Cite as: K. Jiang *et al.*, *Science* 10.1126/science.adr6006 (2024).

Rapid in silico directed evolution by a protein language model with EVOLVEpro

Kaiyi Jiang[1,2,3,4]†, Zhaoqing Yan[1,2,3]†, Matteo Di Bernardo[5]*, Samantha R. Sgrizzi[1,2,3], Lukas Villiger[6], Alisan Kayabolen[1,2,3], B.J. Kim[7], Josephine K. Carscadden[1,2,3], Masahiro Hiraizumi[8], Hiroshi Nishimasu[8,9,10], Jonathan S. Gootenberg[1,2,3]*‡, Omar O. Abudayyeh[1,2,3]*‡

[1]Department of Medicine Division of Engineering in Medicine Brigham and Women's Hospital Harvard Medical School, Boston, MA, USA. [2]Gene and Cell Therapy Institute Mass General Brigham, Cambridge, MA, USA. [3]Center for Virology and Vaccine Research Beth Israel Deaconess Medical Center Harvard Medical School, Boston, MA, USA. [4]Department of Bioengineering Massachusetts Institute of Technology, Cambridge, MA, USA. [5]Whitehead Institute Massachusetts Institute of Technology, Cambridge, MA, USA. [6]Department of Dermatology and Allergology Kantonspital St. Gallen, St. Gallen, Switzerland. [7]Koch Institute for Integrative Cancer Research at MIT Massachusetts Institute of Technology, Cambridge, MA, USA. [8]Department of Chemistry and Biotechnology, Graduate School of Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan. [9]Structural Biology Division, Research Center for Advanced Science and Technology, The University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo, Japan. [10]Inamori Research Institute for Science, 620 Suiginya-cho, Shimogyo-ku, Kyoto, Japan. †These authors contributed equally to this work. ‡These authors contributed equally to this work.

*Corresponding author. E-mail: omar@abudayyeh.science (O.O.A.); jgoot@mit.edu (J.S.G.)

Directed protein evolution is central to biomedical applications but faces challenges like experimental complexity, inefficient multi-property optimization, and local maxima traps. While *in silico* methods using protein language models (PLMs) can provide modeled fitness landscape guidance, they struggle to generalize across diverse protein families and map to protein activity. We present EVOLVEpro, a few-shot active learning framework that combines PLMs and regression models to rapidly improve protein activity. EVOLVEpro surpasses current methods, yielding up to 100-fold improvements in desired properties. We demonstrate its effectiveness across six proteins in RNA production, genome editing, and antibody binding applications. These results highlight the advantages of few-shot active learning with minimal experimental data over zero-shot predictions. EVOLVEpro opens new possibilities for AI-guided protein engineering in biology and medicine.

Protein diversity has been shaped by billions of years of evolutionary pressure, filtering the potential design space for diverse biological functions. Understanding the connection between protein amino acid sequence and function is crucial for advancing biology and developing new therapeutics. Experimental approaches like deep mutational scanning (DMS) can directly measure functional effects of protein mutations (*1*–*3*) but are limited to exploring only a fraction of the possible protein sequence space. Computational approaches leveraging orthologous sequences can reduce the experimental data needed to map fitness landscapes and optimize proteins (*4*–*6*), but these methods require multiple sequence alignments, high ortholog abundance, and minimal sequence length variation.

To overcome these challenges, fitness can be inferred by training broadly across evolutionary diversity. Protein language models (PLMs), such as ESM2 (*7*), are trained across comprehensive protein sequence databases to fill in missing amino acids. PLMs learn informative biological representations (*7*–*12*) protein structure prediction (*7*) and functional annotation (*13*). PLMs have nominated protein mutants with improved activity (*14*, *15*) and generative PLMs (*16*–*18*) have been used to design novel proteins. However, zero-shot predicted mutants have limited success (*14*, *15*) and de novo-designed proteins typically exhibit lower or comparable activity relative to natural wild-type (WT) sequences (*16*, *19*). While zero-shot models can predict antibody mutations to increase binding affinity, they cannot improve other important antibody features, such as developability and immunogenicity (*15*, *20*). These failures of PLMs to substantially improve protein activity in zero-shot settings are driven by their inability to generalize to new contexts due to limited training data (*21*) and the difference between evolutionary fitness and protein function. Therefore, protein optimization and interpretation using PLM-based approaches require additional experimental data to reach design-specific objectives.

Iterative approaches for optimization, such as directed evolution (DE) (*22*), take advantage of the smoothness of protein fitness landscapes to improve function. Although these methods are successful in contexts with suitable activity landscapes and screening methods, they can be labor intensive and fail on rugged landscapes, especially when trapped in local optima. Some improvements can be gained from combining DE with machine learning. Machine learning–directed evolution (MLDE) methods (*23*–*28*) leveraging active learning have effectively improved diverse proteins but at the cost of comprehensive experimental evaluation. Merging active learning with simpler protein representation models, such as recurrent neural networks (*29*, *30*), has simplified the evolution process, but previous attempts at active learning on protein models (*30*) have not generalized well beyond proof-of-concept demonstrations like fluorescent protein engineering due to shortcomings in the protein representation space.

Given the limited success of PLM-based methods to rank variant effects in a zero-shot fashion (*15*) or to iteratively nominate protein mutations (*30*), we reasoned that active learning with high-performance PLMs and small sets of mutants would improve directed evolution performance. Here we present the protein evolution model EVOLVEpro (evolution via language model–guided variance exploration for proteins) to solve these challenges with MLDE. EVOLVEpro nominates high-activity protein variants with active learning and minimal experimental testing, achieving rapid prediction of high-activity mutants. This performance stems from a modular approach that marries an evolutionary-scale PLM with a top-layer regression model to learn a protein's activity landscape and guide the directed evolution process *in silico*. This top-layer model is trained over multiple rounds of evolution, with each round evaluating a small set of EVOLVEpro-predicted mutants using one or more experimental assays. These data are then used to update the model and predict the next round of mutation candidates. We hypothesize that combining the PLM and regressor enables the identification of multiple high-activity regions simultaneously, providing generalization across diverse protein classes with rugged activity landscapes and robustness to local optima due to detrimental or neutral mutants (*28*). Moreover, EVOLVEpro prompting only uses protein sequences and does not require structural information, expert knowledge, or prior data. We demonstrate EVOLVEpro's ability to evolve multiple activities of a protein simultaneously, opening up vast possibilities for its use in biology and medicine.

Development and benchmarking of the EVOLVEpro model

We developed a deep learning-based directed evolution framework, EVOLVEpro, combining (i) a PLM to encode protein sequences into a continuous latent space to facilitate activity optimization and (ii) a top-layer regression model to learn the mapping between latent space and activity from a few number of data points (i.e., the low-N regime). EVOLVEpro actively learns the fitness landscape across multiple rounds of evolution. In each round, the regression model ranks protein sequences according to their predicted activity, selecting top-ranked sequences for experimental validation. Cycles are performed iteratively to improve defined protein activities until they reach desired levels (Fig. 1A).

We first optimized EVOLVEpro's computational framework *in silico* by curating 12 deep mutational scanning (DMS) datasets (*31*–*43*) (table S2 and data S2), allowing the selection of optimal architecture and parameters using simulated runs prior to any experimental testing. This simulation revealed ground truth activity data to EVOLVEpro for only the variants nominated by the model. The twelve DMS datasets selected for model benchmarking span diverse activities, including viral spike proteins, RNA-guided nucleases, DNA-binding proteins, RNA-binding proteins, and kinases, maximizing the generalizability of the model architecture, which would serve as the final EVOLVEpro model for experimental applications throughout the rest of the study.

We optimized the EVOLVEpro architecture across five parameters: (i) the strategy for first-round mutant selection, (ii) the top-layer regression model that learns the activity landscape, (iii) the active learning policy for selecting mutants for the next round, (iv) the data processing for experimentally measured activities, and (v) the PLM embedding vector transformation (table S1 and data S1). We first selected ESM-2 as the base PLM because of its large training data and available model size of >200M proteins and 15B parameters, respectively. Using the ESM-2 15B parameter model, our grid search (see methods) found that the optimal strategy was (i) selecting a random set of first-round variants, (ii) using a random forest regressor discriminatory model to predict protein activities, (iii) using embeddings averaged across all amino acids, and (iv) using a top-N selection strategy in each round of evolution (fig. S1A). This model nominated high frequencies of gain-of-function protein variants in only five rounds (fig. S1, A and B), and both median activity and activity of the nominated top mutant increased rapidly from round to round across all DMS datasets (fig. S1B). We calculated the top mutant's improvement in activity by scaling the activity score of the best mutant in the last round relative to the first round across the 12 DMS datasets. In general, 10 rounds of EVOLVEpro evolution with 16 mutants per round identified top mutants with up to 2.2-fold higher activity than the starting WT sequence (fig. S1B). To understand how the number of variants per round affected performance, we simulated EVOLVEpro evolution with 10 to 100 variants per round. We found that larger rounds increased prediction accuracy without saturation (fig. S1C), indicating that EVOLVEpro can be used for both extremely low-N evolution (<20 mutants per round) for rapid and cheap experimental characterization and medium-N (~100 mutants per round) for quicker and more efficient evolution with fewer rounds.

After optimizing the top-layer model and learning strategies, we surveyed a panel of PLMs, Using the optimal parameters from the grid search, we compared ESM-2 15B with smaller versions of ESM-2 and ESM-1 (*44*), UniRep (*29*, *30*), ProtT5 (*45*), ProteinBERT (*10*), Ankh (*9*), one-hot encoding, and integer-encoded protein representations for the ability to identify the highest activity candidates across the 12 datasets. The ESM-2 15B parameter model outperformed other models on all datasets except two and returned the greatest fraction of high-activity mutants, confirming its final selection for the EVOLVEpro latent space model (Fig. 1B and data S3). Across our panel of embedding methods, only four PLMs had a significantly higher prediction accuracy than one-hot encoding, as determined by one-way analysis of variance (ANOVA), showing the critical importance of the base layer model for EVOLVEpro performance. Given the high dimension of ESM2-15B and the small number of samples seen by the top layer regression model, we explored if the full input dimension was

needed for our observed model performance. Reducing the input dimension of the protein embedding using PCA, we tested a range of input dimensions into the top layer regressor. The original full-length embeddings performed best on 9 of the 12 datasets (fig. S1D), with full-length embeddings contributing more to model accuracy in difficult engineering tasks where there exist fewer high-activity mutants in the population, such as MAPK1 kinase and PafA. In contrast, for easier tasks such as infA and AsCas12f, all dimensionalities saturated activity.

As MLDE methods require pre-training a discriminatory model, we compared augmented EVOLVEpro with various amounts of pre-training (Fig. 1C). Active learning drastically reduced the overall number of mutants required: EVOLVEpro with only 5 rounds of evolution (16 mutants per round) was equivalent in performance to EVOLVEpro pre-trained with 160 mutants, whereas 10 rounds of evolution (16 mutants per round) was equivalent to pre-training with 500 mutants. Moreover, EVOLVEpro outperformed zero-shot prediction methods (Fig. 1C) (*15*). This comparison confirms that the few-shot nature of EVOLVEpro allows for efficient directed evolution with minimal effort and low-N testing per round (Fig. 1C and data S4). To explore if EVOLVEpro benefited from base models that are better at extrapolation of complex landscapes, we compared the performance of a random forest regressor with a Gaussian process regressor and a k-nearest neighbor-based regressor, finding that the random forest regressor performed best in 10 out of 12 datasets (fig. S1E). This finding agrees with the utility of random forest regressors in the low-N regime in other prediction tasks (*46*, *47*).

Lastly, we analyzed the per-round activity improvement for EVOLVEpro compared to one-hot and integer encoding and zero-shot prediction. With 16 mutations per round, EVOLVEpro found variants with significantly enhanced activity by round 5 across every dataset (Fig. 1D and fig. S2). Moreover, the one-hot and integer encoding frameworks often saturated much earlier in the evolution process and never reached the activity levels achieved by EVOLVEpro. Interestingly, we observe a non-linear increase in protein activity after round 3 for some proteins, suggesting more substantial gains in mapping the protein activity landscape as EVOLVEpro evolution proceeds.

Antibody optimization with EVOLVEpro

We used EVOLVEpro to optimize two therapeutically relevant monoclonal antibodies: C143, an antibody against the SARS-CoV-2 spike protein, and aCD71, an antibody against the human transferrin receptor used for delivery of drugs and siRNA to muscle and cardiac cells in vivo (*15*, *48*, *49*). aCD71 has more than 90% sequence homology to Delpacibart, a phase II clinical stage therapy for Myotonic dystrophy. Both antibodies have low nanomolar affinities against their cognate antigen, presenting a challenge for further improvement by EVOLVEpro. We designed a multi-objective optimization with EVOLVEpro on antibody expression levels and binding affinity to the target antigen. Optimization over multiple features shows that the model can jointly evolve antibody binding and yield, as antibody mutations often affect multiple antibody features, including expression, stability, solubility, half-life, or immunogenicity. Critically, zero-shot developability optimization is difficult with protein language or structure-based inverse folding models, as the relationship between sequence and developability or other non-binding features is not directly captured by evolutionary sequence or structure data (*15*, *20*). In our multi-objective directed evolution scheme, we weighed the binding affinity at four times the expression levels (i.e., developability score) to prioritize variants that bind with higher affinity (Fig. 2A).

For C143 evolution, we quantified binding affinity with an enzyme-linked immunosorbent assay (ELISA) against the SP6 stabilizing variants of the SARS-CoV-2 spike (Wuhan strain) protein (*50*). We saw improved binding after 3 rounds of EVOLVEpro evolution, surpassing previous zero-shot approaches (*15*) (Fig. 2B and fig. S3A). At round 4, we found significant improvement with a light chain mutant (N28K), with an IC_{50} of 0.19 nM (Fig. 2C). Using our 4 rounds of single mutations, we had EVOLVEpro design multi-mutant combinations for a fifth round. The best multi-mutant (light chain N28R/Q40K with heavy chain R39K) bound to the SP6 spike antigen with an IC_{50} of 60 pM (Fig. 2C), likely due to the synergistic interaction between N28R on the light chain and R39K on the heavy chain. We stopped at one round of multi-mutant evolution due to the substantial improvements we observed, but in practice, multiple rounds of multi-mutant testing are likely needed to reach convergence on desired properties. We found that many improved binders compromised yields (Fig. 2D), a tradeoff due to the bias toward binding during the multi-objective design. Despite this tradeoff, a subset of C143 mutants, such as R39K, had both an increase in affinity and protein expression, showing that developability can be co-optimized alongside binding affinity.

