AlphaFold 2

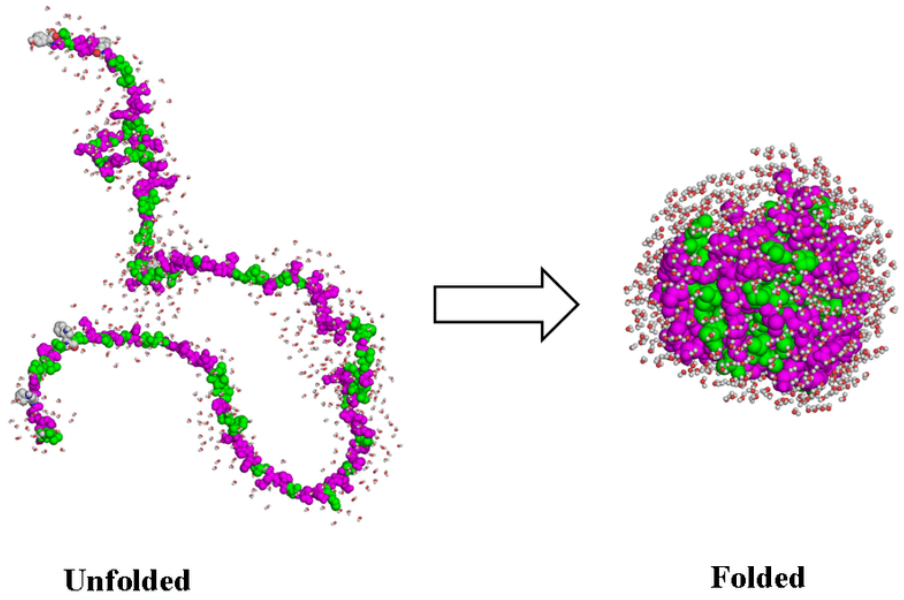
“Method of the Year 2021”

Protein folding

Proteins are chains of amino acids and are the workehorses of living organisms

Proteins provide structure to cells, move and carry molecules, catalyse reactions, etc.

Protein functions are mediated by their **structure**



- **Uniqueness:** Sequences (not always!) map 1:1 to a single 3D structure
- **Function and disease:** the correct structure mediates the function. If a protein misfolds → disease

Experimental methods for protein structure determination

X-ray chrystallography

- 100K– 200K per structure
- Time consuming
- Expensive

Cryo-EM

- Lower resolution

NMR spectroscopy

- Structures pbserved in solution («in vivo»)
- Limited to small proteins
- Time consuming
- Needs large amount of sample

Available datasets: around 200 million sequences (UniProt) but only 18’000 structures (PDB)

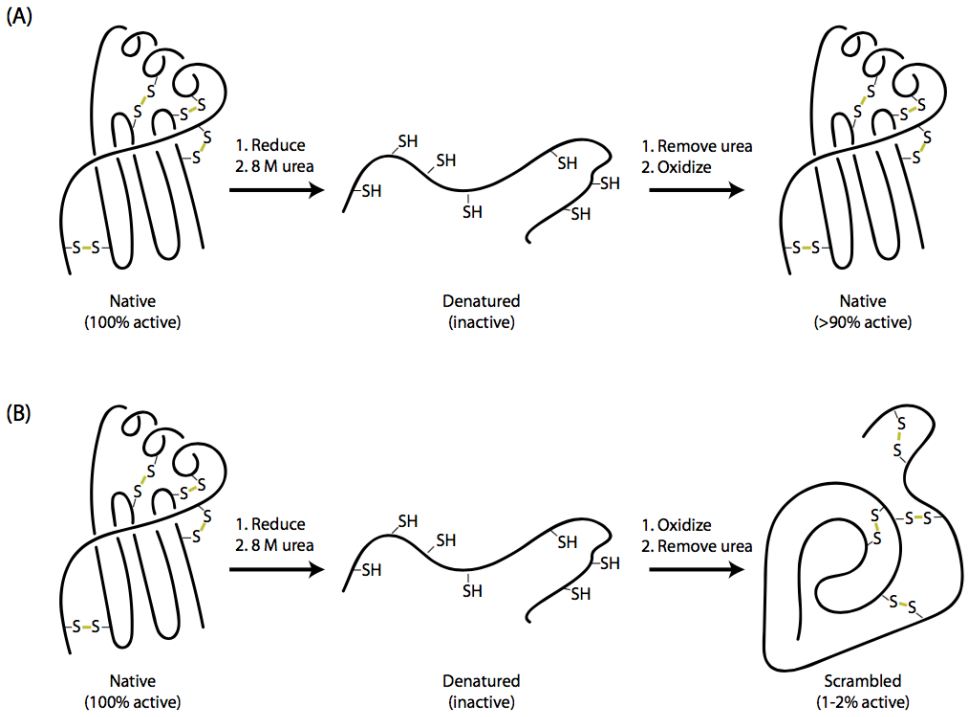
⇩

Computational methods to predict protein structure

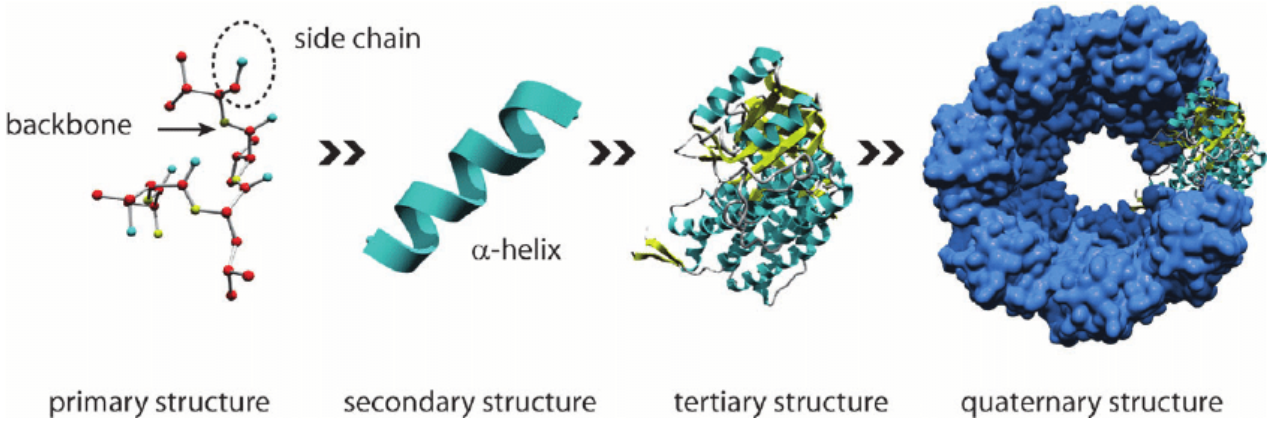
Protein prediction problem

lowest; that is, that the native conformation is determined by the totality of interatomic interactions and hence by the amino acid sequence, in a *given environment*. In terms of natural selection through the "design" of macromole-

Anfinsen, Nobel Prize lecture (1972)



Anfinsen's experiment on ribonuclease A demonstrated that all the information required to fold a protein into its native, lowest-energy conformation is entirely contained within its sequence of amino acids



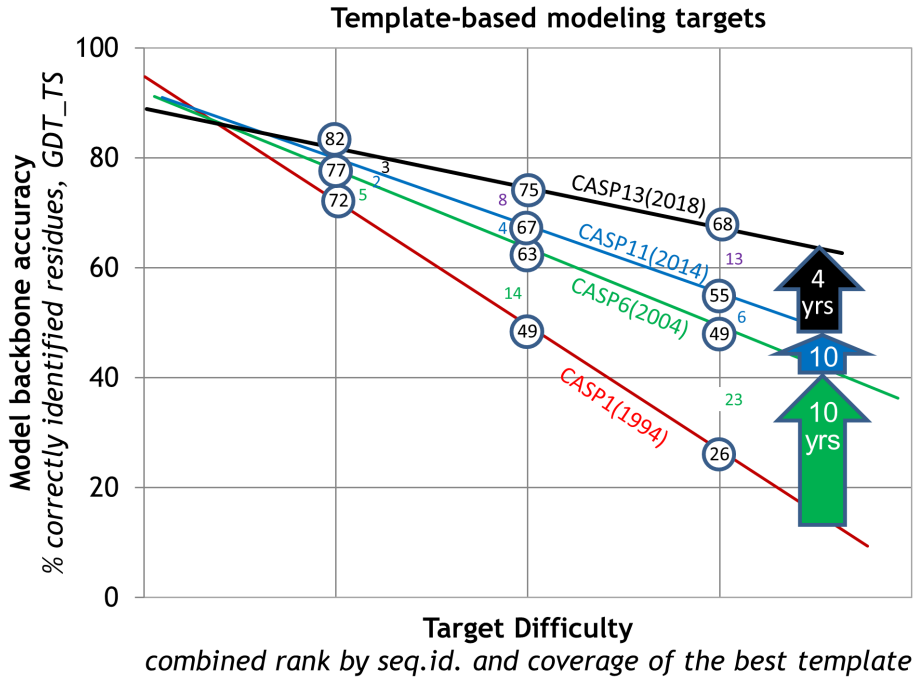
Critical Assessment of Protein Structure Prediction (CASP)

Every 2 years, structural and computational biology reserach groups meet and take part to the CASP competition

The aim of the competition is to evaluate the state of the art in protein structure prediction

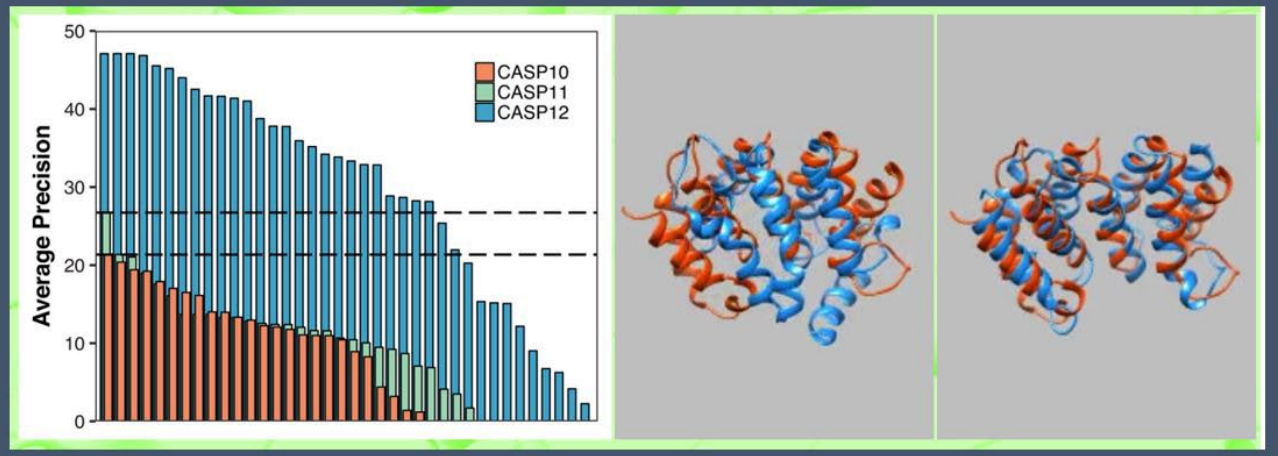
Each group tests its algorithm on protein sequences whose structure has not been published yet

The predictions are based only on the primary sequence of the proteins



CASP13

Critical Assessment of Protein Structure Prediction



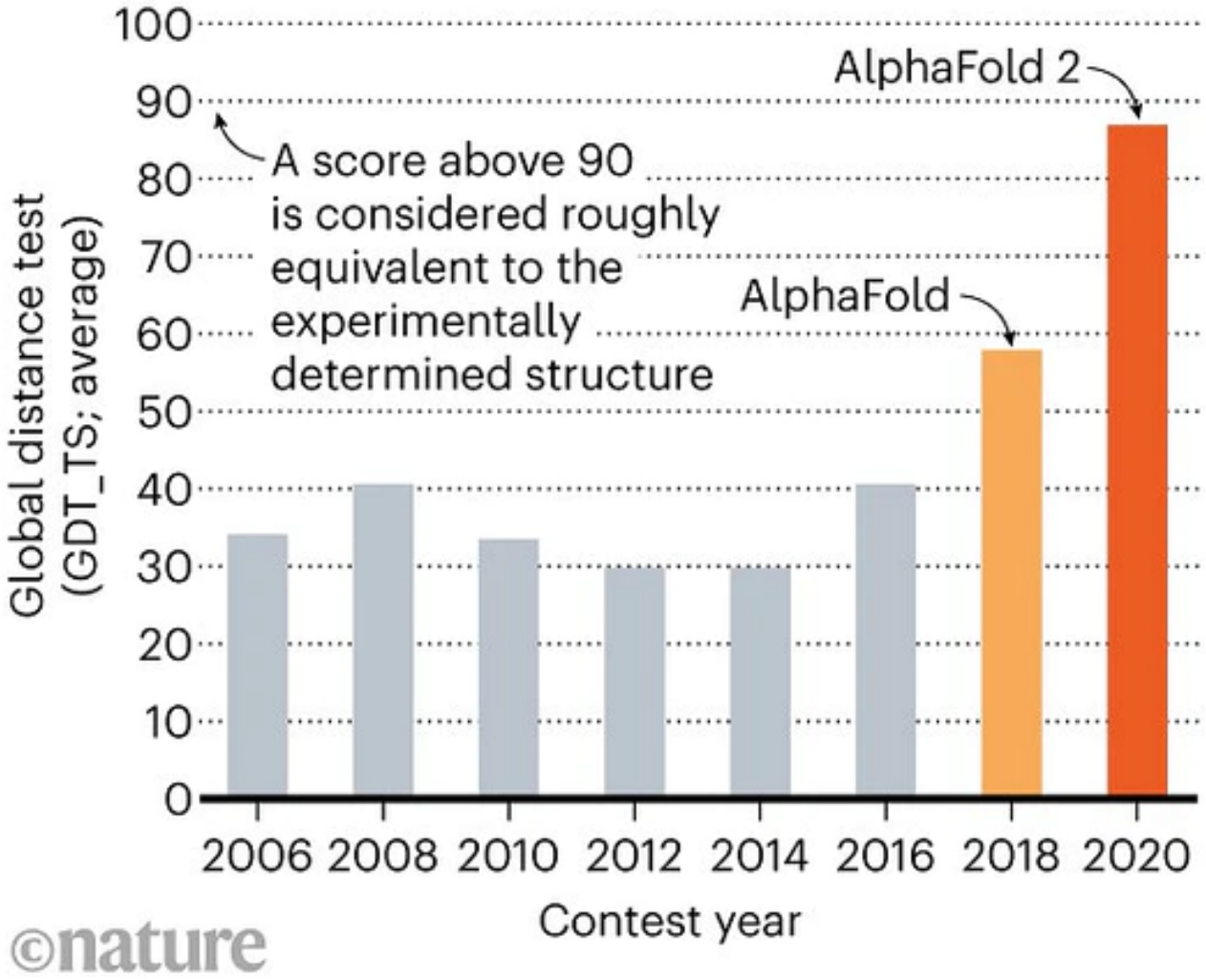
- High Accuracy Modeling
- Topology Prediction
- Contact Prediction
- Model Refinement
- Domain Assembly
- Estimation of Accuracy
- NMR X-link SAXS Data Assisted Modeling
- Biological Relevance

