SPIN-dvEvo: Exploration of vast functional sequence space by directed virtual evolution from a local sequence cluster

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

[1]Tsinghua University, Beijing 100084, China
[2]Shenzhen Medical Academy of Research and Translation (SMART), Shenzhen, 518107, China
[3]Institute of Systems and Physical Biology, Shenzhen Bay Laboratory, Shenzhen, 518107, China
[4]School of Medicine, Southern University of Science and Technology, Shenzhen, 518055, China
[5]Ribopeutic (Shenzhen) Co., Ltd., Futian, Shenzhen, 518000, China
[6]Ribopeutic Inc., Qiantang, 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

Both natural and directed evolution are powerful in improving protein functions but they are slow in exploring the nearly endless sequence space. Here, we present SPIN-dvEvo that couples few-shot low-rank adaptation (LoRA) of an ESM-2 protein language model with a genetic algorithm to quickly evolve functional remote homologs from a local cluster of highly-homologous, binary-labeled sequences. We experimentally tested SPIN-dvEvo on an enzyme (the core deaminase component of adenine base editors, TadA) and an intrinsically disordered protein (antitoxin CcdA). In TadA, virtually evolved sequences with low sequence identity to the starting sequences achieved a 38% success rate (23/60) in the first round and a 51% success rate along with a one-order-of-magnitude improvement in enzymatic activity in the second round, for which SPIN-dvEvo was retrained on first-round labels. Virtual evolution of the disordered protein CcdA was also successful, albeit at low success rate of 2.6%. Thus, SPIN-dvEvo can simulate billions of years of evolution in just minutes, rapidly creating new functional clusters.

Introduction

Directed evolution is a central strategy for engineering functional proteins, enabling stepwise improvement of enzymes and binders directly in the laboratory through iterative cycles of mutagenesis, selection, and amplification. It has produced catalysts with enhanced activity, altered specificity, and improved stability for applications ranging from therapeutics to industrial biocatalysis[1, 2]. Despite these successes, both natural and laboratory evolution remain intrinsically constrained by the locality of accessible mutational steps: most variants that can be efficiently sampled and screened in practice differ from their progenitors by only a small number of substitutions[3], and optimization therefore proceeds as a "local walk" on a rugged fitness landscape[3]. Because protein function is shaped by epistasis and higher-order constraints, such local searches can become trapped on suboptimal peaks, leaving distant but potentially functional better-fitness regions of sequence space systematically underexplored[3, 4]. While deep mutational scanning and deep sequencing expanded access to large-scale sequence–function measurements [5-7], they were also largely limited to local sequence space.

More recently, directed virtual evolution has emerged as a sequence-first alternative that learns surrogate fitness landscapes from sequence–function data and improves sequences by in silico search. Early work established probabilistic surrogate modeling, with Gaussian processes trained on measured activities to guide navigation of fitness landscapes.[8, 9] In parallel, structure-first pipelines such as AiCE integrate inverse-folding models with structural and evolutionary constraints to propose fold-compatible variants when reliable templates are available.[10] These approaches for directed virtual evolution can be grouped by how they obtain supervision and how far they can reliably search. One class relies on regression-style predictors trained on quantitative measurements (activity, fitness, binding, or other continuous phenotypes). These methods include deep supervised models or protein language models (PLM) with evolutionary context and active-learning pipelines for multi-round optimization such as ECNet[11] and EVOLVEpro[12], active-learning evolution of artificial metalloenzymes by Vornholt and colleagues[13], iterative deep-learning–guided directed evolution described by Li and colleagues[14], and Active Learning-assisted Directed Evolution (ALDE) by Yang et al.[15] A second class focuses on improving label efficiency, exemplified by Low-N protein engineering, which trains data-efficient predictors from small labeled sets and then screens large virtual libraries. [16] A third class formalizes "design–test–learn" iteration and can be coupled to automated platforms or biofoundries for higher-throughput cycles, exemplified by the STAR web server[17] and biofoundry-integrated PLM workflows for automated protein evolution.[18] Despite substantial gains in these systems, the learned surrogates continue to be most reliable near the starting scaffold. That is, the search remains implemented as local exploration (for example, small-step mutation moves, Bayesian optimization, or iterative screening) rather than sampling remote, low-identity functional sequences directly.

Another limitation of the above methods is their reliance on strong, quantitative supervision. Many successful workflows train regression-style surrogate landscapes on continuous activity measurements, kinetic readouts, or well-calibrated phenotypes, often requiring hundreds to thousands of assayed variants per round to achieve predictive accuracy that is sufficient to guide search.[12-14, 16, 19] When the available signal is weaker—binary functional labels or enrichment counts from pooled selections—performance can degrade because the training objective becomes less informative per variant and experimental noise from low counts and sampling variance becomes a dominant factor that must be modeled carefully.[6, 7, 20] In practice, this data requirement can be the primary bottleneck early in a campaign, when only a small number of positives exist and quantitative characterization is costly or not yet available.

Here we introduce SPIN-dvEvo (Sequence Prediction with Integrated Neural networks – directed virtual Evolution), a directed virtual evolution framework that unifies discovery and multi-round refinement in a single, sequence-first, low-data workflow. In contrast to conventional generative models that necessitate vast functional datasets, SPIN-dvEvo maintains high precision in label-scarce regimes by harnessing the evolutionary priors embedded within ESM-2[18] and fine-tuning its trajectory through lightweight LoRA.[21] The resulting model was employed as a scorer for directed virtual evolution powered by a genetic algorithm. SPIN-dvEvo was applied to virtually evolve TadA adenosine deaminase activity and the intrinsically disordered antitoxin CcdA. In both systems, only sparse binary functional labels were used to train a LoRA-based activity scorer on a frozen ESM-2 encoder, and candidate sequences were then generated by iterative mutation–crossover search guided by the fixed scorer, with additional feedback round performed for TadA by updating the LoRA head with newly obtained experimental labels. Experimental results confirmed the ability of SPIN-dvEvo to quickly evolve from a local cluster of a few highly homologous sequences to remote functional clusters.

Results

Directed virtual evolution by SPIN-dvEvo

The SPIN-dvEvo framework consists of two tightly coupled components: (i) a LoRA-adapted ESM-2 scoring model that trained task-specific functional scores using a small set of positive and negative sequences, and (ii) a genetic algorithm (GA) that directs sequence sampling and virtual evolution under this learned landscape. Each run starts from an initial seed pool generated by applying 20% random substitutions to the starting sequences. Here, SPIN-dvEvo was fine-tuned using only qualitatively labeled sequences with binary activity labels (1 for active, 0 for inactive) **(Fig. 1)**.

Directed virtual evolution from the neighborhood of an enzyme: TadA

To evaluate the enzyme-evolution capability of SPIN-dvEvo, we selected the tRNA-specific adenosine deaminase TadA as our model enzyme. TadA, originally evolved to target tRNA, has been engineered into adenine base editors that catalyze A•T→G•C conversions in DNA[22]. This system utilizes an R67 DHFR-based codon reversion reporter to rapidly detect the intracellular DNA-editing activity of evolved TadA variants, as in prior studies[23, 24]. In this codon reversion assay, an active variant reverts a premature TAG stop codon to TGG in the reporter, enabling growth under trimethoprim (TMP) selection **(Fig. 2A)**. We quantified intracellular DNA-editing activity as the mutation frequency $f = N_1 / N_0$—the number of TMP-resistant revertants (N_1) divided by the total number of viable cells plated without TMP (N_0)—and converted it into $\mu_{\text{s.p.b.}}$ (per base per generation; See **Methods**).

We compiled a compact set of 10 TadA sequences spanning the wild type from *E. coli* (UniProt ID P68398) (**Supplementary Table S1**) and previously engineered active variants from *E. coli* with 6-20 mutations, (>88.6% sequence identity) and labeled all these sequences as 1. That is, we started with a sequence cluster of close functional neighbors. An equal number of 10 hypothetical inactive sequences were obtained by performing random mutations at 20% of positions in these TadA sequences (See **Methods**).

We then employed SPIN-dvEvo to produce (evolved virtually) 1,000 sequences by starting from the inactive sequences pool (20% random mutations). We confirmed that such virtual evolution started from a tightly clustered sequence region (in red) and quickly expands to other regions according to the t-SNE projection of ESM-2 sequence embeddings (**Fig 2B**).

To examine whether the TadA function was preserved during the virtual evolution, we obtained sequence logos from 1000 natural TadA homologs compiled as in Ref [25] with a median sequence identity of 34.1% and compared them to sequence logos from 1000 evolved sequences (median identity 55.8%) in **Fig 2C**. The sequence motifs found previously[26] in the TadA family such as HAE and PCXXC zinc-dependent deaminase motifs and structural-core signatures EVP and TLE were also conserved in the evolved sequences while allowing substantial variation elsewhere. Thus, essential sequence information preserved in the natural sequences evolved over billions of years was captured by SPIN-dvEvo in a short 12 minutes of computing time with AMD EPYC 9654/ RTX 4090 (24 GB) starting from a local sequence cluster around *E. coli* TadA.

