Supplementary Information for:

Guiding Generative Protein Language Models with Reinforcement Learning

Filippo Stocco[1,2], Maria Artigues-Lleixà[1,2], Andrea Hunklinger[2,3], Talal Widatalla[4,5], Marc Güell[1,6], Noelia Ferruz[2,1,*]

[1]Department of Medicine and Life Sciences, Universitat Pompeu Fabra, Barcelona, Spain
[2]Centre for Genomic Regulation, the Barcelona Institute of Science and Technology, Dr Aiguader 88, Barcelona 08003, Spain
[3]Universitat de Barcelona, Facultat de Farmàcia i Ciències de l'Alimentació, Avda. Diagonal 643, Barcelona 08028, Spain
[4]Stanford University,
[5]Arc Institute
[6]ICREA, Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain
*E-mail: noelia.ferruz@crg.eu

Table of Contents:

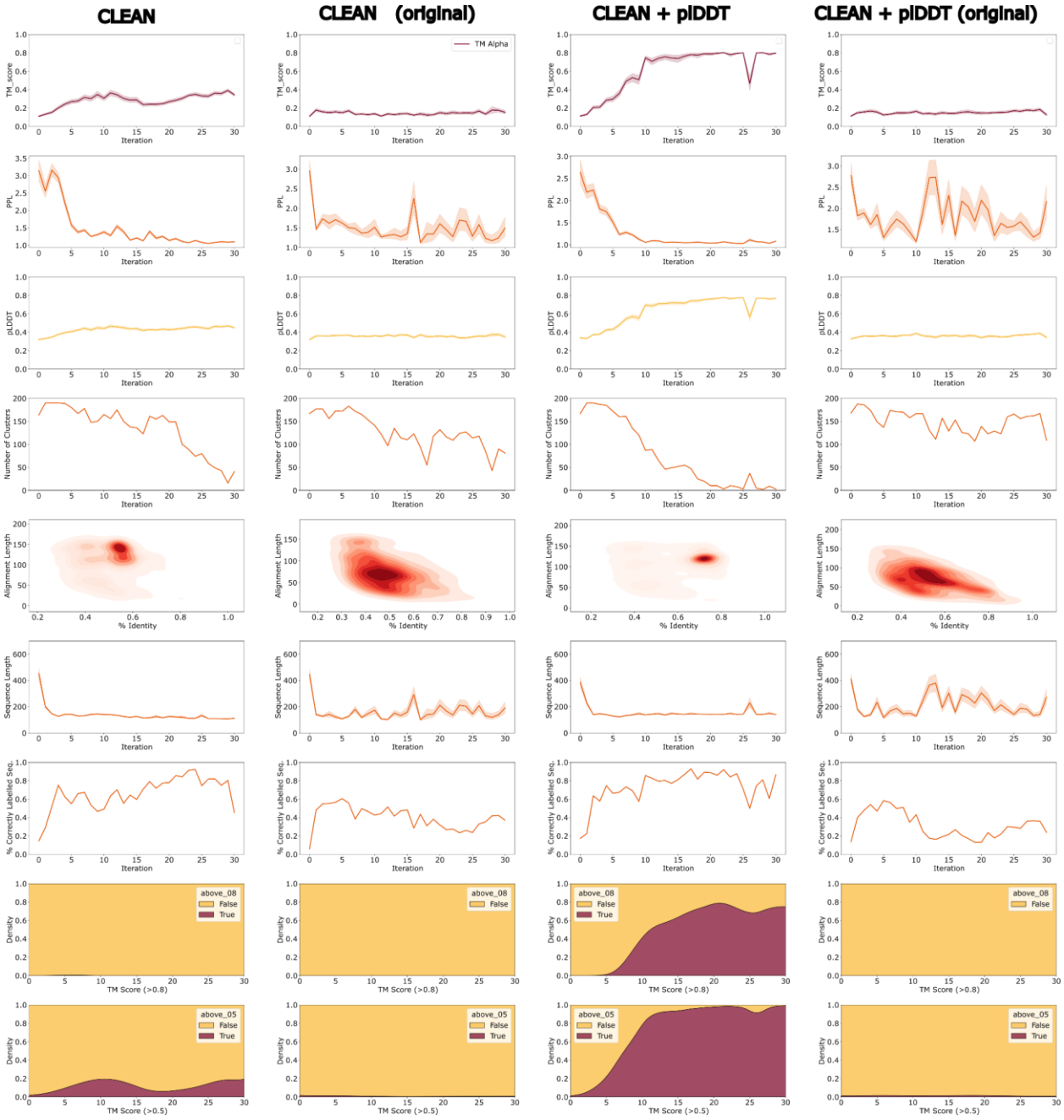

Figure S1: Comparison of different DPO implementations for the optimization of CLEAN-labeled sequences. Runs defined as original correspond to equation (5), otherwise they were obtained with equation (6).

