EVEE: Interpretable variant effect prediction from genomic foundation model embeddings

Michael T. Pearce[*1], Thomas Dooms[*1], Ryo Yamamoto[*1], Joshua Meehl[2], Carl Molnar[2], Mark Bissell[1], Dron Hazra[1], Ching Fang[1], Nam Nguyen[1], Michael Anderson[1], Collin Osborne[2], Patrick Duffy[2], Bridget Toomey[2], Eric Klee[2], Elena Myasoedova[2], Alexander Ryu[2], Shant Ayanian[2], Panos Korfiatis[2], Matt Redlon[2], Archa Jain[1], Daniel Balsam[1], and Nicholas K. Wang[1,†]

[1]Goodfire; [2]Mayo Clinic

Abstract

Predicting the clinical significance of genetic variants remains a central challenge in genomic medicine, with most observed variants classified as variants of uncertain significance. Here we show that representations from Evo 2, a 7-billion-parameter genomic foundation model, support accurate and interpretable pathogenicity prediction across variant types from a single framework. An embedding-based classifier, or "probe", trained on Evo 2 embeddings achieves state-of-the-art performance across single nucleotide variant consequence types (0.997 overall AUROC on 833k ClinVar variants) and generalizes zero-shot to indels (0.991 AUROC), outperforming bioinformatic meta-predictors, protein models, and existing foundation model approaches. Performance is robust across conservation levels and transfers to deep mutational scanning datasets for BRCA1, BRCA2, TP53, and LDLR. To make these predictions interpretable, we train supervised annotation probes to quantify predicted disruptions caused by each variant, then synthesize these disruption profiles into natural language explanations using a frontier reasoning model. We provide pre-computed predictions and on-demand explanations for all 4.2 million ClinVar variants through the Evo Variant Effect Explorer (EVEE), an interactive web resource for the community. This work establishes that representations from genomic foundation models can serve as a unified substrate for both accurate variant effect prediction and mechanistic interpretation, reframing interpretability in computational genomics from a trade-off into a complementary product of learned biological structure.

Main

Clinical interpretation of genetic variants remains a critical bottleneck in genomic medicine. Despite exponential growth in sequencing, the majority of observed variants remain classified as variants of uncertain significance (VUS) (Landrum et al., 2016; Richards et al., 2015). Computational variant effect predictors have proliferated, spanning protein language models such as AlphaMissense (Cheng et al., 2023) and EVE (Frazer et al., 2021), supervised sequence-to-activity models such as AlphaGenome (Avsec et al., 2026) and Nucleotide Transformer v3 (Boshar et al., 2025; Dalla-Torre et al., 2025), evolutionary DNA models such as Evo 2 (Brixi et al., 2026) and GPN-MSA (Benegas et al., 2023), compressed protein language models (Dinh et al., 2026), and classical meta-predictors such as CADD (Schubach et al., 2024). However, these tools face complementary limitations: protein-based methods are restricted to missense variants; AlphaGenome focuses on regulatory noncoding effects; and CADD, while genome-wide, integrates 100+ annotation features with crossed terms and rank transformations that obscure individual feature contributions (Avsec et al., 2026; Schubach et al., 2024; Kircher et al., 2014). Critically, none provide human-readable explanations of why a variant is predicted pathogenic—a key requirement for clinical variant classification under ACMG/AMP guidelines (Richards et al., 2015), which demands categorized evidence rather than opaque scores.

Genomic foundation models, large neural networks pretrained on DNA sequences, learn representations that capture functional constraints across coding and noncoding regions (Dalla-Torre et al., 2025; Brixi et al., 2026; Feng et al., 2025). Evo 2 in particular has demonstrated zero-shot variant effect predictions and internal representations that correspond to biological features such as exon–intron boundaries and protein structural elements (Brixi et al., 2026). However, systematic evaluation of embedding-based classifiers, or "probes", and frameworks for generating mechanistic interpretation from these representations have been lacking.

Here we address both gaps: we introduce a covariance probe that achieves state-of-the-art pathogenicity predic-

[*]These authors contributed equally to this work.
† Correspondence: nick.wang@goodfire.ai

tion across SNV and indel variant types and present a supervised annotation-disruption framework coupled with LLM-based synthesis that produces contextualized, human-readable explanations for each variant's predicted pathogenicity.

Our approach processes reference and alternate DNA sequences through Evo 2, computes the difference in per-position embeddings, and trains a covariance probe on the resulting representation to predict pathogenicity (Figure 1a). We evaluated pathogenicity classification on 833,970 ClinVar SNVs from genes shorter than 100 kb. Because pathogenic fractions vary dramatically across consequence types (nonsense: 98.3%; synonymous: 0.1%), aggregate metrics can be inflated by models that learn consequence-type priors rather than variant-level discrimination. We addressed this distributional confound by stratifying on each consequence type (Feng et al., 2025; Lu et al., 2025) (Figure 1b). An Evo 2 covariance probe achieved 0.997 overall AUROC, with strong performance across all consequence types: 0.971 for missense (n=125,092), 0.961 for synonymous (n=292,369), 0.900 for nonsense (n=36,050), 0.924 for splice (n=21,683), 0.929 for UTR (n=31,911), 0.984 for intronic (n=254,737), and 0.969 for other non-coding variants (n=3,498). This covariance probe outperformed a linear probe on Evo 2 mean-pooled embeddings, Evo 2 loss-based scoring, CADD, AlphaMissense, GPN-MSA, NTv3, and AlphaGenome composite scores across all variant consequence types. Results were confirmed on a deconfounded benchmark that balances pathogenic and benign labels within each consequence type (Supplementary Figure S1), where the covariance probe (0.971 overall AUROC) was matched only by AlphaMissense for missense variants (0.972 AUROC) while exceeding all other methods.

Instead of mean-pooling over the embedding matrix $\mathbf{X} \in \mathbb{R}^{L \times d}$, the covariance probe is trained on the Gram matrix $\mathbf{X}^\top \mathbf{X} \in \mathbb{R}^{d \times d}$. This matrix captures the second-order structure of the embeddings, such as the correlations between model dimensions or co-occurrence of sparse features along the sequence. Since the dimensionality of the matrix is large, in practice we compute a compressed version using linear down-projections (Figure 1a; Methods) (Dooms et al., 2026). This allows for lightweight probe training that improves over logistic regression on mean-pooled embeddings. Since Evo 2 is an autoregressive model, we concatenate the embeddings for the sense and antisense directions along the gene to obtain bidirectional embeddings that are used for probe training throughout.

Strikingly, the Evo 2 covariance probe generalized zero-shot to indels despite being trained exclusively on SNVs. On 73,961 ClinVar indels, it achieved 0.986 overall AUROC, with strong performance across size categories: 0.989 for 1-bp, 0.991 for 2–5 bp, 0.984 for 6–20 bp, and 0.916 for >20 bp indels, as well as 0.991 for insertions and 0.988 for deletions (Figure 1c). This zero-shot transfer outperformed CADD v1.7 InDel (0.980 overall) and substantially exceeded NTv3 (0.828), demonstrating that representations learned from SNV training capture general principles of sequence disruption that extend to insertion–deletion variation. Performance was lower for frameshifts (0.770) due to severe class imbalance (99% pathogenic), though still substantially above NTv3 (0.602).

