Abstract

We investigated whether Nucleotide Transformer v2 (NTv2-500M) contains internal features informative for ClinVar pathogenicity. Using paired reference (REF) and alternate (ALT) sequence contexts (2048 bp; variant centered), we trained a lightweight MLP classifier head on frozen NTv2 embeddings and achieved 72–73% accuracy on a stratified train/validation split (noting strong class imbalance). To obtain token-level interpretability, we trained a sparse autoencoder (SAE) on token activations from NTv2's penultimate layer (4096 latent units) and ranked SAE features by their separation between benign and pathogenic groups (Cohen's d). One feature (Feature 1581) showed both strong group separation and localized REF–ALT activation differences, providing a concrete token-level handle for mechanistic follow-up.

Introduction

Genomic foundation models can predict functional impact from sequence context, but their internal decision pathways are often opaque. Mechanistic interpretability aims to localize which internal features and sequence positions most influence model representations and downstream predictions—especially important for variant interpretation, where subtle local changes can have large biological consequences.

Here, we study NTv2-500M as a general-purpose sequence encoder and ask: Do NTv2 representations contain signal that separates ClinVar benign vs pathogenic variants, and can we localize this signal to specific internal features and token positions? We use paired REF/ALT contexts centered on each variant, enabling allele-level comparisons. Our approach has two components: 1) a baseline classifier head trained on frozen NTv2 embeddings to confirm that NTv2 representations carry label-relevant information, and 2) a sparse autoencoder trained on token activations to produce a sparse feature basis that can be analyzed at token resolution. Our main outcome is not a state-of-the-art classifier, but an interpretable pipeline that identifies specific sparse features whose activations differ systematically across variant groups and localize allele-driven representational changes.

Methods

We used ClinVar variant labels with four classes: *Benign, Likely benign, Likely pathogenic, Pathogenic*. Each example includes paired REF and ALT sequence contexts. We fixed sequence length to 2048 bp with the variant centered to control for positional effects. The dataset contains 24,729 paired samples with strong imbalance (~72% likely benign). We used an 80/10/10 stratified split by class. We encoded REF and ALT

contexts with NTv2-500M and extracted token embeddings. For sequence-level tasks, we applied masked mean pooling over tokens to produce one vector per sequence.

To test whether NTv2 representations contain pathogenicity signal without modifying the backbone, we froze NTv2 and trained an MLP on pooled embeddings. The classifier input concatenated: [ref, alt, (ref - alt), |ref - alt|] and produced logits over the four ClinVar classes. We report overall accuracy on train and validation splits (with the caveat that imbalance can inflate accuracy).

For token-level interpretability, we trained an SAE on token activations from NTv2's 2nd to last layer. The SAE used an encoder: Linear(512 -> 4096) + ReLU, and a decoder: Linear(4096 -> 512), and optimized reconstruction loss with L1 sparsity regularization. We used a schedule that gradually increased the L1 coefficient to achieve sparse solutions while maintaining reconstruction quality.

For each SAE feature, we summarized per-sample activation by averaging across tokens, then ranked features by Cohen's d between (pathogenic + likely pathogenic) vs (benign + likely benign). For top features, we visualized token-level Δ activation = activation(ALT) - activation(REF) to localize allele-driven representational changes.

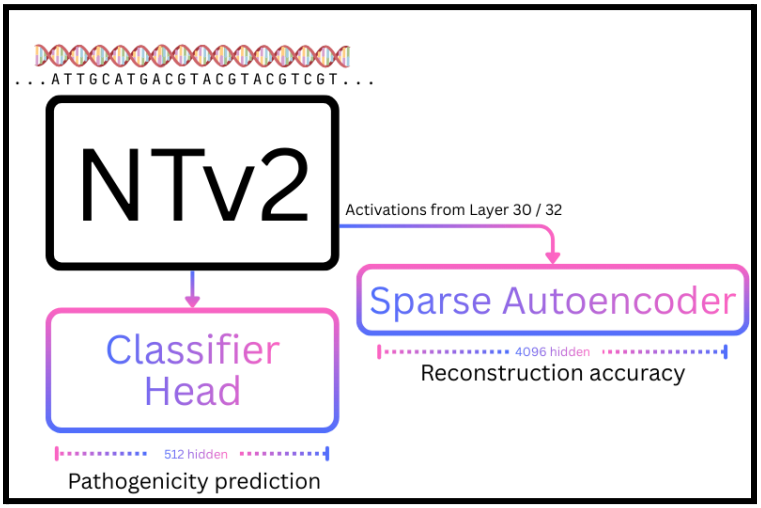

Figure 1. Pipeline diagram

Results

The classifier head trained on frozen NTv2 embeddings achieved 73% training accuracy and 72% validation accuracy. Given class imbalance, this should be interpreted as evidence that NTv2 embeddings contain label-relevant signal rather than as a definitive measure of clinical prediction performance. (Future reporting should include macro-F1 and per-class recall.)

