PHYSIO-DPO: ALIGNING PROTEIN LANGUAGE MODELS WITH LOCAL ENERGY GAPS TO MITIGATE STRUCTURAL HALLUCINATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein language models generate sequences that score well under their own likelihood yet fold into structures an empirical energy function rates as unstable. Direct Preference Optimization cannot separate such a case from a marginal one, because a binary label discards how large the violation is. We propose **Physio-DPO**, which recovers that magnitude. For a backbone-matched pair consisting of a native structure and a physics-perturbed decoy of the same fold, a thresholded and bounded function of the local energy gap scales the DPO update, so severe stereochemical violations drive larger updates than ambiguous ones. Pairs of unrelated folds keep the ordinary binary objective, so no energy magnitude is ever compared across folds. We build PhysioPref-1M, one million preference pairs that combine generation-mined preferences with physics-perturbed hard negatives, and verify the labels by blinded expert review. Across three seeds Physio-DPO reaches an sc-RMSD of 1.32 ± 0.04 Å and a foldability of $92.1 \pm 0.8\%$, ahead of SFT, PPO, DPO, and the strongest energy-aware baseline, and it keeps that lead under folding and scoring oracles unseen during training. The gains reach experimental data as well: on ProteinGym deep mutational scanning assays Physio-DPO raises the average Spearman correlation from 0.67 to 0.72.

1 INTRODUCTION

Protein language models (PLMs), including the ESM series (Lin et al., 2023; Hayes et al., 2024) and ProGen (Madani et al., 2023), have expanded the scope of computational protein design. Trained on large sequence collections, these models learn evolutionary regularities and generate protein-like sequences *de novo*. Their likelihood objectives, however, do not directly assess the geometry or energy of a predicted fold. A sequence may therefore appear plausible to the model while its predicted structure contains steric clashes, poor packing, or exposed hydrophobic residues (Anishchenko et al., 2021; Gopalan & Narayanan, 2025). We use *structural hallucination* to denote this operational mismatch, not to imply experimentally measured instability.

Direct Preference Optimization (DPO) aligns a model offline from ordered preference pairs (Rafailov et al., 2023). Effective as it is in general settings (Das et al., 2025), it encodes only the preference direction, giving a marginal difference and a severe structural violation equal weight. Reference-free variants such as SimPO (Meng et al., 2024) and ORPO (Hong et al., 2024) change the regularization strategy without resolving how magnitude should enter protein LM post-training. AliDiff (Gu et al., 2024) is the closest prior work: it introduced energy-difference-weighted preference optimization for biomolecular generative models, encoding exact energy differences as soft Bradley–Terry targets for target-conditioned ligand diffusion. Two questions remain open for autoregressive protein LMs. An energy gap is informative only when the paired structures are physically comparable, since raw differences across unrelated folds share no global scale. And it is unclear whether the magnitude should set the *target* of the preference loss or its *weight*.

We answer both with **Physio-DPO**. Its fold-gated objective applies a thresholded and bounded energy-gap weight only to backbone-matched native–decoy pairs produced by physical perturbation, so the magnitude signal is used exactly where a within-fold comparison is defined, and every

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.

remaining pair keeps the ordinary binary DPO term. We further construct PhysioPref-1M by combining generation-mined preferences with physics-perturbed hard negatives.

Our contributions are fourfold. **(1)** PhysioPref-1M provides one million preference pairs from two complementary construction routes, supported by pair-level geometry and blinded expert checks. **(2)** Physio-DPO introduces fold-gated, thresholded, and bounded amplification without comparing energy magnitudes across unrelated folds. **(3)** We prove a bounded conditional gradient-modulation property: holding the model state and the pair fixed, the per-pair gradient contribution increases strictly with the local energy gap and stays uniformly bounded, which is what separates our weight from the unbounded linear and margin-offset alternatives. **(4)** Across three random seeds, Physio-DPO achieves an sc-RMSD of 1.32 ± 0.04 **Å** and a predicted foldability of $92.1 \pm 0.8\%$, outperforming matched preference-optimization baselines, and it retains its lead under oracles unseen during training and against experimental fitness labels. All structural evaluation is computational, with experimental evidence entering through ProteinGym fitness measurements.

2 Related Work

Generative Protein Models and Structural Hallucinations. Protein generation began with family-specific alignment statistics such as Potts models (Levy et al., 2017). Transformers (Vaswani et al., 2017) then enabled autoregressive PLMs trained on large unaligned collections (Rives et al., 2021; Elnaggar et al., 2021). The ESM family (Hayes et al., 2024; Lin et al., 2023) and ProtGPT2 (Ferruz et al., 2022) capture long-range evolutionary regularities, yet their token-level objectives never evaluate the geometry or energy of a predicted fold. Statistically plausible sequences may therefore yield predicted structures with steric clashes, poor packing, or unsatisfied hydrogen bonds (Anishchenko et al., 2021). We use *structural hallucination* for this operational mismatch. Structure-generative diffusion models instead model backbones or atomic coordinates directly (Watson et al., 2023; Ingraham et al., 2023; Lisanza et al., 2025), whereas we study post-training of autoregressive sequence generators with structural feedback.

Preference Optimization for Protein Models. Language-model alignment commonly uses reinforcement learning from human feedback (Zhou et al., 2025; Mei et al., 2025), often optimized with Proximal Policy Optimization (PPO) (Schulman et al., 2017). Reinforcement learning has also been applied to biological sequence design (Angermueller et al., 2020), although online optimization can be resource-intensive. Direct Preference Optimization (DPO) provides an offline alternative based on ordered preference pairs (Rafailov et al., 2023). IPO (Azar et al., 2024) and KTO (Ethayarajh et al., 2024) modify the preference formulation, while SimPO (Meng et al., 2024) and ORPO (Hong et al., 2024) remove the explicit reference model. In molecular generation, AliDiff extends preference optimization to target-conditioned ligand diffusion models through Exact Energy Preference Optimization (E2PO), which incorporates energy-derived reward differences into soft preference targets to mitigate over-optimization (Gu et al., 2024).

3 Methodology

Standard DPO is stable, but it treats preferences as binary and discards magnitude (Liu et al., 2025). Where an energy gap measures how badly a structure is violated, that discards the signal saying which errors matter most. **Physio-DPO** keeps it. As Fig. 1 shows, the method has two stages: constructing a preference dataset whose pairs carry a meaningful energy gap, through a Generate–Fold–Score pipeline (Sec. 3.1); and optimizing a magnitude-aware objective that scales each backbone-matched pair's DPO gradient by its local gap (Sec. 3.2).

3.1 PhysioPref-1M Benchmark

Effective preference alignment in protein design requires dense, physically grounded supervision, which is largely absent from existing instruction-tuning datasets. To this end, we introduce PhysioPref-1M, a large-scale preference benchmark comprising 1M protein pairs annotated by energy-based criteria. The dataset is constructed via an adversarial generation and filtration pipeline (Fig. 2) that deliberately induces and identifies hard negatives—structures that appear foldable yet

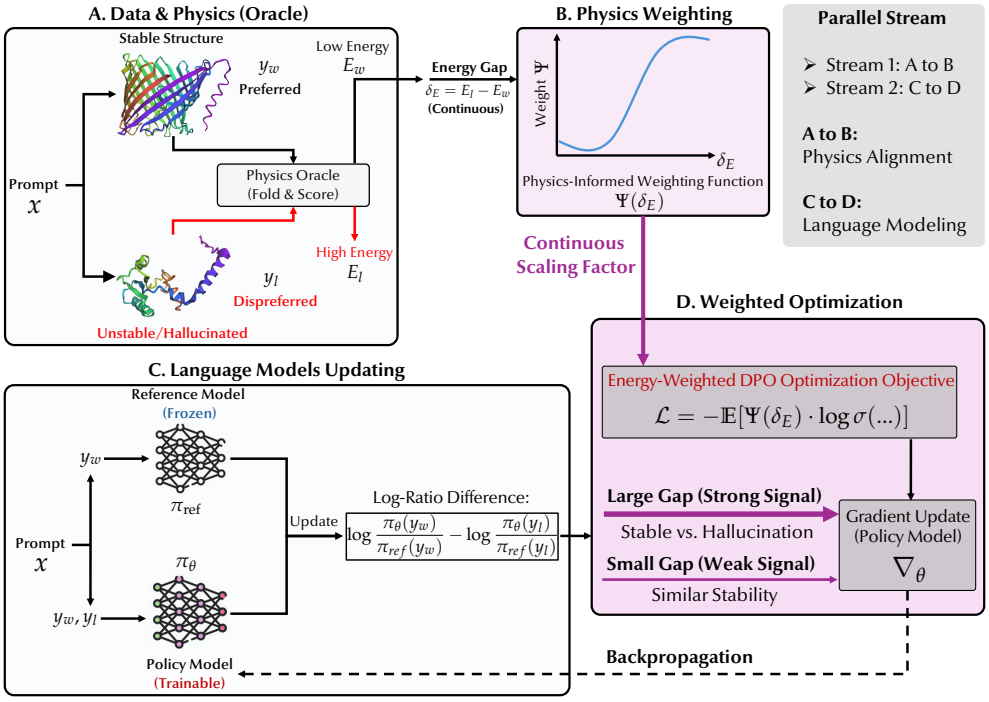

Figure 1: Overview of Physio-DPO. The physics stream folds a preference pair (y_w, y_l) and computes the local energy gap δ_E, which is mapped to a bounded weight $\Psi(\delta_E)$. The language modeling stream computes the DPO log ratio using the policy π_θ and a frozen reference π_{ref}. Physio-DPO reweights the DPO term of each backbone-matched pair by $\Psi(\delta_E)$, so pairs with larger local stability gaps produce larger bounded updates; generation-mined pairs keep the ordinary binary term.

violate physical stability. Preference pairs are formed by contrasting stable proteins against unstable or pathological decoys, ensuring informative energy gaps. A human-in-the-loop evaluation further validates the reliability of the automated labeling. Full dataset details are provided in Appendix D.

3.2 Optimization Objective

Let $\mathcal{E}(y)$ denote the per-residue score of a fixed empirical energy function evaluated on the predicted structure of y. For a preference pair (y_w, y_l) we define the *local* energy gap

$$\delta_E(y_w, y_l) = \mathrm{ReLU}\big(\mathcal{E}(y_l) - \mathcal{E}(y_w)\big) \in [0, \infty), \tag{1}$$

and read it as the severity of the violation that separates the two members of that pair, not as a position on a global stability scale. Standard DPO keeps only the sign of this quantity. We recover its magnitude through a weighting function $\Psi : [0, \infty) \to (0, \lambda_{\max})$,

$$\Psi(\delta_E) = \lambda_{\max} \cdot \sigma\left(\frac{\delta_E - \mu}{\tau}\right), \tag{2}$$

where μ places the transition at the energy gap above which a violation is treated as severe, τ sets how sharp that transition is, and $\lambda_{\max}$ caps the weight. Two properties of Eq. 2 carry the design. It is *thresholded*: for $\delta_E \ll \mu$ the weight falls to $\lambda_{\max}\sigma(-\mu/\tau)$, which at our setting of $\mu = 50$ and $\tau = 10$ is smaller than $\lambda_{\max}$ by a factor of about 150, so near-tied pairs are strongly down-weighted rather than removed. It is also *bounded*: no pair can ever exceed $\lambda_{\max}$, however large its gap, which is what separates Eq. 2 from the linear and margin-offset alternatives whose influence grows without limit on a heavy-tailed energy distribution.

