PROGEN2: EXPLORING THE BOUNDARIES OF PROTEIN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

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 finetuning. 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. We release the PROGEN2 models and code at https://github.com/anonymized-research/progen2.

1 INTRODUCTION

Proteins are the workhorse of life – performing essential and versatile functions critical to sustain 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 (Arnold, 1998), 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 (Saharia et al., 2022; Brown et al., 2020). In particular, the transformer (Vaswani et al., 2017) has proven to be a powerful language model and can serve as a universal computation engine (Lu et al., 2021) across data modalities.

Our research moonshot is to build machine learning models that effectively learn from nature and its underlying principles for functional protein engineering and design. Protein language models have shown promise in representation learning for classification, regression, and generation purposes (Rives et al., 2021; Rao et al., 2021; 2019; Madani et al., 2020; Meier et al., 2021; Elnaggar et al., 2020; Brandes et al., 2022; Lin et al., 2022). Along this pursuit, we perform a study on the effect of very large-scale models and data. In short:

- 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 train sets 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.
- We examine fitness prediction on existing experimental datasets which motivate hypotheses on the role of data distribution and alignment in protein language modeling.

2 Related Work

Large-scale language modeling: 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 (Bahdanau et al., 2014) have shown to be particularly well suited to capture dependency among sequence elements while being capable to scale vast amounts of model parameters (Kaplan et al., 2020; Hoffmann et al., 2022). In this work, we adopt causal LLMs in the form of auto-regressive 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 (Brown et al., 2020) models tasks as auto-regressive 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 (Wei et al., 2022), which motivates the exploration of such capabilities for protein engineering.

Protein sequence generation: 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 (Gumulya et al., 2018), 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 (Russ et al., 2020) to generative adversarial networks (Repecka et al., 2021). More versatile models trained on unaligned and unrelated sequences have emerged (Shin et al., 2021) 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 (Madani et al., 2020; Ferruz et al., 2022). Sequences generated by protein language models (PLMs) are typically predicted to adopt well-folded structures, despite diverging significantly in sequence space. PLMs can be further focused on specific families of interest by finetuning on a subset of relevant proteins. In prior work, finetuning 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 (Madani et al., 2021). Similar strategies have been employed for domain-specific PLMs, such as the antibody-specific IgLM model (Shuai et al., 2021). By conditioning on chain type and species-of-origin, IgLM is capable of generating diverse sets of antibodies resembling those of natural immune repertoires.

Protein fitness prediction: 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 (Hopf et al., 2017; Riesselman et al., 2018; Frazer et al., 2021) 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 significant 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 (Meier et al., 2021) 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 (Shin et al., 2021). These models are intrinsically capable of modeling indels, as well as epistatic mutations. The RITA family of models (Hesslow et al., 2022) demonstrated that not only do autoregressive PLMs effectively estimate protein fitness, but performance also be further improved by scaling model capacity. Tranception (Notin et al., 2022) demonstrated that combining autoregressive language models with retrieval (Borgeaud et al., 2021) capabilities provides a means of enhancing a generalist model with family-specific information from MSAs at inference.

Model Name	Parameters (N)	Test-max90 (ppl)	Test-max50 (ppl)
ProGen2-small	151M	12.9	15.0
ProGen2-medium	764M	11.2	14.3
ProGen2-large	2.7B	11.1	14.4
ProGen2-xlarge	6.4B	9.9	13.9

Table 1: Increasing number of parameters allows the model to better capture the distribution of observed evolutionary sequences. Performance is measured as the perplexity of held-out test sequences at various maximum sequence identity thresholds, i.e. test-max50 is more difficult and out-of-distribution.

3 Methods

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 5 summarizes the model specifications and choice of hyperparameters for the optimization such models. We developed and release the library JAXformer (https://github.com/anonymized-research/jaxformer) for efficient scaling of training with model and data parallelism on TPU. We refer to A.1 for details.

Data: The standard ProGen2 models are pretrained on a mixture of Uniref90 (Suzek et al., 2015) and BFD30 (Steinegger & Söding, 2018) 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 Observed Antibody Space (OAS) database (Olsen et al., 2022a). We refer to A.2 for details.

Evaluation: 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 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 finetuning, and antibody generation after pretraining on only antibody sequences. For universal protein generation, we sampled 5K sequences from the ProGen2-xlarge model. To understand the effects of architecture-specific finetuning on sequence generation, we compared 10K sequences produced by the ProGen2-large model after one and two epochs of finetuning. Antibody sequences were generated using the ProGen2-OAS model after pretraining on a set of variable-fragment sequences from the OAS (Olsen et al., 2022a). Sequences were generated using two prompting strategies: unprompted (52K sequences) and initial-residue prompted (470K sequences).

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 Riesselman et al. (2018) datasets as provided by the authors of Hesslow et al. (2022) 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. (2021) proteins, chorismate mutase proteins from Russ et al. (2020), and the GFP test set proteins from Rao et al. (2019). 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 refer to A.3 for details.