We explored the likelihood of the top EVOLVEpro nominated mutations relative to training data and known antibody variants observed in nature. We analyzed the top 10 mutations for both occurrence in hotspot regions and deviation from the germline sequence. We found none of the top 10 mutations are mutated back to the germline unmutated common ancestor sequence (UCA). The UCA sequence at light chain N28 is a serine (S). However, the affinity-enhancing mutation recommended by EVOLVEpro is either a lysine or arginine, both with likelihoods less than 0.05 when compared to the Uniprot training input, reinforcing the notion that EVOLVEpro mutations are rare and novel. Moreover, this observation highlights the utility of top layer regression model to explore rare mutations not seen in the training input of PLM by promoting exploration into unknown regions of the protein fitness landscape. Furthermore, we found the top single mutant N28R (light chain) happens in the complementarity determining region (CDR), but the majority of the top affinity-

enhancing mutations (7 out of the top 10 mutations) are located in the framework region. This highlights the de novo exploration of EVOLVEpro on the entire variable region of the antibody to find affinity-enhancing mutations that might not seem likely or intuitive in the framework region. To further understand EVOLVEpro's mutational trajectory, we represented the model's attention to particular residues as a cumulative frequency and found residues like K33, R39, and D58 on the heavy chain and S14 and N28 on the light chain are repeatedly explored (fig. S3B).

For using EVOLVEpro to *in silico* evolve an anti-CD71 antibody, we measure the target binding affinity using enzyme-linked immunosorbent assay (ELISA) against the human TfR protein and measured antibody expression with an anti-IgG. We saw improvement in binding after just rounds of evolution with EVOLVEpro (Fig. 2E and fig. S4A). We acquired 10 mutants using the efficient evolution algorithm to benchmark against EVOLVEpro and found that EVOLVEpro nominated mutants 35-fold better than WT whereas efficient evolution's best mutant is only eight-fold better (*15*). At round 4, we found the best single mutant heavy chain S92A to bind to the antigen with an IC_{50} of 29 pM, significantly higher than that of the WT at 551 pM (Fig. 2F). We also asked the model to rank multi-mutants based on the single mutant data from the first four rounds and performed one round of multimutant testing. We improved binding and expression in the multimutant round 5 with heavy chain T70A/S92V mutant. The multimutant binds to the hTfr protein with an IC_{50} of 19 pM (Fig. 2F). Interestingly, most of the mutants nominated after round 1 showed a marked increase in expression profile and binding affinity, showing that EVOLVEpro simultaneously engineered the developability and binding (Fig. 2G). This finding contrasts with the results from the C143 antibody, implying the Pareto frontier of binding and expression likely differs between the two antibodies, with anti-CD71's WT sequence easier to engineer across multiple properties than C143's WT sequence. Future work examining this trade-off between multiple properties for additional antibodies or proteins will enable EVOLVEpro to better traverse Pareto frontiers more efficiently.

Upon analyzing the novelty of the aCD71 antibody mutations, we found only one of the top 10 mutations is mutated back to the germline unmutated common ancestor sequence (UCA) at position V73. The UCA sequence at the site of the best mutation in heavy chain S92 is a threonine (T). However, the affinity-enhancing mutation recommended by EVOLVEpro is either an alanine or valine. S92V mutation has a mutation likelihood of less than 0.05 when compared to the Uniprot training input, highlighting its rarity. This shows EVOLVEpro's ability to insightfully choose novel mutations not seen in the training input of the PLM base layer. Furthermore, we found all top 10 affinity-enhancing mutations are located in the framework region rather than in the CDR that is commonly thought to determine binding affinity. Lastly, to understand EVOLVEpro's mutational trajectory on anti-CD71, we represented the model's attention to particular residues as the cumulative frequency of individual residues being explored by the model and found that multiple residues are repeatedly explored by the model including T70 and S92 on the heavy chain and Q38 on the light chain (fig. S4B).

We used AlphaFold 3 to model the structure of the anti-CD71 and C143 antibodies (Fig. 2, H and I, and data S7). We found two major clusters of exploration by EVOLVEpro on C143 antibody in the framework region with light chain mutations S14, Q40, L50, and K45 co-located and R39, S63, and E89 in close proximity on the heavy chain. These mutations likely alter binding through structural changes in the variable region. Additionally, was a CDR mutation, N28, on the light chain located in the CDR-L1 region that likely directly alters the interaction between the C143 antibody and the antigen, which is not possible to model with AF3 due to a low confidence score of the complex (Fig. 2H). For the anti-CD71 antibody, we found all the best mutations clustered around one region in the heavy chain domain. As they are all in the framework region, they likely alter the binding affinity indirectly, a hypothesis supported by the increase in expression relative to the WT sequence (Fig. 2I).

Lastly, we analyzed each mutant's observed activity versus the PLM-predicted fitness landscape. We calculated the mutant fitness as a predicted marginal masked score within the ESM2 embeddings (pMMS) and found that the activities of EVOLVEpro variants did not correlate with predicted ESM2 fitness (Fig. 2J and fig. S4C). To extrapolate this finding across the entire C143 and anti-CD71 mutational landscape, we projected the base layer PLM fitness score and top-layer random forest predicted fold improvement (pFI) in the latent space for every possible single mutant variant, generating EVOLVEpro determined protein activity landscape (figs. S3C and S4E). There was relatively little overlap between the two distributions, with a negative correlation of -0.16 for C143 antibody and 0.01 for anti-CD71 antibody between predicted fitness and predicted activity, further highlighting ESM2's lack of understanding of protein activity (figs. S3C and S4E). We projected the individual mutants onto the PCA space of the ESM2 embedding and found two opposing directions between higher fitness and higher function (Fig. 2K and fig. S4D). Analyzing the evolution trajectory from round 1 to the final round by calculating the geometric midpoint of each round revealed the top layer model directed the evolutionary process toward the higher side of PCA1 for C143 antibody and the higher side of PCA2 for anti-CD7, which we observe to correlate with higher protein function (figs. S3D and S4F).

Evolution of a miniature RNA-guided CRISPR nuclease with EVOLVEpro

Programmable RNA-guided nucleases have diverse applications in basic biology, therapeutics, and diagnostics. However, commonly used nucleases, such as the Cas9 from *Streptococcus*

pyogenes (SpCas9) are too large to effectively be packaged in common adeno-assocated viral (AAV) vectors, and more compact high-efficiency nucleases, such as the Cas9 from *Staphylococcus aureus* (SaCas9) still preclude the use of larger regulatory elements or protein fusions. Miniature Cas12f nucleases have compact sizes (<700 residues) but suffer from reduced efficiencies, requiring engineering for genome editing applications (*51*). Previous Cas12f engineering efforts relied on DMS or rationally designed mutations to increase the in vitro cleavage activity (*33*, *52*–*55*), requiring extensive screening to find the optimal variant. We tested whether EVOLVEpro could rapidly develop highly active Cas12f variants to accelerate miniature nuclease engineering.

We selected the Cas12f from *Pseudomonas aeruginosa* (PsaCas12f) for evolution with set indel formation at the endogenous *RNF2* locus target site as the optimization metric (Fig. 3A). After four rounds of evolution of 12 single mutants per round, EVOLVEpro yielded point-mutants of PsaCas12f with up to 4.9-fold improvement in indel formation. This top variant, PsaCas12f ^{K333V}, had >40% indel efficiency at the *RNF2* site (Fig. 3B and fig. S5A). To identify synergies between EVOLVEpro nominated mutants, we combined the top-performing variants from previous rounds in a fifth round. We evaluated a set of these multi-mutants and found that PsaCas12f $^{I178A/K333V/K454P}$ generated ~ 50% indel activity at the *RNF2* locus (fig. S5A). Given its performance, we refer to the PsaCas12f $^{I178A/K333V/K454P}$ variant as EVOLVEpro PsaCas12f (epPsaCas12f).

To generalize epPsaCas12f's improved activity, we evaluated the enzyme at 10 different targets across five endogenous genomic loci, comparing to WT PsaCas12f and seven previously characterized Cas12 effectors, AsCas12a, Cas12Φ, UnCas12f1, enAsCas12f, OsCas12f, RhCas12f, and CasMINI (*33*, *53*, *55*–*58*). We observed consistently higher epPsaCas12f activity compared to WT PsaCas12f on 9 of 10 tested targets (Fig. 3C). Moreover, epPsaCas12f edited the 10 targets with a 23.3 ± 16.7% average indel rate, surpassing all tested miniature Cas12f effectors and AsCas12a with 2.2- to 44-fold improvement. Interestingly, epPsaCas12f generated an average deletion of 5-bp across the 10 tested targets (fig. S5B). Together, these data demonstrate that epPsaCas12f is a highly active, compact effector for mammalian genome editing that outperforms other small effectors.

We applied epPsaCas12f for in vivo genome editing applications, using its compact size for single-vector viral delivery in vivo. We designed guides targeting a sequence 5′ of exon 3 in the mouse *PCSK9* gene (Fig. 3D). The PCSK9 protein regulates blood low-density lipoprotein (LDL) by binding to LDL receptors, making it a valuable therapeutic target (*59*). We first tested the efficacy of epPsaCas12f in a murine hepatocyte cell line (Hepa 1-6) by co-transfecting murine codon-optimized epPsaCas12f and sgRNA targeting sequences 5′ of exon 3 in the *PCSK9* gene. Analyses of epPsaCas12f, WT PsaCas12f, and *Staphylococcus pyogenes* Cas9 (SpCas9) revealed that epPsaCas12f robustly edited *PCSK9* in Hepa1-6 cells with ~40% indel formation (comparable levels to SpCas9 and 3-fold higher than the WT PsaCas12f) (Fig. 3D).

After validation of epPsaCas12f in Hepa1-6 cells, we packaged both epPsaCas12f and its sgRNA targeting *PCSK9* in a single AAV2/8 vector (Fig. 3E). AAV-epPsaCas12f was administered at a titer of 1.5×10^{12} viral genome copies per mouse via retro-orbital injection into 3-month-old C57BL/6J mice. We tracked blood PCSK9 levels for 14 days post-injection of AAV and found a significant decrease to around 50% of the original levels after 14 days (Fig. 3F, fig. S5C). We then harvested the liver at day 15, isolated the genomic DNA, and performed next-generation sequencing to survey for indel formation at the *PCSK9* target site (Fig. 5, D and E). We found around 7% on-target indel formation in the AAV-epPsaCas12f injected mice (fig. S5D), demonstrating that epPsaCas12f can be used for single-vector AAV-mediated genome editing. To survey off-targets, we used Cas-OFFinder to predict the top four off-target cleavage sites generated by epPsaCas12f and analyzed the guide-dependent off-target cleavage in the liver (*60*). We only found detectable editing at one of the four sites with a maximum level of 0.27% indels, confirming minimal off-target cleavage triggered by epPsaCas12f (fig. S5F).

To understand the mechanisms of the beneficial mutations nominated by the EVOLVEpro, we used Alphafold3 to predict the structure of PsaCas12f (Fig. 3G and data S7). The predicted structure provides insights into how the PLM-nominated mutations, including I178A/K333V/K454P, contribute to enhancing the DNA cleavage activity (Fig. 3G). The K333V mutation is located in the WED domain, suggesting that it could increase the binding to its RNA guide. The I178A mutation is located in the middle of the long α-helix in the REC domain and forms a hydrophobic core with I245 and L248 in the adjacent α-helix. Given that alanine is a helix-forming residue, the I178A mutation may stabilize the α-helix in the REC domain and thus augment the cleavage activity. The K454P mutation is located at the C terminus of an α-helix in the RuvC domain and forms hydrophobic interactions with A509 and V511 in the adjacent α-helix, suggesting that it also stabilizes the protein conformation.

We then looked at the model's attention to particular residues in the protein by calculating the cumulative frequency of individual residues explored by the model. Multiple residues were repeatedly nominated by the model, including G147 and E451 (Fig. 3H), showing that the model honed attention to specific amino acids. We calculated the pMMS for each nominated mutant to understand the relationship between the base layer PLM's fitness prediction and the actual measured protein activity (Fig. 3I). We found a weak negative correlation between fitness and activity in PsaCas12's local context. We then further projected the base layer PLM's fitness score and the top-layer random forest regressor's activity score in the EMS2 latent space to understand EVOLVEpro's global mutational trajectory (Fig. 3, J to L, and fig. S5G). We found a weak positive correlation of 0.03 between

fitness and activity, further denoting the necessity of a top-layer discrimination model to properly distinguish between high fitness and high activity (Fig. 3I).

Engineering improved prime editors with EVOLVEpro

Many molecular tools, such as next-generation genome editing proteins, function as multiple enzymes acting in concert. Prime editing, which uses an RNA-templated reverse transcriptase to programmably install diverse genome edits, is the fusion of a SpCas9 nicking mutant (nCas9) with an engineered Moloney Murine Leukemia Virus Reverse Transcriptase (M-MLV RT) [D200N, L603W, T306K, W313F, T330P] (termed PE2). We reasoned that EVOLVEpro could improve upon these rational mutations, as optimizations were discovered on M-MLV RT by directed evolution approaches (*61*). As PE-based insertion has difficulty installing longer (>40 nt) edits, we focused on editing outcomes with longer (46 bp) insertions, which have particular utility for programmable gene insertion methods, such as PASTE (*62*). We set up the evolution policy with a previously described twinPE approach, where two overlapping pegRNAs are used in combination to install a 46bp attB site in the NOLC1 loci in murine hepatocyte cell line (Hepa1-6). Editing was quantified at *NOLC1* loci using amplicon sequencing and NGS readout and the top-layer EVOLVEpro model was trained to predict the insertion efficiency.

Over successive rounds of optimization, we found that EVOLVEpro progressively learned the activity landscape of the RT of PE2, yielding improved variants after the initial random selection round and substantially improving upon PE2-based editing by round 4 (Fig. 4A and fig. S6A). To check for bias toward this single locus in the genome that could have developed during the evolution campaign, we tested the top 4 performing variants (A660S, L670C, L670K, and L671R) at three additional genomic loci (human *AAVS1*, human *ALB*, and mouse Factor IX) in two cell lines. At all four sites tested, A660S had statistically significant improvements (Fig. 4B). These results point to the general protein activity improvement by EVOLVEpro, delivering an additional set of RT mutations specifically for larger edits.

Projecting the top mutations onto the AlphaFold3-predicted structure of the RT reveals that most of them are clustered in the C-terminal RNA polymerase H (RNase H) domain (Fig. 4C), which is a surprising result since most PE evolution focuses on RT mutagenesis. We speculate that these mutations could alter the cleavage of the template DNA in the RNA-DNA heteroduplex by the RNaseH domain (*63*), facilitating the completion of the prime editing reaction, a route that has not been explored by traditional engineering of prime editors. Alternatively, they could inactive the RNase H domain, as truncations of this domain have been shown to slightly improve prime editing activity (*64–66*). We then further analyzed EVOLVEpro's residue site preference during evolution and observed attention to residues like L670, L671, and A660, suggesting it was learning that these positions could be quite beneficial for improving activity (Fig. 4D). Analysis of predicted fitness (pMMS) scored by the bottom layer PLM again showed a divergence between fitness and activity for the prime editor (Fig. 4E and fig. S6B), as we found almost no convergence between the two distributions with a negative correlation of 0.08 (Fig. 4E).

Lastly, we try to understand the global mutational trajectory by projecting the activity landscape learned by the random forest regressor and base layer ESM2's protein fitness landscape onto the first two PCAs of the embedding (fig. S6, C and D). This analysis points again to the divergence between the mutational landscape of a protein's activity and the commonly used fitness landscape learned during a foundational model's training on all protein sequences. When we charted the evolution trajectory, we also found a clear movement to the lower side of PCA2 protein embedding dimension, showing that the top layer is pushing the sequences to that region in the latent space (fig. S6E).

