

Frozen structure targets for protein sequence pretraining: a controlled comparison of masked language modelling, latent prediction and distillation

Jacob Davies
University of Warwick
October 2026

Abstract

Latent-prediction objectives in the JEPA family have been reported to degrade protein sequence representations when used without masked language modelling (Ofer et al., 2026). One reason such objectives can fail is that a target encoder learned jointly with the context encoder admits a constant solution. We test whether taking the targets from a frozen, pretrained structure encoder changes this. Four objectives are used to pretrain a 7.41M-parameter encoder with the ESM-2 t6 shape, with identical data (22,839 AlphaFold Database Swiss-Prot chains, 8.90M residues), architecture, optimiser and token budget (about 486M tokens per run): masked language modelling (C1); latent prediction against an EMA target (C2); latent prediction, through a predictor, of frozen ESM-IF1 structure embeddings (C3); and masked distillation of the same ESM-IF1 targets with a linear head (C5b). Eleven runs are evaluated with frozen-feature probes for three- and eight-state secondary structure and zero-shot SCOPe fold retrieval. The EMA condition collapsed as pre-registered: its final encoder RankMe was 37.7 against 187.3 for C3. Frozen targets prevented collapse but did not make latent prediction better than masked language modelling: C3 reached 0.612 ± 0.059 SS3 accuracy against 0.688 ± 0.003 for C1, and its seed-to-seed spread was the largest in the study. Masked distillation without a predictor was the strongest condition on all three tasks (SS3 0.709, SS8 0.565, fold top-1 0.032, against 0.688, 0.542 and 0.022 for C1). All four conditions remain below the public ESM-2 8M model (0.756, 0.611, 0.095), which was trained on far more data. Secondary-structure accuracies are computed on a fixed subset of residues because of a label-conversion error (Section 6), which affects all models equally but makes absolute values not comparable with published CB513 results. At this scale, the structure supervision helped and the predictor did not.

1. Introduction

Most protein sequence encoders are pretrained with masked language modelling (MLM): a fraction of residues is hidden and the model is trained to recover their identities. The objective follows BERT (Devlin et al., 2019), and the ESM-2 family is the standard protein example (Lin et al., 2023). An alternative, borrowed from self-supervised vision, is to predict a *representation* of the hidden positions rather than their tokens. In the joint-embedding predictive architecture (JEPA; LeCun, 2022; Assran et al., 2023) the target representation comes from a second encoder, often an exponential moving average (EMA) of the first, as in BYOL (Grill et al., 2020) and data2vec (Baevski et al., 2022). Ofer et al. (2026) found that JEPA-only training on protein sequences degraded downstream performance in nearly every experiment, and that latent prediction helped only when combined with MLM. With a target encoder learned jointly, a constant embedding minimises the latent loss, and only architectural asymmetries or explicit regularisers rule it out.

This study asks whether that failure is specific to the learned target. If the target is fixed and carries information the sequence encoder does not already have, the constant solution disappears. A natural source of such targets is a structure model. ESM-IF1 (Hsu et al., 2022) is an inverse-folding model whose encoder reads only backbone coordinates; its per-residue embeddings describe local geometry and fold context without encoding the residue's identity. Predicting these embeddings at masked positions asks the sequence encoder to infer structure from sequence, which is the property secondary-structure and fold probes test.

The comparison is built so that only the objective changes. Four conditions share the corpus, the encoder, the corruption scheme, the optimiser, the schedule and the token budget. C1 is MLM. C2 is latent prediction against an EMA target, a replication of the collapse. C3 is latent prediction of frozen ESM-IF1 embeddings through a predictor network. C5b regresses the same targets with only a linear head, so the C3 to C5b difference isolates the predictor. Two hypotheses were fixed before any result was seen: that C2 collapses (H1a), and that C3 matches or beats C1 on at least two of the three structure tasks (H2). Deviations from the pre-registered plan are listed in Section 3.9.

The results are mixed in a specific way. C2 collapsed. C3 did not collapse, but it was worse than C1 on every task and varied widely between seeds. C5b, which uses the same structural information without a predictor, was the best of the four on all three tasks and beat C1 by more than the seed spread. The public ESM-2 8M model, trained on far more data, remained ahead of all of them.

2. Related work

Masked language models for proteins. ESM-2 (Lin et al., 2023) trains transformer encoders with MLM at a range of sizes. Its smallest configuration (6 layers, width 320) defines the encoder shape used here, and the released 8M checkpoint is one of our baselines. ESM-2 random-crops long sequences to its context window, a convention we follow.

