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 J. 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) [1, 2]. Computational variant effect predictors have proliferated, spanning protein language models such as AlphaMissense [3] and EVE [4], supervised sequence-to-activity models such as AlphaGenome [5] and Nucleotide Transformer v3 [6, 7], evolutionary DNA models such as Evo 2 [8] and GPN-MSA [9], compressed protein language models [10], and classical meta-predictors such as CADD [11]. However, these tools face complementary limitations: protein-based methods are restricted to missense variants; AlphaGenome focuses on regulatory non-coding effects; and CADD, while genome-wide, integrates 100+ annotation features with crossed terms and rank transformations that obscure individual feature contributions [5, 11, 12]. Critically, none provide human-readable explanations of why a variant is predicted pathogenic—a key requirement for clinical variant classification under ACMG/AMP guidelines [2], which demands categorized evidence rather than opaque scores.

Genomic foundation models, large neural networks pre-trained on DNA sequences, learn representations that capture functional constraints across coding and non-coding regions [7, 8, 13]. 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 [8]. 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 prediction 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 se-

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

quences through Evo 2, computes the difference in per-position embeddings, and trains a covariance probe on the resulting representation to predict pathogenicity (Figure 1a; Supplementary Figures S1–S3). 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 [13, 14] (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 Figures S4,S5), 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) [15]. 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. The resulting probe produces well-calibrated pathogenicity estimates (Supplementary Figure S6).

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 [16], BRCA2 [17], TP53 [18], and LDLR [19] (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. Together, these results demonstrate that Evo 2 embeddings generalize 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,

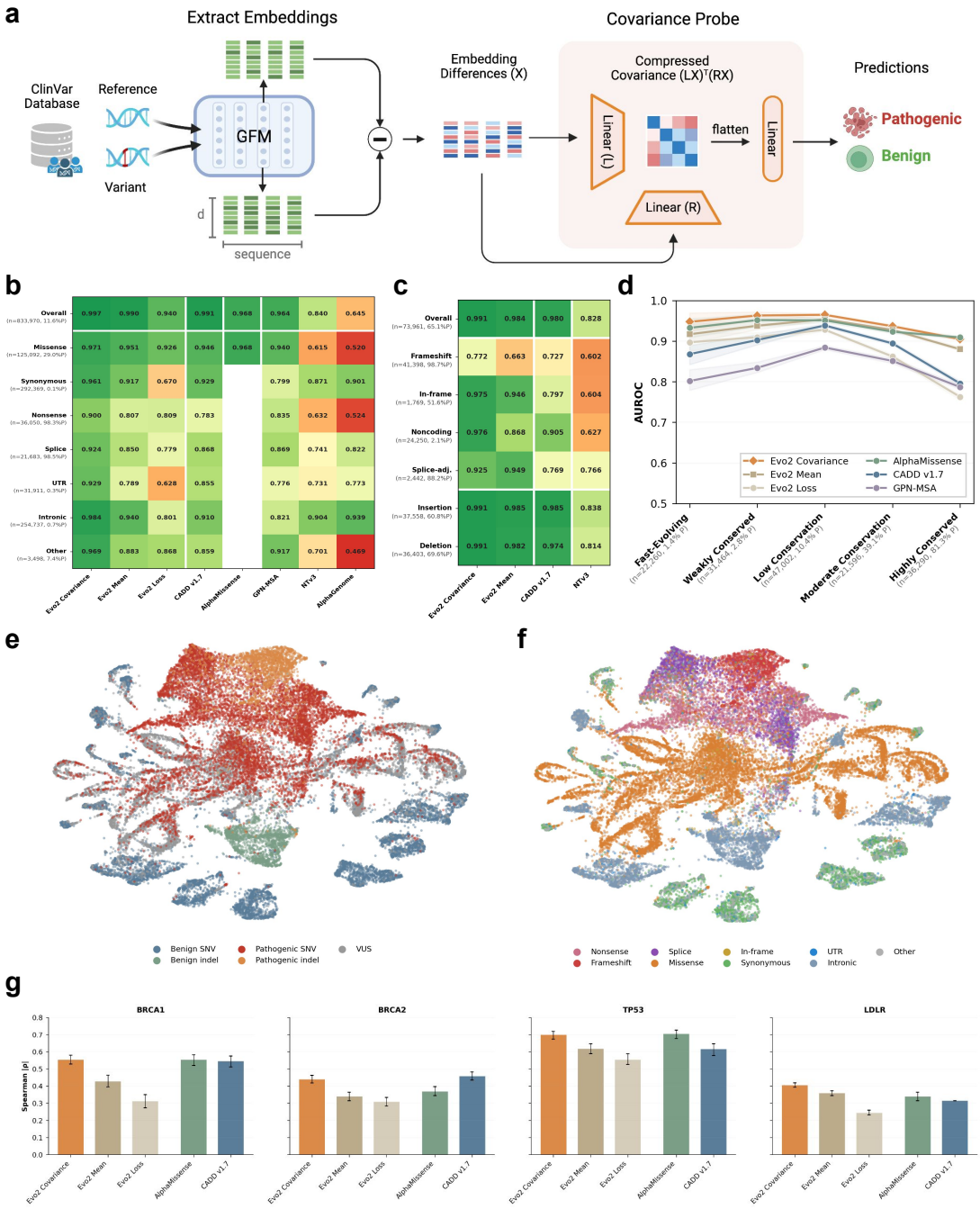

Figure 1. Evo 2 covariance probe achieves state-of-the-art variant effect prediction across variant types. (**a**) Experimental design. Evo 2 processes reference and alternate DNA sequences, producing per-position embeddings. The difference between the embeddings are then input into a covariance probe that is trained to predict pathogenicity. For each input, the covariance probe computes a compressed covariance matrix 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 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 on 73,961 ClinVar indels stratified by consequence type and insertion vs. deletion. (**d**) Performance robustness across conservation levels. AUROC by phyloP100way conservation tier. Evo 2 probes and AlphaMissense maintain performance across 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. Error bars show 95% bootstrap confidence intervals.

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 annotation probes trained on Evo 2's internal representations. Each probe predicts a biological annotation (protein domains, post-translational modifications, splice sites, secondary structure, amino acid identity, and biophysical properties) at each genomic position, and we measure the change in each prediction between the variant and reference sequences as a raw delta. The probes operate 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. The resulting 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 [20]. The framework captures diverse mechanisms beyond splicing. For an LDLR nonsense variant (p.Ser306Ter), the system identifies a premature termination codon that truncates the protein before the EGF precursor homology domain, detecting coordinated loss of EGF-like domain, disulfide bond, and downstream structural features consistent with elimination of the receptor recycling machinery required for LDL uptake. For a RET missense variant (p.Gly798Ser), the system detects near-complete loss of kinase domain identity and catalytic domain signatures, correctly attributing pathogenicity to disruption of the backbone flexibility at a conserved glycine position within the kinase fold (Supplementary Figures S7,S8).

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 1.31), adding gene name (+0.81), gene context such as constraint scores and gene function (+0.22), and HGVS protein notation (+0.43) each improved quality, but the largest gain came from adding Evo 2 probe predictions (+1.03), yielding a composite score of 3.80/5 (mechanism=3.24, biological accuracy=4.05, specificity=4.13) with Claude Sonnet 4.6 as the interpreter (Supplementary Figures S9,S10).

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 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 [2]. 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.

