Mechanistic Interpretability of Nucleotide-Transformer-v2

Hugo Paulat

hugo.paulat@mail.mcgill.ca

CS Department, McGill University, Canada

Introduction

Large Language Models (LLMs), built on the Transformer architecture, have revolutionized Natural Language Processing. Motivated by these successes, researchers have adapted Transformer-based architectures to other forms of sequential data, including genomic sequences (Consens et al. 2025). Models such as DNABERT, EVO, and the Nucleotide Transformer are pretrained on large-scale genomic datasets and demonstrate strong performance across a variety of downstream tasks (Ghosh et al. 2025).

Despite their empirical success, these models remain largely opaque. In genomics, where predictions can influence biological interpretation and decision making, understanding how such models repesent and utilize sequence information is critical (Rudin 2019). As such, mechanistic and representational analyses are necessary to assess what information is captured across model layers and to what extent these representations reflect biologically meaningful signals.

In this study, I analyze Nucleotide Transformer v2, a 500M-parameter model developed by InstaDeep, NVIDIA and TUM, pretrained on over 850 genomes spanning both model and non-model organisms. Using variants from a verified clinical dataset, I embed reference and alternative sequences and probe representations across model layers to evaluate whether mutation-relevant features are linearly accessible within the model's internal representation. I further complement this analysis with an attention-based investigation, examining differences in attention patterns across layers to gain insight into how the model localizes and processes genetic variation.

Methodology

Data Processing

The initial analysis was performed using 151 clinician-verified variants. These were selected from the GV-Rep database, which provides curated records and an accompanying data loader that facilitates sequence extraction and preprocessing. Restricting the analysis to verified entries increases potential signal due to high-confidence annotations, and limits compute requirements.

All sequences were tokenized using the default tokenizer provided by Hugging Face. It operates on 6-mers, enabling the model to process token lengths corresponding to over 12 kbp of sequence. As a compromise between contextual coverage and computational feasibility, I extracted a 3000 bp window centered on each variant. This corresponds to approximately 500 tokens after 6-mer tokenization and remains well within the model's maximum context length, while allowing efficient batch-wise inference on available hardware. Reference and alternative sequences were processed identically to enable direct comparison of their internal representations.

Model Probing on Embeddings

To assess the downstream utility of the learned representations, I evaluated whether the model's embeddings enable discrimination between high-risk (*pathogenic*) and low-risk (*benign*) genetic variants. This analysis follows a probing approach, in which the transformer is treated as a fixed feature extractor and simple classifier heads are trained on its internal representations (Dalla-Torre et al. 2025). Successful performance under this setting indicates that information relevant to the functional impact of SNVs and small indels is linearly accessible within the model's embedding space.

Following the methodology of the original Nucleotide Transformer paper, I extracted hidden states from a representative subset of layers $\{1, 3, 5, 9, 12, 15, 18, 22, 25, 28\}$ spanning the 29-layer architecture. This provides coverage of early, intermediate, and late representations while remaining computationally efficient. To obtain a fixed-size representation for each sequence, token-level hidden states were aggregated via mean pooling across the sequence dimension.

I then probed these representations using a logistic regression classifier, chosen to explicitly test linear accessibility rather than downstream performance. Embeddings were standardized and projected into a lower-dimensional space using PCA ($d = 50$) prior to classification. Given the limited sample size, model performance was evaluated using five-fold stratified cross-validation. To isolate variant-specific effects from background genomic context, I additionally constructed difference embeddings between alternative and reference sequences,

$$\mathbf{v}_{\Delta} = \mathbf{v}_{\text{alt}} - \mathbf{v}_{\text{ref}}$$

This formulation ensures that the classifier probes the perturbation induced by the variant rather than learning locus- or gene-specific identity features (Gereben et al. 2025). As shown in Figure 1, probing performance consistently exceeds random chance across most layers, with some layers achieving ROC-AUC values above 0.70. This indicates that mutation-relevant information is linearly accessible within the model's internal representations.