As protein structures play an essential role in enzymatic functions, we predicted structures of these evolved sequences and compared them to the structure of wild-type TadA (PDB ID:2B3J). We employed PLM-based OmegaFold[27] to make predictions because it does not require homologous sequences for input, and therefore permits fast, large-scale calculations for all 1000 evolved sequences. We obtained the distribution of structural accuracy (measured by TM-score[28], 1 for perfect match and 0 for no match) for predicted structures of those evolved SPIN-dvEvo sequences and compared it to two baseline models PLM-based sequence generators Pinal[29, 30] and structure-based protein-design method ProteinMPNN[31]. ProteinMPNN employed a native structure template; Pinal was prompted with a natural-language TadA functional description (adenosine deaminase/base-editor context; EC 3.5.4.33) together with the wild-type TadA sequence **(Methods).** The results show that most evolved sequences given by SPIN-dvEvo adopted near-native structures (TMscore ~0.8, 89.6% sequences with TM-score>0.5), and was only slightly worse than the structure-based method ProteinMPNN (TMscore ~0.95) (**Fig. 2D**). The baseline sequence-based method Pinal shows a bimodal TM-score distribution, with one major peak at low TM-scores (~0.2–0.3, 53.9% sequences with TM-score<0.5) and another in the near-native range (~0.8–0.9), indicating a mixture of largely off-fold sequences and a smaller subset that retains the TadA fold. An example of a predicted structure for a SPIN-dvEvo sequence is compared to the native structure in **Supplementary Fig. 1**, highlighting near-perfect match, particularly in the regions interacting with a DNA substrate and near catalytic core.

We further selected 60 evolved sequences to validate their enzymatic functions experimentally with the R67 DHFR–based codon reversion assay (**Fig. 2A**). These 60 sequences were selected from the above 1000 evolved sequences according to the high structure-confidence scores (normalized pLDDT> 0.9 given by AlphaFold 3[32] with a single natural MSA to save computing time) and low sequence identity (≤0.5) to the wild type (as shown in **Fig. 2E**). Functional validation identified 23 active variants out of 60 tested (38.3% success rate). Activities spanned more than three orders of magnitude, with several variants matching or exceeding the reference activity of *E. coli* TadA **(Fig. 2F, Supplementary Fig. 2 A, Table S3, Table S4**). More importantly, these individually validated functional sequences span 39–79% amino-acid identity to the *E. coli* TadA wild type, confirming the ability of SPIN-dvEvo to find functional solutions by going significantly beyond the immediate neighborhood of the starting sequences within the identity neighborhood of ≥88% *E. coli* TadA (**Fig. 2E**).

Given 60 newly experimentally tested sequences, we re-trained the LoRA model with the enlarged binary-labelled dataset and performed sequence evolutions again by GA. The newly 1000 evolved sequences (Round II) are now forming new sequence clusters (Fig. 2B). The TMscore distribution of predicted structures for the second-round sequences improves over that of the first round. All predicted structures (100%) are now with TMscore > 0.78 and the highest peak located at TMscore of 0.88, compared to 0.80 in the first round (**Fig. 2D**). We tested 60 new variants chosen according to high AlphaFold 3's pLDDT and low sequence similarity. In this second round, 31 of 60 new variants were active. The higher success rate in Round II than in Round I (51% versus 38.3%) indicates that incorporating new experimental labels with definitive inactive sequences improved the classifier-guided evolution **(Supplementary Fig. 2B, Table S3, Table S5)**. Moreover, the measured activity for the functional sequences in the second round shifted upward relative to the first-round actives by one order of magnitude (**Fig. 2F**). These validated evolved sequences in Round II are more divergent from wild type (29–54% identity, compared to 39–79% in the first round; **Fig. 2E**), confirming the formation of new functional clusters with improved activity (**Fig. 2B**). This is remarkable considering the fact that only binary labels were employed to train SPIN-dvEvo.

A few selected variants are illustrated along with positive and negative controls by plating on TMP-selective medium (dvTadA-55 and dvTadA-56 from round 1; dvTadA-2-02 from round 2). These evolved sequences produced markedly more TMP-resistant colonies than the negative control of expressing only an Xten linker-T7RNAP cassette in place of TadA and thus lacking deaminase activity and were comparable to the positive control (*E. coli* TadA) (**Fig. 2G**), consistent with robust in vivo editing activity.

Directed virtual evolution of intrinsically-disordered binder: anti-toxin CcdA

To test whether SPIN-dvEvo can generalize beyond enzymes with well-defined structures to intrinsically disordered binding proteins, we applied it to the CcdA–CcdB toxin–antitoxin system. In *E. coli*, the antitoxin CcdA is a 72-residue protein. Here we only engineered its C-terminal segment (CcdA$^{36–72}$, 36 residues), which mediates binding to CcdB and thereby blocks CcdB binding to GyrA to neutralize toxicity[33]. This 36-residue C-terminal domain is intrinsically unstructured prior to binding to CcdB[33, 34]. We started from the canonical *E. coli* CcdA (P62552), retrieved CcdA family homologs from closely related *Enterobacterales/ Gammaproteobacteri*a, removed incomplete or atypical entries as well as those sequences at 100% sequence identity cutoff. This yielded 22 close homologs (**Supplementary Table S2**) at 55.2–97.2% sequence identity. A LoRA head on a frozen ESM-2 encoder was fine-tuned on this curated set and then coupled to the GA to generate candidate binders, without introducing any CcdB sequence or structural information during training or sampling. We chose this CcdA-CcdB system because bacterial growth is directly correlated to the ability of the CcdA evolved by SPIN-dvEvo to bind and neutralize CcdB, enabling straightforward functional selection **(Fig. 3A**).

As in the TadA case, we evolved 1000 CcdA variants by SPIN-dvEvo. As shown in **Fig 3B**, these sequences moved far away from the original sequence cluster and formed multiple clusters according to the t-SNE projections of the base ESM-2 embeddings. When we generated the sequence-logo from SPIN-dvEvo sequences (with a median sequence identity of 50.2%), it has similar sequence motifs as those from 100 natural homologs collected by querying the canonical 'Antitoxin CcdA' and filtering to a non-redundant set with a median sequence identity of 38.7% from UniProtKB, suggesting that key binding determinants preserved such as W44, E54,[35] G63, S64, F65, D71 and W72[36, 37](**Fig. 3C,** blue box) in natural CcdA homologs were captured during virtual evolution by SPIN-dvEvo, despite that it was started from a highly local seed set.

To test those sequences experimentally, we synthesized a library of 3,041 evolved CcdA variants and evaluated them using a pooled bacterial growth selection, because the ability for the bacterium to grow is correlated to the ability of the evolved CcdA to neutralize CcdB by binding **(Fig. 3A**). That is, the fitness of activity of CcdA variants can be measured by counting the number of a specific variant pre- and post-selections from high-throughput sequencing[38] (**Fig. 3A**). We estimated enrichment and uncertainty with the DiMSum pipeline [39,40] with Poisson–Delta variance modeling and overdispersion correction. Among 3,041 synthesized CcdA variants, only 2,363 variants were found with >30 reads and a minimum frequency of 10^{-6}in both the pre-selection and post-selection libraries from high-throughput-sequencing data. Further application of an FDR-controlled filter relative to internal stop-codon negative controls of q_value $< 10^{-3}$ yielded 155 statistically significant functional variants (a 6.6% hit rate). We further employed an effect-size threshold to define more robust positives as those variants with $\log_2(\text{fitness}) > 3.0$, resulting in 62 active CcdA variants (a 2.6% hit rate, **Fig. 3D**). These variants contain 26 with $\log_2(\text{fitness}) > 5$ and some comparable to the fitness of *E. coli* CcdA ($\log_2(\text{fitness}) = 8.5$).

To validate the above high-throughput result, we selected four variants around the stringent threshold of 3.0 with $\log_2(\text{fitness}) = 3.3, 3.3, 3.2,$ and $3.0,$ respectively, along with two positive controls *E. coli* CcdA and an evolved variant with $\log_2(\text{fitness})$ =5.3 for in vivo functional testing (**Supplementary Table S7**). As shown in **Fig. 3E** by serial 10-fold dilution spot assays, we confirmed that all variants with $\log_2(\text{fitness})\geq 3.0$ are functional and the variant 878 with a larger fitness value has stronger growth. In particular, the variant 1654 with $\log_2(\text{fitness}) = 3.0$ showed weak growth only at the dilution factor of 10^2. It is noted that sequences with $\log_2(\text{fitness}) \geq 3.0$ retained only ~60–70% sequence identity to the *E. coli* CcdA (**Supplementary Fig. 3**), indicating substantial novelty among functional hits, given that only 36 residues were targeted for virtual evolution.

Discussion

SPIN-dvEvo directly addresses a practical gap in current directed virtual evolution: most existing methods either require substantial labelled datasets to optimize a single scaffold locally, or function as one-shot generators whose sequences are not coupled to an explicit score-and-search loop that can be iterated with newly acquired labels. In contrast, SPIN-dvEvo mimics natural evolution by employing a LoRA adaptor on the top of a frozen ESM-2 encoder to learn functional restraints. We showed that the functional restraints learned from a few dozen positive, binary-labeled samples of a highly homologous sequence cluster are sufficient to drive virtual evolution from dysfunctional sequences to functionally active proteins that are substantially away from original positive sequences by using a genetic algorithm. Some of these sequences, despite low sequence identity, are experimentally validated for their functions on two illustrative cases: enzymatic activity (TadA adenosine deaminase) and toxin-binding intrinsically disordered protein CcdA.