Latent prediction and collapse. Joint-embedding objectives such as BYOL (Grill et al., 2020), data2vec (Baeviski et al., 2022) and I-JEPA (Assran et al., 2023) predict the output of a target encoder rather than raw inputs. When the target encoder is learned, the objective has trivial minimisers and relies on architectural asymmetry (a predictor, an EMA target, stop-gradient) or explicit regularisers to avoid them. VICReg (Bardes et al., 2022) adds a variance criterion for this purpose, and SIGReg (Balestrieri and LeCun, 2025) pushes embeddings towards an isotropic Gaussian. RankMe (Garrido et al., 2023) measures the effective rank of a representation from the entropy of its singular values and is a practical collapse diagnostic. Ofer et al. (2026) added latent prediction to MLM for ESM-2 and other protein models, with a 20% mask rate, 80/10/10 corruption and wall-clock-matched budgets. Their JEPA-only ablation degraded downstream performance in nearly every experiment even with SIGReg in place; latent prediction helped only alongside MLM, and mainly at masked positions. Our C2 is not a reproduction of their recipe: it uses an EMA target and no regulariser at all, and serves as a positive control showing that the collapse failure mode exists in our setup.

Structure-aware sequence models. Several models inject structure into protein language models through discrete tokens: Foldseek's 3Di alphabet (van Kempen et al., 2024) is used as an extra input vocabulary by SaProt (Su et al., 2024b) and as a second language by ProstT5 (Heinzinger et al., 2024). These approaches either require structure at inference time or translate between sequence and 3Di tokens. Here structure enters only as a continuous regression target during pretraining, and the resulting encoder reads sequence alone.

Structure encoders as targets. ESM-IF1 (Hsu et al., 2022) is a GVP-Transformer (Jing et al., 2021) trained for inverse folding on about 12M structures that its authors predicted with AlphaFold2 (Jumper et al., 2021) for UniRef50 sequences, together with experimental CATH structures. Its encoder consumes backbone coordinates (N, C α , C) only. Using its embeddings as regression targets is a form of cross-modal knowledge distillation (Hinton et al., 2015) from structure into a sequence model. Because ESM-IF1 has seen essentially every evaluation protein's sequence family during its own training, any such study inherits an asymmetry in supervision, which we discuss in Section 6.

3. Methods

3.1 Pretraining corpus

Sequences and predicted structures come from the Swiss-Prot subset of the AlphaFold Protein Structure Database (Varadi et al., 2022). We fetched 79,446 unique accessions with mean pLDDT (Jumper et al., 2021) ≥ 70 , excluding disordered chains, whose structure targets would be noisy and near-degenerate. MMseqs2 (Steinegger and Söding, 2017) clustering at 50% sequence identity and 80% coverage left 33,855 cluster representatives. To separate pretraining from evaluation data, we removed any representative with $\geq 30\%$ identity over $\geq 50\%$ coverage to any test sequence in the evaluation sets. This removed 11,016 representatives and left a final corpus of **22,839 chains and 8.90M residues**. Chains shorter than 128 residues make up 9.2% of the corpus; the filter removed them disproportionately, consistent with SCOPe domains being short. We did not refill the short band, since that would mean selecting pretraining data by its overlap with the test set.

Chains shorter than 40 residues are excluded. Chains longer than the 512-residue context are randomly cropped to a 512-residue window, redrawn on each pass, rather than discarded, following ESM-2.

3.2 Structure targets

Per-residue targets are the encoder outputs of ESM-IF1 (esm_if1_gvp4_t16_142M_UR50; Hsu et al., 2022), computed from the backbone coordinates of the full predicted structure. ESM-IF1’s encoder reads no amino-acid identities, so a target carries the local geometry of a masked residue but not its label. Targets are computed once from the complete structure, so a cropped window still receives targets that encode the context of the whole fold.

The 512-dimensional embeddings are reduced by PCA to **256 dimensions**, retaining 0.879 of the variance, and each component is standardised. The effective rank (RankMe; Garrido et al., 2023) of the unwhitened target bank is 229.8 at 256 dimensions and 418.9 for the raw 512-dimensional embeddings. A 128-dimensional reduction retained only 0.720 of the variance and was rejected before any training.

3.3 Model

All conditions share an encoder in the ESM-2 t6 configuration with random initialisation: 6 layers, model width 320, 8 attention heads, feed-forward width 1280, rotary position embeddings (Su et al., 2024a), a 33-token vocabulary and a 512-token context, for **7.41M parameters** (7,408,960). We use 8 heads rather than ESM-2’s 20; the parameter count is unchanged, and a pilot measured 1.24 $\times$ higher throughput. Conditions with a predictor (C2, C3) add a narrow transformer of 2 layers, width 160, 4 heads and feed-forward width 640, whose output is projected to the 256-dimensional target space.

