Title
A language model beats alphafold2 on orphans

Journal
Nature Biotechnology, 40(11)

ISSN
1087-0156

Authors
Michaud, Jennifer M
Madani, Ali
Fraser, James S

Publication Date
2022-11-01

DOI
10.1038/s41587-022-01466-0

Peer reviewed

HHS Public Access

Author manuscript

Nat Biotechnol. Author manuscript; available in PMC 2023 May 01.

Published in final edited form as:

Nat Biotechnol. 2022 November ; 40(11): 1576–1577. doi:10.1038/s41587-022-01466-0.

A language model beats alphafold2 on orphans

Jennifer M. Michaud[1], **Ali Madani**[2], **James S. Fraser**[1]

[1]Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA

[2]Profluent Bio, Oakland, CA

Abstract

Protein structure prediction with a language model improves accuracy for orphan and designed proteins.

Last year, decades of research on protein structure prediction culminated in the publication of two deep-learning methods, AlphaFold2[1] and RoseTTAFold[2], that were nearly as accurate as experimental methods for protein structure determination. But both algorithms consume large computing resources, and because they depend on multiple sequence alignments as input, they are less successful in predicting the structure of so-called 'orphan' proteins — proteins with few or no homologs. Writing in Nature Biotechnology, Chowdhury et al.[3] report substantial progress on both of these challenges. Their recurrent geometric network 2 (RGN2) method, which relies on a protein language algorithm, uses orders-of-magnitude less computing time than AlphaFold2 and RoseTTAFold while outperforming them on average in predicting the structures of orphan proteins. These results highlight the breakneck pace of the field and suggest that further leaps in computational speed lie ahead.

Although orphan proteins may seem rare, they are common in the vast and ever-expanding protein universe that is emerging from large-scale genome sequencing. Approximately 20% of all metagenomic protein sequences[4] and 11% of eukaryotic and viral protein sequences[5] are orphans. Moreover, de novo designed proteins are, by definition, without homologs, and an orthogonal method to predict the structures of proteins designed using traditional knowledge or physics-based forcefields would provide a very useful test of their potential to fold correctly.

AlphaFold2 and RoseTTAFold work less well on orphan proteins because they rely on multiple sequence alignments. Large and diverse alignments are important because correlations in amino acid co-occurrences between positions in multiple sequence alignments are a strong indicator that those positions are near each other in the three-dimensional space of a folded protein. An earlier generation of computational models (AlphaFold1, trRosetta) used these inter-residue distance constraints as input for restrained energy minimization, in a manner similar to how nuclear Overhauser effect restraints are used in NMR spectroscopy for protein structure determination (Figure 1a,b). The most

Correspondence to jaimefraser@gmail.com.

recent deep learning models (AlphaFold2, RoseTTAFold) go a step further and generate a denser network of constraints that is used directly for structure prediction (Figure 1c).

The new RGN2 algorithm of Chowdhury et al. foregoes multiple sequence alignments completely, and it outperforms AlphaFold2 and RoseTTAFold on a set of orphan and designed proteins as measured by the root mean square deviation between predicted and experimental structures. How did the authors achieve such impressive results? The key advance was to employ a deep-learning language model, which is explicitly alignment-free. These models were pioneered for tasks that involve understanding natural language. They are trained to 'fill-in-the-blank' by predicting the most probable word for a given blank (or 'masked token') in a sentence. For example, a language model might complete the sentence "The most exciting language model research is published in ___" with "journals" or "conferences" or "Twitter" and assign a lower probability to the words "space" or "restaurants".

The use of language models is a somewhat recent emergence in the fast-developing space of protein structure prediction. Their utility also exemplifies the theme of how increasing scale has enabled discovery. The ever-expanding protein sequence database provides a large training set for language models and recent advances in GPU computing have made it tractable to train such models with increasing complexity. As the models have grown in size, they have demonstrated increasing power for function prediction[6], evolutionary analyses[7,8], and now structural inference.

A similar task can be formulated to train a language model for protein structure prediction. In this case, the model must 'fill-in-the-blank' by predicting amino-acid-residue probabilities for a masked protein sequence in training. In the natural language example, the model learns a representation in which words that are similar to each other are also close in the embedding space. For protein space, the language model learns a representation that contains information about not only pairwise interactions between residues, but also three-residue, four-residue, and even higher order, interactions between residues. These interactions are the embodiment of the machine learning concepts of local and distant attention, which are also exploited in the analysis of multiple sequence alignments by AlphaFold2 and RoseTTAFold.

Similarly, RGN2 makes use of the concept of local and distant attention, which allows it to learn relationships over a wide range of distances in the one-dimensional input. Attention-based methods infer structure based on many restraints across almost every position in the sequence (Figure 1c). Using the language model, RGN2 goes a step further by learning aspects of these restraints from all proteins, not just those contained in a specific alignment (Figure 1d).

Alongside using language models for guiding which parts of a linear sequence might be near to each other in space, RGN2 also explicitly learns geometric relationships to generate the backbone structure of a protein. It uses a mathematical representation of the polypeptide backbone based on Frenet-Serret formulas that is translationally and rotationally invariant. Although the need for translational and rotational invariance seems obvious — a structure

remains the same no matter how it is positioned in a coordinate system — finding a representation with these properties that is compatible with machine learning is far from trivial.

While RGN2 succeeds in the case of orphan proteins, it does not perform as well as AlphaFold2 or RosettaFold on proteins where multiple sequence alignments can be leveraged as inputs. It is also potentially less useful in iterative protein “hallucination” design applications, where a sequence is selected to converge on a desired structure[9].