The covariance probe also showed robust performance across conservation levels, maintaining high AUROC from fast-evolving to highly conserved sites, whereas CADD and GPN-MSA performance degraded at the extremes of the conservation spectrum (Figure 1d). This suggests that Evo 2 representations encode functional constraint information that is complementary to, rather than redundant with, phylogenetic conservation. UMAP visualization of reference-subtracted Evo 2 embeddings further confirmed that the learned representations capture meaningful biological structure, revealing clear organization by both pathogenicity and variant consequence type (Figure 1e,f).

To validate that Evo 2 representations capture biology beyond ClinVar-specific patterns, we evaluated transfer to deep mutational scanning (DMS) datasets for four well-characterized genes that were held out from probe training: BRCA1 (Findlay et al., 2018), BRCA2 (Sahu et al., 2025), TP53 (Funk et al., 2025), and LDLR (Tabet et al., 2025) (Figure 1g). We assessed correlation between model predictions and continuous DMS functional scores (Spearman $|\rho|$), comparing Evo 2 mean-pooled and covariance-pooled probes trained on ClinVar labels against CADD, AlphaMissense, and Evo 2 loss-based scoring. We also evaluated probes trained directly on DMS data from the other three genes with both pooling strategies.

Across all four genes, ClinVar-trained Evo 2 embedding probes outperformed loss-based scoring, confirming that learned representations encode functional information beyond next-token likelihood. The covariance probe was competitive with AlphaMissense and CADD across genes and exceeded both on TP53 ($|\rho| \approx 0.70$), while CADD showed a particular advantage on BRCA2 ($|\rho| \approx 0.46$ vs. 0.30–0.37 for other methods). Covariance probes consistently outperformed their mean-pooled counterparts, mirroring the gains observed on ClinVar and confirming that second-order embedding structure captures functional information lost under mean pooling. In addition, we trained hold-one-gene-out probes using the DMS datasets that achieved performance comparable to the ClinVar-trained probes, indicating that clinical labels and experimental functional data capture overlapping biology. Together, these results demonstrate that Evo 2 embeddings general-

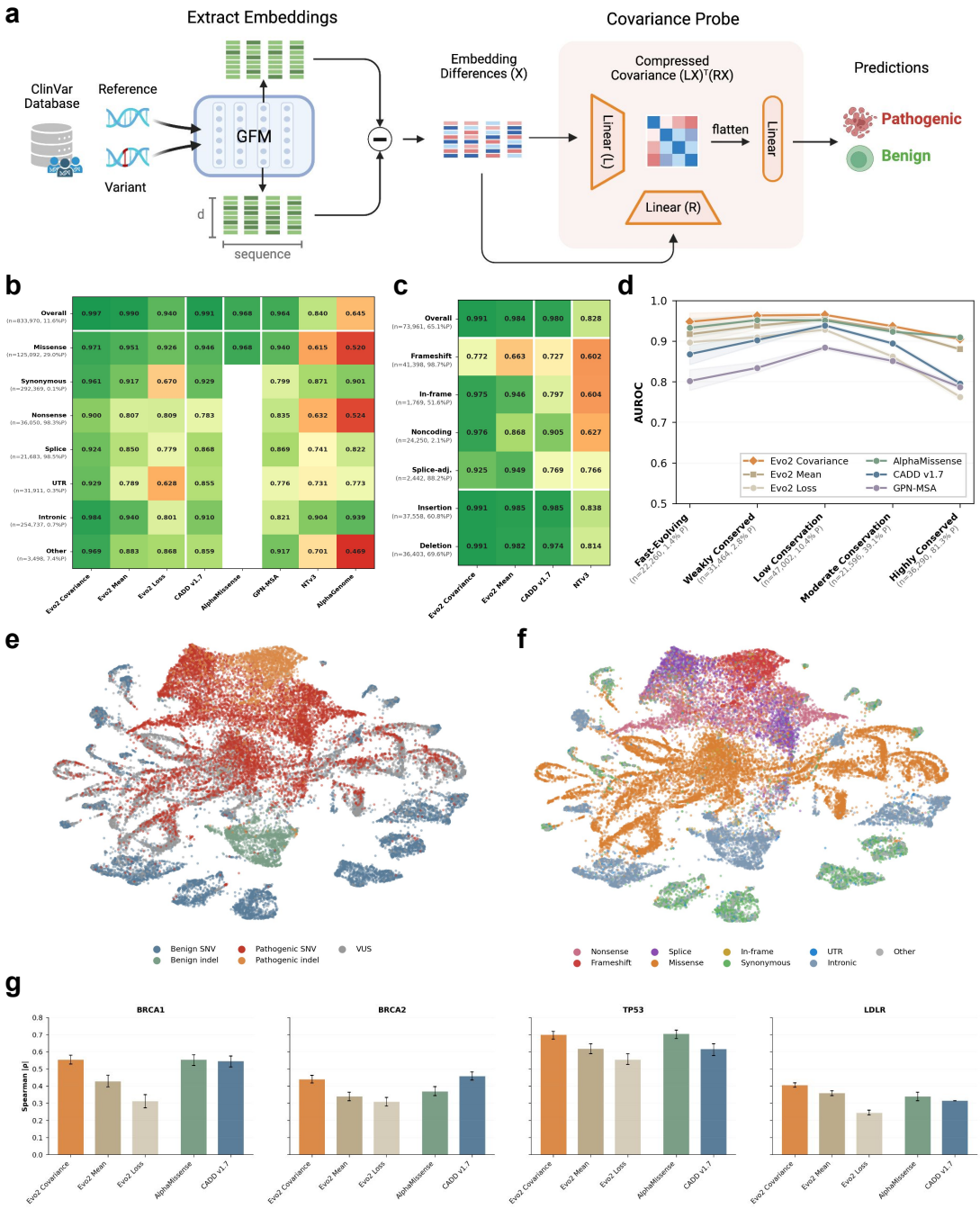

Figure 1. Evo 2 covariance probe achieves state-of-the-art variant effect prediction across variant types. (**a**) Experimental design. Evo 2 processes reference and alternate DNA sequences, producing per-position embeddings. The difference between the embeddings are then input into a covariance probe that is trained to predict pathogenicity. For each input, the covariance probe computes a compressed covariance along the sequence dimension that is flattened and passed through a linear layer to obtain the prediction logits. (**b**) ClinVar pathogenicity prediction across SNV consequence types. AUROC heatmap for Evo 2 covariance probe, Evo 2 mean probe, Evo 2 loss, CADD v1.7, AlphaMissense, GPN-MSA, NTv3, and AlphaGenome composite on 833,970 variants from genes $\leq$100 kb with $\geq$1 star review status. Blank cells indicate methods without predictions for that variant type. (**c**) Zero-shot generalization to indels. AUROC for Evo 2 covariance probe, Evo 2 mean probe, CADD v1.7, and NTv3 probe on 73,961 ClinVar indels stratified by consequence type and insertion vs. deletion. (**d**) Performance robustness across conservation levels. AUROC by phyloP100way conservation tier for Evo 2 probes, AlphaMissense, CADD v1.7, Evo 2 loss, and GPN-MSA. Evo 2 probes and AlphaMissense maintain performance across all bins, while loss-based and annotation-based approaches show degraded performance at conservation extremes. (**e**) UMAP of variant covariance embeddings colored by ClinVar pathogenicity label (benign, VUS, pathogenic) for SNVs and indels. (**f**) UMAP of variant covariance embeddings colored by VEP consequence type, showing organized structure by variant class for both SNVs and indels. (**g**) DMS generalization. Spearman $|\rho|$ between predicted scores and continuous DMS functional readouts for BRCA1, BRCA2, TP53, and LDLR. Methods compared: CADD, AlphaMissense, Evo 2 loss, ClinVar-trained mean-pooled probe (ClinVar probe), ClinVar-trained covariance probe (ClinVar cov probe), DMS-trained mean-pooled probe from held-out genes (DMS probe), and DMS covariance probe (DMS cov probe). Error bars show 95% bootstrap confidence intervals.

ize from clinical pathogenicity labels to direct experimental measurements of variant effects.

