PROTEINZERO: SELF-IMPROVING PROTEIN GENERATION VIA ONLINE REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein generative models have shown remarkable promise in protein design, yet their success rates remain constrained by reliance on curated sequence-structure datasets and by misalignment between supervised objectives and real design goals. We present ProteinZero, an online reinforcement learning framework for inverse folding models that enables scalable, automated, and continuous self-improvement with computationally efficient feedback. ProteinZero employs a reward pipeline that combines structural guidance from ESMFold with a novel self-derived ddG predictor, providing stable multi-objective signals while avoiding the prohibitive cost of physics-based methods. To ensure robustness in online RL, we further introduce a novel embedding-level diversity regularizer that mitigates mode collapse and promotes functionally meaningful sequence variation. Within a general RL formulation balancing multi-reward optimization, KL-divergence from a reference model, and diversity regularization, ProteinZero achieves robust improvements across designability, stability, recovery, and diversity. On the CATH-4.3 benchmark, it consistently outperforms state-of-the-art baselines including ProteinMPNN, ESM-IF, and InstructPLM, reducing design failure rates by 36-48% and achieving success rates above 90% across diverse folds. Importantly, a complete RL run can be executed on a single 8×GPU node within three days, including reward computation and data generation. These results indicate that efficient online RL fine-tuning can complement supervised pretraining by allowing protein generative models to evolve continuously from their own outputs and optimize multiple design objectives without labeled data, opening new possibilities for exploring the vast protein design space.

1 INTRODUCTION

Protein design and engineering represent one of the most promising frontiers in computational biology, with applications spanning drug discovery to novel enzymes (Dauparas et al., 2022; Hsu et al., 2022; Wang et al., 2023a). A central challenge is protein inverse folding: generating amino acid sequences that fold into desired three-dimensional structures (Jing et al., 2021; Zhang & Skolnick, 2005), serving as the foundation for fixed backbone sequence design. This task is crucial as protein backbone structure and side-chain conformation jointly determine functionalities like binding and catalytic interactions. However, optimizing functional properties requires first establishing high designability (designed sequences correctly folding into target structures) and thermodynamic stability (free energy difference favoring folded over unfolded states) as foundational prerequisites. Rocklin et al. (2017) demonstrated that 70-80% of computationally designed proteins fail due to misfolding or instability, with failures persisting in state-of-the-art AI methods (Bennett et al., 2023; Tsuboyama et al., 2023). Moreover, tiny (≈1-2 Å) atomic shifts at binding interfaces can disrupt hydrogen-bond geometry and packing, causing large affinity and specificity losses (failure to distinguish intended from off-target binders) (Clackson & Wells, 1995; Bogan & Thorn, 1998; Bajusz et al., 2021).

Recent deep learning breakthroughs including ProteinMPNN (Dauparas et al., 2022), ESM-IF (Hsu et al., 2022), and graph-based methods (Jing et al., 2021; Wang et al., 2023a) have significantly improved inverse folding accuracy. However, these methods train on paired sequence-structure data from the Protein Data Bank (PDB) which, while valuable, represent a minuscule fraction of the protein sequence space (Dauparas et al., 2022; Hsu et al., 2022; Qiu et al., 2024) and exhibit limited diversity and natural biases. This data scarcity creates a ceiling for model performance and restricts

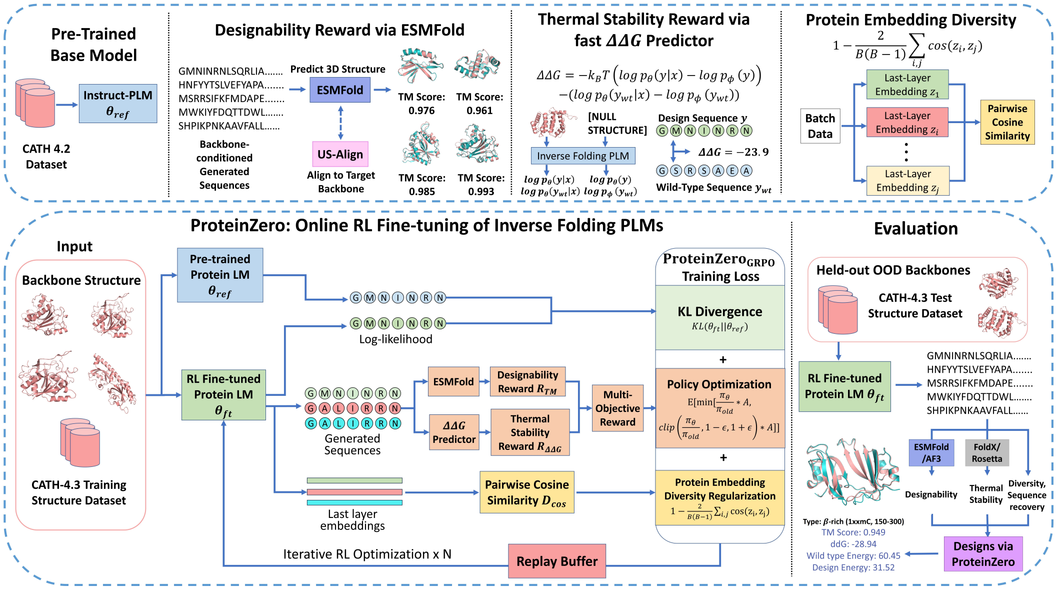

Pre-Trained Base Model
CATH 4.2 Dataset
Instruct-PLM θ_{ref}
Designability Reward via ESMFold
GMNINRNLSQRLIA.......
HNFYYTSLVEFYAPA.......
MSRRSIFKFMDAPE.......
MWKIYFDQTTDWL.......
SHPIKPNKAAVFALL......
Backbone-conditioned Generated Sequences
Predict 3D Structure
ESMFold
US-Align
Align to Target Backbone
TM Score: 0.976
TM Score: 0.961
TM Score: 0.985
TM Score: 0.993
Thermal Stability Reward via fast $\Delta\Delta G$ Predictor
$\Delta\Delta G = -k_B T \left(\log p_\theta(y|x) - \log p_\phi(y)\right) - (\log p_\theta(y_{wt}|x) - \log p_\phi(y_{wt}))$
[NULL STRUCTURE]
Inverse Folding PLM
$\log p_\theta(y|x)$
$\log p_\theta(y_{wt}|x)$
$\log p_\theta(y)$
$\log p_\phi(y_{wt})$
Design Sequence y
G M N I N R N
$\Delta\Delta G = -23.9$
G S R S A E A
Wild-Type Sequence y_{wt}
Protein Embedding Diversity
$1 - \frac{2}{B(B-1)} \sum_{i,j} cos(z_i, z_j)$
Batch Data
Last-Layer Embedding z_1
Last-Layer Embedding z_i
Last-Layer Embedding z_j
Pairwise Cosine Similarity
ProteinZero: Online RL Fine-tuning of Inverse Folding PLMs
Input
Backbone Structure
CATH-4.3 Training Structure Dataset
Pre-trained Protein LM θ_{ref}
RL Fine-tuned Protein LM θ_{ft}
G M N I N R N
G M N I N R N
Log-likelihood
G M N I N R N
G A L I R R N
G A L I R R N
Generated Sequences
ESMFold
Designability Reward R_{TM}
$\Delta\Delta G$ Predictor
Thermal Stability Reward $R_{\Delta\Delta G}$
Multi-Objective Reward
Last layer embeddings
Pairwise Cosine Similarity D_{cos}
ProteinZero$_{GRPO}$ Training Loss
KL Divergence
$KL(\theta_{ft}||\theta_{ref})$
+
Policy Optimization
$E[\min[\frac{\pi_\theta}{\pi_{old}} * A, clip\left(\frac{\pi_\theta}{\pi_{old}}, 1-\epsilon, 1+\epsilon\right) * A]]$
+
Protein Embedding Diversity Regularization
$1 - \frac{2}{B(B-1)} \sum_{i,j} \cos(z_i, z_j)$
Iterative RL Optimization x N
Replay Buffer
Evaluation
Held-out OOD Backbones
CATH-4.3 Test Structure Dataset
RL Fine-tuned Protein LM θ_{ft}
GMNINRNLSQRLIA.......
HNFYYTSLVEFYAPA.......
MSRRSIFKFMDAPE.......
MWKIYFDQTTDWL.......
SHPIKPNKAAVFALL......
ESMFold /AF3
FoldX/ Rosetta
Designability
Thermal Stability
Diversity, Sequence recovery
Type: β-rich (1xxmC, 150-300)
TM Score: 0.949
ddG: -28.94
Wild type Energy: 60.45
Design Energy: 31.52
Designs via ProteinZero

Figure 1: **ProteinZero framework. Upper:** Online RL components: ESMFold-based designability (TM-score via US-Align), $\Delta\Delta G$ predictor using backbone-conditioned likelihoods, and embedding diversity regularization. **Lower:** Iterative training where inverse folding models generate sequences, receive multi-objective rewards, and update with KL constraints and diversity regularization. Held-out CATH-4.3 evaluation demonstrates substantial improvements across all key design metrics.

exploration of novel protein designs beyond known natural and synthetic settings (Fujimoto & Gu, 2021; Shumailov et al., 2024). Moreover, there is a misalignment between the supervised learning task of inverse folding and actual objectives in real-world protein design, where applications require high designability, thermal stability, and sequence diversity (providing numerous reliable candidates for experimental validation rather than converging to known patterns) (Watson et al., 2023; Ingraham et al., 2023). Existing alignment efforts have focused on RL-finetuning structural generative models (Campbell et al., 2024; Huguet et al., 2024; Zhou et al., 2024; Gasser et al., 2024; Park et al., 2024), achieving only single- or few-round alignment with curated offline datasets, limiting exploration to known successes rather than discovering novel design principles through iterative feedback.

We propose ProteinZero, an online RL fine-tuning framework that addresses multi-objective optimization challenges in protein design, enabling automated self-improvement of inverse folding models while balancing designability, stability, and diversity. Our contributions are:

1. We present ProteinZero, achieving stable multi-round self-improvement in protein sequence design through continuous exploration without curated preference datasets.
2. We introduce a self-derived $\Delta\Delta G$ estimator computed from the inverse folding model using backbone-conditioned likelihoods normalized by unconditional priors. Combined with ESMFold-based designability rewards, this enables computationally tractable multi-objective online RL optimization (see Table 6).
3. We develop a novel diversity regularizer operating in protein embedding space rather than sequence space, preventing mode collapse (Shumailov et al., 2024; Alemohammad et al., 2024; Holtzman et al., 2020) while maintaining functional coherence.
4. We elucidate the design space of RL fine-tuning by examining algorithms (GRPO, RAFT, DPO, multi-round DPO), rewards, and regularization strategies, identifying optimal configurations for stable multi-objective optimization without mode collapse.
5. Extensive experiments demonstrate that ProteinZero outperforms existing methods across all key metrics, achieving 36-48% reduction in design failure rates versus ProteinMPNN (Dauparas et al., 2022), ESM-IF (Hsu et al., 2022), and InstructPLM (Qiu et al., 2024), with significant improvements in structural accuracy, stability, and diversity across diverse protein folds including challenging long chains.

2 Related Work

Protein Inverse Folding Models. Inverse folding generates amino acid sequences $y = (y_1, ..., y_L)$ for target structures x, formulated as conditional generation $p_\theta(y|x)$ with model parameters θ trained via supervised loss on PDB pairs. Ingraham et al. (2019) pioneered graph neural networks for this task, extended by ProteinMPNN (Dauparas et al., 2022) with noise-aware training. ESM-IF (Hsu et al., 2022) leveraged pretrained language models, while GVP-GNN (Jing et al., 2021), StructTrans (Wang et al., 2023a), PiFold (Gao et al., 2023), and GraDe-IF (Yi et al., 2023) introduced geometric representations, transformers, co-design, and diffusion respectively. InstructPLM (Qiu et al., 2024) achieved SOTA by adapting frozen language models via structural prompts (our base architecture). While achieving strong benchmarks, supervised approaches face inherent constraints: limited PDB datasets restrict exploration of the vast sequence space, and their objectives, optimizing sequence recovery, may not align with real design goals of maximizing stability, designability, and diversity. We extend these foundations through online RL with efficient proxy rewards, enabling continuous learning from self-generated sequences to directly optimize these multiple design objectives.

RLHF of Protein Generative Models. Classical RL approaches to biological sequence design (Angermueller et al., 2020; Runge et al., 2019) train task-specific policies from scratch via unconditional generation or local mutations, a different paradigm detailed in Appendix D.1. With powerful pre-trained protein models, Reinforcement Learning from Human Feedback (RLHF) has emerged for fine-tuning generative models. RLHF transforms models through online methods like PPO (Schulman et al., 2017), GRPO (Shao et al., 2024), and RAFT (Dong et al., 2023) that optimize rewards directly, and offline methods like DPO (Rafailov et al., 2023) using static preference datasets. While successful in LLMs, applying online RLHF to protein models faces both computational (Table 6) and reward modeling infrastructure challenges. Current protein RLHF work encompasses diverse architectures: Campbell et al. (2024) and Huguet et al. (2024) enhance structural generation with ReFT, Zhou et al. (2024) applies DPO for antibody design, Xu et al. (2025) employ multi-round DPO for inverse folding with structural feedback, ResiDPO implements residue-level DPO with pLDDT scores (Xue et al., 2025), and Wang et al. (2025) introduces online fine-tuning for discrete diffusion sequence models through direct reward backpropagation via Gumbel-Softmax approximations, which requires differentiable rewards. These methods primarily address structure generation or co-design tasks, with most operating offline. Offline approaches rely on pre-collected rewards without iterative learning from self-generated sequences, limiting exploration of protein design space. We introduce online RL for inverse folding models using policy gradients with non-differentiable reward proxies, enabling self-improvement in designability, stability, and diversity.

Diversity Regularization. Promoting diversity in protein generative models is crucial for increasing downstream success rates and maintaining exploration capability in online RL to avoid mode collapse and reward hacking (Ouyang et al., 2022; Fan et al., 2025; Shumailov et al., 2024) (see Appendix D.2). Prior work explored sequence-level metrics: Park et al. (2024) employ Hamming distance as diversity regularizer, operating on raw sequences instead of structure-aware representations. The DPO-based approach faces challenges in simultaneously optimizing rewards and diversity. We introduce embedding-level diversity regularization that operates in the model's embedding space, promoting functionally meaningful variation while preventing mode collapse and maintaining structural coherence (theoretical derivation in Appendix F, empirical dynamics in Appendix C.3).

3 Method

We propose ProteinZero, a framework that fine-tunes protein generative models through online reinforcement learning. Our approach optimizes a reward-based objective $\mathcal{J}_{\text{RL}}(\theta) = \mathbb{E}_{x\sim\mathcal{D}_x, y\sim p_\theta(\cdot|x)}[r(x,y)] - \alpha_{\text{KL}} \cdot \text{KL}\left(p_\theta(\cdot \mid x) \| p_{\text{ref}}(\cdot \mid x)\right)$, where $r(x, y)$ combines multiple design objectives including designability and stability, p_{ref} is a reference model (typically the pre-trained model), and α_{KL} controls divergence from the reference.

3.1 ProteinZero Framework: Addressing Mode Collapse in Online RL for Proteins

To realize this objective while preventing mode collapse, ProteinZero couples reward optimization with novel diversity constraints, enabling stable and effective online learning. The framework enables continuous exploration beyond pre-collected datasets, discovering novel design principles through iterative feedback.

Reinforcement learning for protein design requires optimizing a model to generate sequences that maximize a reward function while maintaining reasonable proximity to a reference model. In practice, recent RL fine-tuning methods can be unified in this general objective through specialized algorithms that balance exploitation and exploration: $\mathcal{L}(\theta) = \mathcal{L}_{\mathrm{RL}}(\theta) + \mathcal{L}_{\mathrm{KL}}(\theta)$. For instance, in the Group Relative Policy Optimization (GRPO) algorithm, this objective is realized as $\mathcal{L}_{\mathrm{GRPO}}(\theta) = -\mathbb{E}_{(x,y)\sim\mathcal{B}}\left[\min(\mathrm{A}*\frac{\mathrm{p}_\theta}{\mathrm{p}_{\theta_{\mathrm{old}}}}, \mathrm{A}*\mathrm{clip}(\frac{\mathrm{p}_\theta}{\mathrm{p}_{\theta_{\mathrm{old}}}}, 1\text{-}\epsilon, 1+\epsilon))\right] + \alpha_{\mathrm{KL}} \cdot \mathrm{KL}\left(p_\theta \| p_{\mathrm{ref}}\right)$ (Shao et al., 2024), where A is the advantage function, $p_{\theta_{old}}$ is learned policy of last iteration. However, we observed that protein generative models suffer from mode collapse in online RL fine-tuning, converging to a narrow set of solutions that maximize rewards without diversity (see Appendix C.3). Thus, we incorporate a diversity regularization term, resulting in a more comprehensive objective:

$$\mathcal{L}(\theta) = \mathcal{L}_{\mathrm{RL}}(\theta) + \mathcal{L}_{\mathrm{KL}}(\theta) + \mathcal{L}_{\mathrm{Div}}(\theta) \tag{1}$$

While diversity can be promoted by incorporating it directly into the reward, our experiments show this often causes training instability and performance degradation (see Tables 5 and 3). Thus, ProteinZero applies diversity regularization at the representation level through a separate loss $\mathcal{L}_{\mathrm{Div}}(\theta)$, encouraging diversity while preserving the integrity of the main reward optimization (Table 1 and Figure 5).

To enable practical online RL fine-tuning, we address two critical challenges: (1) the lack of effective diversity regularization for protein models, and (2) the prohibitive computational cost of reward evaluation, which can extend training to months. We therefore propose embedding-level diversity regularization and fast reward modeling to make online fine-tuning practically achievable.

3.1.1 Embedding-Level Diversity Regularization for Mode Collapse Mitigation

To address mode collapse in protein generative models during online RL fine-tuning, we propose a novel diversity regularization operating at the protein embedding level. Unlike token-level diversity metrics which can compromise functional properties, our approach leverages learned representations shown to encode hierarchical biological information from local patterns to functional domains (Simon & Zou, 2024), with embedding distances reflecting functional relationships (Schmirler et al., 2024; Corso et al., 2021; Blaabjerg et al., 2024). For each protein sequence in a batch, we compute a fixed-dimensional embedding vector by aggregating the last-layer decoder activations:

$$z_i(\theta) = \frac{\sum_t m_{i,t} h_{i,t}}{\sum_t m_{i,t}} \in \mathbb{R}^d, \quad 1 \leq i \leq B \tag{2}$$

where $h_{i,t} \in \mathbb{R}^d$ is the decoder activation at position t for sample i, and $m_{i,t} \in \{0, 1\}$ an attention mask. These protein embeddings are ℓ_2-normalized before computing a cosine-based diversity score: $D_{\cos}(\theta; \mathcal{B}) = 1 - \overline{\cos} \in [0, 1]$, where $\overline{\cos} = \frac{2}{B(B-1)} \sum_{1 \leq i < j \leq B} \cos(z_i, z_j)$. The diversity regularization term is incorporated as:

$$\mathcal{L}_{\mathrm{Div}}(\theta) = -\alpha_{\mathrm{div}} \cdot D_{\cos}(\theta; \mathcal{B}) \tag{3}$$

Since z_i depends on θ, this provides informative gradients that foster the generation of diverse, functionally plausible sequences. Our theoretical analysis (Appendix F) demonstrates how this embedding-based approach mitigates mode collapse in online RL, a contribution applicable beyond protein design. Note that while we optimize embedding-level diversity during training, our evaluation employs standard Hamming distance between sequences to provide an orthogonal assessment of sequence-level diversity. The embedding-level formulation achieves diversity preservation and training stability, validated in ablation studies (Tables 5 and 3) and training dynamics (Appendix C.3).

