CORRUPTED INPUTS MAKE DENSE SUPERVISION INFORMATIVE FOR ZERO-SHOT VARIANT SCORING: A CONTROLLED FACTORIAL STUDY OF COMPACT PROTEIN ENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Masked language modeling, the default pretraining objective for protein encoders, fixes two separable choices: the input (a sequence with mask tokens, or a complete sequence with some residues replaced by substitutes) and the positions at which the loss is computed (selected or all positions). We cross them with architecture, corpus, optimizer and training length fixed, train compact encoders (7.5M and 33.5M parameters) from scratch on 300k Swiss-Prot sequences, and evaluate zero-shot variant effect prediction on 201 ProteinGym substitution assays. The factors interact: over three seeds at 7.5M, computing the loss at all positions increases mean Spearman by $+0.0095$ with masked input and by $+0.0830$ with corrupted input. With masked input almost every visible residue is correct, and the added loss is minimized by copying. Most of the aggregate interaction arises in assays mixing substitution counts, where the count alone predicts fitness; among variants with the same number the interaction is smaller ($+0.0194$) but positive in every seed, and it remains ($+0.0178$) when the corruption rate is matched and no pretrained model is used. In the matched design, with loss at all positions, corrupted input outperforms masked input by $+0.0890$. Under a supervised probe of frozen embeddings, the interaction is not detected in the original design and small in the matched one, so the conclusions concern zero-shot scoring, not learned representations. Sampling substitutes from BLOSUM62 recovers most of the improvement over uniform sampling that a frozen ESM-2 650M teacher gives, and an entropy-preserving permutation of the matrix recovers little.

1 INTRODUCTION

Masked language modeling (MLM) is the default pretraining objective for protein encoders, from ESM-1b (Rives et al., 2021) and ESM-2 (Lin et al., 2023) to ProtBert (Elnaggar et al., 2022). The recipe, inherited from BERT (Devlin et al., 2019), selects 15% of positions, replaces most of them with a `[MASK]` token, and computes the loss only at the selected positions. It fixes two choices that can be made independently: the input, which is either a sequence with mask tokens or a complete sequence in which some residues are replaced by substitutes, and the positions at which the loss is computed, which are either the selected positions or all positions. We ask whether the two choices have separate effects on zero-shot variant scoring or only a joint one.

The scoring rule matters in practice. Wild-type-marginal scoring evaluates every single and multiple substitution of a protein from one forward pass, whereas masked marginals needs one pass per mutated position and scoring each mutant sequence needs one pass per variant; the assays used here contain $12{,}010$ variants on average. A density score from a model of evolutionary sequences also serves as a feature, together with site-specific amino-acid features, in supervised fitness models (Hsu et al., 2022). For compact encoders used for large-scale screening, the quantity of interest is therefore how well one forward pass ranks variants, and this depends on the pretraining objective.

We cross both in a factorial design. Encoders with the ESM-2 architecture (7.5M and 33.5M parameters) are trained from random initialization on 300k Swiss-Prot sequences (The UniProt Consortium,

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.

2025) with the same optimizer, schedule and number of steps, and are evaluated by zero-shot variant effect prediction on 201 of the 217 ProteinGym substitution assays (Notin et al., 2023), all assays within ESM-2's 1,022-residue limit, with one forward pass per protein. The two choices interact: over three seeds at 7.5M, computing the loss at all positions increases mean Spearman correlation by $+0.0095$ with masked input and by $+0.0830$ with corrupted input. Most of this interaction arises in assays that contain variants with different numbers of substitutions; among variants with the same number it is $+0.0194$. With masked input about 98.3% of visible residues are correct, and the all-position loss reaches near zero, consistent with copying. With corrupted input copying cannot drive this loss to zero, because a visible residue may be a substitute. Under a supervised probe on frozen embeddings the interaction is not detected in the original design and is small in the matched design (§3.2), so our conclusions concern zero-shot scoring with the model's output distribution.

Relation to prior objectives. The closest precedent is the efficiency analysis of ELECTRA (Clark et al., 2020, §3.5 and Table 5), which compares masked language modeling with Replace MLM, in which masked positions are filled by generator samples, and All-Tokens MLM, which also predicts every input token; the all-token variant closes most of the gap to ELECTRA, and the authors note that without replacements the model would copy the input at the unmasked positions. COCO-LM (Meng et al., 2021) recovers original tokens from a corrupted sequence with a full-vocabulary head and a copy mechanism, and ProteinBERT (Brandes et al., 2022) and EvoDiff (Alamdari et al., 2023) use substitution corruption for proteins. We do not propose a new objective and do not claim state-of-the-art results. This paper trains all four combinations of input type and supervision density for protein encoders, including the masked all-position configuration that ELECTRA expected to reduce to copying, with the corruption rate and teacher access matched in a second design; it examines the consequence for zero-shot variant scoring, where the output distribution trained by the objective is itself the scoring function; and it decomposes what a frozen teacher contributes as a corruption sampler. All comparisons are between our own configurations, trained for the same number of steps and tokens; released ESM-2 checkpoints, trained on roughly two orders of magnitude more sequences, score higher.

Contributions.

1. A controlled factorial of input type and supervision density (three seeds at 7.5M), with a rate-matched design and a fully matched design in which no teacher information reaches any configuration (§3). The conclusions concern zero-shot single-pass scoring.
2. Measurements of the loss at unaltered positions: with masked input it reaches near zero; without it, and without soft targets, a corrupted-input model scores the substitute above the wild type at almost every single substitution (§3, §3.1).
3. A decomposition of a frozen ESM-2 650M teacher's contribution: BLOSUM62 accounts for most of the difference between teacher and uniform sampling, a permutation control shows that the substitution pattern is required, and partial teacher budgets recover most of the teacher's improvement (single seed; §4).
4. An evaluation finding: in the 69 assays with multi-substitution variants, ranking by the number of substitutions alone gives mean Spearman $+0.3780$, so aggregate Spearman rewards any additive score that decreases with the number of substitutions; we report within-order results for this reason (§3.2).

2 METHOD

The two factors. Figure 1 shows the four combinations and Table 1 lists every configuration used in the main text. Let $x = (x_1, \ldots, x_L)$ be a sequence over the 20 canonical amino acids $\mathcal{A}$, $\tilde{x}$ the input to the encoder, $s_\theta(\cdot \mid \tilde{x}, i)$ the encoder's distribution over $\mathcal{A}$ at position i, and $\ell_i = -\log s_\theta(x_i \mid \tilde{x}, i)$ the cross-entropy against the true residue. The *input type* is masked or corrupted. Masked inputs follow BERT as implemented for ESM-2: 15% of positions are selected, and of these 80% are replaced by `[MASK]`, 10% by a random residue other than the true one, and 10% are left unchanged. Corrupted inputs select a set S of pL positions and replace each with a substitute drawn from a corruption source r_i renormalized over $\mathcal{A} \setminus \{x_i\}$; the input contains no `[MASK]` token and does not mark which positions were replaced. We call positions outside the

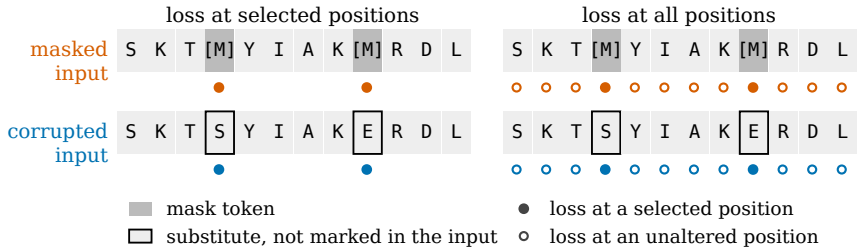


Figure 1: The two factors on a toy sequence with two selected positions. Masked inputs replace selected residues by a mask token; corrupted inputs replace them by substitutes that the input does not mark. Filled dots mark the loss at selected positions and open dots the loss at unaltered positions. Cells: masked input with loss at selected or all positions, MASK and MASK-DENSE; corrupted input, CORRUPT-T-SPARSE or CORRUPT-B-SPARSE-15 (selected) and CORRUPT-T or CORRUPT-B-15 (all). In the experiments 15% or 30% of positions are selected, and masked inputs follow the 80/10/10 rule.

selected set *unaltered positions*. The *supervision density* is sparse or dense: sparse configurations average ℓ_i over the selected positions, dense configurations over all L positions. To vary the weight on unaltered positions continuously we also use a split loss with each group normalized separately,

$$\mathcal{L}_{\text{split}}(\alpha) = \tfrac{1}{|S|}\textstyle\sum_{i\in S}\ell_i \;+\; \alpha\,\tfrac{1}{L-|S|}\sum_{i\notin S}\ell_i, \tag{1}$$

whose case $\alpha = 0$ is the sparse loss (it gives exactly zero gradient at unaltered positions). With $|\mathcal{A}| = 20$, a full output distribution at every position adds little computation.

Loss normalization. The dense loss is not $\mathcal{L}_{\text{split}}(1)$: averaging over all L positions gives $\frac{1}{L}\sum_i \ell_i = p\,\mathcal{L}_{\text{split}}\big((1-p)/p\big)$, the split loss at $\alpha = (1-p)/p$ multiplied by p. AdamW is largely invariant to this global factor, but gradient clipping at max-norm 1.0 is not, so sparse and dense configurations also differ slightly in effective step size. The split-loss configurations separate the two contributions (seed 1337): CORRUPT-T-SPARSE ($\alpha = 0$) scores 0.1718, adding loss at unaltered positions at $\alpha = 0.1$ gives 0.2223 ($p = 7.2 \times 10^{-20}$), and averaging over all positions gives 0.2546. Loss at unaltered positions at one tenth of the weight of the corrupted term thus recovers 61% of the density effect.

Corruption sources and the teacher. The *teacher* source sets $r_i = q_i$, the masked conditional distribution of a frozen ESM-2 650M (Lin et al., 2023) at position i, which depends on context. It is precomputed with four rounds of disjoint 15% masking, covering 60% of positions (Appendix C); teacher-sampled positions are drawn from the covered set, and the teacher is not used at inference. The *BLOSUM62* source sets $r_i \propto \exp(\mathrm{B62}(x_i, \cdot))$, a row of a static substitution matrix (Henikoff & Henikoff, 1992). The *uniform* source has no substitution preferences. The teacher can also provide soft targets: at covered positions within the loss, $\ell_i \leftarrow (1-\lambda)\,\ell_i + \lambda\,\mathrm{KL}\big(q_i \,\|\, s_\theta(\cdot \mid \tilde{x}, i)\big)$. BLOSUM62 corruption with $\lambda = 0$ uses no pretrained model in training; we verified this by pointing the teacher cache at a nonexistent path for those runs.

Designs. In the original design the rows differ in more than input type: the masked row selects 15% of positions and receives no teacher signal ($\lambda = 0$); the corrupted row selects 30%, samples substitutes from the teacher and uses $\lambda = 0.5$, so the dense corrupted cell receives soft targets at every covered position (60% of positions) and the sparse one at every selected position. The *rate-matched* design sets $p = 0.15$ in the corrupted row. The *fully matched* design also samples from BLOSUM62 with $\lambda = 0$, so both rows select 15% of positions and no teacher information reaches any cell. Configurations use $p = 0.30$ and $\lambda = 0.5$ unless a design specifies otherwise (Table 1); these defaults were fixed before any evaluation.