Having established that Evo 2 embeddings support accurate and generalizable pathogenicity prediction, we next asked whether we could make these predictions interpretable. Rather than relying on unsupervised decomposition of model representations, we adopted a supervised approach: we trained annotation probes directly on the same Evo 2 reference-sequence embeddings to predict a comprehensive panel of known biological annotations, then quantified how each variant disrupts these predicted annotations relative to the reference sequence (Figure 2a).

The annotation panel spans 251 properties organized across multiple categories: protein structural features, structural context, regulatory marks, protein domains, post-translational modifications, genomic region identity, and variant-level effects. Probes achieved strong performance: median AUC of 0.919 across a larger set of 357 annotation types on held-out reference positions (Figure 2b), confirming that Evo 2 embeddings encode rich annotation information that can be reliably decoded.

For each variant, we compute a disruption profile using two complementary probes trained on Evo 2's internal representations. A token-level probe predicts biological annotations (protein domains, post-translational modifications, splice sites, secondary structure) at each genomic position, and we measure the change in each prediction between the variant and reference sequences as a raw delta. The probe operates at per-position resolution: in addition to disruptions at the variant site itself, we identify significant changes at up to five flanking positions, capturing long-range effects such as splice site disruption propagating across exon–intron boundaries or domain loss extending tens of base pairs from the mutation. A separate sequence-level probe predicts amino acid biophysical properties (BLOSUM62 substitution score, Grantham physicochemical distance, charge alteration, hydrophobicity and volume changes) and splice disruption probability, providing a complementary view of the variant's effect at the protein level. At coding positions, the token-level probe also predicts amino acid identity, enabling direct comparison of the model's predicted reference and variant residues. The resulting multi-scale disruption profile provides a structured fingerprint of predicted molecular consequences, derived entirely from learned sequence representations rather than curated databases, and transparent at the individual-variant level.

To synthesize disruption profiles into human-readable interpretations, we feed the top 10 disruptions (ranked by magnitude) along with variant metadata (gene name, consequence type, and HGVS protein/coding notation) into an LLM, which generates a contextualized natural-language explanation of the variant's predicted pathogenicity (Figure 2d). For a BRCA1 intronic variant (c.4987−3C>G), the system detects complete loss of splice acceptor recognition at the variant site, with intron-to-exon boundary prediction dropping from 1.00 to 0.00 ($\Delta = -1.00$) and branchpoint region prediction collapsing from 0.99 to 0.07 ($\Delta = -0.92$). The polypyrimidine tract, which guides spliceosome assembly upstream of the acceptor site, similarly drops from 0.98 to 0.12 ($\Delta = -0.86$). Together, these three disruptions indicate that the model recognizes coordinated destruction of the splice acceptor machinery at the exon 15 boundary. RNA studies confirm that this variant produces aberrant, frameshifting transcripts due to weakened splice acceptor recognition (Brandão et al., 2011). The predicted pathogenicity of 71% is consistent with a loss-of-function allele in a haploinsufficiency-intolerant tumor suppressor gene.

We evaluated interpretation quality using an LLM-as-judge approach against expert panel evidence from ClinVar. We selected 154 expert-reviewed ClinVar variants ($\geq$3 stars; 98 benign, 56 pathogenic) and used the expert panel's submission text as ground truth. Interpretations were scored on three axes: mechanism coverage (using a 5-level rubric assessing the specificity of correctly identified molecular mechanisms), biological accuracy (consistency of biological claims with expert evidence), and specificity (whether claims are specific to the variant rather than generic). We systematically varied the information provided to the interpreting LLM to quantify the contribution of each context component (Figure 2c,e). Starting from genomic coordinates alone (composite score 1.47), adding gene name (+0.43), other context such as constraint scores and gene function (+0.40), and HGVS protein notation (+0.57) each improved quality, but the largest gain came from adding Evo 2 probe predictions (+1.02), yielding a composite score of 3.89/5 (mechanism=3.15, biological accuracy=4.27, specificity=4.15) with Claude Opus 4.6 as the interpreter.

We provide all predictions, disruption profiles, and variant interpretations through the Evo Variant Effect Explorer (EVEE), an interactive web application that allows users to browse pre-computed results for 4.2 million ClinVar variants, inspect disruption profiles across annotation categories, and examine LLM-synthesized interpretations.

Our results establish Evo 2 embeddings as a unified foundation for variant effect prediction that addresses three persistent gaps in the current landscape. First, our approach resolves the variant-type coverage gap: a single covariance probe achieves state-of-the-art performance across all SNV consequence types and indels, in contrast to the numerous protein-specific tools restricted to missense or coding variants. Second, our probe is robust across the conservation spectrum, maintaining

