IsoPLM: Isolating the Impacts of Architecture on Protein Language Models

Yash Semlani [*1] **Ruitong Li** [*1] **Frederick Hoffman** [1] **Nauman Javed** [1] **Krithik Ramesh** [1]

Abstract

Since the release of ESM, protein language model (PLM) architectures have proliferated, but new releases simulataneously vary in data, training recipe, and scale, complicating isolation of the architectural primitives which actually drive performance gains. We pre-train a family of PLMs across three scales with data and training recipe held constant, isolating the contributions of mixture-of-experts (MoE) and hybrid attention–SSM primitives under a masked language modeling (MLM) objective. Each model is evaluated on benchmarks spanning protein structure understanding and both in-domain and out-of-domain variant effect prediction, with internal representations probed via linear probes and expert co-activation and specialization analysis. Across the full sweep, hybrid variants generalize most strongly; at smaller and intermediate scales, sparse hybrids further dominate on local structural tasks while dense hybrids dominate on global structural tasks and out-of-domain variant effect prediction. These task-specific advantages become less consistent at the largest scale, suggesting that some benefits of explicit architectural priors are absorbed by scale. Representational analysis shows that hybrids encode a complementary combination of the global priors captured by attention and SSMs, while MoEs encode a sharply localized structural prior. More broadly, this work establishes a controlled-pretraining framework for rigorously evaluating architectural primitives in biological foundation models.

1. Introduction

The release of ESM (Rives et al., 2021; Lin et al., 2023) established encoder-only transformers as the standard architecture for protein language models and demonstrated strong results across a broad range of downstream protein modeling tasks. In its wake, the development of PLM architectures has rapidly accelerated, with promising architectural components such as mixture-of-experts, state space models (SSMs), and hybrid attention–SSM layers seeing use in recent models (Sgarbossa et al., 2024; Sun et al., 2024; Yang et al., 2025).

These components are well aligned with key challenges in protein modeling. MoE layers decouple parameter count from per-token compute, allowing capacity to scale without proportional growth in compute requirements (Shazeer et al., 2017; Fedus et al., 2022). SSMs replace quadratic attention with linear-time sequence mixing, making long-context modeling more efficient (Gu & Dao, 2024; Dao & Gu, 2024). Hybrid stacks combine attention and SSM layers to retain explicit token-pair interactions while improving long-context efficiency (Lieber et al., 2024; Poli et al., 2024). Recent PLMs using these primitives report gains across protein structure understanding, fitness prediction, and protein design relative to Transformer baselines or prior PLMs (Sgarbossa et al., 2024; Sun et al., 2024; Yang et al., 2025).

However, existing works confound architectural claims with variance along three key axes. First, models are trained on different data distributions: some use single protein sequences (Rives et al., 2021; Lin et al., 2023), while others incorporate homologous context, MSAs, structural annotations, metagenomic sequences, or curated families (Elnaggar et al., 2022; Sun et al., 2024; Yang et al., 2025). Second, models differ in training recipe: encoder-only masked language modeling (Devlin et al., 2019; Rives et al., 2021; Lin et al., 2023), autoregressive sequence modeling (Yang et al., 2025), span corruption, contrastive objectives, curriculum schedules, optimizer settings, initialization schemes, masking strategies, and context construction all impose different inductive biases that can interact strongly with architecture. Third, models are evaluated at different scales, including differences in active parameters, total parameters, training tokens, context length, batch size, and compute budget. Furthermore, new models are frequently compared against open-weight baselines such as ESM (Rives et al., 2021; Lin et al., 2023) or ProtTrans (Elnaggar et al., 2022) rather than against baselines pretrained from scratch under the same data, objective, and compute budget. As a result,

[1]Lyra Labs. Correspondence to: Yash Semlani <yash@lyralabs.ai>, Ruitong Li <ruitong@lyralabs.ai>.

Accepted at the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026)

differences in benchmark performance may reflect changes in training corpus, objective, scale, or implementation details rather than the architectural primitive under study.

Isolating architectural effects therefore requires a controlled experimental design: each architecture must be pretrained under the same conditions and evaluated on a common suite of benchmarks. This is substantially more difficult and costly than comparing against existing open-weight baselines. However, without such a study it is difficult to answer the fundamental question for designing next-generation PLMs: which primitives improve protein modeling performance, and under what conditions?

In this work, we perform this rigorous comparison across a set of promising architectural primitives. We pre-train a family of PLMs incorporating attention, SSM, MoE, and hybrid attention–SSM layers, both in isolation and in combination, while fixing the confounding factors identified above. Using TAPE (Rao et al., 2019), ProteinGym (Notin et al., 2023), and FLIP 2 (Didi et al., 2026), we evaluate each architecture on protein structure understanding and both in-distribution and out-of-distribution variant-effect prediction. Finally, we investigate the mechanisms underlying model behavior using linear probes, and expert co-activation and specialization analysis.

Our results indicate that the benefits of architectural primitives depend significantly on scale and on whether the downstream task requires local residue-level or more global sequence-level understanding. At smaller and intermediate scales, sparse hybrid models perform best on local structural tasks, while dense hybrid models are stronger on global structural tasks and out-of-domain variant-effect prediction. However, at the largest scale these gains become less consistent, suggesting that some benefits of explicit architectural priors are absorbed by scale. Through representational analysis, we find that (i) hybrid layers encode a complementary combination of the global structural priors captured by attention and SSMs, and (ii) MoE layers encode a strong localized structural prior.

In summary, our contributions are that:

1. We illuminate the highly task- and scale-dependent benefits of a promising set of model architectures.

2. We show that hybrid- and MoE-based models encode biologically distinct and complementary priors.

3. We provide a framework for comparing architectural components for use in next-generation protein language models.

2. Related Work

Architectural primitives in sequence modeling. We evaluate architectures along two orthogonal axes: *token-mixing* — attention (Vaswani et al., 2017), state-space models such as Mamba (Gu & Dao, 2024), and hybrid stacks that interleave both (Lieber et al., 2024; Poli et al., 2024) — and *parameter allocation* — dense vs. Mixture-of-Experts (Shazeer et al., 2017; Fedus et al., 2022), which dispatches tokens to sparse expert subsets and decouples capacity from per-token compute. Their combinations define the design space we evaluate.

Protein language models and the confound of scale. PLMs have rapidly adopted these primitives: ESM-1/2 (Rives et al., 2021; Lin et al., 2023) established the dense-Transformer baseline, ProtMamba (Sgarbossa et al., 2024) explored SSMs, AIDO (Sun et al., 2024) scaled MoE, and Dayhoff (Yang et al., 2025) combined all three. Each generation simultaneously altered corpus, recipe, and scale. Controlled comparisons in natural language show that architectural advantages at small scales often erode or invert at larger scales (Waleffe et al., 2024); because prior PLMs conflate architecture with corpus and recipe, the isolated impact remains unclear. We close this gap with a compute-matched pretraining pipeline across multiple scales.

Evaluating protein language models. We evaluate architectural representations on two distinct axes: variant-effect prediction and structural understanding. For variant-effect prediction, zero-shot fitness scoring is benchmarked using ProteinGym (Notin et al., 2023) for in-distribution deep mutational scanning assays, and FLIP 2 (Didi et al., 2026) for out-of-distribution generalization. For structural understanding, we assess whether internal representations faithfully encode biochemical priors. Crucially, we separate local structural tasks (e.g., secondary structure) from global ones (e.g., long-range contacts and fold-level topology), as our introduction notes that different architectural priors distinctly benefit local versus global understanding.

Mechanistic interpretability for PLMs. Prior work has probed PLMs via attention analysis (Vig et al., 2021), sparse autoencoders (Simon & Zou, 2025; Adams et al., 2025), and layer-wise linear probes. MoE routing in biological contexts — how experts specialize by amino-acid substitution or structural environment — remains largely unexplored. We pair structural probes with routing analysis to characterize the distinct priors learned by hybrid and MoE architectures.

3. Methods

3.1. Architectures

We train five architecture families at three scales (20M, 150M, 650M): (1) ESM, a dense Transformer baseline; (2) HYDRA (Hwang et al., 2024), a bidirectional SSM baseline; (3) MOE, a sparse Transformer replacing each dense feedforward network with a sparse mixture-of-experts network (Shazeer et al., 2017; Fedus et al., 2022); (4) HYBRID, a dense model interleaving stacks of Hydra blocks with single dense Transformer blocks; and (5) HYBRID-MOE, a sparse hybrid interleaving stacks of Hydra blocks with single Transformer + MoE blocks.

All models use a shared ESM-2-style (Lin et al., 2023) backbone with RoPE positional embeddings, RMSNorm (Zhang & Sennrich, 2019), pre-normalized residual blocks, SwiGLU feedforward activations (Shazeer, 2020), and bias-free feedforward networks. Hybrid models use Hydra to attention ratios of 3:1, 7:1, and 6:1 across the three parameter scales, and MoE feedforward networks use 64 experts with top-k routing where $k = 8$.

3.2. Pre-training recipe

We construct our pretraining dataset using protein sequences sampled from Uniref90 (Suzek et al., 2015) using the ESM-2 sampling protocol (full dataset construction details in Appendix A.2). All models are trained under a masked language modeling objective (Devlin et al., 2019) with 15% token masking under the BERT corruption scheme for 500,000 steps using the AdamW optimizer (Loshchilov & Hutter, 2019) at a global batch size of 2M tokens with a maximum sequence length of 1024 tokens (full training details in Appendix A.1).

3.3. Evaluation suite

We evaluate on three benchmark families: (a) protein structure understanding tasks on TAPE; (b) in-distribution variant-effect prediction on ProteinGym; (c) out-of-distribution variant-effect prediction on FLIP2.

3.3.1. PROTEIN STRUCTURE UNDERSTANDING ON TAPE

We evaluate local and global structural understanding using three TAPE (Rao et al., 2019) tasks: secondary-structure prediction, contact prediction, and remote homology. Secondary structure is reported as pooled Q3 accuracy across CASP12, CB513, and TS115. Contact prediction is reported as precision-at-$L/5$ on medium- and long-range contacts, where contacts are Cα pairs within $8\,$Å. Remote homology is evaluated on SCOP fold classification (Murzin et al., 1995) under the TAPE split, where test sequences are held out at the superfamily level.

We additionally use two layerwise probes to localize structural information inside the pretrained backbones. For remote homology, we freeze each model and train a linear classifier on sequence-pooled representations from each layer to predict the SCOP fold label; probe accuracy as a function of depth measures where fold-level information becomes linearly recoverable. For contacts, we compute a categorical-Jacobian map (Zhang et al., 2024) at each layer by measuring how substitutions at residue j change the representation or logits at residue i, symmetrize the resulting pairwise scores, and evaluate them against native long-range contacts using precision-at-$L/5$. In hybrid models, we report these contact scores separately for SSM and attention layers to test which token-mixing blocks carry pairwise structural signal.

