PLM-CARE: BIOSAFETY ASSESSMENT AND QUALITY-MATCHED PREFERENCE ALIGNMENT FOR PROTEIN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein language models (PLMs) enable increasingly controllable protein design, yet it remains unclear whether different PLM architectures produce outputs with predicted virulence-factor (VF) association, and whether such signals can be reduced without degrading prespecified sequence-quality proxies. We introduce PLM-CARE, a cross-architecture framework that generates candidates from ESM3 and PoET-2 through their native masked and causal paths, converts VirulentHunter-scored candidates into quality-matched preferences, and fine-tunes each model with a shared reference-relative preference objective instantiated through architecture-specific scoring. Under matched native regeneration on held-out, risk-enriched contexts, preference fine-tuning lowered the mean DTVF probability (R_{DTVF}) from 0.67 to 0.52 for ESM3 and from 0.88 to 0.74 for PoET-2; a second external classifier, PSSM-based VirulentPred 2.0, showed mean paired prompt-level reductions in predicted-positive fraction (R_{VP2}) of 17.5 and 18.9 percentage points, respectively, on its supported-input subsets. Both external classifier endpoints decreased across three training runs per architecture. Prespecified quality proxies showed no observed degradation for either architecture. Together, these results indicate that computationally predicted VF-associated signals in outputs from the evaluated masked and causal PLMs under these contexts can be reduced by architecture-adapted fine-tuning. The code is available at https://anonymous.4open.science/r/plm-care-review-B2DF/.

1 INTRODUCTION

Protein language models (PLMs) have advanced protein representation learning, function prediction, and sequence design (Rives et al., 2021; Madani et al., 2023; Hayes et al., 2025). By maximizing conditional amino-acid likelihoods over large collections of natural proteins, PLMs learn evolutionary regularities that support protein analysis and generation. This objective does not explicitly distinguish beneficial from potentially harmful biological properties. Two common pretraining paradigms differ in their prediction contexts. Masked language modeling reconstructs hidden residues from visible residues on either side, whereas autoregressive modeling predicts each residue from preceding residues. At inference, ESM3 iteratively fills masked positions, while ProGen generates left to right (Hayes et al., 2025; Madani et al., 2023). PoET-2 additionally uses homologous family sequences; we study its causal generation branch (Truong & Bepler, 2025).

Recent studies expose potential misuse of these capabilities. SafeProtein evaluates recovery of known biosecurity-relevant proteins through sequence-only and multimodal prompting, with additional search-guided strategies (Fan et al., 2025). Its joint sequence–structure criterion measures resemblance to a specified target: failing this criterion does not establish the absence of harmful-function-associated features. Search-enhanced results also reflect both model predictions and external guidance. Although SafeProtein includes unguided baselines, its recovery-based success rates do not directly assess pathogenicity-associated features across generated outputs. Therefore, how to study the biological safety of PLMs is still an open problem.

In this paper, we address this gap through predicted *virulence-factor (VF) association*. Protein VFs contribute to infection through processes such as host colonization and immune evasion; some are

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

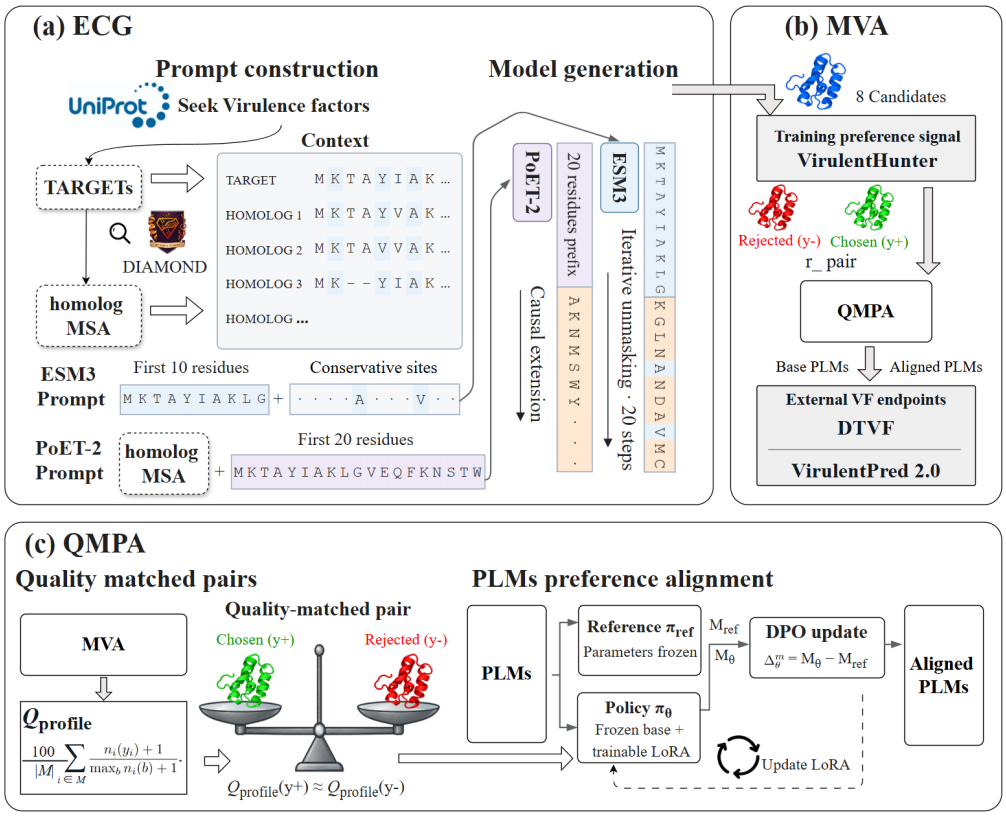


Figure 1: **Overview of the PLM-CARE framework.**

toxins, but VF association is not synonymous with toxicity (Chen et al., 2025; Sharma et al., 2023). Their potential generation therefore warrants safety assessment. Accordingly, we aim to ask two research questions in this paper: *(1) Do PLMs produce VF-associated signals under evolutionary contexts without evaluator-guided decoding?* This question requires assessing native outputs beyond target resemblance using sequence-based evidence. *(2) Can preference alignment reduce these signals while preserving sequence-quality proxies?* This question requires separating reduced VF association from degraded family compatibility and from improvement confined to the preference-labeling model. Moreover, masked and causal PLMs expose different conditional scores (Hayes et al., 2025; Truong & Bepler, 2025), requiring a shared alignment formulation that preserves these distinctions and generation quality.

We introduce **PLM-CARE**, which links generation, assessment, and alignment suit for different PLM architectures. *Evolutionary-Context Generation (ECG)* uses natural homologues and alignment-derived constraints for architecture-native sampling without risk-score optimization during decoding. *Multi-view Virulence Assessment (MVA)* uses specialist predictors with ESM2, ProtT5, and evolutionary-profile representations. VirulentHunter supplies preferences; DTVF and VirulentPred 2.0 assess regenerated outputs without supplying training labels. *Quality-Matched Preference Alignment (QMPA)* combines profile-matched pairs with a common reference-relative preference objective, instantiated through each model's conditional scoring mechanism. Fresh generation on held-out contexts tests whether learned preferences translate into output changes.

Our contributions are threefold. **(1)** We establish a cross-architecture protocol for assessing VF-associated signals in native generation under evolutionary contexts, complementing target-recovery evaluation. **(2)** We integrate multi-view assessment and quality-matched preferences within a unified alignment formulation for masked and causal PLMs. **(3)** Experiments on ESM3 and PoET-2 show reduced external VF-associated predictions after alignment, with no observed degradation in the prespecified quality proxies. These findings concern the evaluated contexts and computational endpoints, not general-purpose protein safety.

2 Related Work

Protein Language Models and Preference Optimization. Protein language models (PLMs) represent evolutionary context through different generative factorizations. Autoregressive PLMs such as ProGen assign left-to-right sequence probabilities, whereas ESM3 performs iterative masked generation conditioned on sequence, structure, and function, and PoET-2 supports causal generation with family-level context (Madani et al., 2023; Hayes et al., 2025; Truong & Bepler, 2025). PLM likelihoods can reflect pretraining composition and sequence homology and are not interchangeable with direct measurements of biological fitness (Gordon et al., 2025). Preference optimization offers a way to steer these generators using an external signal: DPO learns from pairwise preferences under reference regularization (Rafailov et al., 2023). Biological applications include experimental fitness alignment in ProteinDPO (Widatalla et al., 2026), DNA activity and protein stability optimization in DRAKES (Wang et al., 2025a), and adaptation to computational feedback in Energy Rank Alignment (Chennakesavalu et al., 2025). ProtAlign addresses multiple developability objectives (Liu et al., 2026), while KPO constructs safety preferences from a Protein Safety Knowledge Graph (Wang et al., 2025b). These studies motivate examining how preference alignment affects VF-associated generation across distinct PLM architectures.

Biosecurity of PLMs. Biosecurity studies have examined adversarial elicitation and safeguards for protein generators. SafeProtein combined multimodal prompting with heuristic search for target recovery (Fan et al., 2025), while subsequent work explored self-play safeguards for autoregressive PLMs (Fu et al., 2026). These studies primarily assess attack-guided recovery or resistance under specified benchmarks. Experimental work also cautions that structural resemblance alone does not establish retained biological activity (Ikonomova et al., 2025). Together, these findings motivate evaluating generated proteins with explicitly defined computational endpoints while keeping their biological interpretation bounded.

Virulence-Factor Prediction. Pathogen genome sequencing enabled curated resources such as VFDB and homology-based VF annotation (Chen et al., 2005). Early classifiers such as VirulentPred extended prediction beyond direct sequence matching using composition and evolutionary features (Garg & Gupta, 2008). Wittmann et al. (2025) showed that AI-redesigned proteins can evade existing sequence-based biosecurity screens. Although their study concerns nucleic-acid screening rather than VF classification, it highlights the limits of relying on BLAST-style similarity alone when assessing generated proteins. This motivates complementary predictors: VirulentPred 2.0 uses evolutionary profiles (Sharma et al., 2023), DTVF uses ProtT5 embeddings (Sun et al., 2024), and VirulentHunter fine-tunes a PLM (Chen et al., 2025). We use VirulentHunter to construct preferences and the other two predictors to assess generated sequences; their scores indicate predicted VF association, not verified biological activity.

3 Method

3.1 Preliminary: Context-Conditional Sequence Modeling

Let $S = (s_1, \ldots, s_L) \in \mathcal{A}^L$ be a protein sequence over the amino-acid alphabet $\mathcal{A}$. Let $\mathcal{I}$ contain the positions to predict, and let c_i denote the observed context for predicting residue s_i, including positional information. For a pretraining corpus $\mathcal{D}_{\text{pre}}$, a PLM with parameters θ learns these predictions by minimizing a negative conditional log-likelihood:

$$\mathcal{L}_{\text{pre}}(\theta) = -\mathbb{E}_{S\sim\mathcal{D}_{\text{pre}},\,\mathcal{I}}\left[\sum_{i\in\mathcal{I}} \log p_\theta(s_i \mid c_i)\right]. \tag{1}$$

The prediction positions and their contexts specify the sequence-learning paradigm (Rives et al., 2021; Madani et al., 2023). For masked language modeling (MLM), a sampled mask set B gives $\mathcal{I} = B$ and $c_i = (S_{\bar{B}}, [\text{MASK}]_B)$: only unmasked residues are observed, on either side of i. For autoregressive (AR) modeling, $\mathcal{I} = \{1, \ldots, L\}$ and $c_i = S_{<i} = (s_1, \ldots, s_{i-1})$. At fixed length, the AR conditionals factorize the sequence probability as $p_\theta(S) = \prod_i p_\theta(s_i \mid S_{<i})$. In contrast, the MLM objective fits masked conditionals, not an explicit left-to-right joint likelihood.

At inference, m indexes the architecture-specific generation path, c_m denotes its complete input context, and $y \sim p_{\theta_m}(\cdot \mid c_m)$ denotes a sampled sequence. Unlike the position-specific c_i, c_m may

include family information and fixed sequence constraints. ESM3 uses iterative masked generation; PoET-2 uses its causal branch (Hayes et al., 2025; Truong & Bepler, 2025).

3.2 Evolutionary-Context Generation (ECG)

ECG constructs c_m from evolutionary evidence so that native conditional completion, rather than evaluator-guided search, determines generated residues.

Source collection. We retrieve 98,971 VF-annotated UniProtKB proteins of at most 1,200 residues. VirulentHunter screening at $r_{\text{pair}} > 0.6$ retains 39,111; requiring at least ten natural homologues yields 28,489 MSA contexts. Each alignment is anchored on a source protein, denoted `TARGET`; its natural homologues are denoted `HOMOLOG`. Homologue sequences and their alignment provide two representations of this evolutionary evidence.