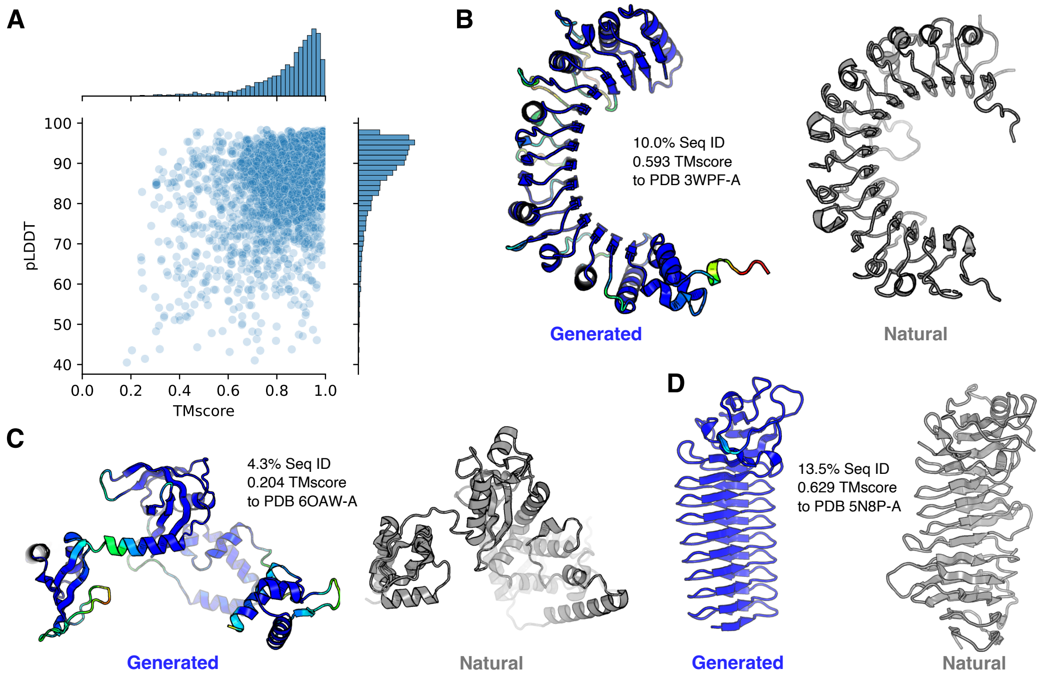

Figure 1: Generating from a pretrained language model trained on a universal protein dataset. (A) Relationship between AlphaFold2 prediction confidence (pLDDT) and similarity to natural protein structures in the PDB (TMscore). (B-D) Comparison of predicted structures for generated sequences (left, colored by pLDDT) and their closest structural counterparts in the PDB (right, gray). Sequence identities and TMscores are calculated against the closest structural matches in the PDB. (B) Solenoid-fold protein generated by the model, with very low sequence identity and high structural similarity to a toll-like receptor protein. The generated protein replaces several alpha helices on the outer edge of the fold with beta sheets, resulting in a looser curvature compared to that of its most similar natural counterpart. (C) Multi-domain α+β-fold generated protein with very low sequential or structural similarity to natural proteins. (D) Generated protein resembling prokaryotic surface protein. The generated protein contains more ordered secondary structure (uniform-length beta sheets, shorter loops) than other beta-roll folds found in the PDB.

4 Results

4.1 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 heldout 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 (Table 1 and Figure 4A). For a sequence $x = (x_1, x_2, \dots, x_n)$ of n tokens, the perplexity is calculated as $ppl(x) = \exp\left(-\frac{1}{n}\sum_{i=1}^{n} \ln p(x_i)\right) = \exp\left(-\frac{1}{n}\sum_{i=1}^{n} \ln(softmax(logits(x)[i]))[x_i]\right)$ 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 was drawn, not necessarily relevant measures of molecular fitness. To be more precise and borrowing notation from Weinstein et al. (2022), let p^∞ be the stationary distribution of 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 $p_0 \neq p^\infty$, and so accurate estimation of the training data distribution p_0 does not necessarily imply accurate estimation of p^∞ or (consequently) f.

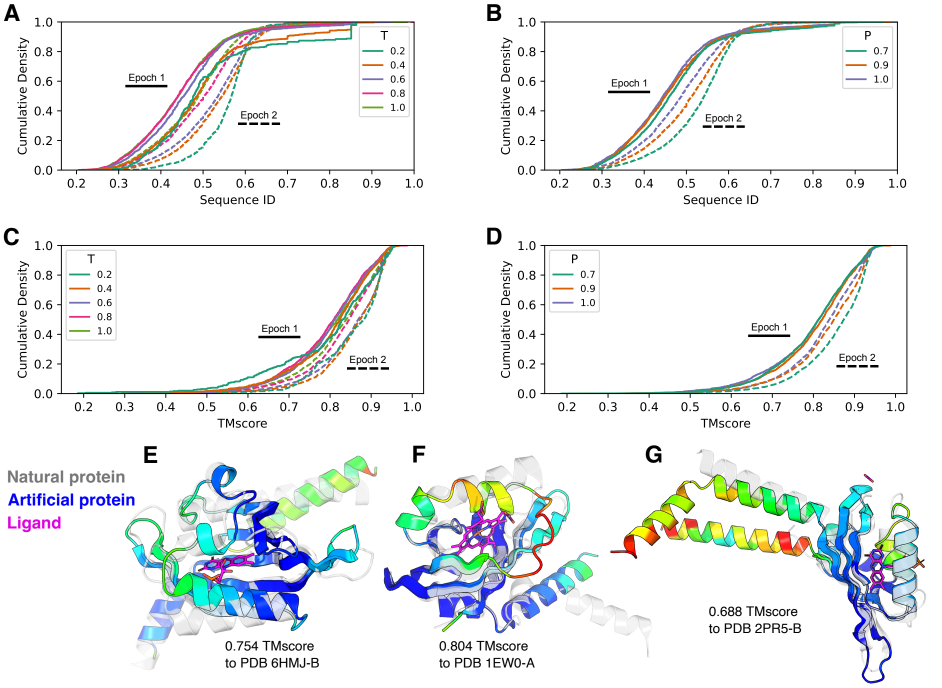

Figure 2: Generating from a language model finetuned on two-layer sandwich architecture proteins. (A-D) Effect of finetuning duration on the sequential and structural similarity of generated proteins to natural proteins. Extended finetuning (two epochs) yields generated sequences more similar to those observed in nature. (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-G) Comparison of predicted structures for sequences generated by the finetuned 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. (E) Generated protein adopting a similar fold to a natural protein binding a flavin mononucleotide ligand. The helical secondary structure of the generated protein matches that of the natural protein near the ligand-binding site, but a shorter loop restricts the space available for binding. (F) 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. (G) 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.

4.2 Generation ability

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 finetuning, and antibody generation after domain-specific pretraining.

Prior work has demonstrated that sequences generated by PLMs can adopt a wide variety of folds, often with significant deviation in sequence from observed proteins (Madani et al., 2020; Ferruz et al., 2022). To assess the generative capacity of ProGen2 models, we generated 5,000 sequences with the ProGen2-xlarge model. The three-dimensional structure of each sequence was predicted using AlphaFold2 (Jumper et al., 2021). For each structure, we identified the most structurally similar natural protein the in the PDB (Berman et al., 2000) using Foldseek (van Kempen et al., 2022). In Figure1, we show the relationship between structural similarity to natural proteins (TMscore) and AlphaFold2 prediction confidence (pLDDT). The majority of structures were confidently predicted (median pLDDT of 90.0) and had structural homologs in the PDB (median TMscore of 0.89). However, closer inspection of predicted structures revealed unique several characteristics of the generated sequences. In Figure1B, we show a generated sequence adopting a solenoid fold. The closest structural homolog in the PDB is the mouse toll-like receptor 9 (PDB ID 3WPF-A), a similarly folding solenoid protein. Interestingly, although the inner face of the generated solendoid fold is composed entirely of beta sheets (as in the natural protein), the outer face combines both alpha helices and beta sheets, resulting in a larger central angle (wider curvature). Further, despite adopting similar folds, the sequence identity between the generated and natural proteins is only 10.0%. For another generated sequence, adopting a multi-domain α+β-fold, no structurally similar proteins were found in the PDB (Figure1C). The most similar natural protein was an uncharacterized protein (PDB ID 6OAW-A), with a low TMscore of 0.204 and little sequence overlap (4.3% identity). In a final case study, we highlight a generated sequence with a predicted structure resembling a prokaryotic RsaA surface protein (PDB ID 5N8P-A). Both structures adopt a similar β-roll fold (TMscore 0.629) yet have a low level of sequence identity (13.5%). Interestingly, we observe that the generated protein resembles an idealized version of the natural protein, with uniform beta sheets and connecting loops. Taken together, these examples illustrate some of the unique properties of sequences generated by PROGEN2. While the generated sequences often fold into structures resembling those produced by nature, they frequently do so with significant sequence deviations, and may adopt novel folds in some cases.