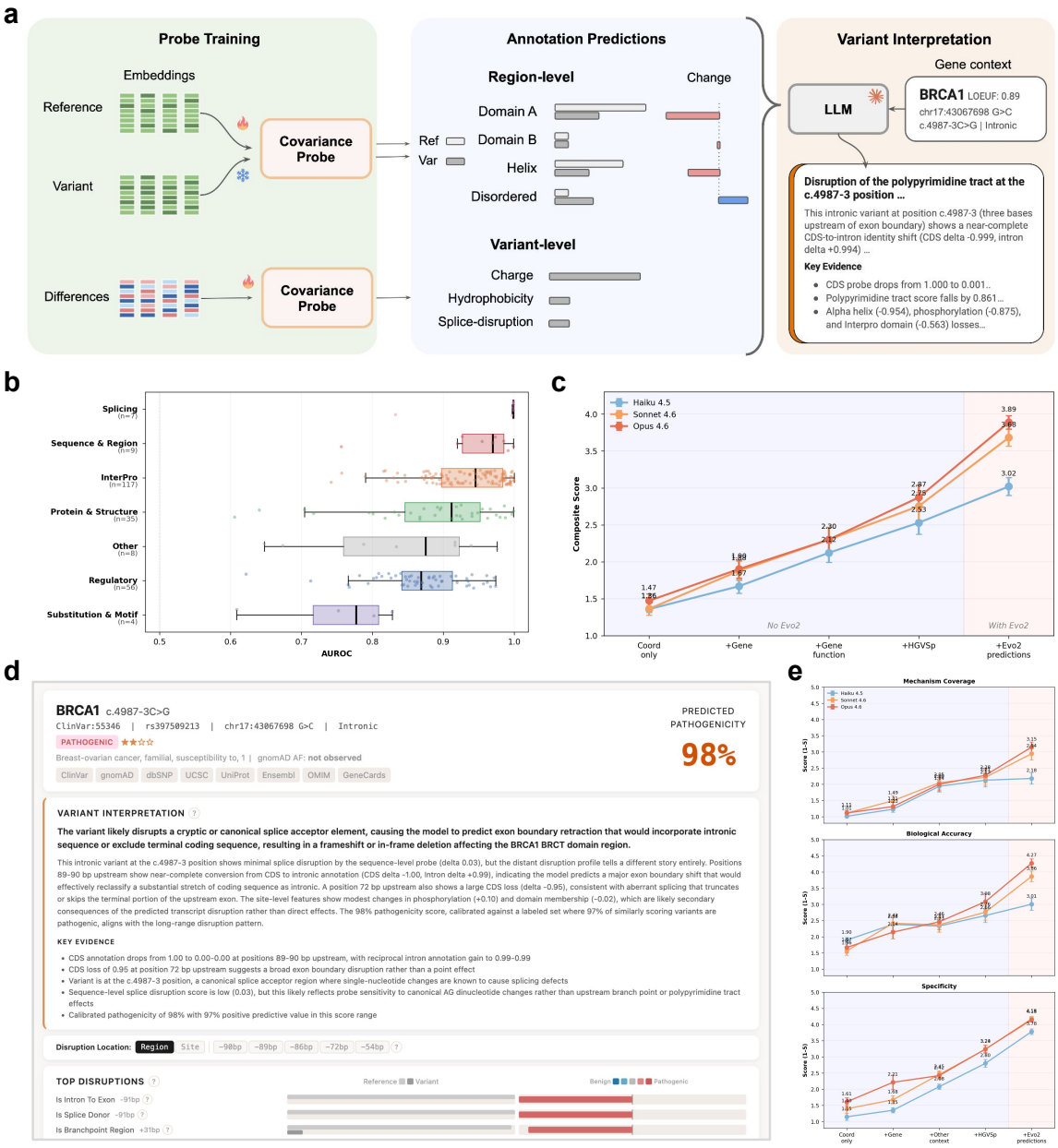

Figure 2. Interpretable predictions through annotation disruption profiling and LLM-based synthesis. (**a**) Schematic of the interpretability framework. Supervised annotation probes are trained on Evo 2 reference-sequence embeddings to predict a panel of biological annotations. For each variant, disruptions (Δ between variant and reference predictions) are computed across all annotations. Top disruptions are synthesized into natural-language explanations by an LLM. (**b**) Annotation probe performance across annotation categories. Horizontal boxplot with jittered scatter of per-head AUROC values, grouped by annotation category for 357 binary annotation probes. (**c**, **e**) Evaluation of variant interpretations using an LLM-as-a-judge approach against ClinVar ($\geq$3 star) expert review summaries. (c) shows composite scores across cumulative context configurations for three Claude model tiers (Haiku 4.5, Sonnet 4.6, Opus 4.6); (e) shows the breakdown by individual criteria: mechanism coverage, biological accuracy, and specificity. Error bars show 95% confidence intervals. (**d**) Example interpretation for a BRCA1 splice-disrupting intronic variant. Disruption profile showing predicted changes in exon–intron boundaries, splicing donor, zinc finger domain, and variant effect annotations, with LLM-generated natural-language explanation contextualizing the disruption profile against known biology.

performance where annotation-dependent methods degrade. Third, the annotation-disruption framework coupled with LLM-based synthesis produces structured, human-readable explanations that move toward the categorized evidence ACMG/AMP guidelines require (Richards et al., 2015). Rather than an opaque score, researchers and clinicians can inspect which specific biological properties are disrupted and how they collectively contribute to a variant's predicted effect.

Several limitations warrant discussion. Evo 2 is trained

on a broad evolutionary prior that captures functional constraints across all domains of life, making it well suited for detecting the strong deleterious effects characteristic of Mendelian pathogenic variants. However, this same prior may be less calibrated for subtler polygenic effects, where variants of small individual effects act in aggregate and may not leave clear evolutionary signatures detectable from local sequence alone. The annotation probes are constrained to predicting known annotations; truly novel molecular mechanisms of pathogenicity would not be captured by this supervised approach. Unsupervised methods such as sparse autoencoders applied to genomic foundation model representations (Brixi et al., 2026; Gorton et al., 2025; Gujral et al., 2025) offer a promising complementary direction for discovering novel functional features that may not correspond to existing annotations, and could be integrated with the disruption-profiling framework in future work. While the LLM-generated interpretations provide valuable insights, they should be treated as hypotheses requiring expert review rather than definitive clinical evidence.

As genomic foundation models continue to scale and diversify, the combination of accurate prediction, broad variant-type coverage, and structured interpretability demonstrated here provides a practical framework for AI-powered variant classification. We anticipate that integration with proteome-wide disease models (Orenbuch et al., 2025), regulatory variant data from massively parallel reporter assays, and clinical cohort validation will further extend the utility of this approach toward complex and polygenic disease contexts.

References

Avsec, Ž., Weilert, M., Shigaki, D., Karollus, A., et al. (2026). Advancing regulatory variant effect prediction with AlphaGenome. *Nature*, 649:1206–1218. doi: 10.1038/s41586-025-10014-0.

Benegas, G., Batra, S. S., and Song, Y. S. (2023). DNA language models are powerful predictors of genome-wide variant effects. *Proceedings of the National Academy of Sciences*, 120(44): e2311219120. doi: 10.1073/pnas.2311219120.

Boshar, S. et al. (2025). A foundational model for joint sequence-function multi-species modeling at scale for long-range genomic prediction. *bioRxiv*. doi: 10.64898/2025.12.22.695963.

Brandão, R. D., van Roozendaal, K., Tserpelis, D., Gómez García, E. B., and Blok, M. J. (2011). Characterisation of unclassified variants in the BRCA1/2 genes with a putative effect on splicing. *Breast Cancer Research and Treatment*, 129:971–982. doi: 10.1007/s10549-011-1599-7.

Brixi, G., Durairaj, J., Gong, A., Atanackovic, L., Feinstein, G., et al. (2026). Genome modelling and design across all domains of life with Evo 2. *Nature*. doi: 10.1038/s41586-026-10176-5.

Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong, L. H., Zielinski, M., Sargeant, T., et al. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, 381(6664):eadg7492. doi: 10.1126/science.adg7492.

Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Carranza, N. L., Grzywaczewski, A. H., Ober, F., Dufter, P., et al. (2025). Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22:287–297. doi: 10.1038/s41592-024-02523-z.

Dinh, T., Jang, S.-K., Zaitlen, N., and Ntranos, V. (2026). Compressing the collective knowledge of ESM into a single protein language model. *Nature Methods*, pages 1–13. doi: 10.1038/s41592-026-03050-9.

Dooms, T., Wang, N. K., and Pearce, M. (2026). Covariance-based sequence pooling. *Goodfire Research*.

Feng, H. et al. (2025). Benchmarking DNA foundation models for genomic and genetic tasks. *Nature Communications*, 16:10780. doi: 10.1038/s41467-025-10780-0.

