PROTEINJEPA: LATENT PREDICTION IMPROVES PROTEIN LANGUAGE MODEL PRETRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein language models are trained primarily with masked language modeling (MLM), which predicts masked amino-acid identities. Joint-embedding predictive architectures (JEPA) instead predict latent representations, but have not been applied to proteins. ProteinJEPA supplements MLM with a cosine loss for predicting the half-depth hidden states of a teacher given the unmasked sequence. On 19 tasks, with ESM2 at 35M and 150M parameters and three pretraining seeds, MLM+JEPA outperforms compute-matched and step-matched MLM-only continued training in 78 and 76 of 114 comparisons (14 losses, 22 ties). The median compute-matched gain is $+0.0106$ on structure- and homology-sensitive tasks versus $+0.0041$ elsewhere, led by SCOPe-40 retrieval and remote homology with improvements of 6.1 percentage points in Recall@1 and 2.7 points in accuracy, respectively. Gains on these tasks increase with model size from 8M to 150M. Against the off-the-shelf checkpoint, MLM+JEPA wins 81 of 114 comparisons (median $+0.0068$) without improving MLM loss. In random initialization the gain is smaller and replicates inconsistently across seeds (p=0.059). The same recipe improves the causal ProGen3 model, beating a compute-matched next-token-prediction control on 12 of 16 tasks. Ablations show that cosine loss beats mean squared error, while adding shallower targets removes most of the task gain. JEPA-only training collapses downstream performance: latent prediction complements MLM rather than replacing it. Code: `https://anonymous.4open.science/r/protJepa-FF24`.

1 INTRODUCTION

Masked language modeling (MLM) is the standard self-supervised objective for protein language models (Brandes et al., 2022; Lin et al., 2023; Rives et al., 2019; Vieira et al., 2025): mask residues and predict their identities. Joint-embedding predictive architectures (JEPA) (LeCun, 2022; Balestriero & LeCun, 2025; Huang et al., 2025; Schmidhuber & Prelinger, 1993) provide a different signal: predicting a hidden representation of the unmasked input rather than the input tokens. Latent prediction has worked well in vision, video, language, and DNA, but has not been applied to proteins (Huang et al., 2025; Balestriero & LeCun, 2025; Larey et al., 2026).

We ask: *can latent prediction improve protein language model pretraining?* To our knowledge, ProteinJEPA is the first application of JEPA-style latent prediction to protein language models. We find it complements rather than replaces MLM: the model jointly predicts masked amino-acid identities and the half-depth hidden states of an exponential-moving-average (EMA) teacher (Sec. 3). Our controlled study applies the recipe to ESM2 at three model sizes (8M, 35M, and 150M; three independent pretraining seeds at 35M and 150M), continued from pretrained checkpoints and trained from random initialization, against step-matched and compute-matched MLM-only controls and the off-the-shelf checkpoint. We extend the recipe to the causal ProGen3-112m model, trained by next-token prediction (NTP). A second, wall-clock-matched study over five backbones tests the breadth of the finding (Sec. 5.5).

Contributions.

- **Latent prediction improves continued pretraining under matched budgets** (Table 2). MLM+JEPA beats step- and compute-matched MLM-only in every pretraining seed, and is

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.

81/20/13 against the off-the-shelf checkpoint where MLM-only continuation is 52/35/27. The same holds for the causal ProGen3 model (Sec. 5.6).

- **The gain concentrates on structural tasks and grows with model size.** SCOPe-40 retrieval and remote homology gain most, and their gains grow from 8M to 150M (Appendix A.2); the structure/homology subset's median gain over the compute-matched control is $+0.0106$, versus $+0.0041$ on the other fifteen tasks (Fig. 1, Sec. 5.2). Peptide–HLA binding, signal-peptide residues, fluorescence, and β-lactamase also gain in at least five of six runs, and per-residue probes improve.
- **Latent prediction complements, but does not replace MLM/NTP.** JEPA-only continuation degrades every pretrained backbone, losing 17 of 19 tasks on ESM2-35M (median -0.124; Sec. 5.7). It also underperforms MLM from random initialization.
- **Continued pretraining is where the gain is largest.** MLM+JEPA wins 78 and loses 14 of 114 comparisons when continuing pretraining, but wins 54 and loses 38 from random initialization (Sec. 5.3).
- **Recipe ablations.** Cosine loss beats mean squared error. Masked- and all-position targets have similar downstream performance, and adding shallower targets to the half-depth one removes most of the structure-task gain (Secs. 5.4 and 5.7).

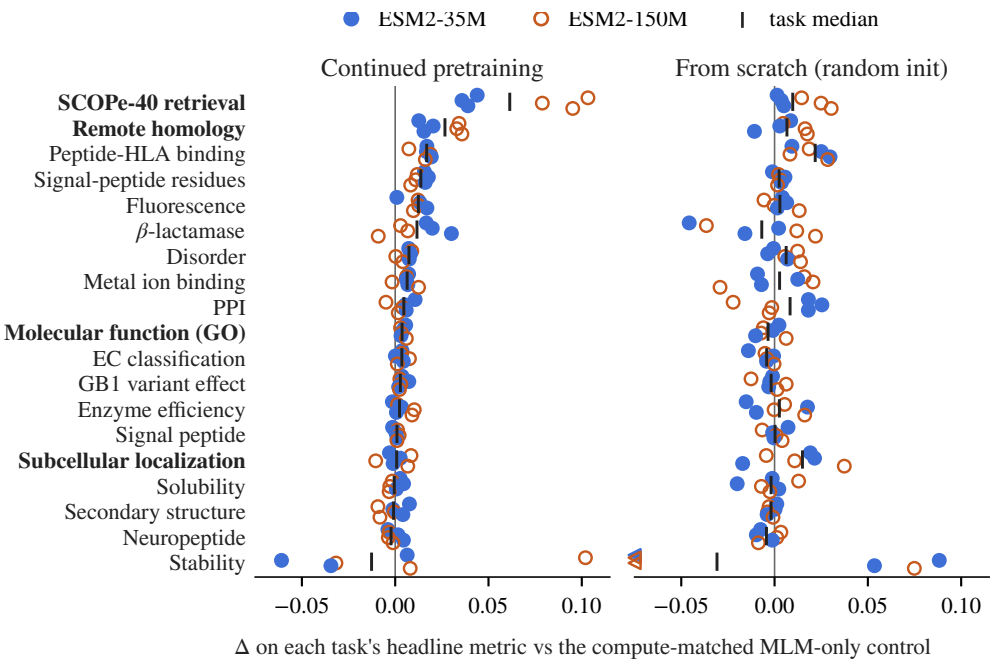

Figure 1: Controlled study: MLM+JEPA minus compute-matched MLM-only on each task's headline metric; positive values favor MLM+JEPA. Each row shows three pretraining seeds at each scale (35M filled, 150M hollow) and their median (black tick), ordered by the continued-pretraining median. Bold labels mark the structure/homology subset (Sec. 4.2). Triangles mark three off-scale from-scratch stability differences (-0.11 to -0.25). Metrics are defined in Table 12.

2 Related Work

Protein language models. Protein language models rely predominantly on token prediction, masked in ESM2 (Lin et al., 2023), ProtTrans (Elnaggar et al., 2020), and AMPLIFY (Fournier et al., 2026) and causal in ProGen2 (Nijkamp et al., 2022). ProteinBERT (Brandes et al., 2022; Michael-Pitschaze et al., 2024), SaProt (Su et al., 2024), and ESM-3 (Hayes et al., 2025) add function or structure channels. Our question is different: *what does a self-supervised latent target add to protein language model pretraining, and is it worth its compute overhead?*

Joint-embedding predictive learning. I-JEPA (Assran et al., 2023) and V-JEPA (Bardes et al., 2024) established latent prediction in vision and video; LLM-JEPA (Huang et al., 2025) and JEPA-DNA (Larey et al., 2026) extend the idea to language and DNA. A latent loss can be satisfied by mapping every input to the same vector (collapse); JEPA methods prevent this with a slowly updated target encoder or a regularizer that keeps embeddings spread out, such as VICReg (Bardes et al., 2022), SIGReg (Balestriero & LeCun, 2025), and VISReg (Wu et al., 2026).

Latent targets for proteins. Continued pretraining further trains an existing model with its original objective (Gururangan et al., 2020) or with new task-specific objectives (Schmirler et al., 2024). Targets other than masked tokens used to pretrain protein models include functional annotations, distillation, and contrastive fine-tuning (Geffen et al., 2022; Ofer et al., 2026; Brandes et al., 2022; Heinzinger et al., 2022); we instead add a data2vec-style EMA-teacher target (Baevski et al., 2022) to MLM/NTP and, in the wall-clock study, a teacher-free variant regularized with SIGReg inspired by LeJEPA (Balestriero & LeCun, 2025) (Sec. 3).

3 METHOD

ProteinJEPA adds a latent prediction loss to MLM at the masked positions:

$$\mathcal{L} = \mathcal{L}_{\text{MLM}} + \lambda\, \mathcal{L}_{\text{JEPA}}^{\text{masked}} + \alpha\, \mathcal{L}_{\text{reg}}, \tag{1}$$

where $\mathcal{L}_{\text{MLM}}$ is masked-token cross-entropy and $\mathcal{L}_{\text{JEPA}}^{\text{masked}}$ averages one minus cosine similarity between student predictions and targets from the *unmasked* input. Both losses use the same masked positions. The student (the model being trained) passes its final-layer hidden states through a small feed-forward predictor with SwiGLU gating; predictions and targets are layer-normalized before comparison. $\mathcal{L}_{\text{reg}}$ discourages collapse, and λ and α weight the latent and regularizer losses (Fig. 2).

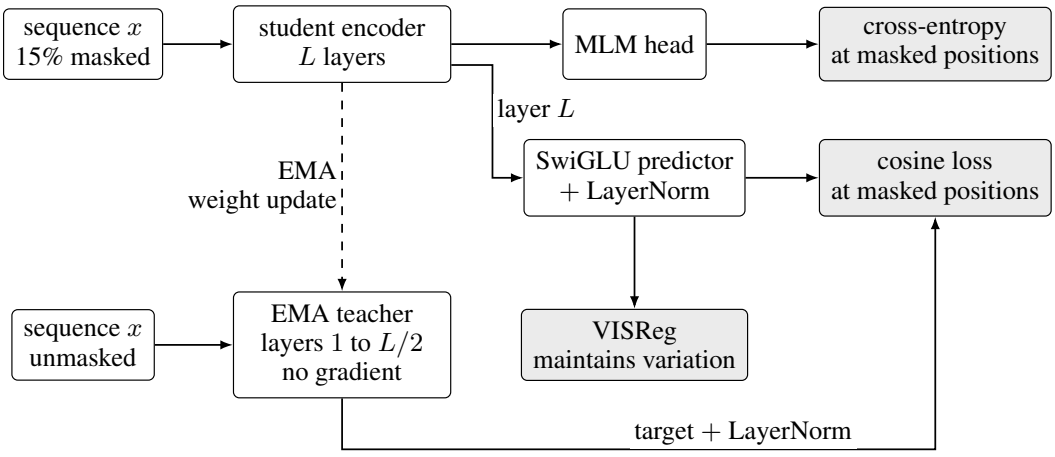

Figure 2: Controlled ProteinJEPA continuation. The student learns from token prediction and a half-depth latent target from an EMA teacher. Both prediction losses use masked positions. VISReg keeps predicted embeddings diverse; the dashed arrow updates teacher weights, without backpropagation.

Why a half-depth target? MLM trains the final layer to recover residue identity. Intermediate layers of protein language models can carry more structure and homology information than the final layer (Ofer et al., 2021; Michael-Pitschaze et al., 2024; Siji et al., 2026; Valeriani et al., 2023; Vig et al., 2020). We therefore predict the half-depth hidden state of the teacher model (layer 6 of 12 at 35M, 15 of 30 at 150M). The teacher is updated by averaging its weights with the student's after each optimizer step and receives the original (unmasked) full input sequence; gradients do not pass through the teacher. Truncating the target layer also reduces its overhead: 17% more compute than MLM-only, compared with about 37% for a last-layer target.

