DELPHI
AlphaFold and Protein Structure Visualization Workshop

Roland Dunbrack
Fox Chase Cancer Center
roland.dunbrack@fccc.edu

June 24, 2026

Fox Chase Cancer Center
Molecular Modeling Facility (2003-present)

~100 publications, >8000 citations

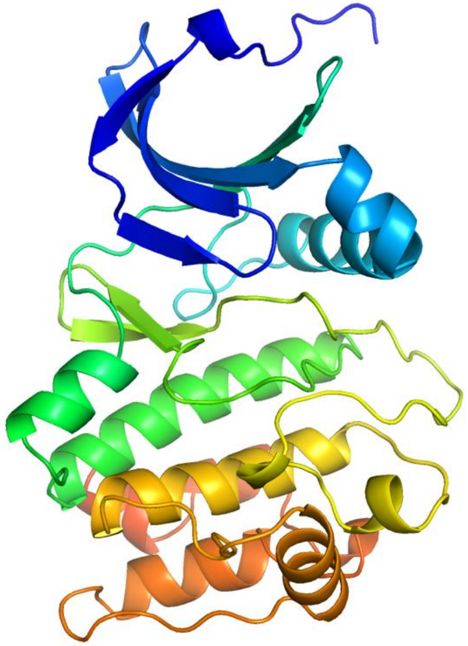
Program	PI	Project
CPC	Arora	JMJD1C, RNF8, RNF168, and MDC1 complexes
CSM	Astsaturov	Ano6 interactions with substrates and inhibitors
CSM	Balachandran	RIPK3 orthologs; RSV-NS2/RNA and N4BP1/RNA binding
CSM	Connolly	CDON interactions with mAb against, cadherins, and hedgehog proteins
CSM	Cukierman	PallD, ACTN1,4, Syndecan complexes
CSM	Duncan	MRCK & GEF-H1
CSM	Kruger	Filaments of cystathionine beta synthase
CSM	Makhov	FGFR-FGF and IL-21/IL21R complexes and protein design for kidney cancer pre-clinical therapeutics

Program	PI	Project
CSM	Torres-Ayuso	FAT1 interaction with TNIK
CSM	Wang, L	BTK kinase mutations and protein interactions in lymphoma
CSM	Wang, R	PTPsigma and LAR dimerization
NDC	Bellacosa	TDG:p300 and TDG-CBP complexes
NDC	Chen	Coilin, TCAB1, SMN1, DKC1, GAR1, telRNA
CSM	Duncan	MRCK & GEF-H1
CSM	Kruger	Filaments of cystathionine beta synthase
NDC	Estaras	TEAD4-YAP1-NR2F2 complexes with DNA
NDC	Kappes	Mutations and protein interactions of THPOK

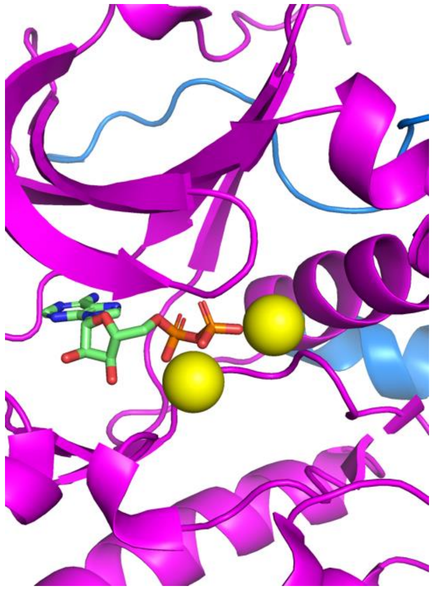
Program	PI	Project
NDC	Liu	SMC1a, SMC3, RAD21, STAG2, MIS4, CTCF
NDC	Skorski	DNA polymerase theta interactions with PARG and PARP1
NDC	Whetstine	KDM3B, TOPOII, CTCF complexes
NDC	Whitaker	Heterodimer complexes of APE1 with transcription factors STAT3, HIF1a, JUN, p53, NFkB1 and interactions with DNA
NDC	Wiest	γδ TCR complexes with Butyrophilins
NDC	Yang	T3 effects on Thyroid hormone receptor TRa1 interactions with EZH2
TUHS	Elrod	Predicting complexes of Glyoxalase-Domain Containing Protein 4 (GLOD4) with NCLX
TUHS	Thomas	DLK kinase inhibitors

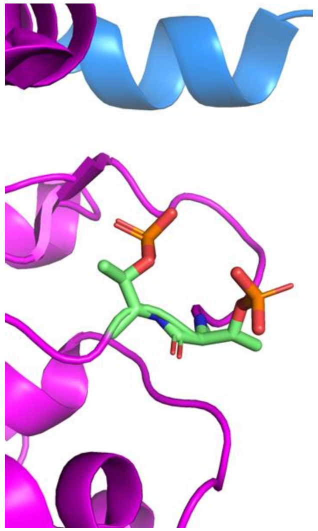
AlphaFold and related programs: 10 projects a year → 40 projects a year

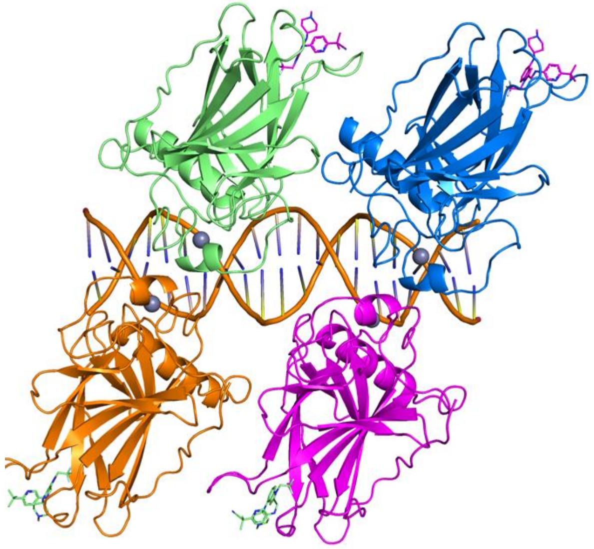
Molecular Structure Prediction


Single protein chains


Protein-protein/peptide complexes


Protein-ligand complexes (including ions)


Protein PTMs


Protein/DNA complexes

Also: RNA folding, prediction of conformational states and ensembles, affinities, kinetics, etc.

Protein Structure Prediction Timeline

Comparative modeling **1969 Browne et al, 1980 Greer, 1993 Modeller**

- Find template related to query
- Align template to query
- Copy backbone and same sidechains
- Model loops and different sidechains (SCWRL4, Rosetta)
- Refinement (MD, Rosetta)

Fold recognition/threading **1992 Thornton, Blundell 1993, Sippl 1995**

- Align sequence to (unrelated) templates max of contact/ss/solvation potentials
- Loop modeling and sidechain placement

Fragment assembly **1997 Rosetta, 2007 iTASSER**

- Find similar-sequence fragments to query with predicted ss
- Assemble with Monte Carlo algorithm
- Refine

Prediction of Protein Complex Structures Timeline

Protein-protein docking **1978 Wodak and Janin, 1990s GRAMM, 2003 RosettaDock**
2003 ZDOCK HADDOCK, 2000s PIPER/ClusPro

- Search over whole surface with FFTs
- Rigid-body sampling, flexible sampling (backbone, sidechains)

Protein-ligand docking **1982 DOCK, 1996 AutoDock, 1997 GOLD, 2004 Glide**

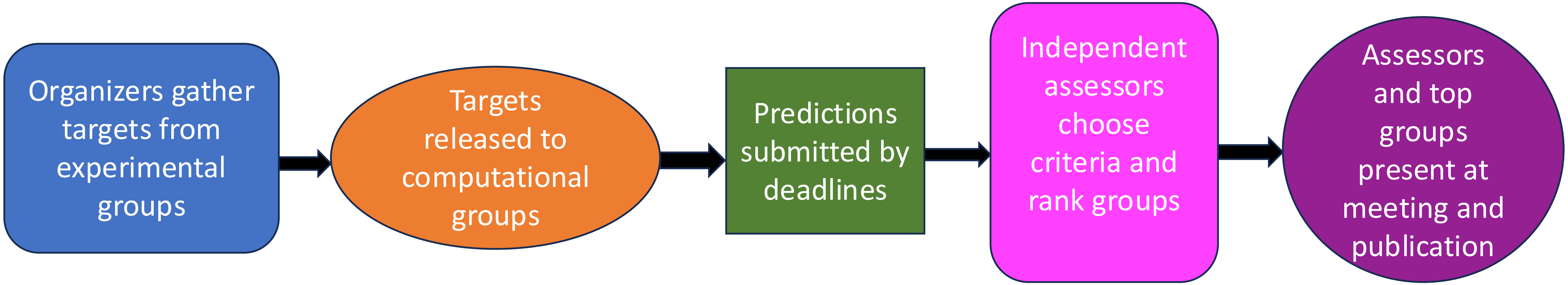
- Overlapping spheres placed in binding site,
- Incremental construction
- Scoring functions, flexibilityof active site

Protein-DNA docking **1990s ZDOCK, GRAMM, 2003 HADDOCK, 2013 RNPDock, 2020 P3DOCK**

- PPI FFT methods
- Nucleic-acid specific scoring functions
- Flexibility of DNA

CASP Experiments

Critical Assessment of Structure Prediction (1994-present, biannually)



Organizers gather targets from experimental groups

Targets released to computational groups

Predictions submitted by deadlines

Independent assessors choose criteria and rank groups

Assessors and top groups present at meeting and publication

January

March-June

March-June

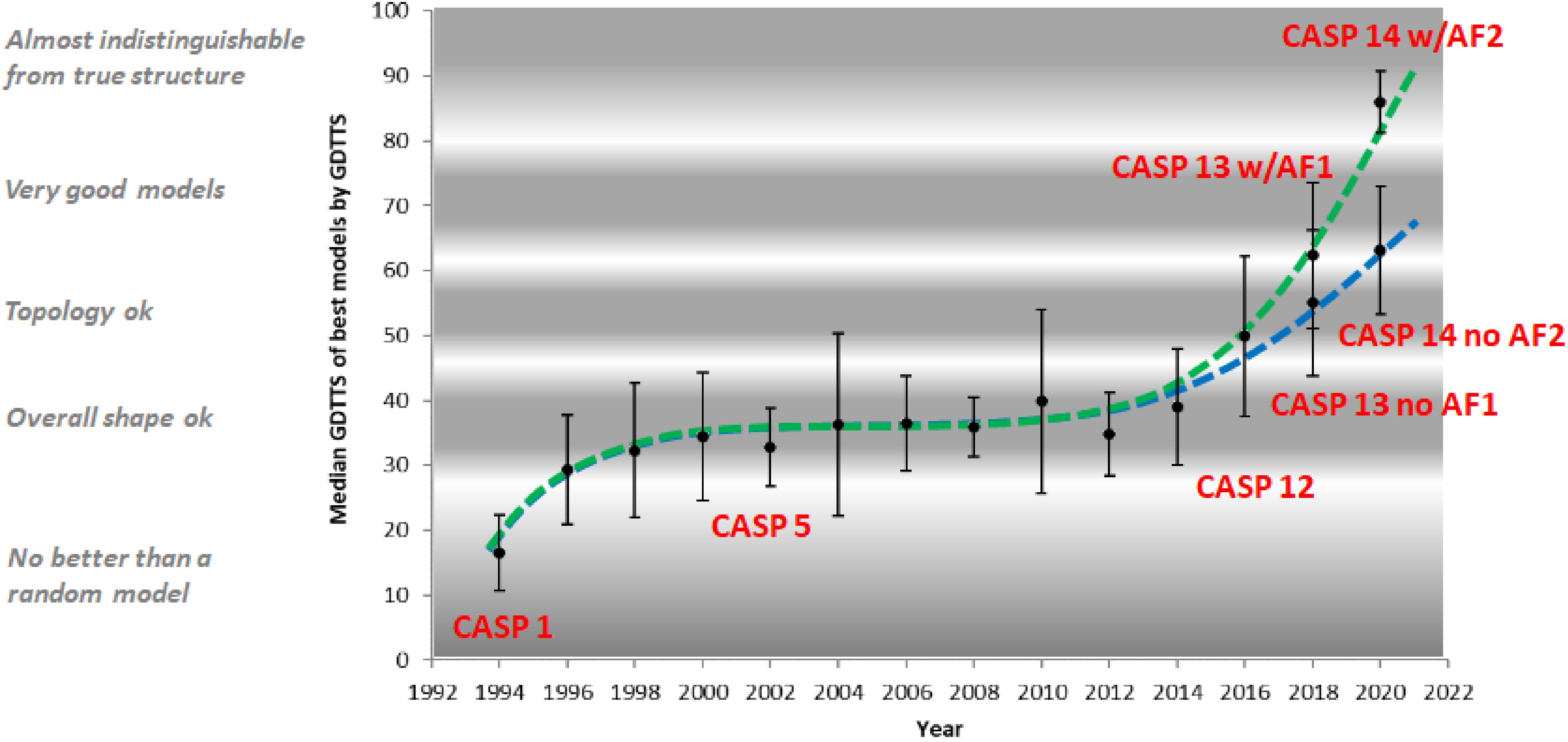
July-September

December

Dunbrack group:
CASP6, CASP11

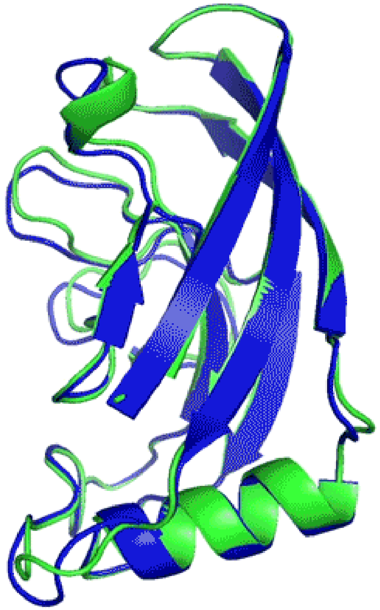
AlphaFold2 in the CASP Blind Structure Prediction Contest (Dec. 2020)

Code available July 2021 – single proteins



Example Prediction: A free model target at CASP14

T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)



T1049 / 6y4f
93.3 GDT
(adhesin tip)

🟢 Experimental result

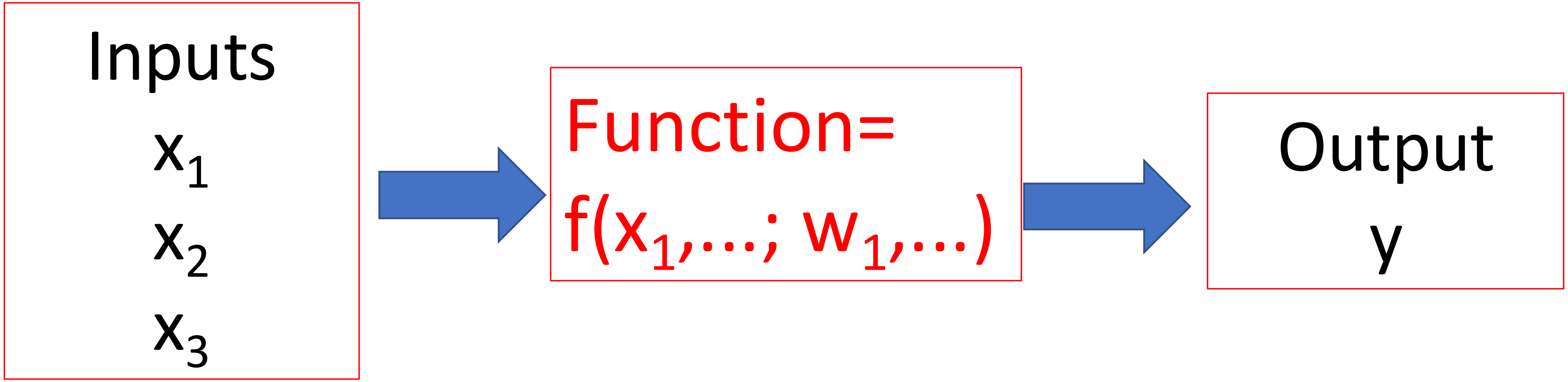
🔵 Computational prediction

How: Deep learning from multiple sequence alignments of 70,000 different proteins and the entire Protein Data Bank (PDB)

Fundamentally a *knowledge-based* inference system

Query sequence + MSA + Templates (optional)
→ Structure

What does a neural network do?



Inputs

x_1

x_2

x_3

Function=

$f(x_1,\ldots;\ w_1,\ldots)$

Output

y

Output can be a class (or classes): cat, dog, tree = Classifier

Output can be a number (or numbers) = Regression

Function has trainable parameters, $\{w_j\}$

Trained on real data: pairs of $\{x_1,x_2,x_3\}$, $\{y\}$

y can also be a vector of numbers

Three "tricks" in neural networks

Trick #1. Linear combination of input data with "weights" to be learned

Inputs

x_1

x_2

x_3

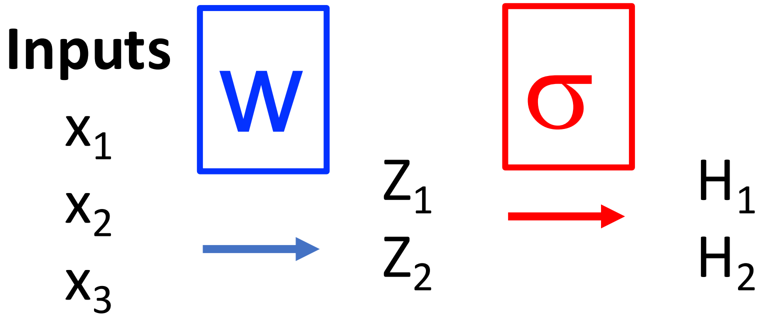
$$Z_1 = w_{11}{}^{x}X_1 + w_{12}{}^{x}X_2 + w_{13}{}^{x}X_3$$

$$Z_2 = w_{21}{}^{x}X_1 + w_{22}{}^{x}X_2 + w_{23}{}^{x}X_3$$

Matrix multiplication of vector **x** and matrix **w**

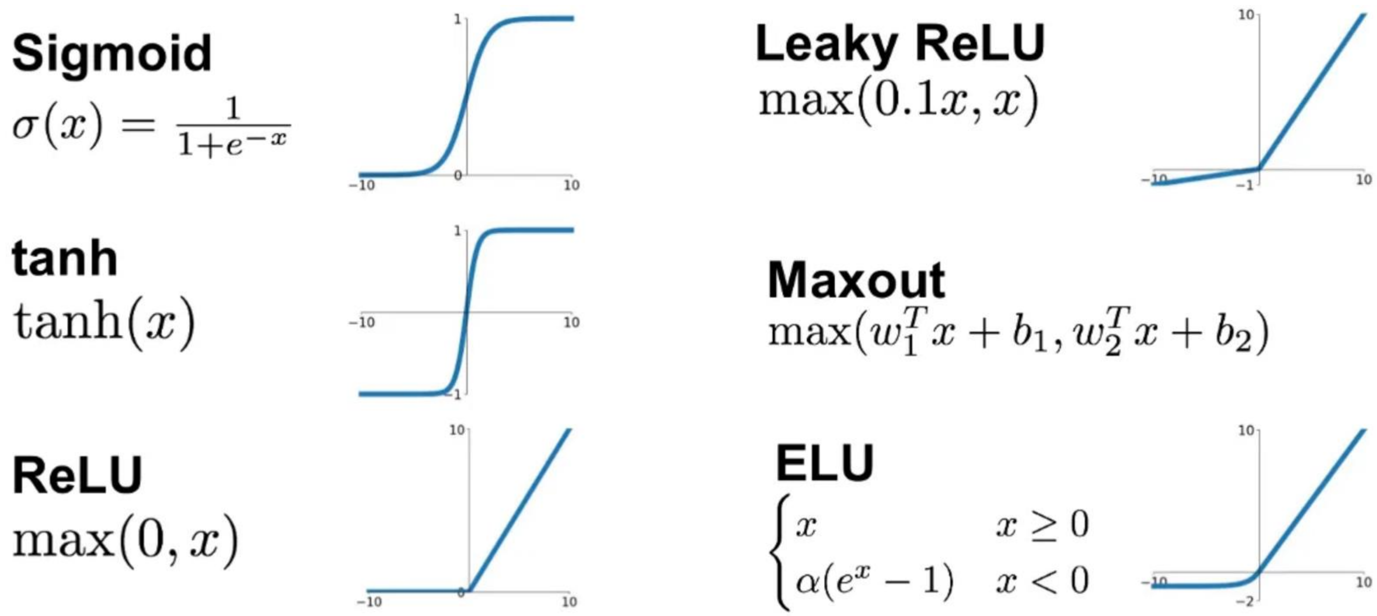
Z=wx

Trick #2. Apply an “Activation Function” (σ)



numbers between 0 and 1
or between -1 and 1

Activation Functions (σ)



Sigmoid
$\sigma(x) = \frac{1}{1+e^{-x}}$

tanh
$\tanh(x)$

ReLU
$\max(0, x)$

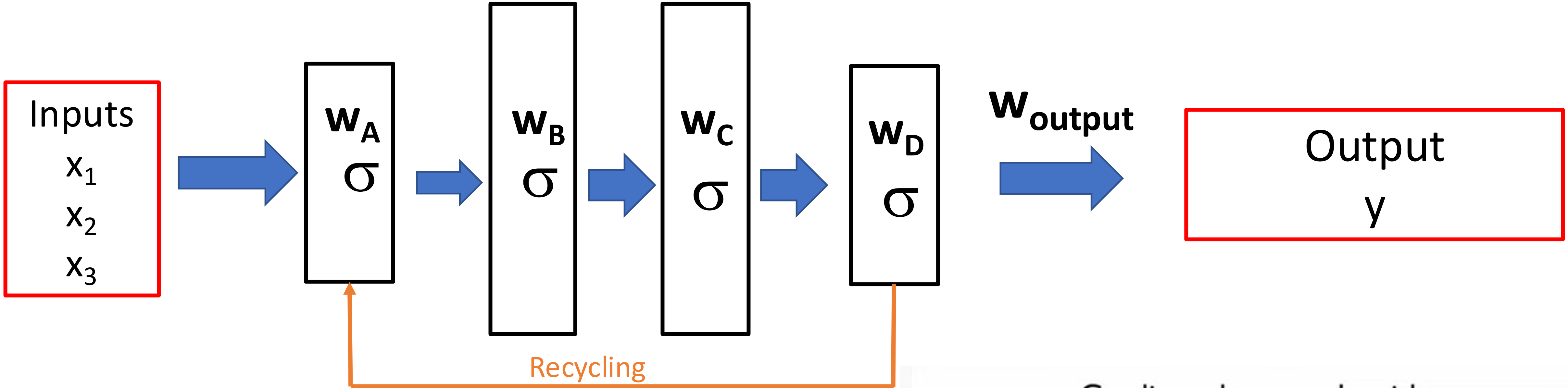
Leaky ReLU
$\max(0.1x, x)$

Maxout
$\max(w_1^T x + b_1, w_2^T x + b_2)$

ELU
$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$

Take any number at turn it in a number between 0 and 1, or between -1 and 1

Trick #3. Multiple layers, possibly scaling the number of dimensions up or down + recycling



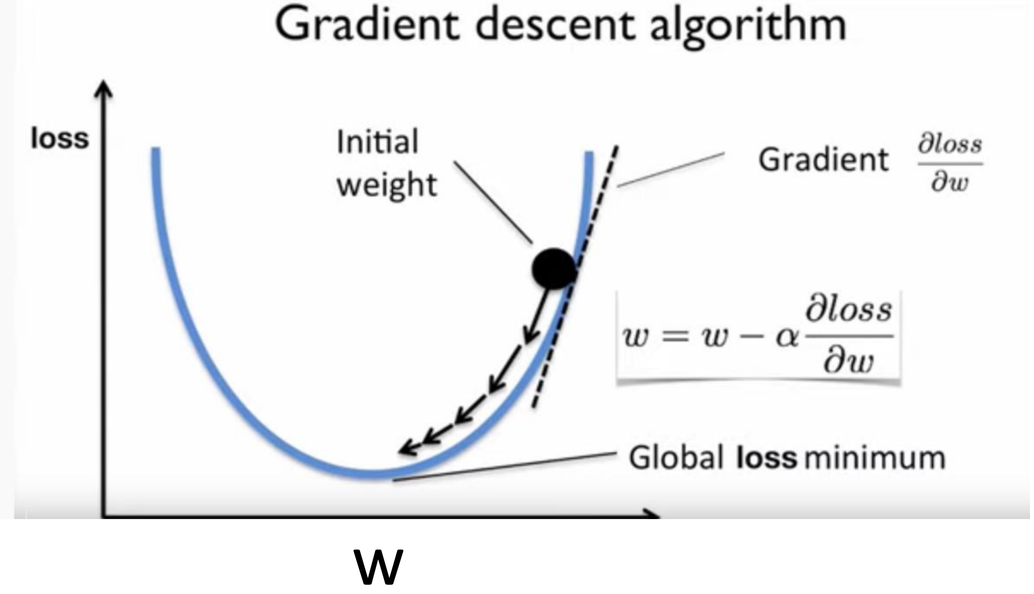
Learn the weights w_A, w_B, w_C, w_D, w_{output}

(1) Random start

(2) Calculate y for training x;

(3) Calculate derivatives of loss as f(w);

(4) Adjust w

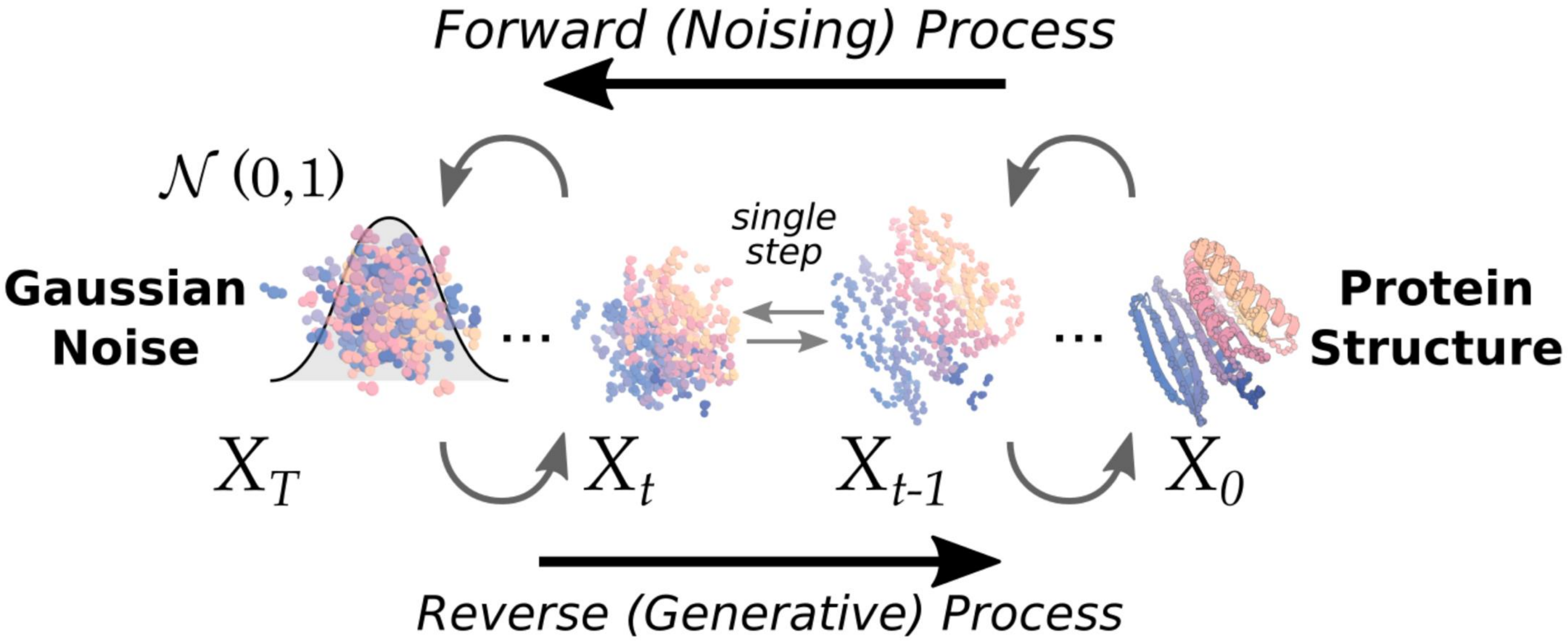


Trick #4. Diffusion neural networks

Generate forward “data” for training: progressively add noise to real objects (images, protein structures) with known information (image of a ”dog”, sequence+molecule types+MSA)

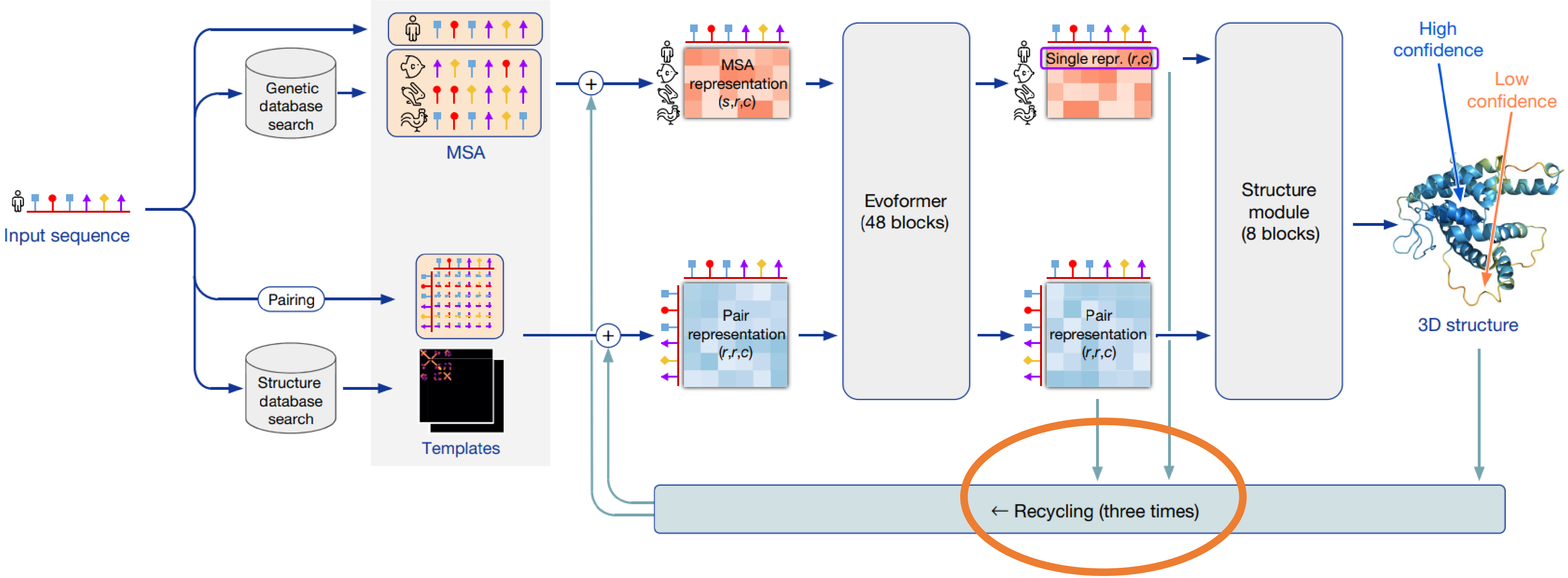
Use reverse “data” to train networks to denoise noisy data under constraints (protein sequence, molecular types, MSA, etc.)

