The machine learning revolution in structural biology

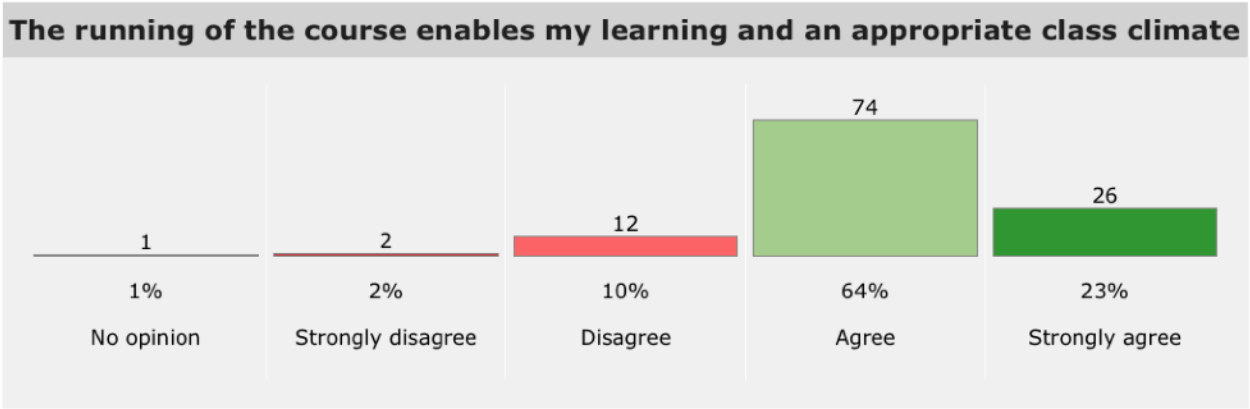
Matteo Dal Peraro
Institute of Bioengineering, School of Life Sciences

made by GPT4 - DALL-E

matteo.dalperaro@epfl.ch

Your feedback

115/182 - 63%



- slides in advance
- course/slides too dense
- not enough TAs

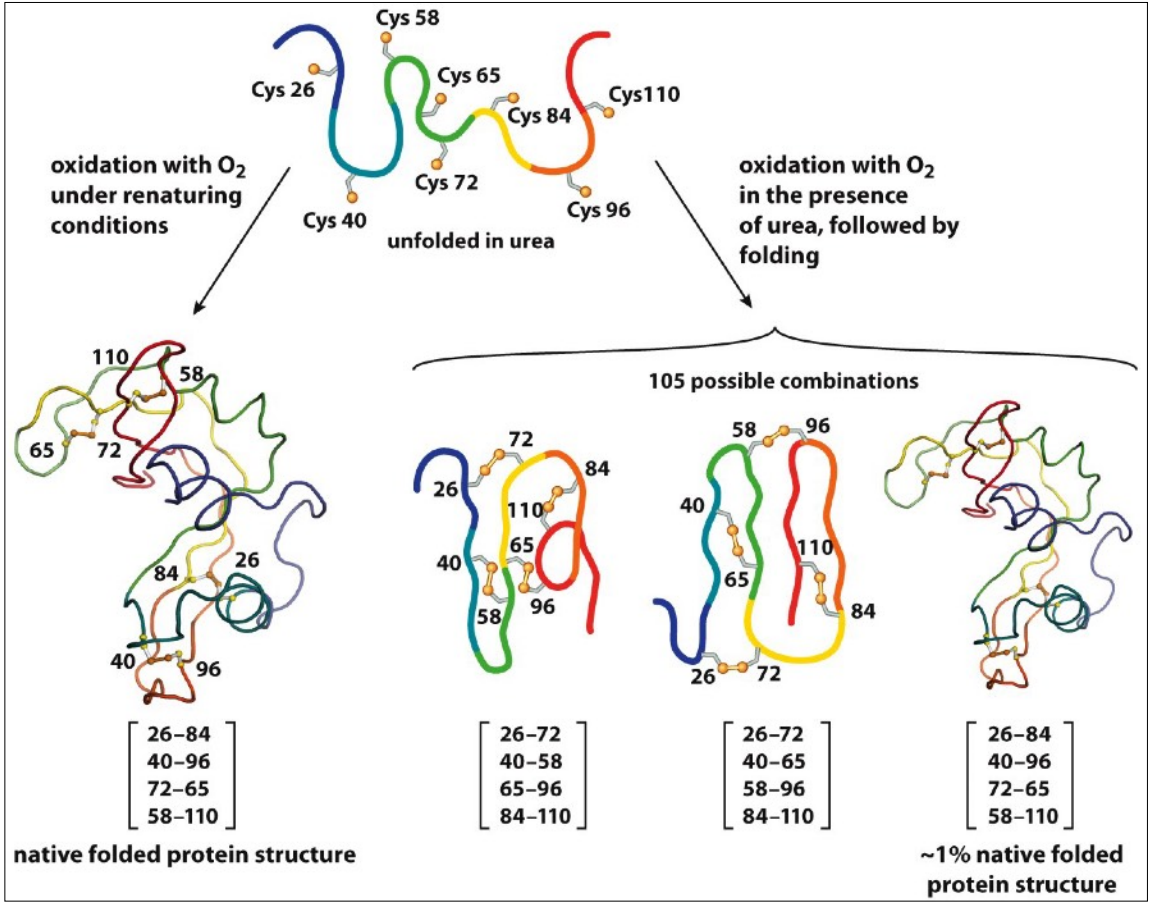
Sequence Determines Structure

The Anfinsen experiment (1954)

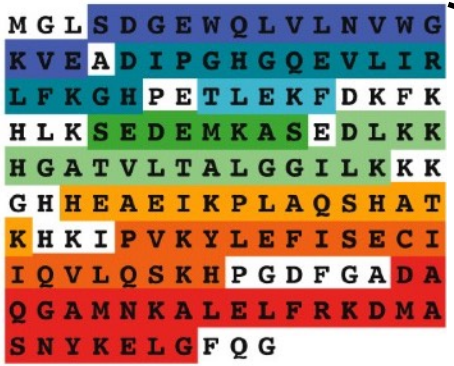
- A Nobel prize (1972) experiment that by measuring enzymatic activity figured out the principles of protein folding

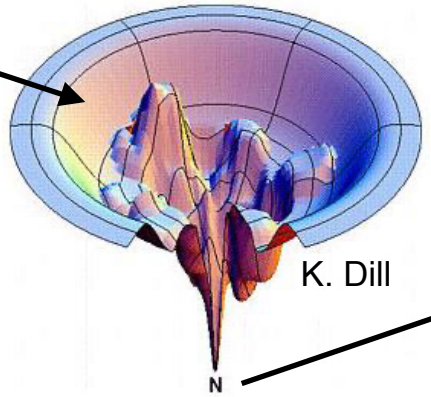
-Ribonuclease refolded and oxidized recovers 100% activity

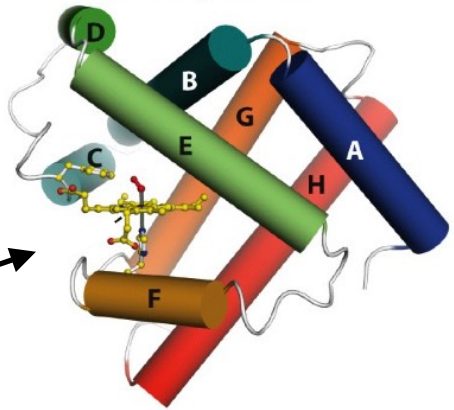
-Ribonuclease oxidized and refolded in presence of urea recovers 1% activity



Protein sequence determines structure







folding funnel

folded structure

- Anfinsen discovered it in a key experiment already in 1954
- This has become the central dogma of structural biology

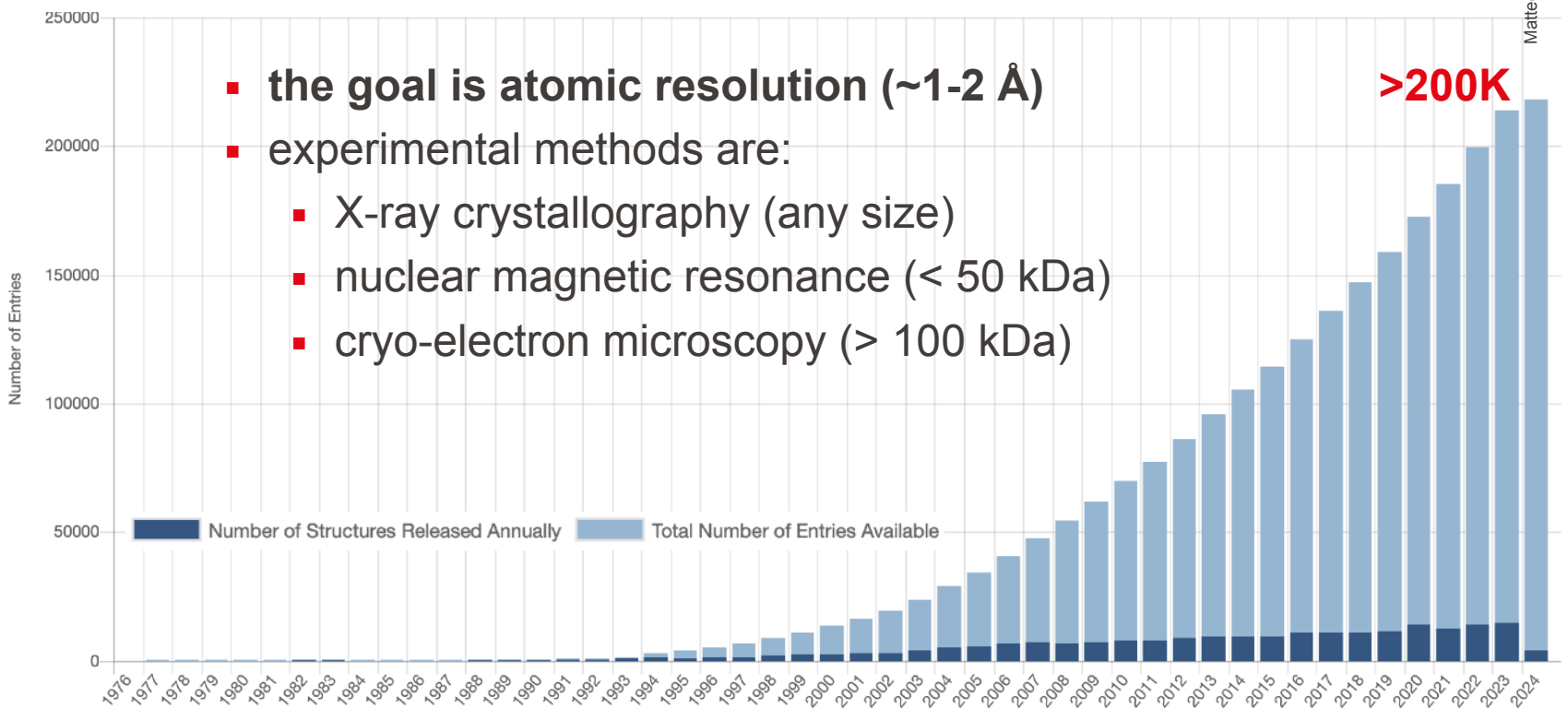
Matteo Dal Peraro

Matteo Dal Peraro

Structure determination is key

- **the goal is atomic resolution (~1-2 Å)**
- experimental methods are:
  - X-ray crystallography (any size)
  - nuclear magnetic resonance (< 50 kDa)
  - cryo-electron microscopy (> 100 kDa)

>200K



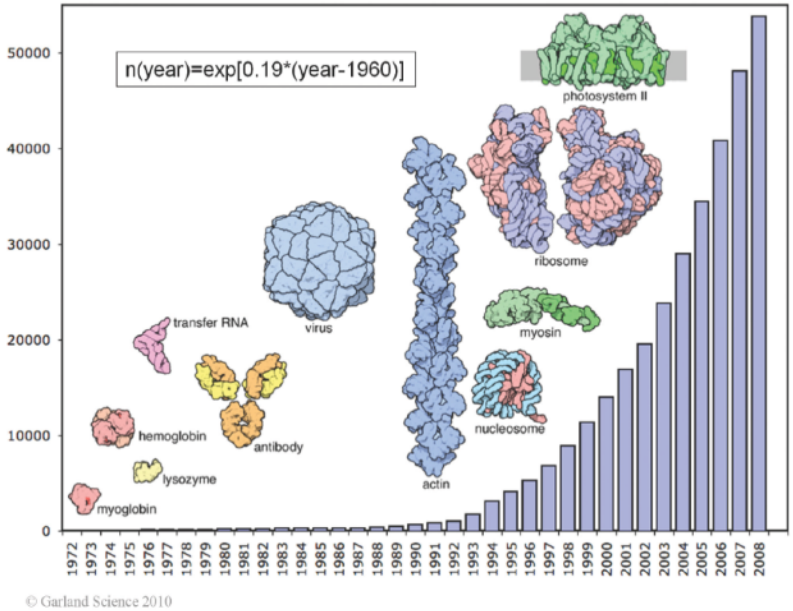
- **The Protein DataBank — PDB** **http://www.rcsb.org**

Biomolecular structure determination

atomic resolution (~1-2 Å, 0.1-0.2 nm - 10^{-10} m)

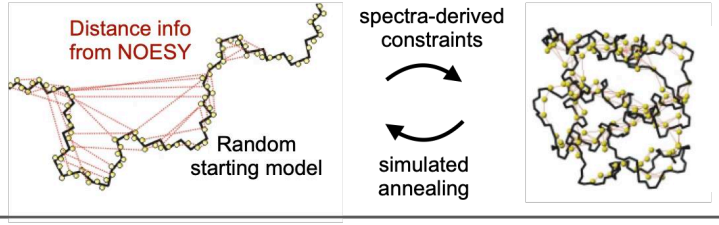
- **X-ray crystallography** (~1-2 Å; any size)
- **nuclear magnetic resonance** (~1-2 Å; < 50 kDa)
- **cryo-electron microscopy** (1-20 Å ; > 100 kDa)

in the **structure** and **dynamics** of any biomolecule in the cell is encoded its **function**



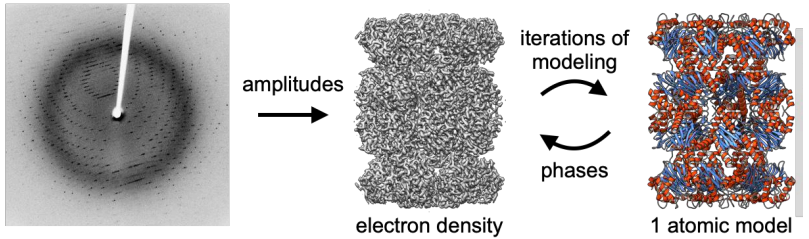
Methods for determining biomolecular structures

NMR



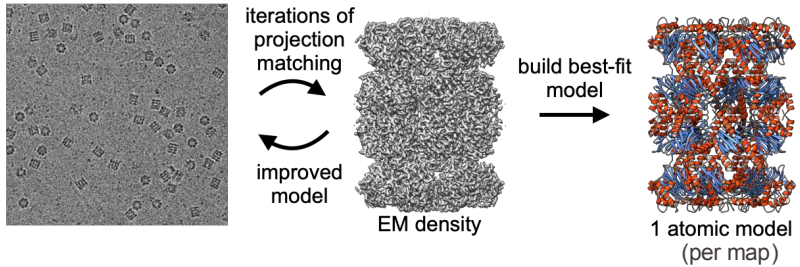
- **Versatile tool for studying protein structure and dynamics**
- Computationally light
- Full structural analysis limited to smaller proteins (<50kDa)
- Requires isotopic labeling
- Results in model ensemble

X-ray



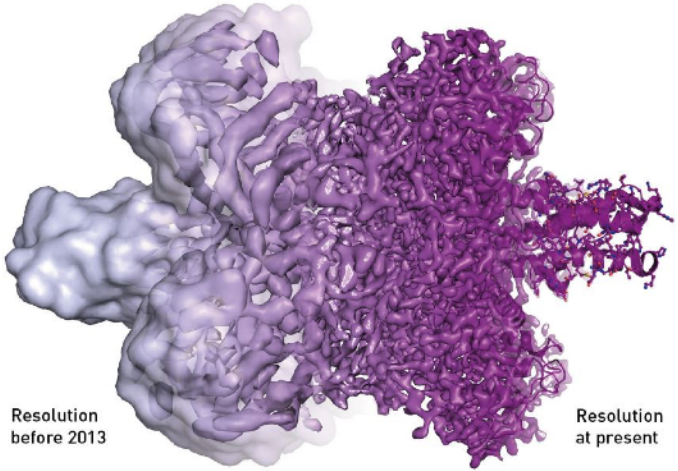
- Gold-standard method for solving protein structures
- **Not limited in size or achievable resolution**
- Computationally light
- Requires highly homogenous, crystallizable sample
- Requires screening of crystallization conditions
- Phase problem
- Results in a single model

CryoEM



- Versatile tool for studying protein assembly, structure, dynamics
- Limited to proteins >40kDa
- No requirement for protein labeling
- **Does not require homogenous samples**
- Grid preparation procedure requires screening
- Real space imaging – no phase problem
- Can be used to study protein dynamics
- Can be expanded to larger assemblies (e.g., viruses and cells)
- Results in 1 or more models per dataset
- Computationally heavy (TBs of data + requirement for GPU processing)

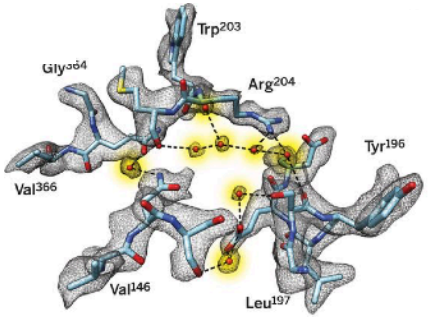
Experimental revolution — cryoEM

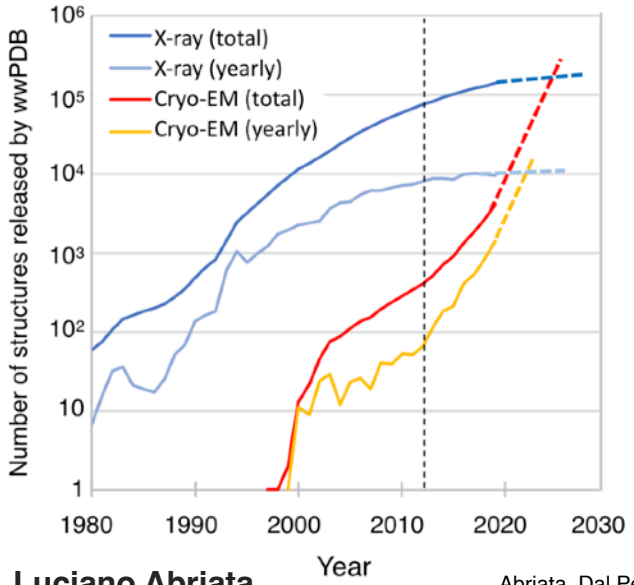


- progress in cryo-electron microscopy led to much improved resolution
- cryo-EM is becoming the new gold standard in structural biology