Notably, performance is not monotonic across depth: strong probing results emerge in early layers, reappear in the middle, and persist at later layers. This suggests progressive refinement of variant-sensitive features rather than a single ”decision layer.” Such behavior is consistent with hierarchical encoding, where early layers capture local perturbations, intermediate layers integrate contextual signals, and later layers aggregate higher-level abstractions.

Based on this profile, I selected representative early, middle, and late layers for further mechanistic analysis. In particular, layers exhibiting both strong and weak probing performance were retained to assess whether differences in linear separability correspond to distinct attention patterns. This motivates the subsequent attention analysis.

Layer	ROC–AUC	Precision	Recall
1	0.676 ± 0.062	0.623 ± 0.064	0.717 ± 0.041
3	0.606 ± 0.053	0.527 ± 0.052	0.567 ± 0.111
5	0.580 ± 0.071	0.556 ± 0.049	0.550 ± 0.172
9	0.708 ± 0.054	0.667 ± 0.055	0.667 ± 0.149
12	0.684 ± 0.067	0.651 ± 0.130	0.667 ± 0.091
15	0.716 ± 0.097	0.679 ± 0.087	0.700 ± 0.041
18	0.681 ± 0.048	0.644 ± 0.080	0.600 ± 0.170
22	0.546 ± 0.077	0.501 ± 0.045	0.500 ± 0.075
25	0.530 ± 0.135	0.546 ± 0.115	0.533 ± 0.085
28	0.708 ± 0.090	0.670 ± 0.158	0.700 ± 0.041

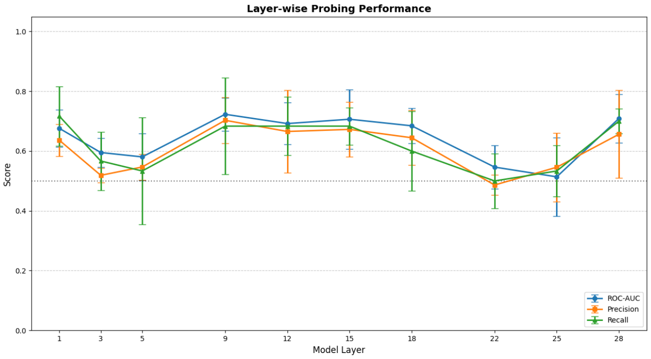

Figure 1: Layer-wise probing performance for ClinVar variant classification using $\mathbf{v}_\Delta$ embeddings. The table reports mean ± standard deviation across cross-validation folds; the plot visualizes these trends.

Attention Maps

To characterize the progression of representation learning across the model architecture, head-averaged self-attention maps were extracted from early (layer 1), intermediate (layers 9 and 18), and late (layer 28) stages. For each variant, paired reference (REF) and alternate (ALT) sequences were analyzed. The attention responses were compared using 2D attention difference maps, defined as $\Delta A = A_{\text{ALT}} - A_{\text{REF}}$, as well as 1D attention profiles extracted at the specific mutation token index to isolate local perturbations in weight distribution (Figure 2).

Across layers, self-attention exhibited a progressive structural evolution. In the early layers (1 and 9), attention was nearly uniform across the sequence, and ALT–REF differences were small and diffuse. Weak center-aligned patterns were occasionally observed, consistent with the construction of fixed-length windows centered on the variant and the effects of tokenization and padding. At this stage, attention appeared to primarily encode low-level sequence structure rather than variant-specific effects.

In intermediate layers (e.g., layer 18), attention became more structured, redistributing toward a limited number of tokens. This was reflected by the emergence of distinct vertical stripes in the attention difference maps, indicating that information was increasingly routed through a small set of “key” tokens. These patterns suggest a transition from broadly distributed attention to more selective information aggregation.

In later layers (22 and 28), attention frequently collapsed onto a single dominant hub at the mutation position. This behavior was consistent across variants and was evident both in the 2D difference maps and in the corresponding 1D attention profiles. Such convergence suggests that, at deeper layers, the model aggregates variant-local information by routing attention through the centered mutation token.