Findlay, G. M., Daza, R. M., Martin, B., Zhang, M. D., Leith, A. P., Gasperini, M., Janizek, J. D., Huang, X., Starita, L. M., and Shendure, J. (2018). Accurate classification of BRCA1 variants with saturation genome editing. *Nature*, 562:217–222. doi: 10.1038/s41586-018-0461-z.

Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J. K., Brock, K., Gal, Y., and Marks, D. S. (2021). Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599: 91–95. doi: 10.1038/s41586-021-04043-8.

Funk, J. S. et al. (2025). Deep CRISPR mutagenesis characterizes the functional diversity of TP53 mutations. *Nature Genetics*, 57:140–153. doi: 10.1038/s41588-024-02039-0.

Gorton, L., Wang, N., Nguyen, N., Deng, M., Ho, E., Balsam, D., and McGrath, T. Interpreting Evo 2: Arc Institute's next-generation genomic foundation model. https://www.goodfire.ai/research/interpreting-evo-2, (2025).

Gujral, O., Bafna, M., Alm, E., and Berger, B. (2025). Sparse autoencoders uncover biologically interpretable features in protein language model representations. *Proceedings of the National Academy of Sciences*, 122:e2506316122. doi: 10.1073/pnas.2506316122.

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3):310–315. doi: 10.1038/ng.2892.

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., et al. (2016). Clinvar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, 44(D1):D862–D868. doi: 10.1093/nar/gkv1222.

Lu, B., Liu, X., Lin, P.-Y., and Brandes, N. (2025). Genomic heterogeneity inflates the performance of variant pathogenicity predictions. *bioRxiv*. doi: 10.1101/2025.09.05.674459.

Orenbuch, R. et al. (2025). Proteome-wide model for human disease genetics. *medRxiv*. doi: 10.1101/2023.11.27.23299062.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, 17 (5):405–424. doi: 10.1038/gim.2015.30.

Sahu, S. et al. (2025). Saturation genome editing-based clinical classification of BRCA2 variants. *Nature*, 638:538–545. doi: 10.1038/s41586-024-08318-0.

Schubach, M., Maass, T., Nazaretyan, L., Röner, S., and Kircher, M. (2024). CADD v1.7: using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions. *Nucleic Acids Research*, 52(D1):D1143–D1154. doi: 10.1093/nar/gkad989.

Tabet, D. R. et al. (2025). The functional landscape of coding variation in the familial hypercholesterolemia gene LDLR. *Science*, 391:eady7186. doi: 10.1126/science.ady7186.

Methods

Genomic foundation model

Embeddings. We used the 7B-parameter Evo 2 model (Brixi et al., 2026) to compute (i) per-position log-likelihood scores for variant effect scoring and (ii) intermediate-layer embeddings for representation analysis. Embeddings were extracted from layer 27, selected based on a layer sweep across all 32 blocks showing optimal pathogenicity classification at layers 24–27 (Supplementary Figure S3).

Harvesting. We extract a fixed-size 65k bp window around each mutation site by evaluating 49k bp upstream context and capturing the 16k bp downstream effect. Our experiments show that accuracy increases with window size, indicating there is a long tail of low-range effects captured by Evo 2 (Supplementary Figure S4). The complete downstream is mean-pooled while only the top 256 diverging position's embeddings are kept, measured by cosine similarity between reference and variant. Our experiments show this storage reduction strategy has no impact over keeping a trailing contiguous window after the mutation site for covariance probes (Supplementary Figure S5).

Dataset. Both the reference and variant sequences are harvested in the sense and anti-sense direction. To handle indels, reference and variant sequences are aligned and any unmatched positions are discarded. The extraction used ∼20,000 H100-hours and the resulting dataset has shape $4{,}252{,}870 \times 2 \times 512 \times 4{,}096$ (variants, var/ref, positions, dimensions) stored as bf16, amounting to ∼34 TB.

Baseline methods

AlphaMissense. Pre-computed pathogenicity scores from the AlphaMissense hg38 genome-wide database (Zenodo); available for missense variants on MANE Select canonical transcripts only (Cheng et al., 2023).

CADD. v1.7 PHRED-scaled scores from pre-computed whole-genome SNV files. Only indels $<$50 bp were scored (Schubach et al., 2024).

GPN-MSA. Pre-computed scores were downloaded from HuggingFace (songlab/gpn-msa-hg38-scores) (Benegas et al., 2023).

NTv3. Nucleotide Transformer v3 (650M parameters) embeddings were extracted from layer 11 (1,536-D) over 8,192 bp context windows centered on the variant. L2-regularized logistic regression trained on the element-wise difference of alternative and reference embeddings (alt $-$ ref) using 5-fold gene-holdout cross-validation (Boshar et al., 2025; Dalla-Torre et al., 2025).

AlphaGenome. Variant effect scores were computed using AlphaGenome (google/alphagenome-all-folds) following the published scoring methodology. Reference and alternative gene-body sequences were passed through the model to predict functional genomic tracks across 9 modalities. Scores were aggregated using CenterMask ($\log_2$ fold-change within variant-centered windows) for chromatin and transcription initiation tracks, and GeneMask (max $|\text{ALT} - \text{REF}|$ or log-fold change over the gene body) for expression and splicing tracks. The final score was the maximum across all modality-specific scores.

Clinical variant datasets

All benchmarks use ClinVar release 2026-03-21.

Full dataset. The full dataset used in EVEE preprocesses the whole of ClinVar by removing multi-allelic, MNV, and unannotated entries (14,118 dropped), then validates against GENCODE v49 for gene boundary and reference-allele agreement (128,586 dropped), leaving 4,252,870 variants. These divide into a labeled set (1,539,596) and an unlabeled complement (2,716,001). The labeled variants are pathogenic or likely-pathogenic / benign or likely-benign with a non-zero star review status, ensuring at least one submitter provided assertion criteria. In contrast, unlabeled contains all variants of uncertain significance, conflicting interpretations, and zero-star submissions.

SNV benchmark. Filtered for SNVs in genes $\leq$100 kb with $\geq$1 star review status. 833,970 variants with consequence-type breakdown: missense (n=125,092, 29.0% pathogenic), synonymous (n=292,369, 0.1% pathogenic), nonsense (n=36,050, 98.3% pathogenic), splice (n=21,683, 98.5% pathogenic), UTR (n=31,911, 0.3% pathogenic), intronic

(n=254,737, 0.7% pathogenic), other (n=3,498, 7.4% pathogenic). Gene-level holdout cross-validation ensured no gene appeared in both training and test sets.

Indel benchmark. 73,961 indels from 3,570 genes $\leq$100 kb with $\geq$1 star review status. Stratified by consequence: frameshift (n=41,398, 99% path.), in-frame (n=1,769, 52% path.), noncoding (n=24,250, 2% path.), splice-adjacent (n=2,442, 88% path.); and by size: 1 bp (n=39,801, 70% path.), 2–5 bp (n=21,682, 60% path.), 6–20 bp (n=9,106, 56% path.), $>$20 bp (n=3,372, 68% path.); insertions (n=37,558, 61% path.) and deletions (n=36,403, 70% path.).

