PROTEINFITNESSBENCH: LEARNING TO EXTRAPOLATE PROTEIN FITNESS ACROSS MUTATION ORDERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein engineering often requires selecting a small number of higher-order variants from models trained mainly on wild-type and low-order mutants. We introduce ProteinFitnessBench, a controlled framework that separates representation, readout, objective, and evaluation, and reports fixed-budget true-top-K recovery (TTK@K) alongside global ranking metrics. Across five seeds and four protein-fitness landscapes, a tested ΔESM-WHT readout package improves global ranking relative to a matched ΔESM-MLP comparator, but does not yield a consistent TTK@10 gain. This exposes a key evaluation mismatch: better global ranking does not necessarily produce better elite retrieval under a finite assay budget. Supporting GB1 analyses under sparse supervision and altered label resolution reinforce this distinction. ProteinFitnessBench provides a controlled way to test whether improvements in protein-fitness models translate into experimentally useful candidate selection.

1 INTRODUCTION

Protein engineering seeks a small number of high-fitness variants within an enormous sequence space. Depending on the campaign, fitness may reflect enzyme activity, fluorescence, binding, expression, or another functional phenotype. Combinatorial mutations make this search non-additive: epistasis occurs when a joint mutational effect departs from an additive expectation, and empirical landscapes show that such interactions can redirect accessible evolutionary and engineering paths (Weinreich et al., 2005; Poelwijk et al., 2007; Wu et al., 2016). In proteins, the measured phenotype can combine active-site interactions, binding interfaces, folding and thermal stability, conformational equilibria, expression and solubility, and long-range allostery. Structure-conditioned sequence-design methods such as ProteinMPNN demonstrate how backbone structure can guide sequence optimization (Dauparas et al., 2022). Separately, increased thermodynamic stability can buffer destabilizing mutations and broaden the mutational neighborhood compatible with function (Bloom et al., 2006). Stability, however, is not synonymous with fitness: the measured phenotype remains a projection of coupled biophysical constraints. Our understanding of the origins and engineering consequences of epistasis therefore continues to expand.

Deep mutational scanning, combinatorial libraries, and high-throughput screening now reveal increasingly large sequence–fitness landscapes (Fowler & Fields, 2014; Yang et al., 2019). Epistasis is consequently not only a theoretical property of these landscapes; it shapes what a learning system must extrapolate from the variants already measured. ML-guided directed-evolution workflows use models trained on assayed variants to prioritize combinatorial libraries for experimental testing (Wu et al., 2019). Under a limited wet-lab budget, the operational problem is candidate prioritization: which sequences should be tested next?

Benchmarking has moved accordingly toward more realistic generalization. FLIP standardized protein-fitness tasks (Dallago et al., 2021), ProteinGym assembled a broad collection of DMS assays and baselines (Notin et al., 2023), and FLIP2 expands evaluation across mutation-count, unseen-mutation, unseen-position, fitness, and sequence-background shifts (Didi et al., 2026). These benchmarks substantially improve the realism of fitness-model evaluation, especially when training data are concentrated near the wild type and evaluation includes higher-order mutants. Distribution-shift realism and experimental decision quality are nevertheless distinct questions: better ranking need not produce better experimental choices.

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.

Global ranking metrics do not answer that decision question in the same way. Spearman correlation summarizes rank agreement across the evaluated set. NDCG is a graded-relevance ranking metric that discounts lower-ranked items (Järvelin & Kekäläinen, 2002), but still evaluates the quality of an extended ranked list. A protein-engineering campaign may rank thousands or millions of candidates but can test only 10, 20, or 50. ProteinFitnessBench therefore reports true-top-K recovery (TTK@K), the fraction of the experimentally best K variants recovered among the model's top-K predictions, alongside global ranking metrics. This separates *global ranking quality* from *fixed-budget elite retrieval*.

Protein language models (pLMs) provide strong pretrained representations and support zero-shot and supervised fitness prediction (Rives et al., 2021; Meier et al., 2021; Lin et al., 2023). Under limited labeled supervision and demanding distribution shifts, however, their advantage is inconsistent, and simple one-hot or Ridge baselines can remain competitive (Biswas et al., 2021; Didi et al., 2026; Fournier et al., 2024; Gordon et al., 2025). Task-irrelevant variation, small-sample overfitting, and downstream capacity are plausible explanations, but benchmark performance alone does not establish them. Structural or multimodal information can help, but is not always available or equally accessible in early-stage or less-characterized systems. Representation quality may therefore be only part of the problem; the mapping from representation to fitness may itself impose an important inductive bias.

To disentangle representation, readout, training objective, distribution shift, and decision quality, we introduce **ProteinFitnessBench**, a modular framework that independently varies the encoder or representation, readout, objective, mutation-order split, and evaluation metric. The framework reports fixed-budget TTK@K alongside conventional global ranking metrics. Within ProteinFitnessBench, Walsh–Hadamard-based readouts are evaluated through a structured readout-package comparison: the transform reorganizes latent/channel coordinates while the underlying pretrained representation is held fixed, rather than performing a genotype-level Walsh epistasis decomposition (O'Donnell, 2014). Figure 1 summarizes the engineering setting and the benchmark's modular evaluation path.

ProteinFitnessBench makes two main contributions. 1) First, it provides a modular mutation-order extrapolation framework that separates representation, readout, objective, split, and evaluation choices, including fixed-budget TTK@K. 2) Second, it couples a controlled frozen-PLM readout comparison with decision-focused diagnostics that separate global ranking from finite-budget candidate selection.

Across these studies, three linked empirical findings emerge. 1) First, improvements in global ranking do not guarantee improved elite retrieval. 2) Second, the remaining retrieval gap is not well described by a single scalar ranking defect: the GB1 diagnostics reveal seed-variable elite coverage, deep-rank blocking, and residual conditional fine-ranking weakness. 3) Third, the tested predictor-side interventions, including elite-focused weighting, pair-difference supervision, and observed four-point supervision, do not yield a reproducible retrieval repair under their evaluated contracts. Together, these findings motivate prediction–decision decoupling as a future direction: fitness prediction and finite-budget candidate selection may require distinct objectives.

2 Related work

Protein fitness benchmarks and distribution shift. TAPE (Rao et al., 2019), FLIP (Dallago et al., 2021), and ProteinGym (Notin et al., 2023) standardized protein sequence evaluation and fitness prediction. FLIP2 explicitly broadens the benchmark to real-world engineering shifts and includes mutation-count splits that test extrapolation across mutational distance (Didi et al., 2026). ProteinBench similarly motivates multidimensional, user-objective-aware evaluation for protein foundation models (Ye et al., 2025); our focus is the complementary supervised fitness-prediction setting, where global ranking and finite-budget retrieval are evaluated separately. ProteinFitnessBench does not claim to introduce mutation-count splitting itself; it builds on this direction by (i) factorizing representation, pooling, readout, and objective under matched controls, (ii) tracking split and execution provenance, and (iii) adding a fixed-budget retrieval view through TTK@K. Top-K retrieval itself is not new: ProteinGym reports top-K recall with K set to the top 10% of each assay, and ECNet evaluates recovery of the true top 100 within a predicted top-K set under a sequencing-budget interpretation (Notin et al., 2023; Luo et al., 2021). Our contribution is the controlled use of explicit small

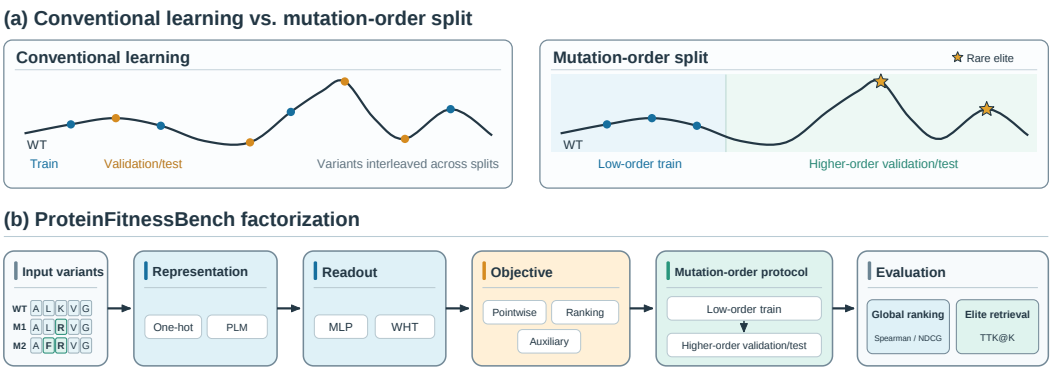


Figure 1: ProteinFitnessBench's engineering question and modular evaluation. (a) Schematic fitness landscape: measured wild-type and low-order variants support training, while a much larger higher-order candidate space may contain rare elite variants that must be prioritized under an assay budget of K. Horizontal position is an explanatory metaphor, not literal sequence-space distance. (b) The framework factorizes representation, readout, objective, mutation-order protocol, and evaluation. Global ranking is summarized by signed Spearman and NDCG; fixed-budget elite retrieval is measured by TTK@K. WHT is one tested readout within the framework.

assay budgets alongside controlled readout-package and mutation-order comparisons, followed by a measurement-aware analysis of the remaining retrieval gap.