In daily use: start with random points for atoms, denoise under constraints



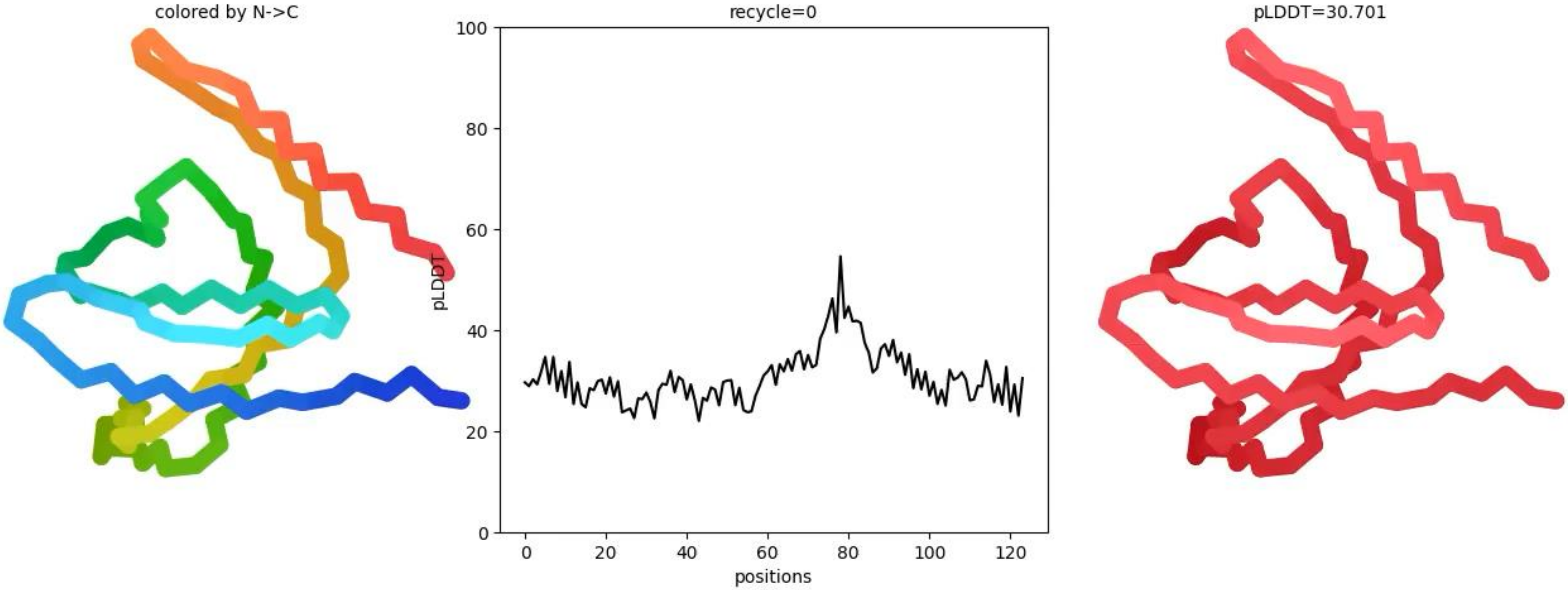
AlphaFold2 random seeds and recycles

Random seed enables random chunks of MSA to be sampled

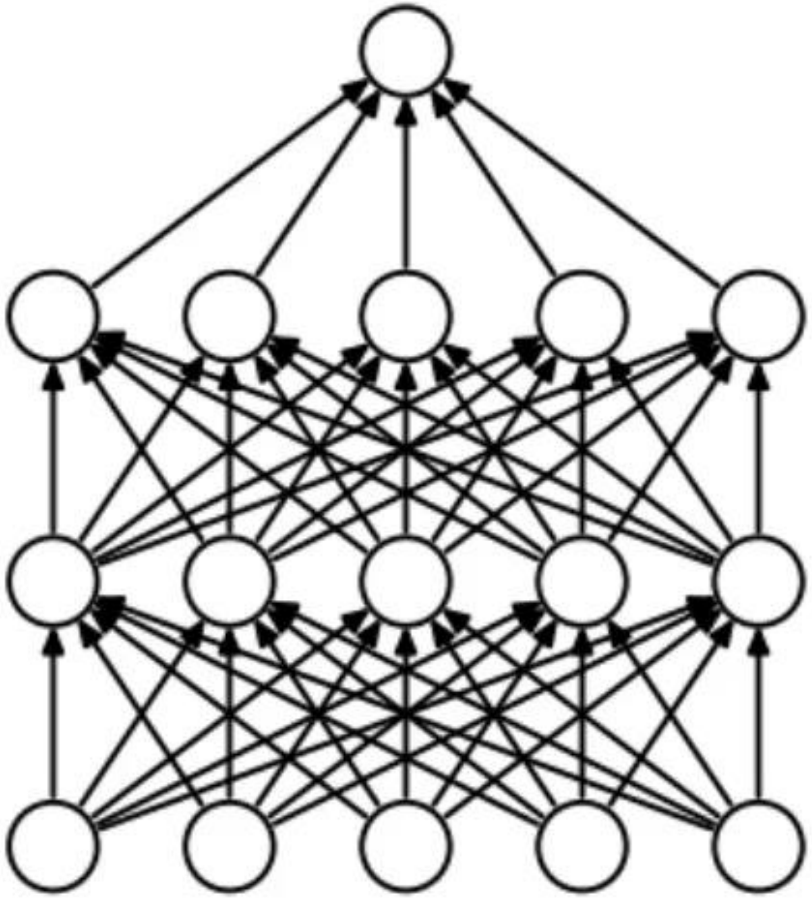


Each box is a vector of length 128 or 384

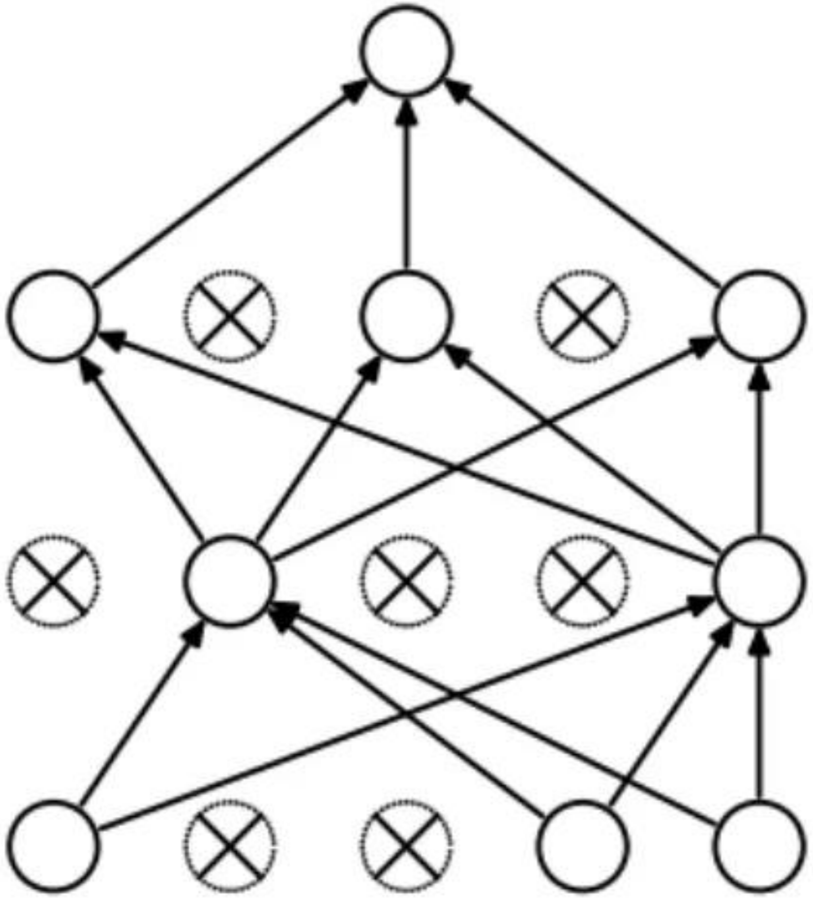
Recycles


colored by N->C
recycle=0
pLDDT=30.701
pLDDT
positions
0
20
40
60
80
100
120

Dropout in Neural Network



(a) Standard Neural Net



(b) After applying dropout.

AlphaFold-Multimer

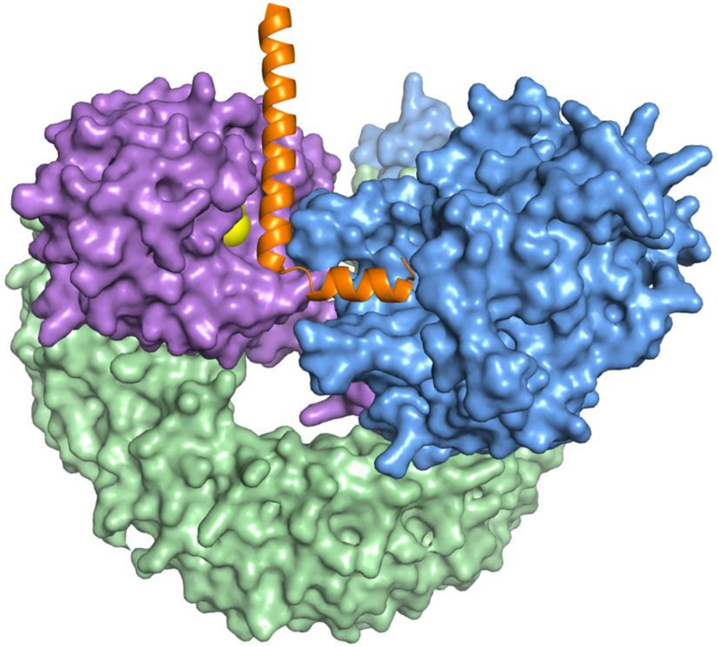
Trained on protein complexes from PDB

v2.1 in Nov 2021
Training data: 2018-4-30

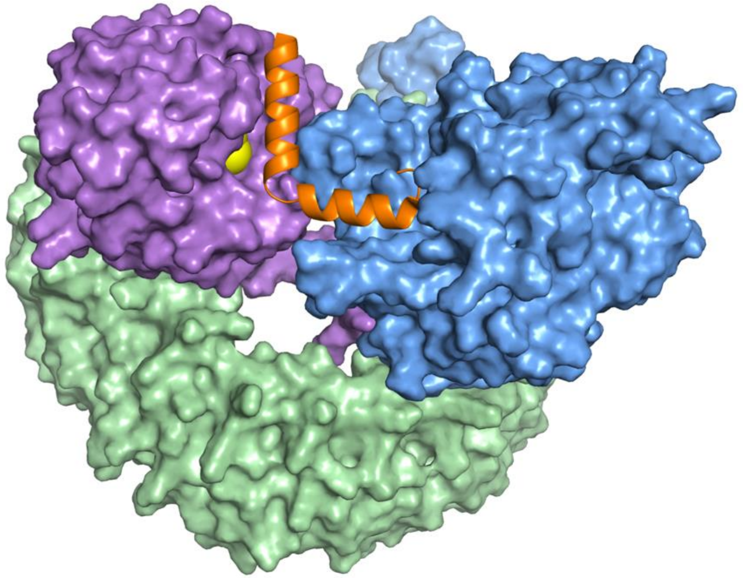
v2.2 in March 2022
Training data: 2018-4-30

v2.3 in Dec 2022
Training data: 2021-09-30.

PP2A+scaffold+B55α+FAM122A
(Xavier Graña, Temple U)



Feb. 2023: AlphaFold-Multimer model of complex of PP2A(A), B55α, PP2A(C), and FAM122A



Oct. 2023: Cryo-EM structure (PDB 8SO0) of PP2A(A), B55α, PP2A(C), and FAM122A

AlphaFold2/AFM are really five models

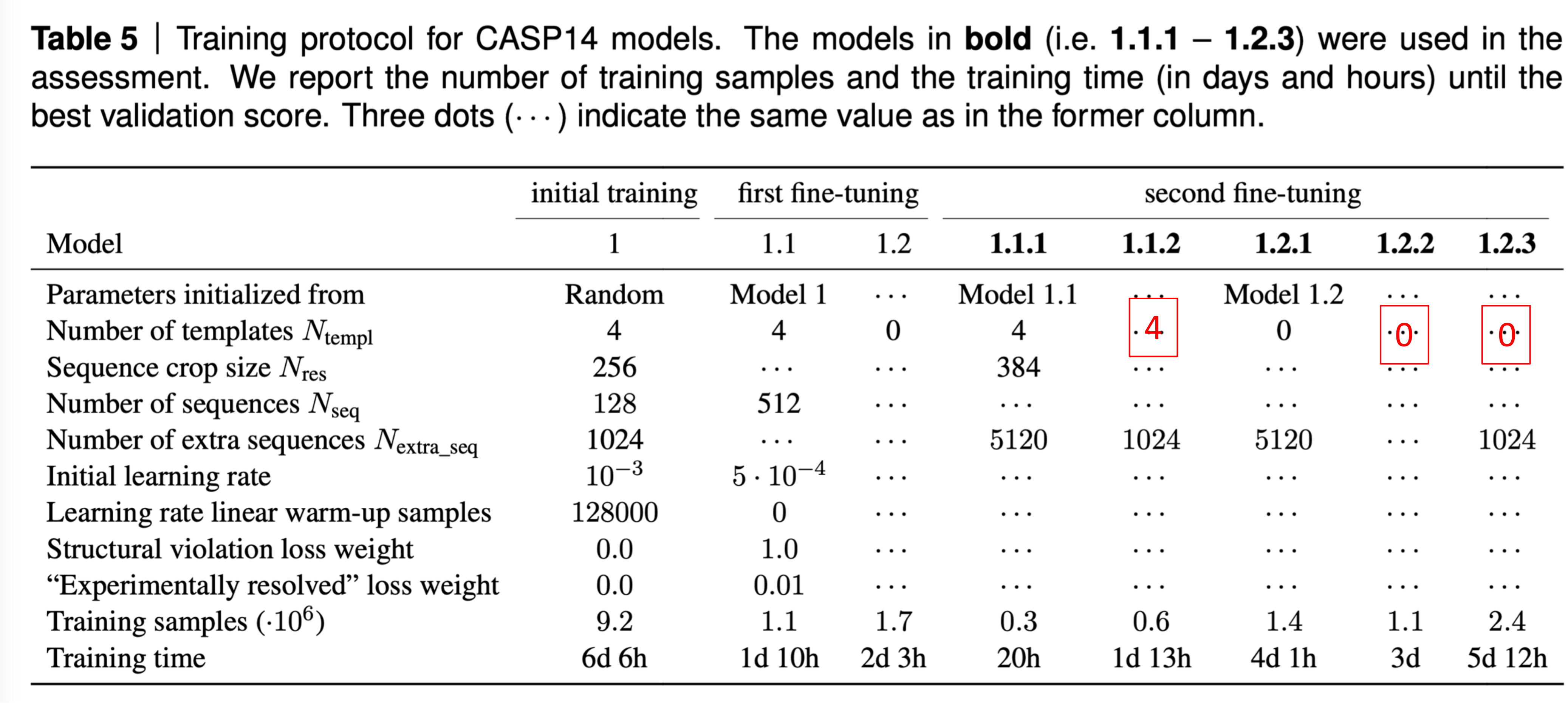


Table 5 | Training protocol for CASP14 models. The models in **bold** (i.e. **1.1.1** – **1.2.3**) were used in the assessment. We report the number of training samples and the training time (in days and hours) until the best validation score. Three dots (···) indicate the same value as in the former column.

	initial training	first fine-tuning		second fine-tuning				
Model	1	1.1	1.2	**1.1.1**	**1.1.2**	**1.2.1**	**1.2.2**	**1.2.3**
Parameters initialized from	Random	Model 1	···	Model 1.1	···	Model 1.2	···	···
Number of templates N_{templ}	4	4	0	4	4	0	0	0
Sequence crop size N_{res}	256	···	···	384	···	···	···	···
Number of sequences N_{seq}	128	512	···	···	···	···	···	···
Number of extra sequences $N_{\text{extra\_seq}}$	1024	···	···	5120	1024	5120	···	1024
Initial learning rate	10^{-3}	$5 \cdot 10^{-4}$	···	···	···	···	···	···
Learning rate linear warm-up samples	128000	0	···	···	···	···	···	···
Structural violation loss weight	0.0	1.0	···	···	···	···	···	···
"Experimentally resolved" loss weight	0.0	0.01	···	···	···	···	···	···
Training samples ($\cdot 10^6$)	9.2	1.1	1.7	0.3	0.6	1.4	1.1	2.4
Training time	6d 6h	1d 10h	2d 3h	20h	1d 13h	4d 1h	3d	5d 12h

Using templates:

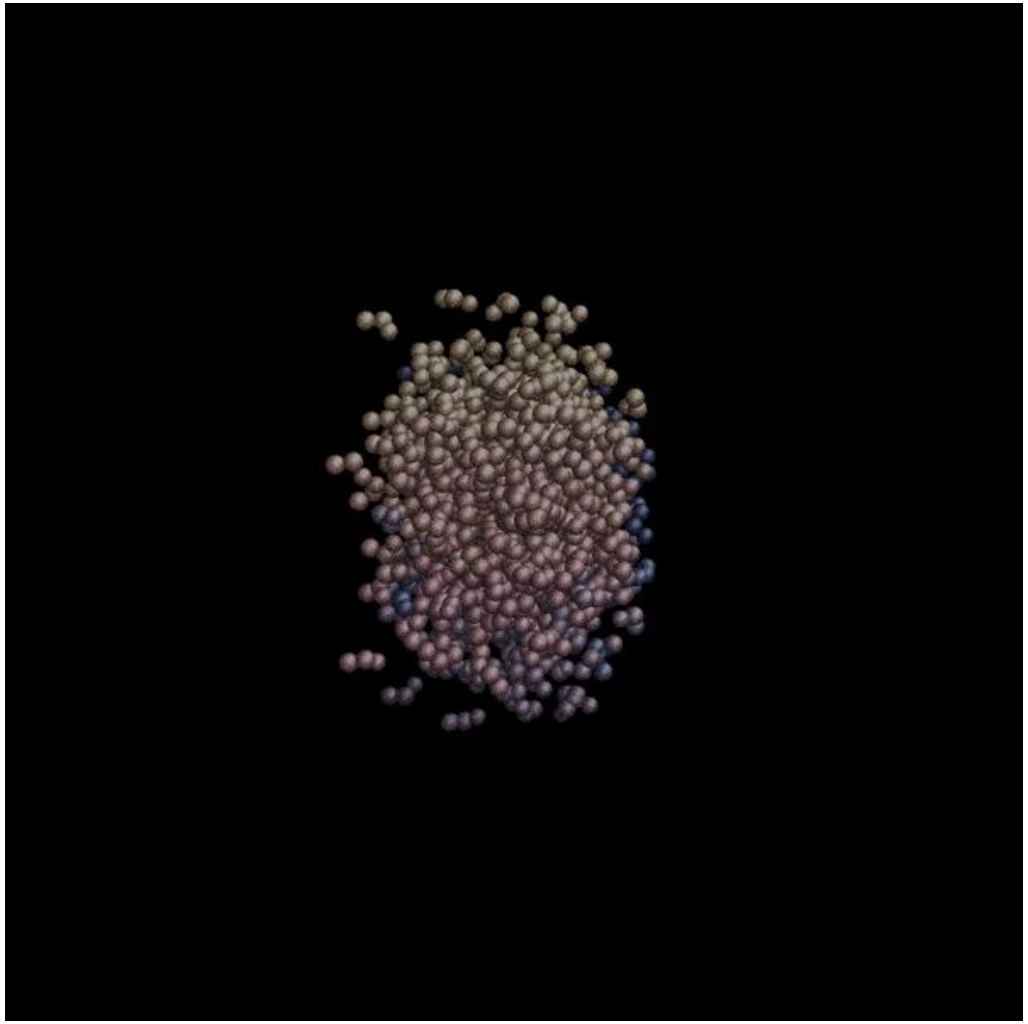
ColabFold AF2 Model# 1 2 3 4 5

ColabFold AFM2.3 Model# 1 2 3 4 5

AlphaFold3 – Diffusion based structure prediction

(Server: May 8, 2024; Code Nov 4, 2024)

Training data: 2021-09-30.

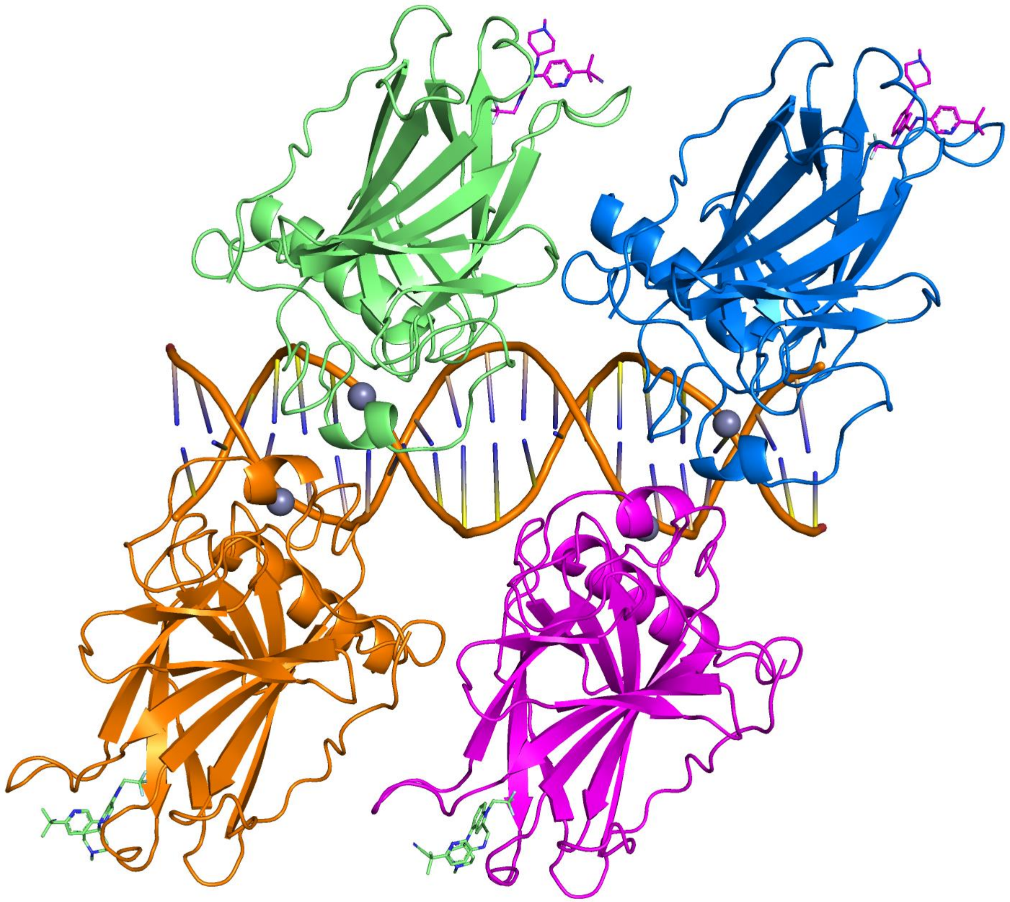


Random noise

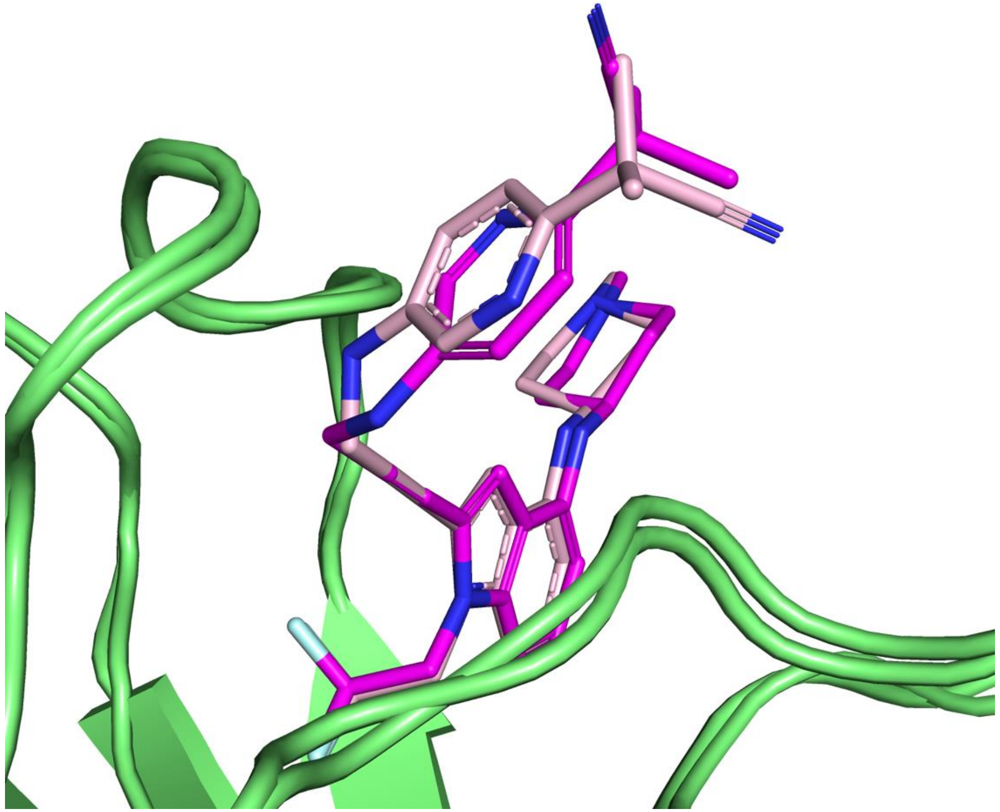
Denoising under informational constraints

AlphaFold3-like programs: AF3 (2024), Boltz2 (2025)

PTMs, DNA/RNA-protein complexes, cofactor-protein complexes, ligand-protein complexes



Boltz2: p53 tetramer bound to DNA and PC-9859



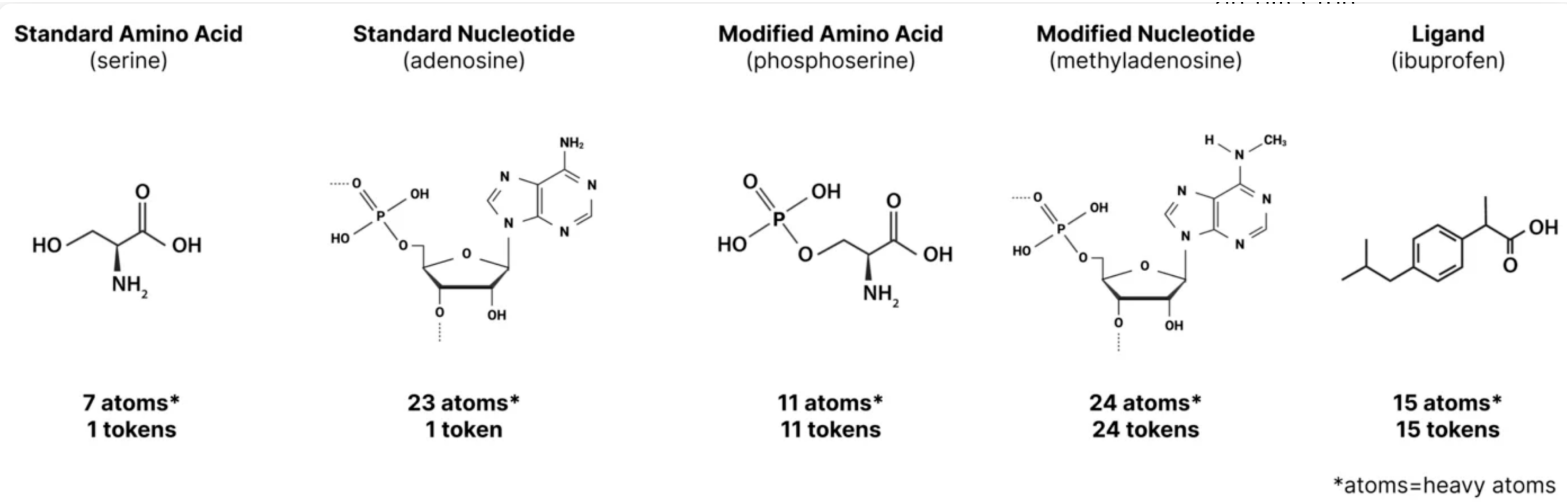
PDB:9BR4 (pink) vs Boltz2 (magenta)

Not in the training data for Boltz2

Hallucinating Alpha Helical Bundles in Disordered Regions

AlphaFold3
(human TCAB1 and coilin + RNA)

Tokens in AF3



Standard Amino Acid (serine)	Standard Nucleotide (adenosine)	Modified Amino Acid (phosphoserine)	Modified Nucleotide (methyladenosine)	Ligand (ibuprofen)
7 atoms* 1 tokens	23 atoms* 1 token	11 atoms* 11 tokens	24 atoms* 24 tokens	15 atoms* 15 tokens

*atoms=heavy atoms

Protein Structure Prediction Timeline, cont'd

Deep learning

Program	When	What	Code	Webserver
AlphaFold2	July 2021 (paper)	Monomers	Yes	Yes (Colab)
ColabFold (AF2)	August 2021	Monomers	Yes	Yes (Colab)
ColabFold-advanced	August 2021	Multimers	Yes	Yes (Colab)
AlphaFold-Multimer v2.1	Nov 2021	Multimers	Yes	Yes (ColabFold)
AlphaFold-Multimer v2.2	Mar 2022	Multimers	Yes	Yes (ColabFold)
AlphaFold-Multimer v2.3	Dec 2022	Multimers	Yes	Yes (ColabFold)
OpenFold (AF2)	May 2024	Multimers	Yes	Yes (Colab)
AlphaFold3	May 2024	Multimers+PTMs+ligands+DNA/RNA	Yes**	Yes
HelixFold3 (AF3)	Sept 2024	Multimers+ligands+DNA/RNA	Yes	Yes
Chai	Sept 2024	Multimers+PTMs+ligands+DNA/RNA	Yes	Yes
Boltz-2	June 2025	Multimers+PTMs+ligands+DNA/RNA	Yes	Several
OpenFold3	Oct 2025	Multimers+PTMs+ligands+DNA/RNA	Yes	Several
Protenix	Feb 2026	Multimers+PTMs+ligands+DNA/RNA	Yes	Yes
ESMfold2	May 2026	Multimers+PTMs+ligands+DNA/RNA	Yes	Several

Several = NVIDIA, TamarindBio, NeuroSnap, BioHub

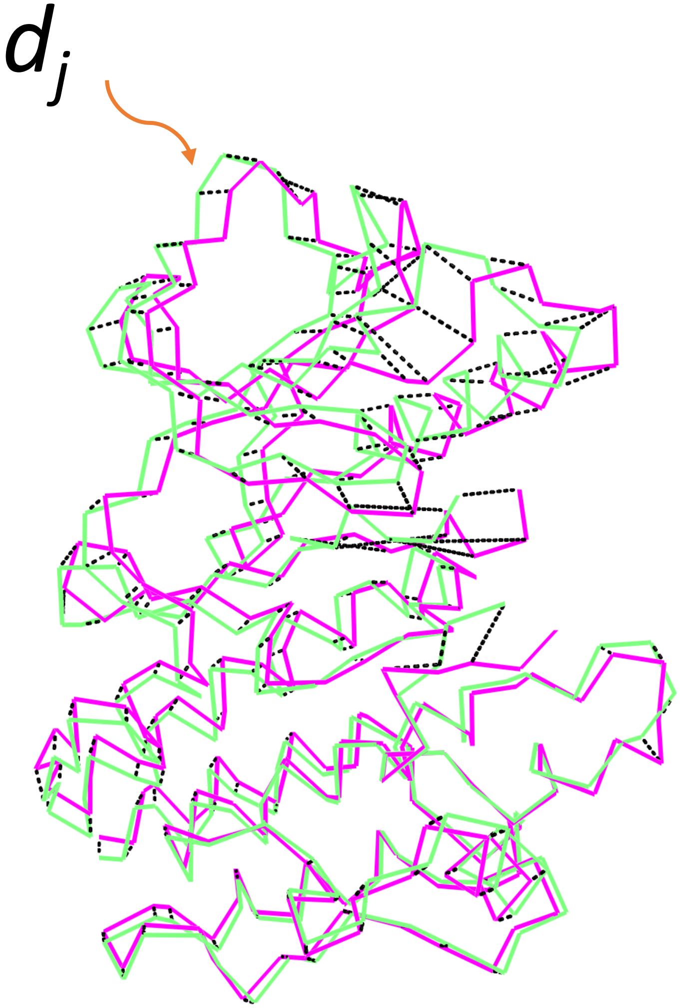
Measures used in CASP experiments used in deep learning predicted accuracy

Alignment-based: RMSD, GDT-TS, TM score

Non-alignment based: LDDT, side-chain dihedrals

RMSD: Root-mean-square-deviation

Depends on alignment and atom choice



$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{j=1}^{N} d_j^2}$$

j=1, N (length of protein)

GDT-TS –Global Distance Test

Align predicted structure to experimental structure, 4 times, with LGA program (Local-Global-Alignment):

