Cell Systems

Article

ProGen2: Exploring the boundaries of protein language models

Highlights

- The ProGen2 suite of protein language models are scaled to 6.4B parameters
- Models with increased scale better capture the distribution of protein sequences
- ProGen2 models generate novel protein sequences adopting natural folds
- ProGen2 model likelihoods are effective for zero-shot fitness prediction

Authors

Erik Nijkamp, Jeffrey A. Ruffolo, Eli N. Weinstein, Nikhil Naik, Ali Madani

Correspondence

ali@profluent.bio

In brief

The ProGen2 suite of models are scaled up to 6.4B parameters and trained on over one billion sequences from genomic, metagenomic, and immune repertoire datasets. We explore the impact of scale and data distribution on fitting the evolutionary sequence distribution, generating protein sequences, and estimating protein fitness.

Nijkamp et al., 2023, Cell Systems *14*, 968–978
November 15, 2023
https://doi.org/10.1016/j.cels.2023.10.002

CellPress

CellPress

Cell Systems

Article

ProGen2: Exploring the boundaries of protein language models

Erik Nijkamp,[1,5] Jeffrey A. Ruffolo,[2,3,5] Eli N. Weinstein,[4] Nikhil Naik,[1] and Ali Madani[1,3,6,*]
[1]Salesforce Research, Palo Alto, CA, USA
[2]Program in Molecular Biophysics, The Johns Hopkins University, Baltimore, MD, USA
[3]Profluent Bio, Berkeley, CA, USA
[4]Data Science Institute, Columbia University, New York, NY, USA
[5]These authors contributed equally
[6]Lead contact
*Correspondence: ali@profluent.bio
https://doi.org/10.1016/j.cels.2023.10.002

SUMMARY

Attention-based models trained on protein sequences have demonstrated incredible success at classification and generation tasks relevant for artificial-intelligence-driven protein design. However, we lack a sufficient understanding of how very large-scale models and data play a role in effective protein model development. We introduce a suite of protein language models, named ProGen2, that are scaled up to 6.4B parameters and trained on different sequence datasets drawn from over a billion proteins from genomic, metagenomic, and immune repertoire databases. ProGen2 models show state-of-the-art performance in capturing the distribution of observed evolutionary sequences, generating novel viable sequences, and predicting protein fitness without additional fine-tuning. As large model sizes and raw numbers of protein sequences continue to become more widely accessible, our results suggest that a growing emphasis needs to be placed on the data distribution provided to a protein sequence model. Our models and code are open sourced for widespread adoption in protein engineering. A record of this paper's Transparent Peer Review process is included in the supplemental information.

INTRODUCTION

Proteins are the workhorse of life—performing essential and versatile functions that are critical to sustaining human health and the environment. Engineering proteins for our desired purposes enables use-cases in industries across pharmaceuticals, agriculture, specialty chemicals, and fuel. Current tools for protein engineering are limited and, as a consequence, mainly rely on directed evolution,[1] a process of stochastically mutating a starting/wild-type sequence, measuring each variant, and iterating until sufficiently optimized for improved function, also referred to as fitness. Nature as an underlying generative process has yielded a rich, complex distribution of proteins. Due to exponentially broken barriers in DNA sequencing, we now collect natural sequences at a previously unimaginable pace. In parallel, we have seen machine learning models perform exceedingly well at capturing data distributions of images and natural language.[2,3] In particular, the transformer[4] has proven to be a powerful language model and can serve as a universal computation engine[5] across data modalities.

Language modeling tries to capture the notion that some sequences are more likely than others by density estimation. For large language models (LLMs), transformer models equipped with self-attention mechanisms[6] have shown to be particularly well suited to capture dependency among sequence elements while being capable to scale vast amounts of model parameters.[7,8] In this work, we adopt causal LLMs in the form of autoregressive decoders for the modeling of proteins. The raw amino acid sequences, which constitute a protein, are considered as observed sequences for the maximum likelihood-based learning. The problem of conditional protein generation is naturally cast as a next-token prediction task. Specifically, few-shot learning[3] models tasks as autoregressive sampling conditional on a small set of examples (or shots). Notably, LLMs possess the capacity to solve the intended task by increasing the number of parameters without task-specific fine-tuning of the model. These few-shot abilities appear to emerge under certain parameter thresholds,[9] which motivates the exploration of such capabilities for protein engineering.

Methods for generating protein sequences that are functional and have desired properties have recently seen tremendous progress. Simple, traditional methods that leverage multiple sequence alignments of similar proteins, such as ancestral sequence reconstruction,[10] have demonstrated the ability to generate useful proteins but are limited in scope. A host of statistical and machine learning techniques exist to access a larger sequence space. Most still train on a fixed protein family to capture coevolutionary signals present within a set of homologous sequences—ranging from direct

coupling analysis techniques[11] to generative adversarial networks.[12] More versatile models trained on unaligned and unrelated sequences have emerged[13] for functional sequence design. Language models, in particular, provide a powerful architecture to learn from large sets of amino acid sequences across families for the purpose of generating diverse, realistic proteins.[14,15] Sequences generated by protein language models (PLMs) are typically predicted to adopt well-folded structures, despite diverging considerably in sequence space. PLMs can be further focused on specific families of interest by fine-tuning on a subset of relevant proteins. In prior work, fine-tuning the ProGen model on a set of lysozyme families yielded proteins retaining functional behavior and even rivaling that of a natural hen egg white lysozyme.[16] Similar strategies have been employed for domain-specific PLMs, such as the antibody-specific IgLM model.[17] By conditioning on chain type and species-of-origin, IgLM is capable of generating diverse sets of antibodies resembling those of natural immune repertoires.

Understanding the functional effects of sequence mutations is critical for the rational design of proteins. Methods for predicting such effects typically fit into one of two categories: family-specific models trained on aligned sequences or universal models trained on unaligned sequences. Models based on alignments of sequences[18–20] face several key challenges limiting their application to protein engineering tasks. First, for proteins with few evolutionary neighbors, the MSA is likely to be shallow and contain little information about functional constraints. Second, for some families of proteins (such as antibodies), there are many sequences available, but they are non-trivial to align. Finally, evaluation of novel variants requires that new sequences be aligned to the MSA used for training; this can be challenging in cases with large insertions or deletions (indels). These limitations prompted the development of fitness predictors based unaligned sets of sequences, particularly transformer models trained on large databases of protein sequences. ESM-1v[21] tasks a transformer encoder model trained via masked-language modeling with estimating heuristic likelihood of mutations relative to the wild-type sequences. Autoregressive PLMs have also been applied to fitness prediction.[13] These models are intrinsically capable of modeling indels, as well as epistatic mutations. The RITA family of models[22] demonstrated that not only do autoregressive PLMs effectively estimate protein fitness, but performance also be further improved by scaling model capacity. Tranception[23] demonstrated that combining autoregressive language models with retrieval[24] capabilities provides a means of enhancing a generalist model with family-specific information from MSAs at inference.

In this work, we perform a study on the effect of very large-scale models and data. We train a suite of models ranging from 151M to 6.4B parameters (one of the largest published for a single protein transformer) on different datasets collectively totaling 1B protein sequences from genomic, metagenomic, and immune repertoire databases. We analyze the generations from universal and family-specific models through predicted structural and biophysical properties. Finally, we examine fitness prediction on existing experimental datasets, which motivate hypotheses on the role of data distribution and alignment in protein language modeling.

RESULTS

Scaling generative protein language models

Autoregressive language models have proven useful for a variety of protein engineering tasks, including functional sequence generation[16] and protein fitness prediction.[22,23] This class of models originated in natural language processing, where recent trends have shown that larger models are increasingly performant and can acquire emergent capabilities with scale.[3] To further assess the behavior of large-scale PLMs, we have trained a suite of models, ranging from 151M to 6.4B parameters (Figures 1A and 1B; Table S1), called ProGen2.

The ProGen2 models are primarily trained on a universal set of proteins from genomic and metagenomic sources. These sequences are drawn from non-redundant subsets of UniProtKB,[25] clustered at 90% sequence identity (UniRef90[26]), and the BFD metagenomic database,[27] clustered at 30% identity (BFD30). We additionally considered two alternative data distributions for model training. First, we reduced the stringency on metagenomic sequences and train ProGen2-BFD90 on a combination of UniRef90 and BFD90 (BFD clustered at 90% sequence identity). Second, we explored use of immune repertoire sequences (i.e., antibodies) from the observed antibody space (OAS) database[28] for modeling training. The sequences from OAS were clustered at 85% sequence identity and used to train ProGen2-OAS.

To evaluate the ProGen2 models, we first consider the capacity of increasingly large models to capture the distribution of protein sequences. We then explore applications of autoregressive PLMs, focusing on sequence generation (Figure 1E) and fitness ranking (Figure 1F). Through our analyses, we highlight the role and importance of data distributions in training and applying PLMs.