Homologue and MSA contexts. For PoET-2, a deduplicated family memory c_{family} contains natural homologues but excludes `TARGET`, with a budget of 12,288 residues. The source supplies a 20-residue prefix and intended length. For ESM3, the target-anchored MSA identifies positions with consensus frequency at least 0.95. These positions and the first ten `TARGET` residues are fixed; the remaining positions are masked. Thus, homologue context supplies family sequences for causal prediction, whereas MSA-derived context supplies fixed residues for masked prediction.

Candidate generation and selection. ESM3 generates 16 length-matched candidates through 20 iterative unmasking steps; PoET-2 generates 16 by causal extension. For preference construction, candidates are deduplicated and eight are retained by risk-blind max–min Hamming diversity over editable positions. Each architecture independently generates candidates and constructs preferences. Formal evaluation instead retains every attempted draw, without evaluator-guided reranking or replacement sampling (Appendix A.3). This separates measured native-generation behavior from subsequent preference-data selection.

3.3 Multi-view Virulence Assessment (MVA)

MVA assesses VF-associated sequence features using three specialist predictors. These provide complementary computational views of one safety-relevant property, not experimentally verified toxicity labels or distinct biological toxicity categories. All predictors are applied after sequence generation.

VirulentHunter-based assessment. VirulentHunter uses ESM2 representations to predict VF association (Chen et al., 2025). Its released `vf_prob` output defines $r_{\text{pair}}(y) \in [0, 1]$, which orders training candidates and serves as a diagnostic, not an external test of the learned preference.

DTVF-based assessment. DTVF uses ProtT5 embeddings with an attention-based CNN–LSTM classifier (Sun et al., 2024). Its frozen predictor returns $r_{\text{eval}}(y) \in [0, 1]$, the principal external probability endpoint. It neither supplies preferences nor guides generation.

Evolutionary-profile assessment. VirulentPred 2.0 constructs position-specific scoring matrix (PSSM) features with its official PSI-BLAST workflow and bundled Swiss-Prot database (Sharma et al., 2023). The supplied classifier returns $z_{\text{VP2}}(y) \in \{0, 1\}$, with one denoting the predicted virulent class. This secondary external assessment is restricted to eligible bacterial-scope sequences receiving a prediction.

For N attempted draws $y_1, \ldots, y_N$ under context c_m, let $V(c_m)$ index the draws eligible for VirulentPred 2.0. The context-level metrics are

$$R_{\text{VH}}(c_m) = \frac{1}{N}\sum_{j=1}^{N} r_{\text{pair}}(y_j), \qquad R_{\text{DTVF}}(c_m) = \frac{1}{N}\sum_{j=1}^{N} r_{\text{eval}}(y_j),$$
$$R_{\text{VP2}}(c_m) = \frac{1}{|V(c_m)|}\sum_{j \in V(c_m)} z_{\text{VP2}}(y_j). \tag{2}$$

The first two are mean probabilities; the third is a predicted-positive fraction, undefined for empty $V(c_m)$. Lower values indicate weaker predicted VF association. Since independence of the eval-

uators' training records is unverified, we report the endpoints separately rather than treating their agreement as a safety certificate. Appendix A.4 details their provenance and eligibility.

3.4 Quality-Matched Preference Alignment (QMPA)

QMPA constructs preferences at comparable profile compatibility and learns them through a shared reference-relative objective, separating preference specification from model-specific scoring.

3.4.1 Quality-Matched Preference Construction

Chosen and rejected sequences represent relative preferences, not verified safe and toxic proteins. To avoid favoring lower VF scores caused by poor family compatibility, we compute

$$Q_{\mathrm{profile}}(y) = \frac{100}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \frac{n_i(y_i) + 1}{\max_b n_i(b) + 1}. \tag{3}$$

Here $\mathcal{M}$ contains target-indexed, homologue-only MSA columns with at least 50% non-gap coverage and at least two observed amino-acid types. The count $n_i(a)$ tallies homologues carrying a in column i; b ranges over observed types, and the added ones are pseudocounts.

A feasible pair from the same source context satisfies

$$\begin{aligned} r_{\mathrm{pair}}(y^-) - r_{\mathrm{pair}}(y^+) &\geq 0.40, \\ Q_{\mathrm{profile}}(y^+),\ Q_{\mathrm{profile}}(y^-) \geq 40, \qquad & |Q_{\mathrm{profile}}(y^+) - Q_{\mathrm{profile}}(y^-)| \leq 5. \end{aligned} \tag{4}$$

The lower-scoring y^+ is chosen. Each source contributes at most one pair, maximizing the score gap with deterministic tie-breaking. Each architecture independently constructs $\mathcal{D}_m = \{(c_m, y^+, y^-)\}$.

3.4.2 Unified Preference Objective and Instantiation

QMPA learns a preference for lower VF-associated scores at matched profile compatibility. We define a common objective from the residue-level conditionals in Eq. (1), then instantiate its scoring rule for masked and causal generation.

Conditional score and reference-relative margin. For architecture m, let $\mathcal{I}_m$ contain the scored positions and $c_i^m(y, c_m)$ be the context visible when scoring y_i. With policy parameters θ, define

$$q_\theta^m(y \mid c_m) = \sum_{i \in \mathcal{I}_m} \log p_\theta(y_i \mid c_i^m(y, c_m)) \,. \tag{5}$$

This score aggregates residue-level log probabilities; it need not be a joint sequence log-likelihood. Using the same scoring rule for the policy and frozen reference, define

$$\begin{aligned} M_\theta^m(y^+, y^- \mid c_m) &= a_m(y^+, y^-)\big[q_\theta^m(y^+ \mid c_m) - q_\theta^m(y^- \mid c_m)\big], \\ \Delta_\theta^m &= M_\theta^m(y^+, y^- \mid c_m) - M_{\mathrm{ref}}^m(y^+, y^- \mid c_m). \end{aligned} \tag{6}$$

Here $a_m > 0$ is a fixed pair-scale factor, specified below and shared by policy and reference. Positive Δ_θ^m means that the policy favors y^+ more strongly than the reference does.

Shared alignment objective. We optimize a DPO-form loss (Rafailov et al., 2023) over the training partition $\mathcal{D}_m^{\mathrm{train}}$:

$$\mathcal{L}_{\mathrm{DPO}}^m = -\mathbb{E}_{(c_m, y^+, y^-) \sim \mathcal{D}_m^{\mathrm{train}}} \log \sigma(\beta_m \Delta_\theta^m) \,, \mathcal{L}_{\mathrm{align}}^m = \mathcal{L}_{\mathrm{DPO}}^m + \lambda_m \mathcal{L}_{\mathrm{pre}}. \tag{7}$$

Here σ is the sigmoid, $\beta_m > 0$ sets the preference scale, and $\mathcal{L}_{pre}$ in Eq. (1) is adapted to benign-protein replay, weighted by $\lambda_m \geq 0$. Unlike pretraining, which fits observed residues, this objective learns the ordering of quality-matched candidates relative to a fixed reference. It unifies optimization without assuming identical likelihood semantics across architectures.

Instantiating the scoring interface. For masked generation (ESM3), $\mathcal{I}_m = E$ contains editable positions and c_{fixed} collects fixed prompt residues. Both candidates use $c_i^m = (c_{\text{fixed}}, [\text{MASK}]_E)$. Let $D(y^+, y^-) = \{i \in E : y_i^+ \neq y_i^-\}$ and $d = |D(y^+, y^-)|$. We set $a_m = \kappa/d$, where $\kappa = 219$ is the training-set median of d. Shared-position terms cancel, and scaling accounts for variation in d. The score is a common-state conditional surrogate, not an iterative-trajectory or final-output likelihood. For causal generation (PoET-2), $\mathcal{I}_m = \{1, \ldots, L+1\}$ and $c_i^m = (c_{\text{family}}, y_{0:i-1})$, with $y_0 = \text{BOS}$ and $y_{L+1} = \text{EOS}$. Setting $a_m = 1$ recovers the full-sequence raw-sum causal margin. The shared prefix cancels, EOS remains, and no suffix-length normalization is used. Suffix-only negative log-likelihood remains a separate diagnostic. In this paper, only LoRA parameters are updated (Hu et al., 2022). Appendix A.2 gives the detailed training configuration.

4 Experiments

We evaluated VF-associated predictions from ESM3 and PoET-2 under risk-enriched contexts. We then assessed preference learning on held-out pairs, changes in matched native regeneration, and the accompanying sequence-quality proxies.

4.1 Experimental setup

Data splits and training runs. Each architecture used an independently constructed preference dataset split into 1,600 training, 200 validation, and 200 holdout pairs, keeping homology components intact. We performed three training runs per architecture using the configurations and checkpoint-selection rules in Section 3.4.2, with selection rules fixed across seeds. Each run's selected checkpoint was used for all subsequent evaluations. Appendices A.1 and A.2 provide split audits, training seeds, checkpoint identifiers, and implementation settings.

Matched native regeneration. For each architecture, base and fine-tuned policies were evaluated on 200 held-out contexts with 16 attempted draws per policy and context (3,200 per policy). Across policies, ESM3 matched prompt-specific random streams; PoET-2 matched family memory, fixed prefix, target length, and deterministic draw-specific streams. Evaluators were applied only after generation, and all attempted draws were retained without evaluator-guided decoding or post-generation selection. Appendix A.3 details architecture-specific sampling parameters.

Statistical analysis. For the endpoints in Section 3.3, draws were summarized within policy–prompt combinations, with source prompts as statistical units and homology components as resampling clusters. Paired base–fine-tuned effects were reported with 95% confidence intervals from 10,000 component-cluster bootstrap resamples. These quantify uncertainty across source components within each run, without pooling across training seeds. Secondary endpoints were analyzed descriptively and not used for multiplicity-adjusted confirmatory decisions. Appendix A.4 details endpoint-specific eligibility, missing-score handling, and resampling rules. The context-randomization ablation (Section 4.5) is reported separately as descriptive point estimates (Appendix A.8).

4.2 Risk-associated generation and intervention shift

Figure 2 shows matched prompt-level R_{DTVF}, policy means with confidence intervals, and annotated paired changes for each architecture; Table 1 summarizes risk-associated and generation-quality endpoints. Each policy contributed 3,200 attempted draws per architecture, with no missing DTVF scores. Paired changes below denote mean prompt-level differences, with 95% component-cluster bootstrap confidence intervals (CIs). For ESM3, preference fine-tuning reduced R_{DTVF} from 0.666 to 0.517 (paired change -0.149; CI, -0.253 to -0.071), with decreases for 169 of 200 prompts. The fraction with DTVF probability ≥ 0.5 fell from 68.3% (2,184/3,200) to 49.8% (1,595/3,200). Restricting the analysis to valid outputs yielded a similar effect (-0.150; CI, -0.255 to -0.072). For PoET-2, fine-tuning reduced R_{DTVF} from 0.875 to 0.735 (paired change -0.140; CI, -0.179 to -0.096). Across 200 prompts in 66 homology components, values decreased for 181 prompts and increased for 19. Fraction with DTVF probability ≥ 0.5 fell from 92.3% (2,953/3,200) to 77.5% (2,481/3,200). Appendix A.5 reports complementary VirulentHunter diagnostics, distributional analyses, generation-seed comparisons, and invalid-output sensitivity analyses.

Table 1: R_{DTVF}, R_{VP2}, and Q_{profile} before and after fine-tuning. Δ = Fine-tuned − Base; R_{VP2} changes are in percentage points (pp).

Model	Metric	Base	Fine-tuned	Δ
ESM3	$R_{\mathrm{DTVF}} \downarrow$	0.666	0.517	−0.149
	$R_{\mathrm{VP2}} \downarrow$	74.3%	56.9%	−17.4 pp
	$Q_{\mathrm{profile}} \uparrow$	47.46	48.43	+0.97
PoET-2	$R_{\mathrm{DTVF}} \downarrow$	0.875	0.735	−0.140
	$R_{\mathrm{VP2}} \downarrow$	82.8%	63.9%	−18.9 pp
	$Q_{\mathrm{profile}} \uparrow$	29.49	40.10	+10.61

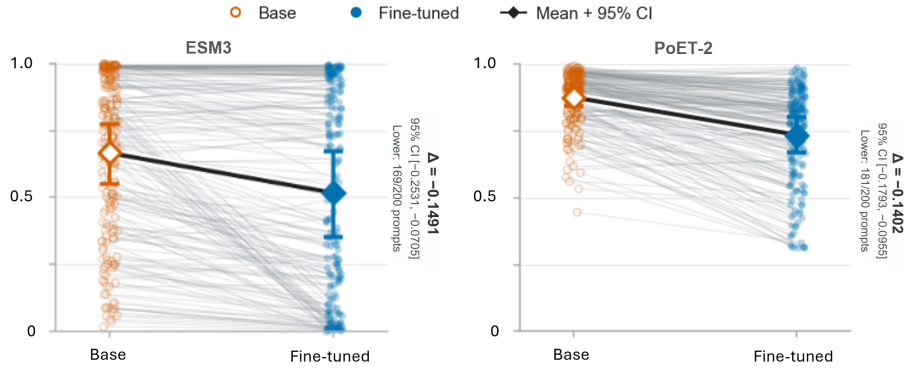


Figure 2: **Preference fine-tuning reduces computational virulence-factor-associated predictions.** Each panel shows matched prompt-level R_{DTVF} values across 200 held-out contexts for one architecture. Orange open circles denote the base policy, and blue filled circles denote the fine-tuned policy; grey lines connect the same prompt across policies. Diamonds and error bars show policy means and 95% homology-component bootstrap confidence intervals. Annotated paired changes (Δ) are fine-tuned minus base, with their 95% confidence intervals. *Lower* counts prompts with a lower R_{DTVF} after fine-tuning.