Fold Gating. A magnitude is only meaningful where the two structures are comparable. PhysioPref-1M contains two kinds of pairs: physics-perturbed pairs $\mathcal{D}_{\text{phys}}$, in which a native structure is contrasted with a PyRosetta decoy built from the same backbone, and generation-mined pairs

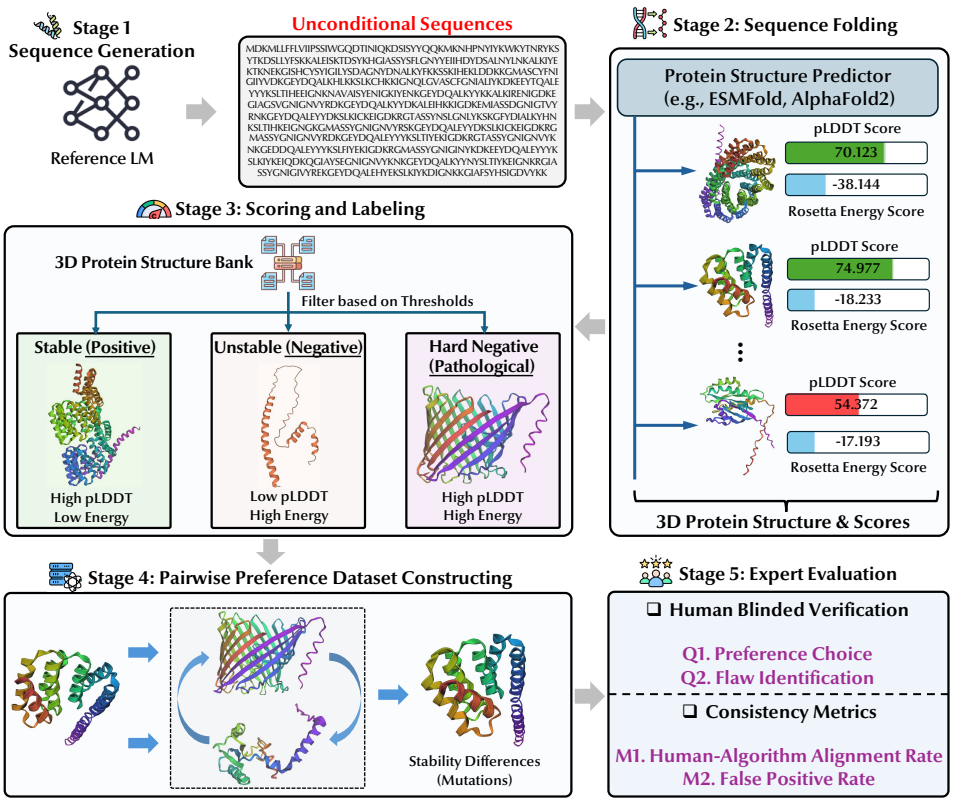

Figure 2: Construction pipeline for PhysioPref-1M. **Step 1:** diverse sequence generation from a reference language model (Ferruz et al., 2022); **Step 2:** structure prediction via folding; **Step 3:** scoring and labeling based on pLDDT confidence (Fang et al., 2025) and Rosetta energy scores (Alford et al., 2017), including the identification of hard negatives with high confidence but poor stability; **Step 4:** construction of preference pairs that maximize stability gaps; and **Step 5:** human-in-the-loop evaluation to verify alignment between labels and biophysical judgment.

$\mathcal{D}_{\text{gen}}$, which are sampled independently and typically hold unrelated folds (median TM-score 0.93 versus 0.28–0.31; Appendix G.5). We therefore gate the weight on pair construction,

$$w_i = \begin{cases} \Psi(\delta_{E,i}), & i \in \mathcal{D}_{\text{phys}}, \\ 1, & i \in \mathcal{D}_{\text{gen}}, \end{cases} \tag{3}$$

so that every energy magnitude entering the loss is a within-fold quantity and every cross-fold pair contributes an ordinary binary DPO term.

The Energy-Weighted Objective. Writing $\Delta_{\theta,i}$ for the DPO log-ratio margin of pair i,

$$\Delta_{\theta,i} = \log \frac{\pi_\theta(y_{w,i} \mid x)}{\pi_{\text{ref}}(y_{w,i} \mid x)} - \log \frac{\pi_\theta(y_{l,i} \mid x)}{\pi_{\text{ref}}(y_{l,i} \mid x)}, \tag{4}$$

the Physio-DPO objective is

$$\mathcal{L}_{\text{Physio}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{i \sim \mathcal{D}}\Big[w_i \cdot \log \sigma\big(\beta \Delta_{\theta,i}\big)\Big]. \tag{5}$$

The target stays binary: the energy gap sets a per-pair cost rather than a target, which is what separates Eq. 5 from soft-label methods such as E2PO. Setting $w_i \equiv 1$ recovers DPO.

3.3 GRADIENT MODULATION

Because $\mathcal{E}$ is a fixed oracle evaluated on precomputed structures, w_i does not depend on θ, and the gradient of Eq. 5 decomposes cleanly. Using $\sigma'(z) = \sigma(z)\sigma(-z)$, the contribution of pair i is

$$\nabla_\theta \mathcal{L}_i = -\beta\, w_i \underbrace{\sigma\big(-\beta\Delta_{\theta,i}\big)}_{\text{error term}} \nabla_\theta \Delta_{\theta,i}, \qquad \big\|\nabla_\theta \mathcal{L}_i\big\| = \beta\, w_i\, \sigma\big(-\beta\Delta_{\theta,i}\big)\, \big\|\nabla_\theta \Delta_{\theta,i}\big\|. \tag{6}$$

The two factors play different roles: the error term $\sigma(-\beta\Delta_{\theta,i})$ decays as a preference is learned, as in DPO, whereas the gate w_i does not decay and on $\mathcal{D}_{\text{phys}}$ reads the local energy gap.

Theorem E.1 in Appendix E turns this decomposition into two guarantees that hold together. Fixing the model state and a physics-perturbed pair and varying only its oracle gap, the norm in Eq. 6 is *strictly increasing* in δ_E, so a more severe local violation always draws more gradient than a milder one; and it is *uniformly bounded* by $\beta\lambda_{\max}\sigma(-\beta\Delta_{\theta,i})\|\nabla_\theta\Delta_{\theta,i}\|$, so no gap, however extreme, lets one pair dominate a minibatch. Obtaining both at once is what the thresholded sigmoid is for: linear weighting and the ODPO offset give the first and lose the second, which is visible in their optimization behaviour in Section 4.2.

4 EXPERIMENTS

We evaluate Physio-DPO on protein generation to answer three questions: **(Q1)** whether local energy-gap magnitudes improve over binary preference alignment; **(Q2)** whether Physio-DPO mitigates structural hallucinations, meaning high model confidence together with poor predicted stability; and **(Q3)** how the weighting function $\Psi(\delta_E)$ and hard-negative mining each contribute.

4.1 SETUP

Data. We train on PhysioPref-1M (Section 3.1), split with MMseqs2 under a 30% sequence-identity ceiling into 900,000 training, 50,000 validation, and 50,000 unseen preference pairs. The validation split selects hyperparameters and checkpoints, and every generation metric is computed on sequences sampled afresh from each trained policy rather than on stored pairs. **Baselines.** We compare against unaligned PLMs (ProGen2-XL 6.4B, ESM-3 Open 1.4B, ESM-2 3B, ProtGPT2 762M), SFT on the preferred (y_w) subset, PPO with a reward model trained on PhysioPref-1M, and the preference-optimization family of DPO, IPO, KTO, SimPO, and ORPO; Section 4.2 adds the energy-aware objectives. **Training.** All models start from ProGen2-XL (Nijkamp et al., 2023) and are fine-tuned with LoRA (Hu et al., 2022) at $r = 16$, $\alpha = 32$ on $4 \times$ NVIDIA A100 (80GB) GPUs using HuggingFace TRL, with $\beta = 0.1$ and $\Psi(\delta_E)$ at $\mu = 50$, $\tau = 10$. **Metrics.** We report structural stability as sc-RMSD, foldability as the fraction of sequences with predicted pLDDT above 70, biophysical validity as mean Rosetta energy per residue, and diversity as perplexity under π_{ref} with maximum sequence identity to the training set; Section 4.3 adds zero-shot fitness prediction against experimental deep mutational scanning data. Appendix F gives baseline configurations, the full hyperparameter table, and formal metric definitions. Unless a table reports three seeds, every result is a single run over 30,000 generated sequences using the ungated variant that weights all pairs, which under the same protocol differs from the fold-gated objective by 0.02 Å and 0.4 points (Table 13).

4.2 MAIN RESULTS

Protein Generation Performance. Table 1 shows that Physio-DPO achieves the strongest generation quality among the aligned models, reducing sc-RMSD by 0.54 Å and raising foldability by 9.2 points over standard DPO. The advantage is most pronounced where structure and diversity normally trade off: Physio-DPO attains the best structural and energy scores at the lowest maximum training-sequence identity of any aligned model, whereas the reference-free variants SimPO and ORPO obtain their structural gains alongside the weakest diversity in the table. PPO shows the same tension in sharper form, improving foldability while both its perplexity and its training-sequence identity rise. This pattern is consistent with a dense but bounded energy signal steering optimization without pulling the policy away from the pretrained prior.

Table 1: Protein generation quality across pre-trained backbones and alignment methods. Every aligned model is initialized from ProGen2-XL.

Method	Structural metrics			Biophysical	Diversity metrics	
	sc-RMSD (Å) ↓	pLDDT ↑	Foldability (%) ↑	Energy (REU) ↓	PPL ↓	Seq-Id (%) ↓
Pre-trained backbones (zero-shot generation)						
ProtGPT2 (762M) (Ferruz et al., 2022)	5.12	48.5	22.1	-0.85	8.2	–
ESM-2 (3B)† (Lin et al., 2023)	4.55	56.2	35.4	-1.05	6.8	–
ESM-3 (1.4B) (Hayes et al., 2024)	3.95	62.5	45.8	-1.31	6.5	–
ProGen2-XL (6.4B) (Nijkamp et al., 2023)	3.25	67.8	52.4	-1.65	**6.1**	–
Alignment methods (backbone: ProGen2-XL 6.4B + LoRA)						
SFT (supervised fine-tuning)	2.35	75.2	71.5	-2.25	7.4	34.1
PPO (RLHF) (Schulman et al., 2017)	2.15	78.5	79.2	-2.48	12.5	39.8
DPO (standard) (Rafailov et al., 2023)	1.82	81.3	83.6	-2.65	8.6	36.5
SimPO (Meng et al., 2024)	1.71	82.6	85.2	-2.73	9.8	41.2
ORPO (Hong et al., 2024)	1.69	82.0	84.7	-2.76	10.3	42.7
IPO (Azar et al., 2024)	1.88	80.8	82.9	-2.58	<u>7.8</u>	35.2
KTO (Ethayarajh et al., 2024)	1.79	82.1	84.1	-2.71	8.1	36.1
Physio-DPO (ours)	**1.28**	**87.5**	**92.8**	**-3.05**	8.2	**33.8**
Improvement vs. standard DPO	*(-29%)*	*(+7.6%)*	*(+11%)*	*(-15%)*	*(-4.7%)*	*(-7.4%)*

†Masked language model; sequences obtained by iterative masked sampling.

