PROTEIN DESIGN

Simulating 500 million years of evolution with a language model

Thomas Hayes[1]†, Roshan Rao[1]†, Halil Akin[1]†, Nicholas J. Sofroniew[1]†, Deniz Oktay[1]†, Zeming Lin[1]†, Robert Verkuil[1]†, Vincent Q. Tran[2,3], Jonathan Deaton[1], Marius Wiggert[1], Rohil Badkundri[1], Irhum Shafkat[1], Jun Gong[1], Alexander Derry[1], Raul S. Molina[1], Neil Thomas[1], Yousuf A. Khan[1], Chetan Mishra[1], Carolyn Kim[1], Liam J. Bartie[2], Matthew Nemeth[2], Patrick D. Hsu[2,3], Tom Sercu[1], Salvatore Candido[1], Alexander Rives[1]*

More than 3 billion years of evolution have produced an image of biology encoded into the space of natural proteins. Here, we show that language models trained at scale on evolutionary data can generate functional proteins that are far away from known proteins. We present ESM3, a frontier multimodal generative language model that reasons over the sequence, structure, and function of proteins. ESM3 can follow complex prompts combining its modalities and is highly responsive to alignment to improve its fidelity. We have prompted ESM3 to generate fluorescent proteins. Among the generations that we synthesized, we found a bright fluorescent protein at a far distance (58% sequence identity) from known fluorescent proteins, which we estimate is equivalent to simulating 500 million years of evolution.

The proteins that exist today have developed into their present forms over the course of billions of years of natural evolution, passing through a vast evolutionary sieve. In parallel experiments conducted over geological time, nature creates random mutations and applies selection, filtering proteins by their myriad sequences, structures, and functions.

As a result, the patterns in the proteins that we observe today reflect the action of the deep hidden variables of the biology that have shaped their evolution across time. Gene sequencing surveys of Earth's natural diversity are cataloging the sequences (*1*–*3*) and structures (*4*, *5*) of proteins, containing billions of sequences and hundreds of millions of structures that illuminate patterns of variation across life. A consensus is developing that underlying these sequences is a fundamental language of protein biology that can be understood using language models (*6*–*11*).

A number of language models of protein sequences have now been developed and evaluated (*5*–*10*, *12*–*17*). It has been found that the representations that emerge within language models reflect the biological structure and function of proteins (*6*–*8*, *18*) and are learned without any supervision on those properties (*19*, *20*), improving with scale (*5*, *21*). In the field of artificial intelligence, scaling laws have been found that predict the growth in capabilities with increasing scale, describing a frontier in compute, parameters, and data (*22*–*24*).

Here, we present ESM3, a frontier multimodal generative model that reasons over the sequences, structures, and functions of proteins. ESM3 is trained as a generative masked language model over discrete tokens for each modality. Structural reasoning is achieved by encoding three-dimensional (3D) atomic structure as discrete tokens rather than with the complex architecture and diffusion in 3D space used in recent predictive (*25*) and generative models (*26*–*28*) of proteins. All-to-all modeling of discrete tokens is scalable and allows ESM3 to be prompted with any combination of its modalities, thus enabling the controllable generation of proteins that respect combinations of prompts. We observed that ESM3 is highly responsive to prompts and finds creative solutions to complex combinations of prompts, including solutions for which we can find no matching structure in nature. Models at all scales can be aligned to better follow prompts, and larger models are far more responsive to alignment, showing greater capability to solve the most difficult prompts after alignment. Using ESM3, we report the generation of a variant of green fluorescent protein (GFP) (*29*, *30*) that is diverged from existing proteins to a degree equivalent to simulating >500 million years of evolution.

ESM3

ESM3 achieves a scalable generative model of the three fundamental properties of proteins, sequence, structure, and function, through language modeling. Previous generative modeling efforts for proteins have focused primarily on individual modalities, leveraging complex architectures and training objectives for structures that represent proteins as 3D objects. To date, the only language models that have been scaled are for protein sequences. In ESM3, sequence, structure, and function are represented through alphabets of discrete tokens. The modalities are input and output as separate sequence tracks that are fused into a single latent space within the model. This simplicity enables ESM3 to leverage a scalable transformer architecture to train up to 98 billion parameters and more than one trillion teraflops of compute, demonstrating the emergence of complex reasoning capabilities over sequence, structure, and function.

ESM3 is trained with a generative masked language modeling objective across all its tracks as described by the following equation:

$$\mathcal{L} = -\mathbb{E}_{x,m}\frac{1}{|m|}\sum_{i\in m}\log p(x_i|x_{\setminus m})$$

A random mask m is applied to the tokens x describing the protein, and the model is supervised to predict the identity of the tokens that have been masked. During training, the mask is sampled using a noise schedule that varies the fraction of positions that are masked so that ESM3 sees many different combinations of masked sequence, structure, and function and predicts completions of any combination of the modalities from any other. This differs from classical masked language modeling (*31*) in that the supervision is applied across all possible masking rates rather than to a single fixed masking rate. This supervision factorizes the probability distribution over all possible predictions of the next token given any combination of previous tokens, thus ensuring that tokens can be generated in any order from any starting point (*32*–*34*).

To generate from ESM3, tokens are iteratively sampled. Starting from a fully or partially masked context, tokens can be sampled one at a time or in parallel and in any order until all positions are fully unmasked (Fig. 1A). In addition to enabling generation, ESM3's training objective is also effective for representation learning. High masking rates improve the generative capability, whereas lower masking rates improve representation learning. We chose to train ESM3 with a noise schedule that balances generative capabilities with representation learning (supplementary materials, section A.2.2).

ESM3 is a bidirectional transformer. Sequence, structure, and function tokens are embedded and fused at the input and then processed through a stack of transformer blocks (Fig. 1B). At the output of the model, shallow multilayer perceptron heads project the final layer representation into token probabilities for each of the tracks. ESM3 uses tokenization, rather than specialized architectural components, to represent the complexity of proteins in a learned multimodal feature space. This approach enables efficient and highly scalable training.

Protein structures are tokenized by a discrete autoencoder (*35*) that is trained to compress 3D structure into discrete tokens (Fig. 1C). We propose an invariant geometric attention mechanism to efficiently process the

[1]EvolutionaryScale, PBC, New York, NY, USA. [2]Arc Institute, Palo Alto, CA, USA. [3]University of California, Berkeley, Berkeley, CA, USA.
*Corresponding author. Email: arives@evolutionaryscale.ai
†These authors contributed equally to this work.