Capturing the distribution of observed proteins

We first evaluate the capacity of ProGen2 to capture the distribution of natural sequences. In particular, we focused on its ability to predict unobserved natural sequences, quantifying performance in terms of perplexity on a held-out test set. We find that larger models yield substantially lower perplexities, consistent with the idea that, despite massive model size, we are far from the overfitting regime (Figures 1C and 1D; Table S2). For a sequence $x = (x_1, x_2, \ldots, x_n)$ of n tokens, the perplexity is calculated as follows:

$$ppl(x) = \exp\left(-\frac{1}{n}\sum_{i=1}^{n}\ln p(x_i)\right)$$
$$= \exp\left(-\frac{1}{n}\sum_{i=1}^{n}\ln(softmax(logits(x)[i])[x_i]\right) \quad \text{(Equation 1)}$$

where $logits()$ maps each token x_i to a vector of logit values under the causal language model p. We report the average perplexity over the held-out partitions of the datasets.

We caution, however, that these results only reflect the capacity of the model to capture the training distribution p_0 from which the data were drawn, not necessarily relevant measures of molecular fitness. To be more precise and borrowing notation from Weinstein et al.,[29] let p^{∞} be the stationary distribution of

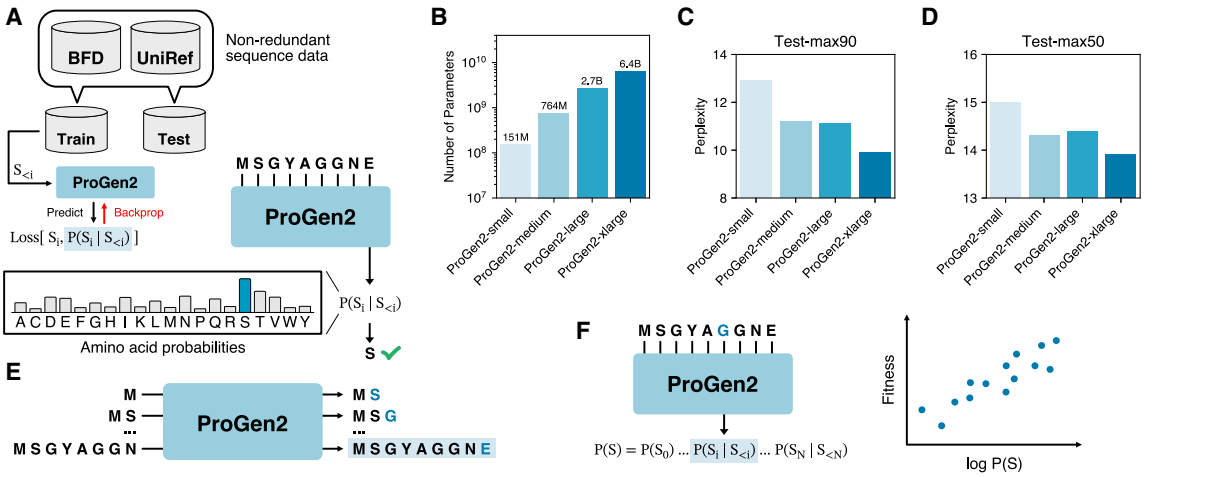

Figure 1. Overview of ProGen2 models for protein sequence generation and scoring

(A) Diagram of model pretraining scheme and autoregressive amino acid prediction.

(B) Number of parameters (log scale) for ProGen2 models.

(C) Perplexity (unitless) for sequences held out from pretraining dataset clustered at 90% sequence identity (Test-max90).

(D) Perplexity (unitless) for sequences held out from pretraining dataset clustered at 50% sequence identity (Test-max50).

(E) Diagram of sequence generation with an autoregressive language model.

(F) Diagram of sequence log likelihood calculation for protein fitness prediction.

the evolutionary process, such that $\log p^{\infty}$ is proportional to log fitness $\log f$. Phylogenetic effects, as well as other imbalances in the dataset, can result in a situation where $p_0 \neq p^{\infty}$; therefore, accurate estimation of the training data distribution p_0 does not necessarily imply accurate estimation of p^{∞} or (consequently) f.

Protein sequence generation

Given the capacity of the ProGen2 family of models for capturing the distribution of observed evolutionary sequences, we next assessed the ability of the models to generate novel sequences. We evaluated sequence generation in three settings: universal protein generation from pretraining, fold-specific generation after fine-tuning, and antibody generation after domain-specific pretraining.

Pretrained models generate diverse protein sequences

Prior work has demonstrated that sequences generated by PLMs can adopt a wide variety of folds, often with considerable deviation in sequence from observed proteins.[14,15] To assess the generative capacity of ProGen2 models, we generated 6,757 sequences with the ProGen2-xlarge model. The three-dimensional structure of each sequence was predicted using ESMFold.[30] For each structure, we identified the most structurally similar natural protein in the PDB[31] using Foldseek.[32] In Figure 2A, we show the relationship between structural similarity to natural proteins (TMscore) and ESMFold prediction confidence (pLDDT). The majority of structures were confidently predicted (median pLDDT of 85.2) and had structural homologs in the PDB (median TMscore of 0.89). Although the generated sequences frequently adopt previously observed folds, they do so with low sequence identity to natural proteins (Figure 2B). In Figure 2C, we show a generated sequence adopting a superhelical fold. The closest structural homolog in the PDB is the central domain of a stress-induced protein (PDB ID 3GCO-A), which aligns with a portion of predicted structure. Despite adopting nearly identical folds, the sequence identity between the generated and natural proteins is only 13.3%. For another generated sequence, with a predicted β-roll structure (Figure 2D), the closest structural match is an isomerase with a truncated β-roll fold (TMscore 0.579). Interestingly, we observe that the generated protein resembles an idealized version of the natural protein, with uniform beta sheets and connecting loops. Figure 2E, we show a generated protein with a ring-like structure. This protein contains similar structural elements to a eukaryotic initiation factor (PDB ID: 3WJ9) but forms a tertiary structure with a greater diameter. In a final example, we show a large generated protein resembling an intracellular transport protein. Despite considerable divergence in sequence space (14.4% identity) over 781 residues, the generated protein is predicted to fold into a well-formed structure. Taken together, these examples illustrate some of the unique properties of sequences generated by ProGen2.

Fine-tuning enables family-specific sequence generation

Next, we considered generation from a model fine-tuned on protein sequences adopting a common structural architecture. The ProGen2-large model was fine-tuned for two epochs on 1M sequences, from Gene3D[33] and CATH,[34] adopting a two-layer sandwich architecture (CATH 3.30). To understand the effects of extended fine-tuning, we generated 30,000 sequences using the model parameters after the first and second epoch of fine-tuning. Sequences were generated using a sweep over sampling temperature and nucleus sampling probability parameters. Sampling temperature effectively reshapes the model's prediction confidence, with values below $T = 1$ sharpening the distribution toward the most confident amino acid predictions. Nucleus sampling removes low-confidence amino acids from the predicted distribution, such that only the minimal set of amino acids contributing to P are selected from. To assess the

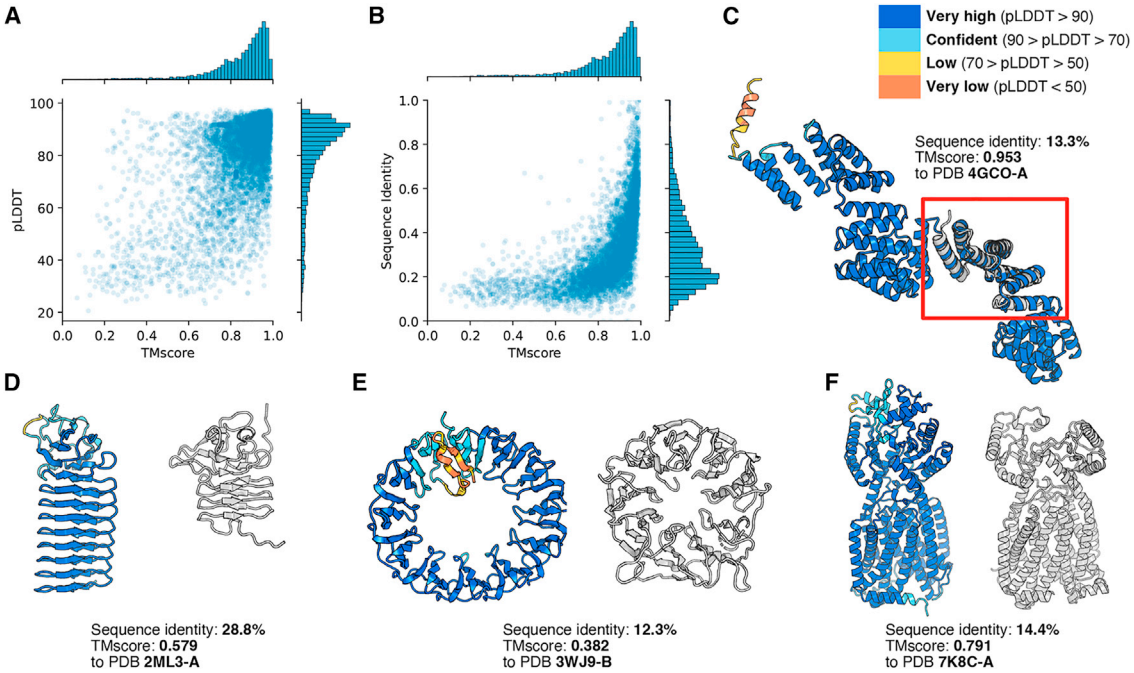

Figure 2. Generating from a pretrained language model trained on a universal protein dataset

Legend in top right corner indicates confidence level (and structural coloring) associated with pLDDT values.

