Reid McNeil – Genomic Interoperability Report – PMCC – Computation Biologist

GitHub: https://github.com/r95m/variant-attention-interpretability.git

Introduction

This technical analysis examines how a pretrained genomic foundation model represents and responds to single-nucleotide variants, with an emphasis on interpretability rather than predictive performance. The goal is to characterize how sequence-level perturbations introduced by variants propagate through the model's internal representations and how these responses vary across biologically meaningful variant classes.

We analyze a nucleotide transformer pretrained on large-scale human reference sequence data, which is not explicitly trained on functional variant annotations. This enables investigation of how sequence constraints and dependencies are learned implicitly from DNA, without conflating interpretability results with supervised labels or task-specific objectives.

Model interpretability is approached by comparing attention patterns between reference and alternate alleles at the same genomic locus. Attention is treated as a diagnostic signal reflecting how the model reallocates contextual importance across the sequence in response to a variant. By quantifying attention perturbations as a function of distance from the variant position and stratifying results by functional class, we provide a structured view of how sequence variation alters internal model behavior. Together, these analyses aim to elucidate representation-level mechanisms by which genomic foundation models encode and respond to single-nucleotide variation.

Methods

Variant data and sequence context

Single-nucleotide variants (SNVs) annotated as Benign or Pathogenic were obtained from ClinVar and restricted to single-base substitutions involving standard DNA nucleotides (A, C, G, T). For each variant, a fixed-length DNA sequence window was extracted from the hg38 reference genome and centered on the variant position. Reference and alternate sequences were generated by substituting the central base while keeping all flanking positions identical. Variants were additionally annotated by functional class (e.g., missense, synonymous, splice, intronic, UTR) for stratified analyses. A total of 1000 variants with a length of 100bp were selected randomly for input into the model. Class balance between clinical significance between benign, variants of uncertain significance (VUS), and pathogenic were instated to prevent skewing.

Model inference and attention extraction

We used the pretrained nucleotide transformer model *InstaDeepAI/nucleotide-transformer-500m-human-ref* in inference-only mode without additional fine-tuning. Variant-centered DNA sequences were tokenized using the model's nucleotide vocabulary and processed individually. Attention matrices were extracted from the final transformer layer. For each attention head, the attention matrix encodes pairwise attention weights between all input tokens. To obtain a per-token measure of attention received, attention weights were summed column-wise across source positions and averaged across heads, yielding a one-dimensional attention-received profile per sequence.

Attention perturbation and variant centrality

Attention perturbation was defined as the difference between attention-received profiles of the alternate and reference sequences ($\Delta\text{attention} = \text{attention}_{\text{alt}} - \text{attention}_{\text{ref}}$). Perturbations were aligned relative to the variant position to enable aggregation across variants. Variant token centrality was assessed by ranking tokens according to absolute Δattention magnitude and computing the fraction of variants for which the variant-associated token ranked among the most perturbed positions.

Distance-based summarization

To characterize the spatial structure of attention perturbations, absolute Δattention values were aggregated as a function of token distance from the variant position. Distance-decay profiles were computed by averaging perturbation magnitudes across variants at each relative position. Locality was summarized using the fraction of total absolute Δattention contained within a ±5-token window around the variant position. All distance-based analyses were performed both across all variants and stratified by functional class.

Activation patching

To test whether variant-token representations are causally responsible for observed attention perturbations, we performed activation patching experiments. For each variant, the model was run on both the reference and alternate sequences, and the hidden-state representation at the variant token position was extracted from a selected intermediate transformer layer. During a subsequent forward pass on the alternate sequence, this variant-token representation was replaced with the corresponding representation from the reference sequence, while all other activations were left unchanged.

Attention patterns from the patched alternate sequence were compared to those from the unpatched alternate and reference sequences. Causal effects were quantified using a local rescue metric, defined as the fractional reduction in absolute Δattention within a ±5-token window around the variant position. Rescue analyses were performed across a large set of variants and stratified by functional class.

