

# Repliement des protéines

## avec la révolution d'AlphaFold2

Dirk Stratmann ([dirk.stratmann@upmc.fr](mailto:dirk.stratmann@upmc.fr))  
<http://www.imPMC.upmc.fr/~stratmann/>

IMPMC, Sorbonne Université

octobre 2022

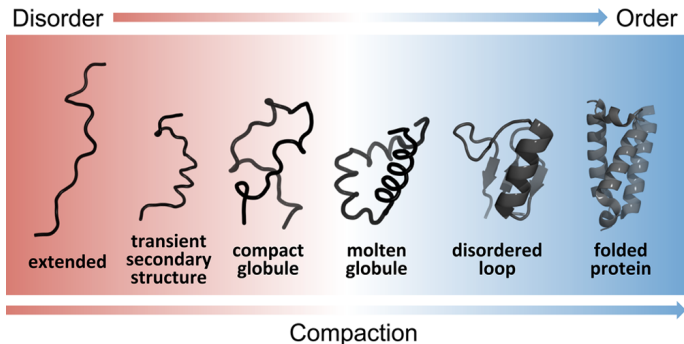
# Plan

- 1 Processus du repliement
- 2 Techniques expérimentales
- 3 Repliement par dynamique moléculaire
- 4 Repliement in vivo
- 5 Protein design
- 6 Alpha Fold
- 7 Bibliographie

- 1 Processus du repliement
  - Processus du repliement
  - État déplié

# Processus du repliement

# Repliement



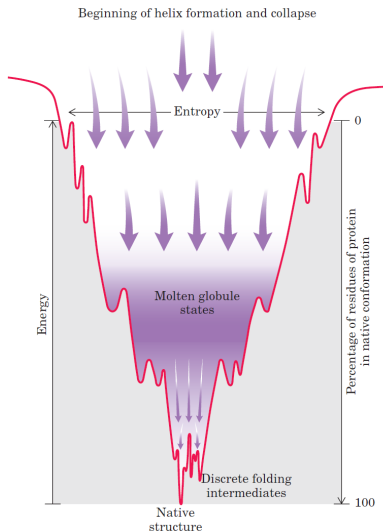
- Molten globule: interactions hydrophobes
- Structure 3D finale: liaisons hydrogènes

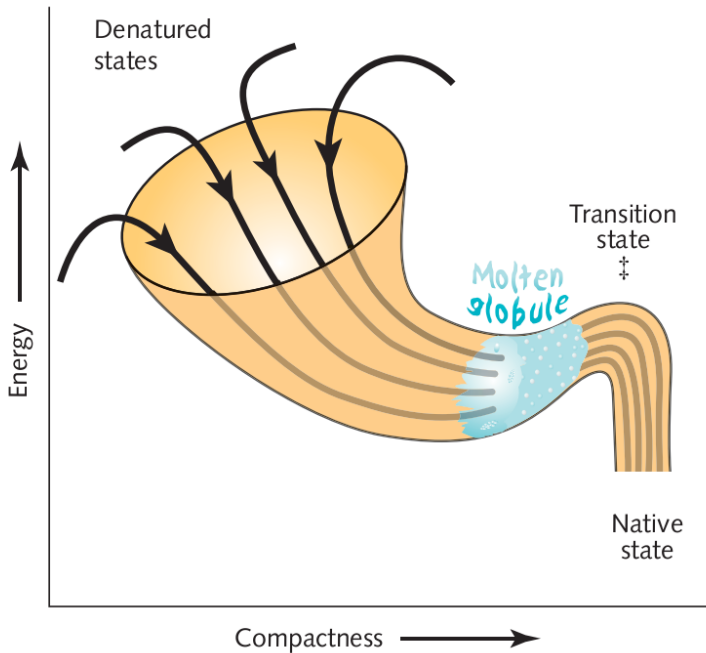
## Pourquoi est le repliement des protéines si rapide ?

Paradoxe de Levinthal (1968):

- chaque résidu n'a que deux possibilité de conformation
- une protéine de 100 résidu aurait  $2^{100} \approx 10^{30}$  conformations possibles.
- Conclusion: une protéine ne peut pas se replier par une recherche au hasard de la conformation native
- Elle doit suivre un chemin de repliement (*folding pathway*) plus efficace.

# Entonnoir du repliement







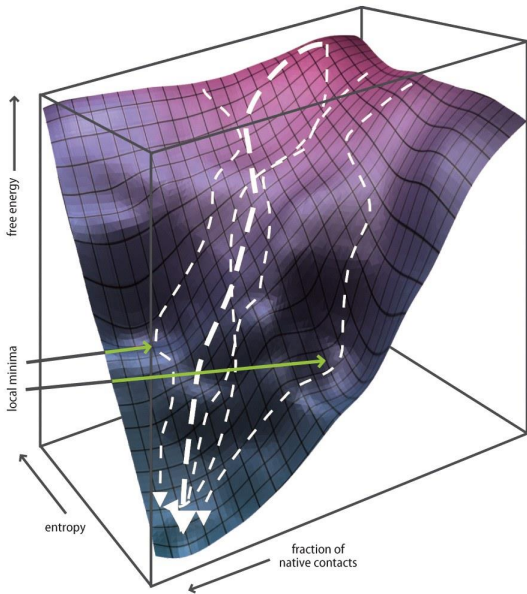
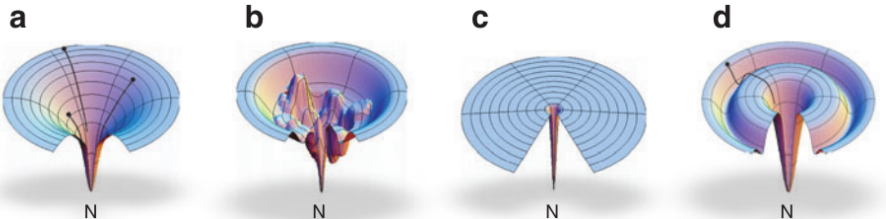


Figure 6.15 How Proteins Work (©2012 Garland Science)

## Entonnoir du repliement



a: fast folder

b: kinetic traps

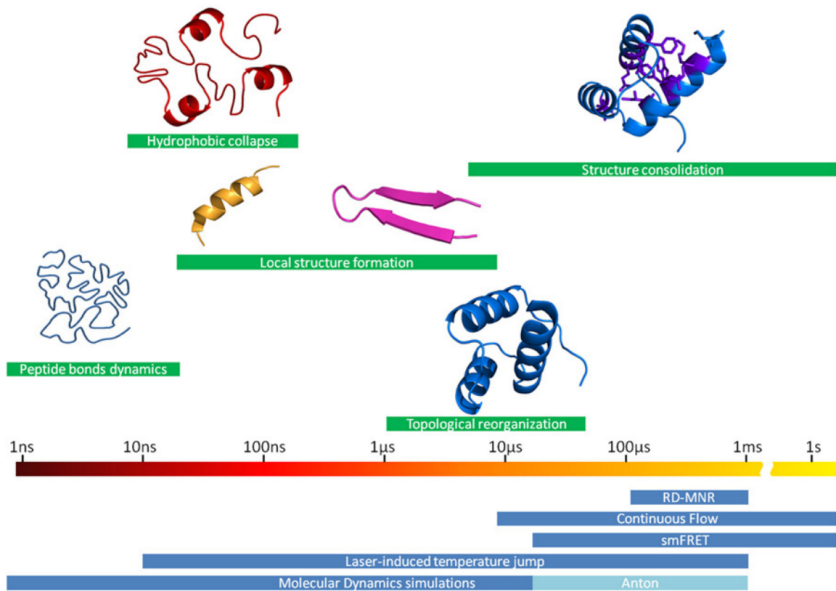
c: diffusional conformational search

d: obligatory intermediate

Ken A. Dill et al. (June 2008). en. In: *Annual Review of Biophysics* 37.1

## Vitesses du repliement

- Hélices se forment plus rapidement que des feuillet
- Hélices  $\Leftrightarrow$  contacts proches dans la séquence
- Feuillet  $\Leftrightarrow$  contacts distants dans la séquence
- Hélices: 0.1 - 1  $\mu\text{s}$ ,  $\beta$ -hairpins: 1 - 10  $\mu\text{s}$
- La vitesse 'limite' du repliement:  $N/50 \mu\text{s}$ , N: nombre de résidus



État déplié

État déplié

## Rayon de giration

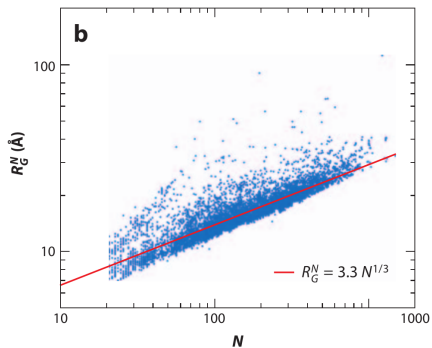
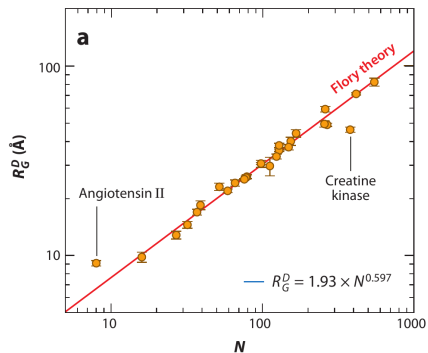
Protein	Number of residues	$R_G$ (native)	$R_G$ (denatured)	Ratio
PI3 kinase, SH3 domain	90	18.6	27.5	0.7
Horse heart cytochrome c	104	17.8	32.6	0.5
Hen egg white lysozyme	129	20.5	34.6	0.6
Yeast triose phosphate isomerase	247	29.7	49.7	0.6

- Si centre de gravité à l'origine, alors:

$$R_G = \sqrt{\frac{\sum_i m_i (x_i^2 + y_i^2 + z_i^2)}{\sum_i m_i}}$$

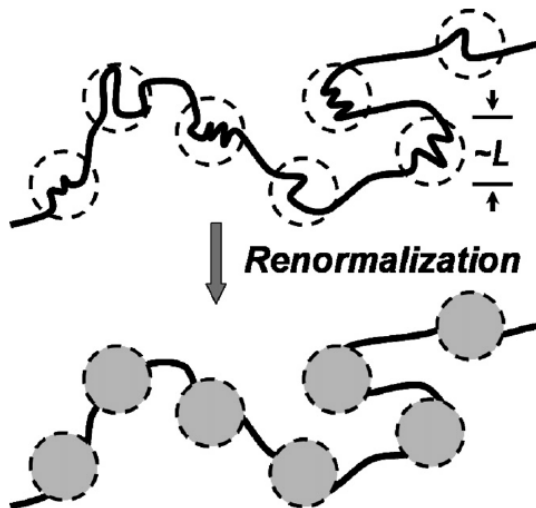
- Random coil:  $R_G \propto N^{0.6}$  avec N nombre de résidus
- Globule compacte:  $R_G \propto N^{0.33}$

## Rayon de giration



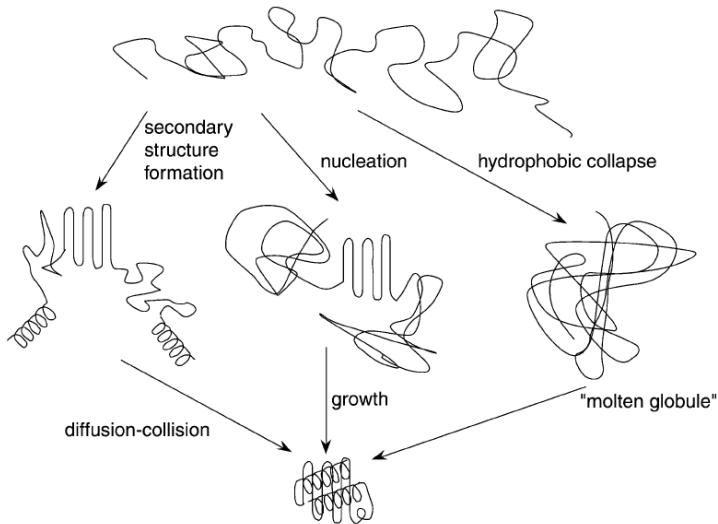
D. Thirumalai et al. (2010). In: *Annual Review of Biophysics* 39.1

# Protéine dépliée = polymer sans interaction





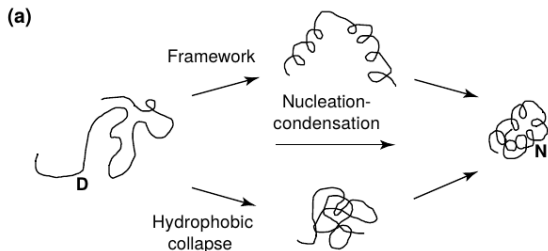
# Modèles classiques pour le repliement



à gauche: "framework" ou "hierarchical",

Alan R. Fersht and Valerie Daggett (2002). In: *Cell*

# Modèles classiques pour le repliement



(b)

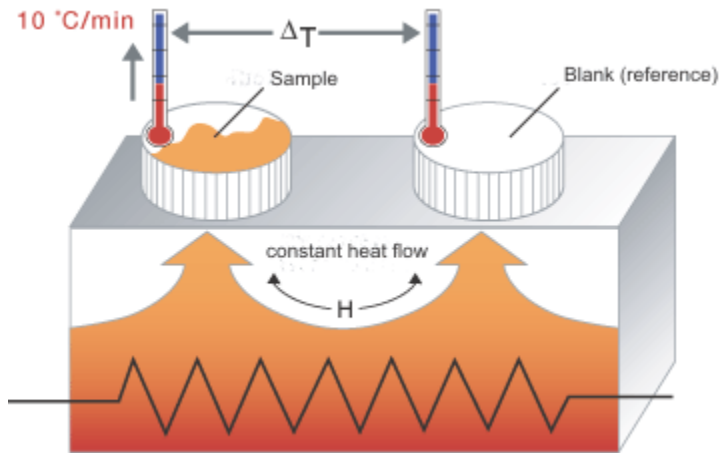
Framework	engrailed homeodomain ( $\alpha$ ) protein A ( $\alpha$ )	} Barnase ( $\alpha/\beta$ ) Hen egg white lysozyme ( $\alpha/\beta$ )
Nucleation- condensation	Cl2 ( $\alpha/\beta$ ) cMyb ( $\alpha$ ) tenascin ( $\beta$ ) FKBP12 ( $\alpha/\beta$ )	
Hydrophobic collapse	?	

## 2 Techniques expérimentales

- Méthodes thermodynamiques
- Techniques sur molécule unique
- RMN
- Méthodes pour cinétique "ultra rapide"

# Méthodes thermodynamiques

## Differential scanning calorimetry (DSC)

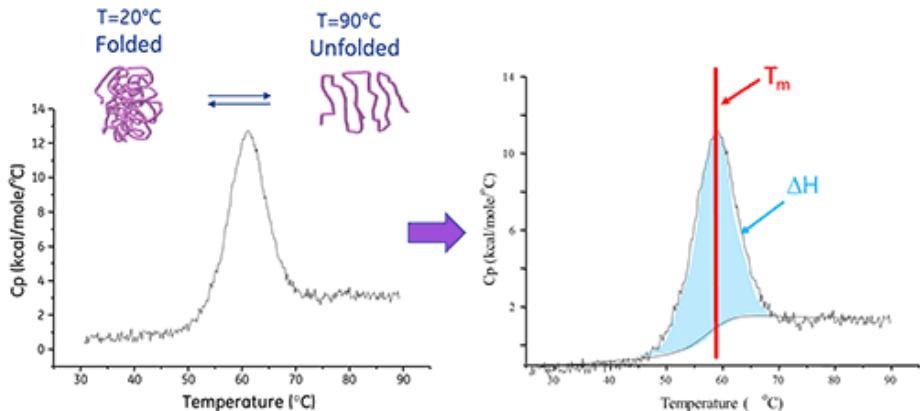


Principe: référence et échantillon sont chauffés simultanément

Protéine absorbe de la chaleur lors du dépliement =>  $\Delta T$

[www.itc.tu-bs.de/Abteilungen/Makro/Methods/dsc.htm](http://www.itc.tu-bs.de/Abteilungen/Makro/Methods/dsc.htm)

# Differential scanning calorimetry (DSC)



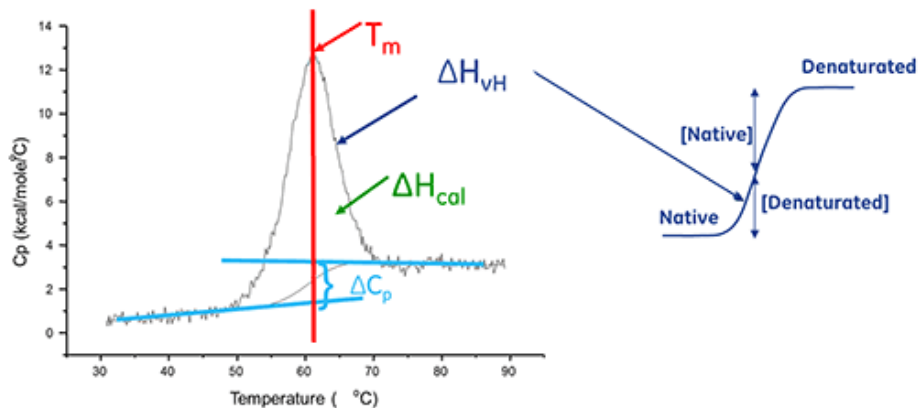
$T_m$ : température médiane de transition thermique (thermal transition midpoint)

$C_p$ : capacité calorifique (heat capacity)

