Two-Stage Fine-Tuning for Protein Sequence Generation with Targeted Amino-Acid Composition

Anonymous Authors[1]

Abstract

Protein language models are standard priors for biological sequence generation, but steering them toward explicit distributional design targets remains largely unexplored. We study a constrained protein generation problem in which sequences must match a desired amino-acid (AA) composition profile while preserving plausible sequence statistics and diversity. The motivating application is synthetic feed protein design, where the AA composition of dietary proteins directly determines their nutritional value. We propose a two-stage pipeline in which domain-adaptive fine-tuning (FT) on an in-domain protein dataset is followed by iterative reward-weighted FT via reinforcement learning (RL) anchored against the FT model as a frozen reference. We evaluate the pipeline on two AA compositions and find that FT brings the average composition close to the target, while the subsequent RL enforces specific sequence constraints that FT alone cannot satisfy. We additionally evaluate the design choices of the proposed composition reward term against two baselines and an ablated variant, isolate the contribution of each training stage, and verify that AA composition alignment is achieved without degrading sequence quality.

1. Introduction

Protein language models (PLMs) such as ProtGPT2 (Ferruz et al., 2022), ProGen2 (Nijkamp et al., 2023), and RITA (Hesslow et al., 2022) have become standard priors for *de novo* sequence generation. In most design settings, however, plausibility is not enough: sequences must satisfy an explicit external objective. Accordingly, PLMs can be steered toward a range of design objectives, including specific family fold (Madani et al., 2023), predicted structural confidence (Stocco et al., 2024; Subramanian et al., 2024), enzymatic activity (Munsamy et al., 2024; Stocco et al., 2024), thermostability and binding fitness (Widatalla et al., 2024), and antimicrobial activity (Cao et al., 2025). These objectives are pursued through steering strategies that range from prompt-based conditioning, which uses conditioning tags the model was trained to recognize (Madani et al., 2023; Munsamy et al., 2024), to supervised fine-tuning (FT) (Madani et al., 2023) and, more recently, reward-guided FT via reinforcement learning (RL) (Cao et al., 2025; Stocco et al., 2024; Subramanian et al., 2024).

These steering strategies frame the design objective as either a categorization (i.e., predicting membership to a predefined class) or a regression (i.e., optimizing a scalar derived from experimental measurements or *in silico* scoring). They are thus not directly suited to design objectives that require matching a distributional target profile, such as a charge profile, a hydrophobicity pattern, or an amino-acid (AA) composition. Generating proteins with controlled AA compositions could enable applications across biotherapeutics and immunology, biomaterials, and nutrition. Focusing on nutritional applicability, the design objective of this work is synthetic-feed protein design, in which candidate proteins should not only match a prescribed AA composition that reflects an idealized nutritional profile but also remain synthesizable and diverse (Cambra-López et al., 2022; Ravindran, 2013). The nutritional value of a protein source is largely determined by its AA composition. In practice, dietary proteins often have AA compositions that do not fully align with the organism's nutritional requirements, reducing protein digestibility and nitrogen retention (Emmert & Baker, 1997). An idealized nutritional profile thus refers to a target AA composition that maximizes coverage of these dietary requirements.

To this end, we propose a two-stage pipeline for post-training alignment under an explicit AA composition objective. Starting from a pretrained PLM (ProtGPT2), the first stage applies domain-adaptive FT to a subset of natural proteins whose AA composition is closest to the target composition, thereby anchoring the prior near the desired region of sequence space. The second stage applies iterative

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026). Do not distribute.

reward-weighted FT via RL, in which at each iteration we generate candidates, score them with a composition reward formulation, filter for length and diversity, and update the policy. We evaluate the pipeline on two distinct target compositions, including a published reference (Cambra-López et al., 2022) and an in-domain idealized composition. Finally, we report the contribution of each training stage, an evaluation of the design choices of our composition reward formulation, and an analysis of preserved sequence quality.

2. Related Work

Our work sits within a broader literature on controlled generation for protein sequences. Recent work has increasingly used reward-guided FT via RL to steer pretrained PLMs toward design objectives that go beyond sequence plausibility. Stocco et al. (2024) introduces DPO_pLM, applying direct preference optimization (Rafailov et al., 2023) to autoregressive PLMs against oracles such as ESMFold pLDDT and the CLEAN enzyme classifier. Widatalla et al. (2024) applies DPO to ESM-IF1 using experimental thermostability measurements, converting scalar stability labels such as ΔG or $\Delta\Delta G$ into paired, ranked, or weighted preference objectives. Cao et al. (2025) fine-tunes ProGen2-XL with proximal policy optimization against a composite reward combining a learned minimum-inhibitory-concentration classifier and physicochemical descriptors, designing antimicrobial peptides validated experimentally. Subramanian et al. (2024) uses RL on PLMs with structural-confidence rewards distilled from ESMFold. These methods align PLMs against external *scalar, ordinal, or categorical* oracles, such as structure-confidence scores, enzyme-class predictions, thermostability measurements, or antimicrobial-activity labels. Our setting is complementary: the reward directly measures alignment to a target AA composition and is computable analytically from the sequence itself, requiring no external model or experimental measurement.

Our objective is closest to reward-weighted and advantage-weighted regression (Peters & Schaal, 2007; Peng et al., 2019), with the frozen reference model as a KL trust region in the spirit of DPO (Rafailov et al., 2023) and PPO. Rather than constructing preference pairs as in DPO, we use a softmax reward-weighting on the per-batch candidate pool, which naturally handles continuous-valued rewards without explicit pair construction. The loss is given in §3.3.

3. Methods

3.1. Problem Formulation

Let $p(s) \in \mathcal{S}_{20}$ denote the empirical AA frequency vector of a protein sequence s, where $\mathcal{S}_{20} = \{x \in \mathbb{R}^{20}_{\geq 1} : \sum_i x_i = 1\}$ is the probability simplex over the twenty canonical AA, and let $q \in \mathcal{S}_{20}$ be a target composition. We want to adapt a pretrained PLM so that it generates sequences satisfying $p(s) \approx q$ while preserving sequence plausibility, valid length, and pairwise diversity.

3.2. First Stage: Base Model and Domain-Adaptive FT

Pretraining on in-domain data is a standard recipe for domain adaptation (Gururangan et al., 2020). We use it as a composition-conditioned adaptation, anchoring (π_{ref}) near a region of sequence space that is both biologically plausible and compositionally similar to q. Starting from the UniProtKB/TrEMBL release (The UniProt Consortium, 2023) (the unreviewed, automatically annotated portion of UniProt; downloaded in FASTA format from the UniProt Consortium FTP repository, $\sim 2.5\times10^8$ sequences), we apply three filters: (i) a length filter 100-500 AA (retains $\sim 1.8\times10^8$ sequences); (ii) a cosine-similarity filter against q at a threshold ≥ 0.95, which retains $\sim 2.5\times10^5$ sequences whose own composition resembles the target composition; and (iii) a sequence-identity filter at $< 70\%$ pairwise identity to remove near-duplicates and avoid bias toward over-represented families, leaving $\sim 1.0\times10^5$ sequences. The same pipeline is applied to two target compositions (q_{A} and q_{B}; §4.1), yielding two distinct FT datasets. The resulting FT dataset contains natural proteins that are both biologically plausible and compositionally close to q, providing a training signal that biases the base model toward the target composition before any reward-guided RL optimization.

The base model is ProtGPT2 (Ferruz et al., 2022). We perform causal-LM domain-adaptive FT on the FT dataset and take the resulting checkpoint (FT prior) as the frozen reference policy π_{ref} for all subsequent RL runs (hyperparameters in App. A).

3.3. Second Stage: Reward-Weighted RL

Given the frozen reference policy π_{ref}, we optimize a trainable policy π_θ with a reward-weighted log-ratio objective. For a batch of sampled sequences $\{s_i\}$, we compute scalar rewards $r(s_i)$, standardize and clip them within the batch to obtain $\tilde{r}_i$, and convert them to weights $w_i = \text{softmax}(\tilde{r}_i)$, so that sequences with higher rewards contribute more to the update. The objective is

$$\mathcal{L}(\theta) = -\sum_i w_i \, \eta\big(\log \pi_\theta(s_i) - \log \pi_{\text{ref}}(s_i)\big) + \lambda_{\text{KL}} \, \text{KL}\big(\pi_\theta \,\|\, \pi_{\text{ref}}\big). \quad (1)$$

where η is a log-ratio scale factor. This is related to reward-weighted and advantage-weighted regression (Peters & Schaal, 2007; Peng et al., 2019), with π_{ref} playing a role analogous to that in DPO (Rafailov et al., 2023). The Kullback-Leibler term $\text{KL}(\pi_\theta \| \pi_{\text{ref}})$ measures how far π_θ has drifted from π_{ref}. Penalizing it acts as a trust region that prevents π_θ from exploiting reward shortcuts at the expense

of sequence plausibility.

Each iteration t: (i) generate a candidate pool of sequences with stochastic decoding; (ii) generate a parallel *diversity-pulse* pool with hotter decoding (higher temperature and top-p; §H); (iii) score both with $r(s)$; (iv) filter sequences by length, de-duplication, and pairwise identity below 0.85 using MMseqs2 (Steinegger & Söding, 2017); (v) re-inject a fraction of the diversity-pulse pool into the candidate pool, with the re-injection fraction increasing across training to counteract late-iteration mode collapse; (vi) construct reward-weighted batches and update π_θ under Eq. 1.

