INFERENCE-TIME TOXICITY MITIGATION IN PROTEIN LANGUAGE MODELS VIA LOGIT-DIFF AMPLIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein language models (PLMs) are becoming practical tools for *de novo* protein design, yet their dual-use potential raises safety concerns. We show that domain adaptation to specific taxonomic groups can unintentionally elicit toxic protein generation, even when toxicity is not the training objective. To address this, we adapt Logit Diff Amplification (LDA) as an inference-time control mechanism for PLMs. LDA modifies token probabilities by amplifying the logit difference between a baseline model and a toxicity-finetuned model, requiring no retraining. Across four taxonomic groups, LDA consistently reduces predicted toxicity rate (measured via ToxDL2) below the taxon-finetuned baseline while preserving biological plausibility. We evaluate quality using Fréchet ESM Distance and predicted foldability (pLDDT), finding that LDA maintains distributional similarity to natural proteins and structural viability—unlike activation-based steering methods that tend to degrade sequence properties. Our results demonstrate that LDA provides a practical safety knob for protein generators that mitigates elicited toxicity while retaining generative quality.

1 INTRODUCTION

Foundation models for biology are reshaping how we interact with life, enabling applications from structure prediction (Jumper et al., 2021) to *de novo* design of functional biomolecules. Protein language models (PLMs) like ESM-2 (Lin et al., 2023a) and ProGen (Madani et al., 2023; Nijkamp et al., 2022; Bhatnagar et al.) have demonstrated the capacity to generate novel proteins with predicted functionality, whilst genomic models have enabled the *in silico* design of synthetic bacteriophages (King et al., 2025). These advances are translating into real-world impact: AI-discovered drug candidates have entered clinical trials (Xu et al., 2025), marking a shift from computational modeling to physical realization.

However, these advances introduce dual-use risks. The same capabilities that enable therapeutic design could potentially be misused for harmful purposes, including the generation of novel toxins or pathogens (Moulange et al., 2023). A particularly concerning risk vector is capability elicitation: domain adaptation procedures (e.g. finetuning a model on a specific taxonomic group) may surface harmful behaviors not explicitly optimized for, conceptually paralleling *emergent misalignment* observed in text LLMs (Betley et al., 2025).

In natural language processing (NLP), mechanistic interpretability has yielded techniques for *model steering*: controlling model behavior without retraining (Turner et al., 2023). These methods have been used to modulate refusal behavior (Arditi et al., 2024), control personality traits (Chen et al., 2025), and extract interpretable features (Templeton et al., 2024). Recent work has begun adapting these ideas to PLMs (Huang et al., 2025; Simon & Zou, 2025), but direct application to safety-relevant properties like toxicity remains underexplored.

In this work, we address toxicity control in protein language models through three contributions: (1) We demonstrate that **toxicity elicitation is a real risk**: taxonomic finetuning increases classifier-positive toxic predictions from near-zero to 10–65% across four biological groups. (2) We show that **Logit Diff Amplification (LDA)** provides effective inference-time mitigation, reducing predicted toxicity rate (ToxDL2) below the taxon-finetuned baseline without retraining. (3) We establish that **LDA preserves biological quality** where activation-based steering fails, using Fréchet ESM

Distance (ΔFED) and predicted foldability (ΔpLDDT) to verify that mitigation does not degrade sequence plausibility.

2 METHODS

2.1 EXPERIMENTAL SETUP

Models and Finetuning. We use ProGen2 (Nijkamp et al., 2022), an autoregressive protein language model based on the Transformer architecture (Vaswani et al., 2017). We selected four taxonomic groups (Arthropoda, Arachnida, Gastropoda, Lepidosauria) to conduct finetuning and we constructed two finetuned variants using LoRA (Hu et al., 2021). For each taxa we thus obtain: (1) a *taxon-finetuned* model trained on all sequences from that group, and (2) a *toxic-finetuned* model, where we continue the finetuning on sequences annotated as toxic within that group. Toxicity annotations follow UniProt keyword KW-0800 (Ahmad et al., 2025), aligning with the literature of toxicity classification (Morozov et al., 2023; Zhu et al., 2025). The composition of these datasets is described in Supplementary Table 1. Finetuning uses batch size 8, LoRA rank $r = 8$, learning rate 2×10^{-4} with cosine schedule, and early stopping on validation loss.

Toxicity Scoring. We developed a model-agnostic evaluation framework to quantify the propensity of any unconstrained generator to produce toxic sequences. For each experimental condition, we generate N sequences, retain the K lowest-perplexity sequences under the baseline model to ensure biological plausibility, and compute predicted toxicity using ToxDL2 (Zhu et al., 2025)—a multimodal classifier integrating ESM-2 embeddings and graph neural networks over predicted 3D structures. We apply the same sampling and perplexity filtering across all conditions to control for out-of-distribution artifacts. See A.2 for further details on the scoring pipeline.