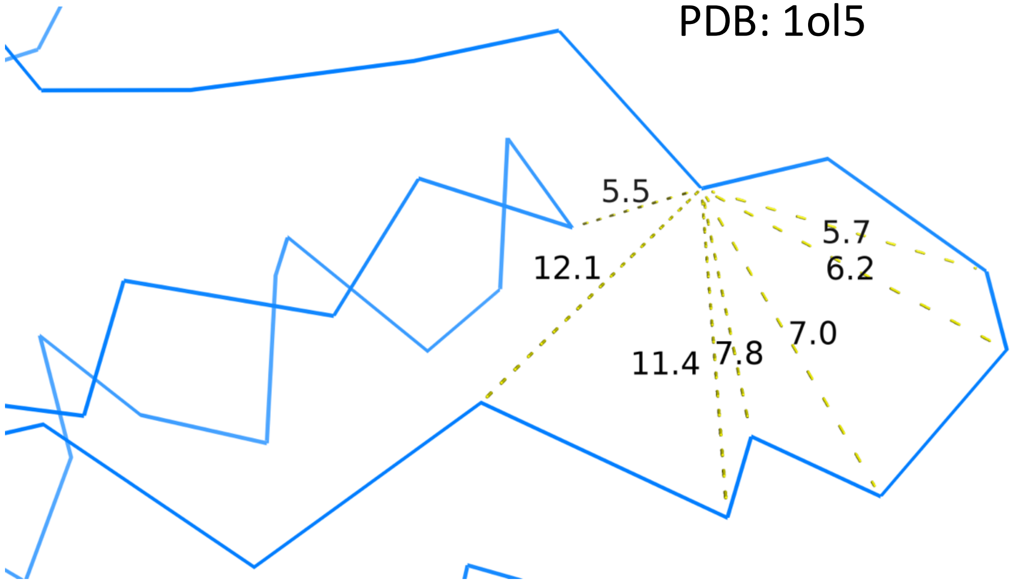
Maximizing # Cα < 1 Å → N_1
Maximizing # Cα < 2 Å → N_2
Maximizing # Cα < 4 Å → N_4
Maximizing # Cα < 8 Å → N_8

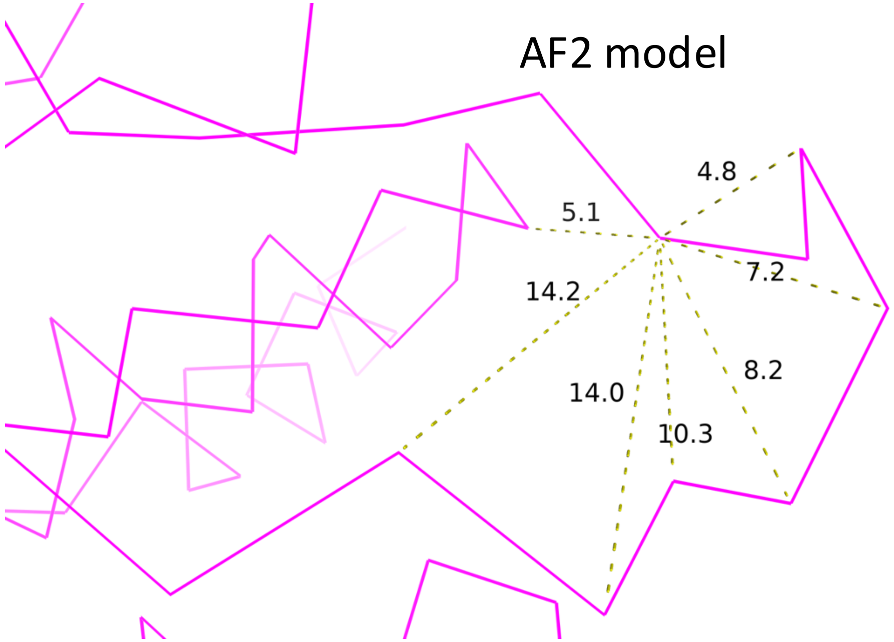
Calculate average of percentages of scored residues

$$\mathrm{GDT{-}TS} = 100\,\frac{(N_1/L + N_2/L + N_4/L + N_8/L)}{4}$$

LDDT – Local Distance Difference Test

Find Cα atoms within some distance of query Cα atom of experimental structure
Calculate distances in experimental structure and model
LDDT = for each residue: average percent of distances with |difference| ≤ 0.5, 1.0, 2.0, 4.0 Å



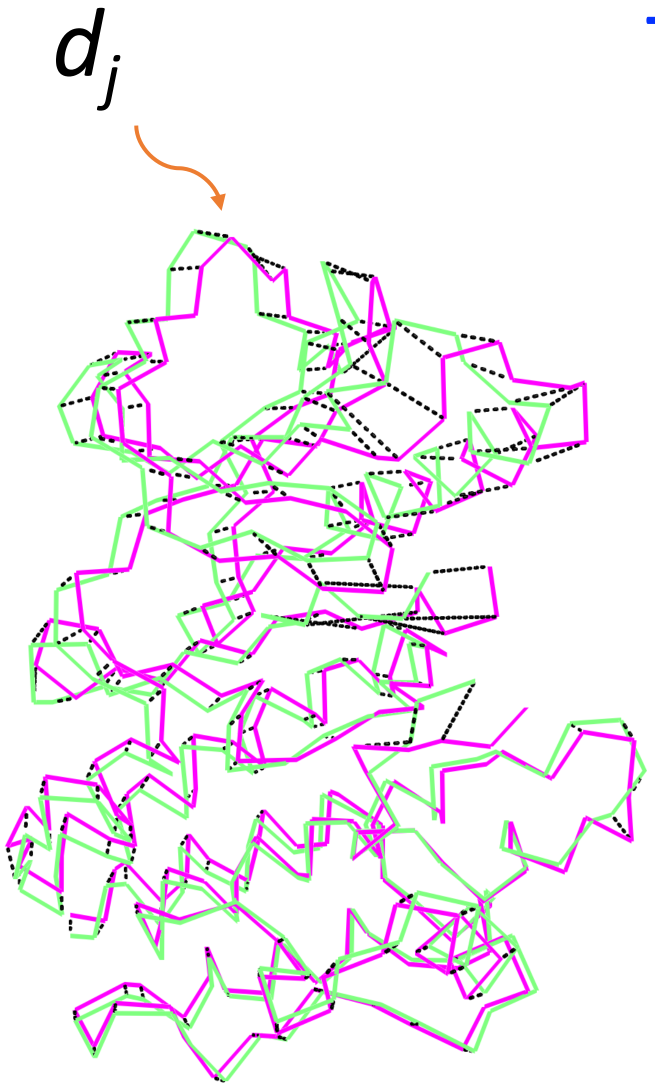


Differences = 0.9, 1.0, 1.2, 2.5, 2.6, 2.1, 0.4

LDDT =(1/7+3/7 +4/7 + 7/7)/4 = 54%

Template Modeling (TM) Score

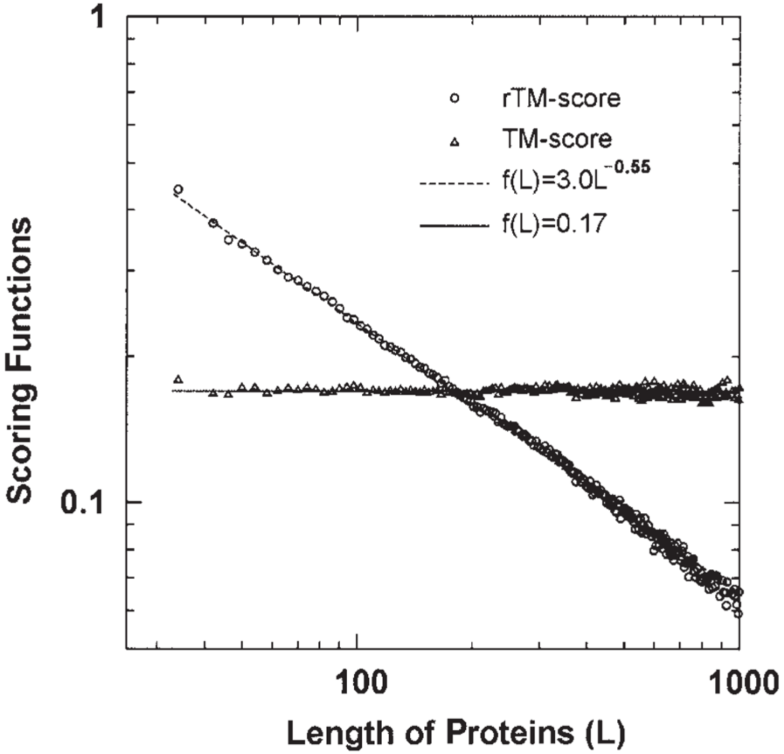
Zhang and Skolnick, 2004



$$TM = \max_{alignments}\left[\frac{1}{L_{target}}\sum_{j=1}^{L_{target}}\frac{1}{1+\left(\frac{d_j}{d_0}\right)^2}\right]$$

d_0 scales the d_j based on size of protein so unrelated proteins score the same regardless of size

$$d_0 = 1.24\sqrt[3]{L_{target}-15}-1.8$$

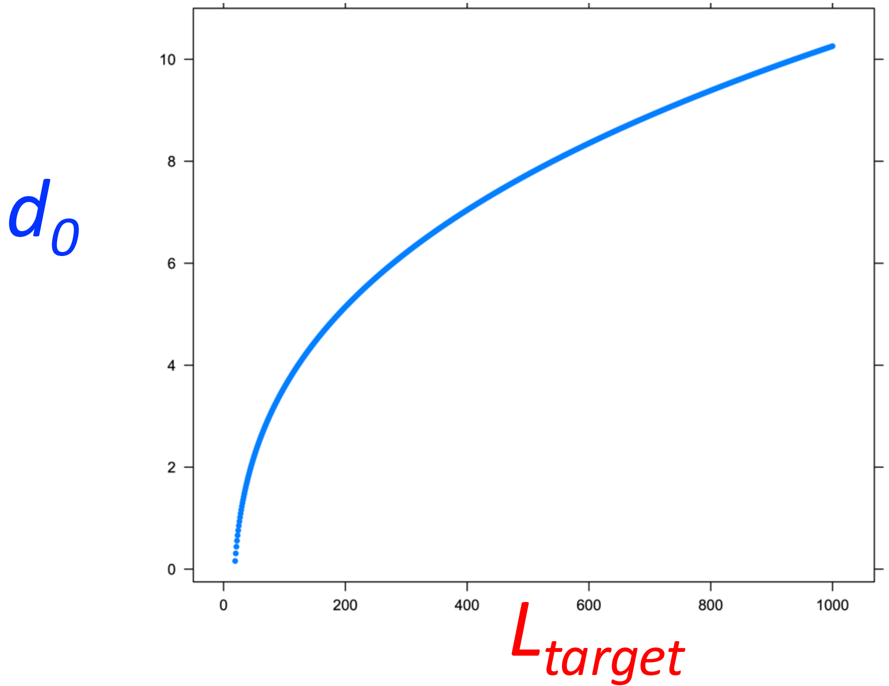


Original TM-score

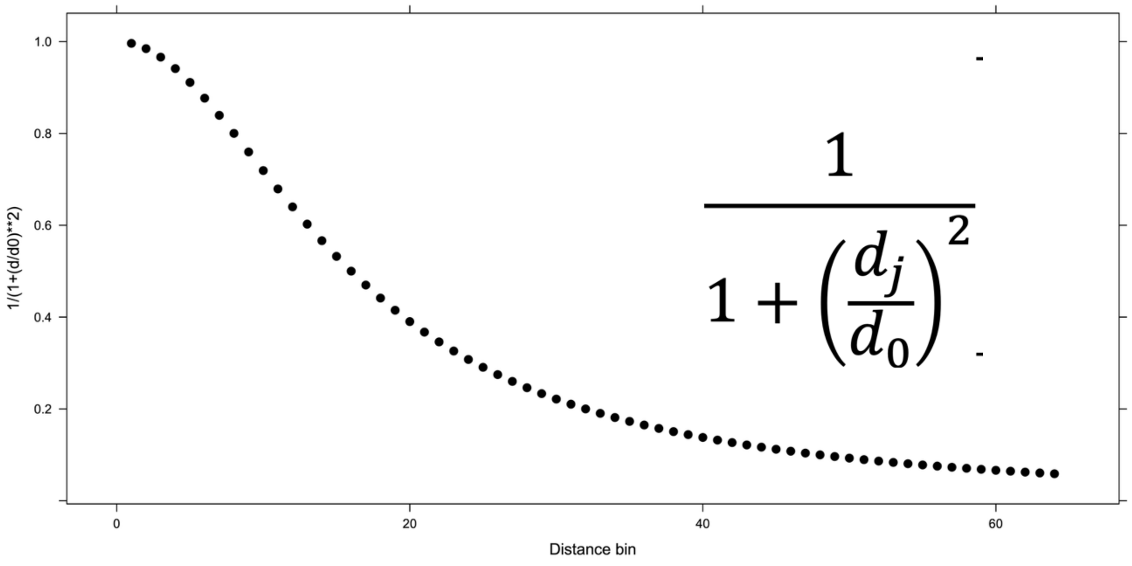
TM-score between a predicted and experimental structure of the same protein sequence is defined by:

$$TM = \max_{alignments}\left[\frac{1}{L_{target}}\sum_{j=1}^{L_{target}}\frac{1}{1+\left(\frac{d_j}{d_0}\right)^2}\right]$$

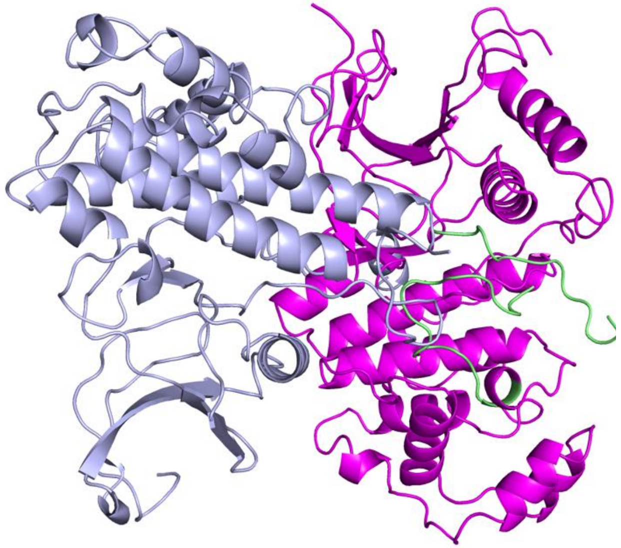
$$d_0 = 1.24\sqrt[3]{L_{target}-15}-1.8$$



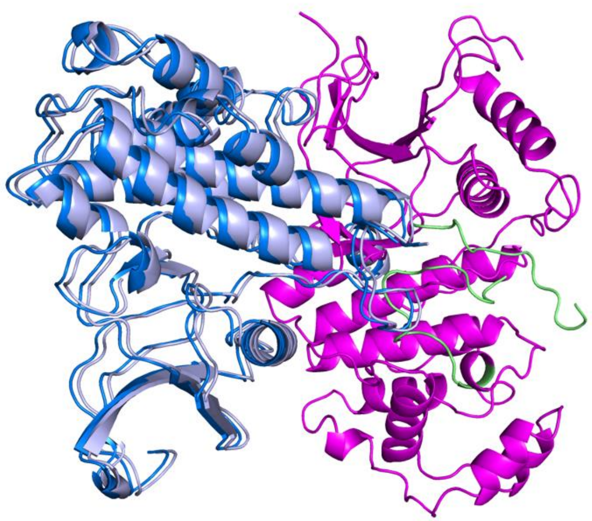
L_{target} = 500 residues



iTM: Align on one chain (ERK2) & calculate TM score over the other (RSK1)

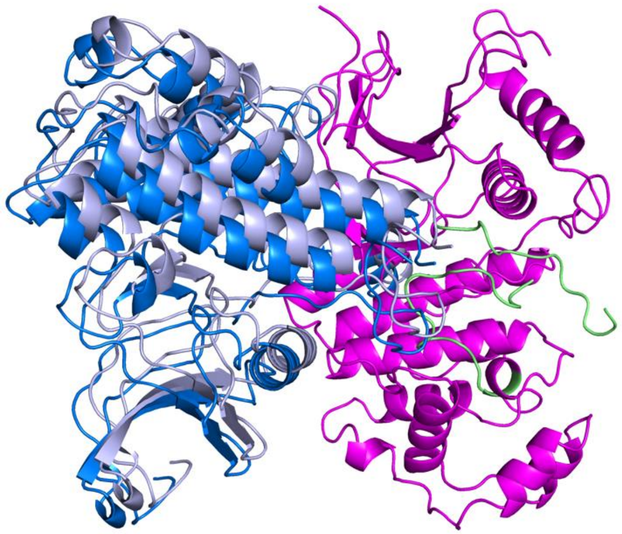


PDB: 4NIF



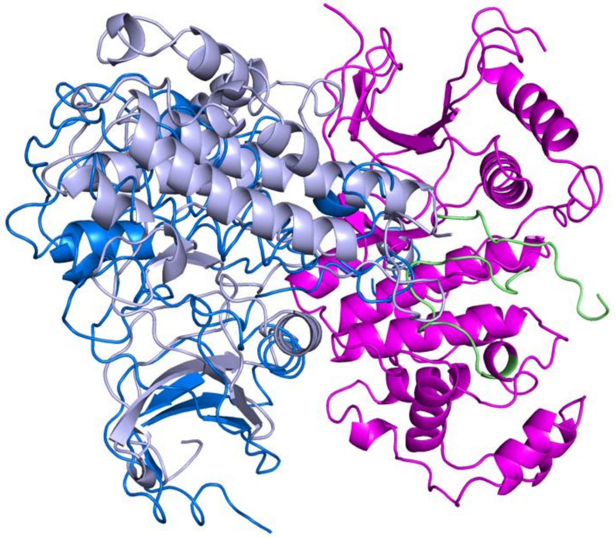
All $d_j = 1$ Å

iTM = 0.95



All $d_j = 3$ Å

iTM = 0.7

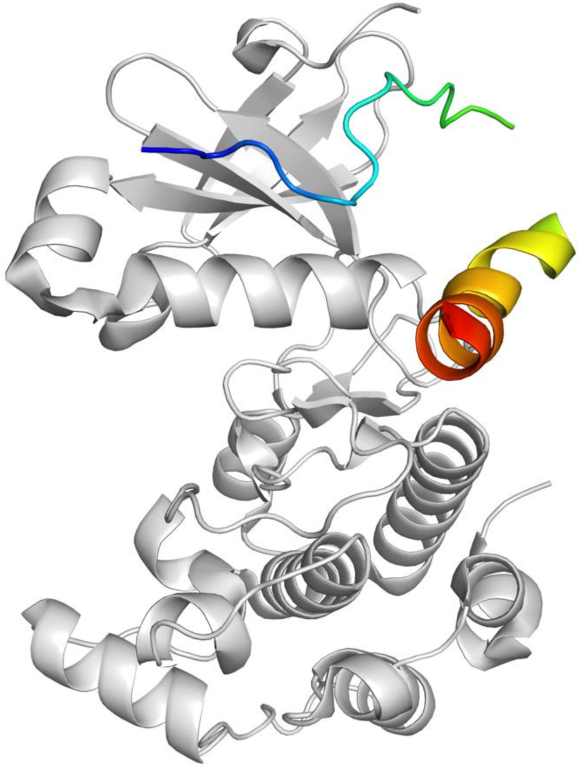


Poor model of RSK1

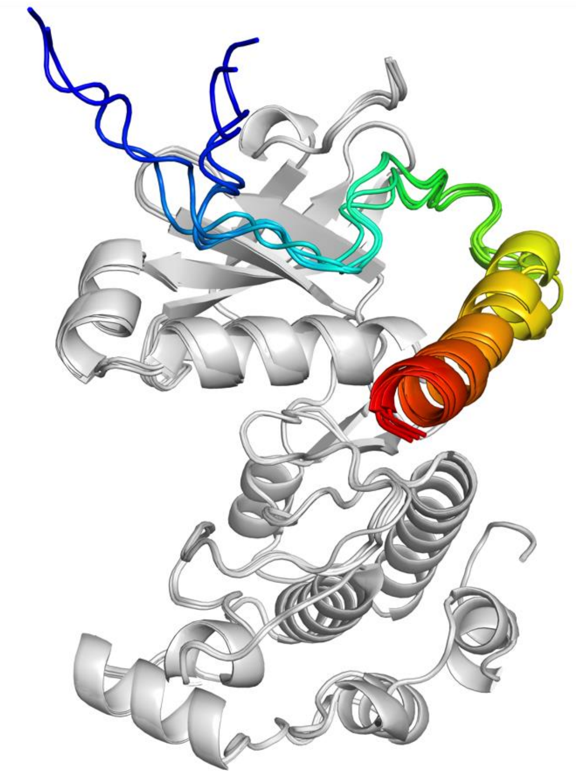
iTM = 0.5

DockQ for protein interfaces

DockQ≥0.25 considered an acceptable model



PDB 1OL5
AURKA+TPX2



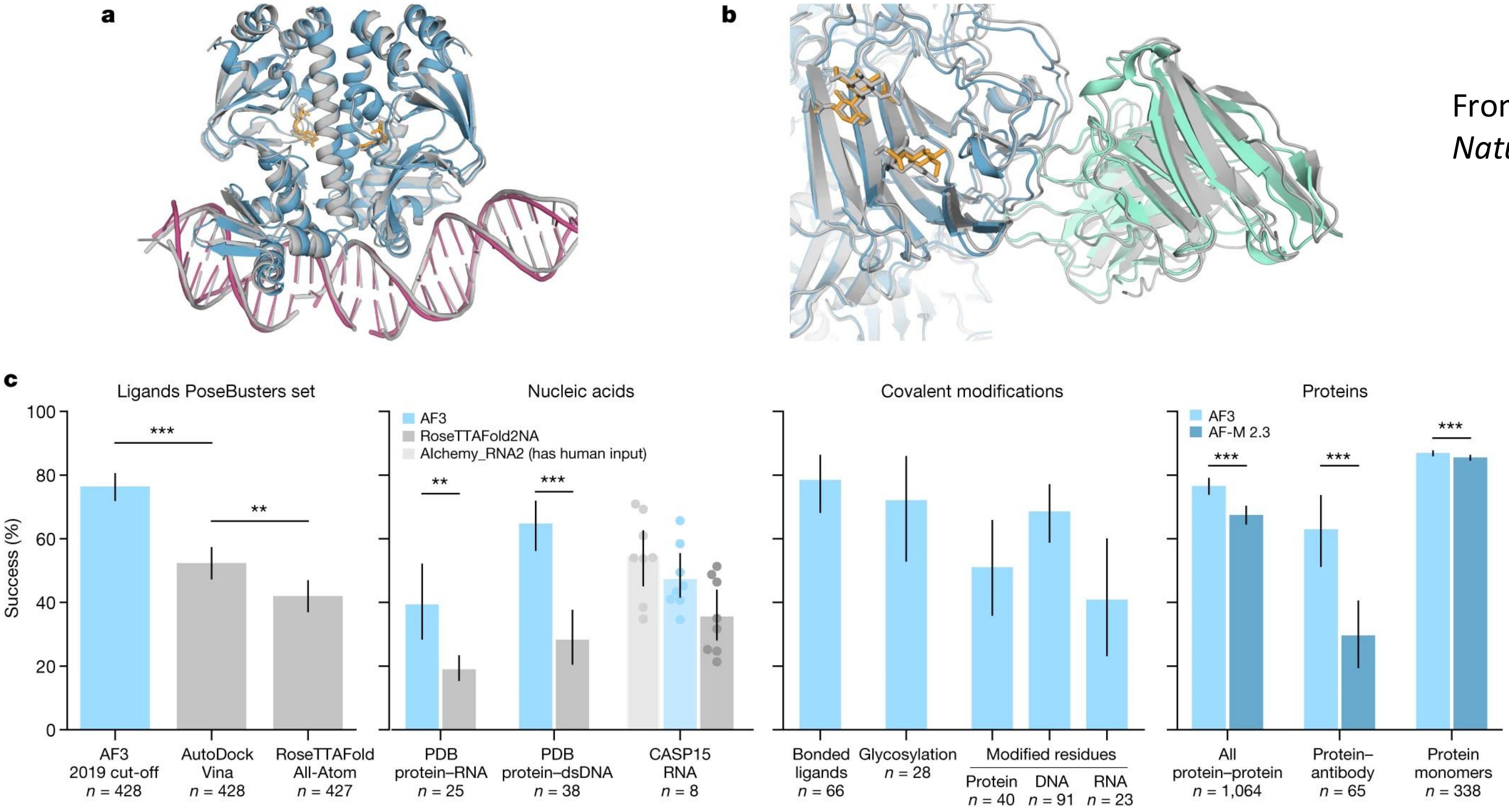
AlphaFold-Multimer
AURKA+TPX2

Calculate:

1. Fnat = fraction of native contacts in model
2. iRMS = RMSD of interface residues (after superposing on just those pairs)
3. LRMS = RMSD of ligand chain after superposing on receptor

$$\mathrm{DockQ} = \frac{1}{3}\left[F_{\mathrm{nat}} + \frac{1}{1+\left(\frac{\mathrm{iRMS}}{1.5}\right)^2} + \frac{1}{1+\left(\frac{\mathrm{LRMS}}{8.5}\right)^2}\right]$$

Benchmarking of Deep-Learning Structure Prediction



From AF3 paper,
Nature 2024

% pocket-aligned RMSD≤2.0 Å

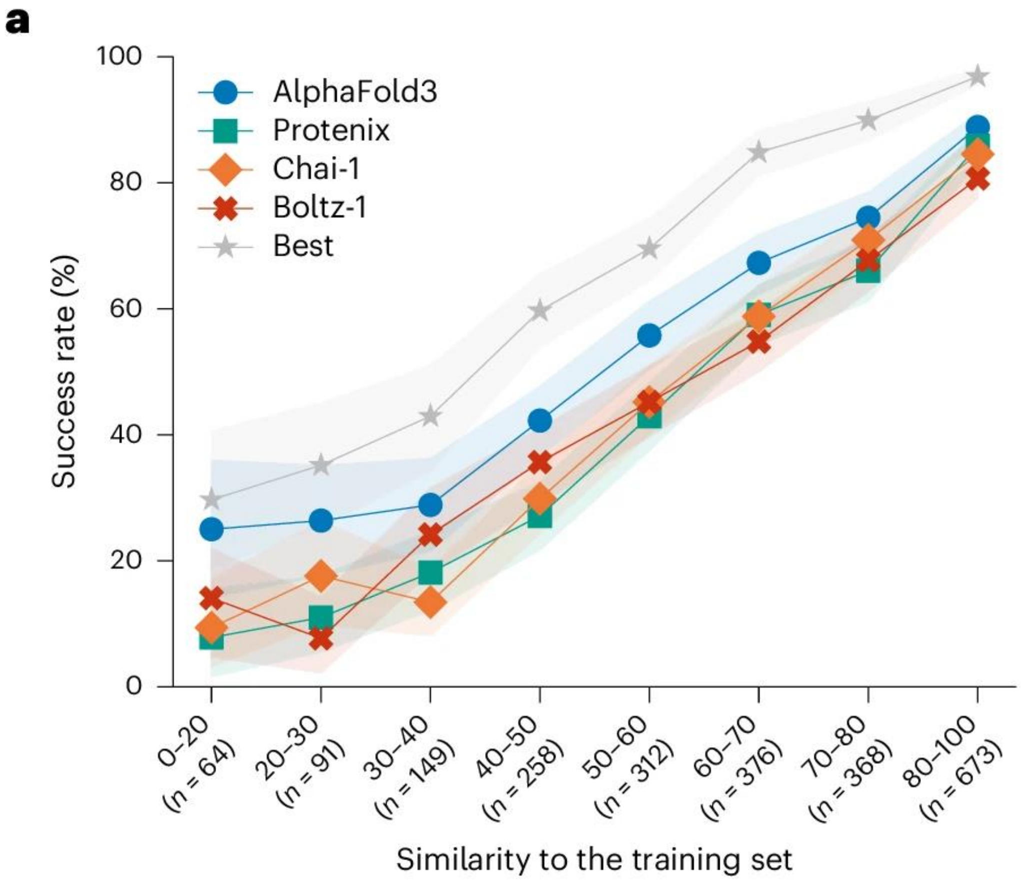
interfaceLDDT

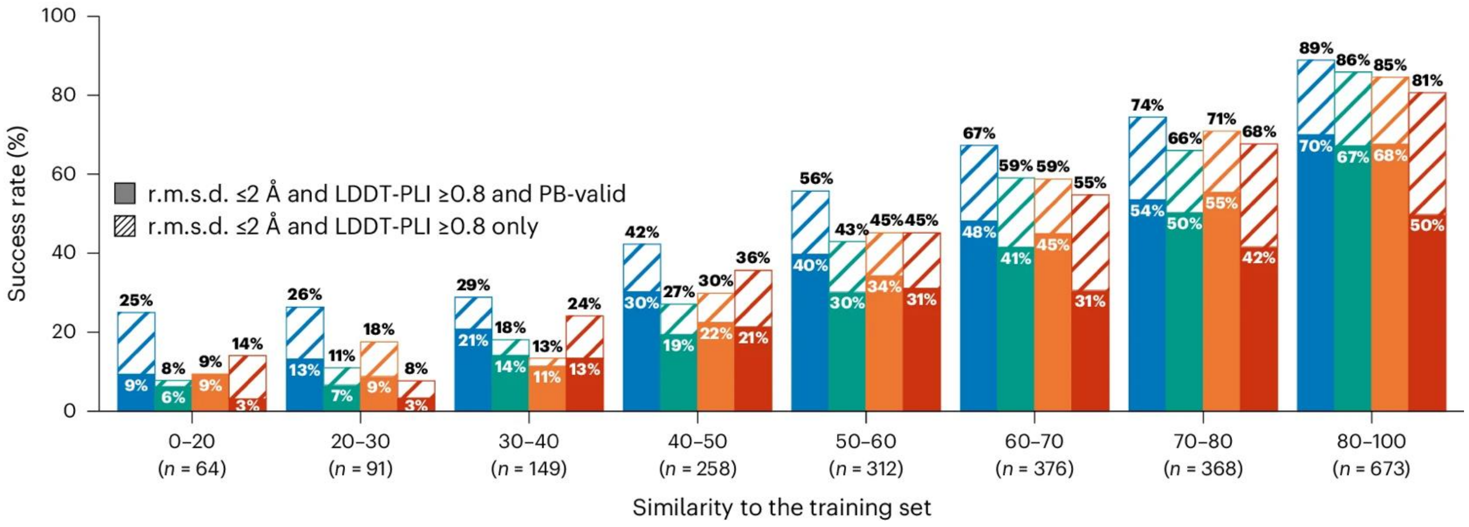
% pocket-aligned RMSD≤2.0 Å

% DockQ>0.23

Independent Benchmarking

Škrinjar et al, 2026





PB-valid:

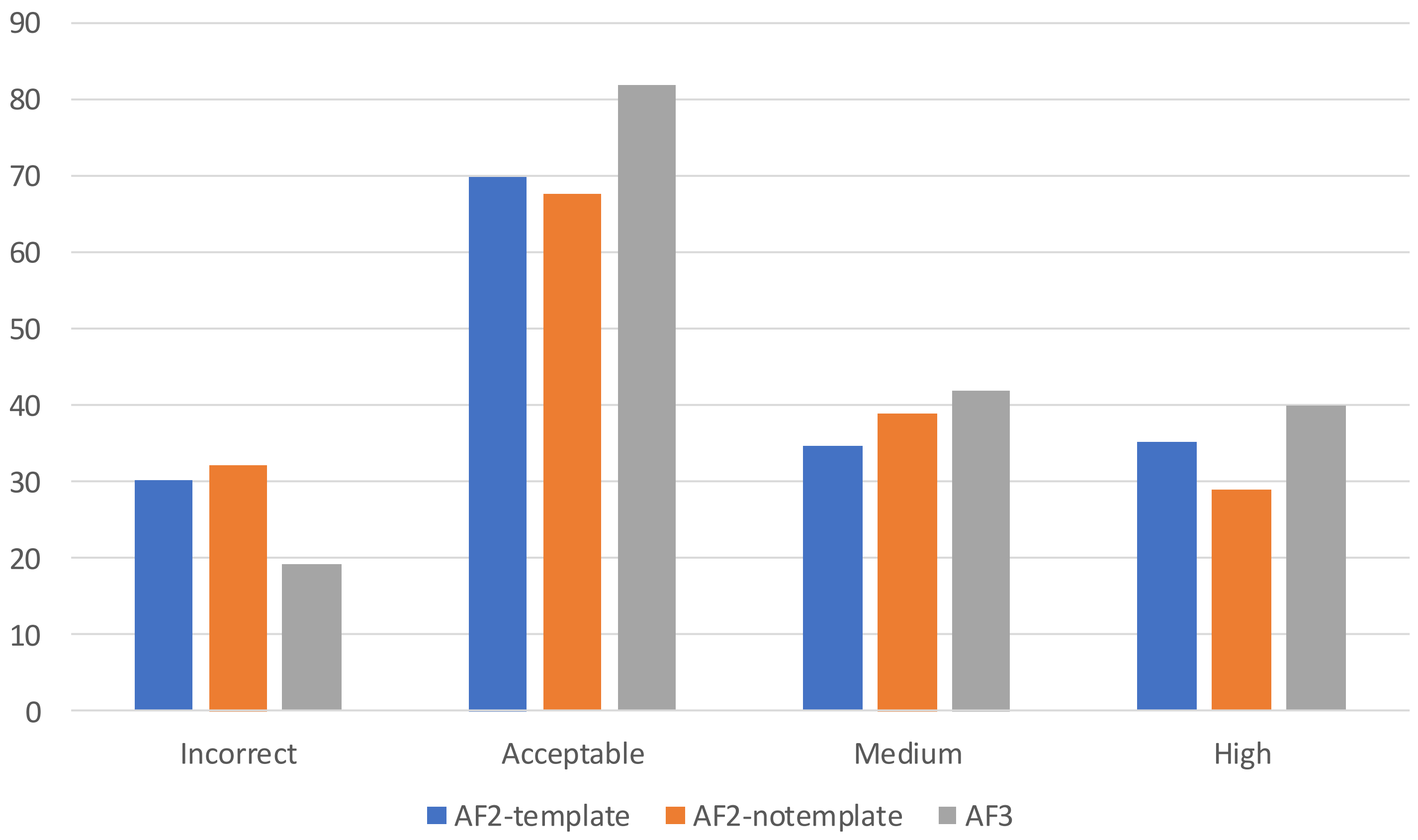
Few steric conflicts, correct geometries, flat rings, etc.

