Transforming Genomic Interpretability: A DNABERT Case Study

Micaela E. Consens [1 2] Nicolas Papernot [1 2] Bo Wang [1 2] Alan Moses [1]

Abstract

While deep learning algorithms, particularly transformers, have recently shown significant promise in making predictions from biological sequences, their interpretability in the context of biology has not been deeply explored. This paper focuses on the recently proposed DNABERT model and explores interpreting it's decisions using modified Layer-wise Relevance Propagation (LRP) methods to determine what the model is learning. This score is then compared to several other interpretability methods commonly applied to transformers, including the attention-score based method proposed by the DNABERT authors. Results of mutagenesis experiments targeting regions identified by different methods show the modified LRP interpretability scores can outperform others at 20 mutations, and also show attention cannot reliably outperform random scores.

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; 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 finetune DNABERT for two classification tasks (identifying TATA and 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 Grad-CAM 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.

[1] Department of Computer Science, University of Toronto, Toronto, Canada [2] Vector Institute, Toronto, Canada. Correspondence to: Micaela E. Consens <micaela.consens@mail.utoronto.ca>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

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

$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_{0,j} a_j w_{jk}} R_k \tag{1}$$

Where j and k are the neurons of consecutive layers, a is the activation of the respective neuron, and w is the weight between two neurons. R_k is then the 'relevance' received by neuron k, which is interpreted as the contribution of that neuron in it's 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 TATA and 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 [1]. 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 $softmax(QK^T/\sqrt{d_k})$

[1] DNABERT Github available here: https://github.com/jerryji1993/DNABERT.

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.

4.4. Motif Analysis

We used the DNABERT's authors code for motif analysis in the non-TATA promoter task, finding contiguous high-scoring regions (specifically for attention, LRP_last and LRP scores), and filtered them by hypergeometric test. The motif instances for each of the three scores in the were aligned and merged to produce position-weight matrices (PWMs). The TOMTOM program (Gupta et al.) was applied to discover motifs compared with the JASPAR 2018 database (Sandelin et al., 2004) as in the DNABERT paper.

5. Results

5.1. Model Performance

On the full test dataset, the models performance for each of the tasks after fine-tuning was 86% accuracy for the human enhancer task, and 93% accuracy for the non-TATA promoter task. For each task we take 500 samples from each class and compare the drop in model accuracy when mutating the 10% most important nucleotides ($n = 20$) to mutating all of them ($n = 200$).

Results in Table 1 show that the fine-tuned DNABERT model for the non-TATA promoter task learned how to identify TATA sequences better than non-TATA (the model accuracy drops significantly on totally random sequences when $n = 200$ compared to $n = 0$ for TATA identification). The fine-tuned DNABERT model for the human enhancer task learned how to identify enhancer sequences better than non-enhancer sequences (the model accuracy drops significantly on totally random sequences when $n = 200$ compared to $n = 0$ for enhancer identification). Interestingly, this aligns with our intuition for what the model should learn: given positive examples of a specific class, and then negative examples that do not belong to the class (but may not follow a specific pattern beyond *not* being part of the class), the model learns only the positive class.

We subset the results for identifying what the model has learned using interpretability scores to the cases of identifying TATA sequences in the non-TATA promoter task, and identifying enhancer sequences in the human enhancer task. The other results are attached in the appendix, see Figure 2.

		$n = 20$		
Task	$n = 0$	LRP*	attention	$n = 200$
Non-TATA	0.878	0.784	0.788	0.678
TATA	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

Table 1. Accuracy on DNABERT fine-tuned models in classifying specific classes with 500 samples. Accuracy is reported on unmutated samples from the test dataset ($n = 0$), 500 sequences with the top 20 identified LRP-based score or attention score positions mutated ($n = 20$), or 500 samples of entirely randomly generated sequences ($n = 200$). Here LRP* indicates the results reported are the best of either LRP or LRP_last. If the interpretability score drops accuracy closer to the accuracy reported at $n = 200$, this indicates the score better explains model decision making.

5.2. Mutagenesis Experiments

The LRP and LRP_last scores outperformed attention and other interpretability scores in the mutagenesis experiment for each task at $n = 20$, as seen in Table 1. In all cases these scores were able to bring the model closer to its performance on random sequence data than attention scores alone.