The SAE trained on penultimate-layer token activations achieved reconstruction loss 12.87, with an average of ~491 active features per token and 0% dead units,

suggesting the model learned a distributed sparse basis over NTv2 activations rather than collapsing onto a small subset of features.

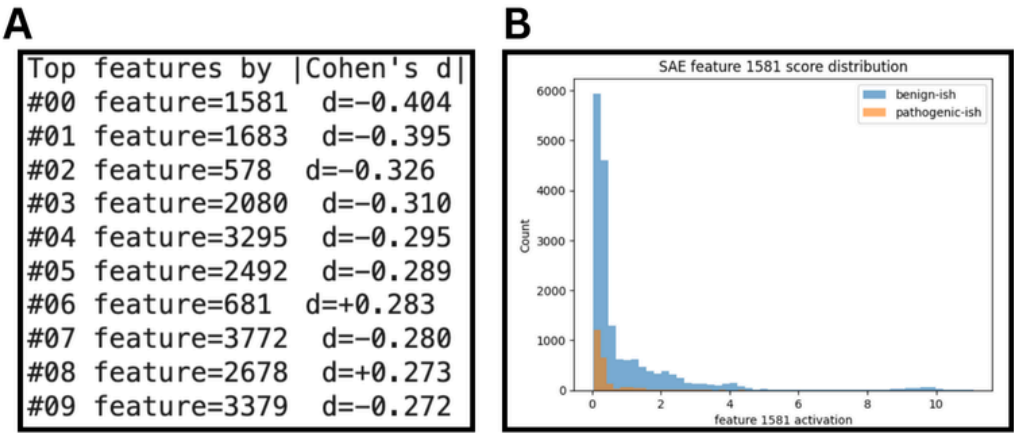

Figure 2. Top SAE features by Cohen's d + distribution plot for Feature 1581

Next, we ranked SAE features by effect size between pathogenic and benign groups (Figure 2). Ranking SAE features by effect size identified several features with consistent differences between pathogenic and benign groups. Feature 1581 had the strongest separation in our analysis (Cohen's d = −0.404, higher activation in benign than pathogenic).

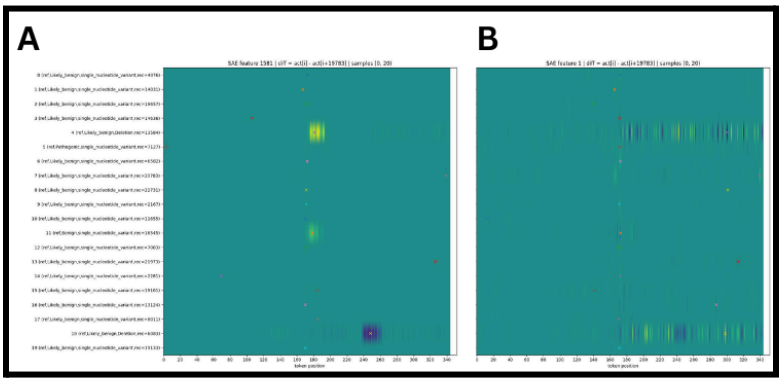

Figure 3. Token Δactivation heatmaps (Feature 1581 vs random feature)

To test whether discriminative features also localize allele effects, we computed token-level Δactivation(ALT - REF) heatmaps. Compared to a randomly chosen feature, Feature 1581 showed sharper, more localized Δactivation hotspots across many variants. (Figure 3) This pattern is consistent with Feature 1581 behaving as a detector for specific local sequence contexts that are disrupted or created by certain alleles.

A key caveat is that insertions/deletions can create downstream token misalignment, producing widespread ALT–REF differences unrelated to biology. This

artifact is visible in non-discriminatory features, which exhibit scattered downstream Δactivation. The observation that Feature 1581 shows localized effects even in non-deletion variants suggests sensitivity beyond pure shift artifacts, motivating causal validation.

Discussion

We present a compact interpretability pipeline for variant analysis with NTv2: a frozen-embedding classifier to confirm label signal, and an SAE to expose sparse token-level features. The main mechanistic handle is that at least one SAE feature (Feature 1581) both separates benign vs pathogenic groups and exhibits localized allele-dependent changes in token activations.

ClinVar labels aggregate heterogeneous mechanisms (splicing, regulatory disruption, coding effects). A plausible next step is to interpret Feature 1581 by extracting maximally activating sequence windows and testing enrichment for recognizable patterns (e.g., splice donor/acceptor consensus, CpG-rich contexts, promoter-like motifs, repeats). The directionality (higher in benign) suggests it may represent “stability-associated” contexts that pathogenic variants disrupt.