Protein language models and conditional utility. Before pLMs, evolutionary-sequence models such as EVmutation and DeepSequence showed that coevolutionary and higher-order sequence constraints can predict mutational effects from homologous sequences (Hopf et al., 2017; Riesselman et al., 2018). Pretrained protein models provide strong representations and zero-shot scores (Rives et al., 2021; Rao et al., 2021; Meier et al., 2021; Notin et al., 2022; Lin et al., 2023). Their utility is not uniform across proteins or shifts. Gordon et al. show that zero-shot fitness performance depends on a pLM's implicit preference for the underlying sequence (Gordon et al., 2025); we therefore report wild-type ESM-2 preference as a descriptive landscape-level covariate in Appendix I, without treating it as an explanatory variable. FLIP2 likewise reports regimes where simpler models match or outperform fine-tuned pLMs (Didi et al., 2026). Our supervised setting complements these findings by holding the frozen pLM representation fixed and testing whether the *readout* changes what transfers under mutation-order shift.

Epistasis and Walsh representations. Walsh–Hadamard expansions provide a natural basis for functions on Boolean genotype spaces and have long been connected to interaction structure in fitness landscapes (O'Donnell, 2014; Poelwijk et al., 2007). That mathematical analogy motivates WHT-based readouts, but it does not by itself license a biological epistasis interpretation. In our implementation, the transform operates in learned feature/channel space. We therefore treat WHT as structured feature mixing rather than a direct epistasis estimator; interaction-oriented supervision is evaluated separately in Section 5.4.

3 ProteinFitnessBench

3.1 Task and mutation-order evaluation

Let $x^{(0)}$ denote the wild type and $k(x)$ the Hamming distance of variant x from $x^{(0)}$. In the strict GB1 protocol, training contains wild type, single, and double mutants; validation contains triple mutants; and test contains quadruple mutants. No test-order genotype is labeled during training. FLIP2 datasets are used under their supplied benchmark contracts where applicable; protocol tiers are reported separately and are never pooled into a single matched leaderboard.

3.2 Factorized predictor and controlled WHT readout

A predictor is written as encoder → pooling/site aggregation → prediction head → objective. Let $h_i^{(28)}(x)$ denote the layer-28 ESM-2 token representation at aligned residue i, and let $S(x) = \{i : x_i \neq x_i^{(0)}\}$ be the substitution sites relative to wild type $x^{(0)}$. The frozen sequence-level Δ feature is $|S(x)|^{-1} \sum_{i \in S(x)} [h_i^{(28)}(x) - h_i^{(28)}(x^{(0)})]$; the wild type maps to the zero vector. We use layer 28 based on the validation-only layer-selection sensitivity (Appendix B). We denote the supervised frozen-feature WHT readout as ΔESM-WHT and its supervised frozen-feature MLP comparator as ΔESM-MLP. The principal comparison uses this ESM-2 (650M) feature under the same data, split, loss, optimizer, seed, and training-budget contract. Both heads use the same objective and loss weights: globally weighted squared error (point weights 10/1/0.1 for the training-set top 5%, ranks 5–50%, and lower half), plus a top-5%-versus-rest margin hinge (margin 0.5; weight 0.5) and ΔNDCG-weighted LambdaRank (weight 0.5); no nested-pair auxiliary is used. ΔESM-WHT concatenates a learned mutation-order embedding, projects to 2,048 dimensions, applies a normalized fast Walsh–Hadamard transform and GELU, then uses a 256-unit ReLU/dropout(0.1) head; the ΔESM-MLP comparator maps the mean Δ feature through a 256-unit GELU hidden layer and receives no separate mutation-order embedding. At validation and test time, ΔESM-WHT indexes its order-embedding table by each variant's mutation count, capped at the configured maximum of 15; in an order-disjoint split, embedding rows for unseen orders have no direct training examples. Thus, the contrast evaluates the complete ΔESM-WHT readout package, not an isolated or parameter-matched FWHT ablation. The transform acts in latent/channel space, not as a direct q-ary genotype Walsh decomposition.

3.3 Metrics and decision alignment

We report signed Spearman ρ, NDCG, TTK@K, and MSE. TTK@K is

$$\mathrm{TTK@}K = \frac{|\operatorname{Top}_K(\hat{y}) \cap \operatorname{Top}_K(y)|}{K}, \tag{1}$$

Because both sets contain K variants, TTK@K equals precision@K and recall@K when the relevant set is the true top-K; the TTK name emphasizes a finite assay budget. Thus, TTK@10 answers a concrete experimental question: if only ten candidates can be tested, what fraction of the true top ten were selected? NDCG and Spearman assess broader ranking quality and can improve even when the identity of the extreme tail does not. This distinction is the basis of the decision-alignment studies below.

4 Experimental protocols

We group the experiments by their scientific role. The held-out full-point comparison uses frozen ESM-2 650M mean mutant-position Δ features (layer 28), complete point coverage, five fixed seeds $\{0, 42, 123, 888, 2026\}$, ten epochs without early stopping, AdamW (learning rate 3×10^{-4}, weight decay 10^{-5}), and batch sizes 64/16/16/8 for GB1/GFP/HIS7/FLIP2-Imine, with immutable split fingerprints. It compares ΔESM-WHT, LSTM-WHT, a ΔESM-MLP comparator, and OneHot-MLP on GB1, GFP, HIS7, and FLIP2-Imine. ΔESM-WHT versus ΔESM-MLP is the primary readout-package comparison under shared frozen features, data, split, loss, and optimization contract; the other internal arms provide context rather than a leaderboard claim.

Two validation-only studies diagnose decision alignment. The split control compares ΔESM-WHT with the ΔESM-MLP comparator under a compact GB1 mutation-order view and a size-matched random view. The objective control keeps ΔESM-WHT fixed while varying uniform versus elite-focused point weights and matched-random versus structured pair-difference streams. An independent four-point confirmation uses seeds $\{13, 42, 77, 123\}$ and compares point-only training with an observed-epsilon triplet auxiliary and a strict epsilon-shuffled control; its outer test remains sealed because the pre-specified promotion gate fails. Appendix A records datasets, seeds, split access, and purpose for each protocol.

Table 1: Readout-package comparison under frozen PLM representations. The held-out full-point protocol covers GB1, GFP, HIS7, and FLIP2-Imine with five seeds. Values are the mean ± SD across seed-level four-landscape macros. ΔESM-WHT versus ΔESM-MLP is the primary comparison; it shares the frozen representation, data, split, loss, optimizer, seeds, and training budget, but is not parameter-matched and does not isolate FWHT from the learned order embedding and projected head. Other rows are contextual controls.

Model	Role	NDCG ↑	Signed Spearman ↑	TTK@10 ↑
ΔESM-WHT	Primary readout package	0.9119 ± 0.0071	0.3897 ± 0.0150	0.015 ± 0.014
ΔESM-MLP	Frozen-feature comparator	0.9059 ± 0.0083	0.3465 ± 0.0073	0.030 ± 0.033
LSTM-WHT	Contextual control	0.9143 ± 0.0073	0.4090 ± 0.0167	0.010 ± 0.014
OneHot-MLP	Contextual control	0.9083 ± 0.0023	0.3810 ± 0.0114	0.050 ± 0.000

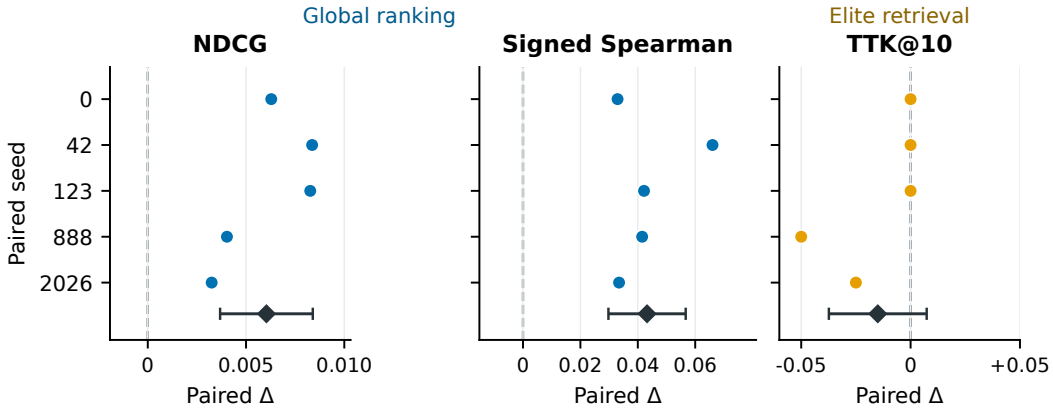


Figure 2: Global-ranking gains do not imply consistent elite-retrieval gains. Each point is one of the same five paired seeds in the four-landscape macro held-out protocol ($n = 5$ paired seeds; descriptive comparison); diamonds and horizontal bars show the mean ± sample SD of paired seed differences. For each metric, the paired difference is ΔESM-WHT minus ΔESM-MLP, so a positive difference denotes improvement. NDCG and signed Spearman have positive paired shifts, whereas TTK@10 has no consistent gain. The negative TTK@10 mean is not interpreted as systematic degradation.

