Protein structural analysis and GPU-enabled tools for predicting protein structure, function, and interactions

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

https://dunbrack.fccc.edu/bioinfo

18 Concepts in Protein Structure and AI-based Tools

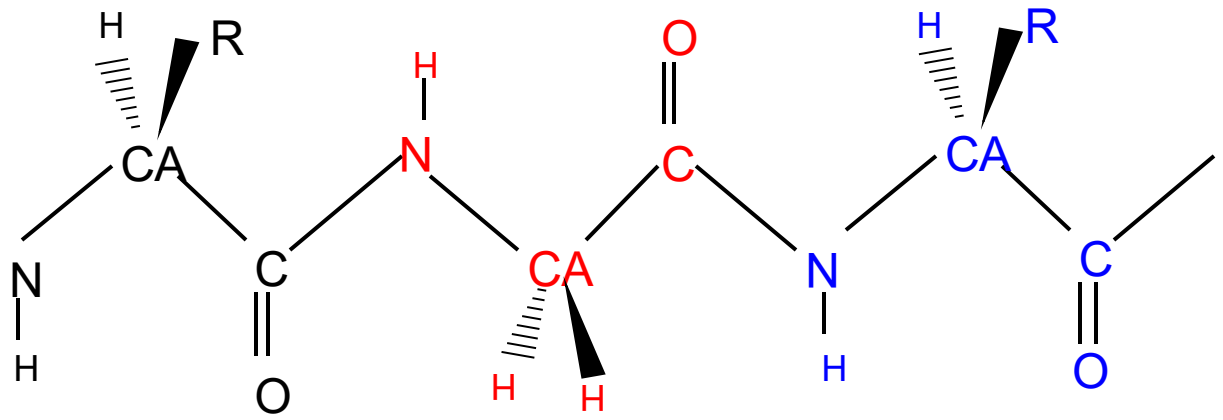
1. Proteins are heteropolymers of 20 amino acids, each a different side group off the backbone (N-CA-C)
2. Their degrees of freedom (structural change) are dihedral angles
3. They contain hydrophobic, charged, and polar side groups, and the backbone is very polar.
4. Hydrophobic side-chains need to be buried, mostly in contact with other hydrophobic groups, because otherwise water interacting with the hydrophobic group loses favorable water-water hbonds and entropy.
5. The polar groups need to be exposed to water (very polar) OR buried when forming intra or intermolecular hydrogen bonds.
6. Proteins have evolved to fold into structures ("folds", "domains") that create surfaces, pockets, grooves to perform binding, catalysis, localization
7. How? By burying (conserved) hydrophobic side-chains, forming backbone-backbone hydrogen bonds in alpha helices and beta sheets, forming side-chain.backbone and side-chain/side-chain hydrogen bonds.
8. Proteins are dynamic, not rocks, as shown by experiments and MD simulations, and that dynamics is *always* relevant to function.
9. About 3000 protein fold families (related by evolution) covers most proteins
10. Some human proteins are entirely disordered (IDPs) and most human proteins have multiple folded domains connected by linkers, which can be disordered (IDRs) and which control domain-domain interactions, act as scaffolds for other proteins, and other functions.
11. Structures are determined by crystallography, NMR, and cryo-EM spectroscopy, which don't see IDRs/IDPs.
12. Structures show assemblies within crystals (more explicit in cryo-EM), including homooligomers which are usually symmetric assemblies and distinct from asymmetric units in crystals.
13. Structure is important for interpreting/predicting function, conservation, inherited/somatic/experimental mutational effects.

Continued...

14. But now we have Alphafold2 and Alphafold3 for predicting structure – fast and accurate structure prediction in the absence of experimental structures and sometimes capable of assessing dynamics or alternate conformations and associations of domains.
15. AF2 and AF3 are deep-learning neural networks which take inputs (sequences and sequence alignments) and output predicted structures via a small number of “tricks”.
16. AF2 and AF3 output “scores: like pLDDT, PAE, ipTM.
17. We developed ipSAE for protein-protein and domain-domain, domain-peptide interactions, which fixes problems in ipTM and works well for miniprotein-binder design.
18. AF is very good for fixing understanding of domain architecture of large human proteins as demonstrated by our analysis of full-length structure prediction of human proteins that contain protein kinase domains.

