Mechanistic Synergy in Multi-Modal VEP: DNA Context Complements PLMs under Biophysical Constraints

Yoojin Kim [1] **Doyeon Ha** [1 †]

Abstract

While Protein Language Models (PLMs) have advanced Variant Effect Prediction (VEP), they can sometimes overlook the complex physical and regulatory contexts of the cell. To address this limitation, we propose a parameter-efficient multi-modal foundation model architecture that integrates DNA signals from a DNA Language Model (DLM) with protein representations from a PLM. Through a systematic analysis across 29 activity-based Deep Mutational Scanning (DMS) datasets from ProteinGym, we demonstrate that the nucleotide modality provides more than a simple ensemble gain; it specifically corrects PLM failure modes localized to charged residues. Ultimately, this work highlights that moving beyond unimodal representations is essential for capturing the mechanistic complexity of biological systems and for accurately identifying experimentally validated functional hotspots.

1. Introduction

Variant Effect Prediction (VEP) aids clinical diagnosis and drug discovery by determining the pathogenicity of genetic variants. Unlike gene-level tasks, VEP must discern subtle functional changes caused by single amino acid substitutions. While Protein Foundation Models (PLMs) are the standard backbone for VEP, ensuring clinical reliability requires a deeper mechanistic understanding of their predictive behaviors.

Despite overall performance improvements, the mechanistic reasons why models fail at specific residues remain underexplored. We hypothesize these blind spots stem from the inherent constraints of unimodal representations. In vivo, proteins fold and interact concurrently during translation. PLMs, which learn from isolated final sequences, might miss this kinetic and multi-protein context.

To bridge this gap, we propose a multi-modal fusion framework. By coupling a PLM with a DNA Language Model (DLM), we explore whether these nucleotide-level signals can compensate for physical contexts invisible to unimodal baselines. Our results quantitatively demonstrate that multi-modal integration effectively corrects specific predictive vulnerabilities, particularly at complex interaction surfaces, even where protein-only models exhibit systematic bias (Jumper et al., 2021; Krishna et al., 2024).

Our main contributions are as follows:

- **Parameter-efficient multi-modal fusion:** We achieve cross-modal synergy using a lightweight architecture that extracts and couples zero-shot geometric scalars from a DLM with protein embeddings, keeping both foundation models completely frozen.
- **Validation of multi-modal synergy across genes:** Through systematic evaluation across 29 ProteinGym targets (Notin et al., 2023), we demonstrate that the nucleotide modality improves upon the unimodal baseline. Crucially, ablation studies confirm that zero-shot geometric priors are the essential drivers of this gain.
- **Variant-level interpretation & literature validation:** Residue-level analysis shows our multi-modal approach corrects unimodal blind spots at complex interaction surfaces. Cross-validation with experimental literature confirms the model accurately targets critical functional sites (e.g., electrostatic networks and PPI hot-spots).

2. Methodology: Multi-modal Fusion Framework

Our framework synergizes protein-level semantics with nucleotide-level regulatory signals through a parameter-efficient fusion strategy (Figure 1).

2.1. Data Acquisition and Multi-modal Representation

Context-aware Back-translation: To align protein variants with their genomic origins, we retrieve validated CDS from the European Nucleotide Archive (ENA; (O'Cathail

[1]3billion, Seoul, Korea. Correspondence to: Doyeon Ha <biologysaves@gmail.com>.

Proceedings of the ICML 2026 3rd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Seoul, Korea. 2026.

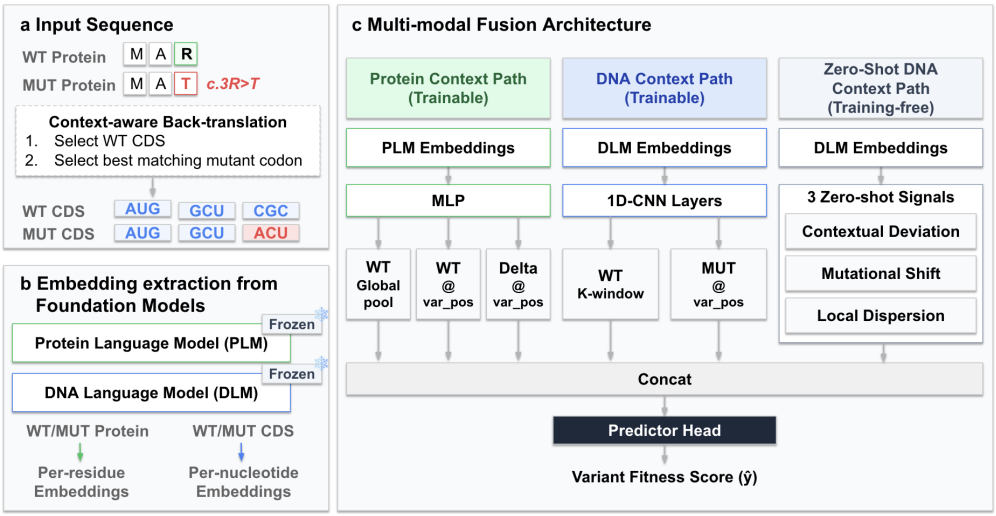

Figure 1. **Schematic of the multi-modal VEP architecture coupling protein and DNA context. (a) Input:** Wild-type (WT) and mutant (MUT) sequences are processed at both protein and DNA levels, with DNA obtained via context-aware back-translation. **(b) Foundation Model Embedding Extraction:** protein and DNA sequences are encoded by frozen PLM and DLM backbones, respectively. **(c) Multi-modal Encoder:** three encoding paths are employed — *(Left)* a trainable protein path projecting ESM3 features into global, WT-local, and mutational-delta views; *(Center)* a trainable DNA path applying a 1D-CNN over DLM's representations to extract windowed and site-specific representations; *(Right)* a training-free path extracting three zero-shot signals directly from the frozen DLM. All features are concatenated and passed through a 2-layer MLP head to predict the variant fitness score ($\hat{y}$)

.

et al., 2025)) via EMBL cross-references in UniProt. For each substitution, mutant codons are selected by minimizing Hamming distance from the wild-type (WT) codon, with ties resolved by gene-specific codon usage frequencies (Figure 1a).

