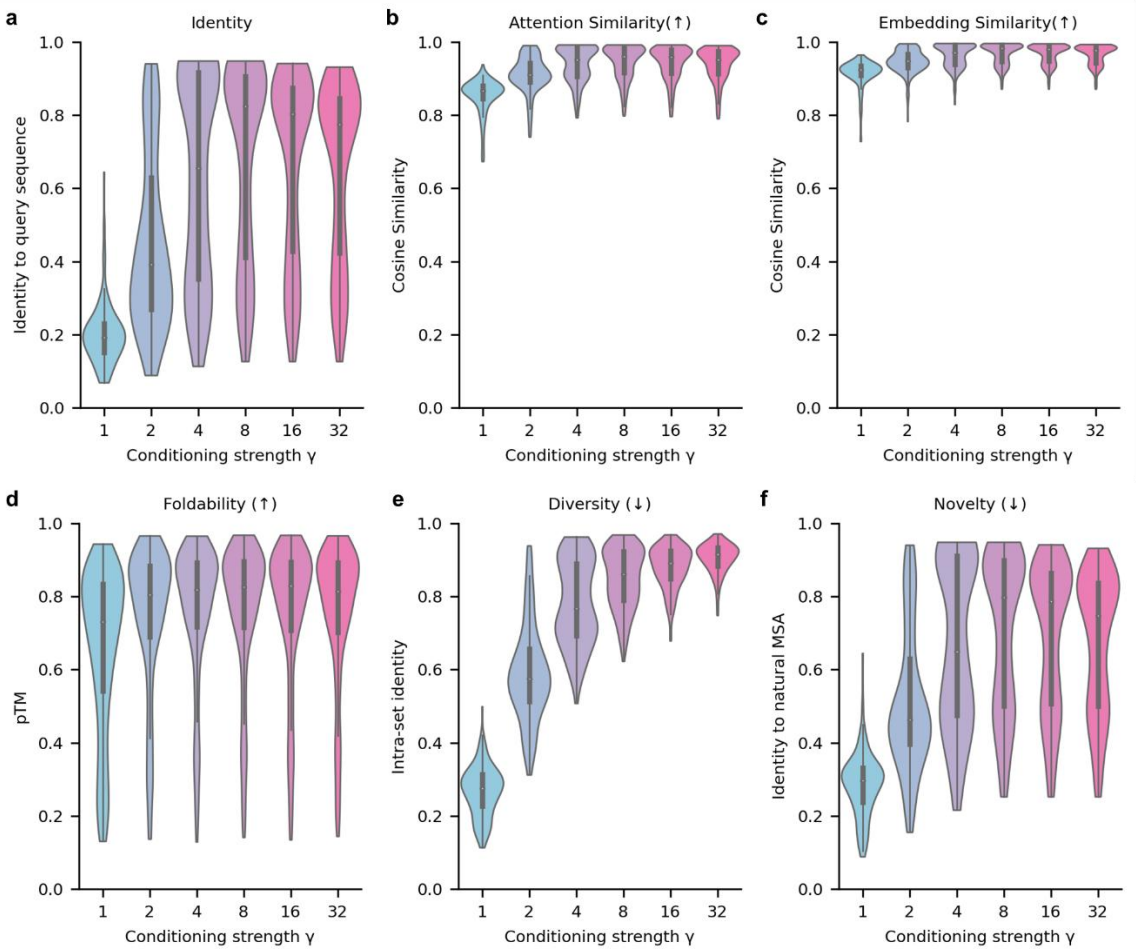
Supplementary Figures



Supplementary Fig. S1. Effects of conditioning strength on sequence similarity, foldability, diversity, and novelty in EvoGUD.

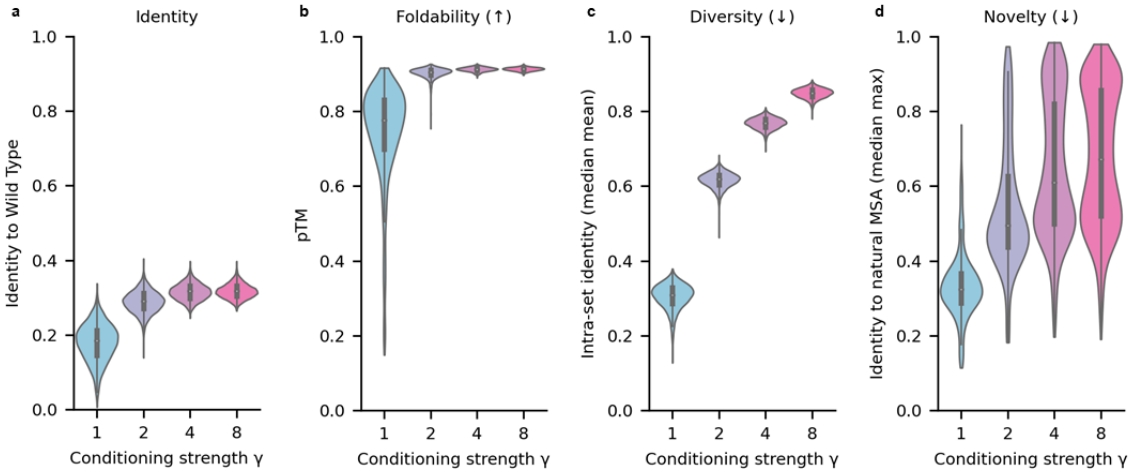
Violin plots summarize sequence- and structure-level statistics of EvoGUD-generated sequences as a function of the conditioning strength γ. Each violin represents 159 targets from the RecentPDB-monomer test set; for each target, the plotted value corresponds to the median across 1,024 generated sequences. Sequences were generated under an identity filter of 0.05–0.95 relative to the query sequence.

(a) sequence identity to the query sequence (same data as Fig. 1d, shown here for reference alongside other generation statistics); (b) cosine similarity between attention maps of generated sequences and the query, computed from the ESM2-3B model; (c) cosine similarity between ESM2-3B embeddings of generated sequences and the query; (d) foldability measured by predicted TM-score (pTM) from ESMFold; (e) intra-set diversity quantified as the mean pairwise sequence identity among generated sequences; (f) novelty measured as the maximum sequence identity of each generated sequence to the closest natural homolog in the corresponding natural MSA.

Metrics annotated with ↑ or ↓ indicate whether higher or lower values are desirable for the intended objective. Distributions reflect variation across targets at each γ, illustrating how conditioning strength controls the trade-off between similarity to the query, structural confidence, sequence diversity, and evolutionary novelty.

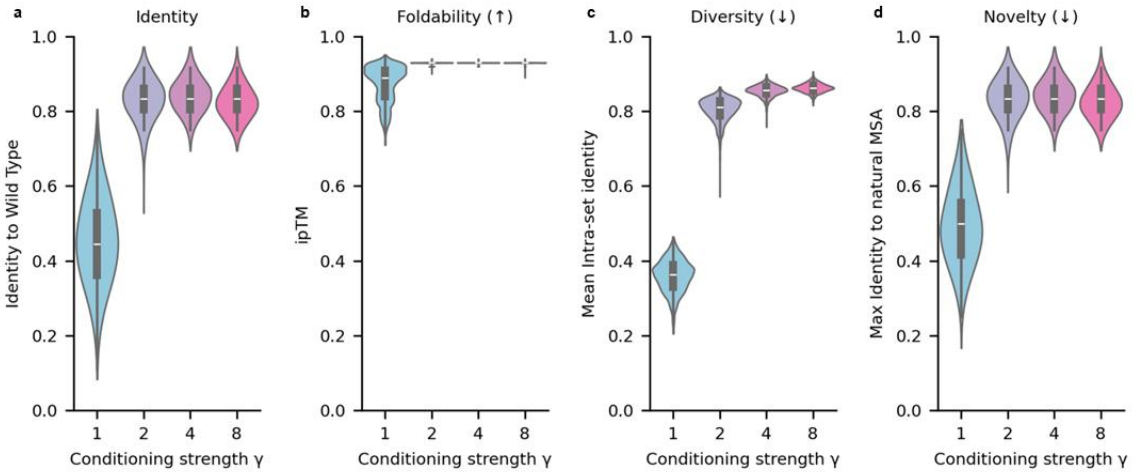
Supplementary Fig. S2. Relationship between sequence identity and embedding similarity for EvoGUD-generated, natural, and random mutated sequences.

Scatter plot showing the relationship between sequence identity to the query (x-axis) and cosine similarity of ESM-2 embeddings (y-axis). EvoGUD data points correspond to 1,024 generated sequences per target across 159 RecentPDB-monomer proteins and six conditioning strengths ($\gamma \in \{1, 2, 4, 8, 16, 32\}$), using the same generated sequences as in Fig. 1d and Supplementary Fig. S1c. Natural MSA sequences were included only when covering at least 80% of query positions to ensure comparable alignment context. Solid lines denote mean trends and shaded regions indicate 95% confidence intervals.



