Supplementary Information

Supplementary Algorithm 1

Algorithm 1 Preference Dataset Generation

Input: Protein Generation Model π, Prompt x,

1: Generate responses set S_0 using π

2: Virtual screening to obtain wet lab test set S_1 and remaining set S_2

3: Obtain the preferred response set S_1^w as the positive feedback sequences from wet lab tests

4: Obtain the unpreferred response set S_2^l as the low-loss low-pLDDT sequences in S_2

5: Construct the datasets as $D = \{(x, y^w, y^l) | y^w \sim S_1^w, y^l \sim S_2^l\}$

Output: D

Supplementary Algorithm 2

Algorithm 2 Web Lab Data In Loop Iterative Optimization

Input: Protein Generation Model π_θ, Prompt x, Number of Iteration T

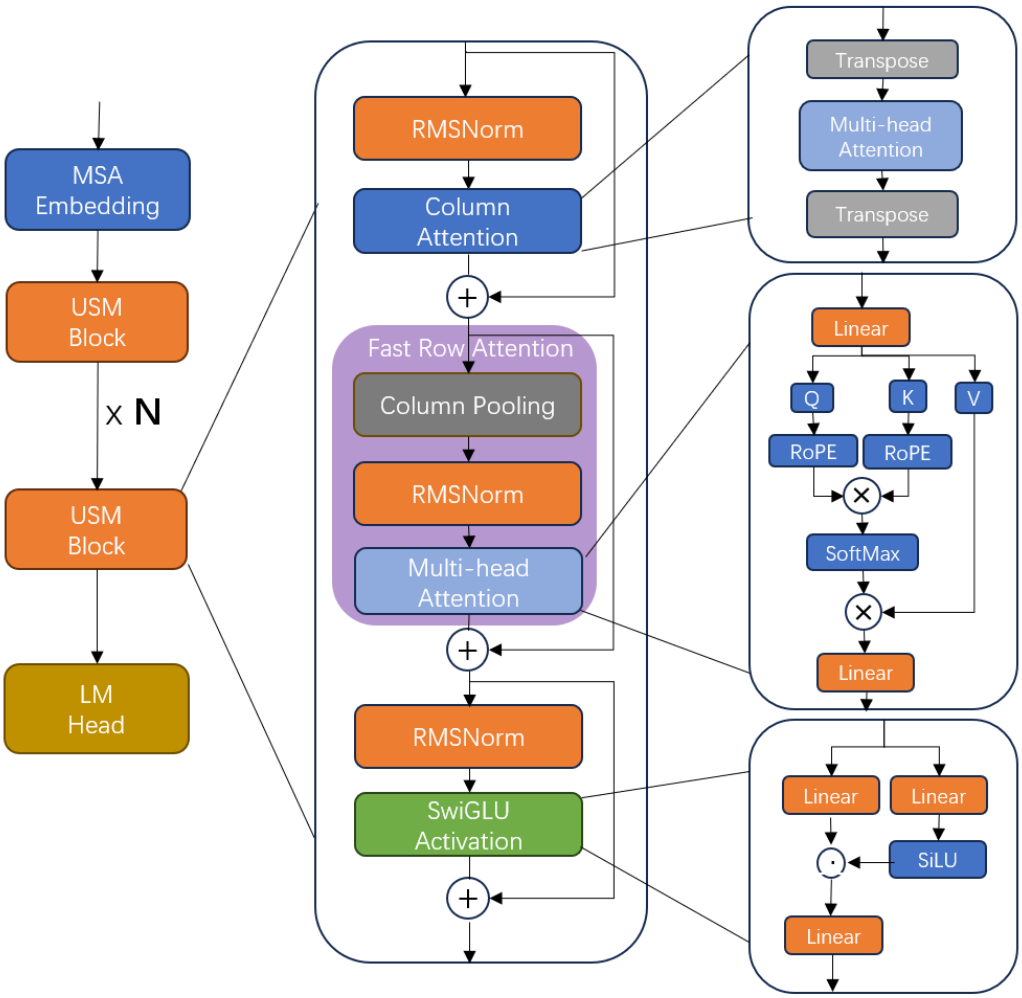
1: **for** t=1, 2, ..., T **do**

2: Assign reference model: $\pi_{\text{ref}} = \pi_\theta$

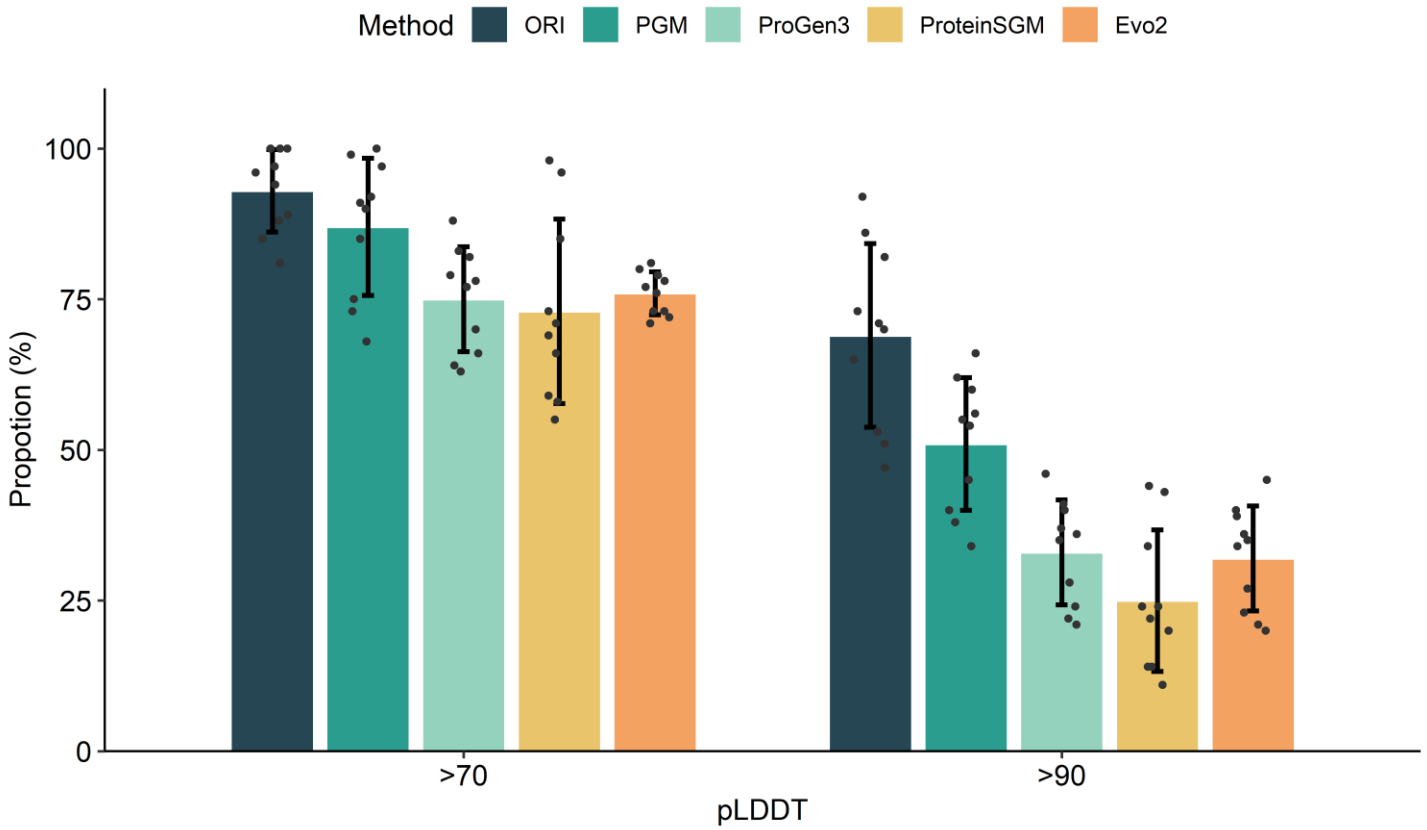
3: Obtain preference dataset D = Preference Dataset Generation (π_{ref}, x)