December 1-4, 2018 Iberostar Paraiso Maya, Riviera Maya, Mexico

Destination airport: Cancun (CUN)

Registration url: predictioncenter.org/casp13/meeting.cgi

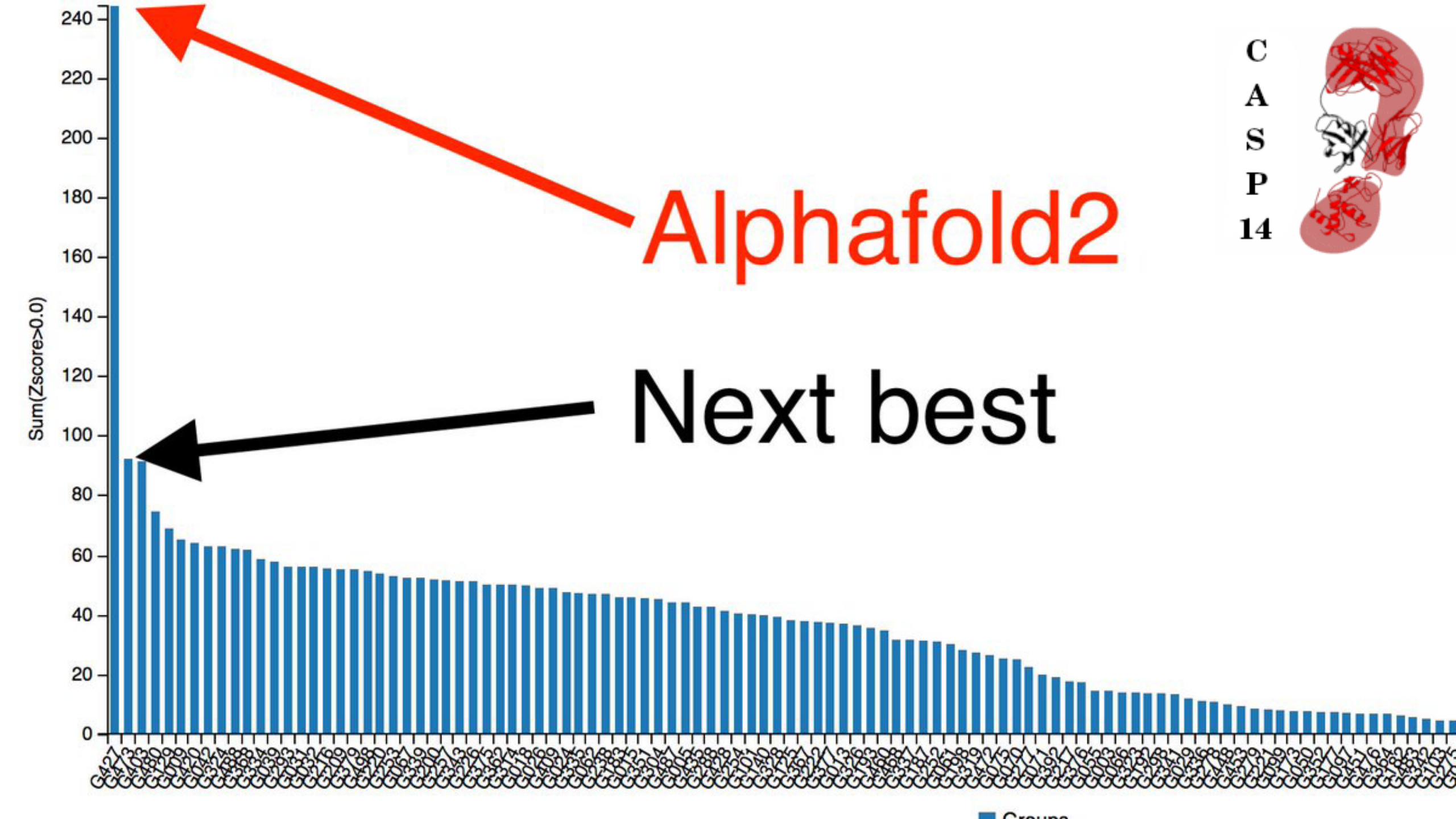
And the winner is… AlphaFold 2



DeepMind

AlphaFold 2 is the new version of an existing AI system (AlphaFold) created by the British company DeepMind, in collaboration with Google and EMBL

The first version of the system already won CASP13 competition, but was not yet reaching excellent evaluation criteria


Alphafold2
Next best
CASP14
Sum(Zscore>0.0)
240
220
200
180
160
140
120
100
80
60
40
20
0

Outlook

- How does AlphaFold 2 work?