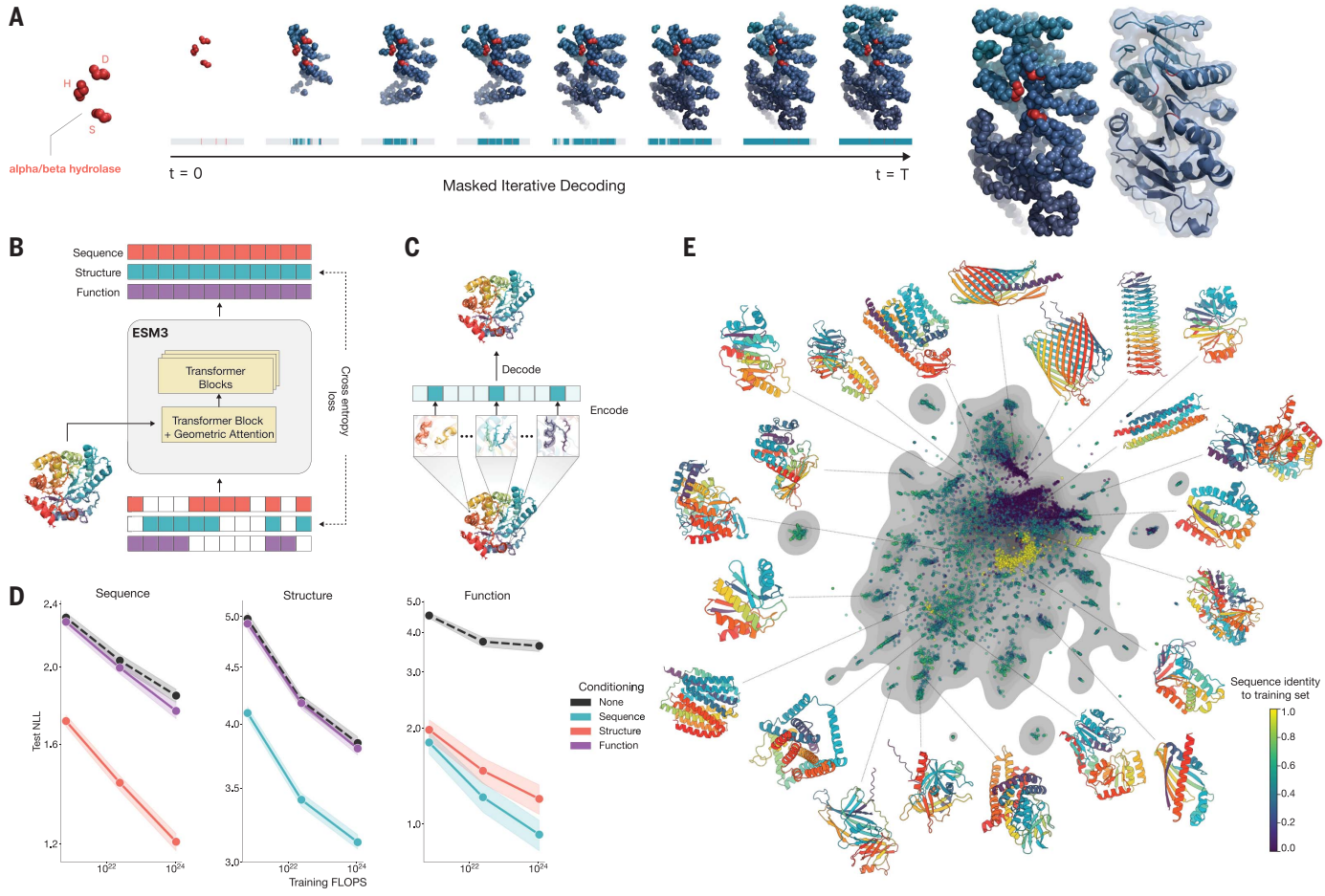

Fig. 1. ESM3 is a generative language model that reasons over the sequence, structure, and function of proteins. (**A**) Iterative sampling with ESM3. Generation of an alpha/beta hydrolase. Sequence, structure, and function can all be used to prompt the model. At each timestep t, a fraction of the masked positions are sampled until all positions are unmasked. (**B**) ESM3 architecture. Sequence, structure, and function are represented as tracks of discrete tokens at the input and output. The model is a series of transformer blocks in which all tracks are fused within a single latent space. Geometric attention in the first block allows conditioning on atomic coordinates. ESM3 is supervised to predict masked tokens. (**C**) Structure tokenization. Local atomic structure around each amino acid is encoded into tokens. (**D**) Models are trained at three scales: 1.4B, 7B, and 98B parameters. Negative log-likelihood (averaged across mask rates) on test set as a function of training FLOPs shows response to conditioning on each of the input tracks, improving with increasing FLOPs (95% confidence interval). (**E**) Unconditional generations from ESM3 98B (colored by sequence identity to the nearest sequence in the training set), embedded by ESM3, and projected by uniform manifold approximation and projection (UMAP) alongside randomly sampled sequences from UniProt (in gray). Generations are diverse, high-quality, and cover the distribution of natural sequences.

3D structure. The mechanism operates in local reference frames defined by the bond geometry at each amino acid and allows local frames to interact globally through a transformation into the global frame (supplementary materials, section A.1.6). The local structural neighborhoods around each amino acid are encoded into a sequence of discrete tokens, one for each amino acid.

When predicting or generating protein structure, the structure tokens output by ESM3 are passed through the decoder, which reconstructs the full atomic structure. The autoencoder is trained to encode and reconstruct atomic coordinates with a geometric loss that supervises the pairwise distances and relative orientations of bond vectors and normals (supplementary materials, section A.1.7.3.1). This tokenization delivers nearly perfect reconstruction of protein structure [<0.5 Å root mean square difference (RMSD) using CAMEO; fig. S3].

Because the local neighborhoods of each structure token contain information about neighboring parts of the structure, we also provided the model with a mechanism to condition on backbone atomic coordinates directly through geometric attention in the first transformer block. To support higher-level abstractions of structure, we included tracks for secondary structure (SS8) tokens and solvent accessible surface area (SASA) tokens. Key words describing biological activity, such as binding, enzymatic function, and domain or fold classifications allow an even higher-level semantic description of protein architecture and function. Derived from free-text descriptions in InterPro (*36*) and Gene Ontology (GO) terms for each residue, these key words are tokenized (supplementary materials, section A.1.8), embedded, and summed at the network input. Residue-level annotations provide multi-hot labeling of the functions of individual residues, such as catalytic sites and posttranslational modifications (supplementary materials, section A.1.8.3).

The largest ESM3 model is trained on 2.78 billion natural proteins collected from sequence and structure databases (*2*, *36*–*39*). Because a small fraction of structures have been experimentally determined relative to sequences, we leveraged predicted structures (*4*, *5*). Sequences were annotated with function key words using a library of hidden Markov models (*40*). We also generated synthetic sequences

with an inverse folding model (supplementary materials, section A.2.1.3) for all structures, including predicted ones. Overall, this increases training data to 3.15 billion protein sequences, 236 million protein structures, and 539 million proteins with function annotations, totaling 771 billion unique tokens. Full details of the training dataset are described in the supplementary materials, section A.2.1.

We trained ESM3 models at three scales: 1.4, 7, and 98 billion parameters (1.4B, 7B, and 98B, respectively). In an initial series of experiments to evaluate representation learning performance in response to architecture hyperparameters, we found a greater response to increasing depth than to width. This behavior informed the choice of relatively deep networks for the final architectures, with the 98B-parameter model incorporating 216 transformer blocks (supplementary materials, section A.1.5).

Scaling ESM3 from 1.4B to 98B parameters results in substantial improvements in the loss for all tracks on the test set, with the greatest improvements observed in sequence loss (Fig. 1D and fig. S11). The gap between unconditional and conditional negative log-likelihoods increases with scale. Conditioning on function keywords primarily constrains sequence at high masking rates, so although responsiveness to key word conditioning is observed at high mask rates, it is less apparent in the averaged negative log-likelihood (fig. S12). These gains in test loss lead to better representation learning (table S8 and fig. S8). In single sequence structure prediction, ESM3 98B surpasses ESMFold [0.880 versus 0.861 mean local distance difference test (LDDT) by the CAMEO test set; table S9). Generating sequences from the model without prompting (unconditional generation) produces high-quality proteins with a mean predicted LDDT (pLDDT) of 0.84 and a predicted template modeling score (pTM) of 0.52, which are diverse in both sequence (mean pairwise sequence identity 0.155) and structure (mean pairwise TM score 0.48), spanning the distribution of known proteins (Fig. 1E and fig. S14).