3.4 Objectives

Every condition selects 20% of positions, as in Ofer et al. (2026), and applies the same 80/10/10 corruption (Devlin et al., 2019) to them: 80% are replaced by `<mask>`, 10% by a random residue and 10% are left unchanged. Because the MLM and latent conditions use the same corruption, their inputs are identically distributed and the comparison isolates the objective. Table 1 lists the four conditions.

Table 1. Pretraining conditions. All losses are computed on masked positions only.

ID	Objective	Target	Predictor	Seeds
C1	masked amino-acid cross-entropy	residue identity	no	3
C2	latent prediction, EMA target	EMA encoder + linear head, uncorrupted input	yes	3
C3	latent prediction, frozen target	ESM-IF1 bank (PCA-256)	yes	3
C5b	masked distillation	ESM-IF1 bank (PCA-256)	linear head	2

The latent objectives (C2, C3, C5b) use cosine distance plus smooth-L1 ($\beta = 1$) on L2-normalised vectors, weighted equally. C2's EMA momentum rises from 0.996 to 1.0 over training. C2 has no anti-collapse term (no centring, whitening or variance penalty), so it is a positive control for collapse rather than a tuned JEPA. C5b is C3 without the predictor.

3.5 Training

The configurations of all conditions are identical outside the objective block. Each run trains for **7,630 steps** of 65,536 token slots (128 sequences of up to 512 tokens), about **486M real tokens** once padding is excluded. The budget is fixed in steps and tokens rather than wall-clock time; Ofer et al. matched wall-clock time, which here would penalise the predictor conditions for reasons unrelated to the objective. Our 512-token context matches theirs.

Optimisation uses AdamW (Loshchilov and Hutter, 2019) with learning rate 4×10^{-4} , 500 warm-up steps then cosine decay to zero, weight decay 0.01, $\beta = (0.9, 0.98)$, gradient clipping at 1.0, bf16 autocast with fp32 master weights and `torch.compile`. Sequences are batched into length buckets (128, 256, 384, 512). A sequence pads up to its bucket when occupancy is at least 0.85 and is otherwise randomly cropped to the bucket below. This gives 95.7% padding efficiency and crops about 14% of residues per pass, with crop windows redrawn each pass; the policy is identical across conditions. The corpus and target bank are held in GPU memory for the whole run. Runs used one RTX 4090 (24 GB). C1, C2 and C3 were trained with seeds 0, 1 and 2, and C5b with seeds 0 and 1, for eleven runs in total.

3.6 Collapse diagnostics

Every 500 steps, and on the final encoder, we compute four statistics on a fixed held-out probe batch: RankMe, the soft rank obtained from the entropy of the normalised singular values; the mean per-dimension standard deviation; the fraction of dimensions with standard deviation below 0.01; and the off-diagonal covariance mass. They are measured on the *encoder* output, which is what the probes consume, not on the predictor output. The target bank's RankMe (229.8) is the reference line.

3.7 Evaluation

All probes use frozen encoder features. A single hyperparameter grid per task is selected on the validation split and applied identically to every condition; nothing is tuned per condition.

Secondary structure (SS3, SS8). NetSurfP-2.0 data (Klausen et al., 2019) as distributed with TAPE (Rao et al., 2019), with DSSP labels (Kabsch and Sander, 1983): probes are trained on the HHblits set (10,792 proteins), selected on TS115 (115 proteins; Yang et al., 2018) and tested on CB513 (511 proteins; Cuff and Barton, 1999; the commonly quoted 513 reflects a curation difference). Per-residue probes are drawn from a grid of three linear and three one-hidden-layer (256-unit) heads, trained for 20 epochs each. The metric is the per-residue accuracy of the selected head on CB513, computed over the subset of residues described in Section 6.

Fold retrieval. SCOPe (Fox et al., 2014) ASTRAL 2.08 (Chandonia et al., 2022) at 40% identity (15,177 domains, 1,257 folds, 2,065 superfamilies). The task is zero-shot: mean-pooled embeddings are ranked by cosine similarity, with queries (1,031) and gallery (14,146) split so that no superfamily appears on both sides. A retrieval is correct when the top-ranked gallery domain shares the query's fold. The metric is top-1 accuracy; mean average precision (MAP) is also reported.

Baselines. One-hot residue features; a randomly initialised encoder of the same architecture; the public ESM-2 t6 8M model; and, for fold retrieval, the raw ESM-IF1 embeddings of the SCOPe domains computed from their ASTRAL coordinates, which indicates how much fold information the targets themselves carry.

3.8 Hypotheses and analysis

Two criteria were fixed before any result was seen.

- **H1a (replication).** C2 collapses: its final encoder RankMe is below 25% of C3's, across all seeds.

- **H2.** $C_3 \geq C_1$ on at least two of the three structure tasks (SS3, SS8, fold retrieval), with a seed-mean gap greater than one standard deviation (the larger of the two conditions' standard deviations).