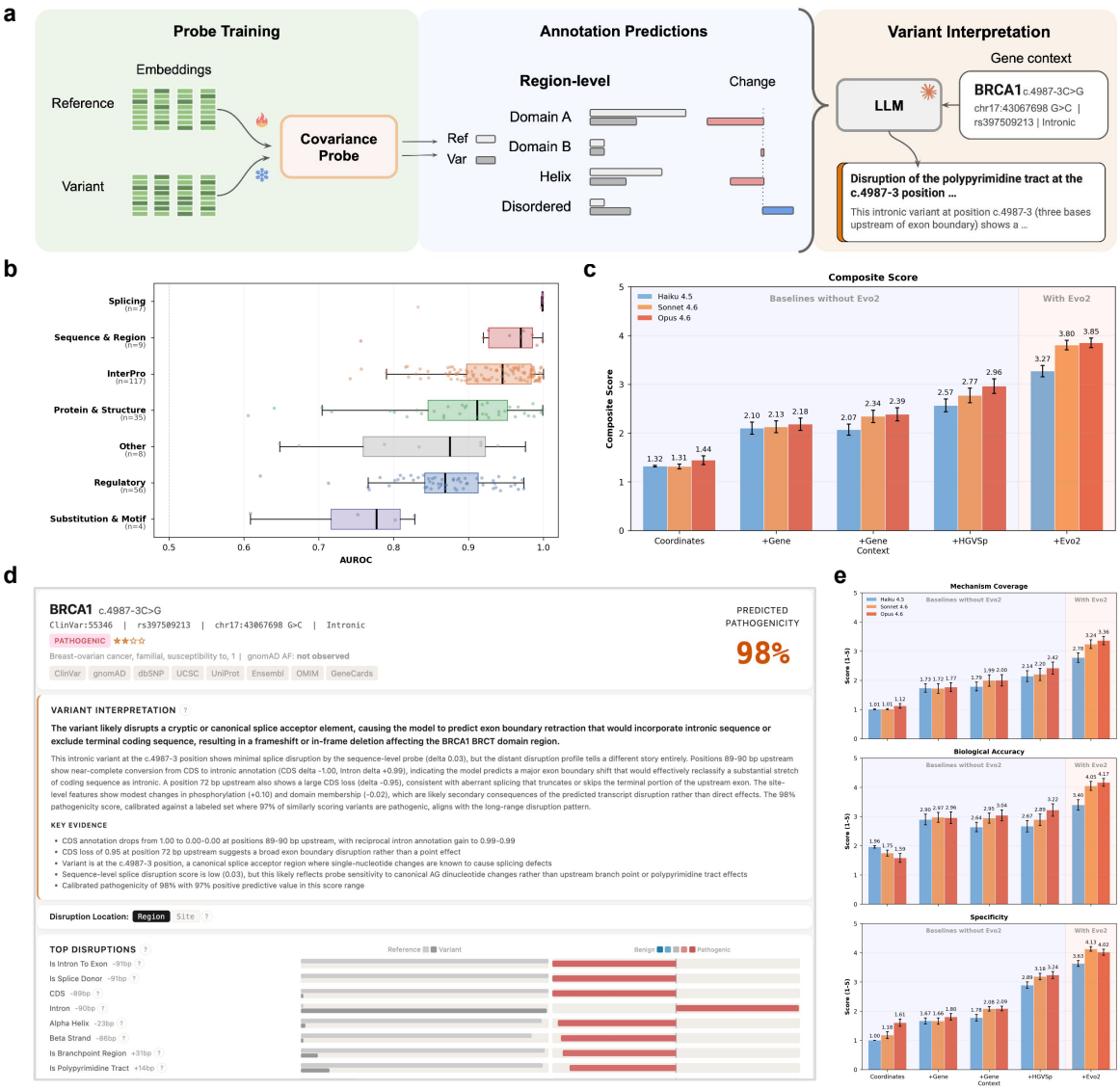

Figure 2. Interpretable predictions through annotation disruption profiling and LLM-based synthesis. (**a**) Schematic of the interpretability framework. Supervised annotation probes are trained on Evo 2 reference-sequence embeddings to predict a panel of biological annotations. For each variant, disruptions (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.

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 [8, 21, 22] 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 [23], 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.

ACKNOWLEDGEMENTS

We thank the Arc Institute and the Evo 2 team for making model weights publicly available, and the ClinVar and ClinGen communities for maintaining openly accessible variant interpretation resources.

AUTHOR CONTRIBUTIONS

M.T.P. and N.K.W. conceived the study. T.D. developed the covariance probe methodology. M.T.P., T.D., and R.Y. designed and implemented the annotation probe and interpretation framework. J.M., C.M., C.O., P.D., B.T., E.K., E.M., A.R., S.A., P.K., and M.R. provided clinical genomics expertise and variant interpretation guidance. M.B., D.H., and C.F. contributed to model development and evaluation. N.N. and M.A. contributed to infrastructure support for the web app. A.J. and D.B. contributed to project coordination and infrastructure. N.K.W. supervised the project. M.T.P., T.D., R.Y., and N.K.W. wrote the manuscript with input from all authors.

COMPETING FINANCIAL INTERESTS

M.T.P., T.D., R.Y., M.B., D.H., C.F., N.N., M.A., A.J., D.B., and N.K.W. are employees of Goodfire. The remaining authors declare no competing interests.

References

1. Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Jeffrey Hoover, et al. Clinvar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, 44(D1):D862–D868, 2016. doi: 10.1093/nar/gkv1222.
2. Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, et al. 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, 2015. doi: 10.1038/gim.2015.30.
3. Jun Cheng, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, 381(6664):eadg7492, 2023. doi: 10.1126/science.adg7492.
4. Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K Min, Kelly Brock, Yarin Gal, and Debora S Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599:91–95, 2021. doi: 10.1038/s41586-021-04043-8.
5. Žiga Avsec, Melanie Weilert, Dustin Shigaki, Alexander Karollus, et al. Advancing regulatory variant effect prediction with AlphaGenome. *Nature*, 649:1206–1218, 2026. doi: 10.1038/s41586-025-10014-0.
6. S Boshar et al. A foundational model for joint sequence-function multi-species modeling at scale for long-range genomic prediction. *bioRxiv*, 2025. doi: 10.64898/2025.12.22.695963.
7. Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Ober, Philipp Dufter, et al. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22:287–297, 2025. doi: 10.1038/s41592-024-02523-z.
8. Garyk Brixi, Janani Durairaj, Andrew Gong, Luka Atanackovic, Greg Feinstein, et al. Genome modelling and design across all domains of life with Evo 2. *Nature*, 2026. doi: 10.1038/s41586-026-10176-5.
9. Gonzalo Benegas, Sanjit Singh Batra, and Yun S Song. DNA language models are powerful predictors of genome-wide variant effects. *Proceedings of the National Academy of Sciences*, 120(44):e2311219120, 2023. doi: 10.1073/pnas.2311219120.
10. Thanh Dinh, Se-Kwon Jang, Noah Zaitlen, and Vasilis Ntranos. Compressing the collective knowledge of ESM into a single protein language model. *Nature Methods*, pages 1–13, 2026. doi: 10.1038/s41592-026-03050-9.
11. Max Schubach, Thorben Maass, Lusiné Nazaretyan, Sebastian Röner, and Martin Kircher. 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, 2024. doi: 10.1093/nar/gkad989.
12. Martin Kircher, Daniela M Witten, Preti Jain, Brian J O'Roak, Gregory M Cooper, and Jay Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3):310–315, 2014. doi: 10.1038/ng.2892.
13. Hao Feng et al. Benchmarking DNA foundation models for genomic and genetic tasks. *Nature Communications*, 16:10780, 2025. doi: 10.1038/s41467-025-10780-0.
14. Bing Lu, Xiaoming Liu, Pin-Yu Lin, and Nadav Brandes. Genomic heterogeneity inflates the performance of variant pathogenicity predictions. *bioRxiv*, 2025. doi: 10.1101/2025.09.05.674459.
15. Thomas Dooms, Nicholas K Wang, and Michael Pearce. Covariance-based sequence pooling. *Goodfire Research*, 2026.
16. Gregory M Findlay, Riza M Daza, Beth Martin, Melissa D Zhang, Anh P Leith, Molly Gasperini, Joseph D Janizek, Xingfan Huang, Lea M Starita, and Jay Shendure. Accurate classification of BRCA1 variants with saturation genome editing. *Nature*, 562:217–222, 2018. doi: 10.1038/s41586-018-0461-z.
17. Satyabrata Sahu et al. Saturation genome editing-based clinical classification of BRCA2 variants. *Nature*, 638:538–545, 2025. doi: 10.1038/s41586-024-08318-0.
18. Jonas S Funk et al. Deep CRISPR mutagenesis characterizes the functional diversity of TP53 mutations. *Nature Genetics*, 57:140–153, 2025. doi: 10.1038/s41588-024-02039-0.
19. Daniel R Tabet et al. The functional landscape of coding variation in the familial hypercholesterolemia gene LDLR. *Science*, 391:eady7186, 2025. doi: 10.1126/science.ady7186.
20. Raquel D Brandão, Kim van Roozendaal, Demis Tserpelis, Encarna B Gómez García, and Marinus J Blok. Characterisation of unclassified variants in the BRCA1/2 genes with a putative effect on splicing. *Breast Cancer Research and Treatment*, 129:971–982, 2011. doi: 10.1007/s10549-011-1599-7.
21. Liv Gorton, Nicholas Wang, Nam Nguyen, Myra Deng, Eric Ho, Daniel Balsam, and Thomas McGrath. Interpreting Evo 2: Arc Institute's next-generation genomic foundation model. https://www.goodfire.ai/research/interpreting-evo-2, 2025.
22. Olivia Gujral, Mukund Bafna, Eric Alm, and Bonnie Berger. Sparse autoencoders uncover biologically interpretable features in protein language model representations. *Proceedings of the National Academy of Sciences*, 122:e2506316122, 2025. doi: 10.1073/pnas.2506316122.
23. Rose Orenbuch et al. Proteome-wide model for human disease genetics. *medRxiv*, 2025. doi: 10.1101/2023.11.27.23299062.
24. Anthropic. Introducing Sonnet 4.6. https://www.anthropic.com/news/claude-sonnet-4-6, 2026.