Next, we considered generation from a model finetuned on protein sequences adopting a common structural architecture. The PROGEN2-large model was finetuned for two epochs on 1M sequences, from Gene3D (Lewis et al., 2018) and CATH (Sillitoe et al., 2021), adopting a two-layer sandwich architecture (CATH 3.30). To understand the effects of extended finetuning, we generated 10,000 sequences using the model parameters after the first and second epoch of finetuning. For all generated sequences, we calculated the sequence identity against the training dataset using MMseqs2 (Steinegger & Söding, 2017). As expected, we observed higher similarity to observed evolutionary sequences with extended finetuning (Figure2A-B). 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). To assess the effect of sampling parameters on structure diversity within the common architecture, we predicted structures for all 20,000 sequences with AlphaFold2 and calculated TMscores against the PDB using Foldseek. A similar trend emerged, with more restrictive sampling parameters typically yielding structures more closely resembling natural proteins (Figure2C-D). Among the more novel structures, the primary source of diversity is in the ligand-binding regions, while the non-binding regions resemble natural proteins (Figure2E-G). In two such cases, the ligand-binding region is less confidently predicted by AlphaFold2 and features rearrangements as compared to the closest natural homologs (Figure2E-F). Interestingly, in both 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 could be due to the truncated AlphaFold2 prediction process (one recycle) or the ligand-agnostic nature of the model itself. In another case, the predicted structure of the generated sequence confidently recapitulates the ligand-binding region (Figure2G). These results demonstrate that the sequences generated by a finetuned model sample diversity at functional regions, while maintaining the common architecture of the training dataset.

Generation of antibody sequences is of particular interest for construction of libraries for therapeutic discovery (Shin et al., 2021; Shuai et al., 2021). However, only relatively small generative models have been trained for this task to date. We investigated the properties of antibody sequences generated by a 764M parameter model pretrained on only natural antibodies. First, we generated 52K 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 (Olsen et al., 2022b). 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) (Shuai et al., 2021). Using this prompting strategy, we generated an additional 470K full-length antibody sequences. In Figure3A-B, we compare the sequence similarity of unprompted and prompted generations to the training distribution. Notably, the prompted sequences share significantly greater sequence identity with the training distribution, likely due to the inclusion of the highly conserved FW1 region that is frequently absent in the N-terminally-truncated unprompted sequences. Intrigu-

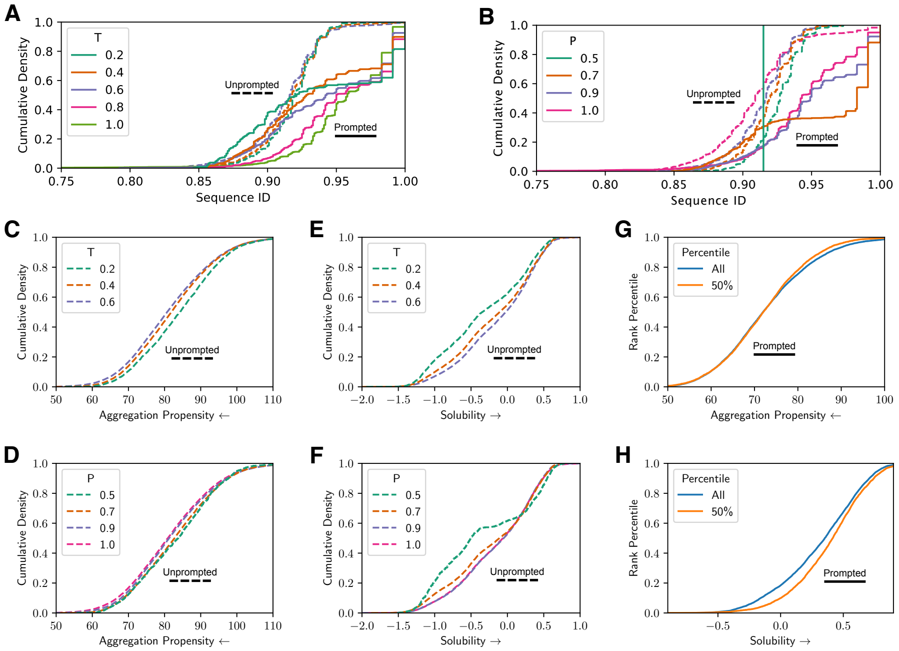

Figure 3: Generating from a pretrained antibody-specific language model. Two strategies were explored for generation of antibody sequences. For unprompted generation, sequences were generated directly by the model without intervention, while for prompted generation the sequence is initialized with three residues (EVQ) commonly observed in human heavy chain antibodies. (A-B) Comparison of sequence identity to the training dataset for unprompted and prompted generations. Prompted generation yields full antibody sequences, resulting in higher sequence identity. (A) Interestingly, higher sampling temperature tends to produce sequences more similar to the training dataset. (B) Lower nucleus sampling probability yields sequences more closely matching the training dataset, as expected. (C-D) Impact of sampling parameters on aggregation propensity of generated sequences. For both temperature sampling (C) and nucleus sampling (D), less restrictive sampling results in lower aggregation propensity for generated sequences. (E-F) Impact of sampling parameters on solubility of sequences. For both temperature sampling (E) and nucleus sampling (F), less restrictive sampling results in more soluble sequences. (G-H) Likelihood ranking of generated antibody sequences with the PROGEN2-base language model. (G) Aggregation propensity is not significantly reduced among the top-50% ranked antibody generations. (H) Solubility is improved by selecting the top 50% of ranked antibody generations.

ingly, we also observe an inverse relationship between more restrictive sampling parameters (lower temperature, higher nucleus probability) and sequence identity to the training dataset.

