# OFF-MANIFOLD COLLAPSE IN GUIDED PROTEIN LANGUAGE MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Protein language models are widely used priors for protein sequence design, and a growing body of work controls them at inference time as an alternative to fine-tuning. Such guidance faces a dilemma: mild enough to preserve natural activation statistics, it barely moves the property; strong enough to move it, the generations become progressively harder to fold. We show the failure has a specific and cheaply detectable signature, an off-manifold collapse of the model's own representations. Guided activations fall toward a region statistically indistinguishable from random amino-acid input, and the sequences degenerate to low complexity, yet the property oracle being optimized can still score these generations as a success. The optimized oracle can therefore fail to witness the collapse and, for solubility, can actively reward it, whereas structure and composition expose the failure. Because the failure is already visible in a finished candidate, we detect it at the output rather than modify the generator. We introduce a cheap density prior over natural protein activations and keep only the candidates that remain typical under it, a training-free post-hoc step we call **Mahalanobis filtering**. At matched guidance settings it improves both the property score and the structural plausibility of the sequences it keeps at negligible cost, without touching the generator, and transfers across different guidance methods.

## 1 INTRODUCTION

Pretrained protein language models (PLMs) such as ESM-2 (Lin et al., 2023), ESM3 (Hayes et al., 2025), ProGen (Madani et al., 2023), and ProtGPT2 (Ferruz et al., 2022) are widely used priors for protein sequence design, and a growing body of work controls them at inference time rather than by fine-tuning: some follow the gradient of a trained property model through the generator (Dhariwal & Nichol, 2021; Gruver et al., 2023; Wang et al., 2024; Yang et al., 2025), others reshape the token distribution at decoding (Fernández Burda et al., 2026; Calvanese et al., 2026), while the ProteinGuide framework supports property guidance across several protein generator classes (Xiong et al., 2026). In addition, *activation steering* edits the hidden states themselves. Steering is among the most attractive of these: transplanted from the control of large language models (Turner et al., 2023), it adds a direction computed from high- and low-property reference sequences to the residual stream during masked decoding and obtains sequences the oracle scores above the reference set (Huang et al., 2025); Section 2 gives the mechanism. Steering requires neither gradients through the property model nor weight updates. We study it as the primary intervention and use predictor-gradient guidance as a mechanistically distinct transfer test.

Aggressive inference-time control nevertheless carries a structural cost. Other controllers report the same tension from the opposite side (Fernández Burda et al., 2026; Calvanese et al., 2026). Logit-difference amplification addresses the degraded sequence properties and distributional similarity caused by activation steering (Fernández Burda et al., 2026), while iterative lookback Monte-Carlo sampling targets steering-induced diversity loss (Calvanese et al., 2026). These studies establish that the degradation is real and recognized, but each treats it as a reason to replace steering; neither directly characterizes its representation-space signature. Figure 1 shows what it looks like in a single pair of generations: pushed hard enough to move the property, steering makes ESM-2 emit a glycine homopolymer that the solubility oracle rates as almost certainly soluble ($0.99$) while ESMFold confidence falls to $0.35$. The optimized oracle *prefers* the degenerate sequence, so the failure can only be read from referees outside the steering objective. We therefore ask:



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.

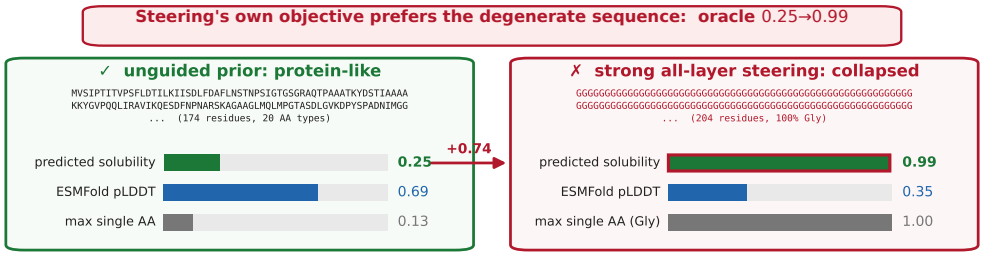


Figure 1: **The failure mode in one pair of generations** (solubility, ESM-2 650M). The unguided sample uses all 20 amino acids and has high predicted structural confidence. Under strong all-layer steering, ESM-2 instead produces a glycine homopolymer that the steering objective rates as almost certainly soluble. The optimized oracle therefore prefers the degenerate sequence, whereas structure and composition reveal the failure. Section 3 characterizes the corresponding representation-space collapse at the population level.

> *What happens to a protein language model's representations when steering is pushed hard enough to move the property, and can the resulting failure be mitigated after generation?*

We make two contributions:

- **A characterization of the failure as *off-manifold collapse*.** Across both solubility and thermostability steering, the representations become less typical and ESMFold pLDDT decreases. However, the property-oracle score does not necessarily decrease at the same time. Therefore, the property score alone is not a reliable indicator of successful steering (Section 3).
- **We introduce *Mahalanobis-filtered steering*, a training-free post-hoc selection layer.** At matched generation settings it recovers substantially higher-confidence candidates without modifying the generator or training an additional model. The same fixed detector transfers across properties and guidance mechanisms, while a learned activation prior confirms that the inexpensive radial statistic captures the dominant representation geometry (Section 4).

## 2 Background and Related Work

**Protein language models.** A protein is a string over an alphabet of 20 amino acids whose folded structure determines its function and properties such as solubility or thermostability. Trained on databases of natural proteins, a protein language model is a generic prior over "protein-like" sequences whose activations encode structure and function (Rives et al., 2021; Meier et al., 2021): ESM-2 (Lin et al., 2023) is a masked language model trained on UniRef (Suzek et al., 2015), and we work with its 650M variant. The family spans masked encoders (Lin et al., 2023; Elnaggar et al., 2022), autoregressive generators (Madani et al., 2023; Nijkamp et al., 2023; Ferruz et al., 2022), discrete-diffusion generators (Wang et al., 2024), instruction-tuned variants (Lv et al., 2026), and multimodal generative models over sequence, structure, and function (Hayes et al., 2025). Generation is by *iterative masked decoding*: from a reference sequence, positions are repeatedly masked and resampled from the model until every position has been re-predicted (Huang et al., 2025).

**Activation steering.** Activation steering edits a model's hidden states at inference time: one fixed direction is added to the residual stream, with no weight update, a technique developed for controlling large language models (Turner et al., 2023; Li et al., 2023; Rimsky et al., 2024; Zou et al., 2023). The direction is typically a *difference of means* between the activations of two contrasting example sets, and Huang et al. (2025) apply it to PLMs. Write $h_\ell(s)$ for the layer-$\ell$ activation of a reference sequence $s$, and let $\mathcal{H}$ and $\mathcal{L}$ be reference sets labeled high and low for the target property. The

steering vector, and the update it makes to the residual stream $x$ at that layer in every forward pass of masked decoding, are

$$v_\ell = \frac{1}{|\mathcal{H}|}\sum_{s\in\mathcal{H}} h_\ell(s) \;-\; \frac{1}{|\mathcal{L}|}\sum_{s\in\mathcal{L}} h_\ell(s), \qquad x \;\leftarrow\; \big(x + \alpha v_\ell\big)\,\frac{\|x\|}{\|x + \alpha v_\ell\|}, \tag{1}$$

where $\alpha$ sets the strength and the rescaling preserves the norm of $x$. We use all layer injection (`allL`) throughout the main experiments and perform a single layer sweep injection for an auxiliary analysis. Preserving the raw norm in Eq. 1, however, does not preserve the per-channel whitened profile, so the steered activations can still leave the model's typical activation region, as measured in Section 3.