Scoring. A variant v, a set of substitutions $a \to b$ at positions i, is scored from one forward pass over the wild type x (wild-type marginals; Meier et al., 2021):

$$\mathrm{score}(v) = \textstyle\sum_{(i,\,a\to b)\in v}\big[\log s_\theta(b \mid x, i) - \log s_\theta(a \mid x, i)\big]. \tag{2}$$

Table 1: Configurations used in the main text. p: fraction of positions selected; loss: positions averaged in the loss; λ: soft-target weight, applied only at teacher-covered positions within the loss. Masked inputs use the 80/10/10 rule. All configurations are 7.5M and trained for 25,000 steps with the same optimizer and schedule; the cells of the original and fully matched designs are repeated at 33.5M with one seed. Seeds and further configurations are listed in Appendix A.

configuration	input	p	sampler	loss	λ	used in
MASK	masked	0.15	—	selected	0	all designs
MASK-DENSE	masked	0.15	—	all L	0	all designs
CORRUPT-T-SPARSE	corrupted	0.30	teacher	selected	0.5	original design
CORRUPT-T	corrupted	0.30	teacher	all L	0.5	original design
CORRUPT-T-SPARSE-15	corrupted	0.15	teacher	selected	0.5	rate-matched design
CORRUPT-T-15	corrupted	0.15	teacher	all L	0.5	rate-matched design
CORRUPT-B-SPARSE-15	corrupted	0.15	BLOSUM62	selected	0	fully matched design
CORRUPT-B-15	corrupted	0.15	BLOSUM62	all L	0	fully matched design
CORRUPT-B-SPARSE-15, $\alpha = 1$	corrupted	0.15	BLOSUM62	split, $\alpha = 1$	0	fully matched design
MASK+KD	masked	0.15	—	selected	0.5	soft targets on masked input
CORRUPT-U	corrupted	0.30	uniform	all L	0.5	teacher decomposition
CORRUPT-B$_{\lambda=0.5}$	corrupted	0.30	BLOSUM62	all L	0.5	teacher decomposition
CORRUPT-B-SHUF	corrupted	0.30	permuted BLOSUM62	all L	0.5	permutation control
CORRUPT-T$_{\lambda=0}$	corrupted	0.30	teacher	all L	0	budget 100%
CORRUPT-B	corrupted	0.30	BLOSUM62	all L	0	budget 0%
HYB-SEQγ	corrupted	0.30	teacher/BLOSUM62 by seq.	all L	0	partial budgets
RTD-BIN	corrupted	0.30	teacher	all L, binary	0	output head

Multi-substitution variants make up 73% of the 2,413,913 rows. The 16 excluded assays are those whose wild type exceeds 1,022 residues (Appendix F). Let C and M denote the corrupted and masked rows and $\bar{\rho}$ the mean over assays of the per-assay Spearman correlation; the interaction, which we test on its per-assay value, is

$$I = (\bar{\rho}_{C,\text{dense}} - \bar{\rho}_{C,\text{sparse}}) - (\bar{\rho}_{M,\text{dense}} - \bar{\rho}_{M,\text{sparse}}), \tag{3}$$

Statistics. We use two-sided Wilcoxon signed-rank tests over paired per-assay values, which test the location of the paired differences; we report mean differences alongside. These tests measure consistency across assays, not variation across training runs, which the three seeds (1337, 2024, 777) address; with three seeds, tests use per-assay seed means, and single-seed results use seed 1337. The primary family consists of the original all-variant interaction, the original and matched within-order interactions (§3.2), and the matched dense-column input effect; with Holm adjustment all four have $p \leq 4.7 \times 10^{-6}$. Averaging per-assay values within each of the 173 proteins before testing leaves all four significant (largest $p = 9.2 \times 10^{-6}$). Other tests are secondary and not corrected for multiplicity. For the fully matched design, the permutation control and the substitution-order analysis, decision rules were fixed before the corresponding runs; all outcomes are reported. All runs use 25,000 steps except the 75,000-step runs in Appendix N.

3 INTERACTION BETWEEN INPUT TYPE AND SUPERVISION DENSITY

Original design. Dense supervision increases mean Spearman by $+0.0095$ with masked input ($p = 0.02$; significant at two of three seeds) and by $+0.0830$ with corrupted input ($p = 2.2 \times 10^{-22}$), an interaction of $+0.0734 \pm 0.0057$ ($p = 1.3 \times 10^{-16}$; three seeds, Table 2). At 33.5M (one seed) all contrasts have the same sign, with an interaction of $+0.0571$. Two scales constitute a replication, not a trend.

Matched corruption rate. The rows of the original design differ in selection rate. At $p = 0.15$ with the teacher (seed 1337), dense supervision increases mean Spearman by $+0.0950$ with corrupted input and by $+0.0028$ with masked input, an interaction of $+0.0922$ ($p = 3.5 \times 10^{-21}$), so the difference in rate does not account for the interaction; on the same-k metric it is $+0.0419$. With the rate and the sparse loss matched, three seeds of CORRUPT-T-SPARSE-15 score $+0.0149$ to $+0.0188$ above MASK+KD (seed 1337; $p \leq 0.011$ for each). Teacher access is not matched in this comparison: every supervised position of CORRUPT-T-SPARSE-15 carries a soft target (fraction without one 0.000), whereas MASK+KD selects positions independently of the 0.600 coverage of the teacher cache.

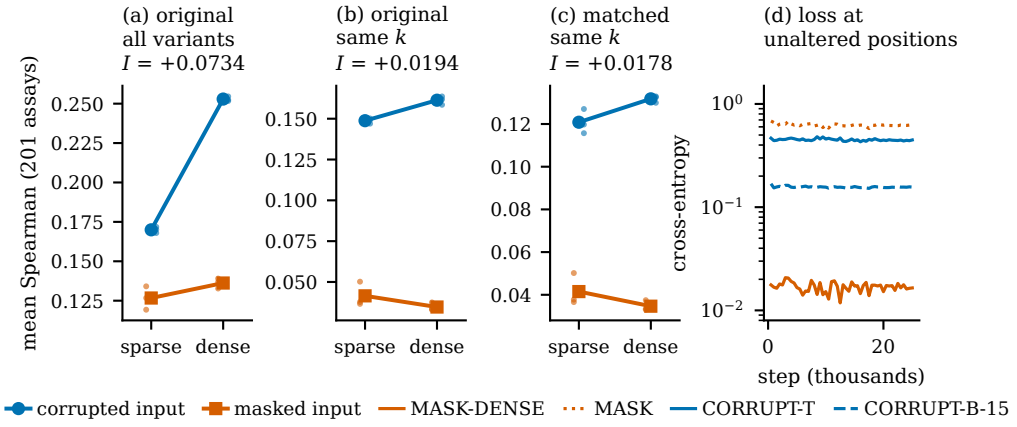


Figure 2: (a) The original design on all variants; (b) the original design and (c) the fully matched design on the same-k metric (Eq. 4). Lines connect three-seed means of the mean per-assay Spearman correlation; small points are individual seeds. The difference between the two slopes is the interaction (Eq. 3; values in Table 2). (d) Training-split cross-entropy at unaltered positions, seed 1337; CORRUPT-B-15 ($p = 0.15$, $\lambda = 0$) is included because CORRUPT-T also has soft targets.

Table 2: Density effects and interaction. Δ_{mask}, Δ_{corr}: dense minus sparse with masked and with corrupted input; I: interaction (Eq. 3) with its Wilcoxon p. All variants: per-assay Spearman over every variant. Same k: order-stratified Spearman (Eq. 4), which ranks only variants with the same number of substitutions. Three-seed rows give seed means ($\pm$ s.d. over seeds for I) and tests on per-assay seed means; cell means are in Appendix G. The sparse corrupted cell of the fully matched design fails on multi-substitution variants (see text).

	all variants			same k		
design	Δ_{mask}	Δ_{corr}	I (p)	Δ_{mask}	Δ_{corr}	I (p)
original, 7.5M (3 seeds)	+0.0095	+0.0830	+0.0734 (1×10^{-16})	−0.0068	+0.0126	+0.0194 ± 0.0094 (2×10^{-6})
original, 33.5M (1 seed)	+0.0204	+0.0774	+0.0571 (1×10^{-12})	+0.0032	+0.0140	+0.0108 (8×10^{-4})
rate-matched, 7.5M (1 seed)	+0.0028	+0.0950	+0.0922 (3×10^{-21})	−0.0167	+0.0252	+0.0419 (1×10^{-11})
fully matched, 7.5M (3 seeds)	+0.0095	+0.2002	+0.1907 (1×10^{-15})	−0.0068	+0.0110	+0.0178 ± 0.0017 (5×10^{-6})
fully matched, 33.5M (1 seed)	+0.0204	+0.2130	+0.1927 (5×10^{-21})	+0.0032	+0.0273	+0.0241 (2×10^{-8})

Matched rate and teacher access. The fully matched design samples substitutes from BLOSUM62 with $\lambda = 0$ at $p = 0.15$, three seeds per cell. On all variants the interaction is +0.1907, but its magnitude is dominated by the failure of the sparse corrupted configuration on multi-substitution variants, so we compare designs on the same-k metric. CORRUPT-B-SPARSE-15 scores the substitute above the wild type at 99.5% of single substitutions (seed 1337), so the summed score of a variant increases with its number of substitutions (mean per-assay Spearman correlation between score and k of +0.6716, against −0.6761 for MASK). One explanation consistent with this is that every supervised position of CORRUPT-B-SPARSE-15 holds a substitute, so without loss at unaltered positions or soft targets the model is trained to assign low probability to the residue it sees; we have not tested this. On the same-k metric the matched interaction is $+0.0178 \pm 0.0017$ ($p = 4.7 \times 10^{-6}$), made up of a small decrease with masked input (−0.0068) and an increase with corrupted input (+0.0110). In the dense column of this design, where rate and teacher access are matched, corrupted input scores +0.0890 above masked input ($p = 2.0 \times 10^{-22}$; same k: +0.0971), positive at

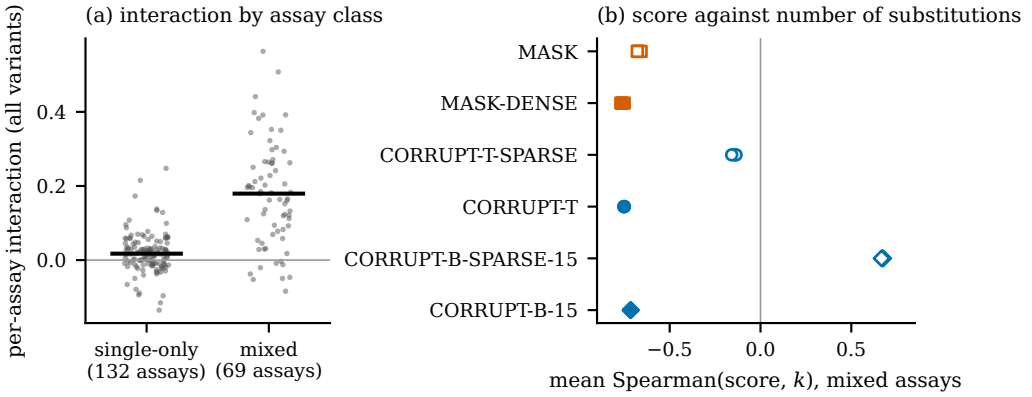


Figure 3: Substitution order. (a) Per-assay interaction on all variants (original design, per-assay seed means) for assays with only single substitutions and for assays that also contain multi-substitution variants; bars are medians. (b) Mean over mixed assays of the per-assay Spearman correlation between a variant's summed score and its number of substitutions k, one point per seed; open markers are sparse, filled markers dense configurations. Ranking by $-k$ alone gives a mean Spearman correlation of $+0.3780$ with fitness in these assays.

every seed. Adding loss at unaltered positions at $\alpha = 1$ to CORRUPT-B-SPARSE-15 raises its mean Spearman from $+0.0304$ to $+0.2185$, close to uniform averaging ($+0.2239$; seed 1337).

3.1 LOSS AT UNALTERED POSITIONS DURING TRAINING

With masked input about 98.3% of visible residues are correct (the random replacements make up 1.7%). MASK-DENSE reaches a cross-entropy at unaltered positions of 0.0177 by step 500 and 0.0165 at 25,000 steps (Figure 2d), consistent with copying; at 33.5M the value is 0.0163. With corrupted input, copying cannot drive the all-position loss to zero. On the validation split we compare each corrupted configuration with the best context-free predictor for its corruption process, which predicts the true residue from the visible residue alone (Appendix H). CORRUPT-B-15 ($p = 0.15$, $\lambda = 0$) has lower cross-entropy than this predictor at unaltered positions (0.1547 against 0.1596) and over all positions (0.6650 against 0.6978), so it uses the surrounding sequence even at residues that were not altered. CORRUPT-T is above its predictor at unaltered positions (0.4431 against 0.3549) and below it overall; CORRUPT-B, with the same p but a BLOSUM62 sampler and no soft targets, is below its predictor at unaltered positions (0.3301 against 0.3453), so the higher unaltered-position loss of CORRUPT-T goes with its soft targets or its sampler. These measurements describe the loss, not the computation, and that models trained with the masked all-position loss improve little downstream is consistent with, but not shown to be caused by, its being minimized by copying.

