nature biotechnology

Article

https://doi.org/10.1038/s41587-022-01618-2

Large language models generate functional protein sequences across diverse families

Received: 12 July 2022

Accepted: 17 November 2022

Published online: 26 January 2023

Check for updates

Ali Madani[1,2] ✉, Ben Krause[1,10], Eric R. Greene[3,10], Subu Subramanian[4,5], Benjamin P. Mohr[6], James M. Holton[7,8,9], Jose Luis Olmos Jr.[3], Caiming Xiong[1], Zachary Z. Sun[6], Richard Socher[1], James S. Fraser[3] & Nikhil Naik[1] ✉

Deep-learning language models have shown promise in various biotechnological applications, including protein design and engineering. Here we describe ProGen, a language model that can generate protein sequences with a predictable function across large protein families, akin to generating grammatically and semantically correct natural language sentences on diverse topics. The model was trained on 280 million protein sequences from >19,000 families and is augmented with control tags specifying protein properties. ProGen can be further fine-tuned to curated sequences and tags to improve controllable generation performance of proteins from families with sufficient homologous samples. Artificial proteins fine-tuned to five distinct lysozyme families showed similar catalytic efficiencies as natural lysozymes, with sequence identity to natural proteins as low as 31.4%. ProGen is readily adapted to diverse protein families, as we demonstrate with chorismate mutase and malate dehydrogenase.

Traditional methods for protein engineering perform iterative mutagenesis and selection of natural protein sequences to identify proteins with desired functional and structural properties. By contrast, rational or de novo protein design methods aim to improve the efficiency and precision of creating novel proteins with desired properties. Structure-based de novo design methods[1–5] employ simulations grounded in biophysical principles, whereas coevolutionary methods[6–10] build statistical models from evolutionary sequence data to specify novel sequences with desired function or stability. Both structural and coevolutionary approaches are not without limitations. Structural methods rely on scarce experimental structure data and difficult or intractable biophysical simulations[3,11]. Coevolutionary statistical models are tailored to specific protein families, frequently rely on multiple sequence alignments, and do not operate well in space outside of the defined multiple sequence alignment[11]. Recently, deep neural networks have shown promise as generative and discriminative models for protein science and engineering[12–20]. Their ability to learn complex representations could be essential to effectively exploit an exponentially growing source of diverse and relatively unannotated protein data—public databases containing millions of raw unaligned protein sequences[21–23].

Inspired by the success of deep-learning-based natural language models trained on large text corpora that generate realistic text with varied topics and sentiments[24–28], we developed ProGen, a protein language model trained on millions of raw protein sequences that generates artificial proteins across multiple families and functions. While prior work has shown that natural-language-inspired statistical representations of proteins are useful for protein informatics tasks such as stability prediction, remote homology detection and secondary structure prediction[11,29–31], we show that the latest advances

[1]Salesforce Research, Palo Alto, CA, USA. [2]Profluent Bio, San Francisco, CA, USA. [3]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA, USA. [4]Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA, USA. [5]Howard Hughes Medical Institute, University of California, Berkeley, Berkeley, CA, USA. [6]Tierra Biosciences, San Leandro, CA, USA. [7]Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [8]Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory, Menlo Park, CA, USA. [9]Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, CA, USA. [10]These authors contributed equally: Ben Krause and Eric R. Greene. ✉e-mail: ali@madani.ai; nnaik@salesforce.com

https://doi.org/10.1038/s41587-022-01618-2

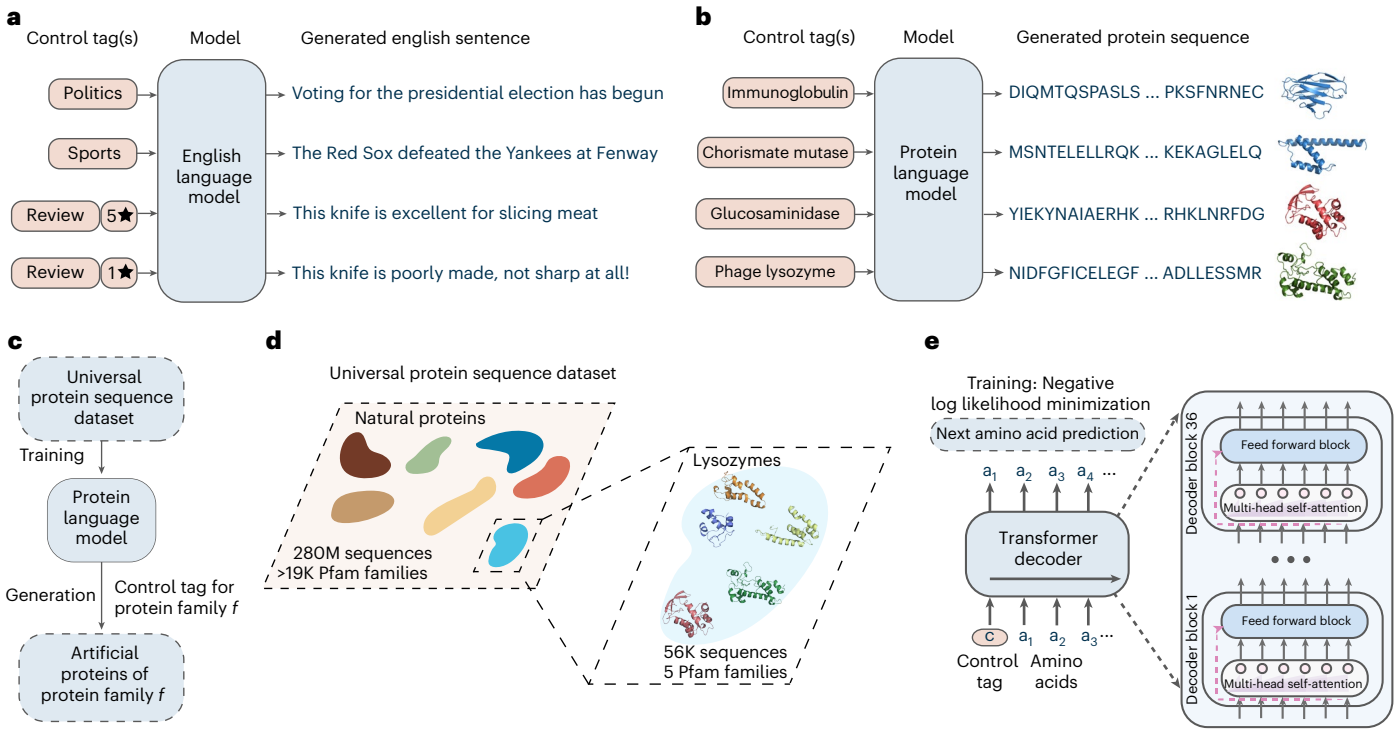

Fig. 1 | Artificial protein generation with conditional language modeling. **a**, Conditional language models are deep neural networks that can generate semantically and grammatically correct, yet novel and diverse natural language text, steerable using input control tags that govern style, topic and other entities. **b**,**c**, Analogous to natural language models, we develop ProGen, a conditional protein language model (**b**) that generates diverse artificial protein sequences across protein families based on input control tags (**c**). **d**, ProGen is trained using a large, universal protein sequence dataset of 280 million naturally evolved proteins from thousands of families, of which five diverse lysozyme families are experimentally characterized in this study. **e**, ProGen is a 1.2-billion-parameter neural network that is based on the Transformer architecture, which uses a self-attention mechanism for modeling comprehensive residue–residue interactions. ProGen is trained to generate artificial sequences by minimizing the loss over the next amino acid prediction problem on the universal protein sequence dataset.

in deep-learning-based language modeling can be adopted to generate artificial protein sequences, from scratch, that function as well as natural proteins.

ProGen is iteratively optimized by learning to predict the probability of the next amino acid given the past amino acids in a raw sequence, with no explicit structural information or pairwise coevolutionary assumptions. Trained in this unsupervised manner from a large, varied protein sequence database (Supplementary Table 1), ProGen learns a universal, domain-independent representation of proteins that subsumes local and global structure motifs, analogous to natural language models learning semantic and grammatical rules. After training, ProGen can be prompted to generate full-length protein sequences for any protein family from scratch, with a varying degree of similarity to natural proteins. In the common case where some sequence data from a protein family is available, we can use the technique of fine tuning pretrained language models[32–35] with family-specific sequences to further improve the ability of ProGen to capture the distribution of local sequence neighborhoods corresponding to the protein family.

ProGen is a 1.2-billion-parameter neural network trained using a publicly available dataset of 280 million protein sequences. A key component of ProGen is conditional generation[28,36–38], that is, sequence generation controlled by property tags (for example, Protein Family: Pfam ID PF16754, Pesticin) provided as input to the language model. In the case of natural language, these control tags may be style, topics, dates and other entities (Fig. 1a). For proteins, the control tags are properties such as protein family, biological process and molecular function, which are available for a large fraction of sequences in public protein databases (Fig. 1b and Supplementary Fig. 1).

Results

We experimentally evaluated the ability of ProGen to generate functional artificial amino acid sequences by testing its generations across five distinct protein families from the lysozyme clan[23,39] (Supplementary Table 2). The protein families contain substantial sequence diversity (Supplementary Table 3) with average sequence length varying between 84–167 across families. The sequences also show large structural diversity and multiple structural folds (Supplementary Fig. 2). As a whole, this represents a challenging design space for a model that is not constrained in generation to local sequence neighborhoods near known functional wild types and also not provided with structural priors. We collected a dataset of 55,948 sequences from these five families from Pfam and UniprotKB sources for obtaining positive controls and for fine tuning[32–35] ProGen.