Regularization. The controlled study combines the EMA teacher (Baevski et al., 2022) with VISReg (variance–invariance–sketching regularization; Wu et al., 2026). As a latent loss alone

can be minimized by predicting the same vector for every input (collapse), VISReg prevents this by projecting the predicted embeddings onto random directions and penalizing each projection's sliced-Wasserstein distance from a standard Gaussian, so predictions must stay spread out; SIGReg (sketched isotropic Gaussian regularization; Balestriero & LeCun, 2025; Maes et al., 2026) applies the same regularization with the Epps–Pulley statistic.

The wall-clock study instead uses the student's own gradient-blocked last layer as the target, with SIGReg and per-backbone hyperparameters listed in Table 14.

Ablated objectives. The wall-clock study also includes **all-position MLM+JEPA**, which applies the latent loss to all non-padding positions, and **JEPA-only**, which removes MLM entirely. Within the controlled study, the **ladder** variant uses targets at layers 3–6 of 12 (35M) or 7, 9, 11, 13, and 15 of 30 (150M): the half-depth target plus shallower ones. Teacher, regularizer, loss weight, and schedule are otherwise unchanged.

4 Experimental Setup

We use two complementary studies. The **controlled study**, ESM2 against step- and compute-matched MLM-only controls, is the primary evidence; the **wall-clock study** (Sec. 5.5) covers five backbones at a fixed 8 h budget. Table 1 lists their differences, and Sec. 5.6 applies the controlled recipe to a causal model. Throughout, an *arm* is one training condition, such as MLM-only or MLM+JEPA.

Table 1: The two studies. Both headline recipes use MLM plus a cosine latent loss at masked positions; they differ in matching, collapse control, and target layer. λ and α weight the latent and regularizer terms of Eq. 1.

	Controlled study	Wall-clock study (Sec. 5.5)
Budget matched on	steps, and separately compute	wall-clock, 8 h
Backbones	ESM2-35M, ESM2-150M	five backbones
Pretraining seeds	three, pooled	pooled, 5–8 runs per arm
Corpus	UniRef50	UniRef30
Target	EMA teacher, half depth	last layer, gradients blocked
EMA decay	$0.996 \rightarrow 0.9995$	— (no teacher network)
Collapse control	VISReg, α=0.008	SIGReg, 256 projections
Predictor / λ	1-layer SwiGLU, 1.0	2-layer SwiGLU, 0.45
Masking rate	15 % cont., 20 % scratch	25 % ESM2, 20 % other

4.1 Backbones

The controlled study uses ESM2 at 35M and 150M parameters (Lin et al., 2023), each either continued from its off-the-shelf checkpoint or trained from random initialization. An ESM2-8M run with the same recipe (one pretraining seed) adds a third size (Appendix A.2). We adapt the recipe to ProGen3-112m (Bhatnagar et al., 2025), a causal next-token-prediction model (Sec. 5.6). The wall-clock study uses ESM2-35M, ESM2-150M, AMPLIFY-120M (Fournier et al., 2026) and ProteinBERT2-35M, a ProteinBERT-style encoder with a ModernBERT backbone (Brandes et al., 2022; Warner et al., 2024).

4.2 Training and evaluation setup

Training. Sequences are truncated or padded to 512 tokens. Both arms of every MLM contrast share the masking rate: 15% for continued runs and 20% from scratch in the controlled study, and the per-backbone rates in Table 14 for the wall-clock study. Runs use AdamW with weight decay 0.01, bfloat16 mixed precision, and FlashAttention-2 (Dao, 2023); hardware details are in Appendix A.11.

The controlled ESM2 study trains on UniRef50 (Suzek et al., 2015), ESM2's original training corpus, for 67,000 steps with an effective batch size of 1024—one pass over the corpus. Peak learning rates are 2×10^{-4} for continuation and 1×10^{-3} for training from scratch (Geiping & Goldstein, 2022), reached after 2,000 warmup steps and decayed by a cosine schedule to 1% of peak. The latent

branch switches on after an MLM-only warm start: at step 2,000 with a 1,000-step ramp on λ for continuation, and at step 20,000 with an 8,000-step ramp from scratch, allowing the model to learn from MLM before introducing latent targets.

The wall-clock study trains on UniRef30 (Mirdita et al., 2017) with a constant schedule after 1,000 warmup steps; the headline grid uses checkpoints at $\{1, 4, 8\}$ hours. Per-checkpoint optimizer-step, sample, and token counts for every run are in Appendix A.7.

Step and compute matching. The latent branch requires an additional target-encoder forward pass. Under *step matching*, both arms take 67,000 optimizer steps; the half-depth MLM+JEPA recipe then uses 17% more compute, measured as wall-clock time per step on the same hardware. Under *compute matching*, MLM-only receives 11,390 additional steps at its final learning rate (1% of peak; 2×10^{-6} for continuation) to make up that 17%.

Both arms are retrained at each seed, so each seed is an independent replication. Within each pair, optimizer, schedule, learning rate, batch, masking rate, and data seed are matched; the compute-matched control differs only by its extra steps.

Benchmark tasks. The tasks fall into five groups. *Fold and function*: SCOPe-40 retrieval (Fox et al., 2014) (retrieve a protein of the same structural family, defined in the Structural Classification of Proteins, SCOPe; family fixes fold and superfamily), remote homology (fold class), molecular function (Gene Ontology, GO, terms), EC classification (Enzyme Commission numbers) and subcellular localization. *Binding*: peptide–HLA binding, metal-ion binding, and protein–protein interaction (PPI). *Activity and variant effects*: enzyme catalytic efficiency, β-lactamase activity, GB1 variant effect (binding fitness), fluorescence, and stability. *Sequence properties*: solubility, signal peptide, and neuropeptide precursor (Karsenty et al., 2014). *Residue level*: eight-state secondary structure, disorder, and signal-peptide residues. Tasks are drawn from TAPE (Rao et al., 2019), PEER (Xu et al., 2022), FLIP (Dallago et al., 2021) and ProteinBERT (Brandes et al., 2022), with further details in Appendix A.9, including metrics, and per-task sources.

Evaluation. The controlled analysis uses 19 tasks: 18 linear probes (logistic or ridge regression on frozen final-layer embeddings of the trained encoder), mean-pooled for the 15 protein-level tasks and per residue for the three residue-level ones, plus zero-shot SCOPe-40 retrieval (Hubbard et al., 1999; Fox et al., 2014). Recall@1 is the fraction of queries whose nearest retrieved protein is in the same SCOPe family, using cosine similarity of L2-normalized mean-pooled embeddings. The wall-clock study uses a separate 19-task suite; Table 12 lists both. To assess the half-depth-target hypothesis, we separately summarize SCOPe-40 retrieval, remote homology, molecular function, and subcellular localization as a structure/homology-sensitive subset.

Statistical reporting. Each task contributes one fixed headline metric, with higher scores indicating better performance. We define Δ as the first-named arm minus its comparison arm and summarize a contrast by wins, losses, and ties (W/L/T), using $|\Delta| < 0.002$ as the tie band, and report one-sided exact sign tests (alternative $\Delta > 0$) together with one-sided Wilcoxon signed-rank tests for the controlled study. Table 2 pools the two model scales and three pretraining seeds (n=114). Because these 114 comparisons reuse the same tasks, they are not independent; per-seed, per-scale, and task-averaged tests support the continued-pretraining result (Appendix A.1).

For the wall-clock study, task-by-backbone comparisons use two-sided t-tests over independent pretraining runs, with Benjamini–Hochberg correction across all 95; arms that share pretraining seeds are paired (Appendix A.11).

5 Results

5.1 MLM+JEPA improves continued training under step and compute matching

After one epoch of continued pretraining, MLM+JEPA beats step-matched MLM-only on 76 of 114 task-by-run comparisons, with 15 losses and 23 ties (p<0.0001), and the compute-matched control on 78, with 14 losses and 22 ties (p<0.0001; Table 2). The extra MLM updates do not close the gap. The result replicates in every pretraining seed ($p \leq 0.0004$ per seed, 38 comparisons each; Appendix A.1).

The latent term does not improve MLM loss: training and validation MLM losses are about 0.002 nats higher with MLM+JEPA at both scales and 0.005 nats higher at 8M, and the extra compute-matching steps change validation loss by at most 0.005 nats (Appendix A.1, Figs. 4 and 5; Table 5).

Table 2: Controlled study: ESM2-35M and 150M, three pretraining seeds, pooled (19 tasks $\times$ 2 scales $\times$ 3 seeds $=$ 114 comparisons per row; the ladder row is one seed and omits EC, 18 tasks $\times$ 2 scales $=$ 36). Δ is the first-named arm minus the second and the median is over the row's comparisons; W/L/T counts use $|\Delta| < 0.002$ as the tie band; p-values are one-sided (alternative $\Delta > 0$). Step-matched MLM-only takes the same 67,000 steps as MLM+JEPA; compute-matched MLM-only receives 11,390 extra MLM steps, matching MLM+JEPA's measured 17% overhead.

Contrast	W/L/T	Median Δ	Sign p	Wilcoxon p
Continued pretraining				
MLM+JEPA vs. step-matched MLM-only	76/15/23	$+0.0051$	<0.0001	<0.0001
MLM+JEPA vs. compute-matched MLM-only	78/14/22	$+0.0049$	<0.0001	<0.0001
MLM+JEPA vs. off-the-shelf checkpoint	81/20/13	$+0.0068$	<0.0001	<0.0001
Step-matched MLM-only vs. off-the-shelf checkpoint	52/35/27	$+0.0012$	0.043	0.15
Ladder (extra shallow targets) vs. single half-depth target	11/15/10	-0.0003	0.84	0.82
From scratch (random init)				
MLM+JEPA vs. step-matched MLM-only	50/41/23	$+0.0013$	0.20	0.036
MLM+JEPA vs. compute-matched MLM-only	54/38/22	$+0.0013$	0.059	0.035

Improvement over the pretrained checkpoint. MLM+JEPA is 81/20/13 against the off-the-shelf ESM2 checkpoint ($p<0.0001$, median $+0.0068$), while MLM-only continuation is 52/35/27 with a median of only $+0.0012$. At 150M, for example, SCOPe-40 Recall@1 is 0.424 off the shelf, 0.404 after compute-matched MLM-only continuation, and 0.507 with MLM+JEPA.

5.2 The largest gains are on structure- and homology-sensitive tasks

On the structure/homology subset defined in Sec. 4.2, the median gain over step-matched MLM-only across both scales and seeds is $+0.0132$ versus $+0.0045$ on the remaining tasks, roughly a threefold difference. Against the compute-matched control the corresponding medians are $+0.0106$ and $+0.0041$ (Fig. 1). Most of the subset gain comes from the two fold tasks: across the six runs the median gain over the compute-matched control is $+0.0614$ on SCOPe-40 retrieval and $+0.0267$ on remote homology, while molecular function gains little but in every run ($+0.0037$) and subcellular localization does not change ($+0.0008$). The next-largest gains are outside the subset: peptide–HLA binding ($+0.0169$), signal-peptide residues ($+0.0137$), fluorescence ($+0.0124$), and β-lactamase ($+0.0117$), each positive in at least five of six runs. The subset gain is larger at 150M than at 35M: median $+0.026$ versus $+0.008$ (step-matched, four tasks, three seeds). Against the compute-matched control, SCOPe-40 gains $+0.015$, $+0.044$, and $+0.103$ at 8M, 35M, and 150M, and remote homology -0.008, $+0.013$, and $+0.034$; the median over the other tasks does not grow ($+0.007$, $+0.007$, $+0.003$; Appendix A.2). The gain concentrates with scale rather than spreading.