4: Update the model π_θ by minimizing $L_{\text{R-DPO}}$

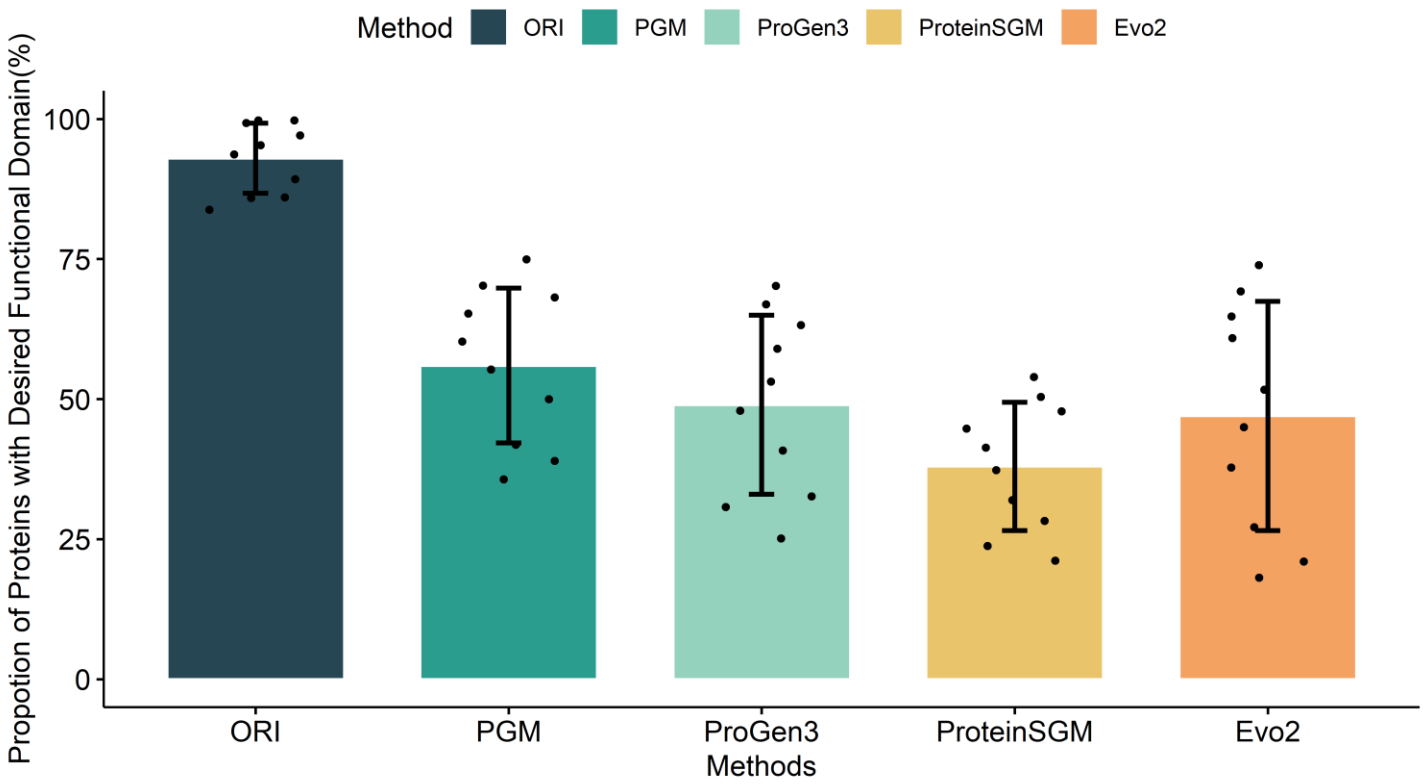
5: **end for Output:** π_θ



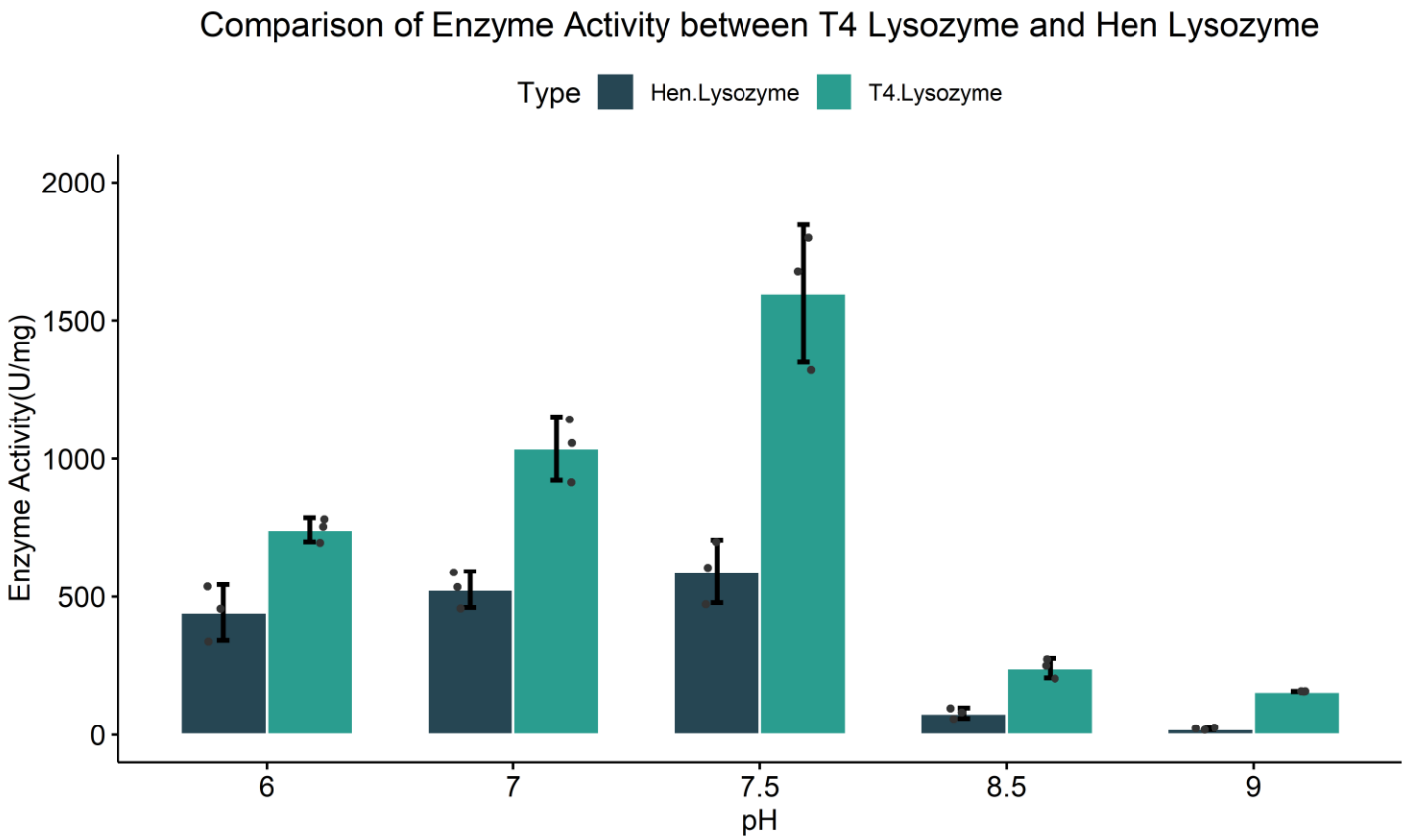
Supplementary Figure 1. The architecture of unified sequence model (USM). The model takes a multiple sequence alignment (MSA) as input. The first layer of the model is the MSA Embedding layer, which embeds the discrete MSA Tokens. The core of the USM model consists of multiple USM Transformer Blocks, which extract features from the MSA. The final output is generated by the language model (LM) head layer. The USM Block consists of three major parts: Column Attention, Fast Row Attention, and SwiGLU Activation. Due to the redundancy along the Column direction in MSA data, the MSA features h also exhibit sparsity along the Column direction. Based on this assumption, we proposed a Fast Row Attention (FRA) method. By applying Average Pooling along the Column direction, the MSA features h can be reduced to $h\_row \in R^{B\times row}$. This operation speeds up the subsequent Multi-Attention by a factor of col, and the computational complexity of Fast Row Attention is $O(hdL^2)$. Thanks to Fast Row Attention, the USM model achieves significantly improved speed and can handle larger parameter sizes. SwiGLU is a type of MLP that consists of Linear and SiLU activation functions. In our approach, Multi-head Attention is combined with Rotary Position Embedding (ROPE) to further enhance the model's ability to extrapolate long sequences.