H: enthalpie

[www.malvern.com/fr/products/technology/differential-scanning-calorimetry](http://www.malvern.com/fr/products/technology/differential-scanning-calorimetry)

# Differential scanning calorimetry (DSC)



$T_m$ : température médiane de transition thermique (thermal transition midpoint)

$C_p$ : capacité calorifique (heat capacity)

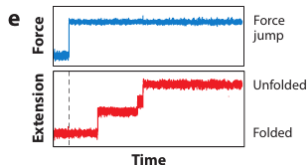
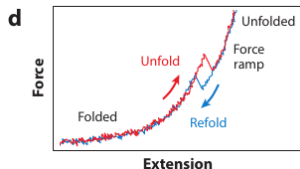
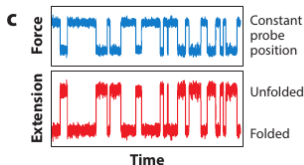
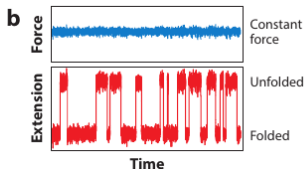
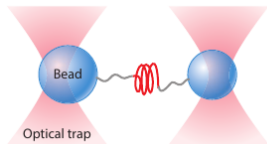
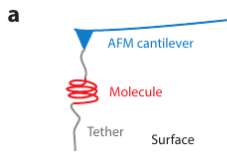
H: enthalpie

[www.malvern.com/fr/products/technology/differential-scanning-calorimetry](http://www.malvern.com/fr/products/technology/differential-scanning-calorimetry)

# Techniques sur molécule unique



# Single molecule force spectroscopy (SMFS)



a) Atomic force microscopy (AFM) and Optical tweezers

b) Constant force mode: extension fluctuates

c) Constant probe position: force and extension fluctuate

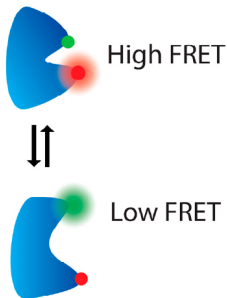
d) Force-ramp mode: elastic stretching is interrupted by a "rip", hysteresis indicates a nonequilibrium process

e) Force-jump mode: extension changes in steps

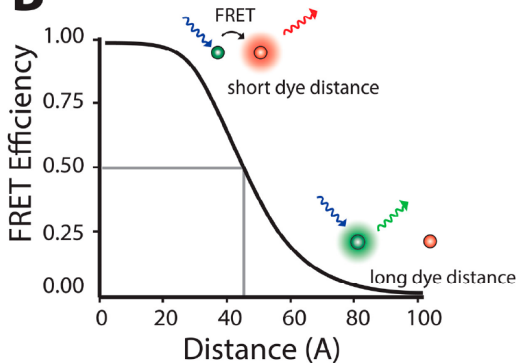
Michael T. Woodside and Steven M. Block (2014). In: *Annual Review of Biophysics* 43.1

# Fluorescence resonance energy transfer (FRET)

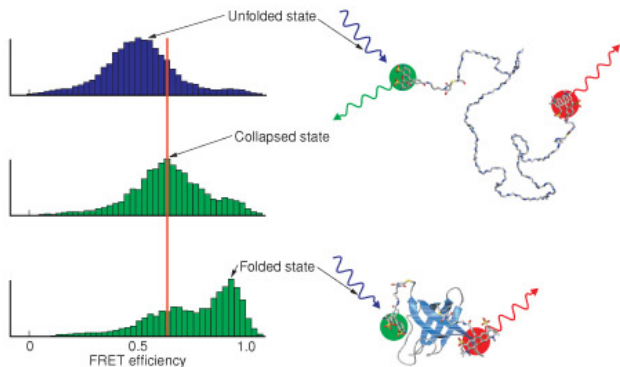
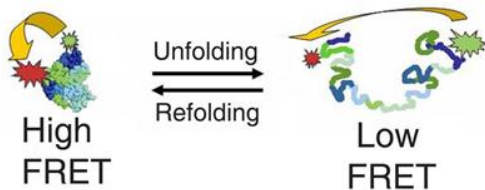
**A**



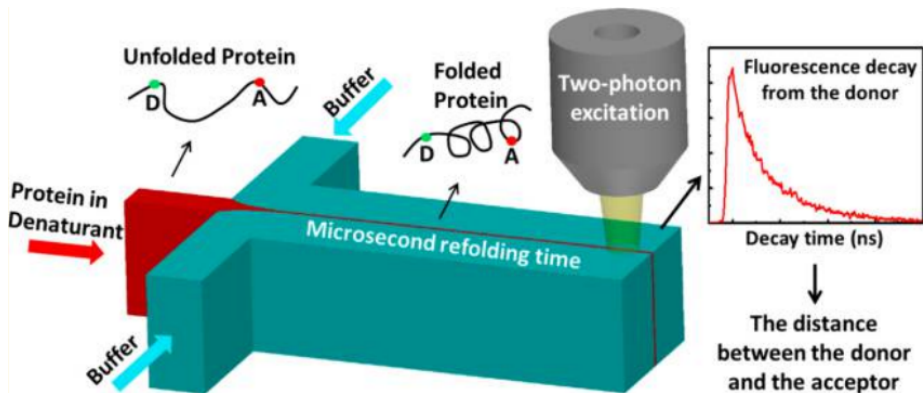
**B**



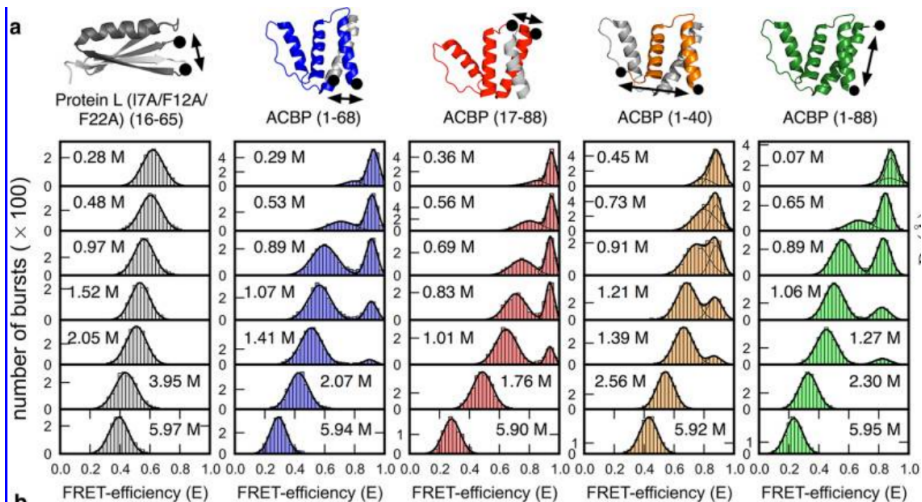
# Fluorescence resonance energy transfer (FRET)



# Time-resolved FRET (TR-FRET)

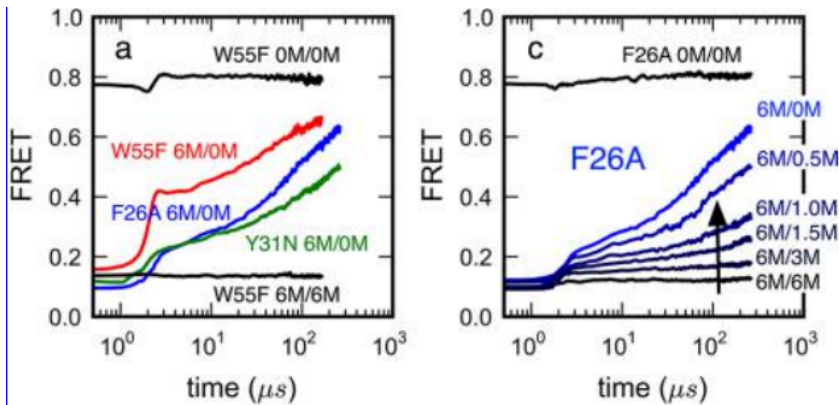


## FRET sur ACBP



## FRET sur ACBP - Collapse très lent

Solution dénaturant (6 M GuHCl) vers solution de repliement (0 M GuHCl):

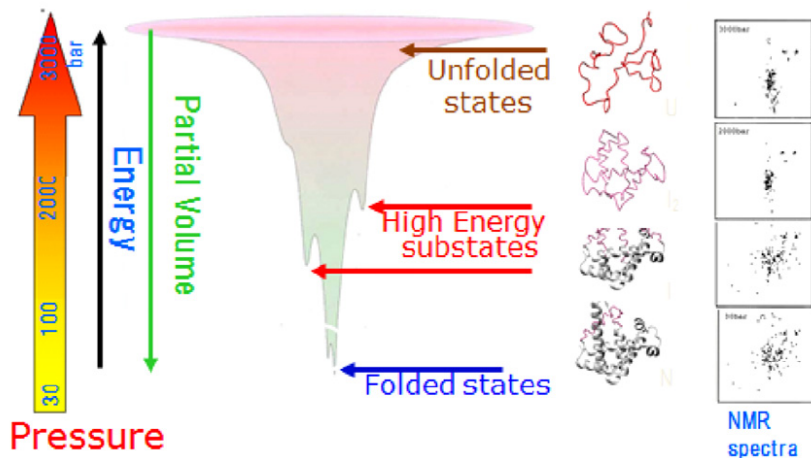


Indication pour: Formation d'une ensemble compact et hétérogènes de structures dépliées

RMN

RMN

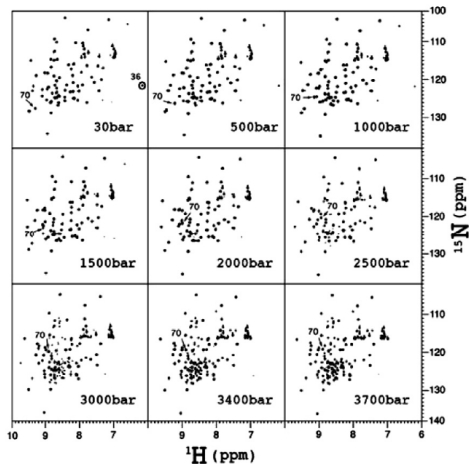
# Pression et RMN



Kazuyuki Akasaka, Ryo Kitahara, and Yuji O. Kamatari (Mar. 2013). In: *Archives of Biochemistry and Biophysics*. Protein Folding and Stability 531.1–2

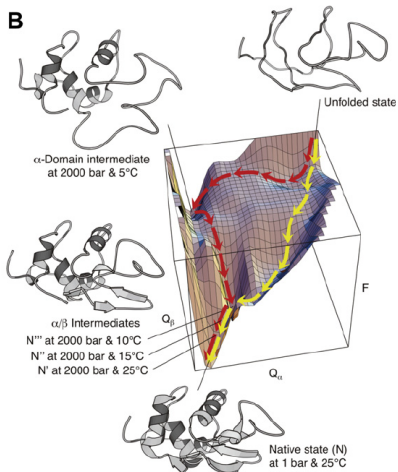


## Pression et RMN



Kazuyuki Akasaka, Ryo Kitahara, and Yuji O. Kamatari (Mar. 2013). In: *Archives of Biochemistry and Biophysics*. Protein Folding and Stability 531.1–2

# Pression et RMN

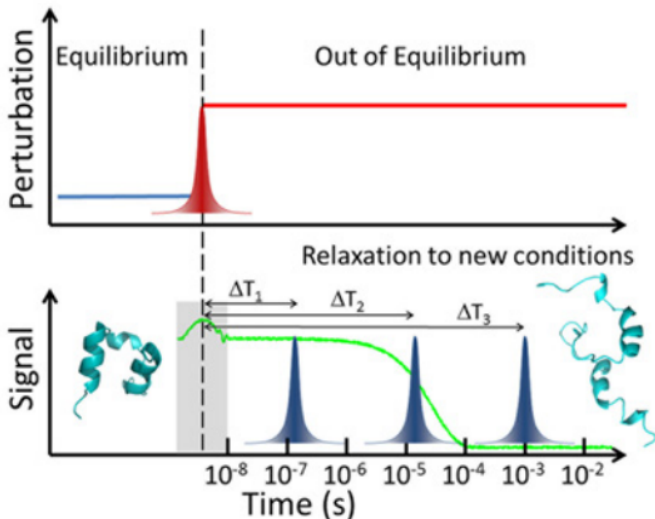


Kazuyuki Akasaka, Ryo Kitahara, and Yuji O. Kamatari (Mar. 2013). In: *Archives of Biochemistry and Biophysics*. Protein Folding and Stability 531.1–2

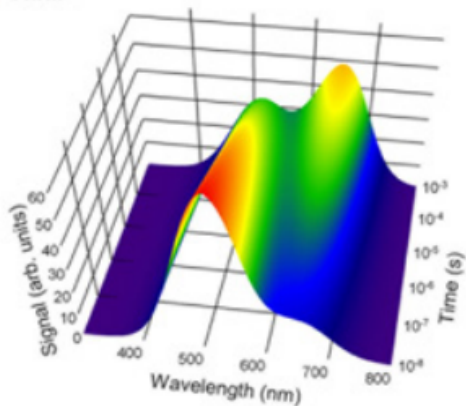
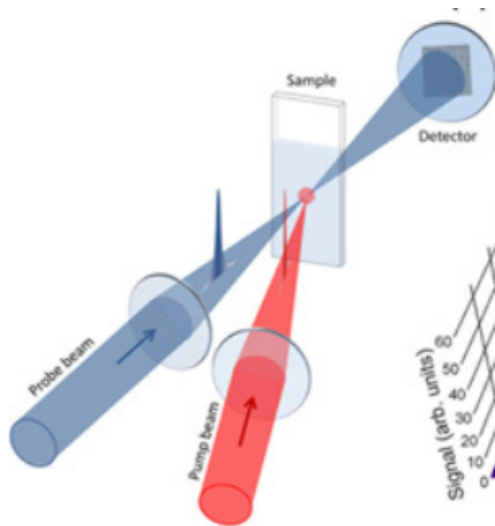
# Méthodes pour cinétique "ultra rapide"

# Laser T-jump

## Ultrafast Kinetic Perturbation Methods



## Laser T-jump



### 3 Repliement par dynamique moléculaire

- Introduction
- Temps de repliement
- Markov State Models (MSM)
- Etat déplié
- Thermodynamique du repliement

# Introduction

## Anton

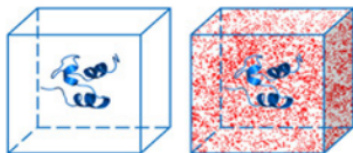
## Molecular Dynamics Simulations

## A. Force Field

$$U = \sum k_b (r - r_0)^2 + \sum k_\theta (\theta - \theta_0)^2 + \sum A [1 + \cos(nT - \phi)] + \sum \sum q_i q_j / r_{ij} + \sum \sum B \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^6 \right]$$

bond stretching
bending
torsional
electrostatics
van der Waals

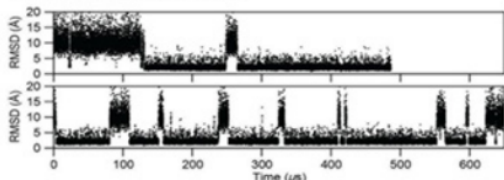
## B. Simulation System



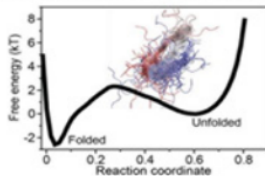
## C. Specialized High Performance Computing