Five of 19 tasks beat the compute-matched control beyond the tie band in all six scale$\times$seed runs: SCOPe-40 retrieval, remote homology, molecular function, peptide–HLA binding, and signal-peptide residues. No task loses in all six. On the three residue-level tasks, the continued JEPA arms of every controlled backbone (ESM2-8M, 35M, 150M, and ProGen3) beat the compute-matched control in 19 of 24 comparisons (Table 6); this gain is largest at 8M and shrinks with size (Appendix A.2).

The SCOPe-40 gain changes little when retrieval is scored after mean-centering or, in the wall-clock study, with Euclidean or centered-and-whitened cosine distance (Appendix A.1); the gain therefore persists beyond the cosine similarity used in training.

5.3 No reliable gain from random initialization

MLM+JEPA does not produce a reliable improvement across the 19 tasks from random weights. Pooled over the six scale$\times$seed runs, it is 50/41/23 against the step-matched control and 54/38/22 against the compute-matched one (sign test p=0.20 and 0.059; Wilcoxon 0.036 and 0.035), and the

result does not replicate across seeds (Appendix A.1). The structure/homology subset trends in the same direction as continued pretraining (median $+0.0059$ versus $+0.0132$ continued, step-matched; p=0.021 uncorrected).

5.4 Additional shallow targets reduce the structural advantage

The ladder ablation keeps the half-depth target and adds shallower targets. Against the half-depth target, the ladder records 11 wins, 15 losses, and 10 ties across 18 tasks at two scales (p=0.84). It loses all eight comparisons on the structure/homology subset, reducing the median gain over step-matched MLM-only from $+0.013$ to $+0.004$.

5.5 The wall-clock study

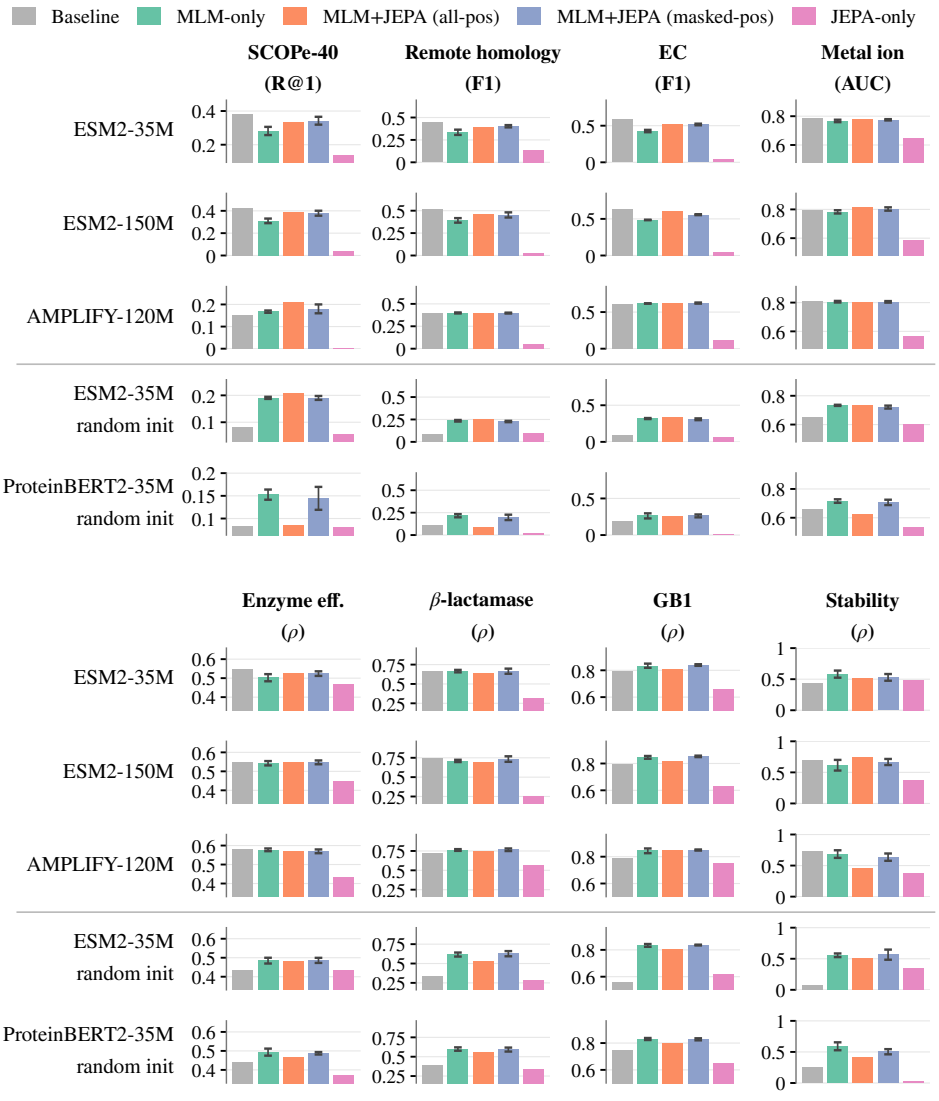

Figure 3: Wall-clock study: eight tasks across five backbones. Error bars: ± 1 standard deviation across pretraining seeds for MLM-only and masked-position MLM+JEPA; other conditions use one run or checkpoint. Baseline: pretrained checkpoint or untrained model. R@1: Recall@1; ρ: Spearman; F1: macro-F1.

The pretrained ESM2 backbones favor MLM+JEPA (10/4/5 at 35M and 12/4/3 at 150M), AMPLIFY-120M is near parity (6/5/8), and the two random-init backbones are at parity or worse (6/8/5 for ESM2-35M, 4/10/5 for ProteinBERT2-35M); Fig. 3 shows eight of the 19 tasks and Appendix Table 11 all of them. Across the 95 task-by-backbone comparisons, ten differences survive Benjamini–Hochberg correction, all positive: SCOPe-40 retrieval, remote homology, and EC classification at both ESM2 scales; enzyme catalytic efficiency, metal-ion binding, and secondary structure at 35M; and disorder on random-init ESM2-35M. Five remain significant under the stricter Holm correction; no negative comparison survives either. On pretrained ESM2 both trained arms end 8 h below the checkpoint on the macro-mean, MLM+JEPA by less (Appendix Table 7). Under wall-clock matching, MLM+JEPA gains occur despite fewer optimizer steps and token exposures than MLM-only (Table 10).

5.6 Next-token prediction shows the same pattern

To test whether the gain extends beyond masked language modeling, we adapt the controlled recipe to ProGen3-112m, a causal mixture-of-experts model trained by next-token prediction (NTP). The latent loss predicts the teacher's half-depth state four positions ahead of the student rather than at masked positions (Appendix A.10). On the 16 tasks scored for this model, continued NTP+JEPA beats step-matched NTP-only on 11 (4 losses) and compute-matched NTP-only on 12 (3 losses; p=0.018). Wins include SCOPe-40 retrieval, remote homology, subcellular localization, and all three residue-level tasks (Table 3); token-prediction loss does not improve (Appendix A.10). Both arms improve over the checkpoint, with median gains of $+0.017$ for NTP+JEPA and $+0.007$ for NTP-only. From random initialization, NTP+JEPA scores 10/3/3 against either control (sign-test p=0.046).

Table 3: Next-token prediction on ProGen3-112m, 16 tasks. Notation as in Table 2; compute-matched NTP-only receives 14,070 extra steps, matching NTP+JEPA's 21% overhead; per-task scores in Appendix A.10.

Contrast	W/L/T	Median Δ	Sign p	Wilcoxon p
Continued pretraining				
NTP+JEPA vs. step-matched NTP-only	11/4/1	$+0.0072$	0.059	0.017
NTP+JEPA vs. compute-matched NTP-only	12/3/1	$+0.0084$	0.018	0.0026
NTP+JEPA vs. off-the-shelf checkpoint	13/1/2	$+0.0169$	0.00092	0.0017
Step-matched NTP-only vs. off-the-shelf checkpoint	12/3/1	$+0.0073$	0.018	0.0055
From scratch (random init)				
NTP+JEPA vs. step-matched NTP-only	10/3/3	$+0.0061$	0.046	0.088
NTP+JEPA vs. compute-matched NTP-only	10/3/3	$+0.0056$	0.046	0.058

5.7 Loss form, target positions, and removing MLM

We ablate three recipe choices: the form of the latent loss, the positions it supervises, and whether MLM is needed at all. A fixed-step ESM2-35M factorial with the wall-clock target and regularizer (Table 1) crossed masked-position versus all-position latent targets with cosine, mean squared error (MSE), and SmoothL1 losses at 25,000 steps (Appendix Table 8). The loss form, not the choice of target positions, is what matters: masked-position and all-position targets are indistinguishable (14/16/18, p=0.856, two-sided), whereas cosine gives the highest mean downstream scores and beats MSE (16/5/11, p=0.027); its difference from SmoothL1 is not significant.

JEPA-only causes large downstream losses on every backbone in the wall-clock study (Appendix Table 7) and on ESM2-35M with the controlled recipe (Appendix A.1), so latent prediction complements rather than replaces token prediction.

6 Discussion and Limitations

Latent prediction improves pretrained models' downstream performance in every seed and at every size tested while its token-prediction loss is 0.002–0.005 nats worse (Sec. 5.1), i.e., language modeling loss does not necessarily track downstream quality in Biology. The concentration of gains on SCOPe-40 retrieval and remote homology supports the hypothesis that intermediate-layer targets help retain

structural information; direct layer-wise measurements are needed to establish that mechanism. This echoes observations that some biological tasks do not improve with model or pretraining data scale, while structure-related protein tasks do (Brandes et al., 2022; Breton et al., 2026; Leib et al., 2026; Li et al., 2024).

The size trend points the same way. Larger checkpoints hold more structural information to begin with (off-the-shelf SCOPe-40 Recall@1 rises from 0.275 at 8M to 0.424 at 150M), continued MLM-only training loses more of it at the larger sizes, and the latent gain over the compute-matched control grows in step, two- to threefold with each roughly fourfold increase in parameters, while median gain on the other tasks does not grow (Table 4). Latent prediction therefore adds most where the model already has a strong representation, and adds little from random initialization, where there is no meaningful representation to update. Three sizes do not establish a scaling law, but the monotone trend suggests one; whether it continues past 150M, and whether more data or larger models close the from-scratch gap, are open.

The controlled recipe ran one epoch over UniRef50 and frozen-representation evaluation only; the wall-clock study broadens the architecture coverage under wall-clock matching alone and UniRef30. We did not test larger models or datasets (such as UniRef100). The EMA teacher and the VISReg regularizer were adopted together; neither was ablated on its own.

7 Conclusion

Latent prediction improves protein language model pretraining when combined with token prediction. The gain holds under both step and compute matching, reproduces across seeds and model sizes, transfers to a causal model, and is largest on structure- and homology-sensitive tasks, where it grows with scale. At 17% extra compute it returns more than the same compute spent on pure MLM. The recipe uses a half-depth EMA-teacher target with a cosine loss at masked positions. Ablations show that cosine beats mean squared error, adding shallower targets reduces structural gains, and removing MLM harms task performance. ProteinJEPA therefore complements rather than replaces MLM and NTP, and is most useful for improving a pretrained model.

Ethics Statement

This work changes a self-supervised training objective and evaluates representations on public benchmarks. As with other biological models, it may support both beneficial and dual-use applications.

Reproducibility Statement

The anonymized repository `https://anonymous.4open.science/r/protJepa-FF24` contains code, benchmark harness, and analysis scripts. Recipes, pretraining seeds, matching criteria, and per-task metrics are documented in Secs. 3–4 and the appendix. Model checkpoints will be released after de-anonymization.

AI Use Statement

Generative AI tools assisted with software development, result interpretation, and manuscript proofing. The authors reviewed the code, analyses, claims, prose and take responsibility for the final content.

References

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture, April 2023. URL `http://arxiv.org/abs/2301.08243`. arXiv:2301.08243 [cs].

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language. In *Proceedings*

of the 39th International Conference on Machine Learning, pp. 1298–1312. PMLR, June 2022. URL https://proceedings.mlr.press/v162/baevski22a.html.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and Scalable Self-Supervised Learning Without the Heuristics, November 2025. URL http://arxiv.org/abs/2511.08544. arXiv:2511.08544 [cs].

