Evolution-guided diffusion generative model enables large-step exploration of functional protein sequence space from single sequences

Xing Zhang[1§], Jinle Tang[1§], Tingkai Zhang[1,2§], Zhihang Chen[3,4], Zhe Zhang[1], Jian Zhan[1,5,6]*, and Yaoqi Zhou[1]*

[1]Institute of Systems and Physical Biology, Shenzhen Bay Laboratory, Shenzhen, 518132, China

[2]School of Medicine, Southern University of Science and Technology, Shenzhen, 518055, China

[3]Shenzhen Medical Academy of Research and Translation (SMART), Shenzhen, 518107, China

[4]Tsinghua University, Beijing 100084, China

[5]Ribopeutic (Shenzhen) Co., Ltd., Shenzhen, 518000, China

[6]Ribopeutic Inc., Hangzhou, 310018, China

[§]co-first authors. These authors are contributed to the manuscript equally.

*Corresponding authors: Yaoqi Zhou, +86-(755) 2684 6275, zhouyq@szbl.ac.cn; Jian Zhan +86-(755) 2684 6275, zhanjian@szbl.ac.cn

Abstract

Protein evolution in nature and in the laboratory proceeds through incremental, largely undirected mutational steps, restricting exploration to local regions of sequence space and limiting access to remote yet potentially functional proteins. We present EvoGUD, a single-sequence–conditioned diffusion framework for large-step exploration of protein sequence space under learned evolutionary constraints. EvoGUD-generated sequences preserve natural-like co-evolutionary structure in representation space despite large sequence divergence. When assembled as virtual multiple sequence alignments, these sequences substantially improve AlphaFold3 single-sequence inference, restoring much of the backbone accuracy and atomic-level side-chain realism for recent deposited protein monomers as well as protein–nucleic-acid complexes, without evolutionary database search. Moreover, EvoGUD enables functional discovery in remote sequence space, yielding active variants of the adenine base-editing enzyme TadA in targeted validation experiments (80% success rate) and large numbers of functional variants of the intrinsically disordered antitoxin CcdA in high-throughput selection assays (19% success rate). Together, these results establish EvoGUD as a single-sequence, evolution-aware generative framework for large-step navigation of protein sequence space, with direct implications for structure modeling and functional protein discovery in previously unexplored sequence space.

Introduction

Protein evolution has generated an extraordinary diversity of molecular structures and biochemical functions, yet the mechanisms by which evolution explores protein sequence space are intrinsically constrained. Natural evolution proceeds through incremental, largely undirected mutations accumulated over long timescales, while laboratory directed evolution accelerates selection but still relies on random, local mutational steps[1,2]. Consequently, both processes tend to explore narrow neighborhoods around existing solutions, leading to uneven sampling of protein fitness landscapes and limited access to distant yet potentially functional regions of sequence space[3]. A central challenge in protein science is therefore to enable large-step exploration of protein sequence space—jumping to remote regions—while preserving the functional and structural constraints characteristic of a target protein family.

Computational approaches to this challenge broadly fall into structure-based and sequence-based strategies. Structure-centric methods have achieved remarkable success in *de novo* protein design, but their reliance on explicit backbone templates or structural hypotheses limits their applicability when reliable structures are unavailable, particularly for intrinsically disordered proteins[4,5]. Sequence-based protein language models provide an alternative by leveraging large-scale evolutionary data to generate functional sequences[6]; however, most operate in an unconditional or weakly conditioned regime and are not designed for targeted exploration of homolog families. Conditional sequence models and evolutionary augmentation approaches can expand sequence diversity but typically depend on pre-training with family-level labels, fine-tuning on known

homologs, or autoregressive sampling schemes that bias generation toward the statistical center of the training distribution[7–9].

Recent diffusion-based generative models offer a complementary paradigm by generating all residues simultaneously through iterative denoising, enabling more effective capture of global context and long-range dependencies than token-by-token autoregressive models[10,11]. In protein modeling, however, most conditional diffusion frameworks function primarily as evolutionary inpainting methods: they rely on a multiple sequence alignment (MSA) as input and recover missing information for the masked query sequence by interpolating within evolutionary boundaries defined by known homologs[12,13]. This inward generation regime benefits from dense evolutionary coordinates supplied by the MSA but inherently limits the novelty of generated sequences. By contrast, single-sequence exploration requires outward extrapolation—inferring latent co-evolutionary constraints from a solitary query sequence to generate a coherent, diverged homolog family *de novo*, without access to an existing alignment. Existing methods lack an explicit mechanism for this type of controlled, query-centric expansion.

Here we introduce EvoGUD (**Evo**lution-**gu**ided **D**iffusion), a single-sequence–conditioned diffusion framework for large-step exploration of protein sequence space. EvoGUD learns the statistical structure of natural homolog families by training on MSAs while conditioning exclusively on a single query sequence, enabling inference without homolog retrieval. A tunable conditioning strength (γ) controls the balance between exploratory breadth and adherence to learned evolutionary constraints, allowing direct generation of substantially diverged yet evolutionarily consistent homolog families

within a single generative process. As a result, EvoGUD-generated sequences preserve natural-like co-evolutionary structure in representation space despite large sequence divergence, closely tracking natural homologs and substantially exceeding identity-matched random controls.

When assembled as virtual MSAs (vMSAs), EvoGUD-generated sequences substantially improve AlphaFold3[14] single-sequence inference, restoring both backbone accuracy and atomic-level side-chain realism for most monomers and protein-nucleic acid complexes, without evolutionary database search. More importantly, EvoGUD enables functional discovery in remote sequence space, yielding active TadA variants (an adenine base-editing enzyme)[15] and large numbers of functional CcdA antitoxin variants (an intrinsically disordered protein interacted with the toxic protein CcdB)[16] under purely sequence-level conditioning. Together, these results establish EvoGUD as a single-sequence–based, evolution-aware generative framework for large-step exploration of protein sequence space under evolutionary constraints, enabling downstream structure modeling and functional discovery.

Results

Generating remote homologs from single protein sequences

EvoGUD was designed to bridge the gap between raw sequence space and the higher-order evolutionary manifold captured by protein language models[17]. Conceptually (Fig. 1a), both natural evolution and conventional laboratory directed evolution rely on small, largely undirected mutational steps, and therefore explore only a limited local neighborhood around a starting sequence[1]. EvoGUD instead enables direct generation of remote homologs by operating in a feature space defined by ESM-2, conditioning sequence generation on the query's per-residue embeddings and pairwise attention patterns[17].

Concretely, EvoGUD was trained as a denoising diffusion model[10,18] to reconstruct natural MSAs (nMSAs) from noise while observing only query-derived ESM-2 features (Fig. 1b). During sampling, the model starts from a random amino-acid probability distribution and iteratively denoises it under conditioning of single query sequence (Fig. 1c). A probability absorption step blends the evolving distribution with the model-predicted denoised distribution, and a single scalar parameter, the conditioning strength γ, modulates the extent to which sampling remains anchored to the conditioning query sequence. As a result, γ provides an explicit and continuous control over the distance of generated sequences from the query sequence in sequence space.

We first quantified how the conditioning strength γ controls sampling distance by measuring the sequence identity between generated sequences and their corresponding query sequences using an independent test set of 159 proteins. This dataset was

constructed from recently released PDB monomers[19] and filtered to be non-redundant with respect to both the training data and within the set itself, applying a 40% sequence-identity cutoff, denoted as RecentPDB-monomer. (see Methods). Across a wide range of γ values, EvoGUD produces smoothly tunable identity distributions, with increasing γ leading to progressively higher similarity to the query sequence (Fig. 1d), along with improved attention similarity (Supplementary Fig. S1b), increased foldability according to pTM from ESMfold (Supplementary Fig. S1c), reduction of diversity according to intra-set sequence identity (Supplementary Fig. S1d) and decreased novelty according to the maximum sequence identity to nMSA sequences (Supplementary Fig. S1f). This demonstrates that EvoGUD does not rely on a fixed exploration regime, but instead enables controlled interpolation between aggressive exploration and conservative refinement.