  Architecture of the deep neural network

- Applications of AlphaFold 2

  Applying structure predictions to complement experimental data

- Limitations of AlphaFold 2

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

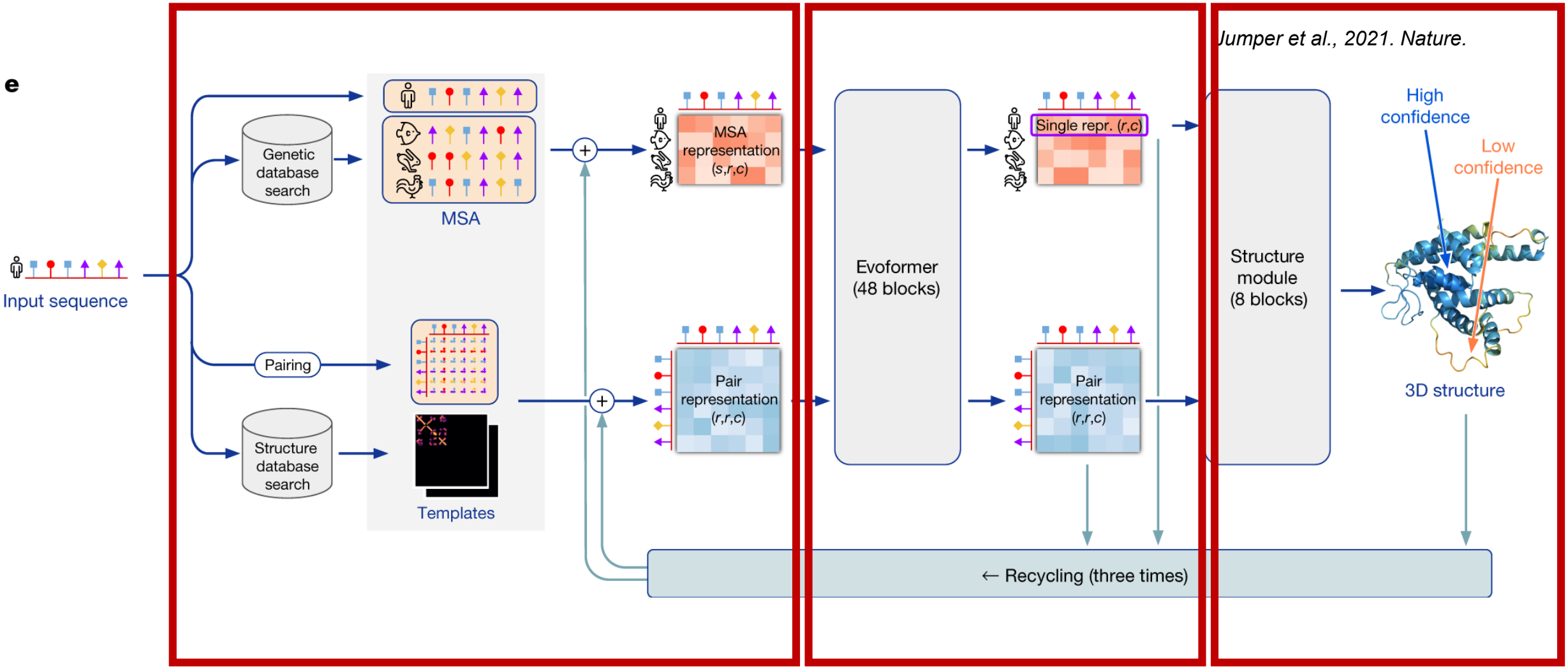
Published online: 15 July 2021

Open access

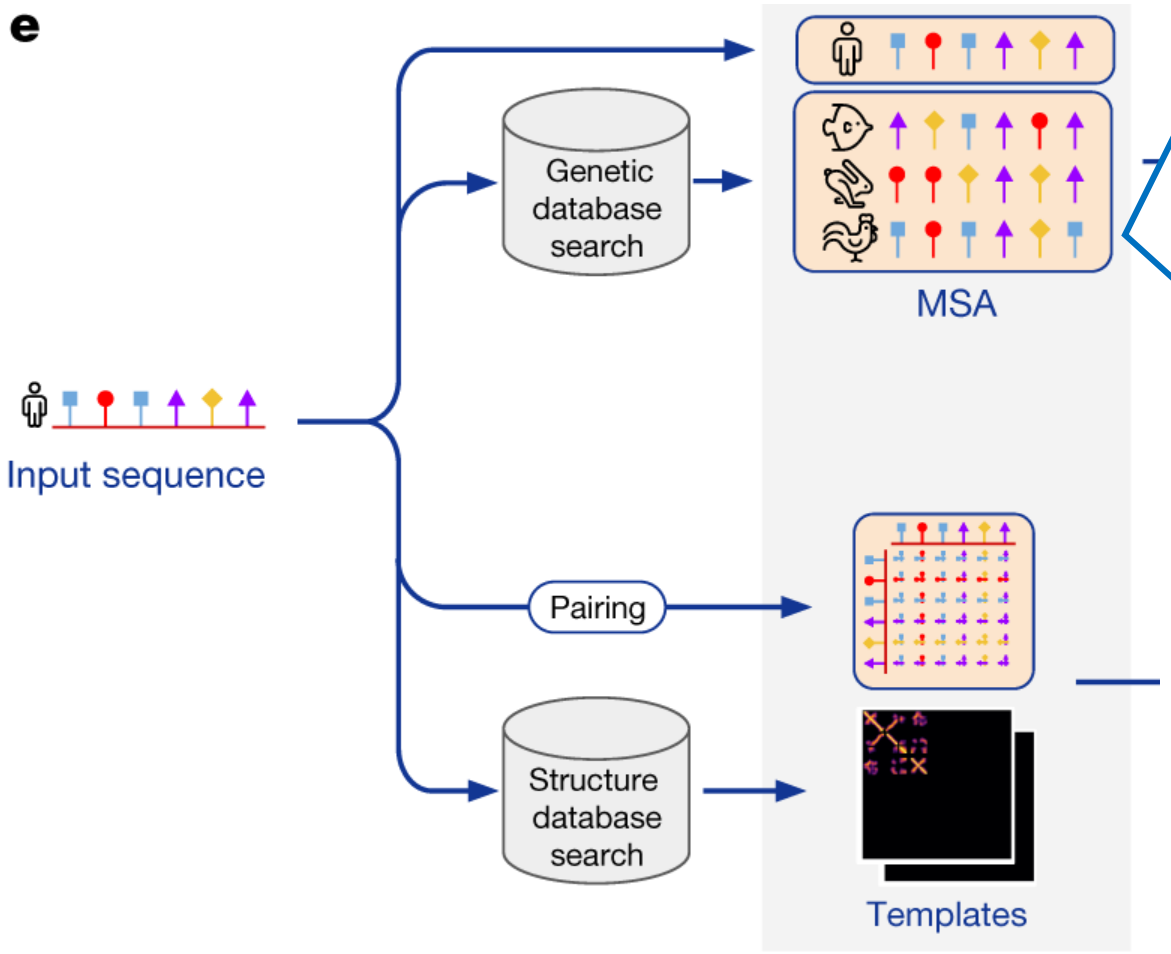
Check for updates

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]✉

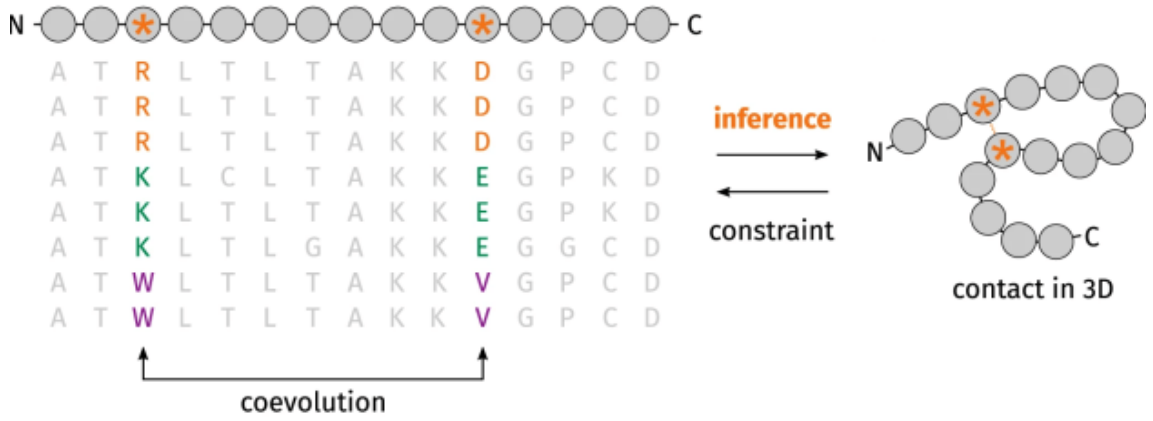
Structure of AlphaFold 2



Pre-processing module

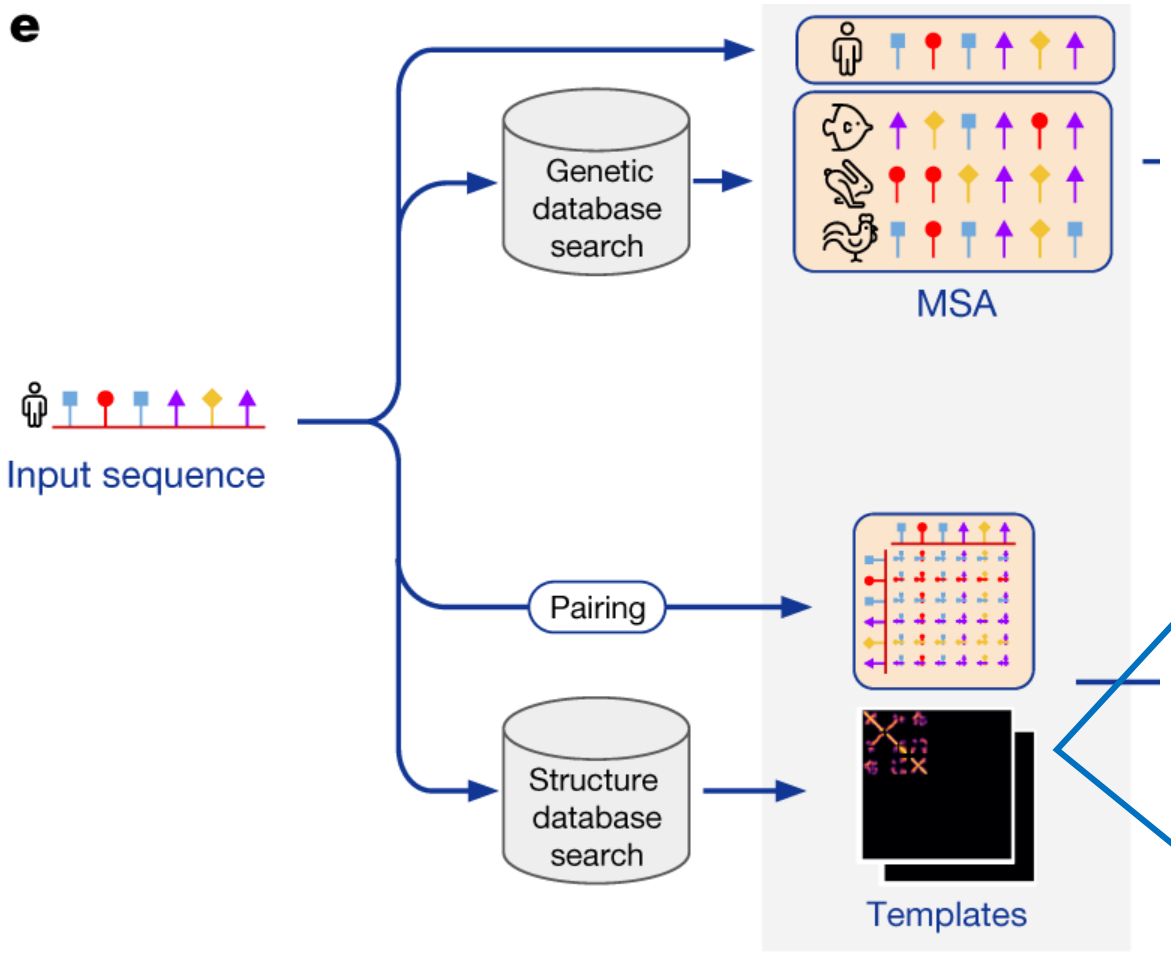


Multiple Sequence Alignment uses the target sequence to extract similar sequences from databases like UniRef.

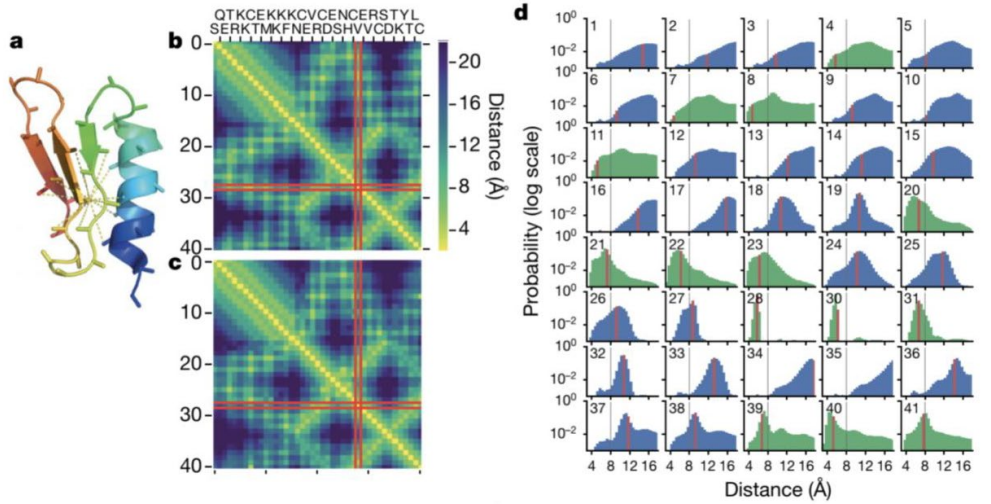


The idea is that if two amino acids are close in the structure, mutations in one amino acid will be closely followed by mutations in the other amino acids, in order to preserve the structure

Pre-processing module



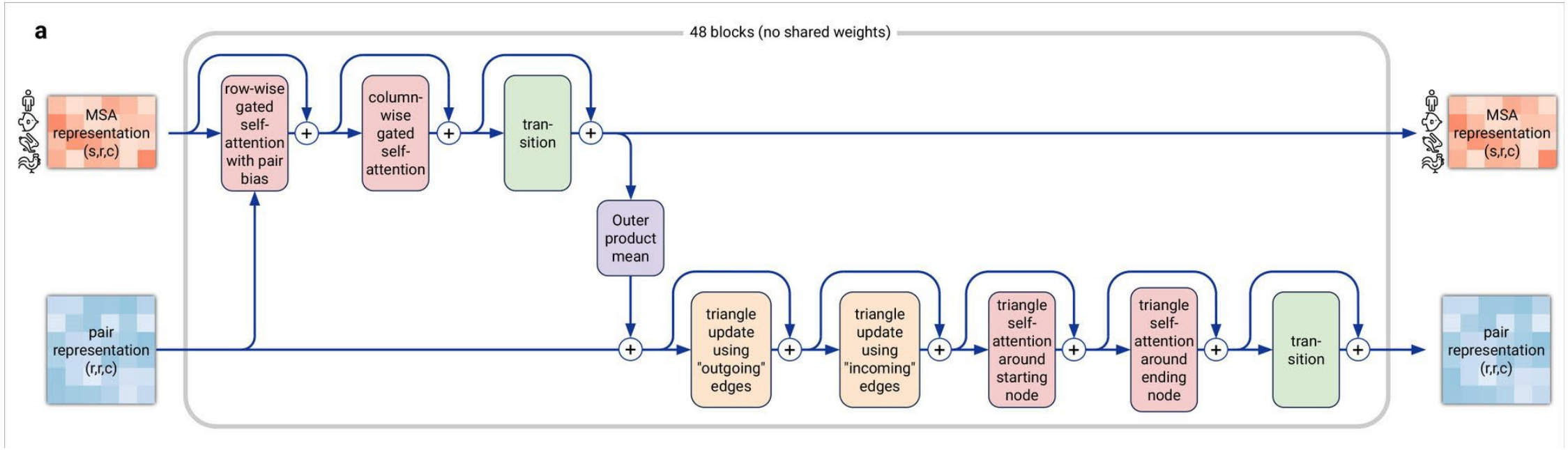
AlphaFold2 also looks for templates that belong to the same family of proteins and might have a similar structure, despite differences in the primary sequences



Then it creates a distogram, which is a representation of the distances between each pair of amino acids.

AF2 reports a distribution over 64 distances, giving an idea of how accurate the model will be

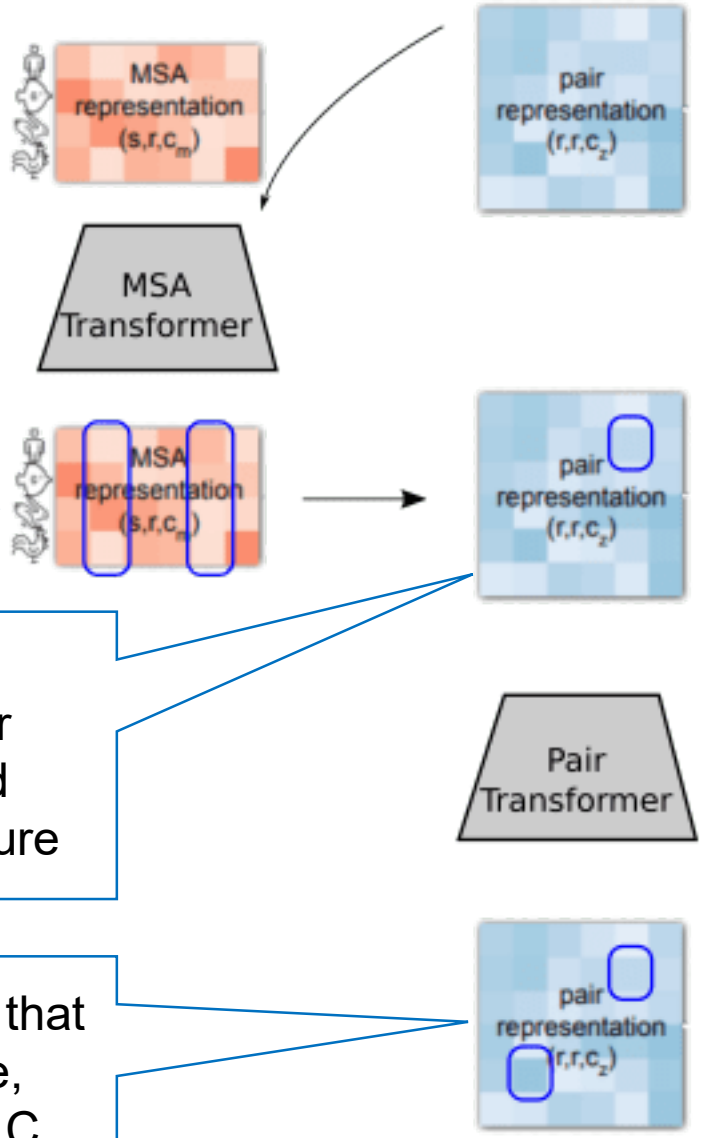
EVOFORMER: The EVOlutionary transFORMER



The Evoformer is composed of 48 non-iterative modules that continuously exchange information between the MSA representation and pair representation.

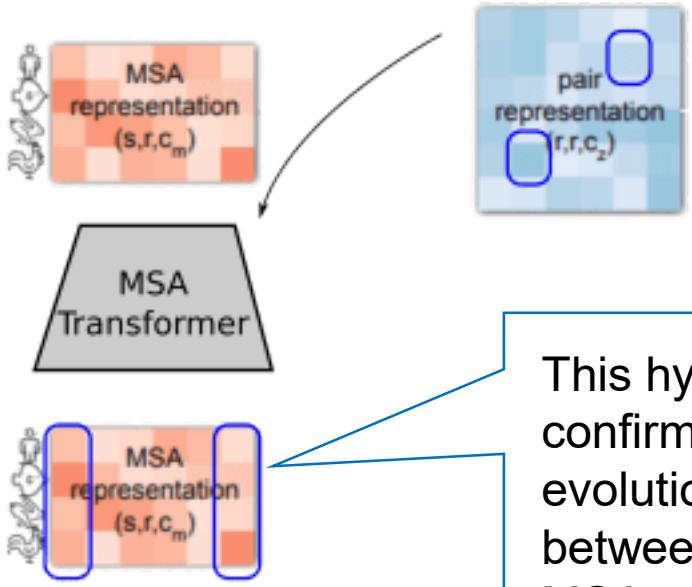
The information extracted from the MSA is used to adjust the pair representation, which in turn improves the information extracted from the MSA. This process is repeated until the system has a solid inference on the target

The MSA finds a correlation between amino acids A and B



This information is included in the pair representation and updates the structure

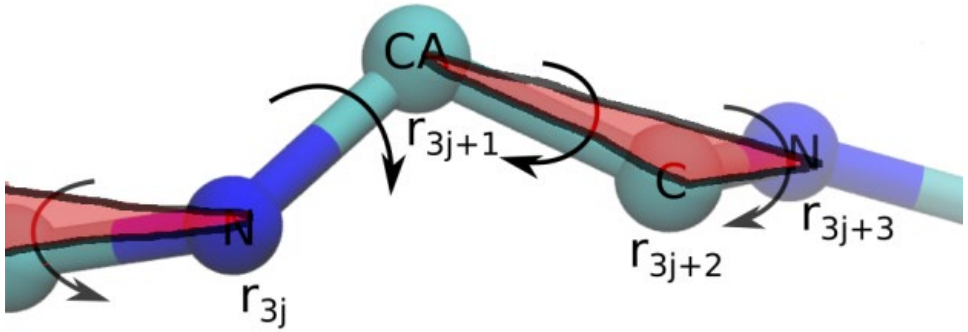
The update shows that if A and B are close, then probably also C and D are related



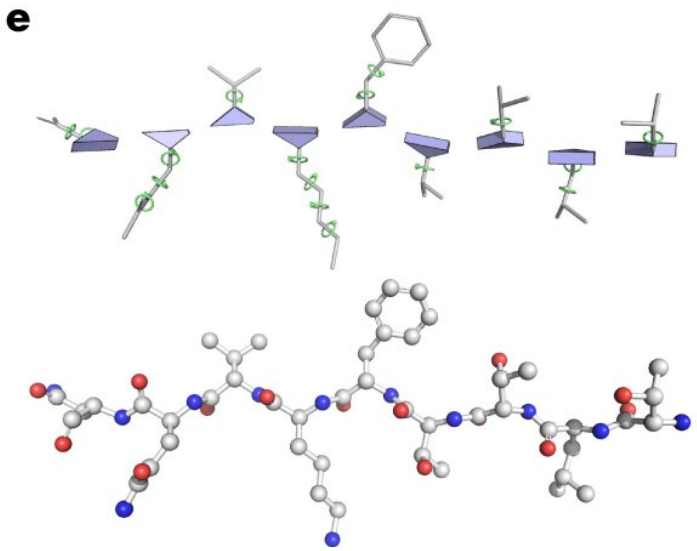
This hypothesis can be confirmed by looking for evolutionary correlations between C and D in the MSA

Repeat for 48 modules!