After fine tuning ProGen using the curated lysozyme dataset, we generated one million artificial sequences using ProGen by providing the Pfam ID for each family as a control tag. Our artificial lysozymes span the sequence landscape of natural lysozymes (Fig. 2a) across five families that contain diverse protein folds, active site architectures and enzymatic mechanisms[40,41]. As our model can generate full-length artificial sequences within milliseconds, a large database can be created to expand the plausible sequence diversity beyond natural libraries (Supplementary Table 3). Although artificial sequences may diverge from natural sequences purely from a sequence identity calculation, (Fig. 2b and Supplementary Fig. 3), they demonstrate similar residue position entropies when forming separate multiple sequence alignments of natural and artificial proteins within each family (Fig. 2c). This indicates that the model has captured

https://doi.org/10.1038/s41587-022-01618-2

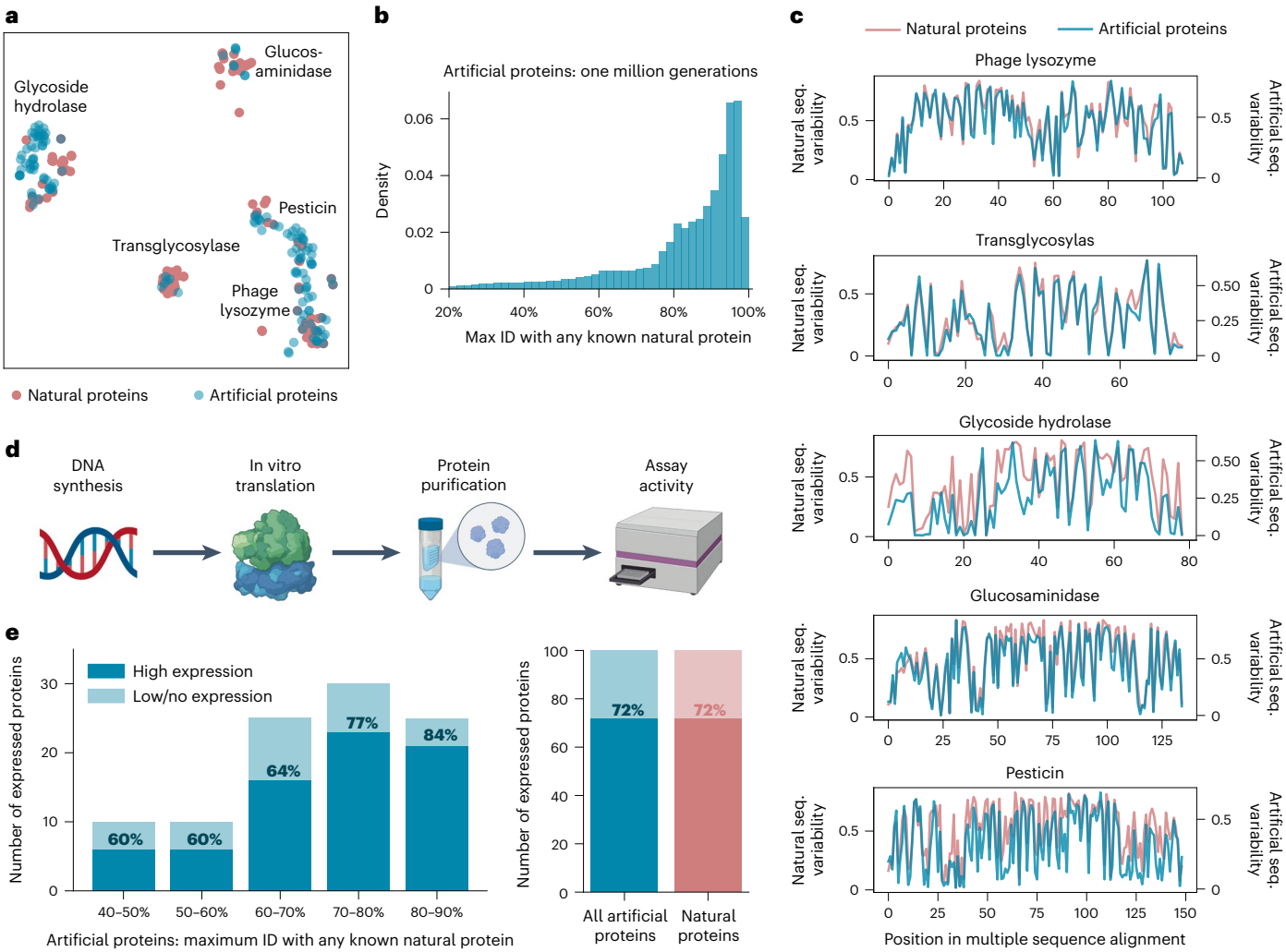

Fig. 2 | Generated artificial antibacterial proteins are diverse and express well in our experimental system. a, When analyzed using *t*-distributed stochastic neighbor embedding (*t*-SNE) as a dimensionality reduction technique for visualization purposes, artificial sequences from our model are shown to span the landscape of natural proteins from five lysozyme families. Each point represents a natural or generated sequence embedded in a two-dimensional *t*-SNE space. **b**, With sufficient sampling, ProGen can generate sequences that are highly dissimilar from natural proteins. Max ID measures the maximum identity of an artificial protein with any publicly available natural protein. **c**, Artificial proteins maintain similar evolutionary conservation patterns as natural proteins across families. Plots demonstrate the variability at each aligned position for a library of proteins. Conserved positions are represented as curve dips. seq., sequence. **d**, From our generated proteins, we select one hundred proteins for synthesis and characterization in our experimental setup. **e**, Artificial proteins express well even with increasing dissimilarity from nature (40–50% max ID) and yield comparable expression quality to one hundred representative natural proteins.

evolutionary conservation patterns without training on explicit alignment information such as with Potts models[42], as implemented in direct coupling analysis[7,43–46].

To experimentally evaluate ProGen performance across a range of sequence divergences from natural proteins, we selected one hundred sequences filtered on the basis of generation quality and diversity to natural sequences, measured as top-hit identities ('max ID') to any protein in our training dataset containing 280 million proteins, which is primarily composed of UniParc[21] (Supplementary Fig. 4). Our selected proteins included 100 artificial sequences (Supplementary Table 2), with a minimum of 8 proteins from each protein family. The average sequence length for artificial proteins varies between 93–179 across families, comparable to natural lysozymes in our curated dataset from Pfam. Artificial proteins included specific amino acids and pairwise interactions never before observed in aligned positions in lysozyme family-specific alignments (Supplementary Tables 4 and 5). We also selected a positive control group from the 55,948 curated lysozyme sequences. We clustered the natural sequences with MMseqs2[47] and chose roughly 20 cluster-representative sequences from each of the five families.

To evaluate function, full-length genes were synthesized and purified via cell-free protein synthesis and affinity chromatography. In the positive control set of 100 natural proteins, 72% were well expressed as measured by chromatography peaks and band visualization. The ProGen-generated proteins express equally well (72/100 total) across all bins of sequence identity to any known natural protein (max ID 40–90%; Fig. 2e). In addition, we designed artificial proteins using bmDCA[7], a statistical model that is based on direct coupling analysis, which explicitly approximates first and second-order residue dependencies. Starting from their publicly available code, we tried to make the bmDCA model converge on the same sequences as ProGen and using additional alignment information and searched over a wide range of hyperparameters. bmDCA was unable to fit three out of the five lysozyme families, and exhibited 60% detectable expression (30/50 proteins) for the remaining two protein families. These results indicate that ProGen can generate artificial proteins that are structurally well

folded for proper expression as compared to a batch of natural proteins, even when sequence alignment size and quality limit the success of alternative approaches.

Next we examined activity on the basis of quench release of fluorescein-labeled *Micrococcus lysodeikticus* cell wall (Molecular Probes EnzChek Lysozyme kit) using 90 randomly chosen proteins out of each expressed set of 100. Proteins were prepared in 96-well plate format to extract fluorescence curves over time (Fig. 3a). Hen egg white lysozyme (HEWL), a naturally evolved exemplar protein, was measured as positive control, in addition to ubiquitin as negative control. Proteins that generated fluorescence one standard deviation above the maximum fluorescence of any negative control were considered functional. Among our artificial proteins, 73% (66/90) were functional and exhibited high levels of functionality across families (Fig. 3c). The representative natural proteins exhibited similar levels of functionality with 59% (53/90) of total proteins considered functional. None of the bmDCA artificial proteins exhibited a detectable level of functionality (Supplementary Fig. 5), which may be due to convergence, sampling, or other specific model run issues further highlighting the versatility of ProGen providing a potentially more robust alternative. These results indicate that ProGen generates protein sequences that not only can express well but also maintain enzymatic function for diverse sequence landscapes across protein families.

In addition to a binary value for functionality, we calculated a relative activity score with respect to HEWL for the in vitro assay. Our artificial proteins match activity levels of natural proteins even at lower levels of sequence identity to any known natural protein, (Fig. 3b and Supplementary Fig. 6). Notably a small number of proteins, both within the natural and artificial proteins, were within an order of magnitude of HEWL, which was substantially more active than all negative controls. These highly active outliers demonstrate the potential for our model to generate sequences that may rival natural proteins that have been highly optimized through evolutionary pressures.

From the 100 artificial proteins, we selected five proteins that spanned a wide range of max IDs (48–89%) to recombinantly express in *Escherichia coli*. Of these, only one, L008, generated no detectable expression (Supplementary Fig. 7). Two (L013 and L038) expressed robustly to inclusion bodies and were not pursued further. Two proteins, L056 (max ID 69.6%) and L070 (max ID 89.2%) expressed well and incurred bactericidal activities towards the E. coli BL21(DE3) strain used during overnight induction at 16 °C. Spent medium harbored enzymatic activity, therefore, enzymes were purified from this material.

While both enzymes purified as monomers at the expected size by size-exclusion chromatography, we also observed a defined later eluting (apparent lower molecular weight, likely owing to binding to the dextran component of the column) species for each enzyme that corresponded to full-length enzyme by SDS-PAGE (Supplementary Fig. 7). The K_M values of both monomers were too weak to be measured using a heterogeneous, fluorescein-labeled *M. lysodeikticus* cell wall substrate (Molecular Probes EnzChek Lysozyme kit); however, both monomers were active using a pseudo-first-order kinetic assay (Supplementary Fig. 8). By contrast, we could readily measure the K_M values for the purified apparent lower molecular weight species, where both L056 and L070 were highly active and had comparable Michaelis–Menten parameters to HEWL (Fig. 3d). Taken together, L056 and L070 harbor potent catalytic activity and bactericidal capabilities that are comparable to HEWL, while diverging from their nearest known natural sequence by 53 and 18 amino acids, respectively. We also found that there is no bias to location or structural element to the mutations that diverge L056 and L070 from their respective nearest sequence homolog in nature. Differing residues are instead uniformly distributed. Some mutations are even found within the active site cleft and within regions that influence conformational state (for L056). Despite having comparable enzymatic activities, L070 and L056 only share 17.9% sequence ID. In sum, these results demonstrate that ProGen can generate artificial proteins with near native activity.

Next, we examined the structural divergence of the artificial proteins. We determined a 2.5-Å resolution crystal of L056 (Fig. 3e and Supplementary Table 6). The global fold was similar to predictions, with a Cα root mean squared deviation (RMSD) of 2.9 Å from the backbone structure predicted by trRosetta and 2.3 Å RMSD from a wild-type T4 lysozyme structure[48,49]. The largest structural divergence occurs in the beta hairpin formed by residues 18–31. This region forms the bottom of the substrate-binding cleft[50] and is part of a hinge binding motion that is important for substrate binding[51]. The structure of the M6I mutant of T4 lysozyme (Protein Data Bank (PDB) accession 150L) is used as a model of the 'open' state of this hinge and more closely resembles the structure of L056 (1.0 Å Cα RMSD). Alignment with a structure featuring a covalently trapped substrate (PDB accession 148L) reveals that the active site cleft is well formed with the key catalytic residue Glu15 (Glu 11 in T4L) and key substrate-binding residue Thr30 (Thr26 in T4L) correctly positioned. In addition, the hydrophobic core of L056 is well packed, with only two small packing voids of less than 5 $Å^3$ in volume, which is typical for structures of this size[52].

