A General Harness for Protein Foundation Model Fitness Prediction

Anonymous authors
Paper under double-blind review

Abstract

Accurate fitness prediction is central to protein engineering and understanding sequence–function relationships. With advances in deep learning, protein foundation models (PFMs) have become widely used for this task. Recent analyses, however, show that these models share preferences reflecting their training corpora, while unreliable inputs can further distort fitness predictions. Family-specific evolutionary evidence and structural context can help address these limitations by providing complementary constraints on model scores, motivating REM-Harness (RH), a general, model-agnostic, training-free Retrieval-Enhanced Mutation harness. It fuses frozen model scores with multiple sequence alignment (MSA) evidence according to model uncertainty, then applies gated background correction and score shrinkage based on structural confidence and solvent exposure. Across 1,211 assays and 3.1 million measured variants from ProteinGym, VenusMutHub, and the newly curated viral benchmark ViroHub, all 71 configurations improve Spearman correlation on all 3 benchmarks by 0.073 on average, with broad gains across 5 metrics. Extended analyses relate retrieval gains to model–MSA preference differences, assess domain-level gains and immune-escape cases, and quantify computational speedups. Built with RH, VenusREM2 is the first to rank highest in all function, taxon, MSA-depth, and mutation-depth categories, with a ProteinGym Average Spearman of 0.556, 0.038 above the prior best.

1 Introduction

Single amino acid substitutions can alter stability, catalysis, recognition, and cellular fitness, making prediction central to variant interpretation and protein engineering (Zhou et al., 2024a; Notin et al., 2024). Deep mutational scanning measures these effects but covers few natural proteins and variants. Computational models scale prediction by ranking mutations from sequence or structure without assay labels (Meier et al., 2021; Frazer et al., 2021; Lin et al., 2023).

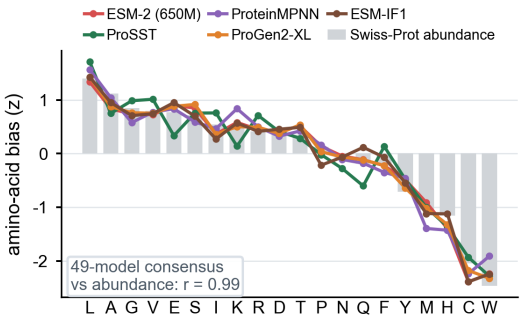


Figure 1: Shared protein-wide amino-acid background preferences across representative models. The 49-model consensus closely follows UniProtKB/Swiss-Prot amino-acid abundance (Pearson $r = 0.99$).

Protein foundation models (PFMs) and homologous sequences provide complementary views of evolutionary constraints. Pretrained models transfer information learned across large sequence collections (Rives et al., 2021), but their shared amino-acid preferences closely track corpus composition (Figure 1). This observation echoes the preference–performance relationship studied by Gordon et al. (2025). Homologs instead record residues tolerated at corresponding family sites (Hopf et al., 2017): an amino acid common across proteins may be disfavored locally. Fewer than 40% of MSA site profiles jointly align with the model and corpus backgrounds in each benchmark (Appendix D.1). Tranception and VenusREM improve frozen predictions using

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.

retrieved homologous sequences (Notin et al., 2022a; Tan et al., 2025b), but fixed weights from parameter search ignore family–corpus preference mismatch.

Homologous sequences describe evolutionary tolerance, while a substitution's physical consequences also depend on its structural environment (Zhou et al., 2024b). Structural inputs make this context explicit: mutations in buried cores are often more destabilizing than surface substitutions, whereas exposed surfaces participate in molecular recognition (Zhang et al., 2024; Beltran et al., 2025). The usefulness of this evidence, however, depends on its reliability (Tan et al., 2024). Prediction accuracy varies across residues, and a single conformation may not represent an assay's molecular environment; structural descriptors and conditioned scores can inherit these limitations (Jumper et al., 2021; Terwilliger et al., 2024). RSALOR weights conservation by accessibility; S3F uses surface geometry and sequence fallback at low confidence (Zhang et al., 2024; Tsishyn et al., 2025). Their weighting and fallback rules remain uncoupled from PFM bias correction.

These observations motivate integrating learned preferences with family constraints while accounting for structural reliability. *How can a general harness integrate evolutionary and structural evidence to mitigate model bias and account for uncertainty in protein fitness prediction?*

Inspired by large language model (LLM) harnesses that coordinate inference for reliable deployment (DeepSeek-AI, 2026), we introduce REM-HARNESS (RH), a training-free Retrieval-Enhanced Mutation harness that provides **a unified and efficient inference layer** around frozen PFMs for fitness prediction. The harness comprises three complementary components. Adaptive MSA fusion derives protein-specific, scale-aware mixing weights from model uncertainty, wild-type support, and model–MSA agreement. Coherence-gated calibration corrects shared model bias while preserving family evidence. Structural attenuation shrinks scores at sites with above-average solvent exposure or prediction uncertainty within each protein, without reversing mutation contrasts.

We evaluate the harness across 71 configurations on ProteinGym (Notin et al., 2024), VenusMutHub (Zhang et al., 2025), and **ViroHub, a newly curated large-scale benchmark for mutation-effect prediction in viral proteins**. Together, these benchmarks contain $1{,}211$ assays and approximately 3.1 million measured variants. All configurations improve Spearman correlation, with mean gains of 0.068, 0.056, and 0.094 on the three benchmarks, respectively. VENUSREM2 combines the harness with 6 variants of PROSST, a structure-aware protein language model with strong zero-shot mutation-effect prediction performance (Li et al., 2024). It achieves an **official ProteinGym Average Spearman of $\mathbf{0.556}$**, exceeding the strongest public baseline in our comparison (0.518) by 0.038, and ranks first in all five functional categories (Figure 3). Extended ablations and domain analyses then examine component contributions and variation in gains across regions and phenotypes, using immune-escape cases as a biological consistency check. Runtime analyses further show $20.7\times$ faster mutation lookup and $62.3\times$ faster MSA construction through native-field reuse.

2 REM-HARNESS: A GENERAL FRAMEWORK FOR FITNESS PREDICTION

Family retrieval supplies local constraints; background calibration adjusts shared native preferences while preserving family evidence. Structural exposure and prediction confidence then modulate site contributions within a common interface across frozen backbones (Figure 2).

2.1 A COMMON SCORE-FIELD INTERFACE

Let $\mathcal{A}$ be the 20 amino acids and $x = (x_1, \ldots, x_L)$ the wild-type sequence. A backbone with frozen parameters θ supplies a score field $R \in \mathbb{R}^{L\times 20}$. Writing its output probability for residue a at site i in context C_i as $q_{\theta,i}(a; C_i)$ gives

$$R^{\mathrm{mask}}_{i,a} = \log q_{\theta,i}(a; x_{\setminus i}), \quad R^{\mathrm{wt}}_{i,a} = \log q_{\theta,i}(a; x), \quad R^{\mathrm{tf}}_{i,a} = \log q_{\theta,i}(a; x_{<i}). \tag{1}$$

Here, $x_{\setminus i}$ denotes x with position i masked, and $x_{<i}$ denotes the prefix preceding position i. Other native fields retain interface-specific scores, including unnormalized or window-aggregated outputs. Required structure enters C_i; C_i and frozen θ stay fixed within each Raw–RH pair.

Let $\mathcal{I}$ contain n query-mapped sites with standard wild-type residues, and $n_{i,v}$ count tokenizer symbol v in retained rows. Equal row weights and no pseudocounts define aligned frequencies and

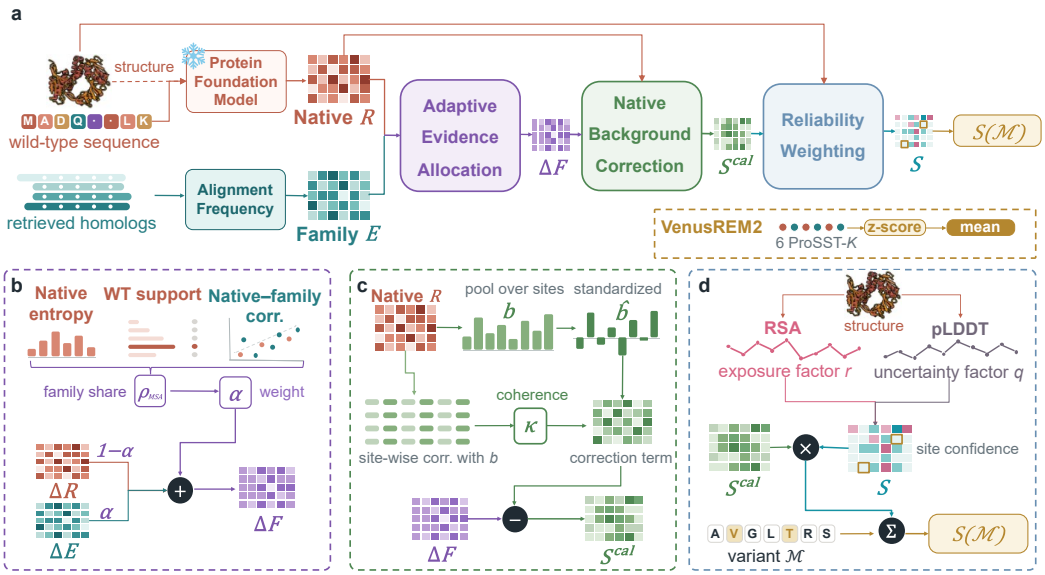


Figure 2: Framework of REM-HARNESS. **(a)** Frozen PFM scores are refined using family evidence and structural context. **(b)** Adaptive MSA fusion adjusts the family contribution to model uncertainty and model–MSA agreement. **(c)** Coherence-gated calibration corrects model background bias while preserving family evidence. **(d)** Solvent exposure and structural confidence guide attenuation at each site before substitution contributions are aggregated into a variant score.

a log-softmax family field; either field X uses wild-type-relative contrasts:

$$f_{i,v} = \frac{n_{i,v}}{\sum_{u\in\mathcal{V}} n_{i,u}}, \qquad E_{i,a} = f_{i,a} - \log \sum_{v\in\mathcal{V}} e^{f_{i,v}}, \quad i \in \mathcal{I}, \tag{2}$$

$$\Delta^X_{i,a} = X_{i,a} - X_{i,x_i}. \tag{3}$$

$\mathcal{V}$ is the tokenizer vocabulary, including padding; E retains amino-acid columns. The site-wise log-normalizer cancels in contrasts. Equation 8 handles relative scales. Set $E_i = R_i$ outside $\mathcal{I}$ so retrieval leaves unmapped sites unchanged.

2.2 ADAPTIVE RETRIEVAL AS SCALE ALLOCATION

A fixed mixing weight gives different effective family contributions across backbones. To allocate the complementary homolog evidence used by TRANCEPTION and VENUSREM (Notin et al., 2022a; Tan et al., 2025b), we distinguish a diagnostic-based target share from its scale-adjusted coefficient. On mapped sites, normalize the native field as

$$p_i(a) = \frac{\exp R_{i,a}}{\sum_{a'\in\mathcal{A}} \exp R_{i,a'}}, \qquad i \in \mathcal{I}. \tag{4}$$

With $\mathrm{clip}(z, 0, 1) = \min\{1, \max\{0, z\}\}$, site entropy H_i and its normalized mean $\bar{H}$ summarize uncertainty:

$$H_i = -\sum_{a\in\mathcal{A}} p_i(a) \log p_i(a), \qquad \bar{H} = \mathrm{clip}\left(\frac{1}{n \log 20} \sum_{i\in\mathcal{I}} H_i, 0, 1\right). \tag{5}$$

WT support π and model–family Pearson alignment $r_{R,E}$ give two normalized retrieval cues:

$$\pi = \frac{1}{n} \sum_{i\in\mathcal{I}} \log p_i(x_i), \qquad \rho_\pi = \mathrm{clip}(-\pi / \log 20, 0, 1), \tag{6a}$$

$$r_{R,E} = \frac{1}{n} \sum_{i\in\mathcal{I}} \mathrm{corr}_{a\in\mathcal{A}}(R_i, E_i), \qquad \rho_{\mathrm{rot}} = \frac{1 - r_{R,E}}{2}. \tag{6b}$$

Correlations use the 20 amino-acid entries, with zero assigned to constant profiles. We set the target family share to

$$\rho_{\mathrm{MSA}} = (1 - \bar{H})\rho_\pi + \bar{H}\rho_{\mathrm{rot}}. \tag{7}$$

Low-entropy fields emphasize wild-type support; high-entropy fields emphasize model–family disagreement. This bounded interpolation is a design choice tested in Section 3.3; its conversion to a mixing coefficient follows from the scale constraint below (Appendix F.2).

Let $s_R, s_E > 0$ be sample standard deviations over all $19n$ non-WT contrasts on $\mathcal{I}$. Requiring the family channel to supply a fraction ρ_{MSA} of the summed weighted scales gives

$$\rho_{\mathrm{MSA}} = \frac{\alpha s_E}{(1-\alpha)s_R + \alpha s_E} \quad \Longrightarrow \quad \alpha = \frac{\rho_{\mathrm{MSA}} s_R}{\rho_{\mathrm{MSA}} s_R + (1 - \rho_{\mathrm{MSA}}) s_E}. \tag{8}$$

The resulting field and contrast are

$$F_{i,a} = (1-\alpha)R_{i,a} + \alpha E_{i,a}, \qquad \Delta^F_{i,a} = (1-\alpha)\Delta^R_{i,a} + \alpha \Delta^E_{i,a}. \tag{9}$$

2.3 Coherence-gated calibration of the native channel

Calibration adjusts the shared native background according to its coherence with site-level preferences, preserving family evidence. The preference perspective of Gordon et al. (2025) motivates this direction; gated and ungated variants test the rule (Section 3.4).

We summarize the background by the log-mean-exp profile of R. Let $\bar{b}$ and s_b be its mean and sample standard deviation across the 20 amino acids:

$$b_a = \log\left(\frac{1}{L}\sum_{i=1}^{L} e^{R_{i,a}}\right), \qquad \widehat{b}_a = \frac{b_a - \bar{b}}{s_b}. \tag{10}$$

Set $\widehat{b} = 0$ for constant backgrounds (Appendix F.3). With $\mathrm{ReLU}(z) = \max\{0, z\}$, mean site-wise coherence gives the gate:

$$c = \frac{1}{L}\sum_{i=1}^{L} \mathrm{corr}_{a\in\mathcal{A}}(R_i, b), \qquad \kappa(R) = \mathrm{ReLU}(c). \tag{11}$$

When their mean alignment is non-positive, $\kappa(R) = 0$ and no global correction is imposed. Otherwise, calibration removes the protein-wide direction only from the native channel:

$$\begin{aligned} S^{\mathrm{cal}}_{i,a} &= (1-\alpha)\big[\Delta^R_{i,a} - \kappa(R)(\widehat{b}_a - \widehat{b}_{x_i})\big] + \alpha\Delta^E_{i,a} \\ &= \Delta^F_{i,a} - (1-\alpha)\kappa(R)(\widehat{b}_a - \widehat{b}_{x_i}). \end{aligned} \tag{12}$$

The factor $1 - \alpha$ preserves the family contribution while calibrating the native background.

2.4 Structure-adjusted output and variant scoring

Beyond accessibility-aware conservation and confidence-dependent fallback (Zhang et al., 2024; Tsishyn et al., 2025), we jointly constrain calibrated scores by exposure and uncertainty relative to protein-specific means (Section 3.4; Table 1).

Let RSA r_i and normalized pLDDT ℓ_i lie in $[0, 1]$, with uncertainty $q_i = 1 - \ell_i$. Define $\bar{r}$ over positive RSA entries and $\bar{q} = 1 - \bar{\ell}$ from mean positive pLDDT. The site weight is

$$\lambda_i = \big(1 - \mathrm{ReLU}(r_i - \bar{r})\big)\big(1 - \mathrm{ReLU}(q_i - \bar{q})\big). \tag{13}$$

Only above-mean exposure or uncertainty attenuates a site, and neither factor can reverse a contrast. The complete readout is

$$S_{i,a} = \lambda_i S^{\mathrm{cal}}_{i,a} = \lambda_i\Big[\Delta^F_{i,a} - (1-\alpha)\kappa(R)(\widehat{b}_a - \widehat{b}_{x_i})\Big]. \tag{14}$$

Missing covariates contribute one (Appendix F.4). For a variant with substitutions $\mathcal{M} = \{(i, a) : x_i \to a\}$, we sum fixed-field main effects:

$$S(\mathcal{M}) = \sum_{(i,a)\in\mathcal{M}} S_{i,a}. \tag{15}$$

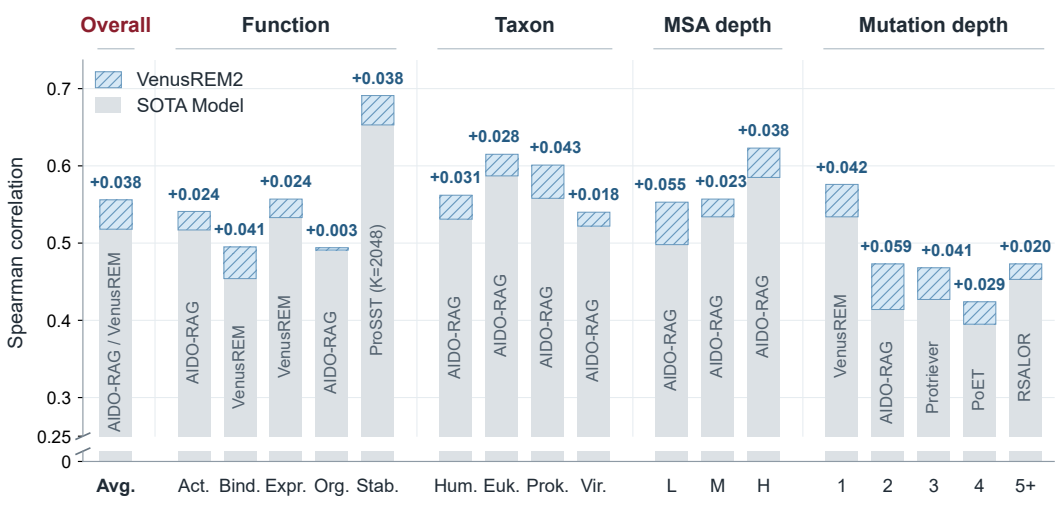


Figure 3: ProteinGym Spearman overall and across 17 strata (Appendix Tables 14–18). Gray: strongest comparator (SOTA Model); blue hatching and labels: VENUSREM2 gains. Vertical labels identify comparators. Avg.: overall; Act./Bind./Expr./Org./Stab.: activity/binding/expression/organismal fitness/stability; Hum./Euk./Prok./Vir.: human/other eukaryotes/prokaryotes/viruses; L/M/H: low/medium/high MSA depth. Mutation depth counts substitutions per variant (5+: at least five). AIDO-RAG and PoET denote AIDO Protein-RAG (16B) and PoET (200M); AIDO-RAG and VenusREM tie overall at reported precision.

VENUSREM2 uses six PROSST vocabularies, $k \in \mathcal{K} = \{20, 128, 512, 1024, 2048, 4096\}$. Member k gives the harness score $S_k(v)$ using α_k. For assay A's candidates $\mathcal{V}_A$, the ensemble is:

$$S_{\text{VenusREM2}}(v; A) = \frac{1}{6} \sum_{k \in \mathcal{K}} \frac{S_k(v) - \mu_{k,A}}{\sigma_{k,A}}. \tag{16}$$

Here $\mu_{k,A}, \sigma_{k,A}$ are means and population standard deviations of predictions on $\mathcal{V}_A$; constant members contribute zero (Appendix F.5).

3 RESULTS

3.1 EXPERIMENTAL SETUP

Benchmarks and paired evaluation. We evaluate ProteinGym (217 substitution assays), VenusMutHub (905 assays; 527 wild-type clusters), and ViroHub (89 assays; 616,690 substitutions), whose experimental measurements do not overlap the other benchmarks. Spearman is primary; NDCG, AUC, MCC, and top-variant recall assess complementary outcomes. Aggregation uses ProteinGym's official Average Spearman, VenusMutHub assay-macro means, and the ViroHub phenotype–protein hierarchy. Appendices B.1 and B.3 detail metrics, valid assay sets, uncertainty, and curation. The 71 configurations span masked and autoregressive language models, structure-aware encoders, and inverse-folding models. Raw–REM-HARNESS pairs fix checkpoints, structural inputs, and native scoring interfaces to isolate the readout (Appendix C).

3.2 CONSISTENT GAINS ACROSS ARCHITECTURES AND BENCHMARKS

Gains across architectures and evaluation metrics. REM-HARNESS improves Spearman for all 71 configurations on all three benchmarks, with benefits extending to discrimination and top-variant recovery. Mean Spearman increases by 0.067 on ProteinGym, 0.056 on VenusMutHub, and 0.094 on ViroHub, under their respective aggregation protocols. Across all five metrics, every one of the 355 model–metric pairs improves on ProteinGym and ViroHub; 353/355 improve or tie at reported precision on VenusMutHub (Figure 4; Appendix I.3). Gains are largest for autoregressive models on ProteinGym and VenusMutHub and masked models on ViroHub, while structure-aware and inverse-folding models also benefit (Appendix E).

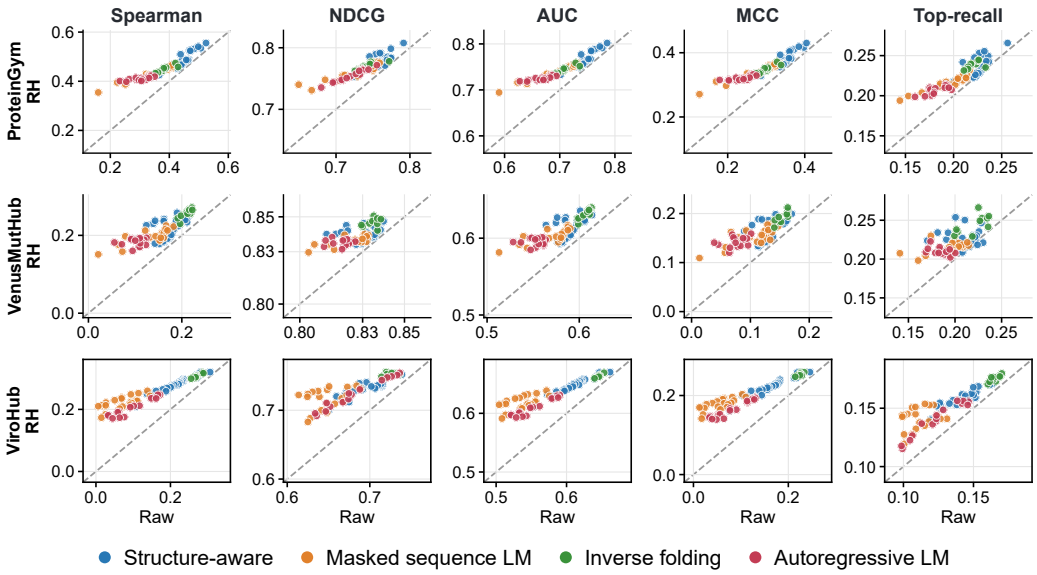


Figure 4: Raw–RH results across five metrics on ProteinGym, VenusMutHub, and ViroHub (top to bottom). Points represent 71 configurations; those above the diagonal indicate gains.

State-of-the-art performance on ProteinGym. Built with REM-HARNESS, the six-member VENUSREM2 ensemble achieves a new state of the art in the official ProteinGym comparison, reaching Average Spearman 0.556, a 0.038 gain over the strongest listed baseline (0.518). It ranks first across all 17 function, taxon, MSA-depth, and mutation-depth strata (Figure 3; Appendix Tables 14–18). The 0.031 gain over the same uncalibrated ensemble isolates improvement beyond model averaging. Functional coverage also extends beyond this ensemble: all 71 configurations improve in all five ProteinGym functional categories ($355/355$ comparisons). The improvement across MSA-depth bins (Appendix C.1), with larger gains at low depth, supports applying the harness to sparsely sampled protein families as well as homolog-rich targets.

3.3 GLOBAL–LOCAL COMPLEMENTARITY AND RETRIEVAL GAINS

Retrieval gains increase with preference separation. Across 148 ProteinGym proteins scored with ESM-2 wild-type readouts, the fraction of globally–locally separated sites correlates with retrieval gain ($\rho = 0.338$; bootstrap interval 0.179–0.481; Figure 5; Appendix D). Proteins in the highest separation quartile gain 0.058 more than those in the lowest (interval 0.033–0.086). Adjusting for sequence length, MSA depth, occupancy, and Raw performance leaves partial $\rho = 0.261$. These results support preference separation as a diagnostic of where family-specific evolutionary evidence can complement pretrained model scores across the evaluated proteins.

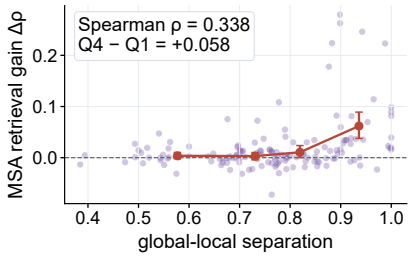


Figure 5: Separation and retrieval gain (148 ProteinGym proteins). Red: quartile means and 95% bootstrap intervals; faint: proteins.

Integration outperforms either evidence source alone. On the 21-assay PROSST-2048 validation split, family evidence alone reaches Spearman 0.4093, the native model 0.5573, and adaptive integration 0.5774. A coefficient sweep shows a broad performance plateau over $\alpha \in [0.6, 0.9]$, which the adaptive readout matches or exceeds (Appendix F.2). The plateau persists when fold means are averaged over all ten folds of the partition, supporting flexible evidence allocation without requiring precise optimization of a single fixed mixing coefficient.

Table 1: PROSST-2048 ablation. PG (ProteinGym) validation/VMH (VenusMutHub): assay-macro; PG full: official Average Spearman; VH (ViroHub): mutant-only phenotype–protein hierarchical Spearman. Δ_{Raw} reports absolute score changes (three decimals). w/o Ent. Weight. fixes $\alpha = 0.8$; w/o Coh. Gating fixes $\kappa = 1$, while w/ Coh. Gating uses $\kappa(R)$.

Component	Setting	PG Validation (21)		PG Full (217)		VMH Full (905)		VH (89)	
		Score	Δ_{Raw}	Score	Δ_{Raw}	Score	Δ_{Raw}	Score	Δ_{Raw}
Raw	Native field	0.5573	—	0.5069	—	0.2126	—	0.2709	—
MSA fusion	w/o Ent. Weight.	0.5742	+0.017	0.5205	+0.014	**0.2341**	**+0.022**	0.2897	+0.019
	w/ Ent. Weight.	**0.5774**	**+0.020**	**0.5246**	**+0.018**	0.2336	+0.021	**0.2901**	**+0.019**
Model calibration	w/o Coh. Gating	0.5641	+0.007	0.5219	+0.015	0.2326	+0.020	0.2841	+0.013
	w/ Coh. Gating	**0.5788**	**+0.022**	**0.5284**	**+0.022**	**0.2369**	**+0.024**	**0.2942**	**+0.023**
Structural attenuation	w/ RSA, w/o pLDDT	0.5826	+0.025	**0.5329**	**+0.026**	**0.2499**	**+0.037**	**0.3085**	**+0.038**
	w/o RSA, w/ pLDDT	**0.5833**	**+0.026**	**0.5329**	**+0.026**	0.2352	+0.023	0.2994	+0.028
REM-HARNESS	Full pipeline	**0.5849**	**+0.028**	**0.5343**	**+0.027**	0.2491	+0.037	**0.3102**	**+0.039**

3.4 CONTRIBUTIONS OF RETRIEVAL, CALIBRATION, AND SHRINKAGE

Later stages add gains beyond retrieval. Retrieval supplies 0.0201 of the 0.0276 PROSST-2048 validation gain; calibration and structural attenuation contribute the remaining 0.0075 (Table 1). All successive stages also improve on the 196 held-out ProteinGym assays, raising official Average Spearman from 0.5072 to 0.5337 and extending the component pattern beyond the operator-selection split. Across all 71 configurations, retrieval accounts for approximately 50%, 55%, and 36% of gains on ProteinGym, VenusMutHub, and ViroHub, respectively, showing that the relative contributions depend on the benchmark (Appendix F).

Gating improves calibration consistency. On ProteinGym, ungated correction improves on retrieval in $43/71$ configurations, versus $71/71$ with gating; mean gains are 0.003 and 0.012 (Table 26). Gating also exceeds ungated correction in $37/71$ VenusMutHub and $68/71$ ViroHub configurations. Excluding the scored site preserves the gate ($\rho = 0.985$; median absolute change 0.010), supporting stability against direct self-inclusion. Together, these checks support improved calibration and stable protein-level control (Figure 14; Appendix F).

Exposure and confidence provide complementary context. Combining the two covariates gives the strongest PROSST-2048 result on ProteinGym and ViroHub (Table 1). On calibrated ProteinGym fields, RSA-only and pLDDT-only shrinkage both reach 0.5329, increasing to 0.5343 when combined. On ViroHub, the combination reaches 0.3102, above RSA alone (0.3085) and pLDDT alone (0.2994). Exposure accounts for most of the structural gain on VenusMutHub, where RSA alone reaches 0.2499 and the combination 0.2491. Both covariates contribute, with benchmark-specific strengths.

3.5 PERFORMANCE HETEROGENEITY ACROSS VIRAL PHENOTYPES

Strong gains in protein expression and activity. Across 71 configurations, mean gains on ViroHub are largest for expression ($+0.194$) and activity ($+0.150$), and smaller for immune escape ($+0.009$) and stability ($+0.007$; Figure 6a). The gain pattern is consistent with the harness refining constraints relevant to protein production and function. The 21 escape assays across 12 viral protein backgrounds provide a complementary biological test: their best mean result is 0.046, motivating analysis of the antibody-specific selection they measure.

Escape assays reveal distinct selective pressures. Among 21 escape assays, 11 show positive signal (up to $+0.21$), 6 remain near zero, and 4 correlate negatively (down to -0.30). Negative cases map to exposed antigenic/receptor-binding regions of SARS-CoV-2 RBD and antibody-contact regions of ZIKV E. Frozen backbones score many escape substitutions as deleterious (Figure 6b); per-site escape and model tolerance anti-correlate, reaching -0.19 for structure-aware models.

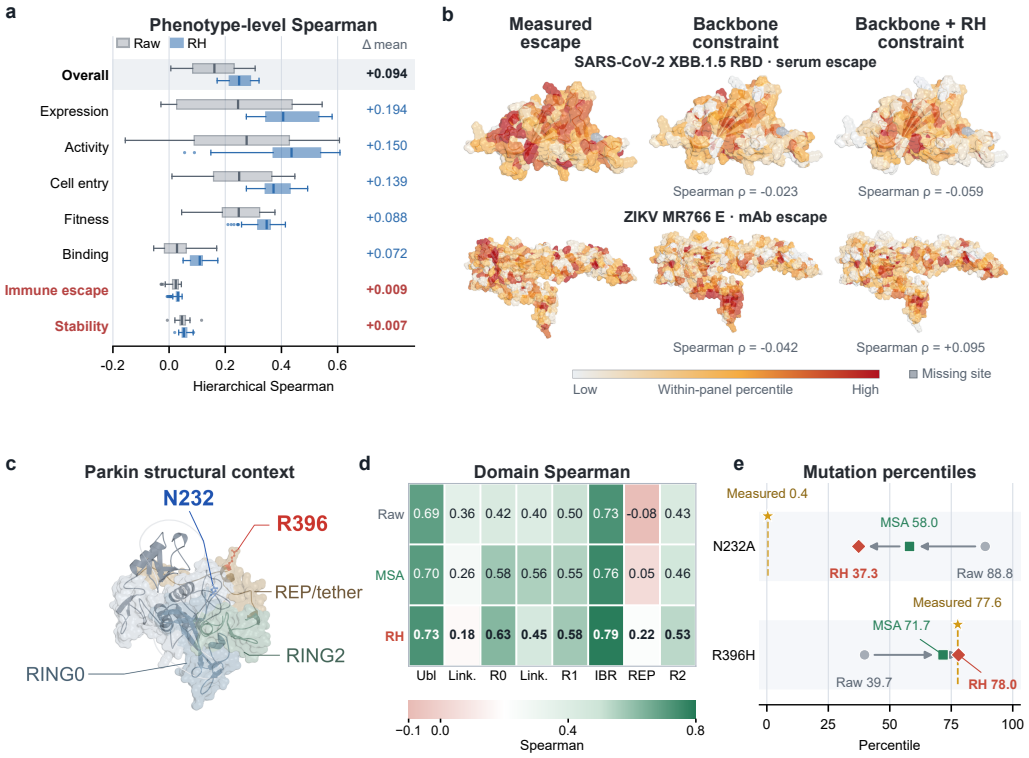


Figure 6: **(a)** Raw/RH distributions by phenotype across 71 configurations (medians, IQRs, 1.5× IQR whiskers, and outliers). Numbers indicate mean gains; red marks unresolved cases. **(b)** Measured escape and predicted constraint for SARS-CoV-2 XBB.1.5 RBD and ZIKV MR766 E; colors show within-panel percentiles and gray marks missing sites. **(c)** Parkin N232 and R396 in structural context. **(d)** Domain-wise Spearman. **(e)** N232A and R396H approach measured ranks.

This localization is consistent with distinct selective pressures: substitutions disfavored by native-protein constraints can evade antibody recognition. The cases link phenotype differences to selection context, motivating antibody- or serum-specific evidence.

3.6 Ranking gains across Parkin domains

Domain-resolved gains support structural interpretation. In the Parkin VAMP-seq abundance landscape (Clausen et al., 2024), Raw, retrieval, and RH reach Spearman 0.503, 0.585, and 0.645 across 8,756 substitutions at 462 sites. The full readout improves folded domains, including RING0 (0.425 to 0.627) and RING2 (0.427 to 0.529; Figure 6d). A cross-protein audit using pLDDT 50 to partition regions finds improvements in 83% of 214 qualifying ordered-region assays (mean $\Delta = +0.056$). This cross-protein agreement supports the domain analysis; Appendix G reports complementary linker and low-confidence-region comparisons.