The reward combines a composition term and a length term,

$$r(s) = w_c \,\mathrm{Comp}(s, q) + w_l \,\mathrm{Len}(s), \tag{2}$$

with static weights $(w_c, w_l) = (0.97, 0.03)$ across all runs. The composition term is the primary signal, while the length term prevents collapse to extremes, motivating the asymmetric weighting. These weights were chosen as a working default after a small exploration. No controlled sweep was performed, so we make no claim of optimality. $\mathrm{Len}(s)$ is a piecewise-linear shaping term that is non-zero on $[70, 400]$ AA with a peak plateau on $[110, 250]$ (App. B). This range used at training does not have to coincide with the FT dataset length filter $[100, 500]$ AA. The length term and weights are identical across all reward variants analyzed in this work, and only $\mathrm{Comp}(s, q)$ differs.

We propose a *differentiated* composition term as our primary $\mathrm{Comp}(s, q)$ term and compare it against one ablation and two baselines at matched compute to assess the contribution of its design choices. The *differentiated* composition term uses an asymmetric per-residue kernel (penalizing deficits in essential AA, those that cannot be synthesized by the organism and must be obtained through diet, more harshly than excesses), two residue pools whose members are treated as biochemically interchangeable (a sulfur pool (Met/Cys) and an aromatic-precursor pool (Phe/Tyr)), and a zero-target residue amplifier (a term that up-weights residues whose target frequency is zero). As an ablation, we evaluate a *symmetric* variant of the *differentiated* term in which the asymmetric per-residue weighting is replaced by uniform absolute deviations. As baselines, we additionally evaluate a *cosine* similarity term and a *global-deviation* L_1 composition term, both of which lack the per-residue structure of the *differentiated* composition term. Full formulas and hyperparameters App. B.

Each composition term has a sharpness coefficient β inside an outer $\exp(-\beta\,[\cdot])$ that maps the per-sequence composition error to a bounded reward, with β ramping during training to progressively sharpen the reward signal. A fixed reference value β_{ref} is used at evaluation, applied to the *differentiated* term so that all variants are compared on a common scoring function (full per-variant ramps in App. B; loop hyperparameters in App. A).

4. Experimental setup

4.1. Target AA Compositions

q_{A} (primary): an experimentally refined poultry-feed AA composition provided by a project partner (details withheld for proprietary reasons). It includes two low-frequency residues (one at exactly zero). We anonymized the residue labels as $\mathrm{aa}_1, \ldots, \mathrm{aa}_{20}$ throughout (sorted by descending target frequency).

q_{B} (published reference): a poultry-feed AA composition derived from Cambra-López et al. (2022), with twenty non-zero target frequencies. Used as a generalization probe.

4.2. Runs

We run all four composition term variants of §3.3 on each target at matched compute and shared seeds. On q_{A}, each variant is run on 30 seeds; on q_{B}, each variant is run on 10 seeds. All other recipe knobs (number of iterations, optimizer, KL schedule, β ramp, candidates per iteration, and the length and identity filters) are identical across variants and across targets. This gives $4{\times}30 + 4{\times}10 = 160$ RL runs in total, plus base ProtGPT2 (no further training) and a domain-adaptive FT checkpoint per target composition, evaluated as four baseline rows in Table 1.

4.3. Metrics

We evaluate each run at the iteration whose candidate pool has the highest mean composition score, recomputed at fixed $\beta_{\text{ref}}{=}20$ through the *differentiated* composition term (App. B), so that selection is independent of each run's β schedule.

We report the *Jensen-Shannon divergence* $\mathrm{JSD}(p, q)$ between the empirical and target composition, the *tolerance count* $N_{\pm 30}(p, q)$, defined as the number of residues whose empirical frequency falls within a 30% relative window of their target frequency, and the composition score at β_{ref}. All three are computed per sequence and then averaged over the candidate pool of the selected iteration. We therefore report pool means for each run, and, for the multi-seed "RL" rows, mean $\pm$ std across seeds of these per-run pool means. Pool sizes range from $\sim$700 to $\sim$2000 sequences depending on iteration, after length and identity filtering. We additionally report three indicators measuring the fraction of sequences in the candidate pool that satisfy a given sequence constraint: *essential-residue coverage* (fraction of sequences in which the ten essential AA each reach at least half of their target frequency), *pool tolerance* (both interchangeable pools within $\pm 30\%$ of their pool target), and *low-target compliance* (at most two occurrences of any residue whose target

Table 1. Pipeline comparison on the primary target q_A and the published reference target q_B. All values are pool means at the run's best iteration (selected by mean composition score at β_{ref}=20). The "RL (n=·)" rows aggregate all seeds of the *differentiated* composition term on the corresponding target (mean ± std across seeds for JSD and Comp.), and the "RL (best run)" rows report the single seed with the lowest pool-mean JSD on that target.

STAGE	TARGET	COMP.↑	JSD↓	$N_{\pm 30}$ ↑
BASE PROTGPT2	q_A	0.007	0.247	3.70
+ DOMAIN-ADAPT. FT	q_A	0.126	0.059	9.17
+ RL (n=30)	q_A	0.397±0.210	0.032±0.020	12.34
+ RL (BEST RUN)	q_A	**0.822**	**0.0044**	**17.40**
BASE PROTGPT2	q_B	0.054	0.202	3.71
+ DOMAIN-ADAPT. FT	q_B	0.207	0.051	8.25
+ RL (n=10)	q_B	0.566±0.123	0.021±0.008	12.21
+ RL (BEST RUN)	q_B	**0.830**	**0.0008**	**20.00**

frequency is at or near zero). The low-target compliance metric is reported only on q_A because q_B has no zero or near-zero frequency residues.

Sequence quality (§5.4) is summarized by NetSolP-predicted solubility (Thumuluri et al., 2022), base ProtGPT2 log-PPL, ESM-2 pPPL, mean length, and intra-pool similarity. Per-variant means use 95% percentile bootstrap CIs over the seed count.

5. Results

5.1. Domain Adaptation Moves the Base Model Toward the Target AA Composition

We first examine the effect of domain-adaptive FT before any reward-weighted RL is applied. Independent FT runs on the respective target-specific FT datasets produce a large shift toward their target compositions (Table 1).

On q_A, held-out perplexity on the FT eval split (a held-out subset of the composition-filtered UniProt sequences, §3.2, App. A) drops by an order of magnitude (5570.6 → 507.3). Concurrently, JSD drops from 0.247 to 0.059 bits, the tolerance count $N_{\pm 30}$ rises from 3.70 to 9.17, and the composition score rises from 0.007 to 0.126 (Table 1). Because this eval split is drawn from the composition-filtered UniProt subset that conditions the FT (§3.2), the held-out perplexity primarily reflects domain adaptation rather than generic protein-likeness. For an independent protein-likeness proxy, we use ESM-2 (150M) pseudo-perplexity (pPPL) (Lin et al., 2023), which we report in full in §5.4 (Table 3) and which shows that FT is slightly worse than base ProtGPT2. On q_B the picture is qualitatively the same (Table 1): JSD drops from 0.202 to 0.051, $N_{\pm 30}$ rises from 3.71 to 8.25, and the composition score from 0.054 to 0.207.

One qualification follows. The average composition is closer to q, but sequence constraint indicators remain weak after FT (essential-residue coverage 0.212, pool tolerance 0.276, low-target compliance 0.098 on q_A, with analogous gaps on q_B; Table 2). This is expected, as the FT dataset is built from natural proteins selected by AA composition cosine similarity (§3.2), and the unconstrained likelihood objective cannot enforce sharp residue constraints that diverge from typical natural composition. The reward-weighted RL stage closes exactly this gap (§5.2). Whether the FT stage is necessary for this gap closure is examined as an ablation in §5.5.2.

5.2. Reward-Weighted RL Closes Sequence Constraints Gap

Having established that domain-adaptive FT improves average composition but leaves sequence constraints largely unsatisfied, we now show that adding the RL stage on top of FT closes the remaining gap. Using the *differentiated* term on q_A results in a mean JSD of 0.032 ± 0.020 across seeds ($n = 30$), compared to 0.059 for FT alone, that is, a roughly 2× reduction (Table 1). The seed distribution is right-skewed: the best run reaches JSD $= 0.0044$ a composition score of 0.822, $N_{\pm 30} = 17.4$ (Table 1), and all sequence constraints indicators (low-target compliance, essential-residue coverage, pool tolerance) are saturated at 1.00, while FT and base ProtGPT2 leave most sequences out of compliance (Table 2). The mean-across-seeds row is intermediate, with typical seeds satisfying roughly half of each sequence constraint indicator. Five of the 30 seeds reach JSD < 0.01. The best-seed numbers are operationally relevant because the best runs are the ones whose candidates will be selected for synthesis. The full top-5 by JSD and by composition score is reported in App. C. On q_B the *differentiated* term reaches a mean JSD $= 0.021 \pm 0.008$ (n=10). The best run achieves JSD $= 0.0008$, a composition score of 0.830, $N_{\pm 30} = 20.0$ and all applicable sequence constraints indicators saturated at 1.00 (Table 1). The q_B best run is closer to its target than the q_A best run, but q_B is a strictly easier task: it has no zero-frequency residues, so the zero-target residue amplifier that dominates the *differentiated* composition term on q_A does not apply. The two "best" numbers are therefore not directly comparable.