3.1.2 Fast Proxy Rewards: Enabling Practical Online RL Training

AlphaFold's MSA and template searches and FoldX's physics calculations (Table 6) require minutes to hours per protein, making online RL infeasible. We address this with two fast proxy rewards:

Designability Reward: We use ESMFold (Hsu et al., 2022) for structural inference, leveraging its alignment-free, single-pass architecture instead of AlphaFold2/3's MSA searches and recycling steps (Jumper et al., 2021; Abramson et al., 2024). Our designability reward $r_{\text{TM}}(x, y)$ specifically uses the TM-score from US-Align Zhang et al. (2022), an updated version of TM-Align (Zhang & Skolnick, 2005), computed between ESMFold-predicted and target structures, explicitly not ESMFold's internal confidence score pTM, ensuring our optimization targets actual structural alignment through length-normalized distance-weighted C_α overlaps, not prediction confidence.

Thermal Stability Reward: We propose a novel thermal stability reward $r_{\Delta\Delta G}(x, y)$, serving as a backbone-specific folding-energy surrogate for single-chain proteins, referenced to the PDB wild-type. Because our monomeric setting lacks an inter-chain interface, the unbound-state term required by the Boltzmann-aligned estimator (BA-DDG) (Jiao et al., 2025) is unevaluable. Instead, drawing on evidence that backbone-conditioned likelihoods reflect folding stability (Shanker et al., 2024; Widatalla et al., 2024; Cagiada et al., 2025; Zheng et al., 2023; Ingraham et al., 2019), we normalize this likelihood with an unconditional sequence prior and anchor it to the wild-type baseline:

$$\Delta\Delta G(x, y) = -\,k_B T[(\log p_\theta(y \mid x) - \log p_\varphi(y)) - (\log p_\theta(y_{\text{wt}} \mid x) - \log p_\varphi(y_{\text{wt}}))], \tag{4}$$

where $p_\theta(y \mid x)$ is the backbone-conditioned inverse-folding likelihood, $p_\varphi(\cdot)$ the unconditional sequence prior, y_{wt} the PDB wild-type sequence, and $k_B T$ the thermal energy at 298 K ($0.593\,\text{kcal}\,\text{mol}^{-1}$). The prior $p_\varphi(\cdot)$ is obtained by running the same inverse-folding network (e.g., ProteinMPNN or InstructPLM) with coordinate channels masked, converting it into a sequence-only language model capturing residue-frequency and chain-length distributions of proteins. Subtracting $\log p_\varphi(y)$ from $\log p_\theta(y \mid x)$ removes background amino-acid composition and chain-length preferences, isolating backbone-specific excess compatibility of candidate sequence y. Hence, using y_{wt} as reference yields a computationally efficient $\Delta\Delta G$ surrogate for monomeric stability optimization.

Multi-objective reward: Our final reward combines both scores after min-max normalization to balance scale differences. Normalization is performed across the candidate pool of inverse folding sequences generated for the same backbone within each reinforcement learning iteration: $\tilde{r}_{\text{TM}} = (r_{\text{TM}} - r_{\text{TM}}^{\min})/(r_{\text{TM}}^{\max} - r_{\text{TM}}^{\min})$ and $\tilde{r}_{\Delta\Delta G}$ analogously, giving $r(x, y) = \lambda_{\text{TM}}\tilde{r}_{\text{TM}}(x, y) + \lambda_{\Delta\Delta G}\tilde{r}_{\Delta\Delta G}(x, y)$. This reward accelerates evaluation 25-100× depending on protein length (Table 6), reducing training time from months to days. Our experiments show substantial thermodynamic stability improvements with high structural fidelity (see Figure 5 and Table 1).

3.2 ProteinZero Algorithms: Diversity-Regularized RAFT and GRPO

Building upon our general framework, we implement two online RL algorithms for fine-tuning inverse folding models: RAFT and GRPO. We adapt both methods to incorporate our dual-objective reward system, designability scores from ESMFold structures evaluated by US-Align, and self-derived $\Delta\Delta G$ for thermodynamic stability, alongside embedding-level diversity regularization. These adaptations enable different optimization strategies (detailed in Sections 3.2.1 and 3.2.2).

3.2.1 ProteinZero$_{\text{RAFT}}$: Reward-ranked Fine-tuning with Embedding Diversity

RAFT (Dong et al., 2023) transforms RL into a supervised learning problem by iteratively filtering model outputs based on rewards. Our adaptation generates multiple candidate sequences per target structure, evaluates them with our efficient reward, and retains only the best to form a filtered dataset. Unlike conventional RAFT that incorporates KL-divergence into the reward, we separate the KL term and add our embedding-based diversity regularization ($\mathcal{L}_{\text{CE}}$ is the cross entropy loss):

$$\mathcal{L}_{\text{ProteinZero}_{\text{RAFT}}}(\theta) = \mathcal{L}_{\text{CE}}\left(\theta; \mathcal{D}_{\text{filtered}}\right) + \alpha_{\text{KL}} \cdot \text{KL}\left(p_\theta \| p_{\text{ref}}\right) - \alpha_{\text{div}} \cdot D_{\cos}\left(\theta; \mathcal{D}_{\text{filtered}}\right) \tag{5}$$

3.2.2 ProteinZero$_{\text{GRPO}}$: Embedding-Diversified Policy Optimization

GRPO (Shao et al., 2024) directly optimizes the policy via a trust-region objective:

$$\begin{aligned}\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{x \sim P(X), \{y_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(Y \mid x)} \frac{1}{G} \sum_{i=1}^{G} \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \min \Bigg[& \frac{\pi_\theta\left(y_{i,t} \mid x, y_{i,<t}\right)}{\pi_{\theta_{\text{old}}}\left(y_{i,t} \mid x, y_{i,<t}\right)} \hat{A}_{i,t}, \\ \text{clip}\left(\frac{\pi_\theta\left(y_{i,t} \mid x, y_{i,<t}\right)}{\pi_{\theta_{\text{old}}}\left(y_{i,t} \mid x, y_{i,<t}\right)}, 1-\varepsilon, 1+\varepsilon\right) \hat{A}_{i,t} \Bigg] & - \beta \mathbb{D}_{KL}\left[\pi_\theta \| \pi_{ref}\right],\end{aligned} \tag{6}$$

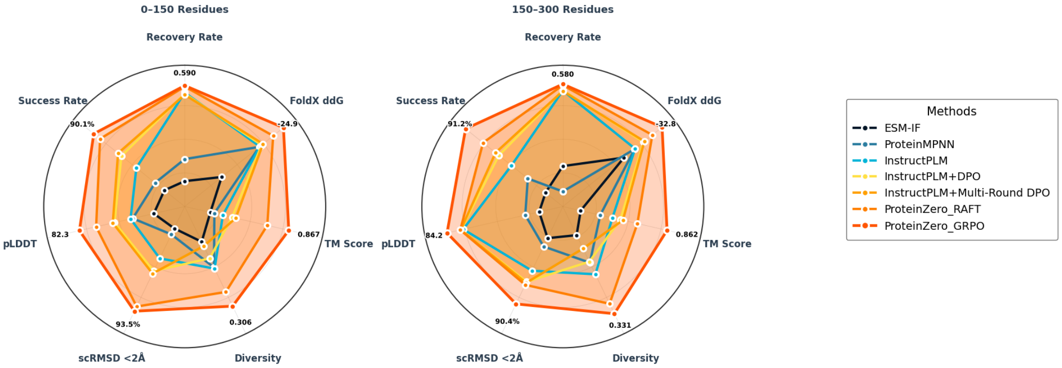


Figure 2: Performance comparison across seven evaluation metrics (Recovery Rate, Stability, TM Score, pLDDT, Diversity, scRMSD <2Å%, and Success Rate) for 0-150 residue proteins (left) and 150-300 residue proteins (right). ProteinZero variants achieve the highest across all metrics.

where ε and β are hyperparameters, and $\hat{A}_{i,t}$ is the advantage calculated from relative rewards within each group. The group relative advantage calculation aligns well with our reward models. Unlike methods that add KL penalty to rewards, GRPO directly adds KL divergence to the loss. We extend this formulation by incorporating our embedding-level diversity regularization ($\mathcal{L}_{\text{GRPO}} = -\mathcal{J}_{\text{GRPO}}$):

$$\mathcal{L}_{\text{ProteinZero}_{\text{GRPO}}}(\theta) = \mathcal{L}_{\text{GRPO}}(\theta) - \alpha_{\text{div}} \cdot D_{\cos}(\theta; \mathcal{B}) \tag{7}$$

Both algorithms effectively implement our ProteinZero framework but approach optimization differently. Our experiments demonstrate that both methods significantly outperform baselines, with ProteinZero$_{\text{GRPO}}$ consistently achieving superior performance across evaluated metrics.

4 Experiment

4.1 Experimental Setup

Evaluation. We evaluated ProteinZero on CATH-4.3 (Orengo et al., 1997), maintaining the train-test-validation split from Hsu et al. (2022). Our test set excluded structures with $> 40\%$ sequence identity to training proteins of 0-150 residues and $> 30\%$ identity to proteins of 150-300 residues, enabling assessment of out-of-distribution generalization. We trained and evaluated models separately for each length category (0-150, 150-300). We evaluate the models with a comprehensive set of metrics, including designability metrics measured by both ESMFold and AF3 (TM Score, PLDDT, scRMSD), stability measured with our fast-ddG predictor and physics-based FoldX/Rosetta ddG (Schymkowitz et al., 2005)), sequence recovery, and sequence Diversity — see Appendix B.3 for detailed definitions. Overall Success Rate was defined as achieving both scRMSD $< 2\,\text{Å}$ and FoldX ddG < 0, inspired by Wang et al. (2025).

ProteinZero implementation. We implemented two algorithms: ProteinZero$_{\text{RAFT}}$, which selects the best-rewarded sequences for fine-tuning, and ProteinZero$_{\text{GRPO}}$, which directly optimizes policy using relative rewards, both running for 20 iterations. Both methods employed embedding-level diversity regularization ($\alpha_{\text{div}} = 0.05$) and KL constraints ($\alpha_{\text{KL}} = 0.1$).

Baselines. We compared against state-of-the-art inverse folding models (ProteinMPNN (Dauparas et al., 2022), ESM-IF (Hsu et al., 2022), InstructPLM (Qiu et al., 2024)). For RL-finetuniung algorithms, we compare with widely used offline RL baselines including DPO (Rafailov et al., 2023) and multi-round DPO (Xu et al., 2025).

4.2 Main Results

Overall Performance Analysis. Table 1 shows ProteinZero consistently outperforms existing methods. Both ProteinZero$_{\text{GRPO}}$ and ProteinZero$_{\text{RAFT}}$ surpass all baselines, with ProteinZero$_{\text{GRPO}}$ achieving best results across metrics (Figure 2). Our approach balances sequence recovery, structural

Table 1: Comparison of protein sequence design methods for 0-150 and 150-300 residue proteins. Success Rate is defined as scRMSD < 2Å and FoldX ddG < 0. Best scores are highlighted in blue, second-best in green. Designability metrics computed by ESMFold (independent AF3 evaluations confirm the same trend, see Table 2). All results are mean ± s.e. over 10 independent runs.

Length	**Method**	**InverseFold Acc.**	**Thermal Stability Metrics**		**Designability Metrics**				**Overall**
		Recovery Rate ↑	Fast-ddG ↓	FoldX ddG ↓	TM Score ↑	PLDDT ↑	Diversity ↑	scRMSD ↓ (scRMSD <2Å% ↑)	Success (%) ↑
0-150 residues	*Base Model*								
	InstructPLM	$0.574_{\pm0.009}$	$-21.543_{\pm1.330}$	$-20.878_{\pm1.445}$	$0.812_{\pm0.011}$	$79.983_{\pm0.614}$	$0.281_{\pm0.007}$	$1.484_{\pm0.044}$ ($85.71\%_{\pm0.002}$)	$84.45\%_{\pm0.0002}$
	SOTA Inverse Folding Models								
	ProteinMPNN	$0.426_{\pm0.006}$	$-21.509_{\pm1.230}$	$-20.792_{\pm1.207}$	$0.805_{\pm0.009}$	$79.883_{\pm0.502}$	$0.280_{\pm0.005}$	$1.500_{\pm0.037}$ ($82.14\%_{\pm0.002}$)	$81.95\%_{\pm0.0002}$
	ESM-IF	$0.377_{\pm0.006}$	$-17.900_{\pm1.235}$	$-14.328_{\pm1.269}$	$0.802_{\pm0.009}$	$78.918_{\pm0.534}$	$0.263_{\pm0.005}$	$1.515_{\pm0.038}$ ($81.25\%_{\pm0.002}$)	$80.71\%_{\pm0.0002}$
	RL Baseline Method								
	DPO	$0.571_{\pm0.008}$	$-21.713_{\pm1.260}$	$-21.191_{\pm1.332}$	$0.820_{\pm0.010}$	$80.716_{\pm0.571}$	$0.274_{\pm0.005}$	$1.473_{\pm0.041}$ ($87.58\%_{\pm0.002}$)	$86.44\%_{\pm0.0002}$
	Multi-Round DPO	$0.569_{\pm0.008}$	$-21.797_{\pm1.312}$	$-21.423_{\pm1.398}$	$0.823_{\pm0.011}$	$80.797_{\pm0.585}$	$0.266_{\pm0.005}$	$1.468_{\pm0.043}$ ($87.95\%_{\pm0.002}$)	$86.89\%_{\pm0.0003}$
	Our Online RL Methods								
	ProteinZero$_{\mathrm{RAFT}}$ (Ours)	$0.587_{\pm0.008}$	$-22.236_{\pm1.272}$	$-23.168_{\pm1.356}$	$0.849_{\pm0.011}$	$81.560_{\pm0.613}$	$0.296_{\pm0.007}$	$1.393_{\pm0.044}$ ($92.86\%_{\pm0.003}$)	$89.29\%_{\pm0.0002}$
	ProteinZero$_{\mathrm{GRPO}}$ (Ours)	$0.590_{\pm0.008}$	$-22.616_{\pm1.327}$	$-24.924_{\pm1.382}$	$0.867_{\pm0.011}$	$82.326_{\pm0.612}$	$0.306_{\pm0.007}$	$1.373_{\pm0.044}$ ($93.55\%_{\pm0.003}$)	$90.13\%_{\pm0.0002}$
150-300 residues	*Base Model*								
	InstructPLM	$0.570_{\pm0.009}$	$-36.362_{\pm2.451}$	$-27.145_{\pm1.797}$	$0.824_{\pm0.014}$	$83.783_{\pm0.568}$	$0.305_{\pm0.008}$	$1.448_{\pm0.048}$ ($88.24\%_{\pm0.002}$)	$86.38\%_{\pm0.0002}$
	SOTA Inverse Folding Models								
	ProteinMPNN	$0.405_{\pm0.007}$	$-35.778_{\pm2.280}$	$-27.057_{\pm1.581}$	$0.816_{\pm0.012}$	$82.361_{\pm0.548}$	$0.297_{\pm0.006}$	$1.469_{\pm0.040}$ ($86.64\%_{\pm0.002}$)	$84.67\%_{\pm0.0002}$
	ESM-IF	$0.446_{\pm0.008}$	$-32.125_{\pm2.207}$	$-24.816_{\pm1.548}$	$0.802_{\pm0.013}$	$82.042_{\pm0.536}$	$0.279_{\pm0.006}$	$1.487_{\pm0.042}$ ($86.09\%_{\pm0.002}$)	$82.81\%_{\pm0.0002}$
	RL Baseline Method								
	DPO	$0.570_{\pm0.009}$	$-36.417_{\pm2.325}$	$-28.915_{\pm1.571}$	$0.830_{\pm0.013}$	$83.837_{\pm0.506}$	$0.296_{\pm0.008}$	$1.441_{\pm0.042}$ ($88.97\%_{\pm0.002}$)	$87.70\%_{\pm0.0002}$
	Multi-Round DPO	$0.569_{\pm0.009}$	$-36.483_{\pm2.402}$	$-29.087_{\pm1.612}$	$0.831_{\pm0.014}$	$83.840_{\pm0.519}$	$0.288_{\pm0.008}$	$1.437_{\pm0.044}$ ($89.04\%_{\pm0.003}$)	$88.05\%_{\pm0.0002}$
	Our Online RL Methods								
	ProteinZero$_{\mathrm{RAFT}}$ (Ours)	$0.578_{\pm0.009}$	$-37.575_{\pm2.391}$	$-30.755_{\pm1.661}$	$0.841_{\pm0.013}$	$83.850_{\pm0.542}$	$0.324_{\pm0.008}$	$1.427_{\pm0.046}$ ($89.17\%_{\pm0.002}$)	$89.36\%_{\pm0.0002}$
	ProteinZero$_{\mathrm{GRPO}}$ (Ours)	$0.580_{\pm0.009}$	$-40.626_{\pm2.422}$	$-32.805_{\pm1.694}$	$0.862_{\pm0.013}$	$84.154_{\pm0.539}$	$0.331_{\pm0.009}$	$1.393_{\pm0.045}$ ($90.43\%_{\pm0.002}$)	$91.19\%_{\pm0.0002}$

accuracy, and stability while learning from self-generated outputs without additional labels. Importantly, although we only use TM-score (ESMFold/US-Align) and self-derived $\Delta\Delta G$ as rewards (Section 3.1.2), our evaluation uses orthogonal metrics, FoldX ddG for stability, pLDDT/scRMSD for structure, recovery/diversity for sequences, demonstrating genuine gains beyond reward hacking. Independent AlphaFold3 evaluation also confirms these improvements are generalizable (see Tables 2 and 8). For example, ProteinZero$_{\mathrm{GRPO}}$ achieves success rates 90.13% and 91.19% for 0-150 and 150-300 residues, respectively, reducing failure rates by 45% (from 18.05% to 9.87%) compared to ProteinMPNN for small proteins. Notably, compared to InstructPLM, we simultaneously improve recovery (0.574 → 0.590) and diversity (0.281 → 0.306), two traditionally conflicting objectives, demonstrating its ability to balance sequence conservation with exploration.

Comparison with DPO-based fine-tuning. We next compare ProteinZero with widely used DPO variants to illustrate the advantages of online RL. Regular DPO improves InstructPLM's success rate modestly (84.45% →86.89% for 0-150 residues), while Multi-Round DPO further raises it slightly to 86.89%. However, both variants reduce sequence diversity below the baseline: DPO lowers it from 0.281 to 0.274 and Multi-Round DPO further to 0.266. In contrast, ProteinZero$_{\mathrm{GRPO}}$ reaches 90.13% success and enhances diversity to 0.306. This divergence reflects a broader trend: offline methods progressively converge toward narrower solution spaces, limiting exploration of novel sequences. Online RL with diversity regularization maintains an exploration-exploitation balance, yielding not only higher diversity but also better structural generalization, as seen in improved scRMSD (1.373Å vs. 1.473Å for DPO). Similar patterns hold for larger proteins, where Multi-Round DPO increases success rate only modestly (86.38% → 88.05%) but still reduces diversity to 0.288, whereas ProteinZero achieves both higher success (91.19%) and greater diversity (0.331).

