Navigating Protein Fitness Landscapes Through Simulated Evolutionary Jumps

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

[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, Guangdong Province 518055, China
[5]Ribopeutic (Shenzhen) Co., Ltd., Futian, Shenzhen, 518000, China
[6]Ribopeutic Inc., Qiantang, Hangzhou, 310018, China

[§]Co-first authors. These authors contributed equally to this work.
*Corresponding authors: Institute of Systems and Physical Biology, Shenzhen Bay Laboratory, Shenzhen, 518107, China, Yaoqi Zhou, +86-(755) 2684 6275, zhouyq@szbl.ac.cn; Jian Zhan, +86-(755) 2684 6275, zhanjian@szbl.ac.cn

Abstract

Natural and directed evolution are powerful for enhancing protein function, but their utility is limited by local exploration, as each mutation must maintain function. Here we introduce SPIN-JEvo (Sequence-based Prediction with Integrated Neural Network by Jump and Evolution), a computational framework that decouples exploration from functional constraint by initiating from functional sequences and jumping to likely nonfunctional variants with ~20% random substitutions. These sequences then evolved using a genetic algorithm to reach remote homologs. Despite training on only a few binary-labeled sequences and starting from a cluster of close sequence neighbors, SPIN-JEvo identified previously inaccessible remote homologs in minutes for both a structured enzyme (tRNA-specific adenosine deaminase) and an intrinsically disordered antitoxin (CcdA)—without requiring structural information during model training or sequence generation. SPIN-JEvo shifts protein engineering from incremental local optimization to global navigation of sequence space, enabling the discovery of novel functional families beyond the reach of traditional methods.

Keywords

Protein engineering, directed evolution, protein language model, fitness landscape, sequence design

Introduction

Directed evolution is a cornerstone methodology for protein engineering, enabling stepwise

improvement of enzymes and binding proteins through iterative cycles of mutagenesis, selection, and amplification. This approach has successfully generated variants with enhanced activity, altered specificity, and improved stability for a wide range of applications in therapeutics and industrial biotechnology[1, 2]. Despite its successes, directed evolution is inherently constrained by the need for the protein to remain functional at each step. Consequently, the accessible sequence space is limited to a "local walk" from the parent sequence, typically exploring variants differing by only a few mutations[3,4]. While deep mutational scanning and next-generation sequencing have vastly increased the throughput of sequence-function measurements [4, 5 6], these explorations remain confined to the immediate neighborhood of known functional sequences, leaving the vast, uncharted regions of sequence space largely inaccessible. Moreover, most practical assays provide coarse readouts (binary active/inactive, survival, or enrichment counts) with a narrow selection window, so the effective signal is sparse and noisy—often insufficient to reliably drive long-range exploration beyond the local mutational neighborhood[4, 7].

Computational approaches offer a potential avenue to overcome this local-search limitation by redesigning sequences based on a structural template[8-10 11]. However, high-resolution structures are not always available, and even when they are, the precise relationship between structure and function can be elusive. This has motivated the development of sequence-based methods that learn surrogate models of the fitness landscape from experimental data, followed by *in silico* sequence optimization. Early efforts employed probabilistic models like Gaussian processes trained on measured activities to guide exploration.[12, 13] More recent strategies leverage deep supervised learning or protein language models (PLMs) to incorporate evolutionary context, often within active-learning pipelines for multi-round optimization[14-20]. Nevertheless, these computational methods, as demonstrated in recent works such as ECNet[14], Low-N[18], ALDE[21] and EVOLVEpro[15], have generally focused on local searches, typically exploring sequences with mutation rates of only ~0.7–2.1%, ~1.3–2.9%, ~2–3%, and ~0.5–1.5%, respectively. Furthermore, while more data-efficient predictors are emerging[18], the majority of these approaches require large, quantitative datasets, limiting their application to a small number of well-characterized protein systems.

Here, we introduce SPIN-JEvo (Sequence-based Prediction with Integrated Neural Network by Jump and Evolution), a virtual evolution framework designed to escape local sequence neighborhoods. In contrast to natural or directed evolution, which require continuous functionality along the evolutionary trajectory, SPIN-JEvo first executes a "jump" to likely nonfunctional variants bearing 20% mutation before evolving toward functional remote clusters. Crucially, it achieves this without requiring structural information or quantitative labels, learning effectively from as few as a handful of qualitative functional sequences. This data efficiency dramatically broadens its potential applicability. We applied SPIN-JEvo to two structurally and dynamically distinct targets: the enzyme TadA and the intrinsically disordered antitoxin CcdA. In both systems, starting from a localized cluster of highly homologous sequences, SPIN-JEvo successfully navigated toward remote functional regions within minutes on a desktop workstation that were previously unreachable.

