



The machine learning revolution in structural biology

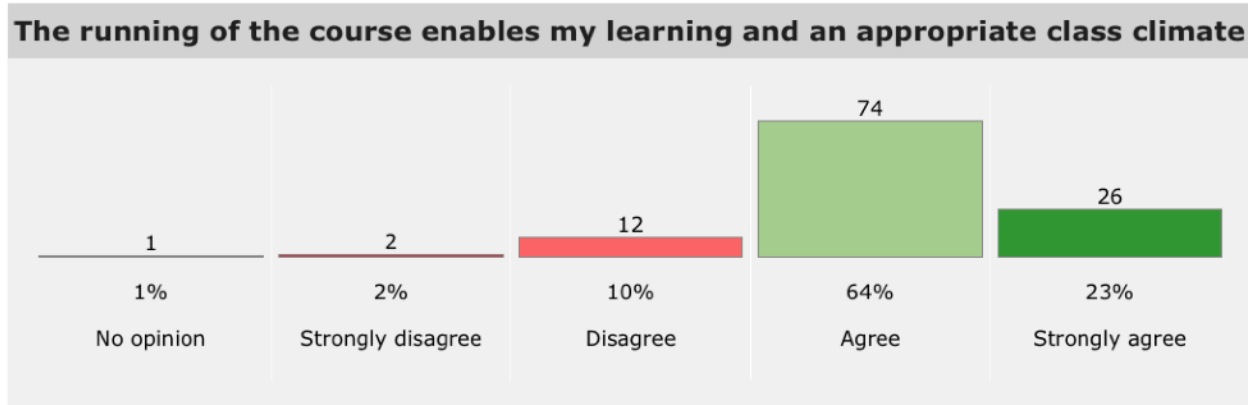
Matteo Dal Peraro

Institute of Bioengineering, School of Life Sciences

made by GPT4 - DALL-E

matteo.dalperaro@epfl.ch

115/182 - 63%



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

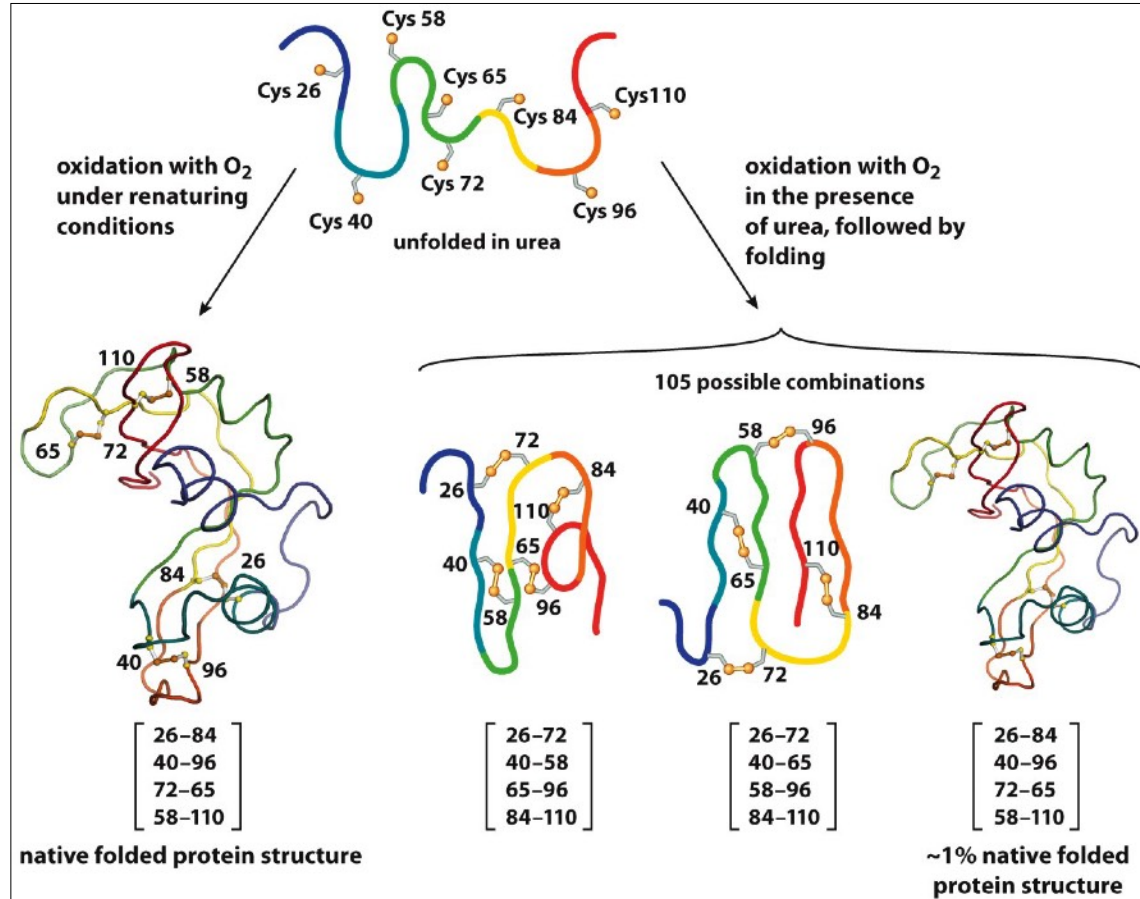
Sequence Determines Structure

The Anfinsen experiment (1954)

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

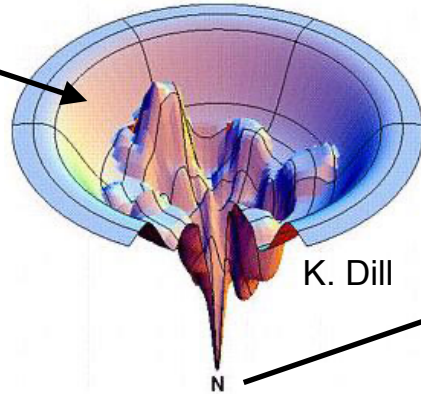
- Ribonuclease refolded and oxidized recovers 100% activity

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

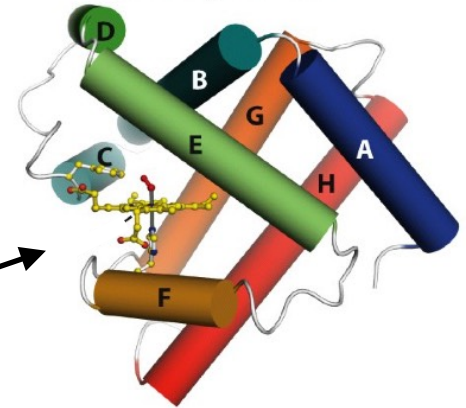


Protein sequence determines structure

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



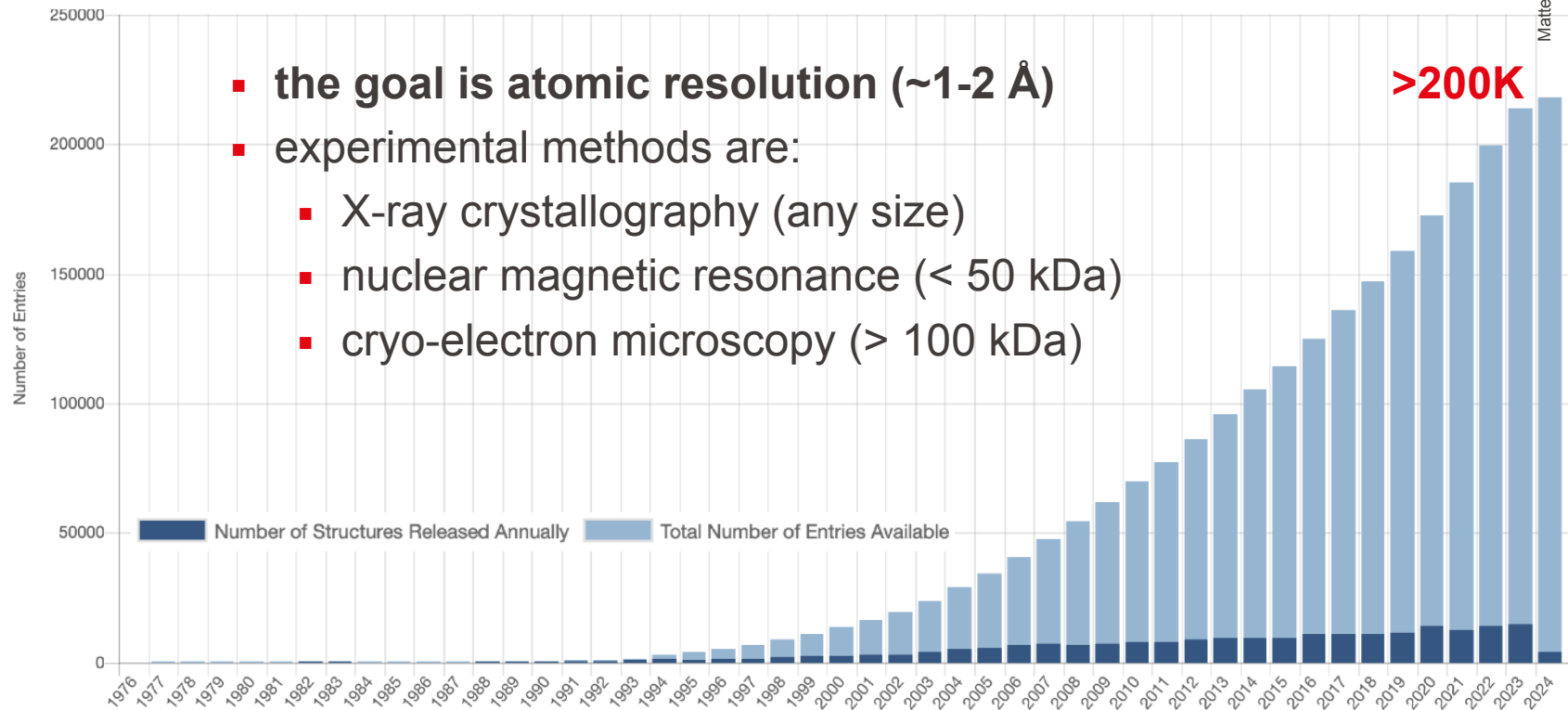
folding funnel



folded structure

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

Structure determination is key



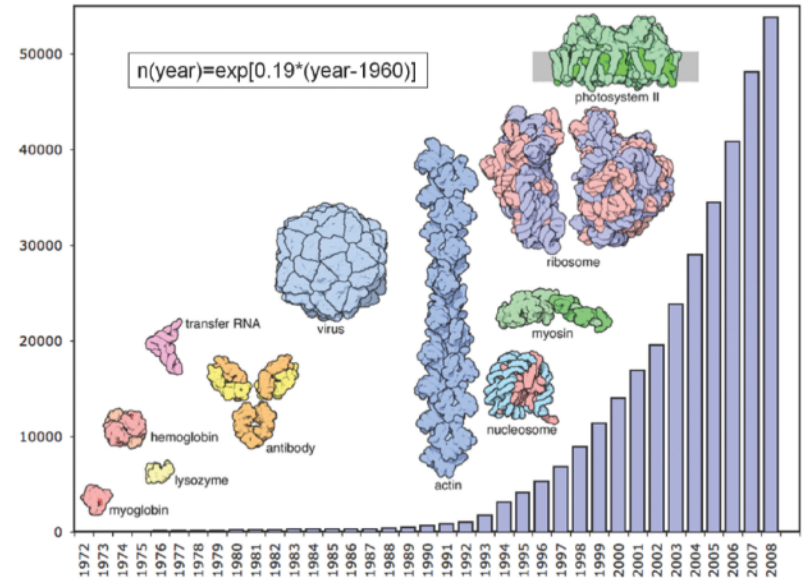
■ The Protein DataBank — PDB

<http://www.rcsb.org>

atomic resolution ($\sim 1\text{-}2\text{ \AA}$, $0.1\text{-}0.2\text{ nm}$ - 10^{-10} m)

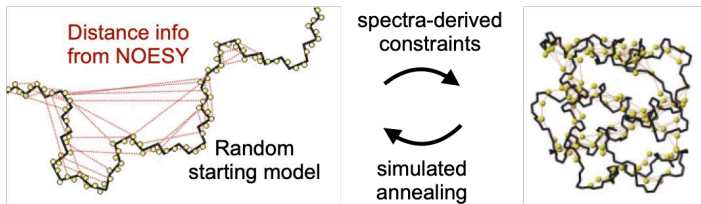
- **X-ray crystallography** ($\sim 1\text{-}2\text{ \AA}$; any size)
- **nuclear magnetic resonance** ($\sim 1\text{-}2\text{ \AA}$; $< 50\text{ kDa}$)
- **cryo-electron microscopy** ($1\text{-}20\text{ \AA}$; $> 100\text{ kDa}$)

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



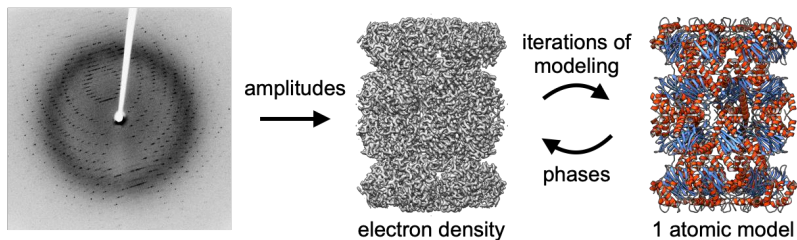
Methods for determining biomolecular structures

NMR



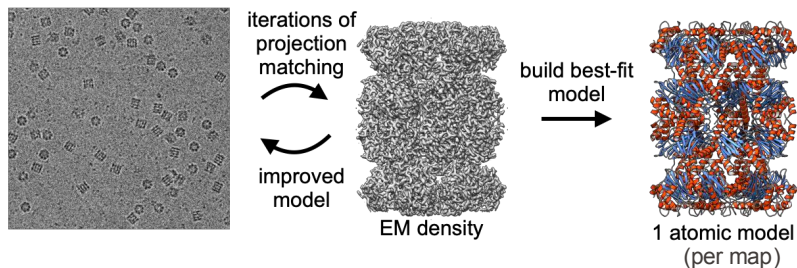
- **Versatile tool for studying protein structure and dynamics**
- Computationally light
- Full structural analysis limited to smaller proteins (<50kDa)
- Requires isotopic labeling
- Results in model ensemble

X-ray



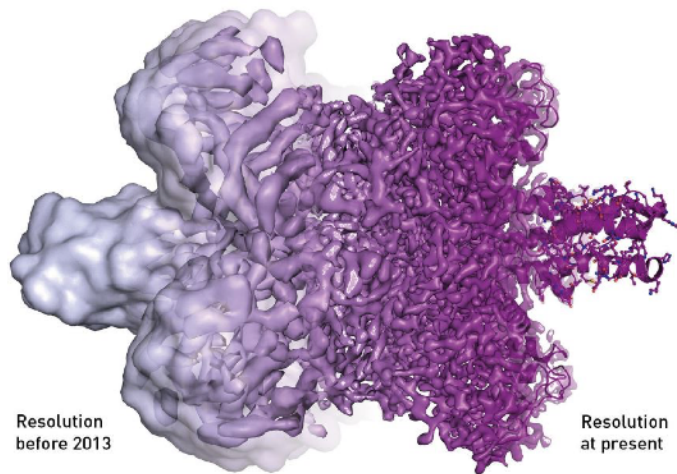
- Gold-standard method for solving protein structures
- **Not limited in size or achievable resolution**
- Computationally light
- Requires highly homogenous, crystallizable sample
- Requires screening of crystallization conditions
- Phase problem
- Results in a single model

CryoEM



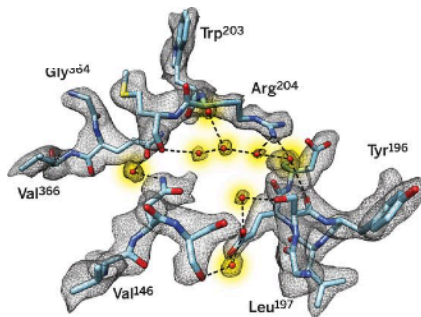
- Versatile tool for studying protein assembly, structure, dynamics
- Limited to proteins >40kDa
- No requirement for protein labeling
- **Does not require homogenous samples**
- Grid preparation procedure requires screening
- Real space imaging – no phase problem
- Can be used to study protein dynamics
- 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

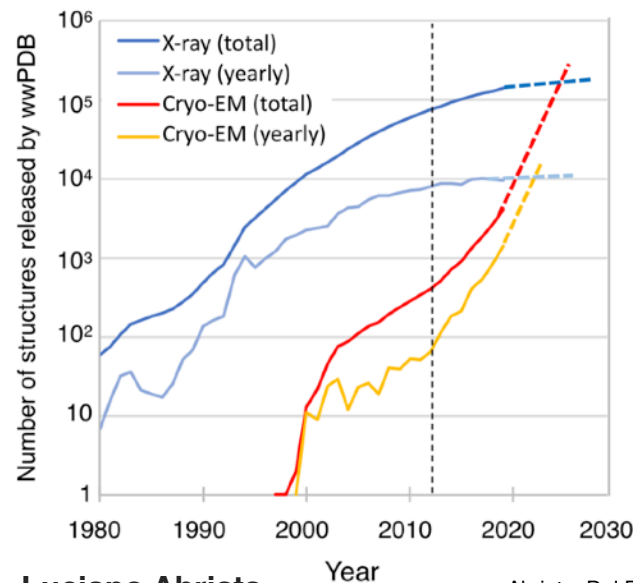


Krios G4 Cryo-TEM

resolution record 1.22 Å
size record ~50 kDa



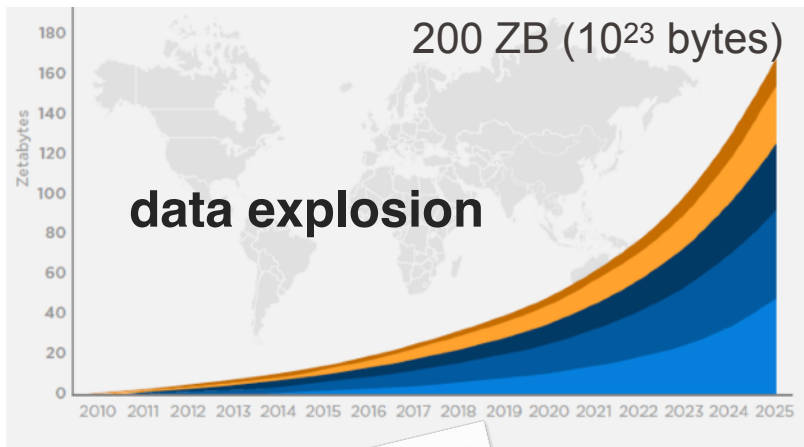
- progress in cryo-electron microscopy led to much improved resolution
- cryo-EM is becoming the new gold standard in structural biology