Corrections address both overestimation and underestimation. N232A, at the start of RING1, is measured at the 0.4th percentile but ranked at 88.8 by Raw; retrieval and RH lower it to 58.0 and 37.3. R396H in REP/tether instead rises from 39.7 to 71.7 and 78.0, approaching its measured 77.6th percentile (Figure 6c,e). These opposite shifts show that the harness can both deprioritize low-abundance substitutions and recover variants with relatively high measured abundance. The same landscape also improves in abundance-class AUC, from 0.769 for Raw to 0.815 with retrieval and 0.857 with the full harness (Appendix G). Together, the rank corrections and class-level separation link improved correlation to more useful candidate prioritization across the abundance landscape.

3.7 SCORING INTERFACES, EFFICIENCY, AND INTERACTION RECOVERY

Gains extend across native scoring interfaces. Masked and wild-type fields both benefit from RH across eight paired checkpoints on all three benchmarks. Full-sequence and native-field scoring are nearly equivalent for ESM-IF1 (0.422 versus 0.421 on ProteinGym), while native teacher forcing raises PROTEINMPNN mean assay Spearman from 0.271–0.290 to 0.376–0.401 across eight checkpoints. At v_48_020, 204 of 217 assays improve, showing that the ProteinMPNN gain extends across proteins rather than being driven by a few large effects (Appendices H.2–H.3).

Field reuse amortizes variant-scoring cost. For PROTEINMPNN, reusing one native field per protein reduces wall-clock cost by 219–601× relative to rebuilding mutant sequences. Separately, after a shared PROSST forward pass, indexed mutation lookup is 20.7× faster and vectorized MSA count-matrix construction 62.3× faster than their loop-based implementations (Figure 7; Appendix H.1). Readout construction costs $O(20L)$ per protein; each M-substitution variant then requires only $O(M)$ lookup and summation, without repeating backbone inference.

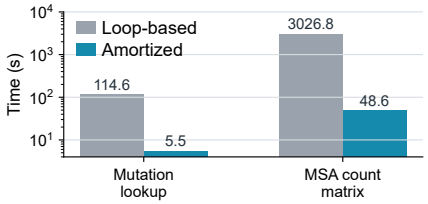


Figure 7: Post-inference runtime on ProteinGym (2.47 million variants): loop-based versus amortized lookup and MSA construction.

Main effects support efficient multi-mutant ranking. On Olson GB1, teacher forcing raises single- and double-mutant Spearman from 0.105 to 0.306 and 0.165 to 0.284. Because a fixed native field is additive, we also rescore substitutions on their partner's mutant background to test interaction recovery. Across 535,917 complete quartets, conditional epistasis correlation is 0.012, versus 0.005 for official scoring; both bootstrap intervals include zero (Appendix H.4). Excluding single or double mutants at the assay floor leaves 400,118 cases: double-mutant Spearman is 0.292 versus 0.152, but conditional epistasis correlation remains only 0.022 (Table 31). The ranking advantage is thus primarily supported by reusable main-effect information.

4 RELATED WORK

Protein models and adaptation. EVMUTATION, GEMME, and EVE infer family constraints (Hopf et al., 2017; Laine et al., 2019; Frazer et al., 2021), complementing pretrained sequence, inverse-folding, and structure-aware models (Rives et al., 2021; Hsu et al., 2022; Li et al., 2024). TRANCEPTION, POET, and VENUSREM incorporate homolog evidence (Notin et al., 2022a; Truong Jr & Bepler, 2023; Tan et al., 2025b); S3F and RSALOR exploit structural context (Zhang et al., 2024; Tsishyn et al., 2025), while homolog fine-tuning addresses likelihood-dependent performance (Gordon et al., 2025).

Evaluation resources. MaveDB archives variant-effect assays (Esposito et al., 2019), and FLIP and ProteinGym benchmark generalization and mutation-effect prediction (Dallago et al., 2021; Notin et al., 2024). PEER and VENUSFACTORY cover broader protein tasks (Xu et al., 2022; Tan et al., 2025a); ViroBench and ViroGym extend evaluation to viral nucleotide and protein applications (Ye et al., 2026; Zhou et al., 2026). Appendix A provides the extended review.

5 DISCUSSION AND CONCLUSION

While LLM harnesses advance into practice, understanding and harnessing biological foundation models remain less explored. REM-HARNESS addresses this gap with training-free evidence integration that improves frozen PFM predictions and enables efficient variant evaluation.

Future work could learn family constraints through trainable MSA representations and fusion, and incorporate protein–protein interaction (PPI) and protein–small-molecule inputs. These extensions would align the supplied context with recognition-dependent phenotypes and test whether mutation-dependent representations improve interaction modeling.

AI USE STATEMENT

In this work, generative AI tools (OpenAI Codex) were used to refine the conceptual framing and hypotheses; to provide feedback on methodology and experimental design; to implement methods; to reformat data; to interpret and present experimental analyses; and to assist with translation. They were not used to generate synthetic datasets; they also assisted in checking algebraic derivations and their assumptions. Additionally, they were used to create and edit scientific figures, to edit code, to search and summarize literature, and to structure, draft, and language-edit the manuscript. The authors reviewed all AI-assisted work, verified reported results against the underlying evaluation artifacts and citations against their primary sources, and take full responsibility for the final text, claims, code, and figures.

ETHICS STATEMENT

This study uses public protein sequence, structure, alignment, and deep-mutational-scanning benchmark data and involves no human participants or personal data. Mutation-effect scores are computational hypotheses that require experimental validation; reuse should follow the licenses and terms of the original datasets and pretrained models.

REPRODUCIBILITY STATEMENT

Section 2 defines the readout and scoring equations. Section 3.1 and Appendix B specify the benchmarks, data provenance, evaluation, validation, and uncertainty estimation. Appendices C–I detail baseline implementations, global–local diagnostics and their cross-benchmark replication, complete results, component ablations, scoring interfaces, runtime measurements, and per-configuration analyses. The source code and data are available at the links below. The repository documents environment setup, optional backbone dependencies, evaluation scripts, and experiment configurations; Appendix H separates scoring-interface, runtime, and epistasis checks. Appendix J describes installation, command-line and dashboard workflows, and interactive score inspection.

GitHub: https://anonymous.4open.science/r/REM-Harness-9C1C
Hugging Face: https://huggingface.co/datasets/anonymous-researcher-123/REM-Harness

REFERENCES

Arjun K. Aditham, Caelan E. Radford, Caleb R. Carr, Naveen Jasti, Neil P. King, and Jesse D. Bloom. Deep mutational scanning of rabies glycoprotein defines mutational constraint and antibody-escape mutations. *Cell Host & Microbe*, 33(6):988–1003.e10, June 2025. ISSN 1931-3128. doi:10.1016/j.chom.2025.04.018. URL http://dx.doi.org/10.1016/j.chom.2025.04.018.

Jenny J Ahn, Timothy C Yu, Bernadeta Dadonaite, Caelan E Radford, and Jesse D Bloom. Influenza hemagglutinin subtypes have different sequence constraints despite sharing extremely similar structures. *Virus Evolution*, 12(1), 2026. ISSN 2057-1577. doi:10.1093/ve/veag018. URL http://dx.doi.org/10.1093/ve/veag018.

Beatriz Álvarez-Rodríguez, Sebastian Velandia-Álvarez, Christina Toft, and Ron Geller. Mapping mutational fitness effects across the coxsackievirus b3 proteome reveals distinct profiles of mutation tolerability. *PLOS Biology*, 22(7):e3002709, July 2024. ISSN 1545-7885. doi:10.1371/journal.pbio.3002709. URL http://dx.doi.org/10.1371/journal.pbio.3002709.

William Bakhache, Walker Symonds-Orr, Lauren McCormick, and Patrick T. Dolan. Deep mutation, insertion and deletion scanning across the enterovirus a proteome reveals constraints shaping viral evolution. *Nature Microbiology*, 10(1):158–168, November 2024. ISSN 2058-5276. doi:10.1038/s41564-024-01871-y. URL http://dx.doi.org/10.1038/s41564-024-01871-y.

Antoni Beltran, Xiang'er Jiang, Yue Shen, and Ben Lehner. Site-saturation mutagenesis of 500 human protein domains. *Nature*, 637(8047):885–894, 2025. doi:10.1038/s41586-024-08370-4. URL https://doi.org/10.1038/s41586-024-08370-4.

Aadyot Bhatnagar, Sarthak Jain, Joel Beazer, Samuel C. Curran, Alexander M. Hoffnagle, Kyle Shan Ching, Michael Martyn, Stephen Nayfach, Jeffrey A. Ruffolo, and Ali Madani. Scaling unlocks broader generation and deeper functional understanding of proteins. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=yvGL2HP7pU.

Yunlong Cao, Fanchong Jian, Jing Wang, Yuanling Yu, Weiliang Song, Ayijiang Yisimayi, Jing Wang, Ran An, Xiaosu Chen, Na Zhang, Yao Wang, Peng Wang, Lijuan Zhao, Haiyan Sun, Lingling Yu, Sijie Yang, Xiao Niu, Tianhe Xiao, Qingqing Gu, Fei Shao, Xiaohua Hao, Yanli Xu, Ronghua Jin, Zhongyang Shen, Youchun Wang, and Xiaoliang Sunney Xie. Imprinted sars-cov-2 humoral immunity induces convergent omicron rbd evolution. *Nature*, 614(7948):521–529, December 2022. ISSN 1476-4687. doi:10.1038/s41586-022-05644-7. URL http://dx.doi.org/10.1038/s41586-022-05644-7.

Caleb Carr and Jesse D. Bloom. Deep mutational scanning of lassa virus gp for escape from columbia/vrc human sera. https://github.com/dms-vep/LASV_Josiah_GP_DMS_Columbia_VRC_sera, 2026. Unpublished dataset; source data repository, commit dabb03a4eca8a33badfc84b71ff52ba057d23616 (accessed 5 September 2026).

Caleb R. Carr, Katharine H.D. Crawford, Michael Murphy, Jared G. Galloway, Hugh K. Haddox, Frederick A. Matsen, Kristian G. Andersen, Neil P. King, and Jesse D. Bloom. Deep mutational scanning reveals functional constraints and antibody-escape potential of lassa virus glycoprotein complex. *Immunity*, 57(9):2061–2076.e11, September 2024. ISSN 1074-7613. doi:10.1016/j.immuni.2024.06.013. URL http://dx.doi.org/10.1016/j.immuni.2024.06.013.

L. Clausen, V. Voutsinos, M. Cagiada, K. E. Johansson, M. Grønbæk-Thygesen, S. Nariya, R. L. Powell, M. K. N. Have, V. H. Oestergaard, A. Stein, D. M. Fowler, K. Lindorff-Larsen, and R. Hartmann-Petersen. A mutational atlas for parkin proteostasis. *Nature Communications*, 15: 1541, 2024. doi:10.1038/s41467-024-45829-4.

Bernadeta Dadonaite and Jesse D. Bloom. Deep mutational scanning of sars-cov-2 omicron ba.2 spike for escape from crowe lab monoclonal antibodies. https://github.com/dms-vep/SARS-CoV-2_Omicron_BA.2_spike_DMS_Crowe_mAbs, 2026. Unpublished dataset; source data repository, commit 2259fd5dfbc067d2122d8e517f980147367bedd5 (accessed 5 September 2026).

Bernadeta Dadonaite, Katharine H.D. Crawford, Caelan E. Radford, Ariana G. Farrell, Timothy C. Yu, William W. Hannon, Panpan Zhou, Raiees Andrabi, Dennis R. Burton, Lihong Liu, David D. Ho, Helen Y. Chu, Richard A. Neher, and Jesse D. Bloom. A pseudovirus system enables deep mutational scanning of the full sars-cov-2 spike. *Cell*, 186(6):1263–1278.e20, March 2023. ISSN 0092-8674. doi:10.1016/j.cell.2023.02.001. URL http://dx.doi.org/10.1016/j.cell.2023.02.001.

Bernadeta Dadonaite, Jenny J. Ahn, Jordan T. Ort, Jin Yu, Colleen Furey, Annie Dosey, William W. Hannon, Amy L. Vincent Baker, Richard J. Webby, Neil P. King, Yan Liu, Scott E. Hensley, Thomas P. Peacock, Louise H. Moncla, and Jesse D. Bloom. Deep mutational scanning of h5 hemagglutinin to inform influenza virus surveillance. *PLOS Biology*, 22(11):e3002916, November 2024a. ISSN 1545-7885. doi:10.1371/journal.pbio.3002916. URL http://dx.doi.org/10.1371/journal.pbio.3002916.

Bernadeta Dadonaite, Jack Brown, Teagan E. McMahon, Ariana G. Farrell, Marlin D. Figgins, Daniel Asarnow, Cameron Stewart, Jimin Lee, Jenni Logue, Trevor Bedford, Ben Murrell, Helen Y. Chu, David Veesler, and Jesse D. Bloom. Spike deep mutational scanning helps predict success of sars-cov-2 clades. *Nature*, 631(8021):617–626, July 2024b. ISSN 1476-4687. doi:10.1038/s41586-024-07636-1. URL http://dx.doi.org/10.1038/s41586-024-07636-1.

Bernadeta Dadonaite, Sheri Harari, Brendan B. Larsen, Lucas Kampman, Alex Harteloo, Anna Elias-Warren, Helen Y. Chu, and Jesse D. Bloom. Spike mutations that affect the function and antigenicity of recent kp.3.1.1-like sars-cov-2 variants. *Journal of Virology*, 99(11), November

2025. ISSN 1098-5514. doi:10.1128/jvi.01423-25. URL http://dx.doi.org/10.1128/jvi.01423-25.

Christian Dallago, Jody Mou, Kadina E Johnston, Bruce Wittmann, Nick Bhattacharya, Samuel Goldman, Ali Madani, and Kevin K Yang. FLIP: Benchmark tasks in fitness landscape inference for proteins. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, 2022.

DeepSeek-AI. DeepSeek Harness: Everything is a plugin. https://github.com/deepseek-ai/deepseek-harness, 2026.

Daniel Esposito, Jochen Weile, Jay Shendure, Lea M. Starita, Anthony T. Papenfuss, Frederick P. Roth, Douglas M. Fowler, and Alan F. Rubin. MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biology*, 20(1):223, 2019. doi:10.1186/s13059-019-1845-6. URL https://doi.org/10.1186/s13059-019-1845-6.

Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K Min, Kelly Brock, Yarin Gal, and Debora S Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, 2021.

Cade Gordon, Amy X. Lu, and Pieter Abbeel. Protein language model fitness is a matter of preference. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=UvPdpa4LuV.

Sheri Harari, Rachel T Eguia, Bernadeta Dadonaite, Caelan E Radford, Cameron Stewart, David Veesler, and Jesse D Bloom. Mutations to the hcov-229e spike have counterbalancing effects on serum antibody neutralization and receptor binding, February 2026. URL http://dx.doi.org/10.64898/2026.02.22.707297.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

Martin Hesslow, Niccolò Zanichelli, Pascal Notin, Iacopo Poli, and Debora Marks. RITA: a study on scaling up generative protein sequence models. *arXiv preprint arXiv:2205.05789*, 2022.

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

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In *International Conference on Machine Learning*, pp. 8946–8970. PMLR, 2022.

Xiaohui Ju, William W. Hannon, Tomasz Kaszuba, Caelan E. Radford, Brendan B. Larsen, Samantha S. Nelson, Christopher A. Nelson, Israel Baltazar-Perez, Ofer Zimmerman, Daved H. Fremont, Michael S. Diamond, and Jesse D. Bloom. Determinants of human versus mosquito cell entry by the chikungunya virus envelope proteins, August 2025. URL http://dx.doi.org/10.1101/2025.08.25.672233.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

Caroline Kikawa, Catiana H. Cartwright-Acar, Jackson B. Stuart, Maya Contreras, Lisa M. Levoir, Matthew J. Evans, Jesse D. Bloom, and Leslie Goo. The effect of single mutations in zika virus envelope on escape from broadly neutralizing antibodies. *Journal of Virology*, 97(11), November 2023. ISSN 1098-5514. doi:10.1128/jvi.01414-23. URL http://dx.doi.org/10.1128/jvi.01414-23.

Atul Kumar, Jacob D. Aguirre, Tara E. C. Condos, R. Julio Martinez-Torres, Viduth K. Chaugule, Rachel Toth, Ramasubramanian Sundaramoorthy, Pascal Mercier, Axel Knebel, Donald E. Spratt, Kathryn R. Barber, Gary S. Shaw, and Helen Walden. Disruption of the autoinhibited state primes the e3 ligase parkin for activation and catalysis. *The EMBO Journal*, 34(20):2506–2521, 2015. doi:10.15252/embj.201592337.

Elodie Laine, Yasaman Karami, and Alessandra Carbone. GEMME: a simple and fast global epistatic model predicting mutational effects. *Molecular Biology and Evolution*, 36(11):2604–2619, 2019.

Brendan B. Larsen, Teagan McMahon, Jack T. Brown, Zhaoqian Wang, Caelan E. Radford, James E. Crowe, David Veesler, and Jesse D. Bloom. Functional and antigenic landscape of the nipah virus receptor-binding protein. *Cell*, 188(9):2480–2494.e22, May 2025. ISSN 0092-8674. doi:10.1016/j.cell.2025.02.030. URL http://dx.doi.org/10.1016/j.cell.2025.02.030.

Brendan B. Larsen, Sheri Harari, Risako Gen, Cameron Stewart, David Veesler, and Jesse D. Bloom. Functional and antigenic constraints on the nipah virus fusion protein. *Proceedings of the National Academy of Sciences*, 123(6), February 2026. ISSN 1091-6490. doi:10.1073/pnas.2529505123. URL http://dx.doi.org/10.1073/pnas.2529505123.

Ivy Larsen and Jesse D. Bloom. Deep mutational scanning of nipah virus rbp for escape from designed minibinder sol23. https://github.com/dms-vep/Nipah_Malaysia_RBP_Minibinder_DMS, 2026. Unpublished dataset; source data repository, commit e46c09a65c9ab40f78ee78b165961e02c8252957 (accessed 5 September 2026).

Mingchen Li, Yang Tan, Xinzhu Ma, Bozitao Zhong, Huiqun Yu, Ziyi Zhou, Wanli Ouyang, Bingxin Zhou, Pan Tan, and Liang Hong. ProSST: Protein language modeling with quantized structure and disentangled attention. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=4Z7RZixpJQ.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8):1099–1106, 2023.

Céline Marquet, Julius Schlensok, Marina Abakarova, Burkhard Rost, and Elodie Laine. Expert-guided protein language models enable accurate and blazingly fast fitness prediction. *Bioinformatics*, 40(11):btae621, 2024.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems*, 34:29287–29303, 2021.

Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. Colabfold: making protein folding accessible to all. *Nature Methods*, 19(6):679–682, 2022.

Erik Nijkamp, Jeffrey A. Ruffolo, Eli N. Weinstein, Nikhil Naik, and Ali Madani. ProGen2: Exploring the boundaries of protein language models. *Cell Systems*, 34(11):968–978, 2023. doi:10.1016/j.cels.2023.10.002.

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 *International Conference on Machine Learning*, pp. 16990–17017. PMLR, 2022a.

Pascal Notin, Lood Van Niekerk, Aaron W Kollasch, Daniel Ritter, Yarin Gal, and Debora Susan Marks. TranceptEVE: Combining family-specific and family-agnostic models of protein sequences for improved fitness prediction. *NeurIPS 2022 Workshop on Learning Meaningful Representations of Life*, 2022b.

Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood Van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, et al. ProteinGym: large-scale benchmarks for protein fitness prediction and design. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

C. Anders Olson, Nicholas C. Wu, and Ren Sun. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Current Biology*, 24(22):2643–2651, 2014. doi:10.1016/j.cub.2014.09.072.

Jeffrey Ouyang-Zhang, Daniel Diaz, Adam Klivans, and Philipp Krähenbühl. Predicting a protein's stability under a million mutations. *Advances in Neural Information Processing Systems*, 36, 2024.

Caelan E Radford and Jesse D Bloom. Comprehensive maps of escape mutations from antibodies 10-1074 and 3bnc117 for envs from two divergent hiv strains, January 2025. URL http://dx.doi.org/10.1101/2025.01.30.635715.

Caelan E. Radford, Philipp Schommers, Lutz Gieselmann, Katharine H.D. Crawford, Bernadeta Dadonaite, Timothy C. Yu, Adam S. Dingens, Julie Overbaugh, Florian Klein, and Jesse D. Bloom. Mapping the neutralizing specificity of human anti-hiv serum by deep mutational scanning. *Cell Host & Microbe*, 31(7):1200–1215.e9, July 2023. ISSN 1931-3128. doi:10.1016/j.chom.2023.05.025. URL http://dx.doi.org/10.1016/j.chom.2023.05.025.

Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. MSA transformer. In *International Conference on Machine Learning*, pp. 8844–8856. PMLR, 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, et al. 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.

Cassandra AL Simonich, Teagan E McMahon, Gavin Juviler, Lucas Kampman, Helen Y Chu, and Jesse D Bloom. Mutational constraints on rsv f and its neutralization by antibodies, February 2026. URL http://dx.doi.org/10.64898/2026.02.12.705519.

Marion Sourisseau, Daniel J. P. Lawrence, Megan C. Schwarz, Carina H. Storrs, Ethan C. Veit, Jesse D. Bloom, and Matthew J. Evans. Deep mutational scanning comprehensively maps how zika envelope protein mutations affect viral growth and antibody escape. *Journal of Virology*, 93 (23), December 2019. ISSN 1098-5514. doi:10.1128/jvi.01291-19. URL http://dx.doi.org/10.1128/jvi.01291-19.

Tyler N. Starr, Allison J. Greaney, William W. Hannon, Andrea N. Loes, Kevin Hauser, Josh R. Dillen, Elena Ferri, Ariana Ghez Farrell, Bernadeta Dadonaite, Matthew McCallum, Kenneth A. Matreyek, Davide Corti, David Veesler, Gyorgy Snell, and Jesse D. Bloom. Shifting mutational constraints in the sars-cov-2 receptor-binding domain during viral evolution. *Science*, 377(6604): 420–424, July 2022a. ISSN 1095-9203. doi:10.1126/science.abo7896. URL http://dx.doi.org/10.1126/science.abo7896.

Tyler N. Starr, Allison J. Greaney, Cameron M. Stewart, Alexandra C. Walls, William W. Hannon, David Veesler, and Jesse D. Bloom. Deep mutational scans for ace2 binding, rbd expression, and antibody escape in the sars-cov-2 omicron ba.1 and ba.2 receptor-binding domains. *PLOS Pathogens*, 18(11):e1010951, November 2022b. ISSN 1553-7374.

doi:10.1371/journal.ppat.1010951. URL http://dx.doi.org/10.1371/journal.ppat.1010951.

Tyler N. Starr, Samantha K. Zepeda, Alexandra C. Walls, Allison J. Greaney, Sergey Alkhovsky, David Veesler, and Jesse D. Bloom. Ace2 binding is an ancestral and evolvable trait of sarbecoviruses. *Nature*, 603(7903):913–918, February 2022c. ISSN 1476-4687. doi:10.1038/s41586-022-04464-z. URL http://dx.doi.org/10.1038/s41586-022-04464-z.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. SaProt: protein language modeling with structure-aware vocabulary. In *The Twelfth International Conference on Learning Representations*, 2023.

Ning Sun, Shuxian Zou, Tianhua Tao, Sazan Mahbub, Dian Li, Yonghao Zhuang, Hongyi Wang, Xingyi Cheng, Le Song, and Eric P. Xing. Mixture of experts enable efficient and effective protein understanding and design. In *NeurIPS 2024 Workshop on AI for New Drug Modalities*. bioRxiv, 2024. doi:10.1101/2024.11.29.625425. URL https://www.biorxiv.org/content/10.1101/2024.11.29.625425v1.

Yang Tan, Mingchen Li, Bingxin Zhou, Bozitao Zhong, Lirong Zheng, Pan Tan, Ziyi Zhou, Huiqun Yu, Guisheng Fan, and Liang Hong. Simple, efficient, and scalable structure-aware adapter boosts protein language models. *Journal of Chemical Information and Modeling*, 2024.

Yang Tan, Chen Liu, Jingyuan Gao, Banghao Wu, Mingchen Li, Ruilin Wang, Lingrong Zhang, Huiqun Yu, Guisheng Fan, Liang Hong, and Bingxin Zhou. VenusFactory: An integrated system for protein engineering with data retrieval and language model fine-tuning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 230–241, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-253-4. doi:10.18653/v1/2025.acl-demo.23. URL https://aclanthology.org/2025.acl-demo.23/.

Yang Tan, Ruilin Wang, Banghao Wu, Liang Hong, and Bingxin Zhou. From high-throughput evaluation to wet-lab studies: advancing mutation effect prediction with a retrieval-enhanced model. *Bioinformatics*, 41, 07 2025b. ISSN 1367-4811. URL https://doi.org/10.1093/bioinformatics/btaf189.

Yang Tan, Bingxin Zhou, Lirong Zheng, Guisheng Fan, and Liang Hong. Semantical and geometrical protein encoding toward enhanced bioactivity and thermostability. *eLife*, 13:RP98033, may 2025c. ISSN 2050-084X. doi:10.7554/eLife.98033. URL https://doi.org/10.7554/eLife.98033.

Yang Tan, Yuanxi Yu, Can Wu, Bozitao Zhong, Mingchen Li, Guisheng Fan, Jiankang Zhu, Yafeng Liang, Nanqing Dong, and Liang Hong. Rank-and-Reason: Multi-agent collaboration accelerates zero-shot protein mutation prediction, 2026. URL https://arxiv.org/abs/2602.00197.

Ashley L. Taylor and Tyler N. Starr. Deep mutational scans of xbb.1.5 and bq.1.1 reveal ongoing epistatic drift during sars-cov-2 evolution. *PLOS Pathogens*, 19(12):e1011901, December 2023. ISSN 1553-7374. doi:10.1371/journal.ppat.1011901. URL http://dx.doi.org/10.1371/journal.ppat.1011901.

Ashley L Taylor and Tyler N Starr. Deep mutational scanning of sars-cov-2 omicron ba.2.86 and epistatic emergence of the kp.3 variant. *Virus Evolution*, 10(1), 2024. ISSN 2057-1577. doi:10.1093/ve/veae067. URL http://dx.doi.org/10.1093/ve/veae067.

Mustafa Tekpinar, Laurent David, Thomas Henry, and Alessandra Carbone. PRESCOTT: a population aware, epistatic and structural model accurately predicts missense effect. *Genome Biology*, 26:113, 2025. doi:10.1186/s13059-025-03581-y.

Thomas C. Terwilliger, Dorothee Liebschner, Tristan I. Croll, Christopher J. Williams, Airlie J. McCoy, Billy K. Poon, Pavel V. Afonine, Robert D. Oeffner, Jane S. Richardson, Randy J. Read, and Paul D. Adams. AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. *Nature Methods*, 21:110–116, 2024. doi:10.1038/s41592-023-02087-4. URL https://doi.org/10.1038/s41592-023-02087-4.

Timothy Truong Jr and Tristan Bepler. PoET: A generative model of protein families as sequences-of-sequences. *Advances in Neural Information Processing Systems*, 36:77379–77415, 2023.

Matsvei Tsishyn, Pauline Hermans, Fabrizio Pucci, and Marianne Rooman. Residue conservation and solvent accessibility are (almost) all you need for predicting mutational effects in proteins. *Bioinformatics*, 41(6):btaf322, 2025. doi:10.1093/bioinformatics/btaf322.

Ruben Weitzman, Peter Mørch Groth, Lood Van Niekerk, Aoi Otani, Yarin Gal, Debora S. Marks, and Pascal Notin. Protriever: End-to-end differentiable protein homology search for fitness prediction. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 66397–66418. PMLR, 2025.

Frances C. Welsh, Rachel T. Eguia, Juhye M. Lee, Hugh K. Haddox, Jared Galloway, Nguyen Van Vinh Chau, Andrea N. Loes, John Huddleston, Timothy C. Yu, Mai Quynh Le, Nguyen T.D. Nhat, Nguyen Thi Le Thanh, Alexander L. Greninger, Helen Y. Chu, Janet A. Englund, Trevor Bedford, Frederick A. Matsen, Maciej F. Boni, and Jesse D. Bloom. Age-dependent heterogeneity in the antigenic effects of mutations to influenza hemagglutinin. *Cell Host & Microbe*, 32(8): 1397–1411.e11, August 2024. ISSN 1931-3128. doi:10.1016/j.chom.2024.06.015. URL http://dx.doi.org/10.1016/j.chom.2024.06.015.

Xinyu Wu, Margareta Go, Julie V. Nguyen, Nathan W. Kuchel, Bernadine G. C. Lu, Kathleen Zeglinski, Kym N. Lowes, Dale J. Calleja, Jeffrey P. Mitchell, Guillaume Lessene, David Komander, Matthew E. Call, and Melissa J. Call. Mutational profiling of sars-cov-2 papain-like protease reveals requirements for function, structure, and drug escape. *Nature Communications*, 15(1), July 2024. ISSN 2041-1723. doi:10.1038/s41467-024-50566-9. URL http://dx.doi.org/10.1038/s41467-024-50566-9.

Minghao Xu, Zuobai Zhang, Jiarui Lu, Zhaocheng Zhu, Yangtian Zhang, Chang Ma, Runcheng Liu, and Jian Tang. PEER: A comprehensive and multi-task benchmark for protein sequence understanding. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=QgTZ56-zJou.

Kevin K Yang, Niccolò Zanichelli, and Hugh Yeh. Masked inverse folding with sequence transfer for protein representation learning. *Protein Engineering, Design and Selection*, 36:gzad015, 2023.

Kevin K Yang, Nicolo Fusi, and Alex X Lu. Convolutions are competitive with transformers for protein sequence pretraining. *Cell Systems*, 15(3):286–294, 2024.

Dongxin Ye, Fang Hu, Han Hu, Shu Hu, Yang Tan, Wanli Ouyang, Stan Z. Li, Jie Cui, and Nanqing Dong. ViroBench: Benchmarking nucleotide foundation models on viral genomics tasks. In *32nd SIGKDD Conference on Knowledge Discovery and Data Mining - AI for Sciences Track*, 2026. doi:10.1145/3770855.3819057. URL https://openreview.net/forum?id=7vstEitzQr.

Timothy C. Yu, Caroline Kikawa, Bernadeta Dadonaite, Andrea N. Loes, Janet A. Englund, and Jesse D. Bloom. Pleiotropic mutational effects on function and stability constrain the antigenic evolution of influenza haemagglutinin. *Nature Ecology & Evolution*, 10(3):452–466, December 2025. ISSN 2397-334X. doi:10.1038/s41559-025-02895-1. URL http://dx.doi.org/10.1038/s41559-025-02895-1.

Yingpu Yu, Maximilian A. Kass, Mengyin Zhang, Noor Youssef, Catherine A. Freije, Kelly P. Brock, Lauren C. Aguado, Leon L. Seifert, Sanjana Venkittu, Xupeng Hong, Amir Shlomai, Ype P. de Jong, Debora S. Marks, Charles M. Rice, and William M. Schneider. Deep mutational scanning of hepatitis b virus reveals a mechanism for cis-preferential reverse transcription. *Cell*, 187(11):2735–2745.e12, May 2024. ISSN 0092-8674. doi:10.1016/j.cell.2024.04.008. URL http://dx.doi.org/10.1016/j.cell.2024.04.008.

Liang Zhang, Hua Pang, Chenghao Zhang, Song Li, Yang Tan, Fan Jiang, Mingchen Li, Yuanxi Yu, Ziyi Zhou, Banghao Wu, et al. Venusmuthub: a systematic evaluation of protein mutation effect predictors on small-scale experimental data. *Acta Pharmaceutica Sinica B*, 2025.

Zuobai Zhang, Pascal Notin, Yining Huang, Aurelie C Lozano, Vijil Chenthamarakshan, Debora Marks, Payel Das, and Jian Tang. Multi-scale representation learning for protein fitness prediction. *Advances in Neural Information Processing Systems*, 37:101456–101473, 2024.

Bingxin Zhou, Yang Tan, Yutong Hu, Lirong Zheng, Bozitao Zhong, and Liang Hong. Protein engineering in the deep learning era. *mLife*, 3(4):477–491, 2024a.

Bingxin Zhou, Lirong Zheng, Banghao Wu, Yang Tan, Outongyi Lv, Kai Yi, Guisheng Fan, and Liang Hong. Protein engineering with lightweight graph denoising neural networks. *Journal of Chemical Information and Modeling*, 2024b.

Yichen Zhou, Jonathan Golob, Amir Karimi, Stefan Bauer, and Patrick Schwab. ViroGym: Realistic large-scale benchmarks for evaluating viral proteins. *arXiv preprint arXiv:2603.06740*, 2026. URL https://arxiv.org/abs/2603.06740.

A Related Work

Protein models. Protein mutation-effect prediction combines family-specific evolutionary models with pretrained protein models. EVmutation, GEMME, and EVE infer homolog-family constraints, while RSALOR and ESCOTT add structural context (Hopf et al., 2017; Laine et al., 2019; Frazer et al., 2021; Tsishyn et al., 2025; Tekpinar et al., 2025). ProteinGym standardizes zero-shot evaluation (Notin et al., 2024) across masked and autoregressive sequence models (Meier et al., 2021; Rives et al., 2021; Hesslow et al., 2022; Lin et al., 2023; Madani et al., 2023; Nijkamp et al., 2023; Yang et al., 2024), structure-aware encoders (Su et al., 2023; Li et al., 2024; Zhang et al., 2024; Hayes et al., 2025; Tan et al., 2025c), and inverse-folding models (Hsu et al., 2022; Dauparas et al., 2022; Yang et al., 2023). These families supply complementary sequence, evolutionary, and geometric priors.

Fitness benchmarks. Fitness benchmarks span variants, whole proteins, and nucleotide sequences. DeepSequence compares unsupervised evolutionary scores with experimental mutation effects; FLIP defines fitness-landscape splits, ProteinGym standardizes substitution and indel evaluation, and MaveDB archives multiplexed variant-effect assays (Riesselman et al., 2018; Esposito et al., 2019; Dallago et al., 2021; Notin et al., 2024). At protein level, PEER covers function, localization, structure, and interactions, while VenusFactory standardizes tasks across more than 40 datasets (Xu et al., 2022; Tan et al., 2025a). Viral benchmarks add complementary modalities: ViroBench evaluates nucleotide foundation models under phylogenetic and temporal shifts, whereas ViroGym evaluates viral-protein fitness, antigenic diversity, and forecasting (Ye et al., 2026; Zhou et al., 2026). These resources motivate explicit reporting of provenance, overlap, aggregation, and phenotype coverage in ViroHub.