Results

Jump and evolution by SPIN-JEvo

The SPIN-JEvo framework consists of two tightly coupled components. These include a task-

specific fitness predictor obtained by LoRA-based fine-tuning (LoRA)[22] of ESM-2[23]. The fitness score was generated from qualitatively labeled sequences with binary activity labels (1 for active, 0 for inactive) **(Fig. 1)**. Only a small set of positive sequences (a family of close homologous sequences) is required, along with synthetic negatives generated by introducing random substitutions at 20% of positions in the positive sequences. The second component is sequence evolution by a genetic algorithm (GA) according to the fitness score. Each run is jump-initialized by applying ~20% random substitutions to the starting positive seed pool (i.e. the training family of close homologs), after which these random variants were evolved by a genetic algorithm to improve the fitness score until convergence. Here, we chose a 20% mutation rate to balance the requirements of functional retention and sequence novelty (See discussion).

To examine whether SPIN-JEvo imposes a genuinely continuous selection pressure during sequence optimization, we tracked the population-wide activity score (p_{act}) across generations in 20 independent genetic algorithm runs. As shown in **Supplementary Fig. 1**, both the mean and median p_{act} increased smoothly over generations and gradually approached saturation at high values. The close agreement between these two summary statistics indicates that the upward shift was a population-level trend rather than being driven by a small number of extreme variants. In parallel, the between-run spread progressively narrowed at later generations, suggesting that the search became increasingly concentrated in high-scoring regions of the learned fitness landscape while preserving broader exploration at earlier stages. Together, these dynamics support that SPIN-JEvo operates through continuous, score-based selection rather than a hard survival cutoff, with incremental improvements being repeatedly retained over the course of evolution.

Jumping virtual evolution from the neighborhood of an enzyme: TadA

To evaluate the enzyme-evolution capability of SPIN-JEvo, 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[24]. 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[25, 26]. 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-JEvo to produce (evolved virtually) 1,000 sequences by starting from the inactive sequences pool (20% random mutations) and evolving for 100 steps (See Methods). We

observed that such virtual evolution started from a tightly clustered sequence region (in red) and quickly expanded 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 [27] 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[28] 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.

As protein structures play an essential role in enzymatic functions, we predicted structures of these evolved sequences and compared them to a TadA reference structure used for TM-score benchmarking (PDB ID:2B3J). We employed PLM-based OmegaFold[29] 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[30], 1 for perfect match and 0 for no match) for predicted structures of those evolved SPIN-JEvo sequences and compared it to two PLM-based sequence generators Pinal[31, 32] and structure-based protein-design method ProteinMPNN[10] as reference controls. 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 **(**see **Supplementary Information).** Most SPIN-JEvo sequences adopted near-native structures (TM-score ~0.8, 89.6% sequences with TM-score>0.5) and were only slightly worse than the structure-based method ProteinMPNN (TM-score ~0.95) (**Fig. 2D**). The 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 misfolded sequences and a smaller subset that retains the TadA fold. An example of a predicted structure for a SPIN-JEvo sequence is compared to the native structure in **Supplementary Fig. 2**, highlighting a 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[33] with a single natural MSA to save computing time) and low sequence identity ($\leqslant$0.5) to the wild type (as shown in **Fig. 2E**). Among 60 variants tested, 23 variants were found active (38.3% success rate). Activities spanned more than three orders of magnitude, with several variants matching or exceeding the reference activity of *E. coli* TadA **(Fig. 2F, Supplementary Fig. 3A, Table S3, Table S4**). These individually validated functional sequences spanned 39–79% amino-acid identity to the *E. coli* TadA wild type, confirming that SPIN-JEvo can effectively "jump" significantly beyond the near neighbors of the initial seeds of a narrow sequence neighborhood ($\geqslant$88% identity) to new functional neighborhoods as shown in **Fig. 2B**.

After incorporating labels from the 60 experimentally tested variants, we re-trained the LoRA model with the enlarged binary-labelled dataset and performed sequence evolutions again by GA. The new 1000 evolved sequences (Round II) formed new sequence clusters (**Fig. 2B**). The TM-score distribution of predicted structures for the second-round sequences improves over that of the first round. All predicted structures (100%) now have TM-score > 0.78 and highest peak at a TM-score of 0.88, compared to 0.80 in the first round (**Fig. 2D**). We tested 60 new variants chosen based on high AlphaFold 3 pLDDT scores 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 refined the navigation map, allowing the evolution to bypass local traps more effectively **(Supplementary Fig. 3B, Table S3, Table S5)**. More importantly, 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**), indicating that iterative refinement enables consecutive jumps into deeper phylogenetic territory while simultaneously climbing toward higher-fitness peaks.

A few selected variants are illustrated along with positive and negative controls by plating on TMP-selective medium (jTadA-55 and jTadA-56 from round 1; jTadA-2-02 from round 2). These evolved sequences produced TMP-resistant colonies comparable to or more numerous than those of the positive control (*E. coli* TadA), whereas no colonies were observed in the negative control (expressing only an Xten linker-T7RNAP cassette in place of TadA and thus lacking deaminase activity) (**Fig. 2G**).

Jumping virtual evolution of an intrinsically disordered antitoxin, CcdA

To test whether SPIN-JEvo 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. We engineered only its C-terminal segment ($CcdA^{36–72}$; 36 residues), which mediates binding to CcdB and neutralizes toxicity by blocking the CcdB–GyrA interaction[34, 35]. We started from the canonical *E. coli* CcdA (P62552) and retrieved CcdA family homologs from closely related *Enterobacterales/ Gammaproteobacteri*a. Incomplete or atypical entries as well as those sequences at 100% sequence identity cutoff were removed. 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-JEvo to bind and neutralize CcdB, enabling straightforward functional selection **(Fig. 3A**).

