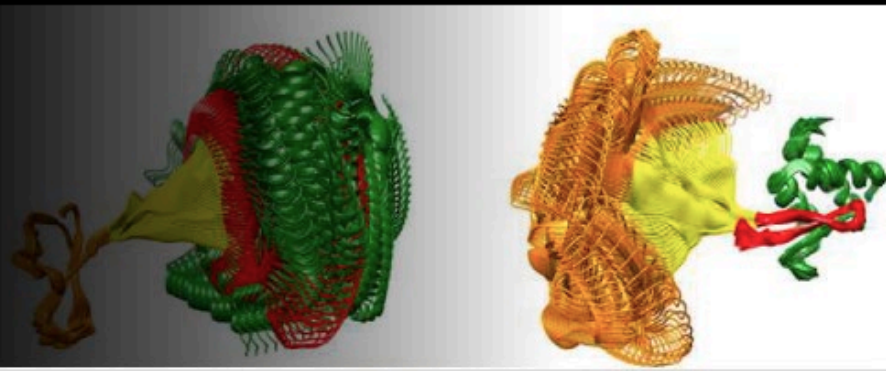
The Impact of Deep Learning for Creating Novel Effectors of Biological Functions

RPI STRUCTURAL
BIOINFORMATICS
LABORATORY

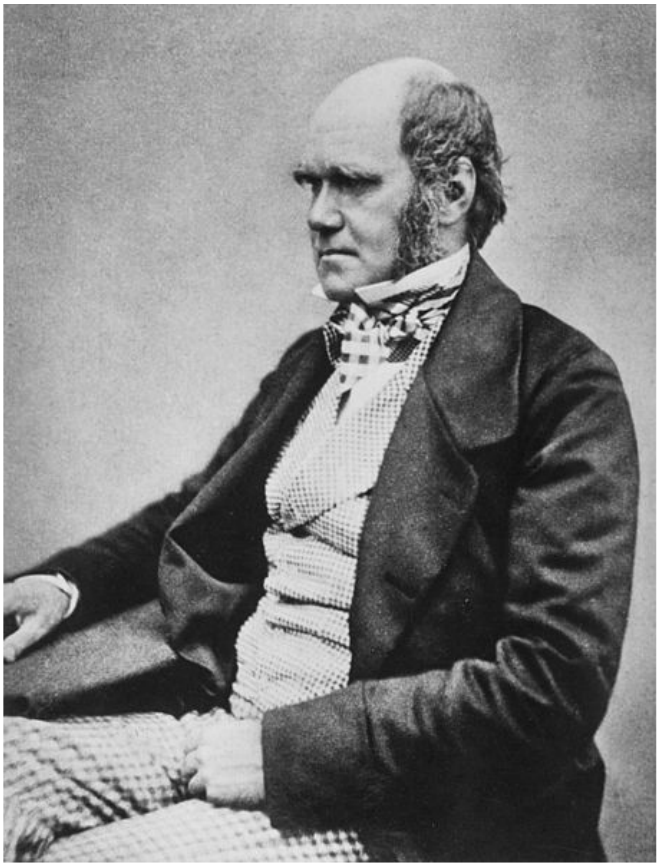
5th National Big Data Health
Science Conference
Columbia, South Carolina

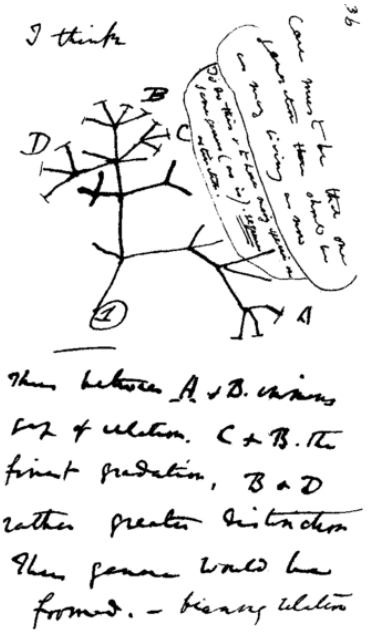
February 3, 2024



BIG DATA

Charles Darwin
(1809-1882)
Transmutation of Species





"Survival of the Fittest"

Charles Darwin – Natural Selection

The concept of fitness is central to natural selection. In broad terms, individuals that are more "fit" have better potential for survival, as in the well-known phrase "survival of the fittest".

Proposed in parallel by Alfred Russell Wallace

The concept of natural selection was originally developed in the absence of a valid theory of heredity; at the time of Darwin's writing, nothing was known of modern genetics.

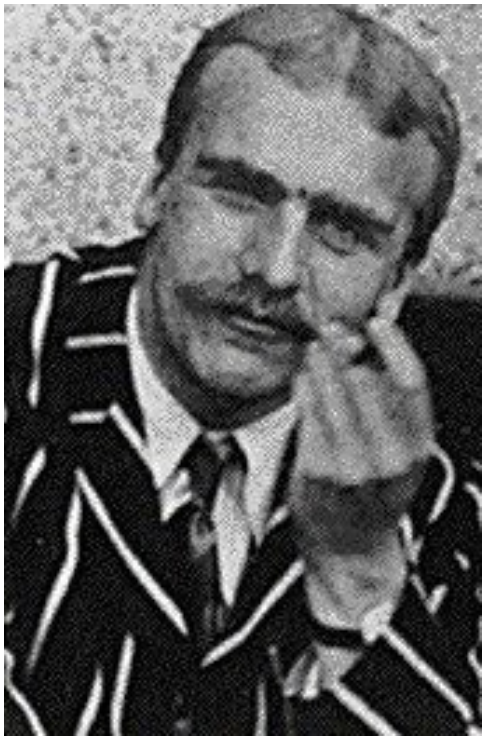
Gregor Mendel
(1822 – 1884)

Mendel demonstrated that the inheritance of certain traits in pea plants follows particular patterns, now referred to as the laws of Mendelian inheritance.

“Particulate” nature of inheritance of traits.

The laws formed the foundation of the modern science of genetics

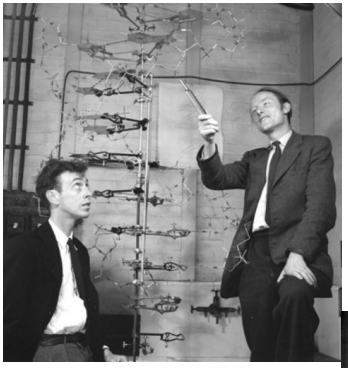
J.B.S. Haldane
(1892– 1964)

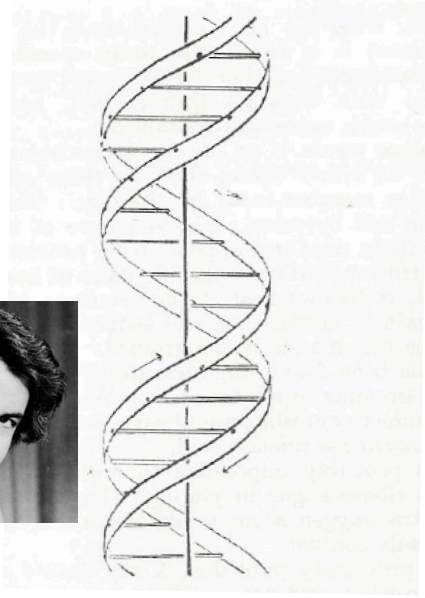


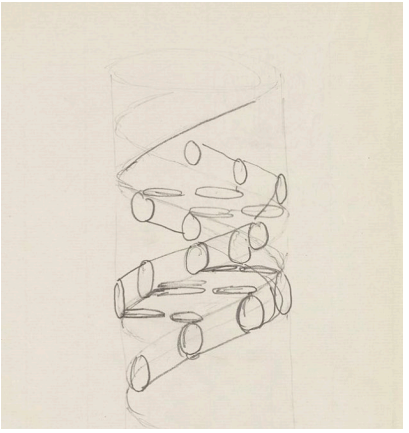
British-born geneticist, protein chemist. and evolutionary biologist generally credited with a central role in the development of neo-Darwinian thinking .

Genetic (gene) – basis of transmutation of species

Mutations of genes give rise to “transmutation of species” by changing the structure of the encoded gene product

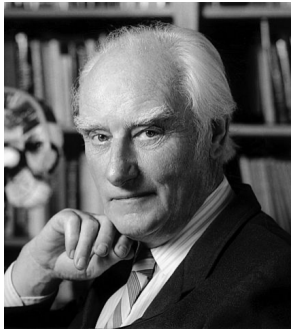






In February 1953, Watson and Crick completed their model, which is now accepted as the first correct model of the double-helix of DNA.

On 28 February 1953 Crick interrupted patrons' lunchtime at The Eagle Pub in Cambridge to announce that he and Watson had "discovered the secret of life"



Francis Crick in 1957

One strand templates the second strand

Francis Crick
James Watson
Rosalind Franklin

“Structural Biology” – using molecular structures to understand biology

Paradigm of Molecular Biology

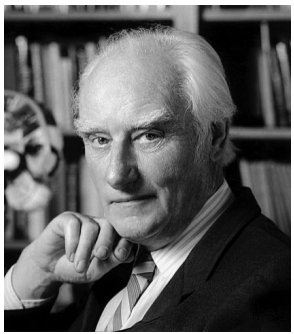
DNA Sequence

↓ transcription

RNA Sequence

↓ translation

Protein Sequence



Francis Crick in 1957

In 1979, Alexander Rich and co-workers at MIT grew a crystal of Z-DNA. This was the first crystal structure of any form of DNA.

Paradigm of Molecular Biology

DNA Sequence

transcription

RNA Sequence

translation / post-translational modification (PTM)

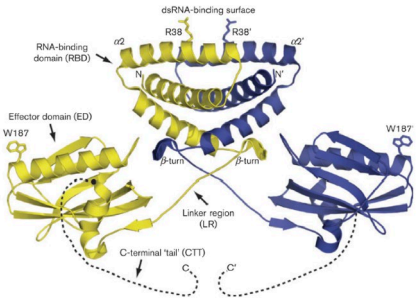
Protein Sequence

protein folding

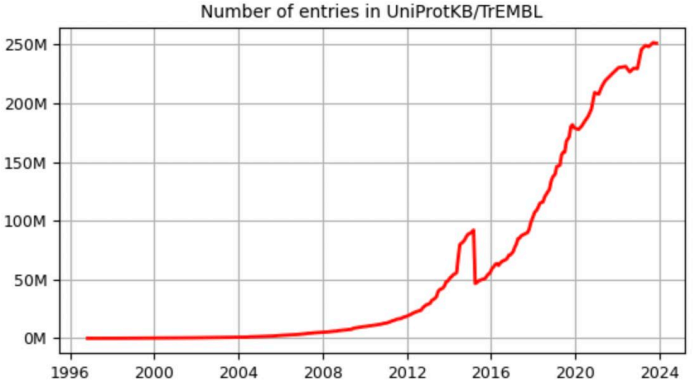
3D Protein Structure

nucleic acid folding

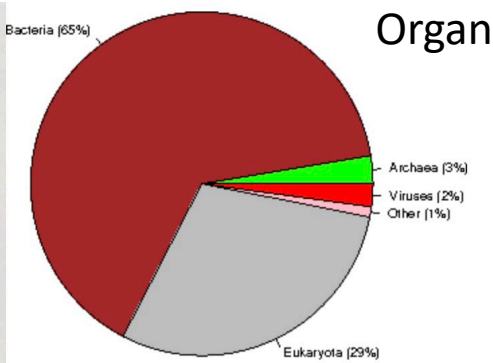
3D Nucleic Acid Structure



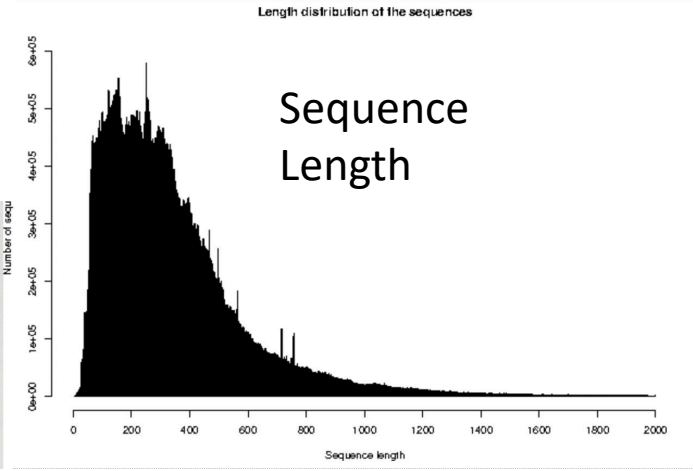
X-ray crystallography
Nuclear Magnetic Resonance (NMR)
Cryo Electron Microscopy



Organism

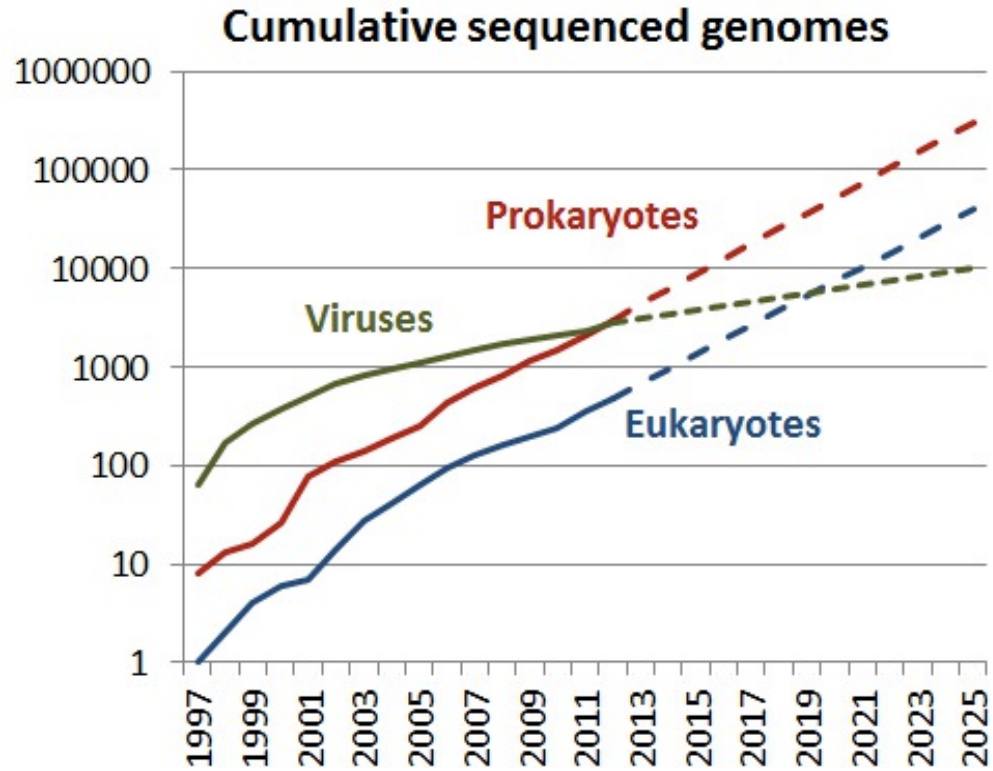


Sequence Length



The number of genomes sequenced increases by > 10-fold every 5 years

Over the next 15 yrs we expect to have 1,000 -fold more sequences than we have today.



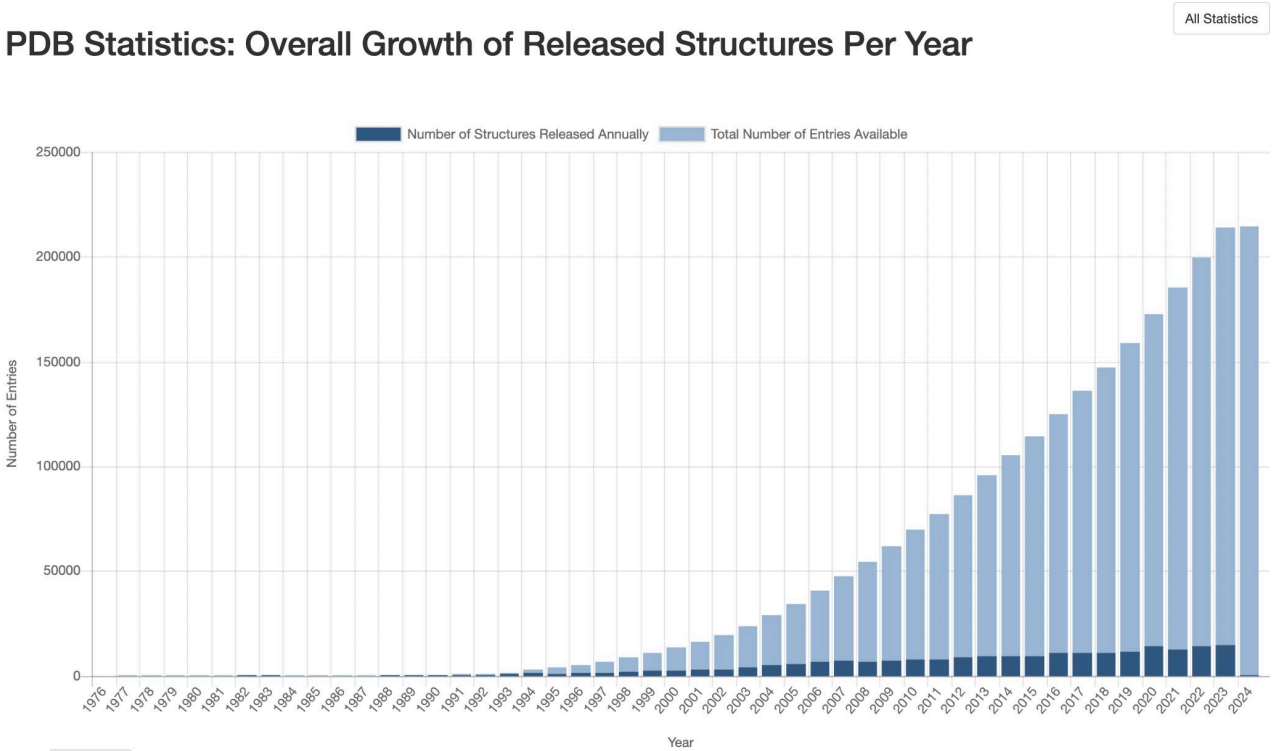
Su, Andrew (2013): Cumulative sequenced genomes. figshare.
http://dx.doi.org/10.6084/m9.figshare.722952

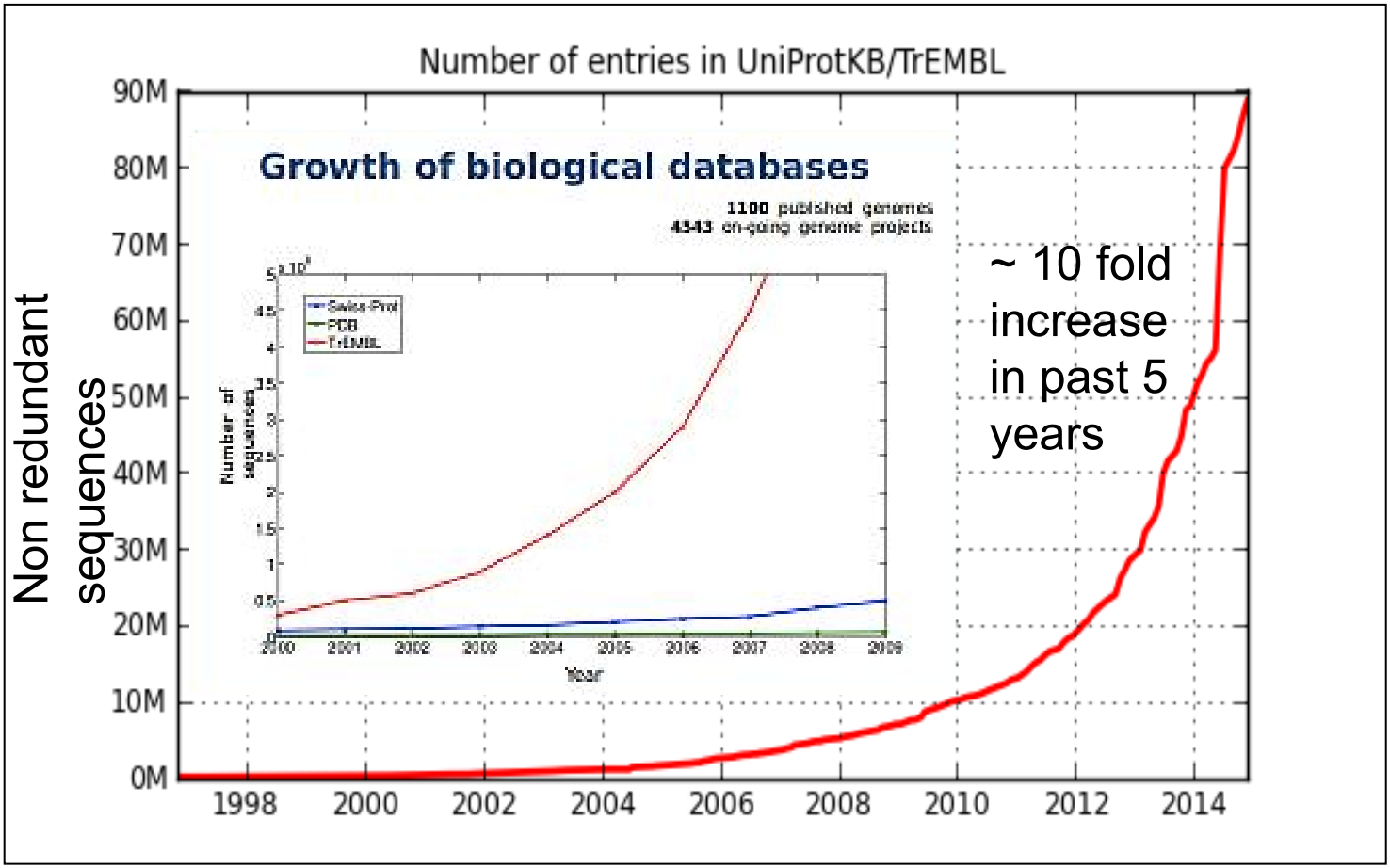
~~1 Gene → 1 Protein~~

1 Gene → Multiple Proteins

- Splicing
- Proteolytic Processing
- Posttranslational modifications
  - glycosylation
  - phosphorylation
  - acetylation
  - methylation
  - ubiquitination and sumylation
  - other modifications

Protein Structure Databank (PDB)




Number of entries in UniProtKB/TrEMBL
Non redundant sequences
90M
80M
70M
60M
50M
40M
30M
20M
10M
0M
1998
2000
2002
2004
2006
2008
2010
2012
2014
Growth of biological databases
1100 published genomes
4543 on-going genome projects
Swiss-Prot
PDB
TrEMBL
Number of sequences
Year
~ 10 fold increase in past 5 years

Paradigm of Molecular Biology

DNA Sequence

transcription

RNA Sequence

translation

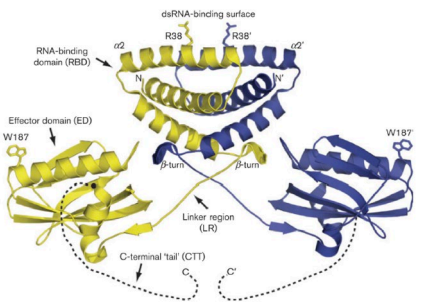
Protein Sequence

protein folding

3D Protein Structure

nucleic acid folding

3D Nucleic Acid Structure



PSI
PSI:BIOLOGY
Funded by the National Institutes of Health

International Protein Structure Initiative

The long-range goal of the Protein Structure Initiative (PSI) is to make the three-dimensional atomic-level structures of most proteins easily obtainable from knowledge of their corresponding DNA sequences.