Given the sophisticated nature of the approaches implemented in RGN2, it should be useful in helping to unpack what makes the deep-learning ‘black box’ work. The substantial reduction of compute time and the ability to iterate quickly could have an invigorating effect on the field by allowing a wider range of model ablations to tease out the contribution of different parts of RGN2. The contributions of Chowdhury et al. to geometric representations are likely to impact other deep-learning structure prediction approaches as well. The direct incorporation of language models in a framework for generating 3D structure predictions is particularly exciting because these models have the ability to extrapolate beyond the training data, even generating novel, functional proteins[10]. Another notable aspect of this study is a new curated set of orphan and designed proteins that can be used as a benchmark for future structure prediction efforts.

Finally, the publication of RGN2 highlights how fast the field is moving. Each day brings new activity in this space, on biorxiv and twitter, where open implementations of alphafold and new interleaved language model/alphafold-type protocols have been announced recently. This sea change was catalyzed in part by Mohammed AlQuraishi, the senior author of Chowdhury et al., who wrote an in-depth blog post on the first AlphaFold breakthrough[11] and another on AlphaFold2[12], months before those works were eventually published in a journal. Even if the speed of advances in protein structure prediction slows, application of the lessons learned to other domains, including protein design and protein-small molecule interactions, has plenty of room to grow. A major lesson of RGN2 is that the most innovative contributions may come from focusing on areas where leading methods fall short and by not abandoning orphan problems.

References

1. Jumper J, et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021). 10.1038/s41586-021-03819-2 [PubMed: 34265844]
2. Baek M et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science 373, 871 (2021). [PubMed: 34282049]
3. Chowdhury R, et al. Single-sequence protein structure prediction using a language model and deep learning. Nat. Biotechnol (2022).
4. Pearson WR An introduction to sequence similarity (‘homology’) searching. Curr Protoc Bioinformatics Chapter 3, Unit3.1 (2013).
5. Perdigão N, et al. Unexpected features of the dark proteome. Proc Natl Acad Sci U S A 112(52):15898–903. (2015) doi: 10.1073/pnas.1508380112. [PubMed: 26578815]
6. Alley EC, et al. Unified rational protein engineering with sequence-based deep representation learning. Nat. Methods 16, 1315–1322 (2019). [PubMed: 31636460]

7. Riesselman AJ, et al. Deep generative models of genetic variation capture the effects of mutations. Nat. Methods 15, 816–822 (2018). [PubMed: 30250057]

8. Hie B, Zhong ED, Berger B, and Bryson B Learning the language of viral evolution and escape. Science 371, 284–288 (2021). [PubMed: 33446556]

9. Anishchenko I, et al. De novo protein design by deep network hallucination. Nature 600, 547–552 (2021). 10.1038/s41586-021-04184-w [PubMed: 34853475]

10. Madani A et al. Deep neural language modeling enables functional protein generation across families. bioRxiv 2021.07.18.452833 (2021) doi:10.1101/2021.07.18.452833.

11. AlQuraishi M AlphaFold @ CASP13: ‘What just happened?’ Some Thoughts on a Mysterious Universe https://moalquraishi.wordpress.com/2018/12/09/alphafold-casp13-what-just-happened/ (2018).

12. AlQuraishi M AlphaFold2 @ CASP14: ‘It feels like one’s child has left home.’ Some Thoughts on a Mysterious Universe https://moalquraishi.wordpress.com/2020/12/08/alphafold2-casp14-it-feels-like-ones-child-has-left-home/ (2020).

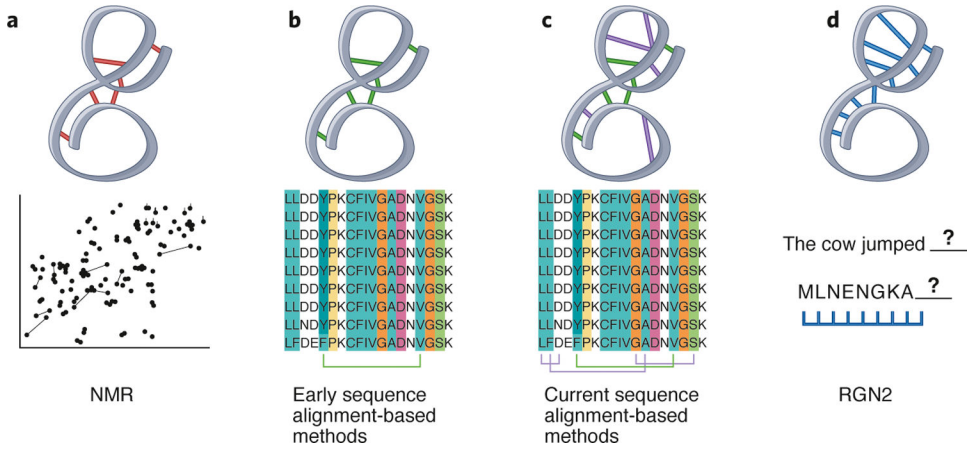

Figure 1: Distance restraints from experiment or language models in protein structure calculations.

a) Protein structure determined by NMR spectroscopy uses nuclear Overhauser effects (shown here as purple lines) to infer residues that are close together in 3D space. **b**) Similarly, early co-evolution-based methods used high correlations from sequence alignments (green) to infer residues that are close together in 3D space. **c**) AlphaFold2 and deep learning models further leverage attention to identify very weak preferences (orange) in sequence alignments to generate a denser network of restraints. **d)** A breakthrough of RGN2 is to learn these preferences (blue) from language models, which are intrinsically alignment-free, improving protein structure prediction for orphan proteins.