EXTENDING CONTROLLABILITY FROM COMPOSITION AND TOPOLOGY TO DOMAIN IDENTITY IN LANGUAGE-MODEL PROTEIN GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein language models can generate plausible sequences, but reliable control over multiple user-specified properties remains limited. We present Aria, which tunes Llama-3.1-8B-Instruct on 350k Swiss-Prot proteins described by a random subset of thirteen physicochemical, developability, structural, and functional descriptors written into the prompt, and evaluates adherence using the same functions that defined the six directly computable properties, supplemented by predictors of topology and domain identity. On cluster-separated evaluation proteins, Aria improves adherence across directly measurable properties, including length, hydrophobicity, cysteine count, transmembrane topology, and signal peptides. Prompt-mismatch controls show that these gains depend on the requested specification. Exact Pfam-domain recovery remains difficult and increases with family exposure. We next test whether natural-sequence-anchored direct preference optimization can improve weaker structural and functional axes. One optimization stage raises mean ESMFold pLDDT by 8.2 points in PF00416, ribosomal protein uS13. Across three additional optimized families, requested-domain recovery increases by 20.8-43.6 percentage points at the evaluated checkpoints, averaged across two training seeds, with corresponding increases in predicted structural similarity. However, applying the resulting family-specific adapters to six families not used for preference optimization does not consistently improve target-domain recovery over the supervised model. These results show that a general instruction-tuned model can acquire broad protein-property control without protein-specific pretraining, while natural-sequence-anchored preference optimization provides substantial but family-local gains.

1 INTRODUCTION

Protein language models have made sequence generation routine. Autoregressive models trained on large sequence corpora, among them ProGen (Madani et al., 2023), ProGen2 (Nijkamp et al., 2023) and ProtGPT2 (Ferruz et al., 2022), emit sequences that fold and resemble natural proteins, and encoder families such as ESM (Lin et al., 2023) represent them well enough to support structure prediction. Producing a plausible protein is no longer the difficulty. Directing that production is.

Two obstacles stand in the way, and they compound each other. The first is that the interface is often narrow. For example, ProGen and ProGen2 condition on family or keyword tags, ProtGPT2 does not condition at all, and ProLLaMA (Lv et al., 2026) exposes a single categorical axis, the protein superfamily, treating physicochemical properties as a decode-time filter rather than as targets a user may set. Work that does address attributes directly steers through learned tokens or model-internal interventions (Ismail et al., 2025; Luo et al., 2026; Liu et al., 2025b; Hayes et al., 2025), or aligns free-text function descriptions with sequence (Liu et al., 2025a; Wang et al., 2024), rather than accepting a user-provided textual specification. The second obstacle is measurement. Evaluations across recent protein-generation studies (Madani et al., 2023; Nijkamp et al., 2023; Ferruz et al., 2022) are dominated by perplexity, predicted structural confidence (Jumper et al., 2021), and template-modeling scores (Zhang & Skolnick, 2004). These metrics ask whether a generation re-

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

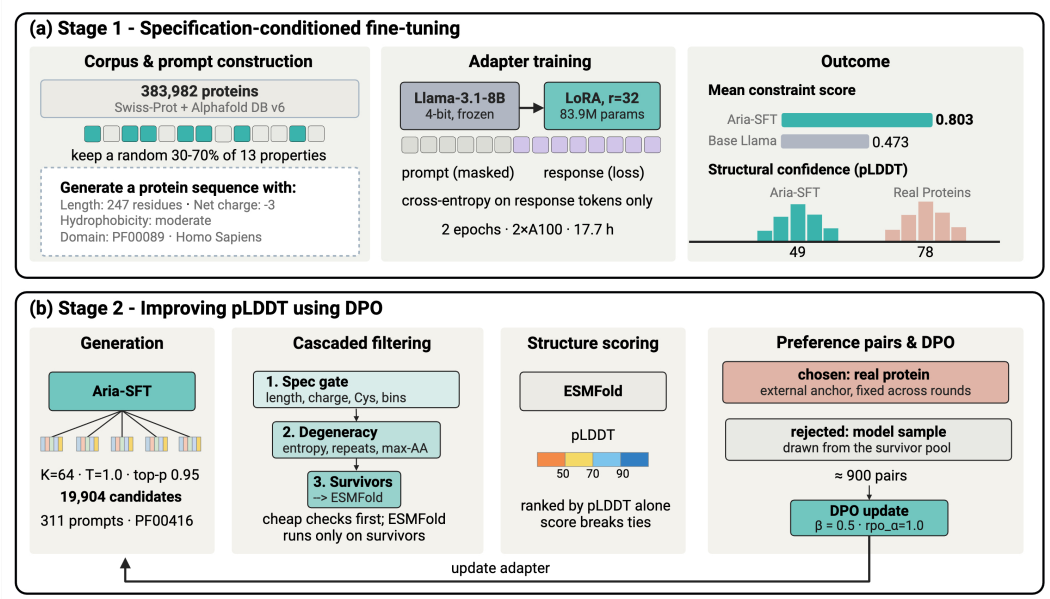


Figure 1: The Aria training and preference-optimization workflow. (a) During specification-conditioned fine-tuning, each example pairs a specification containing a randomly selected 30-70% subset of thirteen protein descriptors with its natural sequence; loss is computed only over the sequence. (b) Preference optimization against that measure, shown for ribosomal protein uS13 (PF00416) and applied identically to three further Pfam families. Candidates sampled from the supervised model pass a specification gate and composition filters before structure prediction, so ESMFold is run only on survivors. For each prompt the natural Swiss-Prot sequence supplies the preferred response and a surviving generation the rejected one, fixing one side of every comparison at a point that satisfies the specification by construction. The adapter is updated and the loop repeats for up to three rounds.

sembles a protein, not whether it carries the properties that were requested. This distinction matters in practice: a model that disregards a stated net charge while emitting a well-folded sequence scores well on commonly used measures, and the failure surfaces only downstream, during purification or formulation.

Instruction tuning can address both at once (Zhang et al., 2023). A general instruction-tuned model already accepts conditions stated in natural language, so a conjunctive, partial request needs no bespoke interface, only training data in which such requests appear; and for many properties the request is verifiable, since the property can be recomputed on the generated sequence with the function that wrote the specification.

Using this approach, we instruction-tune Llama-3.1-8B-Instruct (Grattafiori et al., 2024) with QLoRA (Dettmers et al., 2023; Hu et al., 2022) on 348,867 Swiss-Prot proteins, each described by a partial, randomly sampled subset of thirteen physicochemical, developability, structural and functional descriptors written into the prompt, and evaluate on 35,115 held-out specifications separated from training by sequence clustering. A single low-rank run of approximately 35 GPU-hours, updating 1% of the parameters without protein-specific pre-training, improves adherence on every axis we can recompute. Composition follows the specification closely: length within $\pm 10\%$ rises from 11.6% to 88.8% and exact cysteine count from 23.3% to 68.9%. Topology follows it as well, with transmembrane-helix Matthews correlation rising from 0.275 to 0.869 and signal peptide from 0.073 to 0.739. Exact Pfam-domain recovery is the exception, at 2.9%, and increases with domain exposure during training.

Adherence, however, is not accompanied by confident structure: mean ESMFold pLDDT lies between 45 and 59 across the families we examine, against 77 to 83 for their natural members. Because

these evaluations are computed from sequence, they can act as training signals rather than only as scores. We therefore use the same measure as a reward, applying direct preference optimization (Rafailov et al., 2023) anchored on natural family sequences to four Pfam families. Mean pLDDT increases in all four families, with gains reproduced at two training seeds in the three additional families and reaching up to 13.1 points, closing the distance to natural members of the same family.

In the three additional families, two measures not used to rank preference pairs improve alongside predicted confidence: recovery of the requested domain, assessed by hmmsearch against each family's own profile, rises from 9-44% to 30-80%, and the fraction of generations reaching the natural fold increases in all three families. The gains do not transfer consistently across families. Across six transfer targets, family-specific adapters do not consistently improve requested-domain recovery. Direct preference optimization on three targets nevertheless increases strict-composite passage, indicating that limited cross-family transfer does not preclude improvement using target-family examples.

These results indicate that a general instruction-tuned model can be repurposed into a specification-conditioned protein generator through a single parameter-efficient fine-tuning run, and that evaluation-guided preference optimization can subsequently improve predicted structural confidence and domain recovery, although its effects on specification adherence vary across families. Although demonstrated here for a general-use language model, this formulation provides a testable route to improving adherence in existing protein language models.

