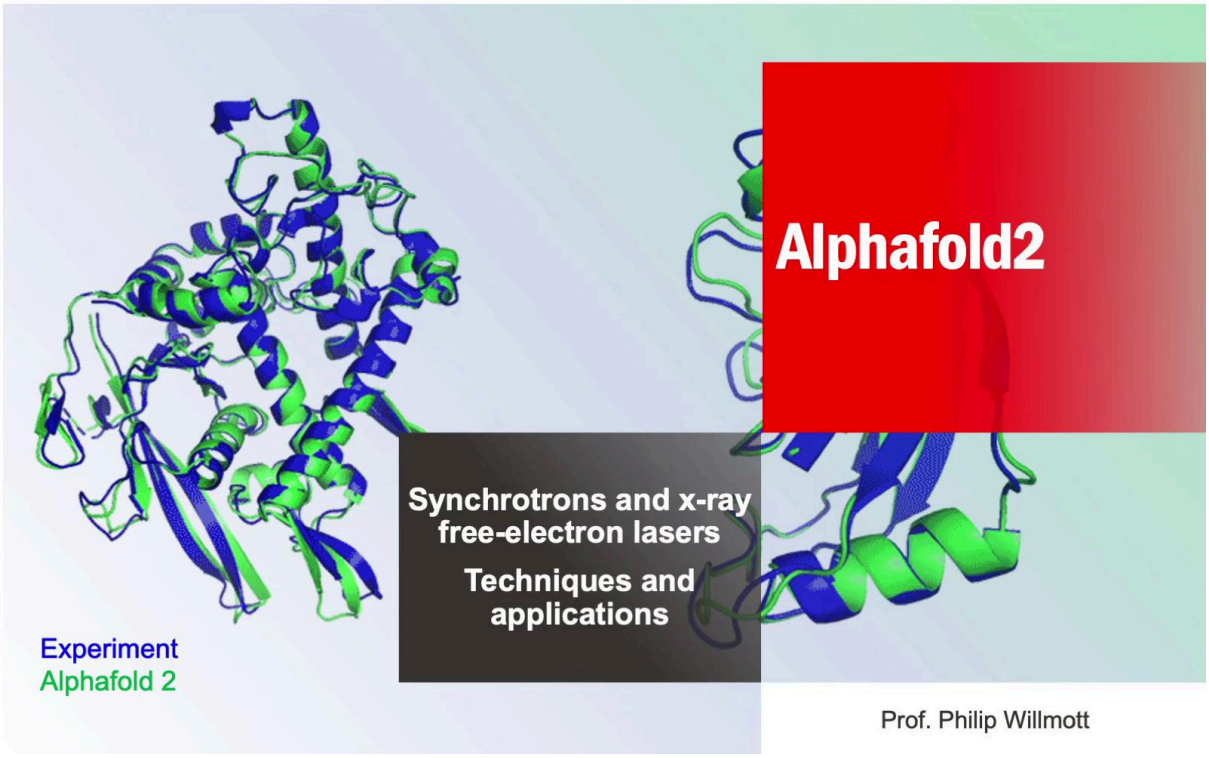

Alphafold2
Synchrotrons and x-ray free-electron lasers
Techniques and applications
Experiment
Alphafold 2
Prof. Philip Willmott


phase
order
structure
ray
crystal
used
time
example
surface
shown
electron
diffraction
synchrotron
around
scattering
distance
using
intensity
technique
object
Bragg
peak
experiment
Note
image
might
part
acid
sequence
function
amino
acid
degree
energy
atom
called
information
detector
Fourier transform
Alphafold
material
video
associated
week
section
angle
another
signal
direction
orbital
within
resolution
approximately
plane
diffraction pattern
recorded
size
small
divided
large
different
Search MOOC
Video

Contents and objectives of this video

- Alphafold2
- Example
- The future of MX

Welcome back. In this video, I will attempt to give a cursory description of Alphafold 2, a machine learning algorithm from Google DeepMind that appears to all intents and purposes to have solved the phase problem in macromolecular crystallography. Note that this disruptive breakthrough was only possible thanks to the enormous knowledge and database harvested over the last seven decades in the fields of protein crystallography and to a lesser extent, other structural solution methods like nuclear magnetic resonance and cryo-EM.