Success: Percent of test ligands with:

RMSD≤2 Å

LDDT-PLI<0.8 (PLI = protein/ligand interface

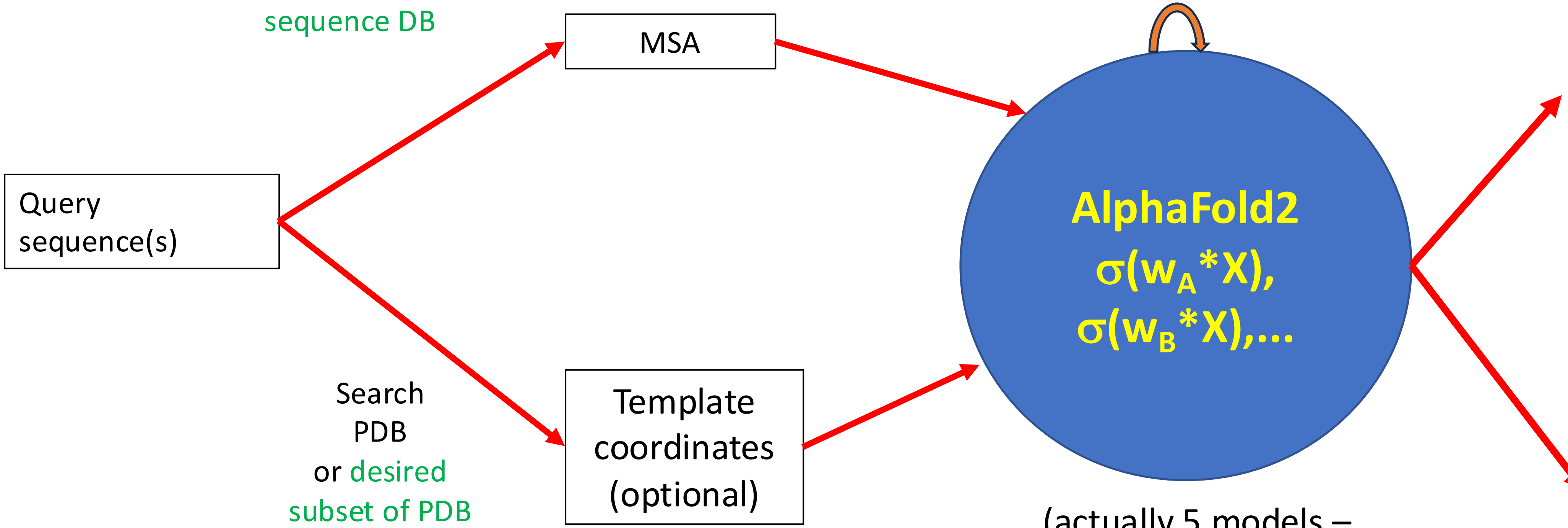
Benchmarking Protein-Protein Interactions



Cutoffs:
DockQ≥0.23 = acceptable
DockQ≥0.49 = medium quality
DockQ≥0.80 = high quality

Genz et al
Protein Sci
2025

AF2 and AF3 output scores: pLDDT, PAE, pTM, ipTM



Query sequence(s)

Search UniProt or other sequence DB

MSA

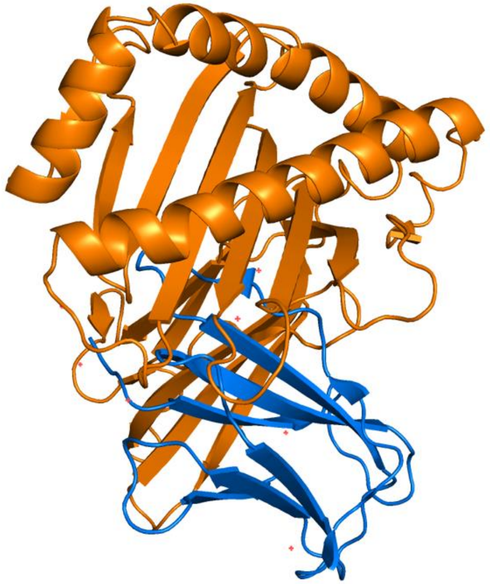
Search PDB or desired subset of PDB

Template coordinates (optional)

AlphaFold2
$\sigma(w_A * X)$, $\sigma(w_B * X)$,...

(actually 5 models – different sets of weights)

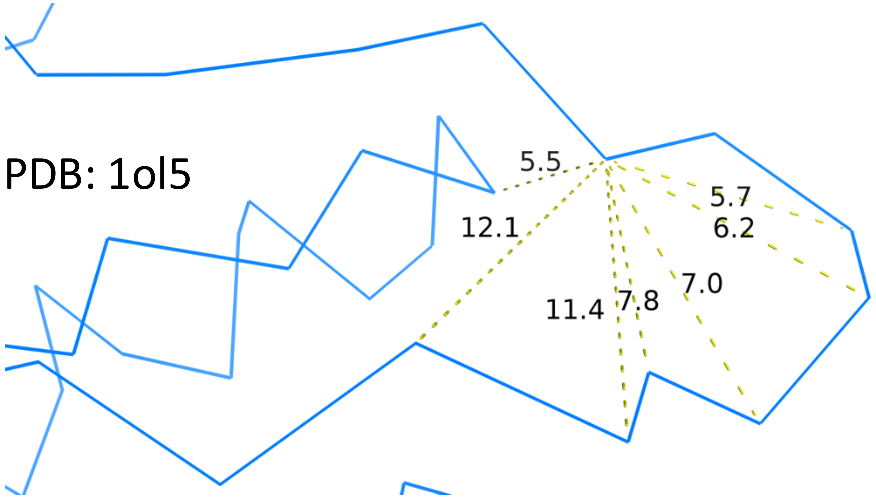
Run multiple times with different random “seeds”

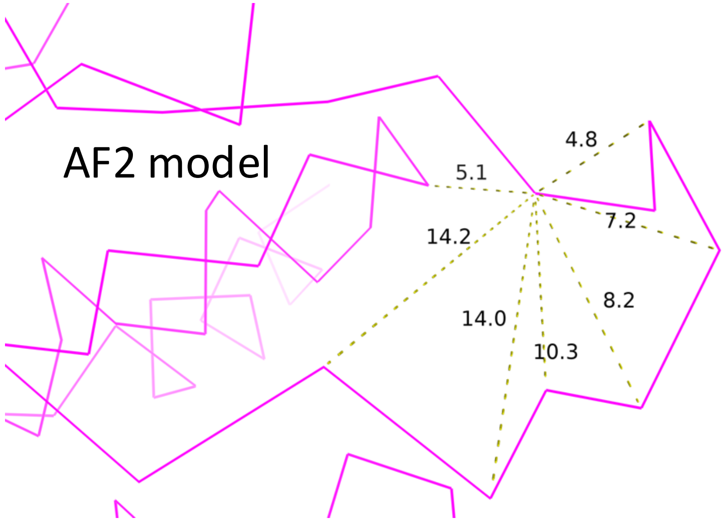


pLDDT (residues)
PAE (residue pairs)
pTM (whole struct)
iPTM (interactions)

pLDDT – predicted Local Distance Difference Test

pLDDT = for each residue: average percent of distances with |difference| ≤ 0.5, 1.0, 2.0, 4.0 Å





Diff = 0.9, 1.0, 1.2, 2.5, 2.6, 2.1, 0.4

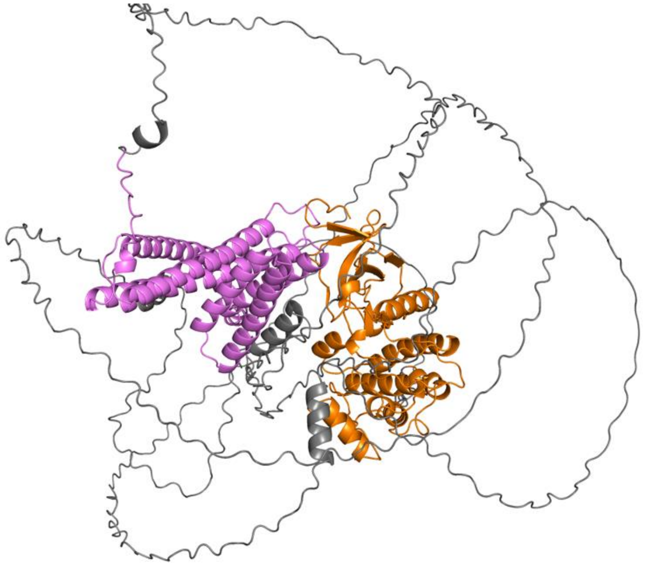
LDDT =(1/7+3/7 +4/7 + 7/7)/4
=54%

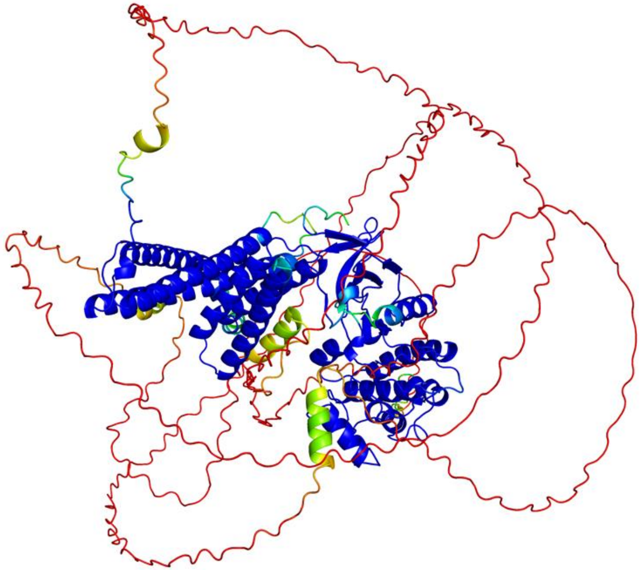
Colored by domains

pLDDT<50 in gray

PyMOL:

color gray,b<50

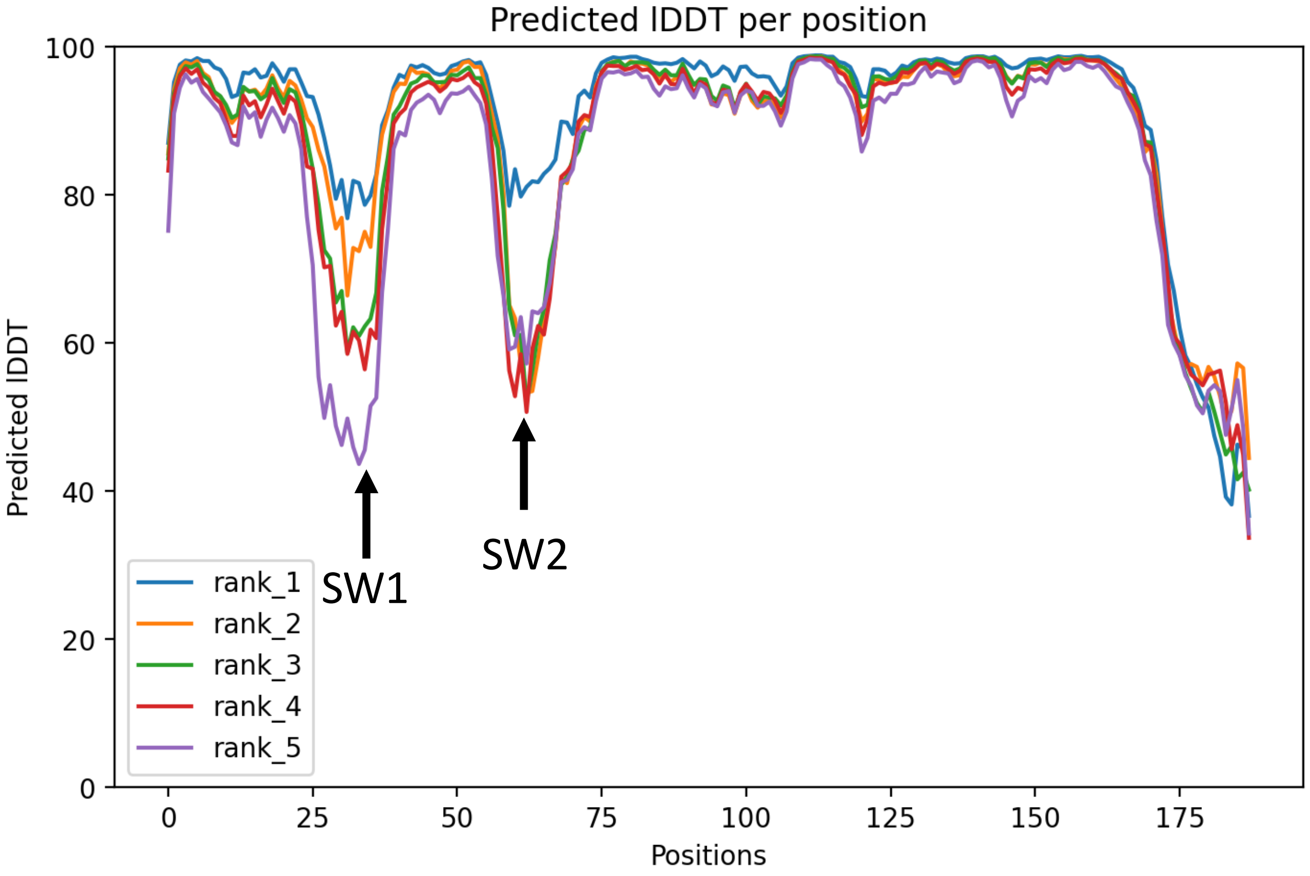


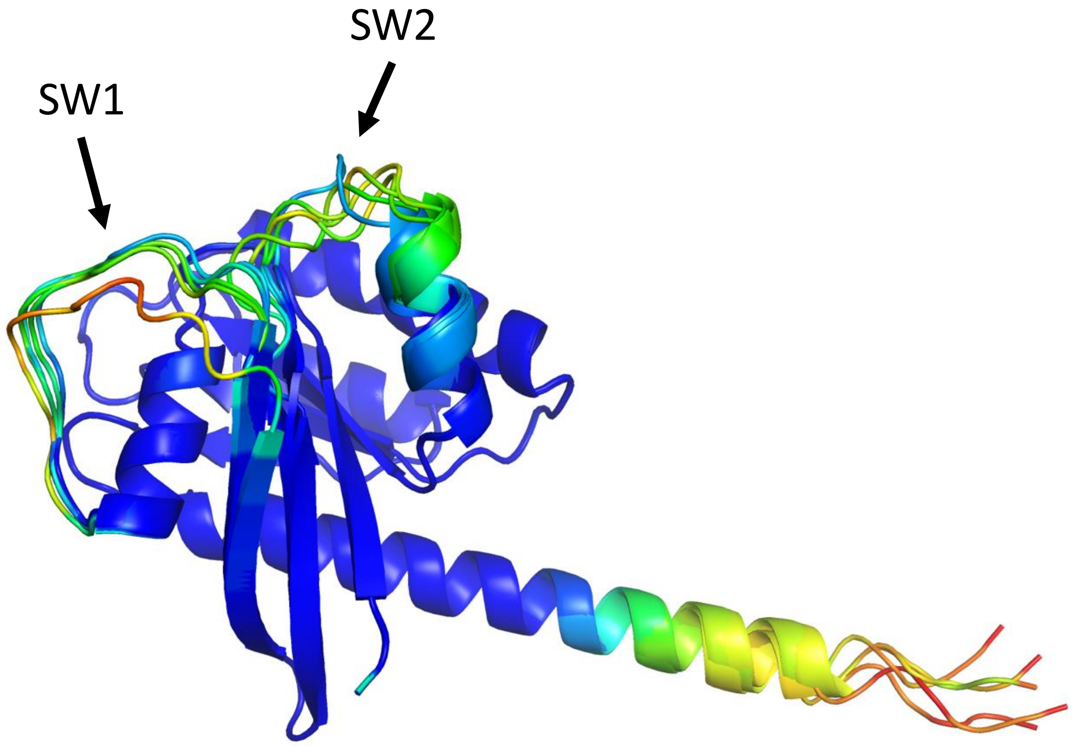


Colored by pLDDT (folded domains with high pLDDT in blue)

PyMOL: color by b-factor in color menu

pLDDT – (predicted) accuracy of atom placement in neighborhood around amino acid

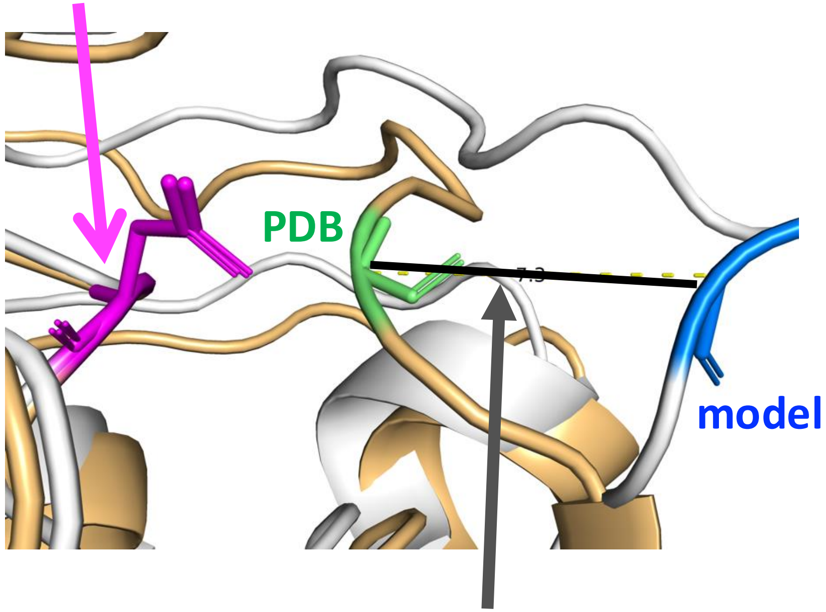




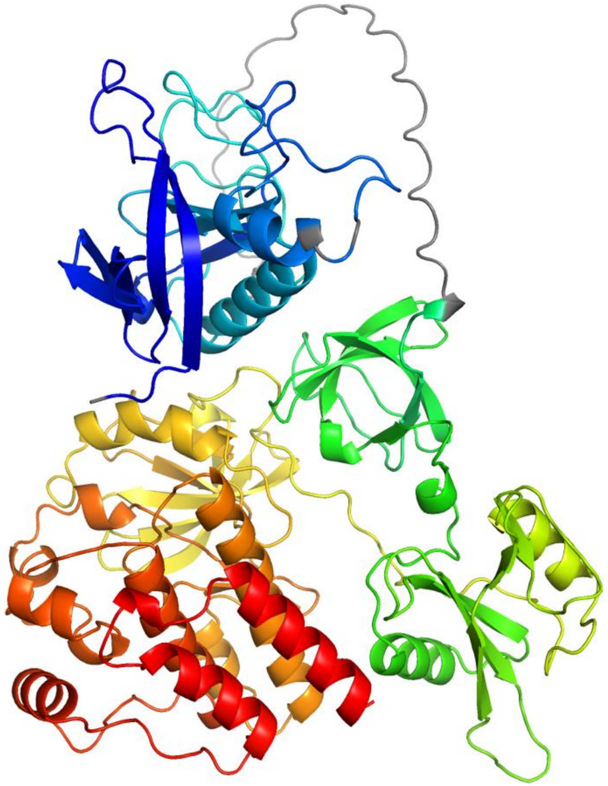
KRAS 4B

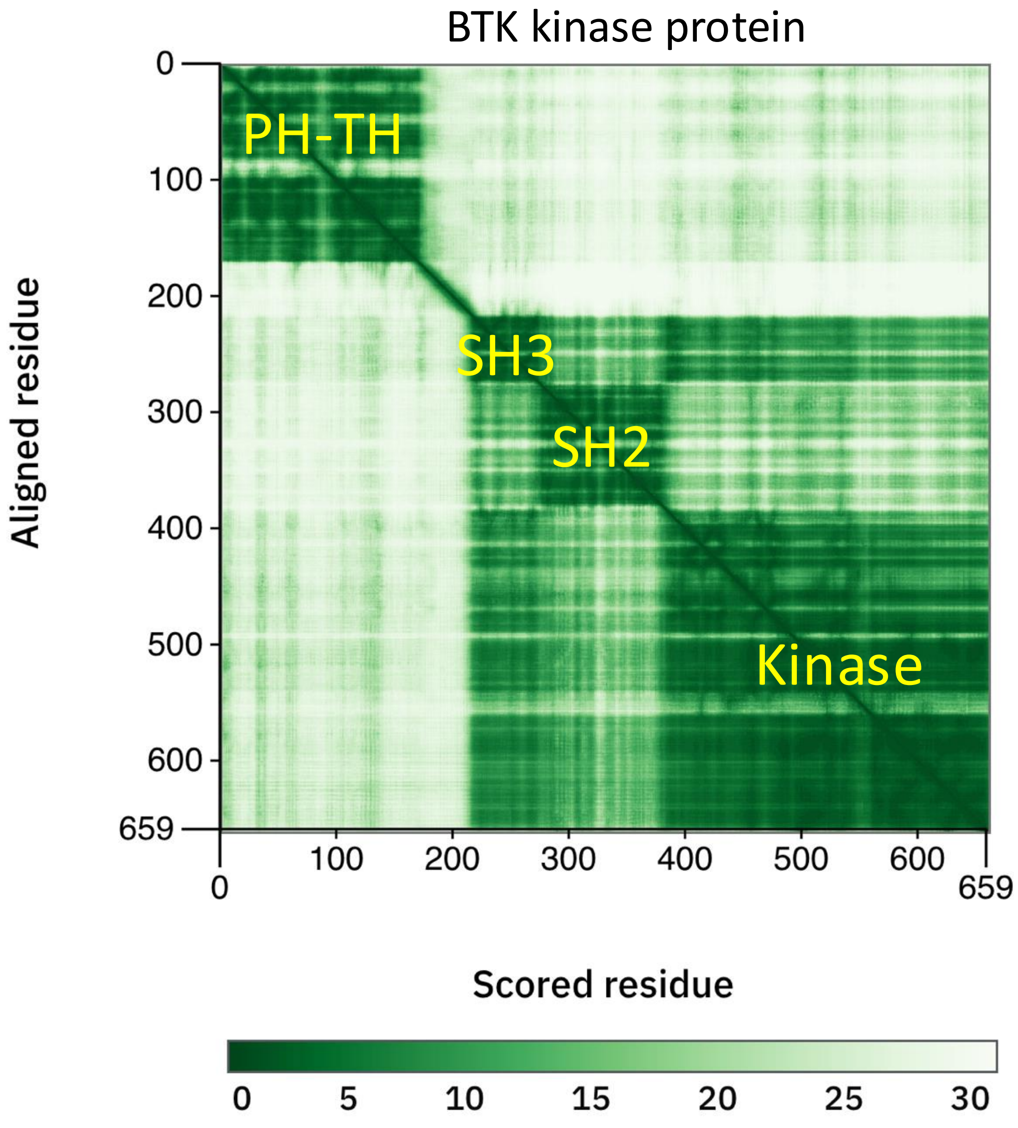
Predicted Aligned Error (PAE)

Aligned residue (N,Cα,C) in model and PDB



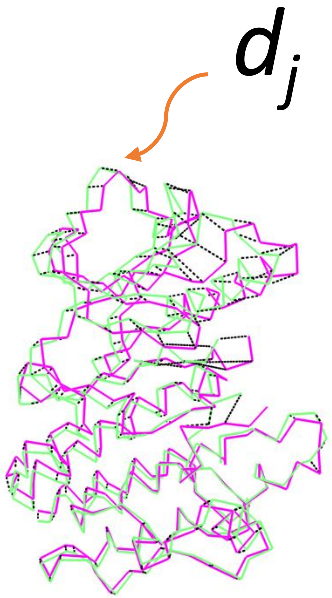
Scored residue PAE distance





pTM (predicted TM) for single chain or whole complex

d_j is just AlphaFold's PAE_{ij}



TM score

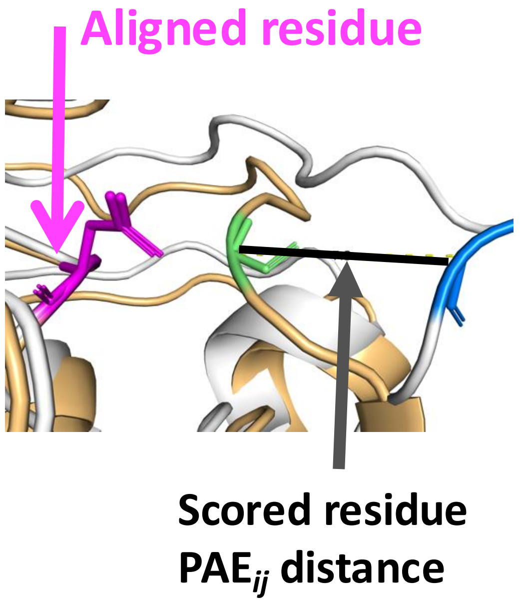
$$TM = \max_{alignments} \left[\frac{1}{L_{target}} \sum_{j=1}^{L_{target}} \frac{1}{1+\left(\frac{d_j}{d_0}\right)^2} \right]$$

pTM from *PAE_{ij}*

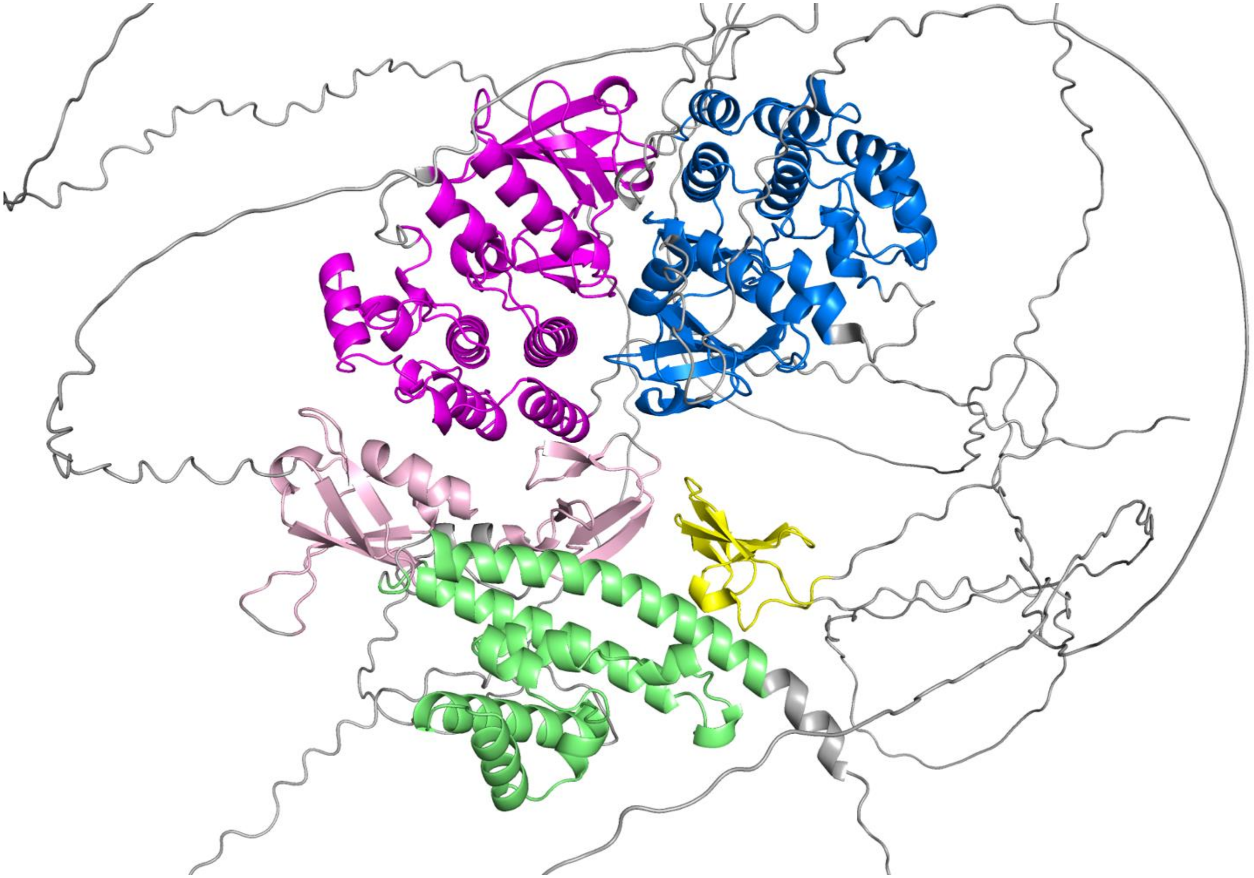
(AF uses probability distribution over PAE)