Structure module



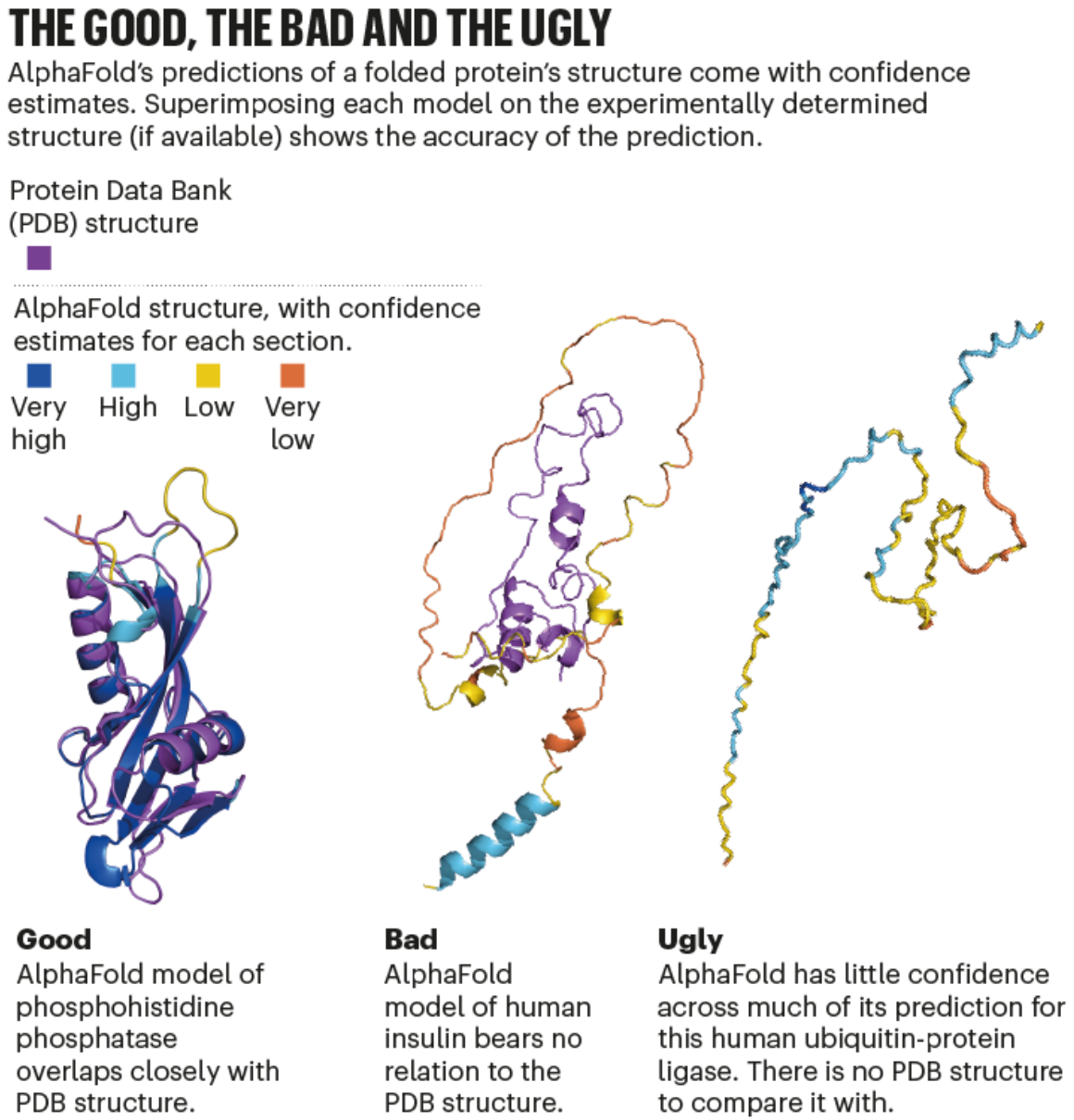
The protein is considered as a "residue gas" and each amino acid is represented as a triangle.



At the beginning all the residues are placed in the same place and then they are moved freely in space without considering physical constraints

The structure module focuses more on obtaining a correct local geometry for each residue rather than having a globally correct folding.

Final predictions



Thornton et al., 2021

The overall quality of the final output is strongly dependent on the abundance of related PDB structures on which the AI has been trained

Can I run AlphaFold 2?

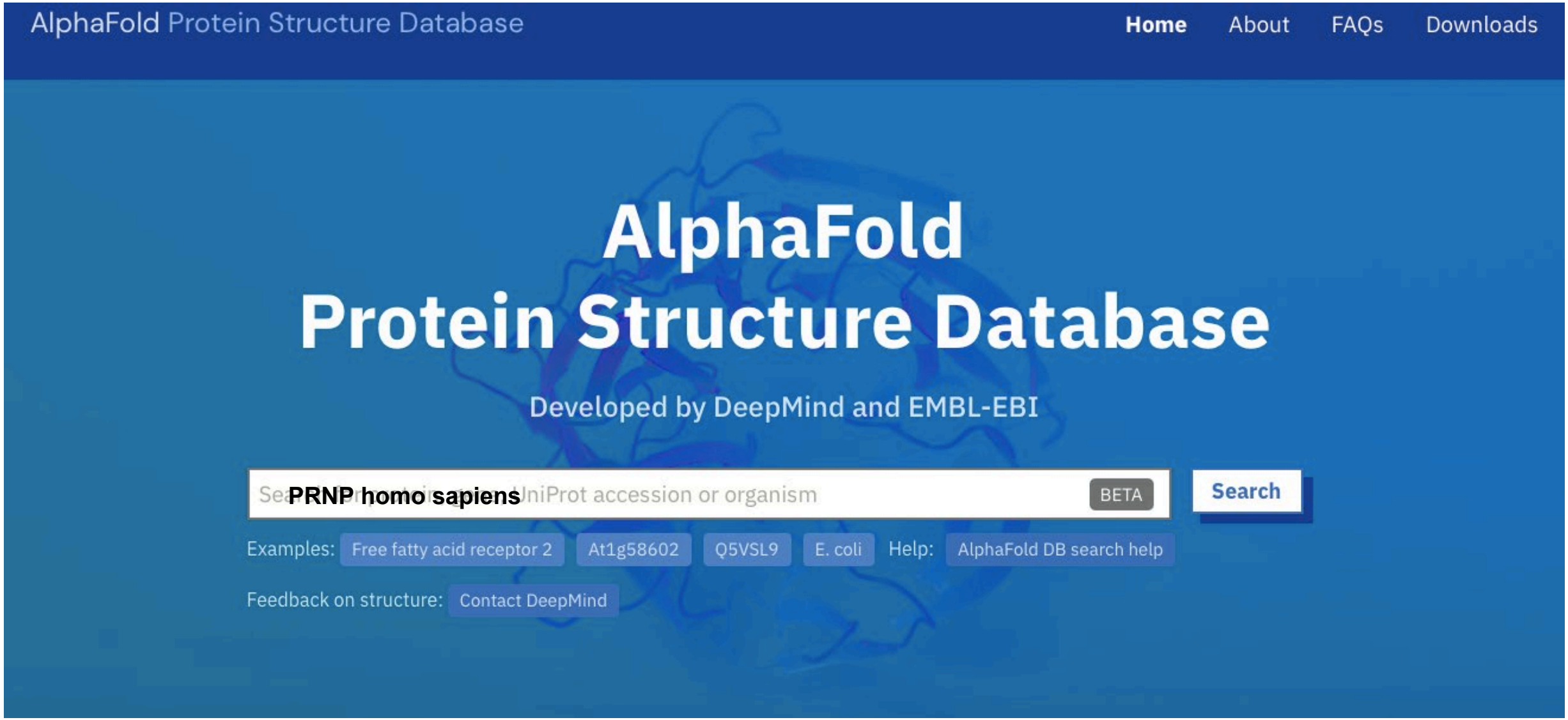
Running AlphaFold

The simplest way to run AlphaFold is using the provided Docker script. This was tested on Google Cloud with a machine using the `nvidia-gpu-cloud-image` with 12 vCPUs, 85 GB of RAM, a 100 GB boot disk, the databases on an additional 3 TB disk, and an A100 GPU.

Not on a standard laptop!

Free notebook environment that runs entirely on the cloud

Everybody can run Python codes through the browser


AlphaFold Protein Structure Database
Home
About
FAQs
Downloads
AlphaFold
Protein Structure Database
Developed by DeepMind and EMBL-EBI
PRNP homo sapiens
UniProt accession or organism
BETA
Search
Examples:
Free fatty acid receptor 2
At1g58602
Q5VSL9
E. coli
Help:
AlphaFold DB search help
Feedback on structure:
Contact DeepMind

3D viewer

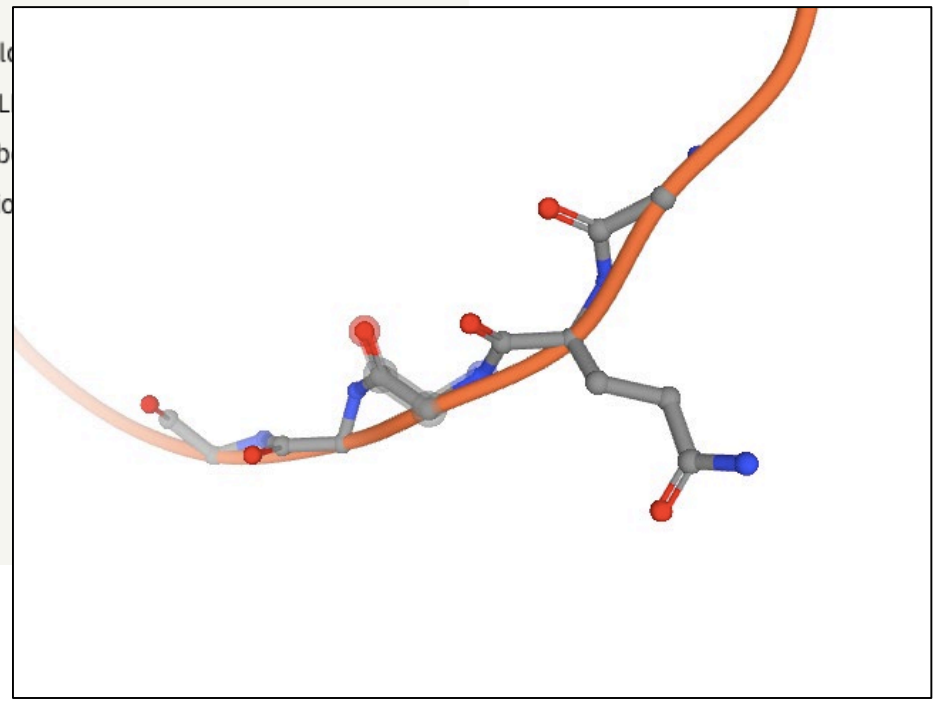
Model Confidence:

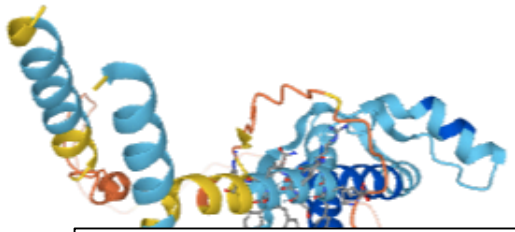
- Very high (pLDDT > 90)
- Confident (90 > pLDDT > 70)
- Low (70 > pLDDT > 50)
- Very low (pLDDT < 50)

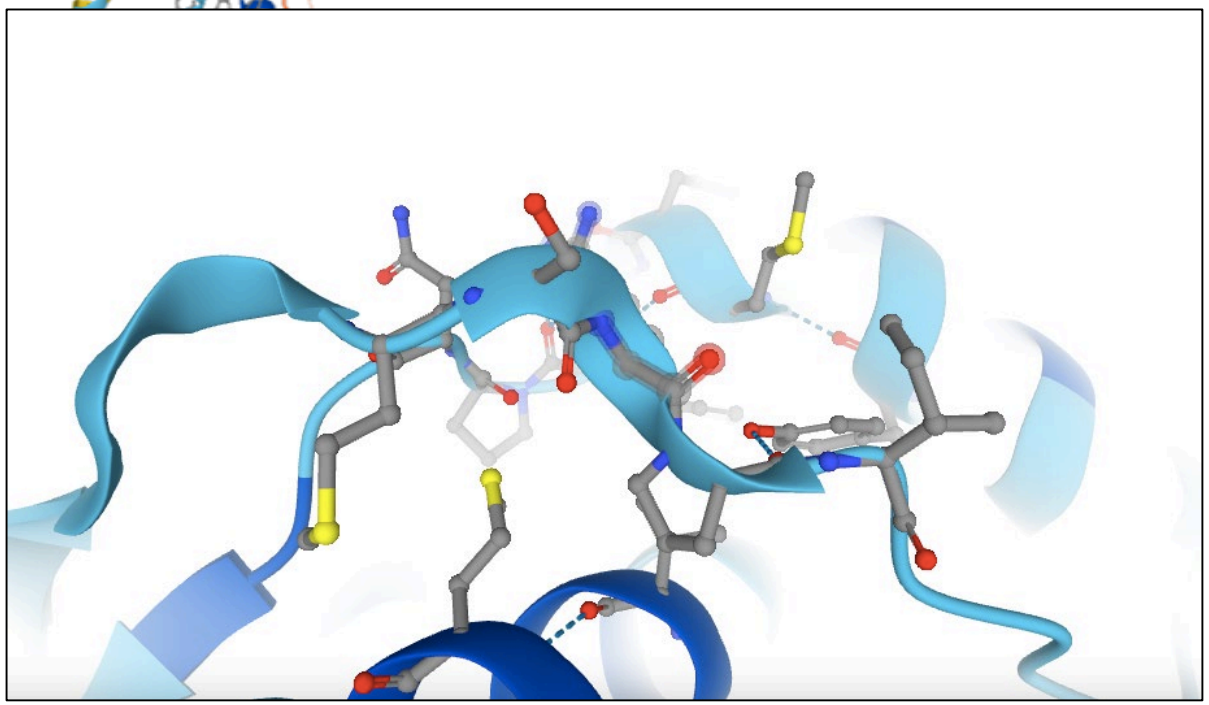
AlphaFol
score (pL
regions b
in isolatio