Feature Extraction: We extract per-residue embeddings from the **`EvolutionaryScale/esm3-sm-open-v1`** model ((Hayes et al., 2025), $d_E = 1536$) and per-nucleotide embeddings from the **`NTv3-650M`** model ((Boshar et al., 2025), $d_N = 768$) for both WT and mutant sequences. These foundation models remain frozen throughout training to preserve their pre-trained biological priors (Figure 1b).

2.2. Fusion Architecture and Training

The architecture integrates three encoding paths to capture the multi-layered biophysical impact of variants (Figure 1c). Specifically, we compute three zero-shot geometric signals from the frozen DLM latent space to incorporate physical priors without additional learning:

1. **Contextual Deviation:** $s_{\text{err}} = \|\mathbf{z}_{wt,p} - \bar{\mathbf{z}}_{ctx,p}\|_2$, capturing codon atypicality.
2. **Mutational Shift:** $s_{\text{shift}} = \|\mathbf{z}_{mut,p} - \mathbf{z}_{wt,p}\|_2$, proxying mutational disruptiveness.
3. **Local Dispersion:** $s_{\text{disp}} = \text{Var}(\mathbf{z}_{wt,p\pm k})$, reflecting local sequence conservation.

The signal vector $\mathbf{s} = [s_{\text{err}}; s_{\text{shift}}; s_{\text{disp}}]$ is concatenated with the learned features and passed to a prediction head to predict the fitness score $\hat{y}$.

To ensure robust generalization, we employ **position-aware 5-fold cross-validation**, partitioning the dataset by residue position rather than individual variants. Detailed hyperparameters are provided in Appendix A.

3. Validation of multi-modal synergy across genes

Before analyzing the mechanistic synergies of the multi-modal signals, we first establish their aggregate efficacy across the 29 ProteinGym targets.

Multi-modal fusion improves VEP. For context, zero-shot ESM3 scoring (without supervised fine-tuning) attains only $\bar{\rho} = 0.296$ on the same 29 targets, indicating that there is room for improvement. Our multi-modal architecture evaluated via a position-aware 5-fold cross-validation protocol achieves a mean Spearman ρ of 0.508. This outperforms the unimodal supervised ESM-ONLY baseline ($\bar{\rho} = 0.497$, paired Wilcoxon $p = 0.039$). While the aggregate multi-modal lift is modest (mean $+0.012$), it is highly heterogeneous across targets, with $\Delta\rho$ ranging from -0.025 to $+0.090$ (**Figure 2**). This heterogeneity indicates that the nucleotide modality does not provide a uniform scalar boost, but rather targets specific predictive vulnerabilities. To explicitly pinpoint the source of this performance gain, a detailed ablation study on the DNA-view contexts is provided in Appendix B. This directly motivates the residue-level

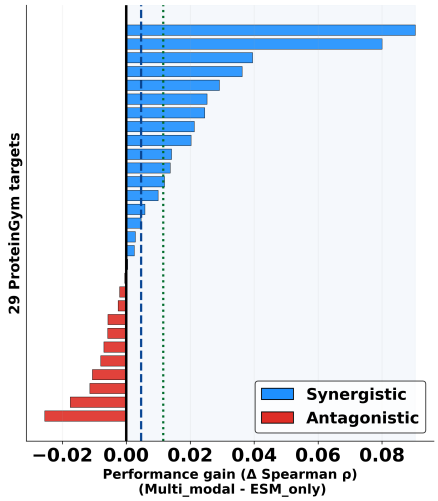

Figure 2. **Overall performance gain of multi-modal fusion.** Per-gene Spearman ρ improvement ($\Delta\rho$) of our multi-modal architecture relative to the unimodal ESM-ONLY baseline across 29 ProteinGym targets. Bars are ordered by performance gain, with blue indicating synergistic improvement (up to $+0.090$) and red indicating antagonistic effects. The dashed line(blue) and the dotted line(green) denote the median and mean ($\Delta\rho$) across the 29 targets (mean = +0.012), respectively. The high heterogeneity motivates the residue-level mechanistic analysis.

mechanistic analysis in Section 4.

4. Residue-Level Analysis: Improvements at Positively Charged Surfaces

While the aggregate lift confirms the efficacy of multi-modal fusion, it obscures *where* the nucleotide modality provides the most critical improvements to the protein-only baseline. To pinpoint these regions, we decomposed the gene-level Spearman improvement into residue-specific contributions. Specifically, we calculated the *rank residual change* for each variant—measuring how much closer the multi-modal prediction moved to the ground-truth DMS rank compared to the ESM-only baseline. By averaging these changes across all substitutions at a given position, we identified the residue locations most sensitive to cross-modal signals. We defined two extreme sets:

- **Synergistic Positions ($\mathcal{P}^+$):** The top 5% of positions where the nucleotide modality most significantly improved predictive accuracy by reducing rank residuals.
- **Antagonistic Positions ($\mathcal{P}^-$):** The bottom 5% of positions where the inclusion of nucleotide features led to the largest relative performance drop.