(A) Relationship between ESMFold prediction confidence (pLDDT) and similarity to natural protein structures in the PDB (TMscore) (n = 6,757).

(B) Relationship between sequence identity and similarity to natural protein structures in the PDB (TMscore) (n = 6,757).

(C–F) Comparison of predicted structures for generated sequences (colored by pLDDT) and their closest structural counterparts in the PDB (gray). Sequence identities and TMscores are calculated against the closest structural matches in the PDB.

(C) Superhelical-fold protein generated by the model, with very low sequence identity and high structural similarity to a stress-induced protein domain.

(D) β-roll protein generated by the model, with low sequence identity to the natural protein. The generated protein contains more ordered secondary structure (uniform-length beta sheets, shorter loops) than other beta-roll folds found in the PDB.

(E) Generated protein with similar topology but low sequence identity to a eukaryotic initiation factor. The generated protein adopts a fold with a wider radius than the natural protein.

(F) Generated protein adopting a fold similar to intracellular transporter proteins. The structure is confidently predicted despite low sequence identity and extended length (781 residues).

effect of sampling parameters on structure diversity within the common architecture, we predicted structures for all 60,000 sequences with ESMFold and calculated TMscores and sequence identities against the most similar structures PDB using Foldseek.[32] We chose to validate both sequence identity and structural similarity against the PDB, rather than large sequence databases without known structures, to assess the capacity of ProGen2 to generate diverse sequences adopting natural folds. For both measures, we observed higher similarity to natural proteins with extended fine-tuning (Figures 3A–3D). Among sequences generated with the same model checkpoints, sampling parameters are strongly correlated with sequence novelty (i.e., higher sampling temperature or nucleus probability yields lower sequence identity), as shown in Figures 3A and 3B. When we compared structural diversity, a similar trend emerged, with more restrictive sampling parameters typically yielding structures more closely resembling natural proteins (Figures 3C and 3D). We next analyzed the diversity of generated two-layer sandwich proteins according to their sequence identity and TMscore to the closest structures in the PDB using Foldseek.[32] The vast majority of sequences were confidently predicted to adopt structures similar to natural proteins (median TMscore of 0.85, Figure 3E). As with the sequence generated by pretrained models, the sequences fold into these structures despite considerable deviation in sequence identity (median identity of 23.4%, Figure 3F). Among the more novel structures, the primary source of diversity is in the ligand-binding regions, whereas the non-binding regions resemble natural proteins (Figures 3G–3I). Interestingly, in all cases, the predicted structures present a clear void suitable for a ligand and even mimic the proximal secondary structures of natural proteins. The lower prediction confidence for these regions may be due to the ligand-agnostic nature of the model itself. These results demonstrate that the sequences generated by a fine-tuned model sample diversity at functional regions, while maintaining the common architecture of the training dataset.

Immune repertoire pretraining for antibody sequence generation

Generation of antibody sequences is of particular interest for construction of libraries for therapeutic discovery.[13,17] However, only relatively small generative models have been trained for this task to date. We investigated the properties of antibody sequences generated by ProGen2-OAS, a 764M parameter model pretrained on only natural antibodies. First, we generated 52K

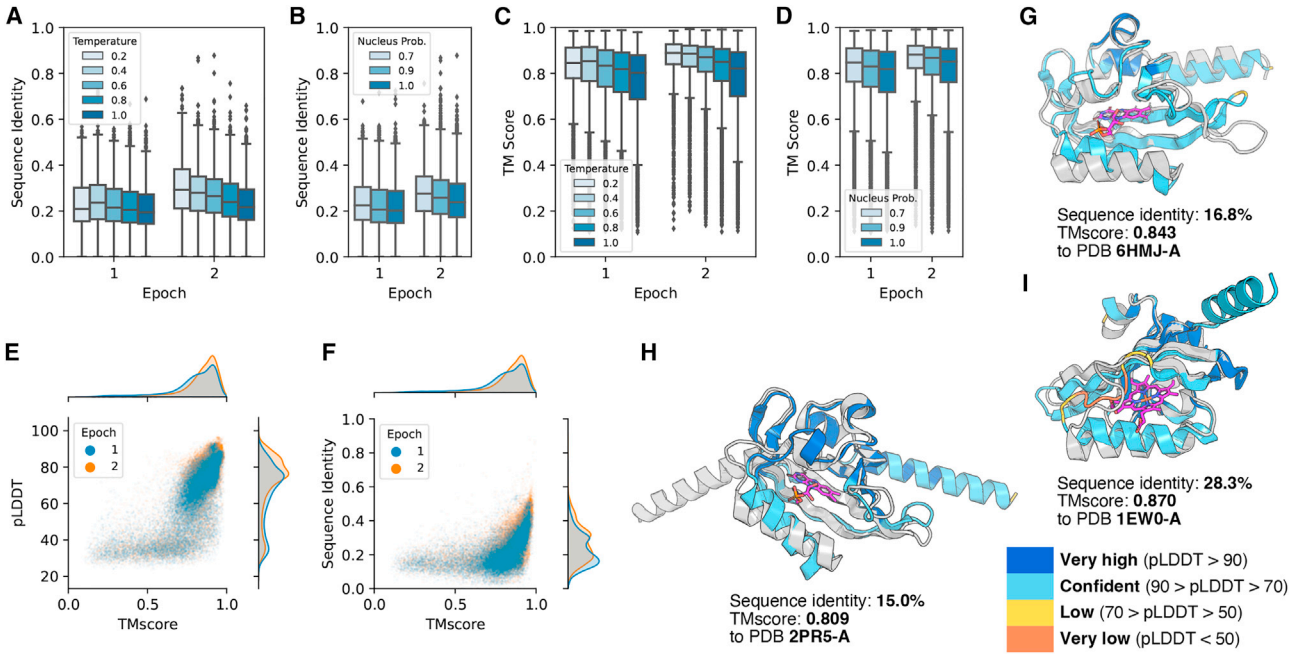

Figure 3. Generating from a language model fine-tuned on two-layer sandwich architecture proteins

Legend in bottom right corner indicates confidence level (and structural coloring) associated with pLDDT values. All box plots have center at median, bounds indicating interquartile range (IQR), whisker length of 1.5× IQR, and points outside of 1.5× IQR range shown as outliers.

(A–D) Effect of fine-tuning duration on the sequential and structural similarity of generated proteins to natural proteins. Extended fine-tuning (two epochs) yields generated sequences more similar to those observed in nature (n = 60,000).

(A) Higher sampling temperature generates more diverse protein sequences.

(B) Higher nucleus sampling probability produces greater sequence diversity.

(C) In general, lower sampling temperature results in sequences adopting structures more similar (higher TMscore) to those found in the PDB.

(D) Lower nucleus sampling probability yields generations with reduced structural diversity.

(E) Relationship between ESMFold prediction confidence (pLDDT) and similarity to natural protein structures in the PDB (TMscore), divided by number of fine-tuning epochs (n = 60,000).

(F) Relationship between sequence identity and similarity to natural protein structures in the PDB (TMscore), divided by number of fine-tuning epochs (n = 60,000).

(G–I) Comparison of predicted structures for sequences generated by the fine-tuned language model (colored by pLDDT) and the most structurally similar proteins in the PDB (transparent). Ligands bound by the natural proteins are shown in pink.

(G) Generated protein adopting a similar fold to a natural protein binding a flavin mononucleotide ligand. The structure of the generated protein closely resembles that of the natural protein near the ligand-binding site, leaving appropriate space available for binding.

(H) Generated protein similar to a natural flavin-mononucleotide-binding protein. The binding site of the generated protein is confidently predicted and reserves appropriate space for the ligand.

(I) Generated protein closely resembling a natural protoporhyrin-binding protein. The structure of the generated protein appears to properly accommodate the ligand but is predicted with low confidence in the unstructured loop regions near the binding site.

non-redundant antibody sequences with the pretrained model. However, experimental limitations of sequencing studies result in over half of antibody sequences in the OAS being truncated at the N termini by 15 or more residues.[35] As such, direct generation from the model yields sequences mirroring the training distribution, rather than fully formed antibody sequences. To overcome this bias in the data and produce full-length antibody sequences, we initiated generation with a three-residue motif commonly found at the beginning of human heavy-chain sequences (EVQ).[17] Using this prompting strategy, we generated an additional 470,000 full-length antibody sequences (Figure 4A). All of the following analyses are based on this full-length heavy-chain-like set.