Krios G4 Cryo-TEM

resolution record 1.22 Å
size record ~50 kDa

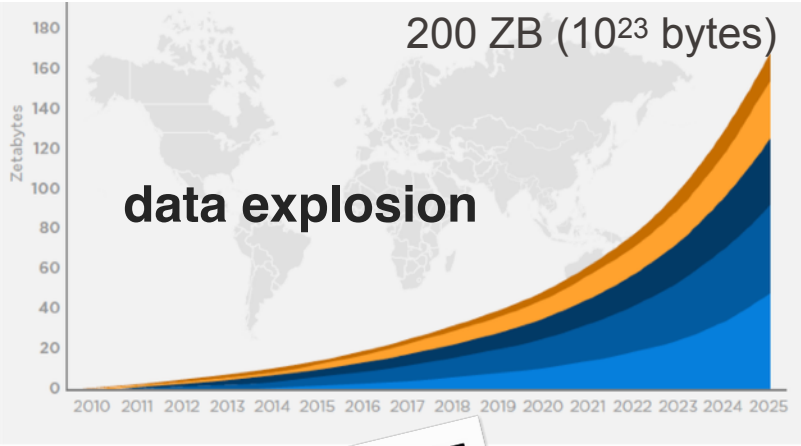




Luciano Abriata

Abriata, Dal Peraro, **JCIM** 2020

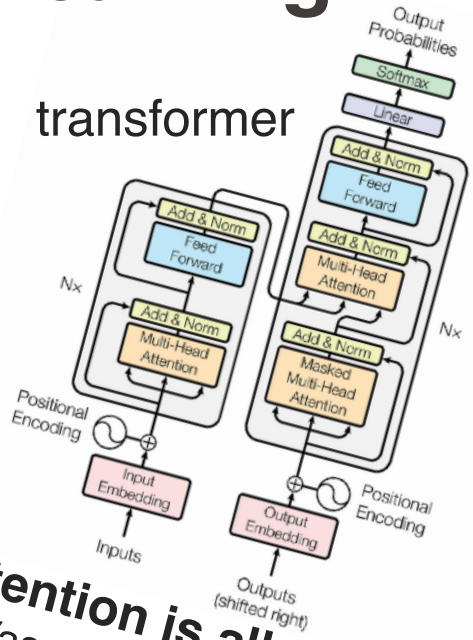
Data revolution — machine learning



powerful hardware (GPU)

ChatGPT

transformer



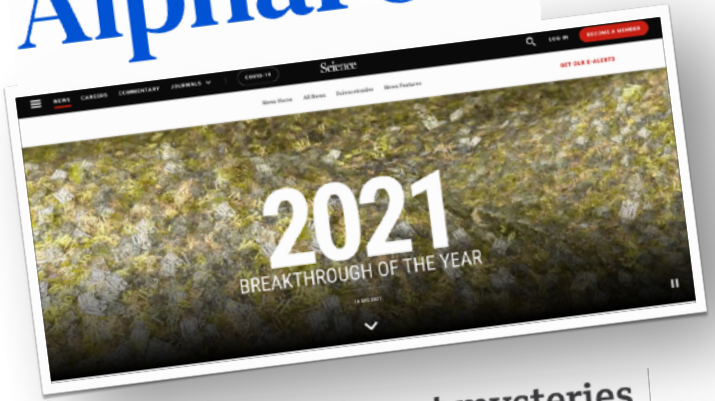
"Attention is all you need"

Vaswani et al. *arXiv* 2017

Machine learning in biology

DeepMind

AlphaFold



FOCUS | 11 JANUARY 2022

Method of the Year 2021: Protein structure prediction

Protein structure prediction is our Method of the Year 2021, for the remarkable levels of accuracy achieved by deep learning-based methods in predicting the 3D structures of proteins and protein complexes, essentially solving this long-standing challenge.

NEWS | BIOLOGY

'The game has changed.' AI triumphs at solving protein structures

In milestone, software predictions finally match structures calculated from experimental data

One of biology's biggest mysteries 'largely solved' by AI

By Helen Briggs
BBC science correspondent

NEWS | 30 November 2020

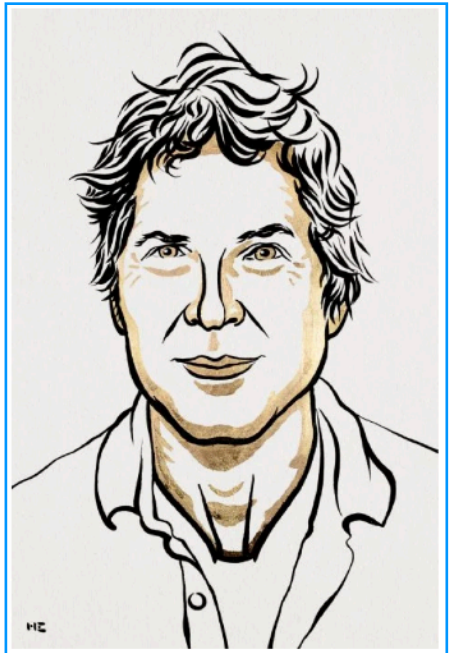
'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures

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

'THE ENTIRE PROTEIN UNIVERSE': AI PREDICTS SHAPE OF NEARLY EVERY KNOWN PROTEIN

DeepMind's AlphaFold tool has determined around 200 million protein structures, which are now available to scientists in a database.

The Nobel Prize in Chemistry 2024 was divided, one half awarded to David Baker "for computational protein design", the other half jointly to Demis Hassabis and John M. Jumper "for protein structure prediction"



Ill. Niklas Elmehed © Nobel Prize Outreach

David Baker

Prize share: 1/2

Ill. Niklas Elmehed © Nobel Prize Outreach

Demis Hassabis

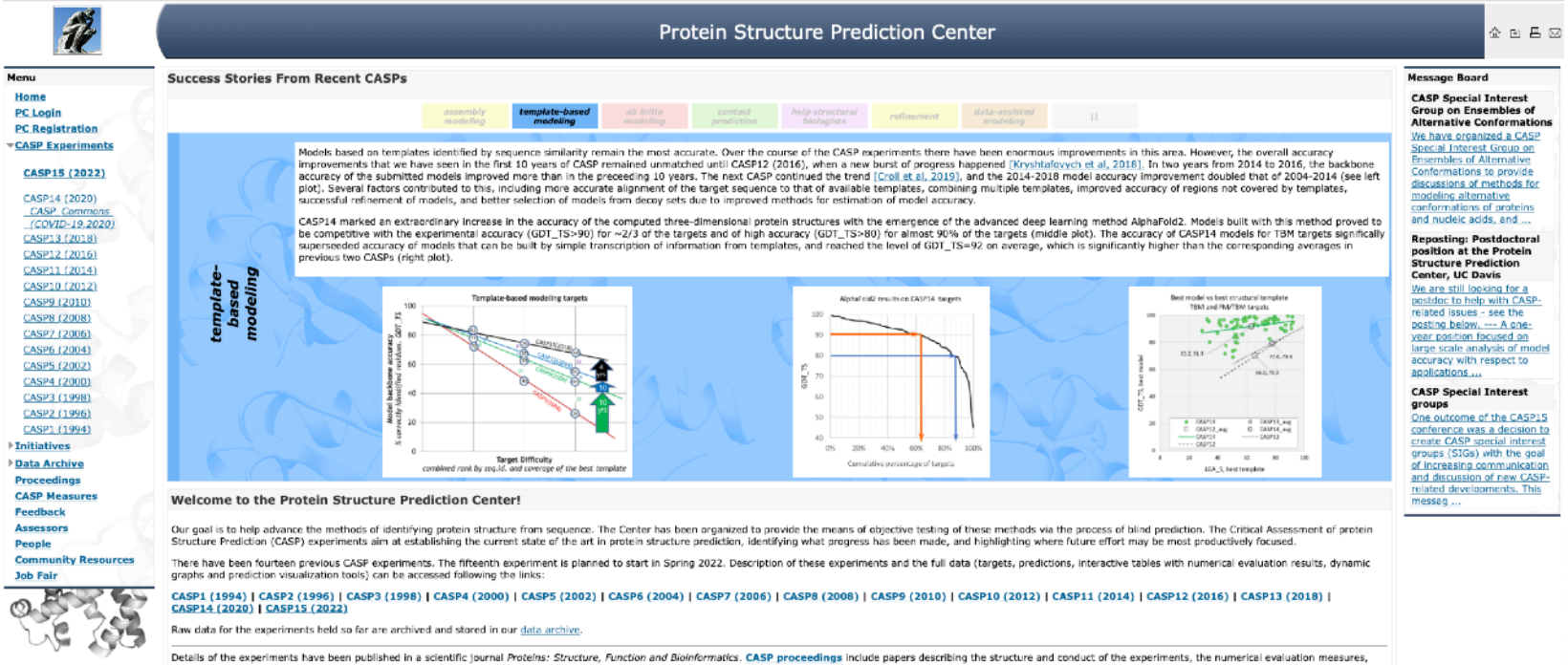
Prize share: 1/4

Ill. Niklas Elmehed © Nobel Prize Outreach

John M. Jumper

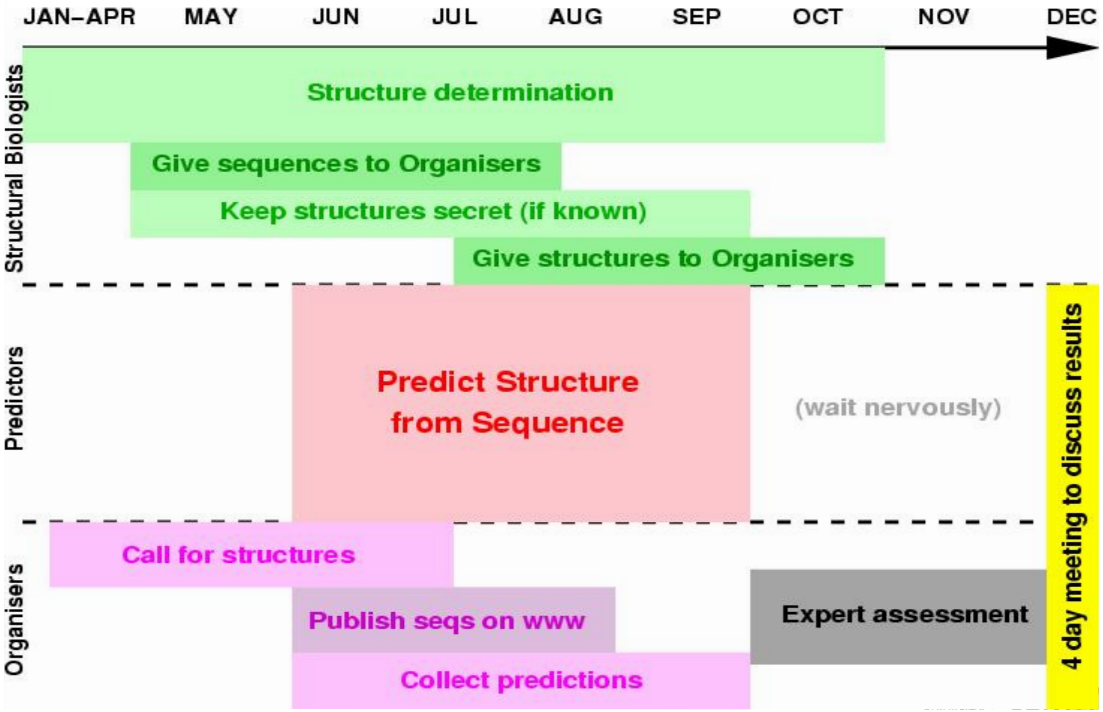
Prize share: 1/4

CASP: Critical Assessment of Techniques for Protein Structure Prediction (now CASP16)



http://predictioncenter.org

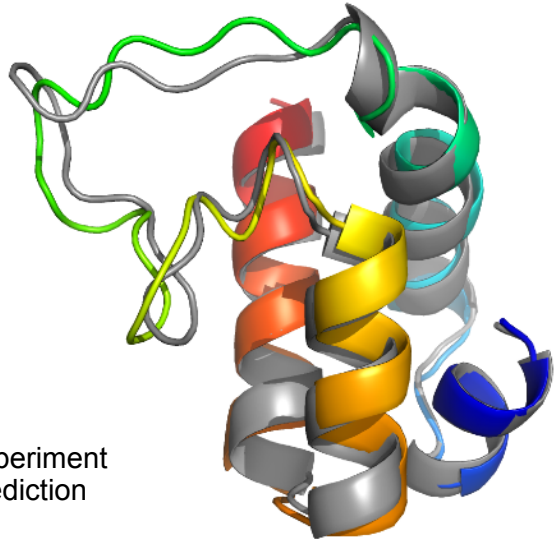
CASP: Critical Assessment of Techniques for Protein Structure Prediction (now CASP16)



http://predictioncenter.org

CASP: Critical Assessment of Techniques for Protein Structure Prediction (now CASP16)

T0990-D1
RMSD 1.6 Å



gray = experiment
rainbow = prediction

T0992-D1
RMSD 1.6 Å

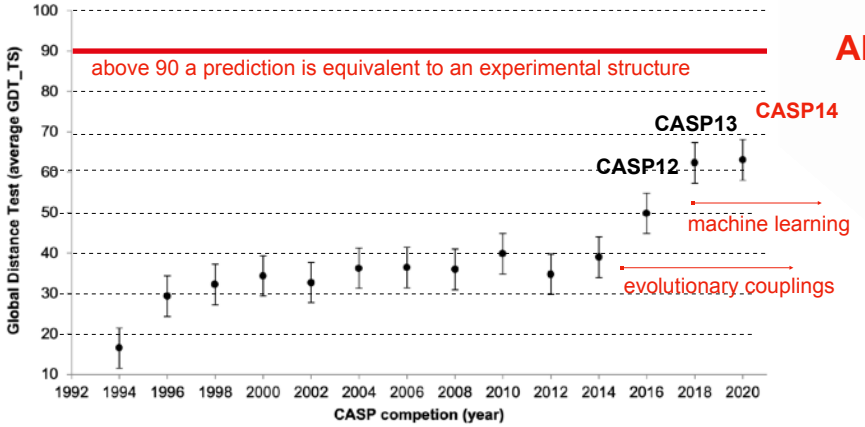
Root Mean Square Displacement :: RMSD defines a measure for similarity:

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d_i^2}$$

http://predictioncenter.org

AlphaFold2 solved a 70yo problem

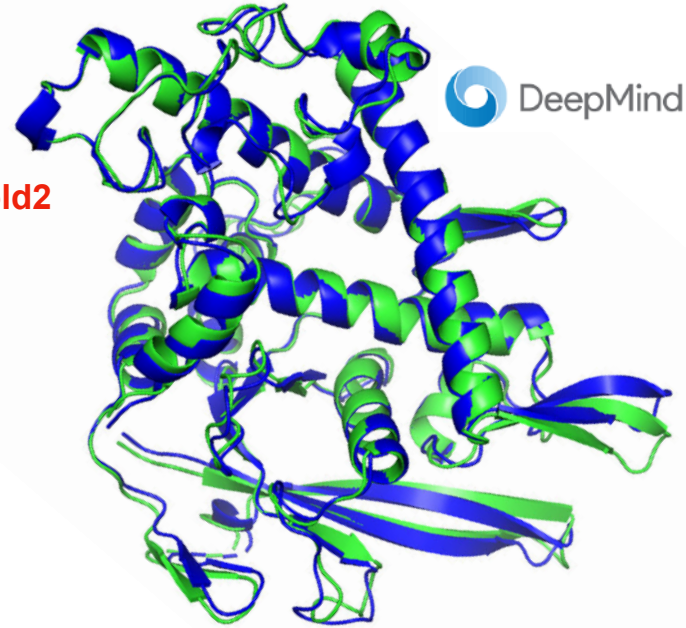
Critical Assessment of protein Structure Prediction — CASP



Abriata and Dal Peraro, **Proteins** 2019
Abriata, Tamo' and Dal Peraro, **Proteins** 2018

Luciano Abriata

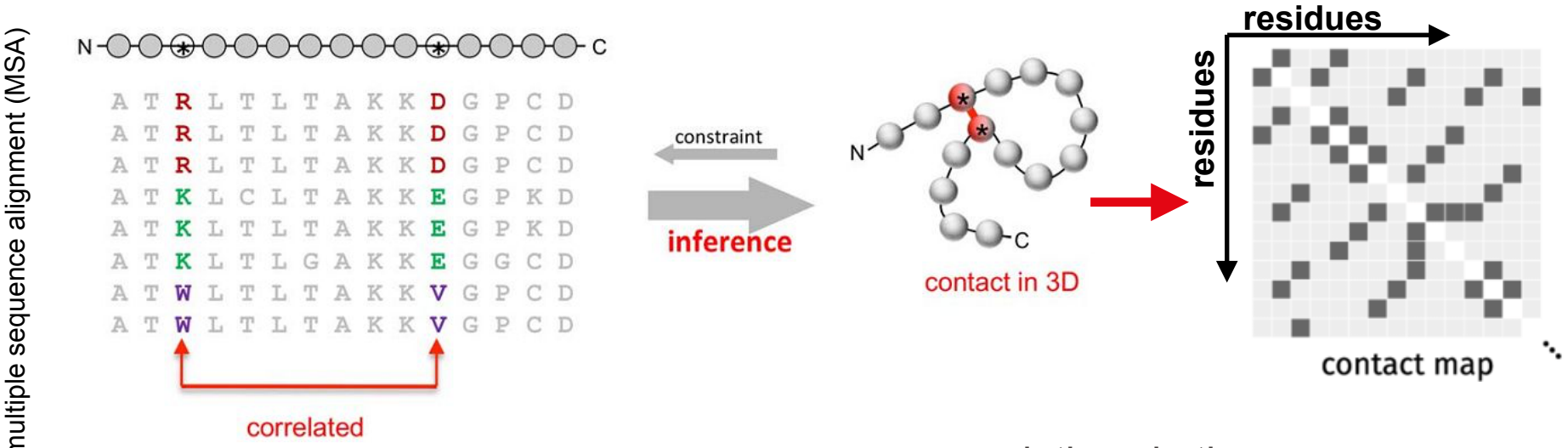
AlphaFold2



experimental structure
computational prediction

- experimental-like accuracy
- >200 M predicted models available in UniProt

Evolutionary couplings for protein prediction



Marks, D. S.; Colwell, L. J.; Sheridan, R.; Hopf, T. A.; Pagnani, A.; Zecchina, R.; Sander, C. *PLoS One* **2011**, *6*, e28766.