For virtual evolution of TadA enzyme, no structural information of was used to train SPIN-dvEvo and to drive evolution. Yet most evolved TadA variants have TadA structural folds (**Fig 2D, Supplementary Fig. 1**) in the first round (89.7% of sequences with predicted structural accuracy >0.5 in TMscore). A minor peak with TMscore<0.5 in the first round was eliminated after including experimental results from 60 variants (still in binary coding). The improved structural similarity to the wild type highlights the importance of a larger and cleaner dataset because in the first round, negatives represented by 20% random mutations may not be negatives. Interestingly, the second-round success rate increased from 38% to 51% along with a one-order-of-magnitude improvement in enzymatic activity, indicating that adding new experimental labels can improve classifier-guided search even for enzymatic activity, despite lacking quantitative labels.

We have selected sequences with high confidence in predicted structures for experimental validations. The high (38% in Round I) but not yet >90% success rate for TadA's virtual evolution illustrates that the structural fold alone is not sufficient as an indicator of enzymatic activity. This is because enzyme function not only requires highly precise active-site geometry and transition-state stabilization, but also depends on compatible conformational dynamics and kinetics that enable efficient substrate binding and product release on a productive timescale.[41-43] More studies are needed to search for a better activity indicator as well as improving scoring for virtual evolution of enzymes.

SPIN-dvEvo evolved functional TadA starting from a 20% randomly mutated (inactive) seed. We kept starting sequences close to the TadA family where the LoRA scorer remains informative. We also tried to start from fully random sequences and found that evolution from these sequences is not productive according to analysis of their predicted structures. This indicates that the sequence space is too large to be located by starting from purely random sequences within practical GA generations. Nevertheless, it can start from one neighborhood of an active sequence to locate other neighborhoods far away from the original sequence cluster as shown in **Fig. 2B** and **Fig 3B**.

However, the success rate of SPIN-dvEvo for a disordered protein CcdA is only 2.6%. This is much lower than virtual evolution of TadA enzyme. Designing an intrinsically disordered protein is a challenging task because activity is typically encoded in an ensemble of rapidly interconverting conformations and mediated by weak, context-dependent interactions, so improvements in fold stability or a single “best” structure provide little guidance. Recent progress has come from explicitly optimizing ensemble-level objectives, for example by using sequence-to-ensemble predictors for IDRs and by combining generative models with biophysical/simulation-based forward models to design sequences that realize targeted disordered-state properties, as well as from diffusion-based binder design strategies that focus the objective on functional binding constraints rather than enforcing an ordered fold.[44] Here, we achieved a success (albeit low success rate) without relying on any information from binding partner CcdB or predicted complex structures.

It is of interest to know how new functional clusters would have been evolved naturally if they were mixed with natural homologs when building phylogenetic trees(see SI). As shown in **Fig. 4A** and **Fig. 4B**, both virtually evolved TadA and CcdA are forming several phylogenetically distinct clusters but do share common ancestors with naturally occurring sisters at different time points. For TadA, this split corresponds to an evolutionary timescale on the order of ~0.2–1.2 Ga, based on TimeTree-derived lineage-age estimates for these taxa[45-47]. Similarly, the estimates for the virtually evolved CcdA clade dating to approximately 2.508 Ga as diverging from a Gammaproteobacteria-associated branc. By comparison, these virtual evolutions took only 713 seconds for TadA and 761 seconds for CcdA by SPIN-dvEvo on a workstation equipped with an AMD EPYC 9654 (96-core, 2.4 GHz) CPU and an NVIDIA RTX 4090 GPU (24 GB).

SPIN-dvEvo was purposefully trained on binary-labeled sequences (1 for functional and 0 for nonfunctional). This is because most proteins with known functions do not have a quantitative functional label. One immediate improvement for SPIN-dvEvo is to employ a regression head, rather than a classification head, when quantitative functional data such as a fitness score, binding affinity, or enzymatic activity is available for a small dataset. A regression head would contain a more accurate evolution direction than a classification head. This is a subject of an ongoing study.

One limitation of SPIN-dvEvo is its reliance on the ESM-2 650M. While ESM-2 is one of the best protein language models available, we did not have the resource to test other language models or utilization of multiple language models that could be potentially more beneficial than ESM-2 in directed virtual evolution. Moreover ESM-2 may be inherently biased toward some protein sequences with large family of homologous sequences as it was indiscriminately trained on all protein sequences.[48, 49] Further studies in this area are needed.

Moreover, current implementation of SPIN-dvEvo is optimized for a single functional objective. A multi-objective model, where functional objectives are optimized alongside other property objectives such as stability, pH tolerance, and thermostability, can be easily implemented. This research is also currently ongoing.

Methods

Data Collection and Curation

For TadA, we compiled 10 functional sequences from previously engineered DNA-editing TadA variants[22] (listed in **Supplementary Table S1**). For CcdA, we constructed the 22-sequence set by sequence-identity clustering of UniProtKB CcdA homologs. Starting from the canonical *E. coli* CcdA (P62552; 36 aa) as the query, we retrieved annotated CcdA family homologs from closely related *Enterobacterales/Gammaproteobacteria*. We then removed incomplete/aberrant entries (e.g., truncated sequences or atypical lengths) and identical sequences (100% sequence identity). This yielded a deduplicated set by keeping only unique amino-acid sequences, yielding 22 non-redundant homologs (accessions in **Supplementary Table S2**). To balance classes during few-shot training, we generated synthetic decoys by randomly mutating 20% of residues in each positive sequence. All positive sequences were labeled as 1 (functional), and all negative sequences—whether randomly generated or literature-confirmed—were labeled as 0 (non-functional).

LoRA-Based Model Adaptation

We adapted ESM-2 (650M parameters) to each task using low-rank adapters (LoRA) while keeping all base model weights frozen. This model size offered a practical trade-off between representation quality and computational cost, allowing training on a single 24–40 GB GPU.

LoRA modules were inserted into the self-attention Q/K/V projection layers of every transformer block. For each pretrained projection $W \in \mathbb{R}^{d\times d}$, LoRA adds a trainable low-rank update $\Delta W = s\,AB$ with rank r and scaling $s = \alpha/r$:

$$\widetilde{W} = W + s\,AB, A \in \mathbb{R}^{d\times r},\ B \in \mathbb{R}^{r\times d},\ s = \alpha/r.$$

We employed $(r, \alpha) = (16,16)$. This setting adds 4,055,040 **LoRA** trainable parameters (excluding the final linear head), corresponding to ~0.62% of the ~650M-parameter ESM-2 base model, and was used throughout this work.

Classification head (binary activity)

For binary activity prediction $y_i \in \{0,1\}$, the frozen ESM-2 produces a sequence representation $h \in \mathbb{R}^d$(pooled from token embeddings), which is mapped to a scalar logit

$$z = u{\top}h + b, score = f(x) = \sigma(z) \in [0,1]$$

The classifier was trained with binary cross-entropy:

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log\, p_i + (1 - y_i)\log\,(1 - p_i)]\,.$$

Only the LoRA parameters (A, B) and the classification head parameters (u, b) were updated during training; all ESM-2 weights remained frozen,

Sequences were truncated to 1,000 amino acids and fine-tuned for 5 epochs using AdamW (learning rate 5×10^{-4}, weight decay 10^{-3}) with a cosine schedule and gradient clipping ($\| \nabla \|_{\max}$= 0.5). LoRA adapters targeted the attention Q/K/V projections (rank $r = 16$, $\alpha = 16$, dropout 0.2; base model frozen) with batch size 4.

Genetic Algorithm Sampling

We performed an iterative mutation–crossover search guided by a fixed LoRA activity scorer. Diversity arose implicitly from uniform parent sampling and stochastic point mutations, and exact duplicate children were removed during population construction. In each generation, parent sequences were sampled uniformly from the current mating pool and recombined to produce a child.

Each sequence was scored by the LoRA-adapted ESM-2 classifier, with the positive-class probability computed from the logits as

$$p_{\mathrm{act}}(x) = \frac{\exp(z_1)}{\exp(z_0) + \exp(z_1)}.$$

Initialization.

The initial population consisted of N sequences (equal to the size of the seed pool), generated by applying 20% random substitutions to a set of positive sequences (natural homologs or previously engineered variants).

Embedding & activity model.

Each sequence was scored by a LoRA-tuned binary activity classifier on a frozen ESM-2 (650M), returning $p_{\mathrm{act}}(x)$. (Sequence embeddings $\phi(x)$ were computed when needed for visualization/analysis, by mean-pooling the last hidden state over non-special tokens followed by L2 normalization.)

Variation & constraints.

Children were generated using a one-point crossover plus point-mutation operator (mutate_crossover). One parent was first chosen as the base; a crossover point $c \in [1, \min(|p_1|, |p_2|) - 1]$ was sampled, and the suffix was swapped with the other parent, yielding a recombinant whose length follows the suffix donor. After crossover, each position was independently mutated with probability 0.02 by substituting a uniformly sampled amino acid from the 20 standard residues. Candidate sequences were filtered with NCBI segmasker[50] to reject sequences containing low-complexity segments longer than 5 residues.