Table 2. Pipeline comparison of sequence constraint indicators on the primary target q_A and the published reference target q_B fraction of sequences in the candidate pool satisfying each indicator). The "RL (n=·)" rows aggregate all seeds of the corresponding target, and the "RL (best run)" rows report the single seed with the lowest pool-mean JSD on that target.

STAGE	TARGET	ESSENT.↑	POOLS↑	LOW-TGT.↑
BASE PROTGPT2	q_A	0.004	0.060	0.226
+ DOMAIN-ADAPT. FT	q_A	0.212	0.276	0.098
+ RL (MEAN, n=30)	q_A	0.470	0.553	0.602
+ RL (BEST RUN)	q_A	**1.000**	**1.000**	**1.000**
BASE PROTGPT2	q_B	0.030	0.032	–
+ DOMAIN-ADAPT. FT	q_B	0.160	0.224	–
+ RL (MEAN, n=10)	q_B	0.724	0.672	–
+ RL (BEST RUN)	q_B	**1.000**	**1.000**	–

The fixed-β_{ref} composition score rises monotonically across iterations (Fig. 1, bottom) for the median of the top six seeds of every composition term variant. Whether the RL gains over FT can be matched by best-of-N selection from the FT prior alone is examined as an ablation in §5.5.3.

5.3. Per-Residue Calibration

Figure 2 shows the per-residue composition calibration of the best RL run relative to the FT prior and base ProtGPT2, revealing where each training stage contributes most. The top panel shows observed vs. target counts for every residue (anonymized as $\text{aa}_1, \ldots, \text{aa}_{20}$, sorted by descending target frequency). The highest-frequency target residues are already well-matched after FT, with RL providing only modest additional sharpening. Intermediate-frequency residues (positions ∼5-15) are where RL contributes the most, often pulling absolute frequency residuals $|p_i - q_i|$ from ~ 0.02 after FT down to < 0.005. Residues whose target frequency deviates substantially from the natural-protein average, including the lowest-target residues, are barely moved by FT and are brought close to their target by the RL stage.

5.4. Composition Alignment Preserves Sequence Quality

We next ask whether target composition alignment comes at the cost of sequence plausibility. Table 4 summarizes three sequence-level quality indicators across all RL runs on both target compositions: predicted solubility (NetSolP), which is relevant because high solubility is a prerequisite for protein digestibility, and two protein-likeness scores, base ProtGPT2 log-PPL (autoregressive, under the original pretrained model) and ESM-2 pPPL (masked-LM, from a different model family).

Mean NetSolP solubility is 0.582 ± 0.030 on q_A and 0.628 ± 0.053 on q_B, both above the NetSolP 0.5 decision threshold of the original NetSolP study, with individual sequences exceeding 0.9 on both targets. NetSolP is itself a predictor, so these scores are an *in-silico* sanity check, not a wet-lab claim. The four composition terms are within 0.04 of each other in mean solubility on q_A (0.560-0.598; Table 7) and within 0.10 on q_B (0.589-0.693; per-variant q_B values in Table 10, App. D).

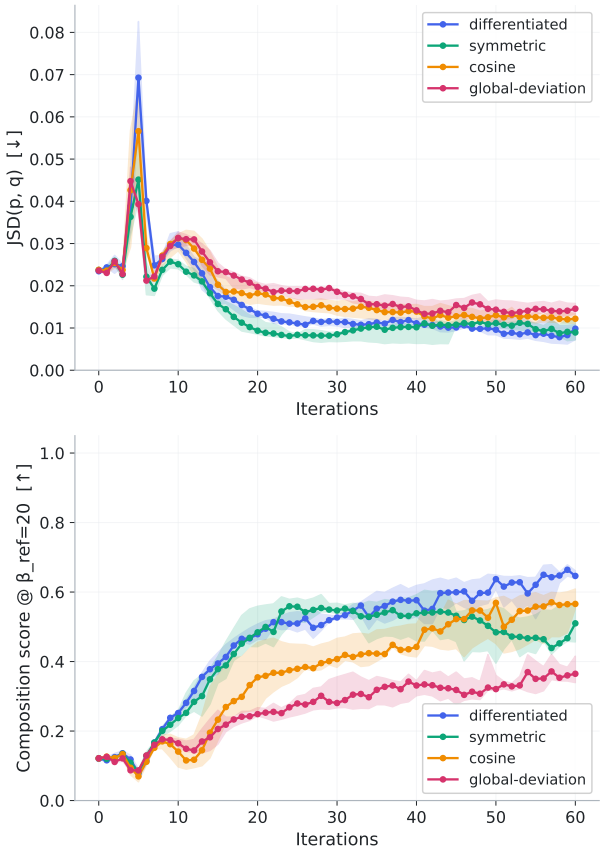

Figure 1. Training dynamics on q_A. Each curve is the median across the top six seeds per composition term variant, with the IQR shaded. (Top) Mean JSD against the target across iterations. (Bottom) Composition score evaluated at a fixed reference temperature β_{ref}=20 using the *differentiated* composition term for all four variants.

Table 3. ESM-2 (150M) pseudo-perplexity (pPPL) as a protein-likeliness measure across pipeline stages, composition term variants, and target compositions; mean±std over up to n=200 unique sequences per condition; lower is better. RL rows use the best iteration of one representative seed per variant. ESM-2 pPPL is length-dependent, so we also report mean sequence length $\bar{L}$.

	q_A		q_B	
CONDITION	PPPL ↓	$\bar{L}$	PPPL↓	$\bar{L}$
BASE PROTGPT2	4.48±1.19	387.9	4.48±1.19	387.9
+ DOMAIN FT	5.74±2.43	375.0	6.04±1.81	262.6
RL COSINE	10.06±0.42	393.0	6.72±0.33	478.9
RL DIFFERENTIATED	9.41±1.88	326.2	12.28±1.19	397.0
RL GLOBAL-DEV.	15.79±1.67	480.8	7.25±0.22	447.6
RL SYMMETRIC	16.65±0.64	495.0	10.98±1.39	470.6

Table 4. Sequence-quality summary, mean±std across all RL runs per target. NetSolP solubility is a predicted score in $[0, 1]$ (higher is more soluble). *base log-PPL* is the mean log-perplexity of the base ProtGPT2 policy. ESM-2 (150M) pseudo-perplexity (pPPL) is computed on up to n=200 unique sequences per RL composition variant; the entry summarizes mean±std across the four primary variants. The *iid floor* column reports the same three metrics on n=100 random sequences sampled iid from q_A at L=400, as a composition-only baseline.

METRIC	q_A (n=120)	q_B (n=40)	IID FLOOR (q_A)
NETSOLP SOLUBILITY ↑	0.582 ± 0.030	0.628 ± 0.053	0.492 ± 0.055
base log-PPL ↓	9.08 ± 1.13	5.97 ± 0.96	9.22 ± 0.24
ESM-2 PPPL ↓	12.98 ± 3.77	9.31 ± 2.74	18.07 ± 0.57

ESM-2 pPPL increases monotonically from base ProtGPT2 (mean 4.48 on both targets, as expected for an unconditional baseline) to the domain-adapted FT prior (5.74-6.04) to the RL policies (6.72-16.65, target- and composition-dependent) (Table 3). For reference, ESM-2 pPPL on natural sequences typically falls in the single-digit range (Lin et al., 2023), so the RL policies sit one tier above, indicating a measurable but moderate plausibility cost. This cost is not uniform across composition terms: the *symmetric* composition term in particular almost triples ESM-2 pPPL relative to the FT prior on q_A, which we interpret as the cost of steering the policy toward a non-natural composition without the biological weighting of the *differentiated* term. Notably, the *differentiated* term (mean ESM-2 pPPL 9.41 on q_A) incurs the smallest increase among the composition variants, suggesting that its biological weighting partially mitigates this plausibility cost. To stress-test the plausibility of the most composition-aligned sequences, we additionally score the top-30 sequences from each policy ranked by composition score (i.e. the most aggressive on-target sequences each policy generates) (Table 15 in App. F). Even on this worst-case slice, no policy collapses to a degenerate high-ESM-2 pPPL regime, and the *differentiated* variant retains its small-fluency-penalty advantage over the other smooth-max variant on q_A.