## D. Molecular Trajectories



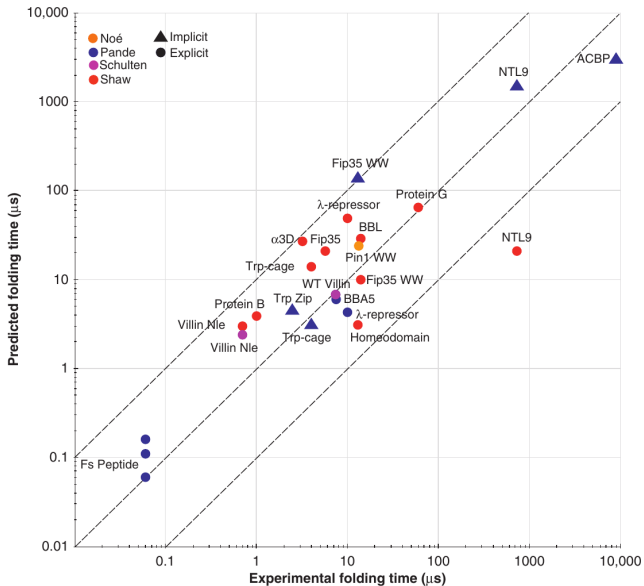
## E. Free Energy Surface calculation



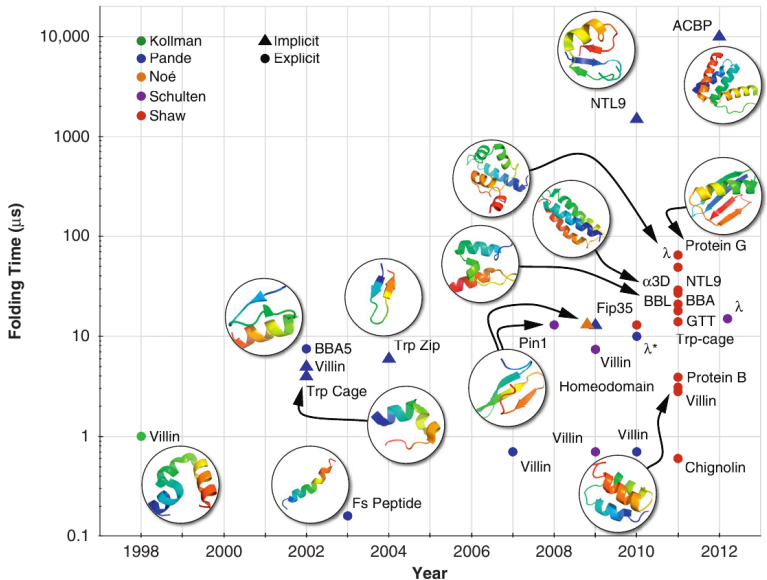


Temps de repliement

Temps de repliement

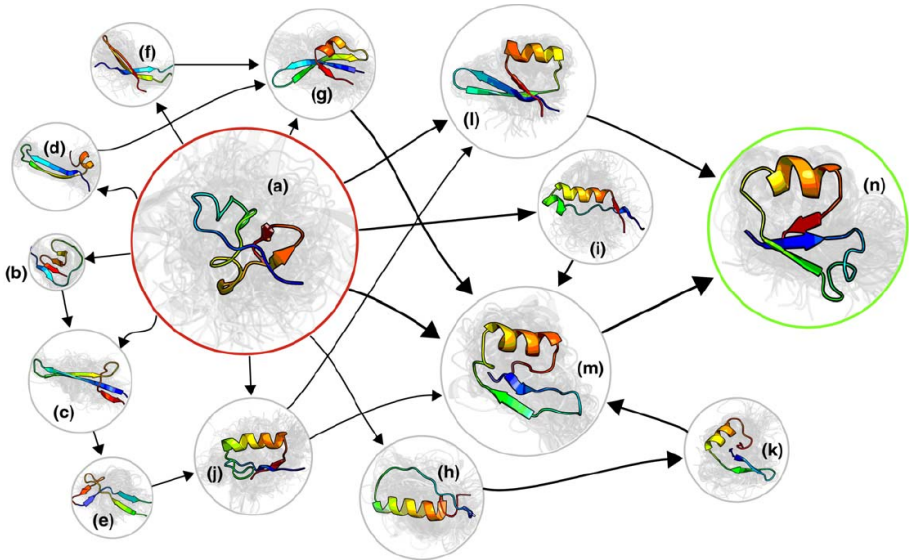


Current Opinion in Structural Biology



Current Opinion in Structural Biology

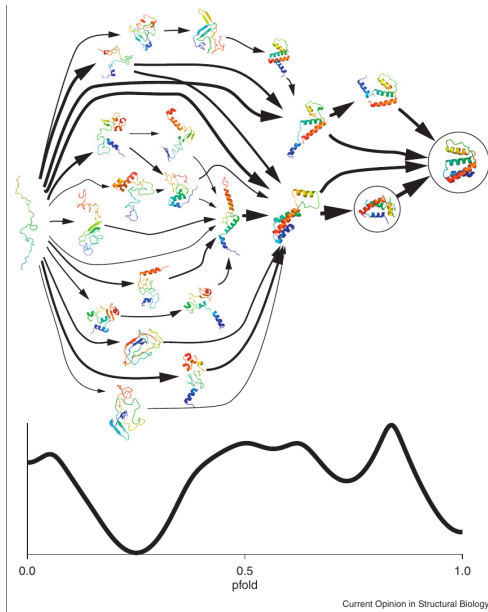
# Markov State Models (MSM)



Current Opinion in Structural Biology

protéine NTL9,  
*Structural Biology* 21.1

Gregory R Bowman, Vincent A Voelz, and Vijay S Pande (Feb. 2011). en. In: *Current Opinion in*



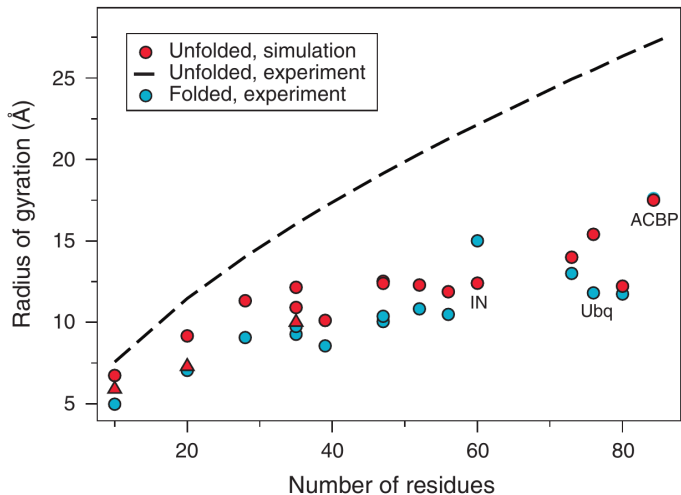
protéine ACBP,

Thomas J Lane et al. (Feb. 2013). en. In: *Current Opinion in Structural Biology* 23.1

Etat déplié

Etat déplié

## Etat déplié: difficile en MD

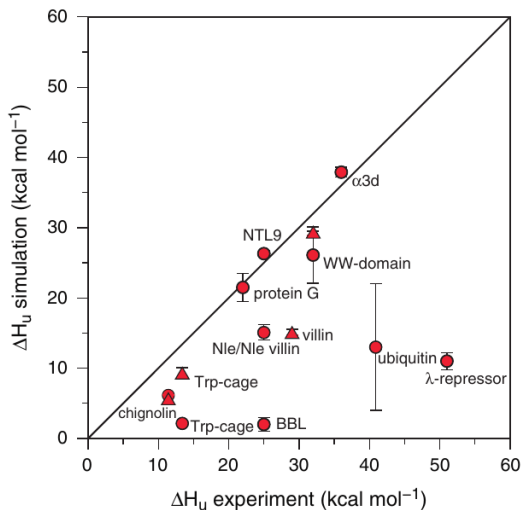


Current Opinion in Structural Biology



# Thermodynamique du repliement

## Prédiction de l'enthalpie du dépliement avec MD



Current Opinion in Structural Biology

## Conclusions de Piana et al.

Peuvent les champs de forces en dynamique moléculaire reproduire les données expérimentales du repliement ?

Oui pour :

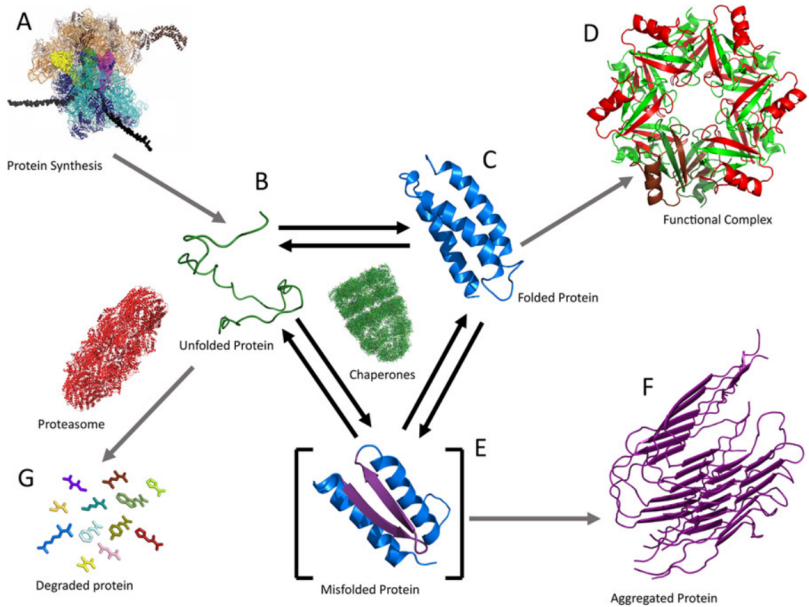
- Structure native, repliée
- Taux de repliement

Non pour:

- Cinétique détaillée
- Structures dépliées
- Enthalpie de l'état repliée plus bas qu'en expérimental

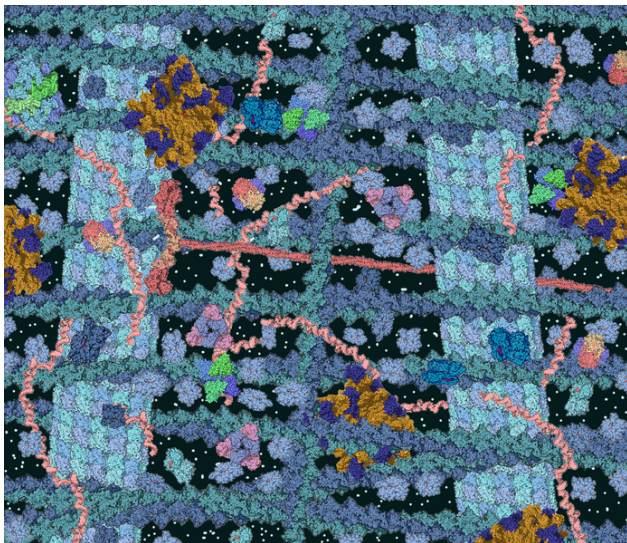
- 4 Repliment in vivo
  - Introduction
  - Environnement cellulaire

# Introduction

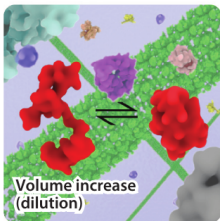
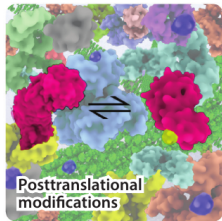
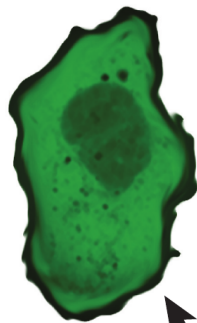
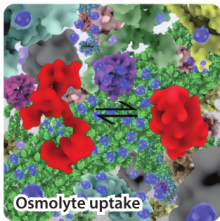
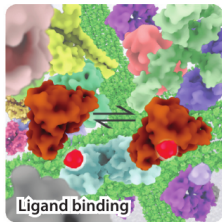
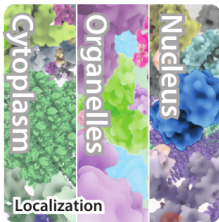
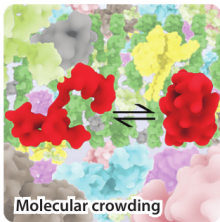


# Environnement cellulaire

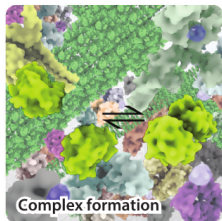
# Molecular crowding







External conditions  
(e.g., temperature,  
pressure, and  
osmolarity)



5

## Protein design

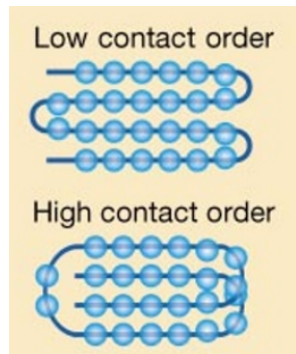
- David Baker - first work
- David Baker - 15 years of protein design
- Peptides cycliques ciblant des interactions protéine-protéine

# David Baker - first work

# Contact order and folding kinetics

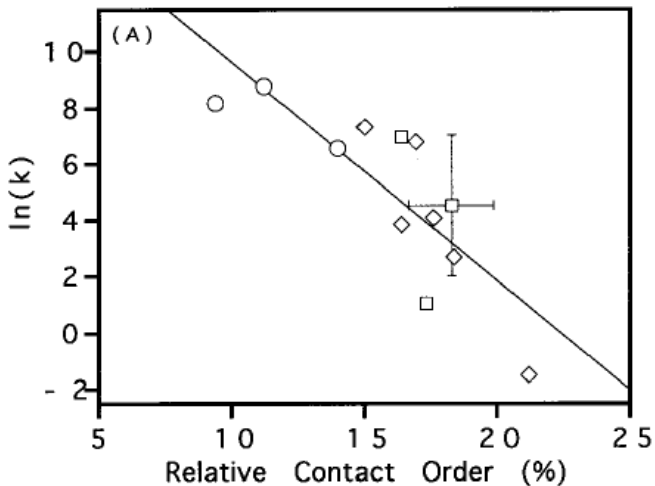
$$CO = \frac{1}{L \cdot N} \sum^N \Delta S_{i,j} \quad (1)$$

where  $N$  is the total number of contacts,  $\Delta S_{i,j}$  is the sequence separation, in residues, between contacting residues  $i$  and  $j$ , and  $L$  is the total number of residues in the protein.

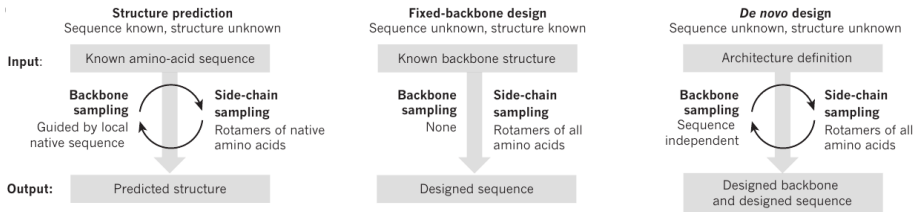


Kevin W Plaxco, Kim T Simons, and David Baker (Apr. 1998). In: *Journal of Molecular Biology* 277.4

## Contact order and folding kinetics

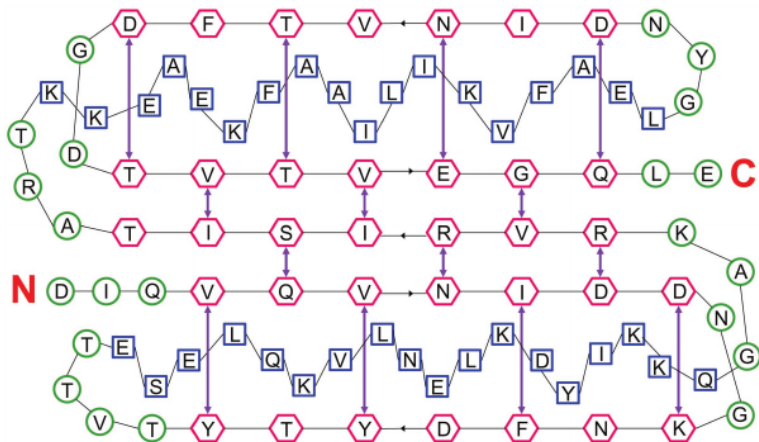


# De novo protein design



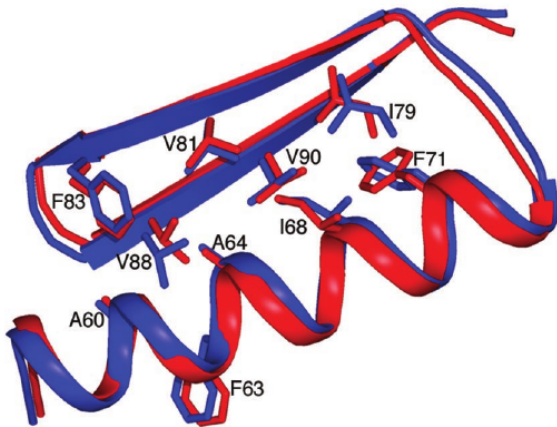
Po-Ssu Huang, Scott E. Boyken, and David Baker (Sept. 2016). en. In: *Nature* 537.7620

# Protein design - first success story



**Fig. 1.** A two-dimensional schematic of the target fold (hexagon, strand; square, helix; circle, other). Hydrogen bond partners are shown as purple arrows. The amino acids shown are those in the final designed (Top7) sequence.