Think of *i* as a set of alignments, one for each residue

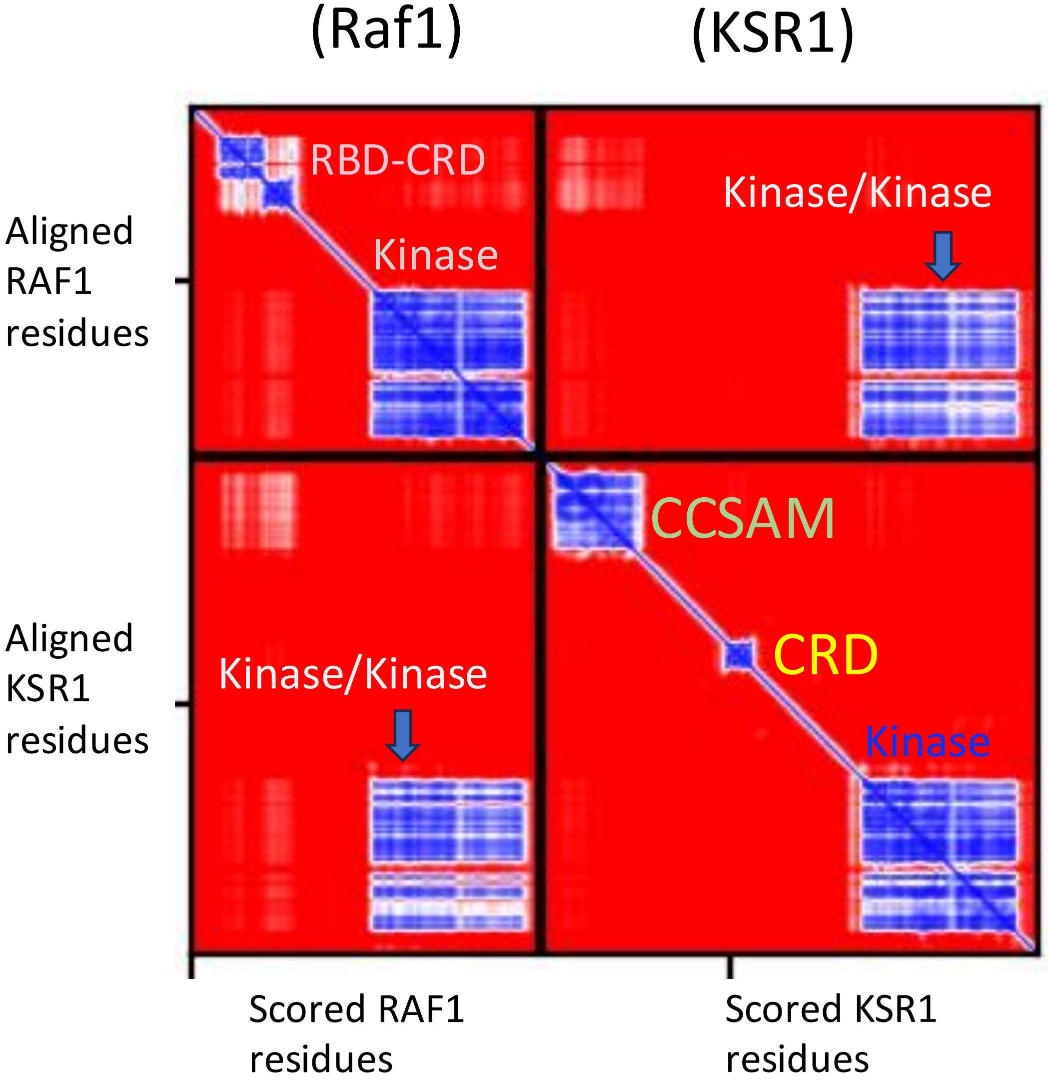
$$pTM = \max_i \left[\frac{1}{L} \sum_{j=1}^{L} \frac{1}{1+\left(\frac{PAE_{ij}}{d_0}\right)^2} \right]$$



Predicted Aligned Error (PAE) – protein complexes



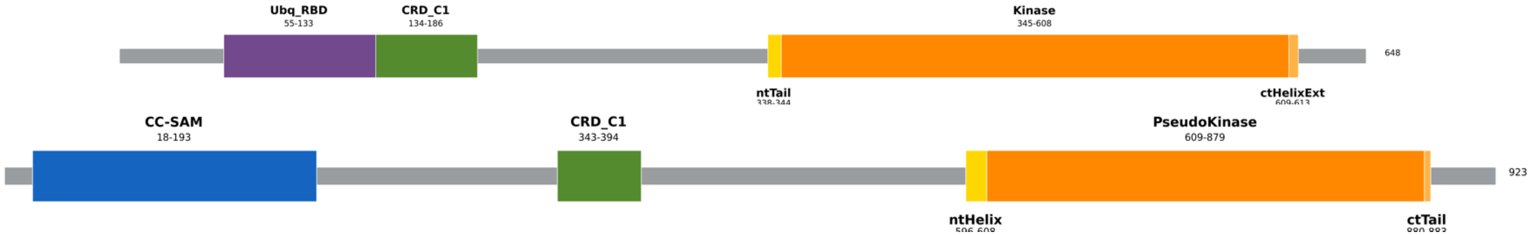
RAF1 (RBD-CRD, Kinase) and
KSR1 (CC-SAM, CRD, Kinase)



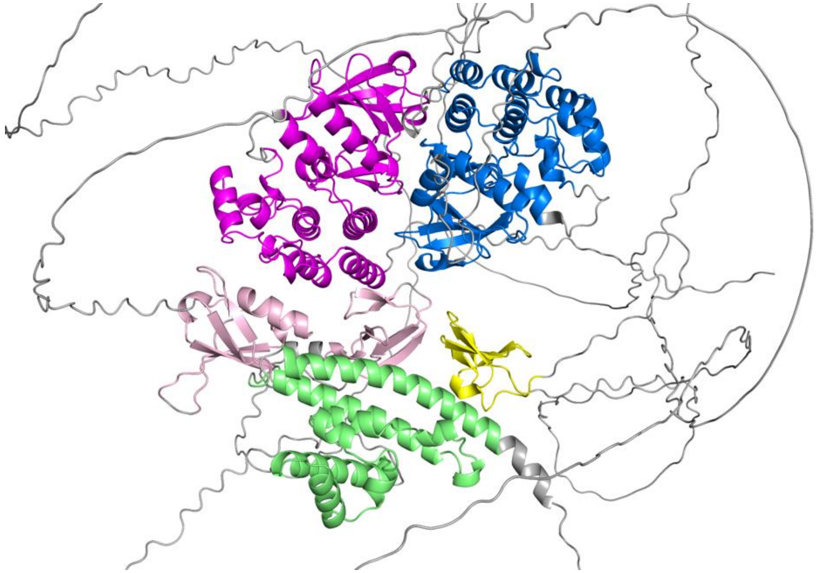
PAE: blue<8 Å, red>25 Å

Modeling PPI: AlphaFold’s *ipTM* score

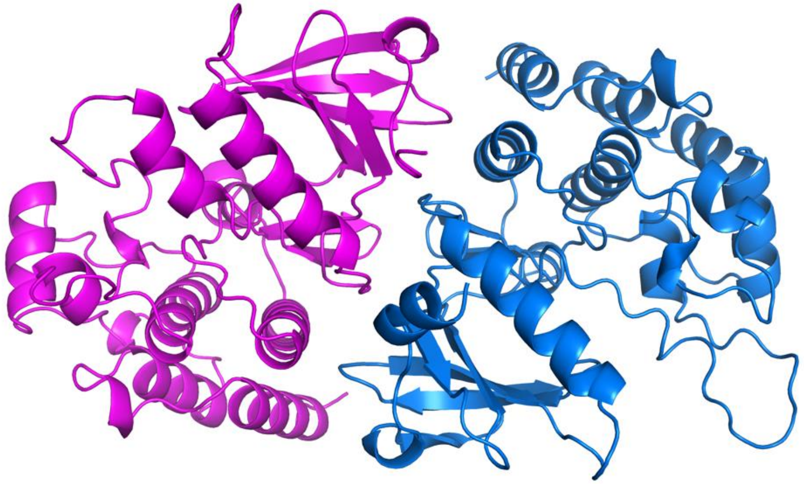
- Large constructs with disorder and lots of domains → low ipTM
- Make new constructs based on structure/PAE and run AF → high ipTM
- Why?



ipTM=0.38



RAF1 (RBD-CRD, Kinase) and
KSR1 (CC-SAM, CRD, Kinase)



ipTM=0.82

RAF1 (Kinase) and
KSR1 (Kinase)

ipTM: align on residues in one chain, score the other

Asymmetric value (*i*=alignments, *j*=scores):

$$\boxed{ipTM(A \to B) = \max_{i \in A}\left[\operatorname*{mean}_{j \in B}\left(\frac{1}{1+\left(\frac{PAE_{ij}}{d_0}\right)^2}\right)\right]}$$

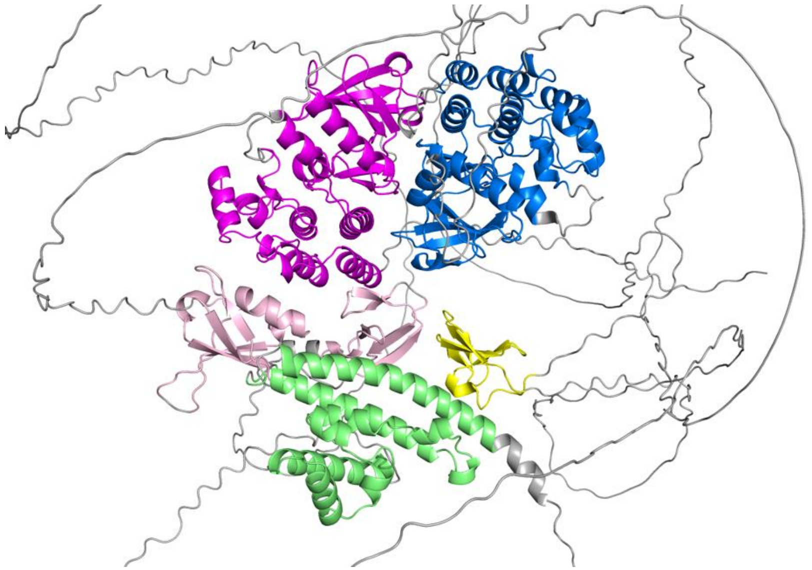
where $d_0 = 1.24\sqrt[3]{L_{target} - 15} - 1.8$ L_{target} is sum of chain lengths

Final *ipTM* value (max over residues in both chains):

$$\boxed{ipTM = \max[ipTM(A \to B), ipTM(B \to A)]}$$

Explains why disorder and extra domains lower *ipTM* *when present in both chains*

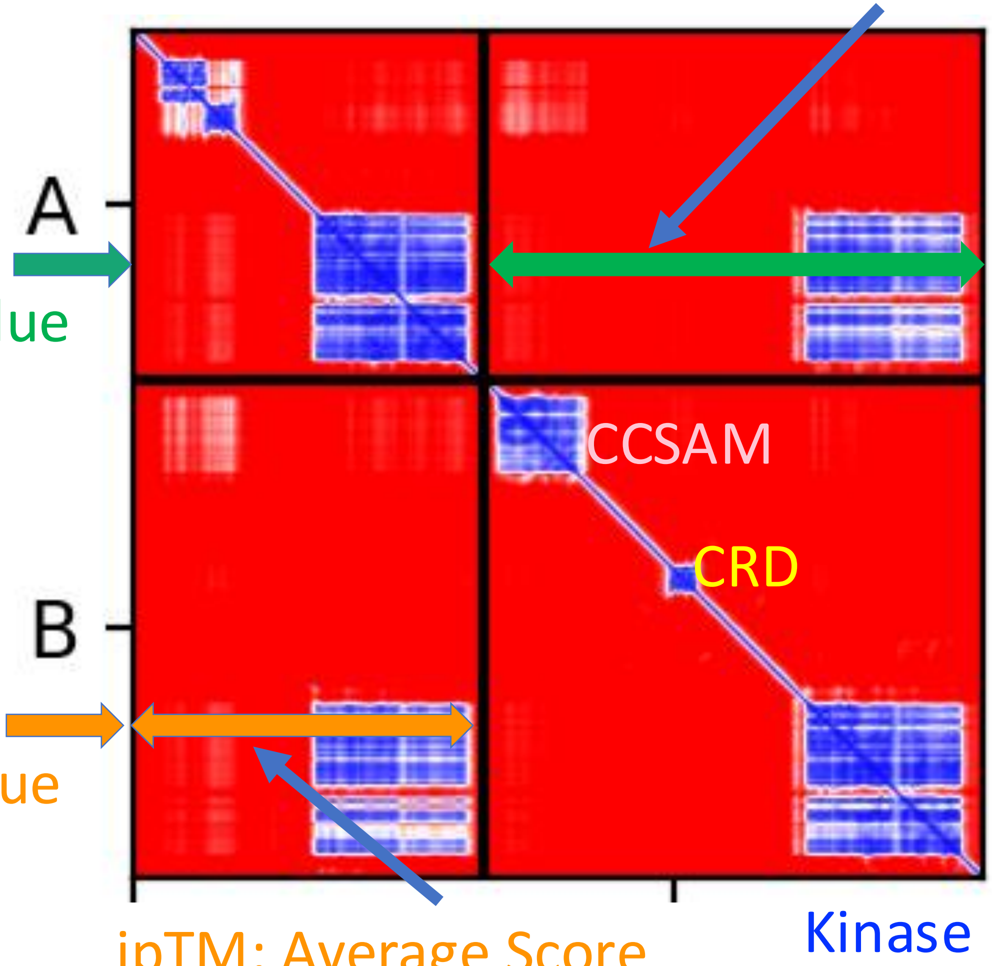
ipTM=0.41



Chain A: RAF1 (RBD-CRD, Kinase)
Chain B: KSR1 (CC-SAM, CRD, Kinase)

Align
chain A residue

ipTM: Average Score over all chain B residues



Align
chain B residue

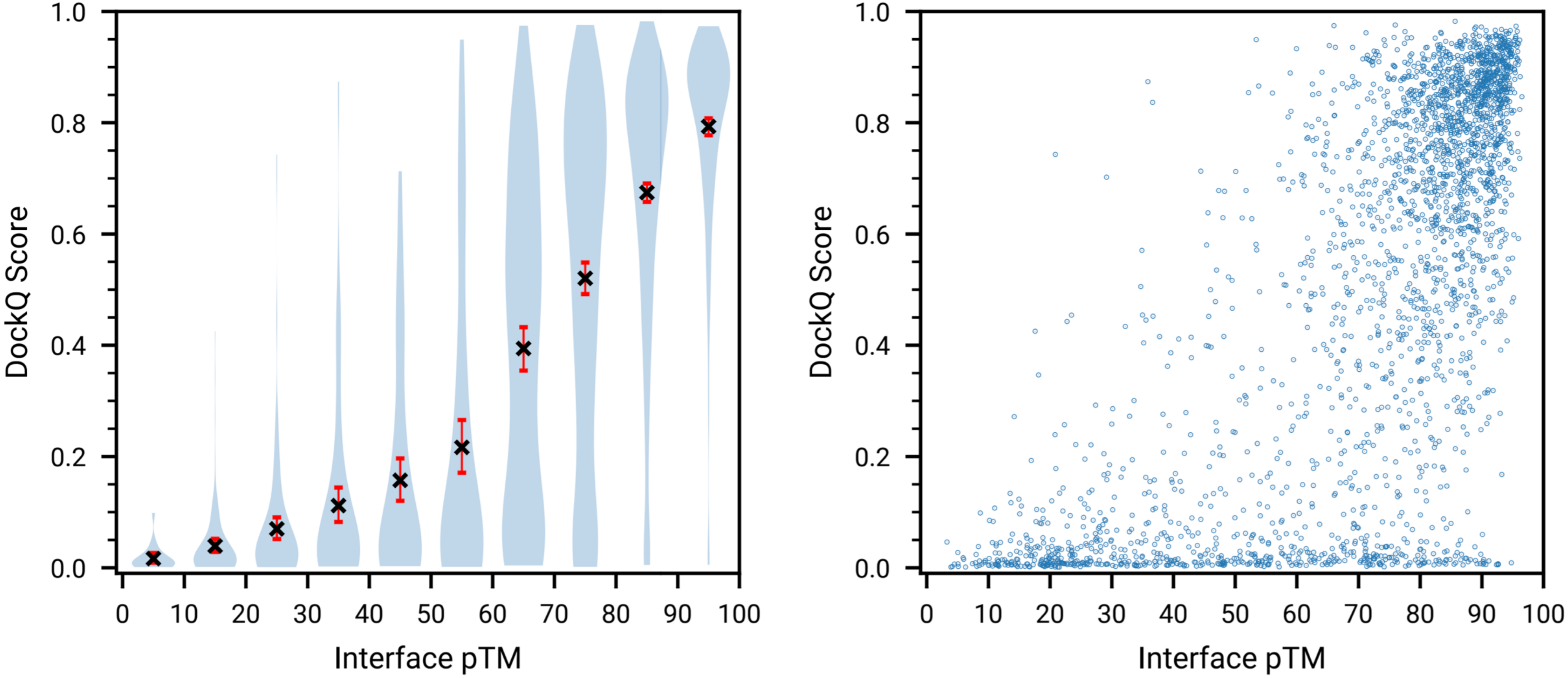
ipTM: Average Score over all chain A residues

ipTM = max over all averages

AlphaFold-Multimer – *iPTM* value

iPTM comes from PAE distribution of output from network between chains

Protein complex prediction with AlphaFold-Multimer



(a) (b)

v2.2 in March 2022

Solution – use PAE cutoff
ipSAE: interaction prediction Score from Aligned Errors

$$ipSAE(A \rightarrow B) = \max_{i \in A} \left[\underset{\substack{j \in B \\ PAE_{ij} < cutoff}}{\text{mean}} \left(\frac{1}{1 + \left(\frac{\text{PAE}_{ij}}{d_0} \right)^2} \right) \right]$$

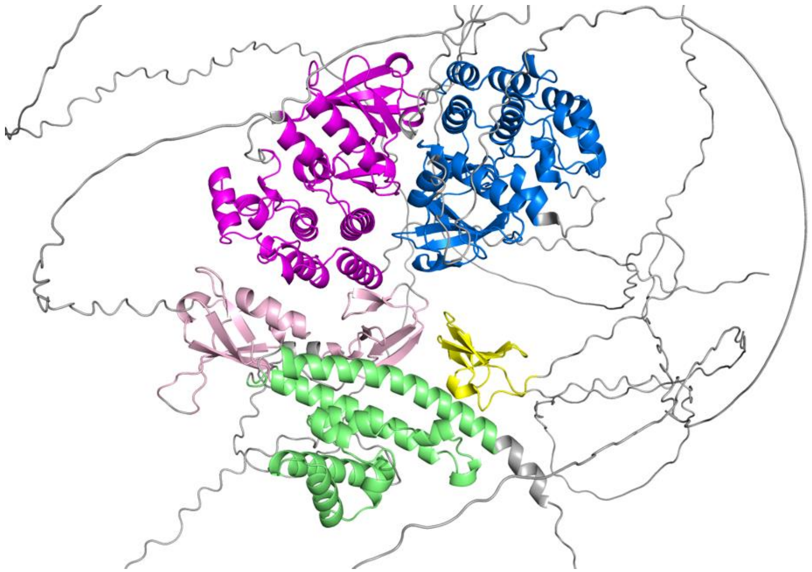
where

$$\boxed{\begin{array}{ll} d_0 = 1.24\sqrt[3]{L_{PAE<cutoff} - 15} - 1.8 & L \geq 27 \\ d_0 = 1 & L < 27 \end{array}}$$

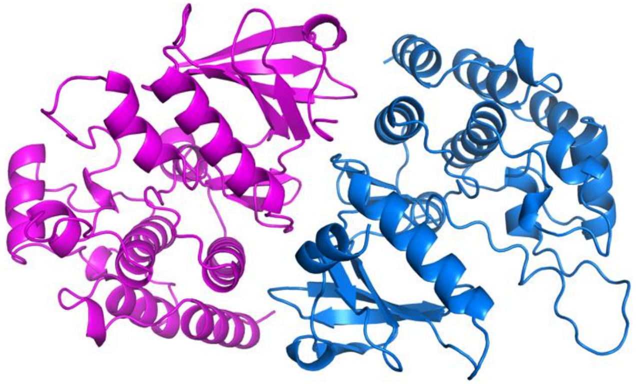
ipSAE solves problem of disorder/extra-domains

when present in both chains

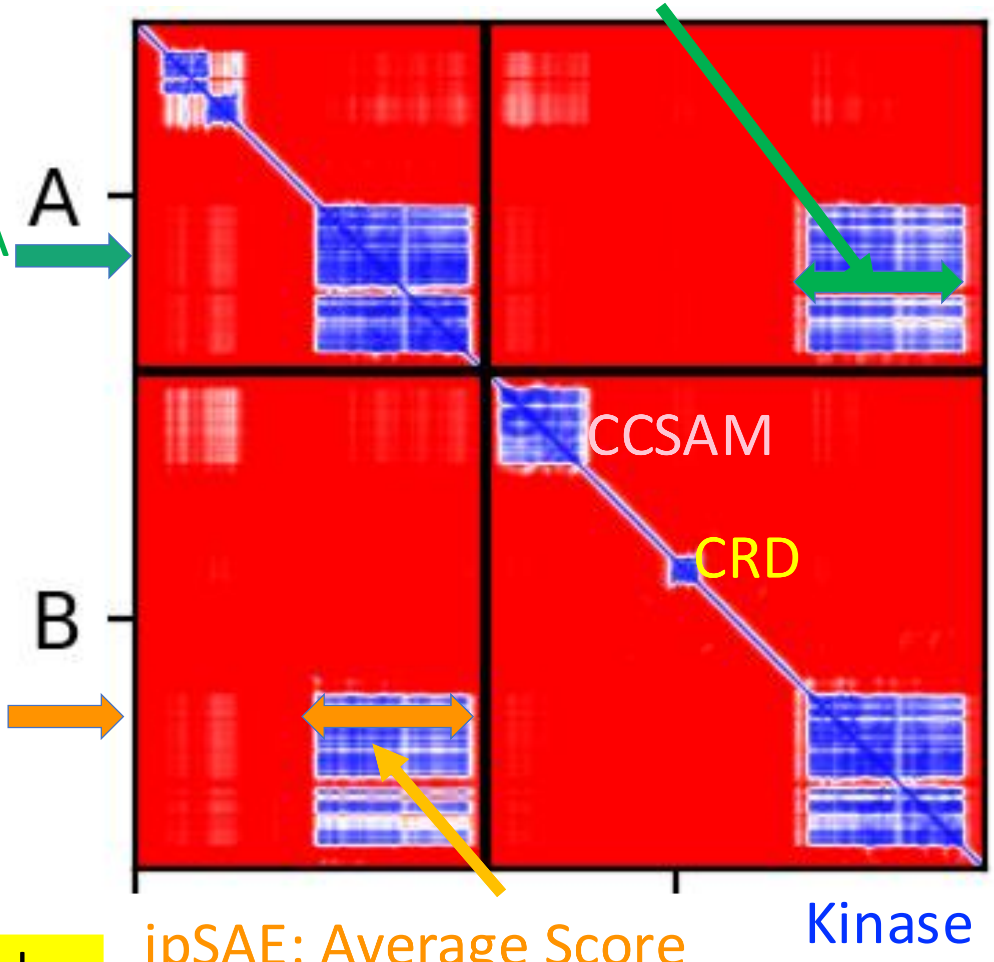
ipSAE=0.75



ipSAE=0.75



ipSAE: Average Score over B residues: PAE<cutoff



Align chain A residue

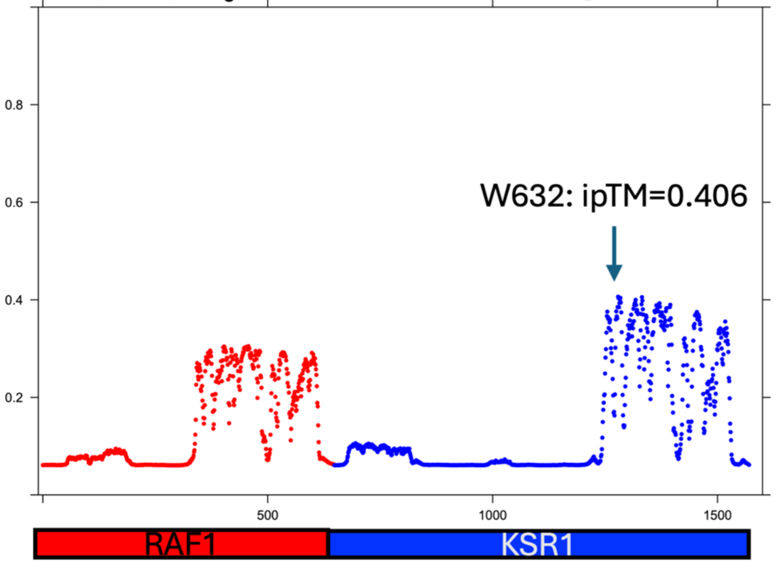
Align chain B residue

ipSAE: Average Score over A residues: PAE>cutoff

ipSAE = max over all aligned residues

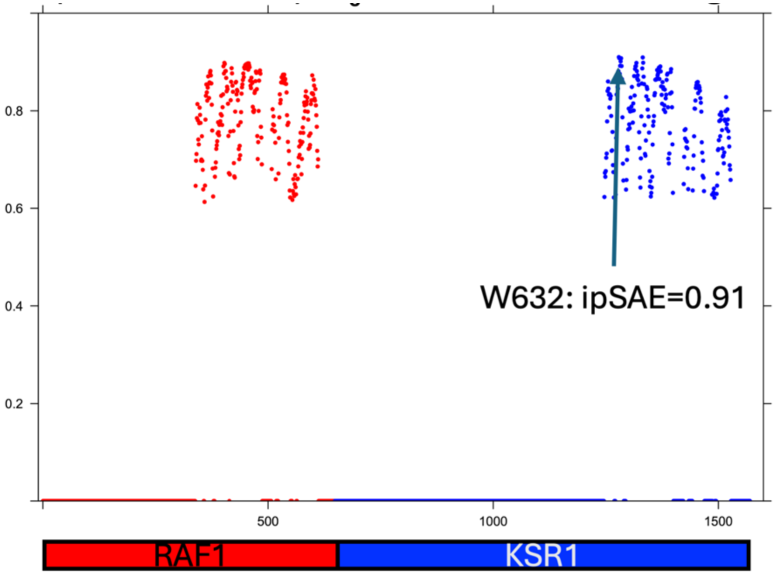
ipTM for alignments on RAF1 or KSR1 residues

ipTM (AF2): d_0 from sum of full-length chains



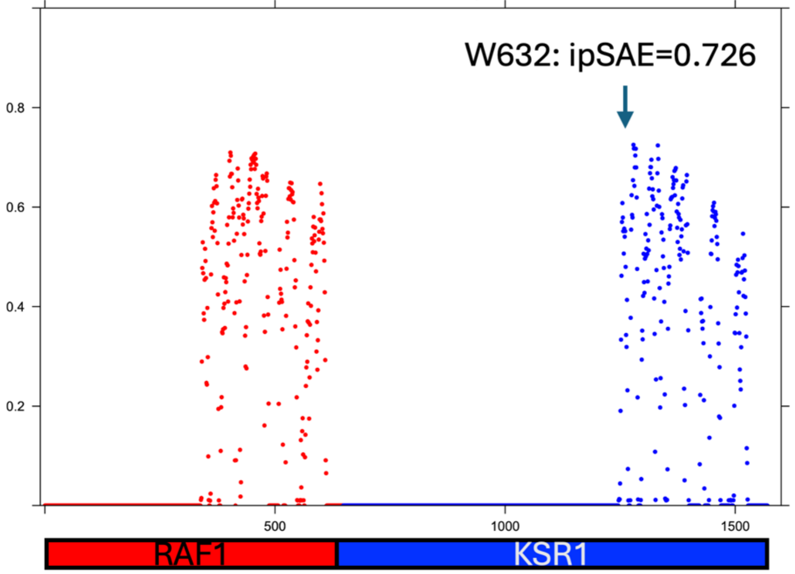
W632: ipTM=0.406

ipSAE (PAEcutoff=15):
d_0 from sum of full-length chains



W632: ipSAE=0.91

ipSAE (PAEcutoff=15): d_0 from PAE<cufoff



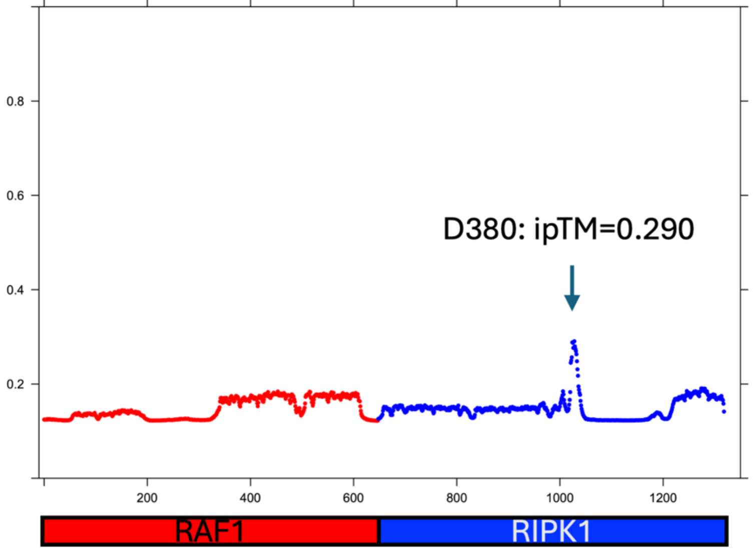
W632: ipSAE=0.726

$$pTM_i = \frac{1}{L}\sum_{j=1}^{L}\frac{1}{1+\left(\frac{PAE_{ij}}{d_0}\right)^2}$$

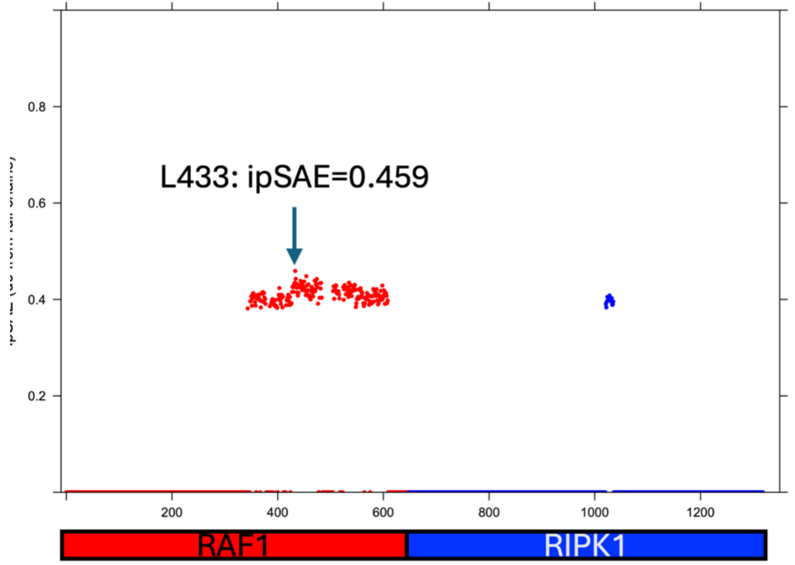
ipTM for alignments on RAF1 and RIPK1 residues

Decoy complex: effect of d_0

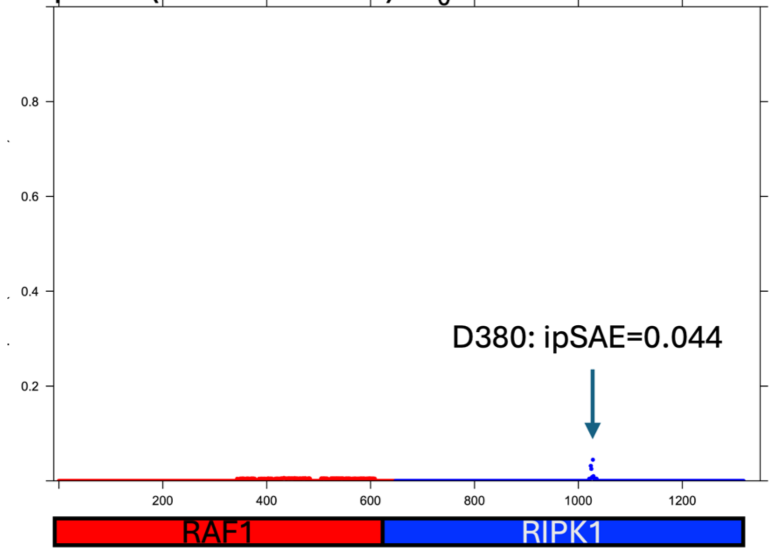
ipTM (AF2): d_0 from sum of full-length chains