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning, January 2022. URL http://arxiv.org/abs/2105.04906. arXiv:2105.04906 [cs.CV].

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting Feature Prediction for Learning Visual Representations from Video, February 2024. URL http://arxiv.org/abs/2404.08471. arXiv:2404.08471 [cs].

Judith Bernett, David B Blumenthal, and Markus List. Cracking the black box of deep sequence-based protein–protein interaction prediction. *Briefings in Bioinformatics*, 25(2):bbae076, January 2024. ISSN 1467-5463, 1477-4054. doi: 10.1093/bib/bbae076. URL https://academic.oup.com/bib/article/doi/10.1093/bib/bbae076/7621029.

Aadyot Bhatnagar, Sarthak Jain, Joel Beazer, Samuel C. Curran, Alexander M. Hoffnagle, Kyle S. Ching, Michael Martyn, Stephen Nayfach, Jeffrey A. Ruffolo, and Ali Madani. Scaling Unlocks Broader Generation and Deeper Functional Understanding of Proteins, October 2025. URL https://www.biorxiv.org/content/10.1101/2025.04.15.649055v2. ISSN: 2692-8205 Pages: 2025.04.15.649055 Section: New Results.

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, April 2022. ISSN 1367-4803. doi: 10.1093/bioinformatics/btac020. URL https://doi.org/10.1093/bioinformatics/btac020.

Lola Le Breton, David Heurtel-Depeiges, Douglas C. Millar, Lara E. Zetzsche, Robert M. Vernon, Christopher James Langmead, Sarath Chandar, and Quentin Fournier. pLM representations unlock metagenomic space beyond homology, July 2026. URL https://www.biorxiv.org/content/10.64898/2026.07.28.739874v1.

Yao Cheng, Haobo Wang, Hua Xu, Yuan Liu, Bin Ma, Xuemin Chen, Xin Zeng, Xianghe Wang, Bo Wang, Carina Shiau, Sergey Ovchinnikov, Xiao-Dong Su, and Chu Wang. Co-evolution-based prediction of metal-binding sites in proteomes by machine learning. *Nature Chemical Biology*, 19(5):548–555, May 2023. ISSN 1552-4469. doi: 10.1038/s41589-022-01223-z. URL https://www.nature.com/articles/s41589-022-01223-z.

Christian Dallago, Jody Mou, Kadina E Johnston, Bruce Wittmann, Nicholas Bhattacharya, Samuel Lucas Goldman, Ali Madani, and Kevin K Yang. FLIP: Benchmark tasks in fitness landscape inference for proteins. preprint, Bioengineering, November 2021. URL http://biorxiv.org/lookup/doi/10.1101/2021.11.09.467890.

Tri Dao. FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning, July 2023. URL http://arxiv.org/abs/2307.08691. arXiv:2307.08691 [cs].

Rupashree Dass, Frans A. A. Mulder, and Jakob Toudahl Nielsen. ODiNPred: comprehensive prediction of protein order and disorder. *Scientific Reports*, 10(1):14780, September 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-71716-1. URL https://www.nature.com/articles/s41598-020-71716-1.

Kieran Didi, Sarah Alamdari, Alex X. Lu, Bruce Wittmann, Kadina E. Johnston, Ava P. Amini, Ali K. Madani, Maya Czeneszew, Christian Dallago, and Kevin K. Yang. FLIP2: Expanding Protein Fitness Landscape Benchmarks for Real-World Machine Learning Applications, February 2026. URL https://www.biorxiv.org/content/10.64898/2026.02.23.707496v3.

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: Towards Cracking the Language of Life's Code Through Self-Supervised Learning. preprint, Bioinformatics, July 2020. URL http://biorxiv.org/lookup/doi/10.1101/2020.07.12.199554.

Quentin Fournier, Robert M. Vernon, Almer van der Sloot, Benjamin Schulz, Sarath Chandar, and Christopher James Langmead. Protein Language Models: Is Scaling Necessary?, January 2026. URL https://www.biorxiv.org/content/10.1101/2024.09.23.614603v2.

Naomi K Fox, Steven E Brenner, and John-Marc Chandonia. SCOPe: Structural Classification of Proteins–extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic acids research*, 42(Database issue):D304–9, January 2014. ISSN 1362-4962. doi: 10.1093/nar/gkt1240. URL http://nar.oxfordjournals.org/content/42/D1/D304.long.

Yaron Geffen, Yanay Ofran, and Ron Unger. DistilProtBert: a distilled protein language model used to distinguish between real proteins and their randomly shuffled counterparts. *Bioinformatics*, 38(Supplement_2):ii95–ii98, September 2022. ISSN 1367-4803. doi: 10.1093/bioinformatics/btac474. URL https://doi.org/10.1093/bioinformatics/btac474.

Jonas Geiping and Tom Goldstein. Cramming: Training a Language Model on a Single GPU in One Day, December 2022. URL http://arxiv.org/abs/2212.14034. arXiv:2212.14034 [cs.CL].

Vladimir Gligorijević, P. Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C. Taylor, Ian M. Fisk, Hera Vlamakis, Ramnik J. Xavier, Rob Knight, Kyunghyun Cho, and Richard Bonneau. Structure-based protein function prediction using graph convolutional networks. *Nature Communications*, 12 (1):3168, May 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-23303-9. URL https://www.nature.com/articles/s41467-021-23303-9.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks, May 2020. URL http://arxiv.org/abs/2004.10964. arXiv:2004.10964 [cs].

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf A. Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, February 2025. doi: 10.1126/science.ads0018. URL https://www.science.org/doi/10.1126/science.ads0018.

Michael Heinzinger, Maria Littmann, Ian Sillitoe, Nicola Bordin, Christine Orengo, and Burkhard Rost. Contrastive learning on protein embeddings enlightens midnight zone. *NAR Genomics and Bioinformatics*, 4(2):lqac043, March 2022. ISSN 2631-9268. doi: 10.1093/nargab/lqac043. URL https://academic.oup.com/nargab/article/doi/10.1093/nargab/lqac043/6605840.

Hai Huang, Yann LeCun, and Randall Balestriero. LLM-JEPA: Large Language Models Meet Joint Embedding Predictive Architectures, October 2025. URL http://arxiv.org/abs/2509.14252. arXiv:2509.14252 [cs].

T. J. P. Hubbard, B. Ailey, S. E. Brenner, A. G. Murzin, and C. Chothia. SCOP: a Structural Classification of Proteins database. *Nucleic Acids Research*, 27(1):254–256, January 1999. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/27.1.254. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/27.1.254.

S. Karsenty, N. Rappoport, D. Ofer, A. Zair, and M. Linial. NeuroPID: a classifier of neuropeptide precursors. *Nucleic Acids Research*, pp. gku363–, May 2014. ISSN 0305-1048. doi: 10.1093/nar/gku363. URL http://nar.oxfordjournals.org/content/early/2014/05/03/nar.gku363.abstract.

Sameer Khurana, Reda Rawi, Khalid Kunji, Gwo-Yu Chuang, Halima Bensmail, and Raghvendra Mall. DeepSol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics*, 34(15):2605–2613, August 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty166. URL https://doi.org/10.1093/bioinformatics/bty166.

Michael Schantz Klausen, Martin Closter Jespersen, Henrik Nielsen, Kamilla Kjærgaard Jensen, Vanessa Isabell Jurtz, Casper Kaae Sønderby, Morten Otto Alexander Sommer, Ole Winther, Morten Nielsen, Bent Petersen, and Paolo Marcatili. NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins: Structure, Function, and Bioinformatics*, 87(6):520–527, 2019. ISSN 1097-0134. doi: 10.1002/prot.25674. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.25674.

Ariel Larey, Elay Dahan, Amit Bleiweiss, Raizy Kellerman, Guy Leib, Omri Nayshool, Dan Ofer, Tal Zinger, Dan Dominissini, Gideon Rechavi, Nicole Bussola, Simon Lee, Shane O'Connell, Dung Hoang, Marissa Wirth, Alexander W. Charney, Nati Daniel, and Yoli Shavit. JEPA-DNA: Grounding Genomic Foundation Models through Joint-Embedding Predictive Architectures, February 2026. URL http://arxiv.org/abs/2602.17162. arXiv:2602.17162 [cs].

Yann LeCun. A path towards autonomous machine intelligence version 0.9.2. *Open Review*, 62(1): 1–62, 2022. URL https://openreview.net/pdf?id=BZ5a1r-kVsf.

Guy Leib, Tal Zinger, Dan Ofer, Raizy Kellerman, Omri Nayshool, Dan Dominissini, Ariel Larey, Jeremy Levy, Yury Nahshan, Elay Dahan, Amit Bleiweiss, Nicole Bussola, Simon Lee, Shane O'Connell, Dung Hoang, Marissa Wirth, Noam D. Beckmann, Alexander W. Charney, Yoli Shavit, Nati Daniel, and Gideon Rechavi. DNT: Diploid Genomic Foundation Model, September 2026. URL https://www.biorxiv.org/content/10.64898/2026.09.05.749576v1. ISSN: 2692-8205 Pages: 2026.09.05.749576 Section: New Results.

Francesca-Zhoufan Li, Ava P. Amini, Yisong Yue, Kevin K. Yang, and Alex Xijie Lu. Feature Reuse and Scaling: Understanding Transfer Learning with Protein Language Models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 27351–27375. PMLR, July 2024. URL https://proceedings.mlr.press/v235/li24a.html.

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. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/10.1126/science.ade2574.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. LeWorldModel: Stable End-to-End Joint-Embedding Predictive Architecture from Pixels, March 2026. URL http://arxiv.org/abs/2603.19312. arXiv:2603.19312 [cs].

Tomer Michael-Pitschaze, Niv Cohen, Dan Ofer, Yedid Hoshen, and Michal Linial. Detecting anomalous proteins using deep representations. *NAR Genomics and Bioinformatics*, 6(1):lqae021, March 2024. ISSN 2631-9268. doi: 10.1093/nargab/lqae021. URL https://doi.org/10.1093/nargab/lqae021.

Milot Mirdita, Lars Von Den Driesch, Clovis Galiez, Maria J. Martin, Johannes Soding, and Martin Steinegger. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Research*, 2017. ISSN 13624962. doi: 10.1093/nar/gkw1081.

Erik Nijkamp, Jeffrey Ruffolo, Eli N. Weinstein, Nikhil Naik, and Ali Madani. ProGen2: Exploring the Boundaries of Protein Language Models, June 2022. URL http://arxiv.org/abs/2206.13517. arXiv:2206.13517 [cs, q-bio].

Dan Ofer and Michal Linial. NeuroPID: a predictor for identifying neuropeptide precursors from metazoan proteomes. *Bioinformatics (Oxford, England)*, 30(7):931–40, March 2014. ISSN 1367-4811. doi: 10.1093/bioinformatics/btt725. URL http://www.ncbi.nlm.nih.gov/pubmed/24336809.

Dan Ofer, Nadav Brandes, and Michal Linial. The language of proteins: NLP, machine learning & protein sequences. *Computational and Structural Biotechnology Journal*, 19:1750–1758, 2021. ISSN 20010370. doi: 10.1016/j.csbj.2021.03.022. URL https://linkinghub.elsevier.com/retrieve/pii/S2001037021000945.

Dan Ofer, Oriel Perets, Michal Linial, and Nadav Rappoport. ProtSent: Protein Sentence Transformers, May 2026. URL http://arxiv.org/abs/2605.06830. arXiv:2605.06830 [cs.LG].

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S. Song. Evaluating Protein Transfer Learning with TAPE. *arXiv:1906.08230 [cs, q-bio, stat]*, June 2019. URL http://arxiv.org/abs/1906.08230. arXiv: 1906.08230.

Alexander Rives, Siddharth Goyal, Joshua Meier, 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. *PNAS*, 2019. doi: 10.1101/622803.