JBC REVIEWS

Structural genomics and the Protein Data Bank

Received for publication, February 17, 2021, and in revised form, April 16, 2021 Published, Papers in Press, May 3, 2021,
https://doi.org/10.1016/j.jbc.2021.100747

Karolina Michalska[1,2] and Andrzej Joachimiak[1,2,3,*]

By the end of the PSI program (2000 – 2016), more than 9400 structures determined, with the majority of them being unique.

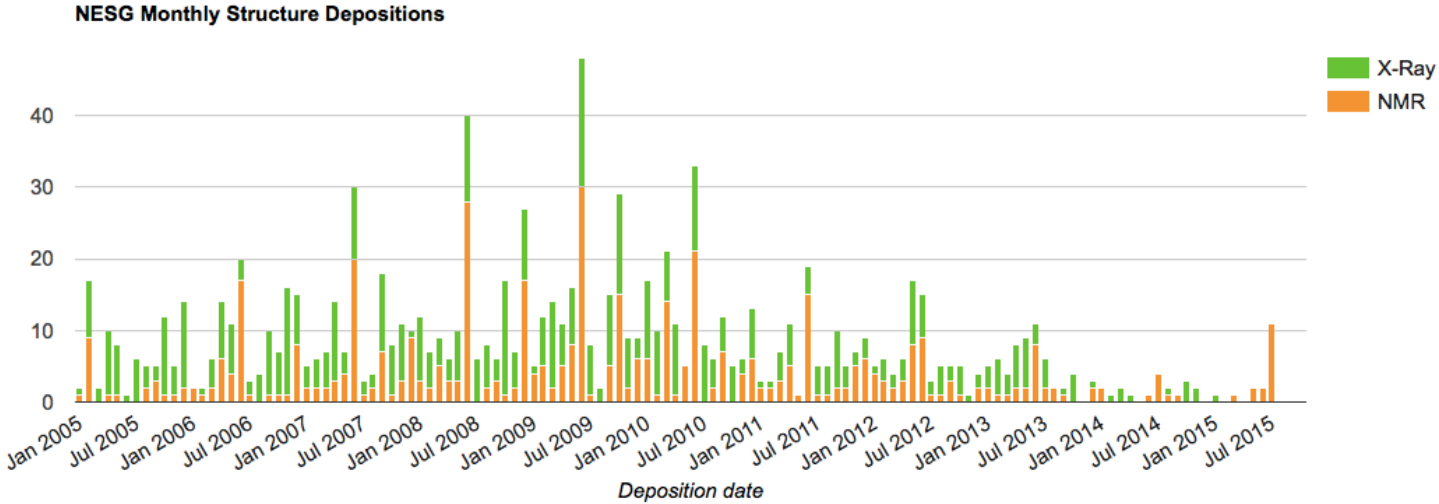
Table 1
Top 20 structural genomics programs

Center	Number of PDB deposits	Origin and funding	Techniques used
RIKEN Structural Genomics/Proteomics Initiative	2746	Japan, government, National Project on Protein Structural and Functional Analyses	NMR, X-ray
Midwest Center for Structural Genomics	1955	USA, PSI/NIH/NIGMS	X-ray, NMR
Structural Genomics Consortium	1896	International/a public–private partnership	X-ray, NMR
Joint Center for Structural Genomics	1601	USA, PSI/NIH/NIGMS	X-ray, NMR
Center for Structural Genomics of Infectious Diseases	1359	USA, NIH/NIAID	X-ray, NMR, cryo-EM
Seattle Structural Genomics Center for Infectious Disease	1355	USA, NIH/NIAID	X-ray, NMR, cryo-EM
Northeast Structural Genomics Consortium	1234	USA, PSI/NIH/NIGMS	X-ray, NMR
New York SGX Research Center for Structural Genomics	1041	USA, PSI/NIH/NIGMS	X-ray, NMR
New York Structural Genomics Research Consortium	364	USA, PSI/NIH/NIGMS	X-ray, NMR
TB Structural Genomics Consortium	344	International worldwide consortium/Various	X-ray, NMR
Center for Eukaryotic Structural Genomics	219	USA, PSI/NIH/NIGMS	X-ray, NMR
Montreal-Kingston Bacterial Structural Genomics Initiative	132	Canada, Canadian Institutes of Health Research	X-ray, NMR
Southeast Collaboratory for Structural Genomics	122	USA, PSI/NIH/NIGMS	X-ray, NMR
Structural Proteomics in Europe	118	European Union	X-ray, NMR
Berkeley Structural Genomics Center	101	USA, PSI/NIH/NIGMS	X-ray
Enzyme Discovery for Natural Product Biosynthesis	91	USA, NIH	X-ray
Structural Genomics of Pathogenic Protozoa Consortium	73	USA, PSI/NIH/NIGMS	X-ray, NMR
New York Consortium on Membrane Protein Structure	70	USA, PSI/NIH/NIGMS	X-ray
Structure 2 Function Project	54	USA, PSI/NIH/NIGMS	X-ray, NMR
GPCR Network	52	USA, PSI/NIH/NIGMS	X-ray

NIAID, National Institute of Allergy and Infectious Diseases; NIGMS, National Institute of General Medical Sciences; NIH, National Institutes of Health; PSI, Protein Structure Initiative.

neSG Northeast Structural Genomics Consortium

NMR: 570 structures
X-Ray: 664 structures
Total: 1,234 structures
177 Euka; 150 Human
> 200,000 homology models

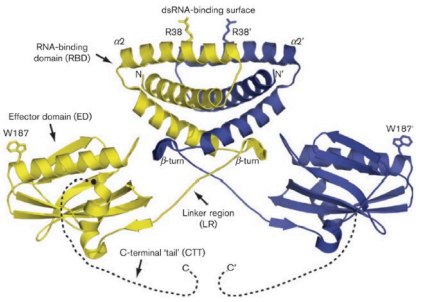


NMR: Arrowsmith, Kennedy, Montelione Powers, Prestegard, Szyperski, Valafar (Aramini, Cort, Eletsky, Gutmanas, Lamack, Lee, Liu, Mani, Mercier, Mills, Pederson, Pulavarti, Ramelot, Rossi, Singarapu, Shen, Stark, Swapna, Tang, Wu, Xu, Yang, Yee and others)

X-ray: Hunt, Tong, Montelione labs

Paradigm of Molecular Biology

Can we do this ALL
in silico



DNA Sequence

↓ transcription

RNA Sequence

↓ translation

Protein Sequence

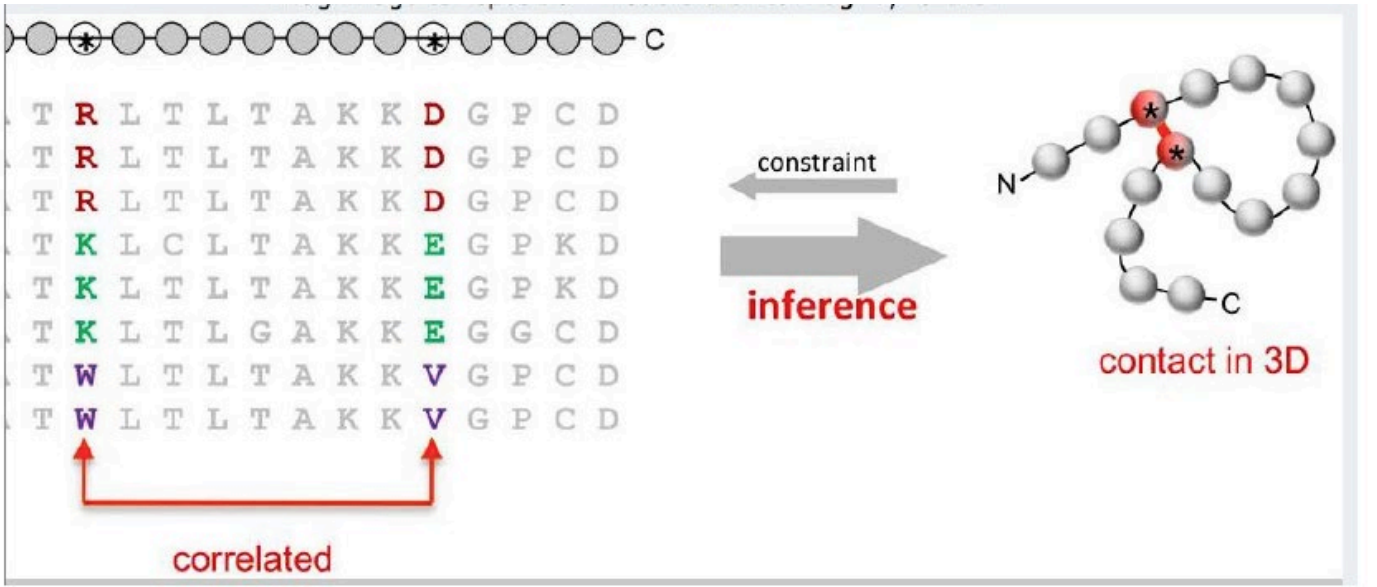
↓ protein folding / PTM

3D Protein Structure

PSI
PSI:BIOLOGY

Funded by the National Institutes of Health

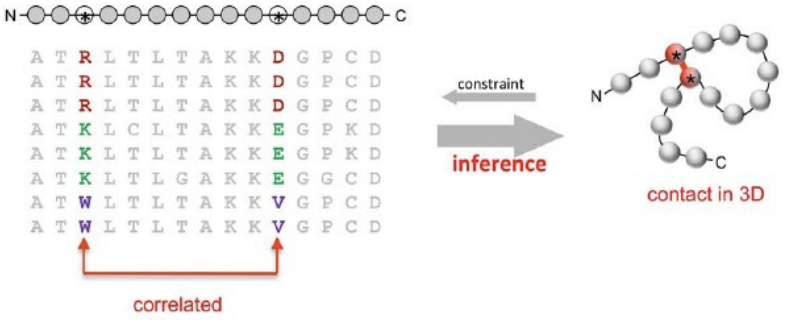
Evolutionary Co-Variance (EC)

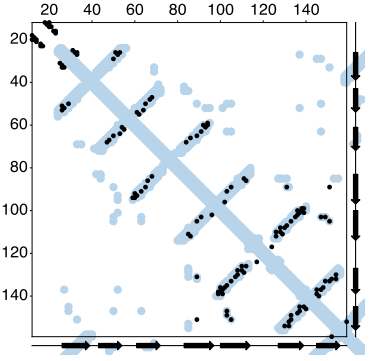


Direct-coupling analysis of residue coevolution captures native contacts across many protein families

Faruck Morcos[a,1], Andrea Pagnani[b,1], Bryan Lunt[a], Arianna Bertolino[c], Debora S. Marks[d], Chris Sander[e], Riccardo Zecchina[b,f], José N. Onuchic[a,g,2], Terence Hwa[a,2], and Martin Weigt[b,h,2]

PNAS, 110, 20533 2011





OPEN ACCESS Freely available online

PLoS one

Protein 3D Structure Computed from Evolutionary Sequence Variation

Debora S. Marks[1*], Lucy J. Colwell[2], Robert Sheridan[3], Thomas A. Hopf[1], Andrea Pagnani[4], Riccardo Zecchina[4,5], Chris Sander[3]

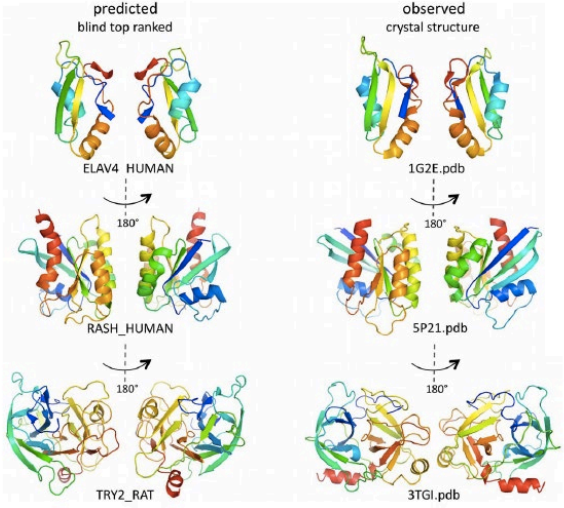


Table 1. Accuracy of predicted proteins.

Target Protein Uniprot ID	Fold	L*	Pfam ID	No. seqs	Blind top Ca-rmsd**	TM***	Best Ca-rmsd**	TM***	TP****	Ref. PDB
RASH_HUMAN	a/b	161	Ras	10K	3.5 (161)	0.7	2.8 (155)	0.76	0.8	5p21
CHEY_ECOLI	a/b	114	Response_reg	72K	2.98 (107)	0.65	2.96 (107)	0.67	0.67	1e6k
THIO_ALIAC	a/b	103	Thioredoxin	13K	3.86 (94)	0.55	3.5 (97)	0.59	0.68	1rqm
RNH_ECOLI	a/b	141	RNase_H	11K	4.0 (110)	0.54	3.5 (114)	0.57	0.68	1f21
TRY2_RAT	b	223	Trypsin	16K	4.27 (186)	0.6	4.27 (186)	0.54	0.81	3tgi
CADH1_HUMAN	b	100	Cadherin	12K	3.8 (88)	0.55	3.86 (96)	0.57	0.86	2o72
YES_HUMAN	b	48	SH3_1	6K	3.6 (47)	0.37	3.35 (43)	0.41	0.52	2hda
O45418_CAEEL	a+b	100	FKBP_C	8K	4.1 (88)	0.48	3.4 (79)	0.53	0.77	1r9h
ELAV4_HUMAN	a+b	71	RRM_1	28K	2.9 (67)	0.57	3.16 (71)	0.59	0.71	1g2e
A8MVQ9_HUMAN	a+b	107	Lectin_C	5K	4.8 (85)	0.39	4.0 (100)	0.53	0.8	2it6
PCBP1_HUMAN	a+b	63	KH_1	9K	4.69 (46)	0.25	4.61 (61)	0.35	0.47	1wvn
OPSD_BOVIN	a tm	258	7tm_1	27K	4.84 (171)	0.5	4.29 (180)	0.55	0.38	1hzx
BPT1_BOVIN	a+b	52	Kunitz_BPTI	2K	2.73 (53)	0.49	2.75 (53)	0.49	0.71	5pti
OMPR_ECOLI	a	77	Trans_reg_C	24K	4.7 (64)	0.35	3.9 (62)	0.45	0.38	1odd
SPTB2_HUMAN	a	108	CH(calp hom)	4K	4.0 (47)	0.37	3.88 (88)	0.5	0.5	1bkr

Backbone RMSD of top model to Xtal structure of 3 – 5 Å

PLoS ONE 6, e28766 (2011).

Genomics-aided structure prediction

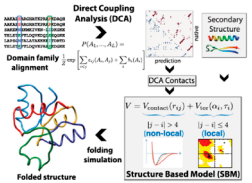
Joanna I. Sułkowska[a,1], Faruck Morcos[a,1,2], Martin Weigt[b], Terence Hwa[a,2], and José N. Onuchic[c,2]

[a]Center for Theoretical Biological Physics, University of California at San Diego, La Jolla, CA 92093-0374; [b]Laboratoire de Gén
Microorganismes, UMR 7238, Université Pierre et Marie Curie, 15 rue de l'École de Médecine, 75006 Paris, France; and
[c]Center for Theoretical Biological Physics, Rice University, Houston, TX 77005-1827

Contributed by José N. Onuchic, May 9, 2012 (sent for review January 20, 2012)

DCA-fold, integrating DCA contacts with an accurate knowledge of local information is sufficient to fold proteins in the range of 1–3 Å accuracy

PNAS 2012



eLIFE elifesciences.org

RESEARCH ARTICLE

2014

Sequence co-evolution gives 3D contacts and structures of protein complexes

Thomas A Hopf[1,2†], Charlotta P I Schärfe[1,3,4†], João P G L M Rodrigues[5†], Anna G Green[1], Oliver Kohlbacher[3,4], Chris Sander[6*], Alexandre M J J Bonvin[5*], Debora S Marks[1*]

eLIFE elifesciences.org

RESEARCH ARTICLE

2014

Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information

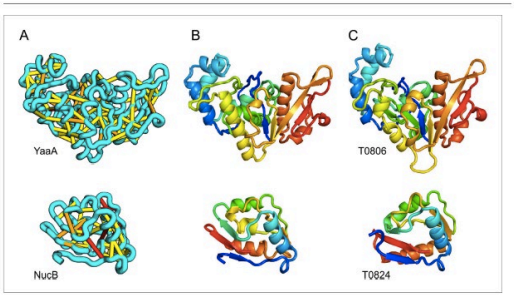
Sergey Ovchinnikov[1,2†], Hetunandan Kamisetty[1,3†], David Baker[1*]

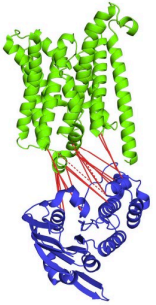
Large-scale determination of previously unsolved protein structures using evolutionary information

Sergey Ovchinnikov[1], Lisa Kinch[2], Hahnbeom Park[1], Yuxing Liao[3], Jimin Pei[2], David E Kim[1], Hetunandan Kamisetty[4], Nick V Grishin[2,3], David Baker[1,5*]

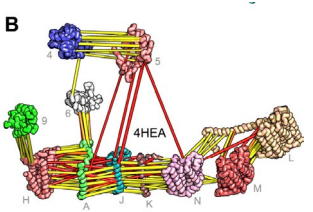
EC restrained Rosetta: Backbone RMSD of top models to Xtal structure of 2.7 – 6 Å

eLife 2015; Science 2017

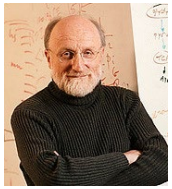




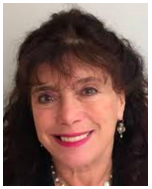
BtuC – BtuD



Can we enable protein structure determination by combining sparse NMR data with EC restraints?