3.3.2. IN-DISTRIBUTION VARIANT-EFFECT PREDICTION ON PROTEINGYM

Zero-shot scoring. We score each of $\sim$2.5M variants in the 217 ProteinGym (Notin et al., 2023) substitution tasks via masked-marginals zero-shot scoring (Meier et al., 2021). For a variant with mutated positions $\mathcal{P}$,

$$\mathrm{LLR}(\mathrm{mut}) = \sum_{p\in\mathcal{P}} \left[\log P\left(\mathrm{mut}_p \mid \mathbf{x}^{\mathrm{wt}}_{\setminus p}\right) - \log P\left(\mathrm{wt}_p \mid \mathbf{x}^{\mathrm{wt}}_{\setminus p}\right)\right], \tag{1}$$

where $\mathbf{x}^{\mathrm{wt}}_{\setminus p}$ is the wildtype sequence with position p masked. We patched our evaluator to also save per-variant records (`mutant`, `dms_score`, `llr`, `n_mutations`, `mut_positions`), so that all stratified analyses below run from a single forward pass. Throughout, residues are indexed by sequence position $i \in \{1, \ldots, L\}$; the sequence separation between i and j is $|i - j|$ and their Cα distance is d_{ij}.

MoE routing-organization analysis. To test whether sparse experts encode residue-level biochemical priors, we track MoE routing on ProteinGym single mutants. For each variant, we substitute the mutant residue into the wildtype sequence and record the top-1 expert at the mutated position in every MoE layer. We group substitutions by source amino-acid class, destination amino-acid class, and burial status, using three residue classes: hydrophobic, polar, and charged. This yields 18 transition buckets.

For each transition bucket t, we identify expert cells whose routing frequency is enriched relative to their global usage, measured as $\log_2[p(\ell, e \mid t)/p(\ell, e)]$ for expert e in layer ℓ. We then compare the overlap between specialist expert groups for transitions sharing the same source class, destination class, or surface/buried status. If experts are organized by a given axis, transitions sharing that axis should reuse the same expert cells more often than unrelated transitions. Full enrichment thresholds and the Jaccard-overlap definition are

given in Appendix A.6.

Helix-cap structural-context probe. The five secondary-structure contexts (`helix_Ncap`, `helix_internal`, `helix_Ccap`, `sheet`, `loop`) crossed with surface/buried define ten *structural-context cells*. We focus on single-mutant substitutions to three residues with a previously-characterized, position-dependent tolerance: *Proline*, tolerated at the helix N-cap but helix-breaking elsewhere (Aurora & Rose, 1998); *Glycine*, favored at the helix C-cap (Richardson & Richardson, 1988); and *Tryptophan*, tolerated in buried hydrophobic contexts. Proline's N-cap tolerance is a single biophysically pre-specified hypothesis; Glycine and Tryptophan test whether any architecture-level effect generalizes. Within each cell, per architecture, we report the across-task mean and 95% bootstrap CI of the per-task Spearman ρ between model LLR and DMS, and a position-normalized tolerance (focal residue minus the median of the other substitutions at the same position) separating residue-specific effects from position difficulty. Full details are given in Appendix A.5.

3.3.3. Out-of-distribution variant-effect prediction on FLIP2

We evaluate supervised out-of-distribution generalization on FLIP 2 (Didi et al., 2026) across the Amylase, IRED, NuCB, TrpB, Hydrolase, Rhomax, and PDZ3 tasks. For each task split, the pretrained backbone is fine-tuned with the same regression head and training recipe across architectures. We report task-level Spearman correlation and aggregate performance as the mean Spearman and mean NDCG across all evaluated splits.

In preliminary experiments we performed supervised fine-tuning on the ProteinGym substitution tasks and found all architectures achieved similar performance on unseen test data at comparable scales. However, FLIP 2 splits are designed to shift mutation count, position, wildtype background, or sequence distance between train and test, allowing us to measure whether models can generalize effectively to data that meaningfully differs from the local training distribution.

4. Protein structure understanding

Across TAPE, hybrids are the most consistent architecture family, with strong performance on both secondary structure and remote homology at small and medium scales. However, contact prediction increasingly favors dense attention at larger scales, suggesting that explicit pairwise modeling remains important for long-range geometry. Adding MoE capacity improves hybrid-MoE models on local secondary-structure tasks but does not reliably improve global tasks. These trends suggest that hybrids encode a strong sequence-level structural prior, while MoEs contribute a more local structural prior.

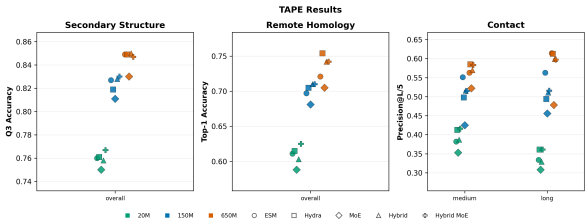


Figure 1. **TAPE Results** We report Q3 accuracy, Top-1 accuracy, and Precision-at-L/5 for Secondary Structure, Remote Homology, and Contact Prediction respectively

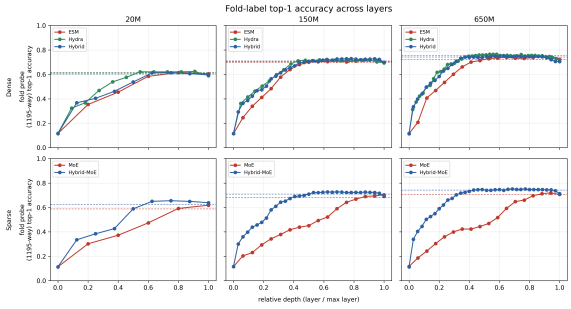


Figure 2. **Remote-homology fold prediction accuracy of linear probes across model depth.** Top-1 accuracy is shown for 1195-way fold-label classification across all scales

Layerwise remote-homology probes show that SSM and hybrid representations make fold-level information linearly recoverable earlier in depth, while MoE models are consistently weaker. Hybrid-MoE remains competitive, suggesting complementarity: SSM-heavy backbones preserve global fold information, whereas sparse experts add capacity that is less effective for global structure in isolation.

Using contact maps derived from intermediate representations, we evaluate how much long-range pairwise residue-interaction information is encoded at each layer. Surprisingly, attention layers inside hybrid models achieve the highest mean precision-at-$L/5$, while attention layers inside transformer-only models consistently achieve the lowest mean precision-at-$L/5$. Together with the lower variance across hybrid attention layers, this suggests that transformer models distribute pairwise interaction modeling across many attention layers, whereas hybrid models concentrate this role in their smaller number of attention layers. In other words, attention appears to specialize more sharply when embedded within an SSM-dominated backbone.

We next examine whether sparse experts specialize for local or global structural information. Across both MoE and Hybrid, we find strong expert specialization aligned with secondary-structure labels. In contrast, specialization is much weaker for remote-homology labels. Experts show

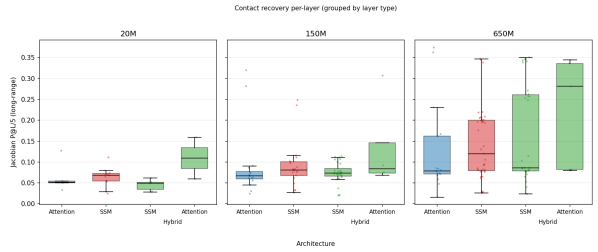


Figure 3. **Categorical Jacobian contact recovery by layer type.** For each model scale, layers from each dense model are grouped by their parent architecture and layer type, with boxes showing the distribution of long-range contact recovery scores (Jacobian P@L/5). Points denote individual layer scores. Hybrid models are split into their SSM and attention layers.

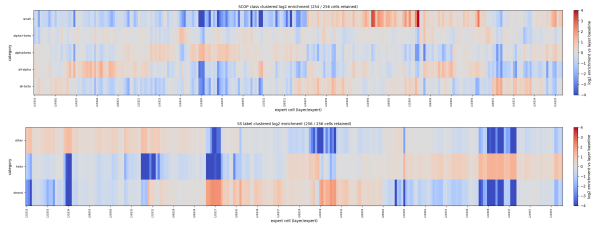


Figure 4. **Hybrid MoE 150M expert enrichment on global and local structural labels.** Log_2 enrichment of experts after TAPE fine-tuning, shown for remote-homology SCOP classes (top) and secondary-structure labels (bottom). Columns correspond to MoE expert cells indexed by layer and expert, and rows correspond to structural categories.

some separation between all-α and all-β classes, but substantially weaker specialization for mixed $\alpha + \beta$ and α/β folds, whose classification depends on the global arrangement of secondary-structure elements. The strongest remote-homology specialization occurs for the small-protein class, suggesting that sparse models can identify rare or distinctive structural regimes, but do not generally organize experts around fold-level topology. This supports the view that MoE layers primarily allocate capacity to local structural motifs rather than global structural organization.

Together, these analyses clarify the structural priors induced by each architecture. SSM-heavy layers expose fold-level information early in depth, while the few attention layers in hybrids concentrate long-range pairwise contact signal. This suggests that hybrids combine sequence-level compression with explicit pairwise interaction modeling, explaining their strong performance. In contrast, MoE layers primarily specialize for local structure: sparse models show strong expert specialization for secondary-structure labels but much weaker specialization for remote-homology classes. This helps explain why sparsity improves residue-level tasks more reliably than global structural tasks.

5. In-distribution variant-effect prediction

5.1. Aggregate scores cluster tightly; the in-house family is competitive with sequence-only public models, and ESM3-1.4B underperforms on viral tasks

All 650M architectures fall within ± 0.01 in mean per-task Spearman ρ (range 0.423–0.433) on 217 ProteinGym substitution tasks; meaningful differences emerge only after structural stratification. Across five public open-weight sequence-only baselines (Appendix A.7), none exceeds our ESM-650M in aggregate NDCG@10; ESM3-1.4B trails by -0.008 despite $\sim 2\times$ the parameters, and only retrieval-augmented E1-300M-RETRIEVAL beats us meaningfully ($+0.027$, attributable to MSA conditioning). The ESM3-1.4B deficit is concentrated on the 31 viral tasks (ΔNDCG@10 $= -0.073$, head-to-head $5/31$) — a viral cluster that appears in every sequence-only panel and disappears under retrieval augmentation (Fig. S1).