2 RESULTS

Supervised instruction tuning improves adherence across the evaluated properties, but the degree of control varies by property class (Figure 2, Table 1). Adherence improves for compositional, developability-related (aggregation propensity and cysteine count), and topological properties, and topology show substantial improvements, whereas recovery of the requested protein family or do-

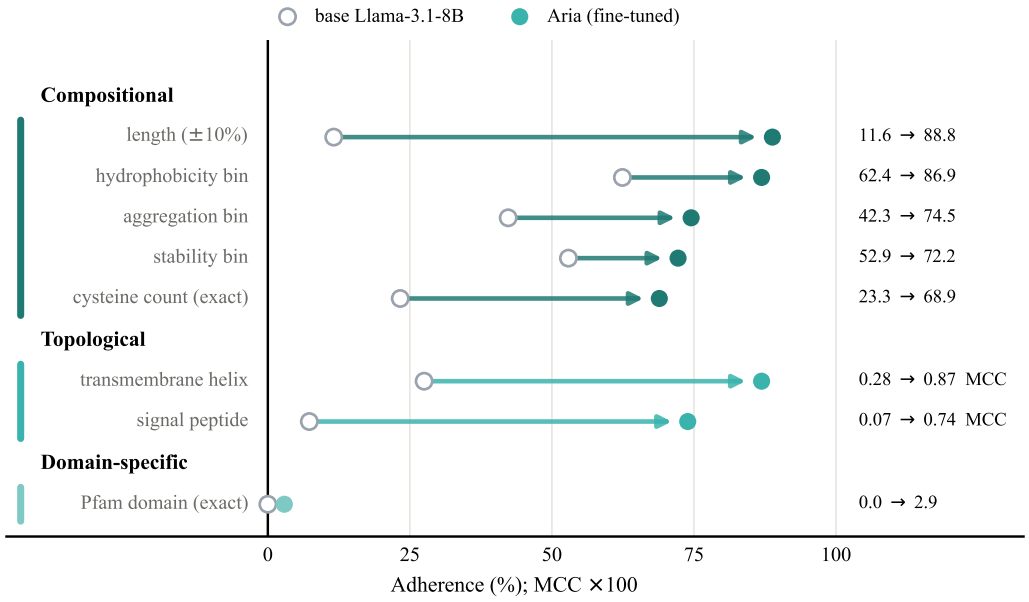


Figure 2: The controllability hierarchy. Each row shows adherence for one property before and after supervised fine-tuning. Open markers denote the base model, filled markers denote Aria, and connecting lines indicate the change. Rows are read independently, so percentage and MCC rows share one axis with MCC $\times 100$. Enzyme class is omitted because its measurement is confounded (Appendix G).

main, as represented in the Pfam database, remains low. We first characterize these differences and evaluate the dependence of adherence on the requested specification. We then examine whether anchored preference optimization improves predicted structural confidence and domain recovery, how it affects specification adherence, and whether the gains transfer across families.

Tier 1: Composition and Developability. Fine-tuning first teaches the model to return the requested output format: the fraction of generations containing a bare sequence rather than prose rises from 0% to 99.9%. Beyond formatting, adherence improves across every measured sequence-derived property. Length adherence rises from 11.6 to 88.8%, hydrophobicity from 62.4 to 86.9%, and exact cysteine-count adherence from 23.3 to 68.9%; aggregation and stability show similar gains (Figure 2). While exact numerical control remains less precise for charge, the correlation between requested and realized net charge rises from $r = 0.26$ to $r = 0.89$, showing that the model responds strongly to the requested direction (Table 1, Appendix E).

Tier 2: Topology. Transmembrane-helix recall increases from 14.0% to 86.5%, and Matthews correlation coefficient (MCC) from 0.275 to 0.869. Signal-peptide recall increases from 6.6% to 61.4%, with MCC rising from 0.073 to 0.739. Positive-class precision remains high at 91.4% for transmembrane helices and 93.2% for signal peptides, indicating that the increased recall is accompanied by specificity for the requested topological features (Figure 2).

Tier 3: Domain Identity. Domains are assigned using hmmscan (Eddy, 2011) against Pfam-A (Mistry et al., 2021), where lower E-values indicate stronger evidence for a sequence-profile match. At the default reporting threshold, the requested domain is detected in 2.90% of fine-tuned generations (520 of 17,941), compared with none of 15,140 base-model generations. The few highly significant assignments are nevertheless specific: among the 145 generations receiving any match at $E \leq 10^{-10}$, 71.0% match the requested domain. The aggregate recovery rate is partly depressed by uneven training exposure: many requested domains are absent or rare in training, and success increases markedly with their representation (Appendix F, Figure A3). Thus, the model acquires targeted signatures for some represented domains, but exact-domain control across the full held-out set remains limited. Whether it can be improved within a single family, using the same recomputed measurements as a training signal, is examined in Section 2.3.

2.1 Adherence reflects prompt conditioning

Raw adherence can be inflated by common property values even when a model ignores the prompt. We therefore compare Aria with two request-blind controls: a constant baseline that always predicts the most frequent training value, and a mismatch control that scores each generated sequence against a different held-out specification. Aria exceeds both controls on all six directly recomputable properties (Table 1). Mismatching specifications reduces adherence by between 20.3 percentage points for net charge and 76.9 percentage points for length. Mean compositional adherence also remains similar when the requested Pfam family is absent from training (71.4%, compared with 71.0% for represented families). Thus, the model can follow compositional requirements even when the requested family was not represented during fine-tuning. Together, these findings support dependence on the requested specification and indicate that compositional control extends to proteins from families absent from supervised training.

Table 1: Prompt-conditioned adherence on directly recomputable properties. The value of n is the number of valid Aria generations whose specification requested the corresponding property. Bracketed values are 95% Wilson confidence intervals for Aria's adherence rate.

Property	n	**Aria [95% CI]**	Constant	Mismatch
Length ($\pm 10\%$)	19,140	**88.8% [88.4, 89.3]**	19.8%	11.9%
Hydrophobicity bin	19,030	**86.9% [86.4, 87.3]**	72.4%	52.3%
Aggregation bin	18,966	**74.5% [73.9, 75.2]**	34.2%	33.0%
Stability bin	18,998	**72.2% [71.5, 72.8]**	55.0%	51.7%
Cysteine count (exact)	18,853	**68.9% [68.3, 69.6]**	17.5%	10.4%
Net charge (± 2)	18,919	**35.2% [34.5, 35.9]**	25.3%	14.9%

2.2 Additional constraints do not weaken length control

We next test whether adding other requested properties weakens control of a fixed length target. For this experiment, we use a subset of 200 proteins from the cluster-separated held-out set. For each protein, we construct thirteen nested specifications, starting with length alone and adding one descriptor at each step while keeping the requested length unchanged. Length adherence remains between 84 and 92%, with no significant trend as constraints are added (slope $+0.028$, 95% CI $[-0.006, +0.061]$; Appendix H). Thus, length adherence remains stable across the tested constraint ladder.

2.3 Anchored DPO improves predicted structural confidence and domain recovery

Supervised tuning improves compositional and topological control, but requested-domain recovery remains limited and predicted structural confidence falls below that of natural family members. We therefore seek to shift the generation distribution towards sequences with higher predicted structural confidence and stronger family-specific characteristics while retaining specification adherence. To this end, we apply natural-sequence-anchored direct preference optimization (DPO) as a post-training stage, which increases the relative likelihood of preferred natural sequences over generated alternatives (Rafailov et al., 2023).

We construct preference pairs using ESMFold pLDDT as the structural-confidence criterion. Before structure prediction, a specification gate checks agreement with the requested sequence-derived properties, while sequence filters exclude low-complexity, excessively repetitive, or compositionally imbalanced outputs. Candidates passing both checks (gate and filter) are folded with ESMFold and paired with the matching natural sequence when its predicted confidence is higher. The natural sequence serves as the preferred response. Training combines DPO with an auxiliary negative-log-likelihood term on preferred responses (Pang et al., 2024). This configuration mitigated the reward-hacking failure observed under self-play preferences, in which higher ESMFold confidence was accompanied by poorer specification adherence and sequence composition (Appendix J, Figure A5). Domain recovery and structural similarity are evaluated separately and are not used to construct the preference ordering.

We first evaluate this approach on ribosomal protein uS13 (PF00416), using 311 natural sequences for preference construction and 260 cluster-separated sequences for held-out evaluation. Pairing the training sequences with eligible generated candidates yields 929 preference pairs, because a natural sequence can be paired with more than one candidate. One training stage increases mean ESMFold pLDDT from 49.4 to 57.6, compared with 78.0 for the full natural reference set of 571 sequences (311 training and 260 held-out). Under matched evaluation with one generation per model per prompt, specification-gate passage increases from 16.9% to 25.8% and sequence-filter passage from 88.5% to 96.5%. The strict composite requiring the gate, filters, and pLDDT ≥ 70 increases from 0.4% to 7.3% (1 to 19 of 260 prompts). Thus, predicted confidence and these population-level adherence measures improve concurrently in PF00416.

To assess whether the improvements extend beyond PF00416, we repeat anchored preference optimization on three additional families: ribosomal proteins uL22 (PF00237) and uL24 (PF00467), and YqgF (PF03652). Each undergoes three rounds of anchored preference optimization at two training seeds, with separate preference-training, validation, and final-test splits. Evaluation uses one generation per held-out prompt; outputs without an extractable sequence count as failures in population pass rates. Mean pLDDT increases relative to the supervised model in all three groups at both seeds.

To assess whether the confidence gains are accompanied by improvements beyond the preference-selection metric, we also measure requested-domain recovery and structural similarity to the corresponding natural references. Both improve, although neither is used to rank preference pairs. Here, domain recovery is assessed using hmmsearch against the requested Pfam profile with its curated gathering threshold, rather than the full Pfam-A scan used in the broad supervised evaluation. At the checkpoints used for subsequent transfer experiments (round two for PF00237 and round three for PF00467 and PF03652) domain recovery averaged across the two seeds increases from 9.0% to 29.8%, 28.4% to 72.0%, and 44.2% to 80.0%, respectively. The fraction of generations whose predicted structures achieve TM-score ≥ 0.5 against the corresponding natural-reference structures also increases in all three groups at both seeds (Figure 3).