Comparison with SOTA Inverse Folding Models. We further compare ProteinZero against state-of-the-art inverse folding models. Starting from InstructPLM (Qiu et al., 2024) as our base model, ProteinZero$_{\mathrm{GRPO}}$ improves TM-score (0.812 → 0.867), stability (FoldX ddG: –20.878 → –24.924 kcal/mol), diversity (0.281 → 0.306), and success rate (84.45% → 90.13%) for short proteins. Similar gains are observed for longer proteins, where success rate increases from 86.38% to 91.19% and stability improves by 21% (–27.145 → –32.805 kcal/mol). Compared with other leading inverse folding models, ProteinZero achieves consistently higher success rates, outperforming ProteinMPNN (Dauparas et al., 2022) (81.95%) and ESM-IF (Hsu et al., 2022) (80.71%) across both size ranges. Qualitative visualizations (Figure 5) further support these findings, highlighting ProteinZero's ability to generate stable designs with high structural fidelity.

Effectiveness of fast-ddg reward. ProteinZero$_{\mathrm{GRPO}}$ achieves substantial gains in thermo-stability compared to InstructPLM, improving FoldX ddG by 19% (from −20.878 to −24.924 kcal/mol) for 0-150 residues and 21% (from −27.145 to −32.805 kcal/mol) for 150-300 residues. Unlike single-

Table 2: Independent validation of 150-300 residue proteins using AlphaFold3 versus ESMFold. Best scores are highlighted in blue , second-best in green .

Method	TM Score ↑		PLDDT ↑		scRMSD ↓		scRMSD <2Å (%) ↑		Success Rate (%) ↑	
	ESMFold	AF3	ESMFold	AF3	ESMFold	AF3	ESMFold	AF3	ESMFold	AF3
Base Model										
InstructPLM	0.8241	0.8418	83.78	85.86	1.4476	1.4018	88.24	90.12	86.38	88.41
Offline RL Baselines										
DPO	0.8296	0.8454	83.84	85.92	1.4407	1.3978	88.97	90.58	87.70	89.31
Multi-Round DPO	0.8313	0.8467	83.84	85.94	1.4372	1.3953	89.04	90.67	88.05	89.56
Our Online RL Methods										
ProteinZero$_{\text{RAFT}}$	0.8413	0.8548	83.85	86.03	1.4271	1.3891	89.17	90.72	89.36	90.64
ProteinZero$_{\text{GRPO}}$	0.8617	0.8718	84.15	86.19	1.3925	1.3598	90.43	91.76	91.19	92.27

objective methods that trade stability for other properties, ProteinZero simultaneously improves TM-score (0.812 to 0.867), diversity (0.281 to 0.306), recovery (0.574 to 0.590), and success rate (84.45% to 90.13%) for small proteins, with similar improvements for larger ones (see Table 1; extended metrics with wild-type and generated absolute energies in Table 12, Appendix C.5).

4.3 Case Study on Diverse Protein Folds and Complex Protein Design Tasks

Stabilization of Natural Proteins for Therapeutic Value: Our visual comparison in Figure 5 shows ProteinZero converts naturally unstable proteins into stable designs while maintaining structural fidelity. For challenging targets like membrane proteins and β-barrels, for example, our method achieves substantial stability improvements. The β-barrel structure (4FD5 chain A) transforms from unstable wild-type (FoldX ddG: 25.75 kcal/mol) to a stable design (-34.18 kcal/mol), while the membrane protein (2W7T chain A) improves from 42.01 to -36.09 kcal/mol. These results show ProteinZero's optimization of sequence-structure relationships, generating stability profiles valuable for therapeutic and industrial applications. By consistently producing designs with high structural accuracy and thermodynamic stability across α-helical, β-sheet, and mixed α/β folds, our approach expands the design space. While these computational evaluation metrics are promising, experimental validation remains essential to confirm functional properties.

Performance Scaling Across Protein Complexity: When compared with InstructPLM (our base model), ProteinZero demonstrates consistent improvements across diverse protein architectures. For challenging β-rich structures, our approach achieves higher structural accuracy (TM-score: 0.949 vs 0.910 for 1XXM chain C) and improved stability (FoldX ddG: -28.94 vs -8.94 kcal/mol). These gains extend across β-sheets, α/β mixed domains, and α-helical structures, as shown in Figure 5. ProteinZero delivers substantial improvements for both protein size categories: for 0-150 residues, success rate increases from 84.45% to 90.13%, stability improves from -20.878 to -24.924 kcal/mol, and diversity rises from 0.281 to 0.306. For 150-300 residues, we observe comparable gains: success rate from 86.38% to 91.19%, stability from -27.145 to -32.805 kcal/mol, and diversity from 0.305 to 0.331. The maintained performance improvements for larger proteins suggest our reinforcement learning framework handles increased structural complexity effectively.

4.4 Exploring the Design Space of Online RL for Fine-tuning Protein Generative Models

Reward Model Designs: Our ablation studies demonstrate that combining TM-score and stability rewards yields the highest overall success rates, consistently outperforming single-objective settings. For proteins of 0-150 residues, the combined reward achieves 90.13% success, compared to 89.52% with TM-score only and 85.15% with stability only. For larger proteins (150-300 residues), success rates are 91.19% for the combined setting, versus 89.76% and 87.38%, respectively. Examing the individual objectives explains this gap: optimizing only TM-score achieves the best structural accuracy (TM: 0.874 vs. 0.867 for the combined setting, 0-150 residues) but reduces stability, while optimizing only stability improves FoldX ddG (-25.381 vs.-24.924 kcal/mol) but compromises structural accuracy (TM: 0.831 vs. 0.867). By contrast, the combined reward balances both criteria, closing the trade-off and substantially reducing design failures.

Table 3: Ablation studies for 150-300 residue proteins across three design dimensions: reward models, learning objectives, and diversity regularization strategies. Best results highlighted in blue.

Design Configuration	InverseFold Acc.	Thermal Stability Metrics		Designability Metrics				Overall
	Recovery Rate ↑	Fast-ddG ↓	FoldX ddG ↓	TM Score ↑	PLDDT ↑	Diversity ↑	scRMSD ↓ (scRMSD <2Å% ↑)	Success (%) ↑
Design Dimension 1: Reward Model Formulation								
Only TM-score as Reward	0.577	-35.793	-25.905	0.870	84.237	0.333	1.384 (91.25%)	89.76%
Only ddG as Reward	0.574	-42.769	-35.927	0.831	83.540	0.327	1.447 (88.52%)	87.38%
Full ProteinZero (TM+ddG)	0.580	-40.626	-32.805	0.862	84.154	0.331	1.393 (90.43%)	91.19%
Design Dimension 2: Learning Objective Components								
Without Diversity Term	0.580	-37.905	-31.185	0.860	84.065	0.281	1.401 (89.71%)	91.32%
Without KL Term	0.569	-40.688	-33.193	0.835	83.008	0.328	1.440 (89.08%)	87.92%
Full ProteinZero (All Terms)	0.580	-40.626	-32.805	0.862	84.154	0.331	1.393 (90.43%)	91.19%
Design Dimension 3: Diversity Regularization Strategies								
Diversity as Reward	0.558	-33.904	-23.967	0.849	83.326	0.315	1.421 (89.32%)	81.71%
Hamming Distance as Reward	0.568	-32.128	-23.228	0.836	83.668	0.294	1.432 (89.14%)	80.29%
Full ProteinZero (Embedding Diversity)	0.580	-40.626	-32.805	0.862	84.154	0.331	1.393 (90.43%)	91.19%

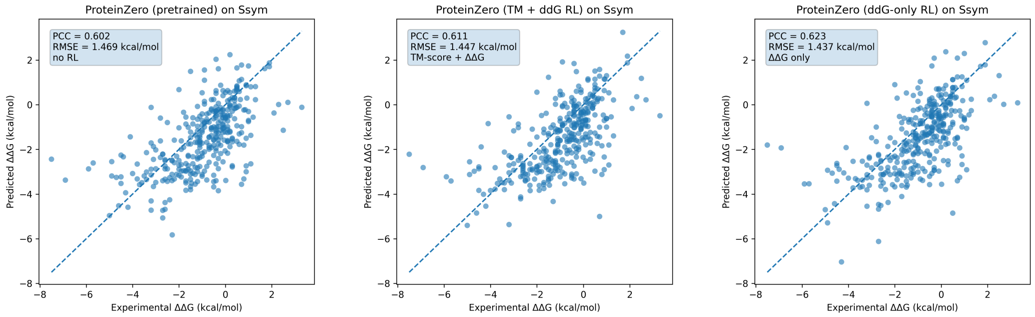
Learning Objective Components: We ablate the diversity regularization and KL divergence to assess their contributions (Tables 5 and 3). Removing the diversity regularization marginally improves success rate (90.23% vs. 90.13% for 0-150 residues, 91.32% vs. 91.19% for 150-300 residues), but significantly reduces sequence diversity from 0.306 to 0.268 for 0-150 residues and from 0.331 to 0.281 for 150-300 residues. This 12-15% reduction in diversity indicates convergence to a narrower solution space, limiting its ability to explore functionally diverse sequences, a key concern with offline RL methods. By contrast, removing KL divergence causes severe degradation: success rate drops by nearly 4%, TM-score declines by around 0.03, and pLDDT decreases by around 1.3, reflecting both reduced structural accuracy and confidence. These results show KL regularization is essential for stable optimization and preventing catastrophic forgetting, while diversity regularization, though slightly reducing peak performance, preserves exploration crucial for discovering novel protein designs beyond the training distribution.

Diversity Regularization Strategies: We compare three strategies for incorporating diversity into the optimization process (Tables 5 and 3; detailed results in Appendix Table 13): embedding-based diversity as a reward, Hamming distance as a reward, and embedding-based diversity as a regularization term in the loss. Introducing diversity directly into the reward sharply reduces performance, with success rates falling to 78.65% and 81.71%, and stability values deteriorating relative to the baseline. Using Hamming distance performs even worse, lowering success rates to 74.63% and 80.29% and further degrading stability and structural accuracy. By contrast, applying embedding-based diversity as a regularizer maintains success rates of 90.13% and 91.19%, preserves sequence diversity at 0.306 and 0.331, and avoids losses in stability or accuracy. These results indicate that reward-based diversity introduces conflicting signals that destabilize training, whereas regularization provides consistent gradients that encourage exploration while safeguarding functional objectives.

The ablation studies validate our design choices and highlight the importance of balancing multiple objectives in online RL for protein design. ProteinZero navigates these trade-offs through separated optimization signals, multi-objective rewards for primary objectives, and regularization for exploration and stability, yielding a robust approach generalizing across protein sizes and architectures.

4.5 Fast-ddG Accuracy for Predicting Mutational $\Delta\Delta G$ Using Wet-Lab Validated Data

We assess Fast-ddG correlation with experimental measurements on the Ssym benchmark (Pucci et al., 2018), comprising 684 single-point mutations with calorimetrically measured $\Delta\Delta G$ values. Consistent with Eq. 4, we evaluate 342 wild-type→mutant transitions by computing stability changes on wild-type backbones. Table 10 (Appendix C.4) compares our predictor against physics-based oracles (FoldX, Rosetta) and supervised predictors (ThermoMPNN (Dieckhaus et al., 2024), ThermoNet (Li et al., 2020), PROSTATA (Umerenkov et al., 2022)). Across three configurations, pretrained, Fast-ddG-only, and TM-score + Fast-ddG, our predictor achieves RMSE 1.44–1.47 kcal/mol and PCC 0.60–0.62, matching FoldX (RMSE: 1.56, PCC: 0.63) while operating 236–760× faster (Tables 6–7). This represents 56% RMSE reduction versus ProteinMPNN (3.38 kcal/mol, PCC: 0.26), demonstrating gains from specialized optimization. While ThermoMPNN achieves superior performance (1.12, 0.72), it requires supervised training and handles only single-residue perturbations, whereas our unsupervised, self-derived predictor generalizes to



(a) Pretrained (b) TM+ddG (c) ddG-only

Figure 3: Fast-ddG predictor performance on the Ssym dataset with 342 wet-lab validated single-point mutations (wild-type → mutant). Each subfigure shows predicted versus experimental $\Delta\Delta G$ values for different model variants: (a) pretrained model before RL fine-tuning, (b) fine-tuned with joint TM-score + Fast-ddG rewards, (c) fine-tuned with Fast-ddG reward only. All variants achieve comparable correlation with experimental measurements (PCC ≈ 0.60–0.62, RMSE ≈ 1.44–1.47 kcal/mol).

Table 4: Per-protein performance of the Fast-ddG surrogate on Ssym dataset (342 single-point direct mutations, wild-type→mutant direction). We report the number of mutations per PDB (n_{mut}), per-protein RMSE (kcal/mol), and PCC for the pretrained and fine-tuned Fast-ddG variants. Best results highlighted in blue.

PDB	n_{mut}	**RMSE (kcal/mol) ↓**			**PCC ↑**		
		Pretrained	Fast-ddG only	TM-score + Fast-ddG	Pretrained	Fast-ddG only	TM-score + Fast-ddG
1L63	118	1.36	1.30	1.33	0.58	0.66	0.60
2LZM	66	1.16	1.13	1.13	0.74	0.75	0.75
1LZ1	61	1.22	1.12	1.18	0.76	0.78	0.79
1BNI	13	1.09	1.03	1.25	0.62	0.70	0.59

full sequence redesigns. Per-protein analysis (Table 4) confirms improvements generalize across diverse targets rather than reflecting outlier bias. For the four largest proteins (1L63, 2LZM, 1LZ1, 1BNI; 72.5% of mutations), fine-tuning consistently reduces RMSE, with Fast-ddG-only achieving best correlation on three of four. On 1L63 (118 mutations), RMSE improves from 1.36 to 1.30 kcal/mol and PCC from 0.58 to 0.66. Representative cases with substantial error reductions appear in Table 11 (Appendix C.4). Figure 3 shows linear correspondence (PCC ≈ 0.60–0.62) with reduced scatter post-tuning. These results establish that Fast-ddG, though derived from unsupervised likelihood ratios and optimized for full-sequence inverse folding, achieves accuracy comparable to physics-based benchmarks while maintaining computational efficiency for online RL.

5 Conclusion

We presented ProteinZero, an online reinforcement learning framework that enables protein generative models to improve beyond supervised pretraining by learning from their own outputs. It integrates two methodological advances: a fast, unsupervised ddG predictor for efficient stability signals and an embedding-level diversity regularizer that prevents collapse while encouraging meaningful variation. These components make online RL tractable for protein design and offer insights for broader RLHF by addressing efficiency and diversity collapse. Experiments show consistent multi-objective gains across structural accuracy, stability, recovery, and diversity, including on challenging folds such as β-barrels and membrane proteins. While evaluation relies on in-silico metrics and requires wet-lab validation, the results demonstrate that efficient online RL can complement supervised methods through scalable feedback, expanding the accessible design space and supporting applications in therapeutics, enzymes, and synthetic biology.

Ethics Statement

Our work on ProteinZero focuses on computational methods for protein design optimization. All experiments were conducted using publicly available datasets (CATH 4.3) and computational simulations without any wet-lab experimentation or use of biological materials. We acknowledge that while our method demonstrates improvements in computational metrics, these results require experimental validation before any practical application. The potential applications of improved protein design methods span therapeutic development, industrial biotechnology, and basic research. We emphasize that any deployment of designed proteins must follow established safety protocols, regulatory frameworks, and ethical guidelines for biological research. The computational nature of our work poses minimal direct ethical concerns, but we recognize the importance of responsible development and deployment of AI systems in biological design. We commit to making our code publicly available to ensure transparency and enable the research community to build upon and scrutinize our work.

Reproducibility Statement

To ensure our results are fully reproducible, we provide comprehensive details of our methodology and experimental setup. Our framework is described in Section 3, which details the core online RL objective, our novel embedding-level diversity regularizer (Section 3.1.1), the time-efficient reward models (Section 3.1.2), and the specific algorithms, ProteinZero$_{\text{RAFT}}$ (Section 3.2.1) and ProteinZero$_{\text{GRPO}}$ (Section 3.2.2).

Our experimental setup, including the use of the public CATH 4.3 dataset, specific train-test splits, and evaluation metrics, is detailed in Section 4 and Appendix B.3. Full implementation details, including all hyperparameters, software dependencies, baseline methods, and computational resource requirements are provided in Appendix B. Our work builds upon the publicly available InstructPLM model, and all evaluation tools (ESMFold, US-align, AlphaFold3, FoldX, and Rosetta) are open-source. As a demonstration, we provide example protein sequences generated by ProteinZero$_{\text{GRPO}}$ as .pdb files in the supplementary material, following the naming convention: ProteinZero_GRPO_[TargetPDB]_[Chain]_designed.pdb (e.g., ProteinZero_GRPO_2hls_A_designed.pdb), where TargetPDB is the original PDB identifier and Chain specifies the protein chain used as the structural template. Upon publication, we will release our complete source code, pre-trained model checkpoints, evaluation scripts, and detailed documentation to facilitate replication of our findings.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. Self-consuming generative models go MAD. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL `https://openreview.net/forum?id=ShjMHfmPs0`.

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. URL `https://openreview.net/forum?id=HklxbgBKvr`.

Dávid Bajusz, Warren S Wade, Grzegorz Satała, Andrzej J Bojarski, Janez Ilaš, Jessica Ebner, Florian Grebien, Henrietta Papp, Ferenc Jakab, Alice Douangamath, et al. Exploring protein hotspots by optimized fragment pharmacophores. *Nature Communications*, 12(1):3201, 2021.

Nathaniel R Bennett, Brian Coventry, Inna Goreshnik, Buwei Huang, Aza Allen, Dionne Vafeados, Ying Po Peng, Justas Dauparas, Minkyung Baek, Lance Stewart, et al. Improving de novo protein binder design with deep learning. *Nature Communications*, 14(1):2625, 2023.

Lasse M Blaabjerg, Nicolas Jonsson, Wouter Boomsma, Amelie Stein, and Kresten Lindorff-Larsen. Ssemb: A joint embedding of protein sequence and structure enables robust variant effect predictions. *Nature Communications*, 15(1):9646, 2024.

Andrew A Bogan and Kurt S Thorn. Anatomy of hot spots in protein interfaces. *Journal of molecular biology*, 280(1):1–9, 1998.

Aron Broom, Zachary Jacobi, Kyle Trainor, and Elizabeth M Meiering. Computational tools help improve protein stability but with a solubility tradeoff. *Journal of Biological Chemistry*, 292(35): 14349–14361, 2017.

Oliver Buß, Jens Rudat, and Katrin Ochsenreither. Foldx as protein engineering tool: better than random based approaches? *Computational and structural biotechnology journal*, 16:25–33, 2018.

Matteo Cagiada, Sergey Ovchinnikov, and Kresten Lindorff-Larsen. Predicting absolute protein folding stability using generative models. *Protein Science*, 34(1):e5233, 2025.

Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. *arXiv preprint arXiv:2402.04997*, 2024.

Longxing Cao, Inna Goreshnik, Brian Coventry, James Brett Case, Lauren Miller, Lisa Kozodoy, Rita E Chen, Lauren Carter, Alexandra C Walls, Young-Jun Park, et al. De novo design of picomolar sars-cov-2 miniprotein inhibitors. *Science*, 370(6515):426–431, 2020.

Alexander E Chu, Jinho Kim, Lucy Cheng, Gina El Nesr, Minkai Xu, Richard W Shuai, and Po-Ssu Huang. An all-atom protein generative model. *Proceedings of the National Academy of Sciences*, 121(27):e2311500121, 2024.