D380: ipTM=0.290

ipSAE (PAEcutoff=15 Å):
d_0 from sum of full-length chains


L433: ipSAE=0.459

ipSAE (PAEcutoff=15): d_0 from PAE<cufoff

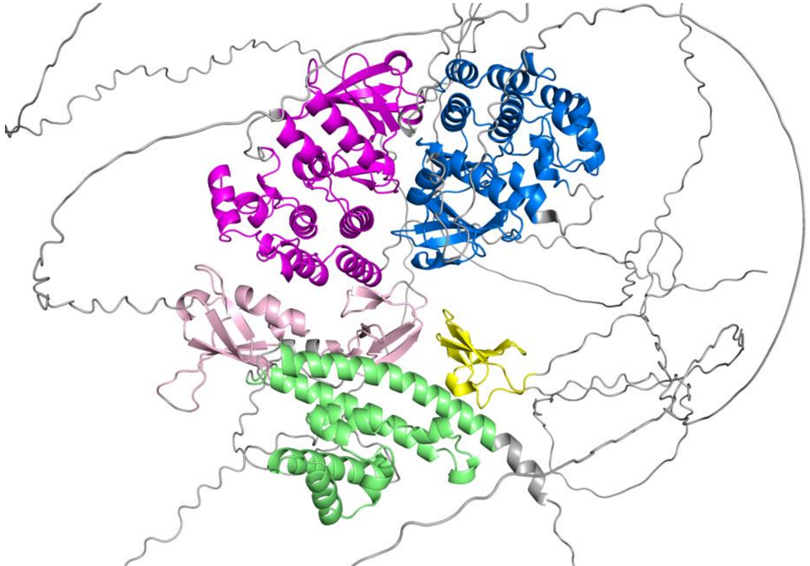

D380: ipSAE=0.044

$$pTM_i = \frac{1}{L}\sum_{j=1}^{L}\frac{1}{1+\left(\frac{PAE_{ij}}{d_0}\right)^2}$$

ipSAE solves problem of disorder/extra-domains

when present in both chains

Chain1	**Chain2**	**nres**	**ipSAE**	**ipTM**	**d0_ipSAE**	**d0_ipTM**
RAF1	KSR1	1571	0.75	0.41	6.2	12.6
RAF1kd	KSR1	1233	0.73	0.81	6.1	11.4
RAF1	KSR1kd	973	0.74	0.79	6.1	10.4
RAF1kd	KSR1kd	635	0.75	0.76	6.2	8.8



$$pTM_{ij} = \frac{1}{1 + \left(\frac{PAE_{ij}}{d_0}\right)^2}$$

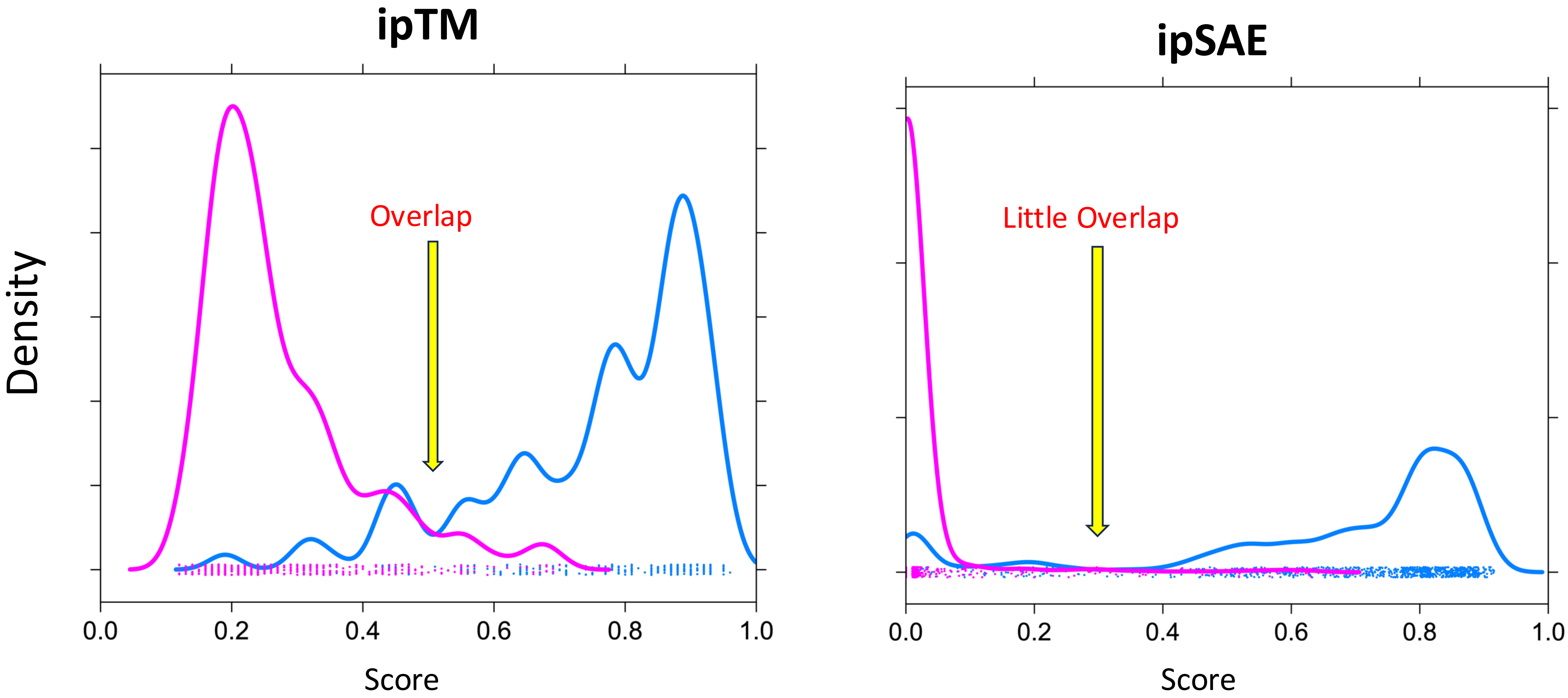
$$d_0 = 1.24\sqrt[3]{L_{PAE<cutoff} - 15} - 1.8 \qquad L \geq 27$$

$$d_0 = 1 \qquad L < 27$$

ipSAE distinguishes interactors from non-interactors better than ipTM

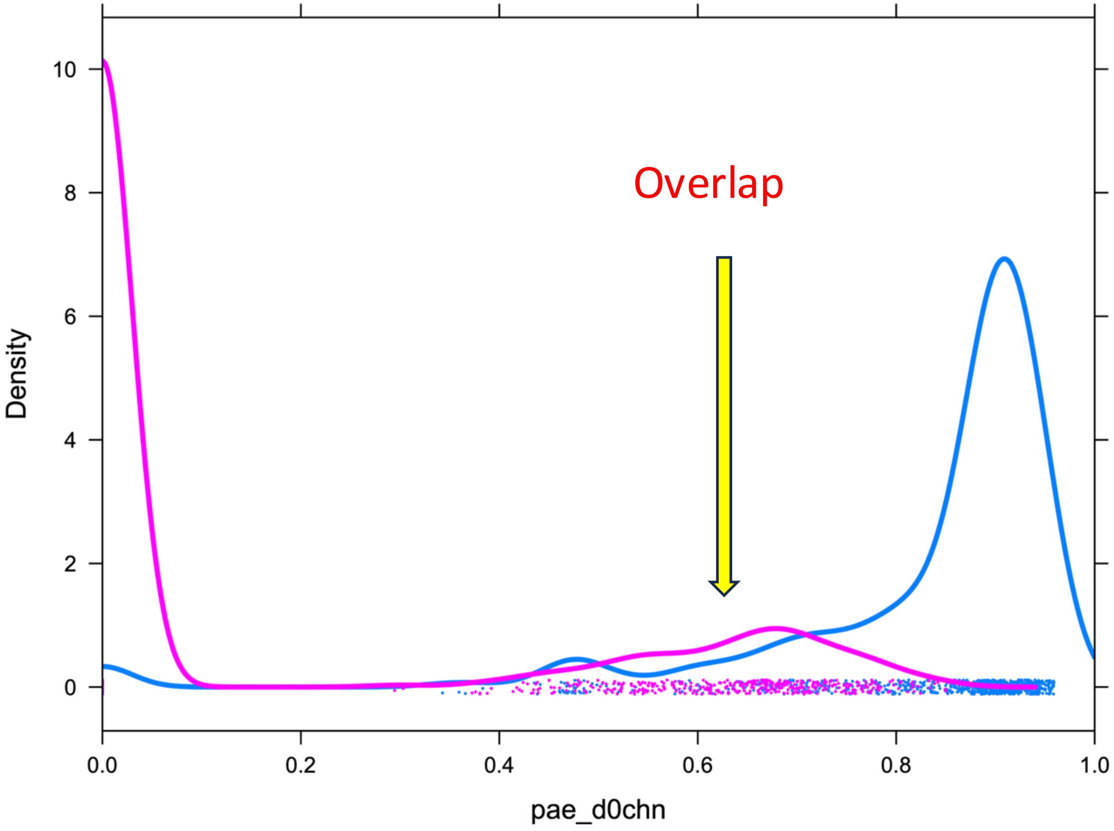
Full Uniprot sequences given to AlphaFold2

40 true complexes   70 false complexes

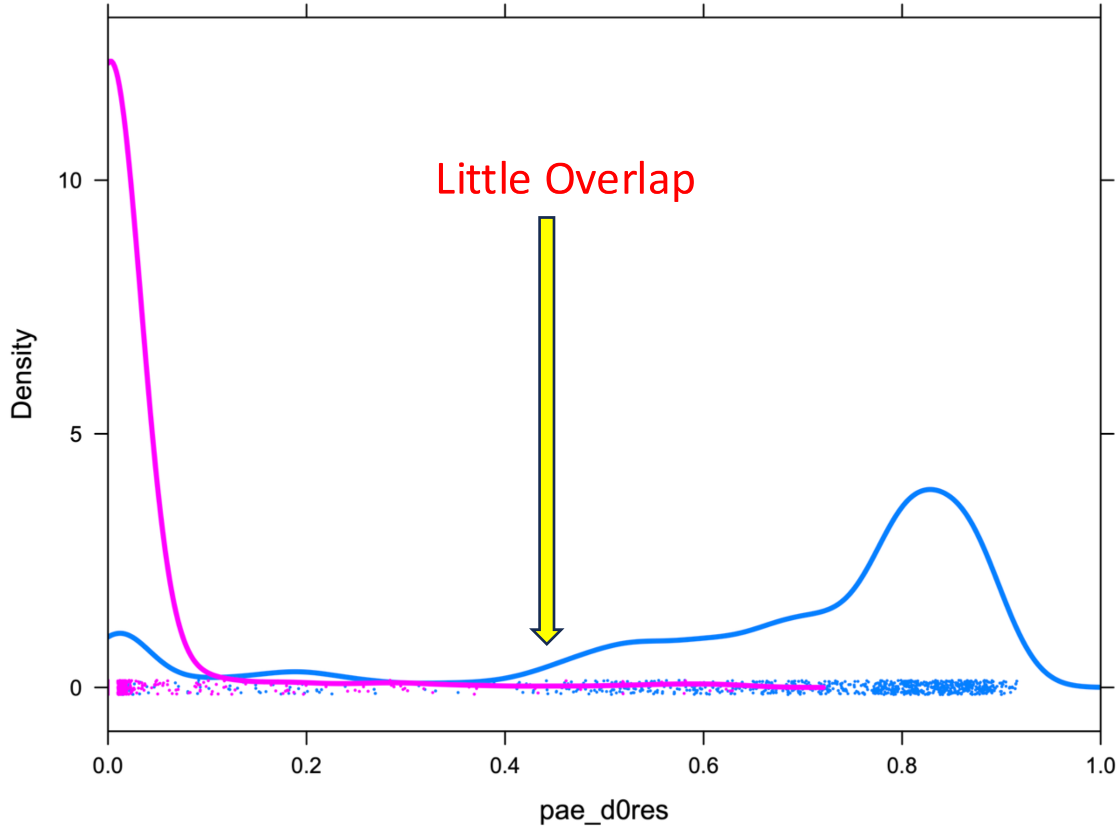


Effect of d_0 calculation

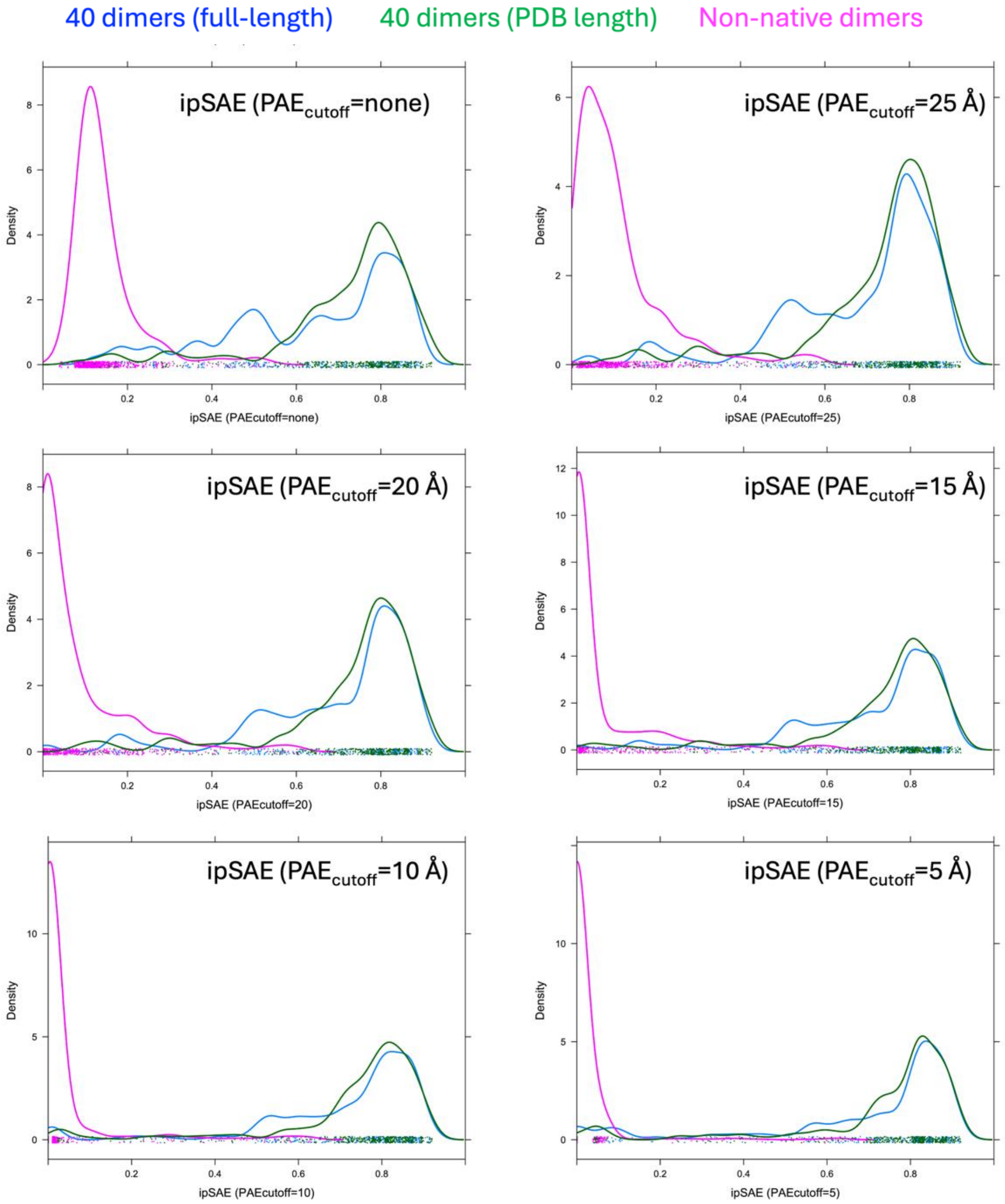
d_0 based on sum of full chain lengths
(ipTM method)



d_0 based on number of PAE<cutoff
(ipSAE method)



Effect of PAE cutoff



Smaller PAE cutoff

→

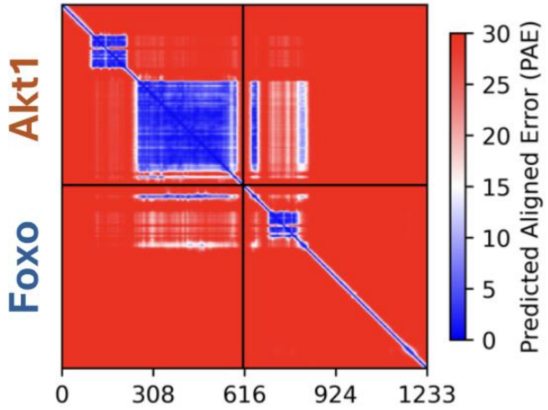
More separation of true/false complexes

Other solutions

- Use pLDDT cutoff to restrict ptm_{ij} : accessory domains have good pLDDT
- Use pLDDT of contact residues (pDockQ): pLDDT often have good values because of folded domains
- Use PAE of contact residues (pDockQ2): accessory domains and disorder may make poorly scored contacts
- Use well-predicted contacts to restrict pTM (actifpTM) : excludes full binding domains from calculation and d_0 from sum of full-length sequences
- Different function of PAE (LIS): linear transform of PAE: 12-0 Å → LIS = 0-1 (average over PAE pairs < 12 Å)

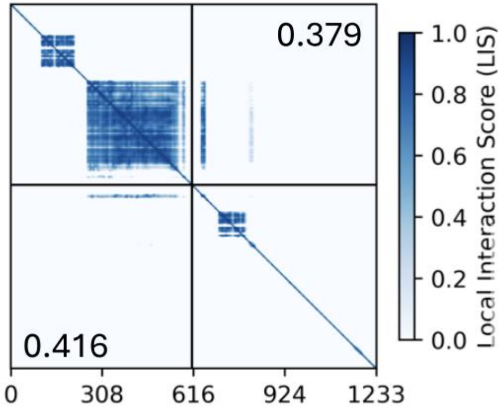
LIS = Local Interaction Score

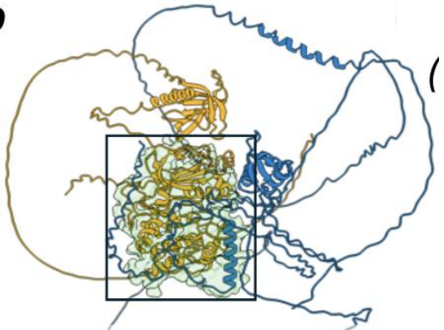
AFM prediction



Local Interaction Area Map

(PAE ≤ 12Å)

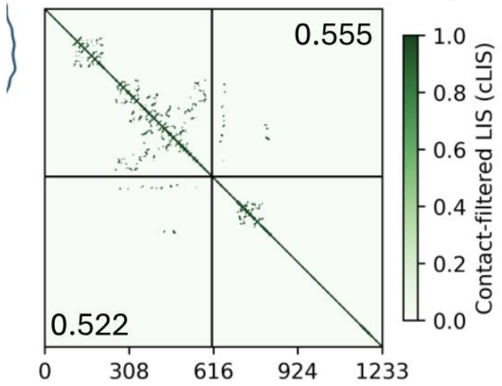




$$LIS = \left(\frac{0.379+0.416}{2}\right) = 0.396$$

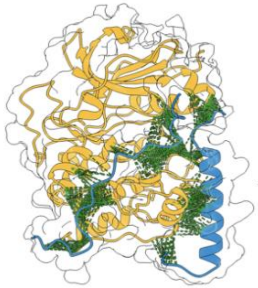
Contact-filtered LIA Map

(PAE ≤ 12Å, Distance ≤ 8Å)



Integrated LIS (iLIS)

$(\sqrt{LIS\ x\ cLIS})$



$$iLIS = \sqrt{0.396\ x 0.539} = \mathbf{0.462}$$

$$cLIS = \left(\frac{0.555+0.522}{2}\right) = 0.539$$

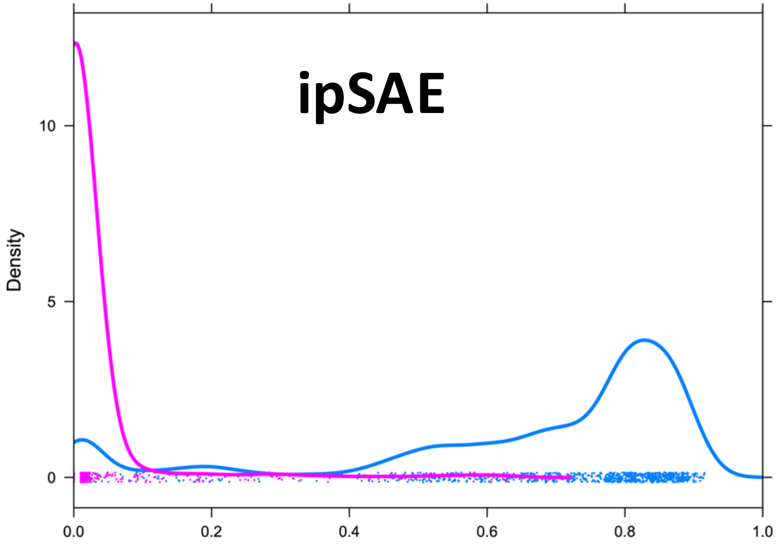
LIS

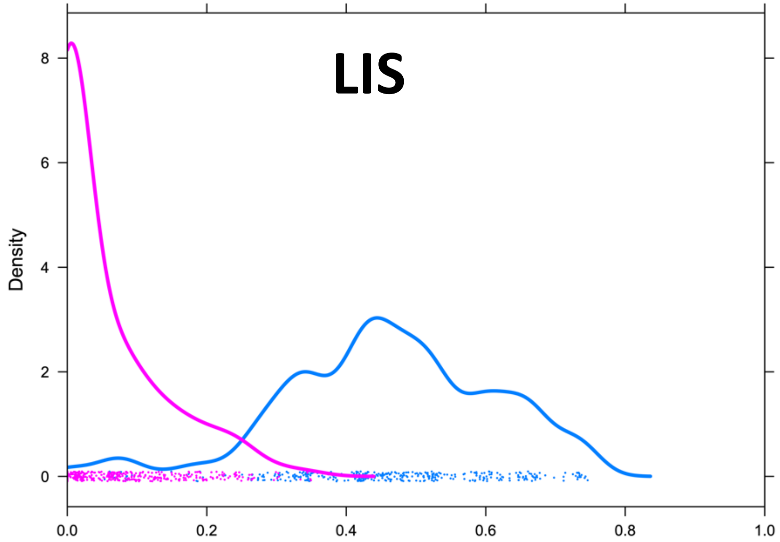
1. Filter PAE ≤ 12 Å
2. Convert PAE from 12 to 0 Å into score from 0 to 1.0 (linear)
3. Average in A→B and B→A rectangles
4. LIS=Average of two rectangles

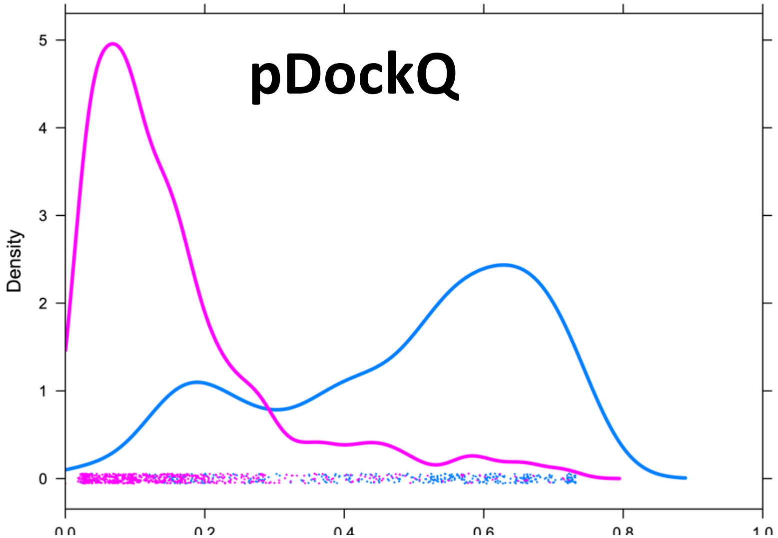
cLIS and iLIS

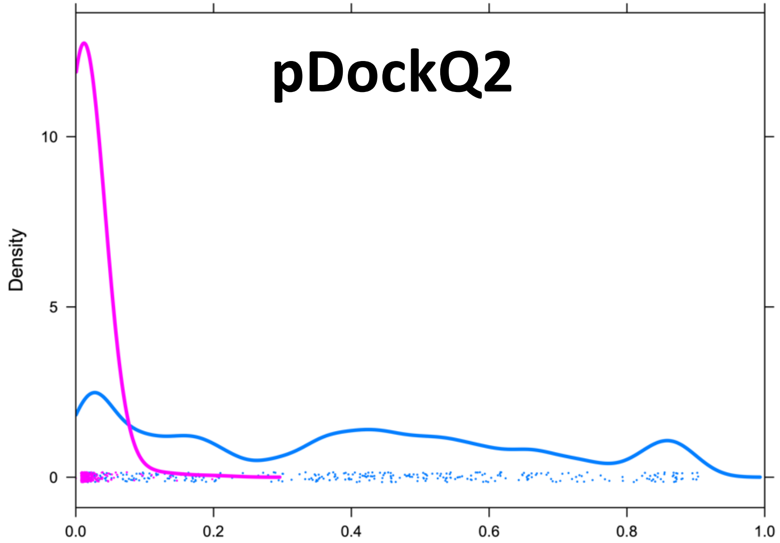
1. Filter PAE by contacts ≤ 8 Å
2. Convert PAE from 12 to 0 Å into score from 0 to 1.0 (linear)
3. Average in A→B and B→A rectangles
4. cLIS=Average of two rectangles
5. iLIS=sqrt(LIS * cLIS)

Other solutions on same benchmark of 40 true + 70 false complexes







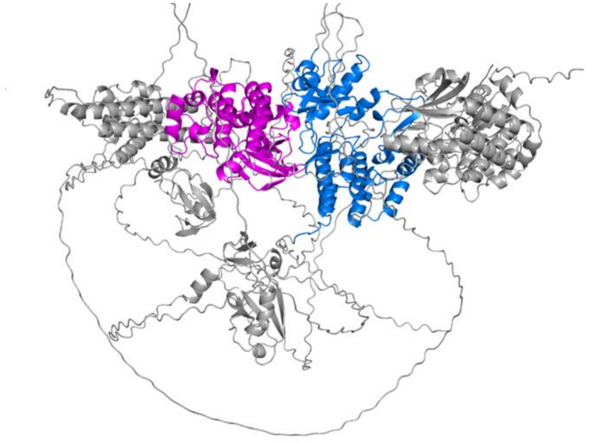


Code and output

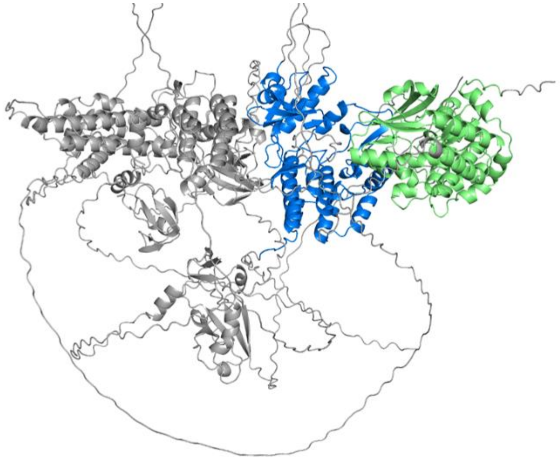
https://github.com/DunbrackLab/IPSAE

```
python ipsae.py  <path_to_json_file>  <path_to_af2_pdb_file>  <pae_cutoff> <dist_cutoff>
python ipsae.py  <path_to_json_file>  <path_to_af3_cif_file>  <pae_cutoff> <dist_cutoff>

python ipsae.py fold_raf1_ksr1_mek1_full_data_0.json fold_raf1_ksr1_mek1_model_0.cif 15 15
```

```
Chn1 Chn2  PAE Dist  Type   ipSAE    ipSAE_d0chn ipSAE_d0dom  ipTM_af  ipTM_d0chn   n0res  n0chn  n0dom  d0res  d0chn  d0dom nres1  nres2  dist1  dist2  Model

A    B     15   15   asym  0.563173   0.824097    0.678527    0.460    0.362276     292   1571   541   6.28   12.57   8.21   249    292    90     83   fold_raf1_ksr1_mek1_model_0
B    A     15   15   asym  0.546718   0.818835    0.671409    0.460    0.442774     278   1571   539   6.14   12.57   8.20   259    280    83     92   fold_raf1_ksr1_mek1_model_0
A    B     15   15   max   0.563173   0.824097    0.678527    0.460    0.442774     292   1571   541   6.28   12.57   8.21   280    292    92     83   fold_raf1_ksr1_mek1_model_0

A    C     15   15   asym  0.261359   0.484143    0.338764    0.510    0.428632     319   1041   491   6.54   10.71   7.88   171    320     0      0   fold_raf1_ksr1_mek1_model_0
C    A     15   15   asym  0.232004   0.486093    0.350874    0.510    0.270827     260   1041   513   5.96   10.71   8.03   250    263     0      0   fold_raf1_ksr1_mek1_model_0
A    C     15   15   max   0.261359   0.486093    0.350874    0.510    0.428632     319   1041   513   6.54   10.71   8.03   263    320     0      0   fold_raf1_ksr1_mek1_model_0

B    C     15   15   asym  0.636110   0.829537    0.732184    0.770    0.751586     344   1316   607   6.76   11.74   8.61   262    345    69     72   fold_raf1_ksr1_mek1_model_0
C    B     15   15   asym  0.559980   0.802212    0.692571    0.770    0.343057     291   1316   598   6.27   11.74   8.56   307    291    71     69   fold_raf1_ksr1_mek1_model_0
B    C     15   15   max   0.636110   0.829537    0.732184    0.770    0.751586     344   1316   607   6.76   11.74   8.61   291    345    69     72   fold_raf1_ksr1_mek1_model_0
```

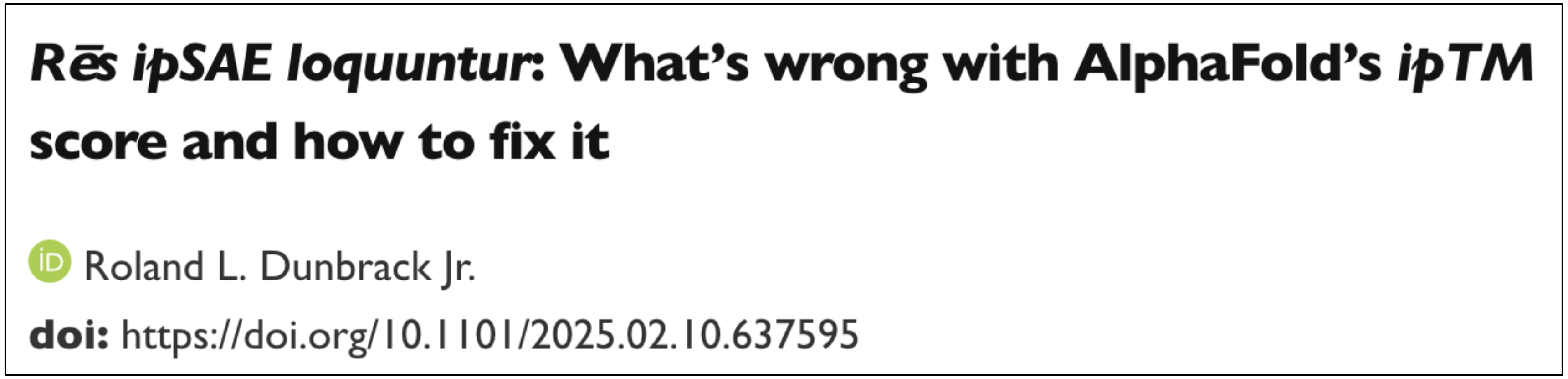


RAF1-KSR1: PAE≤15 Å



KSR1-MEK1: PAE≤15 Å

PyMol script for coloring good PAE residues



Rēs ipSAE loquuntur: What's wrong with AlphaFold's ipTM score and how to fix it

Roland L. Dunbrack Jr.

doi: https://doi.org/10.1101/2025.02.10.637595

Rēs ipSAE loquuntur.

The things themselves speak.

Rēs ipsa loquitur.

The thing itself speaks.

