



# Google It!

AlphaFold and its impact on  
obtaining protein structures

Pablo Ricardo Arantes, PhD.

BIEN 249

# Slides

# Jupyter Notebook

<https://github.com/pablo-arantes/BIEN-249>



 SCAN ME

**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

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

**NEWS** · 30 NOVEMBER 2020

# 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures

Google's deep-learning program for determining the 3D shapes of proteins stands to transform biology, say scientists.

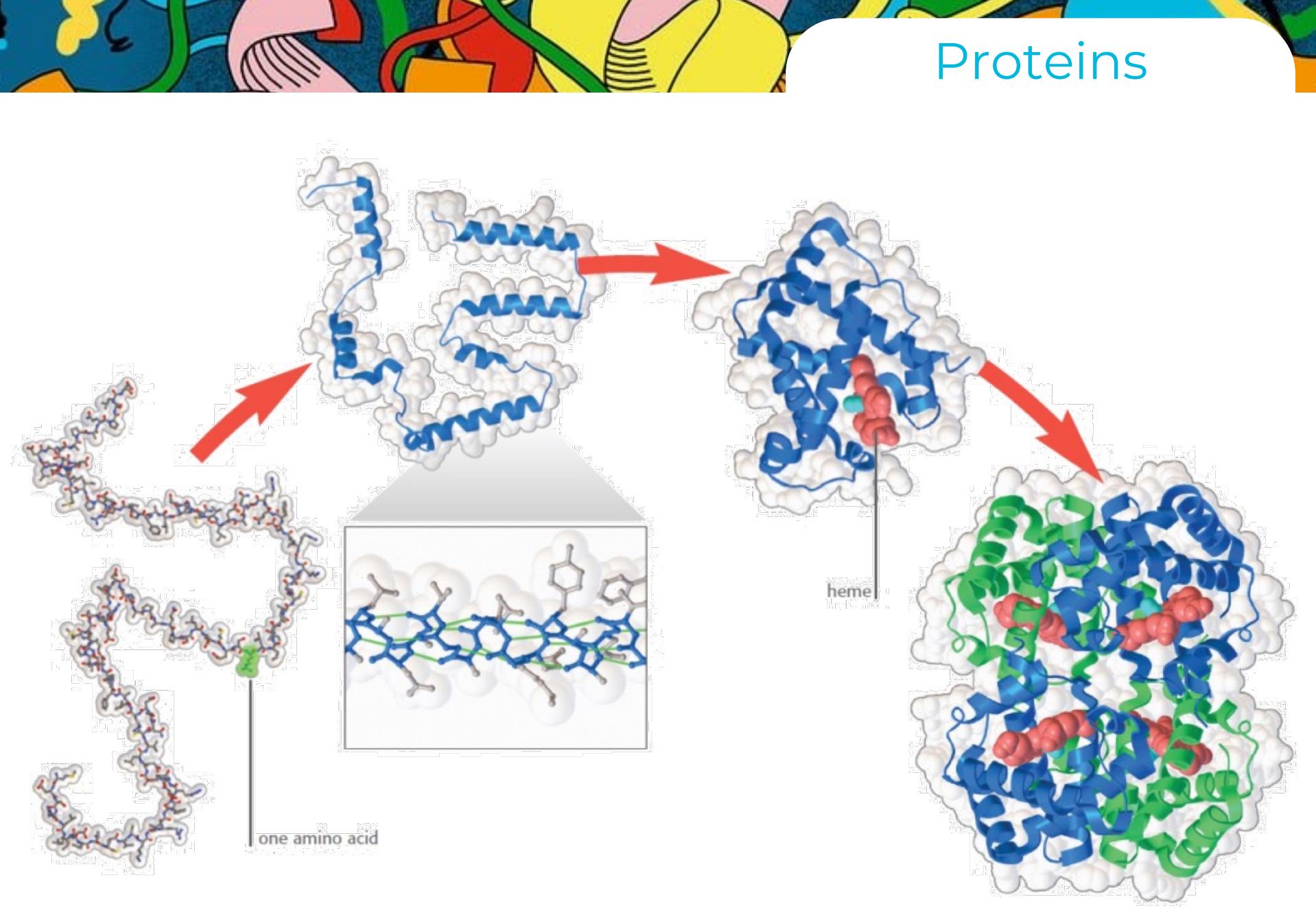
**Ewen Callaway**

By **Robert F. Service** | Nov. 30, 2020, 10:30 AM



<https://www.sciencemag.org/news/2526/1200064/has-changed-ai-triumphs-solving-protein-structures>

# Proteins



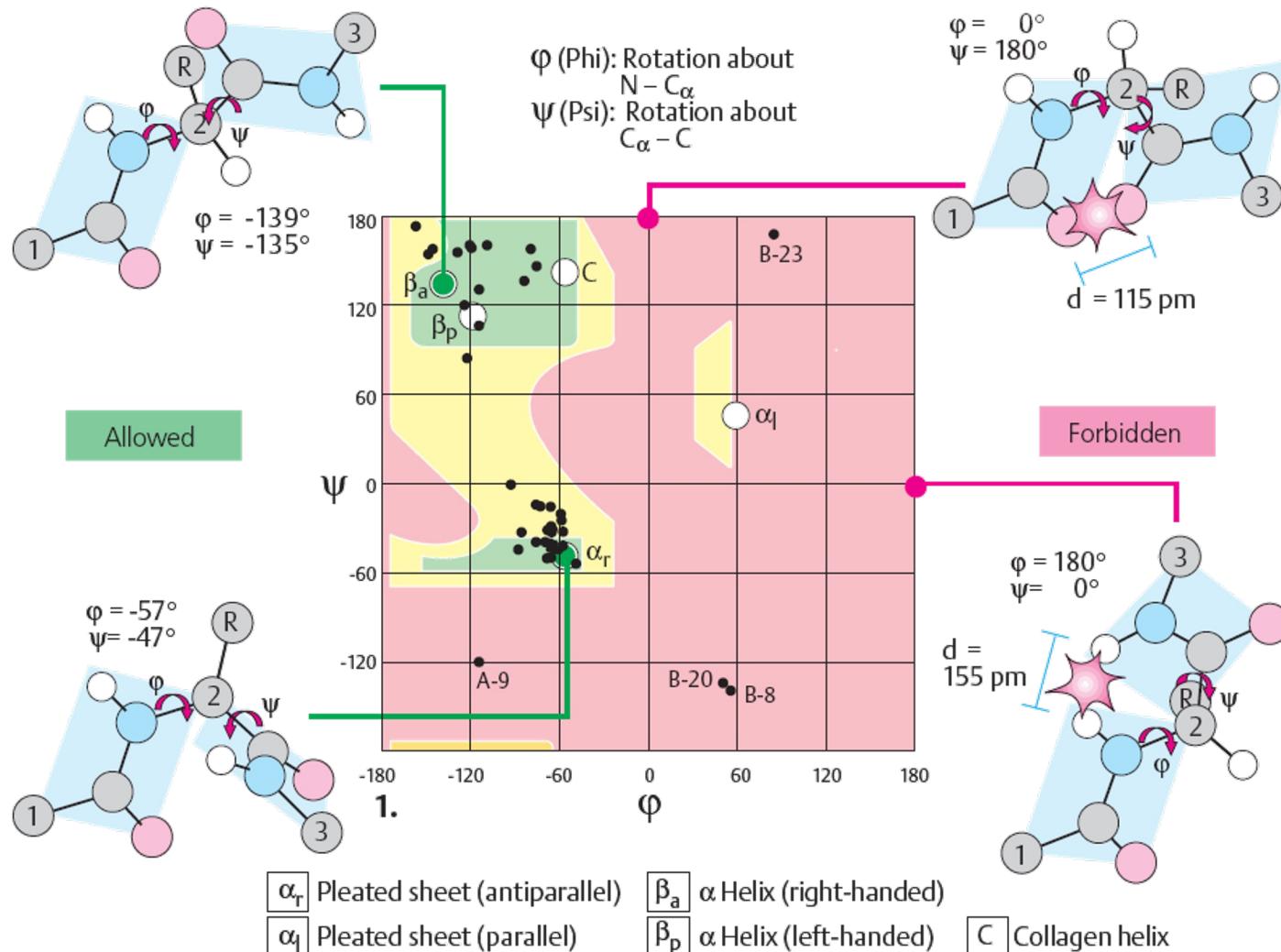
“Structural Biology is the study of the structure and dynamics of macro-biomolecules, particularly proteins and nucleic acids, and how changes in their structure affect their function. It incorporates principles of molecular biology, biochemistry and biophysics.”<sup>(Nature, 2019)</sup>

## Structure = Function

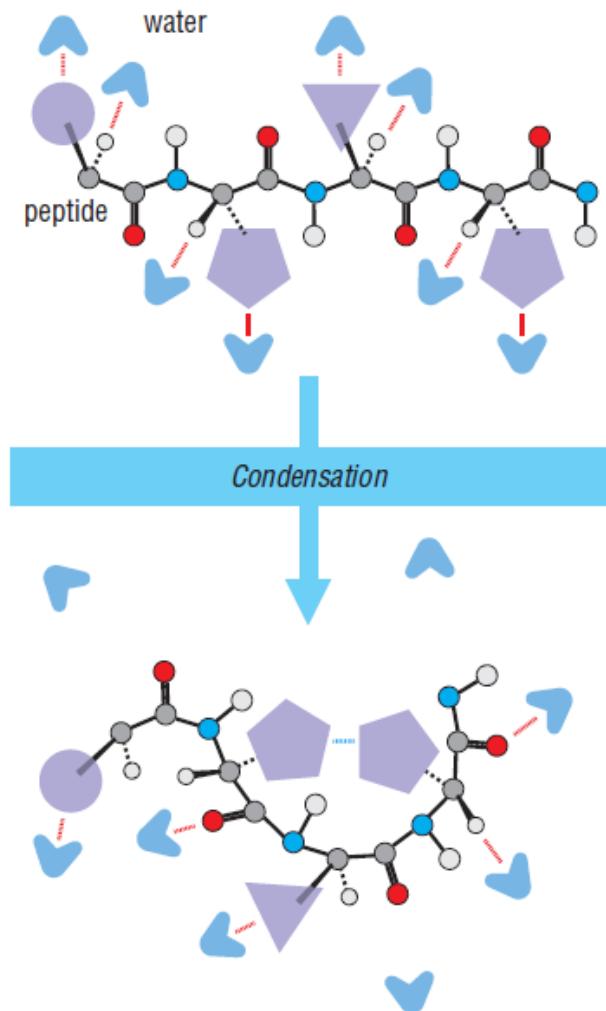
Marcus Vitruvius Pollio (I a.C.)



# Ramachandran Plot



# Folding



**Figure 1-23** Highly simplified schematic representation of the folding of a polypeptide chain in water. In the unfolded chain, side-chain and main-chain groups interact primarily with water, even if they are hydrophobic and the interaction is unfavorable. Burying the hydrophobic groups in the interior of a compact structure enables them to interact with each other (blue line), which is favorable, and leaves polar side chains on the surface where they can interact with water (red lines). The polar backbone groups that are buried along with the hydrophobic side chains must make hydrogen bonds to each other (not shown), as bulk water is no longer available.



# The Protein Folding Problem(s)



## Thermodynamics

How does structure  
emerge from the  
interactions between  
amino acids?

Folding code

## Kinetics

How is structure  
achieved so quickly?  
**Folding rate**

## Computational

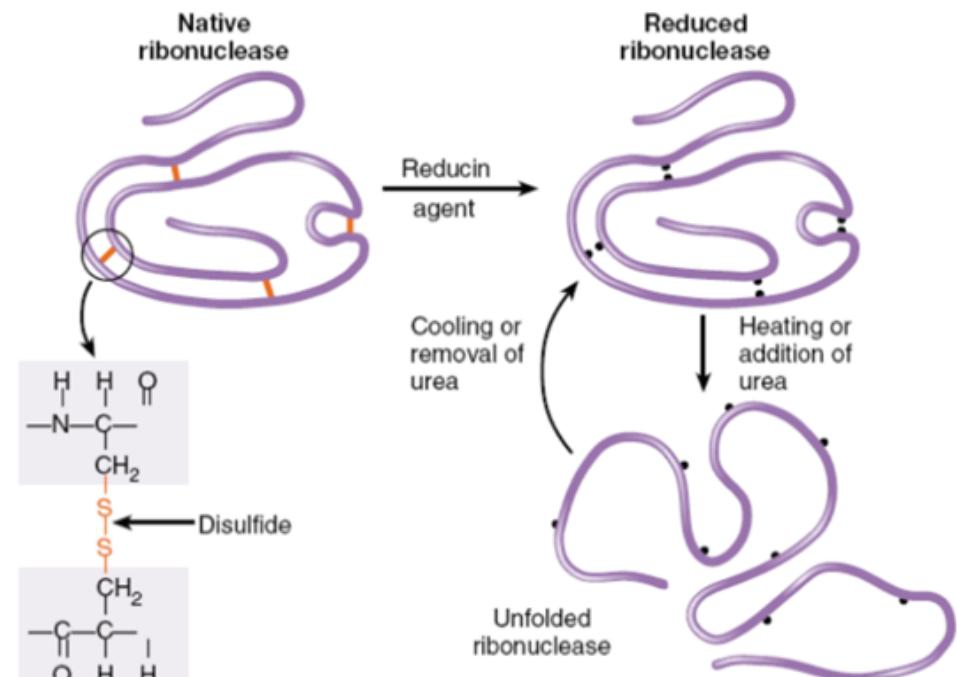
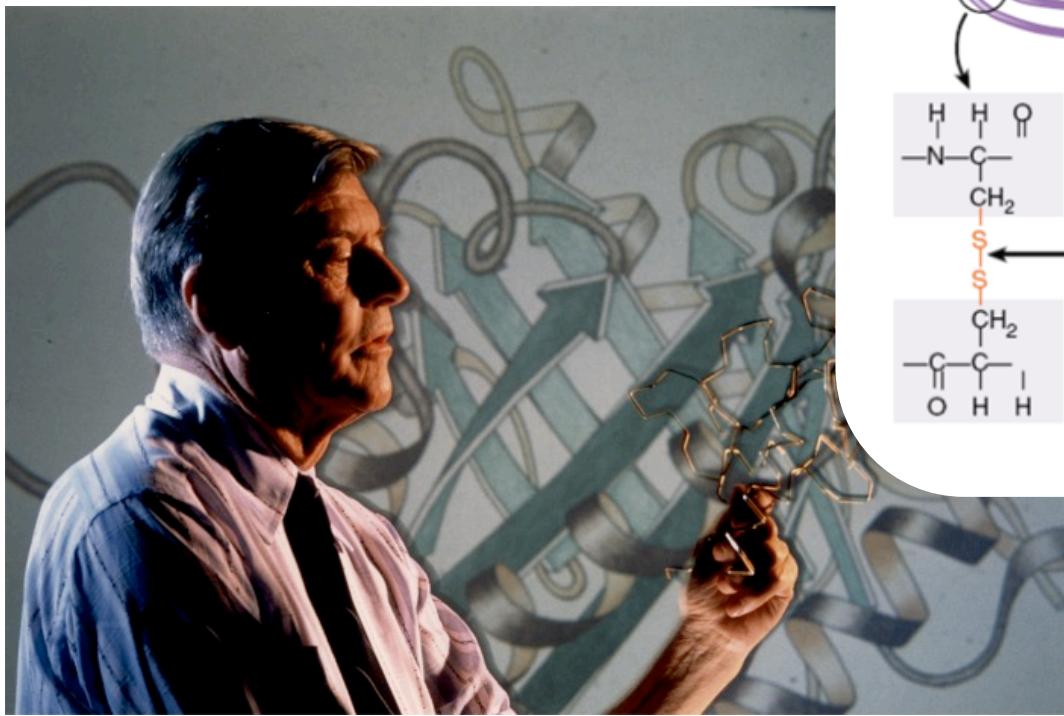
How to predict 3D  
structures starting  
from amino acids  
sequence?  
**Protein Structure  
Prediction**



# The Protein Folding Problem(s)

Anfinsen - 1965  
(Nobel - 1972)

Christian Boehmer Anfinsen, Jr.