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 (Raybould et al., 2019). Here, we focused on quantifying the aggregation propensity and solubility of generated sequences according to their SAP scores (Chennamsetty et al., 2010) and CamSol-intrinsic profiles (Sormanni et al., 2015). We found that for both aggregation propensity and solubility, sequences generated with less restrictive parameters display improved developability (Figure3C-F). Given the effective zero-shot predictive capabilities of PLMs (Hesslow et al., 2022; Notin et al., 2022), we also investigated whether a univerally pretrained model could be used to filter generated antibody libraries and improve their developability profiles. In Figure3G-H, 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, while 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.

model	ProGen2- small	ProGen2- base	ProGen2- large	ProGen2- xlarge	ProGen2- ensemble	
average ρ	0.456	**0.505**	0.485	0.476	**0.518**	
model	RITA-XL	EVE	Transception (no retrieval)	Transception (retrieval)	MSA Transformer	ESM-1v (single)
average ρ	0.443	**0.511**	0.447	**0.503**	0.476	0.475

Table 2: Zero-shot fitness prediction on narrow experimentally-measured fitness landscapes. ProGen2-small outperforms an order of magnitude larger RITA-XL and ProGen2-base is the best performing ProGen2 size– 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 significant role. The average spearman is reported with data and baselines provided by Hesslow et al. (2022)

4.3 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 finetuning in a supervised setting with assay-labeled examples or an unsupervised setting with a focused set of homologous sequences.

For a proper comparison to Hesslow et al. (2022)'s models with a similar architecture to ProGen2 yet trained on a different data distribution, we first characterize zero-shot performance on narrow fitness landscapes from Riesselman et al. (2018) which is comprised mainly of single substitution deep mutational scan experiments. We observe in Table 2 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 is a mixture comprised of an identity-reduced set of sequences from Uniref along with sequences from metagenomic sources. 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.

Intriguingly, 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 (ProGen2-large and ProGen2-xlarge). This stands in contrast to model perplexity, which improves systematically with model scale (Table 1, Figure 4A). Our results are in line with Weinstein et al. (2022) 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 the first evidence that this effect can hold not only in the context of single protein family datasets but also in the context of large-scale datasets containing evolutionarily diverse proteins, and using large-scale transformer models.

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, and so 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 se-

dataset [metric]	ProGen2-small	ProGen2-base	ProGen2-large	ProGen2-xlarge
AAV [AUC]	0.59	0.62	0.65	**0.68**
GFP [AUC]	0.51	0.64	**0.84**	**0.84**
CM [AUC]	0.68	**0.72**	0.66	0.64
GB1 [top100avg]	0.01	0.01	0.24	**0.85**

Table 3: Zero-shot fitness prediction 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.

antibody property	ProGen2-small	ProGen2-base	ProGen2-large	ProGen2-xlarge	ProGen2-OAS
binding [avg ρ]	**0.44**	0.41	0.42	0.40	0.37
general [avg ρ]	0.61	0.73	0.73	**0.74**	0.66

Table 4: Zero-shot fitness prediction on antibody-specific landscapes. Using redundancy-reduced proteins from immune repertoire sequencing studies, OAS Olsen et al. (2022a), 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 as compared to binding affinity. Surprisingly, the binding prediction performance is considerably high considering the associated antigen is not provided to the model.

quence, Table 3. 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 (Wei et al., 2022) in zero-shot identification of the highest fitness variants.

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 4. Samples from immune repertoire sequencing studies seem like an intuitive choice for learning powerful representations useful for antibody fitness prediction tasks (Leem et al., 2022; Ruffolo et al., 2021). However, our ProGen2-OAS model performs poorly as compared to pretrained models trained on universal protein databases. Curiously, the binding prediction performance is non-negligible and may be useful in practical antibody engineering campaigns, even though the corresponding antigen is not provided to the model for likelihood calculation.

5 Conclusion

Protein language models may enable advances in protein engineering and design to solve critical problems for human health and the environment. However, there are many open questions for the field that remain. In our study, the results suggest we can continue to scale model size (>6B parameters) and see appropriate improvements in fitting the distribution of natural sequences. Large protein language models can generate libraries of viable sequences that expand the sequence and structural space of natural proteins. 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. However, our other zero-shot fitness prediction results indicate we need better alignment with respect to data distribution and desired functional predictive ability. Simply using raw sequences as they are collected or reducing redundancy through sequence alignment clustering may not be enough. Worth noting, fitness as defined as an average spearman across the multiple experimental wet lab datasets in this study comes with its own set of biases and may not be the only reliable factor for evaluation of models for protein engineering. We refer the reader to (Dallago et al., 2021; Yang et al., 2022) for further discussion. Lastly, we provide our suite of ProGen2 models available to enable AI-driven protein engineering research at https://github.com/anonymized-research/progen2.

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.

REFERENCES

Frances H Arnold. Design by directed evolution. *Accounts of chemical research*, 31(3):125–131, 1998.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1): 235–242, 2000.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. *arXiv preprint arXiv:2112.04426*, 2021.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL `http://github.com/google/jax`.

Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. Proteinbert: A universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Sidhartha Chaudhury, Sergey Lyskov, and Jeffrey J Gray. Pyrosetta: a script-based interface for implementing molecular modeling algorithms using rosetta. *Bioinformatics*, 26(5):689–691, 2010.

Naresh Chennamsetty, Vladimir Voynov, Veysel Kayser, Bernhard Helk, and Bernhardt L Trout. Prediction of aggregation prone regions of therapeutic proteins. *The Journal of Physical Chemistry B*, 114(19):6614–6624, 2010.

Christian Dallago, Jody Mou, Kadina E Johnston, Bruce J Wittmann, Nicholas Bhattacharya, Samuel Goldman, Ali Madani, and Kevin K Yang. Flip: Benchmark tasks in fitness landscape inference for proteins. *bioRxiv*, 2021.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing. *arXiv preprint arXiv:2007.06225*, 2020.

Noelia Ferruz, Steffen Schmidt, and Birte Höcker. A deep unsupervised language model for protein design. *bioRxiv*, 2022.

Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K Min, Kelly Brock, Yarin Gal, and Debora S Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, 2021.

Yosephin Gumulya, Jong-Min Baek, Shun-Jie Wun, Raine ES Thomson, Kurt L Harris, Dominic JB Hunter, James BYH Behrendorff, Justyna Kulig, Shan Zheng, Xueming Wu, et al. Engineering highly functional thermostable proteins using ancestral sequence reconstruction. *Nature Catalysis*, 1(11):878–888, 2018.