- correlations in the sequence space give structural information
- if you have enough predicted contacts you can fold a protein (similar to NMR)

Evolutionary couplings for protein prediction

Direct-coupling analysis (DCA)

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

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

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

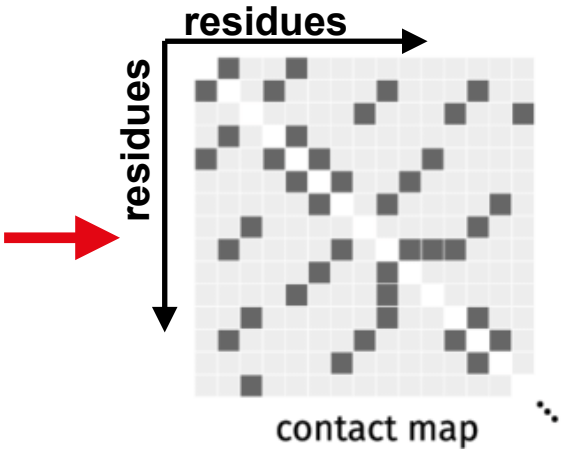
Identify maximally informative pair couplings using **statistical model** of entire protein to infer residue-residue co-evolution

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

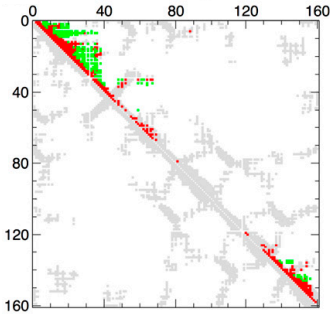
$$DI_{ij} = \sum_{A,B=1}^{q} P_{ij}^{Dir}(A,B)\ln\frac{P_{ij}^{Dir}(A,B)}{f_i(A)f_j(B)}$$

high ranking **transitive** 'indirect correlations'

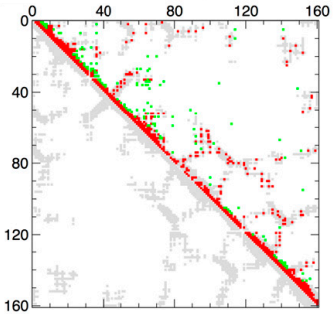
re-ranked correlations 'direct information' = DI



Mutual information



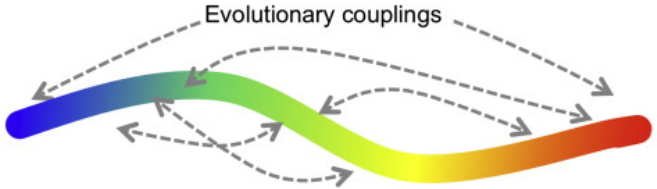
Direct information



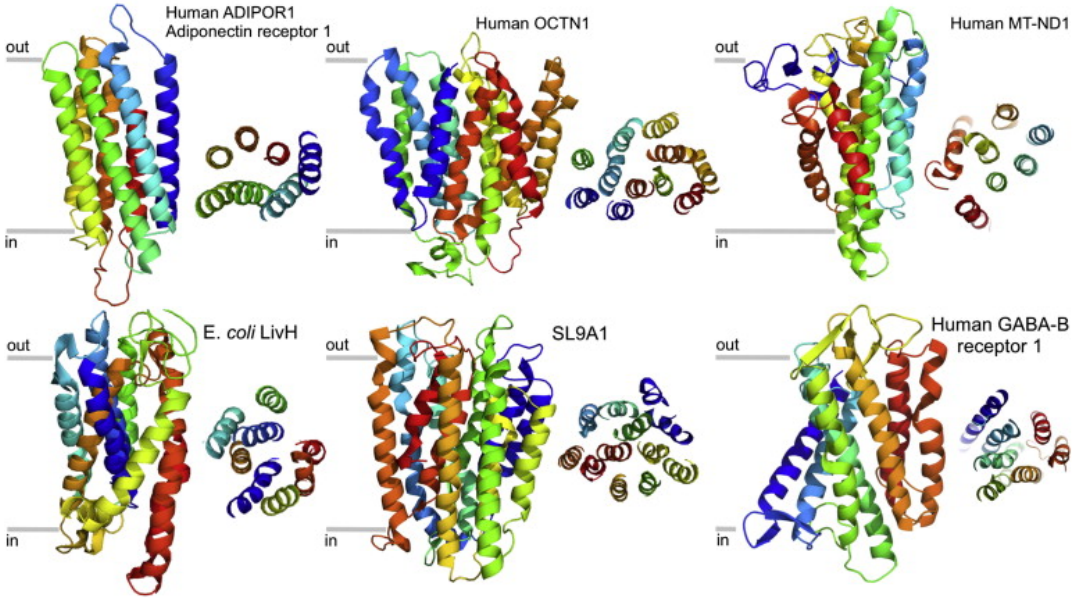
- a statistical physics method were used to crack the problem (ie Potts model)

Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino,Debora S. Marks, Chris Sander, Riccardo Zecchina, José N. Onuchic, Terence Hwa, and Martin Weigt
Direct-coupling analysis of residue coevolution captures native contacts across many protein families, PNAS 2011

EC for membrane protein prediction



- folding a transmembrane protein using EC is easier.
- Why?

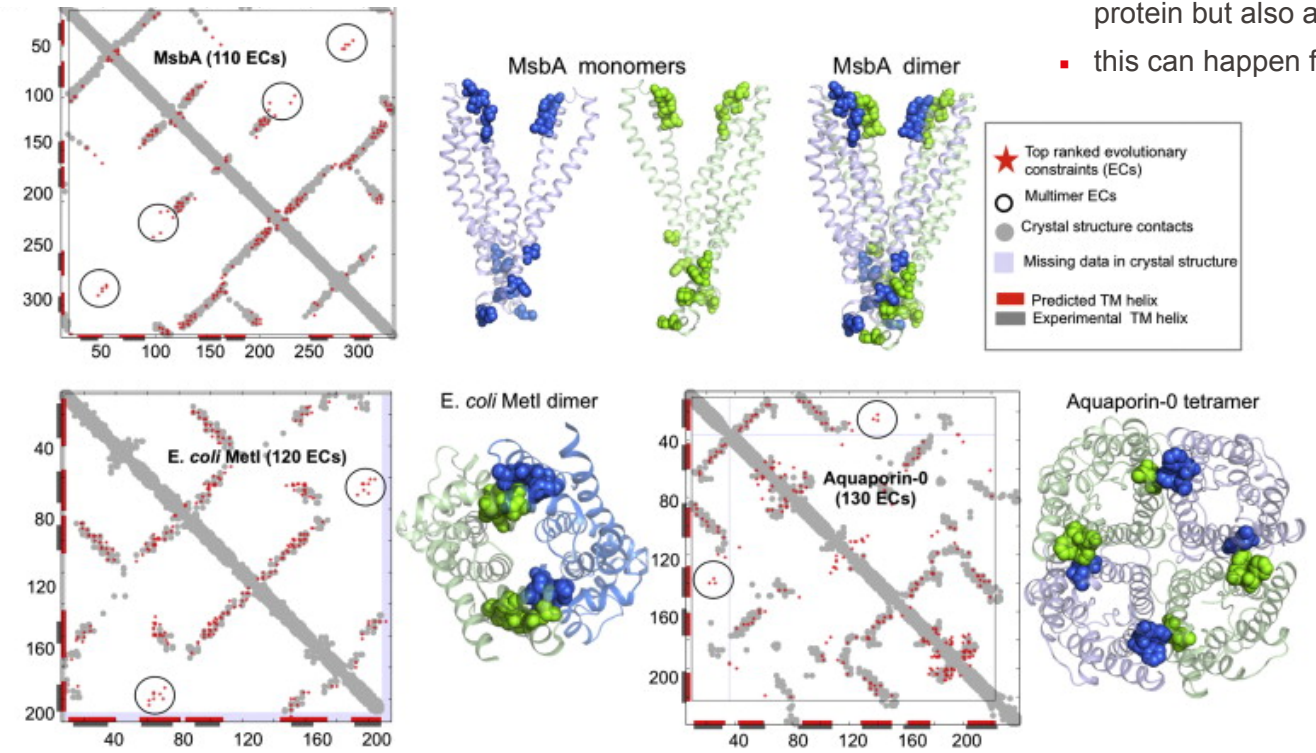


Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing

Thomas A. Hopf, Lucy J. Colwell, Robert Sheridan, Burkhard Rost, Chris Sander, Debora S. Marks, Cell 2012, DOI:https://doi.org/10.1016/j.cell.2012.04.012

EC for protein oligomerization prediction

- ECs are found not only within the same protein but also among protein interfaces
- this can happen for homo or hetero multimers

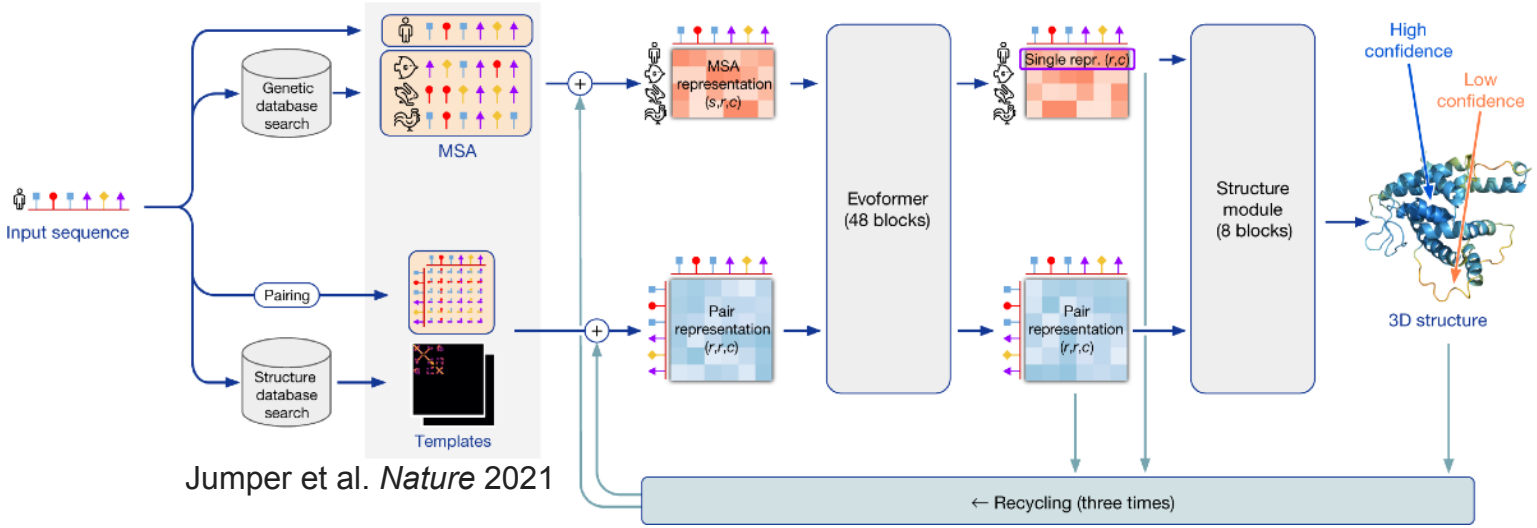


Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing

Thomas A. Hopf, Lucy J. Colwell, Robert Sheridan, Burkhard Rost, Chris Sander, Debora S. Marks, Cell 2012, DOI:https://doi.org/10.1016/j.cell.2012.04.012

AlphaFold2 architecture

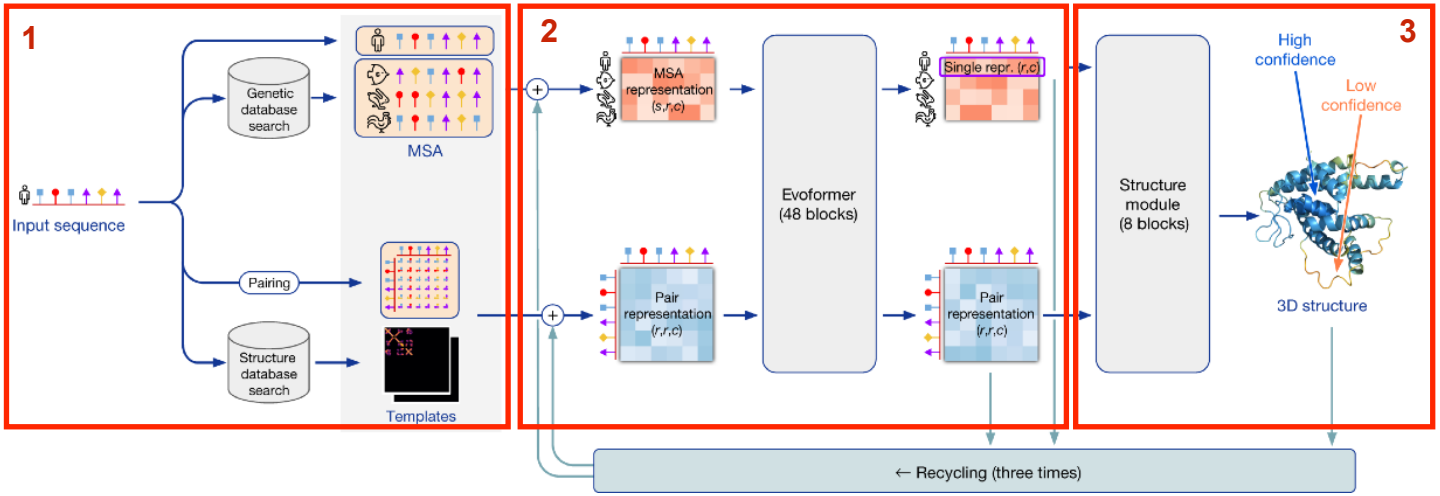
- "AlphaFold greatly improves the accuracy of structure prediction by incorporating novel neural network architectures and training procedures based on the evolutionary, physical and geometric constraints of protein structures".



Jumper et al. *Nature* 2021

- trained on sequence similarity and structural templates from databases (UniProt/metagenomics and PDB)
- end-to-end model produces prediction in one shot using *transformers*

AlphaFold2 architecture



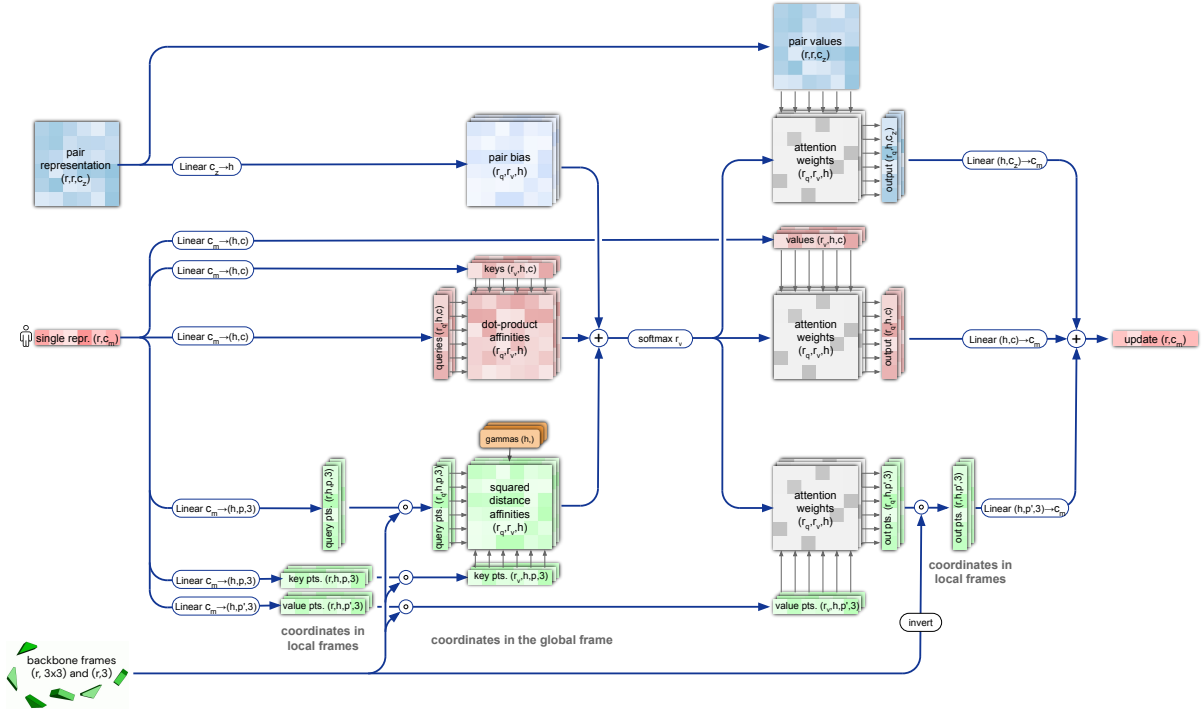
First module:
gather available information like sequence similarity (MSA) and structural templates from databases (UniProt/ metagenomics and PDB) to create a pair representation (which aa are likely in contact with each other)

Second module:
Evoformer transformer which refine the MSA and pair interactions

Third module:
The structure module build the 3D structure based on the MSA and pair interactions information

- end-to-end model produces prediction in one shot
- recycling (3X) to refine further prediction
- huge engineering effort

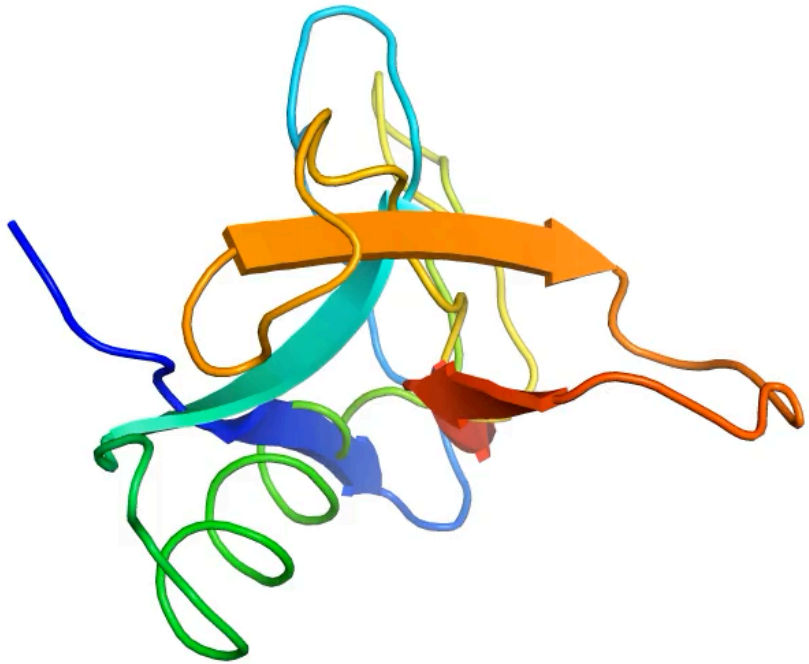
The power of (Google) engineering



Supplementary Figure 8 | Invariant Point Attention Module. (**top, blue arrays**) modulation by the pair representation. (**middle, red arrays**) standard attention on abstract features. (**bottom, green arrays**) Invariant point attention. Dimensions: r: residues, c: channels, h: heads, p: points.