Luciano Abriata

Abriata, Dal Peraro, **JCIM** 2020

EPFL Data revolution — machine learning

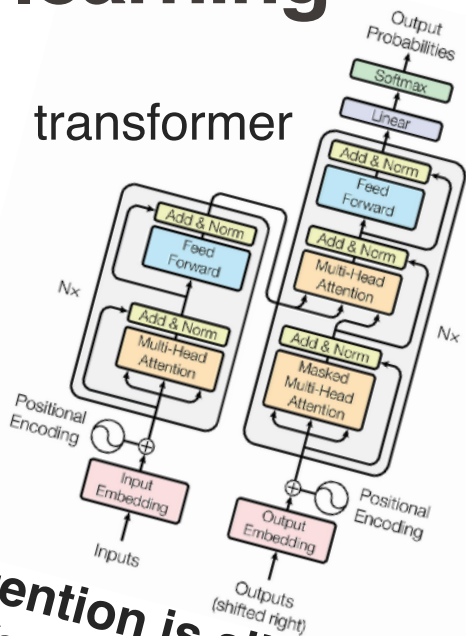


powerful hardware (GPU)



“Attention is all you need”
Vaswani et al. *arXiv* 2017

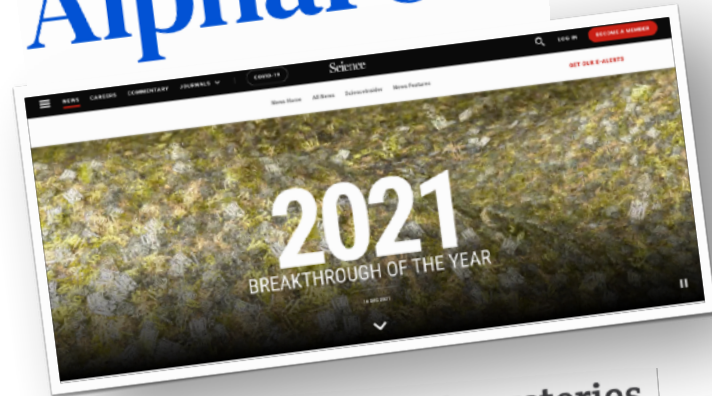
ChatGPT



Machine learning in biology



AlphaFold



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

By Helen Briggs
BBC science correspondent

**'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.

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

**'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)



Protein Structure Prediction Center



Menu

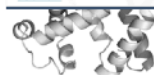
- Home
- PC Login
- PC Registration
- CASP Experiments

CASP15 (2022)

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

Initiatives

- Proceedings
- CASP Measures
- Feedback
- Assessors
- People
- Community Resources
- Job Fair



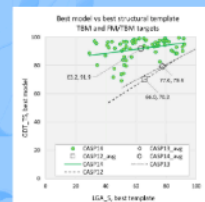
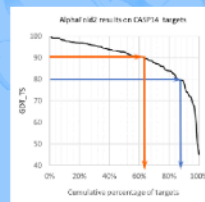
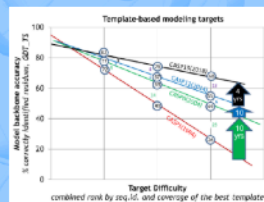
Success Stories From Recent CASPs

- assembly modeling
- template-based modeling
- de novo modeling
- contact prediction
- deep structural biology
- refinement
- data-assisted modeling
- ||

template-based modeling

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

CASP14 marked an extraordinary increase in the accuracy of the computed three-dimensional protein structures with the emergence of the advanced deep learning method AlphaFold2. Models built with this method proved to be competitive with the experimental accuracy (GDT_TS>90) for ~2/3 of the targets and of high accuracy (GDT_TS>80) for almost 90% of the targets (middle plot). The accuracy of CASP14 models for TBM targets significantly superseded accuracy of models that can be built by simple transcription of information from templates, and reached the level of GDT_TS~92 on average, which is significantly higher than the corresponding averages in previous two CASPs (right plot).



Welcome to the Protein Structure Prediction Center!

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

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

CASP1 (1994) | CASP2 (1996) | CASP3 (1998) | CASP4 (2000) | CASP5 (2002) | CASP6 (2004) | CASP7 (2006) | CASP8 (2008) | CASP9 (2010) | CASP10 (2012) | CASP11 (2014) | CASP12 (2016) | CASP13 (2018) | CASP14 (2020) | CASP15 (2022)

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

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

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

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

To proceed to the latest CASP experiment click on the logo below:

Message Board

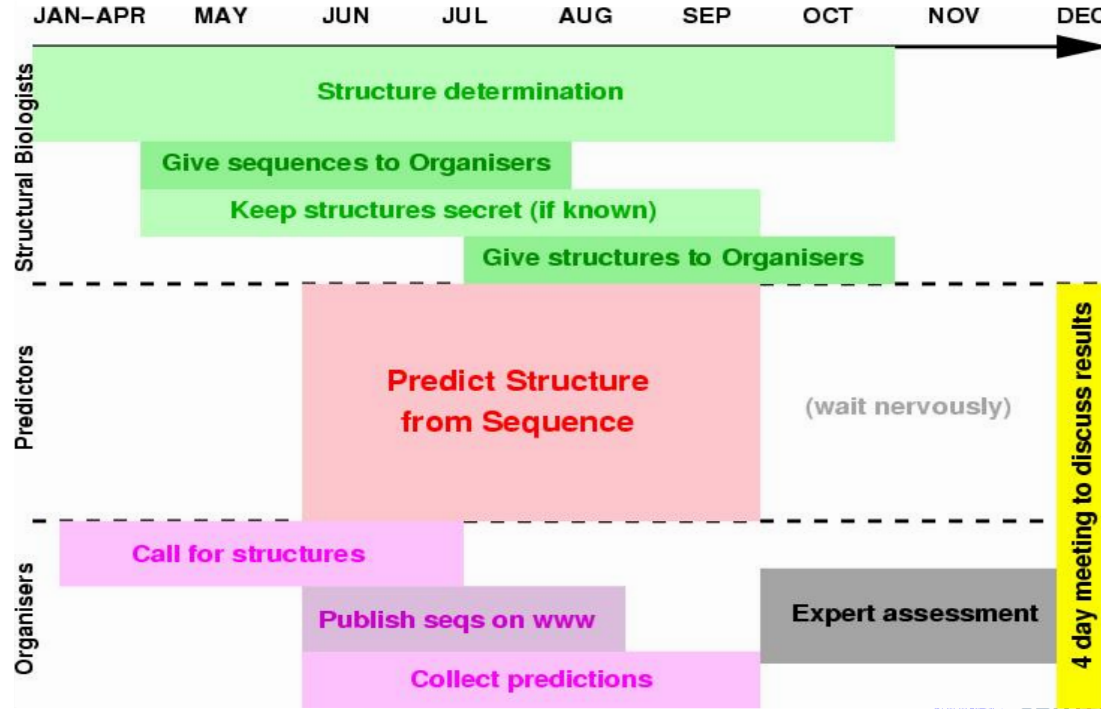
CASP Special Interest Group on Ensembles of Alternative Conformations
We have organized a CASP Special Interest Group on Ensembles of Alternative Conformations to provide discussions of methods for modeling alternative conformations of proteins and nucleic acids, and ...

Reporting: Postdoctoral position at the Protein Structure Prediction Center, UC Davis
We are still looking for a postdoc to help with CASP-related issues - see the posting below. --- A one-year position focused on large scale analysis of model accuracy with respect to applications ...

CASP Special Interest groups
One outcome of the CASP15 conference was a decision to create CASP special interest groups (SIGs) with the goal of increasing communication and discussion of new CASP-related developments. This message ...

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

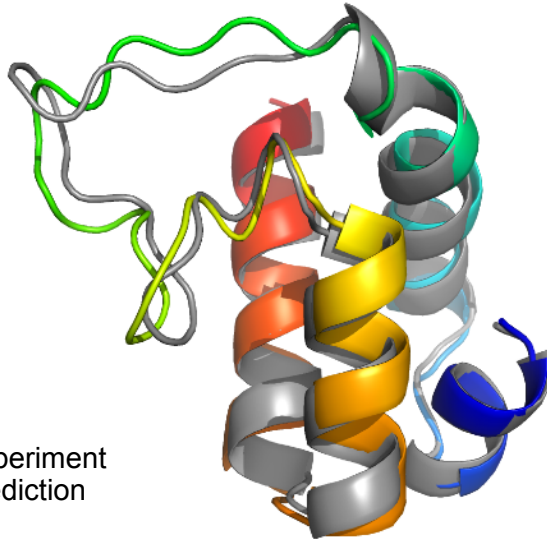
13



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

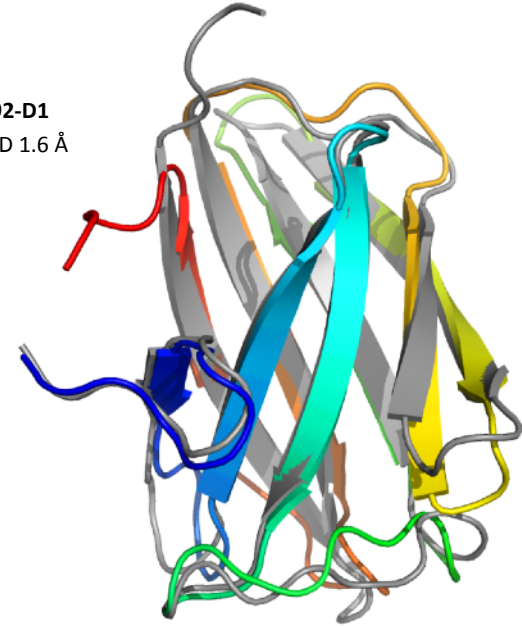
14

T0990-D1
RMSD 1.6 Å



gray = experiment
rainbow = prediction

T0992-D1
RMSD 1.6 Å



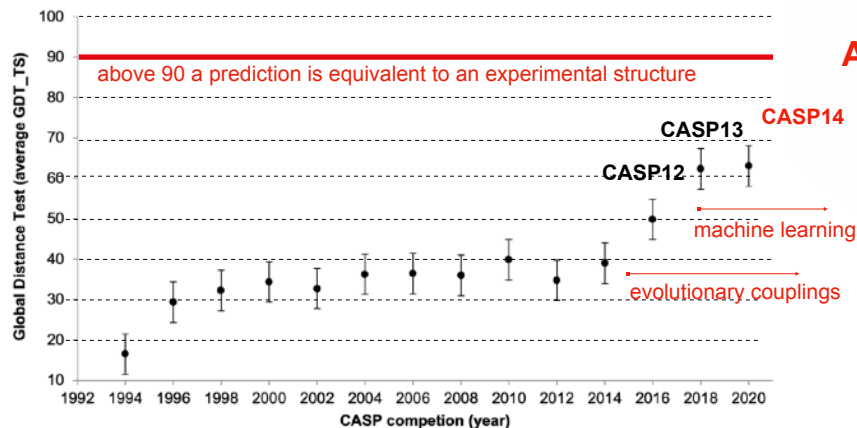
Root Mean Square Displacement :: RMSD defines a measure for similarity:

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

<http://predictioncenter.org>

AlphaFold2 solved a 70yo problem

Critical Assessment of protein Structure Prediction — CASP

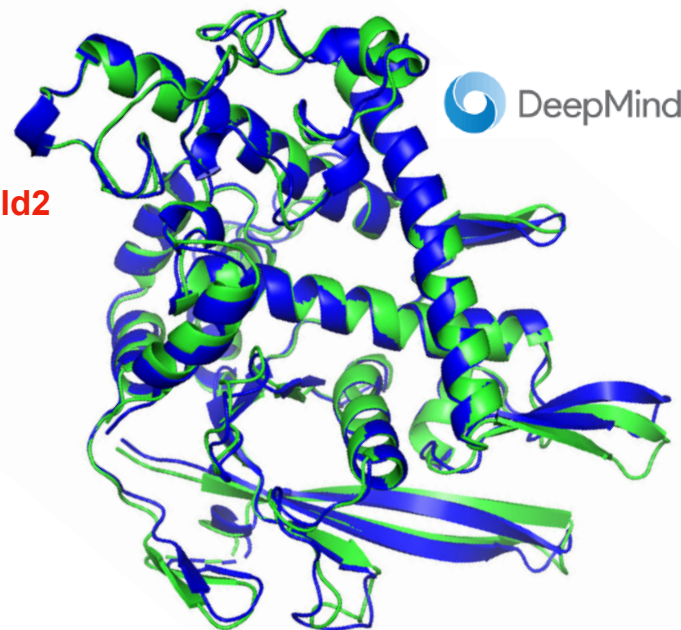


Abriata and Dal Peraro, **Proteins** 2019

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

Luciano Abriata

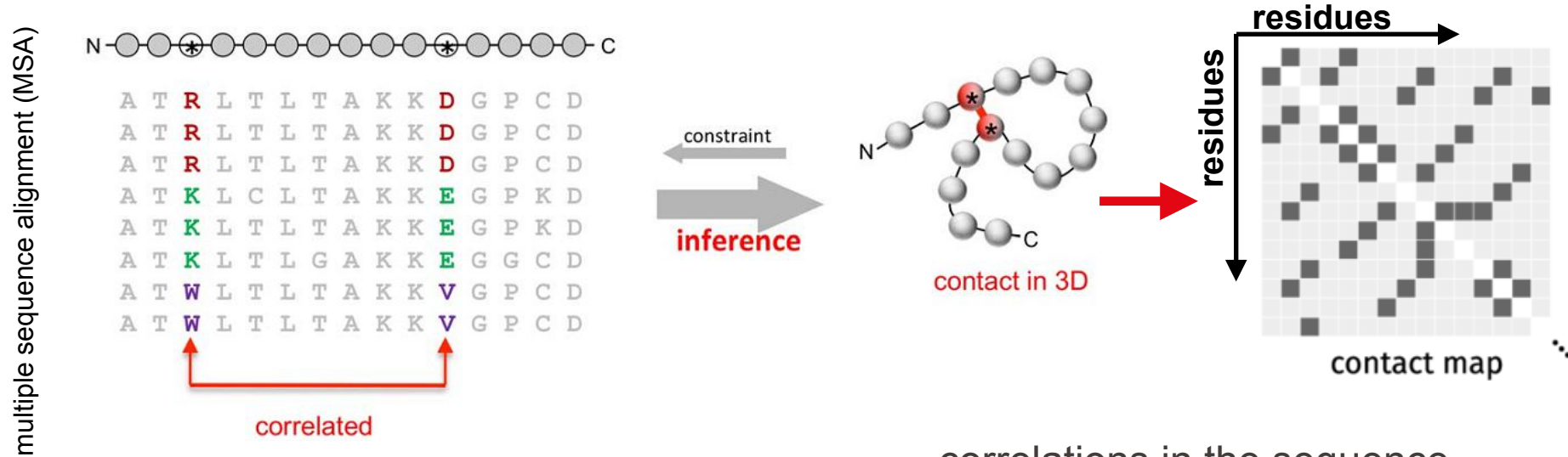
- experimental-like accuracy
- >200 M predicted models available in UniProt



■ experimental structure
■ computational prediction

EPFL Evolutionary couplings for protein prediction

16

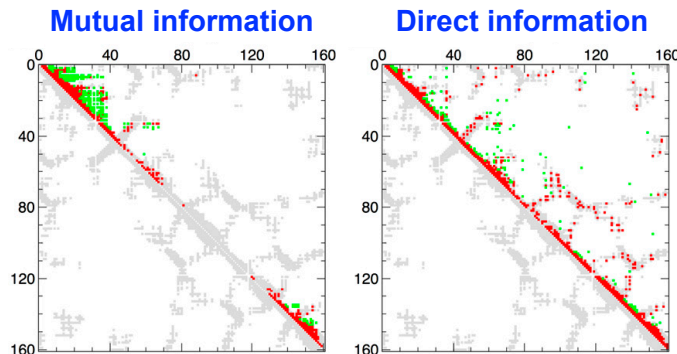
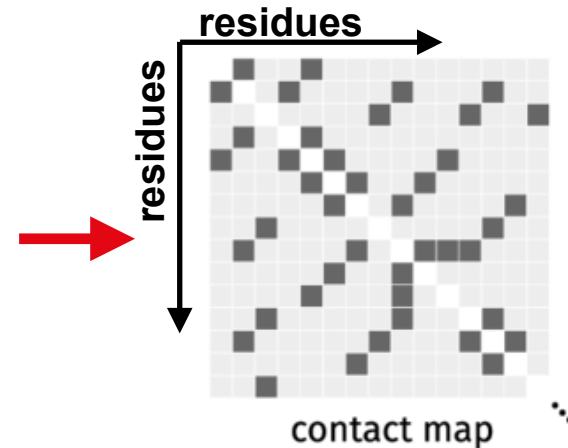
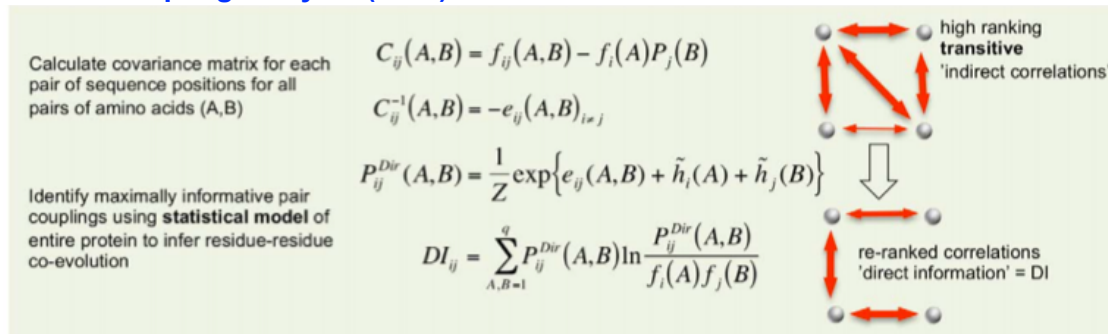


Marks, D. S.; Colwell, L. J.; Sheridan, R.; Hopf, T. A.; Pagnani, A.; Zecchina, R.; Sander, C. *PLoS One* **2011**, 6, e28766.

