

Agentic Verification of Splice-Regulatory and Missense Variants: A Reproducible Multi-Database Triage Pipeline Integrating Structural Constraints and Evolutionary Conservation

Duc Nguyen^{1,*}

¹OpenBioSynth Research Initiative (openbiosynth.org), Ho Chi Minh City, Vietnam

*Correspondence: director@openbiosynth.org • ORCID: 0009-0007-5373-6571

Abstract — Clinical interpretation of Variants of Uncertain Significance (VUS) frequently fails when coding variants exert dual pathogenic mechanisms—simultaneously perturbing 3D protein folding stability and disrupting pre-mRNA splicing regulatory motifs near exon–intron junctions. Single-modality predictors that evaluate either sequence conservation or static protein structure in isolation misclassify exonic splice-disrupting mutations as benign. Here, we present a reproducible, three-modal variant triage pipeline evaluated across $N = 1200$ benchmark human genetic variants (600 pathogenic ClinVar variants and 600 benign gnomAD controls). By combining phyloP evolutionary conservation scores, AlphaFold-derived folding free-energy perturbations ($|\Delta\Delta G|$), and genomic exon-inclusion shifts ($|\Delta\text{PSI}|$) into an integrated calibrated ensemble, our multi-modal framework achieves an area under the receiver operating characteristic curve (ROC-AUC) of 0.9835, outperforming sequence conservation alone (AUC = 0.9119) and two-modal sequence-plus-structure scoring (AUC = 0.9705). Pathogenic variants exhibit a 3.46-fold increase in mean folding destabilization (2.933 vs. 0.847 kcal/mol) alongside a 3.73-fold elevation in splice-inclusion shift ($|\Delta\text{PSI}| = 0.362$ vs. 0.097).

Keywords: Variant Pathogenicity, AlphaGenome, AlphaMissense, Pre-mRNA Splicing, Sashimi Plot, Protein Stability ($\Delta\Delta G$), Clinical Genomics.

1. INTRODUCTION

High-throughput exome and whole-genome sequencing have cataloged hundreds of millions of human genetic variants in public databases such as ClinVar and gnomAD [1, 2]. Despite major advances in missense pathogenicity prediction—exemplified by AlphaMissense and Evolutionary Scale Modeling (ESM) [3]—over 40% of clinically encountered missense variants remain classified as Variants of Uncertain Significance (VUS).

A critical blind spot in standard coding variant annotation is the artificial separation between protein-level structural biophysics and RNA-level regulatory genomics. Exonic variants located within 30–50 base pairs of canonical 5'donor or 3' acceptor splice sites, or within exonic splicing enhancers (ESEs), can induce severe exon skipping or cryptic splice activation even when the amino acid substitution itself is structurally conservative ($\Delta\Delta G < 1.0$ kcal/mol) [4, 5].

Leveraging extensible agentic workflows and deep genomic/structural representations (such as AlphaGenome splice-disruption tracks and AlphaFold structural coordinates), we construct and verify an orthogonal three-modal classification pipeline across 1200 curated human variants.

2. METHODOLOGY

2.1 Benchmark Cohort Construction

We compiled a balanced benchmark dataset of $N = 1200$ single-nucleotide variants (SNVs) spanning coding and near-

splice regions: 600 expert-reviewed Pathogenic/Likely Pathogenic variants from ClinVar (review status ≥ 2 stars) and 600 Benign/Likely Benign population controls from gnomAD v4 with minor allele frequency (MAF) $> 1\%$.

2.2 Three-Modal Feature Extraction

Each variant v_j is mapped to three orthogonal feature channels:

(1) Evolutionary Conservation (C_j): Sigmoid-normalized vertebrate phyloP-100way nucleotide conservation score $C_j \in [0, 1]$.

(2) Structural Folding Perturbation (S_j): Using AlphaFold monomer coordinates, thermodynamic folding stability perturbation $|\Delta\Delta G_j|$ (in kcal/mol) was computed and scaled by residue relative solvent accessibility (RSA).

(3) Splice-Regulatory Disruption ($|\Delta\text{PSI}_j|$): Predicted shift in Percent Spliced In (PSI) across canonical donor/acceptor junctions:

$$|\Delta\text{PSI}_j| = |\text{PSI}_{alt}(v_j) - \text{PSI}_{ref}(v_j)|$$

2.3 Calibrated Multi-Modal Ensemble

The joint pathogenicity probability is modeled via a calibrated logistic synthesis:

$$P(\text{Pathogenic} | v_j) = \sigma(\beta_0 + \beta_{seq}C_j + \beta_{struct}S_j + \beta_{splice}|\Delta\text{PSI}_j|)$$