# The Protein Folding Problem(s)

Levinthal - 1968

Levinthal's paradox

100 amino acids:  $3^{100} = 5 \times 10^{47}$  configurations

Sampling  $10^{13}$  per second ( $3 \times 10^{20}$  per year)

$10^{27}$  years

Universe:  $2 \times 10^{10}$  years

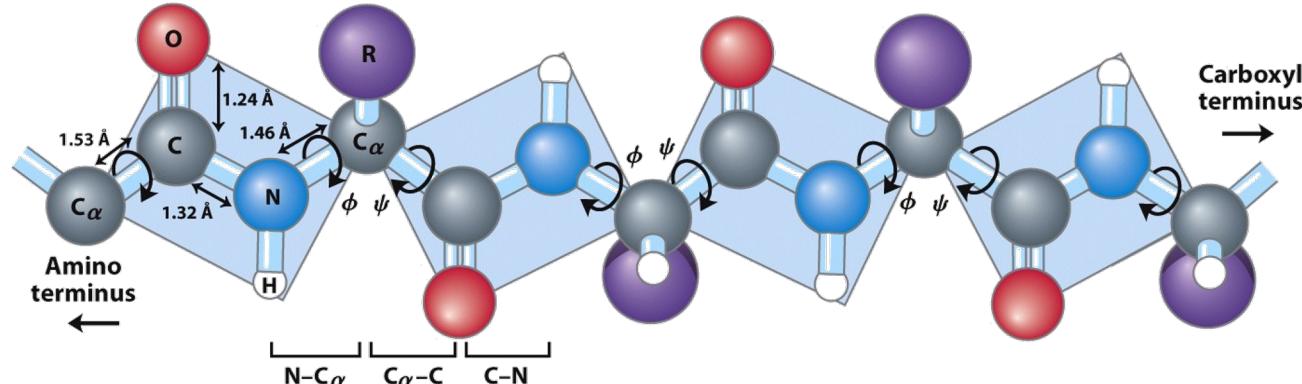
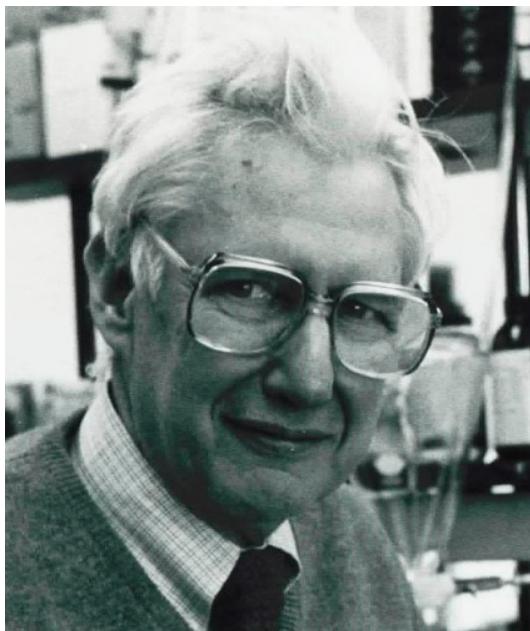
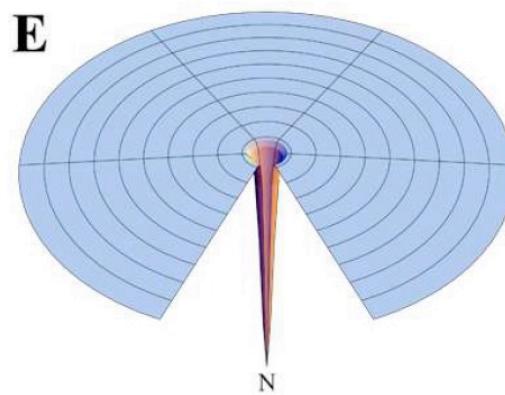
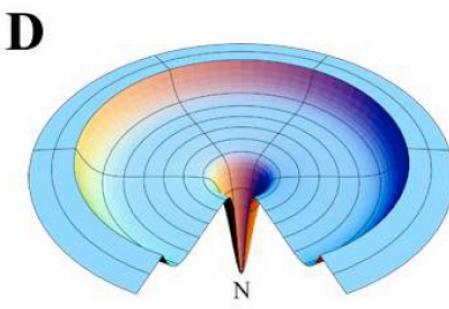
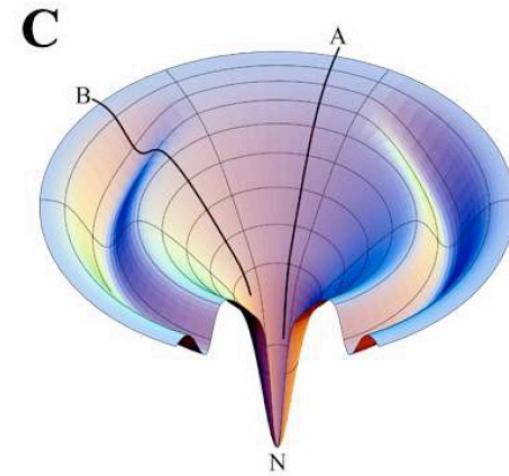
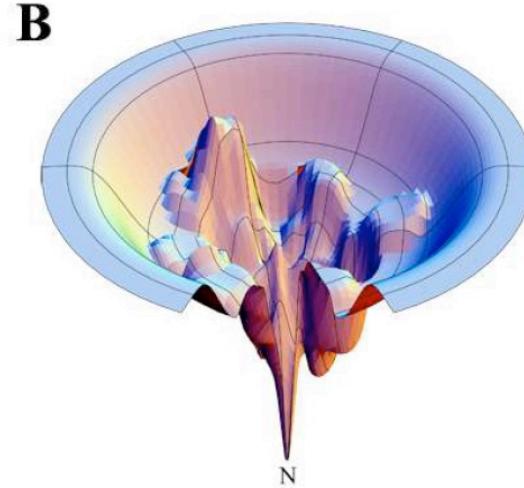
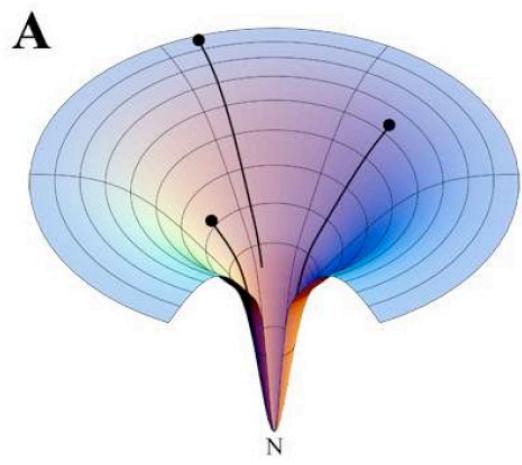


Figure 4-2b  
*Lehninger Principles of Biochemistry, Fifth Edition*  
© 2008 W.H. Freeman and Company

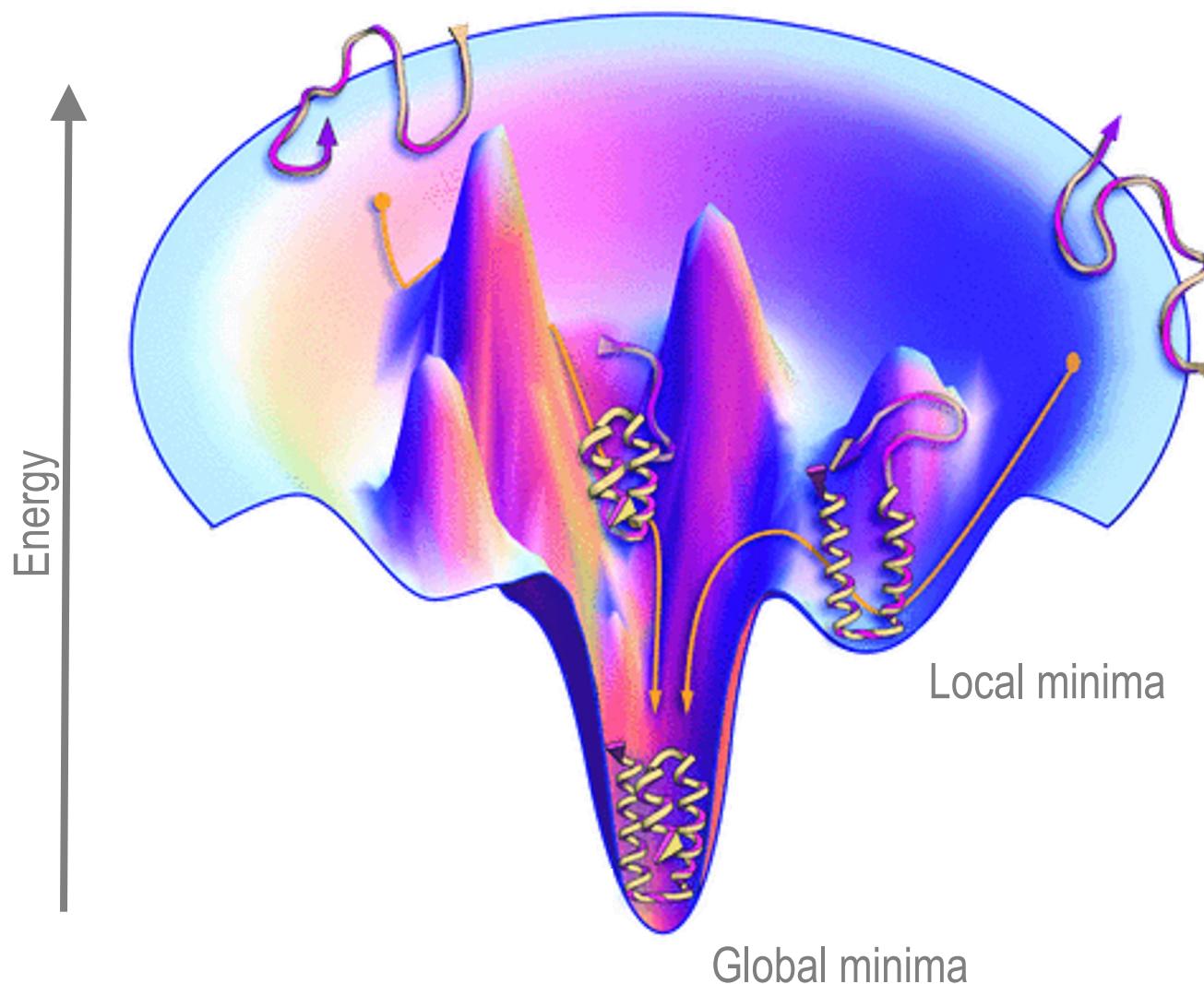




# The Protein Folding Problem(s)



# The Protein Folding Problem(s)



Global minima

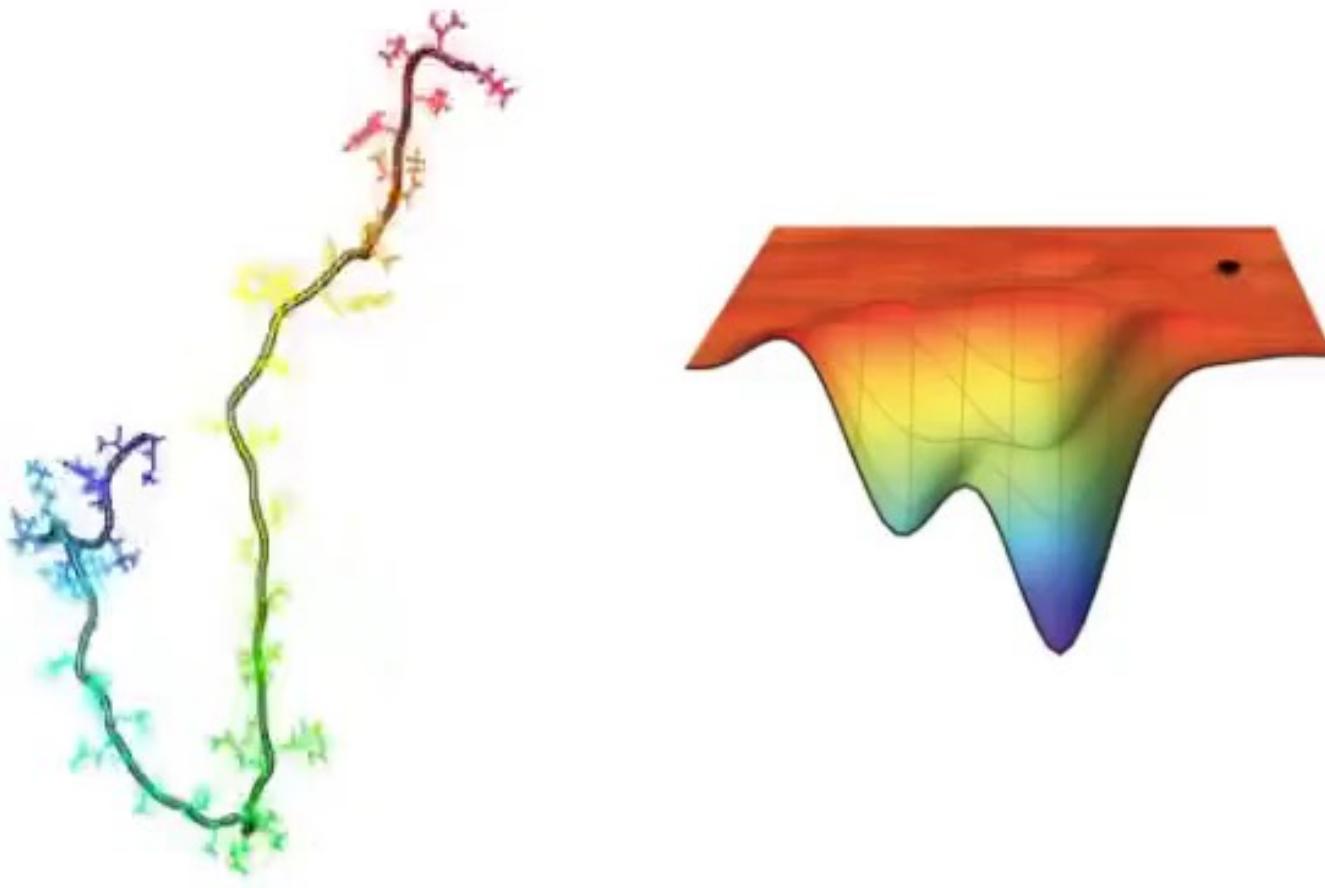
Local minima



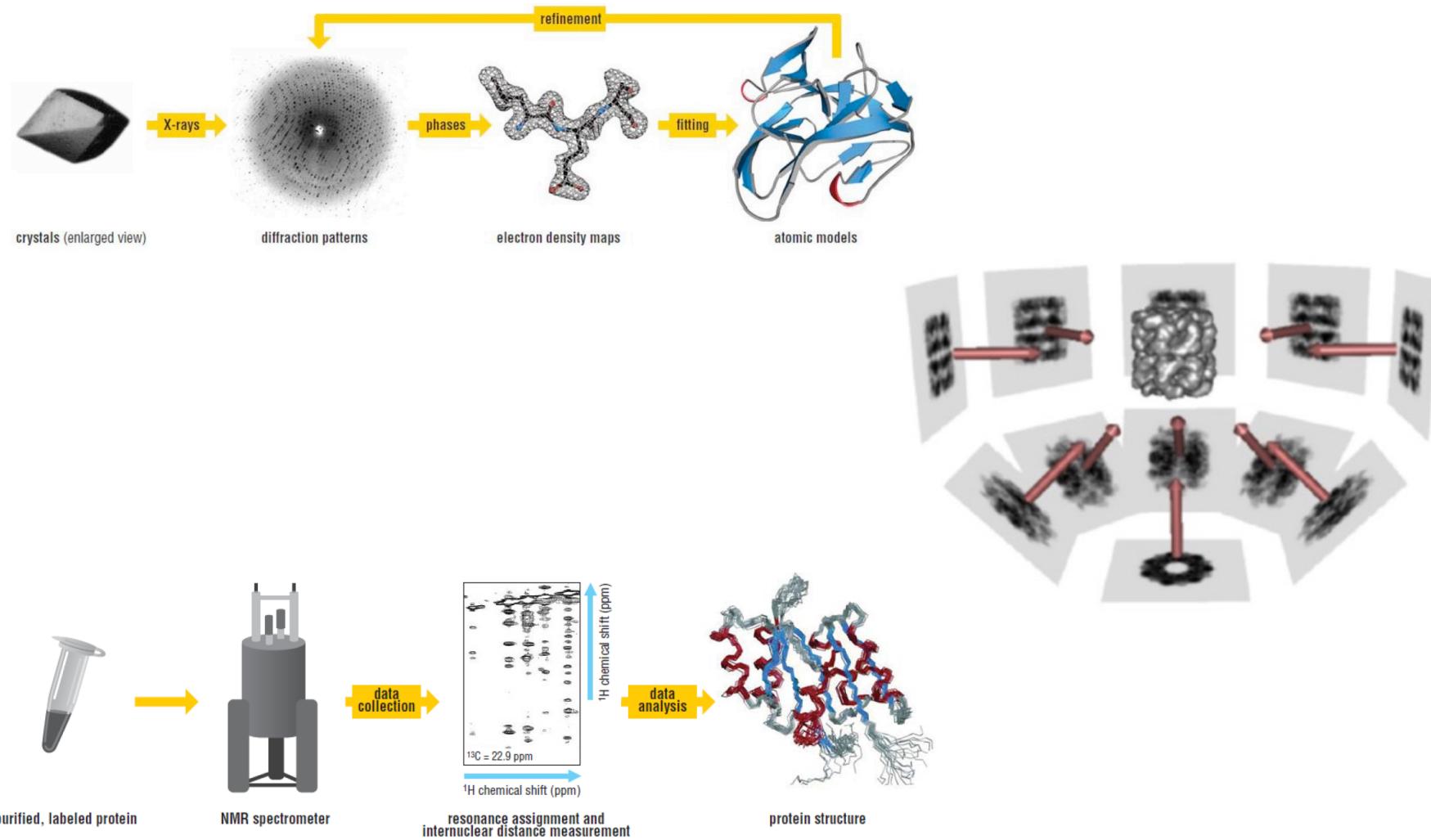
Dill & MacCallum (2012) Science 338:1042-6.