5.2. MoE encodes a local per-residue biochemistry prior, with a narrow proline-specific local-geometry signature and a structurally-coupled pair signature

We propose a potential mechanism that links MoE routing to MoE prediction behavior: at every scale, MoE-bearing models develop sharply localized, per-residue specialists organized by the *destination* amino-acid group of a substitution. This local prior leaves three signatures we test in turn — a routing signature, a (proline-specific, biophysically-bounded) local-geometry signature, and a structurally-coupled pair signature elaborated in the appendix.

Routing: specialists are sharply localized along the destination amino-acid axis, replicated across models. Mutant-forward routing at the mutated site is highly informative: in MOE-20M, the top-1 expert flips on $18/18$ transition buckets at layer 0; the strongest 650M specialists exceed $+4$ $\log_2$ enrichment ($\sim 16\times$ uniform; Fig. S3). Specialist-group sizes peak on the most structurally disruptive substitution (CHARGED$\rightarrow$HYDROPHOBIC@BURIED) across all five MoE-bearing models in our panel (range 20–49 cells). Per-axis Jaccard overlap (Sec. 3.3.2) shows experts cluster by *destination* amino-acid group rather than source: at fixed 650M, $\bar{J}_{\text{dst}}/\bar{J}_{\text{src}} = 0.190/0.064 = 3.0\times$ for HYBRID-MOE-650M; the same direction holds for MOE-20M ($2.3\times$), HYBRID-MOE-150M ($2.2\times$), and the cross-team DAYHOFF-170M ($2.5\times$). Capacity-constrained HYBRID-MOE-20M (2 MoE layers) inverts to $0.45\times$ (Fig. 5), establishing destination-axis organization as a non-trivial choice rather than an artifact of any MoE. The cross-team recovery on DAYHOFF-170M (different architecture, MoE schedule, and pretraining corpus) is the strongest robustness check available.

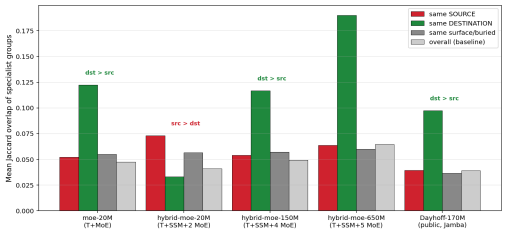


Figure 5. Mean Jaccard overlap between MoE specialist groups for transition pairs sharing each axis (source AA / destination AA / surface-buried) across five MoE-bearing models. All four MoE-rich models show $\bar{J}_{\mathrm{dst}} > \bar{J}_{\mathrm{src}}$; the capacity-constrained HYBRID-MOE-20M (2 MoE layers) inverts.

Local geometry: a proline-specific helix-position signal. A local per-residue prior should help where fitness depends on local geometry, and biophysics pre-specifies the cleanest test. Proline is the only residue lacking a backbone amide hydrogen, so it breaks the $i, i{-}4$ helix hydrogen bond everywhere except the N-terminal turn, whose donors have no intrahelical partner (Aurora & Rose, 1998). Accordingly *→PRO is tolerated only at the helix N-cap: median within-task $z = +0.41$ there vs. -0.46 to -0.58 at helix-internal, helix C-cap and β-sheet — more tolerated than at every other context (rank-sum $p \leq 3 \times 10^{-5}$; Fig. S5). At this one biophysically pre-specified cell, HYBRID-MOE-650M reaches $\rho = 0.390$ ($n = 21$, 95% CI $[0.264, 0.520]$), $+0.10$ above dense ESM-650M (0.294). We then asked whether this generalizes to other residues with a previously-characterized position-dependent tolerance — Glycine, favored at the helix C-cap (Richardson & Richardson, 1988), and Tryptophan in buried contexts. Neither shows a comparable signal: after normalizing each position by its own substitution baseline, Glycine and Tryptophan track that baseline within $\approx 0.1\, z$ across all contexts (Fig. S5), i.e. the position-gated fitness gradient Proline exhibits is absent. With no such signal to learn, no architecture-specific advantage is expected there, and none is observed. The Proline N-cap effect is therefore a specific, biophysically-grounded signature; establishing a more general local-geometry advantage will require validation on other datasets.

Structurally-coupled pairs: P3 surfaces the predicted advantage. A learned per-residue prior should help most when fitness depends on a residue's structural neighborhood — once at a 3D hub, twice at a long-contact pair. **P3** (adding MoE on top of T+SSM) confirms this: three of 19 stratified subsets reach paired Wilcoxon $p < 0.05$, all positive and all on structurally-coupled strata, with the largest effect at long-contact pairs × medium-$|\varepsilon|$ ($\Delta\rho = +0.10$, $n = 17$). Full P1/P2/P3/P4 attribution, the selectivity within long-contact pairs, and the 19-stratum grid are reported in Appendix A.4 (Fig. S6).

6. Out-of-distribution variant-effect prediction

Across 16 splits in seven protein families (Fig. 6; Tab. **??**), no architecture leads universally and parameter scaling yields no monotonic OOD return; family-level winners instead track architectural inductive bias.

Architecture–task alignment. **MoE** variants are most competitive on NUCB (small single-domain nuclease; fitness dominated by local packing and active-site chemistry — the regime our routing prior is most directly applicable to and where Sec. 5.2 finds the sharpest expert specialization). **Hybrid** (T+SSM) leads on AMYLASE and HYDROLASE, both regimes rewarding longer-range capacity: AMYLASE's extended substrate-binding cleft couples sequence-distant residues, and HYDROLASE contains the cross-protein splits requiring transfer across distantly homologous proteins (SSM contributes linear-time long-context propagation, attention contributes explicit pairwise comparison). **Hybrid-MoE** reaches the top RHOMAX score ($\rho = 0.606$, 650M) — the only membrane-protein family, where transmembrane and buried positions dominate fitness and physicochemical-class-crossing substitutions carry large costs, the same regime where our 650M routing analysis finds the sharpest specialization. RHOMAX is one split with wide within-family spread ($\rho \in [0.34, 0.60]$ across Hybrid-MoE 650M), so we treat this result as suggestive.

Scale alone does not improve OOD. 20M→650M is non-monotonic per architecture: NUCB medians *decrease* (dense Hybrid $0.525 \rightarrow 0.367$); HYDROLASE and RHOMAX are median-invariant with larger 650M variance; only AMYLASE trends weakly upward. This parallels the in-house-vs-public contrast on ProteinGym (Sec. 5.1), where ESM3-1.4B trails ESM-650M on the long viral subset — different sources of mismatch (architecture vs. pretraining mix) but the same direction: architectural prior outweighs parameter count at these scales.

Cross-protein transfer collapses uniformly. The three HYDROLASE cross-protein splits (→P06241, →P01053, →P0A9X9) yield median Spearman ≈ 0 across all five architectures and three scales, with negative tails to ~ -0.20 — a clean negative result that no sequence-only prior we evaluate generalizes across distantly homologous proteins, motivating follow-up integrating explicit structural information into the backbone.

Discussion

Our results show that architectural primitives in PLMs induce distinct structural priors whose benefits depend strongly on task type and scale. Hybrid attention–SSM models are the most consistent family at small and inter-

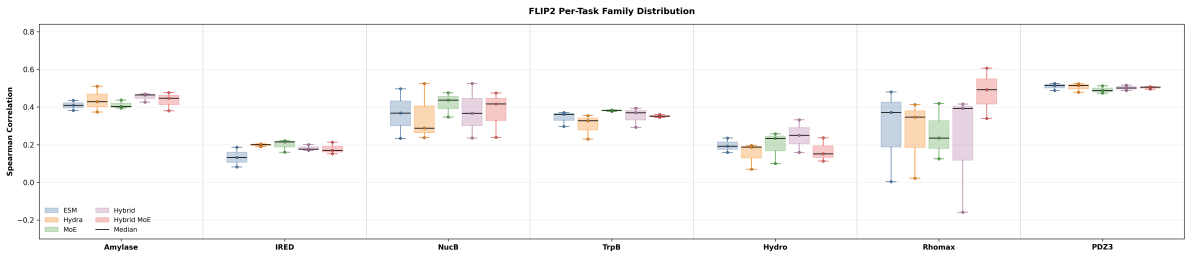


Figure 6. **FLIP2 supervised results.** Results reported across architectural variants and aggregated into box plots across parameter scales.

mediate scales, but their advantages diminish at 650M parameters, where dense attention and SSM baselines close the gap on several global structural tasks. This pattern — explicit priors helping most when capacity is limited, and baselines catching up as capacity grows — mirrors observations in natural language modeling (Waleffe et al., 2024) and suggests that architectural priors are best understood as compute-efficient inductive biases rather than upper bounds on representational quality.

Representational analyses clarify the division of labor inside hybrid models. Layerwise probing shows that SSM-heavy backbones make fold-level information linearly recoverable earlier in depth, while categorical-Jacobian analysis shows that the few attention layers in hybrids concentrate long-range pairwise contact signal — achieving higher mean precision-at-$L/5$ than attention layers in transformer-only models despite being far fewer in number. Attention thus appears to specialize more sharply when scarce: when SSM layers handle global topological compression, the remaining attention layers are freed to concentrate on explicit pairwise interactions. Hybrids combine global fold-level structure with localized pairwise modeling, and this complementarity underlies their consistent benchmark performance.

This division of labor also helps explain the FLIP2 results, where dense hybrids generalize most strongly out of distribution. FLIP2 splits shift mutation count, position, wildtype background, or sequence distance between train and test, so transferable performance requires distribution-invariant structural features rather than recurring training-distribution motifs. Global fold priors combined with concentrated pairwise modeling are a plausibly more transferable signal than the localized capacity that sparse experts allocate to recurring residue-level patterns.

MoE layers contribute a qualitatively different prior. Sparse experts specialize strongly for local secondary-structure labels but much more weakly for remote-homology classes, with the weakest specialization on mixed $\alpha + \beta$ and α/β folds whose classification depends on the global arrangement of secondary-structure elements. This asymmetry matches the downstream pattern: MoE improves residue-level tasks more reliably than global structural tasks and generalizes less well on FLIP2. Sparse routing should therefore not be viewed as a universal expressivity gain; in our setting, it primarily allocates capacity to recurring local structural motifs, with limited transfer to global or out-of-distribution regimes.

Because data, objective, and scale all affect PLM performance, architectural comparisons are difficult to interpret when these factors vary simultaneously. Our controlled setup does not exhaustively explain the performance differences among prior PLMs; rather, it isolates the effect of architectural primitives under a fixed corpus, objective, tokenizer, and training recipe. The resulting picture is more nuanced than a single architecture ranking: the same primitive can help one structural regime while hurting another, and aggregate benchmark scores can obscure these opposite-sign effects. For practitioners designing next-generation PLMs, this argues for hybrid attention–SSM stacks as the strongest general-purpose prior at compute-limited scales, sparse experts as targeted capacity for local residue-level features, and stratified rather than aggregate benchmarking as the appropriate unit of architectural comparison.

