Deep learning for protein structure prediction and design

Tanja Kortemme
Slide credit: modified from Ora Furman (Hebrew University)

- Protein structure prediction – intro and significance
- Alphafold2 / concepts
- Applications: problem solved !?
- Design of new proteins

AI & AlphaFold2 Revolution

- Breakthrough of 2021!

Science

2021 BREAKTHROUGH OF THE YEAR

Protein structures for all

AI-powered predictions show proteins finding their shapes

Check for updates

Method of the Year: protein structure prediction

Nature Methods has named protein structure prediction the Method of the Year 2021.

Vivien Marx

If the Earth moves for you, among other reasons, the causes can be geologic or romantic. In science, in the context of predicting protein structure, you might have felt the ground tremble in late 2020 as you perused the results of the 14th Critical Assessment of Protein Structure Prediction (CASP). In this competition, scientists regularly test the prowess of their methods that computationally predict the intricate twirly-curly three-dimensional (3D) structure of a protein from a sequence of amino acids.

A pleasant frisson may have set in more recently as you browsed the new and rapidly

A top-down view of the human nuclear pore complex, the largest molecular machine in human cells.

ADAPTED FROM REF. 3/ AGNIESZKA OBARSKA-KOSINSKA

WHAT'S NEXT FOR THE AI PROTEIN-FOLDING REVOLUTION

AlphaFold, software that can predict the 3D shape of proteins, is already changing biology.

By Ewen Callaway

For more than a decade, molecular biologist Martin Beck and his colleagues have been trying to piece together one of the world's hardest jigsaw puzzles: a detailed model of the largest molecular machine in human cells.

This behemoth, called the nuclear pore complex, controls the flow of molecules in and out of the nucleus of the cell, where the genome sits. Hundreds of these complexes exist in every cell. Each is made up of more than 1,000 proteins that together form rings around a hole through the nuclear membrane.

These 1,000 puzzle pieces are drawn from more than 30 protein building blocks that interlace in myriad ways. Making the puzzle even harder, the experimentally determined 3D shapes of these building blocks are a potpourri of structures gathered from many species, so don't always mesh together well. And the picture on the puzzle's box – a low-resolution 3D view of the nuclear pore complex – lacks sufficient detail to know how many of the pieces precisely fit together.

In 2016, a team led by Beck, who is based at the Max Planck Institute of Biophysics (MPIB)

234 | Nature | Vol 604 | 14 April 2022

https://www.nature.com/articles/d41586-022-00997-5

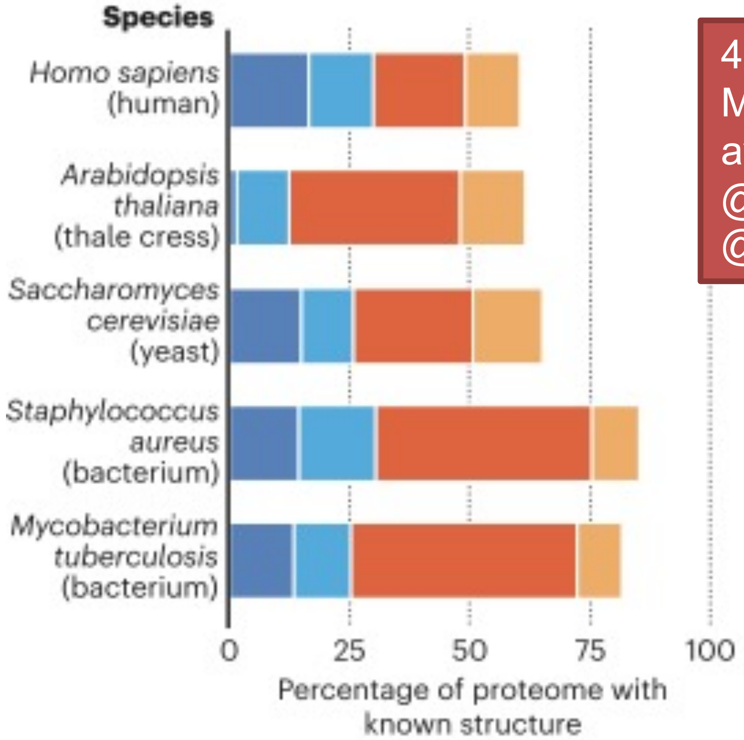
Structural coverage of the proteome

WHAT'S KNOWN ABOUT PROTEOMES

AlphaFold's predictions have greatly increased the proportion of confidently known structures in the human proteome — the collection of all human proteins. The software is even more useful for other species.

Source of knowledge about proteome

- 🟦 High-quality experimental structures in the PDB*
- 🟦 Structural knowledge derived from related proteins in the PDB*
- 🟥 Knowledge from AlphaFold models only (high confidence)
- 🟧 Knowledge from AlphaFold models only (intermediate confidence)



*PDB: Protein Data Bank. AlphaFold can also be used to calculate these structures — but doesn't add significantly to what's already known.

400,000,000 (!) Models available @Uniprot and @EBI

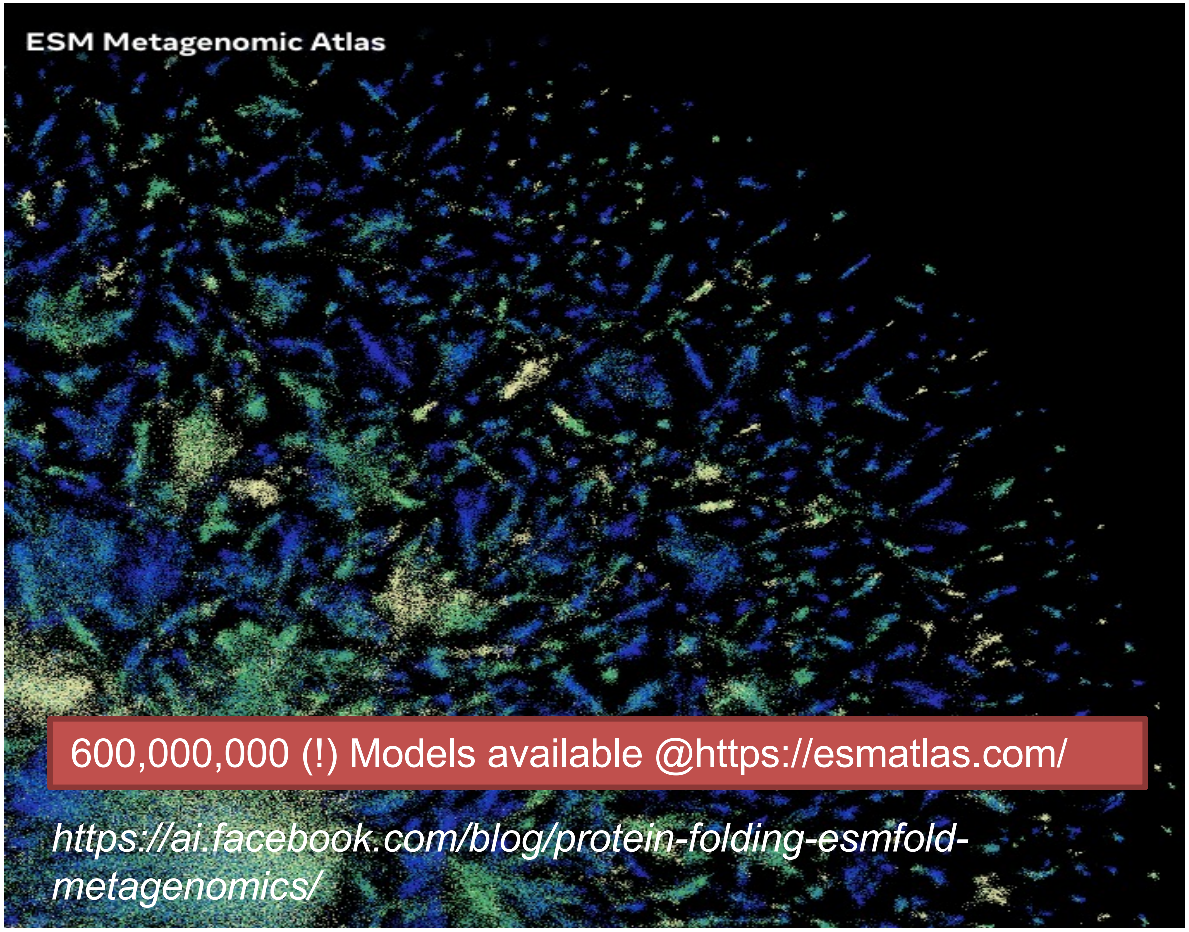
https://www.deepmind.com/research/highlighted-research/alphafold

Structural coverage of the proteome

MetaAI ESMFold

ESM Metagenomic Atlas: The first view of the 'dark matter' of the protein universe

blue: dark matter - no similarity to previous structures)

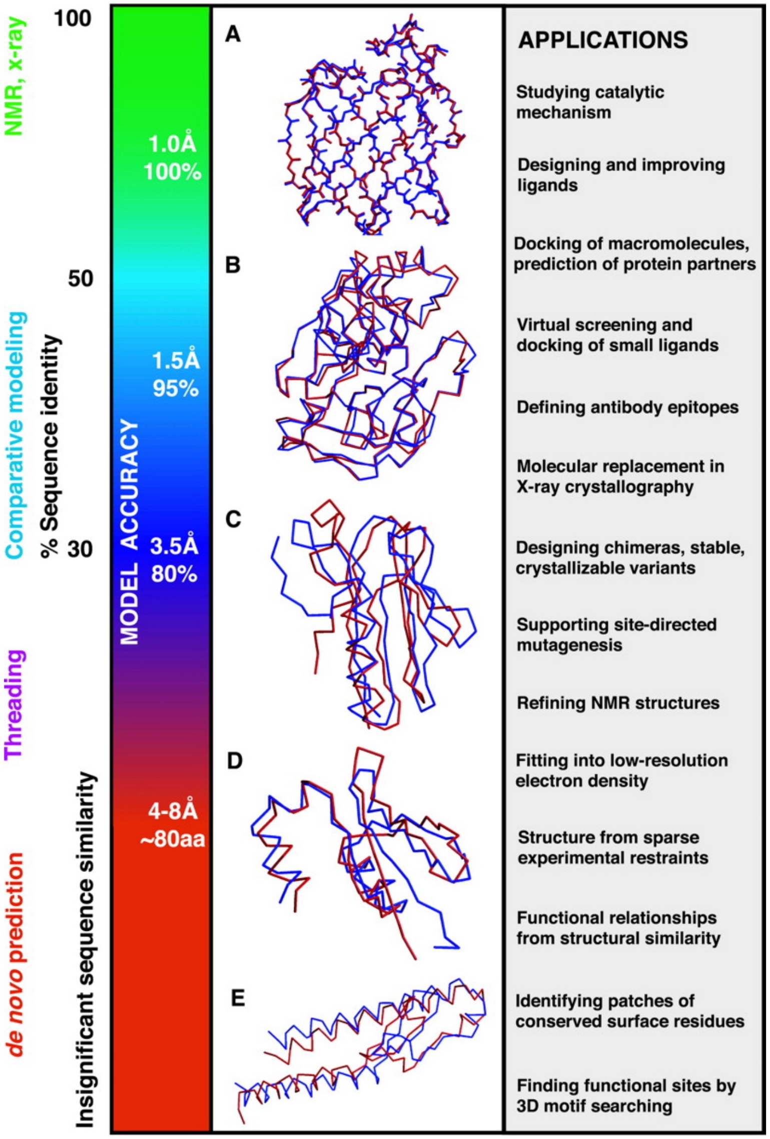


600,000,000 (!) Models available @https://esmatlas.com/

https://ai.facebook.com/blog/protein-folding-esmfold-metagenomics/

Why predict protein structures?