# The Protein Folding Problem

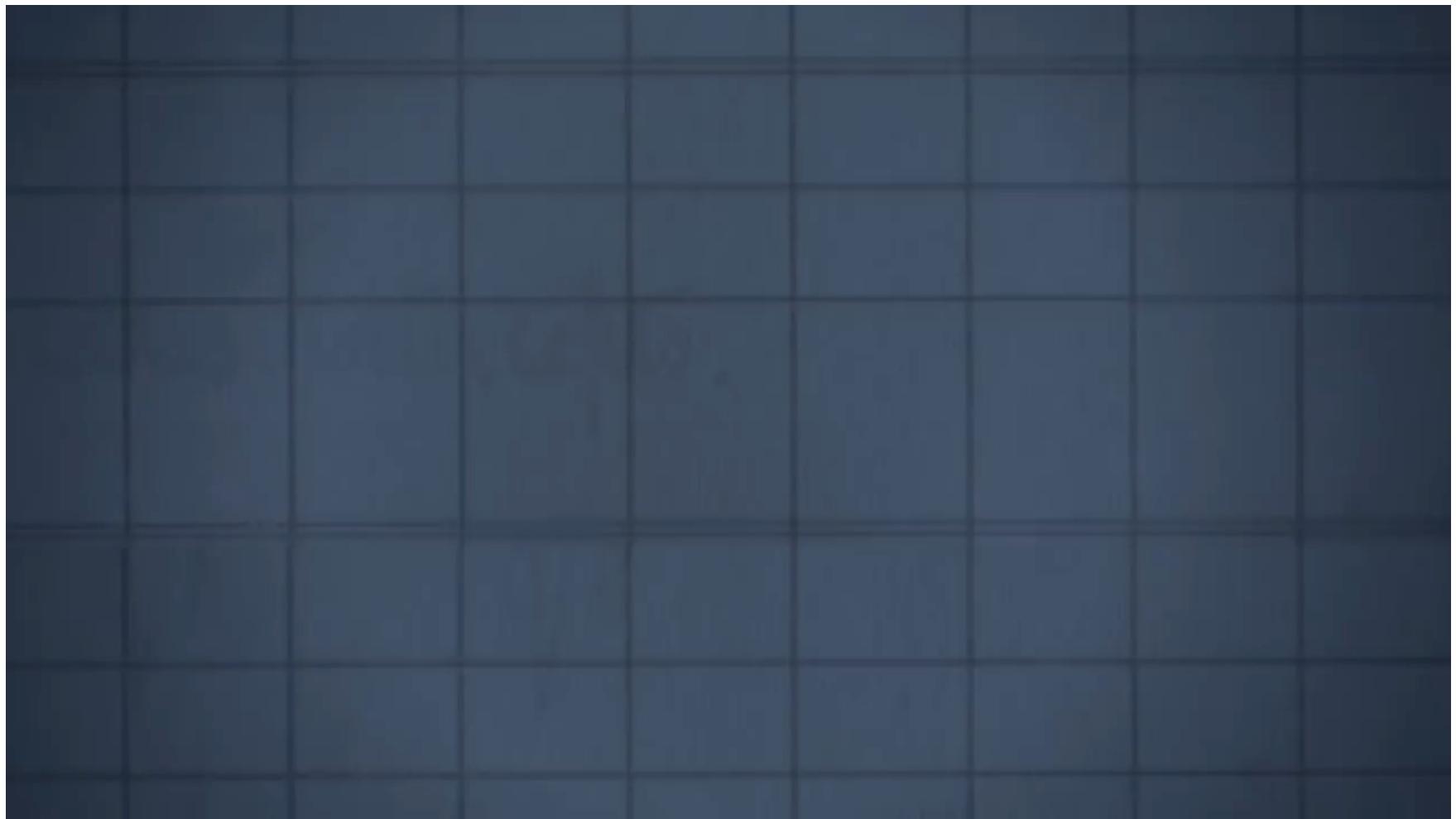


# Structural Solving





# Structural Solving



<https://youtu.be/uqQlwYv8VQI>



Petsko & Ringe, *Protein Structure and Function*, Blackwell (2004); Nogales & Scheres, *Mol Cell* 58, 677-89 (2015).



# Structural Solving

**Transmission Cryo-Electron Microscopy**  
A tool used by structural biologists to study  
molecular nanomachines

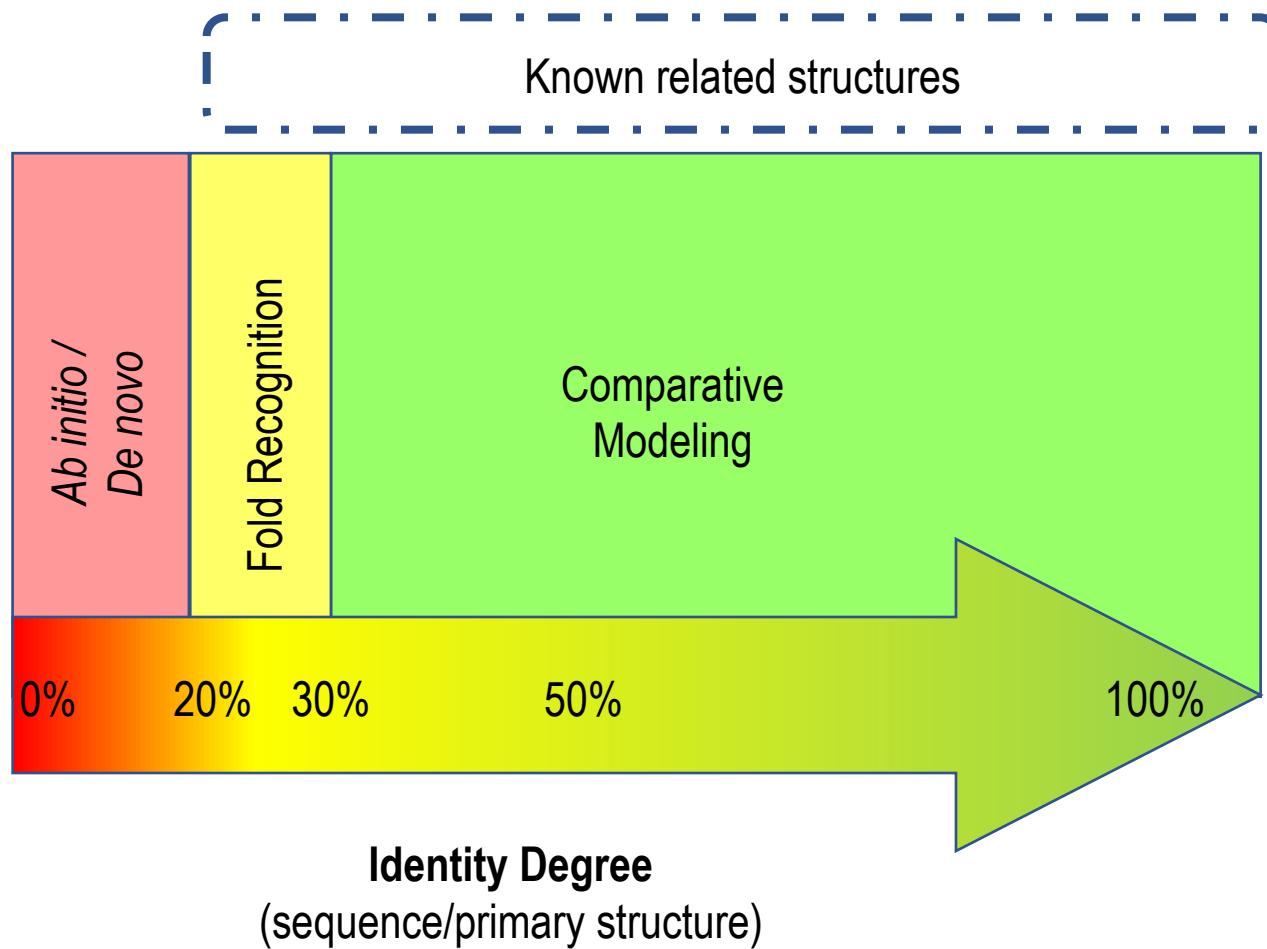
Gabriel Lander, Thesis Defense 2009

<https://youtu.be/BJKkC0W-6Qk>



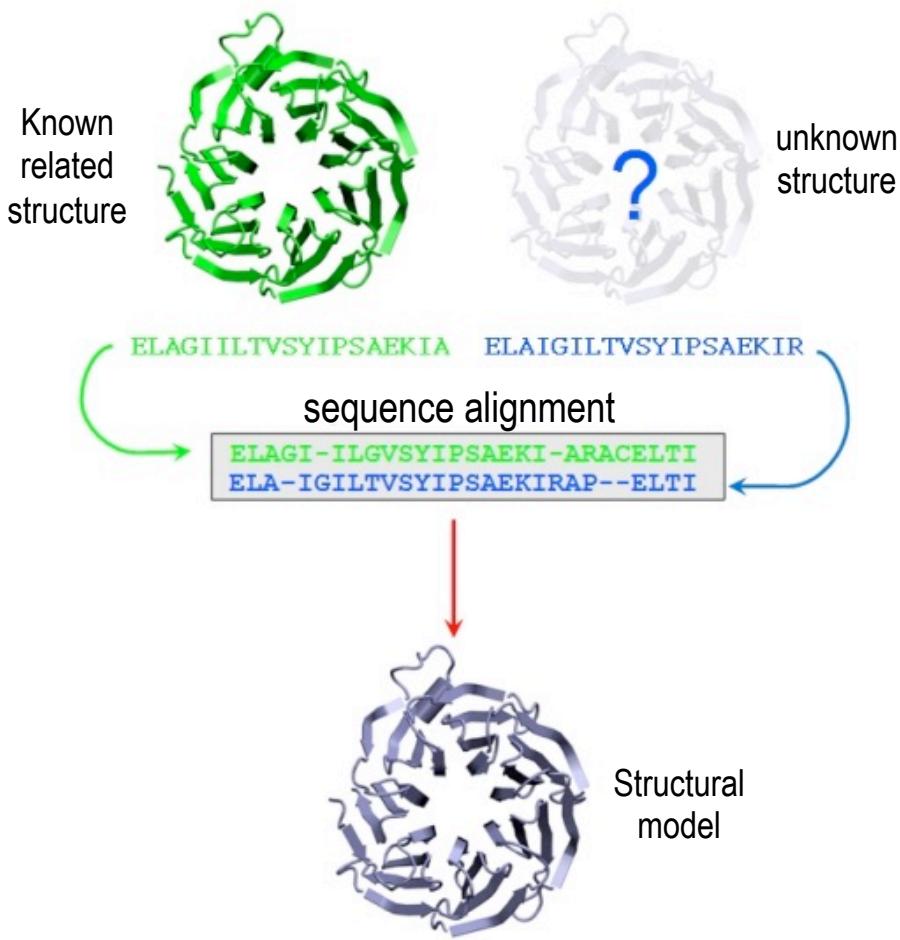
Petsko & Ringe, *Protein Structure and Function*, Blackwell (2004); Nogales & Scheres, *Mol Cell* 58, 677-89 (2015).

# Structural Modeling

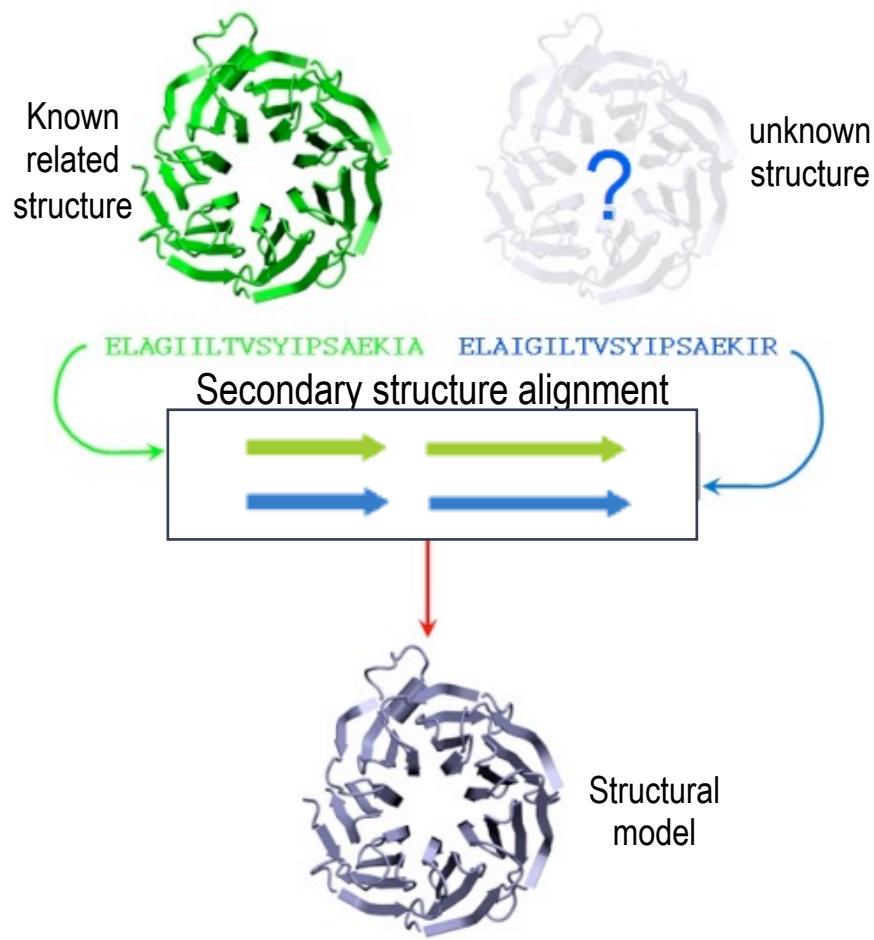


# Structural Modeling

## Comparative Modeling



## Fold Recognition



# Structural Modeling



Swiss Institute of Bioinformatics



## SWISS-MODEL

### Modelling

myWorkspace

Automated Mode

Alignment Mode

Project Mode

### Tools

Template Identification

Domain Annotation

Structure Assessment

Template Library

### Repository

Search by Sequence

Search by AC

SWISS-MODEL is a fully automated protein structure homology-modeling server, accessible via the ExPASy web server, or from the program DeepView (Swiss Pdb-Viewer). The purpose of this server is to make Protein Modelling accessible to all biochemists and molecular biologists WorldWide.

### What's new?

- New automated modeling pipeline with improved hierarchical approach for template selection.
- Increased sensitivity of template detection (sequence to profile search using an adapted HHSearch protocol)
- New tools for model and structure quality assessment: Dfire and Qmean global scores; ProQres residue based assessment scores

### SWISS-MODEL Team

Torsten Schwede: Project Leader

Florian Kiefer: SWISS-MODEL Repository

Lorenza Bordoli: Method Development and user support

Konstantin Arnold: SWISS-MODEL Workspace

### References:

When you publish or report results using SWISS-MODEL, please cite the relevant publications:

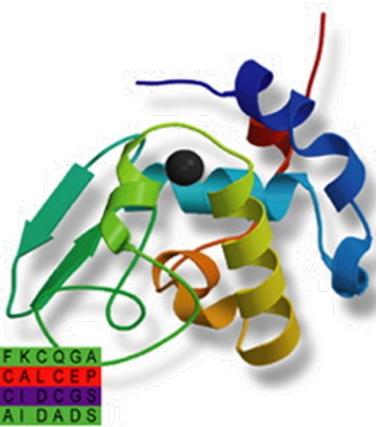
- Arnold K., Bordoli L., Kopp J., and Schwede T. (2006). The SWISS-MODEL Workspace: A web-based environment for protein structure homology modelling. *Bioinformatics*, 22, 195-201.
- Kiefer F., Arnold K., Künzli M., Bordoli L., Schwede T (2009). The SWISS-MODEL Repository and associated resources. *Nucleic Acids Research*. 37, D387-D392.
- Peitsch, M. C. (1995) Protein modeling by E-mail *Bio/Technology* 13: 658-660.

# Structural Modeling

[To main Sali lab pages](#)

## Modeller

Program for Comparative Protein Structure Modelling by Satisfaction of Spatial Restraints



A	I	L	V	G	S	M	P	R	R	D	G	M	E	R	K	D	L	L	K	A	N	V	K	I	F	K	C	Q	G	A
V	E	V	C	P	V	D	C	F	Y	E	G	P	N	F	L	V	I	H	P	D	E	C	I	D	C	A	L	C	E	P
G	A	C	K	P	E	C	P	V	N	I	Q	G	S	-	-	Y	A	I	D	A	D	S	C	I	D	G	S	I	G	
C	-	-	I	A	C	G	A	C	K	P	E	C	P	V	N	I	I	Q	G	S	-	-	I	Y	A	I	D	A	D	S

About MODELLER

MODELLER News

Download & Installation

Release Notes  
Data file downloads

Registration

Accelrys licensing

Discussion Forum

## About MODELLER

MODELLER is used for homology or comparative modeling of protein three-dimensional structures (1,2). The user provides an alignment of a sequence to be modeled with known related structures and MODELLER automatically calculates a model containing all non-hydrogen atoms. MODELLER implements comparative protein structure modeling by satisfaction of spatial restraints (3,4), and can perform many additional tasks, including de novo modeling of loops in protein structures, optimization of various models of protein structure with respect to a flexibly defined objective function, multiple alignment of protein sequences and/or structures, clustering, searching of sequence databases, comparison of protein structures, etc. MODELLER is [available for download](#) for most Unix/Linux systems, Windows, and Mac.