We also report C_3 's RankMe against the target bank's 229.8, without a pass/fail criterion, and compare C_3 with C_5b on all three tasks to measure what the predictor contributes. With three seeds, per-task p -values are not reported. We report the seed mean $\pm$ standard deviation per task, and treat the between-seed spread, set against the between-condition range, as a result in its own right.

3.9 Deviations from the pre-registered plan

The plan was cut to its current scope on 22 October 2026, before any training run of the grid, for cost and time reasons. The changes are:

- **Conditions.** Four of seven conditions were trained (11 runs instead of 23). MLM with an added latent loss (C_4), plain distillation without masking (C_5), predictor without masking (C_5c), the loss-weight sweep, a Foldseek 3Di target and span masking were not run. The hypothesis that C_4 is best (H_3) is therefore untested, and the hypothesis about the predictor (H_4), which needed both C_5b and C_5c , is reported only descriptively from C_3 against C_5b .
- **Tasks.** Three of seven planned tasks were evaluated. Contact prediction, fluorescence, stability and ProteinGym were dropped. H_2 was originally defined over SS3, contact prediction and fold retrieval; with contact dropped, it counts over SS3, SS8 and fold retrieval. SS3 and SS8 are correlated, so this weakens the test.
- **Budget.** About 0.5B nominal tokens per run instead of 1.0B, a reduction the plan allowed in advance.
- **Seeds.** C_5b has two seeds rather than three.
- **Ceiling.** The planned ESM-IF1 secondary-structure ceiling, which needed ESMFold structures for the SS proteins, did not finish within its time cap, so only the fold-retrieval ceiling is reported.

4. Results

4.1 Probe accuracy

Table 2 and Figure 1 give the main results. The ordering is the same on all three tasks: C_5b , then C_1 , then C_3 , then C_2 . C_5b exceeded C_1 by 0.021 on SS3, 0.024 on SS8 and 0.011 on fold top-1, in each case by more than the larger of the two conditions' standard deviations (0.003, 0.005 and 0.004). It also had the highest MAP (0.0204 against 0.0102 for C_1). On MAP, C_3 (0.0130) was above C_1 , unlike on top-1.

Table 2. Frozen-feature probe results on the test splits (CB513 for SS3 and SS8; superfamily-disjoint SCOPe queries for fold retrieval). Conditions show mean $\pm$ standard deviation over seeds; baselines are single models. MAP is the seed mean. Bold marks the best pretraining condition in each column.

Model	Seeds	SS3 acc.	SS8 acc.	Fold top-1	Fold MAP
C_1 masked LM	3	0.688 $\pm$ 0.003	0.542 $\pm$ 0.005	0.022 $\pm$ 0.004	0.0102
C_2 latent, EMA target	3	0.492 $\pm$ 0.006	0.358 $\pm$ 0.004	0.012 $\pm$ 0.003	0.0066
C_3 latent, ESM-IF1 target	3	0.612 $\pm$ 0.059	0.472 $\pm$ 0.049	0.020 $\pm$ 0.002	0.0130
C_5b masked distillation	2	0.709 $\pm$ 0.002	0.565 $\pm$ 0.005	0.032 $\pm$ 0.002	0.0204
ESM-2 8M (public)	–	0.756	0.611	0.095	0.0322
ESM-IF1 embeddings	–			0.161	0.0699
Random-init encoder	–	0.514	0.371	0.016	0.0074
One-hot	–	0.487	0.348	0.016	0.0075