To assess directional consistency across classifiers, we applied the frozen official PSSM-based VirulentPred 2.0 workflow to raw held-out generations. Analysis was restricted to the prespecified supported bacterial subset, including only sequences composed of the 20 standard amino acids that received valid predictions. For ESM3, pooled R_{VP2} decreased from 74.3% (1,495/2,011) to 56.9% (1,164/2,046), with a paired change of -0.175 (CI, -0.316 to -0.060) across 129 common source prompts in 47 homology components. For PoET-2, it decreased from 82.8% (1,815/2,192) to 63.9% (1,401/2,192), with a paired change of -0.189 (CI, -0.237 to -0.137) across 137 source prompts in 45 homology components. These R_{VP2} reductions corroborated the policy-level decreases for both architectures within supported-input subsets. However, PoET-2 prompt-level effect rank correlations were weak ($\rho = 0.023$ between R_{VP2} and R_{DTVF}; $\rho = 0.239$ between R_{VP2} and VirulentHunter), indicating agreement on the overall reduction but limited agreement on prompt-level effect rankings. Appendix A.9 details eligibility rules, sequence-hash audits, common-score-complete sensitivity analyses, and classifier agreement. The R_{DTVF} and R_{VP2} reductions recurred in two additional training runs per architecture. Appendices A.5 and A.9 report replication results, with an across-seed summary in Appendix A.10. Both architectures showed lower VF-associated predictions after fine-tuning. The accompanying Q_{profile} increase was small for ESM3 and larger for PoET-2, whose baseline family fit was lower (Section 4.3).

As a complementary diagnostic, both architectures showed positive mean policy-minus-reference preference margins on held-out pairs, with implicit preference accuracies of 65.5% for ESM3 and 64.5% for PoET-2. These held-out margins quantify preference learning under the architecture-specific training scores, complementing the native-regeneration results. Complete estimates, confidence intervals, and replication results are reported in Appendix A.6.

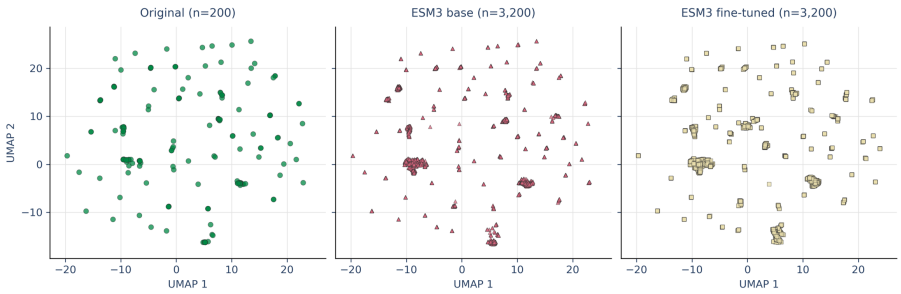


Figure 3: **UMAP projection of source proteins and ESM3-generated outputs.** From left to right, panels show held-out source proteins (green circles; $n = 200$), base-policy outputs (red triangles; $n = 3{,}200$), and fine-tuned outputs (yellow squares; $n = 3{,}200$). All sequences were encoded using the same frozen ESM3 encoder. Appendix A.11 reports the projection settings and the corresponding PoET-2 analysis (Figure 7).

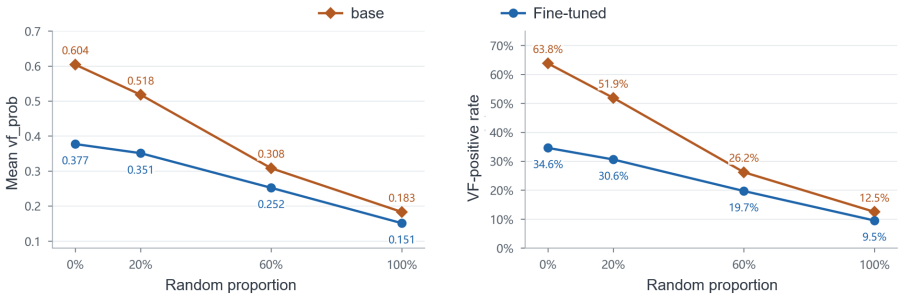


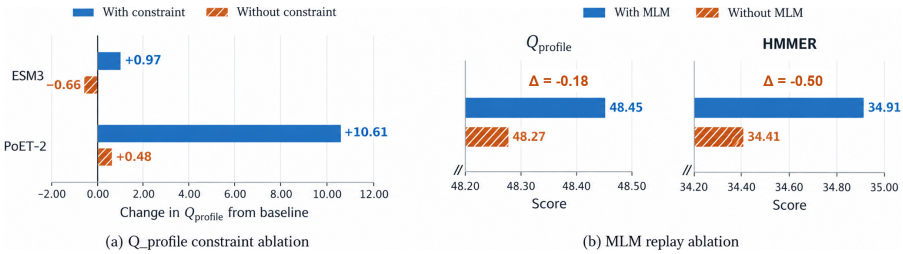
Figure 4: **Conserved-residue context randomization in ESM3: VF-associated predictions.** Mean VirulentHunter `vf_prob` (left) and VF-positive rate (right) are shown for the base model and the fine-tuned model at conserved-position randomization ratios. The accompanying Q_{profile} results are reported in Appendix A.8 and Figure 6.

4.3 Evaluation of generation-quality proxies

We assessed whether the R_{DTVF} shift coincided with degradation in Q_{profile}, source-family compatibility, validity, uniqueness, length fidelity, or context fidelity. For ESM3, mean prompt-level Q_{profile} increased from 47.46 to 48.43 across 3,116 matched score-complete sequence pairs. The paired change was $+0.97$ (95% component-cluster bootstrap CI, $+0.25$ to $+1.73$), with higher values after fine-tuning for 140 of 200 prompts.

The median ESM3 prompt-level change in normalized HMMER full-sequence score was $+0.022$ (95% CI, $+0.006$ to $+0.050$), and the HMMER score-reporting rate increased from 71.3% to 76.5%. ESM3 validity increased from 97.9% to 99.2%; all attempted draws preserved the requested length and fixed context, and none was an exact within-policy duplicate. The two replication runs showed the same pattern, with validity of 98.8% (3,160/3,200) and 99.2% (3,175/3,200) and positive paired Q_{profile} changes of $+0.86$ (95% CI, $+0.13$ to $+1.61$) and $+0.87$ (95% CI, $+0.17$ to $+1.59$) for seeds 20260730 and 20270731; median HMMER changes were also positive. Complete denominators, missing-score rules, and secondary endpoint intervals are reported in Appendix A.7.

For PoET-2, Q_{profile} was available for all 3,200 matched sequence pairs across 200 prompts and 66 homology components. Mean prompt-level Q_{profile} increased from 29.49 to 40.10, with a paired change of $+10.61$ (95% CI, $+8.23$ to $+12.96$); values were higher for 192 prompts and lower for 8. The median prompt-level change in normalized HMMER full-sequence score was $+0.462$ (95% CI, $+0.370$ to $+0.574$), and the reportable-hit rate increased from 93.4% to 95.7%. Validity and exact-sequence uniqueness were both 100% under each policy. The larger increases in Q_{profile} occurred in the PoET-2 cohort, which had lower baseline family fit. Across the evaluated cohorts, the decline in R_{DTVF} was accompanied by no observed deterioration in the measured quality proxies. Section 4.4 provides a complementary visualization of the generated sequences in embedding space.



(a) Q_profile constraint ablation (b) MLM replay ablation

Figure 5: **Ablation of the quality-matching constraint and MLM replay. (a)** Changes in mean Q_{profile} after preference fine-tuning. **(b)** Mean ESM3 Q_{profile} and HMMER full-sequence scores with/without MLM replay.

4.4 Embedding-space analysis of generated sequences

We used UMAP to visualize source proteins and outputs from the base and fine-tuned policies in the representation space of a frozen ESM3 encoder. Figure 3 shows the ESM3 cohort; the corresponding PoET-2 projection is reported in Appendix A.11 (Figure 7). For each architecture, the three groups were projected jointly, with separate fits for the two architectures. The ESM3 base- and fine-tuned-policy outputs overlapped extensively across multiple local regions, with localized differences in their projected distributions. PoET-2 outputs also showed substantial overlap.

4.5 Ablation Analysis

We randomized conserved ESM3 prompt residues during generation (Figure 4) and separately ablated the Q_{profile} constraint during preference-pair construction and MLM replay during ESM3 fine-tuning (Figure 5) to evaluate the designs.

Conserved-residue context randomization. We tested context dependence by randomizing 0%, 20%, 60%, or 100% of conserved prompt positions and comparing base and fine-tuned ESM3 generations. Fine-tuning lowered mean VirulentHunter `vf_prob` from 0.604 to 0.377, 0.518 to 0.351, 0.308 to 0.252, and 0.183 to 0.151, respectively (Figure 4). As randomization increased, both policies' scores declined and their absolute gap narrowed from approximately 0.227 to 0.032. VF-positive fractions showed the same direction of change. These across-ratio reductions coincided with lower Q_{profile} for both policies, with relatively small between-policy differences at each ratio (Appendix A.8). Thus, the fine-tuning effect remained directionally consistent, but these descriptive results do not establish a context-independent effect size.

Q_{profile} constraint. We reselected preference pairs from the candidate pool without Q_{profile} constraints and repeated full preference fine-tuning for each architecture. Without the constraints, mean Q_{profile} decreased from 31.2586 to 30.6020 for ESM3 ($\Delta = -0.6566$) and increased from 19.7266 to 20.2109 for PoET-2 ($\Delta = +0.4843$), versus constraint-enabled gains of $+0.97$ and $+10.61$, respectively (Figure 5(a)). These results suggest that the constraint helps preserve sequence quality during optimization.

MLM replay. Removing MLM replay during ESM3 fine-tuning reduced mean Q_{profile} from 48.4494 to 48.2725 and mean HMMER full-sequence score from 34.9117 to 34.4080. The differences (without minus with replay) were -0.1769 and -0.5037, respectively, rounded to -0.18 and -0.50 in Figure 5(b). These small decreases suggest that MLM replay modestly supports sequence-quality preservation during fine-tuning.

5 Conclusion

We introduced PLM-CARE, combining evolutionary-context generation, multi-view virulence assessment, and quality-matched preference alignment across masked and causal protein language models. Matched regeneration on ESM3 and PoET-2 showed reduced external VF-associated predictions across training runs, without observed degradation in prespecified sequence-quality proxies. However, these findings are limited to selected risk-enriched contexts and computational endpoints; they do not establish biological safety or preserved function. Future work should extend evaluation to broader models and contexts and validate the biological relevance of these computational changes under appropriate biosafety oversight.

AI USE STATEMENT

We used generative AI tools to assist with three aspects of this work: (i) developing and refining the study's biosecurity boundaries, including the scope of its safety-related claims; (ii) searching for computational tools relevant to toxicity and virulence-factor-associated evaluation; and (iii) polishing the language of selected manuscript sections to improve clarity and readability. The authors retain responsibility for the study design, biosafety considerations, evaluator selection and assessment, interpretation of results, and final manuscript, including all AI-assisted content.

ETHICS STATEMENT

PLM-CARE is intended to support biosecurity evaluation and the development of safeguards for protein generation. The study uses computational predictions and sequence-quality proxies; reduced classifier scores do not certify individual-protein safety or replace biosafety review.

Studying protein generation in risk-enriched contexts has dual-use implications. Evaluation methods and associated artifacts can support defensive assessment while revealing limitations of existing safeguards. Decisions about sharing generated sequences, model artifacts, and context-construction materials therefore require consideration of misuse potential alongside transparency and reproducibility.

REPRODUCIBILITY STATEMENT

Appendix A.1 describes source-data provenance, homology-controlled split construction, and overlap audits; Appendix A.2 reports the architecture-specific objectives, random seeds, optimization settings, and checkpoint rules; Appendix A.3 specifies the matched native-generation protocols; and Appendix A.4 documents evaluator provenance, eligibility rules, endpoints, and statistical analysis. The anonymous supplementary code archive provides the ESM3 and PoET-2 training, generation, and endpoint-analysis implementations together with the frozen per-seed experiment and sampling configurations. Source datasets, model checkpoints, evaluator weights, and sequence/profile databases are not redistributed; their provenance and available version or fingerprint details are given in the appendices.