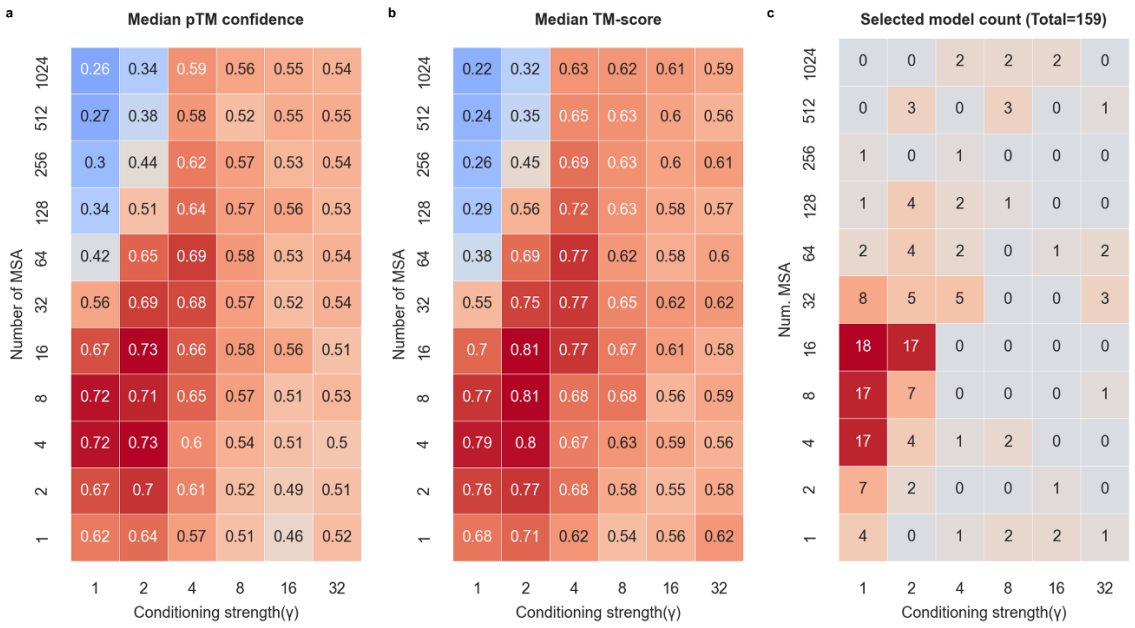
Supplementary Fig. S3. Selection of EvoGUD conditioning strength for TadA generation.

a, Pairwise identity to WT. b, Predicted complex confidence quantified by ipTM. c, Intra-set identity (mean), reported as an inverse proxy for diversity (lower intra-set identity indicates higher diversity). d, Identity to natural homologs (maximum identity), reported as an inverse proxy for novelty (lower identity indicates higher novelty).



Supplementary Fig. S4. Selection of EvoGUD conditioning strength for CcdA generation.

a, Pairwise identity to WT. b, Predicted complex confidence quantified by ipTM. c, Intra-set identity (mean), reported as an inverse proxy for diversity (lower intra-set identity indicates higher diversity). d, Identity to natural homologs (maximum identity), reported as an inverse proxy for novelty (lower identity indicates higher novelty).



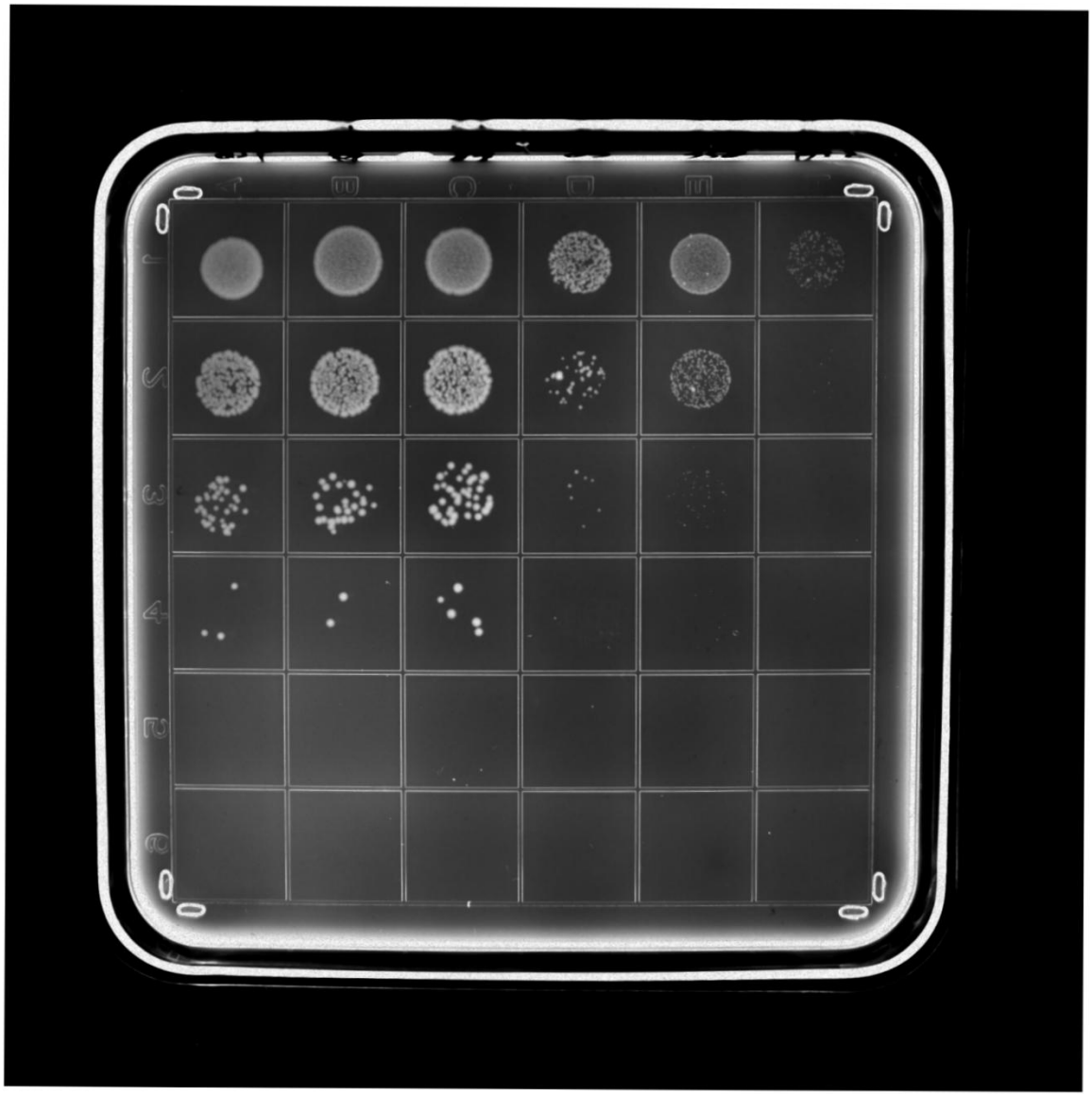
Supplementary Fig. S5. Dependence of EvoGUD + AF3-SS performance on conditioning strength and virtual MSA depth.

a, Median pTM confidence scores predicted by AlphaFold3 for EvoGUD + AF3-SS models across a grid of conditioning strengths γ (x-axis) and virtual MSA (vMSA) depths (y-axis).

b, Corresponding median TM-scores evaluated against experimental structures.

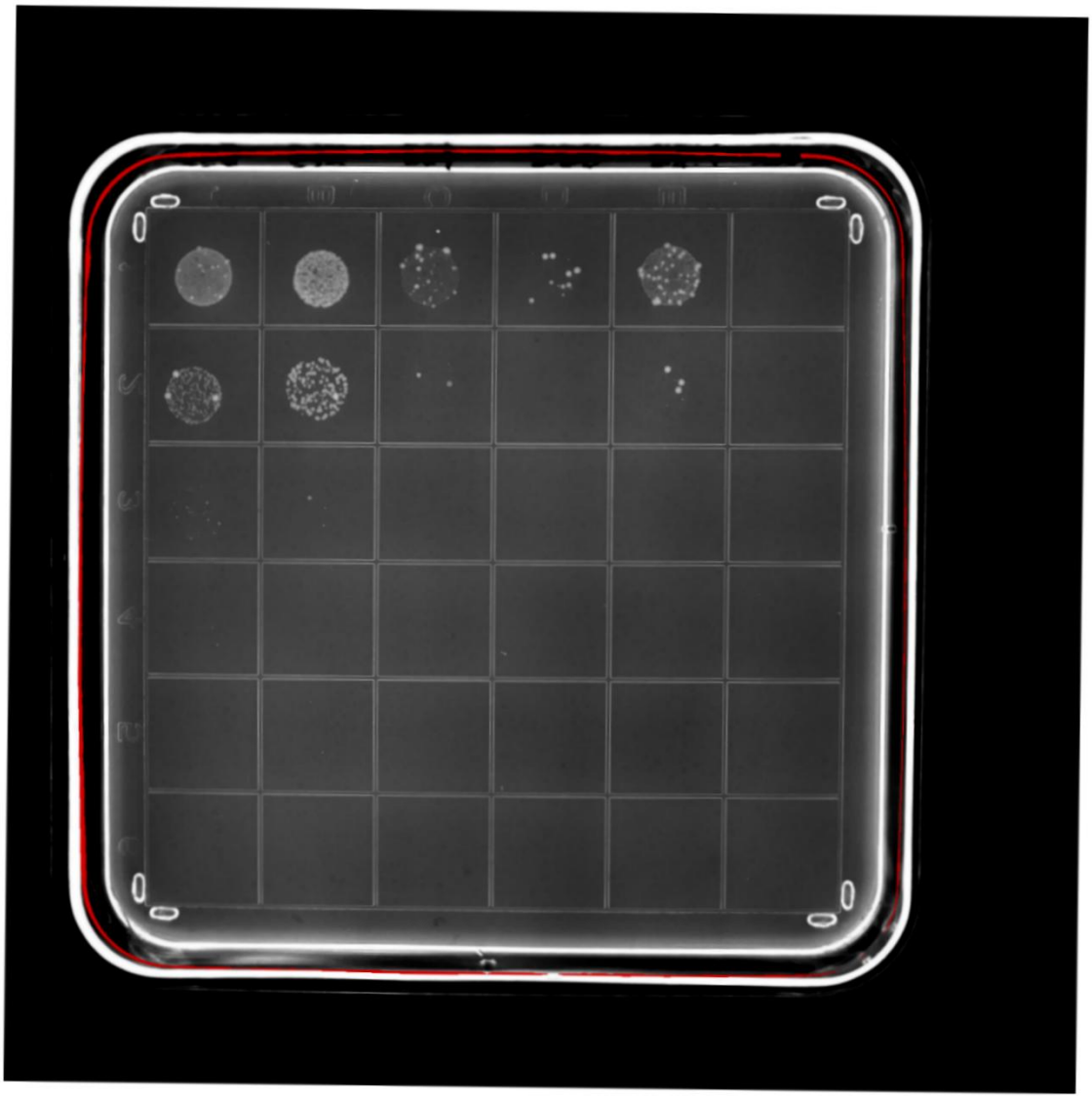
c, Distribution of final model selections across the same parameter grid for the 159 monomer targets. For each target, multiple EvoGUD + AF3-SS models were generated using different combinations of γ and vMSA depth, and the final model was selected based on AlphaFold3 pTM confidence score. Cell values indicate the number of targets whose best-scoring model originated from each parameter setting.