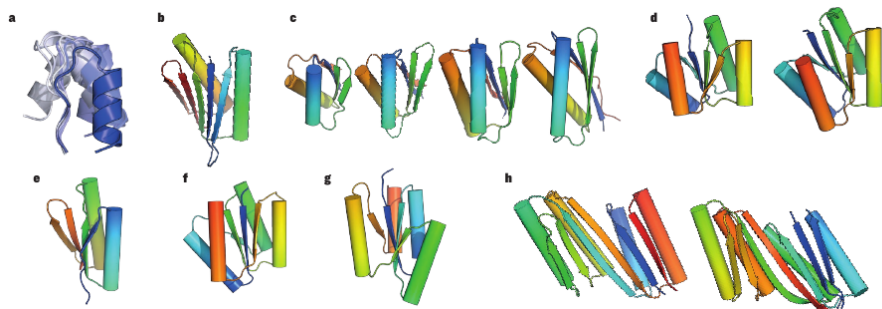
## Protein design - first success story





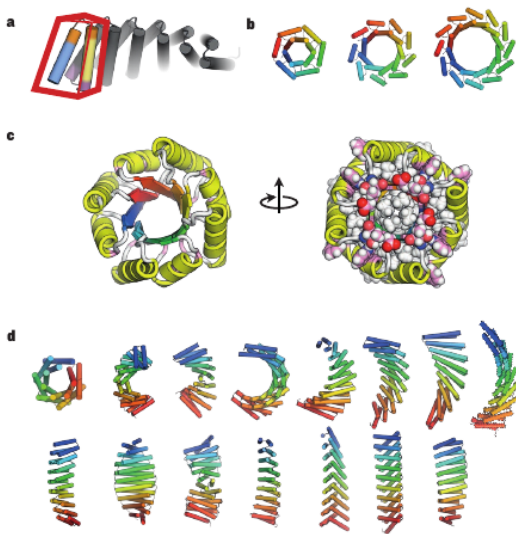
# David Baker - 15 years of protein design

## Protein design - examples

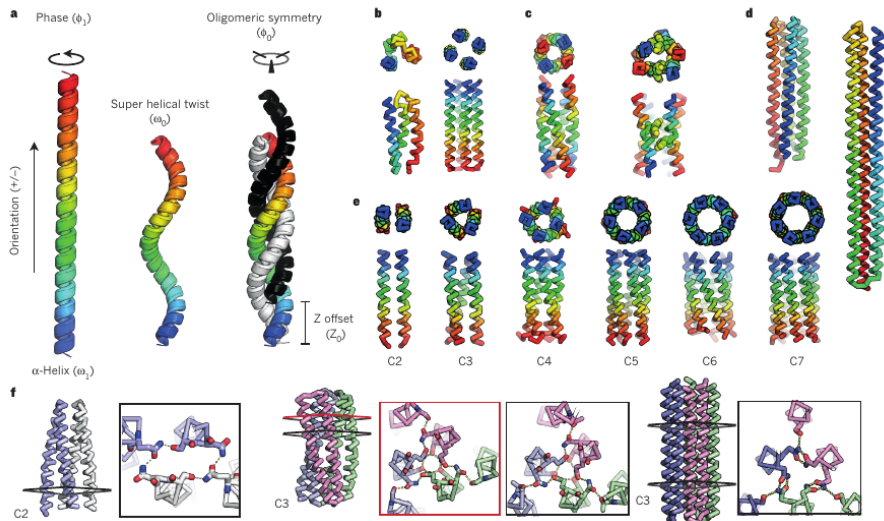


Po-Ssu Huang, Scott E. Boyken, and David Baker (Sept. 2016). en. In: *Nature* 537.7620

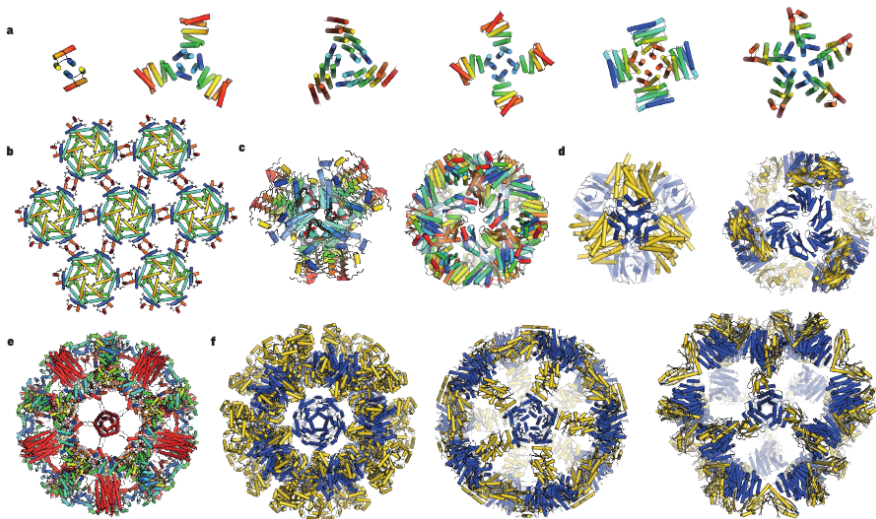
## Protein design - examples



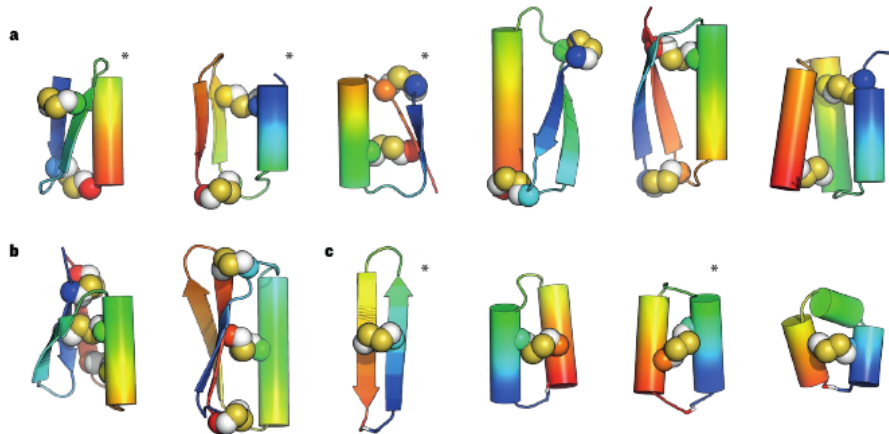
# Protein design - examples



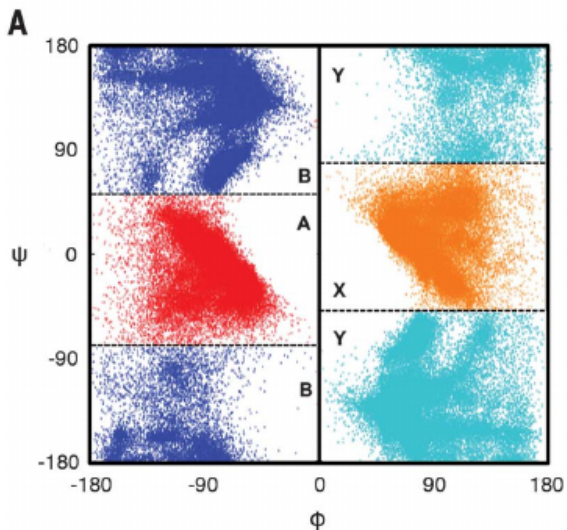
## Protein design - examples



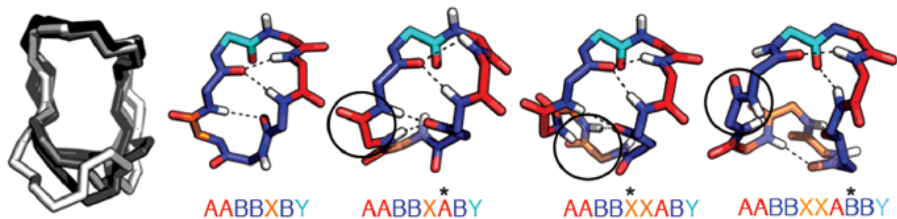
## Protein design - examples



## Protein design - macrocycles



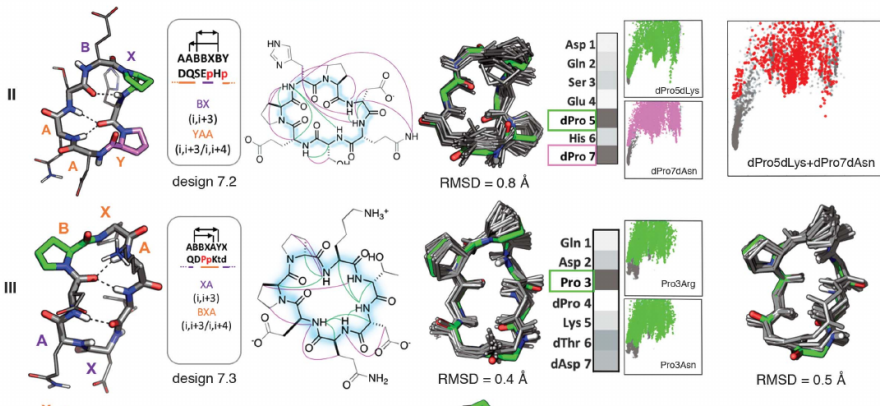
## Protein design - macrocycles



Parisa Hosseinzadeh et al. (Dec. 2017). en. In: *Science* 358.6369

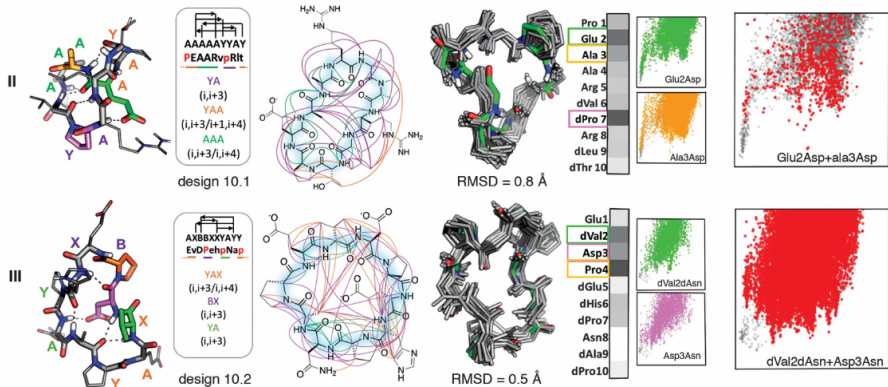


# Protein design - macrocycles



Parisa Hosseinzadeh et al. (Dec. 2017). en. In: *Science* 358.6369

# Protein design - macrocycles

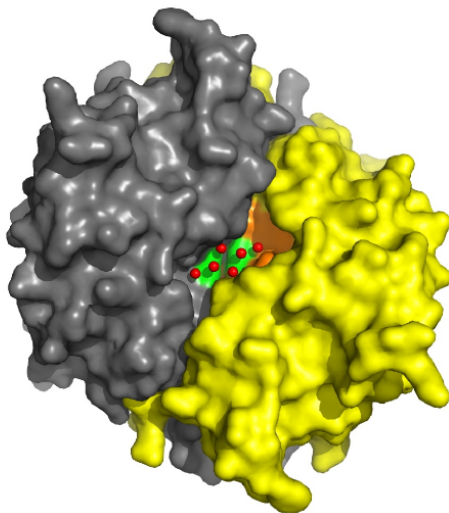


Parisa Hosseinzadeh et al. (Dec. 2017). en. In: *Science* 358.6369

# Peptides cycliques ciblant des interactions protéine-protéine

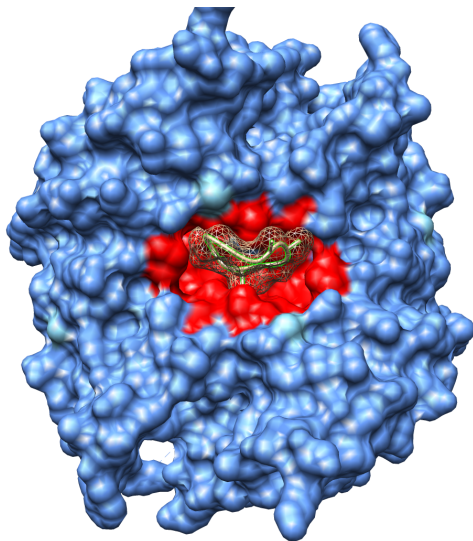
Peptides cycliques ciblant des interactions protéine-protéine

## Drug design: cibler les interactions protéine-protéine

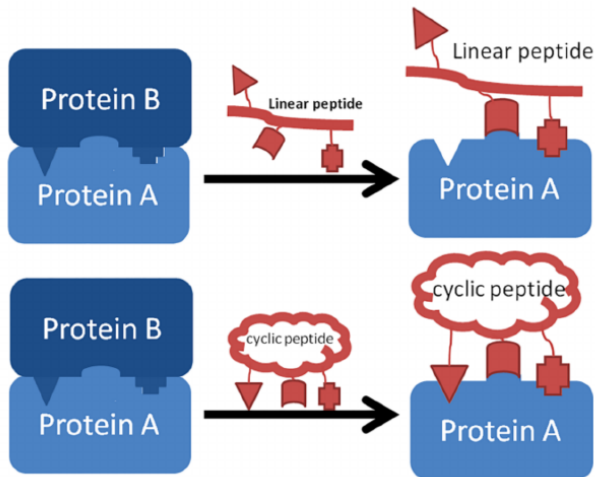


Peptides cycliques ciblant des interactions protéine-protéine

Drug = peptide cyclique



## Drug = peptide cyclique



6

## Alpha Fold

- Alpha Fold - La révolution
- Applications de Alpha Fold
- Autres méthodes utilisant le deep learning
- Comprendre Alpha Fold
- Mutations corrélées, MSA
- Utiliser Alpha Fold
- Pour aller plus loin: EMBL webinar
- L'avenir selon DeepMind

# Alpha Fold - La révolution



## Le press release de CASP du 30 nov 2020

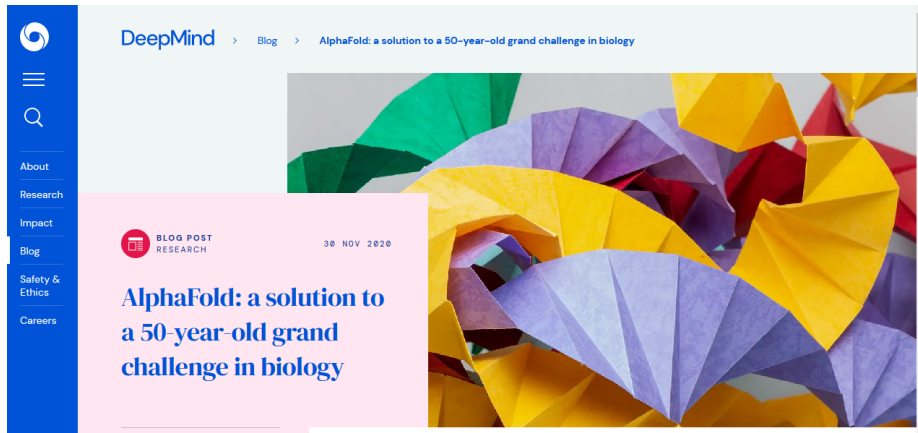
"Artificial intelligence solution to a 50-year-old science challenge could 'revolutionise' medical research"

"Today (Monday) researchers at the 14th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP14) will announce that an artificial intelligence (AI) solution to the challenge has been found."

"Nearly 50 years ago, Christian Anfinsen was awarded a Nobel Prize for showing that it should be possible to determine the shape of proteins based on their sequence of amino acids – the individual building blocks that make up proteins. That's why our community of scientists have been working on the biennial CASP challenge."

[https://predictioncenter.org/casp14/doc/CASP14\\_press\\_release.html](https://predictioncenter.org/casp14/doc/CASP14_press_release.html)

# DeepMind de google basé à Londres



The image shows a screenshot of a web page from DeepMind. On the left is a blue navigation sidebar with icons for a home page, a menu, a search function, and links to 'About', 'Research', 'Impact', 'Blog', 'Safety & Ethics', and 'Careers'. The main content area has a light blue header with the DeepMind logo and breadcrumb navigation: 'DeepMind > Blog > AlphaFold: a solution to a 50-year-old grand challenge in biology'. Below the header is a large, colorful background image of folded paper structures in shades of green, purple, yellow, and red. A pink overlay box on the left side of the image contains the following text: 'BLOG POST RESEARCH' with a red icon, the date '30 NOV 2020', and the main title 'AlphaFold: a solution to a 50-year-old grand challenge in biology' in a large, bold, blue font.

## Le relais immédiat dans Nature de l'annonce de CASP

30 nov 2020, Nature, "It will change everything': DeepMind's AI makes gigantic leap in solving protein structures"

<https://www.nature.com/articles/d41586-020-03348-4>

## La presse

30 nov 2020, The New York Times, London AI claims breakthrough that could accelerate drug discovery

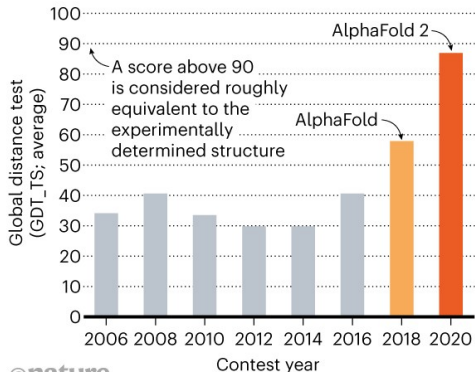
1 déc 2020, France Culture, Le "problème de repliement des protéines" résolu par une intelligence artificielle

11 déc 2020, Les Échos, DeepMind met les chercheurs du monde entier au tapis

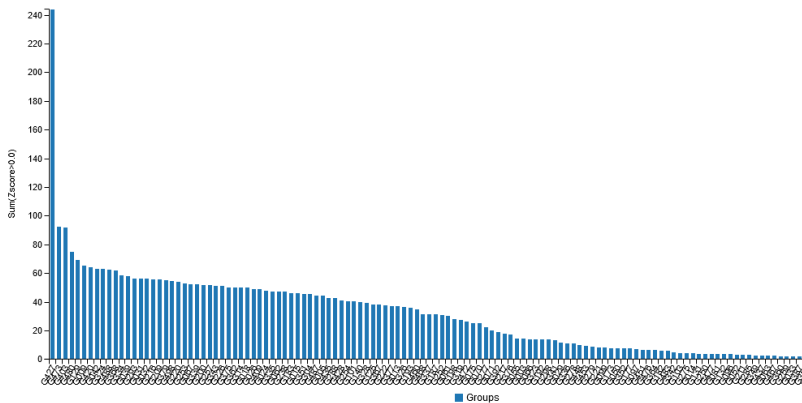
# CASP 14 et AlphaFold 2

## 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.

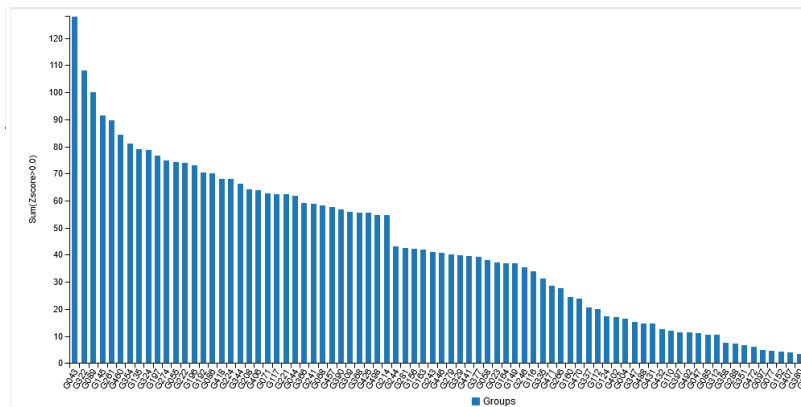


## CASP 14 et AlphaFold 2



#	◆ GR code	◆ GR name	◆ Domains Count	◆ SUM Zscore (>-2.0)	◆ Rank SUM Zscore (>-2.0)	◆ AVG Zscore (>-2.0)	◆ Rank AVG Zscore (>-2.0)	▼ SUM Zscore (>0.0)	▲ Rank SUM Zscore (>0.0)
1	427	AlphaFold2	92	244.0217	1	2.6524	1	244.0217	1
2	473	BAKER	92	90.8241	2	0.9872	2	92.1241	2
3	403	BAKER-experimental	92	88.9672	3	0.9670	3	91.4731	3
4	480	FEIG-R2	92	72.5351	4	0.7884	4	74.5627	4
5	129	Zhang	92	67.9065	5	0.7381	5	68.8922	5