To further assess whether EvoGUD preserves evolutionary features beyond simple sequence identity shown in Fig. 1d, we compared nMSA sequences, EvoGUD-generated sequences, and random sequences to the query sequence in the representation space of ESM-2, according to cosine similarity between ESM-2 attention maps by using the RecentPDB-monomer set. As shown in Fig. 1e, cosine similarity of ESM-2 attention of EvoGUD-generated sequences to the query sequence closely tracks the similarity trajectory of natural homologs (natural MSAs) and remain substantially higher than the similarity of random controls across all identity regimes. Notably, even at low sequence identity (< 0.3), EvoGUD maintains a median attention-space similarity of 0.874, compared to 0.578 for random sequences, indicating that EvoGUD samples realistic regions of the evolutionary manifold rather than merely matching identity statistics.

Comparable trends were observed using ESM-2 embedding similarity instead of attention (Supplementary Fig. S2). Together, these results show that EvoGUD enables controllable generation of structurally plausible remote homologs while preserving higher-order evolutionary constraints, with γ acting as an explicit knob to balance exploration and constraint.

Monomeric structure prediction with vMSA from EvoGUD

We next asked whether vMSAs generated by EvoGUD can replace nMSAs for monomeric protein structure prediction with AlphaFold3[14] under single-sequence conditions. We employed EvoGUD to generate vMSAs across a range of conditioning strengths γ and vMSA depths, and the resulting structures were predicted using AlphaFold3 in single-sequence mode (AF3-SS) using an ensemble of 10 EvoGUD parameter settings (See Methods).

EvoGUD + AF3-SS substantially outperformed AF3-SS on the RecentPDB-monomer test set (Fig. 2a). The mean TM-score[20] increased from 0.476 for AF3-SS to 0.795 for EvoGUD + AF3-SS, approaching the performance of ESMFold (0.827) and AF3 with nMSAs and PDB templates (0.884). The fraction of targets with TM-score ≥ 0.5—a commonly used criterion for correct fold assignment—increased from 40.9% for AF3-SS to 89.3% for EvoGUD + AF3-SS, comparable to ESMFold (92.4%) and approaching AF3 with nMSAs and templates (96.2%). These results indicate that most monomers can be modeled at near-native resolution from a single input sequence using EvoGUD.

To assess generalization beyond RecentPDB-monomer, we evaluated EvoGUD on the CASP15[21] monomer benchmark (N = 71), which was not used during parameter

selection. EvoGUD + AF3-SS again markedly improved backbone accuracy relative to AF3-SS and yielded TM-score distributions comparable to those observed on RecentPDB-monomer (Fig. 2b), demonstrating that the selected vMSA settings generalize to an independent community benchmark without re-tuning. On CASP15, the mean TM-score increased from 0.403 for AF3-SS to 0.593 for EvoGUD + AF3-SS, approaching the performance of ESMFold (0.640) and AF3 with nMSAs and PDB templates (0.740). The fraction of targets with TM-score ≥ 0.5 increased from 25.4% for AF3-SS to 64.8% for EvoGUD + AF3-SS, comparable to ESMFold (66.2%) and approaching AF3 with nMSAs and PDB structure templates (77.5%). A representative example is shown in Fig. 2c, where EvoGUD + AF3-SS recovers the correct overall fold and domain arrangement that is missed by AF3-SS.

Because AlphaFold3 is an all-atom diffusion model, we further evaluated the physical plausibility of the predicted structures by quantifying steric clashes between heavy atoms. ESMFold, despite its strong backbone accuracy, exhibited substantially higher clash counts, with a broad tail of severely mispacked models on both RecentPDB-monomer and CASP15 (Fig. 2d and 2e). In contrast, EvoGUD + AF3-SS markedly reduced steric clashes relative to ESMFold, producing distributions much closer to those of AF3. A representative local side-chain environment is shown in Fig. 2f, where EvoGUD + AF3-SS yields well-packed, stereochemically reasonable side chains, whereas ESMFold displays strained rotamers and steric overlaps[22].

Together, these results show that EvoGUD enables single-sequence AlphaFold3 to recover both high backbone accuracy and AF3-like all-atom realism for monomeric proteins, without requiring time-consuming searches for natural homologs. We next

asked whether similar gains extend beyond monomers to multimeric and protein–nucleic-acid complexes.

Protein-NA complex prediction with vMSA from EvoGUD

We next evaluated whether EvoGUD can extend single-sequence prediction beyond monomeric proteins to multi-chain assemblies, focusing on protein–nucleic-acid (protein–NA) complexes (Fig. 3). We curated a benchmark of 165 experimentally determined protein–DNA/RNA complexes and generated vMSAs for the protein chains only, while keeping the nucleic-acid sequences fixed. These sequences were assembled into 10 sets of vMSA and supplied, together with the original NA sequence, to AF3-SS. For each target we selected the final EvoGUD + AF3-SS model by the highest ipTM confidence score. As shown in Fig. 3a, EvoGUD + AF3-SS substantially improved the accuracy of the protein subunits within the protein–NA complexes (median TM-score = 0.872) relative to AF3-SS (median TM-score = 0.436), close to the distribution of AF3 (median TM-score = 0.928), as in the case of monomeric proteins.

For complex structures evaluated by interface local distance difference test (iLDDT[23]) (as in AF3[14]), AF3-SS frequently produced mis-docked subunits and distorted protein–nucleic-acid contacts, with most targets exhibiting low iLDDT values below 0.4 (Fig. 3b). In contrast, EvoGUD + AF3-SS markedly shifted the distribution toward higher iLDDT values, correctly recovering 55 well-docked complexes out of 90 that are correctly predicted by AF3 using nMSAs for both protein and nucleic acid sequences together with PDB templates—substantially exceeding the 14 complexes recovered by AF3-SS. These results indicate that vMSAs generated from single sequences can

effectively guide accurate docking of protein chains onto DNA and RNA for most cases studied.

To visualize these trends on a per-complex basis, we compared the change in TM-score and iLDDT relative to AF3-SS for each method (Fig. 3c). Most EvoGUD + AF3-SS predictions fall in the quadrant corresponding to simultaneous gains in subunit accuracy and interface quality, demonstrating that vMSAs derived from single sequences can improve both global folding and interfacial organization within the same model.

One structural example illustrates these effects at the level of individual assemblies (Fig. 3d). In the shown protein–NA complex (PDB: 7U7C)[24], AF3-SS fails to correctly position protein subunits relative to the nucleic acid and produces poorly resolved interfaces. In contrast, EvoGUD + AF3-SS accurately recovers the overall architecture and correctly docks the protein chains onto the nucleic-acid scaffold, yielding an interface geometry that closely matches the experimental structure and approaches the AF3 baseline. Thus, vMSAs from EvoGUD can capture the evolution signals not only in monomeric structures but also in interfacial structures for docking.

Locating remote functional enzymes by EvoGUD

To test whether EvoGUD can recover functional enzymes from remote regions of sequence space, we selected TadA, a bacterial tRNA adenosine deaminase that has been repurposed through extensive directed evolution into the catalytic core of adenine base editors (ABEs), enabling programmable A•T→G•C DNA conversion[15]. Although highly effective, previously reported TadA variants remain closely related to their ancestral sequences, motivating exploration of more distant sequence solutions[25,26].

Based on the joint behavior of sequence identity, foldability, diversity, and novelty (Supplementary Fig. S3), we selected a conditioning strength of $\gamma = 2$ and generated 1,024 sequences conditioned on the wild type TadA from *Staphylococcus aureus*[27]. Ten representative TadA variants were selected for experimental evaluation (see Methods). Structural evaluation using AlphaFold3 (AF3) showed that providing EvoGUD-generated sequences as vMSAs substantially improved structure prediction compared with single-sequence input, yielding well-folded TadA-like architectures with markedly higher pTM confidence score and TM-score (shown in Fig. 4a).