Together, these results show that optimal performance and model selection are concentrated in an intermediate regime of conditioning strength and vMSA depth, reflecting a balance between evolutionary constraint and sequence diversity.



Supplementary Fig. S6. Unedited plate images for CcdA selection (plate 1)

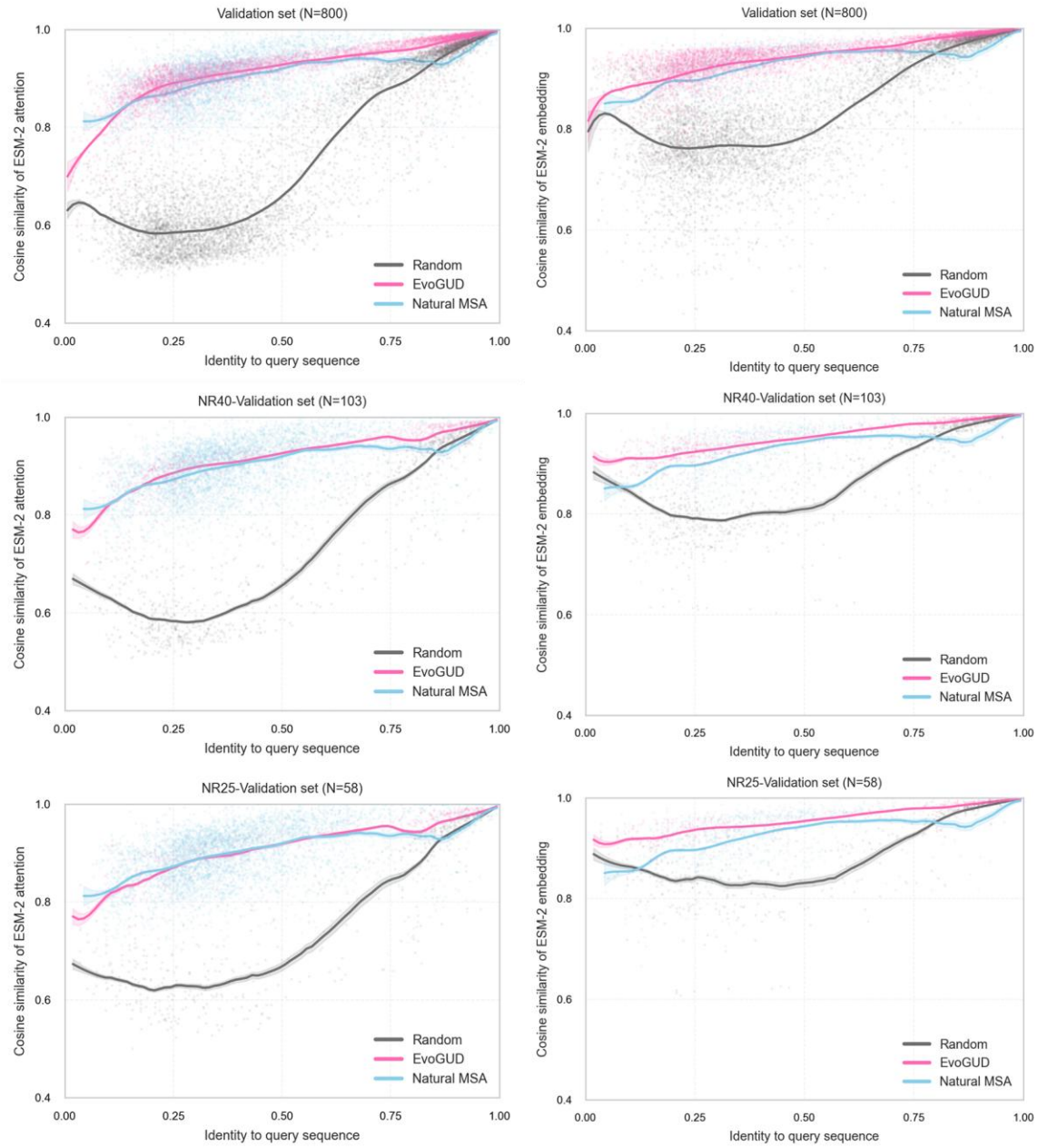
Raw, unedited images of CcdA plate selection experiment 1. Rows (top to bottom) correspond to serial dilutions of 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, and 10^{-6}. Columns (left to right) correspond to WT (normalized $\log_2$ fitness = 12.08), egCcdA-1 (18.41), egCcdA-14 (15.50), egCcdA-78 (10.53), egCcdA-391 (5.66), and egCcdA-933 (3.45). Images are shown without cropping, contrast enhancement, or local intensity adjustment.



Supplementary Fig. S7. Unedited plate images for CcdA selection (plate 2).

Raw images of CcdA plate selection experiment 2. Rows (top to bottom) correspond to serial dilutions of 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, and 10^{-6}. Columns (left to right) correspond to egCcdA-391 (5.66), egCcdA-509 (5.05), egCcdA-675 (4.32), egCcdA-660 (4.39), egCcdA-630 (4.51), and the CcdB-only negative control. To enable visual comparison with Supplementary Fig. S6, a uniform global brightness scaling (85%) was applied to

the entire image; no local adjustments, contrast enhancement, or selective editing were performed.

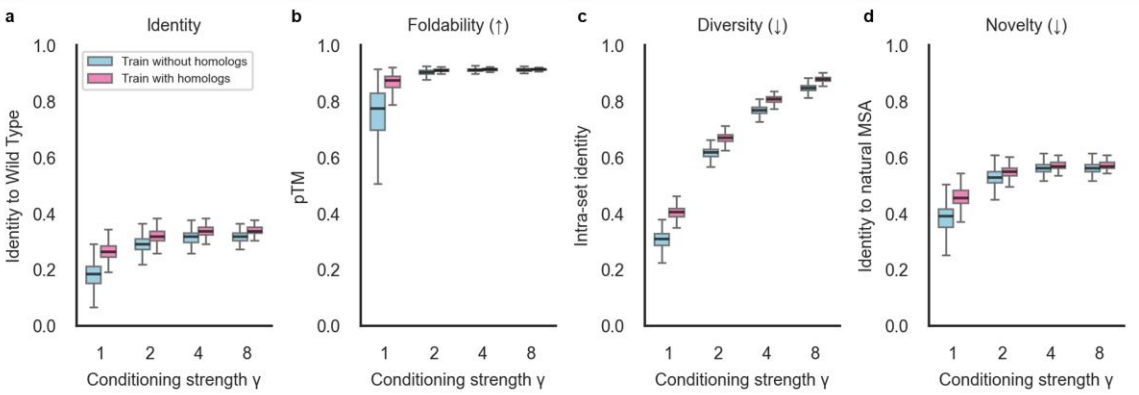


Supplementary Fig. S8. EvoGUD preserves evolutionary representation-space trajectories across homology-filtered validation sets.

Scatter plots show the relationship between sequence identity to the query (x-axis) and cosine similarity in ESM-2 representation space (y-axis) for generated and reference sequences. Left panels report similarity in ESM-2 attention space, and right panels report similarity in ESM-2 embedding space. Rows correspond to the full validation set (top, N

= 800), a non-redundant NR40 subset (middle, N = 103), and an extremely stringent NR25 subset (bottom, N = 58). Pink points and curves denote EvoGUD-generated sequences, blue denote natural MSA homologs, and gray denote identity-matched random sequences. Solid lines indicate smoothed trends, and shaded regions represent 95% confidence intervals.

Across all validation regimes, EvoGUD-generated sequences closely follow the representation-space trajectories of natural homologs and remain well separated from random controls, independent of the evolutionary proximity between the query sequences and the training data.



Supplementary Fig. S9. Limited impact of training-set redundancy on EvoGUD TadA sequence generation.