The remaining variant-bearing sites form the **Background** ($\mathcal{B}$). This spatial decomposition allows us to move beyond aggregate metrics and ask: *What biochemical properties characterize the regions where DNA context is most informative?* Pooling across 29 genes yields $|\mathcal{P}^+| = 482$, $|\mathcal{P}^-| = 482$, and $|\mathcal{B}| = 8,399$ positions. Detailed mathematical derivations are provided in Appendix C.

Spatial Localization: Multi-modal Fusion Specifically Targets Positively Charged Residues. To understand the biological nature of these extreme positions, we tested 16 binary residue-level features (defined by wild-type residue identity and AlphaFold-2((Jumper et al., 2021))-derived relative SASA) for statistical enrichment in $\mathcal{P}^+$ and $\mathcal{P}^-$ relative to $\mathcal{B}$. For each (feature f, side s) pair, the *enrichment ratio* $\mathrm{enr}_{f,s} = \Pr(f \mid s)/\Pr(f \mid \mathcal{B})$ quantifies over-representation. One-sided Fisher exact tests were used with a Bonferroni-corrected threshold ($\alpha_{\mathrm{Bonf}} \approx 1.56 \times 10^{-3}$) to ensure robustness across 32 tests.

As summarized in Table 1, only **positively charged residues (K/R/H)** pass the stringent Bonferroni correction ($p_+ = 7.0 \times 10^{-4}$), exhibiting a clean synergistic signature ($\mathrm{enr}_+ = 1.40$ vs. $\mathrm{enr}_- = 1.02$, Ratio $= 1.37$). This trend is also observed for the subset of these residues located on the protein surface ($\mathrm{enr}_+ = 1.38$ for surface pos. charge). This suggests that the multi-modal lift is localized to residues often involved in electrostatic interactions—potential mediators of protein-nucleic acid binding or structural stability—where the DNA-view may provide critical evolutionary priors that the ESM-only baseline underestimates. Conversely, none of the 16 features passed the Bonferroni correction for the antagonistic positions ($\mathcal{P}^-$). The complete statistical results for all features are detailed in Appendix C.4.

Table 1. **Feature enrichment in extreme positions.** enr_+ measures over-representation in Synergistic positions ($\mathcal{P}^+$); enr_- for Antagonistic positions ($\mathcal{P}^-$). Ratio > 1 indicates a synergy-specific signal. Bold values denote significance under Bonferroni-corrected threshold ($\alpha_{\mathrm{Bonf}} = 0.0016$).

Feature	enr_+	enr_-	Ratio	p_+
Pos. charge (K/R/H)	**1.40**	**1.02**	**1.37**	$\mathbf{7.0 \times 10^{-4}}$
Surface pos. charge	1.38	1.10	1.26	6.8×10^{-3}

5. Structural Case Studies: Resolving PLM Blind Spots

To investigate how multi-modal fusion corrects specific failure modes of ESM-only models, we analyze the KCNJ2 potassium channel (Figure 3). We identify two primary mechanism cases where the multi-modal architecture resolves the systematic "optimistic bias" seen in unimodal predictions—a tendency of ESM-only models to falsely predict deleterious mutations by overlooking variant effects of positively charged residues.

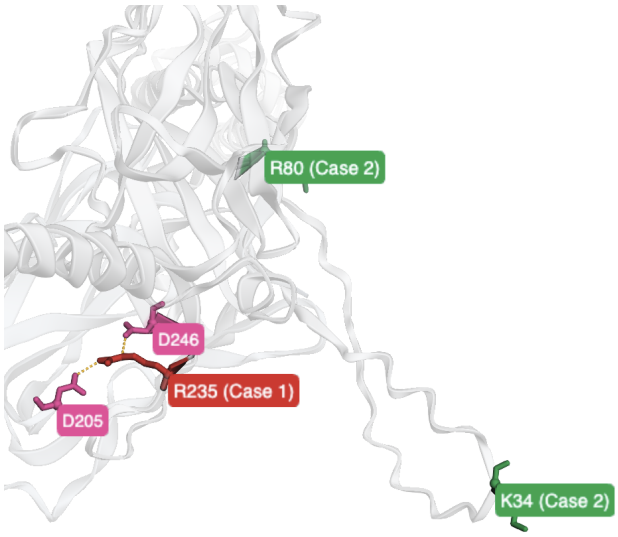

Figure 3. **Structural localization of multi-modal predictive synergy in KCNJ2.** The figure highlights representative case-study residues from the K/R/H subset of KCNJ2's **Synergistic positions** (top-5% $\mathcal{P}^+$). **Red/Pink nodes** (R235 and its interacting partners D205 and D246) illustrate the *Complex Salt-Bridge Network Collapse* (Case 1) mechanism, where multi-modal fusion corrects the failure of protein-only models to capture internal electrostatic stability. **Green nodes** (K34, R80) indicate *Exposed Interfacial PPI Hot-Spots* (Case 2), where DNA-level signals highlight functional interfaces hidden from protein sequences alone. Dashed lines indicate electrostatic contacts within a 4.0 Å cutoff.

(Case 1) Complex Salt-Bridge Network Collapse. Salt bridges—electrostatic interactions between oppositely charged residues—are essential for the structural integrity and gating of Kir channels. A prime example is the positively charged **R235** (highlighted in **red** in Figure 3), which forms a robust intra-subunit salt bridge with the buried, negatively charged **D246** (**pink**) (2.64 Å) while simultaneously establishing an inter-subunit contact with another negatively charged partner, **D205** (**pink**). Specifically, abolishing the salt bridge with D205 has been shown to significantly reduce the basal activity of the Kir2.1 channel compared to the wild-type (WT) state (Borschel et al., 2017). While the ESM-only baseline fails to penalize the uncoupling of these hidden 3D dependencies, the **multi-modal fusion** model correctly identifies the defect by capturing stringent evolutionary constraints encoded in codon usage.