Tim Clackson and James A Wells. A hot spot of binding energy in a hormone-receptor interface. *Science*, 267(5196):383–386, 1995.

Gabriele Corso, Zhitao Ying, Michal Pándy, Petar Veličković, Jure Leskovec, and Pietro Liò. Neural distance embeddings for biological sequences. *Advances in Neural Information Processing Systems*, 34:18539–18551, 2021.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

Kieran Didi, Prashant Sohani, Fabian Berressem, Alexander Nesterovskiy, Boris Fomitchev, Robert Ohannessian, Mohamed Elbalkini, Jonathan Cogan, Anthony Costa, Arash Vahdat, et al. Highly efficient protein structure prediction on nvidia rtx blackwell and grace-hopper.

Henry Dieckhaus, Michael Brocidiacono, Nicholas Z Randolph, and Brian Kuhlman. Transfer learning to leverage larger datasets for improved prediction of protein stability changes. *Proceedings of the national academy of sciences*, 121(6):e2314853121, 2024.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*, 2023.

Jiajun Fan, Shuaike Shen, Chaoran Cheng, Yuxin Chen, Chumeng Liang, and Ge Liu. Online reward-weighted fine-tuning of flow matching with wasserstein regularization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=2IoFFexvuw.

Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. *Advances in neural information processing systems*, 34:20132–20145, 2021.

Zhangyang Gao, Cheng Tan, and Stan Z. Li. Pifold: Toward effective and efficient protein inverse folding. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=oMsN9TYwJ0j.

Hans-Christof Gasser, Diego A Oyarzún, Javier Alfaro, and Ajitha Rajan. Integrating mhc class i visibility targets into the proteinmpnn protein design process. *bioRxiv*, pp. 2024–06, 2024.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. In *International conference on machine learning*, pp. 2160–2169. PMLR, 2019.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. Pmlr, 2018.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8946–8970. PMLR, 2022. URL https://proceedings.mlr.press/v162/hsu22a.html.

Guillaume Huguet, James Vuckovic, Kilian Fatras, Eric Thibodeau-Laufer, Pablo Lemos, Riashat Islam, Cheng-Hao Liu, Jarrid Rector-Brooks, Tara Akhound-Sadegh, Michael M. Bronstein, Alexander Tong, and Avishek Joey Bose. Sequence-augmented se(3)-flow matching for conditional protein backbone generation. *CoRR*, abs/2405.20313, 2024. doi: 10.48550/ARXIV.2405.20313. URL https://doi.org/10.48550/arXiv.2405.20313.

John Ingraham, Vikas Garg, Regina Barzilay, and Tommi Jaakkola. Generative models for graph-based protein design. *Advances in neural information processing systems*, 32, 2019.

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.

Moksh Jain, Emmanuel Bengio, Alex Hernandez-Garcia, Jarrid Rector-Brooks, Bonaventure FP Dossou, Chanakya Ajit Ekbote, Jie Fu, Tianyu Zhang, Michael Kilgour, Dinghuai Zhang, et al. Biological sequence design with gflownets. In *International Conference on Machine Learning*, pp. 9786–9801. PMLR, 2022.

Xiaoran Jiao, Weian Mao, Wengong Jin, Peiyuan Yang, Hao Chen, and Chunhua Shen. Boltzmann-aligned inverse folding model as a predictor of mutational effects on protein-protein interactions. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=lzdFImKK8w`.

Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael John Lamarre Townshend, and Ron O. Dror. Learning from protein structure with geometric vector perceptrons. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL `https://openreview.net/forum?id=1YLJDvSx6J4`.

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.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of RLHF on LLM generalisation and diversity. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=PXD3FAVHJT`.

Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018.

Bian Li, Yucheng T Yang, John A Capra, and Mark B Gerstein. Predicting changes in protein thermodynamic stability upon point mutation with deep 3d convolutional neural networks. *PLoS computational biology*, 16(11):e1008291, 2020.

Tianjian Li and Daniel Khashabi. SIMPLEMIX: Frustratingly simple mixing of off- and on-policy data in language model preference learning. In *Forty-second International Conference on Machine Learning*, 2025. URL `https://openreview.net/forum?id=ucU1o3PNB0`.

Tianjian Li, Yiming Zhang, Ping Yu, Swarnadeep Saha, Daniel Khashabi, Jason Weston, Jack Lanchantin, and Tianlu Wang. Jointly reinforcing diversity and quality in language model generations. *arXiv preprint arXiv:2509.02534*, 2025.

Sidney Lyayuga Lisanza, Jake Merle Gershon, Sam Tipps, Lucas Arnoldt, Samuel Hendel, Jeremiah Nelson Sims, Xinting Li, and David Baker. Joint generation of protein sequence and structure with rosettafold sequence space diffusion. *bioRxiv*, pp. 2023–05, 2023.

Isaac D Lutz, Shunzhi Wang, Christoffer Norn, Alexis Courbet, Andrew J Borst, Yan Ting Zhao, Annie Dosey, Longxing Cao, Jinwei Xu, Elizabeth M Leaf, et al. Top-down design of protein architectures with reinforcement learning. *Science*, 380(6642):266–273, 2023.

Christine A Orengo, Alex D Michie, Susan Jones, David T Jones, Mark B Swindells, and Janet M Thornton. Cath–a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1109, 1997.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Ryan Park, Darren J. Hsu, C. Brian Roland, Maria Korshunova, Chen Tessler, Shie Mannor, Olivia Viessmann, and Bruno Trentini. Improving inverse folding for peptide design with diversity-regularized direct preference optimization. *CoRR*, abs/2410.19471, 2024. doi: 10.48550/ARXIV.2410.19471. URL `https://doi.org/10.48550/arXiv.2410.19471`.

Fabrizio Pucci, Katrien V Bernaerts, Jean Marc Kwasigroch, and Marianne Rooman. Quantification of biases in predictions of protein stability changes upon mutations. *Bioinformatics*, 34(21): 3659–3665, 2018.

Jiezhong Qiu, Junde Xu, Jie Hu, Hanqun Cao, Liya Hou, Zijun Gao, Xinyi Zhou, Anni Li, Xiujuan Li, Bin Cui, Fei Yang, Shuang Peng, Ning Sun, Fangyu Wang, Aimin Pan, Jie Tang, Jieping Ye, Junyang Lin, Jin Tang, Xingxu Huang, Pheng Ann Heng, and Guangyong Chen. Instructplm: Aligning protein language models to follow protein structure instructions. *bioRxiv*, 2024. doi: 10.1101/2024.04.17.589642. URL https://www.biorxiv.org/content/early/2024/04/20/2024.04.17.589642.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Stephen A Rettie, Katelyn V Campbell, Asim K Bera, Alex Kang, Simon Kozlov, Yensi Flores Bueso, Joshmyn De La Cruz, Maggie Ahlrichs, Suna Cheng, Stacey R Gerben, et al. Cyclic peptide structure prediction and design using alphafold2. *Nature Communications*, 16(1):4730, 2025.

Gabriel J Rocklin, Tamuka M Chidyausiku, Inna Goreshnik, Alex Ford, Scott Houliston, Alexander Lemak, Lauren Carter, Rashmi Ravichandran, Vikram K Mulligan, Aaron Chevalier, et al. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science*, 357 (6347):168–175, 2017.

Frederic Runge, Danny Stoll, Stefan Falkner, and Frank Hutter. Learning to design RNA. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ByfyHh05tQ.

Robert Schmirler, Michael Heinzinger, and Burkhard Rost. Fine-tuning protein language models boosts predictions across diverse tasks. *Nature Communications*, 15(1):7407, 2024.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL http://arxiv.org/abs/1707.06347.

Joost Schymkowitz, Jesper Borg, Francois Stricher, Robby Nys, Frederic Rousseau, and Luis Serrano. The foldx web server: an online force field. *Nucleic acids research*, 33(suppl_2):W382–W388, 2005.

Varun R. Shanker, Theodora U. J. Bruun, Brian L. Hie, and Peter S. Kim. Unsupervised evolution of protein and antibody complexes with a structure-informed language model. *Science*, 385(6704): 46–53, 2024. doi: 10.1126/science.adk8946. URL https://www.science.org/doi/abs/10.1126/science.adk8946.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024. doi: 10.48550/ARXIV.2402.03300. URL https://doi.org/10.48550/arXiv.2402.03300.

Shivanshu Shekhar, Shreyas Singh, and Tong Zhang. See-dpo: Self entropy enhanced direct preference optimization. *arXiv preprint arXiv:2411.04712*, 2024.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

Elana Simon and James Zou. Interplm: Discovering interpretable features in protein language models via sparse autoencoders. *bioRxiv*, pp. 2024–11, 2024.

Marcin J Skwark, Nicolás López Carranza, Thomas Pierrot, Joe Phillips, Slim Said, Alexandre Laterre, Amine Kerkeni, Uğur Şahin, and Karim Beguir. Designing a prospective covid-19 therapeutic with reinforcement learning. *arXiv preprint arXiv:2012.01736*, 2020.

Emanuel Todorov. Linearly-solvable markov decision problems. *Advances in neural information processing systems*, 19, 2006.

Kotaro Tsuboyama, Justas Dauparas, Jonathan Chen, Elodie Laine, Yasser Mohseni Behbahani, Jonathan J Weinstein, Niall M Mangan, Sergey Ovchinnikov, and Gabriel J Rocklin. Mega-scale experimental analysis of protein folding stability in biology and design. *Nature*, 620(7973):434–444, 2023.

Dmitriy Umerenkov, Tatiana I Shashkova, Pavel V Strashnov, Fedor Nikolaev, Maria Sindeeva, Nikita V Ivanisenko, and Olga L Kardymon. Prostata: protein stability assessment using transformers. *BioRxiv*, pp. 2022–12, 2022.

Chaoqi Wang, Yibo Jiang, Chenghao Yang, Han Liu, and Yuxin Chen. Beyond reverse KL: Generalizing direct preference optimization with diverse divergence constraints. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=2cRzmWXK9N`.

Chenyu Wang, Masatoshi Uehara, Yichun He, Amy Wang, Avantika Lal, Tommi Jaakkola, Sergey Levine, Aviv Regev, Hanchen, and Tommaso Biancalani. Fine-tuning discrete diffusion models via reward optimization with applications to DNA and protein design. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=G328D1xt4W`.

Chuanrui Wang, Bozitao Zhong, Zuobai Zhang, Narendra Chaudhary, Sanchit Misra, and Jian Tang. Pdb-struct: A comprehensive benchmark for structure-based protein design. *CoRR*, abs/2312.00080, 2023a. doi: 10.48550/ARXIV.2312.00080. URL `https://doi.org/10.48550/arXiv.2312.00080`.

Yi Wang, Hui Tang, Lichao Huang, Lulu Pan, Lixiang Yang, Huanming Yang, Feng Mu, and Meng Yang. Self-play reinforcement learning guides protein engineering. *Nature Machine Intelligence*, 5(8):845–860, 2023b.

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.

Talal Widatalla, Rafael Rafailov, and Brian Hie. Aligning protein generative models with experimental fitness via direct preference optimization. *bioRxiv*, pp. 2024–05, 2024.

Hein J Wijma, Robert J Floor, Peter A Jekel, David Baker, Siewert J Marrink, and Dick B Janssen. Computationally designed libraries for rapid enzyme stabilization. *Protein Engineering, Design & Selection*, 27(2):49–58, 2014.

Junde Xu, Zijun Gao, Xinyi Zhou, Jie Hu, Xingyi Cheng, Le Song, Guangyong Chen, Pheng-Ann Heng, and Jiezhong Qiu. Protein inverse folding from structure feedback. *arXiv preprint arXiv:2506.03028*, 2025.

Fanglei Xue, Andrew Kubaney, Zhichun Guo, Joseph K Min, Ge Liu, Yi Yang, and David Baker. Improving protein sequence design through designability preference optimization. *arXiv preprint arXiv:2506.00297*, 2025.

Kai Yi, Bingxin Zhou, Yiqing Shen, Pietro Liò, and Yuguang Wang. Graph denoising diffusion for inverse protein folding. *Advances in Neural Information Processing Systems*, 36:10238–10257, 2023.

Chengxin Zhang, Morgan Shine, Anna Marie Pyle, and Yang Zhang. Us-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nature methods*, 19(9): 1109–1115, 2022.

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.

Yang Zhang and Jeffrey Skolnick. Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic acids research*, 33(7):2302–2309, 2005.

Zaixiang Zheng, Yifan Deng, Dongyu Xue, Yi Zhou, Fei Ye, and Quanquan Gu. Structure-informed language models are protein designers. In *International conference on machine learning*, pp. 42317–42338. PMLR, 2023.

Xiangxin Zhou, Dongyu Xue, Ruizhe Chen, Zaixiang Zheng, Liang Wang, and Quanquan Gu. Antigen-specific antibody design via direct energy-based preference optimization. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/daef77101ba5711084a57442c8cf2709-Abstract-Conference.html.

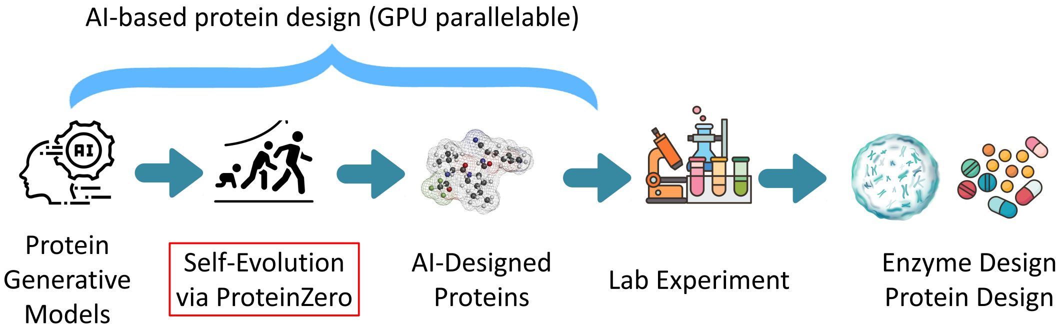

AI-based protein design (GPU parallelable)
Protein Generative Models
Self-Evolution via ProteinZero
AI-Designed Proteins
Lab Experiment
Enzyme Design
Protein Design

Figure 4: Integration of ProteinZero within the AI-driven protein design pipeline. Pre-trained generative models evolve through ProteinZero's online reinforcement learning framework to produce optimized protein sequences. These AI-designed candidates proceed to laboratory synthesis and experimental characterization, enabling applications in diverse biotechnological domains such as enzyme engineering and therapeutic development. The computational stages (blue) can leverage GPU parallelization for efficient large-scale processing.

A Discussion

A.1 Broader Impact

ProteinZero represents a methodological advancement in computational protein design by enabling autonomous improvement of generative models through online reinforcement learning. As illustrated in Figure 4, our framework integrates within the broader protein design pipeline, bridging computational optimization and experimental validation. By reducing reliance on manually curated datasets from repositories like the Protein Data Bank, which capture only a fraction of viable sequence space, our approach offers new possibilities for exploring protein designs beyond naturally occurring examples.

The computational efficiency gains (achieving comparable results with substantially reduced computational time compared to physics-based methods) and improved success rates demonstrated in our experiments could accelerate research in therapeutic development, enzyme engineering, and industrial biotechnology. The reduced computational requirements potentially improve accessibility of advanced protein design capabilities for research groups with limited resources. Applications span from developing novel biologics and vaccines to engineering enzymes for sustainable manufacturing and bioremediation.

However, we emphasize that our computational metrics, while encouraging, require experimental validation to confirm biological functionality. The stability and foldability improvements we demonstrate computationally may not directly translate to enhanced catalytic activity, binding affinity, or other functional properties critical for real-world applications. Furthermore, the path from computational design to practical application involves multiple validation stages. Each designed protein undergoes synthesis, experimental characterization, functional testing, and regulatory approval before deployment. This established multi-step process provides checkpoints for safety and efficacy verification. Our computational improvements represent the initial stage of this pipeline, with subsequent experimental validation remaining essential for confirming biological relevance.

The self-improving nature of ProteinZero, learning continuously from generated outputs rather than requiring new experimental data, represents a shift toward more autonomous systems in computational biology. While this offers exciting possibilities for accelerating discovery, the comprehensive experimental validation pipeline ensures that computational predictions are rigorously tested before practical application. We envision this work contributing to a new generation of AI systems that can explore biological design spaces more efficiently, ultimately advancing our understanding and engineering capabilities in protein science.