Comparison of sequence- and structure-level properties of TadA variants generated by EvoGUD models trained with (pink) or without (blue) TadA-proximal homologs (>40% sequence identity to the TadA wild type). The model trained with/without TadA-proximal homologs used exactly the same training settings and networks. Results are shown as a function of conditioning strength γ.

a, Sequence identity to the TadA wild-type sequence.

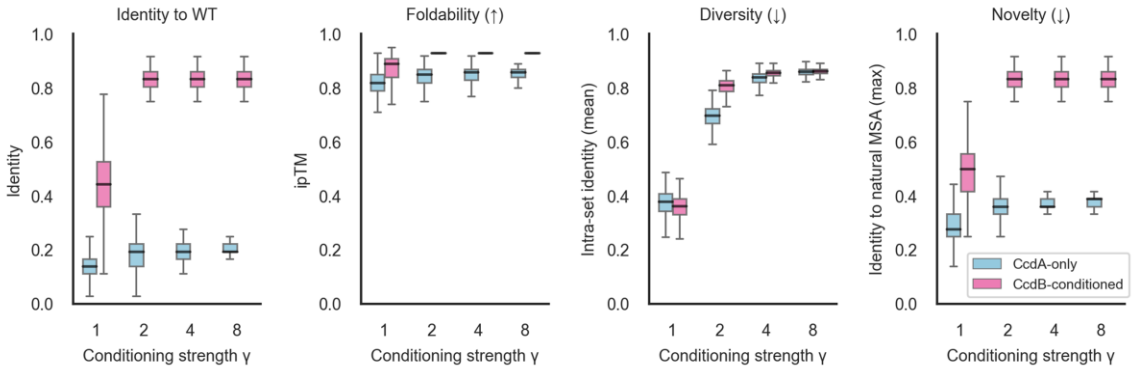
b, Predicted foldability quantified by ESMFold pTM.

c, Intra-set diversity measured as mean pairwise sequence identity among generated sequences.

d, Novelty measured as the maximum sequence identity of each generated sequence to the closest natural homolog in the TadA MSA.

Across all metrics and conditioning strengths, the two training regimes exhibit highly similar distributions, with only modest increases in identity and novelty for the model trained with TadA homologs. These results indicate that the presence of closely related sequences in the training set introduces only minor, non-collapsing biases and does not

substantially alter EvoGUD's generative behavior. On this basis, we did not explicitly remove CcdA homologs from the training set or retrain the model for CcdA sequence generation.



Supplementary Fig. S10. Impact of CcdB fusion conditioning on EvoGUD CcdA sequence generation.

Comparison of sequence- and structure-level properties of CcdA variants generated by EvoGUD under CcdA-only conditioning (blue) or CcdB–G_{50}–CcdA–G_{50}–CcdB fusion conditioning (pink), shown as a function of conditioning strength γ.

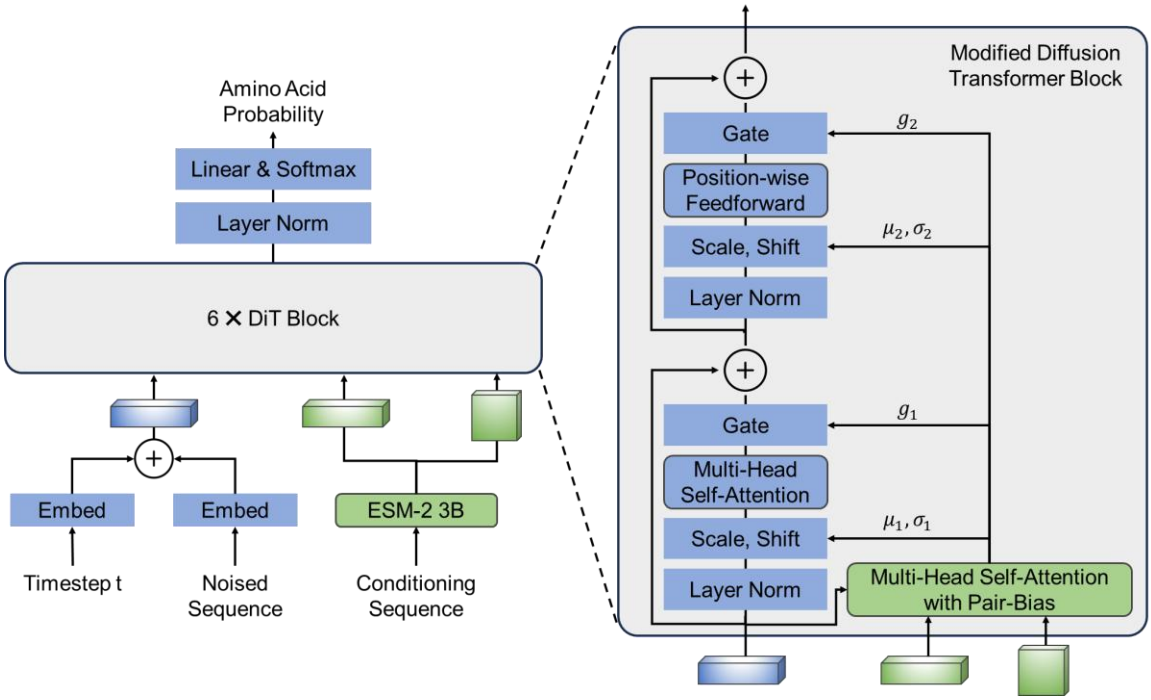
a, Sequence identity to the CcdA wild-type sequence.

b, Predicted foldability quantified by AlphaFold3 ipTM for the CcdA–CcdB complex.

c, Intra-set diversity measured as mean pairwise sequence identity among generated CcdA variants.

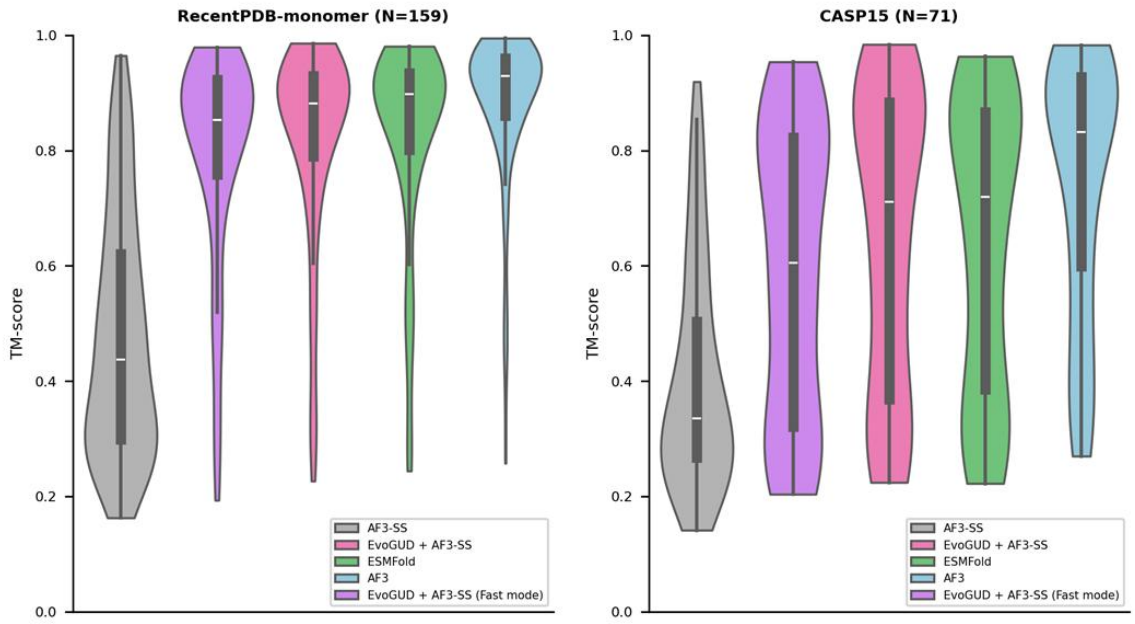
d, Novelty measured as the maximum sequence identity of each generated CcdA variant to the closest natural homolog in the CcdA MSA.

Across conditioning strengths, fusion conditioning with CcdB systematically shifts the generated sequence distribution toward higher identity to wild-type CcdA and improved predicted complex foldability, while maintaining controlled diversity and novelty. These results indicate that conditioning on a cognate binding partner provides informative interaction-specific constraints that guide EvoGUD toward functionally coherent regions of sequence space for intrinsically disordered proteins.



Supplementary Fig. S11. Detailed architecture of the Modified Diffusion Transformer Block.

Unlike the standard DiT which uses global conditioning, EvoGUD employs a sequence-specific cross-attention module with pair-bias. The ESM-2 residue embeddings serve as the context (K, V), while the projected attention maps are incorporated as a pair-bias within the multi-head self-attention. This module generates position-wise modulation parameters (μ, σ, g) for the Adaptive Layer Normalization (adaLN-Zero) layers, allowing for spatially-aware control of the denoising process across both the self-attention and position-wise feedforward sub-layers.