As a composition-only baseline, we score n=100 random sequences whose residues are sampled independently and identically distributed (iid) from q_A at L=400 on all three metrics (Table 4). Any policy scoring close to this iid floor would be statistically indistinguishable from a random composition-matched sequence in terms of solubility and protein-likeness. The iid floor on ESM-2 pPPL is 18.07 (approximately length-independent in the relevant range; App. G), the iid solubility floor is 0.492 ± 0.055, and the iid *base log-PPL* floor is 9.22 ± 0.24. Three observations follow. First, predicted solubility on RL policies (0.582 on q_A) sits about 1.6 floor-std above the iid floor (0.492). The gap is modest but consistent in sign, so policy solubility is not entirely a composition artefact. Second, the RL aggregate on q_A under *base log-PPL* (9.08 ± 1.13) is just below the iid floor (9.22 ± 0.24); per-variant, *differentiated* (8.82) and *cosine* (7.35) sit clearly below the floor, while *global-deviation* (9.98) and *symmetric* (10.18) cross above it. Third, all RL policies remain measurably below the ESM-2 pPPL floor on q_A, so even the worst variant has not collapsed to "random AA" plausibility. The ESM-2 pPPL gap is, however, variant-dependent: the *differentiated* term (9.41) sits roughly midway between the FT prior and the iid floor (about 1.6× the FT pPPL and about half the floor), whereas the *symmetric* term (16.65) and *global-deviation* (15.79) sit close to the floor (within ∼ 8-13% of 18.07) (Table 3). This quantifies the qualitative claim above, that L_p- and *symmetric* composition terms approach the regime of pure-composition random sequences under ESM-2, while the biologically weighted *differentiated* term retains a clear plausibility margin over that floor.

As a diagnostic on the KL anchor of Eq. 1, we report the *reference log-PPL* (policy mean log-perplexity under its own FT prior), a one-sided proxy for the KL term (differs from $\mathrm{KL}(\pi_\theta \| \pi_{\text{ref}})$ by the policy entropy). The RL aggre-

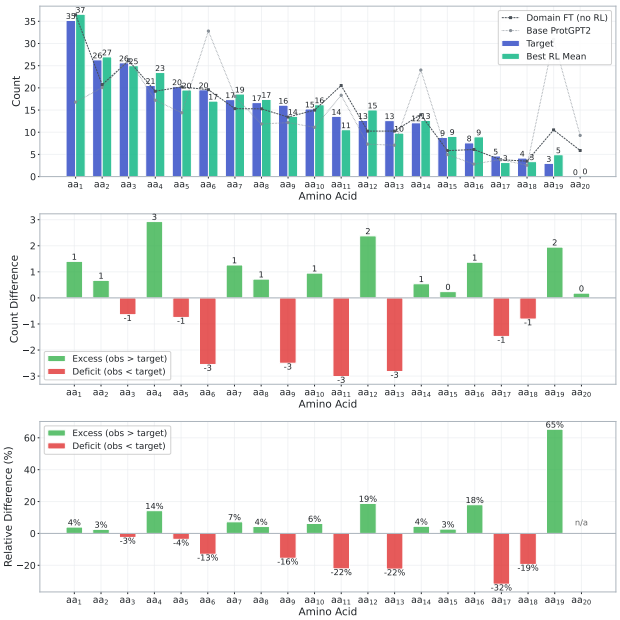

Figure 2. Per-residue calibration on q_A (residues anonymized as $aa_1, \ldots, aa_{20}$, sorted by descending q). Counts are pool-mean frequencies p_i rescaled to a common reference length L=292 AA (the rounded mean sequence length of the best-RL pool). (Top) Target counts (blue) vs. best-RL counts (turquoise); Domain-adaptive FT and base ProtGPT2 are overlaid as dashed and dotted lines. (Middle) Signed count residual (observed vs. target) for the best-RL run. (Bottom) Signed relative residual $(p - q)/q \times 100\%$ for the best-RL run (length-independent), the q=0 residue is marked n/a.

gate is 8.30 ± 1.35 on q_A and 4.83 ± 0.72 on q_B, at or below the iid references under the corresponding FT priors (8.52 ± 0.21 and 8.90 ± 0.17; App. D), so no runaway drift on either target. The larger margin on q_B is consistent with it being the easier target (no zero-target residues). The overlap with the iid reference is expected rather than informative about plausibility, since the FT prior was trained on composition-filtered data, that the same iid-q_A sequences score 9.22 ± 0.24 under base ProtGPT2 (Table 4) vs. 8.52 under the FT model, confirming the FT prior has internalized the composition and is a meaningful KL anchor. Target-independent plausibility is read from base log-PPL and ESM-2 pPPL.

Finally, we verify that the RL stage preserves sequence diversity. Table 5 reports two complementary intra-pool similarity metrics aggregated over all RL runs (lower=more diverse). First, *Aln. id.*, the mean MMseqs2 identity over alignment-survivors at coverage ≥ 0.6, which is alignment-conditional and therefore upward-biased toward the most-similar pair tail, and second, 4-*mer Jacc.*, the mean pairwise Jaccard of 4-mer sets on a fixed 200-sequence subsample, which is defined for every pair and so reads the whole pool. Two anchors interpret the magnitudes (Table 5, q_A / q_B stacked per cell). Pre-RL pools of 500 sequences from the FT prior reach 4-*mer Jacc.* ~ 0.01 on both targets, which we read as the unconditioned-generation floor at $L \sim 300$ AAs. The RL aggregate (0.246 on q_A, 0.483 on q_B) is measurably more concentrated than this floor, as expected for a policy optimized toward a fixed target composition, but stays well clear of the collapsed RL-only seed of §5.5.2 (0.74 4-*mer Jacc.*, 0.97 *Aln. id.*). Per-variant breakdown is in App. D (Table 11), and the diversity-pulse mechanism that maintains the margin during training is detailed in App. H (Fig. 5). Mean sequence length stays inside the target interval $[70, 400]$ AA.

5.5. Ablation Studies

We report three controlled ablations: the contribution of the *differentiated* composition term's design choices relative to ablated and baseline alternatives (*symmetric*, *cosine*, and *global-deviation*) on q_A (§5.5.1), the necessity of the FT stage on q_A (§5.5.2), and best-of-N selection from the FT prior on both targets (§5.5.3).

5.5.1. Composition term formulation

We evaluate the design choices of the *differentiated* composition term against one ablated (*symmetric*) and two baseline (*cosine* and *global-deviation*) composition terms on q_A at matched compute and shared seeds (Table 7). Pairwise Wilcoxon signed-rank tests on the n=30 paired seeds give a clear ranking: every smooth-max (*differentiated*, *symmetric*) or *cosine* term beats the L_1 term (*global-deviation*) on composition score (all surviving Bonferroni correction at $\alpha = 0.05$), and on JSD for *symmetric* over *global-deviation*. The remaining three variants (differentiated, symmetric, cosine) are not pairwise separable on JSD, and only *differentiated* vs. *cosine* separates uncorrected on the composition score. Full pairwise statistics are in App. D.1; per-variant seed distributions in App. D.

On this target, the statistical comparison confirms that any smooth-max- or cosine-based composition term substantially outperforms the L_1 baseline, while the three upper-cluster variants are not reliably separable on aggregate metrics alone. However, the *differentiated* term is our default choice for two reasons that go beyond aggregate composition metrics alone. First, it directly encodes biological constraints, penalizing deficits in essential AA more harshly than excesses, handling interchangeable AA pools as group-level terms, and amplifying deviations on zero-target AA, which are design choices motivated by the nutritional objective rather than arbitrary hyperparameters. Second, it incurs the smallest pPPL penalty among the smooth-max variants (§5.4, Table 3), suggesting that biological weight-

Table 5. Intra-pool sequence similarity (lower=more diverse), aggregated over all RL runs per target (n=120 on q_A, n=40 on q_B; mean±std), and on two pre-RL anchor pools of 500 sequences each (base ProtGPT2 and the domain-adaptive FT prior, q_A / q_B stacked per cell). *Aln. id.* is mean MMseqs2 identity over alignment-survivors at coverage ≥ 0.6 (alignment-conditional, upward-biased; for the pre-RL anchors, the average is over only 28-610 surviving pairs because most pairs do not align at this threshold, so cross-row comparison should rely on 4-mer Jacc.). *4-mer Jacc.* is the mean pairwise Jaccard of 4-mer sets on a fixed 200-sequence subsample (defined for every pair, hence directly comparable across rows).

	RL POLICIES		PRE-RL ANCHORS (n=500)	
METRIC	q_A	q_B	FT PRIOR	BASE PROTGPT2
ALN. ID.↓	0.832 ± 0.121	0.842 ± 0.212	0.44 / 0.54	0.36 / 0.36
4-MER JACC.↓	0.246 ± 0.197	0.483 ± 0.345	0.007 / 0.010	0.011 / 0.010

Table 6. FT→RL vs RL-only on q_A, mean across n=3 paired seeds (same recipe and budget). Comp. is the composition score at β_{ref}=20; Ess. and Pools are sequence constraint indicators (§4).

SETTING	COMP.↑	JSD↓	ESS.↑	POOLS↑	$N_{\pm 30}$ ↑
RL ONLY	0.420	0.032	0.24	0.46	11.9
FT→RL	**0.574**	**0.018**	**0.64**	**0.87**	**13.3**

ing partially mitigates the plausibility cost of composition alignment.