Gabriel J. Rocklin, Tamuka M. Chidyausiku, Inna Goreshnik, Alex Ford, Scott Houliston, Alexander Lemak, Lauren Carter, Rashmi Ravichandran, Vikram K. Mulligan, Aaron Chevalier, Cheryl H. Arrowsmith, and David Baker. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science*, 357(6347):168–175, July 2017. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aan0693. URL https://www.sciencemag.org/lookup/doi/10.1126/science.aan0693.

Karen S. Sarkisyan, Dmitry A. Bolotin, Margarita V. Meer, Dinara R. Usmanova, Alexander S. Mishin, George V. Sharonov, Dmitry N. Ivankov, Nina G. Bozhanova, Mikhail S. Baranov, Onuralp Soylemez, Natalya S. Bogatyreva, Peter K. Vlasov, Evgeny S. Egorov, Maria D. Logacheva, Alexey S. Kondrashov, Dmitry M. Chudakov, Ekaterina V. Putintseva, Ilgar Z. Mamedov, Dan S. Tawfik, Konstantin A. Lukyanov, and Fyodor A. Kondrashov. Local fitness landscape of the green fluorescent protein. *Nature*, 533(7603):397–401, May 2016. ISSN 1476-4687. doi: 10.1038/nature17995. URL https://www.nature.com/articles/nature17995.

Jürgen Schmidhuber and Daniel Prelinger. Discovering Predictable Classifications. *Neural Computation*, 5(4):625–635, July 1993. ISSN 0899-7667. doi: 10.1162/neco.1993.5.4.625. URL https://doi.org/10.1162/neco.1993.5.4.625.

Robert Schmirler, Michael Heinzinger, and Burkhard Rost. Fine-tuning protein language models boosts predictions across diverse tasks. *Nature Communications*, 15(1):7407, August 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-51844-2. URL https://www.nature.com/articles/s41467-024-51844-2.

Aleena Siji, Amir Mohammad Karimi Mamaghan, Ferdinand Kapl, Tobias Höppe, Emmanouil Angelis, Andrea Dittadi, Maurice Brenner, Michael Heinzinger, Karl Henrik Johansson, Kaitlin Maile, Johannes von Oswald, and Stefan Bauer. From Words to Amino Acids: Does the Curse of Depth Persist? March 2026. URL https://openreview.net/forum?id=M9inQg4Fz0.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. SaProt: Protein Language Modeling with Structure-aware Vocabulary, March 2024. URL https://www.biorxiv.org/content/10.1101/2023.10.01.560349v3.

Baris E Suzek, Yuqi Wang, Hongzhan Huang, Peter B McGarvey, and Cathy H Wu. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics (Oxford, England)*, 31(6):926–32, March 2015. ISSN 1367-4811. doi: 10.1093/bioinformatics/btu739. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4375400&tool=pmcentrez&rendertype=abstract.

Felix Teufel, José Juan Almagro Armenteros, Alexander Rosenberg Johansen, Magnús Halldór Gíslason, Silas Irby Pihl, Konstantinos D. Tsirigos, Ole Winther, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nature Biotechnology*, 40(7):1023–1025, July 2022. ISSN 1546-1696. doi: 10.1038/s41587-021-01156-3. URL https://www.nature.com/articles/s41587-021-01156-3.

Vineet Thumuluri, José Juan Almagro Armenteros, Alexander Rosenberg Johansen, Henrik Nielsen, and Ole Winther. DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic Acids Research*, 50(W1):W228–W234, July 2022. ISSN 0305-1048. doi: 10.1093/nar/gkac278. URL https://doi.org/10.1093/nar/gkac278.

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pp. 51234–51252, Red Hook, NY, USA, December 2023. Curran Associates Inc.

Luiz C. Vieira, Morgan L. Handojo, and Claus O. Wilke. Medium-sized protein language models perform well at transfer learning on realistic datasets. *Scientific Reports*, 15(1):21400, July 2025. ISSN 2045-2322. doi: 10.1038/s41598-025-05674-x. URL https://www.nature.com/articles/s41598-025-05674-x.

Jesse Vig, Ali Madani, Lav R. Varshney, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. BERTology Meets Biology: Interpreting Attention in Protein Language Models. June 2020. URL http://arxiv.org/abs/2006.15222. arXiv: 2006.15222.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference, December 2024. URL http://arxiv.org/abs/2412.13663. arXiv:2412.13663 [cs].

Haiyu Wu, Randall Balestriero, and Morgan Levine. VISReg: Variance-Invariance-Sketching Regularization for JEPA training, June 2026. URL http://arxiv.org/abs/2606.02572. arXiv:2606.02572 [cs.CV].

Yejian Wu, Lujing Cao, Zhipeng Wu, Xinyi Wu, Xinqiao Wang, and Hongliang Duan. CcBHLA: pan-specific peptide–HLA class I binding prediction via Convolutional and BiLSTM features, April 2023. URL https://www.biorxiv.org/content/10.1101/2023.04.24.538196v1. Pages: 2023.04.24.538196 Section: New Results.

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. Technical Report arXiv:2206.02096, arXiv, June 2022. URL http://arxiv.org/abs/2206.02096. arXiv:2206.02096 [cs] type: article.

A Additional Results and Reproducibility

A.1 Controlled study: robustness of the headline contrast

The continued-pretraining result holds under changes to the tie band and scale pooling. Alternative metrics preserve the pooled contrasts but weaken some per-scale tests.

Tie band. The table below compares the $|\Delta| < 0.002$ tie band used in Table 2 with exact ties only. Within each contrast, the three rows are the three pretraining seeds in a fixed order:

	Band 0.002		Exact ties only	
Contrast	W/L/T	p	W/L/T	p
Continued vs. step-matched	26/5/7	0.0001	28/10/0	0.0025
	25/6/7	0.0004	29/9/0	0.0008
	25/4/9	0.0001	31/7/0	0.0001
Continued vs. compute-matched	27/5/6	0.0001	30/8/0	0.0002
	26/5/7	0.0001	30/8/0	0.0002
	25/4/9	0.0001	32/6/0	$<$0.0001
From scratch vs. step-matched	15/17/6	0.70	17/21/0	0.79
	17/10/11	0.12	25/13/0	0.037
	18/14/6	0.30	20/18/0	0.44
From scratch vs. compute-matched	16/16/6	0.57	18/20/0	0.69
	20/9/9	0.031	23/15/0	0.13
	18/13/7	0.24	21/17/0	0.31
From scratch, pooled	50/41/23	0.20	62/52/0	0.20

Pooling the two scales. Table 2 pools 19 tasks over two scales and three seeds. Because the scales share a task list, we also report two estimators that do not treat scale–task pairs as independent. The per-task means are reported per seed in the same order; the single-scale rows use the first seed:

Contrast (continued)	Estimator	W/L/T	Sign p
vs. step-matched	per-task mean over scales (n=19)	14/1/4	0.0005
		13/2/4	0.0037
		13/1/5	0.0009
vs. compute-matched	per-task mean over scales (n=19)	14/1/4	0.0005
		13/2/4	0.0037
		14/2/3	0.0021
vs. step-matched	35M alone (n=19)	13/2/4	0.0037
vs. step-matched	150M alone (n=19)	13/3/3	0.011

Retrieval metric. Mean-centering SCOPe-40 embeddings leaves the controlled-study gain nearly unchanged: $+0.0431 \rightarrow +0.0403$ at 35M and $+0.1052 \rightarrow +0.1001$ at 150M. In the wall-clock study, Euclidean and centered-and-whitened cosine scoring move the masked-position gain by less than one percentage point.

Alternative evaluation metrics. The wall-clock study fixed a different metric for nine shared tasks (Table 12). Using those metrics instead, the pooled and per-task-averaged headline contrasts are unchanged. Per-scale contrasts are more sensitive and can lose significance at α=0.05 under those metrics.

JEPA-only at the controlled recipe. The JEPA-only evidence elsewhere in this appendix comes from the wall-clock study, whose runs used an all-position target and different regularization (Table 9). Repeating the ablation at the controlled recipe — masked-position cosine with the EMA teacher,

15% masking, 67,000 steps, continued from pretrained ESM2-35M with the MLM weight set to zero — also degrades downstream performance. Against the matched MLM-only control, W/L/T is 2/17/0 with median $\Delta = -0.124$ ($n = 19$), and against the off-the-shelf checkpoint it is 2/17/0 with median $\Delta = -0.129$. The largest losses are EC (-0.481), remote homology (-0.357), and molecular function (-0.320). Its training MLM loss, monitored but not optimized, climbs to 11.4 nats before settling near 4.5 (Fig. 6).

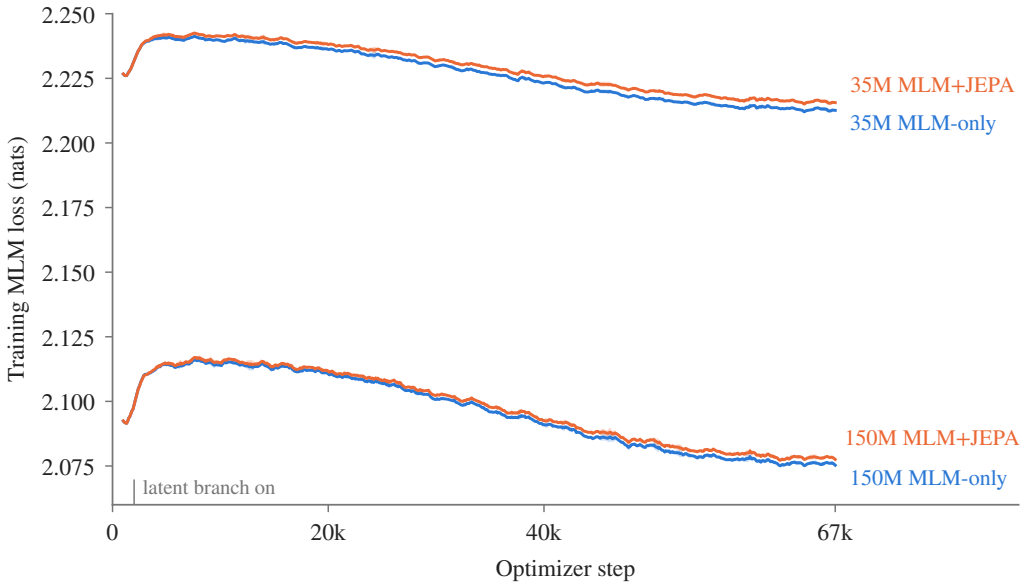

Figure 4: Training MLM loss during continued pretraining (1,000-step moving average; line: mean over the two seeds with complete 67,000-step logs; shading: seed range). Paired arms see the same batches. MLM+JEPA's loss is about 0.002 nats higher at either scale. The latent branch switches on at step 2,000.

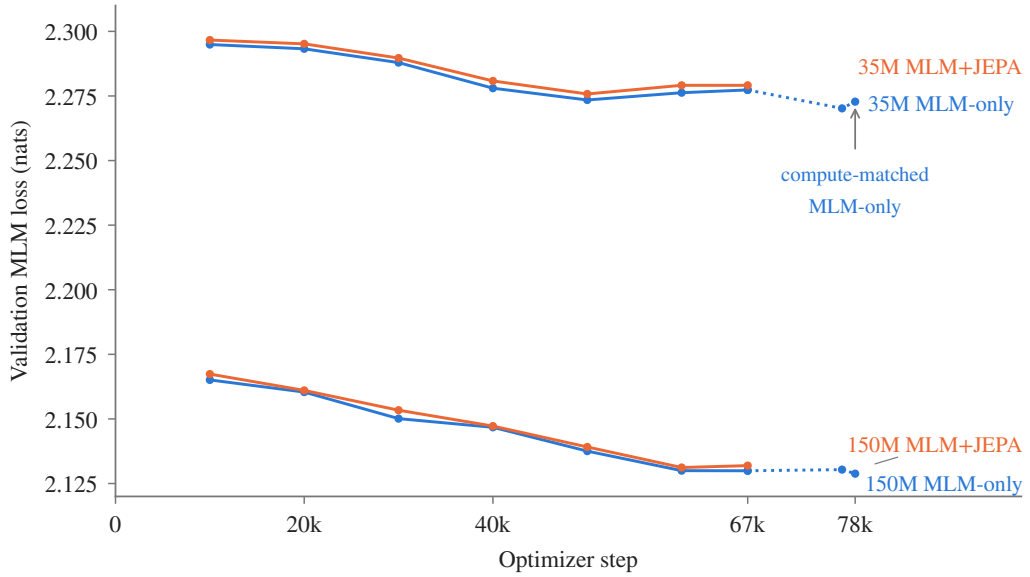

Figure 5: Validation MLM loss (one pretraining seed; 20,000 held-out UniRef50 sequences, every 10,000 steps). Dotted: the compute-matched MLM-only control, 11,390 further steps at the final learning rate, changing loss by at most 0.005 nats. MLM+JEPA's loss is 0.001–0.003 nats higher than MLM-only's.