3.2 SUBSTITUTION ORDER, A SUPERVISED PROBE, AND MASKED MARGINALS

Substitution order. Eq. 2 sums per-position log-ratios and cannot represent epistasis. 69 of the 201 assays contain multi-substitution variants, and in these the number of substitutions k alone predicts fitness: scoring by $-k$ gives mean Spearman $+0.3780$. A model whose summed score decreases with k therefore scores well on the all-variant metric independently of how it ranks variants with the same k. These assays contribute 82% of the sum of per-assay interactions in the original design (three seeds; Figure 3a, Appendix L). The configuration that differs is CORRUPT-T-SPARSE, whose score barely decreases with k (mean correlation with k of -0.1423, against -0.7498 for CORRUPT-T; seed 1337). We therefore also use the order-stratified Spearman correlation of an assay,

$$\rho_{\text{strat}} = \textstyle\sum_k n_k \rho_k \big/ \sum_k n_k, \tag{4}$$

where ρ_k is the Spearman correlation among the n_k variants with k substitutions, over groups with $n_k \geq 20$. On this metric the interaction is $+0.0194 \pm 0.0094$ ($p = 2.1 \times 10^{-6}$), positive at every seed, which meets the criterion fixed before the analysis. CORRUPT-T − MASK is $+0.1240$ on single substitutions ($p = 1.6 \times 10^{-26}$) and $+0.0213$ on multi-substitution variants ($p = 0.30$; three seeds).

Supervised probe. Wild-type marginals scores a residue present in the input, which is what corrupted-input training optimizes, so we also fit a per-assay kernel ridge probe on frozen final-layer embeddings without the output head (seed 1337; Appendix K). CORRUPT-T scores $+0.5204$ and MASK $+0.5067$ ($p = 1.6 \times 10^{-11}$). In the original design the interaction is not detected ($+0.0028$, $p = 0.34$); its 95% bootstrap interval over assays is $[-0.0027, +0.0085]$, so an interaction above $+0.0085$ in the frozen representations is not supported by these data. In the matched design it is small ($+0.0058$; interval $[+0.0005, +0.0109]$). The probe compresses differences between encoders: released ESM-2 8M exceeds MASK by $+0.0386$ under the probe and by $+0.1237$ zero-shot.

Masked marginals. Under masked marginals (L forward passes) no 25,000-step configuration exceeds a mean Spearman of 0.082; the 7.5M interaction remains ($+0.0833$) and the 33.5M interaction reverses (-0.0113). Corrupted configurations never see `[MASK]` in training, so this protocol is out of distribution for them. MASK also scores lower under masked than under wild-type marginals ($+0.0670$ against $+0.1341$); our scoring code reproduces ProteinGym's published ESM-2 8M values at $r = 1.0000$, and we do not have an explanation for this result (Appendix M).

4 DECOMPOSITION OF THE TEACHER'S CONTRIBUTION

Sampler and soft targets. Replacing the teacher as the sampler with uniform substitution (CORRUPT-T −CORRUPT-U, seed 1337) reduces mean Spearman by 0.0921 ($p = 2.5\times10^{-31}$), and removing its soft targets (CORRUPT-T −CORRUPT-T$_{\lambda=0}$) by 0.0138 ($p = 1.6\times10^{-11}$). The ratio between these compares plausible with uniform corruption, not teacher with no teacher. Of the sampler difference, BLOSUM62 accounts for $+0.0808$ (CORRUPT-B$_{\lambda=0.5}$ −CORRUPT-U), a fraction of 0.88. The teacher over BLOSUM62 adds $+0.0113$ (CORRUPT-T −CORRUPT-B$_{\lambda=0.5}$, $p = 2 \times 10^{-4}$); the two samplers differ in context dependence and in the distribution of substitutes. Removing the teacher from both roles reduces mean Spearman by 0.0345 (CORRUPT-T −CORRUPT-B, $p = 1.3\times10^{-12}$); CORRUPT-B, which uses no pretrained model, scores 0.0621 above MASK+KD ($p = 7.4\times10^{-16}$).

Substitution pattern and distribution shape. Uniform and BLOSUM62 corruption differ in entropy and peakedness as well as in which substitutions they favour. A row-permuted BLOSUM62 (CORRUPT-B-SHUF) permutes the probabilities within each row across residues, which preserves each row's entropy and removes the substitution pattern (Appendix I). Over three seeds, of the difference between uniform corruption and BLOSUM62 ($+0.0813$), the permuted matrix reproduces $+0.0098$ and $+0.0714$ ($p = 7.4 \times 10^{-30}$) requires the substitution pattern: a share $\beta = 0.88$, a different quantity from the fraction 0.88 above. By a rule recorded before these runs, a share of at least 0.6 keeps the label substitution plausibility for this term. We define the *hardness* of a source as the mean $-\log q_i(\tilde{x}_i)$ of its substitutes under the teacher's conditional: teacher 1.813, BLOSUM62 3.099, permuted matrix 4.554, uniform 4.751. Permutation keeps the entropy but makes the substitutes almost as implausible to the teacher as uniform ones, which is consistent with the term reflecting the plausibility of the substitutes as judged by a protein language model.

4.1 PARTIAL TEACHER BUDGETS

With soft targets disabled ($\lambda = 0$), CORRUPT-B (0% teacher, $+0.2201$) and CORRUPT-T$_{\lambda=0}$ (100%, $+0.2408$) differ by $+0.0206$. Teacher precomputation accounts for 72% of the end-to-end compute for a single 7.5M model. Sequence-level hybrids (HYB-SEQ) sample the substitutes of a random fraction γ of training sequences from the teacher and the rest from BLOSUM62, so precomputation scales with γ (Appendix J, Figure 4). At $\gamma = 0.25$ the hybrid recovers 0.67 of the full-teacher improvement, and at $\gamma = 0.5$ to 0.9 between 0.86 and 0.90; the 50% and 75% hybrids do not differ significantly from the full teacher ($p = 0.41$, 0.17). Each point is a single seed. Sharpening

BLOSUM62 by a temperature does not reach the teacher's hardness (minimum 2.348), and at the closest temperature (TEMPCAL, $T = 0.05$) the model scores 0.0237 below unmodified BLOSUM62 ($p = 1.6\times10^{-4}$).

4.2 BINARY REPLACED-TOKEN DETECTION HEAD

RTD-BIN replaces the 20-way output with a per-position binary classifier (corrupted or not), as in ELECTRA (Clark et al., 2020); it also differs from CORRUPT-T in λ (0 against 0.5) and in its scoring rule, $-\log P(\text{corrupt} \mid i)$ on the mutant sequence, which requires one forward pass per variant. Its mean Spearman is -0.1938, negative on 195 of 201 assays; reversing the sign gives $+0.1938$, above MASK ($+0.1341$), so its output carries information about fitness with the opposite sign. Released checkpoints are not compute-matched: CORRUPT-T at 7.5M does not differ significantly from ESM-2 8M under wild-type marginals ($p = 0.99$), and our 33.5M configuration scores below ESM-2 35M (Appendix P).

5 DISCUSSION

Interpretation of the factorial results. Loss at unaltered positions trains the probability the model assigns to the residue it sees. With masked input almost every visible residue is correct, and this loss reaches near zero, consistent with copying. With corrupted input a visible residue may be a substitute, so copying cannot drive the loss to zero, and the probability assigned to the visible residue is the quantity that wild-type-marginal scoring reads. Two results are consistent with this reading. Without loss at unaltered positions and without soft targets, the corrupted configuration scores the substitute above the wild type at 99.5% of single substitutions, and adding the loss at $\alpha = 1$ raises its mean Spearman from $+0.0304$ to $+0.2185$. Under the probe, dense supervision does not detectably change performance with either input type in the original design. For encoders intended for single-pass zero-shot scoring, this suggests including loss at unaltered positions of an input that may contain substitutes.

Relation to ELECTRA's attribution. ELECTRA attributes most of its improvement over BERT to learning from all tokens and notes that, without replacements, an all-token loss would be solved by copying (Clark et al., 2020). Our factorial measures both sides of this argument for proteins: the masked all-position configuration reaches near-zero loss at unaltered positions and improves zero-shot scoring little, whereas all-position loss on corrupted input improves it substantially. Two results qualify the attribution in this setting. Most of the improvement arises in assays that mix variants with different numbers of substitutions, and a supervised probe does not detect it in the frozen representations. For zero-shot scoring of proteins, the benefit of all-position loss therefore appears mainly in the output distribution that single-pass scoring reads.

Practical implications. Three points follow for compact encoders used for single-pass scoring. When inputs are corrupted by substitution, the loss should include unaltered positions or soft targets; without either, the model scores the substitute above the wild type at 99.5% of single substitutions. A static substitution matrix provides most of the benefit of a teacher as corruption source: CORRUPT-B, which uses no pretrained model, scores $+0.0860$ above MASK, and a teacher used for a quarter of the training sequences recovers 0.67 of the teacher's improvement (one seed). On assays with multi-substitution variants, additive scores should also be compared within substitution order, because the number of substitutions alone reaches a mean Spearman correlation of $+0.3780$ on these assays.

Limitations. (i) Two scales only: the interaction is $+0.0734$ at 7.5M (three seeds) and $+0.0571$ at 33.5M (one seed); we did not run a third scale. (ii) One benchmark family, zero-shot scoring of ProteinGym substitutions, with no structure, function or fine-tuning task. Under a supervised probe, which is less sensitive than zero-shot scoring, the interaction is not detected in the original design and is small in the matched design, so we restrict our conclusions to zero-shot scoring. Most of the all-variant interaction arises in assays that mix variants with different numbers of substitutions, and CORRUPT-T's advantage over MASK is observed on single substitutions. (iii) Undertrained models: our MASK at 7.5M scores $+0.1341$ against $+0.2578$ for released ESM-2 8M. The 75,000-step runs cover only MASK and CORRUPT-T, in which the difference between them persists (Appendix N); whether the interaction persists under longer training is not tested. (iv) Configurations that use

the teacher receive information from a model trained on far more data. (v) The budget points, the 33.5M cells, the rate-matched design and several secondary contrasts use a single seed. (vi) The training corpus was not filtered against the benchmark proteins; the overlap is identical for every configuration, although its effect could differ between objectives.

6 Related Work

Masked language modeling (Devlin et al., 2019) was adopted for proteins with few changes, in ESM-1b (Rives et al., 2021), ESM-2 (Lin et al., 2023), ProtTrans (Elnaggar et al., 2022) and structure-aware variants (Su et al., 2024b). Zero-shot variant scoring with such models follows Meier et al. (2021) and is benchmarked by ProteinGym (Notin et al., 2023), on which alignment-based (Frazer et al., 2021) and MSA-conditioned (Rao et al., 2021) models perform well. In natural language the mask rate has been re-examined (Wettig et al., 2023), and MacBERT (Cui et al., 2020) replaces selected tokens with similar words while keeping the loss at the selected positions. In encoder–decoder pretraining, Raffel et al. (2020, §3.3.2 and Table 5) compare a BERT-style objective, in which 15% of tokens are corrupted (90% of them replaced by a mask token and 10% by a random token) and the full original sequence is reconstructed, with variants that predict only the corrupted tokens (and replace corrupted spans by sentinel tokens or drop them), and find that all variants perform similarly; this agrees with the small density effect we observe with masked input. ELECTRA (Clark et al., 2020) trains a binary discriminator on all tokens of a corrupted input; its efficiency analysis (Table 5 there) shows that All-Tokens MLM, a full-vocabulary loss at every position of a generator-corrupted input with a copy mechanism, closes most of the gap between BERT and ELECTRA. A ProtElectra model was trained within ProtTrans (Elnaggar et al., 2022). COCO-LM (Meng et al., 2021) trains a binary copy task on all tokens and a language-modeling task on the replaced positions of a corrupted sequence. For proteins, ProteinBERT (Brandes et al., 2022) reconstructs sequences corrupted by uniform random substitution at a 5% rate, and EvoDiff (Alamdari et al., 2023), building on D3PM (Austin et al., 2021), compares uniform and BLOSUM62 transition matrices with masking in a generative diffusion model in which the corruption level is varied and the supervision density is not; its finding that masking gives better reconstruction is compatible with ours, because in our experiments the ordering of masked and corrupted inputs depends on supervision density. Our frozen teacher differs from knowledge distillation (Hinton et al., 2015): its larger contribution is as a corruption sampler, a role that does not arise when the transfer set consists of uncorrupted inputs.