Post-hoc calibration. Post-hoc methods add evidence without redesigning the backbone. Retrieval and family modeling underlie MSA Transformer, Tranception/TranceptEVE, PoET, Protriever, S3F-MSA/S2F-MSA, and VenusREM (Rao et al., 2021; Notin et al., 2022a;b; Truong Jr & Bepler, 2023; Zhang et al., 2024; Weitzman et al., 2025; Tan et al., 2025b). Supervised alternatives fine-tune protein-specific predictors, post-train pretrained representations, or learn multimodal ensemble weights (Zhou et al., 2024b; Ouyang-Zhang et al., 2024; Tan et al., 2026); homolog fine-tuning further links accuracy to wild-type likelihood and family coverage (Gordon et al., 2025). RH instead combines family constraints, frozen scores, and structural reliability in one training-free, architecture-agnostic readout.

B Evaluation, validation, and provenance details

Appendix B.1 defines the metric-specific evaluation sets and aggregation. Appendix B.2 characterizes sequence, alignment, and structural inputs; Appendix B.3 documents viral assay provenance; Appendix B.4 summarizes the resulting benchmark and cross-model outcomes.

B.1 Metric definitions and benchmark aggregation

For an assay with n retained experimental-effect/model-score pairs (y_i, s_i), both oriented so that larger values are preferred, let $u_i = \mathrm{rank}(y_i)$ and $v_i = \mathrm{rank}(s_i)$ use average ranks for ties. Their arithmetic means are $\bar{u}$ and $\bar{v}$. Assay-level Spearman correlation is

$$\rho_{\mathrm{S}}(y, s) = \frac{\sum_{i=1}^{n}(u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum_{i=1}^{n}(u_i - \bar{u})^2}\sqrt{\sum_{i=1}^{n}(v_i - \bar{v})^2}}. \tag{17}$$

ProteinGym and ViroHub use NDCG@10% with linear min–max gains $g_i = (y_i - \min_j y_j)/(\max_j y_j - \min_j y_j)$ and $k = \lfloor 0.1n \rfloor$. Here $r_i^{(s)}$ and $r_i^{(y)}$ are ordinal descending ranks under s and y; the evaluator orders ties with its array-sorting routine, rather than assigning average

ranks. The predicted and ideal discounted gains are

$$\mathrm{DCG}_{10} = \sum_{i:r_i^{(s)} \le k} \frac{g_i}{\log_2(r_i^{(s)}+1)}, \qquad \mathrm{IDCG}_{10} = \sum_{i:r_i^{(y)} \le k} \frac{g_i}{\log_2(r_i^{(y)}+1)}, \\ \mathrm{NDCG}_{10} = \frac{\mathrm{DCG}_{10}}{\mathrm{IDCG}_{10}}. \tag{18}$$

VenusMutHub retains its full-list NDCG convention. With the same normalized relevance g_i, it uses exponential gain $2^{g_i} - 1$ at all n positions:

$$\mathrm{NDCG}_{\mathrm{full}} = \frac{\sum_{i=1}^{n} (2^{g_i}-1)/\log_2(r_i^{(s)}+1)}{\sum_{i=1}^{n} (2^{g_i}-1)/\log_2(r_i^{(y)}+1)}. \tag{19}$$

Its prediction sort is stable, preserving input order for ties. Thus NDCG values are compared between Raw and the harness within a benchmark, with the benchmark-specific definition held fixed.

Let $Q_{0.9}(\cdot)$ be the empirical 90th percentile with linear interpolation, and define $T_y = \{i : y_i \ge Q_{0.9}(y)\}$ and $T_s = \{i : s_i \ge Q_{0.9}(s)\}$. Ties at the threshold are retained, so these sets may exceed 10% of variants. Top-variant recall@10% is

$$\mathrm{Recall}_{10} = \frac{|T_y \cap T_s|}{|T_y|}. \tag{20}$$

For binary discrimination, ProteinGym uses its supplied `DMS_score_bin` labels. VenusMutHub and ViroHub use $z_i = \mathbf{1}\{y_i \ge \mathrm{median}(y)\}$, where $\mathbf{1}\{\cdot\}$ is an indicator. With $n_+ = \sum_i z_i$ and $n_- = n - n_+$, AUC uses continuous model scores and half credit for tied positive–negative pairs:

$$\mathrm{AUC} = \frac{1}{n_+ n_-} \sum_{i:z_i=1} \sum_{j:z_j=0} \left[\mathbf{1}\{s_i > s_j\} + \tfrac{1}{2}\mathbf{1}\{s_i = s_j\} \right]. \tag{21}$$

For MCC, predicted labels are $\hat{z}_i = \mathbf{1}\{s_i \ge \mathrm{median}(s)\}$, giving

$$\mathrm{MCC} = \frac{\mathrm{TP}\,\mathrm{TN} - \mathrm{FP}\,\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}. \tag{22}$$

Here $\mathrm{TP}, \mathrm{TN}, \mathrm{FP}, \mathrm{FN}$ count true positives, true negatives, false positives, and false negatives under $(z_i, \hat{z}_i)$. Spearman requires nonconstant rankings; NDCG requires nonzero relevance range, positive ideal gain, and, for the truncated version, $k \ge 1$; AUC requires both experimental classes. The ViroHub evaluator excludes non-finite pairs, rejects assays with fewer than three pairs or constant scores, and records a zero-denominator MCC as undefined. ProteinGym and VenusMutHub likewise require both experimental classes for AUC/MCC, but their evaluator returns MCC 0 for constant predicted labels. Undefined values are omitted from subsequent means. All metrics are computed per assay before benchmark aggregation; VenusMutHub reports the assay macro-average. For metric q, let $\mathcal{A}_{pb}^{(q)}$ be the assays with a defined value for phenotype p and viral protein b, let $\mathcal{B}_p^{(q)} = \{b : \mathcal{A}_{pb}^{(q)} \ne \varnothing\}$, and let $\mathcal{P}^{(q)} = \{p : \mathcal{B}_p^{(q)} \ne \varnothing\}$. Writing the per-assay value as $m_{pba}^{(q)}$ gives

$$M_{\mathrm{ViroHub}}^{(q)} = \frac{1}{|\mathcal{P}^{(q)}|} \sum_{p \in \mathcal{P}^{(q)}} \frac{1}{|\mathcal{B}_p^{(q)}|} \sum_{b \in \mathcal{B}_p^{(q)}} \frac{1}{|\mathcal{A}_{pb}^{(q)}|} \sum_{a \in \mathcal{A}_{pb}^{(q)}} m_{pba}^{(q)}. \tag{23}$$

This prevents densely assayed viral proteins from dominating. Spearman, NDCG, and top-variant recall use all 89 assays and 52 phenotype–protein cells; AUC and MCC use 82 assays and 45 cells because seven CVB3 fitness assays have single-class labels. ProteinGym leaderboard Spearman follows the official 217-assay evaluator; extended metrics retain the benchmark-specific definitions above.

Official Average Spearman, VenusMutHub macro-averaging, and ViroHub hierarchical aggregation follow Section 3. Uncertainty uses 20,000 paired bootstrap draws over 186 ProteinGym UniProt IDs, 527 VenusMutHub WT-sequence clusters, or the metric-valid ViroHub phenotype–protein cells within each of seven phenotype strata. For ViroHub, seed 0 and the same metric-specific draws are used across all configurations and paired Raw–RH comparisons. Configurations are held fixed, so intervals quantify uncertainty over benchmark units rather than a model population. Appendix D gives the site-profile and separation protocols; Appendix F gives component selection and leave-one-site-out diagnostics.

B.2 Dataset and input distributions

Sequence and alignment distributions. ProteinGym contains 217 assays spanning 186 unique UniProt IDs and 187 unique wild-type (WT) sequences, VenusMutHub contains 905 assays across 527 unique WT clusters, and ViroHub contributes 89 assays across 52 unique WT sequences after duplicate removal. Figure 8a–c and Table 2 compare their inputs. ViroHub sequences are longer on average (559 residues versus 397 and 288) but have shallower MSAs (1,217 sequences versus 8,650 and 9,791); mean MSA occupancy is 0.62, compared with 0.67 for ProteinGym and 0.69 for VenusMutHub.

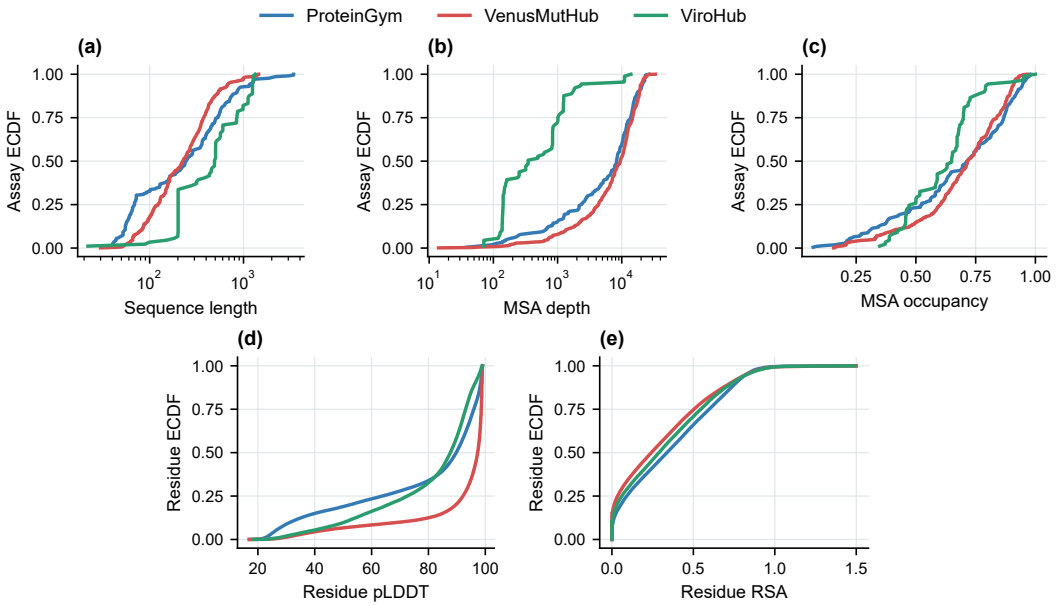


Figure 8: Input distributions for ProteinGym (217 assays), VenusMutHub (905), and ViroHub (89). **(a)** Sequence length. **(b)** MSA depth. **(c)** MSA occupancy. **(d)** Residue pLDDT. **(e)** Residue RSA. Structure panels pool unique WT sequences; all selected structures fully cover the query.

Table 2: Assay-level input statistics for ProteinGym, VenusMutHub, and ViroHub, reported as mean $\pm$ standard deviation. N is the number of assays; MSA occupancy is the fraction of non-gap match-state entries; pLDDT and RSA are assay-level means.

Dataset	N assays	Sequence length	MSA depth	MSA occupancy	Mean pLDDT	Mean RSA
ProteinGym	217	397 $\pm$ 502	8,650 $\pm$ 6,713	0.67 $\pm$ 0.23	86.6 $\pm$ 10.8	0.37 $\pm$ 0.09
VenusMutHub	905	288 $\pm$ 239	9,791 $\pm$ 6,452	0.69 $\pm$ 0.18	92.1 $\pm$ 7.5	0.32 $\pm$ 0.08
ViroHub	89	559 $\pm$ 383	1,217 $\pm$ 2,610	0.62 $\pm$ 0.14	83.9 $\pm$ 9.8	0.33 $\pm$ 0.07

Structure distributions. pLDDT is read from Cα B-factors and RSA from Shrake–Rupley accessibility normalized by residue-specific maximum ASA. Selected AlphaFold2 models cover every query residue in all three benchmarks. ViroHub has the lowest assay-level mean pLDDT (83.9), with 41.1% of residues at pLDDT $\geq$ 90 and 9.7% below 50, while its mean RSA (0.33) lies between ProteinGym (0.37) and VenusMutHub (0.32). Figure 8d–e and Tables 3–4 report the full distributions.

Table 3: AlphaFold2 confidence strata over unique wild-type sequences. All selected structures cover the full query; residue percentages are pooled within each benchmark.

Dataset	Unique WT	Structures	pLDDT < 50	50–70	70–90	≥ 90
ProteinGym	187	187	18.9%	9.0%	22.6%	49.5%
VenusMutHub	527	527	6.7%	3.3%	10.6%	79.5%
ViroHub	52	52	9.7%	13.2%	36.0%	41.1%

Table 4: RSA strata over unique wild-type sequences. Residue percentages are pooled within each benchmark.

Dataset	RSA < 0.2	RSA 0.2–0.5	RSA ≥ 0.5
ProteinGym	36.7%	29.4%	33.9%
VenusMutHub	45.8%	28.9%	25.3%
ViroHub	40.6%	30.2%	29.2%

B.3 ViroHub data curation and provenance

Relation to existing benchmarks. ProteinGym provides a comprehensive benchmark for zero-shot protein fitness prediction and mutation-guided protein design. Its 217 substitution DMS assays span diverse proteins, functions, taxa, and MSA depths, but only 23 are viral (Notin et al., 2024). VenusMutHub contributes 905 high-precision measurements across 527 wild-type sequence clusters, but is not organised around viral phenotypes. ViroGym further assembled 79 viral DMS assays and broadened evaluation to immune escape and other virus-relevant phenotypes (Zhou et al., 2026). However, 23 of these assays overlap with ProteinGym, and most concern SARS-CoV-2 or influenza A. ViroHub complements these resources with distinct viral DMS measurements, wider virus coverage, and stronger representation of immune escape and cell entry.

Data collection. We identified candidate assays from public data portals, repositories, cited studies, and recent literature, then reconstructed each assay from its original materials. We retained quantitative measurements of single-amino-acid substitutions with an accessible mutant-level table, a recoverable wild-type sequence, mappable residue numbering, and approximately 400 or more substitutions after quality control. Assays were distinguished by study, virus strain, protein or scanned region, and phenotype. Mapped coordinates and sequences were used to identify and exclude duplicate measurements.

Data processing. Each assay was converted to a wild-type sequence, single-substitution identifier, and experimental score. Residue coordinates followed the experimental construct, using documented offsets or sequence alignment when needed. We removed missing scores, unmappable substitutions, reference-residue mismatches, and duplicate records. Replicates or experimental conditions were combined using the assay-specific rule recorded in the processing manifest. Score directions were harmonised so that larger values indicate more favourable values of the reported phenotype, while the original numerical scales were retained. The QC files retain $642{,}737$ rows, comprising $616{,}690$ substitutions and $26{,}047$ wild-type controls (XnX). Controls may remain in data and score files but are excluded before computing any evaluation metric; all 89 assays are retained.

Benchmark composition. ViroHub contains $616{,}690$ single-amino-acid substitutions from 89 assays, covering 18 virus types, 52 wild-type sequences, and seven phenotype groups (Table 5). Compared with ViroGym, it expands coverage from 13 to 18 virus types, from 14 to 21 immune-escape assays, and from 11 to 18 cell-entry assays. The manifest records each assay's source, virus strain, experimental construct, coordinate convention, wild-type sequence hash, processing rule, and retained mutation count. Evaluation follows the metric-specific hierarchy in Equation 23.

Assay-level provenance. Tables 6–12 link every retained assay to its experimental readout and original study or public data repository, grouped by readout. Exact assay identifiers, construct coor-

Table 5: Phenotype composition of ViroHub after quality control. Evaluated substitutions exclude wild-type control rows.

Harmonised phenotype	Assays	Evaluated substitutions
Immune escape	21	148,468
Cell entry	18	189,048
Binding	17	90,245
Expression	15	58,120
Fitness	15	114,287
Stability	2	11,187
Activity	1	5,335
Total	89	616,690

dinates, retained mutation counts, coordinate transformations, and assay-specific combination rules are provided in the processing manifest.

Antibody escape is recorded separately for monoclonal antibodies (Table 6) and polyclonal sera (Table 7). This distinction preserves the experimental context of the immune-escape comparisons.

Table 6: ViroHub assays with experimental readout *Monoclonal-antibody escape* (15 assays). Distinct rows from the same study correspond to different viral constructs or measured phenotypes. [†]Public dataset repository without an associated archival article.

Virus and construct	Source
HIV-1 BF520 Env	(Radford & Bloom, 2025)
HIV-1 TRO11 Env	(Radford & Bloom, 2025)
Influenza A/H3N2 HK19 HA	(Welsh et al., 2024)
Influenza A/H3N2 MC22 HA	(Yu et al., 2025)
Lassa virus Josiah GP	(Carr et al., 2024)
Nipah virus F	(Larsen et al., 2026)
Nipah virus Malaysia RBP	(Larsen et al., 2025)
Nipah virus Malaysia RBP	(Larsen & Bloom, 2026)[†]
Rabies virus G	(Aditham et al., 2025)
Respiratory syncytial virus A F	(Simonich et al., 2026)
SARS-CoV-2 Wuhan-Hu-1 RBD	(Cao et al., 2022)
SARS-CoV-2 Omicron BA.2 spike	(Dadonaite & Bloom, 2026)[†]
SARS-CoV-2 KP.3.1.1 spike	(Dadonaite et al., 2025)
SARS-CoV-2 XBB.1.5 spike	(Dadonaite et al., 2024b)
Zika virus E	(Sourisseau et al., 2019; Kikawa et al., 2023)

Table 7: ViroHub assays with experimental readout *Serum-antibody escape* (6 assays).

Virus and construct	Source
HIV-1 BF520 Env	(Radford et al., 2023)
Human coronavirus 229E spike	(Harari et al., 2026)
Influenza A/H5N1 HA	(Dadonaite et al., 2024a)
Lassa virus Josiah GP	(Carr & Bloom, 2026)[†]
Respiratory syncytial virus A F	(Simonich et al., 2026)
SARS-CoV-2 XBB.1.5 RBD	(Dadonaite et al., 2024b)

Cell-entry assays (Table 8) and receptor-binding assays (Table 9) probe related but distinct steps in host interaction; their original readouts remain separate in the manifest.

Table 8: ViroHub assays with experimental readout *Cell entry* (18 assays).

Virus and construct	Source
Chikungunya virus E3/E2/E1	(Ju et al., 2025)
Chikungunya virus E3/E2/E1	(Ju et al., 2025)
HIV-1 BF520 Env	(Radford & Bloom, 2025)
HIV-1 TRO11 Env	(Radford & Bloom, 2025)
Human coronavirus 229E spike	(Harari et al., 2026)
Influenza A/H3N2 MC22 HA	(Yu et al., 2025)
Influenza A/H5N1 HA	(Dadonaite et al., 2024a)
Influenza A/H7N9 Anhui13 HA	(Ahn et al., 2026)
Influenza A/H7N9 Anhui13 HA	(Ahn et al., 2026)
Lassa virus Josiah GP	(Carr et al., 2024)
Nipah virus F	(Larsen et al., 2026)
Nipah virus Malaysia RBP	(Larsen et al., 2025)
Rabies virus G	(Aditham et al., 2025)
Respiratory syncytial virus A F	(Simonich et al., 2026)
SARS-CoV-2 XBB.1.5 RBD	(Dadonaite et al., 2024b)
SARS-CoV-2 Omicron BA.2 spike	(Dadonaite et al., 2024b)
SARS-CoV-2 KP.3.1.1 spike	(Dadonaite et al., 2025)
SARS-CoV-2 XBB.1.5 spike	(Dadonaite et al., 2024b)

Table 9: ViroHub assays with experimental readout *Receptor binding* (17 assays).

Virus and construct	Source
Chikungunya virus E3/E2/E1	(Ju et al., 2025)
Human coronavirus 229E spike	(Harari et al., 2026)
Nipah virus Malaysia RBP	(Larsen et al., 2025)
SARS-CoV-2 Alpha RBD	(Starr et al., 2022a)
SARS-CoV-2 Beta RBD	(Starr et al., 2022a)
SARS-CoV-2 Delta RBD	(Starr et al., 2022a)
SARS-CoV-2 Eta RBD	(Starr et al., 2022a)
SARS-CoV-2 Omicron BA.1 RBD	(Starr et al., 2022b)
SARS-CoV-2 Omicron BA.2.86 RBD	(Taylor & Starr, 2024)
SARS-CoV-2 Omicron BA.2 RBD	(Starr et al., 2022b)
SARS-CoV-2 Omicron BQ.1.1 RBD	(Taylor & Starr, 2023)
SARS-CoV-2 Omicron EG.5 RBD	(Taylor & Starr, 2024)
SARS-CoV-2 Omicron FLip RBD	(Taylor & Starr, 2024)
SARS-CoV-2 Omicron XBB.1.5 RBD	(Taylor & Starr, 2023)
SARS-CoV-2 Omicron BA.2 spike	(Dadonaite et al., 2024b)
SARS-CoV-2 KP.3.1.1 spike	(Dadonaite et al., 2025)
SARS-CoV-2 XBB.1.5 spike	(Dadonaite et al., 2024b)

Table 10 records protein-expression and abundance assays. These measurements provide a complementary readout of protein production and persistence rather than antibody recognition.

Table 10: ViroHub assays with experimental readout *Protein expression and abundance* (15 assays).

Virus and construct	Source
Bat coronavirus PRD0038 RBD	(Starr et al., 2022c)
Bat coronavirus RmYN02 RBD	(Starr et al., 2022c)
Bat coronavirus RsYN04 RBD	(Starr et al., 2022c)
SARS-CoV-2 PLpro	(Wu et al., 2024)
SARS-CoV-2 Alpha RBD	(Starr et al., 2022a)
SARS-CoV-2 Beta RBD	(Starr et al., 2022a)
SARS-CoV-2 Delta RBD	(Starr et al., 2022a)
SARS-CoV-2 Eta RBD	(Starr et al., 2022a)
SARS-CoV-2 Omicron BA.1 RBD	(Starr et al., 2022b)
SARS-CoV-2 Omicron BA.2.86 RBD	(Taylor & Starr, 2024)
SARS-CoV-2 Omicron BA.2 RBD	(Starr et al., 2022b)
SARS-CoV-2 Omicron BQ.1.1 RBD	(Taylor & Starr, 2023)
SARS-CoV-2 Omicron EG.5 RBD	(Taylor & Starr, 2024)
SARS-CoV-2 Omicron FLip RBD	(Taylor & Starr, 2024)
SARS-CoV-2 Omicron XBB.1.5 RBD	(Taylor & Starr, 2023)

Fitness assays (Table 11) and stability or protease-activity assays (Table 12) complete the provenance record. Distinct constructs from one study remain separate assay entries.

Table 11: ViroHub assays with experimental readout *Viral fitness and growth* (15 assays).

Virus and construct	Source
Coxsackievirus B3 2A	(Álvarez-Rodríguez et al., 2024)
Coxsackievirus B3 2B	(Álvarez-Rodríguez et al., 2024)
Coxsackievirus B3 2C	(Álvarez-Rodríguez et al., 2024)
Coxsackievirus B3 3A	(Álvarez-Rodríguez et al., 2024)
Coxsackievirus B3 3B	(Álvarez-Rodríguez et al., 2024)
Coxsackievirus B3 3C	(Álvarez-Rodríguez et al., 2024)
Coxsackievirus B3 3D	(Álvarez-Rodríguez et al., 2024)
Coxsackievirus B3 VP1	(Álvarez-Rodríguez et al., 2024)
Coxsackievirus B3 VP3	(Álvarez-Rodríguez et al., 2024)
Enterovirus A capsid	(Bakhache et al., 2024)
Enterovirus A replicase	(Bakhache et al., 2024)
Hepatitis B virus reverse transcriptase	(Yu et al., 2024)
Influenza A/H3N2 HK19 HA	(Welsh et al., 2024)
SARS-CoV-2 Omicron BA.1 spike	(Dadonaite et al., 2023)
SARS-CoV-2 Delta spike	(Dadonaite et al., 2023)

Table 12: ViroHub assays with experimental readout *Protein stability and protease activity* (3 assays).

Virus and construct	Source
Influenza A/H3N2 MC22 HA	(Yu et al., 2025)
Influenza A/H5N1 HA	(Dadonaite et al., 2024a)
SARS-CoV-2 PLpro	(Wu et al., 2024)

B.4 ViroHub benchmark

Scope and curation. The final benchmark contains 89 viral DMS assays and 616,690 scored substitutions across seven phenotype classes: immune escape, cell entry, binding, expression, fitness, stability, and activity. No retained assay represents the same experimental measurement as an assay in ProteinGym or VenusMutHub.

Hierarchical evaluation. To prevent densely assayed viral proteins from dominating, each metric is first averaged within phenotype–protein cells, then across viral proteins within a phenotype, and finally equally across the seven phenotypes. The 71 Raw–RH pairs use the same frozen checkpoints and five metric definitions as the other benchmarks.

Cross-model transfer. Under this hierarchy, mean Spearman rises from 0.158 to 0.252 ($+0.094$), with positive changes for all 71 configurations. The largest gains occur for masked sequence models ($+0.151$; 0.072 to 0.222) and autoregressive models ($+0.107$; 0.099 to 0.206). Structure-aware encoders improve from 0.215 to 0.280 ($+0.065$), and inverse-folding models from 0.271 to 0.307 ($+0.036$). Thus, family evidence contributes most strongly to sequence-only readouts while remaining complementary to geometric conditioning (Table 24).

Metric and component consistency. Averaged over the same 71 configurations, NDCG increases from 0.684 to 0.729, AUC from 0.584 to 0.634, MCC from 0.125 to 0.202, and top-variant recall from 0.133 to 0.152. Every configuration improves on every metric, yielding 355 of 355 positive model–metric comparisons (Tables 55–58). For PROSST-2048, entropy-weighted retrieval raises Spearman from 0.2709 to 0.2901, and coherence-gated calibration reaches 0.2942. RSA-only and pLDDT-only shrinkage reach 0.3085 and 0.2994, respectively, while the full readout reaches 0.3102, identifying exposure as the larger reliability correction on ViroHub (Table 1).

C BASELINE MODELS AND IMPLEMENTATIONS

Table 13 lists the evaluated baselines, their readout interfaces, and official implementations within the common framework in Figure 2. All backbones are frozen and evaluated with the readout used by the reported run (Equation 1); Appendix D quantifies the shared native-background direction and interface-specific checkpoint deviations induced by these readouts. Configurations from the same family (e.g., the ESM-2 size series) count as distinct model–readout configurations. The CARP series includes 600K, 38M, 76M, and 640M models. PROTSSN covers 9 graph encoders combining $k \in \{10, 20, 30\}$ neighbors with hidden widths $h \in \{512, 768, 1280\}$, alongside their ensemble. Both series use native wild-type fields. The PROSST series varies the structure-vocabulary size K and uses one wild-type forward pass; masked ProSST scoring is not used in the reported experiments. VENUSREM2 scores each of the six K-variants with RH at that variant's own entropy-weighted α from Equation 8, then z-scores each K's per-assay variant-score vector and averages the six standardized vectors. Constant-score members contribute zero after standardization. S3F contributes two readout configurations, wild-type and masked marginals, from one checkpoint but with distinct native fields. SAPROT evaluates 35M-AF2, 650M-AF2, and 650M-PDB checkpoints with both mask and wild-type interfaces. Protein structures are AlphaFold2 predictions (Jumper et al., 2021) and query-matched alignments are built with ColabFold (Mirdita et al., 2022).

The comparison in Figure 3 is detailed in Tables 14 and 15, which report official overall and biological-category scores for one representative per model series, covering sequence, structure, and evolutionary baselines. Table 16 keeps those same rows and replaces the five biological-category columns with official taxon and MSA-depth bins: VENUSREM2 is strongest in every reported stratum, including virus (0.540) and the low-depth bin (0.553). Table 18 completes the comparison by mutation depth, the number of amino-acid substitutions per variant; VENUSREM2 ranks first in all five bins.

Table 13: Evaluated baselines, native readouts, and implementations. Readouts follow Equation 1; mask = masked marginal, wt = wild-type field, and tf = teacher forcing.

Model	Family	Evaluated variants	Readout	Source
ESM-1B	Masked sequence LM	650M	mask, wt	Official repository
ESM-1V	Masked sequence LM	650M	mask, wt	Official repository
ESM-2	Masked sequence LM	8M/35M/150M/650M/3B	mask, wt	Official repository
ESMC	Masked sequence LM	300M, 600M	mask	Official repository
CARP	Masked sequence LM	600K/38M/76M/640M	wt	Official repository
S3F	Structure-aware	S3F	wt, mask	Official repository
PROGEN2	Autoregressive LM	B/S/M/L/XL	tf	Official repository
PROGEN3	Autoregressive LM	112M/219M/339M/762M/1B/3B	tf	Official repository
RITA	Autoregressive LM	S/M/L/XL	tf	Official repository
ESM3	Structure-aware	open-weight (1.4B)	mask	Official repository
PROSST	Structure-aware	K=20/128/512/1024/2048/4096	wt	Official repository
PROTSSN	Structure-aware	k=10/20/30, h=512/768/1280; ensemble	wt	Official repository
SAPROT	Structure-aware	35M-AF2, 650M-AF2, 650M-PDB	mask, wt	Official repository
ESM-IF1	Inverse folding	—	tf	Official repository
PROTEINMPNN	Inverse folding	v_48_002/010/020/030; ProteinMPNN-Soluble ×4	tf	Official repository
MIF-ST	Inverse folding	—	tf	Official repository
VENUSREM	Retrieval-enhanced	PROSST (K=2048) + MSA	wt	Official repository
RSALOR	MSA + structure	conservation × RSA	—	Official repository

Table 14: ProteinGym zero-shot substitution leaderboard: official 217-assay Average Spearman. Bold red, purple, and black mark the top 3 distinct scores per column at reported precision.

Rank	Model	Seq	Str	Evo	Average	Activity	Binding	Expression	Organismal	Stability
—	**VENUSREM2 (ours)**	✓	✓	✓	**0.556**	**0.541**	**0.495**	**0.557**	**0.494**	**0.691**
1	AIDO PROTEIN-RAG (16B) (Sun et al., 2024)	—	✓	✓	**0.518**	**0.517**	0.426	0.522	**0.491**	0.635
2	VENUSREM (Tan et al., 2025b)	✓	✓	✓	**0.518**	0.495	**0.454**	**0.533**	0.459	**0.650**
3	PROSST (K=2048) (Li et al., 2024)	✓	✓	—	**0.507**	0.476	**0.445**	**0.530**	0.431	**0.653**
5	S3F-MSA (Zhang et al., 2024)	—	✓	✓	0.496	**0.502**	0.440	0.479	0.477	0.581
8	PROTRIEVER (Weitzman et al., 2025)	—	—	✓	0.479	0.487	0.396	0.496	**0.479**	0.537
9	ESCOTT (Tekpinar et al., 2025)	—	✓	✓	0.476	0.499	0.389	0.468	0.466	0.557
12	POET (200M) (Truong Jr & Bepler, 2023)	—	—	✓	0.470	0.494	0.396	0.466	0.475	0.519
14	ESM3 open (1.4B) (Hayes et al., 2025)	✓	✓	—	0.466	0.430	0.400	0.470	0.389	0.641
15	RSALOR (Tsishyn et al., 2025)	—	✓	✓	0.465	0.479	0.416	0.427	0.426	0.575
16	VESPAG (Marquet et al., 2024)	✓	—	—	0.458	0.493	0.370	0.456	0.437	0.533
17	SAPROT (650M AF2) (Su et al., 2023)	✓	✓	—	0.457	0.458	0.378	0.488	0.366	0.592
18	TRANCEPTEVE-L (Notin et al., 2022b)	✓	—	✓	0.456	0.487	0.376	0.457	0.459	0.500
20	GEMME (Laine et al., 2019)	—	—	✓	0.455	0.482	0.383	0.438	0.452	0.519
23	PROTSSN ensemble (Tan et al., 2025c)	✓	✓	—	0.449	0.466	0.366	0.449	0.396	0.568
40	ESM-IF1 (Hsu et al., 2022)	—	✓	—	0.422	0.368	0.389	0.407	0.324	0.624
45	ESM-2 (650M) (Lin et al., 2023)	✓	—	—	0.414	0.425	0.337	0.415	0.368	0.523
47	ESM-1V ensemble (Meier et al., 2021)	✓	—	—	0.407	0.420	0.320	0.429	0.386	0.477
53	MIF-ST (Yang et al., 2023)	✓	✓	—	0.400	0.390	0.321	0.438	0.366	0.485
57	ESM-1B (Rives et al., 2021)	✓	—	—	0.394	0.428	0.287	0.406	0.349	0.500
59	PROGEN3 (3B) (Bhatnagar et al., 2025)	✓	—	—	0.392	0.410	0.287	0.428	0.400	0.438
71	RITA-XL (Hesslow et al., 2022)	✓	—	—	0.373	0.366	0.302	0.414	0.384	0.398
72	CARP (640M) (Yang et al., 2024)	✓	—	—	0.369	0.395	0.273	0.397	0.366	0.412
91	PROTEINMPNN (Dauparas et al., 2022)	—	✓	—	0.257	0.197	0.163	0.198	0.164	0.565

Source and notes. Ranked rows use the official ProteinGym substitutions summary; one entry per model series, matching Table 15. VENUSREM2 uses the six-K PROSST ensemble in Table 17.

Table 15: Complete zero-shot ProteinGym substitutions comparison under the official 217-assay Average Spearman protocol, split by biological category. The same models appear by taxon and MSA depth in Table 16. Bold red, purple, and black mark the top 3 distinct scores per column at reported precision.