The ten EvoGUD-generated variants were then evaluated with a trimethoprim (TMP) resistance reversion assay in *E. coli*, in which TadA-mediated A→G editing restores a functional *R67* dihydrofolate reductase (DHFR) reporter gene and as a result, a stronger active TadA variant will grow more colonies (Fig. 4b). Some examples are shown in Fig. 4c. Eight of the ten variants restored TMP resistance, demonstrating robust A•T→G•C DNA-editing activity despite low sequence identity to both the query sequence (0.38–0.40) and any known TadA homologs (0.57–0.62). Quantification of editing activity was made according to the number of colonies before and after TMP selection (see Methods). It revealed a reproducible range of activities across variants, with the activity of the wild type TadA falling within the distribution of EvoGUD-generated sequences as shown in Fig. 4d. Thus, EvoGUD can recover functional TadA DNA-editing enzymes from previously unexplored, remote regions of sequence space.

Application to the intrinsically disordered antitoxin CcdA

To further test the limits of EvoGUD, we examined whether it could generate functional variants of the intrinsically disordered antitoxin CcdA. CcdA lacks a stable structure in isolation and acquires its functional conformation only upon binding its cognate toxin CcdB, posing a stringent challenge for protein design[16,28].

Monomeric CcdA and CcdB sequences were derived from the *E. coli* CcdA–CcdB complex (PDB: 3HPW)[16]. These sequences were then assembled into a symmetrized CcdB–G_{50}–CcdA–G_{50}–CcdB fusion construct. The long flexible link was specifically designed as the query for EvoGUD-based CcdA generation under CcdB conditioning. Based on the joint behavior of identity, foldability, diversity, and novelty (Supplementary Fig. S4), we selected a conditioning strength of $\gamma = 2$ for downstream experiments and generated a pooled library of 5,623 unique CcdA variants for experimental screening (see Methods).

For structural verification, only a vMSA constructed from EvoGUD-generated CcdA sequences was inputted into AlphaFold3, without using any nMSAs or vMSA for CcdB. A representative subset of eight CcdA variants generated at $\gamma = 2$ was selected to form the vMSA, as prediction accuracy decreased when larger numbers of low-γ sequences (> 64) were included (Supplementary Fig. S5). As shown in Fig. 5a, AF3 predictions using this CcdA-only vMSA recover a coherent CcdA–CcdB complex. Relative to the experimentally determined *E. coli* CcdA–CcdB structure, the predicted model shows high agreement at both the subunit and interface levels, as reflected by elevated pTM, ipTM, TM-score, and iLDDT values. Notably, despite containing sequences for CcdA only, the vMSA improves the predicted structures of both the CcdA antitoxin and the flanking

CcdB toxin subunits, as well as their binding interface. In contrast, AF3 single-sequence inference fails to recover the CcdA structure or its docked configuration within CcdB, indicating that EvoGUD-generated CcdA sequence ensembles provide sufficient context for accurate prediction of the full toxin–antitoxin complex, without natural or virtual MSA for CcdB.

The variant library was subjected to an experimental CcdB toxin selection assay, and variant counts were obtained by high-throughput sequencing before and after selection in two independent biological replicates (Fig. 5b). Fitness was inferred from before/after enrichment using early-stop variants as internal negative controls, followed by Benjamini–Hochberg false discovery rate (BH-FDR)[29] filtering ($q \leq 0.01$) and normalization (see Methods). Across the two replicates, 1,110 and 1,153 variants passed the survival test, respectively. Requiring consistent enrichment in both experiments identified 1,072 functional variants, corresponding to an overall success rate of 19% (1,072 of 5,623 variants). Both replicates show a substantial population of variants with normalized $\log_2$ fitness exceeding that of wild-type CcdA (*E. coli* CcdA, or EcCcdA, Fig. 5c), and inferred fitness values for the same variants are highly correlated between replicates (Fig. 5d), enabling robust ranking of functional variants.

To experimentally validate the statistical classification, we evaluated a subset of nine individual variants spanning the fitness range using spot survival assays (Supplementary Fig. S6 and S7). These variants (EvoGUD generated CcdA, denoted as egCcdA) were chosen according to their rank based on average normalized $\log_2$ fitness across replicates between 3 and 18. (egCcdA-1 denotes the highest-ranked variant). Across both plate experiments, all nine tested variants exhibited detectable protection from CcdB toxicity.

No growth was observed in the CcdB-only control, demonstrating that the enrichment in the high-throughput pooled assay reflects genuine antitoxin function. As illustrative examples, a representative subset of five variants is shown in the main text (Fig. 5c and 5e), chosen to illustrate the correspondence between inferred fitness and phenotypic strength across a wide dynamic range. High-ranked variants such as egCcdA-1 and egCcdA-14 displayed activity comparable to or exceeding that of wild-type CcdA at different dilutions (y-axis), whereas egCcdA-78 showed moderately reduced activity, consistent with its lower inferred fitness. As a variant positioned near the statistical decision boundary, egCcdA-933 still exhibited weak but detectable rescue relative to the negative control (Fig. 5f), validating the sensitivity of the fitness-based classification. Thus, EvoGUD can generate large numbers of functional CcdA variants under purely sequence-level conditioning, despite the absence of a stable ground-state fold. Unedited plate images for all tested variants are provided in Supplementary Figs. S6 and S7, ensuring full transparency of the experimental results.

Discussion

EvoGUD was designed to enable large-step exploration of protein sequence space while preserving the higher-order evolutionary constraints that characterize natural protein families. A central finding of this work is that EvoGUD-generated sequences remain embedded within realistic evolutionary manifolds despite substantial sequence divergence. In ESM-2 embedding and attention spaces, generated sequences closely track natural homologs and remain far more consistent with the query in representation space than identity-matched random controls, indicating that EvoGUD captures co-evolutionary structure beyond residue-level conservation. The conditioning strength γ provides continuous control over the balance between exploratory breadth and evolutionary constraint.

These properties translate directly into improved structure prediction. By assembling EvoGUD-generated sequences as vMSAs, AlphaFold3 single-sequence inference recovers much of the accuracy and atomic detail typically associated with natural homolog searches. Across monomer benchmarks, EvoGUD-assisted predictions are near standard AlphaFold3 performance. For protein–DNA and protein–RNA complexes, EvoGUD-derived vMSAs further improve interface geometry, demonstrating that the generated sequence families encode actionable co-evolutionary signals for multimeric recognition.

Besides structure modeling, EvoGUD provides a general framework for functional protein discovery in remote sequence space. For the TadA enzyme, EvoGUD identified highly divergent yet functional DNA-editing variants, revealing functional solutions inaccessible to stepwise directed evolution. Notably, these active variants reside at

approximately 40% sequence identity to the query, corresponding—in natural evolutionary terms—to divergence accumulated over on the order of ~10^9 years (Supplementary Table S4)[30]. Similarly, in the CcdA–CcdB toxin–antitoxin system, EvoGUD enabled large-scale discovery of functional antitoxin variants spanning 54–73% sequence identity, even among low-ranked candidates, indicating robust preservation of context-dependent functional constraints without explicit structural or biophysical scoring.

A key requirement for sequence generative models is generalizability across protein families and evolutionary distances. EvoGUD exhibits robust generalization across multiple validation regimes, including stringent non-redundant subsets, with generated sequences consistently following the same representation-space trajectories as natural homologs irrespective of training-set proximity (Supplementary Fig. S8). Controlled experiments on TadA further show that excluding or retaining close homologs during training produces only minor shifts in generative behavior without evidence of collapse (Supplementary Fig. S9). Notably, the CcdA system provides complementary insight: although CcdA homologs were present in the training data, CcdA-only conditioning yielded low identity and reduced foldability, whereas introducing an unseen fusion context with its cognate binding partner systematically shifted generation toward functionally coherent sequence space (Supplementary Fig. S10). Together, these results indicate that EvoGUD does not rely on memorization of training sequences but is primarily shaped by the evolutionary constraints supplied at inference time.

EvoGUD adopts a modular alternative to end-to-end single-sequence structure prediction pipelines. By decoupling protein sequence feature extraction from structure inference

with intermediate sequence generation, improvements in protein language models or structure predictors can be incorporated by retraining only a lightweight generative adapter, requiring on the order of a single GPU-day. This contrasts sharply with the hundreds of GPU-weeks typically required to train or adapt full-scale structure prediction models[17], enabling rapid iteration and reuse of advances in representation learning. Although EvoGUD's performance is bounded by the fidelity of underlying sequence representations, its modular design provides a scalable and extensible framework for integrating advances in representation learning, enabling large-scale evolutionary exploration that can be naturally combined with local refinement strategies such as directed evolution. In this way, EvoGUD bridges global sequence-space exploration and functional protein engineering while remaining grounded in evolutionary realism.