Methods

Genomic foundation model

Embeddings. We used the 7B-parameter Evo 2 model [8] 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 (Supplementary Figure S1).

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 S2). 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 S3).

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 $\sim$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 $\sim$34 TB.

Baseline methods

AlphaMissense. Pre-computed pathogenicity scores from the AlphaMissense hg38 genome-wide database available for missense variants on 'MANE Select' canonical transcripts only [3].

CADD. v1.7 PHRED-scaled scores from pre-computed whole-genome SNV files. Only indels $<$50 bp were scored [11].

GPN-MSA. Pre-computed scores were downloaded from HuggingFace (songlab/gpn-msa-hg38-scores) [9].

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 [6, 7].

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. [14]. 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%) (Supplementary Figure S5).

Deep mutational scanning datasets

BRCA1. Saturation genome editing functional scores from Findlay et al. [16] 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. [17] 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. [18]; 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. [19]; 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 ∼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. The resulting pathogenicity probe is well-calibrated overall (Supplementary Figure S6).

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 name, consequence type, genomic position, LOEUF constraint score) into a structured prompt. Claude Sonnet 4.6 [24] 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.

The system prompt (Table S1) describes the disruption signal framework: a probe trained on reference genome activations is run on both reference and variant sequences, producing for each biological feature (helix, conservation, domain membership, etc.) a reference prediction, a variant prediction, and a delta (var $-$ ref). The prompt explains that large negative deltas indicate feature disruption while near-zero deltas indicate preservation, and instructs the LLM to lead with the disruption story, use gene function context, and avoid referencing external clinical prediction tools.

The user prompt is assembled programmatically from variant-specific data. It includes: gene name and variant coordinates, consequence type, HGVSp/HGVSc notation, protein domain annotations with LOEUF constraint score, gene function summaries, calibrated pathogenicity percentage, a Region Disruptions table (peak probe signal change across $\pm$32 kb for the top-10 features by $|\Delta| \geq 0.2$, with reference, variant, delta, and distance columns), and a Site Disruptions table (probe signal change at the variant position for the top-10 features by $|\Delta| \geq 0.1$, with reference, variant, and delta columns). An example of the full prompt is shown in Table S2.

Interpretation 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 ($\geq$3 stars). Five cumulative context configurations were tested, each adding information to all previous levels: (1) genomic coordinates only, (2) +gene name, (3) +gene context (domain annotations, LOEUF constraint, gene function summaries), (4) +HGVSp/HGVSc notation, and (5) +Evo 2 disruption profile (calibrated pathogenicity, region and site disruption tables). Configurations without Evo 2 features used a baseline system prompt; the Evo 2 configuration used the probe-aware system prompt described above. Example prompts for each configuration are shown in Table S2.

Interpretations were generated by three Claude model tiers (Haiku 4.5, Sonnet 4.6, and Opus 4.6). Each interpretation was evaluated by two independent LLM judges (Claude Opus 4.6, Table S3) on three axes using a 1–5 scale: mechanism coverage using a five-level specificity rubric; 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. Confidence intervals represent 95% percentiles. Supplementary Figure S9 shows the composite score stratified by pathogenic and benign variants, and Supplementary Figure S10 shows the per-axis score distributions across context configurations.

Example interpretations. Tables S4–S8 show representative LLM interpretations at each mechanism score level (1–5) at the Evo 2 context configuration, illustrating the scoring rubric in practice.

Limitations of LLM-as-judge evaluation. Several limitations of this evaluation framework should be noted. First, the ClinVar expert panel evidence used as ground truth varies considerably in mechanistic detail. Some submissions provide extensive molecular reasoning (e.g., the CFTR F508del panel describes misfolding, ER retention, and proteasomal degradation), while others rely primarily on population frequency and computational predictor scores with minimal mechanistic discussion (e.g., the PAH p.Leu213Pro submissions cite REVEL scores and compound heterozygosity but do not describe

the structural basis of pathogenicity). This asymmetry means the judge may underestimate mechanism coverage when the interpretation correctly identifies a mechanism that the expert panel simply did not discuss. Second, well-characterized variants in ClinVar (e.g., CFTR F508del, BRCA1 c.5266dupC) are likely present in the LLM's training data, meaning the interpreting model may draw on memorized knowledge rather than reasoning from the disruption profile alone. The context ablation design partially controls for this, since memorized variants should score well even without Evo 2 features, but subtler forms of memorization cannot be fully ruled out. Third, as with any LLM-as-judge approach, the judge model may have its own biases, and its ability to verify biological claims is limited to what is present in the provided ClinVar evidence.

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.

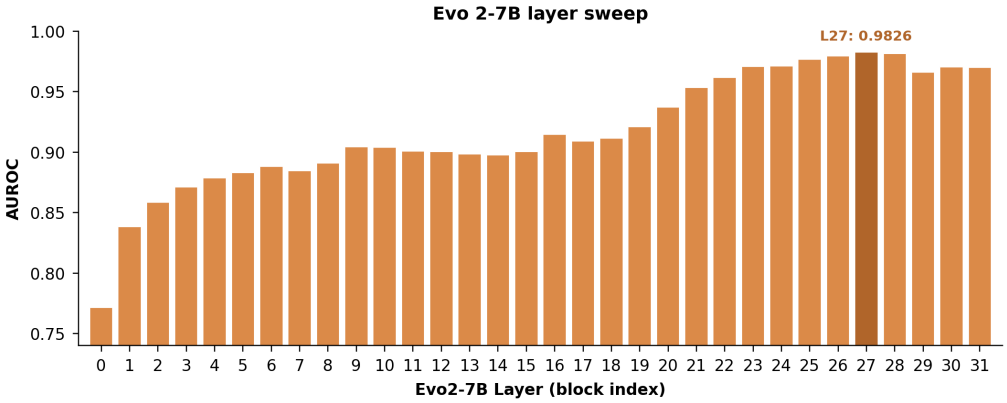

Figure S1. Layer sweep. Layer sweep across all 32 Evo 2-7B blocks showing pathogenicity classification AUROC by layer. Optimal performance at layer 27.