In Figure 1 we can see the results of running mutagenesis experiments for each task with each score mutating $n = 1$, $n = 5$, $n = 10$, $n = 15$ and $n = 20$ positions. In the TATA region identification task, on 500 samples of TATA regions, we see the score that drops the accuracy most is the LRP score. This score consistently remains outside the 95% confidence interval for random mutations, which cannot be said for the others. In the human enhancer task we see the score that drops accuracy the most is the LRP_last score, with LRP and attention just behind.

These preliminary results suggest LRP-based scores can provide better explainability for transformer models even in the context of genomics, tested by mutagenesis rather than ablation. While these results are restricted to two specific tasks DNABERT was fine-tuned on, they demonstrate that attention scores alone are not necessarily capturing all the information the model is using to make decisions in tasks it has correctly learned.

5.3. Motifs

A brief exploration of the motifs in the non-TATA promoter identification task, specifically in identifying TATA promoters, revealed the top motif found by attention was other C4 zinc finger-type factors. The top motif found by LRP_last C2H2 zinc finger factors. LRP scores top-matching motif was homeo domain factors in the JASPAR 2018 database.

6. Conclusion

To the best of our understanding, this is the first paper comparing interpretability methods for transformers beyond simple attention mechanisms in the context of genomics. Preliminary results suggest there is reason to explore using LRP-based scores as a better way to visualize and understand how these models learn genomic data. Further work must be done on more diverse tasks with larger datasets, that contain longer sequences, and across a larger mutational burden window (beyond $n = 20$) to better understand which LRP-based scores are best suited for model interpretability, and to discover if attention scores truly cannot outperform random mutagenesis.

We note that mutating the top $n = 20$ positions of sequences in each task did not converge on random performance ($n = 200$), provide preliminary suggestions that information within genomic sequences is quite distributed after all.

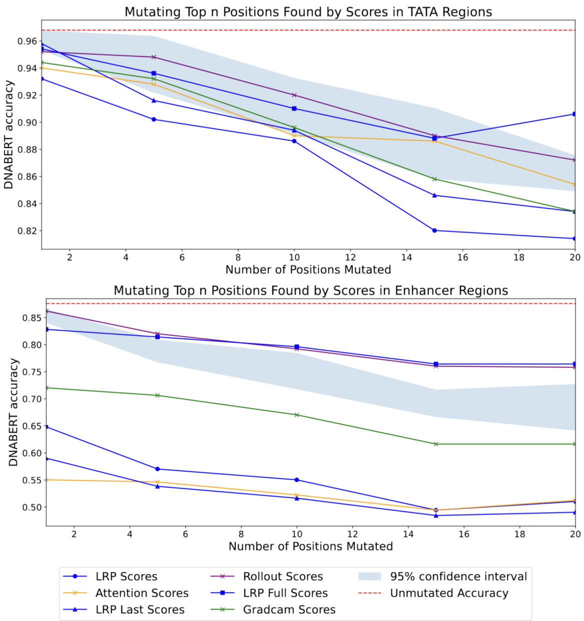

Figure 1. Top graph shows the drop in accuracy of the non-TATA-DNABERT model pictured after mutating the k most important positions identified by each interpretability score. Accuracy drop is the largest for LRP score indicating this score capture best what positions the enhancer-DNABERT model uses to make predictions. Gradcam scores and LRP_last scores do next best. Notably, attention scores alone do no better than random. Bottom graph depicts the drop in accuracy of enhancer-DNABERT model pictured after mutating the k most important positions identified by each interpretability score. Accuracy drop is the largest for LRP_last score, and LRP score and attention are right behind, indicating these scores capture best what positions the enhancer-DNABERT model uses to make predictions.

We leave as future work the analysis of the quality of motifs identified by each score in terms of genomic relevance, i.e. whether more motifs of biological relevance are identified by LRP based scores or other methods.

We acknowledge one of the limitations of applying this LRP mechanism to popular transformer models for genomics like Enformer (Avsec et al., 2021), is that the LRP method employed in this paper propagates relevancies based on a classification. However, recent work has gone into proposing LRP as an interpretability score for regression tasks as well, along with best practices for implementation (Letzgus et al., 2021). This leaves expanding the modified LRP score for transformers specifically for transformers on regession tasks, and applying them to models like Enformer to evaluate their genomic interpretability, for future work.

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. doi: 10.1038/nbt.3300.

Avsec, , Agarwal, V., Visentin, D., Ledsam, J., Grabska-Barwinska, A., Taylor, K., Assael, Y., Jumper, J., Kohli, P., and Kelley, D. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18:1196–1203, 10 2021. doi: 10.1038/s41592-021-01252-x.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. 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. Aug 2022. doi: 10.1101/2022.08.22.504706.