As in the TadA case, we first evolved 1,000 CcdA variants by SPIN-JEvo. 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-JEvo 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 such as W44, E54,[36] G63, S64, F65, D71 and W72[37, 38](**Fig. 3C,** blue box) were preserved in natural CcdA homologs were captured during virtual evolution by SPIN-JEvo, despite that it was started from a highly local seed set.

To test these sequences experimentally, we synthesized 3,041 evolved CcdA variants for plasmid construction. This set included the original 3,000 evolved sequences and 50 additional variants added to offset 9 sequences that failed an in silico reverse translation–translation check. We evaluated the resulting library using a pooled bacterial growth selection, because bacterial growth correlates with the ability of 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[39] (**Fig. 3A**). We estimated enrichment and uncertainty with the DiMSum pipeline[40,41] 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. Applying an FDR-controlled filter relative to internal stop-codon negative controls of $q\_\text{value} < 10^{-3}$ 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**). There are 26 variants with $\log_2(\text{fitness}) > 5$, a few of which are 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 $\log_2(\text{fitness}) = 3.3, 3.3, 3.2, \text{and } 3.0, \text{respectively,}$ 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 individually confirmed that all variants with $\log_2(\text{fitness}) \geq 3.0$ are functional. The functional activities of these variants are consistent with calculated fitness scores from the high-throughput experiment. For example, the variant V878 with a $\log_2$ fitness score of 5.3 value can grow well even at the dilution factor of 10^4 whereas the variant 1654 with $\log_2(\text{fitness}) = 3.0$ can only grow 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 36-amino-acid *E. coli* CcdA segment (**Supplementary Fig. 4**), consistent with discovery of new functional neighborhoods shown in Fig. 3B.

Discussion

SPIN-JEvo 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-JEvo redefines the search process as a navigation across the fitness landscape, 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 "non-local" virtual evolution by "jumping" to remote but nonfunctional sequences prior to

evolving by a genetic algorithm. This initial jump biases the search toward remote regions of sequence space and reduces reliance on strictly local optimization.

We chose 20%-mutated sequences as both the synthetic negatives and the jump-initialized seeds to balance learnability and sequence novelty. If the sequence identity is too low (i.e. the negative sequences are too random), it would be difficult for the LoRA to learn the intrinsic difference between functional and nonfunctional sequences. It would also be difficult to use a score function to bring a purely random sequence back to the functional region. If the sequence identity is too high, the ability to locate remote functional homologs will be limited. To systematically assess this trade-off, we varied the divergence of the jump-initialized seeds (15–40% random substitutions) and evaluated the resulting samples in terms of both structural retention and sequence novelty. Seeding at ~20% random substitutions preserved a predominantly near-native fold distribution, as measured by TM-scores between OmegaFold-predicted structures and the *E. coli* TadA reference (**Supplementary Fig. 5A**), while still shifting the sampled sequences toward substantially lower sequence identity (**Supplementary Fig. 5B**). By comparison, a lower mutation rate (15%) would increase structural recovery (**Supplementary Fig. 5A**) at the expense sequence novelty (**Supplementary Fig. 5B**), whereas a higher mutation rate (25% or more) failed to recover native like structures. This identifies ~20% divergence as a practical operating regime that balances exploration with recoverability under LoRA-guided evolution. By contrast, more aggressive initialization further expanded sequence-space coverage but produced a larger fraction of low-TM-score variants, indicating reduced structural reliability and weaker effective guidance back toward the functional scaffold.

For TadA, no structural information was used to train SPIN-JEvo or to guide sequence evolution. Yet most evolved TadA variants have TadA structural folds (**Fig. 2D, Supplementary Fig. 2**) in the first round (89.7% of sequences with predicted structural accuracy >0.5 in TM-score). A minor peak with TM-score<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, the negatives generated by 20% random mutation may include false negatives. Interestingly, the second-round success rate increased from 38% to 51% along with a one-order-of-magnitude improvement in enzymatic activity. This suggests that SPIN-JEvo can reach the remote regions with improved functions, despite lacking quantitative labels.

We have selected sequences with high confidence in predicted structures for experimental validations. The success rate in the TadA activity for SPIN-JEvo evolved sequences in Round I is reasonably high (38%), considering the fact that SPIN-JEvo is exploring remote sequence space. However, this success rate is much lower than ~90% success rate in achieving TM-score>0.5 in structural accuracy, suggesting that the structural fold alone is not sufficient as an indicator of enzymatic activity. This is understandable 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.[42-44] Future work should incorporate descriptors more directly linked to catalysis, such as active-site geometry, conformational dynamics or experimentally measured

quantitative activity.

Although the success rate for the disordered CcdA is lower than that for the TadA enzyme, it represents a successful navigation through a highly degenerate and noisy fitness landscape without structural supervision. 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.[45] Despite the lower hit rate, SPIN-JEvo recovered functional CcdA variants without using CcdB sequence information or predicted complex structures.