Methods

Training Dataset and Data Preprocessing

EvoGUD was trained on OpenProteinSet-PDB[31], a curated reconstruction of the AlphaFold2 training dataset as implemented in OpenFold[32], comprising 131,487 protein chains with precomputed MSAs. To ensure sequence integrity and computational consistency, chains containing unknown amino acids ("X") were excluded. Sequences were further filtered by length, retaining chains with $30 \leq L \leq 1000$ amino acids. After filtering, the final dataset comprised 117,556 entries, which were partitioned into a training set of 116,756 entries and a validation set of 800 entries.

Model architecture and conditioning

EvoGUD is built on a Diffusion Transformer (DiT) backbone[33] with a modified adaLN-Zero conditioning mechanism (Supplementary Fig. S11). Whereas the original DiT conditions on global features, EvoGUD incorporates sequence-specific evolutionary context derived from a protein language model.

During both training and inference, EvoGUD conditions the denoising process on representations extracted from the ESM-2 3B model, including per-residue embeddings (2,560 dimensions), and pairwise attention maps (36 layers × 40 heads; 1,440 dimensions)[17]. These features are linearly projected into a 128-dimensional latent space. In each DiT block, ESM-2 embeddings are injected via cross-attention as key–value pairs, while projected attention maps are added to the attention logits as a pairwise bias. The resulting representations generate the six modulation parameters (shift, scale, and

gate for attention and feed-forward sublayers) used in the adaLN-Zero operation, enabling position-wise, evolution-aware modulation of the network.

The model comprises 6 DiT blocks with a hidden dimension of 128 (3.06 M parameters total). A global dropout rate of 0.1 was applied. To stabilize early training, the linear layers producing adaLN modulation parameters were zero-initialized, such that each DiT block initially behaves as an identity mapping. Model outputs are projected to a categorical distribution over a 21-token alphabet (20 amino acids plus a gap token).

Diffusion formulation

EvoGUD adopts a continuous-time Gaussian diffusion framework with a cosine noise schedule[34]. The forward process gradually perturbs one-hot encoded amino-acid sequences with Gaussian noise according to a cumulative signal retention coefficient $\bar{\alpha}_t$, defined over normalized time $u \in [0,1]$ as:

$$\bar{a}_u = \frac{f(u)}{f(0)}, \qquad f(u) = \cos^2\left(\frac{\frac{u}{s}+1}{1+s} \cdot \frac{\pi}{2}\right)$$

where s is a small offset to prevent the noise level from becoming too small at $t = 0$. The schedule was discretized into $T = 100$ steps for training. The model is trained to predict the original categorical distribution of clean sequences from noisy inputs, conditioned on ESM-2 features.

Model Training

EvoGUD was trained to reconstruct natural homolog sequences drawn from MSAs using a query-centric batching strategy. For each optimization step, a single query sequence

was paired with 64 target sequences sampled from its associated MSA. For shallow MSAs, target sequences were oversampled; for deep MSAs, 64 members were randomly subsampled. The query sequence was always included in the batch to anchor reconstruction.

Each target sequence was one-hot encoded and independently corrupted by Gaussian noise at a randomly sampled timestep $t_i \in [1, T]$. The model predicted the denoised categorical distribution $p(x_0 | x_{t,i}, cond)$ over the 21-token alphabet.

To emphasize evolutionary diversity rather than trivial sequence conservation, a per-residue weighted cross-entropy loss was applied: positions differing from the query were assigned weight = 1.0, whereas positions identical to the query or corresponding to gaps were down-weighted (weight = 0.1).

Training was performed for 100 epochs using the AdamW optimizer[35] with a learning rate of 1×10^{-3} and automatic mixed precision. Each epoch comprised 10,000 unique queries sampled from the whole training set, totaling ~10^6 optimization steps. Training required approximately 25 h on a single NVIDIA A100 GPU and was implemented in PyTorch 2.2[36].

Single-sequence conditional sampling via probability absorption

At inference time, EvoGUD generates homolog families from a single query sequence using a probability absorption sampling scheme that bridges continuous diffusion dynamics with discrete sequence space.

Sampling begins from isotropic Gaussian noise $x_T \sim \mathcal{N}(0, I)$ and proceeds over T = 100 discrete reverse-diffusion steps. At each step, the model predicts a position-wise amino-

444 acid probability distribution conditioned on ESM-2 features of the query. A discrete
445 sequence is obtained by deterministic argmax decoding, excluding the gap token to
446 generate gap-free sequences.
447 The sampled sequence is projected back into latent space as a centered one-hot
448 representation, scaled by a conditioning strength parameter γ:

449 $$\hat{x}_0 = \gamma \cdot (\text{one\_hot}(S^0) - 0.5)$$

450 The latent state is then updated via the Gaussian reverse transition, blending the absorbed
451 identity signal with stochastic noise. The parameter γ controls the trade-off between
452 evolutionary adherence and exploratory breadth: higher values promote conservative,
453 high-confidence homologs, whereas lower values allow broader exploration of remote
454 sequence space.
455
456 **Benchmark test sets**
457 **RecentPDB-monomer.** To evaluate structure prediction performance, an independent
458 test set was curated from PDB entries released between January 1 and July 1, 2024,
459 following AlphaFold3 benchmarking principles[14]. Only protein-only monomers solved by
460 X-ray crystallography at ≤ 2.0 Å resolution were retained. Chains were filtered to lengths
461 of 30–500 residues and subjected to 40% sequence-identity filtering both within the set
462 and against the training/validation data. The final set comprised 159 non-redundant
463 monomers.
464 **RecentPDB-multimer.** Protein–nucleic-acid complexes were derived from the
465 AlphaFold3 benchmark dataset[14]. After excluding entries lacking protein chains or

containing non-canonical residues, the final set comprised 165 complexes. vMSAs were generated only for protein components, while nucleic-acid sequences were held fixed.

Ensemble settings for EvoGUD + AF3-SS

A grid search over conditioning strength γ and vMSA depth on the RecentPDB-monomer benchmark set revealed a broad operating regime in which EvoGUD-derived vMSAs substantially improved structure prediction accuracy relative to AF3-SS (Supplementary Fig. S5). Within this regime, increasing γ and vMSA depth (# of generated vMSA sequences) led to substantial gains in both predicted confidence (pTM[14]) and backbone accuracy (TM-score), indicating that vMSAs capture much of the evolutionary information normally supplied by natural homologs. Based on this analysis, we selected a small ensemble of 10 EvoGUD parameter settings spanning this robust regime ($\gamma = 1$ with #vMSA $\in \{2,4,8,16\}$ and $\gamma = 2$ with #vMSA $\in \{2,4,8,16,32,64\}$) and fixed these settings for all subsequent analyses.

Generation and validation of TadA variants

Wild-type *Staphylococcus aureus* TadA (PDB: 2B3J)[27] was used as the query for EvoGUD sequence generation. Conditioning strength γ was selected based on predicted foldability, novelty, and diversity ($\gamma = 2$; Supplementary Fig. S3). A total of 1,024 TadA variants were generated under a co-generation identity constraint of $\geqslant$ 35% relative to wild-type TadA, subsequently clustered at 70% sequence identity, and ten representatives from the largest clusters were selected for experimental validation.

TadA activity was quantified using a trimethoprim resistance reversion assay in *E. coli* based on a premature stop-codon reporter. Editing activity was defined as per-base mutation rates estimated from the observed frequency of TMP-resistant colonies, using a Luria–Delbrück approximation[37]. Full experimental protocols are provided in the Supplementary Information.