To examine whether ProGen could generate functional proteins in the 'twilight zone' sequence identity[53] where two proteins are not assumed to share functional similarity[54], we generated 95 new artificial sequences with maximum sequence identities lower than 40% to any known natural protein for two lysozyme families (PF00959 and PF05838). Of the selected sequences, 78 out of the 89 (88%) expressed well and 24 out of the 78 (31%) were soluble (Supplementary Fig. 9). We purified six highly expressed proteins and found that they were all active, but with lower Michaelis–Menten activities than HEWL or the previously generated artificial proteins L056 and L070 (Fig. 3f, Supplementary Fig. 10, and Supplementary Table 7). The protein with the lowest sequence identity to a natural protein, D4 (31.4% ID to a protein from an *Arcobacter nitrofigilis* organism), had a k_{cat}/K_M of 20.2 $M^{-1}s^{-1}$, approximately 200-fold below HEWL. While the activity is substantially lower for these distant proteins, directed evolution could be employed to improve activity. Collectively, these results demonstrate a procedure for generating soluble, active proteins that are distant enough in sequence space that they might not be considered traditional sequence homologs.

To additionally compare across structural representations, we used AlphaFold2 (ref. [14]) to predict the structure of functional artificial sequences. As in the crystal structure of L056, the predicted artificial structures roughly match known structures found in nature (Supplementary Fig. 11) including for low identity (<40%) artificial sequences.

Trained on a universal protein sequence dataset spanning many families, ProGen designs proteins from any family when provided with the corresponding control tag as input. To explore this capability beyond the lysozyme clan, we evaluated the performance of ProGen in generating and predicting functional full-length sequences from families where other methods have previously been applied: chorismate mutase (CM)[7] and malate dehydrogenase (MDH). Generated proteins exhibit similar conservation patterns to natural sequence libraries (Fig. 4a,d). After aligning the generations to a sequence with known structure (Fig. 4b,e), we observed that the conserved positions in generated sequences correlate with ligand-binding and buried residues. Using previously published sequences and their experimentally measured assay data for CM[7] and MDH[55] proteins, we also evaluated the concordance of the ProGen model likelihood for these sequences to their relative activity and compared it with the generative methods used in the original studies—bmDCA[7] and proteinGAN[55]. Specifically, we measured per-token log-likelihoods for artificial sequences using ProGen and used them to predict if artificial sequences should function. On CM function data, ProGen log-likelihoods had an area under the curve (AUC) of 0.85, significantly better ($P < 0.0001$, two-tailed test, $n = 1617$) than bmDCA, which had an AUC of 0.78 (Fig. 4c). On MDH function

https://doi.org/10.1038/s41587-022-01618-2

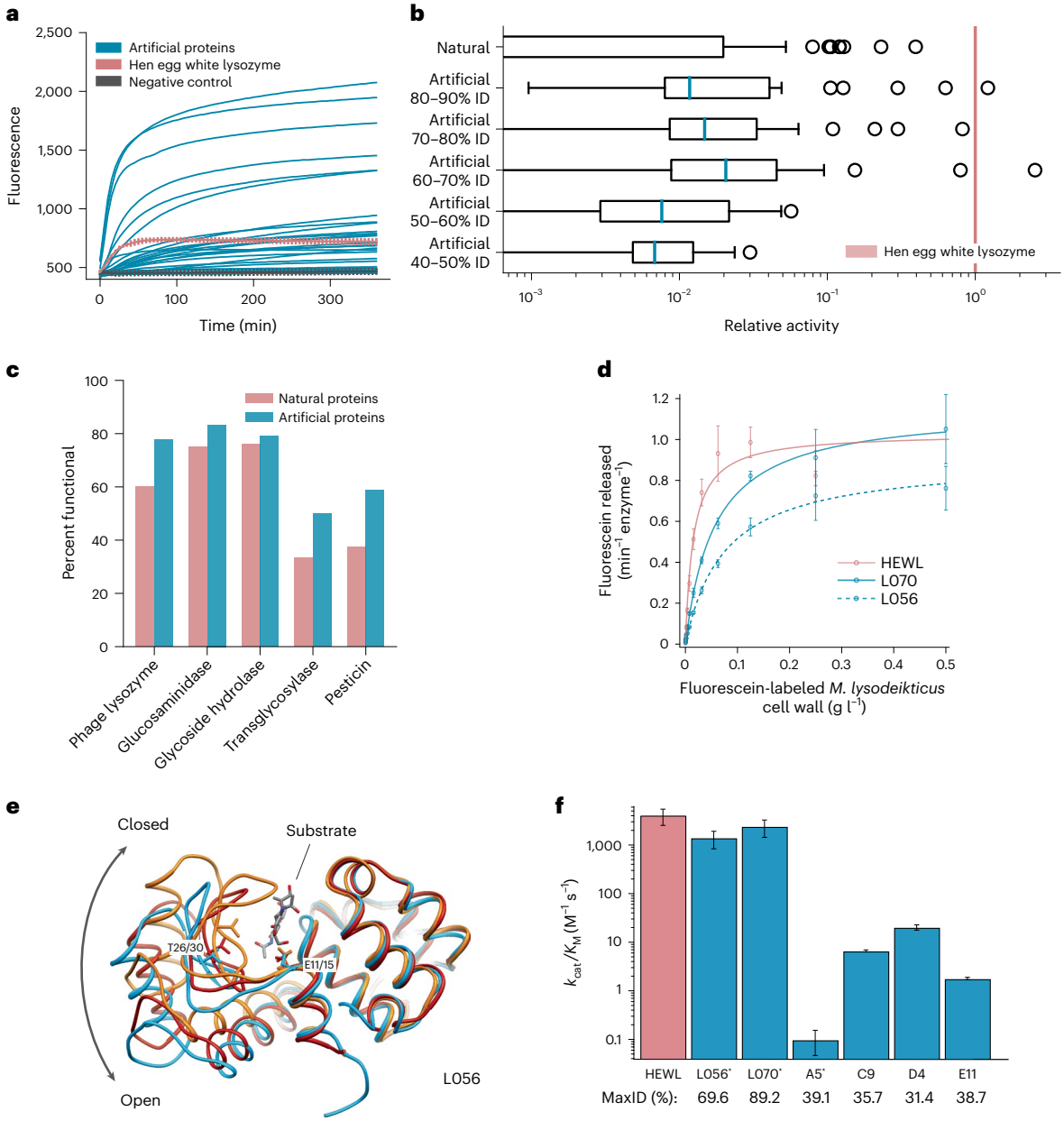

Fig. 3 | Artificial protein sequences are functional while reaching as low as 31% identity to any known protein, exhibit comparable catalytic efficiencies to a highly-evolved natural protein, and demonstrate similar structures to known natural folds. a, Artificial proteins bind well to substrates and exhibit high fluorescence responses over time ($n = 90$). For HEWL and negative (ubiquitin) controls, the minimum and maximum fluorescence range of $n = 3$ replicates are shown as bars. **b**, Artificial proteins remain active even while being dissimilar (40–50% max ID that is, top hit-identity) from known natural proteins. Outliers indicate high activity samples where relative activity is computed with respect to HEWL. Box plots are derived from $n = 90, 23, 28, 22, 8, 9$ samples for each category from top to bottom, respectively. Boxes display the median, first quartile and third quartile with whiskers which extend to 1.5× the inter-quartile range. **c**, Artificial proteins are functional across protein families. Functional is defined as a fluorescence one standard deviation above the maximum value of all negative controls. **d**, Michaelis–Menten kinetics of HEWL natural lysozyme (red) and two generated lysozymes (blue; L056 and L070) against cell wall substrate show comparable performance ($n = 3$ technical replicates where error bars represent standard deviation). **e**, We determined a 2.5-Å resolution crystal of L056 artificial lysozyme. A global overlay of L056 crystal structure with two representative T4 lysozyme conformations is shown with L056 presented in sky blue, 'open' conformation of M6I T4 lysozyme (PDB accession 150L) in dark red, 'closed' conformation of wild-type T4 lysozyme (PDB accession 3FA0) in orange, and substrate (PDB accession 148L) colored by element. Catalytic threonine (T30 in L056 and T26 in T4 lysozyme) and first catalytic glutamate (E15 in L056 and E11 in T4 lysozyme) are represented as sticks. **f**, Bars represent Michaelis–Menten k_{cat}/K_M constants derived for lysozyme variants demonstrating a range of catalytic activities across variants of varied maximal sequence IDs to known natural protein. Error bars represent propagated standard deviations derived from fitting procedure ($n = 3$ for A5, L056 and L070 technical replicates; $n = 4$ for C9 and E11 technical replicates; two biological replicates of each $n = 4$ technical replicates for D4). Asterisk denotes k_{cat}/K_M derived from initial rate analysis and unit converted (Supplementary Table 7).

data, ProGen log-likelihoods had an AUC of 0.94 (Fig. 4f), which was better than ProteinGAN discriminator scores, with an AUC of 0.87 ($P < 0.1$, two-tailed test, $n = 56$). In sum, the model likelihoods of ProGen are better aligned with experimentally measured assay data on two diverse protein datasets—CM and MDH—than the sequence-generation methods from original studies specifically tailored for these families.