Notes

Summary

0m 05s

Alphafold2 – musings from the ‘60s

- John Kendrew, Nobel Lecture 1962

  “… in the very long run, it should only be necessary to determine the amino acid sequence of a protein, and its three-dimensional structure could then be predicted; in my view this day will not come soon, but when it does come the X-ray crystallographers can go out of business, perhaps with a certain sense of relief… It will also be possible to discuss the structures of many important proteins which cannot be crystallized and therefore lie outside the crystallographer’s purview.”

- October 2020: Still not there…

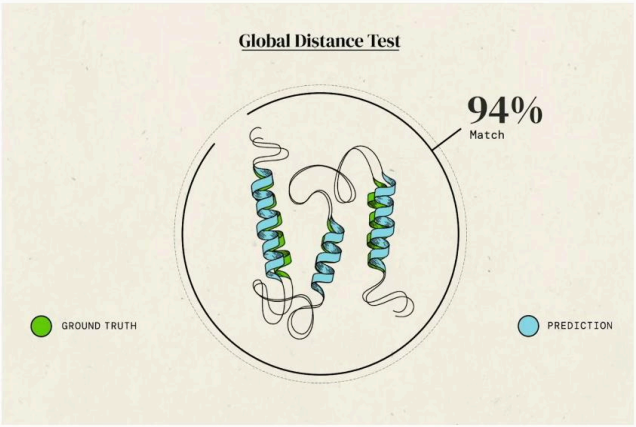
I wonder what John Kendrew would think if he were alive today and had experienced the explosive entrance of Alphafold 2 on the scene in November 2020. In his Nobel lecture in 1962, he would muse, "In the very long run, it should only be necessary to determine the amino acid sequence of a protein, and its three-dimensional structure could then be predicted. In my view, this day will not come soon, but when it does come, the X-ray crystallographers can go out of business, perhaps with a certain sense of relief. It will also be possible to discuss the structures of many important proteins which cannot be crystallised and therefore lie outside the crystallographers' purview." The very long run would turn out to be nearly 60 years. He was also at least partially correct in assuming that such an ability to predict the structure from the amino acid sequence alone would be extremely important for those proteins that do not crystallise easily or at all. However, his prediction that X-ray crystallographers by which he meant those involved in macromolecular crystallography would go out of business does not seem to be coming to pass, at least in the foreseeable future.

Notes

Summary

0m 47s

CASP



Extracted from https://www.youtube.com/watch?v=gg7WjuFs8F4&t=149s

- Critical Assessment of protein Structure Prediction
  - "Help advance the methods of identifying protein structure from (amino-acid) sequence"
  - CASP1: 1994 (biennial)
  - "Double blind" experiments
    - Structure known but kept secret
  - Metric: Global distance test – global score (GDT-TS)
    - Percentage of well-modelled residues w.r.t. target
    - 90% is (was!) the holy grail

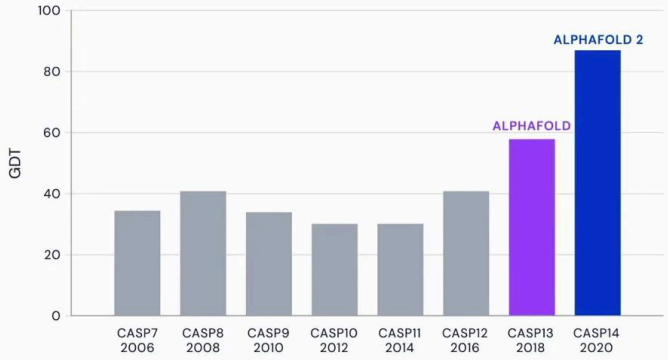
Alphafold 2 is the successor of Alphafold, first featured in CASP 13 in 2018. CASP, short for Critical Assessment of Protein Structure Prediction, is a biannual event that has been following the progress of algorithms that predict the structure of a given protein only from its amino acid sequence. The first meeting was in 1994. Essentially, CASP is an olympiad in which different groups compete for the most accurate algorithm applied to proteins whose structures have not yet been published and have been kept secret, using the metric of the global test distance, which measures the percentage of residues in the sequence that physically match the target structure. In terms of usefulness in biochemistry and the pharmaceutical industry, a value of 90% GDT has long been considered to be the holy grail.