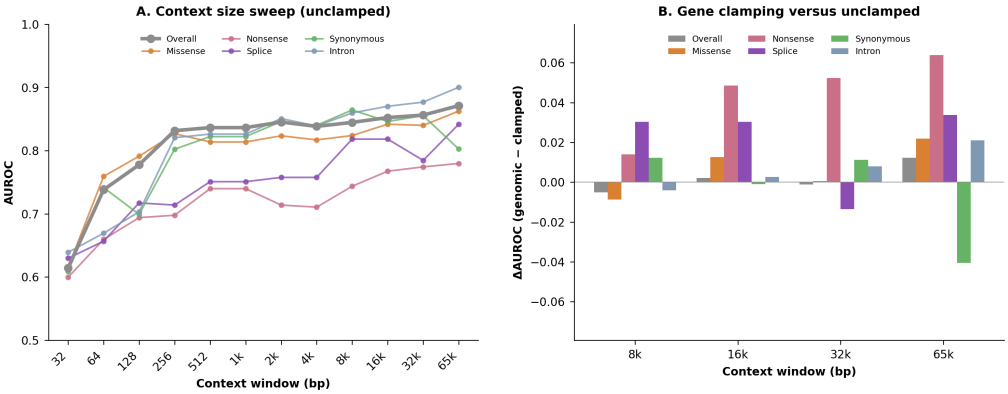

Figure S2. Context window sweep. Context sweeps at layer 27 of Evo 2-7B. Pathogenicity prediction improves sharply with increasing window size below 256 bp, then continues more gradually across all sizes tested. Including flanking sequence beyond gene boundaries outperforms restricting input to the gene body, suggesting Evo 2 captures regulatory signals.

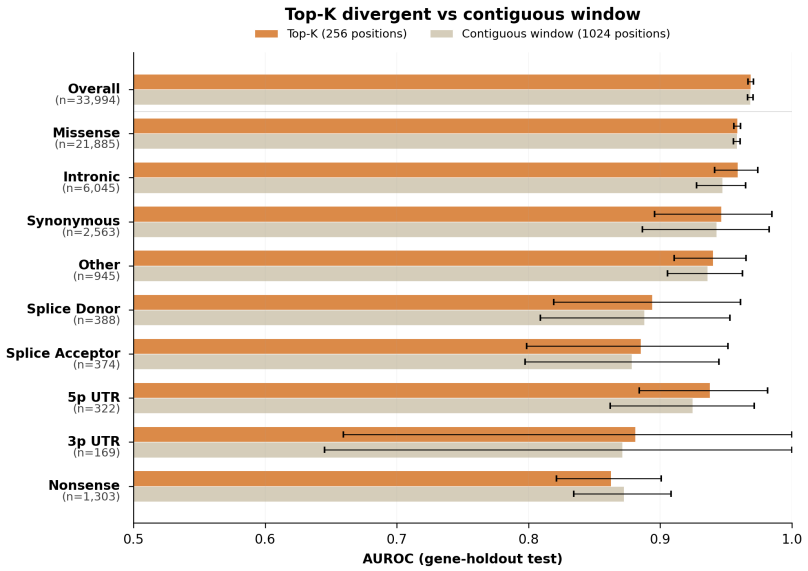

Figure S3. Storage ablation. Comparison of two embedding storage methods: keeping a contiguous window trailing the mutation site or keeping the top-k most divergent positions between reference and alternate. Both perform roughly equally well measured by pathogenicity prediction. However, top-k better captures far-away secondary effects.

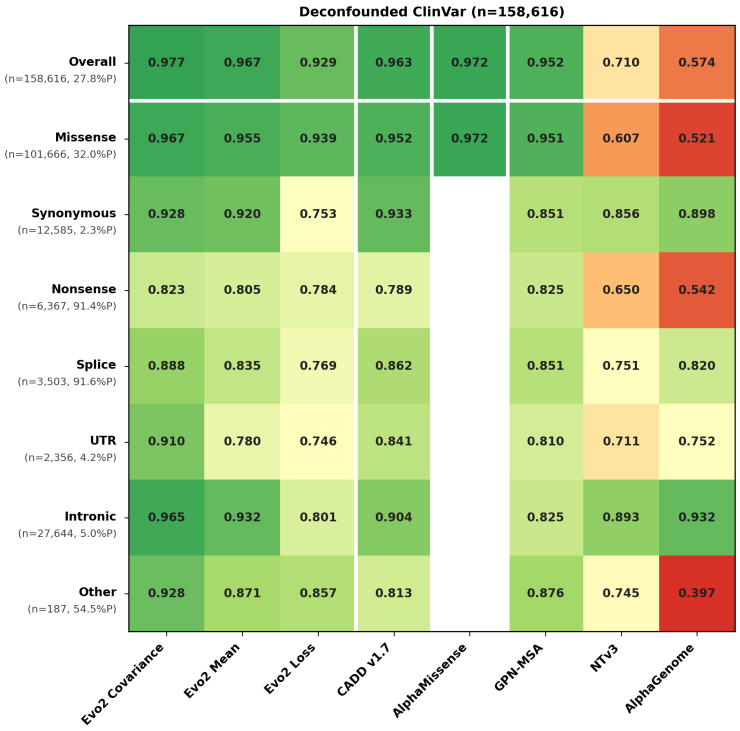

Figure S4. Deconfounded ClinVar benchmark. AUROC by consequence type on a deconfounded benchmark of 158,616 variants with pathogenic and benign labels balanced within each consequence type [14]. Evo 2 covariance probe (0.971 overall) is matched only by AlphaMissense (0.972) while exceeding CADD v1.6 (0.911), CADD v1.7 (0.966), Evo 2 loss (0.932), GPN-MSA (0.952), and NTv3 (0.586).

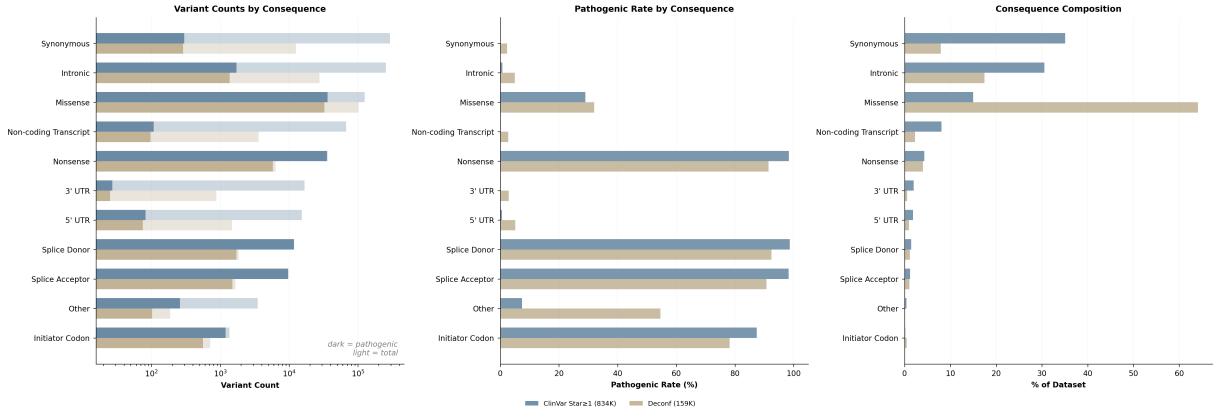

Figure S5. Deconfounded dataset characterization. Variant counts for each consequence type in ClinVar with ≥1 star (gold) and the deconfounded set (grey) for all variants (light bars) and pathogenic variants (dark bars), showing the downsampling rates. Overall pathogenic rates and composition fraction by consequence in the two datasets.