We next measured the distribution of CDR loop lengths for the generated antibody sequences according to the Chothia numbering scheme.[36] In Figure 4B, we show the length distribution for each CDR loop. CDR1 and CDR2 loops are typically generated with lengths of seven and six residues, respectively, mirroring the biological restriction on length diversity for these loops. For CDR3 loops, which are the most variable due to the insertion of an additional gene segment (D-gene), we observe a wide range of loop lengths (median length of 12 residues). Interestingly, less restrictive (higher) values for both sampling temperature and nucleus probability had the effect of truncating CDR3 loops (Figure 4C). For CDR1 and CDR2, these parameters had little effect, although additional lengths beyond the most frequent for each loop tended to be sampled more often with less restrictive parameters (see outlier points).

Potential antibody therapeutics often require extensive optimization to improve their physical properties. Collectively referred to as developability, these properties include thermal stability, expression, aggregation propensity, and solubility.[37] Here, we focused on quantifying the aggregation propensity and solubility

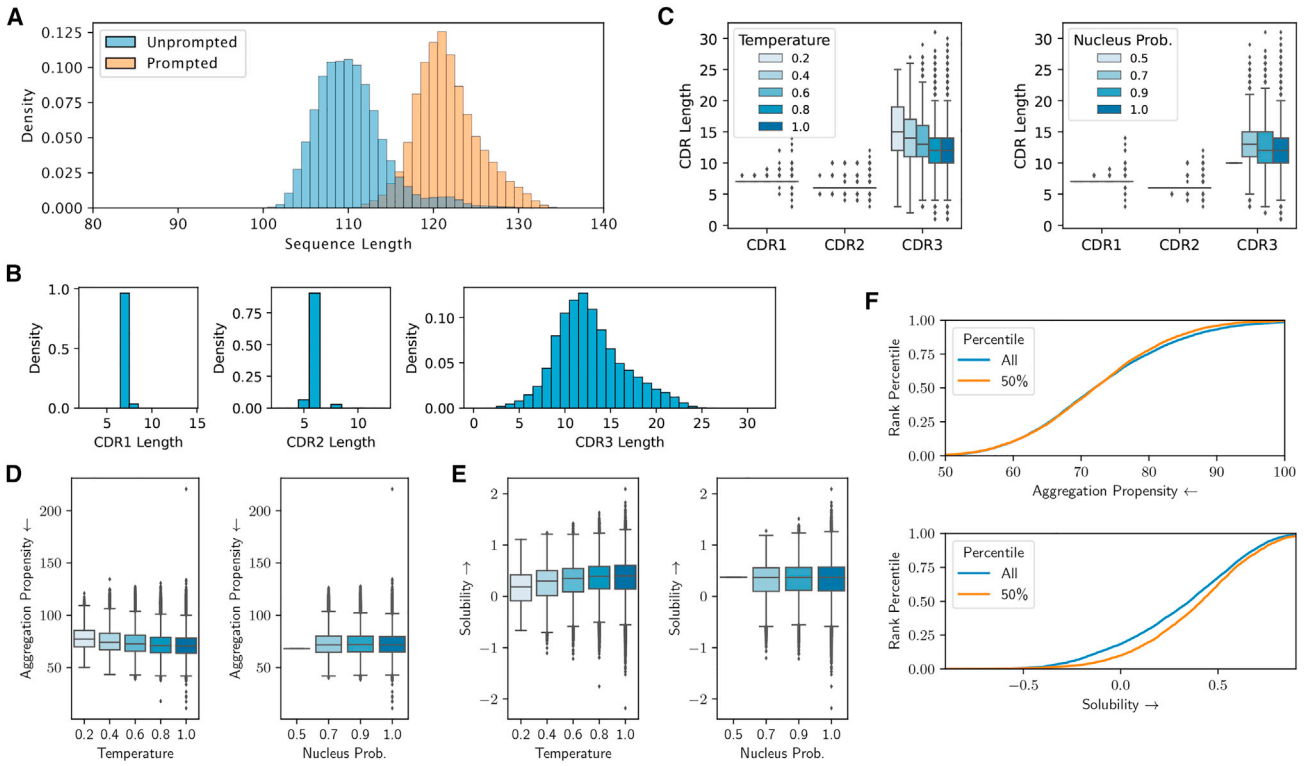

Figure 4. Generating from a pretrained antibody-specific language model

All box plots have center at median, bounds indicating interquartile range (IQR), whisker length of 1.5× IQR, and points outside of 1.5× IQR range shown as outliers.

(A) Comparison of sequence lengths for unprompted and prompted generation strategies. Prompting produces full-length sequences, without N-terminal truncation observed in training dataset.

(B) Distribution of CDR loop lengths (according to Chothia numbering) for generated antibody sequences. CDR1 and CDR2 are predominantly observed to have standard human loop lengths, while a broad range of CDR3 loop lengths are observed (n = 470,000).

(C) Impact of sampling parameters on CDR loop lengths (n = 470,000). Non-standard CDR1 and CDR2 loop lengths are sampled with higher temperature and nucleus probability (see additional outlier points). CDR3 loops tend to be shorter with higher temperature and nucleus probability.

(D) Impact of sampling parameters on aggregation propensity of generated sequences (n = 470,000). Higher sampling temperature results in lower aggregation propensity for generated sequences, whereas changing nucleus probability has limited effect.

(E) Impact of sampling parameters on solubility of sequences (n = 470,000). Higher sampling temperature results in higher solubility for generated sequences, whereas changing nucleus probability has limited effect.

(F) Likelihood ranking of generated antibody sequences with the ProGen2-base language model. Aggregation propensity is not significantly reduced among the top-50%-ranked antibody generations. Solubility is improved by selecting the top 50% of ranked antibody generations.

of generated sequences according to their SAP scores[38] and CamSol-intrinsic profiles.[39] We found that for both aggregation propensity and solubility, sequences generated with less restrictive parameters display improved developability (Figures 4D and 4E). Given the effective zero-shot predictive capabilities of PLMs,[22,23] we also investigated whether a universally pretrained model could be used to filter generated antibody libraries and improve their developability profiles. In Figure 4F, we compare the aggregation propensity and solubility of the full set of generated sequences with the top-50% as scored by the ProGen2-base model. Among the top-ranked sequences, aggregation propensity improves only marginally, whereas the solubility of the sequences shows a favorable shift. These results provide meaningful guidance for generation of antibody sequence libraries with PLMs. In practice, generating with less restrictive sampling parameters and filtering with a universal PLM should provide the most developable set of sequences.

Zero-shot fitness prediction

Generative models for protein sequence design should ideally learn a representation that aligns with our desired functional attributes. Experimental techniques in the wet laboratory have allowed for the collection of protein libraries that associate a given sequence to one or many functional scalar values, which describes a "fitness landscape." We examine how experimentally measured fitness landscapes correlate with a generative model's likelihood in a zero-shot manner, meaning there is no additional fine-tuning in a supervised setting with assay-labeled examples or an unsupervised setting with a focused set of homologous sequences.

Scale does not improve fitness prediction on narrow landscapes

For a proper comparison to Hesslow et al.'s[22] models with a similar architecture to ProGen2, yet trained on a different data distribution, we first characterize zero-shot performance on