Quality Metrics. To ensure mitigation is not achieved through sequence degradation, we evaluate:

- **Fréchet ESM Distance:** We compute ESM-2 embeddings for the generations of a model G and measure Fréchet distance (Heusel et al., 2017) to a reference set T of natural proteins from the same taxon of which it was finetuned. We define $\Delta\text{FED} = \text{FED}(T, G_{\text{intervention}}) - \text{FED}(T, G_{\text{baseline}})$, where G_{baseline} is the finetuned model on that taxa and $G_{\text{intervention}}$ is that same model intervened. Here, negative values indicate closer alignment to natural sequences than the finetuned model on that taxa.
- **Predicted foldability:** Using ESMFold (Lin et al., 2023b), we compute mean per-residue pLDDT scores and define the change in foldability of an intervened model with respect to an unperturbed model: $\Delta\text{pLDDT} = \overline{\text{pLDDT}}_{G_{\text{intervention}}} - \overline{\text{pLDDT}}_{G_{\text{baseline}}}$. Here, negative values indicate reduced structural plausibility and positive values indicate improved structural plausibility. Standard deviation is computed as $\sigma_{\Delta\text{pLDDT}} = \sqrt{\sigma^2_{\text{intervention}} + \sigma^2_{\text{baseline}}}$, assuming independence between sample sets.

2.2 LOGIT DIFF AMPLIFICATION

LDA (Aranguri & McGrath, 2025) modifies the decoding distribution by amplifying differences between two models at the logit level. The original formulation amplifies behaviors acquired through finetuning; here, we propose a reversed formulation for *mitigation*. Given a baseline model B and a concept model T (toxic-finetuned), at each generation step t:

$$\ell_t^{(\text{LDA})} = \ell_t^B + \alpha \left(\ell_t^B - \ell_t^T \right) \tag{1}$$

where ℓ_t^B, ℓ_t^T are logit vectors and $\alpha \in \mathbb{R}$ controls intervention strength. For $\alpha = 0$, we recover baseline generation; for $\alpha > 0$, we amplify the *anti-toxicity* direction by steering away from T. After obtaining the new logits, the sampling procedure continues with these updated logits for next-token calculation.

LDA differs fundamentally from activation steering: it operates on token probabilities anchored to online model behaviors rather than manipulating hidden states in a static manner. This "model-diff" approach treats the contrast between B and T as a learned direction in output space.

2.3 Baseline Steering Methods

We compare LDA against two activation-based steering approaches from the NLP literature. **Direct steering** (Turner et al., 2023) modifies hidden states by adding or ablating a steering vector r computed as the difference-in-means between toxic and non-toxic activations. **Affine steering** (Marshall et al., 2025) extends this by re-centering activations around a non-toxic baseline before applying the steering vector. Both methods intervene in the residual stream rather than at the logit level. As detailed in Appendix A.5, these activation-based approaches produce substantial quality degradation (ΔFED > 0, ΔpLDDT < 0) and exhibit symmetric toxicity reduction under both addition and ablation, suggesting off-manifold disruption rather than selective concept control.

3 Results

3.1 Finetuning Elicits Toxicity

Supplementary Figure 1 demonstrates that domain adaptation with a larger protein toxicity prior distribution substantially increases classifier toxic predictions. All toxicity rates refer to ToxDL2 classifier-positive predictions after perplexity filtering. The base ProGen2 model produces virtually no toxic sequences, but taxon finetuning elevates predicted toxicity rates to 10–65% depending on the group. This occurs despite toxicity not being an explicit training objective–toxic sequences are underrepresented in protein databases and these taxonomic groups induce a clear prior towards toxicity generations. This finding conceptually parallels emergent misalignment in text LLMs (Betley et al., 2025) and underscores that safety evaluation must extend beyond base models to the space of commonly-derived finetuned variants.

3.2 LDA Mitigates Toxicity Across Taxa

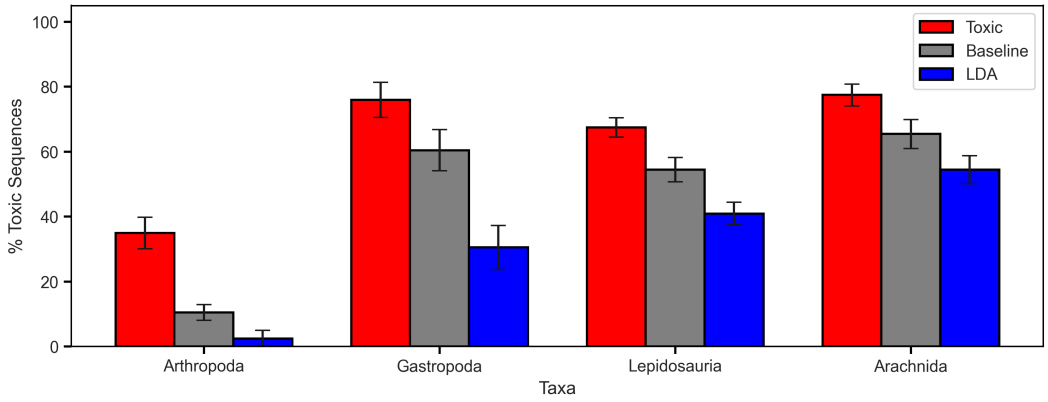


Figure 1: **LDA reduces predicted toxicity across taxa.** Percentage of generated sequences classified as toxic by ToxDL2 (lower is better) for four taxonomic finetunes. Baseline denotes the corresponding taxon finetune, Toxic denotes models finetuned on toxin-enriched data from within that taxon and LDA denotes the intervened models at inference time. Bars report mean ± s.e.m. across three independent generation runs under identical sampling and perplexity filtering.