Bxb1 integrase evolution with EVOLVEpro

Large serine recombinases (LSRs) are enzymes that facilitate precise DNA rearrangements, making them crucial tools for genome editing. Their ability to recognize specific DNA sequences and catalyze targeted recombination events allows for efficient and accurate modifications of genetic material, which is essential for advanced gene therapy, synthetic biology, and genetic research. We recently developed a gene insertion technology, PASTE, that leverages LSRs, specifically the Bxb1 integrase, for programmable gene insertion in eukaryotic cells (*62*). A limitation of Bxb1 integrase, however, is its activity saturates in the 20 to 60% range in cells, limiting the overall integration efficiency that can be achieved. We sought to therefore evolve Bxb1 using EVOLVEpro to improve its activity and demonstrate improved gene integration applications with PASTE in cells.

To evolve Bxb1, we designed a simple integration assay in HEK293FT cells that involved the insertion of an AttP-containing DNA plasmid into an AttB target-containing plasmid (Fig. 4F). Integration can be measured by next-generation sequencing, and the evolution policy is designed to optimize this insertion efficiency. We started evolution with a round of 11 random Bxb1 point mutation variants and then over 9 rounds observed progressively increasing activity resulting in mutants with over 2.6-fold higher activity than WT (Fig. 4G and fig. S7A). As Bxb1 is already fairly active, this fold improvement is expected as we reach near-saturating levels of insertion. To validate the top hits from the evolution campaign, we performed a Bxb1 plasmid titration experiment in a separate cell line (Hela cells) and observed up to fourfold improvement in recombination efficiency under low Bxb1 expression (Fig. 4H). We further validated the top hits by pre-installing attB sites into the genome of HEK293FT cells using lentivirus and then surveyed for integration efficiency of cargo in the genome. We found up to fourfold improvement by

EVOLVEpro's mutants compared to WT (fig. S7B). We termed T166R variant as EVOLVEpro enhanced Bxb1 (epBxb1).

To test whether the epBxb1 variant's improved activity can improve the programmable insertion of cargo DNA into the chromosome, we tested this variant in the context of PASTE and compared it against the WT Bxb1 across five different genomic loci. We found up to ~fourfold improvement in the final large cargo insertion rate into the genome, which highlights the generalizable gain in activity (Fig. 4I and fig. S7C).

An AlphaFold3-predicted model of Bxb1 bound to attachment site DNA indicates that the top beneficial EVOLVEpro mutations clustered in the Bxb1 DNA binding domains, likely increasing the affinity to its DNA targets (fig. S8A). Of these residues, V292S could directly interact with the phosphate backbone of the target DNA based on its positioning relative to the attachment site, whereas the others likely modulate DNA binding via indirect interactions. Analysis of the residue exploration by the model revealed that multiple positions, including F439, V375, and L275, are visited up to 15 times; the DNA-interacting residue V292 was also visited multiple times. Overall, this highlights EVOLVEpro's ability to recognize the functional importance of certain regions in the protein, much like structure-guided engineering approaches (fig. S8B).

We then calculated the relationship between the fitness (pMMS) and activity (observed fold improvement) for Bxb1 integrase and found a weakly positive correlation between the two metrics contrary to the other proteins reported evolves. This likely reflects a subset of protein families where protein stability and fitness as learned by the PLM can predict activity as previously reported (*67*) (fig. S8C). However, given that the relationship is weak, a model like EVOLVEpro is still needed to efficiently and quickly reach high-performing variants without encountering many false positives. Lastly, we found that the global mutation landscape learned by EVOLVEpro was still divergent from the predicted fitness (pMMS) by ESM2 with an even weaker correlation, further highlighting the ability of EVOLVEpro to learn protein activity at a global scale and how stability/fitness prediction is not sufficient for rapid and efficient protein evolution (fig. S8, D to G).

Evolving T7 RNA polymerase for efficient and highly pure RNA production

Multi-objective optimization with EVOLVEpro allows for the evolution of multiple complex activities. We chose to showcase multi-objective optimization on a common and broadly used enzyme with applications across basic biology and therapeutics. We selected the T7 RNA polymerase (RNAP) due to its critical role in RNA production for mRNA therapies, mRNA vaccines, cell engineering, and basic scientific studies. As mRNA production has numerous features characterizing its potency and quality, as opposed to genome editing where one feature matters the most, we designed a multi-objective optimization function to evolve a high-fidelity T7 RNAP for mRNA production with these three parameters: (i) RNA yield measured via UV-vis spectrophotometry, (ii) mRNA translation in a double-stranded RNA (dsRNA) sensitive cell line measured via luciferase translation, and (iii) RNA purity measured via immunogenicity in BJ fibroblast cells by interferon β (IFN-β) RNA production (Fig. 5A). We weighted these features in the EVOLVEpro objective function by 20, 40, and 40%, respectively to prioritize the higher fidelity and lower immunogenicity aspects of this enzyme for clinical applications. To facilitate high throughput variant testing, we relied on SP6 in vitro transcription-translation coupled reaction kits to generate mutant T7 RNAP in a one-pot reaction and subsequently use the produced T7 RNAP to produce co-transcriptionally capped Cypridina luciferase mRNA for downstream in vitro testing.

During the initial two rounds of evolution, fold improvement of top mutants was two- to fourfold. By rounds 3 and 4, we started observing substantial improvements in all features, especially in translation and immunogenicity fold changes over the WT T7 RNAP (Fig. 5B-C, fig. S9A). By the end of round 4, one T7 RNAP mutant, E643G, generated luciferase mRNA that produced 34x more translated luciferase and ~98% less immunogenicity (Fig. 5C). We sought to benchmark E643G against the previously engineered state-of-the-art mutant T7 RNAP with G47A mutation and 884G insertion(G47A/884insG) that has markedly reduced immunogenic byproduct in our in vitro transcription/translation (IVTT) assay (*68*). We found that our E643G mutant produces sevenfold higher translation in cells and approximately twofold less IFNB1 inflammation in BJ fibroblasts (fig. S9B).

To leverage the suite of mutants generated in the first four rounds, we generated multi-mutants with EVOLVEpro, combining up to three previously tested mutations. We also included combinations with the previously identified G47A mutation known to reduce dsRNA formation. In typical rational mutagenesis, single beneficial mutations are combined according to their spatial location under the assumption of synergistic effects of these mutations. Here, we relied on EVOLVEpro's ability to learn the activity landscape to nominate multi-mutants. After two rounds of multi-mutants corresponding to the sixth round of engineering, EVOLVEpro nominated variants with up to ~57x more translation from luciferase mRNA and ~515x less immunogenicity than the original WT T7 RNAP (Fig. 5C). The top variant, T7 RNAP$^{T3M/G47A/E643G}$, was substantially more effective at translation and less immunogenic than the G47A/884insG mutant. This multi-mutant was chosen as the final EVOLVEpro evolution candidate and termed EVOLVEpro enhanced T7 RNAP, or epT7. The wide range of activities we see upon combining mutations points to complex epistatic interaction on the fitness landscape, and further examination of these interactions is needed to understand the mechanism underlying the range of activities seen for different combination mutants.

Given the high throughput testing of mutant T7 RNAPs in the

IVTT reaction, we hypothesized that the unoptimized IVT buffer could change these mutant's mRNA production (*69*) and sought to compare the performance of top mutants in clinically relevant IVT settings with NEB's HiScribe transcription kit, followed by Vaccinia cap-1 capping and polyA tailing. We therefore purified the top performing single mutant (E643G), previously reported state-of-the-art mutant (G47A/884insG) (*68*), and our epT7 (T3M/G47A/E643G) along with WT to compare their performance. We compared the production of six different mRNA sequences, ranging in size from 500 nt to 6500 nt, between epT7, $T7^{E643G}$, and WT T7. Consistent with our IVTT-based experiment, we found that epT7 and E643G produced significantly higher mRNA in a 2-hour transcription scheme than both WT T7 RNAP and the G47A/884insG variant (fig. S9C). Analysis of the three different mRNA products by both E-gel EX and TapeStation gel electrophoresis systems confirmed the presence of a single on-target product across all four enzymes (fig. S9, D to F). Looking at the translation and immunogenicity aspects of the mRNAs produced by these enzymes, we found that in all cases, epT7 produced mRNA had four- to 120-fold higher translation than wild type and four- to 256-fold lower immunogenicity (Fig. 5D and fig. S10A). Functional testing of SpCas9 mRNA also shows significantly higher editing from epT7's produced mRNA in two separate cell lines (fig. S10B). These results validate that the EVOLVEpro derived epT7 mutants are not buffer or template-specific and are genuinely improving the quality of mRNA produced by the polymerase. We next investigated the mechanism of the epT7 performance enhancements by investigating the quality of the RNA. Using an established ELISA for dsRNA, we found that the dsRNA in the epT7-produced mRNA was fivefold lower than WT T7-produced RNA and it performed equally well as the RNA produced by the state-of-the-art G47A/884insG mutant (*68*) (Fig. 5E).

Previous efforts to reduce dsRNA production relied on adding a glycine residue at the C-terminal "foot" region of the enzyme (884G insertion) (*68*). Our model revealed the functional importance of E643 in transcription and, surprisingly, mutating this residue rendered the same effect as 884G insertion (Fig. 5F and fig. S9B). Indeed, analysis of the T7 RNAP structure reveals that E643 is close to the DNA template, suggesting that E643G improves template binding and RNA production (Fig. 5F). However, E643K/E643R did not improve the fidelity of transcription (fig. S9A), suggesting that these bulky residues sterically clash with the template DNA.

To rationalize how EVOLVEpro is exploring the activity landscape it is useful to consider the progression of nominated residues through the first four rounds. E643 was found first in round 3 with the most beneficial mutation being E643N (Fig. 5G). The model quickly zoomed into this region by exploring it 5 more times in round 4, yielding E643G the best single mutant. G47A has been previously reported to increase helix formation, and EVOLVEpro took advantage of this helix-favoring mutation in our multi-mutant generation. The third mutated residue in epT7 is in a disordered region (T3M), suggesting a role independent of DNA template binding. T3M might be involved in improving protein stability or other aspects that can modulate the polymerase activities. These results suggest that EVOLVEpro can be used to identify and interrogate the effect of various mechanisms and determine the right balance biochemically to mutagenize.

We next calculated the relationship between the activity (observed data) and fitness (pMMS) for T7 RNAP and found a negative correlation of 0.13, in this case denoting the lack of association between the two metrics. EVOLVEpro successfully navigated through this divergence by selecting mutants with higher activity but not fitness in later rounds (Fig. 5H). Lastly, we investigated the global evolutionary landscape of epT7 and EVOLVEpro's mutational trajectory. At a high level, as with the previous proteins evolved, the activity map learned by EVOLVEpro diverged from the fitness map predicted by ESM-2, showing that fitness predictions would not be able to predict the mutants that were ultimately discovered to improve protein activity and other parameters (Fig. 5, I to K, and fig. S10C).

Circular RNA production with epT7

Circular RNA has emerged as a promising therapeutic modality for protein replacement therapy thanks to its enhanced stability and prolonged expression of proteins (*70*). Since we observed significantly lower dsRNA production and higher fidelity of transcription with epT7, we hypothesized that epT7 would enhance circular RNA production since the use of RNase R during post-IVT processing typically enriches for both circular RNA and dsRNA species that are immunogenic (Fig. 6A). We thus applied epT7 to the circularization of four different RNA sequences, finding that the translation obtained by circRNA from epT7 is 3 to 30 fold higher than RNA produced by WT T7 RNAP (Fig. 6B and fig. S11, A to D and J). We then used TapeStation gel electrophoresis to quantify the relative ratio of circular products post IVT and found reduced long concatemer formation in the circular RNA produced by epT7 (Fig. 6C). To better understand the mechanism behind better translation of circular RNA made by epT7, we performed gel electrophoresis using 2% E-gel EX as previously validated to check for the relative ratio of precursor, nicked, intermediate and full circular RNA both pre– and post–RNase R treatment (Fig. 6D). We noticed reduced intermediate and nicked byproducts in circular RNA produced by epT7, showing higher fidelity of transcription. We used the gel electrophoresis results to quantify the ratio of circular RNA across three different templates and found significantly higher circular RNA production at around 25% efficiency, which was ~2 fold higher than the efficiency of WT T7 RNAP, higher circRNA purity, and lower concatemer production (Fig. 6E and fig. S11, G to I). Lastly, we used dsRNA ELISA to detect the amount of dsRNA left in the product after RNAse R cleanup. Consistent with our hypothesis, there is a large increase in dsRNA

percentage at around 1.5% from WT T7's produced dsRNA (Fig. 6F). This dsRNA ratio is significantly reduced to 0.2% using epT7, highlighting the fidelity of this variant during long transcription that is needed to accommodate circular RNA production (Fig. 6F). To confirm the higher stability of circular-eGFP RNA, we transfected both WT T7 and epT7's produced circRNA in HEK293FT cells and imaged them 24 hours and 72 hours post-transfection (fig. S11, E to F). We observed higher GFP fluorescence from epT7 than WT T7 RNAP and stable expression of GFP at 72 hours similar to previously reported (*70*).

mRNA for in vivo bioluminescent imaging

Given the high fidelity of epT7, we compared the performance of epT7 with WT T7 RNAP in producing 100% N^1-methylpseudouridine-5′-triphosphate-modified firefly luciferase mRNA that is commonly used for in vivo deep tissue imaging (Fig. 6G). This production process, including the modified bases, mimics the clinical production of therapeutic mRNAs, allowing for a translationally relevant evaluation of epT7. We packaged the produced mRNA with lipid nanoparticles (LNPs) that traffic to the liver for bioluminescent imaging. At 24 hours post-injection of the mRNA-loaded LNPs, we observed ~10-fold higher luminescence for our epT7-produced mRNA compared to mRNA produced by WT T7 RNAP (Fig. 6H). Moreover, we tracked the expression kinetics of both mRNAs for 96 hours and found consistently higher translation with the epT7-produced Fluc mRNA for a longer period of time (Fig. 6I and fig. S11K).

Originality of mutations explored during EVOLVEpro evolution

We analyzed the mutations proposed by EVOLVEpro on the six proteins evolved in this study by calculating the mutational likelihood of individual mutation compared to the training input (Uniprot). We found that the median mutational likelihood for each protein's set of mutations ranges from 0.01 to 0.04 which is well below the 0.05 cutoff for rare mutations (Fig. 6J) (*15*). Most of the mutations explored by the model are uncommon mutations not seen in nature as defined by a probability cutoff of less than 0.1, with 92% of PE2 MLV RT mutations and 77% of Bxb1 integrase mutations below this threshold. Moreover, all the best activity-enhancing single mutants explored during the evolution of the six proteins in this study have a mutational likelihood of less than 0.1. This analysis reveals that the mutation landscape explored by EVOLVEpro is highly original compared to zero-shot-based language models and reinforces the need to search outside naturally occurring mutations to find activity-enhancing mutants (fig. S12).