Results

Attention perturbations are localized and centered on the variant token

To characterize how single-nucleotide variants perturb model attention, we quantified the spatial distribution of absolute attention changes across sequence tokens. Across variants, a substantial fraction of total attention perturbation was concentrated within ±5 tokens of the variant position, exceeding the expected fraction under a uniform distribution (Figure 1A). Ranking tokens by absolute attention change further revealed that variant-associated tokens were consistently among the most perturbed positions, with cumulative centrality curves deviating strongly from the diagonal null expectation (Figure 1B). When aggregated by token distance, mean perturbation magnitude peaked at the variant position and decayed with increasing distance, with a secondary increase at distal positions, indicating structured non-local effects superimposed on a dominant local signal (Figure 1C).

Attention locality is preserved across functional class

We next stratified attention perturbation patterns by variant functional class. Across all classes, the fraction of attention perturbation localized within ±5 tokens remained elevated and broadly comparable (Figure 2A), indicating that strong local effects are a shared property of variant perturbations regardless of annotation. Variant token centrality curves showed similar enrichment across functional classes, with variant-associated tokens disproportionately ranked among the most perturbed positions relative to a null ordering (Figure 2B). Distance-resolved profiles revealed consistent decay of mean attention perturbation with distance from the variant across classes, while preserving class-specific differences in perturbation magnitude and distal structure (Figure 2C). Together, these results indicate that attention locality and variant-centric perturbation are robust features of the model across diverse functional annotations.

Locality and variant token centrality are conserved across functional classes

We next stratified attention perturbation patterns by variant functional class. Across all classes, the fraction of attention perturbation localized within ±5 tokens remained elevated and broadly comparable (Figure 2A), indicating that strong local effects are a shared property of variant perturbations regardless of annotation. Variant token centrality curves showed similar enrichment across functional classes, with variant-associated tokens disproportionately ranked among the most perturbed positions relative to a null ordering (Figure 2B). Distance-resolved profiles revealed consistent decay of mean attention perturbation with distance from the variant across classes, while preserving class-specific differences in perturbation magnitude and distal structure (Figure 2C). Together, these results indicate that attention locality and variant-centric perturbation are robust features of the model across diverse functional annotations.

Variant-token representations causally drive attention redistribution

To test whether variant-token representations causally drive attention redistribution, we performed activation patching across a larger set of variants (N = 250). Replacing the alternate allele's token-level representation with the corresponding reference representation at an intermediate transformer layer consistently reduced local attention perturbation, yielding a positive median rescue across variants (Figure 4A). Stratification by functional class revealed broadly consistent rescue across both coding and non-coding variants, with class-specific variability in rescue magnitude (Figure 4B), indicating that variant-token representations contribute causally to attention redistribution across diverse variant types.

Discussion

In this study, we characterized how single-nucleotide variants perturb attention patterns in a sequence-based genomic model. Across analyses, attention perturbations were strongly localized around the variant position, with variant-associated tokens consistently among the most perturbed. This indicates that the model responds to sequence variation in a spatially structured, variant-centric manner rather than distributing effects diffusely across the input.

To move beyond descriptive analysis, we tested whether the variant-token representation itself is causally responsible for these perturbations using activation patching. Replacing the alternate allele's token-level representation with the corresponding reference representation consistently reduced local attention perturbations, yielding a positive median rescue across a large set of variants. This demonstrates that variant-token representations are causally upstream of a substantial fraction of attention redistribution, rather than merely correlated with it.

Stratification by functional class revealed broadly consistent positive rescue across both coding and non-coding variants, with class-dependent variability in rescue magnitude, suggesting that different variant types engage partially distinct internal mechanisms. At the level of individual variants, both attention perturbation profiles and rescue effects were heterogeneous, indicating that variant interpretation depends on local sequence context and downstream interactions beyond the variant token itself. The partial nature of the rescue suggests that variant effects are distributed across both local token representations and broader sequence context, consistent with the model capturing nuanced, context-dependent responses to sequence variation.

Notably, attention perturbation profiles were broadly similar between coding and non-coding variants, with all classes exhibiting strong localization near the variant position and rapid decay with distance. This suggests that, at the level of attention redistribution, local sequence perturbation dominates over functional annotation, and that biological labels provide contextual interpretation rather than fundamentally distinct attention mechanisms.

Limitations

Several limitations should be considered when interpreting these results. First, analyses were performed on a limited subset of variants and a fixed sequence window length, which may constrain the observed spatial extent of attention perturbations. Expanding both the number of variants and sequence context length may reveal additional long-range effects not captured here.