Notes

Summary

2m 09s

CASP



- Critical Assessment of protein Structure Prediction
  - "Help advance the methods of identifying protein structure from (amino) sequence"
  - CASP1: 1994 (biennial)
  - "Double blind" experiments
    - Structure known but kept secret
  - Metric: Global distance test – global score (GDT-TS)
    - Percentage of well-modelled residues w.r.t. target
    - 90% is (was!) the holy grail

J. Jumper *et al.*, "Highly accurate protein structure prediction with Alphafold"

Since its inception in 1994, best GDT values remained at around 40% or lower. In 2016, Alphafold produced results that approached 60%, a very significant improvement, but not quite there yet. Not yet. Two years later, in the depths of the COVID pandemic, Alphafold 2 burst onto the scene at CASP 14. With a GDT of approximately 90%, Alphafold 2 was recognized as being the first program ever that solved the protein folding problem. This is one of the very few pieces of good news in an otherwise extremely trying year.

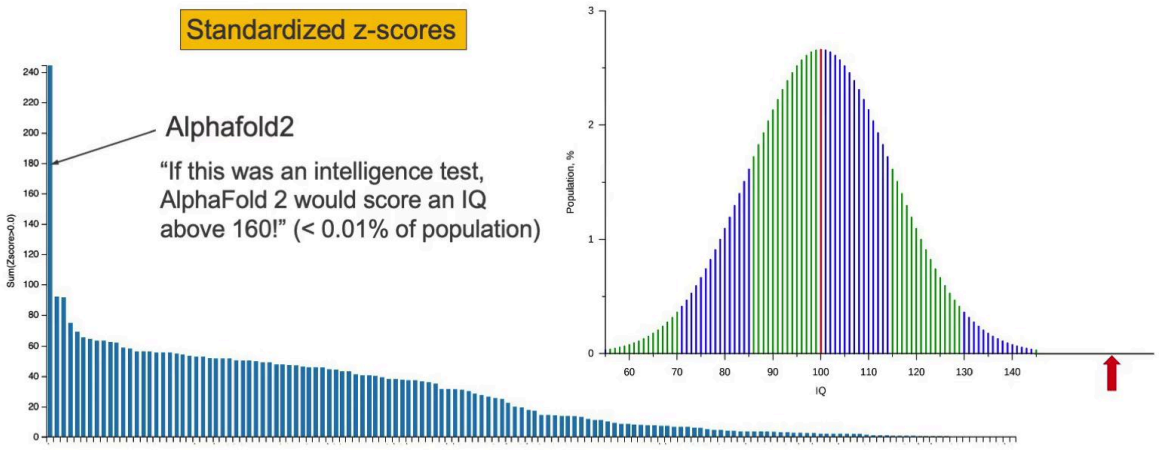
Notes

Summary

3m 11s

CASP14 – Alphafold2 enters stage left

- CASP14, Nov/Dec 2020
- Alphafold2 RMSD |expt - pred| < 1 Å



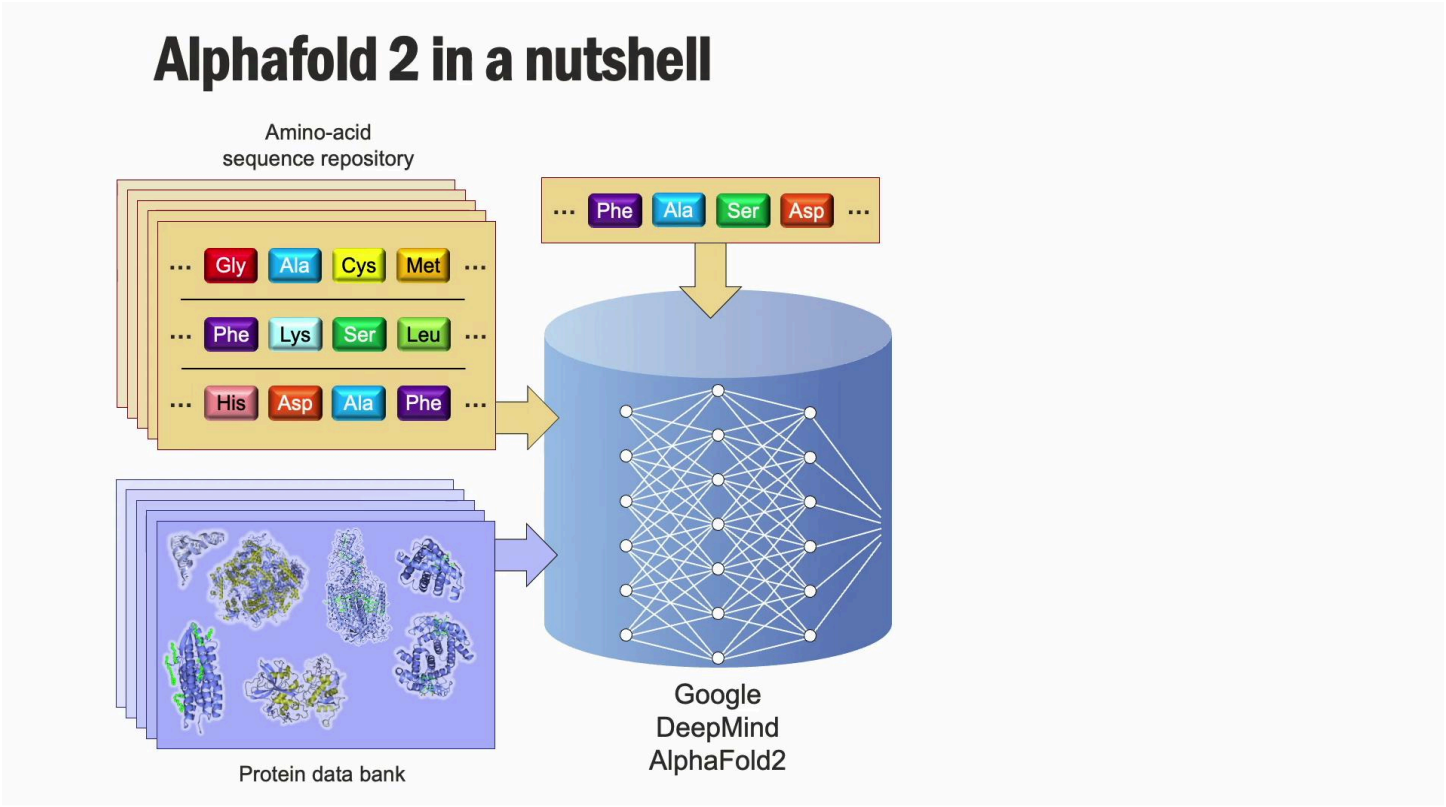
See also: https://www.blopig.com/blog/2020/12/casp14-what-google-deepminds-alphafold-2-really-achieved-and-what-it-means-for-protein-folding-biology-and-bioinformatics/

The root mean square deviation of equivalent atoms was shown to be under one angstrom. It has been stated that if Alphafold 2 were an intelligence test, it would score an IQ of over 160. Alphafold 2 is a machine learning or artificial intelligence algorithm.