Discussion

We demonstrate EVOLVEpro as a model for *in silico* directed evolution of protein activities using few-shot active learning. Over consecutive rounds of improvement, EVOLVEpro yields variants with two- to 515-fold improvements in desired properties, including binding, catalytic efficiency, and immunogenic byproducts. Using both evolutionary scale PLMs and a regression layer, EVOLVEpro learns general rules of protein activity, generating highly active mutants with only a few cycles of evolution. Moreover, because of the rich latent space generated by the PLM and powerful feature selections present in the top-layer module, EVOLVEpro evolution is a low-N learning approach that requires minimal wet lab experimentation. We benchmark EVOLVEpro across 12 different DMS datasets covering 8 protein classes, showing its superiority in the low-N evolution setting. In this benchmarking work, we evaluate all currently available embedding-based PLMs and perform a grid search to optimize over top-layer regression models, active learning selection strategies, and different normalization techniques toward the embeddings and activity measurements. We find that PLMs are essential and their representations of protein sequence outperform traditional encoding methods like one-hot encoding and integer encoding (Fig. 1B). Interestingly, even in the extreme scarcity of data relative to the size of the input vector, dimensionality reduction of the embedding space through PCA did not improve performance, reinforcing the importance of the PLM dimensions in guided *in silico* directed evolution (see methods and data S1). The modular design of EVOLVEpro allows for the integration of future improvements in autoregressive PLMs or next-generation representation models.

The success of EVOLVEpro speaks to the inherent limitations of PLMs, which are trained to learn a masked sequence reconstruction task across evolutionary diversity. As natural sequences do not necessarily select for optimal protein activity, the PLM's learned activity landscape will often not be correlated with a protein's activity landscape (Fig. 6K). In scenarios of correlations between fitness and activity, such as antibodies, zero-shot PLM protein evolution may work with some success (*15*, *20*), but enzyme optimization has proven more challenging. It has been shown that PLMs can scale with increasing parameters just like large language models, but recent analyses have shown saturating scaling effects of PLMs with limited input training datasets (Uniref) on larger models (*71*–*74*). Thus, it is likely that simply increasing the parameters of these PLMs will not enable better prediction of protein activities and other downstream tasks. Alternatively, generative PLMs have yielded functional de novo proteins, such as GFP and CRISPR nucleases (*16*, *19*). These models explore a much larger search space than EVOLVEpro in the initial design phase, but variant designs generated by these methods do not have improved activities relative to WT proteins yet, and the functional success rate of generated proteins is very low. As such, Rufollo *et al*. successfully designed OpenCRISPR, an AI-generated Cas9 protein that has cleavage efficiency comparable with the WT SpCas9 (*19*). We expect generative PLMs to

design fairly active protein sequences in tasks where there doesn't exist a good starting point (i.e., a binder against a previously unknown target) (*75*). These de novo designed sequences may be suitable for combination with EVOLVEpro to create an end-to-end de novo design and evolution framework where de novo generated sequences can be rapidly optimized for state-of-the-art activity and thus real-world deployment. In addition, biophysical based models can also be integrated with the regressive top layer approach established here to further boost prediction accuracy and enable rapid identification of gain-of-function mutants *in silico* (*76*).

Using EVOLVEpro, we present the first comprehensive evaluation of an AI directed evolution model across six therapeutically relevant proteins. These proteins demonstrate a low correlation between observed activity and the PLM-estimated fitness, requiring EVOLVEpro to rapidly navigate the unseen activity landscape. In some cases, we leverage EVOVLEpro for multi-objective feature optimization, allowing evolution of multiple properties simultaneously. Critically, the assays used for measuring protein activity in this work are incompatible with pooled screening approaches, precluding typical directed evolution strategies. Across the multi-mutant landscape of protein activity, EVOLVEpro is able to select highly active single mutants out of more than 16,000 possible sequences and multi-mutants from more than 780 billion possible sequences. We thoroughly validate the six proteins evolved by EVOLVEpro for genome editing, binding, and RNA generation tasks beyond the training set, finding state-of-the-art performance. Structural analysis of top mutations reveals many distinct mechanisms of activity improvement, suggesting future directions for directed evolution of these enzymes. In the context of protein design, EVOLVEpro is a highly capable protein engineering model in that it (i) has high rates of success, (ii) requires no special knowledge about the protein, (iii) can be used for multi-objective function or property optimization, and (iv) is highly modular, allowing for any protein property with a quantifiable assay to be used as an input without extensive finetuning. We anticipate EVOLVEpro will continue to improve with new foundation models and enhanced search strategies and will be broadly useful for protein engineering.

Materials and Methods

Use of ESM2 embeddings

Let $x_i = [a_1, a_2, ..., a_n]$ denote the amino acid sequence of the i-th protein variant, where each a_j represents an individual amino acid and n is the length of the protein. The protein language model embedding transformation (ESM2-15B) maps x_i to a sequence of embeddings, one for each amino acid, where d is the dimensionality of the embedding space (hidden dimension):

$$E_i = \mathrm{PLM}(x_i) \in R^{n \times d} \quad (1)$$

This results in a per-token representation of size $n \times d$. We use the final representation layer of the ESM2-15B model. To reduce the number of features in the low-N setting and obtain a fixed-size representation regardless of protein length, we compute the average embedding vector by taking the mean across all amino acid positions:

$$\overline{e_i} = \frac{1}{n}\sum_{j=1}^{n} E_{ij} \quad (2)$$

This results in a single d-dimensional vector $\overline{e_i} \in R^d$ representing the entire protein variant, which is then used as input for EVOLVEpro.

EVOLVEpro Model

EVOLVEpro utilizes a Random Forest regressor as its top-layer model to learn the functional grammar of variants with respect to their activity. This model operates on information-rich latent space mean embeddings $\overline{e_i}$ generated by a protein language model as described previously, in an active learning setting. The Random Forest regressor uses these embeddings $\overline{e_i}$ as input features to predict the activity or fitness of each variant.

A Random Forest Regressor was employed as the top-layer model in the EVOLVEpro framework. This ensemble learning method combines multiple decision trees to make predictions, offering robustness against overfitting and the ability to capture complex, non-linear relationships in the data.

The Random Forest model was configured with 100 estimators (individual decision trees). The quality of splits was evaluated using the Friedman Mean Squared Error (MSE) criterion.

Let $D = \left\{\left(\overline{e_i}, y_i\right)\right\}_{i=1}^{N}$ be the training dataset, where $\overline{e_i} \in R^d$ are the input features (reduced embeddings) and $y_i \in R$ are the target values (protein fitness).

Decision Trees:

Each tree $h_t\left(\overline{e_i}\right)$ in the forest is trained on a bootstrap sample of the original dataset. At each node, the best split is determined by maximizing the reduction in impurity:

$$\Delta I = I(\mathrm{parent}) - \frac{N_{\mathrm{left}}}{N} I(\mathrm{left}) - \frac{N_{\mathrm{right}}}{N} I(\mathrm{right}) \quad (3)$$

where I is the Friedman MSE impurity measure, calculated as follows:

$$\mathrm{diff} = \bar{y}_{\mathrm{left}} - \bar{y}_{\mathrm{right}} \quad I = \frac{N_{\mathrm{left}} \cdot N_{\mathrm{right}}}{N_{\mathrm{left}} + N_{\mathrm{right}}} \cdot \mathrm{diff}^2 \quad (4)$$

Here, $\bar{y}_{\mathrm{left}}$ and $\bar{y}_{\mathrm{right}}$ are the mean target values in the left and right child nodes, respectively, and N_{left} and N_{right} are the number of samples in each child node.

Random Forest Prediction:

The final prediction of the Random Forest for a new input $\overline{e_i}$ is the average of the predictions from all trees:

$$f\left(\overline{e_i}\right) = \frac{1}{T}\sum_{t=1}^{T} h_t\left(\overline{e_i}\right) \quad (5)$$

where $T = 100$ is the number of trees in the forest.

Active Learning Approach:

For each round of EVOLVEpro, after training the Random Forest on the current dataset, we apply the model to predict fitness values for the remaining, untested protein variants. The active learning step then selects the most promising variants for experimental testing in the next round.

Given the nature of Random Forest regression, all predicted values for untested variants will fall within the range of y-values in the training set:

$$\min\left(y_{train}\right) \le \hat{y}_{predict} \le \max\left(y_{train}\right) \quad (6)$$

where y_{train} are the fitness values in the training set and $\hat{y}_{predict}$ are the predicted fitness values for untested variants.

While we explored various selection strategies, including choosing embeddings most distant from the training set in Euclidean space and selecting a mix of top and bottom predictions, we hypothesize that for the objective of round-over-round optimization of top fitness, the top-n strategy is most effective.

The top-n strategy involves selecting the n variants with the highest predicted fitness values:

$$\text{Selected Variants} = \left\{\overline{e_i} : \hat{y}_i \ge y_{(n)}\right\} \quad (7)$$

where $y_{(n)}$ is the nth highest predicted fitness value.

This strategy aims to iteratively stretch the upper limit of the Random Forest's prediction range:

$$\max\left(y_{train}\right)_{round\ k} \le \max\left(y_{train}\right)_{round\ k} \quad (8)$$

By focusing on the top predictions, we exploit the model's current understanding of high-fitness regions in the embedding space, while also encouraging exploration of nearby areas that may yield even better variants: While this greedy approach might risk overlooking some areas of the fitness landscape, we believe it aligns well with the goal of rapidly identifying and optimizing top-performing protein variants in a limited number of experimental rounds.

Benchmarking on 12 DMS datasets

For model benchmarking, we took 9 existing deep mutational scanning (DMS) datasets which were employed in a previous zero-shot high fitness prediction approach (*15*). From this work, we leveraged a pre-determined cutoff for high-fitness variants for each dataset to select variants that were low and high fitness. To augment the use of these datasets, we also selected three additional DMS datasets: an AsCas12f compact genome editor (*33*), Cov2 viral spike receptor-binding domain (*43*), and Zika virus envelope protein (*42*). For these datasets, cutoff values were set based on the general distribution of high-activity variants. To facilitate downstream work, a tabular format CSV file and a fasta file of all available mutant sequences with activity measurements were generated from each dataset.

EVOLVEpro Parameter Grid search

We conducted an extensive grid search to evaluate various strategies for optimizing fitness in a low number of rounds. The grid search explored the following parameters:

1. Fitness measurement: Raw fitness values from each dataset or min-max normalized fitness.

2. First-round strategy: Random selection of variants or diverse selection using K-medoids clustering on protein language model embeddings.

3. Learning strategies: We compared several strategies for selecting variants in subsequent rounds, including:

a. Random selection

b. Top n predicted fitness variants

c. Top n/2 and bottom n/2 predicted fitness variants

d. Maximizing Euclidean embedding distance from previously selected variants

4. Embedding types: We compared different embedding representations, including raw embeddings, and PCA-reduced embeddings (from 10 to 1000) to account for the fact that this was a high p, low n paradigm. This was entirely done on the largest (15B parameter) ESM2 model.

5. Regression types: We evaluated various regression models for fitness prediction, including ridge regression, lasso regression, elastic net, linear regression, neural networks with a linear last layer, random forest regression, KNN regression, Gaussian processes and gradient boosting regression. These were largely used with default parameters.

For each combination of parameters, we ran three simulations (to vary the first round of selected variants) using 16 variants per round to account for stochastic variability. Performance was assessed using the proportion of high-fitness variants out of the top 16 variants that the updated model would predict. We quantified the overall effectiveness of each parameter value by counting the number of datasets for which it achieved the highest mean fitness binary percentage. This "winning strategy" count provided a simple yet informative summary of which approaches were most successful across diverse protein systems. The "winning strategy" was random first round, raw fitness, top-n selection, random forest regression, and raw embeddings.

REFERENCES AND NOTES

1. D. M. Fowler, S. Fields, Deep mutational scanning: A new style of protein science. *Nat. Methods* **11**, 801–807 (2014). doi:10.1038/nmeth.3027 Medline
2. S. Gelman, S. A. Fahlberg, P. Heinzelman, P. A. Romero, A. Gitter, Neural networks to learn protein sequence-function relationships from deep mutational scanning data. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2104878118 (2021). doi:10.1073/pnas.2104878118 Medline
3. A. Judge, B. Sankaran, L. Hu, M. Palaniappan, A. Birgy, B. V. V. Prasad, T. Palzkill, Network of epistatic interactions in an enzyme active site revealed by large-scale deep mutational scanning. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2313513121 (2024). doi:10.1073/pnas.2313513121 Medline
4. N.-L. Sim, P. Kumar, J. Hu, S. Henikoff, G. Schneider, P. C. Ng, SIFT web server:

Predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* **40**, W452–W457 (2012). doi:10.1093/nar/gks539 Medline
5. T. A. Hopf, J. B. Ingraham, F. J. Poelwijk, C. P. I. Schärfe, M. Springer, C. Sander, D. S. Marks, Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35**, 128–135 (2017). doi:10.1038/nbt.3769 Medline
6. J. Frazer, P. Notin, M. Dias, A. Gomez, J. K. Min, K. Brock, Y. Gal, D. S. Marks, Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91–95 (2021). doi:10.1038/s41586-021-04043-8 Medline
7. Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. Dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, A. Rives, Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023). doi:10.1126/science.ade2574 Medline
8. M. Heinzinger, K. Weissenow, J. G. Sanchez, A. Henkel, M. Mirdita, M. Steinegger, B. Rost, Bilingual Language Model for Protein Sequence and Structure, *bioRxiv* (2024)p. 2023.07.23.550085.
9. A. Elnaggar, H. Essam, W. Salah-Eldin, W. Moustafa, M. Elkerdawy, C. Rochereau, B. Rost, Ankh: Optimized Protein Language Model Unlocks General-Purpose Modelling, arXiv:2301.06568 [cs.LG] (2023).
10. N. Brandes, D. Ofer, Y. Peleg, N. Rappoport, M. Linial, ProteinBERT: A universal deep-learning model of protein sequence and function. *Bioinformatics* **38**, 2102–2110 (2022). doi:10.1093/bioinformatics/btac020 Medline
11. T. Bepler, B. Berger, Learning the protein language: Evolution, structure, and function. *Cell Syst.* **12**, 654–669.e3 (2021). doi:10.1016/j.cels.2021.05.017 Medline
12. T. Bepler, B. Berger, Learning protein sequence embeddings using information from structure, arXiv:1902.08661 [cs.LG] (2019).
13. T. Yu, H. Cui, J. C. Li, Y. Luo, G. Jiang, H. Zhao, Enzyme function prediction using contrastive learning. *Science* **379**, 1358–1363 (2023). doi:10.1126/science.adf2465 Medline
14. Y. He, X. Zhou, C. Chang, G. Chen, W. Liu, G. Li, X. Fan, M. Sun, C. Miao, Q. Huang, Y. Ma, F. Yuan, X. Chang, Protein language models-assisted optimization of a uracil-N-glycosylase variant enables programmable T-to-G and T-to-C base editing. *Mol. Cell* **84**, 1257–1270.e6 (2024). doi:10.1016/j.molcel.2024.01.021 Medline
15. B. L. Hie, V. R. Shanker, D. Xu, T. U. J. Bruun, P. A. Weidenbacher, S. Tang, W. Wu, J. E. Pak, P. S. Kim, Efficient evolution of human antibodies from general protein language models. *Nat. Biotechnol.* **42**, 275–283 (2024). doi:10.1038/s41587-023-01763-2 Medline
16. T. Hayes, R. Rao, H. Akin, N. J. Sofroniew, D. Oktay, Z. Lin, R. Verkuil, V. Q. Tran, J. Deaton, M. Wiggert, R. Badkundri, I. Shafkat, J. Gong, A. Derry, R. S. Molina, N. Thomas, Y. A. Khan, C. Mishra, C. Kim, L. J. Bartie, M. Nemeth, P. D. Hsu, T. Sercu, S. Candido, A. Rives, Simulating 500 million years of evolution with a language model, *bioRxiv* (2024)p. 2024.07.01.600583.
17. N. Ferruz, S. Schmidt, B. Höcker, ProtGPT2 is a deep unsupervised language model for protein design. *Nat. Commun.* **13**, 4348 (2022). doi:10.1038/s41467-022-32007-7 Medline
18. A. Madani, B. Krause, E. R. Greene, S. Subramanian, B. P. Mohr, J. M. Holton, J. L. Olmos Jr., C. Xiong, Z. Z. Sun, R. Socher, J. S. Fraser, N. Naik, Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.* **41**, 1099–1106 (2023). doi:10.1038/s41587-022-01618-2 Medline
19. J. A. Ruffolo, S. Nayfach, J. Gallagher, A. Bhatnagar, J. Beazer, R. Hussain, J. Russ, J. Yip, E. Hill, M. Pacesa, A. J. Meeske, P. Cameron, A. Madani, Design of highly functional genome editors by modeling the universe of CRISPR-Cas sequences. bioRxiv 2024.04.22.590591 [Preprint] (2024); https://doi.org/10.1101/2024.04.22.590591.
20. V. R. Shanker, T. U. J. Bruun, B. L. Hie, P. S. Kim, Unsupervised evolution of protein and antibody complexes with a structure-informed language model. *Science* **385**, 46–53 (2024). doi:10.1126/science.adk8946 Medline
21. F. Ding, J. Steinhardt, Protein language models are biased by unequal sequence sampling across the tree of life. bioRxiv 2024.03.07.584001 [Preprint] (2024); https://doi.org/10.1101/2024.03.07.584001.
22. F. H. Arnold, Design by directed evolution. *Acc. Chem. Res.* **31**, 125–131 (1998). doi:10.1021/ar960017f
23. K. K. Yang, Z. Wu, F. H. Arnold, Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **16**, 687–694 (2019). doi:10.1038/s41592-019-0496-6 Medline
24. N. Thomas, D. Belanger, C. Xu, H. Lee, K. Hirano, K. Iwai, V. Polic, K. D. Nyberg, K. G. Hoff, L. Frenz, C. A. Emrich, J. W. Kim, M. Chavarha, A. Ramanan, J. J. Agresti, L. J. Colwell, Engineering of highly active and diverse nuclease enzymes by combining machine learning and ultra-high-throughput screening, *bioRxiv* (2024)p. 2024.03.21.585615.
25. P. A. Romero, F. H. Arnold, Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.* **10**, 866–876 (2009). doi:10.1038/nrm2805 Medline
26. R. Fox, A. Roy, S. Govindarajan, J. Minshull, C. Gustafsson, J. T. Jones, R. Emig, Optimizing the search algorithm for protein engineering by directed evolution. *Protein Eng.* **16**, 589–597 (2003). doi:10.1093/protein/gzg077 Medline
27. Z. Wu, S. B. J. Kan, R. D. Lewis, B. J. Wittmann, F. H. Arnold, Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 8852–8858 (2019). doi:10.1073/pnas.1901979116 Medline
28. B. J. Wittmann, Y. Yue, F. H. Arnold, Informed training set design enables efficient machine learning-assisted directed protein evolution. *Cell Syst.* **12**, 1026–1045.e7 (2021). doi:10.1016/j.cels.2021.07.008 Medline
29. E. C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi, G. M. Church, Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **16**, 1315–1322 (2019). doi:10.1038/s41592-019-0598-1 Medline
30. S. Biswas, G. Khimulya, E. C. Alley, K. M. Esvelt, G. M. Church, Low-N protein engineering with data-efficient deep learning. *Nat. Methods* **18**, 389–396 (2021). doi:10.1038/s41592-021-01100-y Medline
31. L. Brenan, A. Andreev, O. Cohen, S. Pantel, A. Kamburov, D. Cacchiarelli, N. S. Persky, C. Zhu, M. Bagul, E. M. Goetz, A. B. Burgin, L. A. Garraway, G. Getz, T. S. Mikkelsen, F. Piccioni, D. E. Root, C. M. Johannessen, Phenotypic Characterization of a Comprehensive Set of MAPK1/ERK2 Missense Mutants. *Cell Rep.* **17**, 1171–1183 (2016). doi:10.1016/j.celrep.2016.09.061 Medline
32. P. Notin, A. W. Kollasch, D. Ritter, L. van Niekerk, S. Paul, H. Spinner, N. Rollins, A. Shaw, R. Weitzman, J. Frazer, M. Dias, D. Franceschi, R. Orenbuch, Y. Gal, D. S. Marks, ProteinGym: Large-Scale Benchmarks for Protein Design and Fitness Prediction. bioRxiv 2023.12.07.570727 [Preprint] (2023); https://doi.org/10.1101/2023.12.07.570727.
33. T. Hino, S. N. Omura, R. Nakagawa, T. Togashi, S. N. Takeda, T. Hiramoto, S. Tasaka, H. Hirano, T. Tokuyama, H. Uosaki, S. Ishiguro, M. Kagieva, H. Yamano, Y. Ozaki, D. Motooka, H. Mori, Y. Kirita, Y. Kise, Y. Itoh, S. Matoba, H. Aburatani, N. Yachie, T. Karvelis, V. Siksnys, T. Ohmori, A. Hoshino, O. Nureki, An AsCas12f-based compact genome-editing tool derived by deep mutational scanning and structural analysis. *Cell* **186**, 4920–4935.e23 (2023). doi:10.1016/j.cell.2023.08.031 Medline
34. H. K. Haddox, A. S. Dingens, J. D. Bloom, Experimental Estimation of the Effects of All Amino-Acid Mutations to HIV's Envelope Protein on Viral Replication in Cell Culture. *PLOS Pathog.* **12**, e1006114 (2016). doi:10.1371/journal.ppat.1006114 Medline
35. E. D. Kelsic, H. Chung, N. Cohen, J. Park, H. H. Wang, R. Kishony, RNA Structural Determinants of Optimal Codons Revealed by MAGE-Seq. *Cell Syst.* **3**, 563–571.e6 (2016). doi:10.1016/j.cels.2016.11.004 Medline
36. M. A. Stiffler, D. R. Hekstra, R. Ranganathan, Evolvability as a function of purifying selection in TEM-1 β-lactamase. *Cell* **160**, 882–892 (2015). doi:10.1016/j.cell.2015.01.035 Medline
37. C. J. Markin, D. A. Mokhtari, F. Sunden, M. J. Appel, E. Akiva, S. A. Longwell, C. Sabatti, D. Herschlag, P. M. Fordyce, Revealing enzyme functional architecture via high-throughput microfluidic enzyme kinetics. *Science* **373**, eabf8761 (2021). doi:10.1126/science.abf8761 Medline
38. A. O. Giacomelli, X. Yang, R. E. Lintner, J. M. McFarland, M. Duby, J. Kim, T. P. Howard, D. Y. Takeda, S. H. Ly, E. Kim, H. S. Gannon, B. Hurhula, T. Sharpe, A. Goodale, B. Fritchman, S. Steelman, F. Vazquez, A. Tsherniak, A. J. Aguirre, J. G. Doench, F. Piccioni, C. W. M. Roberts, M. Meyerson, G. Getz, C. M. Johannessen, D. E. Root, W. C. Hahn, Mutational processes shape the landscape of TP53 mutations in human cancer. *Nat. Genet.* **50**, 1381–1387

(2018). doi:10.1038/s41588-018-0204-y Medline
39. E. M. Jones, N. B. Lubock, A. J. Venkatakrishnan, J. Wang, A. M. Tseng, J. M. Paggi, N. R. Latorraca, D. Cancilla, M. Satyadi, J. E. Davis, M. M. Babu, R. O. Dror, S. Kosuri, Structural and functional characterization of G protein-coupled receptors with deep mutational scanning. *eLife* **9**, e54895 (2020). doi:10.7554/eLife.54895 Medline
40. M. B. Doud, J. D. Bloom, Accurate Measurement of the Effects of All Amino-Acid Mutations on Influenza Hemagglutinin. *Viruses* **8**, 155 (2016). doi:10.3390/v8060155 Medline
41. J. M. Lee, J. Huddleston, M. B. Doud, K. A. Hooper, N. C. Wu, T. Bedford, J. D. Bloom, Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human H3N2 influenza variants. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E8276–E8285 (2018). doi:10.1073/pnas.1806133115 Medline
42. M. Sourisseau, D. J. P. Lawrence, M. C. Schwarz, C. H. Storrs, E. C. Veit, J. D. Bloom, M. J. Evans, Deep Mutational Scanning Comprehensively Maps How Zika Envelope Protein Mutations Affect Viral Growth and Antibody Escape. *J. Virol.* **93**, e01291–e19 (2019). doi:10.1128/JVI.01291-19 Medline
43. A. J. Greaney, T. N. Starr, C. O. Barnes, Y. Weisblum, F. Schmidt, M. Caskey, C. Gaebler, A. Cho, M. Agudelo, S. Finkin, Z. Wang, D. Poston, F. Muecksch, T. Hatziioannou, P. D. Bieniasz, D. F. Robbiani, M. C. Nussenzweig, P. J. Bjorkman, J. D. Bloom, Mapping mutations to the SARS-CoV-2 RBD that escape binding by different classes of antibodies. *Nat. Commun.* **12**, 4196 (2021). doi:10.1038/s41467-021-24435-8 Medline
44. A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, R. Fergus, Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2016239118 (2021). doi:10.1073/pnas.2016239118 Medline
45. A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, B. Rost, ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 7112–7127 (2022). doi:10.1109/TPAMI.2021.3095381 Medline
46. J. Funk, L. Machado, S. A. Bradley, M. Napiorkowska, R. Gallegos-Dextre, L. Pashkova, N. G. Madsen, H. Webel, P. V. Phaneuf, T. P. Jenkins, C. G. Acevedo-Rocha, A. I. Proteus, An open-source and user-friendly platform for machine learning-guided protein design and engineering. bioRxiv 2024.10.01.616114 [Preprint] (2024); https://doi.org/10.1101/2024.10.01.616114.
47. J. Deng, Z. Yang, H. Wang, I. Ojima, D. Samaras, F. Wang, Unraveling key elements underlying molecular property prediction: A systematic study. arXiv:2209.13492 [q-bio.QM] (2022).
48. B. Malecova, R. S. Burke, M. Cochran, M. D. Hood, R. Johns, P. R. Kovach, V. R. Doppalapudi, G. Erdogan, J. D. Arias, B. Darimont, C. D. Miller, H. Huang, A. Geall, H. S. Younis, A. A. Levin, Targeted tissue delivery of RNA therapeutics using antibody-oligonucleotide conjugates (AOCs). *Nucleic Acids Res.* **51**, 5901–5910 (2023). doi:10.1093/nar/gkad415 Medline
49. T. Sugo, M. Terada, T. Oikawa, K. Miyata, S. Nishimura, E. Kenjo, M. Ogasawara-Shimizu, Y. Makita, S. Imaichi, S. Murata, K. Otake, K. Kikuchi, M. Teratani, Y. Masuda, T. Kamei, S. Takagahara, S. Ikeda, T. Ohtaki, H. Matsumoto, Development of antibody-siRNA conjugate targeted to cardiac and skeletal muscles. *J. Control. Release* **237**, 1–13 (2016). doi:10.1016/j.jconrel.2016.06.036 Medline
50. C.-L. Hsieh, J. A. Goldsmith, J. M. Schaub, A. M. DiVenere, H.-C. Kuo, K. Javanmardi, K. C. Le, D. Wrapp, A. G. Lee, Y. Liu, C.-W. Chou, P. O. Byrne, C. K. Hjorth, N. V. Johnson, J. Ludes-Meyers, A. W. Nguyen, J. Park, N. Wang, D. Amengor, J. J. Lavinder, G. C. Ippolito, J. A. Maynard, I. J. Finkelstein, J. S. McLellan, Structure-based design of prefusion-stabilized SARS-CoV-2 spikes. *Science* **369**, 1501–1505 (2020). doi:10.1126/science.abd0826 Medline
51. C. Xin, J. Yin, S. Yuan, L. Ou, M. Liu, W. Zhang, J. Hu, Comprehensive assessment of miniature CRISPR-Cas12f nucleases for gene disruption. *Nat. Commun.* **13**, 5623 (2022). doi:10.1038/s41467-022-33346-1 Medline
52. Z. Wu, Y. Zhang, H. Yu, D. Pan, Y. Wang, Y. Wang, F. Li, C. Liu, H. Nan, W. Chen, Q. Ji, Programmed genome editing by a miniature CRISPR-Cas12f nuclease. *Nat. Chem. Biol.* **17**, 1132–1138 (2021). doi:10.1038/s41589-021-00868-6 Medline
53. X. Xu, A. Chemparathy, L. Zeng, H. R. Kempton, S. Shang, M. Nakamura, L. S. Qi, Engineered miniature CRISPR-Cas system for mammalian genome regulation and editing. *Mol. Cell* **81**, 4333–4345.e4 (2021). doi:10.1016/j.molcel.2021.08.008 Medline
54. B. P. Kleinstiver, A. A. Sousa, R. T. Walton, Y. E. Tak, J. Y. Hsu, K. Clement, M. M. Welch, J. E. Horng, J. Malagon-Lopez, I. Scarfò, M. V. Maus, L. Pinello, M. J. Aryee, J. K. Joung, Engineered CRISPR-Cas12a variants with increased activities and improved targeting ranges for gene, epigenetic and base editing. *Nat. Biotechnol.* **37**, 276–282 (2019). doi:10.1038/s41587-018-0011-0 Medline
55. X. Kong, H. Zhang, G. Li, Z. Wang, X. Kong, L. Wang, M. Xue, W. Zhang, Y. Wang, J. Lin, J. Zhou, X. Shen, Y. Wei, N. Zhong, W. Bai, Y. Yuan, L. Shi, Y. Zhou, H. Yang, Engineered CRISPR-OsCas12f1 and RhCas12f1 with robust activities and expanded target range for genome editing. *Nat. Commun.* **14**, 2046 (2023). doi:10.1038/s41467-023-37829-7 Medline
56. L. Zhang, J. A. Zuris, R. Viswanathan, J. N. Edelstein, R. Turk, B. Thommandru, H. T. Rube, S. E. Glenn, M. A. Collingwood, N. M. Bode, S. F. Beaudoin, S. Lele, S. N. Scott, K. M. Wasko, S. Sexton, C. M. Borges, M. S. Schubert, G. L. Kurgan, M. S. McNeill, C. A. Fernandez, V. E. Myer, R. A. Morgan, M. A. Behlke, C. A. Vakulskas, AsCas12a ultra nuclease facilitates the rapid generation of therapeutic cell medicines. *Nat. Commun.* **12**, 3908 (2021). doi:10.1038/s41467-021-24017-8 Medline
57. D. Y. Kim, J. M. Lee, S. B. Moon, H. J. Chin, S. Park, Y. Lim, D. Kim, T. Koo, J.-H. Ko, Y.-S. Kim, Efficient CRISPR editing with a hypercompact Cas12f1 and engineered guide RNAs delivered by adeno-associated virus. *Nat. Biotechnol.* **40**, 94–102 (2022). doi:10.1038/s41587-021-01009-z Medline
58. P. Pausch, B. Al-Shayeb, E. Bisom-Rapp, C. A. Tsuchida, Z. Li, B. F. Cress, G. J. Knott, S. E. Jacobsen, J. F. Banfield, J. A. Doudna, CRISPR-CasΦ from huge phages is a hypercompact genome editor. *Science* **369**, 333–337 (2020). doi:10.1126/science.abb1400 Medline
59. F. A. Ran, L. Cong, W. X. Yan, D. A. Scott, J. S. Gootenberg, A. J. Kriz, B. Zetsche, O. Shalem, X. Wu, K. S. Makarova, E. V. Koonin, P. A. Sharp, F. Zhang, In vivo genome editing using Staphylococcus aureus Cas9. *Nature* **520**, 186–191 (2015). doi:10.1038/nature14299 Medline
60. S. Bae, J. Park, J.-S. Kim, Cas-OFFinder: A fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics* **30**, 1473–1475 (2014). doi:10.1093/bioinformatics/btu048 Medline
61. J. L. Doman, S. Pandey, M. E. Neugebauer, M. An, J. R. Davis, P. B. Randolph, A. McElroy, X. D. Gao, A. Raguram, M. F. Richter, K. A. Everette, S. Banskota, K. Tian, Y. A. Tao, J. Tolar, M. J. Osborn, D. R. Liu, Phage-assisted evolution and protein engineering yield compact, efficient prime editors. *Cell* **186**, 3983–4002.e26 (2023). doi:10.1016/j.cell.2023.07.039 Medline
62. M. T. N. Yarnall, E. I. Ioannidi, C. Schmitt-Ulms, R. N. Krajeski, J. Lim, L. Villiger, W. Zhou, K. Jiang, S. K. Garushyants, N. Roberts, L. Zhang, C. A. Vakulskas, J. A. Walker, A. P. Kadina, A. E. Zepeda, K. Holden, H. Ma, J. Xie, G. Gao, L. Foquet, G. Bial, S. K. Donnelly, Y. Miyata, D. R. Radiloff, J. M. Henderson, A. Ujita, O. O. Abudayyeh, J. S. Gootenberg, Drag-and-drop genome insertion of large sequences without double-strand DNA cleavage using CRISPR-directed integrases. *Nat. Biotechnol.* **41**, 500–512 (2022). Medline
63. A. Baranauskas, S. Paliksa, G. Alzbutas, M. Vaitkevicius, J. Lubiene, V. Letukiene, S. Burinskas, G. Sasnauskas, R. Skirgaila, Generation and characterization of new highly thermostable and processive M-MuLV reverse transcriptase variants. *Protein Eng. Des. Sel.* **25**, 657–668 (2012). doi:10.1093/protein/gzs034 Medline
64. C. Zheng, S.-Q. Liang, B. Liu, P. Liu, S.-Y. Kwan, S. A. Wolfe, W. Xue, A flexible split prime editor using truncated reverse transcriptase improves dual-AAV delivery in mouse liver. *Mol. Ther.* **30**, 1343–1351 (2022). doi:10.1016/j.ymthe.2022.01.005 Medline
65. Z. Gao, S. Ravendran, N. S. Mikkelsen, J. Haldrup, H. Cai, X. Ding, S. R. Paludan, M. K. Thomsen, J. G. Mikkelsen, R. O. Bak, A truncated reverse transcriptase enhances prime editing by split AAV vectors. *Mol. Ther.* **30**, 2942–2951 (2022). doi:10.1016/j.ymthe.2022.07.001 Medline
66. Y. Zong, Y. Liu, C. Xue, B. Li, X. Li, Y. Wang, J. Li, G. Liu, X. Huang, X. Cao, C. Gao, An engineered prime editor with enhanced editing efficiency in plants. *Nat.*