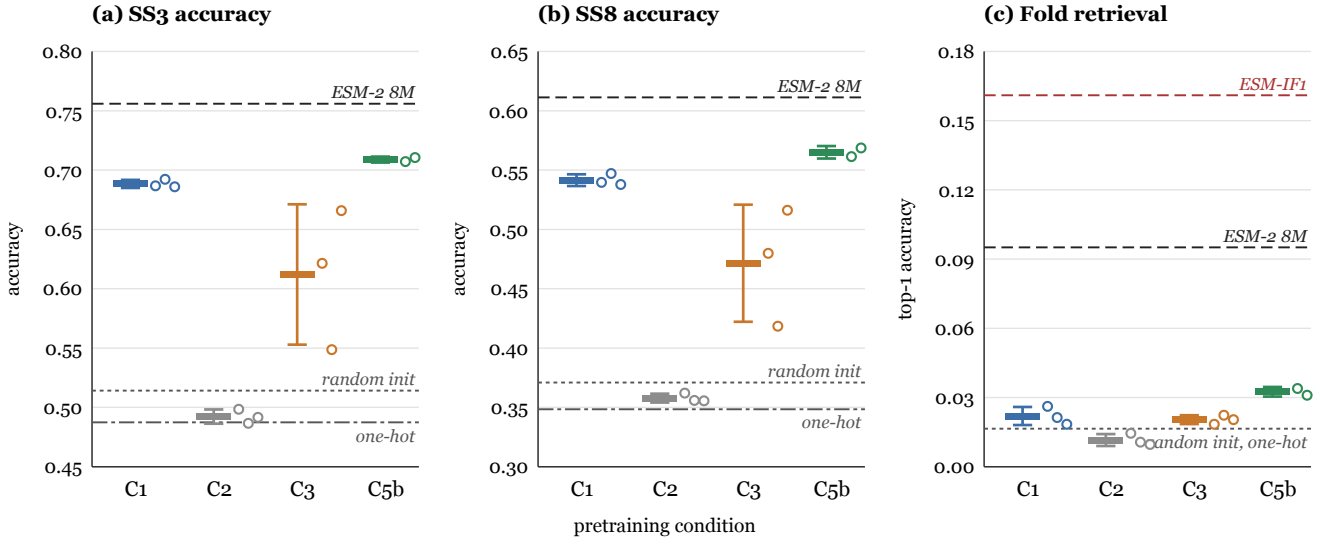


Figure 1. Probe results per condition. The filled bar marks the seed mean, the whisker $\pm$ one standard deviation, and open circles the individual seeds (three for C1–C3, two for C5b). Horizontal lines are single-model baselines. The vertical axes of (a) and (b) do not start at zero. In (c), one query out of 1,031 is worth about 0.001 top-1 accuracy.

C2 scored below the randomly initialised encoder on all three tasks (SS3 0.492 against 0.514; SS8 0.358 against 0.371; fold 0.012 against 0.016) and only slightly above one-hot features on secondary structure. Training against its own EMA target removed information the untrained architecture already provided.

C3 sat between C2 and C1 on secondary structure. On fold retrieval it was level with C1 on top-1 accuracy (0.020 against 0.022) and slightly ahead on MAP. Its seeds spread widely: SS3 accuracies of 0.621, 0.549 and 0.666 for seeds 0, 1 and 2. The best C3 seed was below the worst C1 seed (0.686). The same ordering of seeds appears on the TS115 validation split (0.656, 0.593, 0.699), so the spread is not a feature of the test set.

The public ESM-2 8M model was ahead of every condition, by 0.047 SS3 and 0.046 SS8 over C5b and by a factor of about three on fold retrieval. It has a very similar parameter count but was trained on far more data than our 486M tokens from 8.9M residues, so this gap measures the data budget rather than the objective. On fold retrieval, the raw ESM-IF1 embeddings reached 0.161, five times C5b's 0.032: the sequence encoders recovered only a small part of the fold information present in their targets.

4.2 H2: frozen-target latent prediction against MLM

H2 is not supported. C3 fell below C1 on all three structure tasks (Table 3). None of the gaps clears one standard deviation in C3's favour; on SS3 and SS8 the gaps are negative and larger in magnitude than C3's own standard deviation.

Table 3. The pre-registered H2 comparison. Gap is C3 minus C1 in seed means; sd is the larger of the two conditions' standard deviations. H2 required a positive gap greater than sd on at least two tasks.

Task	C1	C3	Gap	sd	Clears 1 sd
SS3	0.688	0.612	−0.076	0.059	no
SS8	0.542	0.472	−0.070	0.049	no
Fold top-1	0.022	0.020	−0.002	0.004	no

4.3 The predictor: C3 against C5b

C3 and C5b share the targets, the masking and the loss; they differ only in whether a two-layer predictor or a linear head maps encoder outputs to the target space. C5b beat C3 on all three tasks: by 0.097 on SS3 (0.709 against 0.612), 0.094 on SS8 (0.565 against 0.472) and 0.012 on fold top-1 (0.032 against 0.020). A sign test over three tasks gives $p = 0.25$, the smallest value three tasks can produce, so the count alone carries little weight; the size of the secondary-structure gaps, roughly 1.6 to 1.9 times C3's standard deviation, and the fact

that every C5b seed beat every C3 seed on every task, carry more. With a linear head, the encoder output itself must lie close to the target space. With a predictor in between, the encoder can satisfy the loss with representations that need further non-linear processing to reach it, and the frozen-feature probes see only the encoder.

4.4 Collapse

H1a is supported. C2's final encoder RankMe averaged 37.7 across seeds, 0.20 of C3's 187.3, and every C2 seed (37.1 to 38.7) was below 25% of every C3 seed (183.4 to 194.5). Table 4 and Figure 2 give the diagnostics. The collapse was dimensional rather than complete: no condition had any dimension with standard deviation below 0.01, and C2's mean per-dimension standard deviation (0.854) was close to the others. What changed was the spread of variance across directions, visible in C2's off-diagonal covariance mass of 0.966.

