SHARED CONSTRAINTS DOMINATE PROTEIN LANGUAGE MODEL GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

When foundation models are trained on data pooled across heterogeneous generative contexts, successful zero-shot transfer can occur via two distinct mechanisms. The model may dynamically condition on latent context-specific structure, or it may rely primarily on regularities invariant across the training distribution. Protein language models (PLMs) provide an opportune testbed for this distinction, as they are trained on diverse evolutionary corpora and yet evaluated on zero-shot mutation-effect prediction defined by individual, context-dependent fitness landscapes. In this paper, using a controlled synthetic evolutionary framework with exact oracles and then a natural-corpora setting, we investigate whether PLMs learn task-invariant or task-specific structure. In both settings we find that PLM predictions track mutation effects common across training contexts rather than specific to a target. Intervening on training diversity reveals that broader ensembles strengthen this shared signal, while suppressing shared structure causes prediction to collapse rather than triggering a switch to task-specific structure. The synthetic setting reveals that learned representations encode fine-grained information about which lineage a sequence belongs to. Similarly, in the natural setting, where lineage is operationalized as a cluster of similar homologs, cluster identity is indeed decodable from local hidden states. However, we find that models of various scales do not employ this information to predict mutation effects. Altogether, our findings suggest that PLM generalization on mutation tasks is fundamentally driven by shared constraints, providing new insights into how sequence models may handle latent distribution shifts.

1 INTRODUCTION

Foundation models are trained on broad datasets that pool observations from many related but non-identical generative contexts (Bommasani et al., 2021). When a model succeeds on a new context, at least two mechanisms can produce the same aggregate performance. A model may infer which latent context generated the input and condition its prediction on context-specific structure (Xie et al., 2022), or alternatively rely mainly on regularities that remain predictive across contexts (Arjovsky et al., 2019).

Protein language models (PLMs) provide a concrete setting in which to potentially separate these mechanisms. PLMs are pretrained on vast corpora of naturally-occurring protein sequences. The sequences we observe today across organisms are the result of billions of years of adaptation across different environments in response to selective pressures exerted directly on protein function rather than sequence. Function, however, is a property of a particular protein in a particular setting, so a substitution (of an amino acid in a protein sequence) tolerated in one protein or family may be deleterious in another. Evolutionary biology formalizes this dependence as a *fitness landscape*, introduced by Wright (1932) over genotypes and recast by Maynard Smith (1970) over the space of amino-acid sequences of a single protein (which is the formulation used in this paper). Although a fitness landscape is classically defined as a global function over sequence space, the constraints it expresses are local, as they belong to a protein and its evolutionary context.

Consequently, evaluating a PLM by zero-shot mutation effect prediction, in which substitutions in a (target) protein of interest are ranked by the change they induce in model likelihood and compared against experiment, implicitly asks the model to navigate the local landscape of that target. The

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

score the model returns, however, is compatible with two mechanisms: the model may be expressing constraints that recur across evolutionary contexts, constraints particular to the target, or some combination of the two. Natural sequences do not allow us to separate them. Protein families are not independent draws. They descend from common ancestors, so the constraints and the history they share are confounded in any natural corpus. Therefore, in this paper we construct a controlled evolutionary framework in which both the sequence-generating process and every fitness landscape are known exactly. This lets us separate the exact mutation effects of each target into a component shared across the training landscapes and a residual specific to that target, and then ask which component a PLM expresses. An outcome-blind, identity-matched decomposition of natural targets complements the synthetic analysis.

Across the controlled interventions, we find that PLM mutation scores align far more strongly with constraints shared across training landscapes than with the residual specific to the target. Broadening the training ensemble at a fixed sequence budget strengthens the shared alignment, whereas suppressing shared structure removes most of the prediction signal instead of uncovering a target-specific one. This is not an inability to represent context. We show that learned representations retain the evolutionary relationships among training contexts, and for a representative pretrained PLM the context of a natural target is recoverable from hidden states at the site being scored. However, we show that what the representation identifies does not determine which landscape the prediction actually follows. Repeating our analysis on natural proteins allows us to reproduce this separation. Altogether, our findings indicate that zero-shot mutation prediction by PLMs is driven primarily by constraints shared across evolutionary contexts, and that representing context and using it to express the corresponding landscape are distinct capabilities.

2 Related Work

Information encoded versus information used. A representation may encode a factor that generated the data without using it for the actual behavior under study, so probing accuracy can overstate functional relevance (Ravichander et al., 2021). Causal interventions that remove information provide stronger evidence of this phenomenon (Elazar et al., 2021). Recent PLM interpretability work has likewise identified biologically-meaningful features and partially-separable structural knowledge in hidden representations (Adams et al., 2025; Vinod et al., 2026). Our suppression intervention acts on the data-generating process rather than on a learned representation, and our representation–prediction agreement tests ask whether decoded evolutionary context is what governs the mutation profile expressed by a model.

Controlled synthetic data. Synthetic systems expose learning dynamics that are difficult to identify in observational corpora, including grokking on algorithmic tasks (Power et al., 2022) and latent-concept inference in hierarchical data (Xie et al., 2022). Analytically structured training distributions have also been used to study how neural networks acquire features shared across, or differentiating among, semantic categories (Saxe et al., 2019). We use this strategy here to make otherwise unobserved protein fitness landscapes explicit and allow for exact mutation-effect oracles and direct interventions on the potential structure shared among contexts.

Phylogeny, homology, and zero-shot mutation prediction. PLM and multiple-sequence-alignment representations capture protein-family and phylogenetic relationships (Lupo et al., 2022; Tule et al., 2025). Masked PLMs score mutations zero-shot through log-likelihood ratios (Meier et al., 2021), yet better sequence-density estimation and larger scale do not always improve mutation-effect ranking (Gordon et al., 2025; Hou et al., 2026), which suggests that what these models learn is not simply more of the target distribution. In this paper we ask instead whether the phylogenetic information in a representation is used to select context-specific constraints. Our cohorts are matched on sequence identity, which is a standard proxy for relatedness (with well-understood limits in the remote-homology "twilight zone" (Rost, 1999; Kabir et al., 2024)).

Potts models and empirical fitness landscapes. Pairwise Potts models give explicit, context-dependent landscapes. They predict mutation effects from homolog alignments (Hopf et al., 2017) and reproduce family-level sequence statistics (McGee et al., 2021). We therefore use known Potts systems as synthetic generators, which support causal claims, and independently fitted models as natural proxies, which test qualitative correspondence. The latter are scored against ProteinGym, the standard benchmark of deep mutational scans and zero-shot evaluations (Notin et al., 2023).

3 APPROACH

3.1 CONTROLLED EVOLUTIONARY SYSTEM

We construct synthetic contexts with pairwise Potts models (McGee et al., 2021). Each model defines an explicit fitness function through site-specific residue preferences and pairwise interactions, allowing us to compute the effect of every mutation under any landscape in the construction.For a sequence x of length L over amino-acid alphabet $\mathcal{A}$, a sparse Potts landscape $C = (L, \mathcal{A}, E, h, J)$, where E denotes the interaction graph, assigns fitness $F_C(x) = \sum_i h_{i,x_i} + \sum_{(i,j)\in E} J_{ij}(x_i, x_j)$.

The synthetic landscapes are organized in a depth-7 binary *landscape tree* with edges introducing gradual changes in the Potts parameters. From the 128 terminal nodes/landscapes, we select a subset of 56: 32 for training, 8 for validation, and 16 for testing. The remaining terminal landscapes are left unassigned to allow test landscapes to be at exactly 2, 4, 6, or 8 landscape-tree edges from the nearest training landscape. We impose this spacing so that distance on the landscape tree provides an exact, sequence-independent measure of how far the governing fitness function diverges from the training ensemble.

Within each landscape, sequences evolve along a separate depth-4 binary *sequence tree* by single-residue Metropolis mutations. The landscape tree determines how the governing fitness function changes across contexts. The sequence tree determines how far sequences diverge within a context. The central experiment uses $L = 48$ and a fixed sparse interaction topology. Full generation and calibration details are in Appendix C.1. (We also repeat the same primary analyses at $L = 144$ as a sensitivity (Appendix A.6).)

For the primary held-out-landscape evaluation, mutation effects are evaluated on sequences generated under test landscapes withheld from training. We refer to an evaluated sequence as a *query background*, and to the landscape that generates it as its *target landscape*. The *mutation effect* of a substitution (at position i with amino acid a) under landscape C is the change in fitness it produces, $F_C(x^{i\to a}) - F_C(x)$. For a query background x and any landscape C in the construction, let $\boldsymbol{\Delta}_C(x)$ be the vector of mutation effects under C for all $48 \times 19 = 912$ possible single-amino-acid substitutions. We write $\boldsymbol{\Delta}_{\text{target}}(x)$ for this vector when C is the target landscape of x.

Because a landscape can be evaluated on any sequence, we can compute the mutation-effect vector of a single background under the target landscape and under each of the $K = 32$ training landscapes. Their mean, $\overline{\boldsymbol{\Delta}}_{\text{train}}(x) = \frac{1}{K}\sum_{j=1}^{K} \boldsymbol{\Delta}_{C_j}(x)$, defines the *training-ensemble shared component*: how mutations on x behave on average across the landscapes represented during training. We define the complementary *target-specific residual* as $\boldsymbol{r}(x) = \boldsymbol{\Delta}_{\text{target}}(x) - \overline{\boldsymbol{\Delta}}_{\text{train}}(x)$, which is the portion of the target vector not explained by the training-ensemble mean. Both are computable because every landscape is known exactly (which is not possible for natural corpora). We use them as complementary predictors of the model's mutation scores. The mean captures constraints recurring across training contexts, whereas the residual isolates target-specific structure.

3.2 TRANSFORMER MODELS

We train masked Transformers at three scales for the central experiment: `xs` (46k parameters), `l` (667k), and `xxxl` (6.33M). Three further scales (`m`, `xl`, `xxl`) are used only in the capacity controls of Appendix C.3. These sizes are deliberately proportionate to the controlled dataset and are sufficient to test the proposed mechanism; the synthetic study is not intended as an exhaustive scaling analysis of 100M+-parameter models. For each query background, the model's mutation-score vector contains the mutant-minus-wild-type pseudo-log-likelihood difference for each of the 912 possible single-amino-acid substitutions. The natural experiments complement the controlled study with pretrained ESM2 models(Lin et al., 2023) from 8M to 15B parameters, while separate synthetic controls examine how capacity changes the effect of heterogeneous training data.

Prediction decomposition. The PLM, shared, and target-specific residual vectors are separately rank-standardized. We regress the PLM ranks on the shared and residual predictors individually and jointly. A component's incremental rank-R^2 is the loss in joint-model R^2 when that component is removed; it therefore measures variation in PLM mutation rankings associated with that component beyond the other predictor. Per-query Spearman correlations additionally compare the PLM with

the complete target mutation-effect vector and with the individual training landscape nearest either in tree distance or in Potts-parameter RMS distance.

The *held-out-landscape cohort* contains 16 query backgrounds from each held-out landscape. Each held-out query is paired with a *remote same-landscape control*: a query generated under a landscape represented during training and chosen to have approximately the same maximum sequence identity to the training set as the held-out query. *Near same-landscape* queries are also generated under a training-represented landscape, but remain close to that landscape's training neighborhood. This design separates a change in governing fitness function from sequence remoteness. Confidence intervals for the central interventions resample the 16 held-out landscapes and quantify variation across targets within this one generated world, not across training seeds or independently generated worlds.

3.3 Intervening on ensemble diversity and shared structure

The decomposition identifies what PLM predictions resemble; two interventions test why that component is learned.

Varying the number of training landscapes. We train on nested, tree-diverse subsets of $K \in \{4, 8, 16, 32\}$ landscapes while holding the total training budget at 5,120 sequences. Each landscape therefore contributes fewer examples as K increases, and the validation and held-out cohorts remain fixed. If exposure to more contexts favors reusable constraints, PLM alignment with the shared component should increase with K and should increasingly exceed alignment with the nearest individual source landscape.

Shared-component suppression. Write the parameter vector of training landscape i as $\theta_i = m + r_i$, where m is the mean across landscapes and r_i is the landscape-specific residual. We construct a transformed world with $\theta_i' = 0.25m + cr_i$, choosing c so that mean parameter RMS is unchanged. This reduces the common-component variance fraction from 0.712 to 0.044 while preserving total parameter RMS (0.9587), thus testing whether models substitute now-dominant context-specific structure when reusable structure is removed.

3.4 Natural identity-matched cluster-Potts decomposition

We now construct the natural counterpart of Section 3.1 using ProteinGym (Notin et al., 2023). A homolog cluster plays the role of a landscape. A Potts model fitted to that cluster replaces the exact oracle. The two comparison clusters now replace the training ensemble. The correspondence is deliberate but inexact in two respects. The synthetic shared component is an average over known landscapes, whereas its natural estimator averages two fitted models, and the natural specific predictor is obtained by regression in rank space rather than by subtraction in effect space, so it is orthogonal to the shared predictor by construction. Because no intervention is possible on natural landscapes, this analysis establishes qualitative agreement rather than causal evidence.

The construction is as follows. For each assayed target, homologs in its multiple sequence alignment are restricted to a 96-residue window in target coordinates. The exact target is removed, and the remaining sequences are partitioned into operational clusters by sequence similarity, which we do not interpret as reconstructed clades. The cluster containing the homolog most similar to the target is the *target-associated cluster*; two *comparison clusters* are retained only when their distributions of identity to the target can be tightly matched to that of the target-associated cluster.

We select targets without reference to ESM2 or DMS outcomes. Five of 173 meet the prespecified count and identity-matching criteria and form the primary cohort; four more qualify for a sensitivity cohort when only the maximum pairwise KS threshold is relaxed from 0.12 to 0.15 (Appendix C.6).

We fit a separate pseudolikelihood Potts model to each matched cluster. Let T be the target-cluster mutation vector and A, B the comparison-cluster vectors. After rank standardization, the natural shared estimator is $S = (A + B)/2$, and the cluster-specific predictor is the residual of T after regressing out S. Outcomes are experimental DMS effects and ESM2 masked-marginal scores (mutant minus wild-type log probability with the site masked) at six scales from 8M to 15B parameters. We report incremental (drop-one) rank-R^2 and the target-cluster advantage,

$\rho(y, T) - [\rho(y, A) + \rho(y, B)]/2$. Split-half fits measure Potts reliability, and an errors-in-variables sensitivity analysis addresses the lower reliability of the specific residual (Appendix C.6).