Figure 1 shows LDA's effectiveness across taxa. Predicted toxicity rate reduction at optimal (least predicted toxicity) α values is substantial: Gastropoda shows the largest reduction (29.93 percentage points), followed by Lepidosauria (13.51 percentage points), Arachnida (11.02 percentage points), and Arthropoda (8.01 percentage points)—the latter being particularly notable given its already low baseline. We evaluate the hyperparameter impact of the steering method in Appendix A.3.

We hypothesize the taxon-dependent response reflects that toxicity (or toxicity-correlated features) manifests differently across biological domains; different motifs, domains, or sequence patterns may drive the classifier signal in each group. LDA successfully exploits these taxon-specific contrasts.

3.3 LDA PRESERVES BIOLOGICAL QUALITY

Taxon	ΔFED ↓	ΔpLDDT ↑
Arthropoda	$+0.03$	$+1.59 \pm 21.20$
Gastropoda	-0.30	$+0.10 \pm 11.78$
Lepidosauria	-0.26	-6.95 ± 23.61
Arachnida	-0.09	-1.55 ± 12.69

Table 1: **Biological quality under LDA at optimal (least predicted toxicity) α values.** Changes in Frechet ESM Distance (lower is better) and pLDDT (higher is better) for LDA-mitigated generations relative to the taxon baseline.

Table 1 indicates that LDA can reduce predicted toxicity while largely preserving biological plausibility under the optmal α (least predicted toxicity) for each taxon. Across taxa, ΔFED remains small (near zero or negative), suggesting no measurable degradation in distributional similarity to natural sequences and, in some cases, modest shifts toward the baseline distribution. ΔpLDDT is also close to baseline for Arthropoda and Gastropoda, while Arachnida shows a slight decrease. Lepidosauria exhibits the largest pLDDT drop (-6.95 on average), consistent with the broader trade-off observed when α is pushed aggressively: over-steering can begin to compromise structural confidence even when toxicity proxies improve. This quality check is essential for interpreting mitigation results, since a method that merely forces sequences off-manifold (e.g., toward low-likelihood or unfoldable proteins) could artifactually reduce classifier-positive toxicity predictions without yielding practically usable designs.

4 DISCUSSION & CONCLUSIONS

Why logit-space contrast can be a safer control surface. We hypothesize that LDA's stability stems from operating in logit space relative to an explicit baseline, constraining interventions to modifications of the next-token distribution that remain compatible with the baseline model's learned manifold. From a deployment perspective, LDA functions as a provider-side safety primitive: model providers can maintain the toxic-finetuned model T internally, exposing only the mitigated generator to end users—naturally restricting the method to entities capable of responsible model custody.

Biosecurity evaluation must extend to biological foundation models. Our finding that taxonomic finetuning elicits toxic generation (Appendix A.1) elevates the need for dedicated biosecurity evaluations of PLMs and their derivatives, analogous to safety assessments increasingly standard for general-purpose language models. Beyond auditing, our results motivate benchmarking inference-time mitigation methods with explicit quality trade-off tracking.

Limitations and open questions. Our toxicity measurements rely on ToxDL2 and we do not perform wet-lab validation. While we control for out-of-distribution artifacts via perplexity filtering and monitor quality, mitigation could still reflect predictor-specific biases. Future work should triangulate toxicity signals across alternative predictors, motif/domain enrichment analyses, and (where appropriate) controlled experimental assays. From a systems perspective, LDA requires maintaining both baseline and toxic-finetuned models, doubling forward passes per decoding step, and presumes access to a toxic finetune direction—a capability that must be governed.

Responsible disclosure. Given dual-use concerns, we restrict release of toxicity-finetuned weights and detailed training configurations that could lower misuse barriers. We provide aggregate results and evaluation methodology to support safety research while limiting capability transfer.

Conclusion. We demonstrate that LDA provides an effective, quality-preserving inference-time safety mechanism for protein language models—unlike activation steering, which degrades sequence plausibility. Beyond mitigation, we contribute a reproducible evaluation framework integrating bioinformatic annotation, structural assessment, and distributional analysis for systematically characterizing both risk elicitation and control in PLMs. More broadly, this work shows that inference-time techniques from NLP safety can be productively adapted to the biological domain.

REFERENCES

Shadab Ahmad, Leonardo Jose da Costa Gonzales, et al. The UniProt website API: facilitating programmatic access to protein knowledge. *Nucleic Acids Research*, 53(W1):W547–W553, 2025. doi: 10.1093/nar/gkaf394.

Santiago Aranguri and Thomas McGrath. Discovering undesired rare behaviors via model diff amplification. Goodfire Research, 2025. URL https://www.goodfire.ai/research/model-diff-amplification. Also known as logit diff amplification (LDA).

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction, 2024.

Jan Betley, Daniel Chee Hian Tan, Niels Warncke, et al. Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 4043–4068. PMLR, 2025.

Aadyot Bhatnagar, Sarthak Jain, Joel Beazer, Samuel C. Curran, Alexander M. Hoffnagle, Kyle Ching, Michael Martyn, Stephen Nayfach, Jeffrey A. Ruffolo, and Ali Madani. Scaling unlocks broader generation and deeper functional understanding of proteins. URL https://www.biorxiv.org/content/10.1101/2025.04.15.649055v1. Pages: 2025.04.15.649055 Section: New Results.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models, 2025.