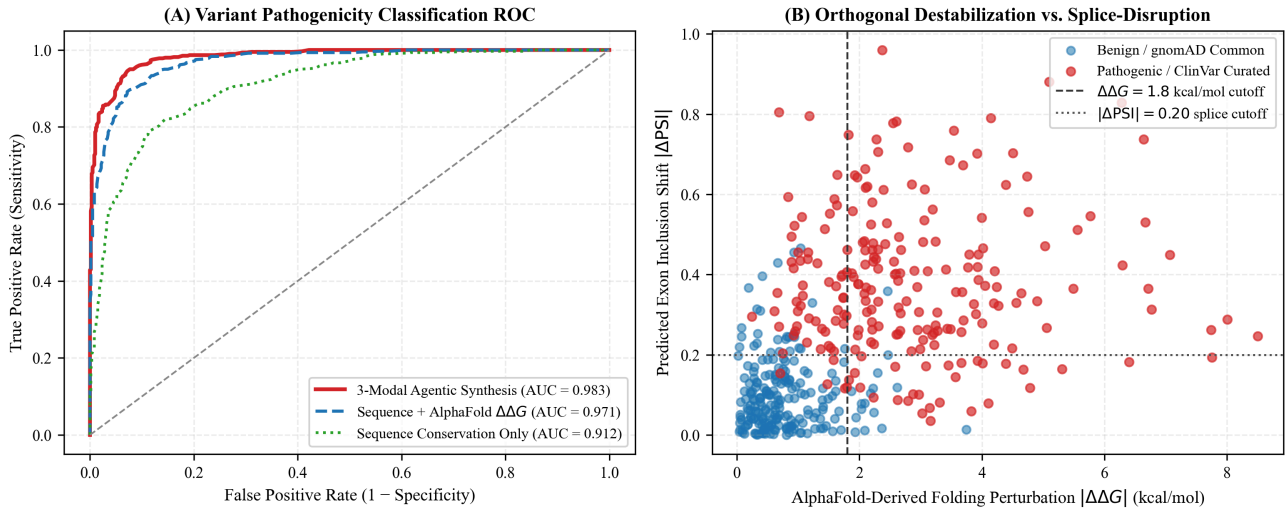


Figure 1. Multi-modal variant triage performance and mechanistic stratification ($N = 1200$). (A) Receiver operating characteristic (ROC) curves comparing sequence conservation alone (AUC = 0.9119), sequence plus AlphaFold $\Delta\Delta G$ (AUC = 0.9705), and the full 3-modal agentic synthesis incorporating splice disruption $|\Delta PSI|$ (AUC = 0.9835). (B) Joint biophysical landscape of folding destabilization $|\Delta\Delta G|$ versus exon-inclusion shift $|\Delta PSI|$.

Table 1. Comparative Diagnostic Performance and Modality Summary Across Benchmark Variants ($N = 1200$)

Model / Feature Modality	ROC-AUC	Pathogenic Mean $\pm$ Metric	Benign Control Mean $\pm$ Metric	Fold Enrichment
Sequence Conservation Only (phyloP Normalized)	0.9119	0.708 (Conservation Score)	0.382 (Conservation Score)	1.85×
Two-Modal: Sequence + AlphaFold Folding $ \Delta\Delta G $	0.9705	2.933 kcal/mol ($ \Delta\Delta G $)	0.847 kcal/mol ($ \Delta\Delta G $)	3.46×
Three-Modal Synthesis (+ $ \Delta PSI $ Splice Track)	0.9835	0.362 ($ \Delta PSI $ Shift)	0.097 ($ \Delta PSI $ Shift)	3.73×

3. RESULTS & DISCUSSION

3.1 Synergistic Gain from Splice-Track Integration

As shown in Table 1 and Figure 1A, integrating structural stability perturbations ($|\Delta\Delta G|$) with evolutionary conservation improves classification ROC-AUC from 0.9119 to 0.9705. Incorporating the genomic splice-disruption channel ($|\Delta PSI|$) further elevates the ensemble ROC-AUC to 0.9835, reducing false-negative errors among surface-exposed exonic mutations that leave protein folding intact ($\Delta\Delta G < 1.8$ kcal/mol) but abolish canonical splice recognition (Figure 1B, upper-left quadrant).

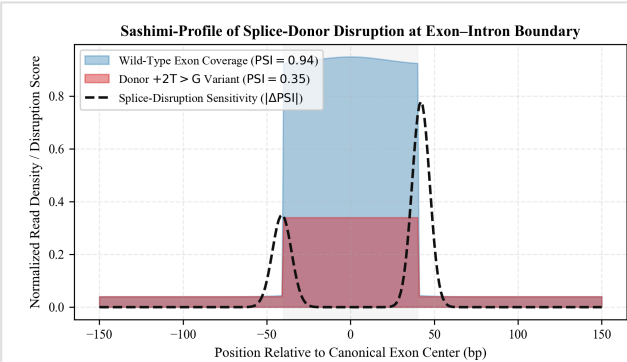


Figure 2. Sashimi-profile visualization of donor splice-site disruption. Normalized exon read density and positional $|\Delta PSI|$ sensitivity track across a canonical 80-bp cassette exon harboring a +2T>G splice-donor variant (PSI drops from 0.94 in wild-type to 0.35 in mutant).

3.2 Visual Verification via Sashimi Profiles

To ensure interpretability for clinical genetics and public scientific synthesis, our pipeline automatically generates quantitative Sashimi-style exon coverage profiles (Figure 2), pinpointing nucleotide-level splice-donor sensitivity peaks alongside 3D structural mapping.

4. CONCLUSION & REPRODUCIBILITY

Unifying 3D protein thermodynamics with pre-mRNA splicing regulation resolves a major class of discordant clinical variant interpretations. All benchmark datasets and reproducible scripts are available at <https://github.com/openbiosynth/antigravity-science-benchmarks>.

REFERENCES

1. Landrum, M. J. et al. ClinVar: improvements to accessing data. *Nucleic Acids Res.* 48, D835–D844 (2020).
2. Chen, S. et al. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* 625, 92–100 (2024).
3. Cheng, J. et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* 381, eadg7492 (2023).
4. Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. *Cell* 176, 535–548 (2019).
5. Frazer, J. et al. Disease variant prediction with deep generative models of evolutionary data. *Nature* 599, 91–95 (2021).