Second, while attention perturbation alone reflects model sensitivity rather than causal importance, we partially addressed this limitation through activation patching experiments. By intervening directly on variant-token representations and measuring the resulting reduction in attention perturbation, we establish a causal link between specific internal representations and downstream attention behavior. However, these analyses do not assess causal effects on model predictions or biological function and should be interpreted as elucidating internal model mechanisms rather than variant pathogenicity.

Third, functional class annotations were treated as coarse categories, and heterogeneity within classes was not explicitly modeled. Future work could incorporate finer-grained annotations or integrate external biological priors to better relate attention perturbation patterns to known regulatory mechanisms. Finally, the observed attention structures are specific to the model architecture and training data used in this study. Whether similar locality and perturbation patterns generalize to other sequence models remains an open question.

Figures

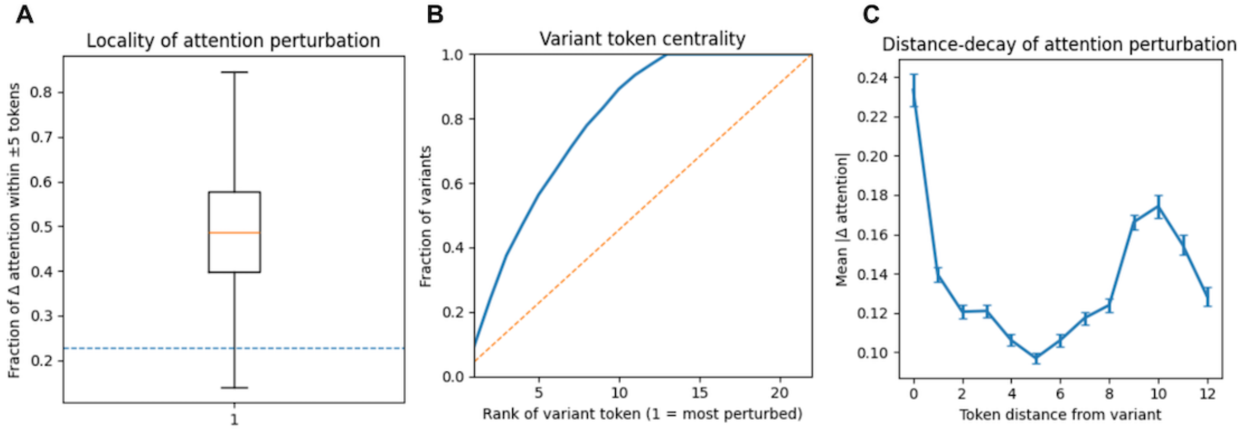


Figure 1: Attention perturbation profiles induced by single-nucleotide variants. (A) Distribution of the fraction of total absolute attention change occurring within a ±5-token window centered on the variant position; the dashed line indicates the expected fraction under a uniform distribution across tokens. (B) Empirical cumulative distribution of the rank of the variant token based on absolute attention change (rank 1 corresponds to the most perturbed token); the dashed line indicates the expected distribution under random ranking. (C) Mean absolute attention change as a function of token distance from the variant position, averaged across all variants; error bars represent the standard error of the mean.