Limitations

Scale. We evaluate three scales up to 650M parameters, where architectural advantages already attenuate. Whether this attenuation continues, plateaus, or inverts at frontier PLM scales (3B–10B+) is an open question. Controlled comparisons in natural language find that architectural advantages can not only erode but reverse at larger scales (Waleffe et al., 2024), and our 650M results are consistent with the early phase of such a transition rather than a stable equilibrium. Conclusions about which primitive is "best" should therefore be read as scale-conditional.

Pretraining objective and corpus. All models are trained with BERT-style masked language modeling on UniRef90 single sequences. Autoregressive objectives, MSA-conditioned objectives, or structure-aware objectives could shift the relative value of attention, SSM, and MoE primitives — for instance, MoE routing under a left-to-right objective may allocate capacity differently than under

bidirectional MLM, potentially altering our conclusion that sparse experts encode a localized structural prior. Our representational claims are thus objective-bound, and extending the controlled-pretraining framework across objectives is a natural next step.

Architectural search space. To keep the comparison tractable we fix hybrid layer ratios, MoE expert count (64) and routing top-k (k=8), normalization scheme, and SSM parameterization rather than optimizing each family independently. MoE behavior in particular is known to be sensitive to expert count, top-k, and load-balancing loss weight, so our characterization of MoEs as encoding a localized structural prior may not generalize to substantially different sparse configurations.

Mechanistic analyses. Our representational claims rely on linear probes and routing statistics, which identify correlational rather than causal structure in representations. Linear probes additionally have a known ceiling: they detect only linearly recoverable information, so weaker probe accuracy in MoE models could reflect either less fold-level information or fold-level information encoded nonlinearly. Causal interventions (e.g., layer-wise ablations, expert knockouts during downstream evaluation) would strengthen the link between representational structure and downstream performance.

Evaluation suite. TAPE, ProteinGym, and FLIP2 cover complementary axes of structural understanding and variant-effect prediction, but TAPE in particular reflects pre-AlphaFold-era splits and contact definitions. Newer structural benchmarks derived from CASP14/15 or AlphaFold-predicted structures might shift the relative ranking of architectures on global structural tasks. Our claims about "global structural" performance should be read as specific to the benchmarks evaluated.

Single Seed Training. Each architecture scale combination is trained once with a single seed due to compute constraints. However, because differences between models are often small and MoE models are known to exhibit training instability, single-seed results make it more difficult to cleanly separate architectural signal from run-to-run noise. This is especially relevant for small performance differences between models which should be interpreted with awareness of potential conflation with training and evaluation variance between seeds.

Broader Impact

PLMs are increasingly deployed in protein engineering, enzyme optimization, and therapeutic discovery, where representational quality shapes downstream design. Our results bear on this deployment context in two ways. Because architectural priors disproportionately benefit some structural regimes over others, aggregate benchmark scores can mask regime-specific failures that matter for real applications. For example, a model competitive on average may underperform sharply on the structural regime relevant to a given desi=gn task (e.g., long-range allosteric coupling, novel folds, or out-of-family variants). This argues for distribution-aware evaluation prior to downstream deployment.

References

Adams, E., Bai, L., Lee, M., Yu, Y., and AlQuraishi, M. From mechanistic interpretability to mechanistic biology: Training, evaluating, and interpreting sparse autoencoders on protein language models. *bioRxiv*, 2025. doi: 10.1101/2025.02.06.636901.

Aurora, R. and Rose, G. D. Helix capping. *Protein Science*, 7(1):21–38, 1998.

Dao, T. and Gu, A. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning (ICML)*, 2024.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

Didi, K., Alamdari, S., Lu, A. X., Wittmann, B., Johnston, K. E., Amini, A. P., Madani, A., Czeneszew, M., Dallago, C., and Yang, K. K. Flip2: Expanding protein fitness landscape benchmarks for real-world machine learning applications. *bioRxiv*, 2026. doi: 10.64898/2026.02.23.707496. URL https://www.biorxiv.org/content/early/2026/02/26/2026.02.23.707496.

Elnaggar, A., Heinzinger, M., Dallago, C., Rihawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., and Rost, B. ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2022. doi: 10.1109/TPAMI.2021.3095381.

Fedus, W., Zoph, B., and Shazeer, N. Switch transformer: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Gu, A. and Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *Conference on Language Modeling (COLM)*, 2024. arXiv:2312.00752.

Hwang, S., Lahoti, A., Dao, T., and Gu, A. Hydra: Bidirectional state space models through generalized matrix mixers. In *NeurIPS*, 2024.

Kabsch, W. and Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983. doi: 10.1002/bip.360221211.

Lieber, O., Lenz, B., Bata, H., Cohen, G., Osin, J., Dalmedigos, I., Safahi, E., Meirom, S., Belinkov, Y., Shalev-Shwartz, S., Abend, O., Alon, R., Asida, T., Bergman, A., Glozman, R., Gokhman, M., Manevich, A., Ratner, N., Rozen, N., Shwartz, E., Zusman, M., and Shoham, Y. Jamba: A hybrid transformer-Mamba language model. *arXiv preprint arXiv:2403.19887*, 2024.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *ICLR*, 2019.

Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., and Rives, A. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021.

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4):536–540, 1995.

Notin, P., Kollasch, A. W., Ritter, D., van Niekerk, L., Paul, S., Spinner, H., Rollins, N., Shaw, A., Orenbuch, R., Weitzman, R., Frazer, J., Dias, M., Franceschi, D., Gal, Y., and Marks, D. S. ProteinGym: Large-scale benchmarks for protein fitness prediction and design. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

Poli, M., Thomas, A. W., Nguyen, E., Ponnusamy, P., Deiseroth, B., Kersting, K., Suzuki, T., Hie, B., Ermon, S., Ré, C., Zhang, C., and Massaroli, S. Mechanistic design and scaling of hybrid architectures. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 40908–40950, 2024.

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., Abbeel, P., and Song, Y. S. Evaluating protein transfer learning with TAPE. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

Richardson, J. S. and Richardson, D. C. Amino acid preferences for specific locations at the ends of alpha helices. *Science*, 240(4859):1648–1652, 1988. doi: 10.1126/science.3381086.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., and Fergus, R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15): e2016239118, 2021. doi: 10.1073/pnas.2016239118.

Sgarbossa, D., Malbranke, C., and Bitbol, A.-F. Protmamba: a homology-aware but alignment-free protein state space model. *bioRxiv*, 2024. doi: 10.1101/2024.05.24.595730. Preprint.

Shazeer, N. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q. V., Hinton, G. E., and Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.

Simon, E. and Zou, J. InterPLM: discovering interpretable features in protein language models via sparse autoencoders. *Nature Methods*, 22(10):2107–2117, 2025. doi: 10.1038/s41592-025-02836-7.

Sun, N., Zou, S., Tan, C., Sun, Y., Wang, P., Hou, L., Wu, L., Liu, P., Zhang, W., et al. Mixture of experts enable efficient and effective protein understanding and design. *bioRxiv*, 2024. doi: 10.1101/2024.11.29.625425. AIDO.Protein. Preprint.

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., and The UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 2015. doi: 10.1093/bioinformatics/btu739.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.

Vig, J., Madani, A., Varshney, L. R., Xiong, C., Socher, R., and Rajani, N. F. BERTology meets biology: Interpreting attention in protein language models. *International Conference on Learning Representations (ICLR)*, 2021.

Waleffe, R., Byeon, W., Riach, D., Norick, B., Korthikanti, V., Dao, T., Gu, A., Hatamizadeh, A., Singh,

S., Narayanan, D., Kulshreshtha, G., Singh, V., Casper, J., Kautz, J., Shoeybi, M., and Catanzaro, B. An empirical study of Mamba-based language models. In *arXiv preprint arXiv:2406.07887*, 2024.

Yang, K. K., Alamdari, S., Lee, A. J., Kaymak-Loveless, K., Char, S., Brixi, G., Domingo-Enrich, C., Wang, C., Lyu, S., Fusi, N., et al. The dayhoff atlas: scaling sequence diversity for improved protein generation. *bioRxiv*, pp. 2025–07, 2025.

Zhang, B. and Sennrich, R. Root mean square layer normalization. In *NeurIPS*, 2019.

Zhang, Z. et al. Protein language models learn evolutionary statistics of interacting sequence motifs. *Proceedings of the National Academy of Sciences*, 2024. doi: 10.1073/pnas.2406285121.

A. Supplementary Methods

A.1. Training recipe

Table 1 gives the pretraining hyperparameters not specified in the main text. We use fused AdamW with $\beta = (0.9, 0.98)$, $\epsilon = 10^{-8}$, weight decay 0.01, and gradient clipping at norm 1.0. The learning rate warms linearly to 4×10^{-4} over 2k optimizer steps, then decays linearly to $0.1\times$ peak over the first 90% of training. We validate and checkpoint every 10k optimizer steps. The nominal token budget is $500{,}000 \times 2048 \times 1024 \approx 1.05$T token slots; the realized number of non-padding tokens is lower because proteins are variable length.

Scale	Steps	Global batch (sequences)	Grad. accum.	Max tokens / step
20M	500k	2048	1	2.1M
150M	500k	2048	1	2.1M
650M	500k	2048	1–4	2.1M

Table 1. Pretraining recipe by scale. Gradient accumulation is increased for some 650M runs to keep the effective global batch fixed.

Compute resources. All pretraining runs were performed on NVIDIA H100 GPUs with 80GB memory. Each reported run used between 32 and 64 H100 GPUs.

A.2. Dataset preprocessing

We pretrain on UniRef90 single-sequence shards prepared following the ESM-2 cluster-based sampling protocol: UniRef90 clustering provides the deduplication unit, and training samples are drawn from cluster-representative sequence pools rather than from a raw redundant sequence dump. Held-out validation shards are constructed before training and are never used for optimization.

Sequences are stored as parquet rows with a `sequence` field. During training, each row is tokenized with the shared amino-acid tokenizer. Sequences that fit in the 1024-token context are encoded as `<cls>` + residues + `<eos>`; longer sequences are randomly cropped on each sample to fit the context window. Training shards are streamed and buffer-shuffled, while validation is evaluated as a finite held-out split.