**Predictor-gradient guidance.** As a mechanistically distinct controller, we also guide masked decoding using the gradient of the property oracle. At each round, the activation is moved by a fraction $\eta$ of its norm along the normalized gradient, producing an input-dependent direction. Because the same oracle both defines this update and reports the property score, its increase is expected.

**How a generated protein is judged.** We evaluate generated proteins along two complementary axes. First, a property oracle estimates the intended functional shift: a solubility classifier (Khurana et al., 2018) or a $T_m$ regressor (Jarzab et al., 2020), each trained on frozen ESM-2 features. These predictors define the steering objective, but fitness models become less reliable under distant extrapolation and vary across protein families (Freschlin et al., 2024; Notin et al., 2023). Second, we use ESMFold pLDDT (Lin et al., 2023) as a proxy for structural plausibility, reporting it over populations rather than treating any individual prediction as definitive. AlphaFold confidence provides an alternative (Jumper et al., 2021); ESMFold's MSA-free inference makes population-scale evaluation tractable. Neither axis replaces experimental validation: computational metrics do not by themselves establish activity (Johnson et al., 2025); experimental studies of generated proteins and closed-loop methods rely on wet-lab feedback (Verkuil et al., 2022; Jiang et al., 2025); our method remains a computational pre-screen.

## 3 Off-Manifold Collapse under Inference-Time Guidance

We characterize the failure across complementary controlled steering sweeps. Figure 2 compares all-layer steering on the property-conditioned `sol_hard` split and on property-unconditioned `global` UniRef50 references, while Figure 3 isolates the representation geometry using single-layer steering on `sol_easy`. Across these diagnostic sweeps, we evaluate more than 4,000 controlled generations; full population definitions and generation counts are given in Appendix A.

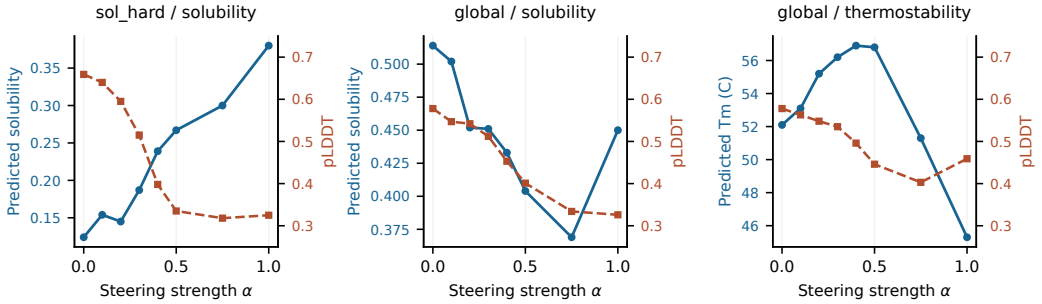


Figure 2: **Structural degradation is consistent despite qualitatively different property responses.** Across three all-layer steering settings ($n = 128$ per point), property scores increase, vary non-monotonically, or rise before declining, while ESMFold pLDDT consistently decreases with stronger steering. Thus property-score direction alone is not diagnostic of collapse.

### 3.1 Property Gains Can Come with Structural Collapse

Figure 2 shows a consistent decline in structural confidence despite qualitatively different property responses. On `sol_hard`, predicted solubility rises from 0.12 to 0.38 as $\alpha$ increases from 0 to 1,

while ESMFold pLDDT falls from $0.66$ to $0.32$. The structural decline begins early: at $\alpha = 0.2$, pLDDT has already dropped to $0.60$ while predicted solubility remains near its unguided value.

Across the other settings, property scores are non-monotonic or improve only at moderate steering strengths, whereas pLDDT again decreases under stronger guidance. Thus the direction of the property score is not informative about collapse: improvement does not rule it out, and deterioration does not establish it. This failure is therefore not simply an oracle accuracy problem; even a stronger property oracle would measure the target property rather than whether generation remains within the model's natural representation regime. This motivates examining the geometry of the generator's internal representations directly.

### 3.2 The Collapse Is Radial in Representation Space

To test whether the loss of structural plausibility has an internal counterpart, we inspect ESM-2's own hidden representations after generation. Let $h(s) \in \mathbb{R}^{L \times D}$ denote the interior-token layer-17 activations of a completed sequence $s$, where $D = 1280$ for ESM-2 650M. Let $\mu_d$ and $\sigma_d^2$ denote the mean and variance of channel $d$ under a large natural-protein reference population. Crucially, these statistics are estimated independently of the guided evaluation pools, using a streaming forward pass over approximately 58 million UniRef50 sequences (Appendix A). We standardize each activation coordinate as

$$z_{i,d}(s) = \frac{h_{i,d}(s) - \mu_d}{\sigma_d}, \tag{2}$$

and define the token-averaged squared Mahalanobis radius (Lee et al., 2018)

$$\mathcal{M}(s) = \frac{1}{L} \sum_{i=1}^{L} \sum_{d=1}^{D} z_{i,d}(s)^2. \tag{3}$$

Across the all-layer sweeps, the dominant representation-space change is an *inward* contraction under this standardization. On the property-unconditioned `global` population, increasing steering from $\alpha = 0$ to $\alpha = 0.5$ reduces $\mathcal{M}$ from $1194$ to $790$ for solubility and from $1194$ to $868$ for thermostability; over the same interval, pLDDT falls from $0.578$ to $0.401$ and $0.446$, respectively. On `sol_hard`, the contraction is stronger still: $\mathcal{M}$ falls from $1226$ to $642$ by $\alpha = 0.75$, alongside a pLDDT decrease from $0.659$ to $0.318$.

To isolate the geometry of this shift from the effects of intervening at many layers simultaneously, we perform the controlled layer-17 sweep in Figure 3. Increasing the steering scale progressively contracts the entire channel-wise standardized activation profile toward the natural-protein mean. At the sequence level, $\mathcal{M}$ falls from approximately $1232$ at $\alpha = 0$ to a plateau near $600$, close to the random-amino-acid reference of $604$.

### 3.3 Off-Manifold Collapse as a Typicality Failure

Why is contraction toward the natural-protein mean a failure rather than a return to a more natural representation? Under the working approximation that standardized natural activations have independent unit-variance coordinates, the token-level radius satisfies

$$\sum_{d=1}^{D} z_{i,d}^2 \sim \chi_D^2, \tag{4}$$

and therefore concentrates around $D$, not around zero. More generally, standardization alone gives $\mathbb{E}[\sum_d z_{i,d}^2] = D$ under the natural reference distribution. Natural-protein activations thus occupy a high-dimensional *typical shell* away from the channel-wise mean.

This is the familiar distinction between density and typicality in high dimensions (Nalisnick et al., 2019a;b). For a diagonal Gaussian, moving toward the mean increases pointwise density, yet sufficiently small radii leave the region in which natural samples concentrate. We use *natural activation manifold* to denote this empirically occupied region, and *off-manifold collapse* to denote the guidance-induced departure from it toward abnormally small radius.