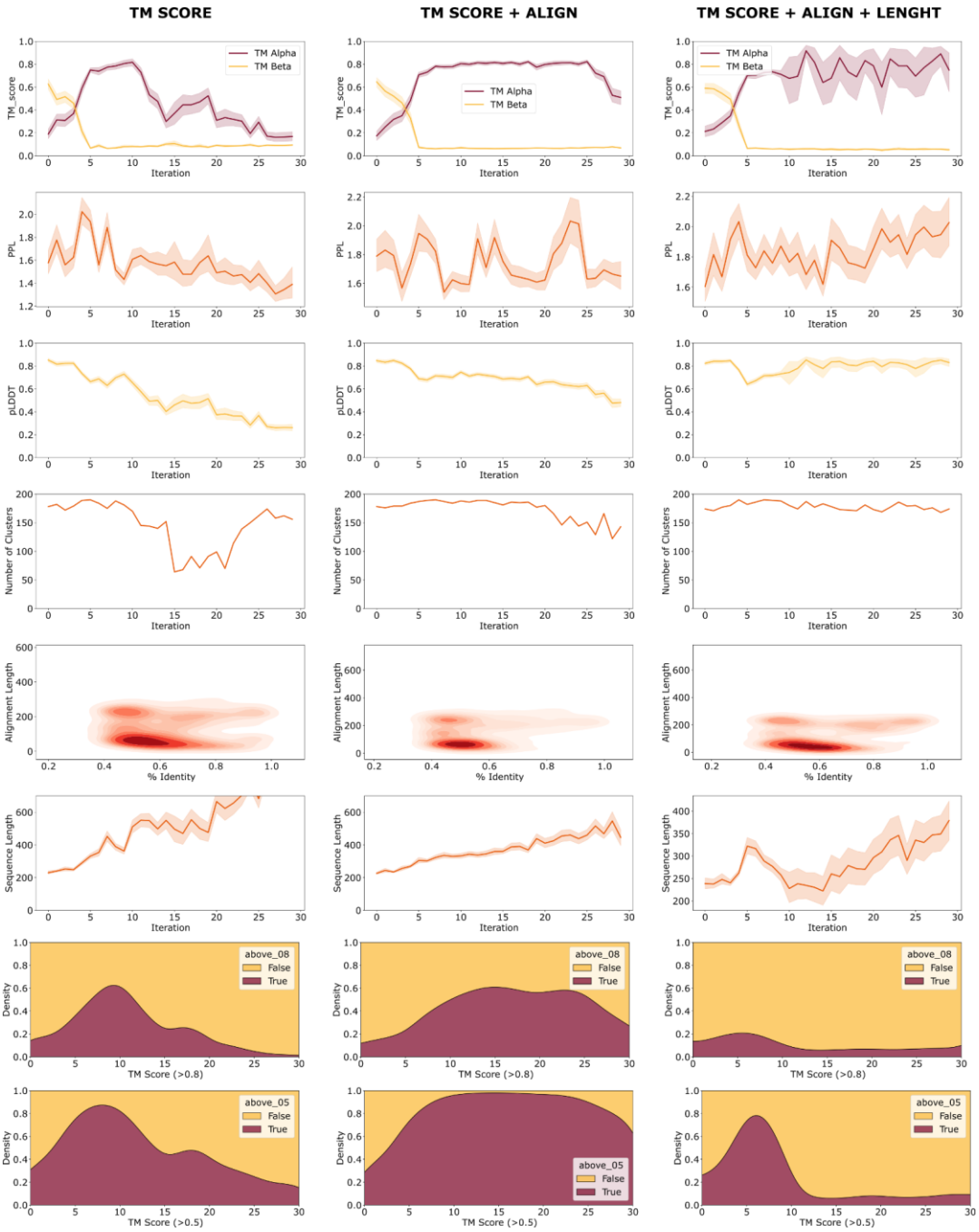

Figure S2: DPO-guided topology generation using TM-score as an oracle. We evaluated the framework at increasing the proportion of generated α CA per iteration, using TM-score as the reward function alone (left), TM-score and alignment length as a composite function (central), and TM-score, a Gaussian equation of length and alignment length (right). We record several properties, namely, in row order: TM-score against α CA pdb 2VVB, perplexity (PPL) of the model, ESMFold plDDT, number of clusters at 80% as computed with MMseqs2, similarity to the training set, sequence length, density of proteins showing TM-scores over 0.8, and over 0.5.