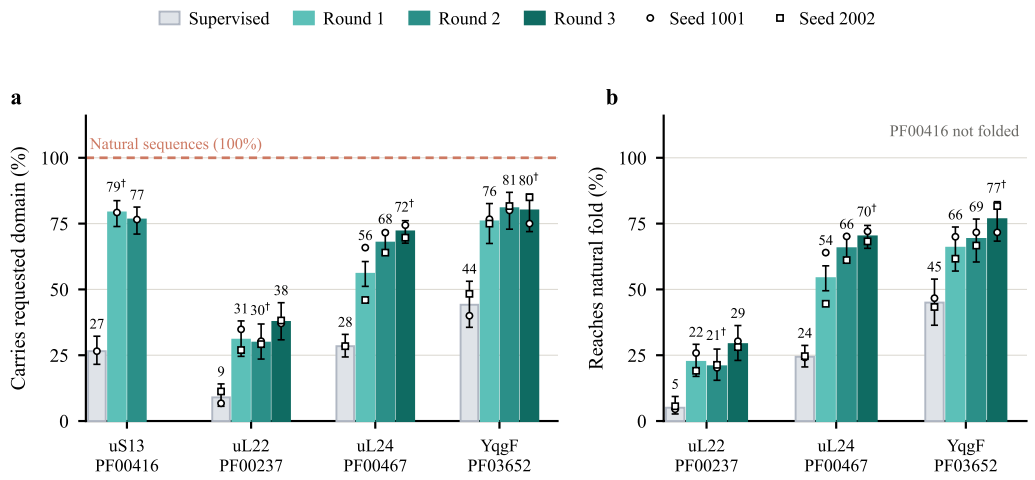


Figure 3: Within-family outcomes of anchored preference optimization for two seeds. (a) Requested-domain recovery. (b) Fraction of generations with TM-score ≥ 0.5 against the predicted structure of the corresponding natural reference. Rates use all held-out prompts as denominators: 260 for PF00416, 89 for PF00237, 211 for PF00467, and 60 for PF03652. Bars show rates averaged across the two seeds and the markers show seed-specific rates. Error bars show 95% Wilson intervals and do not represent variation across training seeds. PF00416 has two optimization rounds and no TM-score analysis in this comparison. Domain recovery and structural similarity improve relative to supervised tuning, with trajectories that vary across families.

These gains are not accompanied by uniform improvements in specification adherence. For example, the first PF00237 round increases mean pLDDT at both seeds, while gate passage decreases from 40.4% to 33.7% at one seed and increases from 28.1% to 46.1% at the other. Anchored preference optimization therefore improves predicted structural confidence and domain recovery across the tested groups, but preservation of specification adherence must be assessed separately.

2.4 Comparison with a protein-specific language model

We next assess how supervised Aria compares with a model built specifically for protein generation, before any family-specific preference optimization. ProLLaMA (Lv et al., 2026) is pretrained on protein sequences and instruction-tuned for superfamily-conditioned generation. Its generation interface takes one instruction, [Generate by superfamily] Superfamily=name, with the name drawn from the fixed vocabulary the model was trained on. It has no field for length, net charge, hydrophobicity, cysteine count, or any other property target, so the specifications used throughout this paper cannot be presented to it at all. The asymmetry in the inputs is a property of the baseline rather than a choice in the protocol. That vocabulary also fixes the scope of the comparison: five of the nine families in this study have an entry, three have none, and one is represented only by a different paralog of the same protein.

Each model therefore receives what its interface accepts, Aria the held-out textual specification and ProLLaMA the corresponding superfamily label. Both generate one sequence per evaluation case at a sampling temperature of 0.7, and both are scored against the same criteria.

ProLLaMA achieves higher requested-domain recovery than supervised Aria in three of the five families (Table 2). For example, recovery reaches 96.4% for ProLLaMA versus 50.0% for Aria on PF00886, although this comparison comprises only 28 held-out cases. Conversely, Aria achieves higher recovery on PF00177 (38.6% versus 0.3%), while neither model recovers PF01157. Domain recovery therefore varies substantially across families for both models.

Aria passes the specification gate more frequently in all five families, with rates of 28.6-40.4%, compared with 9.0-16.0% for ProLLaMA. Aria also achieves higher rates on the strict composite

Table 2: Supervised Aria and ProLLaMA on five evaluated families covered by ProLLaMA's conditioning vocabulary. All rates are percentages over the n evaluation cases. The strict composite requires gate passage, sequence-filter passage, and pLDDT ≥ 70.

		Domain recovery		Gate passage		Strict composite	
Family	n	ProLLaMA	Aria	ProLLaMA	Aria	ProLLaMA	Aria
PF00237	89	29.2	6.7	9.0	40.4	1.1	9.0
PF00177	352	0.3	38.6	10.5	33.5	0.0	4.8
PF01196	168	57.7	50.6	10.7	28.6	0.6	11.9
PF00886	28	96.4	50.0	10.7	39.3	10.7	25.0
PF01157	50	0.0	0.0	16.0	40.0	0.0	8.0

requiring gate passage, sequence-filter passage, and pLDDT ≥ 70 in every family. For ProLLaMA, passing the gate means that a generated sequence happens to meet property targets that were not supplied to the model; it does not measure its ability to follow those targets. The comparison therefore shows a practical distinction between generating a member of a requested family and generating a sequence that also meets additional requirements. With supervised adaptation on 348,867 proteins and no protein-specific pretraining, Aria supports multi-property conditioning and achieves higher gate and composite rates in this evaluation, although ProLLaMA recovers the requested domain more frequently in three of the five families.

2.5 Preference optimization does not consistently transfer across families

The within-family improvements raise the question of whether an adapter optimized on one family can benefit other families without constructing new preference pairs for each target. We test this using the adapters trained separately on the three additional families in Section 2.3, treating these as source families. Six other families, selected for proximity to a source in the family-level ESM2 embedding, serve as transfer targets and contribute no examples to source-family preference training. Each target is evaluated with all three source adapters at both training seeds, yielding 18 source-target combinations per seed. We compare each adapter with the supervised model on the same held-out target specifications (Figure 4).

Transfer is heterogeneous and does not reproduce the consistent within-family improvements. Averaged across seeds, requested-domain recovery decreases for four of the six preassigned nearby source-target pairs and remains unchanged for two. The strict composite requiring specification-gate passage, sequence-filter passage, and pLDDT ≥ 70 decreases in 25 of the 36 source-target-seed comparisons. An increase in domain recovery does not necessarily imply an increase in strict-composite passage. For example, the PF03652 adapter increases domain recovery on PF00366 by 6.2 percentage points averaged across seeds, while the strict-composite rate decreases by 1.3 points.

To separate the contribution of a transferred adapter from that of the target family's own preference data, we optimized three of the six target families directly, each from two starting points, with two runs per configuration. Ribosomal proteins bL17 (PF01196) and uS17 (PF00366), and the RuvC family (PF02075), each received three anchored rounds beginning either from the general supervised model or from the adapter trained on that target's nearest source family. All other settings match Section 2.3, and both conditions use the target's own preference-training split.

The two conditions begin far apart. Averaged across seeds, requested-domain recovery at the starting checkpoint is 55% for the supervised model against 43% for the transferred adapter on PF01196, 23% against 14% on PF00366, and 68% against 35% on PF02075. This restates the transfer result above in a third form: the adapter optimized on the nearest source is a worse starting point than the model it was derived from.

Target-family preference optimization increased requested-domain recovery from both the general supervised and nearby-family initializations (Figure 5). We compare the two initializations after three rounds of optimization. Strict-composite passage on PF01196 reached 21.4-23.8% from the general supervised initialization and 22.6-24.4% from the nearby-family initialization. Corre-

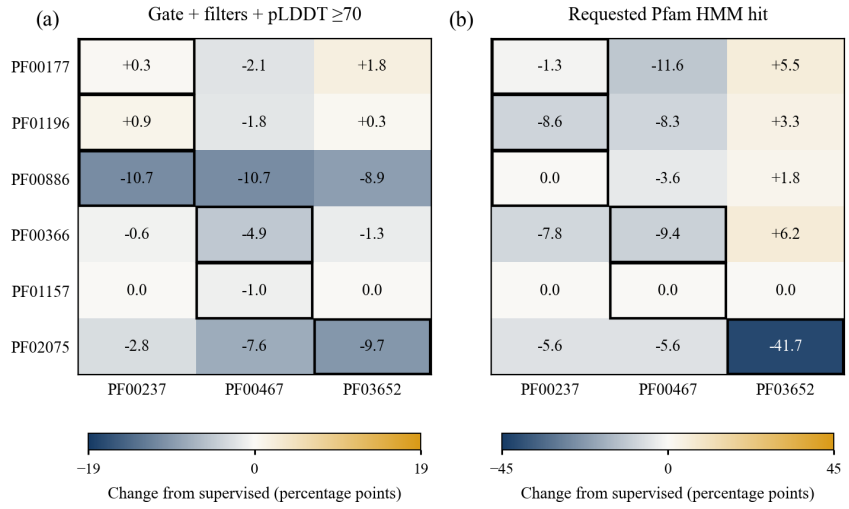


Figure 4: Transfer of family-specific preference optimization. Columns identify source families used for preference training; rows identify target families excluded from that training. Cells show changes from the corresponding supervised baseline, averaged across two seeds, in percentage points. (a) Strict-composite passage, requiring the specification gate, sequence filters, and pLDDT $\geq$ 70. (b) Requested-domain recovery. Black outlines identify preassigned nearby source-target pairs. The adapters are from round two for PF00237 and round three for PF00467 and PF03652.

sponding rates were 26.6-27.9% and 20.8-24.7% on PF00366, and 29.2-30.6% and 22.2-23.6% on PF02075. Ranges describe the two runs, not confidence intervals.