[www.salilab.org/modeller](http://www.salilab.org/modeller)

# Structural Modeling

Standard Mode | [Login](#) for job manager, batch processing, Phyre alarm and other advanced options

Retrieve Phyre Job Id

Fetch

# Phyre<sup>2</sup>

Protein Homology/analogY Recognition Engine V 2.0

Subscribe to Phyre at Google Groups

Email:

[Visit Phyre at Google Groups](#)

 Follow @Phyre2server

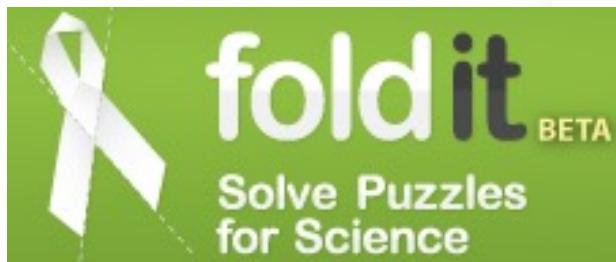


Nominate your protein to be experimentally solved!  
New CASP13 experiment upcoming. [MORE INFO](#)

[EBI 2017 Workshop](#) | [Older Workshops](#) | [Phyre2 paper](#)

[www.sbg.bio.ic.ac.uk/phyre2](http://www.sbg.bio.ic.ac.uk/phyre2)

# Structural Modeling



## SCIENTIFIC AMERICAN™

[Sign In](#) | [Register](#)

[Subscribe](#)

[News & Features](#)

[Topics](#)

[Blogs](#)

[Videos & Podcasts](#)

[Education](#)

[Citizen Science](#)

Technology » News

6 :: Email :: Print



## Foldit Gamers Solve Riddle of HIV Enzyme within 3 Weeks

The online game poses protein-folding puzzles, and participants provided insights recently that solved the structure of an enzyme involved in reproduction of HIV

By Michael J. Coren and Fast Company | September 20, 2011

When video gamers armed with the world's most powerful supercomputers take on science and its most vexing riddles, who wins? Sometimes, it's the gamers.

Humans retain an edge over computers when complex problems require intuition



<https://fold.it>

# Critical Assessment of Techniques for Structure Prediction (CASP) competition



## Protein Structure Prediction Center

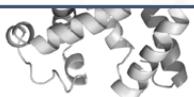
**Menu**

[Home](#)  
[PC Login](#)  
[PC Registration](#)

▼ **CASP Experiments**

[CASP14 \(2020\)](#)  
[CASP Commons \(COVID-19, 2020\)](#)  
[CASP13 \(2018\)](#)  
[CASP12 \(2016\)](#)  
[CASP11 \(2014\)](#)  
[CASP10 \(2012\)](#)  
[CASP9 \(2010\)](#)  
[CASP8 \(2008\)](#)  
[CASP7 \(2006\)](#)  
[CASP6 \(2004\)](#)  
[CASP5 \(2002\)](#)  
[CASP4 \(2000\)](#)  
[CASP3 \(1998\)](#)  
[CASP2 \(1996\)](#)  
[CASP1 \(1994\)](#)

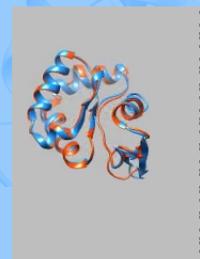
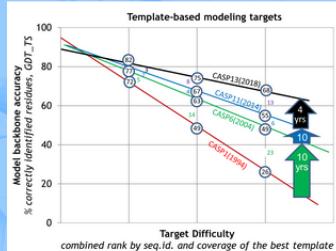
► **Initiatives**  
**Data Archive**  
**Proceedings**  
**CASP Measures**  
**Feedback**  
**Assessors**  
**People**  
**Community Resources**  
**Job Fair**



### Success Stories From Recent CASPs

[template-based modeling](#)   [ab initio modeling](#)   [contact prediction](#)   [help structural biologists](#)   [refinement](#)   [data-assisted modeling](#)   >

Models based on templates identified by sequence similarity remain the most accurate. Over the course of the CASP experiments there have been enormous improvements in this area. However, the overall accuracy improvements that we have seen in the first 10 years of CASP remained unmatched until CASP12 (2016), when a new burst of progress happened [Kryshtafovych et al., 2018]. In two years from 2014 to 2016, the backbone accuracy of the submitted models improved more than in the preceding 10 years. The next CASP continued the trend [Croll et al., 2019], and the 2014–2018 model accuracy improvement doubled that of 2004–2014 (see the plot). Several factors contributed to this, including more accurate alignment of the target sequence to that of available templates, combining multiple templates, improved accuracy of regions not covered by templates, successful refinement of models, and better selection of models from decoy sets due to improved methods for estimation of model accuracy.



target T0868-D1 (orange)  
model 330\_2 (blue): GDT\_TS=87 best template: 2cw6 (seq.id= 4.2%)

### Welcome to the Protein Structure Prediction Center!

Our goal is to help advance the methods of identifying protein structure from sequence. The Center has been organized to provide the means of objective testing of these methods via the process of blind prediction. The Critical Assessment of protein Structure Prediction (CASP) experiments aim at establishing the current state of the art in protein structure prediction, identifying what progress has been made, and highlighting where future effort may be most productively focused.

There have been fourteen previous CASP experiments. The fifteenth experiment is planned to start in Spring 2022. Description of these experiments and the full data (targets, predictions, interactive tables with numerical evaluation results, dynamic graphs and prediction visualization tools) can be accessed following the links:

[CASP1 \(1994\)](#) | [CASP2 \(1996\)](#) | [CASP3 \(1998\)](#) | [CASP4 \(2000\)](#) | [CASP5 \(2002\)](#) | [CASP6 \(2004\)](#) | [CASP7 \(2006\)](#) | [CASP8 \(2008\)](#) | [CASP9 \(2010\)](#) | [CASP10 \(2012\)](#) | [CASP11 \(2014\)](#) | [CASP12 \(2016\)](#) | [CASP13 \(2018\)](#) | [CASP14 \(2020\)](#)

Raw data for the experiments held so far are archived and stored in our [data archive](#).

---

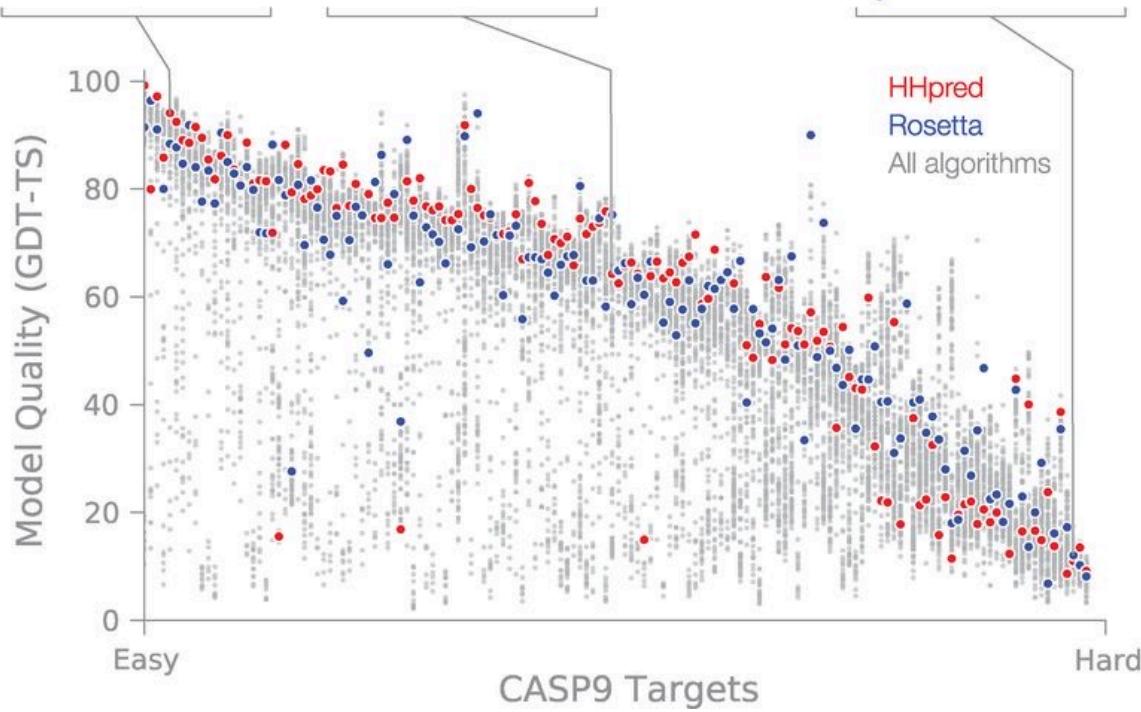
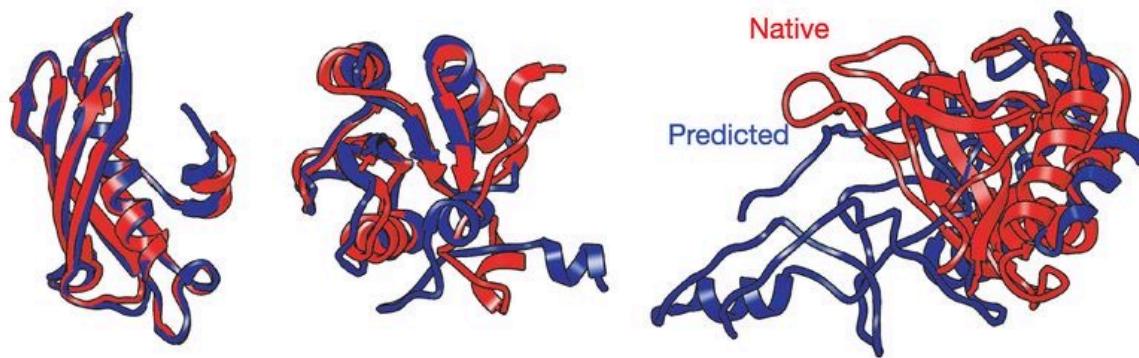
Details of the experiments have been published in a scientific journal *Proteins: Structure, Function and Bioinformatics*. [CASP proceedings](#) include papers describing the structure and conduct of the experiments, the numerical evaluation measures, reports from the assessment teams highlighting state of the art in different prediction categories, methods from some of the most successful prediction teams, and progress in various aspects of the modeling.

---

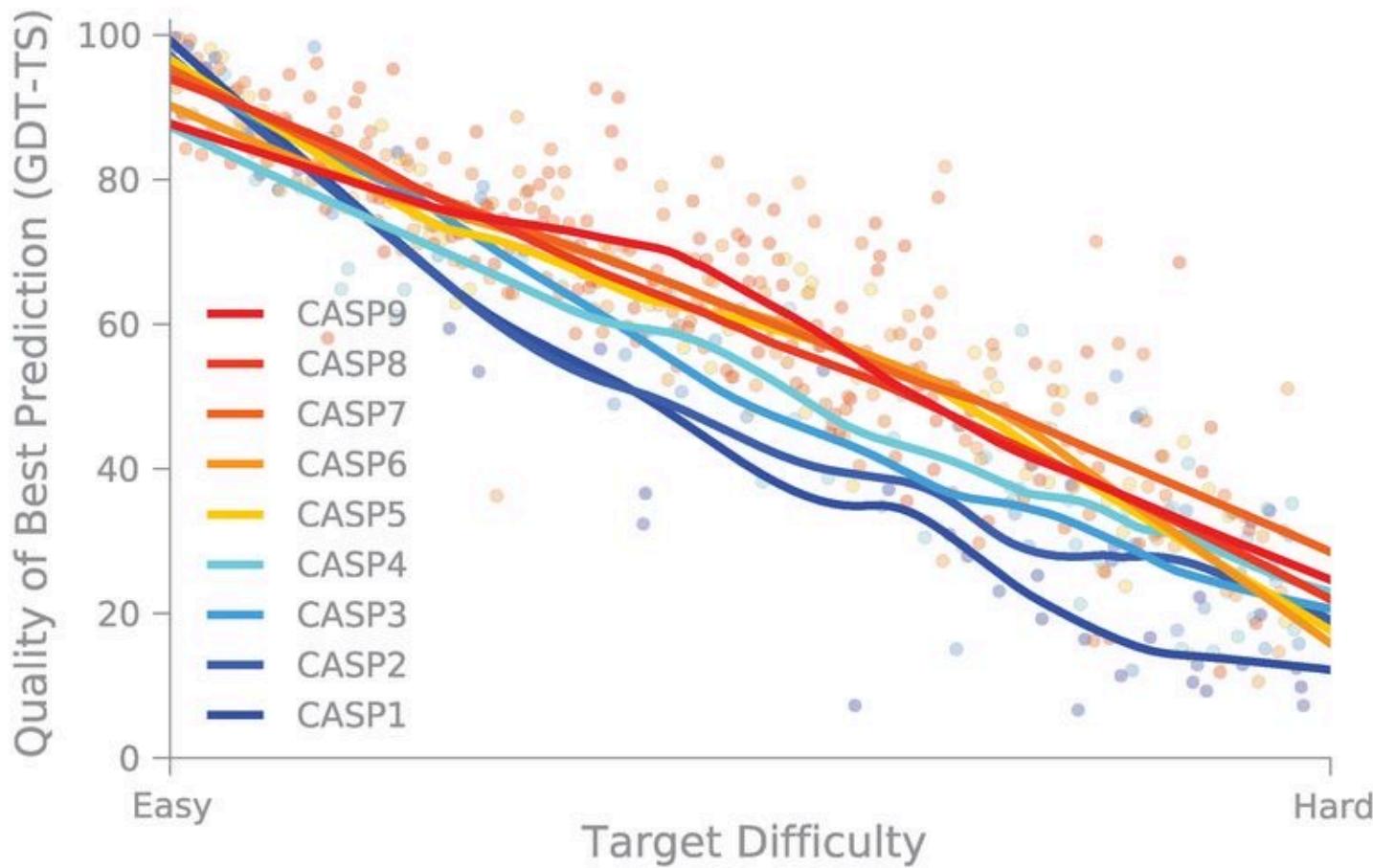
Prediction methods are assessed on the basis of the analysis of a large number of blind predictions of protein structure. Summary of numerical evaluation of the tertiary structure prediction methods tested in the latest CASP experiment can be found [on this web page](#). The main numerical measures used in evaluations, data handling procedures, and guidelines for navigating the data presented on this website are described in [1].

Some of the best performing methods are implemented as [fully automated servers](#) and therefore can be used by public for protein structure modeling.



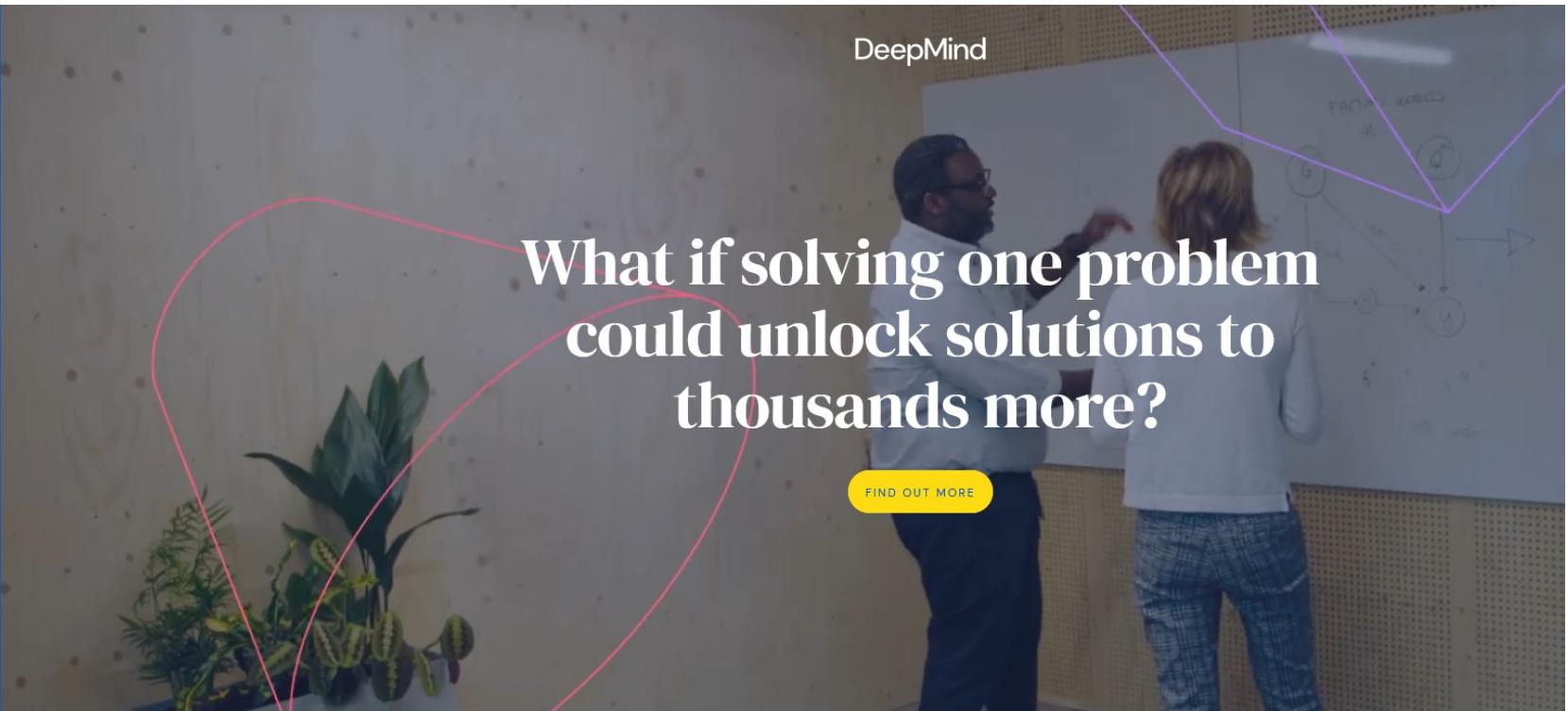
**B Performance in CASP9**

## A Historical CASP Performance





# What about Google?



DeepMind

## What if solving one problem could unlock solutions to thousands more?

[FIND OUT MORE](#)

We research and build safe artificial intelligence systems. Our goal is to solve intelligence and advance scientific discovery for all.