A.3. ProteinGym structural stratification and statistical testing

Structural annotations. For each ProteinGym target, we annotate residues using its AlphaFold2 reference structure. Let $i, j \in \{1, \ldots, L\}$ index residue positions, let d_{ij} denote the Cα–Cα distance between residues i and j, and let $|i - j|$ denote their sequence separation. Per residue, we compute:

- **Burial.** Relative solvent-accessible surface area (rSASA) from DSSP(Kabsch & Sander, 1983). Residues with $\mathrm{rSASA} \geq 0.2$ are *surface*; the rest are *buried*.
- **Contact degree.** Number of partners with $d_{ij} < 8$ Å and $|i - j| \geq 2$.
- **Long-range contact count.** Number of partners with $d_{ij} < 8$ Å and $|i - j| \geq 24$.
- **Secondary-structure context.** DSSP labels reduced to five categories. Helices are maximal contiguous runs of DSSP H/G/I residues; the first and last residue of each run are `helix_Ncap` and `helix_Ccap`, intervening residues are `helix_internal`, DSSP E/B residues are `sheet`, and all other residues are `loop`.

Single-mutant strata. Single-mutant variants are stratified by the burial, contact-degree, long-range-contact, and secondary-structure context of the mutated residue. Contact-degree and long-range-contact strata are computed within each target so that each task contributes variants across comparable low-, medium-, and high-contact regimes where possible.

Double-mutant pair classes. A double mutant at positions (i, j) is assigned to one of four pair classes by sequence separation $|i - j|$ and structural proximity d_{ij}:

$$\begin{aligned} \textbf{local}: &\quad |i-j| < 8, \\ \textbf{medium}: &\quad 8 \le |i-j| < 24, \\ \textbf{long_noncontact}: &\quad |i-j| \ge 24,\ d_{ij} \ge 8\,\text{Å}, \\ \textbf{long_contact}: &\quad |i-j| \ge 24,\ d_{ij} < 8\,\text{Å}. \end{aligned}$$

long_contact is the canonical long-range structural-pair regime: the two mutated residues are sequence-distant but spatially close.

Experimental epistasis. When both single-mutant measurements are present in the same task, the experimental epistasis of a double mutant at (i, j) is

$$\varepsilon_{ij} = y_{ij} - y_i - y_j, \tag{2}$$

where y_{ij} is the measured double-mutant score and y_i, y_j are the corresponding single-mutant scores. Double mutants are stratified into low-, medium-, and high-epistasis groups by within-task terciles of $|\varepsilon_{ij}|$ — the magnitude, because the analysis asks whether the architecture captures non-additive coupling *strength* regardless of sign.

Distance-to-function annotation. Each residue is additionally joined to its nearest UniProt-annotated active or binding site. This annotation feeds only the distance-to-function probe and is not part of the main 19-stratum architecture-ablation grid.

Per-task stratified performance. Within each task t and variant subset s, model M's score is the Spearman correlation between its zero-shot LLR and the experimental DMS measurement,

$$\rho_{t,s}^{M} \;=\; \mathrm{Spearman}\big(\mathrm{LLR}_{t,s}^{M},\ y_{t,s}\big)\,. \tag{3}$$

A task–subset pair is included only when it contains at least five variants.

Pairwise architecture differences. For each architecture pair (A, B) and subset s, the per-task paired difference is $\Delta\rho_{t,s} = \rho_{t,s}^{A} - \rho_{t,s}^{B}$, and we report its mean over eligible tasks,

$$\overline{\Delta\rho}_s \;=\; \frac{1}{|\mathcal{T}_s|} \sum_{t \in \mathcal{T}_s} \Delta\rho_{t,s}, \tag{4}$$

where $\mathcal{T}_s$ is the set of tasks with ≥ 5 variants in subset s.

Uncertainty and significance. For each (subset, pair) we report a 95% confidence interval from a 2,000-iteration paired bootstrap that resamples tasks with replacement from $\mathcal{T}_s$ and recomputes $\overline{\Delta\rho}_s$, plus a paired Wilcoxon signed-rank p-value over the task-level differences. These are descriptive: we use them to identify subsets where the direction of an architectural effect is consistent across tasks, and do not apply a global multiple-testing correction across the 19 stratified subsets.

650M architecture-ablation pairs. The fixed-scale ablation compares four 650M models trained under the same corpus, tokenizer, and recipe. Three minimal-edit pairs isolate each component, and a fourth pair compares the full stack against the dense baseline:

P1 :	HYBRID-650M − ESM-650M	(adds bidirectional SSM to a dense Transformer),
P2 :	HYBRID-650M − HYDRA-650M	(adds attention to an SSM-only backbone),
P3 :	HYBRID-MOE-650M − HYBRID-650M	(adds MoE feedforward layers to the hybrid),
P4 :	HYBRID-MOE-650M − ESM-650M	(full hybrid-MoE stack vs. dense baseline).

P1, P2, P3 are the primary attribution analysis; P4 is a compositional check, since P1 and P3 can have opposite signs on the same subsets and the direct full-stack vs. dense contrast then masks each component's contribution.

A.4. ProteinGym 650M architecture-ablation: full results

P3 (adding MoE on top of T+SSM). A learned per-residue prior should help most when fitness depends on a residue's structural neighborhood — once at a 3D-hub single residue, twice at a long-contact double-mutant pair. P3 surfaces this in the predicted regime: three of 19 stratified subsets reach paired Wilcoxon $p < 0.05$, all positive and all on structurally-coupled strata — many-LR singles ($\Delta\rho = +0.012$, $p = 0.009$); long-contact pair $\times$ medium-$|\varepsilon|$ ($\Delta\rho = +0.10$, $p = 0.015$, $n = 17$, the largest cell anywhere in the design); and long-noncontact pair $\times$ medium-$|\varepsilon|$ ($\Delta\rho = +0.066$, $p = 0.040$). Selectivity within long-contact pairs matters: low-$|\varepsilon|$ pairs (additivity adequate) and high-$|\varepsilon|$ pairs (beyond per-residue reach) do not show the same gain, and the long-contact class aggregated across $|\varepsilon|$ trends positive but is not significant ($+0.028$).

P1 and P2 (architectural controls). P1 (adding SSM alone) and P2 (T+SSM vs. SSM-only) confirm that the P3 signal is MoE-specific: P2 is essentially flat across all 19 strata, ruling out SSM-content as the source of the advantage; P1 does not improve the structurally-coupled regime. The full 19-stratum grid is shown in Fig. S6.

P4 (full stack vs. dense baseline). In aggregate, HYBRID-MOE-650M slightly underperforms the dense baseline ($\overline{\Delta\rho} = -0.007$, paired Wilcoxon $p = 0.003$, $n = 217$). The stratified picture (Fig. S8) decomposes cleanly into the P1 and P3 contributions: pair-level regimes where SSM hurts (P1) dominate the aggregate signal, while regimes where MoE helps (P3) remain visible but attenuated. Significant negative cells include medium pairs (-0.047, $p = 6\times10^{-4}$), local pairs (-0.033, $p = 0.016$), and medium $\times$ med-$|\varepsilon|$ (-0.067, $p = 2\times10^{-3}$). The largest positive cell from P3 (long-contact $\times$ medium-$|\varepsilon|$) survives in direction but attenuates to $+0.039$ (n.s., $n = 17$). This attenuation is exactly why the three-pair ablation is required: the direct full-stack vs. dense contrast masks the opposite-sign contributions of SSM and MoE.

A.5. Helix-cap structural-context probe

We use the helix-cap probe to test whether architectural differences in variant-effect prediction reflect local structural biochemistry rather than only amino-acid identity. The probe focuses on single-mutant variants whose mutant residue is one of three with a previously-characterized, position-dependent tolerance: Proline (tolerated at the helix N-cap, helix-breaking elsewhere), Glycine (favored at the helix C-cap; Richardson & Richardson, 1988), and Tryptophan (tolerated in buried hydrophobic contexts). Proline's N-cap tolerance is a single biophysically pre-specified hypothesis derived from its unique inability to donate a backbone amide hydrogen; Glycine and Tryptophan test whether the effect generalizes to other residues whose preferred positions are known.

We use the same DSSP secondary-structure and rSASA burial annotations defined in Appendix A.3. Crossing five secondary-structure contexts (`helix_Ncap`, `helix_internal`, `helix_Ccap`, `sheet`, and `loop`) with surface/buried status gives ten structural-context cells.

For each architecture and each structural-context cell we compute the per-task Spearman correlation between model LLR and experimental DMS score, restricted to single mutants of the focal residue, and report the across-task mean and 95% CI from a paired bootstrap over tasks. To verify that the underlying fitness gradient exists before attributing any architectural effect, we also compute a position-normalized tolerance $\Delta z(i) = z(i, F) - \mathrm{median}_a\, z(i, a)$, where the median runs over all other substitutions at position i (excluding the focal residue, the wild-type/synonymous residue, and non-standard symbols; ≥ 10 scores required). For Proline we test the N-cap against every other context with a rank-sum (Mann–Whitney) test.

A.6. MoE routing-organization analysis

We define a transition as a triple $t = (g_{\mathrm{wt}}, g_{\mathrm{mut}}, b)$, where $g_{\mathrm{wt}}, g_{\mathrm{mut}} \in \{\mathrm{HYD}, \mathrm{POL}, \mathrm{CHA}\}$ denote the wildtype and mutant amino-acid groups and $b \in \{\mathrm{SUR}, \mathrm{BUR}\}$ denotes burial status. We use the standard grouping HYD = AVILMFYWCGP, POL = STNQ, and CHA = DEKRH, yielding $3 \times 3 \times 2 = 18$ transition buckets.

For every single-mutant variant, we substitute the mutant residue into the wildtype sequence and run a forward pass with MoE-expert tracking. At the mutated position, we record the top-1 expert selected in each MoE layer. Aggregating over variants gives a routing-count tensor indexed by $[\ell, e, g_{\mathrm{wt}}, g_{\mathrm{mut}}, b]$, where ℓ denotes the MoE layer and e denotes the expert.

For each expert cell (ℓ, e) and transition t, we compute $\log_2$ enrichment relative to the cell's marginal usage:

$$\mathrm{enrich}(\ell, e \mid t) = \log_2 \frac{p(\ell, e \mid t)}{p(\ell, e)}. \tag{5}$$

The specialist group $S(t)$ for transition t is the set of cells satisfying

$$\mathrm{enrich}(\ell, e \mid t) \geq 1.5 \quad \text{and} \quad N(\ell, e, t) \geq 50, \tag{6}$$

corresponding to at least $\sim 2.8\times$ the global routing rate and a minimum-evidence threshold of 50 events.