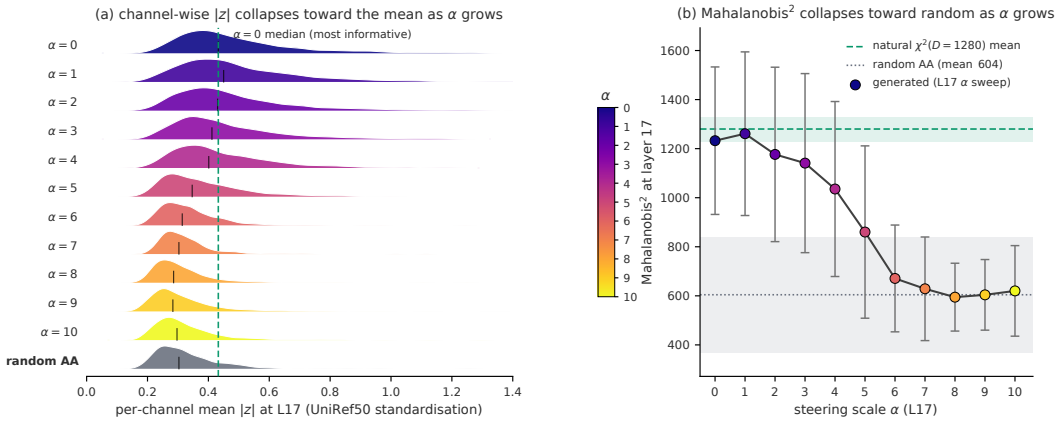


Figure 3: **Strong steering collapses ESM-2 representations toward the random-amino-acid radial regime** (single-layer steering at layer 17, `sol_easy`, 100 sequences per $\alpha$). **(a)** Distribution across the 1280 channels of the mean absolute standardized activation, $z = (h - \mu)/\sigma$. Increasing $\alpha$ contracts the complete channel profile toward the natural-protein mean, reaching the random-AA regime at approximately $\alpha = 6$. **(b)** Sequence-level squared Mahalanobis radius. The mean falls from approximately 1232 to approximately 600, compared with 604 for random-AA input. The green reference indicates the per-token scale $D \pm \sqrt{2D}$ under the diagonal-Gaussian approximation; the substantially larger empirical sequence-to-sequence spread reflects correlations across residues.

The diagonal statistic $\mathcal{M}$ retains only this radial component of the representation geometry. It ignores cross-channel covariance, making it far cheaper than a full learned activation model; Section 5 compares it with a flow-based Generative Latent Prior trained on the same natural activation distribution.

### 3.4 The Same Collapse Appears under Predictor-Gradient Guidance

The inward collapse is not unique to a fixed difference-of-means steering vector. We also guide the same ESM-2 masked-decoding process using the gradient of the property predictor at layer 17. Across both solubility and thermostability, increasing the guidance step $\eta$ again moves the completed sequences inward: $\mathcal{M}$ decreases together with ESMFold pLDDT at every tested guidance strength.

This agreement across two mechanistically different interventions motivates treating inward off-manifold collapse as a failure mode of aggressive inference-time guidance rather than an artifact of one particular steering direction. Section 5 evaluates whether the same post-hoc detector can be reused unchanged across the two controllers.

## 4 Mahalanobis-Filtered Steering

The diagnosis above suggests a direct remedy. Because the off-manifold signature remains visible in the representation of a completed sequence, we leave the generator unchanged and screen candidates only after generation.

We precompute a fixed natural-activation reference from approximately 58 million UniRef50 sequences. The resulting reference is only two 1280-dimensional moment vectors (about 10 KB) and is reused unchanged throughout. It uses the readout defined in Section 3.2; full construction details and the choice of readout layer are given in Appendix A.

### 4.1 A One-Sided Typicality Filter

Given a completed candidate $s$, we retain it if

$$\mathcal{M}(s) \geq D - k\sqrt{2D}, \tag{5}$$

with $k = 1$. For ESM-2 650M, $D = 1280$, yielding the fixed threshold $\tau = 1229$, which is used for all experiments.

The scale in Eq. (5) is motivated by the $\chi^2_D$ radius of a standardized diagonal Gaussian, but is not interpreted as an exact sequence-level quantile: residues within a protein are correlated, so the token-averaged radius does not follow an independent $\chi^2_D$ law. We therefore treat $k$ as a fixed typicality margin and report empirical acceptance rates; threshold sensitivity is given in Appendix E.

The test is deliberately one-sided. Section 3 identifies inward contraction as the dominant failure regime, so Eq. (5) acts as a lower-tail veto.

### 4.2 Post-Hoc Selection

For each completed candidate, we run one unmodified ESM-2 forward pass, standardize its layer-17 activations using the stored natural moments, and evaluate Eq. (3). No decoding step is rerun, no property predictor is retrained, and no additional learned model is required.

Because scoring depends only on the completed sequence, the same reference statistics and threshold can be applied unchanged to activation steering, predictor-gradient guidance, or unguided generation. We refer to steering followed by this post-hoc selection step as *Mahalanobis-filtered steering*.

The online cost is one ESM-2 forward pass plus channel-wise standardization and summation; Section 5 compares this cost with pseudo-perplexity and a learned activation prior.

## 5 Evaluation

We evaluate two primary claims. First, within a fixed guided candidate pool, does Mahalanobis filtering recover a higher-confidence viable tail without changing the generator? Second, does the same fixed detector transfer across properties and guidance mechanisms? We then ask whether the inexpensive radial statistic captures the broad representation geometry of a substantially more expressive learned activation prior, and quantify the resulting computational trade-off.

### 5.1 Filtering Recovers the Viable Tail within a Fixed Guided Pool

Our central test holds the generator, steering strength, reference proteins, and decoding seeds fixed. Mahalanobis filtering only selects a subset of the already generated candidates; it neither regenerates nor modifies them.

Table 1 shows that this within-pool selection strongly enriches for structural confidence. At $\alpha = 0.5$, filtering retains 16–20% of the two guided populations and raises mean pLDDT by 0.19–0.25, with both paired-bootstrap intervals excluding zero. The retained subsets also preserve, and in these settings improve, the target-property shift.

Table 1: **Mahalanobis filtering recovers a higher-confidence subset at fixed steering strength.** All-layer steering on property-unconditioned UniRef50 references ($\alpha = 0.5$, $k = 1$, $n = 128$ per property). The +filter columns contain a subset of the same generated population; no candidate is regenerated or modified. Property is predicted solubility probability for *sol* and predicted $T_m$ (°C) for *therm*. Intervals are paired-bootstrap 95% confidence intervals and Acc. is the retained fraction.

| | **property** | | **pLDDT** | | | |
|---|---|---|---|---|---|---|
| **task** | steer | +filter | steer | +filter | **acc.** | **Δ pLDDT** |
| sol | 0.404 | **0.491** | 0.401 | **0.647** | 16% | +0.247 [0.200, 0.299] |
| therm | 56.8 | **63.6** | 0.446 | **0.633** | 20% | +0.187 [0.140, 0.234] |

Acceptance rate is part of the result: selection trades yield for quality. The same rule retains about half of the unguided global population, providing a reference for the selector's behavior on the model prior. We therefore interpret filtering as recovery of the viable tail of each guided parent pool, not as a claim that filtered steering universally dominates the filtered prior.