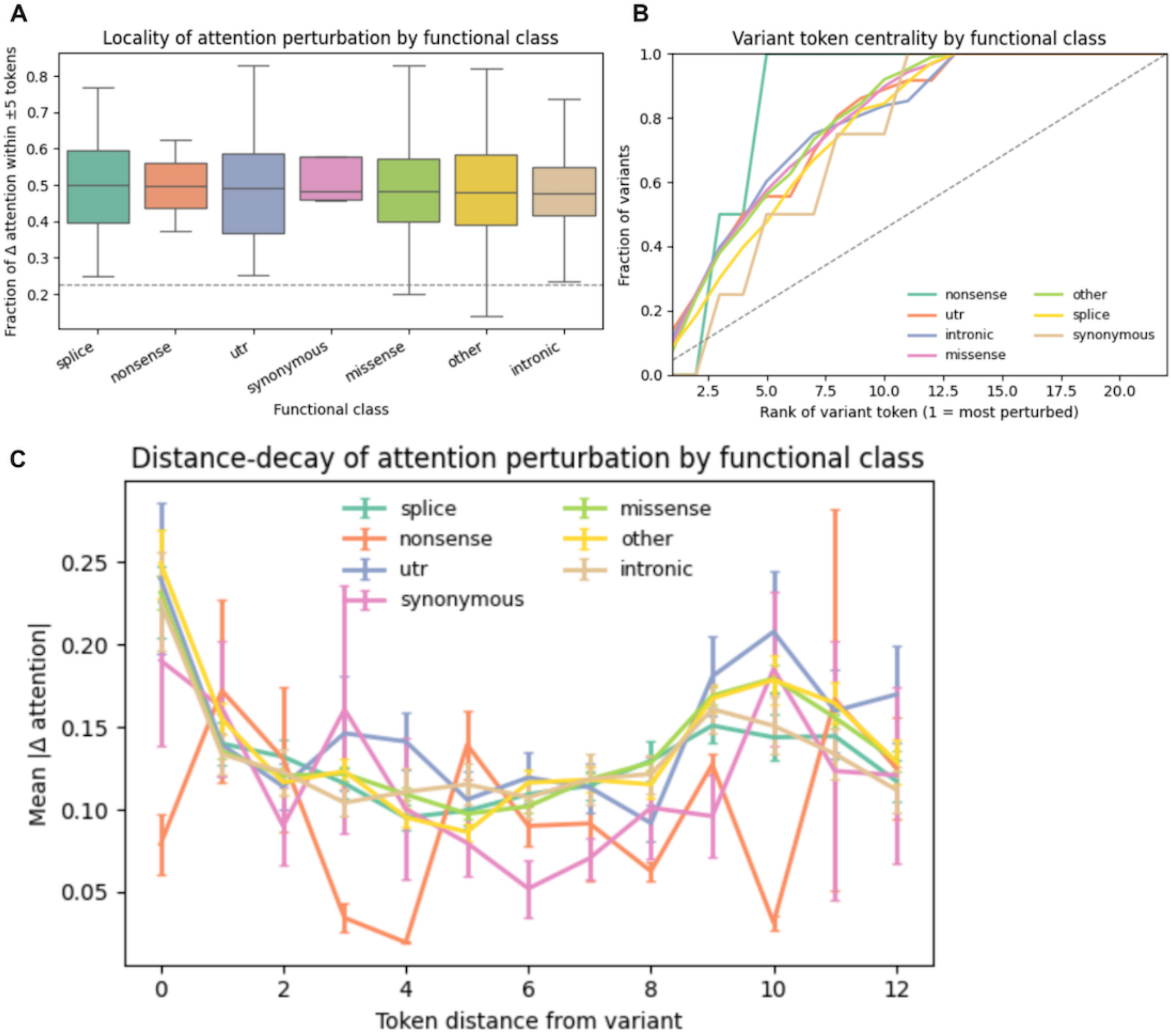


Figure 2: Attention perturbation profiles stratified by functional class. (A) Distribution of the fraction of total absolute attention change occurring within a ±5-token window centered on the variant position, shown separately for each functional class; the dashed line indicates the expected fraction under a uniform distribution across tokens. (B) Empirical cumulative distribution of the rank of the variant token based on absolute attention change (rank 1 corresponds to the most perturbed token), stratified by functional class; the dashed line indicates the expected distribution under random ranking. (C) Mean absolute attention change as a function of token distance from the variant position, averaged within each functional class; error bars represent the standard error of the mean.