Deconfounded benchmark. 158,616 variants with pathogenic and benign labels balanced within each consequence type following the framework of Lu et al. (Lu et al., 2025). The sampling balances two competing objectives: equalizing pathogenic and benign labels within each consequence type, to prevent models from using consequence as a proxy for pathogenicity, and preserving the natural prevalence of consequence types, to avoid overrepresenting rare categories. This is resolved by weighing each consequence type's sampling allocation by both its natural prevalence and its amenability to label balance, then selecting variants within each stratum round-robin over genes. The result is a subset where consequence-only accuracy is 73% (down from 96%).

Deep mutational scanning datasets

BRCA1. Saturation genome editing functional scores from Findlay et al. (Findlay et al., 2018) via MaveDB (urn:mavedb:00000097); 2,803 SNVs spanning 13 exonic regions. Readout: SGE function scores. Variants binarized as functionally abnormal at score < -1.22.

BRCA2. Saturation genome editing functional classification from Sahu et al. (Sahu et al., 2025) via MaveDB (urn:mavedb:00001225); 6,270 SNVs. Readout: SGE function scores. Variants binarized at score < -0.45.

TP53. Deep CRISPR mutagenesis functional scores from Funk et al. (Funk et al., 2025); 1,920 SNVs across exons 5–8. Readout: Enrich2 relative fitness scores. Variants binarized as functionally abnormal at score > -0.97.

LDLR. Functional scores measuring cell-surface abundance and LDL uptake from Tabet et al. (Tabet et al., 2025); 24,999 codon-level variants aggregated to amino acid level by mean pooling. Readout: LDL uptake score. Variants binarized at score < 0.40. Target gene fully excluded from ClinVar training for transfer experiments via gene-holdout cross-validation.

Covariance probe

For each sequence (reference or alternate), Evo 2 produces an $L \times d$ matrix of token embeddings, where L is sequence length of the harvested embeddings and d is embedding dimension. We subtract these position-wise to get the difference $\mathbf{X}$, encoding the variant's effect. Standard mean pooling computes $\bar{\mathbf{x}} = (1/L)\sum \mathbf{x}_i$, yielding a d-dimensional summary. The covariance probe instead uses the Gram matrix $\mathbf{C} = \mathbf{X}^\top \mathbf{X} \in \mathbb{R}^{d \times d}$, which captures pairwise interactions among embedding dimensions accumulated across all sequence positions. Pathogenicity is predicted using a linear classifier on this matrix. This probing setup can be interpreted as a linear attention mechanism that selectively aggregates positional information, rather than unconditionally summing.

Dimensionality. Since the full Gram matrix is too large to use directly as a naive linear probe target, we use a low-dimensional projection $\mathbf{E} = (\mathbf{XL})^\top(\mathbf{XR}) \in \mathbb{R}^{h \times h}$ where $h = 64$ (yielding an equally large embedding as mean pooling, $h^2 = d = 4096$). The result is a low-dimensional linear probe using $\sim$500K parameters (rather than 16M). The embeddings for the forward and backward sequences are concatenated before inputting into the probe.

Training. The covariance probe is trained with the Muon optimizer using a constant learning rate of 0.005, batch size of 512×8 (mini-batch $\times$ shards), weight decay of 0.1 for a single epoch. The labels are ClinVar P/LP versus B/LB, excluding 0-star. We use an 80/20 gene-level split; we do not perform model selection based on hyperparameters. The mean-pooled probe is trained with L2-regularized logistic regression using 5-fold gene-holdout cross-validation. The regularization strength C is selected via inner 5-fold stratified cross-validation over 12 values from 10^{-4} to 10^{1}. Features are mean-centered and scaled to unit variance on the training fold.

Supervised annotation probes

Annotation panel. A panel of 371 biological annotations spanning: protein features (disorder content fraction, secondary structure class, backbone psi angle, solvent accessibility, surface area, pLDDT), structural context (chain membership,

PPI interface, CpG island, disordered region, Ptc region), regulatory features (histone modification breadth: H3K27me3, H3K4me1, H3K4me3, H3K27ac, overall chromatin breadth), protein domains (Pfam/InterPro families), post-translational modifications (glycosylation, phosphorylation), genomic region (segmental duplication identity, transcript copy number, CDS, intron, replication timing period), amino acid substitution properties (hydrophobicity, volume, molecular weight, Grantham distance, BLOSUM62), and variant effects (splice disruption, charge alteration).

Probe training. For each annotation, a covariance probe was trained on Evo 2 layer 27 embeddings at reference-genome positions to predict the annotation value (the covariance of a single position is the outer product of the embeddings). Annotations were treated as either binary (classification, evaluated by AUC), categorical (classification, evaluated by accuracy) or continuous (soft binned classification evaluated by Spearman correlation). Training used $\sim$100M positions from $\sim$18k genes, with 20/80 gene holdout.

Disruption scoring. For each variant, disruption was computed as the change in predicted annotation between alternate and reference sequences. Let f_a be the probe prediction for annotation a; the disruption is $\Delta_a = f_a(\mathbf{E}_{\text{alt}}) - f_a(\mathbf{E}_{\text{ref}})$.

LLM-based interpretation pipeline

For each variant, the top-k disruptions (ranked by $|\Delta|$) were assembled alongside variant metadata (gene, transcript, consequence, genomic position, ClinVar status, gnomAD frequency, LOEUF score) into a structured prompt. Claude Sonnet 4.6 (Anthropic, 2026) was used to synthesize a natural-language interpretation including: a one-sentence summary of the predicted molecular mechanism, a detailed explanation contextualizing the disruption profile against known biology, calibration information (percentage of ClinVar pathogenic variants at similar pathogenicity scores), and bullet-pointed key evidence.

Evaluation. To evaluate the effect of contextual information on LLM-generated variant interpretations, we conducted a context stratification experiment across 154 expert-reviewed ClinVar variants (3+ stars). We defined a cumulative sequence of context configurations, each adding information to the previous level: (1) genomic coordinates only; (2) gene name; (3) gene name plus protein domain annotations, loss-of-function intolerance (LOEUF) scores, gnomAD allele frequencies, and gene function summaries; (4) HGVS protein and coding sequence notation; and (5) Evo 2 foundation model predictions including amino acid substitution probabilities, biophysical property scores (BLOSUM62, Grantham distance, charge, hydrophobicity, volume), splice disruption scores, and disruption profiles across nearby genomic positions. Interpretations were generated by three Claude model tiers (Haiku 4.5, Sonnet 4.6, and Opus 4.6) at temperature 0 with a maximum of 1,024 tokens. Each interpretation was evaluated by an LLM judge (Claude Opus 4.6) on three axes using a 1–5 Likert scale: mechanism coverage using a five-level specificity rubric ranging from gene-level identification (level 1) to multi-mechanism reasoning with biophysical detail (level 5); and biological accuracy and specificity by comparison against expert-submitted ClinVar evidence (SCV records). A composite score was computed as the unweighted mean of the three axes. All judge evaluations used temperature 0 for reproducibility. Confidence intervals represent 95% percentiles.

Evo Variant Effect Explorer (EVEE)

Pre-computed predictions, disruption profiles, and on-demand LLM-based interpretations for 4.2 million ClinVar variants are provided through an interactive web application at `https://evee.goodfire.ai/`.

Data and code availability

Computational methods, including implementations of the covariance probe, and scripts to reproduce the main figures in this paper are publicly available via GitHub at `https://github.com/goodfire-ai/evee-manuscript`.