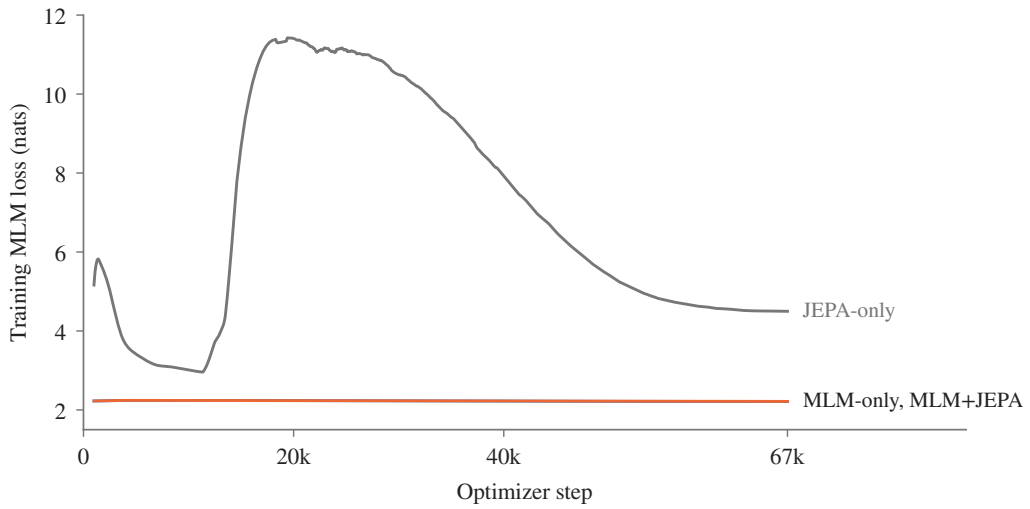

Figure 6: Training MLM loss with JEPA-only continuation (ESM2-35M; MLM weight 0), monitored but not optimized. MLM-only and MLM+JEPA coincide at this scale. JEPA-only reaches 11.4 nats by step 20,000 and settles near 4.5.

A.2 Model-size trend: an ESM2-8M point

To see how the continued-pretraining gain changes with model size we ran the controlled recipe unchanged on ESM2-8M (6 layers, half-depth target at layer 3), including a compute-matched MLM-only control whose 12,080 extra steps come from the 18% overhead measured at this size. Table 4 and Figs. 7–8 compare the three sizes on all 19 tasks; the 35M and 150M entries use one seed so that every size is a single run, and their three-seed ranges appear in Fig. 8. SCOPe-40 retrieval and

remote homology gain more at each step in size, and the 8M point lies below the 35M seed range. The other 17 tasks, including molecular function and subcellular localization, show no such trend. From random initialization no task trends. The three residue-level tasks (Table 6) run the other way: their median gain over the compute-matched control falls from $+0.012$ at 8M to $+0.008$ at 35M and $+0.004$ at 150M. Table 5 lists the 8M held-out losses: MLM+JEPA is 0.003–0.005 nats worse on MLM loss at both initializations, as at 35M and 150M.

Table 4: Model-size trend on the 19 tasks. Δ is MLM+JEPA minus compute-matched MLM-only, one pretraining seed at every size; brackets give the three-seed range at 35M and 150M. The last three rows are absolute SCOPe-40 Recall@1.

	ESM2-8M	ESM2-35M	ESM2-150M
Continued pretraining			
Δ SCOPe-40 retrieval	$+0.015$	$+0.044$ [+0.036, +0.044]	$+0.103$ [+0.079, +0.103]
Δ remote homology	-0.008	$+0.013$ [+0.013, +0.020]	$+0.034$ [+0.033, +0.036]
median Δ, structure/homology (4 tasks)	$+0.005$	$+0.009$	$+0.021$
median Δ, other 15 tasks	$+0.007$	$+0.007$	$+0.003$
W/L/T, all 19 tasks	12/6/1	14/2/3	13/3/3
SCOPe-40 Recall@1, off-the-shelf checkpoint	0.275	0.383	0.424
SCOPe-40 Recall@1, compute-matched MLM-only	0.279	0.368	0.404
SCOPe-40 Recall@1, MLM+JEPA	0.293	0.412	0.507
From random init			
Δ SCOPe-40 retrieval	$+0.011$	$+0.001$	$+0.015$
median Δ, structure/homology (4 tasks)	-0.002	$+0.006$	$+0.000$
median Δ, other 15 tasks	$+0.005$	-0.001	-0.003
W/L/T, all 19 tasks	10/3/6	7/6/6	9/10/0

Table 5: ESM2-8M held-out losses at the end of training (6,810 UniRef50 validation sequences). MLM loss is masked-token cross-entropy in nats; the JEPA loss is one minus cosine similarity at masked positions and is defined only for the MLM+JEPA arms. Compute-matched arms report the end of their 12,080 extra steps.

Arm	Step	Eval MLM loss	Eval JEPA loss
Continued pretraining			
MLM-only, step-matched	67,000	2.391	—
MLM-only, compute-matched	12,080	2.392	—
MLM+JEPA	67,000	2.396	0.072
From random init			
MLM-only, step-matched	67,000	2.445	—
MLM-only, compute-matched	12,080	2.443	—
MLM+JEPA	67,000	2.448	0.051

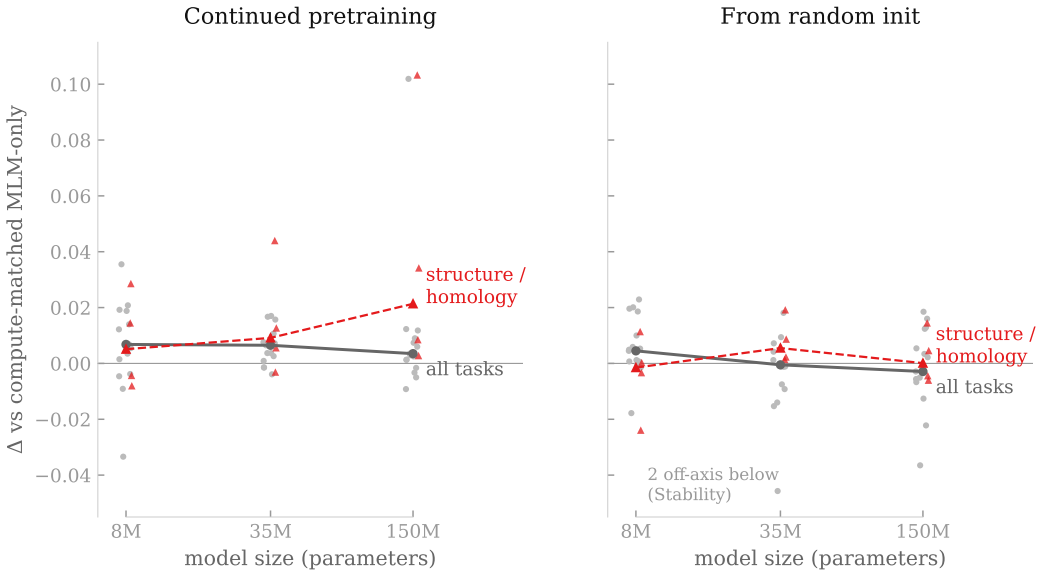

Figure 7: MLM+JEPA minus compute-matched MLM-only against model size (one seed at every size; 19 tasks). Dots are tasks, lines are medians; triangles mark the structure/homology subset. Stability values below -0.055 are clipped to the axis. The subset median rises with size because SCOPe-40 and remote homology do; the other tasks do not.

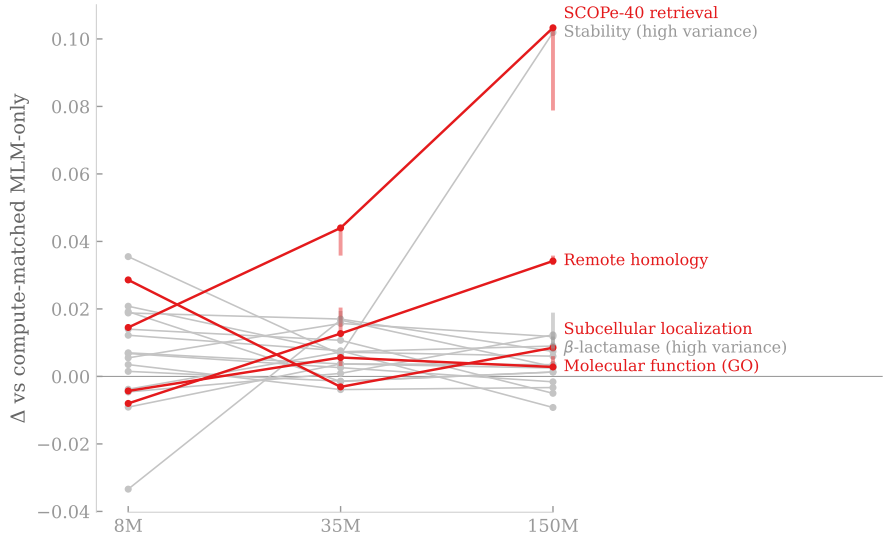

Figure 8: Per-task deltas across sizes, continued pretraining. Vertical ticks at 35M and 150M are three-seed ranges for the five tasks that beat the control in every seed (Sec. 5.2); the 8M SCOPe-40 and remote-homology points fall below the 35M range.

A.3 Per-residue probes

The three residue-level tasks are *eight-state secondary structure* (DSSP labels from NetSurfP-2.0, macro-F1), *intrinsic disorder* (the NetSurfP-2.0 disorder mask, Matthews correlation coefficient, MCC), and *signal peptide* (SignalP 6.0 per-residue signal-peptide labels, macro-F1). Each is a linear probe (a linear head on frozen final-layer residue embeddings), trained on the task's training split and scored on its test split. Table 6 breaks their JEPA-minus-control differences down by backbone,

including ESM2-8M and ProGen3. Under the controlled recipe, the continued JEPA arms beat the step-matched control in 20 of 24 run-by-task comparisons, the compute-matched control in 19, and the off-the-shelf checkpoint in 21 (sign test $p<0.001$ each); from random initialization it is 12/4/8 against the compute-matched control ($p=0.038$). In the wall-clock study the contrast against MLM-only is 21/8/1 on the two pretrained ESM2 backbones ($p=0.012$) and 41/28/9 over all five ($p=0.074$).

Table 6: Per-residue probes: MLM+JEPA (NTP+JEPA for ProGen3) minus its control on three residue-level tasks (eight-state secondary structure, intrinsic disorder, signal peptide), pooled over pretraining runs and tasks. Controlled study: per-backbone rows are against the compute-matched token-prediction control; the last rows of each block pool all four backbones against each control. Wall-clock study: against MLM-only, paired by pretraining seed. Notation as in Table 2; p-values are one-sided.