To determine whether specialist groups are organized by source amino-acid class, destination amino-acid class, or burial, we compute the mean Jaccard overlap between specialist groups of all transition pairs that share each attribute. For an attribute $x \in \{\mathrm{src}, \mathrm{dst}, \mathrm{sb}\}$,

$$\bar{J}_x = \frac{1}{|\mathcal{P}_x|} \sum_{(a,b)\in\mathcal{P}_x} \frac{|S(a) \cap S(b)|}{|S(a) \cup S(b)|}, \quad \mathcal{P}_x = \{(a, b) : a \neq b,\ x(a) = x(b)\}. \tag{7}$$

A model whose experts are organized along axis x should have $\bar{J}_x$ substantially above the all-pairs baseline.

A.7. Comparisons with public open-weight models

For the cross-team ProteinGym comparison (Sec. 5.1), we score every variant under five public open-weight models with their default zero-shot scoring protocol: ESM-C-600M, ESM3-1.4B (sequence-only mode), PROFLUENT E1-300M-SEQ-ONLY, PROFLUENT E1-600M-SEQ-ONLY, and PROFLUENT E1-300M-RETRIEVAL (MSA-conditioned). Per-task NDCG@10 against the in-house ESM-650M baseline pooled across all 217 tasks is shown in Fig. S1; per-taxon stratification is shown in Fig. S2.

B. Supplementary Figures

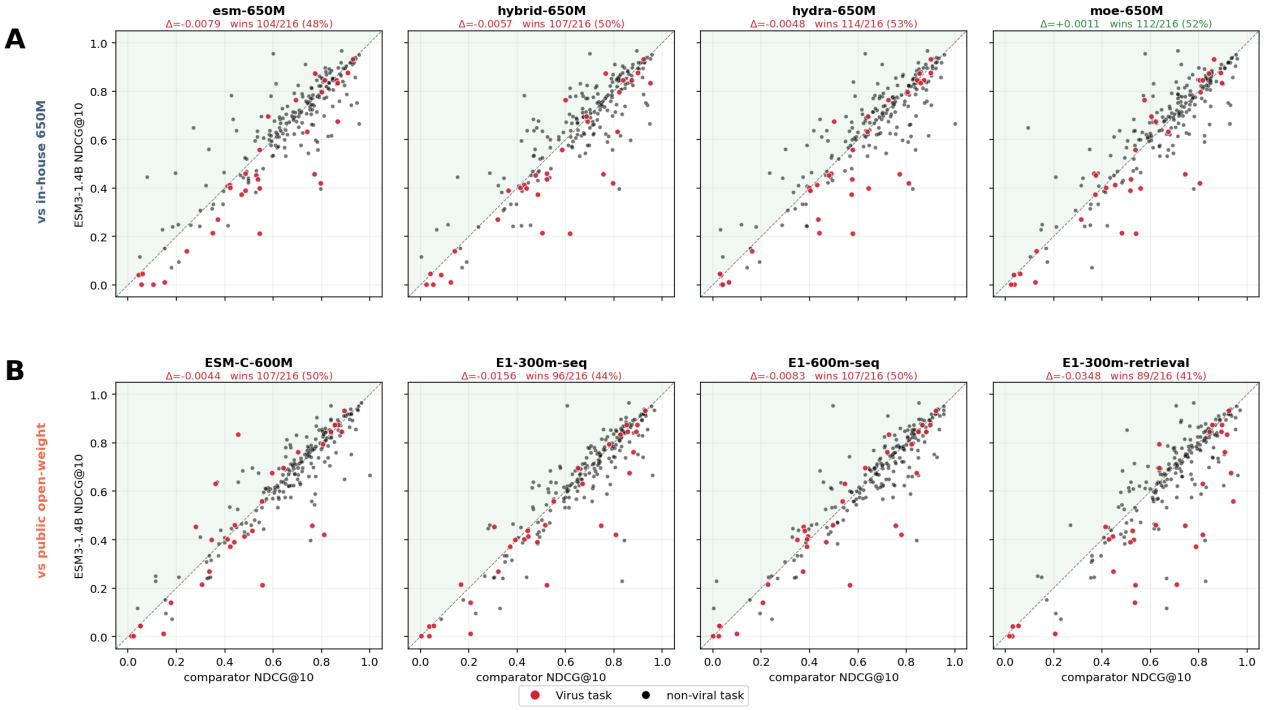


Figure S1. ESM3-1.4B (y-axis) vs. each comparator (x-axis) across all 217 ProteinGym tasks, pooled across taxa. **(A)** ESM3 vs. in-house 650M models. **(B)** ESM3 vs. public open-weight models. Each task is one point; **red** highlights the 31 viral tasks, all other tasks are black. Subtitle reports mean paired Δ (ESM3 $-$ comparator) and head-to-head win count. Viral points cluster below the identity line in every panel, indicating ESM3-1.4B underperforms on the viral subset against every comparator including in-house models with half the parameter budget; the gap is largest against MSA-conditioned E1-300M-RETRIEVAL, identifying retrieval/MSA conditioning rather than architecture as the source of the ceiling above the in-house family.

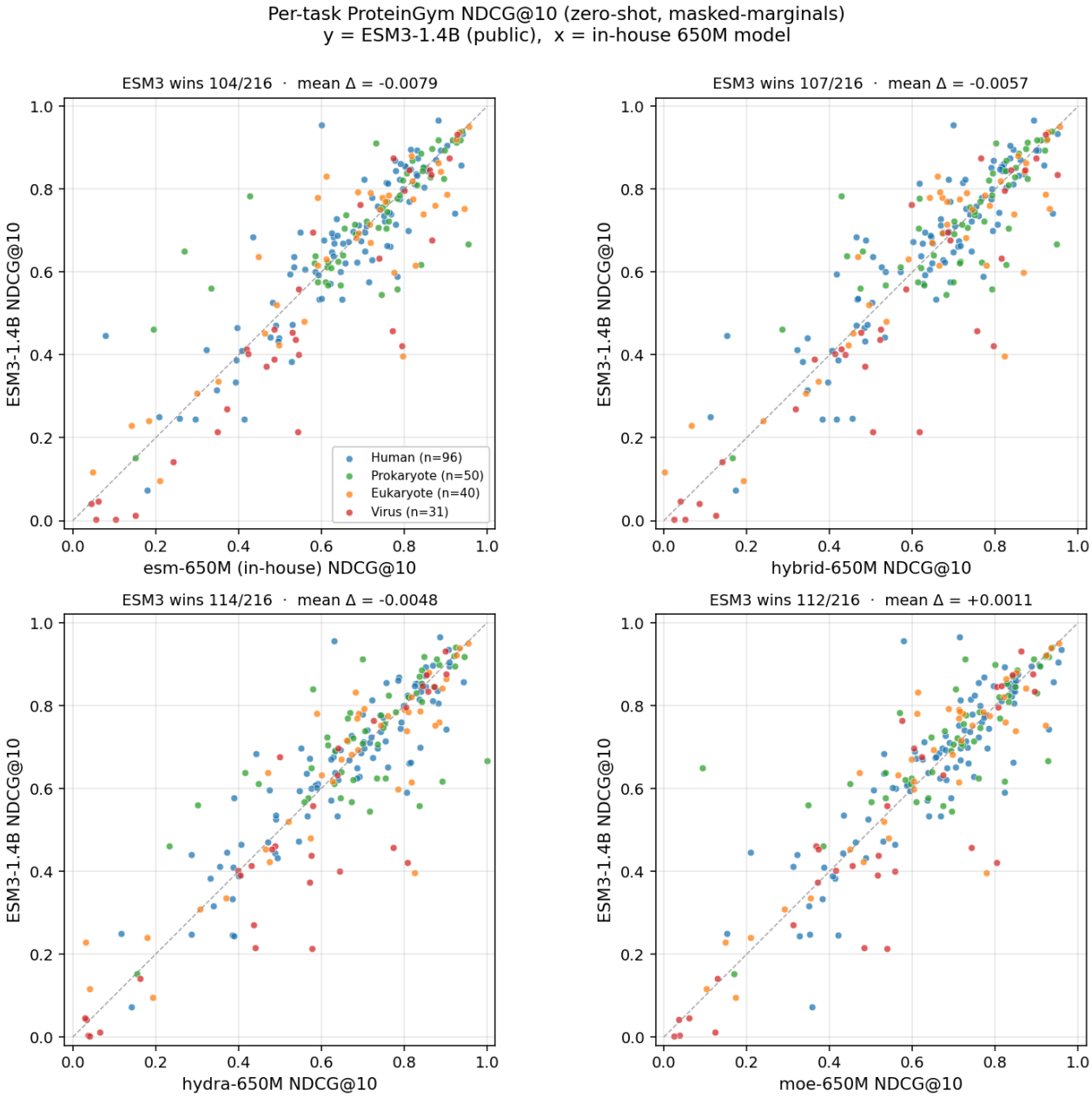


Figure S2. ESM3-1.4B vs. each comparator across 217 ProteinGym tasks, stratified by ProteinGym taxon (Human / Eukaryote / Prokaryote / Virus, columns). Red panel frames mark cells where ESM3 trails the comparator on average. ESM3's viral column is uniformly red against every comparator, confirming the pooled finding in Fig. S1.

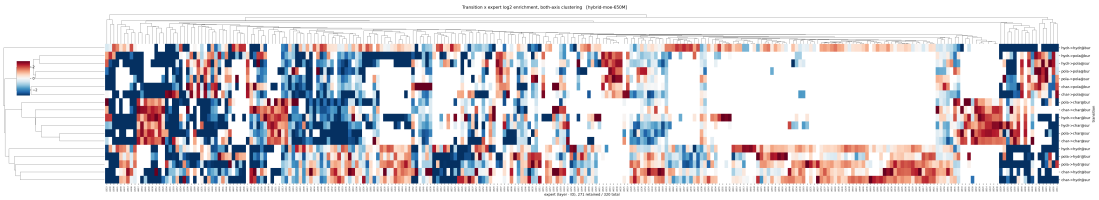


Figure S3. Mutant-forward expert clustermap for HYBRID-MOE-650M. Rows are the 18 mutant transitions; columns are (ℓ, e) MoE cells; color is $\log_2$ enrichment. Block-diagonal row structure (Jaccard distance, average linkage) is organized predominantly by destination amino-acid group.