Rank	Model	Seq	Str	Evo	Average	Activity	Binding	Expression	Organismal	Stability
—	**VENUSREM2 (ours)**	✓	✓	✓	**0.556**	**0.541**	**0.495**	**0.557**	**0.494**	**0.691**
1	AIDO PROTEIN-RAG (16B) (Sun et al., 2024)	—	✓	✓	**0.518**	**0.517**	0.426	0.522	**0.491**	0.635
2	VENUSREM (Tan et al., 2025b)	✓	✓	✓	**0.518**	0.495	**0.454**	**0.533**	0.459	**0.650**
3	PROSST (K=2048) (Li et al., 2024)	✓	✓	—	**0.507**	0.476	**0.445**	**0.530**	0.431	**0.653**
5	S3F-MSA (Zhang et al., 2024)	—	✓	✓	0.496	**0.502**	0.440	0.479	0.477	0.581
8	PROTRIEVER (Weitzman et al., 2025)	—	—	✓	0.479	0.487	0.396	0.496	**0.479**	0.537
9	ESCOTT (Tekpinar et al., 2025)	—	✓	✓	0.476	0.499	0.389	0.468	0.466	0.557
12	POET (200M) (Truong Jr & Bepler, 2023)	—	—	✓	0.470	0.494	0.396	0.466	0.475	0.519
14	ESM3 open (1.4B) (Hayes et al., 2025)	✓	✓	—	0.466	0.430	0.400	0.470	0.389	0.641
15	RSALOR (Tsishyn et al., 2025)	—	✓	✓	0.465	0.479	0.416	0.427	0.426	0.575
16	VESPAG (Marquet et al., 2024)	✓	—	—	0.458	0.493	0.370	0.456	0.437	0.533
17	SAPROT (650M AF2) (Su et al., 2023)	✓	✓	—	0.457	0.458	0.378	0.488	0.366	0.592
18	TRANCEPTEVE-L (Notin et al., 2022b)	✓	—	✓	0.456	0.487	0.376	0.457	0.459	0.500
20	GEMME (Laine et al., 2019)	—	—	✓	0.455	0.482	0.383	0.438	0.452	0.519
23	PROTSSN ensemble (Tan et al., 2025c)	✓	✓	—	0.449	0.466	0.366	0.449	0.396	0.568
40	ESM-IF1 (Hsu et al., 2022)	—	✓	—	0.422	0.368	0.389	0.407	0.324	0.624
45	ESM-2 (650M) (Lin et al., 2023)	✓	—	—	0.414	0.425	0.337	0.415	0.368	0.523
47	ESM-1V ensemble (Meier et al., 2021)	✓	—	—	0.407	0.420	0.320	0.429	0.386	0.477
53	MIF-ST (Yang et al., 2023)	✓	✓	—	0.400	0.390	0.321	0.438	0.366	0.485
57	ESM-1B (Rives et al., 2021)	✓	—	—	0.394	0.428	0.287	0.406	0.349	0.500
59	PROGEN3 (3B) (Bhatnagar et al., 2025)	✓	—	—	0.392	0.410	0.287	0.428	0.400	0.438
71	RITA-XL (Hesslow et al., 2022)	✓	—	—	0.373	0.366	0.302	0.414	0.384	0.398
72	CARP (640M) (Yang et al., 2024)	✓	—	—	0.369	0.395	0.273	0.397	0.366	0.412
91	PROTEINMPNN (Dauparas et al., 2022)	—	✓	—	0.257	0.197	0.163	0.198	0.164	0.565

Source and notes. Ranked rows are from `https://proteingym.org/benchmarks`; one entry per model series, using its highest official Average Spearman. Seq, Str, and Evo denote sequence, structure, and evolutionary inputs.

C.1 STRATIFIED ANALYSES BY TAXONOMY AND MSA DEPTH

Table 17 tests whether the VENUSREM2 gain is confined to one taxonomic or MSA-depth bin. Virus assays gain 0.064, the low-depth bin 0.055, and the high-depth bin 0.011. The larger low-depth gain shows that the aggregate improvement is not restricted to proteins with abundant homologs. These stratified averages characterize where the gains occur; they do not by themselves isolate the effect of retrieval from calibration or structural attenuation.

Table 18 separates variants by the number of substitutions. The same six-member ensemble leads each of the five depth bins, extending the overall comparison to both single- and multiple-substitution variants.

Table 16: Complete zero-shot ProteinGym substitutions comparison for the same models as Table 15, split by official taxon and MSA-depth bins. MSA depth uses N_{eff}/L: Low < 1, Medium $[1, 100)$, High ≥ 100. Bold red, purple, and black mark the top 3 distinct scores per column at reported precision.

Rank	Model	Seq	Str	Evo	Human	Other euk.	Prok.	Virus	Low	Medium	High
—	**VENUSREM2 (ours)**	✓	✓	✓	**0.562**	**0.615**	**0.601**	**0.540**	**0.553**	**0.557**	**0.623**
1	AIDO PROTEIN-RAG (16B) (Sun et al., 2024)	—	✓	✓	**0.531**	**0.587**	**0.558**	**0.522**	**0.498**	**0.534**	**0.585**
2	VENUSREM (Tan et al., 2025b)	✓	✓	✓	**0.529**	**0.582**	**0.549**	0.492	**0.495**	**0.524**	0.577
3	PROSST (K=2048) (Li et al., 2024)	✓	✓	—	0.516	0.573	**0.549**	0.454	0.465	0.507	**0.580**
5	S3F-MSA (Zhang et al., 2024)	—	✓	✓	0.502	0.558	0.521	0.502	0.469	0.509	0.547
8	PROTRIEVER (Weitzman et al., 2025)	—	—	✓	0.480	0.542	0.492	**0.516**	0.464	0.498	0.512
9	ESCOTT (Tekpinar et al., 2025)	—	✓	✓	0.486	0.537	0.500	0.502	0.462	0.496	0.524
12	POET (200M) (Truong Jr & Bepler, 2023)	—	—	✓	0.482	0.541	0.464	0.491	0.478	0.478	0.510
14	ESM3 open (1.4B) (Hayes et al., 2025)	✓	✓	—	0.480	0.548	0.530	0.406	0.397	0.466	0.575
15	RSALOR (Tsishyn et al., 2025)	—	✓	✓	0.473	0.530	0.496	0.477	0.467	0.468	0.529
16	VESPAG (Marquet et al., 2024)	✓	—	—	0.474	0.532	0.484	0.421	0.427	0.472	0.511
17	SAPROT (650M AF2) (Su et al., 2023)	✓	✓	—	0.478	0.529	0.514	0.320	0.394	0.446	0.546
18	TRANCEPTEVE-L (Notin et al., 2022b)	✓	—	✓	0.473	0.513	0.455	0.461	0.436	0.472	0.490
20	GEMME (Laine et al., 2019)	—	—	✓	0.469	0.516	0.467	0.472	0.446	0.474	0.493
23	PROTSSN ensemble (Tan et al., 2025c)	✓	✓	—	0.470	0.528	0.492	0.370	0.409	0.454	0.524
40	ESM-IF1 (Hsu et al., 2022)	—	✓	—	0.417	0.502	0.498	0.389	0.292	0.427	0.549
45	ESM-2 (650M) (Lin et al., 2023)	✓	—	—	0.457	0.486	0.458	0.261	0.338	0.409	0.513
47	ESM-1V ensemble (Meier et al., 2021)	✓	—	—	0.458	0.464	0.413	0.294	0.316	0.409	0.495
53	MIF-ST (Yang et al., 2023)	✓	✓	—	0.399	0.413	0.457	0.405	0.371	0.404	0.453
57	ESM-1B (Rives et al., 2021)	✓	—	—	0.435	0.491	0.438	0.260	0.365	0.396	0.480
59	PROGEN3 (3B) (Bhatnagar et al., 2025)	✓	—	—	0.410	0.433	0.392	0.414	0.325	0.407	0.453
71	RITA-XL (Hesslow et al., 2022)	✓	—	—	0.396	0.399	0.346	0.394	0.300	0.390	0.412
72	CARP (640M) (Yang et al., 2024)	✓	—	—	0.417	0.395	0.381	0.284	0.331	0.374	0.424
91	PROTEINMPNN (Dauparas et al., 2022)	—	✓	—	0.284	0.394	0.348	0.262	0.187	0.271	0.439

Source and notes. Ranked rows use the official ProteinGym substitutions summary; one entry per model series, matching Table 15. VENUSREM2 uses the six-K PROSST ensemble in Table 17.

Table 17: VENUSREM2 ProteinGym Average Spearman stratified by taxon and MSA depth, recomputed with the official ProteinGym evaluator. MSA depth uses N_{eff}/L: Low < 1, Medium $[1, 100)$, High ≥ 100. Δ is RH minus Raw.

Stratification	Group	Raw	RH	Δ
Taxon	Human	0.524	0.562	+0.038
	Other eukaryotes	0.596	0.615	+0.019
	Prokaryote	0.579	0.601	+0.022
	Virus	0.476	0.540	+0.064
MSA depth	Low	0.498	0.553	+0.055
	Medium	0.517	0.557	+0.040
	High	0.612	0.623	+0.011

Table 18: Complete zero-shot ProteinGym substitutions comparison for the same models as Table 15, split by official mutation-depth bins. Depth is the number of amino-acid substitutions per variant; 5+ denotes at least five substitutions. Bold red, purple, and black mark the top 3 distinct scores per column at reported precision.

Rank	Model	Seq	Str	Evo	1	2	3	4	5+
—	**VENUSREM2 (ours)**	✓	✓	✓	**0.576**	**0.473**	**0.468**	**0.424**	**0.473**
1	AIDO PROTEIN-RAG (16B) (Sun et al., 2024)	—	✓	✓	**0.527**	**0.414**	**0.419**	**0.394**	0.414
2	VENUSREM (Tan et al., 2025b)	✓	✓	✓	**0.534**	**0.397**	0.355	0.322	0.368
3	PROSST (K=2048) (Li et al., 2024)	✓	✓	—	0.521	0.394	0.317	0.277	0.332
5	S3F-MSA (Zhang et al., 2024)	—	✓	✓	0.499	0.333	0.378	0.346	0.383
8	PROTRIEVER (Weitzman et al., 2025)	—	—	✓	0.473	0.320	**0.427**	0.387	0.427
9	ESCOTT (Tekpinar et al., 2025)	—	✓	✓	0.476	0.314	0.377	0.385	**0.434**
12	POET (200M) (Truong Jr & Bepler, 2023)	—	—	✓	0.466	0.299	**0.419**	**0.395**	0.418
14	ESM3 open (1.4B) (Hayes et al., 2025)	✓	✓	—	0.487	0.337	0.307	0.285	0.363
15	RSALOR (Tsishyn et al., 2025)	—	✓	✓	0.474	0.292	0.369	0.392	**0.453**
16	VESPAG (Marquet et al., 2024)	✓	—	—	0.456	0.247	0.346	0.323	0.367
17	SAPROT (650M AF2) (Su et al., 2023)	✓	✓	—	0.458	0.313	0.275	0.268	0.336
18	TRANCEPTEVE-L (Notin et al., 2022b)	✓	—	✓	0.446	0.278	0.349	0.331	0.382
20	GEMME (Laine et al., 2019)	—	—	✓	0.448	0.276	0.330	0.341	0.414
23	PROTSSN ensemble (Tan et al., 2025c)	✓	✓	—	0.459	0.291	0.292	0.244	0.298
40	ESM-IF1 (Hsu et al., 2022)	—	✓	—	0.440	0.350	0.301	0.294	0.358
45	ESM-2 (650M) (Lin et al., 2023)	✓	—	—	0.422	0.248	0.205	0.163	0.218
47	ESM-1V ensemble (Meier et al., 2021)	✓	—	—	0.403	0.215	0.173	0.153	0.218
53	MIF-ST (Yang et al., 2023)	✓	✓	—	0.431	0.263	0.327	0.298	0.298
57	ESM-1B (Rives et al., 2021)	✓	—	—	0.383	0.226	0.172	0.150	0.270
59	PROGEN3 (3B) (Bhatnagar et al., 2025)	✓	—	—	0.388	0.179	0.226	0.204	0.256
71	RITA-XL (Hesslow et al., 2022)	✓	—	—	0.357	0.141	0.148	0.160	0.233
72	CARP (640M) (Yang et al., 2024)	✓	—	—	0.391	0.213	0.172	0.159	0.162
91	PROTEINMPNN (Dauparas et al., 2022)	—	✓	—	0.292	0.260	0.174	0.183	0.278

Source and notes. Ranked rows use the official ProteinGym substitutions summary; one entry per model series, matching Table 15. VENUSREM2 uses the six-K PROSST ensemble in Table 17.

D GLOBAL–LOCAL SIGNAL DIAGNOSTICS

Appendix D.1 defines the profile comparison, and Appendix D.2 checks the consistency of model backgrounds. Appendix D.3 summarizes the cross-benchmark evidence; Appendices D.4, D.5, and D.6 examine ProteinGym, VenusMutHub, and ViroHub, respectively, separating shared preferences from dataset-specific family agreement.

D.1 PROFILE CONSTRUCTION AND SEPARATION PROTOCOL

For the background–local comparison in the Introduction, each well-covered site contributes Pearson correlations across the 20 amino-acid entries between its MSA profile and each of two references: log UniProt abundance and the protein-mean native profile. A2M inputs use aligned-FASTA match states; lowercase insertions are converted to uppercase, gaps map to padding, and retained rows are counted without sequence reweighting or pseudocounts. Duplicate assays are collapsed by WT sequence, yielding 27,763, 148,983, and 29,416 complete profiles from 148, 527, and 52 representative sequences in ProteinGym, VenusMutHub, and ViroHub. Joint alignment requires both correlations to be at least 0.3, yielding rates of 19.7%, 24.8%, and 38.1%, respectively, over the complete profile sets. This descriptive threshold measures background–local agreement; it does not establish independence between MSA and model predictions.

For Figure 5, the separation index is the fraction of well-covered sites whose MSA preference is not jointly aligned with log UniProt abundance and the protein-mean native profile at $r \geq 0.3$. All 217 ESM-2 wild-type-interface ProteinGym assays have paired Raw–MSA scores; 177 assays contain at least 20 complete site profiles and span 148 proteins. Shared-protein assays are averaged before

analysis, and the site map retains the assay with the broadest complete profile. The response is the protein-mean paired Spearman gain. Intervals use 10,000 protein-level bootstrap draws; partial rank correlation controls for length, $\log_{10}$ MSA depth, MSA occupancy, and Raw Spearman.

D.2 Native-background diagnostics

Figure 9 extends the shared-preference pattern in Figure 1 with complementary diagnostics on the original 37-model cohort: it overlays model profiles, quantifies pairwise agreement, separates readout scales, and reports protein–configuration heterogeneity in directional alignment. The profiles have median pairwise Pearson $r = 0.995$; the consensus–corpus correlation is 0.993. Alignment c spans 0.01–0.99, motivating a gate that adapts to the queried protein. ESM3 is the clearest checkpoint-level departure, consistent with multimodal pretraining (Hayes et al., 2025).

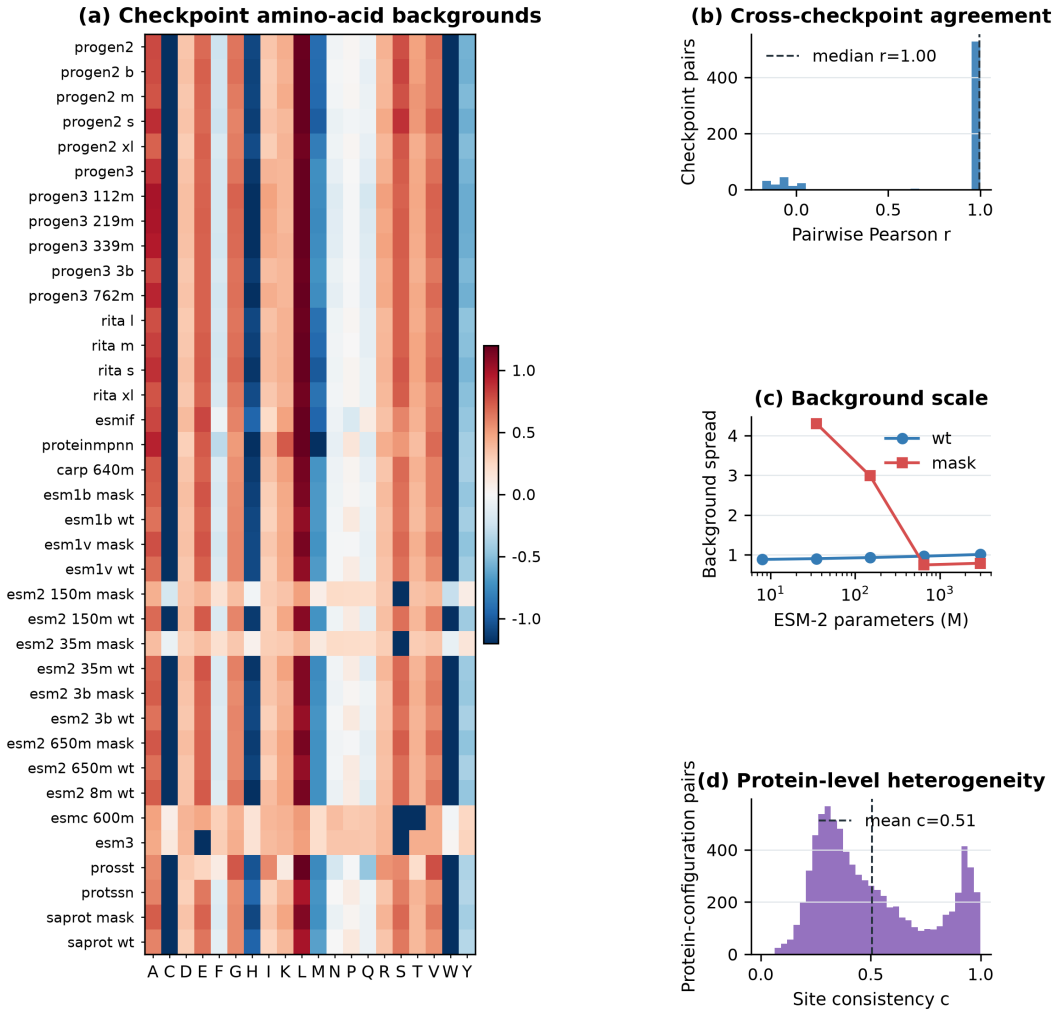


Figure 9: Native-background diagnostics on the original 37-model cohort. **(a)** Protein-level amino-acid profiles versus UniProtKB/Swiss-Prot composition. **(b)** Pairwise Pearson correlations among those profiles. **(c)** Readout-scale comparison across scoring interfaces. **(d)** Mean directional alignment $c = L^{-1} \sum_i \mathrm{corr}_a(R_i, b)$ over protein–configuration pairs.

D.3 Cross-benchmark replication

Table 19 compares the three benchmark-specific checks. Pairwise is the median correlation among model backgrounds (49 on ProteinGym and VenusMutHub, 37 on ViroHub); corpus compares their

consensus with UniProt composition; native–MSA reports assay-level median and pooled correlations for the original 37-model consensus; aligned sites have Pearson $r \geq 0.3$ with both log corpus abundance and the protein-mean native background.

Table 19: Cross-benchmark summary of native-background agreement and local MSA preference separation. Complete sites are unique WT positions with all required profiles.

Dataset	Pairwise	Corpus	Native–MSA		Complete sites	Aligned sites
	r	r	Median r	Pooled r	N	(%)
ProteinGym	0.995	0.991	0.965	0.758	27,763	19.7
VenusMutHub	0.996	0.989	0.964	0.863	148,983	24.8
ViroHub	0.990	0.885	0.976	0.621	29,416	38.1

Checkpoint backgrounds are nearly identical across benchmarks, whereas pooled native–family agreement weakens and most sites fail the joint-alignment criterion. Figures 10–12 retain the dataset-specific distributions and family-level outcomes.

D.4 ProteinGym replication

Figure 10 shows that a strong protein-mean native–MSA correlation coexists with substantial site-level separation: only 19.7% of complete sites align jointly with corpus and native backgrounds (Table 19). Agreement after averaging therefore does not imply interchangeable local preferences. Gains occur across all four backbone families; Appendix E reports their magnitudes.

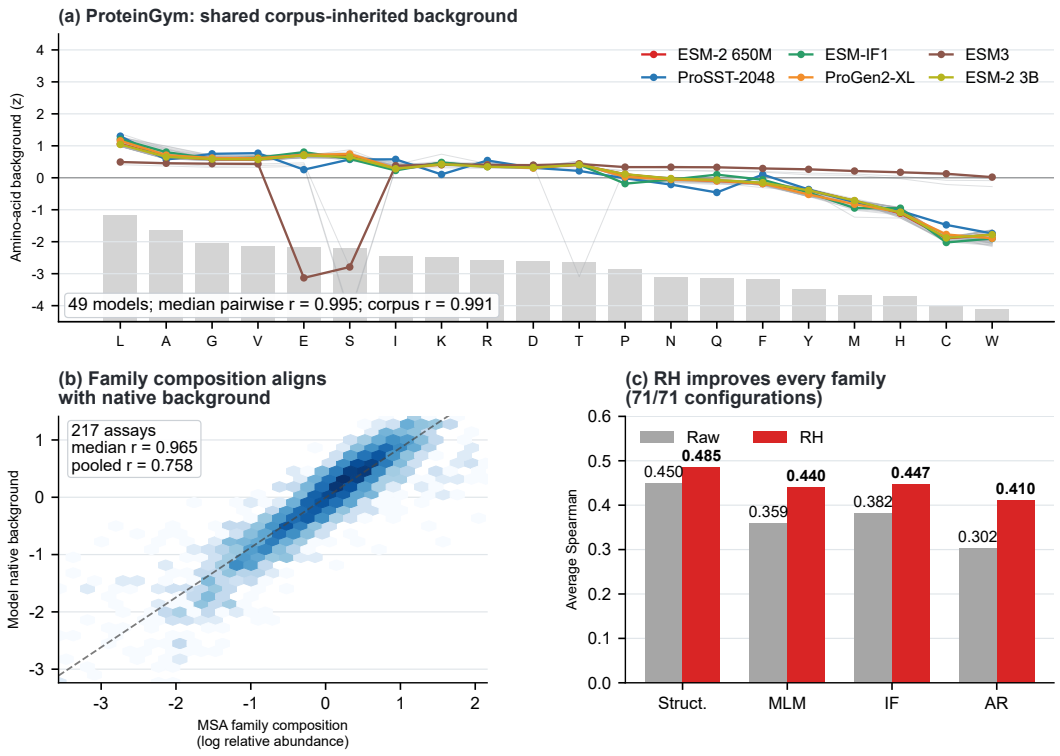


Figure 10: ProteinGym replication. **(a)** Profiles from 49 models and UniProt abundance (gray bars). **(b)** The original 37-model native-background consensus versus MSA abundance (central 99%; correlations use all 217 × 20 pairs). **(c)** Raw–RH Spearman by family across 71 configurations.

D.5 VENUSMUTHUB REPLICATION

VenusMutHub reproduces the separation between shared model backgrounds and local family preferences: only 24.8% of complete sites are jointly aligned (Figure 11; Table 19). Thus the ProteinGym pattern extends to this broader assay collection. Family-level performance in Appendix E provides the corresponding prediction comparison.

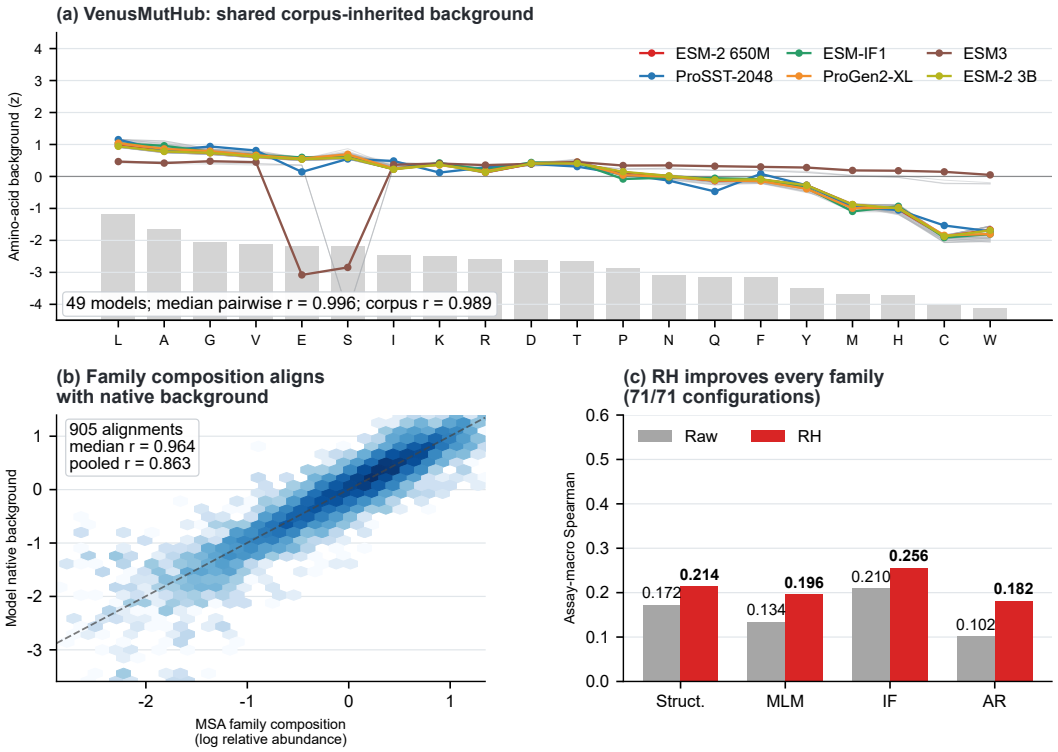


Figure 11: VenusMutHub replication. **(a)** Profiles from 49 models and UniProt abundance (gray bars). **(b)** The original 37-model native-background consensus versus MSA abundance (central 99%; correlations use all 905×20 pairs). **(c)** Raw–RH Spearman by family across 71 configurations.

D.6 VIROHUB REPLICATION

Figure 12 shows the weakest pooled native–MSA agreement ($r = 0.621$) despite a high assay-level median, with 38.1% of complete sites jointly aligned. The median and pooled statistics summarize different levels of variation; strong within-protein agreement can therefore coexist with weaker agreement across the collection. Appendix E reports the gains across model families.

E FULL CROSS-FAMILY RESULTS

All ProteinGym results use one versioned evaluation pipeline. Of 72 configurations, 71 have complete paired Raw–RH coverage on all 217 assays; VENUSREM was not rescored with RH and is excluded from paired comparisons. The comparison is uniform in sign and uneven in magnitude (Table 20): autoregressive models gain the most, and structure-aware encoders still gain after geometry is already in the backbone. Table 21 shows that the PROSST ensemble improves every reported metric, not only Spearman.

The aggregate gain is not confined to one of the five ProteinGym biological categories. Across those categories and 71 configurations, all 355 configuration–category comparisons improve in Spearman. Table 22 reports the corresponding family means; Stability shows the largest overall increase, gaining 0.095 from 0.489 to 0.584.

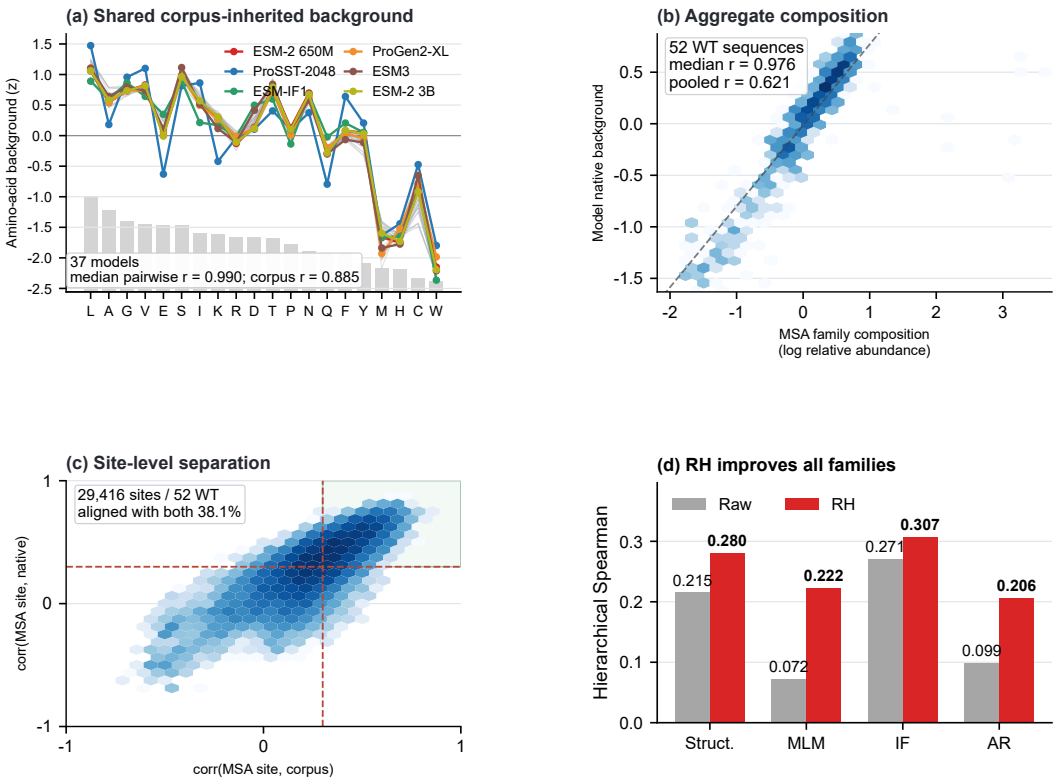


Figure 12: ViroHub replication. **(a)** Profiles from 37 models and UniProt abundance. **(b)** Their native-background consensus versus MSA abundance. **(c)** MSA site preferences versus corpus and native backgrounds. **(d)** Raw–RH Spearman by family across 71 configurations.

Table 20: ProteinGym Average Spearman by backbone family across 71 paired configurations and 217 assays. N is the number of configurations. Mean and median change are paired RH–Raw differences; Wins / N counts configurations with positive change.

Family	N	Raw	RH	Mean change	Median change	Wins / N
Structure-aware	26	0.450	0.485	+0.035	+0.029	26 / 26
Masked sequence LM	20	0.359	0.440	+0.081	+0.062	20 / 20
Inverse folding	10	0.382	0.447	+0.065	+0.070	10 / 10
Autoregressive LM	15	0.302	0.410	+0.108	+0.098	15 / 15

Table 21: VENUSREM2 ProteinGym metrics, reported as assay-macro mean $\pm$ standard deviation over 217 assays. These means are not official Average Spearman. Δ is RH–Raw; the RH column is the six-K PROSST ensemble; top-recall denotes top-variant recall.

Metric	Raw	RH	Δ
Spearman	0.541 ± 0.180	0.575 ± 0.159	+0.034
NDCG	0.794 ± 0.138	0.813 ± 0.130	+0.019
AUC	0.796 ± 0.100	0.815 ± 0.091	+0.019
MCC	0.422 ± 0.164	0.450 ± 0.151	+0.028
Top-recall	0.261 ± 0.131	0.270 ± 0.131	+0.008

VenusMutHub repeats the same family order (Table 23): every family improves, and autoregressive models again gain the most. Table 25 summarizes the paired benchmark units: UniProt IDs

Table 22: Mean ProteinGym Average Spearman by biological category and backbone family across the 71 configurations. Counts in the header are the number of configurations in that family. Structure = structure-aware; Masked LM = masked sequence LM; AR LM = autoregressive LM.

Category	Structure (26)		Masked LM (20)		Inverse folding (10)		AR LM (15)		All (71)	
	Raw	RH	Raw	RH	Raw	RH	Raw	RH	Raw	RH
Activity	0.442	0.486	0.371	0.453	0.318	0.414	0.321	0.434	0.379	0.456
Binding	0.381	0.407	0.288	0.353	0.327	0.384	0.251	0.313	0.320	0.369
Expression	0.459	0.487	0.369	0.433	0.382	0.429	0.334	0.400	0.397	0.445
Organismal Fitness	0.378	0.423	0.295	0.394	0.294	0.373	0.334	0.397	0.334	0.402
Stability	0.589	0.623	0.474	0.567	0.588	0.635	0.272	0.507	0.489	0.584

for ProteinGym, wild-type sequence clusters for VenusMutHub, and phenotype–backbone cells for ViroHub.

Table 23: VenusMutHub macro-averaged assay-level Spearman by backbone family across 71 paired configurations and 905 assays. Column definitions match Table 20.

Family	N	Raw	RH	Mean change	Median change	Wins / N
Structure-aware	26	0.172	0.214	+0.042	+0.032	26 / 26
Masked sequence LM	20	0.134	0.196	+0.062	+0.050	20 / 20
Inverse folding	10	0.210	0.256	+0.046	+0.049	10 / 10
Autoregressive LM	15	0.102	0.182	+0.080	+0.070	15 / 15

ViroHub extends the comparison to a virus-focused distribution with no ProteinGym assay overlap (Table 24). The largest mean lift occurs for masked sequence models, while every configuration improves in every family.

Table 24: ViroHub hierarchically aggregated Spearman by backbone family across 71 paired configurations and 89 assays, excluding wild-type controls. Column definitions match Table 20.

Family	N	Raw	RH	Mean change	Median change	Wins / N
Structure-aware	26	0.215	0.280	+0.065	+0.063	26 / 26
Masked sequence LM	20	0.072	0.222	+0.151	+0.147	20 / 20
Inverse folding	10	0.271	0.307	+0.036	+0.036	10 / 10
Autoregressive LM	15	0.099	0.206	+0.107	+0.110	15 / 15

Table 25: Mean paired RH–Raw changes across 71 configurations. Δ follows each benchmark's primary aggregation: Average Spearman for ProteinGym, assay-macro Spearman for VenusMutHub, and hierarchical Spearman for ViroHub. Units follow the paired bootstrap in Appendix B.1; Wins / N counts positive changes.

Benchmark	Resampling unit	Units	Mean Δ	Wins / N
ProteinGym	UniProt ID	186	+0.068	71 / 71
VenusMutHub	wild-type sequence	527	+0.056	71 / 71
ViroHub	phenotype–backbone cell	52	+0.094	71 / 71

F ADAPTIVE-STAGE AND COMPONENT AUDITS

F.1 STAGE-WISE ABLATIONS

Stages across backbones. Configurations with complete staged fields follow the sequence Raw, entropy-weighted retrieval, coherence-gated correction, RSA shrinkage, and pLDDT shrinkage. Ungated correction provides a separate gate ablation. Component operators are selected on the 21-assay validation split—fold 8 of a ten-fold partition formed after shuffling sorted assay identifiers with seed 42—and then frozen. The per-configuration tables are grouped at the end in Appendix I.2; gated correction raises the aggregate mean on all three benchmarks, while RSA shrinkage improves all 71 configurations on each benchmark.

Controlled PROSST-2048 component isolation. Table 1 applies every downstream operator to the same entropy-weighted MSA field and uses the complementary calibration coefficient $1 - \alpha$. On the random 21-assay validation split, retrieval contributes 0.0201 of the total 0.0276 improvement; gating adds 0.0014, and RSA and pLDDT shrinkage add a further 0.0061, yielding 0.5849 from a Raw value of 0.5573. The held-out 196-assay ProteinGym test set follows the same order under official aggregation: retrieval contributes 0.0169, gating 0.0040, RSA shrinkage 0.0045, and pLDDT shrinkage 0.0011, producing a final score of 0.5337 from 0.5072. The ungated row is a gate intervention rather than a cumulative stage.