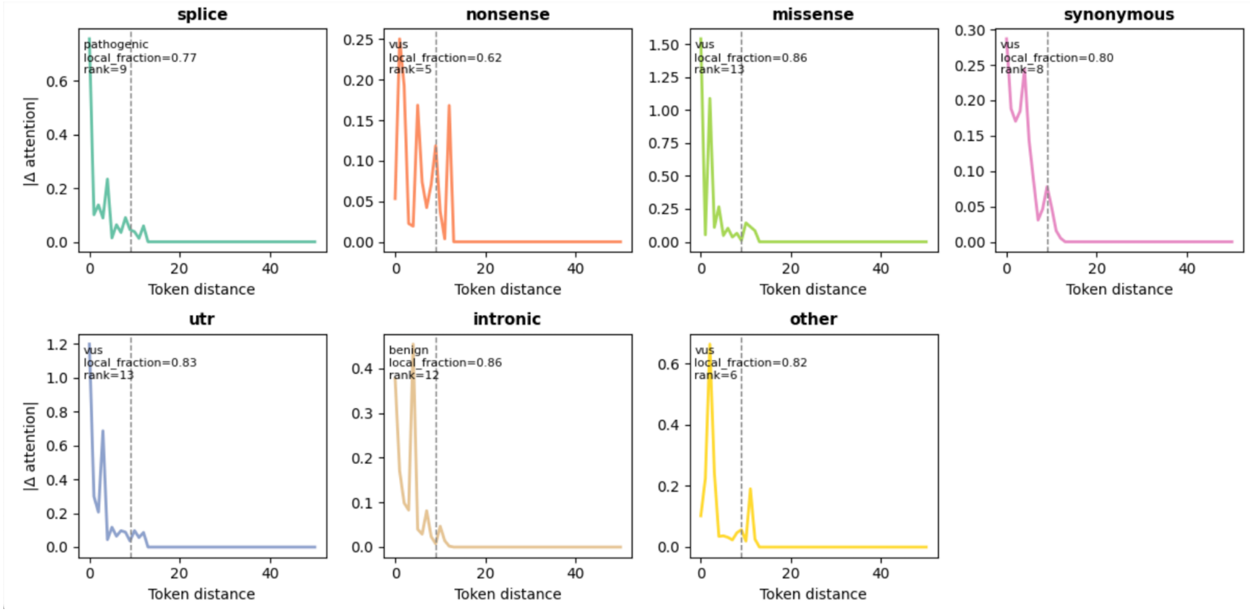


Figure 3: Representative attention perturbation profiles for individual variants. Line plots show the absolute change in attention as a function of token distance from the variant position for representative variants selected from each functional class (splice, nonsense, missense, synonymous, UTR, intronic, and other). Each panel displays a single variant's distance–decay profile, with the dashed vertical line indicating the ±5-token window used to compute local attention metrics. Annotations report the variant's clinical significance, local fraction of total attention change within the ±5-token window, and the rank of the variant token based on absolute attention change.

.