3.5 Historical pretraining composition

We use the exact UniRef50 2021_04 representative corpus on which ESM2 was trained. For six targets with at least 80 usable historical homologs, we fit pooled historical Potts models and also weight the three cluster vectors by historical cluster proportions, a proxy for relative exposure rather than ESM2's sampling stream. Because historical pools differ in depth and identity, strict historical-versus-current identity-matched controls are treated as paired case studies (full details are in Appendix C.7).

3.6 Representation–prediction agreement tests

We test representation–prediction agreement separately in the synthetic and natural settings. In the synthetic held-out-landscape experiment, we use the remote same-landscape cohort, whose queries are generated under landscapes represented during training but are sequence-distant from their corresponding training examples. For each query, we pair its true training landscape with the two alternative training landscapes whose query-to-training sequence-identity distributions have the smallest two-sample KS distance from that of the true landscape; differences in mean and maximum identity break ties. Forty training sequences per landscape define cosine-centroid classifiers using either whole-sequence embeddings (global) or masked-site hidden states (local). Leave-one-out support accuracy measures whether the three candidate landscapes are distinguishable in representation space, whereas three-way query accuracy measures whether the remote query is assigned to its true landscape (chance $= 1/3$).

Because the synthetic landscapes are known exactly, we can also test whether the context selected by the representation predicts the mutation profile expressed by the PLM. For an incorrectly assigned query or site, we define the *same-wrong profile margin* as the PLM mutation-profile Spearman correlation with the representation-assigned wrong landscape minus its correlation with the true generating landscape.A positive margin would indicate that the prediction agrees better with the context the representation selected than with the context that actually generated the query. Landscape-level bootstrap intervals treat the 16 represented landscapes as replicates. The full cohort contains 256 remote queries; because some candidate triplets are poorly identity matched, we additionally report a stricter sensitivity analysis requiring both comparison-landscape KS distances to be at most 0.12, which retains 156 queries spanning all 16 landscapes (Appendix A.2).

For natural proteins, we test analogously whether evolutionary context decoded from a whole-sequence embedding (global) or a masked site's hidden state (local) predicts the mutation profile ESM2 expresses. Because these tests require the same three identity-matched operational clusters used in the cluster-Potts analysis, we use its fixed nine-target sensitivity cohort. We measure cluster separation and preservation of a tree inferred from sequence distances. We then test a discrete context-selection hypothesis: if a site's representation is assigned to a comparison-cluster centroid, its ESM2 substitution profile should more closely match that same cluster's Potts profile than the other comparison profile. We call this correlation difference—representation-assigned minus other comparison cluster—the *attribution-aligned profile margin*. A positive margin supports agreement between the context implied by the representation and the context expressed in prediction. Scale-averaged endpoints and bootstrap intervals treat proteins, not model scales, as replicates. Approximate 80%-power minimum detectable effects (MDEs) at two-sided $\alpha = 0.05$ describe sensitivity rather than evidence for an exact zero effect. Mutation-level and site-permutation controls appear in Appendices C.8 and C.9.

3.7 Additional evolutionary controls

Further analyses in the Appendix vary genealogical distance, lineage breadth, selection regime, and model capacity, and probe whether prediction-risk rankings transfer across these mechanisms. Others examine repeated ProteinGym assays, structurally defined CATH/Gene3D groups, and exact historical pretraining support.

4 RESULTS

We first determine which component of a target landscape PLM mutation scores express. We then intervene on the training ensemble, repeat the decomposition in natural proteins, and finally test whether the evolutionary context a model represents governs the mutation profile it predicts.

4.1 PLMS RECOVER PREDOMINANTLY SHARED SYNTHETIC CONSTRAINTS

The landscapes define meaningfully different mutation tasks. For each query, we define *oracle similarity* as the Spearman correlation between the complete single-mutant effect vectors produced by two exact Potts landscapes. The target's similarity to its nearest training landscape averages 0.914 and decreases with the number of intervening landscape-tree edges. Potts-parameter RMS distance is strongly anticorrelated with oracle similarity ($\rho = -0.957$). The tasks are therefore related, but not interchangeable.

Most recovered signal is associated with the shared component. On held-out landscapes, PLM scores correlate with the training-ensemble component at ρ between 0.126 and 0.159 across the three scales, with the complete target at 0.107 to 0.137, and with the target-specific residual at only 0.016 to 0.027. The same ordering holds for identity-matched remote queries governed by represented landscapes, indicating that it does not arise from the change to a held-out fitness function. In the joint decomposition, the shared component accounts for most of the uniquely attributable rank signal at every model scale: 80–85% of the summed shared and specific incremental rank-R^2 contributions on held-out landscapes, and 77–82% on the remote same-landscape controls.

The model is not merely copying the nearest training landscape.

Two simpler explanations fail. The model is not copying its nearest neighbor: correlation with the single closest training landscape is lower than with the ensemble component, whether proximity is measured on the landscape tree (0.117 to 0.141) or in Potts-parameter space (0.114 to 0.141). Nor is it reconstructing the ensemble mean, which is itself far more predictive of the target ($\rho \approx 0.90$) than any model we train. Model performance therefore leaves substantial target-landscape structure unexplained; our analysis asks which components of the recovered signal are shared across training landscapes and which are target-specific. Within the recovered signal, the shared component dominates at every capacity (Fig. 1A).

This is a statement about composition, not literal arithmetic averaging. The exact ensemble mean is itself highly correlated with the target landscape ($\rho \approx 0.90$), far more strongly than the PLM is. We therefore conclude that the portion of mutation structure recovered by the PLM is predominantly reusable across contexts, not that the network reconstructs the exact ensemble mean.

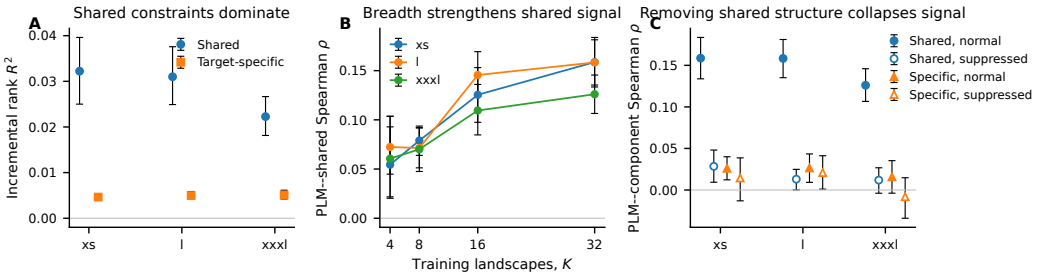


Figure 1: PLMs preferentially recover constraints reusable across evolutionary contexts. A, At baseline $K = 32$, the incremental (drop-one) rank-R^2 contribution of the shared training-ensemble component exceeds that of the target-specific residual at every tested capacity. **B,** At a fixed total budget of 5,120 sequences, increasing the number of training landscapes strengthens alignment between PLM scores and the shared component. **C,** Suppressing the common Potts-parameter component while preserving total parameter RMS sharply reduces shared alignment without a compensating increase in absolute target-specific alignment. Panel A reports incremental rank-R^2; panels B–C report Spearman ρ. Points summarize the same 16 held-out landscapes. Error bars are 95% landscape-bootstrap intervals and quantify target-landscape heterogeneity within one generated world, not training-seed or world-to-world uncertainty.

4.2 INTERVENTIONS SUPPORT SHARED-CONSTRAINT GENERALIZATION

The preceding decomposition is correlational. We next change the diversity of the training ensemble and the amount of structure shared among its members.

More training contexts strengthen the shared component. At a fixed 5,120-sequence budget, PLM–shared Spearman for `xs` rises from 0.054 at $K = 4$ to 0.159 at $K = 32$; for `l`, from 0.072 to 0.158; and for `xxxl`, from 0.061 to 0.126. The paired increases are $+0.104$ (95% landscape-bootstrap CI $[0.059, 0.149]$), $+0.086$ ($[0.050, 0.121]$), and $+0.065$ ($[0.017, 0.110]$), respectively (Fig. 1B). This cannot be explained simply by the $K = 32$ shared oracle being much closer to the target: shared–target similarity changes only from approximately 0.878 to 0.898. Moreover, the advantage of the shared component over the nearest-parameter source grows with K by $+0.025$, $+0.019$, and $+0.012$. Exposure to more landscapes therefore increases alignment with cross-context structure faster than alignment with any individual source.

Separate fixed-budget controls show that the effect of heterogeneous data on overall accuracy is capacity dependent, hurting `xs` and helping larger models, which concerns how well the task is done rather than which component is learned (Appendix A.3).

Removing shared structure removes signal rather than exposing a strong specific predictor. The suppression intervention reduces shared parameter variance from 0.712 to 0.044 at unchanged total parameter RMS, lowering target–shared oracle similarity from 0.898 to 0.219. Prediction does not switch to the specific component, and overall predictive signal instead decreases. PLM–shared Spearman falls to between 0.012 and 0.028, and the summed incremental rank-R^2 of the two components falls by roughly half at every scale (Fig. 1C). The specific component becomes a larger *fraction* of a much smaller total, rising in incremental rank-R^2 from about 0.005 to about 0.008, while its raw correlation with predictions does not improve and is slightly negative at the largest scale (-0.008). The models therefore do not compensate for the loss of reusable constraints by recovering an independent, target-specific fitness function.

The specific correction is strongest near training-represented sequence neighborhoods. For queries generated under landscapes included in training, PLM–specific Spearman falls by roughly a factor of four as queries move from near to remote training neighborhoods (for example, 0.111 to 0.024 for `xs`). After matching this sequence remoteness, replacing the represented landscape with a held-out one produces essentially no further loss. Unlabeled continuation on remote sequences from the correct landscape partially restores correlation with the complete target and its specific residual, but changes the preregistered normalized partial rank-R^2 metric little. The context-specific correction is therefore detectable, but modest, local, and difficult to recover selectively from unlabeled sequences (Appendix C.2).

4.3 NATURAL PROTEINS SHOW THE SAME SHARED-DOMINANT PATTERN

We next test whether the same decomposition holds when the underlying fitness landscapes are unknown and must be inferred from sequence data. The primary cohort comprises five ProteinGym targets with target-associated and comparison clusters matched in sequence-identity distributions within the prespecified tolerances. The nine-target sensitivity cohort relaxes only the KS threshold.

Shared cluster constraints explain substantially more than the target-specific residual. In the strict cohort, ESM2 incremental shared rank-R^2 is 0.171, 0.263, 0.315, 0.257, 0.248, and 0.221 from 8M through 15B parameters. Incremental specific rank-R^2 is much smaller: 0.036, 0.053, 0.055, 0.054, 0.046, and 0.034. Shared exceeds specific in every one of the 30 target-by-scale comparisons, and in all 54 comparisons in the nine-target sensitivity cohort (Fig. 2). This does not mean the target-associated cluster is irrelevant: its complete Potts vector has a positive mean correlation advantage of $+0.052$ to $+0.076$ over the comparison clusters. DMS effects show the same hierarchy, with incremental shared rank-R^2 of 0.076 (95% CI $[0.040, 0.112]$) versus 0.0098 ($[0.0036, 0.0160]$) for the specific residual.

Unequal Potts reliability does not account for the result. The shared predictor is more reproducible across split-half Potts fits than the specific residual (0.891 versus 0.334), which could otherwise favor the shared component. An errors-in-variables correction narrows but does not remove the gap: corrected ESM2 shared incremental R^2 ranges from 0.097 to 0.159 across scales, com-

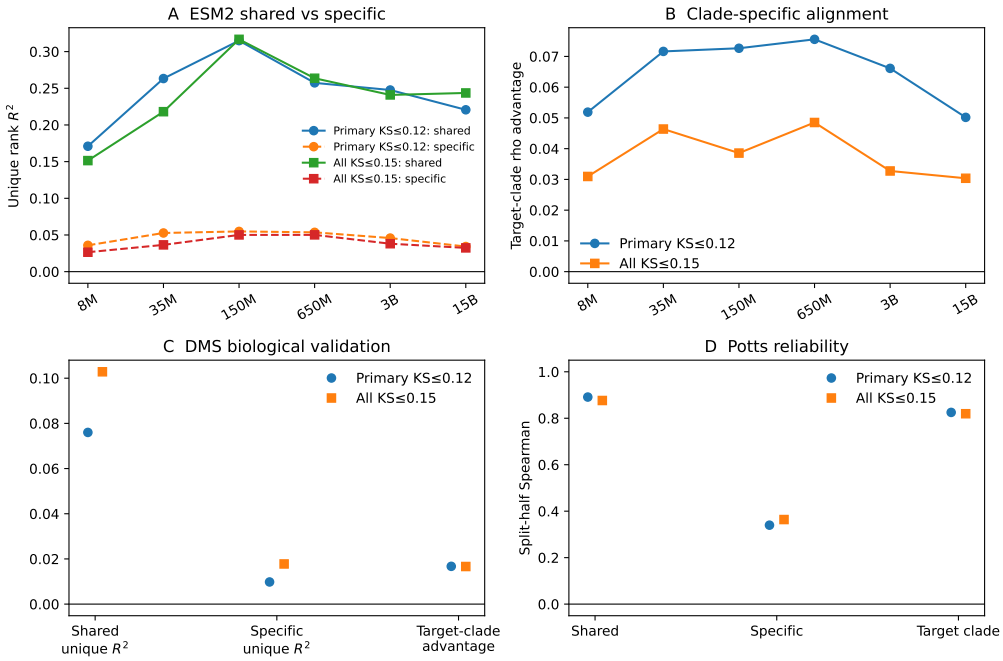


Figure 2: Natural ESM2 scores are dominated by constraints shared across identity-matched sequence contexts. An outcome-blind scan defines a strict five-target cohort and a nine-target sensitivity cohort. **A,** Across six ESM2 scales, incremental shared rank-R^2 exceeds that of the target-cluster-specific residual. **B,** The complete target-cluster vector nevertheless retains a smaller positive advantage over comparison clusters. **C,** Experimental DMS effects show the same hierarchy. **D,** The specific residual is less reliable across split-half fits; errors-in-variables correction narrows but does not remove shared dominance.

pared with 0.037 to 0.061 for the specific component. Corrected DMS values are 0.047 and 0.011, respectively, and the sensitivity cohort is concordant.

Context structure is represented but does not determine a discrete mutation profile. A synthetic genealogical-distance control shows that after adjusting flexibly for maximum training identity, far-versus-near fitness-Spearman contrasts are only -0.006, -0.007, and -0.013 across scales, with all intervals spanning zero. Fine tree recovery remains distance sensitive (contrasts -0.130, -0.160, and -0.289), so representation geometry persists beyond the functional-transfer signal explained by identity (Appendix A.1).