Sequence of AF-P04156-F1 1: Major prion pı A

```
1          11         21         31         41         51         61         71         81         91         101        111        121
MANLGCWMLV LFVATWSDLG LCKKRPKPGG WNTGGSRYPG QGSPGGNRYP PQGGGGWGQP HGGGWGQPHG GGWGQPHGGG WGQPHGGGWG QGGGTHSQWN KPSKPKTNMK HMAGAAAAGA V
           131        141        151        161        171        181        191        201        211        221        231        241
VGGLGGYML GSAMSRPIIH FGSDYEDRYY RENMHRYPNQ VYYRPMDEYS NQNNFVHDCV NITIKQHTVT TTTKGENFTE TDVKMMERVV EQMCITQYER ESQAYYQRGS SMVLFSSPPV IL
           251
LISFLIFLIVG
```

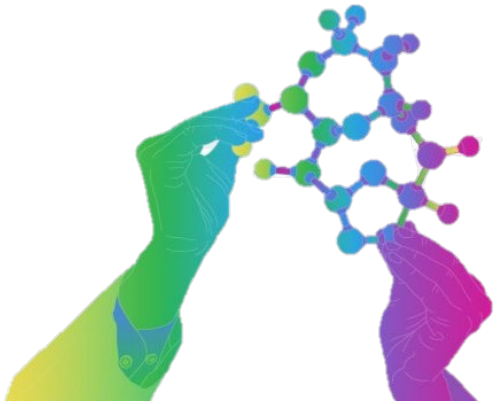




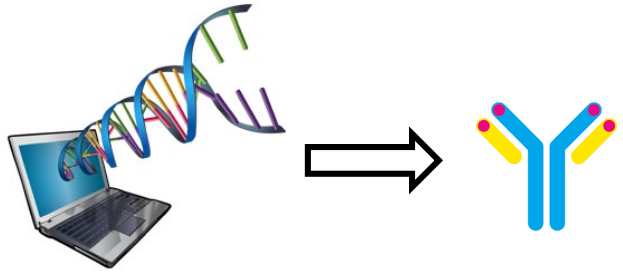


Applications of AlphaFold 2

Structure-based drug discovery



De novo protein design and protein engineering



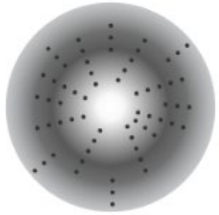
Treatments and drugs

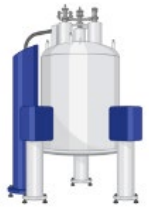
Supplements

Pesticides

Biomaterials

Accelerate structural studies





Complement X-ray and cryoEM data

BRIEF COMMUNICATION
https://doi.org/10.1038/s41594-022-00729-3
nature structural & molecular biology

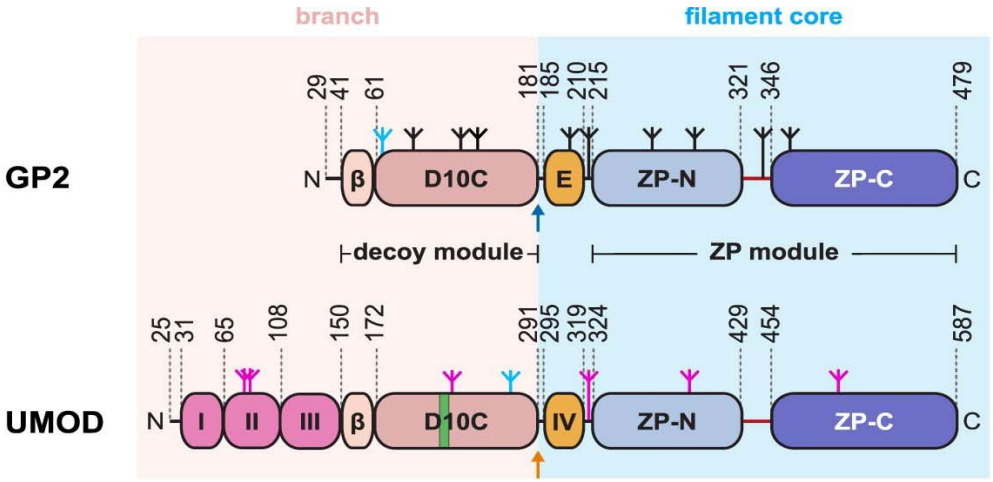
Check for updates

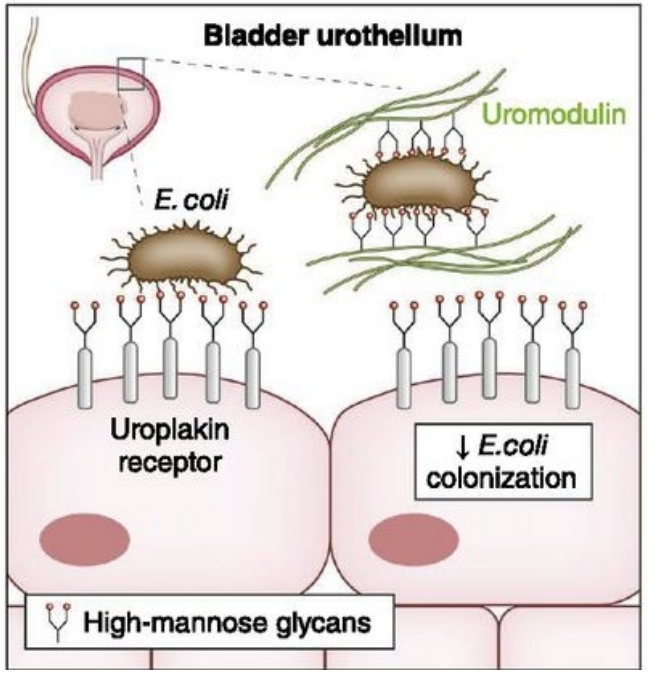
OPEN

Structure of the decoy module of human glycoprotein 2 and uromodulin and its interaction with bacterial adhesin FimH

Alena Stsiapanava[1], Chenrui Xu[2,3], Shunsuke Nishio[1], Ling Han[1], Nao Yamakawa[4], Marta Carroni[5], Kathryn Tunyasuvunakool[6], John Jumper[6], Daniele de Sanctis[7], Bin Wu[2,3] and Luca Jovine[1,2] ✉

Glycoprotein 2 (GP2) and uromodulin protect gastrointestinal and urinary tracts from infection by acting as decoys for the bacterial fimbrial lectin FimH



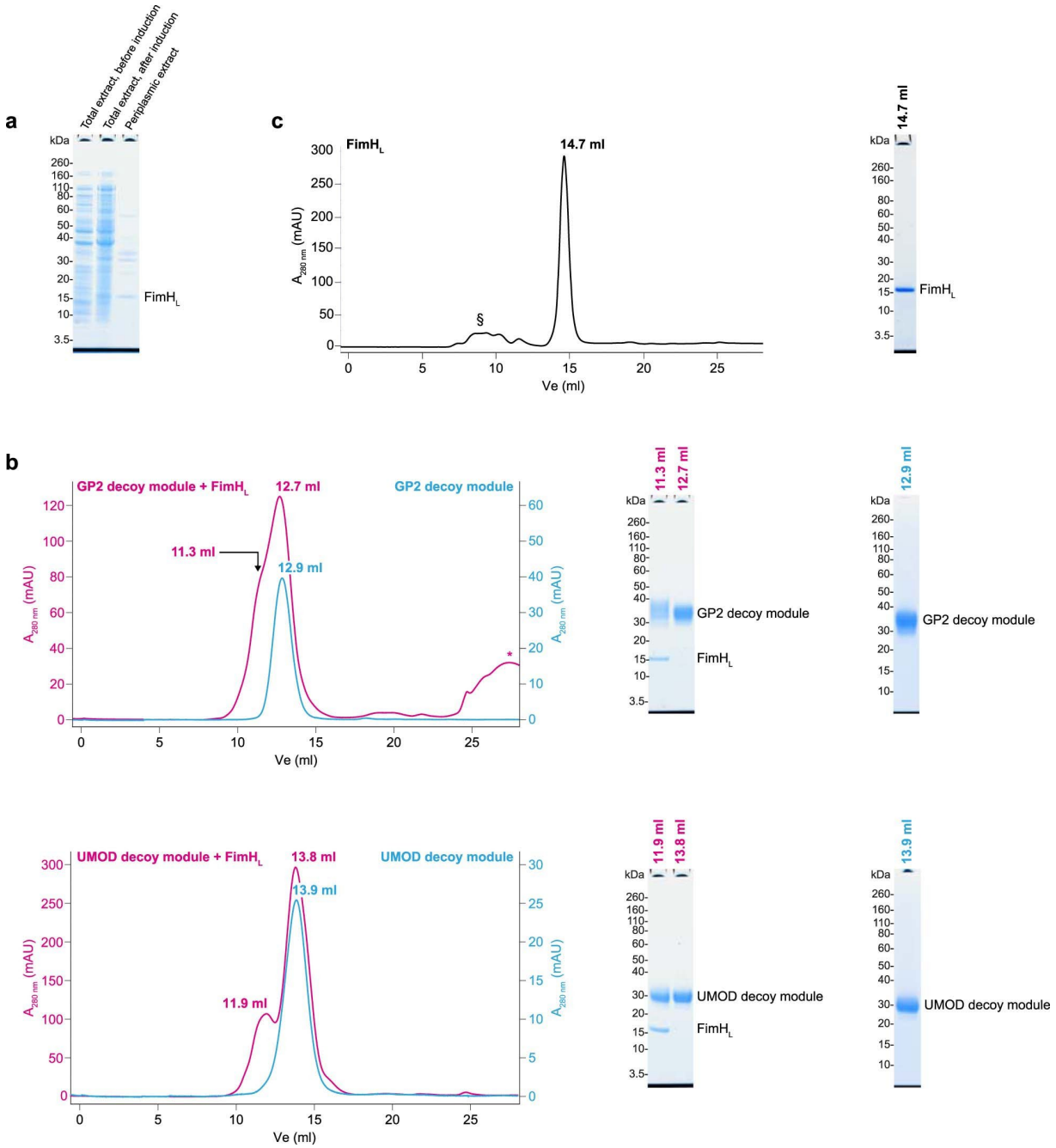


Kipp et al., 2021

The N-terminal region interacts with FimH, but there are no structural information on the binding site

The C-terminal region of uromodulin mediates its polimerization after being secreted from the cells of the urinary epithelium

- What is the structure of the decoy module of GP2 and UMOD?
- Which region of the N-terminal domain is recognized by FimH, and what is the architecture of the sub-domain?

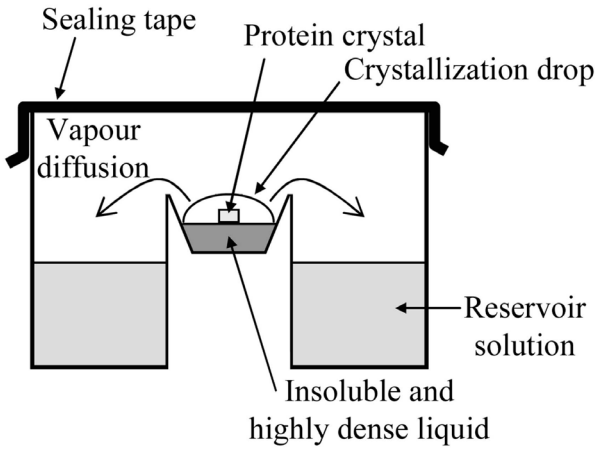


The untagged lectin domain of FimH was expressed in E.coli

The bacterial crude extract was incubated with the purified branch domain of GP2 and UMOD

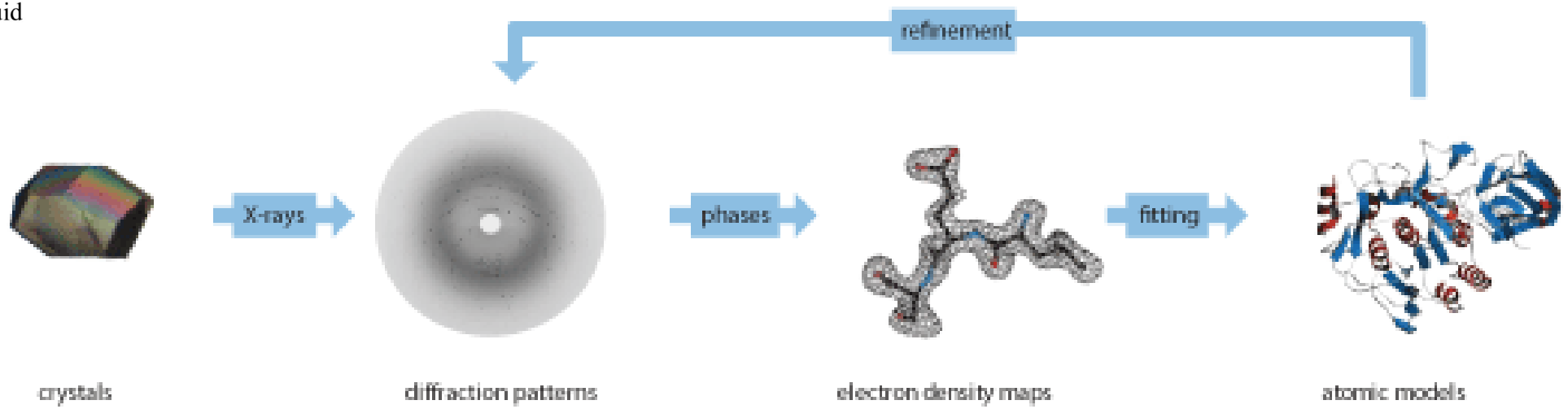
SEC chromatography shows co-elution of each decoy molecule with FimH