5 Results

5.1 Readout-package comparison

Across five paired seed-level four-landscape macros, ΔESM-WHT increases NDCG by 0.006 ± 0.002 and signed Spearman by 0.043 ± 0.014 relative to ΔESM-MLP (mean ± SD of paired differences; Table 1 and the left/center panels of Figure 2). This is a comparison of full readout packages, including the order embedding and projected head; it is neither parameter-matched nor an isolated FWHT test. The contextual controls underscore the scope of this comparison: OneHot-MLP nearly matches ΔESM-WHT in global ranking and attains higher TTK@10.

5.2 Ranking–retrieval misalignment

The same paired comparison separates global ranking from fixed-budget candidate recovery (right panel of Figure 2). Although ΔESM-WHT improves NDCG and signed Spearman, its TTK@10 change is -0.015 ± 0.022 relative to ΔESM-MLP; no consistent TTK@10 gain was observed, and we do not interpret the negative mean as evidence of a systematic decrease. Global metrics evaluate ordering over the candidate landscape, whereas TTK@K asks whether an experimental budget of K assays contains members of the true elite set. We call this empirically observed distinction *ranking–retrieval misalignment*: an improvement in global ordering does not automatically produce an improvement in elite recovery. A separate compact, viability-filtered, validation-only GB1 diagnostic shows this separation outside the severe TTK floor: ΔESM-WHT TTK@10 decreases from 0.300 ± 0.000 in a size-matched random view to 0.080 ± 0.045 in the higher-order view, while signed Spearman is similar (0.749 ± 0.031 versus 0.755 ± 0.026; Appendix D).

5.3 SPARSE-SUPERVISION STRESS TEST

We next stress-tested this distinction under nested reductions of the GB1 HD$\leq$ 2 training pool: D100/D50/D20/D10 contain 2,168/1,084/434/217 training variants, respectively. An additive Ridge null/control baseline is included, consistent with recent analysis emphasizing additive controls when interpreting higher-order generalization (Visani et al., 2026); that analysis motivates the control but does not specify this implementation. All models are evaluated on the fixed HD4 universe (N = 121,174) using the completed seed-42 predictions. Table 2 reports the completed supervised grid. LSTM-WHT has the smallest D100-to-D10 Spearman decrease (-0.058), compared with -0.071 for Additive Ridge, -0.140 for MULTI-evolve, and -0.122 for ΔESM-WHT. This is a seed-42, one-seed stress test, not a matched causal WHT ablation; it provides no multi-seed robustness estimate and has no matched non-WHT ΔESM-MLP arm.

Table 2: Sparse-supervision robustness on the full GB1 HD4 candidate universe (N = 121,174; seed 42; one-seed stress test). Training subsets are nested within the HD$\leq$ 2 pool. TTK@50 is displayed as the most discriminative available evaluated budget because TTK@5/10/20 exhibit strong floor effects. This is not a matched causal WHT ablation. The additive Ridge control is motivated by analysis of additive baselines in higher-order generalization (Visani et al., 2026); this row is not a reproduction of that implementation.

Model	Signed Spearman ρ				$\Delta\rho$	TTK@50	
	D100	D50	D20	D10	D10–D100	D100	D10
Additive Ridge	0.300	0.298	0.301	0.229	-0.071	0.12	0.02
MULTI-evolve (Tran et al., 2026)	0.396	0.364	0.350	0.257	-0.140	0.18	0.00
ΔESM-WHT	0.399	0.364	0.151	0.277	-0.122	0.12	0.02
LSTM-WHT	0.362	0.355	0.351	0.304	-0.058	0.14	0.02

Retrieval does not show the same retention. TTK@50 falls from D100 to D10 for every supervised model: $0.12 \rightarrow 0.02$ for Additive Ridge, $0.18 \rightarrow 0.00$ for MULTI-evolve, $0.12 \rightarrow 0.02$ for ΔESM-WHT, and $0.14 \rightarrow 0.02$ for LSTM-WHT. TTK@5 is at the floor throughout, and TTK@10/20 are strongly floor-limited; TTK@50 is displayed because it is the most discriminative available evaluated budget. Thus, even the strongest observed global-ranking retention does not imply robust extreme-tail retrieval under sparse supervision. The full HD4-only matrix and the distinct joint HD3+HD4 table appear in Appendix C; their TTK values are not directly comparable.

5.4 DECISION-ALIGNMENT INTERVENTIONS

We next asked whether progressively more decision-aligned supervision can repair the retrieval gap. The intervention ladder begins with elite-focused point weighting, adds structured pair differences, and then uses observed four-point closures. Across GB1, GFP, and HIS7, the weighting and pair-difference factorial produces direction changes by landscape rather than a reproducible cross-landscape TTK@10 repair: the uniform structured-pair condition improves all three reported metrics on GFP, but structured pairs reduce TTK@10 on HIS7 and focused weighting reduces signed Spearman in the tested landscapes. These results are bounded to the tested weighting and pair recipes (Table 3).

Table 3: Decision-alignment intervention ladder. Weighting and pair-difference conditions are validation-only on GB1, GFP, and HIS7 with five seeds per condition. Four-point conditions are GB1 validation-only with four independent seeds; the outer test remained sealed after the pre-specified promotion gate failed. Four-point NDCG/Spearman values and design details appear in Appendix F.

Intervention	Data	TTK@10 outcome	Conclusion
Elite weighting	3 datasets, 5 seeds	No consistent cross-landscape gain	No repair
Pair differences	3 datasets, 5 seeds	Mixed across landscapes/weights	No consistent repair
Four-point auxiliary	GB1, 4 seeds	Focused = point-only (0.125 ± 0.050)	Gate failed
Shuffled-ϵ control	GB1, 4 seeds	0.100 ± 0.000	Control

The observed four-point auxiliary uses legal outer-train (WT, A, B, AB) closures with

$$\epsilon = f(AB) - f(A) - f(B) + f(WT). \tag{2}$$

In the independent GB1 validation-only confirmation (four seeds), point-only and focused observed-ϵ training tie at TTK@10 0.125 ± 0.050; the strict epsilon-shuffled control obtains 0.100 ± 0.000. Focused observed-ϵ increases secondary NDCG from 0.916 ± 0.004 to 0.921 ± 0.004, while reducing Spearman from 0.638 ± 0.036 to 0.626 ± 0.025. The pre-specified requirement of a strict focused-over-point TTK@10 improvement therefore fails, and the outer test remains sealed. Taken together, the tested decision-aligned objectives do not reliably close the fixed-budget retrieval gap.

5.5 GB1 ASSAY-RESOLUTION DIAGNOSTIC

We tested lower-tail training-label resolution in a separate GB1 HD0–2 train / HD3 validation / HD4 outer-test diagnostic. Labels below τ times wild-type fitness were set to zero only for training; validation and test labels remained continuous, and the pre-specified $\tau = 0.10$ contrast used Top-K Fitness Gap (TFG@K), evaluated at $K = 20$, as its primary fixed-budget endpoint (Appendix E). On HD3, ΔESM-MLP improved TFG@20 relative to $\tau = 0$ (paired $\Delta = -0.048 \pm 0.048$ SD; four of five training seeds improved) while mean Spearman fell from 0.502 to 0.483; the matched ΔESM-WHT arm did not show the same validation improvement ($+0.018$; one of five improved). On held-out HD4, ΔESM-MLP TFG@20 changed from 0.866 to 0.689 (paired $\Delta = -0.177 \pm 0.172$ SD; five of five improved), whereas ΔESM-WHT showed a smaller mixed change (-0.017 ± 0.038 SD; three of five improved). The result is model-, split-, and metric-specific; the controlled intervention does not simulate a particular DMS detection threshold or support a general censoring recommendation (Appendix Figure E.1).

5.6 GB1 RETRIEVAL ATTRIBUTION

To localize the remaining gap, we performed a GB1 HD4 retrieval attribution using a reconstructed count-posterior fitness reference y^* based on the Wu et al. relative-enrichment normalization (Wu et al., 2016). The analysis uses 500 independent count-posterior draws and five seeds each for ΔESM-MLP and ΔESM-WHT; it is a measurement-aware diagnostic, not an intervention. The observed-fitness and y^* top-10/20/50 identities coincide. Because both references derive from the same underlying pre- and post-selection assay counts, this is cross-processing consistency, not independent validation.