7 Conclusion

For zero-shot variant scoring with compact protein encoders, the input type and the supervision density that masked language modeling fixes together have a non-additive effect: over three seeds, loss at all positions increases mean Spearman correlation by $+0.0095$ with masked input and by $+0.0830$ with corrupted input, and with masked input the added loss reaches near zero, consistent with copying. Most of the interaction arises in assays that mix variants with different numbers of substitutions; among variants with the same number it is smaller but positive at every seed, including when the corruption rate is matched and no pretrained model is used. Under a supervised probe of frozen embeddings it is not detected in the original design and is small in the matched design, so our conclusions concern zero-shot scoring. A frozen teacher contributes mainly as a sampler of substitutes, and BLOSUM62 supplies most of this contribution through its substitution pattern. All conclusions are based on one benchmark, and we do not claim state-of-the-art results.

Use of AI assistants

The authors conceived the research question, designed the factorial comparison, chose the training corpus, the benchmark and the zero-shot scoring protocol, and decided which analyses to run and how to interpret them. Some controls and analyses were first proposed during AI-assisted review of intermediate results, including the fully matched teacher-free design, the row-permuted substitution-matrix control, the supervised probe and the substitution-order analysis, and the authors decided whether to adopt each.

Within this design, we used generative AI tools for implementation, execution and drafting: to write and refactor training, evaluation and analysis code; to run the specified experiments; to generate

figures, tables and the registry of reported numbers; to retrieve and check literature; and to draft and revise the manuscript.

We did not use generative AI tools to generate or impute data, to produce numbers that are not computed by the analysis code from stored results, or to decide endpoints, thresholds or interpretations. The authors reviewed all AI-assisted work and approved every design decision, decision rule, interpretation and claim. Every number in the manuscript is generated from stored results by scripts, without manual transcription; cross-checks between derived tables and their sources identified implementation defects, which were corrected before the reported runs. The authors take full responsibility for the content, including all claims, numerical results and code.

Reproducibility statement

Every configuration is specified in Table 1 and Appendix A; corpus construction, teacher precomputation, architecture, optimization, software versions and the evaluation rule are described in Appendices B–F, with pseudocode in Appendix E. The supplementary material contains the code, the configurations, the per-assay results of every reported configuration and a script that regenerates every benchmark number in the paper from them. The study used 46.21 A100 GPU-hours in total.

Ethics statement

This work trains small protein sequence encoders on public data and evaluates them on a public benchmark of measured variant effects. It involves no human subjects or personal data, and we are not aware of misuse risks beyond those of existing, larger protein language models.

References

Sarah Alamdari, Nitya Thakkar, Rianne Van Den Berg, Neil Tenenholtz, Robert Strome, Alan M Moses, Alex X Lu, Nicolò Fusi, Ava P Amini, and Kevin K Yang. Protein generation with evolutionary diffusion: sequence is all you need. *bioRxiv*, 2023.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17981–17993, 2021.

Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*, 2020.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 657–668, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2022.

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.

Steven Henikoff and Jorja G. Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Chloe Hsu, Hunter Nisonoff, Clara Fannjiang, and Jennifer Listgarten. Learning protein fitness models from evolutionary and assay-labeled data. *Nature Biotechnology*, 40(7):1114–1122, 2022.

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

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. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29287–29303, 2021.

Yu Meng, Chenyan Xiong, Payal Bajaj, Saurabh Tiwary, Paul Bennett, Jiawei Han, and Xia Song. COCO-LM: Correcting and contrasting text sequences for language model pretraining. In *Advances in Neural Information Processing Systems*, volume 34, pp. 23102–23114, 2021.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Roshan M. Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. MSA transformer. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8844–8856. PMLR, 2021.

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

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024a.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. SaProt: Protein language modeling with structure-aware vocabulary. In *International Conference on Learning Representations*, 2024b.

The UniProt Consortium. UniProt: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, 2025.

Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. Should you mask 15% in masked language modeling? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2985–3000, 2023.

A CONFIGURATIONS

Table 3 lists every trained configuration referred to in the paper, with its seeds. The per-assay results file identifies each run by an internal identifier.

Table 3: Trained configurations referred to in the paper. p: selection rate; loss: positions averaged in the loss (split: Equation 1); λ: soft-target weight. Masked inputs use the 80/10/10 rule. Seeds are those present in the per-assay results file. Generated from the run registry by `analysis/make_config_table.py`.

configuration	size	input	sampler	p	loss	λ	seeds
MASK	7.5M	masked	—	0.15	selected	0	1337, 2024, 777
MASK-DENSE	7.5M	masked	—	0.15	all L	0	1337, 2024, 777
MASK+KD	7.5M	masked	—	0.15	selected	0.5	1337
CORRUPT-T-SPARSE	7.5M	corrupted	teacher	0.3	split, $\alpha = 0$	0.5	1337, 2024, 777
CORRUPT-T, $\alpha = 0.1$	7.5M	corrupted	teacher	0.3	split, $\alpha = 0.1$	0.5	1337
CORRUPT-T	7.5M	corrupted	teacher	0.3	all L	0.5	1337, 2024, 777
CORRUPT-T-SPARSE-15	7.5M	corrupted	teacher	0.15	split, $\alpha = 0$	0.5	1337, 2024, 777
CORRUPT-T-15	7.5M	corrupted	teacher	0.15	all L	0.5	1337
CORRUPT-T, $p = 0.50$	7.5M	corrupted	teacher	0.5	all L	0.5	1337
CORRUPT-T, $\lambda = 1$	7.5M	corrupted	teacher	0.3	all L	1	1337
CORRUPT-T$_{\lambda=0}$	7.5M	corrupted	teacher	0.3	all L	0	1337
CORRUPT-U	7.5M	corrupted	uniform	0.3	all L	0.5	1337, 2024, 777
CORRUPT-B$_{\lambda=0.5}$	7.5M	corrupted	BLOSUM62	0.3	all L	0.5	1337, 2024, 777
CORRUPT-B-SHUF	7.5M	corrupted	permuted BLOSUM62	0.3	all L	0.5	1337, 2024, 777
CORRUPT-B	7.5M	corrupted	BLOSUM62	0.3	all L	0	1337
CORRUPT-B-15	7.5M	corrupted	BLOSUM62	0.15	all L	0	1337, 2024, 777
CORRUPT-B-SPARSE-15	7.5M	corrupted	BLOSUM62	0.15	split, $\alpha = 0$	0	1337, 2024, 777
CORRUPT-B-SPARSE-15, $\alpha = 1$	7.5M	corrupted	BLOSUM62	0.15	split, $\alpha = 1$	0	1337
TEMPCAL	7.5M	corrupted	BLOSUM62, $T = 0.05$	0.3	all L	0	1337
HYB-SEQ25	7.5M	corrupted	teacher / BLOSUM62 (25% of sequences)	0.3	all L	0	1337
HYB-SEQ50	7.5M	corrupted	teacher / BLOSUM62 (50% of sequences)	0.3	all L	0	1337
HYB-SEQ75	7.5M	corrupted	teacher / BLOSUM62 (75% of sequences)	0.3	all L	0	1337
HYB-SEQ90	7.5M	corrupted	teacher / BLOSUM62 (90% of sequences)	0.3	all L	0	1337
RTD-BIN	7.5M	corrupted	teacher	0.3	all L, binary head	0	1337
MASK-35M	33.5M	masked	—	0.15	selected	0	1337
MASK-DENSE-35M	33.5M	masked	—	0.15	all L	0	1337
CORRUPT-T-SPARSE-35M	33.5M	corrupted	teacher	0.3	split, $\alpha = 0$	0.5	1337
CORRUPT-T-SPARSE-15-35M	33.5M	corrupted	teacher	0.15	split, $\alpha = 0$	0.5	1337
CORRUPT-T-35M	33.5M	corrupted	teacher	0.3	all L	0.5	1337
CORRUPT-B-SPARSE-15-35M	33.5M	corrupted	BLOSUM62	0.15	split, $\alpha = 0$	0	1337
CORRUPT-B-15-35M	33.5M	corrupted	BLOSUM62	0.15	all L	0	1337

Two further sets of runs are not listed as separate configurations: retraining CORRUPT-T and MASK with ESM's `token_dropout` disabled (Appendix N), and separate 75,000-step runs of CORRUPT-T and MASK at both scales with checkpoints at 25k, 40k, 55k and 75k steps.

B CORPUS

Table 4: Corpus construction: sequences retained at each filtering stage.

stage	sequences
parsed from UniProt Swiss-Prot, release 2026_03	575,503
dropped: length > 512	−105,527
dropped: exact duplicate	−80,056
dropped: length < 30	−5,636
dropped: non-canonical residue	−1,800
retained	382,484
training set	300,000
validation set	5,000
unused	77,484

The training set contains 79,593,015 residues (mean length 265.3, median 258). The split is deterministic given the corpus seed. Sequences containing non-canonical residues (X, B, Z, J, U, O) were dropped, so that every position has a valid 20-way target. The length cap is the largest single filter and shifts the corpus toward shorter proteins; Appendix O examines its effect on the comparison with released checkpoints. The corpus was deduplicated but not filtered against the ProteinGym target proteins, so some benchmark proteins are likely present in training. This is identical for every configuration.

C Teacher Precomputation and Store Validation

The teacher is `facebook/esm2_t33_650M_UR50D`, run frozen in bf16 with SDPA attention. Four rounds of disjoint 15% masking, one forward pass each, cover a measured 60.0% of positions: 47,757,848 covered positions in the training split and 793,932 in validation. Distributions are stored as fp16 vectors over all 20 canonical residues (1.80 GB). Precomputation took 17 min 31 s of wall-clock time on four GPUs (1.17 GPU-hours), with a peak of 4.71 GB per device.

Structural checks over every row. Probabilities sum to one within a maximum deviation of 4.3×10^{-4}; no row has duplicate or out-of-range positions, which confirms that the four rounds are disjoint; and all 305,000 sequences are present exactly once, with identifiers and lengths matching the corpus.

Validation checks. The teacher's top-1 prediction matches the true residue at 51.9% of covered positions (chance: 5%). This detects off-by-one indexing errors: ESM's `<cls>` token shifts residue i to token $i{+}1$, and a misalignment would reduce this figure to chance. Compared with single-position masking on a held-out sample, the mean total-variation distance is 0.103 and top-1 agreement is 81.0%, so masking 15% of a sequence at once changes the conditional at a given position only slightly. Single-position masking is used only in this check.

Coverage and fallback. Per-sequence coverage has mean 0.600, first percentile 0.579 and minimum 0.533. When a requested corruption rate exceeds the covered positions in a sequence, all covered positions are used and the remainder are drawn uniformly. The measured fallback fraction is 0.000% at $p = 0.30$ and $p = 0.50$ and 0.313% at $p = 0.60$, so the corruption rates used here are entirely teacher-sampled. The fallback branch was tested by reducing a test sequence to 18 covered positions of 190.

D Architecture, Optimization, and Environment

All encoders are randomly initialized `EsmForMaskedLM` models built from `EsmConfig` with rotary position embeddings (Su et al., 2024a), with parameter counts identical to the released ESM-2 checkpoints.

Table 5: Encoder architectures.

size	layers	hidden	heads	intermediate	parameters
7.5M	6	320	20	1280	7,512,474
33.5M	12	480	20	1920	33,501,394

ESM's unused contact-prediction head (121 parameters) is frozen because distributed data-parallel training fails on unused parameters; it remains in the state dictionary, so the totals match the released checkpoints. RTD-BIN adds a 321-parameter binary head (7,512,795 in total).

Amino-acid index. The 20 canonical amino acids have token ids `[5,23,13,9,18,6,21,12,15,4,20,17,14,16,10,8,11,7,22,19]` in the ESM vocabulary, which are neither contiguous nor monotonic. The 20-way distribution is therefore read with an explicit index; a contiguous slice such as `logits[...,4:24]` would permute the output relative to the stored teacher distributions without raising an error. The ordering, the absence of duplicate ids, and the round trip through the tokenizer are asserted at load time.