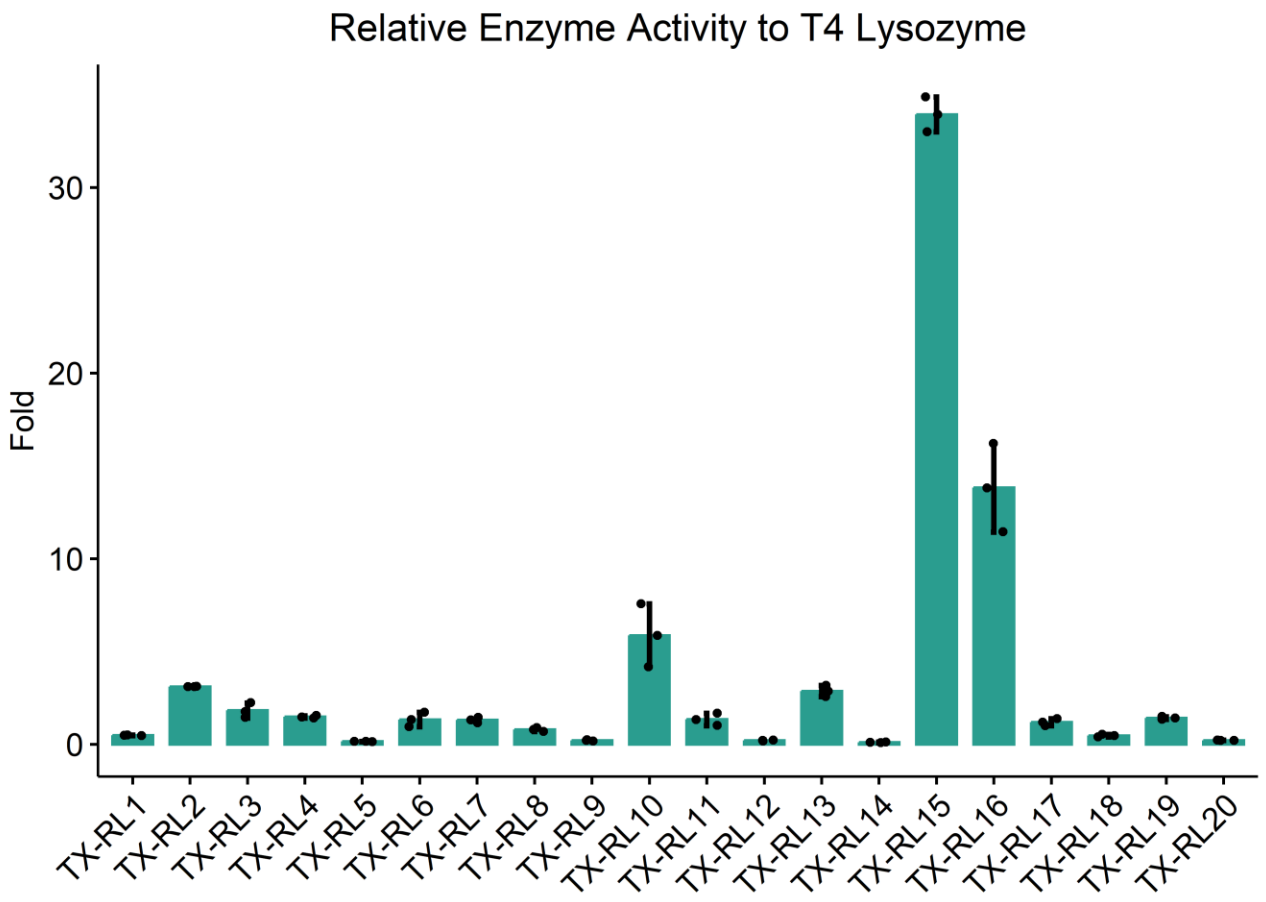
Supplementary Figure 2. Comparison of ORI with PGM, ProGen3, ProteinSGM, and Evo2 in conditioned protein generation. For each method, 1,000 proteins are generated. The y-axis shows the proportion of generated proteins with pLDDT scores above 70 and 90, respectively. The conditioning prompts used for generation include protein descriptions, structural information, and sequences. Although the prompt formats differ across methods, they convey equivalent semantic information. Data are presented as mean values ± SD from ten independent experiments (n = 10), and individual data points are overlaid as jittered dots. Source data are provided as a Source Data file.