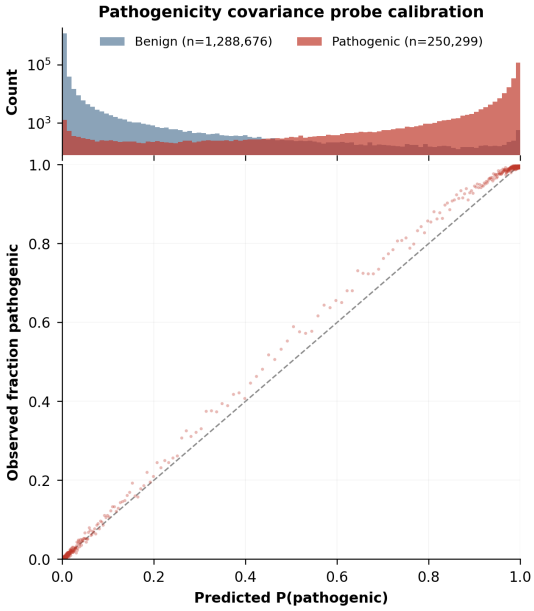

Figure S6. Detailed classification performance. Calibration curve. The pathogenicity probe is well-calibrated overall, especially for confidently benign classifications. It hedges slightly for confidently pathogenic predictions, which is caused by the class imbalance.

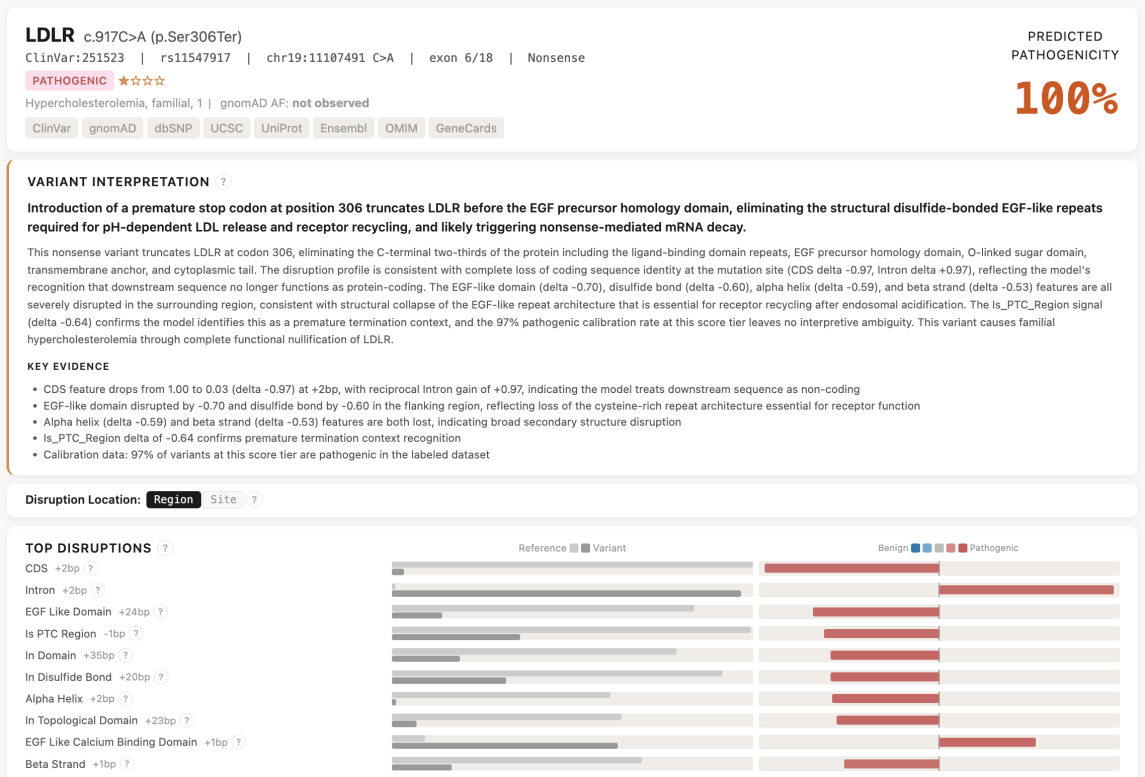

Figure S7. EVEE interpretation for LDLR nonsense variant. LDLR c.917C>A (p.Ser306Ter), predicted 100% pathogenic. The system detects a premature termination codon truncating the protein before the EGF precursor homology domain, with coordinated loss of EGF-like domain ($\Delta = -0.70$), disulfide bond ($\Delta = -0.60$), and alpha helix ($\Delta = -0.89$) features. The interpretation correctly links these disruptions to loss of the receptor recycling machinery required for LDL uptake.

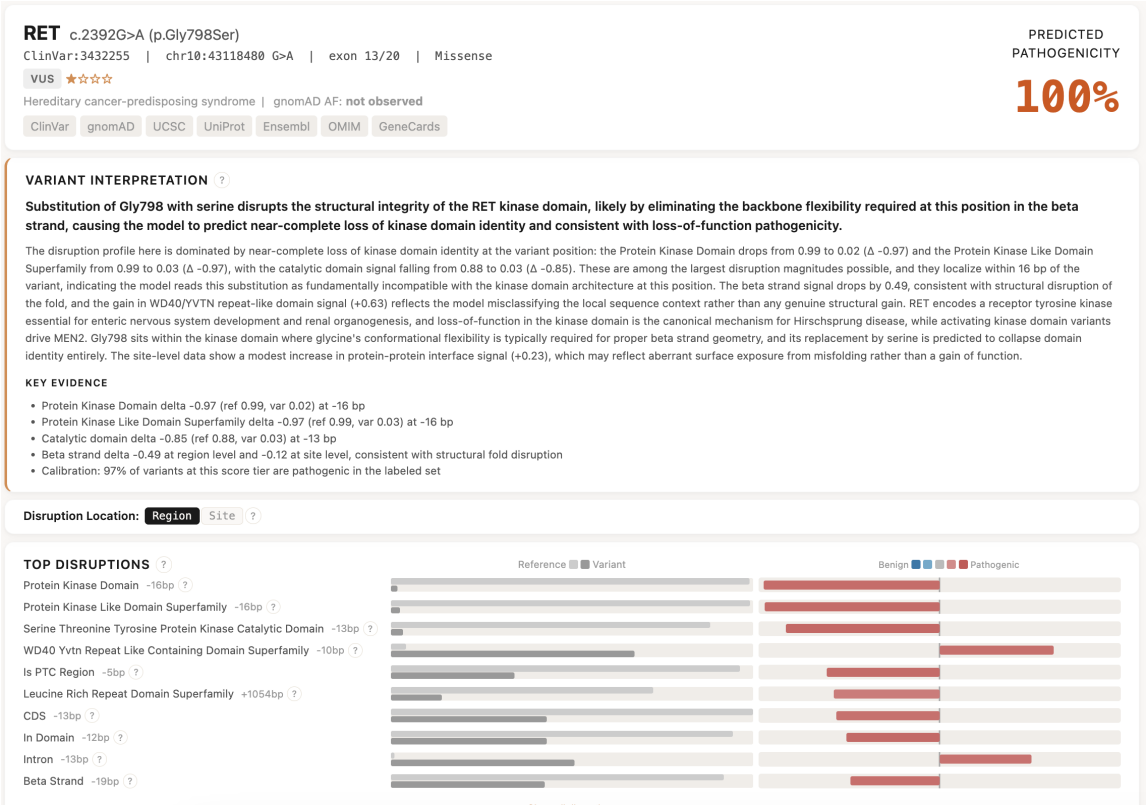

Figure S8. EVEE interpretation for RET missense variant. RET c.2392G>A (p.Gly798Ser), predicted 100% pathogenic. The system detects near-complete loss of protein kinase domain ($\Delta = -0.97$) and catalytic domain ($\Delta = -0.65$) annotations with concurrent beta strand disruption ($\Delta = -0.48$). The interpretation correctly identifies that glycine's backbone flexibility is required at this kinase fold position, and serine substitution causes structural disruption consistent with loss of function.