Optimization, identical for every trained configuration. AdamW with $\beta = (0.9, 0.98)$ and $\varepsilon = 10^{-8}$; learning rate 4×10^{-4}; linear warmup followed by cosine decay to zero, with warmup equal to 6% of the total steps (1,500 for 25,000-step runs and 4,500 for 75,000-step runs); weight decay 0.01, excluding one-dimensional parameters and biases; gradient clipping at max-norm 1.0; bf16 mixed precision; length-bucketed batches of at most 32,768 padded tokens and 128 sequences. A 25,000-step run is about 8.6 epochs (2,897 steps per epoch). Seed 1337 is the primary seed;

2024 and 777 are replicates for the configurations in Appendix G. Training is not bit-reproducible: repeated runs with the same seed and configuration differ by about 10^{-4} in mean Spearman because of nondeterministic CUDA kernels.

Learning-rate scheduler. With `accelerate`, a scheduler passed to `accelerator.prepare()` advances once per process at each step. The scheduler is therefore stepped manually and not passed to `prepare()`.

Environment. Four NVIDIA A100-SXM4-80GB GPUs (CUDA 13.1), Linux 5.15. Python 3.11.8; `torch` 2.7.1; `transformers` 5.9.0; `accelerate` 1.13.0; `numpy` 1.26.4; `pandas` 3.0.3; `pyarrow` 15.0.0; `scipy` 1.17.1; `biopython` 1.86.

Compute. The study used 46.21 A100 GPU-hours in total: 33.21 for the first set of experiments (29.86 for 42 training runs, 1.17 for teacher precomputation, 2.19 for evaluation) and 13.00 for the seed replicates, the fully matched design, the permutation control, the budget runs and the probe. A 7.5M configuration trains in 27 minutes on one A100 (0.45 GPU-hours) and a 33.5M configuration in 67.7 minutes (1.13 GPU-hours). Scoring all 201 assays with a trained 7.5M encoder takes 15 seconds.

E ALGORITHMS

Algorithm 1 Training step for one sequence with corrupted input. No `[MASK]` token enters the encoder, and a corrupted position never receives its true residue.

Require: sequence x of length L; rate p; weight α on unaltered positions; soft-target weight λ; corruption source r; covered set $\mathcal{C}$

$n \leftarrow \max(1, \mathrm{round}(pL))$
if r is the teacher **and** $|\mathcal{C}| \geq n$ **then**
  $S \leftarrow n$ positions drawn uniformly without replacement from $\mathcal{C}$
else if r is the teacher **then**
  $S \leftarrow \mathcal{C} \cup (n - |\mathcal{C}|)$ positions drawn uniformly from $\{1..L\} \setminus \mathcal{C}$ (uniform substitutes at these)
else
  $S \leftarrow n$ positions drawn uniformly without replacement
end if
for $i \in S$ **do**
  $\tilde{x}_i \sim r_i(\cdot)$ restricted to $\mathcal{A} \setminus \{x_i\}$ and renormalized
end for
$\tilde{x}_j \leftarrow x_j$ for all $j \notin S$
$\ell_i \leftarrow -\log s_\theta(x_i \mid \tilde{x}, i)$ for every $i \in \{1..L\}$
for $i \in \mathcal{C}$ **do**
  $\ell_i \leftarrow (1-\lambda)\,\ell_i + \lambda\,\mathrm{KL}(q_i \,\|\, s_\theta(\cdot \mid \tilde{x}, i))$
end for
return split loss $\frac{1}{|S|}\sum_{i\in S}\ell_i + \alpha\,\frac{1}{L-|S|}\sum_{i\notin S}\ell_i$, or dense loss $\frac{1}{L}\sum_i \ell_i$

Algorithm 2 Inference: one forward pass per assay, independent of the number of variants.

Require: wild-type sequence x; set V of variants, each a set of substitutions $(i, a \to b)$

$\ell \leftarrow \log s_\theta(\cdot \mid x, \cdot)$ $\{\ell \in \mathbb{R}^{L\times 20}$, renormalized over the 20 canonical residues$\}$
for each $v \in V$ **do**
  $\mathrm{score}(v) \leftarrow \sum_{(i,\,a\to b)\,\in\,v} [\ell_{i,b} - \ell_{i,a}]$
end for
return $\{\mathrm{score}(v)\}_{v\in V}$

Renormalizing over 20 residues rather than ESM's 33-token vocabulary cancels in the log-ratio and does not change any ranking. Inference is deterministic. RTD-BIN reads $-\log P(\text{corrupt} \mid i)$ from the mutant sequence, and each variant is a different sequence, so it requires one forward pass per variant (12,010 per assay on average).

F Evaluation and Excluded Assays

We evaluate 201 of the 217 ProteinGym v1 substitution assays (retrieved 2026-07-25). One rule is applied before any scoring: assays whose wild-type sequence exceeds 1,022 residues are excluded, which is ESM-2's limit (`max_position_embeddings` 1026, minus `<cls>` and `<eos>`). No assay was excluded for another reason. The retained set spans 37 to 934 residues, 2,413,913 variants, and 28 low, 101 medium and 72 high MSA-depth assays.

Table 6: The 16 assays excluded by the sequence-length rule, with wild-type lengths.

excluded assay	length	excluded assay	length
CAR11_HUMAN_Meitlis_2020_gof	1154	NPC1_HUMAN_Erwood_2022_RPE1	1278
CAR11_HUMAN_Meitlis_2020_lof	1154	CAS9_STRP1_Spencer_2017_positive	1368
KCNH2_HUMAN_Kozek_2020	1159	BRCA1_HUMAN_Findlay_2018	1863
UBE4B_MOUSE_Starita_2013	1173	SCN5A_HUMAN_Glazer_2019	2016
ERBB2_HUMAN_Elazar_2016	1255	POLG_CXB3N_Mattenberger_2021	2185
SPIKE_SARS2_Starr_2020_binding	1273	POLG_HCVJF_Qi_2014	3033
SPIKE_SARS2_Starr_2020_expression	1273	BRCA2_HUMAN_Erwood_2022_HEK293T	3418
NPC1_HUMAN_Erwood_2022_HEK293T	1278	A0A140D2T1_ZIKV_Sourisseau_2019	3423

Validation of the scoring code. Reference scores for released ESM-2 (8M, 35M, 150M, 650M) and ESM-1v under masked marginals are taken from ProteinGym's published table. Our own masked-marginals scoring of ESM-2 8M on the same assays has correlation 1.0000 with the published column and mean absolute difference 0.0003; a ten-assay check of ESM-2 650M gives a maximum absolute difference of 0.0005. Wild-type-marginal scores are not published for any model and were computed by us.

Teacher-free verification. For CORRUPT-B, CORRUPT-B-15, CORRUPT-B-SPARSE-15 and TEMPCAL, the code path that loads teacher data is disabled and the cache directory points to a nonexistent path, so any attempt to read teacher data would have ended the run. All runs completed. TEMPCAL's temperature was fitted offline against the teacher's cached distributions, so it is teacher-free during training but not in its design. The masked configurations (MASK, MASK-DENSE) inherit a teacher cache path from their base configuration and load the cache, but with $\lambda = 0$ and mask-token corruption it enters neither the input nor the loss, so no teacher information reaches any cell of the fully matched design.

G Full Factorial, Decompositions, and Seeds

Table 7: The original 2×2 design at seed 1337, with win counts and Wilcoxon p-values (three-seed values are in Table 9). Row effects are the density effect for each input type; column effects are the input effect at each density.

	sparse	dense	density effect	
7.5M				
masked	MASK 0.1341	MASK-DENSE 0.1369	+0.0028	88/201, $p = 0.60$
corrupted	CORRUPT-T-SPARSE 0.1718	CORRUPT-T 0.2546	+0.0827	158/201, $p = 3.4\times10^{-21}$
input effect	+0.0377	+0.1177	interaction +0.0800	
	130/201, $p = 4.6\times10^{-4}$	175/201, $p = 6.5\times10^{-28}$		
33.5M				
masked	0.1236	0.1440	+0.0204	119/201, $p = 3.9\times10^{-4}$
corrupted	0.1930	0.2704	+0.0774	158/201, $p = 2.3\times10^{-20}$
input effect	+0.0693	+0.1264	interaction +0.0571	
	144/201, $p = 7.3\times10^{-9}$	171/201, $p = 1.5\times10^{-26}$		

Exact decompositions. Table 8 gives four partitions of pairwise differences. Each passes through a shared intermediate configuration, so its terms sum to the total by construction. They describe differences in means, not causal mediation, and single-factor ablations relative to CORRUPT-T are

not jointly additive. The ANOVA decomposition closes exactly (residual 0 at 7.5M and 2.8×10^{-17} at 33.5M); the interaction's share is 66% at 7.5M (seed 1337; 58% on three-seed means, with 34% for input type and 8% for density) and 39% at 33.5M, where the input main effect has 47%.

Table 8: Exact decompositions (seed 1337; 7.5M unless stated).

partition	term	$\Delta\rho$
two-factor, 7.5M (total +0.1205)	input type	+0.0377
	supervision density	+0.0028
	interaction	+0.0800
two-factor, 33.5M (total +0.1468)	input type	+0.0693
	supervision density	+0.0204
	interaction	+0.0571
objective vs. distillation (total +0.1205)	input type and density together (CORRUPT-T −MASK+KD)	+0.0965
	soft targets added to MASK (MASK+KD −MASK)	+0.0240
corruption source (total +0.0921)	BLOSUM62 over uniform (CORRUPT-B$_{\lambda=0.5}$ −CORRUPT-U)	+0.0808
	teacher over BLOSUM62 (CORRUPT-T −CORRUPT-B$_{\lambda=0.5}$)	+0.0113
split loss (total +0.1205)	loss at unaltered positions (CORRUPT-T −CORRUPT-T-SPARSE)	+0.0827
	rate at fixed sparse loss (CORRUPT-T-SPARSE −CORRUPT-T-SPARSE-15)	−0.0017
	input type and teacher access at matched density (CORRUPT-T-SPARSE-15 −MASK)	+0.0394

The term CORRUPT-T −MASK+KD combines input type and supervision density, because CORRUPT-T computes the loss at all L positions and MASK+KD only at its selected 15%. The term CORRUPT-T-SPARSE-15 −MASK (+0.0394) is the only density-matched contrast in the table, but it also changes teacher access, so it is not an isolated input-type effect. The two samplers in CORRUPT-T −CORRUPT-B$_{\lambda=0.5}$ differ in context dependence and in the distribution of substitutes.

Seeds. Table 9 gives all four cells of the original 7.5M design for three seeds and Table 10 the fully matched design. CORRUPT-T versus MASK is significant under each seed-matched pairing: +0.1205 (166/201), +0.1324 (168/201) and +0.1258 (170/201), with the largest p equal to 1.0×10^{-26}. CORRUPT-T-SPARSE-15 (seed values 0.1735, 0.1769, 0.1729) versus MASK+KD (seed 1337), which match the selection rate and the sparse loss, is significant for all three seeds of CORRUPT-T-SPARSE-15. Teacher access differs: every supervised position of CORRUPT-T-SPARSE-15 carries a soft target, whereas MASK+KD selects positions independently of the cache (coverage 0.600). Differences: +0.0154 ($p = 0.0071$), +0.0188 ($p = 0.0024$) and +0.0149 ($p = 0.0104$). CORRUPT-T-SPARSE-15 versus MASK gives +0.0394, +0.0576 and +0.0462.

Seed	MASK	MASK-DENSE	CORRUPT-T-SPARSE	CORRUPT-T	$\Delta^{\text{mask}}_{\text{dens}}$	$\Delta^{\text{corr}}_{\text{dens}}$	Interaction
1337	+0.1341	+0.1369	+0.1718	+0.2546	+0.0028	+0.0827	+0.0800
2024	+0.1193	+0.1327	+0.1679	+0.2517	+0.0135	+0.0838	+0.0703
777	+0.1267	+0.1390	+0.1701	+0.2525	+0.0123	+0.0824	+0.0700
Mean	+0.1267	+0.1362	+0.1700	+0.2529	+0.0095	+0.0830	+0.0734
SD					0.0059	0.0007	0.0057

Table 9: Original 2×2 factorial at 7.5M, all four cells at three pretraining seeds. Cells are mean Spearman ρ over 201 assays (wild-type marginals). Interaction = (CORRUPT-T − CORRUPT-T-SPARSE) − (MASK-DENSE − MASK) per seed; SD is across seeds ($n = 3$). On seed-averaged cells the interaction is positive on 147/201 assays (Wilcoxon $p = 1.3 \times 10^{-16}$).