Biotechnol. **40**, 1394–1402 (2022). doi:10.1038/s41587-022-01254-w Medline
67. J. Meier, R. Rao, R. Verkuil, J. Liu, T. Sercu, A. Rives, Language models enable zero-shot prediction of the effects of mutations on protein function, *bioRxiv* (2021)p. 2021.07.09.450648.
68. A. Dousis, K. Ravichandran, E. M. Hobert, M. J. Moore, A. E. Rabideau, An engineered T7 RNA polymerase that produces mRNA free of immunostimulatory byproducts. *Nat. Biotechnol.* **41**, 560–568 (2023). doi:10.1038/s41587-022-01525-6 Medline
69. Z. J. Kartje, H. I. Janis, S. Mukhopadhyay, K. T. Gagnon, Revisiting T7 RNA polymerase transcription in vitro with the Broccoli RNA aptamer as a simplified real-time fluorescent reporter. *J. Biol. Chem.* **296**, 100175 (2021). doi:10.1074/jbc.RA120.014553 Medline
70. R. Chen, S. K. Wang, J. A. Belk, L. Amaya, Z. Li, A. Cardenas, B. T. Abe, C.-K. Chen, P. A. Wender, H. Y. Chang, Author Correction: Engineering circular RNA for enhanced protein production. *Nat. Biotechnol.* **41**, 293 (2023). doi:10.1038/s41587-022-01393-0 Medline
71. Y. Serrano, Á. Ciudad, A. Molina, Are Protein Language Models Compute Optimal? arXiv:2406.07249 [q-bio.BM] (2024).
72. X. Cheng, B. Chen, P. Li, J. Gong, J. Tang, L. Song, Training Compute-Optimal Protein Language Models. bioRxiv 2024.06.06.597716 [Preprint] (2024); https://doi.org/10.1101/2024.06.06.597716.
73. B. Chen, X. Cheng, P. Li, Y.-A. Geng, J. Gong, S. Li, Z. Bei, X. Tan, B. Wang, X. Zeng, C. Liu, A. Zeng, Y. Dong, J. Tang, L. Song, xTrimoPGLM: Unified 100B-Scale Pre-trained Transformer for Deciphering the Language of Protein. bioRxiv 2023.07.05.547496 [Preprint] (2024); https://doi.org/10.1101/2023.07.05.547496.
74. J. Wintermute, S. Ritter, AminoAcid-0 (AA-0): A Protein LLM Trained with 2 Billion Proprietary Sequences, *Ginkgo Bioworks* (2024). https://www.ginkgobioworks.com/2024/09/17/aa-0-protein-llm-technical-review/.
75. P. Notin, N. Rollins, Y. Gal, C. Sander, D. Marks, Machine learning for functional protein design. *Nat. Biotechnol.* **42**, 216–228 (2024). doi:10.1038/s41587-024-02127-0 Medline
76. S. Gelman, B. Johnson, C. Freschlin, S. D'Costa, A. Gitter, P. A. Romero, Biophysics-based protein language models for protein engineering. bioRxiv 2024.03.15.585128 [Preprint] (2024); https://doi.org/10.1101/2024.03.15.585128.
77. K. Jiang, EVOLVEpro, Zenodo (2024); https://doi.org/10.5281/zenodo.13899358.
78. K. Jiang, J. Koob, X. D. Chen, R. N. Krajeski, Y. Zhang, V. Volf, W. Zhou, S. R. Sgrizzi, L. Villiger, J. S. Gootenberg, F. Chen, O. O. Abudayyeh, Programmable eukaryotic protein synthesis with RNA sensors by harnessing ADAR. *Nat. Biotechnol.* **41**, 698–707 (2023). doi:10.1038/s41587-022-01534-5 Medline
79. R. Chen, S. K. Wang, J. A. Belk, L. Amaya, Z. Li, A. Cardenas, B. T. Abe, C.-K. Chen, P. A. Wender, H. Y. Chang, Engineering circular RNA for enhanced protein production. *Nat. Biotechnol.* **41**, 262–272 (2023). doi:10.1038/s41587-022-01393-0 Medline
80. L. Gieselmann, C. Kreer, M. S. Ercanoglu, N. Lehnen, M. Zehner, P. Schommers, J. Potthoff, H. Gruell, F. Klein, Effective high-throughput isolation of fully human antibodies targeting infectious pathogens. *Nat. Protoc.* **16**, 3639–3671 (2021). doi:10.1038/s41596-021-00554-w Medline
81. X. Wang, S. Liu, Y. Sun, X. Yu, S. M. Lee, Q. Cheng, T. Wei, J. Gong, J. Robinson, D. Zhang, X. Lian, P. Basak, D. J. Siegwart, Preparation of selective organ-targeting (SORT) lipid nanoparticles (LNPs) using multiple technical methods for tissue-specific mRNA delivery. *Nat. Protoc.* **18**, 265–291 (2023). doi:10.1038/s41596-022-00755-x Medline

ACKNOWLEDGMENTS

We would like to thank K. Yang and R. Rajendran for help with cloning, mRNA preparation and next-generation sequencing in this study; E. Boyden for MiSeq instrumentation support; A. Hoshino for providing AsCas12f DMS datasets; B. Berger, D. Marks, J. Yim, A. Kirjner, I. Fiete, M. Yan, B. Li, M. Guo, T. Chen, X. Yang, C.J. Bashor, P. Mehta, and J. Rocks for computational discussions; J. Xie, G. Gao for viral preparation; H. Y. Chang and R. A. Wesselhoeft for circular RNA vector and purification; and Z. Tang, D. Irvine, J. S. Weissman, M. Birnbaum, R. Desimone, and J. Crittenden for support and helpful discussions. We thank the members of the Nishimasu and Abudayyeh-Gootenberg labs for their helpful discussions. **Funding:** J.S.G. and O.O.A. are supported by NIH grants 1R21-AI149694, R01-EB031957, 1R01GM148745, R56-HG011857, and R01AG074932; the K. Lisa Yang and Hock E. Tan Center for Molecular Therapeutics in Neuroscience; Impetus Grants; the Cystic Fibrosis Foundation Pioneer Grant; Google Ventures; Pivotal Life Sciences; MGB Gene and Cell Therapy Institute; and the Yosemite Fund. H.N. is supported by JSPS KAKENHI Grant Numbers 21H05281 and 22H00403, the Takeda Medical Research Foundation, the Inamori Research Institute for Science, and JST, CREST Grant Numbers JPMJCR23B6. M.D. is supported by the NSF Graduate Research Fellowship Program (GRFP) NSF 24-591. **Authors contributions:** Conceptualization: K.J., O.O.A., J.S.G.; Methodology: K.J., Z.Y., M.D.B., M.H., H.N., B.K., J.K.C.; Investigation: K.J., O.O.A., J.S.G., Z.Y., A.K., L.V., S.R.S.; Visualization: K.J., M.D.B., Z.Y.; Funding acquisition: O.O.A., J.S.G., H.N.; Project administration: O.O.A., J.S.G.; Supervision: O.O.A., J.S.G.; Writing – original draft: K.J., O.O.A., J.S.G., Z.Y., M.D.B., M.H., H.N.; Writing – review and editing: K.J., Z.Y., M.D.B., S.R.R., L.V., A.K., B.K., J.K.C., M.H., H.N., J.S.G., O.O.A. **Competing interests:** J.S.G., O.O.A, K.J., MDB, L.V., and Z.Y. have filed patents related to this work with provisional patent (#63/509,139). J.S.G. and O.O.A. are cofounders of Sherlock Biosciences, Tome Biosciences, Doppler Biosciences, and Circle Labs. No companies are currently involved in commercializing this technology. **Data and materials availability:** Expression plasmids are available from Addgene under UBMTA; supporting information is available within this document and supplementary materials. Models and codes are available at Github (https://github.com/mat10d/EvolvePro) and are archived at Zenodo (*77*).

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adr60006
Materials and Methods
Supplementary Text
Figs. S1 to S12
Tables S1 and S2
References (*78–81*)
MDAR Reproducibility Checklist
Data S1 to S7

Submitted 09 July 2024; accepted 12 November 2024
Published online 21 November 2024
10.1126/science.adr6006

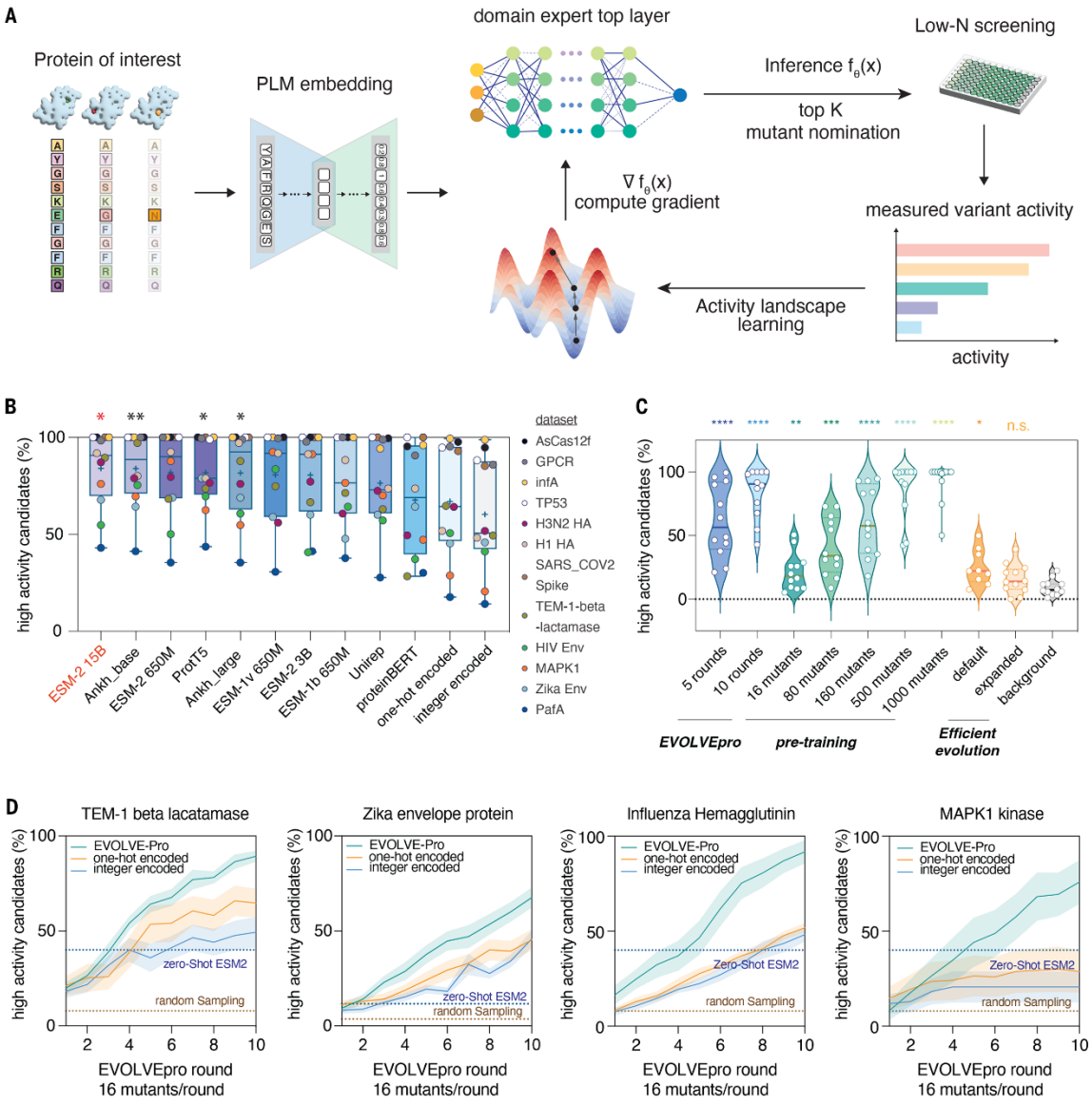

Fig. 1. Developing and benchmarking EVOLVEpro for protein language model-guided engineering. (A) Schematic describing the EVOLVEpro method. Proteins of interest go through iterative rounds of low-N screening. A foundational PLM generates embeddings for all mutants of a protein and the average embedding by pooling across all residues is used as input for the top-layer model. Each mutant's activity is experimentally determined and used to train a domain expert top-layer model with PLM embedding as input. The top-layer model then nominates the top-N mutants for the next round of testing and the weights are updated iteratively in an active learning format. (**B**) Benchmarking of foundational models across a panel of 12 comprehensive deep mutational scanning (DMS) datasets. Each point is a unique protein and its DMS data. ESM2-15B has the highest average percent success in high activity variants prediction. (**C**) Comparison between EVOLVEpro in active learning format, in zero-shot pretraining format, and an existing zero-shot prediction method using protein language model (*15*) across 12 DMS datasets. Each point is a unique protein using its DMS data. (**D**) Performance over 10 rounds of EVOLVEpro with 16 mutants per round, compared to two different non-language model encoding schemes (one-hot encoding and integer encoding). Model performance is benchmarked on four datasets (*31*, *36*, *40*, *42*) and compared to zero-shot ESM2 nomination success rate and background random sampling (*15*). Error bar represents the standard deviation for n=10 random simulations.