FimH is detected also in SDS-PAGE gels of the SEC fractions



Adachi et al., 2003

The his-tagged branched portion of GP2 was subjected to cristallization trials using the «sitting drop vapour diffusion» method and analysed at the synchrotron



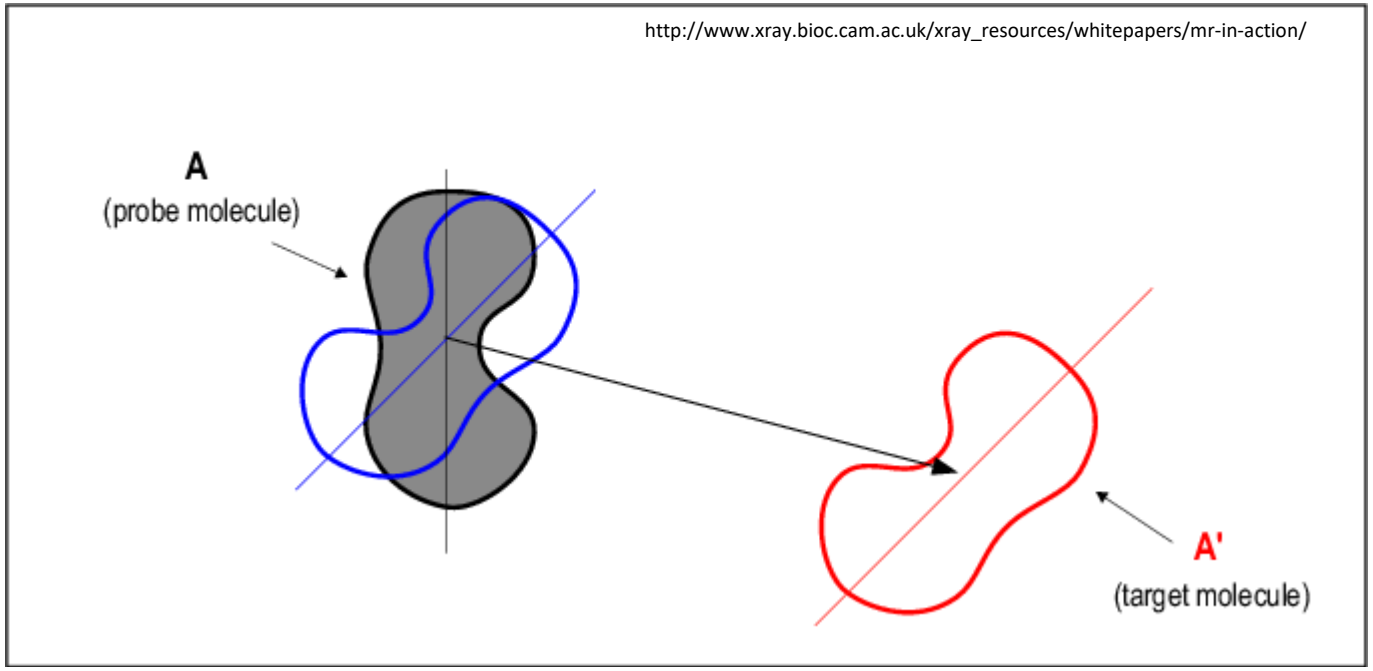
However, the crystal displayed high diffraction disorder and low symmetry.

A correct interpretation of diffracted data is not possible.

Molecular replacement using AlphaFold 2 models

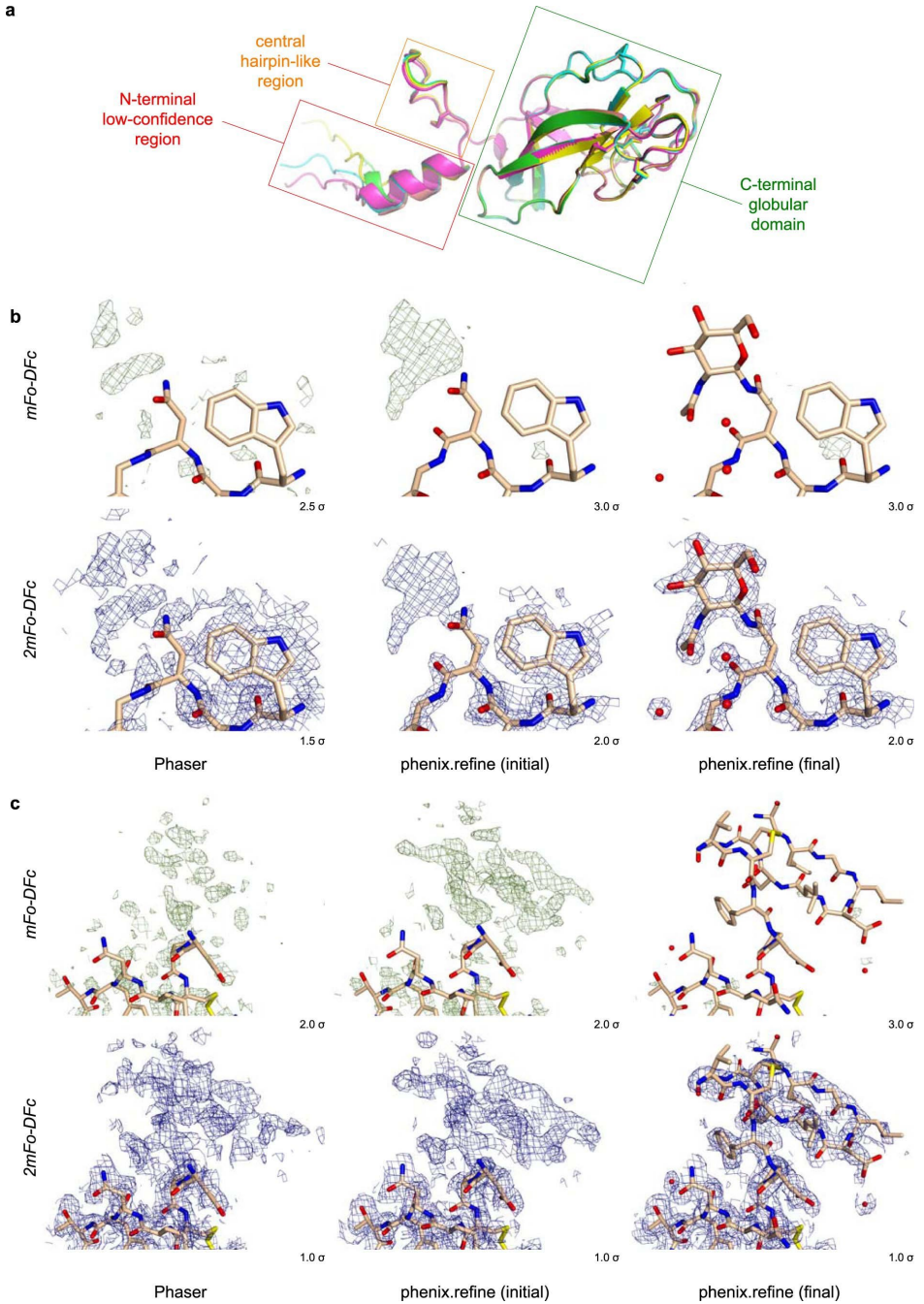
Molecular replacement is a method used in X-rays chrystallography to try to solve the phase problem and obtain the electron density map from the diffraction pattern

Basically, you look for homologous proteins whose structure is known and you use their structure as «template» to build the model of your target protein. Template models are rotated and translated until they fit the experimental data



Which models can we use?

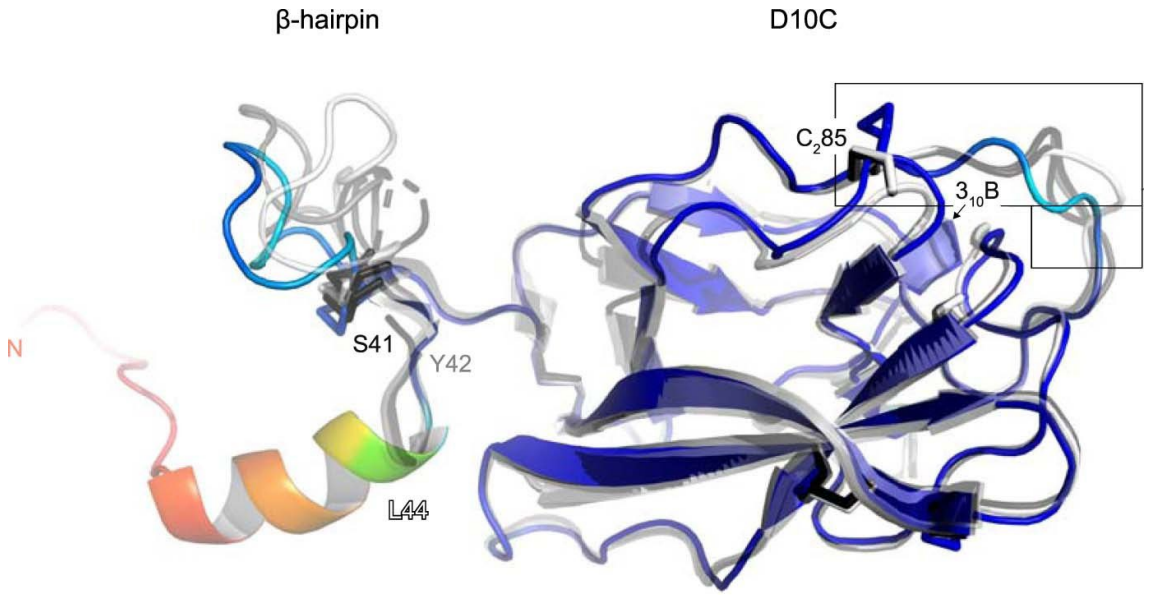
- Structures of known homologous proteins (at least 40% homology)
- Low-quality NMR structures of the same protein
- Predicted models

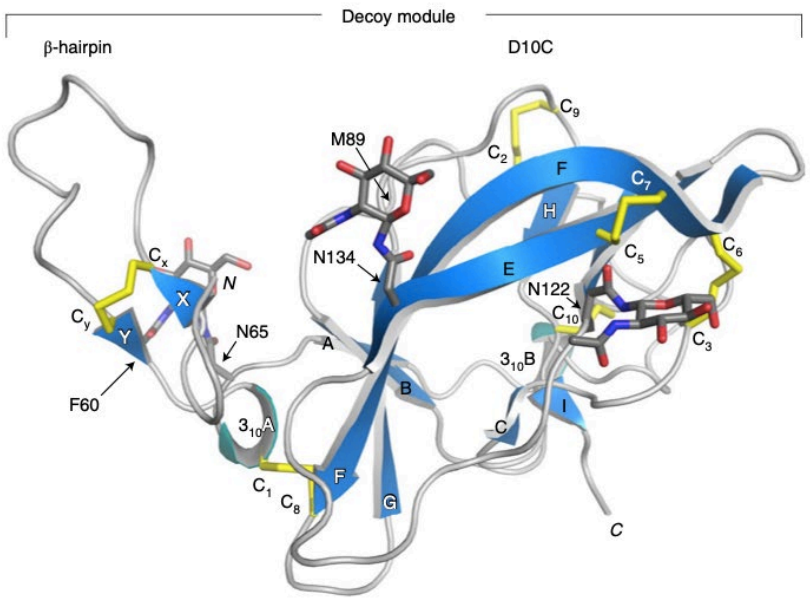


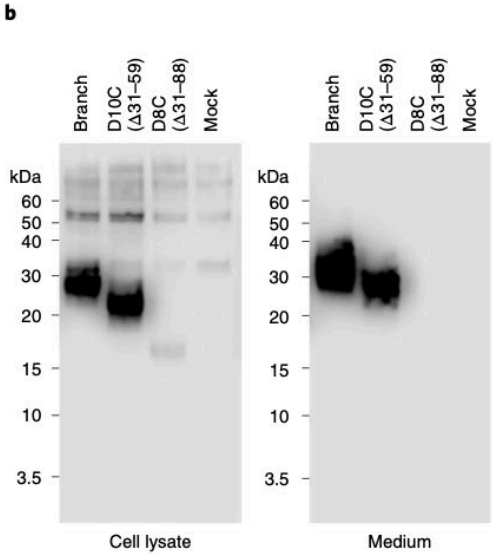
AlphaFold 2 predicted the structure of the decoy portion of GP2

The models were then used to phase the crystals of GP2 previously obtained, and produced an atomic model

The final atomic module of GP2 (in grey) shows very high similarity to its predicted model (in colours)