The coherence gate is evaluated independently of the component table. Relative to the same entropy-weighted adaptive mix, gated correction improves $71/71$, $62/71$, and $71/71$ configurations on ProteinGym, VenusMutHub, and ViroHub, respectively; relative to ungated correction, the corresponding counts are $71/71$, $37/71$, and $68/71$.

F.2 ADAPTIVE SCALE ALLOCATION

Definitions and scale constraint. All logarithms are natural unless a base is shown. For $n = |\mathcal{I}| > 0$, let $\mathcal{D} = \{(i, a) : i \in \mathcal{I},\ a \in \mathcal{A} \setminus \{x_i\}\}$, so $|\mathcal{D}| = 19n$. For $X \in \{R, E\}$, the scales in Equation 8 are

$$\overline{\Delta}^X = \frac{1}{19n} \sum_{(i,a)\in\mathcal{D}} \Delta^X_{i,a}, \qquad s^2_X = \frac{1}{19n - 1} \sum_{(i,a)\in\mathcal{D}} \big(\Delta^X_{i,a} - \overline{\Delta}^X\big)^2. \tag{24}$$

These statistics cover all non-WT amino acids at mapped sites, irrespective of which variants an assay measured. The normalizer in Equation 2 is constant across amino acids at a site, so $\Delta^E_{i,a} = f_{i,a} - f_{i,x_i}$. It also leaves the site-wise Pearson correlation unchanged. Computing the normalizer over just the amino-acid columns would therefore give the same retrieval cues and contrasts, although the fields would differ by row offsets.

Derivation and scope. Write $\rho = \rho_{\mathrm{MSA}}$. Normalized entropy, clipped WT support, and $(1 - r_{R,E})/2$ lie in $[0, 1]$; their convex interpolation therefore gives $\rho \in [0, 1]$. For $s_R, s_E > 0$, multiplying the scale constraint by its positive denominator yields

$$\rho(1 - \alpha)s_R = (1 - \rho)\alpha s_E \quad \Longrightarrow \quad \rho s_R = \alpha\big[\rho s_R + (1 - \rho)s_E\big]. \tag{25}$$

The bracket is positive, giving the unique solution in Equation 8. Its denominator is at least its nonnegative numerator, hence $0 \le \alpha \le 1$; the endpoints are $\alpha(0) = 0$ and $\alpha(1) = 1$. The constraint allocates weighted marginal standard deviations, not a fraction of mixture variance, which would also contain covariance. The entropy interpolation and gate are empirical design choices; this algebra establishes the stated scale allocation, not a guarantee of fitness improvement.

Numerical and degenerate inputs. The implementation floors WT probabilities at 10^{-12} before taking their logarithms and uses zero correlation for constant profiles. The scale solver rejects nonfinite scales, $s_E \le 0$, or a non-positive mapping denominator. With $s_R = 0 < s_E$ and $\rho < 1$, it returns $\alpha = 0$, outside the positive-scale interpretation above. The scoring wrapper uses $\alpha = 0$ when no usable MSA exists; an adaptive-estimation failure with an available MSA invokes the logged fixed fallback $\alpha = 0.8$ and calibration weight 0.2. These are numerical fallback conventions, not solutions to the positive-scale constraint.

Fixed-coefficient family-field sweep. Figure 13 replaces the entropy-adaptive coefficient of Equation 8 with a fixed mixing coefficient $\alpha \in \{0.0, 0.1, \ldots, 1.0\}$, evaluated on the same PROSST-2048 fields and the same 21-assay validation split as Table 1, under both the retrieval-only readout and the full pipeline. Three observations support the adaptive design. First, both retrieval-only endpoints degrade: α=0 reduces to the Raw field and α=1 collapses to 0.4093, so neither channel suffices alone and the frozen backbone remains necessary even where family evidence is available. Second, performance forms a broad plateau over $\alpha \in [0.6, 0.9]$ (0.567–0.574 retrieval-only), showing stable performance across this coefficient range. Third, the per-protein entropy-adaptive coefficient reaches 0.5774 retrieval-only and 0.5849 with the full pipeline, matching or exceeding every fixed coefficient without per-protein tuning. The plateau shape is unchanged when fold means are averaged over all ten folds of the partition (0.531–0.537 retrieval-only over $\alpha \in [0.6, 0.9]$; 0.398 at α=1). Shaded intervals are 95% assay-level bootstrap intervals over the 21 validation assays.

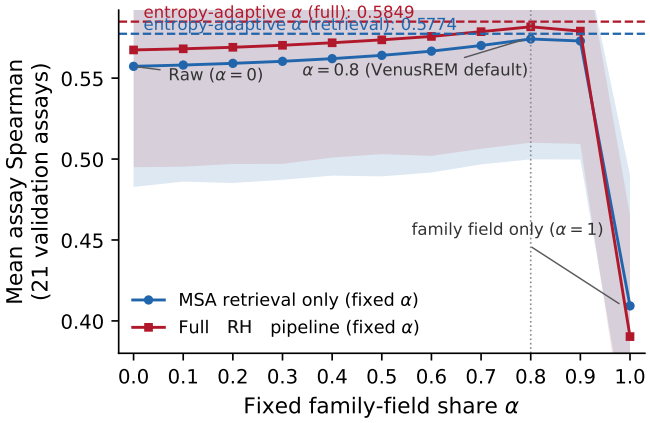


Figure 13: Mean assay Spearman on the 21-assay validation split (PROSST-2048) under a fixed mixing coefficient α, for retrieval only and the full pipeline; bands are 95% assay-bootstrap intervals. Dashed lines: the entropy-adaptive α of Equation 8. Both endpoints degrade, and the adaptive coefficient matches or exceeds the best fixed coefficient.

Table 26 isolates the ProteinGym gate comparison: the fixed ungated correction improves 43 of 71 configurations, whereas coherence gating improves all 71 and yields the larger mean gain over retrieval.

Table 26: Coherence-gate ablation on 71 ProteinGym configurations (official Average Spearman).

Comparison	Mean Δ	Median Δ	Wins / N	Ties
Gated − no correction	+0.012	+0.010	71 / 71	0
Ungated $\kappa = 1$ − no correction	+0.003	+0.003	43 / 71	0
Gated − ungated $\kappa = 1$	+0.010	+0.009	71 / 71	0

F.3 CALIBRATION IDENTITIES AND LEAVE-ONE-SITE-OUT VALIDATION

Native-channel correction. For the background in Equation 10, $\bar{b} = 20^{-1} \sum_{a \in \mathcal{A}} b_a$ and $s_b^2 = 19^{-1} \sum_{a \in \mathcal{A}} (b_a - \bar{b})^2$. The implementation sets $\widehat{b} = 0$ when $s_b \leq 10^{-8}$ and protects Pearson denominators at 10^{-12}. For fixed α and $\kappa = \kappa(R)$, define the corrected field

$$G_{i,a} = (1 - \alpha)(R_{i,a} - \kappa \widehat{b}_a) + \alpha E_{i,a}. \tag{26}$$

Subtracting G_{i,x_i} gives Equation 12: linearity preserves the family term $\alpha \Delta^E_{i,a}$ and scales the native correction by $1 - \alpha$. Since Pearson correlations lie in $[-1, 1]$, $0 \leq \kappa \leq 1$; non-positive mean coherence gives $\kappa = 0$. This establishes which channel is changed, without assuming that the background is purely bias or that its removal always improves prediction.

Excluding the focal site. For $L > 1$, removing site i from background construction gives

$$b_a^{(-i)} = \log\left(\frac{1}{L-1}\sum_{\substack{1\le j\le L\\ j\neq i}} e^{R_{j,a}}\right), \qquad (27)$$

$$c_{\mathrm{LOSO}} = \frac{1}{L}\sum_{i=1}^{L}\mathrm{corr}_{a\in\mathcal{A}}\left(R_i, b^{(-i)}\right), \qquad \kappa_{\mathrm{LOSO}} = \mathrm{ReLU}(c_{\mathrm{LOSO}}).$$

Here LOSO denotes leave-one-site-out; removing direct self-inclusion does not imply statistical independence between sites. Across 217 ProSST-2048 assays and 187 unique WT sequences, leave-one-site-out estimation preserves the full-protein gate (Spearman $\rho = 0.985$, Pearson $r = 0.988$, median $|\Delta\kappa| = 0.010$). Figure 14b compares the assay-level Spearman gain of adaptive calibration over the same MSA retrieval field and reports assay-bootstrap intervals within coherence quartiles. Coherence has no detectable rank association with calibration gain ($\rho = 0.016$, $p = 0.811$), distinguishing stability of the gate estimate from prediction of the gain magnitude.

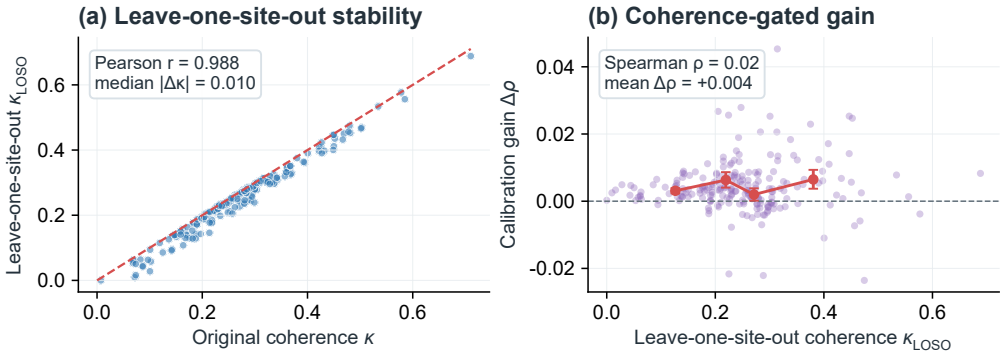


Figure 14: Leave-one-site-out coherence across 217 ProSST-2048 assays. **(a)** κ_{LOSO} preserves the full-protein gate. **(b)** Calibration gain over MSA retrieval versus coherence; red points show quartile means and 95% bootstrap intervals.

F.4 Exposure-and-confidence shrinkage

Normalize pLDDT by $\ell_i = \mathrm{pLDDT}_i/100$. Let $\mathcal{J}_r = \{i : r_i > 0\}$ and $\mathcal{J}_\ell = \{i : \ell_i > 0\}$ after clipping the covariates to $[0, 1]$. The implementation uses

$$\bar{r} = \frac{1}{|\mathcal{J}_r|}\sum_{i\in\mathcal{J}_r} r_i, \qquad \bar{\ell} = \frac{1}{|\mathcal{J}_\ell|}\sum_{i\in\mathcal{J}_\ell} \ell_i, \qquad \bar{q} = 1 - \bar{\ell}, \qquad (28)$$

with a mean of zero for an empty positive-entry set. Each supplied covariate contributes its factor in Equation 13; an absent covariate array contributes one. A zero entry in a supplied array follows the formula, rather than being treated as a missing array. When neither array is supplied, $\lambda_i = 1$.

For any $u, \bar{u} \in [0, 1]$, $0 \le \mathrm{ReLU}(u - \bar{u}) \le 1$, so each factor lies in $[0, 1]$. Their product therefore satisfies $0 \le \lambda_i \le 1$ and $|S_{i,a}| \le |S^{\mathrm{cal}}_{i,a}|$. Positive weights preserve the sign of a site contrast; a zero weight sets it to zero. Because weights vary across sites, this does not imply unchanged cross-site ranks or preservation of the sign of a multi-substitution sum. Table 1 tests the empirical contributions of the two factors.

F.5 Variant aggregation and ensemble standardization

For a fixed protein and member k, write $\mathcal{M}_v$ for variant v's set of substitutions, with at most one substitution per site. Equation 15 gives $S_k(v) = \sum_{(i,a)\in\mathcal{M}_v} S_{k,i,a}$ and a WT score of zero. Consequently, for substitutions m_1, m_2 at distinct sites, $S_k(\{m_1, m_2\}) - S_k(\{m_1\}) - S_k(\{m_2\}) +$

$S_k(\varnothing) = 0$. This is additivity of the fixed-field readout; recomputing fields on a mutant background is a different operation.

Ensembling has also been used in ESM-1v and ProtSSN (Meier et al., 2021; Tan et al., 2025c). For the six-member ensemble, let $N_A = |\mathcal{V}_A| > 0$ and assume complete finite member predictions, as in the evaluated ensembles. Its normalization is

$$\mu_{k,A} = \frac{1}{N_A} \sum_{v \in \mathcal{V}_A} S_k(v), \qquad \sigma^2_{k,A} = \frac{1}{N_A} \sum_{v \in \mathcal{V}_A} \big(S_k(v) - \mu_{k,A}\big)^2, \tag{29a}$$

$$z_{k,A}(v) = \begin{cases} \big(S_k(v) - \mu_{k,A}\big)/\sigma_{k,A}, & \sigma_{k,A} > 0, \\ 0, & \sigma_{k,A} = 0. \end{cases} \tag{29b}$$

Equation 16 averages these six $z_{k,A}$ values with weight $1/6$, including constant members as zero. Only model predictions enter $\mu_{k,A}$ and $\sigma_{k,A}$; experimental fitness values are used for evaluation, not standardization. Individual harness fields depend on the protein, MSA, structure, and backbone. The ensemble additionally depends on the declared candidate set $\mathcal{V}_A$: changing it can change member scales and hence the combined ranking, despite leaving every member prediction fixed.

G PARKIN STABILITY LANDSCAPE

Parkin is a 465-residue RING-between-RING E3 ubiquitin ligase. Its VAMP-seq abundance landscape provides a cellular readout of mutational effects on folding stability and proteostasis. The original study linked reduced abundance to thermodynamic destabilization and proteasomal degradation, while also identifying local degradation signals that need not destabilize the native fold (Clausen et al., 2024). Its ubiquitin-like Ubl, RING0, RING1, IBR, REP/tether, and RING2 architecture couples folded zinc-binding domains to autoinhibitory linkers (Kumar et al., 2015). Of 8,756 measured substitutions across 462 profiled sites, 89.2% exhibit global–local separation. Raw ESM-2 650M with the wild-type interface reaches 0.503 Spearman, MSA retrieval 0.585, and RH 0.645; abundance-class AUC rises from 0.769 to 0.815 and 0.857 (Figure 6).

The rank shift follows domain organization (Figure 6d). RH improves every folded domain, including RING0 from 0.425 to 0.627, IBR from 0.727 to 0.786, and RING2 from 0.427 to 0.529; REP/tether rises from −0.083 to 0.221. The flexible Ubl–RING0 linker instead falls from 0.363 to 0.176. N232A lies at the 0.4th measured percentile but is ranked at the 88.8th Raw percentile; MSA retrieval and RH move it to 58.0 and 37.3. R396H moves in the opposite direction, from 39.7 to 71.7 and 78.0, toward its measured 77.6th percentile (Figure 6e).

Cross-protein region audit. For every ProteinGym assay, we split substitutions by AlphaFold2 confidence—flexible at pLDDT < 50, ordered otherwise—and recompute within-region Spearman under the same ESM-2 650M wild-type readout, retaining subsets with at least 20 variants at five or more sites. Ordered regions improve in 83% of 214 qualifying assays (mean $\Delta = +0.056$; 5.6% degrade by more than 0.02). Among 65 qualifying flexible regions, 57% decrease, with mean $\Delta = -0.020$. The complete readout therefore exhibits a regional contrast: higher average ranking agreement in ordered regions accompanies lower average agreement in low-confidence regions.

H SCORING INTERFACES, RUNTIME, AND EPISTASIS AUDITS

H.1 AMORTIZED SCORE-FIELD INFERENCE

Figure 7 measures post-forward processing after one shared PROSST field has been constructed per protein. Across 2.47 million ProteinGym variants, indexed lookup reduces mutation scoring from 114.56 to 5.53 seconds (20.7×), while vectorized counting reduces MSA count-matrix construction from 3,026.75 to 48.61 seconds (62.3×). Given native and family fields R and E, constructing the complete readout costs $O(20L)$, and scoring an M-substitution variant costs $O(M)$ indexed lookup; the backbone parameters remain frozen. Model-resident execution, batched tokenization, vectorized field construction, and indexed lookup remove repeated work from mutation-wise loops.

H.2 ARCHITECTURE-WIDE SCORING-INTERFACE AUDIT

For a substitution $x_i \to a$, all three interfaces form the contrast $\Delta^R_{i,a} = R_{i,a} - R_{i,x_i}$ from Equation 3, but they differ in the conditioning context used to produce R. A masked marginal uses $R^{\text{mask}}_{i,a} = \log q_{\theta,i}(a; x_{\setminus i})$: the wild-type residue is masked and a full length-L field requires L masked contexts, which can be batched. A wild-type or unmasked field uses $R^{\text{wt}}_{i,a} = \log q_{\theta,i}(a; x)$: a single forward pass scores all sites, but the wild-type residue remains in the input, so this is a ranking interface rather than a masked conditional. Teacher forcing uses $R^{\text{tf}}_{i,a} = \log q_{\theta,i}(a; x_{<i})$ and includes structure when the backbone requires it (Equation 1). It supplies conditionals under a declared factorization, and one native-context field can be reused across variants. These definitions are held fixed within every Raw–RH pair (Table 27); RH changes the readout field, not the conditioning convention.

Table 27: Conditioning and cost of three scoring interfaces. Masked inputs can be batched; teacher forcing hides the target residue at its own conditional.

Interface	Context at site i	WT residue visible	Field construction	Multi-mutant scoring
Masked marginal	$x_{\setminus i}$	No	L masked inputs	Additive native field
Wild-type field	Full native x	Yes	1 forward pass	Additive native field
Teacher forcing	Prefix $x_{<i}$	No (prefix only)	1 forward pass	Additive unless rescored

Bidirectional encoders: masked versus wild-type fields. Across eight paired encoder checkpoints, both masked and wild-type interfaces improve on all three benchmarks, and the same global–local gain holds across conditioning conventions (Table 28).

Table 28: Mean assay Spearman across eight paired checkpoints: ESM-1B, ESM-1V, ESM-2 8M/35M/150M/650M/3B, and SAPROT 650M-AF2. RH uses entropy weighting. ProteinGym uses assay means (not official Average Spearman); both other benchmarks use macro-averages. Wild-type control rows are excluded from ViroHub evaluation.

Dataset	Masked			Wild-type		
	Raw	RH	Δ	Raw	RH	Δ
ProteinGym	0.389	0.468	+0.079	0.395	0.458	+0.062
VenusMutHub	0.157	0.218	+0.061	0.142	0.187	+0.045
ViroHub	0.082	0.239	+0.156	0.076	0.263	+0.188

Autoregressive and inverse-folding models: full-sequence likelihood versus a native field. For these architectures, full-sequence scoring teacher-forces each mutant and aggregates likelihood over the sequence, allowing the substitution to alter later conditionals. Native-field scoring instead teacher-forces the wild-type once and reads the mutant–wild-type contrast only at the mutated position. Table 29 pairs these routes on the same frozen checkpoints and ProteinGym assays.

Table 29: ProteinGym interface comparison on 217 paired assays. Full-sequence values follow the official aggregate; Δ is native field minus full-sequence log-likelihood, and wins are paired at assay level.

Family	Model	Full sequence	Native field	Δ	Wins
Inverse folding	ESM-IF1	0.422	0.421	−0.001	95/217
Autoregressive	PROGEN2-S	0.336	0.288	−0.048	47/217
Autoregressive	PROGEN2-M	0.380	0.337	−0.043	54/217
Autoregressive	PROGEN2-B	0.378	0.330	−0.048	58/217
Autoregressive	PROGEN2-L	0.380	0.327	−0.053	53/217
Autoregressive	PROGEN2-XL	0.391	0.351	−0.040	67/217

The two routes induce almost the same ESM-IF1 ranking (assay-level Spearman 0.966), consistent with structure conditioning localizing the likelihood change. For PROGEN2, however, full-sequence

likelihood retains useful downstream context and exceeds the native field by 0.040–0.053 across the five scales. The reusable field is therefore an architecture-compatible interface for RH, not an assertion that discarding all contextual propagation must improve the backbone score.

H.3 PROTEINMPNN REVERSES THE AUTOREGRESSIVE TREND

Official score versus teacher forcing. Unlike the stable ESM-IF1 comparison and the full-sequence advantage of PROGEN2, official PROTEINMPNN scoring samples a decoding order while rebuilding each mutant sequence, making the score order-dependent. Our native teacher-forcing route evaluates one conditional amino-acid field on the native sequence and applies mutant-minus-wild-type lookup (Table 30). On identical AlphaFold2 structures, this field raises mean assay Spearman from 0.271–0.290 to 0.376–0.401 across eight checkpoints while reducing wall-clock cost by 219–$601\times$; at v_48_020, 204 of 217 assays improve (Figure 15b). The opposite directions of the ProteinMPNN and ProGen2 comparisons show that the effect of native-field scoring depends on the architecture and scoring convention. Conditional rescoring then exposes a non-additive component for the following epistasis diagnostic.

Table 30: PROTEINMPNN official random-order versus native teacher-forcing (TF) scoring.

Route	Context	Evaluations	Order	Multi-mutant score
Official random order	Mutant + structure	Per mutant	Yes	May be non-additive
Native TF	Native + structure	Per protein	No	Additive
Conditional TF	Mutant background + structure	Per background	No	Non-additive

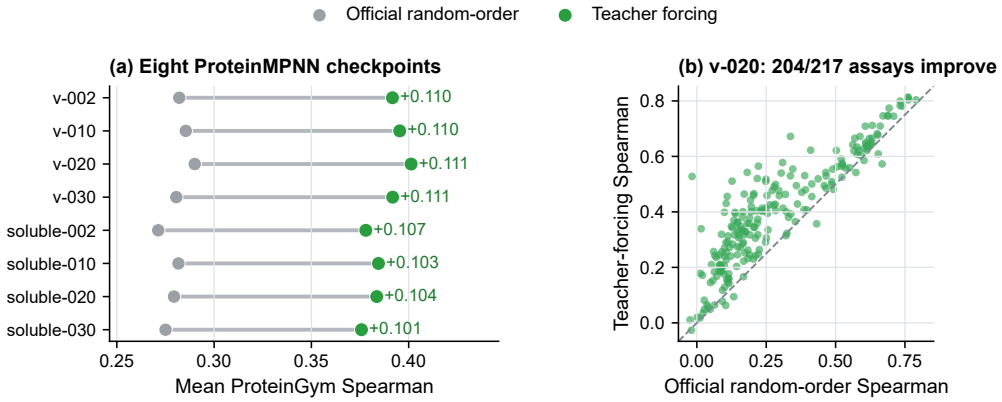


Figure 15: PROTEINMPNN teacher forcing versus official random-order scoring on AlphaFold2 structures. **(a)** Mean assay Spearman across eight checkpoints. **(b)** Per-assay Spearman at v_48_020.

H.4 PAIRWISE EPISTASIS DIAGNOSTIC

PROTEINMPNN epistasis evaluation. We evaluate epistasis on all 535,917 complete wild-type/single/single/double quartets in the GB1 binding landscape of Olson et al. (2014). Teacher-forced variant scores are additive by construction, with maximum numerical deviation 4.8×10^{-7}, so we also rescore mutation A on the B background and vice versa. Table 31 separates main effects from interactions: teacher forcing is substantially stronger for single- and double-mutant fitness, while epistasis correlations remain near zero for both the official random-order score, 0.005 with interval $[-0.002, 0.013]$, and conditional teacher forcing, 0.012 with interval $[-0.005, 0.029]$.

Table 31: PROTEINMPNN v_48_020 on Olson GB1; ρ_ϵ is Spearman correlation with measured pairwise epistasis.

Subset	Quartets	Single TF	Single official	Double TF	Double official	ρ_ϵ official	ρ_ϵ conditional TF
All	535,917	0.306	0.105	0.284	0.165	0.005	0.012
Non-floor	400,118	0.312	0.091	0.292	0.152	0.012	0.022

The focused GB1 diagnostic separates representation from readout: the official and conditional routes recover little pairwise-interaction signal, while the lower-cost additive field retains substantially more main-effect signal under the same frozen backbone and structure inputs.

I ADDITIONAL BENCHMARK TABLES

The following records separate three questions: category-specific ranking in Appendix I.1, component contributions in Appendix I.2, and alternative evaluation criteria in Appendix I.3. They retain the configuration-level evidence behind the compact summaries and mechanistic audits.

I.1 PER-CATEGORY RAW–RH LEADERBOARDS

Tables 32, 33, and 34 give every Raw and RH score for ProteinGym functions, VenusMutHub assay categories, and viral phenotypes, respectively. Aggregation follows each benchmark's primary protocol, and the Overall columns reproduce Tables 20, 23, and 24. ProSST and VENUSREM2 use the wild-type readout described in Appendix C.

Table 32: Unified Raw and RH leaderboard on ProteinGym (official Average Spearman, UniProt-unit aggregation), ranked by the Overall score over the 71 configurations. Each configuration contributes a Raw (×) and a RH (✓) entry. Readout is the native scoring interface (mask = masked marginal, wt = wild-type field, tf = teacher forcing). Categories are the five official selection types; the best value in each column is in red.

Rank	Model	Readout	RH	Overall	Activity	Binding	Expression	Organismal	Stability
1	**VenusREM2**	wt	✓	**0.556**	**0.541**	**0.495**	**0.557**	**0.494**	**0.691**
2	ProSST (K=4096)	wt	✓	0.541	0.516	0.494	0.543	0.475	0.674
3	ProSST (K=2048)	wt	✓	0.534	0.528	0.460	0.541	0.469	0.673
4	ProSST (K=1024)	wt	✓	0.532	0.513	0.467	0.533	0.476	0.670
5	ProSST (K=128)	wt	✓	0.525	0.508	0.472	0.517	0.463	0.663
6	VenusREM2	wt	×	0.524	0.480	0.479	0.538	0.444	0.681
7	ProSST (K=512)	wt	✓	0.524	0.507	0.464	0.520	0.464	0.665
8	ProSST (K=20)	wt	✓	0.510	0.493	0.452	0.507	0.446	0.650
9	ProSST (K=2048)	wt	×	0.507	0.476	0.445	0.530	0.431	0.653
10	ProSST (K=4096)	wt	×	0.498	0.444	0.472	0.507	0.416	0.652
11	S3F (wt)	wt	✓	0.489	0.504	0.391	0.479	0.440	0.631
12	S3F (mask)	mask	✓	0.488	0.496	0.407	0.472	0.443	0.619
13	SaProt (650M AF2, mask)	mask	✓	0.487	0.489	0.408	0.489	0.437	0.612
14	SaProt (650M_PDB, mask)	mask	✓	0.486	0.490	0.392	0.492	0.443	0.615
15	ProSST (K=1024)	wt	×	0.485	0.433	0.436	0.499	0.414	0.642
16	ESM3	mask	✓	0.483	0.484	0.406	0.466	0.427	0.631
17	ProtSSN	wt	✓	0.476	0.489	0.393	0.478	0.416	0.605
18	ESM-IF1	tf	✓	0.474	0.447	0.429	0.459	0.401	0.635
19	ProSST (K=512)	wt	×	0.471	0.423	0.431	0.479	0.394	0.629
20	ProtSSN (k=20, h=1280)	wt	✓	0.470	0.482	0.390	0.469	0.406	0.603
21	ProtSSN (k=20, h=512)	wt	✓	0.470	0.484	0.379	0.472	0.413	0.603
22	ESM-2 650M (mask)	mask	✓	0.470	0.486	0.404	0.446	0.432	0.582

Continued on next page

Table 32 continued

Rank	Model	Readout	RH	Overall	Activity	Binding	Expression	Organismal	Stability
23	ProtSSN (k=20, h=768)	wt	✓	0.470	0.485	0.377	0.473	0.409	0.603
24	ProSST (K=128)	wt	×	0.469	0.417	0.440	0.472	0.387	0.628
25	ProtSSN (k=30, h=512)	wt	✓	0.469	0.482	0.381	0.478	0.408	0.597
26	ESM-2 650M (wt)	wt	✓	0.468	0.489	0.384	0.462	0.411	0.592
27	ProtSSN (k=30, h=768)	wt	✓	0.467	0.481	0.386	0.463	0.407	0.597
28	ProtSSN (k=30, h=1280)	wt	✓	0.466	0.479	0.385	0.460	0.406	0.597
29	S3F (mask)	mask	×	0.466	0.470	0.379	0.470	0.409	0.601
30	ProtSSN (k=10, h=1280)	wt	✓	0.465	0.473	0.383	0.471	0.405	0.592
31	SaProt (35M_AF2, mask)	mask	✓	0.465	0.450	0.384	0.464	0.404	0.622
32	ESMC-600M	mask	✓	0.464	0.480	0.371	0.444	0.439	0.587
33	SaProt (650M_PDB, mask)	mask	×	0.463	0.464	0.370	0.488	0.405	0.590
34	ESM-1v (mask)	mask	✓	0.462	0.482	0.364	0.453	0.440	0.572
35	ESM-2 3B (mask)	mask	✓	0.462	0.483	0.377	0.441	0.437	0.571
36	ProtSSN (k=10, h=512)	wt	✓	0.462	0.479	0.379	0.457	0.402	0.592
37	ESMC-300M	mask	✓	0.460	0.479	0.370	0.433	0.433	0.584
38	S3F (wt)	wt	×	0.460	0.469	0.365	0.453	0.406	0.605
39	MIF-ST	tf	✓	0.458	0.440	0.369	0.465	0.404	0.611
40	SaProt (650M AF2, mask)	mask	×	0.457	0.459	0.384	0.485	0.371	0.588
41	ProtSSN (k=10, h=768)	wt	✓	0.457	0.470	0.368	0.459	0.398	0.589
42	ESM-1v (wt)	wt	✓	0.457	0.480	0.342	0.466	0.425	0.569
43	ProteinMPNN (v_48_020)	tf	✓	0.456	0.415	0.394	0.455	0.369	0.648
44	ESM-1b (mask)	mask	✓	0.455	0.473	0.366	0.448	0.429	0.560
45	ESM-2 150M (wt)	wt	✓	0.455	0.472	0.370	0.454	0.385	0.594
46	ESM-2 3B (wt)	wt	✓	0.455	0.479	0.355	0.455	0.414	0.570
47	SaProt (650M AF2, wt)	wt	✓	0.454	0.464	0.366	0.479	0.359	0.600
48	ProteinMPNN (v_48_030)	tf	✓	0.453	0.416	0.384	0.453	0.367	0.645
49	ESM-2 150M (mask)	mask	✓	0.452	0.466	0.367	0.432	0.409	0.583
50	CARP-640M	wt	✓	0.451	0.465	0.364	0.438	0.428	0.563
51	ProtSSN	wt	×	0.451	0.467	0.371	0.452	0.398	0.570
52	ProteinMPNN (v_48_010)	tf	✓	0.451	0.414	0.387	0.453	0.366	0.635
53	SaProt (650M_PDB, wt)	wt	✓	0.450	0.451	0.346	0.474	0.377	0.604
54	ESM-1b (wt)	wt	✓	0.447	0.466	0.347	0.452	0.409	0.562
55	ProtSSN (k=20, h=1280)	wt	×	0.444	0.459	0.368	0.439	0.385	0.568
56	ProtSSN (k=20, h=512)	wt	×	0.443	0.458	0.357	0.440	0.392	0.567
57	ProteinMPNN (v_48_002)	tf	✓	0.443	0.414	0.370	0.441	0.359	0.630
58	ESM3	mask	×	0.442	0.441	0.381	0.449	0.353	0.586
59	ProtSSN (k=20, h=768)	wt	×	0.442	0.461	0.353	0.443	0.387	0.567
60	ProtSSN (k=30, h=768)	wt	×	0.441	0.459	0.363	0.435	0.386	0.562
61	ProtSSN (k=30, h=512)	wt	×	0.441	0.455	0.355	0.449	0.385	0.560
62	ProtSSN (k=30, h=1280)	wt	×	0.439	0.456	0.361	0.433	0.386	0.561
63	ProSST (K=20)	wt	×	0.438	0.392	0.410	0.447	0.354	0.589
64	CARP-76M	wt	✓	0.436	0.445	0.350	0.435	0.378	0.570
65	ProteinMPNN-Soluble (v_48_010)	tf	✓	0.436	0.399	0.383	0.394	0.367	0.634
66	ProteinMPNN-Soluble (v_48_020)	tf	✓	0.435	0.398	0.371	0.399	0.369	0.640
67	ProtSSN (k=10, h=1280)	wt	×	0.435	0.447	0.356	0.439	0.382	0.552
68	ProteinMPNN-Soluble (v_48_030)	tf	✓	0.433	0.396	0.373	0.386	0.369	0.642
69	ProtSSN (k=10, h=512)	wt	×	0.432	0.453	0.352	0.421	0.377	0.558
70	ProteinMPNN-Soluble (v_48_002)	tf	✓	0.430	0.397	0.379	0.388	0.361	0.626
71	MIF-ST	tf	×	0.430	0.410	0.341	0.441	0.381	0.576
72	ESM-2 35M (mask)	mask	✓	0.430	0.437	0.344	0.410	0.379	0.578
73	SaProt (650M_PDB, wt)	wt	×	0.428	0.429	0.328	0.459	0.355	0.568
74	ESM-2 35M (wt)	wt	✓	0.426	0.434	0.349	0.423	0.347	0.576
75	SaProt (35M_AF2, wt)	wt	✓	0.425	0.403	0.358	0.456	0.312	0.597
76	SaProt (650M AF2, wt)	wt	×	0.424	0.437	0.343	0.459	0.317	0.564
77	ProtSSN (k=10, h=768)	wt	×	0.424	0.441	0.339	0.418	0.370	0.550
78	ProGen2-M	tf	✓	0.423	0.451	0.315	0.427	0.408	0.514