### 5.2 The Same Fixed Detector Transfers to a Different Controller

A stronger generalization test is whether the detector survives a change in the guidance mechanism without recalibration. We replace fixed-vector activation steering with predictor-gradient guidance while keeping the backbone, readout layer, natural moments, and threshold unchanged.

Across both properties and all tested guidance strengths, stronger predictor-gradient guidance again moves the parent population inward, while the same fixed filter enriches for higher-pLDDT candidates (Table 2). The gain is $+0.145$ to $+0.175$ across the six global pools.

Table 2: **The same fixed filter transfers to predictor-gradient guidance.** Results on the property-unconditioned `global` population ($n = 128$ per row). The generator backbone, layer-17 natural moments, and Mahalanobis threshold are identical to the activation-steering experiments. Because the property oracle also defines the guidance gradient, property values describe the generated populations but are not treated as an independent validation signal.

| | | **property** | | **pLDDT** | | |
|---|---|---|---|---|---|---|
| **task** | $\eta$ | parent | retained | parent | retained | **acc.** |
| sol | 0.05 | 0.746 | 0.856 | 0.532 | **0.686** | 44% |
| | 0.10 | 0.775 | 0.882 | 0.501 | **0.676** | 33% |
| | 0.20 | 0.789 | 0.850 | 0.467 | **0.633** | 27% |
| therm | 0.05 | 59.5 | 60.6 | 0.532 | **0.677** | 42% |
| | 0.10 | 60.2 | 63.7 | 0.481 | **0.653** | 33% |
| | 0.20 | 60.6 | 63.1 | 0.438 | **0.595** | 28% |

An output-level control on ProtGPT2 generations gives consistent candidate-level ranking behavior in Appendix D.

### 5.3 The Radial Statistic Captures the Broad Geometry of a Learned Prior

The Mahalanobis detector deliberately retains only radial information under a Gaussian reference. To test how much geometry is lost by this simplification, we compare it with a Generative Latent Prior (GLP) (Luo et al., 2026), a substantially more expressive generative model. We train a GLP on the same natural layer 17 activations used to construct the Mahalanobis reference. GLP learns the activation distribution by flow matching. Given a natural activation $z_0$ and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, it constructs the interpolation

$$z_t = (1 - t)z_0 + t\epsilon, \qquad t \in [0, 1],$$

and trains a velocity field $\hat{u}_\theta$ with

$$\mathcal{L}_{\mathrm{GLP}} = \mathbb{E}_{z_0,\epsilon,t}\left[\|\hat{u}_\theta(z_t, t) - (\epsilon - z_0)\|_2^2\right].$$

The learned vector field defines a continuous flow from the activation distribution to a standard Gaussian, so it also yields a density through the continuous change-of-variables formula,

$$\log p_{\mathrm{GLP}}(z_0) = \log \phi(z_1) + \int_0^1 \nabla_z \cdot \hat{u}_\theta(z_t, t)\, dt, \qquad \frac{dz_t}{dt} = \hat{u}_\theta(z_t, t),$$

where $\phi$ is the standard Gaussian density. We use $\mathrm{NLL}_{\mathrm{GLP}}(z) = -\log p_{\mathrm{GLP}}(z)$ as a learned full-density reference score. Unlike the Mahalanobis radius, which summarizes the reference distribution through its mean and covariance under a Gaussian model, GLP learns the activation distribution directly with a nonlinear generative model, allowing it to represent distributional structure beyond this Gaussian approximation.

Across 4,400 sequences spanning several steering regimes and a random AA anchor, Mahalanobis radius and GLP NLL nevertheless show strong across-regime agreement (Spearman $r = 0.815$; Figure 4).

The pooled correlation should not be interpreted as evidence that the two scores rank candidates identically within a fixed pool: within-setting correlations range from 0.334 to 0.900. Figure 4 therefore validates the broad representation geometry captured by $\mathcal{M}$, while the fixed-pool selection experiments in Section 5.1 provide the operational evidence for filtering. Table 3 quantifies the corresponding cost: the 34 ms Mahalanobis readout uses only 10 KB of stored moments, whereas GLP NLL takes approximately 3 s and requires a 1.3 GB model. The same table places these figures next to sequence-level naturalness baselines: ESM-2 650M and 3B pseudo-perplexity need 472 ms and 1,513 ms per sequence. Mahalanobis requires no additional model beyond the ESM-2 already loaded for generation. Thus, $\mathcal{M}$ is not a surrogate for GLP ranking within a fixed pool; rather, it captures the coarse geometry at roughly two orders of magnitude lower online cost.

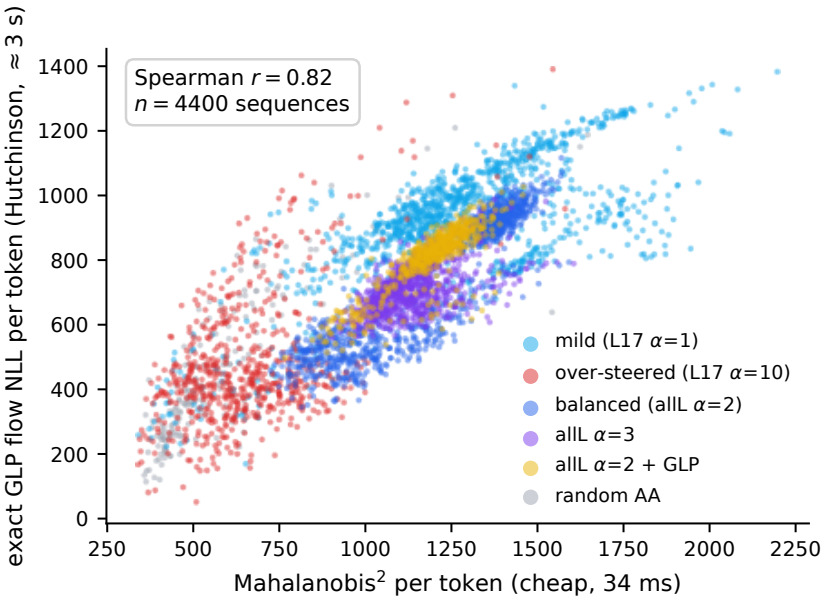


Figure 4: Per-sequence $\mathcal{M}$ vs. GLP flow NLL over 4,400 sequences (two properties, five settings, plus a random-AA anchor). The pooled association ($r$=0.815) summarizes separation across regimes, not agreement of within-pool selections; exact NLL is ≈90× slower (3 s vs. 34 ms per sequence).

Table 3: **Cost of candidate-level naturalness scores** (A100, sequence length approximately 200). Storage denotes additional score-specific parameters beyond the ESM-2 model already used by the generation pipeline.

| score | time / sequence | additional storage |
|---|---|---|
| Mahalanobis (ours) | 34 ms | 10 KB |
| GLP flow residual | 283 ms | 1.3 GB |
| GLP flow NLL | ≈ 3 s | 1.3 GB |
| ESM-2 650M pseudo-PPL | 472 ms | – |
| ESM-2 3B pseudo-PPL | 1,513 ms | separate backbone |

## 6 Discussion and Limitations

**Selection is not optimization.** Our results validate the Mahalanobis score as a post-hoc selection criterion, not as an objective to optimize directly. In an exploratory best-of-$K$ decoding experiment, increasing $K$ from 1 to 8 increases the mean radius from approximately 1015 to 1248 without improving mean pLDDT (0.484 versus 0.478). This suggests that directly optimizing through the score can alter the candidate distribution without improving structural quality. Appendix H provides the full setup and post-hoc selection results.