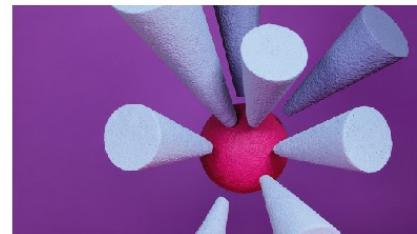




# What about Google?



WaveNet: A generative model for raw audio



Giving doctors a headstart on acute kidney injury



More accurately identifying breast cancer



AlphaStar plays StarCraft II at Grandmaster level



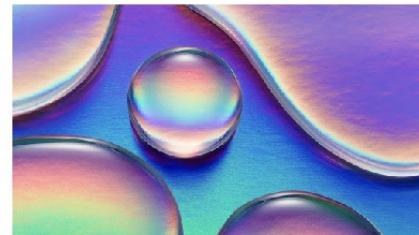
AlphaZero: Shedding new light on chess, shogi, and Go



DQN: Human-level control of Atari games



A neural network with dynamic memory



AlphaGo defeats Lee Sedol in the game of Go

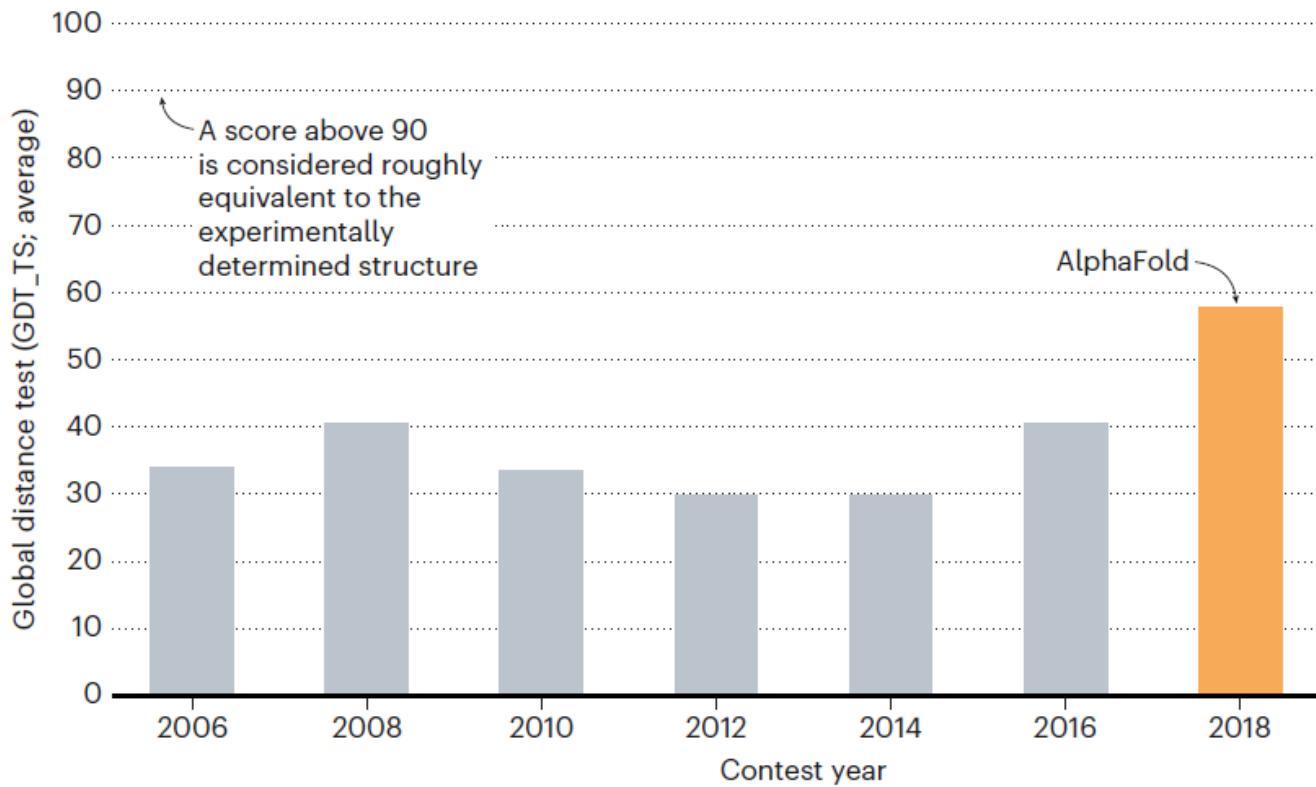


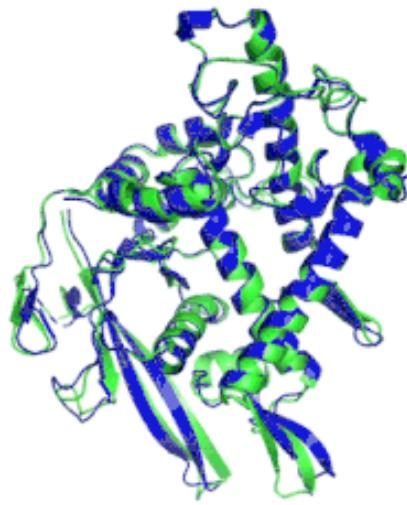
GQN: Neural scene representation and rendering



## 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.





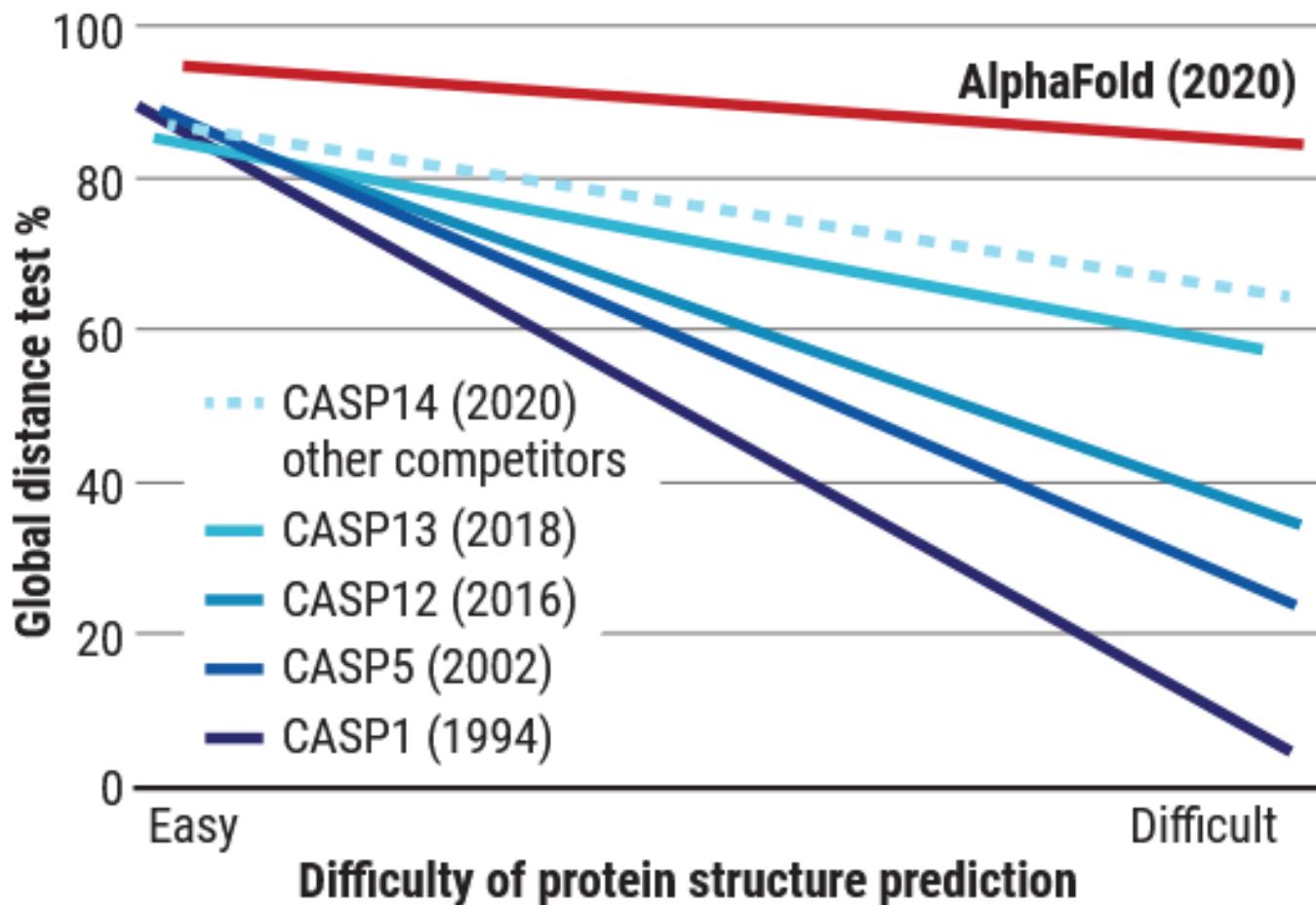
T1037 / 6vr4  
90.7 GDT  
(RNA polymerase domain)



T1049 / 6y4f  
93.3 GDT  
(adhesin tip)