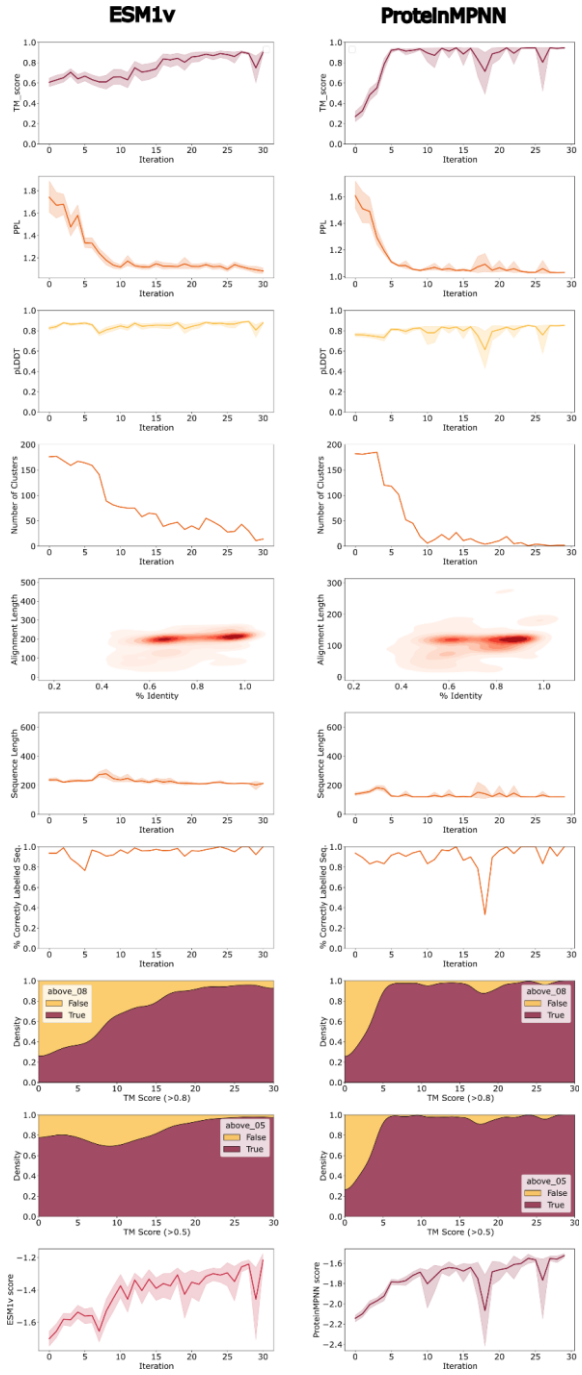

Figure S3: DPO-guided numerical generation using ESM-1v and ProteinMPNN as an oracle. We evaluated the framework at increasing both statistical scores per iteration, using ESM-1v (right) ans ProteinMPNN (left) as the reward function alone. We record several properties, namely, in row order: TM-score against α CA PDB 2VVB, perplexity (PPL) of the model, ESMFold plDDT, number of clusters at 80% as computed with MMseqs2, similarity to the training set, sequence length, density of proteins showing TM-scores over 0.8 and over 0.5 and ESM-1v and ProteinMPNN scores.

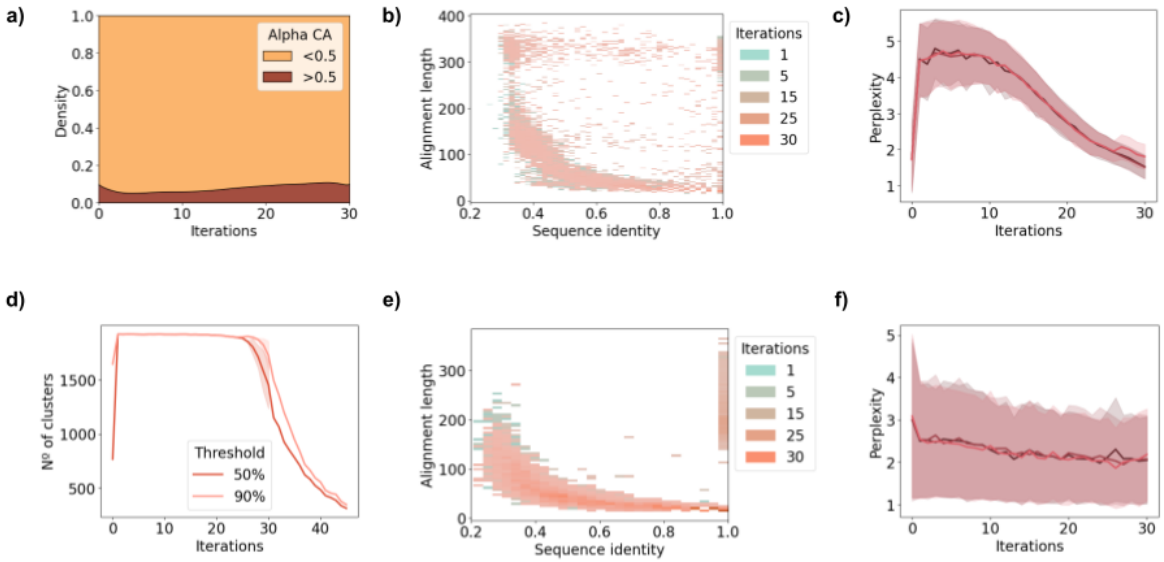

Figure S4: s-FT-guided topology and enzyme functional annotation generation. (a) Proportion of generated sequences with more or less than 0.5 TM-score when superimposed with PDB 2VVB with TM-score as a reward function. **(b)** Distribution of sequence similarity of synthetic sequences generated against the 200 sequences set used to finetune the model the previous iteration for the TM-score reward function. **(c)** Perplexity of the generated sequences with the TM-score reward function across 30 iterations. **(d)** Progression of the number of clusters of sequences with TM-score reward function until 45 iterations. **(e)** Distribution of sequence similarity of synthetic sequences generated against the 200 sequences set used to finetune the model the previous iteration got the CLEAN reward function. **(f)** Perplexity of the generated sequences with the CLEAN reward function across 30 iterations.