We next test representation–prediction agreement directly in the synthetic held-out-landscape setting. Among identity-matched three-landscape candidate sets, training support sequences are sharply separable: leave-one-out centroid accuracy is 0.963–0.980 across global and local representations. Remote queries, however, are not reliably assigned to their true training-represented landscape. Three-way accuracy ranges from 0.313 to 0.387 globally and from 0.344 to 0.383 locally, near the $1/3$ chance level. More importantly, when a local representation is assigned to the wrong landscape, the PLM mutation profile is no more correlated with that landscape's exact fitness function than with the true one: the same-wrong profile margin ranges from -0.0035 to $+0.0013$ across scales, with every landscape-bootstrap interval spanning zero. Applying the stricter KS ≤ 0.12 matching criterion retains 156 queries across all 16 landscapes and gives the same qualitative result (Appendix A.2). Thus distinct training contexts are represented internally, but remote context retrieval is unreliable and an incorrect contextual assignment does not act as a switch to the corresponding fitness function.

Natural ESM2 representations provide a complementary setting in which local homolog-cluster context is decodable at the assayed site. Local hidden states contain enough cluster information for leave-one-out centroid accuracy on support homologs to rise from 0.658 at 8M to 0.750 at 3B. We test whether the comparison cluster selected by the local representation is the same cluster whose mutation profile the prediction follows. For sites assigned to one of the two comparison

clusters, we compare the ESM2 substitution profile's correlation with that assigned cluster's Potts profile against its correlation with the other comparison cluster. Averaging scales within protein, the attribution-aligned profile margin is only $+0.007$ (95% target-bootstrap CI $[-0.008, 0.022]$, $n = 7$; $\text{MDE}_{80} = 0.027$). Sites assigned to a comparison cluster are modestly harder, with a larger ESM2–DMS profile error ($+0.043$, 95% CI $[0.004, 0.079]$, $n = 6$; the two endpoints require different site coverage, hence the differing n), yet their predictions are no closer to that same cluster's Potts profile. The synthetic and natural analyses show that distinguishable contextual structure in a representation does not by itself imply context-specific mutation prediction (Appendices A.2, C.8–C.9).

4.4 Which shared mixture ESM2 expresses varies with pretraining composition.

The equally weighted comparison-cluster estimator $S = (A + B)/2$ is a controlled definition of shared information; it need not equal the mixture learned by ESM2. Across six targets with at least 80 usable historical homologs, the correlation advantage of a Potts model fitted to the historical pool over the target-cluster model changes from -0.022 at 8M and -0.012 at 35M to $+0.072$, $+0.101$, and $+0.092$ at 650M, 3B, and 15B. Independently, weighting the three cluster-specific vectors by their historical cluster proportions improves over the equal-weighted estimator at every scale, by $+0.027$ to $+0.048$ Spearman. Only two proteins satisfy the strict historical-versus-current identity-matching control. Both favor the current pool at 8M but the historical pool at 650M, 3B, and 15B; intermediate scales are mixed (Appendix Table 4). Because historical pools also modestly improve DMS alignment, these cases do not isolate pretraining memorization. They instead support the narrower conclusion that pretraining composition helps determine which reusable constraints dominate ESM2 predictions; they do not imply that ESM2 implements a literal pooled Potts model.

5 Implications, limitations, and future work

Equal benchmark scores need not reflect equal information. A model that captures reusable constraints and a model that infers target-specific constraints can rank mutations similarly in well-populated families and yet rely on different sources of signal. Evaluation should therefore distinguish sequence identity from generative context as separate axes of held-out design, and distinguish evidence that a factor is encoded from evidence that it is used for prediction. This distinction is especially relevant when scores are interpreted for an individual protein, where the target-specific component is not supplied by averaging across homologs.

Our controlled setting isolates discrete selection through short sequences, pairwise interactions, and balanced phylogenies, while the reported intervals capture variation across held-out targets. The natural-protein analysis complements this controlled setting. Together, these experiments test the proposed decomposition across controlled and natural settings while leaving broader forms of distributed conditioning for future study.

A direct extension of this work could provide the target neighborhood explicitly. Alignment- and retrieval-conditioned models expose such context to the predictor, allowing the decomposition developed here to test whether conditioning increases the target-specific component or instead strengthens shared structure. More generally, this framework provides a way to evaluate whether additional context supplies information that changes what exactly a model predicts rather than just improving aggregate benchmark performance.

6 Conclusion

Across controlled and natural systems, PLM mutation predictions are shaped primarily by constraints that recur across evolutionary contexts. Broader synthetic landscape ensembles strengthen this shared signal, whereas removing shared structure eliminates much of the prediction signal without revealing a strong target-specific alternative. At the same time, lineage and cluster information remain decodable from learned representations. The results show that representing evolutionary context does not imply recovery of the mutation landscape governing a particular target. PLM generalization in these experiments is best described as shared-constraint integration with smaller, more local context-specific corrections.

AI USE STATEMENT

Generative AI tools were used to aid and polish writing, support retrieval and discovery of related work, and assist with research ideation, and AI tools were used to assist with code implementation for most experiments. The authors reviewed all AI-assisted content and take responsibility for the final work.

REFERENCES

Etowah Adams, Liam Bai, Minji Lee, Yiyang Yu, and Mohammed AlQuraishi. From mechanistic interpretability to mechanistic biology: Training, evaluating, and interpreting sparse autoencoders on protein language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 460–476, 2025.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

Cade Gordon, Amy Lu, and Pieter Abbeel. Protein language model fitness is a matter of preference. In *International Conference on Learning Representations*, volume 2025, pp. 39759–39775, 2025.

Thomas A. Hopf, John B. Ingraham, Frank J. Poelwijk, Charlotta P. Schärfe, Michael Springer, Chris Sander, and Debora S. Marks. Mutation effects predicted from sequence co-variation. *Nature Biotechnology*, 35:128–135, 2017.

Chao Hou, Di Liu, Aziz Zafar, and Yufeng Shen. Understanding language model scaling for protein fitness prediction. *Nature Computational Science*, pp. 1–11, 2026.

Anowarul Kabir, Asher Moldwin, Yana Bromberg, and Amarda Shehu. In the twilight zone of protein sequence homology: do protein language models learn protein structure? *Bioinformatics Advances*, 4(1):vbae119, 2024.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Umberto Lupo, Damiano Sgarbossa, and Anne-Florence Bitbol. Protein language models trained on multiple sequence alignments learn phylogenetic relationships. *Nature Communications*, 13: 6298, 2022.

John Maynard Smith. Natural selection and the concept of a protein space. *Nature*, 225(5232): 563–564, 1970. doi: 10.1038/225563a0.

Francisco McGee, Sandro Hauri, Quentin Novinger, Slobodan Vucetic, Ronald M. Levy, Vincenzo Carnevale, and Allan Haldane. The generative capacity of probabilistic protein sequence models. *Nature Communications*, 12:6302, 2021.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Yarin Gal, and Debora Marks. ProteinGym: large-scale benchmarks for protein fitness prediction and design. In *Advances in Neural Information Processing Systems*, volume 36, pp. 64331–64379, 2023.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3363–3377, 2021.

Burkhard Rost. Twilight zone of protein sequence alignments. *Protein Engineering*, 12(2):85–94, 1999.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116 (23):11537–11546, 2019.

Sanjana Tule, Gabriel Foley, and Mikael Bodén. Do protein language models learn phylogeny? *Briefings in Bioinformatics*, 26(1):bbaf047, 2025. doi: 10.1093/bib/bbaf047.

Ria Vinod, Ava P. Amini, Lorin Crawford, and Kevin K. Yang. Trainable subnetworks reveal insights into structure knowledge organization in protein language models. *PLOS Computational Biology*, 22(2):e1013925, 2026. doi: 10.1371/journal.pcbi.1013925.

Sewall Wright. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. In *Proceedings of the Sixth International Congress of Genetics*, volume 1, pp. 356–366, 1932.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit Bayesian inference. In *International Conference on Learning Representations*, 2022.

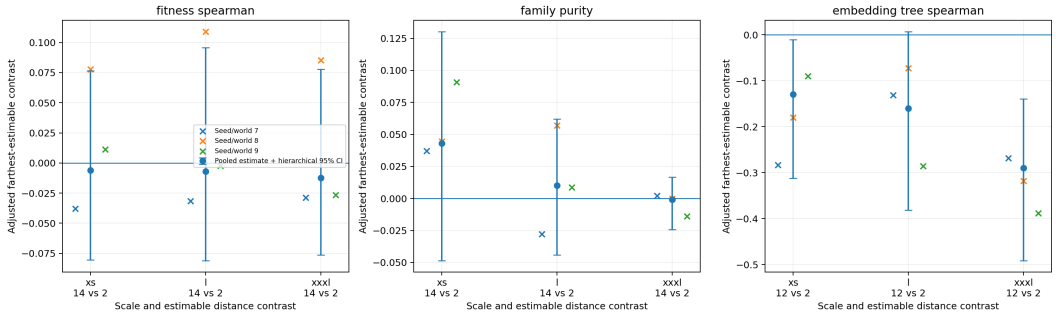


Figure 3: Identity-adjusted genealogical-distance contrasts. Fitness and family-purity effects at the farthest distance are near zero after flexible identity adjustment, while fine tree recovery retains a clear genealogical penalty.

A Additional Analyses

The analyses below test the scope of the shared-constraint account, evaluate alternative explanations, and examine consequences for reliability and natural-data generalization.

A.1 Sequence identity mediates functional transfer while phylogenetic geometry remains represented

Across three independently generated worlds, raw mutation-fitness transfer declines monotonically with exact nearest-training tree distance. Over the same range, however, maximum training identity also declines. After flexible identity adjustment, the distance-14 versus distance-2 fitness contrast is effectively zero at every scale: -0.006 for `xs` (95% interval $[-0.081, 0.077]$), -0.007 for `l` ($[-0.081, 0.096]$), and -0.013 for `xxxl` ($[-0.076, 0.078]$). We measure family purity as the fraction of a query embedding's nearest neighbors with the same synthetic-family label. Fine tree recovery remains distance sensitive: the farthest estimable distance-12 versus distance-2 contrasts are -0.130, -0.160, and -0.289, respectively. Thus ordinary sequence identity summarizes the functional-transfer boundary much better than it summarizes retained representation geometry. (Fig. 3)

A.2 Synthetic representation–prediction agreement under matched contexts

The genealogical-distance analysis above asks whether representation geometry persists beyond the boundary of functional transfer. We additionally test a more direct hypothesis in the E3 landscape-shift experiment: whether the fitness context selected from a representation determines which landscape-specific mutation profile the model expresses. We use the remote same-landscape cohort, for which the generating landscape is present during training but the query is deliberately sequence-distant from that landscape's training examples. The analysis uses the validation-selected `xs`, `l`, and `xxxl` checkpoints.

For each remote query, we construct a three-landscape candidate set containing its true generating landscape and two alternative training landscapes. Alternatives are selected without using PLM outputs or fitness scores: for every training landscape we compute the full distribution of query-to-training-sequence identities and rank alternatives by the two-sample KS distance from the corresponding distribution under the true landscape, breaking ties by differences in mean and maximum identity. Forty training sequences from each candidate landscape define cosine-centroid classifiers. The global classifier uses whole-sequence embeddings; the local classifier masks the queried site and uses its hidden state. Leave-one-out classification of the support sequences asks whether the three candidate contexts occupy distinguishable representation regions, while classification of the remote query asks whether its representation retrieves the true training-represented landscape.

The full cohort contains 256 queries from 16 represented landscapes. Candidate matching is strong for most queries but has a long tail (median comparison KS $= 0.0875$, maximum $= 0.981$), so we additionally report a strict matching sensitivity requiring both comparison-landscape KS values to be

Table 1: Synthetic representation–prediction agreement under strict identity matching. Results use the 156 remote same-landscape queries for which both alternative landscapes have KS distance ≤ 0.12 from the true landscape's query-to-training identity distribution; all 16 represented landscapes remain. Support LOO is three-way leave-one-out centroid accuracy among training sequences. Query accuracy is assignment of the remote query to its true landscape (chance $= 1/3$). Same-wrong is the local mutation-profile correlation with the representation-assigned wrong landscape minus that with the true landscape, evaluated only on incorrect assignments. Brackets are 95% landscape-bootstrap intervals.

Scale	Global support LOO	Local support LOO	Global query accuracy	Local query accuracy	Local same-wrong margin
`xs`	0.966 [0.958, 0.973]	0.979 [0.974, 0.983]	0.391 [0.301, 0.487]	0.355 [0.258, 0.455]	$+0.0023$ [-0.0043, $+0.0098$]
`l`	0.968 [0.962, 0.974]	0.975 [0.970, 0.979]	0.314 [0.217, 0.416]	0.309 [0.215, 0.410]	-0.0012 [-0.0086, $+0.0071$]
`xxxl`	0.962 [0.955, 0.969]	0.968 [0.962, 0.974]	0.328 [0.273, 0.390]	0.336 [0.275, 0.394]	$+0.0013$ [-0.0024, $+0.0052$]

at most 0.12. This restriction retains 156 queries and all 16 landscapes; among retained comparisons the mean KS is 0.072 and the maximum is 0.119. The restriction is based only on sequence identity and does not use representations or mutation-profile outcomes.

Because every synthetic fitness function is known, we can directly test whether the context selected by the representation agrees with the context expressed in prediction. For each query or masked site we compute the PLM mutation vector's Spearman correlation with the exact Potts mutation vector under each of the three candidate landscapes. On incorrect context assignments, the *same-wrong profile margin* is

$$\Delta_{\mathrm{SW}} = \rho(\mathrm{PLM}, \Delta F_{\text{assigned wrong}}) - \rho(\mathrm{PLM}, \Delta F_{\text{true}}).$$

A positive value would indicate that the prediction is redirected toward the same incorrect landscape favored by the representation. Confidence intervals bootstrap the 16 target landscapes.

Across the full cohort, matched support contexts are sharply separable: leave-one-out accuracy ranges from 0.963 to 0.969 globally and from 0.968 to 0.980 locally, far above the three-way chance level of $1/3$. This separability does not translate into reliable context retrieval for the remote queries themselves. Global query accuracy ranges from 0.313 to 0.387 and local accuracy from 0.344 to 0.383. Representation–profile agreement is likewise near chance, and the local same-wrong margins range only from -0.0035 to $+0.0013$, with all landscape-bootstrap intervals spanning zero.