Three complementary retrieval diagnostics characterize the gap. First, early coverage is seed-variable: at a model-predicted top-2,500 cutoff, C_{10}—the fraction of reconstructed top-10 elites that enter that predicted set—spans 0.0–0.6 for ΔESM-MLP and 0.2–0.8 for ΔESM-WHT. Second, missed elites are often preceded by candidates whose reconstructed true ranks are far outside the top 10,000. We define a blocker as a non-elite candidate ranked ahead of a missed reconstructed top-10 variant. Across the five seeds, 83.2–95.0% of ΔESM-MLP blocker rows and 69.3–84.1% of ΔESM-WHT blocker rows have denoised true rank greater than 10,000. Third, conditional fine ranking remains limited: even an oracle-truncated diagnostic at 2,500 candidates, for which O_{10} is the fraction of reconstructed top-10 elites recovered after model ranking within the reconstructed top-2,500 set, yields only 0.0–0.1 for ΔESM-MLP and 0.0–0.2 for ΔESM-WHT. These are complementary diagnostics, not independent mechanisms: coverage loss and blocker burden are related views of missed-elite access, while oracle-truncated conditional ranking probes an additional limitation after conditioning on the reconstructed high-fitness region. Appendix Figure H.1 illustrates these diagnostics for a representative seed-0 model pair; the reported ranges summarize all five seeds.

A separate label-free coordinate diagnostic provides supporting context: across three replicas, raw pair coordinates have effective rank approximately 899–905, versus approximately 56.8 for Walsh pair coordinates; the HD=2 matched-information control is mutually affine to numerical precision. This is consistent with a structured-coordinate inductive bias, not added biological information or a causal biological mechanism (Appendix G).

6 DISCUSSION

Three linked findings organize the results. First, *global ranking quality and fixed-budget elite retrieval are related but non-equivalent*. Under the shared-protocol comparison, the ΔESM-WHT readout package improves NDCG and signed Spearman with frozen PLM features, yet does not consistently improve TTK@10. The one-seed sparse-supervision stress test shows the same separation under reduced supervision, and the compact viability-filtered GB1 validation-only diagnostic shows that higher-order shift can reduce elite recovery much more strongly than global ranking quality. These results therefore argue against treating a single global ranking statistic as a sufficient proxy for experimental candidate-selection quality.

Second, *the retrieval gap is not well described by a single scalar ranking defect*. The GB1 measurement-aware attribution identifies three complementary error signatures: seed-variable early elite coverage, deep-rank blocker burden, and residual conditional fine-ranking weakness. The assay-resolution rank-composition diagnostic provides an important boundary on this interpretation: on HD3, the predicted top-10, top-20, top-50, and top-100 sets contain no candidates in the observed true-rank $> 10{,}000$ bin, so poor retrieval cannot be reduced to a simple picture in which the predicted top list is uniformly flooded by extremely poor variants. At the same time, the full-HD4 retrieval attribution shows that many blockers ahead of missed elites have denoised true rank beyond 10,000, and oracle truncation to the reconstructed high-fitness top-2,500 region does not broadly restore top-10 recovery. These observations are complementary rather than contradictory: retrieval can fail through limited elite access, displacement by lower-quality candidates, and residual ranking errors even after attention is restricted to a high-fitness region. They do not establish a causal biological mechanism.

Third, *the tested predictor-side interventions do not reliably repair the retrieval gap*. Elite-focused point weighting, structured pair-difference supervision, and observed four-point supervision do not produce a reproducible retrieval improvement under their evaluated validation contracts. The separate GB1 assay-resolution diagnostic likewise shows that altering training-label resolution can improve TFG@20 for one model and split while mean Spearman moves in the opposite direction, with no corresponding general improvement for the matched ΔESM-WHT arm. Together, these results suggest that improving the predictive surrogate and improving the downstream selection decision are not necessarily the same optimization problem.

This third finding motivates *prediction–decision decoupling* as a future modeling direction, not an established solution. Fitness estimation and candidate selection need not share an objective. A learned fitness or preference surrogate could estimate relationships among variants from sequence and optional structural, assay, or experimental-history context using pointwise, pairwise, or ranking objectives, while a separate candidate-selection policy allocates a finite assay budget using those imperfect estimates. For a one-shot top-K decision, slate selection or decision-aware ranking may be appropriate; when assay outcomes update the state across experimental rounds, contextual bandits or reinforcement learning become more natural formulations. Such a downstream policy can change how predictive information is used, but it cannot recover information absent from the representation or surrogate. Testing this direction would require its own prospective or sequential evaluation contract.

Limitations. The primary readout comparison spans four landscapes and five seeds, but it evaluates a complete architecture package rather than isolating FWHT or matching parameter count; the sparse-supervision stress test is one seed and lacks a matched non-WHT ΔESM-MLP arm. The retrieval-attribution analysis is GB1-specific and conditional on the reconstructed count-posterior reference y^*; its posterior draws quantify count-measurement uncertainty rather than candidate-resampling uncertainty. The split, objective, and four-point controls are validation-only by design; sealed outer tests prevent post-hoc promotion but also limit deployment claims. External methods retain native training contracts, so cross-method comparisons are contextual rather than causal. The four benchmark wild types also span distinct preference levels under the same frozen ESM-2 650M model (mean masked PLL per residue, -2.805 to -1.132; Appendix I), but with only four landscapes and a mixed, non-monotonic performance pattern, we find no basis to attribute cross-landscape heterogeneity to this covariate. Frozen pLM features may also inherit evolutionary biases from natural-sequence pretraining: MacLean et al. show that pLM probabilities can encode nucleotide-level mutational accessibility alongside fitness-related constraints (MacLean et al., 2026).

Although our assay-defined mutant libraries are not sampled according to natural mutational supply, representation-level bias remains a limitation we do not explicitly disentangle. Finally, TTK@K is intentionally budget-specific: conclusions at one assay budget need not transfer unchanged to another.

7 Conclusion

ProteinFitnessBench establishes mutation-order fitness prediction as both a ranking problem and a fixed-budget candidate-selection problem. In the tested shared-protocol comparison, the ΔESM-WHT readout package improves global ranking under a frozen PLM feature without reliably improving recovery of the elite variants that a finite experimental budget can test. The single-seed GB1 sparse-supervision stress test illustrates the distinction under reduced labels; the tested decision-alignment objectives do not repair the retrieval gap reliably. A measurement-aware retrieval attribution reports three complementary error signatures—coverage, deep-rank blocking, and conditional fine ranking after conditioning on the reconstructed high-fitness region—rather than a single ranking defect. The benchmark therefore provides a controlled setting for determining when conventional ranking gains translate into decision-relevant candidate discovery.

Reproducibility statement

The supplementary material includes the processed datasets, exact split manifests, source code, and frozen artifacts needed to reproduce the main reported analyses.

AI use statement

LLMs assisted with research ideation and conceptual framing, feedback on methodology and experimental design, literature and citation checks, and manuscript drafting and polishing. They were not used to generate experimental data or fabricate results. The authors reviewed and verified all AI-assisted technical content, claims, code, analyses, and interpretations, and take responsibility for the final paper.

References

Thomas Bikias, Evangelos Stamkopoulos, and Sai T. Reddy. PLMFit: Benchmarking transfer learning with protein language models for protein engineering. *Briefings in Bioinformatics*, 26(4): bbaf381, 2025. doi: 10.1093/bib/bbaf381.

Surojit Biswas, Grigory Khimulya, Ethan C. Alley, Kevin M. Esvelt, and George M. Church. Low-N protein engineering with data-efficient deep learning. *Nature Methods*, 18(4):389–395, 2021.

Jesse D. Bloom, Sy T. Labthavikul, Christopher R. Otey, and Frances H. Arnold. Protein stability promotes evolvability. *Proceedings of the National Academy of Sciences*, 103(15):5869–5874, 2006. doi: 10.1073/pnas.0510098103.

Christian Dallago, Jody Mou, Kadina E. Johnston, Bruce Wittmann, Nicholas Bhattacharya, Samuel Goldman, Ali Madani, and Kevin K. Yang. FLIP: Benchmark tasks in fitness landscape inference for proteins. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2021.

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Alexis Courbet, Rob J. de Haas, Neville Bethel, Philip J. Y. Leung, Thomas F. Huddy, Samuel Pellock, Doug Tischer, Frank Chan, Buwei Koepnick, Henry Nguyen, Alex Kang, Bharath Sankaran, Anindita K. Bera, Neil P. King, and David Baker. Robust deep learning-based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, 2022. doi: 10.1126/science.add2187.

Kieran Didi, Sarah Alamdari, Alex X. Lu, Bruce Wittmann, Kadina E. Johnston, Ava P. Amini, Ali Madani, Maya Czeneszew, Christian Dallago, and Kevin K. Yang. FLIP2: Expanding protein fitness landscape benchmarks for real-world machine learning applications. *Forty-third International Conference on Machine Learning*, 2026.

Quentin Fournier, Robert M. Vernon, Almer van der Sloot, Benjamin Schulz, Sarath Chandar, and Christopher James Langmead. Protein language models: Is scaling necessary? *bioRxiv*, 2024. doi: 10.1101/2024.09.23.614603.

Douglas M. Fowler and Stanley Fields. Deep mutational scanning: a new style of protein science. *Nature Methods*, 11(8):801–807, 2014.

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

Thomas A. Hopf, John B. Ingraham, Frank J. Poelwijk, Charlotta P. I. Schärfe, Michael Springer, Chris Sander, and Debora S. Marks. Mutation effects predicted from sequence co-variation. *Nature Biotechnology*, 35(2):128–135, 2017. doi: 10.1038/nbt.3769.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002. doi: 10.1145/582415.582418.

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.