Backbone	n	W/L/T	Median Δ	Sign p	Wilcoxon p
Controlled study, continued pretraining					
ESM2-8M	3	3/0/0	+0.0121	0.125	0.125
ESM2-35M	9	8/0/1	+0.0076	0.004	0.004
ESM2-150M	9	5/2/2	+0.0041	0.227	0.150
ProGen3-112m	3	3/0/0	+0.0082	0.125	0.125
All, vs compute-matched	24	19/2/3	+0.0079	<0.001	<0.001
All, vs step-matched	24	20/2/2	+0.0074	<0.001	<0.001
All, vs off-the-shelf checkpoint	24	21/2/1	+0.0112	<0.001	<0.001
Controlled study, from random init					
ESM2-8M	3	2/0/1	+0.0045	0.250	0.250
ESM2-35M	9	3/2/4	+0.0003	0.500	0.285
ESM2-150M	9	5/2/2	+0.0021	0.227	0.125
ProGen3-112m	3	2/0/1	+0.0046	0.250	0.125
All, vs compute-matched	24	12/4/8	+0.0020	0.038	0.008
All, vs step-matched	24	10/5/9	+0.0016	0.151	0.028
Wall-clock study, 8 h, vs MLM-only					
ESM2-35M	15	11/4/0	+0.0074	0.059	0.018
ESM2-150M	15	10/4/1	+0.0031	0.090	0.084
AMPLIFY-120M	18	7/7/4	+0.0009	0.605	0.695
ESM2-35M (random init)	15	7/6/2	+0.0010	0.500	0.423
ProteinBERT2-35M	15	6/7/2	−0.0014	0.709	0.820
Pretrained ESM2 (35M and 150M)	30	21/8/1	+0.0040	0.012	0.004
All	78	41/28/9	+0.0022	0.074	0.116

A.4 WALL-CLOCK STUDY: THE FULL OBJECTIVE GRID

This section reports the wall-clock study's two remaining objectives — the all-position variant and JEPA-only — across the full 5×3 matrix of five backbones and three budgets.

Table 7: Wall-clock macro-mean linear-probe Δ relative to each backbone's baseline: 16 protein-level tasks at 8 h, 15 (no EC) at 1 h and 4 h. All-pos MLM+JEPA is the all-position MSE control; Masked-pos MLM+JEPA is the masked-position cosine recipe at 8 h. $\Delta\Delta$ is All-pos MLM+JEPA minus MLM-only; a Holm-corrected paired Wilcoxon test finds no rejection (smallest $p_H = 0.15$). Masked-pos pools pretraining seeds; other trained columns are single runs. Table 8 separates target positions from loss form at fixed steps.

Backbone	Time	MLM-only	All-pos MLM+JEPA	$\Delta\Delta$	JEPA-only	Masked-pos MLM+JEPA
ESM2-35M	1h	-0.011	-0.012	-0.001	-0.235	–
	4h	-0.014	-0.007	$+0.007$	-0.222	–
	8h	-0.014	-0.005	$+0.010$	-0.130	-0.001
ESM2-150M	1h	-0.007	-0.004	$+0.003$	-0.208	–
	4h	-0.016	-0.003	$+0.013$	-0.248	–
	8h	-0.028	-0.009	$+0.019$	-0.246	-0.013
AMPLIFY-120M	1h	-0.004	-0.040	-0.036	-0.181	–
	4h	-0.005	-0.012	-0.007	-0.239	–
	8h	-0.005	-0.015	-0.010	-0.201	$+0.004$
ESM2-35M (random init)	1h	$+0.125$	$+0.121$	-0.004	-0.032	–
	4h	$+0.140$	$+0.139$	-0.001	$+0.036$	–
	8h	$+0.151$	$+0.140$	-0.011	$+0.015$	$+0.149$
ProteinBERT2-35M	1h	-0.007	$+0.003$	$+0.010$	-0.076	–
	4h	$+0.032$	$+0.028$	-0.005	-0.112	–
	8h	$+0.012$	$+0.050$	$+0.038$	-0.127	$+0.098$

All-position MLM+JEPA stays close to MLM-only. The macro-mean gap between all-position MLM+JEPA and MLM-only is $|\Delta\Delta| \leq 0.041$ across all 15 backbone–budget pairs and ≤ 0.036 on pretrained backbones (Table 7). No difference is significant after Holm correction. The masked-position recipe wins most matched 8 h (backbone, task) comparisons against this control ($p \approx 2 \times 10^{-4}$), but also changes MSE to cosine; the fixed-step factorial (Table 8) separates these choices.

A.5 Loss-form and target-position ablation

Table 8: Fixed-step ESM2-35M factorial at 25,000 steps with the wall-clock target and regularizer: initialization $\times$ target positions $\times$ loss form. N counts task–configuration comparisons, pooled as indicated; each score averages validation and test results. Ties are $|\Delta| \leq 0.002$. Reported p-values are two-sided; **bold** marks $p < 0.05$.

Comparison	N	W	L	T	Mean Δ	p_{sign}	p_{Wilcox}	p_t
Loss form (pretrained, positions pooled)								
cosine vs. MSE	32	16	5	11	$+0.005$	$\mathbf{0.027}$	$\mathbf{0.019}$	$\mathbf{0.026}$
cosine vs. SmoothL1	32	10	8	14	$+0.003$	0.815	0.331	0.130
MSE vs. SmoothL1	32	5	8	19	-0.002	0.581	0.295	0.111
Target set (pretrained, losses pooled)								
masked vs. all-position	48	14	16	18	$+0.000$	0.856	0.678	0.670
Initialization (all losses, all positions)								
pretrained vs. random-init	96	96	0	0	$+0.361$	$<\mathbf{10^{-15}}$	$<\mathbf{10^{-15}}$	$<\mathbf{10^{-15}}$

A.6 All-position recipe sweep on pretrained ESM2-35M

A matched-wall-clock sweep on pretrained ESM2-35M screened 37 all-position configurations over target construction, loss, predictor, and regularization (selected rows in Table 9). We selected the gradient-blocked target recipe because it needs no teacher network; a sub-sweep over $\lambda \in \{0.10, 0.25, 0.45, 1.00\}$ selected 0.45. The masked-position cosine recipe was not in this sweep.

Table 9: All-position recipe-sweep results on continued ESM2-35M at matched wall-clock. W/L/T are per task ($n = 13$; 12 for JEPA-only). Mean Δ is relative to matched MLM-only.

Recipe	W/L/T	Mean Δ
Baseline ESM2-35M (no continuation)	5/7/1	$+0.0026$
JEPA + EMA teacher	6/7/0	$+0.0075$
JEPA gradient-blocked target, λ=0.45 (selected)	5/8/0	$+0.0033$
JEPA λ=0.25	5/7/1	-0.0029
JEPA + VICReg (Bardes et al., 2022)	4/9/0	-0.0020
JEPA λ=0.10	3/9/1	-0.0001
JEPA λ=1.00	3/9/1	-0.0020
ESM2 JEPA-only (matched)	2/9/1	-0.0095

A.7 Training coverage and compute

Table 10 lists steps, samples, tokens, and analytic FLOP estimates (sequence length 512) for the wall-clock checkpoints, one run per arm. FLOPs are approximate analytic estimates.

Table 10: Training coverage. Save: checkpoint hour. Units: millions of samples, billions of tokens, and 10^{18} FLOPs.

Model	Objective	Save	Steps	Batch	Samp.	Tok.	FLOPs
ESM2-35M	MLM-only	1h	19,633	128	2.51	1.29	0.258
ESM2-35M	MLM-only	4h	80,107	128	10.25	5.25	1.054
ESM2-35M	MLM-only	8h	160,793	128	20.58	10.54	2.116
ESM2-35M	All-pos MLM+JEPA	1h	10,874	128	1.39	0.71	0.199
ESM2-35M	All-pos MLM+JEPA	4h	43,993	128	5.63	2.88	0.805
ESM2-35M	All-pos MLM+JEPA	8h	88,024	128	11.27	5.77	1.612
ESM2-35M	JEPA-only	1h	10,569	128	1.35	0.69	0.193
ESM2-35M	JEPA-only	4h	43,662	128	5.59	2.86	0.797
ESM2-35M	JEPA-only	8h	87,929	128	11.25	5.76	1.604
ESM2-35M	Masked-pos MLM+JEPA	8h	100,000	128	12.80	6.55	1.831
ESM2-150M	MLM-only	1h	4,893	128	0.63	0.32	0.285
ESM2-150M	MLM-only	4h	22,993	128	2.94	1.51	1.338
ESM2-150M	MLM-only	8h	47,161	128	6.04	3.09	2.745
ESM2-150M	All-pos MLM+JEPA	1h	1,673	128	0.21	0.11	0.132
ESM2-150M	All-pos MLM+JEPA	4h	7,016	128	0.90	0.46	0.554
ESM2-150M	All-pos MLM+JEPA	8h	14,433	128	1.85	0.95	1.140
ESM2-150M	JEPA-only	1h	1,693	128	0.22	0.11	0.133
ESM2-150M	JEPA-only	4h	7,135	128	0.91	0.47	0.563
ESM2-150M	JEPA-only	8h	14,661	128	1.88	0.96	1.156
ESM2-150M	Masked-pos MLM+JEPA	8h	44,894	48	2.15	1.10	1.329
AMPLIFY-120M	MLM-only	1h	7,352	208	1.53	0.78	0.556
AMPLIFY-120M	MLM-only	4h	28,165	208	5.86	3.00	2.129
AMPLIFY-120M	MLM-only	8h	57,414	208	11.94	6.11	4.340
AMPLIFY-120M	All-pos MLM+JEPA	1h	5,194	208	1.08	0.55	0.535
AMPLIFY-120M	All-pos MLM+JEPA	4h	19,705	208	4.10	2.10	2.029
AMPLIFY-120M	All-pos MLM+JEPA	8h	40,135	208	8.35	4.27	4.132
AMPLIFY-120M	JEPA-only	1h	5,281	208	1.10	0.56	0.544
AMPLIFY-120M	JEPA-only	4h	19,994	208	4.16	2.13	2.058
AMPLIFY-120M	JEPA-only	8h	40,644	208	8.45	4.33	4.184
ESM2-35M (random init)	MLM-only	1h	15,950	128	2.04	1.05	0.210
ESM2-35M (random init)	MLM-only	4h	65,305	128	8.36	4.28	0.859
ESM2-35M (random init)	MLM-only	8h	134,413	128	17.20	8.81	1.769
ESM2-35M (random init)	All-pos MLM+JEPA	1h	9,514	128	1.22	0.62	0.174
ESM2-35M (random init)	All-pos MLM+JEPA	4h	38,674	128	4.95	2.53	0.708
ESM2-35M (random init)	All-pos MLM+JEPA	8h	78,818	128	10.09	5.17	1.443
ESM2-35M (random init)	JEPA-only	1h	9,601	128	1.23	0.63	0.175
ESM2-35M (random init)	JEPA-only	4h	38,945	128	4.98	2.55	0.710
ESM2-35M (random init)	JEPA-only	8h	79,330	128	10.15	5.20	1.447
ProteinBERT2-35M	MLM-only	1h	13,931	128	1.78	0.91	0.207
ProteinBERT2-35M	MLM-only	4h	57,071	128	7.31	3.74	0.849
ProteinBERT2-35M	MLM-only	8h	114,591	128	14.67	7.51	1.704
ProteinBERT2-35M	All-pos MLM+JEPA	1h	9,167	128	1.17	0.60	0.190
ProteinBERT2-35M	All-pos MLM+JEPA	4h	37,394	128	4.79	2.45	0.774
ProteinBERT2-35M	All-pos MLM+JEPA	8h	75,019	128	9.60	4.92	1.552
ProteinBERT2-35M	JEPA-only	1h	9,196	128	1.18	0.60	0.190
ProteinBERT2-35M	JEPA-only	4h	37,521	128	4.80	2.46	0.776

Continued on next page

Model	Objective	Save	Steps	Batch	Samp.	Tok.	FLOPs
ProteinBERT2-35M	JEPA-only	8h	75,274	128	9.64	4.93	1.557
ProteinBERT2-35M	Masked-pos MLM+JEPA	8h	64,256	128	8.22	4.21	1.329

A.8 MASKED-POSITION MLM+JEPA ACROSS BACKBONES

Table 11 summarizes the per-backbone contrasts shown in Fig. 3.

