Journal Club

Chadi Barakat

2021-04-06

UNIVERSITY OF ICELAND
FACULTY OF INDUSTRIAL ENGINEERING,
MECHANICAL ENGINEERING AND COMPUTER SCIENCE

JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

Joint Research Center for Computational Biomedicine | RWTH AACHEN UNIVERSITY

Article

Highly accurate protein structure prediction with AlphaFold

https://doi.org/10.1038/s41586-021-03819-2

Received: 11 May 2021

Accepted: 12 July 2021

Published online: 15 July 2021

Open access

Check for updates

John Jumper[1,4]✉, Richard Evans[1,4], Alexander Pritzel[1,4], Tim Green[1,4], Michael Figurnov[1,4], Olaf Ronneberger[1,4], Kathryn Tunyasuvunakool[1,4], Russ Bates[1,4], Augustin Žídek[1,4], Anna Potapenko[1,4], Alex Bridgland[1,4], Clemens Meyer[1,4], Simon A. A. Kohl[1,4], Andrew J. Ballard[1,4], Andrew Cowie[1,4], Bernardino Romera-Paredes[1,4], Stanislav Nikolov[1,4], Rishub Jain[1,4], Jonas Adler[1], Trevor Back[1], Stig Petersen[1], David Reiman[1], Ellen Clancy[1], Michal Zielinski[1], Martin Steinegger[2,3], Michalina Pacholska[1], Tamas Berghammer[1], Sebastian Bodenstein[1], David Silver[1], Oriol Vinyals[1], Andrew W. Senior[1], Koray Kavukcuoglu[1], Pushmeet Kohli[1] & Demis Hassabis[1,4]✉

Proteins are essential to life, and understanding their structure can facilitate a mechanistic understanding of their function. Through an enormous experimental effort[1–4], the structures of around 100,000 unique proteins have been determined[5], but this represents a small fraction of the billions of known protein sequences[6,7]. Structural coverage is bottlenecked by the months to years of painstaking effort required to determine a single protein structure. Accurate computational approaches are needed to address this gap and to enable large-scale structural bioinformatics. Predicting the three-dimensional structure that a protein will adopt based solely on its amino acid sequence—the structure prediction component of the 'protein folding problem'[8]—has been an important open research problem for more than 50 years[9]. Despite recent progress[10–14], existing methods fall far short of atomic accuracy, especially when no homologous structure is available. Here we provide the first computational method that can regularly predict protein structures with atomic accuracy even in cases in which no similar structure is known. We validated an entirely redesigned version of our neural network-based model, AlphaFold, in the challenging 14th Critical Assessment of protein Structure Prediction (CASP14)[15], demonstrating accuracy competitive with experimental structures in a majority of cases and greatly outperforming other methods. Underpinning the latest version of AlphaFold is a novel machine learning approach that incorporates physical and biological knowledge about protein structure, leveraging multi-sequence alignments, into the design of the deep learning algorithm.

Cite as: M. Baek *et al.*, *Science* 10.1126/science.abj8754 (2021).

Accurate prediction of protein structures and interactions using a three-track neural network

Minkyung Baek[1,2], Frank DiMaio[1,2], Ivan Anishchenko[1,2], Justas Dauparas[1,2], Sergey Ovchinnikov[3,4], Gyu Rie Lee[1,2], Jue Wang[1,2], Qian Cong[5,6], Lisa N. Kinch[7], R. Dustin Schaeffer[6], Claudia Millán[8], Hahnbeom Park[1,2], Carson Adams[1,2], Caleb R. Glassman[9,10], Andy DeGiovanni[12], Jose H. Pereira[12], Andria V. Rodrigues[12], Alberdina A. van Dijk[13], Ana C. Ebrecht[13], Diederik J. Opperman[14], Theo Sagmeister[15], Christoph Buhlheller[15,16], Tea Pavkov-Keller[15,17], Manoj K. Rathinaswamy[18], Udit Dalwadi[19], Calvin K. Yip[19], John E. Burke[18], K. Christopher Garcia[9,10,11,20], Nick V. Grishin[6,21,7], Paul D. Adams[12,22], Randy J. Read[8], David Baker[1,2,23]*