Daniel Hesslow, Niccoló Zanichelli, Pascal Notin, Iacopo Poli, and Debora Marks. Rita: a study on scaling up generative protein sequence models. *arXiv preprint arXiv:2205.05789*, 2022.

Brian L Hie, Duo Xu, Varun R Shanker, Theodora UJ Bruun, Payton A Weidenbacher, Shaogeng Tang, and Peter S Kim. Efficient evolution of human antibodies from general protein language models and sequence information alone. *bioRxiv*, 2022.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Thomas A Hopf, John B Ingraham, Frank J Poelwijk, Charlotta PI Schärfe, Michael Springer, Chris Sander, and Debora S Marks. Mutation effects predicted from sequence co-variation. *Nature biotechnology*, 35(2):128–135, 2017.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015. URL http://arxiv.org/abs/1412.6980.

Patrick Koenig, Chingwei V Lee, Benjamin T Walters, Vasantharajan Janakiraman, Jeremy Stinson, Thomas W Patapoff, and Germaine Fuh. Mutational landscape of antibody variable domains reveals a switch modulating the interdomain conformational dynamics and antigen binding. *Proceedings of the National Academy of Sciences*, 114(4):E486–E495, 2017.

Jinwoo Leem, Laura S Mitchell, James HR Farmery, Justin Barton, and Jacob D Galson. Deciphering the language of antibodies using self-supervised learning. *Patterns*, pp. 100513, 2022.

Tony E Lewis, Ian Sillitoe, Natalie Dawson, Su Datt Lam, Tristan Clarke, David Lee, Christine Orengo, and Jonathan Lees. Gene3d: extensive prediction of globular domains in proteins. *Nucleic acids research*, 46(D1):D435–D439, 2018.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch. Pretrained transformers as universal computation engines. *arXiv preprint arXiv:2103.05247*, 2021.

Ali Madani, Bryan McCann, Nikhil Naik, Nitish Shirish Keskar, Namrata Anand, Raphael R Eguchi, Po-Ssu Huang, and Richard Socher. Progen: Language modeling for protein generation. *arXiv preprint arXiv:2004.03497*, 2020.

Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. Deep neural language modeling enables functional protein generation across families. *bioRxiv*, 2021.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems*, 34:29287–29303, 2021.

Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. Colabfold: making protein folding accessible to all. *Nature Methods*, pp. 1–4, 2022.

Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. A conversational paradigm for program synthesis. *arXiv preprint arXiv:2203.13474*, 2022.

Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena-Hurtado, Aidan Gomez, Debora S Marks, and Yarin Gal. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. *arXiv preprint arXiv:2205.13760*, 2022.

Tobias H Olsen, Fergus Boyles, and Charlotte M Deane. Observed antibody space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Science*, 31(1):141–146, 2022a.

Tobias H Olsen, Iain H Moal, and Charlotte M Deane. Ablang: An antibody language model for completing antibody sequences. *bioRxiv*, 2022b.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pp. 1310–1318. PMLR, 2013.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32, 2019.

Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. In *International Conference on Machine Learning*, pp. 8844–8856. PMLR, 2021.

Matthew IJ Raybould, Claire Marks, Konrad Krawczyk, Bruck Taddese, Jaroslaw Nowak, Alan P Lewis, Alexander Bujotzek, Jiye Shi, and Charlotte M Deane. Five computational developability guidelines for therapeutic antibody profiling. *Proceedings of the National Academy of Sciences*, 116(10):4025–4030, 2019.

Donatas Repecka, Vykintas Jauniskis, Laurynas Karpus, Elzbieta Rembeza, Irmantas Rokaitis, Jan Zrimec, Simona Poviloniene, Audrius Laurynenas, Sandra Viknander, Wissam Abuajwa, et al. Expanding functional protein sequence spaces using generative adversarial networks. *Nature Machine Intelligence*, 3(4):324–333, 2021.

Adam J Riesselman, John B Ingraham, and Debora S Marks. Deep generative models of genetic variation capture the effects of mutations. *Nature methods*, 15(10):816–822, 2018.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

Jeffrey A Ruffolo, Jeffrey J Gray, and Jeremias Sulam. Deciphering antibody affinity maturation with language models and weakly supervised learning. *arXiv preprint arXiv:2112.07782*, 2021.

Jeffrey A Ruffolo, Lee-Shin Chu, Sai Pooja Mahajan, and Jeffrey J Gray. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *bioRxiv*, 2022.

William P Russ, Matteo Figliuzzi, Christian Stocker, Pierre Barrat-Charlaix, Michael Socolich, Peter Kast, Donald Hilvert, Remi Monasson, Simona Cocco, Martin Weigt, et al. An evolution-based model for designing chorismate mutase enzymes. *Science*, 369(6502):440–445, 2020.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

Jung-Eun Shin, Adam J Riesselman, Aaron W Kollasch, Conor McMahon, Elana Simon, Chris Sander, Aashish Manglik, Andrew C Kruse, and Debora S Marks. Protein design and variant prediction using autoregressive generative models. *Nature communications*, 12(1):1–11, 2021.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

Richard W Shuai, Jeffrey A Ruffolo, and Jeffrey J Gray. Generative language modeling for antibody design. *bioRxiv*, 2021.

Ian Sillitoe, Nicola Bordin, Natalie Dawson, Vaishali P Waman, Paul Ashford, Harry M Scholes, Camilla SM Pang, Laurel Woodridge, Clemens Rauer, Neeladri Sen, et al. Cath: increased structural coverage of functional space. *Nucleic acids research*, 49(D1):D266–D273, 2021.

Pietro Sormanni, Francesco A Aprile, and Michele Vendruscolo. The camsol method of rational design of protein mutants with enhanced solubility. *Journal of molecular biology*, 427(2):478–490, 2015.

Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

Martin Steinegger and Johannes Söding. Clustering huge protein sequence sets in linear time. *Nature communications*, 9(1):1–8, 2018.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

Baris E Suzek, Yuqi Wang, Hongzhan Huang, Peter B McGarvey, Cathy H Wu, and UniProt Consortium. Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 2015.

Michel van Kempen, Stephanie Kim, Charlotte Tumescheit, Milot Mirdita, Johannes Söding, and Martin Steinegger. Foldseek: fast and accurate protein structure search. *bioRxiv*, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax`, May 2021.