Generation and validation of CcdA variants

CcdA variants were generated using a conditional single-chain strategy in which the antitoxin sequence was embedded within a symmetrized CcdB–G_{50}–CcdA–G_{50}–CcdB fusion, with wild type CcdA and CcdB sequences from *E. coli* (PDB: 3HPW)[16]. Here, G_{50} denotes a 50-amino-acid poly-glycine linker that spatially separates the CcdA and CcdB domains while preserving sequence-level context[38]. Residues outside the CcdA region were fixed during sampling. Conditioning strength $\gamma = 2$ was selected according to predicted foldability, novelty, and diversity (Supplementary Fig. S4).

A pooled library of 10,000 CcdA variants was generated under a co-generation identity constraint of ⩽ 75% relative to wild-type CcdA and subsequently clustered at 90% sequence identity, yielding 5,623 unique sequences. Functional selection was performed using a toxin-rescue assay in *E. coli*, followed by deep sequencing.

Variant fitness was estimated using a Poisson-based $\log_2$ enrichment model and false-discovery-rate (FDR) correction[29]. Variants passing FDR ≤ 0.01 in two independent experiments were considered functional. Full experimental protocols are provided in the Supplementary Information.

Reference

1. Arnold, F. H. Design by directed evolution. *Accounts of chemical research* **31**, 125–131 (1998).
2. Romero, P. A. & Arnold, F. H. Exploring protein fitness landscapes by directed evolution. *Nature reviews Molecular cell biology* **10**, 866–876 (2009).
3. Poelwijk, F. J., Tănase-Nicola, S., Kiviet, D. J. & Tans, S. J. Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes. *Journal of theoretical biology* **272**, 141–144 (2011).
4. Huang, P.-S., Boyken, S. E. & Baker, D. The coming of age of de novo protein design. *Nature* **537**, 320–327 (2016).
5. Dauparas, J. *et al.* Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).
6. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118**, e2016239118 (2021).
7. Madani, A. *et al.* Large language models generate functional protein sequences across diverse families. *Nature biotechnology* **41**, 1099–1106 (2023).
8. Nijkamp, E., Ruffolo, J. A., Weinstein, E. N., Naik, N. & Madani, A. Progen2: exploring the boundaries of protein language models. *Cell systems* **14**, 968-978. e3 (2023).
9. Zhang, J. *et al.* Unsupervisedly prompting AlphaFold2 for accurate few-shot protein structure prediction. *Journal of Chemical Theory and Computation* **19**, 8460–8471 (2023).

10. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020).

11. Trippe, B. L. *et al.* Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. *arXiv preprint arXiv:2206.04119* (2022).

12. Alamdari, S. *et al.* Protein generation with evolutionary diffusion: sequence is all you need. *BioRxiv* 2023.09. 11.556673 (2023).

13. Sgarbossa, D., Lupo, U. & Bitbol, A.-F. Generative power of a protein language model trained on multiple sequence alignments. *Elife* **12**, e79854 (2023).

14. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

15. Gaudelli, N. M. *et al.* Programmable base editing of A• T to G• C in genomic DNA without DNA cleavage. *Nature* **551**, 464–471 (2017).

16. De Jonge, N. *et al.* Rejuvenation of CcdB-poisoned gyrase by an intrinsically disordered protein domain. *Molecular cell* **35**, 154–163 (2009).

17. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

18. Austin, J., Johnson, D. D., Ho, J., Tarlow, D. & Van Den Berg, R. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems* **34**, 17981–17993 (2021).

19. Berman, H. M. *et al.* The protein data bank. *Nucleic acids research* **28**, 235–242 (2000).

554 20. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein
555 structure template quality. *Proteins: Structure, Function, and Bioinformatics* **57**, 702–710
556 (2004).
557 21. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment
558 of methods of protein structure prediction (CASP)—Round XV. *Proteins: Structure,*
559 *Function, and Bioinformatics* **91**, 1539–1549 (2023).
560 22. Williams, C. J. *et al.* MolProbity: more and better reference data for improved all-atom
561 structure validation. *Protein Science* **27**, 293–315 (2018).
562 23. Mariani, V., Biasini, M., Barbato, A. & Schwede, T. lDDT: a local superposition-free
563 score for comparing protein structures and models using distance difference tests.
564 *Bioinformatics* **29**, 2722–2728 (2013).
565 24. Chang, C., Lee Luo, C. & Gao, Y. In crystallo observation of three metal ion promoted
566 DNA polymerase misincorporation. *Nature Communications* **13**, 2346 (2022).
567 25. Richter, M. F. *et al.* Phage-assisted evolution of an adenine base editor with improved
568 Cas domain compatibility and activity. *Nature biotechnology* **38**, 883–891 (2020).
569 26. Lapinaite, A. *et al.* DNA capture by a CRISPR-Cas9–guided adenine base editor.
570 *Science* **369**, 566–571 (2020).
571 27. Losey, H. C., Ruthenburg, A. J. & Verdine, G. L. Crystal structure of Staphylococcus
572 aureus tRNA adenosine deaminase TadA in complex with RNA. *Nature structural &*
573 *molecular biology* **13**, 153–159 (2006).
574 28. Loris, R. *et al.* Crystal structure of CcdB, a topoisomerase poison from E. coli. *Journal*
575 *of molecular biology* **285**, 1667–1677 (1999).

29. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57**, 289–300 (1995).

30. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: a resource for timelines, timetrees, and divergence times. *Molecular biology and evolution* **34**, 1812–1819 (2017).

31. Ahdritz, G. *et al.* OpenProteinSet: Training data for structural biology at scale. *Advances in Neural Information Processing Systems* **36**, 4597–4609 (2023).

32. Ahdritz, G. *et al.* OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *Nature methods* **21**, 1514–1524 (2024).

33. Peebles, W. & Xie, S. Scalable diffusion models with transformers. in *Proceedings of the IEEE/CVF international conference on computer vision* 4195–4205 (2023).

34. Nichol, A. Q. & Dhariwal, P. Improved denoising diffusion probabilistic models. in *International conference on machine learning* 8162–8171 (PMLR, 2021).

35. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

36. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32**, (2019).

37. Luria, S. E. & Delbrück, M. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* **28**, 491 (1943).

38. Chen, X., Zaro, J. L. & Shen, W.-C. Fusion protein linkers: property, design and functionality. *Advanced drug delivery reviews* **65**, 1357–1369 (2013).