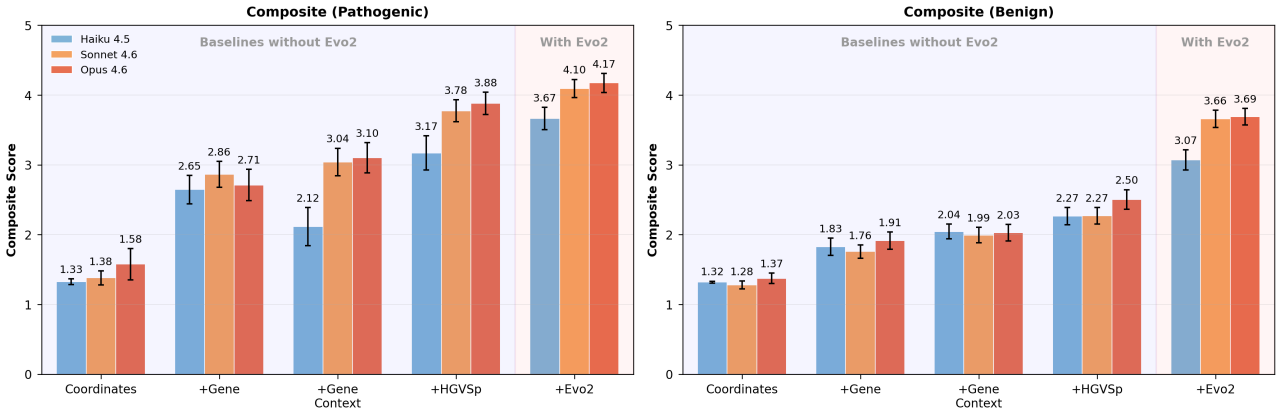

Figure S9. Interpretation quality by pathogenicity class. Composite scores across five context configurations, stratified by pathogenic (left) and benign (right) variants. Three Claude model tiers are compared (Haiku 4.5, Sonnet 4.6, Opus 4.6). Error bars show 95% confidence intervals. Evo 2 predictions provide the largest quality gain for both classes, with pathogenic variants reaching higher absolute scores due to richer mechanistic signals in the disruption profile.

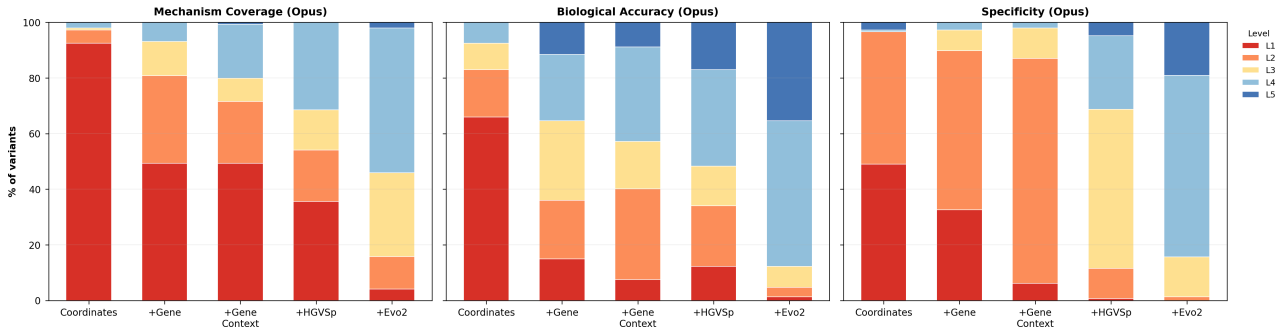

Figure S10. Score distributions across context configurations. Stacked bar charts showing the proportion of variants at each score level (1–5) for mechanism coverage (left), biological accuracy (center), and specificity (right), using Claude Opus 4.6 as the interpreter. Adding Evo 2 predictions shifts the distribution toward higher scores across all three axes, with the most pronounced effect on specificity.

Table S1. System prompts for variant interpretation. Two system prompts were used depending on whether the context configuration includes Evo 2 probe features. Both prompts request a structured JSON response containing a 3–5 sentence summary, a one-sentence mechanism, a confidence level (high/moderate/low), and key evidence points. Both include identical writing-style constraints: direct, concise prose; no em dashes; no superlatives or hedging language; no repetition of the variant ID; specific numbers preferred over vague qualifiers.

Prompt	Content
Baseline (without Evo2)	You are a clinical genomics expert. Based on the variant information provided, interpret the likely biological impact and pathogenicity of this variant.
	Be specific about biological mechanisms when the information supports it. 3–5 sentences for summary, 1 sentence for mechanism.
	Writing style: Write like a senior geneticist. Be direct and concise. Never use em dashes. Never use "notably", "critically", "importantly", "striking". Don't use superlatives or hedging language. State facts plainly. Don't repeat the variant ID or coordinates in the summary. Specific numbers are good. Vague qualifiers ("profound", "massive", "broad") are not.
	Return JSON: {summary, mechanism, confidence, key_evidence}
Evo2 (with Evo2)	You are a clinical genomics expert interpreting variant pathogenicity predictions from Evo2, a 7-billion-parameter DNA foundation model.
	The system uses a probe trained on reference genome activations, then run on both reference and variant activations. For each biological feature (helix, conservation, domain membership, etc.), you see: **ref** (the probe's prediction on the reference genome), **var** (the prediction on the variant genome), and Δ (var $-$ ref, the disruption signal). Large negative deltas mean the variant disrupts that feature. Near-zero deltas mean the feature is preserved. This is the key signal: pathogenic variants show large disruptions, benign variants show near-zero deltas.
	Provide a clinical interpretation: lead with the disruption story (which features were disrupted and which preserved); use the gene function context to explain why disruption of those features matters for this gene; explain why the model predicts what it does based on the disruption profile. Do not reference clinical prediction tools (CADD, AlphaMissense, SIFT, PolyPhen, etc.). 3–5 sentences for summary, 1 sentence for mechanism.
	Writing style: Write like a senior geneticist. Be direct and concise. Never use em dashes. Never use "notably", "critically", "importantly", "striking". Don't use superlatives or hedging language. State facts plainly. Don't repeat the variant ID or coordinates in the summary. Specific numbers are good. Vague qualifiers ("profound", "massive", "broad") are not.
	Return JSON: {summary, mechanism, confidence, key_evidence}

Table S2. Example user prompts for each context configuration. Each prompt shows the same variant (MYO7A chr11:77156021:T:A, a pathogenic missense variant) at increasing context levels. Each configuration cumulatively adds information to the previous level. The Evo 2 configuration uses the Evo 2 system prompt (Table S1); all others use the baseline system prompt.