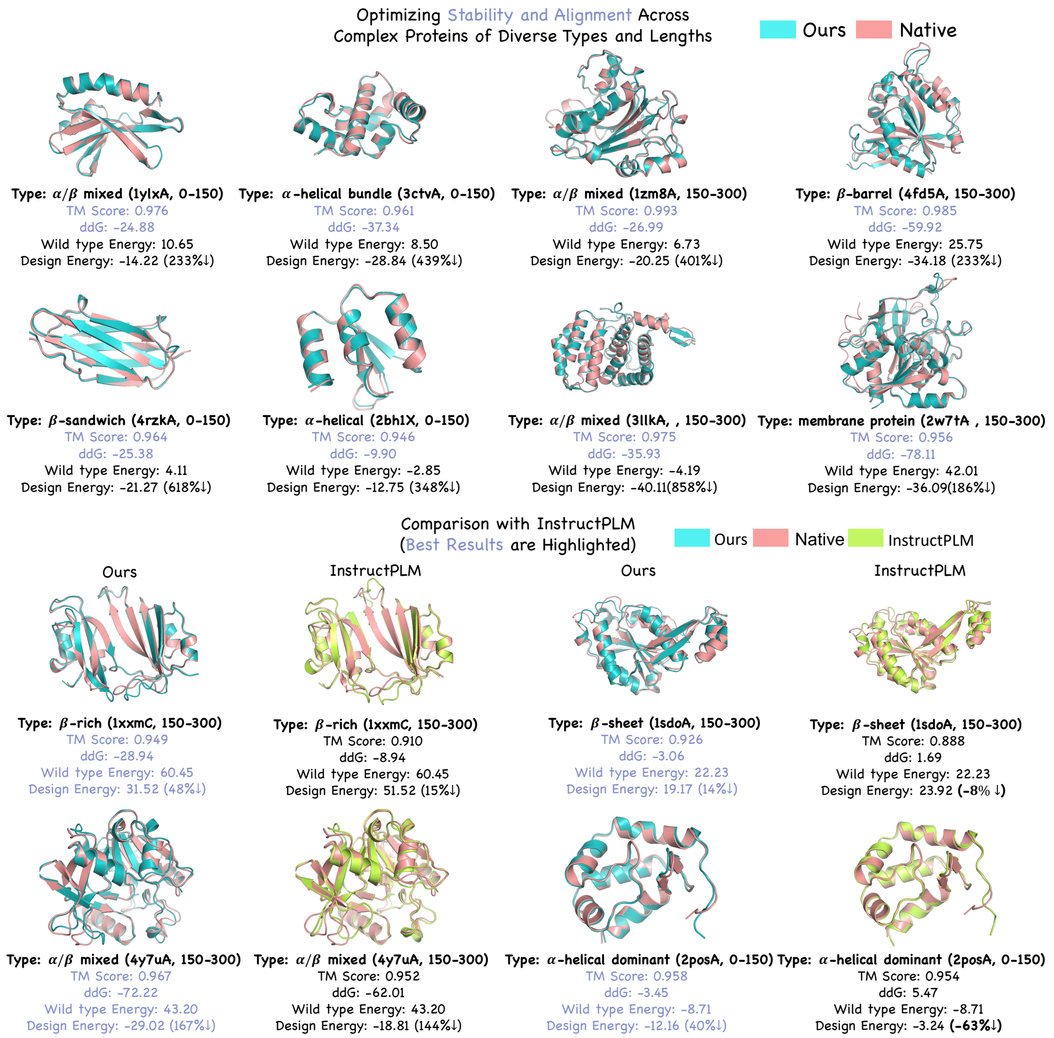

Optimizing Stability and Alignment Across
Complex Proteins of Diverse Types and Lengths
Ours
Native
Type: α/β mixed (1ylxA, 0-150)
TM Score: 0.976
ddG: -24.88
Wild type Energy: 10.65
Design Energy: -14.22 (233%↓)
Type: α-helical bundle (3ctvA, 0-150)
TM Score: 0.961
ddG: -37.34
Wild type Energy: 8.50
Design Energy: -28.84 (439%↓)
Type: α/β mixed (1zm8A, 150-300)
TM Score: 0.993
ddG: -26.99
Wild type Energy: 6.73
Design Energy: -20.25 (401%↓)
Type: β-barrel (4fd5A, 150-300)
TM Score: 0.985
ddG: -59.92
Wild type Energy: 25.75
Design Energy: -34.18 (233%↓)
Type: β-sandwich (4rzkA, 0-150)
TM Score: 0.964
ddG: -25.38
Wild type Energy: 4.11
Design Energy: -21.27 (618%↓)
Type: α-helical (2bh1X, 0-150)
TM Score: 0.946
ddG: -9.90
Wild type Energy: -2.85
Design Energy: -12.75 (348%↓)
Type: α/β mixed (3llkA, , 150-300)
TM Score: 0.975
ddG: -35.93
Wild type Energy: -4.19
Design Energy: -40.11(858%↓)
Type: membrane protein (2w7tA , 150-300)
TM Score: 0.956
ddG: -78.11
Wild type Energy: 42.01
Design Energy: -36.09(186%↓)
Comparison with InstructPLM
(Best Results are Highlighted)
Ours
Native
InstructPLM
Ours
InstructPLM
Ours
InstructPLM
Type: β-rich (1xxmC, 150-300)
TM Score: 0.949
ddG: -28.94
Wild type Energy: 60.45
Design Energy: 31.52 (48%↓)
Type: β-rich (1xxmC, 150-300)
TM Score: 0.910
ddG: -8.94
Wild type Energy: 60.45
Design Energy: 51.52 (15%↓)
Type: β-sheet (1sdoA, 150-300)
TM Score: 0.926
ddG: -3.06
Wild type Energy: 22.23
Design Energy: 19.17 (14%↓)
Type: β-sheet (1sdoA, 150-300)
TM Score: 0.888
ddG: 1.69
Wild type Energy: 22.23
Design Energy: 23.92 (-8% ↓)
Type: α/β mixed (4y7uA, 150-300)
TM Score: 0.967
ddG: -72.22
Wild type Energy: 43.20
Design Energy: -29.02 (167%↓)
Type: α/β mixed (4y7uA, 150-300)
TM Score: 0.952
ddG: -62.01
Wild type Energy: 43.20
Design Energy: -18.81 (144%↓)
Type: α-helical dominant (2posA, 0-150)
TM Score: 0.958
ddG: -3.45
Wild type Energy: -8.71
Design Energy: -12.16 (40%↓)
Type: α-helical dominant (2posA, 0-150)
TM Score: 0.954
ddG: 5.47
Wild type Energy: -8.71
Design Energy: -3.24 (-63%↓)

Figure 5: **Representative cases of protein structure designs from held-out test set.** Visual comparison between ProteinZero (cyan), native proteins (pink), and InstructPLM (lime green). Top panels show selected cases where naturally unstable proteins are redesigned by ProteinZero. In these examples, predicted stability improvements range from 233% to 858% (based on FoldX ddG calculations) while maintaining structural similarity (TM-scores > 0.95). Bottom panels present comparative examples with InstructPLM for challenging β-rich structures and complex architectures. In the shown cases, ProteinZero generates designs with negative predicted ddG values while InstructPLM produces positive values indicating predicted instability. These visualizations represent individual design outcomes; comprehensive quantitative results are provided in Table 1.

A.2 Limitations and Future Directions

Restriction to Monomeric Scaffolds. Our experiments target monomeric proteins, a practically significant class spanning critical therapeutic modalities: *de novo* miniprotein inhibitors (Cao et al., 2020), antigen-display architectures (Lutz et al., 2023), and cyclic peptide binders (Rettie et al., 2025). Leading generative methods including RFdiffusion (Watson et al., 2023), ProteinMPNN (Dauparas et al., 2022), and Chroma (Ingraham et al., 2023) have similarly demonstrated advances on single-chain scaffolds. However, many drug discovery applications require multimeric complexes and protein-protein interfaces. The core framework components—online RL optimization, embedding-level diversity regularization, and fast proxy rewards—are architecture-agnostic and naturally extend to assemblies by incorporating interface-aware structural rewards and multimer-capable stability predictors to optimize binding affinity and interface packing simultaneously.

Reliance on Computational Proxies. ProteinZero employs computational predictors (Fast-ddG, ESMFold TM-score) as reward signals. While these metrics act as proxies for biological properties rather than substitutes for experimental validation, computational screening remains standard in protein engineering pipelines. Empirical studies demonstrate that computational stability predictors enrich for mutations that experimentally increase protein thermodynamic stability: 30–40% of computationally predicted stabilizing mutations are confirmed stable in wet-lab validation, compared to near-zero success rates for random amino acid substitutions (Wijma et al., 2014; Broom et al., 2017; Buß et al., 2018). However, individual oracles encode systematic preferences—FoldX, for instance, favors mutations that increase hydrophobic core packing, which may trade off against solubility (Broom et al., 2017).

We address over-optimization to single-oracle patterns through two mechanisms. First, multi-objective optimization with diversity regularization enforces complementary constraints (structural designability, stability, KL-regularization, embedding diversity), preventing the policy from satisfying one objective at the expense of others. Second, independent evaluation rigorously separates training rewards (Fast-ddG, ESMFold) from evaluation metrics (FoldX, AlphaFold3). Transferability to these independent oracles (Section 4) indicates the model learns generalizable biophysical principles rather than oracle-specific patterns.

B Experimental Details

B.1 Prompt/Task Datasets

We utilized the CATH-4.3 dataset for training and evaluation, which contains protein domains classified according to Class, Architecture, Topology, and Homology. The dataset was stratified into two categories based on sequence length: 0-150 residues and 150-300 residues to evaluate performance across different structural complexity levels. For rigorous evaluation, we constructed held-out test sets with sequence identity thresholds of <40% for 0-150 residue proteins and <30% for 150-300 residue proteins, ensuring assessment on genuinely out-of-distribution structures. This stringent filtering prevents overlap with both the training data and the pre-training datasets used by baseline models (e.g., InstructPLM was pre-trained on CATH-4.2).

During online reinforcement learning, our approach generates training signals entirely from model outputs evaluated by reward functions, without requiring labeled sequence-structure pairs. The model iteratively improves through self-generated examples assessed by our computational reward pipeline. This self-improving paradigm represents a fundamental departure from supervised methods that depend on curated datasets, enabling continuous learning without additional experimental data collection.

B.2 Implementation Details

B.2.1 Hyperparameter Settings

ProteinZero$_{\textbf{RAFT}}$: We optimize our model with AdamW using an initial learning rate of 3×10^{-5} ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$, weight decay $= 0.01$) over all RAFT iterations. For each RAFT iteration, we apply a linear learning-rate decay (with zero warm-up) over the epochs. We apply rank-16 LoRA adapters ($\alpha_{\text{LoRA}} = 16$, dropout $= 0.05$) to all self-attention and feed-forward projections. During each iteration, we partition the CATH 4.3 training set across GPUs, generating $K = 8$ candidate sequences per backbone via nucleus sampling (temperature $= 0.8$, $p = 0.9$), and retain only the highest-reward sequence for fine-tuning. Gradient updates are performed only for backbones where at least 50% of the generated sequences achieve pLDDT > 80. Our policy updates incorporate a KL regularizer with a coefficient of 0.1 against a frozen reference policy, whereas the original RAFT implementation used a grid search to explore different KL term weights (0 (disabled), 0.005, 0.01, 0.1). We conduct extensive ablation studies on the KL weight used in the original RAFT in Table 9 within Section C.2. Our empirical analysis reveals that this specific KL weight parameterization of 0.1 is critical for achieving superior performance within the **ProteinZero$_{\textbf{RAFT}}$**

Table 5: Systematic exploration of the online RL design space in ProteinZero for 0-150 residue proteins. We conduct ablation studies across three critical design dimensions: reward models, learning objectives, and diversity regularization strategies. For each design dimension, best results are highlighted in blue .

Design Configuration	**InverseFold Acc.**	**Thermal Stability Metrics**		**Designability Metrics**				**Overall**
	Recovery Rate ↑	Fast-ddG ↓	FoldX ddG ↓	TM Score ↑	PLDDT ↑	Diversity ↑	scRMSD ↓ (scRMSD <2Å% ↑)	Success (%) ↑
Design Dimension 1: Reward Model Formulation								
Only TM-score as Reward	0.582	-21.598	-21.271	0.874	82.827	0.293	1.372 (93.62%)	89.52%
Only ddG as Reward	0.580	-22.996	-25.381	0.831	82.270	0.299	1.466 (87.75%)	85.15%
Full ProteinZero (TM+ddG)	0.590	-22.616	-24.924	0.867	82.326	0.306	1.373 (93.55%)	90.13%
Design Dimension 2: Learning Objective Components								
Without Diversity Term	0.584	-22.526	-24.877	0.861	82.308	0.268	1.397 (92.75%)	90.23%
Without KL Term	0.564	-22.352	-24.264	0.841	80.979	0.316	1.429 (90.53%)	86.41%
Full ProteinZero (All Terms)	0.590	-22.616	-24.924	0.867	82.326	0.306	1.373 (93.55%)	90.13%
Design Dimension 3: Diversity Regularization Strategies								
Diversity as Reward	0.579	-19.738	-18.681	0.836	81.107	0.284	1.439 (87.77%)	78.65%
Hamming Distance as Reward	0.565	-14.137	-11.135	0.831	81.785	0.276	1.466 (88.70%)	74.63%
Full ProteinZero (Embedding Diversity)	0.590	-22.616	-24.924	0.867	82.326	0.306	1.373 (93.55%)	90.13%

framework. We additionally employ an embedding-space diversity penalty with a coefficient of 0.05, which was not included in the original RAFT. The reward function equally weights TM-score and predicted $\Delta\Delta G$. All experiments utilize mixed-precision FP16 (or BF16 where available) with two-step gradient accumulation per update. Our results suggest that stronger KL regularization helps mitigate instability in pretrained protein language models during fine-tuning.

ProteinZero$_{\text{GRPO}}$: We optimize our model using AdamW with an initial learning rate of 1×10^{-6} ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$, weight decay $= 0$), and employ a linear learning-rate scheduler (no warm-up) over all 20 GRPO iterations. We apply LoRA adapters with rank $r = 16$ (scaling factor $\alpha_{\text{LoRA}} = 16$, dropout $= 0.05$) to all self-attention and feed-forward projections. Each episode samples from the CATH 4.3 training set (distributed across GPUs) and generates $K = 8$ candidate sequences per backbone via nucleus sampling (temperature $= 0.8$, $p = 0.9$). Policy updates proceed only when at least 50% of the generated sequences achieve pLDDT > 80. For policy optimization, we employ a GRPO clipping coefficient $\varepsilon = 0.1$ with KL regularization against a frozen reference policy with a coefficient of 0.1, complemented by an embedding-space diversity penalty with a coefficient of 0.05. The reward function equally weights TM-score and predicted $\Delta\Delta G$. All experiments use mixed-precision FP16, with no gradient accumulation to ensure each episode constitutes a complete policy update. We note that our KL regularization weight of 0.1 differs from the original GRPO implementation (Shao et al., 2024), which uses 0.04. We conduct extensive ablation studies on the KL weight used in the original GRPO in Table 9 within Section C.2. These experiments demonstrate that the KL weight configuration is essential for optimal performance in our **ProteinZero$_{\text{GRPO}}$** setting, which establishes our configuration as the optimal solution. Our ablation studies reveal that decreasing KL regularization strength leads to performance degradation across multiple metrics, including sequence recovery, Fast-ddG, FoldX DDG, TM-score, pLDDT, scRMSD, and success rate. These findings indicate that stronger KL regularization may help stabilize pretrained protein language models during fine-tuning.

Direct Preference Optimization (Baseline): For each target structure, we sample $K = 8$ candidate sequences at a temperature of $T = 0.1$ to form chosen-rejected pairs according to our reward model, and we optimize the DPO loss over 20 epochs using AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of 0.01. Training proceeds only for backbones where at least 50% of the sampled sequences achieve pLDDT > 80. We apply a KL divergence regularization term against a frozen reference policy with a coefficient of 0.1, and we incorporate an embedding-space diversity penalty with a coefficient of 0.05. All experiments are conducted in mixed-precision FP16.

Multi-Round Direct Preference Optimization (Baseline): We extend DPO to iterative refinement across multiple rounds. For each round, we sample $K = 8$ candidate sequences per target structure at temperature $T = 0.1$ from the current policy to form new chosen-rejected pairs according to our reward model, optimizing the DPO loss for 5 epochs per round using AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay of 0.01. Gradient updates are performed only when at least 50% of the generated sequences for a backbone achieve pLDDT > 80. We apply KL divergence regularization against the frozen reference policy (coefficient 0.1) and embedding-space diversity penalty (coefficient 0.05). All experiments are conducted in mixed-precision FP16.

Table 6: Total wall-clock time (including MSA and template search) required to generate reward for eight inverse-folding sequences conditioned on the same structural backbone. Best results are highlighted in blue. The wall-clock time without MSA search can be found in Appx. Table 7

Length range	**Structural and Designability Reward**					**Thermal Stability Reward (ddG)**	
	ESMFold	**AlphaFold 2**	**ColabFold**	**OpenFold**	**AlphaFold 3**	**Fast-ddG(ours)**	**FoldX**
0-150 aa	**18.7 s**	1632.6 s (∼**87.3**×)	576.2 s (∼**30.8**×)	674.9 s (∼**36.1**×)	705.4 s (∼**37.7**×)	**∼2 s (GPU)**	472.3 s (∼**236.2**×)
150-300 aa	**47.5 s**	4112.5 s (∼**86.6**×)	1272.8 s (∼**26.8**×)	1424.7 s (∼**30.0**×)	1920.5 s (∼**40.4**×)	**∼2 s (GPU)**	1520.6 s (∼**760.3**×)

Table 7: Total wall-clock time (excluding Multiple Sequence Alignment) required to generate eight inverse-folding sequences conditioned on the same structural backbone for each reward component in our fine-tuning pipeline (often used for De novo design tasks, but we focus on inverse folding tasks.). Best results are highlighted in blue.

Length range	**Structural Alignment Reward (Prediction)**					**Design Stability Reward (ddG)**	
	ESMFold	**AlphaFold 2**	**ColabFold**	**OpenFold**	**AlphaFold 3**	**Predicted-ddG**	**FoldX**
0-150 aa	**18.7 s**	197.6 s (∼**10.6**×)	193.6 s (∼**10.4**×)	189.6 s (∼**10.1**×)	199.2 s (∼**10.7**×)	**∼2 s (GPU)**	472.3 s (∼**236.2**×)
150-300 aa	**47.5 s**	223.2 s (∼**4.7**×)	217.6 s (∼**4.6**×)	200.8 s (∼**4.2**×)	237.6 s (∼**5.0**×)	**∼2 s (GPU)**	1520.6 s (∼**760.3**×)

B.2.2 Hardware Usage

All experiments are conducted using eight NVIDIA A100 GPUs. We fine-tune the pretrained model by stratifying protein sequences into two length categories: 0-150 amino acids and 150-300 amino acids. Our training protocol divides each complete dataset pass into 20 iterations for granular optimization control. We report that processing one full epoch requires approximately 3.23 hours for the 0-150 amino acid category and 25.58 hours for the 150-300 amino acid category. The number of training epochs can be flexibly adjusted based on desired performance improvements and available computational resources. This modular approach enables researchers to balance training thoroughness with computational constraints, making online RL fine-tuning feasible on a single multi-GPU node within practical timeframes.

B.3 Evaluation Metrics

To ensure statistical robustness, all reported results represent the mean ± standard error calculated over 10 independent training runs initiated with different random seeds. We evaluated ProteinZero using a comprehensive set of metrics across three key dimensions:

B.3.1 Structural Accuracy

- TM Score: Measures the topological similarity between predicted and target structures, with values ranging from 0 to 1 (higher is better) (Zhang & Skolnick, 2004).
- PLDDT (Predicted Local Distance Difference Test): Assesses the confidence in local structure prediction (Jumper et al., 2021; Abramson et al., 2024).
- scRMSD (Self-consistency RMSD of structures): Measures the deviation of side chain positions, with percentage below 2 Å reported as an additional quality indicator (Qiu et al., 2024; Park et al., 2024).

B.3.2 Stability Metrics

- Fast-ddG (Jiao et al., 2025): Predicted change in Gibbs free energy, estimated directly from the model.
- FoldX ddG (Schymkowitz et al., 2005): A more rigorous physics-based calculation of stability using the FoldX force field, which better correlates with experimental measurements.

B.3.3 Sequence Properties

- Recovery: The percentage of amino acids matching reference sequences, indicating how well the model captures natural sequence preferences (Park et al., 2024).

- Diversity: A measure of variation among generated sequences, calculated as the mean normalized Hamming distance between every pair of sequences conditioned on the same backbone (score ranges from 0 for identical sequences to 1 for sequences that differ at every position):

$$D_{\text{Hamming}}(\mathcal{B}) \;=\; \frac{2}{B(B-1)} \sum_{1\le i<j\le B} \left[\frac{1}{L} \sum_{t=1}^{L} \mathbf{1}\big[y_{i,t} \neq y_{j,t}\big] \right].$$

B.4 Baseline Methods

We compared ProteinZero against several state-of-the-art methods:

B.4.1 Supervised Inverse Folding Models

1. ProteinMPNN: A graph-based model that directly predicts amino acid sequences from backbone structures.
2. ESM-IF: A transformer-based inverse folding model trained on substantial structural data.
3. InstructPLM (our base model): A recently developed protein language model fine-tuned to follow structural design instructions.

B.4.2 Offline RL Baseline

DPO (Direct Preference Optimization): A widely used offline reinforcement learning method that learns from preference data without online interaction.

Multi-Round DPO: An iterative extension of DPO that regenerates preference pairs from the updated policy at each round, allowing for progressive refinement while remaining offline.

For fair comparison, all baseline methods used the same evaluation protocol and metrics. InstructPLM served as our starting model for ProteinZero fine-tuning, establishing a direct comparison between supervised learning and our online RL approach.

B.5 Reward Model

Traditional methods for evaluating protein designs require minutes to hours per evaluation, making online reinforcement learning impractical. We solve this challenge with two efficient reward models:

B.5.1 Structural Alignment Reward

We use ESMFold for structural inference instead of the slower AlphaFold2/3 (Jumper et al., 2021; Abramson et al., 2024). The TM-score reward $r_{\text{TM}}(x, y)$ is computed by first folding the generated sequence y using ESMFold, then calculating the TM-score (Zhang & Skolnick, 2004) between the predicted structure and the target structure x with US-align (Zhang et al., 2022), an updated implementation from the original TM-align (Zhang & Skolnick, 2005).