[1]Department of Biochemistry, University of Washington, Seattle, WA 98195, USA. [2]Institute for Protein Design, University of Washington, Seattle, WA 98195, USA. [3]Faculty of Arts and Sciences, Division of Science, Harvard University, Cambridge, MA 02138, USA. [4]John Harvard Distinguished Science Fellowship Program, Harvard University, Cambridge, MA 02138, USA. [5]Eugene McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, TX, USA. [6]Department of Biophysics, University of Texas Southwestern Medical Center, Dallas, TX, USA. [7]Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, TX, USA. [8]Department of Haematology, Cambridge Institute for Medical Research, University of Cambridge, Cambridge, UK. [9]Program in Immunology, Stanford University School of Medicine, Stanford, CA 94305, USA. [10]Department of Molecular and Cellular Physiology, Stanford University School of Medicine, Stanford, CA 94305, USA. [11]Department of Structural Biology, Stanford University School of Medicine, Stanford, CA 94305, USA. [12]Molecular Biophysics & Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [13]Department of Biochemistry, Focus Area Human Metabolomics, North-West University, 2531 Potchefstroom, South Africa. [14]Department of Biotechnology, University of the Free State, 205 Nelson Mandela Drive, Bloemfontein 9300, South Africa. [15]Institute of Molecular Biosciences, University of Graz, Humboldtstrasse 50, 8010 Graz, Austria. [16]Medical University of Graz, Graz, Austria. [17]BioTechMed-Graz, Graz, Austria. [18]Department of Biochemistry and Microbiology, University of Victoria, Victoria, BC, Canada. [19]Life Sciences Institute, Department of Biochemistry and Molecular Biology, The University of British Columbia, Vancouver, BC, Canada. [20]Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, CA 94305, USA. [21]Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, TX, USA. [22]Department of Bioengineering, University of California, Berkeley, Berkeley, CA 94720, USA. [23]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA.

***Corresponding author. Email: dabaker@uw.edu**

DeepMind presented remarkably accurate predictions at the recent CASP14 protein structure prediction assessment conference. We explored network architectures incorporating related ideas and obtained the best performance with a three-track network in which information at the 1D sequence level, the 2D distance map level, and the 3D coordinate level is successively transformed and integrated. The three-track network produces structure predictions with accuracies approaching those of DeepMind in CASP14, enables the rapid solution of challenging X-ray crystallography and cryo-EM structure modeling problems, and provides insights into the functions of proteins of currently unknown structure. The network also enables rapid generation of accurate protein-protein complex models from sequence information alone, short circuiting traditional approaches which require modeling of individual subunits followed by docking. We make the method available to the scientific community to speed biological research.

Background Information

- The “protein folding problem” increases in complexity as the peptide chain grows.
- Current go-to methods include X-ray Crystallography, Fluorescence Spectroscopy, and Nuclear Magnetic Resonance Spectroscopy.
- CASP: Critical Assessment of Protein Structure, a biennial competition to test protein structure computation methods on newly decoded.