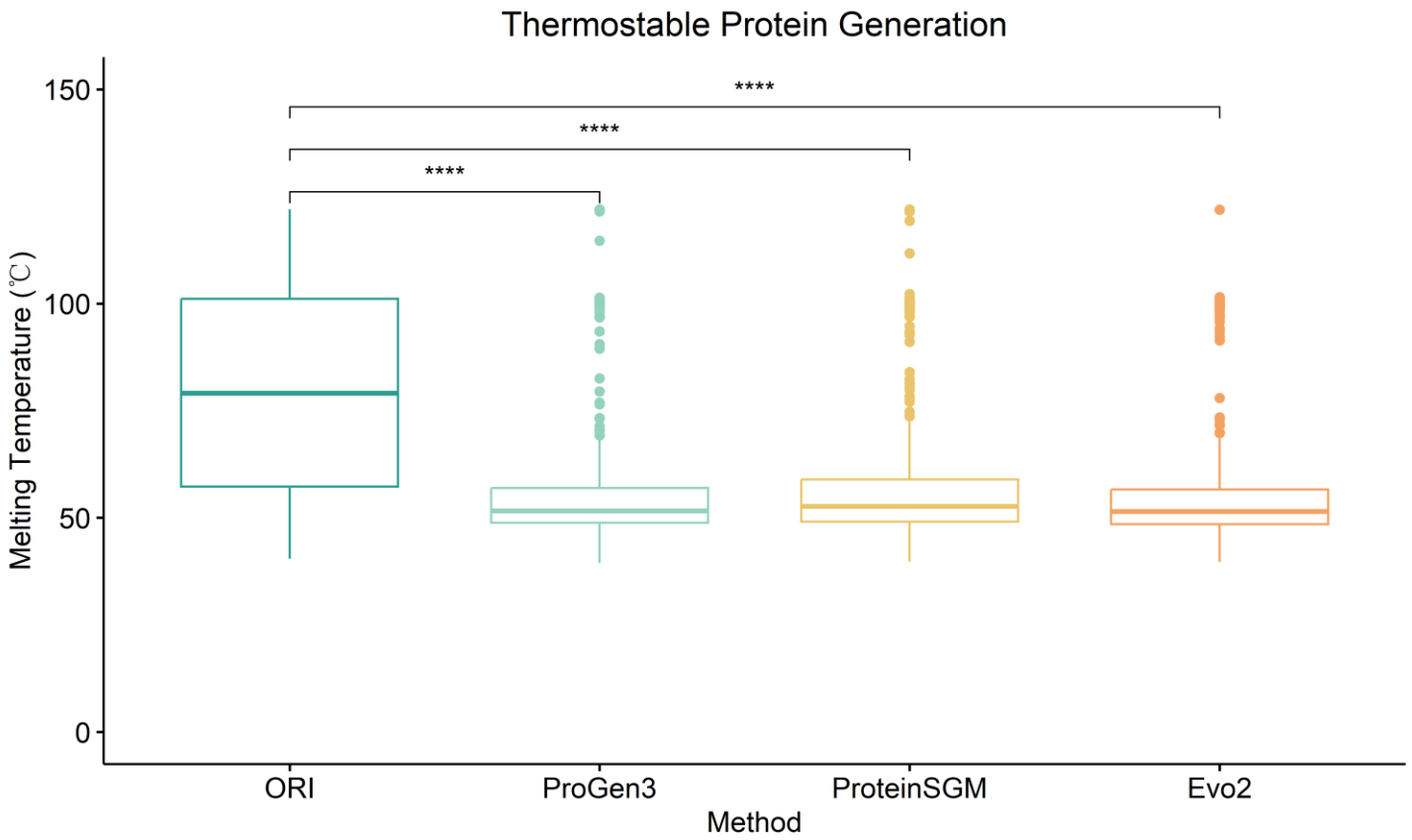
Supplementary Figure 3. Comparison of ORI with PGM, ProGen3, ProteinSGM, and Evo2 in generating functional proteins. The y-axis shows the proportion of generated proteins with the desired functional domain. Conditioning prompts for generation were provided as descriptions, structure-based, or sequence-based cues for specific protein functions, depending on the requirements of each method. Although the prompt formats differ across methods, they convey equivalent semantic information. Data are presented as mean values ± SD from ten independent experiments (n = 10), and individual data points are overlaid as jittered dots. Source data are provided as a Source Data file.



