y. S. DNA language models are powerful zero-shot predictors of non-coding variant effects. *Genome Biology*, 2022.
- Chefer, H., Gur, S., and Wolf, L. Transformer Interpretability Beyond Attention Visualization. *arXiv preprint arXiv:2012.09838*, 2020.
- Clauwaert, J., Menschaert, G., and Waegeman, W. Explainability in transformer models for functional genomics. *Briefings in Bioinformatics*, 22(5), 2021.
- Eraslan, G., Avsec, Ž., Gagneur, J., and Theis, F. J. Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*, 20(7):389–403, 2019.
- Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L., and Noble, W. S. Quantifying similarity between motifs. *Genome Biology*, 8:R24, 2007.
- Hendrycks, D., and Gimpel, K. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.
- Howe, K. L., Achuthan, P., Allen, J., et al. Ensembl 2021. *Nucleic Acids Research*, 49(D1):D884–D891, 2021.

- Ji, Y., Zhou, Z., Liu, H., and Davuluri, R. V. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.
- Jia, H., Park, S.-J., and Nakai, K. A semi-supervised deep learning approach for predicting the functional effects of genomic non-coding variations. *BMC Bioinformatics*, 22, 2021.
- Kelley, D., Snoek, J., and Rinn, J. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, 2015.
- Kelley, D. R., Reshef, Y. A., Bileschi, M., et al. Sequential regulatory activity prediction across chromosomes with Convolutional Neural Networks. *Genome Research*, 28(5):739–750, 2018.
- Letzgus, S., Wagner, P., Lederer, J., et al. Toward explainable AI for regression models. *arXiv preprint arXiv:2112.11407*, 2021.
- Martinek, V., Cechak, D., Gresova, K., Alexiou, P., and Simecek, P. Fine-Tuning Transformers For Genomic Tasks. *bioRxiv*, 2022.
- Molnar, C., Casalicchio, G., and Bischl, B. Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges. *ECML PKDD 2020 Workshops*, pp. 417–431, 2020.
- Montavon, G., Lapuschkin, S., Binder, A., Samek, W., and Müller, K.-R. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, 65:211–222, 2017.
- Novakovsky, G., Dexter, N., Libbrecht, M. W., Wasserman, W. W., and Mostafavi, S. Obtaining genetics insights from deep learning via Explainable Artificial Intelligence. *Nature Reviews Genetics*, 2022.
- Quang, D., and Xie, X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Research*, 2015.
- Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W., and Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research*, 32:D91–D94, 2004.
- Serrano, S., and Smith, N. A. Is Attention Interpretable? *arXiv preprint arXiv:1906.03731*, 2019.
- Umarov, R. K., and Solovyev, V. V. Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. *PLoS One*, 12(2):e0171410, 2017.
- Visel, A., Minovitsky, S., Dubchak, I., and Pennacchio, L. A. VISTA Enhancer Browser—a database of tissue-specific human enhancers. *Nucleic Acids Research*, 35:D88–D92, 2007.

- Voita, E., Talbot, D., Moiseev, F., Sennrich, R., and Titov, I. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*, 2019.
- Wang, M., Tai, C., E, W., and Wei, L. Define: Deep convolutional neural networks accurately quantify intensities of transcription factor-DNA binding and facilitate evaluation of functional non-coding variants. *Nucleic Acids Research*, 46(11), 2018.
- Wolf, T., Debut, L., Sanh, V., et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv preprint arXiv:1910.03771*, 2019.
- Zhou, J., and Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12(10):931–934, 2015.
- Zhou, J., Theesfeld, C. L., Yao, K., et al. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature Genetics*, 50(8):1171–1179, 2018.
- Zou, J., Huss, M., Abid, A., et al. A primer on deep learning in genomics. *Nature Genetics*, 51(1):12–18, 2019.

In practice, these results suggest that attribution faithfulness tests based on sparse mutagenesis can be strongly confounded by dataset construction, especially by the structure of the negative class, and may fail to distinguish interpretability signals even when models perform well. This motivates analyses that tie model internals to explicit biological annotations. In the next chapter, we develop a framework that relates attention patterns to curated genomic features across architectures and training stages, enabling more direct diagnosis of which biological structures are represented and used.