- Experimental result
- Computational prediction



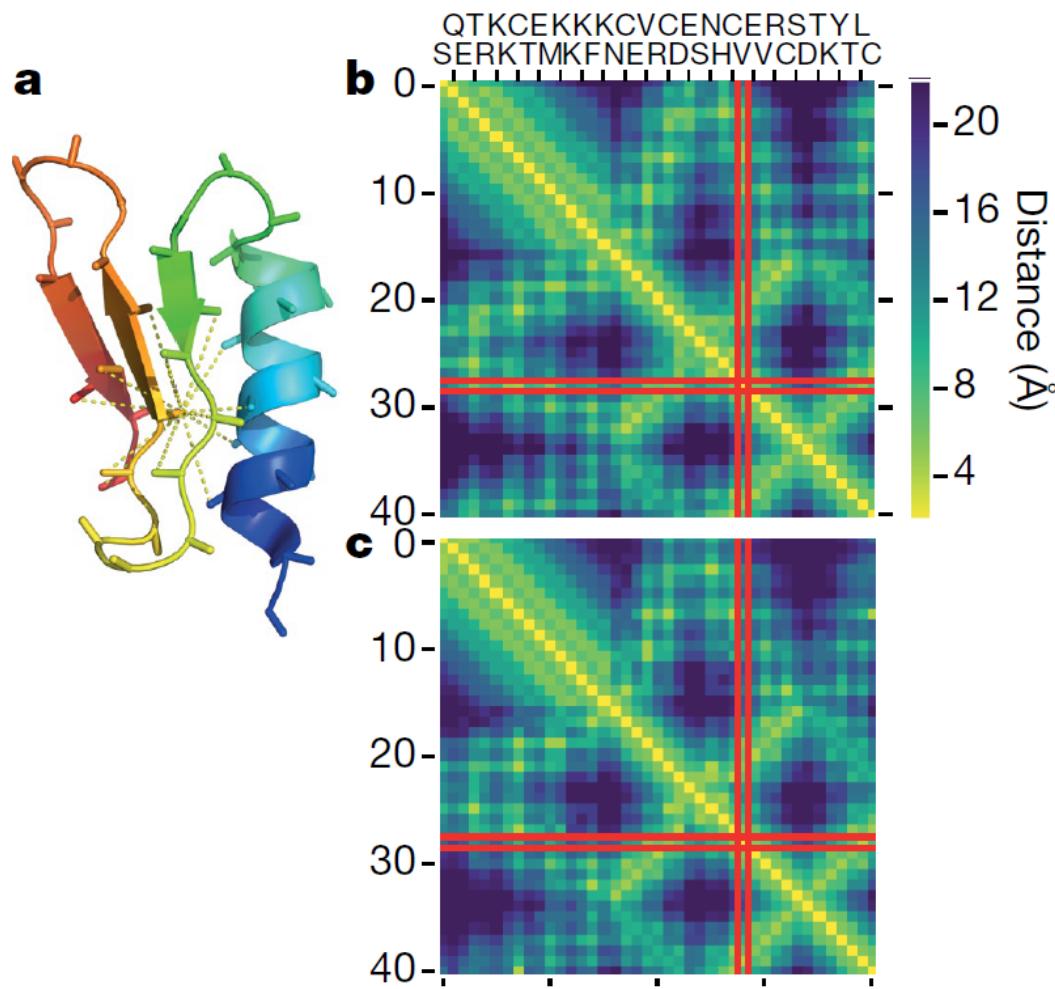




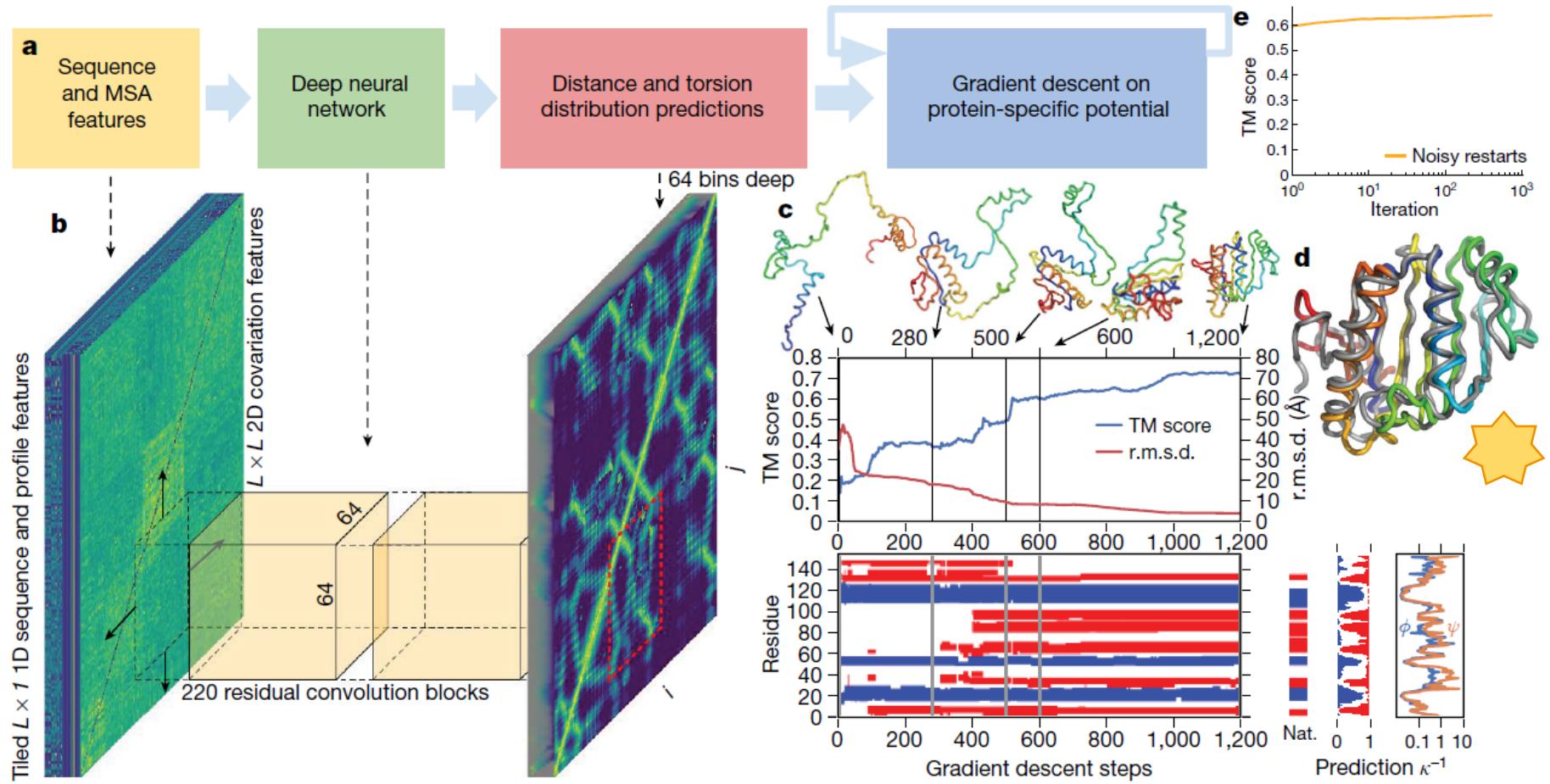
# AlphaFold2



<https://youtu.be/gg7WjuFs8F4>



# AlphaFold2



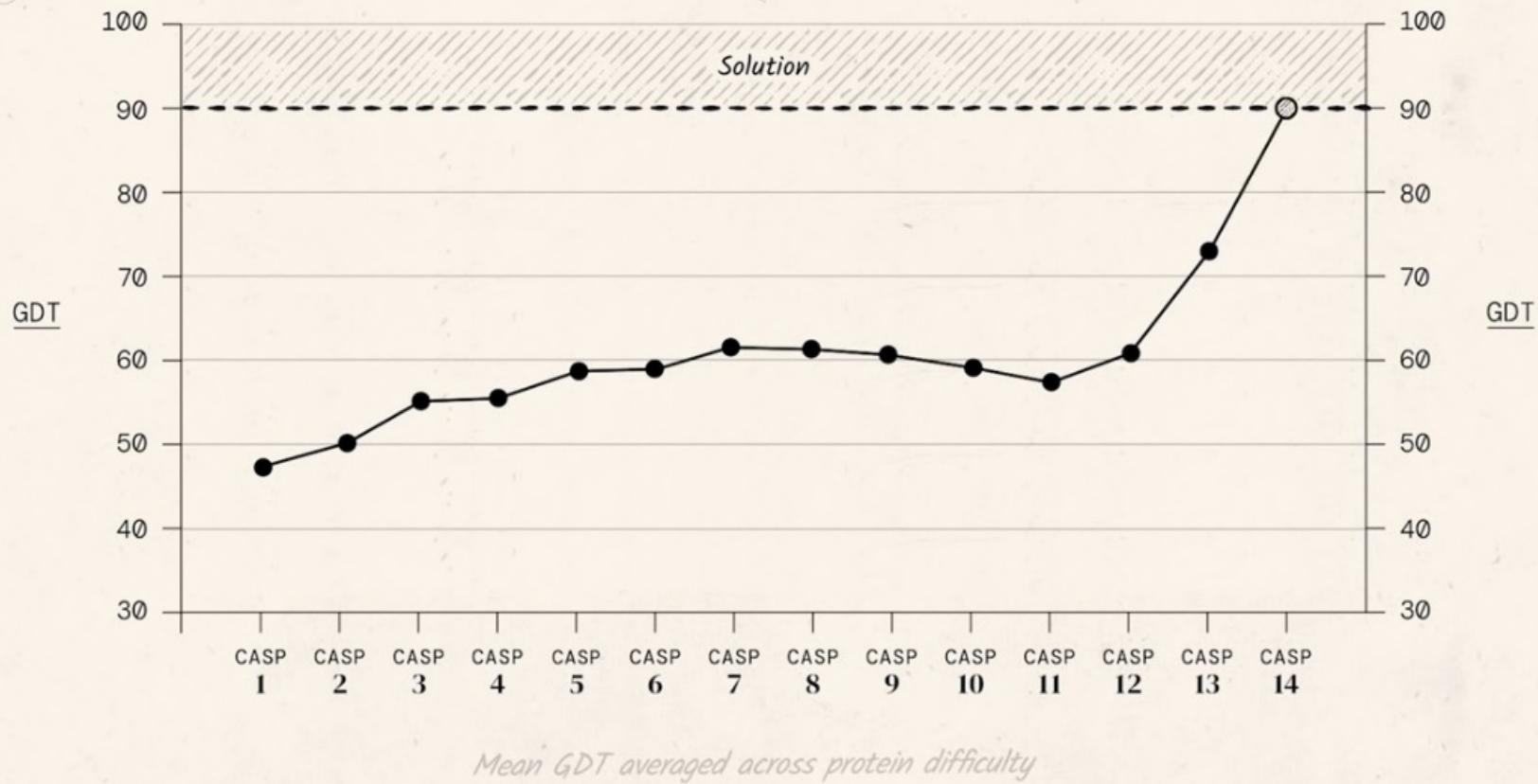


AlphaFold2



<https://youtu.be/gg7WjuFs8F4>

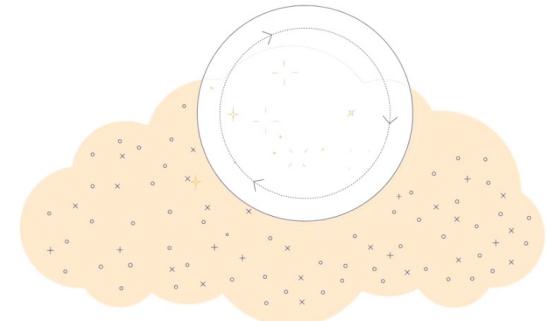
## Critical Assessment of Structure Prediction





Potential

# Basic Science Drug Design Protein Engineering





Are we done?

“Being able to investigate the shape of proteins quickly and accurately has the potential to revolutionize life sciences.

**Now that the problem has been largely solved for single proteins**, the way is open for development of new methods for determining the shape of protein complexes – collections of proteins that work together to form much of the machinery of life, and for other applications.”



## Wrap up & future outlook

© 2020 DeepMind Technologies Limited

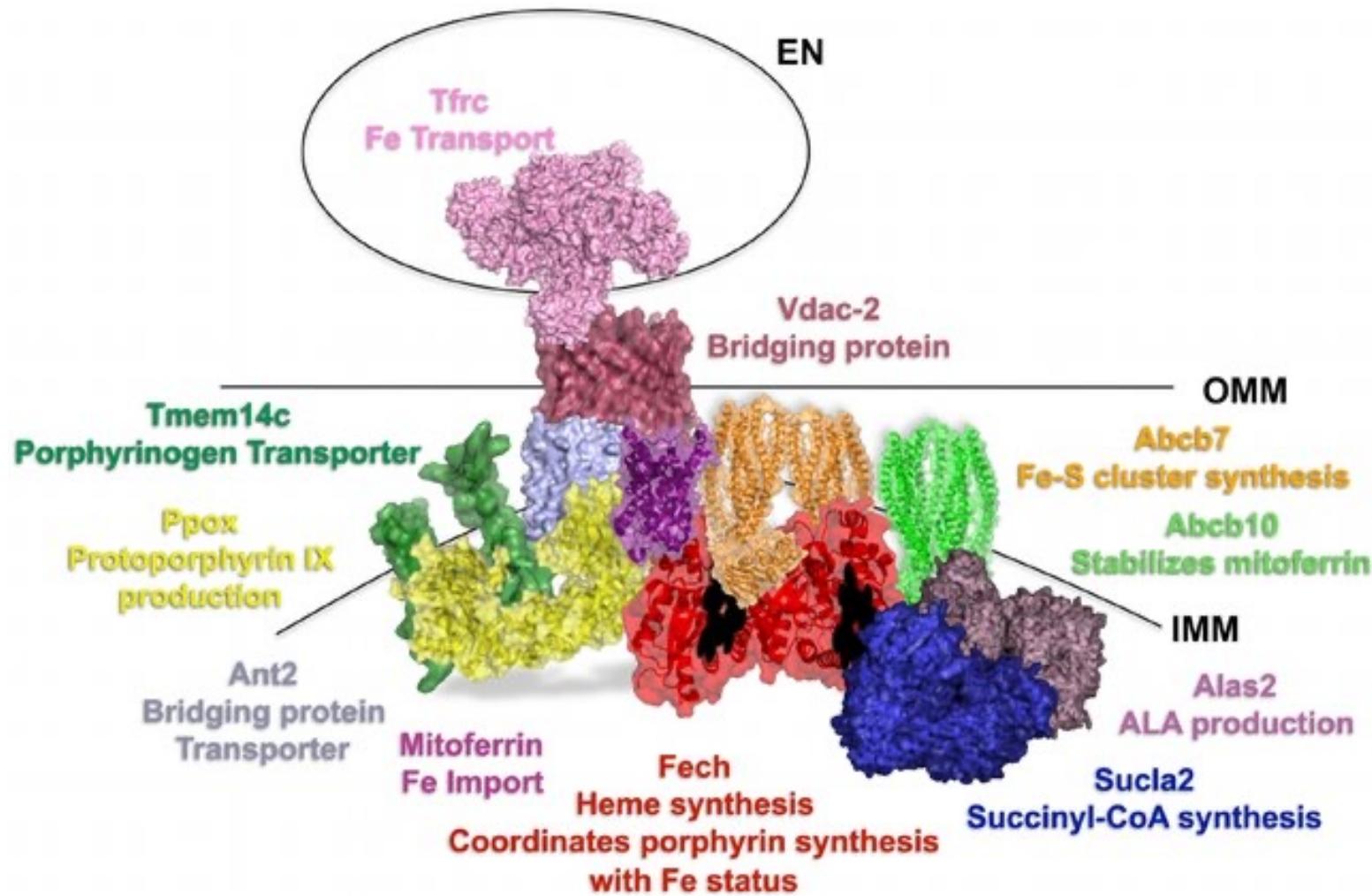
- We have built a system that confidently predicts accurate structures for most proteins – and knows when it is wrong
- As for CASP13<sup>1,2</sup>, we'll publish a peer-reviewed paper
- We're also working on providing broad access to our work
- Lots of exciting work ahead for the field: Complexes, conformational change etc

[1] Senior, A. W., et al. "Improved protein structure prediction using potentials from deep learning." *Nature* 577.7792 (2020): 706-710.

[2] Senior, A. W., et al. "Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13)." *Proteins* 87.12 (2019): 1141-1148.



# Protein Complexes/Oligomers



Are we done?



bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.500000; this version posted July 19, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

...

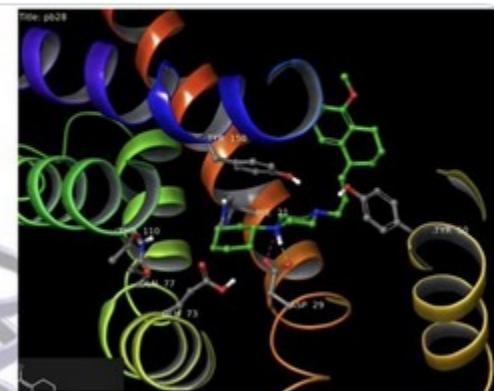
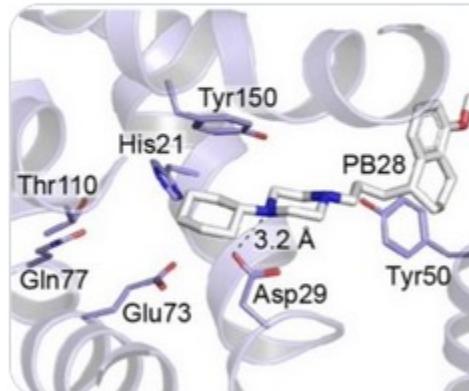
I'm pretty impressed with AlphaFold2. Predicted structure of sigma2/tmem97 on left vs. unpublished x-ray structure on right. Keep in mind, there are no good templates for homology models.



bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.500000; this version posted July 19, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

...

Had someone dock pb28 into the PDB for me. It is strikingly similar to the published structure:



2



3



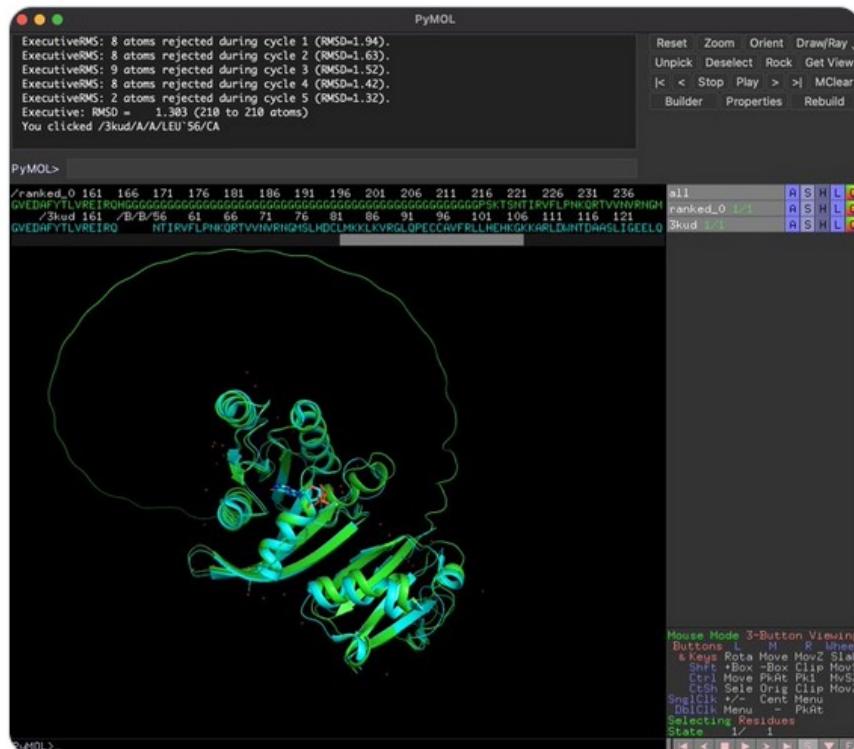
1:37 AM · Jul 19, 2021 · Twitter Web App

# Are we done?

# Protein Complexes/Oligomers



AlphaFold2 can also predict heterocomplexes. All you have to do is input the two sequences you want to predict and connect them with a long linker.



7:07 AM · Jul 19, 2021 · Twitter Web App



Homooligomeric prediction in `#alphafold` works a little too good. So far worked on nearly every case we (me & ~~the team~~) tried. Going beyond dimers! Seems `@deeplmind` accidentally "solved" the homooligomeric prediction problem (w/ MSA input) 😂



4:52 PM · Jul 21, 2021 · Twitter Web App

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

<sup>1</sup>Department of Biochemistry, University of Washington, Seattle, WA 98195, USA. <sup>2</sup>Institute for Protein Design, University of Washington, Seattle, WA 98195, USA. <sup>3</sup>Faculty of Arts and Sciences, Division of Science, Harvard University, Cambridge, MA 02138, USA. <sup>4</sup>John Harvard Distinguished Science Fellowship Program, Harvard University, Cambridge, MA 02138, USA. <sup>5</sup>Eugene McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, TX, USA.

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

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

First release: 15 July 2021

Article

# Highly accurate protein structure prediction for the human proteome

<https://doi.org/10.1038/s41586-021-03828-1>

Received: 11 May 2021

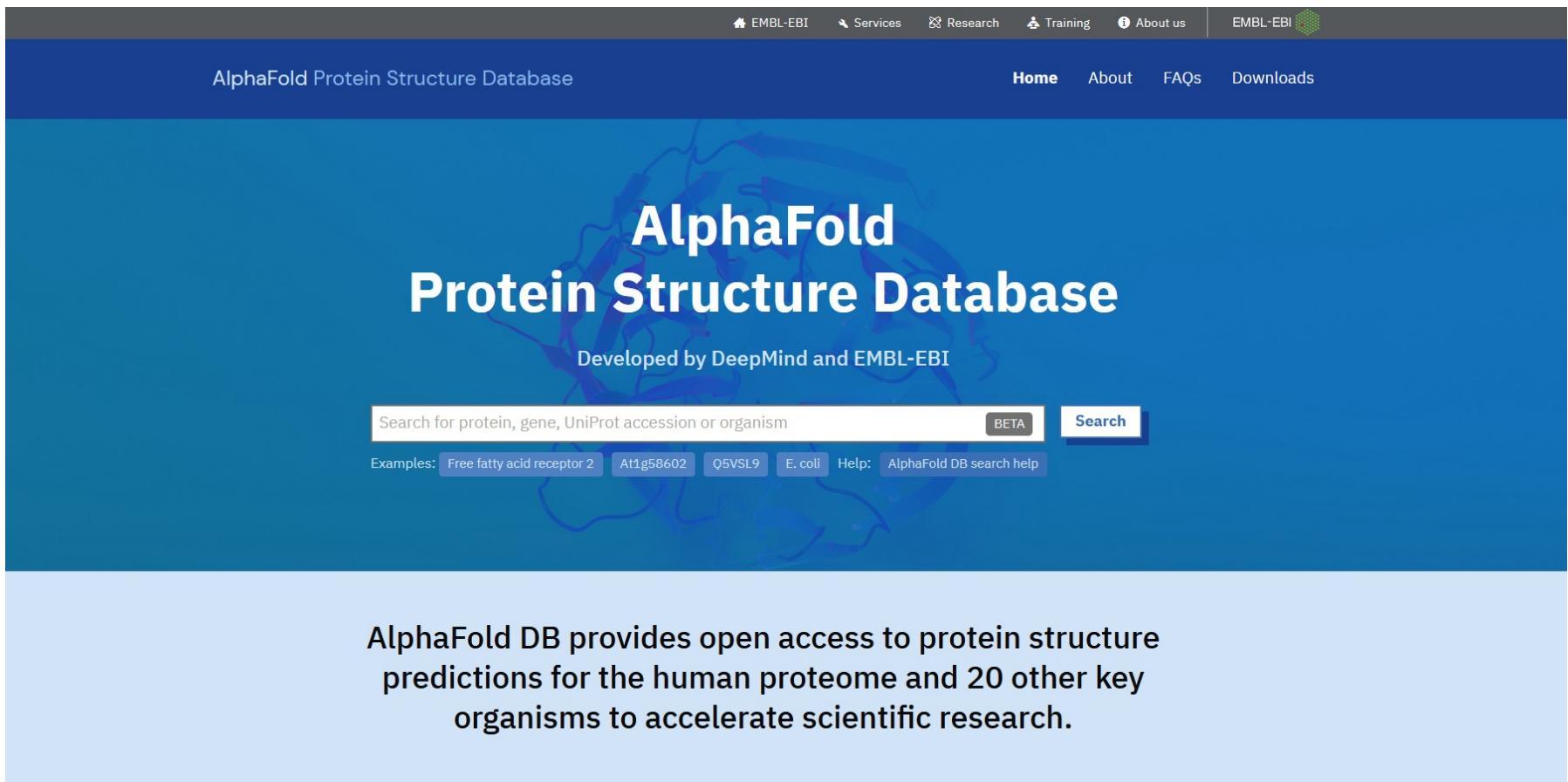
Accepted: 16 July 2021

Published online: 22 July 2021

Kathryn Tunyasuvunakool<sup>1</sup>✉, Jonas Adler<sup>1</sup>, Zachary Wu<sup>1</sup>, Tim Green<sup>1</sup>, Michal Zielinski<sup>1</sup>, Augustin Žídek<sup>1</sup>, Alex Bridgland<sup>1</sup>, Andrew Cowie<sup>1</sup>, Clemens Meyer<sup>1</sup>, Agata Laydon<sup>1</sup>, Sameer Velankar<sup>2</sup>, Gerard J. Kleywegt<sup>2</sup>, Alex Bateman<sup>2</sup>, Richard Evans<sup>1</sup>, Alexander Pritzel<sup>1</sup>, Michael Figurnov<sup>1</sup>, Olaf Ronneberger<sup>1</sup>, Russ Bates<sup>1</sup>, Simon A. A. Kohl<sup>1</sup>, Anna Potapenko<sup>1</sup>, Andrew J. Ballard<sup>1</sup>, Bernardino Romera-Paredes<sup>1</sup>, Stanislav Nikолов<sup>1</sup>, Rishabh Jain<sup>1</sup>, Ellen Clancy<sup>1</sup>, David Reiman<sup>1</sup>, Stig Petersen<sup>1</sup>, Andrew W. Senior<sup>1</sup>, Koray Kavukcuoglu<sup>1</sup>, Ewan Birney<sup>2</sup>, Pushmeet Kohli<sup>1</sup>, John Jumper<sup>1,3</sup>✉ & Demis Hassabis<sup>1,3</sup>✉