5.5.2. DOMAIN-ADAPTIVE FT VS RL-ONLY

We assess the necessity of the domain-adaptive FT stage by running RL directly on base ProtGPT2 (RL-only), on n=3 seeds shared with the FT → RL runs (Table 6). Mean paired differences (FT→RL minus RL-only across the three seeds) favor FT → RL on every metric: composition +0.15, essential-residue coverage +0.40, pool tolerance +0.41, $N_{\pm 30}$+1.4, and JSD −0.014 (lower is better). Given only n=3 paired seeds, we report direction and magnitude without formal significance testing. The FT→RL values of this subsection are restricted to the three seeds matched to the RL-only version, and it is therefore not directly comparable to the n=30 mean reported in §5.2. Per-seed inspection reveals a characteristic reward-hacking failure mode that emerges when RL is run without the FT prior, that is, one of the three RL-only seeds collapsed onto a very narrow AA palette, satisfying low-target compliance perfectly (no zero-target AA ever exceeded its cap), but in doing so, it dropped essential-residue coverage to zero. Despite this, it still scored a high composition score of 0.70 because the average frequency vector remained close to q_A.

The collapse is also visible as a diversity signature. The failed RL-only seed reaches a mean intra-pool pairwise identity of 0.97 (near saturation; *Aln. id.*) and *4-mer Jaccard* of 0.74, versus 0.84/0.89 *Aln. id.* and 0.21/0.31 *4-mer Jacc*. for the surviving RL-only seeds, and 0.72-0.82 *Aln. id.* (0.52-0.57 4-mer Jacc.) for the matched FT→RL seeds, suggesting on both metrics that, without the FT prior, the policy contracts onto a narrow palette of nearly-identical sequences. This single seed inflates the RL-only mean composition score in Table 6, but the other two RL-only seeds average composition ≈ 0.29. FT prevents this failure mode in our sweep, though we note this is based on a single RL-only failure case and a limited $n = 3$ paired comparison. We read the mechanism as a reward-landscape effect rather than a property of the FT prior in isolation. The FT prior is itself narrower than base ProtGPT2 (its training set is composition-filtered), but it sits much closer to q (JSD 0.059 vs. 0.247), so reward-weighted updates are spread across many moderately-rewarding directions and the KL penalty against π_{ref} (§3.3) actively pulls the policy back toward this broad composition-conditioned region. From base ProtGPT2, the same KL budget cannot reach a comparably-rewarded region without latching onto a few high-reward modes, which is the collapse signature observed here.

5.5.3. BEST-OF-N FROM THE FT PRIOR

To test whether reward-weighted RL updates merely reproduce best-of-N selection from the FT prior, we draw 500 sequences from the FT-only model and from base ProtGPT2, and rank them by the same fixed-β_{ref} composition score used throughout. On q_A, the best-of-500 from FT-only reaches JSD = 0.018 (composition score= 0.529), while the best FT→RL run reaches JSD = 0.0044 (composition score = 0.822). This is a 4× reduction in JSD that no amount of selection from the FT prior recovers in 500 draws. On q_B, best-of-500 from FT-only reaches JSD = 0.0111 vs. 0.0008 for the best FT→RL run, a 14× gap. Base ProtGPT2 is two orders of magnitude further from either target even at the best-of-500 tail (Table 14 in App. E). The RL stage, therefore, moves the policy into a region of sequence space that the FT prior does not reach by oversampling alone.

6. Discussion

This work demonstrates that steering a PLM toward an explicit distributional target AA composition benefits from two training stages addressing different aspects of the prob-

Table 7. Reward-formulation comparison on q_{A} (mean±std across $n = 30$ seeds, matched compute). Solub. is mean NetSolP predicted solubility (higher=more soluble). Best per column in bold.

REWARD	COMP.↑	JSD↓	$N_{\pm 30}$↑	L_1↓	SOLUB.↑
DIFFERENTIATED	**0.397±0.210**	0.032±0.020	12.34±2.57	0.259±0.091	**0.598±0.031**
SYMMETRIC	0.305±0.167	**0.028±0.012**	**12.89±2.02**	**0.247±0.069**	0.582±0.027
COSINE	0.280±0.188	0.033±0.013	12.51±2.53	0.258±0.091	0.586±0.032
GLOBAL-DEVIATION	0.140±0.028	0.041±0.003	11.44±0.49	0.312±0.013	0.560±0.010

lem. Domain-adaptive FT shifts the base model toward an average composition close to q, but cannot enforce the sequence constraints needed to select individual candidates for synthesis. This per-residue gap is largest on residues whose target frequency is far from the natural protein average. The RL stage closes this gap, taking the FT prior from weak sequence constraint satisfaction to near-saturation under the best aligned policy, a transition that oversampling from the FT prior alone cannot reproduce. Removing the FT stage degrades every metric on every paired seed, and one of the three RL-only seeds exhibits a discrete failure mode in which the policy reward-hacks by collapsing onto a narrow residue palette that satisfies the target composition on average while suppressing essential residue coverage. This failure mode is absent from all FT→RL runs in our sweep, supporting the view that the domain-adaptive FT provides a stable initialization that the reward signal alone cannot guarantee.

A key contribution of this work is the *differentiated* composition term, a novel reward formulation tailored to the specific task of designing digestible proteins with a nutrition-oriented AA composition. To evaluate its design choices, we compare it against an ablated variant (*symmetric*) and two existing baselines (*cosine*, *global-deviation*) in a controlled seed-matched experiment. The results show that the *global-deviation* baseline is clearly insufficient while the remaining three terms form a broadly comparable upper cluster. A weak signal suggests that smooth-max variants may outperform the *cosine* term, though this does not survive Bonferroni correction at the seed counts used here. Within the upper cluster, the *differentiated* term attains the highest mean composition score on the q_A target composition while its best run saturates all sequence constraints (essential-residue coverage, pool tolerance, low-target compliance), making it the most faithful to the nutritional design objective. Additionally, *differentiated* carries the smallest ESM-2 pPPL penalty, indicating that the composition gains are not bought at a disproportionate plausibility cost, and suggesting that biologically grounded reward design may inherently produce more natural-looking sequences than uniform alternatives.

Several limitations should be noted. First, the pipeline does not enforce sequence plausibility beyond the regularization provided by the KL penalty against the FT reference model, and we do not measure structure or function directly. Second, seed counts are modest (30 on q_{A} and 10 on q_{B}), so we report bootstrap CIs rather than asymptotic null tests. Third, biological follow-up, including digestibility-related proxies, is necessary to determine whether improved target composition alignment also yields biologically plausible candidates. Addressing these limitations represents the natural next step for this line of work.

To the best of our knowledge, this is the first work to steer a PLM toward an explicit target AA composition as a primary design objective. While matching a frequency vector over 20 AA may appear to be a simple distributional matching problem, it is better understood as a multi-objective sequence design task: generated sequences must simultaneously satisfy a range length, essential-residue coverage, interchangeable pool balance, and zero-target compliance, without drifting away from the manifold of plausible protein sequences. That the best aligned policies satisfy all these constraints simultaneously while retaining reasonable ESM-2 plausibility demonstrates that the proposed two-stage pipeline, domain-adaptive FT followed by reward-weighted RL with a biologically grounded composition term, is a viable approach to this class of constrained generative objectives.

7. Impact Statement

This work develops a method for steering PLMs toward an explicit target AA composition. The motivating application is feed-protein design, but the pipeline is broadly applicable to any composition-constrained protein generation task. All generated sequences are computational proposals that require further validation before wet-lab synthesis or downstream deployment. Depending on the application, this includes structure prediction, disorder analysis, digestibility assays for feed proteins, or activity assays for functional proteins. Potential positive impacts include better-aligned dietary proteins for animal feed and a reduced environmental cost of feed production.

References

Cambra-López, M., Marín-García, P. J., Lledó, C., Cerisuelo, A., and Pascual, J. J. Biomarkers and de novo protein design can improve precise amino acid

nutrition in broilers. *Animals*, 12(7):935, 2022. doi: 10.3390/ani12070935.

Cao, H., Torres, M. D. T., Zhang, J., Gao, Z., Wu, F., Gu, C., Leskovec, J., Choi, Y., de la Fuente-Nunez, C., Chen, G., and Heng, P.-A. A deep reinforcement learning platform for antibiotic discovery. *bioRxiv*, 2025. doi: 10.1101/2025.09.23.678086. Preprint.

Emmert, J. L. and Baker, D. H. Use of the ideal protein concept for precision formulation of amino acid levels in broiler diets. *Journal of Applied Poultry Research*, 6(4): 462–470, 1997. doi: 10.1093/japr/6.4.462.

Ferruz, N., Schmidt, S., and Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications*, 13:4348, 2022.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pp. 8342–8360. Association for Computational Linguistics, 2020.

Hesslow, D., Zanichelli, N., Notin, P., Poli, I., and Marks, D. RITA: A study on scaling up generative protein sequence models. *arXiv preprint arXiv:2205.05789*, 2022.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Madani, A., Krause, B., Greene, E. R., Subramanian, S., Mohr, B. P., Holton, J. M., Olmos, J. L., Xiong, C., Sun, Z. Z., Socher, R., Fraser, J. S., and Naik, N. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41:1099–1106, 2023.

Munsamy, G., Lindner, S., Lorenz, P., and Ferruz, N. Conditional language models enable the efficient design of proficient enzymes. *bioRxiv*, 2024. doi: 10.1101/2024.05.03.592223. Preprint.

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

Peng, X. B., Kumar, A., Zhang, G., and Levine, S. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.

