Mechanistic Interpretability of a Genomic Language Model for Splice- and Expression-QTL Sequences

Rana A. Barghout
University of Toronto
rana.barghout@mail.utoronto.ca

January 15, 2026

Abstract

We present a mechanistic interpretability analysis of DNA_bert_6, a 12-layer genomic language model [1], applied to sequences spanning splice-QTL (sQTL) and expression-QTL (eQTL) variants. Using attention analysis, activation patching, and sparse autoencoder (SAE) feature analysis, we probe how the pretrained encoder represents significant versus non-significant variants in its final embeddings. Attention statistics show structured heads in mid-to-late layers but only weak, noisy differences in CLS-to-variant attention between classes. Activation patching provides a causal test of information flow and reveals an extreme bottleneck at the final transformer layer, while position-wise patching indicates that variant-associated signal is encoded diffusely across the local context window rather than localized to the variant token. SAE features learned on layer-12 CLS activations are highly sparse with only modest class-dependent shifts and do not yield clear motif-like, disentangled directions. Overall, QTL-associated signal appears distributed with a strong late-layer bottleneck but no robust evidence for splice- or expression-specific sequence regions. All code and figures for this work can be found here: https://github.com/ranaabarghout/genomic-mechanistic-interpretability.

1 Introduction

Genomic language models such as DNABERT [1] achieve strong performance on diverse sequence-based tasks, yet their internal computation remains difficult to interpret. Understanding *how* these models represent and propagate variant information is important for assessing biological plausibility and guiding more interpretable architectures.

We ask: **What internal mechanisms does DNA_bert_6 use to encode the functional impact of genetic variants in QTL sequence contexts?** Specifically, we test whether the model develops biologically meaningful mechanisms—such as heads or latent features that focus on splice donors/acceptors (GT–AG), branch points, polypyrimidine tracts (sQTLs), or promoter/enhancer-proximal elements (eQTLs)—or whether variant information is represented in a diffuse and distributed way.

Initial experiments used DNABERT-2 [2] fine-tuned on sQTLs because it offers architectural and pretraining improvements over the original DNABERT and strong reported performance on the Genomic Understanding Evaluation (GUE) benchmark. However, unresolved weight-loading issues made it difficult to guarantee that probed checkpoints matched the fine-tuned model. We therefore pivoted to the original DNABERT with 6-mer tokenization (DNA_bert_6), a 12-layer BERT-base transformer pretrained as a masked language model on genomic 6-mers [1]. All analyses use the pretrained encoder without a supervised head.

2 Methods

2.1 Model and Data

Model: DNA_bert_6 [1].

Data: GTEx v8 splice-QTL (sQTL) and expression-QTL (eQTL) subsets from the GV-Rep framework [3], derived from the GTEx project [4, 5].

Each example is a 1024 bp reference sequence centered on the variant with a binary label.

2.2 Attention Analysis

We run DNA_bert_6 with `output_attentions=True`. For each attention head we compute entropy, variance, and CLS-to-variant attention, compared across classes using a two-sided t-test and Cohen's d.

2.3 Activation Patching

Activation patching provides a causal test of information flow. For matched significant/non-significant pairs we patch hidden states at layer l and measure the change in the layer-12 CLS embedding.

2.4 Sparse Autoencoder

We train an overcomplete ReLU SAE with 2048 units on layer-12 CLS embeddings. We compute mean activation, differential activation Δ, and sparsity. Distribution plots are saved in the repository as `2_feature_analysis.png` under `outputs/<dataset>_analysis/...`.

3 Results

3.1 Attention Patterns

Per-head statistics for eQTLs show that mid-to-late layers (7-10) have lower entropy, higher variance, and higher max attention than early or final layers, indicating more structured, peaked CLS attention in this band (Fig. 1). CLS-to-variant comparisons reveal only weak class dependence.

One apparent outlier, layer 7 head 5, shows highly significant CLS-to-variant differences, but its position-wise attention is dominated by the [CLS] and final tokens for both classes, making this statistic unreliable (Fig. 2). Most heads (e.g. L0H0) show broad, low-amplitude attention with overlapping class means.