Selection & replacement.

For each parent sequence x with score $p_{\mathrm{act}}(x)$, a child x' was proposed and evaluated to obtain $p_{\mathrm{act}}(x')$. The acceptance ratio was computed as

$$r = \frac{p_{\mathrm{act}}(x')}{p_{\mathrm{act}}(x)}.$$

The child was accepted if $r \geq 1$; otherwise, it was accepted with probability $0.125 \times r$. After iterating this accept/reject update across the population, sequences were ranked by score (by p_{act} in probability-only mode) and the top 25% sequences (ranked by score) were retained as the mating pool for the next generation. Unless stated otherwise, virtual evolutions were conducted for a pre-specified number of generations (default is 100) and the per-generation mean score was logged.

Parallel runs. Each run outputs N evolved sequences (set by the seed pool size). Larger libraries were obtained by launching multiple independent runs in parallel with different random seeds and by aggregating the resulting sequences.

Sequence sampling of baseline models: ProteinMPNN and Pinal

ProteinMPNN

A structure-templated baseline library was generated using ProteinMPNN in fixed-backbone design mode with the experimental TadA reference structure as the input template (PDB: **2B3J**, Chain **A**).

The structure file was preprocessed to retain only the designed protein chain (non-protein atoms were removed) and was provided to ProteinMPNN to compute per-position amino-acid distributions conditioned on the backbone coordinates. 1,000 sequences were then sampled stochastically from the model using temperature-controlled decoding (temperature = 0.1) with otherwise default ProteinMPNN settings. Sampled sequences were post-processed to remove exact duplicates and were written to FASTA for downstream structure prediction and evaluation.

Prompt for Pinal Sequence Generation

TadA Prompt：TadA (tRNA adenosine deaminase) is an enzyme that catalyzes the deamination of adenosine to inosine at the wobble position (A34) of tRNA molecules, thereby expanding codon recognition during translation, adenosine$_{34}$ in tRNA + H2O + H^+ = inosine$_{34}$ in tRNA + $NH4^+$. EC:3.5.4.33. Through the introduction of two key mutations, A106V and D108N, the substrate specificity of *E.coli* TadA has been reprogrammed, enabling the enzyme to catalyze adenosine and cytosine deamination directly on DNA substrates. These engineered TadA variants are incorporated into adenine base editors (ABEs), facilitating the precise conversion of A•T base pairs to G•C in DNA without introducing double-strand breaks. This strategy offers an efficient and high-fidelity tool for genome editing, particularly for the correction of disease-associated point mutations.

CcdA Prompt：CcdA is a bacterial antitoxin protein that functions as part of the CcdA–CcdB type II toxin-antitoxin system encoded by the F plasmid in *Escherichia coli*. The CcdA protein comprises 72 amino acids and adopts a two-domain structure: an N-terminal dimerization and DNA-binding domain, followed by a C-terminal domain that binds to the CcdB toxin. In the absence of CcdB, the C-terminal domain of CcdA is intrinsically disordered. Upon binding to CcdB, CcdA undergoes a conformational change, forming a stable CcdA–CcdB complex that neutralizes the toxicity of CcdB. This complex also acts as a transcriptional repressor of the ccd operon by binding to the operator region. The CcdA–CcdB interaction is dynamic, with varying stoichiometries leading to different complex formations, including (CcdA)2–(CcdB)2 and (CcdA)2–(CcdB)4 complexes. The balance between CcdA and CcdB concentrations regulates the stability of the complex and the repression of the operon. CcdA is subject to degradation by the Lon protease, which modulates the levels of the antitoxin and, consequently, the activity of the toxin.",

Structure prediction for SPIN-dvEvo sequences

SPIN-dvEvo sequences were evaluated by two complementary structure-prediction pipelines with distinct roles. For high-throughput, distribution-level benchmarking across large libraries, we used the MSA-free, PLM-based OmegaFold (v2.3.2)[27] to predict structures for all sequences, and quantified global fold similarity to experimental references using TM-align (TM-score). For TadA, PDB 2B3J (tRNA adenosine deaminase from *Staphylococcus aureus* in complex with RNA) was used as the reference structure, because it provides a substrate-bound, catalytically relevant conformation for a consistent TM-score fold-similarity benchmark; in contrast, the available *E. coli* TadA structure PDB 1Z3A is apo and does not capture the RNA-engaged state[51]. TM-scores reported in the main text refer to alignments between the native structure (PDB 2B3J) and OmegaFold-predicted structures for SPIN-dvEvo-evolved variants.

Separately, we used AlphaFold3 (AF3) to obtain model confidence estimates for experimental prioritization. To reduce the computational time for MSA retrieval, sequences were clustered at 80% pairwise identity; a representative sequence per cluster was used to query the AF3 MSA database, and the resulting MSAs were reused for all members of that cluster during the batch inference. For TadA, per-chain pLDDT was used as the confidence metric.

TadA experimental methods

Reagents and Strains

All PCR reactions for cloning restriction sites and generating recombineering targeting cassettes were performed using 2 × Phanta UniFi Master Mix DNA Polymerase (Vazyme, Nanjing, China, P516-02). Colony PCR reactions for subsequent sequencing were conducted using Premix Taq^{TM} DNA Polymerase (Takara, Dalian, China, R901A). Homologous recombination was performed using the CloneExpress II One Step Cloning Kit (Vazyme, C112-02). All primers were synthesized by GENEWIZ (Suzhou, China). Gene sequences for R67, which confers resistance to trimethoprim (TMP), and engineered TadA variants were synthesized by GENERAL BIOL (Anhui, China). Antibiotics, including ampicillin sodium (Sangon Biotech, Shanghai, China, A100339-0025) and chloramphenicol, along with L-arabinose, were obtained from commercial sources. Chemically competent *E. coli* DH5α cells were purchased from AlpalifeBio (Beijing, China), and chemically competent *E. coli* DH10B cells were obtained from Biomed (Beijing, China).

Plasmid construction

Engineered TadA variants used in this study are detailed in **Tables S3.** Expression plasmids for these variants and T7 RNA polymerase (T7RNAP) were constructed using the pMuta088 vector backbone. This backbone, derived from pDae079, carries the tandem PmCDA1-T7 RNA polymerase and uracil glycosylase inhibitor (UGI).

For this study, expression plasmids for the engineered TadA variants were constructed by replacing the PmCDA1 gene in the pMuta088 scaffold with the specific TadA sequences via homologous recombination. A negative control plasmid (pT7RNAP-ΔTadA), expressing only an Xten-linker–T7RNAP cassette, was constructed using the same strategy.[23]

TadA editing activity was quantified by measuring the frequency of trimethoprim-resistant revertants following the general MutaT7/eMutaT7 workflow with minor modifications as detailed below.[52] To characterize the A•T-to-G•C editing activity of TadA variants via antibiotic resistance reversion, a reporter plasmid was developed. The *R67* gene, encoding dihydrofolate reductase (DHFR) which confers resistance to trimethoprim (TMP), was cloned into a low-copy-number plasmid (T7 promoter + terminators reporter plasmid). This was achieved by replacing the existing *neoR*/ *kanR* gene (from Tn5) in a precursor plasmid via homologous recombination. In the final reporter construct (pReporter-R67), expression of the *R67* gene is driven by a T7 promoter and transcription is terminated by a tandem array of ten T7 terminators. Subsequently, site-directed mutagenesis was employed to convert the tryptophan codon (TGG) at position 23 into a premature stop codon (TAG), resulting in the final reporter construct pReporter-R67^{W23*}. In this system, TadA-mediated adenine deamination reverts the stop codon to wild-type, thereby restoring functional R67 expression and conferring TMP resistance.

Evaluation of TadA Variant Activity in *E. coli*

To quantitatively characterize intracellular DNA-editing activity, the mutation (editing) frequency was defined as the ratio of the total TMP-resistant revertants to the total viable cell population.

To perform this assay, chemically competent *E. coli* DH10B cells were co-transformed with two plasmids: (1) The reporter plasmid (AmpR) pReporter-R67^{W23*}; (2) a chloramphenicol-resistant (CmR) expression plasmid (pDae079 derivative) encoding either pT7RNAP-ΔTadA (negative control), wild-type TadA (positive control), or an engineered TadA variant.

Transformants were selected on LB agar plates containing 100 μg/mL ampicillin and 25 μg/mL chloramphenicol, followed by incubation at 37°C for 12–16 hours. Individual colonies were then inoculated directly into 10 mL of LB broth supplemented with 100 μg/mL ampicillin, 25 μg/mL chloramphenicol, and 0.2% (w/v) L-arabinose, followed by overnight incubation (16 h) at 37°C with shaking at 220 rpm to initiate TadA expression and mutation accumulation.

On the following day, the overnight cultures were diluted 1:100 into fresh LB medium containing the same concentrations of ampicillin, chloramphenicol, and L-arabinose. To promote the fixation of mutations during active growth, these cultures were incubated for 4 hours at 37℃ with shaking at 220 rpm.