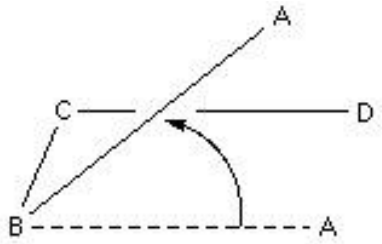
1. Proteins are heteropolymers of 20 amino acids, each a different side group off the backbone (N-Cα–C)

$(NH\text{-}C\alpha(R)\text{-}CO)^n$

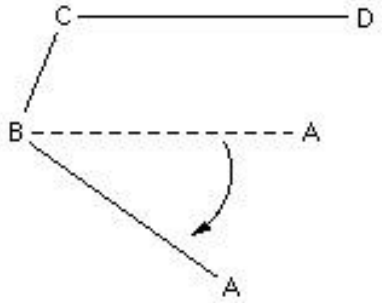


2. Their degrees of freedom (structural change) are dihedral angles

Bond lengths and bond angles are relatively fixed



Positive dihedral



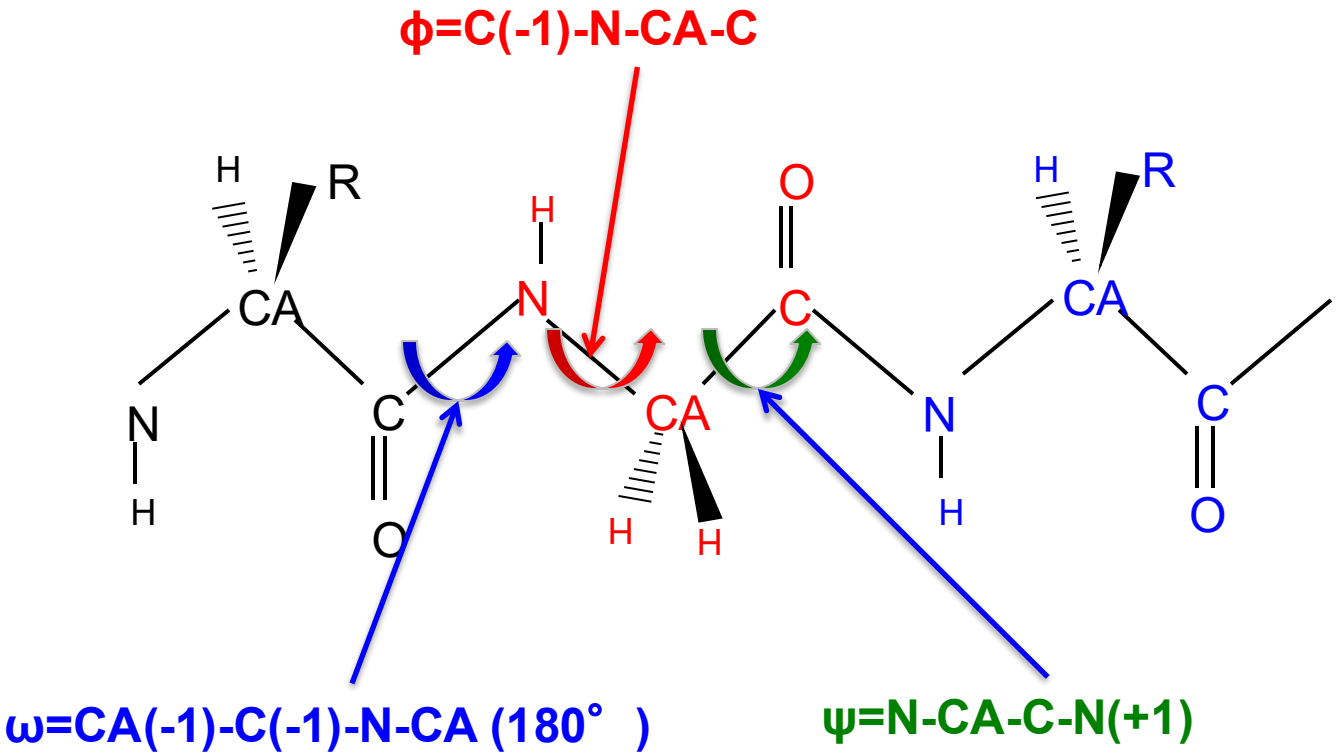
Negative dihedral

0°

180°

Protein backbone dihedral angles

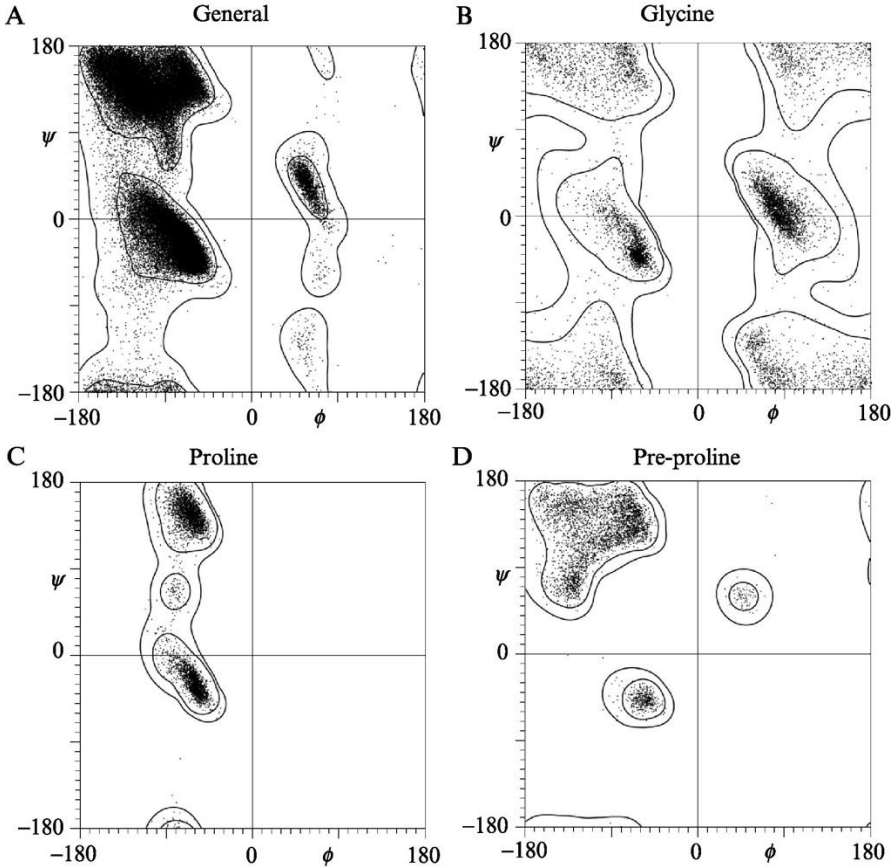
Bond lengths (N-Cα) and bond angles (N-Cα-C) are relatively fixed