Seed	MASK	MASK-DENSE	CORRUPT-B-SPARSE-15	CORRUPT-B-15	$\Delta^{\text{mask}}_{\text{dens}}$	$\Delta^{\text{corr}}_{\text{dens}}$	Interaction
1337	+0.1341	+0.1369	+0.0304	+0.2239	+0.0028	+0.1935	+0.1908
2024	+0.1193	+0.1327	+0.0234	+0.2255	+0.0135	+0.2021	+0.1886
777	+0.1267	+0.1390	+0.0212	+0.2262	+0.0123	+0.2050	+0.1927
Mean	+0.1267	+0.1362	+0.0250	+0.2252	+0.0095	+0.2002	+0.1907
p (seed-avg.)					0.02	5.1×10^{-21}	1.4×10^{-15}
33.5M (1337)	+0.1236	+0.1440	+0.0182	+0.2313	+0.0204	+0.2130	+0.1927

Table 10: Fully matched, teacher-free factorial: $p = 0.15$ and $\lambda = 0$ in every cell, no teacher information reaches any cell; only the input type differs between rows (masked: 80/10/10 `[MASK]`; corrupted: BLOSUM62, $T = 1$) and only the loss support and normalization between columns (selected positions vs. uniform over all L). 7.5M, three seeds; p from two-sided Wilcoxon over 201 assays on seed-averaged cells. Last row: 33.5M, seed 1337 only.

H LOSS AT UNALTERED POSITIONS BY CHECKPOINT

Table 11: Training-split cross-entropy and accuracy at unaltered positions, by training step. The validation split gives 0.0166, 0.4438 and 0.5933 at step 25,000 for the first three rows.

configuration	500	5,000	12,500	25,000
MASK-DENSE 7.5M	0.0177 / 1.0000	0.0166 / 1.0000	0.0118 / 1.0000	0.0165 / 1.0000
CORRUPT-T 7.5M	0.4662 / 1.0000	0.4523 / 0.9996	0.4687 / 0.9850	0.4480 / 0.9973
MASK 7.5M	0.6840 / 0.9966	0.6339 / 0.9949	0.6125 / 0.9848	0.6239 / 0.9912
CORRUPT-B-15 7.5M	0.1686 / 1.0000	0.1573 / 0.9999	0.1563 / 0.9999	0.1565 / 0.9999
MASK-DENSE 33.5M	—	—	—	0.0163 / 1.0000

The geometric-mean probability assigned to the true residue at unaltered positions is about 98% for MASK-DENSE and about 64% for CORRUPT-T at step 25,000, at similar accuracy.

Loss decomposition on the validation split. Table 12 applies each configuration's own corruption procedure to the 5,000 validation sequences (the validation seed of the trainer) and computes cross-entropy against the true residue, without the soft-target term, for the final 25,000-step checkpoints (seed 1337). For each corrupted configuration, the best context-free predictor is the empirical distribution of the true residue given the visible residue, estimated from the same corrupted validation set; it uses the residue frequencies, p and the corruption source, but not the surrounding sequence. The comparison rule was recorded before these quantities were computed.

Table 12: Cross-entropy at unaltered positions, at selected positions and over all positions on the validation split (final checkpoints, seed 1337), for the model and for the best context-free predictor of its corruption process.

				model			context-free predictor		
configuration	p	sampler	λ	unaltered	selected	all	unaltered	selected	all
CORRUPT-T	0.30	teacher	0.5	0.4431	3.1892	1.2669	0.3549	3.7033	1.3594
CORRUPT-B	0.30	BLOSUM62	0	0.3301	2.9201	1.1071	0.3453	3.0804	1.1658
CORRUPT-B-15	0.15	BLOSUM62	0	0.1547	3.5568	0.6650	0.1596	3.7472	0.6978
MASK-DENSE	0.15	—	0	0.0167	2.8031	0.4347	—	—	—

CORRUPT-B-15 is below its context-free predictor both at unaltered positions and overall. CORRUPT-T is above its predictor at unaltered positions and below it overall, whereas CORRUPT-B, with the same p but a different sampler and no soft targets, is below its predictor at unaltered positions. For MASK-DENSE the selected positions include the mask tokens.

I ROW-PERMUTED BLOSUM62 CONTROL

CORRUPT-B-SHUF uses BLOSUM62 at $T = 1$ with the probabilities in each row (over the 19 residues other than the wild type) permuted by a fixed random permutation (one per row), so every row keeps its entropy and its set of values while the assignment of values to residues is lost. The configuration is otherwise identical to CORRUPT-B$_{\lambda=0.5}$ ($p = 0.30$, $\lambda = 0.5$). CORRUPT-B$_{\lambda=0.5}$, CORRUPT-U and CORRUPT-B-SHUF were each trained with three seeds, and the partition uses seed means. Because each row keeps its sorted probability vector, its entropy and maximum probability are unchanged.

Arm / term	1337	2024	777	Seed mean	Wins	p
CORRUPT-U	+0.1625	+0.1638	+0.1612	+0.1625		
CORRUPT-B-SHUF	+0.1730	+0.1717	+0.1723	+0.1723		
CORRUPT-B$_{\lambda=0.5}$	+0.2433	+0.2431	+0.2449	+0.2438		
Total: B − U	+0.0808	+0.0793	+0.0837	+0.0813	179/201	6.1×10^{-30}
Shape: B-Shuf − U	+0.0105	+0.0078	+0.0111	+0.0098	117/201	1.3×10^{-4}
Pattern: B − B-Shuf	+0.0703	+0.0714	+0.0726	+0.0714	179/201	7.4×10^{-30}
Substitution-pattern share β	0.87	0.90	0.87	0.88		

Table 13: Corruption-source partition with a shuffled-BLOSUM control (7.5M, $p = 0.30$, $\lambda = 0.5$). CORRUPT-B-SHUF permutes each source residue's 19 BLOSUM62 replacement probabilities among its 19 targets (fixed permutation, identical for all seeds): every row keeps its entropy and sorted probability vector, the substitution pattern is removed. Wins and p (two-sided Wilcoxon, 201 assays) on seed-averaged cells; $\beta = (B - \text{B-Shuf})/(B - U)$ on seed means.

J TEACHER BUDGET AND CORRUPTION HARDNESS

Allocation. A sequence-level hybrid HYB-SEQγ samples all corruptions of a random fraction γ of training sequences from the teacher and those of the other sequences from BLOSUM62 ($T = 1$). Teacher precomputation is needed only for the teacher sequences, so it falls to $\gamma \times 1.17$ GPU-hours. All budget configurations use $\lambda = 0$, because the soft-target term would require the teacher's distributions for every sequence.

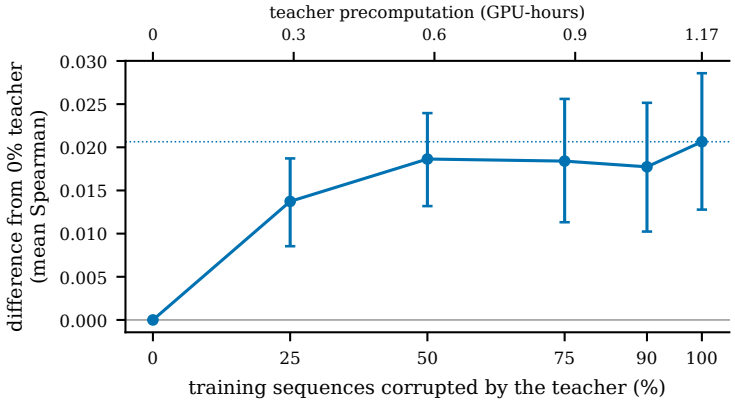


Figure 4: Partial teacher budgets ($\lambda = 0$, seed 1337, one run per point). Difference in mean Spearman from CORRUPT-B (0% teacher) as a function of the fraction of training sequences whose substitutes are sampled from the teacher; 100% is CORRUPT-T$_{\lambda=0}$. Error bars are paired bootstrap 95% confidence intervals over assays; the dotted line marks the full-teacher value. The top axis gives the teacher precomputation required.

Arm	Teacher seqs	Teacher GPU-h	Mean ρ	Δ vs 0%	Wins	p
CORRUPT-B (no teacher)	0%	0.000	+0.2201	–	–	–
HYB-SEQ25	25%	0.292	+0.2339	+0.0137	131/201	6.8×10^{-8}
HYB-SEQ50	50%	0.585	+0.2388	+0.0186	143/201	2.1×10^{-10}
HYB-SEQ75	75%	0.877	+0.2385	+0.0184	132/201	1.2×10^{-7}
HYB-SEQ90	90%	1.053	+0.2379	+0.0177	130/201	1.0×10^{-6}
CORRUPT-T$_{\lambda=0}$	100%	1.170	+0.2408	+0.0206	127/201	2.0×10^{-7}

Table 14: Teacher budget (sequence-level allocation, $\lambda = 0$, 7.5M, seed 1337; single seed per point). Teacher GPU-h required = budget $\times$ 1.17.

Offline hardness analysis. Over 632,228 covered positions from 4,000 sequences, teacher-sampled corruptions have a mean teacher probability of 0.2916 (median 0.1818) and BLOSUM62 corruptions at $T = 1$ of 0.1496 (median 0.0590); in $-\log q$ terms (hardness, §4), 1.813 and 3.099. The row-permuted matrix has hardness 4.554 and uniform corruption 4.751. As $T \to 0$ the BLOSUM62 value decreases to about 2.348 and does not reach the teacher's 1.813, so corruption difficulty cannot be matched by a temperature. The closest point ($T^* = 0.05$) leaves a gap of 0.535, with histogram total-variation distance 0.143. TEMPCAL trains at this temperature and scores 0.1965, 0.0237 below CORRUPT-B (75/201, $p = 1.6 \times 10^{-4}$).

For a single 7.5M model, teacher precomputation (1.17 GPU-hours) is 72% of the end-to-end compute (1.62 GPU-hours).

K SUPERVISED PROBE

The probe uses the final-layer embeddings of each frozen encoder, which are the input to the output head; neither the head nor the log-ratio score is used. For a variant v with mutated positions $M(v)$ and mutant residues m_i, the feature vector is $\phi(v) = \sum_{i \in M(v)} \hat{h}_i \otimes e_{m_i}$, where $\hat{h}_i$ is the wild-type embedding at position i and e_{m_i} a one-hot vector over the 20 residues. A kernel ridge regression is fitted per assay with five-fold cross-validation (folds identical across encoders, regularization chosen by leave-one-out on each training fold), and the Spearman correlation of the out-of-fold predictions is reported. At most 5,000 variants per assay are used (a fixed subset). The one-hot baseline replaces $\hat{h}_i$ with a one-hot indicator of position i and uses no encoder.

Encoder	Mean ρ	95% CI	Δ vs MASK	Wins	p
One-hot (no model)	+0.1917	[0.1455, 0.2403]	−0.3150	5/201	3.3×10^{-34}
MASK	+0.5067	[0.4779, 0.5379]	–	–	–
MASK-DENSE	+0.5034	[0.4748, 0.5353]	−0.0033	91/201	0.20
CORRUPT-T-SPARSE	+0.5209	[0.4921, 0.5517]	+0.0142	142/201	1.0×10^{-9}
CORRUPT-T	+0.5204	[0.4919, 0.5517]	+0.0137	140/201	1.6×10^{-11}
CORRUPT-B-SPARSE-15	+0.5047	[0.4766, 0.5361]	−0.0020	86/201	0.15
CORRUPT-B-15	+0.5072	[0.4784, 0.5383]	+0.0005	96/201	0.50
CORRUPT-B	+0.5126	[0.4848, 0.5431]	+0.0059	112/201	0.0080
ESM-2 8M (released)	+0.5453	[0.5175, 0.5751]	+0.0386	155/201	5.2×10^{-17}
Interaction, original 2×2	+0.0028			105/201	0.34
Interaction, matched 2×2 (seed 1337)	+0.0058			118/201	0.0024

Table 15: Supervised linear probe on frozen final-layer embeddings (the input of the LM head; no LM head, no log-ratios). Per assay: $\phi(v) = \sum_{i \in M(v)} \hat{h}_i \otimes e_{m_i}$, kernel ridge with α chosen by leave-one-out on the training fold, 5-fold CV with folds identical across encoders, Spearman of the out-of-fold predictions; up to 5,000 variants per assay (fixed subset); seed-1337 students. 95% CI: bootstrap over assays. Wins/p: paired two-sided Wilcoxon over 201 assays (interaction rows: vs. 0).