C Sander

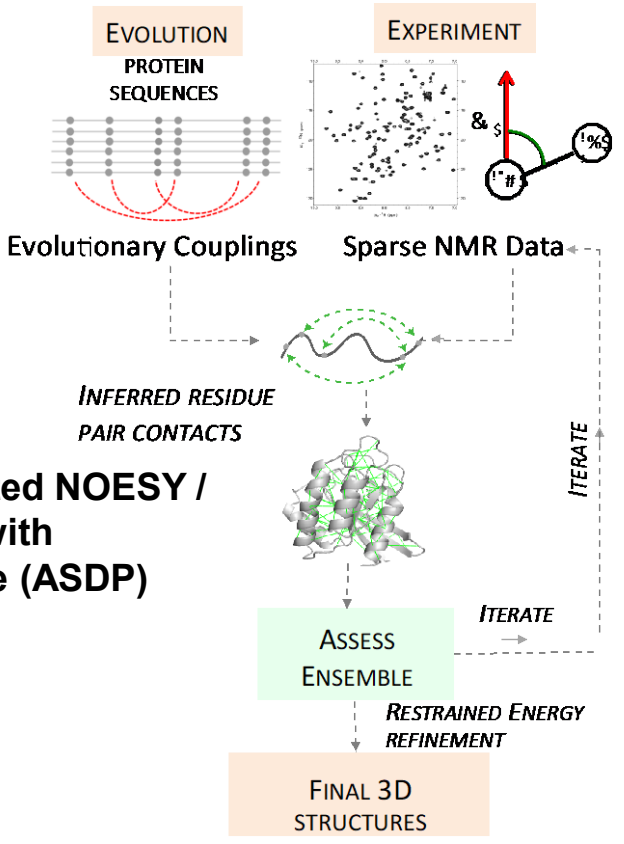


D Marks

Yuefeng Tang
Janet Huang
Thomas Hopf
Chris Sander
Debora Marks

EC-NMR Method

EVFold Server



Structure generation

Fully automated NOESY / EC analysis with AutoStructure (ASDP)

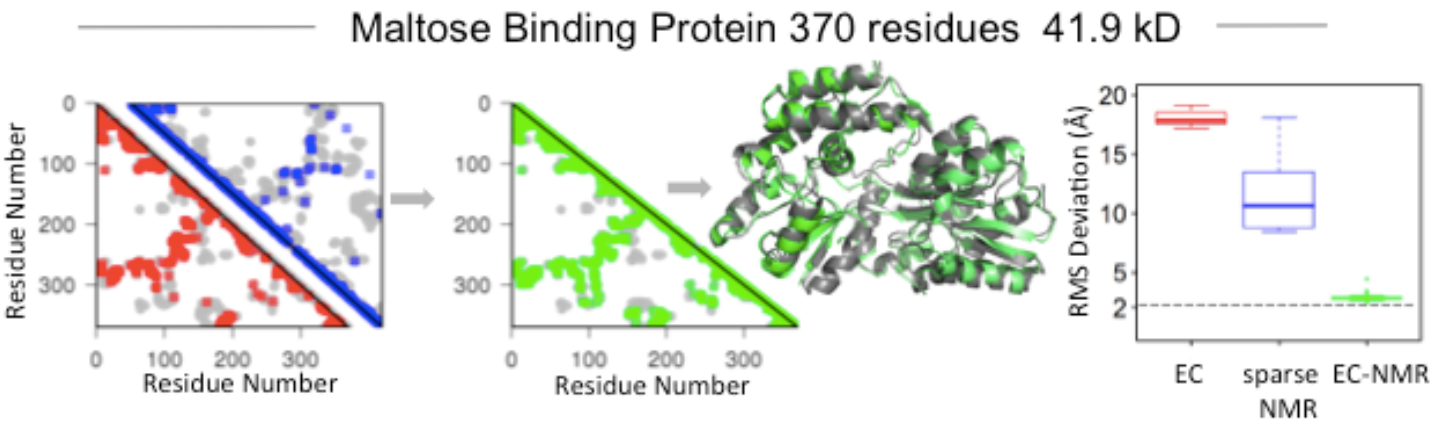
Refined Residue Pair Contacts -> Atom Pair Contacts

Tang, Huang et al
Nature Methods 2015

Restrained Rosetta

Hybrid EC-NMR Method

H^N, ILV Me Assignments
1 ^{15}N-^{1}H RDC alignment
N_{eff} = 12,416 N_{eff} / L = 33



NOESY cross peak assignment ambiguity is iteratively resolved

False positive ECs are identified and removed

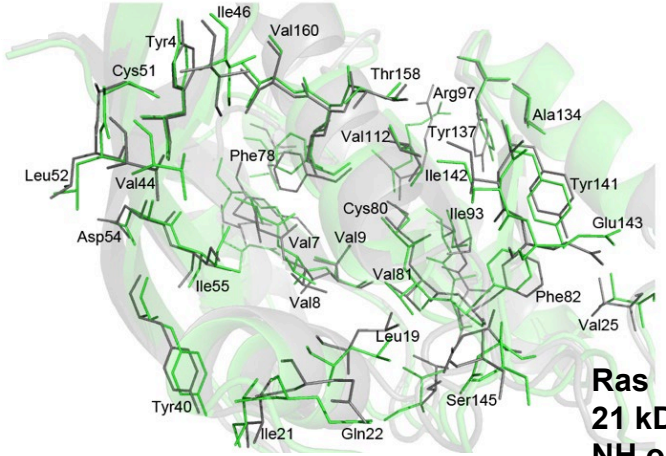
ECs are deconvolved into atom-atom distance restraints.

More complete and precise distance restraints.

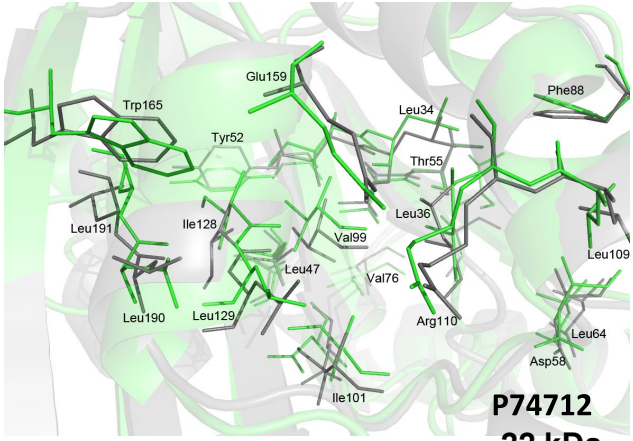
MBP – beta cyclodextrin complexes

Tang et al
Nature Methods 2015

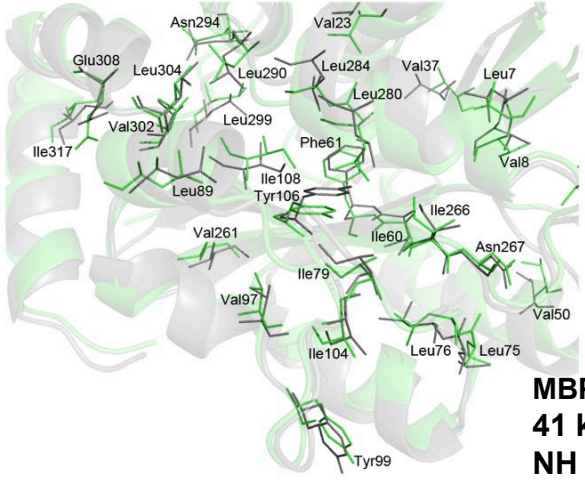
EC-NMR structures have accurate core sidechain structures



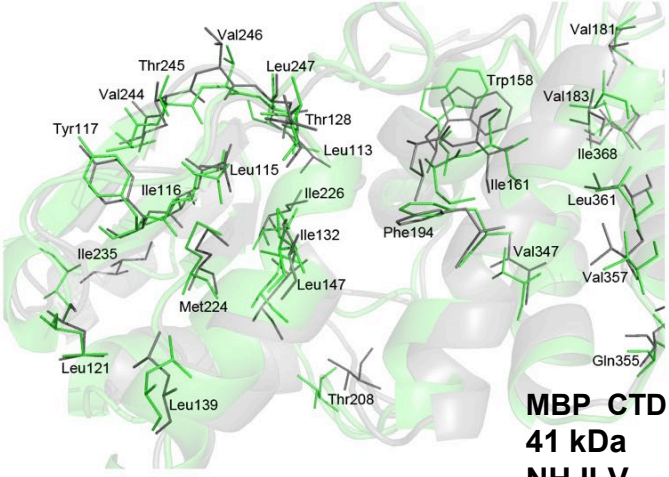
Ras
21 kDa
NH only



P74712
22 kDa
NH ILV



MBP NTD
41 kDa
NH ILV



MBP CTD
41 kDa
NH ILV

Tang et al Nat Methods

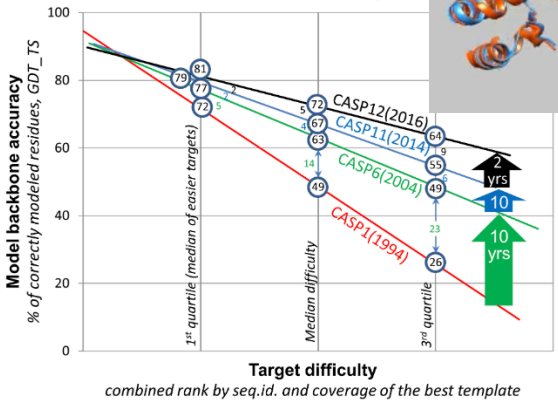
CASP – Critical Assessment of Protein Structure Prediction

A community-wide, worldwide experiment for protein structure prediction taking place every
two years since 1994

CASP provides research groups with an opportunity to objectively test their structure prediction methods and delivers independent assessment of the state of the art in protein structure modeling

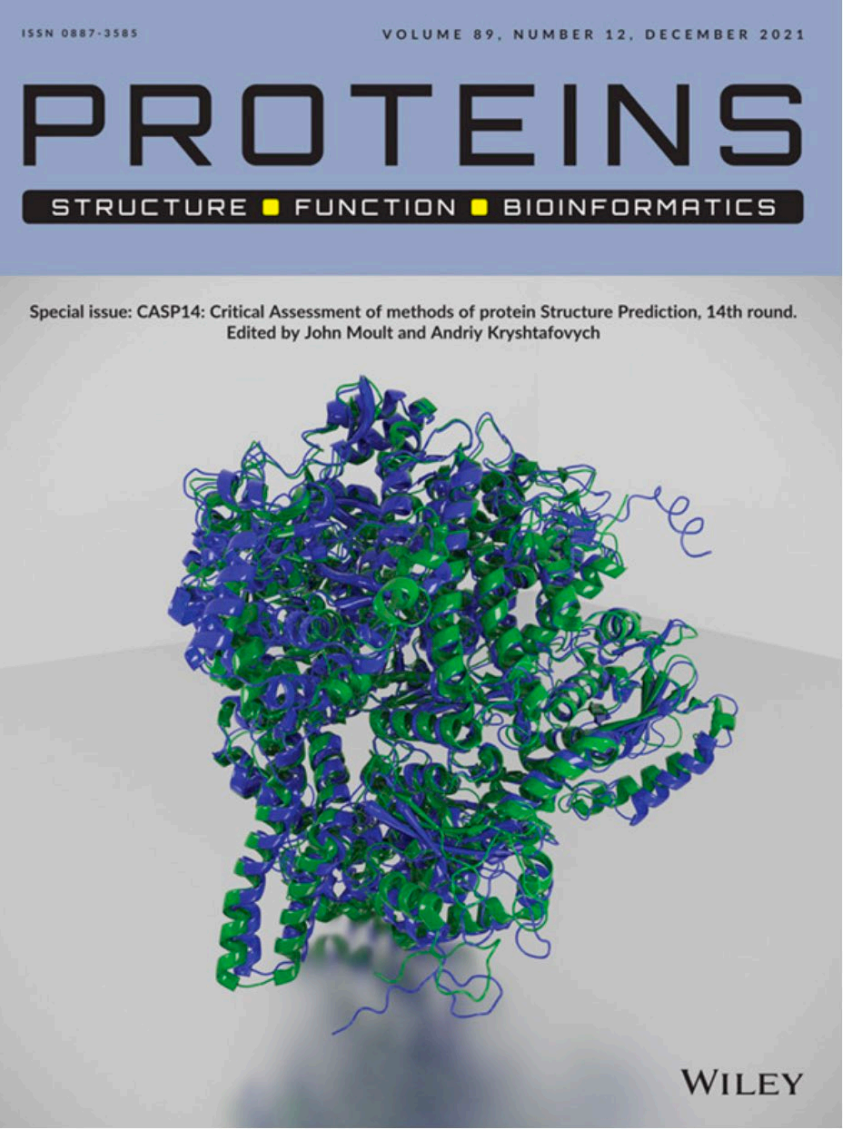
More than 100 research groups from all over the world participate in CASP on a regular basis

http://predictioncenter.org/



Categories:

- Template-based Modeling
- ab initio modeling
- contact prediction
- Refinement
- Data-assisted prediction



CASP – Critical Assessment of Protein Structure Prediction

- Experimental Structures
- Sequences Distributed
- Blind Predictions
- Third-party Assessors

CASP1 – 1994

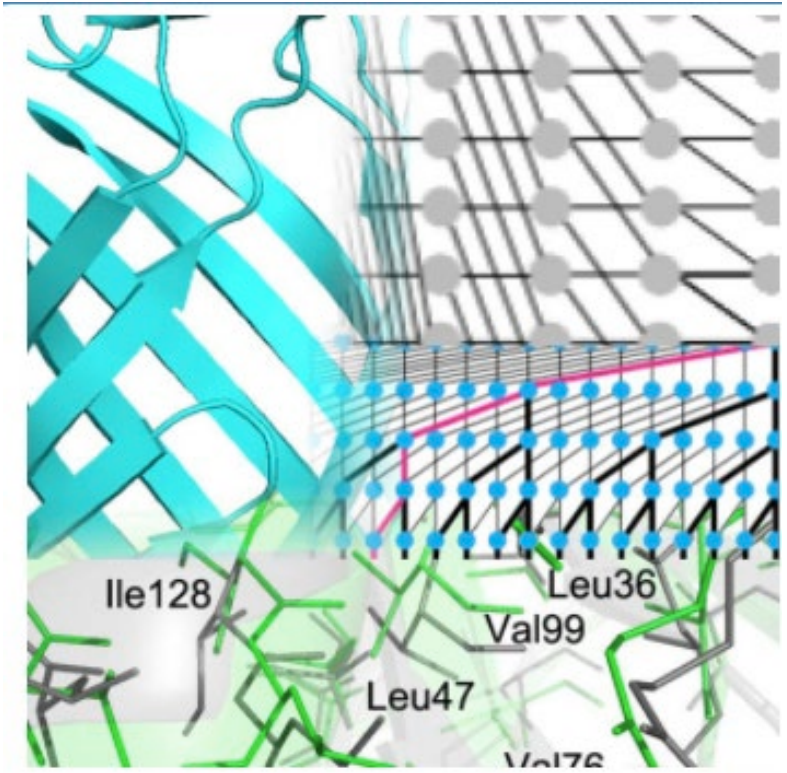
…

CASP15 - 2022

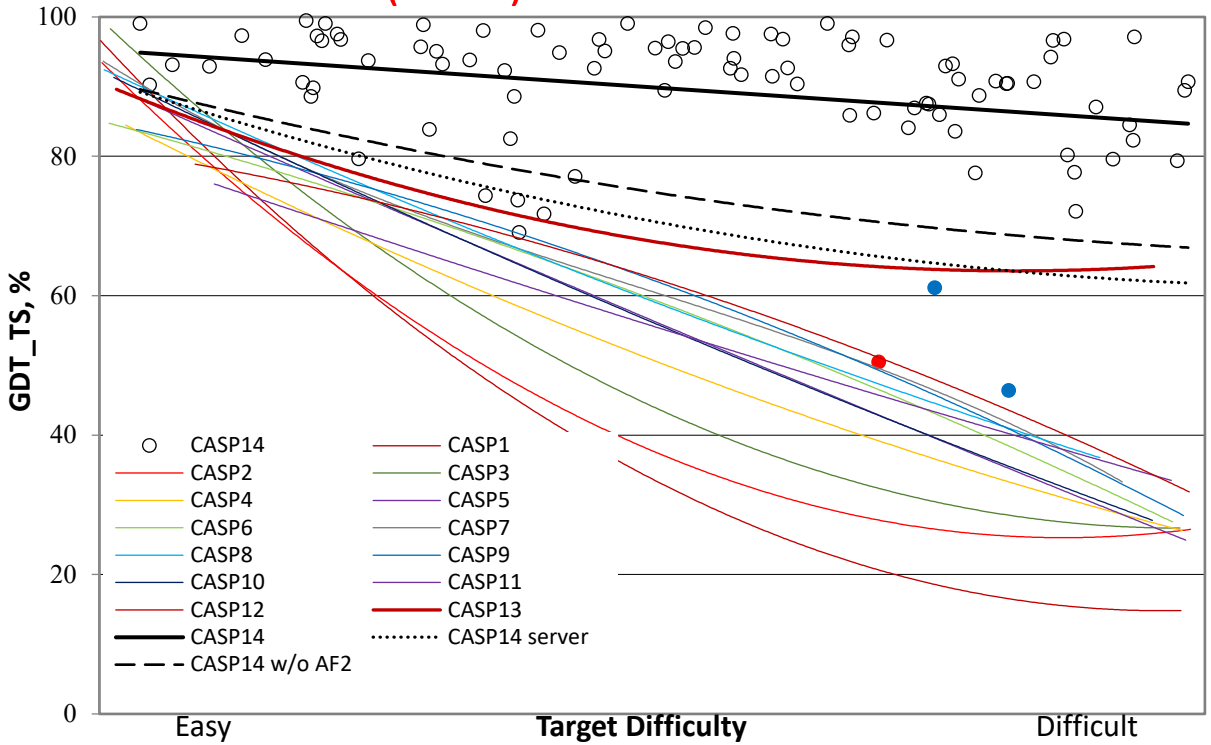
rmsd – root-mean square deviation of atomic coordinates.
2 crystal structures < 1.5 Å

GDT – global distance test. Percent of atoms < 1 Å rmsd

Deep Learning



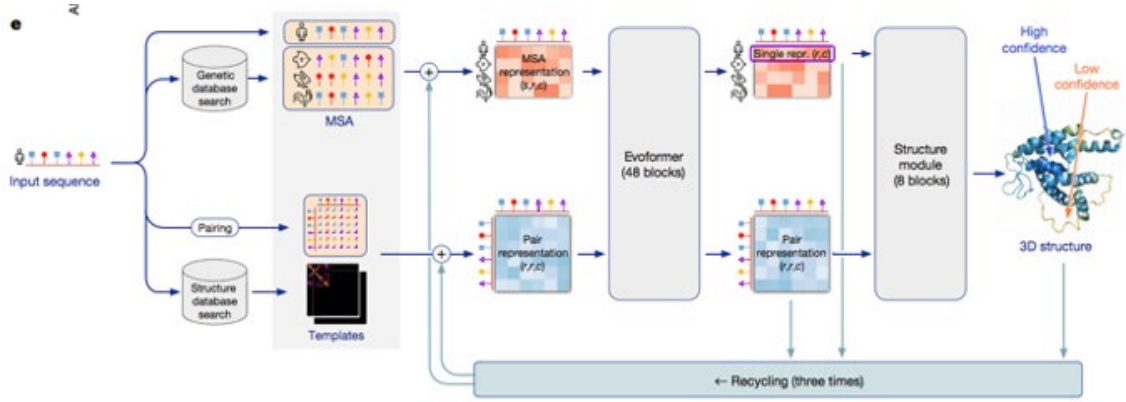
Remarkable Performance of AlphaFold in CASP14 (2020) – for 96 Evaluation Units



For 96 CASP14 targets AF2 models had a mean GDT_TS of 0.88 ± 0.1, corresponding to RMSD between predicted and experimental protein structures of about 1.5 Å

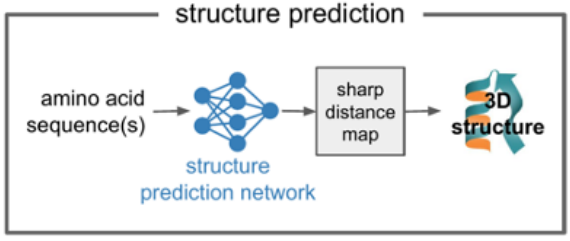
Kryshtafovych, Fidelis, Moult et al, 2021

Highly Accurate Protein Structure Prediction with AlphaFold: Deep Learning from the Protein Structure DataBase



Jumper et al Nature 2021

Deep Mind, Inc



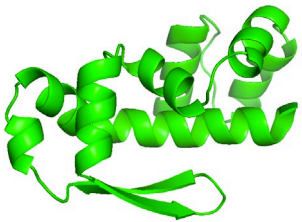
AlphaFold – Convolutional Neural Network

AlphaFold2 – Attention-based Network

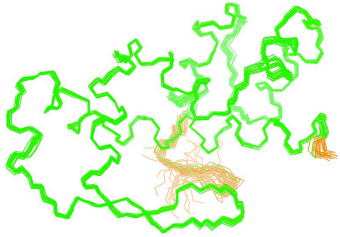
AlphaFold 2020

CASP14 Target T1055

well-defined region:
residues 305 - 426



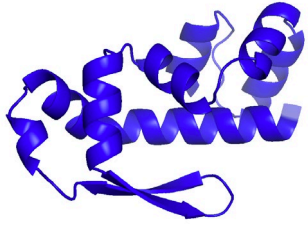
NMR Model



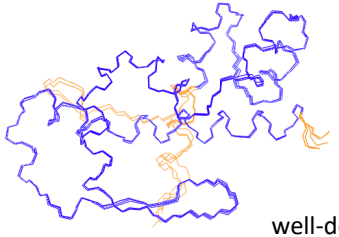
NMR Ensemble

Common Region for RMSD / GDT comparison: 310-426

<RMSD> = 0.97 Å
GDT = 0.90

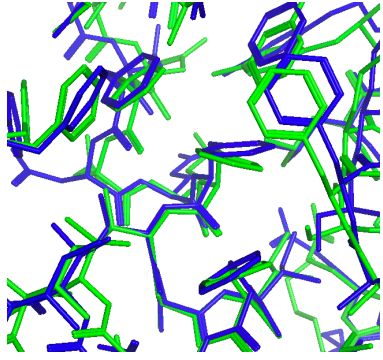


AF Model

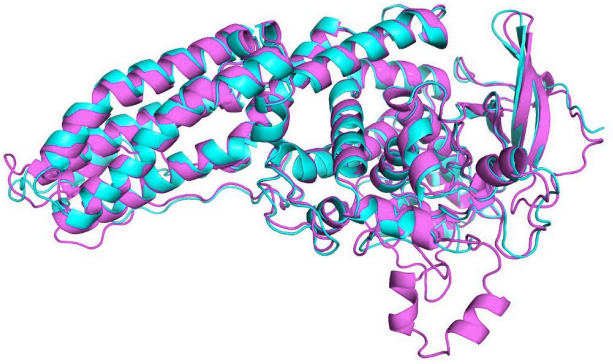


well-defined region:
residues 310-428

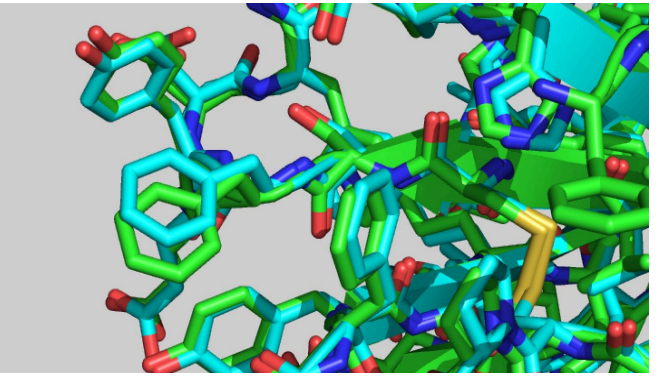
AF Ensemble



Huang YJ, Zhang N, Bersch B, Fidelis K, Inouye M, Ishida Y, Kryshtafovych A, Kobayashi N, Kuroda Y, Liu G, LiWang A, Swapna GVT, Wu N, Yamazaki T, Montelione GT. Assessment of prediction methods for protein structures determined by NMR in CASP14: Impact of AlphaFold2. ***Proteins.*** *2021*



CASP target T1053, a two-domain bacterial kinase. Both domains are difficult modeling targets (FM/TBM category).