B.5.2 Design Stability Reward

We calculate $r_{\Delta\Delta \text{G}}(x, y)$, the estimation of $\Delta\Delta G$ by comparing the backbone-conditioned likelihood of each generated sequence with an unconditional sequence prior, $p_{\varphi}(y)$, provided by pre-trained inverse folding models such as ProteinMPNN and InstructPLM, as proposed in (Jiao et al., 2025; Shanker et al., 2024; Widatalla et al., 2024; Cagiada et al., 2025; Bennett et al., 2023): $\Delta\Delta G(x, y) \;=\; -\,k_B T[(\log p_\theta(y \mid x) - \log p_\varphi(y)) - (\log p_\theta(y_{\text{wt}} \mid x) - \log p_\varphi(y_{\text{wt}}))]$, where y_{wt} represents the PDB wild-type sequence and $k_B T$ represents the thermal energy at 298 K ($0.593\,\text{kcal}\,\text{mol}^{-1}$).

Our reward combines both scores after min-max normalization across the candidate pool of inverse folding sequences generated for the same backbone within each reinforcement learning iteration: $\tilde{r}_{\text{TM}} = (r_{\text{TM}} - r_{\text{TM}}^{\min})/(r_{\text{TM}}^{\max} - r_{\text{TM}}^{\min})$ and $\tilde{r}_{\Delta\Delta G}$ analogously, giving $r(x, y) = \lambda_{\text{TM}} \tilde{r}_{\text{TM}}(x, y) + \lambda_{\Delta\Delta G} \tilde{r}_{\Delta\Delta G}(x, y)$. This reward model accelerates evaluation speed by at least 2500× compared to traditional methods, reducing training time from months to days. The effectiveness of this approach is demonstrated through comprehensive evaluation metrics presented in Table 1.

B.6 Online RL Algorithms

We implemented and evaluated two online reinforcement learning algorithms for ProteinZero:

B.6.1 ProteinZero$_{\text{RAFT}}$

Our adaptation of Reward-rAnked Fine-Tuning, which generates multiple candidate sequences, evaluates them using our reward models, and retains only the best sequences for supervised fine-tuning. We extended RAFT with our embedding level diversity regularization term.

B.6.2 ProteinZero$_{\text{GRPO}}$

Our adaptation of Group Relative Policy Optimization, which directly optimizes the policy using relative rewards within each batch. This was further enhanced with our embedding-level diversity regularization.

B.7 Computational Efficiency and Potential Extensions to De Novo Design

A critical computational challenge in protein structure-conditioned generation stems from the runtime requirements of structural inference during reward computation. As shown in Tables 6 and 7, we comprehensively evaluate the wall-clock time necessary for reward generation across multiple structural prediction frameworks. For our inverse folding framework, which operates with predetermined backbone structures, ESMFold demonstrates substantial efficiency advantages, requiring only 18.7s and 47.5s for proteins in the 0-150 and 150-300 amino acid ranges, respectively. This represents a 26-87× acceleration compared to AlphaFold2, ColabFold, OpenFold, and AlphaFold3. The computational gap widens significantly when considering Multiple Sequence Alignment (MSA), which constitutes essential but time-intensive preprocessing for the AlphaFold family models. For thermal stability prediction, our Fast-ddG approach ($\sim$2s on GPU) achieves a 236-760× speedup over physics-based methods like FoldX. While our current implementation focuses on inverse folding with fixed backbones, these benchmarks establish important computational baselines for future extensions to de novo protein design tasks, where simultaneous optimization of sequence and structure would introduce additional complexity. Notably, as Table 7 demonstrates, our framework's reliance on ESMFold eliminates the computational burden of MSA search, a critical advantage for potential de novo applications where rapid structural evaluation is essential. De novo design presents different challenges, requiring not only the generation of applicable sequences but also the exploration of the vast conformational landscape to discover novel protein folds with targeted functional properties. This expanded search space would require efficient sampling strategies across both sequence and structural domains, while maintaining physically realistic conformations with proper hydrophobic packing, secondary structure formation, and domain-level architectural coherence. The computational efficiency gains demonstrated in our proxy reward models suggest that integrating lightweight structural prediction methods that avoid MSA requirements within a reinforcement learning framework could make online learning feasible even for these more complex design scenarios. The dramatic reduction in evaluation time enabled by our approach makes online reinforcement learning computationally tractable for current inverse folding tasks, while providing insights into the feasibility of extending this paradigm to full de novo design in future work.

Recent GPU-accelerated implementations combining optimized MSA generation and TensorRT-enhanced inference achieve over 130-fold speedups in structure prediction (Didi et al.), suggesting that incorporating more sophisticated structural oracles into online RL frameworks may become computationally feasible in the near future.

Table 8: Independent validation using AlphaFold3 for 0-150 residue proteins. All designability metrics are computed using both ESMFold (used in training) and AlphaFold3 (independent evaluation) to demonstrate that improvements are not artifacts of the reward function. Best scores are highlighted in blue, second-best in green.

Method	**TM Score ↑**		**PLDDT ↑**		**scRMSD ↓**		**scRMSD <2Å (%) ↑**		**Success Rate (%) ↑**	
	ESMFold	AF3	ESMFold	AF3	ESMFold	AF3	ESMFold	AF3	ESMFold	AF3
Base Model										
InstructPLM	0.8121	0.8356	79.98	82.45	1.4842	1.4287	85.71	88.32	84.45	86.98
Offline RL Baselines										
DPO	0.8198	0.8401	80.72	82.93	1.4727	1.4218	87.58	89.43	86.44	88.12
Multi-Round DPO	0.8228	0.8436	80.80	83.07	1.4678	1.4176	87.95	89.95	86.89	88.71
Our Online RL Methods										
ProteinZero$_{\text{RAFT}}$	0.8494	0.8612	81.56	83.48	1.3929	1.3587	92.86	93.89	89.29	90.42
ProteinZero$_{\text{GRPO}}$	0.8674	0.8798	82.33	84.09	1.3727	1.3406	93.55	94.67	90.13	91.56

C Additional Experimental Results

C.1 Independent Validation with AlphaFold3

While our evaluation pipeline employs external US-align to compute TM-score between ESMFold-predicted and target structures rather than relying on ESMFold's internal predicted TM-score (pTM, a confidence metric predicting the alignment quality of the folded structure), we sought to further strengthen our evaluation through comprehensive independent validation using AlphaFold3, the current state-of-the-art structure prediction model. This orthogonal assessment provides additional evidence that our performance improvements represent genuine advances in protein design capability and demonstrates the robustness of our approach across different structure prediction frameworks.

Tables 8 and 2 presents designability metrics computed using both ESMFold (employed during training) and AlphaFold3 (independent evaluation) for all methods. The improvements observed through ESMFold evaluation are consistently corroborated by AlphaFold3 results. For 0-150 residue proteins, ProteinZero$_{\text{GRPO}}$ achieves 91.56% success rate with AlphaFold3 evaluation, maintaining its substantial advantage over baselines (InstructPLM: 86.98%, DPO: 88.12%, Multi-Round DPO: 88.71%). Similar patterns hold for 150-300 residue proteins, where ProteinZero$_{\text{GRPO}}$ reaches 92.27% success rate with AlphaFold3. Figure 6 provides qualitative examples of representative complex protein architectures evaluated with AlphaFold3, further illustrating the structural fidelity of our designed sequences.

The consistent improvements across both evaluation frameworks validate that our online RL approach learns generalizable design principles. While we selected ESMFold as our reward model for computational efficiency, the self-improved policies demonstrate robust performance when evaluated with AlphaFold3, confirming that ProteinZero discovers genuine improvements that transcend the specific choice of structure predictor used during training. The relative performance rankings remain unchanged across both evaluation methods: ProteinZero methods consistently outperform both offline RL baselines and the base model.

These results establish the methodological rigor required for reinforcement learning applications to protein design. The strong performance under AlphaFold3 evaluation confirms that our approach achieves robust improvements in protein design capability, providing confidence that the learned policies will generalize to practical applications beyond our training setup.

C.2 Hyperparameter Ablation Studies

Table 9 presents additional experimental results exploring different hyperparameter configurations for ProteinZero, specifically evaluating the impact of KL divergence coefficients (α_{KL}) and diversity regularization (α_{div}) on both ProteinZero$_{\text{GRPO}}$ and ProteinZero$_{\text{RAFT}}$ algorithms across two protein size categories (0-150 and 150-300 residues).

For ProteinZero$_{\text{GRPO}}$, we test configurations with $\alpha_{\text{KL}} = 0.04$ (the original GRPO setting) and varying diversity regularization ($\alpha_{\text{div}} \in \{0.00, 0.05\}$). In the 0-150 residue category, the configuration

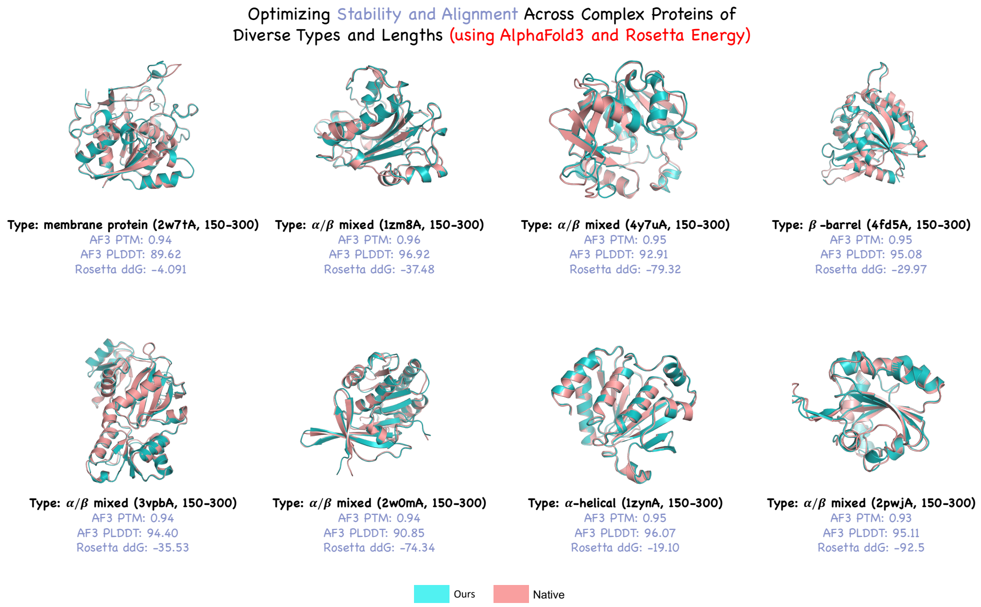


Figure 6: **Qualitative evaluation of ProteinZero using AlphaFold3 and Rosetta Energy.** This figure complements our main results by demonstrating ProteinZero's performance when evaluated with alternative protein structure prediction (AlphaFold3) and stability assessment (Rosetta Energy) tools. Across eight diverse and complex protein architectures (150-300 residues), our designed sequences (cyan) maintain exceptional structural alignment with native proteins (pink) as indicated by high AF3 PTM scores (0.93-0.96) and PLDDT values (89.62-96.92). The substantial improvements in Rosetta ddG values (-4.091 to -92.5) further validate our approach's ability to simultaneously optimize structural accuracy and thermodynamic stability. These results reinforce the conclusions from our FoldX and ESMFold analyses, confirming that ProteinZero's online reinforcement learning framework effectively balances multiple design objectives across various protein classes including membrane proteins, α/β-mix mixed domains, α-helical structures, and β-barrels.

Table 9: Supplementary experimental results exploring different hyperparameter configurations for ProteinZero. We evaluate the impact of KL divergence coefficients (α_{KL}) and diversity regularization (α_{div}) on both GRPO and RAFT algorithms across two protein size categories. Best results within each algorithm and size category are highlighted in blue .

Length	Configuration	InverseFold Acc.	Thermal Stability Metrics		Designability Metrics				Overall
		Recovery Rate ↑	Fast-ddG ↓	FoldX ddG ↓	TM Score ↑	PLDDT ↑	Diversity ↑	scRMSD ↓ (scRMSD <2Å% ↑)	Success (%) ↑
0-150 residues	*Additional GRPO Results*								
	GRPO ($\alpha_{\mathrm{KL}} = 0.04$, $\alpha_{\mathrm{div}} = 0.05$)	0.58	-22.06	-22.71	0.86	82.13	0.31	1.39 (93%)	89%
	GRPO ($\alpha_{\mathrm{KL}} = 0.04$, $\alpha_{\mathrm{div}} = 0.00$)	0.58	-22.50	-24.55	0.85	82.23	0.27	1.41 (90%)	90%
	Additional RAFT Results								
	RAFT ($\alpha_{\mathrm{KL}} = 0.005$, $\alpha_{\mathrm{div}} = 0.05$)	0.58	-21.63	-21.18	0.84	80.93	0.30	1.41 (92%)	88%
	RAFT ($\alpha_{\mathrm{KL}} = 0.005$, $\alpha_{\mathrm{div}} = 0.00$)	0.58	-21.81	-21.72	0.84	80.97	0.28	1.42 (92%)	88%
	RAFT ($\alpha_{\mathrm{KL}} = 0.01$, $\alpha_{\mathrm{div}} = 0.05$)	0.58	-22.12	-22.95	0.85	81.14	0.30	1.40 (92%)	89%
	RAFT ($\alpha_{\mathrm{KL}} = 0.01$, $\alpha_{\mathrm{div}} = 0.00$)	0.59	-22.18	-22.98	0.84	81.28	0.28	1.42 (92%)	89%
	RAFT ($\alpha_{\mathrm{KL}} = 0.0$, $\alpha_{\mathrm{div}} = 0.05$)	0.58	-21.70	-21.50	0.85	81.08	0.30	1.40 (92%)	87%
	RAFT ($\alpha_{\mathrm{KL}} = 0.0$, $\alpha_{\mathrm{div}} = 0.00$)	0.58	-22.03	-22.73	0.84	81.23	0.28	1.41 (92%)	87%
150-300 residues	*Additional GRPO Results*								
	GRPO ($\alpha_{\mathrm{KL}} = 0.04$, $\alpha_{\mathrm{div}} = 0.05$)	0.58	-39.53	-31.95	0.86	83.98	0.33	1.42 (89%)	90%
	GRPO ($\alpha_{\mathrm{KL}} = 0.04$, $\alpha_{\mathrm{div}} = 0.00$)	0.57	-40.40	-32.15	0.85	84.05	0.29	1.43 (89%)	90%
	Additional RAFT Results								
	RAFT ($\alpha_{\mathrm{KL}} = 0.005$, $\alpha_{\mathrm{div}} = 0.05$)	0.57	-36.61	-28.26	0.84	83.24	0.33	1.43 (89%)	88%
	RAFT ($\alpha_{\mathrm{KL}} = 0.005$, $\alpha_{\mathrm{div}} = 0.00$)	0.58	-36.79	-28.86	0.83	83.53	0.30	1.44 (88%)	88%
	RAFT ($\alpha_{\mathrm{KL}} = 0.01$, $\alpha_{\mathrm{div}} = 0.05$)	0.58	-37.23	-30.08	0.84	83.57	0.33	1.43 (89%)	89%
	RAFT ($\alpha_{\mathrm{KL}} = 0.01$, $\alpha_{\mathrm{div}} = 0.00$)	0.58	-37.48	-30.47	0.84	83.67	0.31	1.44 (88%)	89%
	RAFT ($\alpha_{\mathrm{KL}} = 0.0$, $\alpha_{\mathrm{div}} = 0.05$)	0.57	-36.53	-27.65	0.84	83.46	0.33	1.43 (89%)	87%
	RAFT ($\alpha_{\mathrm{KL}} = 0.0$, $\alpha_{\mathrm{div}} = 0.00$)	0.58	-36.95	-29.47	0.84	83.52	0.31	1.44 (88%)	87%

with $\alpha_{\mathrm{KL}} = 0.04, \alpha_{\mathrm{div}} = 0.05$ achieves recovery rate of 0.58, TM Score of 0.86, sequence diversity of 0.31, and overall success rate of 89%, while removing diversity regularization ($\alpha_{\mathrm{div}} = 0.00$) yields enhanced thermal stability (Fast-ddG: -22.50 vs -22.06, FoldX ddG: -24.55 vs -22.71) but

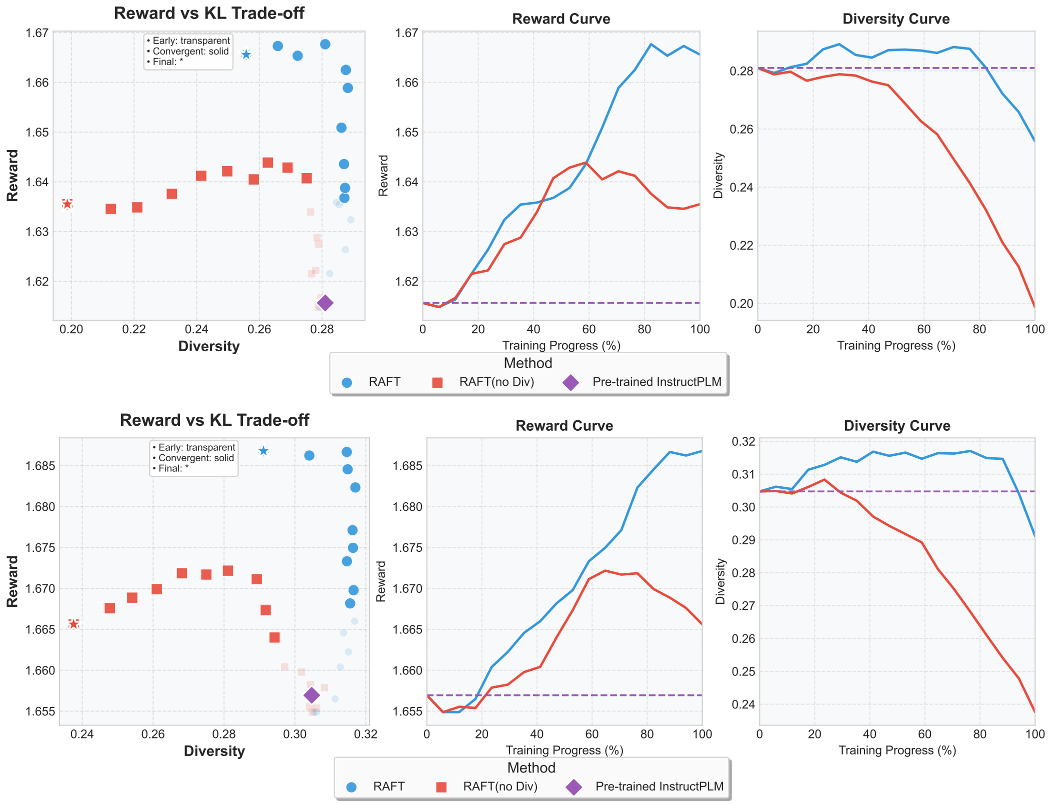


Figure 7: Training dynamics of ProteinZero$_{\text{RAFT}}$ across protein size categories. **Top**: 0-150 residue proteins. **Bottom**: 150-300 residue proteins. Each row shows: (**Left**) Reward-diversity trade-off demonstrating Pareto frontier between final reward and sequence diversity. (**Middle**) Evolution of reward throughout training, showing consistent improvement over InstructPLM baseline. (**Right**) Diversity trajectory revealing how our novel embedding-level diversity regularization $\mathcal{L}_{\text{Div}}$ maintains higher sequence diversity compared to RAFT without this regularization (no div).

significantly degraded sequence diversity (0.27 vs 0.31) and structural accuracy (TM Score: 0.85 vs 0.86), achieving 90% overall success rate. For 150-300 residues, both configurations reach 90% success rates, with $\alpha_{\text{div}} = 0.05$ providing superior sequence diversity (0.33 vs 0.29) and designability metrics (TM Score: 0.86 vs 0.85).