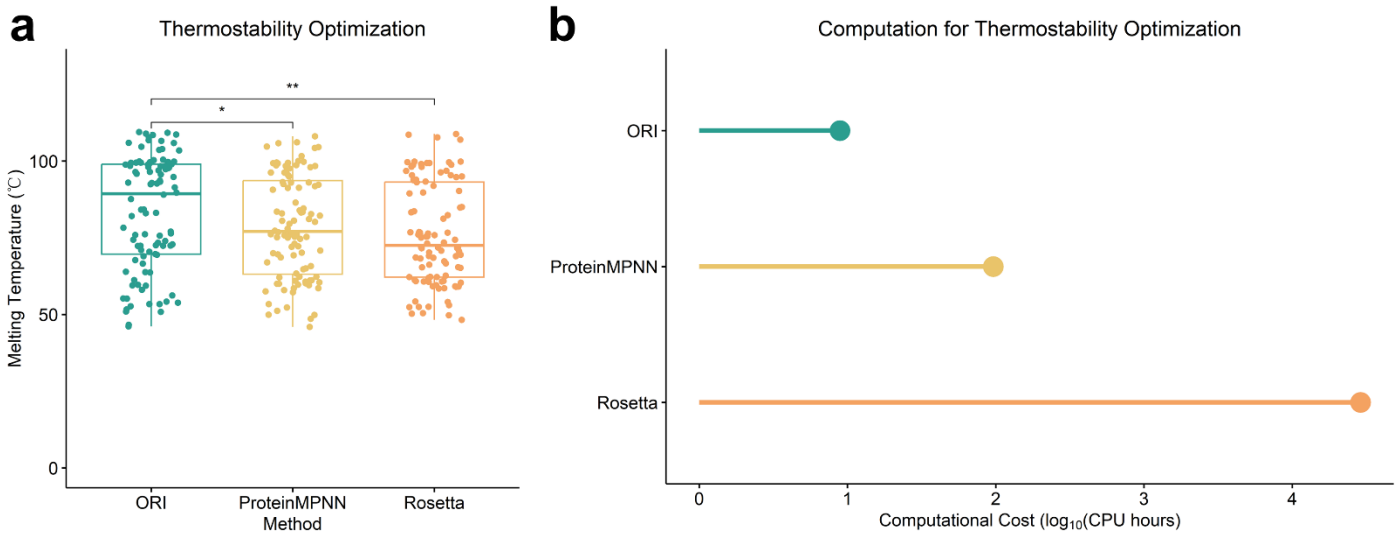
Supplementary Figure 4. Comparison of the assayed enzyme activity of T4 lysozyme and hen egg-white lysozyme across different pH levels. Data are presented as mean values ± SD from three independent experiments (n = 3), and individual data points are overlaid as jittered dots. Source data are provided as a Source Data file.



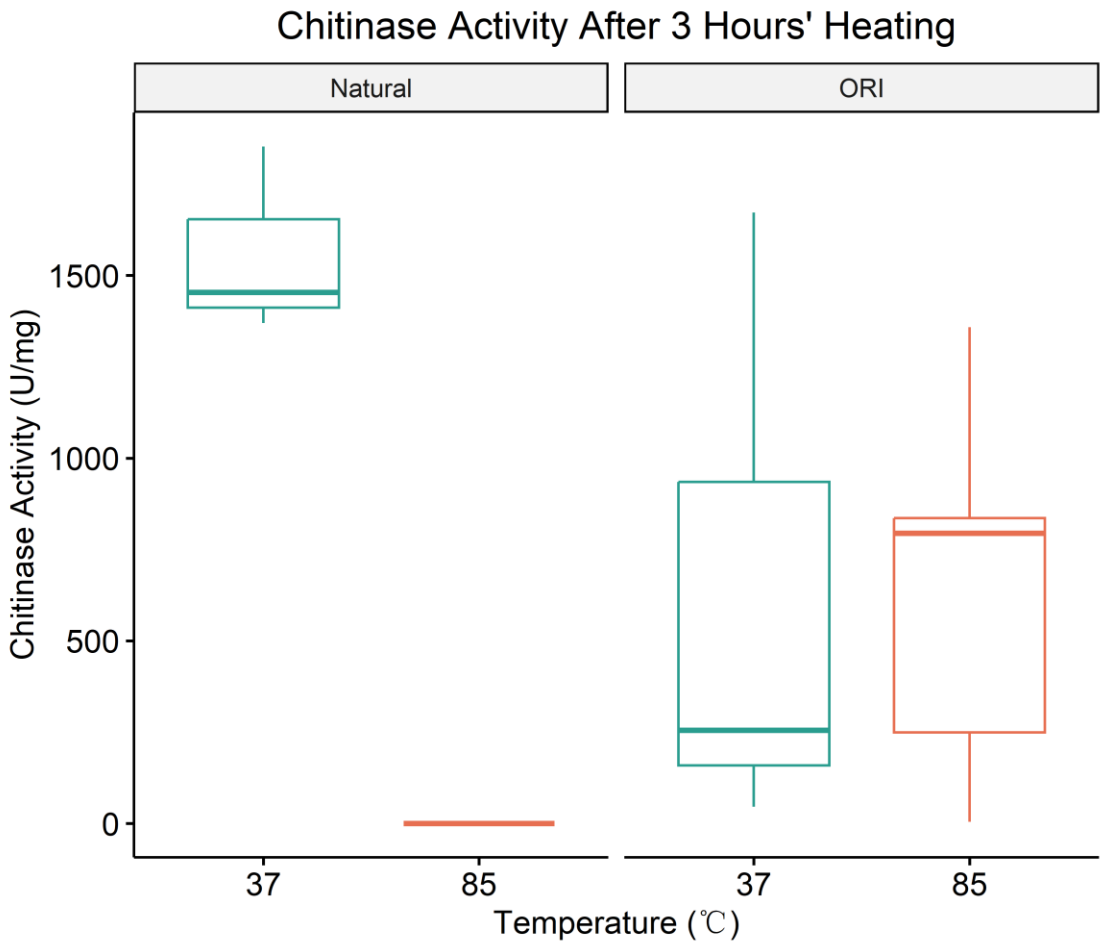
Supplementary Figure 5. Comparison of the enzyme activity between ORI-generated proteins and T4 lysozyme. The y-axis indicates the fold change in lysozyme activity of ORI-generated proteins relative to T4 lysozyme. Data are presented as mean values ± SD from three independent experiments (n = 3), and individual data points are overlaid as jittered dots. Source data are provided as a Source Data file.