Protein structures can provide invaluable information, both for reasoning about biological processes and for enabling interventions such as structure-based drug development or targeted mutagenesis. After decades of effort, 17% of the total residues in human protein sequences are covered by an experimentally-determined structure<sup>1</sup>. Here we dramatically expand structural coverage by applying the state-of-the-art machine learning method, AlphaFold<sup>2</sup>, at scale to almost the entire human proteome (98.5% of human proteins). The resulting dataset covers 58% of residues with a confident prediction, of which a subset (36% of all residues) have very high confidence. We introduce several metrics developed by building on the AlphaFold model, and use them to interpret the dataset, identifying strong multi-domain predictions as well as regions likely to be disordered. Finally, we provide some case studies illustrating how high-quality predictions may be used to generate biological hypotheses. Importantly, we are making our predictions freely available to the community via a public database (hosted by the European Bioinformatics Institute at <https://alphafold.ebi.ac.uk/>). We anticipate that routine large-scale and high-accuracy structure prediction will become an important tool, allowing new questions to be addressed from a structural perspective.

Are we done?



The image shows the homepage of the AlphaFold Protein Structure Database. The header features a colorful abstract background with various organic shapes in green, pink, yellow, and blue. The top navigation bar includes links for EMBL-EBI, Services, Research, Training, About us, and the EMBL-EBI logo. Below the header, a secondary navigation bar has links for Home, About, FAQs, and Downloads. The main title "AlphaFold Protein Structure Database" is prominently displayed in large white font against a blue background. Below the title, it says "Developed by DeepMind and EMBL-EBI". A search bar at the bottom left contains the placeholder text "Search for protein, gene, UniProt accession or organism" and a "BETA" button next to a "Search" button. Below the search bar, there is a row of example queries: "Free fatty acid receptor 2", "At1g58602", "Q5VSL9", "E. coli", "Help:", and "AlphaFold DB search help". The main content area below the search bar contains a descriptive paragraph about the database's purpose.

AlphaFold Protein Structure Database

Home    About    FAQs    Downloads

EMBL-EBI

AlphaFold Protein Structure Database

Developed by DeepMind and EMBL-EBI

Search for protein, gene, UniProt accession or organism

BETA

Search

Examples: Free fatty acid receptor 2   At1g58602   Q5VSL9   E. coli   Help: AlphaFold DB search help

AlphaFold DB provides open access to protein structure predictions for the human proteome and 20 other key organisms to accelerate scientific research.



<https://www.alphafold.ebi.ac.uk/>

# Are we done?

RCSB PDB Deposit Search Visualize Analyze Download Learn More Documentation Careers MyPDB Contact us

**RCSB PDB** PROTEIN DATA BANK 201,196 Structures from the PDB 1,068,577 Computed Structure Models (CSM)

3D Structures Enter search term(s), Entry ID(s), or sequence Include CSM Advanced Search | Browse

PDB-101 wwPDB EMDDataResource NUCLEIC ACID DATABASE Foundation New: More Computed Structure Models

Welcome Deposit Search Visualize Analyze Download Learn

RCSB Protein Data Bank (RCSB PDB) enables breakthroughs in science and education by providing access and tools for exploration, visualization, and analysis of:

- Experimentally-determined 3D structures from the **Protein Data Bank (PDB)** archive
- Computed Structure Models (CSM)** from AlphaFold DB and ModelArchive

These data can be explored in context of external annotations providing a structural view of biology.

**COVID-19 CORONAVIRUS Resources**

**200000** Structures in the Protein Data Bank

To include/exclude Computed Structure Models (CSMs) in the search results, toggle the “Include CSM” switch on/off.

**February Molecule of the Month**

SARS-CoV-2 Nucleocapsid and Home Tests



<https://www.rcsb.org/>



# Are we done?

☰ README.md

## ColabFold - v1.5.1

```
+ 04Feb2023: v1.5.0 - ColabFold updated to use AlphaFold v2.3.1!
+ 06Feb2023: v1.5.1 - fixing --save-all/--save-recycles option
```

For details of what was changed in v1.5, see [change log!](#)



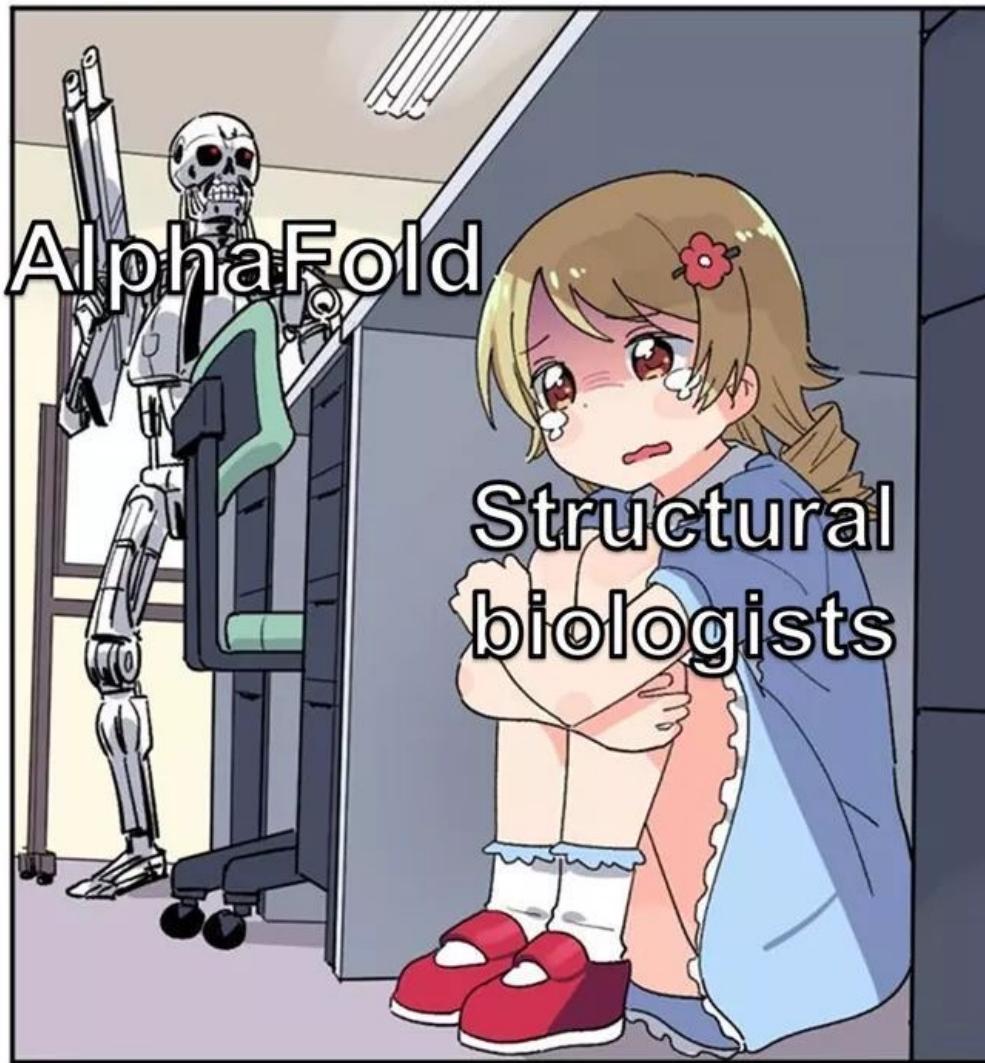
Making Protein folding accessible to all via Google Colab!

Notebooks	monomers	complexes	mmseqs2	jackhmmer	templates
<a href="#">AlphaFold2_mmseqs2</a>	Yes	Yes	Yes	No	Yes
<a href="#">AlphaFold2_batch</a>	Yes	Yes	Yes	No	Yes
<a href="#">RoseTTAFold</a>	Yes	No	Yes	No	No
<a href="#">AlphaFold2</a> (from Deepmind)	Yes	Yes	No	Yes	No
<a href="#">ESMFold</a>	Yes	Maybe	No	No	No



<https://github.com/sokrypton/ColabFold>

Are we done?



## Current limitations of the prediction method

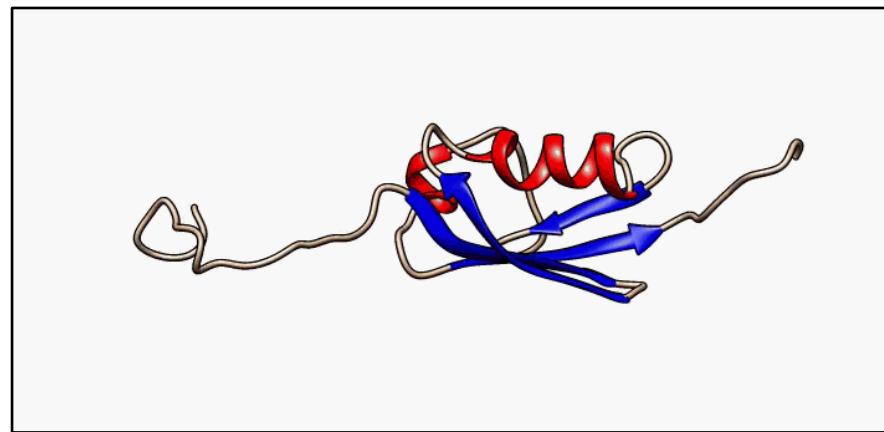
Although the availability of predicted 3D models for the known “protein universe” is an exciting prospect with huge impact, there are nevertheless limitations to the AlphaFold method and resource, some of which may be addressed in the future:

- Many proteins function as **complexes** with other proteins, nucleic acids (DNA or RNA) or ligands. AlphaFold does not currently predict 3D structures for protein-protein or protein-DNA/RNA/ligand complexes. In some cases, the single-chain prediction may correspond to the structure adopted in a complex. In other cases (especially if the protein is structured only upon binding partner molecules) the missing context from surrounding molecules may lead to an uninformative prediction.
- Proteins are dynamic systems and adopt different structures depending on their environment or state within a functional cycle. Where a protein is known to have multiple conformations AlphaFold will usually only produce one of them. This leaves open many interesting questions about the **conformational dynamics** of proteins, crucial for understanding biological function, and this will remain a very active area of research.
- For regions that are **intrinsically disordered or unstructured** in isolation, AlphaFold is expected to produce a low-confidence prediction and the predicted structure will have an extended, ribbon-like appearance. AlphaFold may be of use as a tool for identifying such regions, but the prediction makes no statement about the relative likelihood of different conformations (in biophysical terms: it is not a sample from the Boltzmann distribution). Furthermore, AlphaFold does not claim to predict the “folding pathway”.
- AlphaFold has not been trained or validated for predicting the **effect of mutations**. In particular, it is not expected to capture the effect of point mutations that destabilise a protein.
- **Ligands** are not included in the structures since AlphaFold does not make any predictions about any of the non-protein components that are often observed in experimental structures (such as cofactors, metals, ligands including drug-like molecules, ions, carbohydrates and other post-translational modifications).
- As with experimental structures, predicted structures may (or may not) lead to hypotheses about the function of the protein and the mechanism underlying that function, but such hypotheses then have to be tested by further experimentation.



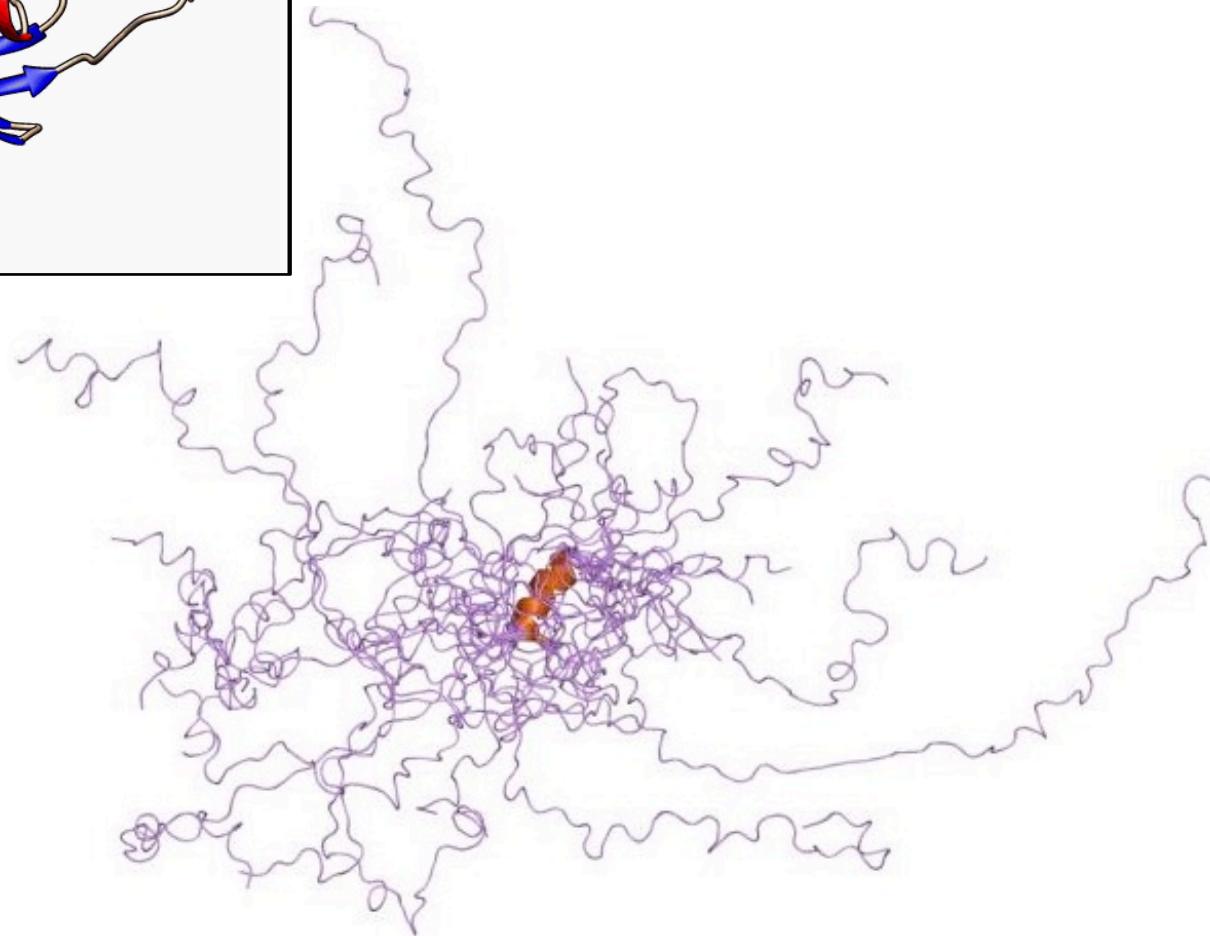


Are we done?