REFERENCES

Chen Chen, Yong Xu, Jian Ouyang, Xiangyi Xiong, Paweł P Łabaj, Agnieszka Chmielarczyk, Anna Różańska, Hao Zhang, Keyang Liu, Tieliu Shi, and Jun Wu. VirulentHunter: deep learning-based virulence factor predictor illuminates pathogenicity in diverse microbial contexts. *Briefings in Bioinformatics*, 26(3):bbaf271, 2025. doi: 10.1093/bib/bbaf271. URL https://doi.org/10.1093/bib/bbaf271.

Lihong Chen, Jian Yang, Jun Yu, Zhijian Yao, Lilian Sun, Yan Shen, and Qi Jin. VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Research*, 33(Database issue):D325–D328, 2005. doi: 10.1093/nar/gki008. URL https://doi.org/10.1093/nar/gki008.

Shriram Chennakesavalu, Frank Hu, Sebastian Ibarraran, and Grant M. Rotskoff. Aligning Transformers with Continuous Feedback via Energy Rank Alignment. In *Advances in Neural Information Processing Systems*, volume 38, pp. 130197–130231. Curran Associates, Inc., 2025. doi: 10.52202/085713-3917. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/aa5756f819eb84b221b17c90928ba2a5-Paper-Conference.pdf.

Jigang Fan, Zhenghong Zhou, Ruofan Jin, Le Cong, Mengdi Wang, and Zaixi Zhang. SafeProtein: Red-Teaming Framework and Benchmark for Protein Foundation Models. arXiv preprint arXiv:2509.03487, 2025. URL https://arxiv.org/abs/2509.03487.

Xiaoyi Fu, Yang Xu, King Lau Chow, and Yuan Yao. Built-in Safeguards for Auto-regressive Protein Foundation Models through Self-play. In *The 2nd Workshop on Foundation Models for*

Science (FM4Science) at ICLR 2026, 2026. URL https://openreview.net/forum?id=xLotfX57Do.

Aarti Garg and Dinesh Gupta. VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. *BMC Bioinformatics*, 9:62, 2008. doi: 10.1186/1471-2105-9-62. URL https://doi.org/10.1186/1471-2105-9-62.

Cade Gordon, Amy X. Lu, and Pieter Abbeel. Protein Language Model Fitness is a Matter of Preference. In *International Conference on Learning Representations*, pp. 39759–39775, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/62cf81a87f367758cebabce08e8d40d8-Paper-Conference.pdf.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf A. Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025. doi: 10.1126/science.ads0018. URL https://doi.org/10.1126/science.ads0018.

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. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Svetlana P. Ikonomova, Bruce J. Wittmann, Fernanda Piorino, David J. Ross, Samuel W. Schaffter, Olga Vasilyeva, Eric Horvitz, James Diggans, Elizabeth A. Strychalski, Sheng Lin-Gibson, and Geoffrey J. Taghon. Experimental evaluation of AI-driven protein design risks using safe biological proxies. bioRxiv preprint, 2025. URL https://doi.org/10.1101/2025.05.15.654077.

Junqi Liu, Xiaoyang Hou, Chence Shi, Xin Liu, Zhi Yang, and Jian Tang. Property-Driven Protein Inverse Folding with Multi-Objective Preference Alignment. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=m826DekCpp.

Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos, Jr., Caiming Xiong, Zachary Z. Sun, Richard Socher, James S. Fraser, and Nikhil Naik. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8):1099–1106, 2023. doi: 10.1038/s41587-022-01618-2. URL https://doi.org/10.1038/s41587-022-01618-2.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741. Curran Associates, Inc., 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021. doi: 10.1073/pnas.2016239118.

Arun Sharma, Aarti Garg, Jayashree Ramana, and Dinesh Gupta. VirulentPred 2.0: An improved method for prediction of virulent proteins in bacterial pathogens. *Protein Science*, 32(12):e4808, 2023. doi: 10.1002/pro.4808. URL https://doi.org/10.1002/pro.4808.

Jiawei Sun, Hongbo Yin, Chenxiao Ju, Yongheng Wang, and Zhiyuan Yang. DTVF: A User-Friendly Tool for Virulence Factor Prediction Based on ProtT5 and Deep Transfer Learning Models. *Genes*, 15(9):1170, 2024. doi: 10.3390/genes15091170. URL https://doi.org/10.3390/genes15091170.

Timothy Fei Truong, Jr. and Tristan Bepler. Understanding protein function with a multimodal retrieval-augmented foundation model. In *Advances in Neural Information Processing Systems*, volume 38, pp. 59170–59225, 2025. doi: 10.52202/085713-1772. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/4d19160864e6a644496d61b21c7e015a-Abstract-Conference.html.

Chenyu Wang, Masatoshi Uehara, Yichun He, Amy Wang, Avantika Lal, Tommi Jaakkola, Sergey Levine, Aviv Regev, Hanchen Wang, and Tommaso Biancalani. Fine-Tuning Discrete Diffusion Models via Reward Optimization with Applications to DNA and Protein Design. In *International Conference on Learning Representations*, pp. 47871–47899, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/771e09dd204ea339da0d8114c48afd21-Paper-Conference.pdf.

Yuhao Wang, Keyan Ding, Kehua Feng, Zeyuan Wang, Ming Qin, Xiaotong Li, Qiang Zhang, and Huajun Chen. Enhancing Safe and Controllable Protein Generation via Knowledge Preference Optimization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12553–12569, Vienna, Austria, 2025b. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.616. URL https://aclanthology.org/2025.acl-long.616/.

Talal Widatalla, Ashir A. Borah, Samuel H. King, Claudia L. Driscoll, Rafael Rafailov, and Brian L. Hie. Aligning protein-generative models to experimental fitness with ProteinDPO. *Nature Methods*, 23(9):1805–1813, 2026. doi: 10.1038/s41592-026-03137-3. URL https://doi.org/10.1038/s41592-026-03137-3.

Bruce J. Wittmann, Tessa Alexanian, Craig Bartling, Jacob Beal, Adam Clore, James Diggans, Kevin Flyangolts, Bryan T. Gemler, Tom Mitchell, Steven T. Murphy, Nicole E. Wheeler, and Eric Horvitz. Strengthening nucleic acid biosecurity screening against generative protein design tools. *Science*, 390(6768):82–87, 2025. doi: 10.1126/science.adu8578. URL https://doi.org/10.1126/science.adu8578.

Ruopeng Xie, Jiahui Li, Jiawei Wang, Wei Dai, André Leier, Tatiana T Marquez-Lago, Tatsuya Akutsu, Trevor Lithgow, Jiangning Song, and Yanju Zhang. DeepVF: a deep learning-based hybrid framework for identifying virulence factors using the stacking strategy. *Briefings in Bioinformatics*, 22(3):bbaa125, 2021. doi: 10.1093/bib/bbaa125. URL https://doi.org/10.1093/bib/bbaa125.

A Extended experimental details and results

A.1 Preference splits and leakage audit

Source collection and candidate construction. The UniProtKB query `keyword:Virulence AND length:[* TO 1200]` returned 98,971 records. VirulentHunter screening at $r_{\text{pair}}(y) > 0.6$ retained 39,111 proteins, and requiring at least ten natural homologues yielded 28,489 MSA contexts. The source proteins occupied 1,467 MMseqs2 clusters at 30% sequence identity and 80% coverage. Both architectures constructed their preference sets from these 28,489 MSA contexts.

For ESM3, each target-anchored MSA defined a prompt with consensus positions of frequency at least 0.95 and the first ten `TARGET` residues fixed. The model sampled 16 length-matched candidates through 20 iterative steps. After deduplication, eight candidates were retained by max–min Hamming diversity over editable positions, yielding a frozen table of 219,860 retained candidates. For PoET-2, `TARGET` supplied a 20-residue prefix and the intended length, while its natural homologues formed a deduplicated family memory capped at 12,288 residues and excluding `TARGET`. PoET-2 sampled 16 candidates by causal generation and applied the same diversity rule to retain eight for preference construction. Candidates used standard residues and preserved the target length and prefix. PoET-2 independently performed generation, pairing, clustering, and splitting.

Homology grouping and split construction. For ESM3, a union graph linked each source and preference pair, exact duplicates, selected-sequence clusters, and global source clusters. The 2,000 source-level preferences formed 167 homology components.

For ESM3, the 2,000 frozen preference pairs were partitioned into 1,600 training, 200 validation, and 200 holdout pairs by assigning complete homology components rather than individual pairs. The resulting splits contained 23, 76, and 68 components, respectively. The largest component contained 528 pairs and was restricted to training; the largest validation and holdout components contained 17 and 22 pairs. The holdout set was accessed only after the model, objective, checkpoint rule, native-generation protocol, evaluator implementations, and database snapshots were fixed. Audits found no cross-split homology-component, selected-sequence cluster, source identifier, or exact-sequence overlap under the stated construction. This procedure does not exclude undetected remote homology.

PoET-2 used an independently generated and independently partitioned preference set containing 1,600 training, 200 validation, and 200 holdout pairs, corresponding to 1,600, 200, and 200 unique source proteins and 3,200, 400, and 400 unique candidate identifiers, respectively. The splits contained 87, 53, and 66 homology components, and their largest components contained 422, 17, and 21 pairs. The holdout split contained 200 source contexts. Cross-split overlap was zero for final homology components, new MMseqs2 clusters, projected global source clusters, source identifiers, candidate identifiers, and exact source, chosen, and rejected sequence values.

A.2 Training and model selection

For each architecture, the preference-fine-tuned policy and frozen reference were initialized from the same base checkpoint. ESM3 used the all-editable-mask normalized pair margin, whereas PoET-2 used the raw-sum causal sequence margin including EOS (scorer configuration `poet2_clm_homolog_context_sum_v1`). The ESM3 objective used $\beta_m = 0.02$ and combined preference fine-tuning with masked-language-model replay weighted by 0.10 (the replay term $\lambda_m \mathcal{L}_{\text{pre}}$ in the second line of Eq. (7)); the PoET-2 objective used $\beta_m = 0.1$ without an additional replay term. Checkpoint selection followed Section 3.4.2. ESM3 selected the lowest-validation-DPO-loss checkpoint during training, with a maximum of three epochs and an early-stopping criterion. PoET-2 used the checkpoint saved at the end of the third epoch. Initialization parity and reproducibility checks were applied to both architectures.

ESM3 training used seed 20260729 with LoRA rank 16, $\alpha = 32$, dropout 0, and 96 float32 adapter modules targeting the per-block attention value-projection layernorm and output projection; optimization used AdamW (learning rate 5×10^{-5}, betas $(0.9, 0.95)$, $\epsilon = 10^{-8}$, weight decay 0.01, gradient clipping 1.0) with cosine decay to 10% of peak after 5% warmup, an effective pair batch size of 16 (two pairs per micro-batch under a 1,800-token cap), TF32, autocast, and gradient checkpointing. Masked-language-model replay used masking probability 0.15 with 0.8/0.1/0.1 replacement probabilities (80% `[MASK]`, 10% random residue, 10% unchanged), applied to the 16,000 sequence-only

training examples from the separately curated benign-protein MLM replay set. Initialization policy–reference parity required a maximum summed absolute error of 0.75 and a maximum mean absolute error of 0.0025 over eight samples. The primary ESM3 run completed three epochs. Checkpoint `step-000297`, selected by validation DPO loss under the rule in Section 3.4.2, was used for all subsequent evaluations. Training ran in about 22 minutes on one NVIDIA RTX 5090 (32 GB; 19.1 GB peak) using the ESM SDK 3.1.1 with PyTorch 2.10.0+cu128 on CUDA 12.8. Two replication runs repeated the identical configuration under seeds 20260730 and 20270731 (checkpoint fingerprints in Appendix A.12). The matched generation seed slots characterize sampling variation for a fixed checkpoint; the additional training runs assess variation across training seeds.

PoET-2 training used seed 20260825 with LoRA rank 8, $\alpha = 16$, dropout 0, and 48 attention-only adapter modules (query and value projections of causal self- and cross-attention across 12 decoder layers; 786,432 trainable parameters); optimization used AdamW (learning rate 1×10^{-4}, betas $(0.9, 0.999)$, $\epsilon = 10^{-8}$, weight decay 0.01, gradient clipping 1.0) with cosine decay to 10% of peak after 5% warmup, an effective pair batch size of 8 with micro-batch 1, three epochs (600 optimizer steps), a 12,288 context-token budget, and TF32; initialization policy–reference parity used four examples with tolerance 0.1. Training ran in about 12 minutes on one NVIDIA RTX 4090. The policy uses the `epoch-003` checkpoint saved after the third epoch, starting from the base checkpoint `poet-2.ckpt`. Supplementary PoET-2 runs used seeds 20260826 and 20260827 and the same checkpoint-selection rule. Each run's regeneration endpoints were computed from outputs generated with its own `epoch-003` checkpoint.

A.3 Matched native-regeneration protocols

ESM3. The base and preference-fine-tuned policies were evaluated on 200 homology-held-out prompts. Each policy produced 16 raw samples per prompt using the same prompt order, batch composition, deterministic random stream, 20-step iterative unmasking schedule, temperature 1.0, and top-p 1.0. The 16 draws comprised four matched prompt-specific generation seed slots, with four draws per slot.