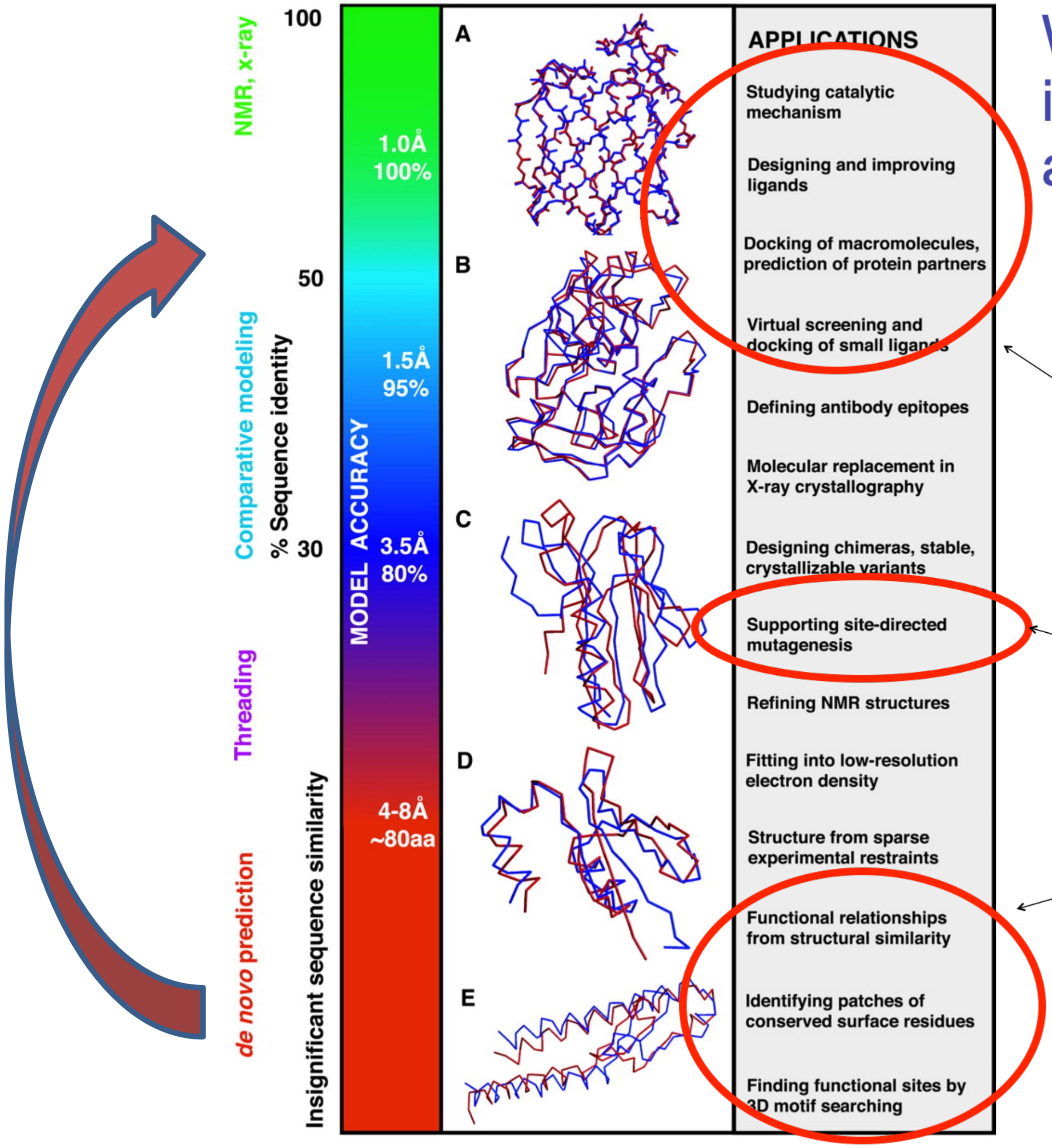
(and what accuracy is needed?)



APPLICATIONS

- Studying catalytic mechanism
- Designing and improving ligands
- Docking of macromolecules, prediction of protein partners
- Virtual screening and docking of small ligands
- Defining antibody epitopes
- Molecular replacement in X-ray crystallography
- Designing chimeras, stable, crystallizable variants
- Supporting site-directed mutagenesis
- Refining NMR structures
- Fitting into low-resolution electron density
- Structure from sparse experimental restraints
- Functional relationships from structural similarity
- Identifying patches of conserved surface residues
- Finding functional sites by 3D motif searching

Protein structure prediction and structural genomics.
Baker D, Sali A. Science. 2001 Oct 5;294(5540):93-6.



APPLICATIONS

Studying catalytic mechanism

Designing and improving ligands

Docking of macromolecules, prediction of protein partners

Virtual screening and docking of small ligands

Defining antibody epitopes

Molecular replacement in X-ray crystallography

Designing chimeras, stable, crystallizable variants

Supporting site-directed mutagenesis

Refining NMR structures

Fitting into low-resolution electron density

Structure from sparse experimental restraints

Functional relationships from structural similarity

Identifying patches of conserved surface residues

Finding functional sites by 3D motif searching

Which modeling accuracy is useful depends on the application

- Drug & protein design
- Docking

Design mutations for experimental tests

Hypotheses for function, effects of genetic variation

Protein structure prediction and structural genomics.
Baker D, Sali A. Science. 2001 Oct 5;294(5540):93-6.

15th Community Wide Experiment on the
Critical Assessment of Techniques for Protein Structure Prediction

http://www.predictioncenter.org/casp15/

CASP

-Blind structure prediction experiment

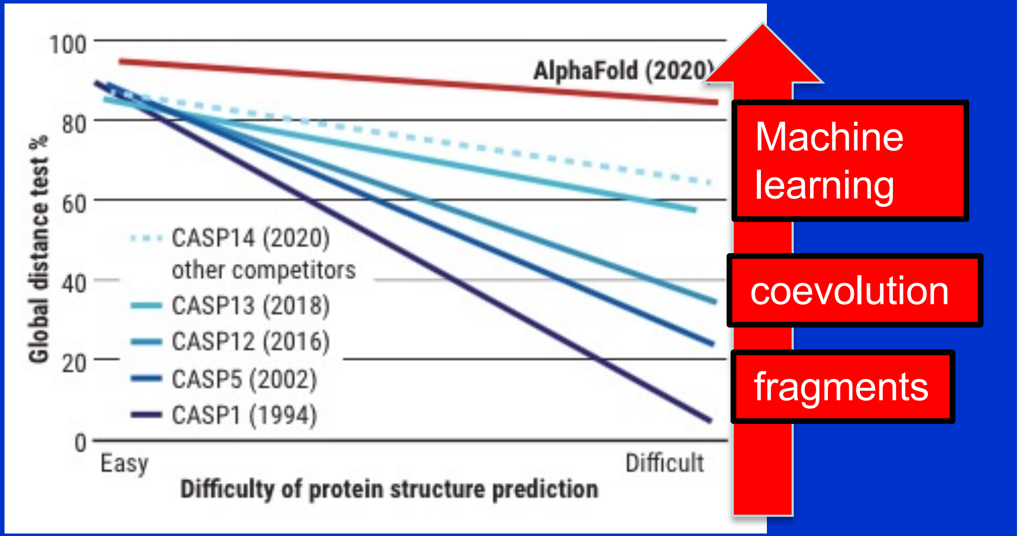
allows assessment of different approaches

- *every 2 years; summer 2022: CASP15*

Identification of major “winner strategies”:

• CASP4: **fragments** (Rosetta)

• CASP11&12: **coevolution** and contact prediction methods (contact-assisted modeling)



Machine learning

coevolution

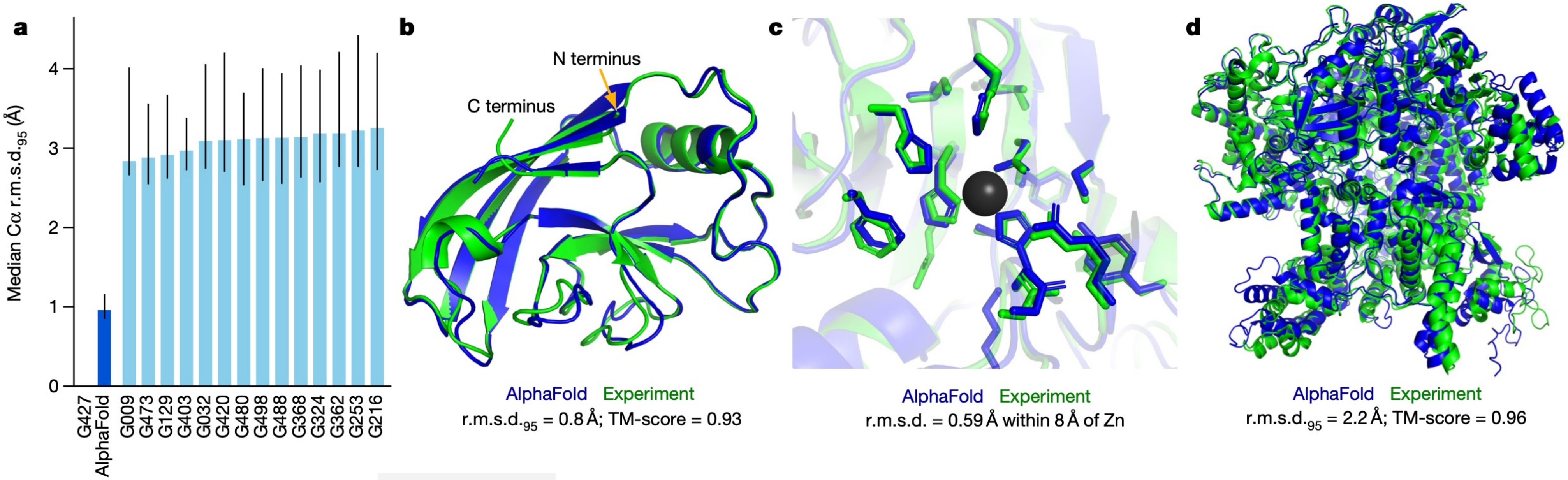
fragments

Starting CASP13: Deep learning (Google alphafold)

CASP14 (2020): Google alphafold deepmind (AF2) “solved the problem”

CASP15 (2022): AF2-**based** methods lead; new, faster approaches using natural language processing models (e.g., ESMFold) accelerate predictions

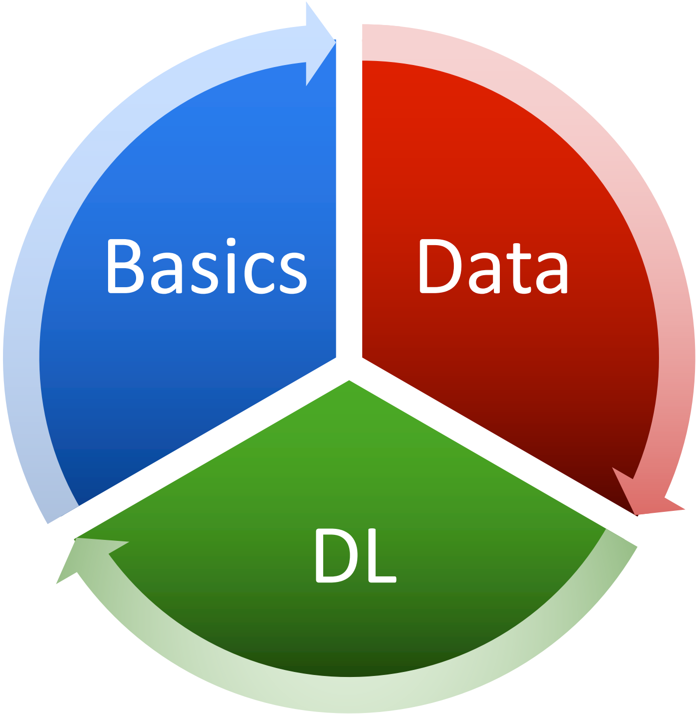
Alphafold2: a game changer (CASP14 – 2020)



Highly accurate protein structure prediction with AlphaFold.
Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. Nature. 2021 Aug;596(7873):583-589. doi: 10.1038/s41586-021-03819-2. Epub 2021 Jul 15.