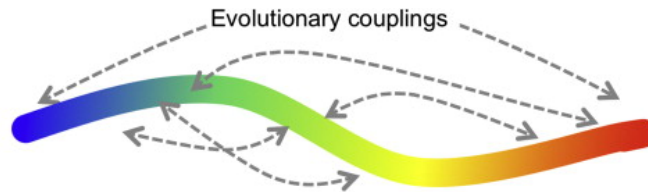
- correlations in the sequence space give structural information
- if you have enough predicted contacts you can fold a protein (similar to NMR)

Evolutionary couplings for protein prediction

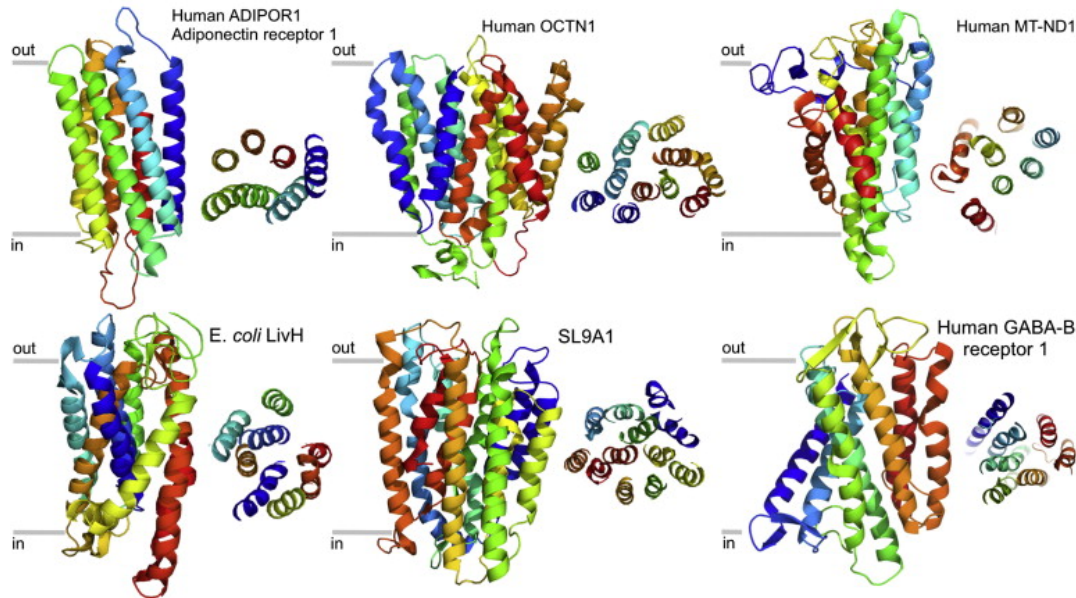
Direct-coupling analysis (DCA)



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



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

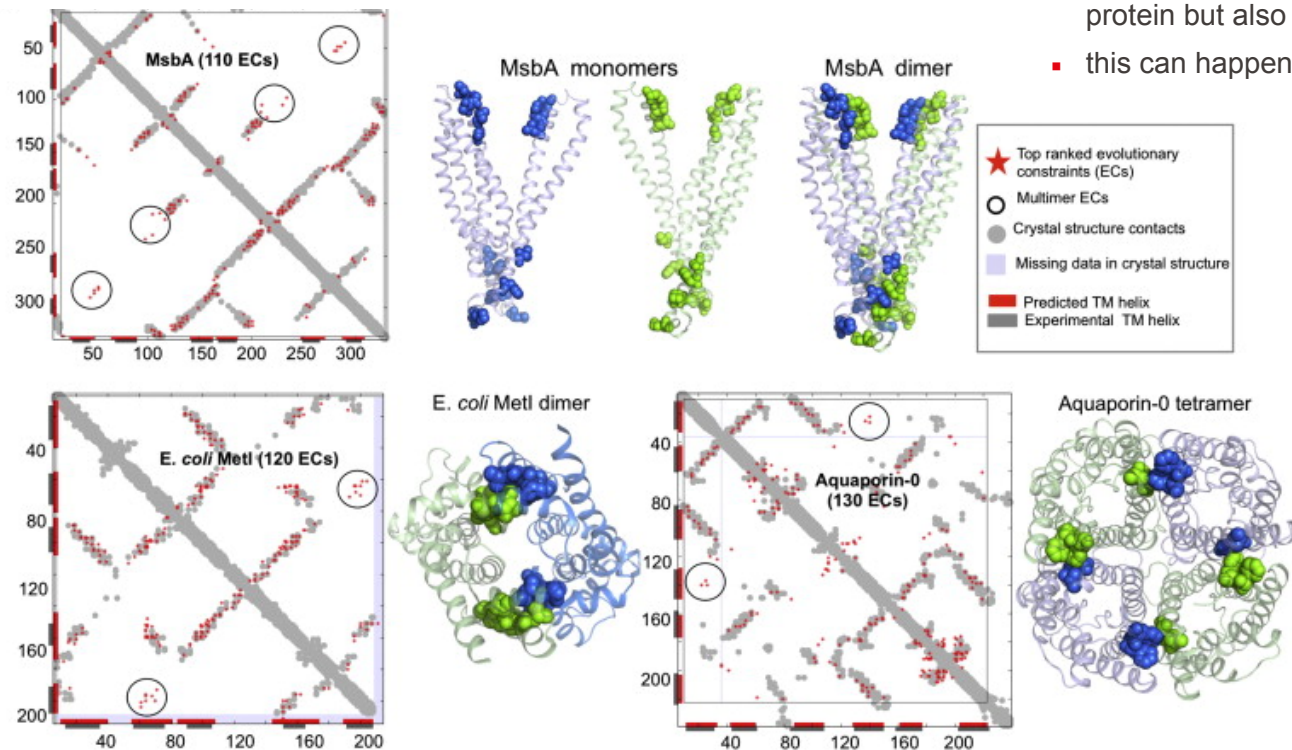


Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing

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

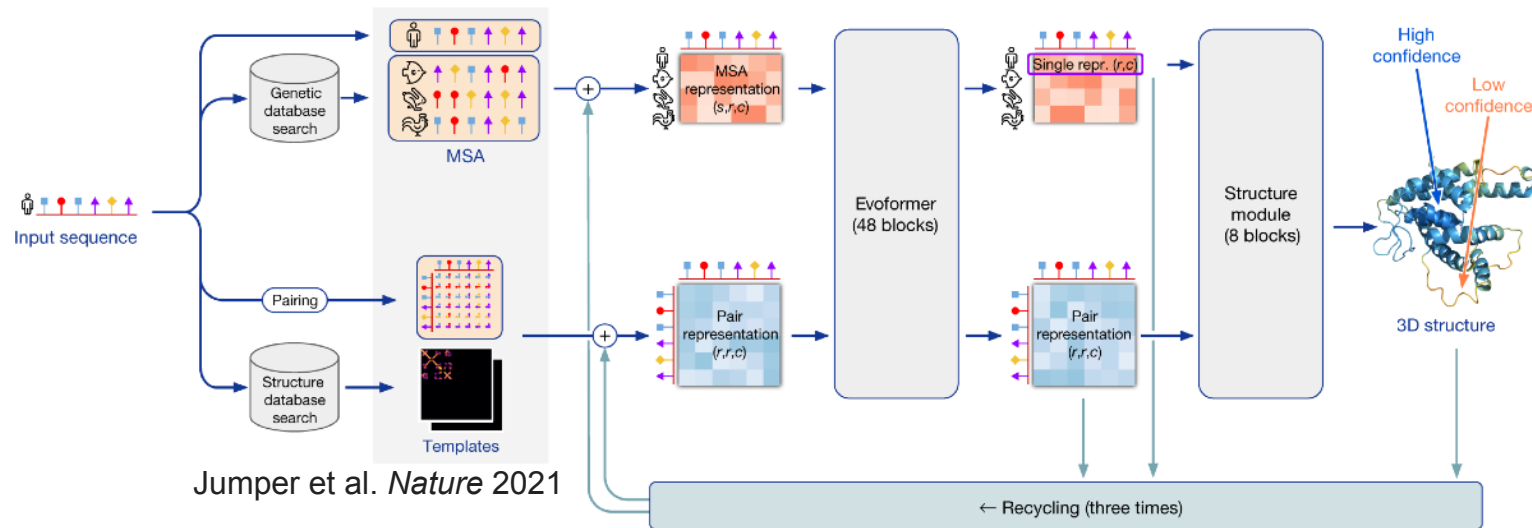


Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing

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>

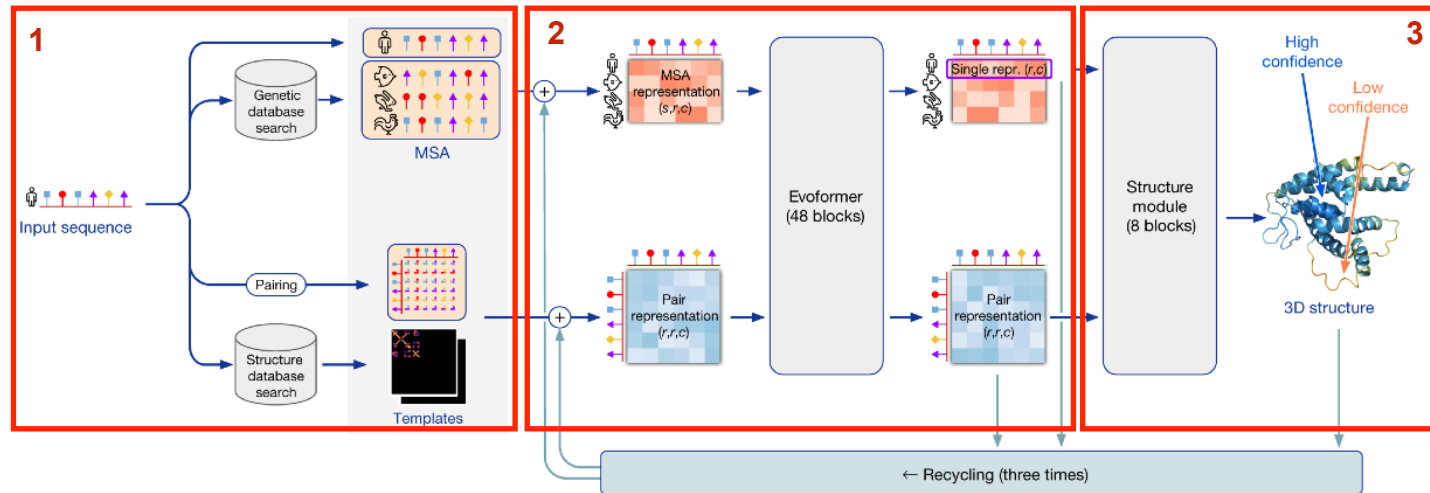
AlphaFold2 architecture

- “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”.



- 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:

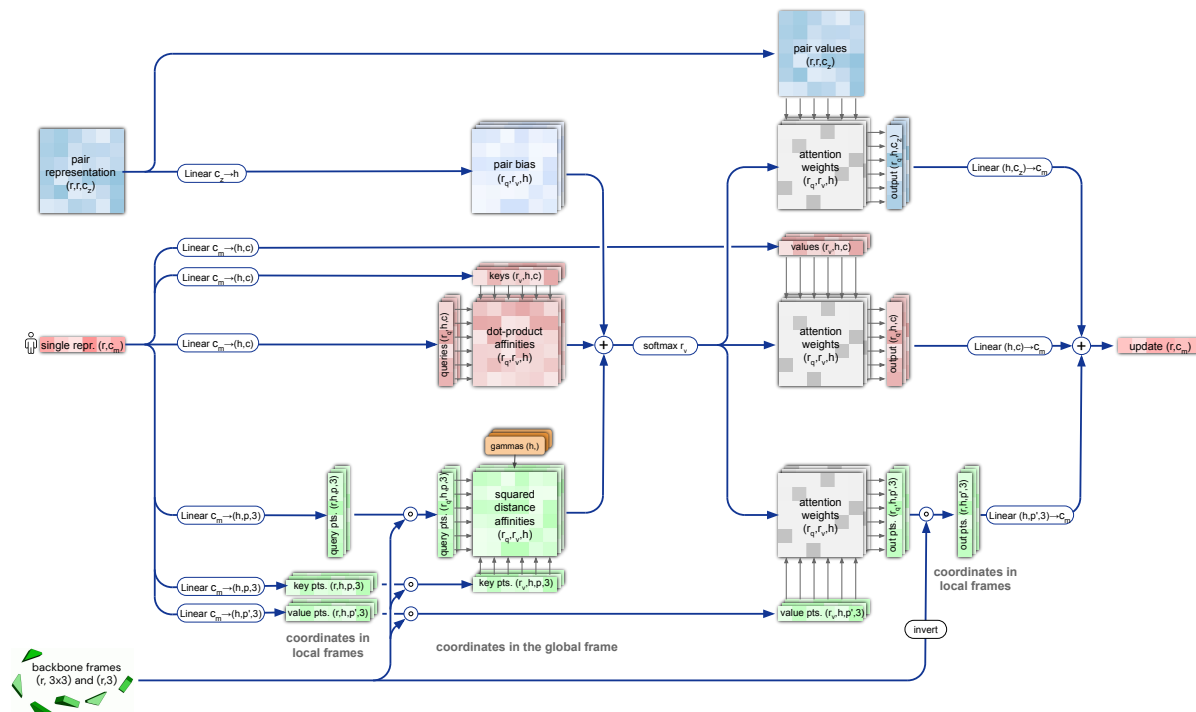
Evoformer transformer which refine the MSA and pair interactions

Third module:

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

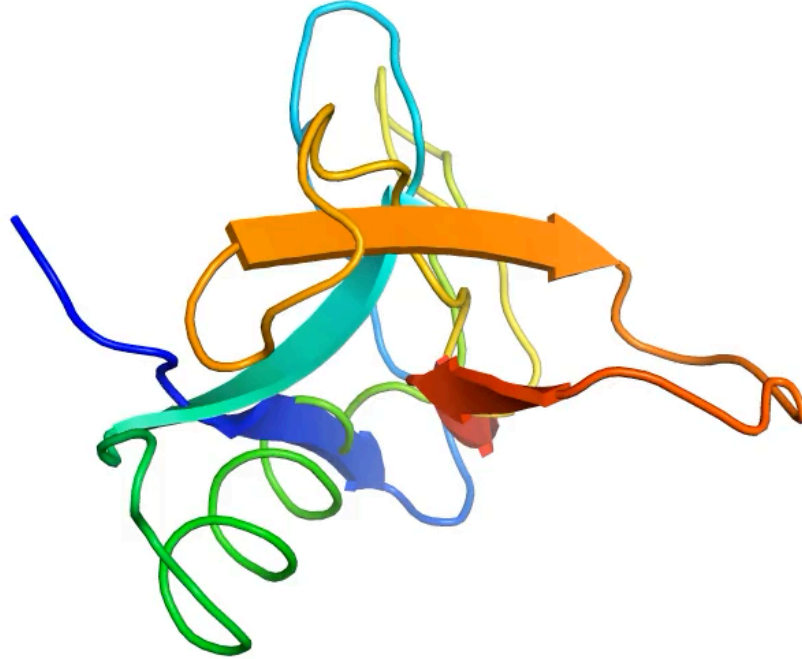
- end-to-end model produces prediction in one shot
- recycling (3X) to refine further prediction
- huge engineering effort

The power of (Google) engineering



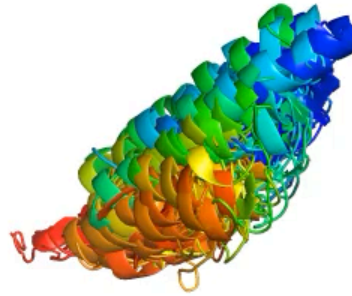
Supplementary Figure 8 | Invariant Point Attention Module. (**top, blue arrays**) modulation by the pair representation. (**middle, red arrays**) standard attention on abstract features. (**bottom, green arrays**) Invariant point attention. Dimensions: r: residues, c: channels, h: heads, p: points.

AlphaFold2 at work



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

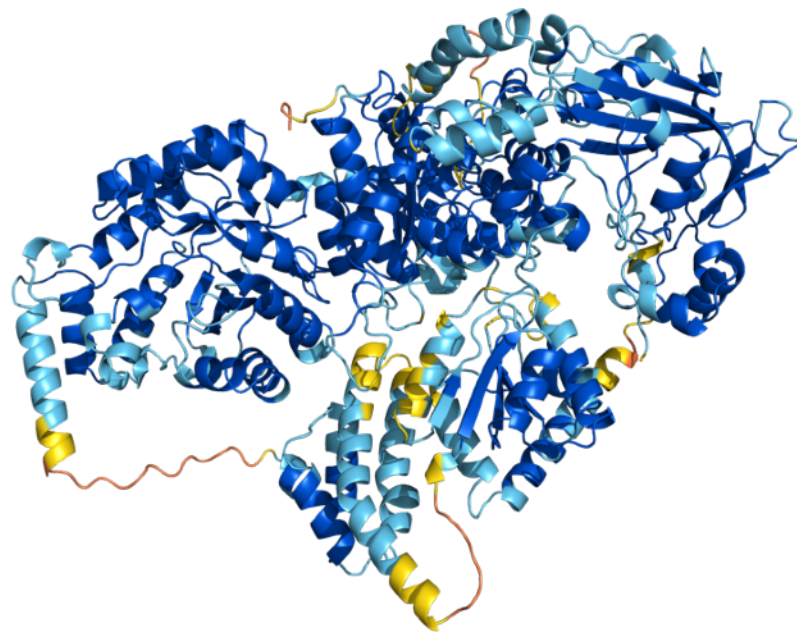
AlphaFold2 at work



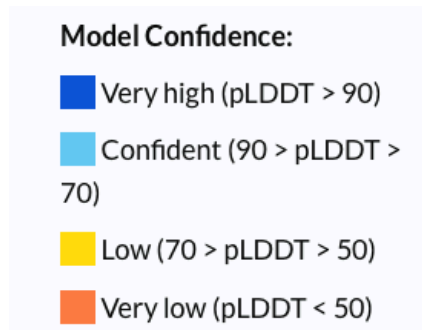
Recycling iteration 0, block 01
Secondary structure assigned from the final prediction

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

Search for protein, gene, UniProt accession or organism

BETA

Search

Examples: Free fatty acid receptor 2 At1g58602 Q5VSL9 E. coli

See search help

PAE: predicted alignment error

3D viewer

Sequence of AF-P35247-F1 Chain 1: Pulmonary su A

MLLFLLSALVLLTQPLGLYLEARMKTYSHRTNPSACTLVMSVESGLPGRDGRDREGPRGEK
 GDPLPGAAGQAGMPQAGPVGPKGDNQSVGEPPGPDPGSGPPGPDPVGPAGRGFLGKQ
 GNIGPQGPKEAGPKGEVGPAGMGQSGAGARGLAGPKGERGVGERGVPGNTGAAGSAGAM

Quick Styles: Default, Stylized, Illustrative

Components: AF-P35247-F1

Preset: + Add

Polymer: Cartoon, Eye icon, ...