Energy-Aware Objectives. All magnitude-aware objectives here apply δ_E only to backbone-matched pairs and share the backbone, budget, and seeds. Table 2 shows that using the magnitude at all is what separates these objectives from standard DPO, and that Physio-DPO is the strongest among them. Within the magnitude-aware group the ordering tracks how the magnitude enters the loss rather than how much of it is used: the two unbounded routes, linear weighting and the ODPO margin offset (Amini et al., 2024), trail the bounded ones, and sampling the label from the gap trails both. In contrast, conservative DPO softens the target at a fixed rate that carries no magnitude information and remains indistinguishable from standard DPO (Rafailov et al., 2023). The pattern supports treating the local energy gap as a bounded weight on a binary target rather than as the target itself, which is the point at which E2PO caps its own winner-pushing gradient.

Table 2: Objectives that use the energy gap, over three seeds (mean ± 95% CI) under matched settings.

Method	sc-RMSD (Å)↓	Fold. (%)↑
DPO (standard)	1.85 ± 0.06	83.1 ± 1.1
cDPO (smoothed)	1.81 ± 0.07	83.7 ± 1.2
Stochastic BT-DPO	1.64 ± 0.09	85.9 ± 1.5
Linear weighting	1.57 ± 0.07	87.6 ± 1.2
ODPO	1.52 ± 0.06	88.6 ± 1.1
Linear, replacing Ψ	1.48 ± 0.06	88.9 ± 1.0
E2PO from AliDiff	1.44 ± 0.05	89.8 ± 0.9
Physio-DPO	$\mathbf{1.32 \pm 0.04}$	$\mathbf{92.1 \pm 0.8}$

4.3 Analysis

Ablation Studies. Table 3 shows that both components are necessary, and that the binding constraint is the data rather than the objective. Replacing mined hard negatives with random ones is the most damaging change, raising sc-RMSD from 1.28 to 2.12 Å, whereas removing the weighting term alone returns performance to standard DPO. Neither a sampled label nor an unbounded weight recovers the difference: Stochastic BT-DPO lands between the two, and linear weighting stops at 1.45 Å with roughly twice the reward variance measured over training. This indicates that the magnitude becomes usable only when the paired structures are close enough for the gap to mean something, and stable only when its weight is bounded, as Theorem E.1 requires. The mined negatives are not memorized fingerprints: holding out one perturbation type and testing only on it costs Physio-DPO under 3.5 foldability points against over 5 for DPO (Appendix D).

Training Dynamics. Figure 3 shows that Physio-DPO keeps improving throughout training while the baselines either destabilize or stall. PPO destabilizes after roughly 2K steps, its KL divergence rising sharply as energy degrades, whereas standard DPO converges quickly and then saturates. The

Table 3: Contribution of negative selection and energy-aware weighting. Each row modifies a single component of Physio-DPO, and *w/o* denotes without.

Ablation	Change	sc-RMSD ↓	Foldability ↑
Physio-DPO	*Sigmoid + hard negatives*	**1.28**	**92.8%**
w/o weighting	Standard DPO	1.82	83.6%
w/o hard negatives	Random negatives	2.12	76.5%
w/ linear weighting	$\Psi(\delta_E) \propto \delta_E$	1.45	89.1%
w/ stochastic preference	Stochastic BT-DPO	1.61	86.3%

saturation is what a binary label predicts: once a pair is confidently ordered its error term decays and the pair stops contributing, however severe its violation. Physio-DPO continues to lower energy and sc-RMSD with KL divergence bounded near 1.5 nats, which is consistent with a gate that supplies gradient independently of how confident the model has become.

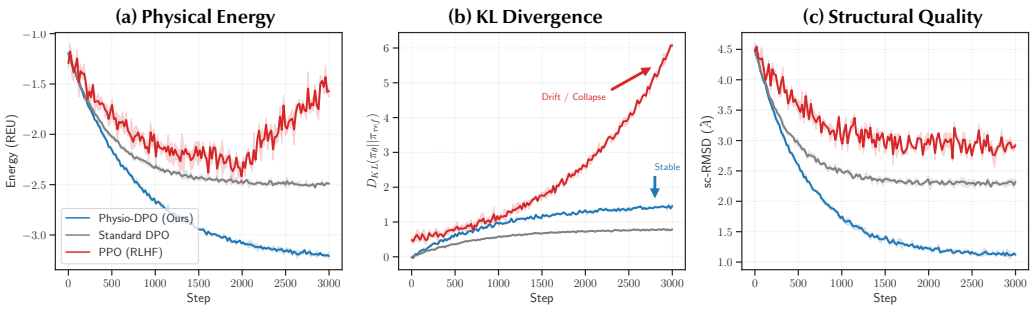

Figure 3: Training dynamics: (a) oracle energy, (b) KL from π_{ref}, (c) sc-RMSD.

Mitigating Hallucinations. We treat low-perplexity sequences with high oracle energy as structural hallucinations. Figure 4 shows a distinct high-confidence, high-energy population under standard DPO. This population is largely absent under Physio-DPO, and more samples move into the high-confidence, low-energy region, indicating a closer correspondence between model confidence and oracle-derived energy. Section 4.3 and Appendix C test whether that correspondence survives when the oracle is replaced.

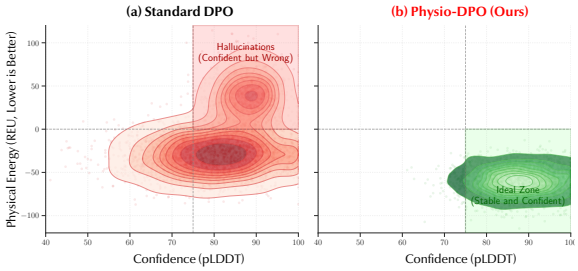

Figure 4: Energy versus confidence (pLDDT).

Structural Corrections. Figure 5 shows what the energy gap buys at atomic resolution. The SFT baseline usually preserves global topology while violating fine-grained constraints, and the corrections group into three kinds. Physio-DPO improves conformational stability, compacting disordered loops into well-formed helices (Fig. 5a), burying hydrophobic residues into a tight core (Fig. 5b), and eliminating internal voids (Fig. 5h). It restores atomic interactions that baseline generations disrupt: hydrogen-bond networks in β-sheet regions (Fig. 5c), disulfide geometry (Fig. 5e), and salt bridges between oppositely charged residues (Fig. 5f). And it suppresses stereochemical violations, correcting non-physical backbone distortions (Fig. 5d), forbidden torsion configurations (Fig. 5g), chain connectivity breaks (Fig. 5i), and steric clashes (Fig. 5j). These are precisely the defect classes that the physics-perturbed decoys are built to expose.

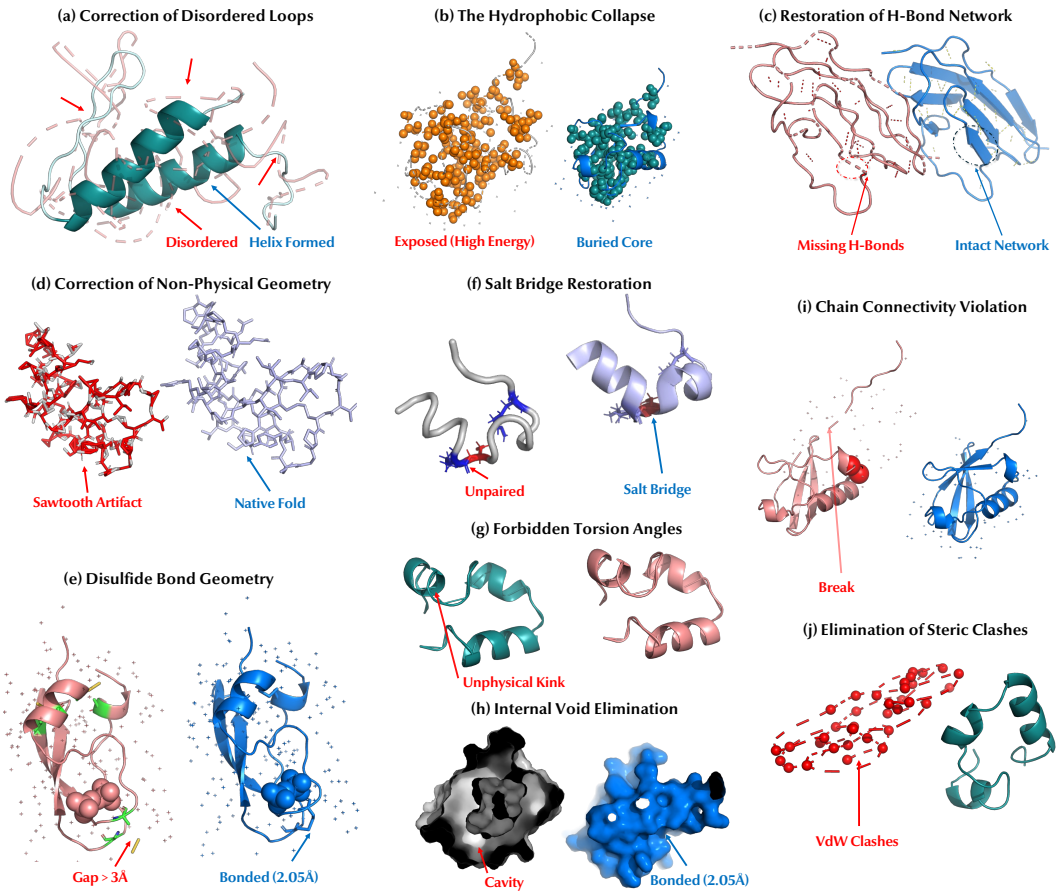

Figure 5: Comprehensive qualitative analysis of biophysical validity. We compare structures generated by the SFT Baseline (Red/Pink/Grey, Left) and Physio-DPO (Blue/Teal, Right). **(a)** Physio-DPO compacts disordered loops into stable helices. **(b)** Exposed hydrophobic residues (Orange) in SFT are buried into a tight core (Teal) by Physio-DPO. **(c, e, f)** Restoration of critical atomic interactions: hydrogen bond networks in β-sheets, precise disulfide bond geometry (2.05Å), and electrostatic salt bridges. **(d, g, i)** Correction of severe geometrical violations, including non-physical "sawtooth" backbones, forbidden torsion kinks, and chain connectivity breaks. **(h, j)** Optimization of packing density by eliminating destabilizing internal voids and steric clashes (Red).

Zero-shot Generalization. Because δ_E is produced by the same family of tools used for structural evaluation, the gains have to be shown against labels no oracle produced. ProteinGym (Notin et al., 2023) supplies them, scoring log-likelihoods under π_θ against wet-lab deep mutational scanning measurements (Table 4). Physio-DPO raises the average Spearman correlation from 0.67 to 0.72, with the largest gains on the stability-driven GFP and P53 assays, so agreement with Rosetta cannot by itself account for the improvement; PPO degrades zero-shot performance while Physio-DPO preserves it. The one assay where Physio-DPO trails a retrieval-augmented baseline is GB1, whose fitness is binding-driven while our energy oracle scores monomers. That is a property of the oracle, not of the objective, which never inspects how δ_E is computed: substituting Rosetta InterfaceAnalyzer binding energies and adding 60K interface pairs built from PDB dimers raises GB1 from 0.63 to 0.69, the largest move in the table, while every monomeric assay stays within 0.01.