Yunan Luo, Guangde Jiang, Tianhao Yu, Yang Liu, Lam Vo, Hantian Ding, Yufeng Su, Wesley Wei Qian, Huimin Zhao, and Jian Peng. ECNet is an evolutionary context-integrated deep learning framework for protein engineering. *Nature Communications*, 12(1):5743, 2021. doi: 10.1038/s41467-021-25976-8.

Oscar A. MacLean, Kieran Lamb, Laura Mojsiejczuk, Spyros Lytras, Ke Yuan, Joseph Hughes, and David L. Robertson. Protein language models learn underlying mutation biases alongside fitness landscapes. *bioRxiv*, 2026. doi: 10.64898/2026.07.06.736753.

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, pp. 29287–29303, 2021.

Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena-Hurtado, Aidan N. Gomez, Debora Marks, and Yarin Gal. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16990–17017, 2022.

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, Datasets and Benchmarks Track*, 2023.

Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.

Frank J. Poelwijk, Daniel J. Kiviet, Daniel M. Weinreich, and Sander J. Tans. Empirical fitness landscapes reveal accessible paths to the evolution of drug resistance. *Nature*, 445(7126):863–866, 2007.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun S. Song. Evaluating protein transfer learning with TAPE. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Roshan Rao, Jason Liu, Robert Verkuil, Joshua Meier, John F. Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. MSA transformer. In *Proceedings of the 38th International Conference on Machine Learning*, ICML, 2021.

Adam J. Riesselman, John B. Ingraham, and Debora S. Marks. Deep generative models of genetic variation capture the effects of mutations. *Nature Methods*, 15(10):816–822, 2018. doi: 10.1038/s41592-018-0138-4.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

Vincent Q. Tran, Matthew Nemeth, Liam J. Bartie, Sita S. Chandrasekaran, Alison Fanton, Hyungseok C. Moon, Brian L. Hie, Silvana Konermann, and Patrick D. Hsu. Rapid directed evolution guided by protein language models and epistatic interactions. *Science*, 392(6798), 2026. doi: 10.1126/science.aea1820.

Gian Marco Visani, Aayush Verma, and William S. DeWitt. Additive baselines furnish no evidence for epistasis learning by MULTI-evolve. *bioRxiv*, 2026. doi: 10.64898/2026.04.23.719915.

Daniel M. Weinreich, Robert A. Watson, and Lin Chao. Perspective: Sign epistasis and genetic constraint on evolutionary trajectories. *Evolution*, 59(6):1165–1174, 2005.

Nicholas C. Wu, Lei Dai, C. Anders Olson, James O. Lloyd-Smith, and Ren Sun. Adaptation in protein fitness landscapes is facilitated by indirect paths. *eLife*, 5:e16965, 2016. doi: 10.7554/eLife.16965.

Zachary Wu, S. B. Jennifer Kan, Russell D. Lewis, Bruce J. Wittmann, and Frances H. Arnold. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proceedings of the National Academy of Sciences of the United States of America*, 116(18):8852–8858, 2019. doi: 10.1073/pnas.1901979116.

Kevin K. Yang, Zachary Wu, and Frances H. Arnold. Machine-learning-guided directed evolution for protein engineering. *Nature Methods*, 16(8):687–694, 2019.

Fei Ye, Zaixiang Zheng, Dongyu Xue, Yuning Shen, Lihao Wang, Yiming Ma, Yan Wang, Xinyou Wang, Xiangxin Zhou, and Quanquan Gu. ProteinBench: A holistic evaluation of protein foundation models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/49b9d06c2dae5af46125490e0a6f9e32-Abstract-Conference.html.

A Experiment map

Table A.1: Map of analyses, datasets, repeats, purposes, and result locations.

Analysis	Data	Repeats	Purpose	Results
Primary readout comparison	4 landscapes	5 seeds	Compare readout packages	Table 1; Figure 2
Split control	GB1	5 seeds	Compare mutation-order and size-matched random evaluation	Appendix D
Objective control	3 landscapes	5 seeds	Test weighting and pair-difference supervision	Table 3
Four-point confirmation	GB1	4 seeds	Test observed-ϵ auxiliary	Table 3; Appendix F
Sparse-supervision stress test	GB1	1 seed	Stress-test reduced supervision	Table 2; Appendix C
Assay-resolution diagnostic	GB1	5 seeds	Test training-label resolution	Section 5.5; Appendix E
Retrieval attribution	GB1	5 seeds/model	Characterize retrieval errors	Section 5.6; Appendix H
Geometry diagnostic	GB1 pair features	3 replicas	Characterize coordinate geometry	Section 5.6; Appendix G

B ESM LAYER-SELECTION SENSITIVITY

Layer selection was a validation-only Ridge sensitivity screen over ESM2-650M layers 0–33 on GB1 and HIS7. It was treated as a task-specific feature-extraction choice rather than a universal default (Bikias et al., 2025). The selected layer was locked lexicographically by validation macro TTK@10, then NDCG, then signed Spearman; neither test performance nor MSE selected the layer.

Table B.1: Validation-only ESM2-650M layer-selection sensitivity. A Ridge screen over layers 0–33 selected the layer lexicographically by macro TTK@10, then NDCG, then signed Spearman. Test performance and MSE did not select the layer. This comparison documents sensitivity and selection provenance, not a universal best-layer claim.

Probe	L28	L33
Validation macro TTK@10	0.050	0.000
Validation macro NDCG	0.9194	0.9213
Validation macro Spearman	0.6693	0.6862

A full-training Ridge follow-up provides a cross-check rather than an additional selection criterion. On GB1, L28 obtains TTK@5/10 of $0.40/0.20$, versus $0.20/0.10$ for L33; on HIS7, no corresponding intermediate-layer benefit appears and L33 ranks slightly better. In downstream paired comparisons, ΔESM-WHT and LSTM-WHT have better ranking at L28 across the two studied landscapes. Layer sensitivity does not establish a universal optimal layer.

C SPARSE-SUPERVISION FULL METRICS

C.1 HD4-ONLY EVALUATION

The full seed-42 sparse-supervision results are evaluated on 121,174 HD4 candidates. ESM2-LLR is zero-shot and therefore appears once. D100/D50/D20/D10 contain 2,168/1,084/434/217 training variants, respectively.

Table C.1: Sparse-supervision results on HD4 candidates ($N = 121{,}174$; seed 42).

Model	Fraction	ρ	NDCG	TTK@5	TTK@10	TTK@20	TTK@50
ESM2-LLR	zero-shot	-0.1905	0.5888	0.00	0.00	0.00	0.00
Additive Ridge	D100	0.3000	0.7276	0.00	0.00	0.00	0.12
Additive Ridge	D50	0.2979	0.7288	0.00	0.00	0.05	0.06
Additive Ridge	D20	0.3013	0.7179	0.00	0.00	0.00	0.02
Additive Ridge	D10	0.2293	0.7008	0.00	0.00	0.00	0.02
MULTI-evolve	D100	0.3964	0.8286	0.00	0.10	0.10	0.18
MULTI-evolve	D50	0.3639	0.7570	0.00	0.00	0.00	0.02
MULTI-evolve	D20	0.3500	0.7410	0.00	0.00	0.00	0.00
MULTI-evolve	D10	0.2568	0.6865	0.00	0.00	0.00	0.00
ΔESM-WHT	D100	0.3989	0.8199	0.00	0.00	0.00	0.12
ΔESM-WHT	D50	0.3636	0.7621	0.00	0.00	0.00	0.02
ΔESM-WHT	D20	0.1510	0.7185	0.00	0.00	0.00	0.04
ΔESM-WHT	D10	0.2770	0.7358	0.00	0.00	0.00	0.02
LSTM-WHT	D100	0.3617	0.8048	0.00	0.00	0.00	0.14
LSTM-WHT	D50	0.3552	0.8051	0.00	0.00	0.00	0.10
LSTM-WHT	D20	0.3514	0.7585	0.00	0.00	0.00	0.04
LSTM-WHT	D10	0.3041	0.7233	0.00	0.00	0.00	0.02

Absolute TTK values should not be compared with Table C.2 because the ranked candidate sets differ.

C.2 JOINT HD3+HD4 EVALUATION

This joint view ranks a different candidate set from Table C.1; its absolute TTK values are therefore not directly comparable.

Table C.2: Sparse-supervision results on the joint HD3+HD4 candidate set ($N = 147{,}193$; seed 42).