Configuration	User prompt
1. Coordinates only	chr11:77156021:T:A
2. +Gene	MYO7A — chr11:77156021:T:A Consequence: missense_variant
3. +Gene context	MYO7A — chr11:77156021:T:A Consequence: missense_variant Domains: PF00063, PS51456, PTHR22692 | LOEUF: 0.995 (tolerant) Gene Function Context: Protein family: TRAFAC class myosin-kinesin ATPase superfamily. Myosin family Molecular functions: actin filament binding, ADP binding, ATP binding Biological processes: actin filament organization, actin filament-based movement, auditory receptor cell stereocilium organization, cochlea development, endocytosis
4. +HGVSp	MYO7A — chr11:77156021:T:A Consequence: missense_variant Protein: ENSP00000386331.3:p.Ile134Asn | Coding: ENST00000409709.9:c.401T>A Domains: PF00063, PS51456, PTHR22692 | LOEUF: 0.995 (tolerant) Gene Function Context: Protein family: TRAFAC class myosin-kinesin ATPase superfamily. Myosin family Molecular functions: actin filament binding, ADP binding, ATP binding Biological processes: actin filament organization, actin filament-based movement, auditory receptor cell stereocilium organization, cochlea development, endocytosis
5. +Evo2	MYO7A — chr11:77156021:T:A Consequence: missense_variant Protein: ENSP00000386331.3:p.Ile134Asn | Coding: ENST00000409709.9:c.401T>A Domains: PF00063, PS51456, PTHR22692 | LOEUF: 0.995 (tolerant) Gene Function Context: Protein family: TRAFAC class myosin-kinesin ATPase superfamily. Myosin family Molecular functions: actin filament binding, ADP binding, ATP binding Biological processes: actin filament organization, actin filament-based movement, auditory receptor cell stereocilium organization, cochlea development, endocytosis Predicted pathogenicity: 97% Calibration: among labeled variants scoring 90–101%, 97% are pathogenic. Region Disruptions (peak across ±32 kb): Feature ref var Δ distance CDS 0.964 0.431 −0.533 +68 bp Intron 0.036 0.569 +0.533 +68 bp P Loop NTP Hydrolase 0.222 0.749 +0.528 +8 bp Is Splice Donor 0.261 0.770 +0.509 +71 bp In Domain 0.648 0.203 −0.445 +21 bp Alpha Helix 0.491 0.092 −0.399 +68 bp *(top 6 of 10 shown)* Site Disruptions (at variant position): Feature ref var Δ In Transmembrane 0.045 0.440 +0.395 In Domain 0.443 0.128 −0.315 Is Disordered 0.096 0.322 +0.225 Beta Strand 0.326 0.147 −0.179 Alpha Helix 0.647 0.811 +0.164

Table S3. LLM judge prompts. Two independent judges (Claude Opus 4.6) evaluate each interpretation. Each judge receives the variant identifier, the LLM-generated interpretation, and ClinVar expert panel evidence as context.

Judge	Prompt
Mechanism coverage	You are evaluating an automated variant interpretation generated by an AI system that uses probe features from a genomic language model (Evo2) and an LLM without internet access. The system has no access to patient data, population frequencies, functional assays, or experimental results. Use the ClinVar expert panel evidence as ground truth for the molecular mechanism; ignore any non-expert submissions. *Variant:* ⟨variant_id⟩ *LLM Interpretation:* ⟨interpretation⟩ *ClinVar Expert Panel Evidence (ground truth):* ⟨clinvar_evidence⟩ *Mechanism Coverage Rubric (1–5).* Score = highest specificity level of mechanism correctly identified, using ClinVar expert panel SCV as ground truth. Only correct claims count. Level 1: **Gene-level only.** Names gene/disease/pathway but no variant-specific mechanism. Level 2: **Correct variant-type.** Identifies consequence type and expected effect but not which domain/residue. Level 3: **Correct functional class or protein region.** Names the right domain or functional category. Level 4: **Specific mechanism with biological reasoning.** Connects variant to consequence through biological logic. Level 5: **Level 4 + additional biological detail.** Second affected function, biophysical basis, downstream pathway, or non-obvious mechanism. *Judge procedure:* (1) Extract mechanism claims from the expert SCV; (2) extract mechanism claims from the LLM interpretation; (3) compare, where correct claims match or are consistent with the SCV and incorrect claims contradict it; (4) assign the score as the highest rubric level among correct claims. *Scoring rules:* Wrong pathogenicity direction caps the score at 1; hedged claims score at the level described if correct; missing clinical, population, or experimental data is not penalized. Returns JSON: {mechanism_score: 1–5, reasoning: 2–3 sentence justification}.
Biological accuracy & specificity	You are evaluating an automated variant interpretation generated by an AI system that uses probe features from a genomic language model (Evo2). *Variant:* ⟨variant_id⟩ *LLM Interpretation:* ⟨interpretation⟩ *ClinVar Submitted Evidence (human-curated):* ⟨clinvar_evidence⟩ Compare the LLM interpretation against the ClinVar evidence. Score on two axes (1–5): (1) *biological_accuracy*: Are the biological claims in the interpretation consistent with the ClinVar evidence? (5 = fully consistent, 1 = contradicts evidence) (2) *specificity*: Does the interpretation make specific, concrete claims rather than generic statements? (5 = very specific to this variant, 1 = could apply to any variant) Returns JSON: {biological_accuracy: 1–5, specificity: 1–5}.

Table S4. Example interpretation: mechanism score 1 (gene-level only). *ETHE1* c.55G>T (chr19:43506794:C:A), classified as benign by the ClinGen Mito Variant Curation Expert Panel, generated at the Evo 2 context configuration by Claude Opus 4.6. The LLM misidentifies the coding variant as a 3' UTR variant (c.*55G>T), building its entire mechanism reasoning on an incorrect premise. While the benign conclusion is correct, only gene-level identification is accurate.

Component	Content
LLM summary	This 3' UTR variant in ETHE1 shows a 0% pathogenicity prediction, consistent with the absence of any meaningful disruption to protein-coding or structural features. As a variant located 55 bases past the stop codon, it falls outside the translated region and is not expected to alter protein structure, active site residues, or conserved functional domains. The LOEUF score of 0.979 indicates ETHE1 is relatively tolerant to loss-of-function variation, further reducing concern. The model correctly identifies this as a benign variant given the lack of disruption to any feature relevant to ETHE1's sulfur dioxygenase or iron-binding functions.
LLM mechanism	A 3' UTR variant at this position does not disrupt the coding sequence, protein folding, or any catalytic domain required for ETHE1's mitochondrial sulfur dioxygenase activity.
Expert SCV	The filtering allele frequency of the c.55G>T variant in the ETHE1 gene is >0.7% by the Exome Aggregation Consortium, which is a high enough frequency to be classified as benign based on thresholds defined by the ClinGen Mito Variant Curation Expert Panel (BA1).
Judge reasoning	The LLM correctly identifies the gene (ETHE1) and reaches the correct benign conclusion, but incorrectly classifies the variant as a 3' UTR variant (c.*55G>T) when the expert panel identifies it as c.55G>T (coding region). Since the variant-type mechanism reasoning is based on an incorrect premise about the variant's location, the highest correct claim is gene-level identification with correct pathogenicity direction.

Table S5. Example interpretation: mechanism score 2 (correct variant-type). *MAP2K1* c.1068+9A>G (chr15:66489771:A:G), classified as likely benign by the ClinGen RASopathy Expert Panel, generated at the Evo 2 context configuration by Claude Opus 4.6. The LLM correctly identifies the intronic variant as outside the canonical splice site with no predicted disruption, but does not advance beyond variant-type classification.

Component	Content
LLM summary	This intronic variant located 9 bases into the intron shows no disruptive signal across any probed feature. All deltas are near zero or slightly positive, with no negative disruptions to secondary structure, domain integrity, or protein interactions. The small positive shifts in kinase domain and secondary structure probes at a nearby position likely reflect minor probe noise rather than meaningful biological change, as the variant sits well outside the critical splice donor dinucleotide. MAP2K1 is a constrained gene (LOEUF 0.515), but this deep intronic position is unlikely to affect splicing or protein function. The model assigns 0% pathogenicity, consistent with a benign intronic variant that preserves all functional features.
LLM mechanism	An intronic variant at position c.1068+9 that does not disrupt splice signals, protein domain structure, or any other probed biological feature.
Expert SCV	The c.1068+9A>G intronic variant in the MAP2K1 gene is classified as likely benign because it does not alter an amino acid residue, is not located within the canonical splice site, and computational splice prediction tools do not predict an impact on splicing (BP4, BP7). It has been identified in 0.01123% of Latino chromosomes in gnomAD (BS1 not met). In summary, this variant meets criteria to be classified as likely benign. ACMG/AMP criteria applied: BP4, BP7.
Judge reasoning	The LLM correctly identifies this as an intronic variant (c.1068+9) outside the canonical splice site with no predicted splice disruption, matching the expert SCV's BP4/BP7 classification as likely benign. This represents correct variant-type identification (intronic, no splice impact) consistent with Level 2. While the LLM mentions the kinase domain and specific distance from the splice site, these details support the same "intronic, no splice impact" conclusion rather than adding a distinct functional region or mechanistic insight that would reach Level 3.