Kryshtafovych et al, 2021

attention-based machine learning

largest CASP14 target was 949 res

AF2 and RosettaFold source freely available

Dec 2020 CASP14

Jul 2021 AlphaFold (AF2) Published in *Nature*

Jul 2021 RoseTTAFold Published in *Science*

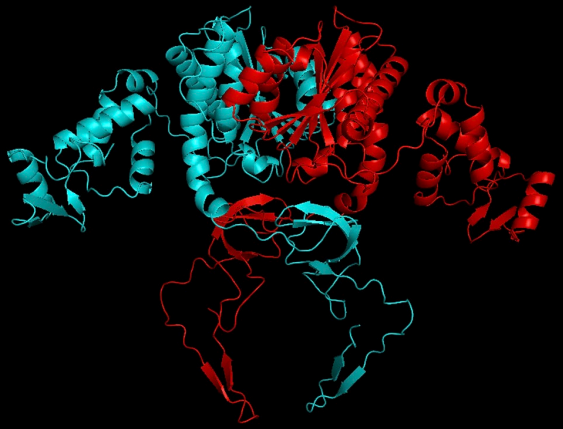
Aug 2021 AF and RoseTTAFold Servers *Google Cloud*

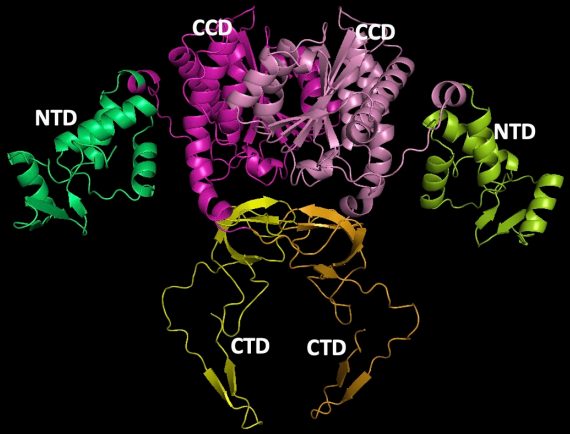
Aug 2021 20 Proteomes published in *Science*

Oct 2021 AF Running at RPI on CCI GPU Clusters

Jul 2022 200 M structures released by Deep Mind

MLV Integrase
408 aa X 2 = 816 aa ~ 90 kDa
Chain A in cyan and chain B in red





Eury DsrAB Disulfide Reductase
1400 residues ~ 155 kDa

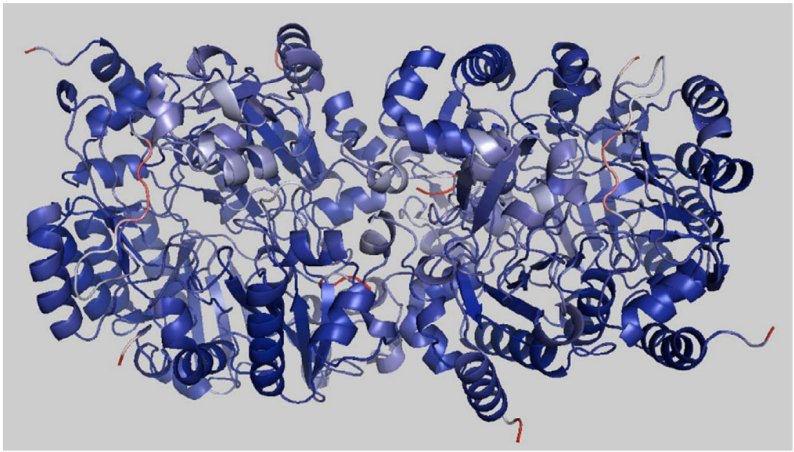


FIGURE 4 AlphaFold2 model of MV2.Eury DsrAB. The sequence is colored for pLDDT score (described in the text), with dark blue corresponding to high confidence prediction (96%) and red to low confidence prediction (43%). The view is a top view of the structure

GVT Swapna
C Royer

"We're releasing now the structures for the whole protein universe,"
said Demis Hassabis, founder and CEO of DeepMind, at a press conference in London.

Science Magazine "Breakthrough of 2021"

Science

Current Issue First re

HOME > SCIENCE > VOL. 374, NO. 6574 > PROTEINS, PROTEINS EVERYWHERE

EDITORIAL

Proteins, proteins everywhere

H. HOLDEN THORP

SCIENCE · 16 Dec 2021 · Vol 374, Issue 6574 · p. 1415 · DOI: 10.1126/science.abn5795

NEWS | TECHNOLOGY

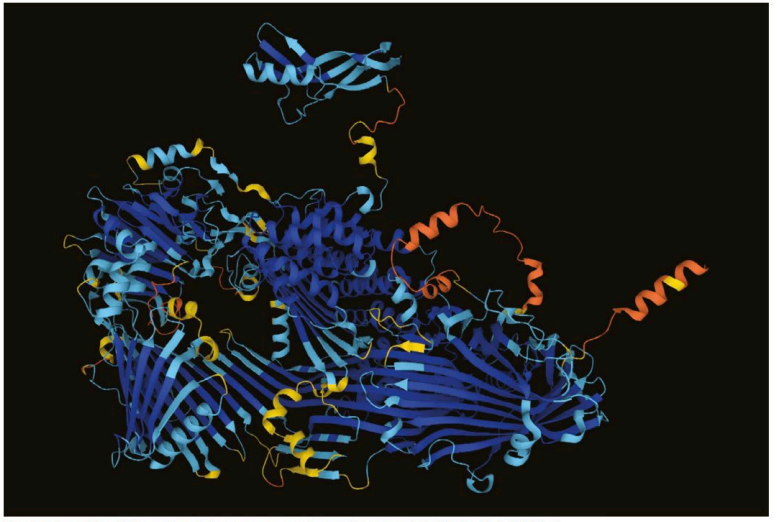
'New era in digital biology': AI reveals structures of nearly all known proteins

Advance from DeepMind's AlphaFold software could revolutionize biology and medicine

29 JUL 2022 · 11:25 AM · BY JOHN TRAVIS

The world this week

News in focus



The structure of the vitellogenin protein — a precursor of egg yolk — as predicted by the AlphaFold tool.

'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.

By Ewen Callaway

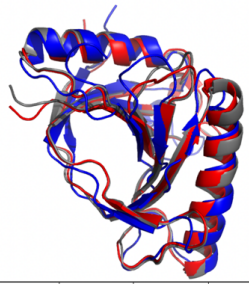
available for free in a database set up by Deep-Mind – the London-based AI company, owned

The 3D shape, or structure, of a protein is what determines its function in cells. Most

NESG NMR, X-ray Pairs

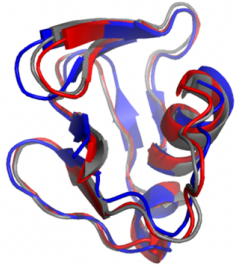
NMR
AF
Xray

CtR107



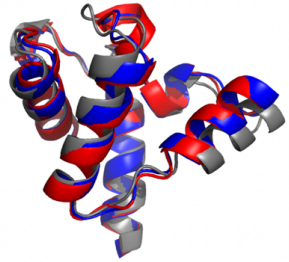
CtR107	2KCU	AF	X-ray
NMR 2KCU	2.11 Å	64.7	67.0
AF	3.89 Å	0.25 Å	97.3
X-ray 3E0H	3.62 Å	1.09 Å	----

GmR137



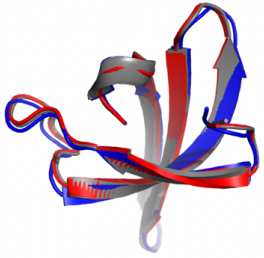
GmR137	2K5P	AF	X-ray
NMR 2K5P	0.78 Å	82.5	81.1
AF	1.52 Å	0.21 Å	98.8
X-ray 3CWI	1.59 Å	0.64 Å	----

RpR324



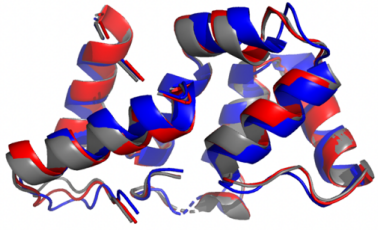
RpR324	7TZD	2LPK*	AF	X-ray
NMR 7TZD	0.54 Å	92.6	92.9	90.3
NMR 2LPK*	1.10 Å	0.45 Å	99.4	86.4
AF	1.13 Å	0.59 Å	0.27 Å	86.6
X-ray 3LMO	1.22 Å	1.68 Å	1.74 Å	----

SgR42



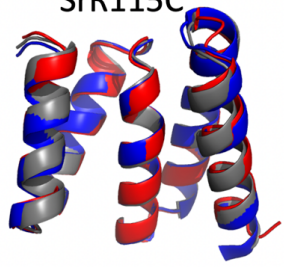
SgR42	2JZ2	AF	X-ray
NMR 2JZ2	0.66 Å	94.2	93.3
AF	0.86 Å	0.09 Å	99.6
X-ray 3C4S	0.97 Å	0.48 Å	----

SgR209C



SgR209C	2L06	7TZ8*	AF	X-ray
NMR 2L06	0.84 Å	82.7	82.7	82.2
NMR 7TZ8*	1.42 Å	0.94 Å	91.3	88.4
AF	1.51 Å	1.16 Å	0.11 Å	99.2
X-ray 3OSJ	1.53 Å	1.32 Å	0.52 Å	----

SrR115C



SrR115C	2KCL	2KCV*	AF	X-ray
NMR 2KCL	0.47 Å	92.9	93.4	87.4
NMR 2KCV*	1.03 Å	0.43 Å	95.6	95.8
AF	1.27 Å	1.28 Å	0.14 Å	94.7
X-ray 3MA5	1.35 Å	0.78 Å	1.13 Å	----

Tejero et al, Frontiers in Mol. Biol. 2022

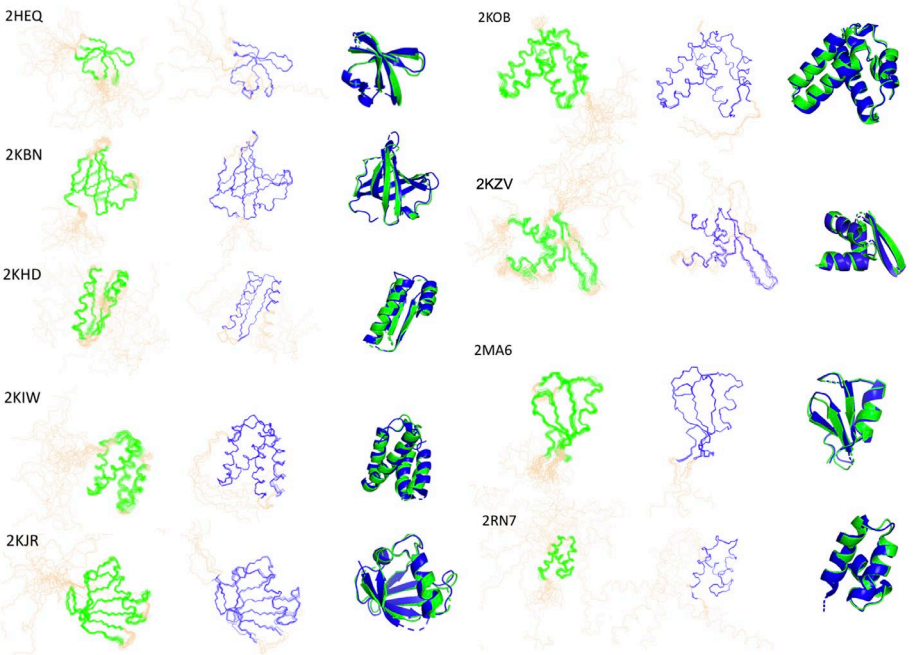
Contents lists available at ScienceDirect

Journal of Magnetic Resonance

journal homepage: www.elsevier.com/locate/jmr

Blind assessment of monomeric AlphaFold2 protein structure models with experimental NMR data

Ethan H. Li [a], Laura E. Spaman [a], Roberto Tejero [a], Yuanpeng Janet Huang [a], Theresa A. Ramelot [a], Keith J. Fraga [a], James H. Prestegard [b], Michael A. Kennedy [c], Gaetano T. Montelione [a,*]



Homo sapiens

Thrombomodulin

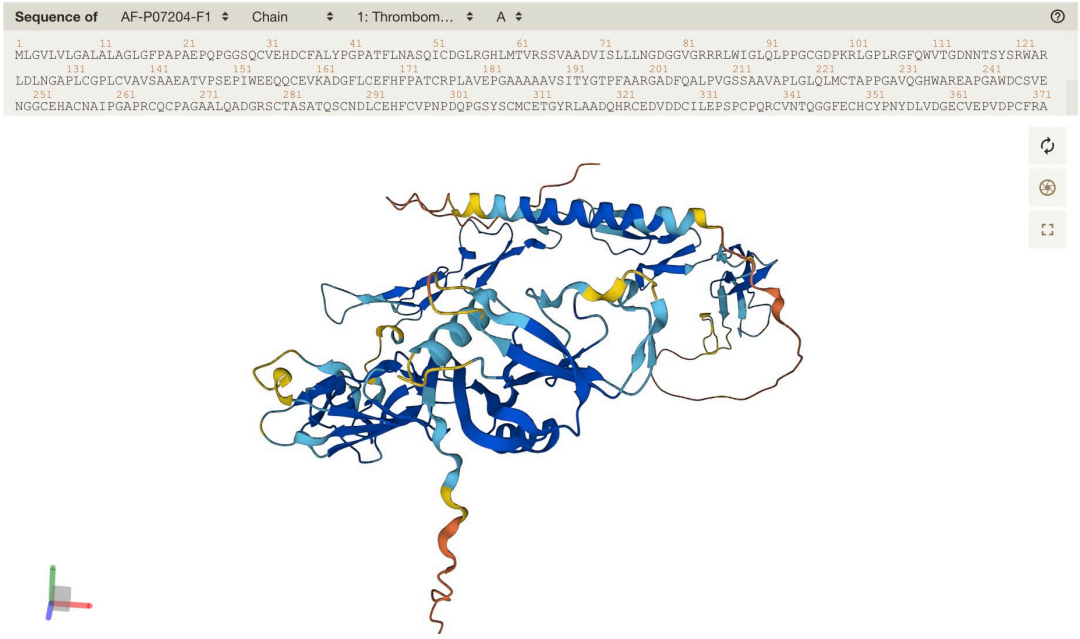
AlphaFold structure prediction

3D viewer

Model Confidence:

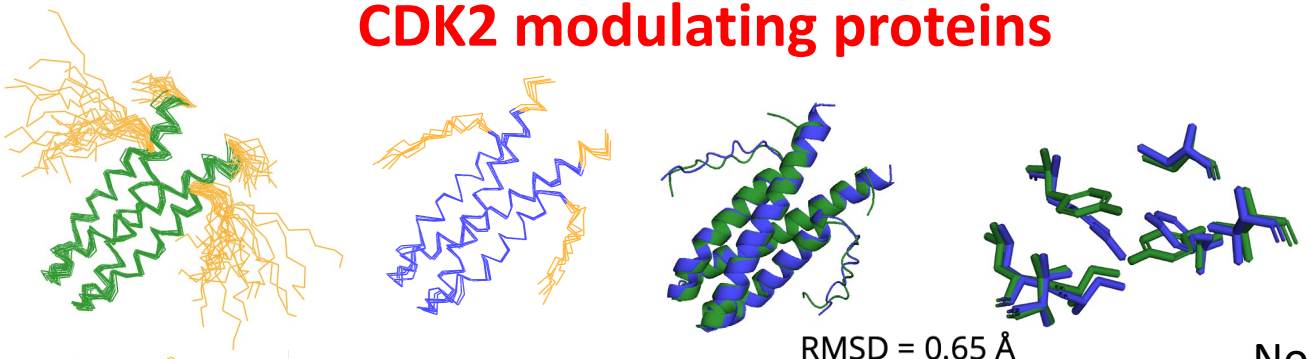
- Very high (pLDDT > 90)
- Confident (90 > pLDDT > 70)
- Low (70 > pLDDT > 50)
- Very low (pLDDT < 50)

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



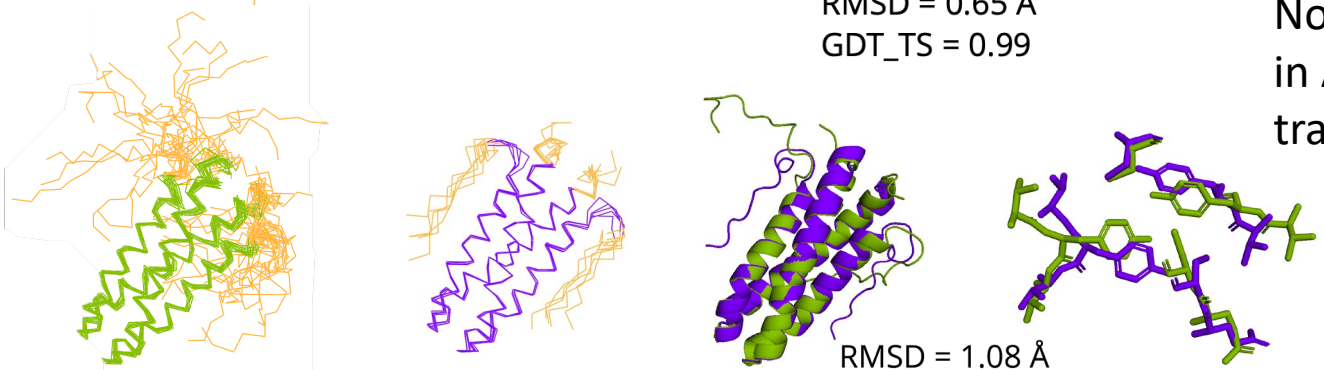
Jumper, J et al. Highly accurate protein structure prediction with AlphaFold. Nature (2021).
Varadi, M et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Research (2021).