Our results show that scaling with language modeling, which is enabled by tokenization, efficient architectures, and masked token prediction, yields continued improvements in both representational and generative applications. This approach allows the model to build

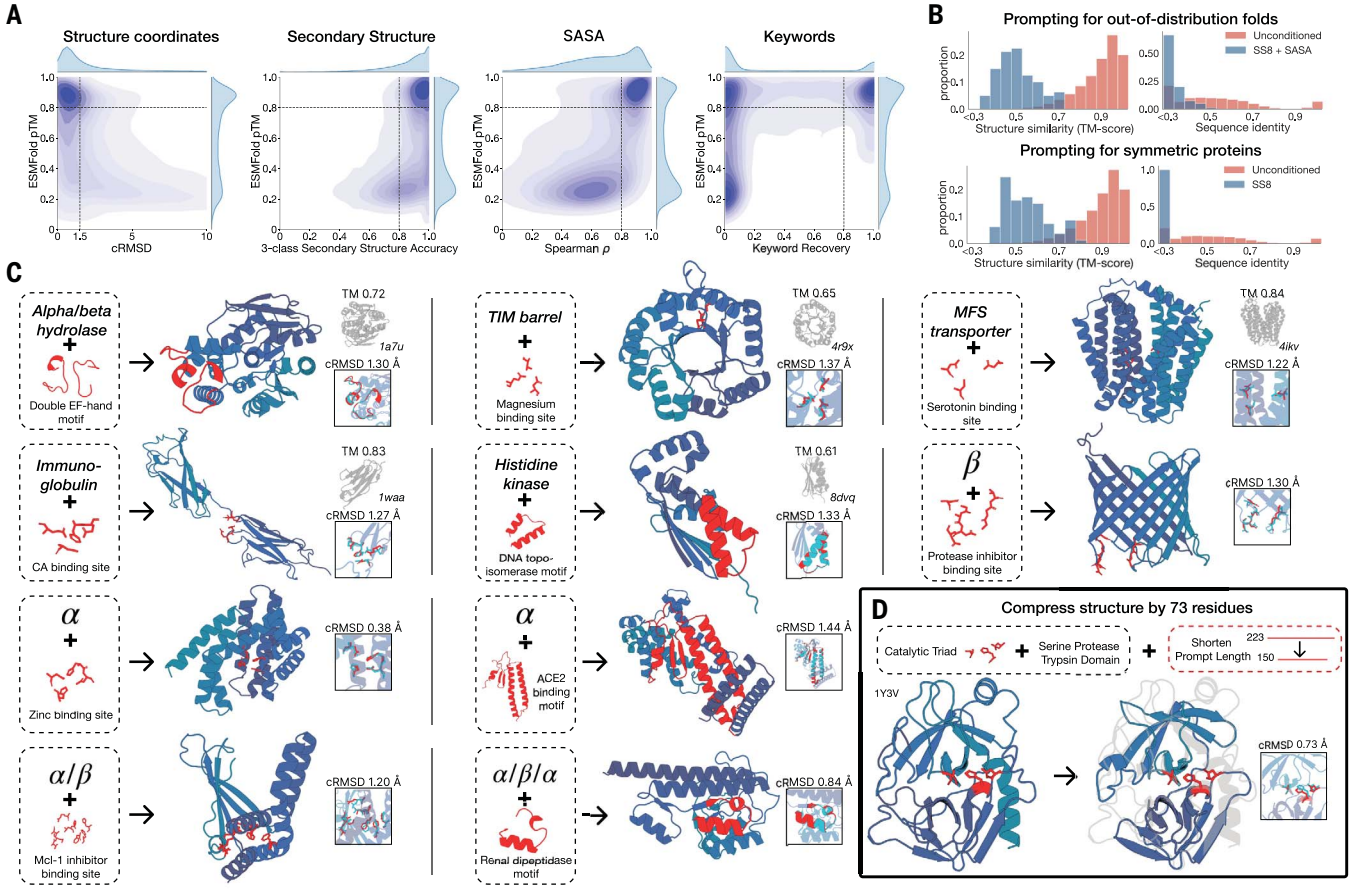

Fig. 2. Generative programming with ESM3. (**A**) ESM3 can follow prompts from each of its input tracks. Density of faithfulness to prompting for each of the tracks is shown. Generations achieve consistency with the prompt (backbone cRMSD, SS3 accuracy, SASA Spearman ρ, and key word recovery) and high structure prediction confidence (pTM). (**B**) ESM3 can be prompted to generate proteins that differ in structure (left) and sequence (right) from the training set and natural proteins. Prompted generations (blue) shift toward a more novel space versus unconditional generations (red) in response to prompts derived from out-of-distribution natural structures (top) and computationally designed symmetric proteins (bottom). (**C**) ESM3 generates creative solutions to a variety of combinations of complex prompts. We show compositions of atomic-level motifs with high-level instructions specified through key words or secondary structure prompts. Fidelity to the prompt is shown through similarity to a reference structure (for key word prompts) and all-atom RMSD (for motif prompts). Solutions differ from the scaffolds where the motif prompt was derived (median TM score 0.36 ± 0.14), and for many motifs (e.g., serotonin, calcium, protease inhibitor, and Mcl-1 inhibitor binding sites), we could find no significant similarity to other proteins that contain the same motif. (**D**) Example of especially creative behavior. ESM3 compresses a serine protease by 33% while maintaining the active site structure.

a shared multimodal representation space that is learned from the data rather than being explicitly hardcoded into its architecture. Given increasing compute and data, the model could learn an increasingly richer and more general feature space. In the following sections, we show that this approach achieves high fidelity for the controllable generation of proteins.

Programmable design with ESM3

We explored the ability of ESM3 to follow complex prompts with different compositions. ESM3 can be prompted with instructions from each of its input tracks: sequence, structure coordinates, SS8, SASA, and function key words. This allows prompts to be specified at multiple levels of abstraction, from atomic-level structure to high-level key words describing the function and fold topology.

We evaluated ESM3's ability to follow prompts in each of the tracks independently (Fig. 2A). A set of prompts are constructed for each of the tracks using a temporally held out test set of natural proteins (supplementary materials, section A.3.8). The resulting generations are evaluated using ESMFold for consistency with the prompt and confidence of structure prediction (pTM). We defined four consistency metrics for each track: (i) constrained site RMSD (cRMSD), the RMSD between the coordinates of the prompt, i.e., the positions of the backbone atoms, and the corresponding coordinates in the generation; (ii) SS3 accuracy, the fraction of residues where three-class secondary structure between the prompt and generations match; (iii) SASA Spearman ρ, the correlation between the SASA prompt and the corresponding region of the generation; and (iv) key word recovery, the fraction of prompt key words recovered by InterProScan (*40*). Across all tracks, the 7B parameter ESM3 finds solutions that follow the prompt and have structures that are confidently predicted by ESMFold (pTM > 0.8). Some mode switching is observed, including under key word prompting, where a fraction of the generations have confidently predicted structures that do not recover the key words.

Unconditional generations reflect the distribution of natural proteins. Because we observed that ESM3 can faithfully follow prompts, we reasoned that prompting could steer the model to generate proteins that differ from the training set and natural proteins. First, we tested the ability of the model to follow out-of-distribution prompts. We constructed a set of prompts combining SS8 and SASA from held-out structures (TM <0.7 to training set). Under these prompts, although the model continues to generate coherent globular structures (mean pTM 0.85 ± 0.03 under ESM3 7B; supplementary materials, section A.3.9), the distribution of similarities to the training set (as measured by TM score and sequence identity) shifts to be more novel (average sequence identity to nearest training set protein <20% and mean TM score 0.48 ± 0.09; Fig. 2B, top). To test the ability to generalize to structures beyond the distribution of natural proteins, we used secondary structure prompts derived from a dataset of artificial symmetric protein designs distinct from the natural proteins found in the training dataset (supplementary materials, section A.3.9). Similarly, ESM3 produces high-confidence generations (pTM > 0.8, pLDDT > 0.8) with low sequence and structure similarity to proteins in the training set (sequence identity <20% and TM score 0.52 ± 0.10; Fig. 2B, bottom), indicating that the model can be used to generate protein sequences and structures highly distinct from those that exist in nature.