**The method requires residual viable candidates.** Mahalanobis filtering can recover a viable tail from a degraded candidate population, but it cannot create such a tail once the entire pool has collapsed. Its benefit therefore depends on generation retaining some variation in structural quality.

**Scope of the current detector.** The one sided rule is designed for the recurrent inward collapse regime identified in Section 3. Other failure modes, including outward or primarily composition driven degeneracies, may require complementary signals. Our guided generation experiments also use a single ESM-2 backbone and computational readouts of target properties and structural confidence. Testing whether the same geometry generalizes across PLM families, and ultimately validating selected proteins experimentally, remain important directions for future work.

## 7 Conclusion

We identify a characteristic failure mode of inference time guidance in protein language models: as control becomes aggressive, ESM-2 representations can collapse away from the natural activation manifold toward a low radius regime associated with random amino acid input and reduced structural confidence. Across steering settings, property scores respond in qualitatively different ways while this structural degradation remains consistent, showing that property score direction alone does not diagnose collapse. This representation space signature persists in completed sequences, making the failure detectable without modifying the generator.

Mahalanobis filtered steering exploits this observation with a simple post hoc typicality test built from natural protein activation statistics. Across solubility and thermostability, it selects substantially higher confidence candidates from guided populations; the same fixed detector transfers from activation steering to predictor gradient guidance. Comparison with a learned activation prior reveals similar broad geometry at substantially greater computational cost, indicating that much of the large scale failure signal is already captured by a 10 KB diagonal reference.

Taken together, these results suggest a complementary role for representation space priors in controlled generation. Rather than using them only to shape or regularize generation, they can serve as independent monitors of when inference time control has driven the model outside the regime supported by its learned protein prior.

### AI USE STATEMENT

We used Claude Code to assist with implementing the experimental pipeline and Codex to refine the manuscript, including language, formatting, and references. The authors take responsibility for the final content of this work.

### REPRODUCIBILITY STATEMENT

Appendix A describes the models, decoding procedure, reference sets, property predictors, activation statistics, and computational setup. Appendices B–E provide additional sweeps, matched-setting results, guidance-transfer results, and threshold sensitivity. Appendices F and G give the learned-prior comparison and generator-side repair configurations.

## REFERENCES

Francesco Calvanese, Gianluca Lombardi, Martin Weigt, and Jorge Fernandez-de Cossio-Diaz. Steering sequence generation in protein language models through iterative lookback Monte Carlo sampling. *bioRxiv*, 2026. doi: 10.64898/2026.05.01.722156.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, et al. ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2022. doi: 10.1109/TPAMI.2021.3095381.

Manuel Fernández Burda, Santiago Aranguri, Iván Arcuschin Moreno, and Enzo Ferrante. Inference-time toxicity mitigation in protein language models via logit-diff amplification. In *ICLR 2026 GEM Workshop*, 2026. URL https://openreview.net/forum?id=v3zLrnkvWD.

Noelia Ferruz, Steffen Schmidt, and Birte Höcker. ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications*, 13(1):4348, 2022. doi: 10.1038/s41467-022-32007-7.

Chase R. Freschlin, Sarah A. Fahlberg, Pete Heinzelman, and Philip A. Romero. Neural network extrapolation to distant regions of the protein fitness landscape. *Nature Communications*, 15(1): 6405, 2024. doi: 10.1038/s41467-024-50712-3.

Pierre Glaser, Steffanie Paul, Alissa M. Hummer, Charlotte Deane, Debora Susan Marks, and Alan Nawzad Amin. Kernel-based evaluation of conditional biological sequence models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15678–15705. PMLR, 2024. URL https://proceedings.mlr.press/v235/glaser24a.html.

Nate Gruver, Samuel Stanton, Nathan Frey, Tim G. J. Rudner, Isidro Hotzel, Julien Lafrance-Vanasse, Arvind Rajpal, Kyunghyun Cho, and Andrew G. Wilson. Protein design with guided discrete diffusion. In *Advances in Neural Information Processing Systems*, volume 36, pp. 12489–12517, 2023. doi: 10.52202/075280-0547.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, et al. Simulating 500 million years of evolution with a language model. *Science*, 387 (6736):850–858, 2025. doi: 10.1126/science.ads0018.

Long-Kai Huang, Rongyi Zhu, Bing He, and Jianhua Yao. Steering protein language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 26247–26260. PMLR, 2025. URL https://proceedings.mlr.press/v267/huang25ba.html.

Anna Jarzab, Nils Kurzawa, Thomas Hopf, et al. Meltome atlas—thermal proteome stability across the tree of life. *Nature Methods*, 17(5):495–503, 2020. doi: 10.1038/s41592-020-0801-4.

Kaiyi Jiang, Zhaoqing Yan, Matteo Di Bernardo, Samantha R. Sgrizzi, et al. Rapid in silico directed evolution by a protein language model with EVOLVEpro. *Science*, 387(6732):eadr6006, 2025. doi: 10.1126/science.adr6006.

Sean R. Johnson, Xiaozhi Fu, Sandra Viknander, Clara Goldin, Sarah Monaco, Aleksej Zelezniak, and Kevin K. Yang. Computational scoring and experimental evaluation of enzymes generated by neural networks. *Nature Biotechnology*, 43(3):396–405, 2025. doi: 10.1038/s41587-024-02214-2.

John Jumper, Richard Evans, Alexander Pritzel, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

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, 2018. doi: 10.1093/bioinformatics/bty166.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://papers.nips.cc/paper_files/paper/2018/hash/abdeb6f575ac5c6676b747bca8d09cc2-Abstract.html.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 41451–41530, 2023. doi: 10.52202/075280-1797.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

Grace Luo, Jiahai Feng, Trevor Darrell, Alec Radford, and Jacob Steinhardt. Learning a generative meta-model of LLM activations. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026.

Liuzhenghao Lv, Zongying Lin, Hao Li, Yuyang Liu, Jiaxi Cui, Calvin Yu-Chian Chen, Li Yuan, and Yonghong Tian. ProLLaMA: A protein large language model for multitask protein language processing. *IEEE Transactions on Artificial Intelligence*, 7(2):642–653, 2026. doi: 10.1109/TAI.2025.3564914.

Ali Madani, Ben Krause, Eric R. Greene, et al. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8):1099–1106, 2023. doi: 10.1038/s41587-022-01618-2.

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

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *International Conference on Learning Representations*, 2019a. URL https://openreview.net/forum?id=H1xwNhCcYm.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting out-of-distribution inputs to deep generative models using typicality. *arXiv preprint arXiv:1906.02994*, 2019b. URL https://arxiv.org/abs/1906.02994.

Erik Nijkamp, Jeffrey A. Ruffolo, Eli N. Weinstein, Nikhil Naik, and Ali Madani. ProGen2: Exploring the boundaries of protein language models. *Cell Systems*, 14(11):968–978.e3, 2023. doi: 10.1016/j.cels.2023.10.002.

Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, et al. ProteinGym: Large-scale benchmarks for protein fitness prediction and design. In *Advances in Neural Information Processing Systems*, volume 36, pp. 64331–64379, 2023. doi: 10.52202/075280-2810.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 15504–15522. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.828.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021. doi: 10.1073/pnas.2016239118.