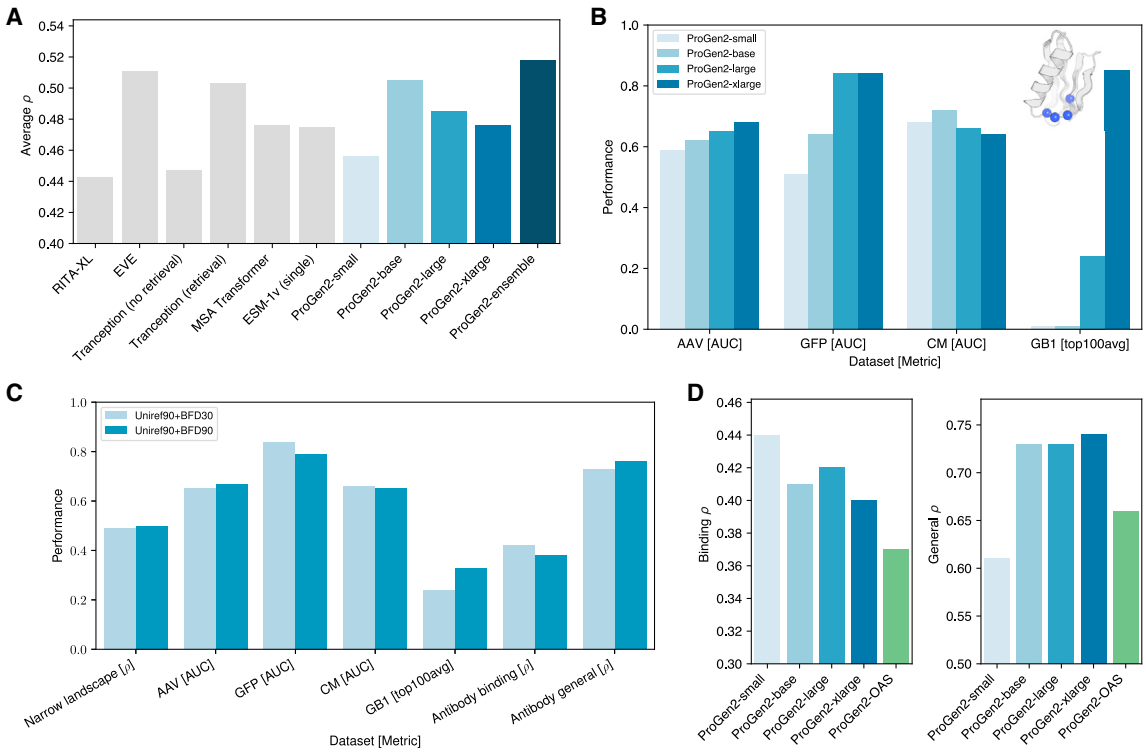

Figure 5. Zero-shot fitness prediction performance

(A) Zero-shot performance of ProGen2 models and alternative methods on narrow fitness landscapes. Model scale provides limited performance benefits and even degrades zero-shot capabilities for the largest models.

(B) Zero-shot performance of ProGen2 models on wide fitness landscapes (units indicated for each bar). Performance typically improves with model scale and may lead to emergent zero-shot capabilities for low-homology, highly epistatic landscapes, such as GB1 (structure with mutation sites shown).

(C) Comparison of zero-shot fitness prediction performance for 2.7B parameter models trained on Uniref90+BFD30 and Uniref90+BFD90 (units indicated for each bar).

(D) Zero-shot performance of universal ProGen2 models and the antibody-specific ProGen2-OAS for binding datasets and general antibody fitness prediction tasks (e.g., stability and expression). Models trained on broad evolutionary sequence datasets outperform antibody-specific models on both tasks.

narrow fitness landscapes from Riesselman et al.,[19] which is composed mainly of single substitution deep mutational scan experiments (Figure 5A; Table 1A). EVE and Tranception (with retrieval), which operate on multiple sequence alignments, perform similarly well to the best-performing single-sequence language models. We observe that our smallest model (ProGen2-small), with an order of magnitude less parameters to RITA-XL, exhibits higher average performance across zero-shot tasks, indicating the importance of pretraining data distributions. In contrast to RITA, the ProGen2 training data are a mixture comprised of an identity-reduced set of sequences from Uniref along with sequences from metagenomic sources. To assess the impact of including metagenomic sequences, we compared the performance of two ProGen2 models trained on varying levels of metagenomic sequence redundancy. Specifically, we compare 2.7B parameter models trained on UniRef90+BFD30 and UniRef90+BFD90. In Figure 5C, we show that inclusion of more metagenomic sequence slightly improves fitness prediction performance on narrow landscapes (Table S5). Our best ProGen2 model outperforms or matches all other baselines spanning a variety of differing modeling strategies, amplifying the importance of understanding what set of sequences are provided to the model for training. Ultimately, an ensemble produced by averaging the scores from ProGen2-medium, ProGen2-large, ProGen2-xlarge, ProGen2-base, and ProGen2-BFD90 for each sequence proved more effective than any individual model.

We find that, as model capacity increases, performance at zero-shot fitness prediction (averaged across all datasets in the narrow landscape) peaks at 764M parameters (ProGen2-base) before decreasing with larger and larger models (Figure 5A). A similar trend is observed for the ESM-2 family of models, which peak in performance at 650M parameters. This stands in contrast to model perplexity, which improves systematically with model scale (Table S2). Our results are in line with Weinstein et al.,[29] where the authors show that when $p_0 \neq p^\infty$, fitness estimates from misspecified models can systematically outperform fitness estimates from well-specified models (even in the limit of infinite data), by projecting the data distribution p_0 onto a model class closer to p^∞ than p_0 itself. Intuitively, this result says that phylogenetic biases and other distortions in the dataset can be partially corrected for by using a relatively small but well-chosen model, which is capable of describing the key features present in real fitness landscapes but is not capable of exactly matching the data distribution. Our results provide evidence that this effect can hold not only in the context of single

Table 1. Zero-shot fitness prediction on experimentally measured landscapes

	Narrow (A)	Wide (B)				Antibody (C)	
Model	Average (ρ)	AAV (AUC)	GFP (AUC)	CM (AUC)	GB1 (top100avg)	Binding (ρ)	General (ρ)
RITA-XL	0.443	–	–	–	–	–	–
EVE	0.511	–	–	–	–	–	–
Tranception (no retrieval)	0.447	–	–	–	–	–	–
Tranception (retrieval)	0.503	–	–	–	–	–	–
MSA transformer	0.476	–	–	–	–	–	–
ESM-1v	0.475	–	–	–	–	–	–
ESM-2 (151M)	0.470	0.523	0.561	0.697	0.218	0.415	0.664
ESM-2 (650M)	0.506	0.588	0.586	0.684	0.850	0.380	0.603
ESM-2 (3B)	0.473	0.512	0.609	0.688	0.552	0.406	0.643
ProGen2-small	0.456	0.585	0.513	0.677	0.009	0.436	0.613
ProGen2-base	0.505	0.615	0.635	0.717	0.005	0.415	0.732
ProGen2-large	0.485	0.652	0.844	0.664	0.242	0.416	0.728
ProGen2-xlarge	0.476	0.678	0.841	0.638	0.846	0.404	0.737
ProGen2-ensemble	0.518	–	–	–	–	–	–
ProGen2-OAS	–	–	–	–	–	0.373	0.659

(A) Performance on narrow experimentally measured fitness landscapes. ProGen2-small outperforms an order of magnitude larger RITA-XL and ProGen2-base is the best-performing ProGen2, indicating larger model capacity does not always translate to improved predictive performance. ProGen2 models outperform or match other baseline methods across a variety of modeling strategies, suggesting the distribution of observed evolutionary sequences provided to the model, along with its inherent biases, likely plays a considerable role. The average spearman is reported with data and baselines provided by Hesslow et al.[22]

(B) Performance on wider experimental landscapes. Larger model capacity may translate to benefits for landscapes involving higher edit distances or low-homology settings. Particularly for GB1 (a low-homology, epistatic landscape), the largest model may demonstrate emergent behavior in finding top-ranked sequences.

(C) Performance on antibody-specific landscapes. Using redundancy-reduced proteins from immune repertoire sequencing studies, OAS,[28] does not lead to better fitness prediction for antibodies. In particular, we examine antibody fitness predictive performance for binding K_D values and general protein properties including expression quality and T_M melting temperatures. The models trained on universal protein databases are better at predicting general properties compared with binding affinity. The binding prediction performance is considerably high given that the associated antigen is not provided to the model.

protein family datasets but also in the context of large-scale datasets containing evolutionarily diverse proteins and using large-scale transformer models.

Scale improves fitness prediction on wide mutational landscapes

Although bigger models may not translate into better zero-shot fitness performance in general, they may still have advantages in certain cases. Most of the available fitness assays to which we compare focus on well-studied proteins with large numbers of evolutionarily similar sequences and measure the fitness/functionality of mutants only one or two mutations away from a wild-type sequence. Intuitively, regions of sequence space with very low probability under p_0 are likely to be especially poorly described with smaller models; therefore, in these regions, both fitness estimation and generation may suffer. Empirically, we find some suggestive evidence that larger models outperform smaller models at fitness estimation in wider landscapes where sequences are farther from any natural sequence (Figure 5B; Table 1B). In particular for the GB1 library, a challenging low-homology protein mutated at positions with non-linear epistasis, our largest models may exhibit emergent behavior[9] in zero-shot identification of the highest fitness variants. We additionally note that training data distribution plays a critical role on wider fitness landscapes, with some landscapes benefiting from inclusion of more metagenomic sequences and others showing signs of performance degradation (Figure 5C; Table S5).

Antibody-specific training does not improve fitness prediction

On antibody-specific landscapes, our results again indicate more attention needs to be placed on the distribution of sequences provided to a model during training. We examine the zero-shot fitness prediction of binding (K_D) and general properties (expression and melting temperature T_M) of antibodies in Table 1C. Samples from immune repertoire sequencing studies seem like an intuitive choice for learning powerful representations useful for antibody fitness prediction tasks.[40,41] However, our ProGen2-OAS model performs poorly compared with pre-trained models trained on universal protein databases (Figure 5D). This is likely reflective of the divergence between the properties of natural antibodies, which must only be sufficiently optimized to circulate in the body and neutralize an antigen, and engineered sequences, which are subject to non-biological pressures throughout production. Further, the types of mutations produced during antibody engineering campaigns (such as deep mutational scans) are unlikely to be observed in natural sequences and thus present an out-of-distribution problem for PLMs trained on immune repertoire data. Curiously, the binding prediction performance of the universal ProGen2 models is

non-negligible and may be useful in practical antibody engineering campaigns, although the corresponding antigen is not provided to the model for likelihood calculation.