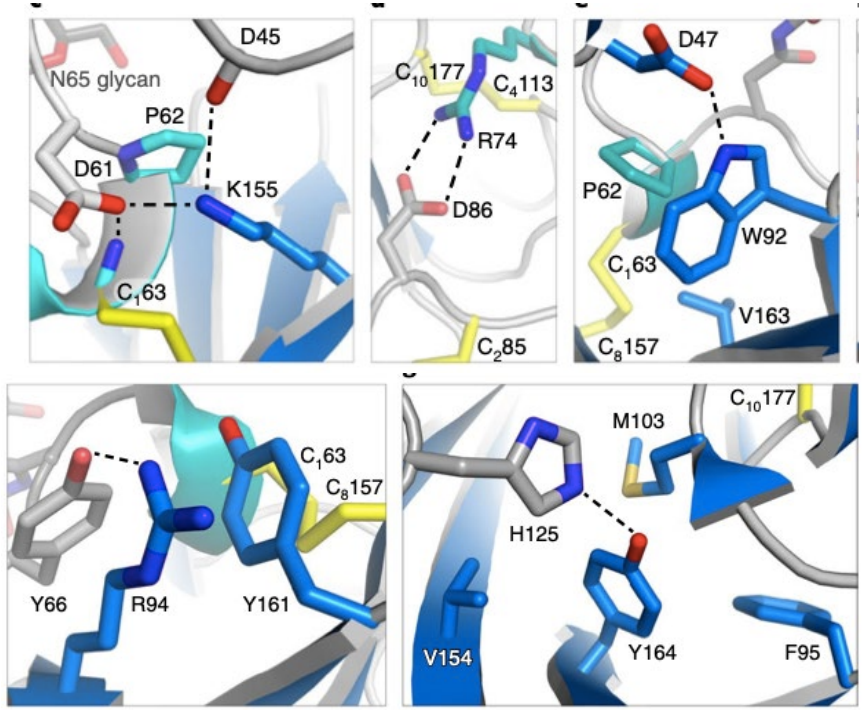




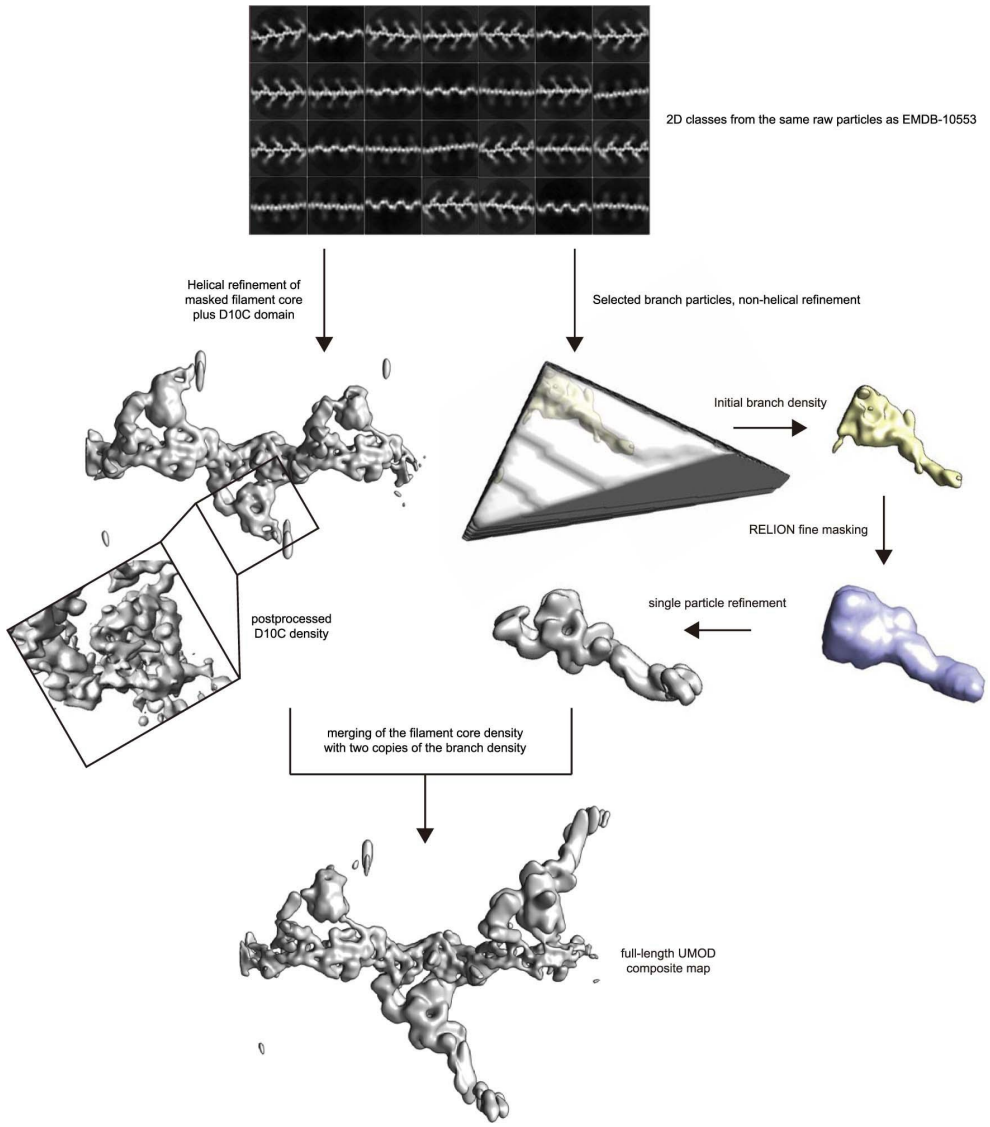


Structural data revealed that the interacting domain had 10 cysteine residues and not 8, as it was suggested previously.

Given the high sequence homology (>60%) between GP2 and UMOD, the branch region structure of GP2 was used to model the effect of pathogenic mutations of UMOD affecting invariant amino acids

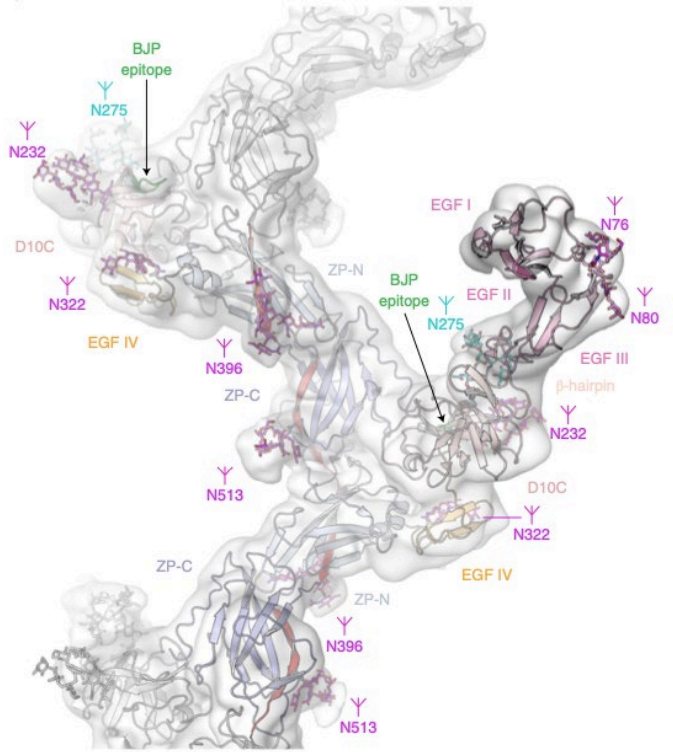


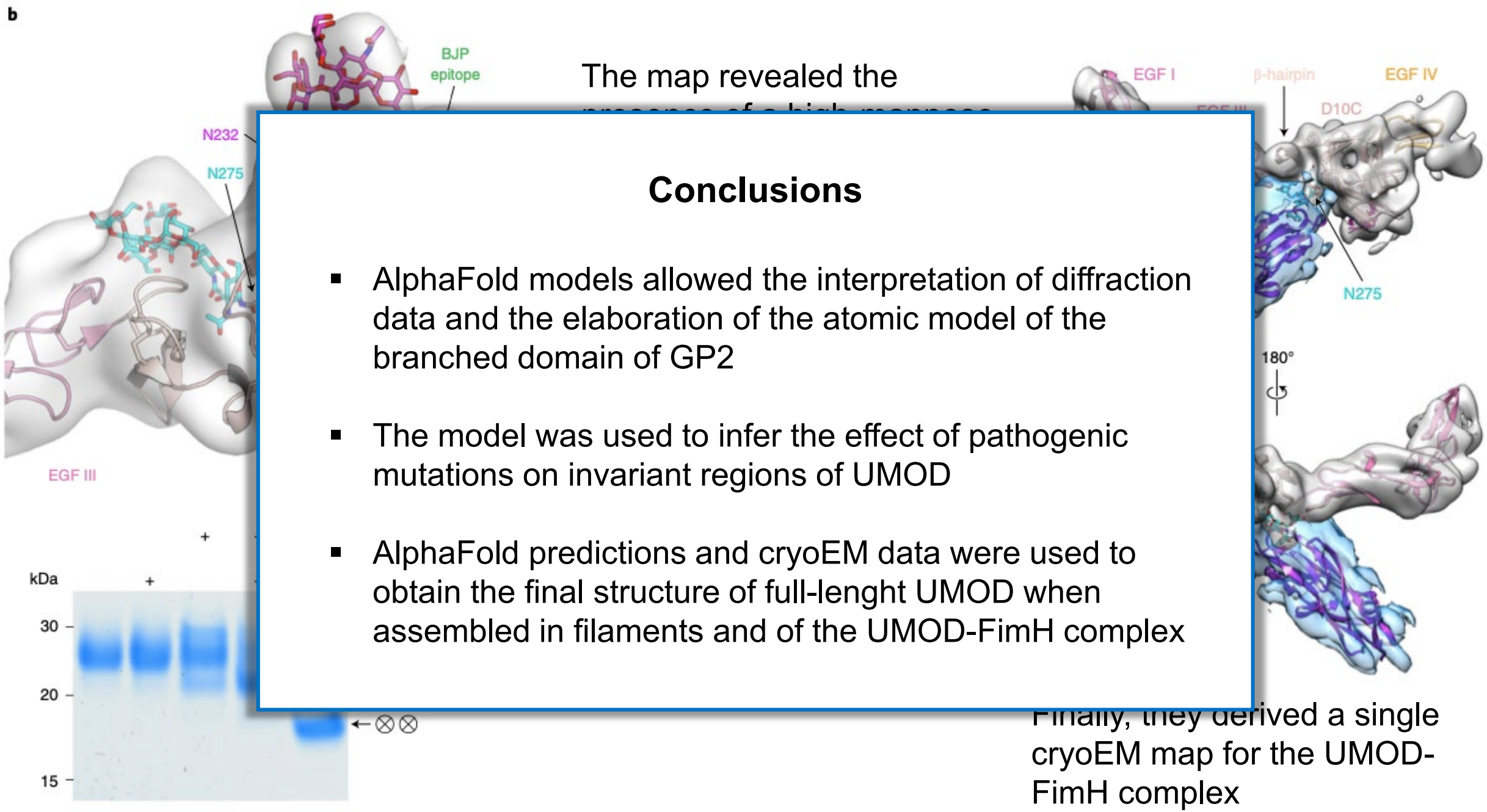
Next, they investigated the structure of the full-lenght UMOD protein arranged into fibers to identify the binding site of FimH



Cryo-EM analysis of UMOD fibers yielded a composite map of the full-lenght protein; however, only the filaments core could be interpreted with high confidence.

Previously obtained GP2 structure of the branch region was combined with AlphaFold predictions to generate a model that helped in the interpretation of the cryoEM map



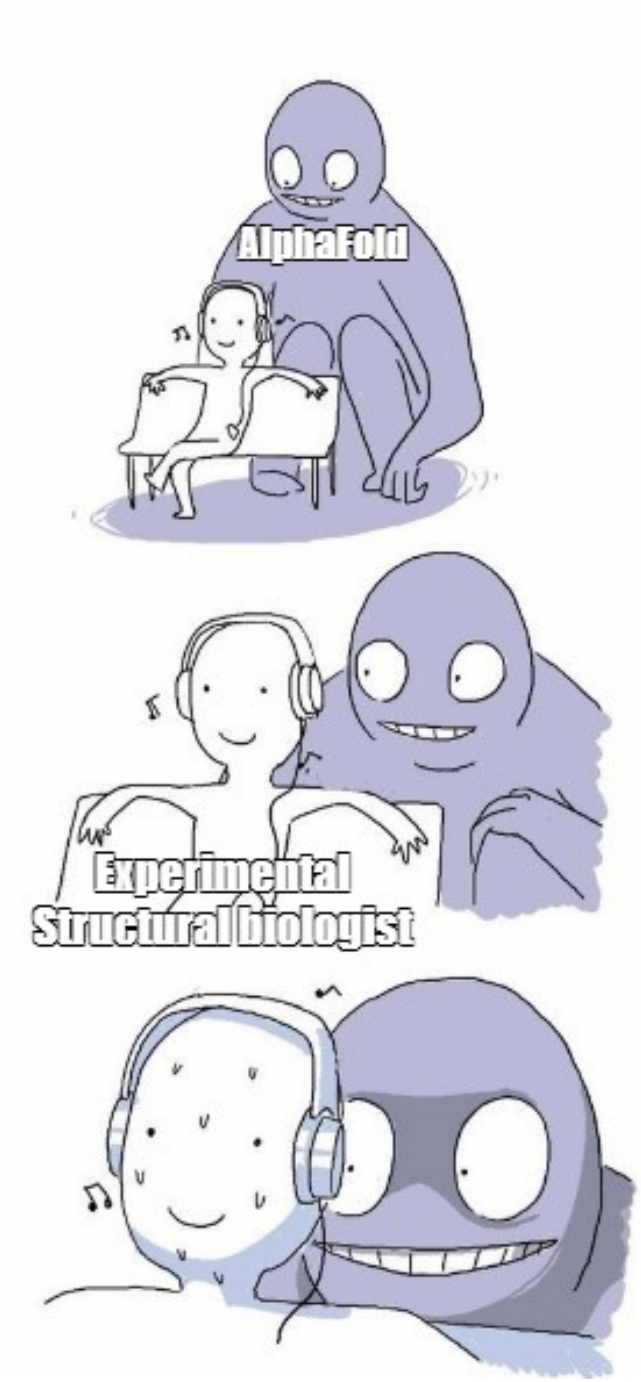


The map revealed the

Conclusions

- AlphaFold models allowed the interpretation of diffraction data and the elaboration of the atomic model of the branched domain of GP2
- The model was used to infer the effect of pathogenic mutations on invariant regions of UMOD
- AlphaFold predictions and cryoEM data were used to obtain the final structure of full-lenght UMOD when assembled in filaments and of the UMOD-FimH complex

Finally, they derived a single cryoEM map for the UMOD-FimH complex



Is this the end for experimental structural biology?