# CASP 13 et AlphaFold 1



#	GR code	GR name	Domains Count	SUM Zscore (>-2.0)	Rank SUM Zscore (>-2.0)	AVG Zscore (>-2.0)	Rank AVG Zscore (>-2.0)	SUM Zscore (>0.0)	Rank SUM Zscore (>0.0)
1	043	A7D	104	120.4307	1	1.1580	1	128.0693	1
2	322	Zhang	104	107.5948	2	1.0346	2	108.1948	2
3	089	MULTICOM	104	99.4661	3	0.9564	3	99.9886	3
4	145	QUARK	104	90.9915	4	0.8749	4	91.5625	4
5	261	Zhang-Server	104	88.9540	5	0.8553	5	89.7597	5

## La presse en 2021

**3 oct 2021, Forbes: "AlphaFold is the most important achievement in AI - Ever"**

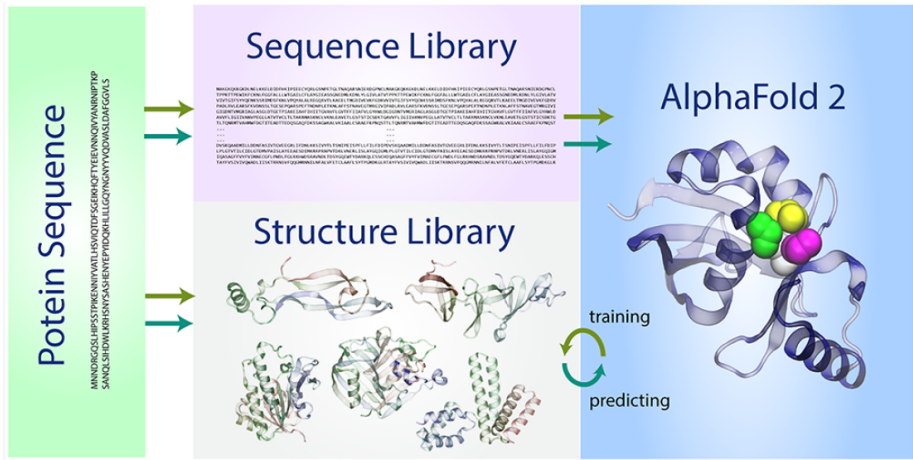
<https://www.forbes.com/sites/robtoews/2021/10/03/alphafold-is-the-most-important-achievement-in-ai-ever/?sh=2857ff656e0a>

**22 juillet 2021, Fortune: "In giant leap for biology, DeepMind's A.I. reveals secret building blocks of human life"**

<https://fortune.com/2021/07/22/deepmind-alphafold-human-proteome-database-proteins/>



# Principe de Alpha Fold 2



# 15 juillet 2021: Détails de la méthode et code source

## 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

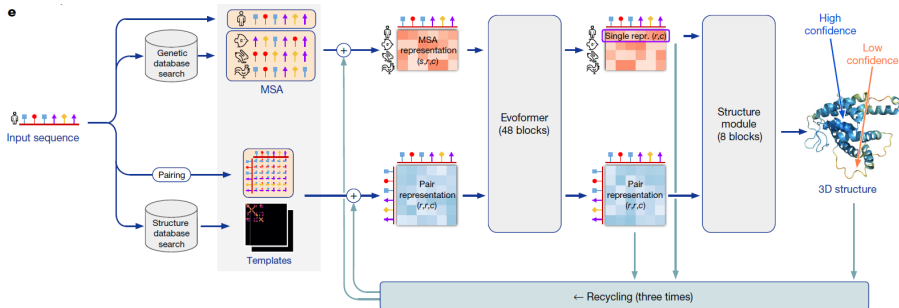
Open access

 Check for updates

John Jumper<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 Židek<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>, Rishub 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> 

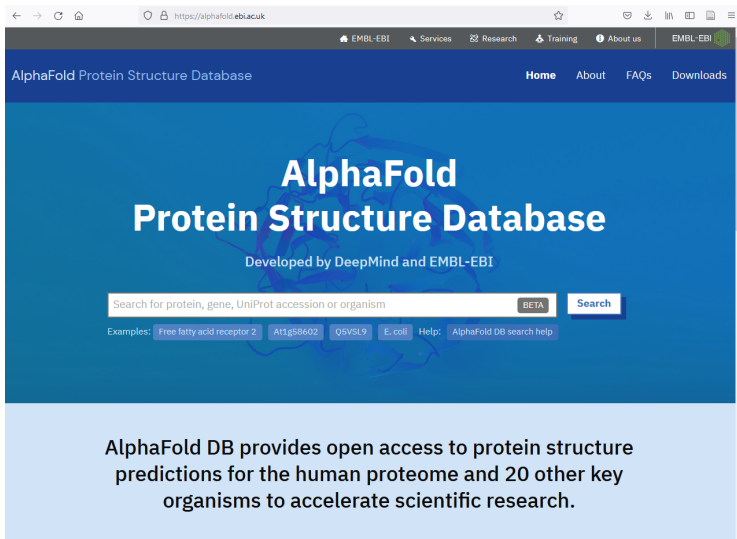
+5000 citations sur google scholar à ce jour (6 oct 2022) !

## Principe de Alpha Fold 2



# Applications de Alpha Fold

# Alpha Fold DB



AlphaFold Protein Structure Database

Home About FAQs Downloads

## AlphaFold Protein Structure Database

Developed by DeepMind and EMBL-EBI

Search for protein, gene, UniProt accession or organism BETA Search

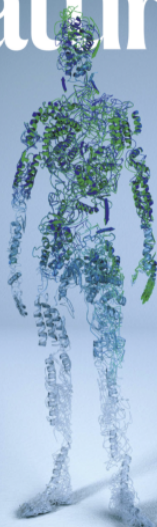
Examples: [Free fatty acid receptor 2](#) [AT1g58602](#) [Q5VSL9](#) [E. coli](#) Help: [AlphaFold DB search help](#)

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

The international journal of science / 26 August 2021

outlook  
Sickle-cell  
disease

# nature



## PROTEIN POWER

AI network predicts highly accurate 3D structures for the human proteome

**Troubled waters**  
The race to save the Great Barrier Reef from climate change

**Coronavirus**  
Time is running out to find the origins of SARS-CoV-2

**Storage hunting**  
Quantifying carbon held in Africa's montane forests

# Autres méthodes utilisant le deep learning

## Le pionnier: RaptorX de Jinbo Xu et al.

Serveur RaptorX:

<http://raptorx.uchicago.edu/>

CASP 14: score 38 vs 244 pour AlphaFold 2

"This server was officially ranked 1st in contact prediction in both CASP12 and CASP13 and initiated the revolution of protein structure prediction by deep learning."

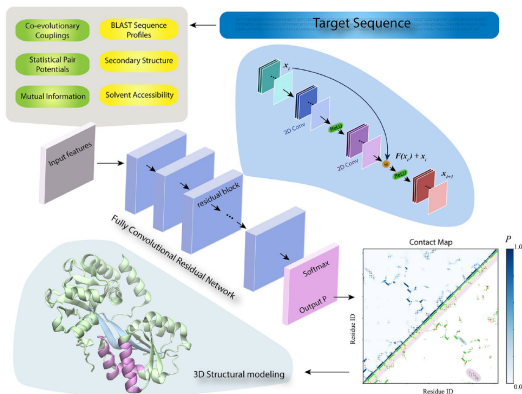


# DESTINI de Jeffrey Skolnick et al.

Serveur DESTINI:

<https://sites.gatech.edu/cssb/destini/>

CASP 14: score 29 vs 244 pour AlphaFold 2



# Le rattrapage de David Baker



Search...

[News](#) [Research](#) [People](#) [Audacious Project](#) [Get Involved](#) [Impact & Outreach](#) [In the Media](#) [About](#)

JULY 15, 2021



RoseTTAFold: Accurate protein structure prediction accessible to all

# Le rattrapage de David Baker

Science

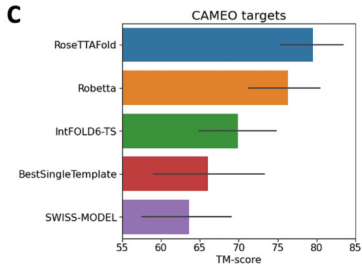
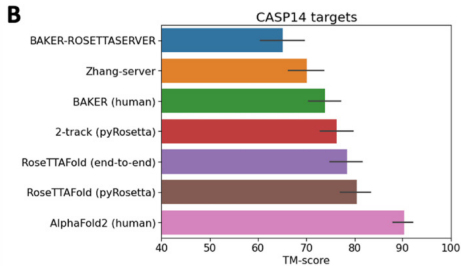
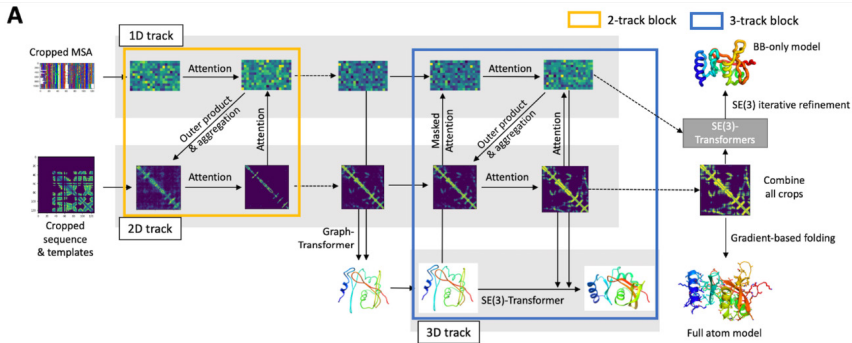
RESEARCH ARTICLES

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>

publié le même jour que le Nature sur AlphaFold2 (15 juillet 2021) !  
mais "que" 1120 citations sur google scholar à ce jour (6 oct 2022)



# Next frontier: Oligomères

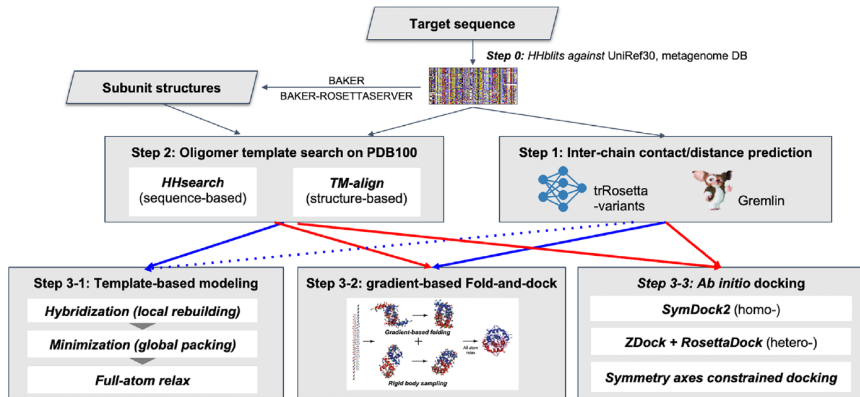


FIGURE 1 The oligomer structure modeling procedure used by the BAKER-experimental group

## RosettaFold: github et serveur

<https://github.com/RosettaCommons/RoseTTAFold>

<https://rosetta.bakerlab.org/>

**voir aussi ici:**

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

# Comprendre Alpha Fold

## Le blog de "Oxford Protein Informatics Group"

[https://www.blopig.com/blog/2021/07/  
alphafold-2-is-here-whats-behind-the-structure-predict](https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-predict)

"... we have many new questions. What is the secret sauce before the news splash, and why is it so effective? Is it a piece of code that the average user can actually run? What are AlphaFold 2's shortcomings? And, most important of all, what will it mean for computational biology? And for all of us?"



## Le blog de Mohammed AlQuraishi

<https://moalquraishi.wordpress.com/2020/12/08/alphafold2-casp14-it-feels-like-ones-child-has-left-home-amp/>

# Mutations corrélées, MSA

# Mutations corrélées, MSA

OPEN ACCESS Freely available online

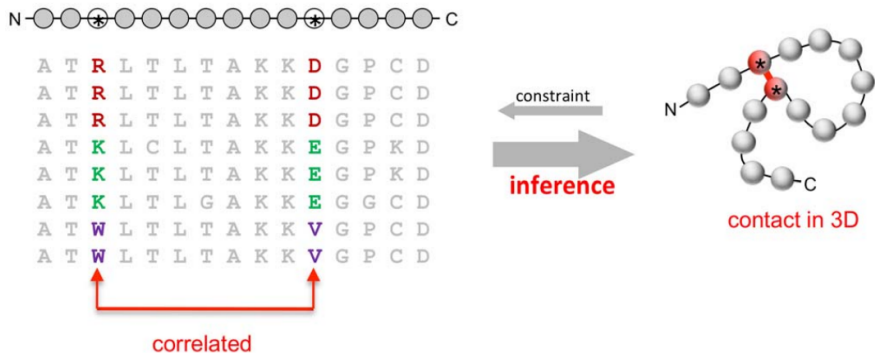


## Protein 3D Structure Computed from Evolutionary Sequence Variation

**Debora S. Marks<sup>1\*</sup>, Lucy J. Colwell<sup>2,3</sup>, Robert Sheridan<sup>3</sup>, Thomas A. Hopf<sup>1</sup>, Andrea Pagnani<sup>4</sup>, Riccardo Zecchina<sup>4,5</sup>, Chris Sander<sup>3</sup>**

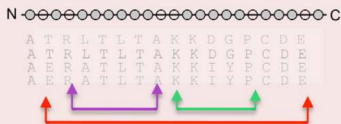
**1** Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, United States of America, **2** MRC Laboratory of Molecular Biology, Hills Road, Cambridge, United Kingdom, **3** Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York, New York, United States of America, **4** Human Genetics Foundation, Torino, Italy, **5** Politecnico di Torino, Torino, Italy

## Principe



## Protocole de 2011 sans deep-learning

Align evolutionary diverged sequences



Calculate covariance matrix for each pair of sequence positions for all pairs of amino acids (A,B)

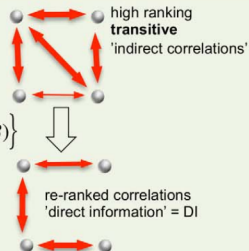
$$C_{ij}(A,B) = f_{ij}(A,B) - f_i(A)P_j(B)$$

$$C_{ij}^{-1}(A,B) = -e_{ij}(A,B)_{i \neq j}$$

$$P_{ij}^{Dir}(A,B) = \frac{1}{Z} \exp\{e_{ij}(A,B) + \tilde{h}_i(A) + \tilde{h}_j(B)\}$$

Identify maximally informative pair couplings using **statistical model** of entire protein to infer residue-residue co-evolution

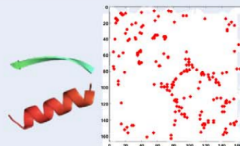
$$DI_{ij} = \sum_{A,B=1}^q P_{ij}^{Dir}(A,B) \ln \frac{P_{ij}^{Dir}(A,B)}{f_i(A)f_j(B)}$$



## Protocole de 2011 sans deep-learning

Analyze the highest scoring pairs to produce ranked list of residue pairs which we predict to be close in 3D space. Use these pairs as predicted close "evolutionary inferred contacts"<sup>3</sup>, EICs, in folding calculations

```
assign (resid 143 and name CA) (resid 123 and name CA) 4 4 3
assign (resid 16 and name CA) (resid 10 and name CA) 4 4 3
assign (resid 141 and name CA) (resid 82 and name CA) 4 4 3
assign (resid 129 and name CA) (resid 87 and name CA) 4 4 3
assign (resid 92 and name CA) (resid 11 and name CA) 4 4 3
assign (resid 116 and name CA) (resid 81 and name CA) 4 4 3
```

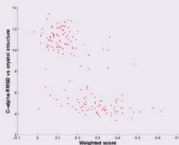


predicted contacts (EICs)

Start with extended structure use **distance geometry** and **simulated annealing** with predicted constraints, EICs, to fold the chain



Rank predicted structures using quality measure of backbone alpha torsion and beta sheet twist



good scores



bad scores



# Utiliser Alpha Fold

## google colab

[https://colab.research.google.com/notebooks/intro.ipynb?utm\\_source=scs-index](https://colab.research.google.com/notebooks/intro.ipynb?utm_source=scs-index)

<https://www.youtube.com/watch?v=inN8seMm7UI>

**Exemple pour faire de la MD ou aussi AlphaFold2 + MD:**

<https://github.com/pablo-arantes/Making-it-rain>



# Alpha Fold avec google colab

**une version amélioré d'alphaFold2:**

<https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb>

**ou l'original:**

<https://colab.research.google.com/github/deepmind/alphafold/blob/main/notebooks/AlphaFold.ipynb>

## Alpha Fold en local

On peut l'installer aussi sur une machine en local, mais il faut qu'elle soit très puissante:

<https://github.com/deepmind/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."