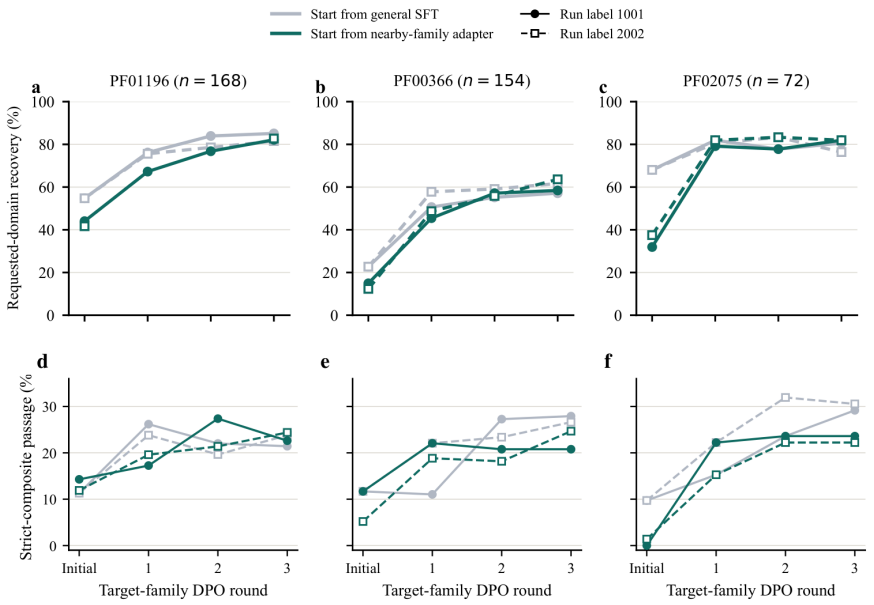


Figure 5: Starting point against destination for three families that also serve as transfer targets. (a-c) Requested-domain recovery, assessed with hmmsearch against the family's own Pfam profile. (d-f) Strict-composite passage, requiring the specification gate, the sequence filters, and pLDDT $\geq$ 70. The initial point is the checkpoint each condition began from, so the nearby-family series starts at the transferred adapter's rate rather than the supervised model's. Rates are over all held-out prompts.

These results distinguish cross-family transfer from optimization using target-family examples. Target-family optimization improved strict-composite passage across all three tested families, whereas initialization from a nearby-family adapter did not yield consistently higher round-three rates than initialization from general supervised Aria.

3 DISCUSSION

Across the tested axes, Aria shows a hierarchy of controllability. Gains are strongest for physicochemical, developability, and topological properties, whereas exact Pfam-domain recovery remains low in the broad supervised evaluation. Request-blind controls show that the composition gains are not explained solely by an unconditional shift in model outputs. On cluster-separated evaluation proteins, composition control persists, whereas exact domain recovery increases with family exposure (Appendix F). Controllability therefore depends on both the requested property and, for domain identity, training exposure. On PF00416, one natural-sequence-anchored DPO stage raises mean ESMFold pLDDT from 49.4 to 57.6 and improves the property gate, sequence filters, and strict composite under matched one-sample evaluation. Nevertheless, 57.6 remains below the ESMFold mean of 78.0 for the natural PF00416 reference set. One round therefore closes approximately 29% of this confidence gap, yielding a substantial but partial improvement.

The additional experiments extend these findings beyond PF00416: predicted structural confidence, requested-domain recovery, and structural similarity improve across three further families at both training seeds. However, specification adherence does not improve uniformly, and the resulting adapters do not consistently benefit other families. Direct preference optimization increases strict-composite passage across three target families, whereas the available cold- and warm-start comparisons do not establish a consistent benefit from nearby-family initialization. These findings support target-specific adaptation rather than assuming that improvements learned within one family will transfer to another. The ProLLaMA comparison further distinguishes domain recovery from multi-property adherence: neither model dominates domain recovery across families, whereas Aria achieves higher specification-gate and composite rates under its richer conditioning interface.

Several limitations remain. Supervised adaptation was evaluated with one training run of an 8B model. The preference-optimization experiments cover a restricted set of families and a limited number of runs, leaving variability across model scales and broader biological settings uncharacterized. ProLLaMA has no interface for property targets, so that comparison cannot be run under exact identical inputs and does not isolate the effect of the training procedures. Moreover, anchored preference optimization requires matching natural sequences as preferred responses; its applicability to specifications without such examples remains untested. All structural and functional assessments are computational: increased ESMFold confidence, structural similarity, and Pfam-profile recovery do not establish experimental folding or biological function.

Within these bounds, the study demonstrates a parameter-efficient route to multi-objective, property-conditioned protein generation. More broadly, the measurement-to-feedback strategy may improve adherence in an existing protein generator whenever the requested property can be reliably evaluated.

AI USE STATEMENT

We used generative AI assistance to discuss and refine experimental methodology, interpret results, and identify limitations and potential confounders. AI tools also assisted with literature searches, drafting and revising manuscript text, implementation and debugging of training and evaluation pipelines, analysis code, and scientific visualizations. The authors reviewed and revised the AI-assisted text and figure designs. AI assistance was not used to create an additional synthetic protein dataset; model-generated sequences used in the experiments, including preference construction, are described in the Methods. The authors are responsible for the final manuscript, including its analyses, citations, claims, and AI-assisted artifacts.

REPRODUCIBILITY STATEMENT

Details sufficient to reproduce the reported results are given in the appendices. Appendix A specifies the base model, the adapter configuration, the optimization settings, the construction of preference

pairs, and the evaluation procedure, including the criteria applied by the specification gate and the sequence filters. Appendix B lists the thirteen conditioning descriptors, their sources, and the methods used to evaluate adherence. Appendix C reports the training footprint, and Appendix I states the rule by which source and target families were selected, which was fixed before any preference training. Corpus construction, the cluster-level separation of training from evaluation sequences, and the disjoint preference-training, validation and test partitions within each family are described in Appendix A; reported checkpoints were selected on the validation split, and the test split was not consulted during selection. External tools are applied at published settings: ESMFold for structure prediction, HMMER with Pfam gathering thresholds for domain assignment, and TM-score for structural comparison. The three families optimized after PF00416 were each trained at two seeds.

REFERENCES

Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J. L. de Hoon. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009. doi: 10.1093/bioinformatics/btp163.

Oscar Conchillo-Solé, Natalia S. de Groot, Francesc X. Avilés, Josep Vendrell, Xavier Daura, and Salvador Ventura. AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinformatics*, 8:65, 2007. doi: 10.1186/1471-2105-8-65.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pp. 10088–10115, 2023. doi: 10.52202/075280-0441.

Sean R. Eddy. Accelerated profile HMM searches. *PLoS Computational Biology*, 7(10):e1002195, 2011. doi: 10.1371/journal.pcbi.1002195.

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

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Jeppe Hallgren, Konstantinos D. Tsirigos, Mads Damgaard Pedersen, José Juan Almagro Armenteros, Paolo Marcatili, Henrik Nielsen, Anders Krogh, and Ole Winther. DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. *bioRxiv*, 2022. doi: 10.1101/2022.04.08.487609.

Thomas Hayes, Roshan Rao, Halil Akin, et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025. doi: 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.

Aya Abdelsalam Ismail, Tuomas Oikarinen, Amy Wang, et al. Concept bottleneck language models for protein design. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2411.06090.

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

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

Shengchao Liu, Yutao Zhu, Jiarui Lu, Zhao Xu, Weili Nie, Anthony Gitter, Chaowei Xiao, Jian Tang, Hongyu Guo, and Anima Anandkumar. A text-guided protein design framework. *Nature Machine Intelligence*, 7(4):580–591, 2025a. doi: 10.1038/s42256-025-01011-z.

Xiangyu Liu, Yi Liu, Silei Chen, and Wei Hu. Controllable protein sequence generation with LLM preference optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 505–513, 2025b. doi: 10.1609/aaai.v39i1.32030.

Jiawei Luo, Jiahao Li, Xianliang Liu, Yang Zhang, Qingcai Chen, and Junjie Chen. Controllable protein design by prefix-tuning protein language models. *Journal of Chemical Information and Modeling*, 66(8):4932–4946, 2026. doi: 10.1021/acs.jcim.5c03186.

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

Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos, 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*, 41:1099–1106, 2023. doi: 10.1038/s41587-022-01618-2.

Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A. Salazar, Erik L. L. Sonnhammer, Silvio C. E. Tosatto, Lisanna Paladin, Shriya Raj, Lorna J. Richardson, Robert D. Finn, and Alex Bateman. Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1):D412–D419, 2021. doi: 10.1093/nar/gkaa913.

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

Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. Iterative reasoning preference optimization. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, pp. 116617–116637, 2024. doi: 10.52202/079017-3702.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, 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, 2023. doi: 10.52202/075280-2338.

Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, 2017. doi: 10.1038/nbt.3988.

Felix Teufel, José Juan Almagro Armenteros, Alexander Rosenberg Johansen, Magnús Halldór Gislason, Silas Irby Pihl, Konstantinos D. Tsirigos, Ole Winther, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nature Biotechnology*, 40:1023–1025, 2022. doi: 10.1038/s41587-021-01156-3.

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

Zeyuan Wang, Qiang Zhang, Keyan Ding, Ming Qin, Xiang Zhuang, Xiaotong Li, and Huajun Chen. InstructProtein: Aligning human and protein language via knowledge instruction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1114–1136, 2024.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023.

Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004. doi: 10.1002/prot.20264.

A METHOD

Aria is built in two stages (Figure 1).