Taken together, these observations indicate that self-attention in the model progressively focuses on the mutation site as depth increases. Notably, this qualitative behavior was largely shared across variants, suggesting that attention primarily reflects where information is aggregated rather than encoding discriminative signals related to variant pathogenicity. Additionally, the strong late-layer concentration at the mutation site implies that, for single-nucleotide variants, very large contextual windows may be less critical, as attention ultimately converges on local sequence information.

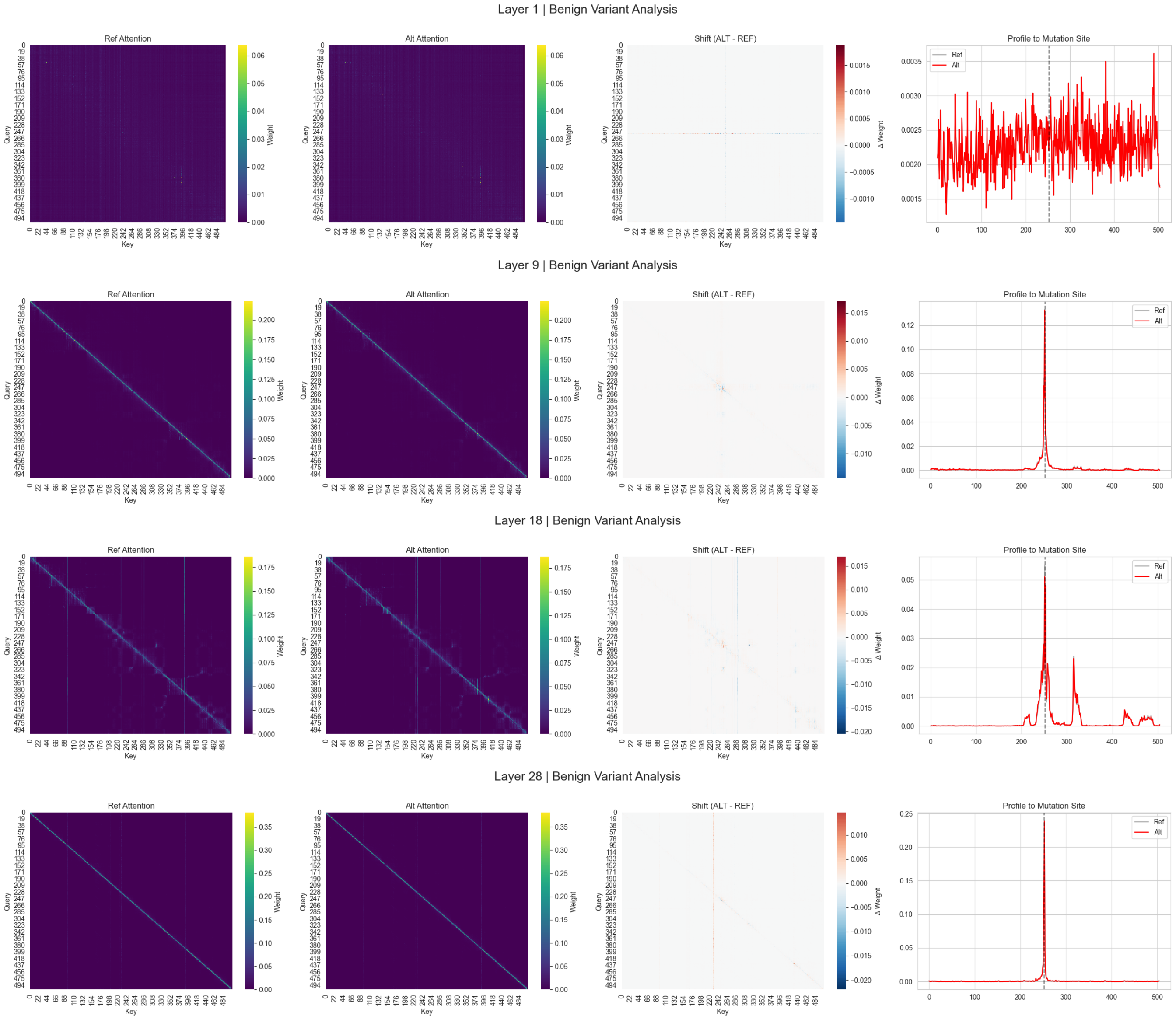

Figure 2: Impact of Benign Variant on Model Attentions. Comparative heatmaps show the attention distribution for reference and mutated sequences. The Shift (ALT-REF) map isolates the change in model focus, while the 1D Profile tracks the attention magnitude assigned to the mutation site across the entire sequence length.

Limitations

This study is subject to several limitations that constrain the interpretation and generalizability of the results.

First, the analysis is conducted on a relatively small set of clinician-verified variants. While restricting the dataset to high-confidence entries reduces label noise, it also limits statistical power and may obscure subtler representational differences, particularly in the attention-based analysis. As a result, the findings should be interpreted as exploratory rather than definitive.

Second, the probing experiments rely on linear classifiers applied to pooled hidden representations. Although this approach is standard for assessing linear accessibility of information, it does not capture nonlinear decision boundaries that may be exploited by downstream fine-tuned models. Consequently, the reported probing performance likely underestimates the full discriminative capacity of the model's representations. A future study could increase the dataset by pulling more entries from GV-Rep, with potential to train an MLP classification head.

Third, the attention analysis is primarily qualitative. While attention difference maps and mutation-centered profiles provide intuitive insight into how information is routed across layers, attention weights are not guaranteed to correspond to feature importance or causal influence. As such, observed attention patterns should be interpreted as reflecting information aggregation mechanisms rather than direct explanations of model predictions.