Figures

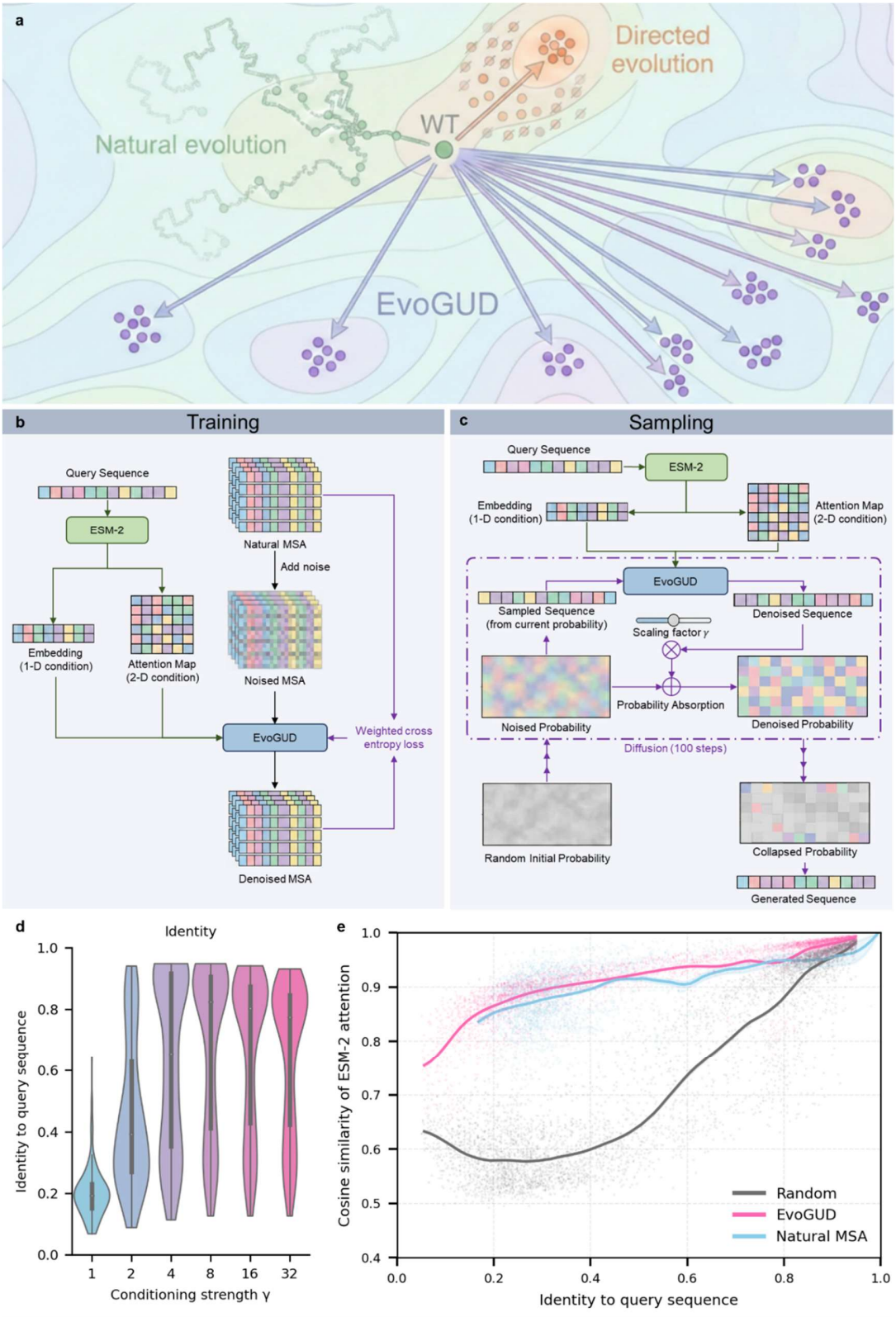


Figure 1. EvoGUD enables controllable, large-step exploration of protein sequence space from a single query.

a, Conceptual landscape illustrating the limitations of natural evolution and laboratory directed evolution, and how EvoGUD enables large-step sampling toward remote functional sequence regions.

b, Training of EvoGUD using nMSAs: the model is trained to denoise corrupted MSAs conditioned on query-derived ESM-2 embeddings and pairwise attention maps.

c, Sampling procedure: starting from a random probability distribution, EvoGUD iteratively denoises sequence probabilities under query conditioning, with a scaling factor γ controlling the strength of probability absorption before collapsing to a discrete sequence.

d, Distribution of sequence identity to the conditioning query as a function of γ, demonstrating tunable control over sampling distance in sequence space. Additional validation metrics are shown in Supplementary Figures S1.

e, Evolutionary consistency beyond identity: cosine similarity of ESM-2 attention maps versus sequence identity for EvoGUD-generated sequences, compared with identity-matched random controls and natural MSA (nMSA) homologs, demonstrating preservation of natural-like co-evolutionary geometry across divergent regimes. 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. S1b. Natural MSA sequences were included only when covering at least 80% of query positions to ensure comparable

623 alignment context. Solid lines denote mean trends and shaded regions indicate 95%

624 confidence intervals.

625

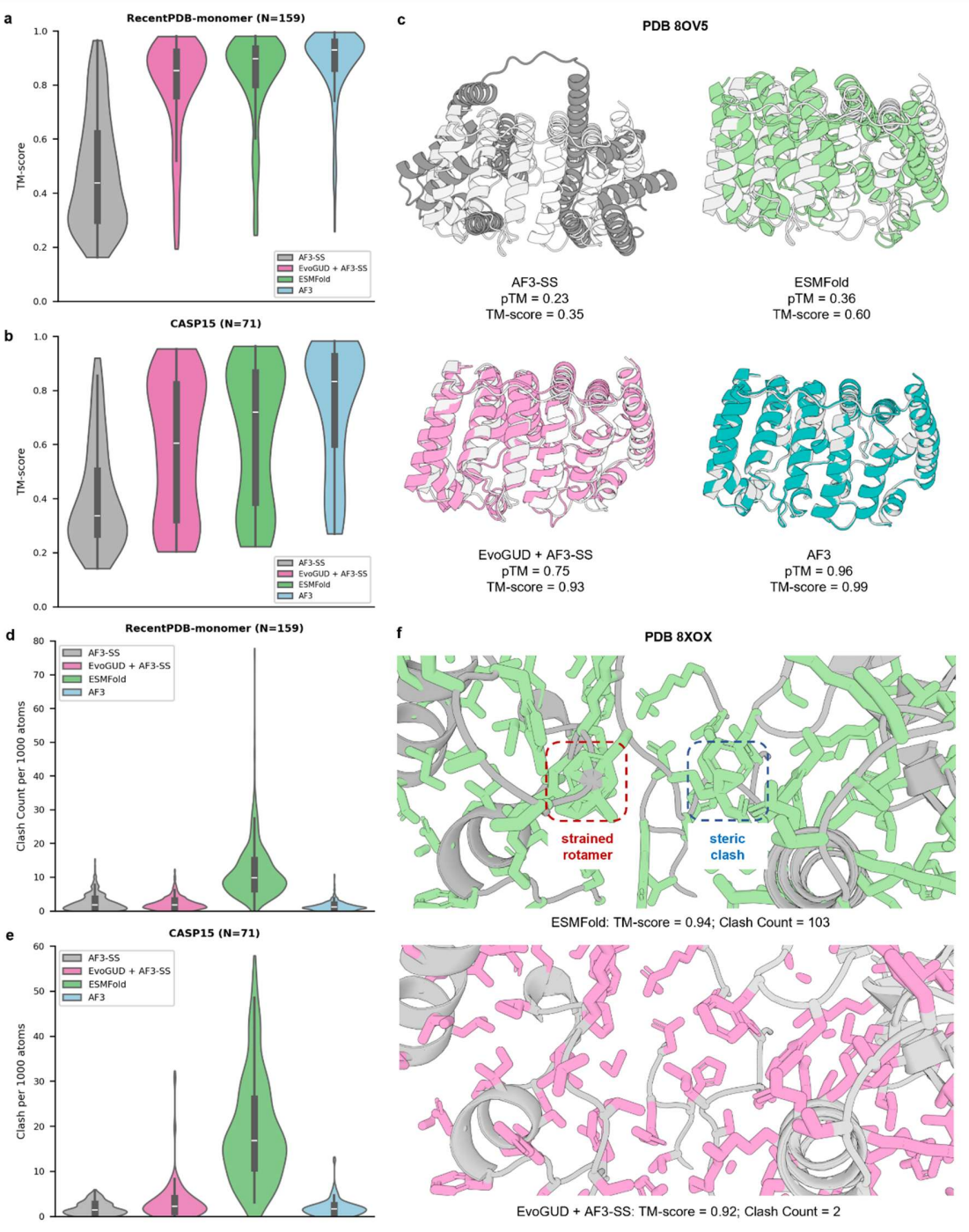


Figure 2. EvoGUD restores MSA-level monomer performance and improves all-atom quality from a single sequence.

a, TM-score distributions on the RecentPDB-monomer test set (N = 159), evaluated using the selected EvoGUD ensemble settings (Supplementary Fig. S5). Predictions from AF3-SS (single sequence), EvoGUD + AF3-SS, ESMFold, and AF3 with nMSAs are compared. EvoGUD substantially improves backbone accuracy relative to AF3-SS and approaches the performance of ESMFold and AF3.

b, TM-score distributions on the CASP15 monomer benchmark (N = 71), evaluated using the same EvoGUD ensemble settings selected on RecentPDB-monomer. EvoGUD + AF3-SS generalizes to this independent benchmark without retuning.

c, Representative monomer example (PDB: 8OV5) illustrating global backbone accuracy. The experimental structure (white) is compared with predictions from AF3-SS (gray), ESMFold (green), EvoGUD + AF3-SS (pink), and AF3 (blue). EvoGUD + AF3-SS recovers the correct overall topology and domain arrangement.

d, All-atom steric clash counts per 1,000 atoms on the RecentPDB-monomer benchmark. Clash counts are computed as heavy-atom contacts closer than the sum of van der Waals radii with a 0.6 Å tolerance. EvoGUD + AF3-SS significantly reduces steric clashes relative to ESMFold.

e, All-atom steric clash counts on the CASP15 monomer benchmark. EvoGUD + AF3-SS maintains low clash rates comparable to AF3, indicating improved side-chain packing without nMSAs search.

f, Local side-chain environment example (PDB: 8XOX). ESMFold (green sticks, top) shows strained rotamers and steric clashes despite a high-quality backbone, whereas EvoGUD + AF3-SS (pink sticks, bottom) produces well-packed, stereochemically reasonable side chains.