L SUBSTITUTION ORDER

For each assay we recomputed the Spearman correlation separately on its single-substitution variants and on its multi-substitution variants, keeping only assays with at least 50 variants of that type (199 and 69 assays), and computed an order-stratified correlation: variants are grouped by their number of substitutions k, the Spearman correlation is computed within each group with at least 20 variants, and the assay value is the group-size-weighted mean (for assays with only single substitutions it equals the all-variant value). Scores are unchanged: a multi-substitution variant is scored as the sum of its per-position log-ratios. These definitions and the decision rules for Table 17 were recorded before the quantities were computed.

	Single substitutions	Multiple substitutions
MASK	$+0.0592$	$+0.0574$
MASK-DENSE	$+0.0465$	$+0.0062$
CORRUPT-T-SPARSE	$+0.1648$	$+0.0397$
CORRUPT-T	$+0.1790$	$+0.0571$
CORRUPT-B-SPARSE-15	$+0.1463$	-0.0783
CORRUPT-B-15	$+0.1531$	$+0.0037$
Density effect, masked	-0.0127 ($p = 0.0013$)	-0.0511 ($p = 0.0010$)
Density effect, corrupted (T)	$+0.0142$ ($p = 1.2 \times 10^{-5}$)	$+0.0174$ ($p = 0.46$)
Density effect, corrupted (B, p=0.15)	$+0.0068$ ($p = 7.0 \times 10^{-4}$)	$+0.0820$ ($p = 0.21$)
Interaction, original	$+0.0270$ ($p = 1.9 \times 10^{-7}$)	$+0.0685$ ($p = 5.3 \times 10^{-4}$)
Interaction, matched	$+0.0195$ ($p = 3.2 \times 10^{-5}$)	$+0.1331$ ($p = 3.0 \times 10^{-6}$)
CORRUPT-T $-$ MASK	$+0.1198$ ($p = 3.0 \times 10^{-25}$)	-0.0003 ($p = 0.94$)
Assays ($\geq$50 rows of the type)	199	69

Table 16: Wild-type-marginal Spearman computed separately on single- and multi-substitution variants (seed 1337; only assays with at least 50 rows of that type). A multi-substitution variant is scored as the sum of its per-position log-ratios, which cannot represent epistasis. p: two-sided Wilcoxon over the assays in the column.

Design / metric	$\Delta^{\text{mask}}_{\text{dens}}$	$\Delta^{\text{corr}}_{\text{dens}}$	1337	2024	777	Interaction (mean $\pm$ s.d.)	p	n
Original (CORRUPT-T), 7.5M, three seeds								
all variants	$+0.0095 \pm 0.0059$	$+0.0830 \pm 0.0007$	$+0.0800$	$+0.0703$	$+0.0700$	$+0.0734 \pm 0.0057$	1.3×10^{-16}	201
single substitutions only	-0.0042 ± 0.0077	$+0.0119 \pm 0.0024$	$+0.0270$	$+0.0073$	$+0.0141$	$+0.0161 \pm 0.0100$	3.4×10^{-5}	199
order-stratified	-0.0068 ± 0.0087	$+0.0126 \pm 0.0029$	$+0.0302$	$+0.0129$	$+0.0151$	$+0.0194 \pm 0.0094$	2.1×10^{-6}	201
Fully matched (p=0.15, λ=0, BLOSUM62), 7.5M, three seeds								
all variants	$+0.0095 \pm 0.0059$	$+0.2002 \pm 0.0060$	$+0.1908$	$+0.1886$	$+0.1927$	$+0.1907 \pm 0.0020$	1.4×10^{-15}	201
single substitutions only	-0.0042 ± 0.0077	$+0.0153 \pm 0.0083$	$+0.0195$	$+0.0137$	$+0.0254$	$+0.0195 \pm 0.0058$	4.8×10^{-8}	199
order-stratified	-0.0068 ± 0.0087	$+0.0110 \pm 0.0073$	$+0.0196$	$+0.0163$	$+0.0175$	$+0.0178 \pm 0.0017$	4.7×10^{-6}	201
33.5M, seed 1337 only								
Original, all variants			$+0.0571$				1.0×10^{-12}	201
Original, single substitutions only			$+0.0153$				1.5×10^{-4}	199
Original, order-stratified			$+0.0108$				8.1×10^{-4}	201
Matched, all variants			$+0.1927$				4.6×10^{-21}	201
Matched, single substitutions only			$+0.0327$				2.4×10^{-11}	199
Matched, order-stratified			$+0.0241$				2.0×10^{-8}	201

Table 17: Interaction by substitution order. All variants: per-assay Spearman over every variant (as in the main results). Single substitutions only: assays with $\geq$50 single-substitution rows. Order-stratified: per assay, Spearman within each group of variants with the same number of substitutions k ($n_k \geq 20$), n_k-weighted mean over groups, so variants of different orders are never ranked against each other. Density effects: seed mean $\pm$ s.d.; interaction per seed, mean $\pm$ s.d.; p: two-sided Wilcoxon of the per-assay interaction on seed-averaged cells (7.5M) or seed 1337 (33.5M); n: assays.

Quantity	Single-only (132)	Mixed (69)	Share from mixed	Mean (201)	Median (201)
Original: interaction	+0.0197	+0.1762	82%	+0.0734	+0.0300
Original: density effect (corrupted row)	+0.0197	+0.2039	84%	+0.0830	+0.0305
Matched: interaction	+0.0094	+0.5375	97%	+0.1907	+0.0287
Matched: density effect (corrupted row)	+0.0094	+0.5652	97%	+0.2002	+0.0211

Table 18: Where the all-variant effect comes from (7.5M, seed-averaged cells over three seeds). Single-only assays contain no multi-substitution variants; mixed assays contain at least one. Share = sum of per-assay values over mixed assays / sum over all 201 assays.

Model	Seed	Mean log-ratio (singles)	Frac. singles above WT	Mean ρ(score, k) (mixed assays)
MASK	1337	−3.4716	0.005	−0.6761
MASK	2024	−3.4372	0.006	−0.6577
MASK	777	−3.4687	0.005	−0.6783
MASK-DENSE	1337	−5.8982	0.000	−0.7687
MASK-DENSE	2024	−5.7017	0.000	−0.7506
MASK-DENSE	777	−6.1314	0.000	−0.7629
CORRUPT-T-SPARSE	1337	−0.7870	0.271	−0.1423
CORRUPT-T-SPARSE	2024	−0.8136	0.263	−0.1344
CORRUPT-T-SPARSE	777	−0.7809	0.277	−0.1587
CORRUPT-T	1337	−3.9187	0.000	−0.7498
CORRUPT-T	2024	−3.9034	0.000	−0.7532
CORRUPT-T	777	−3.9068	0.000	−0.7499
CORRUPT-B-SPARSE-15	1337	+5.4403	0.995	+0.6716
CORRUPT-B-SPARSE-15	2024	+5.4381	0.996	+0.6753
CORRUPT-B-SPARSE-15	777	+5.3955	0.995	+0.6668
CORRUPT-B-15	1337	−6.1437	0.000	−0.7144
CORRUPT-B-15	2024	−6.1566	0.000	−0.7159
CORRUPT-B-15	777	−6.1425	0.000	−0.7142
CORRUPT-B-SPARSE-15, α=1	1337	−4.4409	0.005	−0.6094
MASK-35M	1337	−3.1199	0.007	−0.6437
MASK-DENSE-35M	1337	−5.4154	0.000	−0.7611
CORRUPT-T-SPARSE-35M	1337	−0.8682	0.273	−0.1829
CORRUPT-T-35M	1337	−3.9948	0.000	−0.7457
CORRUPT-B-SPARSE-15-35M	1337	+5.7743	0.994	+0.6598
CORRUPT-B-15-35M	1337	−6.1843	0.000	−0.7137
Order baseline (score $= -k$)	–	–	–	+0.3780 (vs. fitness, 69 assays)

Table 19: Score offset under wild-type marginals. Mean log-ratio $\log p(\text{mt}) - \log p(\text{wt})$ over single-substitution variants; fraction of single substitutions scored above the wild type; mean over mixed assays of the Spearman correlation between a variant's summed score and its number of substitutions k. Last row: Spearman between $-k$ and measured fitness (no model), averaged over mixed assays.

M MASKED MARGINALS AND CROSS-PROTOCOL RESULTS

Table 20: Mean and median Spearman under wild-type marginals (WT, one pass per assay) and masked marginals (MM, L passes per assay). Released ESM-2 masked-marginal values are ProteinGym's published scores; all other values were computed by us.

configuration	WT mean	WT median	MM mean	MM median
CORRUPT-T	0.2546	0.2479	0.0760	0.0568
CORRUPT-T-15	0.2685	0.2595	0.0338	0.0163
CORRUPT-T, $\lambda = 1$	0.2646	0.2616	−0.0024	−0.0134
CORRUPT-T$_{\lambda=0}$	0.2408	0.2234	0.0584	0.0399
CORRUPT-B$_{\lambda=0.5}$	0.2433	0.2414	−0.0360	−0.0612
CORRUPT-B	0.2201	0.2170	−0.0545	−0.0585
CORRUPT-U	0.1625	0.1389	0.0347	0.0309
CORRUPT-T-SPARSE	0.1718	—	0.0104	—
CORRUPT-T-SPARSE-15	0.1735	—	0.0431	—
CORRUPT-T-35M	0.2704	0.2637	0.0558	0.0338
CORRUPT-T-SPARSE-35M	0.1930	—	0.0742	—
CORRUPT-T-SPARSE-15-35M	0.1906	—	0.0817	—
MASK+KD	0.1581	0.1208	0.0630	0.0528
MASK-DENSE	0.1369	0.0852	0.0493	0.0448
MASK-DENSE-35M	0.1440	0.0815	0.0587	0.0482
MASK	0.1341	0.0848	0.0670	0.0521
MASK-35M	0.1236	0.0901	0.0658	0.0513
ESM-2 8M	0.2578	0.2460	0.2111	0.1940
ESM-2 35M	0.3450	0.3780	0.3300	0.3580
ESM-2 150M	0.4172	0.4495	0.4096	0.4550
ESM-2 650M	0.4516	0.4922	0.4460	0.4930

No 25,000-step configuration exceeds a masked-marginals mean of 0.082. The 7.5M interaction under masked marginals is $+0.0833$ ($p = 8.4 \times 10^{-21}$); the 33.5M interaction is -0.0113 ($p = 0.03$). Corrupted configurations never receive a `[MASK]` token during training, so masked marginals presents them with an input they have not seen, and their scores under it do not bear on the comparison in either direction. RTD-BIN is not listed because its score is read from the mutant sequence and corresponds to neither protocol.

The masked configurations also score lower under masked marginals, the protocol that matches their training, than under wild-type marginals (MASK: 0.1341 and 0.0670). This is not an error in our scoring code, which reproduces the published ESM-2 8M masked-marginals scores (Appendix F). We examined ESM's `token_dropout` rescaling as a cause and ruled it out: the embedding scale differs between the two protocols by 0.4%, which cannot account for the difference. We do not have an explanation for this result. In the separate 75,000-step run of MASK at 7.5M, the masked-marginals score at 25k, 40k, 55k and 75k steps is $+0.0621 \to +0.0762 \to +0.0842 \to +0.0829$, and the difference between wild-type and masked marginals is $+0.0669 \to +0.0601 \to +0.0572 \to +0.0597$: the masked-marginals score increases with training, but the difference does not decrease to zero by 75,000 steps.

For released ESM-2, the advantage of wild-type over masked marginals decreases with size: $+0.0467$ at 8M, $+0.0151$ at 35M, $+0.0077$ at 150M and $+0.0056$ at 650M. At 35M the mean favours wild-type marginals although wild-type marginals scores higher on only 64 of 201 assays ($p = 7 \times 10^{-4}$).