DOC1 and DOC1R – Deleted in Oral Cancer. CDK2 modulating proteins



RMSD = 0.65 Å
GDT_TS = 0.99

Not used in AF training



RMSD = 1.08 Å
GDT_TS = 0.94

L Spaman

Protein	Method	<DP>[a]	DP_{avg} [a]	R_{avg} [a]	P_{avg} [a]	F_{avg} [a]	ProCheck - bb[b]	ProCheck - all[b]	Mol Probity [b]	Rama-chandran Statistics [c]
CDK2AP1	NMR	0.77	0.71	0.97	0.90	0.93	+2.44	+2.48	-0.84	99.3/0.6
	AF	0.69	0.68	0.95	0.86	0.90	+1.85	+1.89	+1.53	91.5/8.5
CDK2AP2	NMR	0.80	0.72	0.96	0.92	0.94	+2.36	+1.83	+0.08	98.9/1.1
	AF	0.77	0.73	0.96	0.90	0.93	+1.89	+1.42	+0.41	89.4/10.6

nature methods

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › nature methods › brief communications › article

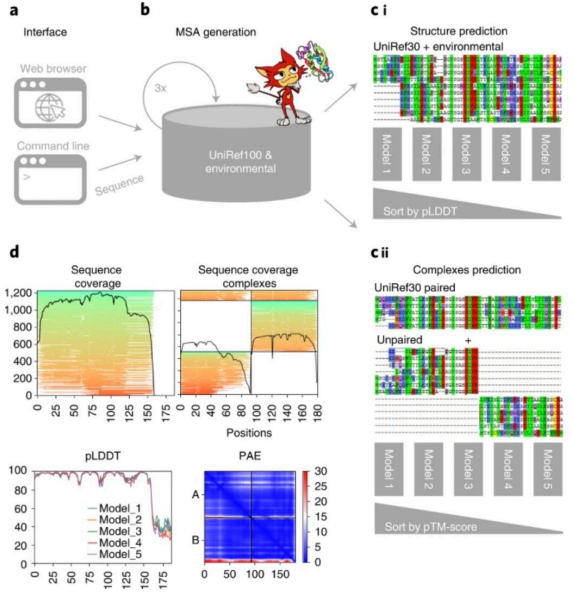
Brief Communication | Open access | Published: 30 May 2022

ColabFold: making protein folding accessible to all

Milot Mirdita ✉, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov ✉ & Martin Steinegger ✉

Nature Methods **19**, 679–682 (2022) | Cite this article

Fig. 1: Schematic diagram of ColabFold.



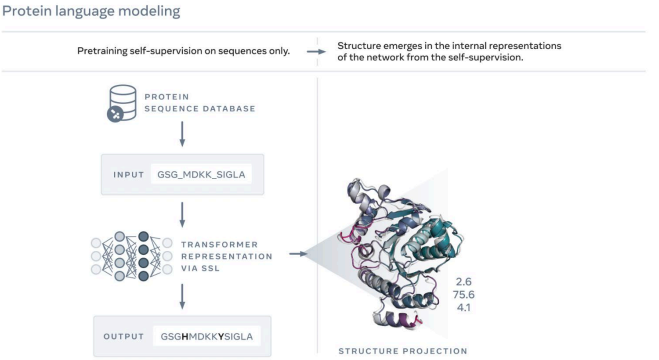
a,**b**, ColabFold has a web and a command line interface (**a**) that send FASTA input sequence(s) to an MMseqs2 server (**b**) searching two databases, UniRef100 and a database of environmental sequences,

Web based
AF2 modeling

Democatization
of
Structural
Biology

Evolutionary-scale prediction of atomic-level protein structure with a language model

ZEMING LIN, HALIL AKIN, ROSHAN RAO, BRIAN HIE, ZHONGKAI ZHU, WENTING LU, NIKITA SMETANIN, ROBERT VERKUIL, ORI KABELI, [...], AND ALEXANDER RIVES +5 authors Authors Info & Affiliations



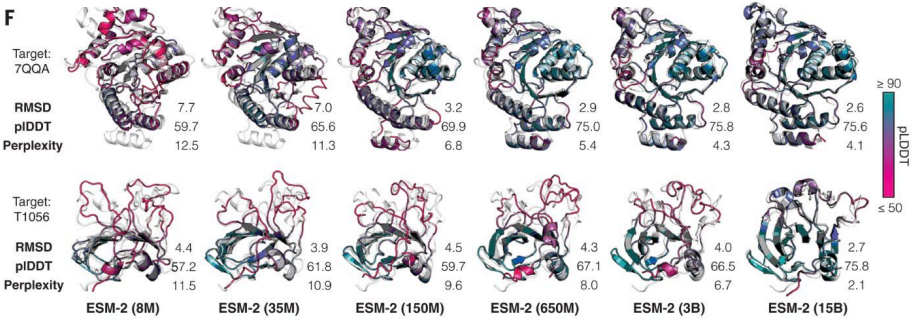


Fig. 1. Emergence of structure when scaling language models to 15 billion parameters.

The CASP15 rankings

#1 PEZYFoldings AF2-based. Diverse MSAs. Custom, fine-tuned AF2 refinement

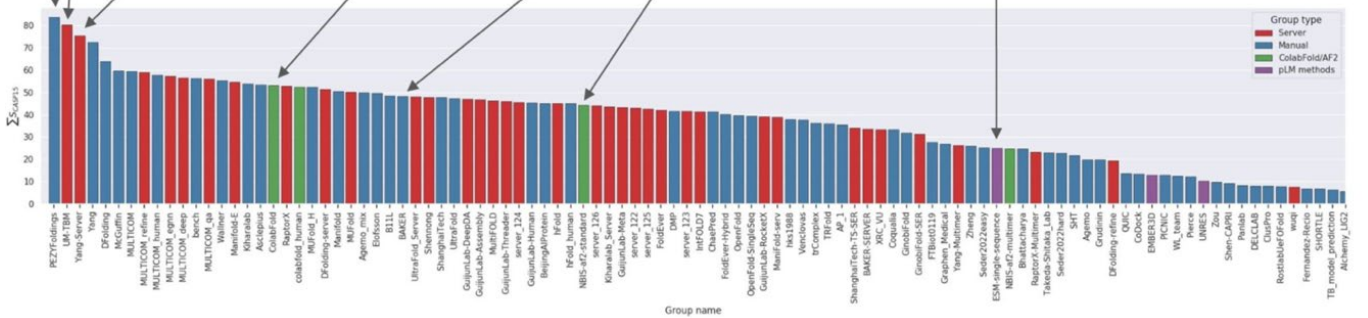
#2 UM-TBM Diverse MSAs. Threading then AF2 predictions guide I-TASSER REMC

#3 Yang-Server Diverse MSAs. AF2 predictions fed to trRosettaX2

ColabFold and NBIS-af2-standard

BAKER top non-AF2 method

ESM-singlesequence is the top pure pLM method by this metric



OpenFold

Democratizing AI for Biology

Why OpenFold?

Open Source

An open-source project that can be used and improved by academics and companies alike.

Weights Available

Utilize the existing pre-trained weights to get quickly get started fine-tuning your model.

Permissive License

A permissively licensed model that allows commercial and non-commercial use.

Training Pipeline

Provides the tools used to train the model under the same license.

Optimized for Performance

Optimized performance for use on state-of-the-art and widely available GPUs.

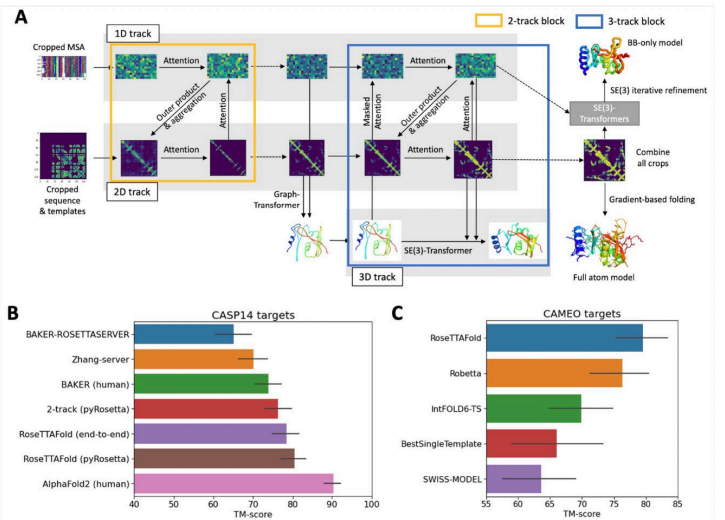
PyTorch-Based

A supercomputer scale, distributed training, PyTorch-based training framework

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



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

Protein-Protein Complexes

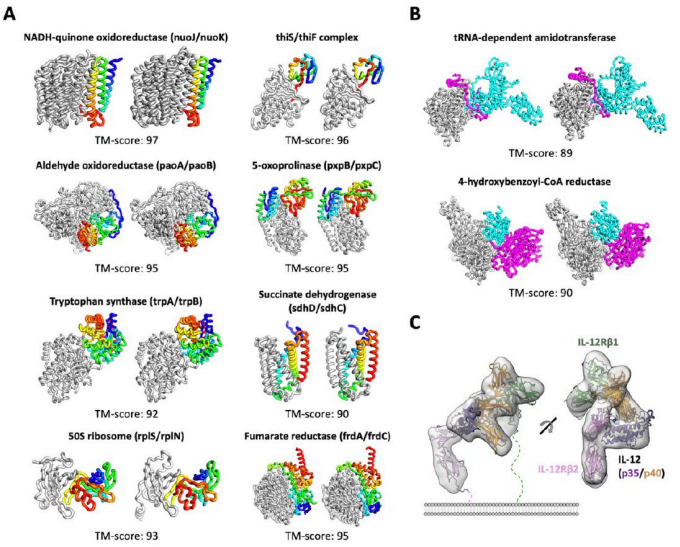


Fig. 4. Complex structure prediction using RoseTTAFold. (**A** and **B**) Prediction of structures of *E.coli* protein complexes from sequence information. Experimentally determined structures are on the left, RoseTTAFold models, on the right; the TM-scores below indicate the extent of structural similarity. (A) Two chain complexes. The first subunit is colored in gray, and the second subunit is colored in a rainbow from blue (N-terminal) to red (C-terminal). (B) Three chain complexes. Subunits are colored in gray, cyan, and magenta. (**C**) IL-12R/IL-12 complex structure generated by RoseTTAFold fits the previously published cryo-EM density (EMD-21645).

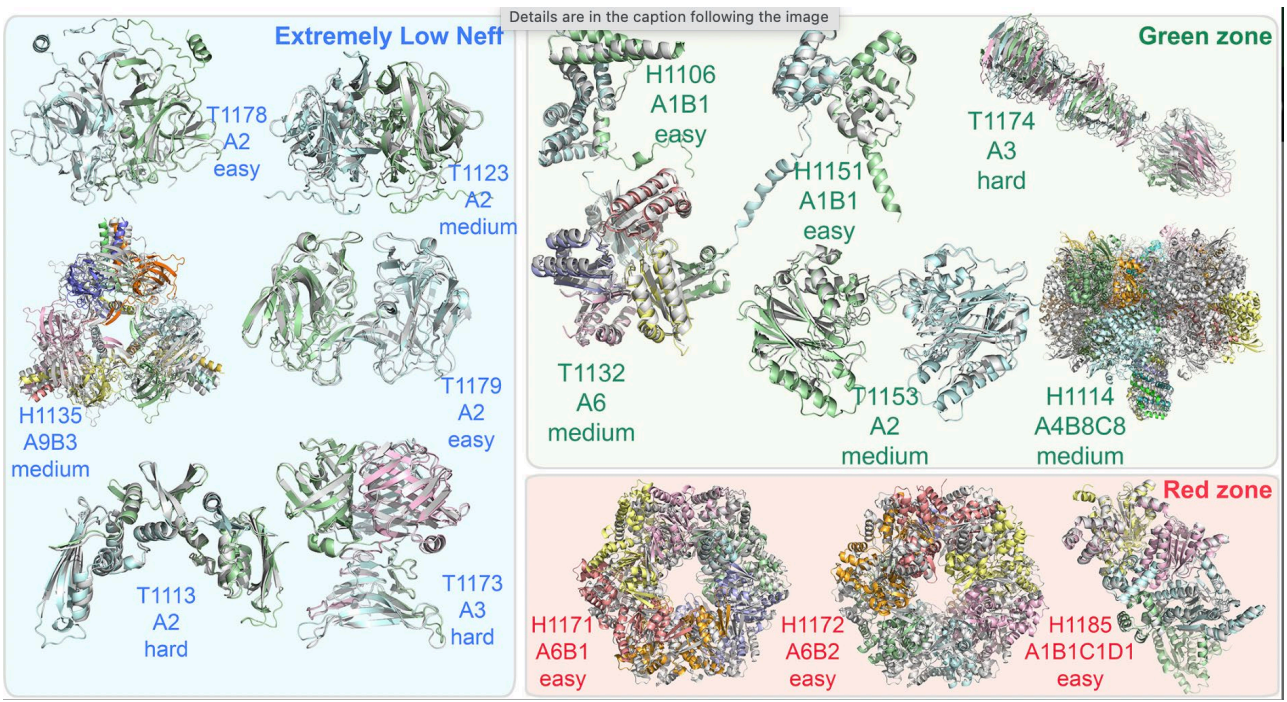
PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

RESEARCH ARTICLE | Open Access

The impact of AI-based modeling on the accuracy of protein assembly prediction: Insights from CASP15

Burcu Ozden, Andriy Kryshtafovych, Ezgi Karaca

First published: 20 October 2023 | https://doi.org/10.1002/prot.26598 | Citations: 2



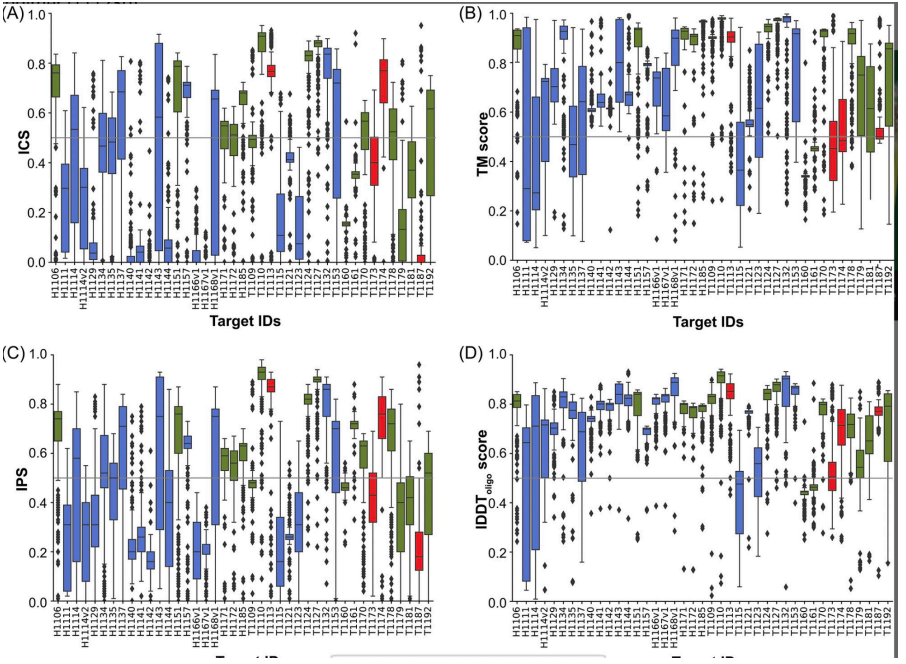
PROTEINS
STRUCTURE FUNCTION BIOINFORMATICS

RESEARCH ARTICLE | Open Access

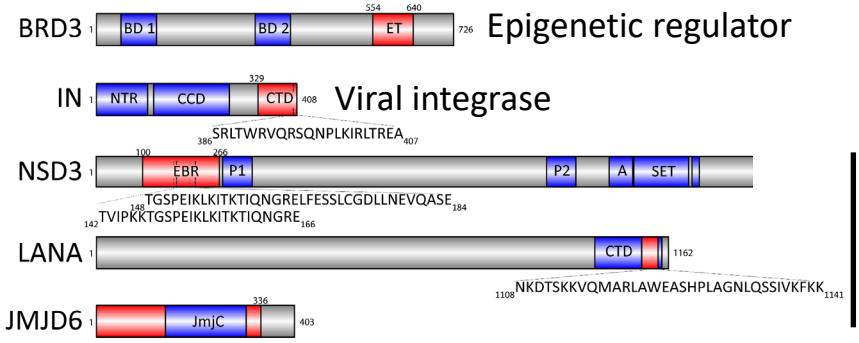
The impact of AI-based modeling on the accuracy of protein assembly prediction: Insights from CASP15

Burcu Ozden, Andriy Kryshtafovych, Ezgi Karaca

First published: 20 October 2023 | https://doi.org/10.1002/prot.26598 | Citations: 2



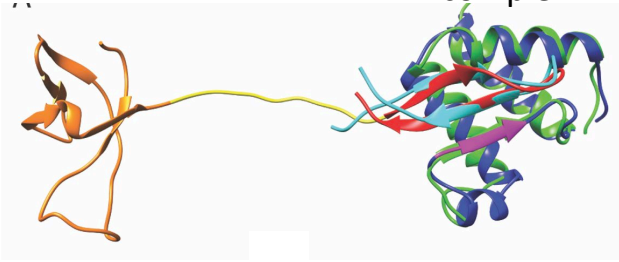
BET Proteins: ET : Peptide Complexes



Epigenetic regulators and viral proteins

MLV IN CTD

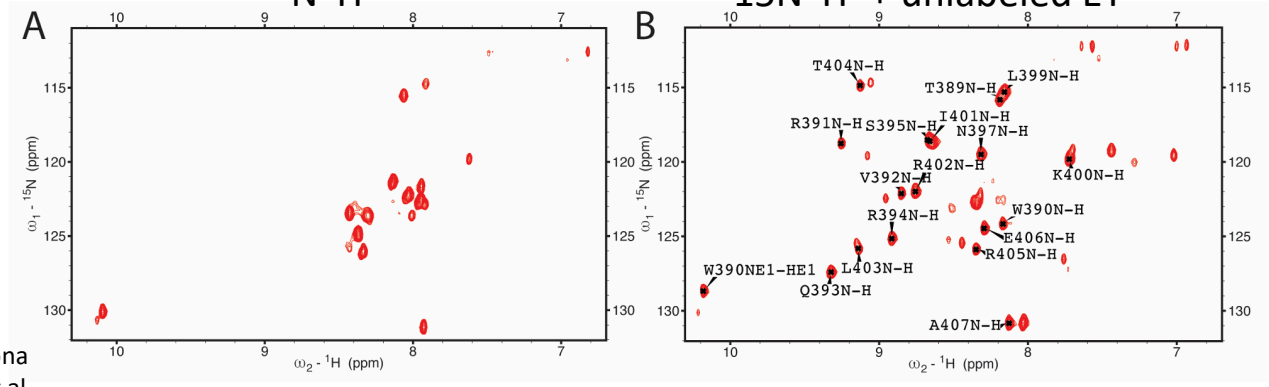
ET – TP complex



Disorder -> order

^{15}N-TP

15N-TP + unlabeled ET



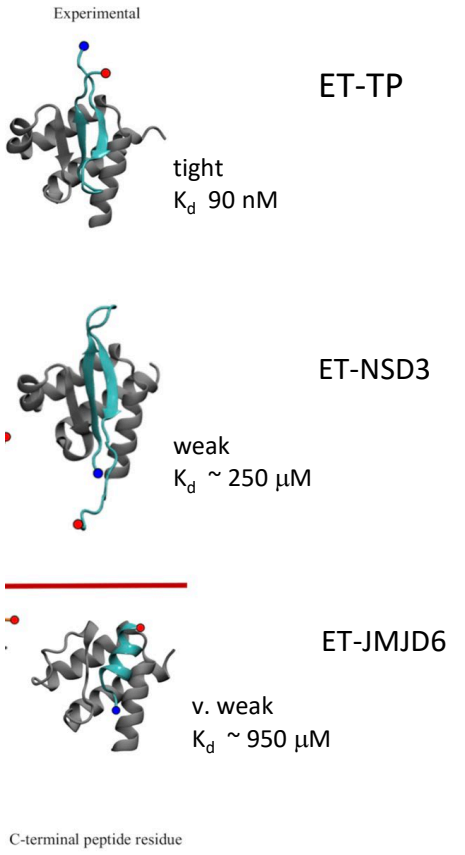
GVT Swapna
M Roth et al

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Structure Determination of Challenging Protein–Peptide Complexes Combining NMR Chemical Shift Data and Molecular Dynamics Simulations

Arup Mondal, G.V.T. Swapna, Maria M. Lopez, Laura Klang, Jingzhou Hao, LiChung Ma, Monica J. Roth,* Gaetano T. Montelione,* and Alberto Perez*