Anthropic. Introducing Sonnet 4.6. https://www.anthropic.com/news/claude-sonnet-4-6, (2026).

Benegas, G., Batra, S. S., and Song, Y. S. (2023). DNA language models are powerful predictors of genome-wide variant effects. *Proceedings of the National Academy of Sciences*, 120(44):e2311219120. doi: 10.1073/pnas.2311219120.

Boshar, S. et al. (2025). A foundational model for joint sequence-function multi-species modeling at scale for long-range genomic prediction. *bioRxiv*. doi: 10.64898/2025.12.22.695963.

Brixi, G., Durairaj, J., Gong, A., Atanackovic, L., Feinstein, G., et al. (2026). Genome modelling and design across all domains of life with Evo 2. *Nature*. doi: 10.1038/s41586-026-10176-5.

Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong, L. H., Zielinski, M., Sargeant, T., et al. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, 381(6664):eadg7492. doi: 10.1126/science.adg7492.

Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Carranza, N. L., Grzywaczewski, A. H., Ober, F., Dufter, P., et al. (2025). Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22:287–297. doi: 10.1038/s41592-024-02523-z.

Findlay, G. M., Daza, R. M., Martin, B., Zhang, M. D., Leith, A. P., Gasperini, M., Janizek, J. D., Huang, X., Starita, L. M., and Shendure, J. (2018). Accurate classification of BRCA1 variants with saturation genome editing. *Nature*, 562:217–222. doi: 10.1038/s41586-018-0461-z.

Funk, J. S. et al. (2025). Deep CRISPR mutagenesis characterizes the functional diversity of TP53 mutations. *Nature Genetics*, 57:140–153. doi: 10.1038/s41588-024-02039-0.
Lu, B., Liu, X., Lin, P.-Y., and Brandes, N. (2025). Genomic heterogeneity inflates the performance of variant pathogenicity predictions. *bioRxiv*. doi: 10.1101/2025.09.05.674459.
Sahu, S. et al. (2025). Saturation genome editing-based clinical classification of BRCA2 variants. *Nature*, 638:538–545. doi: 10.1038/s41586-024-08318-0.
Schubach, M., Maass, T., Nazaretyan, L., Röner, S., and Kircher, M. (2024). CADD v1.7: using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions. *Nucleic Acids Research*, 52(D1):D1143–D1154. doi: 10.1093/nar/gkad989.
Tabet, D. R. et al. (2025). The functional landscape of coding variation in the familial hypercholesterolemia gene LDLR. *Science*, 391:eady7186. doi: 10.1126/science.ady7186.

Supplementary Figures

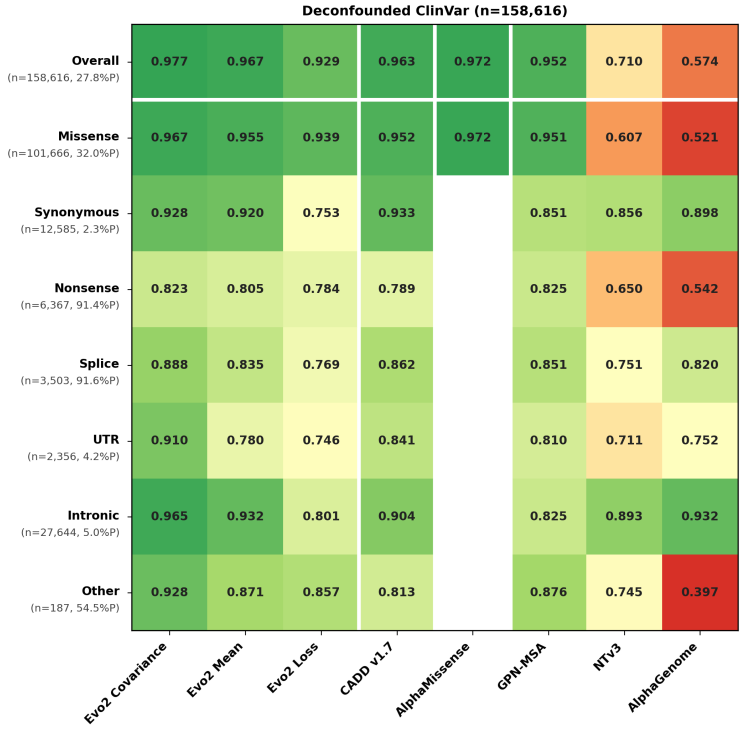

Figure S1. Deconfounded ClinVar benchmark. AUROC by consequence type on a deconfounded benchmark of 158,616 variants with pathogenic and benign labels balanced within each consequence type (Lu et al., 2025). Evo 2 covariance probe (0.971 overall) is matched only by AlphaMissense (0.972) while exceeding CADD v1.6 (0.911), CADD v1.7 (0.966), Evo 2 loss (0.932), GPN-MSA (0.952), and NTv3 (0.586).

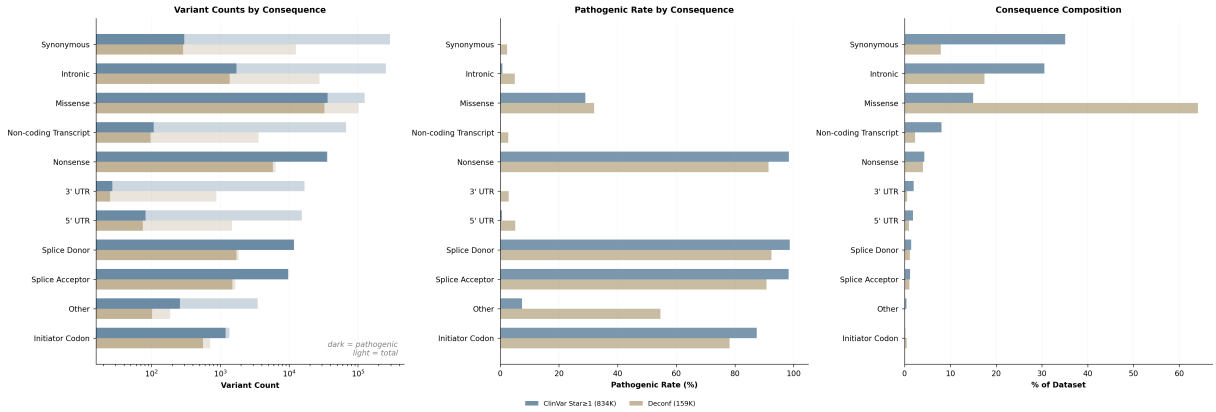

Figure S2. Deconfounded dataset characterization. Variant counts for each consequence type in ClinVar with ≥1 star (gold) and the deconfounded set (grey) for all variants (light bars) and pathogenic variants (dark bars), showing the downsampling rates. Overall pathogenic rates and composition fraction by consequence in the two datasets.