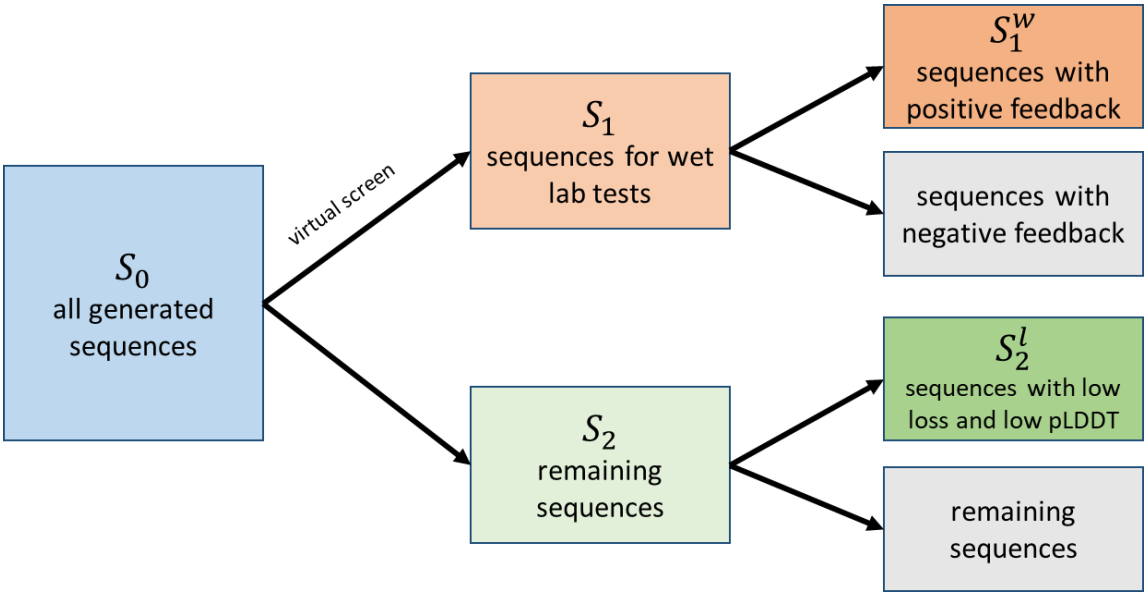
Supplementary Figure 6. Comparison of thermostable protein generation by ORI, ProGen3, ProteinSGM, and Evo2. The y-axis represents the melting temperatures of proteins generated by each method. In the box plots, the centre line indicates the median; the bounds of the box represent the 25th and 75th percentiles (interquartile range); the whiskers extend to the minima and maxima (excluding outliers); while additional solid points represent statistical outliers located beyond the whiskers. Statistical significance between ORI and each baseline is indicated by stars (*p < 0.05, **p < 0.01, ***p < 0.001, **p < 0.0001), calculated using the two-sided Wilcoxon rank-sum test without multiple comparisons adjustment.



Supplementary Figure 7. Comparison of protein thermostability optimization by ORI, ProteinMPNN, and Rosetta. (a) Boxplots of the melting temperatures of proteins optimized by each method. In the box plots, the central line indicates the median; the bounds of the box represent the 25th and 75th percentiles (interquartile range); the whiskers extend to the minima and maxima (excluding outliers); and individual data points are overlaid as jittered dots. Statistical significance was determined using the two-sided Wilcoxon rank-sum test without multiple comparisons adjustment. Statistical significance is indicated by stars (*p < 0.05, **p < 0.01, ***p < 0.001, **p < 0.0001). (b) Computational cost of each method for optimizing the proteins shown in panel (a). Computational cost is reported as the $\log_{10}$ value of CPU hours, with GPU runtimes converted into equivalent CPU hours for consistency.