Model	Fraction	ρ	NDCG	TTK@5	TTK@10	TTK@20	TTK@50
ESM2-LLR	zero-shot	-0.1099	0.6522	0.00	0.00	0.00	0.00
Additive Ridge	D100	0.3263	0.7696	0.00	0.00	0.00	0.06
Additive Ridge	D50	0.3228	0.7666	0.00	0.00	0.05	0.04
Additive Ridge	D20	0.3226	0.7566	0.00	0.00	0.00	0.00
Additive Ridge	D10	0.2478	0.7350	0.00	0.00	0.00	0.02
MULTI-evolve	D100	0.4535	0.8692	0.00	0.00	0.00	0.02
MULTI-evolve	D50	0.4200	0.8280	0.00	0.00	0.00	0.02
MULTI-evolve	D20	0.4030	0.8104	0.00	0.00	0.00	0.00
MULTI-evolve	D10	0.3139	0.7623	0.00	0.00	0.00	0.00
ΔESM-WHT	D100	0.4388	0.8609	0.00	0.00	0.00	0.02
ΔESM-WHT	D50	0.3807	0.8208	0.00	0.00	0.00	0.00
ΔESM-WHT	D20	0.1714	0.7923	0.00	0.00	0.00	0.00
ΔESM-WHT	D10	0.3138	0.7805	0.00	0.00	0.00	0.00
LSTM-WHT	D100	0.3864	0.8475	0.00	0.00	0.00	0.04
LSTM-WHT	D50	0.3752	0.8417	0.00	0.00	0.00	0.04
LSTM-WHT	D20	0.3847	0.8163	0.00	0.00	0.00	0.00
LSTM-WHT	D10	0.3309	0.7786	0.00	0.00	0.00	0.00

D COMPACT DISTRIBUTION-SHIFT DIAGNOSTIC

This validation-only diagnostic uses a viability-filtered GB1 view with train $N = 1{,}288$, HD3 $N = 1{,}000$, and HD4 $N = 1{,}500$; it is not a full-GB1 result. The ΔESM-WHT versus ΔESM-MLP comparison likewise evaluates the full readout package rather than an isolated transform.

Table D.1: Compact matched distribution-shift diagnostic (GB1 validation only; five seeds).

Readout	View	TTK@10	NDCG	Signed Spearman
ΔESM-WHT	Size-matched random	0.300 ± 0.000	0.8938 ± 0.0206	0.7487 ± 0.0308
ΔESM-WHT	Higher-order	0.080 ± 0.045	0.9077 ± 0.0057	0.7546 ± 0.0256
ΔESM-MLP	Size-matched random	0.140 ± 0.055	0.8582 ± 0.0040	0.6307 ± 0.0484
ΔESM-MLP	Higher-order	0.080 ± 0.084	0.8869 ± 0.0108	0.5736 ± 0.1011

E ASSAY-RESOLUTION DIAGNOSTIC

This independent GB1 diagnostic uses the frozen HD0–2 train / HD3 validation / HD4 outer-test split, five training seeds $\{0, 42, 123, 888, 2026\}$, pointwise squared error, and train-only hard censoring $y < \tau y_{WT} \mapsto 0$ on a verified wild-type-normalized linear scale. The primary threshold was $\tau = 0.10$; 0.01, 0.05, and 0.20 were HD3 sensitivity thresholds only. HD4 used only $\tau = 0$ and 0.10 after the branch-level gate passed through ΔESM-MLP. HD3 and HD4 labels used for evaluation were not censored. For a predicted top-K selected set S_K, the reported Top-K Fitness Gap is

$$\mathrm{TFG@}K = 1 - \frac{\frac{1}{K}\sum_{i \in S_K} y_i - \bar{y}}{\frac{1}{K}\sum_{i \in \mathrm{Top}_K(y)} y_i - \bar{y}}, \tag{3}$$

where $\bar{y}$ is the mean observed fitness over the evaluated candidate universe. $\mathrm{TFG@}K = 0$ corresponds to the oracle top-K mean fitness, whereas $\mathrm{TFG@}K = 1$ corresponds to a selected set whose mean fitness equals the candidate-set mean. Lower values are better. TTK@K measures exact elite identity recovery, whereas TFG@K measures the fitness shortfall of the selected set relative to the oracle top-K. The reported $\pm$ values are sample SD across five training seeds, not biological replicate uncertainty or SEM.

Table E.1: HD3 validation threshold sensitivity. Values are mean ± sample SD across five training seeds. Only $\tau = 0.10$ entered the pre-specified promotion gate; the other nonzero thresholds are descriptive sensitivity points.

Model	τ	Spearman ↑	TFG@20 ↓
ΔESM-MLP	0.00	0.5019 ± 0.0522	0.4391 ± 0.0301
ΔESM-MLP	0.01	0.5150 ± 0.0463	0.4841 ± 0.1055
ΔESM-MLP	0.05	0.4787 ± 0.0542	0.4575 ± 0.1181
ΔESM-MLP	0.10	0.4834 ± 0.0746	0.3908 ± 0.0339
ΔESM-MLP	0.20	0.4657 ± 0.0192	0.4790 ± 0.0900
ΔESM-WHT	0.00	0.6040 ± 0.0242	0.3301 ± 0.0431
ΔESM-WHT	0.01	0.6365 ± 0.0265	0.3207 ± 0.0494
ΔESM-WHT	0.05	0.6297 ± 0.0212	0.3343 ± 0.0286
ΔESM-WHT	0.10	0.6067 ± 0.0155	0.3479 ± 0.0438
ΔESM-WHT	0.20	0.6186 ± 0.0086	0.3439 ± 0.0251

Table E.2: Per-seed HD3 validation at the primary assay-resolution contrast. Each row is one training seed; Δ is $\tau = 0.10$ minus $\tau = 0$ in TFG@20. ΔESM-MLP passed the HD3 model-specific gate; ΔESM-WHT did not.

Model	Seed	TFG@20 ↓		ΔTFG@20	Spearman ρ ↑	
		$\tau = 0$	$\tau = 0.10$	$\tau = .10 - 0$	$\tau = 0$	$\tau = 0.10$
ΔESM-MLP	0	0.4027	0.3861	-0.0166	0.4407	0.5699
ΔESM-MLP	42	0.4270	0.4271	+0.0002	0.5800	0.5404
ΔESM-MLP	123	0.4842	0.3725	-0.1117	0.4801	0.4650
ΔESM-MLP	888	0.4484	0.4217	-0.0267	0.4880	0.4617
ΔESM-MLP	2026	0.4334	0.3465	-0.0869	0.5209	0.3799
ΔESM-WHT	0	0.3430	0.2844	-0.0586	0.6077	0.5916
ΔESM-WHT	42	0.3451	0.3720	+0.0268	0.6415	0.5963
ΔESM-WHT	123	0.2548	0.3220	+0.0672	0.5883	0.6188
ΔESM-WHT	888	0.3645	0.3688	+0.0043	0.5780	0.5996
ΔESM-WHT	2026	0.3430	0.3920	+0.0490	0.6045	0.6275

Table E.3: Per-seed HD4 outer-test results for the assay-resolution contrast. Δ is $\tau = 0.10$ minus $\tau = 0$ in TFG@20. LSTM-WHT is an HD4-only exploratory sensitivity arm with no separate HD3 promotion gate.

Model	Seed	TFG@20 ↓		ΔTFG@20	Spearman ρ ↑	
		$\tau = 0$	$\tau = 0.10$	$\tau = .10 - 0$	$\tau = 0$	$\tau = 0.10$
ΔESM-MLP	0	0.7873	0.7815	-0.0058	0.3148	0.3072
ΔESM-MLP	42	0.7489	0.5011	-0.2478	0.3726	0.4109
ΔESM-MLP	123	0.9460	0.8880	-0.0580	0.2473	0.3345
ΔESM-MLP	888	0.9994	0.8661	-0.1333	0.2457	0.3478
ΔESM-MLP	2026	0.8467	0.4079	-0.4389	0.2830	0.3058
ΔESM-WHT	0	0.4266	0.3722	-0.0544	0.4170	0.4109
ΔESM-WHT	42	0.4830	0.4303	-0.0526	0.4240	0.3619
ΔESM-WHT	123	0.4102	0.3896	-0.0207	0.4034	0.4486
ΔESM-WHT	888	0.4038	0.4257	+0.0219	0.3827	0.4327
ΔESM-WHT	2026	0.4045	0.4260	+0.0215	0.4192	0.4329
LSTM-WHT	0	0.6431	0.6831	+0.0400	0.4052	0.3986
LSTM-WHT	42	0.6017	0.5437	-0.0580	0.3986	0.3842
LSTM-WHT	123	0.5514	0.6090	+0.0576	0.3763	0.3593
LSTM-WHT	888	0.6370	0.5930	-0.0440	0.3925	0.3781
LSTM-WHT	2026	0.7083	0.6776	-0.0307	0.3823	0.3664

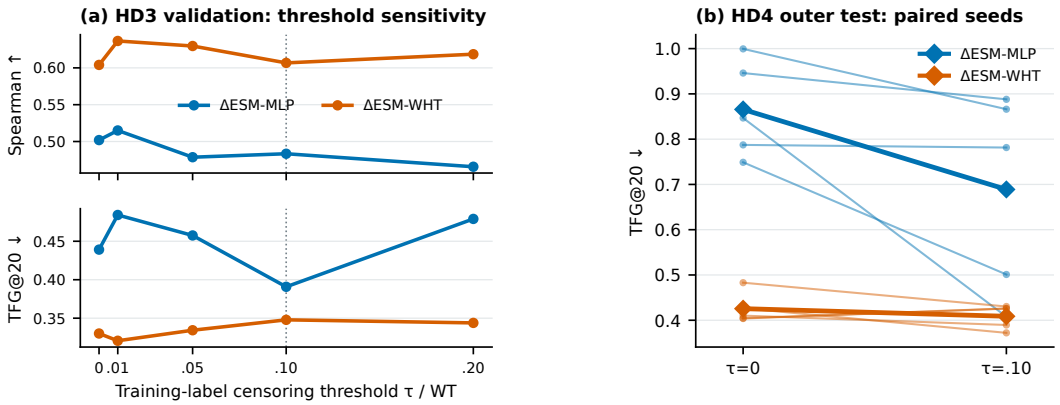


Figure E.1: Training-label resolution sensitivity on GB1. (a) HD3 validation means across five training seeds for both models at the frozen threshold grid; lower TFG@20 and higher Spearman are better. The dotted line marks the pre-specified $\tau = 0.10$. (b) HD4 outer-test TFG@20 for the same five paired training seeds at $\tau = 0$ and 0.10; thin lines are seed pairs, and diamonds are means. Mean paired differences ($\tau = 0.10$ minus 0) are -0.177 ± 0.172 SD for ΔESM-MLP and -0.017 ± 0.038 SD for ΔESM-WHT. This is a descriptive, model-specific comparison.