Matteo Dal Peraro

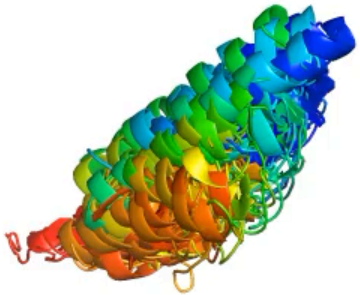
AlphaFold2 at work



Recycling iteration 0, block 01
Secondary structure assigned from the final prediction

Jumper et al. *Nature* 2021

AlphaFold2 at work



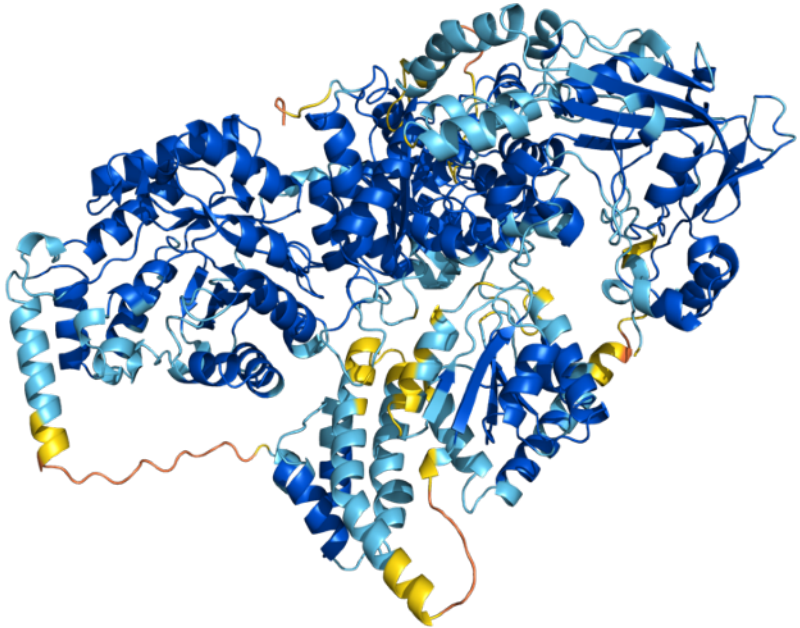
Recycling iteration 0, block 01
Secondary structure assigned from the final prediction

Jumper et al. *Nature* 2021

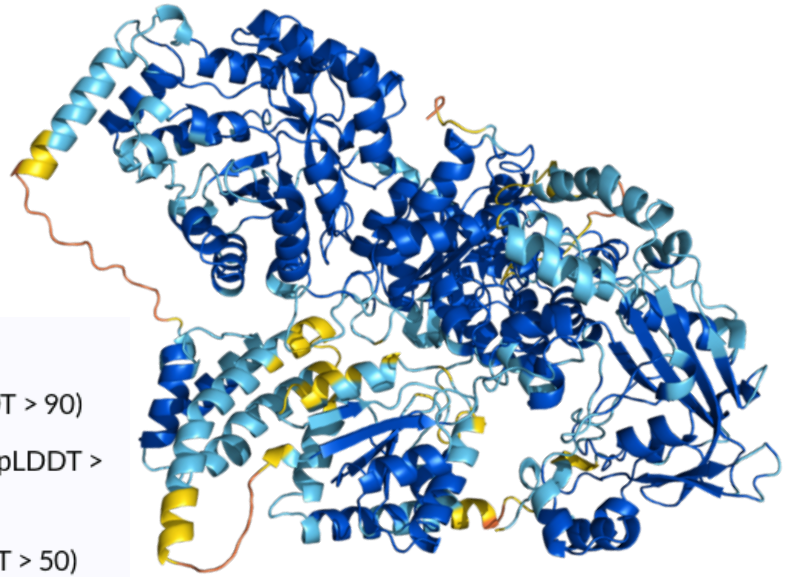
AlphaFold2 database

- > 200 million protein structure predictions
- Almost all catalogued proteins (UniProt)
- Over 1 million organisms
- Freely and openly available
- Collaboration DeepMind and EMBL-EBI
- 35.2% predictions with mean pLDDT > 90
- 79.1% predictions with mean pLDDT > 70

- https://www.alphafold.ebi.ac.uk/
- https://uniprot.org/

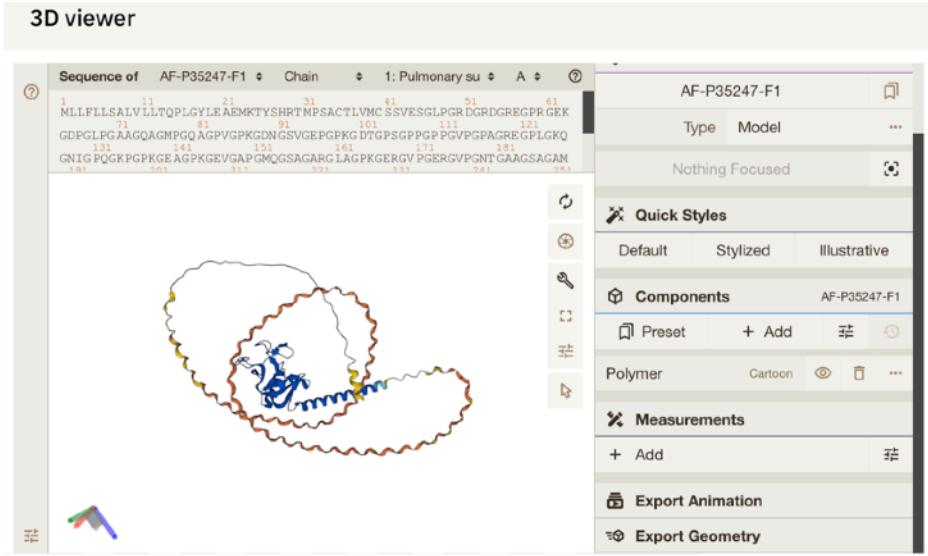


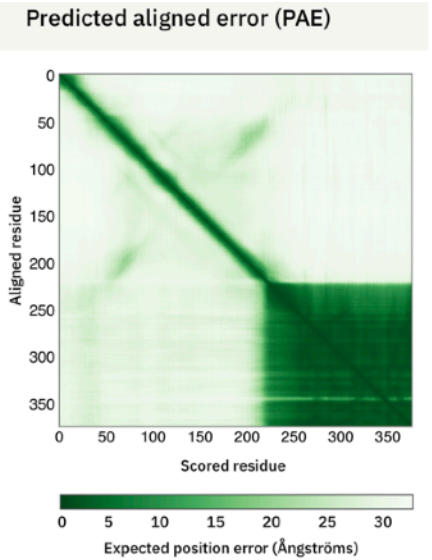
AlphaFold2 is self-assessing



- **pLDDT**: predicted local distance difference test score
- prediction of the local distances between pairs of residues in the predicted structure compared to a reference or ground truth structure.
- low score (<50) indicates that the region is disordered or AF2 does not have enough information

PAE: predicted alignment error





Model Confidence

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

AlphaFold produces a per-residue model confidence score (pLDDT) between 0 and 100. Some regions below 50 pLDDT may be unstructured in isolation.

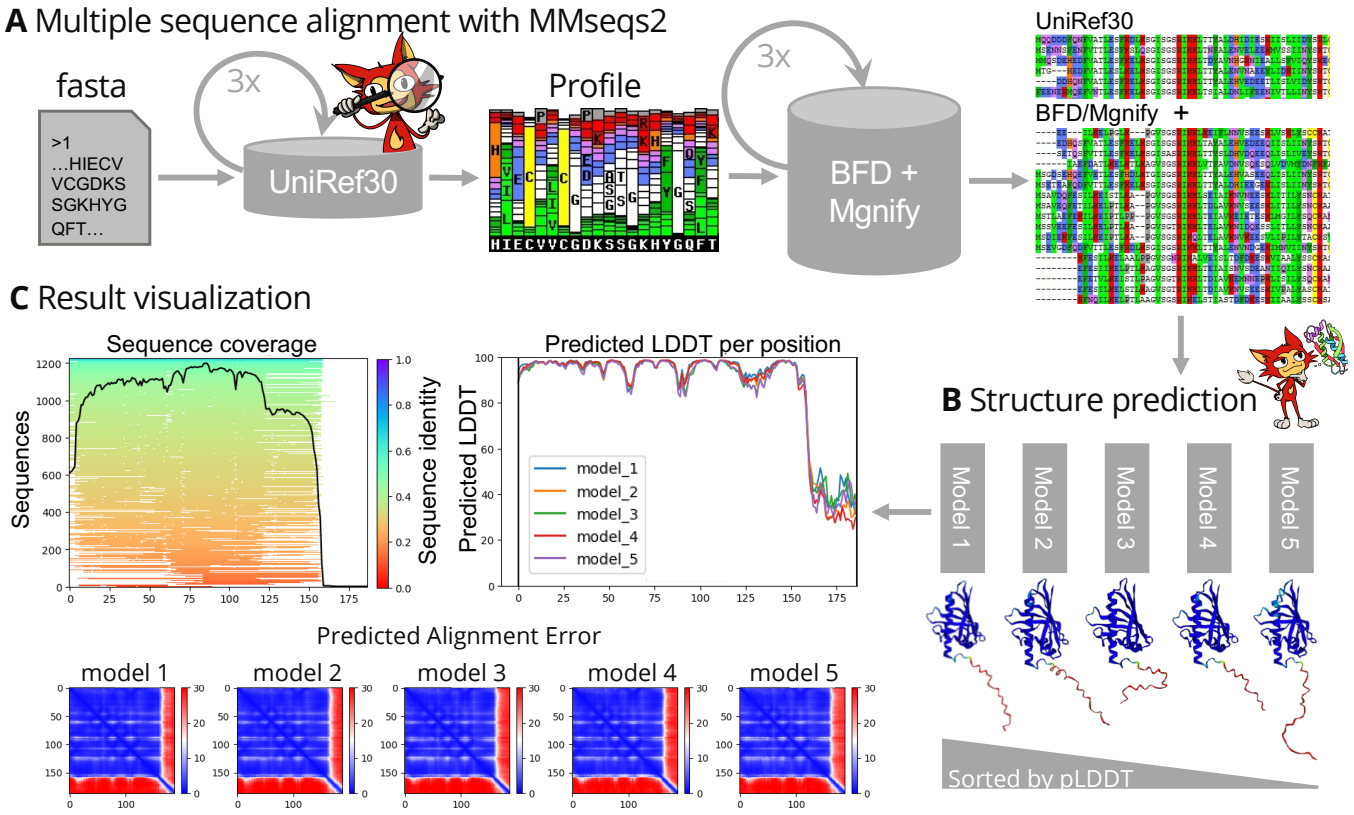
Click and drag a box on the PAE viewer to select regions of the structure and highlight them on the 3D viewer.

PAE data is useful for assessing inter-domain accuracy – go to Help section below for more information.

- pLDDT is a good score only at short/local distances
- it cannot give you good estimation of the quality of a prediction with different domains
- their reciprocal orientation cannot be estimated by pLDDT
- PAE is the solution for this scenario

ColabFold - Making protein prediction accessible to all

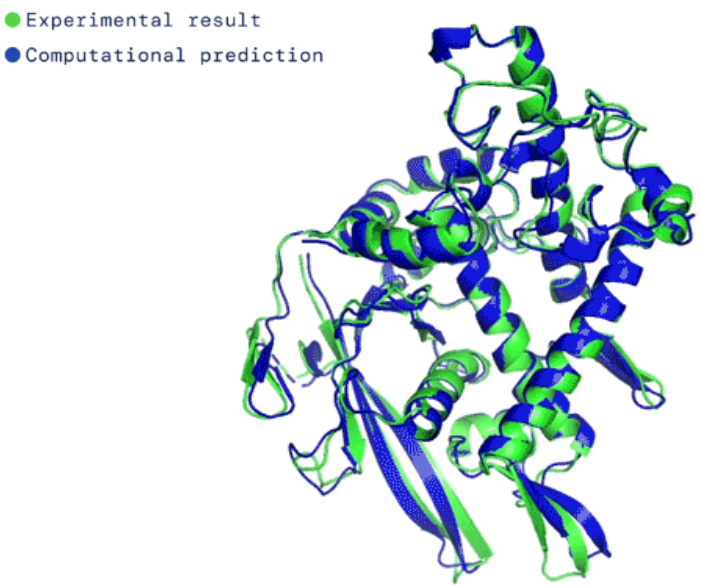
Matteo Dal Peraro



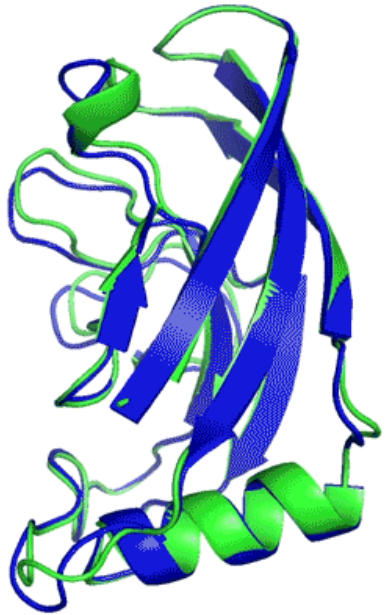
Nature Methods | VOL 19 | JUNE 2022 | 679–682 |

Matteo Dal Peraro

AlphaFold2 is a new SB methods



T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)



T1049 / 6y4f
93.3 GDT
(adhesin tip)

~200K (PDB)

>200M (UniProt)

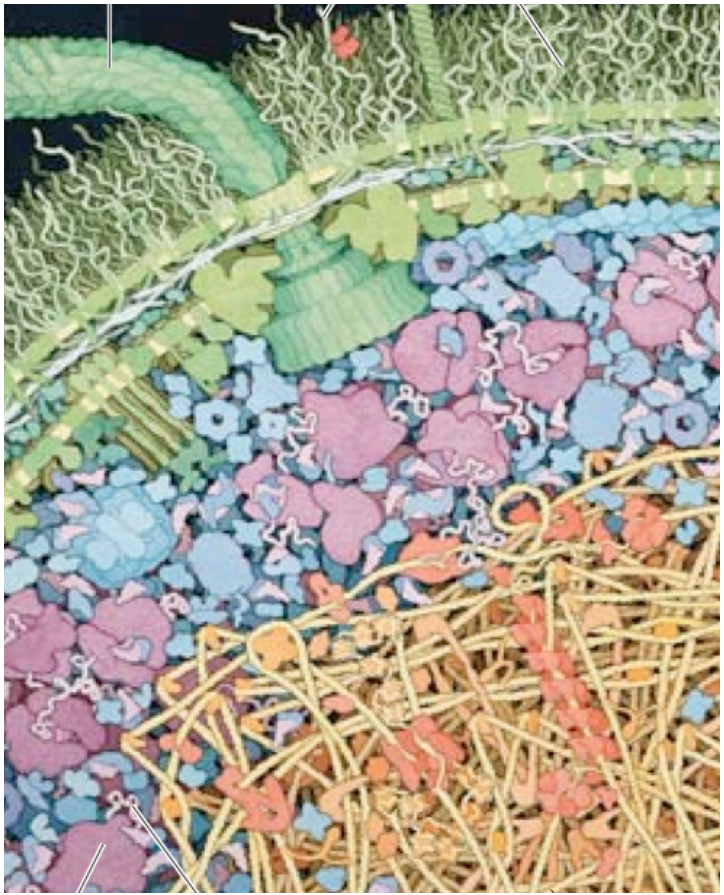
Jumper et al. *Nature* 2021

AlphaFold2 is not the ultimate oracle

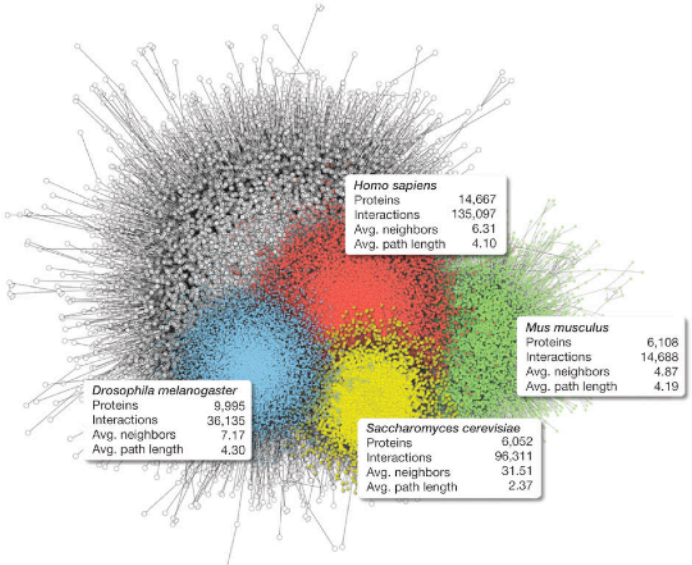
- **Limitations**
  - need for a deep MSA (>30 sequences) to create accurate models
  - not all the models are highly accurate as an experimental structure
  - it does not account for dynamics and multiple states
  - does not account for the post-translational modifications

- **Other potential benefits**
  - AF triggered many other developments
  - can assist experimental structure determination
  - eg, in molecular replacement in X-ray crystallography
  - eg, in cryoEM fitting and model building
  - it is a means to look at protein-protein networks

Proteins form assemblies and networks

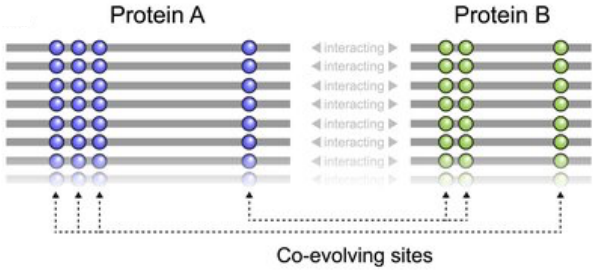


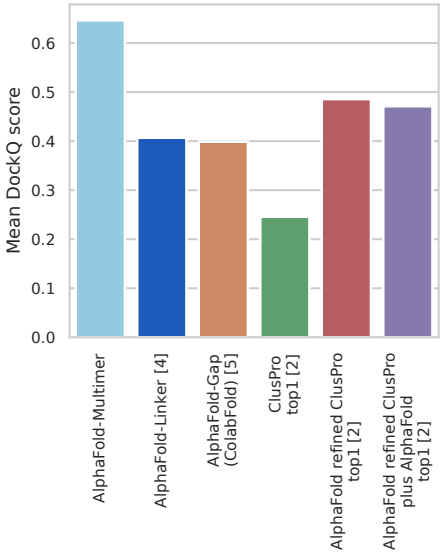
- Proteins can then arrange into supramolecular assemblies
- Interacting with nucleic acid, metabolites, membranes, etc.
- They create large network of interactions



AlphaFold Multimer

coevolution coupling holds for protein-protein interactions





+ **Support for multiple chains**
+ Multi-chain features
+ Various architectural modifications
+ Paired MSAs
+ Training on complexes