The Ramachandran Map
ϕ,ψ Distributions differ by residue type

Points outside the contours are probably wrong

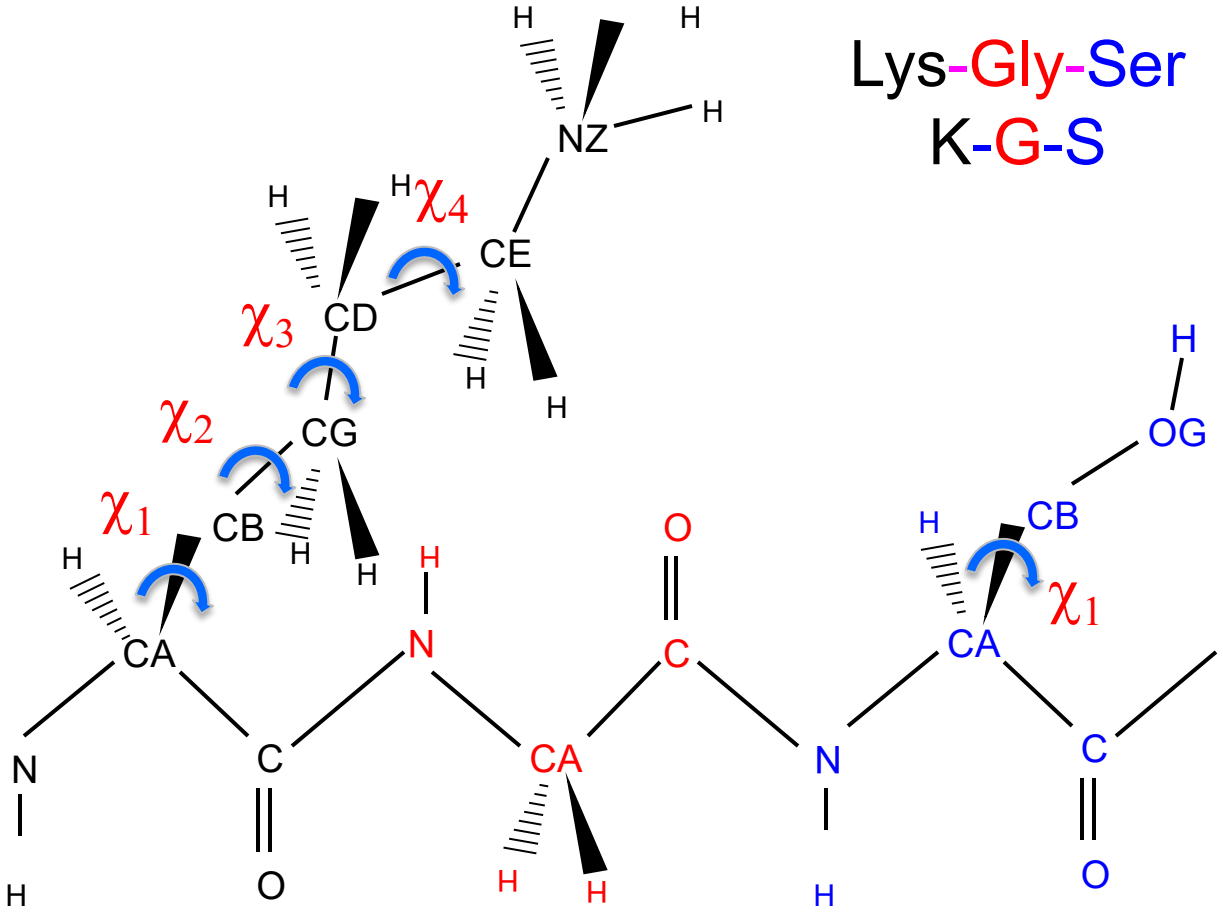
http://molprobity.biochem.duke.edu



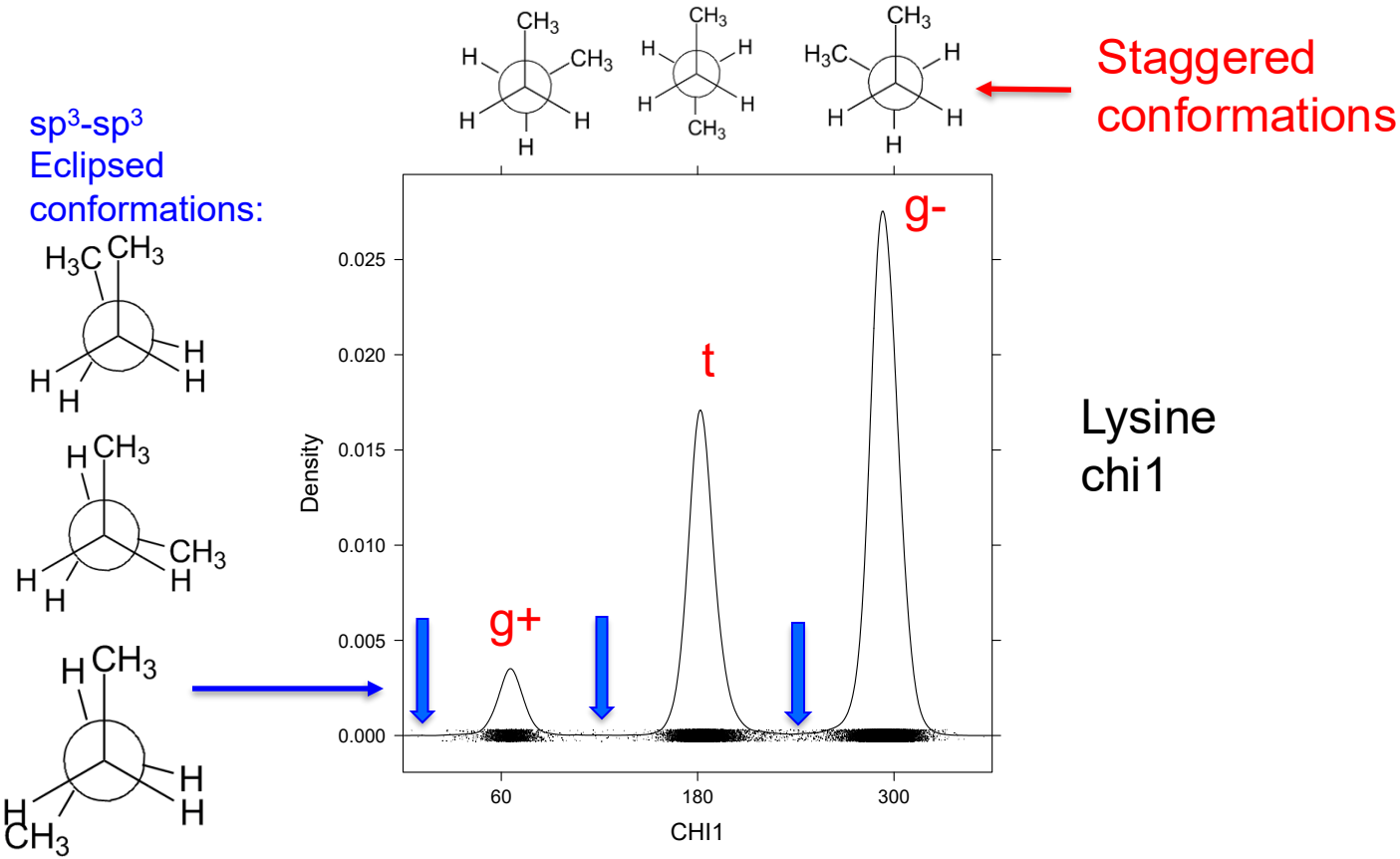
Gly is "flexible"

Pro is "rigid"

Side-chain Degrees of Freedom