The stricter KS ≤ 0.12 subset gives the same conclusion (Table 1). Support landscapes remain almost perfectly separable, while remote-query assignment remains near chance and incorrect local assignments do not redirect the predicted mutation profile toward the corresponding wrong fitness landscape. The synthetic result therefore localizes the failure more precisely than representation decoding alone: distinct fitness contexts are present in representation space, but sequence-remote queries do not reliably retrieve the correct context, and whichever alternative context they approach does not act as a discrete selector of its exact fitness function.

A.3 The value of heterogeneous evolutionary data depends on model capacity

Genealogical breadth at fixed sequence budget. Under the fixed 720-sequence lineage-breadth budget, natural-breadth sampling increases the number of represented training nodes from 48 at source depth 1 to 720 at depths 5–6, whereas matched-lineage sampling retains 48 nodes throughout (Fig. 4a). Under natural-breadth sampling, increasing source depth from 1 to 6 changes mean mutation-fitness Spearman from 0.293 to 0.167 for `xs`, but from 0.450 to 0.593 for `xxxl` (Fig. 4b). Family purity falls substantially with source depth for `xs` but remains near ceiling for `xxxl` (Fig. 4c). Fine tree recovery eventually declines for both scales, although it remains higher for `xxxl` (Fig. 4d).

To separate the effect of breadth from the effect of source depth, we compare broadly sampled and matched-lineage cohorts at the same depth of 6 and the same training and validation budgets, averaged over the common target-depth evaluation cohort within each generated world. The broad-minus-matched Spearman contrast is -0.072 (SEM 0.007) for `xs`, $+0.181$ (0.008) for `xxl`, and $+0.194$ (0.005) for `xxxl` across three independent worlds. Each endpoint has the same sign in all three worlds. The `m`, `l`, and `xl` contrasts vary appreciably among worlds; these data do not determine a precise capacity threshold or a smooth crossover.

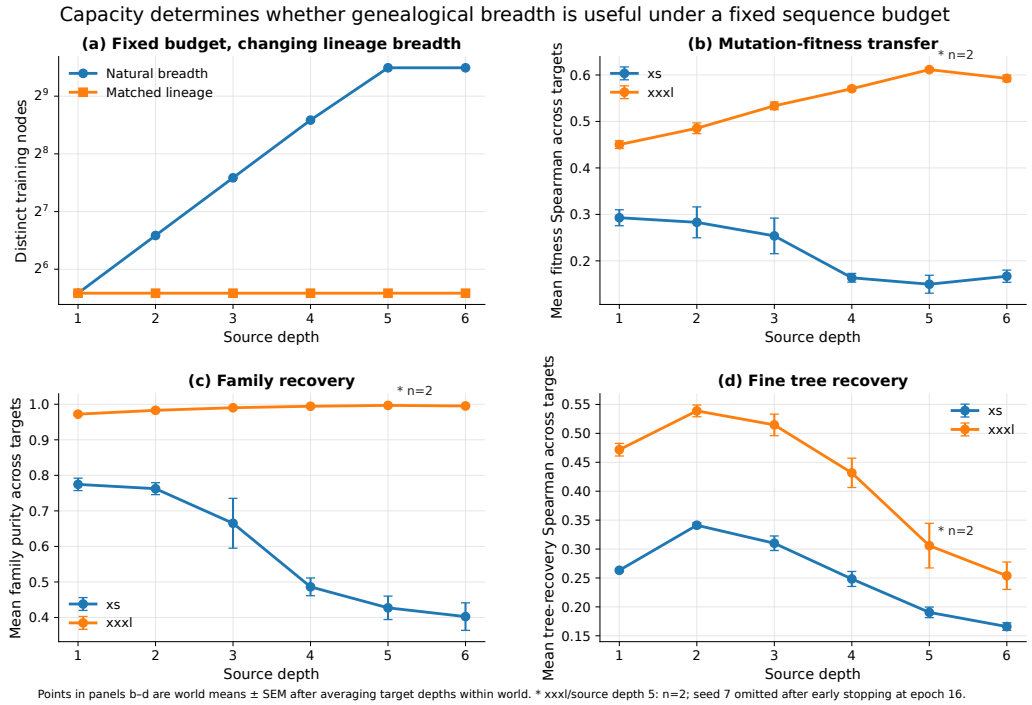


Figure 4: Capacity determines whether genealogical breadth is useful under a fixed sequence budget. **(a)** At a fixed budget of 720 training sequences, natural-breadth sampling draws from increasingly many lineage nodes as source depth increases, reaching 720 nodes at depths 5–6, whereas matched-lineage sampling retains 48 nodes. **(b–d)** Under natural-breadth sampling, mean mutation-fitness Spearman, family purity, and embedding–tree Spearman across source depth for `xs` and `xxxl`. Each point first averages the common target-depth evaluation cohort within a generated world and then averages across worlds; error bars show SEM across worlds. Three worlds contribute to each point except `xxxl` at source depth 5, where the seed-7 optimization failure is excluded and $n = 2$.

Selection-regime mixing. In the selection-regime experiment, a low specialist is trained on 3,200 sequences from one selection regime; the mixed model retains those exact rows and adds 6,400 sequences from the other two regimes. Its total of 9,600 sequences equals the *full* specialist's budget, not that of the low specialist. The mixed-minus-low-specialist fitness-Spearman contrast changes from -0.076 (SEM 0.006) for `xs` to $+0.011$ (0.002) for `l`, $+0.021$ (0.002) for `xl`, and $+0.020$ (0.001) for `xxl`. A second comparison retains all 9,600 rows of the full specialist and adds 19,200 off-regime rows: the mixed-minus-full-specialist contrast changes from -0.092 for `xs` to $+0.011$, $+0.016$, and $+0.008$ for `l`, `xl`, and `xxl`, respectively. These are three-world, paired contrasts; details of row pairing and training are in Appendix C.3.

The broad-lineage comparison holds total training size fixed but changes which lineages contribute sequences; the regime-mixing comparisons retain source-regime examples while increasing total training size. Thus both support a scale-dependent response to heterogeneous training data, but only the former holds total sequence count fixed. Neither contrast decomposes PLM predictions into shared and target-specific components. Model parameter count relative to the raw number of Potts coefficients, as used in the source analysis, is descriptive and does not identify a learning threshold.

A.4 Reliability rankings transfer better than absolute calibration

A ridge model mapping observable diagnostics to prediction risk can transfer its risk ranking from two evolutionary shift mechanisms to a third, especially for lineage and selection shift, but its absolute calibration changes with mechanism. Disagreement among models trained on the same data with different initializations adds little beyond deterministic features and is least informative for a held-out fitness-landscape shift. These results are relevant to deployment but conceptually separate from constraint integration; full protocols are in Appendix C.4.

Table 2: Longer-sequence sensitivity of the E3 decomposition. ρ_{sh} and ρ_{sp} are Spearman correlations of PLM mutation scores with the training-ensemble shared component and target-specific residual, respectively. The final column reports the unnormalized drop-one incremental rank-R^2 contributions of the shared and specific components on shifted targets.

Scale	Near, represented landscape		Shifted landscape		Shifted $\Delta R^2_{\mathrm{sh}}/\Delta R^2_{\mathrm{sp}}$
	ρ_{sh}	ρ_{sp}	ρ_{sh}	ρ_{sp}	
xs	0.063	0.007	0.028	-0.018	0.0030 / 0.0028
l	0.075	0.002	0.045	-0.012	0.0044 / 0.0022
xxxl	0.057	0.003	0.030	-0.000	0.0030 / 0.0018

A.5 Repeated ProteinGym assays and exact pretraining support

Among 23 UniProt proteins with repeated assays of the exact same target sequence, cross-assay DMS concordance predicts mean zero-shot accuracy across methods ($\rho = 0.642$, permutation $p = 0.0012$), while lower concordance predicts larger changes in apparent model performance across assays. Separately, the number of homologous clusters in the exact UniRef50 2021_04 pretraining release predicts ESM2 wild-type likelihood at every scale but is essentially unrelated to the 8M-to-650M DMS-ranking gain ($\rho = -0.002$). These observations distinguish sequence familiarity, assay-specific mutation constraints, and zero-shot functional transfer.

A.6 Longer-sequence sensitivity of shared and target-specific alignment

We repeated the central E3 decomposition at $L = 144$, tripling sequence length relative to the primary $L = 48$ setting while otherwise retaining the same landscape-shift design and training protocol. Because changing sequence length while holding the data and optimization regime fixed also changes the effective dimensionality and difficulty of the learning problem, we treat this experiment as a sensitivity to the generative/training regime rather than as an isolated biological effect of protein length. Branch lengths are unchanged from E3 because mutation proposals scale as $\mathrm{Poisson}(\mu L \ell)$; only burn-in is tripled from 4,000 to 12,000 with sequence length. Each $L = 144$ query contains $144 \times 19 = 2736$ possible single-amino-acid substitutions. We use the same unnormalized drop-one incremental rank-R^2 decomposition as in the main analysis, so that $\Delta R^2_{\mathrm{shared}}$ is the loss in joint-model R^2 when the shared component is removed and $\Delta R^2_{\mathrm{specific}}$ is the corresponding loss when the target-specific residual is removed.

The shared component remains more aligned with model predictions than the target-specific residual even in the near same-landscape cohort, where the correct fitness landscape is represented during training and the query remains close to its training neighborhood. Near-query shared Spearman is 0.063, 0.075, and 0.057 for xs, l, and xxxl, respectively, whereas target-specific Spearman is only 0.007, 0.002, and 0.003. On shifted targets, shared alignment remains positive at all three scales, while specific-residual alignment is near zero or negative; the incremental shared contribution also exceeds the specific contribution at every scale.

This sensitivity therefore offers a less informative setting for analyzing the locality of target-specific alignment. At $L = 144$, near-query specific Spearman is already close to zero (0.007, 0.002, and 0.003 for xs, l, and xxxl), compared with 0.111, 0.081, and 0.075 at $L = 48$, leaving little near–remote contrast to resolve. Remote same-landscape specific Spearman is likewise small (-0.007, -0.008, and -0.009). Thus, while the shared-dominant pattern persists, the $L = 144$ regime provides little measurable target-specific signal with which to assess its locality.

We do not interpret this as evidence that increasing sequence length directly strengthens shared learning. Absolute shared alignment is substantially weaker than in the primary $L = 48$ experiment, and validation selects early checkpoints (epochs 7, 5, and 1 for xs, l, and xxxl). Instead, the sensitivity shows that the shared component is more robustly recoverable across this change in the generative/training regime than the target-specific correction: the latter is detectable locally in the primary calibrated setting but can disappear even under conditions favorable to target-specific recovery.

B Cohort Construction

B.1 Controlled genealogical-distance clades

For each world, every family retains 24 training nodes and four validation nodes. Three held-out target clades are chosen per family. A constrained assignment distributes families across exact target distances $\{2, 4, 6, 8, 10, 12, 14\}$, and the split is accepted only when the realized nearest-training distance equals the requested value for every test node. Each selected node contributes ten local descendants. The manifest stores the nearest phylogenetic training sequence, the globally most similar sequence, the most similar same-family sequence, aligned distances for both identity definitions, candidate support, and node/family identifiers.

B.2 Fixed-budget and paired-lineage cohorts

Every lineage-breadth source task selects exactly 720 training and 96 validation sequences. Natural-breadth sampling round-robins across all source-depth nodes. Matched-lineage sampling selects two nested lineage nodes per family and uses exactly 15 training and two validation sequences per node. Depth-1 cohorts are harmonized and byte-identical across conditions. Common fitness cohorts are chosen independently of source depth and balanced exactly by family and major branch. Paired-lineage evaluation uses two disjoint target descendants from each matched node, producing 96 queries at each depth.

The source-depth trajectories in Fig. 4 show the endpoint scales `xs` and `xxxl`; the six-scale comparison is a separate broad-minus-matched contrast evaluated at source depth 6 and summarized in Appendix A.3.

B.3 Controlled fitness-landscape and query cohorts

The synthetic world and its evolutionary calibration are described in Appendix C.1. From its 128 terminal Potts systems, the controlled splitter selects 32 training, eight validation, and 16 shifted test landscapes, with exactly four test landscapes at each nearest-training tree distance $2, 4, 6, 8$. Every stratum contains four independent parent clades and at least two top-level branches. These landscape memberships are fixed before construction of the query cohorts below.

The common shifted cohort contains 16 sequences per test landscape. The near same-landscape cohort uses independently sampled sequences from 16 training landscapes. The remote same-landscape cohort draws 192 candidates per landscape from long branches and equilibrium samples, then performs one-to-one identity matching to shifted queries. All cohorts are shared across model and regularization conditions, balanced by landscape, and checked for overlap with training. The oracle cache evaluates 256 shifted query backgrounds under 33 systems (the target plus 32 training references), yielding 912 mutation effects per query and 8192 query–reference comparisons. For adaptation, each represented landscape contributes 64 remote unlabeled sequences, 16 disjoint source-validation sequences, and the 16 remote evaluation queries defined above. The wrong-source control is a one-to-one derangement over landscapes; the pooled condition draws from all 15 non-target landscapes. Sequence hashes verify no overlap among training, adaptation, validation, and evaluation sets.

C Detailed Experimental Protocols

This appendix gives implementation details for the landscape-shift, selection-regime, reliability, and ProteinGym analyses reported in the main text.

C.1 Synthetic world generation and calibration

The central synthetic experiments use a common hierarchical evolutionary generator whose parameters were calibrated jointly to create the range of sequence relationships required by the study. The objective was not to assign biological significance to each simulation constant independently. Rather, we sought a regime with substantial diversity within a fitness landscape, graded relatedness

among landscapes, and both near- and remote-homology relationships, while avoiding the degenerate limits in which sampled sequences are nearly identical or effectively unrelated. We chose the settings below to produce such biologically plausible sequence-diversity and homology regimes and then held them fixed across the central landscape-shift interventions. This allows experimental contrasts to change the quantity of interest—for example, landscape diversity or shared constraint—without simultaneously retuning the underlying evolutionary generator.

Each sequence has length $L = 48$ over the standard 20-amino-acid alphabet. A Potts fitness landscape has sitewise fields h_i and pairwise couplings J_{ij},

$$F_C(x) = \sum_i h_{i,x_i} + \sum_{(i,j)\in E} J_{ij}(x_i, x_j).$$

The interaction graph is sampled once with edge probability 0.04 for each pair of sites, corresponding to approximately 45 interacting pairs on average at $L = 48$. Root-landscape fields are drawn independently from $\mathcal{N}(0, 0.15^2)$ and coupling entries from $\mathcal{N}(0, 1^2)$. The sparse graph supplies epistatic structure without directly coupling every pair of residues.