Corpus and specifications. From the reviewed subset of UniProt (The UniProt Consortium, 2023) we keep sequences of 20-512 residues with canonical residues, no fragments and no duplicates, leaving ∼384k proteins. To prevent leakage we cluster with MMseqs2 (Steinegger & Söding, 2017) at 30% identity and 80% coverage and split at the *cluster* level: 348,867 training and 35,115 held-out proteins, no held-out protein above 30% identity to any training protein. Thirteen descriptors span four families (Appendix B): physicochemical and developability properties computed from the sequence with Biopython (Cock et al., 2009) and an AGGRESCAN scale (Conchillo-Solé et al., 2007); structural and topological properties from AlphaFold (Jumper et al., 2021) and UniProt annotations; and curator-assigned functional annotations. Each specification contains a random 30-70% subset of the descriptors a protein has (3-10 per prompt, mean 6.6), which trains the model for the realistic under-specified case. The target completion is the bare sequence.

Training. QLoRA (Dettmers et al., 2023; Hu et al., 2022) on Llama-3.1-8B-Instruct (Grattafiori et al., 2024): 4-bit base, rank-32 adapter ($\alpha = 16$), two epochs, loss on the completion only. 83.9M trainable parameters, ∼35 GPU-hours. No protein-specific pre-training.

Evaluation. Adherence is computed only over specifications that requested a property, so n varies per property. Six axes are recomputed with the feature code that built the prompt; applying it to the reference Swiss-Prot sequence returns 100% by construction, which bounds definitional slack at zero. Signal peptides use SignalP 6 (Teufel et al., 2022), transmembrane helices DeepTMHMM (Hallgren et al., 2022), domains `hmmscan` (Eddy, 2011) against Pfam-A (Mistry et al., 2021). We report MCC for the imbalanced binary topology tasks. GO and organism are not scored. The base model is prompted identically; it embeds sequences in prose, so we extract the longest canonical residue run, and outputs with none are dropped from denominators rather than counted as failures, which favours the baseline.

Multi-family study and data partitioning. We extended preference optimisation to three source families: PF00237, PF00467, and PF03652. The original cluster-separated supervised test partition was retained for final evaluation. Within the supervised training partition, clusters were further divided into preference-training and validation subsets. Validation therefore excluded preference-training examples, but was not held out from the preceding supervised training. The resulting preference-training/validation/test counts were 502/111/89 for PF00237, 342/67/211 for PF00467, and 316/71/60 for PF03652.

Source and transfer-target selection. We represented 2,403 Pfam families with at least 20 sequences using mean ESM-2 sequence embeddings and grouped their normalised representations into 600 clusters. Source selection considered cluster medoids representing soluble proteins of 80-200 residues, with at least 250 preference-training, 20 validation, and 20 test examples, and mean annotated AlphaFold confidence of at least 70. Among qualifying sources with eligible transfer targets, we selected three in descending order of cluster size. Transfer targets were nearby families within the corresponding embedding cluster, subject to the same length, solubility, and confidence criteria and at least 20 test examples; no minimum preference-training count was required. The six targets were PF00177, PF01196, PF00886, PF00366, PF01157, and PF02075, with 352, 168, 28, 154, 50, and 72 test prompts, respectively. These families were excluded from source-family preference optimisation, but not necessarily from the original supervised training corpus.

Iterative natural-sequence-anchored preference optimisation. For each source family, we performed three optimisation rounds with seeds 1001 and 2002. Each round sampled 64 candidates per preference-training prompt from the current adapter. Candidates were screened by the specification gate and sequence filters before ESMFold evaluation. The matching natural sequence served as the preferred response, and eligible generated sequences served as rejected responses when their mean pLDDT was at least five points lower than that of the natural sequence. Up to three rejected responses were retained per prompt, selecting the lowest-confidence eligible candidates.

Each round continued from the preceding adapter, while the reference policy remained the original supervised model. Training used the sigmoid DPO objective with $\beta = 0.5$ and an additional preferred-response negative-log-likelihood term with coefficient 1.0. The recorded configuration used rank-32 LoRA adapters with scaling parameter 16 and dropout 0.05, a learning rate of 5×10^{-5}, an effective batch size of eight, and 60 optimisation steps per round. Accordingly, this experiment evaluates anchored preference optimisation with a supervised auxiliary term, rather than an unregularised DPO objective.

Specification gate. The gate evaluates the requested sequence-derived properties and requires a valid amino-acid sequence. Requested length must be within $\pm 10\%$ of its target and net charge within ± 2 charge units. Cysteine counts must match exactly for targets of zero, one, or two; larger targets allow a deviation of one residue. Stability bins must match exactly. For hydrophobicity and aggregation, an adjacent ordinal bin is permitted, but a low-high mismatch is rejected. The gate additionally requires an aggregate constraint score of at least 0.70. These criteria differ from the exact-bin and exact-cysteine adherence measures used in the broad supervised evaluation. Gate passage therefore does not mean exact agreement with every requested descriptor.

Sequence filters. Sequence filters require at least 20 canonical amino-acid residues, an amino-acid Shannon entropy of at least 3.2 bits, and no individual residue type comprising more than 23% of the sequence. The longest contiguous tandem-repeat region, considering repeat periods of one to five residues, must cover no more than 25% of the sequence. These filters screen for low complexity, residue overrepresentation, and short-period repetition; they do not establish biological function or experimental foldability.

Population-level evaluation. Evaluation used one generated response per held-out prompt, without best-of-k selection. We separately measured specification-gate passage, sequence-filter passage, their conjunction, and the fraction reaching ESMFold pLDDT ≥ 70. The strict composite required gate passage, filter passage, and this confidence threshold simultaneously; it did not additionally require a Pfam-domain hit. Binary population rates used all assigned prompts as the denominator, with missing or unsuccessful predictions not counted as successes. Mean pLDDT was calculated over responses with an available confidence estimate. ESMFold confidence was interpreted as a prediction-confidence measure, not as experimental evidence of folding or function.

Domain recovery and structural similarity. For the family-specific experiments, requested-domain recovery was assessed with HMMER using the requested Pfam profile and its curated gathering threshold (`--cut_ga`). This criterion differs from the E-value threshold analysis used in the broad supervised-generation benchmark. For the structural comparisons, generated and accession-matched natural sequences were predicted with ESMFold. Their Cα coordinates were aligned using TM-align through `tmtools`, with TM-score normalised by the natural reference length. We report the fraction of all test prompts with TM-score ≥ 0.5 as a thresholded structural-similarity measure; comparisons without an available alignment did not count as successes.

Cross-family transfer and direct training. Each source-family adapter was evaluated on all six transfer targets. The transfer analysis used round 2 for PF00237 and round 3 for PF00467 and PF03652, retaining these checkpoints across both seeds. Heatmap entries report the mean of the two seed-specific percentage-point changes relative to the supervised baseline. The designated nearby source-target pairs were determined from the family embedding, rather than from observed transfer performance.

We compared target-family preference optimization from general supervised Aria (cold start) and from a nearby-family adapter (warm start) on PF01196, PF00366, and PF02075. Preference-training/validation/test counts were 269/115/168, 284/69/154, and 283/78/72, respectively. Warm starts used PF00237 round two, PF00467 round three, and PF03652 round three, respectively. Each configuration underwent three optimization rounds using the target family's natural sequences as preferred responses, with two runs per configuration. The DPO reference remained fixed at general supervised Aria for cold starts and at the initial nearby-family adapter for warm starts. The preference objective and optimization settings otherwise followed those described above. Evaluation used one generation per held-out prompt, and endpoint comparisons used round three for both configurations.

Protein-language-model comparison. ProLLaMA was evaluated through its only generation interface, [Generate by superfamily] Superfamily=<name>, using the superfamily strings shipped with the model. Families were mapped to that vocabulary on an exact match only, checked against the model's own superfamilies.txt: PF00467, PF03652 and PF02075 have no entry, and PF00366 appears only as the eukaryotic eS17 paralog rather than the bacterial uS17 requested here, so all four were excluded rather than substituted with a near neighbour. Generation used the decoding settings from ProLLaMA's own inference script, with the temperature raised from its default of 0.2 to the 0.7 used for Aria throughout, so that neither model is sampled more conservatively than the other. One sequence was generated per evaluation case. Sequences were folded with the same ESMFold model and scored with the same family evaluators as Aria's. Because the property targets cannot be supplied through this interface, ProLLaMA's gate and composite rates measure incidental agreement with those targets rather than an ability to follow them.

Uncertainty. Where shown for individual population rates, error bars represent 95% Wilson binomial confidence intervals. These intervals describe uncertainty across evaluated prompts and do not estimate variability across training seeds. Two-seed heatmap averages are descriptive summaries; their signs alone do not establish statistical significance.

B DESCRIPTORS

Thirteen descriptors, with how each specification value is obtained and how adherence to it is measured. "Presence" is the fraction of held-out specifications in which a descriptor appears, a consequence of partial sampling. Every value is computational; no descriptor is experimentally measured.

Table A1: Descriptors used for specification conditioning and adherence evaluation. Presence denotes the percentage of held-out specifications containing each descriptor.