Rien que la carte graphique A100 coûte le prix d'une petite voiture:

<https://fr.aliexpress.com/item/1005002408111365.html>

Pour aller plus loin: EMBL webinar

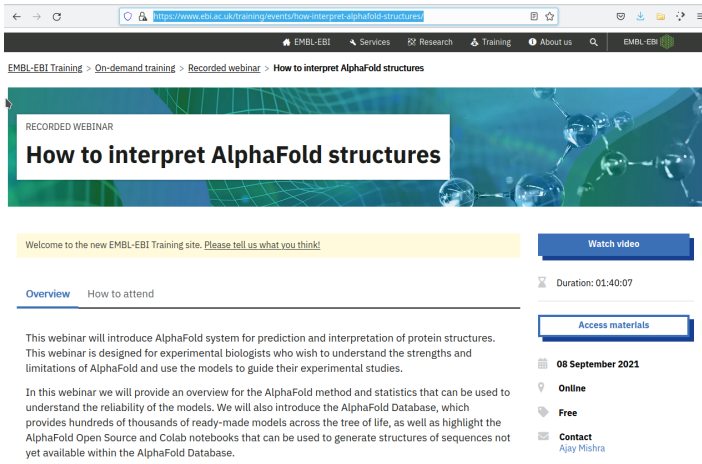
Pour aller plus loin: EMBL webinar

Pour aller plus loin: EMBL webinar

# EMBL webinar "How to interpret AlphaFold structures"

8 sept 2021

<https://www.ebi.ac.uk/training/events/how-interpret-alphafold-structures/>



The screenshot shows a web browser displaying the EMBL-EBI website. The address bar shows the URL: <https://www.ebi.ac.uk/training/events/how-interpret-alphafold-structures/>. The navigation menu includes: EMBL-EBI, Services, Research, Training, About us, and EMBL-EBI logo. The breadcrumb trail is: EMBL-EBI Training > On-demand training > Recorded webinar > How to interpret AlphaFold structures.

The main content area features a banner with the text: **RECORDED WEBINAR** and **How to interpret AlphaFold structures**. Below the banner is a yellow message box: "Welcome to the new EMBL-EBI Training site. [Please tell us what you think!](#)".

There are two tabs: **Overview** (selected) and **How to attend**.

The main text under the **Overview** tab reads: "This webinar will introduce AlphaFold system for prediction and interpretation of protein structures. This webinar is designed for experimental biologists who wish to understand the strengths and limitations of AlphaFold and use the models to guide their experimental studies." and "In this webinar we will provide an overview for the AlphaFold method and statistics that can be used to understand the reliability of the models. We will also introduce the AlphaFold Database, which provides hundreds of thousands of ready-made models across the tree of life, as well as highlight the AlphaFold Open Source and Colab notebooks that can be used to generate structures of sequences not yet available within the AlphaFold Database."

On the right side, there are several buttons and information: **Watch video** (blue button), **Duration: 01:40:07**, **Access materials** (blue button), **08 September 2021**, **Online**, **Free**, and **Contact Ajay Mishra**.

DeepMind

# Introduction to AlphaFold

**Presenter:** Kathryn Tunyasuvunakool  
Research Scientist at DeepMind



# How does it work? (the short version)

## Article Highly accurate protein structure prediction with AlphaFold

bioRxiv preprint doi: <https://doi.org/10.1101/2021.03.03.430812>; this version posted May 20, 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 aCC-BY-NC-ND 4.0 International license.

Protein structure prediction is a long-standing challenge in computational biology. The development of deep learning methods has led to significant improvements in the accuracy of protein structure prediction. In this paper, we present AlphaFold, a deep learning-based method for protein structure prediction. AlphaFold achieves a state-of-the-art performance on the CASP14 competition, demonstrating high accuracy in predicting protein structure from sequence alone. The success of AlphaFold is attributed to its novel architecture, which combines a deep neural network for sequence representation with a transformer-based attention mechanism for capturing long-range dependencies in the protein sequence. AlphaFold's performance is evaluated on a large set of protein structures, showing that it can accurately predict the structure of proteins that have never been solved before. This work represents a major step forward in the field of protein structure prediction, and it has the potential to revolutionize our understanding of protein function and disease.

The development of computational methods to predict the structure of proteins from their amino acid sequences is a long-standing challenge in computational biology. In this paper, we present AlphaFold, a deep learning-based method for protein structure prediction. AlphaFold achieves a state-of-the-art performance on the CASP14 competition, demonstrating high accuracy in predicting protein structure from sequence alone. The success of AlphaFold is attributed to its novel architecture, which combines a deep neural network for sequence representation with a transformer-based attention mechanism for capturing long-range dependencies in the protein sequence. AlphaFold's performance is evaluated on a large set of protein structures, showing that it can accurately predict the structure of proteins that have never been solved before. This work represents a major step forward in the field of protein structure prediction, and it has the potential to revolutionize our understanding of protein function and disease.

Supplementary Information for: Highly accurate protein structure prediction with AlphaFold

## Supplementary Information for: Highly accurate protein structure prediction with AlphaFold

John Jumper<sup>1</sup>, Richard Evans<sup>1</sup>, Alexander Pritzel<sup>1</sup>, Tom Green<sup>1</sup>, Michael Figurno<sup>1</sup>, Iain Tunyasubunwong<sup>1</sup>, Kathryn Tomasevsk<sup>1</sup>, Ronne Byvatn<sup>1</sup>, Andrew Dalgaard<sup>1</sup>, Anna Ringgaard<sup>1</sup>, Clément Mouton<sup>1</sup>, Sina A. Kohli<sup>1</sup>, Andrew J. Senior<sup>1</sup>, Andrew Cowan<sup>1</sup>, Benjamin Rieker<sup>1</sup>, Patrick Shaker<sup>1</sup>, Nikhil Joshi<sup>1</sup>, Jasnoor Kaur<sup>1</sup>, Trevor Beech<sup>1</sup>, Srijan Pichay<sup>1</sup>, David Reineer<sup>1</sup>, Ilkka Clancy<sup>1</sup>, Michal Zelikovsk<sup>1</sup>, Maria Skworek<sup>1</sup>, Michaluk Pachkral<sup>1</sup>, Tamas Bagshaw<sup>1</sup>, Sebastian Boudry<sup>1</sup>, David Shcherb<sup>1</sup>, Orit Viriyath<sup>1</sup>, Andrew W Senior<sup>1</sup>, Karel Kazianka<sup>1</sup>, Pabhanu Kaul<sup>1</sup> and Denis Hassel<sup>1</sup>

<sup>1</sup>DeepMind, London, UK

<sup>2</sup>School of Biological Sciences and Artificial Intelligence Institute, Seoul National University, Seoul, South Korea

<sup>3</sup>These authors contributed equally

\*Corresponding author: John Jumper, Denis Hassel

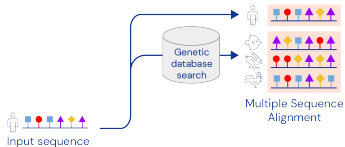
## Contents

1	Supplementary Methods	4
1.1	Notation	4
1.2	Data pipeline	5
1.2.1	Training	5
1.2.2	Dataset search	5
1.2.3	Targeting search	5
1.2.4	Training data	6
1.2.5	Validation data	6
1.2.6	MSA block definition	6
1.2.7	MSA clustering	6
1.2.8	Residue pairing	7
1.2.9	Residue contact prediction	7
1.3	Self-supervised dataset	7
1.4	AlphaFold-Multimer	10
1.4.1	Input and output	10
1.4.2	Architecture	11
1.4.3	MSA on-line guide and interface with Hmmer	14

See Jumper et al. 2021 (especially the SI) for details

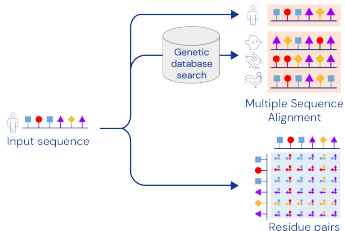


# Inputs



A key AlphaFold input is the MSA, containing sequences evolutionarily related to the target. Related sequences are found using standard tools and public databases.

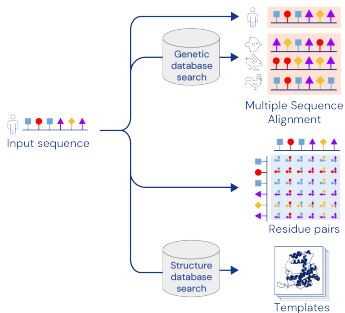
# Inputs



The input sequence is used to create an array of numbers representing all residue pairs.

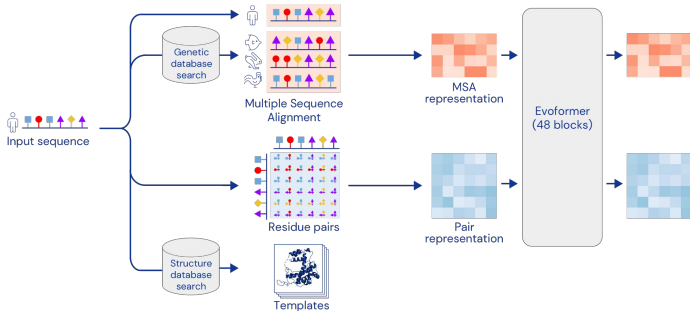


# Inputs



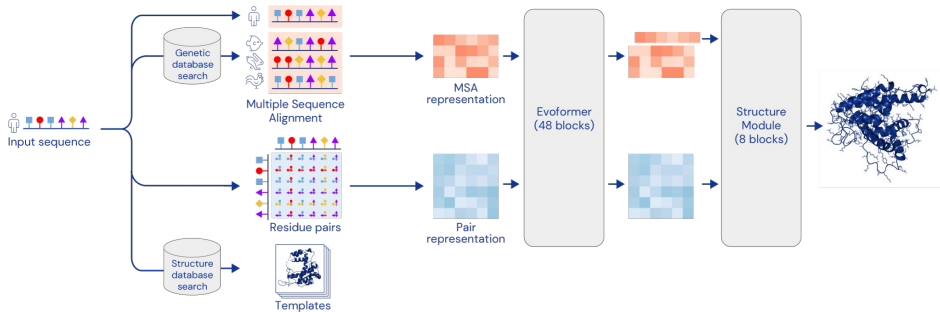
AlphaFold can also use template structures from the PDB, found using standard tools. However, it often produces accurate predictions without a template.

# Network



The Evoformer blocks extract information about the relationship between residues.  
The MSA representation can update the pair representation and vice versa.

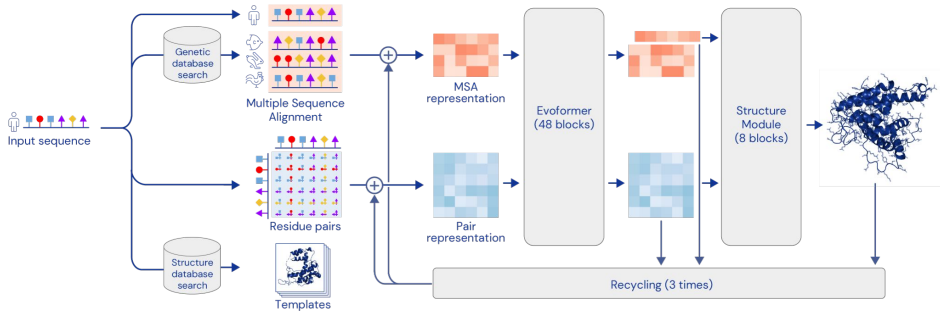
# Network



The Structure Module predicts a rotation + translation to place each residue.  
A small network predicts side chain chi angles. The final structure is run through a relaxation process.

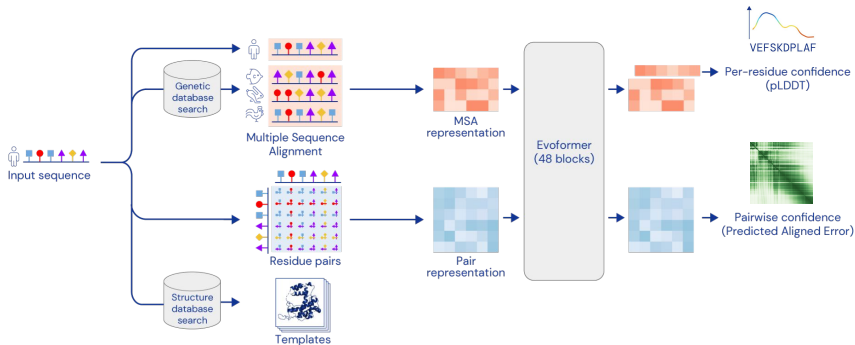
# Network

Private & Confidential



Feeding certain outputs back through the network again improves performance

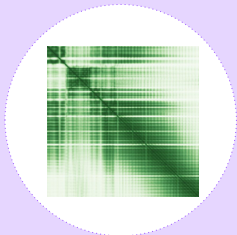
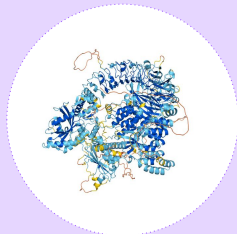
## Other outputs



As well as a predicted structure, AlphaFold produces two confidence estimates

# Interpreting predictions

The short version: use **both** confidence metrics

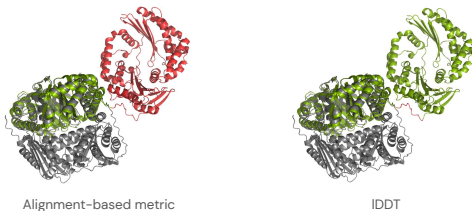


## Predicted LDDT: definition

AlphaFold's per-residue prediction of its IDDT-Ca score\*

Roughly, IDDT measures the percentage of correctly predicted interatomic distances, not how well the predicted and true structures can be superimposed.

It rewards **locally correct** structures, and **getting individual domains right**.  
pLDDT behaves similarly, as a measure of **local confidence**



\*Mariani et al. 2013

## Predicted LDDT: format

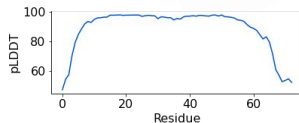
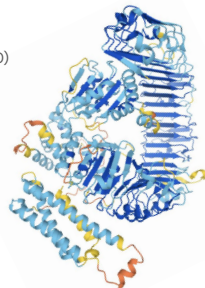
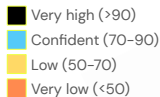
pLDDT ranges from 0 to 100 (100 is most confident)

We use a consistent “confidence bands” color scheme when displaying predictions

A pLDDT plot is also displayed by some of our tools

Prediction files always contain pLDDT in the B-factors  
Therefore a **higher** B-factor is better!

MODEL		0									
ATOM	1	N	MET	A	1	-9.212	-5.798	33.490	1.00	63.75	N
ATOM	2	CA	MET	A	1	-10.075	-6.130	32.322	1.00	63.75	C
ATOM	3	C	MET	A	1	-11.469	-6.615	32.714	1.00	63.75	C
ATOM	4	CB	MET	A	1	-9.419	-7.112	31.341	1.00	63.75	C
ATOM	5	O	MET	A	1	-12.429	-6.075	32.184	1.00	63.75	O
ATOM	6	CG	MET	A	1	-8.311	-6.411	30.547	1.00	63.75	C
ATOM	7	SD	MET	A	1	-7.766	-7.280	29.061	1.00	63.75	S
ATOM	8	CE	MET	A	1	-7.045	-8.751	29.798	1.00	63.75	C
ATOM	9	N	ALA	A	2	-11.624	-7.579	33.634	1.00	66.38	N
ATOM	10	CA	ALA	A	2	-12.951	-8.096	34.007	1.00	66.38	C





# Predicted LDDT: usage

Private & Confidential

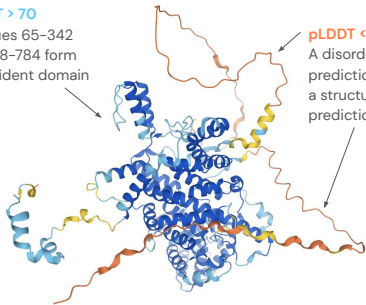
Identifying domains & possible disordered regions

**pLDDT > 70**

Residues 65-342  
and 418-784 form  
a confident domain

**pLDDT < 50**

A disorder  
prediction not  
a structure  
prediction



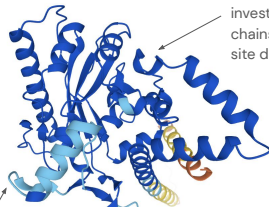
Assessing confidence within a domain

**pLDDT > 90**

Reasonable to  
investigate side  
chains / active  
site details

**pLDDT > 70**

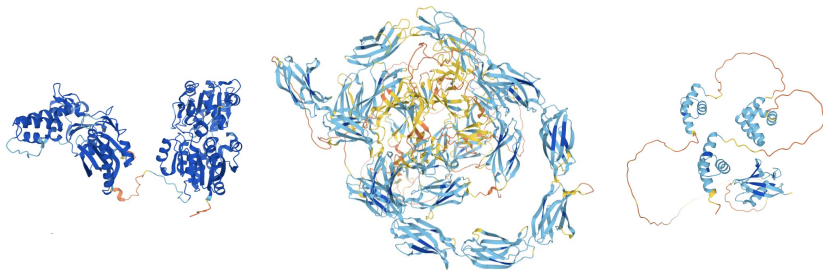
Lower confidence on  
these specific parts



## Predicted LDDT: pitfalls

Private & Confidential

High pLDDT on all domains does **not** imply AlphaFold is confident of their relative positions



## Predicted Aligned Error: definition

AlphaFold's prediction of its position error at residue  $x$ ,  
if the predicted and the true structures were aligned on residue  $y$

PAE aims to measure confidence in the **relative positions** of **pairs of residues**

Mainly used to assess relative domain positions, but applicable whenever pairwise confidence is relevant

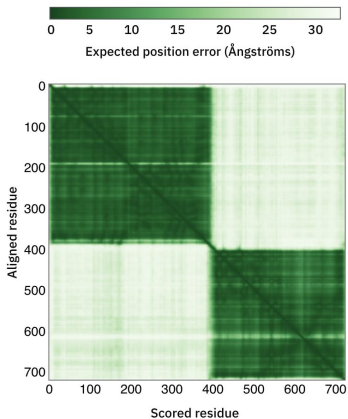
# Predicted Aligned Error: format

Private & Confidential

PSE is displayed as a 2D plot.