ESM3 is able to follow complex prompts and has the ability to compose prompts from different tracks and at different levels of abstraction. To evaluate this ability, we prompted ESM3 with motifs that require solving for spatial coordination of individual atoms, including atoms participating in tertiary contacts between residues far apart in the sequence, such as catalytic centers and ligand-binding sites. We combined the atomic-level motif prompts with high-level prompts, either secondary structure prompts or key word prompts that specify the fold architecture. For each unique combination of atomic-level motif and high-level prompt, we generated sequences until there was a success (for atomic-level prompts, when all-atom RMSD <1.5 Å; for fold architecture key word prompts, when TM was >0.6 to a representative structure; for secondary structure prompts, when SS3 accuracy was >80%; and for all prompts, when pTM was >0.8 and pLDDT was >0.8 for the entire generated protein).

We found that ESM3 is able to solve a wide variety of such tasks (Fig. 2C). It does so without retrieving the motif's original scaffold (median TM score of 0.40 ± 0.10; supplementary materials, section A.3.10). In some cases, the scaffolds are transferred from existing proteins that have similar motifs (for example, the ESM3-designed alpha-helical scaffold for the zinc-binding motif has high similarity to Ni^{2+}-binding proteins, PDB: 5DQW, 5DQY; Fig. 2C, row 3, column 1). For many motifs (e.g., binding sites for serotonin, calcium, protease inhibitor, and Mcl-1 inhibitor), Foldseek (*41*) finds no other proteins that contain the same motif. In these cases, we observed that sometimes the motif has been grafted into entirely different

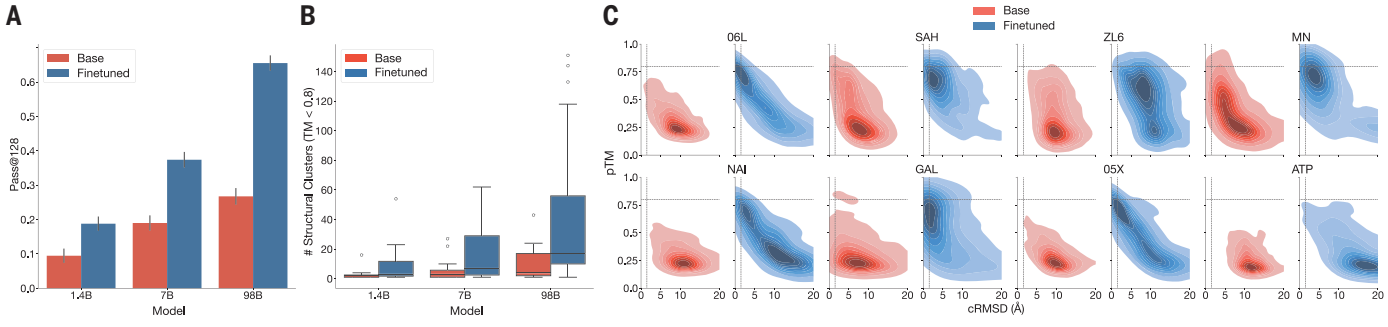

Fig. 3. The ability to solve complex tasks increases with scale through alignment. ESM3 was aligned to follow tertiary coordination prompts with a dataset of preference pairs constructed from prompted generations, where positive samples with good scores for desired properties (high pTM, low cRMSD) are paired with negative samples with worse scores. The preference tuning loss encourages the model to put higher likelihood on the positive samples. After training, models are evaluated by prompting with the backbone atomic coordinates of residues in tertiary contact. (**A**) Effect of fine-tuning on the fraction of tasks solved with 128 generations (Pass@128; error bars indicate 2 SDs). A large gap opens between the models with scale. The response to alignment shows a latent capability to solve complex tasks in the largest model. (**B**) Number of distinct solutions (clustered at TM > 0.8) generated for each tertiary motif. After fine-tuning, there are often many unique solutions for ligands where there are successes. (**C**) Densities of prompted generations are shown for the base model (left) and the aligned model (right) at the 98B scale for a number of randomly selected ligands. After alignment, the fidelity to the prompt (backbone cRMSD) and quality of generations (pTM) tends to improve substantially.

folds (e.g., a protease inhibitor binding site motif in a beta-barrel that is most similar to a membrane-bound copper transporter, PDB: 7PGE; Fig. 2C, row 3, column 3). At other times, the scaffold has low structural similarity to all known proteins in the PDB, ESMAtlas, and AlphaFold databases (maximum TM score <0.5; Fig. 2C, row 4, column 1), such as for an alpha/beta protein designed to scaffold the Mcl-1 inhibitor binding motif. Overall, the generated solutions have high designability, i.e., confident recovery of the original structure after inverse folding with ESM-IF1 (*42*) and refolding with ESMFold (median pTM 0.80 ± 0.08; scTM 0.96 ± 0.04; supplementary materials, section A.3.10).

Through experiments with prompt engineering, we have observed especially creative responses to prompts. Here, we highlight an example of protein compression (Fig. 2D). Starting from a natural trypsin (PDB 1Y3V), we prompted with the sequence and coordinates of the catalytic triad and functional key words describing trypsin but reduced the overall generation length by a third (from 223 to 150 residues). The ESM3 design maintains the coordination of the active site (all-atom RMSD 0.73 Å) and the overall fold with high designability (pTM 0.84, scTM mean 0.97, SD 0.006) despite the considerable reduction in sequence length and the fold only being specified by the function key word prompt (supplementary materials, section A.3.11).

These examples illustrate ESM3's ability to find creative solutions to prompts specified in any of its input tracks, individually or in combination. This capability enables a rational approach to protein design, providing control at various levels of abstraction, from high-level topology to atomic coordinates, using a generative model to bridge the gap between the prompt and biological complexity.

Biological alignment

Although we have observed meaningful increases in the performance of the base models with scale, larger models could have even greater latent capabilities that we did not observe. The base ESM3 models can be prompted to perform difficult tasks such as tertiary motif scaffolding and composition of prompts despite the fact that the models have not been explicitly optimized for these objectives. Because the properties that we evaluated generative outputs on, such as adherence to the prompt or the confidence of the scaffold, are only seen by the model indirectly during pretraining, aligning the model directly to the generative task with fine-tuning could elicit even greater capability differences with larger models.

We studied how the base models could be aligned (*43*, *44*) to generate proteins that satisfy challenging prompts. For each model, we constructed a dataset of backbone atomic coordinate prompts consisting of contiguous spans of residues and tertiary motifs (which also specify the identities of the contacting amino acids). We generated multiple protein sequences for each prompt and fold each of the sequences using ESM3, scoring for consistency with the prompt (backbone cRMSD) and structure prediction confidence (pTM). High-quality samples were paired with low-quality samples for the same prompt to construct a preference dataset (supplementary materials, section A.4). ESM3 was then fine-tuned with a preference optimization loss (*45*, *46*), which causes the model to put higher likelihood on the high-quality samples relative to the low-quality samples.