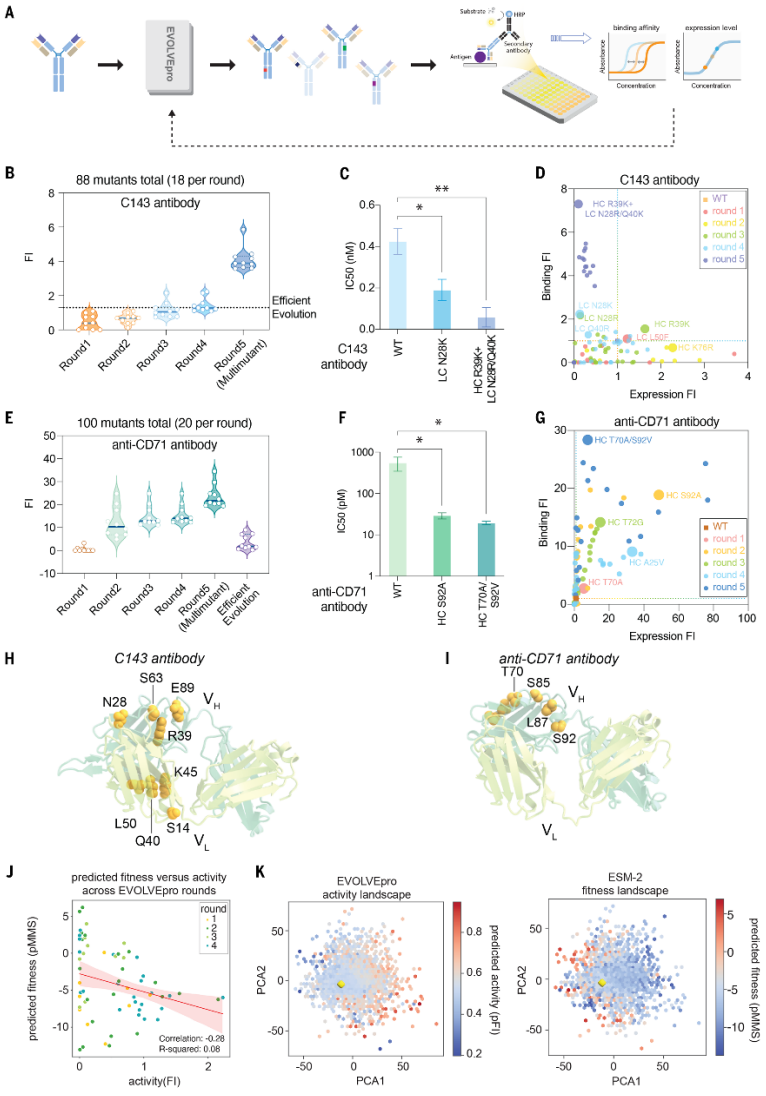

Fig. 2. Evolution of two monoclonal antibodies with EVOLVEpro. (**A**) Schematic of the evolution strategy with EVOLVEpro for engineering two monoclonal antibodies across two parameters (binding affinity and antibody expression). (**B**) Engineering of the C143 antibody over five rounds of EVOLVEpro. Data shows cumulative top 10 mutants' fold improvement over WT binding affinity to the target antigen across 5 evolution rounds. (**C**) IC_{50} value estimated from ELISA binding data for the WT C143 antibody, the best single mutant (LC N28K) and the best multi-mutant (LC N28R/Q40K+HC R39K). Error bars represent standard error of mean with n=3 technical replicates. A one way ANOVA was run between the three groups (*p<0.05, **p<0.01). (**D**) Scatter plot showing each individual mutant's expression fold improvement versus binding affinity improvement for the C143 antibody. The best mutant in each round is highlighted with a larger circle. (**E**) Engineering of the aCD71 over five rounds of EVOLVEpro. Data shows cumulative top 10 mutants' fold improvement over WT binding affinity to the target antigen across 5 evolution rounds. (**F**) IC_{50} value estimated from ELISA binding data for the WT anti-CD71 antibody, the best single mutant (S92A) and the best multi-mutant (T70A_S92V). Y axis is shown on log 10 scale. Error bars represent standard error of mean with n=3 technical replicates. A one way ANOVA was run between the three groups (*p<0.05, **p<0.01). (**G**) Scatter plot showing each individual mutant's expression fold improvement versus binding affinity improvement for the aCD71 antibody. The best mutant in each round is highlighted with a larger circle. (**H** and **I**) Mapping of the top mutations on the predicted structure of C143 (H) and anti-CD71 (I) respectively (AF3). (**J**) Scatter plot comparing the predicted naive ESM-2 C143 protein fitness (predicted masked marginal score) and scaled tested activity of nominated mutants across evolution. Scatter points are colored by rounds in evolution. The correlation and linear regression line are shown in red and the R square of the correlation is reported. (**K**) Comparison of the C143 embedding latent space with either predicted naive ESM-2 protein fitness landscape or EVOLVEpro protein activity landscape. Yellow rhombus denotes WT sequence.

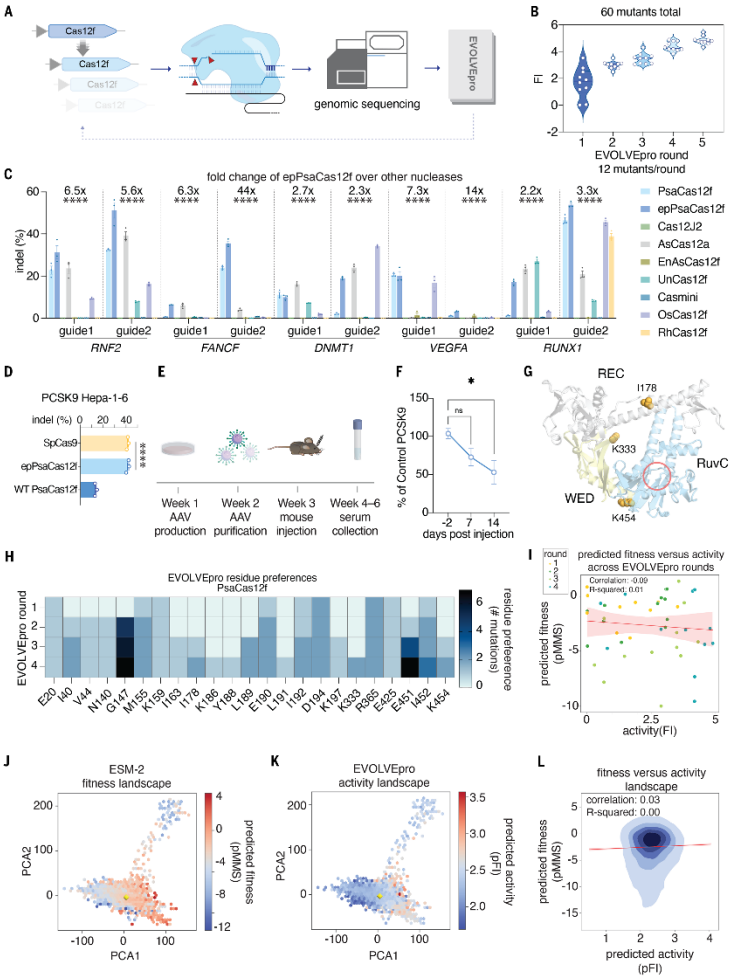

Fig. 3. Evolution of highly active miniature CRISPR nucleases with EVOLVEpro. (**A**) Schematic of the evolution strategy with EVOLVEpro for engineering a miniature Cas12f. (**B**) Engineering of PsaCas12f over four rounds of EVOLVEpro and a rational combination multi-mutant round. Data shows cumulative top 10 mutants from current and preceding rounds, as measured by fold improvement of indel activity at the endogenous RNF2 genomic locus. (**C**) Indel activities of WT PsaCas12f, epPsaCas12f, and a panel of published Cas12a and Cas12f nucleases on 10 different genomic targets across five genes (two guides per gene). The fold change on top of each guide denotes the relative fold increase of epPsaCas12f compared to the average of the other published Cas12a and Cas12f nucleases. A one-way ANOVA is performed for each guide sequence shown (****p<0.0001). Error bars represent standard error of mean with n=3 biological replicates. (**D**) Next-generation sequencing quantified indel formation at murine PCSK9 genomic loci by epPsaCas12f, WT PsaCas12f, and SpCas9. A one-way ANOVA is performed for each guide sequence shown (****p<0.0001). Error bars represent standard error of mean with n=3 biological replicates. (**E**) Schematic of the in vivo validation assay for EnPsaCas12f editing at the murine PCSK9 locus for PCSK9 reduction. (**F**) Serum PCSK9 levels at three different time points from −2 days of injection to +14 days. The percent of control PCSK9 was calculated by normalizing to the control group with PBS injected. A two-sided Student's *t* test was run on each time point relative to −2 days' baseline PCSK9 level (ns, non-significant; *p<0.05). Error bars represent standard error of mean with n=3 biological replicates. (**G**) Mapping of the top mutations on the AlphaFold3 model of PsaCas12f. The RuvC active site is indicated by a red circle. (**H**) Heatmap showing most common PsaCas12f mutations explored by EVOLVEpro over rounds of evolution. Any position explored more than once is shown on a cumulative basis across rounds. (**I**) Scatter plot comparing the predicted naive ESM-2 protein fitness (predicted masked marginal score) and scaled tested activity of nominated mutants across evolution, scatter points are colored by rounds in evolution. (**J** and **K**) Comparison of the PsaCas12f embedding latent space with either predicted naive ESM-2 protein fitness landscape (J) or EVOLVEpro protein activity landscape (K). Yellow rhombus denotes WT sequence. (**L**) A kernel density estimate plot of protein fitness as predicted by ESM-2 versus protein activity as predicted by EVOLVEpro. The correlation and linear regression line are shown in red and the R square of the correlation is reported.