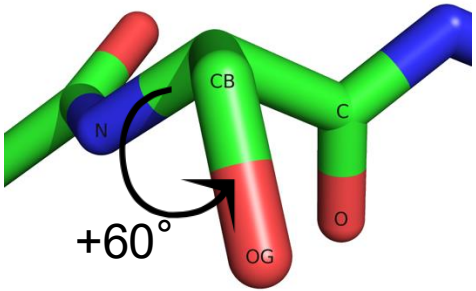


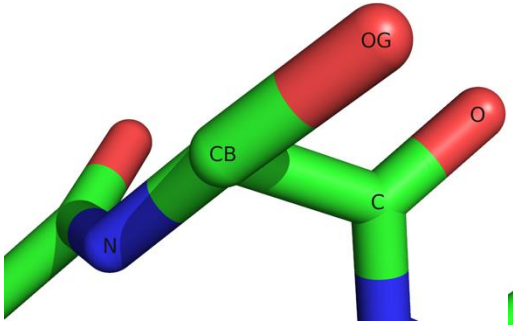
Side-chain Degrees of Freedom Are (mostly) Discrete -- Rotamers

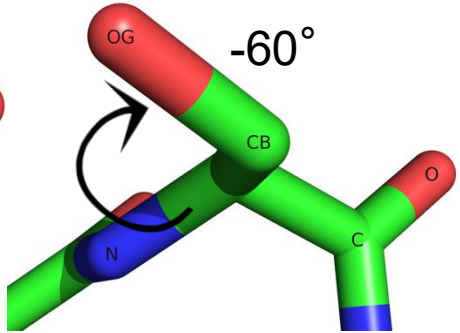


Side-chain rotamers

χ_1=N-CA-CB-CG
χ_1=N-CA-CB-CG1 (VAL, ILE)
χ_1=N-CA-CB-OG (SER)
χ_1=N-CA-CB-OG1 (THR)
χ_1=N-CA-CB-SG (CYS)