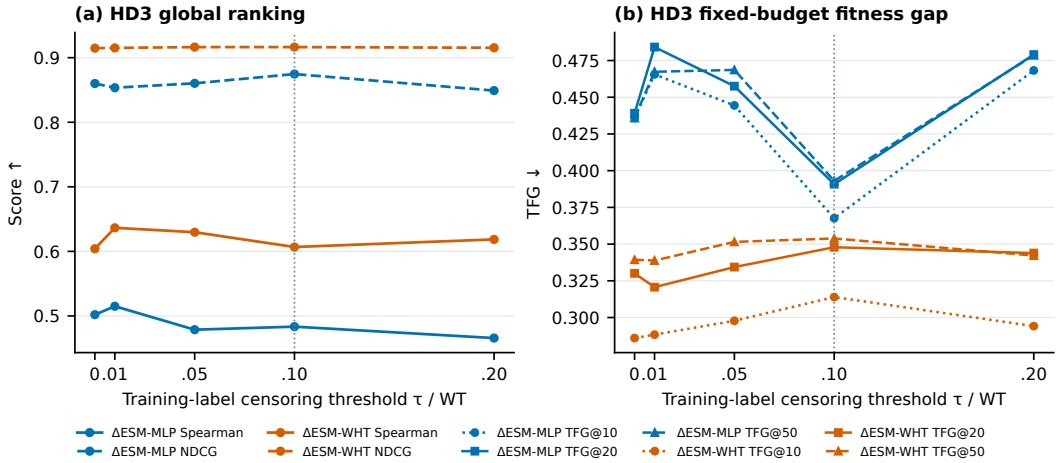


Figure E.2: Full HD3 validation sensitivity at the five frozen censoring thresholds. (a) Global ranking by Spearman and NDCG, where higher is better. (b) Fixed-budget TFG@10/20/50, where lower is better. Lines are means across five training seeds; no uncertainty band is plotted. The vertical dotted line marks the pre-specified $\tau = 0.10$ contrast.

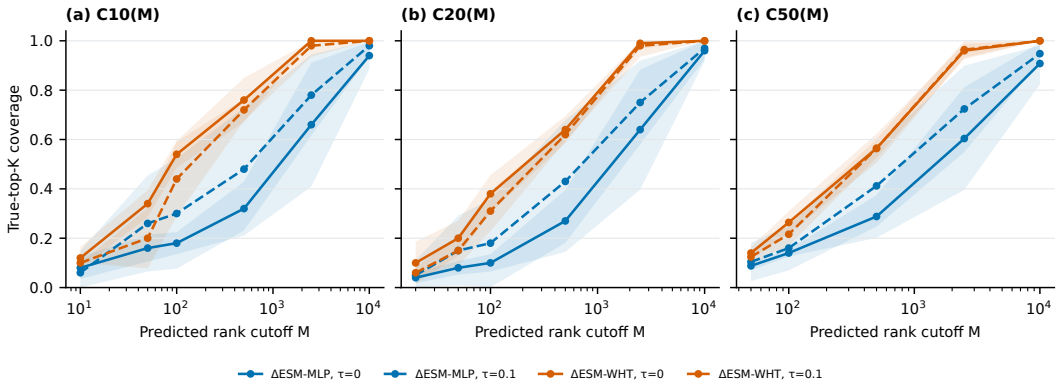


Figure E.3: HD3 validation elite coverage $C_K(M) = |\operatorname{Top}_K(y) \cap \operatorname{Top}_M(\hat{y})|/K$ for (a) $K = 10$, (b) $K = 20$, and (c) $K = 50$. Curves are means across five training seeds; shaded regions are mean $\pm$ sample SD, clipped to $[0, 1]$. Each displayed M is unique and satisfies $M \geq K$. The true-top-K set uses original continuous validation labels.

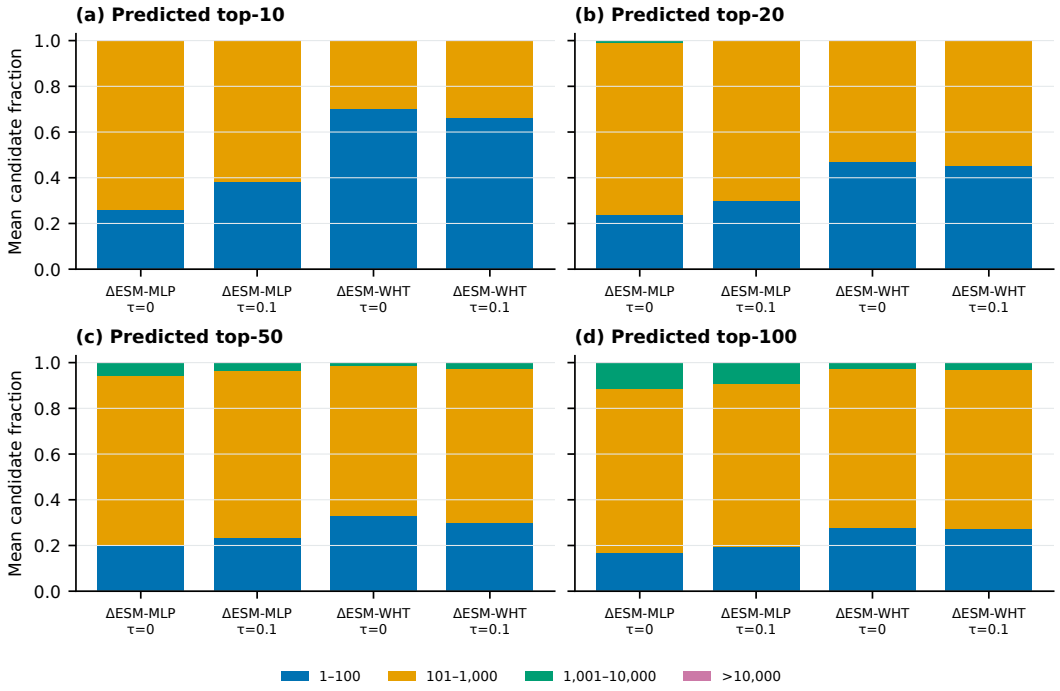


Figure E.4: HD3 validation true-rank composition of predicted (a) top-10, (b) top-20, (c) top-50, and (d) top-100 candidates, averaged across five training seeds. Mutually exclusive observed-label rank bins cover every selected candidate, so each stack sums to one. The $> 10{,}000$ bin is absent in every HD3 cell. These are observed true-rank categories, distinct from reconstructed-reference blocker rows in the retrieval-attribution analysis.

In the HD4 assay-resolution diagnostic, severe-blocker rate is the fraction of predicted top-ten candidates with observed true rank $> 10{,}000$ (denominator ten per seed), and for ΔESM-MLP its HD4 mean is 0.280 at $\tau = 0$ and 0.100 at $\tau = .10$. This reduction coincides with the TFG@20 improvement; it is not established as its cause. The retrieval-attribution blocker diagnostic instead counts non-elite rows ahead of missed elites under a reconstructed count-posterior reference, with a different denominator and reference; its percentages cannot be pooled with those from the assay-resolution diagnostic.

Reported HD3 and HD4 metrics were independently recomputed from preserved raw predictions. The exact source snapshot associated with the first HD4 execution is unavailable; however, a subse-

quently source-pinned package contains the same core configurations, predictions, training histories, scientific metrics, and HD4 sequence fingerprints. The shared units are used for provenance confirmation rather than treated as independent replications.

F Decision-alignment details

All 12 independent units use frozen ESM2-650M Δ features, ΔESM-WHT, GB1, ten epochs, and seeds $\{13, 42, 77, 123\}$. The three arms are point-only, focused observed-epsilon triplet, and epsilon-shuffled triplet with identical focused closure weights. All arms use the same 2,091 legal outer-train closures; only epsilon targets are permuted in the strict shuffle control. Test metrics were not accessed.

Table F.1: Independent validation-only four-point confirmation. The primary endpoint is TTK@10; NDCG and signed Spearman are secondary.