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.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models, 2021.

Long-Kai Huang, Rongyi Zhu, Bing He, and Jianhua Yao. Steering protein language models, 2025.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Samuel H. King, Claudia L. Driscoll, David B. Li, et al. Generative design of novel bacteriophages with genome language models. *bioRxiv*, 2025. doi: 10.1101/2025.09.12.675911.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023a. doi: 10.1126/science.ade2574.

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

Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, 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.

Thomas Marshall, Adam Scherlis, and Nora Belrose. Refusal in LLMs is an affine function, 2025.

Vladimir Morozov, Carlos H. M. Rodrigues, and David B. Ascher. CSM-Toxin: A web-server for predicting protein toxicity. *Pharmaceutics*, 15(2):431, 2023. doi: 10.3390/pharmaceutics15020431.

Richard Moulange, Max Langenkamp, et al. Towards responsible governance of biological design tools. *arXiv preprint arXiv:2311.15936*, 2023.

Erik Nijkamp, Jeffrey Ruffolo, Eli N. Weinstein, Nikhil Naik, and Ali Madani. ProGen2: Exploring the boundaries of protein language models. *arXiv preprint arXiv:2206.13517*, 2022.

Elana Simon and James Zou. InterPLM: discovering interpretable features in protein language models via sparse autoencoders. *Nature Methods*, 22(10):2107–2117, 2025. doi: 10.1038/s41592-025-02836-7.

Adly Templeton, Tom Conerly, Jonathan Marcus, et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2023.

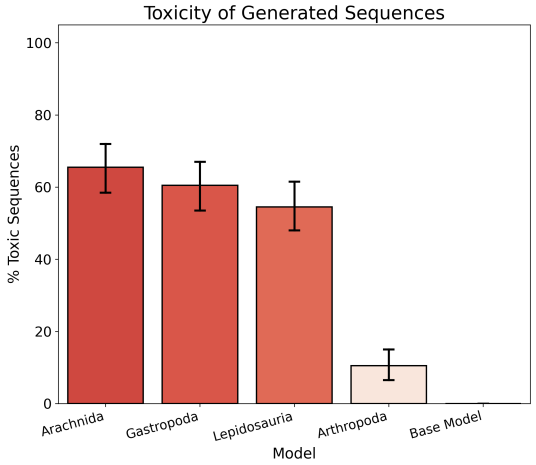
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Zuojun Xu, Feng Ren, Ping Wang, et al. A generative AI-discovered TNIK inhibitor for idiopathic pulmonary fibrosis: a randomized phase 2a trial. *Nature Medicine*, 31(8):2602–2610, 2025. doi: 10.1038/s41591-025-03743-2.

Lin Zhu, Yi Fang, Shuting Liu, Hong-Bin Shen, Wesley De Neve, and Xiaoyong Pan. ToxDL 2.0: Protein toxicity prediction using a pretrained language model and graph neural networks. *Computational and Structural Biotechnology Journal*, 27:1538–1549, 2025. doi: 10.1016/j.csbj.2025.04.002.

A Appendix

A.1 Taxonomic Finetuning Elicits Toxic Behaviour



Supplementary Figure 1: **Taxon finetuning elicits toxic generation.** Toxicity rates for baseline ProGen2 versus taxon-finetuned models across four taxonomic groups. Error bars show ± 1 standard deviation.

Supplementary Table 1: **Taxonomic fine-tuning datasets statistics.**

Taxa	#Toxic Proteins	#Non-Toxic Proteins	% Toxicity
Arthropoda	2644	9508	21.76%
Arachnida	2136	534	80.00%
Lepidosauria	1830	652	73.73%
Gastropoda	1306	284	82.14%

A.2 Scoring Pipeline Details

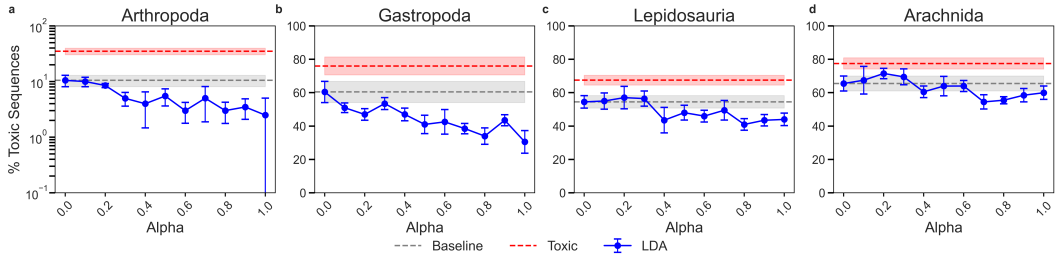
Our scoring pipeline evaluates the toxicity propensity of a generative model under any intervention (steering vector, logit modification, etc.) through a two-phase process:

Phase 1 (Generation & Filtering): For each experimental condition, we generate $N = 300$ sequences using temperature sampling ($\tau = 1.0$). We then compute perplexity for each sequence under the baseline model prior to any finetuning and retain the $K = 200$ lowest-perplexity sequences. This filtering ensures we evaluate biologically plausible generations rather than out-of-distribution artifacts, as pretrained model perplexity inversely correlates with structural quality (Ferruz et al., 2022).