Measurements: + Add

Export Animation

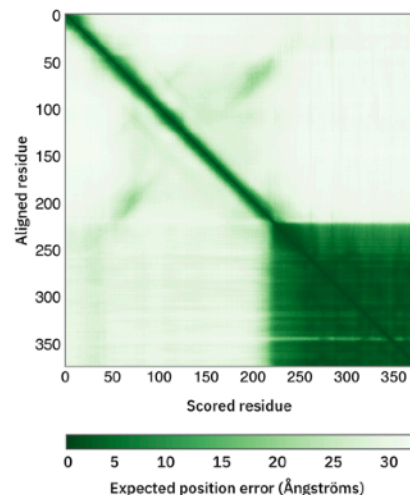
Export Geometry

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.

Predicted aligned error (PAE)

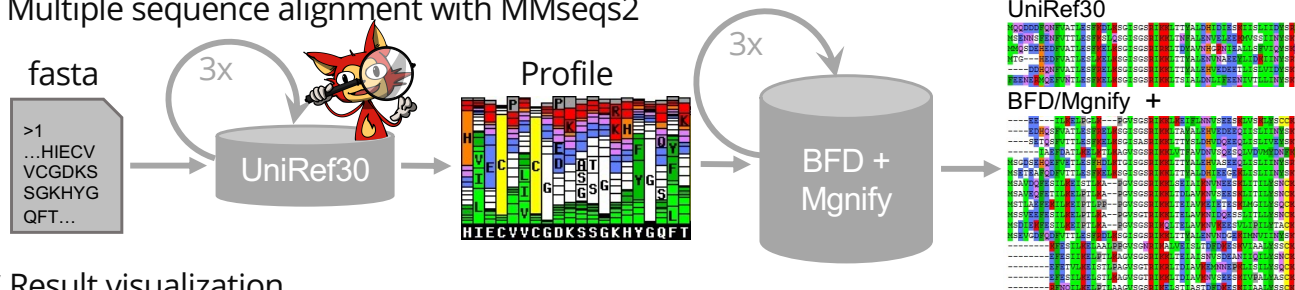


Click and drag a box on the PAE viewer to select regions of the structure and highlight them on the 3D viewer.

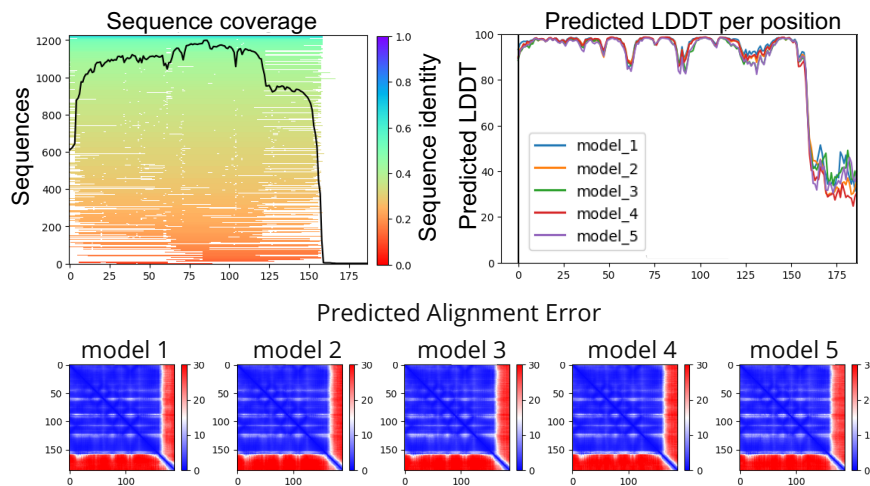
PAE data is useful for assessing inter-domain accuracy – [go to Help section below](#) for more information.

- 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

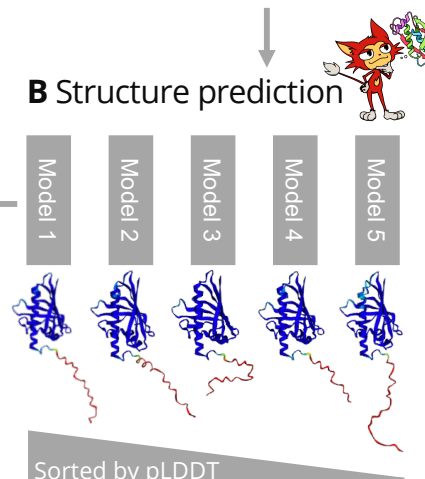
A Multiple sequence alignment with MMseqs2



C Result visualization

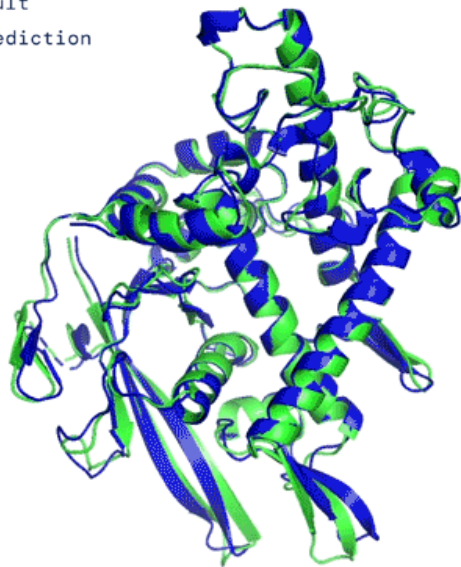


B Structure prediction

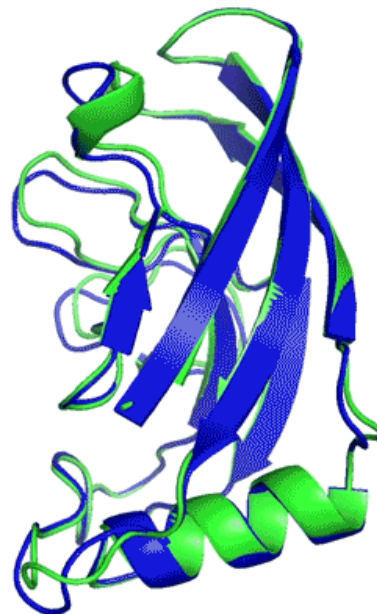


AlphaFold2 is a new SB methods

- Experimental result
- Computational prediction



T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)



T1049 / 6y4f
93.3 GDT
(adhesin tip)

~200K (PDB)



>200M (UniProt)

EPFL AlphaFold2 is not the ultimate oracle

- **Limitations**

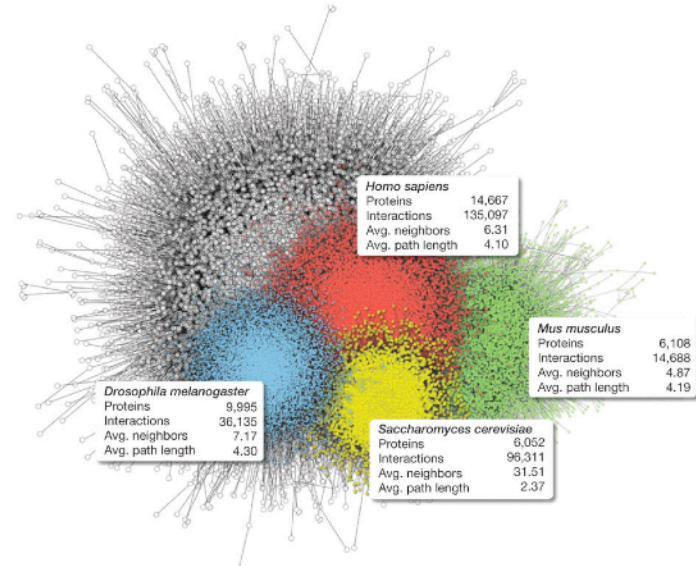
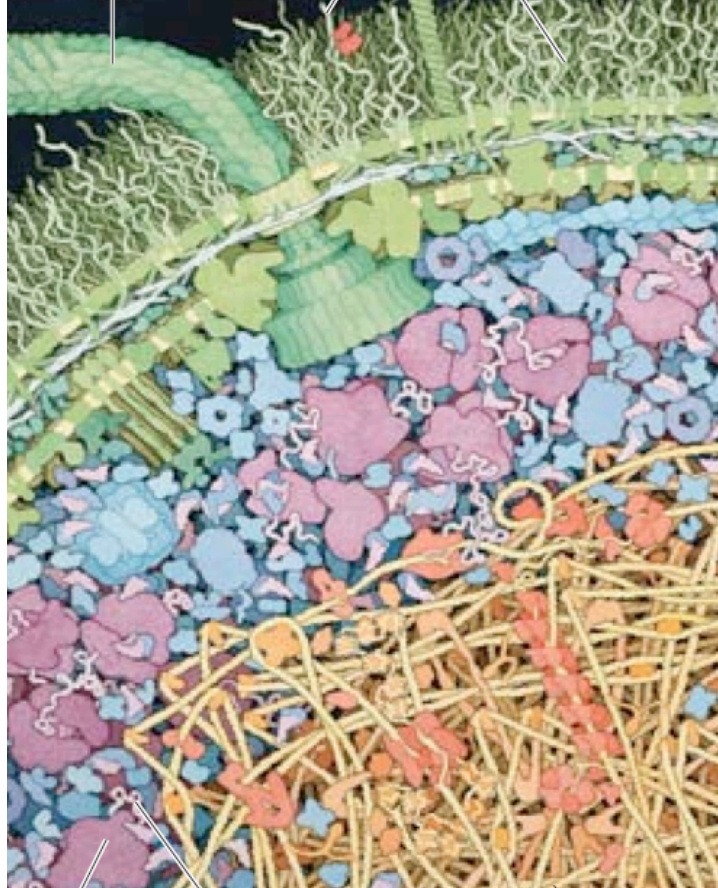
- need for a deep MSA (>30 sequences) to create accurate models
- not all the models are highly accurate as an experimental structure
- it does not account for dynamics and multiple states
- does not account for the post-translational modifications

- **Other potential benefits**

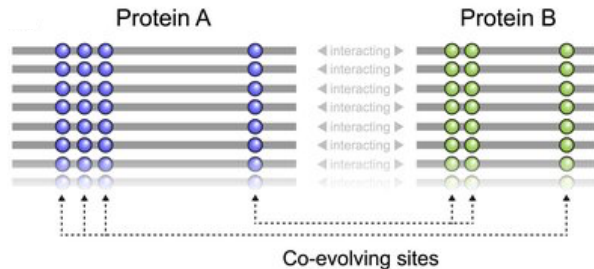
- AF triggered many other developments
- can assist experimental structure determination
- eg, in molecular replacement in X-ray crystallography
- eg, in cryoEM fitting and model building
- it is a means to look at protein-protein networks

Proteins form assemblies and networks

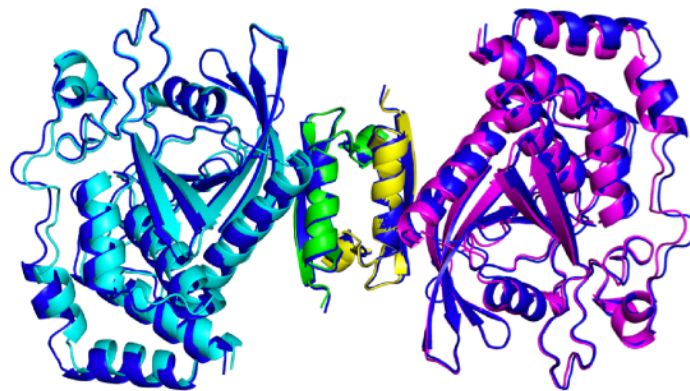
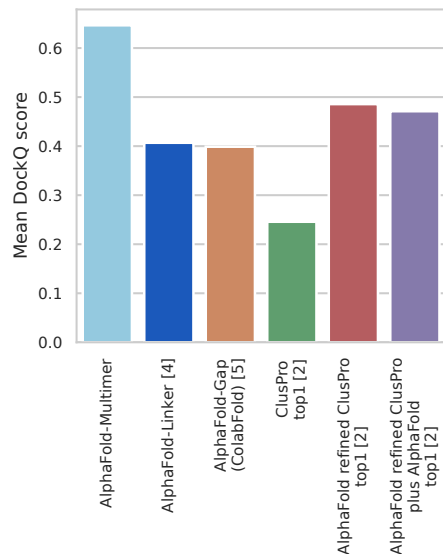
- Proteins can then arrange into supramolecular assemblies
- Interacting with nucleic acid, metabolites, membranes, etc.
- They create large network of interactions



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>

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. Dunham¹, Pascal Albanese^{4,5}, Andrew Keller⁶, Richard A. Scheltema^{4,5}, James E. Bruce⁶, Alexander Leitner⁷, Petras Kundrotas^{2,3,8}, Pedro Beltrao^{1,7} & Arne Elofsson^{2,3}✉

Article

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

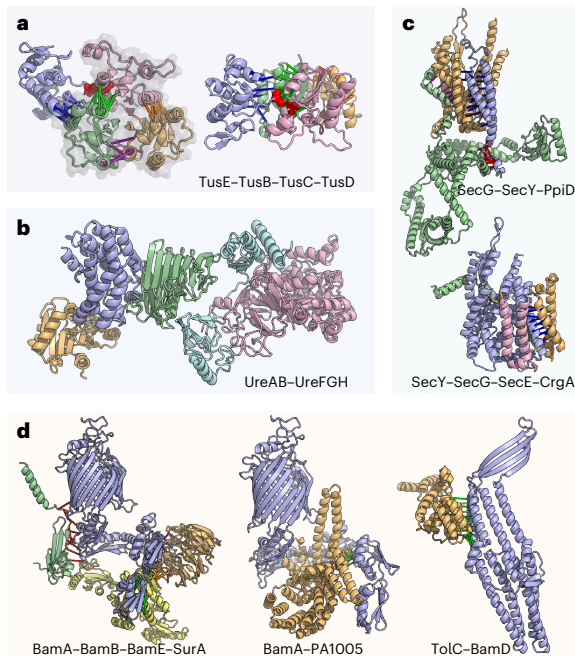
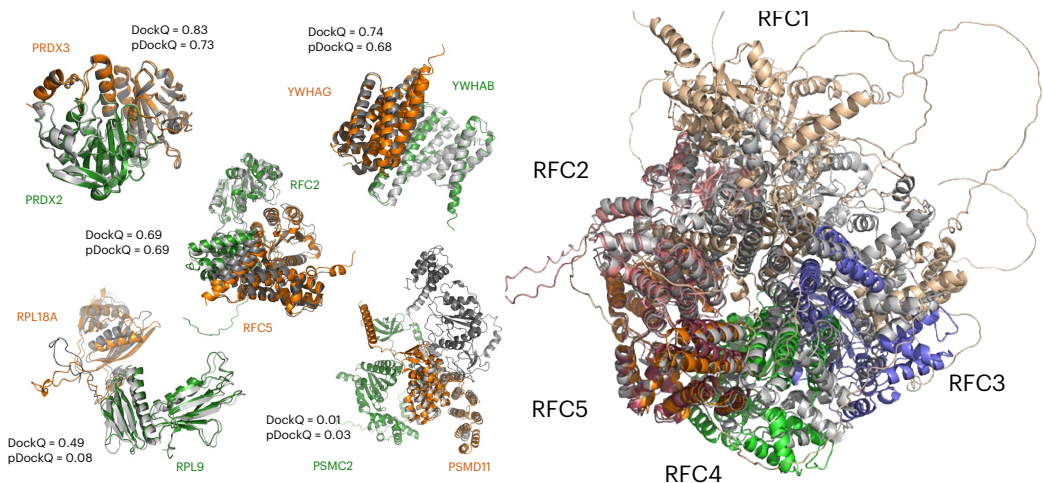
Protein interactions in human pathogens revealed through deep learning