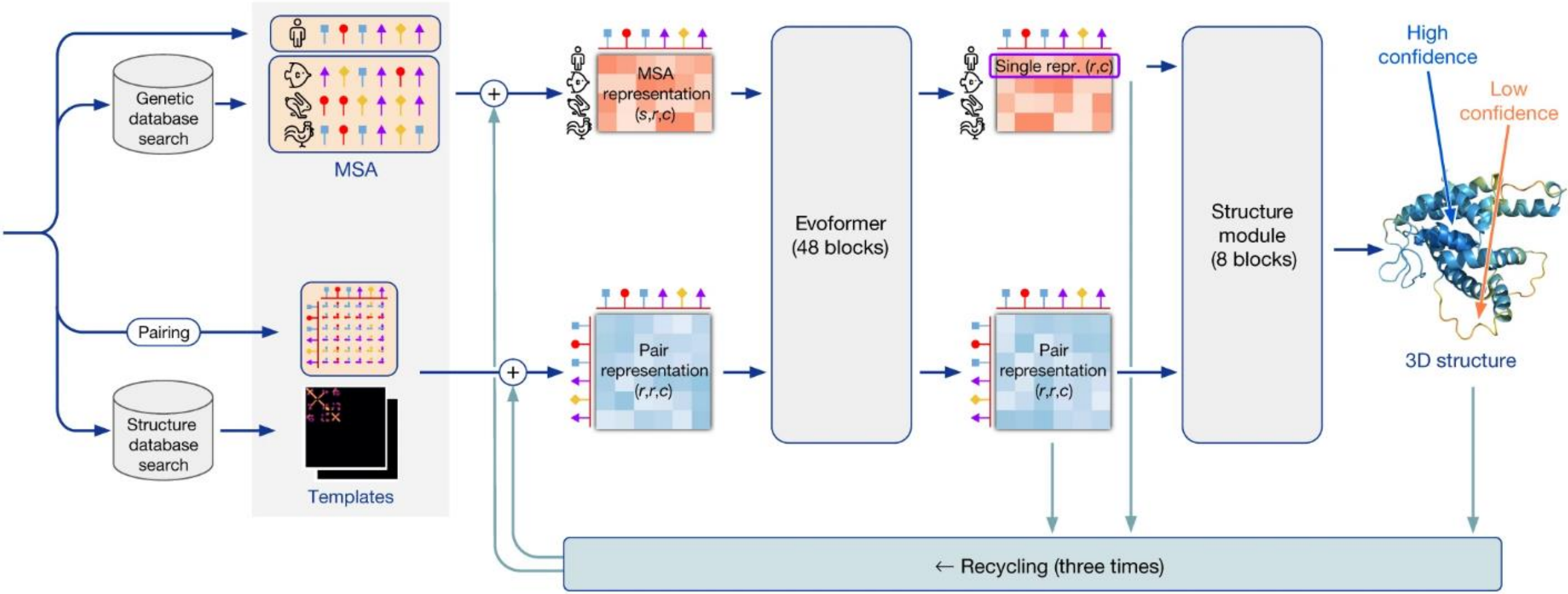
DeepMind and AlphaFold

- UK-based company purchased by Google.
- Developed AlphaGO which is famous for being the first AI to beat the world Go champion.
- Participated in CASP13 with the first Iteration of AlphaFold and met limited success.
- Participated in CASP14 with AlphaFold2 and outperformed all competing methods.
  - 0.96 Å r.m.s.d. backbone accuracy vs. 2.8 Å r.m.s.d. second best performer.
  - 1.5 Å r.m.s.d. all-atom accuracy vs. 3 Å r.m.s.d. second best performer.

AlphaFold2

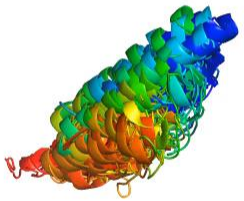
- Neural Net that treats the protein structure as a Graph Inference problem in 3D.
- Consisting of two main stages:
  - Trunk with “EvoFormer” blocks that have parallel channels:
    - (N_{res} x N_{seq}) which represents the Multiple Sequence Alignment (MSA)
    - (N_{res} x N_{res}) which represents the residues along the chain.
  - Structure Model that performs rotations and translations on the resulting chain in order to output 3D shape.
- Up to 48 EvoFormer blocks.
- End-to-end prediction of protein structure

AlphaFold2 Structure



AlphaFold in Action

RNA Polymerase of a crAss-like Phage



Recycling iteration 0, block 01
Secondary structure assigned from the final prediction