Notes

Summary

3m 54s

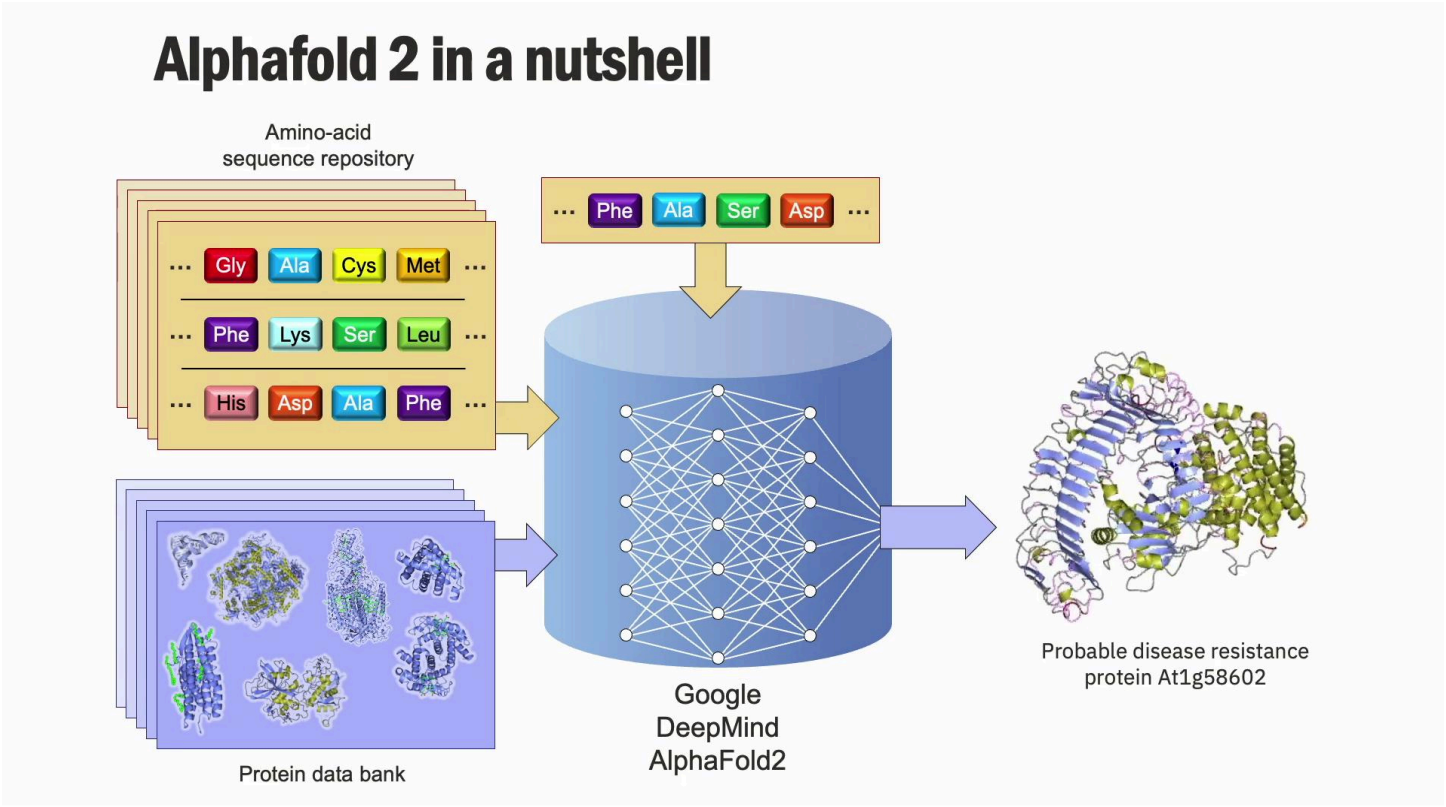


Explained extremely roughly, Alphafold 2 works by taking as its initial input for every known protein structure, both the amino acid sequence, that is the daisy chain composed of the 20 naturally occurring amino acids, plus its determined structure from their respective repositories. This totals in excess of 160,000 independent structures. It uses this information using a so-called neural network to identify trends between all the different parts of the structures and their neighbourhoods, seeing how these have an impact on distances and relative orientations, thereby building up general rules and probabilities. The neighbourhood need not only be neighbours in the amino acid sequence, but also other amino acids that happen to be nearby in space due to folding, even if they might be very far away on the amino acid sequence. In this manner, the initial intelligence is generated that Alphafold 2 needs to then predict new structures. This need only be done once for a given set of rules that Alphafold 2 uses. Armed with this deep knowledge, a new amino acid sequence is then entered into the algorithm and the probable structure is determined.

Notes

Summary

4m 16s



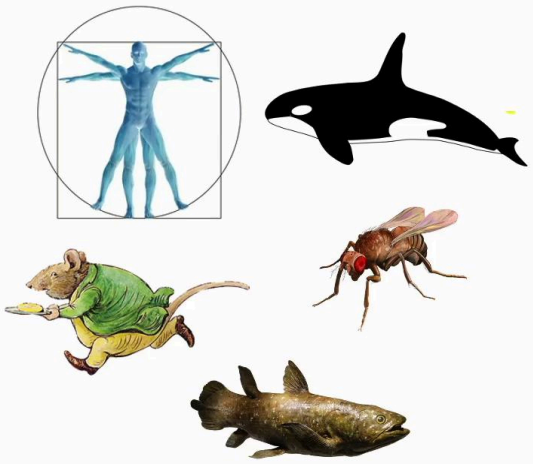
To achieve this, Alphafold 2 uses the input amino acid sequence to the database of protein sequences and constructs a so-called multiple sequence alignment or MSA. An MSA identifies similar but not identical sequences that have been found in living organisms, which enables the identification of those parts of the sequence that are more likely to mutate and allows the algorithm to detect correlations between them. Alphafold 2 also tries to identify proteins that may have a similar structure to the input and constructs an initial representation of the structure in a manner not dissimilar to molecular replacement. This delivers a probabilistic model of which amino acids are likely to be in contact with one another. This can be reentered iteratively until a final structure is output.

Notes

Summary

5m 44s

Alphafold 2 in summary