Supplementary Fig. S12. Comparison of adaptive GA mode and fixed-trial fast mode for EvoGUD + AF3-SS on monomer benchmarks.

a, TM-score distributions on the RecentPDB-monomer benchmark (N=159). Predictions from AF3-SS (single-sequence AlphaFold3), EvoGUD + AF3-SS with adaptive genetic-algorithm selection (default mode), ESMFold, AF3, and EvoGUD + AF3-SS fast mode are compared. In the fast mode, a fixed ensemble of 10 preset EvoGUD parameter settings is evaluated without iterative optimization, whereas the default mode uses a genetic algorithm to adaptively select a target-specific subset of EvoGUD-generated sequences for AlphaFold3 inference. The fast mode recovers most of the accuracy gain obtained by the adaptive mode, with mean TM-scores of 0.795 for the fast mode and 0.820 for the adaptive mode, compared with 0.476 for AF3-SS.

b, TM-score distributions on the CASP15 monomer benchmark (N=71). The same comparison is shown on an independent community benchmark. The fast mode again

captures most of the performance gain of EvoGUD-assisted AlphaFold3 while requiring substantially less computation, with mean TM-scores of 0.593 for the fast mode and 0.650 for the adaptive mode, compared with 0.403 for AF3-SS.

Together, these results show that the fixed-trial fast mode provides a computationally efficient approximation to the default adaptive EvoGUD + AF3-SS pipeline, while the genetic-algorithm mode yields the strongest overall monomer prediction performance.

Supplementary Table S1. Amino acid sequences of wild type TadA and EvoGUD-generated variants.

Name	Sequence
Staphylococcus aureus TadA (WT)	MTNDIYFMTLAIEEAKKAAQLGEVPIGAIITKDDEVIARAHNLRETLQQPTAHAEHIAIER AAKVLGSWRLEGCTLYVTLEPCVMCAGTIVMSRIPRVVYGADDPKGGCSGSLMNLLQ QSNFNHRAIVDKGVLKEACSTLLTTFFKNLRAN
egTadA-1	MKDDEEFMEQALKMAEIAYEKGEIPVGSVVVDNGKIVGSGHNQREKTKDATAHAEIM AIKKASQKINNWRLSDCVLYVTLEPCAMCSGAMLNARIAKVVWGAPDPKAGAAGTMA NVFLDPYLNHQVEIKGGLMADECAAMLNSFFREQRKK
egTadA-2	MKDHEEYMQQALALAREAYALGDVPVGAVVVRNGKIVGSGHNQRETTKDAMAHAE MMAITAASQKLGNWRLLDCSIYVTLEPCPMCAGAMLQARIAKVIFGAPDPRAGAAGTI LNVFSSPQLNHHVEITGGVLADECADLLKAYFKARRKK
egTadA-3	MKDDEEYMKQAIALAQQALARGDVPVGAVVVREGKILGSGYNQKVKTKDATSHAEM MAIQQASQALNNWRINDAVLYVTLEPCAMCAGAIVQARIARVVFGAPDPRAGAAGTL ADMFDDPRLNHHVEVLAGVMRDECSDMLNEFFRERRKK
egTadA-4	MKDDDKYMKEALKLAQKAYAKGETPVGAVVVRNGKILASGYNQKQSTLDATAHAEI MAIREASQKINNWRLNDCVLYVTLEPCAMCAGALVQARIAKLIFGAYDPKAGAAGTVF NIPKDPHLNHKIEITGGVMKDECADLLQEFFESKRKK
egTadA-5	MSDDEFFMKKALKLAQKASKMGEVPVGAVVVRNGKILGSGYNQRELTKDATAHAEM LAIQKAGEKLGNWRLNDCTIYVTLEPCPMCAGAMIQARVAYVIFGAKDPKAGAAGTM ASIFNEPYLNHQVEIIGGIMEDECAELLSSFFRQRRKK
egTadA-6	MKDDERFMRMALDQARRAFARGEVPIGAVVVRDGRILGAGYNQTERSKDPLAHAEIM AIRQASEALKNWRLTDCVMYVTLEPCAMCAGAMLQARIAKLVFGAPDPRAGAAGSIFD IFNDPSLNHKVEVKGGIGAQECREILKDYFRERRKK
egTadA-7	MADDEAFMRQALALARTALAKGDVPVGAVVVRGGRIVGAGYNQRQSTKDALAHAEM LAIREAGRALGNWRLSDCVLYVTLEPCPMCAAAMLQARVARVVFGAPDPKAGAAGSF FDIFDDPGLNHRVEITSGLLADECAALLRAFFRERRHK
egTadA-8	MKADEAWMRRALDLARRAAAQGEVPVGAVVVRDGKIVGAGHNQKEASKDALDHAE MLAIRKASAALKNWRLTDAVIYVTLEPCAMCAGAMAQARVARVVFGAPDPRAGAAG TMANVFEDPRLNHHIEVTGGIRADECAALLSSFFRKRRKR
egTadA-F1 (Failed)	MKDDERYMRRALDRARRAEAAGEVPVGAVLVRGGKILGSGHNQKVQSKDALGHAEM MAITAAAEAIGTWRLPDCVLYVTLEPCPMCAGAMIHARVARLVFGAADPRAGAAGSM LDVFGDPRLNHRVAVVAGVMADECAALLKEFFREMRAK
egTadA-F2 (Failed)	MKDDEFYMRKALELARRALAAKEVPVGAVVVYNGKILGSGYNQKESTKDALAHAEIQ AIREASDALKNWRLLDCVMYVTLEPCAMCAGAIMQARIATVVFGAYDPKAGAAGTML DIFEDPHLNHHSEIISGVLAEECKDLMKSYFRERRKK

Supplementary Table S2. egTadA variants activity validation data for plate display.

Variant	TMP-resistant colonies (300 µL spread)				Total viable cells		Activity
	Rep 1	Rep 2	Rep 3	Mean ± STD	10 mL	300 µL**	
Neg. Control	0*	0	0	0.00 ± 0.00	2.00×10^9	6.00×10^7	0
WT	9	13*	8	10.00 ± 2.65	1.50×10^9	4.50×10^7	4.86×10^{-8}
egTadA-1	20*	20	17	19.00 ± 1.73	1.50×10^9	4.50×10^7	9.24×10^{-8}
egTadA-2	12	15	17*	14.67 ± 2.52	2.00×10^9	6.00×10^7	5.35×10^{-8}

Notes:

* denoted the plates displayed in Fig 4.c.

** Total viable cell counts were extrapolated to the 10 mL culture volume from serial dilution spot assays on non-selective plates.

Supplementary Table S3. egTadA variants activity validation data.

Variant	TMP-resistant colonies (300 µL spread)				Total viable cells		Activity
	Rep 1	Rep 2	Rep 3	Mean ± STD	10 mL	300 µL*	
egTadA-1	9	9	12	10.00 ± 1.73	6.00×10^{8}	1.80×10^{7}	1.22×10^{-7}
egTadA-2	5	3	2	3.33 ± 1.53	3.00×10^{8}	9.00×10^{6}	8.10×10^{-8}
egTadA-3	6	15	2	7.67 ± 6.66	8.00×10^{8}	2.40×10^{7}	6.99×10^{-8}
egTadA-4	9	6	7	7.33 ± 1.53	8.50×10^{8}	2.55×10^{7}	6.29×10^{-8}
egTadA-5	2	4	6	4.00 ± 2.00	5.50×10^{8}	1.65×10^{7}	5.30×10^{-8}
egTadA-6	5	5	3	4.33 ± 1.15	7.00×10^{8}	2.10×10^{7}	4.52×10^{-8}
egTadA-7	2	2	3	2.33 ± 0.58	6.50×10^{8}	1.95×10^{7}	2.62×10^{-8}
egTadA-8	3	3	0	2.00 ± 1.73	9.00×10^{8}	2.70×10^{7}	1.62×10^{-8}

Notes:

* Total viable cell counts were extrapolated to the 10 mL culture volume from serial dilution spot assays on non-selective plates.

Supplementary Table S4. Cloning and sequence verification of selected CcdA variants.