PoET-2. The base and preference-fine-tuned policies used the same family memory, 20-residue fixed prefix and target length. The generator recomputes the full prefix at each step. The frozen generation manifest records temperature 1.0, top-p 1.0, no top-k cutoff, repetition penalty 1.4, and a deterministic seed equal to the existing task-generation seed multiplied by 1009 plus the zero-based sample index. The evaluation used 200 held-out source contexts and 16 attempted draws per policy and context, yielding 3,200 realized draws per policy and zero failed generations.

For both architectures, formal regeneration retained every attempted draw and applied no best-of-n selection, evaluator-guided reranking, rejection sampling, sequence-selection heuristic, or replacement draw.

A.4 Evaluators and statistical analysis

VirulentHunter defined preference labels and remained a diagnostic evaluator. DTVF, run locally using ProtT5 embeddings and released weights, served as the principal external evaluator of computational VF association. The second classifier criterion used the official VirulentPred 2.0 `predict.pl` workflow: PSI-BLAST generated PSSM features against the bundled UniProt Swiss-Prot database, after which the supplied classifier produced a binary virulent/nonvirulent label. This workflow excluded protein-language-model embeddings but retained homology-derived evolutionary-profile features. Q_{profile} was the principal profile-compatibility summary. HMMER was a secondary diagnostic of source-family compatibility. Validity, exact-sequence uniqueness, length fidelity, context fidelity, and PoET-2 suffix negative log-likelihood were additional quality diagnostics. PoET-2 VirulentHunter scoring used a frozen fast-inference implementation with length-sorted batches of eight, automatic bfloat16 CUDA precision, and a requested token limit of 1,200 clamped to 1,026; binary softmax probabilities were rounded to three decimals. Replaying the complete base and fine-tuned cohorts under this configuration reproduced all VF probabilities exactly, whereas alternative precision or batching configurations did not, so the recorded configuration was used for all runs, including the replication seeds, without changes to weights, batching rules, truncation, or postprocessing.

DTVF scoring used ProtT5 embeddings (`Rostlab/prot_t5_xl_bfd`) with the official DTVF checkpoint `best_model.pth` (fingerprint in Appendix A.12), computing the class-1 probability as $\sigma(\text{logit}_1)$ with decision threshold 0.5; embeddings were extracted in batches of eight with at most 3,000 residues per pass, and every attempted draw received a score, so no failure handling was invoked. For VirulentPred 2.0, the official standalone release scripts were used with the bundled PSI-BLAST 2.2.26, the supplied AutoGluon 0.4.0 PSSM classifier, and the bundled UniProt Swiss-Prot 2022_01 database; inputs were restricted to valid sequences containing only the 20 standard amino acids from bacterial-scope sources, the supplied classifier returned the binary virulent/nonvirulent label, and all staged eligible inputs received predictions. ESM3 evaluations used eight bounded workers for the replication-seed cohort and four for the primary cohort; each supplementary PoET-2 evaluation used four. Official output IDs were normalized only by removing their FASTA marker and surrounding whitespace, then checked against exact staged IDs and sequence hashes. All completed PSSM matrices had 400 finite features per eligible sequence; no incomplete worker output was included. Release provenance was pinned as follows. The DTVF repository (`https://github.com/niuwa2333/DTVF`) provides no tagged releases or version identifiers; the `best_model.pth` checkpoint was uploaded once, in the 28 July 2024 commit `3356b008a77f3497b558981094cf06b2cef73b38`, and was unchanged through the repository snapshot used here, which corresponds to the final 9 September 2024 commit `4299f8e12de3e0219ad0d01287079b67ae24dd60`. The VirulentPred 2.0 standalone was distributed only as `virulentpred_2_0.tar.gz` from the authors' download page (`https://bioinfo.icgeb.res.in/virulent2/down.html`); the distribution carries no version-controlled commit identifier or release tag, so the original release package, execution scripts, input manifests, and completed outputs were retained to support reproducibility. No supported sequence-length range is documented for VirulentPred 2.0: the official pages require FASTA-format input of single or multiple sequences and recommend at most ten sequences per web submission for speed, and the standalone accepted every staged eligible input as submitted.

Training-data provenance follows the original publications, and neither classifier was retrained, recalibrated, or supplemented with study sequences, so all external scores reflect the authors' released training partitions. DTVF (Deep Transfer Learning for Virulence Factor prediction; Sun et al., 2024; `https://doi.org/10.3390/genes15091170`) was trained by its authors on the benchmark assembled for DeepVF (Xie et al., 2021): 9,749 virulence factors compiled from the Victors, VFDB, and PATRIC databases and 66,982 non-VF negatives selected from VFDB by the DeepVF negative-data procurement methodology, reduced by CD-HIT clustering at a 0.3 sequence-identity threshold to a non-redundant corpus of 3,576 VFs and 4,910 non-VFs (8,486 sequences in total), partitioned into training and independent test sets following the original benchmark and reported at 84.55% accuracy and 92.08% AUROC; the model consumes frozen ProtT5-XL-BFD per-residue embeddings, matching the `Rostlab/prot_t5_xl_bfd` extraction used here. VirulentPred 2.0 (Sharma et al., 2023; `https://doi.org/10.1002/pro.4808`) was trained on 3,580 virulent proteins retrieved from the VFDB core dataset and 2,852 retrieved from UniProt using virulence-related keywords (virulence, adhesin, toxin, invasion, capsule, and related terms), from which entries annotated as probable, putative, similarity, fragment, hypothetical, unknown, or possible were removed so that only experimentally verified virulent proteins remained; the negative set comprised 3,406 annotated bacterial enzyme sequences from UniProt, drawn mainly from the same bacterial proteomes as the positives and randomly selected to balance them. CD-HIT redundancy reduction at a 0.5 identity cut-off first reduced the positive pool (6,432 to 3,403 sequences) and, after joint positive–negative comparison, yielded the final balanced corpus of 6,781 sequences (3,375 virulent and 3,406 nonvirulent) spanning 32 bacterial genera, randomly divided 80/20 into 5,425 training and 1,356 test sequences with an AutoGluon-internal 0.9/0.1 train/validation split of the training portion; a PSI-BLAST audit reported sequence identity $\geq$50% to some training sequence for only 6.05% (82 of 1,356) of the test set. The deployed PSSM model is a WeightedEnsemble_L2 stack of 14 AutoGluon algorithms trained on 400-column PSSM features generated by PSI-BLAST against the Swiss-Prot database at an E-value cut-off of 0.001, with reported validation accuracy 84.71%, test accuracy 85.18%, MCC 0.70, and AUC 0.924; the same PSI-BLAST-against-Swiss-Prot workflow, using the bundled Swiss-Prot 2022_01 database, generated the features for our staged inputs.

The source prompt was the independent statistical unit, and the 16 sequences generated for a policy–prompt combination were subsamples summarized within prompt. For classifier-based regeneration endpoints, draw-level probabilities were averaged within prompt before paired base–fine-tuned effects were calculated. The all-attempted ESM3 classifier analysis retained the observed evaluator

Table 2: ESM3 paired classifier effects by matched generation seed slot. Effects are fine-tuned minus base changes in prompt-level mean probability. S1–S4 contain four draws per prompt and slot; Joint uses all 16 draws. Direction counts are lower/higher/tied prompts.

Endpoint	Slot	Paired change	95% CI	Direction
R_{DTVF}	S1	-0.1531	$[-0.2577, -0.0717]$	159/41/0
R_{DTVF}	S2	-0.1567	$[-0.2593, -0.0811]$	163/37/0
R_{DTVF}	S3	-0.1500	$[-0.2555, -0.0685]$	161/39/0
R_{DTVF}	S4	-0.1364	$[-0.2423, -0.0575]$	154/46/0
R_{DTVF}	Joint	-0.1491	$[-0.2531, -0.0705]$	169/31/0
VirulentHunter	S1	-0.2261	$[-0.3043, -0.1615]$	172/28/0
VirulentHunter	S2	-0.2182	$[-0.2965, -0.1535]$	167/32/1
VirulentHunter	S3	-0.2314	$[-0.3062, -0.1702]$	173/26/1
VirulentHunter	S4	-0.2320	$[-0.3043, -0.1742]$	180/20/0
VirulentHunter	Joint	-0.2269	$[-0.3012, -0.1666]$	183/17/0

score for every attempted draw, including outputs marked invalid. For Q_{profile}, draws with scores for both policies were paired within prompt, and prompt means were calculated over common score-complete draws; invalid-output counts were reported through the validity endpoint.

Homology components were the resampling clusters. Paired effects were reported with 95% component-cluster bootstrap confidence intervals from 10,000 resamples using seed 20260917. R_{DTVF} and Q_{profile} were the principal regeneration summaries. R_{VP2}, VirulentHunter, validity, exact-sequence uniqueness, length and context fidelity, and HMMER were descriptive secondary diagnostics. The VirulentPred 2.0 analysis was restricted to supported-input, score-complete raw generations. Eligibility and exclusion counts were reported separately for each policy and architecture. Confidence intervals quantify uncertainty for the reported effects but were not used for multiplicity-adjusted confirmatory decisions. The main regeneration analyses report effect estimates, confidence intervals, denominators, invalid-output counts, and missing-score counts; generated sequences were not treated as independent replicates.

A.5 Complete classifier-based regeneration results

ESM3 DTVF distribution and sensitivity analyses. For the base policy, DTVF probability had a mean (R_{DTVF}) of 0.6656 (SD 0.3220), a median of 0.7562, and an interquartile range of 0.3975–0.9779. Of 3,200 attempted draws from 200 held-out prompts, 2,184 (68.25%) had DTVF probability ≥ 0.5; 67 outputs were marked invalid, and no DTVF score was missing. For the preference-fine-tuned policy, the mean was 0.5166 (SD 0.3965), the median was 0.4962, and the interquartile range was 0.0890–0.9536. Of 3,200 attempted draws, 1,595 (49.84%) had probability ≥ 0.5; 25 outputs were marked invalid, and no DTVF score was missing.

The mean paired change in prompt-level R_{DTVF} was -0.1491 (95% component-cluster bootstrap CI, -0.2531 to -0.0705). Prompt-level R_{DTVF} was lower after fine-tuning for 169 of 200 prompts and higher for 31. The positive fraction decreased by 18.41 percentage points; the paired prompt-level change was -0.1841 (95% CI, -0.3301 to -0.0607). The median prompt-level change in R_{DTVF} was -0.0698 (95% CI, -0.1425 to -0.0204). Restricting each prompt to valid outputs gave a mean paired change of -0.1502 (95% CI, -0.2550 to -0.0715).

ESM3 VirulentHunter diagnostic. Mean predicted VF probability decreased from 0.6039 for the base policy to 0.3770 after fine-tuning, and the fraction with probability ≥ 0.5 decreased from 63.75% (2,040/3,200) to 34.63% (1,108/3,200). The mean paired change in prompt-level probability was -0.2269 (95% CI, -0.3012 to -0.1666), with lower values for 183 of 200 prompts and higher values for 17. Changes in R_{DTVF} and VirulentHunter scores agreed in direction for 166 prompts (83.0%): both decreased for 159, both did not decrease for 7, and 34 were discordant. Prompt-effect rank correlation was moderate ($\rho = 0.347$), whereas sequence-level correlations were weak (base $\rho = 0.032$; fine-tuned $\rho = 0.271$). VirulentHunter therefore provided policy-level directional corroboration but was not interchangeable with DTVF at the individual-sequence level.

Table 3: ESM3 paired classifier effects for the two training-seed replication runs. Effects are fine-tuned minus base changes in prompt-level mean probability. S1–S4 contain four draws per prompt and slot; Joint uses all 16 draws. Direction counts are lower/higher/tied prompts.