Continued on next page

Table 32 continued

Rank	Model	Readout	RH	Overall	Activity	Binding	Expression	Organismal	Stability
79	ProGen2-B	tf	✓	0.422	0.454	0.303	0.431	0.407	0.514
80	ESM-IF1	tf	×	0.421	0.357	0.382	0.435	0.313	0.617
81	ProGen2-XL	tf	✓	0.419	0.446	0.317	0.415	0.407	0.508
82	ESM-2 650M (wt)	wt	×	0.418	0.454	0.325	0.412	0.356	0.541
83	ProGen2-L	tf	✓	0.416	0.452	0.283	0.426	0.403	0.514
84	ESM-2 650M (mask)	mask	×	0.415	0.431	0.337	0.415	0.368	0.523
85	CARP-38M	wt	✓	0.415	0.422	0.346	0.406	0.351	0.547
86	ProGen3-339M	tf	✓	0.414	0.431	0.340	0.388	0.398	0.514
87	ProGen3-762M	tf	✓	0.413	0.441	0.310	0.391	0.409	0.516
88	ProGen3-3B	tf	✓	0.413	0.437	0.292	0.416	0.400	0.521
89	ESM-1v (wt)	wt	×	0.410	0.445	0.293	0.430	0.381	0.500
90	ESMC-300M	mask	×	0.409	0.429	0.320	0.406	0.367	0.523
91	ProGen3-219M	tf	✓	0.409	0.422	0.347	0.377	0.391	0.508
92	SaProt (35M_AF2, mask)	mask	×	0.408	0.369	0.360	0.441	0.294	0.576
93	RITA-L	tf	✓	0.407	0.432	0.308	0.409	0.400	0.486
94	ESM-2 3B (mask)	mask	×	0.407	0.423	0.321	0.403	0.378	0.509
95	ESMC-600M	mask	×	0.407	0.427	0.295	0.413	0.372	0.528
96	ESM-1v (mask)	mask	×	0.407	0.421	0.320	0.429	0.386	0.477
97	ESM-2 3B (wt)	wt	×	0.406	0.440	0.300	0.404	0.366	0.521
98	ProGen2-S	tf	✓	0.406	0.425	0.328	0.381	0.385	0.512
99	ProGen3-1B	tf	✓	0.406	0.436	0.274	0.395	0.409	0.514
100	RITA-M	tf	✓	0.405	0.432	0.304	0.395	0.397	0.495
101	RITA-XL	tf	✓	0.401	0.423	0.298	0.418	0.386	0.482
102	ProGen3-112M	tf	✓	0.401	0.413	0.340	0.365	0.379	0.509
103	RITA-S	tf	✓	0.398	0.415	0.336	0.370	0.379	0.491
104	ESM-2 8M (mask)	mask	✓	0.395	0.389	0.322	0.380	0.347	0.538
105	ESM-1b (wt)	wt	×	0.394	0.428	0.287	0.406	0.350	0.500
106	ESM-2 150M (wt)	wt	×	0.393	0.422	0.316	0.400	0.297	0.528
107	CARP-640M	wt	×	0.390	0.403	0.289	0.401	0.372	0.486
108	ProteinMPNN (v_48_020)	tf	×	0.389	0.316	0.332	0.411	0.287	0.601
109	ESM-1b (mask)	mask	×	0.389	0.406	0.297	0.409	0.354	0.480
110	ESM-2 150M (mask)	mask	×	0.388	0.400	0.326	0.402	0.304	0.510
111	ESM-2 8M (wt)	wt	✓	0.387	0.376	0.323	0.401	0.307	0.531
112	ProteinMPNN (v_48_010)	tf	×	0.385	0.316	0.329	0.410	0.284	0.584
113	ProteinMPNN (v_48_030)	tf	×	0.381	0.309	0.330	0.404	0.278	0.584
114	ProteinMPNN (v_48_002)	tf	×	0.378	0.318	0.312	0.396	0.276	0.587
115	SaProt (35M_AF2, wt)	wt	×	0.366	0.334	0.320	0.398	0.231	0.545
116	ProteinMPNN-Soluble (v_48_010)	tf	×	0.363	0.295	0.317	0.336	0.283	0.587
117	ProteinMPNN-Soluble (v_48_020)	tf	×	0.361	0.288	0.309	0.339	0.286	0.585
118	CARP-76M	wt	×	0.360	0.378	0.290	0.372	0.273	0.487
119	ProteinMPNN-Soluble (v_48_002)	tf	×	0.357	0.291	0.308	0.333	0.275	0.577
120	CARP-600K	wt	✓	0.354	0.358	0.251	0.378	0.279	0.507
121	ProteinMPNN-Soluble (v_48_030)	tf	×	0.353	0.278	0.312	0.316	0.278	0.580
122	ProGen2-XL	tf	×	0.351	0.374	0.269	0.351	0.371	0.387
123	ProGen2-M	tf	×	0.337	0.353	0.254	0.376	0.364	0.339
124	ProGen3-3B	tf	×	0.331	0.365	0.226	0.363	0.360	0.340
125	ProGen2-B	tf	×	0.330	0.355	0.242	0.381	0.351	0.320
126	ESM-2 35M (wt)	wt	×	0.329	0.333	0.282	0.335	0.208	0.487
127	ProGen2-L	tf	×	0.327	0.368	0.216	0.373	0.355	0.323
128	ESM-2 35M (mask)	mask	×	0.321	0.316	0.291	0.343	0.217	0.439
129	ProGen3-1B	tf	×	0.320	0.354	0.218	0.347	0.365	0.318
130	CARP-38M	wt	×	0.314	0.319	0.282	0.306	0.220	0.444
131	RITA-L	tf	×	0.309	0.321	0.252	0.351	0.352	0.267
132	RITA-XL	tf	×	0.305	0.316	0.233	0.360	0.340	0.277
133	ProGen3-762M	tf	×	0.304	0.328	0.244	0.330	0.346	0.274
134	RITA-M	tf	×	0.297	0.308	0.251	0.336	0.348	0.241

Continued on next page

Table 32 continued

Rank	Model	Readout	RH	Overall	Activity	Binding	Expression	Organismal	Stability
135	ProGen2-S	tf	×	0.288	0.300	0.272	0.309	0.291	0.269
136	ProGen3-339M	tf	×	0.286	0.304	0.280	0.311	0.319	0.217
137	ProGen3-219M	tf	×	0.264	0.274	0.279	0.288	0.296	0.185
138	RITA-S	tf	×	0.254	0.248	0.267	0.277	0.289	0.190
139	ESM-2 8M (wt)	wt	×	0.251	0.217	0.244	0.264	0.131	0.399
140	ProGen3-112M	tf	×	0.232	0.244	0.260	0.253	0.265	0.139
141	ESM-2 8M (mask)	mask	×	0.223	0.191	0.260	0.266	0.139	0.262
142	CARP-600K	wt	×	0.159	0.142	0.090	0.167	0.065	0.331

Table 33 retains the VenusMutHub task breakdown. Its within-benchmark ordering shows which configurations support each task beyond the assay-macro average.

Table 33: Unified Raw and RH leaderboard on VenusMutHub (assay-macro Spearman), ranked by the Overall score over the 71 configurations. Each configuration contributes a Raw (×) and a RH (✓) entry. Readout is the native scoring interface (mask / wt / tf). Categories are the five assay tasks; the best value in each column is in red.

Rank	Model	Readout	RH	Overall	stability	activity	PPI bind.	selectivity	DTI bind.
1	ProteinMPNN (v_48_020)	tf	✓	**0.271**	**0.399**	0.108	0.024	0.012	0.198
2	ESM-IF1	tf	✓	0.267	0.356	0.152	0.108	0.018	0.269
3	ProteinMPNN-Soluble (v_48_020)	tf	✓	0.265	0.396	0.106	0.036	0.000	0.093
4	ProteinMPNN-Soluble (v_48_030)	tf	✓	0.265	0.383	0.120	0.039	0.022	0.168
5	ProteinMPNN (v_48_030)	tf	✓	0.259	0.388	0.106	0.019	-0.020	0.137
6	**VenusREM2**	wt	✓	0.258	0.351	0.122	0.113	-0.017	0.274
7	ProteinMPNN-Soluble (v_48_010)	tf	✓	0.256	0.383	0.104	0.023	-0.025	0.133
8	ProteinMPNN (v_48_010)	tf	✓	0.256	0.383	0.095	0.038	-0.023	0.122
9	ProteinMPNN-Soluble (v_48_002)	tf	✓	0.250	0.379	0.075	0.033	-0.016	0.152
10	ProSST (K=2048)	wt	✓	0.249	0.348	0.118	0.103	-0.048	0.204
11	ProteinMPNN (v_48_002)	tf	✓	0.244	0.370	0.069	0.021	0.020	0.134
12	ProSST (K=1024)	wt	✓	0.242	0.334	0.120	0.083	-0.014	0.236
13	SaProt (650M_PDB, mask)	mask	✓	0.240	0.319	0.139	0.090	-0.014	0.282
14	SaProt (650M AF2, mask)	mask	✓	0.240	0.316	0.161	0.091	0.004	0.200
15	ProSST (K=20)	wt	✓	0.238	0.327	0.083	0.122	-0.036	0.312
16	ProSST (K=4096)	wt	✓	0.237	0.324	0.117	0.096	**0.031**	0.177
17	SaProt (35M_AF2, mask)	mask	✓	0.237	0.312	0.171	0.068	-0.040	0.256
18	S3F (mask)	mask	✓	0.234	0.316	0.140	0.095	-0.030	0.195
19	ProSST (K=128)	wt	✓	0.230	0.305	0.117	0.121	-0.015	0.261
20	MIF-ST	tf	✓	0.230	0.352	0.044	0.079	-0.106	0.169
21	ESM3	mask	✓	0.226	0.302	0.150	0.106	-0.034	0.142
22	ProSST (K=512)	wt	✓	0.226	0.313	0.102	0.083	-0.037	0.264
23	ESM-2 150M (mask)	mask	✓	0.225	0.291	0.173	0.079	-0.002	0.206
24	S3F (wt)	wt	✓	0.224	0.304	0.111	0.123	-0.022	0.175
25	ESM-2 3B (mask)	mask	✓	0.223	0.277	**0.180**	0.113	-0.081	0.295
26	ProteinMPNN-Soluble (v_48_030)	tf	×	0.222	0.357	0.026	-0.030	0.030	0.131
27	ProteinMPNN (v_48_020)	tf	×	0.221	0.366	0.006	-0.033	-0.001	0.116
28	ProteinMPNN (v_48_010)	tf	×	0.218	0.360	0.013	-0.034	-0.050	0.155
29	ESM-2 150M (wt)	wt	✓	0.216	0.288	0.127	0.110	-0.015	0.186
30	ESM-IF1	tf	×	0.216	0.328	0.074	0.040	0.018	0.021
31	ProteinMPNN-Soluble (v_48_020)	tf	×	0.216	0.367	0.004	-0.049	0.000	0.031
32	SaProt (650M_PDB, wt)	wt	✓	0.215	0.304	0.059	0.078	-0.034	0.319
33	ESM-2 650M (mask)	mask	✓	0.213	0.280	0.149	0.100	-0.063	0.197
34	ProSST (K=2048)	wt	×	0.213	0.315	0.076	0.064	-0.065	0.137
35	ESM-2 35M (mask)	mask	✓	0.212	0.280	0.158	0.049	0.003	0.185
36	ProteinMPNN-Soluble (v_48_010)	tf	×	0.212	0.357	-0.004	-0.042	-0.070	0.167

Continued on next page

Table 33 continued

Rank	Model	Readout	RH	Overall	stability	activity	PPI bind.	selectivity	DTI bind.
37	S3F (mask)	mask	×	0.211	0.295	0.109	0.043	-0.007	0.203
38	ESM-1b (mask)	mask	✓	0.211	0.265	0.163	0.126	-0.048	0.211
39	ProteinMPNN (v_48_030)	tf	×	0.210	0.357	0.009	-0.060	-0.022	0.067
40	SaProt (650M AF2, mask)	mask	×	0.207	0.275	0.152	0.058	0.014	0.143
41	ESM-1v (mask)	mask	✓	0.206	0.268	0.165	0.081	-0.082	0.201
42	CARP-640M	wt	✓	0.206	0.271	0.151	0.076	-0.094	0.236
43	SaProt (650M_PDB, mask)	mask	×	0.205	0.275	0.116	0.048	-0.013	0.285
44	SaProt (650M AF2, wt)	wt	✓	0.202	0.286	0.093	0.057	-0.035	0.175
45	ProteinMPNN-Soluble (v_48_002)	tf	×	0.198	0.340	-0.021	-0.039	-0.054	0.135
46	ESMC-300M	mask	✓	0.198	0.254	0.151	0.106	-0.059	0.181
47	ESM-2 8M (mask)	mask	✓	0.198	0.256	0.144	0.067	-0.045	0.248
48	ProtSSN (k=20, h=1280)	wt	✓	0.198	0.282	0.072	0.083	-0.034	0.162
49	ProtSSN (k=10, h=512)	wt	✓	0.197	0.286	0.064	0.079	-0.050	0.164
50	MIF-ST	tf	×	0.196	0.317	0.017	0.029	-0.110	0.135
51	ProtSSN	wt	✓	0.195	0.281	0.079	0.084	-0.050	0.110
52	CARP-76M	wt	✓	0.194	0.259	0.119	0.083	-0.034	0.196
53	ProGen2-B	tf	✓	0.194	0.252	0.142	0.100	-0.089	0.195
54	ESMC-600M	mask	✓	0.194	0.248	0.142	0.108	-0.069	0.210
55	ESM-2 3B (wt)	wt	✓	0.194	0.262	0.100	0.082	-0.060	0.242
56	ESM-1b (wt)	wt	✓	0.193	0.255	0.098	**0.131**	-0.069	0.246
57	RITA-L	tf	✓	0.193	0.250	0.138	0.101	-0.056	0.188
58	S3F (wt)	wt	×	0.193	0.271	0.082	0.068	-0.024	0.182
59	ProGen2-S	tf	✓	0.192	0.246	0.153	0.086	-0.070	0.209
60	ProtSSN (k=20, h=512)	wt	✓	0.192	0.270	0.078	0.095	-0.046	0.156
61	ProtSSN (k=10, h=1280)	wt	✓	0.192	0.279	0.072	0.090	-0.044	0.070
62	ProtSSN (k=30, h=512)	wt	✓	0.191	0.276	0.073	0.077	-0.055	0.142
63	ProteinMPNN (v_48_002)	tf	×	0.191	0.328	-0.021	-0.032	0.001	0.064
64	ProtSSN (k=20, h=768)	wt	✓	0.191	0.271	0.078	0.098	-0.044	0.110
65	ProGen3-219M	tf	✓	0.190	0.242	0.155	0.101	-0.059	0.157
66	VenusREM2	wt	×	0.190	0.294	0.018	0.006	-0.049	0.267
67	ProGen3-339M	tf	✓	0.189	0.242	0.162	0.081	-0.071	0.169
68	ProGen2-L	tf	✓	0.188	0.243	0.146	0.084	-0.088	0.224
69	ESM-1v (wt)	wt	✓	0.188	0.256	0.112	0.097	-0.068	0.133
70	SaProt (35M_AF2, mask)	mask	×	0.188	0.258	0.153	-0.017	-0.047	0.180
71	ESM-2 650M (wt)	wt	✓	0.187	0.259	0.089	0.081	-0.056	0.188
72	RITA-M	tf	✓	0.186	0.239	0.138	0.115	-0.101	0.203
73	ESM-2 35M (wt)	wt	✓	0.186	0.265	0.092	0.055	-0.024	0.120
74	ESM3	mask	×	0.185	0.264	0.128	-0.002	-0.032	0.106
75	ProtSSN (k=30, h=768)	wt	✓	0.185	0.262	0.077	0.092	-0.042	0.113
76	ProtSSN (k=30, h=1280)	wt	✓	0.183	0.262	0.055	0.095	-0.042	0.167
77	ESM-2 3B (mask)	mask	×	0.182	0.228	0.161	0.058	-0.082	0.274
78	SaProt (650M_PDB, wt)	wt	×	0.182	0.256	0.050	0.033	-0.025	0.363
79	RITA-S	tf	✓	0.182	0.228	0.134	0.121	-0.041	0.177
80	ProGen2-M	tf	✓	0.182	0.240	0.130	0.094	-0.117	0.187
81	ProGen3-762M	tf	✓	0.180	0.228	0.159	0.099	-0.077	0.122
82	SaProt (35M_AF2, wt)	wt	✓	0.179	0.250	0.098	0.033	-0.023	0.171
83	ProtSSN (k=10, h=768)	wt	✓	0.178	0.258	0.060	0.077	-0.041	0.118
84	SaProt (650M AF2, wt)	wt	×	0.178	0.249	0.092	0.031	-0.023	0.190
85	ProGen3-112M	tf	✓	0.177	0.219	0.130	0.121	-0.038	0.203
86	ESM-2 8M (wt)	wt	✓	0.176	0.239	0.079	0.047	-0.052	0.318
87	ProGen2-XL	tf	✓	0.176	0.232	0.136	0.095	-0.111	0.131
88	ProtSSN	wt	×	0.174	0.251	0.079	0.056	-0.048	0.111
89	ProtSSN (k=20, h=1280)	wt	×	0.174	0.247	0.073	0.067	-0.035	0.138
90	ProGen3-1B	tf	✓	0.173	0.220	0.127	0.121	-0.096	0.175
91	ProGen3-3B	tf	✓	0.170	0.228	0.110	0.113	-0.115	0.126
92	ESM-2 650M (mask)	mask	×	0.169	0.224	0.134	0.019	-0.054	0.220

Continued on next page

Table 33 continued

Rank	Model	Readout	RH	Overall	stability	activity	PPI bind.	selectivity	DTI bind.
93	ProtSSN (k=30, h=512)	wt	×	0.169	0.244	0.076	0.030	-0.057	0.182
94	ProtSSN (k=10, h=512)	wt	×	0.169	0.247	0.061	0.044	-0.053	0.168
95	ProtSSN (k=20, h=512)	wt	×	0.168	0.236	0.076	0.064	-0.043	0.171
96	ESM-2 150M (mask)	mask	×	0.167	0.221	0.144	-0.004	-0.033	0.190
97	ProtSSN (k=10, h=1280)	wt	×	0.166	0.245	0.073	0.051	-0.047	0.062
98	ESM-1b (mask)	mask	×	0.166	0.214	0.129	0.067	-0.048	0.174
99	ProtSSN (k=20, h=768)	wt	×	0.165	0.233	0.079	0.077	-0.046	0.109
100	ESM-2 150M (wt)	wt	×	0.165	0.223	0.120	-0.004	-0.029	0.225
101	ProtSSN (k=30, h=1280)	wt	×	0.162	0.232	0.056	0.061	-0.044	0.169
102	ProSST (K=1024)	wt	×	0.162	0.265	-0.016	0.004	-0.043	0.184
103	ESM-1v (mask)	mask	×	0.162	0.218	0.141	0.003	-0.093	0.180
104	ProSST (K=4096)	wt	×	0.161	0.250	0.013	0.004	0.029	0.155
105	ESM-2 3B (wt)	wt	×	0.161	0.218	0.092	0.030	-0.058	0.269
106	RITA-XL	tf	✓	0.160	0.216	0.112	0.069	-0.100	0.162
107	ProtSSN (k=30, h=768)	wt	×	0.160	0.230	0.075	0.056	-0.050	0.097
108	CARP-38M	wt	✓	0.158	0.216	0.075	0.065	-0.042	0.205
109	CARP-640M	wt	×	0.157	0.221	0.116	0.005	-0.119	0.187
110	ESMC-600M	mask	×	0.155	0.205	0.114	0.060	-0.074	0.164
111	ESM-1v (wt)	wt	×	0.153	0.212	0.109	0.015	-0.075	0.172
112	ESM-1b (wt)	wt	×	0.153	0.201	0.092	0.072	-0.057	0.209
113	ESM-2 650M (wt)	wt	×	0.152	0.210	0.087	0.015	-0.060	0.241
114	ProtSSN (k=10, h=768)	wt	×	0.152	0.221	0.052	0.053	-0.037	0.132
115	CARP-600K	wt	✓	0.151	0.212	0.040	0.074	-0.071	0.260
116	ESMC-300M	mask	×	0.144	0.195	0.123	0.011	-0.073	0.132
117	ProSST (K=20)	wt	×	0.143	0.230	-0.031	0.013	-0.087	0.300
118	SaProt (35M_AF2, wt)	wt	×	0.142	0.201	0.086	-0.031	-0.037	0.234
119	ProSST (K=128)	wt	×	0.139	0.225	-0.020	0.009	-0.044	0.204
120	CARP-76M	wt	×	0.131	0.182	0.092	-0.013	-0.033	0.168
121	ProGen2-B	tf	×	0.124	0.172	0.111	0.004	-0.109	0.114
122	ProSST (K=512)	wt	×	0.124	0.219	-0.045	-0.017	-0.089	0.180
123	ProGen2-XL	tf	×	0.123	0.170	0.105	0.030	-0.132	0.113
124	ESM-2 35M (mask)	mask	×	0.122	0.179	0.106	-0.095	0.025	0.078
125	ProGen2-L	tf	×	0.120	0.157	0.114	-0.011	-0.108	0.227
126	ProGen2-M	tf	×	0.115	0.168	0.086	-0.004	-0.129	0.104
127	ProGen3-762M	tf	×	0.111	0.132	0.178	-0.020	-0.087	0.090
128	RITA-L	tf	×	0.110	0.158	0.091	-0.018	-0.066	0.080
129	ProGen3-339M	tf	×	0.110	0.133	0.173	-0.051	-0.077	0.146
130	ProGen3-3B	tf	×	0.110	0.160	0.082	0.039	-0.123	0.009
131	ESM-2 35M (wt)	wt	×	0.110	0.169	0.065	-0.092	-0.037	0.175
132	ProGen3-1B	tf	×	0.108	0.151	0.112	0.013	-0.099	0.001
133	RITA-M	tf	×	0.100	0.137	0.081	-0.021	-0.068	0.184
134	RITA-XL	tf	×	0.095	0.132	0.081	-0.011	-0.100	0.142
135	ProGen3-219M	tf	×	0.092	0.122	0.111	-0.071	-0.070	0.192
136	ProGen2-S	tf	×	0.091	0.137	0.067	-0.053	-0.066	0.127
137	ESM-2 8M (mask)	mask	×	0.079	0.110	0.055	-0.085	-0.057	0.317
138	CARP-38M	wt	×	0.073	0.108	0.020	-0.042	-0.054	0.245
139	ProGen3-112M	tf	×	0.072	0.086	0.105	-0.048	-0.044	0.156
140	ESM-2 8M (wt)	wt	×	0.065	0.102	0.017	-0.115	-0.076	**0.368**
141	RITA-S	tf	×	0.056	0.068	0.083	-0.038	-0.064	0.143
142	CARP-600K	wt	×	0.021	0.057	-0.048	-0.104	-0.103	0.272

Table 34 retains the viral phenotype breakdown under hierarchical aggregation. Low immune-escape correlations remain visible alongside stronger intrinsic-fitness and expression scores.

Table 34: Unified Raw and RH leaderboard on ViroHub (mutant-only hierarchical Spearman: means within phenotype–backbone cells, then across the seven phenotypes), ranked by the Overall score over the 71 configurations. Each configuration contributes a Raw (×) and a RH (✓) entry. Readout is the native scoring interface (mask / wt / tf); the best value in each column is in red.

Rank	Model	Readout	RH	Overall	fitness	activity	expression	cell entry	binding	stability	immune esc.
1	SaProt (650M_PDB, mask)	mask	✓	**0.320**	0.394	0.577	0.569	0.484	0.145	0.045	0.026
2	S3F (wt)	wt	✓	0.318	0.404	0.582	0.540	**0.493**	0.111	0.052	**0.046**
3	ESM-IF1	tf	✓	0.318	0.359	0.603	0.566	0.439	**0.173**	0.060	0.026
4	MIF-ST	tf	✓	0.317	0.391	0.542	0.553	0.476	0.144	0.068	0.042
5	ProSST (K=2048)	wt	✓	0.310	0.366	0.570	0.561	0.469	0.126	0.048	0.032
6	SaProt (650M_PDB, wt)	wt	✓	0.307	0.343	**0.608**	0.564	0.432	0.131	0.052	0.022
7	ProteinMPNN (v_48_010)	tf	✓	0.306	0.352	0.544	0.565	0.441	0.143	0.066	0.033
8	SaProt (650M_PDB, mask)	mask	×	0.306	0.360	0.606	0.545	0.445	0.130	0.045	0.015
9	ProteinMPNN (v_48_020)	tf	✓	0.306	0.357	0.536	0.567	0.444	0.144	0.060	0.032
10	ProteinMPNN (v_48_002)	tf	✓	0.305	0.346	0.543	0.567	0.438	0.146	0.064	0.034
11	ProteinMPNN-Soluble (v_48_002)	tf	✓	0.305	0.347	0.545	0.560	0.438	0.145	0.066	0.036
12	ProteinMPNN-Soluble (v_48_020)	tf	✓	0.304	0.360	0.528	0.559	0.440	0.145	0.061	0.035
13	ProteinMPNN (v_48_030)	tf	✓	0.304	0.352	0.538	0.570	0.441	0.142	0.049	0.035
14	ProteinMPNN-Soluble (v_48_010)	tf	✓	0.304	0.345	0.541	0.557	0.438	0.146	0.064	0.035
15	S3F (mask)	mask	✓	0.302	**0.414**	0.511	0.509	0.475	0.118	0.053	0.036
16	ProteinMPNN-Soluble (v_48_030)	tf	✓	0.299	0.357	0.514	0.556	0.438	0.137	0.055	0.035
17	VenusREM2	wt	✓	0.297	0.353	0.480	**0.580**	0.462	0.126	0.046	0.033
18	ProtSSN	wt	✓	0.294	0.357	0.565	0.502	0.435	0.117	0.044	0.036
19	ProtSSN (k=20, h=512)	wt	✓	0.291	0.360	0.545	0.502	0.435	0.119	0.042	0.033
20	ProtSSN (k=20, h=1280)	wt	✓	0.289	0.352	0.547	0.499	0.426	0.114	0.047	0.034
21	MIF-ST	tf	×	0.288	0.374	0.480	0.485	0.448	0.114	0.074	0.044
22	ProtSSN (k=30, h=512)	wt	✓	0.288	0.357	0.560	0.484	0.430	0.110	0.037	0.037
23	ProtSSN (k=20, h=768)	wt	✓	0.287	0.352	0.551	0.485	0.428	0.114	0.046	0.034
24	ProtSSN (k=30, h=768)	wt	✓	0.286	0.357	0.547	0.487	0.421	0.111	0.044	0.033
25	ESM-IF1	tf	×	0.285	0.248	0.606	0.542	0.343	0.169	0.065	0.024
26	ProtSSN (k=10, h=1280)	wt	✓	0.284	0.352	0.544	0.466	0.426	0.113	0.052	0.035
27	SaProt (650M AF2, mask)	mask	✓	0.281	0.360	0.501	0.495	0.426	0.118	0.035	0.032
28	S3F (wt)	wt	×	0.279	0.377	0.549	0.451	0.433	0.055	0.047	0.044
29	ProtSSN (k=10, h=768)	wt	✓	0.279	0.347	0.522	0.472	0.413	0.118	0.044	0.037
30	ProtSSN (k=30, h=1280)	wt	✓	0.279	0.346	0.533	0.474	0.417	0.107	0.036	0.037
31	ProSST (K=1024)	wt	✓	0.275	0.329	0.436	0.538	0.425	0.116	0.056	0.028
32	SaProt (650M_PDB, wt)	wt	×	0.273	0.312	0.546	0.492	0.395	0.098	0.057	0.013
33	ProtSSN (k=10, h=512)	wt	✓	0.272	0.333	0.523	0.447	0.410	0.106	0.049	0.034
34	ProSST (K=4096)	wt	✓	0.271	0.315	0.405	0.548	0.429	0.121	0.045	0.035
35	ProSST (K=2048)	wt	×	0.271	0.307	0.550	0.497	0.407	0.080	0.026	0.030
36	ProteinMPNN (v_48_010)	tf	×	0.271	0.250	0.538	0.501	0.383	0.121	0.067	0.035
37	ProteinMPNN (v_48_020)	tf	×	0.270	0.256	0.526	0.507	0.386	0.123	0.059	0.033
38	ProteinMPNN (v_48_002)	tf	×	0.269	0.241	0.533	0.508	0.377	0.126	0.065	0.036
39	S3F (mask)	mask	×	0.269	0.373	0.477	0.442	0.433	0.077	0.049	0.036
40	ProSST (K=512)	wt	✓	0.269	0.324	0.418	0.543	0.418	0.111	0.035	0.033
41	ProteinMPNN-Soluble (v_48_002)	tf	×	0.269	0.238	0.532	0.502	0.377	0.122	0.069	0.040
42	ProteinMPNN-Soluble (v_48_010)	tf	×	0.268	0.241	0.533	0.499	0.377	0.122	0.066	0.037
43	ProteinMPNN (v_48_030)	tf	×	0.266	0.241	0.529	0.515	0.374	0.119	0.045	0.038
44	ProteinMPNN-Soluble (v_48_020)	tf	×	0.265	0.249	0.515	0.494	0.376	0.120	0.060	0.038
45	ProSST (K=128)	wt	✓	0.264	0.326	0.395	0.519	0.415	0.120	0.048	0.028

Continued on next page

Table 34 continued

Rank	Model	Readout	RH	Overall	fitness	activity	expression	cell entry	binding	stability	immune esc.
46	ESM-2 3B (wt)	wt	✓	0.260	0.362	0.521	0.391	0.396	0.073	0.050	0.027
47	ProSST (K=20)	wt	✓	0.259	0.312	0.413	0.528	0.396	0.119	0.020	0.028
48	SaProt (650M AF2, wt)	wt	✓	0.259	0.268	0.542	0.487	0.333	0.100	0.049	0.031
49	ProteinMPNN-Soluble (v_48_030)	tf	×	0.256	0.237	0.497	0.489	0.366	0.109	0.054	0.037
50	ESM-2 650M (wt)	wt	✓	0.252	0.364	0.457	0.392	0.369	0.095	0.053	0.036
51	ESM3	mask	✓	0.249	0.320	0.462	0.406	0.369	0.099	0.059	0.030
52	ESM-1b (wt)	wt	✓	0.248	0.320	0.445	0.404	0.348	0.098	0.086	0.036
53	RITA-XL	tf	✓	0.247	0.359	0.420	0.406	0.371	0.107	0.058	0.010
54	RITA-L	tf	✓	0.247	0.363	0.449	0.390	0.356	0.113	0.052	0.007
55	SaProt (35M_AF2, mask)	mask	✓	0.245	0.325	0.403	0.424	0.382	0.097	0.049	0.032
56	ESM-2 3B (mask)	mask	✓	0.243	0.393	0.456	0.314	0.403	0.050	0.053	0.036
57	ESM-1v (wt)	wt	✓	0.238	0.287	0.396	0.405	0.363	0.114	0.069	0.029
58	RITA-M	tf	✓	0.237	0.370	0.420	0.374	0.352	0.089	0.058	-0.007
59	VenusREM2	wt	×	0.236	0.242	0.420	0.500	0.351	0.072	0.034	0.032
60	RITA-S	tf	✓	0.236	0.378	0.420	0.354	0.345	0.092	0.063	-0.004
61	ProGen2-XL	tf	✓	0.235	0.350	0.392	0.396	0.367	0.106	0.039	-0.003
62	ESM-2 150M (wt)	wt	✓	0.235	0.291	0.451	0.395	0.308	0.107	0.062	0.032
63	ProtSSN (k=20, h=512)	wt	×	0.233	0.352	0.420	0.361	0.381	0.050	0.037	0.033
64	ProtSSN	wt	×	0.233	0.352	0.434	0.350	0.376	0.042	0.041	0.036
65	ProtSSN (k=20, h=1280)	wt	×	0.230	0.347	0.419	0.357	0.366	0.044	0.044	0.034
66	CARP-76M	wt	✓	0.230	0.265	0.405	0.407	0.316	0.109	0.076	0.029
67	ESMC-600M	mask	✓	0.228	0.318	0.409	0.335	0.365	0.071	0.059	0.041
68	ESM-2 650M (mask)	mask	✓	0.227	0.388	0.376	0.309	0.372	0.058	0.049	0.041
69	ProtSSN (k=20, h=768)	wt	×	0.226	0.345	0.420	0.329	0.367	0.042	0.044	0.034
70	ProtSSN (k=30, h=512)	wt	×	0.225	0.347	0.433	0.325	0.369	0.032	0.033	0.038
71	ProtSSN (k=30, h=768)	wt	×	0.223	0.350	0.413	0.335	0.357	0.037	0.041	0.031
72	ESM-1b (mask)	mask	✓	0.223	0.355	0.362	0.314	0.355	0.063	0.075	0.039
73	SaProt (650M AF2, mask)	mask	×	0.222	0.255	0.477	0.398	0.299	0.075	0.025	0.026
74	CARP-38M	wt	✓	0.222	0.244	0.378	0.404	0.312	0.112	0.073	0.030
75	ESMC-300M	mask	✓	0.220	0.311	0.370	0.339	0.356	0.070	0.054	0.038
76	ProtSSN (k=10, h=1280)	wt	×	0.220	0.344	0.407	0.303	0.363	0.039	0.047	0.036
77	CARP-640M	wt	✓	0.215	0.374	0.286	0.299	0.361	0.065	0.087	0.034
78	ProtSSN (k=10, h=768)	wt	×	0.214	0.335	0.384	0.309	0.345	0.045	0.040	0.038
79	ESM-2 8M (wt)	wt	✓	0.214	0.231	0.371	0.392	0.298	0.114	0.059	0.029
80	ESM-2 35M (wt)	wt	✓	0.213	0.245	0.359	0.390	0.292	0.109	0.069	0.026
81	ProGen2-M	tf	✓	0.212	0.368	0.295	0.349	0.341	0.079	0.044	0.006
82	ProGen3-3B	tf	✓	0.211	0.366	0.270	0.367	0.359	0.078	0.034	0.004
83	CARP-600K	wt	✓	0.210	0.221	0.354	0.395	0.293	0.116	0.069	0.025
84	ProtSSN (k=30, h=1280)	wt	×	0.210	0.336	0.392	0.305	0.345	0.025	0.032	0.036
85	SaProt (35M_AF2, wt)	wt	✓	0.209	0.210	0.421	0.391	0.275	0.059	0.072	0.036
86	ESM-2 150M (mask)	mask	✓	0.209	0.320	0.380	0.303	0.304	0.069	0.054	0.035
87	ProGen3-1B	tf	✓	0.207	0.336	0.280	0.335	0.344	0.094	0.056	0.003
88	ProSST (K=1024)	wt	×	0.204	0.210	0.351	0.436	0.296	0.063	0.049	0.025
89	ESM-1v (mask)	mask	✓	0.202	0.326	0.271	0.289	0.356	0.069	0.074	0.032
90	ProtSSN (k=10, h=512)	wt	×	0.202	0.320	0.376	0.277	0.338	0.028	0.043	0.033
91	ProSST (K=4096)	wt	×	0.200	0.186	0.318	0.460	0.303	0.073	0.029	0.032
92	ProGen3-762M	tf	✓	0.192	0.316	0.249	0.317	0.341	0.063	0.038	0.023
93	ProSST (K=512)	wt	×	0.191	0.194	0.314	0.441	0.281	0.055	0.021	0.032
94	ProGen3-339M	tf	✓	0.191	0.291	0.248	0.315	0.337	0.065	0.067	0.013
95	SaProt (650M AF2, wt)	wt	×	0.188	0.213	0.423	0.349	0.231	0.043	0.033	0.024
96	ESM-2 35M (mask)	mask	✓	0.184	0.274	0.280	0.285	0.288	0.070	0.060	0.032
97	ProGen3-112M	tf	✓	0.182	0.247	0.243	0.303	0.324	0.065	0.064	0.027
98	ProGen3-219M	tf	✓	0.181	0.275	0.242	0.282	0.335	0.063	0.048	0.021
99	ProSST (K=128)	wt	×	0.179	0.196	0.266	0.393	0.272	0.059	0.042	0.025
100	ProGen2-B	tf	✓	0.175	0.371	0.055	0.333	0.343	0.071	0.048	0.006