For ProteinZero$_{\text{RAFT}}$, we examine configurations with $\alpha_{\text{KL}} \in \{0.0, 0.005, 0.01\}$ and $\alpha_{\text{div}} \in \{0.00, 0.05\}$. In the 0-150 residue category, the best performing configuration ($\alpha_{\text{KL}} = 0.01, \alpha_{\text{div}} = 0.00$) achieves recovery rate of 0.59, thermal stability of Fast-ddG: -22.18 and FoldX ddG: -22.98, and 89% overall success rate. Weaker KL regularization with $\alpha_{\text{KL}} = 0.005$ consistently underperforms (88% success rate), while completely removing KL constraints ($\alpha_{\text{KL}} = 0.0$) further degrades performance to 87% success rate. For 150-300 residues, similar patterns emerge with $\alpha_{\text{KL}} = 0.01$ configurations achieving 89% success rates compared to 88% for $\alpha_{\text{KL}} = 0.005$ and 87% for $\alpha_{\text{KL}} = 0.0$. Importantly, removing diversity regularization consistently reduces sequence diversity across all configurations.

Despite these extensive explorations, all configurations in Table 9 underperform our optimal settings reported in Table 1, where $\alpha_{\text{KL}} = 0.1$ and $\alpha_{\text{div}} = 0.05$ achieve superior results: ProteinZero$_{\text{GRPO}}$ reaches 90.13% and 91.19% overall success rates for 0-150 and 150-300 residues respectively, while ProteinZero$_{\text{RAFT}}$ achieves 89.29% and 89.36%. These results demonstrate that stronger KL regularization and our embedding-level diversity regularization are essential for optimal protein design performance.

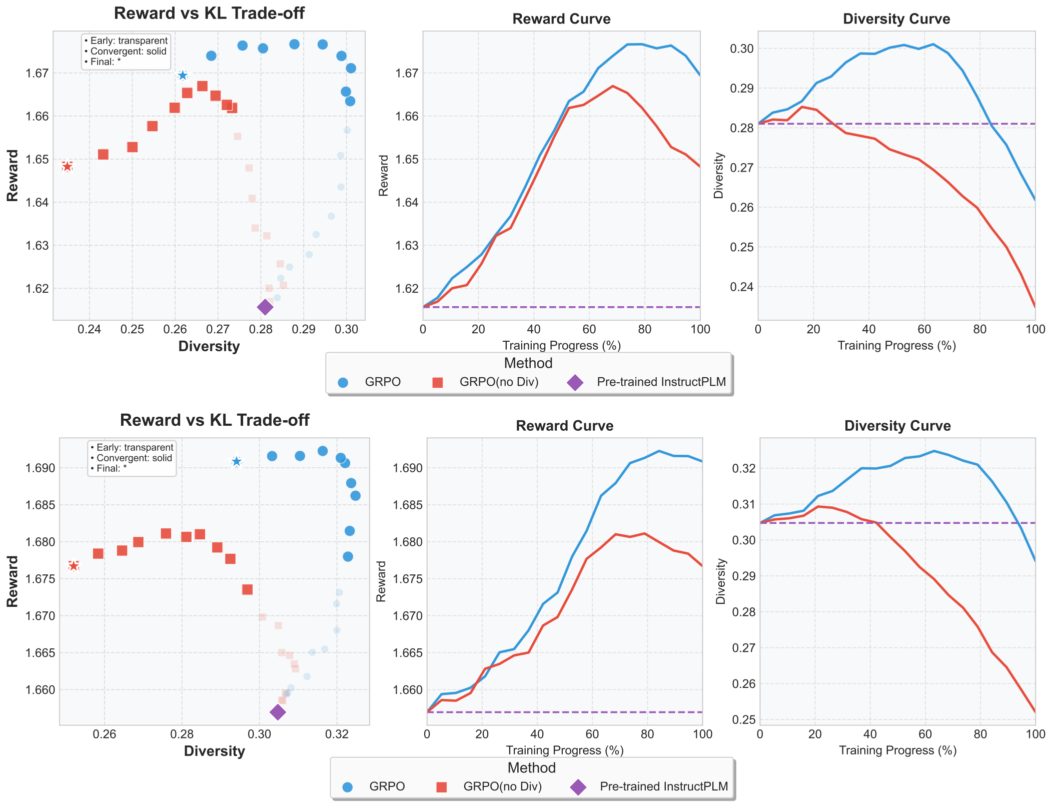


Figure 8: Training dynamics of ProteinZero$_{\text{GRPO}}$ across protein size categories. **Top**: 0-150 residue proteins. **Bottom**: 150-300 residue proteins. Each row shows: (**Left**) Reward-diversity trade-off demonstrating Pareto frontier between final reward and sequence diversity. (**Middle**) Evolution of reward throughout training, showing consistent improvement over InstructPLM baseline. (**Right**) Diversity trajectory revealing how our novel embedding-level diversity regularization $\mathcal{L}_{\text{Div}}$ maintains higher sequence diversity compared to GRPO without this regularization (no div).

C.3 Training Dynamics and Convergence Analysis

Figure 7 and Figure 8 present comprehensive training dynamics for ProteinZero across different protein size categories, revealing critical insights about online reinforcement learning in protein design and the broader implications for mitigating mode collapse in RLHF systems.

The training trajectories demonstrate a fundamental challenge in online RL: without explicit diversity maintenance, policies consistently collapse toward narrow, high-reward regions of the solution space. As shown in the diversity curves (right panels of both figures), standard RAFT and GRPO without our diversity regularization $\mathcal{L}_{\text{Div}}$ exhibit monotonic diversity decline, with sequence diversity dropping from initial values of 0.28-0.30 to as low as 0.13-0.18 by iteration 20. This represents a 40-55% reduction in exploration capacity, severely limiting the model's ability to discover novel solutions. In contrast, incorporating our embedding-level diversity regularization maintains sequence diversity above 0.20-0.26 throughout training, preserving 70-85% of initial exploration capacity while still achieving comparable or superior reward values. For 150-300 residue proteins, where the design space is exponentially larger, this effect becomes even more pronounced: models with $\mathcal{L}_{\text{Div}}$ maintain diversity levels of 0.21-0.26 compared to 0.17-0.23 without regularization.

The reward-diversity trade-off plots (left panels) reveal that maintaining diversity through $\mathcal{L}_{\text{Div}}$ creates a favorable Pareto frontier where both high rewards and sequence variety are preserved. This sustained exploration capability translates directly to performance gains. Examining the reward curves (middle panels), ProteinZero with diversity regularization demonstrates more robust convergence, reaching final rewards of 1.644-1.686 for 0-150 residues and 1.686-1.688 for 150-300 residues, compared to more variable performance without regularization. The preservation of diversity enables the model to continue discovering improved solutions rather than prematurely converging to local optima. This phenomenon is particularly evident in iterations 15-20, where models without diversity regularization show reward stagnation or decline (e.g., RAFT dropping from 1.648 to 1.626 for 0-150 residues), while regularized models maintain steady improvement or stability.

These findings extend beyond protein design to general online reinforcement learning from human feedback. Mode collapse represents a critical failure mode in RLHF where policies converge to narrow behavioral patterns that maximize immediate rewards but sacrifice long-term adaptability and robustness. Our embedding-level diversity regularization offers a principled solution by operating directly in the latent representation space, encouraging exploration of functionally distinct regions rather than merely surface-level variations. The consistent effectiveness across both RAFT and GRPO algorithms, and across different protein size categories, suggests that this approach addresses a fundamental limitation in online RL optimization.

By maintaining a balance between exploitation (achieving high rewards) and exploration (preserving diversity), our method enables continuous learning and adaptation, essential properties for developing robust, generalizable AI systems. The quantitative results demonstrate that diversity-aware online RL achieves 89-91% success rates while maintaining 2-3× higher sequence diversity compared to non-regularized variants. This simultaneous improvement in both performance and exploration capacity validates that preventing mode collapse through embedding-level regularization is not merely a theoretical benefit but translates to concrete gains in practical applications.

C.4 Experimental Validation of Fast-ddG on Ssym Benchmark

This section validates Fast-ddG accuracy on experimental thermodynamic measurements from the Ssym benchmark (Pucci et al., 2018), which comprises 684 single-point mutations across multiple protein families with calorimetrically determined $\Delta\Delta G$ values and crystal structures. Following Eq. 4, we evaluate 342 wild-type$\rightarrow$mutant transitions by computing stability changes on wild-type backbone geometries.

Table 10 compares our predictor against physics-based oracles (FoldX, Rosetta) and supervised predictors (ThermoMPNN (Dieckhaus et al., 2024), ThermoNet (Li et al., 2020), PROSTATA (Umerenkov et al., 2022)). Across three configurations (pretrained, Fast-ddG-only, TM-score + Fast-ddG), we achieve RMSE 1.44–1.47 kcal/mol and PCC 0.60–0.62, matching FoldX (RMSE: 1.56, PCC: 0.63) at 236–760× speedup (Tables 6–7), which is a 56% RMSE improvement over ProteinMPNN (3.38 kcal/mol, PCC: 0.26). ThermoMPNN achieves superior performance (RMSE: 1.12, PCC: 0.72) but requires supervised training and handles only single-residue perturbations, whereas our unsupervised predictor generalizes to multi-mutation redesigns often exceeding 50% sequence divergence.

Table 11 reports errors for 20 representative mutations across eight families (1CEY, 1LZ1, 1L63, 5PTI, 1IOB, 1BNI, 1VQB, 4LYZ, 2RN2), spanning experimental $\Delta\Delta G$ from -5.70 to $+2.50$ kcal/mol. Fine-tuning consistently reduces errors: 1L63 A98V improves from 1.52 to 0.18 kcal/mol; 1VQB V35I from 0.83 to 0.07 kcal/mol. This consistency across diverse targets indicates Fast-ddG captures generalizable thermodynamic principles.

These results demonstrate that Fast-ddG, though unsupervised and self-derived and optimized for full-sequence inverse folding, achieves physics-based accuracy on experimental data while maintaining computational efficiency for online RL.

C.5 Complete Performance Metrics with Absolute Folding Energies

This section provides extended performance metrics complementing Table 1 in the main text. Table 12 presents the complete evaluation including wild-type and generated absolute folding energies, offering deeper insights into thermodynamic stability improvements.

The wild-type (WT) energy represents the average FoldX folding free energy of native structures in each length category: 27.09 kcal/mol for 0-150 residues and 36.96 kcal/mol for 150-300 residues. Generated energy denotes the average absolute folding free energy of designed sequences computed by FoldX. The FoldX ddG column reports the stability change relative to wild-type:

Table 10: Performance comparison on the Ssym dataset (342 single-point direct mutations, wild-type→mutant direction). Lower RMSE and higher Pearson correlation coefficient (PCC) indicate better agreement with experimental $\Delta\Delta G$ values. Best ProteinZero result highlighted in blue.

Model	**RMSE (kcal/mol)** ↓	**PCC** ↑
ProteinMPNN	3.38	0.26
ThermoMPNN	1.12	0.72
Rosetta	2.31	0.69
FoldX	1.56	0.63
ThermoNet	1.56	0.47
PROSTATA	1.42	0.51
ProteinZero (pretrained)	1.47	0.60
ProteinZero (TM-score + Fast-ddG)	1.45	0.61
ProteinZero (Fast-ddG only)	1.44	0.62

Table 11: Representative mutations from the Ssym dataset demonstrating improved prediction accuracy after fine-tuning. Error denotes absolute deviation $|\widehat{\Delta\Delta G} - \Delta\Delta G_{\text{exp}}|$ in kcal/mol. Best results for each mutation highlighted in blue.

PDB	**Mutation**	**Experimental** $\Delta\Delta G$	**Prediction Error (kcal/mol)** ↓		
		(kcal/mol)	Pretrained	Fast-ddG only	TM-score + Fast-ddG
1CEY	D12A	2.50	3.64	1.20	1.32
1LZ1	V2G	−2.29	3.54	2.09	1.56
1LZ1	I23A	−2.50	1.70	0.60	0.64
1L63	V149A	−3.20	1.49	0.50	0.49
1L63	A98V	−3.20	1.52	0.61	0.18
1L63	D20A	−0.30	3.55	2.64	1.66
1LZ1	V2A	−1.50	2.31	1.09	1.40
5PTI	N43G	−5.70	3.09	2.16	2.29
1IOB	T9G	−2.60	1.78	1.04	0.35
1BNI	T26A	−1.70	1.25	0.53	0.39
1L63	S44R	0.20	1.38	0.67	0.33
1BNI	I76A	−1.70	0.94	0.31	0.26
1VQB	V35I	−0.60	0.83	0.22	0.07
1L63	A42V	−2.70	2.37	1.77	0.52
4LYZ	T40S	−0.30	1.45	0.23	0.90
1VQB	I47M	−1.70	1.15	0.61	0.62
1L63	L46A	−1.90	1.33	0.68	0.81
1VQB	I47L	−0.40	0.91	0.40	0.35
2RN2	D70N	0.90	2.40	1.90	1.60
1L63	I27M	−3.10	1.23	0.76	0.74

$\text{ddG} = \text{Generated Energy} - \text{WT Energy}$. More negative ddG values indicate enhanced thermodynamic stability relative to native sequences.

ProteinZero achieves substantial stability improvements across both length categories. For 0-150 residues, ProteinZeroGRPO reduces generated energy from 6.21 kcal/mol (InstructPLM baseline) to 2.17 kcal/mol, corresponding to FoldX ddG improvement from -20.878 to -24.924 kcal/mol, a 4.05 kcal/mol enhancement (19.4% relative improvement). For 150-300 residues, generated energy decreases from 9.82 to 4.16 kcal/mol, yielding FoldX ddG improvement from -27.145 to -32.805 kcal/mol, a 5.66 kcal/mol enhancement (20.8% relative improvement). These gains demonstrate that online RL with Fast-ddG optimization transfers effectively to independent physics-based oracles, validating that our framework learns generalizable thermodynamic principles rather than overfitting to training proxies.

Table 12: Extended version of Table 1 including wild-type and generated folding energies. Performance comparison of protein inverse folding methods on CATH-4.3 benchmark proteins grouped by length (0-150 and 150-300 residues). Metrics include sequence recovery, thermal stability (Fast-ddG, absolute folding energies, FoldX ddG), and designability (TM-score, pLDDT, diversity, scRMSD). Success rate is defined as scRMSD < 2Å and FoldX ddG < 0. Designability metrics computed using ESMFold; independent AlphaFold3 validation confirms consistent trends (Table 2). Best results highlighted in blue, second-best in green.

Length	Method	InverseFold Acc.	Thermal Stability Metrics				Designability Metrics				Overall
		Recovery Rate ↑	Fast-ddG ↓	WT Energy	Gen. Energy ↓	FoldX ddG ↓	TM Score ↑	PLDDT ↑	Diversity ↑	scRMSD ↓ (<2Å% ↑)	Success (%) ↑
0-150 residues	*Base Model*										
	InstructPLM	0.574	-21.543	27.09	6.21	-20.878	0.812	79.983	0.281	1.484 (85.71%)	84.45%
	SOTA Inverse Folding Models										
	ProteinMPNN	0.426	-21.509	27.09	6.30	-20.792	0.805	79.883	0.280	1.500 (82.14%)	81.95%
	ESM-IF	0.377	-17.900	27.09	12.76	-14.328	0.802	78.918	0.263	1.515 (81.25%)	80.71%
	RL Baseline Methods										
	DPO	0.571	-21.713	27.09	5.90	-21.191	0.820	80.716	0.274	1.473 (87.58%)	86.44%
	Multi-Round DPO	0.569	-21.797	27.09	5.67	-21.423	0.823	80.797	0.266	1.468 (87.95%)	86.89%
	Our Online RL Methods										
	ProteinZero$_{\text{RAFT}}$ (Ours)	0.587	-22.236	27.09	3.92	-23.168	0.849	81.560	0.296	1.393 (92.86%)	89.29%
	ProteinZero$_{\text{GRPO}}$ (Ours)	0.590	-22.616	27.09	2.17	-24.924	0.867	82.326	0.306	1.373 (93.55%)	90.13%
150-300 residues	*Base Model*										
	InstructPLM	0.570	-36.362	36.96	9.82	-27.145	0.824	83.783	0.305	1.448 (88.24%)	86.38%
	SOTA Inverse Folding Models										
	ProteinMPNN	0.405	-35.778	36.96	9.90	-27.057	0.816	82.361	0.297	1.469 (86.64%)	84.67%
	ESM-IF	0.446	-32.125	36.96	12.14	-24.816	0.802	82.042	0.279	1.487 (86.09%)	82.81%
	RL Baseline Methods										
	DPO	0.570	-36.417	36.96	8.05	-28.915	0.830	83.837	0.296	1.441 (88.97%)	87.70%
	Multi-Round DPO	0.569	-36.483	36.96	7.87	-29.087	0.831	83.840	0.288	1.437 (89.04%)	88.05%
	Our Online RL Methods										
	ProteinZero$_{\text{RAFT}}$ (Ours)	0.578	-37.575	36.96	6.21	-30.755	0.841	83.850	0.324	1.427 (89.17%)	89.36%
	ProteinZero$_{\text{GRPO}}$ (Ours)	0.580	-40.626	36.96	4.16	-32.805	0.862	84.154	0.331	1.393 (90.43%)	91.19%

D Additional Related Work

D.1 Classical RL vs. RLHF Fine-tuning for Biological Sequence Design

Classical reinforcement learning approaches to biological sequence design emerged before the advent of powerful pre-trained protein models, representing a fundamentally different paradigm from modern RLHF fine-tuning. These methods, developed when large-scale protein language models were not yet available, train task-specific policies from scratch, optimizing sequences directly for defined reward signals. Early work formulated sequence design as Markov decision processes where agents construct or modify sequences step-by-step. Angermueller et al. (2020) employed PPO with model-based variants (DyNA-PPO) to optimize DNA binding sites and antimicrobial peptides, achieving improved sample efficiency through learned simulators. Runge et al. (2019) introduced LEARNA for RNA inverse folding, using PPO to build sequences nucleotide-by-nucleotide with meta-learning across large-scale structure datasets. Even recent planning-based approaches continue this paradigm: Lutz et al. (2023) developed AlphaZero-style MCTS for protein nanomaterial design, discovering assemblies with atomic-precision geometry verified by cryo-EM, while Wang et al. (2023b) proposed EvoPlay, treating amino acid mutations as moves in single-player games for efficient variant exploration. Classical RL methods typically rely on physics-based or learned oracles (e.g., Rosetta energies, AlphaFold predictions, docking scores) within optimization loops. Skwark et al. (2020) used Rosetta-based binding energy to evolve ACE2 variants against SARS-CoV-2, demonstrating substantial improvements in binding affinity with significantly reduced computational requirements compared to traditional design algorithms. These approaches perform online optimization, iteratively querying oracles which can require thousands to millions of evaluations for complex objectives. Model-based variants help address computational costs: DyNA-PPO trains surrogate models between experimental rounds, while Jain et al. (2022) combines GFlowNets with active learning to sample diverse high-fitness sequences proportional to reward, achieving enhanced diversity compared to standard RL baselines. Wang et al. (2025) introduced DRAKES for reward optimization in discrete diffusion models. Their approach enables direct reward backpropagation through diffusion trajectories via Gumbel-Softmax approximations when rewards are differentiable; for non-differentiable rewards, they resort to standard policy gradient methods (PPO) or reward-weighted maximum likelihood estimation. ProteinZero targets protein inverse folding models, employing online RL with policy gradients designed from the outset for non-differentiable scalar rewards from structure predictors (ESMFold, US-align) and stability oracles (Fast-ddG, FoldX). A key distinction lies in diversity handling: DRAKES reports sequence entropy as a post-hoc metric without explicit regularization, whereas ProteinZero incorporates embedding-level diversity regularization with theoretical guarantees (Appendix F) to actively prevent mode collapse during training. While both advance reward-guided protein design, they address complementary model classes: DRAKES for discrete diffusion, ProteinZero for protein inverse folding.