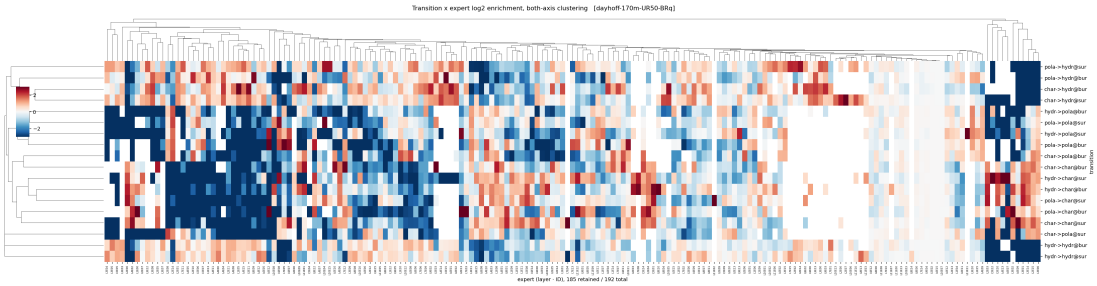
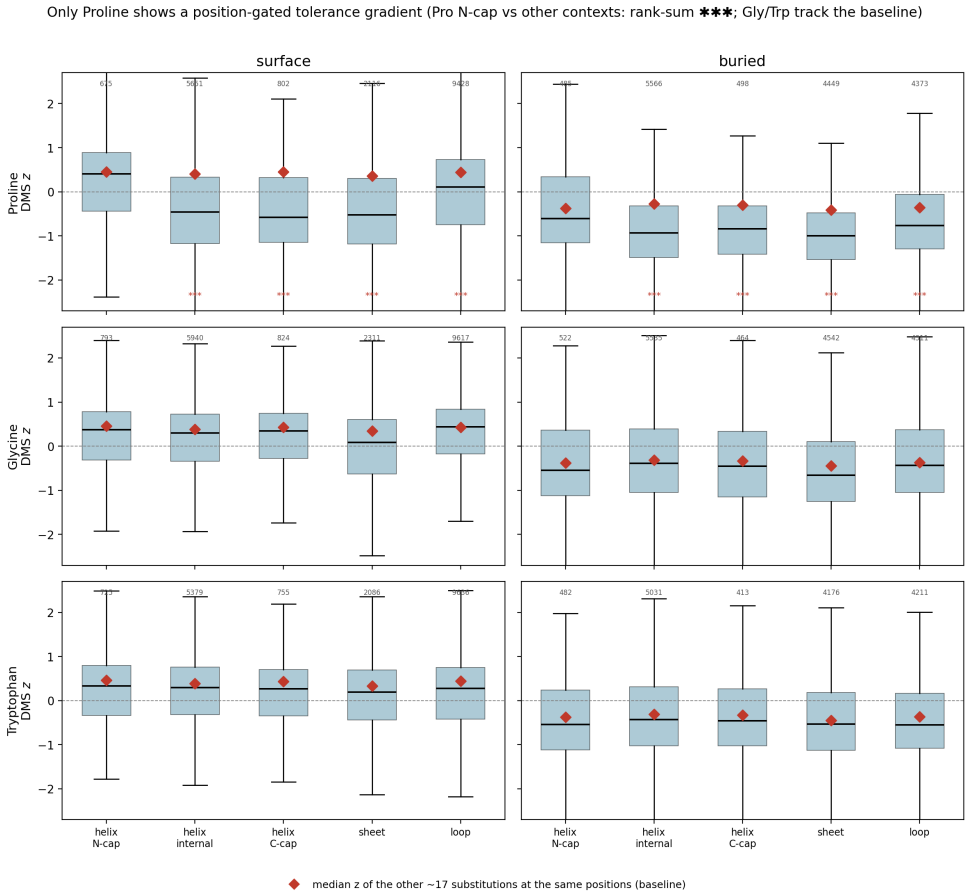


Figure S4. Mutant-forward expert clustermap for DAYHOFF-170M (independent team, Jamba architecture, UR50 corpus, narrow MoE). The destination-axis block structure recovers despite the change in architecture, MoE schedule, and pretraining corpus, supporting the cross-team replication claim in Sec. 5.2.



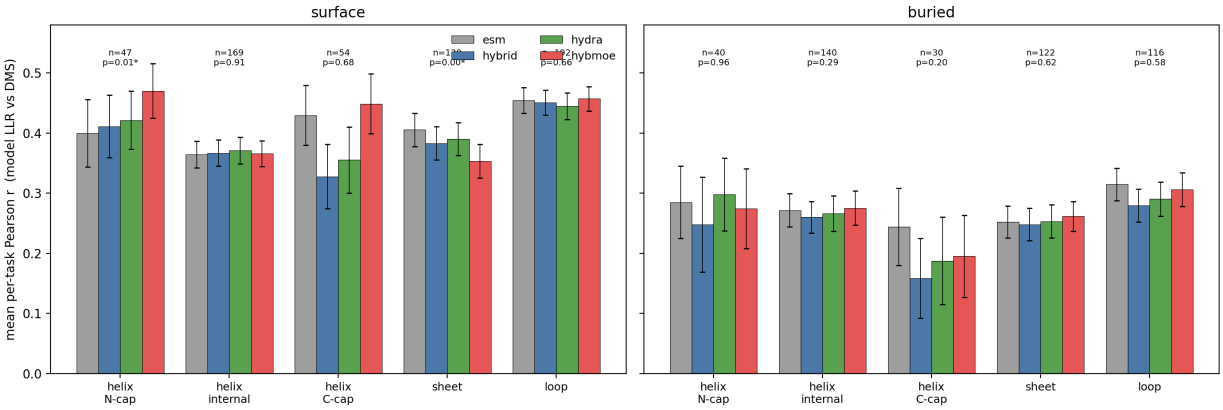


Figure S5. Top: within-task z-scored experimental DMS by structural-context cell for *→PRO, *→GLY and *→TRP (box), with the per-position substitution baseline (red ♦). Only Proline shows a position-gated gradient — tolerated at the helix N-cap and penalized elsewhere (rank-sum vs. every other context $p \leq 3 \times 10^{-5}$). Glycine (favored at the C-cap in natural proteins) and Tryptophan (buried) track the baseline in all contexts, so no position-specific signal is present to be learned. Bottom: mean per-task correlation (model LLR vs. DMS) for the four 650M architectures across the ten cells. At the pre-specified Proline surface helix N-cap, HYBRID-MOE-650M leads dense ESM-650M by $+0.10$; the remaining cells are exploratory.

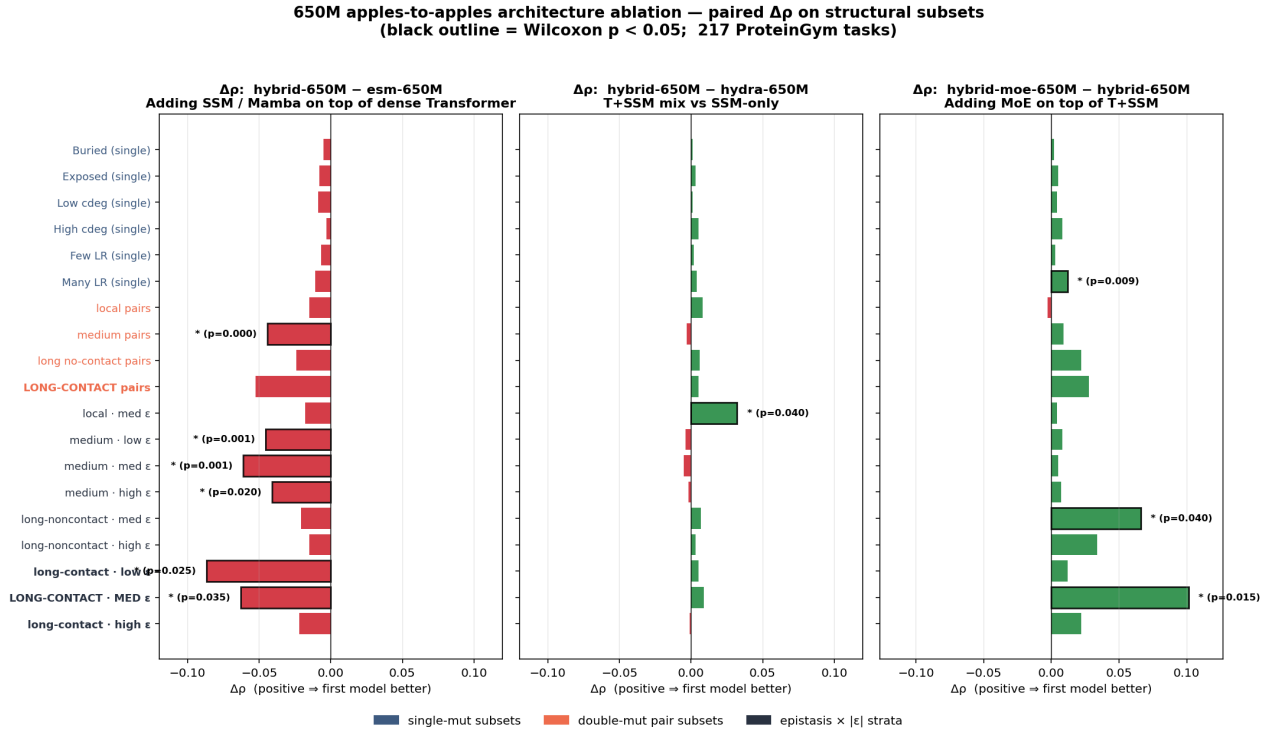


Figure S6. Full 19-stratum grid for the three 650M pairs P1, P2, P3. Rows are stratified subsets (single-mutant burial / contact-degree / long-range-contact count; double-mutant pair classes; double-mutant pair classes crossed with $|\varepsilon|$ terciles); columns are pairs P1 (red, adding SSM), P2 (gray, T+SSM vs. SSM-only), and P3 (green, adding MoE). Black outlines mark cells with paired Wilcoxon $p < 0.05$. P3 is consistently positive on long-range rows; P1 is consistently negative on the same rows; P2 is empty across nearly all cells. Headline subset of seven cells appears in Fig. S7.