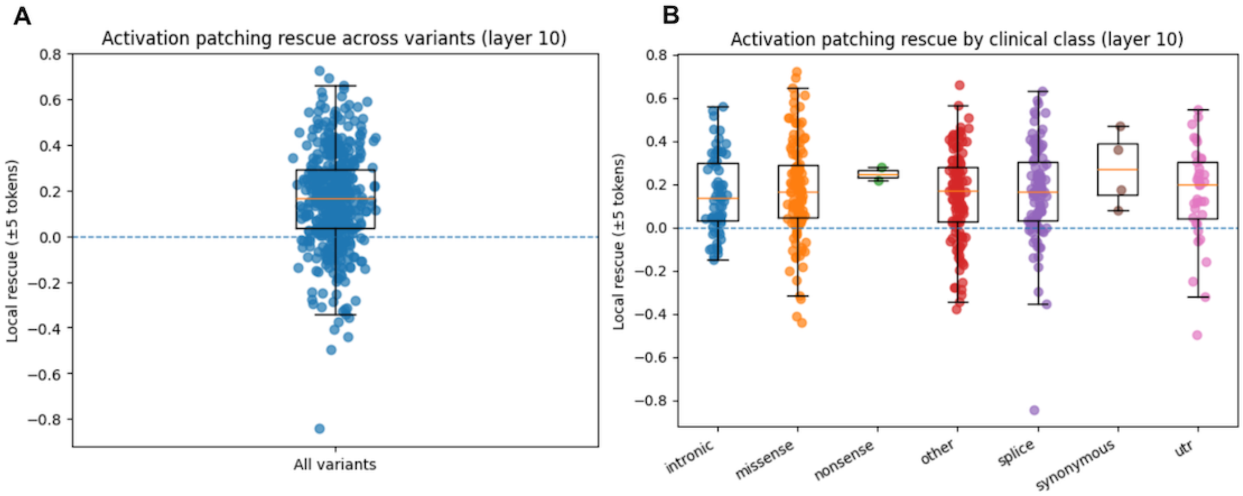
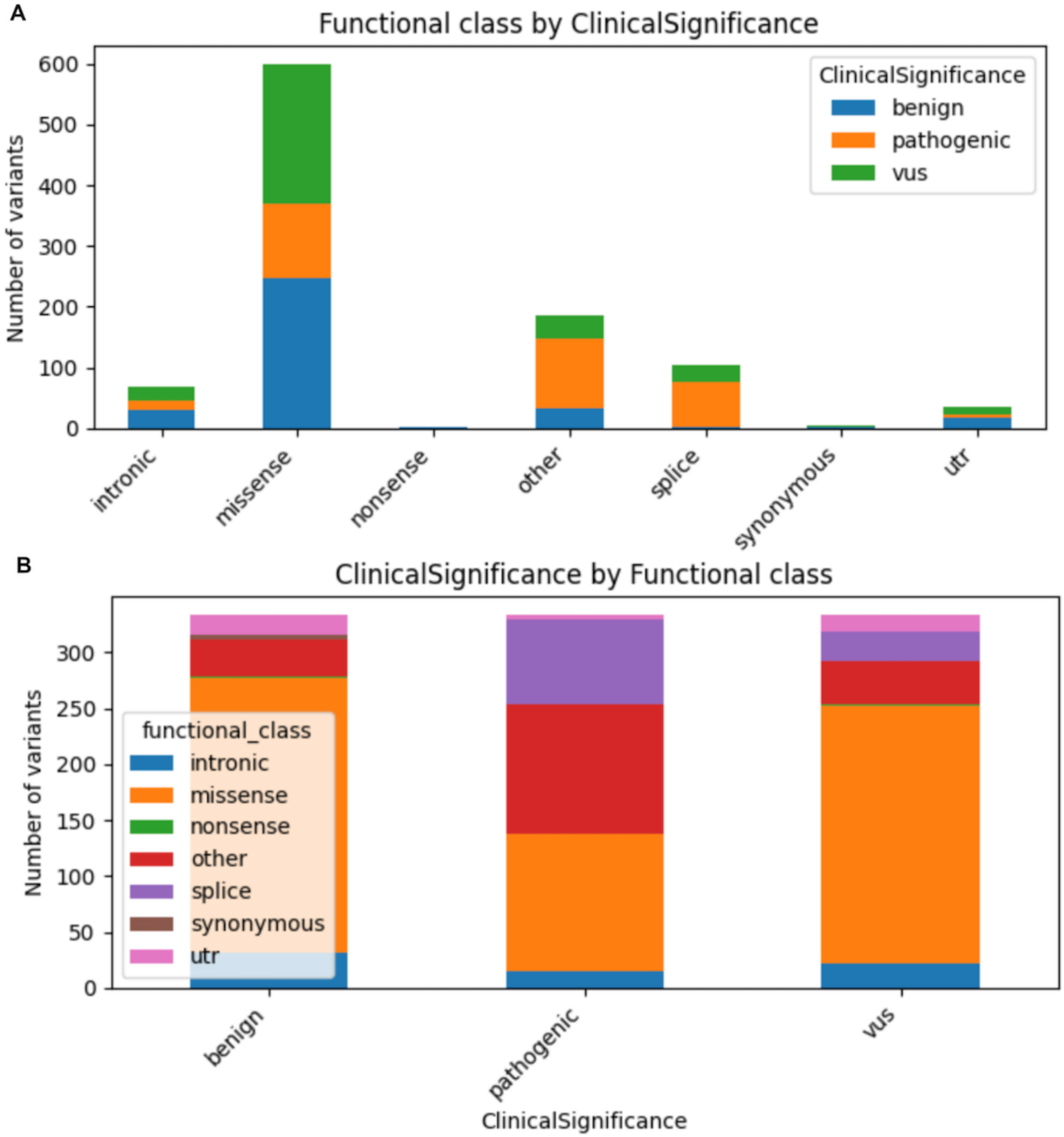


Figure 4: Activation patching rescue across variants and functional classes. (A) Local rescue scores (±5 tokens) quantify the reduction in variant-induced attention perturbation after replacing the alternate allele's token-level representation with the reference representation at layer 10. Each point represents one variant (N = 250). (B) Local rescue scores stratified by functional class show consistent positive medians across coding and non-coding variants, indicating that variant-token representations contribute causally to attention redistribution across diverse variant types.



Supplementary Figure 1: Composition of variant dataset used for analysis. Distribution of single-nucleotide variants included in the analysis by functional class and clinical significance. (A) Counts of variants stratified by functional class (intronic, missense, nonsense, splice, synonymous, UTR, other), with bars colored by Clinvar clinical significance (benign, pathogenic, VUS). (B) Complementary view showing the distribution of functional classes within each clinical significance category.