After aligning each of the base models, we evaluated their absolute performance and the shift in the distribution of generations. We focused on a series of challenging prompts that require coordination of the backbone atoms of residues in tertiary contact. We used ESMFold to evaluate the ability to generate high-quality scaffolds (pTM >0.8) that follow the prompt with high resolution (backbone cRMSD <1.5Å). We prompted each model with amino acid identities and backbone atomic coordinates from a held-out dataset of 46 ligand-binding motifs (supplementary materials, section A.4.5). For each motif, we created 1024 prompts by permuting the order of the residues, varying their position in the sequence, and varying the length of the sequence. A single protein was generated per prompt. The 1024 generations for each motif were used to construct an unbiased estimator of the fraction of tertiary coordination tasks solved after 128 generations (Pass@128; supplementary materials, section A.4.5).

Aligned models solve double the tertiary coordination tasks compared with base models (Fig. 3A). Although the base models show differences in the percentage of tasks solved (9.5% for 1.4B, 19.0% for 7B, 26.8% for 98B; Fig. 3A), a much larger capability difference was revealed through alignment (increasing from 9.5 to 18.8%, 19.0 to 37.4%, and 26.8 to 65.5% for the 1.4B, 7B, and 98B models, respectively). Preference-tuned models not only solve a greater proportion of tasks, but also find a greater number of solutions per task, as evaluated by the number of distinct structural clusters (TM >0.8) with backbone cRMSD <1.5Å and pTM >0.8 (Fig. 3B). A shift in the distribution of ESMFold pTM and backbone cRMSD for each ligand binding motif was observed (Fig. 3C and fig. S18). At the 98B scale, the fine-tuned model produced more distinct successful clusters than the base model on 37 of the 46 tested ligands, whereas the remaining nine ligands were not solved by either the base or aligned model, indicating that alignment almost universally improves the faithfulness to the prompt and confidence of the structure prediction for the generated proteins. These results represent state-of-the-art motif scaffolding performance (table S16). Compared with a supervised fine-tuning baseline, which only maximizes the likelihood of the positive examples, preference tuning leads to larger improvements at all scales (supplementary materials, section A.4.6).

Our experiments with alignment reveal a considerable difference in capabilities between model scales. The largest aligned model improves substantially relative to the base model before alignment and compared with the smaller models after alignment. Through alignment, the models learn to generalize from a small number of examples; the distribution of generations shifts to improve the quality of scaffolds and consistency with prompts, thus increasing the fraction of tasks solved and the number of distinct solutions.

Alignment requires the models to learn by example. The ability to identify the underlying properties that are illustrated by the fine-tuning examples, and to generalize those demonstrations to new tasks, implies that there is an internal representation of the properties that the fine-tuning accesses. This representation space is learned through the process of pretraining, in which the model is trained on proteins across evolution, suggesting that it reflects and contains the immense variety and complexity of protein biology. Such a representation space is likely to contain features that support the generalization of many biological properties. The greater responsiveness of larger models to alignment suggests that their internal representation space better approximates those underlying properties, which is evidence of a deep capability for transfer through the features learned in pretraining that improves with scale.

Generating a distant fluorescent protein

We sought to understand whether the base-pretrained ESM3 model has sufficient biological fidelity to generate functional proteins. We set out to create a functional GFP with low sequence similarity to existing ones. We chose the functionality of fluorescence because it is difficult to achieve, easy to measure, and one of the most beautiful mechanisms in nature.

Proteins in the GFP family are responsible for the fluorescence of jellyfish and the vivid colors of coral (*47*), and are unique in their ability to form a fluorescent chromophore without cofactors or substrates (*30*). This property allows the GFP sequence to be inserted into the genomes of other organisms to visibly label molecules, cellular structures, or processes, providing a foundational toolkit that has been broadly applied across the biosciences.

The GFP family has been the subject of decades of protein engineering efforts, but by far most of the known sequence diversity of GFPs