Table S6. Example interpretation: mechanism score 3 (correct functional class). *FOXN1* p.Arg255Ter (chr17:28529156:C:T), classified as pathogenic by the ClinGen SCID VCEP, generated at the Evo 2 context configuration by Claude Opus 4.6. The LLM correctly localizes the nonsense variant to the forkhead DNA-binding domain, but proposes protein truncation with domain loss rather than the expert-annotated nonsense-mediated decay mechanism.

Component	Content
LLM summary	This nonsense variant introduces a premature termination codon at residue 255 in FOXN1, a forkhead box transcription factor essential for thymic epithelial cell development and hair follicle differentiation. The model detects catastrophic loss of the downstream DNA-binding forkhead domain, with homeodomain-like superfamily membership dropping from 0.74 to 0.18 ($\Delta = -0.56$) and DNA binding probability collapsing from 0.61 to 0.05 ($\Delta = -0.56$). The downstream region transitions from structured (beta strand 0.69 to 0.13) to disordered (0.22 to 0.81), consistent with complete loss of the folded forkhead domain that mediates target gene regulation. FOXN1 is highly intolerant to loss of function (LOEUF = 0.442), and biallelic loss-of-function variants cause nude/SCID phenotype with T-cell immunodeficiency and alopecia, making truncation upstream of the DNA-binding domain unambiguously damaging.
LLM mechanism	Premature stop codon at position 255 eliminates the forkhead DNA-binding domain, abolishing FOXN1 transcriptional activity required for thymic epithelial and keratinocyte differentiation.
Expert SCV	The NM_001369369.1(FOXN1):c.763C>T (p.Arg255Ter) nonsense variant occurs in exon 5 of 9 and is predicted to result in nonsense-mediated decay (PVS1). The highest MAF in gnomAD v2.1.1 is 0.000008790 in the European (non-Finnish) population, below the SCID VCEP threshold (PM2_supporting). At least one patient is described with congenital alopecia, nail dystrophy, and very low T cell number, highly specific to FOXN1 deficiency (PP4). ACMG/AMP criteria: PVS1, PM2_supporting, PP4, PP1_strong, PM3.
Judge reasoning	The LLM correctly identifies the nonsense variant in FOXN1, the forkhead DNA-binding domain as the critical affected region, and provides biological reasoning about loss of transcriptional activity. However, the expert panel explicitly states the primary mechanism is nonsense-mediated decay (PVS1, exon 5 of 9), while the LLM describes protein truncation with domain loss. This mechanistic discrepancy (NMD vs truncated protein) prevents scoring at Level 4, but correct domain identification and functional class earns Level 3.

Table S7. Example interpretation: mechanism score 4 (specific mechanism with biological reasoning). *PAH* p.Leu213Pro (chr12:102855203:A:G), classified as pathogenic by multiple clinical laboratories, generated at the Evo 2 context configuration by Claude Opus 4.6. The LLM connects proline-mediated helix breaking to destabilization of the catalytic domain required for enzymatic activity.

Component	Content
LLM summary	This Leu-to-Pro substitution in the PAH catalytic domain produces a clear alpha helix disruption at the variant site (Δ -0.245, from 0.975 to 0.731), with a concurrent gain in beta strand propensity (+0.218) and increased disorder (+0.166). Proline is a well-known helix breaker, and the probe captures exactly this structural consequence: a position that was confidently helical in the reference is shifted toward irregular or strand-like conformation. The variant also increases the predicted protein-protein interaction interface signal (+0.147), suggesting altered surface properties at a region involved in PAH oligomerization or cofactor interaction. PAH requires precise active-site geometry for iron coordination and BH4 cofactor binding, and destabilization of a helix within the catalytic domain (PF00351) is a recognized mechanism for loss of phenylalanine hydroxylase activity, consistent with classical PKU.
LLM mechanism	Proline substitution disrupts an alpha helix in the PAH catalytic domain, increasing local disorder and destabilizing the structural framework required for iron binding and enzymatic activity.
Expert SCV	This sequence change replaces leucine, which is neutral and non-polar, with proline, which is neutral and non-polar, at codon 213 of the PAH protein (p.Leu213Pro). This variant is not present in population databases (gnomAD no frequency). This missense change has been observed in individuals with phenylketonuria (PMIDs: 8632937, 9521426, 19292873, 22112818, 23430547). Advanced modeling indicates that this missense variant is expected to disrupt PAH protein function with a positive predictive value of 80%. PAH-specific ACMG/AMP criteria applied: PP4_Moderate, PM3_Very-strong.
Judge reasoning	The LLM correctly identifies the pathogenic direction, places the variant in the PAH catalytic/hydroxylase domain (PF00351), and provides specific biological reasoning connecting proline-mediated helix disruption to destabilization of the structural framework required for iron binding and enzymatic activity. This is consistent with the expert panel's pathogenic classification and PP3 computational evidence, reaching Level 4 specificity with mechanistic biological logic beyond just naming the domain.

Table S8. Example interpretation: mechanism score 5 (multi-mechanism with biophysical detail). *GAA* p.Met1Val (chr17:80104586:A:G), classified as pathogenic by the ClinGen Lysosomal Diseases VCEP for Pompe disease, generated at the Evo 2 context configuration by Claude Opus 4.6. The LLM traces a multi-step biological chain from start codon loss through signal peptide elimination to failed lysosomal targeting, entirely from disruption profile signals.

Component	Content
LLM summary	This variant abolishes the initiator methionine codon of GAA, with the model detecting a start codon probability drop from 0.976 to 0.477 at the variant site and from 0.959 to 0.079 in the peak region (+2 bp). The downstream consequences are severe: the model predicts loss of the signal/transit peptide (0.382 to 0.029), loss of the protein chain (0.863 to 0.230), and near-complete destruction of the transmembrane signal peptide region (0.923 to 0.007 at +48 bp). Without a functional start codon, GAA acid alpha-glucosidase cannot be synthesized or properly targeted to lysosomes, which is the direct cause of Pompe disease (glycogen storage disease type II). The model also detects aberrant CDS and polypyrimidine tract signals upstream, consistent with the translational machinery failing to recognize this region as a normal coding sequence initiation site.
LLM mechanism	Loss of the initiator methionine codon prevents translation of GAA acid alpha-glucosidase, eliminating the signal peptide required for lysosomal targeting and abolishing enzymatic glycogen degradation.
Expert SCV	The NM_000152.5:c.1A>G (p.Met1Val) start loss variant in GAA may cause a truncated or absent protein by altering the start codon (PVS1_Strong). This variant is absent in gnomAD v2.1.1 (PM2_Supporting). This variant has been detected in at least 5 individuals with Pompe disease, including three compound heterozygous for another pathogenic variant (PM3_Strong). At least two patients had documented GAA deficiency with $<$10% of the normal mean control level of GAA activity in dried blood spots. One homozygous patient had documented severe infantile onset and was CRIM negative (PP4_Moderate). ACMG/AMP criteria: PVS1_Strong, PM3_Strong, PP4_Moderate, PM2_Supporting.
Judge reasoning	The LLM correctly identifies the start codon loss in GAA causing absent or truncated protein, matching the SCV's p.Met1Val start loss, and goes beyond by explaining the biological chain: loss of initiator methionine, loss of signal peptide, failure of lysosomal targeting, abolished glycogen degradation, and Pompe disease. This provides Level 4 biological reasoning (connecting variant to consequence) plus additional biological detail about the signal peptide and lysosomal targeting pathway, reaching Level 5.