Table 4. Collapse diagnostics on the final encoder output, mean over seeds (RankMe also $\pm$ sd). The encoder width is 320; the target bank's RankMe is 229.8 at 256 dimensions.

Condition	RankMe	Mean dim. sd	Dims with sd < 0.01	Off-diag. cov. mass
C1 masked LM	179.3 $\pm$ 7.2	0.881	0	0.854
C2 latent, EMA target	37.7 $\pm$ 0.9	0.854	0	0.966
C3 latent, ESM-IF1 target	187.3 $\pm$ 6.3	0.865	0	0.925
C5b masked distillation	214.4 $\pm$ 2.3	0.946	0	0.817

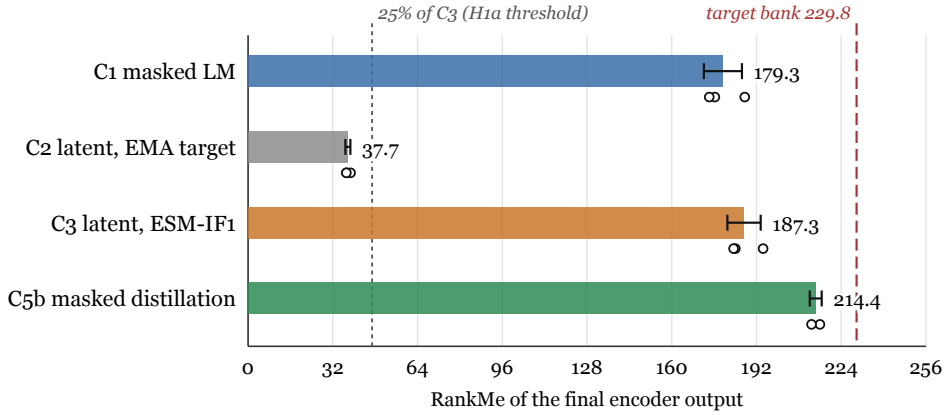


Figure 2. RankMe of the final encoder output. Bars show the seed mean, the black whisker $\pm$ one standard deviation, and open circles the individual seeds. The red dashed line is the RankMe of the PCA-256 target bank (229.8); the grey dotted line is 25% of C3's mean, the H1a threshold.

Frozen targets prevented collapse, as they must by construction: a fixed, non-degenerate target leaves no constant solution. C3's rank (187.3) stayed below the target bank's 229.8, and C5b's (214.4) came closest to it. Rank did not predict probe quality between C1 and C3: C3's RankMe was slightly higher than C1's, yet it was worse on every task. Within C3, the seed with the highest off-diagonal covariance mass (seed 1, 0.960) had the lowest SS3 accuracy, and the seed with the lowest (seed 2, 0.891) the highest. Three seeds are too few to call this a relationship, but it fits the reading that C3's weaker seeds concentrated variance in fewer directions.

4.5 Seed variance

Table 5 sets the mean between-seed standard deviation of each task against the range of seed means across conditions. On every task the between-condition range is eight to thirteen times the seed noise, so the condition ordering is not an artefact of seed choice. The averages hide one exception: C3's standard deviation on SS3 (0.059) and SS8 (0.049) is nearly ten times the largest of C1, C2 and C5b, and is comparable to the gap between C3 and C1.

Table 5. Between-condition range of seed means against mean between-seed standard deviation, per task.

Task	Between-condition range	Mean between-seed sd
SS3	0.217	0.018
SS8	0.207	0.016
Fold top-1	0.021	0.003

5. Discussion

The study set out to separate two things that a frozen-target JEPA combines: supervision from a structure model, and the latent-prediction machinery of a predictor network. The four conditions let us look at each. Taking C1 as the reference, adding structure supervision through a linear head (C5b) improved all three tasks, while adding it through a predictor (C3) made secondary structure worse and left fold top-1 accuracy unchanged. The useful ingredient was the target, and at this scale the predictor got in the way.

Frozen targets did fix the problem Ofer et al. identified. C2 collapsed to an effective rank near 38, below the untrained encoder in probe accuracy, and C3 did not. But removing collapse is a weaker property than learning a good representation. C3's rank was comparable to C1's and its probe accuracy was not, and the C3 seeds differed far more than any other condition's. One explanation consistent with the data is that the predictor gives the encoder freedom it does not need: the loss is satisfied jointly by the encoder and the predictor, and how much of the mapping to structure space each takes on is not fixed by the objective. Different seeds may settle on different divisions of that work, and a probe on the encoder alone sees the result. C5b removes that freedom by making the encoder output a linear function of the target. We have not tested this explanation directly, and other factors, such as the predictor's learning dynamics under a 7,630-step schedule, could contribute.

