Protein engineering

https://doi.org/10.1038/s41587-022-01634-2

Hallucinating functional protein sequences

David Belanger & Lucy J. Colwell

Check for updates

Functional proteins with limited homology to natural proteins are designed using a large language model.

Natural proteins cover just a tiny fraction of the space of viable protein sequences. How can we unlock the potential of proteins not generated by evolution? As Frances Arnold notes, “today we can for all practical purposes read, write, and edit any sequence of DNA, but we cannot compose it”[1]. Writing in *Nature Biotechnology*, Madani et al.[2] show that one solution to this challenge lies in using the wealth of protein sequence data and functional metadata assembled by the community to train machine learning models (Fig. 1). The authors’ large language model ProGen exploits this resource to design functional novel sequences that are strikingly diverse, with <50% identity to any known natural protein, without explicit biophysical or biochemical modeling. The ability to design new protein sequences with specific functional activities could have an enormous impact on our ability to produce food, combat climate change and cure diseases.

Directed evolution has proven remarkably successful at finding variants of known proteins with enhanced properties[1]. Yet designing proteins that are not homologous to those found in nature is extremely challenging. The strategy of walking uphill on a rugged protein fitness landscape can stall at local optima, making it hard to discover diverse functional variants. Techniques such as DNA shuffling recombine parental variants and allow larger moves in sequence space, but diverse variants are rarely generated because the synthesis process favors sequence-similar parents.

A more recent strategy is sequence design guided by models. These approaches aim to characterize the relationship between amino acid sequence and functional activity to guide the experimental exploration of sequence space[3,4]. An accurate model can make it possible to escape local optima and teleport across valleys in the fitness landscape, opening up previously inaccessible regions of sequence space that cannot be reached via paths that maintain function.

But what kind of modeling approach is suitable? Structure-based design, recently accelerated by deep learning, finds sequences that fold to a desired structure. This can work well if the structure of a protein

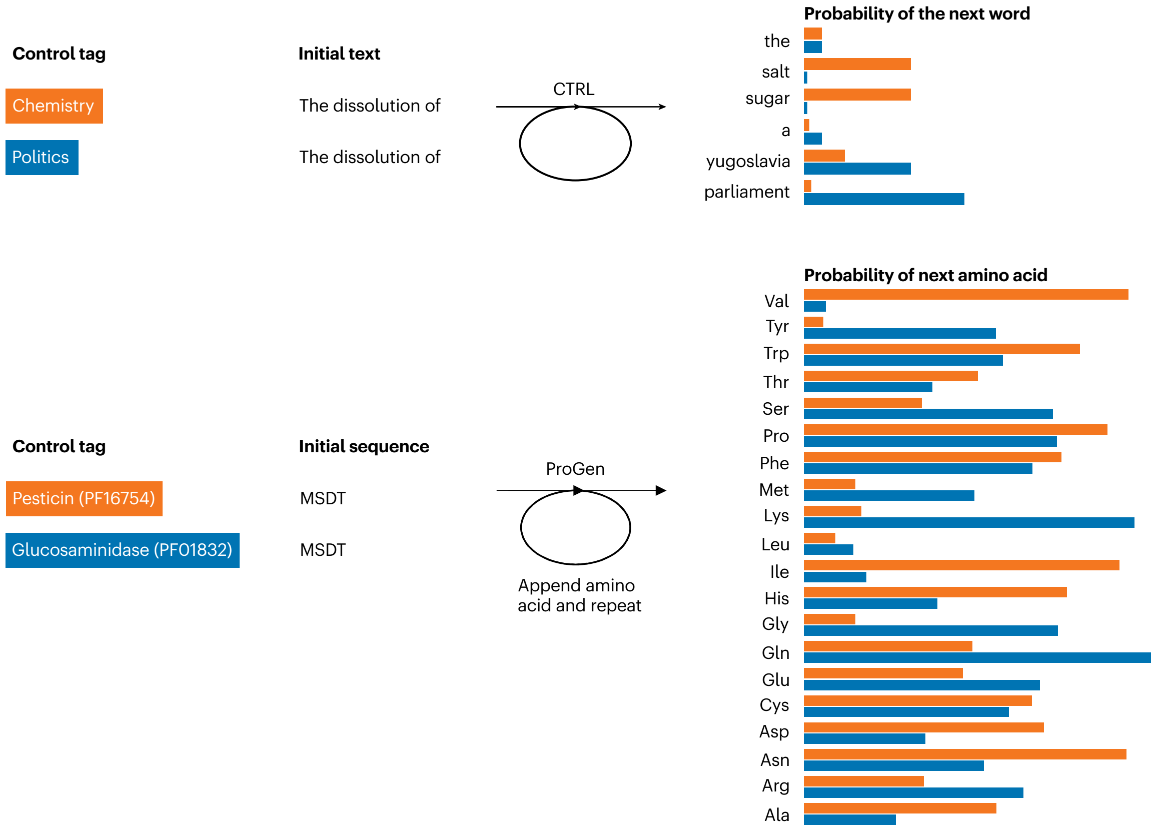

Fig. 1 | ProGen. In line with deep models for controllable generation of natural language text such as CTRL[10], ProGen samples protein sequences one amino acid at a time; the probability of each amino acid is influenced by the prefix of the sequence generated so far and a set of control tags that specify a desired protein function. Given the same prefix but different control tags, the resulting distribution may be quite different (orange vs. blue histograms).