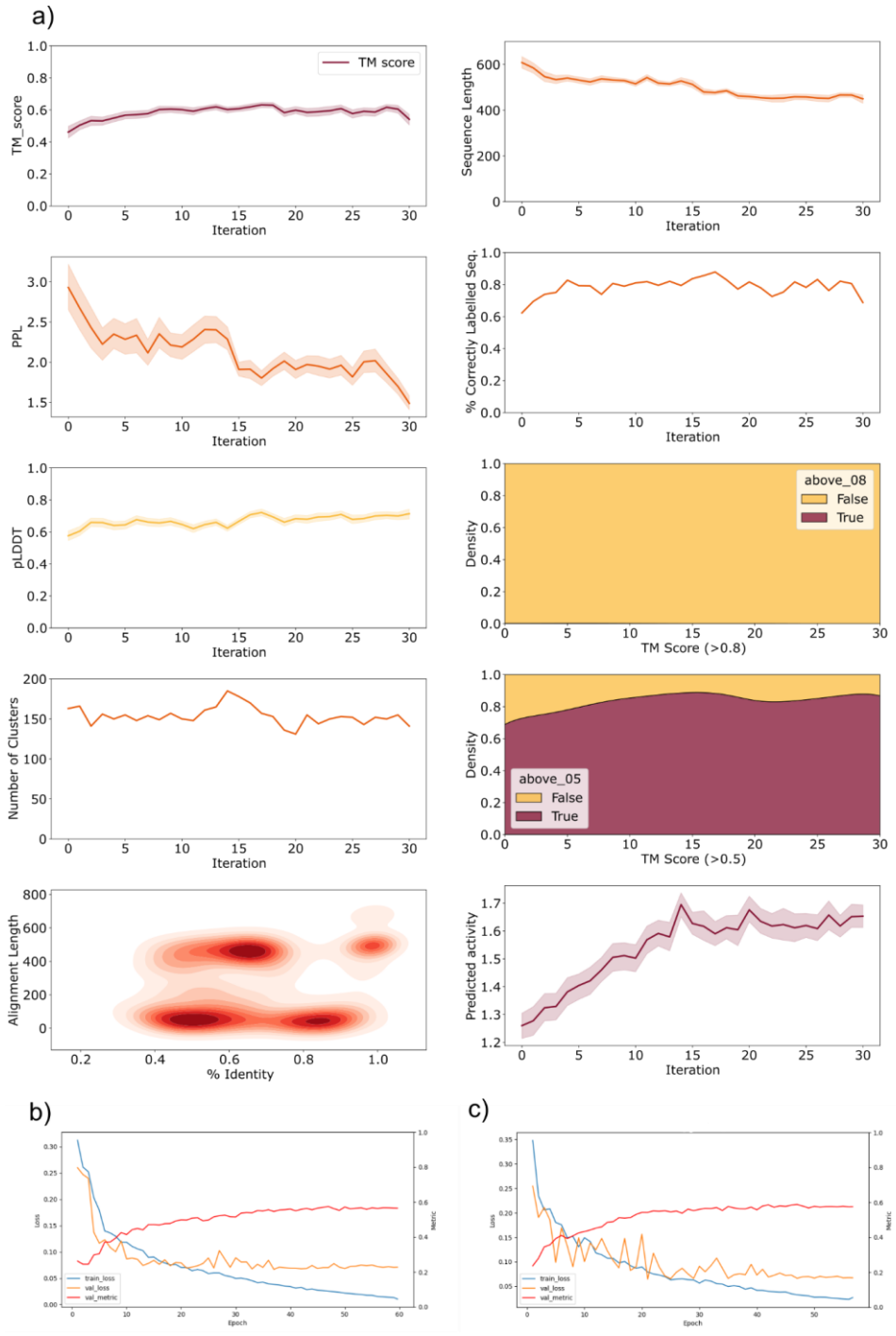

Figure S5: DPO Optimized for Activity Prediction a) Progression of key metrics during the reinforcement campaign, highlighting the maximization of activity as the objective. Notably, no substantial decline is observed in metrics such as TM score, perplexity (ppl), or pLDDT. **(b), (c)** Training curves for regression models predicting activity from input sequences. Models were constructed using ESM2 or ESM1v architectures and fine-tuned with LoRA for a single-label output. In both cases, the Spearman correlation coefficient (red line) achieved a value of 0.6.