Table 11: Masked-position MLM+JEPA versus matched MLM-only continuation after 8 h on 19 tasks. W/L/T counts use $|\Delta| < 0.002$ as the tie band. Median Δ is relative to matched MLM-only and the sign-test p-value is one-sided over wins and losses. Scores for both arms pool pretraining seeds.

Backbone	W	L	T	Median Δ	Sign p
Pretrained continuation					
ESM2-35M	10	4	5	$+0.004$	0.090
ESM2-150M	12	4	3	$+0.006$	0.038
AMPLIFY-120M	6	5	8	-0.001	0.500
Random init (from scratch)					
ProteinBERT2-35M	4	10	5	-0.002	0.971
ESM2-35M	6	8	5	-0.001	0.788

A.9 BENCHMARK TASKS AND SOURCES

Table 12 lists the dataset identifier and the metric fixed in advance for each task in each study. Both suites mix TAPE-derived tasks (Rao et al., 2019) with public benchmark datasets, plus the ProteinBERT-style SignalP binary setup, the NeuroPID neuropeptide benchmark (Karsenty et al., 2014; Ofer & Linial, 2014), and the SCOPe-40 retrieval benchmark.

The two suites are both 19 tasks and overlap in 18 of them: the wall-clock study scores CheZoD disorder and the controlled study scores molecular function (GO) in its place. Nine shared tasks carry different metrics, so their scores are not comparable across studies.

Fold and function. *Remote homology* predicts a protein's fold class; *SCOPe-40* retrieves a protein of the same SCOPe family, with Recall@1 measuring whether the nearest retrieved sequence belongs to that family. *Molecular function* predicts Gene Ontology (GO) activity labels, while *EC classification* predicts Enzyme Commission reaction labels. *Subcellular localization* predicts the cellular compartment in which a protein resides (Thumuluri et al., 2022; Gligorijević et al., 2021).

Binding and interactions. *Peptide–HLA binding* classifies whether a peptide binds a human leukocyte antigen (HLA) molecule (Wu et al., 2023). *Metal-ion binding* predicts whether a protein binds a metal ion (Cheng et al., 2023). *Protein–protein interaction* classifies interacting versus non-interacting protein pairs (Bernett et al., 2024).

Quantitative activity and variant effects. *Enzyme catalytic efficiency* is the benchmark's name for predicting enzyme turnover (k_{cat}). *β-lactamase* predicts measured activity of enzyme variants; *GB1* predicts variant binding-fitness scores; *fluorescence* predicts fluorescent-protein brightness (Sarkisyan et al., 2016); and *stability* predicts resistance to proteolysis (Rocklin et al., 2017; Xu et al., 2022; Dallago et al., 2021; Didi et al., 2026).

Sequence properties. *Solubility* classifies soluble versus insoluble proteins (Khurana et al., 2018). *Signal peptide* predicts the presence of a secretion signal (Teufel et al., 2022), and *neuropeptide precursor* identifies proteins that yield neuropeptides (Karsenty et al., 2014; Ofer & Linial, 2014). The wall-clock-only *CheZoD* task predicts a disorder score derived from NMR chemical shifts (Dass et al., 2020).

Residue-level. *Secondary structure* assigns each residue one of eight DSSP states; *disorder* labels each residue as ordered or disordered (NetSurfP-2.0) (Klausen et al., 2019); and *signal peptide* labels each residue as inside or outside a signal peptide (SignalP 6.0) (Teufel et al., 2022). Each is a linear probe on frozen per-residue embeddings.

Table 12: Benchmark sources and headline metrics. Identifiers are Hugging Face datasets unless a URL is given; – marks an absent task; AP is average precision. EC is multilabel, hence micro-averaged in the controlled suite.

		Metric	
Task	Dataset / source identifier	Controlled	Wall-clock
Remote Homology (Fold)	`biomap-research/fold_prediction`	Accuracy	F1-Macro
SCOPe-40 Retrieval	`tattabio/scope40_test`	Recall@1	Recall@1
Subcellular Localization	`proteinea/deeploc`	F1-Macro	AUC
Molecular Function (GO)	`AI4Protein/GO_MF`	F1-Macro	–
EC Classification	`AI4Protein/EC`	F1-Micro	F1-Macro
Enzyme Catalytic Efficiency	`biomap-research/enzyme_catalytic_efficiency`	Spearman	Spearman
β-lactamase-PEER	`SaProtHub/Dataset-Beta_Lactamase-PEER`	Spearman	Spearman
Variant Effect (GB1)	`biomap-research/fitness_prediction`	Spearman	Spearman
Fluorescence (TAPE)	`cradle-bio/tape-fluorescence`	Spearman	Spearman
Stability	`biomap-research/stability_prediction`	Spearman	Spearman
CheZoD Disorder	`PeptoneLtd/contrastive-finetuning-plms`	–	Spearman
Peptide-HLA Binding	`biomap-research/peptide_HLA_MHC_affinity`	AP	AUC
Metal Ion Binding	`biomap-research/metal_ion_binding`	AP	AUC
Protein–protein interaction (PPI)	`Synthyra/bernett_gold_ppi`	AP	AUC
Solubility (DeepSol)	`proteinea/solubility`	AP	AUC
Signal Peptide Prediction	ProteinBERT SignalP_Binary – `https://github.com/nadavbra/protein_bert`	AP	AUC
Neuropeptide Precursor	ProteinBERT – `https://github.com/nadavbra/protein_bert`	AP	AUC
Residue-level (both studies)			
Secondary Structure (8-state)	`GleghornLab/SS8`	F1-Macro	F1-Macro
Disorder	`agemagician/NetSurfP-SS3`	MCC	MCC
Signal Peptide (per residue)	`SaProtHub/Dataset-Signal-Peptides`	F1-Macro	F1-Macro

A.10 Next-token prediction on ProGen3-112m

Model and objective. ProGen3-112m (Bhatnagar et al., 2025) is a causal protein language model with sparse mixture-of-experts feed-forward blocks (10 layers, hidden size 384, 8 experts, top-2 routing). We use a PyTorch implementation with converted pretrained weights. Half of the training sequences are reversed to train both generation directions. Both arms minimize next-token cross-entropy plus the model's router load-balancing loss. NTP+JEPA replaces the masked-position latent loss of Eq. 1 with a future-position one: the predictor maps the student's final hidden state at position t to the EMA teacher's layer-5 (of 10) state at $t{+}4$, giving the target access to four residues beyond the student's prefix. Predictor, loss, EMA schedule, VISReg, and λ follow the controlled study.

Training. Training follows the controlled study (Sec. 4.2): UniRef50, 67,000 steps, batch 1024, length 512, and the same learning rates and latent-branch schedule. Validation NTP loss at step

65,000 is 2.504 (NTP-only) versus 2.507 (NTP+JEPA) continued, and 2.523 versus 2.532 from scratch. NTP+JEPA costs 21% more time per step, so the compute-matched control continues each NTP-only model for 14,070 steps at its final learning rate; validation loss changes by at most 0.001 nats.

Evaluation. We evaluate 16 tasks, excluding PPI, GO, and EC to limit evaluation cost. Unlike the ESM2 probes, regularization strength is selected per task on the validation split (or a 10% training holdout) from $C \in \{10^{-6}, 10^{-4}, \ldots, 10^{6}\}$ before refitting on the training data, so scores are not comparable with Table 2. Embeddings are mean-pooled final-layer states; mean pooling outscored last-token pooling on average for every arm evaluated with both. Table 13 lists per-task scores.

Table 13: ProGen3-112m per-task scores (headline metric; CM: compute-matched).

Task	Off-the-shelf	NTP	NTP (CM)	NTP+JEPA	Scratch NTP	Scratch NTP (CM)	Scratch NTP+JEPA
Remote homology	0.500	0.515	0.514	0.521	0.502	0.502	0.512
Solubility	0.675	0.681	0.680	0.692	0.698	0.700	0.712
Signal peptide	0.960	0.973	0.973	0.980	0.983	0.982	0.988
Neuropeptide precursor	0.967	0.986	0.986	0.974	0.989	0.988	0.987
β-lactamase	0.615	0.634	0.666	0.676	0.608	0.598	0.627
Peptide–HLA	0.771	0.808	0.809	0.835	0.813	0.814	0.844
Metal-ion binding	0.794	0.781	0.749	0.772	0.769	0.768	0.750
Subcellular loc.	0.462	0.452	0.454	0.467	0.480	0.480	0.512
GB1	0.804	0.825	0.825	0.820	0.814	0.814	0.813
Fluorescence	0.578	0.575	0.576	0.576	0.591	0.586	0.577
Stability	0.607	0.608	0.600	0.634	0.604	0.594	0.401
Enzyme efficiency	0.518	0.524	0.525	0.518	0.485	0.487	0.492
SCOPe-40 R@1	0.296	0.304	0.301	0.309	0.257	0.259	0.293
Secondary structure	0.228	0.231	0.231	0.238	0.244	0.243	0.245
Disorder	0.268	0.275	0.274	0.306	0.292	0.291	0.296
Signal-peptide residues	0.590	0.602	0.601	0.609	0.605	0.605	0.617

A.11 Reproducibility

Pretraining uses NVIDIA A100 80GB GPUs—one at 35M and two to four at 150M, with effective batch size held fixed. Controlled-study settings are in Sec. 4.2; Table 14 lists the wall-clock settings. Both studies start from the official ESM2 weights; the wall-clock study loads them through Synthyra's weight-identical FastPLMs packaging (`Synthyra/ESM2-35M`, `Synthyra/ESM2-150M`).

Table 14: Wall-clock optimizer and objective settings. Each within-family contrast shares batch size, masking rate, and schedule. All backbones use $\lambda = 0.45$ and a constant schedule after 1,000 warmup steps.

Backbone	LR (JEPA / MLM-only)	Eff. batch	Mask	α	λ ramp
ESM2-35M (pretrained)	3×10^{-4} / 5×10^{-4}	128	25 %	10^{-3}	1k steps
ESM2-150M (pretrained)	3×10^{-4} / 5×10^{-4}	128	25 %	10^{-3}	1k steps
ESM2-35M (random init)	3×10^{-4} / 3×10^{-4}	128	25 %	10^{-3}	1k steps
AMPLIFY-120M	5×10^{-5} / 5×10^{-5}	208	20 %	1.0	2k steps, 25 % dropout
ProteinBERT2-35M	1×10^{-3} / 1×10^{-3}	128	20 %	1.0	5k delay, 5k ramp, 25 % dropout

The all-position control uses MSE, target layer normalization, SIGReg, gradient-blocked clean-input targets, and no JEPA ramp. The masked-position recipe uses the ramps in Table 14 and replaces all-position MSE with cosine loss at masked tokens. Probes use three seeds, distinct from pretraining seeds.

Controlled-study implementation deviations. In the first seed, predictor heads were initialized before seeding, preventing exact replay; the from-scratch pair also used `torch.compile` only for MLM+JEPA. A cross-seed check found no detectable effect on scores.

Wall-clock pretraining seeds. Both arms of the masked-position contrast use independent pretraining seeds (counts below). These runs provide Table 11's W/L/T, Fig. 3's error bars, and pooled

SCOPe-40 Recall@1. Tests pair shared seeds on four backbones; ProteinBERT2 uses an unpaired test.

Backbone	Masked-pos MLM+JEPA	MLM-only
ESM2-35M (pretrained)	8	8
ESM2-150M (pretrained)	7	6
AMPLIFY-120M	7	7
ESM2-35M (random init)	7	7
ProteinBERT2-35M	5	6

EC uses three runs per arm and the residue-level tasks five to seven. The all-position EC scores for ESM2-35M and ESM2-150M in Fig. 3 come from independently trained checkpoints of the same recipe, rather than the checkpoints used for the other tasks.

The all-position, JEPA-only, recipe-sweep, and 25,000-step runs use one pretraining seed each; their statistics are task-level.