https://doi.org/10.1038/s41587-022-01618-2

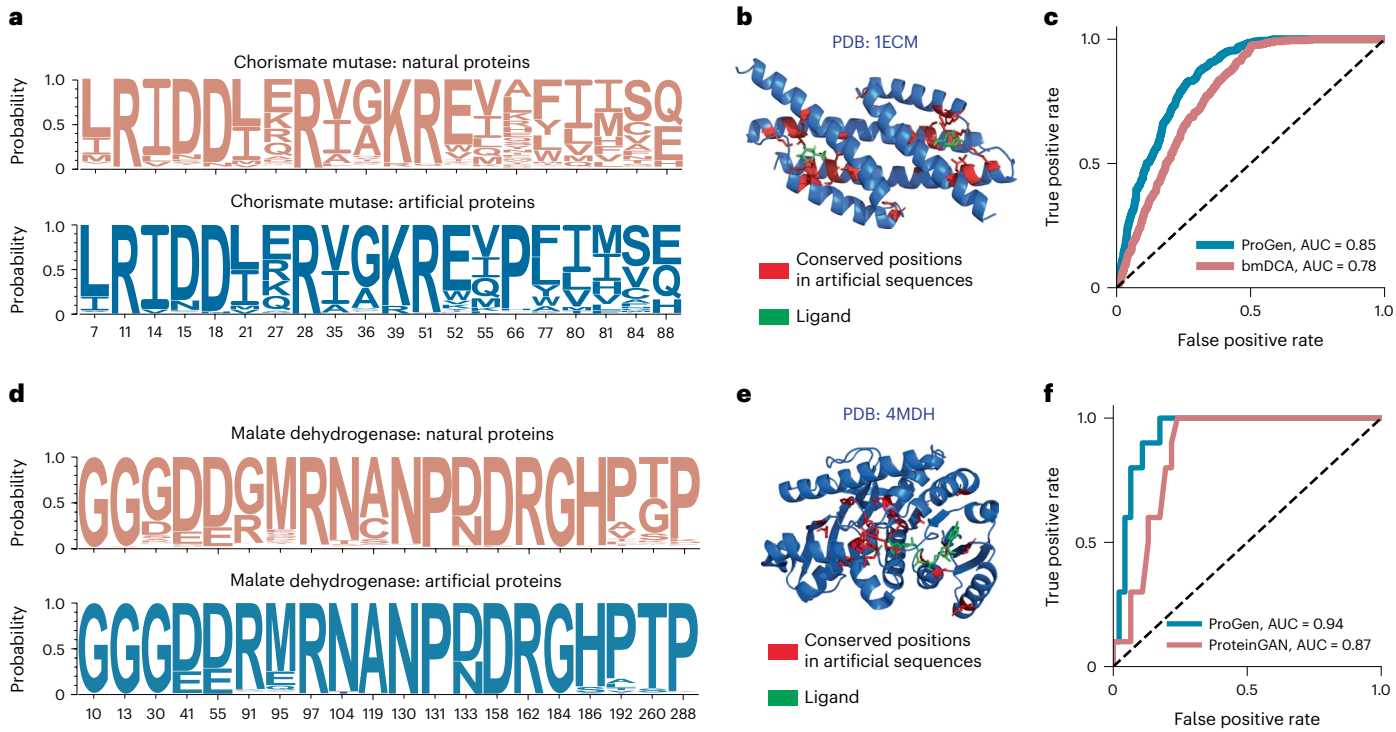

Fig. 4 | Applicability of conditional language modeling to other protein systems. a,**b**, Using the appropriate control tag, our language model, ProGen, can generate sequences for distinct protein families. Here we show that ProGen can generate CM enzymes that exhibit a similar residue distribution to nature (**a**) and the conserved residues among generated sequences correlate to ligand-binding sites (**b**). **c**, The model likelihoods of ProGen can also accurately predict the functionality of CM variants from published data, slightly better than the coevolutionary bmDCA[7] algorithm from the original study. **d**, ProGen can also generate MDH proteins that exhibit a similar residue distribution to nature. **e**, The conserved residues among generated sequences correlate to buried residues. **f**, The model likelihoods of ProGen are also accurate in predicting functionality of published variants of MDH, similar to the generative proteinGAN[55] model used in the original study.

To understand the relative impact of the universal sequence dataset and of sequences from the protein family of interest on the generation ability of ProGen, we perform two ablation studies using the CM and MDH experimentally measured assay data. First, we evaluated the performance of ProGen trained only with the universal sequence dataset. We measured per-token log-likelihoods for artificial sequences for this version of ProGen using control tags for CM and MDH. These likelihoods showed a significant drop in AUC of 0.18 for CM ($P < 0.0001$, two-tailed test, $n = 1{,}617$) and 0.08 for MDH ($P < 0.1$, two-tailed test, $n = 56$), as compared to fine-tuned ProGen when predicting if an artificial sequence should function. Conversely, the ProGen architecture trained on CM and MDH protein sequences alone without the benefit of initial training on the universal sequence dataset also showed a significant drop in performance as compared to fine-tuned ProGen—the AUC reduced by 0.11 ($P < 0.0001$, two-tailed test, $n = 1{,}617$) and 0.04 ($P < 0.05$, two-tailed test, $n = 56$) on the CM and MDH data, respectively.

These results indicate that both components of our training strategy—initial training on the universal sequence dataset and fine tuning on the protein family of interest—contribute significantly to final model performance. Training with the universal sequence dataset containing many protein families enables ProGen to learn a generic and transferable sequence representation that encodes intrinsic biological properties. Fine tuning on the protein family of interest steers this representation to improve generation quality in the local sequence neighborhood. Similar to the adaptability shown by neural networks trained on large datasets using transfer learning and fine tuning in natural language processing[25,34,56] and computer vision[57,58], protein language models have the potential to be a versatile tool for generating tailored proteins with desired properties. In Supplementary Fig. 12, the distribution of available sequences for different protein families indicates there is a large portion of the protein universe where our current technique would be useful. We extrapolate that it may be possible to successfully generate artificial proteins with functional activity without fine tuning, especially for larger protein families; however, it would likely do so at a small success rate. We did not attempt to experimentally test generated sequences without additional fine tuning in our study.

Discussion

In conclusion, our study shows that a state-of-the-art transformer-based conditional language model trained only with evolutionary sequence data generates functional artificial proteins across protein families. Additional analyses suggest that our model has learned a flexible protein sequence representation that can be applied to diverse families such as lysozymes, CM, and MDH. While we do not expect our language model to generate proteins that belong to a completely different distribution or domain (for example, creating a new fold that catalyzes an unnatural reaction), it can substantially expand the space of protein sequences from those sampled by evolution. Combining biophysical modeling with generative models could further help us explore data distributions that are completely distinct from those sampled by evolution[17,59,60]. Applications of our model could include generating synthetic libraries of highly likely functional proteins for discovery or iterative optimization. In combination with ever-increasing sources of sequence data and more expressive control tags, our work points to the potential for the use of deep-learning-based language models for precise de novo design of proteins to solve problems in biology, medicine, and the environment.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-022-01618-2.

References

1. Koga, N. et al. Principles for designing ideal protein structures. *Nature* **491**, 222–227 (2012).
2. Lin, Y.-R. et al. Control over overall shape and size in de novo designed proteins. *Proc. Natl Acad. Sci. USA* **112**, E5478–E5485 (2015).
3. Huang, P.-S., Boyken, S. E. & Baker, D. The coming of age of de novo protein design. *Nature* **537**, 320–327 (2016).
4. Huang, P.-S. et al. De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nat. Chem. Biol.* **12**, 29–34 (2016).
5. Boyken, S. E. et al. De novo design of protein homo-oligomers with modular hydrogen-bond network–mediated specificity. *Science* **352**, 680–687 (2016).
6. Lapedes, A. S., Bertrand, G. G., LonChang, L. & Stormo, G. D. Correlated mutations in models of protein sequences: Phylogenetic and structural effects. *Lect. Notes Monogr. Ser.* **33**, 236–256 (1999).
7. Russ, W. P. et al. An evolution-based model for designing chorismate mutase enzymes. *Science* **369**, 440–445 (2020).
8. Hopf, T. A. et al. The EVcouplings Python framework for coevolutionary sequence analysis. *Bioinformatics* **35**, 1582–1584 (2019).
9. Morcos, F. et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl Acad. Sci. USA* **108**, E1293–E1301 (2011).
10. Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *eLife* **3**, e02030 (2014).
11. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **16**, 1315–1322 (2019).
12. Wu, Z. et al. Signal peptides generated by attention-based neural networks. *ACS Synth. Biol.* **9**, 2154–2161 (2020).
13. Shin, J.-E. et al. Protein design and variant prediction using autoregressive generative models. *Nat. Commun.* **12**, 2403 (2021).
14. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
15. Bryant, D. H. et al. Deep diversification of an AAV capsid protein by machine learning. *Nat. Biotechnol.* **39**, 691–696 (2021).
16. Das, P. et al. Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nat. Biomed. Eng.* **5**, 613–623 (2021).
17. Anishchenko, I. et al. De novo protein design by deep network hallucination. *Nature* **600**, 547–552 (2021).
18. Moffat, L., Kandathil, S. M. & Jones, D. T. Design in the DARK: Learning deep generative models for De Novo Protein Design. Preprint at bioRxiv https://doi.org/10.1101/2022.01.27.478087 (2022).
19. Ferruz, N., Schmidt, S. & Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nat. Commun.* **13**, 4348 (2022).
20. Huang, B. et al. A backbone-centred energy function of neural networks for protein design. *Nature* **602**, 523–528 (2022).
21. Leinonen, R. et al. UniProt archive. *Bioinformatics* **20**, 3236–3237 (2004).
22. Bairoch, A. et al. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **33**, D154–D159 (2005).
23. Finn, R. D. et al. Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–D230 (2014).
24. Vaswani, A. et al. Attention is all you need. In *31st Conference on Neural Information Processing Systems* (NIPS, 2017).
25. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (NAACL-HLT, 2019).
26. Brown, T. B. et al. Language models are few-shot learners. In *34th Conference on Neural Information Processing Systems* (NeurIPS, 2020).
27. Zellers, R. et al. Defending against neural fake news. In *33rd Conference on Neural Information Processing Systems* (NeurIPS, 2019).
28. Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C. & Socher, R. CTRL: a conditional transformer language model for controllable generation. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1909.05858 (2019).
29. AlQuraishi, M. The future of protein science will not be supervised. *Some Thoughts on a Mysterious Universe* https://moalquraishi.wordpress.com/2019/04/01/the-future-of-protein-science-will-not-be-supervised/ (2019).
30. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl Acad. Sci. USA* **118**, e2016239118 (2021).
31. Elnaggar, A. et al. ProtTrans: towards cracking the language of lifes code through self-supervised deep learning and high performance computing. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 7112–7127 (2021).
32. Peters, M. E. et al. Deep contextualized word representations. In *Proceedings of North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (NAACL-HLT, 2018).
33. Howard, J. & Ruder, S. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (ACL, 2018).
34. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving language understanding by generative pre-training. Preprint at https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf (2018).
35. Biswas, S., Khimulya, G., Alley, E. C., Esvelt, K. M. & Church, G. M. Low-N protein engineering with data-efficient deep learning. *Nat. Methods* **18**, 389–396 (2021).
36. Pfaff, C. W. Constraints on language mixing: Intrasentential code-switching and borrowing in Spanish/English. *Language* **55**, 291–318 (1979).
37. Poplack, S. Sometimes I'll start a sentence in Spanish Y TERMINO EN ESPAÑOL: toward a typology of code-switching. *Linguistics* **18**, 581–618 (1980).
38. Dathathri, S. et al. Plug and play language models: a simple approach to controlled text generation. In *8th International Conference on Learning Representations* (ICLR, 2020).
39. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **405**, 442–451 (1975).
40. Broendum, S. S., Buckle, A. M. & McGowan, S. Catalytic diversity and cell wall binding repeats in the phage-encoded endolysins. *Mol. Microbiol.* **110**, 879–896 (2018).
41. Love, M. J., Abeysekera, G. S., Muscroft-Taylor, A. C., Billington, C. & Dobson, R. C. J. On the catalytic mechanism of bacteriophage endolysins: opportunities for engineering. *Biochim. Biophys. Acta. Proteins Proteom.* **1868**, 140302 (2020).