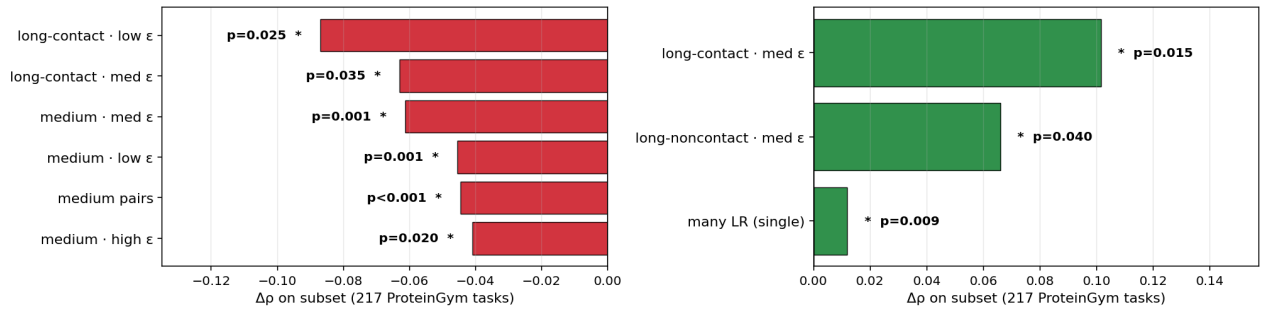


Figure S7. Significant subsets of the 650M architecture comparison ($p < 0.05$, paired Wilcoxon) across 217 ProteinGym tasks, split by pair and sorted by effect-size magnitude within each panel. Each panel's x-axis is restricted to its own sign for legibility. **Right:** P3 (HYBRID-MOE-650M − HYBRID-650M; adding MoE on top of T+SSM) is significantly positive on the long-range structurally-coupled subsets — consistent with MoE's per-residue local prior recruited once at a 3D-hub single-mutant residue (many-LR singles) and at both ends of a long-contact double-mutant pair. **Left:** P1 (HYBRID-650M − ESM-650M; adding SSM on top of dense Transformer) serves as an architectural control showing the structurally-coupled regime is not improved by adding Mamba alone. P2 (T+SSM vs. SSM-only; not shown) has no significant cells, ruling out SSM-content as the source of the advantage. P1 and P3 share the long-contact × medium-$|\varepsilon|$ subset, so comparing all three pairs — rather than a head-to-head full-stack vs. dense contrast — is the appropriate attribution.

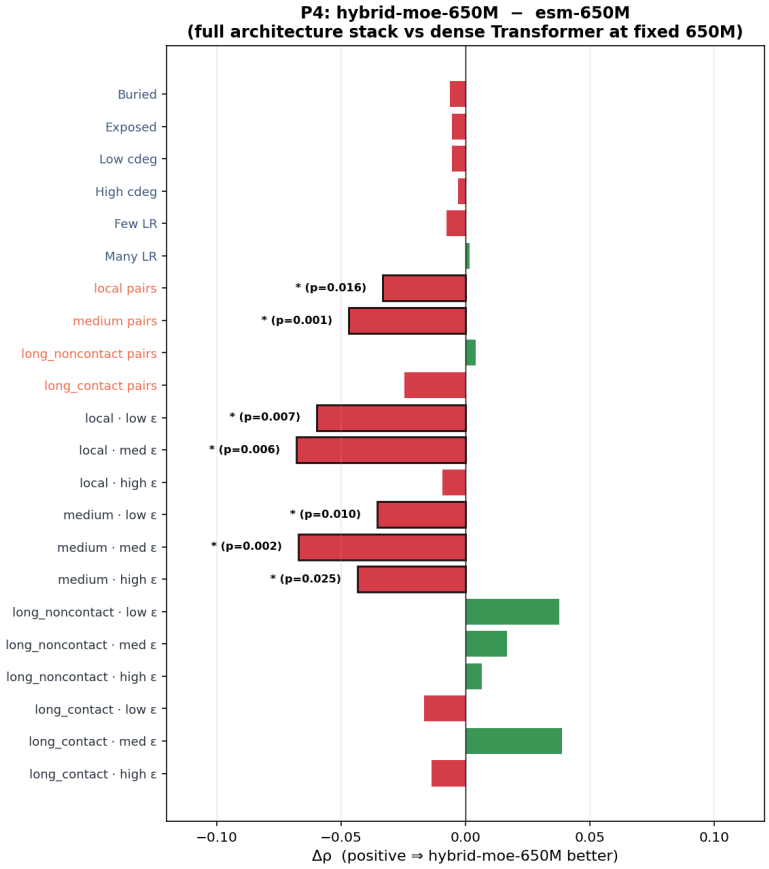


Figure S8. P4 forest plot: HYBRID-MOE-650M − ESM-650M per-stratum $\Delta\rho$ at 650M scale. Significant cells (paired Wilcoxon $p < 0.05$) are outlined; all seven significant cells are negative (red); positive cells exist on long-noncontact and long-contact $\times$ medium-ε but are attenuated by the SSM penalty (P1) and do not pass significance. Discussed in Appendix A.4.

C. Supplementary Tables

	Secondary Structure				Remote Homology				Contact P@L/5	
	ts115	cb513	casp12	Overall	Superfam.	Fold	Family	Overall	Med.	Long
Dense										
Transformer-20M	0.783	0.756	0.717	0.760	0.499	0.225	0.941	0.611	0.382	0.334
SSM-20M	0.785	0.759	0.696	0.761	0.507	**0.247**	0.930	0.615	0.413	**0.361**
Hybrid-20M	0.780	0.753	**0.722**	0.758	0.492	0.225	0.926	0.603	0.387	0.329
Sparse										
Transformer-20M	0.776	0.746	0.697	0.750	0.462	0.200	0.931	0.588	0.353	0.308
Hybrid-20M	**0.787**	**0.765**	0.719	**0.767**	**0.513**	0.238	**0.953**	**0.625**	**0.416**	**0.361**
Dense										
Transformer-150M	0.837	0.828	**0.772**	0.827	0.643	0.308	0.972	0.697	**0.551**	**0.563**
SSM-150M	0.831	0.823	0.741	0.819	0.667	0.288	**0.978**	0.705	0.498	0.494
Hybrid-150M	0.838	0.831	0.764	0.828	**0.689**	0.288	0.969	**0.710**	0.516	0.511
Sparse										
Transformer-150M	0.823	0.813	0.742	0.811	0.615	0.270	**0.978**	0.681	0.425	0.456
Hybrid-150M	**0.842**	**0.833**	0.768	**0.830**	0.667	**0.319**	0.975	**0.710**	0.515	0.516
Dense										
Transformer-650M	0.854	0.851	**0.808**	0.849	0.699	0.288	**0.987**	0.721	0.563	**0.614**
SSM-650M	**0.855**	0.852	0.803	0.849	**0.753**	**0.345**	0.985	**0.754**	**0.585**	0.613
Hybrid-650M	0.854	**0.855**	0.795	**0.850**	0.726	0.338	0.985	0.742	0.570	0.599
Sparse										
Transformer-650M	0.839	0.832	0.774	0.830	0.656	0.295	0.986	0.705	0.522	0.478
Hybrid-650M	**0.855**	0.851	0.775	0.847	0.734	0.326	0.986	0.742	0.583	0.597

Table S1. **TAPE structural-understanding results across architectural variants and parameter scales.** Three task families (secondary structure, remote-homology fold classification, and contact P@L/5) for the five architectures at 20M, 150M, and 650M. **Bold**: best on each task *within scale*. Underline: best among dense models within that scale. Ties are marked jointly.

	Overall		Amylase				IRED	NucB	TrpB			Hydrolase					Rhomax	PDZ3
	NDCG	ρ	1→Many	Close→Far	Far→Close	By-Mut.	2→Many	2→Many	1→Many	2→Many	By-Pos.	3→Many	Low→High	→P06241	→P01053	→P0A9X9	By-WT	Single→Double
Dense																		
Transformer-20M	0.938	0.331	0.505	0.175	0.313	0.646	0.187	0.234	0.465	0.504	0.115	0.469	0.377	0.115	**0.296**	−0.077	**0.481**	0.488
SSM-20M	**0.942**	**0.352**	**0.677**	**0.340**	**0.381**	0.644	**0.201**	0.524	0.345	0.509	**0.211**	0.381	0.332	0.186	0.141	−0.059	0.347	0.478
Hybrid-20M	0.940	0.333	0.355	0.336	0.370	0.644	0.200	**0.525**	**0.470**	**0.515**	0.196	**0.642**	0.333	0.164	−0.079	−0.262	0.416	0.501
Sparse																		
Transformer-20M	0.937	0.336	0.556	0.163	0.227	0.632	0.160	0.476	0.468	0.492	0.170	0.502	**0.379**	0.209	0.175	**0.022**	0.235	**0.512**
Hybrid-20M	0.938	0.324	0.356	0.178	0.339	**0.647**	0.170	0.417	0.372	0.506	0.177	0.442	0.273	**0.261**	0.291	−0.086	0.340	0.495
Dense																		
Transformer-150M	0.938	0.315	0.689	0.133	0.312	0.607	0.132	0.368	0.442	0.507	0.164	**0.553**	0.258	0.041	0.012	−0.069	0.372	0.514
SSM-150M	**0.940**	0.294	0.478	**0.288**	**0.356**	0.592	0.190	0.238	0.040	0.499	0.153	0.545	0.230	**0.224**	−0.080	0.018	0.414	0.514
Hybrid-150M	0.939	**0.322**	0.662	0.283	0.352	0.561	0.170	0.236	0.191	0.495	**0.190**	0.540	**0.411**	0.200	**0.135**	**0.373**	−0.158	**0.517**
Sparse																		
Transformer-150M	0.935	0.310	**0.711**	0.253	0.300	0.485	**0.221**	**0.437**	**0.459**	0.502	0.188	0.530	0.232	−0.062	−0.024	−0.173	0.420	0.474
Hybrid-150M	**0.940**	**0.322**	0.693	0.210	0.351	**0.650**	0.214	0.239	0.350	**0.508**	0.178	0.476	0.376	0.079	−0.155	−0.017	**0.492**	0.506
Dense																		
Transformer-650M	0.932	0.280	0.570	0.277	0.134	0.546	0.082	**0.497**	0.260	**0.507**	0.126	**0.603**	0.257	0.038	−0.060	**0.123**	0.003	**0.525**
SSM-650M	0.930	0.242	0.501	0.184	0.206	0.603	0.204	0.288	0.282	0.499	**0.201**	0.554	0.228	−0.172	−0.286	0.025	0.022	0.524
Hybrid-650M	**0.941**	**0.354**	0.570	**0.328**	**0.328**	**0.653**	0.177	0.367	0.425	0.506	0.178	0.462	**0.451**	**0.341**	**0.105**	−0.106	0.393	0.488
Sparse																		
Transformer-650M	0.933	0.319	**0.714**	0.151	0.145	0.603	**0.215**	0.347	**0.478**	0.499	0.176	0.546	0.242	0.244	0.099	0.039	0.125	0.488
Hybrid-650M	0.937	0.324	0.617	0.312	0.260	0.598	0.152	0.475	0.457	0.479	0.147	0.528	0.277	−0.127	−0.137	0.025	**0.606**	0.508

Table S2. **FLIP2 supervised out-of-distribution generalization across architectural variants and parameter scales.** Overall columns report mean NDCG and mean Spearman ρ across all 16 evaluated splits; per-task columns report Spearman ρ only. Splits are designed to shift mutation count, position, wildtype background, or sequence distance between train and test, so per-task scores measure architectural transfer beyond the local training distribution. **Bold**: best on each task *within scale*. Underline: best among dense models within that scale.