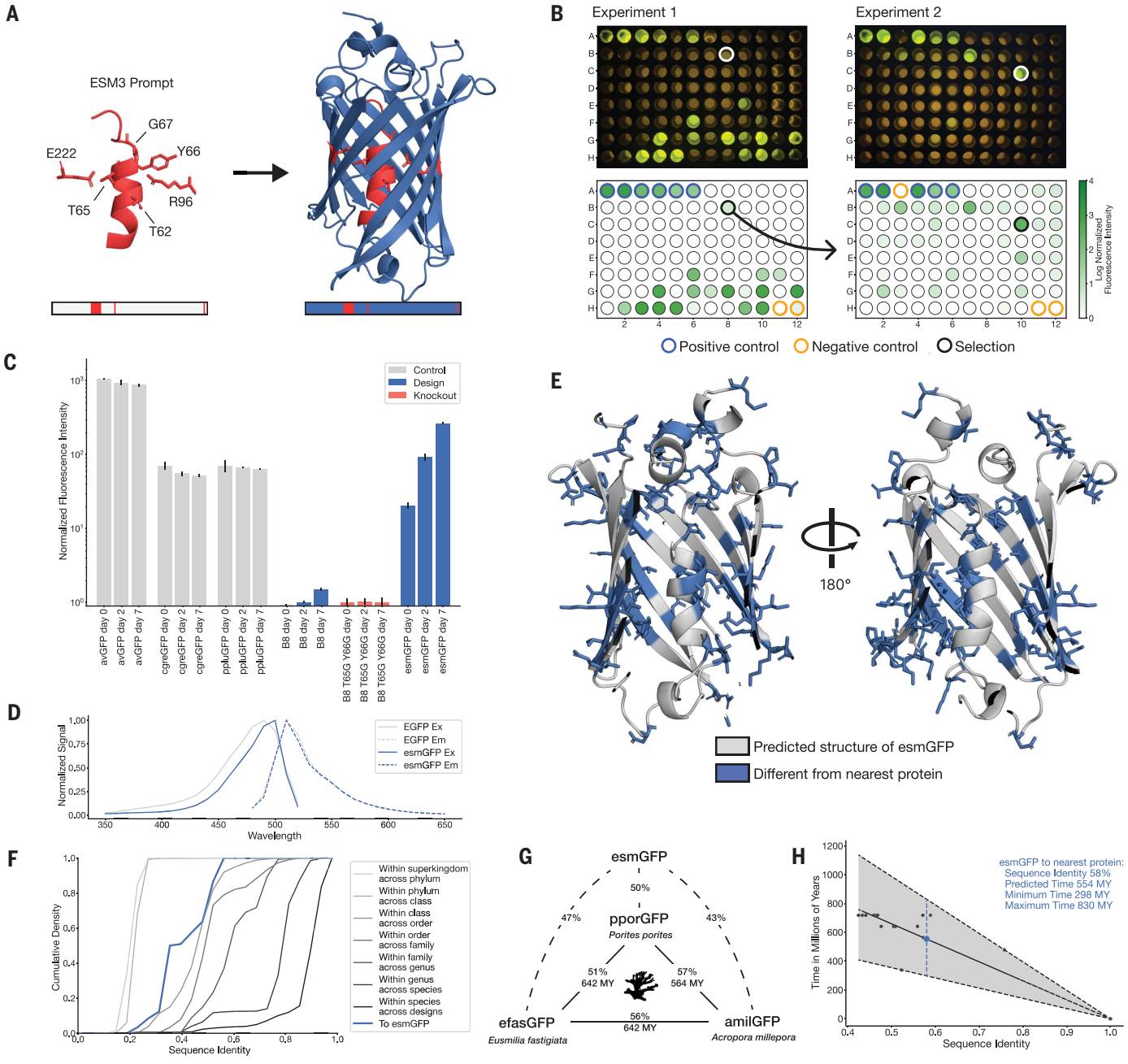

Fig. 4. Generating a distant fluorescent protein with a chain of thought. (**A**) We prompted ESM3 with the sequence and structure of residues required for forming and catalyzing the chromophore reaction, as well as the structure of part of the central alpha helix from a natural GFP (left). Through a chain of thought, ESM3 generates design candidates (right). (**B**) ESM3 found a bright GFP distant from other known GFPs in two experiments. We measured fluorescence in *Escherichia coli* lysate. Top row, photograph of plates. Bottom row, plate reader fluorescence quantification. Positive controls of known GFPs are marked with blue circles, and negative controls with no GFP sequence or no *E. coli* are marked with orange circles. In the first experiment (left), we expressed designs with a range of sequence identities. A notable design with low sequence identity (57%) to known fluorescent proteins appears in the well-labeled B8 (highlighted by a black circle at the bottom and a white circle at the top). We continued the chain of thought from the protein in B8 for the second experiment (right). A bright design appears in the well-labeled C10 (58% sequence identity to known fluorescent proteins; again, highlighted by a black circle at the bottom and a white circle at the top), which we designate esmGFP. (**C**) esmGFP exhibits fluorescence intensity similar to common GFPs. Normalized fluorescence is shown for a subset of proteins in experiment 2. (**D**) Excitation and emission spectra for esmGFP overlaid on the spectra of EGFP. (**E**) Two cutout views of the central alpha helix and the inside of the beta barrel of a predicted structure of esmGFP. The 96 mutations that esmGFP has relative to its nearest neighbor, tagRFP, are shown in blue. (**F**) Cumulative density of sequence identity between fluorescent proteins across taxa. esmGFP has the level of similarity to all other FPs that is typically found when comparing sequences across orders but within the same class. (**G**) Evolutionary distance by time in millions of years (MY) and sequence identities for three example anthozoan GFPs and esmGFP. (**H**) Estimator of evolutionary distance by time (MY) from GFP sequence identity. We estimate that esmGFP is >500 million years of natural evolution removed from the closest known protein.

has come from prospecting the natural world, because protein engineering efforts have for the most part explored only a few mutations starting from naturally fluorescent sequences. Rational design and mutagenesis have yielded GFP sequences with improved properties, such as higher brightness or stability or differently colored variants, that incorporated small numbers of mutations (typically five to 15 of the total 238 amino acid coding sequence). In a few cases, leveraging high-throughput experimentation and machine learning, scientists have been able to introduce up to 40 to 50 mutations (i.e., 80% sequence identity) while retaining fluorescence (*48*–*50*).

Generating an engineered GFP with considerable sequence distance from natural variants would require materialization of the complex biochemistry and physics that underlie its fluorescence. In all GFPs, an autocatalytic process forms the chromophore from three key amino acids in the core of the protein. The structure of GFP, a kinked central alpha helix surrounded by an 11-stranded beta barrel with inward-facing coordinating residues, enables this reaction (*51*). Once formed, the chromophore must not just absorb light but also emit it to be fluorescent. Light emission is highly sensitive to the local electronic environment of the chromophore. The fitness landscape of GFP reflects the precise configuration of both the active site and the surrounding tertiary interactions required to achieve its function, because a few random mutations are sufficient to reduce fluorescence to zero (*48*, *52*).

In an effort to generate GFP sequences, we directly prompted the base-pretrained 7B parameter ESM3 to generate a 229-residue protein conditioned on the positions Thr^{62}, Thr^{65}, Tyr^{66}, Gly^{67}, Arg^{96}, Glu^{222}, which are critical residues for generating the chromophore (Fig. 4A). We additionally conditioned on the structure of residues 58 through 71 from the experimental structure in 1QY3, which are known to be structurally important for the energetic favorability of chromophore formation (*53*). Specifically, sequence tokens, structure tokens, and atomic coordinates of the backbone are provided at the input, and generation begins from a nearly completely masked array of tokens corresponding to 229 residues, except for the token positions used for conditioning.

We generated designs using a chain-of-thought procedure as follows. The model first generates structure tokens, effectively creating a protein backbone. Backbones that have sufficiently good atomic coordination of the active site but differentiated overall structure from the 1QY3 backbone pass through a filter to the next step of the chain. We added the generated structure to the original prompt to generate a sequence conditioned on the new prompt. We then performed an iterative joint optimization, alternating between optimizing the sequence and the structure. We rejected chains of thought that lose atomic coordination of the active site (supplementary materials, section A.5.1). We drew a computational pool of tens of thousands of candidate GFP designs from the intermediate and final points in the iterative joint optimization stage of the generation protocol. We bucketed the designs by sequence similarity to known fluorescent proteins and filtered and ranked designs using a variety of metrics (supplementary materials, section A.5.1.5).

We performed a first experiment with 88 designs on a 96-well plate, evaluating the top generations in each sequence similarity bucket. Each generated protein was synthesized, expressed in *E. coli*, and measured for fluorescence activity at an excitation wavelength of 485 nm (Fig. 4B, left). We measured brightness similar to positive controls from a number of designs that have higher sequence identity with naturally occurring GFPs. We also identified a design in well B8 (highlighted in a black circle) with only 36% sequence identity to the 1QY3 sequence and 57% sequence identity to the nearest existing fluorescent protein, tagRFP. This design was 50× less bright than natural GFPs, and its chromophore matured over the course of a week, instead of in under a day, but it presents a signal of function in a part of sequence space that to our knowledge has not been found in nature or through protein engineering.

We continued the chain of thought starting from the sequence of the design in well B8 to generate a protein with improved brightness using the same iterative joint optimization and ranking procedure as above. We created a second 96-well plate of designs and, using the same plate reader assay, we found that a few designs in this cohort have a brightness in the range of GFPs found in nature. The best design, located in well C10 of the second plate (Fig. 4B, right), we designated as esmGFP.

We found that esmGFP exhibits brightness in the distribution of natural GFPs. We evaluated the fluorescence intensity at 0, 2, and 7 days of chromophore maturation and plotted these measurements for esmGFP, a replicate of B8, a chromophore knockout of B8, along with three natural GFPs: avGFP, cgreGFP, and ppluGFP (Fig. 4C). esmGFP takes longer to mature than the known GFPs that we measured but achieves a comparable brightness after 2 days. To validate that fluorescence was mediated by the intended Thr^{65} and Tyr^{66}, we showed that B8 and esmGFP variants in which these residues were mutated to glycine lost fluorescence activity (fig. S22).

Analysis of the excitation and emission spectra of esmGFP revealed that its peak excitation occurs at 496 nm, which is shifted 7 nm relative to the 489-nm peak for EGFP, but both proteins emit at a peak of 512 nm (Fig. 4D). The shapes of the spectra indicated a narrower full width at half maximum (FWHM) for the excitation spectrum of esmGFP (39 nm for esmGFP versus 56 nm for EGFP), whereas the FWHMs of their emission spectra were highly comparable (35 and 39 nm, respectively). Overall, esmGFP exhibits spectral properties consistent with known GFPs.

We next sought to understand how esmGFP compares with known proteins. A BLAST (*54*) search against the nonredundant protein sequences database and an MMseqs (*55*) search of ESM3's training set reported the same top hit, tagRFP, which was also the nearest neighbor to B8, with 58% sequence identity representing 96 mutations throughout the sequence. tagRFP is a designed variant, and the closest wild-type sequence to esmGFP from the natural world is eqFP578, a red fluorescent protein that differs from esmGFP by 107 sequence positions (53% identity). Sequence differences between esmGFP and tagRFP occur throughout the structure (Fig. 4E), with 22 mutations occurring in the protein's interior, which is known to be highly sensitive to mutations due to chromophore proximity and a high density of interactions (*56*).