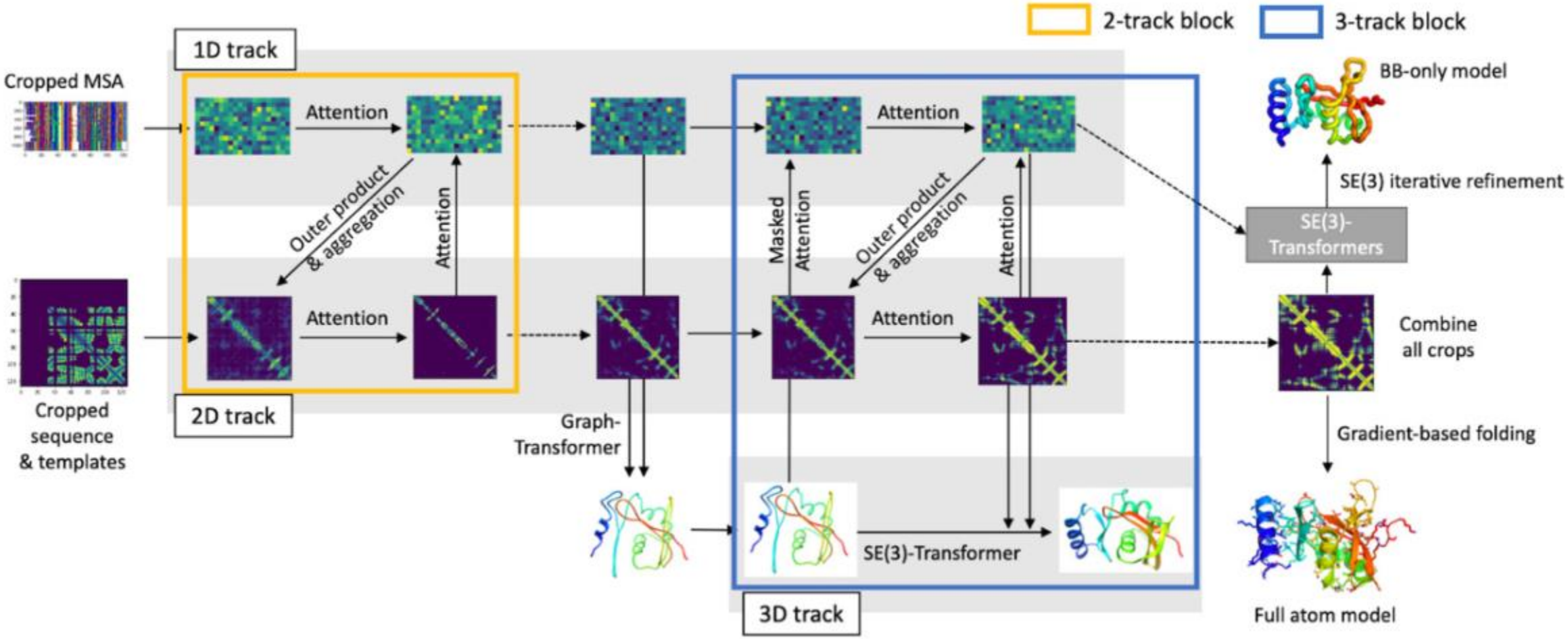
T1091

Recycling iteration 0, block 01
Secondary structure assigned from the final prediction

RoseTTAFold

- Developed by research group at University of Washington after CASP14.
- Tried two approaches and decided on a 3-Track approach:
  1. 1D sequence alignment.
  2. 2D distance matrix.
  3. 3D backbone coordinates.
- Constant back-and-forth communication between the 3 tracks.
- Simultaneously performs calculations on all 3 tracks instead of taking a hierarchical approach.