42. Martin, P. P. *Potts Models And Related Problems In Statistical Mechanics* (World Scientific, 1991).
43. Thomas, J., Ramakrishnan, N. & Bailey-Kellogg, C. Graphical models of residue coupling in protein families. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **5**, 183–197 (2008).
44. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein–protein interaction by message passing. *Proc. Natl Acad. Sci. USA* **106**, 67–72 (2009).
45. Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S.-I. & Langmead, C. J. Learning generative models for protein fold families. *Proteins* **79**, 1061–1078 (2011).
46. Stein, R. R., Marks, D. S. & Sander, C. Inferring pairwise interactions from biological data using maximum-entropy probability models. *PLoS Comput. Biol.* **11**, e1004182 (2015).
47. Mirdita, M., Steinegger, M., Breitwieser, F., Söding, J. & Levy Karin, E. Fast and sensitive taxonomic assignment to metagenomic contigs. *Binformatics* **37**, 3029–3031 (2021).
48. Mooers, B. H. M., Tronrud, D. E. & Matthews, B. W. Evaluation at atomic resolution of the role of strain in destabilizing the temperature-sensitive T4 lysozyme mutant Arg 96 → His. *Protein Sci.* **18**, 863–870 (2009).
49. Baase, W. A., Liu, L., Tronrud, D. E. & Matthews, B. W. Lessons from the lysozyme of phage T4. *Protein Sci.* **19**, 631–641 (2010).
50. Kuroki, R., Weaver, L. H. & Matthews, B. W. A covalent enzyme–substrate intermediate with saccharide distortion in a mutant T4 lysozyme. *Science* **262**, 2030–2033 (1993).
51. Mchaourab, H. S., Oh, K. J., Fang, C. J. & Hubbell, W. L. Conformation of T4 lysozyme in solution. Hinge-bending motion and the substrate-induced conformational transition studied by site-directed spin labeling. *Biochemistry* **36**, 307–316 (1997).
52. Kim, J.-K. et al. BetaCavityWeb: a webserver for molecular voids and channels. *Nucleic Acids Res.* **43**, W413–W418 (2015).
53. Rost, B. Twilight zone of protein sequence alignments. *Protein Eng.* **12**, 85–94 (1999).
54. Pearson, W. R. An introduction to sequence similarity ('homology') searching. *Curr. Protoc. Bioinforma.* **3**, 3.1 (2013). ChapterUnit.
55. Repecka, D. et al. Expanding functional protein sequence spaces using generative adversarial networks. *Nat. Mach. Intell.* **3**, 324–333 (2021).
56. Ruder, S., Peters, M. E., Swayamdipta, S. & Wolf, T. Transfer learning in natural language processing. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics* (eds Jill Burstein, J., Doran, C. & Solorio T.) (Association for Computational Linguistics, 2019).
57. Huh, M., Agrawal, P. & Efros, A. A. What makes ImageNet good for transfer learning? Preprint at *arXiv* https://doi.org/10.48550/arXiv.1608.08614 (2016).
58. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
59. Norn, C. et al. Protein sequence design by conformational landscape optimization. *Proc. Natl Acad. Sci. USA* **118**, e2017228118 (2021).
60. Anand, N. et al. Protein sequence design with a learned potential. *Nat. Commun.* **13**, 746 (2022).

Methods

Training data curation

To train ProGen, we collected a universal protein sequence dataset containing 281 million non-redundant protein sequences (from >19,000 Pfam[23] families) and associated metadata (as control tags) from UniParc[21], UniprotKB[22], Pfam[23] and NCBI taxonomic information[61] (Fig. 1d and Supplementary Table 1). The amino acid vocabulary consisted of the standard 25 amino acids designations in IUPAC[62]. The control tags were divided into two categories: (1) keyword tags and (2) taxonomic tags. Following the definitions laid out in the UniprotKB controlled, hierarchical vocabulary of keywords (many of which are derived from Gene Ontology (GO) terms[63]), the control keyword tags included 1,100 terms ranging from cellular component, biological process, and molecular function terms. The taxonomic tags include 100,000 terms from the NCBI taxonomy across the eight standard taxonomic ranks. The aggregated dataset was split into a training set of size 280 million and two test sets, an out-of-distribution test set (OOD-test) of size 100,000 from 20 protein families and a randomly sampled in-domain test set (ID-test) of size 1 million, that were held out for training and used for evaluation. After model training on the training database, the model was further trained, that is fine tuned, to the following datasets for generation and classification tasks.

For fine tuning on lysozyme proteins, five protein families from the Pfam database were selected, phage lysozyme (PF00959), pesticin (PF16754), glucosaminidase (PF01832), glycoside hydrolase family 108 (PF05838) and transglycosylase (PF06737), yielding a total of 55,948 sequences. Proteins were provided to the model during fine tuning as unaligned protein sequences with one control tag prepended for the protein family designation. For fine tuning on CM proteins, a search with HHBlits and blastp was performed with residues 1–95 of EcCM (the CM domain of the *E. coli* CM-prephenate dehydratase, the P-protein) yielding 20,214 sequences. For fine tuning on MDH proteins, the L-lactate/MDH protein family from Interpro IPR001557 was selected with 17,094 sequences.

Conditional language modeling

Let $a = (a_1, ..., a_{na})$ be a sequence of amino acids that specifies a protein of length $n_a - 1$ appended with an 'end of sequence' token. Let $c = (c_1, ..., c_{nc})$ be an associated set of descriptors such as protein family or source organism, that is, 'control tags', through which we would like to control generation of amino acid sequences. Let $x = [c;a]$ be the sequence formed by prepending a control tag sequence to an amino acid sequence. The probability over such a combined sequence of length $n = n_a + n_c$ is then $P(x)$. Language modeling decomposes the problem of generating x into a next-token prediction problem[64], where a token can either be an amino acid or a control tag. We train a neural network with parameters θ to minimize the negative log-likelihood over a dataset $D = \{x^1, \ldots, x^{|D|}\}$

$$L(D) = -\frac{1}{|D|}\sum_{k=1}^{|D|}\frac{1}{n^k}\sum_{i=1}^{n^k}\log p_\theta\left(x_i^k | x_{<i}^k\right) \quad (1)$$

A new protein $\underline{a}$ of length m_a with desired properties encoded by a control tag sequence $\underline{c}$ of length m_c can then be generated by sequentially sampling its constituent tokens: $p_\theta(a_1|\underline{c})$, $p_\theta(a_2|\underline{a_1}, \underline{c})$,…,$p_\theta(a_j|\underline{a_{<j}}, \underline{c})$ (ref. [65]). Generation continues until the model generates an 'end of sequence' token.

We use a transformer-based[24] neural network architecture for constructing ProGen. The transformer learns long-range context within sequences using a series of stacked layers, each containing a self-attention mechanism (Fig. 1e). The self-attention mechanism in each layer infers pairwise interaction relationships between all positions in its input sequence. Stacking multiple self-attention layers allows us to learn multiple-residue interactions[66]. The transformer-based approach has been shown to be related to coevolutionary methods for sequence design such as MRFs[67], Potts models[68] and Hopfield networks[69]. In contrast to transformer-based language models that encode amino acid sequences for discriminative protein prediction tasks[30,70,71], ProGen is a decoder transformer tailored for autoregressive generation: it generates a sequence in a left-to-right manner, token-by-token, where the next token is conditioned on all previously generated tokens.

The transformer architecture of ProGen has 36 layers, and 8 self-attention heads per layer and a total of 1.2 billion trainable neural network parameters. We trained ProGen to minimize the negative log-likelihood defined in Eq. 1 using this dataset with a batch size of 2,048 for 1 million iterations. Training was performed across 256 Google Cloud TPU v3 cores for 2 weeks. Once trained, ProGen could be used to generate protein sequences from scratch by specifying a control tag (for example, protein family identifier from Pfam; Fig. 1c).

ProGen training

For training, we included each sequence and its reverse. We prepended each sequence with a corresponding subset of control tags. For a given sequence, there can be multiple versions across databases, each with their own associated control tags. We randomly sampled which set of control tags to use, but biased sampling toward SwissProt tags as they are manually verified. Additionally, we always included a sample with the sequence alone without control tags so that ProGen could be used to complete proteins using sequence data alone. We truncated all sequences to a maximum length of 512. Sequences of length less than 512 were padded, and padded tokens were excluded from the cost function used for training. The average token length of control tags during pretraining was eight. Our model was implemented in TensorFlow and trained with a global batch size of 2048 distributed across 256 cores of a Cloud TPU v3 Pod for a fixed number of 1 million iterations, with no specific stopping criterion. The perplexity on a held-out test set was monitored and did not exceed training set perplexity throughout model training. Training took approximately 2 weeks using Adagrad with linear warmup from 0 to 1×10^{-2} over the initial 40,000 steps with a linear decay for the remainder of training. The model was initialized with pretrained weights of CTRL[28], which was trained on an English language corpus.

Lysozyme generation

Fine tuning involves making limited, computationally inexpensive, gradient updates to the parameters of the trained model. We fine tuned ProGen to the 55,948-sequence fine tuning dataset using the conditional language modeling loss function introduced in Eq. 1, using a separate control tag for each of the five lysozyme families. The fine tuning dataset was clustered at 80% sequence identity and 10% of the clusters were held-out as a validation set for hyperparameter optimization and stopping criteria. The model was fit for 4 epochs using the Adam optimizer[72] with a learning rate of 0.0001, batch size of 2, gradient norm clipping[73] threshold of 0.25, and a dropout[74] rate of 0.1. We then applied sampling using the final checkpoint of the fine-tuned model. We generated 1 million artificial sequences from the learned conditional probability distribution $p_\theta(a_i|a_{<i}, c)$ using each of the five lysozyme families as a control tag c, and applying top-p sampling[75], which zeros out the probability of the tail of the distribution during sampling, and uses a hyperparameter p to determine what fraction of the original distribution to keep. Lower p values result in sequences with a higher likelihood under the model, but lower diversity. We generated a batch of 1 million synthetic sequences (Supplementary Fig. 3) using p values that varied in [0.25,0.50,0.75], and applied the sequence selection criteria in the next section to determine which sequences to synthesize.