Robustness to the Scoring Oracle. The structural ranking likewise does not depend on the tools used to build PhysioPref-1M. Re-scoring generations with FoldX on OmegaFold structures, with AGAPE, and with AlphaFold3, none of which took part in dataset construction or checkpoint selection, preserves the ordering, and AlphaFold3 places Physio-DPO 7.7 foldability points above standard DPO. AlphaFold3 carries the most weight here, being a learned structure predictor rather

Table 4: Zero-shot fitness prediction on ProteinGym: Spearman correlation (ρ) between model log-likelihoods and experimental fitness. **Bold** and underlined mark best and second-best.

Method *Spearman's ρ*	**GFP** *(Stability)*	**GB1** *(Binding)*	**AAV2** *(Viral)*	**TEM-1** *(Resistance)*	**P53** *(Suppressor)*	**Avg.**
Prior baselines						
ESM-2 (3B) (Lin et al., 2023)	0.62	0.55	0.70	0.63	0.51	0.60
Tranception (Notin et al., 2022)	0.65	0.68	0.72	0.69	0.48	0.64
ProGen2-XL (6.4B) backbones						
ProGen2-XL (base)	0.64	0.56	0.71	0.65	0.53	0.62
SFT	0.66	0.55	0.70	0.67	0.56	0.63
PPO (RLHF)	0.62	0.53	0.65	0.61	0.54	0.59
DPO (standard)	0.71	0.59	0.73	0.70	0.61	0.67
Physio-DPO (ours)	**0.78**	0.63	**0.75**	**0.76**	0.70	0.72
+ interface oracle	**0.78**	**0.69**	**0.75**	**0.76**	**0.71**	**0.74**

than an empirical energy function and the least correlated with the training signal at Spearman $\rho = 0.29$, against 0.54 for FoldX and 0.48 for AGAPE. That the two independent energy functions agree with each other at $\rho = 0.79$ while each correlates only moderately with Rosetta indicates the improvement is not an artefact of a single scorer. Appendix C reports the full cross-oracle evaluation.

Hyperparameter Sensitivity. We evaluate the robustness of Physio-DPO with respect to two key hyperparameters: the KL-penalty coefficient (β) and the physics weighting scale (μ). Figure 6(a) compares sc-RMSD for Physio-DPO and DPO across $\beta \in [0.01, 1.0]$. Physio-DPO remains stable and achieves consistently low sc-RMSD even at small β values ($\beta = 0.01$), indicating that the dense energy-gap supervision carried by $\Psi(\delta_E)$ effectively regularizes training and mitigates catastrophic forgetting. The μ governs the strength of energy-dependent modulation. As shown in Figure 6(b), small values of μ (< 10) yield SFT-like behavior with higher energy, while excessively large values (> 100) overemphasize physical energy and degrade language modeling performance, reflected by increased perplexity. We identify a broad optimal range of $\mu \in [20, 50]$, where Physio-DPO achieves low physical energy while preserving strong generative quality.

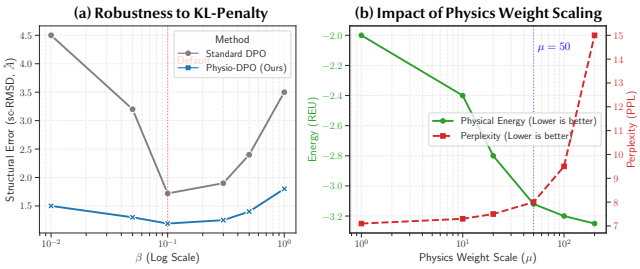

Figure 6: Hyperparameter sensitivity analysis results.

Additional Analysis. Appendix G reports the checks that bear directly on the claims above. The pair-geometry statistics justifying fold gating are in Table 12, and a sweep of the sigmoid sharpness τ (Table 16) shows the thresholded weight is not narrowly tuned. Physio-DPO preserves fold diversity and reduces memorization relative to standard DPO, whereas naive energy minimization collapses both (Table 11), and the gains extend past foldability to predicted solubility and aggregation propensity (Table 15). Blinded expert review confirms the automated labels (Table 7).

5 Conclusion

We propose Physio-DPO, a physics-informed preference optimization framework for aligning large protein language models with local energy-gap supervision. We show that discrete preference modeling is insufficient in biophysical settings, as it neglects the continuous structure of energy landscapes. By incorporating energy magnitudes directly into the alignment objective, Physio-DPO guides optimization toward physically meaningful distinctions. Our results demonstrate that embedding physical principles at the alignment stage enables large-scale protein language models to internalize fine-grained biophysical constraints without sacrificing generative or linguistic capacity.

AI USE STATEMENT

In this work, we used large language models to polish grammar and wording, and to check mathematical notation and proofs. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the final content of this paper.

REPRODUCIBILITY STATEMENT

Appendix F specifies the backbone, adapter configuration, optimizer, schedule, batch size, step budget, and every objective-specific hyperparameter for Physio-DPO and for all baselines, including β, μ, and τ for Ψ, and the τ_E used by the soft-target and sampled-label variants. Appendix D documents the PhysioPref-1M construction pipeline, including sampling temperature, the pLDDT and energy thresholds that define each subset, the PyRosetta perturbation protocol with its energy function and RMSD filter, the subset proportions, and the blinded human verification procedure. Appendix F.2 gives formal definitions and computation procedures for every reported metric. Upon acceptance we will release PhysioPref-1M together with the interface pairs, the dataset construction code, and the training and evaluation code for every method reported here.

REFERENCES

Rebecca F Alford, Andrew Leaver-Fay, Jeliazko R Jeliazkov, Matthew J O'Meara, Frank P DiMaio, Hahnbeom Park, Maxim V Shapovalov, P Douglas Renfrew, Vikram K Mulligan, Kalli Kappel, et al. The rosetta all-atom energy function for macromolecular modeling and design. *Journal of chemical theory and computation*, 13(6):3031–3048, 2017.

Afra Amini, Tim Vieira, and Ryan Cotterell. Direct preference optimization with an offset. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9954–9972, 2024.

Christof Angermueller, David Dohan, David Belanger, Ramya Deshpande, Kevin Murphy, and Lucy Colwell. Model-based reinforcement learning for biological sequence design. In *International Conference on Learning Representations*, 2020.

Ivan Anishchenko, Samuel J Pellock, Tamuka M Chidyausiku, Theresa A Ramelot, Sergey Ovchinnikov, Jinghan Hao, Vikas Bafna, Christoffer Norn, Alex Kang, Asim K Bera, et al. De novo protein design by deep network hallucination. *Nature*, 600(7889):547–552, 2021.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pp. 4447–4455. PMLR, 2024.

Shreyas Chaudhari, Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, Ameet Deshpande, and Bruno Castro da Silva. Rlhf deciphered: A critical analysis of reinforcement learning from human feedback for llms. *ACM Computing Surveys*, 58(2):1–37, 2025.

Amitava Das, Suranjana Trivedy, Danush Khanna, Yaswanth Narsupalli, Basab Ghosh, Rajarshi Roy, Gurpreet Singh, Vinija Jain, Vasu Sharma, Aishwarya Naresh Reganti, et al. Dpo kernels: A semantically-aware, kernel-enhanced, and divergence-rich paradigm for direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 22174–22270, 2025.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rabaan, Florian Burkhardt, and Burkhard Rost. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2021.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

Ziqi Fang, Hongbiao Ran, YongHan Zhang, Chensong Chen, Ping Lin, Xiang Zhang, and Min Wu. Alphafold 3: an unprecedent opportunity for fundamental research and drug development. *Precision Clinical Medicine*, 8(3):pbaf015, 2025.

Noelia Ferruz, Steffen Schmidt, and Birte Höcker. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348, 2022.

Shreya Gopalan and Sundar Narayanan. Hallucinations in alphafold3 for intrinsically disordered proteins with disorder in biological process residues. *arXiv preprint arXiv:2510.15939*, 2025.

Siyi Gu, Minkai Xu, Alexander Powers, Weili Nie, Tomas Geffner, Karsten Kreis, Jure Leskovec, Arash Vahdat, and Stefano Ermon. Aligning target-aware molecule diffusion models with exact energy optimization. *Advances in Neural Information Processing Systems*, 37:44040–44063, 2024.

Thomas Hayes, Roshan Rao, et al. Simulating 500 million years of evolution with a language model. *bioRxiv*, 2024.

Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*, 2024.

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.

John B Ingraham, Max Baranov, Zak Costello, Karl W Barber, Wujie Wang, Ahmed Ismail, Vincent Frappier, Dana M Lord, Christopher Ng-Thow-Hing, Erik R Van Vlack, et al. Illuminating protein space with a programmable generative model. *Nature*, 623(7989):1070–1078, 2023.

Ronald M Levy, Allan Haldane, and William F Flynn. Potts hamiltonian models of protein covariation, free energy landscapes,and evolutionary fitness. *Current Opinion in Structural Biology*, 43:55–62, 2017.

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

Sidney Lyayuga Lisanza, Jacob Merle Gershon, Samuel WK Tipps, Jeremiah Nelson Sims, Lucas Arnoldt, Samuel J Hendel, Miriam K Simma, Ge Liu, Muna Yase, Hongwei Wu, et al. Multistate and functional protein design using rosettafold sequence space diffusion. *Nature biotechnology*, 43(8):1288–1298, 2025.

Shunyu Liu, Wenkai Fang, Zetian Hu, Junjie Zhang, Yang Zhou, Kongcheng Zhang, Rongcheng Tu, Ting-En Lin, Fei Huang, Mingli Song, et al. A survey of direct preference optimization. *arXiv preprint arXiv:2503.11701*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Ali Madani, Ben Krause, Eric R Greene, S Subramadian, et al. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8):1099–1106, 2023.

Zhiyu Mei, Wei Fu, Kaiwei Li, Guangju Wang, Huanchen Zhang, and Yi Wu. Real: Efficient rlhf training of large language models with parameter reallocation. *Proceedings of Machine Learning and Systems*, 7, 2025.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235, 2024.

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, 2023.

Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena-Hurtado, Aidan N Gomez, Debora Marks, and Yarin Gal. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. In *International Conference on Machine Learning*, pp. 16990–17017. PMLR, 2022.

Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood Van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, et al. Proteingym: Large-scale benchmarks for protein fitness prediction and design. *Advances in Neural Information Processing Systems*, 36: 64331–64379, 2023.

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. *Advances in neural information processing systems*, 36:53728–53741, 2023.

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

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Yijia Xiao, Wanjia Zhao, Junkai Zhang, Yiqiao Jin, Han Zhang, Zhicheng Ren, Renliang Sun, Haixin Wang, Guancheng Wan, Pan Lu, et al. Protein large language models: A comprehensive survey. *arXiv preprint arXiv:2502.17504*, 2025.

Jiayi Zhou, Jiaming Ji, Josef Dai, and Yaodong Yang. Sequence to sequence reward modeling: Improving rlhf by language feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 27765–27773, 2025.