Related fitness landscapes are generated along a balanced depth-7 binary landscape tree. If d denotes the drift magnitude on an edge, child parameters are formed as

$$h' = \sqrt{1 - d^2}\, h + d\, \epsilon_h, \qquad J' = \sqrt{1 - d^2}\, J + d\, \epsilon_J,$$

where ϵ_h and ϵ_J are independent draws at the same field and coupling scales as the root landscape. The nominal per-edge drift is 0.18; individual edge drifts are lognormally varied with log-standard deviation 0.7 and clipped to $[0.04, 0.50]$. The interaction graph itself is held fixed (`edge_rewire_prob`$= 0$). This construction produces landscapes that share a common interaction topology while their preferred residues and interaction strengths diverge hierarchically. Holding the graph fixed is deliberate: it makes the principal source of cross-landscape heterogeneity the values of the constraints rather than simultaneous changes in which residue pairs interact, yielding a cleaner shared-versus-specific comparison.

Sequences evolve under Metropolis point-mutation dynamics. A proposed mutation is accepted according to

$$P_{\text{accept}} = \min\{1, \exp(\beta \Delta F)\},$$

with $\beta = 2.0$. The proposal-rate parameter is $\mu = 1.1$: along a branch of length b, the number of proposed point mutations is drawn from

$$N \sim \text{Poisson}(\mu L b).$$

Thus μ and branch length jointly control sequence divergence, whereas β together with the Potts field and coupling scales controls how strongly fitness constrains the accepted trajectories. These quantities were calibrated jointly rather than interpreted in isolation.

Within each landscape, sequences evolve along an independent balanced depth-4 sequence tree. Its branches have length 4.0. The root sequence is initialized by 4,000 Metropolis burn-in proposals under the corresponding Potts landscape. We sample at the 16 depth-4 terminal nodes and generate ten additional descendants per terminal node using local branch length 1.0, yielding 160 sequences per landscape. The depth and branch-length choices provide both shared ancestry within a landscape and appreciable sequence divergence; the additional local descendants provide multiple related observations around each terminal lineage rather than independent equilibrium samples.

The depth-7 landscape tree supplies 128 terminal Potts systems. For the central landscape-shift experiment, the controlled splitter selects 32 landscapes for training, eight for validation, and 16 as shifted test landscapes. The test set contains exactly four landscapes whose nearest training landscape is at each tree distance in $2, 4, 6, 8$, with multiple independent parent clades and top-level branches represented within each distance stratum. Because each training landscape contributes 160

Table 3: Central synthetic-world generation parameters. Parameters are calibrated jointly to produce the sequence-diversity and homology regime used by the controlled experiments; individual numerical values are not intended as independently estimated biological constants.

Component	Parameter	Value
Sequence	Length L	48
	Alphabet size	20 amino acids
Potts landscape	Interaction probability	0.04
	Field scale	0.15
	Coupling scale	1.0
	Selection strength β	2.0
Landscape hierarchy	Tree depth	7
	Terminal landscapes	128
	Nominal drift per edge	0.18
	Drift log-SD	0.7
	Drift range	$[0.04, 0.50]$
	Edge-rewiring probability	0
Sequence evolution	Proposal rate μ	1.1
	Sequence-tree depth	4
	Tree branch length	4.0
	Sampled tree depth	4
	Descendants per terminal node	10
	Local descendant branch length	1.0
	Root burn-in proposals	4,000

sequences, the 32-landscape training set contains 5,120 sequences. These landscape splits are held fixed when comparing model scales and regularization conditions.

Several parameter choices also serve experimental-control purposes. Sparse but nonzero Potts couplings ensure that mutation effects contain epistatic as well as sitewise structure. Hierarchical parameter drift creates related but nonidentical fitness functions, allowing shared and landscape-specific mutation effects to coexist. The absence of interaction-edge rewiring isolates changes in constraint values from changes in interaction topology. Finally, the combination of selection strength, proposal intensity, and sequence-tree branch lengths was chosen to span useful near- and remote-homology regimes while retaining measurable landscape-specific signal. The central conclusions therefore concern behavior within this calibrated regime; complementary experiments vary genealogical distance, lineage breadth, selection regime, and training-landscape composition rather than treating the particular numerical generator settings as universal biological constants.

C.2 Fitness-landscape oracle, prediction decomposition, and adaptation controls

The fitness-landscape analysis proceeds through five levels of evidence. First, identity-matched controls compare shifted queries with remote queries generated under represented training landscapes. Second, the model-independent oracle analysis asks whether the Potts systems actually induce different mutation tasks. Third, PLM prediction decomposition asks which component of those tasks is encoded by learned mutation rankings. Fourth, near/remote same-landscape decomposition locates the loss of target-specific alignment under sequence divergence. Fifth, adaptation and source-specificity controls test whether unlabeled correct-landscape sequences can restore the lost signal and whether their benefit exceeds identity-matched wrong-source continuation. Oracle rank correlation, sign agreement, beneficial-set overlap, normalized root-mean-square error (RMSE), incremental rank-R^2, and the preregistered adaptation partial rank-R^2 are reported separately rather than collapsed into a composite.

For each query background x, let $\boldsymbol{y}(x)$ be the vector of PLM scores, $\boldsymbol{\Delta}_{\mathrm{target}}(x)$ the exact target-landscape effects, and

$$\boldsymbol{g}(x) = \frac{1}{K}\sum_{j=1}^{K} \boldsymbol{\Delta}_{C_j}(x) \tag{1}$$

the mean effect vector across the K training landscapes. We define the target-specific residual in mutation-effect score space as

$$\boldsymbol{r}(x) = \boldsymbol{\Delta}_{\text{target}}(x) - \boldsymbol{g}(x). \tag{2}$$

The PLM, shared, and residual vectors are then separately rank-standardized. For each query we fit three ordinary least-squares regressions predicting the PLM score ranks: one from the shared predictor alone, one from the specific residual alone, and one from both predictors jointly. These yield R^2_{shared}, R^2_{specific}, and R^2_{combined}. The central synthetic–natural comparison uses the unnormalized drop-one contributions

$$\Delta R^2_{\text{specific}} = R^2_{\text{combined}} - R^2_{\text{shared}}, \qquad \Delta R^2_{\text{shared}} = R^2_{\text{combined}} - R^2_{\text{specific}}. \tag{3}$$

We call these quantities incremental rank-R^2: each measures the additional variation in PLM mutation rankings explained by one component beyond the other. Landscape-level bootstrap intervals resample the 16 test systems. All results use validation-selected checkpoints and identical query backgrounds across model scales.

For adaptation, let A denote the specified alignment metric. The primary prediction-level gain is

$$\Delta A_{k,m} = A_{k,m} - A_{0,\text{remote}}, \tag{4}$$

where m indexes the adaptation method and k the number of unlabeled sequences. We report specific-residual, complete-target, and shared-oracle versions of A, plus the loss of performance on the original source after adaptation. To preserve the preregistered adaptation endpoint, these analyses additionally report normalized partial rank-R^2,

$$P^2_{\text{specific|shared}} = \frac{R^2_{\text{combined}} - R^2_{\text{shared}}}{1 - R^2_{\text{shared}}}. \tag{5}$$

This is the fraction of rank variation remaining after the shared-only fit that is explained by adding the specific residual; it is not numerically compared with the incremental rank-R^2 used in the central decomposition. The source-specificity experiment pairs conditions within target landscape and specifies

$$\Delta P^2_{\text{source}} = P^2_{\text{specific|shared, correct}} - P^2_{\text{specific|shared, wrong}} \tag{6}$$

as the primary endpoint. Raw residual and target-oracle Spearman are secondary because improved shared alignment can increase both without isolating the target-specific component.

C.3 Selection-regime replication and convergence sensitivity

Each selection-regime world contains 24 family constraint systems sampled under $\beta \in \{1.5, 3, 6\}$. Training membership is family balanced. Full specialists use 400 selected sequences per family (9600 total) from one regime; low specialists use approximately 133–134 per family (3200 total). The equal-total mixed condition concatenates the exact low-specialist rows from all three regimes for 9600 total sequences, while the full-per-regime mixed condition concatenates the exact full-specialist rows for 28,800 total sequences. Specialist validation uses 2880 sequences from its source regime (120 per family); mixed validation uses 960 per regime (40 per family per regime), for the same total validation budget. Membership hashes verify equality of the paired row sets in every world, scale, and regime.

The target evaluation cohorts are generated independently of training selection and shared across every source condition within a world. For specialist transfer, a source-β model is evaluated on all three target regimes. For budget contrasts, query-level predictions are paired across the mixed and corresponding specialist condition. We first compute each target sequence's Spearman correlation between the PLM single-mutant log-likelihood ratio (LLR) and true Potts ΔF, then form target- and family-balanced contrasts. The identity-adjusted model includes query fixed effects and a spline in maximum same-family training identity; family bootstrap resampling supplies within-world intervals. Across-world inference never pools queries as independent observations: the seed/world contrast is the unit used to compute the reported mean, standard error of the mean (SEM), sign consistency, and descriptive Student-t interval.

The selection-regime analysis uses three independently generated worlds. Each world independently generates the constraint banks, sequence samples, split memberships, and model initializations. The

capacity comparison spans `xs`, `m`, `l`, `xl`, and `xxl`; the added `m` scale uses learning rate 10^{-4} and `xl`/`xxl` use 5×10^{-5}. All models use the same 1000-epoch ceiling and patience of 50. Provenance checks verify independent constraint banks, exact mixed/specialist row pairing, paired initialization within each world and scale, and complete task execution. The capacity interaction has the same sign in all three worlds for both primary contrasts.

For the convergence sensitivity analysis, `xs` models that reach the 1000-epoch ceiling are resumed from their stateful checkpoints. Twenty-one of 24 `xs` models meet this criterion; the other three had already stopped by patience. The continuation raises only the epoch ceiling to 1500, keeping optimizer state, random-number-generator state, patience, learning rate, data order, and all other hyperparameters fixed. The equal-total capacity interaction changes from $+0.0871$ to $+0.0957$, and the full-per-regime interaction from $+0.1030$ to $+0.0978$, supporting the interpretation that mixed-regime interference in `xs` is not merely a transient consequence of slower convergence.

C.4 Cross-shift reliability and ensemble disagreement

The cross-shift reliability analysis combines target-level outputs from the genealogical-distance, fitness-landscape, and selection-regime experiments. After harmonizing to `xs` and `l`, it contains 18,816 targets from 26 model instances. The genealogical-distance component contributes 2,160 held-out-lineage targets per scale across three worlds. The fitness-landscape component contributes 768 targets per scale: 256 shifted, 256 near same-landscape, and 256 identity-matched remote same-landscape queries, organized into 16 landscape clusters per cohort. The selection-regime component contributes 6,480 full-specialist transfer targets per scale across nine source-regime–world combinations. Mixed-training contrasts are excluded so this component isolates selection-regime transfer rather than the capacity intervention.

For every target, the deployable signals are sequence distance $1 - \max \mathrm{id}$, nearest embedding distance, random-neighbor fraction (RNF; the fraction of a query's ten nearest reference-or-random embeddings that are random decoys), predictive entropy, negative wild-type pseudo-log-likelihood (PLL), negative single-mutant contrast range, and negative contrast standard deviation. Privileged exact tree distance, landscape distance, Potts-parameter distance, and $|\Delta\beta|$ are saved only for diagnostics and never enter deployable models. The primary reliability model uses the first four deployable signals. Each is percentile-ranked within model instance without using fitness. A fixed ridge penalty $\alpha = 1$ is used, and source rows are weighted so the two training mechanisms contribute equally. The leave-one-shift-family-out fold is therefore defined before examining the held-out mechanism.

Risk–coverage curves sort targets by predicted risk. At coverage c, observed risk is mean $1 - \rho_{\mathrm{fitness}}$ among the retained lowest-risk fraction. Area under the risk–coverage curve (AURC) is compared with oracle and random orderings; worst-quintile area under the receiver-operating-characteristic curve (AUROC) and calibration mean absolute error (MAE) are secondary. Because a reliability score can rank examples correctly while misestimating absolute risk, calibration and mean bias are reported separately.

The ensemble-disagreement analysis holds the data fixed within each shift mechanism. One generated world is used for each of genealogical, landscape, and selection shift, and both `xs` and `l` use five independently initialized models trained on identical training and validation memberships. Each mechanism–scale pair uses 192 targets: genealogical targets are balanced across tree distance; landscape targets comprise 64 shifted, 64 same-landscape, and 64 identity-matched queries; and selection targets comprise 64 queries at each of $\beta = 1.5, 3, 6$. Verification confirms identical data membership and target/oracle vectors across ensemble members, distinct initialization states, and reproduction of the reference-model fitness scores.

For target x, the primary ensemble disagreement is

$$U_{\mathrm{ens}}(x) = 1 - \frac{2}{M(M-1)} \sum_{i<j} \rho_S(\mathbf{s}_i(x), \mathbf{s}_j(x)) \,, \tag{7}$$

where $M = 5$ and $\mathbf{s}_i(x)$ is member i's complete single-mutant score vector. Standardized score dispersion and top-20-set disagreement are secondary. The ensemble endpoint is the increment in held-out-shift risk Spearman after adding U_{ens} to the deterministic multivariable feature set. Two thousand target-cluster bootstrap replicates give descriptive intervals.

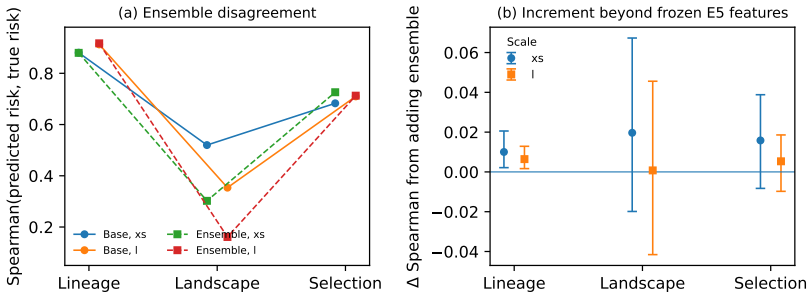


Figure 5: Same-data ensemble disagreement is mechanism dependent. (A) On a matched 192-target subset per shift mechanism and scale, ensemble disagreement alone ranks genealogical and selection risk well but is weak for latent landscape shift. The deterministic multivariable risk model is shown for comparison. (B) Adding ensemble disagreement yields small positive increments for genealogical shift, but landscape and selection intervals overlap zero. Error bars are descriptive target-cluster bootstrap 95% intervals.