Arm	TTK@10	NDCG	Spearman
Point-only	0.125 ± 0.050	0.9157 ± 0.0041	0.6376 ± 0.0359
Focused observed-ϵ	0.125 ± 0.050	0.9213 ± 0.0037	0.6260 ± 0.0247
Strict shuffled-ϵ	0.100 ± 0.000	0.9127 ± 0.0015	0.6631 ± 0.0199

G Structured-coordinate geometry diagnostic

The completed diagnostic reproduces the geometry separation across three replicas. The raw pair representation has effective rank 899–905, whereas the Walsh pair representation has effective rank approximately 56.8. Reaching 90% explained variance requires about 810–815 raw components but only about 50 Walsh components, and the separation survives centering and PCA checks. In the HD=2 (double-mutant) matched-information control, the raw and Walsh pair coordinates have identical observed state support and are mutually affine up to numerical residuals of order 10^{-14}. The result therefore supports a coding-induced low-effective-dimensional geometry and inductive-bias interpretation, not additional biological information, genotype-space epistasis recovery, or a causal mechanism.

Orthogonal-reparameterization control. This control compares native ten-fold MULTI-evolve with the same model after a deterministic, invertible, same-width Hadamard-derived orthogonal channel transform on the frozen GB1 order-shift task. The CPU run gives $\rho = 0.463$ native versus 0.342 transformed; matched GPU confirmation gives 0.468 versus 0.344. NDCG and MSE also worsen, and TTK does not improve. The transformed condition therefore provides no evidence of an advantage from generic orthogonal reparameterization. This experiment is an orthogonal-reparameterization control only; it does not identify genotype Walsh coefficients or test biological epistasis recovery.

H Retrieval-attribution details

The main-text retrieval attribution (Section 5.6) evaluates GB1 HD4 predictions against the reconstructed count-posterior reference y^* using 500 independent count-posterior draws and five neural seeds per model. The reconstruction is a measurement-aware diagnostic reference, not a claim to recover unobserved fitness and not a training target. Appendix Figure H.1 shows one representative seed-0 ΔESM-MLP/ΔESM-WHT pair; the text reports five-seed ranges. Its oracle-truncated curves condition on the reconstructed high-fitness region and are diagnostic rather than deployable retrieval. Utility@10 is the normalized mean reconstructed fitness of a model's predicted top ten: 0 corresponds to the mean reference fitness across the candidate universe, and 1 to the reconstructed top-ten mean. The pre-specified diagnostic reliability gate required each low-order-minus-neural Utility@10 gap to be positive in at least 95% of the 500 count-posterior draws and to have a fifth percentile above 0.10; every comparison passed (minimum fifth-percentile gap 0.215).

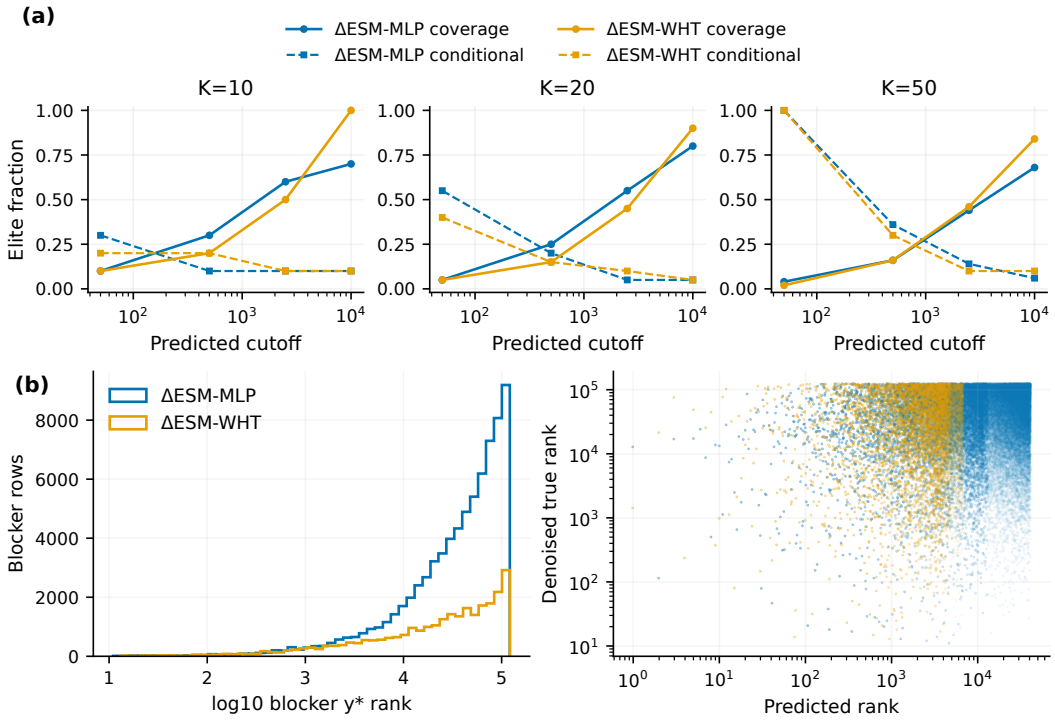


Figure H.1: GB1 retrieval diagnostics under reconstructed count-posterior reference y^*. (a) Coverage and oracle-truncated conditional-ranking curves across cutoffs for $K \in \{10, 20, 50\}$. (b) Blocker true-rank composition and predicted-versus-denoised true ranks. Panels show one representative seed-0 ΔESM-MLP/ΔESM-WHT pair; quantitative five-seed ranges in Section 5.6 summarize all five seeds. Oracle truncation is a diagnostic conditional on the reconstructed high-fitness region, not deployable retrieval.

Reference construction. For variant v, let $c_{v,s}$ and $c_{v,i}$ denote its selected and input read counts, and let $c_{WT,s}$ and $c_{WT,i}$ denote the corresponding wild-type counts. We use independent posteriors $\lambda_{u,q} \mid c_{u,q} \sim \text{Gamma}(c_{u,q} + 1/2, \text{rate} = 1)$ for $u \in \{v, WT\}$ and $q \in \{s, i\}$. The relative-enrichment reference is the posterior geometric mean

$$y_v^* = \exp\left(\mathbb{E}\left[\log \frac{\lambda_{v,s}/\lambda_{v,i}}{\lambda_{WT,s}/\lambda_{WT,i}} \,\middle|\, \mathbf{c}\right]\right).$$

The 500 independent posterior draws quantify count-measurement uncertainty; they are not candidate-resampling draws.

Blocker definition. For a missed elite $e \in \text{Top}_{10}(y^*) \setminus \text{Top}_{10}(s)$, its blockers are non-elites $j \notin \text{Top}_{10}(y^*)$ ranked ahead of it by model score s. Each (e, j) pair contributes one blocker row, so the same candidate can contribute multiple rows if it precedes multiple missed elites.

I WILD-TYPE ESM-2 PREFERENCE DIAGNOSTIC

For each wild-type/reference sequence x, we computed exact masked pseudo-log-likelihood (PLL), $\text{PLL}(x) = \sum_i \log P(x_i \mid x_{\setminus i})$, by masking one residue at a time and evaluating the masked-language-model head of `EsmForMaskedLM` at checkpoint `facebook/esm2_t33_650M_UR50D`. We report $\text{PLL}(x)/|x|$, the mean masked PLL per residue; this is distinct from ESM-LLR, the frozen ΔESM hidden-state encoding, and any supervised predictor. Wild-type/reference provenance and numerical quality checks passed for all four landscapes. The WT-PLL diagnostic was run after the primary model comparisons were frozen and was not used for architecture, hyperparameter, split, or test-set selection. It involved no retraining or new test-set model selection. The associated held-out full-point metrics were recovered from the existing frozen V4.5-PF-Refresh family; re-aggregation of its raw records reproduced the four-landscape macro results in Table 1.

The four wild types span -2.805 to -1.132 mean PLL per residue (Table I.1). Their associated held-out full-point performance pattern is mixed and non-monotonic: the least-preferred GFP wild type does not show uniformly poor ΔESM performance, and WHT-minus-MLP ranking differences change sign across landscapes. With only four landscapes, this descriptive PLM-preference covariate does not support an explanatory relationship with benchmark performance or a mechanism for ranking–retrieval misalignment.

Table I.1: Wild-type ESM-2 preference and held-out full-point performance by landscape. WT preference is exact masked pseudo-log-likelihood per residue under ESM-2 650M. Performance values come from the frozen five-seed V4.5-PF-Refresh full-point protocol underlying Table 1. Differences are ΔESM-WHT minus ΔESM-MLP means, computed before display rounding. The four-landscape pattern is mixed and non-monotonic; this table is descriptive and is not used for statistical attribution.

Landscape	WT mean PLL per residue	ΔESM-MLP Spearman	ΔESM-WHT Spearman	ΔSpearman	ΔNDCG	ΔTTK@10
GB1	-1.856	0.323	0.389	+0.067	+0.037	-0.080
GFP	-2.805	0.412	0.369	-0.043	-0.018	0.000
HIS7	-1.132	0.577	0.670	+0.093	+0.004	+0.020
FLIP2-Imine	-1.362	0.075	0.131	+0.056	+0.001	0.000