Fourth, the attention maps are head-averaged, which simplifies visualization but may obscure specialized behaviors of individual attention heads. Prior work has shown that distinct heads can encode different functional roles, and future analyses could benefit from head-specific attention analysis.

Finally, no explicit linkage was made between learned features or attention patterns and known genomic annotations such as regulatory elements, chromatin accessibility, transcription factor binding, or evolutionary conservation. As a result, it remains unclear whether the observed representations correspond to biologically interpretable signals or reflect abstract regularities learned during pretraining. Future work would benefit from integrating external annotations and perturbation-based analyses to assess whether attention redistribution and embedding separability align with known functional genomic phenomena.

References

Consens, Micaela E., Cameron Dufault, Michael Wainberg, Duncan Forster, Mehran Karimzadeh, Hani Goodarzi, Fabian J. Theis, Alan Moses, and Bo Wang. 2025. "Transformers and genome language models." *Nature Machine Intelligence* 7 (3): 346–362. https://doi.org/10.1038/s42256-025-01007-9.

Dalla-Torre, Hugo, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, et al. 2025. "Nucleotide Transformer: building and evaluating robust foundation models for human genomics." *Nature Methods* 22:287–297. https://doi.org/10.1038/s41592-024-02523-z. https://www.nature.com/articles/s41592-024-02523-z.

Gereben, Orsolya, Hedvig Tordai, Lana Khamisi, and Tamás Hegedűs. 2025. "pLM-SAV: A Δ-Embedding Approach for Predicting Pathogenic Single Amino Acid Variants." Preprint, *bioRxiv,* https://doi.org/10.1101/2025.05.24.655916. https://www.biorxiv.org/content/10.1101/2025.05.24.655916v1.

Ghosh, Nimisha, Daniele Santoni, Indrajit Saha, and Giovanni Felici. 2025. "A review on the applications of Transformer-based language models for nucleotide sequence analysis." *Computational and Structural Biotechnology Journal* 27:1244–1254. https://doi.org/10.1016/j.csbj.2025.03.024. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11984569/.

Rudin, Cynthia. 2019. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence* 1 (5): 206–215. https://doi.org/10.1038/s42256-019-0048-x.