To further confirm newly found active sequences as new functional clusters, we constructed new phylogenetic trees using both natural functional sequences collected previously (1000 for TadA[27] and 100 for CcdA[46]) and experimentally validated, SPIN-JEvo sequences. As shown in **Fig. 4A** and **Fig. 4B**, both virtually evolved TadA and CcdA formed 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[47-49](See methods in SI, **Fig. 4**)). Similarly, the virtually evolved CcdA clade was estimated to have diverged approximately 2.508 Ga ago from a Gammaproteobacteria-associated branch (**Fig. 4**). By comparison, these virtual evolutions took only 713 seconds for TadA and 761 seconds for CcdA by SPIN-JEvo on a workstation equipped with an AMD EPYC 9654 (96-core, 2.4 GHz) CPU and an NVIDIA RTX 4090 GPU (24 GB). The ability to traverse such immense evolutionary distances in approximately 12 minutes signifies a paradigm shift from local neighborhood search to global exploration of sequence space through a 20% mutation jump.

SPIN-JEvo 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. By demonstrating that sparse binary signals are sufficient to guide long-range jumps, SPIN-JEvo shifts the bottleneck of protein engineering from data collection to hypothesis testing. One immediate improvement for SPIN-JEvo 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 could provide a more accurate evolution direction than a classification head. This is a subject of an ongoing study.

One limitation of SPIN-JEvo 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 more beneficial than ESM-2 in the jump and evolution scheme. 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.[50,

[51] Further studies in this area are needed.

Moreover, current implementation of SPIN-JEvo 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[24] (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 (w, 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 $= 5 \times 10^{-4}$, 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[52] 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.

Structure prediction for SPIN-JEvo sequences

SPIN-JEvo 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)[29] 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[53]. TM-scores reported in the main text refer to alignments between the native structure (PDB 2B3J) and OmegaFold-predicted structures for SPIN-JEvo-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™ 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 the engineered TadA variants and T7 RNA polymerase (T7RNAP) were constructed based on the pDae079 vector system. Since the original pDae079 vector contains two deaminase domains, we modified the backbone to retain only a single deaminase component. Specifically, the gene sequence encoding the SPIN-JEvo-evolved TadA variant was inserted to replace the original deaminase moieties via homologous recombination. A negative control plasmid (pT7RNAP-ΔTadA), expressing only an Xten linker-T7RNAP cassette in place of TadA and thus lacking deaminase activity, was also constructed using homologous recombination in the pDae079 backbone.[25].

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.[54] 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℃ for 12–16 hours. Individual colonies were then inoculated into 10 mL of LB broth supplemented with 100 μg/mL ampicillin, 25 μg/mL chloramphenicol, followed by overnight incubation at 37℃ with shaking at 220 rpm.

On the following day, the overnight cultures were diluted 1:100 into fresh LB medium containing the same concentrations of ampicillin and chloramphenicol, supplemented with 0.2% (w/v) L-arabinose to induce TadA expression. To promote the accumulation of mutations during active

growth, these cultures were maintained by serial 1:100 dilutions into identical fresh induction medium every 4 hours over a 24-hour period, with incubation at 37℃ and shaking at 220 rpm

Editing activity Assay

At the 24-h 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 $E[f] \approx \mu \ln(R_{\text{eff}})$. Each 4-h propagation round used a 1:100 reinoculation followed by regrowth to saturation, corresponding to ~ 6.6 generations (G). Assuming binary fission ($R_{\text{eff}} = 2^G$), $\ln(R_{\text{eff}}) = G \ln 2 \approx 4.57$. 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 $R67^{W23*}$ 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 S6**); 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 S6**.

Library Construction, Selection and High-Throughput Sequencing

The SPIN-JEvo-evolved ccdA$^{36\text{-}72}$ 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$^{36\text{-}72}$ 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-JEvo-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-JEvo-evolved variants by matching the variable region to the SPIN-JEvo-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.[40]

Fitness estimation and statistical analysis

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

$$N_{\mathrm{pre}} = \sum_{s} c_{\mathrm{pre}}\,(s), N_{\mathrm{post}} = \sum_{s} c_{\mathrm{post}}\,(s).$$

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

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

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

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

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:

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

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-JEvo source code and LoRA model weights for TadA and CcdA are publicly available at https://github.com/chenzh-hash/SPIN-JEvo.

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 National Natural Science Foundation of China (Grant No. 92370202) for support. 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.

Competing interests

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

References

1. Arnold, F.H. Directed Evolution: Bringing New Chemistry to Life. *Angew. Chem. Int. Ed.* **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. Fowler, D.M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11**, 801-807 (2014).
5. 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).