Disorder -> order



bioRxiv 2022

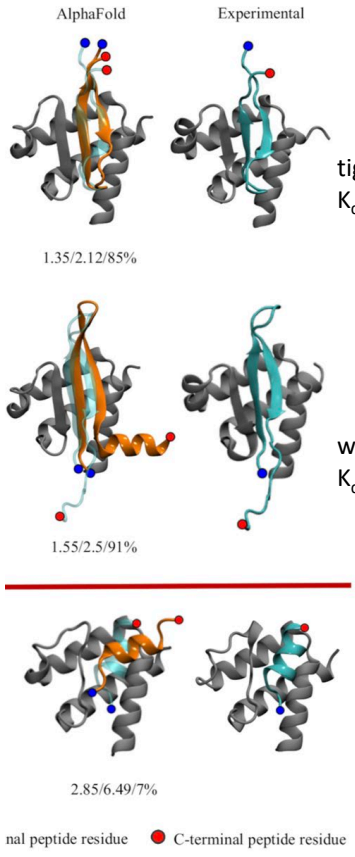
JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Structure Determination of Challenging Protein–Peptide Complexes Combining NMR Chemical Shift Data and Molecular Dynamics Simulations

Arup Mondal, G.V.T. Swapna, Maria M. Lopez, Laura Klang, Jingzhou Hao, LiChung Ma, Monica J. Roth,* Gaetano T. Montelione,* and Alberto Perez*

Disorder -> order

AF2 models of the peptide-receptor complex are an excellent match to experimental structures



ET-TP

tight
K_d 90 nM

ET-NSD3

weak
K_d ~ 250 μM

ET-JMJD6

v. weak
K_d ~ 950 μM

bioRxiv 2022

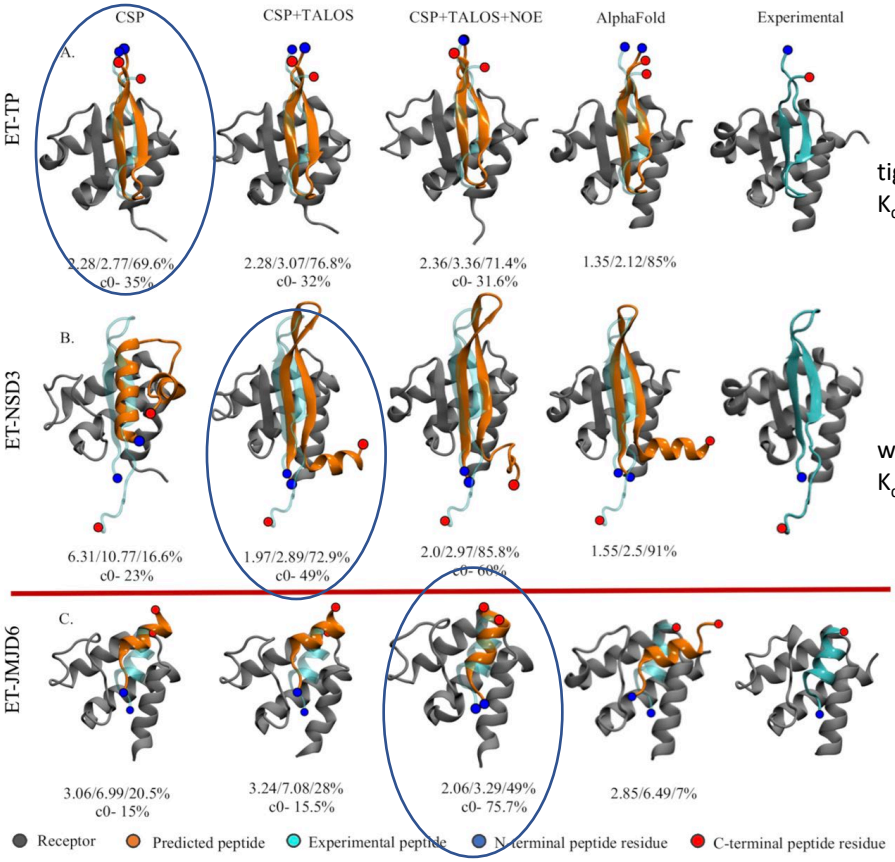
JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Structure Determination of Challenging Protein–Peptide Complexes Combining NMR Chemical Shift Data and Molecular Dynamics Simulations

Arup Mondal, G.V.T. Swapna, Maria M. Lopez, Laura Klang, Jingzhou Hao, LiChung Ma, Monica J. Roth,* Gaetano T. Montelione,* and Alberto Perez*

Combining MD replica exchange and NMR chemical shift perturbation data through Bayesian inference

Comparing MELD + NMR vs AlphaFold

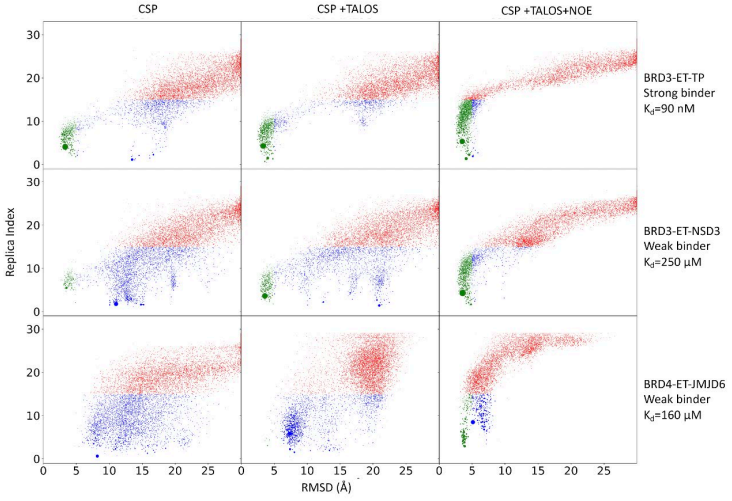


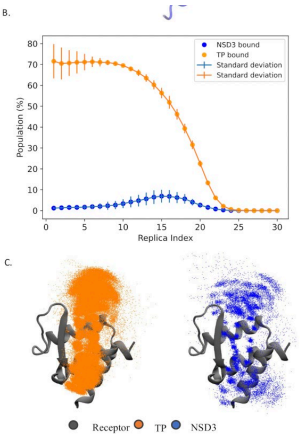
bioRxiv 2022

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Structure Determination of Challenging Protein–Peptide Complexes Combining NMR Chemical Shift Data and Molecular Dynamics Simulations

Arup Mondal, G.V.T. Swapna, Maria M. Lopez, Laura Klang, Jingzhou Hao, LiChung Ma, Monica J. Roth,* Gaetano T. Montelione,* and Alberto Perez*





Predicting relative binding affinities

$\Delta\Delta G_{pred}$ = - 2.4 kcal / mol

$\Delta\Delta G_{exp}$ = - 4.7 kcal / mol

AF2 for Docking -> MELD for docking free energy

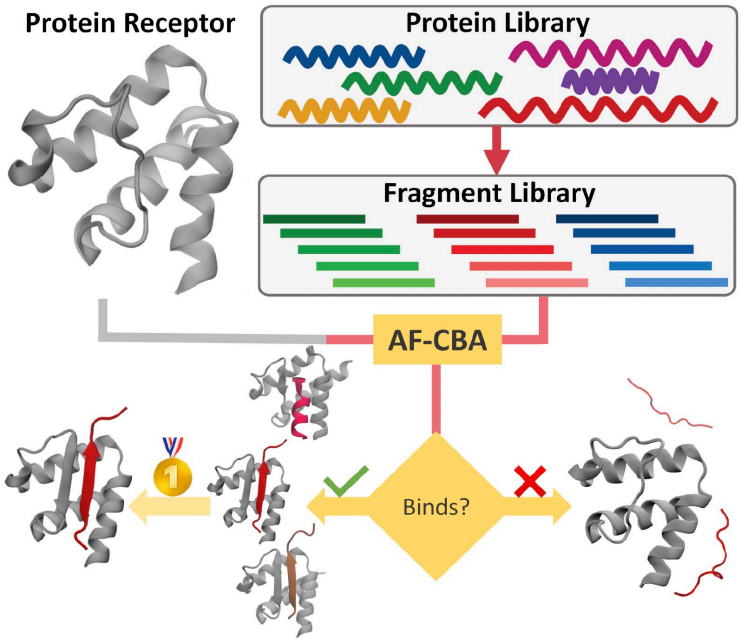
We are designing synthetic peptides that bind to ET and compete with natural ET binding proteins, by conventional de novo design and hallucination approaches.

- Antiviral or anticancer properties
- Create novel ET-protein complexes to modulate gene expression

Cold Spring Harbor Laboratory
bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

Sifting Through the Noise: A Computational Pipeline for Accurate Prioritization of Protein-Protein Binding Candidates in High-Throughput Protein Libraries

Arup Mondal, Bhumika Singh, Roland H. Felkner, Anna De Falco, GVT Swapna, Gaetano T. Montelione, Monica J. Roth, Alberto Perez



AF-CBA:

AlphaFold Competition Binding Assay

For 11 of 15 cases tested; AF-CBA identified correct region of the protein to bind to receptor, and correct experimental binding mode

A. Perez
A. Mondal
Univ of Florida

M. Roth - Rutgers

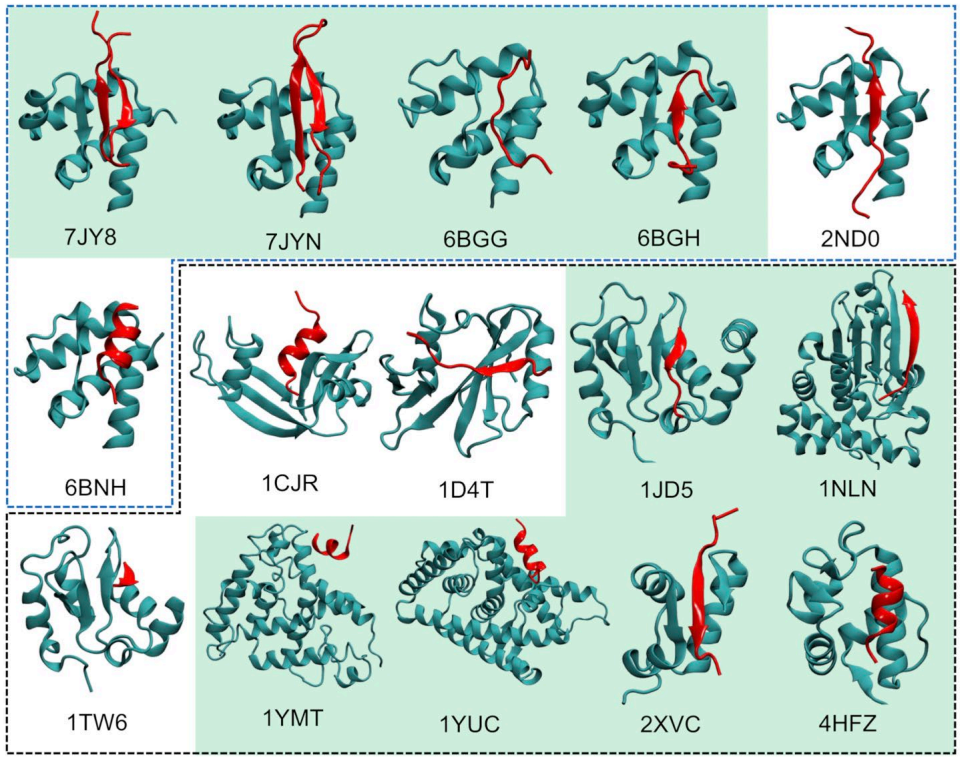
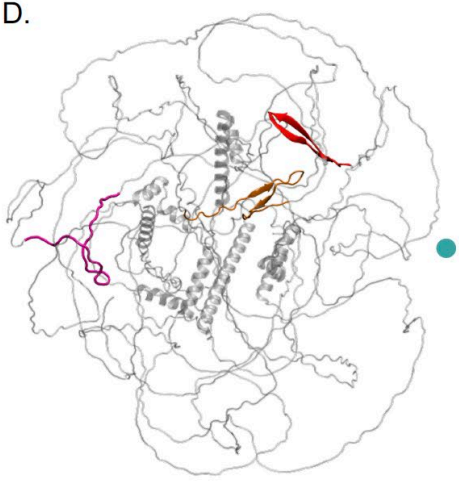
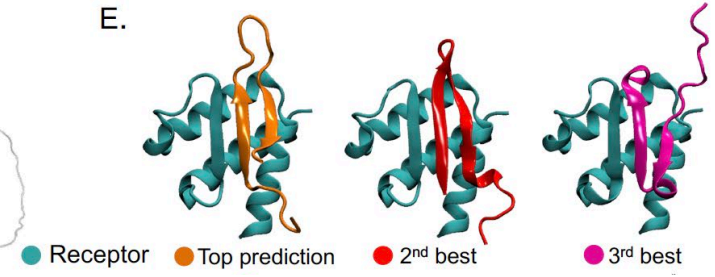


Figure 3. Test cases for AF-CBA pipeline. Experimental structures of protein complexes used in this study with their PDB code. The receptor protein is shown in cyan and peptides in red. The blue dashed line denotes systems involving the ET receptor (Proof of concept 1), whereas those inside the black dashed line show transferability to other systems (Proof of concept 2). Successful cases are highlighted with green background, new insights and limitations are discussed in the text for those with a white background. These representations are rendered with VMD[26].

D.

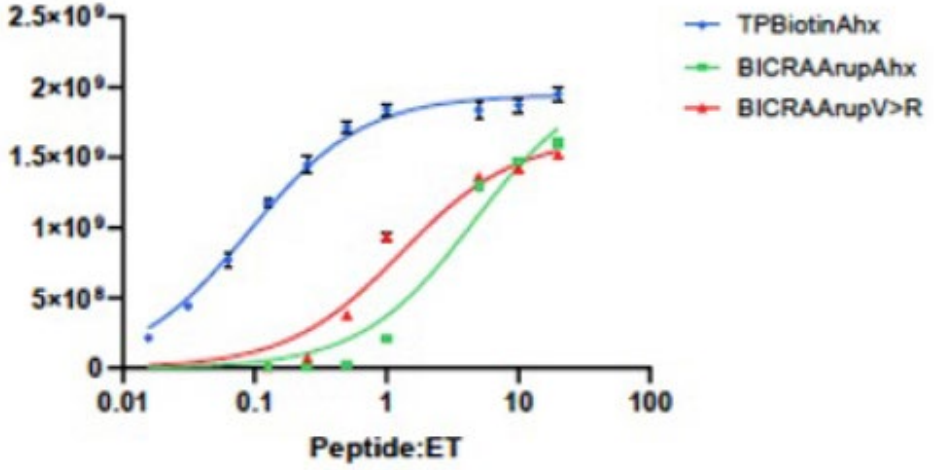


E.



Predicted BIRCA binding peptides with ET receptor

BICRA BET-binding protein identified by pull down



NESG BIOMEDICAL THEME PROJECT

Human Cancer Protein Interaction Network (HCPIN)

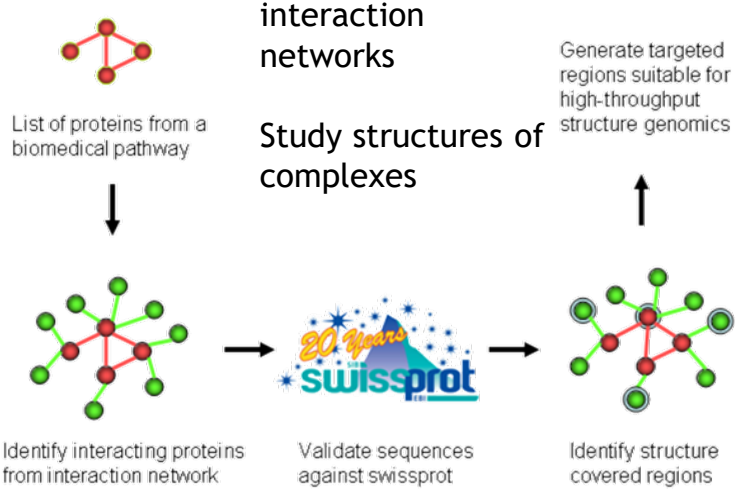
KEGG Pathways

- Apoptosis
- TGF-beta
- Innate Immune Response
- EGF receptor pathway
- NF-κB
- JAK/Stat
- MAPK

Protein Interaction Data from HPRD: Manually curated db of protein-protein interactions based largely on literature, describing physical protein- protein interactions

Systematic structural coverage of pathways and interaction networks

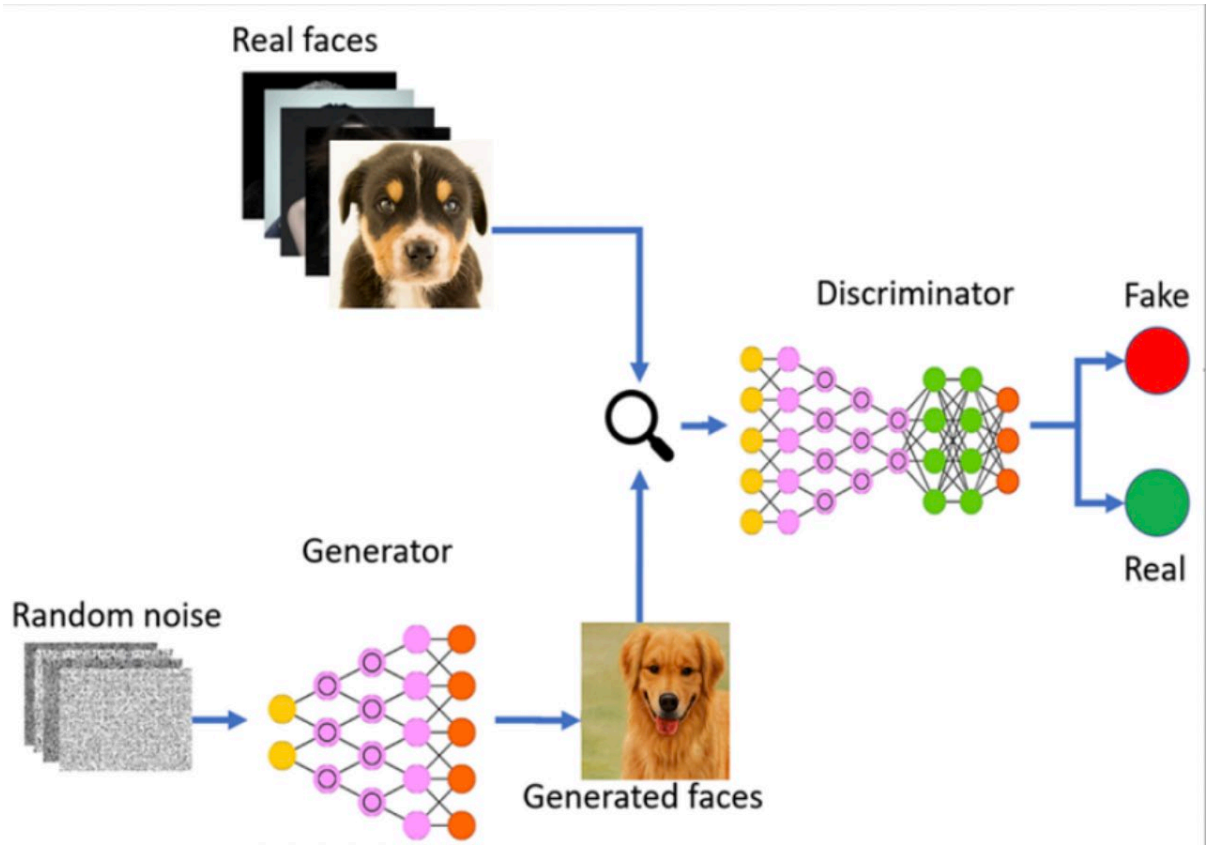
Study structures of complexes



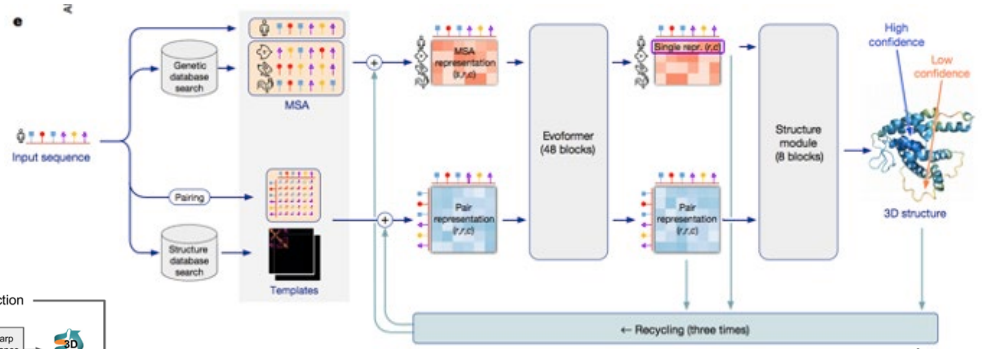
Targeting the Human Cancer Pathway Protein Interaction Network by Structural Genomics*

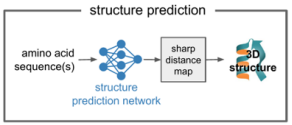
Yuanpeng Janet Huang‡, Dehua Hang‡, Long Jason Lu§¶, Liang Tong‖, Mark B. Gerstein§**, and Gaetano T. Montelione‡ ‡‡