Editing activity Assay

At the endpoint, cultures were serially diluted (10-fold). To determine the total viable cell population (N_0), 10 μL aliquots of each serial dilution were spotted onto a single non-selective LB agar plate (containing 100 μg/mL ampicillin and 25 μg/mL chloramphenicol). To enumerate the TMP-resistant population (N_1), 300 μL aliquots of undiluted culture were spread onto three selective LB agar plates containing 20 μg/mL TMP (supplemented with the same antibiotics). Plating for N_1 was performed in triplicate. Colony counts were extrapolated to the full 10 mL culture volume to derive the total viable cells (N_0, scaled from the 10 μL spot and dilution factors) and total TMP-resistant revertants (N_1, scaled from the 300 μL spread). The frequency f was calculated as the ratio N_1 / N_0.

Mutation-rate calculation.

For cross-study comparison to prior eMutaT7 reports, endpoint TMP-reversion frequencies were converted to per-base, per-generation mutation rates using the Luria–Delbrück rare-mutation approximation, where the expected mutant frequency satisfies $E[f] \approx \mu \ln(R_{\mathrm{eff}})$. Although induction was maintained for 16 h, the calculation was normalized to the effective population expansion of the final outgrowth step, as mutation fixation is replication-dependent. This single 4 h propagation propagation round used a 1:100 reinoculation followed by regrowth to saturation, corresponding to ~ 6.6 generations (G). Assuming binary fission ($R_{\mathrm{eff}} = 2^G$), $\ln(R_{\mathrm{eff}}) = G \ln 2 \approx 4.57$. Because TMP-resistance restoration of the R67 reporter requires a single-base reversion, the effective target size was set to $S = 1$ and rates were reported as site-specific values (not normalized by the 192-bp reporter length):

$$\mu_{s.p.b.} = \frac{f}{G\, ln\, 2} \approx \frac{f}{4.57}\ (per\ base\ per\ generation)$$

Verification of R67 Gene Reversion

To confirm that TMP resistance resulted from the targeted A•T-to-G•C edit in the *R67* gene, colony PCR was performed. For a representative subset of TadA variants tested, five independent TMP-resistant colonies were randomly picked from the selective agar plates for each selected variant. The *R67* gene locus was PCR-amplified from these colonies. The resulting amplicons were purified and subjected to Sanger sequencing (GENEWIZ, Suzhou, China). The obtained sequences were aligned with the reference *R67*$^{\mathrm{W23*}}$ sequence and the wild-type *R67* gene sequence to identify the specific A-to-G reversion at codon 23 and any other potential off-target mutations within the amplified region.

CcdA library generation, selection, and validation

The Plasmids Construction

The pUC57-Kan-ccdA/B plasmid was constructed to co-express the CcdA$^{36–72}$ domain and ccdB in *E. coli*. In this generation, the forward strand carries the J23119 promoter–driven CcdA$^{36–72}$ cassette, and the reverse strand carries the AmpR promoter–driven ccdB gene. A 21-bp spacer was inserted between the two stop codons to facilitate PCR amplification. Both ccdA$^{36–72}$ and ccdB were codon-optimized for *E. coli*, synthesized by General Biosystems, and subcloned into pUC57-Kan using PciI and NdeI restriction sites. For construction of the ccdA mutant library, we generated pUC57-Kan-2BspQI-ccdB by inserting two BspQI sites using primers BspQI-FP and BspQI-RP (**Supplementary Table S4**); this cloning step was performed in DB3.1 competent cells, which are resistant to ccdB toxicity. All plasmids were verified by Sanger sequencing, and complete vector and primer sequences are provided in **Supplementary Table S4**.

Library Construction, Selection and High-Throughput Sequencing

The SPIN-dvEvo-evolved ccdA$^{36-72}$ variants, codon-optimized for *E. coli*, were synthesized as an oligo pool containing the BspQI site by GenScript (China). The oligo pool was first amplified using PrimerSTAR HS DNA polymerase (Takara) and subsequently digested with BspQI. The digested fragments were then ligated into the BspQI-linearized pUC57-Kan-2BspQI-ccdB vector using T4 DNA ligase (Takara). Finally, the ligation products were purified and eluted in nuclease-free water, ready for electroporation.

The ligation products were electroporated into electrocompetent DB3.1 cells using a Bio-Rad Micropulser according to the manufacturer's protocol. Transformants were recovered in 10 mL of LB medium at 37°C for 1 hour. To estimate the library size, a portion of the culture was serially diluted, plated on LB agar containing kanamycin, and incubated for colony counting. Meanwhile, kanamycin was added to the main culture to a final concentration of 50 µg/mL, followed by incubation at 37°C for 10 hours. Subsequently, 100 µL of this culture was inoculated into 10 mL of fresh LB medium for amplification and subsequent plasmid extraction. The remainder of the overnight culture was harvested, resuspended in LB medium with 15% glycerol, and stored at -80°C. The initial, unselected ccdA library consisted of plasmids extracted from the CcdB-resistant DB3.1 strain. To perform functional selection, this library was electroporated into the CcdB-sensitive DH5α strain. Plasmids successfully recovered from DH5α transformants then represented the selected ccdA library. The CcdA$^{36-72}$ gene was PCR-amplified from both libraries using INDEX-containing primers. The amplicons were gel-purified and sequenced by Salus Pro platform (ShenZhen Salus Biomed Ltd)..

In vivo functional analysis of the SPIN-dvEvo-evolved CcdA variants

Selected CcdA variants (see **Supplementary Table S7**), encompassing a range of fitness scores, were cloned into a pUC57-Kan-ccdA/B expression vector. All gene sequences were synthesized and subsequently confirmed by DNA sequencing (General Biol). To evaluate *in vivo* function, 80 ng of each plasmid construct was transformed into the ccdB-sensitive *Escherichia coli* strain DH5α. Transformants were selected on LB agar plates supplemented with kanamycin. A ten-fold serial dilution series of each transformation was plated to enable quantitative assessment. After incubation (37 °C, 20 h), colony-forming units (CFUs) were counted at matched dilution factors and reported as relative survival/growth under co-expression of ccdB, where functional CcdA variants rescue colony formation (**Supplementary Fig. 6**).

Sequencing data processing

Raw reads were demultiplexed, adapter-trimmed, and quality-filtered. Reads were assigned to SPIN-dvEvo-evolved variants by matching the variable region to the SPIN-dvEvo-evolved dictionary (allowing ≤1 mismatch to tolerate sequencing error; ambiguous matches were discarded). For each variant i counts c_i^{pre} and c_i^{post} were tabulated. Samples with $<10^6$ total mapped reads were excluded. Unless noted, a small pseudocount (α=0.5) was used only for descriptive normalization of very low counts; final fitness estimates and uncertainty were obtained from DiMSum.[39]

Fitness estimation and statistical analysis

After read mapping and quality filtering, 2,363 SPIN-dvEvo-evolved variants were retained for downstream analysis. For each variant s, we denote the pre-selection and post-selection read counts as $c_{\text{pre}}(s)$and $c_{\text{post}}(s)$, with total library depths

$$N_{\text{pre}} = \sum_s c_{\text{pre}}(s), N_{\text{post}} = \sum_s c_{\text{post}}(s).$$

Counts were library-size normalized, and per-variant enrichment was defined as

$$ES(s) = \frac{c_{\text{post}}(s)/N_{\text{post}}}{c_{\text{pre}}(s)/N_{\text{pre}}}.$$

Variant fitness was then defined as the $\log_2$ enrichment without any wild-type normalization:

$$F(s) = \log_2 ES(s) = \log_2 \left(\frac{c_{\text{post}}(s)}{c_{\text{pre}}(s)}\right) - \log_2 \left(\frac{N_{\text{post}}}{N_{\text{pre}}}\right).$$

Fitness (log2 enrichment) and associated uncertainty were estimated with DiMSum (Poisson–Delta model with overdispersion correction), consistent with the definition above.

To identify significantly enriched variants, we applied an FDR-controlled significance filter based on DiMSum-reported q-values:

$$q\_\text{value} < 10^{-3},$$

For effect-size stratification, we labeled variants with log2 enrichment F(s) > 3.0 as functional and those with F(s) > 5.0 as wild-type-like

Code availability

The SPIN-dvEvo source code and the LoRA model weights for TadA and CcdA will be soon publicly available

Data availability

All data generated or analyzed in this study are included in the main text and Supplementary Information. Input and output sequence files (including training seeds, natural homolog sets, and evolved sequence libraries), as well as analysis-ready intermediate results, are publicly available at https://zhouyq-lab.szbl.ac.cn/download/. Additional materials are available from the corresponding authors upon reasonable request.

Author Contributions

ZC collected data, built the models, and performed sequence-based computational evolution. JT, TZ, and QN designed experiments and performed experimental validations. XZ helped with computational design. JZ and YZ initiated and supervised the project. YZ provided the funding support. YZ and ZC drafted the initial manuscript. All authors contributed to subsequent manuscript revision and approved the final version.

Acknowledgements

We thank the support for the National Natural Science Foundation of China (Grant Nos. 92370202) and the National Key R&D Program of China (Grant No. 2021YFF1200400). We also 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 computational support. Figure 2A was created with BioRender.com.

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.