The C5b result is a distillation result. It shows that, for a fixed sequence-model budget of about 486M tokens, regressing ESM-IF1 embeddings at masked positions produced better secondary-structure and fold features than predicting the masked residues. It does not show that structure targets are a better use of information than MLM, because C5b's targets come from a model that has itself seen far more data (Section 6). The fold-retrieval numbers make the size of what is left on the table clear: the targets alone reach 0.161 top-1, and C5b reaches 0.032.

None of the conditions approaches ESM-2 8M. That model shares the architecture's scale and is trained with MLM, so the gap between it and C1 (0.068 SS3) shows how far our token budget and corpus are from saturating the encoder. Whether the C5b advantage over C1 persists, grows or disappears with more data is not something this study can answer.

6. Limitations

Secondary-structure scoring subset. The evaluation's label conversion read the NetSurfP-2.0 `cb513_mask` column, a space-separated list such as 1.0 1.0 0.0, character by character rather than as numbers. Residue i was kept when character i of that string was neither 0 nor a space. In the training set and TS115, where every mask value is 1.0, this keeps exactly the residues with $i \bmod 4$ in $\{0, 1\}$, about half of each split. On CB513 it keeps 56,147 of 144,011 residues, against 83,991 under the standard mask; 19,678 of the kept residues lie outside the standard mask, and only 43% of the standard residues are scored. Every condition and baseline was trained, selected and scored on the same residues, so comparisons between them are unaffected, but the absolute SS3 and SS8 accuracies are not comparable to published CB513 numbers. A corrected evaluation is left to a revision.

Asymmetric supervision. C3 and C5b receive supervision from a 142M-parameter structure model trained on far more data than C1 sees. ESM-IF1 was trained on AlphaFold structures spanning UniRef50, so essentially every evaluation protein lies within its training distribution. Our pretraining/test filter removes

overlap between the pretraining corpus and the test sets only; it cannot remove ESM-IF1's own exposure. Matching tokens does not correct this asymmetry. The claim the design supports is about how best to spend a fixed sequence-model budget given access to a pretrained structure encoder, not that one objective beats another at equal information.

Seeds and statistics. C1, C2 and C3 have three seeds and C5b has two, so per-task p -values are not meaningful and are not reported; the C5b standard deviations rest on two values each. C3 is unstable across seeds (sd 0.059 on SS3), which makes its mean a less reliable summary than the others'. SS3 and SS8 are computed from the same residues and are correlated, so H2's count over three tasks is less independent than three unrelated tasks would be (Section 3.9), and the sign test over tasks treats them as independent.

Scale. The encoder has 7.41M parameters, each run sees about 486M tokens from an 8.90M-residue corpus, and the context is 512 tokens. Fold-retrieval accuracies are low in absolute terms for every pretrained condition, and differences of 0.002 correspond to about two queries out of 1,031. Conclusions about the predictor may not hold for larger models or longer training, where a predictor has more room to be useful.

Matching. Runs are matched by steps and tokens, not wall-clock time. Conditions with a predictor do more computation per step, so a wall-clock-matched comparison would favour C1 and C5b further.

7. Conclusion

Using frozen ESM-IF1 embeddings as latent-prediction targets removes the collapse that an EMA target produces in protein sequence pretraining, but it does not make latent prediction competitive with masked language modelling at 7.41M parameters and about 486M tokens. Regressing the same targets through a linear head did better than both, on secondary structure and fold retrieval alike. The structural information in the targets was what helped; the predictor added instability and cost accuracy. A public ESM-2 model of similar size, trained on far more sequence data, remains well ahead.

References

- Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y. and Ballas, N. (2023). Self-supervised learning from images with a joint-embedding predictive architecture. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15619–15629.
- Baevski, A., Hsu, W.-N., Xu, Q., Babu, A., Gu, J. and Auli, M. (2022). data2vec: A general framework for self-supervised learning in speech, vision and language. *International Conference on Machine Learning (ICML)*, PMLR 162, 1298–1312.
- Balestrieri, R. and LeCun, Y. (2025). LeJEPA: Provable and scalable self-supervised learning without the heuristics. arXiv:2511.08544.
- Bardes, A., Ponce, J. and LeCun, Y. (2022). VICReg: Variance-invariance-covariance regularization for self-supervised learning. *International Conference on Learning Representations (ICLR)*.
- Chandonia, J.-M., Guan, L., Lin, S., Yu, C., Fox, N. K. and Brenner, S. E. (2022). SCOPe: improvements to the structural classification of proteins – extended database to facilitate variant interpretation and machine learning. *Nucleic Acids Research* 50(D1), D553–D559.
- Cuff, J. A. and Barton, G. J. (1999). Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins* 34(4), 508–519.
- Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171–4186.
- Fox, N. K., Brenner, S. E. and Chandonia, J.-M. (2014). SCOPe: Structural Classification of Proteins – extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Research* 42(D1), D304–D309.
- Garrido, Q., Balestrieri, R., Najman, L. and LeCun, Y. (2023). RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. *International Conference on Machine Learning*