A Problem Setup and DPO Formulation

Direct Preference Optimization (DPO). Direct Preference Optimization (DPO) (Rafailov et al., 2023) is a framework for aligning language models with preference data without reinforcement learning. Given a reference policy π_{ref} and a trainable policy π_θ, it optimizes π_θ to prefer a winner y_w over a loser y_l by maximizing the probability:

$$P(y_w \succ y_l) = \sigma(r(x, y_w) - r(x, y_l)), \tag{7}$$

where σ is the sigmoid and $r(x, y)$ follows the Bradley–Terry model, yielding a stable preference-based objective regularized by π_{ref}.

Problem Formulation. We formulate protein design as an unconditional sequence generation problem. Let x denote a generic prefix (e.g., a start token); an autoregressive PLM parameterized by θ defines a policy $\pi_\theta(y|x)$ over amino acid sequences $y \in \mathcal{Y}$. While pretrained PLMs capture evolutionary plausibility, they do not explicitly enforce biophysical validity. We assume access to an empirical energy oracle $\mathcal{E}(y)$, where lower energy indicates higher predicted stability. Our objective is to align π_θ to favor low-energy sequences while remaining close to a reference model π_{ref} to keep diversity.

B Extended Results

This appendix reports the matched multi-seed evaluation. The companion analyses referenced from the main text appear in Appendix C (unseen computational oracles), Appendix G.5 (pair geometry and fold gating), Appendix G.7 (sigmoid sharpness), and Appendix G.6 (novelty, solubility, and aggregation). Across three random seeds under identical settings, Physio-DPO achieves the strongest performance on both primary metrics. Its 95% confidence intervals do not overlap those of any baseline and are the narrowest in both columns (Table 5). These intervals summarize variation across three seed-level estimates and are reported as descriptive uncertainty.

Table 5: Matched generation results over three random seeds (mean $\pm$ 95% confidence interval; 30,000 generated sequences per seed). Lower sc-RMSD and higher foldability are better.

Method	**sc-RMSD (Å)** $\downarrow$	**Foldability (%)** $\uparrow$
SFT	2.38 ± 0.07	71.1 ± 1.4
DPO (standard) (Rafailov et al., 2023)	1.85 ± 0.06	83.1 ± 1.1
SimPO (Meng et al., 2024)	1.74 ± 0.08	84.8 ± 1.3
ORPO (Hong et al., 2024)	1.72 ± 0.07	84.3 ± 1.2
Stochastic BT-DPO	1.64 ± 0.09	85.9 ± 1.5
Physio-DPO	$\mathbf{1.32 \pm 0.04}$	$\mathbf{92.1 \pm 0.8}$

C Unseen Computational Oracles

To assess whether the observed improvements are specific to the Rosetta-based signal used during preference construction, we evaluated the generated sequences with energy functions and structure predictors that were not used to construct PhysioPref-1M or to select training checkpoints. Table 6 reports the resulting cross-oracle evaluation.

Physio-DPO retains its advantage across every unseen evaluation. FoldX and AGAPE assign lower energies to Physio-DPO generations, while AlphaFold3 identifies 90.1% of the evaluated sequences as foldable, compared with 82.4% for standard DPO. The OmegaFold-derived measurements also indicate a larger buried hydrophobic fraction and a higher fraction of structured residues.

The unseen scores are not restatements of the training signal. On the evaluated samples, the Spearman correlations between Rosetta energy and FoldX, AGAPE, and AlphaFold3 pLDDT are 0.54, 0.48, and 0.29, so none of the three is redundant with the signal Physio-DPO was trained on, and

Table 6: Cross-oracle evaluation using computational predictors unseen during training. FoldX, AGAPE, SASA, and DSSP are evaluated on 1,000 generated sequences per model; AlphaFold3 is evaluated on 500 sequences per model. The final two rows report changes relative to standard DPO.

Unseen evaluation	**Standard DPO**	**Physio-DPO**
FoldX ΔG on OmegaFold structures (kcal/mol) $\downarrow$	-24.9	**-33.1**
AGAPE energy $\downarrow$	-2.05	**-2.71**
AlphaFold3 foldability, pLDDT > 70 (%) $\uparrow$	82.4	**90.1**
Buried hydrophobic fraction, SASA < 20 Å^2, change relative to DPO	–	**+19.2%**
Structured-residue fraction from DSSP, change relative to DPO	–	**+12.5%**

AlphaFold3 is close to decorrelated from it. FoldX and AGAPE nevertheless agree with each other at $\rho = 0.79$, so the ranking they produce is not an artefact of either one.

D PHYSIOPREF-1M BENCHMARK

Effective preference alignment for protein language models requires dense, physically grounded supervision, which is largely absent from existing instruction-tuning datasets. To address this gap, we introduce PhysioPref-1M, a large-scale preference benchmark comprising 1M protein pairs annotated by energy-based criteria. A central challenge in constructing such data is that random mutations typically yield trivially unfolded structures, resulting in uninformative preference signals. To mine physically meaningful comparisons, particularly hard negatives that expose structural hallucinations, we design an adversarial generation–filtration pipeline (Figure 2) that integrates large-scale sampling, physics-based scoring, and expert validation.

Adversarial Generation and Filtering Pipeline We first sample approximately 50M sequences from the reference policy π_{ref} using high-temperature nucleus sampling ($T = 1.2, p = 0.9$), encouraging exploration of the distribution tail where hallucinations are likely to occur. Each sequence is then folded into a 3D structure using ESMFold (Lin et al., 2023), and evaluated with a joint biophysical score vector:

$$\mathbf{v} = [\text{pLDDT}(S), \mathcal{E}_{\text{Rosetta}}(S)],$$

where pLDDT measures folding confidence and $\mathcal{E}_{\text{Rosetta}}$ denotes the Rosetta Energy Unit (REU) score. Based on this joint assessment, structures are categorized into three classes: (1) Stable Positives, with high confidence (pLDDT > 75) and low energy; (2) Unstable Negatives, with low confidence (pLDDT < 50) and high energy; and (3) Hard Negatives, a critical subset exhibiting high folding confidence yet unfavorable physical energy, often due to defects such as exposed hydrophobic cores or loose packing. Preference pairs (y_w, y_l) are constructed by contrasting Stable positives against Unstable or Hard Negative samples, with the final dataset stratified to ensure 30% Hard Negatives, forcing models to learn fine-grained biophysical distinctions beyond simple foldability. Dataset statistics are summarized in Table 9.

Physics-Based Hard Negative Construction. To further enrich Hard Negatives, we apply explicit physics-based perturbations to high-resolution native structures from the CATH dataset using PyRosetta. We adopt the ref2015 energy function, which accounts for van der Waals interactions, solvation, hydrogen bonding, and electrostatics, all measured in Rosetta Energy Units (REU). For each native structure S_{native}, we generate a decoy S_{decoy} via a combination of: (1) Side-chain repacking on 15% of residues using high-temperature rotamer sampling to induce steric clashes; (2) Backbone torsion perturbations by adding Gaussian noise ($\sigma = 5^\circ$) to loop regions, disrupting hydrogen bond networks; and (3) Rare chirality inversions of a single C_α atom to create subtle but severe stereochemical violations. Decoys are filtered to ensure an RMSD within $[1.0, \text{Å}, 4.0, \text{Å}]$ relative to the native structure, yielding pairs that are structurally similar yet physically distinguishable, a defining property of Hard Negatives.

Table 7: Blinded human verification results. Alignment rate denotes the percentage of Physio-DPO labels confirmed by human experts.

Subset	HAA rate (%)	Expert consensus (κ)
Trivial pairs	96.5%	0.88
Energy-dominant	88.2%	0.71
Hard negatives	82.7%	0.65
Overall	**89.1%**	**0.74**

Human-in-the-Loop Validation. To assess the reliability of automated physics-based labeling, we conduct a human-in-the-loop evaluation on a stratified sample of 10K preference pairs. Three domain experts perform blinded side-by-side comparisons, judging structural plausibility and identifying specific defects. As reported in Table 7, the dataset achieves an overall Human–Algorithm Alignment (HAA) rate of 89.1%. Notably, for Hard Negative pairs, expert judgments agree with energy-based preferences in 82.7% of cases, primarily citing exposed hydrophobic residues and poor packing. Inter-annotator agreement is substantial, confirming the reliability of PhysioPref-1M.

Generalization Beyond Synthetic Hard Negatives. A potential concern is that Physio-DPO may overfit to synthetic perturbation fingerprints rather than learning general biophysical failure modes. To assess this, we conduct three complementary analyses, summarized in Table 8. Across out-of-distribution ProteinGym assays, perturbation-source holdout evaluation, and failure-mode analysis of high-energy *de novo* generations, Physio-DPO consistently shows stronger transfer than standard DPO. These results suggest that synthetic hard negatives act as scaffolds for exposing general physical violations, rather than shortcuts that can be memorized.

Table 8: Evidence that Physio-DPO generalizes beyond synthetic hard negatives.

Evaluation	Protocol	DPO	Physio-DPO	Conclusion
OOD generalization (ProteinGym DMS)	Zero-shot fitness prediction on GFP and P53	0.71 / 0.61	**0.78 / 0.70**	Transfers to experimental fitness
Perturbation-source ablation	Omit one perturbation type during training; test only on that type	Foldability drop $> 5\%$	Foldability drop $< 3.5\%$	Robust across perturbation sources
Failure-mode diversity	Analyze 50 high-energy *de novo* generations	Chirality-heavy	Exposed cores (**45%**), clashes (**30%**), and chirality (**6%**)	Captures diverse physical defects

E Theoretical Analysis of Physio-DPO

This appendix proves the gradient-modulation property invoked in Section 3.3, records how the weight reads at initialization, and gives the Boltzmann argument that motivated using energy as a source of preference strength.

E.1 Preliminaries

Definition E.1 (Implicit reward)**.** *Given a parameterized policy $\pi_\theta(y \mid x)$ and a frozen reference policy $\pi_{\mathrm{ref}}(y \mid x)$, the implicit reward is the scaled log-likelihood ratio*

$$r_\theta(x, y) = \beta \log \frac{\pi_\theta(y \mid x)}{\pi_{\mathrm{ref}}(y \mid x)}, \tag{8}$$

where $\beta > 0$ is the KL-penalty coefficient.

Table 9: Statistics of the proposed PhysioPref-1M dataset. Here, y_w and y_l denote the preferred and dispreferred sequences, respectively.