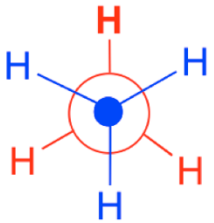


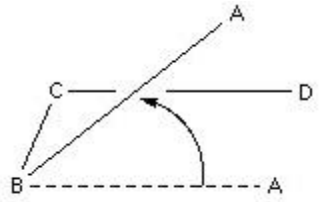




g+ (+60°) Trans (180°) g- (-60°)

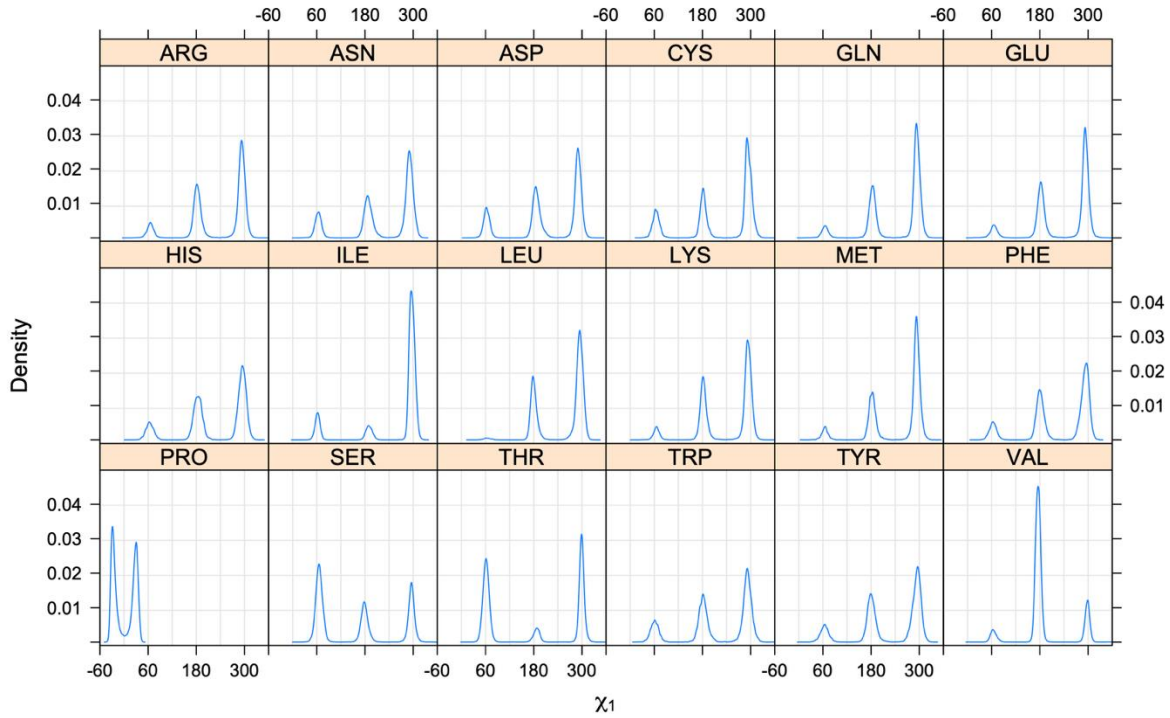
the Newman projection





Positive dihedral

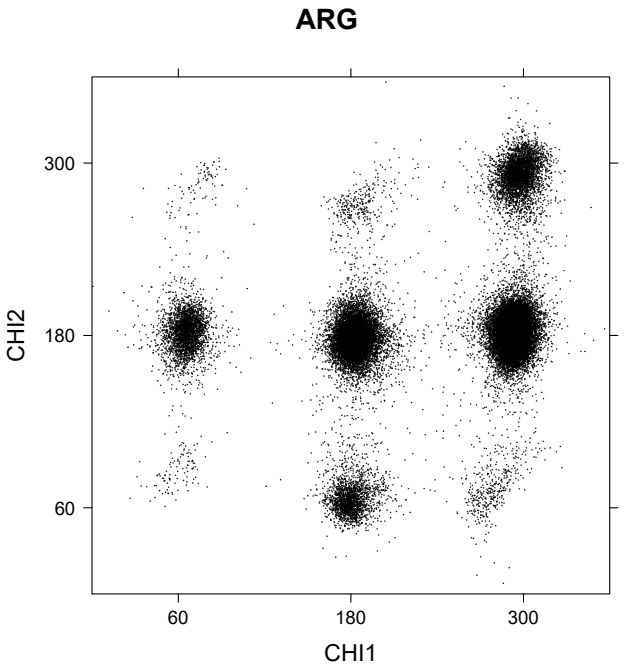
Chi1 is always rotameric