- (ICML), PMLR 202, 10929–10974.
- Grill, J.-B., Strub, F., Alth  , F., Tallec, C., Richemond, P. H., Buchatskaya, E. et al. (2020). Bootstrap your own latent: A new approach to self-supervised learning. *Advances in Neural Information Processing Systems* 33, 21271–21284.
- Heinzinger, M., Weissenow, K., Gomez Sanchez, J., Henkel, A., Mirdita, M., Steinegger, M. and Rost, B. (2024). Bilingual language model for protein sequence and structure. *NAR Genomics and Bioinformatics* 6(4), lqae150.
- Hinton, G., Vinyals, O. and Dean, J. (2015). Distilling the knowledge in a neural network. arXiv:1503.02531.
- Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., Lerer, A. and Rives, A. (2022). Learning inverse folding from millions of predicted structures. *International Conference on Machine Learning (ICML)*, PMLR 162, 8946–8970.
- Jing, B., Eismann, S., Suriana, P., Townshend, R. J. L. and Dror, R. (2021). Learning from protein structure with geometric vector perceptrons. *International Conference on Learning Representations (ICLR)*.
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O. et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589.
- Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22(12), 2577–2637.
- Klausen, M. S., Jespersen, M. C., Nielsen, H., Jensen, K. K., Jurtz, V. I., S  nderby, C. K. et al. (2019). NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins* 87(6), 520–527.
- LeCun, Y. (2022). A path towards autonomous machine intelligence. Version 0.9.2, OpenReview.
- Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W. et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 379(6637), 1123–1130.
- Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. *International Conference on Learning Representations (ICLR)*.
- Ofer, D., Shahaf, D. and Linial, M. (2026). ProteinJEPA: Latent prediction complements protein language models. *3rd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, ICML 2026*. arXiv:2605.07554 (v1; v2 retitled “ProteinJEPA: Latent prediction improves protein language model pretraining”).
- Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., Abbeel, P. and Song, Y. S. (2019). Evaluating protein transfer learning with TAPE. *Advances in Neural Information Processing Systems* 32.
- Steinegger, M. and S  ding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* 35, 1026–1028.
- Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W. and Liu, Y. (2024a). RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing* 568, 127063.
- Su, J., Han, C., Zhou, Y., Shan, J., Zhou, X. and Yuan, F. (2024b). SaProt: Protein language modeling with structure-aware vocabulary. *International Conference on Learning Representations (ICLR)*.
- van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L. M., S  ding, J. and Steinegger, M. (2024). Fast and accurate protein structure search with Foldseek. *Nature Biotechnology* 42, 243–246.
- Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G. et al. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research* 50(D1), D439–D444.
- Yang, Y., Gao, J., Wang, J., Heffernan, R., Hanson, J., Paliwal, K. and Zhou, Y. (2018). Sixty-five years of the long march in protein secondary structure prediction: the final stretch? *Briefings in Bioinformatics* 19(3), 482–494.

Appendix: per-run results

Table A1. Every run. SS columns are CB513 test accuracy with TS115 validation accuracy in parentheses; fold is SCOPe top-1; RankMe is on the final encoder output.

Run	SS3 test (val)	SS8 test (val)	Fold top-1	Fold MAP	RankMe
C1 seed 0	0.687 (0.717)	0.540 (0.596)	0.026	0.0103	176.3
C1 seed 1	0.692 (0.727)	0.547 (0.604)	0.021	0.0097	187.5
C1 seed 2	0.686 (0.717)	0.538 (0.597)	0.018	0.0105	174.2
C2 seed 0	0.499 (0.537)	0.362 (0.428)	0.015	0.0068	37.2
C2 seed 1	0.487 (0.536)	0.356 (0.432)	0.011	0.0065	38.7
C2 seed 2	0.492 (0.538)	0.356 (0.429)	0.010	0.0066	37.1
C3 seed 0	0.621 (0.656)	0.480 (0.545)	0.018	0.0122	184.0
C3 seed 1	0.549 (0.593)	0.418 (0.494)	0.022	0.0122	183.4
C3 seed 2	0.666 (0.699)	0.516 (0.575)	0.020	0.0147	194.5
C5b seed 0	0.707 (0.740)	0.561 (0.625)	0.034	0.0202	216.0
C5b seed 1	0.711 (0.744)	0.569 (0.624)	0.031	0.0206	212.8