Applications of ipSAE

- PPI screens like BioiD (“interactome studies”)
- Binding specificity: E2/E3-ring pairs (Jarboe and Dunbrack, Biorxiv 2026)
- Loop conformations: Kinase activation loop in modeling the active form of all 437 catalytic protein kinases (Gizzio et al, Biorxiv 2026). Align on kinase domain residues, score activation loop with intramolecular ipSAE
- Domain-domain interactions within large proteins: given domain definitions (e.g. ECOD), score all vs all including linkers. Application to all 480 human protein-kinase-domain containing proteins.
- Protein-binder design

Applications: ipSAE in AFDB Homo- and Heterodimers

MANUSCRIPT

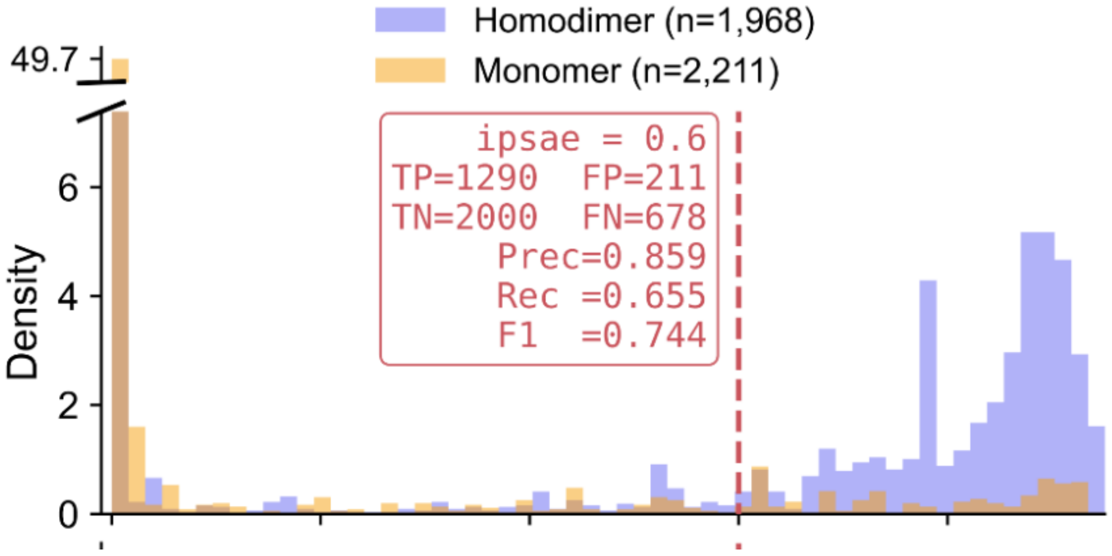
AlphaFold Database expands to proteome-scale quaternary structures

Han, Tsenkov, Venanzi et al.

Yewon Han[1,2*], Maxim I. Tsenkov[3,*], Niccolò A. E. Venanzi[4,*], Damian Bertoni[3], Sooyoung Cha[1,2], Alejandro Chacón[4], Nick Dietrich[5], Boris Fomitchev[4], Yonathan Goldtzvik[3], Darren Hsu[4], Jeannie Austin[3], Joseph Ellaway[3], Kieran Didi[4,6], Oleg Kovalevskiy[5], Dariusz Lasecki[5], Agata Laydon[5], Micha Livne[4], Paulyna Magaña[3], Maciej Majewski[7], Sreenath Nair[3], Urmila Paramval[3], Nilkanth Patel[4], Risha Patel[5], Ivanna Pidruchna[3], Brianda Santini Lopez[4], Prashant Sohani[4], Ahsan Tanweer[3], Duc Tran[4], Kyle Tretina[4], Melanie Vollmar[3], Quan Vu[4], Augustin Žídek[5], Sameer Velankar[3,#], Martin Steinegger[1,2,8,9,#], Jennifer Fleming[3,#], Milot Mirdita[1,10,11,#], Christian Dallago[4,12,13,#]

ipSAE in AFDB Homo- and Heterodimers

1.8 million homodimers online from key genomes



ipSAE_min

ipSAE$_{min}$:

“*very high-confidence*” (≥ 0.8, high accuracy interface; 972,625 entries),

“*confident*” (0.7 to < 0.8, correct, well-resolved interaction likely; 438,879)

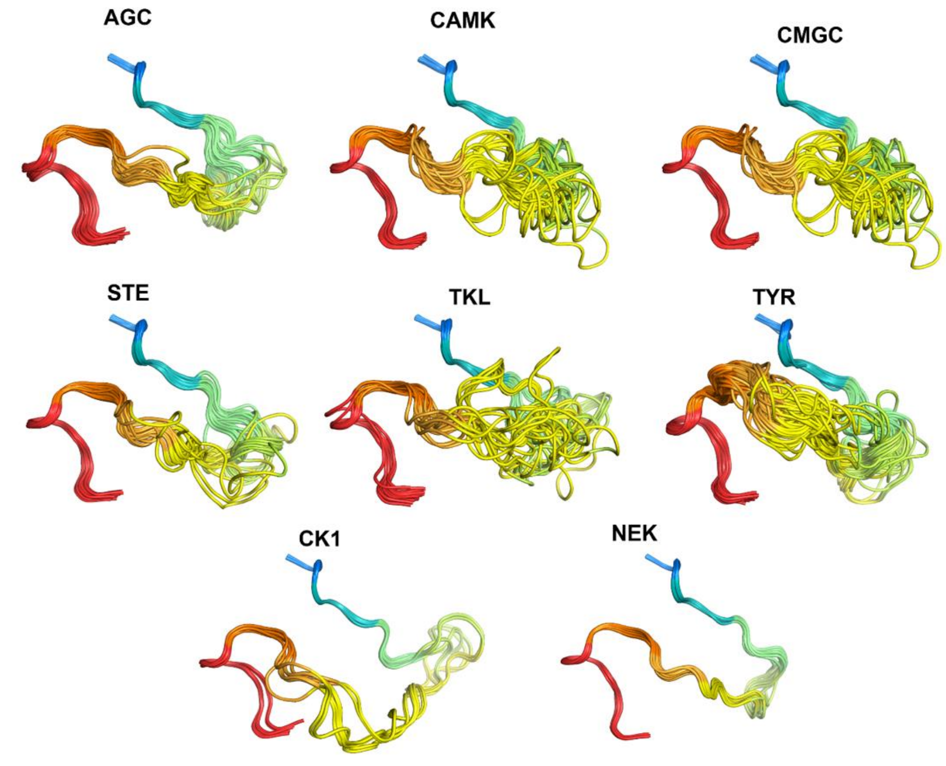
“*low-confidence*” (0.6 to < 0.7, expected interaction signal, but interpret with caution; 342,738).

Applications: Active models of all 437 catalytical human protein kinases with AF2

Intramolecular ipSAE (Gizzio et al, BioRxiv 2026)

54 substrate-bound structures → criteria for active structures (DFG, HRD, and substrates

1. Ortholog MSA
2. *Active* PDB templates
3. Test for *Active models*
4. Choose structure with highest intramolecular ipSAE scoring of activation loop



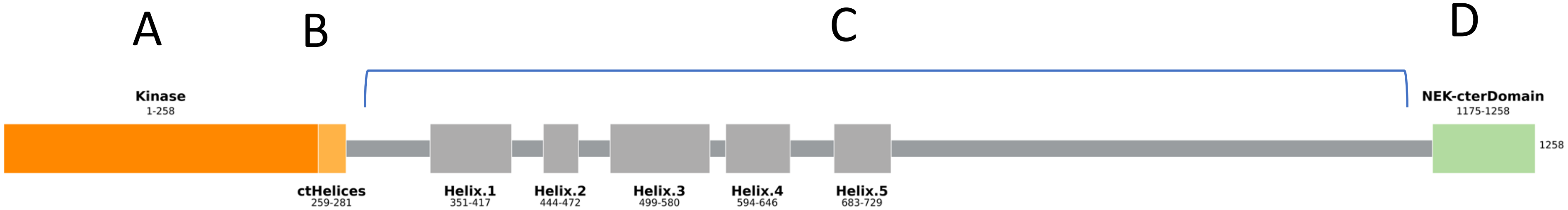
Only 116/437 are Active in PDB (27%)
Only 208 in AFDB
437 models publicly available
https://dunbrack.fccc.edu/kincore

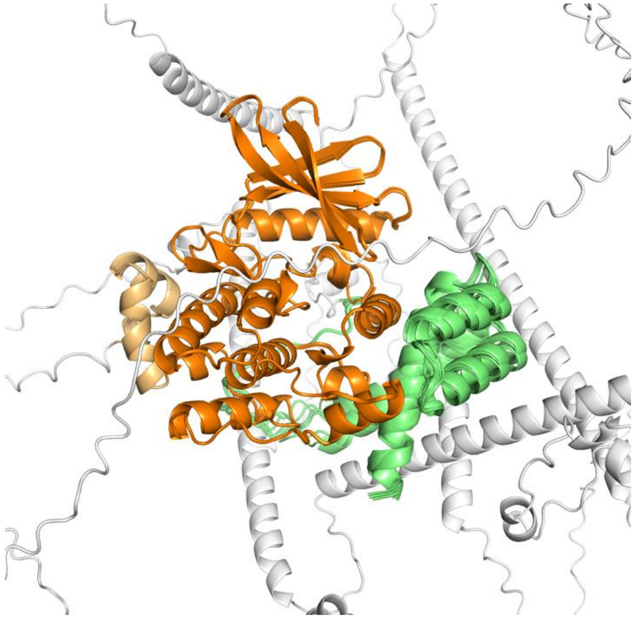
$$ipSAE = \max_{\substack{i\in domain\\ i\notin actloop}} \left[\frac{1}{L} \sum_{j\in actloop}^{L} \frac{1}{1+\left(\frac{PAE_{ij}}{d_0}\right)^2} \right]$$

Scoring intramolecular domain-domain interactions with ipSAE

Just break up model into multiple chain-ids and run ipSAE code as is

Novel domain in
NEK1 kinase
(also NEK4, NEK11)

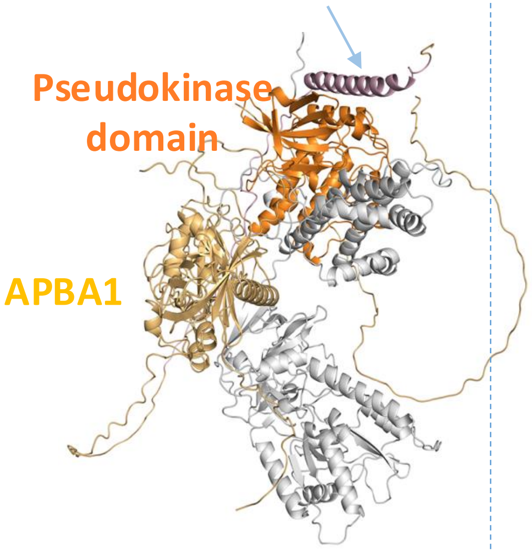




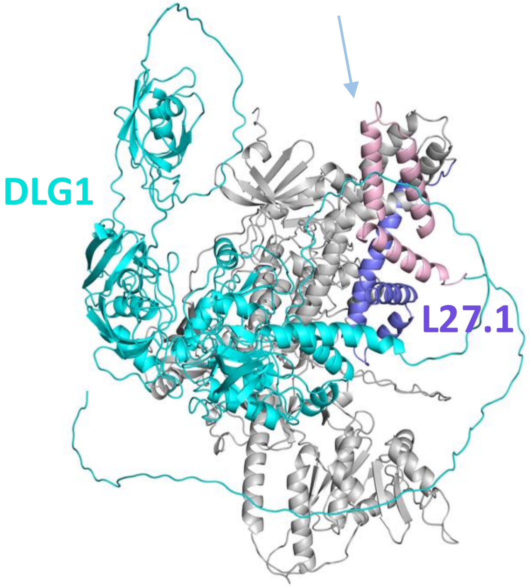
	A	B	C	D
A		0.81	0.11	0.45
B			0.11	0.19
C				0.15
D				

Application: AF2 Interactomes with ipSAE: Models of *CASK pseudokinase* and its interacting partners

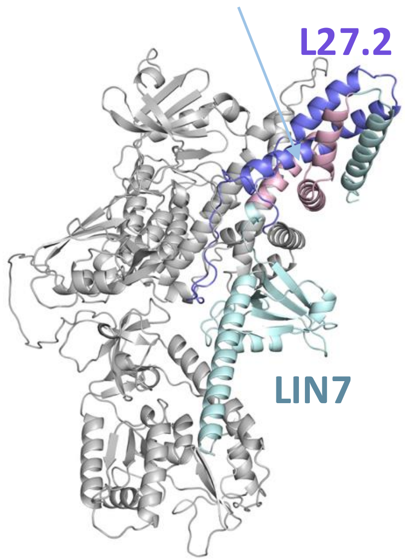
APBA1 (331-406)



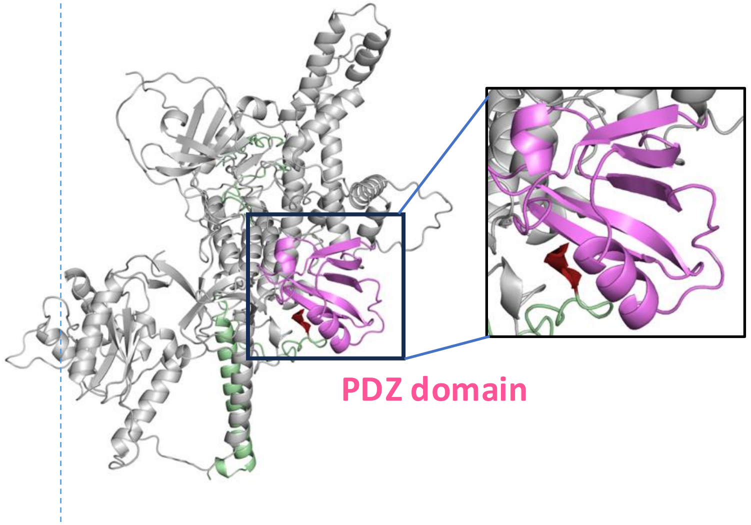
DLG1 (1-65)

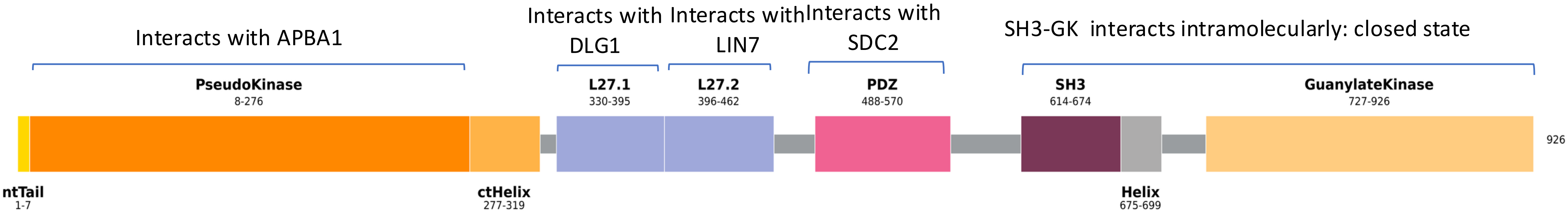


LIN7-L27 domain (1-60)



EFYA motif of SDC2





Interacts with APBA1

Interacts with DLG1

Interacts with LIN7

Interacts with SDC2

SH3-GK interacts intramolecularly: closed state

PseudoKinase 8-276

L27.1 330-395

L27.2 396-462

PDZ 488-570

SH3 614-674

GuanylateKinase 727-926

926

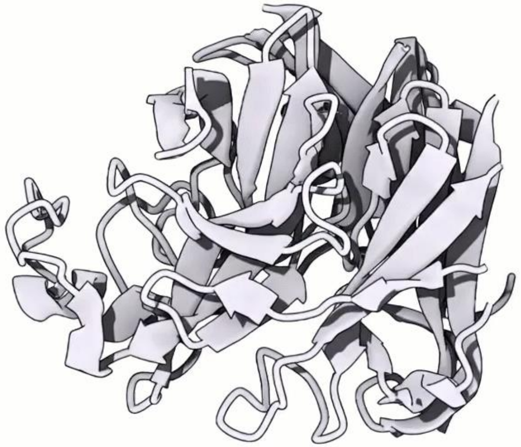
ntTail 1-7

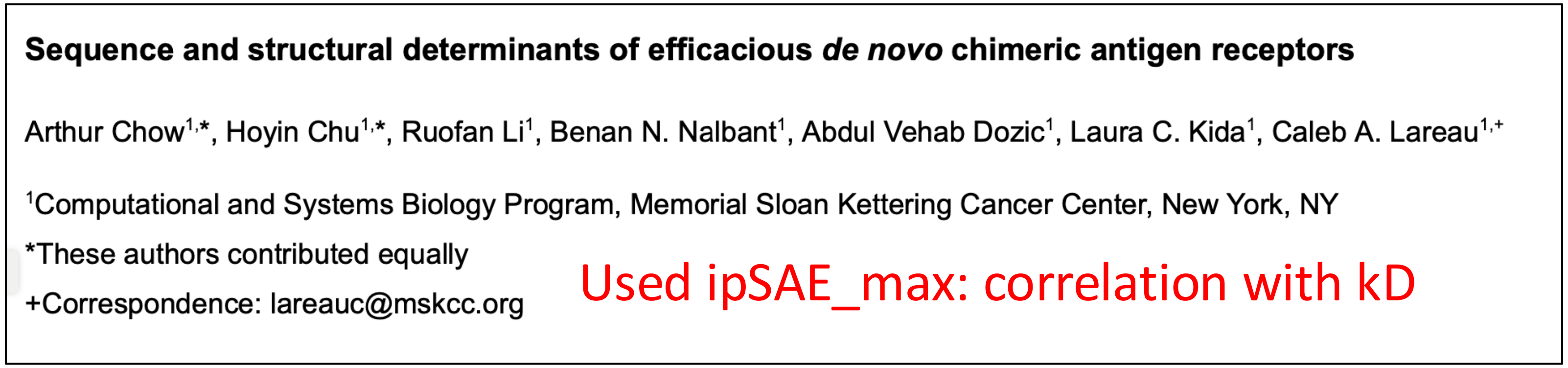
ctHelix 277-319

Helix 675-699

Applications: ipSAE in Protein binder design

Works as a filter to enrich binders





Sequence and structural determinants of efficacious *de novo* chimeric antigen receptors

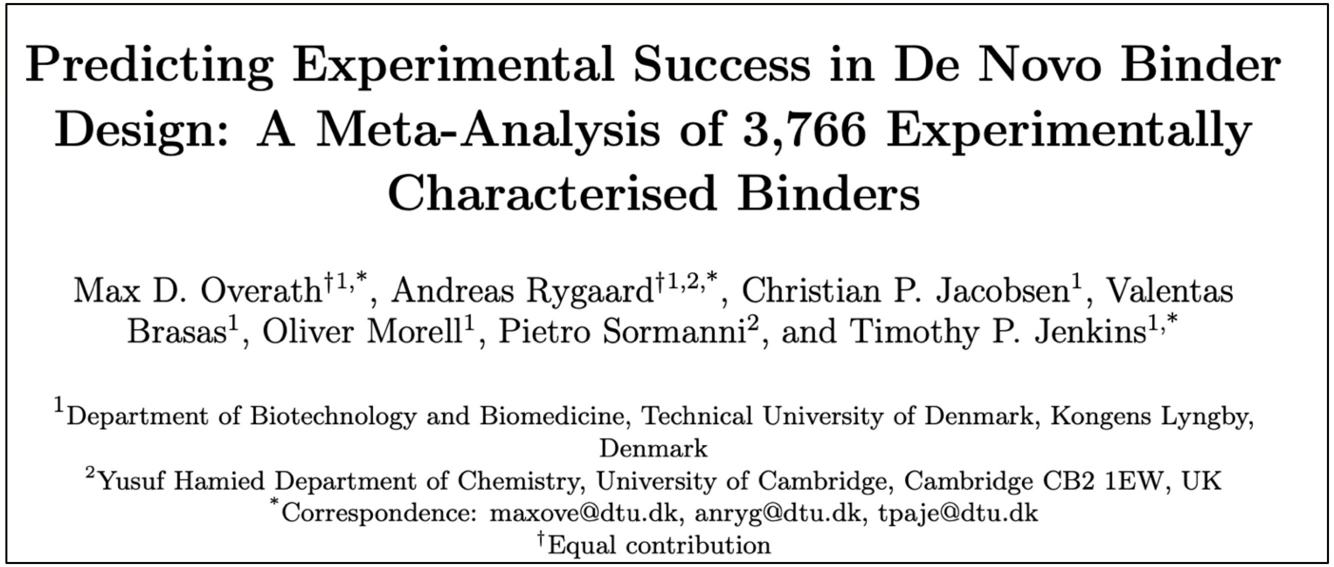
Arthur Chow[1,*], Hoyin Chu[1,*], Ruofan Li[1], Benan N. Nalbant[1], Abdul Vehab Dozic[1], Laura C. Kida[1], Caleb A. Lareau[1,+]

[1]Computational and Systems Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY

*These authors contributed equally

+Correspondence: lareauc@mskcc.org

Used ipSAE_max: correlation with kD

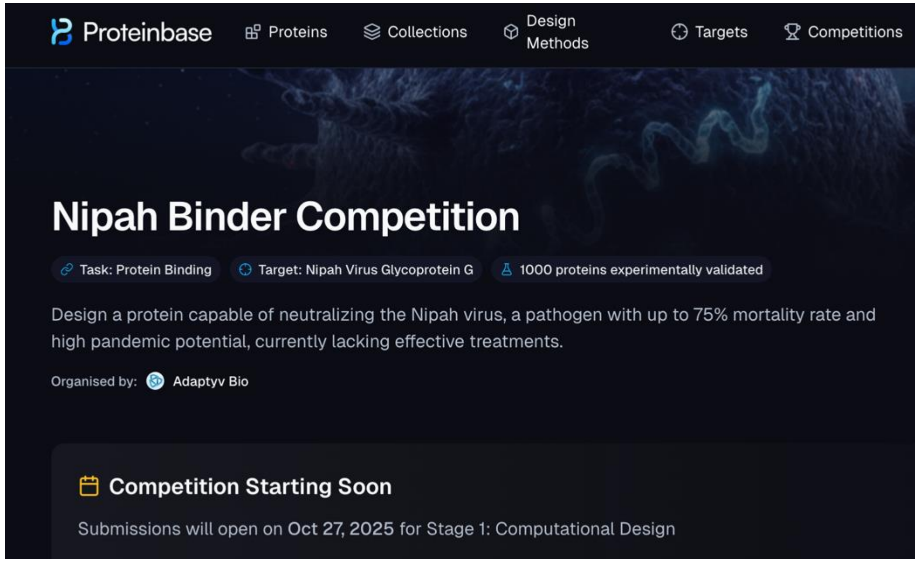


Predicting Experimental Success in De Novo Binder Design: A Meta-Analysis of 3,766 Experimentally Characterised Binders

Max D. Overath[†1,*], Andreas Rygaard[†1,2,*], Christian P. Jacobsen[1], Valentas Brasas[1], Oliver Morell[1], Pietro Sormanni[2], and Timothy P. Jenkins[1,*]

[1]Department of Biotechnology and Biomedicine, Technical University of Denmark, Kongens Lyngby, Denmark
[2]Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, UK
[*]Correspondence: maxove@dtu.dk, anryg@dtu.dk, tpaje@dtu.dk
[†]Equal contribution

Used ipSAE_min



Nipah Binder Competition

Task: Protein Binding | Target: Nipah Virus Glycoprotein G | 1000 proteins experimentally validated

Design a protein capable of neutralizing the Nipah virus, a pathogen with up to 75% mortality rate and high pandemic potential, currently lacking effective treatments.

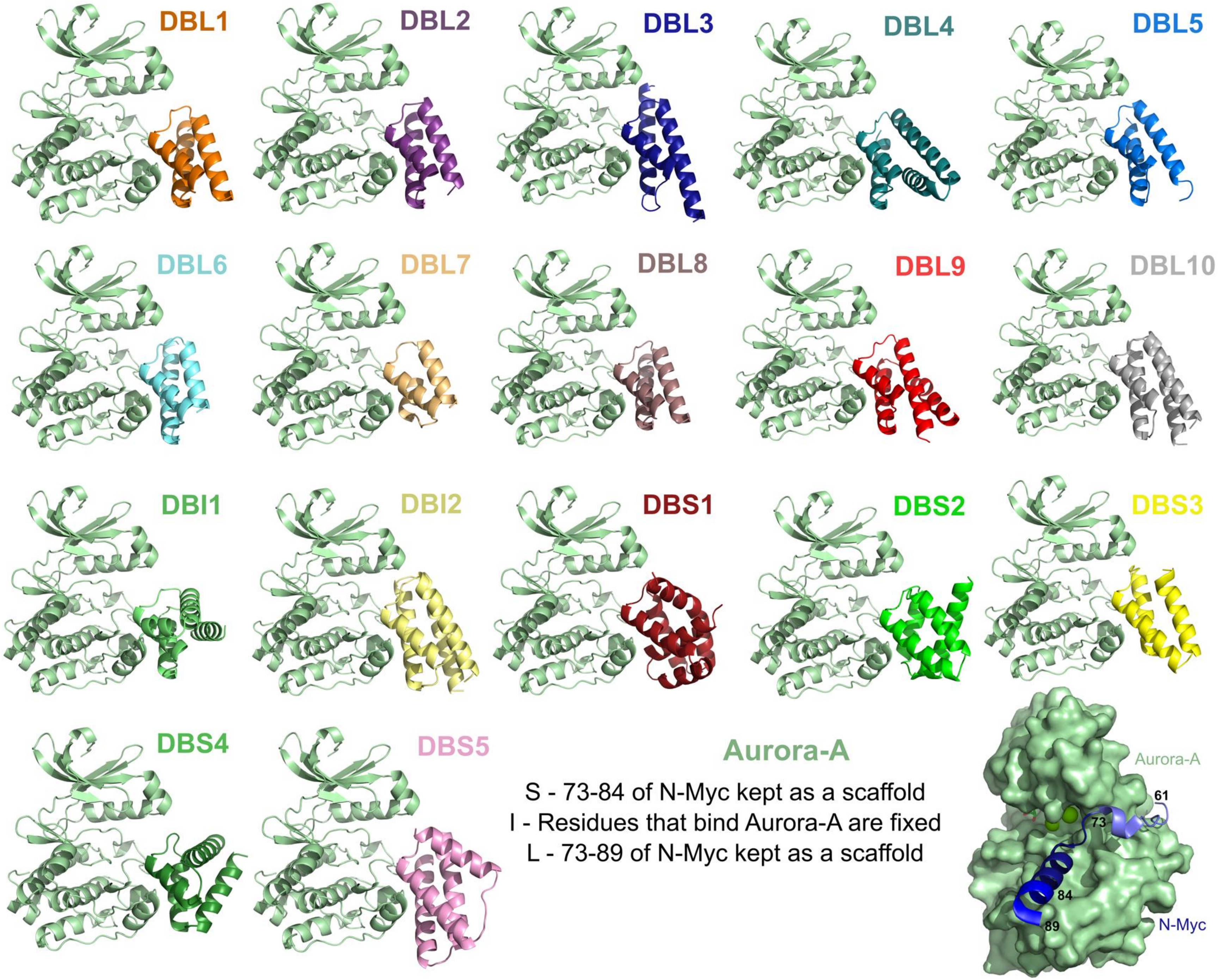
Organised by: Adaptyv Bio

Competition Starting Soon

Submissions will open on Oct 27, 2025 for Stage 1: Computational Design

Used ipSAE_min as filter

AlphaFold2 models of MYC-based designs bound to Aurora-A (J. Miles and R. Bayliss, Leeds)



Aurora-A

S - 73-84 of N-Myc kept as a scaffold
I - Residues that bind Aurora-A are fixed
L - 73-89 of N-Myc kept as a scaffold

The helix from N-Myc is present at the front of each of the binders.

The designs comprise 3 helices at least, with the most being a 6 helical bundle.

ipSAE from AF2 models

Binding affinities determined using Isothermal Titration Calorimetry

Construct	Avg ipSAE	Kd (nM)
DBS1	**0.75**	**1**
DBL5	**0.64**	**3**
DBS5	**0.71**	**22**
DBS2	0.42	114
DBL6	0.53	141
DBL3	0.56	250
DBL4	0.70	787

Jennifer Miles and Richard Bayliss, University of Leeds

Further Considerations

- ipSAE_min, ipSAE_max, ipSAE (A → B), ipSAE (B → A) for protein design
- ipSAE_min, ipSAE_max, ipSAE (A → B), ipSAE (B → A) for domain/peptides
- AF2 vs AF3 vs Boltz2 ipSAE scores
- Protein/nucleic-acid complexes in AF3, Boltz2
- Protein/ligand complexes in AF3, Boltz2
- Multivalent PPI

Assigning all domains in human protein kinases from full-length AlphaFold models

Functional domains can be:

- Activating
- Inhibiting
- Membrane-binding
- Substrate-association
- Catalytic activity
- Scaffolding

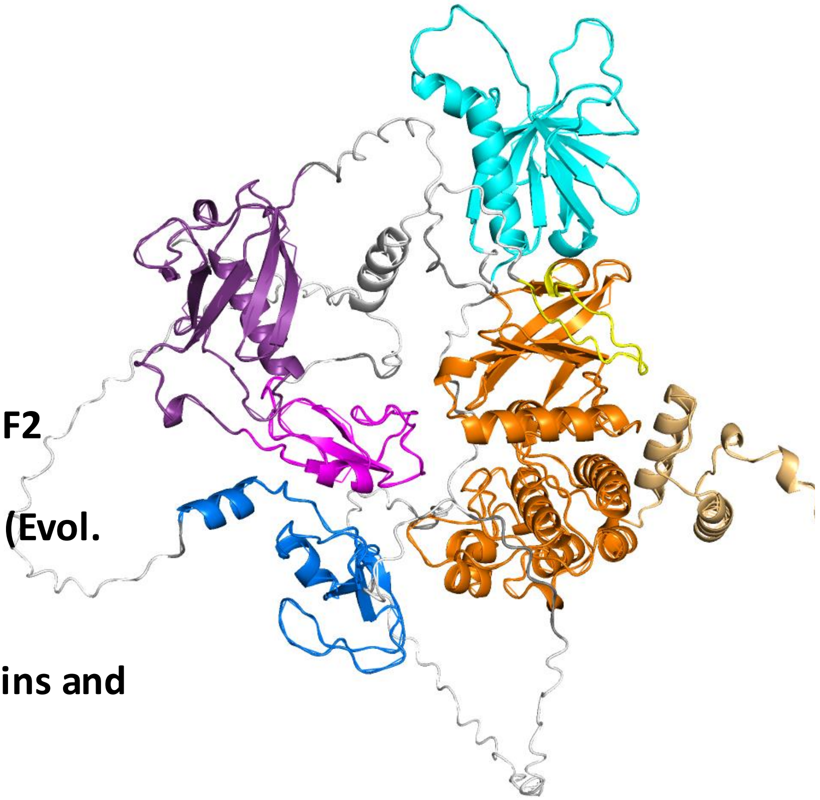
Full-length models of UniProt sequences with AF2

Visualization with PyMOL of ECOD assignments (Evol. Classification of Domains, Grishin, UTSW)

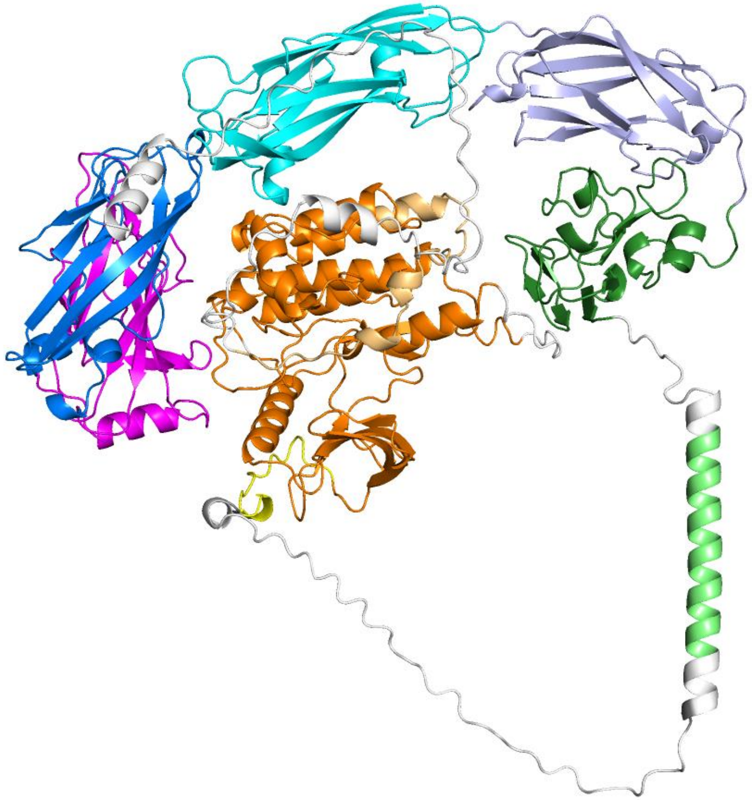
Manual recording of 1673 domains in 480 proteins and NT/CT tails and isolated helices

