Controlling *Repetition* in Protein Language Models

Jiahao Zhang[1] · Zeqing Zhang[1] · Di Wang[2] · Lijie Hu[1]
[1]Mohamed bin Zayed University of AI [2]KAUST *Equal contribution (jiahao.zhang@mbzuai.ac.ae)

ICLR 2026

01 Motivation — Repetition Collapses the Fold

PLMs like **ESM-3** and **ProtGPT2** generate sequences that collapse into **homopolymers** or **motif loops**. Unlike text repetition, these **destroy the predicted** fold. AlphaFold pLDDT drops from 85+ to under 50 across repetitive regions.

*Blue = pLDDT > 90 (confident); **orange/red** = < 70 (low). Repetitive residues fold poorly.*

02 Measuring Repetition · R(x) and U(x)

We unify three complementary signals into a single repetition score in [0, 1] (higher = less repetition):

R_{ngram} — Token entropy, catches **global** collapse.

$Distinct\text{-}n$ — $n \in \{1, 2\}$ catches **motif loops**.

R_{TopCF} — Penalizes runs of φ 4 identical residues.

$$R(x) = \text{harm}_{w=1/3}(\text{H}_2(S_2)/S_2,\ R_{ngram}) \cdot \text{EXP}(-\text{MAX}(0, 100 - \varphi)/60)$$

Across ESM-3 and ProtGPT2 on **CATH**, **UniRef50**, **SCOP**, low R reliably predicts low U — repetition is not cosmetic.

```
↑R ⇒ less repetition  ↑U ⇒ better fold
```

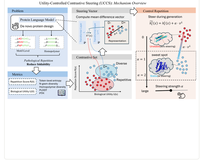
03 Why not just penalize repetition?

Every decoding penalty or mechanism edit we tested either **barely moves R** or **trades R for a large U drop. No baseline simultaneously improves both.**

METHOD FAMILY	R	U	JOINTLY ↑?
Temperature / top-p	○	○	✗
Entropy-based sampling	○ ↑	○ ↓	✗
Repetition / n-gram penalty	↑	○	✗
Neuron deactivation	↑	○	✗
Probe steering	○	○	✗
UCCS (ours)	↑	↑	✓

04 UCCS · Utility-Controlled Contrastive Steering

UCCS extracts a **single steering vector** from a utility-matched contrastive dataset and injects it at inference — **no retraining, no decoding penalty**.



STEP 1
Utility-matched pairs
Build D^+ (low R) and D^- (high R) so that $D(x)$ is a write $D(\cdot)$ is maximized. Pairs (x^+, x^-) share utility but differ in expansion.

STEP 2
Mean-difference vector
Pool per-layer representations — mean for MLMs, last-token for AR LMs. One forward pass per sequence.

$$v^l = \mathbb{E}_{D^+}[h_l^{\text{pool}}(x)] - \mathbb{E}_{D^-}[h_l^{\text{pool}}(x)]$$

STEP 3
Plug-and-play injection
Add αv^l to hidden states at inference. α = 7 works across backbones.

$$h_l^{\text{new}}(x) = h_l^{\text{base}}(x) + \alpha \cdot v^l$$

$(x^+, x^-) \in \text{original} \cdot \text{AA} \cdot \text{MD} \cdot \text{AE} \cdot \text{AV}$

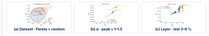
Takeaway. Expansion and utility are linearly separable in PLM activations, and a single vector can steer between them. Mean-difference (MD) and activation-addition (AE) both work, but AE is more brittle — it overshoots at the same α that MD handles cleanly. Random-direction steering (RN) collapses utility, confirming the steering is not just noise. Steering the un-controlled variant fails to improve R. On D"(0)" pools, **R metrics separate sharply** while **U distributions align**, confirming the disentanglement.

05 Main Results · ESM-3, unconditional, CATH

+52%	+13%	✓
R ↑ vs ORIGINAL	U ↑ vs ORIGINAL	R ↑ & R_{ngram} Pareto

METHOD	R ↑	U ↑	R ↑ & R_{ngram} ↑
Original	0.423	0.570	—
Temperature	0.751	0.566	✗
Entropy sampling	**0.804**	0.477	✗
Neuron deactivation	0.531	0.507	✗
Probe steering	0.476	0.587	✓
UCCS (ours)	**0.643**	**0.652**	✓

06 Ablations · Dataset · α · Layer



(a) Dataset · Pareto > random **(b) α · peak ≈ 1~1.5** **(c) Layer · last 3-8 %**

07 UCCS Generations · Foldable & Diverse Across Backbones

ESM-3 (MLM) · **ProtGPT2 (AR)** · All examples at **pLDDT** > centered. UCCS preserves secondary structure and tertiary packing across diverse lengths and topologies — not collapsing into one complexity bucket.

ICLR 2026

Takeaway. Repetition is PLMs' dominant failure mode — a utility-disentangled steering vector controls it without retraining, improving foldability rather than trading it.

Limits. steering vectors are model-specific; the cross-model transfer yet; contrastive pairs currently rely on AlphaFold pLDDT — perplexity-based filtering is a cheaper alternative.

Paper · Code · Checkpoints · Data · Discussion