Peters, J. and Schaal, S. Reinforcement learning by reward-weighted regression for operational space control. In *Proceedings of the 24th International Conference on Machine Learning (ICML 2007)*, pp. 745–750. ACM, 2007.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023.

Ravindran, V. Feed enzymes: The science, practice, and metabolic realities. *Journal of Applied Poultry Research*, 22(3):628–636, 2013.

Steinegger, M. and Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, 2017.

Stocco, F., Artigues-Lleixà, M., Hunklinger, A., Widatalla, T., Güell, M., and Ferruz, N. Guiding generative protein language models with reinforcement learning. *arXiv preprint arXiv:2412.12979*, 2024. Preprint.

Subramanian, J., Sujit, S., Irtisam, N., Sain, U., Islam, R., Nowrouzezahrai, D., and Ebrahimi Kahou, S. Reinforcement learning for sequence design leveraging protein language models. *arXiv preprint arXiv:2407.03154*, 2024. Preprint.

The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1): D523–D531, 2023. doi: 10.1093/nar/gkac1052.

Thumuluri, V., Almagro Armenteros, J. J., Johansen, A. R., Nielsen, H., and Winther, O. NetSolP: Predicting protein solubility in *Escherichia coli* using language models. *Bioinformatics*, 38(4):941–946, 2022. doi: 10.1093/bioinformatics/btab801.

Widatalla, T., Rafailov, R., and Hie, B. Aligning protein generative models with experimental fitness via direct preference optimization. *bioRxiv*, 2024. doi: 10.1101/2024.05.20.595026. Preprint.

A. Reproducibility details

This appendix consolidates the practical settings used to produce every number in the main text. All runs used a single NVIDIA H100 GPU and a single shared conda environment, and software versions and random seeds are recorded in the released run-level dataset.

Code availability. The full training pipeline (composition-conditioned UniProt filtering, FT, and reward-weighted RL stages) and evaluation code will be released as an open-source repository upon acceptance.

Dataset construction. Filters and final dataset sizes (§3.2): UniProtKB/TrEMBL (The UniProt Consortium, 2023) (FASTA from the UniProt Consortium FTP) → length filter $[100, 500]$ aa → per-sequence cosine similarity to $q \geq 0.95$ → MMseqs2 redundancy at $< 70\%$ pairwise identity, yielding $\sim 1.0 \times 10^5$ sequences split 50/50 into train and eval per target.

Stage 1 (FT). ProtGPT2 base; ProtGPT2 tokenizer; block size 512; AdamW with learning rate 5×10^{-5}, weight decay 0.01, 3% linear warm-up; effective batch size 32 blocks; up to 40 epochs with early stopping on eval loss (patience 7, $\min \Delta = 10^{-3}$). The selected checkpoint for q_{A} is at epoch 21; total wall-clock $\sim$13.5 h. The same checkpoint is reused as π_{ref} for every RL run on q_{A} while q_{B} uses its own FT checkpoint trained with identical hyperparameters on the q_{B} filtered dataset.

Stage 2 (RL). Settings shared across all reported runs (Table 8).

Table 8. Stage-2 (RL) hyperparameters. Values are fixed across all 160 runs reported in this paper.

SETTING	VALUE
OUTER ITERATIONS	50
CANDIDATES PER ITERATION	2000
DIVERSITY FILTER	MMSEQS2, < 0.85 PAIRWISE IDENTITY
LENGTH FILTER	TARGET INTERVAL, OUTLIERS DROPPED
REWARD WEIGHTS (w_c, w_l)	$(0.97, 0.03)$ STATIC
SHARPNESS COEFFICIENT β	RAMP 15→36 DURING TRAINING
EVALUATION β_{REF}	20 (FIXED)
POLICY LR	3×10^{-6}
OPTIMIZER	ADAMW, DEFAULT BETAS
KL WEIGHT λ_{KL}	$0.05 \rightarrow 0.15$ OVER FIRST 55%
REFERENCE MODEL	FT CHECKPOINT, FROZEN

Run breakdown. The full $4 \times 30 + 4 \times 10 = 160$ runs of §4.2 share the loop and optimizer settings of Table 8. The only differences between runs are the random seed, the composition variant, and the target.

Evaluation protocol. For each run, we select the iteration that maximizes the fixed β_{ref}=20 composition score on its candidate pool, and report all metrics on that pool. Per-variant confidence intervals are seed-level 95% percentile bootstraps with 10^4 resamples.

B. Composition term formulas

We give explicit formulas for all composition terms used at training and evaluation time. Let $p \in \mathcal{S}_{20}$ be the observed AA frequency vector of a candidate sequence (over the 20 standard AA) and $q \in \mathcal{S}_{20}$ the target frequency vector. Let L denote sequence length and let $\mathcal{E}$ denote the essential-AA index set used in the *differentiated* composition.

Per-residue error (*differentiated*). For each AA i,

$$e_i = \begin{cases} w^{-}(q_i - p_i) & i \in \mathcal{E},\ p_i < q_i, \\ w^{+}(p_i - q_i) & i \in \mathcal{E},\ p_i \geq q_i, \\ w^{\mathrm{ne}}\,|p_i - q_i| & i \notin \mathcal{E}, \end{cases}$$

with $w^{-} = 3.0$, $w^{+} = 0.35$, $w^{\mathrm{ne}} = 1.0$, plus a zero-target amplifier $e_i \leftarrow \alpha_0\, e_i$ whenever $q_i = 0$ and $p_i > 0$ (with $\alpha_0 = 3$). Per-residue errors are clipped to ≤ 2. Group-level errors $w^{\mathrm{grp}}\,|\sum_{i \in G} p_i - t_G|$ for interchangeable AA groups G (with $w^{\mathrm{grp}} = 1.5$; on q_{A} these are the sulfur and aromatic-precursor pools, §4) are appended to the error vector e.

***Differentiated* score.** With $\alpha = 60$, $w_{\text{rms}} = 1$, $w_{\text{sm}} = 0.9$, and β ramped 15→36 across training iterations, and letting $\text{rms}(e) = \sqrt{\frac{1}{n}\sum_i e_i^2}$ and $\text{smax}_\alpha(e) = \frac{1}{\alpha}\log\big(\frac{1}{n}\sum_i e^{\alpha e_i}\big)$,

$$\text{Comp}_{\text{diff}}(p;q) = \exp\Big(-\beta\big[w_{\text{rms}}\,\text{rms}(e) + w_{\text{sm}}\,\text{smax}_\alpha(e)\big]\Big).$$

The second bracket term is a smooth-max (log-sum-exp) over the per-residue errors with sharpness α. Evaluation uses $\beta_{\text{ref}} = 20$.

***Symmetric* score.** Same RMS + smooth-max kernel, but with uniform absolute deviation $e_i = |p_i - q_i|$ (no essential split, no group terms). The remaining hyperparameters $(\alpha, w_{\text{rms}}, w_{\text{sm}}, \alpha_0)$ are kept identical to the *differentiated* variant; only the outer sharpness β is ramped on a different schedule, 30→80 across training iterations, to match the loss scale of the simpler kernel:

$$\text{Comp}_{\text{sym}}(p;q) = \exp\Big(-\beta\big[w_{\text{rms}}\,\text{rms}(e) + w_{\text{sm}}\,\text{smax}_\alpha(e)\big]\Big).$$

differentiated and *symmetric* therefore share the same RMS + smooth-max kernel; the *differentiated* variant adds biologically-motivated essential/non-essential weighting and the interchangeable-group terms, while *symmetric* uses uniform $|p_i - q_i|$ throughout.

***Cosine* score.**

$$\text{Comp}_{\text{cos}}(p;q) = \exp\Big(-\beta\big[1 - \tfrac{p\cdot q}{\|p\|\,\|q\|}\big]\Big),$$

with β ramped 5→25 across training iterations. Scale-invariant: matches direction in $\mathbb{R}^{20}$, not magnitude, which is why it lags on the L_1 metric.

***Global-deviation* score.**

$$\text{Comp}_{\text{gd}}(p;q) = \exp\Big(-\beta\,\tfrac{1}{20}\textstyle\sum_i |p_i - q_i|\Big),$$

with β ramped 20→100 across training iterations. The simplest baseline: mean-L_1 deviation, no smooth-max, no essential split.

Per-variant β ramps. The four ramps above (15→36, 30→80, 5→25, 20→100) were chosen based on the best results that were achieved. Within each variant, β is linearly ramped over a fixed window of iterations (starting at iteration 9, target value reached by iteration 33, then held constant).

Reference-β re-scoring. For all aggregate comparisons, we rescore each candidate pool with the *differentiated* formula at β_{ref}=20, regardless of which variant was used during training. This puts every variant on a single common scoring function and removes confounding from the per-variant β schedule.

Length term. A piecewise-linear shaping term over the support interval $[L_{\min}, L_{\max}]$ with peak plateau $[L_a, L_b]$:

$$R_\ell(L) = \begin{cases} 0 & L \le L_{\min} \text{ or } L \ge L_{\max}, \\ \dfrac{L - L_{\min}}{L_a - L_{\min}} & L_{\min} < L < L_a, \\ 1 & L_a \le L \le L_b, \\ \dfrac{L_{\max} - L}{L_{\max} - L_b} & L_b < L < L_{\max}, \end{cases}$$

with $L_{\min} = 70$, $L_a = 110$, $L_b = 250$, $L_{\max} = 400$ AA. The non-zero support $[70, 400]$ AA is intentionally wider than the dataset length filter $[100, 500]$ AA (§3.2) so that on-target sequences just outside the dataset window still receive partial credit.

C. Top-5 runs

Table 9 lists the top RL runs on q_{A} across all n=120 runs, with rank columns for both JSD (lower better) and fixed-β_{ref} composition score (higher better) so that disagreements between the two rankings are visible.

Table 9. Top-5 RL runs on q_A (across all n=120 runs) ranked by JSD (lower is better).

RANK-JSD	RANK-COMP	REWARD	COMP.↑	JSD↓	$N_{\pm 30}$ ↑	ESSENT.↑	POOLS↑	LOW-TGT.↑
1	1	DIFFERENTIATED	0.822	0.0044	17.37	1.000	1.000	1.000
2	–	SYMMETRIC	0.618	0.0049	19.00	1.000	1.000	1.000
3	3	SYMMETRIC	0.761	0.0049	17.03	1.000	1.000	1.000
4	2	COSINE	0.797	0.0060	17.00	1.000	1.000	1.000
5	5	DIFFERENTIATED	0.669	0.0073	16.75	0.588	0.591	0.977
–	4	DIFFERENTIATED	0.690	0.0086	16.38	1.000	1.000	1.000

D. Composition-variant breakdown

This appendix collects the per-variant figures referenced from §5.5.1, and the analogous composition-variant comparison on q_B (n=10 seeds per variant; Table 10).

Table 10. Reward-formulation comparison on q_B (mean±std across n=10 seeds, matched compute). Same metrics, scoring, and selection rule as Table 7. Best per column in bold.

REWARD	COMP.↑	JSD↓	$N_{\pm 30}$↑	L_1↓	SOLUB.↑
DIFFERENTIATED	0.566±0.117	0.021±0.007	12.21±2.18	0.257±0.065	0.589±0.036
SYMMETRIC	**0.672±0.165**	**0.009±0.008**	**17.11±2.31**	**0.117±0.071**	**0.693±0.019**
COSINE	0.524±0.171	0.016±0.010	14.02±3.12	0.204±0.090	0.628±0.024
GLOBAL-DEVIATION	0.460±0.131	0.023±0.009	12.12±2.86	0.265±0.081	0.603±0.048

Table 11. Per-variant intra-pool sequence-similarity breakdown on q_A (n=30 seeds per variant) and q_B (n=10 seeds per variant), mean±std (lower=more diverse). *Aln. id.* is mean MMseqs2 identity over alignment-survivors at coverage ≥ 0.6. *4-mer Jacc.* is the mean pairwise Jaccard over 4-mer sets on a fixed 200-sequence subsample.

TARGET	REWARD	ALN. ID.↓	4-MER JACC.↓
q_A	DIFFERENTIATED	0.849±0.092	0.334±0.244
q_A	SYMMETRIC	**0.837±0.106**	0.288±0.165
q_A	COSINE	0.848±0.185	0.307±0.271
q_A	GLOBAL-DEVIATION	0.862±0.081	**0.143±0.018**
q_B	DIFFERENTIATED	0.873±0.090	0.411±0.312
q_B	SYMMETRIC	0.855±0.291	0.766±0.268
q_B	COSINE	**0.794±0.280**	0.484±0.348
q_B	GLOBAL-DEVIATION	0.847±0.072	**0.272±0.239**

D.1. Pairwise Wilcoxon: full statistics

Table 13 reports the median paired difference Δ, 95% percentile-bootstrap CI, and Wilcoxon p-value for every pair and every aggregate metric on q_A.

E. Best-of-N vs. RL

Table 14 expands the rejection-sampling argument of §5.2. We score 500 samples from each prior (base ProtGPT2 and FT-only) on each target with the fixed-β_{ref}=20 composition score and report the best and top-10% JSD alongside the best composition score. The best-of-500 from the FT prior is closer to the target than from base ProtGPT2 by a factor of two to three on JSD, confirming that domain-adaptive FT is a real prior shift, but it is still 4× (q_A) to 14× (q_B) further from the target than the best RL run. The gap is largest on the per-sequence composition score, where the FT prior simply does not contain sequences that satisfy the harder per-residue constraints.

Table 12. Log-perplexity of each RL policy under its FT prior (one-sided proxy for $\mathrm{KL}(\pi_\theta \| \pi_{\text{ref}})$, equal to $\mathbb{E}_{\pi_\theta}[-\log \pi_{\text{ref}}]$ up to the policy entropy, §5.4). Each cell is mean±std across seeds of per-run pool means. Values are *not* directly comparable between targets because each target uses a different FT prior. The *Aggregate* row pools all four variants per target. The *iid floor* row reports the same metric on n=100 random sequences sampled iid from the corresponding target at L=400, scored under that target's FT prior.

CONDITION	q_A	q_B
RL COSINE	8.28 ± 0.63 (n=30)	4.96 ± 0.79 (n=10)
RL DIFFERENTIATED	8.10 ± 2.35 (n=30)	4.50 ± 0.81 (n=10)
RL GLOBAL-DEV.	8.36 ± 0.80 (n=30)	4.85 ± 0.43 (n=10)
RL SYMMETRIC	8.47 ± 0.88 (n=30)	5.02 ± 0.76 (n=10)
AGGREGATE	8.30 ± 1.35 (n=120)	4.83 ± 0.72 (n=40)
IID FLOOR	8.52 ± 0.21 (n=100)	8.90 ± 0.17 (n=100)

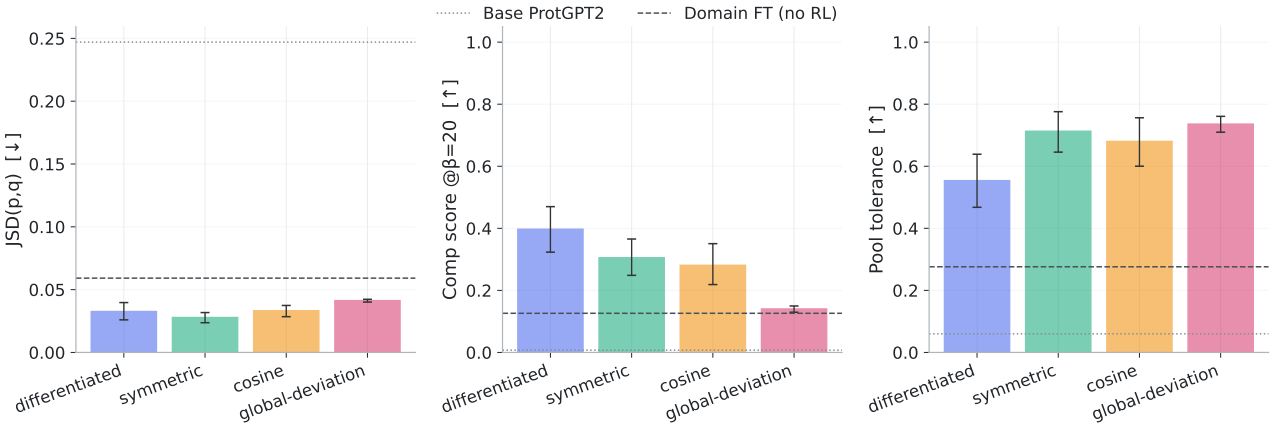

Figure 3. Per-variant means with seed-level 95% bootstrap confidence intervals on JSD, composition score (re-scored at fixed β_{ref}=20), and pool tolerance (q_A, n=30 seeds per variant). Dotted line: base ProtGPT2; dashed line: domain-adaptive FT (no RL).

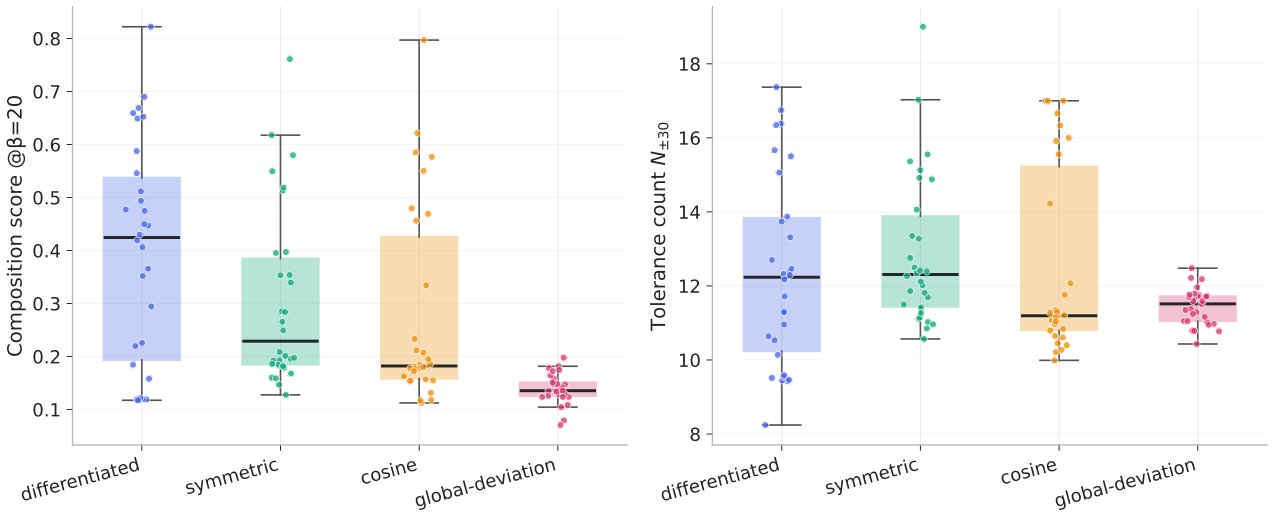

Figure 4. Seed variance per composition term variant on the composition score (left, at β_{ref}=20) and the tolerance count $N_{\pm 30}$ (right) on q_A, n=30 seeds per variant. Boxes summarize the seed distribution; jittered points show individual seeds.