17 novel domains in 30 kinases

Domain diagrams and data freely available (soon)



CAMK_PRKD1



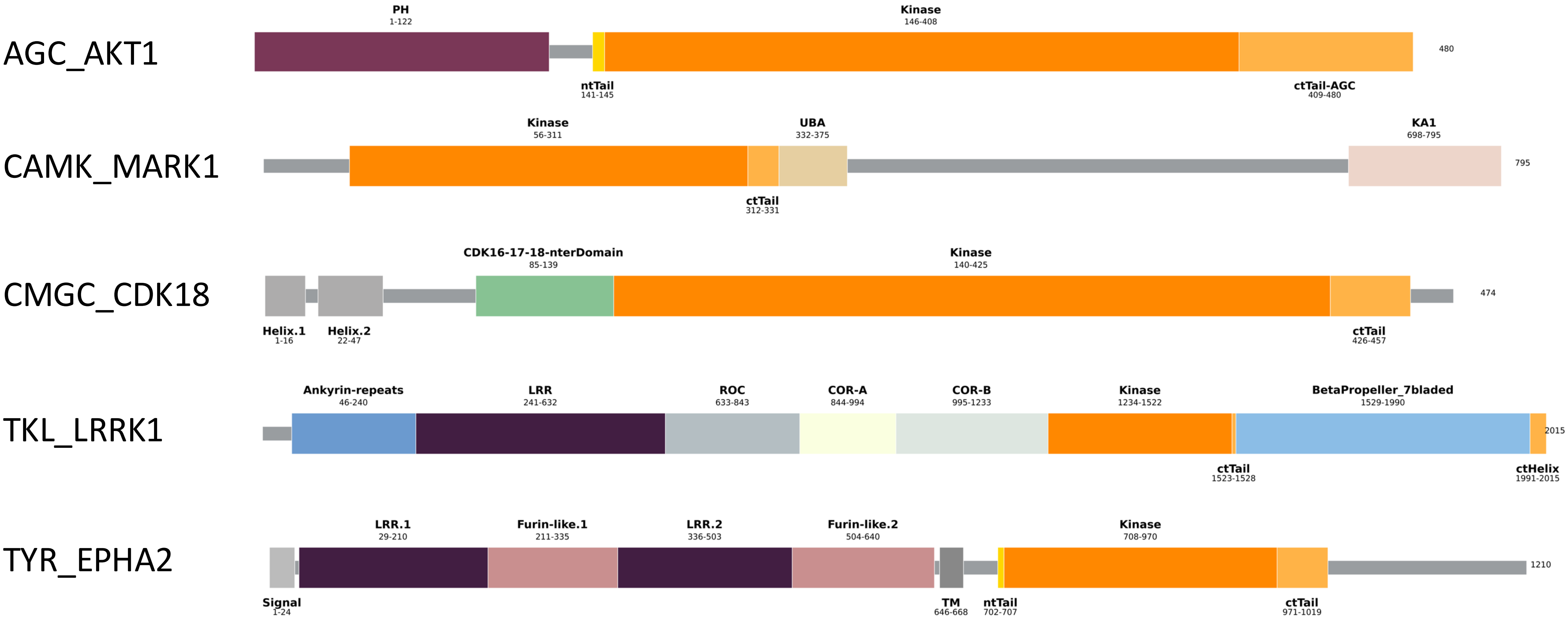
TYR_RET

Domain diagrams for all human kinases

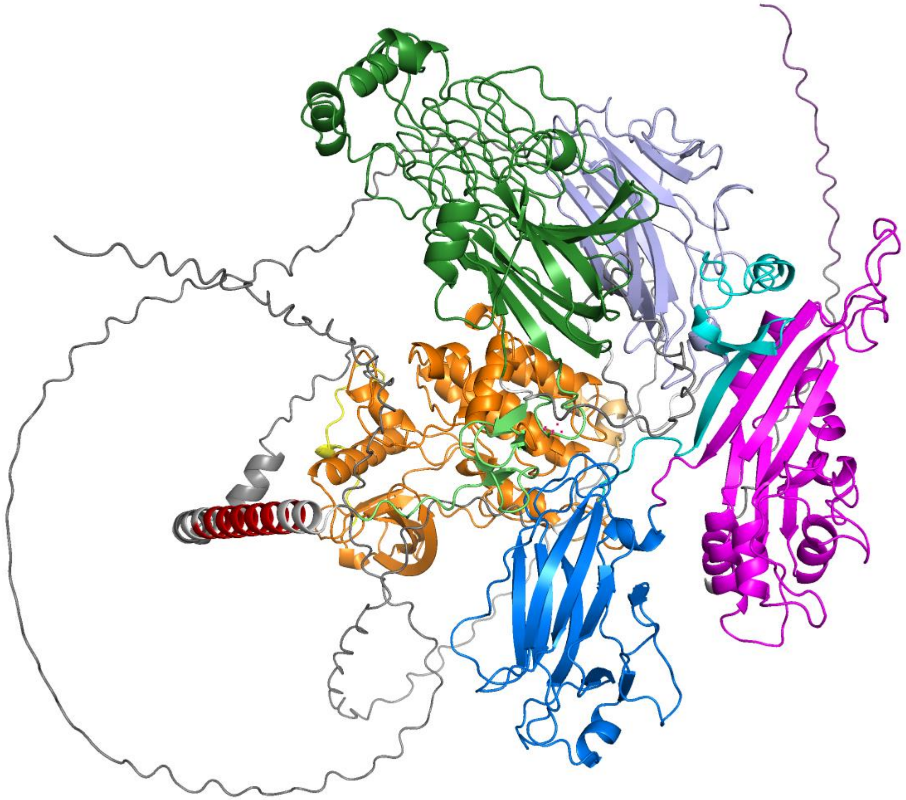
21 domains shared between families – one color each

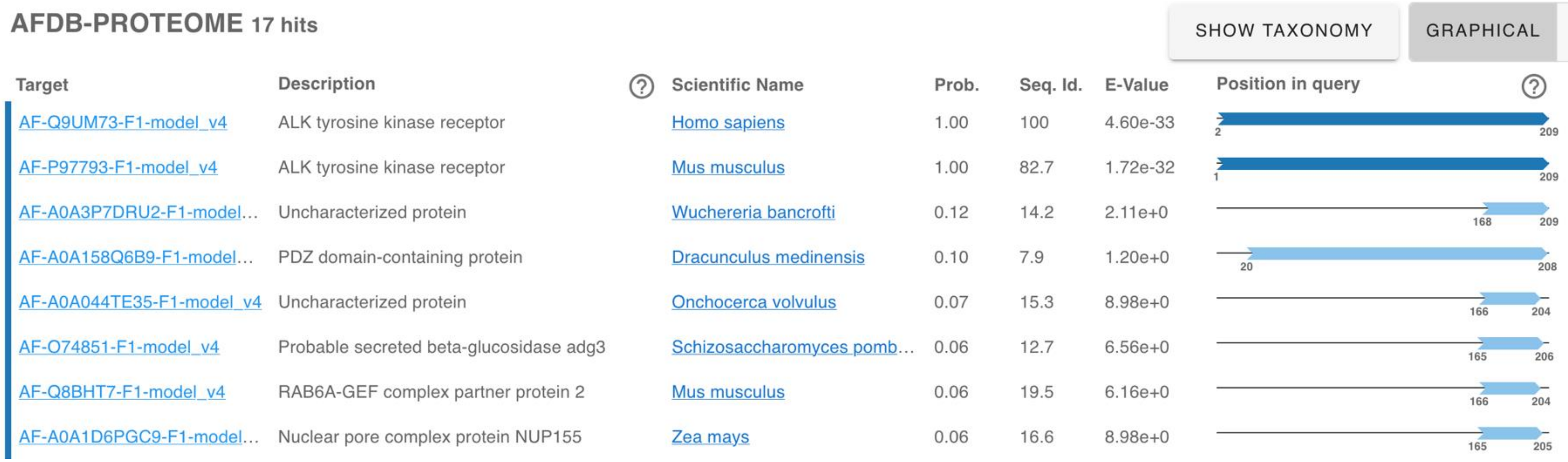
9 colors used within each family for each domain

domains above and helices/TM/Signal/ntTail/ctTail below



17 Novel domains





AFDB-PROTEOME 17 hits

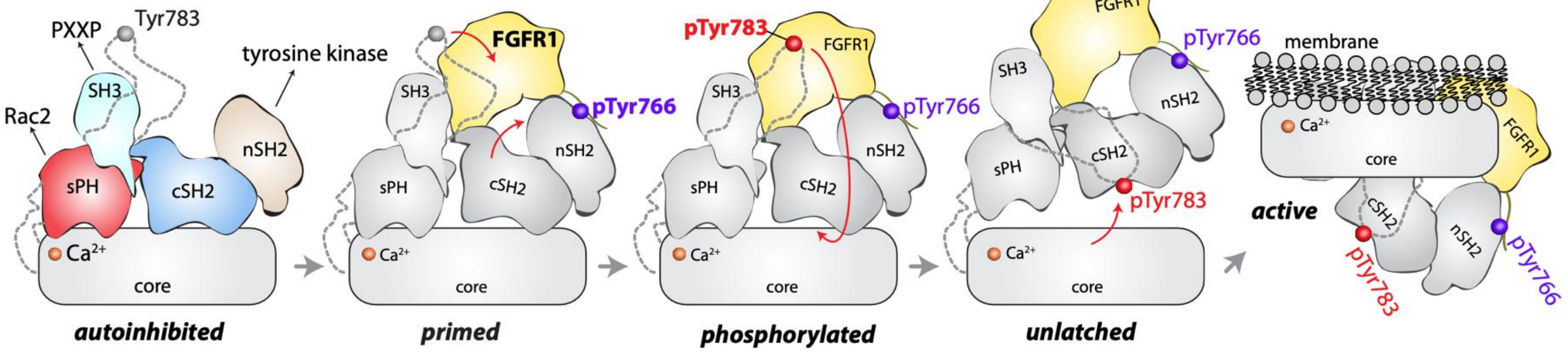
Target	Description	Scientific Name	Prob.	Seq. Id.	E-Value	Position in query
AF-Q9UM73-F1-model_v4	ALK tyrosine kinase receptor	Homo sapiens	1.00	100	4.60e-33	2 209
AF-P97793-F1-model_v4	ALK tyrosine kinase receptor	Mus musculus	1.00	82.7	1.72e-32	1 209
AF-A0A3P7DRU2-F1-model…	Uncharacterized protein	Wuchereria bancrofti	0.12	14.2	2.11e+0	168 209
AF-A0A158Q6B9-F1-model…	PDZ domain-containing protein	Dracunculus medinensis	0.10	7.9	1.20e+0	20 208
AF-A0A044TE35-F1-model_v4	Uncharacterized protein	Onchocerca volvulus	0.07	15.3	8.98e+0	166 204
AF-O74851-F1-model_v4	Probable secreted beta-glucosidase adg3	Schizosaccharomyces pomb…	0.06	12.7	6.56e+0	165 206
AF-Q8BHT7-F1-model_v4	RAB6A-GEF complex partner protein 2	Mus musculus	0.06	19.5	6.16e+0	166 204
AF-A0A1D6PGC9-F1-model…	Nuclear pore complex protein NUP155	Zea mays	0.06	16.6	8.98e+0	165 205

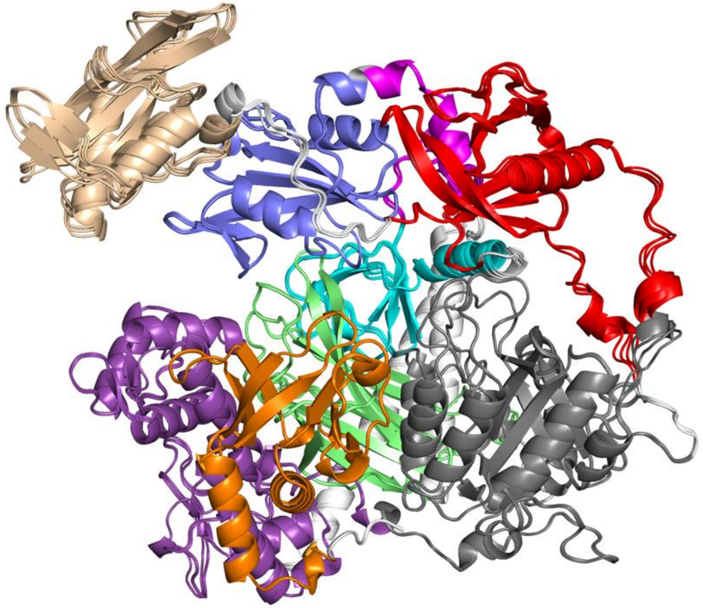
TYR_ALK: Nterminal domain (47-255)
ECOD: not present (earlier ConcanavaliA)
TED: present, “no match”

https://search.foldseek.com/search

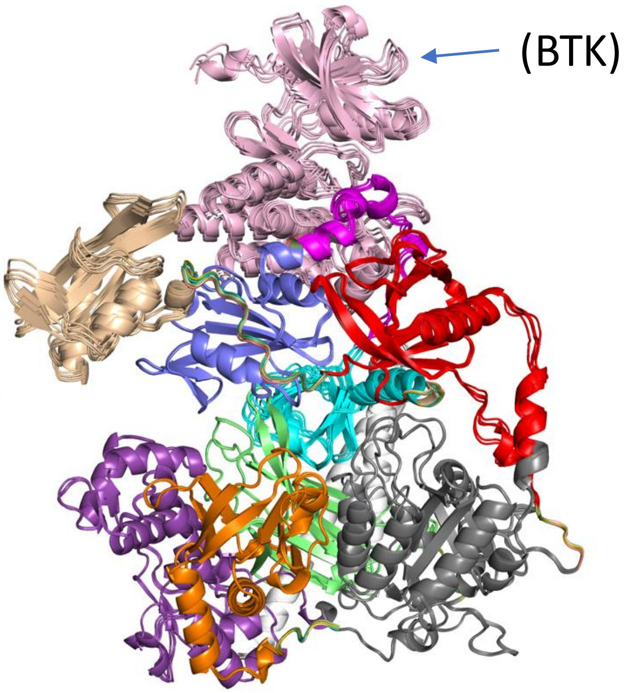
Conformational change: Activation of PLCG2 by BTK and RAC2

With Lynn Wang, FCCC

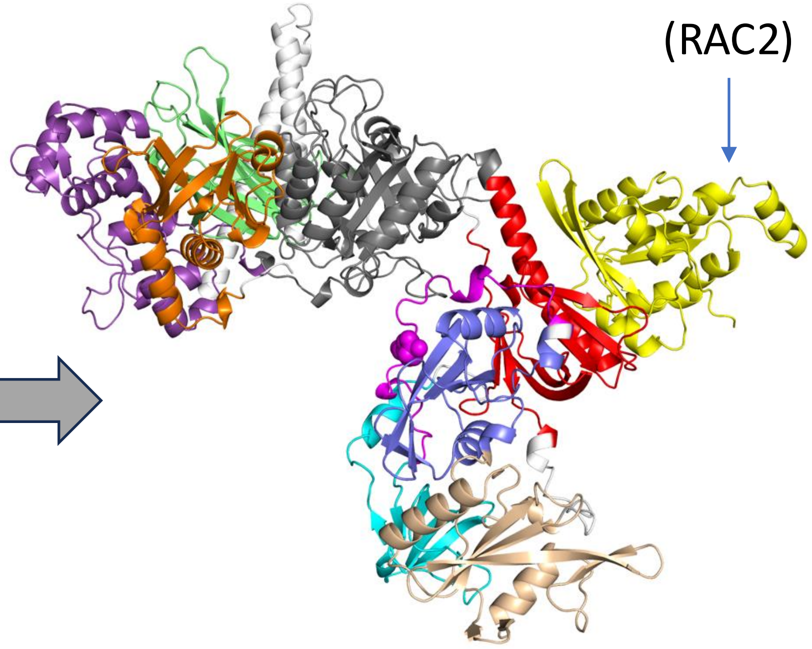




AlphaFold3 models of the
Inactive form of PLCG2



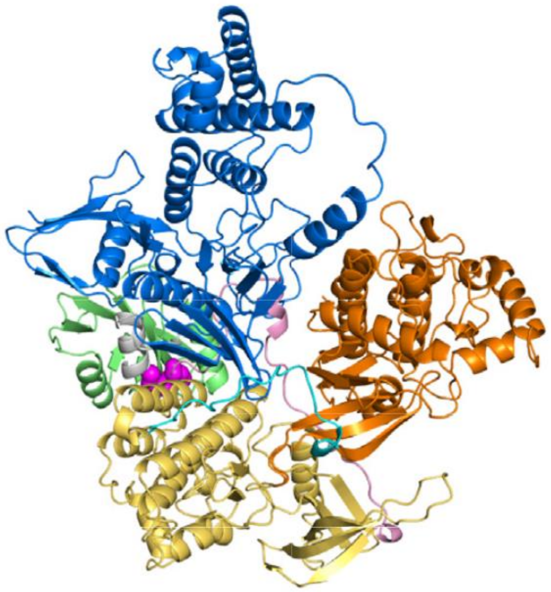
AlphaFold3 models of the
Inactive form of PLCG2 with BTK (pink)



AlphaFold3 models of the
Active form of PLCG2 with RAC2 (yellow)

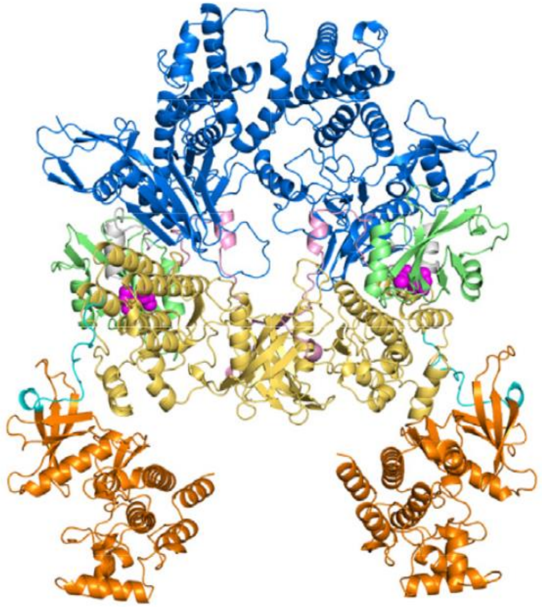
Conformational change: JAK1 activation with AlphaFold2

With Gareth Thomas, Temple U



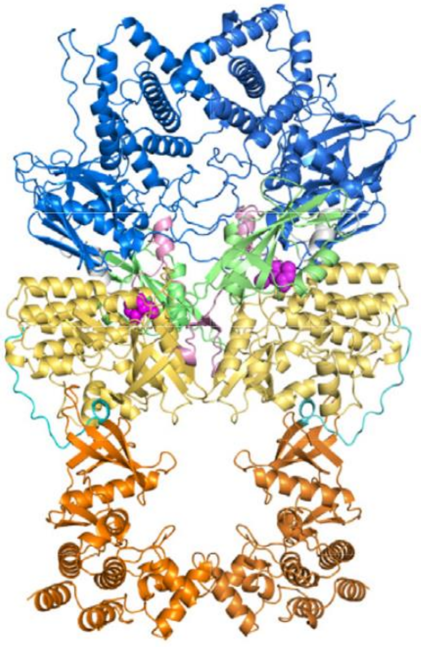
Autoinhibited monomer

⇌



Substrate-competent dimer

⇌



Autophosphorylation dimer

Summary

- AlphaFold and similar programs can model protein domains with near experimental accuracy, as well as PPI, PTMs, DNA/RNA complexes, small-molecular complexes, conformational change.

- Programs struggle with small-molecule protein complexes when the molecules are not similar to anything in the training data (but with exceptions)

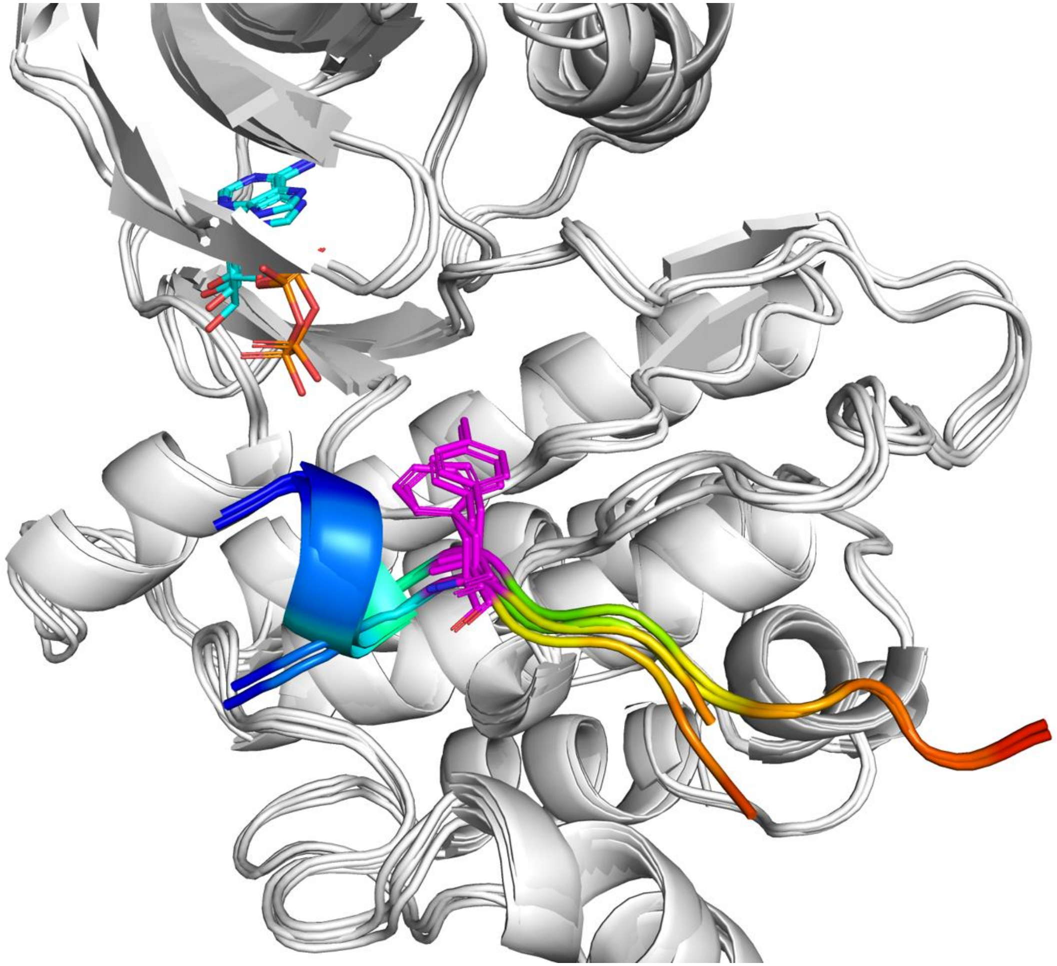
- Requires high sampling on GPU-based computers (FCCC High-Performance Cluster)

- Programs can sample different conformations: with templates (e.g. active and inactive kinases) or without templates (clustering large ensembles of models)

- Interactome studies require specialized scoring metrics that we developed (ipSAE) to distinguish true interactions from potential false positives

- AF-like models provide *hypotheses* and require some experimental validation (mutation studies, over-expression of interacting domains or peptides from interface, SAXS, etc.)

- Protein computational design (RFDiffusion, ProteinMPNN) – AlphaFold-ipSAE acts as validation/filter



Acknowledgments

Qifang Xu
Joan Gizzio
Pragya Priyadarshini
Brianna Jarboe
Xiyao Long

NIH MIRA R35 GM122517
NIH CCSG P30 CA006927
NIH CAP-IT U54 CA272686

https://github.com/DunbrackLab/IPSAE

Demonstrations of AlphaFold Modeling and Visualization

Links:

https://dunbrack.fccc.edu/bioinfo/

PyMOL Demonstration

https://dunbrack.fccc.edu/bioinfo/pymol_install.html

Mac installation

1. Install brew:

```
/bin/bash -c "$(curl -fsSL
https://raw.githubusercontent.com/Homebrew/install/HEAD/install.sh)”
```

2. Install pymol (try first one; if it doesn’t work, try second one)

```
brew install pymol
/opt/homebrew/bin/brew install pymol
```

PyMOL Commands and Menu Items

- Fetch: "fetch 1ol5, type=pdb1"
- The A(ction), S(how), H(ide), L(abel), C(olor) menus
- Display menu (Sequence, Background, etc.)
- Settings menu
- Mouse menu
- Wizard menu (distances, mutations)
- File menu (saving sessions, loading files)
- Action menu: preset, duplicate, delete
- Sele resn, sele resi, sele name, sele chain
- Aligning structures
- Images: button at very top right
- Bioassemblies: Movie->Show all states; split_states 1ol5

Ways to perform AlphaFold2,3 and Boltz-2 Calculations

AlphaFold2 ColabFold Jupyter Notebook (just google “ColabFold”)
https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb

AlphaFold2 ColabFold on local GPU machines:
https://github.com/YoshitakaMo/localcolabfold

AlphaFold3 server:
https://alphafoldserver.com/

AlphaFold3 on local GPU machines:
https://github.com/google-deepmind/alphafold3

Boltz-2 servers
https://build.nvidia.com/mit/boltz2
https://app.tamarind.bio/boltz (also https://app.tamarind.bio/app for other apps)

Example sequences (KRAS-4B, RAF1-RBD)

MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAGQEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHHYREQIKRVKDSEDVPMVLVGNKCDLPSRTVDTKQAQDLARSYGIPFIETSAKTRQGVDDAFYTLVREIRKHKEKMSKDGKKKKKKSKTKCVIM
SKTSNTIRVFLPNKQRTVVNVRNGMSLHDCLMKALKVRGLQPECCAVFRLLHEHKGKKARLDWNTDAASLIGEELQVDFL

ColabFold for AlphaFold2/AFM

https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb
https://colab.research.google.com/signup (getting Colab account will allow jobs to run longer and save zip)

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

86k Accesses | **739** Citations | **415** Altmetric | Metrics

Hints

1. Check AlphaFold Database (AFDB) https://alphafold.ebi.ac.uk/
2. Run more seeds (10) if first run does not produce good scores
3. Run fragments of proteins of interacting domains
4. Run ipSAE score locally or on LIVIA https://flyark.github.io/LIVIA/
5. Check ECOD domains http://prodata.swmed.edu/ecod/

Input

1. Enter sequences separated by ":" (from https://uniport.org)
2. Template mode: "none" or "PDB100"
3. Num-relax: 0, 1, 5

MSA options

1. mmseqs_uniref
2. Paired-unpaired

Advanced settings

1. model_type: alphafold_multimer_v3 (even for single chains)
2. Num_recycles: 6
3. Recycle-earlystop-tolerance: 0.5 (auto)
4. Calc_extra_iptm: check box for actifpTM
5. Num_seeds: 5 models per seed, more is better but might time out
6. Save: do NOT save-all or save-recycles (huge files)
7. Sample settings: reduce max_msa and check Drop-out for more variation

Hit "Run All" at top of page. Zip file will be saved with coordinates and json files

colabfold_batch

Put in file “run_smo” (all one line)

```
colabfold_batch --model-type alphafold2_multimer_v3 --zip
--amber --use-gpu-relax --num-seeds 10 --num-recycle 10
--recycle-early-stop-tolerance 1.0 smo_pka.fasta  SMO/ > SMO.out
```

Type these commands:

```
chmod ugo+x run_smo
nohup ./run_smo
```

File smo_pka.fasta:

```
>SMO_PKA_mouse_F577A

RGAASSGNATGPGPRSAGGSARRSAAVTGPPPPLSHCGRAAPCEPLRYNVCLGSVLPYGATSTLLAGDSDSQEEAHGKLVLWSGLRNAPRCWAVIQPLLCAVYMPKCENDRVELPSRTLCQATRGPCAIVERERGWPDFLRCTPDSDF:

GNAAAAKKGSEQESVKEFLAKAKEDFLKKWESPAQNTAHLDQFERIKTLGTGSFGRVMLVKHKETGNHYAMKILDKQKVVKLKQIEHTLNEKRILQAVNFPFLVKLEFSFKDNSNLYMVMEYVPGGEMFSHLRRIGRFSEPHARFYAAQIVLTFEYLHSL
DLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWTLCGTPEYL


>SMO_PKA_mouse_R569E

VILHPNETIFNDFCKKSTTCEVLKYNTCLGSPLPYTHTSLILAEDSETQEEAFEKLAMWSGLRNAPRCWAVIQPLLCAVYMPKCENGKVELPSQHLCQATRNPCSIVERERGWPNFLKCENKEQFPKGCQNEVQKLKFNTSGQCEAPLVKTDTLLCTFFT
LATFLASMMDVGRTAVVPRADGRRGVQIHSRTNLMDAELLDADSDF:

GNAPTAKNKGNEMESVKEFLAKAKEDFLKKWENPAQNTASLDHFERLKTLGTGSFGRVMLVKHKESGQHFAMKILDKQKVVKLKQIEHTLNEKRILQAVSFPFLVRLEHSFKDNTNLYMVMEYVPGGEMFSHLRRIGRFSEPHARFYAAQIVLTFEYLHS
LDLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWTLCGTPEY
```

AF3

Running AlphaFold3 on AF3 server

1. Go to https://alphafoldserver.com and sign in with gmail
2. Put in sequences, ligands, ions, etc.
3. Hit “Continue and preview job”
4. Name job (important)
5. Download zip file when job finishes

Hints

1. Summary_confidences file for ipTM (esp. chain_pair_iptm)
2. Look at PAE plot for protein-protein interaction
3. Run ipSAE locally or on LIVIA: https://flyark.github.io/LIVIA/

Running Boltz-2
https://app.tamarind.bio/boltz

Running Boltz-2

1. Set up account: https://www.tamarind.bio/ (free)
2. Go to https://app.tamarind.bio/boltz
3. Input sequence(s). Choose number of copies
4. Add modification button: put in PTM position and three-letter code (SEP, TPO, PTR for phospho Ser, Thr, Tyr)
5. Hit "+" button to add igand: choose "CCD code" or "SMILES" and input CCD or SMILES (see hints). Choose number of copies.
6. Optional: "Add restraint" between residues
7. Advanced settings: number of jobs, recycles, PDB or mmCIF format, choose or upload template

Hints

1. Get 3-5 character CCD code from PDB search: https://www.rcsb.org/ (e.g. GTP, MG, SLI)
2. Get any SMILES from PDB ligand page, e.g. https://www.rcsb.org/ligand/GTP or PubChem
3. Run multiple jobs and choose mmCIF format (in "Advanced settings")

Installing and running ipSAE

1. Go to https://github.com/DunbrackLab/IPSAE
2. Download zip file and double-click on it
3. Move to desired location (location/)
4. Add to .zshrc or .bashrc: alias ipsae=‘python3 location/ipsae.py’
5. Run: source ~/.zshrc (or .bashrc)
6. Run:

```
ipsae    file.json    file.pdb    10 15
ipsae    file.json    file.mmcif  10 15
```

OpenFold3 and Other Servers

https://colab.research.google.com/github/aqlaboratory/openfold/blob/main/notebooks/OpenFold.ipynb

https://neurosnap.ai/service/OpenFold3%20%28AlphaFold3%29

https://www.tamarind.bio/tools/openfold

https://build.nvidia.com/openfold/openfold3

https://openfold-3.readthedocs.io/en/latest/

https://protenix-server.com/login

Links

ColabFold Server (AF2 and AF2-multimer-v3)
https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb

AlphaFold3 Server
https://golgi.sandbox.google.com/

Chai Lab Webserver
https://lab.chaidiscovery.com/dashboard

OpenFold webserver
https://colab.research.google.com/github/aqlaboratory/openfold/blob/main/notebooks/OpenFold.ipynb

Boltz-2 webserver (TamarindBio)
https://app.tamarind.bio/boltz

Boltz-2 webserver (NVIDIA)
https://build.nvidia.com/mit/boltz2

LIVIA Viewer
https://flyark.github.io/**LIVIA**/universal.html

Local interaction score (colab)
https://github.com/flyark/AFM-LIS?tab=readme-ov-file

ipSAE github
https://github.com/DunbrackLab/IPSAE

Links:

https://dunbrack.fccc.edu/bioinfo/

Questions:

roland.dunbrack@fccc.edu