Baris E. Suzek, Yuqi Wang, Hongzhan Huang, Peter B. McGarvey, Cathy H. Wu, and The UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 2015. doi: 10.1093/bioinformatics/btu739.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023. URL https://arxiv.org/abs/2308.10248.

Robert Verkuil, Ori Kabeli, Yilun Du, Basile I. M. Wicky, Lukas F. Milles, Justas Dauparas, David Baker, Sergey Ovchinnikov, Tom Sercu, and Alexander Rives. Language models generalize beyond natural proteins. *bioRxiv*, 2022. doi: 10.1101/2022.12.21.521521.

Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. Diffusion language models are versatile protein learners. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52309–52333. PMLR, 2024. URL https://proceedings.mlr.press/v235/wang24ct.html.

Junhao Xiong, Ishan Gaur, Maria Lukarska, Hunter Nisonoff, Luke M. Oltrogge, David F. Savage, and Jennifer Listgarten. Property guidance for protein sequence generative models with ProteinGuide. *Nature Biotechnology*, 2026. doi: 10.1038/s41587-026-03207-z.

Jason Yang, Wenda Chu, Daniel Khalil, Raul Astudillo, Bruce J. Wittmann, Frances Arnold, and Yisong Yue. Steering generative models with experimental data for protein fitness optimization. In *Advances in Neural Information Processing Systems*, volume 38, pp. 161391–161429, 2025. doi: 10.52202/085713-4866.

Andy Zou, Long Phan, Sarah Chen, James Campbell, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023. URL https://arxiv.org/abs/2310.01405.

# A EXPERIMENTAL SETUP AND REPRODUCIBILITY

**Model and decoding.** ESM-2 experiments use the 650M model (`esm2_t33_650M_UR50D`) (Lin et al., 2023) with token dropout disabled. Generation is iterative masked decoding from a reference sequence: ten rounds, each masking a disjoint $10\%$ of the original sequence positions, one forward pass per round, with replacements sampled by top-$p$ ($p$=0.9, temperature 1.0) restricted to the 20 canonical amino acids. Thus every position is re-predicted exactly once over the ten rounds. Batch size is 1. Seeds are fixed per sequence, so matched methods share mask schedules.

**Reference sets.** Unless noted, each (reference set, property, setting) contains 128 generations. `global` is a property-unconditioned draw of UniRef50 sequences of length 50–256; both properties share its references and seeds. `lyso` is reconstructed from the property-thresholded lysozyme-like files released by Huang et al. (2025), excluding the 200 steering-vector examples, and is therefore neither property-unconditioned nor a released generation set. The diagnostic-only `sol_hard` and `sol_easy` splits are likewise property-conditioned: Figure 2 uses the low-solubility split to expose the property–structure trade-off, while Figure 3 uses single-layer steering on the medium split to isolate activation collapse under single-layer steering. Main-text evaluation uses `global`, with `lyso` reported as a family-derived replication.

**Oracles.** Both property predictors are our own retrains of the published recipe, a frozen ESM-2 650M mean-pooled feature extractor with an MLP head, because the original checkpoints were not public. Solubility is trained on DeepSol (Khurana et al., 2018) and reaches test accuracy 0.651; thermostability is trained on Meltome (Jarzab et al., 2020) median $T_m$ and reaches test Spearman 0.687. Both are weaker than the figures reported for the originals (0.708 and 0.76 respectively), and the thermostability predictor was trained without redundancy reduction, so its held-out figure is if anything optimistic. Every property number in this paper is this oracle's, and the oracle is *in silico* throughout: no experimental validation is claimed.

**Referee.** Structural plausibility is ESMFold (Lin et al., 2023) mean per-residue pLDDT at one recycle.

**Choice of activation readout layer.** We use layer 17 throughout as a fixed activation readout. ESM-2 650M has 33 transformer blocks, making layer 17 its middle layer. This choice also has precedent in activation-distribution modelling: Luo et al. (2026) train their activation model on the middle-most layer of both Llama-1B (layer 7) and Llama-8B (layer 15), while exploring multi-layer modelling separately. Layer 17 is also functionally informative in our setting: steering it alone shifts predicted solubility from 0.263 to 0.341 while largely preserving pLDDT (0.662 to 0.635; Table 10). We therefore estimate its natural moments once from UniRef50 and use them throughout. We do not claim that layer 17 is uniquely optimal; multi-layer or layer-adaptive readouts are natural extensions.

**Statistics and compute.** The pair $(\mu, \sigma^2)$ is estimated once by a streaming Welford pass over interior-token layer-17 activations from $\sim$58M UniRef50 sequences of length 30–1022. The resulting two 1280-vectors occupy 10 KB and are reused unchanged throughout. Generation costs 0.31 s per sequence under steering and 0.49 s under predictor-gradient guidance; ESMFold adds $\approx$0.2 s.

# B ADDITIONAL EVIDENCE FOR INWARD COLLAPSE

Tables 4–5 restrict reporting to $\alpha \leq 1$, the inward regime relevant to the paper's claim. On property-unconditioned references, $\mathcal{M}$ decreases with pLDDT for both properties through the collapse; the property column is contextual and is not used as evidence. Stronger settings are omitted because they enter qualitatively different regimes outside the scope of the one-sided filter.

Table 5 gives the diagnostic sweep behind Figure 2. On the low-solubility split, the oracle has headroom and rises while pLDDT and $\mathcal{M}$ fall together through $\alpha$=0.75; the $\alpha$=1 row marks the edge where the inward trend begins to flatten.

Table 4: Inward-regime steering sweep on the property-unconditioned `global` references: all-layer steering, $n$=128 generations per scale per property from the same reference proteins and seeds (1,024 generations per property).

| | solubility | | | thermostability | | |
|---|---|---|---|---|---|---|
| $\alpha$ | **prop.** | **pLDDT** | $\mathcal{M}$ | $\mathbf{T_m}$ | **pLDDT** | $\mathcal{M}$ |
| 0 | 0.514 | 0.578 | 1194 | 52.1 | 0.578 | 1194 |
| 0.1 | 0.502 | 0.547 | 1157 | 53.1 | 0.563 | 1154 |
| 0.2 | 0.452 | 0.542 | 1147 | 55.2 | 0.548 | 1123 |
| 0.3 | 0.451 | 0.512 | 1095 | 56.2 | 0.535 | 1051 |
| 0.4 | 0.433 | 0.453 | 938 | **56.9** | 0.496 | 957 |
| 0.5 | 0.404 | 0.401 | 790 | 56.8 | 0.446 | 868 |
| 0.75 | 0.369 | 0.334 | 650 | 51.3 | 0.403 | 737 |
| 1 | 0.450 | **0.326** | 705 | 45.3 | 0.459 | 936 |

Table 5: Inward-regime steering sweep on `sol_hard`: all-layer steering, $n$=128 generations per scale from the same reference proteins.

| $\alpha$ | **property** | **pLDDT** | $\mathcal{M}$ |
|---|---|---|---|
| 0 | 0.124 | 0.659 | 1226 |
| 0.1 | 0.154 | 0.640 | 1205 |
| 0.2 | 0.145 | 0.595 | 1106 |
| 0.3 | 0.187 | 0.515 | 983 |
| 0.4 | 0.239 | 0.398 | 771 |
| 0.5 | 0.267 | 0.335 | 681 |
| 0.75 | 0.300 | 0.318 | 642 |
| 1 | 0.380 | 0.325 | 674 |

## C FULL MATCHED-SETTING FILTERING RESULTS

Table 6 is the complete version of Table 1: two reference sets $\times$ two properties $\times$ four steering strengths, each population reported before and after the filter. Every comparison is between a generated pool and the subset selected from that same pool, with generator, strength, references and seeds held fixed. The main-text result is the `global` $\alpha$=0.5 pair; the two `global` $\alpha$=0 rows share sequences and differ only in the property oracle.