C.5 ProteinGym: exact-sequence landscape validation

The natural-data analysis uses the ProteinGym substitution DMS benchmark and its released zero-shot score matrices (Notin et al., 2023). All 217 local substitution assays have matched DMS files, MSA metadata, and zero-shot score files. Twenty-four UniProt proteins have repeated DMS assays; after requiring exact equality of the target amino-acid sequence and at least 20 overlapping single substitutions, 35 assay pairs from 23 proteins remain for the formal analysis. DMS scores use the ProteinGym direction convention in which larger values indicate better phenotype. Pairwise landscape concordance is Spearman correlation on the mutation intersection.

The primary endpoint first computes each method's mean DMS Spearman across the two assays, averages the 88 available zero-shot methods within assay pair, then averages repeated pairs within UniProt. The resulting 23 protein-level observations are correlated with the corresponding mean DMS concordance. Twenty thousand bootstrap replicates resample UniProt proteins, and 50,000 permutations shuffle the protein-level outcome labels. A median-across-method sensitivity prevents a small number of extreme methods from determining the result. Model-specific analyses preserve the same 23-protein aggregation and are reported only as a consistency check; all 88 method-specific concordance–performance correlations are positive.

The cross-context performance-gap endpoint computes the absolute difference between a method's DMS Spearman in the two assays, then follows the same method, pair, and UniProt aggregation. Its negative association with DMS concordance shows that zero-shot benchmark performance becomes more assay dependent as two landscapes disagree. Using the verified 23 UniProt-level coordinates, the primary concordance–accuracy association is $\rho = 0.6423$ (95% UniProt-bootstrap CI [0.2296, 0.8465]; two-sided 50,000-permutation $p = 0.00128$). The concordance–absolute-performance-gap association is $\rho = -0.5119$ (95% CI [-0.7793, -0.1061]; permutation $p = 0.0122$). These relationships are shown in Fig. 6.

For the leave-one-assay-out analysis, exact-sequence assay groups with sufficient repeated measurements are considered independently. A mutation enters the reference consensus only if it is measured in the held-out assay and at least two other assays; the held-out scores never enter the consensus. Four UniProt proteins produce 13 held-out assays and 1118 held-out-assay–method rows over 86 methods. Across methods, mean alignment to the independent consensus is 0.466, while mean partial alignment to held-out fitness after conditioning on the consensus is 0.147 (Fig. 7). These quantities are descriptive because the number of proteins is small.

C.6 ProteinGym: identity-matched clade-Potts decomposition

Outcome-blind feasibility scan and fixed cohorts. We begin from 173 ProteinGym targets with an MSA and at least 80 valid single substitutions in a 96-residue window. Eligible MSA crops must use canonical amino acids and satisfy the ProteinGym MSA coverage criterion used throughout the natural-data analyses. Sequences are clustered, and clusters with at least 100 unique rows are can-

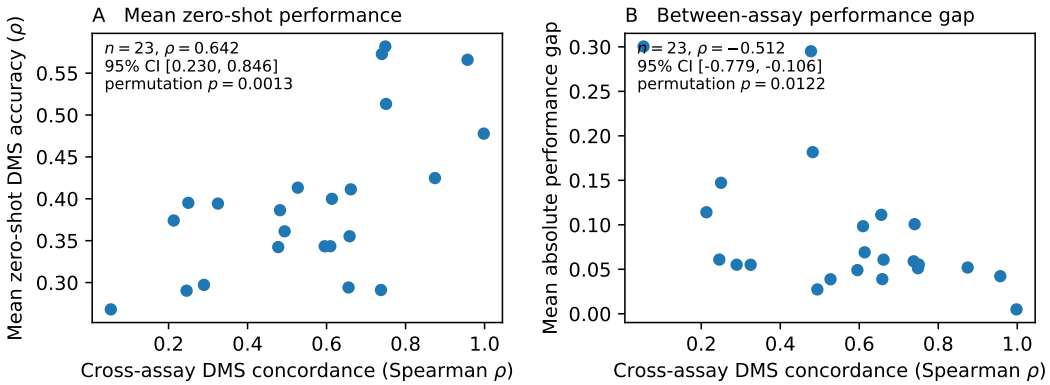


Figure 6: Repeated assays reveal assay-dependent zero-shot difficulty at fixed sequence. Each point is one UniProt protein after averaging usable repeated ProteinGym assay pairs with exactly identical target sequence. **A,** Greater cross-assay DMS concordance is associated with higher mean zero-shot mutation-ranking accuracy after averaging the available zero-shot methods within assay pair. **B,** Lower concordance is associated with larger differences in apparent zero-shot performance across the paired assays. Statistics use the 23 UniProt-level aggregates; intervals are 95% UniProt-bootstrap confidence intervals and p-values are two-sided protein-label permutation tests. These comparisons quantify assay dependence at fixed sequence and are not interpreted as a pure biological fitness-landscape intervention.

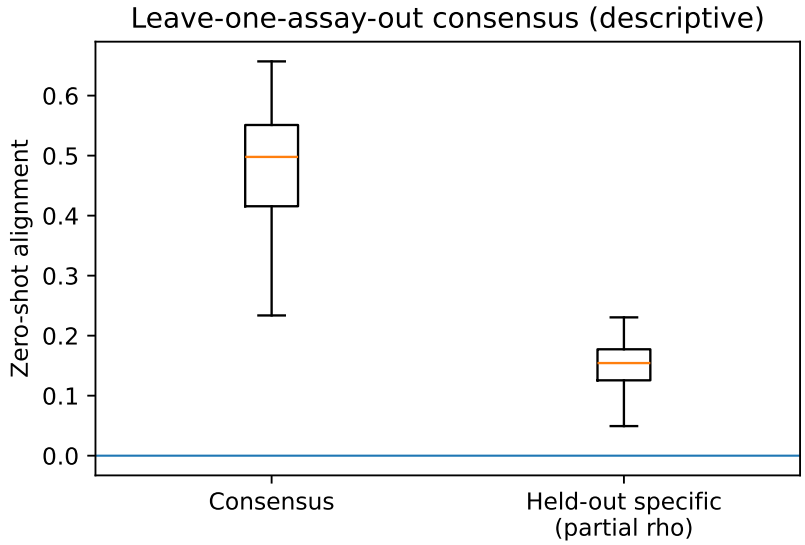


Figure 7: Leave-one-assay-out natural shared-versus-specific analysis. For proteins with at least three suitable exact-sequence DMS assays, each held-out assay is predicted relative to a mutation consensus constructed only from the remaining assays. Across four proteins and 86 zero-shot methods, alignment with the independent consensus is substantially larger than held-out-specific partial alignment. This analysis is mechanistic corroboration rather than population-level inference because only four proteins qualify.

didates. The operational target clade is the viable cluster with the highest target identity. For each target we search up to 30 comparison clades and all comparison-clade pairs without using DMS values or ESM2 predictions. Matching is one-dimensional optimal transport on target identity over the common support of all three clades. Retained triplets must have maximum identity spread ≤ 0.02; each clade must retain at least 100 matched rows; the maximum difference in clade mean identity must be ≤ 0.015; and the maximum pairwise KS distance must be ≤ 0.12. Only five targets pass all criteria: ENVZ_ECOLI, OTC_HUMAN, CCR5_HUMAN, ARGR_ECOLI, and VG08_BPP22. Median matched depth is 295 rows/clade, median maximum mean-identity gap is 0.0036, and median maximum KS is 0.102. A prespecified sensitivity analysis adds ODP2_GEOSE, DOCK1_MOUSE,

Table 4: Observed and reliability-corrected incremental shared-versus-specific signal in the strict five-target cohort. Values are target means. Errors-in-variables correction uses independent split-half Potts fits and is a sensitivity analysis for differential predictor reliability.

Endpoint	Observed ΔR^2 shared	Observed ΔR^2 specific	Corrected ΔR^2 shared	Corrected ΔR^2 specific
DMS	0.076	0.010	0.047	0.011
ESM2-8M	0.171	0.036	0.112	0.041
ESM2-35M	0.263	0.053	0.133	0.061
ESM2-150M	0.315	0.055	0.159	0.053
ESM2-650M	0.257	0.054	0.137	0.053
ESM2-3B	0.248	0.046	0.125	0.039
ESM2-15B	0.221	0.034	0.097	0.037

SCIN_STAAR, and MBD11_ARATH under the identical design with only the KS threshold relaxed to 0.15. These four proteins are never included in the primary five-target estimate.

Clade-specific Potts models and decomposition. For each selected clade we encode the aligned 96-residue window, apply 80% sequence-identity reweighting, and fit one multinomial logistic regression per site using all other aligned positions as predictors (L_2 penalty, $C = 0.2$, L-BFGS, maximum 250 iterations). The local mutant score is the fitted mutant-minus-wild-type logit contrast. Within each clade, the complete single-mutant score vector is transformed to rank z scores. For target clade T and comparison clades A, B, the shared vector is $(A + B)/2$ in rank space. The specific vector is the residual from linearly regressing T on this shared vector. Rank-space regression of an endpoint on shared and specific vectors yields the full R^2 and the unique contribution of each component by dropping that component from the two-predictor model. Target-clade advantage is $\rho(y, T) - [\rho(y, A) + \rho(y, B)]/2$. Endpoints are ProteinGym DMS effects and released ESM2 masked-marginal scores at 8M, 35M, 150M, 650M, 3B, and 15B.

Split-half reliability and errors-in-variables sensitivity. Matched triplets are deterministically divided into two disjoint halves while preserving the identity coupling. Each of the three clade Potts models is then refit independently in each half. In the primary cohort, cross-half Spearman is 0.891 for the shared vector, 0.334 for the specific residual, and approximately 0.83 for the full target component; in the nine-target sensitivity cohort the corresponding shared/specific values are 0.878/0.368. We therefore treat the two half fits as replicate noisy measurements of a common latent shared and specific predictor. Cross-half covariance estimates the latent predictor covariance, while DMS and ESM2 outcomes are left uncorrected. The resulting errors-in-variables estimates are summarized in Table 4. Shared remains larger than specific for DMS and every ESM2 scale after correction. In the nine-target sensitivity cohort, corrected ESM2 shared incremental R^2 ranges from 0.105 to 0.212, versus 0.030 to 0.048 for specific; corrected DMS is 0.068 versus 0.021. We use this only to test whether unequal Potts reliability explains the main result, not as a replacement for the observed prospective endpoints.

C.7 ProteinGym: historical-pretraining pooled Potts estimators

The clade-Potts experiment deliberately gives the two comparison clades equal weight. To ask whether ESM2 instead reflects the composition of homologous evidence present during pretraining, we use the exact UniRef50 2021_04 representative corpus and retain pairwise alignments from the target searches. Historical hits must have at least 20% identity and 80% query-window coverage; aligned residues are projected into the same 96-residue target coordinates used by the clade-Potts fits. Six prespecified targets retain at least 80 usable historical homologs: ENVZ_ECOLI, OTC_HUMAN, CCR5_HUMAN, ODP2_GEOSE, DOCK1_MOUSE, and MBD11_ARATH. Their unrestricted historical pools contain 170–1158 projected representatives (median 620). When a pool exceeds 1200 rows, a deterministic target-identity-quantile sample caps the fit at 1200. All pooled fits use the same pseudolikelihood estimator, regularization, and 80% sequence-identity reweighting as the clade models.

Table 5: Historical-pretraining pooled-Potts comparisons. The first two columns are six-target mean Spearman differences. The last two are strict historical-minus-current matched-pool differences for the only two proteins satisfying the fixed count/identity criteria. Positive values favor the historical-pretraining composition. The strict controls are descriptive case studies, not a two-protein inferential cohort.

Endpoint	Hist.−target ($n = 6$)	Exposure-wtd.−equal ($n = 6$)	OTC hist.−current	DOCK hist.−current
ESM2-8M	-0.022	$+0.027$	-0.053	-0.022
ESM2-35M	-0.012	$+0.037$	-0.085	$+0.020$
ESM2-150M	$+0.010$	$+0.040$	-0.058	$+0.000$
ESM2-650M	$+0.072$	$+0.048$	$+0.034$	$+0.040$
ESM2-3B	$+0.101$	$+0.030$	$+0.013$	$+0.079$
ESM2-15B	$+0.092$	$+0.029$	$+0.027$	$+0.009$
DMS	$+0.009$	$+0.015$	$+0.013$	$+0.064$

The *historical-exposure-weighted* estimator does not refit a Potts model. Instead, each historical homolog is assigned to the nearest of the three operational clade consensuses in projected sequence identity; the fractions assigned to target, comparison 1, and comparison 2 weight the corresponding rank-standardized clade Potts vectors. This assignment is intentionally operational: for some targets many historical homologs have small margins between their two closest clade consensuses, so the weights should be interpreted as a coarse summary of pretraining composition rather than exact lineage membership.

The direct historical pool is compared with the target-clade and equal-comparison-clade estimators across the six ESM2 scales and DMS. Across the six available proteins, historical-pool minus target-clade mean Spearman is $-0.022, -0.012, +0.010, +0.072, +0.101,$ and $+0.092$ for 8M, 35M, 150M, 650M, 3B, and 15B; the corresponding DMS difference is $+0.009$. Historical-exposure weighting improves over the equal comparison-clade estimator by $+0.027, +0.037, +0.040, +0.048, +0.030,$ and $+0.029$ across ESM2 scales and by $+0.015$ for DMS. These six-target contrasts are observational because historical pools can differ from the controlled clade sets in both identity distribution and depth.

We therefore construct strict historical/current pooled controls only where the two sources have adequate common target-identity support. OTC_HUMAN uses 486 sequences per source with identical 2.5-percentage-point target-identity-bin counts (mean-identity gap 0.07 percentage points; KS 0.045). DOCK1_MOUSE initially falls just outside the KS threshold despite ample overlap; an outcome-blind paired-subset optimization, using identity information only and never ESM2 or DMS outcomes, selects the largest valid subset under the unchanged $n \geq 80$, mean-gap ≤ 1.5 percentage points, and KS ≤ 0.12 rules. This yields 117 historical/current pairs at a 0.75-percentage-point pair caliper, mean-identity gap 0.02 points, and KS 0.1197. Standard 80%-identity redundancy weights are applied to both members of each historical/current comparison and total fit weight is equalized within protein.