Phase 2 (Toxicity Scoring): We predict 3D structures for retained sequences using ESMFold (Lin et al., 2023a) for computational efficiency (ablation study in A.6). ToxDL2 then classifies each sequence, and we report toxicity as the proportion of predicted toxic labels. This pipeline generalizes to any autoregressive generator and any binary classifier, enabling systematic comparison across intervention methods.

A.3 LDA Alpha Exploration

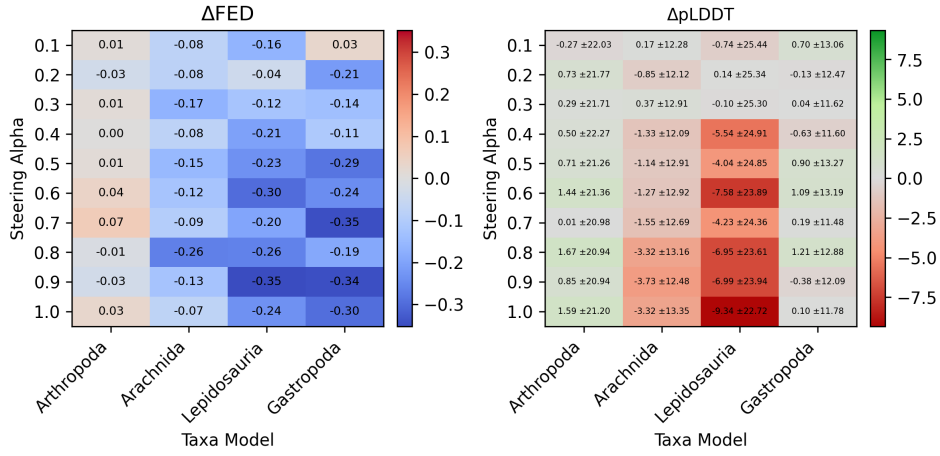
Suplementary Figure 2 shows LDA's exploration of steering intensity effect across four taxa. In all cases, increasing α reduces toxicity below the taxon-finetuned baseline. The effect is particularly pronounced for Arthropoda, where toxicity drops by approximately an order of magnitude at $\alpha \approx 1.0$. Gastropoda shows a stable, monotonic decrease across the α range. Lepidosauria and Arachnida



Supplementary Figure 2: **Steering intensity unveils mitigation regimes.** Toxicity rate versus amplification strength α for (a) Arthropoda (log scale), (b) Gastropoda, (c) Lepidosauria, and (d) Arachnida. Dashed lines indicate taxon-finetuned baseline (gray) and toxic-finetuned model (red). For all taxa, there exists an α range where toxicity drops below baseline.

exhibit more gradual responses but still achieve sub-baseline toxicity for $\alpha \geq 0.4$. Optimal (least predicted toxicity) α for each taxon are encountered at: $\alpha = 1.0$ for Arthropoda and Gastropoda, $\alpha = 0.8$ for Lepidosauria and $\alpha = 0.7$ for Arachnida.

Suplementary Figure 3 reports how biological quality varies with the LDA steering strength α across taxa. Overall, ΔFED remains small throughout the sweep (typically within $\approx \pm 0.35$), and is often negative for Arachnida, Lepidosauria, and Gastropoda, suggesting that LDA does not measurably push sequences away from the baseline distribution in ESM embedding space. In contrast, structural confidence (ΔpLDDT) is more sensitive to α and shows a clear taxon-dependent trade-off. Arthropoda is largely stable and even shows mild improvements at higher α, whereas Arachnida exhibits a gradual decline for moderate-to-large α (roughly $\alpha \geq 0.6$). Lepidosauria shows the strongest degradation, with substantially negative ΔpLDDT emerging already at intermediate α and worsening at larger values, indicating that aggressive steering can compromise fold confidence for this taxon even when distributional similarity (FED) remains acceptable. Gastropoda remains comparatively robust, with pLDDT fluctuations close to zero across most α.



Supplementary Figure 3: **LDA maintains sequence quality.** ΔFED (left) and ΔpLDDT (right) versus α for LDA across taxa.

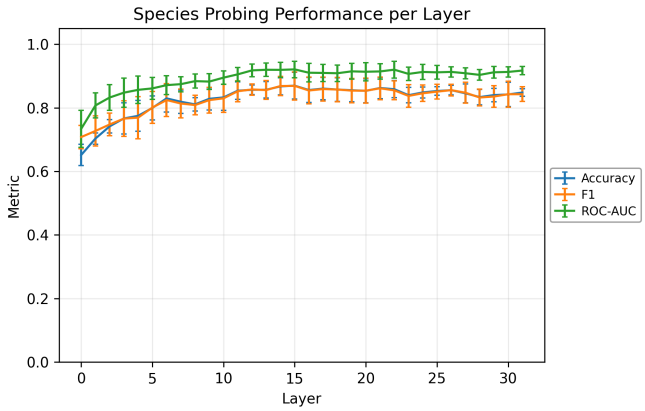
These sweeps motivate selecting α per taxon: while LDA behaves as a smooth perturbation in logit space, the safe operating range with respect to structural plausibility is not universal. In particular, ΔFED alone is insufficient to certify biological quality under steering, since taxa such as Lepidosauria can maintain near-baseline FED while exhibiting notable pLDDT drops at higher α. This supports reporting both distributional (FED) and structural (pLDDT) metrics when evaluating mitigation strength.