Protein complex prediction with AlphaFold-Multimer
Richard Evans,…, John Jumper, Demis Hassabis
bioRxiv 2021.10.04.463034; doi: https://doi.org/10.1101/2021.10.04.463034

nature structural & molecular biology

Article

https://doi.org/10.1038/s41594-022-00910-8

Towards a structurally resolved human protein interaction network

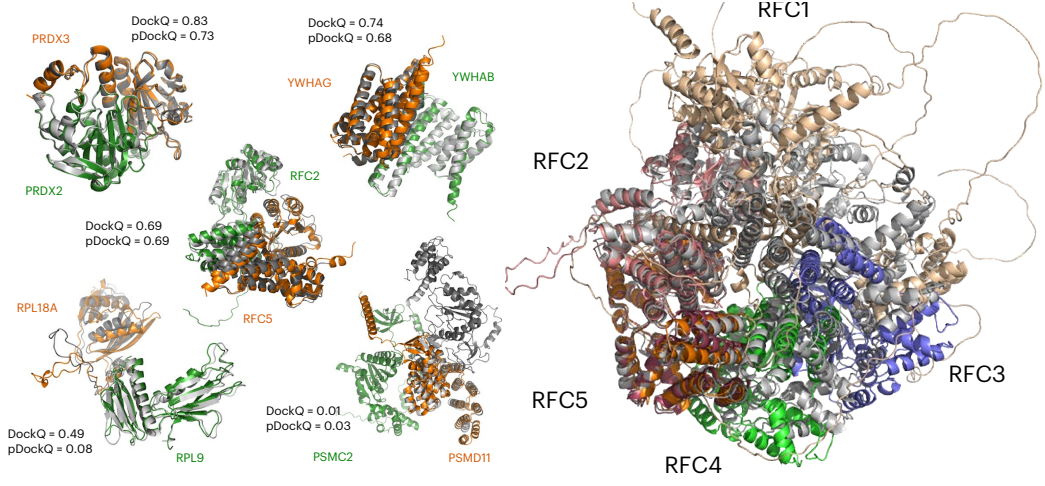
Received: 11 February 2022

Accepted: 14 December 2022

Published online: 23 January 2023

Check for updates

David F. Burke[1,9], Patrick Bryant[2,3,9], Inigo Barrio-Hernandez[1,9], Danish Memon[1,9], Gabriele Pozzati[2,3,9], Aditi Shenoy[2,3], Wensi Zhu[2,3], Alistair S. Dunham[1], Pascal Albanese[4,5], Andrew Keller[6], Richard A. Scheltema[4,5], James E. Bruce[6], Alexander Leitner[7], Petras Kundrotas[2,3,8] ✉, Pedro Beltrao[1,7] ✉ & Arne Elofsson[2,3] ✉



nature microbiology

Article

https://doi.org/10.1038/s41564-024-01791-x

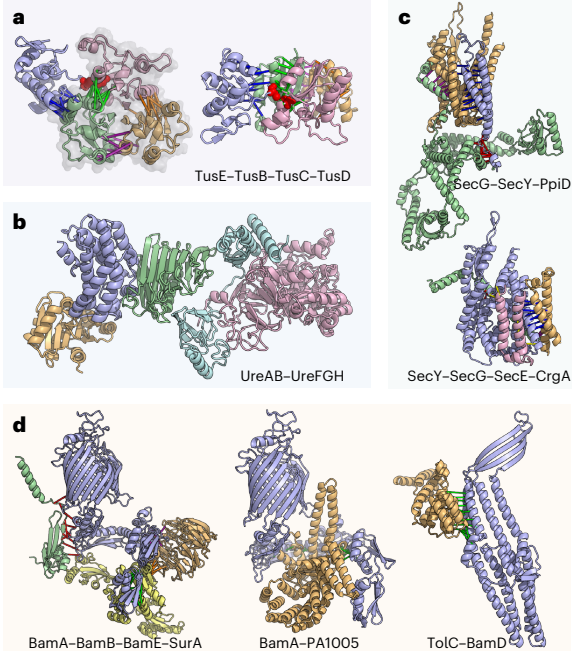
Protein interactions in human pathogens revealed through deep learning

Received: 28 April 2023

Accepted: 23 July 2024

Published online: 18 September 2024

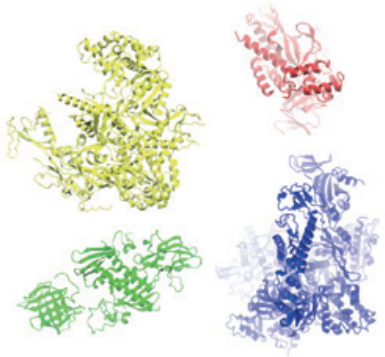
Ian R. Humphreys[1,2,13], Jing Zhang[3,4,5,13], Minkyung Baek[6,13] ✉, Yaxi Wang[7,13], Aditya Krishnakumar[1,2], Jimin Pei[3,4,5], Ivan Anishchenko[1,2], Catherine A. Tower[7], Blake A. Jackson[7], Thulasi Warrier[8,9,10], Deborah T. Hung[8,9,10], S. Brook Peterson[7], Joseph D. Mougous[7,11,12], Qian Cong[3,4,5] ✉ & David Baker[1,2,11] ✉



Integrative structural biology

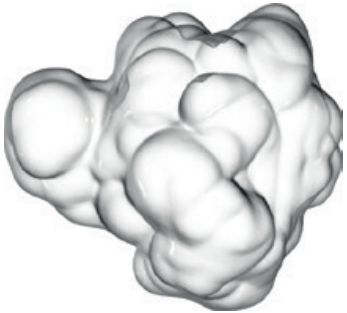
Individual subunits

- X-ray crystallography
- NMR
- Cryo-EM
- AI-based predictions



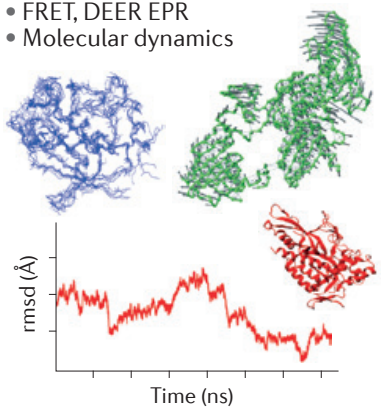
Volumetric maps

- Cryo-EM
- Electron tomography
- SAXS, SANS
- AFM



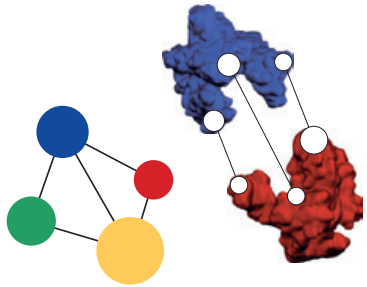
Structural flexibility

- Side-chain and backbone sampling
- Elastic network models
- NMR ensembles
- FRET, DEER EPR
- Molecular dynamics



Spatial connectivity

- Mutagenesis
- Evolutionary couplings
- Chemical crosslinking
- Proteomics
- H/D exchange
- ChIP–seq and ChIP–exo
- 3C, 4C, 5C and Hi-C

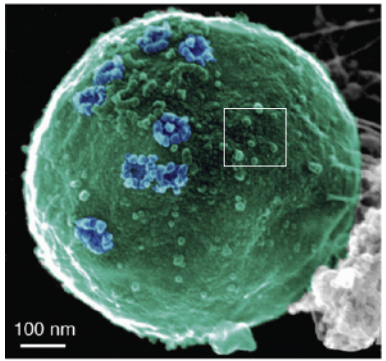


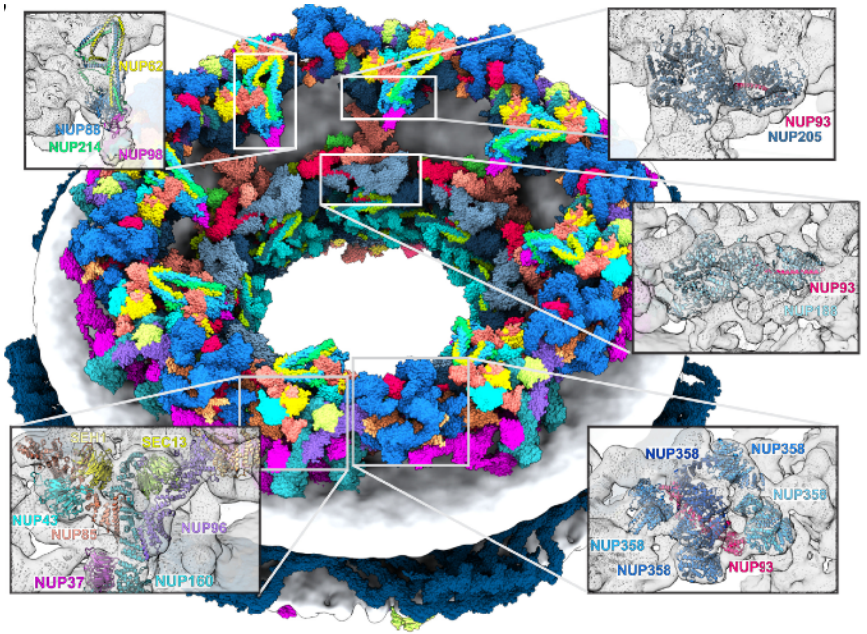
data integration and model building

Matteo Dal Peraro

AF2 affects integrative structural biology

e.g., the Nuclear Pore Complex





- **yeast:** ~52 MDa, ~550 proteins
- **human:** ~120 MDa, ~100 proteins

Mosalaganti et al. *Science* 2022

https://alphafoldserver.com

AlphaFold Server

Powered by AlphaFold 3

Abramson, J et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature (2024)

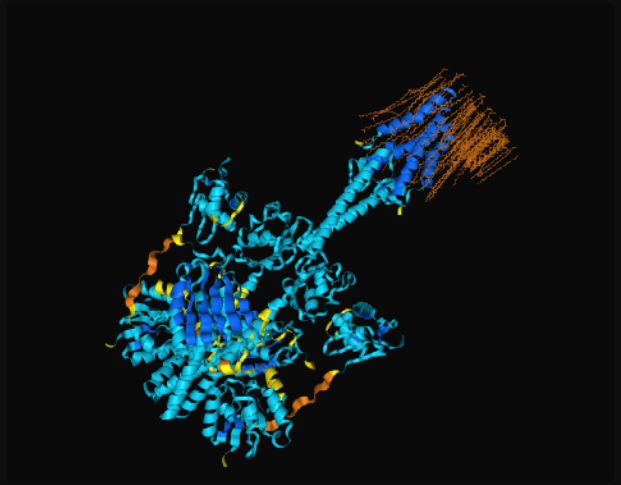
AlphaFold Server BETA Server About FAQs

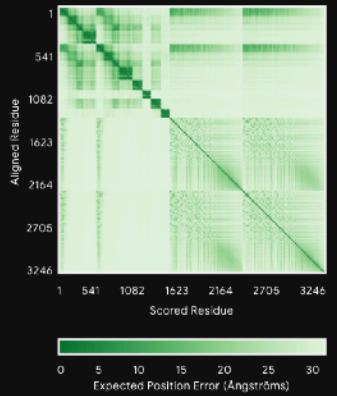
PhoR-PhoB

← Back Download Clone and reuse Feedback on structure

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

ipTM = 0.45 pTM = 0.46 learn more



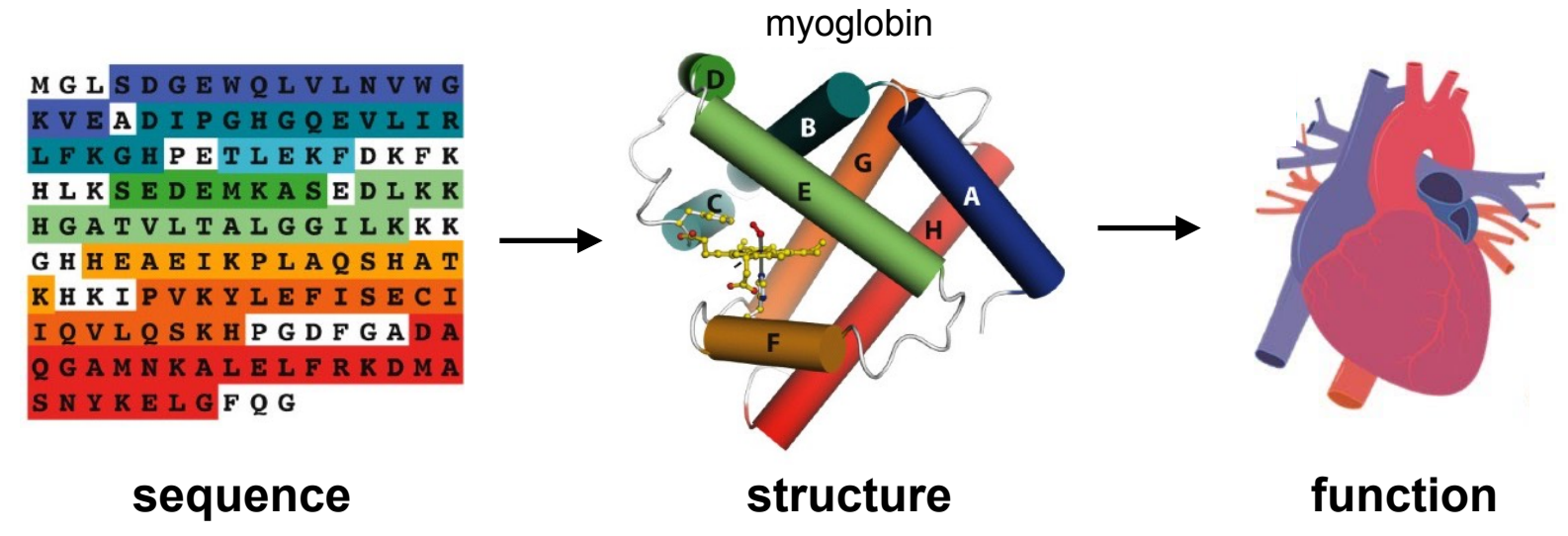


Type	Copies	Sequence
Protein	2	MLERLSWKRL VLELLLCCLP AFILGAFFGY LPWFLLASV RSMTPPPGRG SWEPLLYGLH QMQLRNKKRR RELGNLIKR FWCNGLAQQI LGLRWPEDNG QNILNLLRYP EFTQYLKTR PYTHKQLLMV ARDVTQMHQL EGARRNFFAN VSHELRTPL KALHTMREQT QRMEGLVKQL LTLSKIEAAP THLLNEKVD FTFEIDNGLK VSGNEDQLRS AISNLVYNAV NHTPEGTHI IAPEHIPRLT ERFYRVDKAR SRQTGGSGLG LAIVKHAVN IPERLIAKNS D PTMs 213H: ND1-Phosphohistidine
Ligand	50	PLM – Palmitic acid
Ligand	50	OLA – Oleic acid
Protein	2	MARRILVVED EAPIREMVCF VLEQNGFQPV EAEDYDSAV GIQFIKHLKR ESMTRDIPVV MLTARGEEED RVRGLETGA RRISPMAVEE VIEMQGLSLD PTSHRVMAGE EPLEMGPTE NHVWGTNVYV EDRTVDVHIR RLRKALEPGG HDRMVQTVR

Seed: 1026411006

The folding paradigm

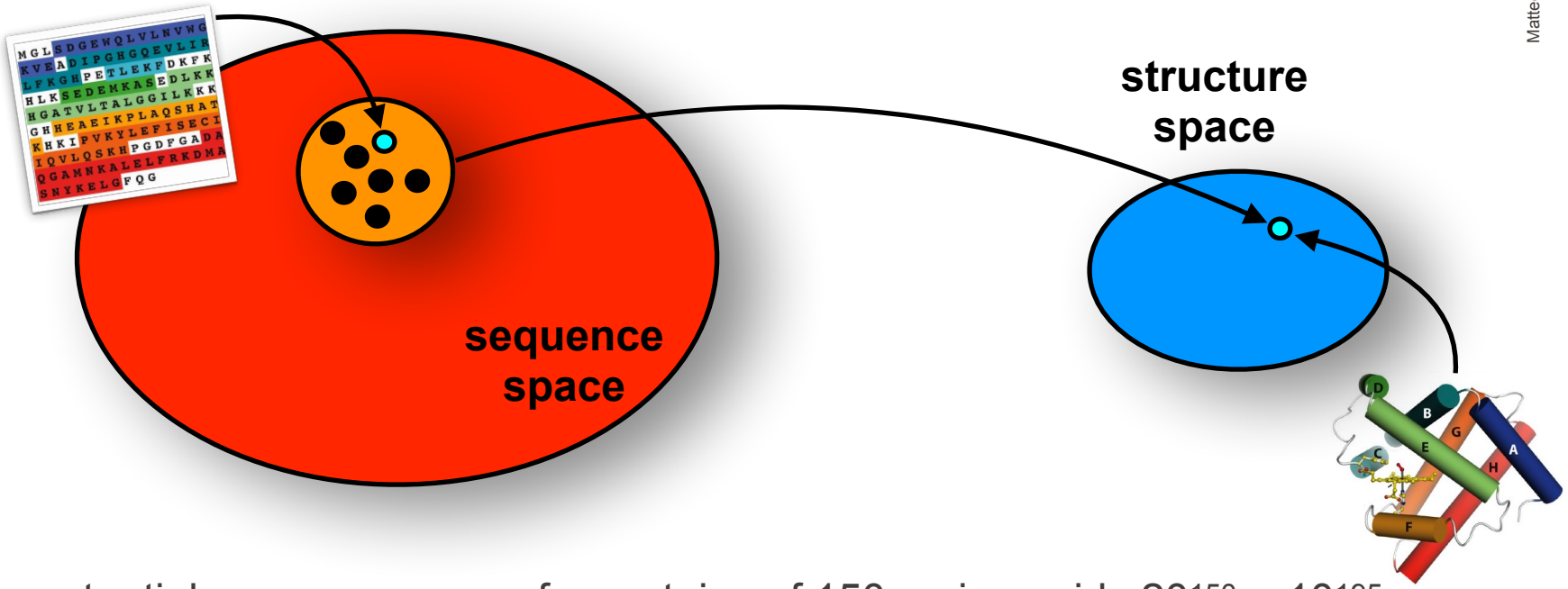
Matteo Dal Peraro



- Prediction of final structure and binding helps discovering new biology
- Not all the questions are answered though by AF2 !!