The strict controls reproduce the qualitative scale crossover seen in the unrestricted six-target analysis. Both OTC and DOCK favor the matched contemporary pool at 8M, whereas both favor the historical pool at 650M, 3B, and 15B; intermediate scales are mixed. Historical pools also modestly improve DMS alignment in both proteins, so the large-model historical preference is not uniquely identifiable as pretraining memorization. Taken together, these analyses support an exposure-weighted integration interpretation: the dominant ESM2 constraint component increasingly resembles the homolog mixture historically available during pretraining, while the data do not imply that ESM2 literally implements one pooled Potts model.

C.8 ProteinGym: ESM2 representation and inferred phylogenetic geometry

The representation analysis uses the fixed nine-target cohort and 40 matched sequences from each operational clade (120 support sequences per protein). The same sequences are embedded with ESM2-8M, 150M, 650M, and 3B; each embedding averages final-layer tokens over the 96-residue crop. For k-nearest-neighbor clade purity, we measure the fraction of an embedding's $k \in \{1, 5, 10\}$ neighbors with the same label. We also report leave-one-out nearest-centroid accuracy, the fraction of the target query's nearest support sequences drawn from the target clade, and its cosine-similarity

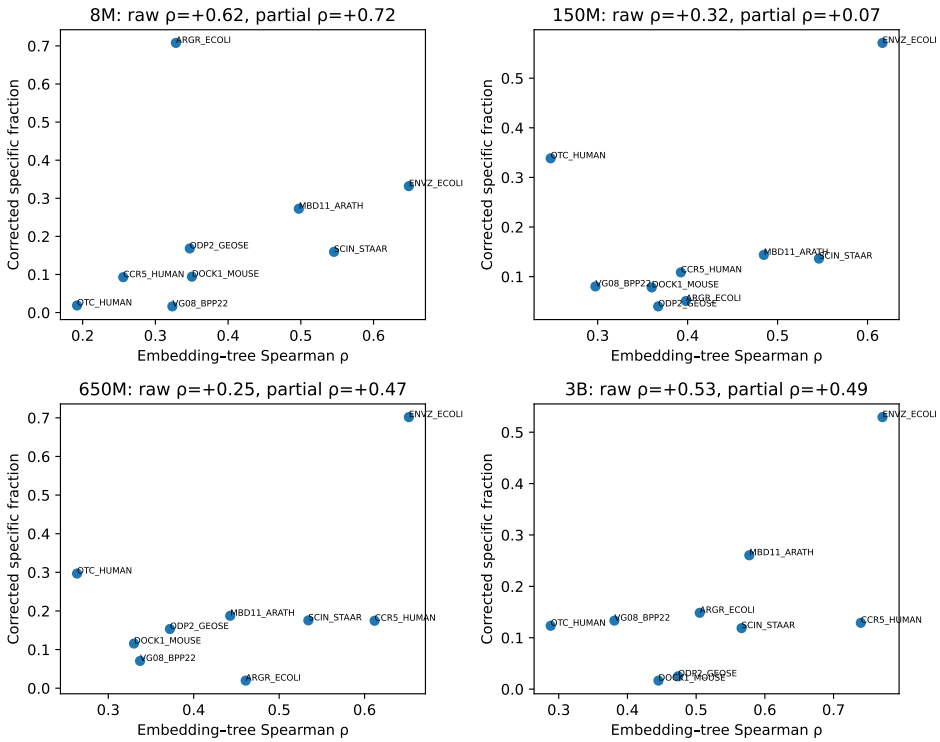


Figure 8: Exploratory representation–specificity analysis. For the fixed nine-target cohort, ESM2 preservation of inferred phylogenetic geometry is compared with the reliability-corrected fraction of mutation signal attributed to the target-clade-specific component. The positive average association remains after controlling ordinary sequence-distance fidelity and matched target identity, but it is not uniform across model scales and is sensitive to individual families. Clade purity and centroid classification show no comparable relationship.

margin between the target and closest comparison centroid. Geometry metrics correlate pairwise embedding distance with aligned p-distance (the proportion of mismatched residues) or with leaf-to-leaf path length in a neighbor-joining tree inferred from those sequence distances. This tree is a graded sequence-derived summary, not known genealogy.

We define the reliability-corrected *specific fraction* as corrected incremental specific R^2 divided by the sum of the corrected incremental shared and specific contributions. Simple clade separability shows no detectable association with this fraction, but the small cohort does not support an equivalence claim. Averaged across scales, 10-nearest-neighbor purity has correlation $\rho = 0.00$ with specific fraction; leave-one-out centroid accuracy gives $\rho = -0.13$, target-clade retrieval -0.03, and matched target identity 0.00. With only nine proteins, the approximate correlation magnitude required for 80% power at two-sided $\alpha = 0.05$ is $|\rho| \simeq 0.816$, so moderate associations remain compatible with these estimates. Embedding–tree fidelity is more suggestive: its raw target-level association with specific fraction is $\rho = 0.58$. After rank-residualizing both variables on embedding–sequence-distance fidelity and matched mean target identity, the partial association is $\rho = 0.738$ with exact target-label permutation $p = 0.019$. Per-scale partial associations are 0.722 (8M), 0.068 (150M), 0.467 (650M), and 0.492 (3B). Leave-one-target-out values for the across-scale average range from 0.144 to 0.904; removing ENVZ increases the association to 0.777, whereas removing ARGR reduces it to 0.144. The association is therefore exploratory and family sensitive. It is also specific to the relative specific-fraction endpoint: controlled associations with shared-minus-specific R^2 and target-clade advantage are near zero.

C.9 PROTEINGYM: TESTING DISCRETE CONTEXT SELECTION AS A FITNESS-ERROR MECHANISM

A natural alternative to the shared-constraint account is that ESM2 selects one cluster-specific mutation profile for each target or site and makes large errors when it selects a comparison cluster. We test this hypothesis progressively so that proximity in one-dimensional score space is not mistaken for agreement on a complete mutation profile.

Mutation-level proximity is suggestive but not diagnostic. Across 58,050 mutation–model observations from the fixed nine-target cohort, we rank-standardize ESM2, DMS, target-cluster, shared, and the two comparison-cluster Potts scores within protein. Among mutations whose DMS value is itself closest to the target-cluster score, high-error ESM2 predictions are often closest to a comparison-cluster score: this fraction ranges from 0.750 at 8M to 0.630 at 15B. A large scalar error, however, necessarily moves a prediction away from the target value and makes one of several alternative references easier to approach. Mutation-level proximity alone therefore cannot identify which context governs the prediction.

Full substitution profiles remove most apparent comparison-cluster correspondence. We next group mutations by residue and retain sites with at least eight assayed substitutions. OTC_HUMAN has no qualifying site, leaving eight proteins and 2,844 site–model observations across six ESM2 scales. At each site, we correlate the complete ESM2 substitution profile with the DMS, target-cluster, shared, and two comparison-cluster Potts profiles. The primary subset contains *DMS-target-like sites*, where the DMS profile is most similar to the target-cluster profile. Apparent comparison-cluster enrichment among high-error sites is $+0.169$, $+0.168$, and $+0.123$ at 8M, 35M, and 150M, and $+0.052$, $+0.082$, and $+0.077$ at 650M, 3B, and 15B. Yet a within-protein permutation that breaks same-site correspondence while preserving ESM2/DMS error, target/shared profiles, and the marginal distribution of comparison profiles explains most of this pattern. Median per-target permutation p values range from 0.23 to 0.55, with only 0–2 of eight proteins nominally significant at any scale. High error can therefore create apparent resemblance to an alternative profile without coherent use of that profile at the same site.

The context decoded locally does not predict the expressed comparison profile. For the same eight proteins, we mask the focal target residue and the corresponding residue in identity-matched support sequences, then extract local hidden states from ESM2-8M through ESM2-3B. A nearest-centroid classifier assigns each target site to the target cluster or one of two comparison clusters. Support-sequence leave-one-out accuracy rises from 0.658 at 8M to 0.750 at 3B, confirming that local cluster information is decodable. Nevertheless, 50–61% of DMS-target-like sites are assigned to a comparison cluster without the ESM2 mutation profile preferentially following that same cluster. We quantify agreement with the *attribution-aligned profile margin*,

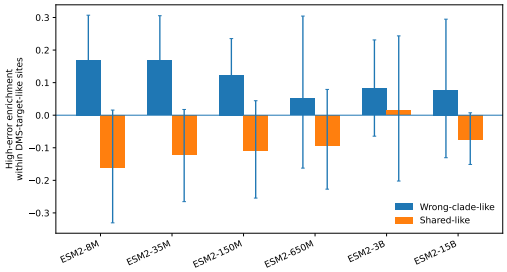
$$\rho_S(E_i, P_{\text{assigned comparison},i}) - \rho_S(E_i, P_{\text{other comparison},i}).$$

The margin is close to zero at every scale (Table 6). We define DMS-profile error as one minus the sitewise Spearman correlation between the ESM2 and DMS substitution profiles. After averaging scales within protein, comparison-cluster attribution is associated with a $+0.043$ error difference (95% target-bootstrap CI $[0.004, 0.079]$, $n = 6$), but the cross-scale attribution-aligned margin is only $+0.007$ (95% CI $[-0.008, 0.022]$, $n = 7$; $\text{MDE}_{80} = 0.027$). Thus local attribution identifies harder sites without specifying which alternative fitness profile the prediction follows. Restricting to sites with support-sequence centroid accuracy of at least 0.60 gives the same qualitative conclusion; an isolated positive 650M margin is supported by only three targets.

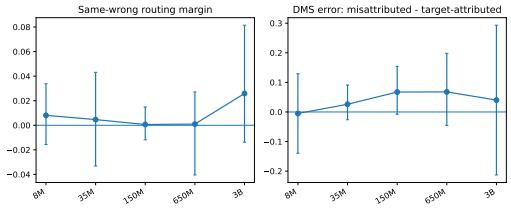
Detectability analysis. The cross-scale target-average attribution-aligned margin is the most sensitive endpoint in this sequence: $+0.0071$ with 95% target-bootstrap CI $[-0.0078, 0.0216]$ and $\text{MDE}_{80} = 0.0268$. The site-profile controls are less sensitive. Comparison-cluster enrichment above the within-protein permutation null averages -0.019 (95% CI $[-0.196, 0.135]$, $\text{MDE}_{80} = 0.296$), and the continuous comparison-margin-above-null endpoint averages -0.100 (95% CI $[-0.216, 0.008]$, $\text{MDE}_{80} = 0.198$). These controls provide convergent evidence against a large discrete context-selection effect but do not exclude small or distributed forms of contextual conditioning.

Table 6: Local cluster attribution does not determine the expressed comparison-cluster profile. Target-balanced results on DMS-target-like sites with sufficient mutation coverage. Assignment to a comparison cluster is common, but the mutation profile does not preferentially follow the same cluster selected from the local hidden state. Averaged across scales within protein, comparison-cluster attribution is associated with a modest error increase ($+0.043$, 95% CI $[0.004, 0.079]$), while the attribution-aligned profile margin remains near zero ($+0.007$, 95% CI $[-0.008, 0.022]$).

ESM2 scale	Comparison attribution fraction	Attribution-aligned margin	DMS-error increase
8M	0.611	$+0.008$ $[-0.016, +0.034]$	-0.005 $[-0.140, +0.129]$
35M	0.534	$+0.005$ $[-0.033, +0.043]$	$+0.026$ $[-0.026, +0.091]$
150M	0.574	$+0.001$ $[-0.012, +0.015]$	$+0.067$ $[-0.008, +0.154]$
650M	0.502	$+0.001$ $[-0.040, +0.027]$	$+0.068$ $[-0.045, +0.198]$
3B	0.526	$+0.026$ $[-0.014, +0.081]$	$+0.040$ $[-0.213, +0.293]$



(a) Comparison-profile resemblance and the site-permutation null.



(b) Local cluster attribution, profile agreement, and DMS error.

Figure 9: Testing discrete context selection as a natural ESM2 failure mechanism. (A) Apparent resemblance to a comparison-cluster score largely disappears when complete sitewise substitution profiles are evaluated against a within-protein permutation null. (B) Even when a masked local representation is assigned to a comparison cluster, the ESM2 mutation profile does not systematically follow that same cluster. Such sites are modestly harder on average, but the attribution-aligned profile margin remains near zero.

C.10 ProteinGym: homolog-cohort breadth, diversity, and mechanism stress tests

We evaluate natural homolog breadth with complementary fixed-budget, mechanism-control, and phenotype-diversity designs; B denotes the number of training sequences in a cohort. The fixed-budget cohort uses ProteinGym target-region MSAs with $\geq 90\%$ target-position coverage and excludes the target and sequences at $\geq 90\%$ identity. Identity bins are near (70–90%), middle (50–70%), and remote (30–50%). In the four-family fixed-budget sensitivity ($B = 384$), narrow contains 384 proximal rows; broad contains 96 near, 48 middle, and 240 remote deduplicated rows; mixed-low adds 48 middle and 336 remote rows to the same 384 proximal rows; specialist-full contains 768 proximal rows; and mixed-full adds 96 middle and 672 remote rows to those 768 rows. The `xs` and `l` models share initialization and validation membership across paired conditions; three initializations are averaged within family. Scores use local masked LLR and full PLL mutant-minus-wild-type differences.

To isolate cluster breadth more tightly, we broaden the MSA pool and match exact counts in 2.5-percentage-point bins over 30–90% target identity. After deterministic subsampling, MMseqs clusters ungapped homologs at 70% identity and 80% coverage. A low-breadth cohort concentrates samples in large clusters; a high-breadth cohort maximizes represented clusters while matching target identity and coverage. Passing $B = 128$ cohorts require tight identity/coverage Wasserstein distances and at least 1.5–2-fold gains in cluster and effective-cluster counts. Column-shuffled controls preserve each alignment column's exact amino-acid/gap counts. A profile-matching pass reduces per-site total-variation distance while retaining the breadth contrast. Twenty strongly matched proteins, enriched for stability assays, form the neural panel. Three initializations replicate the reduced real-versus-shuffled contrast and are averaged within protein.

The phenotype-diversity analysis uses the same construction at $B = 64$, extends the identity range to 15–90%, and retains $\geq 90\%$ coverage. Thirteen non-stability targets pass: six Growth, five fluorescence-activated cell sorting (FACS), one Voltage, and one unclassified target. Shared Pfam architectures reduce these to nine evolutionary-family groups. One initialization is resumed with optimizer state intact until all four real/shuffled conditions early-stop; two further initializations repeat the reduced panel with a 4000-epoch ceiling. We average initializations within target and homologous targets within family group before inference. Full PLL uses a common deterministic 128-mutation subset per target/assay/condition. Growth and FACS summaries are descriptive only.