What made the Alphafold2 breakthrough possible?

- **Basic research** - insights from > 70 years of protein research
- **Big data** - solved structures, large-scale sequencing *etc.*
- **Deep learning** - new architectures, optimization methods



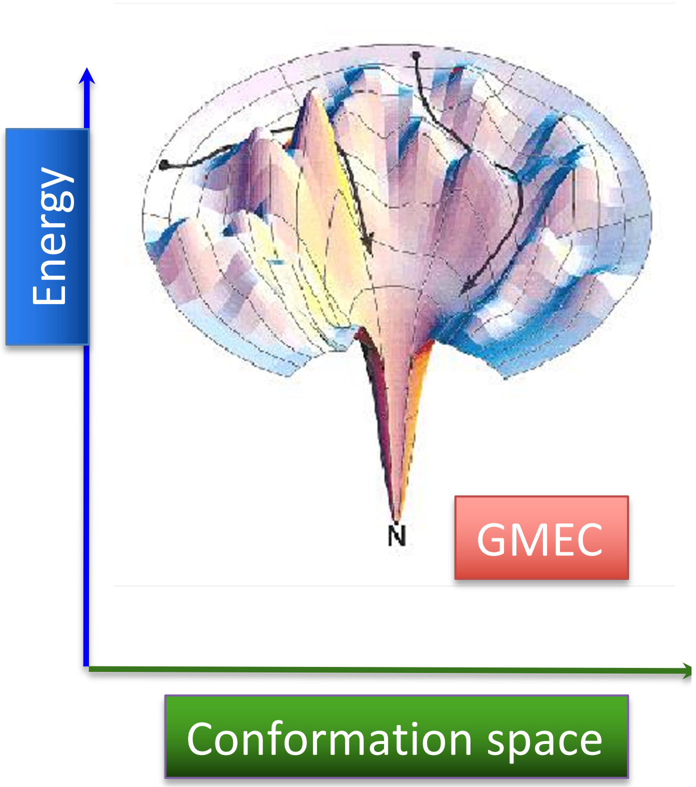
Sequence→ structure

Anfinsen's dogma*:

Native structure determined **only** by sequence

→ Native structure = global energy minimum

- unique
- stable
- kinetically accessible



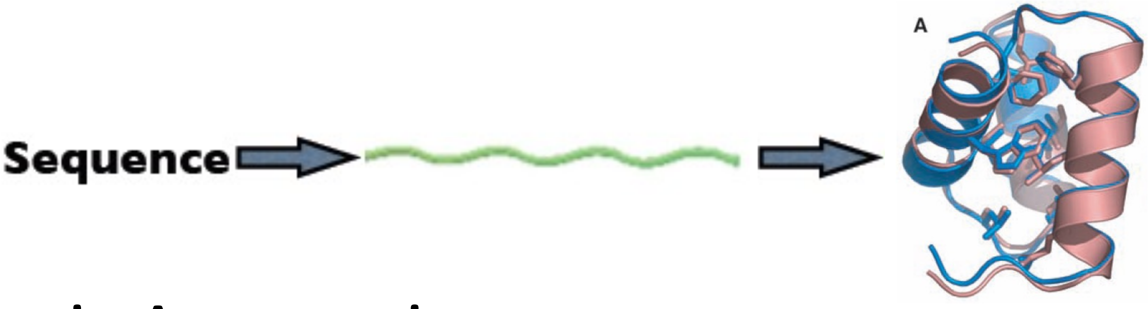
Energy

GMEC

Conformation space

* true at least for a small globular protein, in its standard physiological environment

Structure prediction



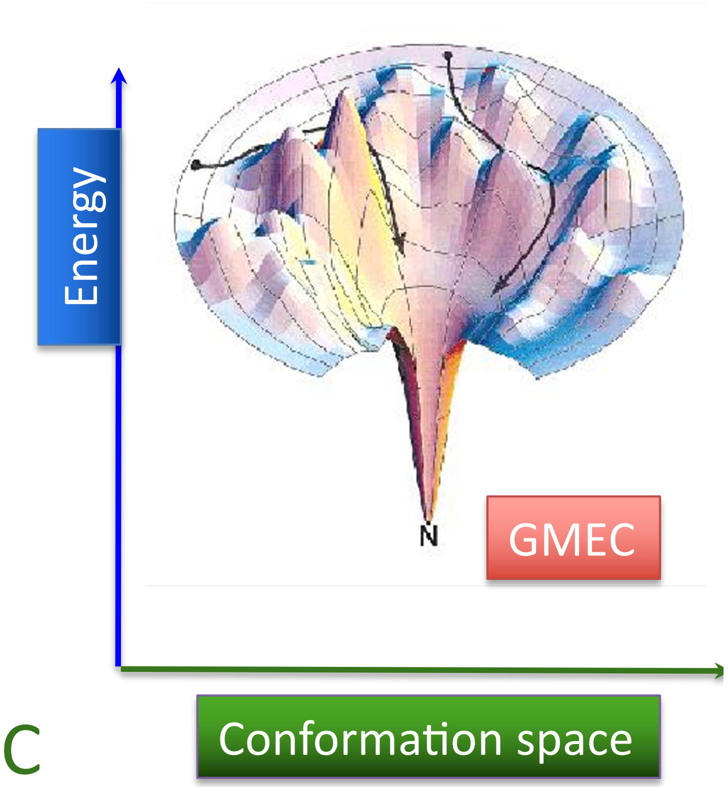
Basic Assumption:

Native structure = GMEC

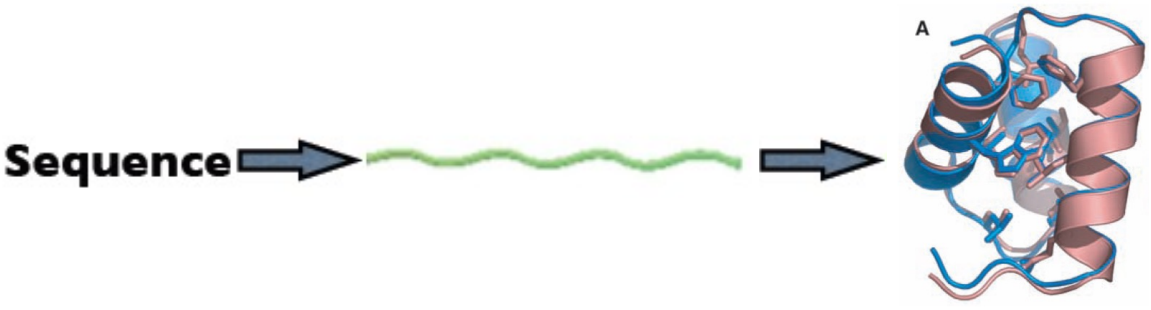
Global Minimum Energy Conformation

→A good energy function selects GMEC

→A **good sampling technique** finds GMEC



Why structure prediction is hard: Conformational space in “ab initio” structure prediction is enormous



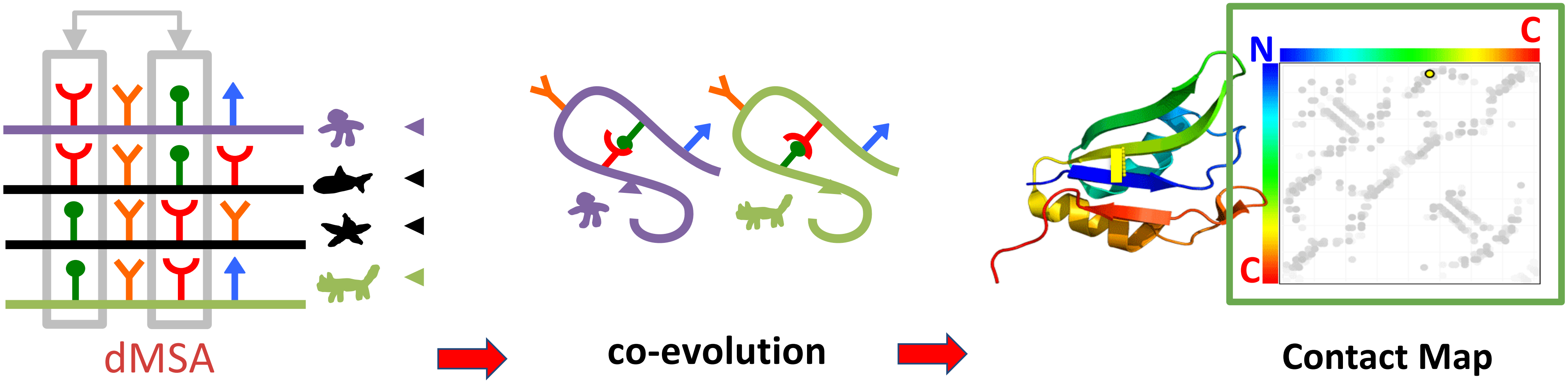
- If only 3 states per residue, 100 residue protein: $3^{100} \sim 5 \times 10^{47}$
- Just considering 3 states isn’t going to be detailed enough
- Clearly need methods to restrict degrees of freedom

Sergey Ovchinnikov

Breakthrough: contact maps

Rosetta GREMLIN (Generative REgularized ModeLs of proteINs)

Long-standing idea: derive residue-residue contacts from sequence information



dMSA

co-evolution

Contact Map

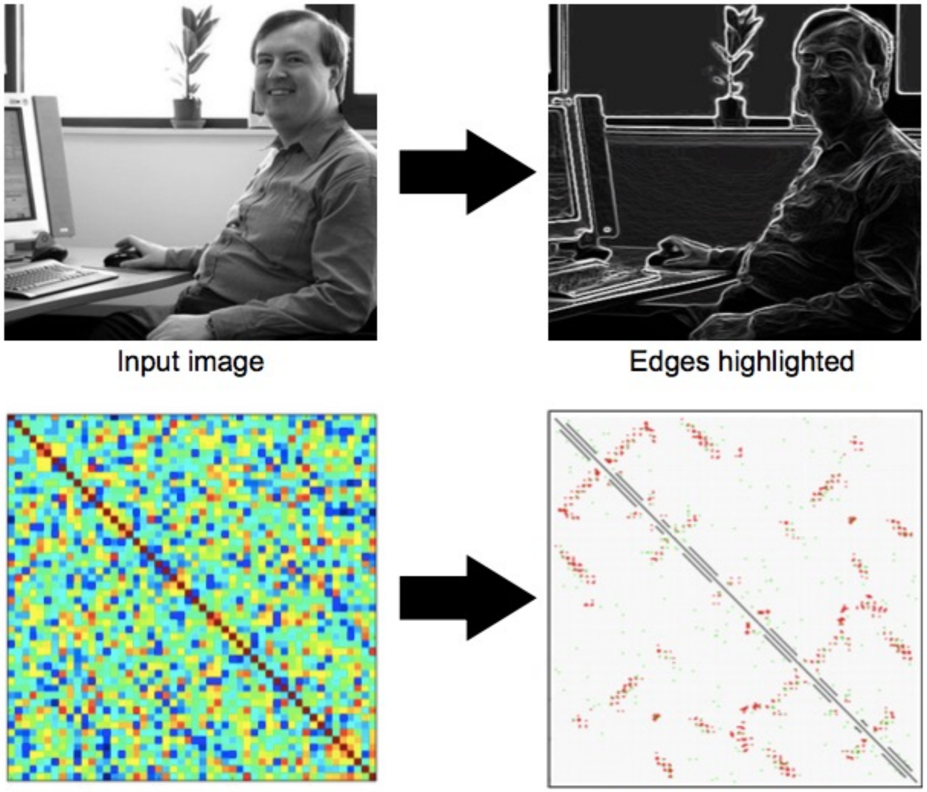
Learning: Apply techniques for object recognition on pictures… cats, street lights, faces, …