(Case 2) Exposed Interfacial PPI Hot-Spots. Synergistic positions are also prominent in functional interfaces, such as the positively charged residues **R80** and **K34** (highlighted in **green** in Figure 3). Structurally located at the exposed N-terminus of the outer helix, **R80** is essential for channel activation as it directly binds the 1' phosphate of PIP_2 (Hansen et al., 2011). While MSAs and protein-only models often treat such exposed interfacial regions as mutation-resilient, the **multi-modal fusion** model leverages DNA-view signals to identify the stringent regulatory pressures specific to these critical lipid-protein interaction hot-spots.

Quantifying the Improvement of PLM Blind Spots. The impact of multi-modal integration is most evident in *charge-flip* mutations (K/R/H $\rightarrow$ D/E), which maximally disrupt electrostatic networks. As shown in Table 2, the ESM-only model exhibits a severe optimistic bias (mean residual: $+1{,}404$ ranks), consistently failing to detect variant lethality. In contrast, the multi-modal fusion eliminates this bias, achieving near-perfect calibration(-40 ranks). On the broader K/R/H set ($n = 141$), the same fusion delivers a **4.4× reduction** in mean rank residual ($+1{,}617 \rightarrow +366$). At binary resolution (predictions binarized at a model-inferred WT baseline; ground-truth from ProteinGym `DMS_score_bin`, 0=LOF, 1=functional), this corresponds to a $\mathbf{4.9}\times$ improvement in LOF sensitivity ($\mathbf{15.4}\% \rightarrow \mathbf{75.6}\%$, AUC $0.46 \rightarrow 0.63$). This confirms that nucleotide-level signals provide the necessary context to restore predictive accuracy where functional partner-disruption is most severe.

Table 2. **Improvement of rank residuals at synergistic positions in KCNJ2.** Positive values denote an over-prediction of variant fitness. By integrating DNA-level signals, the multi-modal model effectively neutralizes the optimistic bias of the ESM-only baseline, particularly for charge-flip mutations. All values are mean rank residuals rounded to the nearest integer.

Metric (Mean Rank Residual)	ESM-ONLY	MULTI-MODAL
Overall K/R/H Variants ($n = 141$)	$+1,617$	$+366$
Charge-flip (K/R/H $\rightarrow$ D/E)	$\mathbf{+1,404}$	$\mathbf{-40}$

6. Conclusion

In this study, we addressed the limitations of unimodal PLMs by introducing a parameter-efficient multi-modal fusion framework. By coupling protein semantics with zero-shot geometric signals from a DLM, we achieved performance improvements across 29 ProteinGym targets without massive parameter updates.

Crucially, our residue-level analysis revealed that this cross-modal synergy specifically corrects the "optimistic bias" of unimodal models at positively charged residues. The structural case studies suggest that our model accurately targets hidden biophysical constraints, such as delicate electrostatic networks and exposed protein-protein interaction (PPI) hot-spots in charged residues. The observed improvements are consistent with critical functional sites validated in literature, including the KCNJ2 inter-subunit salt-bridge.

While our current framework does not explicitly decouple the specific contributions of translation kinetics from DNA-level evolutionary conservation, the spatial localiza-

tion of these improvements to critical functional sites is compelling. Ultimately, our findings highlight that incorporating DNA-level signals is essential for capturing the mechanistic complexity of biological systems.

Acknowledgements

Impact Statement

This paper presents a machine learning approach for predicting protein fitness under sequence mutations. While our work is primarily aimed at advancing computational biology and protein engineering, such predictive models could potentially be misused in harmful biological design contexts. We emphasize that our study is based on publicly available data and is intended solely for beneficial scientific applications. We do not identify any immediate societal risks beyond those generally associated with advances in machine learning and biotechnology.

References

Borschel, W. F., Wang, S., Lee, S., and Nichols, C. G. Control of kir channel gating by cytoplasmic domain interface interactions. *Journal of General Physiology*, 149(5):561–576, 2017.

Boshar, S., Evans, B., Tang, Z., Picard, A., Adel, Y., Lorbeer, F. K., Rajesh, C., Karch, T., Sidbon, S., Emms, D., et al. A foundational model for joint sequence-function multi-species modeling at scale for long-range genomic prediction. *bioRxiv*, pp. 2025–12, 2025.

Hansen, S. B., Tao, X., and MacKinnon, R. Structural basis of pip2 activation of the classical inward rectifier k+ channel kir2. 2. *Nature*, 477(7365):495–498, 2011.

Hayes, T., Rao, R., Akin, H., Sofroniew, N. J., Oktay, D., Lin, Z., Verkuil, R., Tran, V. Q., Deaton, J., Wiggert, M., et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Krishna, R., Wang, J., Ahern, W., Sturmfels, P., Venkatesh, P., Kalvet, I., Lee, G. R., Morey-Burrows, F. S., Anishchenko, I., Humphreys, I. R., et al. Generalized biomolecular modeling and design with rosettafold all-atom. *Science*, 384(6693):eadl2528, 2024.

Notin, P., Kollasch, A., Ritter, D., Van Niekerk, L., Paul, S., Spinner, H., Rollins, N., Shaw, A., Orenbuch, R., Weitzman, R., et al. Proteingym: Large-scale benchmarks for protein fitness prediction and design. *Advances in neural information processing systems*, 36:64331–64379, 2023.