Subset Type	Count	Average Length		Average Metrics of y_l		Physical Gap
		y_w	y_l	pLDDT	Energy (REU)	$\overline{\Delta\mathcal{E}}$ (REU)
Trivial Pairs *Stable vs. unstable*	200K (20%)	256	254	42.5	+150.2	185.4
Energy-Dominant *Stable vs. high-energy*	500K (50%)	310	308	68.1	+24.5	65.8
Hard Negatives *Stable vs. hallucinated*	300K (30%)	285	288	**78.4**	**+45.1**	92.3
Total / Weighted Average	**1M**	291.7	291.2	66.1	+55.8	97.7

Definition E.2 (Preference margin)**.** *For a preference pair $i = (x, y_w, y_l)$, the margin $\Delta_{\theta,i}$ is given by Eq. 4, so that $\beta\Delta_{\theta,i} = r_\theta(x, y_w) - r_\theta(x, y_l)$.*

Definition E.3 (Local energy gap and weighting)**.** *Let $\mathcal{E} : \mathcal{Y} \to \mathbb{R}$ be a fixed empirical energy function evaluated on predicted structures. The local energy gap of pair i is $\delta_{E,i} = \mathrm{ReLU}\big(\mathcal{E}(y_{l,i}) - \mathcal{E}(y_{w,i})\big) \in [0, \infty)$ as in Eq. 1, and the weighting function is $\Psi(\delta_E) = \lambda_{\max}\,\sigma\big((\delta_E - \mu)/\tau\big)$ as in Eq. 2, with $\lambda_{\max} > 0$, $\mu \in \mathbb{R}$, and $\tau > 0$. The fold-gated weight w_i is defined by Eq. 3.*

Remark E.1 (Range of Ψ)**.** *On the domain $[0, \infty)$ the argument $(\delta_E - \mu)/\tau$ ranges over $[-\mu/\tau, \infty)$, so Ψ maps $[0, \infty)$ into $\big[\lambda_{\max}\sigma(-\mu/\tau),\ \lambda_{\max}\big)$. The lower endpoint is attained at $\delta_E = 0$ and the upper endpoint is a strict supremum: Ψ is strictly positive everywhere and never reaches $\lambda_{\max}$. At $\mu = 50$ and $\tau = 10$ we have $\Psi(0) = \lambda_{\max}\sigma(-5) \approx 6.7 \times 10^{-3}\lambda_{\max}$, so thresholding down-weights near-tied pairs by a factor of roughly 150 rather than discarding them.*

Throughout, w_i is independent of θ. This holds because $\mathcal{E}$ is a fixed oracle evaluated on structures predicted before training, $\lambda_{\max}$, μ, and τ are fixed, and membership in $\mathcal{D}_{\mathrm{phys}}$ or $\mathcal{D}_{\mathrm{gen}}$ is determined by dataset construction.

E.2 Bounded Conditional Gradient Modulation

Theorem E.1 (Bounded conditional gradient modulation)**.** *Let $\ell_i(\theta) = -\,w_i \log\sigma\big(\beta\Delta_{\theta,i}\big)$ be the contribution of pair i to Eq. 5, and write $s_i = \sigma(-\beta\Delta_{\theta,i}) \in (0, 1)$ and $G_i = \|\nabla_\theta\Delta_{\theta,i}\|$. Then*

$$\big\|\nabla_\theta\ell_i(\theta)\big\| = \beta\, w_i\, s_i\, G_i. \tag{9}$$

Fix the model state θ and the two sequences of a pair $i \in \mathcal{D}_{\mathrm{phys}}$, and regard the norm as a function of that pair's oracle gap $\delta_{E,i}$ alone. If $G_i > 0$, this function is strictly increasing, with

$$\frac{\partial}{\partial\delta_E}\big\|\nabla_\theta\ell_i\big\| = \beta\, s_i\, G_i\, \frac{\lambda_{\max}}{\tau}\,\sigma\left(\frac{\delta_E - \mu}{\tau}\right)\sigma\left(-\frac{\delta_E - \mu}{\tau}\right) > 0, \tag{10}$$

and it is uniformly bounded on the whole domain,

$$\beta\,\lambda_{\max}\,\sigma(-\mu/\tau)\,s_iG_i \;\le\; \big\|\nabla_\theta\ell_i\big\| \;<\; \beta\,\lambda_{\max}\,s_iG_i. \tag{11}$$

Every pair $j \in \mathcal{D}_{\mathrm{gen}}$ has $w_j = 1$, hence $\|\nabla_\theta\ell_j\| = \beta s_j G_j$, in which no energy gap appears.

Proof. **Step 1 (norm identity).** Since w_i does not depend on θ, the chain rule applied to $\Delta \mapsto -\log\sigma(\beta\Delta)$ gives

$$\nabla_\theta\ell_i(\theta) = -\,w_i\,\frac{\sigma'(\beta\Delta_{\theta,i})}{\sigma(\beta\Delta_{\theta,i})}\,\beta\,\nabla_\theta\Delta_{\theta,i} = -\,\beta\,w_i\,\sigma\big(-\beta\Delta_{\theta,i}\big)\,\nabla_\theta\Delta_{\theta,i}, \tag{12}$$

where the second equality uses the sigmoid identity $\sigma'(z) = \sigma(z)\big(1 - \sigma(z)\big) = \sigma(z)\sigma(-z)$. The three scalars β, w_i, and s_i are strictly positive, the first two by assumption and s_i because σ takes values in $(0, 1)$ at every finite argument. Positive scalars factor out of a norm, which gives Eq. 9.

Step 2 (strict monotonicity). With θ and the two sequences held fixed, neither s_i nor G_i depends on $\delta_{E,i}$, so differentiating Eq. 9 gives $\partial_{\delta_E}\|\nabla_\theta \ell_i\| = \beta s_i G_i \Psi'(\delta_E)$. Writing $z = (\delta_E - \mu)/\tau$ and applying the chain rule together with the same sigmoid identity yields $\Psi'(\delta_E) = (\lambda_{\max}/\tau)\,\sigma(z)\sigma(-z)$. Because $\lambda_{\max} > 0$, $\tau > 0$, and $\sigma(z)\sigma(-z) > 0$ at every finite z, we obtain $\Psi'(\delta_E) > 0$ on all of $[0, \infty)$, which is Eq. 10; it is strictly positive exactly when $s_i G_i > 0$, and since $s_i > 0$ always, the only requirement is $G_i > 0$.

Step 3 (bounds). By the Remark on the range of Ψ, $\Psi(\delta_E) \in [\lambda_{\max}\sigma(-\mu/\tau), \lambda_{\max})$ for every $\delta_E \in [0, \infty)$. Multiplying through by the strictly positive factor $\beta s_i G_i$ preserves both inequalities and gives Eq. 11.

Step 4 (gated pairs). Substituting $w_j = 1$ into Eq. 12 and taking norms gives $\|\nabla_\theta \ell_j\| = \beta s_j G_j$, which contains no energy term. □

Remark E.2 (Why boundedness is the load-bearing half)**.** *Replacing Ψ by the linear weight $\Psi_{\mathrm{lin}}(\delta_E) = c\,\delta_E$ leaves Step 2 intact but destroys Step 3: the same computation gives $\|\nabla_\theta \ell_i\| = \beta c\,\delta_{E,i} s_i G_i$, which is unbounded in $\delta_{E,i}$, so a single pair with an extreme oracle gap can contribute an arbitrarily large share of a minibatch gradient. Under the heavy-tailed gap distribution of PhysioPref-1M (Table 9) this is the regime the linear and margin-offset baselines operate in. Theorem E.1 bounds the worst-case per-pair contribution; the resulting reduction in reward variance and the optimization stability in Figure 3 are measured quantities.*

Remark E.3 (Scope)**.** *Theorem E.1 compares one pair against itself under a counterfactual change of its oracle gap. Across two different pairs the factors s_i and G_i also differ, so the theorem does not order the update magnitudes of distinct pairs by their energy gaps. It is a statement about the gradient contribution of a pair, not about the realized parameter step, which additionally depends on the optimizer; with plain gradient descent the two coincide up to the learning rate. Finally, the theorem concerns the magnitude of updates and not the function the implicit reward converges to. Because the preference target in Eq. 5 remains the hard label, ℓ_i is strictly decreasing in $\Delta_{\theta,i}$ for every pair, so the weight sets how fast a margin is driven up and not the value it approaches. Soft-target methods such as E2PO make the opposite choice and thereby pin the margin to a calibrated function of δ_E; Section 4.2 compares the two empirically.*

E.3 The Weight at Initialization

The next statement records what w_i does to the first update, and is the sense in which the weighting acts as a curriculum over pairs.

Assumption E.1 (Initialization proximity)**.** *At θ_0 the policy coincides with the reference policy, $\pi_{\theta_0} = \pi_{\mathrm{ref}}$, so $\Delta_{\theta_0,i} = 0$ for every pair i.*

Theorem E.2 (Weighted initial descent direction)**.** *Under Assumption E.1, and with $J(\theta) = \mathbb{E}_i\big[w_i\,\beta\Delta_{\theta,i}\big]$,*

$$\nabla_\theta \mathcal{L}_{\mathrm{Physio}}\big|_{\theta_0} = -\tfrac{1}{2}\,\nabla_\theta J(\theta)\big|_{\theta_0} = -\tfrac{\beta}{2}\,\mathbb{E}_i\big[w_i\,\nabla_\theta \Delta_{\theta,i}\big]\big|_{\theta_0}, \tag{13}$$

as an exact identity. The initial descent direction is therefore a conical combination of the per-pair margin-ascent directions in which pair i carries relative weight $w_i/\sum_j w_j$.

Proof. By Assumption E.1, $\Delta_{\theta_0,i} = 0$, hence $s_i = \sigma(0) = \frac{1}{2}$ exactly, for every pair simultaneously. Substituting this value into Eq. 12 and averaging over the training set, which is a finite sum and so may be differentiated term by term, gives the second equality in Eq. 13. Since w_i and β are constants in θ, $\nabla_\theta J(\theta) = \beta\,\mathbb{E}_i[w_i \nabla_\theta \Delta_{\theta,i}]$, and comparing the two expressions at θ_0 gives the first equality. The final claim follows because the coefficients $\frac{\beta}{2} w_i$ are strictly positive; normalizing them gives $w_i/\sum_j w_j$. No Taylor truncation is used: the evaluation $\sigma(0) = \frac{1}{2}$ is exact at this point. □

Remark E.4 (Reading of Theorem E.2)**.** *Theorem E.2 describes the update direction at one parameter value. It does not characterize a stationary point, and by Remark E.3 no proportionality between r_θ and $-\mathcal{E}$ follows from it. Assumption E.1 also holds exactly only when training starts from π_{ref}; in our runs all preference methods are initialized from the SFT checkpoint (Appendix F.1), so Theorem E.2 characterizes the idealized initialization and the empirical results rest on Theorem E.1, which assumes nothing about initialization.*

E.4 Boltzmann-Inspired Motivation

This section gives the reasoning that motivated using physical energy differences as a source of preference strength. It relates lower energy to higher policy probability, and it is a statement about a hypothetical reward function rather than about the objective we optimize.