Continued on next page

Table 34 continued

Rank	Model	Readout	RH	Overall	fitness	activity	expression	cell entry	binding	stability	immune esc.
101	ESM-2 8M (mask)	mask	✓	0.173	0.258	0.240	0.275	0.280	0.074	0.054	0.033
102	ProGen2-L	tf	✓	0.173	0.358	0.090	0.300	0.340	0.076	0.042	0.007
103	RITA-XL	tf	×	0.172	0.327	0.236	0.272	0.270	0.043	0.065	-0.008
104	ProGen2-S	tf	✓	0.171	0.300	0.149	0.301	0.320	0.077	0.050	0.001
105	RITA-L	tf	×	0.166	0.323	0.253	0.241	0.249	0.052	0.060	-0.014
106	ProGen2-XL	tf	×	0.162	0.308	0.240	0.265	0.259	0.052	0.037	-0.029
107	ProSST (K=20)	wt	×	0.161	0.179	0.276	0.373	0.232	0.052	-0.006	0.023
108	RITA-M	tf	×	0.155	0.335	0.229	0.227	0.230	0.020	0.068	-0.026
109	SaProt (35M_AF2, mask)	mask	×	0.149	0.189	0.303	0.246	0.205	0.030	0.042	0.025
110	RITA-S	tf	×	0.148	0.319	0.234	0.206	0.205	0.025	0.072	-0.021
111	ESM-2 3B (mask)	mask	×	0.145	0.330	0.352	0.032	0.290	-0.054	0.048	0.018
112	ESM-2 3B (wt)	wt	×	0.138	0.321	0.342	0.019	0.277	-0.055	0.046	0.017
113	ESM3	mask	×	0.137	0.203	0.338	0.168	0.186	0.013	0.037	0.015
114	ProGen2-M	tf	×	0.117	0.307	0.121	0.191	0.161	0.012	0.038	-0.011
115	ESM-2 650M (wt)	wt	×	0.114	0.321	0.232	-0.004	0.210	-0.040	0.050	0.030
116	ESM-2 650M (mask)	mask	×	0.111	0.326	0.199	-0.001	0.218	-0.040	0.042	0.031
117	ESMC-600M	mask	×	0.108	0.190	0.263	0.077	0.180	-0.015	0.036	0.024
118	SaProt (35M_AF2, wt)	wt	×	0.108	0.122	0.272	0.193	0.117	-0.025	0.048	0.026
119	ProGen3-3B	tf	×	0.101	0.274	0.011	0.191	0.222	-0.006	0.030	-0.013
120	ESMC-300M	mask	×	0.097	0.181	0.220	0.074	0.165	-0.020	0.033	0.023
121	ESM-1b (wt)	wt	×	0.094	0.248	0.190	0.025	0.125	-0.025	0.066	0.027
122	ESM-1b (mask)	mask	×	0.093	0.253	0.190	0.020	0.129	-0.031	0.063	0.029
123	ProGen3-1B	tf	×	0.092	0.225	0.030	0.139	0.187	0.017	0.056	-0.009
124	CARP-640M	wt	×	0.092	0.291	0.066	-0.003	0.174	-0.026	**0.116**	0.024
125	ProGen2-B	tf	×	0.076	0.311	-0.143	0.169	0.162	-0.003	0.049	-0.013
126	ESM-2 150M (mask)	mask	×	0.071	0.193	0.214	-0.007	0.054	-0.024	0.050	0.020
127	ESM-1v (wt)	wt	×	0.069	0.192	0.073	-0.014	0.166	-0.017	0.069	0.016
128	ESM-2 150M (wt)	wt	×	0.069	0.188	0.207	-0.006	0.047	-0.023	0.049	0.019
129	ProGen2-L	tf	×	0.064	0.293	-0.156	0.122	0.156	0.008	0.032	-0.009
130	ESM-1v (mask)	mask	×	0.064	0.193	0.031	-0.023	0.173	-0.020	0.073	0.017
131	ProGen3-762M	tf	×	0.062	0.181	0.010	0.068	0.155	-0.013	0.026	0.008
132	ProGen3-339M	tf	×	0.058	0.146	0.009	0.067	0.138	-0.012	0.063	-0.003
133	CARP-76M	wt	×	0.050	0.129	0.106	0.014	0.057	-0.016	0.049	0.013
134	ProGen2-S	tf	×	0.045	0.132	-0.028	0.080	0.089	0.002	0.045	-0.008
135	ProGen3-112M	tf	×	0.034	0.060	0.001	0.029	0.112	-0.011	0.043	0.006
136	ProGen3-219M	tf	×	0.033	0.102	-0.018	-0.005	0.122	-0.018	0.040	0.005
137	CARP-38M	wt	×	0.027	0.088	0.028	-0.013	0.047	-0.019	0.043	0.013
138	ESM-2 35M (wt)	wt	×	0.026	0.098	0.052	-0.018	0.015	-0.024	0.048	0.012
139	ESM-2 35M (mask)	mask	×	0.026	0.106	0.038	-0.025	0.024	-0.022	0.046	0.014
140	ESM-2 8M (wt)	wt	×	0.017	0.065	0.021	-0.026	0.021	-0.017	0.041	0.012
141	ESM-2 8M (mask)	mask	×	0.015	0.066	0.002	-0.029	0.020	-0.016	0.048	0.012
142	CARP-600K	wt	×	0.006	0.045	-0.032	-0.015	0.010	-0.014	0.043	0.006

I.2 Per-configuration staged component scores

The following tables report staged component scores for all 71 configurations on each benchmark, separating structure-aware, masked sequence, inverse-folding, and autoregressive models. They provide the configuration-level source for the stage-wise conclusions in Appendix F.1. The columns show the entropy-adaptive MSA mix, background calibration without or with the coherence gate (alternative corrections on the mix), and RSA and pLDDT shrinkage stacked on the gated branch; the final column is the full RH readout.

ProteinGym: structural readouts. Tables 35 and 36 trace the inverse-folding and structure-aware configurations, respectively. The stages test whether family evidence and background correction still contribute when the backbone already conditions on structure.

Table 35: Per-configuration staged component scores for 10 inverse-folding configurations on ProteinGym (official 217-assay Average Spearman).

Model	Raw	Adaptive mix	+ Ungated CCD	+ Gated CCD	+ RSA	+ pLDDT (Full)
ESM-IF1	0.421	0.466	0.457	0.467	0.473	0.474
MIF-ST	0.430	0.440	0.441	0.444	0.456	0.458
ProteinMPNN (v_48_002)	0.378	0.425	0.417	0.431	0.441	0.443
ProteinMPNN (v_48_010)	0.385	0.431	0.424	0.439	0.449	0.451
ProteinMPNN (v_48_020)	0.389	0.434	0.430	0.444	0.454	0.456
ProteinMPNN (v_48_030)	0.381	0.428	0.431	0.440	0.451	0.453
ProteinMPNN-Soluble (v_48_002)	0.357	0.409	0.408	0.418	0.428	0.430
ProteinMPNN-Soluble (v_48_010)	0.363	0.412	0.415	0.423	0.433	0.436
ProteinMPNN-Soluble (v_48_020)	0.361	0.413	0.414	0.422	0.433	0.435
ProteinMPNN-Soluble (v_48_030)	0.353	0.408	0.413	0.419	0.430	0.433

Table 36: Per-configuration staged component scores for 26 structure-aware configurations on ProteinGym (official 217-assay Average Spearman).

Model	Raw	Adaptive mix	+ Ungated CCD	+ Gated CCD	+ RSA	+ pLDDT (Full)
ESM3	0.442	0.461	0.460	0.469	0.481	0.483
ProSST (K=1024)	0.485	0.517	0.513	0.522	0.530	0.532
ProSST (K=128)	0.469	0.506	0.503	0.512	0.522	0.525
ProSST (K=20)	0.438	0.486	0.481	0.495	0.507	0.510
ProSST (K=2048)	0.507	0.525	0.522	0.528	0.533	0.534
ProSST (K=4096)	0.498	0.525	0.522	0.531	0.538	0.541
ProSST (K=512)	0.471	0.507	0.503	0.513	0.522	0.524
ProtSSN	0.451	0.453	0.433	0.454	0.474	0.476
ProtSSN (k=10, h=1280)	0.435	0.438	0.420	0.438	0.462	0.465
ProtSSN (k=10, h=512)	0.432	0.435	0.418	0.435	0.458	0.462
ProtSSN (k=10, h=768)	0.424	0.426	0.408	0.427	0.454	0.457
ProtSSN (k=20, h=1280)	0.444	0.446	0.427	0.446	0.467	0.470
ProtSSN (k=20, h=512)	0.443	0.445	0.428	0.445	0.467	0.470
ProtSSN (k=20, h=768)	0.442	0.444	0.424	0.445	0.466	0.470
ProtSSN (k=30, h=1280)	0.439	0.441	0.422	0.442	0.463	0.466
ProtSSN (k=30, h=512)	0.441	0.443	0.426	0.444	0.466	0.469
ProtSSN (k=30, h=768)	0.441	0.443	0.428	0.444	0.464	0.467
S3F (mask)	0.466	0.475	0.477	0.479	0.486	0.488
S3F (wt)	0.460	0.471	0.466	0.474	0.485	0.489
SaProt (35M_AF2, mask)	0.408	0.444	0.448	0.452	0.463	0.465
SaProt (35M_AF2, wt)	0.366	0.373	0.376	0.381	0.419	0.425
SaProt (650M_PDB, mask)	0.463	0.478	0.476	0.480	0.487	0.486
SaProt (650M_PDB, wt)	0.428	0.432	0.430	0.436	0.451	0.450
SaProt (650M AF2, mask)	0.457	0.474	0.474	0.478	0.486	0.487
SaProt (650M AF2, wt)	0.424	0.427	0.427	0.433	0.452	0.454
VenusREM2	0.524	0.542	0.538	0.550	0.554	0.556

ProteinGym: sequence readouts. Tables 37 and 38 trace the autoregressive and masked sequence configurations, respectively. The full readout improves on Raw for every configuration, while the intermediate changes show how retrieval, calibration, and structural attenuation contribute differently across models.

Table 37: Per-configuration staged component scores for 15 autoregressive language model configurations on ProteinGym (official 217-assay Average Spearman).

Model	Raw	Adaptive mix	+ Ungated CCD	+ Gated CCD	+ RSA	+ pLDDT (Full)
ProGen2-B	0.330	0.377	0.387	0.396	0.419	0.422
ProGen2-L	0.327	0.371	0.382	0.390	0.413	0.416
ProGen2-M	0.337	0.380	0.391	0.398	0.420	0.423
ProGen2-S	0.288	0.362	0.378	0.386	0.404	0.406
ProGen2-XL	0.351	0.373	0.375	0.385	0.415	0.419
ProGen3-112M	0.232	0.341	0.367	0.381	0.398	0.401
ProGen3-1B	0.320	0.368	0.373	0.380	0.403	0.406
ProGen3-219M	0.264	0.354	0.378	0.389	0.406	0.409
ProGen3-339M	0.286	0.364	0.385	0.395	0.412	0.414
ProGen3-3B	0.331	0.370	0.375	0.385	0.411	0.413
ProGen3-762M	0.304	0.369	0.382	0.392	0.411	0.413
RITA-L	0.309	0.358	0.368	0.377	0.404	0.407
RITA-M	0.297	0.358	0.370	0.379	0.402	0.405
RITA-S	0.254	0.344	0.366	0.376	0.395	0.398
RITA-XL	0.305	0.352	0.358	0.367	0.398	0.401

Table 38: Per-configuration staged component scores for 20 masked sequence language model configurations on ProteinGym (official 217-assay Average Spearman).

Model	Raw	Adaptive mix	+ Ungated CCD	+ Gated CCD	+ RSA	+ pLDDT (Full)
CARP-38M	0.314	0.342	0.356	0.367	0.409	0.415
CARP-600K	0.159	0.221	0.249	0.271	0.345	0.354
CARP-640M	0.390	0.427	0.433	0.437	0.450	0.451
CARP-76M	0.360	0.380	0.391	0.399	0.432	0.436
ESM-1b (mask)	0.389	0.430	0.436	0.440	0.453	0.455
ESM-1b (wt)	0.394	0.407	0.414	0.418	0.445	0.447
ESM-1v (mask)	0.407	0.436	0.444	0.450	0.461	0.462
ESM-1v (wt)	0.410	0.419	0.428	0.433	0.455	0.457
ESM-2 150M (mask)	0.388	0.424	0.434	0.438	0.450	0.452
ESM-2 150M (wt)	0.393	0.406	0.416	0.421	0.451	0.455
ESM-2 35M (mask)	0.321	0.395	0.407	0.414	0.427	0.430
ESM-2 35M (wt)	0.329	0.355	0.367	0.377	0.419	0.426
ESM-2 3B (mask)	0.407	0.436	0.444	0.445	0.459	0.462
ESM-2 3B (wt)	0.406	0.417	0.425	0.426	0.451	0.455
ESM-2 650M (mask)	0.415	0.445	0.454	0.457	0.468	0.470
ESM-2 650M (wt)	0.418	0.429	0.438	0.440	0.465	0.468
ESM-2 8M (mask)	0.223	0.345	0.367	0.378	0.392	0.395
ESM-2 8M (wt)	0.251	0.291	0.307	0.323	0.380	0.387
ESMC-300M	0.409	0.438	0.440	0.446	0.458	0.460
ESMC-600M	0.407	0.438	0.445	0.449	0.462	0.464

VenusMutHub: structural readouts. Tables 39 and 40 trace the inverse-folding and structure-aware configurations, respectively. The stages test whether family evidence and background correction still contribute when the backbone already conditions on structure.

Table 39: Per-configuration staged component scores for 10 inverse-folding configurations on VenusMutHub (assay-macro Spearman over 905 assays).

Model	Raw	Adaptive mix	+ Ungated CCD	+ Gated CCD	+ RSA	+ pLDDT (Full)
ESM-IF1	0.216	0.260	0.243	0.260	0.268	0.267
MIF-ST	0.196	0.209	0.203	0.210	0.228	0.230
ProteinMPNN (v_48_002)	0.191	0.233	0.232	0.241	0.245	0.244
ProteinMPNN (v_48_010)	0.218	0.240	0.246	0.247	0.254	0.256
ProteinMPNN (v_48_020)	0.221	0.257	0.254	0.262	0.269	0.271
ProteinMPNN (v_48_030)	0.210	0.244	0.251	0.251	0.260	0.259
ProteinMPNN-Soluble (v_48_002)	0.198	0.236	0.231	0.241	0.249	0.250
ProteinMPNN-Soluble (v_48_010)	0.212	0.246	0.241	0.248	0.256	0.256
ProteinMPNN-Soluble (v_48_020)	0.216	0.249	0.253	0.257	0.266	0.265
ProteinMPNN-Soluble (v_48_030)	0.222	0.250	0.253	0.255	0.264	0.265

Table 40: Per-configuration staged component scores for 26 structure-aware configurations on VenusMutHub (assay-macro Spearman over 905 assays).

Model	Raw	Adaptive mix	+ Ungated CCD	+ Gated CCD	+ RSA	+ pLDDT (Full)
ESM3	0.185	0.210	0.214	0.213	0.226	0.226
ProSST (K=1024)	0.162	0.223	0.236	0.231	0.241	0.242
ProSST (K=128)	0.139	0.213	0.225	0.221	0.231	0.230
ProSST (K=20)	0.143	0.213	0.216	0.224	0.237	0.238
ProSST (K=2048)	0.213	0.234	0.233	0.237	0.250	0.249
ProSST (K=4096)	0.161	0.225	0.228	0.230	0.238	0.237
ProSST (K=512)	0.124	0.210	0.220	0.217	0.226	0.226
ProtSSN	0.174	0.175	0.158	0.174	0.197	0.195
ProtSSN (k=10, h=1280)	0.166	0.168	0.155	0.167	0.191	0.192
ProtSSN (k=10, h=512)	0.169	0.171	0.157	0.171	0.196	0.197
ProtSSN (k=10, h=768)	0.152	0.156	0.143	0.155	0.178	0.178
ProtSSN (k=20, h=1280)	0.174	0.176	0.162	0.175	0.198	0.198
ProtSSN (k=20, h=512)	0.168	0.171	0.158	0.171	0.192	0.192
ProtSSN (k=20, h=768)	0.165	0.169	0.155	0.168	0.189	0.191
ProtSSN (k=30, h=1280)	0.162	0.164	0.144	0.162	0.184	0.183
ProtSSN (k=30, h=512)	0.169	0.172	0.158	0.172	0.193	0.191
ProtSSN (k=30, h=768)	0.160	0.165	0.158	0.164	0.185	0.185
S3F (mask)	0.211	0.221	0.228	0.224	0.233	0.234
S3F (wt)	0.193	0.208	0.212	0.209	0.225	0.224
SaProt (35M_AF2, mask)	0.188	0.224	0.228	0.228	0.236	0.237
SaProt (35M_AF2, wt)	0.142	0.148	0.153	0.156	0.181	0.179
SaProt (650M_PDB, mask)	0.205	0.219	0.220	0.220	0.237	0.240
SaProt (650M_PDB, wt)	0.182	0.185	0.187	0.188	0.217	0.215
SaProt (650M AF2, mask)	0.207	0.225	0.230	0.225	0.239	0.240
SaProt (650M AF2, wt)	0.178	0.180	0.189	0.183	0.203	0.202
VenusREM2	0.190	0.243	0.255	0.248	0.257	0.258

VenusMutHub: sequence readouts. Tables 41 and 42 trace the autoregressive and masked sequence configurations, respectively. The full readout improves on Raw for every configuration, while the intermediate changes show how retrieval, calibration, and structural attenuation contribute differently across models.

Table 41: Per-configuration staged component scores for 15 autoregressive language model configurations on VenusMutHub (assay-macro Spearman over 905 assays).

Model	Raw	Adaptive mix	+ Ungated CCD	+ Gated CCD	+ RSA	+ pLDDT (Full)
ProGen2-B	0.124	0.168	0.183	0.180	0.193	0.194
ProGen2-L	0.120	0.159	0.175	0.175	0.189	0.188
ProGen2-M	0.115	0.154	0.173	0.167	0.182	0.182
ProGen2-S	0.091	0.149	0.179	0.179	0.191	0.192
ProGen2-XL	0.123	0.148	0.149	0.153	0.178	0.176
ProGen3-112M	0.072	0.141	0.160	0.162	0.176	0.177
ProGen3-1B	0.108	0.152	0.160	0.158	0.174	0.173
ProGen3-219M	0.092	0.150	0.184	0.181	0.191	0.190
ProGen3-339M	0.110	0.154	0.177	0.174	0.190	0.189
ProGen3-3B	0.110	0.149	0.147	0.153	0.173	0.170
ProGen3-762M	0.111	0.160	0.174	0.172	0.184	0.180
RITA-L	0.110	0.160	0.182	0.179	0.193	0.193
RITA-M	0.100	0.155	0.178	0.175	0.188	0.186
RITA-S	0.056	0.148	0.173	0.171	0.181	0.182
RITA-XL	0.095	0.132	0.142	0.139	0.157	0.160

Table 42: Per-configuration staged component scores for 20 masked sequence language model configurations on VenusMutHub (assay-macro Spearman over 905 assays).

Model	Raw	Adaptive mix	+ Ungated CCD	+ Gated CCD	+ RSA	+ pLDDT (Full)
CARP-38M	0.073	0.102	0.136	0.132	0.159	0.158
CARP-600K	0.021	0.063	0.120	0.115	0.150	0.151
CARP-640M	0.157	0.187	0.191	0.194	0.206	0.206
CARP-76M	0.131	0.149	0.172	0.174	0.192	0.194
ESM-1b (mask)	0.166	0.198	0.200	0.200	0.214	0.211
ESM-1b (wt)	0.153	0.166	0.174	0.171	0.194	0.193
ESM-1v (mask)	0.162	0.188	0.193	0.193	0.206	0.206
ESM-1v (wt)	0.153	0.162	0.165	0.168	0.187	0.188
ESM-2 150M (mask)	0.167	0.203	0.221	0.216	0.226	0.225
ESM-2 150M (wt)	0.165	0.177	0.192	0.191	0.216	0.216
ESM-2 35M (mask)	0.122	0.177	0.207	0.199	0.212	0.212
ESM-2 35M (wt)	0.110	0.134	0.169	0.159	0.187	0.186
ESM-2 3B (mask)	0.182	0.205	0.209	0.208	0.221	0.223
ESM-2 3B (wt)	0.161	0.170	0.170	0.174	0.194	0.194
ESM-2 650M (mask)	0.169	0.197	0.199	0.200	0.214	0.213
ESM-2 650M (wt)	0.152	0.163	0.167	0.165	0.187	0.187
ESM-2 8M (mask)	0.079	0.158	0.189	0.187	0.197	0.198
ESM-2 8M (wt)	0.065	0.103	0.149	0.146	0.176	0.176
ESMC-300M	0.144	0.177	0.181	0.182	0.196	0.198
ESMC-600M	0.155	0.172	0.181	0.176	0.193	0.194

ViroHub: structural readouts. Tables 43 and 44 trace the inverse-folding and structure-aware configurations, respectively. The stages test whether family evidence and background correction still contribute when the backbone already conditions on structure.

Table 43: Per-configuration staged component scores for 10 inverse-folding configurations on ViroHub: mutant-only hierarchical Spearman over 89 assays, 52 cells, and 7 phenotypes; calibration and shrinkage use $\beta = 1 - \alpha$.

Model	Raw	Adaptive mix	+ Ungated CCD	+ Gated CCD	+ RSA	+ pLDDT (Full)
ESM-IF1	0.285	0.305	0.300	0.307	0.316	0.318
MIF-ST	0.288	0.299	0.304	0.303	0.316	0.317
ProteinMPNN (v_48_002)	0.269	0.285	0.283	0.291	0.304	0.305
ProteinMPNN (v_48_010)	0.271	0.285	0.285	0.292	0.305	0.306
ProteinMPNN (v_48_020)	0.270	0.284	0.285	0.291	0.304	0.306
ProteinMPNN (v_48_030)	0.266	0.280	0.284	0.288	0.302	0.304
ProteinMPNN-Soluble (v_48_002)	0.269	0.284	0.285	0.292	0.304	0.305
ProteinMPNN-Soluble (v_48_010)	0.268	0.281	0.285	0.289	0.302	0.304
ProteinMPNN-Soluble (v_48_020)	0.265	0.280	0.282	0.288	0.302	0.304
ProteinMPNN-Soluble (v_48_030)	0.256	0.272	0.277	0.281	0.297	0.299

Table 44: Per-configuration staged component scores for 26 structure-aware configurations on ViroHub: mutant-only hierarchical Spearman over 89 assays, 52 cells, and 7 phenotypes; calibration and shrinkage use $\beta = 1 - \alpha$.

Model	Raw	Adaptive mix	+ Ungated CCD	+ Gated CCD	+ RSA	+ pLDDT (Full)
ESM3	0.137	0.183	0.196	0.211	0.244	0.249
ProSST (K=1024)	0.204	0.243	0.247	0.250	0.272	0.275
ProSST (K=128)	0.179	0.223	0.232	0.234	0.261	0.264
ProSST (K=20)	0.161	0.211	0.213	0.224	0.254	0.259
ProSST (K=2048)	0.271	0.290	0.284	0.294	0.309	0.310
ProSST (K=4096)	0.200	0.237	0.242	0.244	0.268	0.271
ProSST (K=512)	0.191	0.231	0.239	0.241	0.265	0.269
ProtSSN	0.233	0.236	0.208	0.237	0.288	0.294
ProtSSN (k=10, h=1280)	0.220	0.223	0.204	0.224	0.278	0.284
ProtSSN (k=10, h=512)	0.202	0.206	0.180	0.207	0.265	0.272
ProtSSN (k=10, h=768)	0.214	0.217	0.192	0.218	0.273	0.279
ProtSSN (k=20, h=1280)	0.230	0.233	0.208	0.234	0.283	0.289
ProtSSN (k=20, h=512)	0.233	0.237	0.211	0.237	0.285	0.291
ProtSSN (k=20, h=768)	0.226	0.229	0.205	0.230	0.281	0.287
ProtSSN (k=30, h=1280)	0.210	0.214	0.181	0.215	0.273	0.279
ProtSSN (k=30, h=512)	0.225	0.229	0.205	0.229	0.282	0.288
ProtSSN (k=30, h=768)	0.223	0.227	0.199	0.228	0.280	0.286
S3F (mask)	0.269	0.280	0.284	0.285	0.301	0.302
S3F (wt)	0.279	0.291	0.291	0.296	0.314	0.318
SaProt (35M_AF2, mask)	0.149	0.207	0.216	0.217	0.242	0.245
SaProt (35M_AF2, wt)	0.108	0.121	0.126	0.127	0.199	0.209
SaProt (650M_PDB, mask)	0.306	0.307	0.309	0.309	0.320	0.320
SaProt (650M_PDB, wt)	0.273	0.280	0.280	0.283	0.307	0.307
SaProt (650M AF2, mask)	0.222	0.253	0.259	0.259	0.279	0.281
SaProt (650M AF2, wt)	0.188	0.195	0.200	0.201	0.254	0.259
VenusREM2	0.236	0.263	0.267	0.274	0.294	0.297

ViroHub: sequence readouts. Tables 45 and 46 trace the autoregressive and masked sequence configurations, respectively. The full readout improves on Raw for every configuration, while the intermediate changes show how retrieval, calibration, and structural attenuation contribute differently across models.

Table 45: Per-configuration staged component scores for 15 autoregressive language model configurations on ViroHub: mutant-only hierarchical Spearman over 89 assays, 52 cells, and 7 phenotypes; calibration and shrinkage use $\beta = 1 - \alpha$.

Model	Raw	Adaptive mix	+ Ungated CCD	+ Gated CCD	+ RSA	+ pLDDT (Full)
ProGen2-B	0.076	0.123	0.132	0.136	0.172	0.175
ProGen2-L	0.064	0.121	0.136	0.137	0.169	0.173
ProGen2-M	0.117	0.152	0.159	0.165	0.208	0.212
ProGen2-S	0.045	0.114	0.128	0.133	0.167	0.171
ProGen2-XL	0.162	0.179	0.179	0.186	0.232	0.235
ProGen3-112M	0.034	0.117	0.152	0.159	0.178	0.182
ProGen3-1B	0.092	0.149	0.171	0.177	0.204	0.207
ProGen3-219M	0.033	0.115	0.152	0.158	0.177	0.181
ProGen3-339M	0.058	0.133	0.162	0.167	0.187	0.191
ProGen3-3B	0.101	0.154	0.174	0.180	0.208	0.211
ProGen3-762M	0.062	0.132	0.159	0.165	0.189	0.192
RITA-L	0.166	0.181	0.184	0.189	0.245	0.247
RITA-M	0.155	0.175	0.176	0.182	0.234	0.237
RITA-S	0.148	0.176	0.179	0.184	0.233	0.236
RITA-XL	0.172	0.184	0.187	0.191	0.244	0.247

Table 46: Per-configuration staged component scores for 20 masked sequence language model configurations on ViroHub: mutant-only hierarchical Spearman over 89 assays, 52 cells, and 7 phenotypes; calibration and shrinkage use $\beta = 1 - \alpha$.

Model	Raw	Adaptive mix	+ Ungated CCD	+ Gated CCD	+ RSA	+ pLDDT (Full)
CARP-38M	0.027	0.067	0.096	0.112	0.213	0.222
CARP-600K	0.006	0.047	0.087	0.098	0.201	0.210
CARP-640M	0.092	0.162	0.185	0.190	0.212	0.215
CARP-76M	0.050	0.090	0.113	0.130	0.221	0.230
ESM-1b (mask)	0.093	0.169	0.192	0.197	0.220	0.223
ESM-1b (wt)	0.094	0.127	0.150	0.159	0.240	0.248
ESM-1v (mask)	0.064	0.140	0.173	0.180	0.200	0.202
ESM-1v (wt)	0.069	0.099	0.124	0.137	0.229	0.238
ESM-2 150M (mask)	0.071	0.153	0.175	0.182	0.205	0.209
ESM-2 150M (wt)	0.069	0.103	0.122	0.136	0.226	0.235
ESM-2 35M (mask)	0.026	0.119	0.152	0.158	0.180	0.184
ESM-2 35M (wt)	0.026	0.063	0.092	0.099	0.203	0.213
ESM-2 3B (mask)	0.145	0.199	0.216	0.218	0.240	0.243
ESM-2 3B (wt)	0.138	0.164	0.183	0.186	0.251	0.260
ESM-2 650M (mask)	0.111	0.177	0.198	0.201	0.224	0.227
ESM-2 650M (wt)	0.114	0.142	0.163	0.168	0.243	0.252
ESM-2 8M (mask)	0.015	0.108	0.140	0.149	0.170	0.173
ESM-2 8M (wt)	0.017	0.053	0.082	0.097	0.203	0.214
ESMC-300M	0.097	0.164	0.184	0.193	0.216	0.220
ESMC-600M	0.108	0.173	0.191	0.200	0.224	0.228

I.3 Per-configuration five-metric results

Tables 47–58 are the configuration-level source of Figure 4: all 355 model–metric pairs improve on ProteinGym and ViroHub, and 353 of 355 improve or remain unchanged at the reported precision on VenusMutHub.

ProteinGym: structural models. Tables 47 and 48 compare the five metrics for structure-aware and inverse-folding models, respectively. The paired columns distinguish changes in ranking, binary discrimination, and recovery of top variants under the ProteinGym definitions.

Table 47: ProteinGym structure-aware models with complete five-metric Raw–RH exports.

Model	Spearman		NDCG		AUC		MCC		Top-recall	
	Raw	RH	Raw	RH	Raw	RH	Raw	RH	Raw	RH
ESM3	0.442	0.483	0.767	0.780	0.744	0.765	0.346	0.377	0.228	0.236
ProSST (K=1024)	0.485	0.532	0.761	0.790	0.764	0.788	0.372	0.410	0.230	0.251
ProSST (K=128)	0.469	0.525	0.754	0.789	0.757	0.786	0.363	0.407	0.227	0.253
ProSST (K=20)	0.438	0.510	0.745	0.786	0.739	0.777	0.336	0.393	0.210	0.242
ProSST (K=2048)	0.507	0.534	0.757	0.778	0.777	0.791	0.398	0.416	0.236	0.249
ProSST (K=4096)	0.498	0.541	0.774	0.798	0.773	0.794	0.385	0.419	0.232	0.255
ProSST (K=512)	0.471	0.524	0.759	0.791	0.757	0.785	0.360	0.404	0.222	0.248
ProtSSN	0.451	0.476	0.766	0.777	0.748	0.762	0.353	0.373	0.228	0.231
ProtSSN (k=10, h=1280)	0.435	0.465	0.762	0.776	0.740	0.756	0.343	0.365	0.226	0.231
ProtSSN (k=10, h=512)	0.432	0.462	0.759	0.772	0.738	0.754	0.341	0.362	0.221	0.224
ProtSSN (k=10, h=768)	0.424	0.457	0.760	0.773	0.732	0.750	0.332	0.357	0.226	0.228
ProtSSN (k=20, h=1280)	0.444	0.470	0.762	0.775	0.744	0.759	0.350	0.369	0.228	0.230
ProtSSN (k=20, h=512)	0.443	0.470	0.764	0.777	0.743	0.758	0.346	0.367	0.227	0.231
ProtSSN (k=20, h=768)	0.442	0.470	0.763	0.775	0.742	0.757	0.344	0.366	0.225	0.229
ProtSSN (k=30, h=1280)	0.439	0.466	0.759	0.773	0.742	0.756	0.344	0.364	0.223	0.224
ProtSSN (k=30, h=512)	0.441	0.469	0.765	0.776	0.741	0.757	0.341	0.365	0.223	0.228
ProtSSN (k=30, h=768)	0.441	0.467	0.766	0.776	0.742	0.757	0.345	0.364	0.226	0.228
S3F (mask)	0.466	0.488	0.770	0.785	0.755	0.766	0.365	0.383	0.237	0.242
S3F (wt)	0.460	0.489	0.773	0.784	0.752	0.767	0.361	0.384	0.233	0.240
SaProt (35M_AF2, mask)	0.408	0.465	0.736	0.763	0.724	0.754	0.318	0.362	0.215	0.236
SaProt (35M_AF2, wt)	0.366	0.425	0.726	0.753	0.702	0.733	0.286	0.328	0.200	0.218
SaProt (650M_PDB, mask)	0.463	0.486	0.775	0.783	0.756	0.767	0.365	0.381	0.231	0.237
SaProt (650M_PDB, wt)	0.428	0.450	0.766	0.771	0.736	0.749	0.336	0.355	0.221	0.223
SaProt (650M AF2, mask)	0.457	0.487	0.769	0.780	0.752	0.767	0.359	0.383	0.234	0.243
SaProt (650M AF2, wt)	0.424	0.454	0.757	0.770	0.735	0.750	0.335	0.355	0.223	0.226
VenusREM2	0.524	0.556	0.791	0.808	0.786	0.802	0.405	0.430	0.256	0.266

Table 48: ProteinGym inverse-folding models with complete five-metric Raw–RH exports.