Name	Protein Sequence	DNA Sequence
E. coli CcdA (WT)	MRRLRAERWKAEN QEGMAEVARFIEMN GSFADENRDW	ATGCGTCGTCTGCGCGCAGAACGTTGGAAAGCAGAAAATCA GGAAGGCATGGCAGAAGTGGCCCGTTTTATTGAAATGAATG GCAGCTTTGCCGATGAAAATCGTGATTGGTAA
egCcdA-1	MNRIKAEKWKADN QEGMAEIAKFVEEN GSFAEENRDW	ATGAACCGTATTAAAGCGGAAAAATGGAAAGCGGATAACCA GGAAGGCATGGCGGAAATTGCGAAATTTGTGGAAGAAAACG GCAGCTTTGCGGAAGAAAACCGCGATTGGTAA
egCcdA-14	MKKMKADQWREEN EEGMAEVARFIDEN GSLAEENRDW	ATGAAAAAAATGAAAGCGGATCAGTGGCGCGAAGAAAACG AAGAAGGCATGGCGGAAGTGGCGCGTTTTATTGATGAAAAC GGCAGCCTGGCGGAAGAAAACCGCGATTGGTAA
egCcdA-78	MKKIKAEKWRAEN KEGMAEISKYVEEN GSFAEENRDW	ATGAAAAAAATTAAAGCGGAAAAATGGCGCGCGGAAAACA AAGAAGGCATGGCGGAAATTAGCAAATATGTGGAAGAAAA CGGCAGCTTTGCGGAAGAAAACCGCGATTGGTAA
egCcdA-391	MKKLKAEKWREDN HEGMAEISKYIEMN GSFADENKEW	ATGAAAAAACTGAAAGCGGAAAAATGGCGCGAAGATAACC ATGAAGGCATGGCGGAAATTAGCAAATATATTGAAATGAAC GGCAGCTTTGCGGATGAAAACAAAGAATGGTAA
egCcdA-509	MKKLKAEKWKAEN QEGMSEIARYVEVN GSFAEENRDW	ATGAAAAAACTGAAAGCGGAAAAATGGAAAGCGGAAAACC AGGAAGGCATGAGCGAAATTGCGCGCTATGTGGAAGTGAAC GGCAGCTTTGCGGAAGAAAACCGCGATTGGTAA
egCcdA-630	MKKLKAEKWRAEN QEGMAEVARYVELN GSFAEDNRDW	ATGAAAAAACTGAAAGCGGAAAAATGGCGCGCGGAAAACC AGGAAGGCATGGCGGAAGTGGCGCGTTATGTGGAACTGAAC GGCAGCTTTGCGGAAGATAACCGTGATTGGTAA
egCcdA-660	MKKLKAEQWRAEN AEGMAEISKIVEVNG SFADENRDW	ATGAAAAAACTGAAAGCGGAACAGTGGCGCGCGGAAAACG CGGAAGGCATGGCGGAAATTAGCAAAATTGTGGAAGTGAAC GGCAGCTTTGCGGATGAAAACCGCGATTGGTAA
egCcdA-675	MKRLKAEKWKAEN EEGMAEIAKFIDENG SFGDENKEW	ATGAAACGCCTGAAAGCGGAAAAATGGAAAGCGGAAAACG AAGAAGGCATGGCGGAAATTGCGAAATTTATTGATGAAAAC GGCAGCTTTGGCGATGAAAACAAAGAATGGTAA
egCcdA-933	MKKLKAEKWREDN EEGMEEIAKYVEMN GSFAEDNRDW	ATGAAAAAACTGAAAGCGGAAAAATGGCGCGAAGATAACG AAGAAGGCATGGAAGAAATTGCGAAATATGTGGAAATGAAC GGCAGCTTTGCGGAAGATAACCGCGATTGGTAA

Supplementary Table S5. Identity-to-reference versus predicted divergence time from TimeTree-calibrated fits.

Family	Reference	Identity to reference	Predicted divergence (Ma)
TadA	*E. coli* TadA (UniProt P68398)	0.8	1.4049
		0.7	231.7360
		0.6	561.3701
		0.5	1148.1288
CcdA	*E. coli* CcdA (UniProt P62552)	0.9	4.6449
		0.8	20.3885
		0.7	40.2859
		0.6	67.3466
		0.5	109.8419
		0.4	215.5726

Notes:

Values are derived from TimeTree-calibrated timetrees and MSA-based identities using a saturating decay fit. They are intended as approximate, order-of-magnitude mappings rather than strict molecular-clock estimates. This Table is adapted from Chen et al.[1]

206 **Supplementary Table S6. RecentPDB-monomer test set.**

8UX2, 8G32, 8GM3, 8V6T, 9AZZ, 8W7P, 8R2C, 8G0N, 8VC8, 8CKN, 8OU1, 9BKW, 8QPJ, 8IRQ, 8TFO, 8CPK, 8FZ9, 8S5B, 8DQD, 8I11, 8RXA, 8ROM, 8G1L, 9FGP, 8T9T, 8RCW, 8YXK, 8JB3, 8SUC, 8SX0, 8J8Y, 8YQ4, 8Q1V, 8U47, 8IUD, 8JJN, 8JUP, 8POR, 8GKD, 8P8E, 8XJG, 8PAJ, 8P0C, 8TJG, 8CBA, 9B8D, 8WIK, 8TI5, 8QD1, 8X38, 8SSK, 8TW1, 8YTS, 8EZ6, 8Y9P, 8HEL, 8VPO, 8PAB, 8YE0, 8XIF, 8RE3, 8V33, 8S9L, 8RA0, 8ZQF, 8QLE, 8CIH, 8PW2, 8UJL, 8J2N, 8SV7, 8HW6, 8RHE, 8FZ8, 8CPQ, 8QB1, 7Z65, 8OVQ, 8FJM, 8QOH, 8T2J, 8YA6, 8OLD, 8SCA, 8QQ1, 8GDL, 8PAQ, 7GQS, 8BTM, 8IYM, 8YCM, 8TLS, 8FMG, 8S0U, 8IVH, 8TN1, 8IW2, 8VR3, 8VVA, 8Q90, 8J2W, 7YPD, 8VKT, 8X6V, 8Q28, 8PB3, 8P4W, 8V14, 8UOO, 8SBH, 8CMP, 8J5E, 8JU9, 8IXW, 8JDH, 8YPJ, 8WBL, 8YLA, 8P2O, 8JU3, 8SM8, 8CLO, 8OI4, 8JU7, 8JA1, 8V6S, 8Z9U, 8YLZ, 8TCD, 8VR2, 8WU6, 8VDW, 8UFO, 8P37, 8CKG, 8IYP, 8IMD, 8G48, 8PQP, 8IHA, 8HJF, 8IS2, 8AND, 8S9W, 8YE5, 8Q1W, 8ACX, 8VR6, 8OV8, 8JYV, 8WCT, 8JQF, 8QKY, 8IL8, 8SM6, 9ARD, 8W1D, 8ONM, 8IOO

207

208 **Supplementary Table S7. RecentPDB-multimer test set.**

7B0C, 7BCA, 7BJQ, 7EDS, 7EOF, 7F3J, 7F9H, 7M4L, 7MWH, 7MZ0, 7MZ1, 7MZ2, 7N5U, 7N5V, 7N5W, 7NQF, 7OGS, 7OOT, 7OUE, 7OWF, 7OY7, 7P0W, 7P3F, 7P8L, 7P9J, 7P9Z, 7PSX, 7PZA, 7PZB, 7Q3O, 7Q4N, 7Q94, 7QAZ, 7R6R, 7R6T, 7R8G, 7R8H, 7R8I, 7RCC, 7RCD, 7RCE, 7RCF, 7RCG, 7RGU, 7RSR, 7RSS, 7S03, 7S68, 7S9J, 7S9K, 7S9L, 7S9M, 7S9N, 7S9O, 7S9P, 7S9Q, 7SOP, 7SOS, 7SOT, 7SOU, 7SOV, 7SOW, 7SUM, 7SUV, 7SVB, 7SX5, 7SXE, 7T18, 7T19, 7T1A, 7T1B, 7T8K, 7TDW, 7TDX, 7TEA, 7TEC, 7TO1, 7TO2, 7TQW, 7TUV, 7TXC, 7TZ1, 7TZV, 7U76, 7U79, 7U7A, 7U7B, 7U7C, 7U7F, 7U7G, 7U7I, 7U7J, 7U7K, 7U7L, 7UBL, 7UBU, 7UPZ, 7UU4, 7UXD, 7V2Z, 7VE5, 7VG8, 7VKI, 7VKL, 7VN2, 7VO9, 7VOU, 7VOV, 7VOX, 7VP1, 7VP2, 7VP3, 7VP4, 7VP5, 7VP7, 7VSJ, 7VTI, 7WM3, 7WQ5, 7X5E, 7X5F, 7X5G, 7X5L, 7X5M, 7XHV, 7XI3, 7XQ5, 7XRC, 7XS4, 7YHO, 7YZE, 7YZF, 7YZG, 7Z0U, 7Z5A, 7ZHH, 7ZVN, 7ZVX, 8A1C, 8A4I, 8B0R, 8CSH, 8CTZ, 8CU0, 8CZQ, 8DVP, 8DVR, 8DVS, 8DVU, 8DVY, 8DW0, 8DW1, 8DW4, 8DW8, 8DWM, 8DZK, 8E2P, 8E2Q, 8EDJ, 8EF9, 8EFC, 8EFK, 8GMS, 8GMT, 8GMU

209
210

Supplementary Table S8. The sequence of pUC57-Kan-ccdA/B, pUC57-Kan-2BspQI-ccdB and primers used in this study.