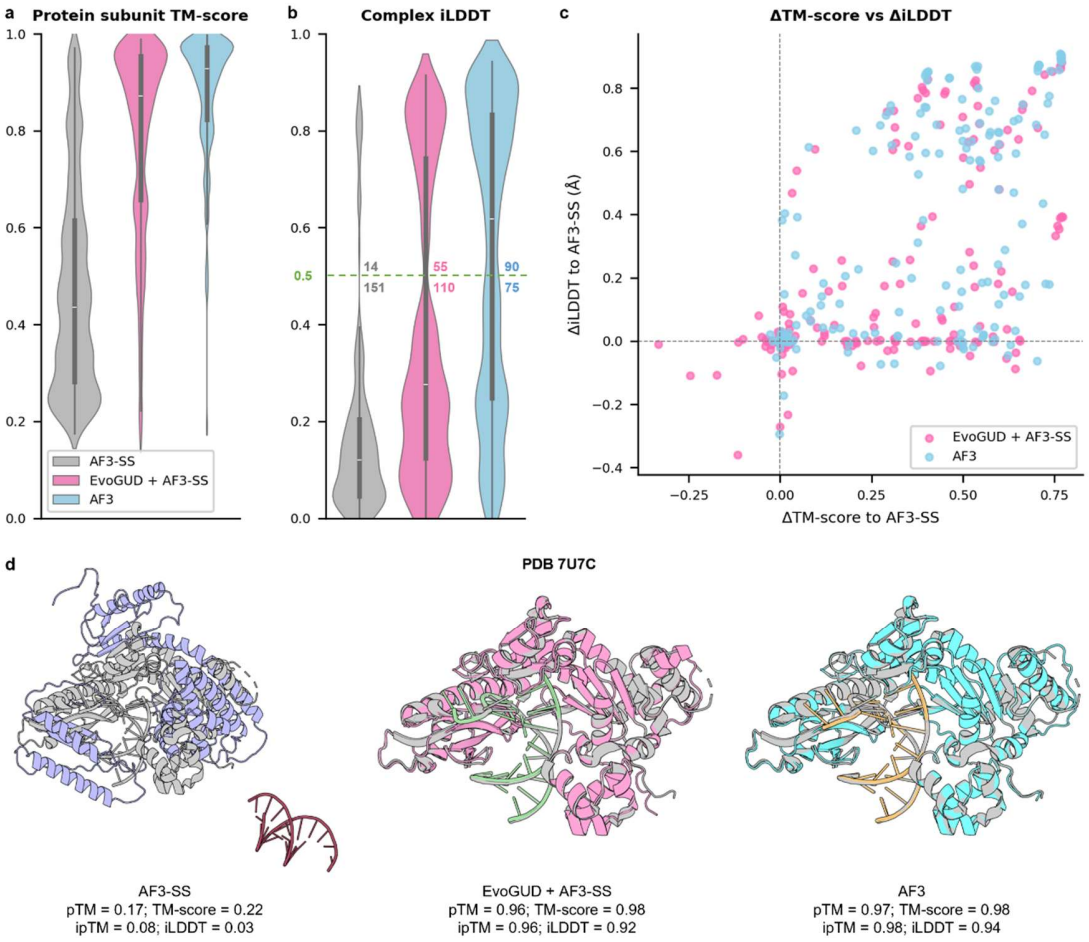


Figure 3. EvoGUD enables single-sequence AlphaFold3 to model protein–nucleic-acid complexes.

a, TM-score distributions of protein subunits extracted from predicted complexes for a benchmark of 165 protein–DNA/RNA assemblies, comparing AF3-SS (single sequence), EvoGUD + AF3-SS (vMSAs generated from single sequences), and AF3 with nMSAs and PDB template search ("AF3"). TM-scores are computed on protein subunits only, quantifying the correctness of individual protein folds within the predicted complexes. b, Interface LDDT (iLDDT) distributions for the same complexes, computed over protein–protein and protein–nucleic-acid interfaces following the AlphaFold3 evaluation

protocol. The horizontal dashed line at iLDDT = 0.5 is shown as a visual reference to illustrate the separation between lower- and higher-quality interface predictions observed in this benchmark. Numbers above and below the line report the counts of complexes on either side of this reference.

c, Per-complex changes in protein-subunit TM-score and interface iLDDT relative to AF3-SS. Each point corresponds to one complex, with ΔTM-score on the x-axis and ΔiLDDT on the y-axis (positive values indicate improved interfaces).

d, Representative protein–nucleic-acid complex example (PDB:7U7C).