Shira Warszawski, Aliza Borenstein Katz, Rosalie Lipsh, Lev Khmelnitsky, Gili Ben Nissan, Gabriel Javitt, Orly Dym, Tamar Unger, Orli Knop, Shira Albeck, et al. Optimizing antibody affinity and stability by the automated design of the variable light-heavy chain interfaces. *PLoS computational biology*, 15(8):e1007207, 2019.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Eli N Weinstein, Alan N Amin, Jonathan Frazer, and Debora S Marks. Non-identifiability and the blessings of misspecification in models of molecular fitness and phylogeny. *bioRxiv*, 2022.

Kevin K Yang, Alex X Lu, and Nicolo K Fusi. Convolutions are competitive with transformers for protein sequence pretraining. *bioRxiv*, 2022.

Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.

A APPENDIX

A.1 MODEL PARAMETERS

Our models are autoregressive transformers with next-token prediction language modeling as the learning objective. The family of PROGEN2 models is trained in various sizes with 151M, 764M, 2.7B, and 6.4B parameters.

The architecture follows a standard transformer decoder with left-to-right causal masking. For the positional encoding, we adopt rotary positional encodings (Su et al., 2021). For the forward pass, we execute the self-attention and feed-forward circuits in parallel for improved communication overhead following (Wang & Komatsuzaki, 2021), that is, $x_{t+1} = x_t + \mathrm{mlp}(\mathrm{ln}(x_t + \mathrm{attn}(\mathrm{ln}(x_t))))$ is altered to $x_{t+1} = x_t + \mathrm{attn}(\mathrm{ln}(x_t)) + \mathrm{mlp}(\mathrm{ln}(x_t))$ for which the computation of self-attention, $\mathrm{attn}()$, and feed-forward, $\mathrm{mlp}()$, with layer-norm, $\mathrm{ln}()$, is simultaneous.

Table 5 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. (2020), 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. (2020); Wang & Komatsuzaki (2021); Nijkamp et al. (2022), these variations introduce insignificant degradation of perplexity for sufficiently large models, while significantly improving computational efficiency.

	Model				
Hyper-parameter	PROGEN2-small	PROGEN2-medium	PROGEN2-base	PROGEN2-large	PROGEN2-xlarge
Number of params	151M	764M	764M	2.7B	6.4B
Number of layers	12	27	27	32	32
Number of heads	16	16	16	32	16
Head dimensions	64	96	96	80	256
Context length	1,024	1,024	2,048	1,024	1,024
Batch size	500k	500k	500k	500k	1M
Learning rate	6.0e-4	2.5e-4	2.0e-4	0.8e-4	0.1e-4
Weight decay	0.1	0.1	0.1	0.1	0.1
Grad norm clip	1.0	1.0	0.8	0.8	0.8
Warm-up steps	3,000	3,000	10,000	10,000	10,000
Total steps	350,000	350,000	400,000	400,000	350,000

Table 5: Choice of hyper-parameters for model specification and optimization for the family of PROGEN2 causal language models for protein engineering.

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 Bradbury et al. (2018). For parallel evaluation in JAX the $pjit()$[1] 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.[2] 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 Shoeybi et al. (2019); Wang & Komatsuzaki (2021).

[1] https://jax.readthedocs.io/en/latest/_modules/jax/experimental/pjit.html

[2] https://jax.readthedocs.io/en/latest/jax-101/06-parallelism.html

For the pre-training of the ProGen2 models, Table 5 summarizes the hyper-parameters. We adopt the Adam (Kingma & Ba, 2015) optimizer with $(\beta_1, \beta_2, \epsilon) = (0.9, 0.999, 1\text{e}{-}08)$ and global gradient norm clipping (Pascanu et al., 2013) of 0.8 and 1.0. The learning rate function over time follows GPT-3 (Brown et al., 2020) 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. (2020). 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 ℓ_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 fine-tuning 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 fine-tuning covers at most two epochs over the fine-tuning dataset to avoid over-fitting.

A.2 Training Data

The standard ProGen2 models are pretrained on a mixture of Uniref90 (Suzek et al., 2015) and BFD30 (Steinegger & Söding, 2018) 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 Observed Antibody Space (OAS) database (Olsen et al., 2022a). 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 significant 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 (Steinegger & Söding, 2018), 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.

A.3 Evaluation Methods

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 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 finetuning, and antibody generation after pretraining on only antibody sequences. For universal protein generation, we sampled 5,000 sequences from the ProGen2-xlarge model. To understand the effects of architecture-specific finetuning on sequence generation, we compared 10,000 sequences produced by the ProGen2-large model after one and two epochs of finetuning. In both generation settings, we varied the sampling temperature and nucleus sampling probability to produce a diverse set of sequences. Structures were predicted for a subset of generated sequences using AlphaFold2 (Jumper et al., 2021), and the similarity to known structures in the PDB was measured with Foldseek (van Kempen et al., 2022).

Antibody sequences were generated using the PROGEN2-OAS model after pretraining on a set of variable-fragment sequences from the OAS (Olsen et al., 2022a). Sequences were generated using two prompting strategies: unprompted (52K sequences) and initial-residue prompted (470K sequences). For initial-residue prompting, we began generation with a three-residue sequence motif commonly observed in human heavy chain sequences (EVQ). For both prompting strategies, we generate a diverse set of sequences by varying the sampling temperature and nucleus sampling probability. Structures for all generated antibody sequences were predicted using IgFold (Ruffolo et al., 2022). To investigate the therapeutic developability of generated antibody sequences, aggregagation propensity (Chennamsetty et al., 2010) and solubility (Sormanni et al., 2015) were calculated for all sequences.

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 Riesselman et al. (2018) datasets as provided by the authors of Hesslow et al. (2022) 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. (2021) proteins, chorismate mutase proteins from Russ et al. (2020), and the GFP test set proteins from Rao et al. (2019).

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 (Koenig et al., 2017). From a second DMS study, we collected binding enrichment measurements for variants of the d44 anti-lysozyme antibody (Warszawski et al., 2019). 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 (Hie et al., 2022). We combined measurements for the mAb114 and mAb114UCA antibodies from the original study into a single fitness dataset because the parent sequences shared significant overlap.