6. 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).
7. Yang, K.K., Wu, Z. & Arnold, F.H. Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **16**, 687-694 (2019).
8. Khersonsky, O. et al. Optimization of the in-silico-designed kemp eliminase KE70 by computational design and directed evolution. *J. Mol. Biol.* **407**, 391-412 (2011).
9. Goldenzweig, A. et al. Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability. *Mol. Cell* **63**, 337-346 (2016).
10. Dauparas, J. et al. Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49-56 (2022).
11. Gomez de Santos, P. et al. Repertoire of Computationally Designed Peroxygenases for Enantiodivergent C-H Oxyfunctionalization Reactions. *J. Am. Chem. Soc.* **145**, 3443-3453 (2023).
12. 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).
13. Bedbrook, C.N. et al. Machine learning-guided channelrhodopsin engineering enables minimally invasive optogenetics. *Nat. Methods* **16**, 1176-1184 (2019).
14. Luo, Y. et al. ECNet is an evolutionary context-integrated deep learning framework for protein engineering. *Nat. Commun.* **12**, 5743 (2021).
15. Jiang, K. et al. Rapid in silico directed evolution by a protein language model with EVOLVEpro. *Science* **387**, eadr6006 (2025).
16. Vornholt, T. et al. Enhanced Sequence-Activity Mapping and Evolution of Artificial Metalloenzymes by Active Learning. *ACS Central Science* **10**, 1357-1370 (2024).
17. Li, X. et al. An iterative deep learning-guided algorithm for directed protein evolution. *iScience* **28**, 113324 (2025).
18. Biswas, S., Khimulya, G., Alley, E.C., Esvelt, K.M. & Church, G.M. Low-N protein engineering with data-efficient deep learning. *Nat. Methods* **18**, 389-396 (2021).
19. 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).
20. Zhang, Q. et al. Integrating protein language models and automatic biofoundry for enhanced protein evolution. *Nat Commun* **16**, 1553 (2025).
21. Yang, J. et al. Active learning-assisted directed evolution. *Nat. Commun.* **16**, 714 (2025).
22. Hu, J.E. et al. LoRA: Low-Rank Adaptation of Large Language Models. *arxiv* **abs/2106.09685** (2021).
23. Lin, Z. et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv* 2022.07.20.500902 (2022).
24. 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).
25. 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).
26. 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).
27. Zhang, S. et al. TadA orthologs enable both cytosine and adenine editing of base editors. *Nat. Commun.* **14**, 414 (2023).

28. 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).
29. Wu, R. et al. High-resolution de novo structure prediction from primary sequence. *bioRxiv*, 2022.07.21.500999 (2022).
30. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302-2309 (2005).
31. Dai, F. et al. Pinal: Toward *De Novo* Protein Design from Natural Language. *bioRixv*, 2024.08.01.606258 (2025).
32. Su, J., Han, C., Zhou, Y., Shan, J., Zhou, X. & Yuan, F. SaProt: Protein Language Modeling with Structure-aware Vocabulary. *ICLR 2024* (2024).
33. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493-500 (2024).
34. De Jonge, N. et al. Rejuvenation of CcdB-Poisoned Gyrase by an Intrinsically Disordered Protein Domain. *Molecular Cell* **35**, 154-163 (2009).
35. Aghera, N.K. et al. Mechanism of CcdA-Mediated Rejuvenation of DNA Gyrase. *Structure* **28**, 562-572.e564 (2020).
36. Bajaj, P., Manjunath, K. & Varadarajan, R. Structural and functional determinants inferred from deep mutational scans. *Protein Science* **31**, e4357 (2022).
37. Chandra, S., Manjunath, K., Asok, A. & Varadarajan, R. Mutational scan inferred binding energetics and structure in intrinsically disordered protein CcdA. *Protein Science***32**, e4580 (2023).
38. De Jonge, N. et al. Rejuvenation of CcdB-poisoned gyrase by an intrinsically disordered protein domain. *Mol. Cell* **35**, 154-163 (2009).
39. Chandra, S. et al. The High Mutational Sensitivity of ccdA Antitoxin Is Linked to Codon Optimality. *Mol. Biol. Evol.* **39** (2022).
40. 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).
41. Faure, A.J. et al. Mapping the energetic and allosteric landscapes of protein binding domains. *Nature* **604**, 175-183 (2022).
42. Lienhard, G.E. Enzymatic catalysis and transition-state theory. *Science* **180**, 149-154 (1973).
43. Hanson, J.A. et al. Illuminating the mechanistic roles of enzyme conformational dynamics. *Proc Natl Acad Sci U S A* **104**, 18055-18060 (2007).
44. 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).
45. Lotthammer, J.M., Ginell, G.M., Griffith, D., Emenecker, R.J. & Holehouse, A.S. Direct prediction of intrinsically disordered protein conformational properties from sequence. *Nat. Methods* **21**, 465-476 (2024).
46. Wu, A.Y., Kamruzzaman, M. & Iredell, J.R. Specialised functions of two common plasmid mediated toxin-antitoxin systems, ccdAB and pemIK, in Enterobacteriaceae. *PloS one* **15**, e0230652 (2020).
47. Kumar, S. et al. TimeTree 5: An Expanded Resource for Species Divergence Times. *Mol. Biol. Evol.* **39** (2022).
48. 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).