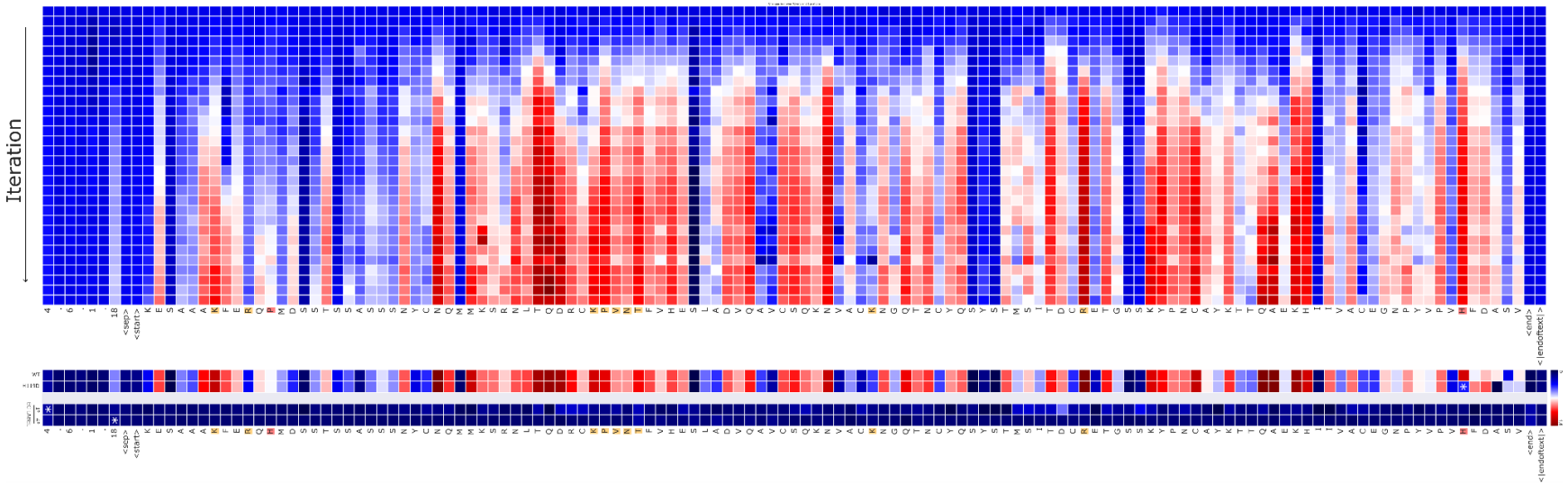

Figure S6: DPO can extract relevant features at the token level. Wild-type sequence for Pancreatic Ribonuclease (EC 4.6.1.18, UniProt: P67926). At the top variation of the reward over the iterations over time for each amino acid. The model identifies key features with higher reward (red color). At the bottom, experiments analyzing single-token variations and their effects compared to the wild-type (WT). Point variation (white asterisks) induce localized changes in the reward, whereas alterations of tokens at the EC label result in a complete drop in the reward across the sequence to zero.