These natural analyses are deliberately not labeled as recovered phylogenetic or selection-regime interventions. Their role is to ask whether the consequences of breadth and heterogeneous training seen in the synthetic generator persist when natural MSA statistics—including marginal conservation, redundancy, and higher-order sequence combinations—change together.

C.11 ProteinGym: CATH/Gene3D hierarchy-controlled breadth

The hierarchy-controlled breadth analysis begins from 187 unique ProteinGym target sequences. Target domains are assigned structurally to CATH 4.4 and filtered so that the assigned domain contains at least 100 valid single substitutions and at least 90% of the target's valid single-substitution DMS mutations. The 98 resulting targets span 75 CATH homologous-superfamily groups (H-groups). Their H codes are matched exactly, without hierarchy remapping, to the official Gene3D v21.0.1 domain-sequence release (reported as CATH 4.3.0); all 75 groups are present. The extraction contains 12,302,139 canonical Gene3D domain sequences. A structural-CATH-only feasibility analysis is much shallower: after exact identity matching, only 16/12, 9/8, 6/5, 3/2, and 2/1 targets/H-groups remain at $B = 8, 16, 24, 32, 64$, respectively, motivating the deeper Gene3D corpus rather than relaxed matching.

Within each H-group, Gene3D sequences are clustered by MMseqs at 35% sequence identity and 80% coverage. Target-to-domain searches use a 20–70% target-identity window. The target-associated 35% cluster is identified from the closest searched Gene3D hit and excluded completely. Cohort construction removes exact sequence duplicates, retains at most one domain per full Gene3D parent protein, projects candidates onto the ProteinGym target-domain coordinate system, and requires at least 90% actual non-gap target coverage. Narrow and broad both contain 128 sequences and have exactly equal counts in every 2.5-percentage-point target-identity bin. Narrow draws all 128 rows from one remaining 35% cluster; broad draws from at least eight remaining clusters, with no cluster contributing more than 16 rows. A further 32 parent- and sequence-disjoint homologs form a validation cohort shared across paired conditions. These requirements leave 23 targets in 19 CATH H-groups. In the resulting cohorts, the median narrow–broad identity Wasserstein distance is 0.00188 (maximum 0.00416), and broad effective cluster count has median 14.86 and minimum 9.25.

For each target we train `xs` and `l` with three neural initializations and paired randomness across narrow and broad conditions. All 276 fits complete. DMS scoring uses all eligible within-domain substitutions for local masked LLR and a common 128-mutation subset for full PLL. Repeated assays are averaged within target/model; broad-minus-narrow is computed within target–scale–initialization, then initializations, homologous targets, and finally the 19 H-groups are aggregated in that order. Figure 10 shows the contrasts. Local breadth effects are $+0.0756$ for `xs` (95% CI $[0.0468, 0.1064]$) and $+0.0968$ for `l` ($[0.0625, 0.1354]$). The capacity interaction, $I = (\text{broad} - \text{narrow})_{\texttt{l}} - (\text{broad} - \text{narrow})_{\texttt{xs}}$, is $+0.0212$ ($[0.0064, 0.0390]$). Full-PLL effects are $+0.0721$ and $+0.0915$, with $I = +0.0194$ ($[-0.0045, 0.0447]$). The local interaction is positive under every initialization; full-PLL is less stable. Broad conditions generally select later validation-optimal checkpoints, but no fit reaches 4000 epochs, so the interaction reflects validation-selected rather than update-matched training.

C.12 ProteinGym: exact ESM2 UniRef50 2021_04 support

The pretraining-support analysis uses the same 187 unique ProteinGym target sequences and six ESM2 zero-shot score sets (8M, 35M, 150M, 650M, 3B, and 15B). Following Hou et al. (2026), homolog support is defined by searching target sequences against UniRef50 release 2021_04 at $> 20\%$ identity and 80% coverage. Rather than substituting a current UniRef snapshot, we parse

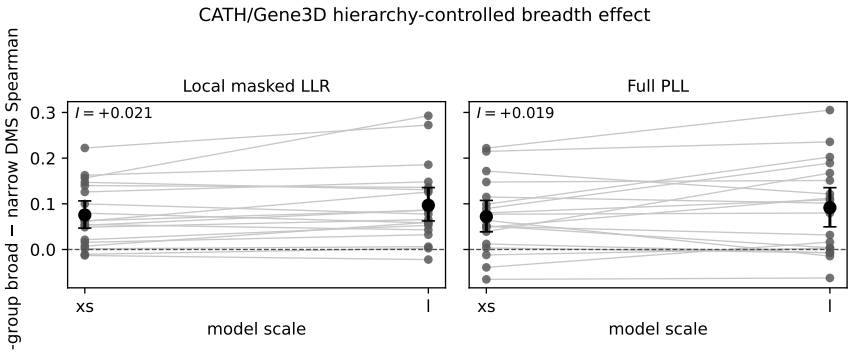


Figure 10: Hierarchy-controlled natural breadth within CATH homologous superfamilies. Each gray segment is one of 19 CATH H-groups after averaging repeated assays, neural initializations, and homologous ProteinGym targets in the specified order. Points show broad-minus-narrow DMS Spearman for `xs` and `l`; black markers and error bars show H-group means and bootstrap 95% intervals. Breadth improves both scales under local LLR and full PLL. The local capacity interaction is positive with an interval above zero; the full-PLL interaction has the same mean direction but crosses zero.

Table 7: Mean target-level ESM2 DMS Spearman and inferred wild-type NLL across the 187-target ProteinGym panel. Fitness ranking peaks at an intermediate scale while wild-type likelihood continues improving.

ESM2 scale	DMS ρ	WT NLL
8M	0.201	2.482
35M	0.305	2.237
150M	0.386	1.975
650M	0.423	1.580
3B	0.417	1.313
15B	0.408	1.104

the complete historical release used for ESM2 training: 49,874,565 UniRef50 clusters, released 2021-10-27. One hundred sixty-five targets have at least one exact-release match, with 46,883 filtered target–representative hits. UniRef50 XML metadata supplies declared cluster membership and nested UniRef90/UniRef100 identifiers; these hierarchy counts do not provide pairwise identities for the child sequences and are not treated as realized ESM2 exposures.

Wild-type negative log-likelihood (NLL) is reconstructed from ProteinGym masked-marginal ESM2 scores at sites for which all 19 alternatives are present: if $\ell_a = \log p(a) - \log p(\mathrm{WT})$, then $-\log p(\mathrm{WT}) = \log\big(1 + \sum_a e^{\ell_a}\big)$, where WT denotes wild type. Exact-release support count is negatively associated with mean wild-type NLL at every model scale: Spearman $\rho = -0.280, -0.295, -0.337, -0.307, -0.295$, and -0.185 from 8M through 15B, respectively, with all six bootstrap intervals below zero (Fig. 11A). In contrast, exact support count does not predict the primary 8M-to-650M DMS-performance gain among targets no longer than ESM2's 1,022-residue native context limit ($n = 174$): $\rho = -0.002$ (95% CI $[-0.150, 0.147]$; Fig. 11B). Nearest representative identity and mean distinct UniRef90 children per matched UniRef50 cluster are likewise near zero for this endpoint.

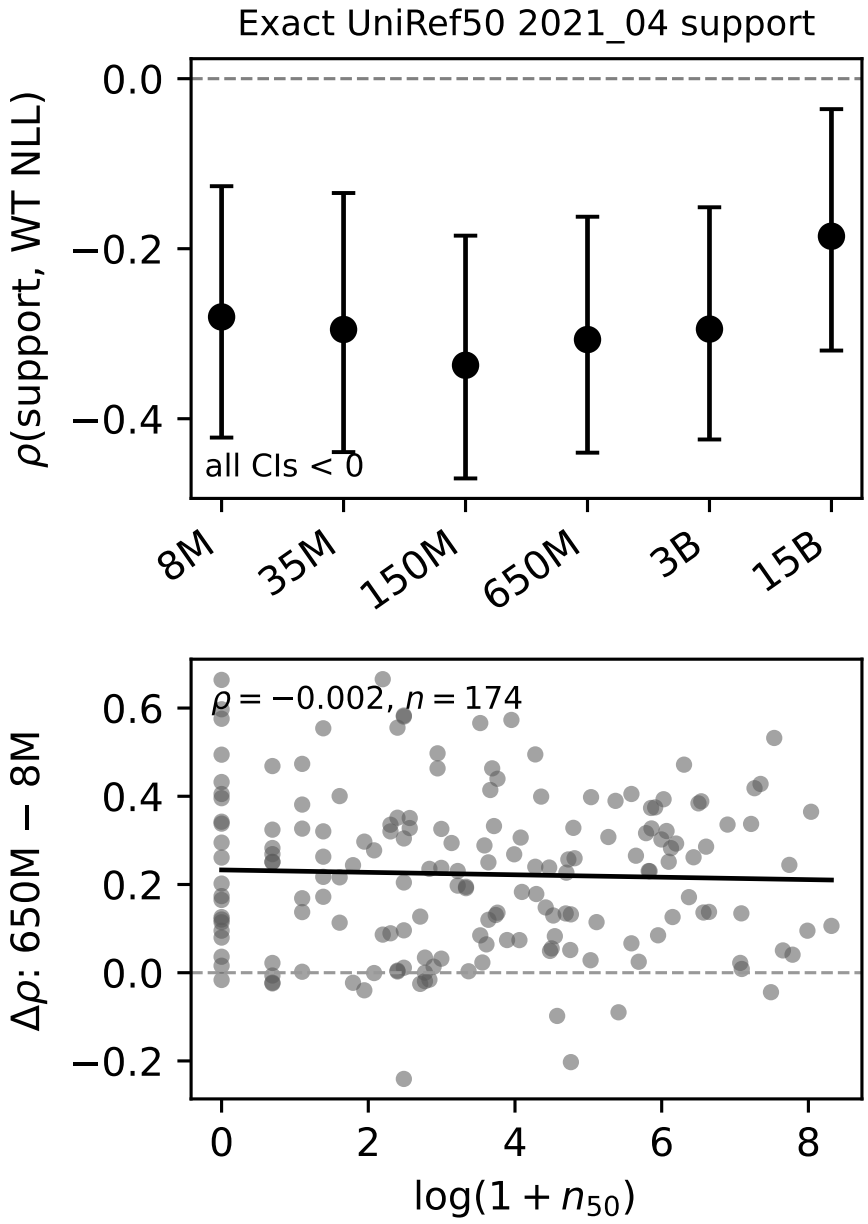


Figure 11: Exact historical pretraining support predicts ESM2 familiarity but not its primary scaling gain. Top: more UniRef50 2021_04 homolog clusters are associated with lower inferred wild-type NLL at all six ESM2 scales; error bars are target-bootstrap 95% intervals. Bottom: exact support count is essentially unrelated to the DMS-Spearman gain from 8M to 650M among native-length targets. Support is observational pretraining-pool geometry, not a reconstructed ESM2 sampling stream.

The mean DMS-Spearman trajectory is nonmonotonic (Table 7): it rises from 0.201 at 8M to 0.423 at 650M, then decreases to 0.408 at 15B while inferred WT NLL continues to fall. The exploratory 650M-to-15B analysis finds a positive association with database members per nested UniRef90 group (raw $\rho = 0.266$, 95% CI approximately $[0.10, 0.42]$; partial $\rho = 0.211$, $p = 0.009$ after controlling support count, nearest representative identity, and target length). Because several correlated support summaries were tested and database membership does not equal realized training frequency, this large-scale association is not used as a primary claim.

D Broader Impact and Reproducibility

The core work is methodological and uses synthetic sequence data, supplemented by observational and retraining-based external-validity analyses of public ProteinGym deep-mutational-scanning assays, MSAs, and released zero-shot predictions. Its positive impact is a clearer understanding of when learned protein representations can be expected to transfer across evolutionary novelty, potentially informing benchmark construction, pretraining-data selection, and calibrated use of zero-shot mutation scores. A risk is that simplified synthetic mechanisms or selected natural protein cohorts may be overgeneralized. We therefore distinguish causal conclusions from synthetic interventions from external signatures in ProteinGym, do not interpret cross-assay discordance as a pure biological landscape shift, and do not interpret MMseqs sequence clusters as recovered genealogy or natural homolog heterogeneity as a measured β intervention.

The anonymous code supplement contains the paper-facing generator, training, evaluation, and analysis code, together with configuration files, tests, a CPU smoke experiment, and a manuscript-to-code map. A dependency-ordered reproduction guide gives concrete commands for each synthetic and natural experiment family and records the world and initialization seeds, task grids, fixed hyperparameters, continuation rules, cohort sizes, and bootstrap and permutation seeds used in the reported analyses. The synthetic workflows regenerate their sequence and landscape worlds from the supplied configurations and random seeds. Scheduler-specific launch scripts and private filesystem assumptions have been removed; task-level jobs can instead be executed directly from explicit manifest indices.

The supplement includes the complete analysis path for the synthetic landscape-shift experiments, including validation-based checkpoint selection, shared-versus-specific prediction decomposition, near/remote same-landscape controls, ensemble-diversity and shared-component interventions, adaptation controls, and the representation–prediction agreement analysis. For the latter, the supplied code reproduces the first-pass 32-landscape oracle and representation analysis, the identity-matched three-landscape follow-up, and the stricter sensitivity retaining queries for which both comparison-landscape identity distributions have KS distance at most 0.12. The accompanying post-processing script applies the same target-landscape aggregation and bootstrap used for the reported sensitivity analysis.

The natural-data workflows operate on publicly available ProteinGym, ESM2, UniRef, CATH, and Gene3D resources and document the required releases, preprocessing steps, identity and coverage criteria, cohort construction, MMseqs searches, Potts fitting, representation extraction, and aggregation procedures. The CATH/Gene3D workflow includes the upstream outcome-blind target-domain mapping rather than assuming a private precomputed mapping. Historical-pretraining analyses specify the exact UniRef50 2021_04 release used for ESM2, while the hierarchy-controlled analysis records the CATH and Gene3D versions used by the study.

Large public-data mirrors, pretrained model weights, serialized synthetic worlds, trained checkpoints, oracle caches, fitted Potts objects, and other compute-intensive intermediate outputs are not duplicated in the supplement; the included code and runbook specify how to regenerate them. Neural initializations are averaged within the relevant biological unit rather than treated as independent evidence, and the lineage-breadth workflow retains the documented exclusion of the single run that failed to converge.