Biological expectation. If the encoder had learned explicit regulatory detectors, we would expect attention heads that preferentially attend to canonical motifs—GT–AG splice junctions, branch points, or promoter/enhancer elements—and that show consistent class-conditional shifts. The absence of such specialization suggests that pretrained attention does not align with known regulatory grammars.

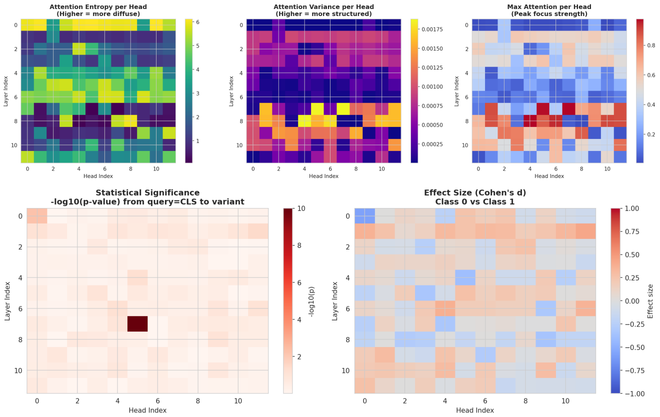

Figure 1: **Attention statistics for eQTL sequences.** *Top:* Entropy, variance, and maximum CLS attention per head and layer quantify how focused and structured each head is. Mid-to-late layers exhibit more peaked, organized attention. *Bottom:* $-\log_{10} p$ values and Cohen's d for CLS-to-variant attention comparing significant vs non-significant eQTLs show that class differences are weak and noisy across heads.

For sQTLs, mid-layer heads show similar structured attention with weak class dependence and no heads focusing on splice junctions or polypyrimidine tracts (Fig. 3).

3.2 Activation Patching

Layer-wise patching shows negligible effects for layers 0–11 and a large effect at layer 12, revealing an extreme final-layer bottleneck in the CLS embedding. Position-wise patching within layer 12 yields an approximately flat effect across a 20-token window centered on the variant, indicating that variant-associated signal is distributed across nearby context rather than localized.

3.3 Sparse Autoencoder Features

SAE features learned from layer-12 CLS embeddings exhibit tightly centered differential-activation distributions (Δ) with minimal separation between significant and non-significant variants, indicating that QTL-associated signal is captured only weakly and in a distributed manner rather than by discrete, motif-like latent units. In other words, although the SAE successfully decomposes the CLS embed-