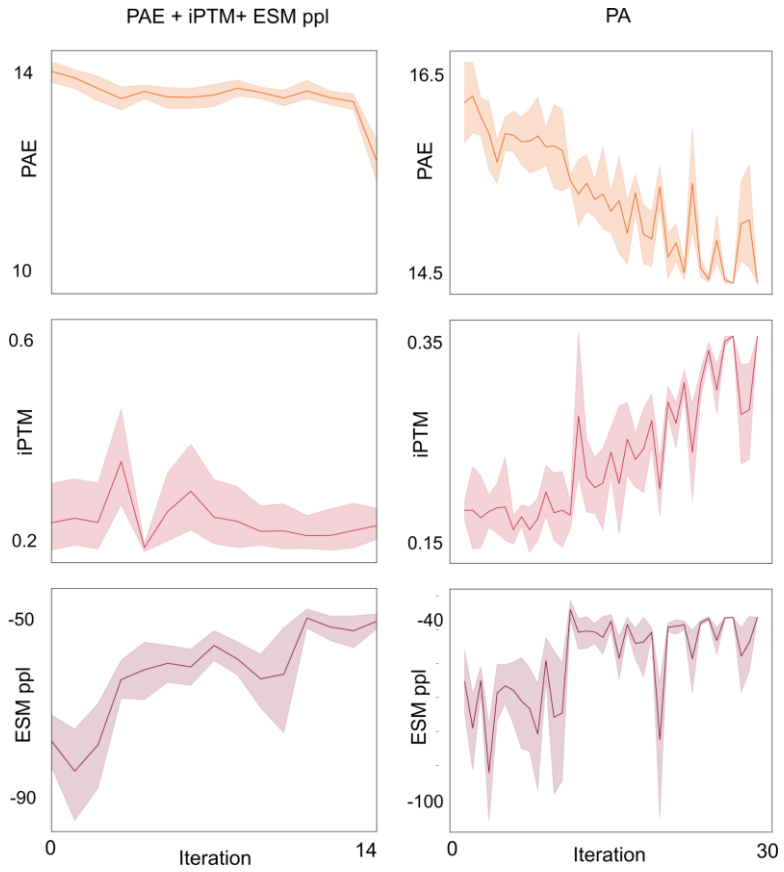

Figure S7: Reinforcement Learning campaign for binders design. The first round was approached as a multi-objective optimization, utilizing ESM perplexity (ppl), Predicted Aligned Error (PAE), and inter-protein TM-score (iPTM) as rewards. In the second round, only PAE was employed to further refine the fine-tuned ZymCTRL model. Binders submitted to the AdaptiveBio competition were sourced from iterations 0, 3, and 12 of the first round.

EC number	Objective	Scoring function	Experiments
4.2.1.1	From beta to alpha CA	(TM_norm_query + align/100)*length_g where: **TM_norm_query** is the TM score normalized for the query, **align** is the length of the alignment for the corresponding TM score, and **length_g** equals the gaussian of the ratio between the WT sequence and the studied sequence centered on 1, as follows: **length_g** = math.exp(-((((len(sequence)/len(ref_seq))-1)**2)/(0.5**2)))	DPO, s-FT
4.6.1.18	From 10% to 100%	torch.nn.CosineSimilarity(ref_clean_emb,target_clean_emb)*length_g	DPO, s-FT
4.6.1.18	From 10% to 100% (with increased pLDDT)	torch.nn.CosineSimilarity(ref_emb,target_emb)*length_g*plddt	DPO
2.7.11.5	From 0% to 100%	torch.nn.CosineSimilarity(ref_emb,target_emb)*length_g	DPO
5.4.99.5	Increase ProteinMPNN	ProteinMPNN score	DPO
4.2.1.1	Increase ESM1v	ESM1v score	DPO
1.3.3.18	Increase PAE, ESM PPL, iPTM	(-pae+(iptm*10)+(esm_ppl/10))*(length_g) or ((-pae)+(iptm*10)+(-esm_ppl/100))*(length_g)	DPO

3.2.1.1	Increase activity (SAPI)	cosine_similarity(seq_emb, reference_emb)*plddt*(length_g)	DPO

Table S1: Different scoring functions used in this study.

Hyperparameters (DPO)	
β	0.01
Seed number	1998
Learning rate	1×10-7
Batch size	5
epochs	5
Train/Test split	0.2
Adams	(0.9, 0.98)
ϵ	1× 10-8
Adam decay	0.1

Table S2: Hyperparemeters used for the training of DPO_pLM unless otherwise specified in the text.