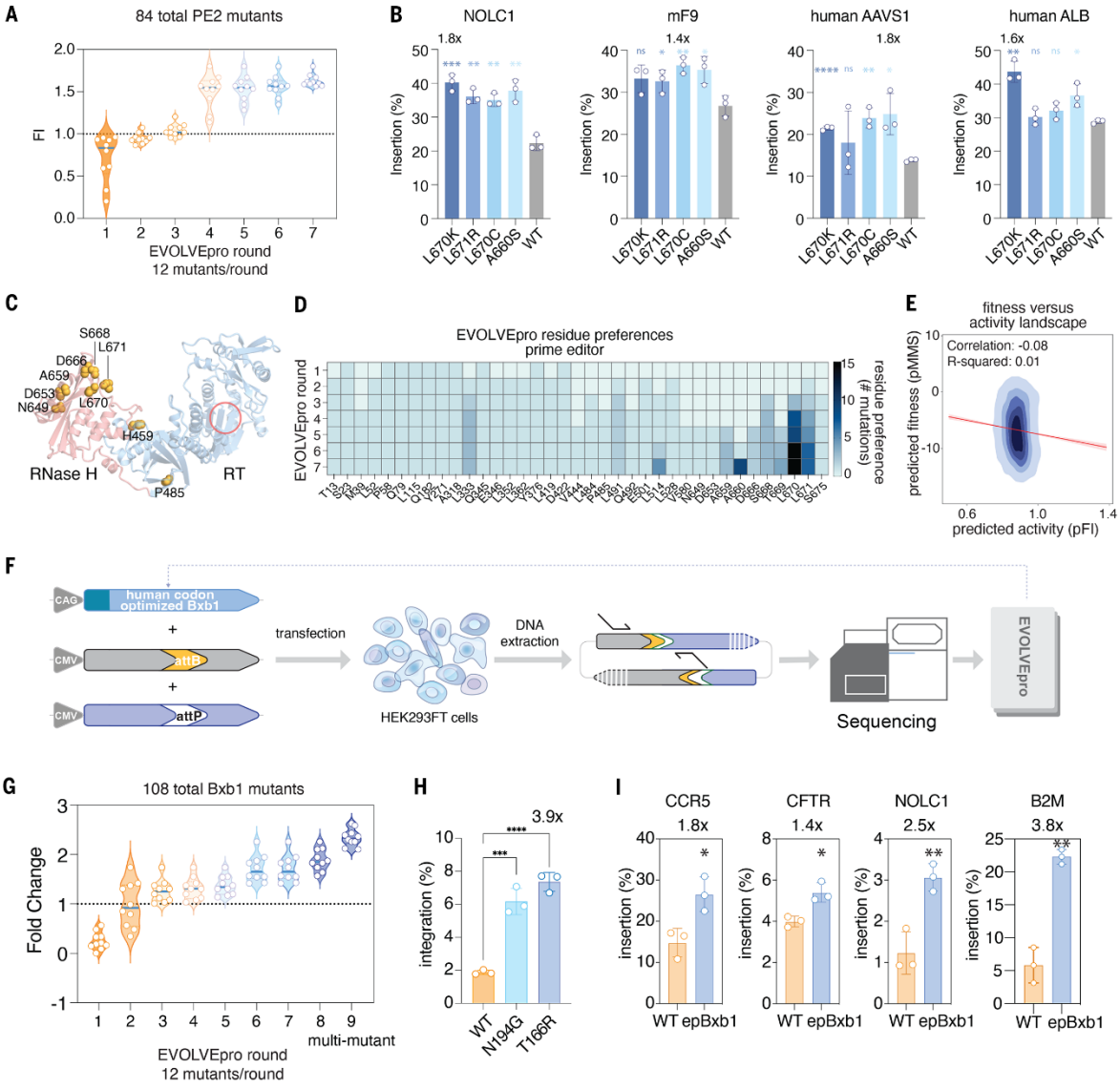

Fig. 4. Evolution of prime editor with EVOLVEpro. (**A**) Engineering of the prime editor PE2 with twinPE guides over seven rounds of EVOLVEpro. Data shows cumulative top 10 mutants from current and preceding rounds, as measured by fold improvement of prime editing activity to install a 46 bp AttB site at the murine NOLC1 genomic locus. (**B**) Validation of 4 evolved prime editors in the installation of attB sites at four different endogenous sites in either mouse or human genomes. A two-sided unpaired *t* test was run between WT and each evolved prime editor (ns, non-significant; *p<0.05; **p<0.01; ***p<0.001; ****p<0.0001). Fold change over WT PE2 is shown for the best mutant on each genomic locus. Error bars represent standard deviation with n=3 biological replicates. (**C**) Mapping of the top mutations on the AlphaFold3 model of M-MLV RT. The RT active site is indicated by a red circle. (**D**) Heatmap showing most common PE2 mutations explored by EVOLVEpro over rounds of evolution. Any position explored more than once is shown on a cumulative basis across rounds. (**E**) Scatter plot comparing the predicted naive ESM-2 protein fitness (predicted masked marginal score) and scaled tested activity of nominated mutants across evolution, scatter points are colored by rounds in evolution. (**F**) Schematic of the evolution strategy for evolving the Bxb1 serine integrase from the Mycobacteriophage. (**G**) Engineering of the Bxb1 integrase over 8 rounds of EVOLVEpro. Data shows cumulative top 10 mutants from current and preceding rounds, as measured by fold improvement of plasmid integration over WT. (**H**) Performance of top Bxb1 mutants for plasmid recombination with low Bxb1 expression in Hela cell. A two-sided Student's *t* test was run between WT and each evolved Bxb1 integrase (***p<0.001, ****p<0.0001). Fold change over WT Bxb1 is shown for the best mutant. Error bars represent standard deviation with n=3 biological replicates. (**I**) Validation of epBxb1 with PASTE at four genomic sites across human and mice genomes. A two-sided Student's *t* test was run between WT and each evolved Bxb1 integrase (*p<0.05, **p<0.01). Fold change over WT Bxb1 integrase is shown for each genomic locus. Error bars represent standard deviation with n=3 biological replicates.

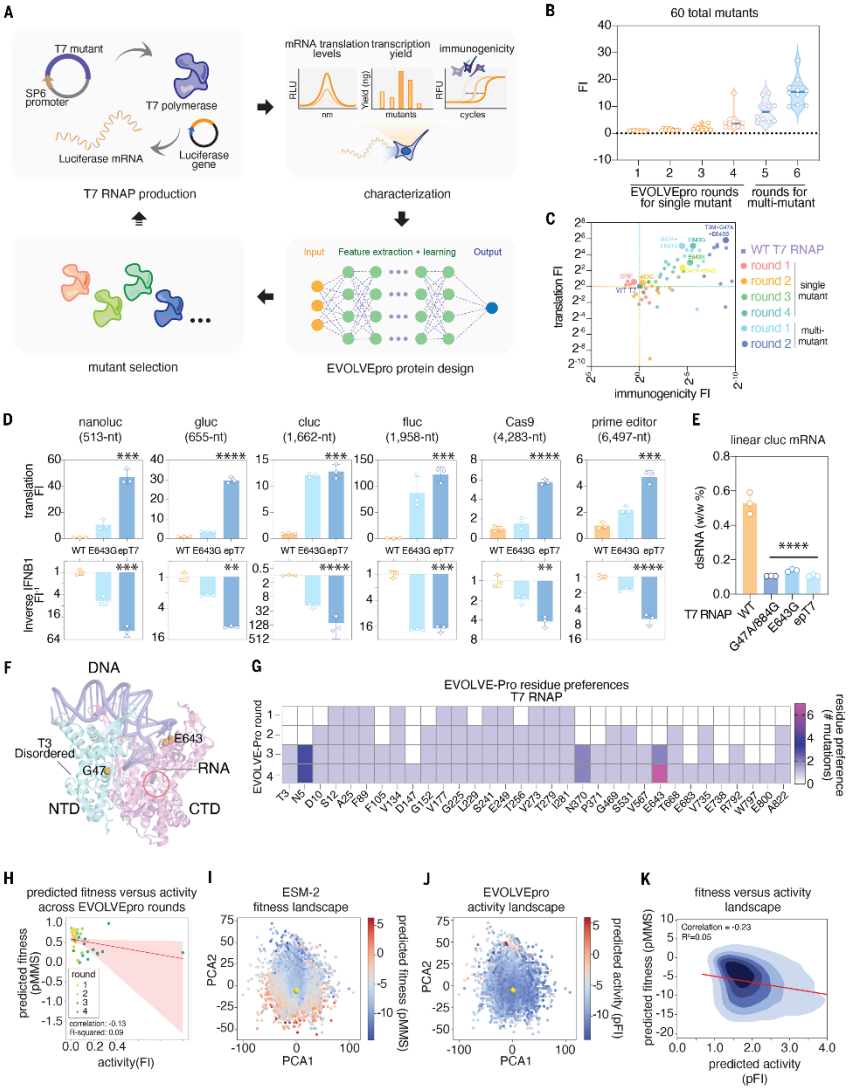

Fig. 5. Engineering RNA polymerases for high yield and low immunogenicity mRNA production. (**A**) Schematic of the strategy for high-throughput T7 RNA polymerases mutant testing and evolution policy setup for evolving a high fidelity T7 RNAP. (**B**) Engineering of T7 RNAP over six rounds of EVOLVEpro. Data shows the top 10 mutants from current and preceding rounds, as measured by fold improvement of transcription fidelity over WT. (**C**) Performance of T7 mutants from six EVOLVEpro rounds and previously engineered G47A/884insG SOTA T7 RNAP in Cluc mRNA translation and immunogenicity in BJ Fibroblast cells. (**D**) Validation of epT7 for production of 6 mRNA sequences ranging from 513nt to 6496nt. Purified WT or mutant RNAP is used to produce these sequences, and they were transfected into BJ fibroblast cells for either protein translation readout or targeted IFNB1 gene expression analysis using quantitative polymerase chain reaction 24 hours after transfection. A two-sided Student's *t* test was run between WT and each evolved T7 RNAP (**p<0.01, ***p<0.001, ****p<0.0001). Error bars represent standard deviation with n=3 biological replicates. (**E**) dsRNA ELISA is used to analyze the amount of dsRNA during transcription of a 1662 nt Cypridina luciferase mRNA. 500 ng of posttranscription product is used as input for the dsRNA ELISA. A two-sided Student's *t* test was run between WT and each evolved T7 RNAP (****p<0.0001). Error bars represent standard deviation with n=3 biological replicates. (**F**) Mapping of the top mutations on the T7 RNAP structure (PDB ID 3E2E). The active site is indicated by a red circle. (**G**) Heatmap showing most common T7 RNAP mutations explored by EVOLVEpro over rounds of evolution. Any position explored more than once is shown on a cumulative basis across rounds. (**H**) Scatter plot comparing the predicted ESM-2 protein fitness score versus experimentally measured T7 RNAP transcription fidelity scaled score across evolution rounds. The correlation and linear regression line are shown in the plot. (**I** and **J**) Comparison of the T7 RNAP latent space with either predicted ESM-2 protein fitness (masked marginal score) or EVOLVEpro protein activity fold improvement. Yellow rhombus denotes WT sequence. (**K**) A kernel density estimate of protein fitness as predicted by ESM-2 versus protein activity as predicted by EVOLVEpro. The correlation and linear regression line are shown in red and the R^2 of correlation is reported.

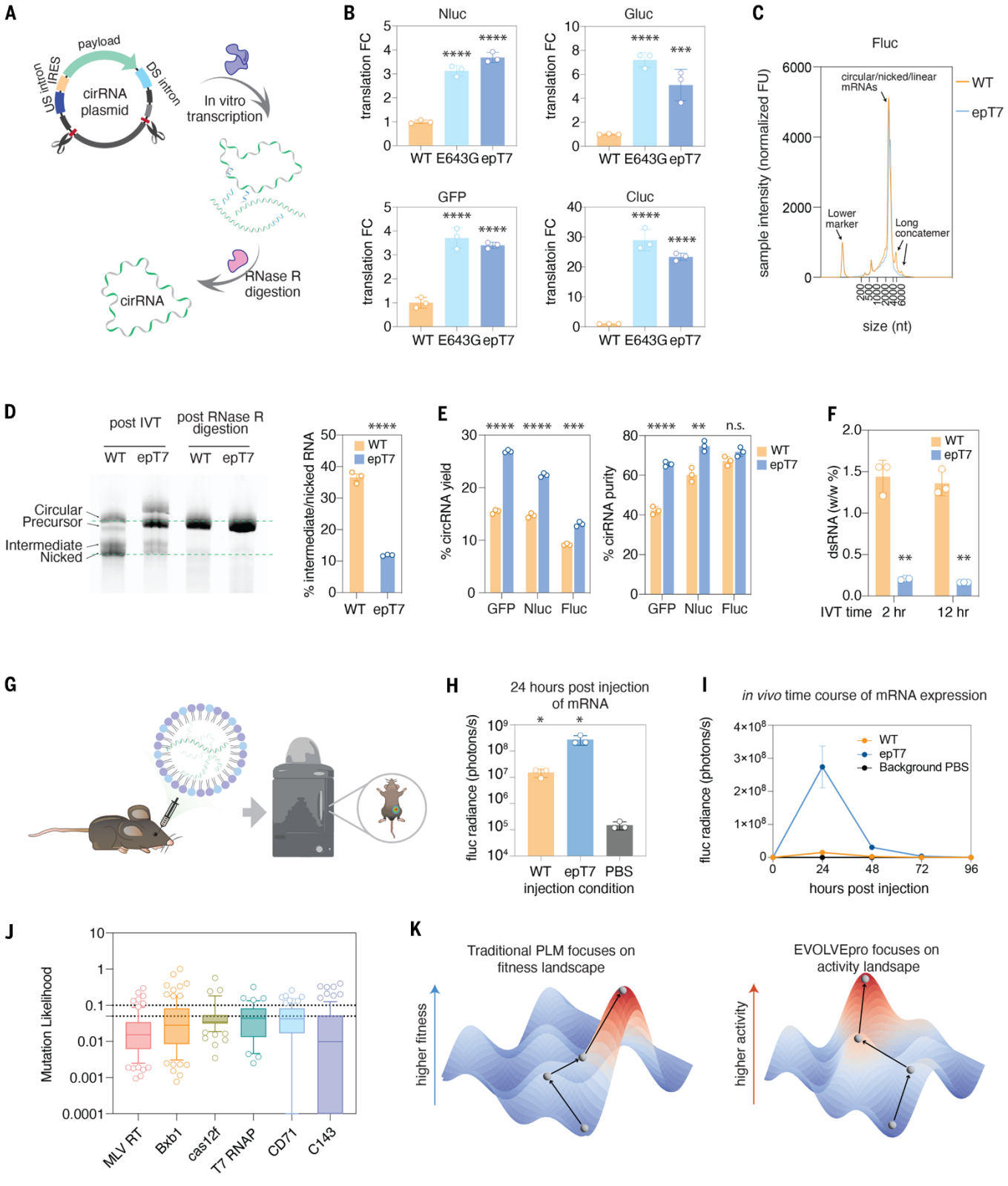

Fig. 6. Application of epT7 for circular RNA production and in vivo bioluminescence. (**A**) Schematic of circular RNA production. (**B**) Validation of epT7 produced circRNA on four different template sequences compared to both $T7^{E643G}$ and WT T7. Translation of each protein is measured in HEK293FT cells 48 hours after transfection. A two-sided Student's *t* test was run between WT and each evolved T7 RNAP (***p<0.001, ****p<0.0001). Error bars represent standard deviation with n=3 biological replicates. (**C**) Tapestation gel electrophoresis analysis of circular Fluc RNA produced by either epT7 or WT RNAP. epT7 shows reduced concatemer production. (**D**) Comparison of RNA products for Fluc circRNA produced by epT7 compared to WT T7 via gel electrophoresis using 2% E-gel EX at different steps in the production process: post-initial IVT and post–RNase R processing. The panel on the right shows quantification of intermediate and nicked RNA ratio in the post IVT samples. Error bars represent standard deviation with n=3 biological replicates. (**E**) Comparison of purified GFP, nanoluc (Nluc), and Fluc circRNA yield by epT7 compared to WT T7 after the initial RNase R clean-up. The panel on the left shows the raw mass percentage left after the cleanup. The panel on the right shows the purity of the circular RNA in the post clean-up reaction as determined by quantification using a TapeStation analysis. A two-sided Student's *t* test was run between WT and epT7 (**, p<0.01, ****, p<0.0001). (**F**) Comparison of dsRNA content for nanoluc circRNA produced by epT7 compared to WT T7 using either 2 hours of IVT or 12 hours of IVT. Input into the dsRNA ELISA assay involves 500 ng of post–RNase R cleaned-up samples. A two-sided Student's *t* test was run between WT and evolved T7 RNAP (**, p<0.01). Error bars represent standard deviation with n=3 biological replicates. (**G**) Schematic of the in vivo mRNA assay for measuring mRNA expression in the liver via non-invasive luminescent imaging. (**H**) In vivo luminescent signal detected 24 hours post-injection in mice injected with mRNA produced by either epT7 or WT T7 or PBS controls. A two-sided Student's *t* test was run between WT, WT T7 RNAP, and epT7 (*, p<0.05). Error bars represent standard deviation with n=3 biological replicates. (**I**) Time-course of in vivo luminescent signal detected up to 96 hours post-injection of LNP-mRNA produced by either epT7 or WT T7, or PBS controls. A two-sided paired Student's *t* test was run between WT, WT T7 RNAP, and epT7 (*p<0.05) for each time point. Error bars represent the standard error of mean with n=3 biological replicates. (**J**) A box plot of mutational likelihood for each individual mutant nominated by EVLOVEpro shown for each of the six proteins in this study. A dashed line at 0.05 and 0.1 are shown to denote the threshold for rare mutations and uncommon mutations, respectively. (**K**) A schematic showing the evolution of higher activity variants with EVOLVEpro versus traditional PLM evolution approaches. The mutagenesis landscape of proteins is often conceptualized as a complex terrain with numerous potential paths. Shown here is a gray road that conceptualizes the protein mutagenesis landscape where traversing upwards results in higher protein activity and traversing downwards reduces protein fitness. Traditional frameworks of evolutionary plausibility attempt to navigate this terrain based on natural selection, which is constrained by historical and environmental factors.