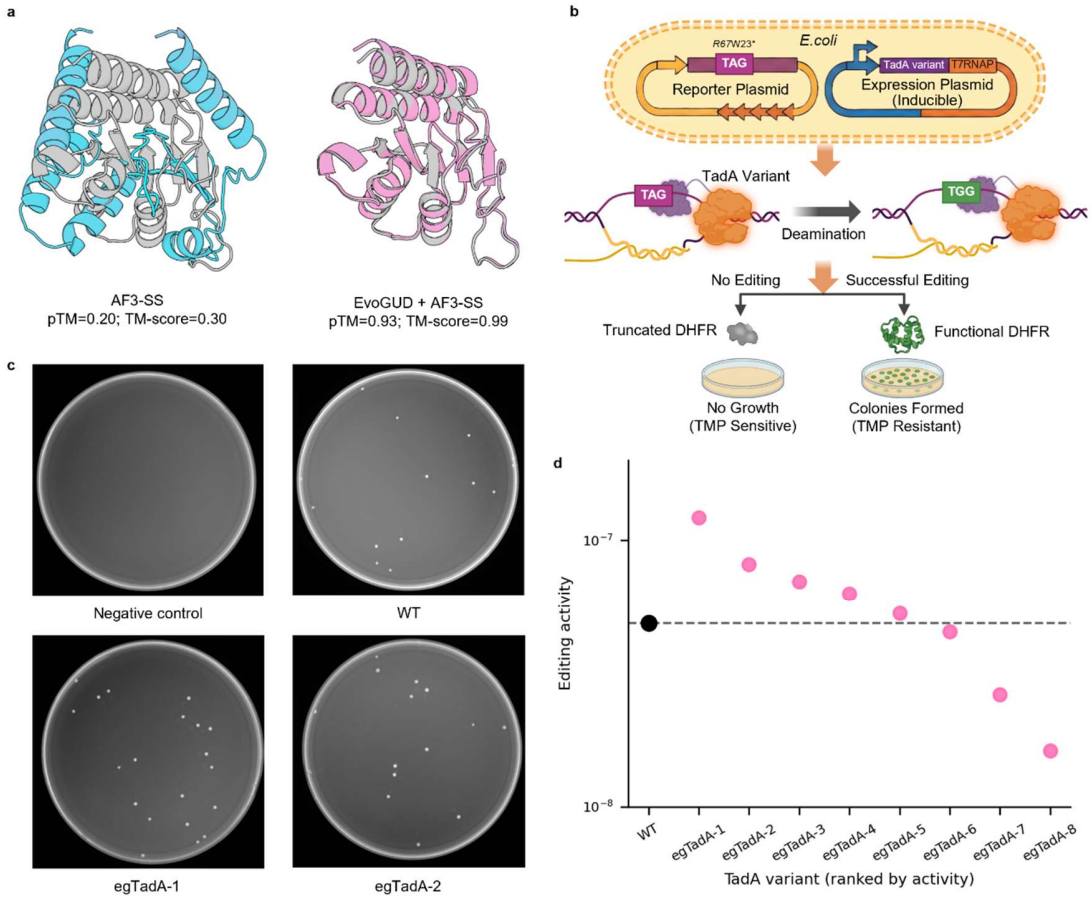


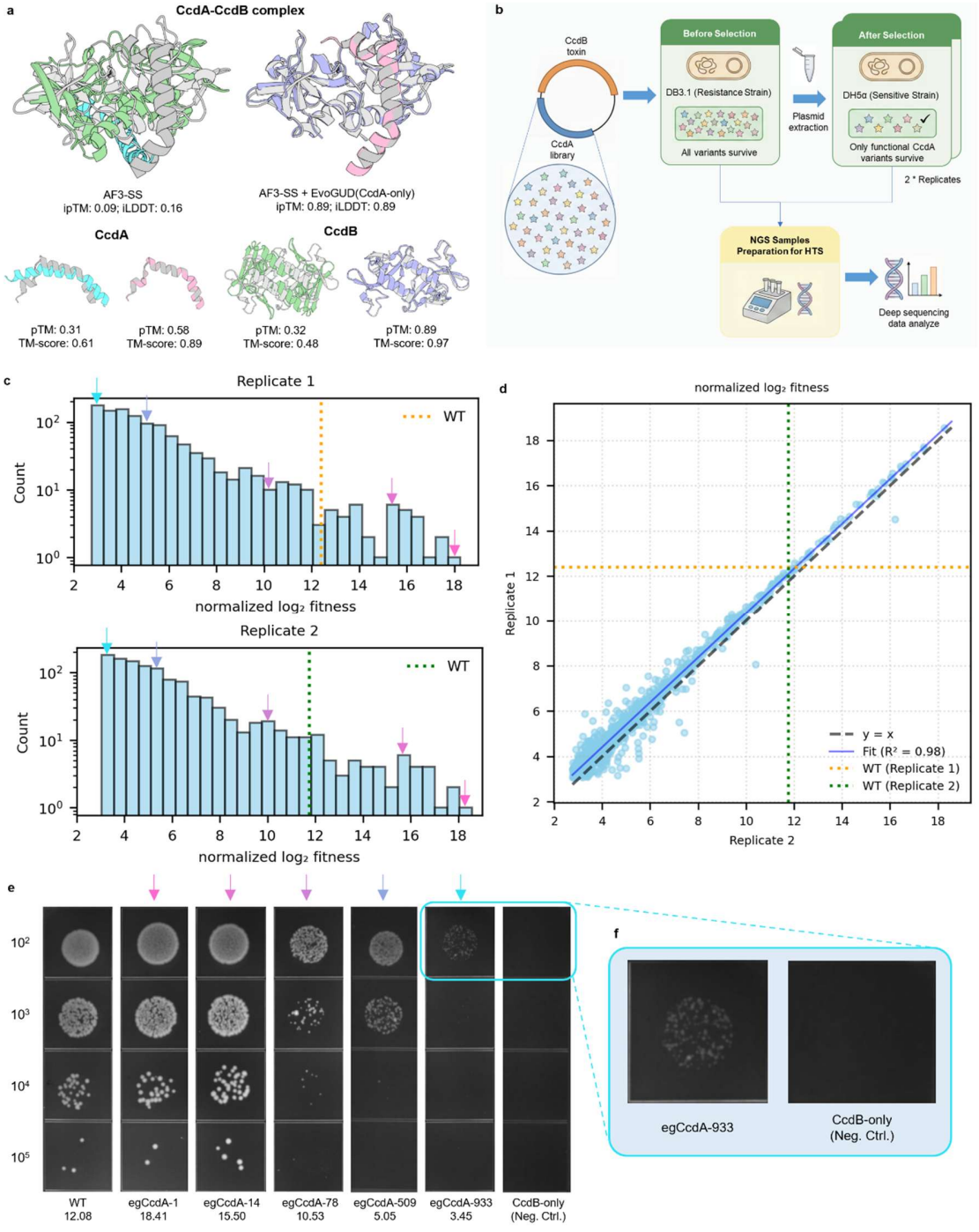
Figure 4. EvoGUD generates remotely homologous yet functional TadA variants.

a, AlphaFold3 (AF3) structure predictions for wild type *Staphylococcus aureus* TadA with EvoGUD-generated variants as vMSA, compared with AF3 predictions obtained from single-sequence input without MSA (AF3-SS). Aligned on reference PDB (2B3J) structure (gray).

b, Schematic of the trimethoprim (TMP) resistance reversion assay used to evaluate TadA DNA-editing activity in E. coli. TadA-mediated A•T→G•C editing reverts a premature TAG stop codon in an *R67* dihydrofolate reductase (DHFR) reporter gene, restoring the functional TGG codon and conferring TMP resistance.

c, Representative agar plate images showing TMP-resistant colony growth. A reporter-only strain lacking TadA expression serves as the negative control, wild type (WT) TadA is shown as a positive control, and two EvoGUD-generated variants (egTadA-1 and egTadA-2) illustrate functional recovery.

d, Quantification of DNA-editing activity for EvoGUD-generated TadA variants measured by TMP-resistance reversion. Each point represents one variant; WT is shown in black and EvoGUD-generated variants in pink.



689

690 **Figure 5.** Sequence-only conditional generation of the intrinsically disordered antitoxin

691 CcdA and high-throughput functional screening.

a, AlphaFold3 (AF3) predictions of the CcdA–CcdB complex for EvoGUD-generated CcdA variants. Left: AF3 single-sequence inference (AF3-SS) using the wild-type CcdA sequence. Right: AF3-SS using EvoGUD-generated CcdA sequences provided as vMSA ($\gamma = 2$, #MSA = 8). CcdA is shown in cyan (AF3-SS) or pink (AF3-SS + EvoGUD), and CcdB is shown in green (AF3-SS) or blue (AF3-SS + EvoGUD). Below, predicted structures of the CcdA and CcdB subunits are shown separately using the same color scheme. Reported ipTM, iLDDT, pTM, and TM-score values are indicated beneath each model.

b, Pooled selection workflow. A pooled oligo library of CcdA variants was constructed by cloning into a ccdB expression vector, pUC57-Kan-2BspQI-ccdB. The library was first propagated in the CcdB-resistant strain DB3.1 (Before selection), then subjected to selection in the CcdB-sensitive strain DH5α (After selection; two biological replicates). Plasmids were extracted and deep-sequenced to infer variant fitness.

c, Distributions of normalized $\log_2$ fitness for two biological replicates. Fitness is computed from before/after sequencing using early-stop variants as negative controls to estimate a baseline distribution (median and MAD), followed by BH-FDR filtering ($q \leq 0.01$) and normalization to obtain normalized $\log_2$ fitness. Dotted lines mark WT; arrows indicate variants chosen for plate assays.

d, Cross-validation of normalized $\log_2$ fitness between replicates (N = 1072), with $y = x$ reference, fitted trend, and WT reference lines.

e, Plate-based validation of selected variants. Serial dilution spot assays for WT, selected variants, and a CcdB-only negative control; numbers under each label denote normalized $\log_2$ fitness used for selection.

715 f, Zoomed comparison highlighting egCcdA-933 versus the CcdB-only negative control.

716

Data availability

The source code, sampling script, and model weight are soon publicly available at https://github.com/EricZhangSCUT/EvoGUD.

Author contributions

XZ developed the computational methods, performed computational evaluation, and managed the overall project. JT and TZ designed experiments and performed experimental validations. ZC and ZZ performed NGS data processing together with XZ. JZ and YZ initiated and supervised the project, and YZ provided funding support. YZ and XZ drafted the initial manuscript. All authors contributed to manuscript revision and approved the final version.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 92370202) and the National Key R&D Program of China (Grant No. 2021YFF1200400). We acknowledge the High-Performance Computing Cluster at Shenzhen Bay Laboratory (SZBL) and the high-performance computing resources of the Shenzhen Medical Academy of Research and Translation (SMART) for providing computational support.

Conflict of Interest

All authors declare no financial interest. Jian Zhan is the founder and CEO of Ribopeutic, and Yaoqi Zhou is the scientific founder of Ribopeutic.