1 million species

Images: creative commons

- Phase problem solved!
- MIR, MAD, SAD, etc. no longer needed
- MR still workhorse
- Refinement of Alphafold2 predictions
  - Is crystal structure = in-vivo structure?
- July 2022: Number of proteins solved by Alphafold2 since November 2020…

> 200’000’000!!

https://www.nature.com/articles/d41586-022-02083-2

In summary, for the large majority of protein structures, Alphafold 2 has solved the phase problem in macromolecular crystallography. Techniques such as multiple isomorphous replacement, multi-wavelength anomalous dispersion, and single-wavelength anomalous diffraction are now largely redundant. Molecular replacement, now in tandem with Alphafold 2, are the main tools for validation of experimentally mined structures. Interestingly, some differences seen between diffraction-derived structures and Alphafold 2 structures are thought to be not due to inaccuracies in Alphafold 2, but possibly due to the protein structure in a crystalline environment not wholly representing the structure in in-vivo conditions. By the summer of 2022, less than two years after Alphafold was introduced, over 200 million proteins from approximately 1 million species had been predicted. This represents an increase in known structures by a factor of over 1000 and is liable to entirely revolutionise molecular biology and medical research. Of these, 35% have been deemed highly accurate, while another 45% are thought to be sufficiently accurate for many applications.

Notes

Summary

6m 46s

In the next video...

In the following video, we finish this week by considering an example of macromolecular structural studies on plastic-eating enzymes, which includes both more conventional approaches, pre-Alphafold, and modified structures derived by deep learning artificial intelligence, before speculating how structural biology is likely to develop in the next years in this new scientific landscape in which macromolecular crystallography has now been joined by both machine learning algorithms and electron microscopies as tools to tackle this problem.

Notes

Summary

8m 16s