O'Cathail, C., Ahamed, A., Burgin, J., Cummins, C., Devaraj, R., Gueye, K., Gupta, D., Gupta, V., Haseeb, M., Ihsan, M., Ivanov, E., Jayathilaka, S., Kadhirvelu, V., Kumar, M., Lathi, A., Leinonen, R., McKinnon, J., Meszaros, L., Pauperio, J., Pesant, S., Rahman, N., Rinck, G., Selvakumar, S., Suman, S., Sunthornyotin, Y., Ventouratou, M., Waheed, Z., Woollard, P., Yuan, D., Zyoud, A., Burdett, T., and Cochrane, G. The european nucleotide archive in 2024. *Nucleic Acids Research*, 53(D1):D49–D55, 01 2025. ISSN 1362-4962. doi: 10.1093/nar/gkae975. URL https://doi.org/10.1093/nar/gkae975.

A. Extended Methodology

A.1. Detailed Multi-modal Architecture

The core of our architecture relies on the integration of protein semantics and DNA-level context. The model processes these inputs through three distinct paths before fusing them into a final prediction.

Protein Context Path (Trainable). To capture protein-level semantics, pre-trained ESM3 embeddings are first normalized using LayerNorm and projected into a lower-dimensional latent space via a linear layer. From this projected space, we extract three distinct representational views to capture both macro and micro structural impacts:

- **WT Global pool:** The masked average of the entire wild-type (WT) sequence, providing the macroscopic structural context.
- **WT @ var_pos:** The specific WT embedding precisely at the mutation site, capturing the local baseline state.
- **Delta @ var_pos:** The explicit representational shift induced by the mutation, calculated as the difference between the mutant (MUT) and WT embeddings at the variant position.

DNA Context Path (Trainable). To capture the DNA-level context, nucleotide embeddings from DLM(NTv3) are processed through a two-layer dilated 1D-CNN. This trainable path distills the sequence into two targeted spatial features:

- **WT K-window:** A window-averaged representation of the local nucleotide neighborhood surrounding the variant, capturing regional translation-level constraints.
- **MUT @ var_pos:** The site-specific CNN feature extracted precisely at the variant position.

Zero-Shot DNA Context Path (Training-free). To integrate strict evolutionary and geometric priors, we extract three zero-shot signals directly from the frozen DLM latent space:

- **Contextual Deviation:** Computes how much the wild-type nucleotide at the variant position deviates from what the surrounding spatial context predicts, acting as a proxy for codon atypicality.
- **Mutational Shift:** Measures the raw distance between the WT and MUT nucleotide embeddings at the variant position, capturing the disruptiveness of the mutation.
- **Local Dispersion:** Calculates the embedding variance within a local window, reflecting the degree of local sequence conservation.

Fusion Head. The architecture combines these diverse signals by concatenating the five latent vectors (three from the PLM path, two from the trainable DNA path) with the three normalized zero-shot geometric scalars into a unified multi-modal vector. A 2-layer Multi-Layer Perceptron (MLP) with dropout regularization then maps this fused representation to the final variant fitness score ($\hat{y}$), effectively integrating high-level protein semantics with nucleotide-level context.

A.2. Detailed Training Strategy

Position-aware Evaluation. By partitioning the dataset based on residue positions rather than individual variants, we ensure the model never encounters mutations at the same site during both training and testing, thereby preventing data leakage.

Parameter-efficient Supervised Learning. Optimization is performed via AdamW with a Mean Squared Error (MSE) loss. During training, both the PLM and DLM backbones remain strictly frozen, ensuring parameter efficiency. The model is early-stopped based on the validation Spearman ρ**—calculated on a random 10% holdout from each fold's training partition—**and final performance is reported as the aggregate mean across all five folds.

B. Ablation study of DNA-view

Geometric Scalars Drive the Lift. To pinpoint the source of this performance gain, we conducted a leave-one-out ablation study on the DLM-derived inputs. Crucially, removing the three zero-shot signals resulted in a significant performance

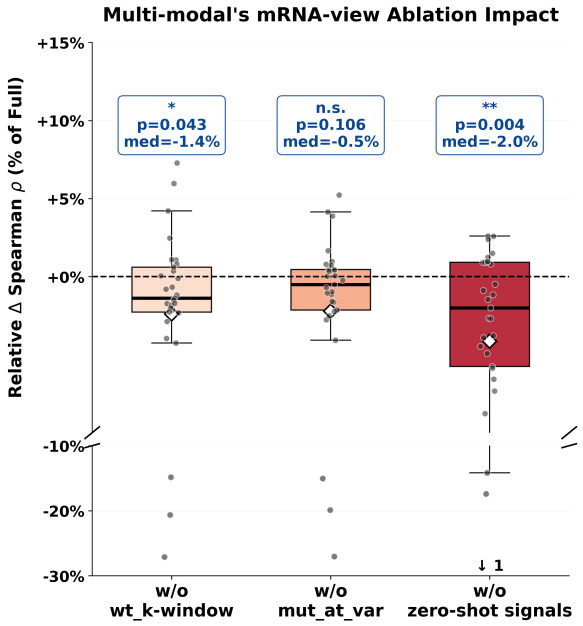

Figure 4. **Overall performance and ablation analysis.** Per-gene Spearman ρ improvement.

degradation ($p = 4.0 \times 10^{-3}$) **(Figure 4)**. In contrast, *w/o wt_k-window* showed a significant but smaller impact ($p = 0.043$), while *mut_at_var* was statistically insignificant ($p = 0.106$). This confirms that the geometric priors are the essential drivers of the cross-modal synergy.