49. Konaté, M.M. et al. Molecular function limits divergent protein evolution on planetary timescales. *eLife* **8** (2019).
50. Ding, F. & Steinhardt, J. Protein language models are biased by unequal sequence sampling across the tree of life. *ICLR* 2024.03.07.584001 (2024).
51. Notin, P. et al. in Proceedings of the 39th International Conference on Machine Learning, *Proc. Mach. Learn. Res* **162**. 16990-17017 (2022).
52. Madden T, C.C. BLAST+ features. *National Center for Biotechnology Information* (2008).
53. 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).
54. Park, H. & Kim, S. Gene-specific mutagenesis enables rapid continuous evolution of enzymes in vivo. *Nucleic acids research* **49**, e32 (2021).

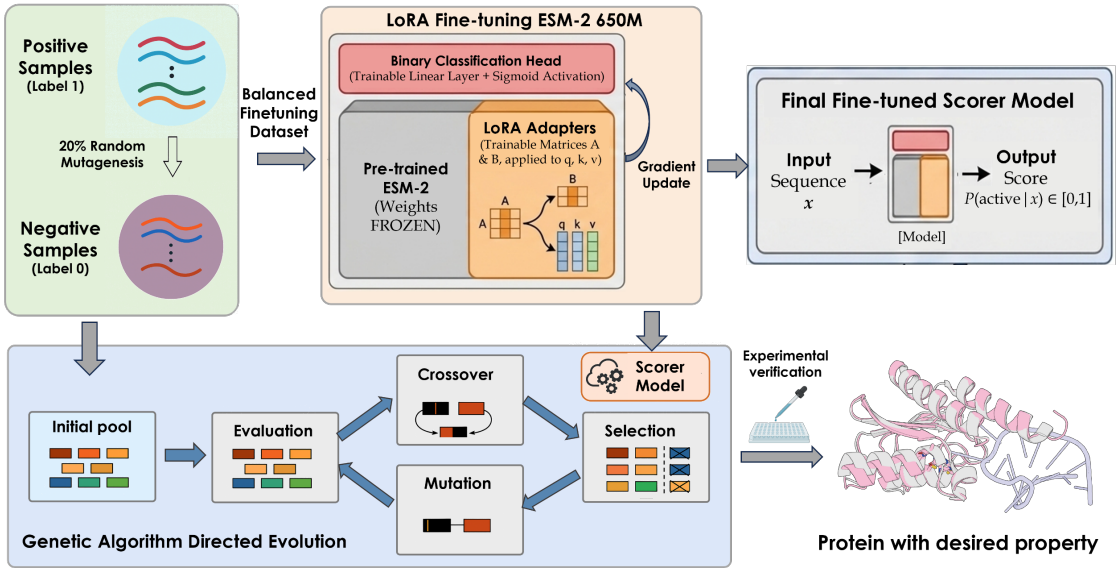

Figure 1. **Schematic overview of the framework for virtual jump evolution: SPIN-JEvo**. A LoRA-adapted ESM-2 model is fine-tuned utilizing only a few curated positive and randomly generated negative (binary) samples. The model is then integrated into a genetic algorithm as a scorer to iteratively evolve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