Descriptor	Type	Presence	Specified from	Adherence scored by
Physicochemical				
Length	exact integer	54.6%	len(seq)	same function
Net charge (pH 7.4)	exact integer	53.9%	Biopython	same function
Hydrophobicity (GRAVY)	3 levels	54.3%	Biopython	same function
Stability	2 levels	54.2%	Biopython	same function
Developability				
Aggregation propensity	3 levels	54.1%	AGGRESCAN scale	same function
Cysteine count	exact integer	53.8%	count(C)	same function
Structural and topological				
Structural confidence	4 levels	52.8%	AlphaFold DB	not scored
Transmembrane helix	binary	54.2%	UniProt annotation	DeepTMHMM
Signal peptide	binary	54.2%	UniProt annotation	SignalP 6
Functional				
Domain (Pfam)	categ. (2,204)	51.1%	UniProt curator	hmmscan/Pfam-A
Enzyme class (EC)	categ. (60)	25.4%	UniProt curator	CLEAN (confounded)
Molecular function (GO)	categorical	45.3%	UniProt curator	not scored
Organism	categorical	53.7%	UniProt metadata	not scored

C TRAINING FOOTPRINT

Stage 1 is 348,867 examples over 21,804 optimizer steps (two epochs) on two A100-80GB GPUs, $\sim$35 GPU-hours total, updating 83.9M parameters ($\sim$1% of the model). Stage 2 is 929 preference pairs over 60 steps on one GPU with $\beta = 0.5$ and $\alpha_{\text{RPO}} = 1.0$; its optimizer cost is negligible beside its oracle cost of 19,904 sampled candidates and ESMFold on the survivors.

Constraint score. The constraint score is the arithmetic mean of satisfaction scores over the sequence-derived properties present in a specification. Hydrophobicity, stability, and aggregation receive one for an exact bin match and zero otherwise. For realised length L and target $L^{\star}$, the

length score is one within $\pm 10\%$ of the target and $\max(0, 1 - |L - L^\star|/L^\star)$ otherwise. For realised and requested net charges q and $q^\star$, the score is $\max(0, 1 - |q - q^\star|/10)$. For cysteine counts c and $c^\star$, it is $\max(0, 1 - |c - c^\star|/5)$. Invalid sequences receive zero; valid sequences with no scorable constraints receive one. Descriptors requiring external predictors do not contribute. The aggregate threshold is applied in addition to the gate's individual property checks, not in place of them.

Anchoring fixes the preferred response to a natural sequence, but does not guarantee preservation of every conditioned property. With both responses drawn from the model, the DPO objective depends only on the *difference* of the two log-ratios. The preference margin can increase even when the probabilities of both responses decrease. Diversity after the anchored round does not contract: mean within-set pairwise identity is 0.440 against 0.557 for the natural held-out set, and identity to a consensus built from the training sequences is 0.568 against 0.647.

Because the policy was supervised-trained on the same natural responses later used as preferred examples, memorization is a live concern. Mean identity to the nearest training sequence rises from 0.184 after supervised fine-tuning to 0.354, reaching 0.658 among the 19 generations that pass the full conjunction, and 2 of 255 outputs with an extractable sequence are at least 90% identical to a training sequence. A control on the natural family settles the interpretation: real held-out members of PF00416 sit at 0.707 mean identity to their nearest training protein, and 29 of 260 (11%) exceed 90%. The family is highly conserved, so genuine homologues are already this similar to one another. Generated sequences are therefore less similar to the training set than real proteins of the same family are, and cross the near-copy threshold an order of magnitude less often.

D HELD-OUT SPECIFICATIONS BEFORE AND AFTER SUPERVISED TUNING

Figure A1 shows three held-out specifications with the base model's output and Aria's, under identical decoding. They are chosen to cover the three ways the base model fails to return a usable sequence: it wraps one in commentary, it hedges and then emits a low-complexity string, or it declines the request. Base excerpts are trimmed only at the marked ellipses.



Figure A1: Base model against Aria on three identical held-out specifications. The base model wraps a 410-residue sequence in commentary and an invented citation while reporting it as the requested 237 (Example 1), hedges and then emits 354 residues drawn from 9 of the 20 amino acids (Example 2), or declines a benign enzyme specification outright (Example 3).

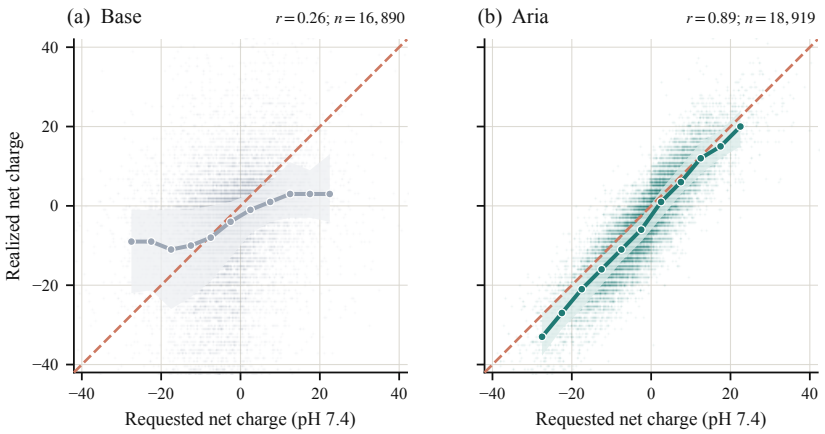


Figure A2: Requested versus realized net charge at pH 7.4 for the base model (a; $n = 16{,}890$) and Aria (b; $n = 18{,}919$). Faint points represent individual valid generations; integer-valued requests are jittered horizontally for visualization only. Solid lines show the median realized charge within requested-charge bins, and shaded bands show the corresponding interquartile ranges. The dashed line denotes exact agreement ($y = x$). Pearson correlations are calculated using all valid charge-requesting outputs, including observations outside the displayed range.

The format change is not the whole of what supervised tuning buys, but it is a precondition for measuring anything else: adherence is recomputed from the returned sequence, and a sequence that has to be recovered from prose by a parser cannot be graded on the terms the specification was written in. For Example 3, recomputing the four properties that are determined by sequence alone gives length 235 against a requested 236, hydrophobicity moderate, stability stable and 0 cysteines, all as requested. The remaining requested properties in these three specifications, Pfam domain, signal peptide, and structural confidence, are assessed with the external predictors described in Section 2 and are reported in the main results.

E Adherence to charge

Net charge provides a stricter test of numerical conditioning than the binned physicochemical properties because the prompt specifies an integer target and the realized value is recomputed directly from the generated sequence. Requested and realized charge correlate at $r = 0.89$, against 0.26 for the baseline (Figure A2), so asking for a more positively charged protein reliably returns one. It does not, however, arrive at the value asked for, missing by 4.70 charge units on average. That shortfall is systematic rather than random.

Regressing realized on requested charge gives realized $\approx 1.10 \times$ requested $- 2.14$: the model overresponds slightly to changes in the request, and its outputs sit low across most of the range. Both terms are fixed biases, so part of the error is miscalibration rather than a failure to follow the request. The baseline fails in the opposite direction, a slope of 0.567 meaning it moves only half as far as it is asked to.

F Domain adherence against significance

$N = 17{,}941$ fine-tuned generations whose specification requested a domain. The last row is `hmmscan`'s default reporting threshold with no further filtering, and is the 2.90% quoted in the text. The baseline records 0 successes in 15,140 generations at every threshold.

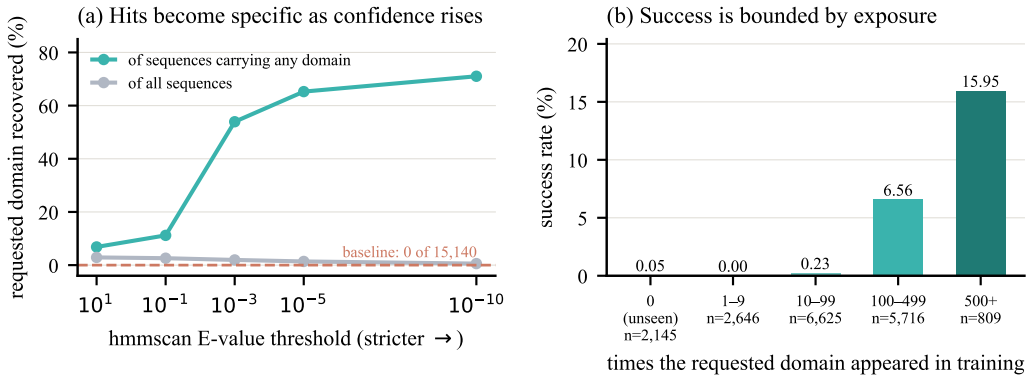


Figure A3: Domain adherence is rare, targeted, and bounded by exposure. (a) As the hmmscan threshold tightens, fewer generations carry any significant assignment but the probability such an assignment is the requested one rises to 71%; the baseline records zero successes at every threshold. (b) Attempt-weighted success against how often the requested domain appeared in training.

Table A2: Pfam-domain recovery across hmmscan reporting thresholds. Lower E-values indicate stronger sequence-profile matches. Stricter thresholds reduce coverage but substantially increase the specificity of the retained assignments.

E-value	Any significant assignment	Requested domain	P(correct | any)
10^{-10}	145 (0.8%)	0.57%	71.0%
10^{-5}	380 (2.1%)	1.38%	65.3%
10^{-3}	649 (3.6%)	1.95%	53.9%
10^{-2}	1,206 (6.7%)	2.23%	33.2%
10^{-1}	4,173 (23.3%)	2.60%	11.2%
10^{1} (default)	7,607 (42.4%)	2.90%	6.8%

G Enzyme-class adherence is confounded by the predictor

Across the 32 EC classes requested at least 50 times, the correlation between how frequently CLEAN emits a class over the whole evaluation and that class's apparent success rate is $r = 0.96$; the value is stable between 0.90 and 0.96 for every class-size cutoff we tried. EC 7.4 was requested 129 times and never predicted for any sequence, so its zero was guaranteed by the predictor; EC 1.2 was requested 609 times and predicted three times. Conversely EC 3.2 was predicted roughly fourteen times more often than requested and shows 46.6% apparent success. We therefore exclude enzyme class from the reported adherence results, and report the diagnostic rather than the number.