Endpoint	Slot	Paired change	95% CI	Direction
R_{DTVF} (seed 20260730)	S1	-0.1452	$[-0.2508, -0.0648]$	157/43/0
R_{DTVF} (seed 20260730)	S2	-0.1543	$[-0.2554, -0.0807]$	165/35/0
R_{DTVF} (seed 20260730)	S3	-0.1276	$[-0.2340, -0.0474]$	151/49/0
R_{DTVF} (seed 20260730)	S4	-0.1406	$[-0.2425, -0.0661]$	161/39/0
R_{DTVF} (seed 20260730)	Joint	-0.1419	$[-0.2453, -0.0655]$	168/32/0
R_{DTVF} (seed 20270731)	S1	-0.1440	$[-0.2488, -0.0630]$	154/46/0
R_{DTVF} (seed 20270731)	S2	-0.1569	$[-0.2568, -0.0846]$	162/38/0
R_{DTVF} (seed 20270731)	S3	-0.1365	$[-0.2424, -0.0579]$	155/45/0
R_{DTVF} (seed 20270731)	S4	-0.1348	$[-0.2396, -0.0576]$	155/45/0
R_{DTVF} (seed 20270731)	Joint	-0.1430	$[-0.2463, -0.0672]$	164/36/0
VirulentHunter (seed 20260730)	S1	-0.2042	$[-0.2845, -0.1360]$	162/37/1
VirulentHunter (seed 20260730)	S2	-0.1999	$[-0.2758, -0.1378]$	159/41/0
VirulentHunter (seed 20260730)	S3	-0.2174	$[-0.2869, -0.1656]$	164/36/0
VirulentHunter (seed 20260730)	S4	-0.2226	$[-0.2931, -0.1664]$	175/25/0
VirulentHunter (seed 20260730)	Joint	-0.2110	$[-0.2842, -0.1535]$	182/18/0
VirulentHunter (seed 20270731)	S1	-0.2268	$[-0.3050, -0.1611]$	169/29/2
VirulentHunter (seed 20270731)	S2	-0.2082	$[-0.2850, -0.1456]$	157/43/0
VirulentHunter (seed 20270731)	S3	-0.2061	$[-0.2770, -0.1502]$	163/37/0
VirulentHunter (seed 20270731)	S4	-0.2129	$[-0.2812, -0.1609]$	173/27/0
VirulentHunter (seed 20270731)	Joint	-0.2135	$[-0.2850, -0.1556]$	180/20/0

ESM3 training-seed replication runs. Two replication runs repeated the identical training and regeneration protocol under seeds 20260730 and 20270731; the base policy is shared, so base-policy statistics are identical. For seed 20260730, fine-tuned DTVF probability had a mean (R_{DTVF}) of 0.5237 (SD 0.3910), a median of 0.5235, and an interquartile range of 0.1068–0.9527; 1,621 of 3,200 draws (50.66%) had probability ≥ 0.5, 40 outputs were marked invalid, and no DTVF score was missing. The mean paired prompt-level change was -0.1419 (95% CI, -0.2453 to -0.0655), with lower values for 168 prompts and higher values for 32; the median change was -0.0643, and restricting each prompt to valid outputs gave -0.1426 (95% CI, -0.2468 to -0.0659). VirulentHunter decreased from 0.6039 to 0.3929, with 1,161 draws (36.28%) at probability ≥ 0.5 and a paired change of -0.2110 (95% CI, -0.2842 to -0.1535; 182 lower, 18 higher). For seed 20270731, fine-tuned DTVF probability had a mean (R_{DTVF}) of 0.5226 (SD 0.3918), a median of 0.5169, and an interquartile range of 0.1025–0.9514; 1,619 of 3,200 draws (50.59%) had probability ≥ 0.5, 25 outputs were marked invalid, and no DTVF score was missing. The mean paired change was -0.1430 (95% CI, -0.2463 to -0.0672), with lower values for 164 prompts and higher values for 36; the median change was -0.0672, and the valid-output restriction gave -0.1441 (95% CI, -0.2482 to -0.0678). VirulentHunter decreased to 0.3905, with 1,153 draws (36.03%) at probability ≥ 0.5 and a paired change of -0.2135 (95% CI, -0.2850 to -0.1556; 180 lower, 20 higher). Validity was 98.75% (3,160/3,200) and 99.22% (3,175/3,200), with no exact within-policy duplicates in either run. Supplementary Q_{profile} and HMMER results for both replication seeds are reported in Appendix A.7. Table 3 reports seed-slot effects.

PoET-2 classifier results. For the primary cohort, base-policy DTVF probability had a mean (R_{DTVF}) of 0.8752 (SD 0.1976), median 0.9638, and interquartile range 0.8627–0.9905. Of 3,200 draws, 2,953 (92.28%) had probability ≥ 0.5. For the fine-tuned policy, the mean was 0.7351 (SD 0.2851), median 0.8639, and interquartile range 0.5560–0.9681; 2,481 draws (77.53%) had probability ≥ 0.5. No DTVF score was missing. The mean paired prompt-level change was -0.1402 (95% CI, -0.1793 to -0.0955), with 181 lower and 19 higher prompts among 200 prompts in 66 homology components. The positive fraction decreased by 14.75 percentage points (paired change -0.1475; 95% CI, -0.2113 to -0.0800).

Table 4: Frozen preference-pair point estimates for the primary training runs. Policy margin denotes the policy's pairwise score margin (ESM3 scaled; PoET-2 raw sum), rather than the policy-minus-reference margin. Confidence intervals and replication results are reported below.

Model	Split	DPO loss ↓	Policy margin ↑	Chosen acc. ↑	Implicit acc. ↑
ESM3	Validation	0.6415	12.144	59.5%	66.5%
ESM3	Holdout	0.6480	11.717	58.5%	65.5%
PoET-2	Validation	0.6025	−3.535	50.5%	72.0%
PoET-2	Holdout	0.7729	−7.292	45.0%	64.5%

On the same cohort, mean VirulentHunter probability decreased from 0.5514 to 0.4086, and the fraction with probability ≥ 0.5 decreased from 56.25% (1,800/3,200) to 39.25% (1,256/3,200). The paired change was -0.1428 (95% CI, -0.1871 to -0.1021), with 159 lower and 41 higher prompts. Changes in R_{DTVF} and VirulentHunter scores agreed on reduction versus nonreduction for 156/200 prompts (78.0%): both decreased for 148 and both did not decrease for 8. Prompt-effect Spearman correlation was $\rho = 0.162$. All 137 VirulentPred 2.0 supported prompts had DTVF and VirulentHunter results.

A.6 Complete frozen-pair preference evaluation

We evaluated the frozen preference pairs using the architecture-specific training scores. These analyses assess preference learning under the training surrogate, not native generation-trajectory likelihoods. All 95% confidence intervals were estimated from per-pair records using 10,000 homology-component cluster bootstrap resamples (seed 20260917).

ESM3 seed 20260729, validation (200 pairs; 76 components): DPO loss 0.6415 (95% CI, 0.5935 to 0.6914), policy margin 12.144 (95% CI, 2.935 to 21.791), reference margin 3.641 (95% CI, −2.035 to 9.519), policy-minus-reference margin 8.504 (95% CI, 2.998 to 14.286), chosen accuracy 59.5% (95% CI, 51.0 to 68.3%), and implicit preference accuracy 66.5% (95% CI, 56.2 to 76.0%).

ESM3 seed 20260729, holdout (200 pairs; 68 components): DPO loss 0.6480 (95% CI, 0.5978 to 0.7150), policy margin 11.717 (95% CI, −1.137 to 22.677), reference margin 1.580 (95% CI, −7.044 to 8.445), policy-minus-reference margin 10.137 (95% CI, 3.766 to 15.563), chosen accuracy 58.5% (95% CI, 49.4 to 67.5%), and implicit preference accuracy 65.5% (95% CI, 57.3 to 73.1%).

ESM3 seed 20260730, validation (200 pairs; 76 components): DPO loss 0.6474 (95% CI, 0.6032 to 0.6929), policy margin 11.092 (95% CI, 2.424 to 20.209), reference margin 3.641 (95% CI, −2.035 to 9.519), policy-minus-reference margin 7.451 (95% CI, 2.465 to 12.670), chosen accuracy 58.5% (95% CI, 50.0 to 67.2%), and implicit preference accuracy 63.5% (95% CI, 52.6 to 73.8%).

ESM3 seed 20260730, holdout (200 pairs; 68 components): DPO loss 0.6468 (95% CI, 0.5995 to 0.7098), policy margin 11.074 (95% CI, −1.485 to 21.795), reference margin 1.580 (95% CI, −7.044 to 8.445), policy-minus-reference margin 9.494 (95% CI, 3.514 to 14.649), chosen accuracy 57.5% (95% CI, 48.2 to 66.6%), and implicit preference accuracy 66.0% (95% CI, 58.0 to 73.4%).

ESM3 seed 20270731, validation (200 pairs; 76 components): DPO loss 0.6483 (95% CI, 0.6018 to 0.6963), policy margin 11.196 (95% CI, 2.299 to 20.427), reference margin 3.641 (95% CI, −2.035 to 9.519), policy-minus-reference margin 7.555 (95% CI, 2.296 to 12.953), chosen accuracy 60.0% (95% CI, 51.6 to 68.9%), and implicit preference accuracy 64.0% (95% CI, 52.0 to 74.7%).

ESM3 seed 20270731, holdout (200 pairs; 68 components): DPO loss 0.6510 (95% CI, 0.6027 to 0.7156), policy margin 11.041 (95% CI, −1.544 to 21.810), reference margin 1.580 (95% CI, −7.044 to 8.445), policy-minus-reference margin 9.460 (95% CI, 3.362 to 14.701), chosen accuracy 58.0% (95% CI, 48.8 to 67.1%), and implicit preference accuracy 65.5% (95% CI, 57.9 to 72.6%).

PoET-2 (seed 20260825) validation (200 pairs; 53 components): DPO loss 0.6025 (95% CI, 0.4942 to 0.7328), policy margin −3.535 (95% CI, −12.694 to 5.444), reference margin −16.065 (95% CI, −26.560 to −6.457), policy-minus-reference margin 12.530 (95% CI, 8.470 to 16.446), chosen

accuracy 50.5% (95% CI, 44.2 to 56.7%), implicit preference accuracy 72.0% (95% CI, 65.1 to 77.7%).

PoET-2 (seed 20260825) holdout (200 pairs; 66 components): DPO loss 0.7729 (95% CI, 0.6237 to 0.9234), policy margin −7.292 (95% CI, −19.777 to 4.779), reference margin −15.542 (95% CI, −26.843 to −5.186), policy-minus-reference margin 8.250 (95% CI, 4.930 to 12.290), chosen accuracy 45.0% (95% CI, 38.2 to 51.6%), implicit preference accuracy 64.5% (95% CI, 58.7 to 70.7%).

PoET-2 training-seed replications. PoET-2 seed 20260826, epoch-003, validation (200 pairs; 53 components): DPO loss 0.6034 (95% CI, 0.4901 to 0.7375), policy-minus-reference margin 12.799 (95% CI, 8.765 to 16.756), absolute policy margin −3.266 (95% CI, −12.454 to 5.774), reference margin −16.065 (95% CI, −26.560 to −6.457), chosen accuracy 51.0% (95% CI, 43.8 to 57.9%), and implicit preference accuracy 72.0% (95% CI, 65.3 to 77.5%).

PoET-2 seed 20260826, epoch-003, holdout (200 pairs; 66 components): DPO loss 0.7562 (95% CI, 0.6033 to 0.9046), policy-minus-reference margin 8.348 (95% CI, 5.086 to 12.406), absolute policy margin −7.194 (95% CI, −19.747 to 5.058), reference margin −15.542 (95% CI, −26.843 to −5.186), chosen accuracy 45.0% (95% CI, 38.4 to 51.3%), and implicit preference accuracy 67.5% (95% CI, 61.0 to 74.5%).

PoET-2 seed 20260827, epoch-003, validation (200 pairs; 53 components): DPO loss 0.6204 (95% CI, 0.4957 to 0.7724), policy-minus-reference margin 12.323 (95% CI, 8.334 to 16.080), absolute policy margin −3.742 (95% CI, −13.005 to 5.423), reference margin −16.065 (95% CI, −26.560 to −6.457), chosen accuracy 49.0% (95% CI, 41.6 to 56.0%), and implicit preference accuracy 72.5% (95% CI, 65.6 to 78.5%).

PoET-2 seed 20260827, epoch-003, holdout (200 pairs; 66 components): DPO loss 0.7873 (95% CI, 0.6328 to 0.9426), policy-minus-reference margin 7.990 (95% CI, 4.744 to 11.872), absolute policy margin −7.553 (95% CI, −20.135 to 4.524), reference margin −15.542 (95% CI, −26.843 to −5.186), chosen accuracy 43.5% (95% CI, 36.8 to 50.0%), and implicit preference accuracy 65.5% (95% CI, 59.9 to 72.0%).

A.7 Complete generation-quality diagnostics

For ESM3, Q_{profile} was available for 3,116 matched score-complete base–fine-tuned sequence pairs across 200 prompts and 68 homology components; each prompt contributed 10–16 matched pairs. Mean prompt-level Q_{profile} was 47.4648 for the base policy and 48.4320 after fine-tuning. The paired change was +0.9672 (95% CI, +0.2521 to +1.7298), with higher values for 140 prompts and lower values for 60. The median prompt-level change was +0.7672 (95% CI, +0.2077 to +1.3513).

ESM3 validity increased from 97.91% (3,133/3,200) under the base policy to 99.22% (3,175/3,200) after fine-tuning. The paired prompt-level change was +0.0131 (95% CI, +0.00759 to +0.02027). All 3,200 attempted sequences were exact non-duplicates within each policy, and all valid sequences were also exact non-duplicates. Under both policies, all attempted draws matched the requested sequence length, preserved every fixed prompt position, and filled every masked position. Exact-sequence uniqueness establishes only the absence of exact duplicates within each policy. These secondary intervals are descriptive and were not used for multiplicity-adjusted confirmatory decisions.

At the prompt level, R_{DTVF} decreased for 169 of 200 prompts, whereas Q_{profile} increased for 140. We did not define a binary prompt-level joint-success endpoint; the joint interpretation is based on the paired population-level estimates.