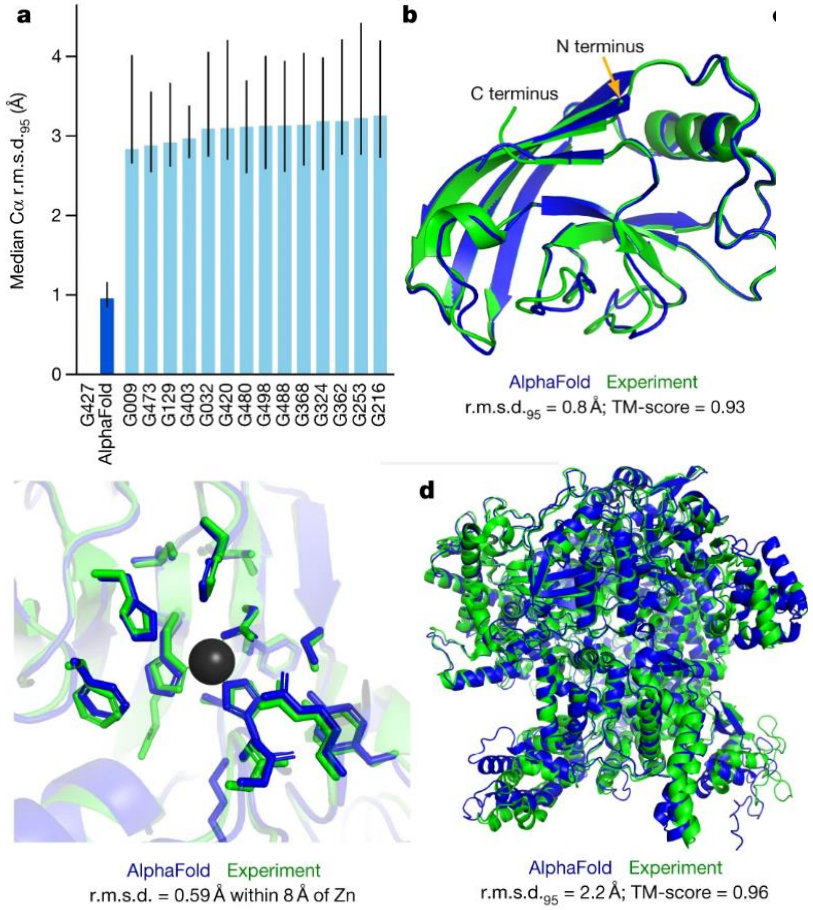
RoseTTAFold Structure

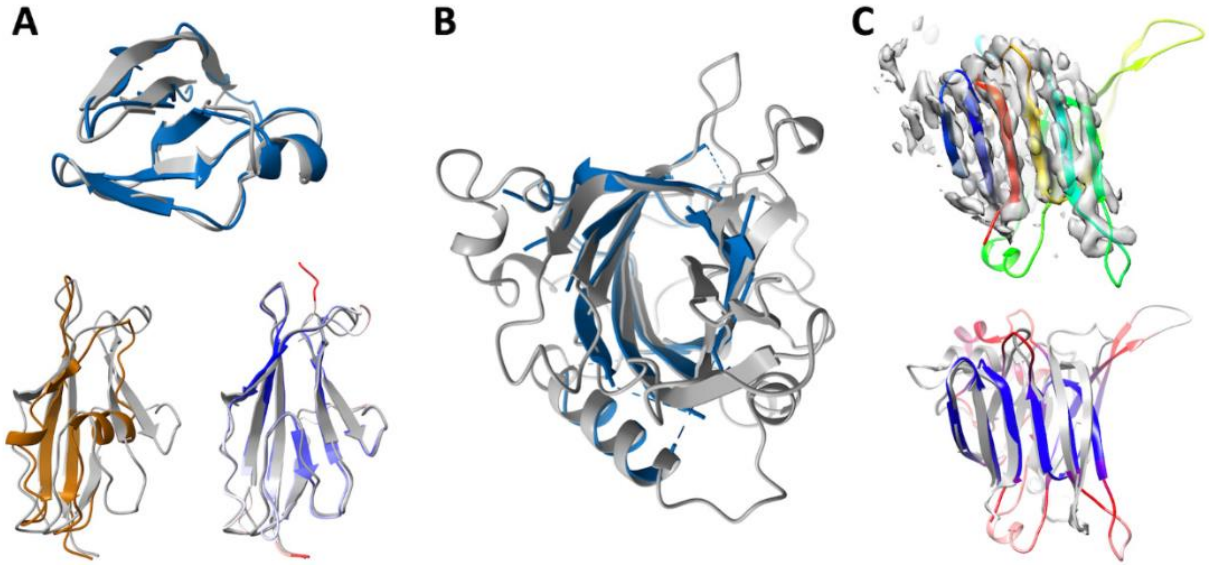


Why it’s worth mentioning RoseTTAFold

- Due to hardware limitations, the networks takes small snippets of the whole chain at a time.
  - Seems to produce better results than end-to-end
- Whole model trains more efficiently than the described training of AlphaFold2.
- Being able to deal with disconnected snippets has the side-effect of allowing RoseTTAFold to deal with protein-protein complexes.

Side by side of the results





Takeaways

- Sometimes trying to work around a problem yields some pleasant little surprises.
- Finding inspiration in the work of others, communicating, collaborating, and competing.
- And most importantly: open source.

generation of accurate protein-protein complex models from sequence information alone, short circuiting traditional approaches which require modeling of individual subunits followed by docking. We make the method available to the scientific community to speed biological research.

Code availability

Source code for the AlphaFold model, trained weights and inference script are available under an open-source license at https://github.com/deepmind/alphafold.

Thank you

Resources

- https://www.nature.com/articles/s41586-021-03819-2
- https://github.com/deepmind/alphafold
- https://www.science.org/doi/abs/10.1126/science.abj8754
- https://github.com/RosettaCommons/RoseTTAFold