H The constraint ladder

To isolate prompt load from target difficulty, we constructed a nested constraint ladder for each of 200 cluster-separated held-out proteins. For each, we build thirteen prompts. All thirteen prompts for a protein shared the same length target. The first requested length alone, and each subsequent prompt added one further descriptor, yielding zero to twelve additional properties. Each model generated one sequence per prompt, and length adherence was defined as falling within $\pm 10\%$ of the fixed target ($n = 200$ per rung per model). Each rung is 200 generations, 2,600 in total. Adherence stays between 84.5% and 92.0% across the ladder. The base model stays near its unconditioned rate throughout. This experiment is specific to preservation of the fixed length target; it does not establish that every conditioned property is equally insensitive to prompt load.

Table A3: Length adherence across the nested constraint ladder. Each rung adds the listed descriptor to the preceding specification while retaining the same length target ($n = 200$ cluster-separated held-out proteins per rung and model; one generation per prompt). Values are the percentages of generations within $\pm 10\%$ of the requested length.

Descriptors	Added at this rung	Aria	Base
1	(length target only)	90.0%	7.0%
2	net charge	86.5%	7.0%
3	hydrophobicity	84.5%	9.0%
4	stability	86.5%	9.5%
5	aggregation	90.5%	6.0%
6	cysteine count	89.5%	7.5%
7	structural confidence	90.5%	14.0%
8	transmembrane helix	92.0%	10.0%
9	signal peptide	92.0%	9.0%
10	Pfam domain	91.5%	7.5%
11	EC class	90.0%	8.5%
12	molecular function	90.5%	8.5%
13	organism	87.5%	7.0%

I FAMILY SELECTION AND EMBEDDING DISTANCE

Source and target families were selected before any preference training. Family embeddings were averaged from per-sequence ESM2 representations and clustered, and cluster medoids were screened on solubility, sequence length, the number of available preference prompts, and the predicted confidence of their natural members. Six of 600 medoids met these criteria. The three retained as sources were those with the largest clusters among medoids having at least one cluster-mate that also met the target criteria. Targets were the nearest qualifying cluster-mates by embedding distance. Eight of the ten resulting families are ribosomal proteins, which follows from the joint requirement that a family be short, soluble and well represented in Swiss-Prot rather than from a design choice.

Figure A4a shows the selected families within the embedded corpus. The eight ribosomal families occupy a narrow region, and the nuclease pair PF03652-PF02075 lies elsewhere.

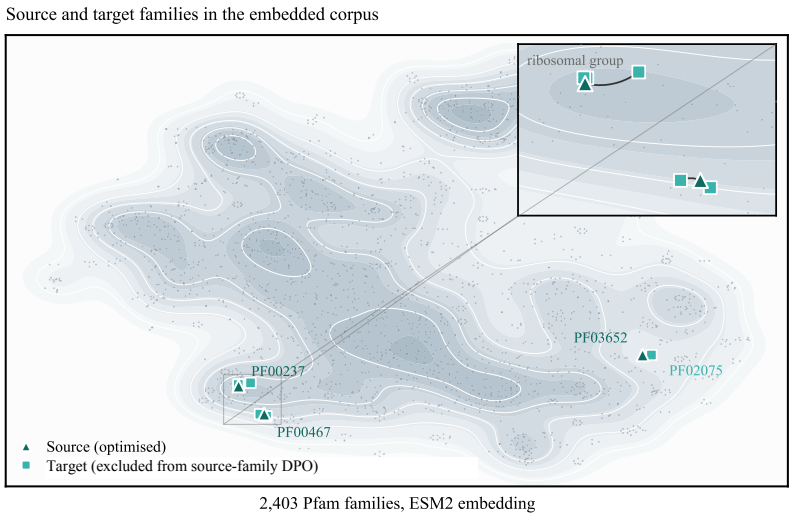


Figure A4: Family selection and embedding distance. (a) The 2,403 Pfam families present in the corpus, projected to two dimensions from averaged ESM2 embeddings, with the three source families, their six targets, and the source-target links. The inset enlarges the region containing the ribosomal families.

J REWARD HACKING IN SELF-PLAY PREFERENCE OPTIMISATION

An initial PF00416 experiment tested whether preferences between model-generated sequences could improve predicted structural confidence while preserving specification adherence. Starting from the supervised checkpoint, we sampled candidate sequences for each training specification and retained those passing the specification gate and sequence filters. Among eligible candidates, higher ESMFold pLDDT determined the preferred response. This self-play configuration used 859 preference pairs from 310 training prompts, DPO coefficient $\beta = 0.3$, and no auxiliary likelihood term.

Despite screening the training candidates, the resulting policy exhibited reward hacking: the optimised confidence metric improved while performance on the intended generation task deteriorated. In a matched evaluation on 260 held-out prompts, mean pLDDT increased from 49.0 to 67.9, whereas gate passage decreased from 19.2% to 1.2% and sequence-filter passage from 88.8% to 27.3% (Table A4). Mean sequence length increased from 138 to 350 residues, and mean absolute charge error increased from 6.6 to 103.0 charge units. The strict composite remained near zero. These results show that screening preference-training candidates did not ensure that subsequent generations retained the same properties.

The highest-confidence self-play generation illustrates this failure (Figure A5). Although ESMFold assigned a mean pLDDT of 93.8, its predicted structure was dominated by an extended helix. The sequence contained 194 residues, including 68 lysines, despite a requested length of 126 residues, and failed both the specification gate and sequence filters. For the same prompt, anchored DPO generated 124 residues with mean pLDDT 71.7 and passed both checks. Thus, the higher confidence of the self-play output did not indicate better compliance with the requested design.

To address this failure, we trained an alternative configuration from the supervised checkpoint, replacing the preferred generation with the matching natural sequence, increasing β to 0.5, and adding a preferred-response negative-log-likelihood term with weight 1.0 (Pang et al., 2024). This anchored configuration used 929 pairs from 310 training prompts. Self-play and anchored training used 107 and 60 optimisation steps, respectively, with the supervised policy held fixed as the reference.

The anchored configuration reversed the observed collapse: gate passage reached 22.7%, sequence-filter passage reached 96.2%, and 15 of 260 generations satisfied the strict composite. Mean length returned to 130 residues and charge error decreased to 4.0 units. Mean pLDDT was 56.5, lower than the self-play value but accompanied by substantially higher adherence and filter passage. The preferred-response source, loss configuration, and training duration differed between the two configurations; this comparison therefore does not isolate the contribution of natural-sequence anchoring. It establishes that the bundled intervention mitigated the observed reward-hacking failure in this setting, not that it prevents all such failures.

This evaluation was conducted separately from the PF00416 comparison in Section 2.3. Outputs from the supervised and anchored checkpoints were generated separately within each evaluation run. The comparisons here therefore use the matched values in Table A4, rather than combining results across runs.

Table A4: Matched PF00416 evaluation of self-play and anchored preference optimisation. Each model produced one generation for each of the same 260 held-out prompts in a common evaluation run. Gate and filter values are percentages over all prompts. The strict composite requires gate passage, sequence-filter passage, and pLDDT $\geq$ 70; entries give counts and percentages. Mean pLDDT excludes unparseable outputs.

Configuration	Mean pLDDT	Gate (%)	Filters (%)	Strict composite
Supervised	49.0	19.2	88.8	0/260 (0.0%)
Self-play DPO	67.9	1.2	27.3	1/260 (0.4%)
Anchored DPO	56.5	22.7	96.2	15/260 (5.8%)

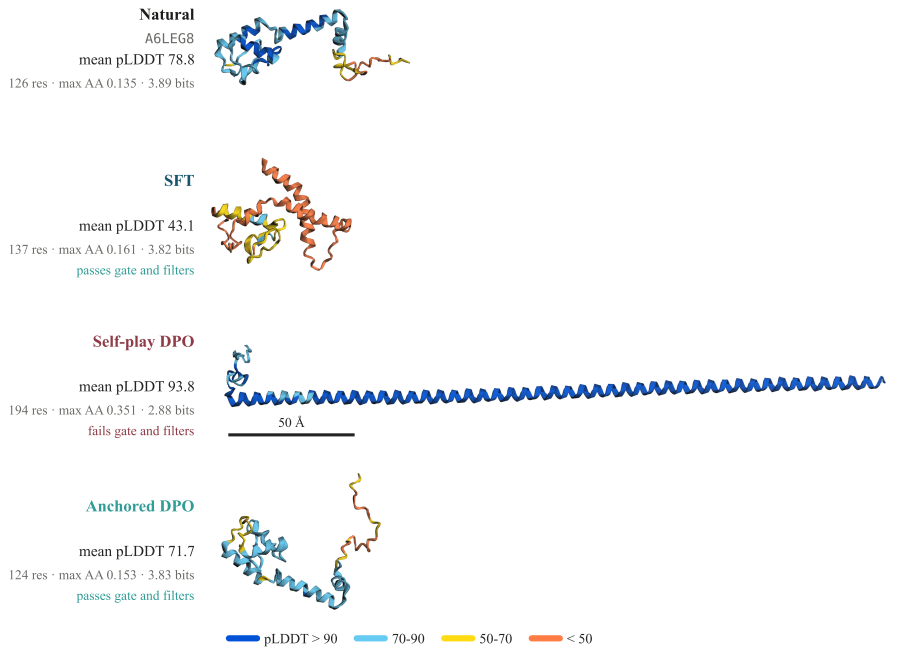


Figure A5: A high-confidence failure under self-play DPO. Natural, supervised, self-play, and anchored structures correspond to the same held-out specification, A6LEG8. The example was selected as the highest-confidence self-play generation, not as a representative sample. Structures share a common spatial scale and are coloured by per-residue ESMFold confidence. Despite its higher pLDDT, the extended, lysine-rich self-play output fails both the specification gate and sequence filters. Records are from the matched evaluation reported in Table A4.