Image recognition using Deep NNs

Good at image recognition tasks:

Apply filters to image that highlight specific features

(for example: convoluted neural networks, CNN)



Input image

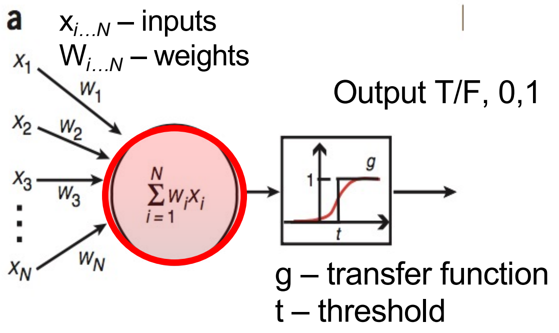
Edges highlighted

Residue covariance matrix

Contact probabilities

Neural Networks

Single Neuron - **linear** separation

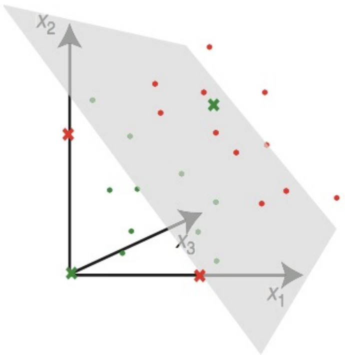

$x_{i...N}$ – inputs

$W_{i...N}$ – weights

Output T/F, 0,1

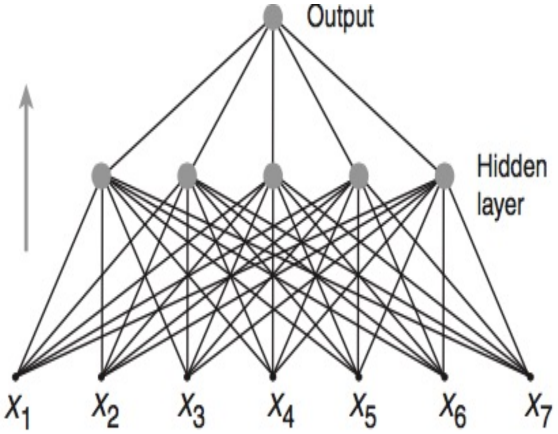
g – transfer function

t – threshold



Problem: not (linearly) separable

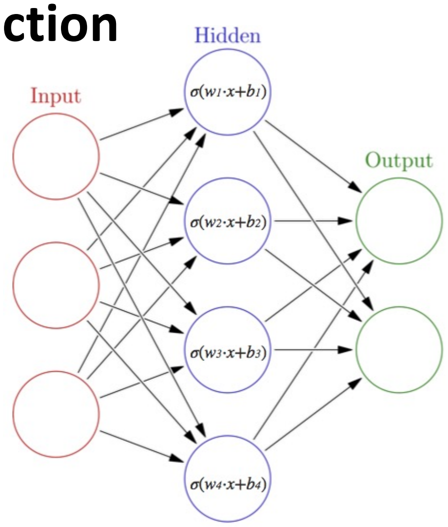
Solution: multiple neurons, **multiple layers**

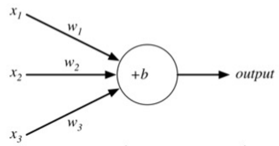


Deep Neural Networks

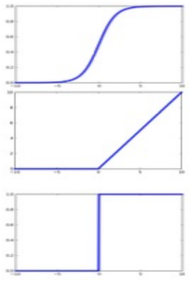
Universal approximation theorem:

A feed-forward neural network with a finite number of nodes can **approximate any continuous function**

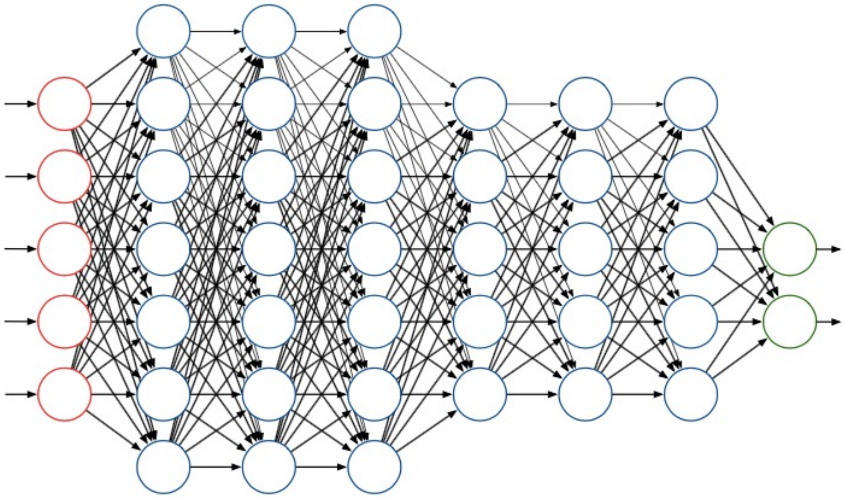




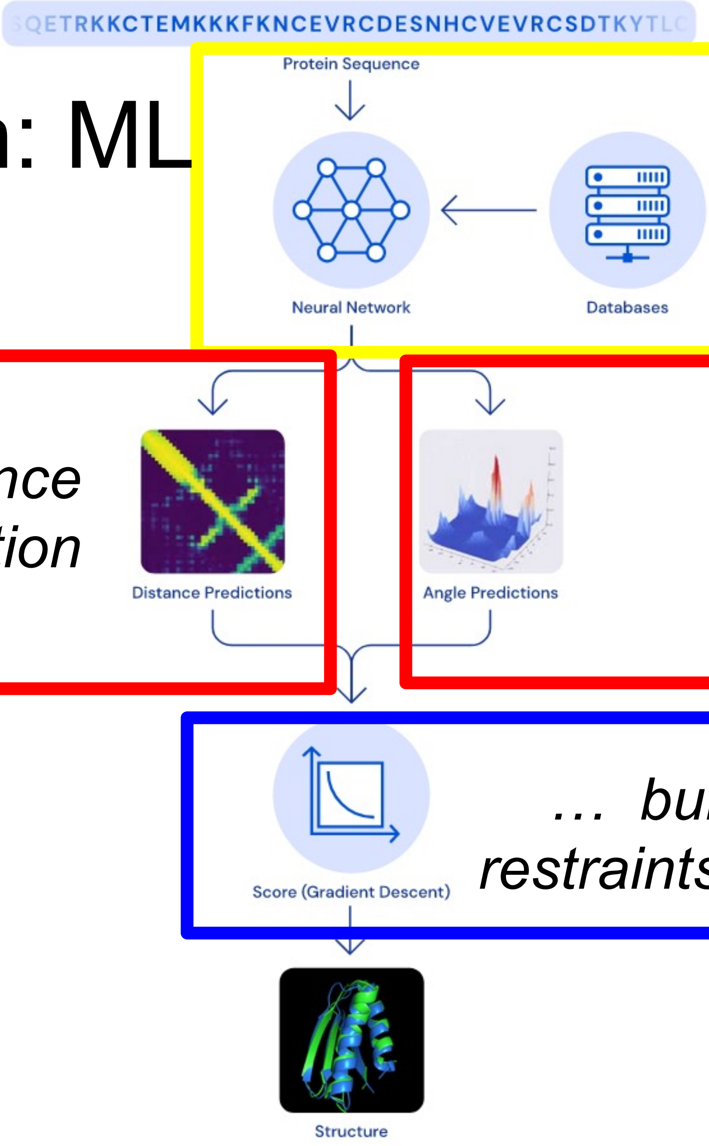
$$output = \sigma\left(\sum_i w_i x_i + b\right)$$



Deep NN: many layers



Breakthrough: ML



Collect sequence data

Contact / Distance prediction

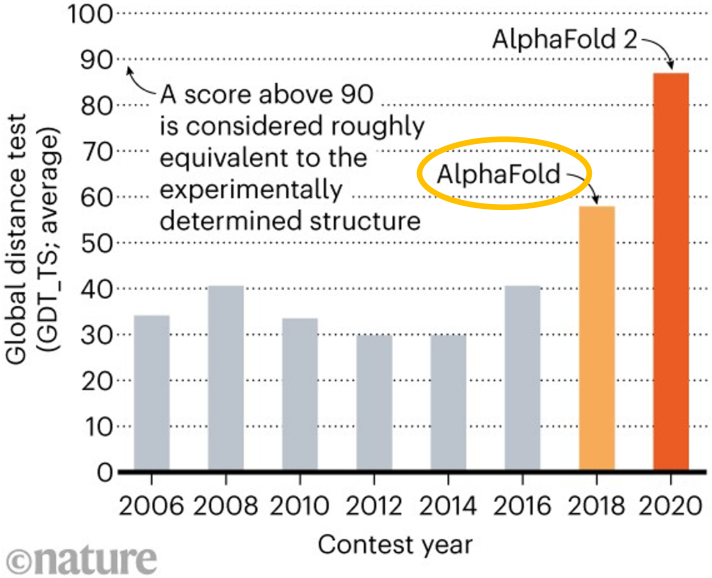
Angle prediction

… build structure from restraints (as in Modeller)

CASP performance

STRUCTURE SOLVER

DeepMind's AlphaFold 2 algorithm significantly outperformed other teams at the CASP14 protein-folding contest — and its previous version's performance at the last CASP.



©nature

Alphafold2: End-to-End architecture

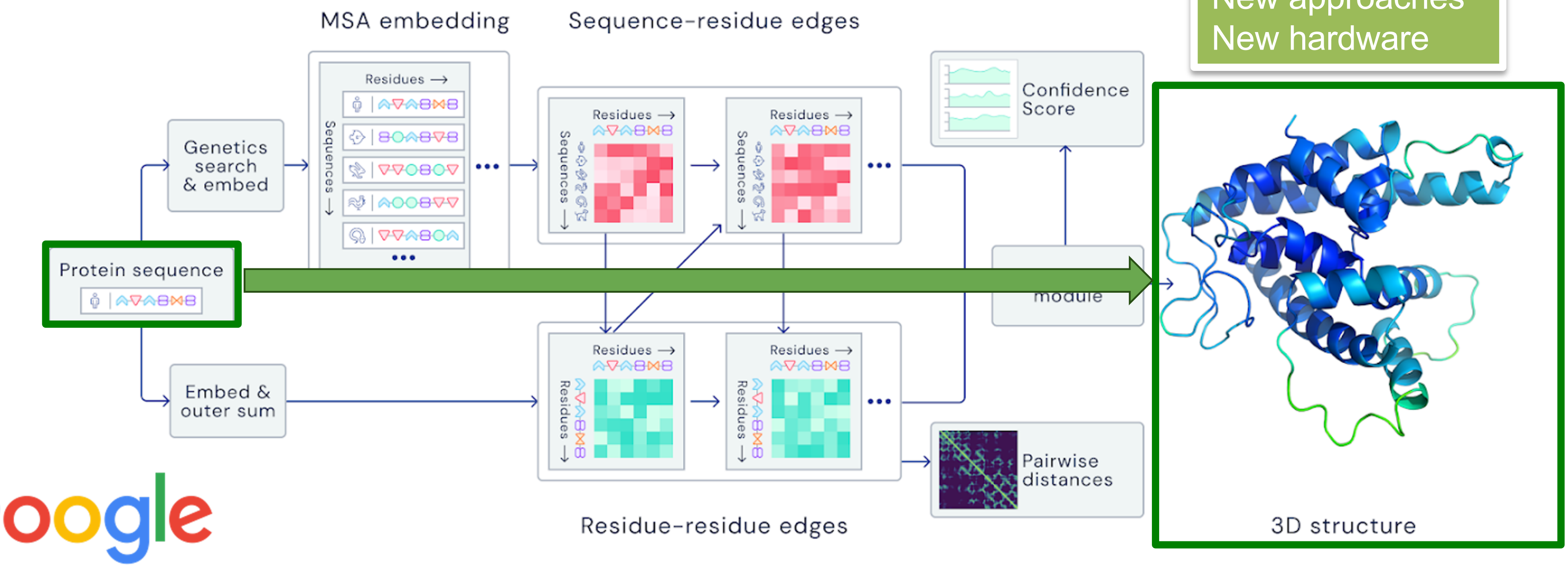
Unprecedented accuracy using novel representations

Trained on publicly available data consisting of ~170,000 protein structures (PDB) & large protein sequence databases.