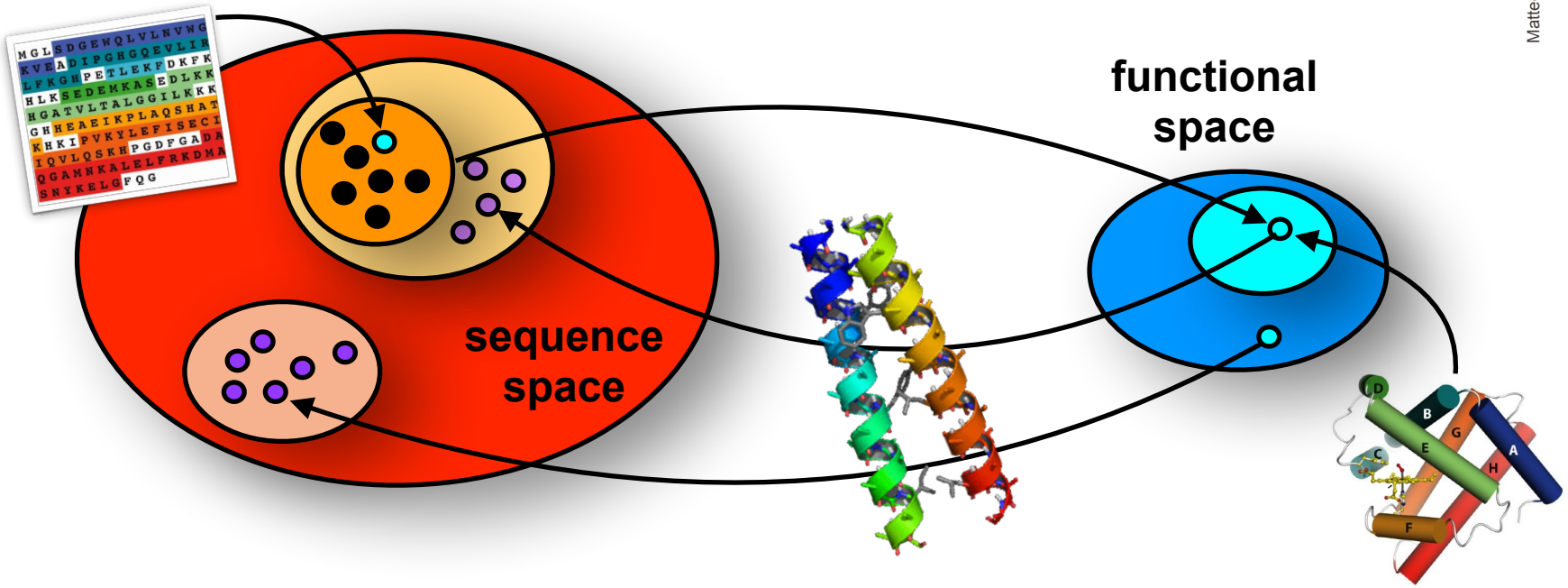
The sequence space is enormous

Matteo Dal Peraro



- potential sequence space for proteins of 150 amino acids $20^{150} \sim 10^{195}$
- atoms in the observed universe $\sim 10^{80}$
- the sequences explored by evolution are much less ($\sim 10^{10\text{-}20}$), structures lesser

The inverse folding problem — design



- Application to study protein evolution and function
- Protein engineering for therapeutics, synthetic biology and (bio)technology

Matteo Dal Peraro

The origins: the Paracelsus challenge (‘94)

- Rose and Creamer: convert a protein to another fold changing no more than 50% of its sequence

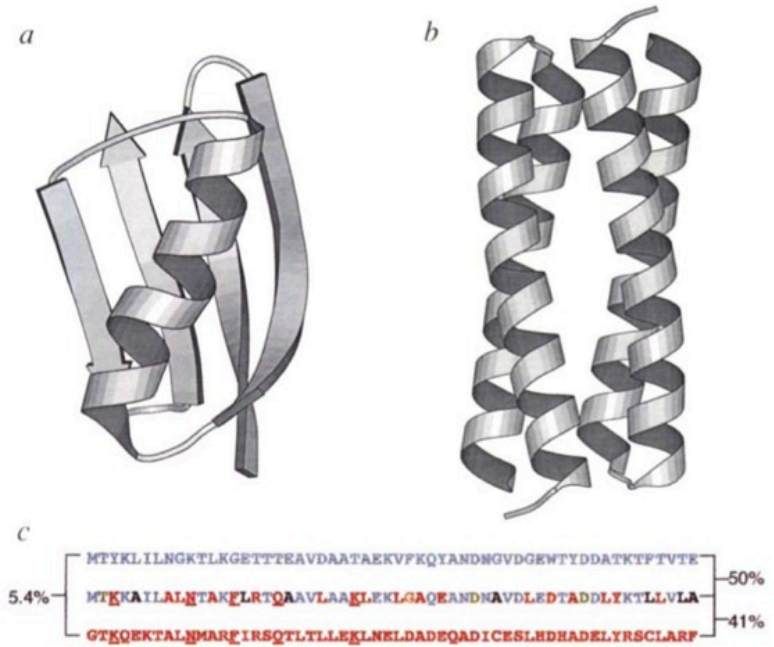


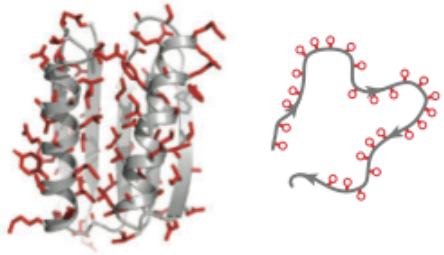
Fig. 1 Ribbon representation[29] of the folds of ***a***, the B1 domain of IgG-binding protein G[5] and ***b***, Rop[6]. ***c***, An alignment of the sequences of the B1 domain (blue), Rop (red) and Janus. Residues in Janus are coded as follows: blue, residues from B1; red, residues from Rop; underlined red, RNA-binding residues in Rop[13]; green, residues that are conserved in both Rop and B1; black, ‘a’ and ‘d’ position residues that are different from those in wild-type Rop; orange, the first residue of the turn between Helix 1 and Helix 2. The D30G mutation was introduced in the turn of Janus because a previous study demonstrated that this point mutation increases the stability of Rop[30]. The percent identity between the different sequences are indicated. The seven amino acid, unstructured C-terminal tail of Rop (Gly–Asp–Asp–Gly–Glu–Asn–Leu) extends beyond the sequence depicted for both Rop and Janus and is also not shown in (*b*). It was retained in Janus because it increases the solubility of wild type Rop[31].

Dalal et al., Protein alchemy: Changing β-sheet into α-helix, Nature Structural & Molecular Biology 1997

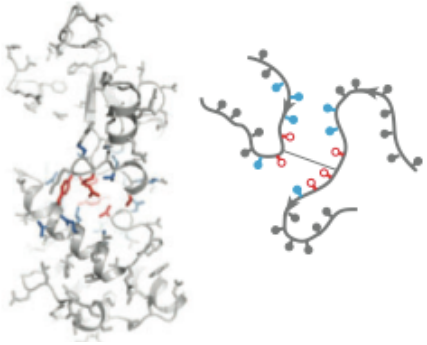
Multiple tasks for protein design

- create de novo proteins
- explore new folds
- embed new functions

Protein design



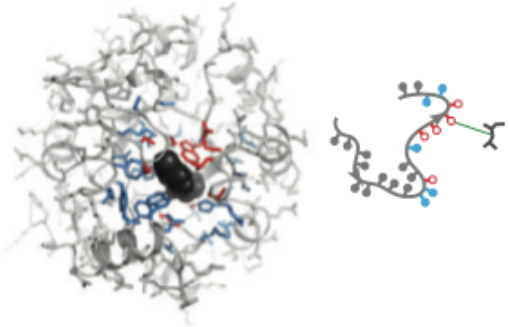
Protein-protein interface design



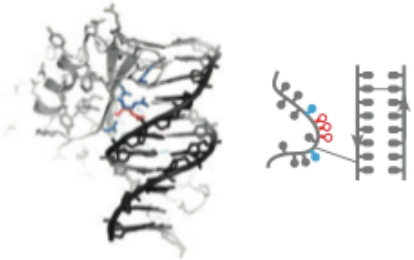
- create high affinity binders
- therapeutic biologics
- artificial sensors/probes

- tailor enzymatic function
- improve thermostability
- catalyse new reactions

Enzyme design



Protein-DNA interface design



- explore DNA interactions
- new therapeutic solutions

• Filled colored circles - flexible side chains
○ empty colored circles – flexible amino acid: design

Machine learning for protein design

RESEARCH ARTICLES

Science

Cite as: J. Dauparas *et al.*, *Science* 10.1126/science.add2187 (2022)

Robust deep learning–based protein sequence design using ProteinMPNN

Article

De novo design of protein structure and function with RFdiffusion

Baker lab - UW

Chroma - **Generate**:Biomedicines

Article

Illuminating protein space with a programmable generative model

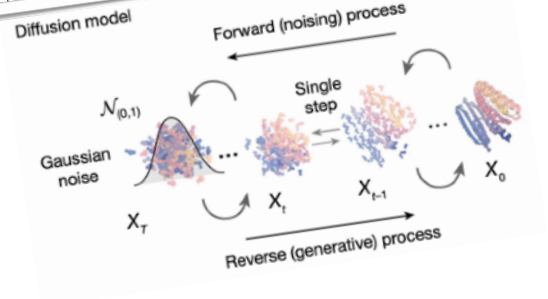
https://doi.org/10.1038/s41586-023-06728-8

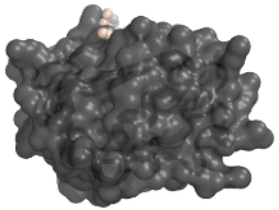
Received: 20 December 2022

Accepted: 6 October 2023

Published online: 15 November 2023

John B. Ingraham[1], Max Baranov[1], Zak Costello[1], Karl W. Barber[1], Wujie Wang[1], Ahmed Ismail[1], Vincent Frappier[1], Dana M. Lord[1], Christopher Ng-Thow-Hing[1], Erik R. Van[illegible], Vincent Xue[1], Sarah C. Cowles[1], Alan Leung[1], João V. Rodrigues[1], Claudio L. Morales-Perez[1], Alex M. Ayoub[1], Robin Green[1], Katherine [illegible], Fritz Obermeyer[1], Adam R. [illegible]





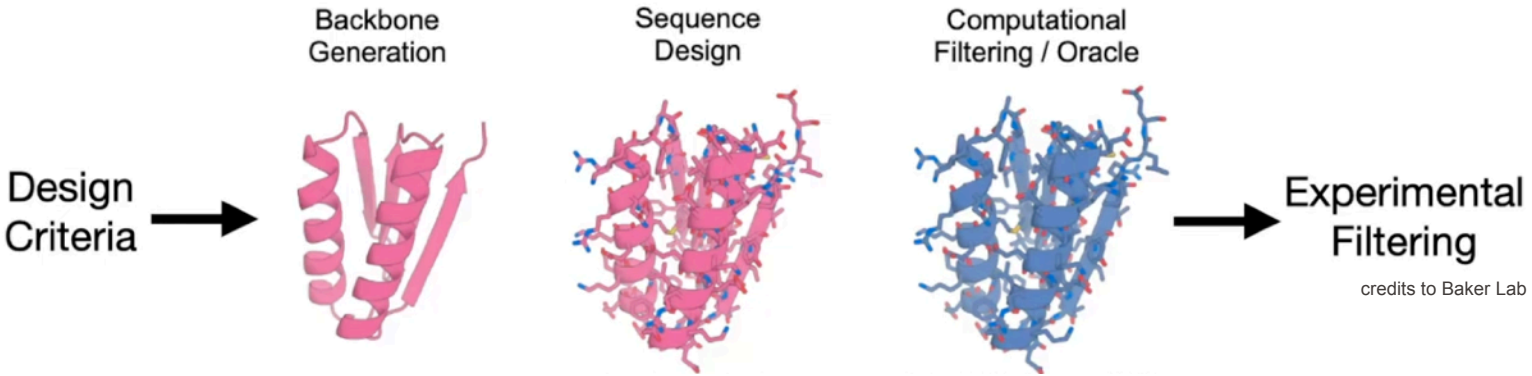
- Application to study protein evolution and function
- Protein engineering for therapeutics, synthetic biology and biotechnology

… leading to molecular engineering

- Application to study protein evolution and function
- Protein engineering for therapeutics, synthetic biology and biotechnology

Matteo Dal Peraro

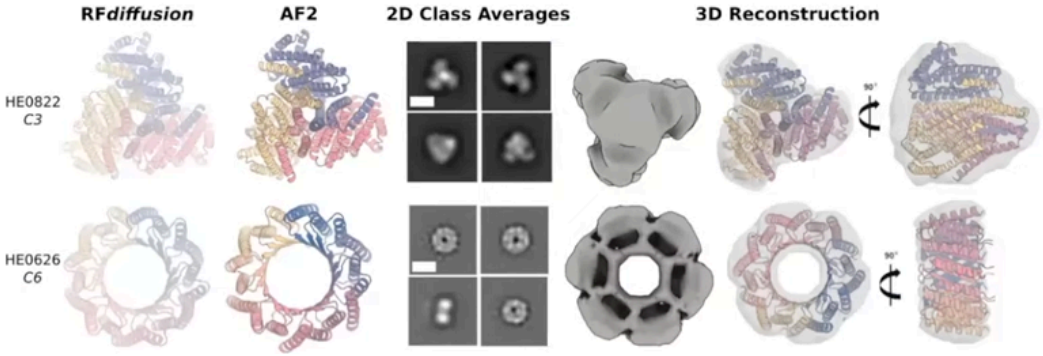
Pipeline of today's protein design



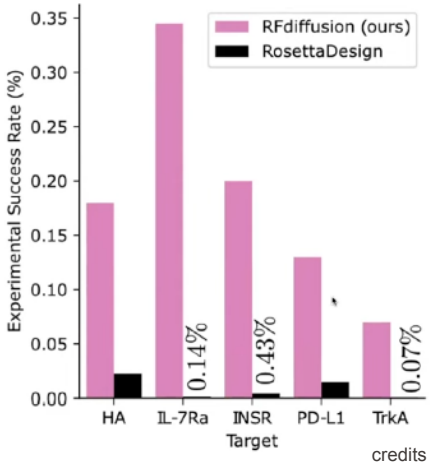
- AF2 has been key to filter potentially good protein designs
- Experimental testing is the ultimate validation of designs
- AI methods enhanced the experimental rate of success
- Protein engineering is now feasible for therapeutics, synthetic biology and biotechnology

Matteo Dal Peraro

Pipeline of today's protein design



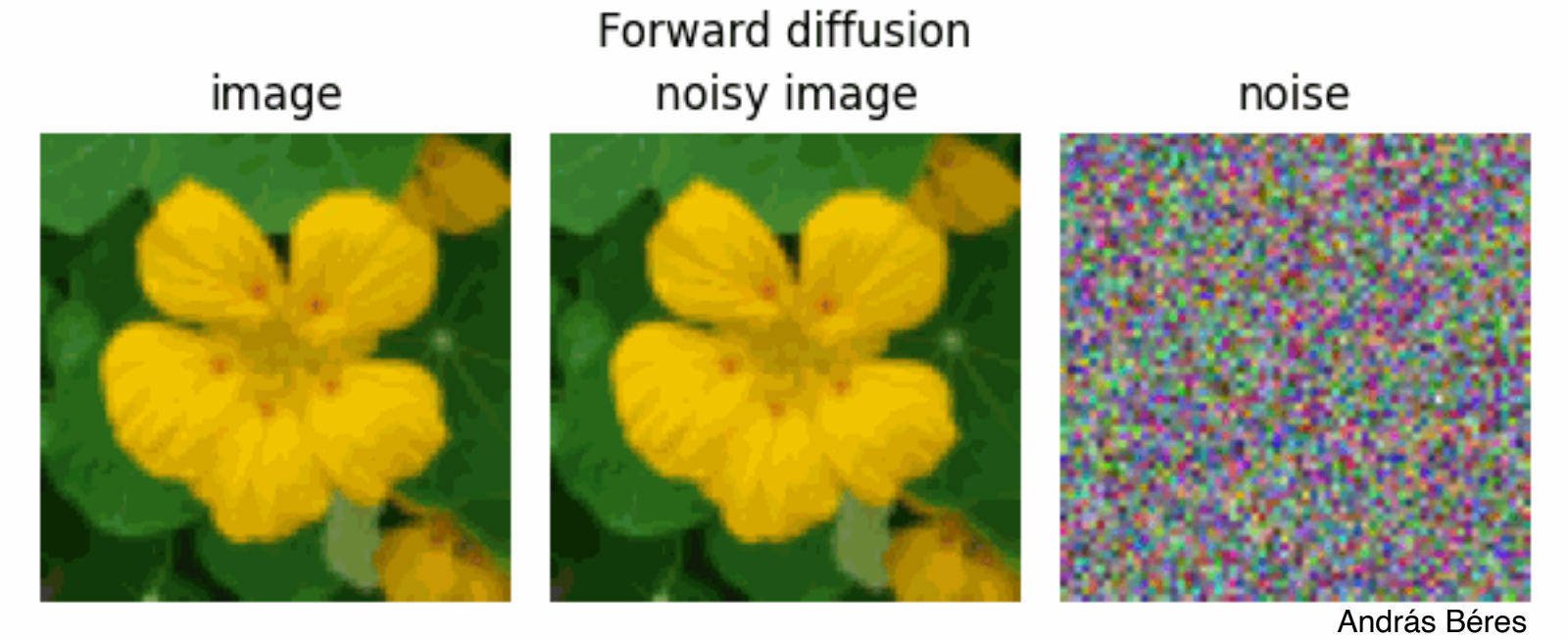
validation using cryoEM



credits to Baker Lab

- AF2 has been key to filter potentially good protein designs
- Experimental testing is the ultimate validation of designs
- AI methods enhanced the experimental rate of success
- Protein engineering is now feasible for therapeutics, synthetic biology and biotechnology

Machine learning for protein design



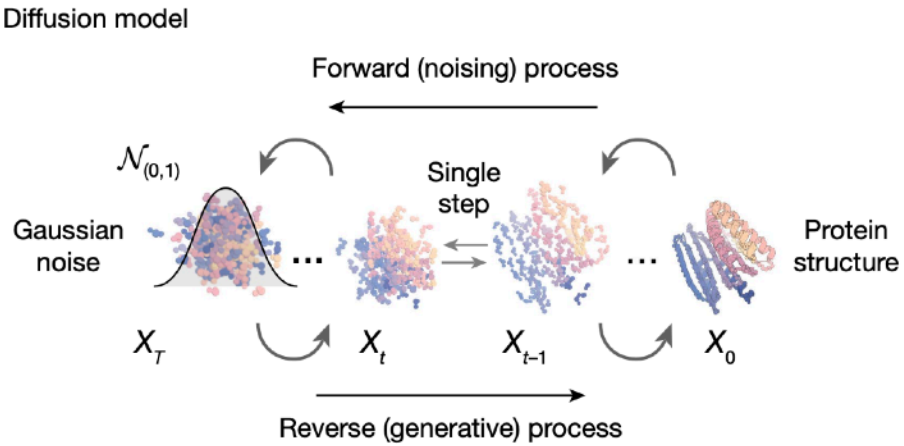
- Denoising Diffusion Models - as those used in DALL-E
- Trained to denoise noisy images, they can generate images by iteratively denoising pure noise