A.4 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 finetuning, and antibody generation after pretraining on only antibody sequences. For universal protein generation, we sampled 5,000 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, 10,000 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. The structures of all generated sequences were predicted with AlphaFold2 (Jumper et al., 2021). For universal generations from the pretrained model, structures were predicted using ColabFold (Mirdita et al., 2022) with twelve recycles (other parameters set to their default values). For generations after finetuning, structures were predicted using DeepMind's implementation of AlphaFold2 with single-sequence inputs (no MSAs), structural templates from the PDB (Berman et al., 2000), and only one recycle. All structures were predicted with the full five-model ensemble (using the pTM models) and the top-ranked structures for each sequence were considered for structural analysis. Similarity of predicted structures to observed proteins in the PDB was measured by calculating the TMscore (Zhang & Skolnick, 2004) using Foldseek (van Kempen et al., 2022). For universal generations, we report the sequence identity against the most structurally similar protein reported by Foldseek. For finetuned generations, we calculated the sequence identity against the finetuning dataset using MMseqs2 (Steinegger & Söding, 2017).

Antibody sequences were generated using the PROGEN2-OAS model after pretraining on a set of variable-fragment sequences from the OAS. We evalauted 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 470K 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 (Steinegger & Söding, 2017). IgFold (Ruffolo et al., 2022) was used to predict structures for all generated antibody sequences. The full four-model ensemble of IgFold models was used for predictions, with PyRosetta (Chaudhury et al., 2010) refinement applied to model outputs. Aggregation propensities of generated sequences were measured by calculating the SAP score (Chennamsetty et al., 2010) of the predicted structures. Solubility profiles were calculated based on sequence using the public CamSol-intrinsic (Sormanni et al., 2015) web server.

A.5 Zero-shot Prediction Results

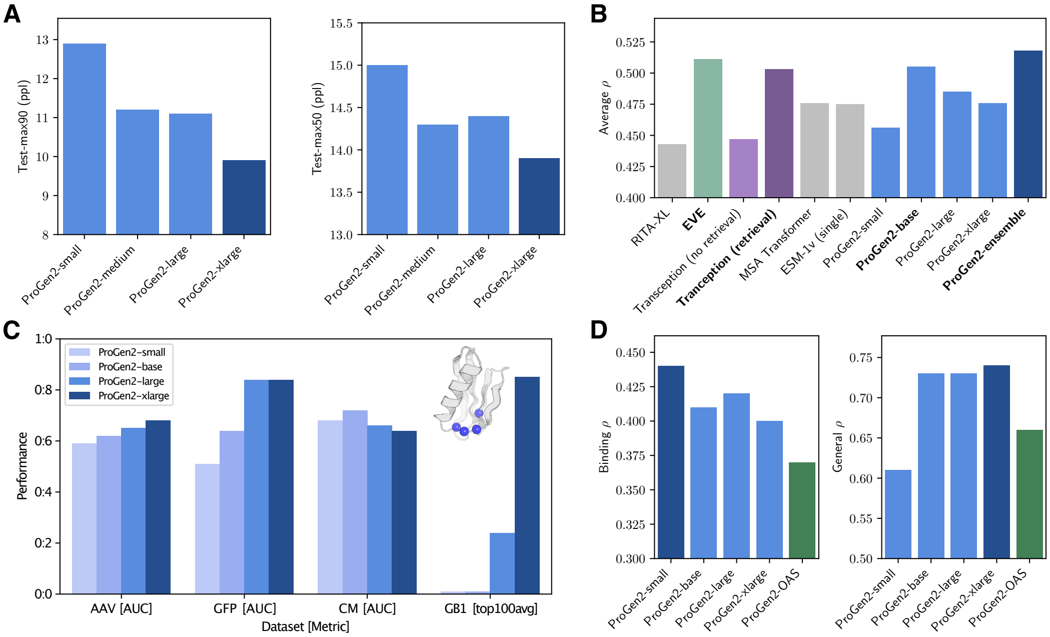

Figure 4: Test perplexity and zero-shot performance. (A) Performance of models on held-out test sets measured by perplexity at various maximum sequence identities. Larger models better capture the distribution of evolutionary sequences, as indicated by lower perplexity values. (B) Zero-shot performance 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. (C) Zero-shot performance of ProGen2 models on wide fitness landscapes. Performance typically improves with model scale, and may lead to emergent zero-shot capabilities for low-homology, highly epistatic landscapes like GB1 (structure with mutation sites shown). (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.

dataset	ProGen2-base	tranception	tranception (retrieval)	wavenet
A0A1J4YT16_9PROT_Davidi_2020	**0.195**	0.178	**0.191**	0.117
B1LPA6_ECOSM_Russ_2020	**0.405**	0.321	**0.415**	0.385
BLAT_ECOLX_indels	**0.664**	0.296	0.357	0.546
PTEN_HUMAN_Mighell_2018_deletions	0.641	0.563	0.598	**0.699**
CAPSD_AAV2S_Sinai_2021_indels	0.392	0.549	**0.586**	0.457
HIS7_YEAST_Pokusaeva_2019_indels	**0.702**	**0.707**	**0.692**	0.68
P53_HUMAN_Kotler_2018_deletions	**0.398**	**0.395**	**0.401**	0.001
AVERAGE	**0.485**	0.430	0.463	0.412

Table 6: Zero-shot fitness prediction on experimental studies evaluating indels collected by Notin et al. (2022). ProGen2 outperforms baselines including models with inference-time retrieval.

antibody property	ProGen2-small	ProGen2-base	ProGen2-large	ProGen2-xlarge	ProGen2-OAS
g6_bind	**0.164**	**0.153**	**0.152**	0.135	0.032
g6_exp	**0.411**	0.386	0.368	0.352	0.184
d44_bind	0.290	**0.426**	**0.421**	0.366	0.260
C143_Kd	0.162	**0.197**	0.132	0.085	0.006
C143_Tm	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
MEDI8852UCA_Kd	0.466	**0.913**	**0.916**	**0.902**	0.798
MEDI8852UCA_Tm	0.314	**0.829**	**0.829**	**0.829**	0.771
MEDI8852_Kd	**0.504**	0.007	0.021	0.043	0.157
MEDI8852_Tm	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
REGN10987_Kd	**0.520**	0.327	0.377	0.462	0.396
REGN10987_Tm	0.238	**0.810**	0.762	0.667	0.667
S309_Kd	**0.728**	0.526	0.586	0.581	0.600
S309_Tm	**0.830**	0.770	0.758	0.794	0.794
mAb114-mAb114UCA_Kd	0.658	**0.769**	0.727	0.658	0.738
mAb114-mAb114UCA_Tm	0.500	0.333	0.383	**0.517**	0.200
AVERAGE	0.519	**0.563**	**0.562**	**0.559**	0.507

Table 7: Full results of zero-shot fitness prediction for the antibody landscapes evaluated in this study.