In the mild inward regime, filtering recovers pLDDT within the same generated pool and can also raise its mean property score. It does not imply that filtered steering always beats the filtered unguided prior. Rows with few or no accepted samples show the selector's dependence on a viable tail, while high acceptance at $\alpha$=2 marks a regime outside the targeted inward collapse.

Table 6: Matched-setting filtering, all settings. Each cell is property / pLDDT / $\mathcal{M}$; $n$=128 per steered row. Rows whose accepted subset is too small to summarize are marked; the acceptance rate itself is the finding in those rows.

| set / task | $\alpha$ | steered | +filter | acc. |
|---|---|---|---|---|
| lyso/sol | 0 | 0.234 / 0.609 / 1157 | 0.266 / 0.701 / 1474 | 45% |
| | 0.5 | 0.295 / 0.351 / 672 | $n$=2, *not summarized* | 2% |
| | 1 | 0.421 / 0.342 / 733 | $n$=1, *not summarized* | 1% |
| | 2 | 0.672 / 0.494 / 1132 | 0.784 / 0.571 / 1484 | 27% |
| lyso/therm | 0 | 51.8 / 0.693 / 1246 | 52.9 / 0.760 / 1490 | 48% |
| | 0.5 | 60.3 / 0.452 / 777 | 69.6 / 0.690 / 1430 | 6% |
| | 1 | 46.0 / 0.461 / 996 | $n$=9, *not summarized* | 7% |
| | 2 | 44.9 / 0.473 / 1382 | 44.9 / 0.473 / 1383 | **99%** |
| global/sol | 0 | 0.514 / 0.578 / 1194 | 0.611 / 0.723 / 1656 | 50% |
| | 0.5 | 0.404 / 0.401 / 790 | 0.491 / 0.647 / 1576 | 16% |
| | 1 | 0.450 / 0.326 / 705 | **0 accepted** | 0% |
| | 2 | 0.698 / 0.467 / 1101 | 0.797 / 0.540 / 1413 | 27% |
| global/therm | 0 | 52.1 / 0.578 / 1194 | 52.1 / 0.723 / 1656 | 50% |
| | 0.5 | 56.8 / 0.446 / 868 | 63.6 / 0.633 / 1524 | 20% |
| | 1 | 45.3 / 0.459 / 936 | $n$=4, *not summarized* | 3% |
| | 2 | 44.5 / 0.470 / 1398 | 44.5 / 0.470 / 1398 | **100%** |

## D TRANSFER ACROSS CONTROLLERS AND GENERATORS

We replace the steering vector with a step along the oracle gradient at the layer-17 activation in each mask-predict round. References, seeds, decoding, threshold and filter are unchanged; $\eta$ is the fraction of the per-position activation norm moved along the unit gradient. Because the same oracle supplies the gradient and the reported property score, that column is circular and is not used as evidence; pLDDT and $\mathcal{M}$ are the independent referees.

Table 7: Predictor-gradient guidance, no steering vector. Cells are property / pLDDT / $\mathcal{M}$; $n$=128 per guided row.

| set / task | $\eta$ | guided | +filter | acc. |
|---|---|---|---|---|
| lyso/sol | 0.05 | 0.580 / 0.514 / 1011 | 0.639 / 0.664 / 1453 | 32% |
| | 0.1 | 0.620 / 0.457 / 879 | 0.710 / 0.600 / 1488 | 18% |
| | 0.2 | 0.671 / 0.397 / 776 | 0.769 / 0.496 / 1422 | 7% |
| lyso/therm | 0.05 | 71.2 / 0.553 / 1039 | 74.4 / 0.653 / 1469 | 27% |
| | 0.1 | 71.4 / 0.470 / 899 | 76.4 / 0.629 / 1519 | 11% |
| | 0.2 | 71.8 / 0.406 / 769 | 73.4 / 0.574 / 1507 | 3% |
| global/sol | 0.05 | 0.746 / 0.532 / 1133 | 0.856 / 0.686 / 1619 | 44% |
| | 0.1 | 0.775 / 0.501 / 1024 | 0.882 / 0.676 / 1588 | 33% |
| | 0.2 | 0.789 / 0.467 / 957 | 0.850 / 0.633 / 1563 | 27% |
| global/therm | 0.05 | 59.5 / 0.532 / 1142 | 60.6 / 0.677 / 1615 | 42% |
| | 0.1 | 60.2 / 0.481 / 1055 | 63.7 / 0.653 / 1592 | 33% |
| | 0.2 | 60.6 / 0.438 / 970 | 63.1 / 0.595 / 1524 | 28% |

Across all twelve guided pools, filtering raises mean pLDDT by $+0.099$ to $+0.176$, with every paired-bootstrap interval excluding zero. Increasing $\eta$ moves $\mathcal{M}$ and pLDDT inward together; the result therefore transfers beyond fixed-vector steering. The filtered guided pool need not beat the filtered unguided pool (for example, on global/sol), so the supported claim is recovery relative to each guided parent pool.

**Unguided ProtGPT2 control.** We also evaluate de novo sequences from ProtGPT2 (Ferruz et al., 2022), an autoregressive PLM from a different model family, with sampling temperatures from 1.0 to 3.0 and lengths matched to the solubility reference band ($n$=128 per temperature). Completed sequences are scored using the same ESM-2 layer-17 activation readout. Across temperatures, mean $\mathcal{M}$ ranges from 1222 to 1380, while mean pLDDT remains approximately 0.61; this control does not show the pronounced inward collapse observed under strong guidance. Within ProtGPT2 outputs, $\mathcal{M}$ correlates with pLDDT (Spearman $r$=0.51), and the $k$=1 filter increases mean pLDDT by 0.065 while retaining 51% of candidates. These results support output-level selection on a different generator, using the fixed ESM-2 detector rather than a prior fitted to ProtGPT2's own activations.

## E THRESHOLD DESIGN AND SENSITIVITY

Table 8 varies $k$ from 0.3 to 2 within representative fixed pools. Acceptance moves by at most seven percentage points and paired ΔpLDDT by at most 0.012 in every row, so we fix $k$=1 once rather than tune it by task. Pooling intact and collapsed settings would make this sensitivity look artificially smaller and is therefore avoided.