Suppose residue  $y$  were aligned to the true structure and we measured the position error at residue  $x$ . The color at  $(x, y)$  is the predicted position error.

In this case the square correspond to two domains.



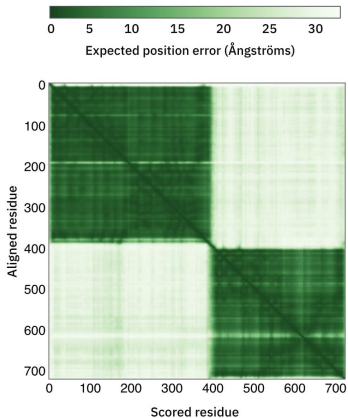
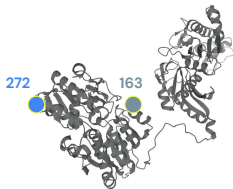
# Predicted Aligned Error: format

Private & Confidential

PSE is displayed as a 2D plot.

Suppose residue  $j$  were aligned to the true structure  
and we measured the position error of residue  $i$ .  
The color of  $(i, j)$  is the predicted position error.

AlphaFold is confident in relative positions within each domain.



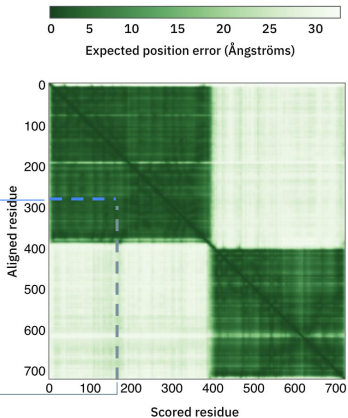
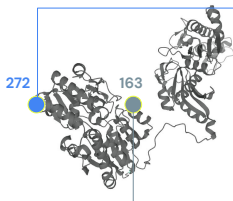
# Predicted Aligned Error: format

Private & Confidential

PSE is displayed as a 2D plot.

Suppose residue  $y$  were aligned to the true structure  
and we measured the position error at residue  $x$ .  
The color at  $(x, y)$  is the predicted position error.

AlphaFold is confident in relative positions within each domain.



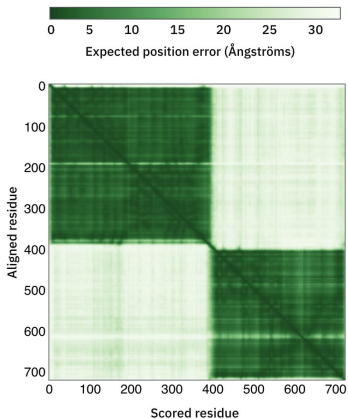
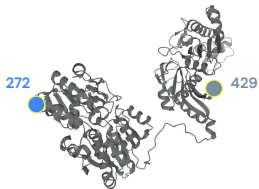
# Predicted Aligned Error: format

Private & Confidential

PSE is displayed as a 2D plot.

Suppose residue  $j$  were aligned to the true structure  
But we measured the position error at residue  $i$ .  
The color at  $(i, j)$  is a likelihood prediction of the error.

Just read between domains.



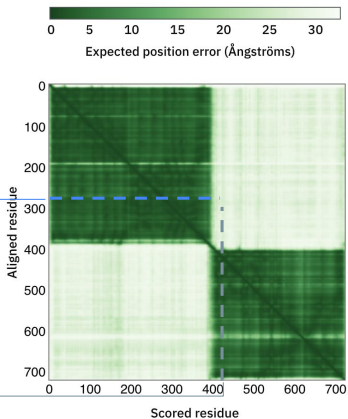
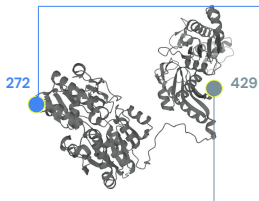
# Predicted Aligned Error: format

Private & Confidential

RAE is displayed as a 2D plot.

Suppose residue  $y$  were aligned to the true structure  
and we measured the position error of residue  $x$ .  
The color of  $(x, y)$  is a statistical prediction of the error.

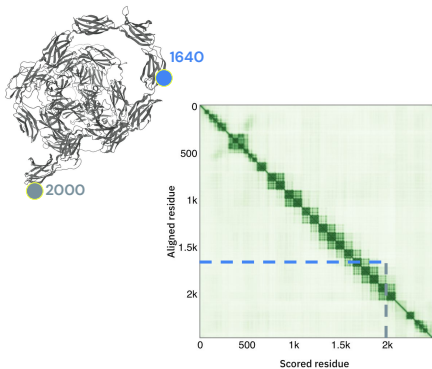
Just read between domains.



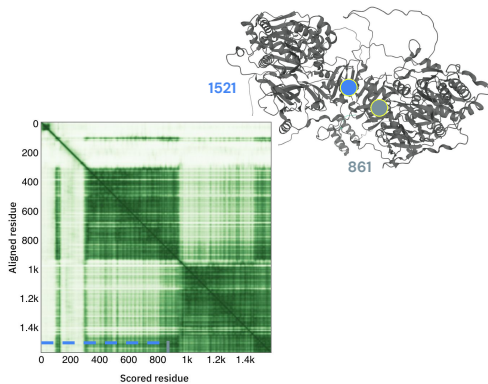


# Predicted Aligned Error: usage

No confidence in relative domain positions



Predicted domain packing



# Things to be aware of

## Uncertain domain placement

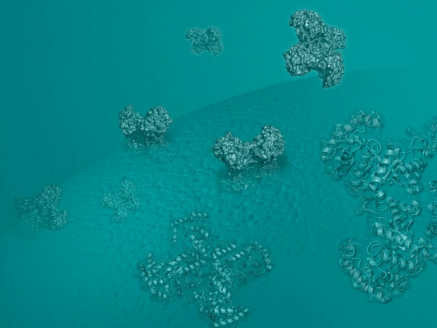
- If AlphaFold is uncertain, it won't necessarily place domains sensibly relative to each other
  - Membrane proteins won't leave space for the cell membrane
  - Clashes can occur

## Complexes

- For proteins that exist in complex, AlphaFold is missing context about their binding partners
  - Heteromers more problematic than homomers
  - Worst case: the protein is flexible in isolation
- Some have had success predicting complexes by joining 2 sequences with a linker
  - We think it is possible to extend the ideas in AlphaFold to complexes
  - However, this linker setup remains to be benchmarked

# AlphaFold database

Sameer Velankar  
EMBL-EBI



## AlphaFold Database

- ~365K predicted structures for proteins from 21 model organisms
- For the organisms currently covered, predicted structures available for sequences in the UniProt reference proteome that are between 16 and 2700 amino acids long and contain only standard amino acids
- Only one prediction (out of 5 independent predictions) is made available in the release
- Accession – AF-P12345-F[1-N]
- Files
  - AF-P12345-F[1-N]-model\_V[1-N].[pdb, mmCIF]
  - AF-P12345-F[1-N]-predicted-aligned-error\_V[1-N].json

# AlphaFold web pages

- Basic search system
- Allows search using UniProt accession, UniProt id, protein name, gene name and organism
- Clear indication that the structure shown is a prediction
- Allow easy download of structure data
- Basic information about protein
- Clearly indicates if there are experimental structures available
- Display residue-quality information in 3D viewer (pLDDT – predicted Local Distance Difference Test)
- Predicted Aligned Error (PAE viewer)

Showing all search results for **catpain**  
8 / 100 of 100 results

**Filter by:**

**Organism**

- Trypanosoma cruzi (Genus: Chlamydomonas)
- Chlamydomonas (2)
- Arabidopsis thaliana (2)
- Leishmania infantum (1)
- Mus musculus (1)
- Rattus norvegicus (1)
- Caenorhabditis elegans (1)
- Escherichia coli (1)
- Helicobacter pylori (1)
- Zooniverse (1)
- Arabidopsis thaliana (1)

**Calpain small subunit 1-like**  
Accession: P06608 (1)

**Protein:** Calpain small subunit 1-like  
**Gene:** CALP1  
**Source Organism:** Rattus norvegicus - search this organism or UniProt: A0A022K647 - go to UniProt or

**Calpain 15**  
Accession: P06609 (1)

**Protein:** Calpain 15  
**Gene:** CALP2  
**Source Organism:** Rattus norvegicus - search this organism or UniProt: A0A022K648 - go to UniProt or

**Probable disease resistance protein At1g58602**  
AlphaFold structure prediction

Download [PDB file](#) [swaCIF file](#) [Predicted aligned error](#)

**Information**

**Protein:** Probable disease resistance protein At1g58602  
**Gene:** At1g58602  
**Source organism:** Arabidopsis thaliana - go to search or  
**UniProt:** Q0V292 - go to UniProt or  
**Experimental structures:** None available in the PDB  
**Biological function:** Potential disease resistance protein - go to UniProt or

**3D Viewer**

Sequence of At1g58602: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100

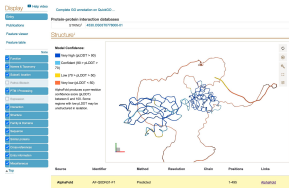
**Model Confidence:**

- Very High (pLDDT > 90)
- Confident (50 < pLDDT <= 90)
- Low (70 < pLDDT <= 50)
- Very Low (pLDDT < 50)

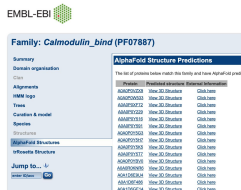
AlphaFold produces a per-residue confidence score (pLDDT) between 0 and 100. Some regions below 50 pLDDT may be worth cautious inspection.

# Models available across EBI resources

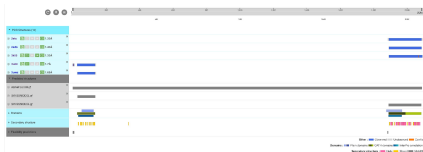
UniProt



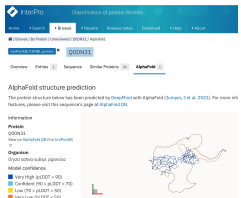
Pfam



PDBe-KB



InterPro



## AlphaFold Database – limitations

- Information on complexes with other proteins, nucleic acids (DNA or RNA) or ligands. In some cases, the single-chain prediction may correspond to the structure adopted in a complex. The missing context from surrounding molecules may lead to an uninformative prediction
  - AlphaFold does not make any predictions about any of the non-protein components such as cofactors, metals, ligands including drug-like molecules, ions, carbohydrates and other post-translational modifications
- Protein dynamics - AlphaFold will usually only produce one of multiple conformations
- AlphaFold has not been trained or validated for predicting the effect of mutations
- May (or may not) lead to hypotheses about protein function – any hypotheses have to be tested by further experimentation

## What's next – under discussion

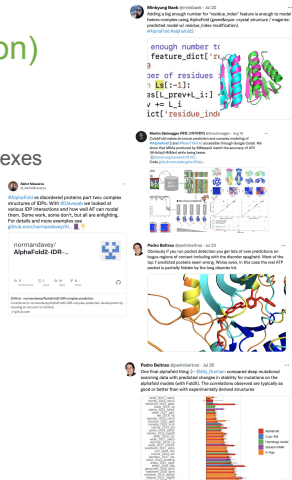
- Remove signal peptides from predictions
- Making 5 independent predictions available for each protein
- Additional metadata
  - MSA – need to consider data size
  - information on templates
  - quality criteria e.g. predicted TM score
- Updating database to UniRef90 dataset (~130 million structures)



# Impact of AlphaFold database on life science research

# Structural bioinformatics – (structure/function)

- Predicting complexes between macromolecules
  - Homo- and Hetero- Protein-protein; Protein-nucleic acid complexes
  - Intrinsically disordered proteins
- Provide information on protein dynamics
  - Relevant conformational states
- Functionally important residues
  - Impact of mutation; Binding sites; Conformationally important residues
  - Interfaces
- Ligand prediction – What binds?
  - What might bind in a pocket



# Structural biology

- Accelerating structure studies
  - Improved construct design
  - Starting model for structure determination
  - Fitting models in low resolution EM maps
  - Time resolved studies to understand mechanism



PRE-AP POST-AP

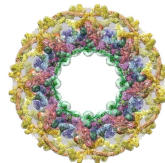
Justin M. Gonzalez  
@JMGonzalez

Just happened today I wanted to do a crystal with a small fragment with a small His tag at the C-terminus. I had over 100 mg of protein. This morning got the AP prediction, obviously adjustment to fit into the map a refinement and right... [@AlphaFold](#) is a game changer!




## Structural biology

- Integrative/hybrid methods
  - Models for individual components
- Combination of sparse experimental data and predicted model may lead to actionable data to test hypothesis
  - Chemical foot printing
  - Hydrogen-Deuterium exchange
  - smFRET - Single molecule fluorescence resonance energy transfer



I/H Methods Structures  
552-protein yeast Nuclear Pore Complex  
Kim et al. (2018) *Nature* 555, 475-82  
PDBDEV\_00000010; PDBDEV\_00000011; PDBDEV\_00000012



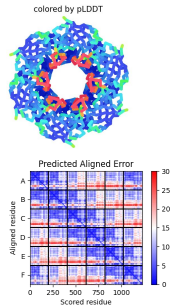
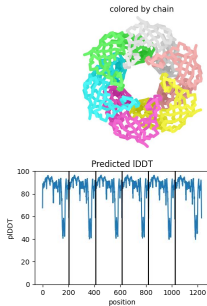
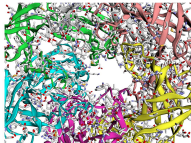
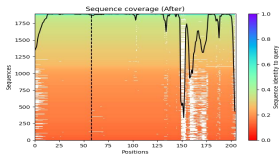
**Jochem Smit** @Jhsmit\_ · 23 Jul  
Replying to @eitan\_jerner  
If we had a curated **smFRET** / structure database it could probably serve as an input to a modified **AlphaFold** which might give us structures of transient species or ratios of subpopulations  
(and/or HDX-MS etc)  
2 3

**Dina Grohmann** @DinaGrohmann · 23 Jul  
I'm amazed by the accuracy of the predicted structure in the Mid/PIWI lobe. Apart from that, I think that **AlphaFold** could help us **smFRET** folks to find suitable positions for dye engineering.  
1 3

Show this thread

# ColabFold

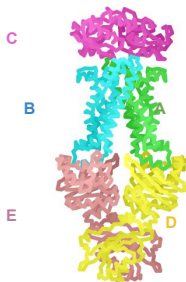
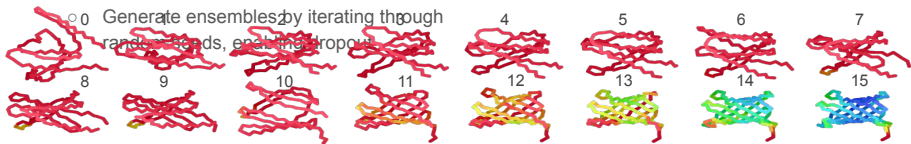
Making Protein folding accessible to all via Google Colab  
(and the unintended uses of AlphaFold)



[github.com/sokrypton/ColabFold](https://github.com/sokrypton/ColabFold)

# ColabFold - Advanced options

- Modify MSA input
  - Custom or MMseqs2 (much faster)
  - Trim
- **Complexes**
  - **Homo-oligomers**
  - **Hetero-oligomers**
- Fine control
  - Number of recycles
- Sample (Output more than 5 models)





# Can predict protein-protein/peptide interactions

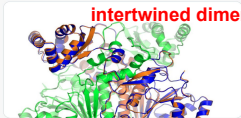
**Cesar Ramirez-Sarmiento** @cxarramirez · Jul 22  
Replying to @cxarramirez @sokrypton and 3 others  
OMG 🤯 the monomers in the predicted "homodimer" of FoxP1 (cyan, green) show very similar orientations when compared to the monomers in the domain-swapped structure (red) 🤖

dimer-swap

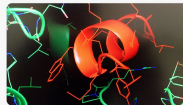


**James Murray** @jem\_imperial · Jul 24  
Replying to @drpetarmood  
There are already notebooks to predict heterodimers and homooligomers. [github.com/sokrypton/Colo](https://github.com/sokrypton/Colo). For my unpublished intertwined dimer, the monomer and dimer predictions were essentially identical, also matching the crystal structure exactly except for a few rotamers.

intertwined dimer

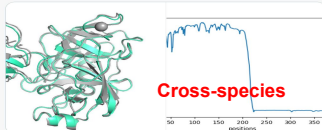


**Eugene Vekov** @eugenevekov · Jul 26  
After days of running [AlphaFold2](https://github.com/sokrypton/Colo), I am still astounded by IneqsIt provides. Here, it predicted mode of peptide binding "completely consistent" with biochemical data and mutagenesis and gave additional clues which we will explore!



consistent w/  
biochem data

**padhorny** @padhorny · Jul 20  
Replying to @sokrypton and @minkbaek  
Amazing stuff. Seemingly can even do cross-species complexes (at least the strong binders). Here is what it gave me for [rcsb.org/structure/1AVX](https://rcsb.org/structure/1AVX) (not the top model though):



## Preprints rolling in...

Can AlphaFold2 predict protein-peptide complex structures accurately?