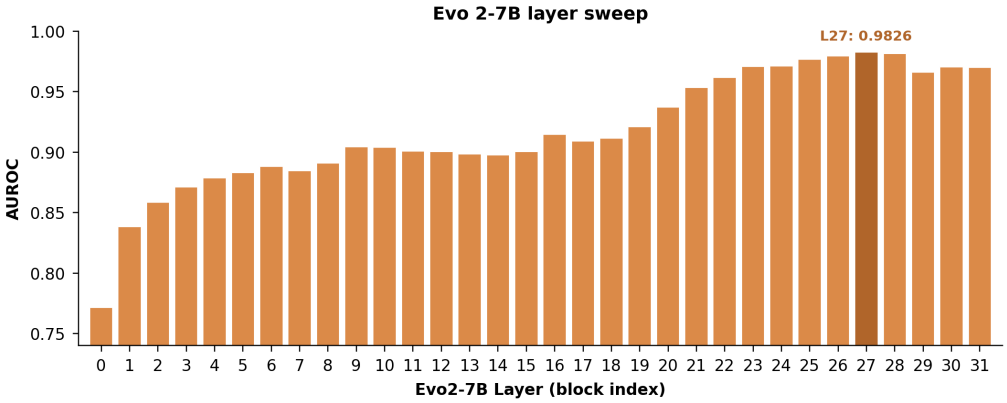

Figure S3. Layer sweep. Layer sweep across all 32 Evo 2-7B blocks showing pathogenicity classification AUROC by layer. Optimal performance at layer 27.

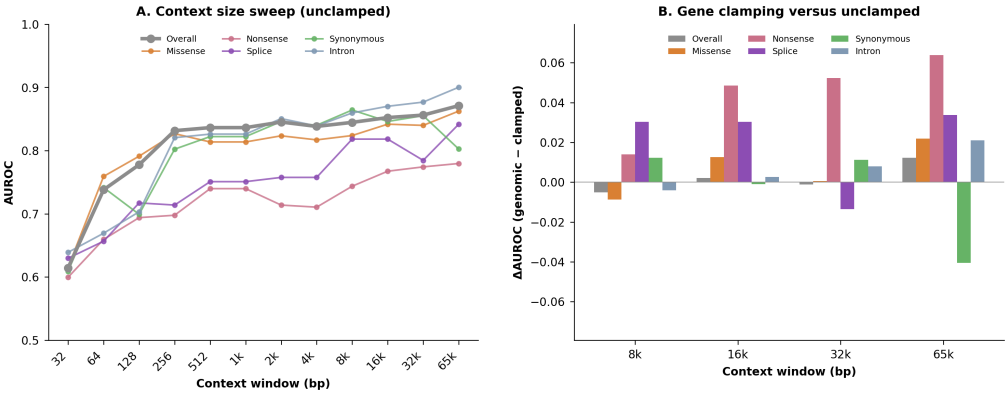

Figure S4. Context window sweep. Context sweeps at layer 27 of Evo 2-7B. Pathogenicity prediction improves sharply with increasing window size below 256 bp, then continues more gradually across all sizes tested. Including flanking sequence beyond gene boundaries outperforms restricting input to the gene body, suggesting Evo 2 captures regulatory signals.

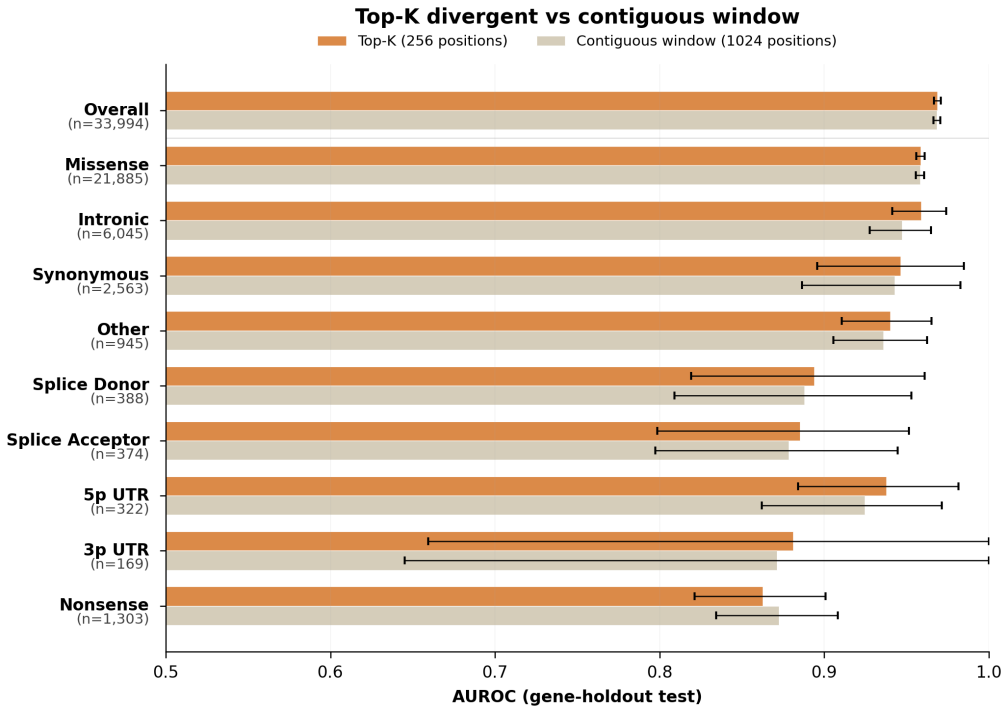

Figure S5. Storage ablation. Comparison of two embedding storage methods: keeping a contiguous window trailing the mutation site or keeping the top-k most divergent positions between reference and alternate. Both perform roughly equally well measured by pathogenicity prediction. However, top-k better captures far-away secondary effects.

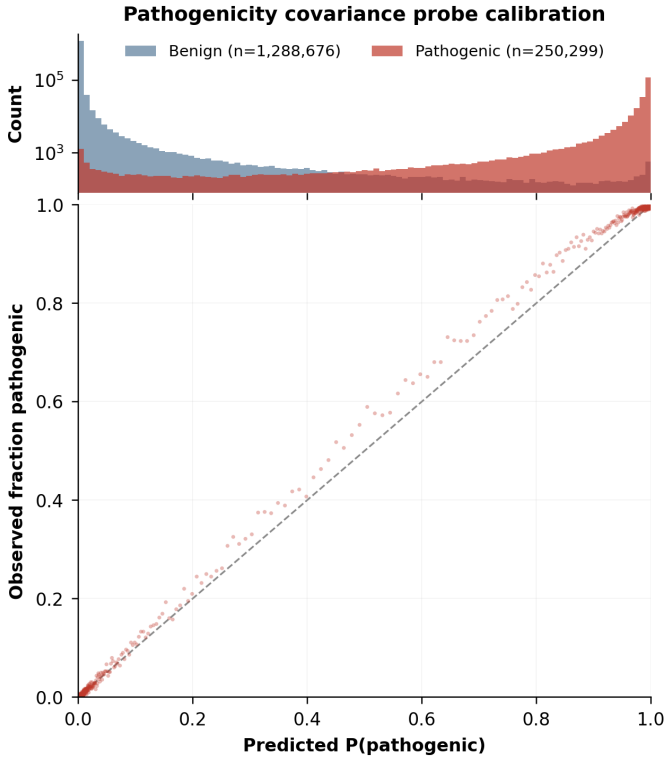

Figure S6. Detailed classification performance. Calibration curve. The pathogenicity probe is well-calibrated overall, especially for confidently benign classifications. It hedges slightly for confidently pathogenic predictions, which is caused by the class imbalance.