Reference

1. Arnold, F.H. Directed Evolution: Bringing New Chemistry to Life. *Angew Chem Int Ed Engl.* **57**, 4143-4148 (2018).
2. Bloom, J.D. & Arnold, F.H. In the light of directed evolution: Pathways of adaptive protein evolution. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 9995-10000 (2009).
3. Tracewell, C.A. & Arnold, F.H. Directed enzyme evolution: climbing fitness peaks one amino acid at a time. *Curr. Opin. Chem. Biol.* **13**, 3-9 (2009).
4. Romero, P.A. & Arnold, F.H. Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.* **10**, 866-876 (2009).

5. Fowler, D.M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nature Methods* **11**, 801-807 (2014).
6. Wrenbeck, E.E., Faber, M.S. & Whitehead, T.A. Deep sequencing methods for protein engineering and design. *Curr. Opin. Struct. Biol.* **45**, 36-44 (2017).
7. Wong, T.S., Roccatano, D., Zacharias, M. & Schwaneberg, U. A Statistical Analysis of Random Mutagenesis Methods Used for Directed Protein Evolution. *J. Mol. Biol.* **355**, 858-871 (2006).
8. Romero, P.A., Krause, A. & Arnold, F.H. Navigating the protein fitness landscape with Gaussian processes. *Proc Natl Acad Sci U S A* **110**, E193-201 (2013).
9. Bedbrook, C.N. et al. Machine learning-guided channelrhodopsin engineering enables minimally invasive optogenetics. *Nat. Methods* **16**, 1176-1184 (2019).
10. Fei, H. et al. Advancing protein evolution with inverse folding models integrating structural and evolutionary constraints. *Cell* **188**, 4674-4692.e4619 (2025).
11. Luo, Y. et al. ECNet is an evolutionary context-integrated deep learning framework for protein engineering. *Nat Commun* **12**, 5743 (2021).
12. Jiang, K. et al. Rapid in silico directed evolution by a protein language model with EVOLVEpro. *Science* **387**, eadr6006 (2025).
13. Vornholt, T. et al. Enhanced Sequence-Activity Mapping and Evolution of Artificial Metalloenzymes by Active Learning. *ACS Central Science* **10**, 1357-1370 (2024).
14. Li, X. et al. An iterative deep learning-guided algorithm for directed protein evolution. *iScience* **28**, 113324 (2025).
15. Yang, J. et al. Active learning-assisted directed evolution. *Nat Commun* **16**, 714 (2025).
16. Biswas, S., Khimulya, G., Alley, E.C., Esvelt, K.M. & Church, G.M. Low-N protein engineering with data-efficient deep learning. *Nature Methods* **18**, 389-396 (2021).
17. Yang, L., Liang, X., Zhang, N. & Lu, L. STAR: A Web Server for Assisting Directed Protein Evolution with Machine Learning. *ACS Omega* **8**, 44751-44756 (2023).
18. Zhang, Q. et al. Integrating protein language models and automatic biofoundry for enhanced protein evolution. *Nat Commun* **16**, 1553 (2025).
19. Yang, J. et al. Active learning-assisted directed evolution. *Nat Commun* **16**, 714 (2025).
20. Rubin, A.F. et al. A statistical framework for analyzing deep mutational scanning data. *Genome Biol.* **18**, 150 (2017).
21. Hu, J.E. et al. LoRA: Low-Rank Adaptation of Large Language Models. *arxiv* **abs/2106.09685** (2021).
22. 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).
23. Seo, D., Koh, B., Eom, G.-e., Kim, H.W. & Kim, S. A dual gene-specific mutator system installs all transition mutations at similar frequencies in vivo. *Nucleic Acids Res.* **51**, e59-e59 (2023).
24. Moore, C.L., Papa, L.J., III & Shoulders, M.D. A Processive Protein Chimera Introduces Mutations across Defined DNA Regions In Vivo. *J. Am. Chem. Soc.* **140**, 11560-11564 (2018).
25. Zhang, S. et al. TadA orthologs enable both cytosine and adenine editing of base editors. *Nat Commun* **14**, 414 (2023).
26. Yokobori, S.-i., Kitamura, A., Grosjean, H. & Bessho, Y. Life without tRNAArg–adenosine deaminase TadA: evolutionary consequences of decoding the four CGN codons as arginine in Mycoplasmas and other Mollicutes. *Nucleic Acids Res.* **41**, 6531-6543 (2013).
27. Wu, R. et al. High-resolution de novo structure prediction from primary sequence. *bioRxiv*,

2022.07.21.500999 (2022).

28. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302-2309 (2005).
29. Dai, F. et al. Pinal: Toward *De Novo* Protein Design from Natural Language. *bioRixv*, 2024.2008.2001.606258 (2025).
30. Yuan, J.S.a.C.H.a.Y.Z.a.J.S.a.X.Z.a.F. SaProt: Protein Language Modeling with Structure-aware Vocabulary. *ICLR 2024* (2024).
31. Dauparas, J. et al. Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49-56 (2022).
32. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493-500 (2024).
33. Aghera, N.K. et al. Mechanism of CcdA-Mediated Rejuvenation of DNA Gyrase. *Structure* **28**, 562-572.e564 (2020).
34. De Jonge, N. et al. Rejuvenation of CcdB-Poisoned Gyrase by an Intrinsically Disordered Protein Domain. *Molecular Cell* **35**, 154-163 (2009).
35. Bajaj, P., Manjunath, K. & Varadarajan, R. Structural and functional determinants inferred from deep mutational scans. *Protein science : a publication of the Protein Society* **31**, e4357 (2022).
36. Chandra, S., Manjunath, K., Asok, A. & Varadarajan, R. Mutational scan inferred binding energetics and structure in intrinsically disordered protein CcdA. *Protein science : a publication of the Protein Society* **32**, e4580 (2023).
37. De Jonge, N. et al. Rejuvenation of CcdB-poisoned gyrase by an intrinsically disordered protein domain. *Mol. Cell* **35**, 154-163 (2009).
38. Chandra, S. et al. The High Mutational Sensitivity of ccdA Antitoxin Is Linked to Codon Optimality. *Mol. Biol. Evol.* **39** (2022).
39. Faure, A.J., Schmiedel, J.M., Baeza-Centurion, P. & Lehner, B. DiMSum: an error model and pipeline for analyzing deep mutational scanning data and diagnosing common experimental pathologies. *Genome Biol.* **21**, 207 (2020).
40. Faure, A.J. et al. Mapping the energetic and allosteric landscapes of protein binding domains. *Nature* **604**, 175-183 (2022).
41. Lienhard, G.E. Enzymatic catalysis and transition-state theory. *Science* **180**, 149-154 (1973).
42. Hanson, J.A. et al. Illuminating the mechanistic roles of enzyme conformational dynamics. *Proc Natl Acad Sci U S A* **104**, 18055-18060 (2007).
43. Acevedo, O. & Jorgensen, W.L. Advances in quantum and molecular mechanical (QM/MM) simulations for organic and enzymatic reactions. *Acc. Chem. Res.* **43**, 142-151 (2010).
44. Lotthammer, J.M., Ginell, G.M., Griffith, D., Emenecker, R.J. & Holehouse, A.S. Direct prediction of intrinsically disordered protein conformational properties from sequence. *Nature Methods* **21**, 465-476 (2024).
45. Kumar, S. et al. TimeTree 5: An Expanded Resource for Species Divergence Times. *Mol. Biol. Evol.* **39** (2022).
46. Feng, D.-F., Cho, G. & Doolittle, R.F. Determining divergence times with a protein clock: Update and reevaluation. *Proc Natl Acad Sci U S A* **94**, 13028-13033 (1997).
47. Konaté, M.M. et al. Molecular function limits divergent protein evolution on planetary timescales. *eLife* **8** (2019).
48. Ding, F. & Steinhardt, J. Protein language models are biased by unequal sequence sampling across

the tree of life. *ICLR 2024 Workshop on Generative and Experimental Perspectives for Biomolecular Design*, 2024.2003.2007.584001 (2024).

49. Notin, P. et al. in Proceedings of the 39th International Conference on Machine Learning, *Proceedings of Machine Learning Research* **162**. 16990--17017 (2022).
50. Madden T, C.C. BLAST+ features. *National Center for Biotechnology Information (US)* (2008).
51. Rallapalli, K.L., Ranzau, B.L., Ganapathy, K.R., Paesani, F. & Komor, A.C. Combined Theoretical, Bioinformatic, and Biochemical Analyses of RNA Editing by Adenine Base Editors. *The CRISPR journal* **5**, 294-310 (2022).
52. Park, H. & Kim, S. Gene-specific mutagenesis enables rapid continuous evolution of enzymes in vivo. *Nucleic acids research* **49**, e32 (2021).