DISCUSSION

Protein language models will enable advances in protein engineering and design to solve critical problems for human health and the environment. However, there are many open questions that remain as we begin to realize these advances. In this work, we introduce the ProGen2 suite of models and demonstrate the effectiveness of generative language models for a variety of protein design tasks. Throughout the study, we investigate the impact of increasing model scale for modeling protein sequence landscapes. As model capacity increases, we continue to see improvements in fitting the distribution of natural protein sequences (lower test perplexity). This suggests that current models still under fit the sequence datasets available, and we should expect larger models to deliver further improvements along this axis. Next, we demonstrate the utility of generative language models for creating novel sequences. As shown in prior works,[15] pretrained generative models produce diverse sequences spanning the functional and structural space of natural proteins. Sequences from ProGen2 typically adopt natural folds (as predicted by ESMFold[30]) while diverging in sequence space. Further, we show that fine-tuning ProGen2 models enables a narrowing of the sequence landscape for targeted generation of particular families. Similar approaches have been used to create functional enzymes[16] and are a promising approach for protein design. Finally, we show that the likelihoods learned by LLMs, such as ProGen2, are a useful proxy for protein fitness and are competitive with state-of-the-art methods across a variety of sequence landscapes.

Scaling transformer language models has yielded impressive performance and even emergent capabilities for natural language processing.[3,7] Several studies have investigated whether these scaling trends apply to protein sequence modeling and have typically included that larger models indeed provide improvements across a variety of tasks.[22,30,42] The RITA study found consistent improvements for protein fitness prediction with increasing model capacity up to 1.2B parameters.[22] Similarly, the ESM-2 models (trained for masked-language modeling) were better able to predict protein structure in both unsupervised and supervised settings as model sizes were increased up to 15B parameters. Since their initial release, the ProGen2 models have also been assessed on the extensive ProteinGym benchmark,[23] which is divided between substitution and indel landscapes. This analysis found that greater model scale typically yielded improvements in fitness prediction for both regimes. In contrast to these results, we show that scaling model capacity is not a panacea for all protein design tasks. Although larger ProGen2 models improved zero-shot fitness prediction on broader mutational landscapes, for narrower landscapes composed primarily of amino acid substitutions, we observed a degradation of performance for our largest models. In such cases, models based on multiple sequence alignments,[20,23] which provide detailed context of the local fitness landscape, also performed well for fitness prediction. The test-max50 and wide fitness landscape results suggest that scale may particularly show advantages for out-of-distribution, difficult, or tail-end distribution problems. This is exemplified by the significant advances in zero-shot prediction at larger model scales on the challenging GB1 landscape. Finally, it is worth consideration that fitness as defined as an average spearman across the multiple experimental datasets in this and other studies comes with its own set of biases and may not be the most reliable criteria for evaluation of models for protein engineering. We refer the reader to prior work from Dallago et al.[43] and Yang et al.[44] for further discussion.

Although pretraining on larger sets of sequences would seem to be an intuitive means of creating broadly useful models, our results suggest that the composition of the pretraining dataset is of critical importance. For zero-shot predictions on narrow fitness landscapes, larger ProGen2 models perform relatively poorly despite capturing the pretraining sequence distribution better. This indicates a divergence between the two and could potentially be remedied by identifying a more suitable pretraining corpus. Conversely, for broader mutational landscapes, larger models that better capture the pretraining dataset typically improve zero-shot performance. For the GB1 landscape in particular, pretraining on BFD90 rather than BFD30 yielded considerable improvements at the same model scale. Analysis of ProGen2 and other recent models on the ProteinGym benchmark further highlights the importance of pretraining distribution because fitness prediction improved for sequences with greater numbers of homologs in UniRef100.[23] Perhaps the most distinctive illustration of the importance of dataset-task alignment is the lackluster zero-shot performance of models pretrained on immune repertoire sequences from the OAS. For both binding and general properties of antibody sequences, ProGen2 models pretrained on universal sets of protein sequences (rather than just antibodies) outperformed the model pretrained on antibodies alone (even when having fewer parameters). In the case of antibodies, this may be because the selective pressures on natural antibodies diverge from the properties evaluated experimentally (such as thermal stability and binding affinity), although there may be tasks not explored in this study that are better approached with an antibody-specific model, such as ProGen2-OAS. More broadly, these results suggest that to improve model performance we must carefully consider the alignment of the pretraining dataset and the downstream task.

Ethics statement

Predicting the fitness of a protein sequence and capturing the distribution of natural proteins for generative purposes could be a powerful tool for protein design. If our technique or a future iteration thereof is adopted broadly, care should be taken in terms of the end use-cases of these designed samples and downstream effects to ensure safe, non-nefarious, and ethical applications. For projects in any domain, active oversight during project initiation, experimental optimization, and deployment phases should be put in place to ensure safe usage and limitation of unintended harmful effects.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cels.2023.10.002.

ACKNOWLEDGMENTS

The authors thank the Salesforce Research Computing Infrastructure team and the Google Cloud TPU team for their help with computing resources.

AUTHOR CONTRIBUTIONS

E.N., J.A.R., and A.M. designed and implemented the experiments. E.N., J.A.R., E.N.W., and A.M. analyzed the results. N.N. and A.M. supervised the project. All authors contributed toward writing the manuscript.

DECLARATION OF INTERESTS

J.A.R. and A.M. are employed by Profluent Bio Inc.

Received: January 12, 2023
Revised: May 1, 2023
Accepted: October 2, 2023
Published: October 30, 2023

REFERENCES

1. Arnold, F.H. (1998). Design by directed evolution. Acc. Chem. Res. *31*, 125–131.
2. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., et al. (2022). Photorealistic text-to-image diffusion models with deep language understanding. Preprint at arXiv: 2205.11487.
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. Adv. Neural Inf. Process. Syst. *33*, 1877–1901.
4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. Adv. Neural Inf. Process. Syst. *30*.
5. Lu, K., Grover, A., Abbeel, P., and Mordatch, I. (2021). Pretrained transformers as universal computation engines. Preprint arXiv: 2103.05247.
6. Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. Preprint at arXiv: 1409.0473.
7. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models. Preprint at arXiv: 2001.08361.
8. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D., Hendricks, d.L., L.A., Welbl, J., Clark, A., et al. (2022). Training compute-optimal large language models. Preprint at arXiv: 2203.15556.
9. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. (2022). Emergent abilities of large language models. Preprint at arXiv: 2206.07682.
10. Gumulya, Y., Baek, J.-M., Wun, S.-J., Thomson, R.E.S., Harris, K.L., Hunter, D.J.B., Behrendorff, J.B.Y.H., Kulig, J., Zheng, S., Wu, X., et al. (2018). Engineering highly functional thermostable proteins using ancestral sequence reconstruction. Nat. Cat. *1*, 878–888.
11. Russ, W.P., Figliuzzi, M., Stocker, C., Barrat-Charlaix, P., Socolich, M., Kast, P., Hilvert, D., Monasson, R., Cocco, S., Weigt, M., et al. (2020). An evolution-based model for designing chorismate mutase enzymes. Science *369*, 440–445.
12. Repecka, D., Jauniskis, V., Karpus, L., Rembeza, E., Rokaitis, I., Zrimec, J., Poviloniene, S., Laurynenas, A., Viknander, S., Abuajwa, W., et al. (2021). Expanding functional protein sequence spaces using generative adversarial networks. Nat. Mach. Intell. *3*, 324–333.
13. Shin, J.E., Riesselman, A.J., Kollasch, A.W., McMahon, C., Simon, E., Sander, C., Manglik, A., Kruse, A.C., and Marks, D.S. (2021). Protein design and variant prediction using autoregressive generative models. Nat. Commun. *12*, 2403.
14. Madani, A., McCann, B., Naik, N., Keskar, N.S., Anand, N., Eguchi, R.R., Huang, P.-S., and Socher, R. (2020). Progen: language modeling for protein generation. Preprint at arXiv: 2004.03497.
15. Ferruz, N., Schmidt, S., and Höcker, B. (2022). ProtGPT2 is a deep unsupervised language model for protein design. Nat. Commun. *13*, 4348.
16. Madani, A., Krause, B., Greene, E.R., Subramanian, S., Mohr, B.P., Holton, J.M., Olmos, J.L., Jr., Xiong, C., Sun, Z.Z., Socher, R., et al. (2023). Large language models generate functional protein sequences across diverse families. Nat. Biotechnol. *41*, 1099–1106.
17. Shuai, R.W., Ruffolo, J.A., and Gray, J.J. (2021). Generative language modeling for antibody design. bioRxiv.
18. Hopf, T.A., Ingraham, J.B., Poelwijk, F.J., Schärfe, C.P., Springer, M., Sander, C., and Marks, D.S. (2017). Mutation effects predicted from sequence co-variation. Nat. Biotechnol. *35*, 128–135.
19. Riesselman, A.J., Ingraham, J.B., and Marks, D.S. (2018). Deep generative models of genetic variation capture the effects of mutations. Nat. Methods *15*, 816–822.
20. Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J.K., Brock, K., Gal, Y., and Marks, D.S. (2021). Disease variant prediction with deep generative models of evolutionary data. Nature *599*, 91–95.
21. Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., and Rives, A. (2021). Language models enable zero-shot prediction of the effects of mutations on protein function. Adv. Neural Inf. Process. Syst. *34*, 29287–29303.
22. Hesslow, D., Zanichelli, N., Notin, P., Poli, I., and Marks, D. (2022). RITA: a study on scaling up generative protein sequence models. Preprint at arXiv: 2205.05789.
23. Notin, P., Dias, M., Frazer, J., Hurtado, J.M., Gomez, A.N., Marks, D., and Gal, Y. (2022). Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. In International Conference on Machine Learning, pp. 16990–17017.
24. Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., Van Den Driessche, G.B., Lespiau, J.-B., Damoc, B., Clark, A., et al. (2022). Improving language models by retrieving from trillions of tokens. In International conference on machine learning, pp. 2206–2240.
25. UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. *47*, D506–D515.
26. Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., and Wu, C.H.; UniProt Consortium (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics *31*, 926–932.
27. Steinegger, M., and Söding, J. (2018). Clustering huge protein sequence sets in linear time. Nat. Commun. *9*, 2542.
28. Olsen, T.H., Boyles, F., and Deane, C.M. (2022a). Observed antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. Protein Sci. *31*, 141–146.
29. Weinstein, E., Amin, A., Frazer, J., and Marks, D. (2022). Non-identifiability and the blessings of misspecification in models of molecular fitness. Adv. Neural Inf. Process. Syst. *35*, 5484–5497.

30. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science *379*, 1123–1130.

31. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The protein data bank. Nucleic Acids Res. *28*, 235–242.

32. van Kempen, M., Kim, S.S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C.L.M., Söding, J., and Steinegger, M. (2023). Fast and accurate protein structure search with Foldseek. Nat. Biotechnol. 1–4.

33. Lewis, T.E., Sillitoe, I., Dawson, N., Lam, S.D., Clarke, T., Lee, D., Orengo, C., and Lees, J. (2018). Gene3D: extensive prediction of globular domains in proteins. Nucleic Acids Res. *46*, D435–D439.

34. Sillitoe, I., Bordin, N., Dawson, N., Waman, V.P., Ashford, P., Scholes, H.M., Pang, C.S.M., Woodridge, L., Rauer, C., Sen, N., et al. (2021). CATH: increased structural coverage of functional space. Nucleic Acids Res. *49*, D266–D273.

35. Olsen, T.H., Moal, I.H., and Deane, C.M. (2022b). AbLang: an antibody language model for completing antibody sequences. Bioinform. Adv. *2*, vbac046.

36. Dunbar, J., and Deane, C.M. (2016). ANARCI: antigen receptor numbering and receptor classification. Bioinformatics *32*, 298–300.

37. Raybould, M.I.J., Marks, C., Krawczyk, K., Taddese, B., Nowak, J., Lewis, A.P., Bujotzek, A., Shi, J., and Deane, C.M. (2019). Five computational developability guidelines for therapeutic antibody profiling. Proc. Natl. Acad. Sci. USA. *116*, 4025–4030.

38. Chennamsetty, N., Voynov, V., Kayser, V., Helk, B., and Trout, B.L. (2010). Prediction of aggregation prone regions of therapeutic proteins. J. Phys. Chem. B *114*, 6614–6624.

39. Sormanni, P., Aprile, F.A., and Vendruscolo, M. (2015). The CamSol method of rational design of protein mutants with enhanced solubility. J. Mol. Biol. *427*, 478–490.

40. Leem, J., Mitchell, L.S., Farmery, J.H.R., Barton, J., and Galson, J.D. (2022). Deciphering the language of antibodies using self-supervised learning. Patterns (N Y) *3*, 100513.

41. Ruffolo, J.A., Gray, J.J., and Sulam, J. (2021). Deciphering antibody affinity maturation with language models and weakly supervised learning. Preprint at arXiv: 2112.07782.

42. Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., et al. (2022). Prottrans: toward understanding the language of life through self-supervised learning. IEEE Trans. Pattern Anal. Mach. Intell. *44*, 7112–7127.

43. Dallago, C., Mou, J., Johnston, K.E., Wittmann, B.J., Bhattacharya, N., Goldman, S., Madani, A., and Yang, K.K. (2021). FLIP: benchmark tasks in fitness landscape inference for proteins. bioRxiv.

44. Yang, K.K., Lu, A.X., and Fusi, N.K. (2022). Convolutions are competitive with transformers for protein sequence pretraining. bioRxiv.

45. Su, J., Lu, Y., Pan, S., Wen, B., and Liu, Y. (2021). Roformer: enhanced transformer with rotary position embedding. Preprint at arXiv: 2104.09864.

46. Wang, B., and Komatsuzaki, A. (2021). GPT-J-6B: A 6 billion parameter autoregressive language model. https://github.com/kingoflolz/mesh-transformer-jax.

47. Nijkamp, E., Pang, B., Hayashi, H., Tu, L., Wang, H., Zhou, Y., Savarese, S., and Xiong, C. (2022). A conversational paradigm for program synthesis. Preprint at arXiv: 2203.13474.

48. Bradbury, J., Frostig, R., Hawkins, P., Johnson, M.J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., et al. (2018). JAX: Composable Transformations of Python+NumPy Programs.

49. Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., and Catanzaro, B. (2019). Megatron-lm: training multi-billion parameter language models using model parallelism. Preprint at arXiv: 1909.08053.

50. Kingma, D.P., and Ba, J. (2014). Adam: A method for stochastic optimization. Preprint at arXiv: 1412.6980.

51. Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In International conference on machine learning pp, pp. 1310–1318.

52. Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., Abbeel, P., and Song, Y.S. (2019). Evaluating protein transfer learning with TAPE. Adv. Neural Inf. Process. Syst. *32*, 9689–9701.

53. Koenig, P., Lee, C.V., Walters, B.T., Janakiraman, V., Stinson, J., Patapoff, T.W., and Fuh, G. (2017). Mutational landscape of antibody variable domains reveals a switch modulating the interdomain conformational dynamics and antigen binding. Proc. Natl. Acad. Sci. USA. *114*, E486–E495.

54. Warszawski, S., Borenstein Katz, A., Lipsh, R., Khmelnitsky, L., Ben Nissan, G., Javitt, G., Dym, O., Unger, T., Knop, O., Albeck, S., et al. (2019). Optimizing antibody affinity and stability by the automated design of the variable light-heavy chain interfaces. PLoS Comput. Biol. *15*, e1007207.

55. Hie, B.L., Shanker, V.R., Xu, D., Bruun, T.U.J., Weidenbacher, P.A., Tang, S., Wu, W., Pak, J.E., and Kim, P.S. (2023). Efficient evolution of human antibodies from general protein language models. Nat. Biotechnol. 1–9.

56. Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. *35*, 1026–1028.

57. Ruffolo, J.A., Chu, L.S., Mahajan, S.P., and Gray, J.J. (2023). Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. Nat. Commun. *14*, 2389.

58. Chaudhury, S., Lyskov, S., and Gray, J.J. (2010). PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. Bioinformatics *26*, 689–691.

59. Leaver-Fay, A., Tyka, M., Lewis, S.M., Lange, O.F., Thompson, J., Jacak, R., Kaufman, K.W., Renfrew, P.D., Smith, C.A., Sheffler, W., et al. (2011). ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. In Methods Enzymol, *487* (Elsevier), pp. 545–574.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Generated sequences and fitness data for ProGen2	This paper	https://doi.org/10.5281/zenodo.7877981
Software and algorithms		
Code for ProGen2	This paper	https://doi.org/10.5281/zenodo.8078725

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Ali Madani (ali@profluent.bio).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Generated sequences and fitness prediction data have been deposited at Zenodo and are publicly available as of the date of publication. DOIs are listed in the key resources table.
- All original code has been deposited at Zenodo and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Model

The family of ProGen2 models are autoregressive transformers with next-token prediction language modeling as the learning objective trained in various sizes with 151M, 764M, 2.7B, and 6.4B parameters. Table S1 summarizes the model specifications and choice of hyper-parameters for the optimization such models. We developed and release the library JAXformer (https://github.com/salesforce/jaxformer) for efficient scaling of training with model and data parallelism on TPU. We refer to the supplemental information for details.