with the desired functional activity is known[5]. In contrast, models that directly relate amino acid sequence to functional activity do not make this assumption and so can be applied more broadly. Well-known examples such as hidden Markov models and position-specific scoring matrices cannot capture the dependency between sequence positions, or epistasis, which often determines functional activity. This can cause their accuracy to diminish rapidly outside the training data. Coevolutionary methods, which explicitly incorporate pairwise dependencies between positions, may also exhibit this weakness[6]. Exciting recent progress has been made using experimental measurements of large sets of protein sequences collected specifically for the functional activity of interest to train machine learning models[4]. However, collecting these data can be costly and difficult, and the high-throughput assays required may not be feasible for many protein functional activities.

A complementary approach is suggested by the wealth of protein sequence data and corresponding functional annotations that the community has systematically assembled into publicly accessible databases. Recent work showed that deep learning language models trained using these data outperform existing alignment-based methods in tasks such as predicting function or structural contacts[7,8].

Madani et al. apply their language model in a generative mode rather than a predictive mode (Fig. 1). In contrast to other language models for protein sequence generation, ProGen does not just use amino acid sequences for training but also incorporates tags corresponding to functional keywords and taxonomic terms. This means that one can choose a specific function when generating new sequences by selecting the corresponding tag as input.

ProGen is trained across millions of diverse evolutionary protein sequences using the self-supervised task of predicting the next amino acid. ProGen is then fine-tuned with unaligned sequences from specific protein families, demonstrated in the paper for five different lysozyme families. The stacked self-attention layers that make up the model architecture allow ProGen to learn patterns that involve interactions between multiple sequence positions.

Madani et al. sampled 100 top-ranked sequences with 40–90% identity to the nearest training sequence for experimental validation. They found that these sequences successfully expressed and exhibited functional activity as frequently as a set of 100 positive control natural proteins. In contrast, sequences generated by a coevolutionary model fit using the same protein family data were less likely to express and had no detectable functional activity. Even sequences with 20–40% identity to the nearest training sequence expressed robustly, albeit with lower levels of functional activity.

To demonstrate ProGen's performance across diverse protein families, Madani et al. used literature data, demonstrating that ProGen log-likelihoods had a substantially higher area under the receiver operating characteristic curve (AUROC) for prediction of binary function labels for chorismate mutase variants than the coevolutionary model and a higher AUROC than ProteinGAN for malate dehydrogenase protein variants[6,9]. Removing either the large-scale initial training or the subsequent family-specific fine-tuning step resulted in significant performance drops, suggesting that both steps play important roles.

The results in this paper are exciting and suggest a variety of paths forward for the field, which will likely involve important technical challenges. For example, ProGen is trained only with evolutionary sequence data. An iterative active learning strategy in which the model is employed and updated over multiple rounds of experiments could improve performance, especially for functional activities that are not well represented among natural sequences. Moreover, the analysis of Madani et al. suggests that while ProGen designs are diverse in sequence, they retain structural homology to examples used to fine-tune the model. Extending the model to generate sequences with novel structures will likely require more innovation. Finally, the authors note that they do not expect ProGen to generate sequences with completely new functions, which poses another exciting challenge to the field.

David Belanger [1] **& Lucy J. Colwell** [1,2] ✉
[1]Google Research, Mountain View, CA, USA. [2]Department of Chemistry, Cambridge University, Cambridge, UK.
✉e-mail: ljc37@cam.ac.uk

Published online: 26 January 2023

References

1. Arnold, F. H. *Angew. Chem. Int. Ed. Engl.* **58**, 14420–14426 (2019).
2. Madani, A. et al. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-022-01618-2 (2023).
3. Romero, P. A., Krause, A. & Arnold, F. H. *Proc. Natl Acad. Sci. USA* **110**, E193–E201 (2013).
4. Bryant, D. H. et al. *Nat. Biotechnol.* **39**, 691–696 (2021).
5. Dauparas, J. et al. *Science* **378**, 49–56 (2022).
6. Russ, W. P. et al. *Science* **369**, 440–445 (2020).
7. Rives, A. et al. *Proc. Natl Acad. Sci. USA* **118**, e2016239118 (2021).
8. Dohan, D. et al. In *Proc. 27th ACM SIGKDD Conf. Knowledge Discovery & Data Mining* 2782–2791 (ACM, 2021).
9. Repecka, D. et al. *Nat. Mach. Intell.* **3**, 324–333 (2021).
10. Keskar, N. S. et al. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1909.05858 (2019).

Competing interests

D.B. and L.J.C. have performed research as part of their employment at Google LLC. Google is a technology company that sells machine learning services as part of its business.