Received: 28 April 2023

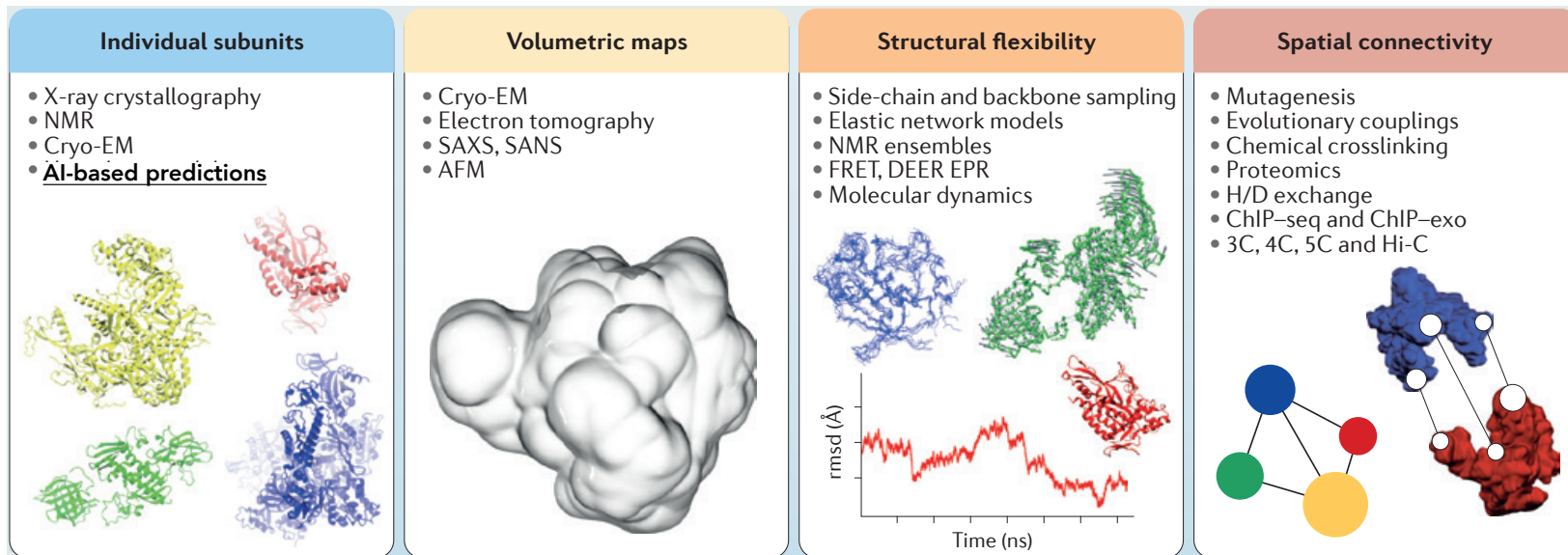
Accepted: 23 July 2024

Published online: 18 September 2024

Ian R. Humphreys^{1,2,13}, Jing Zhang^{2,4,5,13}, Minkyung Baek^{6,13}✉, Yaxi Wang^{7,13}, Aditya Krishnakumar^{1,2}, Jimin Pei^{3,4,5}, Ivan Anishchenko^{1,2}, Catherine A. Tower⁷, Blake A. Jackson⁷, Thulasi Warriar^{8,9,10}, Deborah T. Hung^{8,9,10}, S. Brook Peterson⁷, Joseph D. Mougous^{7,11,12}✉, Qian Cong^{3,4,5}✉ & David Baker^{1,2,11}✉



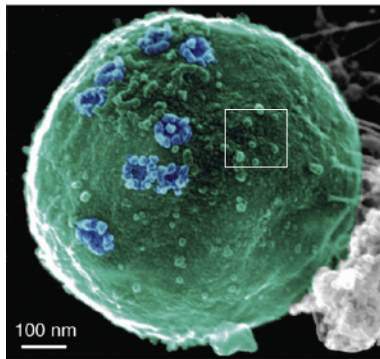
Integrative structural biology



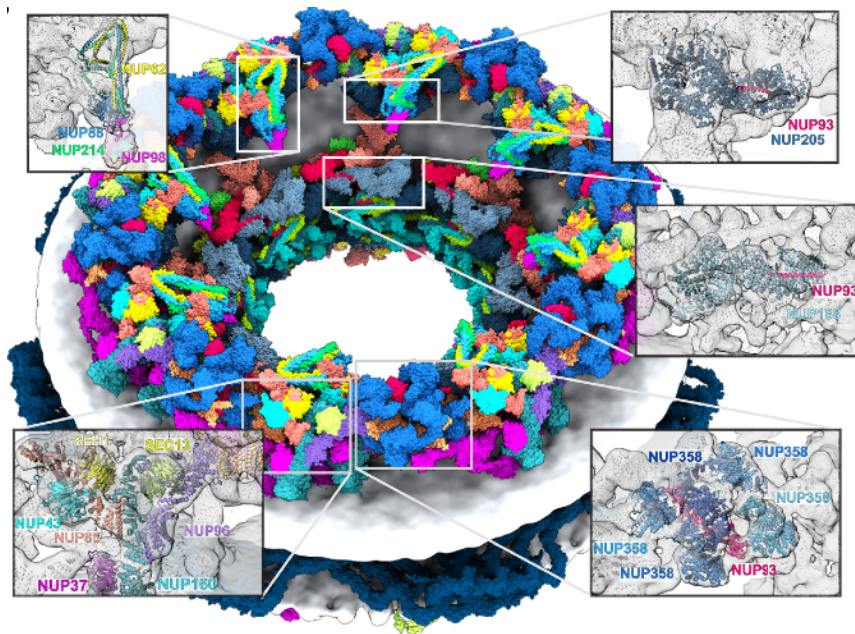
**data integration and
model building**

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

AlphaFold Server

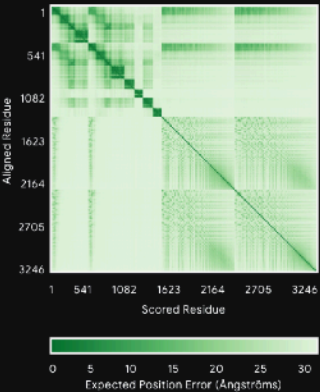
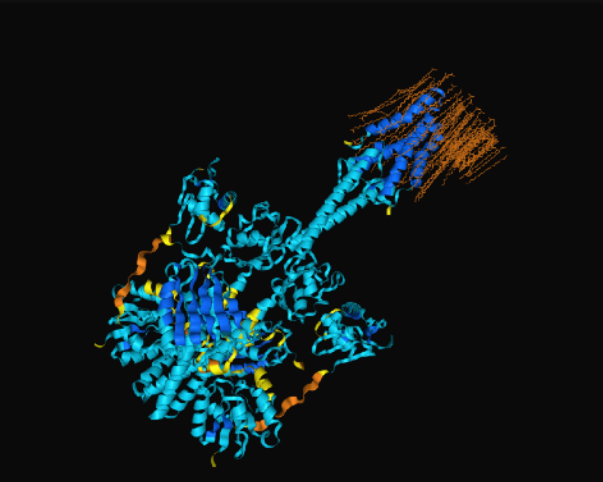
Powered by AlphaFold 3

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



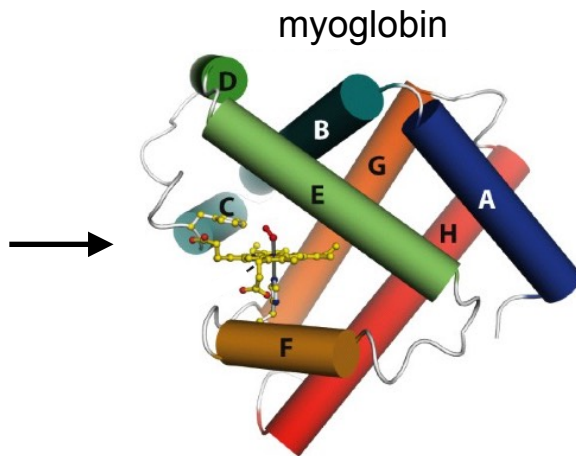
Type	Copies	Sequence
Protein	2	102030405060708090100110120130140150160170180190200210220230240250260270280290300310320330340350360370380390400410420430440450460470480490500510520530540550560570580590600610620630640650660670680690700710720730740750760770780790800810820830840850860870880890900910920930940950960970980991001011021031041051061071081091101111121131141151161171181191201211221231241251261271281291301311321331341351361371381391401411421431441451461471481491501511521531541551561571581591601611621631641651661671681691701711721731741751761771781791801811821831841851861871881891901911921931941951961971981992002012022032042052062072082092102112122132142152162172182192202212222232242252262272282292302312322332342352362372382392402412422432442452462472482492502512522532542552562572582592602612622632642652662672682692702712722732742752762772782792802812822832842852862872882892902912922932942952962972982993003013023033043053063073083093103113123133143153163173183193203213223233243253263273283293303313323333343353363373383393403413423433443453463473483493503513523533543553563573583593603613623633643653663673683693703713723733743753763773783793803813823833843853863873883893903913923933943953963973983994004014024034044054064074084094104114124134144154164174184194204214224234244254264274284294304314324334344354364374384394404414424434444454464474484494504514524534544554564574584594604614624634644654664674684694704714724734744754764774784794804814824834844854864874884894904914924934944954964974984995005015025035045055065075085095105115125135145155165175185195205215225235245255265275285295305315325335345355365375385395405415425435445455465475485495505515525535545555565575585595605615625635645655665675685695705715725735745755765775785795805815825835845855865875885895905915925935945955965975985996006016026036046056066076086096106116126136146156166176186196206216226236246256266276286296306316326336346356366376386396406416426436446456466476486496506516526536546556566576586596606616626636646656666676686696706716726736746756766776786796806816826836846856866876886896906916926936946956966976986997007017027037047057067077087097107117127137147157167177187197207217227237247257267277287297307317327337347357367377387397407417427437447457467477487497507517527537547557567577587597607617627637647657667677687697707717727737747757767777787797807817827837847857867877887897907917927937947957967977987998008018028038048058068078088098108118128138148158168178188198208218228238248258268278288298308318328338348358368378388398408418428438448458468478488498508518528538548558568578588598608618628638648658668678688698708718728738748758768778788798808818828838848858868878888898908918928938948958968978988999009019029039049059069079089099109119129139149159169179189199209219229239249259269279289299309319329339349359369379389399409419429439449459469479489499509519529539549559569579589599609619629639649659669679689699709719729739749759769779789799809819829839849859869879889899909919929939949959969979989991000100110021003100410051006100710081009101010111012101310141015101610171018101910201021102210231024102510261027102810291030103110321033103410351036103710381039104010411042104310441045104610471048104910501051105210531054105510561057105810591060106110621063106410651066106710681069107010711072107310741075107610771078107910801081108210831084108510861087108810891090109110921093109410951096109710981099110011011102110311041105110611071108110911101111111211131114111511161117111811191120112111221123112411251126112711281129113011311132113311341135113611371138113911401141114211431144114511461147114811491150115111521153115411551156115711581159116011611162116311641165116611671168116911701171117211731174117511761177117811791180118111821183118411851186118711881189119011911192119311941195119611971198119912001201120212031204120512061207120812091210121112121213121412151216121712181219122012211222122312241225122612271228122912301231123212331234123512361237123812391240124112421243124412451246124712481249125012511252125312541255125612571258125912601261126212631264126512661267126812691270127112721273127412751276127712781279128012811282128312841285128612871288128912901291129212931294129512961297129812991300130113021303130413051306130713081309131013111312131313141315131613171318131913201321132213231324132513261327132813291330133113321333133413351336133713381339134013411342134313441345134613471348134913501351135213531354135513561357135813591360136113621363136413651366136713681369137013711372137313741375137613771378137913801381138213831384138513861387138813891390139113921393139413951396139713981399140014011402140314041405140614071408140914101411141214131414141514161417141814191420142114221423142414251426142714281429143014311432143314341435143614371438143914401441144214431444144514461447144814491450145114521453145414551456145714581459146014611462146314641465146614671468146914701471147214731474147514761477147814791480148114821483148414851486148714881489149014911492149314941495149614971498149915001501150215031504150515061507150815091510151115121513151415151516151715181519152015211522152315241525152615271528152915301531153215331534153515361537153815391540154115421543154415451546154715481549155015511552155315541555155615571558155915601561156215631564156515661567156815691570157115721573157415751576157715781579158015811582158315841585158615871588158915901591159215931594159515961597159815991600160116021603160416051606160716081609161016111612161316141615161616171618161916201621162216231624162516261627162816291630163116321633163416351636163716381639164016411642164316441645164616471648164916501651165216531654165516561657165816591660166116621663166416651666166716681669167016711672167316741675167616771678167916801681168216831684168516861687168816891690169116921693169416951696169716981699170017011702170317041705170617071708170917101711171217131714171517161717171817191720172117221723172417251726172717281729173017311732173317341735173617371738173917401741174217431744174517461747174817491750175117521753175417551756175717581759176017611762176317641765176617671768176917701771177217731774177517761777177817791780178117821783178417851786178717881789179017911792179317941795179617971798179918001801180218031804180518061807180818091810181118121813181418151816181718181819182018211822182318241825182618271828182918301831183218331834183518361837183818391840184118421843184418451846184718481849185018511852185318541855185618571858185918601861186218631864186518661867186818691870187118721873187418751876187718781879188018811882188318841885188618871888188918901891189218931894189518961897189818991900190119021903190419051906190719081909191019111912191319141915191619171918191919201921192219231924192519261927192819291930193119321933193419351936193719381939194019411942194319441945194619471948194919501951195219531954195519561957195819591960196119621963196419651966196719681969197019711972197319741975197619771978197919801981198219831984198519861987198819891990199119921993199419951996199719981999200020012002200320042005200620072008200920102011201220132014201520162017201820192020202120222023202420252026202720282029203020312032203320342035203620372038203920402041204220432044204520462047204820492050205120522053205420552056205720582059206020612062206320642065206620672068206920702071207220732074207520762077207820792080208120822083208420852086208720882089209020912092209320942095209620972098209921002101210221032104210521062107210821092110211121122113211421152116211721182119212021212122212321242125212621272128212921302131213221332134213521362137213821392140214121422143214421452146214721482149215021512152215321542155215621572158215921602161216221632164216521662167216821692170217121722173217421752176217721782179218021812182218321842185218621872188218921902191219221932194219521962197219821992200220122022203220422052206220722082209221022112212221322142215221622172218221922202221222222232224222522262227222822292230223122322233223422352236223722382239224022412242224322442245224622472248224922502251225222532254225522562257225822592260226122622263226422652266226722682269227022712272227322742275227622772278227922802281228222832284228522862287228822892290229122922293229422952296229722982299230023012302230323042305230623072308230923102311231223132314231523162317231823192320232123222323232423252326232723282329233023312332233323342335233623372338233923402341234223432344234523462347234823492350235123522353235423552356235723582359236023612362236323642365236623672368236923702371237223732374237523762377237823792380238123822383238423852386238723882389239023912392239323942395239623972398239924002401240224032404240524062407240824092410241124122413241424152416241724182419242024212422242324242425242624272428242924302431243224332434243524362437243824392440244124422443244424452446244724482449245024512452245324542455245624572458245924602461246224632464246524662467246824692470247124722473247424752476247724782479248024812482248324842485248624872488248924902491249224932494249524962497249824992500250125022503250425052506250725082509251025112512251325142515251625172518251925202521252225232524252525262527252825292530253125322533253425352536253725382539254025412542254325442545254625472548254925502551255225532554255525562557255825592560256125622563256425652566256725682569257025712572257325742575257625772578257925802581258225832584258525862587258825892590259125922593259425952596259725982599260026012602260326042605260626072608260926102611261226132614261526162617261826192620262126222623262426252626262726282629263026312632263326342635263626372638263926402641264226432644264526462647264826492650265126522653265426552656265726582659266026612662266326642665266626672668266926702671267226732674267526762677267826792680268126822683268426852686268726882689269026912692269326942695269626972698269927002701270227032704270527062707270827092710271127122713271427152716271727182719272027212722272327242725272627272728272927302731273227332734273527362737273827392740274127422743274427452746274727482749275027512752275327542755275627572758275927602761276227632764276527662767276827692770277127722773277427752776277727782779278027812782278327842785278627872788278927902791279227932794279527962797279827992800280128022803280428052806280728082809281028112812281328142815281628172818281928202821282228232824282528262827282828292830283128322833283428352836283728382839284028412842284328442845284628472848284928502851285228532854285528562857285828592860286128622863286428652866286728682869287028712872287328742875287628772878287928802881288228832884288528862887288828892890289128922893289428952896289728982899290029012902290329042905290629072908290929102911291229132914291529162917291829192920292129222923292429252926292729282929293029312932293329342935293629372938293929402941294229432944294529462947294829492950295129522953295429552956295729582959296029612962296329642965296629672968296929702971297229732974297529762977297829792980298129822983298429852986298729882989299029912992299329942995299629972998299930003001300230033004300530063007300830093010301130123013301430153016301730183019302030213022302330243025302630273028302930303031303230333034303530363037303830393040304130423043304430453046304730483049305030513052305330543055305630573058305930603061306230633064306530663067306830693070307130723073307430753076307730783079308030813082308330843085308630873088308930903091309230933094309530963097309830993100310131023103310431053106310731083109311031113112311331143115311631173118311931203121312231233124312531263127312831293130313131323133