Limitations.

(i) Accuracy is inflated by class imbalance; per-class evaluation is needed. (ii) ClinVar labels are not a single mechanistic target, so feature-label associations may reflect correlations. (iii) Insertions/Deletions and tokenization complicate tokenwise REF–ALT comparisons due to alignment/shift artifacts.

Next Steps

(1) Establish causality by ablating or clamping top SAE features and measuring classifier logit shifts. (2) Use per-variant Δfeature = feature(ALT) - feature(REF) to reduce global confounds. (3) Add an indel-aware control by comparing only aligned windows around the variant.

Overall, the results suggest NTv2 internal activations can be decomposed into sparse features that meaningfully localize allele-driven representational differences.

Works Referenced

- Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., & Henighan, T. (2024). Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. Transformer Circuits Thread. Retrieved January 13, 2026, from https://transformer-circuits.pub/2024/scaling-monosemanticity/
- Olah, C. (2024, July). The Dark Matter of Neural Networks? In Circuits Updates — July 2024 (A. Jermyn, Ed.). Transformer Circuits Thread. Retrieved January 13, 2026, from https://transformer-circuits.pub/2024/july-update/index.html#dark-matter
- Welch Labs. (2024, December 23). The Dark Matter of AI [Mechanistic Interpretability] [Video]. YouTube. https://www.youtube.com/watch?v=UGO_Ehywuxc
- University of Toronto Mississauga. (2023–2024). CSC311H5: Introduction to machine learning.
- University of Toronto Mississauga. (2023–2024). CSC413H5: Neural networks and deep learning.
- Dalla-Torre, H., et al. (2025). Nucleotide Transformer: Building and evaluating robust foundation models for human genomics. Nature Methods. https://doi.org/10.1038/s41592-024-02523-z
- Dalla-Torre, H., et al. (2023). The Nucleotide Transformer. bioRxiv. https://doi.org/10.1101/2023.01.11.523679
- Boshar, S., Evans, B., Tang, Z., Picard, A., Adel, Y., Lorbeer, F. K., Rajesh, C., Karch, T., Sidbon, S., Emms, D., Mendoza-Revilla, J., Al-Ani, F., Seitz, E., Schiff, Y., Bornachot, Y., Hernandez, A., Lopez, M., Laterre, A., Beguir, K., Koo, P., Kuleshov, V., Stark, A., de Almeida, B. P., & Pierrot, T. (2025, December 23). A foundational model for joint sequence-function multi-species modeling at scale for long-range genomic prediction. InstaDeep. https://instadeep.com/research/paper/a-foundational-model-for-joint-sequence-function-multi-species-modeling-at-scale-for-long-range-genomic-prediction/
- Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., & Maglott, D. R. (2014). ClinVar: Public archive of relationships among sequence variation and human phenotype. Nucleic Acids Research, 42(1), D980–D985. https://doi.org/10.1093/nar/gkt1113

Supplementary information

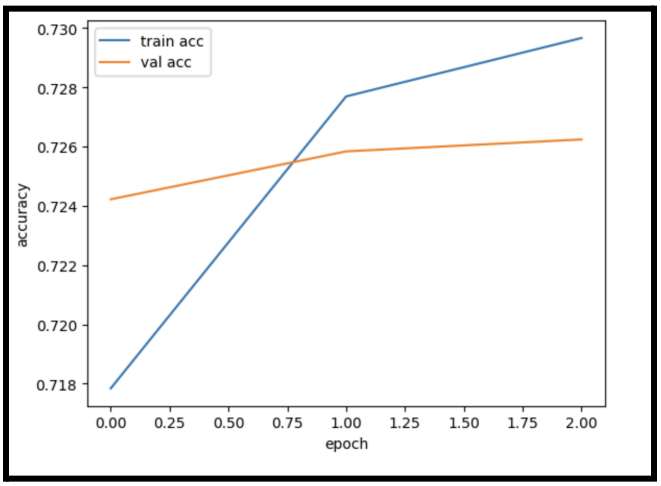

Figure S1. Training curve of Classifier Model. Training accuracy is the number of correct classifications of the Classification head divided by the total number of samples in the training sample. Validation accuracy is similar but with the validation dataset. The Classifier Model is a 1-layer MLP with 512 hidden neurons. As input, it takes the embeddings for the reference sequence, alternate sequence, difference between the sequences, and absolute value of differences.

Figure S2. Training curve of SAE model. Reconstruction loss (rcl) is blue, L1 regularization loss (L1) is orange. As L1-Regularization loss decreases (the model learns to activate neurons more sparsely), reconstruction loss increases (the model fails to encode/decode accurately). We continuously increased the L1-Regularization coefficient, with stronger increases proportional to the number of neurons activated per token.