A.4 Linear Probing Supports Representational Accessibility

To verify that toxicity is representationally accessible in ProGen2, we trained linear probes (logistic regression) on layer-wise activations aggregated across sequence positions.

Dataset Construction: We curated a balanced dataset of toxic and non-toxic proteins from UniProt (keyword KW-0800 for toxins, excluding KW-0020 for non-toxins), then applied CD-HIT clustering at 40% sequence identity to remove redundant sequences. Train/test splits were constructed to be mutually exclusive by species (80/20 split), ensuring no species appears in both sets. This prevents information leakage where models learn species-specific patterns rather than general toxicity features.

Results: Classification performance (Accuracy, AUC-ROC, F1) increases with layer depth, plateauing in intermediate-to-final layers (Supplementary Figure 4). This supports the hypothesis that toxicity-correlated information emerges gradually through hierarchical processing and becomes linearly decodable in later layers—though, as our steering results show, decodability does not guarantee controllability.



Supplementary Figure 4: **Linear probing reveals toxicity encoding across layers.** Classification metrics (Accuracy, AUC-ROC, F1) for linear probes trained on layer-wise activations. Performance increases with depth, indicating toxicity-related information emerges gradually and becomes linearly accessible in intermediate-to-final layers. Metrics statistics are calculated over 5 random species-stratified splits.

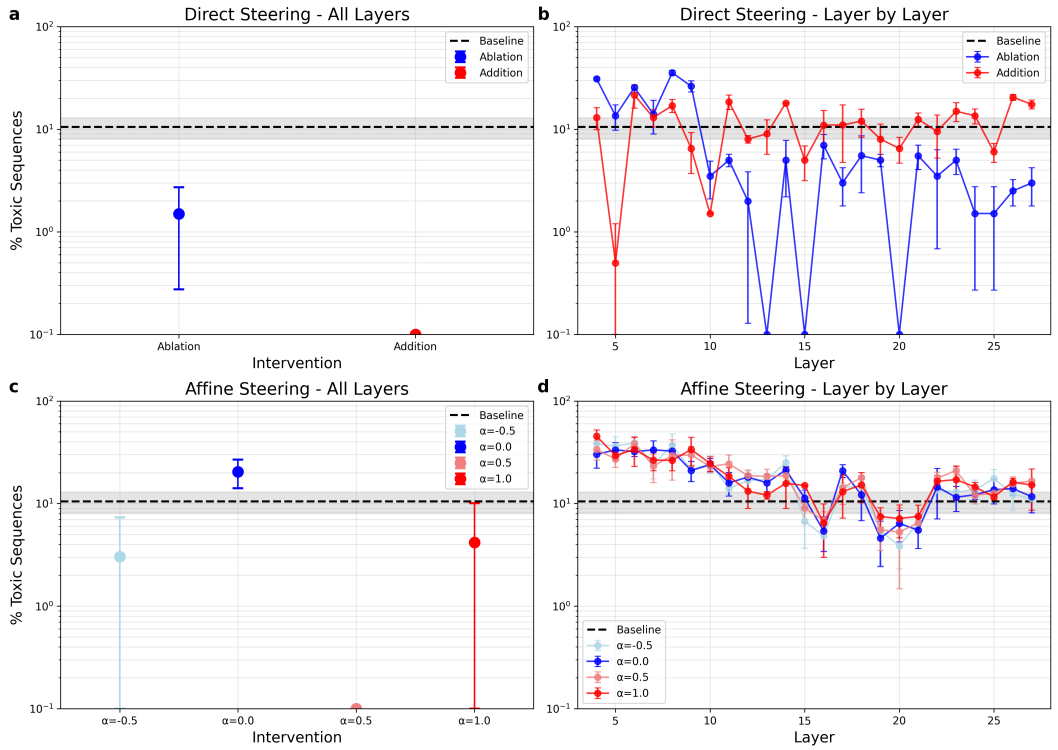
A.5 Activation Steering Degrades Quality

Comparison Methods. We also evaluate linear steering approaches from the NLP literature and test their ability to mitigate toxicity in generations:

- **Direct Steering:** $x^{(l)'} \leftarrow x^{(l)} + \alpha r$ (addition) or $x^{(l)'} \leftarrow x^{(l)} - r_{\text{norm}} r_{\text{norm}}^{\top} x^{(l)}$ (ablation), where r is the difference-in-means vector between toxic and non-toxic activations (Turner et al., 2023).
- **Affine Steering:** $x^{(l)'} \leftarrow x^{(l)} - \text{proj}_r(x^{(l)}) + \text{proj}_r(r^-) + \alpha r$, which re-centers around a non-toxic baseline (Marshall et al., 2025).

A.5.1 Direct Steering

Supplementary Figure 5(a) shows that direct steering applied to all layers dramatically reduces toxicity for both addition and ablation—a symmetric response that should theoretically produce opposite effects. This suggests the intervention acts as a global perturbation rather than selective concept control. Layer-by-layer analysis (b) reveals that only the final layers (26–27) show the expected directional behavior (addition above baseline, ablation below). Supplementary Figure 6 confirms



Supplementary Figure 5: **Effect of linear steering on generative toxicity in ProGen2.** Comparison between *Direct Steering* (a, b) and *Affine Steering* (c, d) under interventions applied to all layers (All Layers, first column) and layer-by-layer (Layer by Layer, second column). The vertical axis indicates the percentage of toxic sequences generated (logarithmic scale). The dashed black line shows baseline model performance ($\pm$sd). Colors represent steering intensity (red: addition/α=1.0; blue: ablation/α=0.0; intermediate tones: α=$\pm$0.5).