C. Extended Mechanism Analysis Pipeline

This section provides the mathematical formulations for the variant-level decomposition and statistical testing referenced in Section 4.

C.1. Variant- and Position-Level Contribution Aggregation

Per-variant Spearman contribution. For a single-residue variant i, we define the rank residual as $e_i^m = r_i^m - r_i^{\text{DMS}}$ for model $m \in \{\text{ESM\_only}, \text{Multi-modal}\}$. The per-variant contribution is calculated as:

$$\Delta\rho_i = \frac{6}{n(n^2-1)}\left[(e_i^{\text{ESM\_only}})^2 - (e_i^{\text{Multi-modal}})^2\right]$$

This ensures $\sum_{i=1}^{n} \Delta\rho_i = \rho^{\text{Multi-modal}} - \rho^{\text{ESM\_only}}$. A positive $\Delta\rho_i$ indicates that Multi-modal ranked variant i closer to its true DMS rank than ESM_only did.

Position-level aggregation. To eliminate confounding from varying substitution counts, we compute the *mean* contribution for each residue position p:

$$\overline{\Delta\rho}_p = \frac{1}{k_p} \sum_{i \in \mathcal{V}(p)} \Delta\rho_i$$

where k_p is the number of observed variants at position p.

C.2. Selection of Synergetic & Antagonistic Positions

To account for significant variations in gene length, we select top positions proportionally. For $L_{\text{var}} \leq L$ variant-bearing positions, the threshold is:

$$K_{\text{gene}} = \max\left(1, \left\lceil L_{\text{var}} \cdot 0.05 \right\rceil\right)$$

We extract two extreme sets per gene: $\mathcal{P}^{+}_{\text{gene}}$ (Top K_{gene} positions where $\overline{\Delta\rho}_p > 0$) and $\mathcal{P}^{-}_{\text{gene}}$ (Bottom K_{gene} positions where $\overline{\Delta\rho}_p < 0$). Aggregating across 29 genes yields $|\mathcal{P}^{+}| = 482$ and $|\mathcal{P}^{-}| = 482$. The background set for statistical testing strictly excludes these extremes ($|\mathcal{B}| = 8{,}399$) to ensure mutually exclusive sets for the contingency tables.

C.3. Statistical Testing

We evaluate 16 binary biochemical features $\mathcal{F}$ for both extreme sides $s \in \{\text{top}_+, \text{top}_-\}$. The enrichment rate is calculated as:

$$\text{enr}_{f,s} \;=\; \frac{n^{\text{ext}}_{f,s} \,/\, n^{\text{ext}}_{s}}{n^{\mathcal{B}}_{f} \,/\, |\mathcal{B}|} \;=\; \frac{\Pr(f \mid s)}{\Pr(f \mid \mathcal{B})} \tag{1}$$

To distinguish Multi-modal's synergetic signals from universally sensitive positions, we compute the asymmetry ratio:

$$\text{ratio}_f \;=\; \text{enr}_{f,\text{top}_+} \,/\, \text{enr}_{f,\text{top}_-} \tag{2}$$

Statistical significance is determined via a one-sided Fisher's exact test. To strictly control the Family-Wise Error Rate (FWER) across the 32 hypotheses (16 features $\times$ 2 sides), we apply the Bonferroni threshold:

$$\alpha_{\text{Bonf}} \;=\; \frac{0.05}{32} \;\approx\; 1.56 \times 10^{-3} \tag{3}$$

Only the `is_pos_charged` feature on the top-positive side survived this highly conservative threshold ($p = 7.0 \times 10^{-4}$).

C.4. Full Feature Enrichment Analysis

Table 3 reports the full enrichment results for all 16 binary residue-level features tested in §4. Of the 32 Fisher exact tests (16 features $\times$ 2 sides), exactly one passes Bonferroni correction at $\alpha_{\text{Bonf}}{=}1.56 \times 10^{-3}$: `is_pos_charged` on the top-positive (multi-modal synergetic) side ($p_+{=}7.0 \times 10^{-4}$). Five additional tests fall below the *uncorrected* 0.05 threshold but do *not* survive multiple-testing correction (`surface_pos_chg`$_+$, `is_charged`$_+$, `is_surface`$_+$, `surface_special` on both sides, and `is_special`$_-$); their raw p-values are reported in the table for transparency but are not interpreted as discoveries. Notably, `surface_pos_chg` ($p_+{=}6.8 \times 10^{-3}$, $\text{enr}_+{=}1.38$) yields a nearly identical enrichment to the Bonferroni-significant `is_pos_charged` ($\text{enr}_+{=}1.40$); the loss of statistical significance in the surface-restricted feature is attributable to reduced sample size ($\sim$76% of K/R/H residues in $\mathcal{B}$ are surface-exposed, so the restriction discards $\sim$24% of the data while preserving the effect size). Together, these results indicate that the K/R/H signal is driven primarily by surface-exposed K/R/H residues, consistent with their canonical role in protein–partner electrostatic contacts (§5).

Reading the table. The Bonferroni-corrected significance threshold is $\alpha_{\text{Bonf}} = 0.05/32 = 1.56 \times 10^{-3}$; only `is_pos_charged`$_+$ ($p_+{=}7.0 \times 10^{-4}$) clears it. Without correction, six tests are nominally significant — but only the K/R/H result remains decisive after multiple-comparison adjustment.

The Ratio column distinguishes four mechanistic regimes (cf. §4):

- **Multi-modal Synergetic** (Ratio > 1, enrichment confined to $+$ side): `is_pos_charged` (1.37), `surface_pos_chg` (1.25), `is_charged` (1.17). The enrichment for charged residues is driven specifically by the positive subset (K/R/H); the negatively-charged subset (D/E) is at baseline (`is_neg_charged` $\text{enr}_+{=}0.99$).
- **Multi-modal Antagonistic** (Ratio < 1, enrichment confined to $-$ side): `is_special` (0.78), `buried_special` (0.45). Conformationally restrictive residues (G/P) where ESM3's evolutionary prior is strong attract.
- **Bidirectional** (both enr > 1, Ratio ≈ 1): `surface_special` ($\text{enr}_+{=}1.44$, $\text{enr}_-{=}1.55$). G/P residues exposed at the surface are sites of large model disagreement in either direction.
- **Depleted on both sides** (both enr < 1): `buried_structural` (rSASA$<0.2 \wedge$ helix/strand, partial overlap with `buried_hydro` enr$=0.80$), `is_hydrophobic` (0.89/0.91), `is_buried` (0.84/0.88), `surface_hydro` (0.94/0.94).

Robustness. The single Bonferroni-significant result (`is_pos_charged`$_+$) is robust to single-gene removal: $27/29$ leave-one-gene-out re-tests retain Bonferroni significance, and all 29 retain uncorrected $p < 0.05$. The two genes whose removal pushes p marginally above the Bonferroni threshold are KCNJ2 ($p{=}5.2 \times 10^{-3}$) and MET_HUMAN ($p{=}1.9 \times 10^{-3}$). The full LOGO table is given in Table 4.

Table 3. All 16 features tested for enrichment in extreme cases (top-5% per gene, pooled across 29 genes; $|\mathcal{P}^+|=|\mathcal{P}^-|=482$, $|\mathcal{B}|=8{,}399$). n_+ and n_- are residue counts carrying feature f within $\mathcal{P}^+$ and $\mathcal{P}^-$, respectively. $\mathrm{enr}_\pm$ and $p_\pm$ are computed as in §4. Ratio= $\mathrm{enr}_+/\mathrm{enr}_-$; values $\gg 1$ are multi-modal synergetic, $\ll 1$ are multi-modal antagonistic, ≈ 1 with both $\mathrm{enr} > 1$ are bidirectional. **Bold**: passes Bonferroni correction ($p < \alpha_{\text{Bonf}}=1.56 \times 10^{-3}$, controlling family-wise error rate across 32 tests). All raw p-values are reported for transparency; values in the range $\alpha_{\text{Bonf}} \leq p < 0.05$ do *not* survive multiple-testing correction. Features grouped by definitional category.

Feature	n_+	n_-	enr_+	enr_-	Ratio	p_+	p_-
Amino-acid class indicators							
is_pos_charged (K/R/H)	**92**	67	**1.40**	1.02	**1.37**	$\mathbf{7.0\times10^{-4}}$	0.45
is_neg_charged (D/E)	59	62	0.99	1.04	0.95	0.55	0.39
is_charged (K/R/H/D/E)	151	129	1.21	1.03	1.17	5.7×10^{-3}	0.36
is_polar (S/T/N/Q)	90	94	0.96	1.00	0.96	0.70	0.53
is_hydrophobic (A/V/I/L/M/F/W/Y/C)	187	190	0.89	0.91	0.98	0.98	0.97
is_special (G/P)	54	69	1.02	1.30	0.78	0.46	1.7×10^{-2}
SASA indicators							
is_surface (rSASA>0.3)	252	246	1.10	1.07	1.02	2.8×10^{-2}	0.08
is_buried (rSASA<0.2)	176	184	0.84	0.88	0.96	0.999	0.99
AA class ∧ surface							
surface_pos_chg	64	51	1.38	1.10	1.25	6.8×10^{-3}	0.26
surface_neg_chg	39	44	0.90	1.01	0.89	0.77	0.49
surface_charged	103	95	1.15	1.06	1.08	0.08	0.29
surface_polar	55	54	1.01	0.99	1.02	0.49	0.55
surface_hydro	55	55	0.94	0.94	1.00	0.71	0.71
surface_special	39	42	1.44	1.55	0.93	1.8×10^{-2}	4.6×10^{-3}
AA class ∧ buried							
buried_hydro	108	108	0.80	0.80	1.00	0.997	0.997
buried_special	10	22	0.45	0.99	0.45	0.999	0.56

Table 4. Leave-one-gene-out (LOGO) Fisher exact tests for the $\texttt{is\_pos\_charged}_+$ enrichment across 29 genes. Each row excludes one gene from both $\mathcal{P}^+ \cup \mathcal{P}^-$ and $\mathcal{B}$, then re-runs the aggregate Fisher exact test. Genes are sorted by gene-level $\Delta\rho_g$ (descending). $n_{\text{K/R/H}}$ is the count of K/R/H positions the removed gene contributed to $\mathcal{P}^+$. Status: *Bonf* = passes $\alpha_{\text{Bonf}}{=}1.56 \times 10^{-3}$; *uncorr.* = passes $p < 0.05$ but not Bonferroni; *n.s.* = not significant. **Bold** marks the two genes whose removal pushes p marginally above the Bonferroni threshold (still strongly significant in absolute terms). The aggregate finding is robust: $27/29$ LOGO tests pass Bonferroni, and $29/29$ pass uncorrected $p < 0.05$.

Gene removed	$\Delta\rho_g$	$n_{\text{K/R/H}}$	OR	p-value	Status
ENVZ_ECOLI_Ghose_2023	$+0.109$	0	1.51	0.0006	*Bonf*
RL40A_YEAST_Roscoe_2014	$+0.086$	2	1.47	0.0012	*Bonf*
Q8WTC7_9CNID_Somermeyer_2022	$+0.040$	2	1.51	0.0006	*Bonf*
KCNJ2_MOUSE_Coyote-Maestas_2022_function	$+0.028$	8	1.40	0.0052	*uncorr.*
CCDB_ECOLI_Adkar_2012	$+0.026$	2	1.48	0.0011	*Bonf*
SRC_HUMAN_Ahler_2019	$+0.026$	4	1.47	0.0015	*Bonf*
SRC_HUMAN_Chakraborty_2023_binding-DAS_25uM	$+0.019$	3	1.49	0.0010	*Bonf*
RNC_ECOLI_Weeks_2023	$+0.019$	3	1.49	0.0010	*Bonf*
PAI1_HUMAN_Huttinger_2021	$+0.015$	3	1.51	0.0007	*Bonf*
AMIE_PSEAE_Wrenbeck_2017	$+0.010$	1	1.54	0.0004	*Bonf*
Q6WV12_9MAXI_Somermeyer_2022	$+0.010$	2	1.50	0.0008	*Bonf*
PTEN_HUMAN_Mighell_2018	$+0.010$	4	1.52	0.0006	*Bonf*
KCNE1_HUMAN_Muhammad_2023_function	$+0.007$	2	1.48	0.0010	*Bonf*
LGK_LIPST_Klesmith_2015	$+0.006$	4	1.49	0.0011	*Bonf*
RASH_HUMAN_Bandaru_2017	$+0.005$	3	1.47	0.0014	*Bonf*
SC6A4_HUMAN_Young_2021	$+0.005$	4	1.48	0.0012	*Bonf*
MET_HUMAN_Estevam_2023	$+0.003$	5	1.45	0.0019	*uncorr.*
HEM3_HUMAN_Loggerenberg_2023	$+0.001$	2	1.54	0.0004	*Bonf*
GFP_AEQVI_Sarkisyan_2016	$+0.001$	2	1.51	0.0007	*Bonf*
PHOT_CHLRE_Chen_2023	$+0.001$	2	1.48	0.0011	*Bonf*
OXDA_RHOTO_Vanella_2023_activity	$+0.001$	3	1.52	0.0006	*Bonf*
PPARG_HUMAN_Majithia_2016	-0.001	3	1.55	0.0004	*Bonf*
A0A247D711_LISMN_Stadelmann_2021	-0.003	1	1.49	0.0008	*Bonf*
D7PM05_CLYGR_Somermeyer_2022	-0.003	2	1.51	0.0007	*Bonf*
CASP7_HUMAN_Roychowdhury_2020	-0.007	4	1.48	0.0011	*Bonf*
ADRB2_HUMAN_Jones_2020	-0.008	2	1.53	0.0005	*Bonf*
OTC_HUMAN_Lo_2023	-0.009	2	1.53	0.0005	*Bonf*
S22A1_HUMAN_Yee_2023_activity	-0.016	4	1.48	0.0012	*Bonf*
CAS9_STRP1_Spencer_2017_positive	-0.025	13	1.63	0.0002	*Bonf*

Pass Bonferroni: $27/29$; pass uncorrected $p < 0.05$: $29/29$.

D. Extended Case Study Analysis for KCNJ2

D.1. Variant-Level Quantification Framework

To quantify the predictive gap, we define the *rank residual* for each variant i as: $e_i^m = r_i^m - r_i^{\text{DMS}}$, where r_i^m is the predicted rank and r_i^{DMS} is the empirical fitness rank. A positive residual ($e_i^m > 0$) indicates an **optimistic bias**, where the model fails to recognize a deleterious variant. For instance, the KCNJ2 **R67D** variant is highly lethal ($r^{\text{DMS}} = 318/8,117$), yet ESM-only predicts a functional rank of $5,317$ ($e^{\text{ESM}} = +4,999$). Multi-modal significantly recalibrates this to 610, aligning it with the LOF regime.

D.2. Detailed Residual Stratification

Table 5 provides the complete stratification of residuals by mutation class and LOF status. The results show that ESM-only's optimism is most acute in genuinely deleterious (LOF) variants, where the mean residual inflates to $+4,103$ (roughly half the total rank range). Multi-modal consistently halves this bias across all critical categories.

Table 5. Full variant-level rank residuals for KCNJ2 K/R/H positions ($n = 141$).

Metric	ESM-ONLY	MULTI-MODAL
Mean residual $\bar{e}^m$ (ranks)	$+1,617$	$+366$
Mean magnitude $\overline{\lvert e^m \rvert}$ (ranks)	$3,284$	$1,823$
Optimism rate $\Pr[e_i^m > 0]$	72%	56%
Stratified by Mutation Class		
Charge-flip (K/R/H → D/E)	$+1,404$	-40
Lose-charge (K/R/H → neutral)	$+1,658$	$+379$
Same-charge (K/R/H → K/R/H)	$+1,490$	$+695$
Stratified by LOF Status		
LOF (DMS < median)	$+4,103$	$+1,832$
Functional (DMS ≥ median)	-834	$-1,079$

E. Limitations

While our multi-modal framework improves variant effect prediction at critical functional sites, we cannot definitively prove the underlying biological mechanism of this cross-modal synergy. Because foundation model embeddings encode highly entangled information, we cannot explicitly decouple co-translational folding kinetics from other nucleotide-level factors, such as stability or general sequence conservation. Future work will focus on applying mechanistic interpretability techniques to better isolate and understand the exact DNA-level drivers captured within the model's latent space.