Consider the energy-tilted distribution

$$q_{\mathrm{B}}(y \mid x) = \frac{1}{Z_{\mathrm{B}}(x)}\, \pi_{\mathrm{ref}}(y \mid x) \exp\!\left(-\frac{\mathcal{E}(y)}{T_{\mathrm{eff}}}\right), \tag{14}$$

where T_{eff} is an effective temperature and $Z_{\mathrm{B}}(x)$ the normalizing constant. For KL-regularized reward maximization, the policy that is optimal for a reward $r(x, y)$ takes the form (Chaudhari et al., 2025)

$$\pi_r(y \mid x) = \frac{1}{Z_r(x)}\, \pi_{\mathrm{ref}}(y \mid x) \exp\!\left(\frac{r(x, y)}{\beta}\right). \tag{15}$$

Proposition E.3 (Necessary condition)**.** *Suppose $\pi_r(\cdot \mid x) = q_{\mathrm{B}}(\cdot \mid x)$. Then, for every y with $\pi_{\mathrm{ref}}(y \mid x) > 0$,*

$$r(x, y) = -\frac{\beta}{T_{\mathrm{eff}}}\, \mathcal{E}(y) + c(x), \qquad c(x) = \beta \log \frac{Z_r(x)}{Z_{\mathrm{B}}(x)}. \tag{16}$$

Proof. Equating Eqs. 14 and 15 and cancelling the common factor $\pi_{\mathrm{ref}}(y \mid x)$, which is legitimate exactly on $\{y : \pi_{\mathrm{ref}}(y \mid x) > 0\}$ since both distributions are supported there, gives $\exp(r(x, y)/\beta)/Z_r(x) = \exp(-\mathcal{E}(y)/T_{\mathrm{eff}})/Z_{\mathrm{B}}(x)$. Both sides are strictly positive, so taking logarithms is valid and yields $r(x, y)/\beta - \log Z_r(x) = -\mathcal{E}(y)/T_{\mathrm{eff}} - \log Z_{\mathrm{B}}(x)$. Rearranging gives Eq. 16, and the additive term depends on x only. Outside the support of $\pi_{\mathrm{ref}}(\cdot \mid x)$ both densities vanish and r is unconstrained, so the conclusion holds π_{ref}-almost everywhere. □

Proposition E.3 is what makes energy a natural carrier of preference strength: any KL-regularized policy that reproduces an energy-tilted target must have a reward that is affine in $-\mathcal{E}$. The condition is necessary and not sufficient, and it constrains a hypothetical r rather than the r_θ produced by Eq. 5. Physio-DPO neither estimates q_{B} nor minimizes a divergence to it; the properties it does have are Theorems E.1 and E.2, and their empirical consequences are examined through the fold-gating ablations, the sigmoid-sensitivity sweep, the matched multi-seed comparison, and the unseen computational oracles of Appendix B.

F Experimental Details

F.1 Baselines

We compare Physio-DPO against a diverse set of strong baselines covering unaligned models, supervised fine-tuning, reinforcement learning, and preference-based alignment. All methods use the same backbone and LoRA configuration to ensure fair comparison.

Unaligned PLMs. We evaluate the off-the-shelf ProGen2-XL (6.4B) to establish zero-shot generation performance. For cross-architecture reference, we additionally report results for ESM-3 Open (1.4B) and ProtGPT2 (762M), representative causal protein language models.

Supervised Fine-Tuning (SFT). The backbone model is fine-tuned on the Stable (y_w) subset of PhysioPref-1M using standard cross-entropy loss. We adopt the same learning rate as preference-based methods to isolate the effect of the alignment objective. This SFT model serves as the initialization point for all subsequent reinforcement learning and preference optimization methods.

Reinforcement Learning (PPO). We apply standard RLHF with Proximal Policy Optimization. A separate reward model (initialized from ProGen2-Medium) is trained on the PhysioPref-1M dataset to estimate biophysical quality. The policy is initialized from the SFT checkpoint. We use a clip ratio $\epsilon = 0.2$ and an adaptive KL penalty coefficient to maintain a target KL divergence of approximately 0.1 relative to the SFT reference.

Preference Optimization Methods. We compare against representative preference-based approaches, including Standard DPO, IPO, KTO, SimPO, and ORPO. All models are initialized from the SFT checkpoint. Standard DPO follows the official TRL implementation with $\beta = 0.1$. For IPO and KTO, we utilize the default hyperparameters recommended in their respective implementations. These methods use the same preference pairs as Physio-DPO but treat them as binary labels (i.e., $\Psi(\delta_E) = 1$), ignoring the continuous magnitude of the energy gap.

F.2 Evaluation Metrics

This section provides formal definitions and computation procedures for all evaluation metrics reported in the main paper.

Structural Stability: Self-consistency RMSD To measure structural stability, we adopt the self-consistency RMSD (sc-RMSD) protocol (Xiao et al., 2025). Given a generated amino acid sequence y, we first predict its folded structure $\mathcal{S}(y)$ using a structure prediction model (e.g., ESMFold). We then apply inverse folding to obtain a sequence $\hat{y}$ from $\mathcal{S}(y)$, followed by a second folding step to obtain $\mathcal{S}(\hat{y})$. The sc-RMSD is defined as:

$$\text{sc-RMSD}(y) = \text{RMSD}\big(\mathcal{S}(y), \mathcal{S}(\hat{y})\big), \tag{17}$$

where RMSD is computed after optimal rigid-body alignment. Lower sc-RMSD values indicate higher structural self-consistency and stability.

Foldability. Foldability is measured as the proportion of generated sequences whose predicted structure attains a high confidence score. Specifically, given a set of generated sequences $\mathcal{Y}$, foldability is defined as:

$$\text{Foldability} = \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \mathbb{I}\big[\text{pLDDT}(y) > 70\big], \tag{18}$$

where $\text{pLDDT}(y)$ denotes the average predicted Local Distance Difference Test score of the folded structure of y. The threshold of 70 follows standard practice and indicates confident prediction.

Biophysical Validity: Energy Score. Biophysical validity is quantified using Rosetta Energy Units (REU) (Alford et al., 2017). For a generated sequence y with predicted structure $\mathcal{S}(y)$, we compute its total Rosetta energy $\mathcal{E}_{\text{tot}}(y)$ using the standard Rosetta scoring function. To normalize across sequence lengths, we report the per-residue energy:

$$\mathcal{E}(y) = \frac{\mathcal{E}_{\text{tot}}(y)}{|y|}, \tag{19}$$

where $|y|$ denotes the sequence length. Lower energy values correspond to higher predicted stability under this scoring function.

F.3 Diversity Metrics

We evaluate diversity along both distributional and sequence-level dimensions.

Perplexity. Distributional diversity is measured by the perplexity (PPL) of generated sequences under the pretrained reference model π_{ref}. For a sequence $y = (y_1, \ldots, y_T)$, perplexity is defined as:

$$\text{PPL}(y) = \exp\left(-\frac{1}{T}\sum_{t=1}^{T} \log \pi_{\text{ref}}(y_t \mid y_{<t})\right). \tag{20}$$

Lower perplexity indicates higher likelihood under the training distribution.

Sequence Identity. To assess novelty and prevent memorization, we compute the maximum sequence identity between each generated sequence and the training set. Sequence identity is defined as the fraction of identical residues in an optimal pairwise alignment. Lower maximum identity values indicate greater novelty while maintaining plausibility.

F.4 Detailed Implementation Settings

In this section, we provide comprehensive details regarding the model architecture, training infrastructure, dataset construction, and baseline configurations to ensure reproducibility.

F.4.1 Model Architecture and Training Infrastructure

Table 10: Hyperparameter settings for Physio-DPO and baselines trained on ProGen2-XL.

Hyperparameter	Value
Model and adaptation	
Base model	**ProGen2-XL (6.4B)**
Architecture	Decoder-only causal LM
Precision	`bfloat16`
LoRA rank (r)	16
LoRA alpha (α)	32
LoRA dropout	0.05
Target modules	`q_proj, k_proj, v_proj, o_proj, fc_in, fc_out`
Optimization	
Optimizer	AdamW
Learning rate	5×10^{-6}
Learning-rate scheduler	Cosine decay with warmup
Warmup ratio	0.03
Weight decay	0.0
Batch size	64
Gradient accumulation steps	2
Maximum sequence length	512
Training steps	5,000
DPO and Physio-DPO	
β (KL coefficient)	0.1
μ (energy-gap threshold)	50
τ (sigmoid sharpness)	10
Label smoothing	0.0

Table 11: Additional analyses of diversity collapse and memorization. Physio-DPO preserves structural diversity and novelty relative to standard DPO. In the first three rows the comparison column is standard DPO; in the final row it is a naive energy-minimization policy, which is the configuration that collapses.

Analysis	Metric	DPO	Physio-DPO (ours)	Interpretation
Structural diversity	Fold clusters from TM-score clustering (10K samples)	13	**14**	Preserves fold diversity
Diversity concentration	Cluster-size Gini coefficient ↓	0.34	**0.32**	Comparable but slightly more uniform diversity
Structural novelty	Sequences without a close PDB analog (TM-score < 0.6) ↑	69%	**73%**	Improves structural novelty
Naive energy minimization	sc-RMSD ↓ / maximum training sequence identity ↓	1.81 Å 48.9%	**1.28 Å** **33.8%**	Achieves lower sc-RMSD without increasing similarity to training sequences

Base Model. We instantiate our policy network using ProGen2-XL (Nijkamp et al., 2023), a state-of-the-art autoregressive PLM trained on over 1 billion sequences from genomic databases. We employ LoRA to fine-tune the model efficiently. Specifically, we apply low-rank adapters to the query, key, value, and output projection layers of the self-attention mechanism, as well as the inner projection layers of the feed-forward networks, ensuring robust adaptation while maintaining the generalizability of the pre-trained weights.

Compute Environment. All experiments are conducted on a cluster of 4 $\times$ NVIDIA A100 (80GB) GPUs interconnected via NVLink. We utilize the HuggingFace ecosystem (`transformers`, `peft`, and `trl`) for training. Given the 6.4B parameter size of ProGen2-XL, the model fits comfortably within the VRAM of A100 GPUs. Therefore, we employ DeepSpeed ZeRO-2 optimization to maximize training throughput and support larger batch sizes, without necessitating CPU offloading strategies required for larger models. Mixed-precision training (`bfloat16`) is enabled to accelerate convergence and reduce memory footprint.

Hyperparameters. We optimize the model using *AdamW* (Loshchilov & Hutter, 2017) with a cosine learning rate decay and a linear warm-up phase covering 3% of total steps. The effective batch size is maintained at 64. For the Physio-DPO objective, the scaling parameters (μ, τ) were selected based on preliminary ablation studies to balance the sensitivity to energy gaps. Detailed hyperparameters are listed in Table 10.

G Additional Analysis

This section provides additional analysis of our proposed Physio-DPO, including scaling laws of physical alignment, Ramachandran density analysis, and sequence length analysis. Note that the setup is the same as in the main paper.

G.1 Scaling Laws of Physical Alignment

As shown in Figure 7, we investigate the scaling properties of physical alignment across the **ProGen2 model family** (151M, 764M, 2.7B, and 6.4B). Our proposed Physio-DPO (Blue) exhibits a significantly steeper scaling trajectory compared to the SFT baseline (Red). While smaller models (151M, 764M) show limited gains, the performance gap widens distinctively at the 6.4B scale. This suggests that sufficient model capacity is a prerequisite for effectively internalizing complex thermodynamic constraints from continuous energy-weighted supervision.

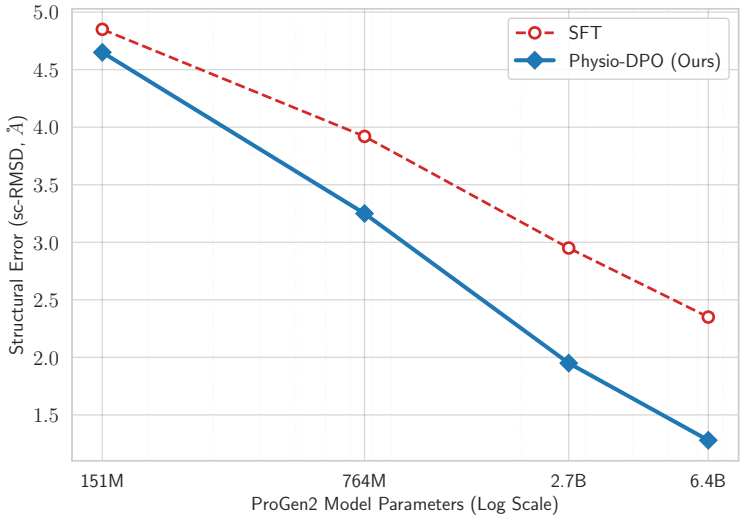

Figure 7: Scaling laws of physical alignment.