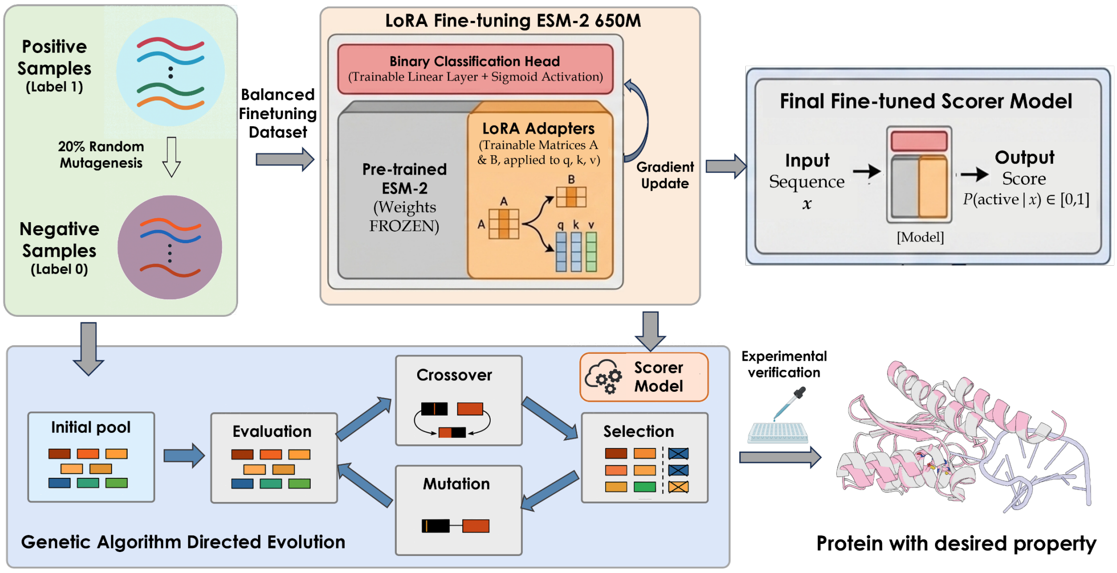

Positive Samples (Label 1)
20% Random Mutagenesis
Negative Samples (Label 0)
Balanced Finetuning Dataset
LoRA Fine-tuning ESM-2 650M
Binary Classification Head (Trainable Linear Layer + Sigmoid Activation)
Pre-trained ESM-2 (Weights FROZEN)
LoRA Adapters (Trainable Matrices A & B, applied to q, k, v)
A
B
q k v
Gradient Update
Final Fine-tuned Scorer Model
Input Sequence x
Output Score $P(\text{active} \mid x) \in [0,1]$
[Model]
Initial pool
Evaluation
Crossover
Mutation
Scorer Model
Selection
Genetic Algorithm Directed Evolution
Experimental verification
Protein with desired property

Figure 1. **Schematic overview of the framework for directed virtual evolution: SPIN-dvEvo**. A LoRA-adapted ESM-2 model is fine-tuned utilizing only a few curated positive and randomly generated negative (binary) samples. The model is then integrated into a genetic algorithm as a scorer to iteratively evolve sequences toward desired functionality but away from the original sequence cluster.

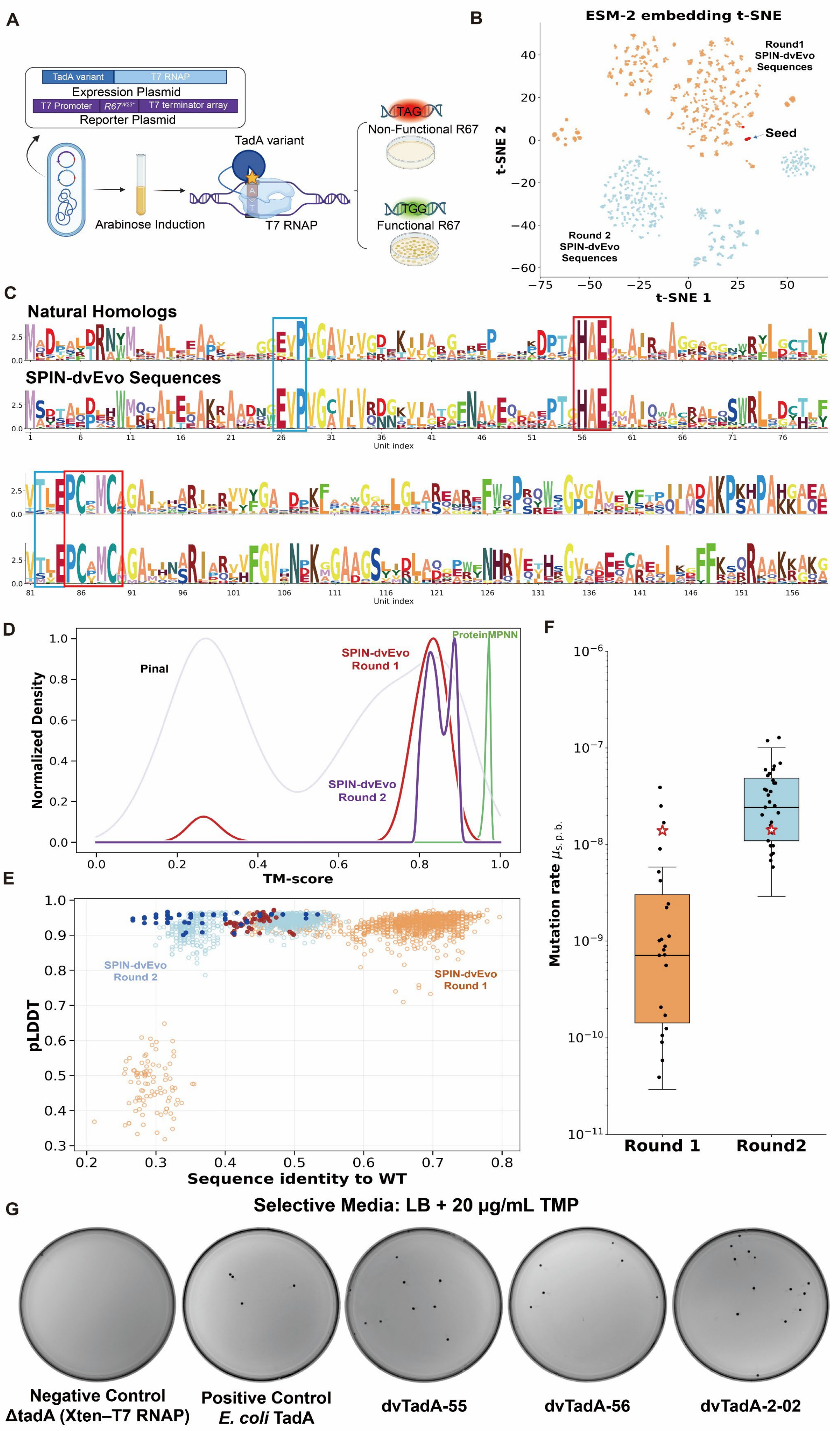

A
TadA variant
T7 RNAP
Expression Plasmid
T7 Promoter
R67W23*
T7 terminator array
Reporter Plasmid
Arabinose Induction
TadA variant
T7 RNAP
TAG
Non-Functional R67
TGG
Functional R67
B
ESM-2 embedding t-SNE
Round1 SPIN-dvEvo Sequences
Seed
Round 2 SPIN-dvEvo Sequences
t-SNE 1
t-SNE 2
C
Natural Homologs
SPIN-dvEvo Sequences
Unit index
D
Pinal
SPIN-dvEvo Round 1
SPIN-dvEvo Round 2
ProteinMPNN
Normalized Density
TM-score
E
SPIN-dvEvo Round 2
SPIN-dvEvo Round 1
pLDDT
Sequence identity to WT
F
Mutation rate $\mu_{s.p.b.}$
Round 1
Round2
G
Selective Media: LB + 20 μg/mL TMP
Negative Control ΔtadA (Xten–T7 RNAP)
Positive Control E. coli TadA
dvTadA-55
dvTadA-56
dvTadA-2-02

Figure 2. Validation of virtually evolved enzyme TadA from sequence motifs, predicted structures and experiments. (**A**) Schematic of the experimental reporter system employed for quantifying A•T-to-G•C editing activity. TadA-mediated reversion of a premature TAG stop codon to a TGG codon in the *R67* gene confers resistance to trimethoprim (TMP), enabling selection of active variants. (**B**) Newly emerged clusters from directed virtual evolution by SPIN-dvEvo according to the t-SNE projections of the base ESM-2 embeddings of the 10 starting TadA sequences to 1000 evolved sequences in Round 1 and Round 2. (**C**) Similar conserved functional and structural core motifs between virtual evolved sequences and natural homologs (top). (**D**) The accuracy for the predicted structures (according to TMscore) for 1000 TadA variants generated by four models (sequence generators Pinal and structure-based designs ProteinMPNN) compared to those given by SPIN-dvEvo in two rounds. (**E**) Scatter plot of the predicted confidence score pLDDT versus sequence identity to the wild type (*E. coli* TadA) for 1000 evolved sequences by SPIN-dvEvo in Round 1 and Round 2. The 60 experimentally tested sequences selected from Round 1 and the 60 from Round 2 are highlighted as filled points. (**F**) Boxplots comparing experimental activities of validated first- and second-round evolved TadA sequences, showing an upward-shifted distribution after including the first-round result in training. (**G**) Illustrative examples of the plates from the R67 DHFR–based *E. coli* reporter assay on TMP-selective medium. Shown are the negative control (ΔTadA cells only expressing Xten linker–T7RNAP), a positive-control TadA variant (*E. coli* TadA), and cells expressing SPIN-dvEvo-evolved TadA variants dvTadA-55, dvTadA-56 and dvTadA-2-02.