Lysozyme sequence selection

We selected sequences for synthesis by ranking them using the combination of an adversarial discriminator[27,76] and generative model log-likelihood scoring[77]. First, we trained an adversarial discriminator to distinguish between natural lysozymes and ProGen-generated

lysozymes. A higher discriminator score indicates a protein sequence that is 'semantically' and 'grammatically' closer to natural sequences, but not necessarily one of high sequence identity to natural proteins. To train the discriminator, we generated a batch of samples from fine-tuned ProGen (with nucleus sampling turned off, or $p = 1$) that was the same size and distribution of families as our dataset of natural lysozymes. The discriminator architecture was a fine-tuned TAPE-BERT[71]. For robustness, we trained three discriminators using different random seeds. We assigned each sequence a discriminator score as the geometric mean of the probability of the sample being a natural sequence as predicted by the three discriminators. We also assigned each sequence a log-likelihood score as the average per-token log-likelihood for each sample computed using the fine-tuned ProGen model and conditioned on the control tag used to generate the sequence, given by

$$\text{Score}\,(a) = \frac{1}{n_a}\sum_{i=1}^{n_a}\log p_\theta\left(a_i|a_{<i}, c\right) \quad (2)$$

A higher log-likelihood score indicates a sequence close to the probability distribution of sequences seen in training. Model log-likelihoods are directly correlated with perplexity as a language modeling evaluation metric. We selected artificial sequences using separate rankings based on the discriminator and log-likelihood scores. We separately ranked candidate sequences in maximum sequence identity ranges of 40–50%, 50–60%, 60–70%, 70–80% and 80–90%. For each range, we added the top discriminator-ranked sequences, skipping any sequences that were >80% identical to any previously selected sequence, for a total of 90 sequences. Ten more sequences were added on the basis of ranking by generative model log-likelihood scores in each range, again skipping any sequences with >80% identity to any previously selected sequence.

Evaluating ProGen on other protein systems

We also evaluated ProGen on generation of CM and MDH proteins. We separately fine tuned ProGen on datasets of CM and MDH proteins using the Adam optimizer, a learning rate of 1×10^{-4}, a gradient norm clipping threshold of 0.25, and a dropout rate of 0.1. We also prepended the CM and MDH data with control tags that corresponded to CM and MDH families in original training of ProGen. After fine tuning, we generated a set of 64,000 sequences using top-p sampling ($p = 0.75$) from the CM and MDH fine-tuned models, respectively. We measured concordance of the log-likelihoods of our model with protein function data on CM and MDH sequences, and compared with bmDCA[7] and ProteinGAN[55] baselines, respectively. We computed the AUC in receiver operating characteristic (ROC) curves for predicting binary function labels from model scores. We computed model scores for each sequence in both CM and MDH by using the per-token model log-likelihood in Eq. 2. We used model scores for bmDCA given by negative energy of each CM sequence provided by the authors of the study[7]. We also applied thresholding at 0.42 norm relative enrichment to obtain binary labels for CM function, which roughly corresponds to the cutoff point between two modes that exist in CM function data, to be used for ROC curves, following the original study[7].

Since model likelihoods for GANs are intractable, we used discriminator scores corresponding to the probability at which the ProteinGAN discriminator predicted each sample was real as a ProteinGAN model score for each MDH sequence. The MDH functional labels are binary, so no thresholding was needed to compute AUCs. For an ablation study on ProGen, we also evaluate: i) a randomly initialized LM that has the same architecture as ProGen and is fine tuned to the same task-specific data as ProGen (CM or MDH), but is not pretrained on a larger dataset; and ii) ProGen without task-specific fine tuning, conditioning on control tags for CM or MDH from the original ProGen pretraining data. After measuring the AUC of each model for each dataset, we used bootstrapping to compute the statistical significance of the difference in AUC of fine-tuned ProGen versus the reference method (bmDCA and ProGen ablations for CM, ProteinGAN and ProGen ablations for MDH). At each bootstrapping iteration, we resampled a new dataset of fitness and model score pairs the same size as the original dataset by randomly selecting data points from the original dataset with replacement. For each sample dataset, we compute the difference in AUC score between fine-tuned ProGen and the reference method. We drew a total of 10,000 bootstrapping samples, and the P value is given by the percentage of the samples where the baseline achieves an AUC greater than or equal to fine-tuned ProGen, multiplied by two to give two-tailed.

Materials

All reagents were purchased from Thermo Fisher Scientific unless otherwise noted. DNAs used for in vitro translation were purchased from Twist Bioscience and DNAs used for *E. coli* expression and purification were purchased from VectorBuilder.

High-throughput cell-free expression of lysozymes

Lysozymes were expressed using the Tierra Bioscience cell-free expression platform. Cell-free extracts for protein expression were prepared according to the methods of Sun et al.[78] with the following modifications: Terrific Broth was used in lieu of 2xYT, cells were lysed in a single pass by French press at 10,000 p.s.i, dithiothreitol was omitted from wash buffers, and run-off and dialysis steps were removed to streamline extract processing. Expression reactions were composed of cell-free extract, an energy buffer and a linear DNA template containing a promoter sequence, the protein sequence of interest, the sequence of a strep purification tag and a terminator sequence; reactions were carried out at 29 °C for 6 hours. Expression reactions for screening optimal affinity purification tag terminus were performed in 10 µL volumes; selected reactions with good expression were then scaled to 200 µL. Lysozymes were purified from expression reactions by affinity chromatography with elution by enzymatic cleavage with 3 C protease leaving a small sequence scar.

High-throughput screening of lysozyme activity

Purified cell-free synthesized lysozymes were assayed with the EnzChek Lysozyme Assay Kit (Thermo Fisher Scientific). The assay was performed according to protocol with minimal modifications. HEWL standards and purified proteins in buffer (100 mM Tris pH 7.4, 150 mM NaCl, 2 mM TCEP, 20% glycerol) were brought to 50 µl with reaction buffer (100 mM sodium phosphate pH 7.5, 100 mM NaCl, 2 mM NaN_3) in a 96-well plate. Fifty microliters of DQ lysozyme substrate, fluorescein conjugate (1 mg ml^{-1}) was added to each well and fluorescence (excitation 485/20; emission 528/20) was collected every 5 min with a Synergy 2 multi-mode microplate reader (BioTek) for 6 h at 37 °C.

For each 96-well plate, three random wells were dedicated for HEWL controls and three wells were dedicated for a negative control of ubiquitin expressed and purified on the Tierra Biosciences cell-free expression platform. A purified protein was considered functional if it exhibited a higher fluorescence than one standard deviation above the maximum fluorescence value of all negative controls. The relative activity for each protein was calculated by the following equation:

$$\text{Relative activity} = \frac{r_{\text{protein}} - r_{\text{negative}}}{r_{\text{HEWL}} - r_{\text{negative}}} \times \frac{m_{\text{HEWL}}}{m_{\text{protein}}} \quad (3)$$

Where r is the linear rate of fluorescence increase in the initial 20 min of the fluorogenic assay and m is the mass of protein as determined by Bradford assay concentration and measured volumes.

E. coli expression of lysozyme variants

We chose five generated lysozyme variants (L008, L013, L038, L056, L070) for expression in *E. coli* on the basis of strength of signal in the in vitro assay, expression level in the cell-free system and max ID to natural proteins. Generated lysozyme variants, were codon

https://doi.org/10.1038/s41587-022-01618-2

optimized for *E. coli* (Integrated DNA Technologies) with an HRV3C protease site N-terminal of the open reading frame. DNA was synthesized and cloned in-frame with a 5′ His_6-tag in a pET vector and transformed into BL21(DE3) (Vectorbuilder). One liter of Terrific Broth (Fisher) was prewarmed to 37 °C before being inoculated with 10 ml overnight starter culture. Cultures were grown to $0.6 < OD_{600} < 1.0$ before temperature was dropped to 16 °C for expression. Cultures were induced with 0.5 mM isopropyl β-D-1-thiogalactopyranoside (source) and protein expression was allowed to continue overnight. For induced cultures of L056 and L070, turbidity was observed in the spent medium after cells were pelleted at 3,500 r.c.f. for 30 min at 4 °C. Spent medium also harbored lysozyme activity as ascertained through fluorescence increase over time of the fluorescein-labeled *M. Lysodeikticus* cell wall substrate (EnzChek kit; Thermo Fisher). Spent medium was saved for protein purification (outlined below) and cell pellet frozen and stored at −20 °C. Variant L008 did not express under multiple different conditions. L013 and L038 expressed highly to inclusion bodies.

Purification of L056 and L070 from spent medium

Medium was split into two 0.5 l pools each. The first pools were loaded onto a 5 ml HisTrap FF NiNTA column (GE) using a peristaltic pump at room temperature. Columns were washed with 200 ml 30 mM HEPES pH 7.6, 150 mM NaCl, 25 mM imidazole, 0.5 mM TCEP. Columns were eluted with 25 ml 30 mM HEPES pH 7.6, 150 mM NaCl, 250 mM imidazole, 0.5 mM TCEP. Eluates were concentrated to 8–10 ml and dialyzed against 30 mM HEPES pH 7.6, 150 mM NaCl, 0.5 mM TCEP with HRV3C protease added overnight at 4 °C. Dialyzed protein was put through an ortho 5 ml HisTrap FF NiNTA column (GE) to remove HRV3C protease and uncleaved lysozyme. Though highly pure by SDS-PAGE analysis, protein was further purified by size-exclusion chromatography and loaded on an S75 10/300 gl column pre-equilibrated with 30 mM HEPES pH 7.6, 150 mM NaCl, 0.5 mM TCEP. Two peaks were resolved for each variant that harbored lysozyme activity against the fluorescein-labeled *M. Lysodeikticus* cell wall substrate (EnzChek kit; Thermo Fisher). Individual peaks were pooled and protein concentration determined either by Bradford assay (Biorad) or by SDS-PAGE using colloidal coomassie (Thermo Fisher) and HEWL in-gel standards.

The second spent medium pools were batch bound to 5 ml HisPur NiNTA resin (Thermo Fisher) at 4 °C for 1 h before protein-bound resin was pelleted through centrifugation at 3,000 r.c.f. for 5 min at 4 °C. Protein-bound resin was resuspended with 25 ml 30 mM HEPES pH 7.6, 150 mM NaCl, 25 mM imidazole, 0.5 mM TCEP and applied to a gravity flow column (BioRad) at room temperature. Columns were washed with 200 ml 30 mM HEPES pH 7.6, 150 mM NaCl, 25 mM imidazole, 0.5 mM TCEP. Columns were eluted with 25 ml 30 mM HEPES pH 7.6, 150 mM NaCl, 250 mM imidazole, 0.5 mM TCEP. Eluates were concentrated to 8–10 ml and dialyzed against 30 mM HEPES pH 7.6, 150 mM NaCl, 0.5 mM TCEP with HRV3C protease added overnight at 4 °C. Lysozyme was separated from HRV3C protease by size-exclusion chromatography on an S75 10/300 gl column pre-equilibrated with 30 mM HEPES pH 7.6, 150 mM NaCl, 0.5 mM TCEP. Two peaks were resolved for each variant that harbored lysozyme activity against the fluorescein-labeled *M. Lysodeikticus* cell wall substrate (EnzChek kit; Thermo Fisher) that corresponded to peaks observed in the first pool purification. Individual peaks were pooled and protein concentration determined either by Bradford assay (Biorad) or by SDS-PAGE using colloidal coomassie (Thermo Fisher) and HEWL in-gel standards.