	Uniref90+BFD30	Uniref90+BFD90
test set [ppl]	12.7	**12.6**
antibody binding [rho]	**0.42**	0.38
antibody general [rho]	0.73	**0.76**
narrow landscape avg [rho]	0.49	**0.50**
AAV [auc]	0.65	**0.67**
GFP [auc]	**0.84**	0.79
CM [auc]	**0.66**	0.65
GB1 [top100avg]	0.24	**0.33**

Table 8: Comparing language modeling and zero-shot fitness prediction performance for two 2.7B parameter models trained with differing amounts of proteins from metagenomic sources. BFD30 and BFD90 are clustered at 30% and 90% sequence identity respectively. The Uniref90+BFD90 database is majority metagenomic.

dataset	PROGEN2- small	PROGEN2- base	PROGEN2- large	PROGEN2- xlarge	PROGEN2- ensemble	rita-xl	eve	transcept	transcept (retrieval)	msa transf	esm-1v (single)
A0A192B1T2 Haddox 2018	0.487	0.441	0.483	0.460	0.490	0.500	0.510	0.510	0.509	0.510	0.490
A0A2Z5U3Z0 9INFA Doud 2016	0.486	0.573	0.555	0.563	0.584	0.540	0.534	0.539	0.549	0.160	0.510
AMIE PSEAE Wrenbeck 2017	0.584	0.558	0.580	0.549	0.593	0.560	0.558	0.501	0.585	0.610	0.610
B3VI55 LIPST Klesmith 2015	0.425	0.584	0.538	0.596	0.591	0.450	0.436	0.491	0.468	0.530	0.480
BLAT ECOLX Deng 2012	0.451	0.533	0.480	0.439	0.529	0.400	0.506	0.381	0.480	0.560	0.530
BLAT ECOLX Firnberg 2014	0.606	0.684	0.602	0.505	0.672	0.560	0.741	0.521	0.678	0.740	0.680
BLAT ECOLX Jacquier 2013	0.564	0.636	0.574	0.478	0.647	0.560	0.742	0.539	0.683	0.700	0.680
BLAT ECOLX Stiffler 2015	0.586	0.666	0.569	0.470	0.651	0.570	0.740	0.518	0.667	0.730	0.690
BRCA1 HUMAN Findlay 2018	0.245	0.441	0.425	0.455	0.460	0.070	0.322	0.538	0.574	0.400	0.440
CALM1 HUMAN Weile 2017	0.292	0.331	0.349	0.354	0.356	0.270	0.244	0.306	0.283	0.254	0.246
DLG4 RAT McLaughlin 2012	0.440	0.456	0.364	0.351	0.417	0.370	0.523	0.304	0.446	0.507	0.565
GAL4 YEAST Kitzman 2015	0.349	0.466	0.437	0.569	0.517	0.350	0.511	0.326	0.526	0.583	0.441
HSP82 YEAST Mishra 2016	0.528	0.559	0.499	0.494	0.536	0.530	0.537	0.503	0.530	0.482	0.568
IF1 ECOLI Kelsic 2016	0.485	0.513	0.477	0.492	0.522	0.420	0.525	0.548	0.509	0.227	0.538
KKA2 KLEPN Melnikov 2014	0.323	0.538	0.534	0.584	0.593	0.560	0.597	0.584	0.584	0.576	0.614
MK01 HUMAN Brenan 2016	0.182	0.057	0.076	0.068	0.048	0.050	0.251	0.034	0.139	0.153	0.183
MTH3 HAEAE Rockah 2015	0.479	0.711	0.684	0.710	0.717	0.680	0.710	0.673	0.655	0.687	0.701
P84126 THETH Chan 2017	0.501	0.573	0.560	0.657	0.635	0.550	0.567	0.533	0.541	0.631	0.546
PABP YEAST Melamed 2013	0.555	0.633	0.639	0.644	0.654	0.680	0.639	0.641	0.689	0.662	0.665
PA I34A1 Wu 2015	0.301	0.572	0.564	0.569	0.572	0.540	0.539	0.541	0.572	0.383	0.054
POLG HCVJF Qi 2014	0.524	0.522	0.513	0.589	0.562	0.440	0.614	0.525	0.577	0.600	0.605
Q2N0S5 9HIV1 Haddox 2018	0.510	0.372	0.423	0.319	0.412	0.350	0.496	0.412	0.492	0.490	0.496
Q59976 STRSQ Romero 2015	0.708	0.759	0.756	0.756	0.771	0.650	0.647	0.645	0.659	0.674	0.506
RASH HUMAN Bandaru 2017	0.429	0.410	0.375	0.319	0.408	0.400	0.454	0.377	0.447	0.415	0.359
RL401 YEAST Mavor 2016	0.482	0.454	0.478	0.460	0.486	0.430	0.364	0.331	0.368	0.378	0.297
RL401 YEAST Roscoe 2013	0.552	0.547	0.549	0.539	0.566	0.490	0.421	0.392	0.419	0.434	0.314
RL401 YEAST Roscoe 2014	0.426	0.399	0.379	0.364	0.408	0.390	0.401	0.343	0.404	0.365	0.254
SUMO1 HUMAN Weile 2017	0.532	0.536	0.513	0.434	0.513	0.390	0.531	0.424	0.488	0.423	0.430
TPK1 HUMAN Weile 2017	0.181	0.360	0.335	0.382	0.380	0.290	0.230	0.313	0.314	0.268	0.284
TPMT HUMAN Matreyek 2018	0.451	0.494	0.478	0.441	0.513	0.510	0.548	0.445	0.522	0.508	0.540
TRPC SACS2 Chan 2017	0.503	0.559	0.525	0.517	0.573	0.570	0.577	0.551	0.585	0.629	0.606
TRPC THEMA Chan 2017	0.475	0.377	0.415	0.474	0.425	0.450	0.420	0.453	0.436	0.474	0.472
UBC9 HUMAN Weile 2017	0.481	0.525	0.507	0.499	0.524	0.450	0.538	0.433	0.485	0.503	0.479
UBE4B MOUSE Starita 2013	0.452	0.429	0.410	0.292	0.420	0.300	0.476	0.256	0.388	0.347	0.462
YAP1 HUMAN Araya 2012	0.368	0.404	0.344	0.281	0.386	0.190	0.438	0.218	0.359	0.071	0.281
AVERAGE	0.456	**0.505**	0.485	0.476	**0.518**	0.443	**0.511**	0.447	**0.503**	0.476	0.475

Table 9: Full results of zero-shot fitness prediction for narrow landscapes.