G.2 Ramachandran Density Analysis

We visualize the distribution of backbone torsion angles (ϕ, ψ) for 100,000 residues, as shown in Figure 8. While SFT generates outliers in biologically forbidden regions (white space), Physio-DPO recovers the precise allowable manifolds of natural proteins (α-helical and β-sheet regions), demonstrating implicit learning of stereochemical rules.

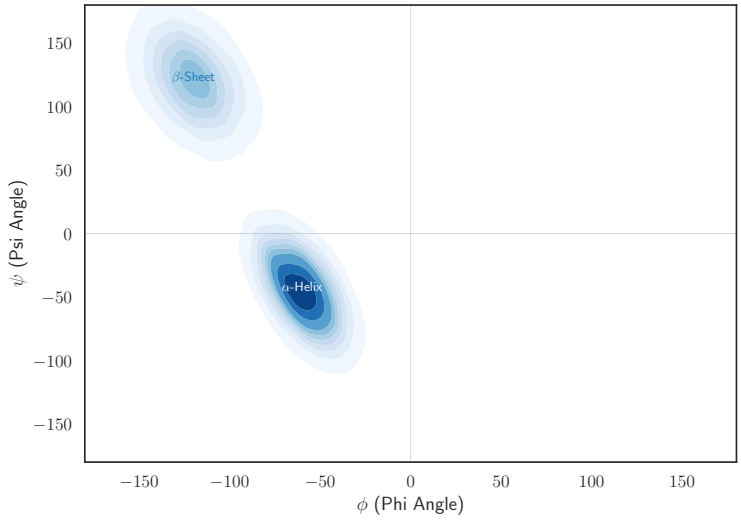

Figure 8: Ramachandran density of Physio-DPO.

G.3 Sequence Lengths

Transformer are easily distracted by long sequences, leading to collapsed long protein structures. We evaluate the robustness of Physio-DPO on sequences of different lengths, as shown in Figure 9. SFT degrade significantly for long sequences (> 400 residues) due to attention diffusion. Physio-DPO maintains consistent structural quality across all length bins, effectively mitigating the "long-range collapse" problem.

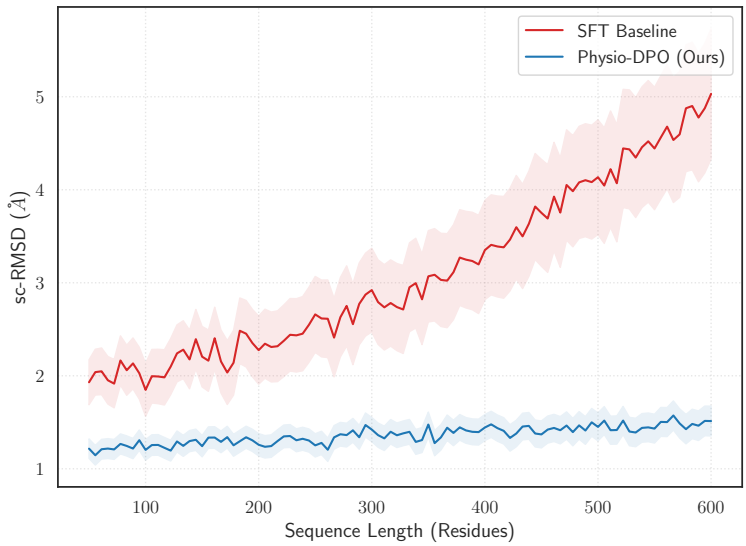

Figure 9: Sequence length analysis.

G.4 Diversity Collapse and Memorization Analysis

A potential concern is that energy-informed optimization may collapse diversity toward a narrow set of low-energy motifs or increase memorization of the training set. To evaluate this, we compare Physio-DPO against standard DPO on structural diversity, diversity concentration, and structural novelty, and further include a naive energy minimization ablation. As shown in Table 11, Physio-DPO preserves fold diversity, yields a slightly more uniform cluster distribution, and improves structural novelty relative to standard DPO. In contrast, naive energy minimization leads to clear diversity collapse and substantially higher memorization, whereas Physio-DPO avoids this pathology while improving structural plausibility.

G.5 Pair Geometry and Fold-Gating Ablations

Pair geometry. PhysioPref-1M contains generation-mined and physics-perturbed preference pairs. Generation-mined pairs are constructed from independently sampled sequences and therefore need not share a common fold. In contrast, each physics-perturbed pair contains a native structure and a PyRosetta decoy derived from the same backbone. Table 12 quantifies the geometric differences between these two construction strategies.

Table 12: Pair geometry in PhysioPref-1M, measured between the preferred and dispreferred structures. TM-score and RMSD are reported as medians.

Pair construction	**Backbone matched**	**Median TM-score**	**Median RMSD (Å)**	**CATH topology agreement**
Generation-mined (stable vs. unstable)	No	0.28	12.4	6.1%
Generation-mined (stable vs. high-energy)	No	0.31	11.2	7.8%
Physics-perturbed (native vs. PyRosetta decoy)	**Yes**	**0.93**	**2.1**	**100%**

Only the physics-perturbed subset supports a within-fold interpretation of the energy gap. Its pairs have a median TM-score of 0.93, a median RMSD of 2.1 Å, and complete CATH-topology agreement. By contrast, the generation-mined pairs typically contain unrelated folds. Per-residue energy normalization controls for sequence length but does not resolve this topological mismatch.

Fold-gated objective. The weighting is therefore applied only to the backbone-matched, physics-perturbed subset, through the gate w_i of Eq. 3 entering the objective of Eq. 5. Generation-mined pairs retain the ordinary binary DPO term and receive no cross-fold energy-magnitude weighting.

Ablation over weighted pair subsets. To isolate the effect of fold gating, we compare standard DPO, energy-gap weighting over all preference pairs, weighting restricted to the physics-perturbed subset, and the final fold-gated objective. All configurations use the same model, training data, and evaluation protocol. The results are reported in Table 13.

Weighting only the physics-perturbed pairs recovered most of the improvement over standard DPO. It reduced sc-RMSD from 1.82 to 1.33 Å and increased foldability from 83.6% to 91.6%. Retaining the generation-mined pairs through ordinary binary DPO further improved sc-RMSD to 1.30 Å and foldability to 92.4%.

Weighting all pairs produced slightly stronger single-run values, but it assigned energy-gap magnitudes to generation-mined pairs that often contained unrelated folds. The fold-gated objective removed this cross-fold weighting at a cost of 0.02 Å in sc-RMSD and 0.4 percentage points in foldability. Unless otherwise stated, all subsequent Physio-DPO experiments use the fold-gated objective.

G.6 Broader Evaluation Beyond Structural Metrics

To further assess whether Physio-DPO improves broader generalization beyond the structural metrics reported in the main text, we conduct two additional analyses. First, we evaluate *de novo* fold designability by measuring how often stable generated proteins remain structurally dissimilar to known proteins in the PDB (Table 14). Second, we assess orthogonal biophysical properties related

Table 13: Ablation over the preference-pair subsets receiving energy-gap weighting. Single run, 30,000 generated sequences. Bold marks the strongest numerical result and underlining the selected fold-gated configuration.

Training objective	**Use of δ_E**	**sc-RMSD (Å) ↓**	**Foldability (%) ↑**	**Energy (REU) ↓**
Standard DPO (all pairs)	Not used	1.82	83.6	-2.65
Standard DPO, physics-perturbed pairs only	Not used	1.79	84.0	–
Physio-DPO, generation-mined pairs only	Cross-fold pairs only	1.66	86.1	–
Physio-DPO, Ψ on all pairs	All pairs, including cross-fold pairs	**1.28**	**92.8**	**-3.05**
Physio-DPO, physics-perturbed pairs only	Backbone-matched pairs only	1.33	91.6	-2.98
Physio-DPO, fold-gated	Ψ on backbone-matched pairs; DPO on remaining pairs	<u>1.30</u>	<u>92.4</u>	<u>-3.02</u>

to practical viability, including predicted solubility and aggregation propensity (Table 15). Together, these analyses provide complementary evidence that the gains of Physio-DPO are not limited to foldability alone, but extend to novelty and biophysical usability.

Table 14: Novelty and *de novo* fold designability evaluated on 3,000 unconditional samples. "Novel and stable" denotes sequences with pLDDT > 70 and a maximum TM-score < 0.5 to any known PDB structure.

Method	**Maximum TM-score to PDB ↓**	**Novel and stable rate (%) ↑**
SFT	0.72	12.4
Standard DPO	0.65	18.2
Physio-DPO	**0.54**	**31.5**

Table 15: Orthogonal evaluation of biophysical viability. Protein-Sol predicts protein solubility, whereas A3D measures aggregation propensity. A Protein-Sol score > 0.45 is classified as soluble.

Method	**Protein-Sol score ↑**	**Soluble fraction (%) ↑**	**A3D score ↓**
SFT	0.42	41.5	1.85
Standard DPO	0.46	52.3	1.42
Physio-DPO	**0.58**	**78.6**	**0.84**

G.7 Sigmoid Sensitivity

Sensitivity to sigmoid sharpness. We evaluated the sensitivity of Physio-DPO to the sigmoid sharpness τ by testing $\tau \in \{2, 5, 10, 20, 50\}$ with the threshold fixed at $\mu = 50$. As shown in Table 16, performance changed little across the intermediate settings $\tau \in \{5, 10, 20\}$: sc-RMSD ranged from 1.30 to 1.34 Å, and foldability ranged from 91.8% to 92.4%. The best values for both metrics were obtained at $\tau = 10$. Halving or doubling this value increased sc-RMSD by at most 0.04 Å and reduced foldability by at most 0.6 percentage points, suggesting that performance is relatively insensitive to the precise choice of τ within this range.

Both extreme settings yielded poorer results. At $\tau = 2$, sc-RMSD increased to 1.41 Å and foldability fell to 90.1%; at $\tau = 50$, these values reached 1.46 Å and 89.3%, respectively. This pattern is consistent with the role of τ in controlling the transition around the threshold. A small τ makes the weights change abruptly near μ, whereas a large τ flattens the weighting function and reduces the differences between weights. The intermediate settings retain a smoother transition while preserving

Table 16: Sensitivity of the fold-gated Physio-DPO objective to the sigmoid sharpness τ at $\mu = 50$. Results are obtained from a single run. The setting used in the main experiments is shown in bold.

τ	**2**	**5**	**10**	**20**	**50**
sc-RMSD (Å) ↓	1.41	1.34	**1.30**	1.32	1.46
Foldability (%) ↑	90.1	91.8	**92.4**	92.1	89.3

these differences. We therefore use $\tau = 10$ in the main experiments, as it achieved the lowest sc-RMSD and highest foldability among the tested settings.