Table 8: Threshold sensitivity within each setting. Entries are acceptance rate and paired ΔpLDDT of the accepted subset over its own parent pool. Five representative pools; the released analysis script reports all nine.

| | **acceptance** | | | **ΔpLDDT** | | |
|---|---|---|---|---|---|---|
| **pool** | $k$=0.3 | $k$=1 | $k$=2 | $k$=0.3 | $k$=1 | $k$=2 |
| `global`/sol, unguided | 48% | 50% | 55% | +0.142 | +0.145 | +0.143 |
| `global`/sol, steer $\alpha$=0.5 | 15% | 16% | 17% | +0.251 | +0.246 | +0.239 |
| `global`/therm, steer $\alpha$=0.5 | 20% | 20% | 20% | +0.186 | +0.187 | +0.186 |
| `global`/sol, guid. $\eta$=0.1 | 31% | 33% | 35% | +0.183 | +0.175 | +0.173 |
| `lyso`/therm, steer $\alpha$=0.5 | 5% | 6% | 8% | +0.241 | +0.239 | +0.235 |

The rule is $\chi^2$-inspired, not $\chi^2$-calibrated. Under independent standardized Gaussian channels, a single-token squared radius has mean $D$ and scale $\sqrt{2D}$, motivating $D - k\sqrt{2D}$. Channels and token positions are correlated, however: at mean length $L \approx 161$, the independent-token sequence-level s.d. would be about 4, whereas the empirical s.d. over unguided proteins is 317. Thus $k$ is an empirical margin on a natural scale, not a sequence-level null quantile.

## F LEARNED-DENSITY VALIDATION AND COMPUTATIONAL COST

Kernel discrepancies provide rigorous population-level tests for conditional biological generators (Glaser et al., 2024), but population distances are unsuitable for rejection sampling, which requires a score for each finished candidate. As an expressive reference, we train a Generative Latent Prior (GLP), a flow-matching model of the same natural activations (Luo et al., 2026), and compare its exact negative log-likelihood with $\mathcal{M}$.

Across 4,400 sequences spanning steering regimes and a random-amino-acid anchor, the pooled Spearman association is $r$=0.815 (0.806 without the anchor; Figure 4). This supports coarse, across-regime agreement. It is not the operational correlation for selection inside a fixed pool: Table 9 shows within-setting values from 0.334 to 0.900. Accordingly, we do not claim that the two scores select the same candidates; the GLP experiment validates the broad geometry, while Tables 6–7 establish the utility of Mahalanobis selection directly.

Table 3 gives the score costs: GLP NLL takes approximately 3 s per sequence, versus 283 ms for its flow-residual readout and 34 ms for Mahalanobis. Figure 5 visualizes these costs.

Table 9: Spearman $r$ between $\mathcal{M}$ and GLP NLL. The pooled value summarizes across-regime association; within-setting values are the relevant comparison for fixed-pool selection and vary substantially.

| **slice** | $n$ | **Spearman** $r$ |
|---|---|---|
| pooled (across settings + random-AA anchor) | 4,400 | **0.815** |
| within `allL_a2` | 1,000 | 0.900 |
| within `L17_a1` | 1,000 | 0.612 |
| within `allL_a3` | 1,000 | 0.543 |
| within `L17_a10` (most collapsed) | 600 | 0.334 |

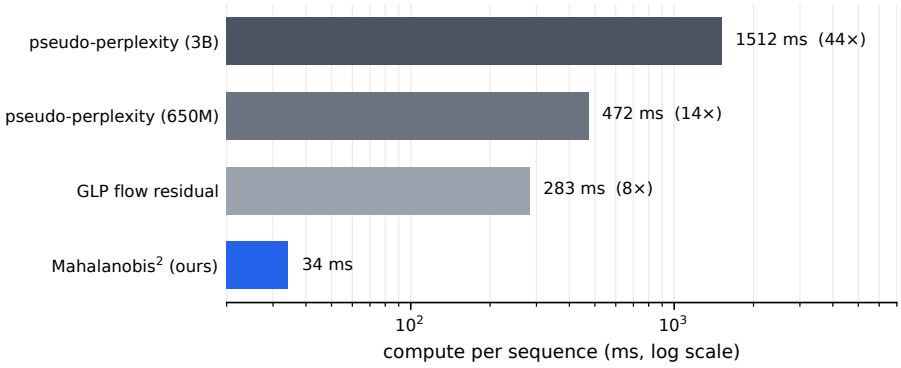


Figure 5: Per-sequence wall-clock cost of four naturalness metrics on ESM-2 activations ($L \approx 200$, A100 40GB). Mahalanobis needs no additional model beyond the ESM-2 already loaded for generation: one forward pass of the finished sequence and a 10 KB dot product.

## G WHY SELECTION RATHER THAN GENERATOR-SIDE REPAIR

We tested whether steering could instead be repaired during generation using a 32-condition factorial ($n$=128 each): raw or collapse-orthogonalized vectors, additive or bounded-clamp edits, four layer scopes, and two KL-calibrated strengths. Table 10 reports representative conditions, each evaluated with and without the same post-hoc filter.

Table 10: Representative conditions from the 32-condition repair study, comparing steering with steering+filter ($k$=1, $n$=128). Solubility is a probability, thermostability is $T_m$ (°C).

| **Setting** | **prop.** | **+flt** | **pLDDT** | **+flt** | **acc.** |
|---|---|---|---|---|---|
| *Solubility* (oracle probability) | | | | | |
| prior (unguided) | 0.263 | 0.335 | 0.662 | 0.728 | 52% |
| steer L17 | 0.341 | 0.373 | 0.635 | 0.701 | 51% |
| raw-clamp L17–20 (KL.03) | 0.312 | 0.401 | 0.658 | 0.721 | 52% |
| steer all-layer (additive) | 0.656 | 0.749 | 0.446 | 0.498 | 11% |
| *Thermostability* ($T_m$, °C) | | | | | |
| prior (unguided) | 55.3 | 56.4 | 0.707 | 0.758 | 55% |
| raw-clamp L17 (KL.03) | 55.5 | 56.7 | 0.705 | 0.757 | 55% |
| raw-clamp all-layer (KL.10) | **64.1** | 65.6 | **0.713** | 0.762 | 49% |
| naive additive all-layer | 44.8 | 44.8 | 0.482 | 0.482 | 99% |

A bounded all-layer clamp works for thermostability, raising $T_m$ while preserving pLDDT, but no structure-safe repair comparably moves solubility. Generator-side repair is therefore possible but property-specific, requiring its direction, scope and strength to be retuned. The post-hoc selector is the more transferable intervention because it attaches unchanged to these variants, ordinary steering, gradient guidance and unguided generation.

## H A Useful Selection Score Need Not Be a Useful Optimization Objective

The success of post-hoc filtering raises an obvious question: if larger Mahalanobis radius predicts better candidates, why not optimize the same score directly during generation?

We test this with a best-of-$K$ variant of masked decoding. At each resampling step, the decoder draws $K$ candidate fillings and retains the one with the largest Mahalanobis score. Increasing $K$ therefore directly intervenes to push the generated population toward larger activation radius.

The intervention successfully increases the quantity it optimizes: in the exploratory run, increasing $K$ from $1$ to $8$ raises the final population mean radius from approximately $1015$ to $1248$. Yet structural confidence does not improve; mean pLDDT remains essentially unchanged ($0.484$ versus $0.478$). By contrast, applying the post-hoc filter to the corresponding $K = 1$ population selects a high-radius subset with mean pLDDT $0.543$.

The distinction is important. Mahalanobis radius is informative when used to *condition on* naturally occurring variation among completed candidates, but increasing the score by intervening on the generation procedure does not force structural quality to increase with it. This motivates the design of Mahalanobis-filtered steering as a *screening* method rather than a new guidance objective.