Table 13. Full pairwise Wilcoxon signed-rank statistics on q_A, n=30 paired seeds. Bold p survives Bonferroni at α=0.05 within each metric.

METRIC	A	B	Δ	95% CI	p
6*COMP.	DIFF	SYM	$+0.069$	$[-0.059, +0.166]$	0.10
	DIFF	COS	$+0.118$	$[+0.003, +0.252]$	0.039
	DIFF	GD	$+0.271$	$[+0.111, +0.351]$	$\mathbf{< 10^{-5}}$
	SYM	COS	$+0.031$	$[-0.024, +0.173]$	0.40
	SYM	GD	$+0.114$	$[+0.050, +0.214]$	$\mathbf{< 10^{-6}}$
	COS	GD	$+0.054$	$[+0.030, +0.084]$	$\mathbf{< 10^{-4}}$
6*JSD	DIFF	SYM	$+0.0006$	$[-0.0059, +0.0172]$	0.30
	DIFF	COS	-0.0034	$[-0.0125, +0.0107]$	0.77
	DIFF	GD	-0.0130	$[-0.0207, +0.0020]$	0.017
	SYM	COS	-0.0076	$[-0.0165, +0.0000]$	0.088
	SYM	GD	-0.0120	$[-0.0184, -0.0045]$	$\mathbf{< 10^{-6}}$
	COS	GD	-0.0033	$[-0.0069, +0.0008]$	0.024
6*$N_{\pm 30}$	DIFF	SYM	-0.40	$[-1.99, +1.19]$	0.30
	DIFF	COS	-0.05	$[-1.66, +1.35]$	0.75
	DIFF	GD	$+0.99$	$[-1.13, +1.86]$	0.21
	SYM	COS	$+0.67$	$[+0.05, +1.70]$	0.37
	SYM	GD	$+0.86$	$[+0.30, +1.68]$	$\mathbf{< 10^{-4}}$
	COS	GD	-0.08	$[-0.55, +0.81]$	0.40

Table 14. Rejection-sampling baseline. For each target, we draw N=500 sequences from base ProtGPT2 and from the FT-only model and rank them by the same fixed-β_{ref} composition score used elsewhere. We report the best and top-10% JSD alongside the best composition score, compared against the best RL run on each target. Even with N=500 draws, the FT prior alone does not reach the JSD or score that the RL stage reaches in a single run, on either target. Values for the two "RL best" rows are the per-run pool means at the best iteration (n=19, n=39 valid sequences after filtering on q_A and q_B respectively).

TARGET	SOURCE	N	$\text{SCORE}_{\text{BEST}} \uparrow$	$\text{JSD}_{\text{BEST}} \downarrow$	$\text{JSD}_{\text{P10}} \downarrow$
q_A	BASE PROTGPT2 (B/N)	500	0.146	0.041	0.091
q_A	FT ONLY (B/N)	500	0.529	0.018	0.028
q_A	RL BEST RUN	19	**0.822**	**0.004**	–
q_B	BASE PROTGPT2 (B/N)	500	0.446	0.024	0.076
q_B	FT ONLY (B/N)	500	0.619	0.011	0.023
q_B	RL BEST RUN	39	**0.830**	**0.001**	–

F. Top-30 ESM-2 pPPL

The per-condition top-30 slice referenced in §5.4 is computed by ranking each policy's candidate pool (the same generation pool that backs Table 3) by the fixed-β_{ref} composition score and scoring the top 30 with ESM-2 (150M) pPPL (Table 15). The top-30 sequences are those most aggressive in matching q from each policy, so this is the worst-case fluency snapshot (rather than the average over the whole pool reported in Table 3). Even on the per-sequence top-30 slice, no variant collapses to a degenerate-low-perplexity regime: all RL policies remain in the same order of magnitude as their respective FT prior, and the differentiated reward stays closest to the FT prior on q_A.

G. iid floor: length dependence

Table 16 reports ESM-2 pPPL of the composition-only iid baseline at $L \in \{100, 250, 400\}$ AA on both targets. The floor shrinks by less than 1 pPPL unit between L=100 and L=400 on each target, supporting the use of the L=400 value in Table 4 as an approximately length-independent floor for the RL policies (whose mean lengths fall inside this range, Table 3).

Table 15. ESM-2 (150M) pseudo-perplexity (pPPL) on the top-30 composition-scoring sequences from each policy. Lower is more natural-protein-like.

	q_A		q_B	
Source	pPPL↓	$\bar{L}$	pPPL↓	$\bar{L}$
base ProtGPT2	4.52	352	4.23	360
FT only	6.13	349	6.39	258
RL global-deviation	15.42	462	7.24	431
RL cosine	9.80	364	6.85	466
RL symmetric	17.00	440	10.76	489
RL differentiated	8.48	307	12.29	403

Table 16. ESM-2 (150M) pseudo-perplexity (pPPL) of the composition-only iid floor at three lengths, per target (n=100 sequences each, residues sampled iid from the corresponding target AA distribution; mean±std).

L	IID PPPL (q_A)	IID PPPL (q_B)
100	18.97 ± 1.06	19.27 ± 1.14
250	18.46 ± 0.74	18.77 ± 0.80
400	18.07 ± 0.57	18.50 ± 0.55

H. Diversity pulse and re-injection

The diversity-pulse mechanism introduced in §3.3 is parameterized as follows. The baseline candidate pool is sampled at temperature T=0.80 and top-p=0.90, and the diversity-pulse pool is sampled at T=0.88 and top-p=0.93 (i.e. both knobs slightly hotter). Both pools pass through the same length and MMseqs2 identity filters before merging. Two fractions schedule the merge across training:

- **Re-inject fraction** (carry-over of merged pool to the next iteration's prompt set): 0.35 for the first 20% of iterations, 0.45 during the early-boost window, and 0.55 during the late ($\geq 80\%$) window.
- **Pulse mix-in fraction** (share of the diversity-pulse pool added to the baseline pool inside one iteration): 0.20, 0.30, and 0.35 on the same three windows.

Hotter decoding alone would degrade reward. The re-injection schedule is what turns the diversity pulse into a useful signal, and the schedule is identical across all 160 runs reported in the main text.

Figure 5 summarizes the per-iteration composition score (re-scored at fixed $\beta_{\text{ref}} = 20$) of sequences sampled by the two decoders, aggregated across the top 6 *differentiated* composition seeds on q_A (matching the seed selection used in Fig. 1). The diversity-pulse pool tracks the baseline pool throughout training, lagging it by no more than a few hundredths of a composition-score unit during the middle of training and converging back to it once the policy has contracted around the target. The pulse, therefore, acts as a controlled exploration channel. It does not collapse reward relative to the baseline pool, but it keeps a wider per-sequence distribution available to the reward-weighted update during the regime where the policy is still moving.

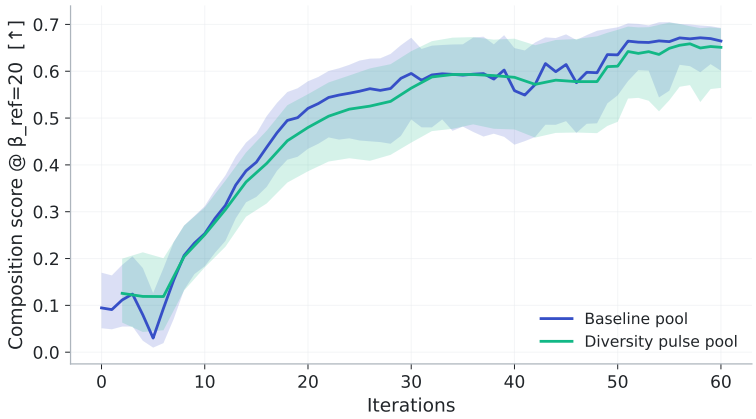

Figure 5. Per-iteration composition score (re-scored at fixed $\beta_{\text{ref}} = 20$ with the *differentiated* variant) of the baseline pool versus the diversity-pulse pool, aggregated across the top 6 *differentiated*-variant seeds on q_{A}. Lines: per-iteration median; shading: per-iteration 25-75% inter-quartile range across all surviving candidate sequences. The pulse pool tracks the baseline pool throughout training, lagging it slightly during the contraction regime (roughly the first 30 iterations) and converging back to it afterwards.