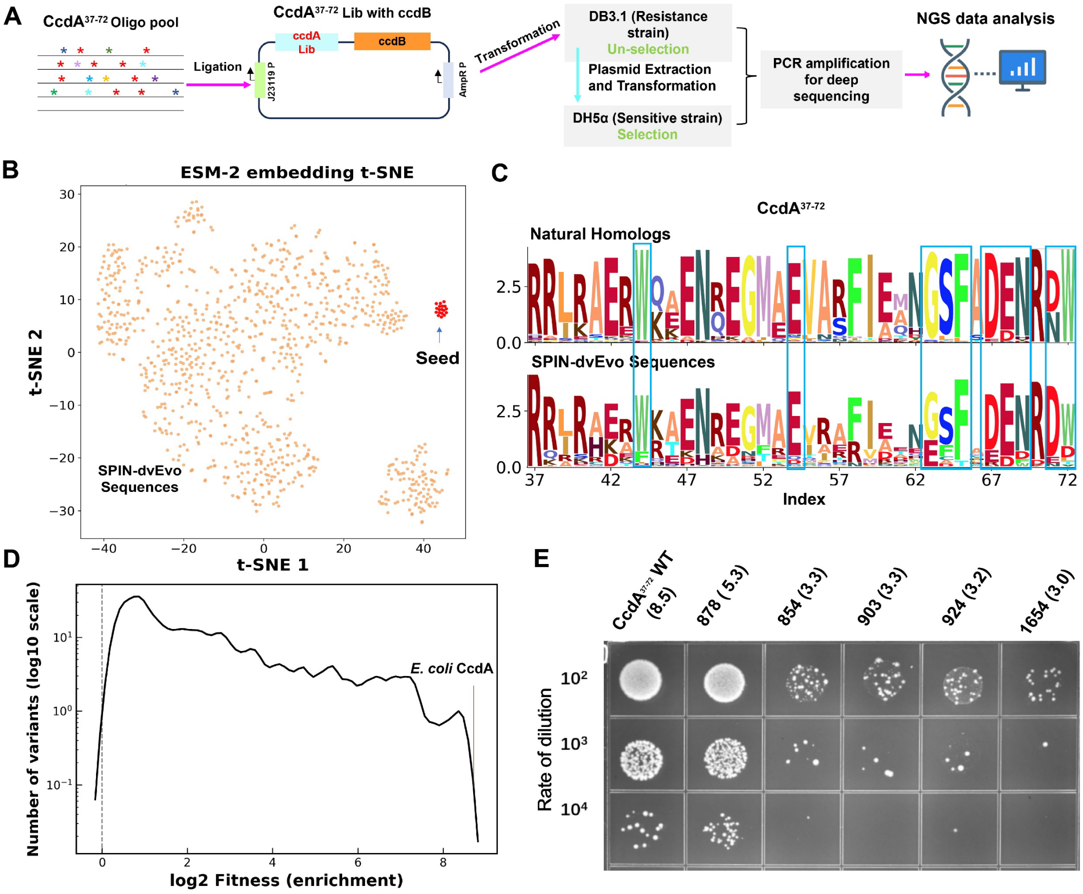
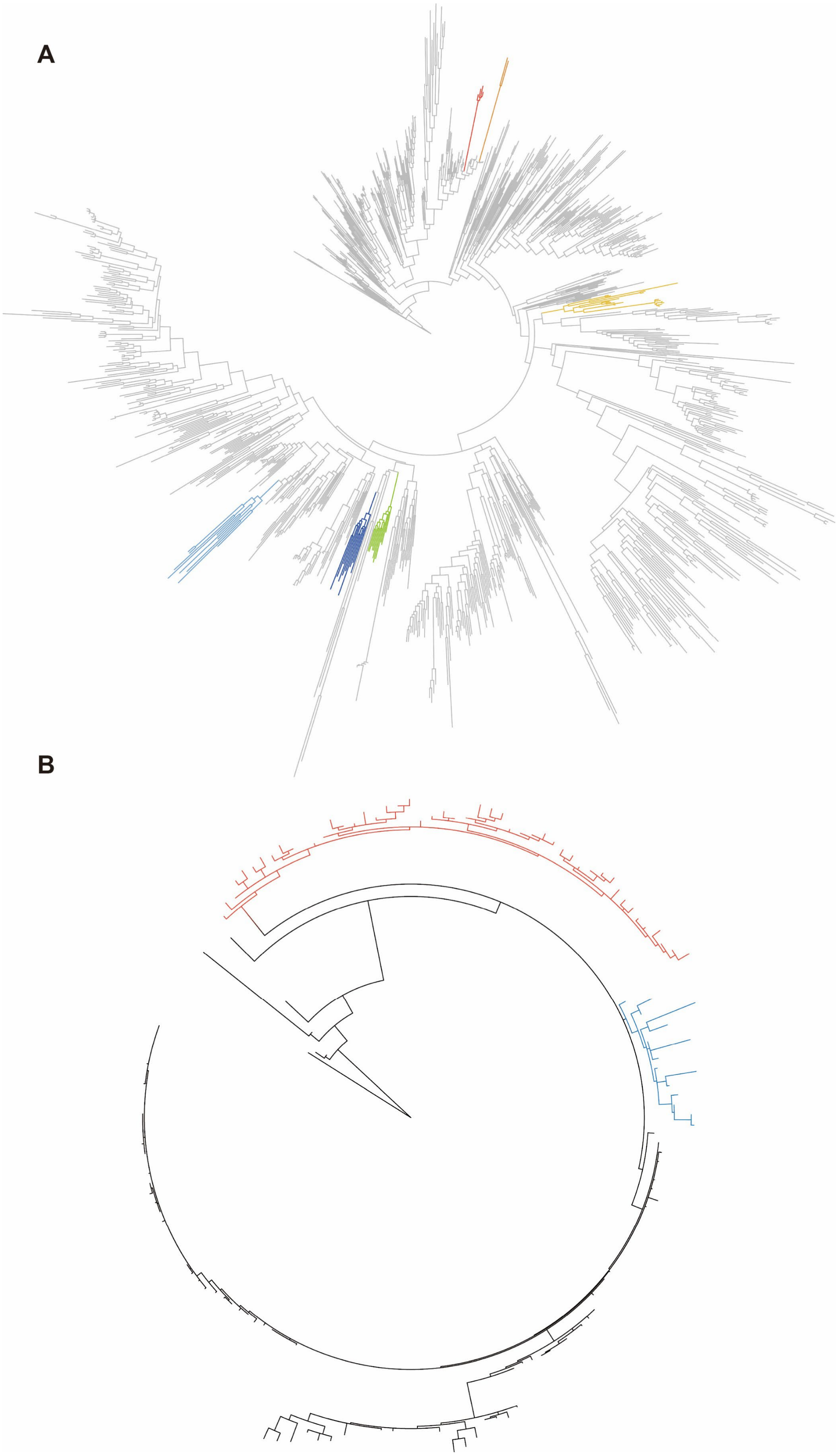


Figure 3. **Experimental validation of evolved variant library of intrinsically disordered protein CcdA.**

(**A**) Schematic diagram for high-throughput validation of evolved CcdA according to the ability of a CcdA variant that can neutralize CcdB toxin in *E. coli* growth, measured by sequence counts pre and post selections. (**B**) Emergence of new clusters in SPIN-dvEvo sequences evolved from the starting 22 natural CcdA input sequences according to the t-SNE projections of the base ESM-2 embeddings. (**C**) Sequence motifs from SPIN-dvEvo sequences are highly similar to those obtained from natural homologs according to key conserved residues highlighted in blue boxes. (**D**) The distribution in number of variants as measured fitness scores (Log_2 fitness distributions normalized by the library size). (**E**) Activity confirmation of selected variants according to their fitness. Serial 10-fold dilution spot assay showing CcdA WT from *E. coli* and five CcdA variants (1654 (Log_2 fitness = 3.0), 924 (Log_2 fitness = 3.2), 903 (Log_2 fitness = 3.3), 854 (Log_2 fitness = 3.3), and 878 (Log_2 fitness = 5.3) along with the wild type (Log_2 fitness = 8.5)) for rescuing toxin CcdB at a dilution factor of 10^2–10^4. Higher colony counts indicate stronger neutralization activity.



818

Figure 4. Phylogenetic novelty of SPIN-dvEvo TadA and CcdA variants in joint natural–evolved trees.

Maximum-likelihood phylogenies inferred from multiple sequence alignments containing natural homologs and experimentally validated SPIN-dvEvo evolved variants (sequences combined prior to alignment and tree building). Triangles denote nodes with bootstrap support in the 70–100 range. **(A)** TadA**:** alignment includes 1000 natural TadA homologs and 54 dvTadA variants. Highlighted sectors mark major, evolve-enriched dvTadA branches separated from dominant natural clades, supporting phylogenetically distinct lineages beyond the initial natural neighborhood. **(B)** CcdA: alignment includes 100 natural CcdA homologs and 62 dvCcdA variants. Light-blue and red sectors highlight two major evolved dvCcdA branches, indicating phylogenetically distinct lineages relative to the bulk of natural homologs.