Examination of a sequence alignment of 648 natural and designed GFP-like fluorescent proteins revealed that esmGFP has the level of similarity to all other FPs that is typically found when comparing sequences across taxonomic orders but within the same taxonomic class (Fig. 4F). For example, the difference between esmGFP and other FPs is similar to the level of difference between FPs belonging to the orders of Scleractinia (stony corals) and Actiniaria (sea anemones), both of which belong to the larger class Anthozoa of marine invertebrates (Fig. 4G). The closest FPs to esmGFP come from the Anthozoa class (corals and anemones; average sequence identity 51.4%), but esmGFP also shares some sequence identity with FPs from the Hydrozoa (jellyfish), in which avGFP was discovered (average sequence identity 33.4%; fig. S23).

We can draw insight from evolutionary biology on the amount of time that it would take for a protein with similar sequence identity to arise through natural evolution. In Fig. 4G, we show esmGFP alongside three anthozoan GFPs. We used a time-calibrated phylogenetic analysis of the anthozoans (*57*) that estimated the millions of years ago (MYA) to last common ancestors to estimate evolutionary time between each pair of these species. Using a larger dataset of six anthozoan GFPs and species for which we have accurate MYA to last common ancestors and GFP sequence identities, we constructed a simple estimator that correlates sequence identity between FPs to MY of evolutionary time between the species (Fig. 4H) to calibrate against natural evolution. On the basis of this analysis, we estimate that esmGFP

represents an equivalent of >500 million years of evolution from the closest protein that has been found in nature.

Discussion

We have found that language models can reach a design space of proteins that is distant from the space explored by natural evolution and can generate functional proteins that would take evolution hundreds of millions of years to discover. Protein language models do not explicitly work within the physical constraints of evolution, but instead can implicitly construct a model of the multitude of potential paths that evolution could have followed.

Proteins can be seen as existing within an organized space where each protein is neighbored by every other protein that is one mutational event away (*58*). The structure of evolution appears as a network within this space, connecting all proteins by the paths that evolution can take between them. The paths that evolution can follow are the ones by which each protein transforms into the next without the collective loss of function of the system of which it is a part.

It is in this space that a language model sees proteins. It sees the data of proteins as filling this space, densely in some regions and sparsely in others, revealing the parts that are accessible to evolution. Because the next token is generated by evolution, it follows that to solve the training task of predicting the next token, a language model must predict how evolution can move through the space of possible proteins.

Simulations are computational representations of reality. In that sense, a language model that can predict possible outcomes of evolution can be said to be a simulator of it. ESM3 is an emergent simulator that has learned from solving a token prediction task on data generated by evolution. It has been theorized that neural networks discover the underlying structure of the data that they are trained to predict (*59*, *60*). In this way, solving the token prediction task would require the model to learn the deep structure that determines which steps evolution can take, i.e., the fundamental biology of proteins.

In ESM3's generation of a fluorescent protein, it is the first chain of thought to B8 that is the most intriguing. At 96 mutations to B8's closest neighbor, there are $\binom{229}{96} \times 19^{96}$ possible proteins, of which only a vanishingly small fraction can have function because fluorescence falls off sharply even after just a few random mutations. The existence of C10 and other bright designs in the neighborhood of B8 confirms that in the first chain of thought to B8, ESM3 found a part of the space of proteins that, although unexplored by nature, is dense with fluorescent proteins.