N TRAINING LENGTH AND TOKEN DROPOUT

Table 21: Mean Spearman at each checkpoint of the separate 75,000-step runs.

configuration	25k	40k	55k	75k
CORRUPT-T 33.5M	0.2817	0.2950	0.2998	0.3014
CORRUPT-T 7.5M	0.2562	0.2681	0.2755	0.2779
MASK 33.5M	0.1422	0.1543	0.1641	0.1674
MASK 7.5M	0.1291	0.1363	0.1414	0.1427

No early stop was triggered, and training and validation loss decreased at every checkpoint for all four runs, so these models are undertrained rather than limited by the data. Gains decrease with training but do not stop. Because warmup is 6% of the total steps, these runs are separate from the 25,000-step configurations rather than continuations of them. Within them, the difference between CORRUPT-T and MASK changes from $+0.1272$ to $+0.1352$ at 7.5M and from $+0.1395$ to $+0.1340$ at 33.5M. The 25,000-step values in the main text ($+0.1205$ and $+0.1468$) are from the standalone runs. At 75,000 steps, CORRUPT-T at 7.5M is $+0.0200$ above ESM-2 8M under wild-type marginals (111/201, $p = 0.059$, not significant), and CORRUPT-T at 33.5M is 0.0436 below ESM-2 35M (79/201, $p = 3.4\times10^{-4}$).

Embedding scale under `token_dropout`. ESM's `token_dropout` sets `[MASK]` embeddings to zero and rescales the remaining ones by $(1 - 0.12)/(1 - \text{observed mask ratio})$. MASK uses BERT's 80/10/10 rule, with a measured mask ratio of 0.11867, giving a scale of 0.99849 during training and 0.88000 at wild-type-marginals evaluation; CORRUPT-T uses no mask tokens and has scale 0.88000 in both. Re-scoring MASK at its training-time scale changes its mean Spearman by $+0.0002$. Retraining both configurations with `token_dropout` disabled gives a difference of $+0.1760$ (188/201, $p = 1.4\times10^{-32}$), larger than the reported $+0.1205$; the reported difference is therefore the smaller of the two.

O STRATIFIED ANALYSES

Table 22: Mean Spearman by MSA-depth stratum (assay counts in parentheses).

configuration	low (28)	medium (101)	high (72)
CORRUPT-T	0.2289	0.2400	0.2850
CORRUPT-B	0.1993	0.2069	0.2467
MASK+KD	0.1312	0.1365	0.1988
MASK	0.0927	0.1100	0.1841
ESM-2 8M (WT)	0.1965	0.2154	0.3412
ESM-2 8M (MM, published)	0.1595	0.1859	0.2665
ESM-2 650M (WT)	0.3507	0.4363	0.5123

Corrupted configurations score above masked ones in every stratum. Relative to released ESM-2 8M under the same protocol, CORRUPT-T is higher on low-depth ($+0.0324$) and medium-depth ($+0.0246$) assays and lower on high-depth assays (-0.0562).

Table 23: Mean Spearman by wild-type length, split at the 512-residue training cap (assay counts in parentheses).

configuration	$\leq$512 (162)	>512 (39)	change
CORRUPT-T 7.5M	0.2708	0.1870	−0.0838
CORRUPT-B 7.5M	0.2324	0.1692	−0.0632
MASK 7.5M	0.1623	0.0169	−0.1454
CORRUPT-T 33.5M	0.2886	0.1950	−0.0935
MASK 33.5M	0.1457	0.0319	−0.1138
ESM-2 8M	0.2754	0.1849	−0.0905
ESM-2 35M	0.3699	0.2417	−0.1282
ESM-2 150M	0.4468	0.2946	−0.1522

Our encoders were trained on sequences of at most 512 residues and evaluated up to 1,022, so longer assays require length extrapolation for our models but not for released ESM-2. Within strata, CORRUPT-T at 7.5M versus ESM-2 8M is -0.0046 ($p = 0.83$) at $\leq$512 and $+0.0022$ ($p = 0.95$) above; at 33.5M the difference is -0.0813 ($p = 1.8\times10^{-9}$) and -0.0467 ($p = 0.13$), and at the 75k checkpoint -0.0526 and -0.0065. The differences do not increase in the longer stratum. With 39 assays in that stratum these tests have low power, so this is an absence of evidence for a length penalty rather than evidence of its absence. Masked configurations decrease more on long sequences than corrupted ones.

P RELEASED CHECKPOINTS

Table 24: Our configurations and released checkpoints on the 201 assays. MM: masked marginals, published by ProteinGym; all other rows computed by us. Our configurations are trained on 300k Swiss-Prot sequences for 25,000 steps and are not compute-matched to the released checkpoints, which were trained on UniRef50. Masked marginals requires L passes per assay, RTD-BIN one pass per variant, and all other rows one pass per assay.

model / configuration	protocol	parameters	mean ρ	median ρ
ESM-2 650M	WT	651M	0.4516	0.4922
ESM-2 650M	MM	651M	0.4460	0.4930
ESM-1v (ensemble)	MM	5×650M	0.4412	0.4770
ESM-2 150M	WT	149M	0.4172	0.4495
ESM-2 150M	MM	149M	0.4096	0.4550
ESM-1v (single)	MM	650M	0.4090	0.4250
ProtBert	WT	420M	0.3766	0.3847
ESM-2 35M	WT	33.5M	0.3450	0.3780
ESM-2 35M	MM	33.5M	0.3300	0.3580
CORRUPT-T-35M	WT	33.5M	0.2704	0.2637
ESM-2 8M	WT	7.5M	0.2578	0.2460
CORRUPT-T	WT	7.5M	0.2546	0.2479
CORRUPT-B	WT	7.5M	0.2201	0.2170
ESM-2 8M	MM	7.5M	0.2111	0.1940
CORRUPT-U	WT	7.5M	0.1625	0.1389
MASK+KD	WT	7.5M	0.1581	0.1208
MASK-DENSE	WT	7.5M	0.1369	0.0852
MASK	WT	7.5M	0.1341	0.0848
MASK-35M	WT	33.5M	0.1236	0.0901
RTD-BIN	mutant sequence	7.5M	−0.1938	−0.1310

CORRUPT-T at 7.5M does not differ significantly from ESM-2 8M under wild-type marginals (-0.0033, 102/201, $p = 0.99$); the point estimate is slightly in favour of ESM-2. At 33.5M our configuration is lower than ESM-2 35M (-0.0746, 65/201, $p = 3.8\times10^{-9}$). CORRUPT-U and MASK+KD differ by 0.0044, less than the seed standard deviation of MASK (0.0074), so the separation between corrupted and masked inputs holds only for corrupted configurations with a

non-uniform corruption source. ProtBert required reading its token map from `vocab.txt` and constructing `BertForMaskedLM` explicitly under `transformers` 5.x. ProtT5 is not in ProteinGym's published zero-shot table and was not run.

Q Additional Figures

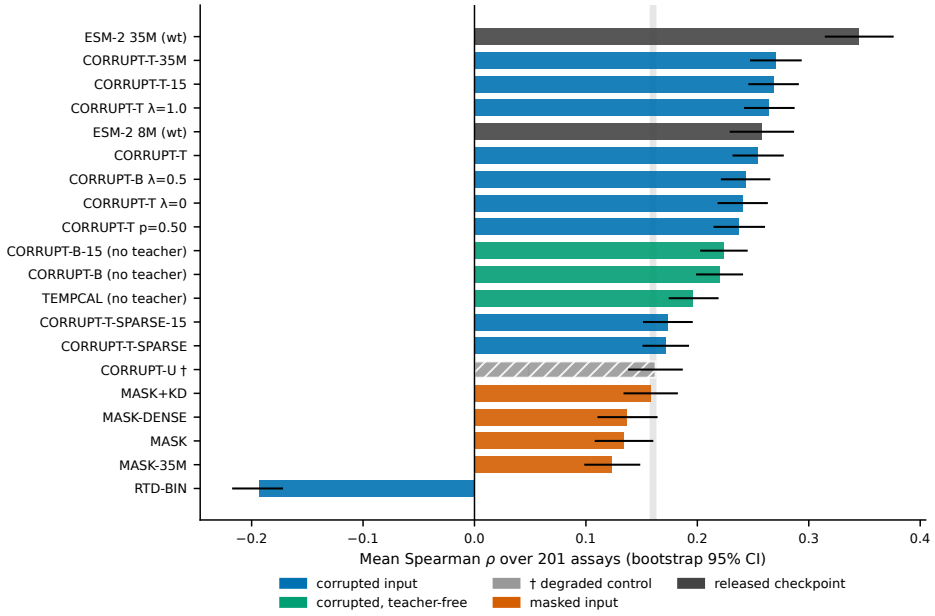


Figure 5: All 7.5M configurations and selected others by mean Spearman over 201 assays, with bootstrap 95% confidence intervals, coloured by input type. CORRUPT-U is a control with a uniform corruption source; its interval overlaps that of MASK+KD.

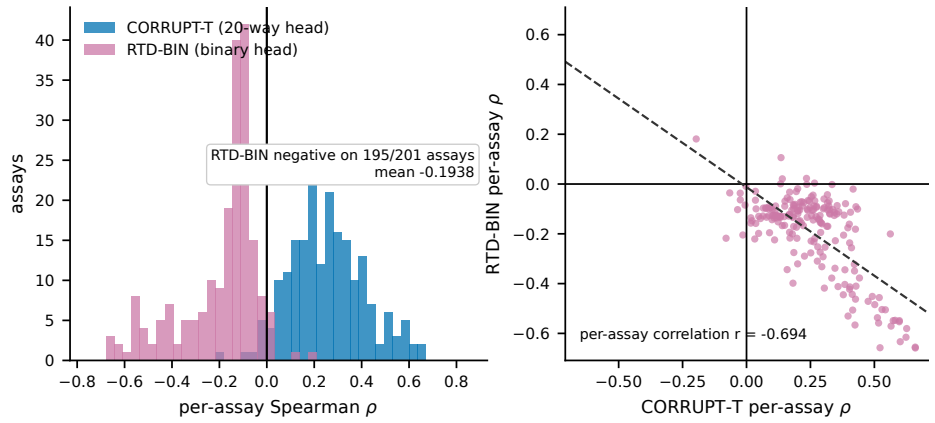


Figure 6: RTD-BIN against CORRUPT-T. Left: distribution of per-assay Spearman. Right: per-assay Spearman of RTD-BIN against CORRUPT-T ($r = -0.694$).

R Additional Limitations and Negative Results

Teacher approximation. Teacher conditionals were computed with four rounds of disjoint 15% masking rather than per-position masking, covering 60% of positions. The approximation is small

by the measurements in Appendix C, but it is an approximation, and uncovered positions receive the full cross-entropy loss regardless of λ.

Seed coverage. Three seeds were run for the four cells of the original 7.5M design, the cells of the fully matched design, CORRUPT-T-SPARSE-15, and the three configurations of the permutation control (Table 3). All other configurations, including RTD-BIN, the 33.5M cells and all budget configurations, use a single seed. Seed stability was not verified for the single-seed configurations.

Gradient scale. The sparse and dense configurations differ in a global loss factor p as well as in density (§2). Because gradient clipping is not scale-invariant, their effective step sizes differ slightly. This can affect the size of the density effects; the split-loss comparison at $\alpha = 0.1$, which changes density without changing the global factor, is consistent with the factor not accounting for them.

Hyperparameters. The corruption rate $p = 0.30$ and soft-target weight $\lambda = 0.5$ were fixed before any benchmark evaluation. A sweep finds $p = 0.15$ (0.2685; CORRUPT-T scores higher on only 65/201 assays, $p = 1.4\times10^{-10}$) and $\lambda = 1.0$ (0.2646; 68/201, $p = 3.9\times10^{-7}$), both above the default. Configurations use the defaults unless a design specifies otherwise (Table 1), and no comparison mixes tuned and untuned settings. No baseline was tuned. The higher score at $p = 0.15$ does not indicate that a lower corruption rate is better: at a fixed sparse loss the rate has no measurable effect (-0.0017, $p = 0.31$), and $p = 0.15$ scores higher in the sweep because uniform averaging over all positions gives a larger relative weight to unaltered positions at a lower rate.

Negative results. Sharpening BLOSUM62 by a temperature to approach the teacher's corruption hardness lowers mean Spearman (TEMPCAL, -0.0237; Appendix J). Standardizing scores per position lowers mean Spearman for every method, including released ESM-2. At 25,000 steps, uniform corruption gives a higher training-split loss at unaltered positions than teacher corruption (CORRUPT-U 0.5229, CORRUPT-T 0.4480). ESM's `token_dropout` rescaling does not explain the masked-marginals result for MASK (Appendix M). Against released ESM-2 8M, CORRUPT-T scores higher on low- and medium-depth assays and lower on high-depth assays (Appendix O). RTD-BIN requires one forward pass per variant, not one per protein.