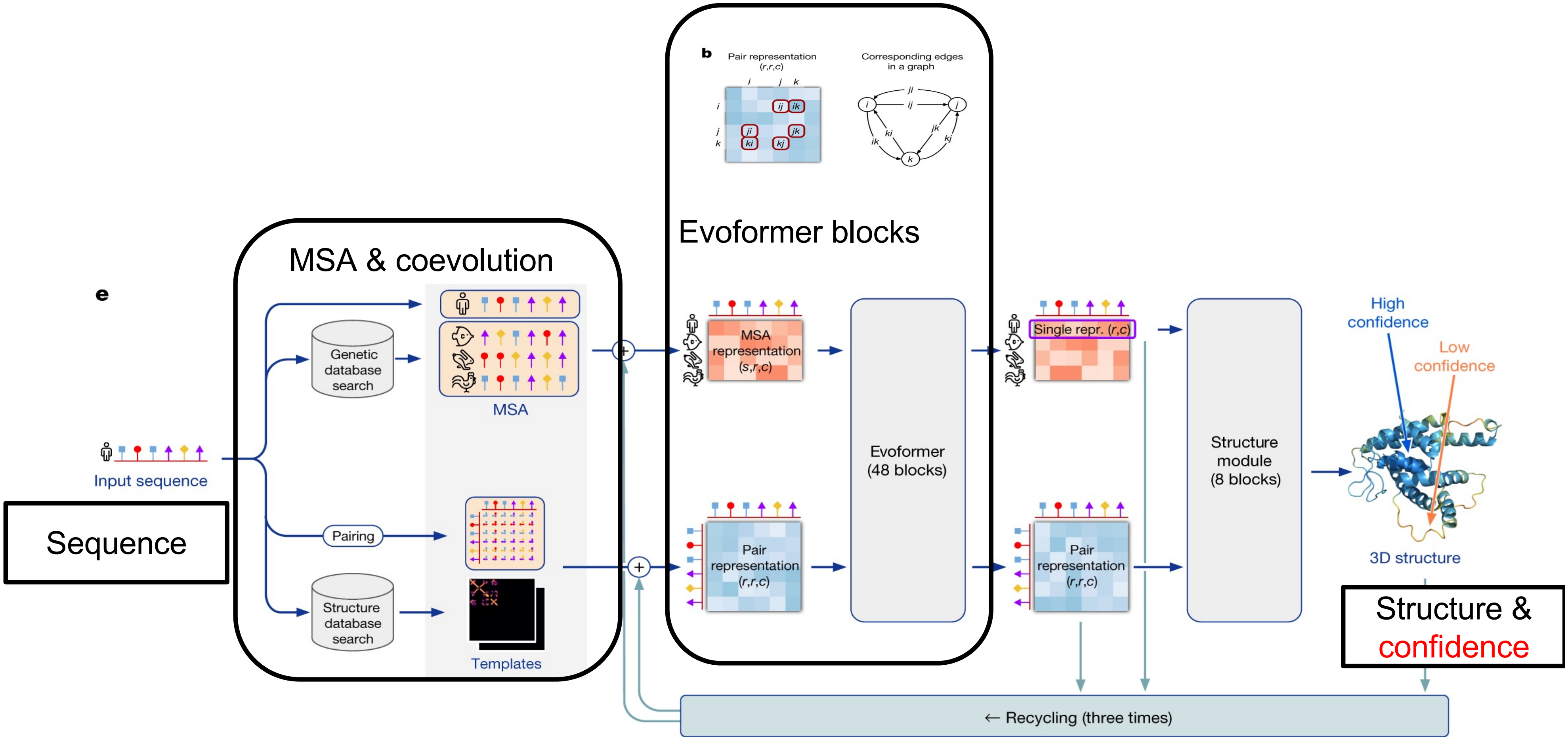
Uses ~ 16 TPUv3s (= 128 TPUv3 cores ~100-200 GPUs) run over a few weeks

Provides confidence value & iterative improvement

New architectures
New approaches
New hardware



Alphafold2 architecture in a nutshell



MSA & coevolution

Evoformer blocks

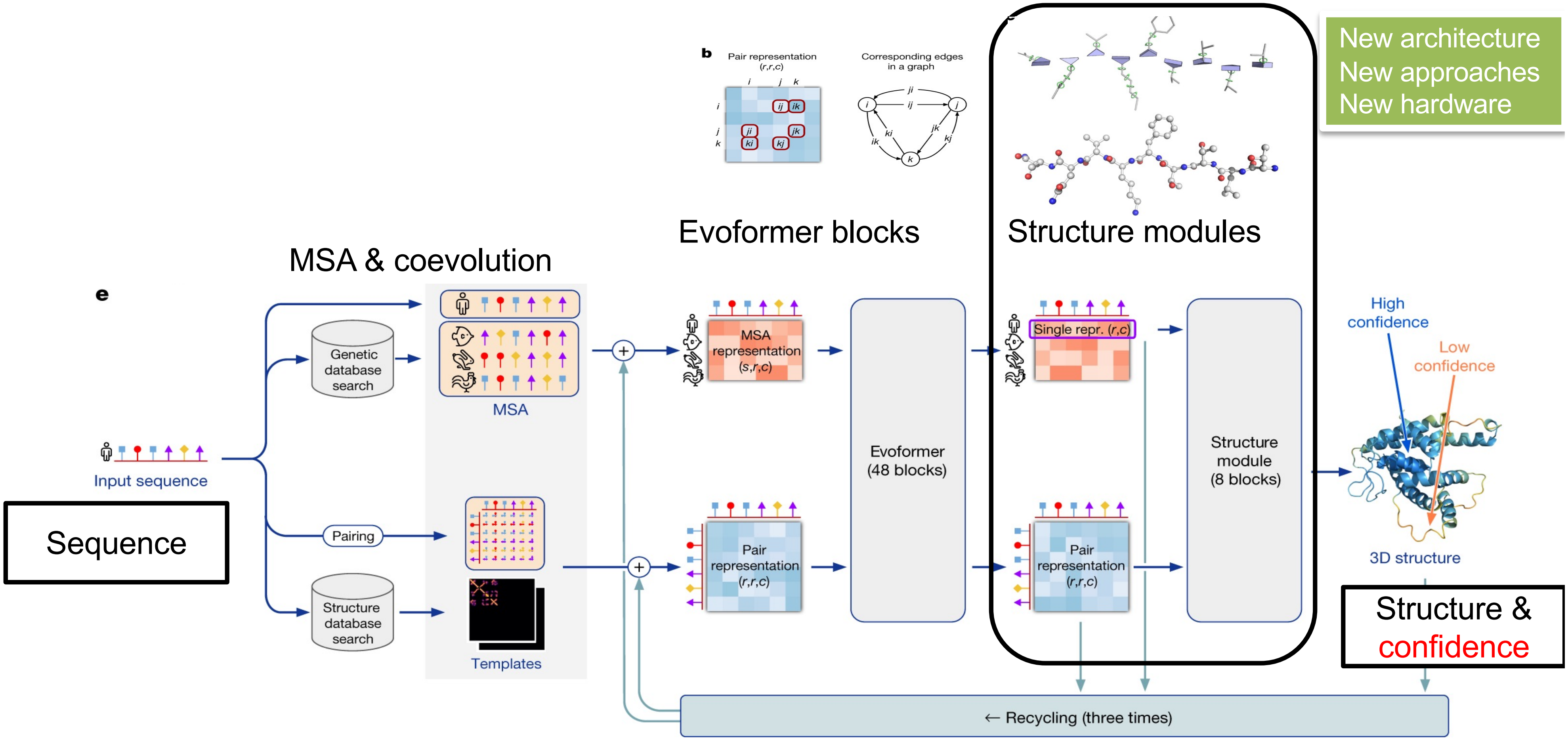
Sequence

Structure & confidence

b
Pair representation (r,r,c)
i j k
ij ik
ji jk
ki kj
Corresponding edges in a graph
i j k
ji ij
ki ik
jk kj
e
Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
+
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

Alphafold2 architecture in a nutshell

New architecture
New approaches
New hardware



MSA & coevolution

Evoformer blocks

Structure modules

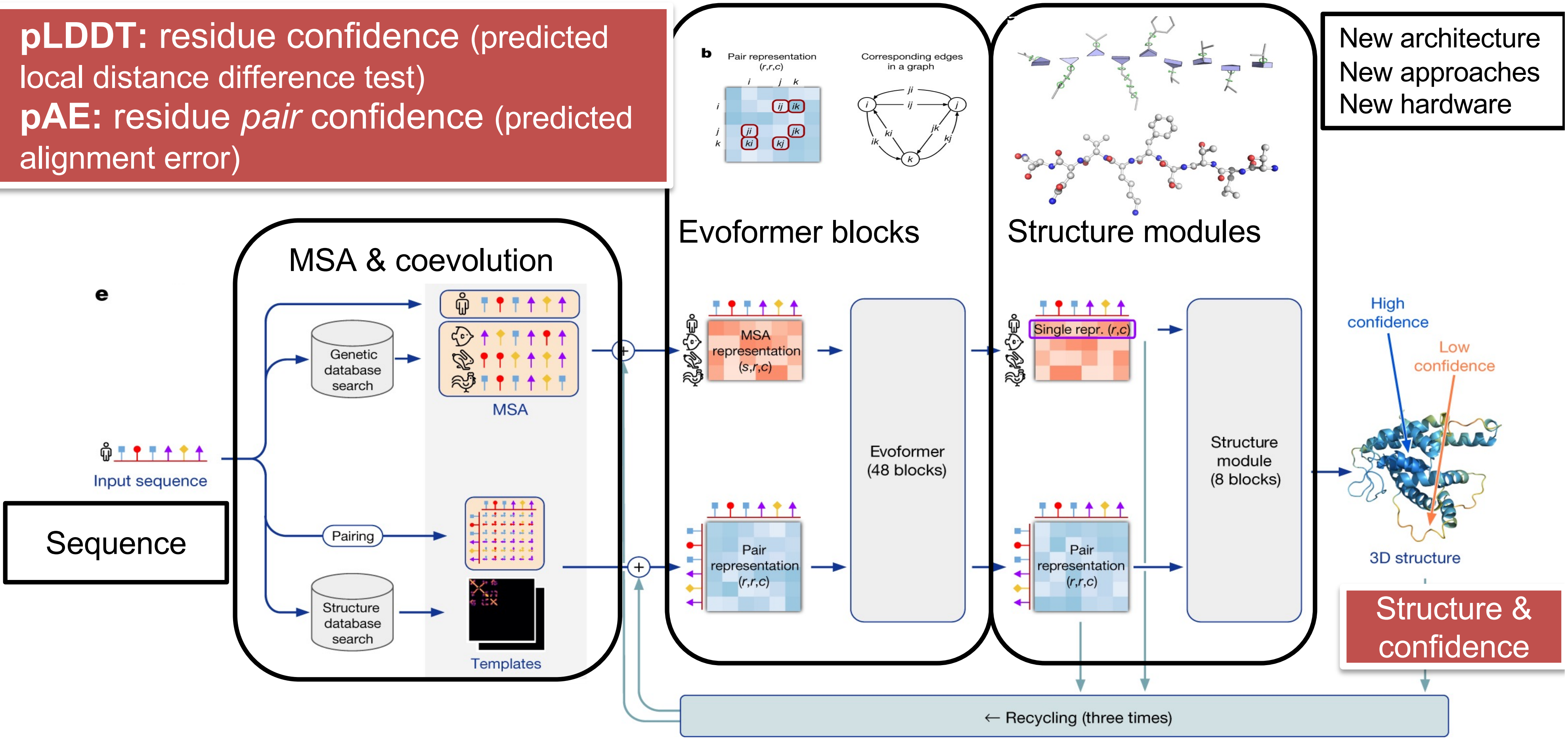
Sequence

Structure & confidence

b
Pair representation (r,r,c)
i
j
k
ij
ik
ji
jk
ki
kj
Corresponding edges in a graph
e
Input sequence
Genetic database search
Pairing
Structure database search
MSA
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