Chefer, H., Gur, S., and Wolf, L. Transformer interpretability beyond attention visualization, 2020. URL https://arxiv.org/abs/2012.09838.

Clauwaert, J., Menschaert, G., and Waegeman, W. Explainability in transformer models for functional genomics. *Briefings in Bioinformatics*, 22(5), 04 2021. ISSN 1477-4054. doi: 10.1093/bib/bbab060. URL https://doi.org/10.1093/bib/bbab060. bbab060.

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., Bailey, T., and Noble, W. Quantifying similarity between motifs (2007) genome biol, 8, pp. *R24*.

Hendrycks, D. and Gimpel, K. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, 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, 02 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab083. URL https://doi.org/10.1093/bioinformatics/btab083.

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, 06 2021. doi: 10.1186/s12859-021-03999-8.

Kelley, D., Snoek, J., and Rinn, J. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. pp. 028399, 10 2015. doi: 10.1101/028399.

Kelley, D. R., Reshef, Y. A., Bileschi, M., Belanger, D., McLean, C. Y., and Snoek, J. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Research*, 28(5):739–750, 2018. doi: 10.1101/gr.227819.117.

Letzgus, S., Wagner, P., Lederer, J., Samek, W., Müller, K.-R., and Montavon, G. 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*, pp. 2022–02, 2022.

Molnar, C., Casalicchio, G., and Bischl, B. Interpretable machine learning – a brief history, state-of-the-art and challenges. In *ECML PKDD 2020 Workshops*, pp. 417–431. Springer International Publishing, 2020. doi: 10.1007/978-3-030-65965-3_28. URL https://doi.org/10.1007%2F978-3-030-65965-3_28.

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, may 2017. doi: 10.1016/j.patcog.2016.11.008. URL https://doi.org/10.1016%2Fj.patcog.2016.11.008.

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. doi: 10.1038/s41576-022-00532-2.

Quang, D. and Xie, X. Danq: a hybrid convolutional and recurrent deep neural network for quantifying the function of dna sequences. pp. 032821, 12 2015. doi: 10.1101/032821.

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(suppl_1):D91–D94, 2004.

Serrano, S. and Smith, N. A. Is attention interpretable?, 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(suppl_1): 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. doi: 10.1093/nar/gky215.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771, 2019. URL http://arxiv.org/abs/1910.03771.

Zhou, J. and Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12(10):931–934, 2015. doi: 10.1038/nmeth.3547.

Zhou, J., Theesfeld, C. L., Yao, K., Chen, K. M., Wong, A. K., and Troyanskaya, O. G. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature Genetics*, 50(8):1171–1179, 2018. doi: 10.1038/s41588-018-0160-6.

Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., and Telenti, A. A primer on deep learning in genomics. *Nature genetics*, 51(1):12–18, 2019.

A. Appendix

Results for incorrectly learned regions:

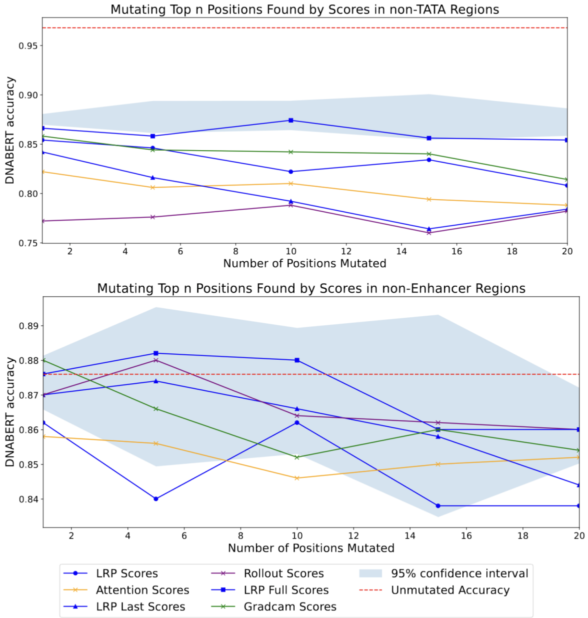

Figure 2. Top graph shows the drop in accuracy of the non-TATA-DNABERT pictured after mutating the k most important positions identified by each interpretability score. Bottom graph shows the drop in accuracy of the enhancer-DNABERT pictured after mutating the k most important positions identified by each interpretability score.