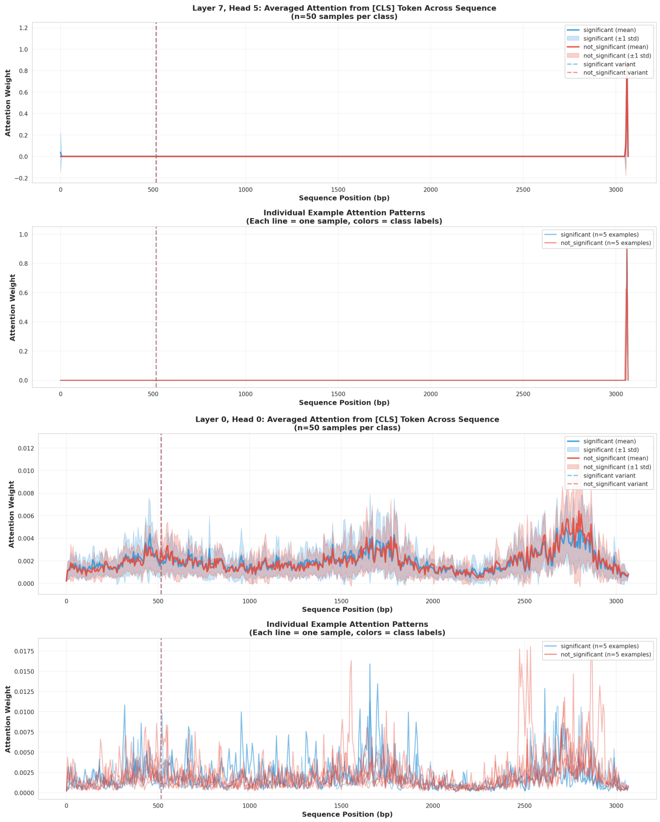

Figure 2: **Position-resolved CLS attention for two representative eQTL heads.** *Top:* Head L7H5 appears highly significant by scalar metrics but attends almost exclusively to [CLS] and the final token, indicating a degenerate pattern. *Bottom:* A typical head (L0H0) spreads attention broadly across the sequence with overlapping class means, consistent with diffuse encoding.

ding into sparse directions, most features are either inactive for the majority of variants or respond similarly across classes.

This is illustrated by the two most class-biased SAE units for eQTLs (Fig. 4). Features 1428 and 140 have the highest mean activation and the largest $|\Delta|$ among all units, yet their activation distributions for significant and non-significant variants remain highly overlapping. Even the strongest latent directions therefore fail to produce clear bimodality or sharp class separation, which would be expected if the SAE had isolated interpretable regulatory factors such as splice-site or promoter-like motifs.

Together, these results suggest that the pretrained DNA_bert_6 CLS space encodes QTL-associated information in weak, approximately linear directions that are superimposed across many features, rather than in a small number of disentangled, biologically meaningful components. This further supports the view that regulatory structure does not emerge cleanly in the absence of supervised fine-tuning on QTL labels.

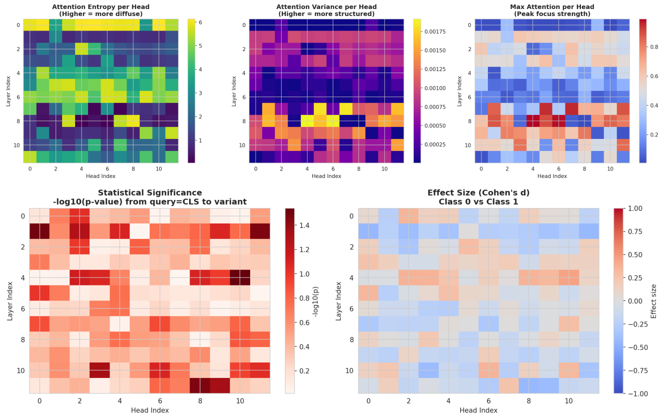

Figure 3: **Attention statistics for sQTL sequences.** Entropy, variance, and CLS-to-variant attention show the same mid-layer band of structured heads as in eQTLs, but with weak and noisy class-conditional effects and no heads specialized for splice-regulatory motifs.

4 Discussion

DNA_bert_6 encodes weak QTL-associated signal in a distributed manner with a strong late-layer bottleneck. While regulatory biology predicts specialization around splice and promoter elements, no robust head-, position-, or feature-level specialization is observed in the pretrained setting, consistent with objective mismatch.

Fine-tuning on QTL labels should induce sharper specialization, testable via repeated mechanistic analysis and experimental assays such as minigene splicing or reporter constructs.

Experimental and mechanistic validation. The mechanistic patterns observed here generate testable biological hypotheses. If the weak and distributed QTL-associated signal identified in the pretrained encoder reflects a lack of learned regulatory grammar, then fine-tuning on supervised sQTL and eQTL prediction should induce sharper specialization. This can be evaluated by repeating the same attention, patching, and SAE analyses after fine-tuning. In particular, we predict that a task-aligned