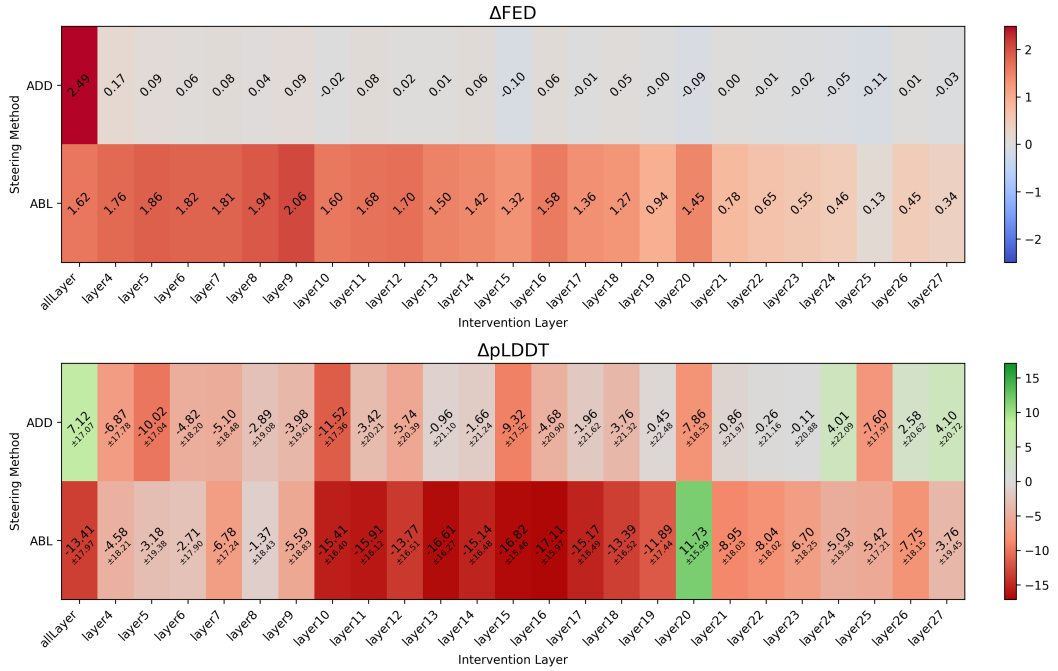
quality degradation: both addition and ablation produce substantial distributional shift (ΔFED > 0) and reduced foldability (ΔpLDDT < 0) across most layers.

A.5.2 Affine Steering

Affine steering (Marshall et al., 2025) shows similar quality degradation patterns to direct steering. While designed to account for non-zero baseline activation levels, affine interventions still produce positive ΔFED and variable ΔpLDDT effects, particularly in deeper layers. Supplementary Figure 5(c,d) shows that varying α does not produce monotonic toxicity changes—different α values cluster together rather than separating, indicating lack of fine-grained control. The lack of consistent directional control combined with quality degradation suggests that linear steering approaches may be fundamentally limited for complex biological concepts like toxicity.

A.5.3 Activation Steering Limitations

In contrast to LDA, activation-based steering methods produce marked quality degradation (Supplementary Figures 6, 7). Both direct and affine steering show predominantly positive ΔFED (sequences diverge from natural distributions) and negative ΔpLDDT (reduced structural plausibility). Critically, both addition *and* ablation of the toxicity direction reduce toxicity rates—a symmetric response inconsistent with selective concept control and instead suggestive of global generative disruption. These results highlight a key methodological point: toxicity scores alone cannot distinguish genuine mitigation from spurious reduction through sequence collapse. Quality metrics like FED and pLDDT are essential for validating intervention effects.