Michaelis–Menten kinetics of lysozyme variants using fluorescein-labeled *M. lysodeikticus* cell wall

Fluorescein-labeled *M. Lysodeikticus* cell wall substrate (EnzChek kit; Thermo Fisher) was reconstituted in 30 mM HEPES pH 7.6, 150 mM NaCl to 1 mg ml^{-1}, aliquoted and stored at −20 °C until use. A serial two-fold dilution series of substrate was prepared in 30 mM HEPES pH 7.6, 150 mM NaCl and treated as a 2× solution for enzymatic assays. Enzyme concentration was calculated either through Bradford assay (Bio-Rad) or by SDS-PAGE, in-gel using Novex or Abcam Colloidal Coomassie stain against a HEWL standard (Alfa Aesar). Enzymes were diluted to between 10 and 100 nM in 30 mM HEPES pH 7.6, 150 mM NaCl (HEWL) or 30 mM HEPES pH 7.6, 150 mM NaCl, 0.5 mM TCEP (L056 and L070) and these stocks treated as a 2× solution for enzymatic assays. Kinetic assays were performed in a Tecan Spark 10 M plate reader using monochrometers with a fixed 20 nm bandpass filter in a 384-well black-bottom plate (Corning) at 10 µl final volume. Reactions were initiated by pipetting 5 µl of substrate into appropriate wells followed immediately by 5 µl of enzyme, mixed by pipetting before starting data acquisition. The dead time from reaction initiation to acquisition of first read was measured to be 24 s. For reactions carried out above ambient temperature (25 °C), the plate was preincubated at temperature for at least 5 min before reaction initiation. Initial velocities were calculated through linearly fitting fluorescence intensity (a.u.) versus time for the first 2 min of each reaction. Finally, velocities were converted from a.u. to fluorescein liberated through application of a fluorescein (Sigma) standard curve (Supplementary Fig. 7) and normalized to enzyme concentration. Averaged data ($n = 3$ technical replicates) were non-linearly fit to the Michaelis–Menten model (Eq. 4) in IgorPro 7 to report k_{cat} in units of fluorescein liberated $enzyme^{-1}$ min^{-1} and K_M in units of g l^{-1} (the average molecular weight of the fluorescein-labeled *M. Lysodeikticus* cell wall substrate was unknown and likely heterogeneous).

$$v_o = \frac{k_{cat} * [\text{substrate}]}{K_M + [\text{substrate}]} \quad (4)$$

For low ID lysozyme A5, the above protocol was altered slightly to accommodate lower catalytic activity of these variants: reaction volumes were increased to 20 µl, the plate was covered with an optically transparent seal (Microseal 'B' seal; BioRad) to mitigate sample evaporation, fluorescence reads were taken every 5 min for 16 h with 5 s of linear plate shaking before each measurement to minimize photobleaching of substrate and ensure substrate maintained homogeneous dispersion during longer reactions. The rate of substrate photobleaching was measured using a buffer-only control and used as a background rate subtraction for initial rate determination.

Lysozyme k_{cat}/K_M extrapolation from pseudo-first-order kinetic data

For the higher molecular weight L056 and L070 species whose K_M values were beyond the concentration regime of the fluorescein-labeled *M. Lysodeikticus* cell wall substrate (EnzChek kit; Thermo Fisher), the ratio k_{cat}/K_M was measured through pseudo-first-order kinetics where when [Enyzme] » Substrate the Michaelis–Menten model simplifies to Eq. 5. Fluorescein-labeled *M. Lysodeikticus* cell wall substrate was diluted to 0.01 g l^{-1} and this stock was treated as 2× for kinetic assays. Kinetic assays were performed in a Tecan Spark 10M plate reader using monochrometers with a fixed 20-nm bandpass filter in a 384-well black-bottom plate (Corning) at 10 µl final volume. The dead time from reaction initiation to acquisition of first read was measured at 24 s and the 0 s fluorescence intensity was measured through dilution of substrate with buffer. Reactions were initiated by pipetting 5 µl 2× enzyme into 5 µl 2× substrate in a prewarmed 384-well black assay plate (Corning). Five technical replicates were performed across four enzyme concentrations. The resultant data were not described by a single exponential model but were described by a double exponential model (Eq. 6), likely owing to the heterogeneity of the substrate, and all data were fit in IgorPro 7. The reciprocal of the weighted sum of each tau component was taken to estimate a single k_{obs} value for subsequent analysis (Eq. 7). To estimate k_{cat}/K_M, k_{obs} values were plotted against enzyme concentration where the slope of a linear fitting is equal to k_{cat}/K_M.

$$k_{obs} = \frac{k_{cat}}{K_M} \times [\text{Enzyme}] \quad (5)$$

$$y = y_o + \text{Amplitude}_1 \times e^{\left(-\frac{x}{tau_1}\right)} + \text{Amplitude}_2 \times e^{\left(-\frac{x}{tau_2}\right)} \quad (6)$$

$$k_{obs}^{-1} = \text{tau}_1 \times \frac{\text{Amplitude}_1}{\text{Amplitude}_2} + \text{tau}_2 \times \frac{\text{Amplitude}_2}{\text{Amplitude}_1} \quad (7)$$

For lowID lysozyme variants the above protocol was altered slightly to accommodate lower catalytic activity of these variants: reaction volumes were increased to 20 µl where 2 µl 0.05 mg ml^{-1} (0.005 mg ml^{-1} final) of fluorescein-labeled *M. lysodeikticus* cell wall substrate was diluted with 18 µl lysozyme variant to initiate reactions, plate was covered with an optically transparent seal (Microseal 'B' seal; BioRad) to mitigate sample evaporation, fluorescence reads were taken every 5 min for 16 h with 5 s of linear plate shaking before each measurement to minimize photobleaching of substrate and ensure substrate maintained homogeneous dispersion during longer reactions. At least four enzyme concentrations were tested. Initial rates from these data (first 2 h of reaction) were also collected to determine enzyme relative activity according to Eq. 3 (Supplementary Fig. 10).

Crystallization and structure determination of L056

Purified L056 was concentrated to 18.6 mg ml^{-1} in 30 mM HEPES pH 7.6, 150 mM NaCl, 0.5 mM TCEP. Crystals were identified from sitting drop vapor diffusion experiments set at 20 °C with a 1:1 ratio of 200 nl protein and 200 nl well solution (0.1 M CHES 9.5 pH, 30 %w/v PEG 3000). Diffraction data were collected from a single crystal at Beamline 8.3.1. at the Advanced Light Source. Data were processed using XDS[79] and a molecular replacement solution was identified using phaser[80] with a trRosetta model of L056 as a search model. Significant translational non-crystallographic symmetry and differences with the search model resulted in maps that were initially hard to interpret. The initial model was improved using Refmac jelly body refinement[81] using rebuilding using phenix.autobuild[82] and the CCP4 buccaneer_pipeline[83]. The model was finalized and iteratively improved with multiple rounds of manual modification in Coot[84] and refinement using phenix.refine[85]. The model is deposited as PDB accession 7RGR.

Low max ID lysozyme sequence selection, expression and assay

To evaluate whether ProGen can generate low max ID sequences, we generated an additional batch of sequences selected to have maximum sequence identities under 40% with respect to any natural protein. Since we could only test a limited number of proteins in vitro for this experiment, we modified our earlier generation procedure to bias the distribution of generations towards lysozyme families with higher measured functionality in previous experiments. We fine tuned an ensemble of four ProGen models only to lysozymes in PF00959 and PF05838 families. During generation, we used control tags for the two families, as well as control tags to indicate proteins with at least a 30% sequence similarity to L056 and L070, two proteins that we were able to successfully measure catalytic efficiency for in the previous batch. We then used a geometric ensemble of these four models to generate 1 million samples across these control tag settings with varying top-p values. We only kept generations with maximum sequence identities between 20–40%, and ranked these generations using discriminator scores using the same methodology as before, except with a larger 5B parameter discriminator that was pretrained as the T5[86] model, instead of TAPE-BERT. Our final batch included 12 sequences with the PF00959 control tag, 13 with the PF05838 control tag, 20 with the 'L056 similar' control tag, 20 with the 'L070 similar' control tag, 13 across control tags with under 30% maximum sequence identity and 20 sequences from the 1 million generated for the original batch (with 10 at least 30% similar to L056 or L070, and 10 not similar), ranked by both the TAPE-BERT and T5 discriminators.

High-throughput expression testing of low max ID lysozyme variants

Variant sequences were appended with an N-terminal His_6 and HRV3C tagged on their N-termini, codon optimized (VectorBuilder), cloned into a pET vector (VectorBuilder), transformed into BL21(DE3) and shipped from VectorBuilder as a glycerol stock in 96-well block. Variants were inoculated into 1 ml ZYM-5052 autoinduction medium[87] supplemented with 100 µg ml^{-1} carbenicillin in a 96-well deep block, covered with a gas-permeable seal and allowed to grow and expressed by shaking at 37 °C overnight (16 h). High-density expressed cultures were lysed by addition of detergent (Promega Fast Break Cell Lysis Reagent) supplemented with lysis buffer (30 mM HEPES pH 7.6, 150 mM NaCl, 0.5 mM TCEP, cOmplete mini EDTA free protease inhibitor cocktail (Roche), benzonase nuclease) with incubation under gentle shaking for at least 15 min at room temperature before whole expression SDS-PAGE gel samples were taken. Individual wells from 96-well block were transferred to microcentrifuge tubes, centrifuged at 21,000g for 5 min at room temperature, and the soluble fraction was transferred to a new 96-well block for soluble protein SDS-PAGE sample collection.

Expression and purification of lowID lysozyme variants

Variants A5, B6, C9, D4, D10 and E11 were chosen for follow up biochemical characterization on the basis of their high expression and solubility (Supplementary Fig. 10). Variants were inoculated into 50–200 ml ZYM-5052 autoinduction medium[87] supplemented with 100 µg ml^{-1} carbenicillin and allowed to grow and express constructs overnight (16 h) at 37 °C. High-density cell culture was pelleted by centrifugation at 4,000g for 20 min at 4 °C and resuspended to half the total culture volume in 30 mM HEPES pH 7.6, 150 mM NaCl, 0.5 mM TCEP, cOmplete mini EDTA free protease inhibitor cocktail (Roche), benzonase nuclease. Resuspended cells were being lysed by addition of detergent (Promega Fast Break Cell Lysis Reagent) by rotating end-over-end at 4 °C for at least 15 min. Lysate was clarified by centrifugation at 4,000g for 20 min at 4 °C. Clarified lysate was batch bound to 0.5–1 ml dry volume of HisPur NiNTA resin (Thermo Fisher) for 45 min at 4 °C by rotating end-over-end. NiNTA bound variants were purified by either gravity or vacuum flow by washing resin with 75–125 ml 30 mM HEPES pH 7.6, 150 mM NaCl, 0.1 mM TCEP, 25 mM imidazole before eluting with 4 ml 30 mM HEPES pH 7.6, 150 mM NaCl, 0.5 mM TCEP, 250 mM imidazole. His_6 tags were removed through addition of HRV3C protease and cleavage was allow to proceed either at room temperature for 2 h followed by buffer exchange using EconoPac 10 DG desalting columns (BioRad) equilibrated with 30 mM HEPES pH 7.6, 150 mM NaCl, 0.5 mM TCEP or dialyzed overnight at 4 °C against 30 mM HEPES pH 7.6, 150 mM NaCl, 0.5 mM TCEP. If total protein concentration was low, protein was concentrated in 3 kDa molecular weight cutoff Amico centrifugal filters. In-gel Coomassie quantification against HEWL standard curve was performed for all preparations and used for variant enzymology.