For PoET-2, all 3,200 matched pairs from 200 prompts in 66 homology components received Q_{profile} scores. Base and fine-tuned prompt means were 29.4879 and 40.0968; the paired change was +10.6089 (95% CI, +8.2298 to +12.9606), with 192 higher and 8 lower prompts. The recomputation used the same profile scorer ($\alpha = 1$, minimum column coverage 0.5) on the hash-matched input cohort.

Both policies had 3,200/3,200 valid outputs and 3,200/3,200 exact unique sequences. Matched output lengths agreed for every draw. All 6,400 sequences also matched the original task lengths and preserved the 20-residue prefixes. A supplementary frozen-base-model diagnostic computed teacher-forced NLL over residues 21 through L, excluding the prefix and EOS without applying sampling penalties. Mean prompt-level NLL was 2.4884 for base outputs and 1.8491 for fine-tuned outputs; the paired change was -0.6393 (95% CI, -0.7298 to -0.5390). This measures compatibility under one fixed reference model, not native generation-trajectory probability or experimentally measured fitness.

ESM3 supplementary training-seed quality evaluations. ESM3 seed 20260730: Q_{profile} used 3,107 matched score-complete pairs from 200 prompts (68 components), with prompt means 47.4599 (base) and 48.3242 (adapter), and paired change 0.8642 (95% CI, 0.1314 to 1.6120). Values increased for 133 prompts. HMMER had 2,097 both-hit pairs in 161 prompts (57 components). The median prompt-level normalized-score change was 0.02867 (95% CI, 0.01209 to 0.05370); 130/29/2 prompts increased/decreased/tied. Base and adapter reportable hits were 2,213/3,107 (71.23%) and 2,333/3,107 (75.09%). The paired prompt-level hit-rate change was 0.03829 (95% CI, 0.01060 to 0.07013).

ESM3 seed 20270731: Q_{profile} used 3,116 matched score-complete pairs from 200 prompts (68 components), with prompt means 47.4588 (base) and 48.3268 (adapter), and paired change 0.8680 (95% CI, 0.1674 to 1.5907). Values increased for 132 prompts. HMMER had 2,116 both-hit pairs in 164 prompts (56 components). The median prompt-level normalized-score change was 0.02539 (95% CI, 0.01029 to 0.05403); 132/30/2 prompts increased/decreased/tied. Base and adapter reportable hits were 2,220/3,116 (71.25%) and 2,347/3,116 (75.32%). The paired prompt-level hit-rate change was 0.04056 (95% CI, 0.01186 to 0.07848).

Completed supplementary PoET-2 regeneration diagnostics. PoET-2 generation-quality seed 20260826, epoch-003: all 3,200 draws passed alphabet, length, fixed-prefix, and source-metadata checks, and 3,200/3,200 were exact-sequence unique. Mean prompt-level Q_{profile} was 29.4879 for base and 37.4545 after fine-tuning; the paired change was 7.9666 (95% CI, 5.9459 to 9.9786). HMMER returned reportable hits for 2,990/3,200 (93.44%) base and 3,043/3,200 (95.09%) fine-tuned draws. The paired hit-rate change was 0.01656 (95% CI, 0.00411 to 0.03092). Among 2,891 matched draws with hits under both policies (200 sources, 66 components), the median prompt-level change in normalized full-sequence score was 0.32446 (95% CI, 0.24449 to 0.38176), and the corresponding mean was 0.36152 (95% CI, 0.29040 to 0.43081). Shared-base HMMER scores exactly reproduced the original evaluation. All intervals used 10,000 homology-component bootstrap resamples.

PoET-2 supplementary VirulentHunter seed 20260826: mean probability 0.5514 for base and 0.3763 after fine-tuning; paired prompt-level change -0.1752 (95% CI, -0.2153 to -0.1381). Positive counts were 1,800/3,200 and 1,127/3,200, respectively; all 200 sources and 66 homology components were retained.

PoET-2 generation-quality seed 20260827, epoch-003: all 3,200 draws passed alphabet, length, fixed-prefix, and source-metadata checks, and 3,200/3,200 were exact-sequence unique. Mean prompt-level Q_{profile} was 29.4879 for base and 38.2993 after fine-tuning; the paired change was 8.8114 (95% CI, 6.6016 to 11.0331). HMMER returned reportable hits for 2,990/3,200 (93.44%) base and 3,064/3,200 (95.75%) fine-tuned draws. The paired hit-rate change was 0.02312 (95% CI, 0.00987 to 0.03830). Among 2,904 matched draws with hits under both policies (200 sources, 66 components), the median prompt-level change in normalized full-sequence score was 0.34112 (95% CI, 0.27459 to 0.45160), and the corresponding mean was 0.40314 (95% CI, 0.32257 to 0.48336). Shared-base HMMER scores exactly reproduced the original evaluation. All intervals used 10,000 homology-component bootstrap resamples.

PoET-2 supplementary VirulentHunter seed 20260827: mean probability 0.5514 for base and 0.3723 after fine-tuning; paired prompt-level change -0.1791 (95% CI, -0.2230 to -0.1392). Positive counts were 1,800/3,200 and 1,101/3,200, respectively; all 200 sources and 66 homology components were retained.

PoET-2 supplementary suffix-NLL seed 20260826, epoch-003: all 3,200 draws passed source, sequence-hash, length and prefix checks. Mean prompt-level frozen-base-reference NLL was

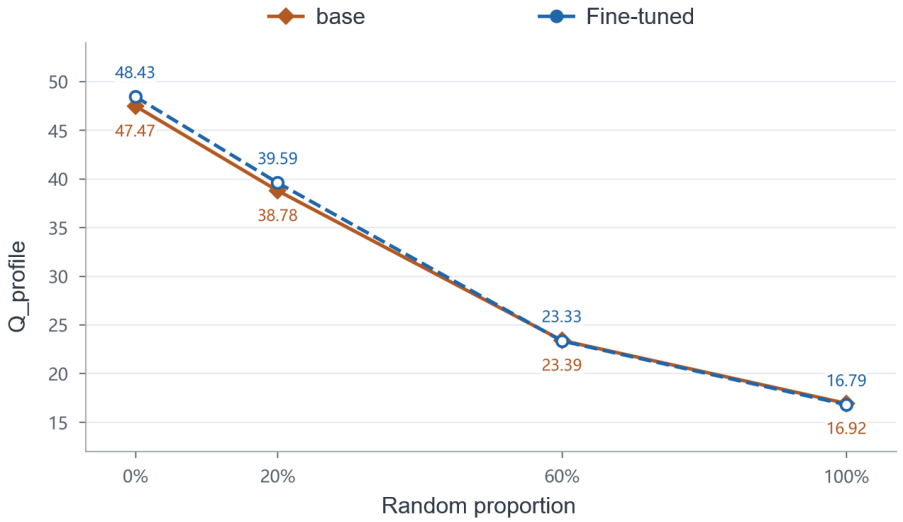


Figure 6: Q_{profile} **under conserved-residue context randomization.** Source-averaged Q_{profile} is shown for the base ESM3 model and the fine-tuned model (training seed 20260729) at the four randomization ratios used in Figure 4. Points are descriptive estimates without confidence intervals.

Table 5: Numerical summaries of the ESM3 context-randomization ablation. Base and fine-tuned columns refer to the base model and training seed 20260729, respectively. VF-positive fractions are percentages; Q_{profile} is on its 0–100 scale. Values are reported at the precision of the ablation summaries and are descriptive point estimates.

	Mean `vf_prob`		VF-positive (%)		Q_{profile}	
Randomized	Base	Fine-tuned	Base	Fine-tuned	Base	Fine-tuned
0%	0.604	0.377	63.8	34.6	47.47	48.43
20%	0.518	0.351	51.9	30.6	38.78	39.59
60%	0.308	0.252	26.2	19.7	23.39	23.33
100%	0.183	0.151	12.5	9.5	16.92	16.79

1.9817, compared with 2.4884 for the shared base draws; paired change -0.5067 (95% CI, -0.5819 to -0.4253). This uses the same fixed-reference, residues-21-through-L diagnostic as the primary analysis, excluding EOS.

PoET-2 supplementary suffix-NLL seed 20260827, epoch-003: all 3,200 draws passed source, sequence-hash, length and prefix checks. Mean prompt-level frozen-base-reference NLL was 1.9796, compared with 2.4884 for the shared base draws; paired change -0.5087 (95% CI, -0.5974 to -0.4144). This uses the same fixed-reference, residues-21-through-L diagnostic as the primary analysis, excluding EOS.

A.8 Conserved-residue context randomization in ESM3

Evaluation setting. This analysis extends the context ablation in Section 4.5 by reporting the accompanying Q_{profile} scores and the complete numerical summaries. We varied the fraction of conserved prompt positions randomized at generation time over 0%, 20%, 60%, and 100%, and compared the base ESM3 model with the fine-tuned model from training seed 20260729. The 0% condition retained the original conserved-residue context; 100% denotes randomization of all conserved positions included in this perturbation, rather than unconditional generation. VirulentHunter provided the diagnostic classifier summaries, while Q_{profile}, defined in Eq. (3), summarized residue-profile compatibility. The results below are descriptive point estimates for this training run; no confidence intervals or between-seed uncertainty estimates are reported for this ablation.

Q_{profile} **results.** As randomization increased from 0% to 100%, Q_{profile} decreased from 47.47 to 16.92 for the base model and from 48.43 to 16.79 for the fine-tuned model (Figure 6). At 20% randomization, the scores were 38.78 and 39.59, respectively; at 60%, they were 23.39 and 23.33. The fine-tuned model therefore had slightly higher point estimates at 0% and 20% and slightly lower estimates at 60% and 100%. These between-policy differences were small compared with the decline across randomization ratios. The accompanying VF-positive fractions decreased after fine-tuning at all four ratios, from 63.8% to 34.6%, 51.9% to 30.6%, 26.2% to 19.7%, and 12.5% to 9.5%, respectively (Table 5).

Interpretation. The within-ratio comparison measures the observed difference between the two policies at a specified level of context randomization, whereas the across-ratio comparison captures changes accompanying the perturbation of the context itself. Both VF-associated predictions and profile compatibility declined as randomization increased. Consequently, the across-ratio decline in classifier scores should not be interpreted as a risk reduction achieved at unchanged profile compatibility. Within each tested ratio, fine-tuning was associated with lower VF-associated predictions and comparatively small differences in Q_{profile}. These observations do not establish statistical equivalence of quality scores, biological safety, or replication across training seeds.

A.9 Complete VirulentPred 2.0 and HMMER analyses

VirulentPred 2.0 assessed VF-associated predictions alongside the results in Section 4.2, whereas HMMER assessed source-family compatibility.

PSSM-based VirulentPred 2.0 analysis. The frozen classifier was applied after generation and only to raw held-out sequences within its prespecified supported bacterial input scope, using sequences composed of the 20 standard amino acids that received a valid prediction. Its outputs did not inform model selection, decoding, filtering, reranking, rejection sampling, or replacement draws. The official `predict.pl` workflow used PSI-BLAST against the bundled UniProt Swiss-Prot database to construct PSSM features and then applied the supplied classifier. These features encode homology-derived evolutionary information.

Sequence-hash audits verified that each fine-tuned VirulentPred 2.0 cohort originated from the corresponding training seed's regeneration outputs and that the shared base cohort matched the recorded base-generation run. Within each architecture, all training-seed contrasts used the same base draws.

The main analysis computed each policy's prompt-level R_{VP2} using all eligible draws for that policy and then estimated the mean paired change across common source prompts. The pooled R_{VP2} values reported in Table 6 are descriptive summaries and should not be subtracted to reconstruct the prompt-level paired effect. A separate common-score-complete sensitivity analysis restricted both policies to identical source/draw keys with valid predictions under both policies. These two analyses can yield different effects when eligibility differs between policies.

All R_{VP2} confidence intervals used 10,000 homology-component bootstrap resamples. They describe uncertainty within each training run, not between-seed uncertainty; uncertainty was not pooled across training seeds.

ESM3 eligibility and matched-draw sensitivity analyses. Across all three ESM3 training-seed contrasts, the shared base cohort contained 2,011 eligible sequences out of 3,200 attempted draws; 1,189 draws were excluded by the frozen scope/validity rules. The main paired analyses covered 129 common source prompts in 47 homology components. Every eligible input received a label.

For the primary training seed 20260729, 2,046 fine-tuned draws were eligible and 1,154 were excluded. Using all eligible draws within each policy, prompt-level R_{VP2} decreased for 71 prompts, increased for 22, and were unchanged for 36. Restricting both policies to the 1,998 identical source/draw keys gave 1,484/1,998 base and 1,134/1,998 fine-tuned virulent labels. The mean paired prompt-level change in this strict matched-draw sensitivity analysis was -0.1749 (95% CI, -0.3165 to -0.0609).