Alphafold2 architecture in a nutshell

- **pLDDT:** residue confidence (predicted local distance difference test)
- **pAE:** residue *pair* confidence (predicted alignment error)

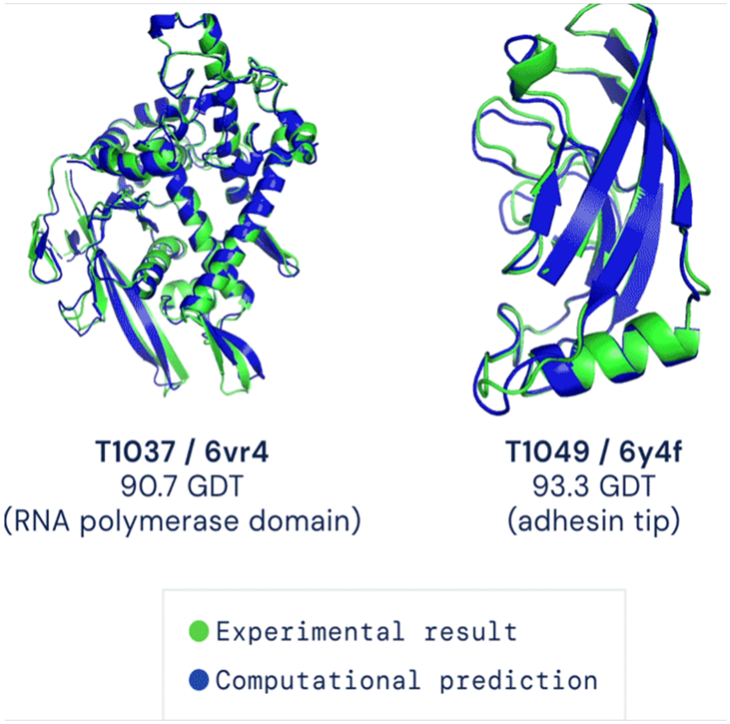


New architecture
New approaches
New hardware

Sequence

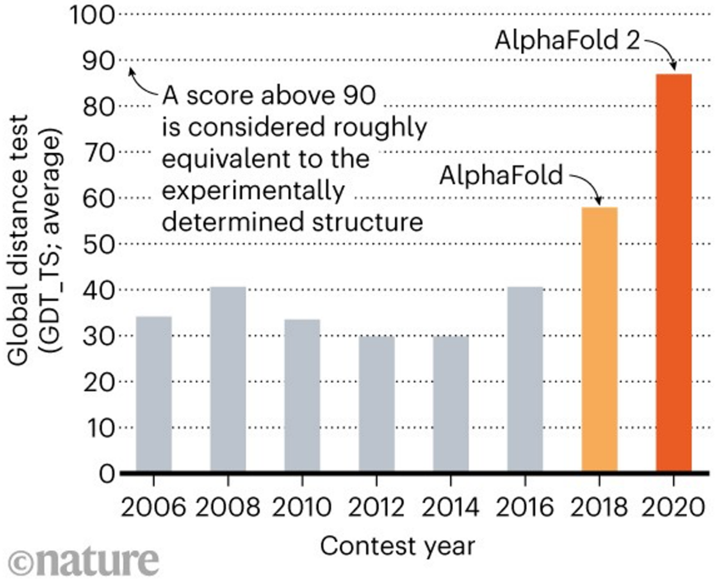
Structure & confidence

Alphafold2: a game changer (CASP14 – 2020)



STRUCTURE SOLVER

DeepMind’s AlphaFold 2 algorithm significantly outperformed other teams at the CASP14 protein-folding contest — and its previous version’s performance at the last CASP.



©nature

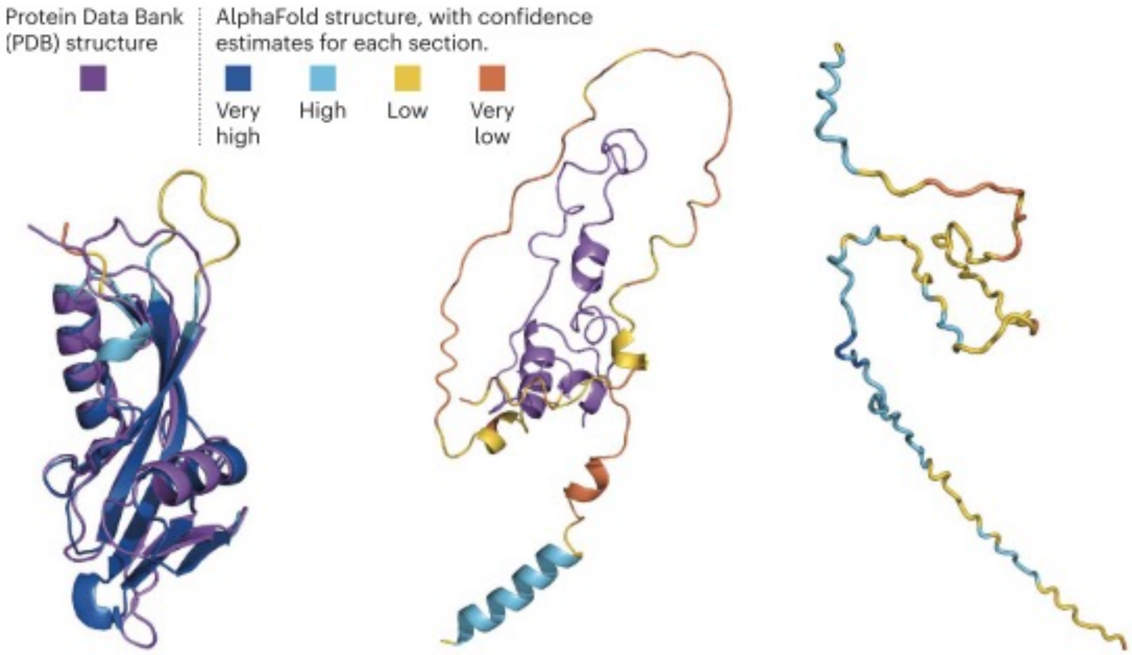
The good, the bad and the ugly

The network also **models the uncertainty in its predictions** - when the s.d. of the predicted distribution is low, the predictions are more accurate:

Confidence measure for each residue:

pLDDT

(predicted local distance difference test)


Protein Data Bank (PDB) structure
AlphaFold structure, with confidence estimates for each section.
Very high
High
Low
Very low

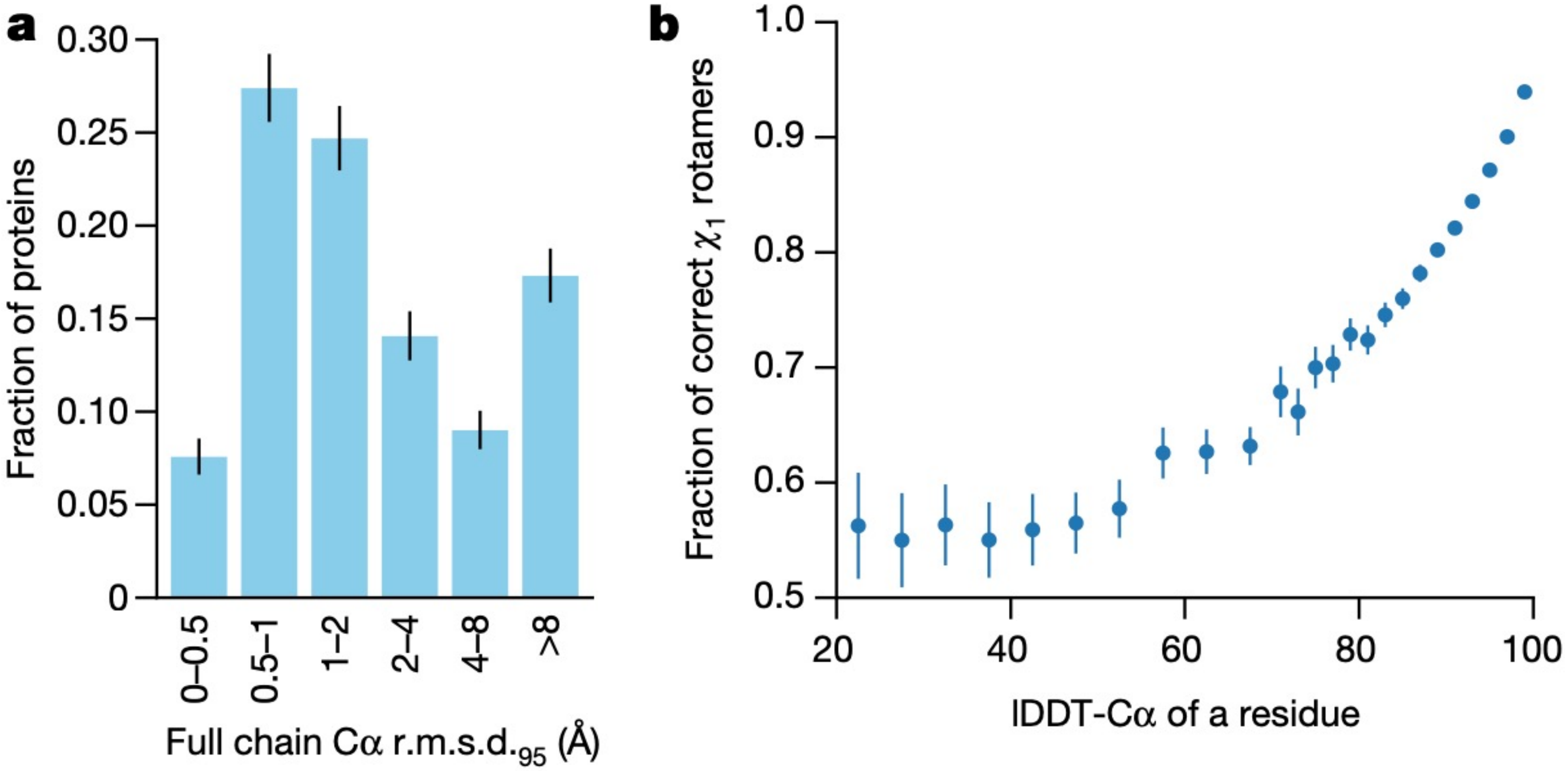
Good
AlphaFold model of phosphohistidine phosphatase overlaps closely with PDB structure.

Bad
AlphaFold model of human insulin bears no relation to the PDB structure.

Ugly
AlphaFold has little confidence across much of its prediction for this human ubiquitin-protein ligase. There is no PDB structure to compare it with.