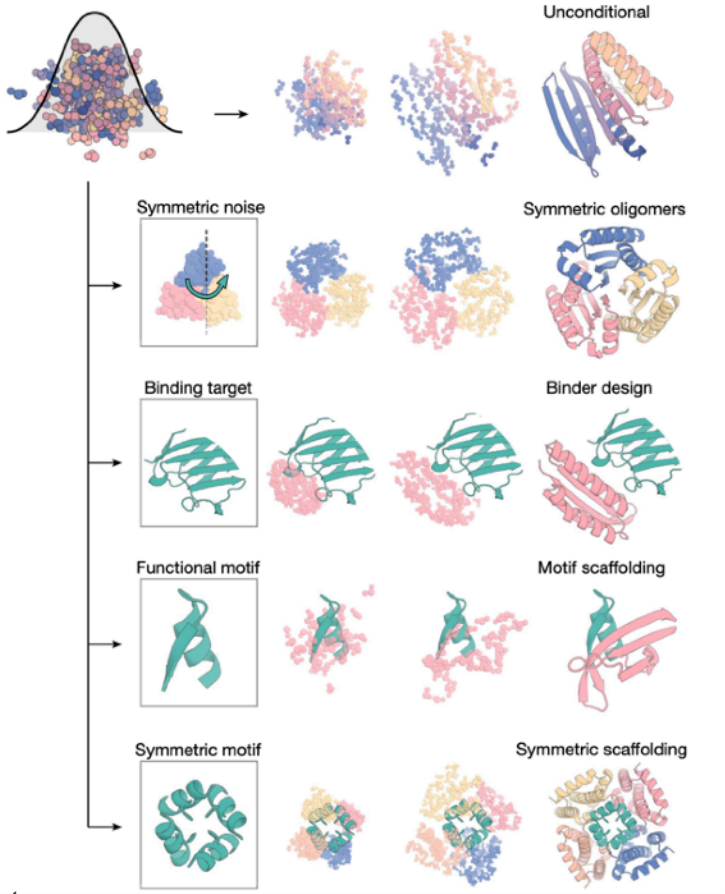
Article

De novo design of protein structure and function with RFdiffusion

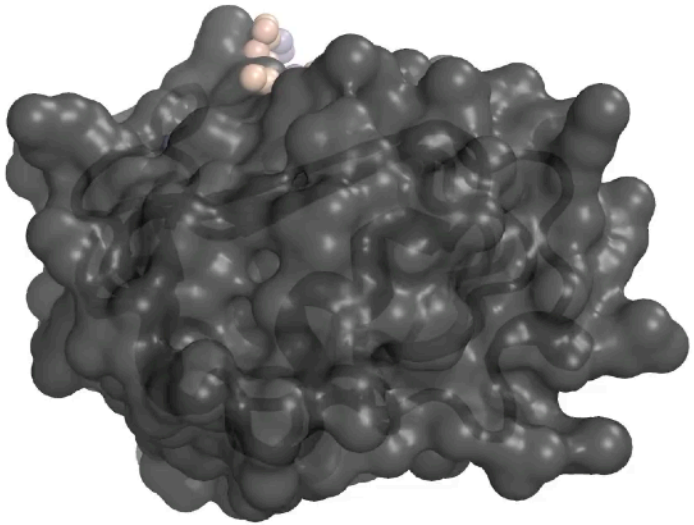
https://doi.org/10.1038/s41586-023-06415-8
Received: 14 December 2022
Accepted: 7 July 2023
Published online: 11 July 2023
Open access

Joseph L. Watson[1,2,15], David Juergens[1,2,3,15], Nathaniel R. Bennett[1,2,3,15], Brian L. Trippe[2,4,5,15], Jason Yim[2,6,15], Helen E. Eisenach[1,2,15], Woody Ahern[1,2,7,15], Andrew J. Borst[1,2], Robert J. Ragotte[1,2], Lukas F. Milles[1,2], Basile I. M. Wicky[1,2], Nikita Hanikel[1,2], Samuel J. Pellock[1,2], Alexis Courbet[1,2,8], William Sheffler[1,2], Jue Wang[1,2], Preetham Venkatesh[1,2,9], Isaac Sappington[1,2,9], Susana Vázquez Torres[1,2,9], Anna Lauko[1,2,9], Valentin De Bortoli[8], Emile Mathieu[10], Sergey Ovchinnikov[11,12], Regina Barzilay[6], Tommi S. Jaakkola[6], Frank DiMaio[1,2], Minkyung Baek[13] & David Baker[1,2,14]✉





- the reverse process is learned using a neural network
- its loss function encourages the reverse process to accurately estimate how the data transitions from one noisy step to the previous step.



https://www.bakerlab.org/2022/11/30/diffusion-model-for-protein-design/

Cite as: J. Dauparas *et al.*, *Science* 10.1126/science.add2187 (2022).

Robust deep learning–based protein sequence design using ProteinMPNN

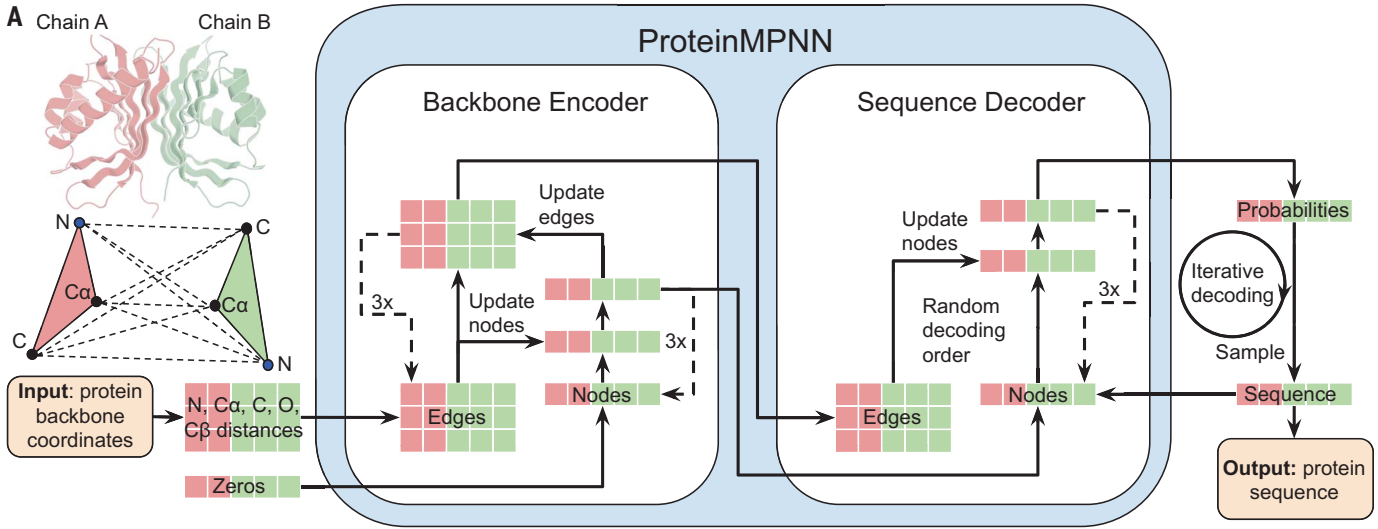
J. Dauparas[1,2], I. Anishchenko[1,2], N. Bennett[1,2,3], H. Bai[1,2,4], R. J. Ragotte[1,2], L. F. Milles[1,2], B. I. M. Wicky[1,2], A. Courbet[1,2,4], R. J. de Haas[5], N. Bethel[1,2,4], P. J. Y. Leung[1,2,3], T. F. Huddy[1,2], S. Pellock[1,2], D. Tischer[1,2], F. Chan[1,2], B. Koepnick[1,2], H. Nguyen[1,2], A. Kang[1,2], B. Sankaran[6], A. K. Bera[1,2], N. P. King[1,2], D. Baker[1,2,4]*

[1]Department of Biochemistry, University of Washington, Seattle, WA, USA. [2]Institute for Protein Design, University of Washington, Seattle, WA, USA. [3]Molecular Engineering Graduate Program, University of Washington, Seattle, WA, USA. [4]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA. [5]Department of Physical Chemistry and Soft Matter, Wageningen University and Research, Wageningen, Netherlands. [6]Berkeley Center for Structural Biology, Molecular Biophysics and Integrated Bioimaging, Lawrence Berkeley Laboratory, Berkeley, CA, USA.

*Corresponding author. Email: dabaker@uw.edu

While deep learning has revolutionized protein structure prediction, almost all experimentally characterized de novo protein designs have been generated using physically based approaches such as Rosetta. Here we describe a deep learning–based protein sequence design method, ProteinMPNN, with outstanding performance in both in silico and experimental tests. On native protein backbones, ProteinMPNN has a sequence recovery of 52.4%, compared to 32.9% for Rosetta. The amino acid sequence at different positions can be coupled between single or multiple chains, enabling application to a wide range of current protein design challenges. We demonstrate the broad utility and high accuracy of ProteinMPNN using X-ray crystallography, cryoEM and functional studies by rescuing previously failed designs, made using Rosetta or AlphaFold, of protein monomers, cyclic homo-oligomers, tetrahedral nanoparticles, and target binding proteins.

- Backbone distances are encoded and processed using a message-passing neural network (Encoder) to obtain graph node and edge features.
- The encoded features, together with a partial sequence, are used to generate amino acids iteratively in a random decoding order.

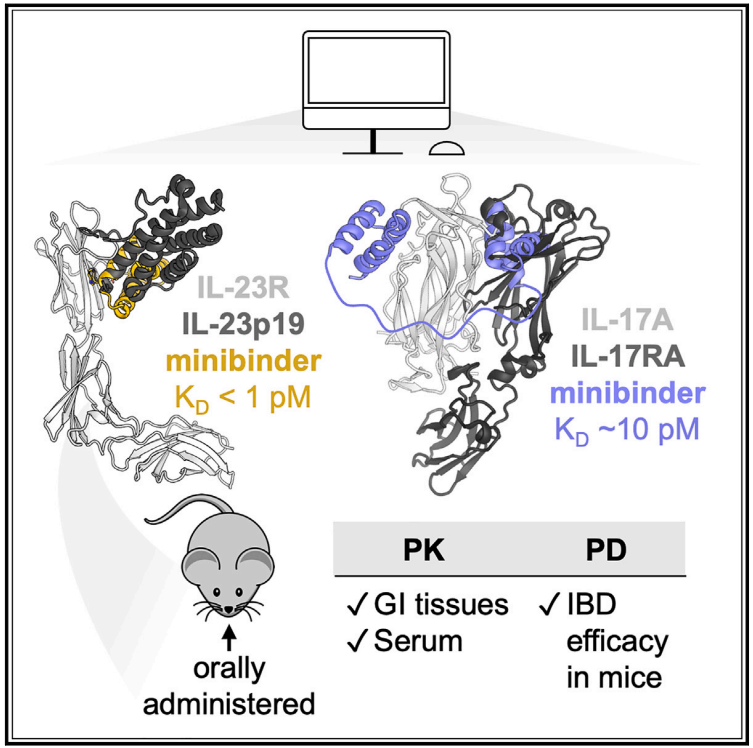


Biological Chemistry 1 - BIO-212, Lecture 6 2024

Cell

Preclinical proof of principle for orally delivered Th17 antagonist miniproteins

Graphical abstract



Authors

Stephanie Berger, Franziska Seeger, Ta-Yi Yu, ..., Matthias Siebeck, Roswitha Gropp, David Baker

Correspondence

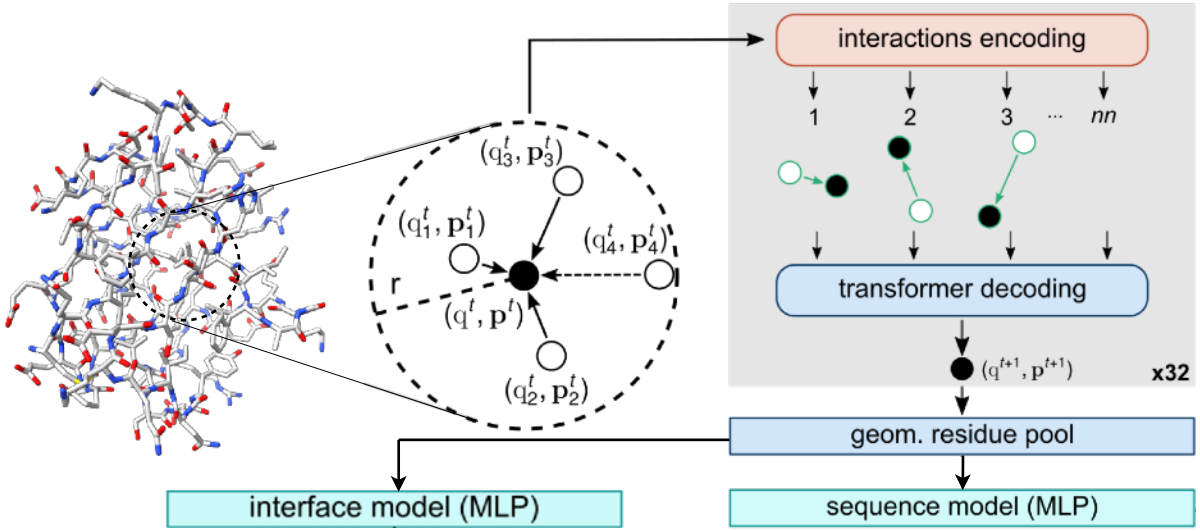
berger389@gmail.com (S.B.), dabaker@uw.edu (D.B.)

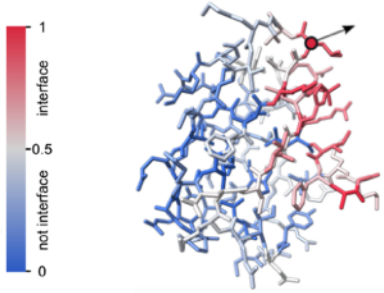
Highlights

- Computational design yielded low- and sub-pM minibinders of IL-17A and IL-23R
- IL-23R minibinders are extremely resistant to heat, acid, and proteolysis
- Oral IL-23R minibinder is as effective as a clinical mAb in mouse colitis

Berger et al., 2024, Cell *187*, 4305–4317
August 8, 2024
https://doi.org/10.1016/j.cell.2024.05.052

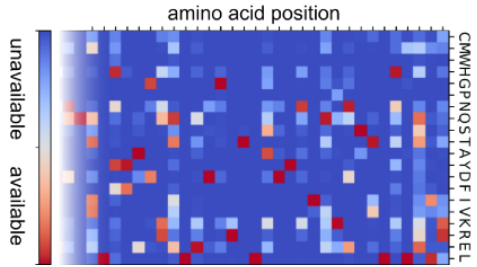
Protein Structure Transformer @LBM





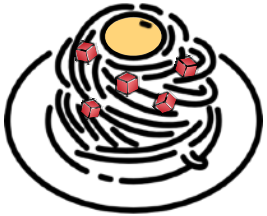
PeSTo: binding interfaces

Krapp et al. *Nat Comms* 2023



CARBonAra: molecular design

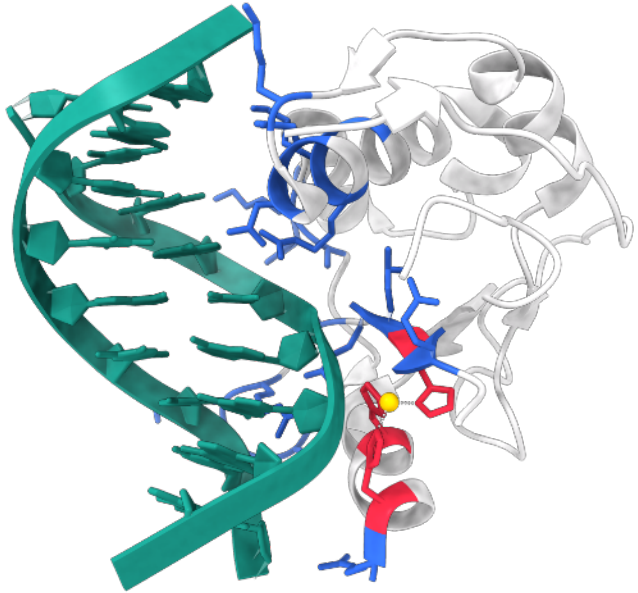
Krapp et al. *Nat Comms* 2024



Matteo Dal Peraro

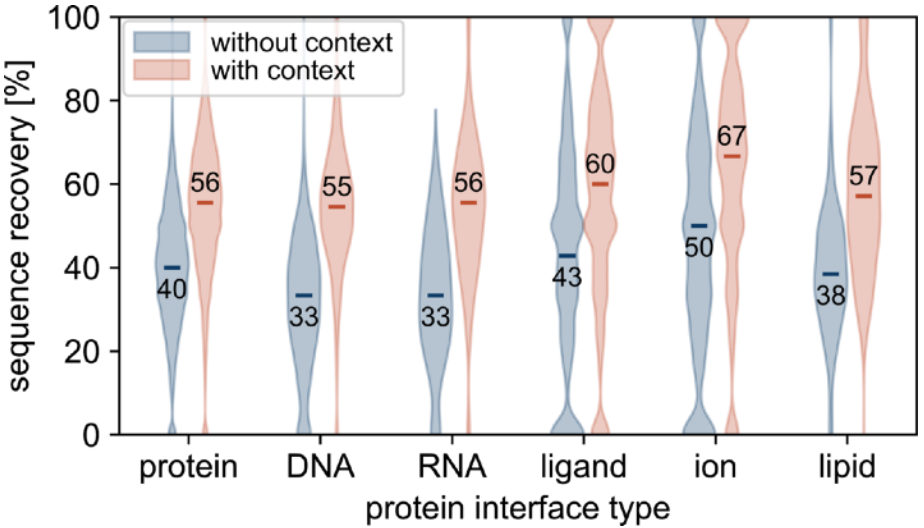
Unique ability — context awareness

- example with context



colicin E7

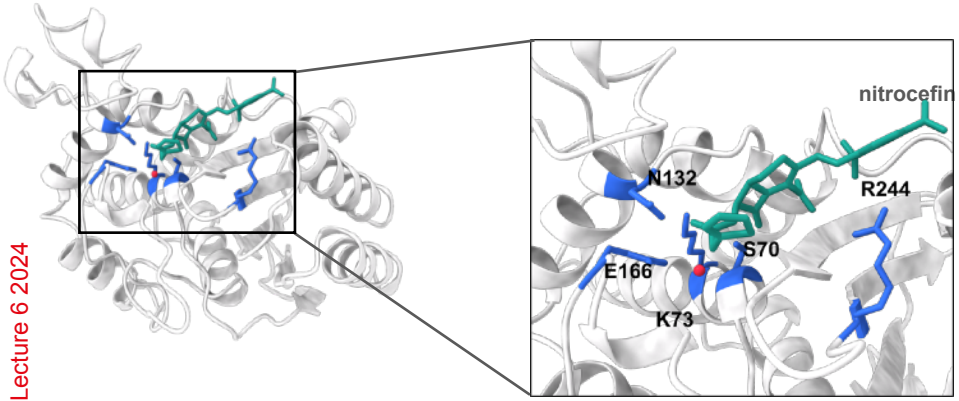
- large-scale benchmark



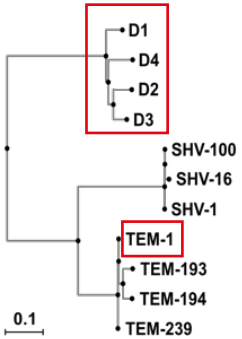
1000 structures sampled with maximum 30% sequence identity and separate C.A.T.H. classification from training set

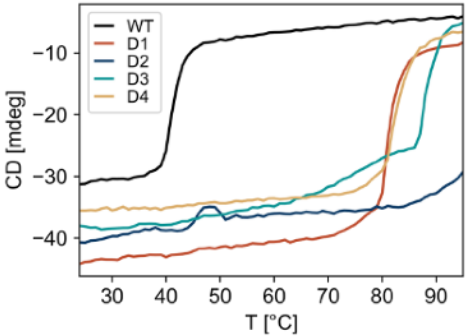
Can we re-engineer an enzyme?