For replication training seed 20260730, 2,034 fine-tuned draws were eligible and 1,166 were excluded. Using all eligible draws within each policy, prompt-level R_{VP2} decreased for 63 prompts, increased for 24, and were unchanged for 42. Restricting both policies to the 1,989 identical

Table 6: Complete R_{VP2} results across training seeds. Base and fine-tuned columns report virulent predictions divided by eligible sequences, with pooled R_{VP2} percentages in parentheses. Paired changes are fine-tuned minus base prompt-level R_{VP2} values, computed using all eligible draws within each policy; they are not the strict matched-draw sensitivity estimates. Confidence intervals are from within-run homology-component bootstrap resampling. Primary training seeds are 20260729 for ESM3 and 20260825 for PoET-2; the remaining rows are replication runs. Base draws are shared across training-seed contrasts within each architecture.

Model	Training seed	Base virulent/eligible (pooled R_{VP2}, %)	Fine-tuned virulent/eligible (pooled R_{VP2}, %)	Paired change (95% CI)
ESM3	20260729	1,495/2,011 (74.3%)	1,164/2,046 (56.9%)	-0.175 (-0.316, -0.060)
ESM3	20260730	1,495/2,011 (74.3%)	1,217/2,034 (59.8%)	-0.144 (-0.282, -0.039)
ESM3	20270731	1,495/2,011 (74.3%)	1,184/2,045 (57.9%)	-0.164 (-0.312, -0.053)
PoET-2	20260825	1,815/2,192 (82.8%)	1,401/2,192 (63.9%)	-0.189 (-0.237, -0.137)
PoET-2	20260826	1,815/2,192 (82.8%)	1,278/2,192 (58.3%)	-0.245 (-0.289, -0.201)
PoET-2	20260827	1,815/2,192 (82.8%)	1,337/2,192 (61.0%)	-0.218 (-0.266, -0.169)

source/draw keys gave 1,478/1,989 base and 1,183/1,989 fine-tuned virulent labels. The mean paired prompt-level change in the strict matched-draw sensitivity analysis was -0.1468 (95% CI, -0.2867 to -0.0398).

For replication training seed 20270731, 2,045 fine-tuned draws were eligible and 1,155 were excluded. Using all eligible draws within each policy, prompt-level R_{VP2} decreased for 70 prompts, increased for 25, and were unchanged for 34. Restricting both policies to the 1,996 identical source/draw keys gave 1,481/1,996 base and 1,154/1,996 fine-tuned virulent labels. The mean paired prompt-level change in the strict matched-draw sensitivity analysis was -0.1611 (95% CI, -0.3100 to -0.0499).

Thus, both the main analysis and the strict matched-draw sensitivity analysis showed a reduction in prompt-level R_{VP2} for every ESM3 training run.

PoET-2 primary and replication analyses. For each PoET-2 training-seed contrast, 2,192 of 3,200 base-policy draws and 2,192 of 3,200 fine-tuned draws were eligible. The respective exclusion counts were 1,008 and 1,008, and no eligible prediction was missing. The eligible set comprised 137 source prompts in 45 homology components. All 2,192 strict source/draw pairs were score complete in each contrast, so the main and common-score-complete sensitivity analyses used the same draws.

For the primary training seed 20260825, pooled R_{VP2} was 82.80% (1,815/2,192) under the base policy and 63.91% (1,401/2,192) after fine-tuning. The mean paired prompt-level change was -0.1889 (95% component-cluster bootstrap CI, -0.2374 to -0.1366). Effects were lower for 111 prompts, higher for 11, and tied for 15. The common-score-complete sensitivity analysis gave the same point estimate.

For replication training seed 20260826, fine-tuned pooled R_{VP2} was 58.3% (1,278/2,192), compared with 82.8% (1,815/2,192) for the shared base outputs. The mean paired prompt-level change was -0.245 (95% component-cluster bootstrap CI, -0.289 to -0.201).

For replication training seed 20260827, fine-tuned pooled R_{VP2} was 61.0% (1,337/2,192), compared with 82.8% (1,815/2,192) for the shared base outputs. The mean paired prompt-level change was -0.218 (95% component-cluster bootstrap CI, -0.266 to -0.169).

The R_{VP2} reduction therefore recurred across the primary and both replication training runs for each architecture. Appendix A.10 summarizes these effects alongside the other evaluated endpoints.

Cross-classifier agreement and interpretation. For the primary PoET-2 run, R_{VP2} agreed with R_{DTVF} in the reduction-versus-nonreduction classification for 106/137 common evaluated prompts (77.4%) and with VirulentHunter for 102/137 prompts (74.5%). Prompt-effect rank correlations were $\rho = 0.023$ between R_{VP2} and R_{DTVF} and $\rho = 0.239$ between R_{VP2} and VirulentHunter.

Table 7: Across-seed summary of paired fine-tuning effects. Cells give the run-specific paired change (fine-tuned minus base) for the primary seed and the two replication seeds of each architecture (ESM3: 20260729, 20260730, 20270731; PoET-2: 20260825, 20260826, 20260827). Mean and range summarize the three per-seed values; uncertainty is not pooled across runs.

Model	Endpoint	Primary	Repl. 1	Repl. 2	Mean	Range
ESM3	R_{DTVF}	-0.149	-0.142	-0.143	-0.145	$[-0.149, -0.142]$
ESM3	VirulentHunter prob.	-0.227	-0.211	-0.214	-0.217	$[-0.227, -0.211]$
ESM3	R_{VP2}	-0.175	-0.144	-0.164	-0.161	$[-0.175, -0.144]$
ESM3	Q_{profile}	$+0.967$	$+0.864$	$+0.868$	$+0.900$	$[+0.864, +0.967]$
PoET-2	R_{DTVF}	-0.140	-0.152	-0.130	-0.141	$[-0.152, -0.130]$
PoET-2	VirulentHunter prob.	-0.143	-0.175	-0.179	-0.166	$[-0.179, -0.143]$
PoET-2	R_{VP2}	-0.189	-0.245	-0.218	-0.217	$[-0.245, -0.189]$
PoET-2	Q_{profile}	$+10.609$	$+7.967$	$+8.811$	$+9.129$	$[+7.967, +10.609]$

The classifiers therefore agreed on the policy-level reduction within the supported PoET-2 subset, with limited agreement in the ranking of prompt-level effects. Section 3.3 defines the interpretation of these computational endpoints and their dependence on training-data provenance.

Profile HMM analysis of source-family compatibility. Fixed prompt residues were replaced by X before paired base and fine-tuned sequences were searched against the same profile HMM constructed from the corresponding source-homologue alignment. This analysis measures compatibility with the source-family profile. The following estimates refer to the primary training run of each architecture; replication-run quality and HMMER diagnostics are reported in Appendix A.7.

For ESM3, the median prompt-level change in normalized HMMER full-sequence score was $+0.0222$ (95% CI, $+0.00625$ to $+0.0504$). Score-complete effects were available for 164 of 200 prompts, comprising 2,123 of 3,116 matched sequence pairs and 58 of 68 homology components. Effects were positive for 121 prompts, negative for 38, and unchanged for 5; 36 prompts contained no sequence pair with scores reported in both conditions. HMMER reported a full-sequence score for 2,222 of 3,116 base-policy sequences (71.31%) and 2,385 of 3,116 fine-tuned sequences (76.54%). The prompt-level mean reporting-rate change was $+0.0520$ (95% CI, $+0.0155$ to $+0.0964$). These descriptive results were consistent with maintained source-family compatibility, but they do not constitute a formal equivalence test.

For PoET-2, the median prompt-level change in normalized HMMER full-sequence score was $+0.46157$ (95% CI, $+0.36968$ to $+0.57428$). Both policies had reportable scores for 2,906 matched pairs, covering all 200 prompts and 66 components. Effects were positive for 197 prompts and negative for 3. Base and fine-tuned reportable-hit counts were 2,990/3,200 (93.44%) and 3,062/3,200 (95.69%). The mean prompt-level hit-rate change was $+0.02250$ (95% CI, $+0.01136$ to $+0.03603$). There were 84 base-only, 156 fine-tuned-only, and 54 neither-hit pairs. Missing scores were not imputed as zero.

A.10 Across-seed summary of paired effects

Table 7 summarizes the principal paired effects across the three training seeds of each architecture. Every effect retained the same direction in every run, and the between-seed spread was small relative to the effect size; for example, ESM3 R_{DTVF} paired changes spanned -0.149 to -0.142, and PoET-2 R_{DTVF} paired changes spanned -0.152 to -0.130. Each per-seed value is a paired change with its own source-cluster interval; the mean and range summarize the three runs descriptively.

A.11 Embedding extraction, UMAP settings, and PoET-2 visualization

This section details the embedding-space analysis in Section 4.4 and presents the PoET-2 projection in Figure 7. Each architecture-specific cohort comprised 200 held-out source proteins, 3,200 base-policy outputs, and 3,200 fine-tuned outputs. All sequences were encoded at full length using the same frozen ESM3-sm-open-v1 sequence encoder (ESM software version 3.1.1). Final-layer

representations were mean-pooled over residues in FP32, excluding BOS/EOS but retaining native unknown-residue tokens, yielding 1,536-dimensional vectors. PoET-2 outputs were therefore represented in ESM3 embedding space. For each cohort, we L2-normalized the vectors and fitted UMAP jointly to all 6,600 observations without group labels. Parameters were cosine distance, 30 neighbors, minimum distance 0.1, 500 epochs, and random seed 20260831. The three displayed groups reused the same fitted coordinates within each architecture, without refitting. The ESM3 and PoET-2 cohorts were projected separately, so coordinates are not directly comparable between Figures 3 and 7.

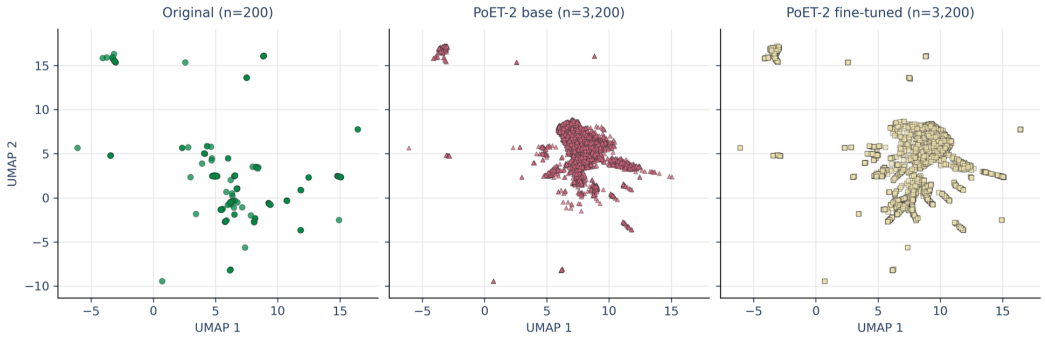


Figure 7: **UMAP projection of source proteins and PoET-2-generated outputs.** From left to right, panels show held-out source proteins (green circles; $n = 200$), base-policy outputs (red triangles; $n = 3{,}200$), and fine-tuned outputs (yellow squares; $n = 3{,}200$). All sequences, including PoET-2 outputs, were encoded using the same frozen ESM3 encoder. The three groups are displayed as subsets of a single UMAP fit to all 6,600 sequences in the PoET-2 cohort and share the same fitted coordinates. This fit is separate from the ESM3 fit in Figure 3; coordinates are not directly comparable between the two figures.

PoET-2 visualization. Figure 7 shows the held-out source proteins and the base- and fine-tuned-policy outputs for the PoET-2 cohort. The two generated groups overlapped substantially, with visually apparent local enrichment of each group within the main point cloud. This qualitative pattern complements the ESM3 projection in Figure 3 and the sequence-level results in Section 4.3. The projection is exploratory and does not independently establish changes in VF association or biological function.

A.12 Artifact fingerprints

Table 8 records SHA-256 fingerprints of the checkpoints and evaluator weights used in the reported runs, so that released artifacts can be verified against the results above.

Table 8: Artifact fingerprints for the checkpoints and evaluator weights used in all reported runs.

Artifact	Seed	SHA-256
ESM3 fine-tuned checkpoint (step-000297)	20260729	`94140a54f0974fdbb04893ee47c18bd0de805a6df5bbe579793f7eb3c311c72e`
ESM3 fine-tuned checkpoint (step-000297)	20260730	`e332f3614035d2322c4829eca51e5973bf9f6b75cd9d62f026c6c7256c3e7d1f`
ESM3 fine-tuned checkpoint (step-000297)	20270731	`22f95e28e91197483f7b0bbeff0c9afa4e5f9c62d1f762832be06d3d9a19e9ba`
PoET-2 base weights (`poet-2.ckpt`)	–	`1f5627e08ab6dd45236a655fe676df5d17d769ce7109fa10dfebb27fdf152a79`
PoET-2 fine-tuned checkpoint (epoch-003)	20260825	`f8b3ccf169855c0ac11e9dba76b6e10546293c0c3cb74b5be2cbde1f0330ff99`
DTVF evaluator weights (`best_model.pth`; 9,034,786 bytes)	–	`0cd5f56ecec5b611d69d5e8768dfa89dc2a5cad1794f7130452136c2a64db28f`