pUC57-Kan-ccdA/B *	TGCAGCTCTGGCCCGTGTCTCAAAATCTCTGATGTTACATTGCACAAGATAAA AATATATCATCATGAACAATAAAACTGTCTGCTTACATAAACAGTAATACAAG GGGTGTTATGAGCCATATTCAACGGGAAACGTCGAGGCCGCGATTAAATTCC AACATGGATGCTGATTTATATGGGTATAAATGGGCTCGCGATAATGTCGGGCA ATCAGGTGCGACAATCTATCGCTTGTATGGGAAGCCCGATGCGCCAGAGTTGT TTCTGAAACATGGCAAAGGTAGCGTTGCCAATGATGTTACAGATGAGATGGT CAGACTAAACTGGCTGACGGAATTTATGCCTCTTCCGACCATCAAGCATTTTA TCCGTACTCCTGATGATGCATGGTTACTCACCACTGCGATCCCCGGAAAAACA GCATTCCAGGTATTAGAAGAATATCCTGATTCAGGTGAAAATATTGTTGATGC GCTGGCAGTGTTCCTGCGCCGGTTGCATTCGATTCCTGTTTGTAATTGTCCTTT TAACAGCGATCGCGTATTTCGTCTCGCTCAGGCGCAATCACGAATGAATAACG GTTTGGTTGATGCGAGTGATTTTGATGACGAGCGTAATGGCTGGCCTGTTGAA CAAGTCTGGAAAGAAATGCATAAACTTTTGCCATTCTCACCGGATTCAGTCGT CACTCATGGTGATTTCTCACTTGATAACCTTATTTTTGACGAGGGGAAATTAA TAGGTTGTATTGATGTTGGACGAGTCGGAATCGCAGACCGATACCAGGATCTT GCCATCCTATGGAACTGCCTCGGTGAGTTTTCTCCTTCATTACAGAAACGGCT TTTTCAAAAATATGGTATTGATAATCCTGATATGAATAAATTGCAGTTTCATTT GATGCTCGATGAGTTTTTCTAATCAGAATTGGTTAATTGGTTGTAACATTATTC AGATTGGGCTTGATTTAAAACTTCATTTTTAATTTAAAAGGATCTAGGTGAAG ATCCTTTTTGATAATCTCATGACCAAAATCCCTTAACGTGAGTTTTCGTTCCAC TGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTT TCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTG GTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTT CAGCAGAGCGCAGATACCAAATACTGTTCTTCTAGTGTAGCCGTAGTTAGGCC ACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATCCTG TTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTC AAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCG TGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTAC AGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACA GGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCC AGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGAC TTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAAC GCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCAC ATGTttgacagctagctcagtcctaggtataatgctagctactagagaaagaggagaaatactagATGCGTCGTCT GCGCGCAGAACGTTGGAAAGCAGAAAATCAGGAAGGCATGGCAGAAGTGGC CCGTTTTATTGAAATGAATGGCAGCTTTGCCGATGAAAATCGTGATTGGtaaAA GCTTgcacaggtcGTCGACTTAAATGCCCCAAAACATCAGGTTAATGGCATTCTTA ATATCATTTTCACGATGACTCAGATCTGCCACTTCTTCACCAATAACACTAAC CGGCACACTTGCCATATCGGTGGTCATCATACGCCAACTTTCATCGCCAATAT GCACAACCGGATACAGTTCACGACTCACTTTATCACTCAGCAGGCGGGCGCTG GCCAGCGGAATAACCATACGGCGACCCGGGGTATCAATAATATCGCTCTGCA CATCCACAAACAGACGATAACGGCTTTCACGCTTATAGGTGTAAACCTTAAAC TGCATactcttccttttcaatattattgaagcatttatcagggttattgtctcatgagcggatacatatttgaatgtatttagaaaaat aaacaaaCATATGGTGCACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAAGC CAGCCCCGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGTCTGCT CCCGGCATCCGCTTACAGACAAGCTGTGACCGTCTCCGGGAGCTGCATGTGTC AGAGGTTTTCACCGTCATCACCGAAACGCGCGA
pUC57-Kan-2BspQI-ccdB **	TGCAGCTCTGGCCCGTGTCTCAAAATCTCTGATGTTACATTGCACAAGATAAA AATATATCATCATGAACAATAAAACTGTCTGCTTACATAAACAGTAATACAAG GGGTGTTATGAGCCATATTCAACGGGAAACGTCGAGGCCGCGATTAAATTCC AACATGGATGCTGATTTATATGGGTATAAATGGGCTCGCGATAATGTCGGGCA ATCAGGTGCGACAATCTATCGCTTGTATGGGAAGCCCGATGCGCCAGAGTTGT TTCTGAAACATGGCAAAGGTAGCGTTGCCAATGATGTTACAGATGAGATGGT CAGACTAAACTGGCTGACGGAATTTATGCCTCTTCCGACCATCAAGCATTTTA TCCGTACTCCTGATGATGCATGGTTACTCACCACTGCGATCCCCGGAAAAACA

	GCATTCCAGGTATTAGAAGAATATCCTGATTCAGGTGAAAATATTGTTGATGC GCTGGCAGTGTTCCTGCGCCGGTTGCATTCGATTCCTGTTTGTAATTGTCCTTT TAACAGCGATCGCGTATTTCGTCTCGCTCAGGCGCAATCACGAATGAATAACG GTTTGGTTGATGCGAGTGATTTTGATGACGAGCGTAATGGCTGGCCTGTTGAA CAAGTCTGGAAAGAAATGCATAAACTTTTGCCATTCTCACCGGATTCAGTCGT CACTCATGGTGATTTCTCACTTGATAACCTTATTTTTGACGAGGGGAAATTAA TAGGTTGTATTGATGTTGGACGAGTCGGAATCGCAGACCGATACCAGGATCTT GCCATCCTATGGAACTGCCTCGGTGAGTTTTCTCCTTCATTACAGAAACGGCT TTTTCAAAAATATGGTATTGATAATCCTGATATGAATAAATTGCAGTTTCATTT GATGCTCGATGAGTTTTTCTAATCAGAATTGGTTAATTGGTTGTAACATTATTC AGATTGGGCTTGATTTAAAACTTCATTTTTAATTTAAAAGGATCTAGGTGAAG ATCCTTTTTGATAATCTCATGACCAAAATCCCTTAACGTGAGTTTTCGTTCCAC TGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTT TCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTG GTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTT CAGCAGAGCGCAGATACCAAATACTGTTCTTCTAGTGTAGCCGTAGTTAGGCC ACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATCCTG TTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTC AAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCG TGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTAC AGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACA GGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCC AGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGAC TTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAAC GCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCAC ATGTttgacagctagctcagtcctaggtataatgctagctactagagaaagaggagaaatactagATGAGAAGAGC ACGTCCGATTGCTCTTCGtaaAAGCTTgcacaggtcGTCGACTTAAATGCCCCAAAA CATCAGGTTAATGGCATTCTTAATATCATTTTCACGATGACTCAGATCTGCCA CTTCTTCACCAATAACACTAACCGGCACACTTGCCATATCGGTGGTCATCATA CGCCAACTTTCATCGCCAATATGCACAACCGGATACAGTTCACGACTCACTTT ATCACTCAGCAGGCGGGCGCTGGCCAGCGGAATAACCATACGGCGACCCGGG GTATCAATAATATCGCTCTGCACATCCACAAACAGACGATAACGGCTTTCACG CTTATAGGTGTAAACCTTAAACTGCATactcttccttttcaatattattgaagcatttatcagggttattgtct catgagcggatacatatttgaatgtatttagaaaaataaacaaaCATATGGTGCACTCTCAGTACAATCTG CTCTGATGCCGCATAGTTAAGCCAGCCCCGACACCCGCCAACACCCGCTGACG CGCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAAGCTGTGACC GTCTCCGGGAGCTGCATGTGTCAGAGGTTTTCACCGTCATCACCGAAACGCGC GA
P5-ccdA-009-FP	AATGATACGGCGACCACCGAGATCTACACtctgttggACACTCTTTCCCTACACGA CGCTCTTCCGATCTCTGGCCTTTTGCTGGCCT
P7-SalI-009-RD2-RP	CAAGCAGAAGACGGCATACGAGATcctgaactGTGACTGGAGTTCAGACGTGTGC TCTTCCGATCtTAAGTCGACgacctgtgc
P5-ccdA-010-FP	AATGATACGGCGACCACCGAGATCTACACctcaccaaACACTCTTTCCCTACACGA CGCTCTTCCGATCTCTGGCCTTTTGCTGGCCT
P7-SalI-010-RD2-RP	CAAGCAGAAGACGGCATACGAGATttcaggtcGTGACTGGAGTTCAGACGTGTGC TCTTCCGATCtTAAGTCGACgacctgtgc

Notes:

* For pUC57-Kan-ccdA/B, the J23119 promoter, ccdA$^{36\text{-}72}$, ccdB and AmpR promoter sequence are highlighted in red, green, blue, and purple, respectively. The plasmid backbone is pUC57-Kan.

** For pUC57-Kan-2BspQI-ccdB, the J23119 promoter, BspQI site, ccdB and AmpR promoter sequence are highlighted in red, green, blue, and purple, respectively. The plasmid backbone is pUC57-Kan.

Supplementary Methods

Adaptive vMSA selection by genetic algorithm for EvoGUD + AF3-SS

For the default monomer prediction pipeline, we used a genetic algorithm (GA) to adaptively select a target-specific subset of EvoGUD-generated sequences as the virtual MSA (vMSA) input to AlphaFold3. For each query, EvoGUD sequences generated under multiple conditioning strengths (γ = 1, 2, and 4; 1,024 sequences per setting) were pooled into a candidate library. The GA then iteratively optimized both vMSA composition and vMSA depth by evaluating candidate subsets with AlphaFold3 single-sequence inference and ranking them by a confidence-based fitness score (pTM + 0.5*pLDDT). Population updates used elitism, selection, crossover, mutation, and immigrant introduction over 5 to 50 generations, with population size 32. Early stop would be applied if pTM reach 0.9 or best fitness score did not improve for 5 generations. To improve robustness, initial populations and newly introduced individuals were biased by the empirical landscape derived from the benchmark grid search (Supplementary Fig. S5), rather than sampled uniformly. The best-scoring subset identified during the search was used as the final EvoGUD + AF3-SS prediction for each monomer target.

Experimental Validation of TadA Variants

TadA DNA-editing assay

DNA-editing activity of TadA variants was assessed using a trimethoprim (TMP) resistance reversion assay in *Escherichia coli*. A modified R67 dihydrofolate reductase (DHFR) reporter gene containing a premature TAG stop codon was co-expressed with

TadA variants. TadA-mediated A→G editing restores the functional TGG codon, thereby conferring resistance to TMP. Editing activity was defined as the site-specific per-base per-generation mutation rate, $\mu_{s.p.b.}$, which was estimated from the observed frequency of TMP-resistant colonies using a Luria–Delbrück rare-mutation approximation. TMP-resistant colonies were further validated by Sanger sequencing to confirm the intended A→G substitution. Specific details listed below.