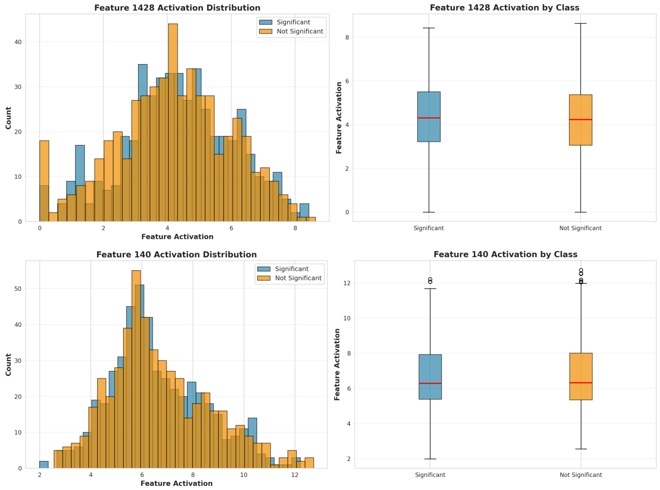

Figure 4: **Top sparse autoencoder features for eQTL CLS embeddings.** Features 1428 and 140 exhibit the largest mean activation and differential activation across classes, yet their distributions overlap substantially, indicating that even the most informative SAE units encode only weak class separation rather than discrete regulatory motifs.

model will exhibit (i) attention heads that preferentially focus on splice junctions, branch points, or polypyrimidine tracts for sQTLs, (ii) promoter- or enhancer-proximal focus for eQTLs, and (iii) SAE features with more bimodal class separation and recognizable sequence patterns among top-activating variants.

To connect model interpretations to real biology, variants identified as strongly influential by attribution methods (e.g., integrated gradients or patching sensitivity) could be tested in minigene splicing assays for sQTLs or reporter-gene assays for eQTLs. Targeted *in silico* mutagenesis of nucleotides highlighted by the model would further allow direct testing of whether predicted regulatory elements causally modulate splicing or expression, providing a bridge between model-based interpretability and experimental validation. For an experiment like this, an analysis like integrated gradients would be most appropriate, as it can directly identify basepairs and positions that influence output features, while inherently considering the effect of genetic variants in the computation. Previous work done [https://github.com/ranaabarghout/IGLOO] on this has revealed interesting patterns in protein variant function prediction and can be implemented for genomic language models.

5 Conclusion

In this work, we applied mechanistic interpretability tools to a pretrained genomic language model to probe how splice- and expression-QTL variants are represented in sequence embeddings. Across attention analysis, causal activation patching, and sparse autoencoder feature discovery, we find that DNA_bert_6 encodes only weak QTL-associated signal in a highly distributed manner, with a dominant final-layer bottleneck and little evidence of biologically aligned specialization for splicing or regulatory motifs.

A central implication of these findings is that pretraining alone is insufficient to induce interpretable regulatory structure in the model. Fine-tuning on supervised sQTL and eQTL prediction is likely necessary for biologically meaningful mechanisms to emerge. Repeating the same mechanistic analyses after fine-tuning provides a concrete and falsifiable path forward for testing whether regulatory grammar becomes localized in attention heads, internal representations, or sparse latent features.

Several assumptions and limitations underlie this analysis. Attention weights are an imperfect proxy for importance and can be misleading in degenerate heads. Activation patching measures changes in representation rather than downstream prediction, since no classification head is attached. SAE features are constrained by reconstruction and sparsity objectives rather than biological interpretability, and the CLS embedding may not preserve all local variant effects. Finally, QTL effects are often tissue- and context-specific, whereas the model operates on sequence alone.

Despite these limitations, this work demonstrates how mechanistic tools can be used to audit genomic foundation models and to generate biologically grounded hypotheses about how sequence variation is represented internally. Coupling such analyses with task-specific fine-tuning and experimental validation offers a promising route toward building interpretable and biologically faithful models of regulatory genomics.

References

[1] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers

model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.

[2] Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*, 2023.

[3] Zehui Li, Vallijah Subasri, Guy-Bart Stan, Yiren Zhao, and Bo Wang. Gv-rep: A large-scale dataset for genetic variant representation learning, 2024.

[4] Latarsha J Carithers and Helen M Moore. The genotype-tissue expression (gtex) project. *Biopreservation and Biobanking*, 13(5):307, 2015.

[5] GTEx Consortium. The gtex consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, 2020.