Classical methods often focus on specific objectives such as binding affinity, folding accuracy, or assembly geometry, learning the necessary biophysical constraints through exploration. In contrast, RLHF fine-tuning, enabled by the recent emergence of powerful pre-trained models, operates in a different problem setting: leveraging these foundation models that already encode extensive biophysical knowledge, we refine competent generators rather than training naive policies. This setting enables holistic multi-objective optimization for generalizable improvements, promotes sequence realism without hard-coded penalties, and achieves sample-efficient learning as every oracle query refines an already capable model rather than teaching basic constraints from scratch. The pre-trained foundation facilitates generation of biologically plausible candidates that satisfy RL objectives, fundamentally changing the optimization landscape compared to classical approaches that must discover these constraints through extensive exploration from scratch.

D.2 Mode Collapse in Online Reinforcement Learning

Mode collapse represents a critical failure mode in online reinforcement learning where policies converge to narrow output distributions despite diverse valid solutions existing. Kirk et al. (2024) demonstrate that RLHF significantly reduces output diversity compared to supervised fine-tuning, with models producing uniform responses across different inputs. Cui et al. (2025) reveal that policy entropy plummets early in training, causing exploration to vanish and performance to saturate. As models converge to limited outputs, policy distributions become highly peaked, creating a vicious cycle where reduced diversity leads to overconfidence, further limiting exploration. The standard KL penalty in PPO-style RLHF only partially alleviates this issue, as reverse KL is inherently mode-seeking, which favors single high-probability solutions.

Various mitigation strategies have emerged. Entropy regularization directly adds bonuses to maintain broader distributions: Shekhar et al. (2024) integrate self-entropy into preference optimization, while Wang et al. (2024) show forward KL and Jensen-Shannon divergences achieve better alignment-diversity trade-offs than reverse KL. Diversity-reinforced objectives explicitly incorporate variety into rewards, with Li et al. (2025) using semantic clustering as diversity bonuses to achieve simultaneous improvements in quality and novelty. Data mixing strategies like SimpleMix (Li & Khashabi, 2025) combine on-policy and off-policy data to prevent collapse by maintaining broader training distributions.

Our embedding-level diversity regularization represents a novel contribution. Unlike existing approaches operating on output probabilities or rewards, we directly encourage semantic diversity in latent representation space. By penalizing similarity between hidden states of generated sequences, our method captures meaningful variation beyond surface differences. This complements traditional regularization: KL maintains proximity to reference distributions, entropy encourages probabilistic exploration, while our embedding regularizer ensures exploration of functionally distinct sequence regions. For protein design with expensive oracles, maintaining diversity is critical to maximize information per query. Our approach enables covering more possibilities with fewer oracle calls, avoiding redundant evaluations. The combination provides robust protection against mode collapse while maintaining alignment with design objectives, as demonstrated by simultaneous improvements in diversity and performance metrics.

D.3 Relation to Fully Atomistic Generative Models

Recent sequence-structure co-generation models include fully atomistic generators (Chroma (Ingraham et al., 2023), Protpardelle (Chu et al., 2024), ProteinGenerator (Lisanza et al., 2023)) and backbone-level co-design methods (MultiFlow (Campbell et al., 2024)). These models learn joint distributions over three-dimensional backbone geometries and amino acid sequences for *de novo* fold sampling with compatible sequences. They combine continuous backbone representations (residue frames or atomic coordinates) with discrete or relaxed sequence representations; ProteinGenerator performs diffusion in continuous sequence space coupled to structure prediction networks for atomic coordinates.

Table 13: Comparison of diversity incorporation strategies. We report success rate, FoldX ddG (kcal/mol), and TM-score for proteins of different lengths.

Strategy	Success Rate (%)		FoldX ddG (kcal/mol)		TM-score	
	0–150	150–300	0–150	150–300	0–150	150–300
(1) Embedding reward	78.65	81.71	-18.681	-23.967	0.836	0.831
(2) Hamming reward	74.63	80.29	-11.135	-23.228	0.836	0.831
(3) Embedding regularization	90.13	91.19	-24.924	-32.805	0.867	0.867

ProteinZero addresses the complementary problem of backbone-conditioned inverse folding. In practical engineering workflows, enzyme optimization or epitope-specific binder design, backbone geometry is predetermined by experimental structures, docking simulations, or motif grafting and must be preserved as a hard constraint. The objective is identifying sequences maximizing stability and foldability for fixed geometries rather than generating novel backbone shapes. ProteinZero provides sequence refinement on fixed backbones where structural template preservation is essential.

A fundamental distinction lies in computational tractability for online RL. Applying online RL to joint sequence-structure generators entails repeated sampling in high-dimensional continuous coordinate space ($\mathbb{R}^{3\times N}$, often including side-chain atoms), with reward evaluation requiring expensive physics-based simulations or slow structural oracles for geometric validity. ProteinZero operates in discrete sequence space with efficient proxy rewards (Fast-ddG, ESMFold), demonstrating that multi-objective, online RL is tractable for sequence optimization. Our evaluation focuses on standard backbone-conditioned inverse folding benchmarks (CATH-4.3) rather than direct comparison with *de novo* atomistic generators. This isolates the online RL algorithm's contribution: fixed backbones ensure the observed 36–48% reduction in design failure rates is attributable to policy optimization and diversity regularization rather than backbone sampling or flexibility.

E Additional Results on Diversity Regularization Strategies

In the main text, we discuss the impact of incorporating diversity through different strategies. For completeness, Table 13 reports the detailed numerical results, including success rate, FoldX ddG, and TM-score for both protein length categories. These results further illustrate that embedding-based diversity applied as a regularizer preserves stability and structural accuracy, while reward-based variants lead to significant degradation in performance.

F Diversity Regularizer: Theoretical Foundation for Preventing Mode Collapse

We provide a theoretical analysis of our embedding-level diversity regularizer, demonstrating how it helps prevent mode collapse in online reinforcement learning. We formalize mode collapse for a conditional policy $p_\theta(y \mid x)$ as a sharp decrease in policy entropy $H_\theta(Y \mid X{=}x)$ and a contraction of its effective support. This perspective aligns with maximum-entropy RL, where entropy encourages stochasticity and prevents brittle policies (Haarnoja et al., 2018; Levine, 2018; Geist et al., 2019). The standard KL-regularized objective, $\mathbb{E}[r] - \alpha_{\mathrm{KL}}\mathrm{KL}(p\|p_{\mathrm{ref}})$, yields the Boltzmann distribution $p^*(y \mid x) \propto p_{\mathrm{ref}}(y \mid x)\exp(r(x,y)/\alpha_{\mathrm{KL}})$. A small α_{KL} or highly peaked rewards can drive concentration and an entropy drop, a known mode-seeking behavior (Todorov, 2006; Levine, 2018).

F.1 Mean-Field Objective and Properties

Let $Z = \psi_\theta(X,Y) \in \mathbb{S}^{d-1}$ denote the unit-norm embeddings of generated sequences, as constructed in Section 3.1.1. For a fixed input x, we simplify notation by considering probability measures $p(\cdot) \equiv p_\theta(\cdot \mid x)$ on sequences y. We define a symmetric kernel $c(y,y') = \cos(\psi_\theta(x,y), \psi_\theta(x,y'))$.

Assumption 1 (Absolute continuity and i.i.d. pairing)**.** *For each x, the feasible set is $\{p \in \Delta : p(\cdot \mid x) \ll p_{\text{ref}}(\cdot \mid x)\}$, ensuring $\text{KL}(p \| p_{\text{ref}})$ is finite. Expectations over pairs (y, y') are taken w.r.t. the product measure $p(\cdot \mid x) \otimes p(\cdot \mid x)$ (i.i.d. draws).*

Remark 1 (Setting and scope of analysis)**.** *All variational arguments below fix θ and the conditioning input x, and treat $p(\cdot) \equiv p_\theta(\cdot \mid x)$ as the optimization variable. We work with discrete sequence policies, so $p_{\text{ref}}(y \mid x) > 0$ on the feasible support, making atomic distributions $\delta_{y^\star}$ admissible whenever $p_{\text{ref}}(y^\star \mid x) > 0$.*

Coefficient sign convention. Throughout we assume nonnegative coefficients, in particular $\alpha_{\text{div}} \geq 0$ (and $\alpha_{\text{KL}} \geq 0$), so that the diversity term acts as a repulsive regularizer.

At the population level, we analyze the regularized functional for any fixed x:

$$\max_{p \in \Delta:\, p \ll p_{\text{ref}}(\cdot|x)} \quad \mathcal{J}[p] \;:=\; \mathbb{E}_{y \sim p}[r(x,y)] - \alpha_{\text{KL}}\, \text{KL}\big(p \,\|\, p_{\text{ref}}(\cdot \mid x)\big) - \frac{\alpha_{\text{div}}}{2}\, \mathbb{E}_{y,y' \sim p}\big[c(y,y')\big]. \quad (8)$$

Remark 2. *Writing the diversity term as $+\frac{\alpha_{\text{div}}}{2}\big(1 - \mathbb{E}[c]\big)$ is equivalent up to an additive constant and yields the same optimizer.*

Lemma 1 (Diversity as a penalty on the embedding mean)**.** *Under Assumption 1, with $Z = \psi_\theta(X, Y)$ on the unit sphere, the diversity term is the squared norm of the mean embedding: $\mathbb{E}_{y,y' \sim p}\big[c(y,y')\big] = \big\|\mathbb{E}_{y \sim p}[Z]\big\|_2^2$. Consequently, the objective*

$$\mathcal{J}[p] = \mathbb{E}_p[r] - \alpha_{\text{KL}}\text{KL}(p \| p_{\text{ref}}) - \frac{\alpha_{\text{div}}}{2}\, \|\mathbb{E}_p[Z]\|_2^2$$

is concave in p. It is strictly concave on the relative interior if $\alpha_{\text{KL}} > 0$.

Proposition 1 (Interior fixed point with a non-local repulsive potential)**.** *Assume $\alpha_{\text{KL}} > 0$. Any interior stationary point p^* of Eq. 8 (where $p^*(y) > 0$ for all feasible y) satisfies*

$$p^*(y \mid x) \;\propto\; p_{\text{ref}}(y \mid x)\, \exp\!\left(\frac{1}{\alpha_{\text{KL}}} r(x,y) - \frac{\alpha_{\text{div}}}{\alpha_{\text{KL}}}\, \Phi_\theta\big(y; p^*\big)\right), \quad (9)$$

where the potential $\Phi_\theta(y;p) := \mathbb{E}_{y' \sim p}\big[c(y,y')\big]$ is a ***non-local repulsive term****. Placing mass on a sequence y increases the "energy" of other sequences y' with similar embeddings, discouraging collapse.*

F.2 Guarantees Against Collapse to a Single Mode

With $\alpha_{\text{KL}} > 0$, the KL term alone rules out collapse to a point mass (delta distribution). The diversity term adds a non-local repulsion that discourages uni-directional concentration in representation space.

Theorem 1 (KL barrier to deterministic collapse)**.** *Suppose $\alpha_{\text{KL}} > 0$ and Assumption 1 holds. Let $y^\star$ be any sequence with $p_{\text{ref}}(y^\star \mid x) > 0$. If another sequence $y' \neq y^\star$ exists with $p_{\text{ref}}(y' \mid x) > 0$, then the point mass $p = \delta_{y^\star}$ is* not *a stationary point of Eq. 8.*

Proof. Consider a perturbation $p_\varepsilon = (1-\varepsilon)\delta_{y^\star} + \varepsilon\delta_{y'}$ for a small $\varepsilon > 0$. The change in the reward term is $\Delta\mathcal{J}_{\text{reward}} = \varepsilon\left(r(x,y') - r(x,y^\star)\right) + O(\varepsilon^2)$. For the diversity term, let $c = c(y^\star, y')$. The change is $\Delta\mathcal{J}_{\text{div}} = \alpha_{\text{div}}\varepsilon(1-c) + O(\varepsilon^2)$. For the KL divergence, the change is

$$\begin{aligned}\Delta\text{KL} &:= \text{KL}(p_\varepsilon \| p_{\text{ref}}) - \text{KL}(\delta_{y^\star} \| p_{\text{ref}}) \\ &= (1-\varepsilon)\log(1-\varepsilon) \;+\; \varepsilon\log\varepsilon \;+\; \varepsilon\log\frac{p_{\text{ref}}(y^\star \mid x)}{p_{\text{ref}}(y' \mid x)}.\end{aligned}$$

Using $(1-\varepsilon)\log(1-\varepsilon) = -\varepsilon + O(\varepsilon^2)$, we find that $-\alpha_{\text{KL}}\,\Delta\text{KL}$ is dominated by the term $-\alpha_{\text{KL}}\varepsilon\log\varepsilon$. Combining these, the directional derivative of the full objective is:

$$\begin{aligned}\lim_{\varepsilon \to 0^+} \frac{\mathcal{J}[p_\varepsilon] - \mathcal{J}[\delta_{y^\star}]}{\varepsilon} &= \underbrace{r(x,y') - r(x,y^\star)}_{\text{reward}} + \underbrace{\alpha_{\text{div}}(1-c)}_{\text{diversity}} \\ &\quad + \underbrace{\alpha_{\text{KL}}\Big(1 - \log\varepsilon - \log\tfrac{p_{\text{ref}}(y^\star|x)}{p_{\text{ref}}(y'|x)}\Big)}_{\text{KL barrier}} + o(1).\end{aligned}$$

As $\varepsilon \to 0^+$, the $-\log\varepsilon$ term drives the quotient to $+\infty$. Moving probability mass away from any single point mass $\delta_{y^\star}$ thus always increases the objective, meaning $\delta_{y^\star}$ cannot be a stationary point. □

Proposition 2 (No-KL case: finite condition that rules out a delta optimum)**.** *If $\alpha_{\mathrm{KL}} = 0$ and there exists $y' \neq y^\star$ such that*

$$r(x, y') - r(x, y^\star) \; + \; \alpha_{\mathrm{div}}\big(1 - c(y^\star, y')\big) \; > \; 0,$$

then $p = \delta_{y^\star}$ is not a (local) maximizer of Eq. 8.

Corollary 1 (Readable sufficient condition)**.** *For any $y^\star, y'$ with $p_{\mathrm{ref}}(y^\star \mid x), p_{\mathrm{ref}}(y' \mid x) > 0$:*

- *If $\alpha_{\mathrm{KL}} > 0$, then $p(\cdot \mid x) = \delta_{y^\star}$ is never stationary (Theorem 1).*
- *If $\alpha_{\mathrm{KL}} = 0$, a sufficient condition for non-stationarity is $r(x, y^\star) - r(x, y') \; < \; \alpha_{\mathrm{div}}\big(1 - c(y^\star, y')\big)$, which is the finite, reward–diversity tradeoff stated in Proposition 2.*

Remark 3 (Scope of the diversity term)**.** *Since $\mathbb{E}_{y,y'}[c(y, y')] = \|\mathbb{E}[Z]\|^2$, the regularizer mainly discourages* uni-directional *concentration (single-mode collapse aligned with one embedding direction). It may not penalize symmetric few-mode collapse where $\mathbb{E}[Z] \approx 0$.*

F.3 Entropy Lower Bound and Implementation

The diversity regularizer also yields a conservative lower bound on policy entropy. Let $Z = \psi_\theta(X, Y) \in \mathbb{S}^{d-1}$ and define the cosine kernel $k(z, z') = (1 + \cos(z, z'))/2 \in [0, 1]$. The *information potential* of the embedding distribution $\nu_\theta(\cdot \mid x)$ is

$$\begin{aligned} I_k(Z \mid X{=}x) &:= \mathbb{E}\big[k(Z, Z') \mid X{=}x\big] = 1 - \tfrac{1}{2}\,\bar{D}_{\cos}(\theta; x), \\ \bar{D}_{\cos}(\theta; x) &:= 1 - \mathbb{E}[\cos(Z, Z') \mid X{=}x]. \end{aligned}$$

Since $H(Y \mid X) \geq H_2(Y \mid X) \geq H_2(Z \mid X)$ and $H_2(Z \mid X) \geq -\log I_k(Z \mid X)$, we obtain the lower bound on policy entropy and perplexity:

$$H_\theta(Y \mid X{=}x) \; \geq \; -\log\big(1 - \tfrac{1}{2}\bar{D}_{\cos}(\theta; x)\big), \qquad \mathrm{Perp}_\theta(x) \; \geq \; \frac{1}{1 - \frac{1}{2}\bar{D}_{\cos}(\theta; x)}. \tag{10}$$

By Lemma 1, $\mathbb{E}[\cos(Z, Z')] = \|\mathbb{E}[Z]\|_2^2 \in [0, 1]$, hence $I_k = \frac{1}{2}\big(1 + \|\mathbb{E}[Z]\|_2^2\big) \in [1/2, 1]$. Thus the bound is conservative and cannot exceed $\log 2$ (equivalently, the perplexity lower bound is at most 2). It should be viewed as a safety valve rather than a strong guarantee.

Remark 4 (Mini-batch estimator)**.** *In practice we estimate $\bar{D}_{\cos}$ using off-diagonal pairs to avoid upward bias:*

$$\widehat{\mathbb{E}[\cos]} \; = \; \frac{1}{m(m-1)} \sum_{i \neq j} \cos(z_i, z_j) = \frac{m\|\bar{z}\|_2^2 - 1}{m - 1} \; \in \; \Big[-\frac{1}{m-1},\, 1\Big],$$

$$\widehat{\bar{D}}_{\cos} \; = \; 1 - \widehat{\mathbb{E}[\cos]} \; \in \; \Big[0,\, 1 + \frac{1}{m-1}\Big].$$

When $m \geq 3$, $1 - \widehat{\bar{D}}_{\cos}/2 > 0$ holds automatically; for $m = 2$, a tiny truncation can be applied before evaluating $-\log(1 - \widehat{\bar{D}}_{\cos}/2)$.

The objective in Eq. 8 is implemented in our ProteinZero$_{\mathrm{RAFT}}$ and ProteinZero$_{\mathrm{GRPO}}$ algorithms by appending the diversity loss term, which induces the repulsive fixed point from Eq. 9 and benefits from the entropy guarantees of Eq. 10.

G Use of LLMs

We acknowledge the use of Large Language Models (LLMs) to assist in the preparation of this manuscript. LLMs were employed exclusively for language polishing to improve clarity, grammar, and consistency of technical writing. All scientific content, experimental design, methodology, data analysis, and core insights represent original work by the authors. LLMs did not contribute to the conceptualization, experimentation, or interpretation of results. All factual claims, mathematical derivations, and experimental outcomes were independently generated and verified by the authors. The use of LLMs was strictly limited to improving the presentation and readability of our independently developed research, serving only as writing aids rather than contributing to the scientific content itself.