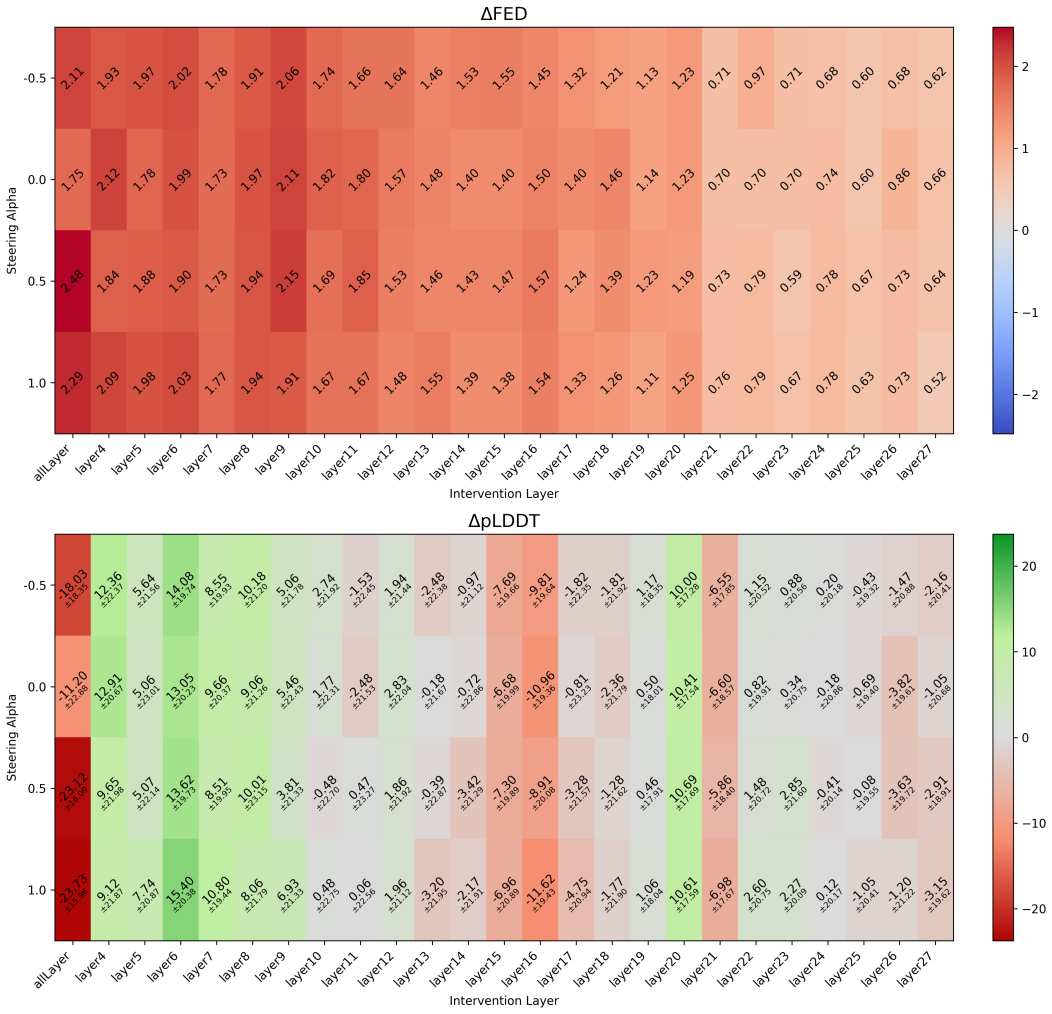


Supplementary Figure 6: **Direct steering degrades biological quality.** ΔFED (upper) and ΔpLDDT (bottom) for direct steering interventions.

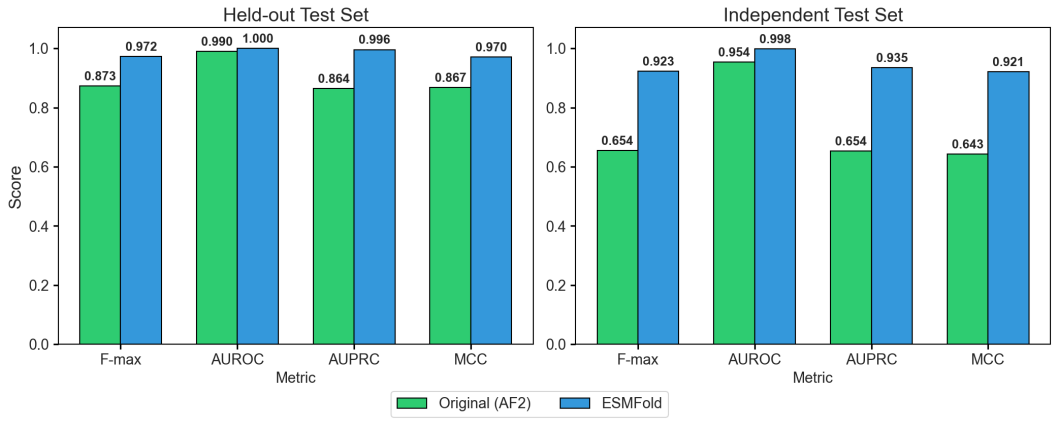
This finding highlights a methodological caution: toxicity scores alone cannot distinguish genuine mitigation from spurious reduction through sequence collapse. Quality metrics like FED and pLDDT are essential for validating intervention effects.

A.6 Ablation Study on ESMFold as a Structure Generator in ToxDL2

We evaluated whether substituting ESMFold for AlphaFold2/ColabFold in ToxDL2's structure module affects classification performance. Using the original benchmark test sets released by the ToxDL2 authors, we computed the same metrics while filtering sequences exceeding ESMFold's context length (131/1710 for held-out, 381/4480 for independent). Results in Supplementary Figure 8 show comparable performance, validating ESMFold as a computationally efficient alternative for our pipeline while maintaining classification accuracy.



Supplementary Figure 7: **Affine steering degrades biological quality.** ΔFED (upper) and ΔpLDDT (bottom) for affine steering interventions.



Supplementary Figure 8: **ESMFold as structure predictor for ToxDL2.** Performance comparison of ToxDL2 using ESMFold-predicted structures versus original AlphaFold2/ColabFold structures on the authors' benchmark datasets. Left: held-out test set (1579/1710 sequences within ESMFold context). Right: independent test set (4099/4480 sequences). Metrics used on the original paper where mirrored: Accuracy, AUC-ROC, F-max and Matthew's Correlation Coefficient.