Model	Spearman		NDCG		AUC		MCC		Top-recall	
	Raw	RH	Raw	RH	Raw	RH	Raw	RH	Raw	RH
ESM-IF1	0.421	0.474	0.747	0.777	0.729	0.759	0.328	0.372	0.223	0.239
MIF-ST	0.430	0.458	0.772	0.778	0.736	0.752	0.339	0.362	0.233	0.235
ProteinMPNN (v_48_002)	0.378	0.443	0.741	0.770	0.706	0.741	0.302	0.345	0.218	0.239
ProteinMPNN (v_48_010)	0.385	0.451	0.747	0.774	0.709	0.745	0.303	0.350	0.225	0.244
ProteinMPNN (v_48_020)	0.389	0.456	0.746	0.774	0.710	0.747	0.304	0.353	0.226	0.245
ProteinMPNN (v_48_030)	0.381	0.453	0.736	0.769	0.707	0.746	0.296	0.352	0.218	0.240
ProteinMPNN-Soluble (v_48_002)	0.357	0.430	0.731	0.765	0.695	0.734	0.282	0.333	0.213	0.232
ProteinMPNN-Soluble (v_48_010)	0.363	0.436	0.735	0.767	0.697	0.737	0.286	0.337	0.215	0.236
ProteinMPNN-Soluble (v_48_020)	0.361	0.435	0.731	0.764	0.696	0.737	0.283	0.335	0.215	0.233
ProteinMPNN-Soluble (v_48_030)	0.353	0.433	0.726	0.762	0.691	0.735	0.273	0.336	0.211	0.231

ProteinGym: sequence models. Tables 49 and 50 compare the five metrics for masked and autoregressive models, respectively. The paired columns distinguish changes in ranking, binary discrimination, and recovery of top variants under the ProteinGym definitions.

Table 49: ProteinGym masked protein LMs with complete five-metric Raw–RH results.

Model	Spearman		NDCG		AUC		MCC		Top-recall	
	Raw	RH	Raw	RH	Raw	RH	Raw	RH	Raw	RH
CARP-38M	0.314	0.415	0.709	0.755	0.675	0.729	0.248	0.322	0.186	0.212
CARP-600K	0.159	0.354	0.649	0.740	0.590	0.694	0.127	0.271	0.144	0.194
CARP-640M	0.390	0.451	0.747	0.767	0.716	0.748	0.307	0.355	0.208	0.223
CARP-76M	0.360	0.436	0.728	0.759	0.699	0.740	0.282	0.339	0.196	0.214
ESM-1b (mask)	0.389	0.455	0.742	0.765	0.715	0.750	0.306	0.360	0.202	0.218
ESM-1b (wt)	0.394	0.447	0.747	0.768	0.719	0.747	0.311	0.353	0.202	0.214
ESM-1v (mask)	0.407	0.462	0.750	0.763	0.724	0.753	0.320	0.361	0.211	0.221
ESM-1v (wt)	0.410	0.457	0.752	0.771	0.727	0.752	0.324	0.357	0.210	0.219
ESM-2 150M (mask)	0.388	0.452	0.729	0.754	0.715	0.748	0.304	0.355	0.205	0.222
ESM-2 150M (wt)	0.393	0.455	0.736	0.763	0.718	0.751	0.310	0.355	0.210	0.225
ESM-2 35M (mask)	0.321	0.430	0.705	0.746	0.676	0.735	0.250	0.335	0.186	0.215
ESM-2 35M (wt)	0.329	0.426	0.712	0.756	0.682	0.734	0.257	0.329	0.192	0.217
ESM-2 3B (mask)	0.407	0.462	0.756	0.777	0.724	0.754	0.321	0.364	0.213	0.228
ESM-2 3B (wt)	0.406	0.455	0.759	0.776	0.725	0.751	0.324	0.358	0.213	0.221
ESM-2 650M (mask)	0.415	0.470	0.748	0.774	0.729	0.758	0.328	0.369	0.217	0.232
ESM-2 650M (wt)	0.418	0.468	0.751	0.775	0.732	0.758	0.334	0.366	0.214	0.226
ESM-2 8M (mask)	0.223	0.395	0.667	0.731	0.623	0.716	0.170	0.310	0.153	0.200
ESM-2 8M (wt)	0.251	0.387	0.685	0.748	0.641	0.713	0.196	0.297	0.167	0.204
ESMC-300M	0.409	0.460	0.751	0.766	0.728	0.754	0.322	0.363	0.213	0.222
ESMC-600M	0.407	0.464	0.750	0.770	0.726	0.755	0.321	0.364	0.204	0.221

Table 50: ProteinGym autoregressive LM models with complete five-metric Raw–RH exports.

Model	Spearman		NDCG		AUC		MCC		Top-recall	
	Raw	RH	Raw	RH	Raw	RH	Raw	RH	Raw	RH
ProGen2-B	0.330	0.422	0.736	0.765	0.682	0.731	0.259	0.332	0.193	0.212
ProGen2-L	0.327	0.416	0.737	0.764	0.681	0.728	0.258	0.328	0.192	0.214
ProGen2-M	0.337	0.423	0.736	0.764	0.687	0.732	0.266	0.333	0.191	0.213
ProGen2-S	0.288	0.406	0.708	0.747	0.658	0.721	0.223	0.314	0.175	0.207
ProGen2-XL	0.351	0.419	0.755	0.772	0.694	0.730	0.275	0.329	0.198	0.207
ProGen3-112M	0.232	0.401	0.680	0.735	0.628	0.719	0.179	0.315	0.160	0.199
ProGen3-1B	0.320	0.406	0.732	0.757	0.677	0.722	0.251	0.316	0.191	0.210
ProGen3-219M	0.264	0.409	0.698	0.745	0.643	0.722	0.203	0.316	0.175	0.210
ProGen3-339M	0.286	0.414	0.708	0.748	0.656	0.726	0.222	0.322	0.176	0.208
ProGen3-3B	0.331	0.413	0.744	0.765	0.683	0.727	0.257	0.323	0.195	0.210
ProGen3-762M	0.304	0.413	0.719	0.753	0.667	0.726	0.237	0.321	0.181	0.209
RITA-L	0.309	0.407	0.727	0.757	0.672	0.725	0.241	0.321	0.180	0.204
RITA-M	0.297	0.405	0.714	0.751	0.664	0.722	0.230	0.315	0.177	0.200
RITA-S	0.254	0.398	0.696	0.743	0.641	0.718	0.203	0.314	0.170	0.199
RITA-XL	0.305	0.401	0.731	0.762	0.670	0.722	0.241	0.318	0.179	0.203

VenusMutHub: structural models. Tables 51 and 52 compare the five metrics for structure-aware and inverse-folding models, respectively. These assay-macro results retain the VenusMutHub metric definitions; comparisons concern Raw and the harness within each configuration rather than absolute values across benchmarks.

Table 51: VenusMutHub structure-aware models with complete five-metric Raw–RH exports.

Model	Spearman		NDCG		AUC		MCC		Top-recall	
	Raw	RH	Raw	RH	Raw	RH	Raw	RH	Raw	RH
ESM3	0.185	0.226	0.837	0.842	0.598	0.619	0.147	0.187	0.219	0.219
ProSST (K=1024)	0.162	0.242	0.822	0.845	0.585	0.627	0.131	0.187	0.201	0.254
ProSST (K=128)	0.139	0.230	0.817	0.840	0.575	0.620	0.103	0.176	0.181	0.234
ProSST (K=20)	0.143	0.238	0.820	0.841	0.576	0.627	0.114	0.189	0.191	0.225
ProSST (K=2048)	0.213	0.249	0.840	0.847	0.614	0.630	0.170	0.199	0.228	0.246
ProSST (K=4096)	0.161	0.237	0.823	0.842	0.585	0.623	0.120	0.188	0.193	0.234
ProSST (K=512)	0.124	0.226	0.813	0.840	0.564	0.617	0.089	0.164	0.170	0.224
ProtSSN	0.174	0.195	0.836	0.839	0.588	0.601	0.134	0.148	0.220	0.220
ProtSSN (k=10, h=1280)	0.166	0.192	0.834	0.838	0.584	0.598	0.126	0.141	0.221	0.224
ProtSSN (k=10, h=512)	0.169	0.197	0.835	0.839	0.585	0.600	0.129	0.145	0.217	0.222
ProtSSN (k=10, h=768)	0.152	0.178	0.830	0.835	0.579	0.593	0.115	0.133	0.207	0.212
ProtSSN (k=20, h=1280)	0.174	0.198	0.836	0.838	0.590	0.605	0.127	0.154	0.208	0.209
ProtSSN (k=20, h=512)	0.168	0.192	0.835	0.838	0.586	0.600	0.122	0.143	0.212	0.216
ProtSSN (k=20, h=768)	0.165	0.191	0.835	0.840	0.583	0.598	0.125	0.146	0.222	0.224
ProtSSN (k=30, h=1280)	0.162	0.183	0.832	0.836	0.582	0.594	0.126	0.142	0.215	0.216
ProtSSN (k=30, h=512)	0.169	0.191	0.835	0.838	0.585	0.599	0.131	0.152	0.210	0.216
ProtSSN (k=30, h=768)	0.160	0.185	0.833	0.838	0.581	0.595	0.113	0.134	0.199	0.210
S3F (mask)	0.211	0.234	0.839	0.843	0.606	0.621	0.156	0.178	0.230	0.221
S3F (wt)	0.193	0.224	0.837	0.841	0.596	0.614	0.129	0.163	0.226	0.235
SaProt (35M_AF2, mask)	0.188	0.237	0.832	0.841	0.597	0.624	0.162	0.186	0.208	0.227
SaProt (35M_AF2, wt)	0.142	0.179	0.827	0.834	0.576	0.596	0.116	0.140	0.209	0.215
SaProt (650M_PDB, mask)	0.205	0.240	0.836	0.844	0.604	0.623	0.151	0.179	0.220	0.230
SaProt (650M_PDB, wt)	0.182	0.215	0.836	0.840	0.594	0.610	0.127	0.156	0.224	0.218
SaProt (650M AF2, mask)	0.207	0.240	0.837	0.842	0.606	0.623	0.161	0.184	0.223	0.230
SaProt (650M AF2, wt)	0.178	0.202	0.832	0.838	0.591	0.604	0.138	0.153	0.222	0.228
VenusREM2	0.190	0.258	0.830	0.846	0.601	0.636	0.143	0.198	0.211	0.250

Table 52: VenusMutHub inverse-folding models with complete five-metric Raw–RH exports.

Model	Spearman		NDCG		AUC		MCC		Top-recall	
	Raw	RH	Raw	RH	Raw	RH	Raw	RH	Raw	RH
ESM-IF1	0.216	0.267	0.833	0.846	0.613	0.640	0.164	0.206	0.203	0.235
MIF-ST	0.196	0.230	0.837	0.842	0.599	0.618	0.141	0.172	0.219	0.229
ProteinMPNN (v_48_002)	0.191	0.244	0.830	0.844	0.598	0.626	0.149	0.183	0.199	0.230
ProteinMPNN (v_48_010)	0.218	0.256	0.836	0.848	0.611	0.631	0.163	0.199	0.229	0.253
ProteinMPNN (v_48_020)	0.221	0.271	0.835	0.851	0.611	0.638	0.164	0.194	0.225	0.266
ProteinMPNN (v_48_030)	0.210	0.259	0.835	0.848	0.603	0.628	0.152	0.183	0.227	0.249
ProteinMPNN-Soluble (v_48_002)	0.198	0.250	0.829	0.845	0.599	0.626	0.141	0.186	0.202	0.238
ProteinMPNN-Soluble (v_48_010)	0.212	0.256	0.837	0.846	0.604	0.630	0.148	0.193	0.236	0.242
ProteinMPNN-Soluble (v_48_020)	0.216	0.265	0.837	0.849	0.609	0.634	0.162	0.199	0.235	0.253
ProteinMPNN-Soluble (v_48_030)	0.222	0.265	0.839	0.849	0.610	0.637	0.163	0.212	0.236	0.255

VenusMutHub: sequence models. Tables 53 and 54 compare the five metrics for masked and autoregressive models, respectively. These assay-macro results retain the VenusMutHub metric definitions; comparisons concern Raw and the harness within each configuration rather than absolute values across benchmarks.

Table 53: VenusMutHub masked protein LMs with complete five-metric Raw–RH results.

Model	Spearman		NDCG		AUC		MCC		Top-recall	
	Raw	RH	Raw	RH	Raw	RH	Raw	RH	Raw	RH
CARP-38M	0.073	0.158	0.816	0.831	0.543	0.587	0.057	0.120	0.195	0.206
CARP-600K	0.021	0.151	0.804	0.830	0.513	0.581	0.013	0.109	0.160	0.198
CARP-640M	0.157	0.206	0.830	0.838	0.584	0.609	0.128	0.152	0.192	0.212
CARP-76M	0.131	0.194	0.826	0.837	0.567	0.600	0.099	0.152	0.200	0.224
ESM-1b (mask)	0.166	0.211	0.831	0.838	0.589	0.612	0.137	0.162	0.208	0.215
ESM-1b (wt)	0.153	0.193	0.831	0.836	0.582	0.604	0.123	0.156	0.204	0.221
ESM-1v (mask)	0.162	0.206	0.830	0.838	0.581	0.606	0.124	0.165	0.198	0.215
ESM-1v (wt)	0.153	0.188	0.831	0.837	0.577	0.595	0.122	0.147	0.199	0.210
ESM-2 150M (mask)	0.167	0.225	0.831	0.840	0.588	0.617	0.136	0.179	0.197	0.221
ESM-2 150M (wt)	0.165	0.216	0.833	0.840	0.588	0.615	0.129	0.177	0.206	0.216
ESM-2 35M (mask)	0.122	0.212	0.821	0.840	0.565	0.612	0.093	0.165	0.174	0.230
ESM-2 35M (wt)	0.110	0.186	0.823	0.835	0.559	0.600	0.080	0.151	0.194	0.215
ESM-2 3B (mask)	0.182	0.223	0.832	0.838	0.590	0.613	0.147	0.174	0.208	0.218
ESM-2 3B (wt)	0.161	0.194	0.832	0.836	0.581	0.600	0.111	0.142	0.215	0.217
ESM-2 650M (mask)	0.169	0.213	0.832	0.839	0.588	0.610	0.130	0.166	0.213	0.221
ESM-2 650M (wt)	0.152	0.187	0.830	0.836	0.579	0.598	0.118	0.140	0.204	0.216
ESM-2 8M (mask)	0.079	0.198	0.807	0.834	0.539	0.602	0.067	0.162	0.141	0.207
ESM-2 8M (wt)	0.065	0.176	0.814	0.834	0.535	0.592	0.058	0.131	0.169	0.204
ESMC-300M	0.144	0.198	0.828	0.836	0.575	0.602	0.119	0.156	0.198	0.212
ESMC-600M	0.155	0.194	0.830	0.837	0.580	0.601	0.134	0.146	0.197	0.210

Table 54: VenusMutHub autoregressive LM models with complete five-metric Raw–RH exports.

Model	Spearman		NDCG		AUC		MCC		Top-recall	
	Raw	RH	Raw	RH	Raw	RH	Raw	RH	Raw	RH
ProGen2-B	0.124	0.194	0.823	0.837	0.564	0.598	0.101	0.159	0.184	0.210
ProGen2-L	0.120	0.188	0.823	0.835	0.560	0.598	0.091	0.147	0.187	0.210
ProGen2-M	0.115	0.182	0.825	0.835	0.558	0.595	0.078	0.133	0.189	0.206
ProGen2-S	0.091	0.192	0.815	0.838	0.546	0.600	0.069	0.157	0.174	0.223
ProGen2-XL	0.123	0.176	0.827	0.836	0.561	0.591	0.083	0.134	0.195	0.202
ProGen3-112M	0.072	0.177	0.812	0.833	0.537	0.594	0.045	0.137	0.169	0.209
ProGen3-1B	0.108	0.173	0.821	0.833	0.561	0.592	0.096	0.135	0.186	0.207
ProGen3-219M	0.092	0.190	0.816	0.835	0.548	0.600	0.073	0.149	0.183	0.212
ProGen3-339M	0.110	0.189	0.821	0.837	0.560	0.601	0.094	0.149	0.183	0.215
ProGen3-3B	0.110	0.170	0.823	0.835	0.553	0.588	0.064	0.120	0.200	0.208
ProGen3-762M	0.111	0.180	0.822	0.834	0.556	0.592	0.088	0.150	0.191	0.209
RITA-L	0.110	0.193	0.823	0.837	0.555	0.600	0.071	0.151	0.194	0.212
RITA-M	0.100	0.186	0.821	0.836	0.552	0.598	0.075	0.140	0.179	0.208
RITA-S	0.056	0.182	0.812	0.836	0.528	0.595	0.038	0.141	0.171	0.213
RITA-XL	0.095	0.160	0.820	0.831	0.547	0.585	0.065	0.130	0.182	0.208

ViroHub: structural models. Tables 55 and 56 compare the five metrics for structure-aware and inverse-folding models, respectively. All scores exclude WT controls. AUC and MCC use 82 assays in 45 phenotype–protein cells, while the other metrics use all 89 assays in 52 cells.

Table 55: ViroHub structure-aware models with mutant-only hierarchical phenotype–backbone means; AUC/MCC use 82 assays (45 cells), and the other metrics use 89 assays (52 cells).

Model	Spearman		NDCG		AUC		MCC		Top-recall	
	Raw	RH	Raw	RH	Raw	RH	Raw	RH	Raw	RH
ESM3	0.137	0.249	0.690	0.733	0.573	0.632	0.103	0.200	0.124	0.152
ProSST (K=1024)	0.204	0.275	0.673	0.727	0.608	0.645	0.157	0.219	0.125	0.153
ProSST (K=128)	0.179	0.264	0.671	0.725	0.596	0.640	0.139	0.212	0.127	0.152
ProSST (K=20)	0.161	0.259	0.669	0.721	0.586	0.637	0.122	0.212	0.126	0.155
ProSST (K=2048)	0.271	0.310	0.691	0.730	0.643	0.664	0.223	0.252	0.135	0.157
ProSST (K=4096)	0.200	0.271	0.674	0.723	0.609	0.646	0.156	0.220	0.127	0.150
ProSST (K=512)	0.191	0.269	0.681	0.728	0.603	0.643	0.145	0.214	0.127	0.151
ProtSSN	0.233	0.294	0.716	0.737	0.622	0.657	0.181	0.235	0.150	0.159
ProtSSN (k=10, h=1280)	0.220	0.284	0.714	0.739	0.615	0.652	0.171	0.225	0.149	0.162
ProtSSN (k=10, h=512)	0.202	0.272	0.702	0.735	0.605	0.646	0.155	0.215	0.141	0.157
ProtSSN (k=10, h=768)	0.214	0.279	0.710	0.736	0.612	0.650	0.166	0.224	0.144	0.161
ProtSSN (k=20, h=1280)	0.230	0.289	0.713	0.736	0.620	0.654	0.179	0.232	0.147	0.158
ProtSSN (k=20, h=512)	0.233	0.291	0.714	0.736	0.622	0.655	0.179	0.233	0.145	0.156
ProtSSN (k=20, h=768)	0.226	0.287	0.712	0.737	0.619	0.654	0.175	0.230	0.150	0.161
ProtSSN (k=30, h=1280)	0.210	0.279	0.706	0.732	0.609	0.649	0.159	0.221	0.146	0.156
ProtSSN (k=30, h=512)	0.225	0.288	0.713	0.735	0.617	0.653	0.172	0.229	0.151	0.159
ProtSSN (k=30, h=768)	0.223	0.286	0.708	0.734	0.616	0.652	0.173	0.228	0.144	0.156
S3F (mask)	0.269	0.302	0.718	0.752	0.640	0.659	0.212	0.244	0.154	0.169
S3F (wt)	0.279	0.318	0.739	0.752	0.646	0.669	0.219	0.258	0.162	0.170
SaProt (35M_AF2, mask)	0.149	0.245	0.660	0.720	0.579	0.630	0.112	0.198	0.133	0.154
SaProt (35M_AF2, wt)	0.108	0.209	0.650	0.699	0.555	0.612	0.076	0.162	0.119	0.140
SaProt (650M_PDB, mask)	0.306	0.320	0.727	0.753	0.662	0.670	0.245	0.259	0.164	0.171
SaProt (650M_PDB, wt)	0.273	0.307	0.714	0.735	0.644	0.663	0.218	0.244	0.152	0.164
SaProt (650M AF2, mask)	0.222	0.281	0.687	0.739	0.618	0.649	0.174	0.226	0.150	0.168
SaProt (650M AF2, wt)	0.188	0.259	0.675	0.712	0.597	0.638	0.144	0.201	0.138	0.153
VenusREM2	0.236	0.297	0.696	0.740	0.626	0.658	0.183	0.240	0.142	0.162

Table 56: ViroHub inverse-folding models with mutant-only hierarchical phenotype–backbone means; AUC/MCC use 82 assays (45 cells), and the other metrics use 89 assays (52 cells).

Model	Spearman		NDCG		AUC		MCC		Top-recall	
	Raw	RH	Raw	RH	Raw	RH	Raw	RH	Raw	RH
ESM-IF1	0.285	0.318	0.720	0.755	0.648	0.666	0.226	0.258	0.160	0.171
MIF-ST	0.288	0.317	0.736	0.752	0.653	0.669	0.234	0.259	0.166	0.173
ProteinMPNN (v_48_002)	0.269	0.305	0.729	0.754	0.647	0.663	0.229	0.252	0.170	0.180
ProteinMPNN (v_48_010)	0.271	0.306	0.728	0.753	0.647	0.664	0.225	0.253	0.169	0.178
ProteinMPNN (v_48_020)	0.270	0.306	0.722	0.749	0.647	0.663	0.225	0.255	0.164	0.175
ProteinMPNN (v_48_030)	0.266	0.304	0.718	0.747	0.644	0.662	0.221	0.251	0.161	0.173
ProteinMPNN-Soluble (v_48_002)	0.269	0.305	0.725	0.752	0.647	0.663	0.228	0.250	0.168	0.177
ProteinMPNN-Soluble (v_48_010)	0.268	0.304	0.726	0.754	0.646	0.662	0.225	0.251	0.169	0.177
ProteinMPNN-Soluble (v_48_020)	0.265	0.304	0.721	0.753	0.644	0.662	0.221	0.251	0.165	0.176
ProteinMPNN-Soluble (v_48_030)	0.256	0.299	0.714	0.749	0.640	0.660	0.215	0.247	0.163	0.176

ViroHub: sequence models. Tables 57 and 58 compare the five metrics for masked and autoregressive models, respectively. All scores exclude WT controls. AUC and MCC use 82 assays in 45 phenotype–protein cells, while the other metrics use all 89 assays in 52 cells.

Table 57: ViroHub masked LMs with mutant-only hierarchical phenotype–backbone means; AUC/MCC use 82 assays (45 cells), and the other metrics use 89 assays (52 cells).

Model	Spearman		NDCG		AUC		MCC		Top-recall	
	Raw	RH	Raw	RH	Raw	RH	Raw	RH	Raw	RH
CARP-38M	0.027	0.222	0.630	0.727	0.515	0.621	0.028	0.178	0.103	0.145
CARP-600K	0.006	0.210	0.614	0.722	0.504	0.615	0.013	0.170	0.099	0.143
CARP-640M	0.092	0.215	0.658	0.709	0.547	0.614	0.071	0.170	0.115	0.135
CARP-76M	0.050	0.230	0.637	0.728	0.527	0.625	0.040	0.181	0.109	0.151
ESM-1b (mask)	0.093	0.223	0.649	0.710	0.545	0.616	0.064	0.180	0.111	0.138
ESM-1b (wt)	0.094	0.248	0.648	0.730	0.545	0.632	0.064	0.192	0.115	0.154
ESM-1v (mask)	0.064	0.202	0.648	0.697	0.537	0.608	0.057	0.165	0.106	0.128
ESM-1v (wt)	0.069	0.238	0.651	0.735	0.539	0.630	0.061	0.191	0.106	0.151
ESM-2 150M (mask)	0.071	0.209	0.641	0.700	0.535	0.609	0.052	0.172	0.109	0.132
ESM-2 150M (wt)	0.069	0.235	0.639	0.720	0.533	0.626	0.050	0.185	0.109	0.145
ESM-2 35M (mask)	0.026	0.184	0.628	0.689	0.513	0.597	0.020	0.152	0.100	0.128
ESM-2 35M (wt)	0.026	0.213	0.624	0.720	0.514	0.616	0.023	0.168	0.099	0.144
ESM-2 3B (mask)	0.145	0.243	0.689	0.732	0.576	0.630	0.110	0.196	0.124	0.143
ESM-2 3B (wt)	0.138	0.260	0.684	0.736	0.572	0.639	0.103	0.203	0.122	0.153
ESM-2 650M (mask)	0.111	0.227	0.672	0.722	0.556	0.620	0.079	0.181	0.118	0.139
ESM-2 650M (wt)	0.114	0.252	0.673	0.734	0.557	0.634	0.082	0.196	0.119	0.153
ESM-2 8M (mask)	0.015	0.173	0.625	0.683	0.508	0.591	0.017	0.142	0.101	0.120
ESM-2 8M (wt)	0.017	0.214	0.625	0.724	0.509	0.616	0.017	0.170	0.100	0.144
ESMC-300M	0.097	0.220	0.668	0.712	0.550	0.615	0.073	0.177	0.126	0.141
ESMC-600M	0.108	0.228	0.666	0.717	0.557	0.621	0.085	0.187	0.131	0.141

Table 58: ViroHub autoregressive LMs with mutant-only hierarchical phenotype–backbone means; AUC/MCC use 82 assays (45 cells), and the other metrics use 89 assays (52 cells).

Model	Spearman		NDCG		AUC		MCC		Top-recall	
	Raw	RH	Raw	RH	Raw	RH	Raw	RH	Raw	RH
ProGen2-B	0.076	0.175	0.674	0.716	0.543	0.596	0.064	0.141	0.123	0.141
ProGen2-L	0.064	0.173	0.675	0.724	0.537	0.594	0.049	0.139	0.124	0.144
ProGen2-M	0.117	0.212	0.688	0.730	0.562	0.612	0.090	0.164	0.128	0.149
ProGen2-S	0.045	0.171	0.648	0.711	0.528	0.593	0.039	0.140	0.113	0.137
ProGen2-XL	0.162	0.235	0.715	0.744	0.588	0.627	0.128	0.190	0.146	0.153
ProGen3-112M	0.034	0.182	0.636	0.692	0.521	0.597	0.035	0.148	0.099	0.115
ProGen3-1B	0.092	0.207	0.667	0.712	0.549	0.609	0.077	0.162	0.121	0.136
ProGen3-219M	0.033	0.181	0.634	0.695	0.521	0.597	0.034	0.144	0.099	0.118
ProGen3-339M	0.058	0.191	0.647	0.698	0.533	0.601	0.056	0.153	0.105	0.123
ProGen3-3B	0.101	0.211	0.678	0.723	0.552	0.611	0.077	0.167	0.123	0.139
ProGen3-762M	0.062	0.192	0.652	0.705	0.534	0.602	0.052	0.151	0.106	0.127
RITA-L	0.166	0.247	0.733	0.751	0.586	0.630	0.124	0.191	0.142	0.154
RITA-M	0.155	0.237	0.726	0.749	0.581	0.626	0.114	0.183	0.143	0.156
RITA-S	0.148	0.236	0.721	0.748	0.579	0.625	0.113	0.186	0.138	0.156
RITA-XL	0.172	0.247	0.736	0.754	0.590	0.631	0.130	0.194	0.146	0.156

J USAGE OF REM-HARNESS

The `vrh` package exposes REM-HARNESS through a command-line interface and a local browser-based dashboard. The workflow below summarizes the repository README: prepare inputs, choose a backbone, inspect ranked predictions, and export selected variants. Figures 16–18 reproduce the three README screenshots of the dashboard preview.

Installation and launch. In an environment with a compatible PyTorch installation, run the following commands from the repository root:

```
pip install -e .
vrh doctor
vrh dashboard
```

The command-line interface is invoked with `vrh`; the `prosst` extra supplies dependencies for VENUSREM2. The dependency check reports the installed backbone dependencies and cache status. Install any optional dependencies required by the chosen backbone; the README lists these extras. The dashboard opens locally at `http://127.0.0.1:8765`. The optional `vrh demo` command tests the installation on the bundled example.

1. Prepare inputs and start a prediction. Open **New prediction** (Figure 16). Upload the sequence and any required structure, or enter a UniProt/PDB identifier and use **Fetch**. Supply an MSA when available. The readiness checklist records which inputs are present. Select the model family and configuration, review the scoring scope and recipe, and click **Start scoring**. The **Demo** button fills the form with the bundled example for an initial walkthrough.

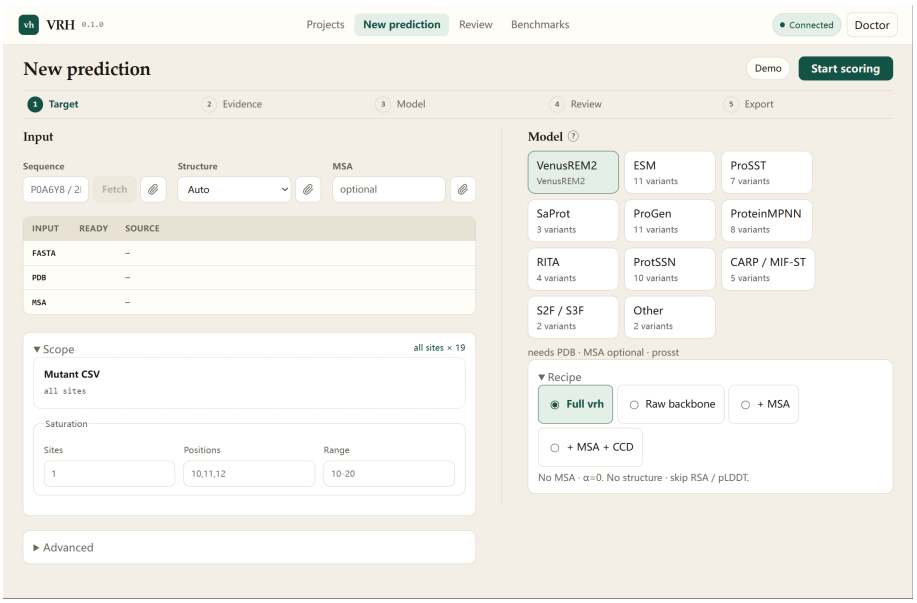


Figure 16: **New prediction.** The README dashboard preview places sequence, structure, and MSA inputs on the left and model selection on the right. The readiness checklist and **Start scoring** button guide submission of a prediction.

2. Inspect predictions and export a shortlist. Open **Review** after scoring (Figure 17). Search or filter the ranked table and select a variant to inspect its position in the structure, its rank, and the available residue evidence. Tick the desired rows and use **Add to CSV** to export the shortlist. Scores express relative model preference, not $\Delta\Delta G$. Experimental PDB B-factors are not pLDDT; confidence coloring requires a predicted structure with confidence values.

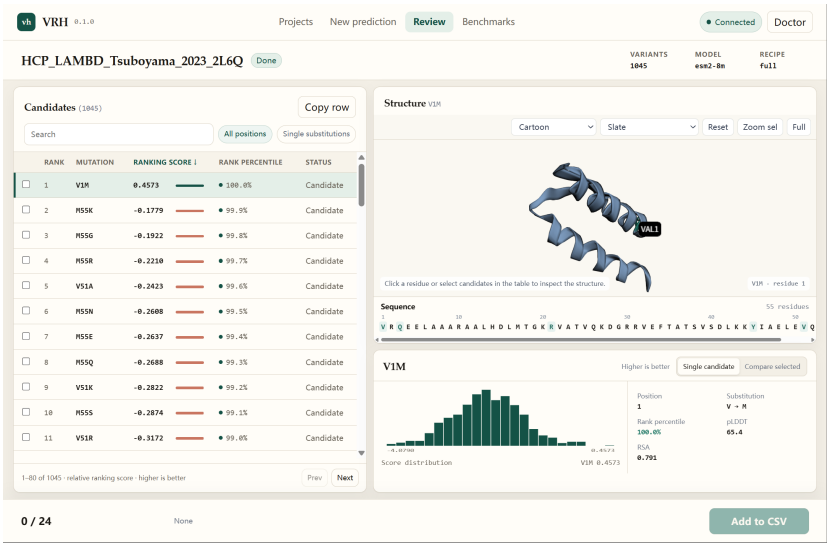


Figure 17: **Review and export.** The ranked candidate table is linked to the structure viewer and residue-level evidence. Selected rows can be exported using **Add to CSV**. Screenshot reproduced from the README preview.

3. Browse benchmark comparisons. Open **Benchmarks** (Figure 18), choose a dataset, and filter by metric, functional property, or input type. Switch between **Raw** and the harness-adjusted view to inspect corresponding model results. The available datasets and model entries depend on the benchmark tables installed with the dashboard.

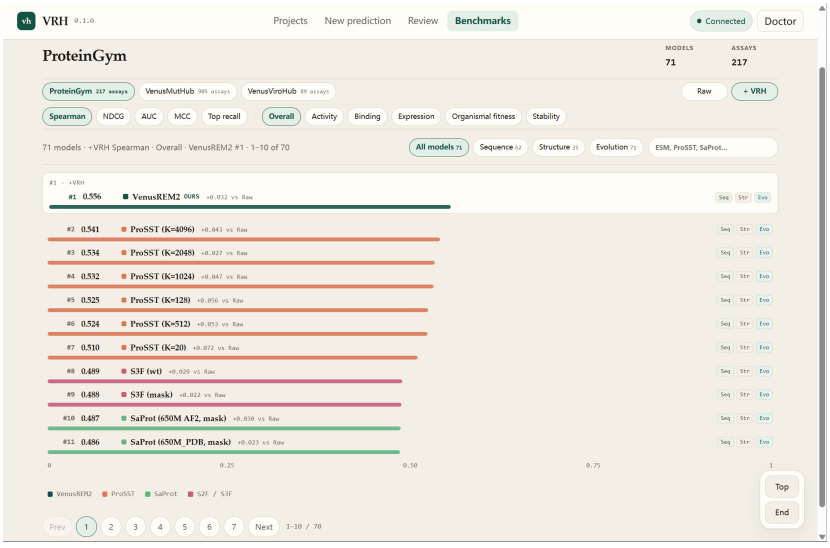


Figure 18: **Benchmark browsing.** Dataset, metric, property, and input-type filters support inspection of Raw and harness-adjusted scores. The README screenshot shows the 71-configuration ProteinGym comparison.