Generative Adversarial Networks (GANS) - Hallucination



Goodfellow et al 2014





Jumper et al Nature 2021

nature

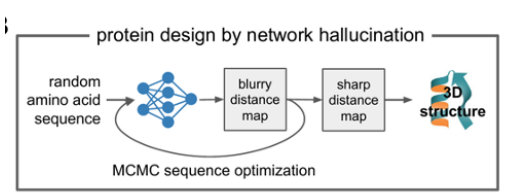
Explore content · About the journal · Publish with us

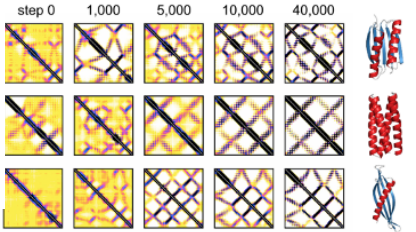
nature › articles › article

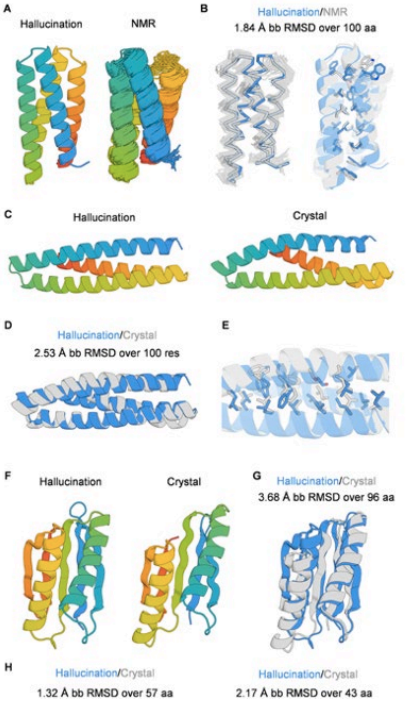
Article | Published: 01 December 2021

De novo protein design by deep network hallucination

Ivan Anishchenko, Samuel J. Pellock, Tamuka M. Chidyausiku, Theresa A. Ramelot, Sergey Ovchinnikov, Jingzhou Hao, Khushboo Bafna, Christoffer Norn, Alex Kang, Asim K. Bera, Frank DiMaio, Lauren Carter, Cameron M. Chow, Gaetano T. Montelione & David Baker





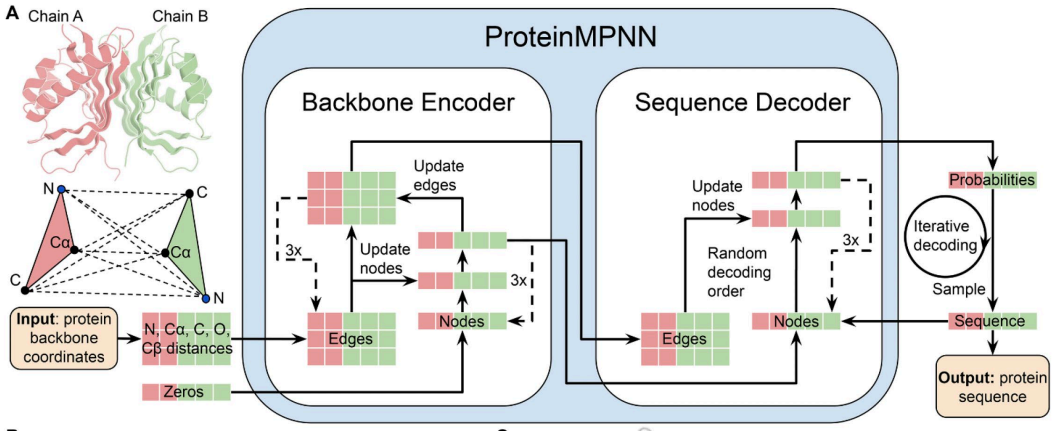


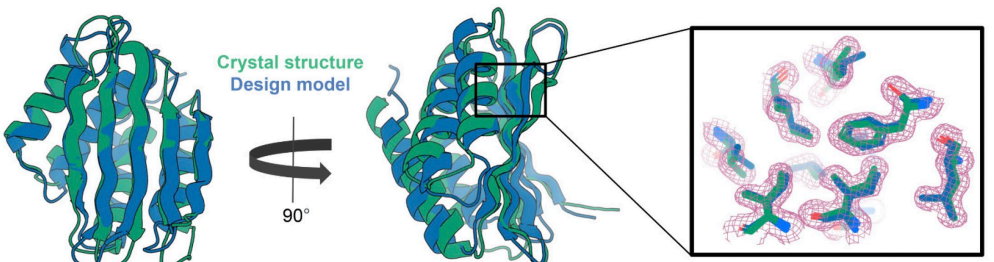
T Ramelot

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]*



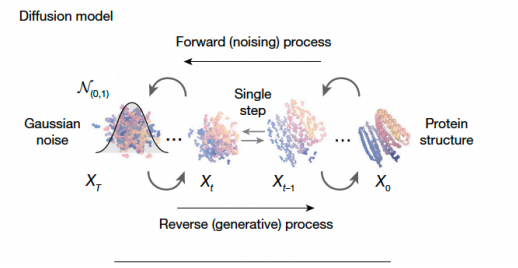


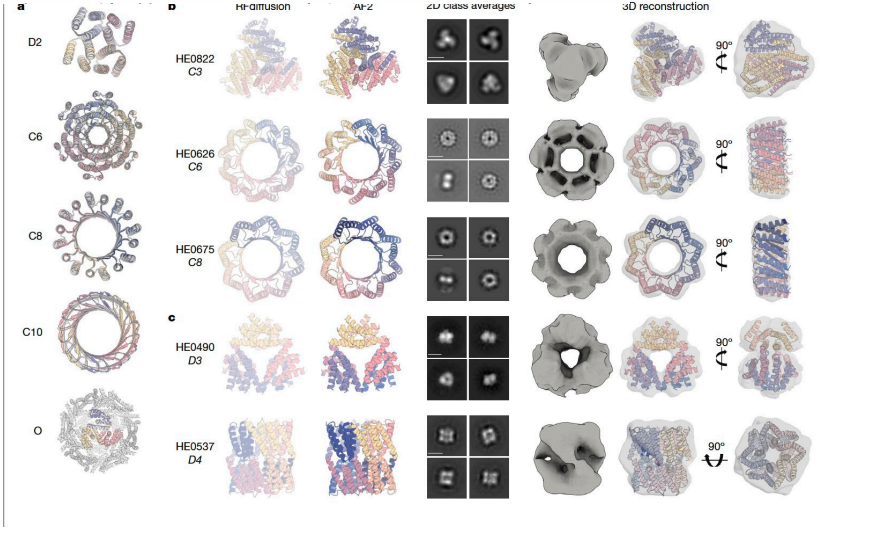
Article

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
Check for updates

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] ✉





Article

De novo design of protein structure and function with RFdiffusion

https://doi.org/10.1038/s41586-023-06415-8

Received: 14 December 2022

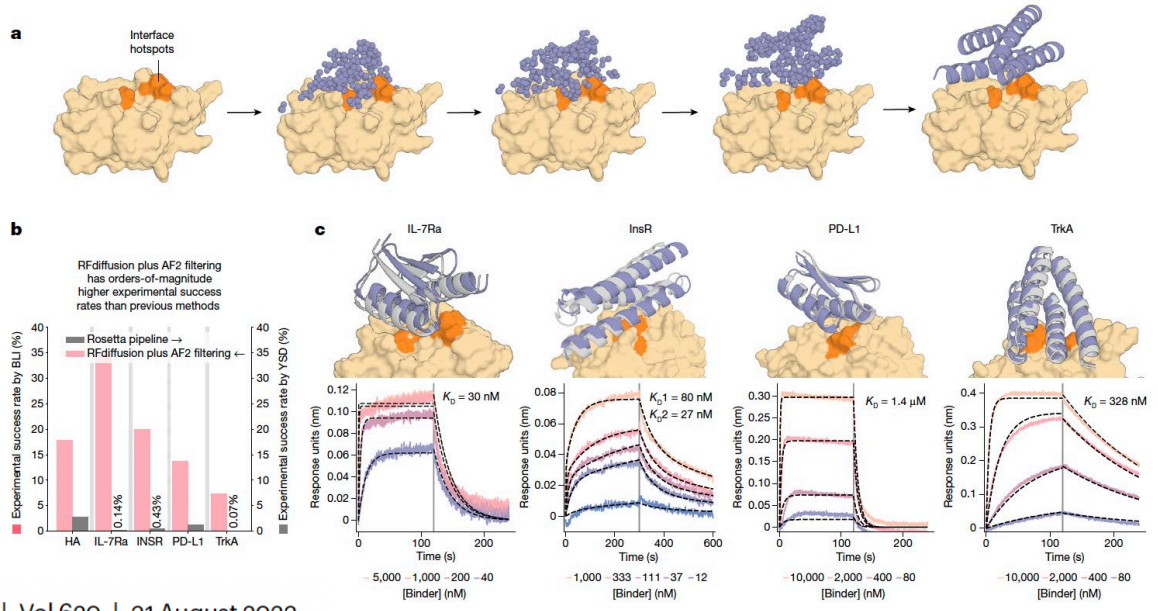
Accepted: 7 July 2023

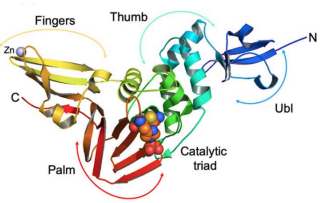
Published online: 11 July 2023

Open access

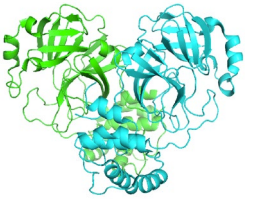
Check for updates

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] ✉





SARS CoV2 PL^{pro}

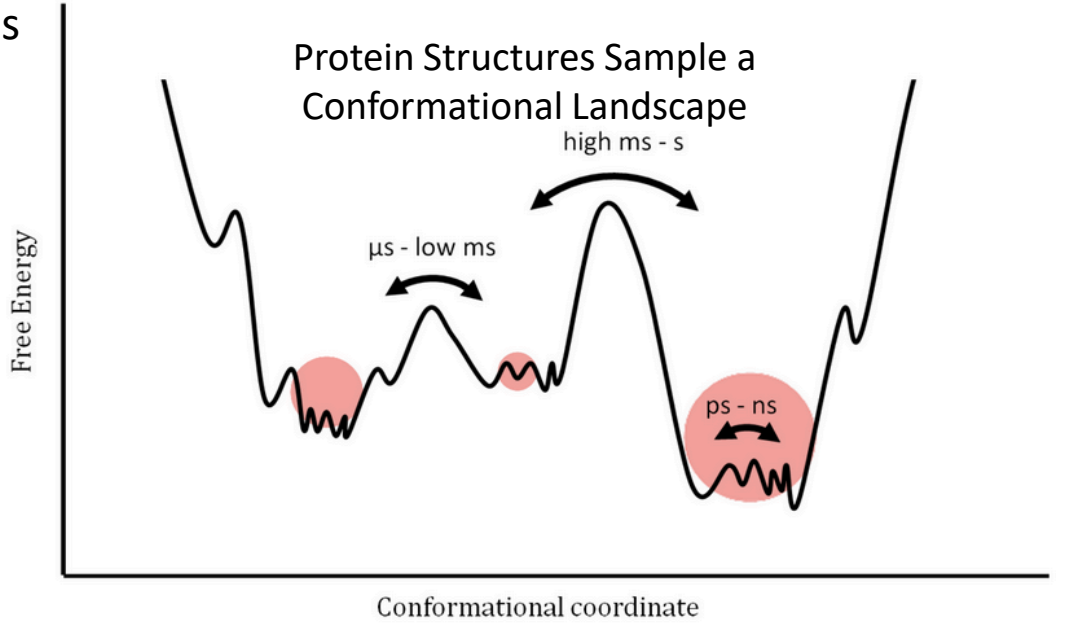


SARS CoV2 M^{pro}

4D Protein Structures

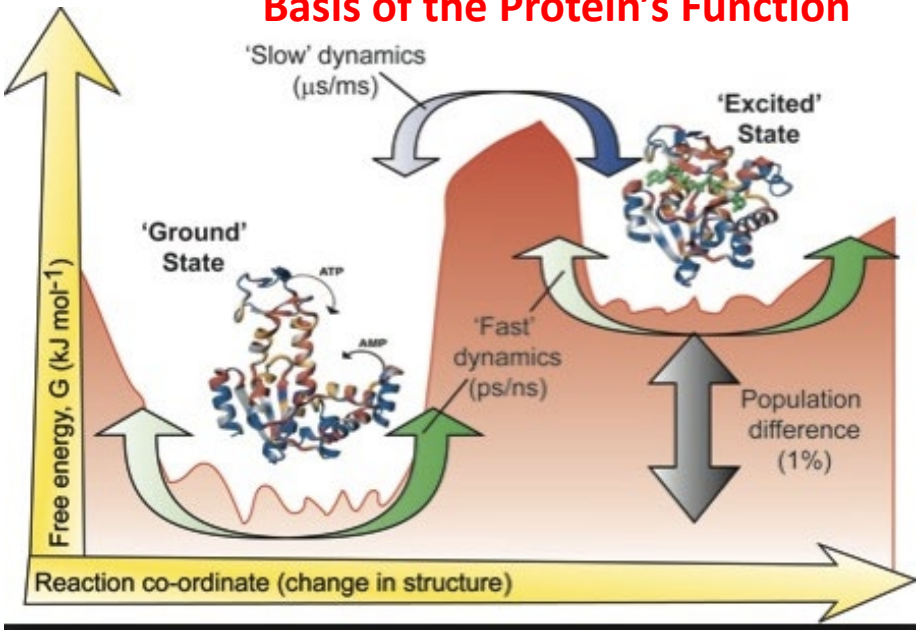
Protein Dynamics

Protein Structures Sample a Conformational Landscape

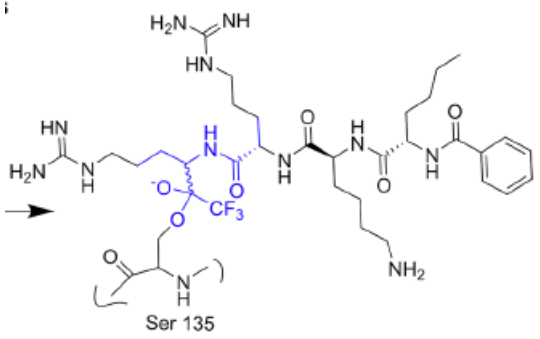


www.mdpi.com

The Conformational Landscape Provides the Basis of the Protein's Function



Andrew Baldwin
research.chem.ox.ac.uk



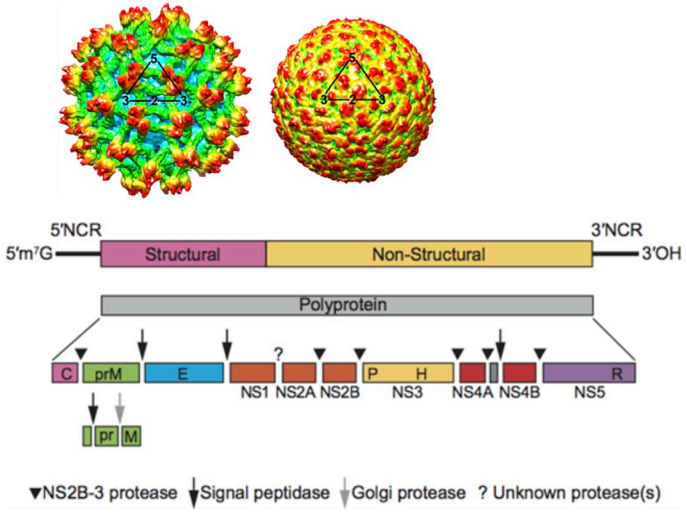


Figure 1 Three-dimensional cryo-electron microscopic reconstructions of immature (left) and mature (right) particles of an isolate of dengue virus (courtesy of M. Rossmann). Shown is a surface rendering of immature dengue virus at 12.5A resolution (left) and mature DENV at 10A resolution (right). The viruses are depicted to scale, but not colored to scale. Triangles outline one icosahedral unit.

Al Gibbs, Ruth Steele
Janssen Pharma

Gaohua Liu
Nexomics Biosciences

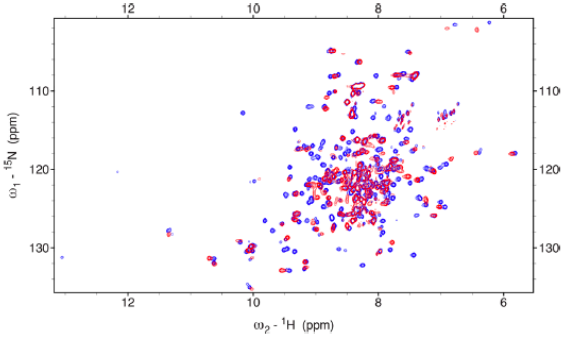
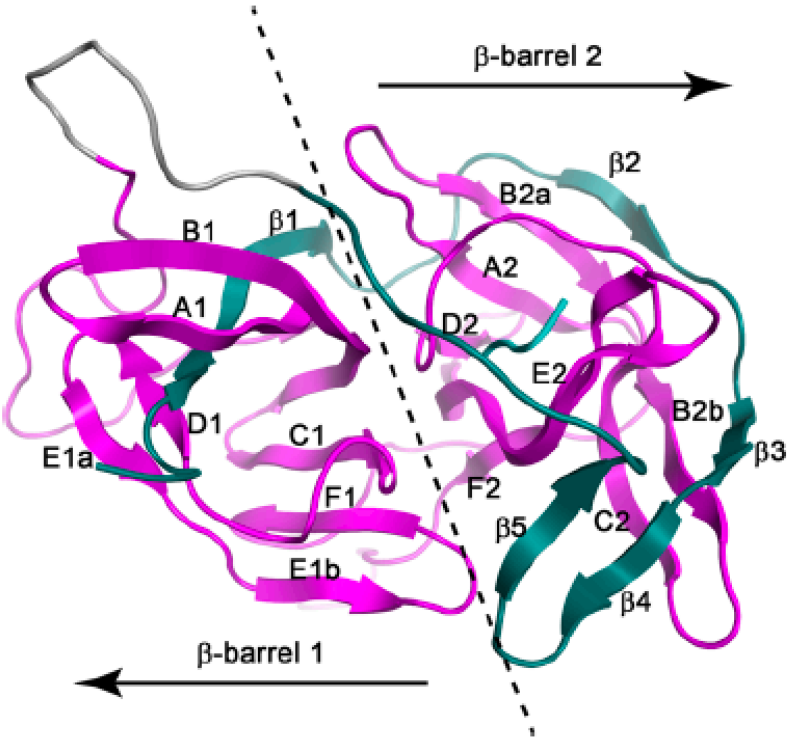
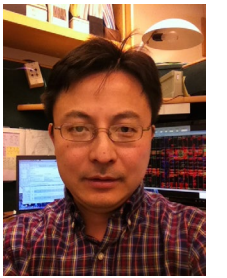


Figure 4. Comparison of [^{15}N–^{1}H]-TROSY-HSQC spectra of apo (red) and inhibitor-bound (blue) NS2B-SN3pro. Extensive exchange broadening limits the quality of the spectrum of apo NS2B-NS3pro, while significantly more well-resolved resonances are observed in the inhibitor-bound spectra.