Structure prediction methods

To predict structure for the functional artificial sequences, we used AlphaFold2[14] in single-sequence mode (without multiple sequence alignment (MSA) information), with PDB templates, and 12 recycles. We performed structure prediction without an MSA as input so as to not heavily bias the structure prediction toward a known natural mode. The highest ranked predicted structure among five models was used. We attempted structure prediction without templates under varying settings (1–48 recycles) using three different implementations (AlphaFold2 run locally, ColabFold[88] run on Google Colab and ColabFold run locally), however all predictions for our functional artificial sequences yielded unreliable results with predicted local distance difference test (pLDDT) scores below 60.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All sequence databases used in this study are publicly available and include UniprotKB, UniParc, NCBI Taxonomy, Pfam, Uniref30, NCBI nr database and Interpro. Please refer to Supplementary Table 1 for more details. Sequences and activity data for natural and artificial lysozymes tested are in the Supplementary Material. Evaluation data for the CM experiments can be found in Russ et al.[6]. Evaluation data for the MDH experiments can be found in Repecka et al.[52]. The crystal structure datasets generated during the current study are available under PDB accession 7RGR. Source data are provided with this paper.

Code availability

Our code and checkpoints are publicly available on Zenodo and can be reproduced using the details provided in the Methods section on data preparation, model architecture and training protocol. Major components of our model architecture and training protocol can be reproduced using CTRL (https://github.com/salesforce/ctrl). The most updated and supported codebase can be found at https://github.com/salesforce/progen.

References

61. Federhen, S. The NCBI Taxonomy database. *Nucleic Acids Res.* **40**, D136–D143 (2012).
62. Pettit, L. D. The IUPAC stability constants database. *Chem. Int.* **28**, 14–15 (2006).
63. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
64. Bengio, Y., Ducharme, R., Vincent, P. & Janvin, C. A neural probabilistic language model. *J. Mach. Learn. Res.* **3**, 1137–1155 (2003).
65. Madani, A. et al. ProGen: language modeling for protein generation. Preprint at arXiv https://doi.org/10.1101/2020.03.07.982272 (2020).
66. Vig, J. et al. BERTology meets biology: Interpreting attention in protein language models. In International Conference on Learning Representations (ICLR, 2020).
67. Goyal, K., Dyer, C. & Berg-Kirkpatrick, T. Exposing the implicit energy networks behind masked language models via metropolis–hastings. In *10th International Conference on Learning Representations* (ICLR, 2022).
68. Bhattacharya, N. et al. Single layers of attention suffice to predict protein contacts. Preprint at *bioRxiv* https://doi.org/10.1101/2020.12.21.423882 (2020).
69. Ramsauer, H. et al. Hopfield Networks is All You Need. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2008.02217 (2020).
70. Alley, E., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **16**, 1315–1322 (2019).
71. Rao, R. et al. Evaluating protein transfer learning with TAPE. *Adv. Neural Inf. Process. Syst.* **32**, 9689–9701 (2019).
72. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. Preprint *arXiv* https://doi.org/10.48550/arXiv.1412.6980 (2014).
73. Pascanu, R., Mikolov, T. & Bengio, Y. On the difficulty of training recurrent neural networks. In *Proc. 30th International Conference on Machine Learning* (eds. Dasgupta, S. & McAllester, D.) 1310–1318 (PMLR, 2013).
74. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
75. Holtzman, A., Buys, J., Du, L., Forbes, M. & Choi, Y. The curious case of neural text degeneration. In *8th International Conference on Learning Representations* (ICLR, 2020).
76. Goodfellow, I. J. et al. Generative adversarial networks. In *28th Conference on Neural Information Processing Systems* (NIPS, 2014).
77. Koehn, P. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. in *Machine Translation: From Real Users to Research* 115–124 (Springer, 2004).
78. Sun, Z. Z. et al. Protocols for implementing an *Escherichia coli* based TX-TL cell-free expression system for synthetic biology. *J. Vis. Exp.* **16**, e50762 (2013).
79. Kabsch, W. XDS. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 125–132 (2010).
80. McCoy, A. J. et al. Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
81. Kovalevskiy, O., Nicholls, R. A., Long, F., Carlon, A. & Murshudov, G. N. Overview of refinement procedures within REFMAC5: utilizing data from different sources. *Acta Crystallogr D Struct. Biol.* **74**, 215–227 (2018).
82. Terwilliger, T. C. et al. Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. *Acta Crystallogr. D Biol. Crystallogr.* **64**, 61–69 (2008).
83. Hoh, S. W., Burnley, T. & Cowtan, K. Current approaches for automated model building into cryo-EM maps using Buccaneer with CCP-EM. *Acta Crystallogr D Struct. Biol.* **76**, 531–541 (2020).
84. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).
85. Afonine, P. V. et al. Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr. D Biol. Crystallogr.* **68**, 352–367 (2012).
86. Raffel, C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1910.10683 (2019).
87. Studier, F. W. Protein production by auto-induction in high density shaking cultures. *Protein Expr. Purif.* **41**, 207–234 (2005).
88. Mirdita, M. et al. ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).

Acknowledgements

We thank B. McCann, C. Gee, E. Procko, K. Trego, L. Varshney, N. Shirish Keskar and S. Savarese for their feedback at various stages of this project. We thank A. Cook, D. Lo and V. Nemali for operational support. Thanks also to the Salesforce Research Computing Infrastructure team and the Google Cloud TPU team for their help with computing resources, in addition to Twist Bioscience for DNA synthesis support. Beamline 8.3.1 at the Advanced Light Source is operated by the University of California Office of the President, Multicampus Research Programs and Initiatives grant MR-15-328599, the National Institutes of Health (R01 GM124149 and P30 GM124169), Plexxikon Inc. and the Integrated Diffraction Analysis Technologies program of the US Department of Energy Office of Biological and Environmental Research. The Advanced Light Source (Berkeley, CA) is a national user facility operated by Lawrence Berkeley National Laboratory on behalf of the US Department of Energy under contract number DE-AC02-05CH11231, Office of Basic Energy Sciences. Icons in one figure were created using BioRender (https://biorender.com). E.R.G. is supported by NIH F32-GM144982-01. J.S.F. was supported by NIH GM123159, NIH GM145238 and a Sanghvi-Agarwal Innovation Award.

Author contributions

A.M. conceived and designed the study in collaboration with S.S. A.M. and B.K. designed and performed machine learning modeling,

generation and scoring. B.P.M. performed the cell-free expression and activity assay and was supervised by Z.Z.S. E.R.G. performed the cell-based expression and kinetics assay and was supervised by J.S.F. J.M.H., J.L.O., J.S.F. performed the structure determination. A.M., S.S., B.K. and N.N. performed computational analysis, and were advised by C.X. R.S. provided advice on machine learning and computational methods. A.M., J.S.F. and N.N. wrote the manuscript with feedback and contributions from all authors, in particular from E.G. and B.K. N.N. supervised and managed the project.

Competing interests

A.M. is a co-founder of Profluent Bio. All other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-022-01618-2.

Correspondence and requests for materials should be addressed to Ali Madani or Nikhil Naik.

Peer review information *Nature Biotechnology* thanks the anonymous reviewers for their contribution to the peer review of this work.

nature portfolio

Corresponding author(s): Ali Madani, Nikhil Naik

Last updated by author(s): Nov 9, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	mmseqs2 was used for clustering and searching. pytorch 1.6 and tensorflow 1.14 was used to format data for model training. The progen code used in this study can be found at https://zenodo.org/record/7296780 and the newest version can be found in https://github.com/salesforce/progen
Data analysis	alphafold2 and trrosetta2 for predicting structures, pymol 2.4.0 was used for structure visualization and alignment, scikit-learn 0.24.1 and matplotlib 3.3.1 were used for figure creation, tools for structure determination were XDS79, phaser80, Refmac jelly body refinement81, phenix.autobuild82, CCP4 buccaneer_pipeline83, Coot84, phenix.refine85

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All sequence databases used in this study are publicly available and include UniprotKB (https://www.uniprot.org/uniprot/), UniParc (https://www.uniprot.org/uniparc/), NCBI Taxonomy (https://www.ncbi.nlm.nih.gov/taxonomy), Pfam (https://pfam.xfam.org/), Uniref30 (https://www.uniprot.org/uniref/), NCBI nr database (https://www.ncbi.nlm.nih.gov/refseq/about/nonredundantproteins/), and Interpro (https://www.ebi.ac.uk/interpro/). Sequences and activity data for natural and artificial lysozymes tested are in the Supplementary Material. Evaluation data for the chorismate mutase experiments can be found in Russ et al (https://www.science.org/doi/10.1126/science.aba3304). Evaluation data for the malate dehydrogenase experiments can be found in Repecka et al (https://www.nature.com/articles/s42256-021-00310-5). The crystal structure datasets generated during the current study are available in the Protein Data Bank repository, under accession 7RGR (https://www.rcsb.org/structure/7RGR).

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	*Use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design whether sex and/or gender was determined based on self-reporting or assigned and methods used. Provide in the source data disaggregated sex and gender data where this information has been collected, and consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.*
Population characteristics	*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were not chosen based on a predetermined statistical method. It was chosen based on 96 well plates with 6 wells dedicated for controls. Five samples were chosen for cell-based expression. Should be sufficient to support a claim that our method has the capability to engineer highly-active enzymes that are far in sequence space.
Data exclusions	No data was excluded.
Replication	All samples were replicated three times within a trial. Eight samples were replicated as an independent trial for high-throughput activity data by re-performing DNA synthesis, in vitro expression, and activity measurement. The samples characterized in cell-based assay were also present in the cell-free setting. All attempts at replication were successful.
Randomization	Natural samples were selected at random but was ensured to note have 80% sequence identity with any other natural sequence. Artificial sequences were selected in defined sequence identity bins and prioritized by generative model likelihood and adversarial discriminator scores. No two sequences across artificial sequences shared greater than 80% identity overlap.
Blinding	Experimentalists performing synthesis and characterization were blinded until completion of measurement.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging