Auditing Gene-Identity Shortcuts and the Equilibrium Structural Ceiling in Missense Gain- versus Loss-of-Function Prediction

Dale Richards-Smith[1,*]

[1]RS9 Systems (rs9.com.au) · [*]Correspondence: lab@rs9.com.au

November 2026

Plain English Summary

Two single-letter DNA typos inside the *same* human disease gene can have opposite physical effects: one jams a molecular switch **stuck ON** (*gain-of-function*, requiring an inhibitory blocker), while another breaks it **switched OFF** (*loss-of-function*, where giving a blocker harms the patient). Recent protein AI models trained on 112,399 human mutations (`GLOF`) reported ∼75% accuracy at telling these apart. Here we audit that benchmark and find that the 75% score was a statistical shortcut: standard models silently chop off every protein past amino acid 1,022 (throwing away 26.0% of variants and 8.4% of all known stuck-ON mutations), and when we erase the mutation completely and show the AI *only the healthy unmutated gene*, it still scores 75.6% simply by memorizing which genes tend to harbor ON versus OFF mutations. When asked to separate ON from OFF mutations **inside the same gene**, its real accuracy is 46.1%—worse than flipping a coin.

Removing whole-protein shortcuts and reading un-chopped 2,046-letter windows with a 6-billion-parameter protein model (`ESMC-6B`) lifts true same-gene accuracy from 46.1% to 63.9% (with explicit mutated-minus-healthy sequence differences adding no extra signal). We then ask whether generating 348 flexible 3D protein conformations (`Boltz-2`/`ConforFlux` + `OpenMM` + `ThermoMPNN` across 38 genes) or assembling 4-subunit potassium channel pores (140 variants) breaks past ∼64%. While 3D ensembles modestly beat single static structures on their own (60.6% vs. 56.0%, uncorrected $p = 0.119$), combining them with `ESMC-6B` plateaus at 62.5%–64.7%, and 4-chain pore assembly improves single-subunit folding (8.58 → 5.33 Å) without improving average zero-shot discrimination over static AlphaFold2 (69.7% vs. 71.8%). Together, these results establish ∼62%–65% as the empirical ceiling of unliganded, aqueous-stability-scored models and show why breaking it requires paired open-versus-closed conformational states in explicit cell-membrane potentials.

Abstract

Distinguishing gain-of-function (GoF) from loss-of-function (LoF) missense variants within the same human disease gene is critical for precision pharmacology, yet recent protein language model (pLM) classifiers evaluated under gene-disjoint splits of `GLOF`—notably `JENNIE`, which concatenates global mean-pooled `ESM-2` reference and mutant embeddings truncated at 1,022 residues—report cross-gene AUROCs near 0.75. Auditing 112,399 `GLOF` missense entries reveals three benchmark artifacts and two ablation confounders: (1) enforcing a 1,022-residue cap silently discards 29,258 variants (26.03%), including 264 pathogenic GoF alleles (8.44% of all GoF) enriched in C-terminal catalytic/regulatory domains (*MTOR*, *DICER1*, *ALK*, *SMARCA2*, *MET*, *ABCC8*, *SCN9A*, *PIK3CA*); (2) under 5-fold gene-disjoint `GroupKFold` CV ($N = 14{,}037$ pathogenic variants), feeding **only the unmutated wild-type `ESM-2` vector** achieves 0.7555 ± 0.0613 **AUROC** versus 0.7527 ± 0.0541 for reference-plus-mutant concatenation, suffers an -8.02 **pp** paralog drop under <30% sequence-identity `MMseqs2` splits (0.6725), collapses to 0.4610–0.4884 AUROC within 60 bidirectional genes (≥2 GoF, ≥2 LoF), and yields exact per-fold C-terminal randomness (0.4994 `GeneCV`, 0.5000 `HomCV`) masked by cross-fold base-rate shifts (0.4366 / 0.5241 pooled); and (3) pairing residue pLDDT_i with $\text{pLDDT}_i - \overline{\text{pLDDT}}_{\text{protein}}$ reconstructs $\overline{\text{pLDDT}}_{\text{protein}}$ ($R^2 = 1.0000$, inflating pilot `GeneCV` by +4.46 pp), while unmatched pure-LoF background cohorts (24.7k vs. 14.0k) distort mutant-delta ablations.

Eliminating whole-protein tabular constants ($R^2 < 0.08$ verified via Ridge unit tests) and expanding residue-centered wild-type representations from `ESMC-600M` (1,022-aa windows) to **May-2026** `ESMC-6B` (2,046-aa windows) lifts <30% homology cross-cluster AUROC from 0.6617 to 0.6956 (0.7052 ± 0.0432 with zero-shot predictors) and raises 60-gene out-of-homology **within-gene** AUROC from 0.4610 (`JENNIE`) and 0.5572 (`ESMC-600M`) to 0.6391 (`true_af2_local`, +17.81 pp over `JENNIE`). On matched cohorts ($N = 14{,}037$), 384-aa mutant-minus-wild-type `ESMC-6B` deltas do not improve over wild-type site context (0.6256 vs. 0.6467 across 38 structural genes, two-sided Wilcoxon $p = 0.184$). Finally, we test whether unliganded conformational ensembles break this ∼0.64 ceiling by generating 348 `Boltz-2`/`ConforFlux` flow-matching particles ($K = 6$/crop across 58 domain crops, 100% ColabFold MSA coverage) with `OpenMM` relaxation and `ThermoMPNN` $\Delta\Delta G$ scoring across 38 bidirectional genes (1,347 variants), plus a paired **4-chain `[A,B,C,D]` homo-tetrameric K^+ pore control** (140 variants across *KCNJ11*, *KCNQ2*, *KCNJ2*, *KCNA2*). Only 182/348 (52.3%) monomeric particles fold within Core-RMSD ≤ 7.5 Å. Strict-gated zero-shot $-\min_k \Delta\Delta G_k$ reaches 0.5983 within-gene AUROC (0.6099 unfiltered vs. 0.5666 static AF2), and a 6-feature pure-structural ensemble classifier reaches 0.6059 vs. 0.5595 static AF2 (+4.64 pp, 95% CI [+0.0044, +0.0954], two-sided $p = 0.119$; nonsignificant after Holm–Bonferroni correction across 24 Part B tests), but Stage-2 late fusion with `ESMC-6B` is flat (0.6249 vs. 0.6229, $p = 0.560$). In Part C, 4-chain assembly lowers chain-A Core-RMSD (8.58 → 5.33 Å; 7/24 → 18/24 particles ≤7.5 Å) and lifts single-conformer *KCNJ11* AUROC (0.6923 → 0.8308 at 8.995 Å), yet collapses ensemble diversity on tightly folded *KCNA2* (0.776 Å, 0.5333 → 0.4000), yielding no net 3-channel gain over static AF2 (0.6967 tetramer vs. 0.7011 monomer vs. 0.7178 static). These negative and boundary results bound the equilibrium sequence-and-soluble-structure ceiling at ∼0.62–0.65 within-gene AUROC.

1 Introduction

Pathogenic missense mutations in the same human gene frequently diverge into opposite functional mechanisms with mutually exclusive clinical interventions [1–4]. In *KCNJ11* (encoding the Kir6.2 pore-forming subunit of pancreatic K_{ATP} channels), activating GoF mutations impair ATP-mediated channel closure and cause neonatal diabetes mellitus treatable with oral sulfonylureas, whereas inactivating LoF mutations cause congenital hyperinsulinism [5]. Analogous bidirectional phenotypes govern clinical decision-making in *SCN9A* pain channelopathies, *STAT1* primary immunodeficiencies, *GNAO1* neurodevelopmental movement disorders, *MORC2* axonal neuropathies, and *RET* / *MTOR* / *PIK3CA* oncogenic versus developmental syndromes [1, 4]. Standard zero-shot pathogenicity predictors—including `AlphaMissense` [6], `EVE` [7], and `ESM-1b/ESM-2` evolutionary log-likelihood ratios [8, 9]—estimate the strength of purifying selection against a substitution. Because both hypermorphic GoF and hypomorphic LoF alleles depress reproductive fitness, evolutionary conservation strongly separates pathogenic from benign variants while remaining largely orthogonal to functional *directionality* within a single protein [4, 10].

To bridge this gap, Maricato, Schlesinger, and de Souza Moura curated the `GLOF` resource integrating functionally characterized human GoF and LoF variants with structural and evolutionary annotations [4], and the companion `JENNIE` study evaluated classifiers on concatenated global mean-pooled `ESM-2` (650M) wild-type ($\bar{\mathbf{h}}_{\mathrm{WT}}$) and mutant ($\bar{\mathbf{h}}_{\mathrm{Mut}}$) sequence embeddings under 5-fold gene-disjoint cross-validation, reporting cross-gene AUROCs of 0.74–0.76 [11]. Parallel work in cancer genomics (`OncoCal`, Lee 2026) demonstrated that pan-gene decision thresholds systematically conflate *gene-level* mechanism priors (e.g., oncogenes versus tumor suppressors) with *variant-level* functional effects [10], while MacLean et al. (2026) showed that genetic-code mutational accessibility (1- vs. 2–3-nucleotide codon transitions) independently shapes the clinical enrichment of specific missense substitutions [12].

Here we ask two connected questions: first, whether global pLM mean pooling genuinely resolves within-gene GoF versus LoF mechanisms or merely memorizes gene and paralog priors; and second, whether un-truncated foundation pLMs and unliganded 3D conformational ensembles can overcome static ground-state limitations. Auditing 112,399 `GLOF` variants and scaling evaluation across **May-2026** `ESMC-6B` [13] and `Boltz-2` [14] particle-guided `ConforFlux` conformational ensembles [15] scored with `ThermoMPNN` [16], we establish four findings—combining a decisive benchmark audit (Part A) with rigorously instrumented boundary and negative results on equilibrium conformational sampling (Parts B and C):

1. **Cross-gene pLM mean pooling is a whole-protein shortcut:** Global `ESM-2` pooling ($\leq$1,022 aa) discards 26.03% of `GLOF` variants (8.44% of all GoF), drops -8.02 pp under $<$30% sequence-identity homology splits, performs slightly better when the mutant sequence is deleted entirely (0.7555 WT-Only vs. 0.7527 Ref+Mut), and collapses to 0.4610–0.4884 AUROC within bidirectional genes.

2. **Hidden leakage and cohort traps in residue-level ablations:** Pairing residue pLDDT_i with protein-centered $\mathrm{pLDDT}_i - \overline{\mathrm{pLDDT}}_{\mathrm{protein}}$ reconstructs $\overline{\mathrm{pLDDT}}_{\mathrm{protein}}$ ($R^2 = 1.0000$), inflating pilot cross-gene AUROC by +4.46 pp. Moreover, because site embeddings still identify protein neighborhoods across genes, **within-gene AUROC** is the sole strictly gene-prior-immune endpoint, and comparing models across unmatched pure-LoF cohorts (24.7k vs. 14.0k) creates spurious ablation shifts.

3. **Joint context-window and representation expansion (`ESMC-600M`/1,022 aa $\rightarrow$ `ESMC-6B`/2,046 aa) lifts within-gene AUROC to $\sim$0.64, while explicit mutant deltas add no net gain:** Across zero-constant tabular tiers, moving from 1,022-aa `ESMC-600M` to un-truncated 2,046-aa `ESMC-6B` wild-type site windows raises 60-gene out-of-homology within-gene AUROC from 0.5572 to 0.6190 (`seq_codon_only`) and 0.6391 (`true_af2_local`, +17.81 pp over `JENNIE`), whereas 384-aa mutant-minus-wild-type `ESMC-6B` deltas yield no additive improvement on matched cohorts (0.6256 vs. 0.6467, $p = 0.184$).

4. **Unliganded conformational ensembles and homo-tetrameric pores expose an equilibrium aqueous-energetics ceiling at $\sim$0.62–0.65:** Across 38 bidirectional genes (1,347 variants, 58 domain crops, 348 particles), 52.3% (182/348) of monomeric `Boltz-2`/`ConforFlux` particles fold within 7.5 Å. While pure-structural ensemble features lift within-gene AUROC by +4.64 pp over static AF2 (0.6059 vs. 0.5595, two-sided $p = 0.119$), no Part B comparison survives multiple-testing correction, late fusion with `ESMC-6B` is flat (0.6249 vs. 0.6229, $p = 0.560$), isolated membrane subunits suffer aqueous interface exposure (*CHRNE* -44.0 pp, *KCNJ2* -20.7 pp despite folded backbones), and 4-chain homo-tetramer pore assembly lowers chain-A RMSD (8.58 $\rightarrow$ 5.33 Å) without improving pooled 3-channel zero-shot discrimination over static AF2 (0.6967 vs. 0.7178).

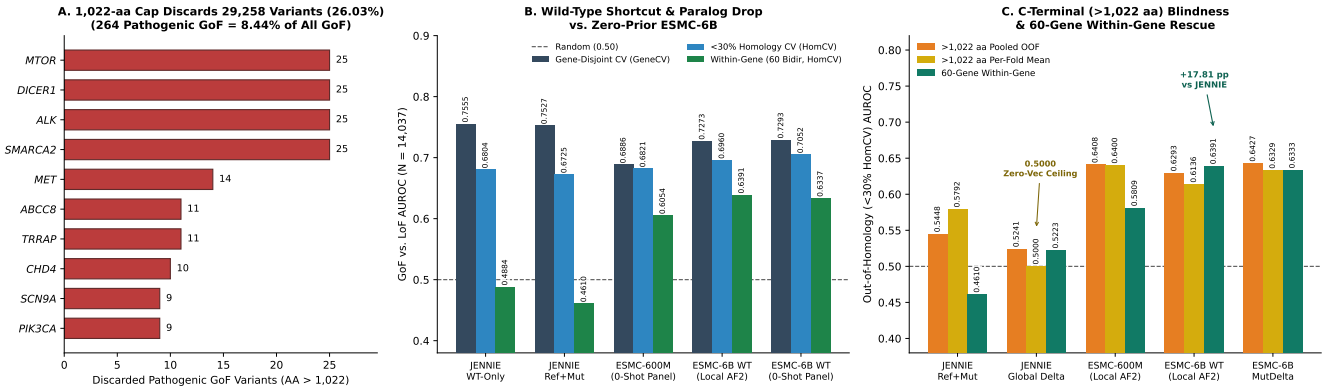


Figure 1. Three benchmark artifacts in global 1,022**-aa pLM GoF/LoF classification (**`GLOF`**,** $N = 112{,}399$**). (A)** Silent C-terminal truncation (`AA_POSITION` $>$ 1,022) discards 29,258 variants (26.03% of `GLOF`), including 264 pathogenic GoF alleles (8.44% of all known human GoF variants) across major clinical genes (*MTOR*, *DICER1*, *ALK*, *SMARCA2*, *MET*, *ABCC8*, *TRRAP*, *CHD4*, *SCN9A*, *PIK3CA*). **(B)** Under $N = 14{,}037$ pathogenic GoF/LoF variants (60 bidirectional genes with $\geq$2 GoF and $\geq$2 LoF), `JENNIE` trained on the **unmutated wild-type sequence alone** (`WT-Only`, 0.7555) matches or exceeds `Ref+Mut` (0.7527), drops −8.02 pp under <30% sequence-identity `MMseqs2` homology splits (0.6725), and collapses below random (0.4610) within individual bidirectional genes, whereas un-truncated `ESMC-6B` site windows achieve 0.6960–0.7052 <30% `HomCV` and 0.6337–0.6391 within-gene AUROC. **(C)** Out-of-homology C-terminal (>1,022 aa) pooled OOF AUROC vs. per-fold mean and 60-gene mean within-gene AUROC, highlighting the exact 0.5000 per-fold random ceiling of `JENNIE Global Delta` above residue 1,022 and the +17.81 pp within-gene rescue by `ESMC-6B` (`true_af2_local`).

2 Results

2.1 Audit of the `GLOF` Benchmark and `JENNIE` Paradigm (112,399 Variants)

We audited all 112,399 human missense entries in `GLOF` (3,128 pathogenic GoF, 13,864 pathogenic LoF, and 95,407 neutral/population variants across 1,352 genes) prior to model training (**Figure 1**; **Table 1**).

First, standard `ESM-2` positional encodings impose a hard 1,024-token ceiling (1,022 amino acids excluding `<cls>` and `<eos>`). Across `GLOF`, 29,258 **of** 112,399 **variants (**26.03%**) reside at** `AA_POSITION` $>$ 1,022, comprising 264 pathogenic GoF (8.44% of all GoF alleles in the database), 3,621 pathogenic LoF, and 25,373 neutral variants. These discarded variants are heavily enriched in clinically actionable C-terminal kinase, RNase III, ATPase, and pore-regulatory domains, led by *MTOR* (25 GoF), *DICER1* (25 GoF), *ALK* (25 GoF), *SMARCA2* (25 GoF), *MET* (14 GoF), *ABCC8* (11 GoF), *TRRAP* (11 GoF), *CHD4* (10 GoF), *SCN9A* (9 GoF), and *PIK3CA* (9 GoF).

Second, to hold downstream classifier capacity strictly constant across representations, we evaluated `JENNIE`'s 2,560-D global mean-pooled concatenation $[\bar{\mathbf{h}}_{\mathrm{WT}} \,\|\, \bar{\mathbf{h}}_{\mathrm{Mut}}]$ from `facebook/esm2_t33_650M_UR50D` (truncated to 1,022 residues) alongside two diagnostic controls on $N = 14{,}037$ UniProt-mapped pathogenic GoF/LoF variants (retaining all variants in any gene with $\geq$1 GoF allele and deterministically capping pure-LoF genes with zero GoF alleles at $\leq$10 variants; 60 genes harbor $\geq$2 GoF and $\geq$2 LoF variants for within-gene AUROC evaluation). Each representation was evaluated under both 5-fold gene-disjoint `GroupKFold` CV (`GeneCV`) and 5-fold <30% **sequence-identity homology-cluster** `GroupKFold` **CV** (`HomCV`, `MMseqs2 easy-cluster -min-seq-id 0.30 -c 0.8`) (**Table 1**):

- **Wild-type-only shortcut (`JENNIE WT-Only`):** Feeding *only* the unmutated wild-type vector $\bar{\mathbf{h}}_{\mathrm{WT}} \in \mathbb{R}^{1280}$—so every GoF and LoF variant in the same gene receives the exact same feature vector—yields 0.7555 ± 0.0613 `GeneCV` **AUROC**, slightly outperforming `JENNIE Ref+Mut` (0.7527 ± 0.0541). On the 60 bidirectional genes ($\geq$2 GoF, $\geq$2 LoF), both `JENNIE WT-Only` (0.4822 `GeneCV` / 0.4884 `HomCV`) and `JENNIE Ref+Mut` (0.4713 `GeneCV` / 0.4610 `HomCV`) score **below random** (0.5000) within genes. Because global mean pooling divides single-residue perturbations by $L \sim 500$–2,500, learners ignore the $O(1/L)$ mutation signal and predict each protein's GoF:LoF class ratio from its wild-type sequence family.
- **Paralog family leakage:** Transitioning from gene-disjoint splits (`GeneCV`) to <30% sequence-identity homology clusters (`HomCV`) reduces `JENNIE Ref+Mut` AUROC by −8.02 **pp** ($0.7527 \rightarrow 0.6725 \pm 0.0899$) and `JENNIE WT-Only` by −7.51 **pp** ($0.7555 \rightarrow 0.6804 \pm 0.0873$). Under gene-disjoint CV, close paralogs with conserved mechanism priors (e.g., Na^+/K^+ channel families *SCN1A/2A/8A/9A* and *KCNQ1/2/3*, or receptor tyrosine kinases

Table 1. Cross-gene (GeneCV), <30% sequence-identity homology-cluster (HomCV), C-terminal (>1,022 aa), and within-gene discrimination across $N = 14{,}037$ pathogenic GoF/LoF variants (60 bidirectional genes with ≥2 GoF and ≥2 LoF). All ESMC models enforce zero-gene-prior *tabular* features (out-of-sample Ridge $R^2 < 0.08$ against whole-protein length L and $\overline{\text{pLDDT}}_{\text{protein}}$) and strictly fold-internal median imputation + 96-component PCA inside every GroupKFold split. Note that ESMC-600M uses 1,022-aa sliding windows whereas ESMC-6B uses 2,046-aa sliding windows; cross-cluster metrics still reflect sequence-neighborhood identity in the 96-D PCA block, making within-gene AUROC (rightmost column) the sole strictly gene-prior-immune endpoint.

Model / Feature Representation	Tabular Tier	GeneCV AUROC	<30% HomCV AUROC	Paralog Drop (Δ)	HomCV >1,022 aa Pooled (Fold)	60-Gene Within-Gene AUROC (HomCV)
JENNIE WT-Only ($\bar{\mathbf{h}}_{\text{WT}}$, no mutation info)	Global ESM-2 (≤1022)	0.7555 ± 0.0613	0.6804 ± 0.0873	−0.0751	0.5541 (0.5844)	0.4884
JENNIE Ref+Mut ($[\bar{\mathbf{h}}_{\text{WT}} \,\|\, \bar{\mathbf{h}}_{\text{Mut}}]$)	Global ESM-2 (≤1022)	0.7527 ± 0.0541	0.6725 ± 0.0899	−0.0802	0.5448 (0.5792)	0.4610
JENNIE Global Delta ($\bar{\mathbf{h}}_{\text{Mut}} - \bar{\mathbf{h}}_{\text{WT}}$)	Global ESM-2 (≤1022)	0.6197 ± 0.0292	0.5875 ± 0.0583	−0.0322	0.5241 (0.5000)	0.5223
ESMC-600M WT-Site (1,022-aa sliding window)	seq_codon_only	0.6694 ± 0.0446	0.6617 ± 0.0582	−0.0077	0.6385 (0.6387)	0.5572
ESMC-600M WT-Site (1,022-aa sliding window)	true_af2_local	0.6768 ± 0.0454	0.6641 ± 0.0617	−0.0127	0.6408 (0.6400)	0.5809
ESMC-600M WT-Site (1,022-aa sliding window)	full_zeroshot_panel	0.6886 ± 0.0474	0.6821 ± 0.0607	−0.0065	**0.6533** (**0.6511**)	0.6054
ESMC-6B WT-Site (2,046-aa sliding window)	seq_codon_only	0.7262 ± 0.0387	0.6956 ± 0.0361	−0.0306	0.6281 (0.6125)	0.6190
ESMC-6B WT-Site (2,046-aa sliding window)	true_af2_local	0.7273 ± 0.0364	0.6960 ± 0.0432	−0.0313	0.6293 (0.6136)	**0.6391**
ESMC-6B WT-Site (2,046-aa sliding window)	full_zeroshot_panel	0.7293 ± 0.0425	$\mathbf{0.7052 \pm 0.0432}$	−0.0241	0.6210 (0.6014)	0.6337
ESMC-6B WT + Nested MutDelta Probe	seq_codon_only	0.7308 ± 0.0395	0.6946 ± 0.0337	−0.0362	0.6303 (0.6154)	0.6259
ESMC-6B WT + Nested MutDelta Probe	true_af2_local	0.7300 ± 0.0379	0.6952 ± 0.0402	−0.0348	0.6427 (0.6329)	0.6333
ESMC-6B WT + Nested MutDelta Probe	full_zeroshot_panel	$\mathbf{0.7316 \pm 0.0419}$	0.7044 ± 0.0391	−0.0272	0.6280 (0.6092)	0.6285
ESMC-6B Raw Concat PCA (64 + 48 + 32) Ablation	full_zeroshot_panel	0.7282 ± 0.0435	0.7029 ± 0.0435	−0.0253	0.6497 (0.6290)	0.6270

FGFR1/2/3 and *RET/ALK/MET*) split across train and test folds, allowing wild-type embeddings to transfer paralog class priors.

- **Zero-vector C-terminal blindness and fold-intercept pooling:** Subtracting wild-type global vectors (JENNIE Global Delta, $\bar{\mathbf{h}}_{\text{Mut}} - \bar{\mathbf{h}}_{\text{WT}}$) eliminates the wild-type identity vector, causing GeneCV AUROC to drop −13.30 pp to 0.6197 ± 0.0292 (0.5875 ± 0.0583 HomCV). For all variants at AA_POSITION > 1,022, 1,022-aa truncation renders wild-type and mutant prefixes identical ($\Delta\bar{\mathbf{h}} \equiv \mathbf{0}$). Notably, evaluating pooled out-of-fold probabilities across CV folds with unequal GoF base rates produces spurious non-0.50 pooled AUROCs (0.4366 GeneCV, 0.5241 HomCV) on identical zero vectors solely due to per-fold intercept shifts, whereas per-fold C-terminal AUROC (computed across folds containing ≥4 C-terminal variants spanning both classes) confirms strict randomness (0.4994 **GeneCV,** 0.5000 **HomCV**). Crucially, within-gene AUROC is mathematically immune to this cross-fold intercept artifact because GroupKFold places all variants of any single gene into the same held-out fold.

2.2 Zero-Constant Tabular Tiers, ESMC-6B Context+Scale Expansion, and Matched-Cohort MutDelta Ablations

During adversarial auditing of our preliminary (v2 pilot) feature pipeline, we uncovered two subtle methodological traps that distort GoF/LoF ablations (**Figure 2**):

1. **Linear-combination gene-prior leakage (ΔpLDDT):** Even when explicit protein length L, relative position i/L, and protein-mean $\overline{\text{pLDDT}}_{\text{protein}}$ are excluded, including both residue pLDDT_i and its protein-centered offset $\Delta\text{pLDDT}_i = \text{pLDDT}_i - \overline{\text{pLDDT}}_{\text{protein}}$ allows axis-aligned tree splits and linear models to recover $\overline{\text{pLDDT}}_{\text{protein}} = \text{pLDDT}_i - \Delta\text{pLDDT}_i$ exactly ($R^2 = 1.0000$). Removing ΔpLDDT_i between our pre-v3 pilot (0.7332 GeneCV) and the permanently sanitized v3 pipeline (0.6886 GeneCV on ESMC-600M full_zeroshot_panel) revealed that +4.46 pp of apparent cross-gene performance had come from whole-protein disorder priors rather than local residue context. Importantly, while our automated Ridge regression unit test verifies $R^2 < 0.08$ against L and $\overline{\text{pLDDT}}_{\text{protein}}$ across all *tabular* feature tiers, the 96-component PCA projection of the ESMC residue embedding still encodes local domain/motif identity in cross-cluster CV; consequently, **within-gene AUROC** is the only metric completely immune to protein-level priors.
2. **Separating local AlphaFold2 geometry from external foundation-model scores:** Benchmarks such as GLOF bundle external variant predictors (AlphaMissense, ESM-IF1, ESM-1v, ESM-2, ESM-3, Evo-2) alongside structural coordinates. Grouping them under a single "Static AF2" label obscures how much signal comes from sequence context versus local 3D coordinates versus external evolutionary conservation.

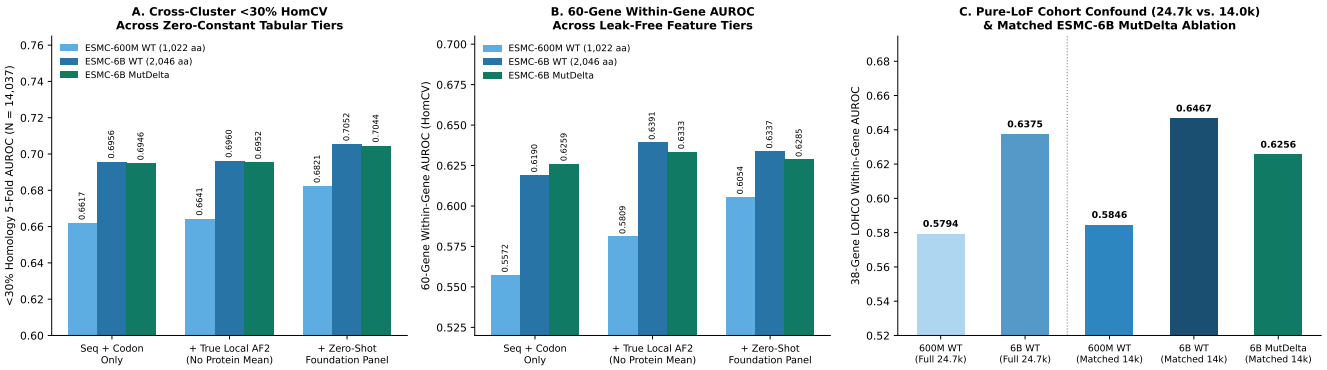


Figure 2. Zero-constant tabular-tier decomposition and matched-cohort resolution of the `ESMC-6B` mutant-delta (`MutDelta`) confound. **(A)** <30% homology-cluster (`HomCV`) cross-cluster AUROC across `seq_codon_only`, `true_af2_local`, and `full_zeroshot_panel` tiers ($N = 14{,}037$). **(B)** Out-of-homology 60-gene within-gene AUROC across the same three tiers, peaking at 0.6391 for `ESMC-6B WT + true_af2_local`. **(C)** Controlled Stage-1 out-of-homology within-gene AUROC on the 38 structural benchmark genes ($N = 1{,}347$ variants): expanding `ESMC-600M` (1,022 aa) → `ESMC-6B WT` (2,046 aa) delivers +5.81 pp on the full 24.7k cohort (0.5794 → 0.6375) and +6.21 **pp** on the matched 14.0k cohort (0.5846 → 0.6467), whereas comparing matched 14.0k `WT` (0.6467) against `MutDelta` (0.6256, −2.11 pp, two-sided $p = 0.184$) shows that wild-type site context—not explicit 384-aa mutant embedding subtraction—drives within-gene generalization.

We therefore evaluated `biohub/ESMC-600M` ($d = 1{,}152$, extracted over 1,022-aa sliding tiles snapped to 256-aa steps) and **May-2026** `biohub/ESMC-6B` ($d = 2{,}560$, extracted over 2,046-aa sliding tiles snapped to 512-aa steps; both projected to 96 fold-internal `PCA` components) across three nested zero-constant tabular tiers (**Table 1**, **Figure 2A–B**). Because window length (1,022 → 2,046 aa) and model capacity (600M → 6B) advance together between these two configurations, the gains below reflect **joint representation-scale and context-window expansion** rather than parameter count at fixed window size:

- **Tier 1 (`seq_codon_only`, 52 columns = 46 sequence/codon + 6 WT site-logit scalars):** Pure amino-acid one-hot, physicochemical substitution properties, genetic-code minimum nucleotide distance [12], and wild-type site entropy/log-odds scalars. Using *zero* structural coordinates and *zero* external predictors, expanding from `ESMC-600M` (1,022 aa) to `ESMC-6B` (2,046 aa) lifts `GeneCV` by +5.68 **pp** (0.6694 → 0.7262), <30% `HomCV` by +3.39 **pp** (0.6617 → 0.6956), and 60-gene out-of-homology **within-gene** AUROC by +6.18 **pp** (0.5572 → 0.6190; 0.6008 → 0.6470 under `GeneCV`).
- **Tier 2 (`true_af2_local`, +5 local AF2 features):** Adding *only* local static AlphaFold2 geometry (`plddt_at_pos` and ±5/±15-residue local window offsets, with zero whole-protein mean) lifts `ESMC-6B` out-of-homology within-gene AUROC to 0.6391 (+17.81 pp above `JENNIE Ref+Mut` 0.4610 and +5.82 pp above `ESMC-600M` `true_af2_local` 0.5809).
- **Tier 3 (`full_zeroshot_panel`, +8 conservation/structure interaction features):** Adding `AlphaMissense`, `ESM-IF1`, `ESM-1v`, `ESM-2`, `ESM-3`, `Evo-2`, and disorder-interaction terms lifts <30% cross-cluster `HomCV` AUROC to 0.7052 ± 0.0432 (+3.27 pp over `JENNIE Ref+Mut`) while slightly *lowering* within-gene ranking (0.6391 → 0.6337), directly illustrating that stronger evolutionary conservation improves cross-protein pathogenicity calibration without resolving within-gene functional sign.

Matched-cohort resolution of `ESMC-6B` mutant-delta perturbations (Figure 2C): To test whether explicit contextual mutant-minus-wild-type pLM deltas ($\Delta \mathbf{h}_i = \mathbf{h}_{\mathrm{Mut},i} - \mathbf{h}_{\mathrm{WT},i}$ over exact 384-aa windows) add signal beyond wild-type site embeddings plus exact amino-acid substitution features, we evaluated both a nested inner-4-fold `GroupKFold` ℓ_2-logistic probe over 80 fold-internal `PCA` components (48 site-delta + 32 local-pocket-delta) plus epistatic scalar invariants, and a direct 144-D raw `PCA` concatenation control (64 WT + 48 site-delta + 32 pocket-delta, bypassing the nested probe). In Part A ($N = 14{,}037$), `ESMC-6B MutDelta` yields a +0.69 pp lift under `seq_codon_only` (0.6190 → 0.6259) and +1.34 pp on C-terminal variants under `true_af2_local` (0.6293 → 0.6427), but neither the nested-probe model (0.6285) nor the raw concatenation control (0.6270) improves within-gene AUROC over `ESMC-6B WT` once local AF2 geometry or zero-shot scores are included (0.6391 / 0.6337). On the 38 Part B structural genes ($N = 1{,}347$ variants), training Stage-1 on all 24,700 mapped variants (∼87% pure-LoF) yields 0.6375 (`6B WT`) vs. 0.5794 (`600M WT`), whereas training on the matched $N = 14{,}037$ cohort lifts `ESMC-6B`

Table 2. Domain-cropped `Boltz-2/ConforFlux` + `ThermoMPNN` biophysical evaluation across all 38 bidirectional structural benchmark genes (1,347 missense variants). `LOHCO` holds out the target gene's entire <30% sequence-identity cluster; `LOGO` holds out only the target gene (10,000 paired gene-level bootstrap resamples; two-sided Wilcoxon signed-rank p-values reported throughout). Stage-1 rows evaluate a single precomputed out-of-homology log-odds column and are therefore mathematically invariant between `LOHCO` and `LOGO`. Across all 24 bootstrap comparisons in the Part B evaluation family, **no comparison survives Holm–Bonferroni multiple-testing correction** ($\min p_{\text{two-sided}} = 0.119$).

Evaluation Layer & Method ($N = 38$ **Bidirectional Genes:** 17 **Soluble/1-TM** + 21 **Multi-Pass TM**)	<30% **Homology** (`LOHCO`) **AUROC**	**Leave-One-Gene-Out** (`LOGO`) **AUROC**	**Difference vs. Baseline** (`LOHCO` 95% **CI; Two-Sided** p)
1. Unsupervised Zero-Shot Within-Gene Ranking (No Supervised Fitting)			
−AlphaMissense (pathogenicity inversion baseline)	0.5158	0.5158	−0.0508 vs. Static AF2
$-\Delta\Delta G_{\text{Static AF2}}$ (`ThermoMPNN` on static AF2)	0.5666	0.5666	Baseline (0.0000)
$-\min_k \Delta\Delta G_k$ (`TM-Topology-Gated`, $n_{\text{TM}} \geq 2 \rightarrow$ static)	0.5774	0.5774	+0.0108 vs. Static AF2
$-\min_k \Delta\Delta G_k$ (Strict ≤7.5 Å Gate, pooled with $\Delta\Delta G_{\text{Static}}$)	0.5983	0.5983	+0.0317 vs. Static AF2
$-\min_k \Delta\Delta G_k$ (Unfiltered 6-Particle Ensemble, includes > 7.5 Å decoys)	0.6099	0.6099	+0.0433 (includes unfolded decoys)
2. Pure Structural Classifier (2 Static vs. 6 Strict vs. 5 TM-Gated Biophysical Features, Zero Sequence Input)			
Static AF2 Classifier ($\Delta\Delta G_{\text{static}}$, 10 Å Cα neighbors; 2 cols)	0.5595	0.5601	Baseline (0.0000)
TM-Topology-Gated Structural Classifier (5 cols; $n_{\text{TM}} \geq 2$ gated)	0.5606	0.5587	+0.0011 ([−0.0179, +0.0220], $p = 0.920$)
Domain-Cropped Strict (≤7.5 Å) Ensemble Classifier (6 cols)	**0.6059**	**0.6059**	+0.0464 ([+0.0044, +0.0954], $p = 0.119$; **1-sided** 0.060)
3. Stage-1 Sequence Context Alone (1 Out-of-Homology Log-Odds Feature; `LOHCO` ≡ `LOGO`)			
`ESMC-600M` WT-Site (1,022 aa; `Full 24.7k` / `Matched 14.0k`)	0.5794 / 0.5846	0.5794 / 0.5846	Cohort shift +0.0052
`ESMC-6B` WT-Site (2,046 **aa;** `Full 24.7k` **/** `Matched 14.0k`)	0.6375 / **0.6467**	0.6375 / **0.6467**	Cohort shift +0.0092 ([−0.0222, +0.0390], $p = 0.330$)
`ESMC-6B` WT + `MutDelta` (`Matched 14.0k`, Stage-1 input to Layer 4)	0.6256	0.6256	−0.0211 vs. `6B WT` ([−0.0451, +0.0017], $p = 0.184$)
4. Stage-2 Sequence + Structure Late Fusion (3 Static vs. 5 Ensemble Features)			
Stage-1 `ESMC-6B` + Static AF2 `ThermoMPNN` (3 cols)	0.6229	0.6245	Baseline (−0.0027 vs. Stage-1, $p = 0.934$)
Stage-1 `ESMC-6B` + Unfiltered Monomer Ensemble (5 cols)	0.6223	0.6211	−0.0006 ([−0.0469, +0.0436], $p = 0.860$)
Stage-1 `ESMC-6B` + Monomer Strict ≤7.5 Å Gate (5 cols)	0.6164	0.6143	−0.0066 ([−0.0425, +0.0259], $p = 0.678$)
Stage-1 `ESMC-6B` + TM-Topology-Gated Ensemble (5 cols, $n_{\text{TM}} \geq 2$)	0.6249	0.6268	+0.0020 ([−0.0191, +0.0249], $p = 0.560$)
↪ All Soluble & Single-Pass Genes ($N = 17$: 14 folded + 3 unfolded)	0.5685 → 0.5734	0.5696 → 0.5747	+0.0049 overall (3 unfolded revert to static)
↪ Folded Soluble & Single-Pass Subset ($N = 14$ genes)	0.5767 → 0.5860	0.5779 → 0.5876	+0.0094 ([−0.0449, +0.0683], $p = 0.946$)
↪ Multi-Pass TM Subset ($N = 21$: 19 folded + 2 unfolded; Static → Strict / TM-Gate)	0.6669 → 0.6519 / 0.6665	0.6690 → 0.6479 / 0.6689	Strict hurts *CHRNE* (−0.440) & *KCNJ2* (−0.207)
↪ Large Proteins Subset (>1,022 aa, $N = 11$ genes; Strict / TM-Gate)	0.5232 / 0.5318	0.5232 / 0.5318	Remains near-random on giant multi-domain proteins

`WT` to 0.6467 (+6.21 pp over matched `600M WT` 0.5846) versus 0.6256 for `ESMC-6B MutDelta` (−2.11 pp, 95% CI [−0.0451, +0.0017], two-sided Wilcoxon $p = 0.184$). We note one implementation nuance in the nested probe (Section 4.3): training rows receive out-of-fold inner-probe log-odds whereas test rows receive full-training-fold probe log-odds, which can mildly attenuate learner weight on the probe column; however, the probe-free 144-D raw concatenation control (0.6270 vs. 0.6337 WT) independently confirms the absence of an additive `MutDelta` benefit.

2.3 Domain-Cropped `Boltz-2/ConforFlux` Ensembles and the Equilibrium Late-Fusion Ceiling (38 Genes, 1,347 Variants)

Static sequence context and single-state AlphaFold2 coordinates represent a single equilibrium minimum. In principle, an activating GoF mutation in an autoinhibitory latch could destabilize an inactive conformation ($\Delta\Delta G_{\text{inactive}} > 0$) while sparing an alternative conformation ($\min_k \Delta\Delta G_k \leq 0$). To test rigorously whether unliganded flow-matching ensembles capture this effect in practice, we identified all 38 UniProt-mapped human disease genes in `GLOF` with length ≤2,600 aa, ≥3 pathogenic GoF, and ≥3 pathogenic LoF variants (1,347 variants total; 0 genes dropped during single-fragment AlphaFold DB coordinate matching). Across proteins up to 2,521 residues (*PIEZO1*, *CACNA1A*, *FN1*, *F5*, *CACNA1I*, *ABCC8*), our pLDDT-guided domain cropper produced 58 deduplicated crops (≤360 aa linker-snapped crops, or a single structured crop up to 402 aa when trimmed span ≤410 aa), achieving 58/58 (100.0%) `ColabFold` MSA coverage (0 single-sequence fallbacks). For each crop, we generated $K = 6$ particle-guided conformers via `Boltz-2` [14] (348 conformer particles total), relaxed them in `OpenMM` (`Amber14`), and scored all 1,347 variants via `ThermoMPNN` (**Table 2**; **Figure 3A**).

Auditing both structural fidelity and statistical uncertainty across **Table 2** yields three candid lessons:

1. **Particle-level folding fidelity (52.3%) is much stricter than gene-level coverage (86.8%):** While 33/38 (86.84%) genes produce *at least one* particle with Core-RMSD ≤ 7.5 Å (e.g., *TSHR* 1.18 Å [12/12], *PRPS1* 1.87 Å [6/6], *CASR* 3.75 Å [10/12], *F5* 4.86 Å [10/12], *CACNA1I* 6.83 Å [8/12], *MORC2* 7.59 Å [3/6]), across all 348 generated particles only 182/348 (52.3%) satisfy Core-RMSD ≤ 7.5 Å. Ten of the 33 "covered" genes exhibit mean Core-RMSD > 10 Å due to unfolded particles in two-crop or multi-TM topologies (*KCNQ2* 20.29 Å [2/12 kept], *KCNQ1* 17.91 Å [3/12], *FN1* 16.49 Å [2/12], *ATP1A1* 14.45 Å [4/12], *CLCN2* 13.70 Å [3/12], *RET* 12.50 Å [6/12], *FGFR2* 12.38 Å [6/12], *FGFR1* 11.94 Å [6/12], *CACNA1A* 11.83 Å [1/12], and *MYBPC3* 10.11 Å [2/12]). Five genes yielded 0 particles ≤7.5 Å: lipid-free *APOA1* (37.30 Å), 38-TM *PIEZO1* (24.14 Å), heterotrimeric *SCNN1B* (25.78 Å), pro-domain *TGFB1* (19.00 Å), and soluble chromatin subunit *SMARCB1* (385 aa, 26.18 Å,

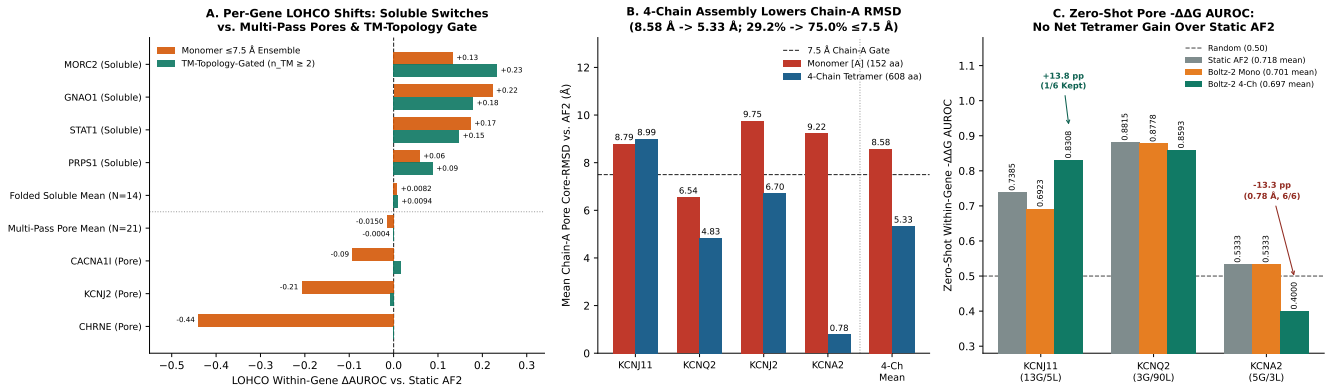


Figure 3. Multi-pass membrane aqueous-exposure artifacts and the 4-chain homo-tetramer K^+ channel pore control. (A) Per-gene `LOHCO` within-gene AUROC shift (Δ over static AF2) across soluble/single-pass switches (*MORC2*, *GNAO1*, *STAT1*, *PRPS1*; 14-gene mean +0.0094, $p = 0.946$) versus multi-pass pores (*CHRNE*, *KCNJ2*, *CACNA1I*), where folded monomers suffer severe regressions unless masked by the $n_{\mathrm{TM}} \geq 2$ topology gate. **(B)** Paired monomeric (152 aa) vs. 4-chain `[A, B, C, D]` homo-tetrameric (608 aa) chain-A Core-RMSD across the 4 symmetric K^+ channels (8.58 Å [7/24, 29.2%] → 5.33 Å [18/24, 75.0%] ≤7.5 Å). **(C)** Zero-shot within-gene $-\Delta\Delta G$ AUROC across the 3 bidirectional pore modules, showing **no net 3-channel gain over static AF2** (0.7178 static > 0.7011 monomer > 0.6967 tetramer): single-conformer *KCNJ11* rises to 0.8308 (1/6 kept, 8.995 Å mean) while tightly folded *KCNA2* (0.776 Å, 6/6 kept) loses ensemble diversity and drops to 0.4000.

a clear monomeric flow-matching failure). Restrained `OpenMM` minimization reduced mean heavy-atom clashes from 7.4 → 1.9/gene overall, though high-density crops retained residual clashes in *FSHR* (19.9 → 15.5), *F9* (30.8 → 9.0), and *FN1* (10.1 → 7.0).

2. **Pure-structural ensemble signals are modest, sensitive to decoy variance, and statistically underpowered across** 38 **genes:** Without sequence embeddings, strict ≤7.5 Å zero-shot $-\min_k \Delta\Delta G_k$ (which pools kept conformers with $\Delta\Delta G_{\mathrm{static}}$) achieves 0.5983 vs. 0.5666 static AF2 (+3.17 pp), and the 6-feature pure-structural classifier reaches 0.6059 **vs.** 0.5595 (+4.64 pp `LOHCO`, CI [+0.0044, +0.0954], one-sided $p = 0.0595$, two-sided $p = 0.1191$; +4.57 pp `LOGO`, CI [+0.0036, +0.0943], two-sided $p = 0.1241$). However, two caveats are essential. First, across the 24 bootstrap tests in Part B, no difference survives Holm–Bonferroni correction, and because the 6-feature strict model adds four ensemble columns over the 2-feature static baseline, part of its +4.64 pp lift may reflect feature-set capacity. Second, unfiltered zero-shot $-\min_k \Delta\Delta G_k$ (0.6099, including 18–37 Å unfolded decoys) slightly exceeds strict-gated (0.5983) and TM-gated (0.5774) ranking: in partially unfolded decoys, buried hydrophobic core sites become solvent-exposed so $\Delta\Delta G_k \to 0$, artificially mimicking "conformational stability relief" on destabilizing mutations.

3. **Late fusion plateaus at** ∼0.625**, and membrane regressions stem from aqueous interface exposure rather than monomer unfolding:** Fusing conformational features with Stage-1 `ESMC-6B` yields **null overall lift** (0.6164 strict [$p = 0.678$], 0.6249 TM-gated [$p = 0.560$] vs. 0.6229 static AF2 and 0.6256 Stage-1 alone; on the 11 giant proteins >1,022 aa, late fusion reaches only 0.5232–0.5318). On the 14 folded soluble/single-pass genes, individual switches show large apparent shifts (*MORC2* 0.3333 → 0.5667, *GNAO1* 0.6467 → 0.8261, *STAT1* 0.3391 → 0.4870, *PRPS1* 0.6544 → 0.7426), yet the subset mean lift is +0.94 pp with wide uncertainty (CI [−0.0449, +0.0683], two-sided $p = 0.946$). Conversely, on the 21 multi-pass transmembrane genes ($n_{\mathrm{TM}} \geq 2$), strict monomeric ensembles drop mean `LOHCO` AUROC from 0.6669 to 0.6519, led by pentameric nicotinic receptor subunit *CHRNE* (1.0000 → 0.5600, −44.00 **pp**) and tetrameric K^+ channel *KCNJ2* (0.7931 → 0.5862, −20.69 **pp**). Crucially, our structural diagnostics falsify "monomer unfolding" as the cause: *CHRNE* (2.92 Å mean Core-RMSD, 6/6 particles kept) and *KCNJ2* (6.50 Å, 3/6 kept) **remained well folded**, but scoring isolated transmembrane subunits in `ThermoMPNN`'s implicit aqueous potential exposes inter-subunit and lipid-facing surfaces to water. While gating out all $n_{\mathrm{TM}} \geq 2$ ensembles (`TM-Topology-Gated`) prevents these regressions, it is a blunt post-hoc filter that reverts 21/38 (55%) of the cohort—including well-folded monomeric 7-TM GPCRs such as *TSHR* (1.18 Å, 12/12 kept), *FSHR*, and *CASR*—to static AF2.

Table 3. Paired monomeric `[A]` versus 4-chain `[A, B, C, D]` homo-tetrameric `Boltz-2/ConforFlux` + `OpenMM` + `ThermoMPNN` pore control (`Part C`). Core-RMSD measures chain-A backbone Cα deviation against the monomeric AF2 pore crop. Pair count $N_{\text{pairs}} = N_{\text{GoF}} \times N_{\text{LoF}}$ sets the exact discrete AUROC step size $1/N_{\text{pairs}}$. †In *KCNJ2*, 3 of 4 GoF alleles lie in the cytoplasmic domain outside residues 58–209, leaving 1 pore GoF variant (AUROC undefined). ‡Monomer `[A]` pools kept conformers with static AF2 $\Delta\Delta G$, whereas 4-chain tetramer scores $\min_k \Delta\Delta G_k$ over kept tetramer conformers only (1 conformer on *KCNJ11*).

K$^+$ Channel Gene (Subunit)	Pore Crop (152 aa)	GoF / LoF (N_{pairs})	Monomer `[A]` Chain-A RMSD (Kept)	4-Chain `[A,B,C,D]` Chain-A RMSD (Kept)	Inter-Chain Cα (<10 Å/res)	Zero-Shot $-\Delta\Delta G$ AUROC‡ Static / Mono / Tetramer
KCNJ11 (Kir6.2, K_{ATP})	49–200	13 / 5 (65)	8.789 Å (1/6)	8.995 Å (1/6)	3.12	0.7385 / 0.6923 / 0.8308 (+9/65 pairs vs. mono)
KCNQ2 (Kv7.2, M-ch.)	194–345	3 / 90 (270)	6.544 Å (4/6)	**4.830** Å (6/6)	3.28	**0.8815** / 0.8778 / 0.8593 (−5/270 pairs)
KCNJ2 (Kir2.1, IRK1)	58–209	1 / 20 (20)†	9.754 Å (0/6)	**6.704** Å (5/6)	3.40	N/A (1 pore GoF)†
KCNA2 (Kv1.2, Shaker)	291–442	5 / 3 (15)	9.220 Å (2/6)	**0.776** Å (6/6)	2.37	**0.5333** / **0.5333** / 0.4000 (−2/15 pairs)
4-Channel Mean / Total	608 aa (4×152)	22 / 118	**8.577** Å (7/24, 29.2%)	**5.326** Å (18/24, 75.0%)	**3.04**	**0.7178** / 0.7011 / 0.6967 (3-ch mean)

2.4 Direct Quaternary Control: 4-Chain `[A, B, C, D]` Homo-Tetrameric K$^+$ Channel Pore Ensembles (`Part C`)

Because *KCNJ2* and *CHRNE* failed as isolated subunits in aqueous `ThermoMPNN`, we tested whether explicit homo-oligomeric pore assembly rescues zero-shot within-gene $\Delta\Delta G$ discrimination. We selected all 4 symmetric homo-tetrameric K$^+$ channels in our cohort—inward rectifiers **Kir6.2 (*KCNJ11*)** and **Kir2.1 (*KCNJ2*)**, and voltage-gated **Kv7.2 (*KCNQ2*)** and **Kv1.2 (*KCNA2*)** (excluding *ABCC8*/SUR1, an obligate Kir6.2$_4$SUR1$_4$ hetero-octamer)—and extracted each channel's 152-aa pore module (`S5-P-loop-S6` / `M1-P-loop-M2`, 140 clinical pore variants). We ran paired `Boltz-2/ConforFlux` ($K = 6$) + restrained `OpenMM` + `ThermoMPNN` ensembles on the **isolated `[A]` monomer (**152 **aa)** versus the **4-chain `[A, B, C, D]` homo-tetramer (**608 **aa)** with 100% ColabFold MSA coverage (**Figure 3B–C**; **Table 3**).