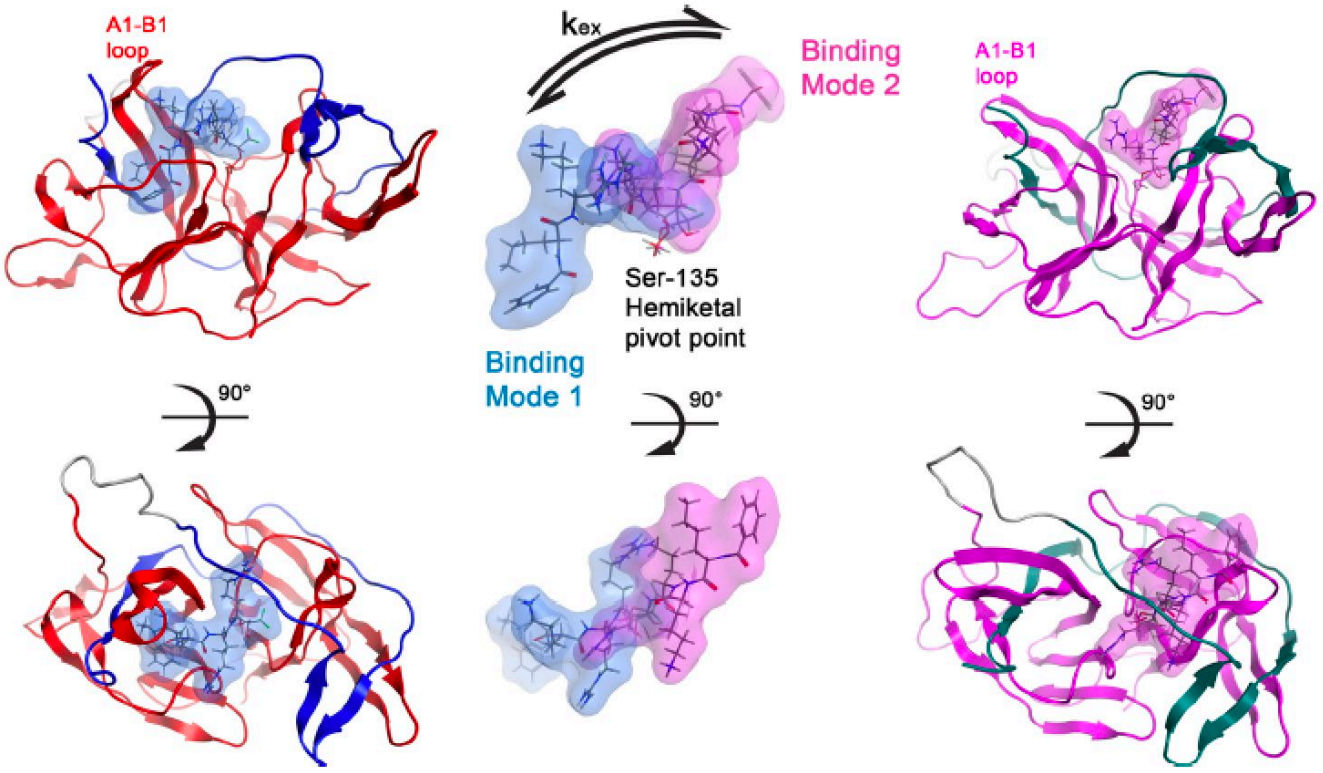
Solution NMR Structure of DENV2-NS2B-NS3pro Protease Complex $^{2}H,^{15}N,^{13}C$, ILVA Me Labeled

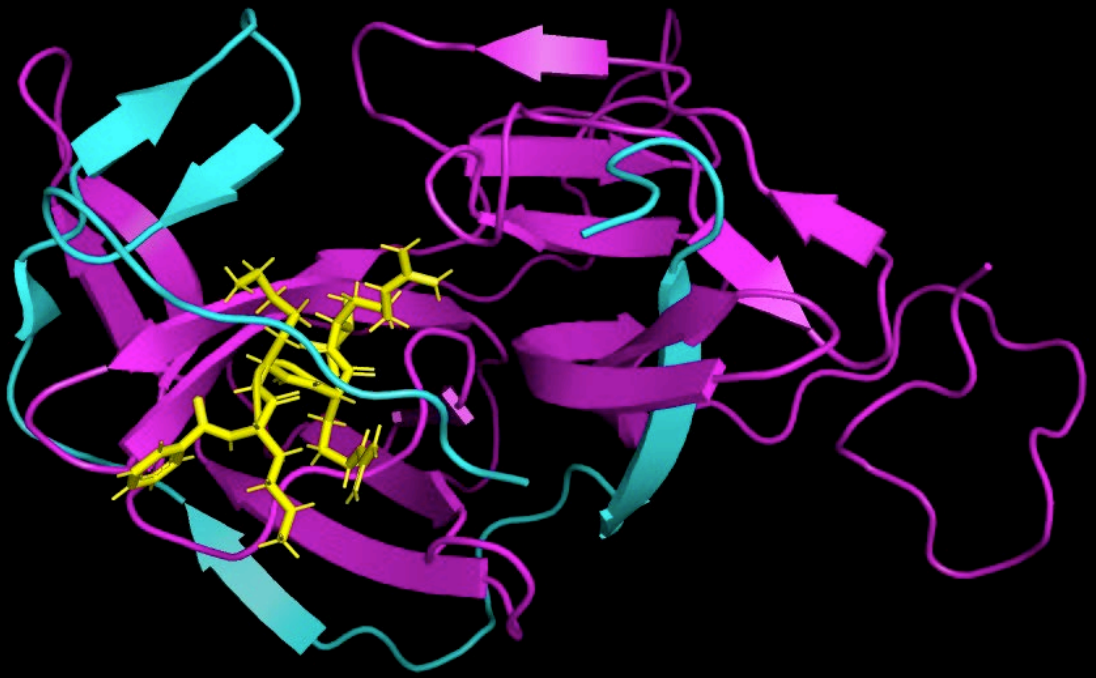




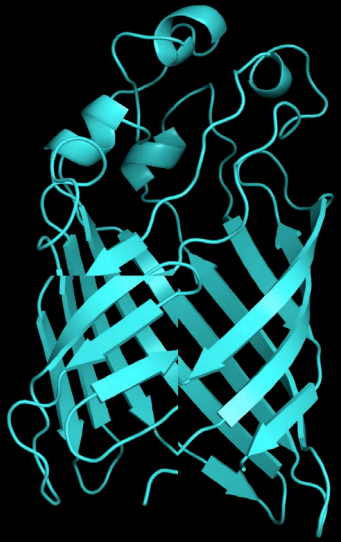
G. Liu

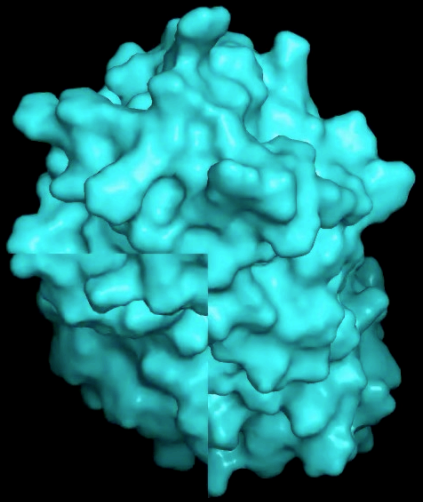
NMR Reveals Two Non-Overlapping Inhibitor Binding Sites in DENV2-NS2B-NS3pro Protease Complex





Klebsiella pneumoniae MipA in detergent micelles



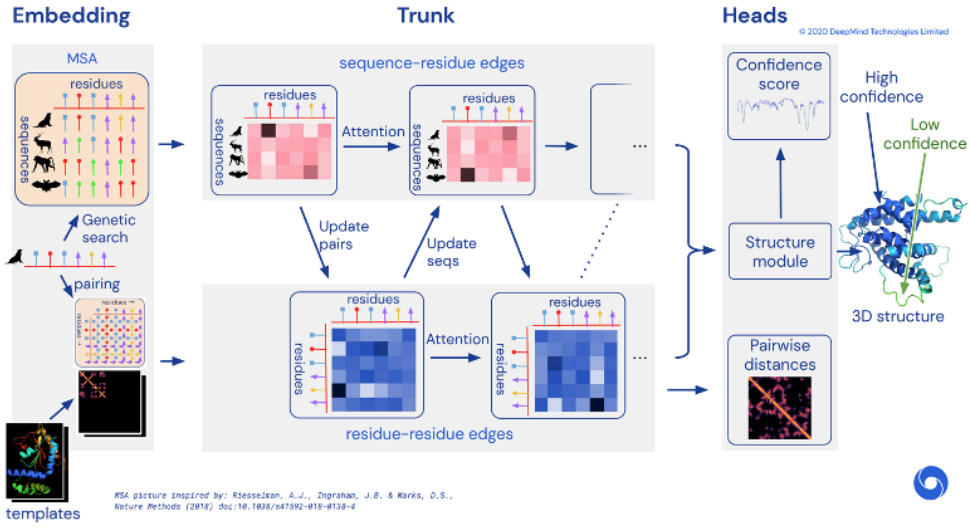


YJ Huang
G Liu
Y Ishida
GVT Swapna
S. McCallum

Can we use AlphaFold to Predict the Multiple Conformational States of MipA?

attention-based machine learning

Use shallow MSAs to provide subsets of ECs



Jumper et al 2021

John Jumper
CASP14 Talk Dec 2020

eLife

NEWSLETTER ABOUT

HOME MAGAZINE COMMUNITY INNOVATION

Short Report
Structural Biology and Molecular Biophysics

Sampling alternative conformational states of transporters and receptors with AlphaFold2

Diego del Alamo, Davide Sala, Hassane S Mchaourab, Jens Meiler

AF-Alt

nature

Article | Open access | Published: 13 November 2023

Predicting multiple conformations via sequence clustering and AlphaFold2

Hannah K. Wayment-Steele, Adedolapo Ojoawo, Renee Otten, Julia M. Apitz, Warintra Pitsawong, Marc Hömberger, Sergey Ovchinnikov, Lucy Colwell & Dorothee Kern

Nature **625**, 832–839 (2024) | Cite this article

AF-cluster

Bioinformatics

AFsample: improving multimer prediction with AlphaFold using massive sampling

Björn Wallner

Bioinformatics, Volume 39, Issue 9, September 2023, btad573,

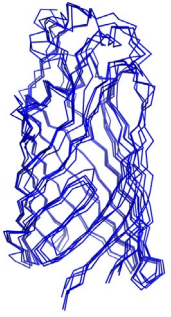
AF-sample

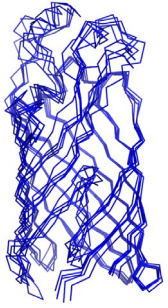
AF_ALT generates three clusters of models structural variation in the strand β2, β3, β4, β5

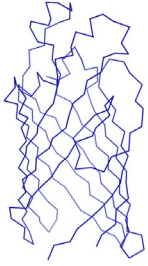
AF_ALT_1
cluster 1

AF-ALT_2
cluster 2

AF_ALT_3
cluster 3

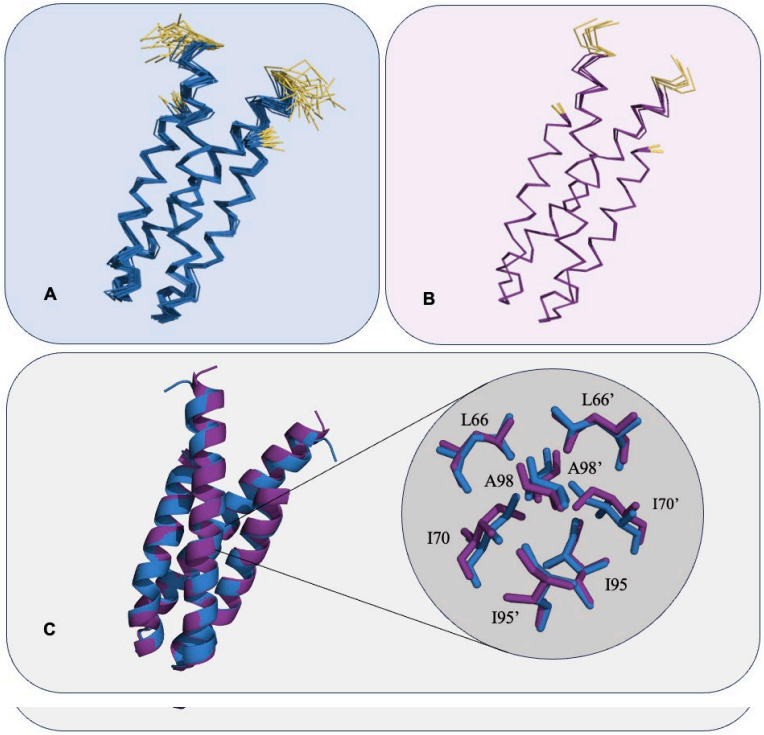


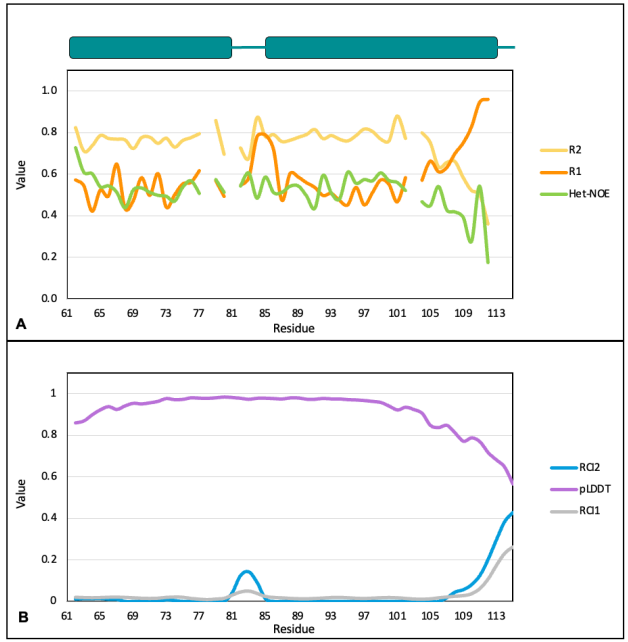


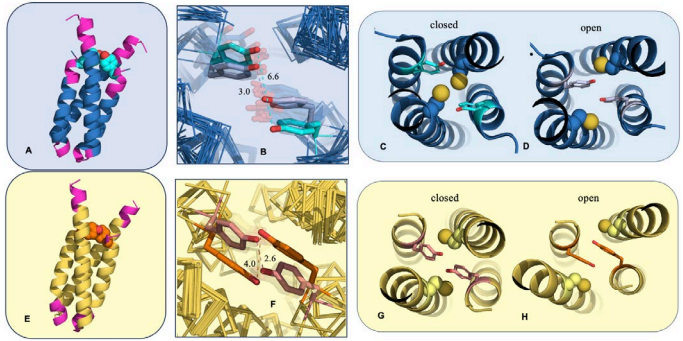


YJ Huang

Human DOC1 – Deleted in Oral Cancer





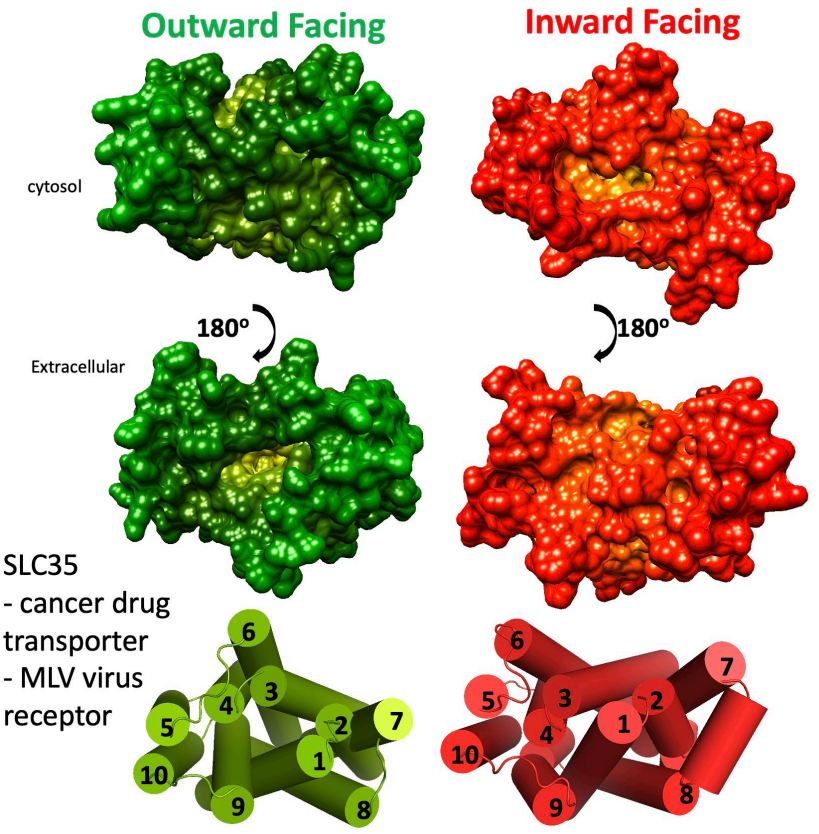


AlphaFold2-sample – models conformational dynamics that match well to experimental data

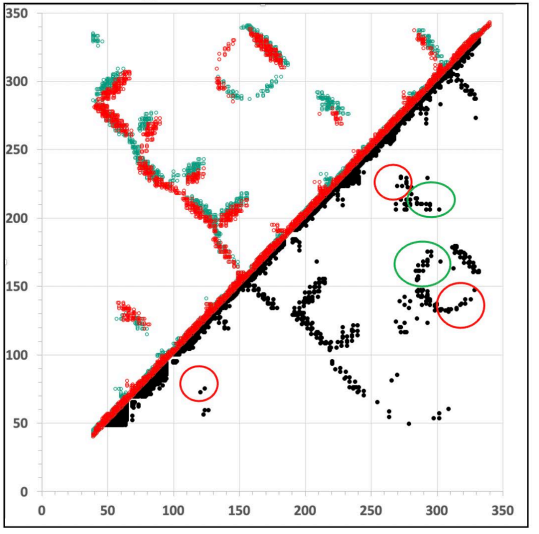
L Spaman
YJ Huang
R Tejero

Modeling Alternative Conformational States of Pseudo-Symmetric Membrane Protein Transporters using Methods from Machine Learning

G.V.T. Swapna, Namita Dube, Monica J. Roth, and Gaetano T. Montelione



AF-alt predicts both outward and inward facing conformations



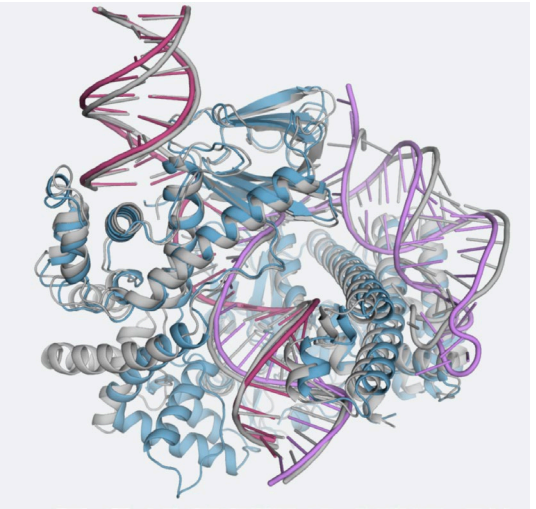
Some ECs are consistent only with outward facing – some only with inward facing

Mutations based on ECs can be used to shift preference from outward to inward

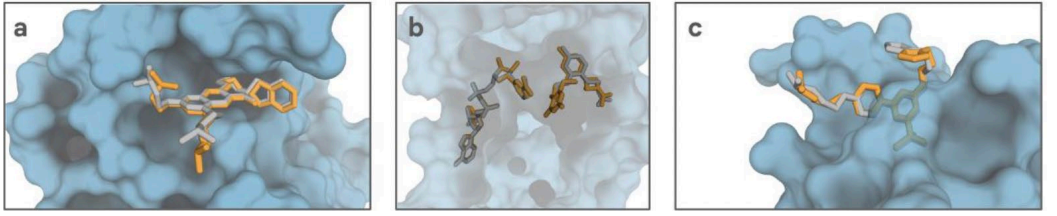
Performance and structural coverage of the latest, in-development AlphaFold model

Google DeepMind AlphaFold Team[1] and Isomorphic Labs Team[2]
[1]DeepMind, London, UK, [2]Isomorphic Labs, London, UK

bioRxiv
Nov 2023

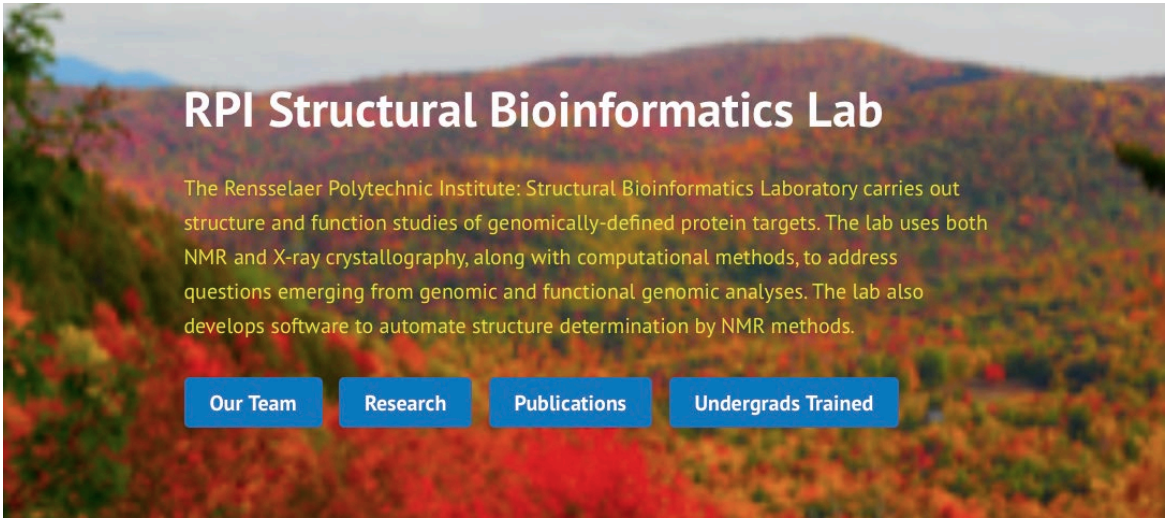


Protein:Nucleic Acid Complex: AF3 vs ground truth



Small molecule binding: AF3 vs ground truth

BIG DATA



YJ Huang
T Ramelot
GVT Swapna
R Tejero
G Liu
T Acton

B Shurina
L Spaman
N Dube
A DeFalco
L Klang
R Greene-Cramer

A Perez
A Mondal

M Roth
S Aiyer

C Sander
D Marks
K Brock

D Baker
S Pellock
I Anishchenko