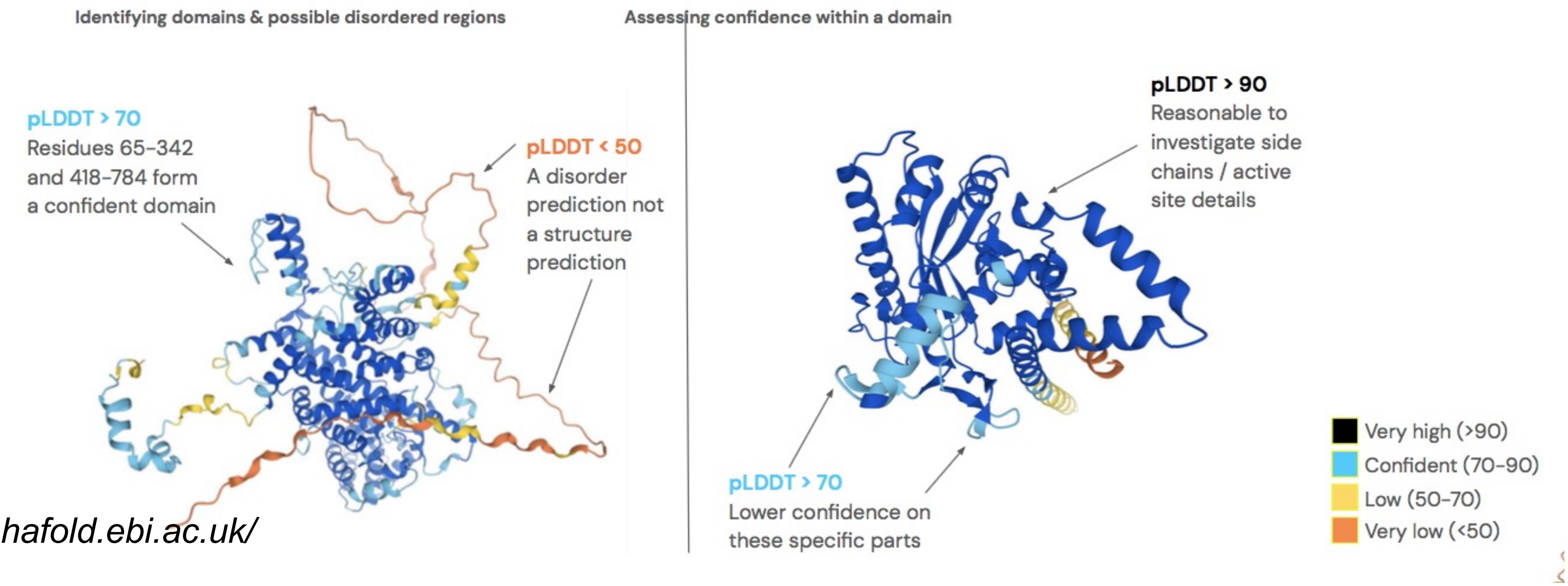
https://www.nature.com/articles/d41586-022-00997-5

Accuracy on recent PDB structures



Interpretation of Alphafold2 models

- pLDDT: predicted **local** distance from solved structure [0..1] > 0.7 precise



Identifying domains & possible disordered regions

pLDDT > 70
Residues 65-342 and 418-784 form a confident domain

pLDDT < 50
A disorder prediction not a structure prediction

Assessing confidence within a domain

pLDDT > 90
Reasonable to investigate side chains / active site details

pLDDT > 70
Lower confidence on these specific parts

■ Very high (>90)
🟦 Confident (70-90)
🟨 Low (50-70)
🟧 Very low (<50)

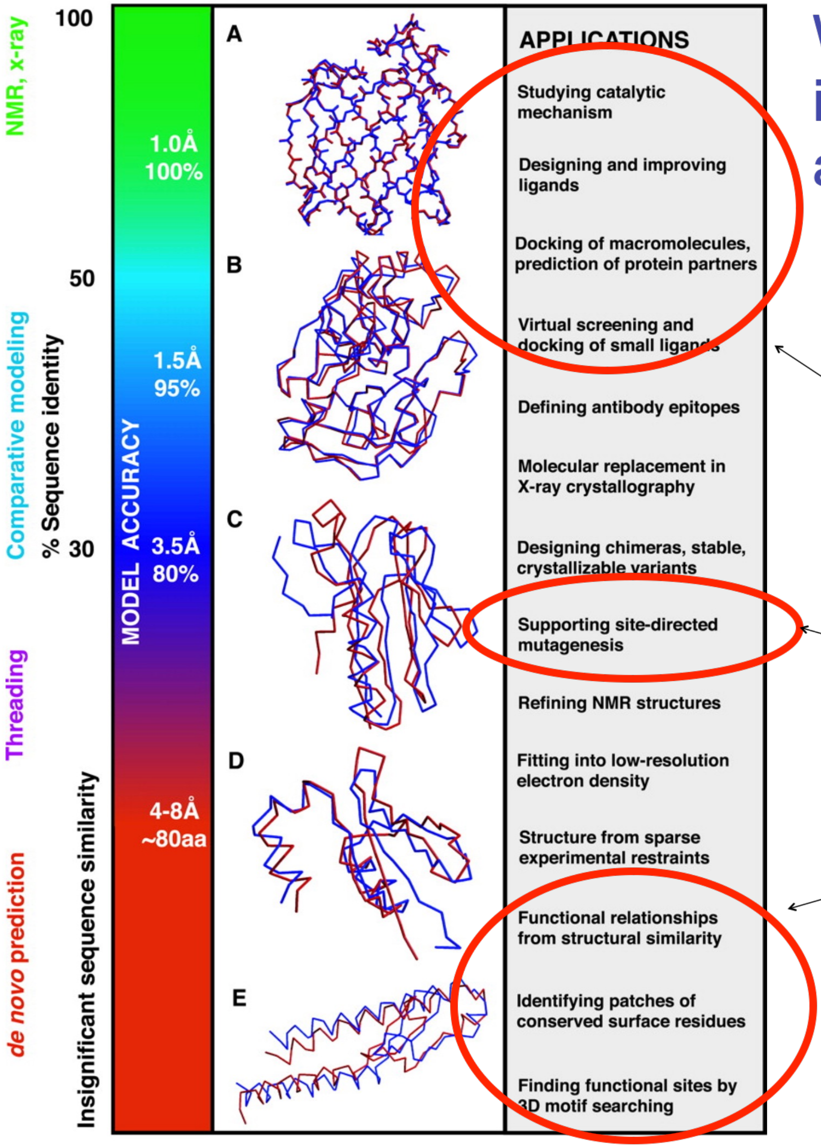
https://alphafold.ebi.ac.uk/

What can be done now? (and what is difficult)

- Combine AF2 predictions with experimental data to create models of complex proteins and assemblies
- Predict structures of complexes (limitation: MSA!)
- In some cases: use predictions for ligand docking

- Disorder? Some indication from pLDDT
- ”Orphan” sequences and de novo proteins – accuracy?
- Prediction of effect of mutations? Difficult!

Which modeling accuracy is useful depends on the application


NMR, x-ray
Comparative modeling
Threading
de novo prediction
100
50
30
% Sequence identity
Insignificant sequence similarity
MODEL ACCURACY
1.0Å
100%
1.5Å
95%
3.5Å
80%
4-8Å
~80aa
A
B
C
D
E
APPLICATIONS
Studying catalytic mechanism
Designing and improving ligands
Docking of macromolecules, prediction of protein partners
Virtual screening and docking of small ligands
Defining antibody epitopes
Molecular replacement in X-ray crystallography
Designing chimeras, stable, crystallizable variants
Supporting site-directed mutagenesis
Refining NMR structures
Fitting into low-resolution electron density
Structure from sparse experimental restraints
Functional relationships from structural similarity
Identifying patches of conserved surface residues
Finding functional sites by 3D motif searching

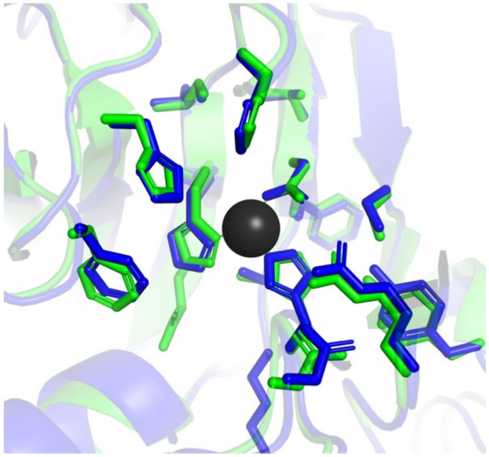
- Drug & protein design
- Docking

Design mutations for experimental tests

Hypotheses for function, effects of genetic variation

Protein structure prediction and structural genomics.
Baker D, Sali A. Science. 2001 Oct 5;294(5540):93-6.

Why are AI models often insensitive to mutations?



AlphaFold Experiment
r.m.s.d. = 0.59 Å within 8 Å of Zn

- In the example, the metal binding site is predicted accurately even though the metal was not included!
- Methods trained on metal-bound structures recognize the pattern of a metal binding site (even if a structure unfolds in the absence of the metal)

Summary : Structure prediction

Enormous recent progress, enabled by:

large databases of sequences and structures, AI methods from other fields, new deep learning network architectures, hardware, computing power

  - Informative and large sequence alignment is (typically) critical, but many sequences are available today (metagenomic data)

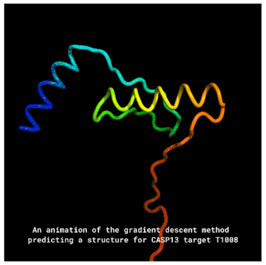
- ML & END-TO-END models (Alphafold2, ESMfold and more to come !)
- Language models to learn the Protein language (fast, perhaps more general?)

Accessible to all:

  - Models available in Uniprot, EBI, MetaAI
  - Modeling made easy on COLAB

Challenges: multiprotein assemblies, disordered proteins, mutations

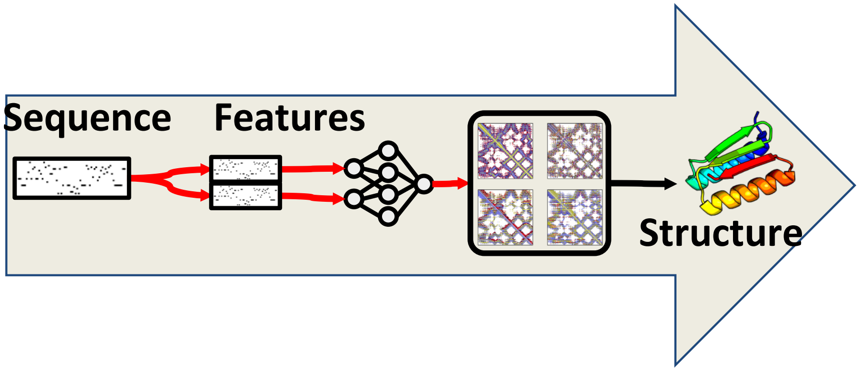
Outlook


An animation of the gradient descent method predicting a structure for CASP13 target T1008