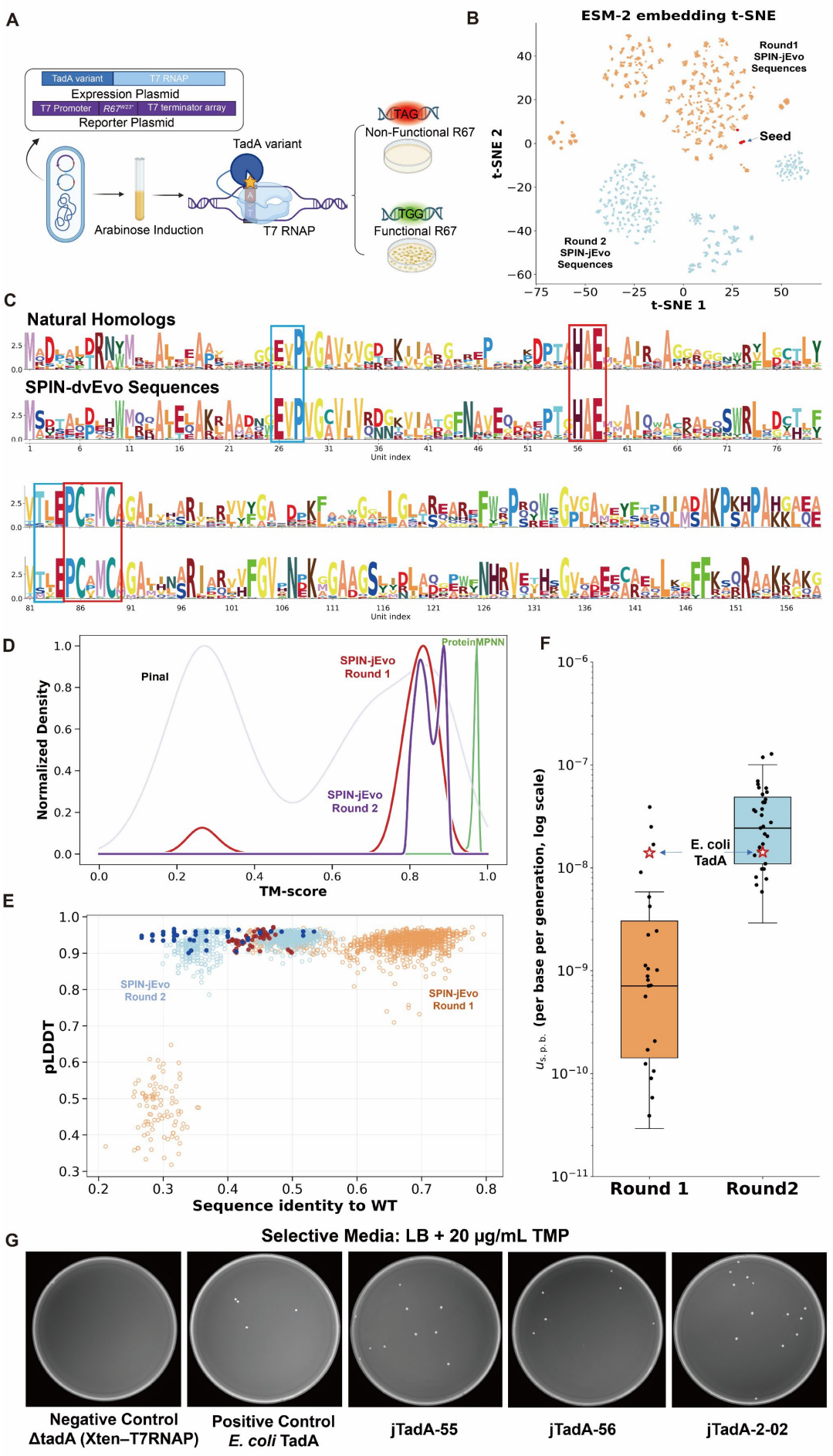

Figure 2. Validation of virtually evolved enzyme TadA from sequence motifs, predicted structures and experiments. (**A**) Schematic of the experimental reporter system employed for quantifying A•T-to-G•C editing activity. TadA-mediated reversion of a premature TAG stop codon to a TGG codon in the *R67* gene confers resistance to trimethoprim (TMP), enabling selection of active variants. (**B**) Newly emerged clusters from virtual jump evolution by SPIN-JEvo 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 TM-score) for 1000 TadA variants generated by four models (sequence generators Pinal and structure-based designs ProteinMPNN) compared to those given by SPIN-JEvo 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-JEvo 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-JEvo-evolved TadA variants jTadA-55, jTadA-56 and jTadA-2-02.

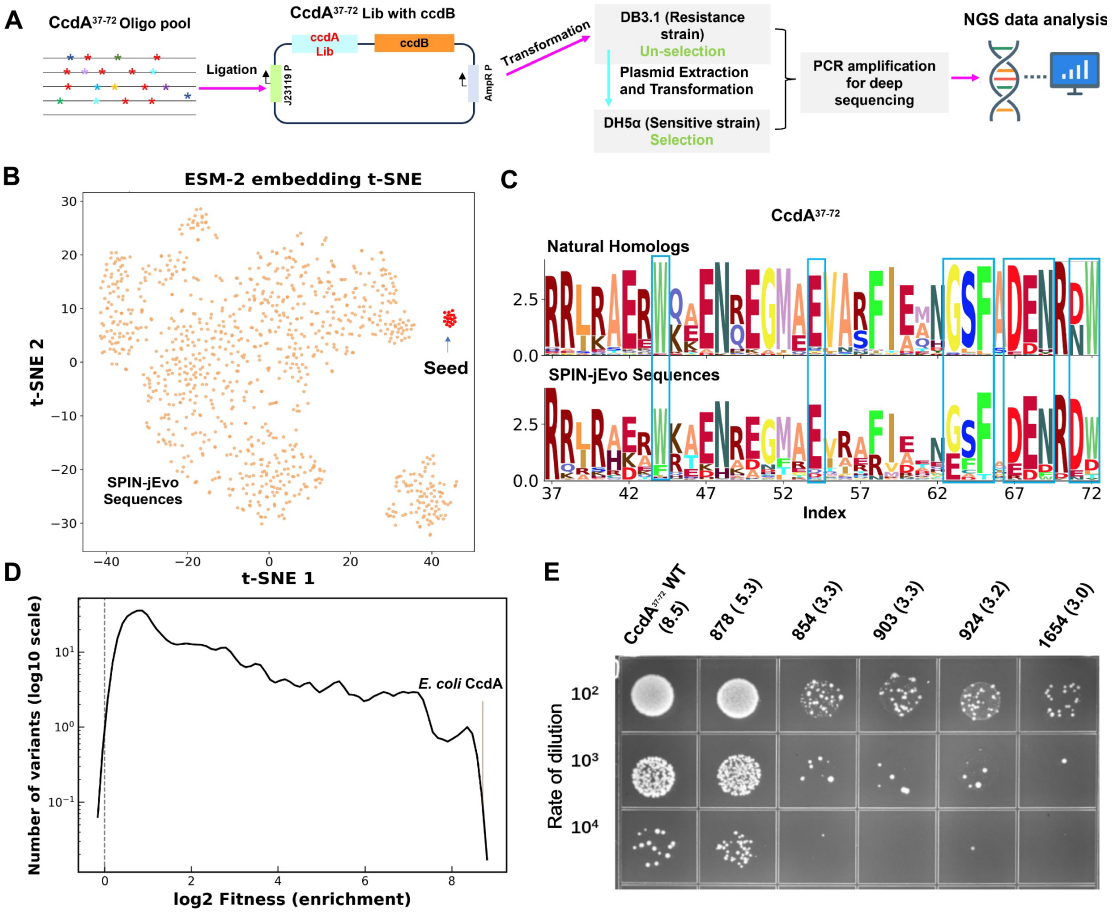

Figure 3. Experimental validation of evolved variant library of intrinsically disordered protein CcdA.