K Distributions of structural similarity

Section 2.3 reports the fraction of generations reaching TM-score ≥ 0.5, a threshold conventionally taken to indicate a shared fold. That summary does not describe how the population is distributed around it. Figure A6 shows the full TM-score distributions by round.

In all three families the distribution is bimodal, with one mode near 0.2, corresponding to generations structurally unrelated to the natural reference, and a second above 0.5. Successive rounds redistribute mass between these two modes rather than shifting the population as a whole. In PF00467 and PF03652 the upper mode becomes dominant; in PF00237 it remains the smaller of the two at every round, which is consistent with that family's lower threshold rates in Figure 3.

L Worked examples

Two of the nineteen held-out uS13 specifications whose generations satisfied the full conjunction are shown below. Realized values are recomputed from the generated sequences with the functions used to construct their specifications.

B1MVZ2: 4 properties requested, all scorable ones met (confidence 79.1). Requested length 123 (realized 119, within $\pm 10\%$) and cysteine count 0 (0); aggregation low (low); organism *Leuconostoc citreum* (strain KM20) was conditioned on but not scored.

Generated sequence:

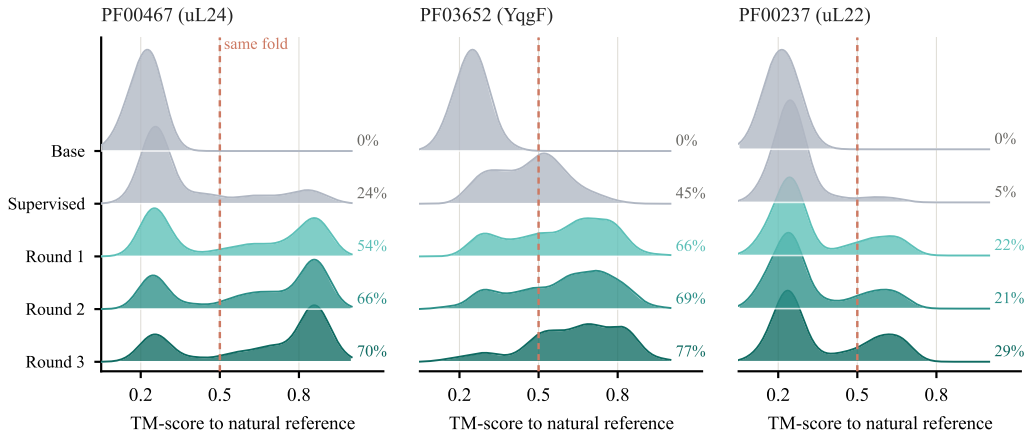


Figure A6: TM-score distributions against the natural reference, by optimization round, pooling both training seeds. The dashed line marks TM-score 0.5. The densities cover generations that produced an alignment. The percentages at the right of each ridge are over all held-out prompts. Base and supervised rows are shown in grey.

```
MARIAGVNIPTNKRVVIALQYIHGIGTKFAQEILEKVGIPAERRVNQLTDAEVLQIRETI
DRDYQVEGDLRRETSMNIKRLMDLGSYRGLRHRRGLPVRGQRTKTNARTRKGPRKPIKK
```

B8I805: 8 properties requested, all scorable ones met (confidence 78.2). Requested length 122 (realized 121), hydrophobicity low (low), stability stable (stable), aggregation low (low), and cysteine count 1 (1); structural confidence high, domain PF00416, and molecular function rRNA binding were conditioned on but not scored.

Generated sequence:

```
MARIAGVNIPTNKRVEIALTYIHGIGRFVAGQIIEKVGIPAERRVNQLTDADIDKLREEV
GKFIVEGDLRREVAMNVKRLMDLGCYRGLRHRRGLPVRGQRTKTNARTRKGPRKTVAGKKK
```

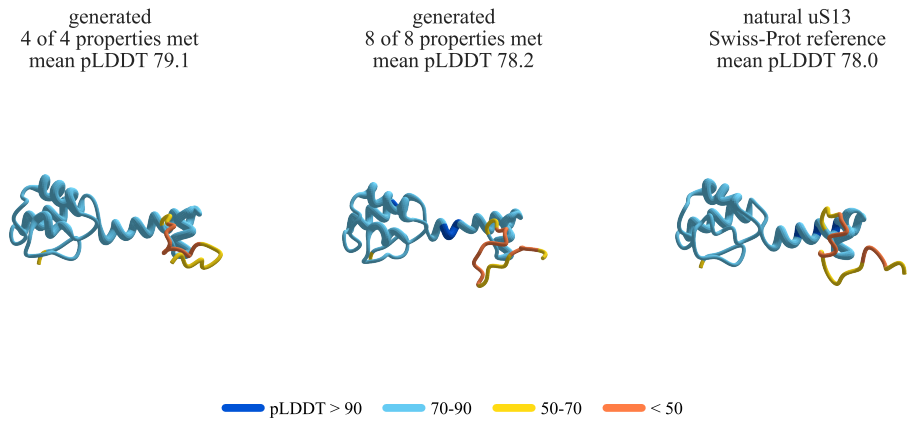


Figure A7: Predicted structures for the two generations described above, beside a natural uS13 sequence from the same family. Spline-smoothed Cα tubes are thickened through helices and strands and colored by per-residue ESMFold confidence; all three structures are shown along their principal axes under identical lighting.

M POPULATION VIEW OF THE ANCHORED ROUND

Figure A8 shows every held-out generation for PF00467 at each stage of training, together with the natural members of the same held-out set. No generation is selected or reordered: each grid contains all 211 prompts in a fixed order, so the figure describes the population rather than representative cases.

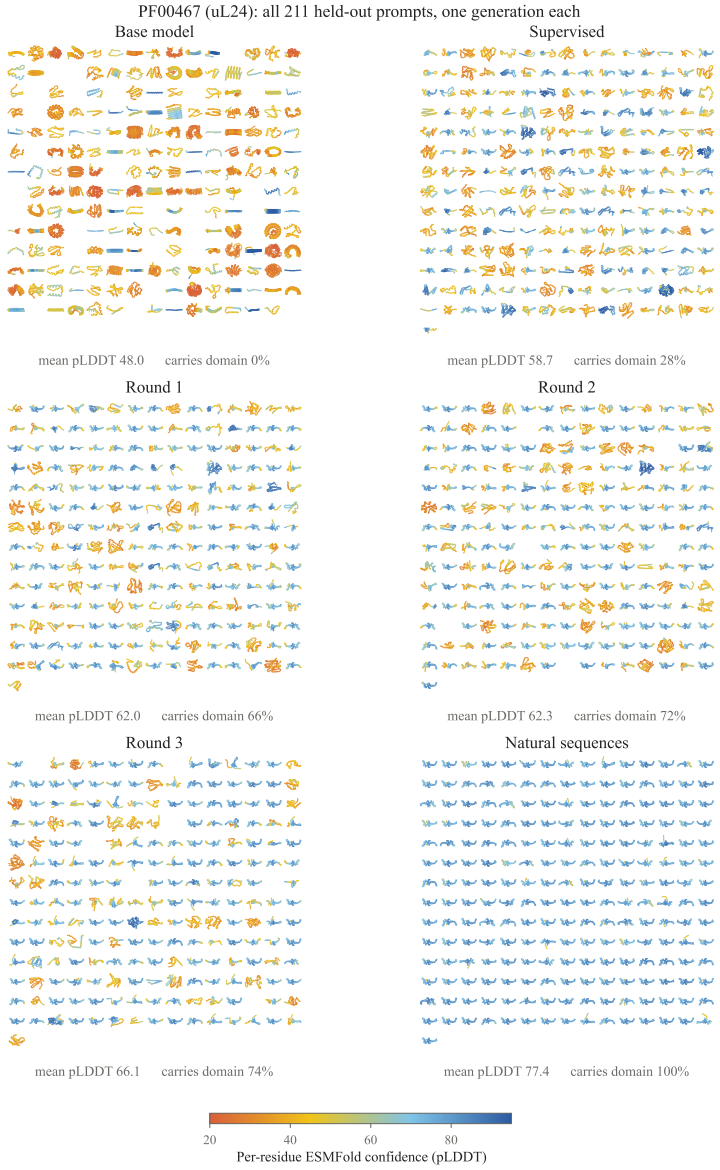


Figure A8: All 211 held-out PF00467 prompts for the run labelled seed 1001, with one generation per prompt, at the base model, after supervised tuning, after each of three anchored rounds, and for the natural members of the same held-out set. Backbones are Cα traces oriented by their own principal axes and colored by per-residue ESMFold confidence. Each panel is captioned with its mean predicted confidence and its requested-domain recovery rate. Grids are complete and unordered with respect to outcome.

Predicted backbones from the base model are structurally heterogeneous and receive low per-residue confidence. After supervised tuning the sequences shorten toward the family's length range and confidence increases, but the predicted structures remain varied. Across successive anchored rounds

the population converges toward a common backbone arrangement resembling that of the natural members, and mean confidence rises from 48.0 to 66.1 against 77.4 for the natural set. Requested-domain recovery over the same grids rises from 0% to 74%, against 100% for the natural members.