The folding paradigm

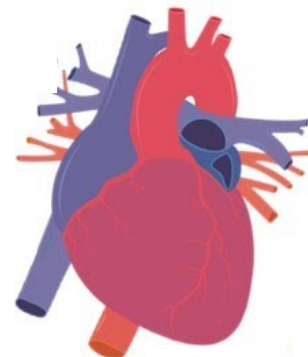
```

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

sequence



structure

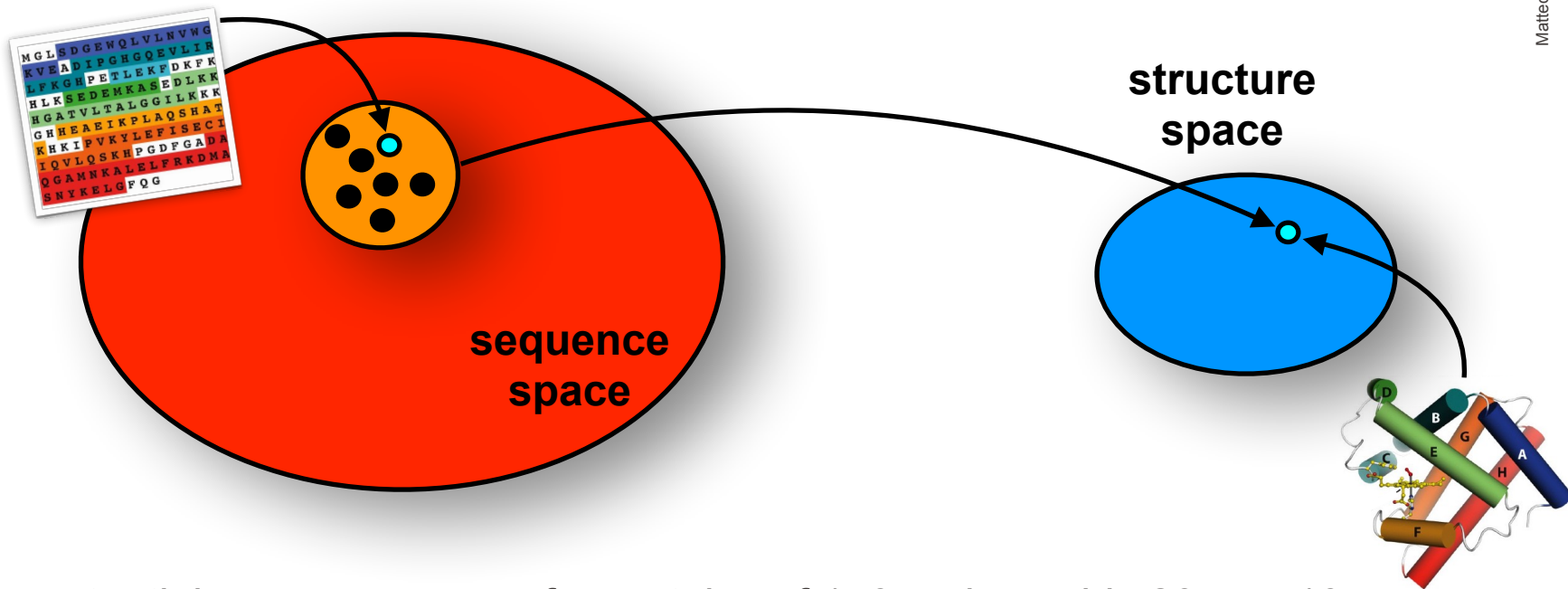


function

evolution (billion year)

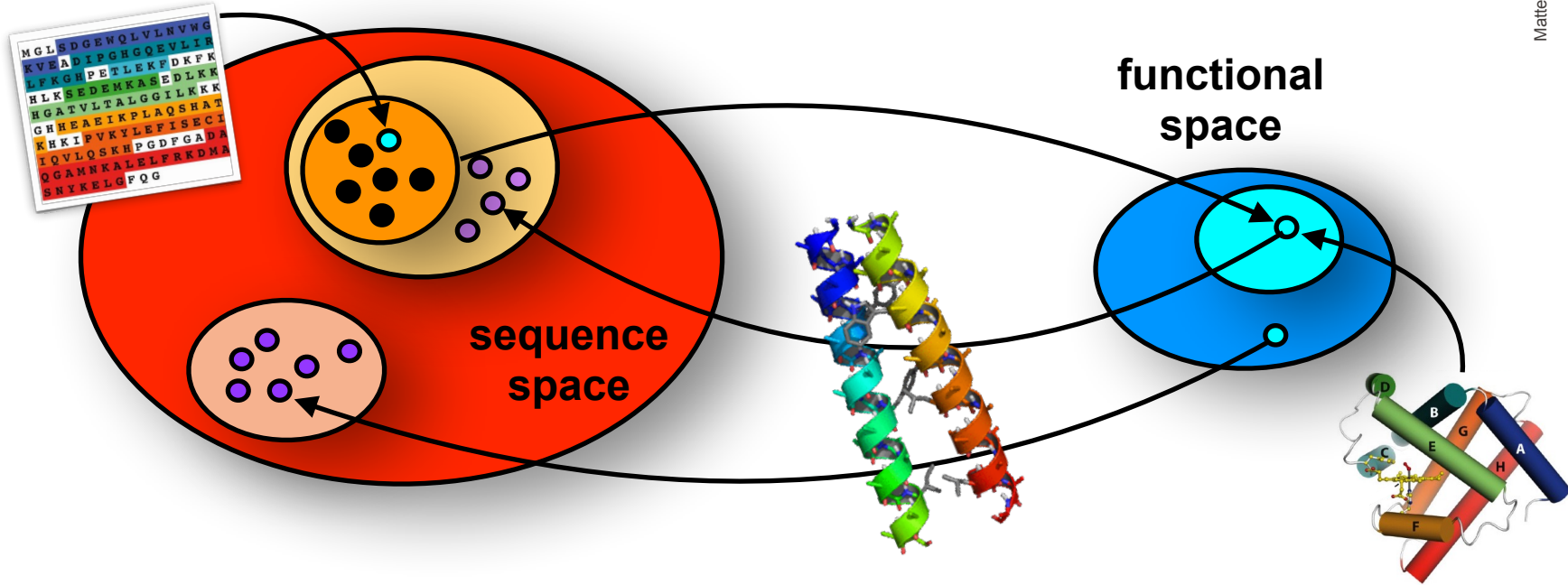
- Prediction of final structure and binding helps discovering new biology
- Not all the questions are answered though by AF2 !!

The sequence space is enormous



- 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 ($\sim 10^{10-20}$), structures lesser

EPFL The inverse folding problem — design



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

EPFL The origins: the Paracelsus challenge ('94)

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

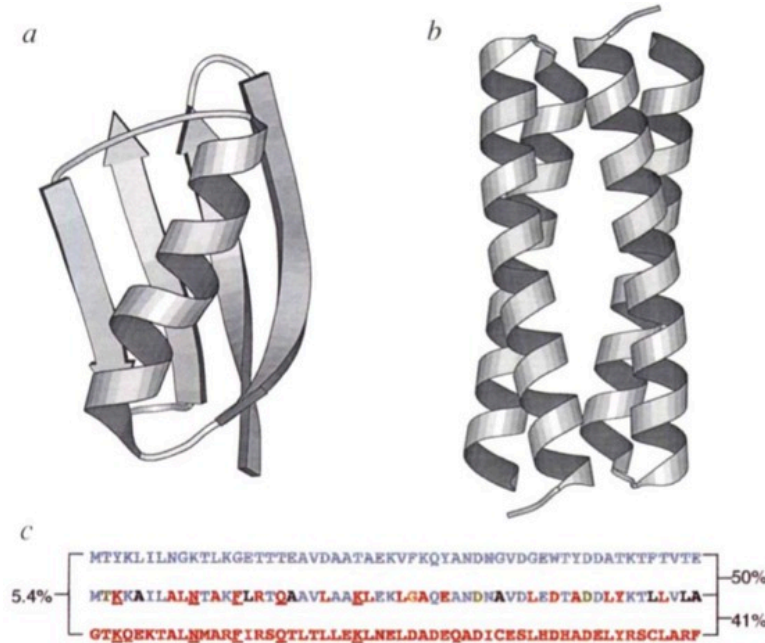
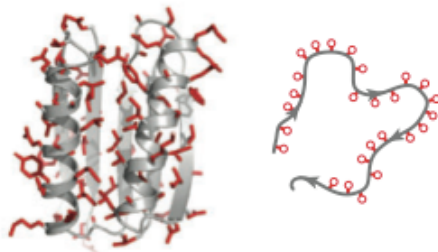


Fig. 1 Ribbon representation²⁹ of the folds of **a**, the B1 domain of IgG-binding protein G⁵ and **b**, Rop⁶. **c**, An alignment of the sequences of the B1 domain (blue), Rop (red) and Janus. Residues in Janus are coded as follows: blue, residues from B1; red, residues from Rop; underlined red, RNA-binding residues in Rop¹³; 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 Rop³⁰. 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 Rop³¹.

Multiple tasks for protein design

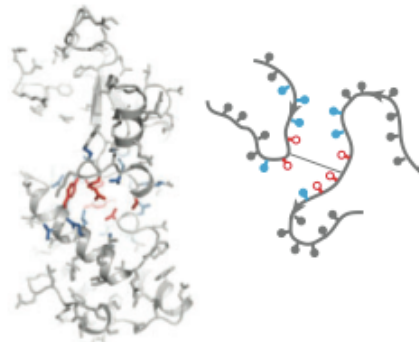
Protein design

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



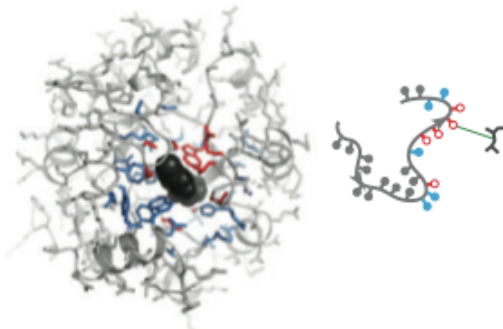
Protein-protein interface design

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



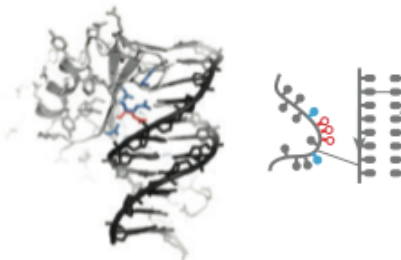
Enzyme design

- tailor enzymatic function
- improve thermostability
- catalyse new reactions



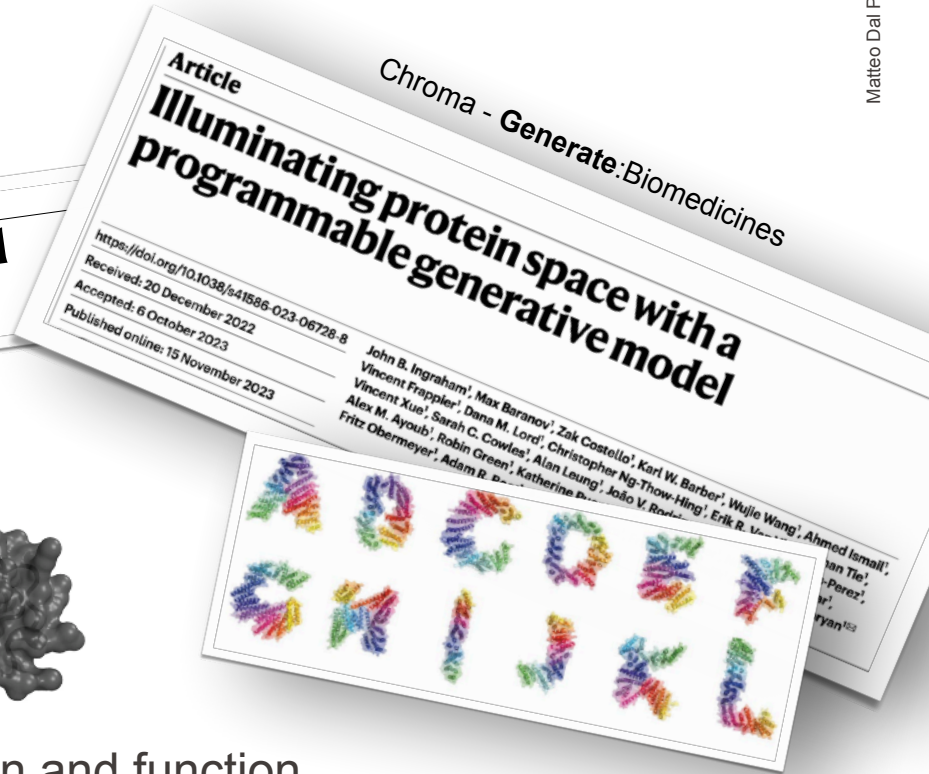
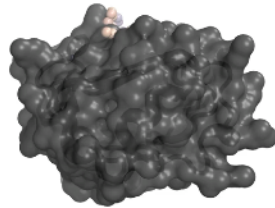
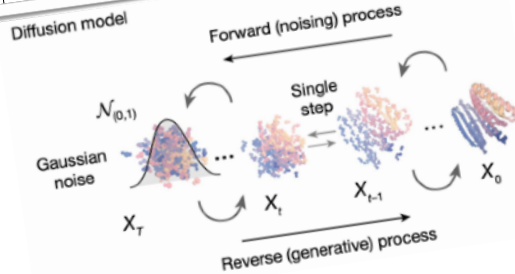
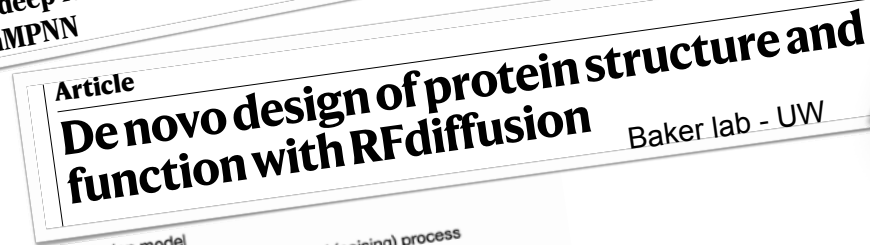
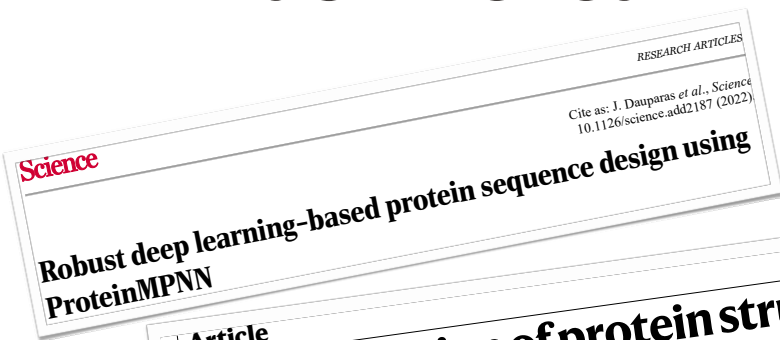
Protein-DNA interface design

- explore DNA interactions
- new therapeutic solutions



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

EPFL Machine learning for protein design



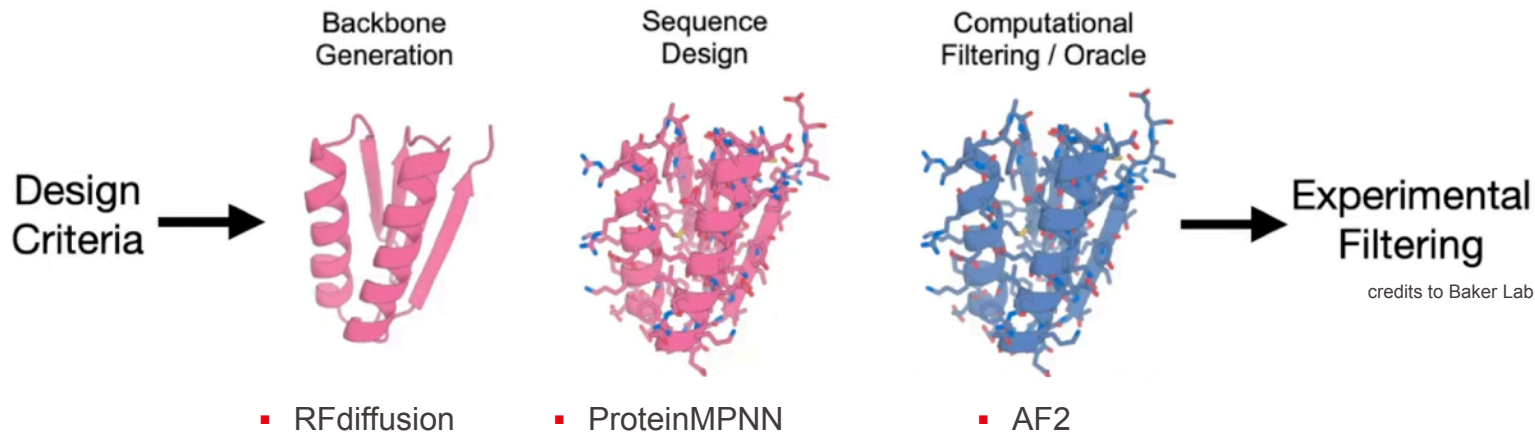
- Application to study protein evolution and function
- Protein engineering for therapeutics, synthetic biology and biotechnology

... leading to molecular engineering



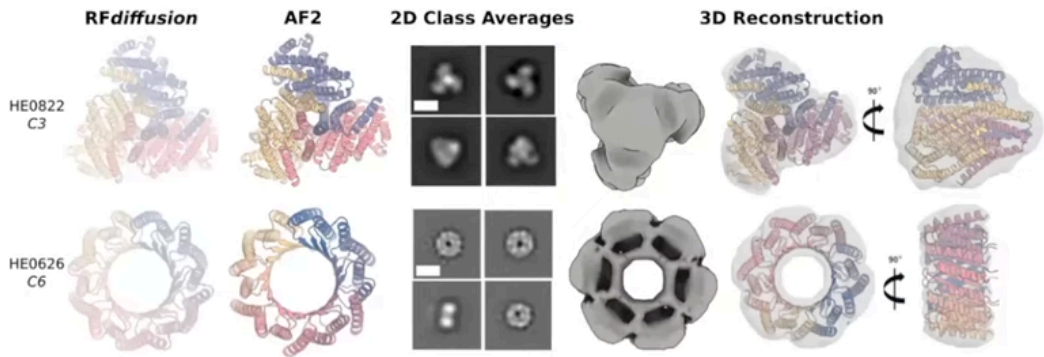
- Application to study protein evolution and function
- Protein engineering for therapeutics, synthetic biology and biotechnology

EPFL Pipeline of today's protein design

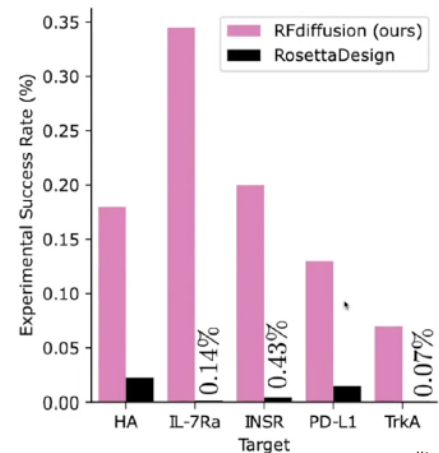


- AF2 has been key to filter potentially good protein designs
- Experimental testing is the ultimate validation of designs
- AI methods enhanced the experimental rate of success
- Protein engineering is now feasible for therapeutics, synthetic biology and biotechnology