(**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-JEvo 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-JEvo 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.

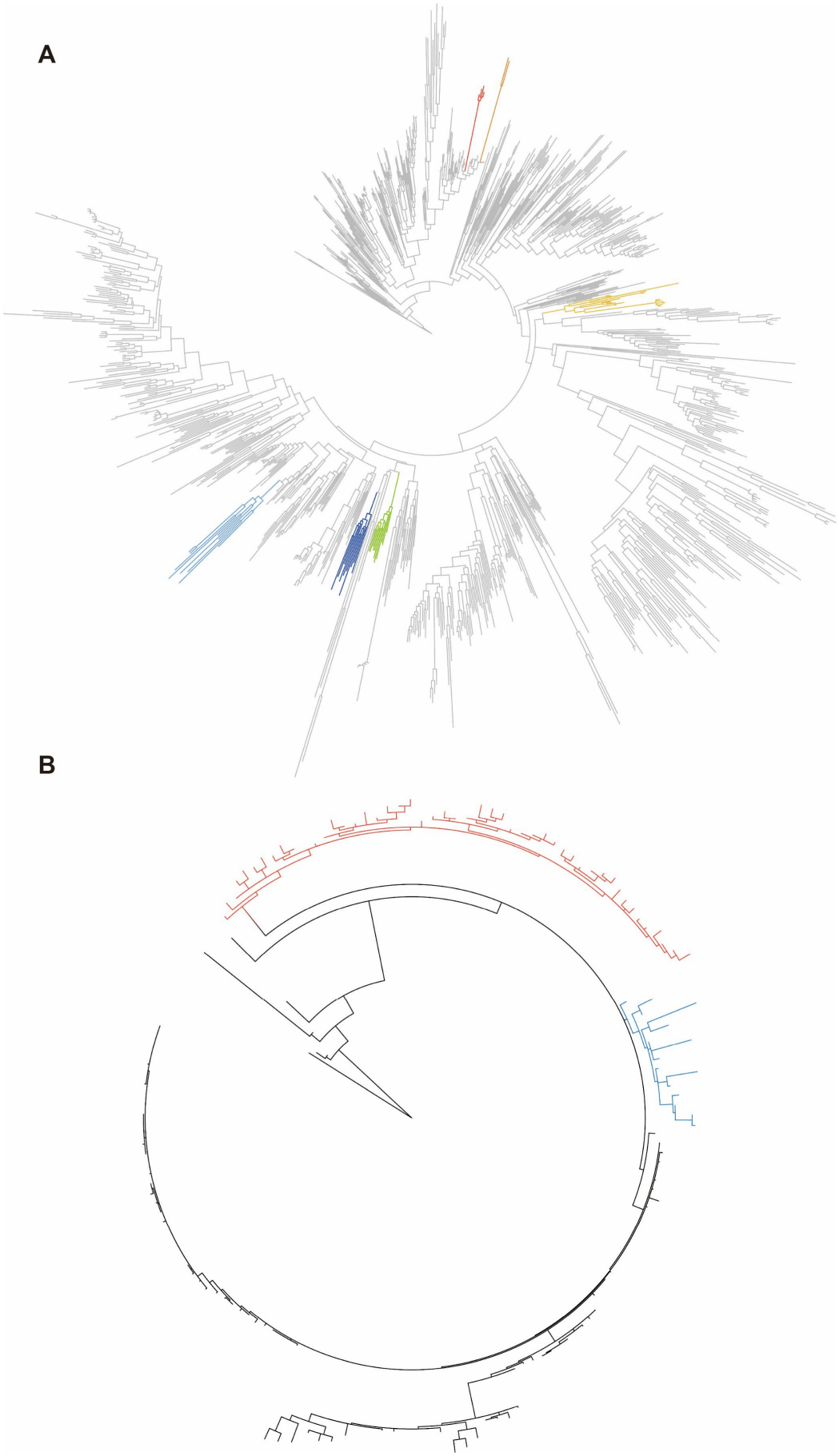

Fig.4. Phylogenetic novelty of SPIN-JEvo TadA and CcdA variants in joint natural–evolved trees.

Maximum-likelihood phylogenies inferred from multiple sequence alignments containing natural homologs and experimentally validated SPIN-JEvo 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 jTadA variants. Highlighted sectors mark major, evolve-enriched jTadA 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 jCcdA variants. Light-blue and red sectors highlight two major evolved jCcdA branches, indicating phylogenetically distinct lineages relative to the bulk of natural homologs.