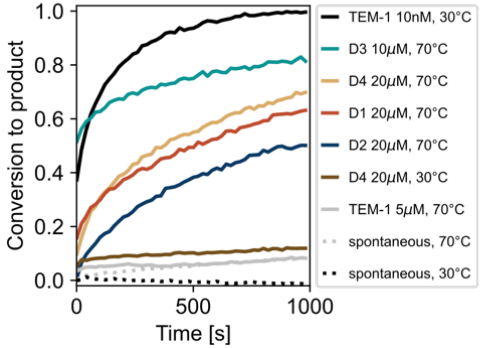
- TEM-1 serine β-lactamase



only 50% sequence identity







- sequences generation with substrate as constraint
- selected 10 top-ranked predictions based on plDDT
- 4/10 designs are soluble and monomeric
- they are folded and more thermostable than wild-type TEM-1
- catalytically active at high T - not as the wild-type yet
- represent a separate subclass of β-lactamases

Krapp et al. *Nat Comms* 2024

Matteo Dal Peraro

The future is bright and exciting …

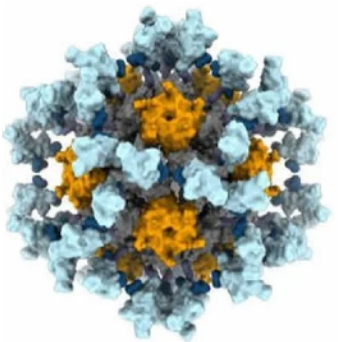
… biomolecular design will address many societal needs

- **Medicine**

  vaccines & antivirals

  smart medicines

  drug delivery



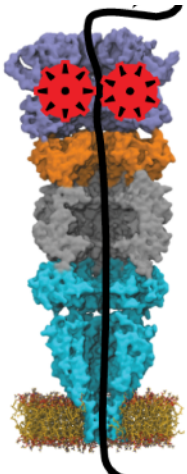
SARS-CoV-2 RBD
nanoparticle immunogen (Cell 2020)

- **Biotechnology**

  protein-silicon devices

  bio-based computers

  nanoscale manufacturing



SM proteomics with
biological nanopores
(Nat Chem 2021)

- **Sustainability**

  artificial photosynthesis

  CO_2 sequestration

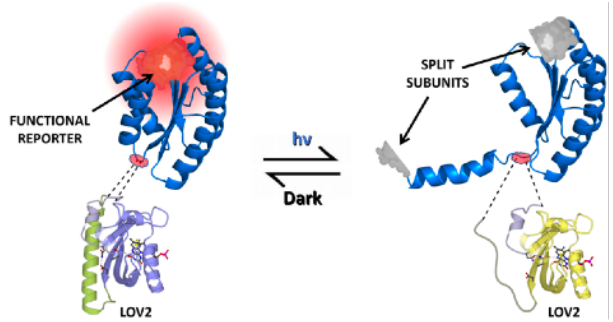
  plastic degradation

FAST-PETase
(Nature 2022)

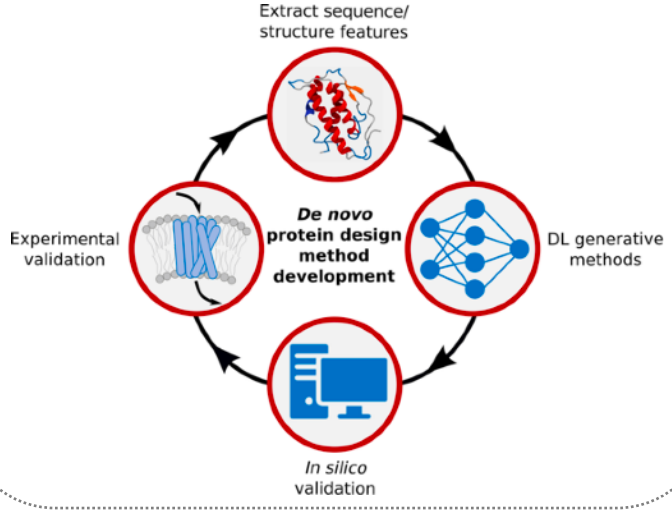
Laboratory of Protein and Cell Engineering

AI FOR PROTEIN DESIGN

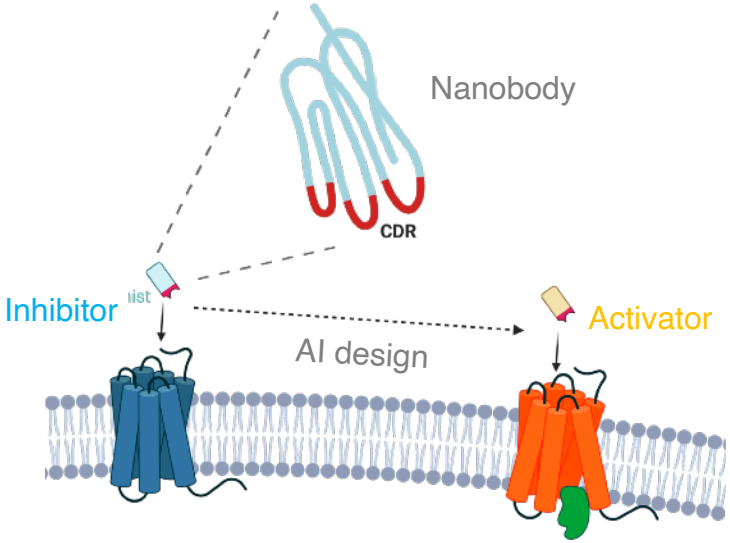
De novo optogenetic switches



De novo membrane protein design



De novo therapeutic protein ligands

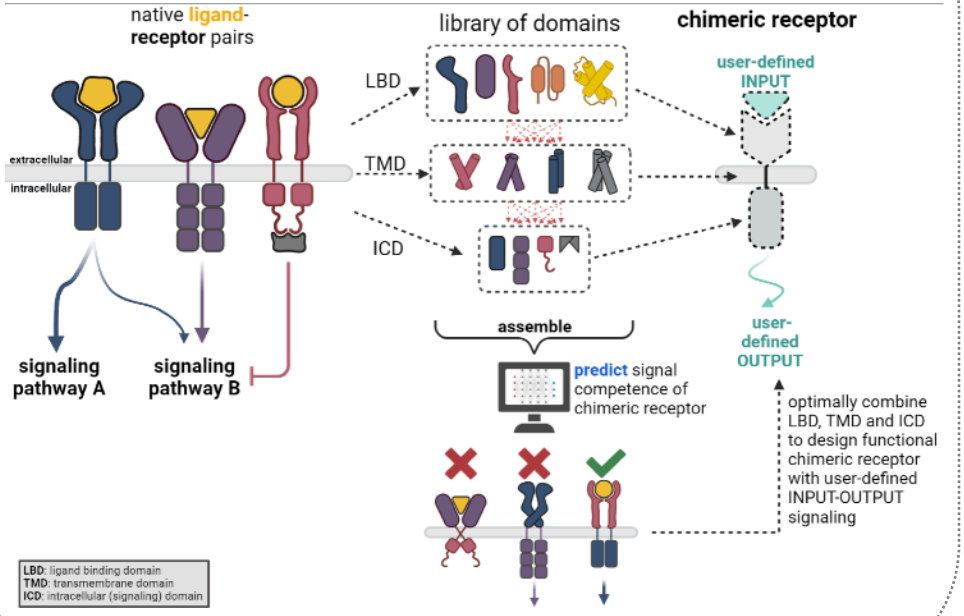


Barth Lab

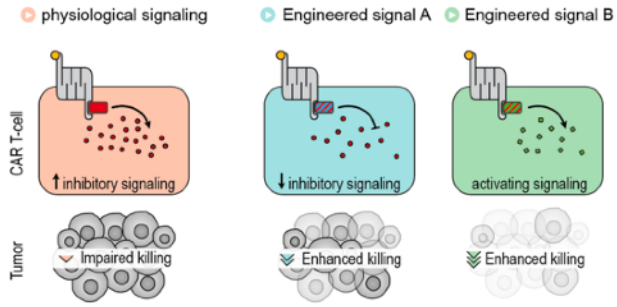
Laboratory of Protein and Cell Engineering

IMMUNOENGINEERING

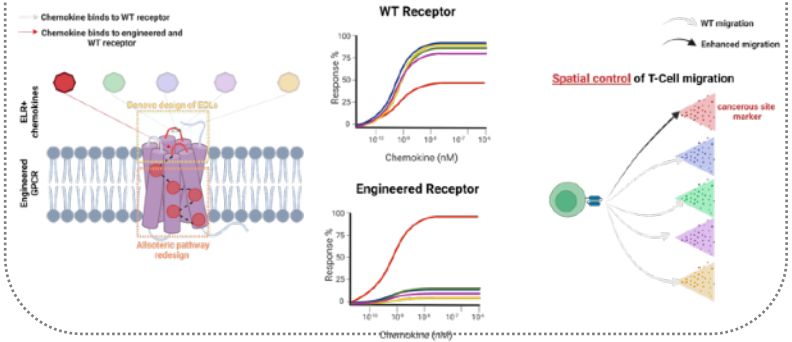
Allosteric biosensors for engineered cell therapies



G-protein signal rewiring to boost CAR T-cells



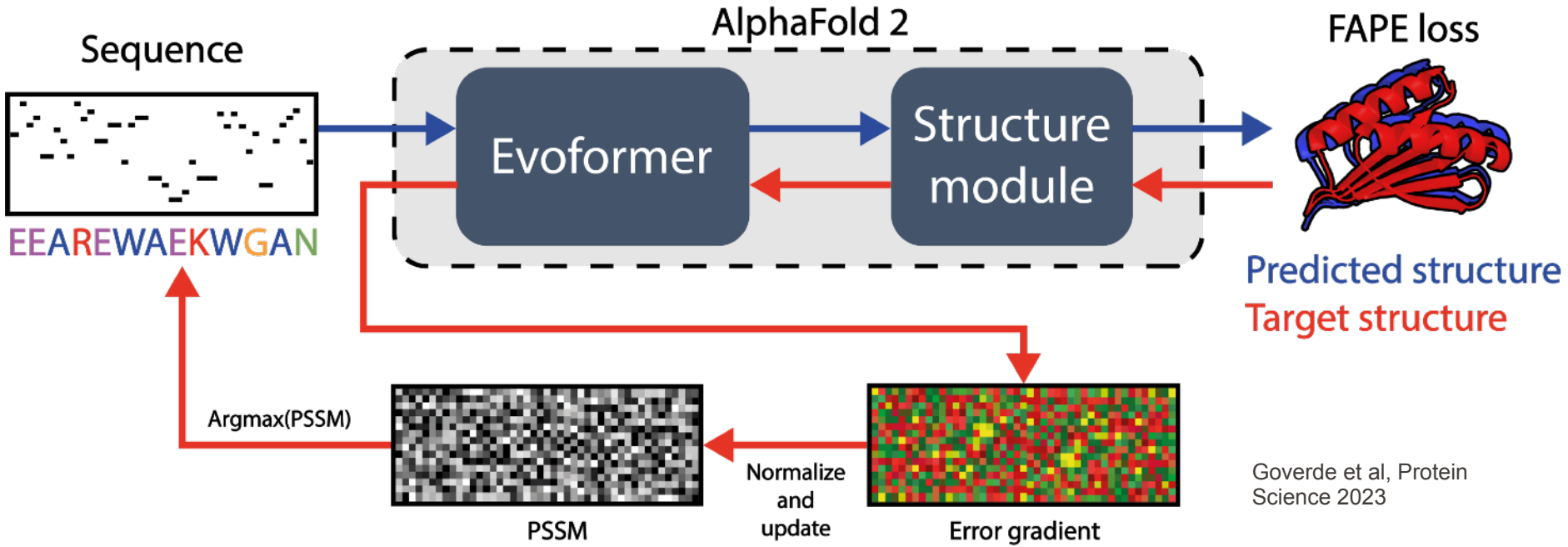
Redirecting cell migration with designed receptors



Barth Lab

Inverting AlphaFold for protein design

Bruno Correia, Laboratory of Protein Design and Immunoengineering

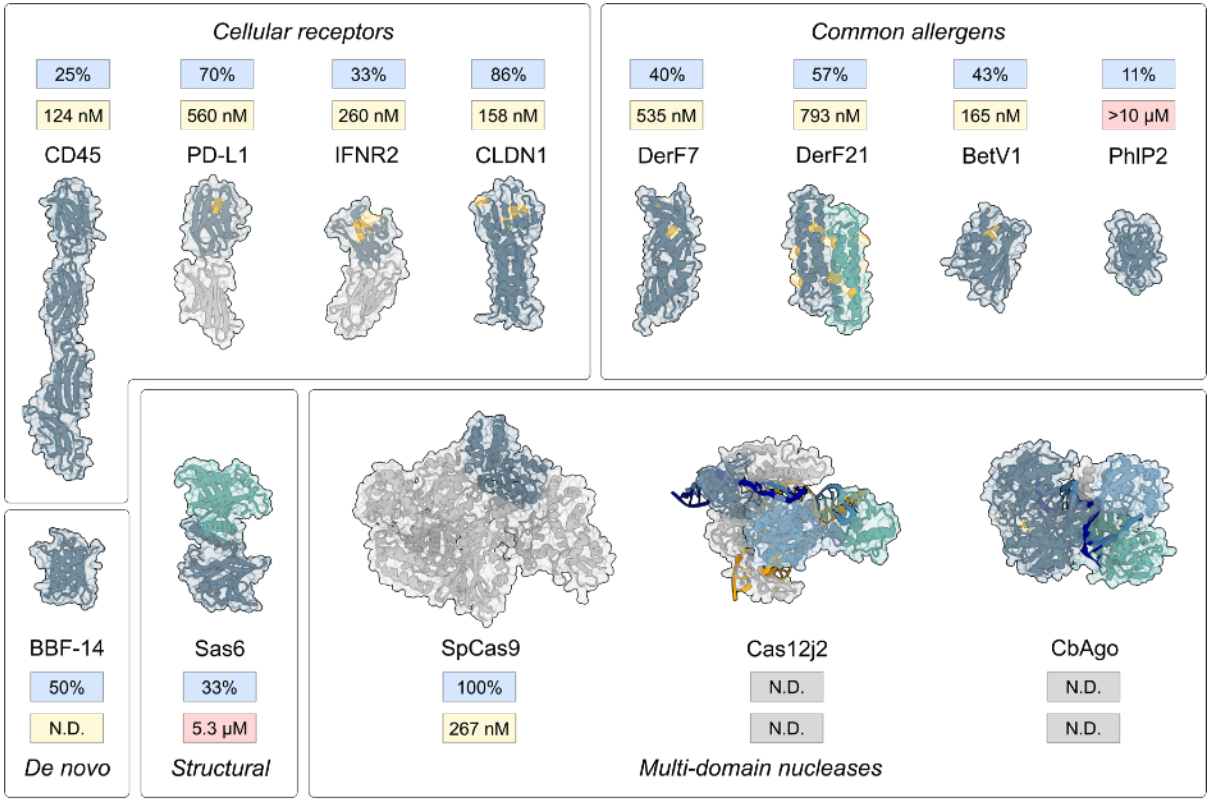


Goverde et al, Protein Science 2023

-Final sequences designed with proteinMPNN on AF2 generated backbones

High experimental success rates in binder design

Bruno Correia, Laboratory of Protein Design and Immunoengineering



Pacesa & Nickel, et al., *unpublished*

We weren't alone !!!!

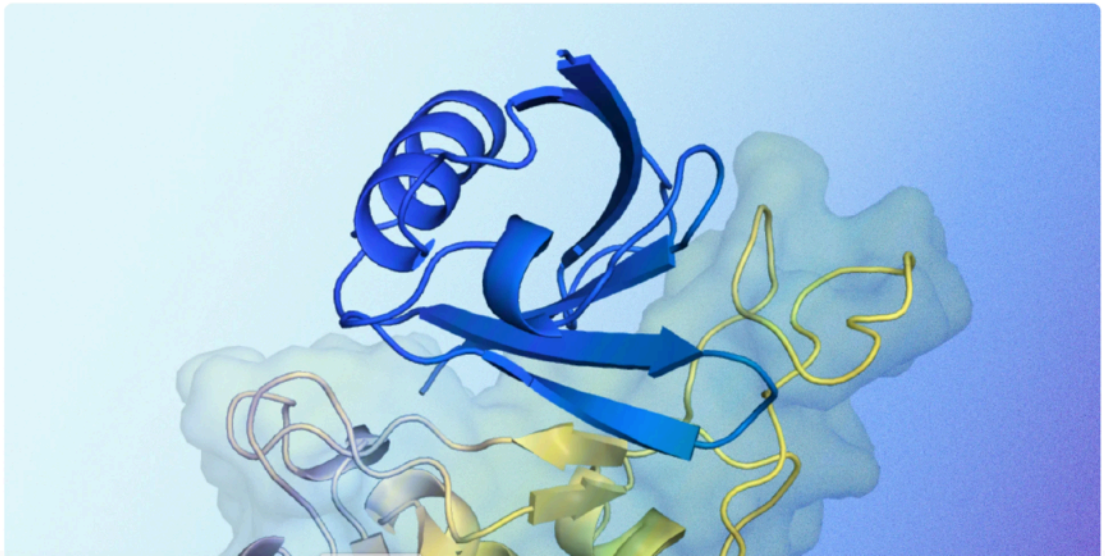
Google DeepMind About Research Technologies Impact Discover

AlphaProteo generates novel proteins for biology and health research

5 SEPTEMBER 2024

Protein Design and Wet Lab teams

Share



Designing Life with AI

https://www.designinglifewithai.com/

contact the MAKE team for ongoing projects offered by labs

We're thrilled to introduce "Designing Life with AI" at EPFL, where AI and protein design intersect, involving faculty, professors, and 40 students collaborating on topics like binder design and phosphosite engineering to kinase remodeling. After a year of innovative research, our projects are now being tested in the wet-lab, and we're working on creating a pipeline and resources for new students, aiming to expand our project and make EPFL a hub for protein design.