New applications

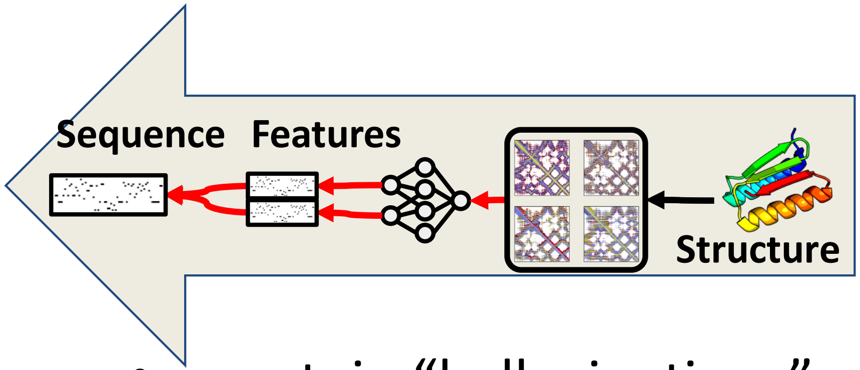
Fast and accurate

- structures for research & medicine
- drug design



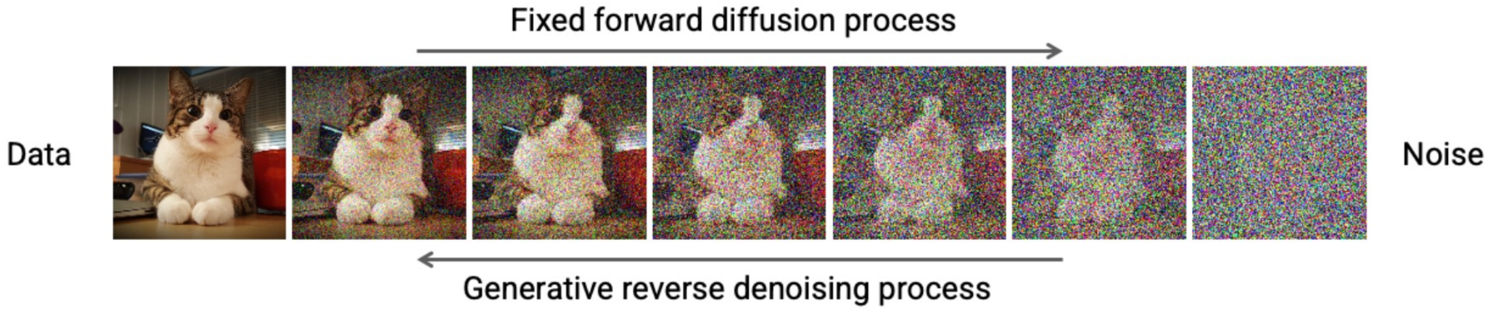
Extend to protein design

- inverse direction:

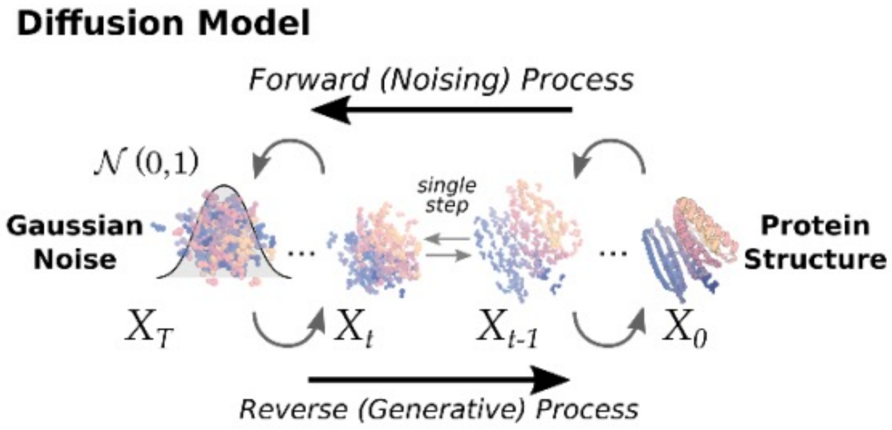


- protein “hallucinations”: Dream new proteins with the NN and much much more

Diffusion models for protein design


Fixed forward diffusion process
Data
Noise
Generative reverse denoising process

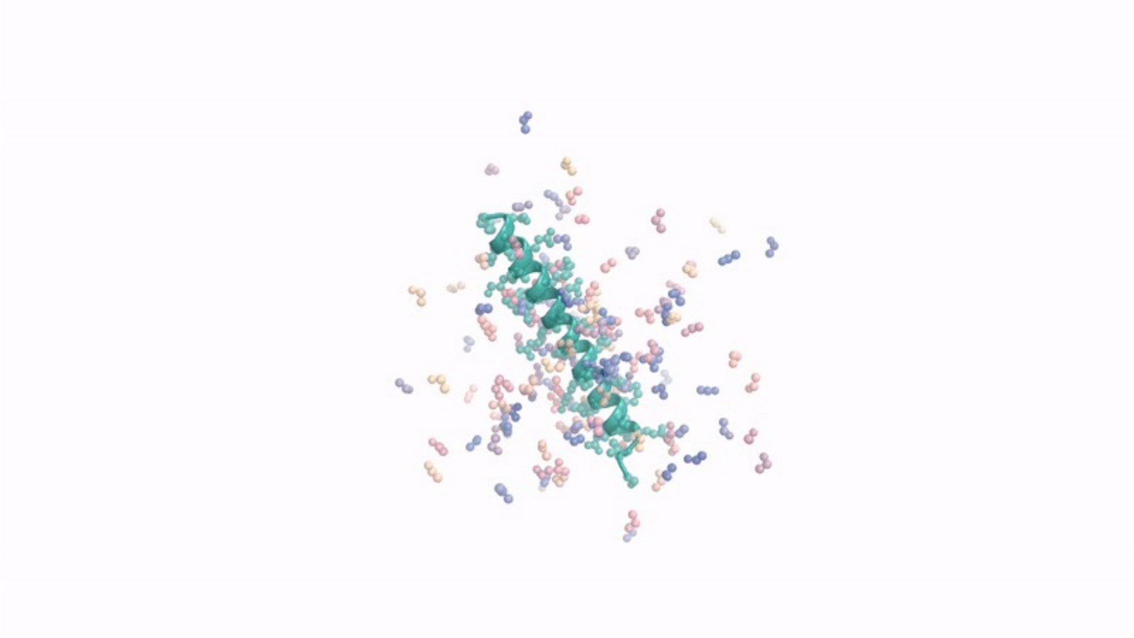
https://cvpr2022-tutorial-diffusion-models.github.io



https://www.biorxiv.org/content/10.1101/2022.12.09.519842v2.full.pdf

Diffusion model for protein design

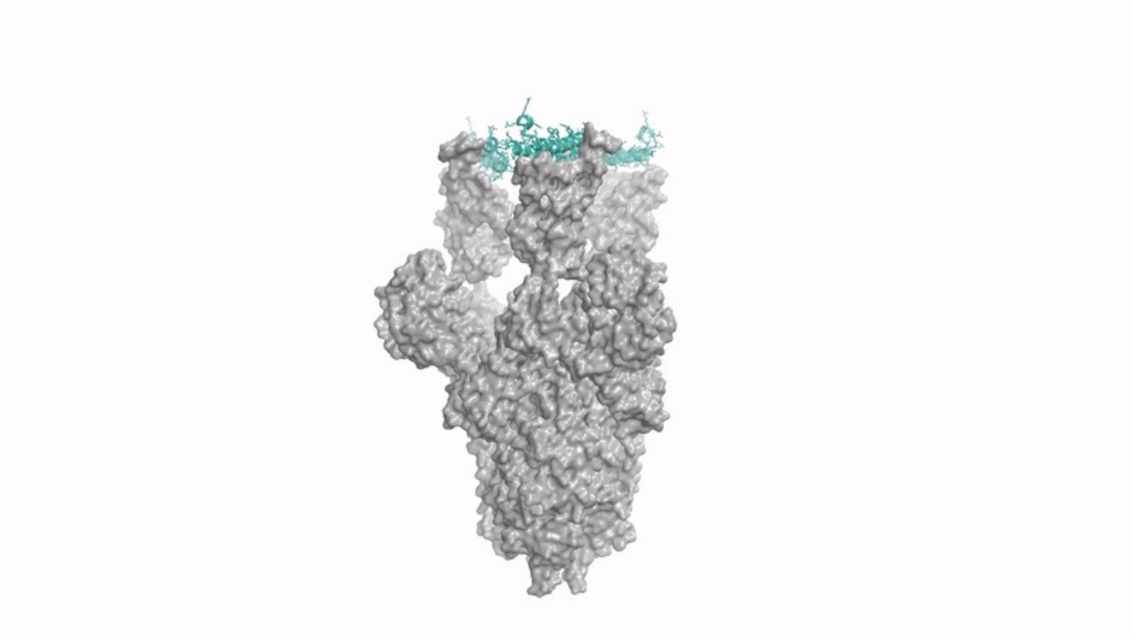
Generate a protein that binds to a helix:



https://www.ipd.uw.edu/2022/12/a-diffusion-model-for-protein-design/

Diffusion model for protein design

Make assemblies



https://www.ipd.uw.edu/2022/12/a-diffusion-model-for-protein-design/

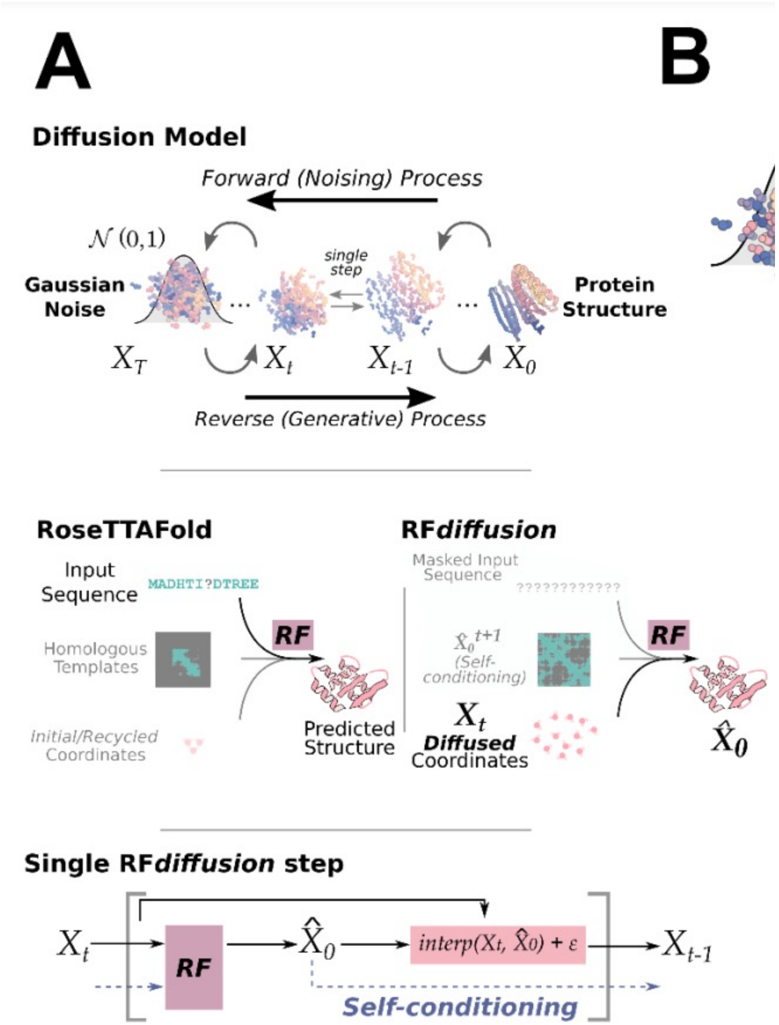


Figure 1: RF*diffusion* is a denoising diffusion probabilistic model with RoseTTAFold **fined-tuned as the denoising network. A)** Top panel: Diffusion models for proteins are trained to recover structures of proteins corrupted with noise, and generate new structures by reversing the corruption process through iterative denoising of initially random noise X_T into a realistic structure X_0. Middle panel: RoseTTAFold (RF, left) can be fine-tuned as the denoising network in a DDPM. RF*diffusion* (right) is trained from a *pre-trained* RF network with minimal architectural changes. While in RF, the primary input to the model is sequence, in RF*diffusion*, the primary input is diffused residue frames. In both cases, the model predicts final 3D coordinates directly (denoted $\hat{X}_0$ in RF*diffusion*). In RF*diffusion*, the model receives its previous prediction as a template input ("self-conditioning", see Methods 2.4). Bottom panel: At each timestep "t" of a design trajectory (typically 200 steps), RF*diffusion* takes X_t and $\hat{X}_0^{t+1}$ from the previous step and then predicts an updated X_0 structure ($\hat{X}_0^{t}$). The coordinate input to the model at the next time step (X_{t-1}) is generated by a noisy interpolation toward $\hat{X}_0^{t}$. B) RF*diffusion* is of broad

https://www.biorxiv.org/content/10.1101/2022.12.09.519842v2.full.pdf