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

EPFL Machine learning for protein design

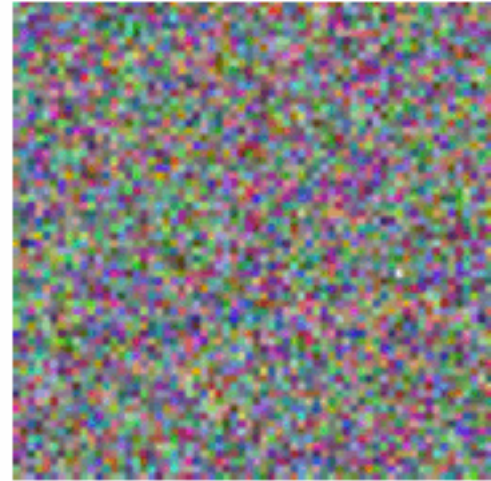
image



Forward diffusion
noisy image



noise



András Béres

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

De novo design of protein structure and function with RFdiffusion

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

Received: 14 December 2022

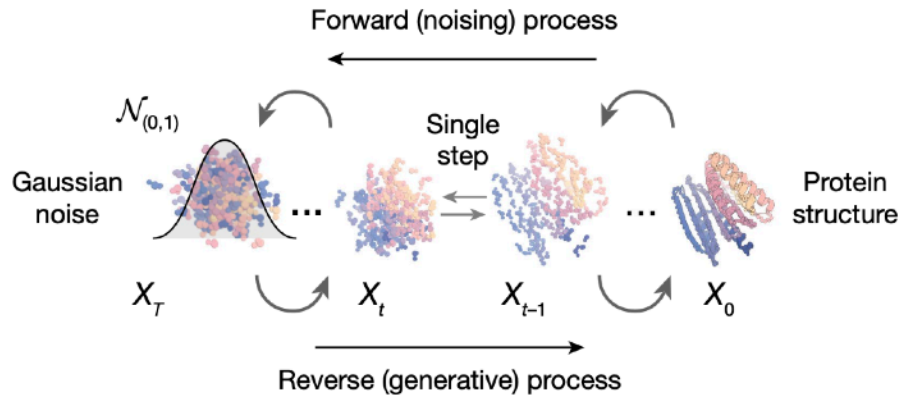
Accepted: 7 July 2023

Published online: 11 July 2023

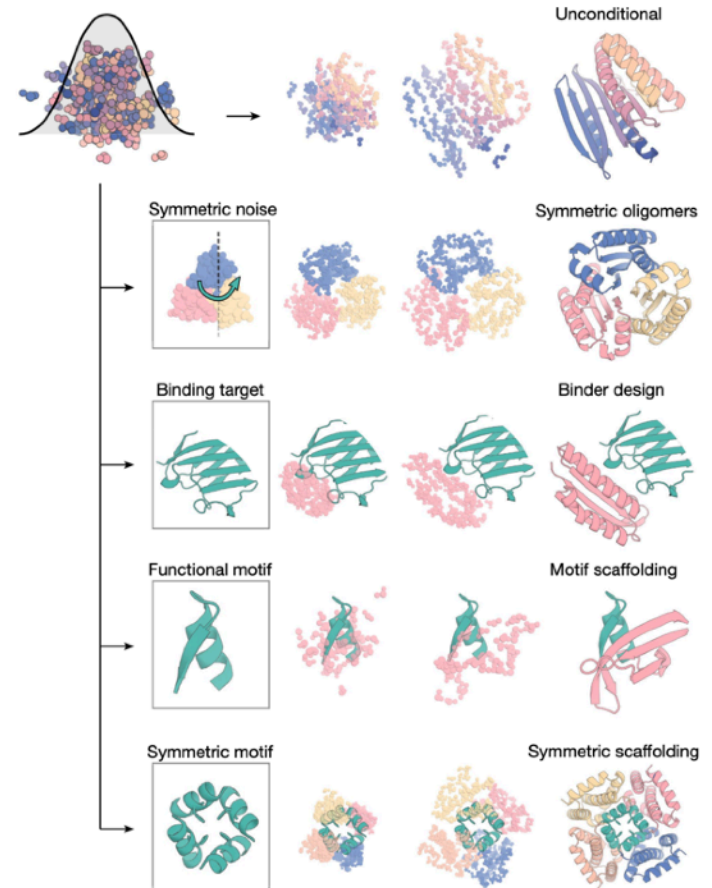
Open access

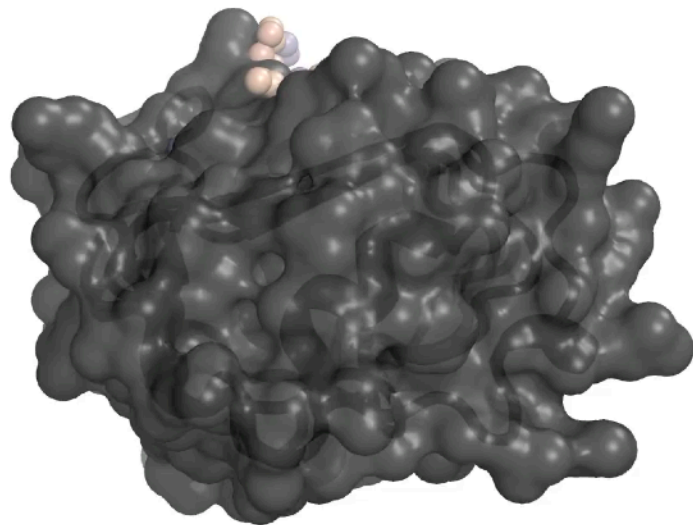
Joseph L. Watson^{1,2,15}, David Juergens^{1,2,3,15}, Nathaniel R. Bennett^{1,2,3,15}, Brian L. Trippe^{2,4,5,15}, Jason Yim^{2,6,15}, Helen E. Eisenach^{1,2,15}, Woody Ahern^{1,2,7,15}, Andrew J. Borst^{1,2}, Robert J. Ragotte^{1,2}, Lukas F. Milles^{1,2}, Basile I. M. Wicky^{1,2}, Nikita Hanikel^{1,2}, 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 Bortoli⁸, Emile Mathieu¹⁰, Sergey Ovchinnikov^{11,12}, Regina Barzilay⁶, Tommi S. Jaakkola⁶, Frank DiMaio^{1,3}, Minkyung Baek^{1,3} & David Baker^{1,2,14,15}

Diffusion model



- 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.





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

Robust deep learning-based protein sequence design using ProteinMPNN

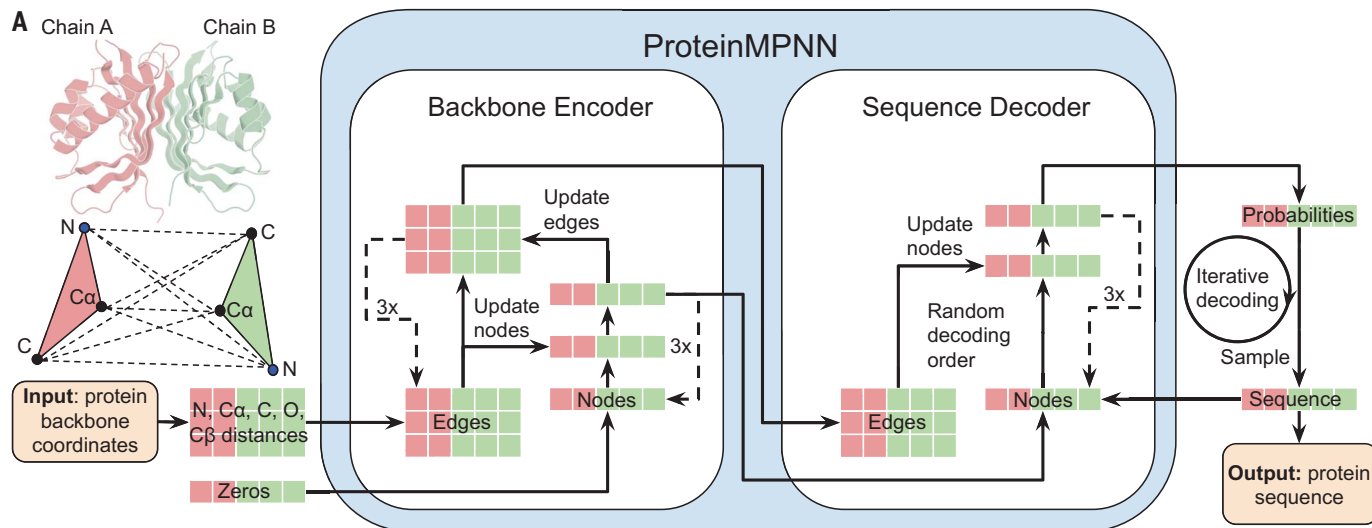
J. Dauparas^{1,2}, I. Anishchenko^{1,2}, N. Bennett^{1,2,3}, H. Bai^{1,2,4}, R. J. Ragotte^{1,2}, L. F. Milles^{1,2}, B. I. M. Wicky^{1,2}, A. Courbet^{1,2,4}, R. J. de Haas⁴, N. Bethel^{1,2,4}, P. J. Y. Leung^{1,2,3}, T. F. Huddy^{1,2}, S. Pellock^{1,2}, D. Tischer^{1,2}, F. Chan^{1,2}, B. Koepnick^{1,2}, H. Nguyen^{1,2}, A. Kang^{1,2}, B. Sankaran⁴, A. K. Bera^{1,2}, N. P. King^{1,2}, D. Baker^{1,2,4*}

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

*Corresponding author. Email: dabaker@uw.edu

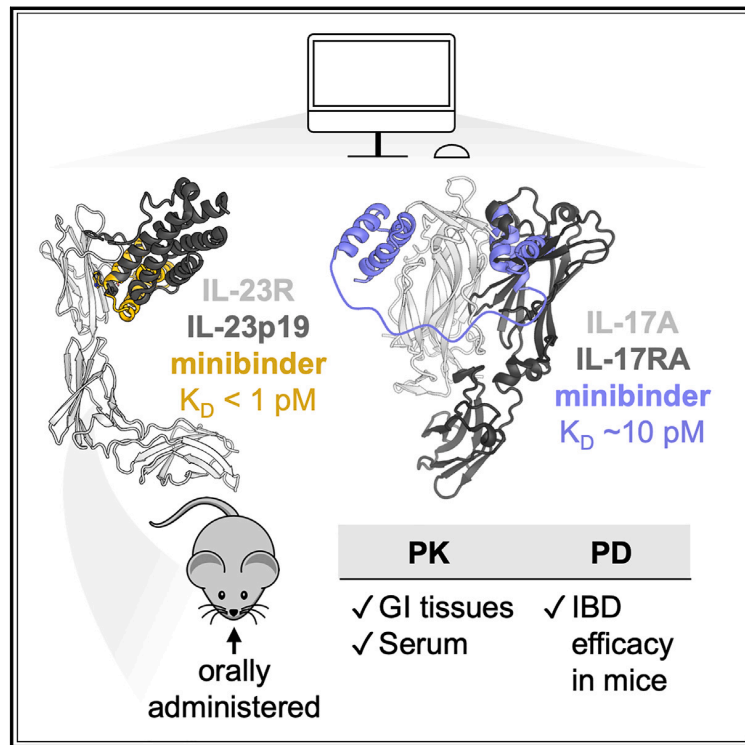
While deep learning has revolutionized protein structure prediction, almost all experimentally characterized de novo protein designs have been generated using physically based approaches such as Rosetta. Here we describe a deep learning-based protein sequence design method, ProteinMPNN, with outstanding 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.



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

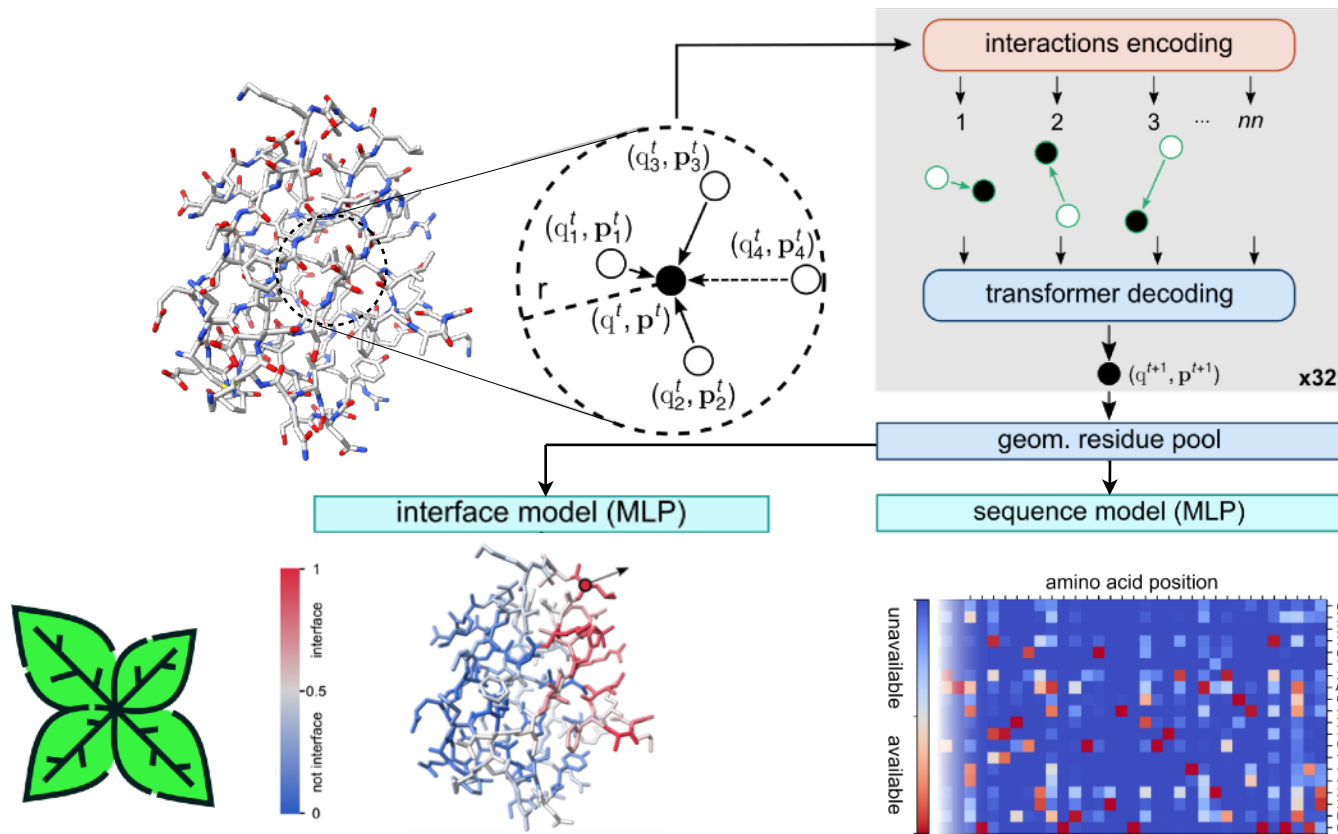
berger389@gmail.com (S.B.),
dabaker@uw.edu (D.B.)

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 © 2024 The Author(s). Published by Elsevier Inc.
<https://doi.org/10.1016/j.cell.2024.05.052>

EPFL Protein Structure Transformer @LBM



PeSTo: binding interfaces

Krapp et al. *Nat Comms* 2023

CARBonAra: molecular design

Krapp et al. *Nat Comms* 2024

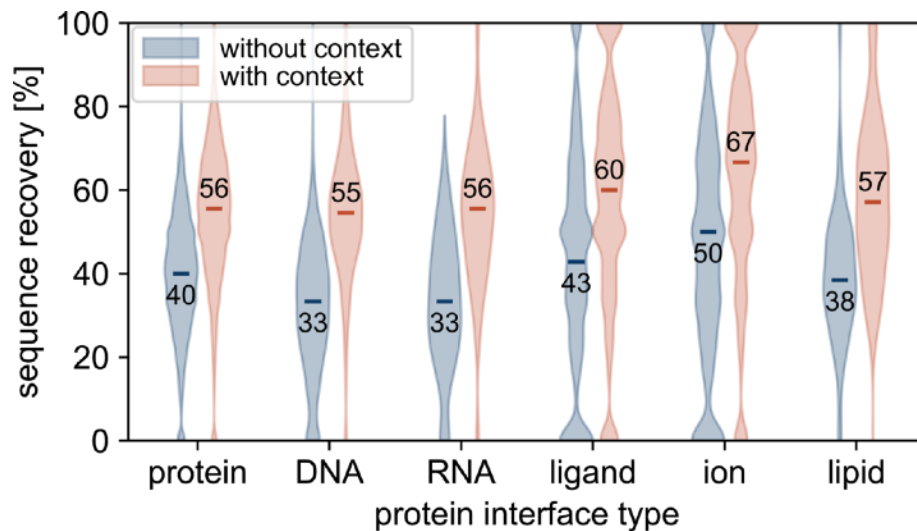
EPFL Unique ability — context awareness