Junsu Ko, Juyong Lee

bioRxiv 2021.07.27.453972; doi: <https://doi.org/10.1101/2021.07.27.453972>

Harnessing protein folding neural networks for peptide-protein docking

Tomer Tsaban, Julia Varga, Orly Avraham, Ziv Ben-Aharon, Alisa Khrumshin, Ora Schueler-Furman

bioRxiv 2021.08.01.454656; doi: <https://doi.org/10.1101/2021.08.01.454656>

Improved Docking of Protein Models by a Combination of AlphaFold2 and ClusPro

Usman Ghani, Israel Desta, Akhil Jindal, Omeir Khan, George Jones, Sergey Kotelnikov, Dzmitry Padhorny, Sandor Vajda, Dima Kozakov

bioRxiv 2021.09.07.459290; doi: <https://doi.org/10.1101/2021.09.07.459290>



# ColabFold se diversifie

## Making Protein folding accessible to all via Google Colab!

<b>Notebooks</b>	<b>monomers</b>	<b>complexes</b>	<b>mmseqs2</b>	<b>jackhmmmer</b>	<b>templates</b>
<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 (from Deepmind)</a>	Yes	Yes	No	Yes	No
<b>BETA (in development) notebooks</b>					
<a href="#">OmegaFold</a>	Yes	No	No	No	No
<a href="#">AlphaFold2_advanced</a>	Yes	Yes	Yes	Yes	No
<b>OLD retired notebooks</b>					
<a href="#">AlphaFold2_complexes</a>	No	Yes	No	No	No
<a href="#">AlphaFold2_jackhmmmer</a>	Yes	No	Yes	Yes	No
<a href="#">AlphaFold2_noTemplates_noMD</a>					
<a href="#">AlphaFold2_noTemplates_yesMD</a>					

# ColabDesign

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

## ColabDesign

---

### Making Protein Design accessible to all via Google Colab!

```
pip install git+https://github.com/sokrypton/ColabDesign.git
```

- [TrDesign](#) - using TrRosetta for design (support for TrMRf coming soon)
- [AfDesign](#) - using AlphaFold for design

(WIP) Not yet fully integrated into ColabDesign

- [MSA\\_transformer](#)
- [Potts models](#) (GREMLIN, mFDCA, arDCA, plmDCA, bmDCA, etc)
- [ProteinMPNN](#)
- [RfDesign](#) - using RoseTTAFold for design

### Presentations

[Slides Talk](#)

### Contributors:

- [Sergey Ovchinnikov](#) @sokrypton
- [Justas Dauparas](#) @dauparas
- [Weikun.Wu](#) @guyujun Levinthal.bio
- [Shihao Feng](#)

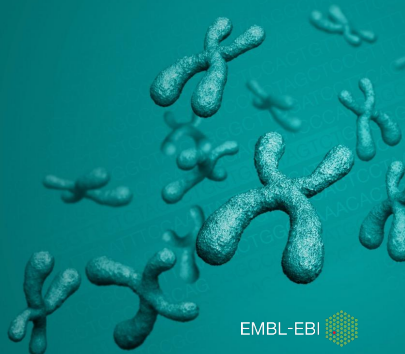
Pour aller plus loin: EMBL webinar

# AlphaFold-Multimer de DeepMind

Richard Evans et al. (Oct. 2021). en. preprint. Bioinformatics

# Some AlphaFold use cases

Alex Bateman



## Just because AlphaFold can fold it doesn't mean nature can

- Pfam use case 2: CPB\_BcsS family (PF17036)
- AlphaFold prediction of region matched by Pfam identifies incomplete domain
- Structurally similar to MBB clan

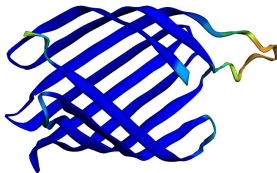
There are 129 sequences with the following architecture: CPB\_BcsS

[A0A3S2XL24\\_9RHIZ](#) [Methylobacterium sp. TER-1] Cellulose biosynthesis protein BcsS (ECO:0000313|EMBL:RVU17441.1) (241 residues)

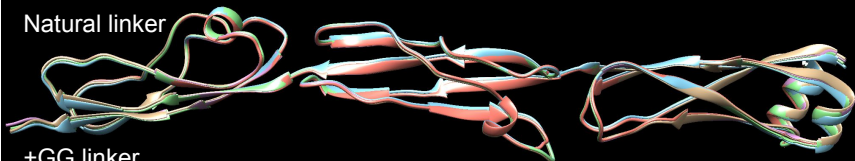


[Show](#) all sequences with this architecture.

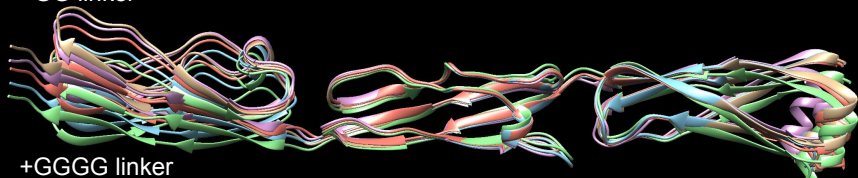
- This will not be stable in vitro!



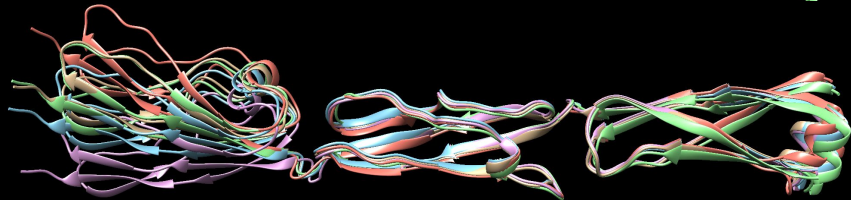
Natural linker



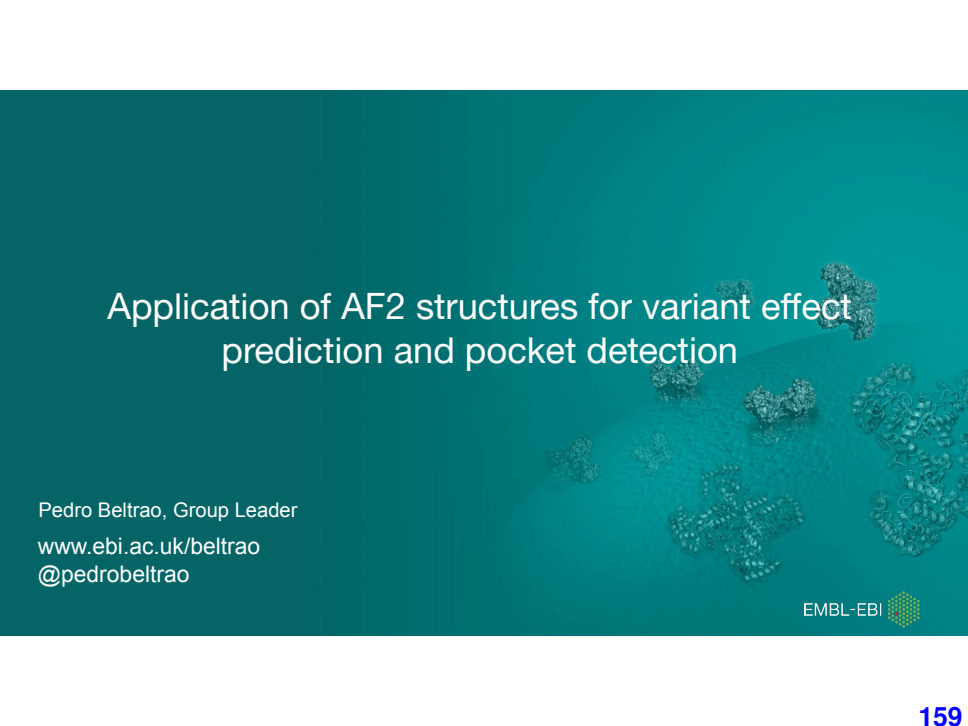
+GG linker



+GGGG linker



# Application of AF2 structures for variant effect prediction and pocket detection

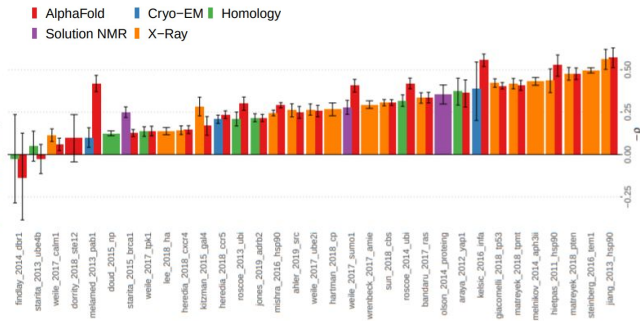


Pedro Beltrao, Group Leader

[www.ebi.ac.uk/beltrao](http://www.ebi.ac.uk/beltrao)

@pedrobeltrao

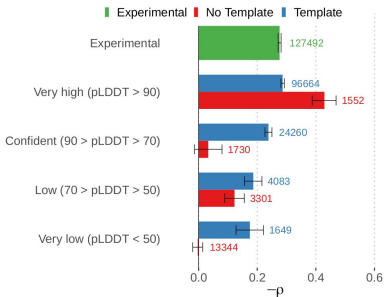
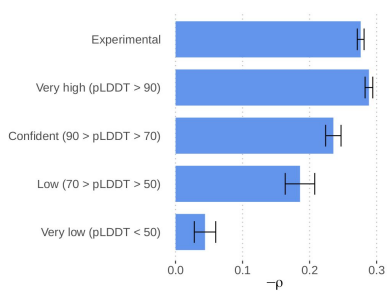
# Comparing experimental vs structure based prediction of missense mutations



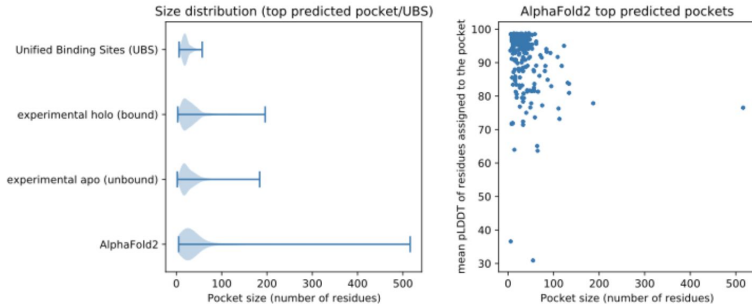
- Experimental impact of mutations from deep mutational scanning experiments (30 proteins)
- Predicted ddG of mutation using FoldX on alphafold structures or experimental structures
- Alphafold structures give equal or better predictions and it holds for regions with no templates



# Comparing experimental vs structure based prediction of missense mutations

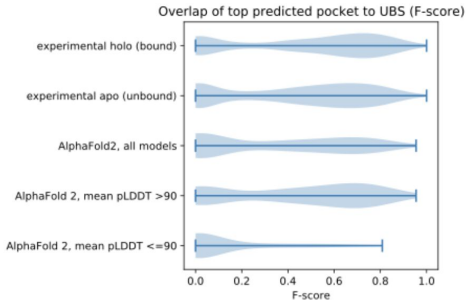


## Pocket detection and how to filter the models



We retained 230 of 304 proteins from a dataset by Clark et al., 2020. Pocket detection was performed using ghecom (Kawabata, 2010), as done previously in (Clark et al., 2020).

## Pocket detection and how to filter the models



Filtering the pocket residues by confidence (likely also predicted aligned residue) improves pocket detection

# AlphaFold2

for detecting intrinsically  
disordered protein regions

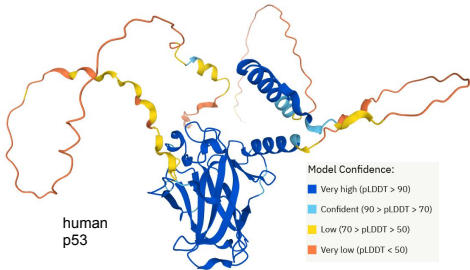


Bálint Mészáros  
EMBL Heidelberg  
08/09/2021



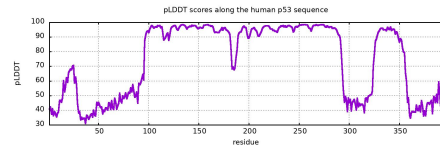
# AlphaFold2 indicates the presence of IDRs

AF2 generates coordinates for every residue, even ones that have no fixed structure



## Two interpretations for low confidence:

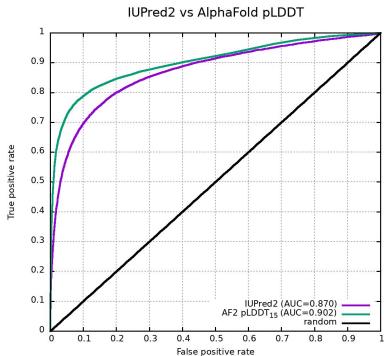
- AF2 isn't good enough to predict the structure
- There is no structure to predict



pLDDT is a good indicator of disordered regions (in this case)

Let's test the generic case – binary disordered prediction

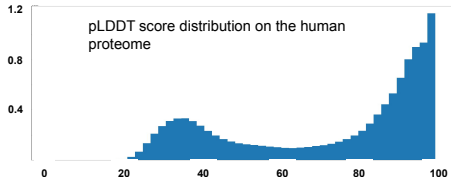
# AlphaFold2 as a disorder prediction method



false positive



false negative



# L'avenir selon DeepMind

## Une petite vidéo pour terminer...

[https://www.youtube.com/watch?time\\_continue=8&v=KpedmJdrTpY&feature=emb\\_title](https://www.youtube.com/watch?time_continue=8&v=KpedmJdrTpY&feature=emb_title)

à regarder chez vous (regarder sur les écrans, vous reconnaissez le logiciel de visualisation que les chercheurs de DeepMind utilisent ?):

<https://www.youtube.com/watch?v=gg7WjuF's8F4>



# The end

- MERCI pour votre attention!

## 7 Bibliographie

# Bibliographie I



Akasaka, Kazuyuki, Ryo Kitahara, and Yuji O. Kamatari (Mar. 2013). “Exploring the folding energy landscape with pressure”. In: *Archives of Biochemistry and Biophysics*. Protein Folding and Stability 531.1–2, pp. 110–115.



Bowman, Gregory R, Vincent A Voelz, and Vijay S Pande (Feb. 2011). “Taming the complexity of protein folding”. en. In: *Current Opinion in Structural Biology* 21.1, pp. 4–11.



Chen, Yiwen et al. (Jan. 2008). “Protein folding: Then and now”. In: *Archives of Biochemistry and Biophysics*. Highlight Issue: Protein Folding 469.1, pp. 4–19.



Daggett, Valerie and Alan R. Fersht (2003). “Is there a unifying mechanism for protein folding?” In: *Trends in biochemical sciences* 28.1, pp. 18–25.



Dill, Ken A. et al. (June 2008). “The Protein Folding Problem”. en. In: *Annual Review of Biophysics* 37.1, pp. 289–316.



Evans, Richard et al. (Oct. 2021). *Protein complex prediction with AlphaFold-Multimer*. en. preprint. Bioinformatics.



Fersht, Alan R. and Valerie Daggett (2002). “Protein folding and unfolding at atomic resolution”. In: *Cell* 108.4, pp. 573–582.



Hosseinzadeh, Parisa et al. (Dec. 2017). “Comprehensive computational design of ordered peptide macrocycles”. en. In: *Science* 358.6369, pp. 1461–1466.

# Bibliographie II



Huang, Po-Ssu, Scott E. Boyken, and David Baker (Sept. 2016). “The coming of age of *de novo* protein design”. en. In: *Nature* 537.7620, pp. 320–327.



Kuhlman, Brian et al. (Nov. 2003). “Design of a Novel Globular Protein Fold with Atomic-Level Accuracy”. en. In: *Science* 302.5649, pp. 1364–1368.



Lane, Thomas J et al. (Feb. 2013). “To milliseconds and beyond: challenges in the simulation of protein folding”. en. In: *Current Opinion in Structural Biology* 23.1, pp. 58–65.



Muñoz, Victor and Michele Cerminara (Sept. 2016). “When fast is better: protein folding fundamentals and mechanisms from ultrafast approaches”. In: *Biochem J* 473.17, pp. 2545–2559.



Piana, Stefano, John L Klepeis, and David E Shaw (Feb. 2014). “Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations”. en. In: *Current Opinion in Structural Biology* 24, pp. 98–105.



Plaxco, Kevin W, Kim T Simons, and David Baker (Apr. 1998). “Contact order, transition state placement and the refolding rates of single domain proteins” Edited by P. E. Wright”. In: *Journal of Molecular Biology* 277.4, pp. 985–994.



Thirumalai, D. et al. (2010). “Theoretical Perspectives on Protein Folding”. In: *Annual Review of Biophysics* 39.1, pp. 159–183.

# Bibliographie III



Voelz, Vincent A. et al. (Aug. 2012). “Slow Unfolded-State Structuring in Acyl-CoA Binding Protein Folding Revealed by Simulation and Experiment”. In: *J. Am. Chem. Soc.* 134.30, pp. 12565–12577.



Woodside, Michael T. and Steven M. Block (2014). “Reconstructing Folding Energy Landscapes by Single-Molecule Force Spectroscopy”. In: *Annual Review of Biophysics* 43.1, pp. 19–39.