# IDPs

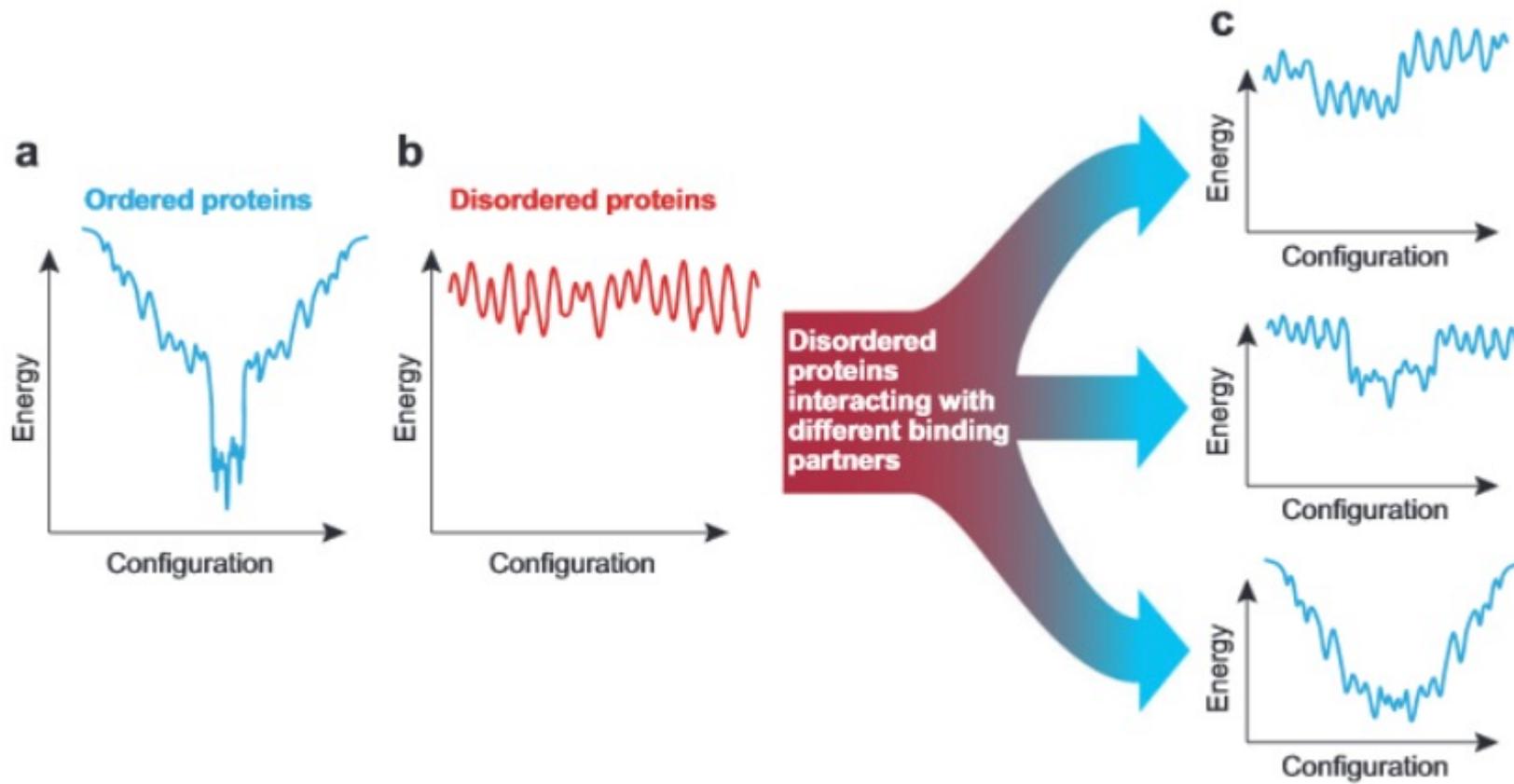
intrinsically  
disordered  
proteins





Are we done?

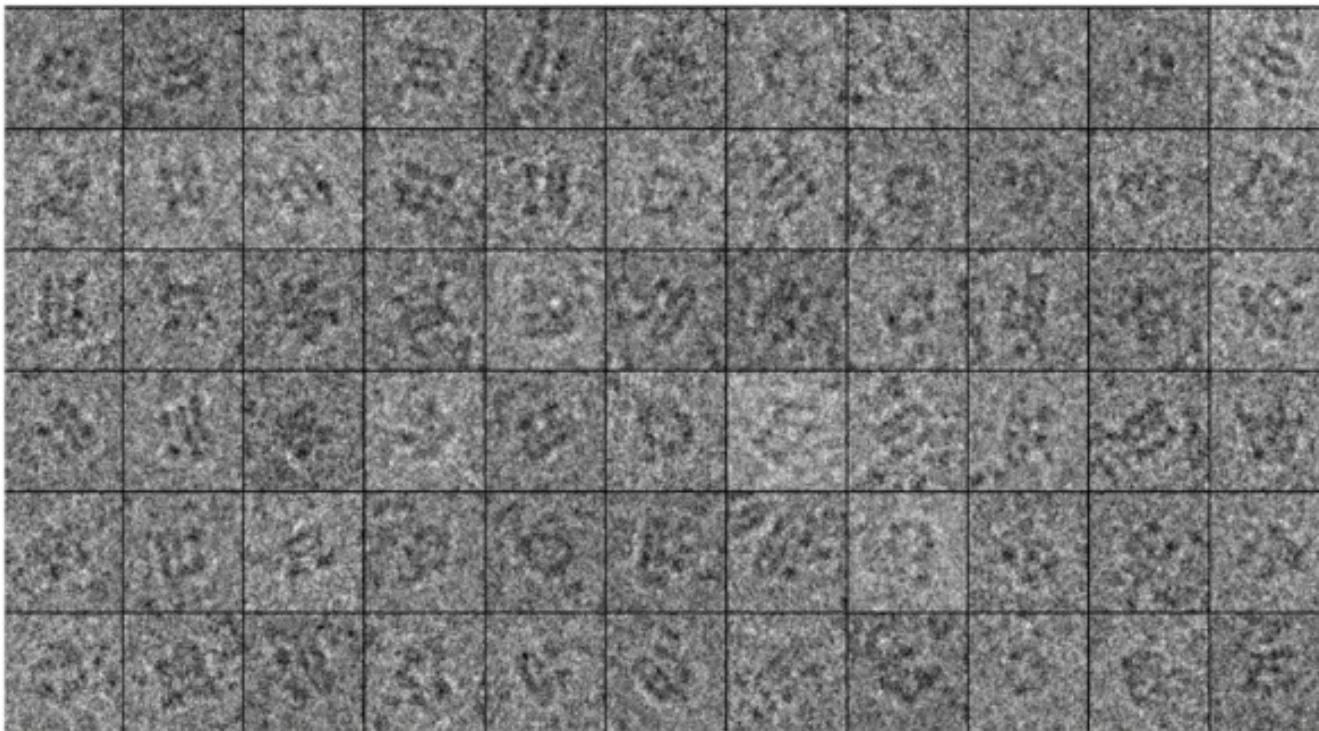
## IDPs intrinsically disordered proteins





What about the  
future?

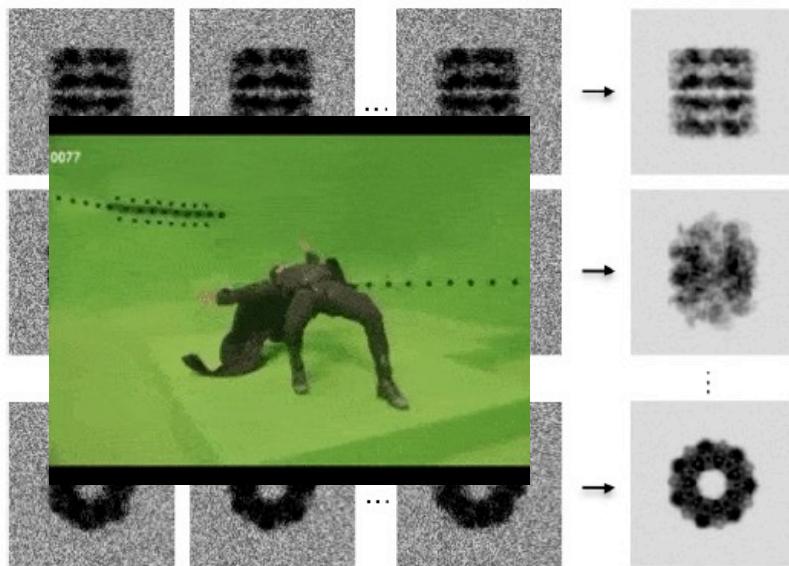
# CryoEM



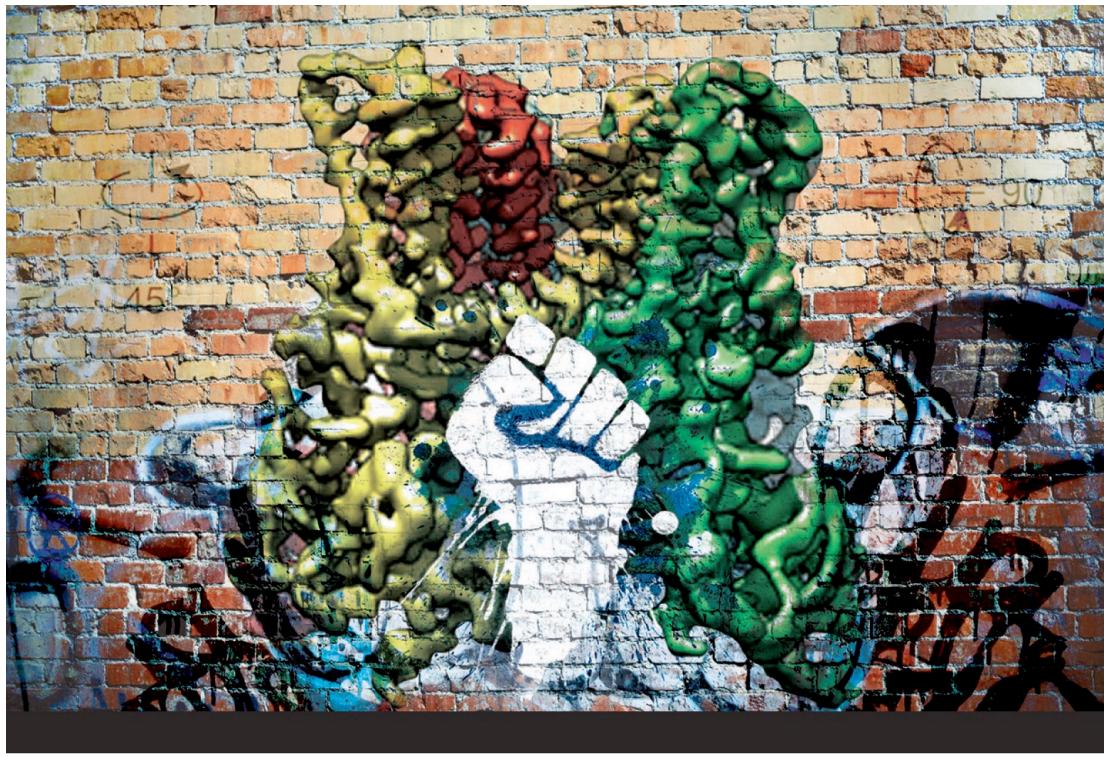


What about the  
future?

# CryoEM



Current status



**THE REVOLUTION WILL NOT BE  
CRYSTALLIZED**



Callaway, *Nature* 525, 172-174 (2015).

# Try it!

[https://colab.research.google.com/github/pablo-arantes/BIEN-249/blob/main/Lecture2\\_BIEN249.ipynb](https://colab.research.google.com/github/pablo-arantes/BIEN-249/blob/main/Lecture2_BIEN249.ipynb)

Lecture2\_BIEN249.ipynb ☆

File Edit View Insert Runtime Tools Help All changes saved

+ Code + Text

## Hello there!

This is a Jupyter notebook to predict protein structure and complex using [AlphaFold2](#) and [Alphafold2-multimer](#). Sequence alignments/templates are generated through [MMseqs2](#) and [HHsearch](#). For more details, see the instructions, checkout the [ColabFold GitHub](#) and read their manuscript.

This notebook is part of BIEN 249 class of bioengineering course at University of California, Riverside.

---

### Acknowledgments

- I would like to thank the OpenMM team for developing an excellent and open source engine.
- We would like to thank the AlphaFold team for developing an excellent model and open sourcing the software.
- [Söding Lab](#) for providing the computational resources for the MMseqs2 server
- Credit to Sergey Ovchinnikov ([@sokrypton](#)), Milot Mirdita ([@milot\\_mirdita](#)) and Martin Steinegger ([@theSteinegger](#)) for their fantastic [ColabFold](#)
- A Making-it-rain team, **Pablo R. Arantes** ([@pablitoarantes](#)), **Marcelo D. Polôto** ([@mdpoloto](#)), **Conrado Pedebos** ([@ConradoPedebos](#)) and **Rodrigo Ligabue-Braun** ([@ligabue\\_braun](#)).
- Also, credit to [David Koes](#) for his awesome [py3Dmol](#) plugin.
- For related notebooks see: [Making-it-rain](#)



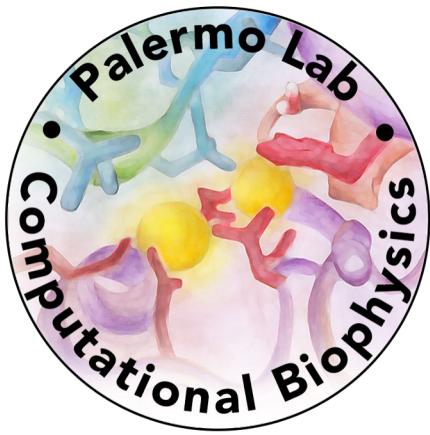
**Hen egg-white lysozyme (Monomer) =**

KVFGRCLEAAAMKRHGLNYRGYSLGNWVCAAKFESNFNTQATNRNDGSTDYGILQINSRWWCNDGRTPGSRNLCNIPCSALLSSDITAS  
VNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL

**tRNA-specific adenosine deaminase (Complex) =**

MSEVEFSHEYWMRHALTLAKRARDEREVPVGAVLVLNNRVIGEGWNRAIGLHDPTAHAEIMALRQGGLVMQNYRLYDATLYSTFEPCVMC  
AGAMIHSRIGRVVFGRVNAKTGAAGSLMDVLHHPGMNHRVEITEGILADECACALLCRFFRMPRRVFNAQKKAQSSTD:MSEVEFSHEYWM  
RHALTLAKRARDEREVPVGAVLVLNNRVIGEGWNRAIGLHDPTAHAEIMALRQGGLVMQNYRLYDATLYSTFEPCVMCAGAMIHSRIGRVV  
FGVRNAKTGAAGSLMDVLHHPGMNHRVEITEGILADECACALLCRFFRMPRRVFNAQKKAQSSTD

# Acknowledgements



<https://palermolab.com/>

Department of Bioengineering  
University of California Riverside  
223 Materials Science & Engineering  
900 University Ave. | Riverside, CA 92521

## Acknowledgements

website

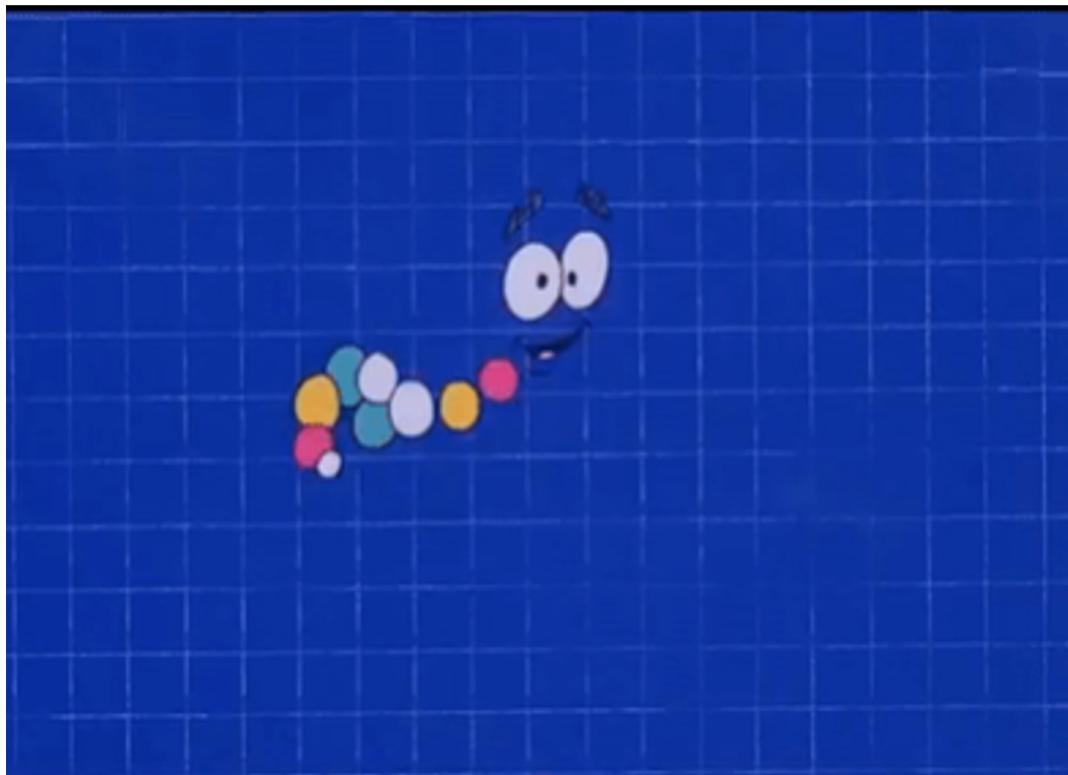


SCAN ME

pablitoarantes@gmail.com

pabloa@ucr.edu

<https://pablo-arantes.github.io>



twitter



SCAN ME

# Obrigado (Thank you)!





• Cristina Spano

# Google It!

AlphaFold and its impact on  
obtaining protein structures

Pablo Ricardo Arantes, PhD.