Supplementary Figure 8. Chitinase activity comparison of natural and ORI-generated proteins after 3 hours of heating at different temperatures. Boxplots of the melting temperatures of proteins optimized by each method. In the box plots, the central line indicates the median; the bounds of the box represent the 25th and 75th percentiles (interquartile range); the whiskers extend to the minima and maxima (excluding outliers).



Supplementary Figure 9. Illustration of preference dataset construction. The preference dataset is utilized in an iterative optimization process with wet lab feedback to refine the model using Direct Preference Optimization (DPO).

Supplementary Table 1. Comparison of USM with other models in computational efficiency.

Model	**MSA Size**	**Parameters**	**FLOPS(G)**
USM-100M	128 x 128	113. 32M	2939
USM-3B	128 x 128	2.83B	73452
ESM2-100M	1 x 128	113.43M	29.81
ESM2-3B	1 x 128	2.95B	730.94
MSA Transformer	64 x 128	114.44M	1976

Supplementary Table 2. Comparison of USM with other models in predicting protein characteristics.

Model		**Performance**			
		Spearman's Correlation		**ACC**	
Name	**Parameters**	**Fluorescence**	**Thermostability**	**Antibiotic Resistance**	**Metal Ion Binding**
USM	100M	**0.68**	**0.81**	**0.98**	**0.80**
ESM2	15B	0.64	0.76	**0.98**	**0.80**
MSA-Transformer	100M	0.64	0.67	0.91	0.71

Supplementary Table 3. The protein expression and enzyme activity assays of novel proteins.

Protein	Basic Information			Protein Expression(mg/ml)	Enzyme Activity	
	Sequence Identity*	pLDDT	Length		pH	Activity(U/mg)
TX-L1	0.55	76.10	84	0.18	6.0	186.65
					7.0	235.42
					7.5	401.86
					8.0	111.54
					8.5	N.A.
					9.0	N.A.
TX-L2	0.53	83.62	146	0.09	6.0	527.87
					7.0	632
					7.5	252.54
					8.0	N.A.
					8.5	N.A.
					9.0	N.A.
TX-L3	0.82	81.08	194	0.10	6.0	280.96
					7.0	187.88
					7.5	N.A.
					8.0	N.A.
					8.5	N.A.
					9.0	N.A.
TX-L4	0.67	73.14	199	0.07	6.0	359.45
					7.0	261.73
					7.5	N.A.
					8.0	N.A.
					8.5	N.A.
					9.0	N.A.
TX-L5	0.62	83.74	94	0.16	6.0	237.02
					7.0	200.38
					7.5	N.A.
					8.0	N.A.
					8.5	N.A.
					9.0	N.A.
TX-L6	0.54	79.48	73	0.05	6.0	741.74
					7.0	1037.5
					7.5	1598.39
					8.0	248.05
					8.5	240.85
					9.0	156.71

*The sequence identity to the closest natural lysozyme proteins.

Supplementary Table 4. The sequence of ORI-engineered proteins.

Name	Type	Sequence
TX-L6	Protein with lysozyme activity	MLAPRTVAGGVAGAAMWDQVA QCESSGDWQNGFHGGLQFTQST WETAFGGLYAPSADKASRPDLAS RSQQIAVAERVLAGQGPGAWPN CGGRLP
TX-RL15	Protein with lysozyme activity	MWERVAACESGGRWHLDTGNG YFGGLQLAKATWAAFGGTGSAA EHSRARQIEIAKQVRAARGTRPW PRCTAGLGLAGG
TX-SC2	Thermostable protein with chitinase activity	MSPKRIIAYFPEWKVKDEYLNYS VEDIPWKLLTHINYAFAKIVEGK LHPIEEDLFYSNMEKIKEYKKEY RDVKVLISVGGWTDSGEFSDVAL NEENRKKFAKSALELVKEFNLDG VDIDWEFPCTEIDKENFTLLLKTL RDVLKEENENYLLTIAAPASYTQI HNTEPDKYHIFLDFINLMTYDFH GIWDKYTNHHSPLYGNPKDPDE KSRERANCDFAVKEYLKFGIPPE KIVLGVPFYGKGWICEDDGHNGL FAKVKGIPYGIFDNEDHPSGSNPF FFIKGVIENDPNYIKFRDKYAKVP WLWNPKEKIMYSYDDEESILEKC NYVIKNNLGGIMFWEVTQDYPF KGHTLVKLIHQKFWEG
TX-ME	Protein with lysozyme and chitinase activity	APPPTLVVYWGQNGSEGTLAEA CATGRYDLVNIAFLDVFGRGSPP PVITLDGHCDDPDDTGCPGLAAL KSCQAKGIKVLLSIGGGSGGYGL SSPEDAKSVANYLWDNFLGGSSS SRPLGDAVLDGIDLDIEKGGATG HWDDLARFLKAYSGVGRRVYLT AAPQCPFPDASLGPALDTGLFDY VWVQFYNNPPCQYSSGNINNLV KAWNQWTTSIPARVFLGLPAAPE AAGSGYIPPDVLTSQILPAVKTSA KYGGVMLWSRYYDELTGYSSKI KHHV