REFERENCES AND NOTES

1. UniProt Consortium, *Nucleic Acids Res.* **43**, D204–D212 (2015).
2. I. V. Grigoriev *et al*., *Nucleic Acids Res.* **40**, D26–D32 (2012).
3. A. L. Mitchell *et al*., *Nucleic Acids Res.* **48**, D570–D578 (2020).
4. M. Varadi *et al*., *Nucleic Acids Res.* **52**, D368–D375 (2024).
5. Z. Lin *et al*., *Science* **379**, 1123–1130 (2023).
6. E. C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi, G. M. Church, *Nat. Methods* **16**, 1315–1322 (2019).
7. M. Heinzinger *et al*., *BMC Bioinformatics* **20**, 723 (2019).
8. A. Rives *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2016239118 (2021).
9. A. Madani *et al*., *Nat. Biotechnol.* **41**, 1099–1106 (2023).
10. N. Ferruz, S. Schmidt, B. Höcker, *Nat. Commun.* **13**, 4348 (2022).
11. R. Verkuil *et al*., Language models generalize beyond natural proteins. bioRxiv 521521 [Preprint] (2022); https://doi.org/10.1101/2022.12.21.521521.
12. A. Elnaggar *et al*., *IEEE Trans. Pattern Anal. Mach. Intell.* **14**, 7112–7127 (2022).
13. D. Hesslow, N. Zanichelli, P. Notin, I. Poli, D. Marks, RITA: A study on scaling up generative protein sequence models. arXiv: 2205.05789 [q-bio.QM] (2022).
14. E. Nijkamp, J. A. Ruffolo, E. N. Weinstein, N. Naik, A. Madani, *Cell Syst.* **14**, 968–978.e3 (2023).
15. S. Alamdari *et al*., Protein generation with evolutionary diffusion: sequence is all you need. bioRxiv 556673 [Preprint] (2023); https://doi.org/10.1101/2023.09.11.556673.
16. M. Heinzinger *et al*., Bilingual language model for protein sequence and structure. bioRxiv 550085 [Preprint] (2024); https://doi.org/10.1101/2023.07.23.550085.
17. J. Su *et al*., SaProt: Protein language modeling with structureaware vocabulary. bioRxiv 560349 [Preprint] (2023); https://doi.org/10.1101/2023.10.01.560349.
18. J. Meier *et al*., Language models enable zero-shot prediction of the effects of mutations on protein function. bioRxiv 450648 [Preprint] (2021); https://doi.org/10.1101/2021.07.09.450648.
19. J. Vig *et al*., BERTology meets biology: Interpreting attention in protein language models. arXiv:2006.15222 [cs.CL] (2020).
20. R. Rao, J. Meier, T. Sercu, S. Ovchinnikov, A. Rives, Transformer protein language models are unsupervised structure learners. bioRxiv 422761 [Preprint] (2021); https://doi.org/10.1101/2020.12.15.422761.
21. B. Chen *et al*., xTrimoPGLM: Unified 100B-scale pre-trained transformer for deciphering the language of protein. bioRxiv 547496 [Preprint] (2023); https://doi.org/10.1101/2023.07.05.547496.
22. J. Kaplan *et al*., Scaling laws for neural language models. arXiv:2001.08361 [cs.LG] (2020).
23. T. B. Brown *et al*., Language models are few-shot learners. arXiv:2005.14165 [cs.CL] (2020).
24. J. Hoffmann *et al*., Training compute-optimal large language models. arXiv:2203.15556 [cs.CL] (2022).
25. J. Abramson *et al*., *Nature* **630**, 493–500 (2024).
26. J. L. Watson *et al*., *Nature* **620**, 1089–1100 (2023).
27. J. B. Ingraham *et al*., *Nature* **623**, 1070–1078 (2023).
28. Y. Lin, M. Lee, Z. Zhang, M. AlQuraishi, Out of many, one: Designing and scaffolding proteins at the scale of the structural universe with Genie 2, May 2024. arXiv:2405.15489 [q-bio.BM] (2024).
29. O. Shimomura, F. H. Johnson, Y. Saiga, *J. Cell. Comp. Physiol.* **59**, 223–239 (1962).
30. R. Y. Tsien, *Annu. Rev. Biochem.* **67**, 509–544 (1998).
31. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 [cs.CL] (2018).
32. H. Chang, H. Zhang, L. Jiang, C. Liu, W. T. Freeman, Maskgit: Masked generative image transformer. arXiv:2202.04200 [cs.CV] (2022).
33. B. Uria, I. Murray, H. Larochelle, A deep and tractable density estimator. arXiv:1310.1757 [stat.ML] (2014).
34. J. Austin, D. D. Johnson, J. Ho, D. Tarlow, R. van den Berg. Structured denoising diffusion models in discrete state-spaces. arXiv:2107.03006 [cs.LG] (2023).
35. A. van den Oord, O. Vinyals, K. Kavukcuoglu, Neural discrete representation learning. arXiv:1711.00937 [cs.LG] (2017).
36. B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu; UniProt Consortium, *Bioinformatics* **31**, 926–932 (2015).
37. L. Richardson *et al*., *Nucleic Acids Res.* **51**, D753–D759 (2023).
38. T. H. Olsen, F. Boyles, C. M. Deane, *Protein Sci.* **31**, 141–146 (2022).
39. S. K. Burley *et al*., *Nucleic Acids Res.* **47**, D464–D474 (2019).
40. T. Paysan-Lafosse *et al*., *Nucleic Acids Res.* **51**, D418–D427 (2023).
41. M. van Kempen *et al*., Foldseek: fast and accurate protein structure search. bioRxiv 479398 [Preprint] (2022); https://doi.org/10.1101/2022.02.07.479398.
42. C. Hsu *et al*., Learning inverse folding from millions of predicted structures. bioRxiv 487779 [Preprint] (2022); https://doi.org/10.1101/2022.04.10.487779.
43. D. M. Ziegler *et al*., Fine-tuning language models from human preferences. arXiv:1909.08593 [cs.CL] (2019).
44. L. Ouyang *et al*., Training language models to follow instructions with human feedback. arXiv:2203.02155 [cs.CL] (2022).
45. R. Rafailov *et al*., Direct preference optimization: Your language model is secretly a reward model. arXiv:2305.18290 [cs.LG] (2023).
46. R. Y. Pang *et al*., Iterative reasoning preference optimization. arXiv:2404.19733 [cs.CL] (2024).
47. Y. A. Labas *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **99**, 4256–4261 (2002).
48. L. Gonzalez Somermeyer *et al*., *eLife* **11**, e75842 (2022).
49. S. Biswas *et al*., Toward machine-guided design of proteins. bioRxiv 337154 [Preprint] (2018); https://doi.org/10.1101/337154.
50. S. Biswas, G. Khimulya, E. C. Alley, K. M. Esvelt, G. M. Church, *Nat. Methods* **18**, 389–396 (2021).
51. M. Ormö *et al*., *Science* **273**, 1392–1395 (1996).
52. K. S. Sarkisyan *et al*., *Nature* **533**, 397–401 (2016).
53. D. P. Barondeau, C. D. Putnam, C. J. Kassmann, J. A. Tainer, E. D. Getzoff, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 12111–12116 (2003).
54. C. Camacho *et al*., *BMC Bioinformatics* **10**, 421 (2009).
55. M. Steinegger, J. Söding, *Nat. Biotechnol.* **35**, 1026–1028 (2017).
56. J. Y. Weinstein *et al*., *Nat. Commun.* **14**, 2890 (2023).
57. A. M. Quattrini *et al*., *Nat. Ecol. Evol.* **4**, 1531–1538 (2020).
58. J. M. Smith, *Nature* **225**, 563–564 (1970).
59. U. Kamath, J. Liu, J. Whitaker, in *The Philosophy of Artificial Intelligence*, G. E. Hinton, J. L. McClelland, D. E. Rumelhart, Eds. (Springer, 1986); pp. 203–261.
60. N. Tishby, F. C. Pereira, W. Bialek, The information bottleneck method. arXiv:physics/0004057 [physics.data-an] (1999).
61. EvolutionaryScale, "Esm3 source code," Zenodo (2024); https://doi.org/10.5281/zenodo.

ACKNOWLEDGMENTS

We thank E. Schreiter, K. Svoboda, and S. Turaga for feedback on the properties of esmGFP; I. Holmes for feedback on the evolutionary analysis of esmGFP; M. Iskander, V. Kher, and the Andromeda cluster team for support on compute infrastructure; A. Pawluk for assistance with manuscript preparation; the experts who provided feedback on our approach to responsible development; and the experts who participated in the review of the risks and benefits of releasing ESM3-open. Y.A.K. was an intern with EvolutionaryScale during the course of this study. **Funding:** This research was funded by EvolutionaryScale. **Author contributions:** Data: H.A., Z.L., R.R., A.R., T.S., N.T., R.V.; Pretraining: H.A., S.C., J.D., T.H., Z.L., D.O., R.R., A.R., T.S., I.S., R.V., M.W.; Posttraining: H.A., S.C., A.D., J.G., T.H., D.O., R.R., A.R., M.W.; Evaluation and analysis: R.B., J.D., A.D., T.H., Y.A.K., C.K., Z.L., R.S.M., A.R., N.J.S; Open model and responsible development: J.G., I.S., N.J.S., T.S., R.S.M., Z.L., R.R., A.R., N.T.; API and deployment: J.G., C.M., R.S.M., Z.L., T.S.; GFP computation: S.C., T.H., N.J.S., A.R., R.V.; GFP experimental validation: L.J.B., M.N., P.D.H., Y.A.K., N.J.S., N.T., V.Q.T.; Writing: S.C., T.H., R.R., A.R.; N.J.S.; Supplementary materials: H.A., R.B., L.J.B., S.C., J.D., A.D., T.H., C.K., Z.L., R.S.M., D.O., R.R., A.R., N.J.S., T.S., I.S., N.T., V.Q.T., R.V., M.W.; Overall scientific direction: A.R. **Competing interests:** H.A., R.B., S.C., J.D., A.D., J.G., T.H., C.K., Z.L., R.S.M., C.M., D.O., R.R., A.R., N.J.S., T.S., I.S., N.T., R.V., and M.W. are employees of EvolutionaryScale, PBC. S.C., A.R., and T.S. are officers and members of the board of directors of EvolutionaryScale. P.D.H. is a cofounder of Stylus Medicine, Circle Labs, and Spotlight

Therapeutics; serves on the board of directors at Stylus Medicine; is a board observer at EvolutionaryScale, Circle Labs, and Spotlight Therapeutics; is a scientific advisory board member at Arbor Biosciences and Veda Bio; and is an advisor to NFDG, Varda Space, and Vial Health. The remaining authors declare no competing interests. Patents have been filed related to aspects of this work. **Data and materials availability:** Weights and code for ESM3-open are provided for academic research use at https://github.com/evolutionaryscale/esm and are permanently archived at Zenodo (*61*). The ESM3-open model was reviewed by a committee of technical experts who found that the benefits of releasing the model greatly outweighed any potential risks. ESM3 models are available through API with a free access tier for academic research. The sequence of esmGFP (along with the other GFPs generated for this work) is committed to the public domain. Plasmids for esmGFP-C10 and esmGFP-B8 have been deposited with Addgene.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.ads0018
Materials and Methods
Figs. S1 to S24
Tables S1 to S17
References (*62*–*117*)
MDAR Reproducibility Checklist

Submitted 24 July 2024; accepted 7 January 2025
Published online 16 January 2025
10.1126/science.ads0018