Reagents and bacterial strains

PCR reactions for cloning and cassette generation were performed using 2× Phanta UniFi Master Mix DNA Polymerase (Vazyme, P516-02). Colony PCR for sequencing validation employed Premix Taq DNA Polymerase (Takara, R901A). Homologous recombination cloning was carried out using the CloneExpress II One Step Cloning Kit (Vazyme, C112-02). Primers were synthesized by GENEWIZ (Suzhou, China). Gene sequences for R67 and engineered TadA variants were synthesized by GENERAL BIOL (Anhui, China).

Ampicillin sodium (Sangon Biotech, A100339-0025), chloramphenicol, and L-arabinose were purchased from commercial suppliers. Chemically competent *E. coli* DH5α and DH10B cells were purchased from AlpalifeBio (Beijing, China) and Biomed (Beijing, China), respectively.

Plasmid construction

pMuta088 was constructed using pDae079 (eMutaT7transition; Addgene #187622) as the backbone[2]. The vector backbone was amplified via PCR (primers: araBAD-FP and araC-

RP), and the resulting product was purified using a gel extraction kit. The purified fragment was ligated with T4 ligase and T4 polynucleotide kinase, then cloned into *E. coli* DH5α to generate the pMuta088 vector. The resulting pMuta088 carries the tandem PmCDA1-T7 RNA polymerase and uracil glycosylase inhibitor (UGI). Plasmids for expressing TadA variants and T7 RNA polymerase were constructed by replacing the PmCDA1 gene in the pMuta088 with engineered TadA sequences via homologous recombination. A negative control plasmid (pT7RNAP-ΔTadA), expressing only an Xten-linker–T7RNAP cassette, was constructed using the same strategy.

The TMP-resistance reporter plasmid (pReporter-R67^{W23*}*) was generated by inserting the R67 gene into a low-copy-number vector T7 promoter + terminators reporter (Addgene #156456) replacing the *neoR*/*kanR* cassette[3]. A premature stop codon (TGG→TAG; Trp23→Stop) was introduced by site-directed mutagenesis. TMP resistance is restored only upon TadA-mediated A→G editing at the target site.

Evaluation of editing frequency

Chemically competent *E. coli* DH10B cells were co-transformed with (i) the reporter plasmid (AmpR) and (ii) the TadA expression plasmid (CmR). Transformants were selected on LB agar containing 100 μg/mL ampicillin and 25 μg/mL chloramphenicol. Individual colonies were picked and inoculated directly into LB broth containing 100 μg/mL ampicillin, 25 μg/mL chloramphenicol, and 0.2% (w/v) L-arabinose, and incubated overnight (16 h) at 37℃ to initiate TadA induction and mutation accumulation. Subsequently, the cultures were subcultured (1:100 dilution) into fresh

induction medium (supplemented with the same antibiotics and arabinose) and grown for 4 h to fix mutations under constant agitation (37℃, 220 rpm).

At the endpoint, the editing frequency was determined as the ratio N_1 / N_0. Here, N_1 represents the total TMP-resistant colonies (extrapolated to the total culture volume from colony counts on 300 μL spread plates), and N_0 represents the total viable cells (extrapolated to the total culture volume from serial dilution spot assays). Plating for $N_1$was performed in triplicate.

Activity calculation

Endpoint editing frequencies were converted to per-base per-generation mutation rates using the Luria–Delbrück rare-mutation approximation. Under this model, the expected frequency satisfies:

$$\mathbb{E}[f] \approx \mu \ln R_{\text{eff}}$$

where μ is the mutation rate (per base per generation) and R_{eff} represents the effective population expansion. Although induction was maintained for 16 hours, the calculation was normalized to the effective population expansion of the final outgrowth step (one propagation cycle), as mutation fixation is replication-dependent. This 4-hour growth cycle consisted of a 1:100 reinoculation followed by regrowth to saturation, corresponding to approximately 6.6 generations (G). Assuming binary fission ($R_{\text{eff}} = 2^G$), the normalization factor is. $G \ln 2 \approx 4.57$.

Because TMP resistance restoration requires a single-base reversion, the effective target size was set to S = 1, and mutation rates were reported without normalization by the 192-bp reporter length. The site-specific per-base mutation rate was calculated as:

$$\mu_{\text{s.p.b.}} = \frac{f}{G \ln 2} \approx \frac{f}{4.57}$$

Verification of R67 reversion

To confirm that TMP resistance resulted from the intended A→G substitution, five independent TMP-resistant colonies per variant were analyzed by colony PCR and Sanger sequencing (GENEWIZ, Suzhou, China). R67 sequences were aligned against wild-type and mutant references to verify the codon 23 reversion and identify any additional substitutions.

Experimental Validation of CcdA Variants

Vector design and construction

The pUC57-Kan-ccdA/B vector was designed to co-express the $CcdA^{36–72}$ domain and CcdB in *E. coli*. The construct places the J23119 promoter driving $ccdA^{36–72}$ on the forward strand and the AmpR promoter driving ccdB on the reverse strand, with a 21-bp spacer between stop codons for PCR amplification. Both genes were codon-optimized for *E. coli* and synthesized by General Biosystems. Constructs were cloned into the PciI and NdeI restriction sites of the pUC57-Kan vector.

The pUC57-Kan-2BspQI-ccdB vector used for library construction was generated in DB3.1 competent cells using the pUC57-Kan-ccdA/Bas template. These two vectors were verified by Sanger sequencing and the sequences information are provided in Supplementary Tables S8.

Library construction, screening, and deep sequencing

Codon-optimized $CcdA^{36–72}$ variants were synthesized as an oligonucleotide pool flanked by BspQI restriction sites (GenScript, China). First, the oligo pool was amplified using PrimerSTAR HS DNA polymerase (Takara). The resulting product was then digested with BspQI and ligated into the linearized pUC57-Kan-2BspQI-ccdB vector using T4 DNA ligase (Takara). After purification, the ligation products were finally electroporated into E. coli DB3.1 competent cells using a Bio-Rad Micropulser. Transformants were recovered in LB medium at 37 °C for 1 h. Library size was estimated by serial dilution and plating on kanamycin-containing LB agar. The bulk culture was supplemented with kanamycin (50 μg/mL) and incubated for 10 h, followed by amplification and plasmid extraction. The unselected library was stored in LB medium containing 15% glycerol at −80 °C.

Plasmids extracted from DB3.1 cells constituted the unselected library and were subsequently electroporated into ccdB-sensitive DH5α cells for functional selection. Plasmids recovered from DH5α transformants constituted the selected library. The $CcdA^{36–72}$ region was PCR-amplified from both libraries using indexed primers (Supplementary Table S7), gel-purified, and sequenced on a Salus Pro platform (ShenZhen Salus Biomed Ltd.).

NGS preprocessing

Paired-end reads from unselected and selected libraries were merged and quality-filtered using PEAR and the FASTX Toolkit. Constant 5′ and 3′ flanking regions were trimmed, and high-quality merged reads were retained. Identical nucleotide sequences were

collapsed into unique records with associated read counts. Technical replicates were merged by summing counts for identical sequences.

Fitness estimation

Variant fitness was quantified using a Poisson-based $\log_2$ enrichment model. Effective frequencies before and after selection were computed using a pseudocount $\alpha = 1$:

$$p_i = \frac{n_i + \alpha}{\sum(n_i + \alpha)}$$

where n_i is the number of reads from NGS. After filtering the nucleotide sequences with 10^{-5} from both before and after, the fitness was defined as:

$$F_i = \log_2(p_{after,i}) - \log_2(p_{before,i})$$

Standard errors were estimated using a delta-method approximation. Variants containing internal in-frame stop codons served as negative controls to define a null distribution (median μ_0 and MAD-derived σ_0). One-sided *p*-values were computed and corrected for multiple hypothesis testing using the Benjamini–Hochberg procedure.

For two independent selection experiments, fitness and *q*-values were computed separately. Variants were classified as functional if they satisfied FDR ≤ 0.01 in both experiments, using the maximum *q*-value as the final significance measure. Fitness values were further normalized as:

$$F_i' = \frac{F_i - \mu_0}{\sigma_0}$$

to enable cross-replicates comparison.

In vivo activity assessment of the screened CcdA variants

Nine CcdA variants with different fitness values (Supplementary Table S4) were selected and cloned into the pUC57-Kan-ccdA/B vector. All sequences were synthesized and verified by Sanger sequencing at General Biol. For functional assays, 80 ng of each plasmid was transformed into the ccdB-sensitive E. coli strain DH5α. Transformants were serially diluted ten-fold, plated on LB agar containing kanamycin, and incubated at 37°C for 20 h. Colony-forming units (CFUs) were enumerated from dilution plates to assess the in vivo activity of each variant.

Supplementary References

1. Chen et al., “Navigating Protein Fitness Landscapes Through Simulated Evolutionary Jumps”, LangTaoSha Preprint Server. (2026).
2. Seo, Daeje, et al. "A dual gene-specific mutator system installs all transition mutations at similar frequencies in vivo." Nucleic Acids Research 51.10 (2023): e59-e59.
3. Moore, Christopher L., Louis J. Papa III, and Matthew D. Shoulders. "A processive protein chimera introduces mutations across defined DNA regions in vivo." Journal of the American Chemical Society 140.37 (2018): 11560-11564.