Limitations of AlphaFold 2

AlphaFold 2 principles

- Anfinsen's dogma: one sequence – one structure
- Principle of co-evolution
- AlphaFold 2 does not follow folding pathways

⇨ **No prediction of protein complexes**

DeepMind

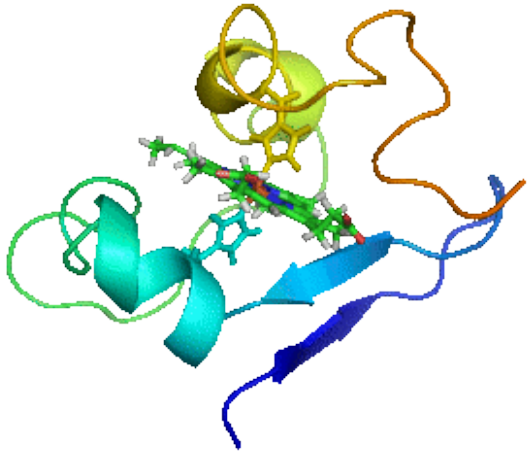
2021-10-04

Protein complex prediction with AlphaFold-Multimer

A new version of the code is being released now

Cofactors and prosthetic groups are still not included

No prediction of protein dynamics



Cytochrome C adopts 17 different conformations in solution (NMR data)

AlphaFold 2 only predicts one conformation

No prediction of the impact of missense mutations

AlphaFold2 can predict structural and phenotypic effects of single mutations

John M. McBride,[1,*] Konstantin Polev,[1,2] Vladimir Reinharz,[3] Bartosz A. Grzybowski,[1,4,†] and Tsvi Tlusty[1,4,‡]

[1]*Center for Soft and Living Matter, Institute for Basic Science, Ulsan 44919, South Korea*
[2]*Departments of Biomedical Engineering, Ulsan National Institute of Science and Technology, Ulsan 44919, South Korea*
[3]*Université du Québec à Montréal, Canada*
[4]*Departments of Physics and Chemistry, Ulsan National Institute of Science and Technology, Ulsan 44919, South Korea*

Check for updates

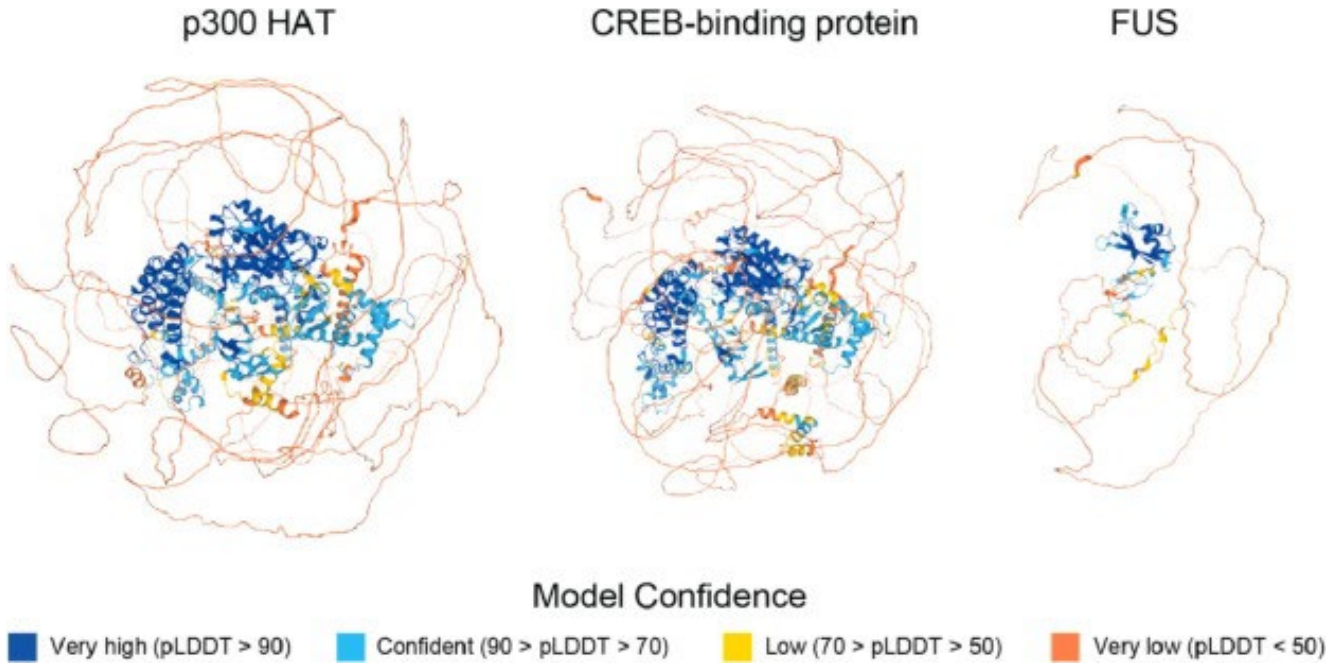
correspondence

Can AlphaFold2 predict the impact of missense mutations on structure?

green fluorescent protein[10]. This lack of correlation agrees with our case studies, illustrating the inability of AlphaFold2 to predict the effects of point mutations on protein structure. This limitation probably arises because it predicts structures based on those available in the PDB, rather than by fundamental driving forces of protein folding. It is possible that the merger of

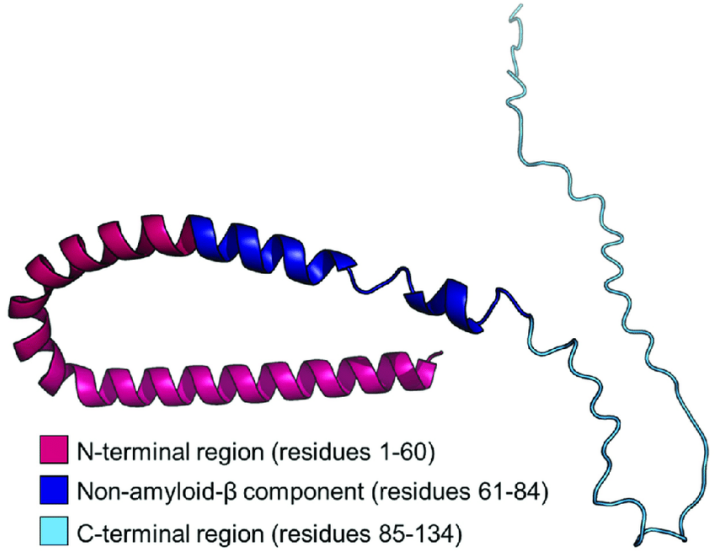
No prediction of intrinsically disordered regions

Intrinsically disordered regions (IDRs) are modeled as ribbon-like structures and are considered as low-confidence prediction regions



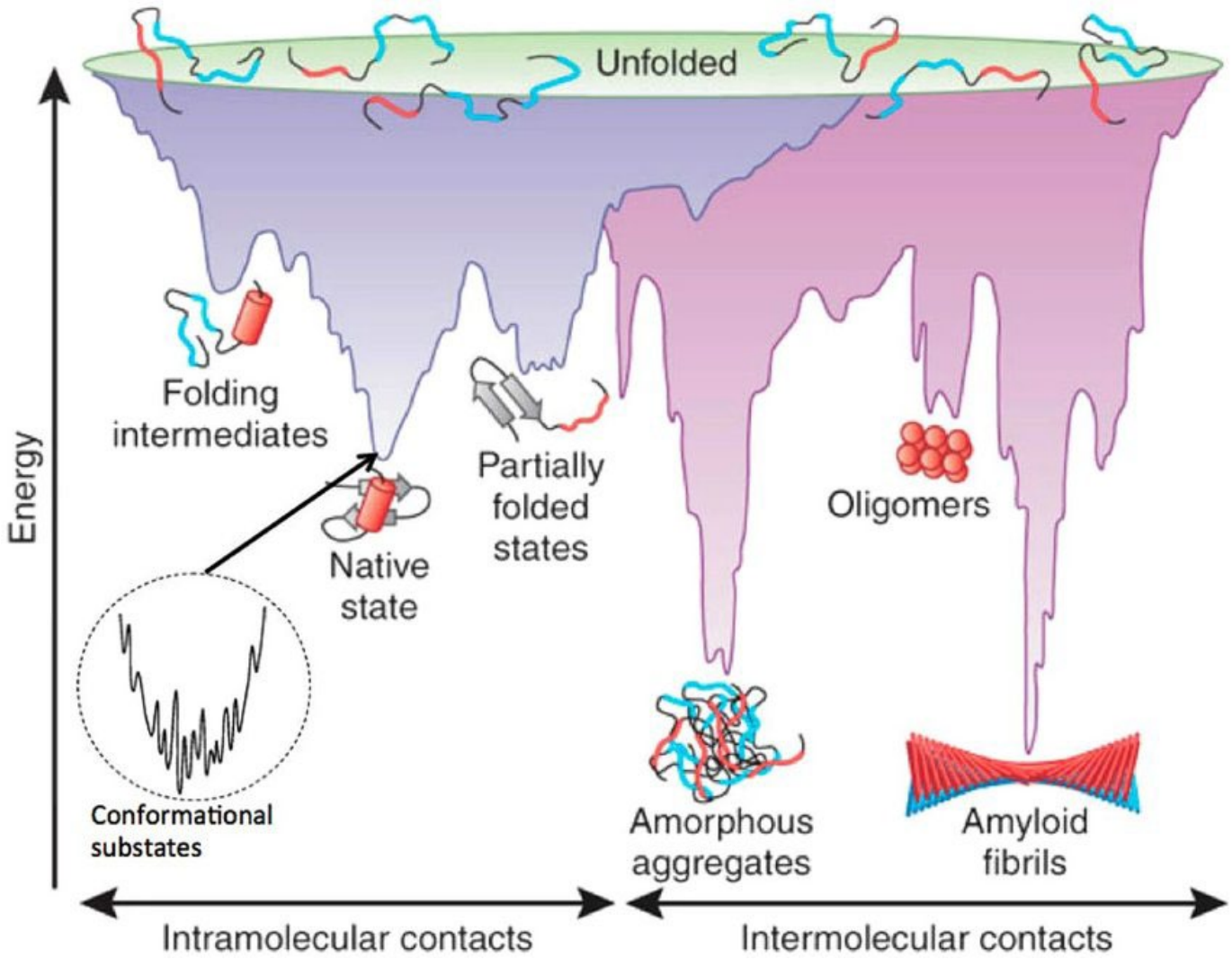
Ruff et al., 2021

However, the prediction makes no statement about the likelihood of different «transient» conformations



Hyashi et al., 2022

⇨ No prediction of amyloid fibrils



Raskatov et al., 2017

Aggregation-prone proteins have a multi-funneled energetic lansdcape

Oligomers and aggregation intermediates might not occupy energetic minima in the aggregation landscape

Aberrant aggregation is extrinsic to evolutionary selection, which is one of the working principles of AlphaFold 2

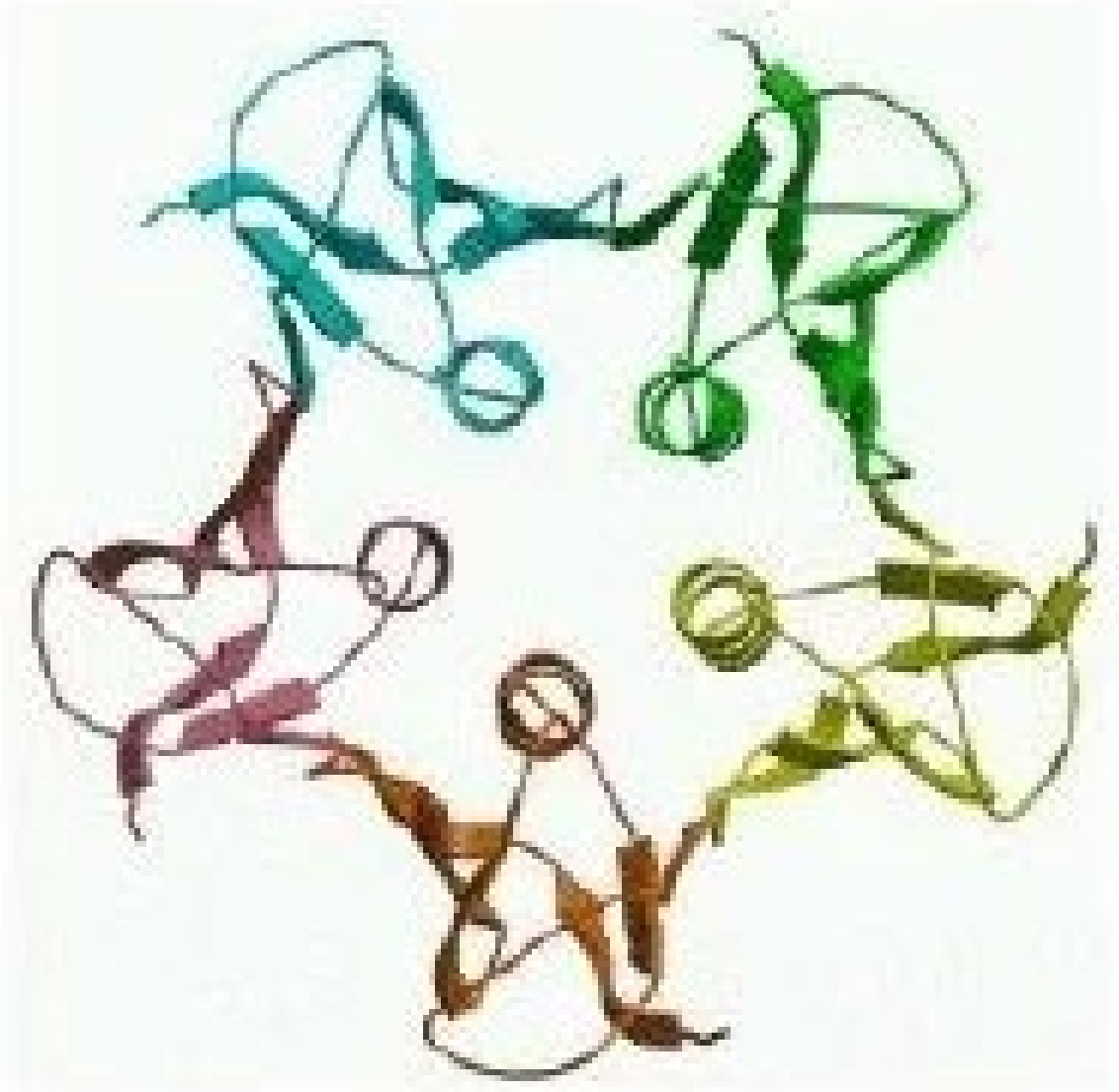
Conclusions

AlphaFold 2 generates high-confidence structural predictions for almost all globular proteins starting only from the amino acid sequence

«AlphaFold will provide new insights and understanding of fundamental processes related to health and disease, with applications in biotechnology, medicine, agriculture, food science and bioengineering. It will probably take one or two decades until the full impact of this development can be properly assessed»

EMBL

The code and the training database are publicly available, meaning that everybody can use it to predict the structure of a protein of interest



Thank You for Your Attention !