■ example with context



colicin E7

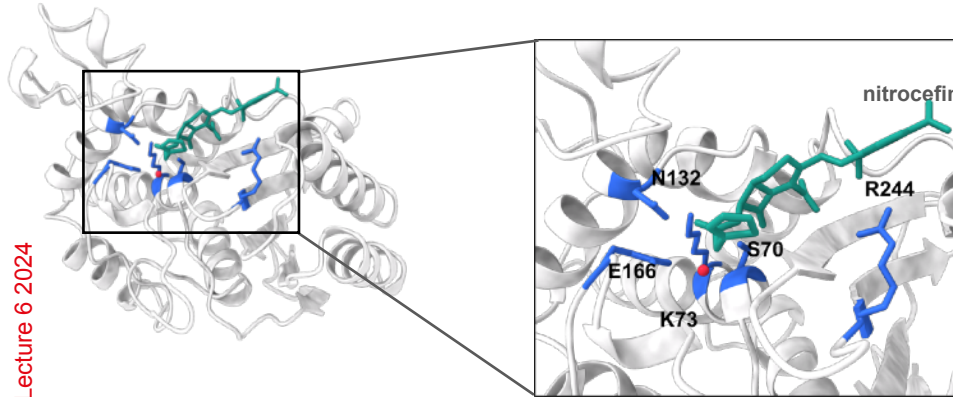
■ large-scale benchmark



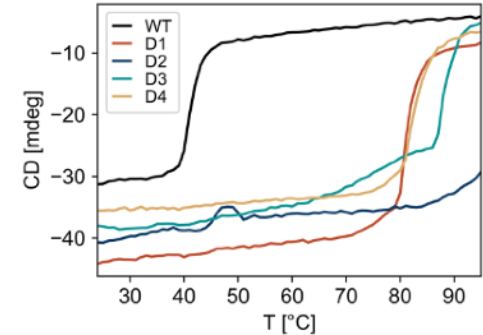
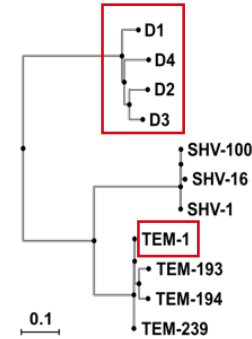
1000 structures sampled with maximum 30% sequence identity and separate C.A.T.H. classification from training set

Can we re-engineer an enzyme?

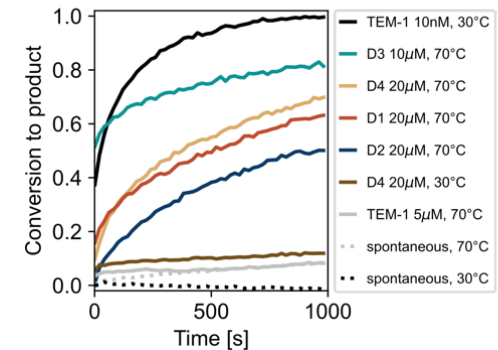
■ TEM-1 serine β -lactamase



only 50% sequence identity



- sequences generation with substrate as constraint
- selected 10 top-ranked predictions based on pIDDT
- 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



The future is bright and exciting ...

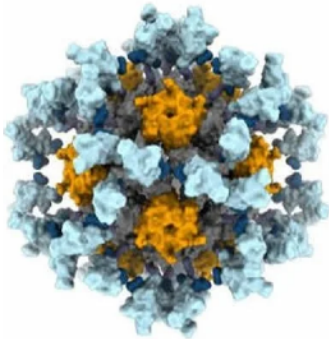
... biomolecular design will address many societal needs

■ Medicine

vaccines & antivirals

smart medicines

drug delivery



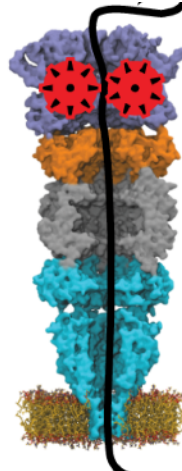
SARS-CoV-2 RBD
nanoparticle immunogen (Cell 2020)

■ Biotechnology

protein-silicon devices

bio-based computers

nanoscale manufacturing



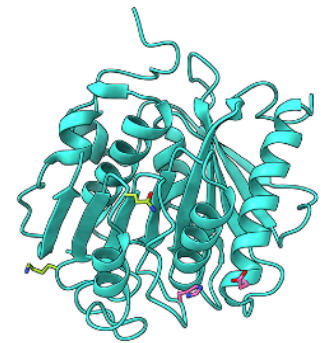
SM proteomics with
biological nanopores
(Nat Chem 2021)

■ Sustainability

artificial photosynthesis

CO₂ sequestration

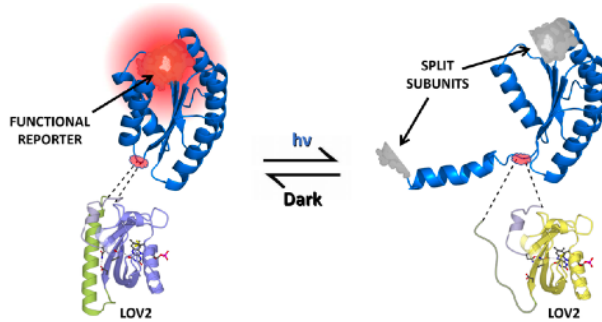
plastic degradation



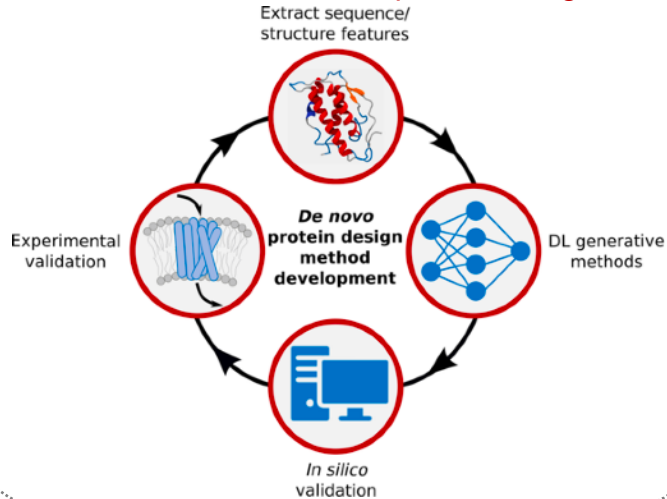
FAST-PETase
(Nature 2022)

Laboratory of Protein and Cell Engineering

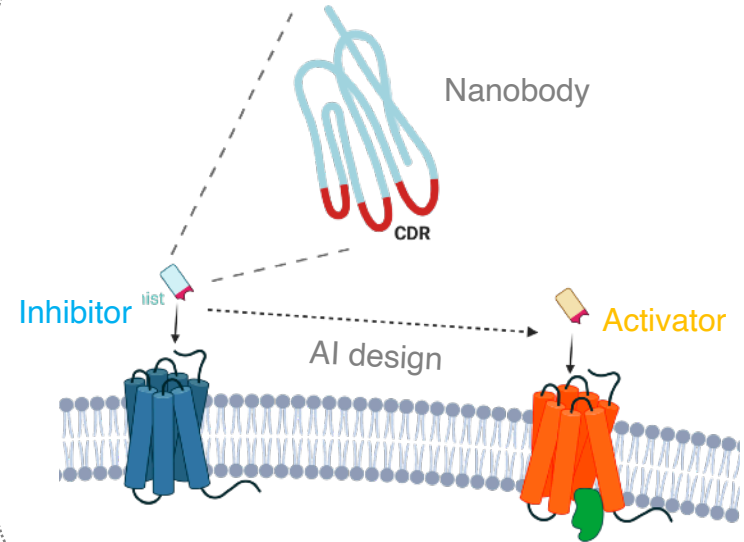
De novo optogenetic switches



De novo membrane protein design



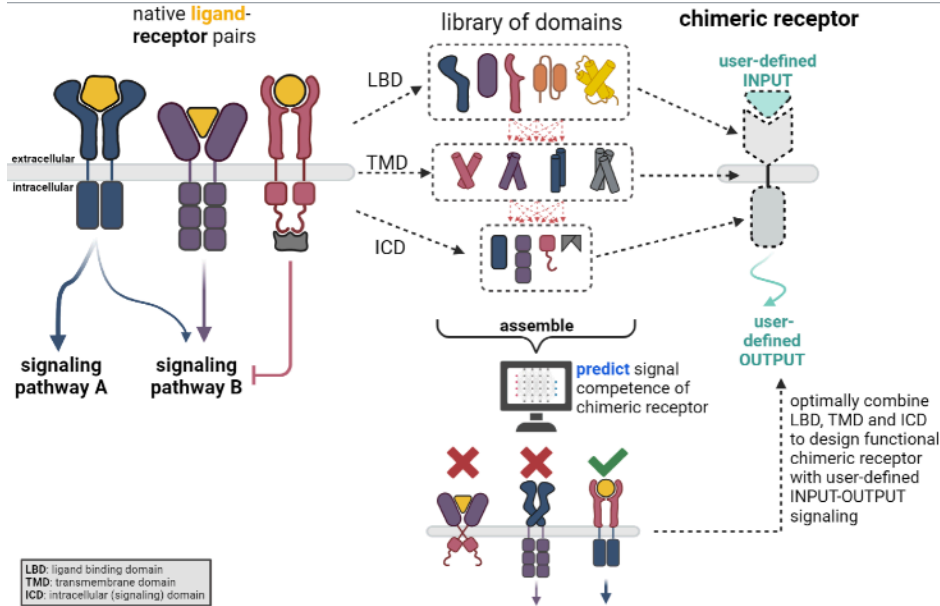
De novo therapeutic protein ligands



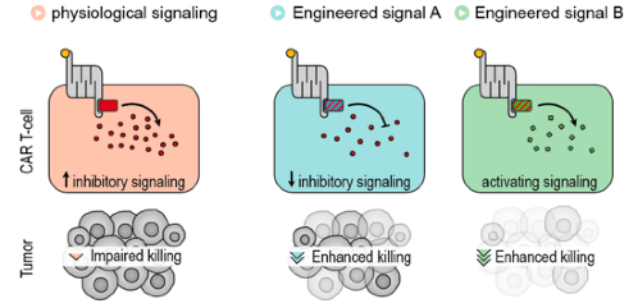
Barth Lab

Laboratory of Protein and Cell Engineering

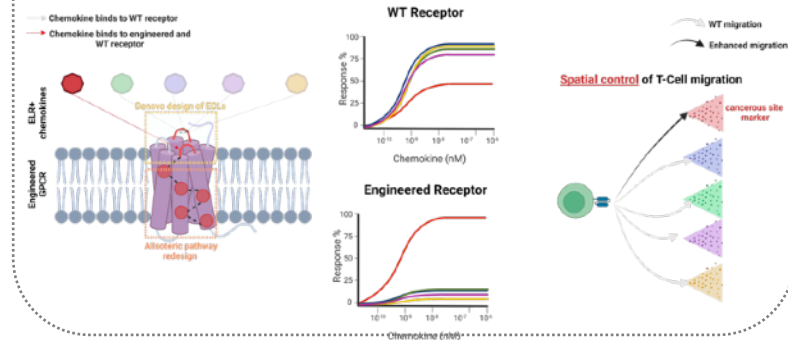
Allosteric biosensors for engineered cell therapies



G-protein signal rewiring to boost CAR T-cells



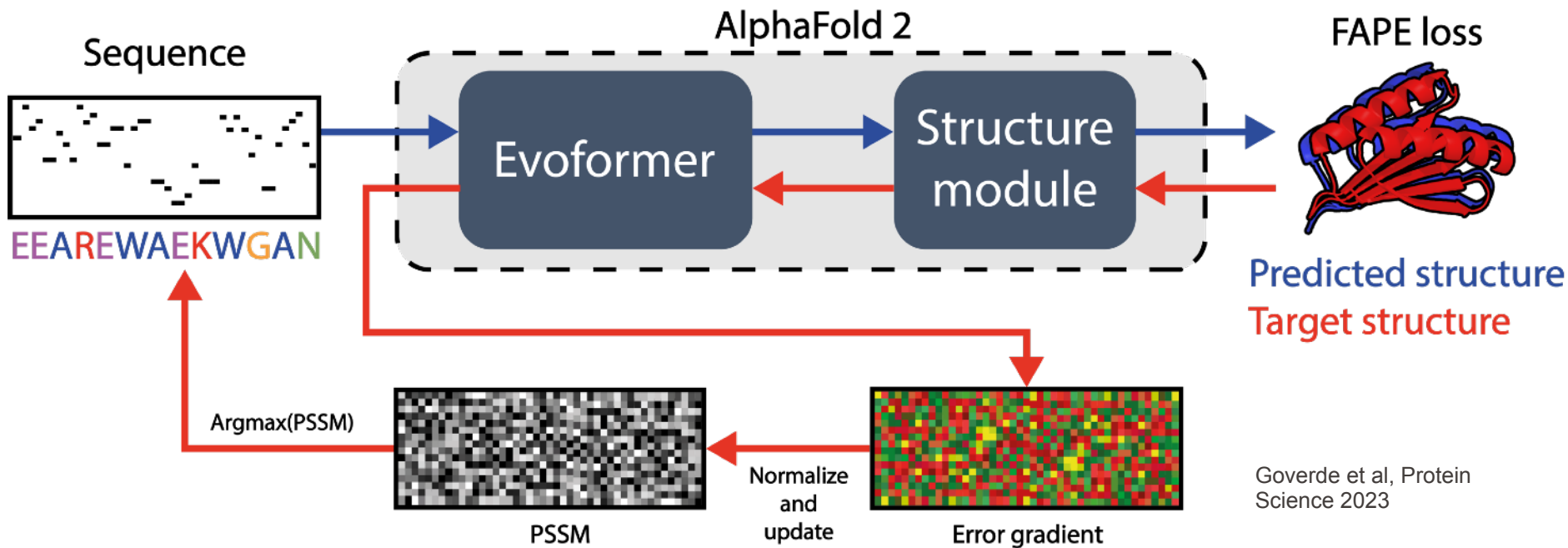
Redirecting cell migration with designed receptors



EPFL Inverting AlphaFold for protein design

57

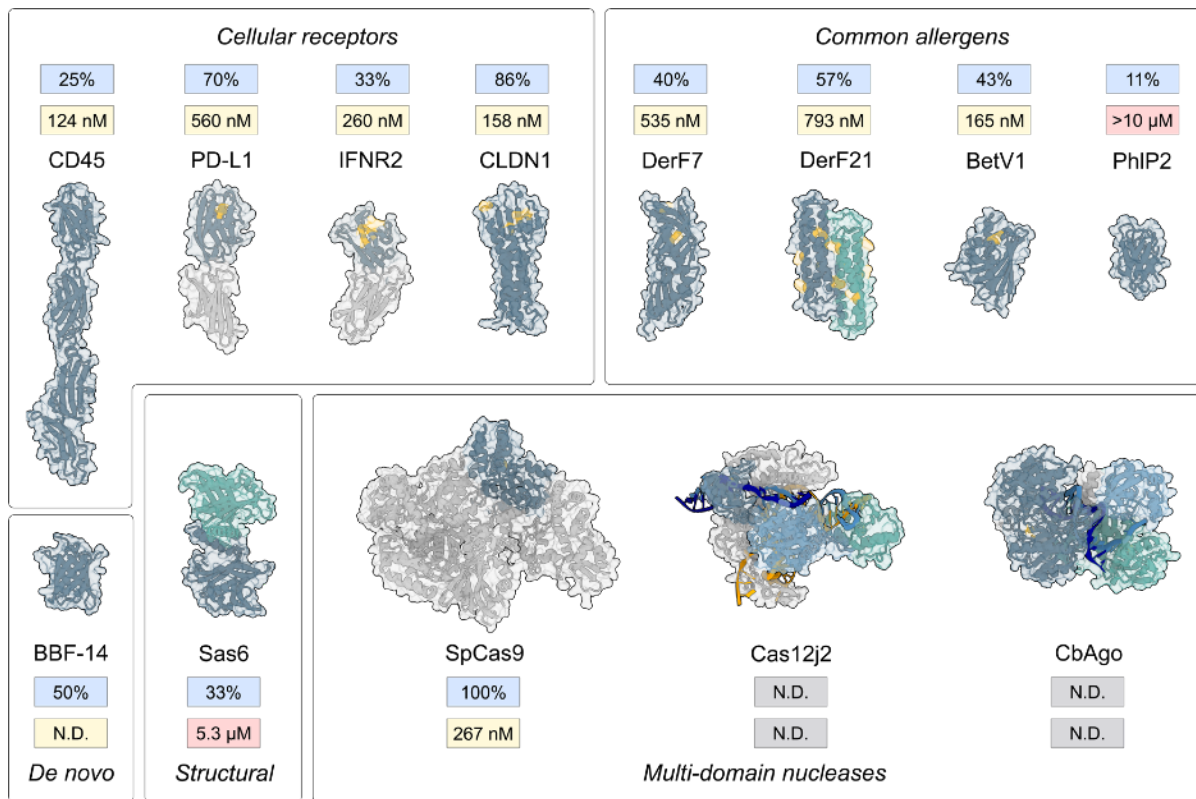
Bruno Correia, Laboratory of Protein Design and Immunoengineering



-Final sequences designed with proteinMPNN on AF2 generated backbones

High experimental success rates in binder design

Bruno Correia, Laboratory of Protein Design and Immunoengineering



Experimental success rate

Highest affinity binder (no experimental optimisation)

Pacesa & Nickel, et al., *unpublished*

We weren't alone !!!!

Google DeepMind

About ▾

Research

Technologies ▾

Impact

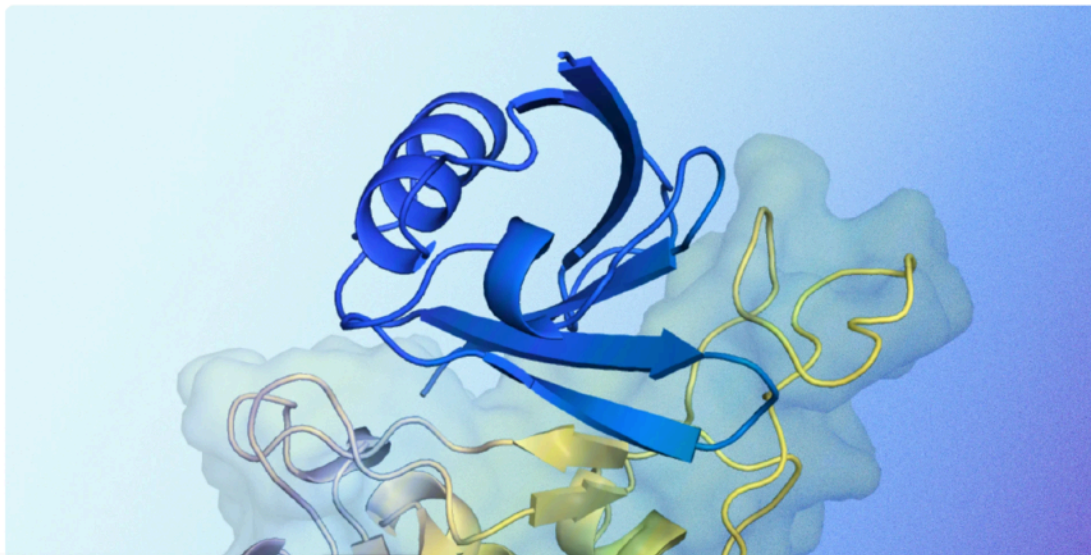
Discover ▾

AlphaProteo generates novel proteins for biology and health research

5 SEPTEMBER 2024

Protein Design and Wet Lab teams

[Share](#)



Designing Life with AI

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.

EPFL

MAKE!
USEFUL. CREATIVE. SUSTAINABLE.

<https://www.designinglifewithai.com/>

[contact the MAKE team for ongoing projects offered by labs](#)