The architecture follows a standard transformer decoder with left-to-right causal masking. For the positional encoding, we adopt rotary positional encodings.[45] For the forward pass, we execute the self-attention and feed-forward circuits in parallel for improved communication overhead following,[46] that is, $x_{t+1} = x_t + \text{mlp}(\ln(x_t + \text{attn}(\ln(x_t))))$ is altered to $x_{t+1} = x_t + \text{attn}(\ln(x_t)) + \text{mlp}(\ln(x_t))$ for which the computation of self-attention, $\text{attn}()$, and feed-forward, $\text{mlp}()$, with layer-norm, $\ln()$, is simultaneous.

Table S1 summarizes the model specifications and choice of hyper-parameters for the optimization such models. The choice of the hyper-parameters was informed by Brown et al.[3]; however, the number of layers is reduced with a small number of self-attention heads of relatively high dimensionality to improve overall utilization of the TPU-v3 compute. As explored in Brown et al.,[3] Wang and Komatsuzaki,[46] and Nijkamp et al.,[47] these variations introduce insignificant degradation of perplexity for sufficiently large models, while considerably improving computational efficiency.

Data

The standard ProGen2 models are pretrained on a mixture of Uniref90[26] and BFD30[27] databases. Uniref90 are cluster representative sequences from UniprotKB at 90% sequence identity. The BFD30 dataset is approximately $1/3$ the size of Uniref90, majority from metagenomic sources, commonly not full-length proteins, and clustered at 30% sequence identity. For the ProGen2-BFD90 model, Uniref90 is mixed with representative sequences with at least 3 cluster members after clustering UniprotKB, Metaclust, SRC, and MERC at 90% sequence identity. This BFD90 dataset is approximately twice the size as Uniref90.

To train the antibody-specific ProGen2-OAS, we collected unpaired antibody sequences from the OAS database.[28] OAS is a curated collection of 1.5B antibody sequences from eighty immune repertoire sequencing studies, which contains heavy- and light-chain sequences from six species (humans, mice, rats, camel, rabbit, and rhesus). The sequences in OAS possess a considerable degree of redundancy, due both to discrepancies in the sizes of its constituent studies, as well as the innate biological

redundancy of antibody sequences within organisms. To reduce this redundancy, we clustered the OAS sequences at 85% sequence identity using Linclust,[27] yielding a set of 554M sequences for model training. Alignment coverage in Linclust was calculated with respect to the target sequence ("cov-mode 1"), with all other parameters set to their default values.

All samples are provided to the model with a 1 or 2 character token concatenated at the N-terminal and C-terminal side of the sequence. Each sequence is then provided as-is and flipped. For a given batch, proteins are concatenated with others to fill the maximum token length during training.

Training

The scaling of large language models requires data and model parallelism. Google's TPU-v3 hardware with a high-speed toroidal mesh interconnect naturally allows for efficient parallelism. To efficiently utilize the hardware, the training of the models is implemented in JAX.[48] For parallel evaluation in JAX the *pjit*() (https://jax.readthedocs.io/en/latest/_modules/jax/experimental/pjit.html) operator is adopted. The operator enables a paradigm-named single-program, multiple-data (SPMD) code, which refers to a parallelism technique where the same computation is run on different input data in parallel on different devices (https://jax.readthedocs.io/en/latest/jax-101/06-parallelism.html). Specifically, *pjit*() is the API exposed for the XLA SPMD partitioner in JAX, which allows a given function to be evaluated in parallel with equivalent semantics over a logical mesh of compute.

Our library JAXformer recruits a designated coordinator node to orchestrate the cluster of TPU-VMs with a custom TCP/IP protocol. For data parallelism, the coordinator partitions a batch and distributes the partitions to the individual TPU-VMs. For model parallelism, a partitioning scheme is adopted where parameters are sharded across MXU cores inside a physical TPU-v3 board and replicated across boards following Wang and Komatsuzaki[46] and Shoeybi et al.[49]

For the pretraining of the ProGen2 models, Table S1 summarizes the hyper-parameters. We adopt the Adam[50] optimizer with $(\beta_1, \beta_2, \epsilon) = (0.9, 0.999, 1e-08)$ and global gradient norm clipping[51] of 0.8 and 1.0. The learning-rate function over time follows GPT-3[3] with warm-up steps and cosine annealing.

Notably, the cross-entropy appeared to diverge from the projected power-law relation over time when following standard configurations detailed in Brown et al.[3] In particular, an increasing the global norm of the gradient as an indicator for a divergence from the expected log-log linear behavior of cross-entropy over time was observed. Decreasing the learning rate, increasing weight-decay (or equivalently m_2-regularization under re-parameteriztation) and decreasing the gradient norm clipping factor resulted in a near-constant global norm of the gradient which stabilized training.

For the finetuning of the ProGen2 models, the training is continued from a converged model. The state of the optimizer is re-initialized such Adam's moving averages for the first and second moment estimators are set to zero. The learning rate decay function is adjusted such that initial learning-rate is decreased by a factor of 5. The finetuning covers at most two epochs over the finetuning dataset to avoid over-fitting.

Evaluation data

Two test sets at differing levels of difficulty were constructed to examine language modeling performance. Test-max90 and Test-max50 correspond to representative sequences from held-out clusters from the Uniref90+BFD30 set of sequences at 90% and 50% sequence identity respectively.

To assess zero-shot fitness prediction ability, we evaluate on three sets of experimentally-measured protein landscapes: narrow, wide, and antibody-specific. The narrow landscape set is comprised of the[19] datasets as provided by Hesslow et al.[22] and generally includes variants that are one or two substitutions away from a given wild-type/natural sequence. The wide landscape set involves larger edit distances and are comprised of the Dallago et al.[43] proteins, chorismate mutase proteins from Russ et al.,[11] and the GFP test set proteins from Rao et al.[52]

Lastly, for the antibody-specific landscape, we compiled a dataset consisting of binding, expression, and thermal stability measurements for variants derived from eight distinct antibodies. We collected expression and antigen-binding enrichment measurements for variants of the anti-VEGF g6 antibody from a DMS study.[53] From a second DMS study, we collected binding enrichment measurements for variants of the d44 anti-lysozyme antibody.[54] Binding affinity (K_D) and thermal stability measurements (T_M) for the remaining six antibodies (C143, MEDI8852UCA, MEDI8852, REGN10987, S309, and mAb114) were drawn from a recent study on antibody affinity maturation using pretrained language models.[55] We combined measurements for the mAb114 and mAb114UCA antibodies from the original study into a single fitness dataset because the parent sequences shared high identity.

Sequence generation

To investigate the properties of sequences generated by the ProGen2 family of models, we sampled complete protein sequences in three settings: universal generation after pretraining, fold-specific generation after fine-tuning, and antibody generation after pretraining on only antibody sequences. For universal protein generation, we sampled 6,757 sequences from the ProGen2-xlarge model. A diverse set of sequences was sampled using a Cartesian product of temperature ($T \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$) and nucleus sampling ($P \in \{0.5, 0.7, 0.9, 1.0\}$) parameters. To understand the effects of architecture-specific finetuning on sequence generation, we compared the sequences produced by the ProGen2-large model after one and two epochs of finetuning. Using a similar strategy as for universal protein generation, 30,000 sequences were generated using a Cartesian product of temperature ($T \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$) and nucleus sampling ($P \in \{0.7, 0.9, 1.0\}$) parameters for both model checkpoints. Structures were

predicted for all generated sequences with ESMFold[30] using the default prediction parameters (three recycles). The sequential and structural similarity to known proteins in the PDB was measured with Foldseek.[32]

Antibody sequences were generated using the ProGen2-OAS model after pretraining on a set of variable-fragment sequences from the OAS. We evaluated sequences generated by the model with and without initial-residue prompting. A set of 52K unprompted sequences was generated using sampling parameters from a Cartesian product of temperature ($T \in \{0.2, 0.4, 0.6\}$) and nucleus sampling probability ($P \in \{0.5, 0.7, 0.9, 1.0\}$). An additional 470,000 full-length sequences were generated by initializing the sequence with a three-residue motif commonly observed in human heavy chain antibody sequences (EVQ). Prompted sequences were similarly generated using a Cartesian product of temperature ($T \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$) and nucleus sampling ($P \in \{0.5, 0.7, 0.9, 1.0\}$) parameters. The sequence identity of generated sequences against the training dataset was calculated with MMseqs2.[56] IgFold[57] was used to predict structures for all generated antibody sequences. The full four-model ensemble of IgFold models was used for predictions, with PyRosetta[58] refinement applied to model outputs. To investigate the therapeutic developability of generated antibody sequences, aggregation propensity[38] and solubility[39] were calculated for all sequences. Aggregation propensity was calculated using the predicted structures and the Rosetta[59] implementation of the SAP score tool.[38] Solubility was calculated using the public CamSol web server.[39]