Rather than confirming a simple "quaternary rescue" of functional directionality, **Table 3** reveals an instructive **dissociation—and across individual channels, an anti-correlation—between protomer folding improvement and zero-shot $\Delta\Delta G$ AUROC**:

- **Protomer folding improves on 3 of 4 channels, but quaternary arrangement is unverified against cryo-EM:** Including all four chains lowers mean chain-A Core-RMSD (measured against the monomeric AF2 pore crop) from 8.577 Å (7/24 [29.2%] particles ≤7.5 Å) to 5.326 Å (18/24 [75.0%] kept), with 2.37–3.40 inter-chain Cα contacts (<10 Å) per residue: *KCNA2* tightens from 9.220 → 0.776 Å (2/6 → 6/6), *KCNJ2* from 9.754 → 6.704 Å (0/6 → 5/6), and *KCNQ2* from 6.544 → 4.830 Å (4/6 → 6/6), whereas *KCNJ11* gets slightly *worse* (8.789 → 8.995 Å, only 1/6 particle ≤7.5 Å). We emphasize that chain-A RMSD against a monomeric crop verifies subunit folding, not native quaternary pore geometry relative to experimental cryo-EM tetramers (e.g., Kir6.2 `6BAA`/`6C3O`, Kir2.1 `3SPI`, Kv1.2 `2A79`/`3LUT`, Kv7.2 `7CR0`); notably, *KCNA2* exhibits the lowest inter-subunit contact density (2.37/res) despite the tightest chain-A RMSD (0.776 Å).
- **Pooled 3-channel zero-shot AUROC does not beat static AF2 (**0.7178 > 0.7011 > 0.6967**):** Averaged across the three bidirectional pore modules, 4-chain tetramer scoring (0.6967) trails both isolated monomer ensembles (0.7011) and static monomeric AF2 (0.7178). On *KCNJ11* (65 GoF×LoF pairs), AUROC rises to 0.8308 vs. 0.7385 static (+6/65 pairs) and 0.6923 monomer (+9/65 pairs), yet this value rests on a **single kept tetramer conformer (**1/6**)** from an ensemble with 8.995 Å mean RMSD, and reflects a scoring asymmetry in our harness whereby monomer `[A]` pooled kept conformers with $\Delta\Delta G_{\text{static}}$ whereas the tetramer arm took $\min_k \Delta\Delta G_k$ over kept tetramer conformers alone. Conversely, on *KCNA2* (15 pore pairs), tightly converging all 6/6 particles to 0.776 Å **collapses conformational ensemble diversity to zero**, eliminating the $\min_k \Delta\Delta G_k$ variance premise and dropping pore AUROC by 2 swapped pairs (0.5333 → 0.4000; whereas full-gene sequence+static late fusion achieves 1.0000 on *KCNA2* in Part B by leveraging S4 voltage-sensor positional priors). Moreover, scoring single-subunit substitutions averaged across C4-equivalent chains inside soluble-trained `ThermoMPNN` violates `Megascale`'s monomeric aqueous training domain on both oligomerization and lipid-bilayer axes.

3 Discussion

3.1 Mandatory Controls for Variant Directionality Benchmarks

Evaluating missense mechanism predictors solely by gene-disjoint cross-validation AUROC is invalid whenever disease genes harbor skewed GoF:LoF ratios. Any global protein vector (`ESM-2` mean pooling) or linear-combination constant leak (pLDDT_i plus $\text{pLDDT}_i - \overline{\text{pLDDT}}_{\text{protein}}$) achieves 0.73–0.76 cross-gene AUROC while performing **below coin-flipping (**0.46–0.49**)** within the same patient gene. Future GoF/LoF benchmarks should require: (1) **mutation-blinded wild-type-only (`WT-Only`) controls** (noting as a limitation that while we executed this control on `ESM-2` here, future releases should also log a zero-tabular, WT-embedding-only `ESMC-6B` control to quantify residual motif-level gene identification in cross-cluster CV); (2) <30% **sequence-identity `MMseqs2` splits**; (3) **within-gene AUROC** (alongside exact pair-count CIs) and **per-fold** C-terminal (>1,022 aa) stratification; and (4) **automated Ridge linear-span checks** plus **matched pure-LoF cohorts** across all ablations.

3.2 Why Larger Sequence Models (98B–256B) and Unliganded Aqueous Ensembles Plateau Near ∼0.64

Two complementary principles explain why within-gene discrimination saturates around 0.62–0.65 across both foundation pLMs and unliganded structural ensembles:

1. **Evolutionary orthogonality (and literature-contamination hazards at 32B–256B):** Masked protein and genomic language models estimate evolutionary sequence likelihood $P(\text{seq} \mid \text{evolution})$. Because natural selection purifies against *both* hypermorphic GoF and hypomorphic LoF disease alleles, functional *sign* is nearly orthogonal to evolutionary conservation once regional domain context (e.g., autoinhibitory latch vs. hydrophobic core) is resolved by `ESMC-6B`. Indeed, adding six external zero-shot foundation predictors (`AlphaMissense`, `ESM-IF1`, `ESM-1v/2/3`, `Evo-2`) *increases* cross-cluster AUROC (0.6960 → 0.7052) while *decreasing* within-gene AUROC (0.6391 → 0.6337). Scaling further to multimodal 98B protein models or 40B genomic models re-encodes the same purifying-selection axis, while prompting 32B–256B general text LLMs on `GLOF` introduces severe **training-corpus memorization** of published ClinVar/OMIM clinical alleles unless strictly restricted to post-cutoff prospective variants.
2. **Equilibrium folding stability vs. state-differential membrane energetics:** `Boltz-2/ConforFlux` samples unconditioned equilibrium backbones around a single low-energy state, and `ThermoMPNN` predicts soluble monomeric thermal unfolding ($\Delta\Delta G_{\text{unfold}}$). Breaking the ∼0.64 ceiling requires computing **explicit state-differential free energies** ($\Delta\Delta G_{\text{open/active}} - \Delta\Delta G_{\text{closed/inactive}}$) across paired cryo-EM or multi-state templates (including hetero-oligomers such as *CHRNE* $\alpha_2\beta\delta\epsilon$ and *ABCC8* $\text{Kir6.2}_4\text{SUR1}_4$) using membrane-aware potentials, coupled with deep mutational scanning (DMS) directionality datasets to overcome the ±4–6 pp small-sample variance of clinical bidirectional cohorts.

4 Methods

4.1 Dataset Curation, Truncation Audit, and <30% Homology Splitting (`GLOF`)

We ingested `GLOF` (Maricato et al. 2026 [4]; 112,399 human missense variants: 3,128 GoF, 13,864 LoF, 95,407 neutral) and audited truncation at `AA_POSITION` > 1,022. Mapping pathogenic variants to canonical UniProt sequences yielded N = 24,700 GoF/LoF variants. For Part A balanced benchmarking, we retained all variants in any gene with ≥1 GoF allele and deterministically subsampled pure-LoF genes (GoF = 0) to ≤10 variants (`seed` = 42), yielding N = 14,037 variants. Within-gene AUROC in Part A was evaluated across all 60 genes harboring ≥2 GoF and ≥2 LoF variants. Canonical wild-type sequences were clustered via `MMseqs2` (`easy-cluster -min-seq-id 0.30 -c 0.8 -cov-mode 0` [17]). Both 5-fold gene-disjoint (`GeneCV`) and 5-fold homology-cluster-disjoint (`HomCV`) splits were generated via `GroupKFold(n_splits=5)`. Per-fold C-terminal (>1,022 aa) AUROC averaged folds containing ≥4 C-terminal variants covering both classes.

4.2 `ESM-2`, `ESMC-600M`, and May-2026 `ESMC-6B` Window Extraction

For `JENNIE` baselines, we averaged final-layer residue representations (`layers[-1]`, residues 1..1,022, excluding `<cls>`/`<eos>`) from `facebook/esm2_t33_650M_UR50D` (d = 1,280) [8] for $\bar{\mathbf{h}}_{\text{WT}}$, $[\bar{\mathbf{h}}_{\text{WT}} \,\|\, \bar{\mathbf{h}}_{\text{Mut}}]$, and $\bar{\mathbf{h}}_{\text{Mut}} - \bar{\mathbf{h}}_{\text{WT}}$. For local site embeddings on `NVIDIA H200 NVL 140GB` (`bfloat16`), we ran `biohub/ESMC-600M` (d = 1,152) over tiled

1,022-**aa wild-type windows** (`snap` = 256 aa) and **May-2026** `biohub/ESMC-6B` (d = 2,560) [13] over tiled 2,046-**aa wild-type windows** (`snap` = 512 aa), extracting target-site $\mathbf{h}_{\mathrm{WT},i}$ plus 6 site-logit/context-cosine scalars (±7-residue local context). For `MutDelta`, we ran exact 384-**aa centered wild-type and single-mutant windows** through `ESMC-6B`, extracting site deltas ($\mathbf{h}_{\mathrm{Mut},i} - \mathbf{h}_{\mathrm{WT},i}$), local pocket-window deltas, and scalar perturbation invariants.

4.3 Zero-Constant Tabular Tiers and Fold-Internal Blended Learner Protocol

Tabular features excluded L, i/L, global composition, $\overline{\mathrm{pLDDT}}_{\mathrm{protein}}$, and $\Delta\mathrm{pLDDT}_i = \mathrm{pLDDT}_i - \overline{\mathrm{pLDDT}}_{\mathrm{protein}}$; an automated unit test (`Ridge(alpha=1e-4)`) verified out-of-sample $R^2 < 0.08$ against synthetic L and $\overline{\mathrm{pLDDT}}_{\mathrm{protein}}$ across all three tiers: `seq_codon_only` (46 sequence/codon columns + 6 site scalars), `true_af2_local` (+5 local AF2 pLDDT_i / $\pm5, \pm15$ window columns), and `full_zeroshot_panel` (+8 zero-shot conservation and disorder-interaction columns). Inside every `GroupKFold` split, custom training-fold median imputation (`impute_fold_medians`) and `PCA` (96 WT components; or 64 WT + 48 site-delta + 32 pocket-delta = 144 components in raw concatenation) were fit strictly on training rows. For `MutDelta`, 48 site-delta and 32 pocket-delta `PCA` components (80-D) were compressed via a nested inner 4-fold `GroupKFold` ℓ_2 logistic probe (`LogisticRegression(C=0.05, class_weight="balanced")`) into a single log-odds column (training rows received inner-OOF log-odds; test rows received full-training-fold probe log-odds). Predictions in every fold used a fixed 0.65/0.35 probability blend of `HistGradientBoostingClassifier(max_iter=250, learning_rate=0.04, max_leaf_nodes=31, min_samples_leaf=25, l2_regularization=1.5, random_state=42)` and `make_pipeline(StandardScaler(), LogisticRegression(C=0.25, max_iter=500, class_weight="balanced"))` (`max_iter=260` in 10-fold `clu30`-grouped Stage-1 log-odds extraction).

4.4 Domain-Cropped `Boltz-2/ConforFlux`, `OpenMM` Relaxation, `ThermoMPNN`, and Late Fusion (`Part B`)

All 38 UniProt-mapped genes with $L \leq 2{,}600$ aa, ≥3 GoF, and ≥3 LoF variants (1,347 variants) were matched to AlphaFold DB coordinates [18] (version fallback `v4` → `v6` → `v3`). After trimming disordered N/C tails (pLDDT < 55), spans ≤410 aa formed a single crop (up to 402 aa) and larger proteins were split into ≤2 linker-snapped 360-aa crops (<45% overlap; 58 crops total, 58/58 [100%] `ColabFold` MMseqs2 server MSAs). Conformational ensembles (K = 6 particles/crop, 348 total) were generated via **`Boltz-2`** [14] implementing `ConforFlux`-style guided flow matching [15] (`boltz predict -use_msa_server -num_particles 6 -sigma 1.8 -alpha_s 0.02 -alpha_z 0.02 -recycling_steps 2 -sampling_steps 80 -no_kernels`), relaxed in **`OpenMM 8.x`** [19] (`amber14-all.xml`, Cα restraint k = 500 kJ mol^{-1}nm^{-2}, 120 L-BFGS iterations), filtered by structured-core (pLDDT ≥ 70) Kabsch Core-RMSD ≤ 7.5 Å, and scored via **`ThermoMPNN`** [16] (in-crop variants used domain-crop static AF2 $\Delta\Delta G_{\mathrm{static}}$; out-of-crop variants fell back to full-chain static AF2 $\Delta\Delta G_{\mathrm{static}}$; strict-gated ensemble statistics pooled kept conformers with $\Delta\Delta G_{\mathrm{static}}$).

Under `LOGO` and <30%-homology `LOHCO`, `eval_stage2_protocol` fit gene-and-class-balanced (`compute_gene_class_balanced_weights`) standardized `LogisticRegression(C=0.5, max_iter=500)` models on: (i) pure-structural matrices (2 static columns: `static_ddg`, `ca_neighbors_10A`; 6 strict columns adding `gated_ddg_min`, `gated_min_minus_static_ddg`, `gated_delta_conf_ddg`, `gated_selectivity_ratio`; or 5 TM-gated columns), (ii) 1-column Stage-1 out-of-homology log-odds, and (iii) Stage-2 late fusion (3 static columns adding `z_seq_matched_mut6b`, or 5 ensemble columns adding `*_min_minus_static_ddg` and `*_delta_conf_ddg`). Note that while held-out test clusters in Stage-2 never appear in their own Stage-1 training folds, Stage-2 training rows reuse precomputed 10-fold `clu30` OOF log-odds rather than refitting Stage-1 inside each of the 38 leave-one-cluster-out splits. Gate comparisons were evaluated across the 38-gene cohort without a separate gate-selection split; uncertainty was quantified via 10,000 paired gene-level bootstrap resamples and two-sided Wilcoxon signed-rank tests alongside Holm–Bonferroni family-wise error control across all 24 Part B bootstrap comparisons.

4.5 4-Chain `[A, B, C, D]` Homo-Tetrameric K^+ Channel Pore Control (`Part C`)

For *KCNJ11* (49–200), *KCNQ2* (194–345), *KCNJ2* (58–209), and *KCNA2* (291–442) (152 aa/chain, 140 pore variants), we ran paired monomeric (`[A]`, K = 6) and 4-chain homo-tetrameric (`[A, B, C, D]`, 608 aa, K = 6) `Boltz-2` ensembles (100% ColabFold MSAs) and multi-chain `OpenMM` relaxation. Chain-A Cα Core-RMSD was measured against the monomeric AF2 pore crop, inter-subunit contacts counted Cα atoms within 10 Å on chains `B, C, D`, and `ThermoMPNN` scored single-subunit substitutions in the 4-chain coordinate frame averaged across the four C4-equivalent protomers (taking $\min_k \Delta\Delta G_k$ over kept ≤7.5 Å tetramer conformers vs. static-pooled kept conformers in the monomer arm).

Data and Code Availability

All benchmark audit scripts, zero-leakage unit tests, `ESMC-6B` extractors, domain-cropped and 4-chain homotetrameric `Boltz-2`/`ConforFlux` + `OpenMM` + `ThermoMPNN` harnesses (`run_conforgof_v3_paper.py`), figure scripts (`generate_figures_v3.py`), figures (`figures/`), LaTeX source and PDF (`MANUSCRIPT_BIORXIV.tex`, `MANUSCRIPT_BIORXIV.pdf`), adversarial referee audit (`OPUS_ADVERSARIAL_REVIEW_V3.md`), logs (`run_v3.log`), and machine-readable outputs (`conforgof_v3_paper_results.json`) are archived under `benchmarks/soto-smoke-tests/conforgof/`.

References

[1] Backwell, L. & Marsh, J. A. Diverse molecular mechanisms underlying pathogenic protein mutations: beyond loss of function. *Annu. Rev. Genomics Hum. Genet.* **23**, 475–508 (2022).

[2] Gerasimavicius, L., Livesey, B. J. & Marsh, J. A. Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure. *Nat. Commun.* **13**, 3895 (2022).

[3] Sevim Bayrak, C. *et al.* Identification of discriminative gene-level and protein-level features associated with pathogenic gain- and loss-of-function variants. *Am. J. Hum. Genet.* **108**, 2301–2318 (2021).

[4] Maricato, J., Schlesinger, D. & de Souza Moura, L. `GLOF`: a comprehensive multi-scale resource of human gain- and loss-of-function missense variants. *bioRxiv* `10.64898/2026.06.05.729843` (2026).

[5] Pipatpolkai, T., Usher, S., Stansfeld, P. J. & Ashcroft, F. M. New insights into K_{ATP} channel gene mutations and neonatal diabetes mellitus. *Nat. Rev. Endocrinol.* **16**, 378–393 (2020).

[6] Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).

[7] Frazer, J. *et al.* Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91–95 (2021).

[8] Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model (`ESM-2`). *Science* **379**, 1123–1130 (2023).

[9] Brandes, N., Goldman, G., Wang, C. H., Ye, C. J. & Ntranos, V. Genome-wide prediction of disease variant effects with a deep protein language model (`ESM-1b`). *Nat. Genet.* **55**, 1512–1522 (2023).

[10] Lee, S. `OncoCal`: within-gene calibration eliminates gene-level prior confounding in cancer driver and mechanism variant classification. *bioRxiv* `10.64898/2026.07.16.739080` (2026).

[11] Maricato, J. *et al.* `JENNIE`: protein language model classification of human gain- and loss-of-function missense variants (`jennie-paper` companion benchmark to `GLOF`). *bioRxiv* (2026).

[12] MacLean, C. *et al.* Genetic-code mutational accessibility shapes the clinical spectrum of pathogenic missense substitutions. *bioRxiv* `10.64898/2026.07.06.736753` (2026).

[13] EvolutionaryScale & Chan Zuckerberg Biohub. `ESMC-6B`: scaling masked protein language representations (`biohub/ESMC-6B`, May 2026 release). HuggingFace Hub / Technical Report (2026).

[14] Passaro, S., Corso, G., Wohlwend, J. *et al.* `Boltz-2`: towards accurate and efficient binding affinity and biomolecular structure prediction. *bioRxiv* `10.1101/2025.06.14.659707` (2025).

[15] Suzuki, K. & Amagasa, T. `ConforFlux`: fast protein conformational ensemble generation via SE(3) flow matching. *Adv. Neural Inf. Process. Syst. (NeurIPS) / bioRxiv* `10.64898/2026.05.16.725138` (2026).

[16] Dieckhaus, H., Brocidiacono, M., Randolph, N. Z. & Kuhlman, B. Transfer learning to leverage larger datasets for improved prediction of protein stability changes (`ThermoMPNN`). *Proc. Natl. Acad. Sci. USA* **121**, e2314853121 (2024).

[17] Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).

[18] Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

[19] Eastman, P. *et al.* OpenMM 8: molecular dynamics simulation with machine learning potentials. *J. Phys. Chem. B* **128**, 109–116 (2024).