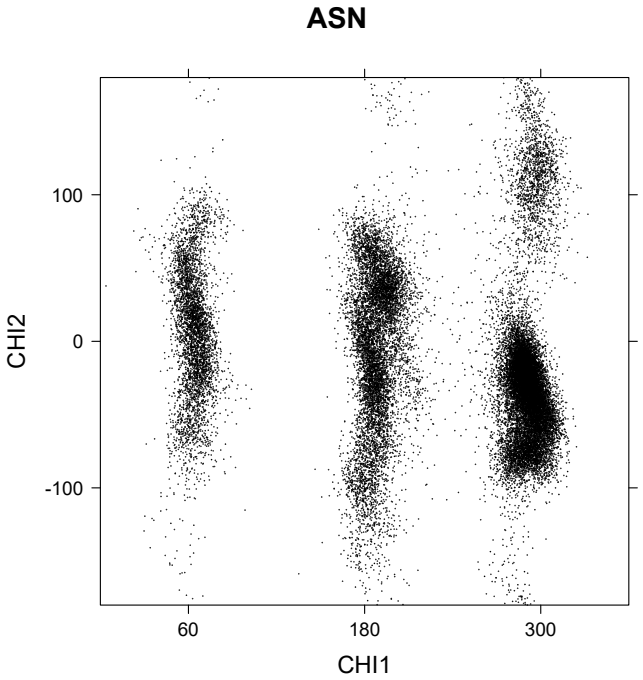
https://dunbrack.fccc.edu/bbdep2010/ConformationalAnalysis.php

Rotameric and Non-rotameric Degrees of Freedom

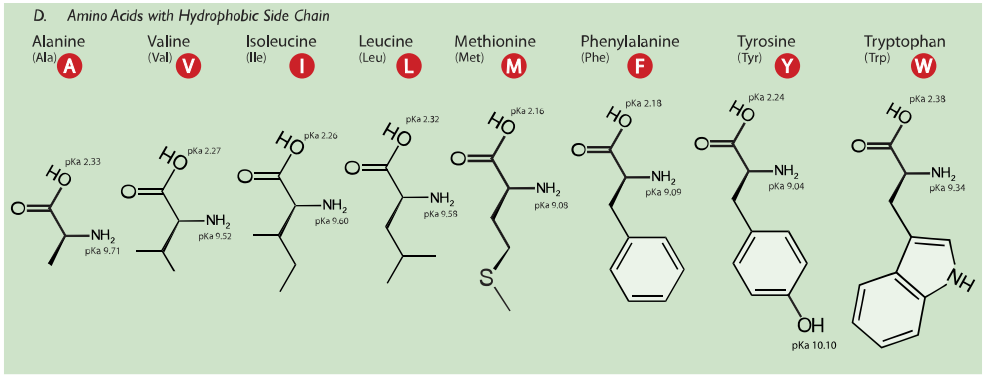
χ_1 rotameric
χ_2 rotameric

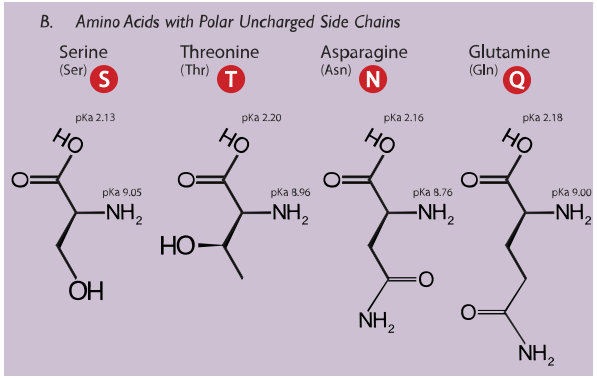


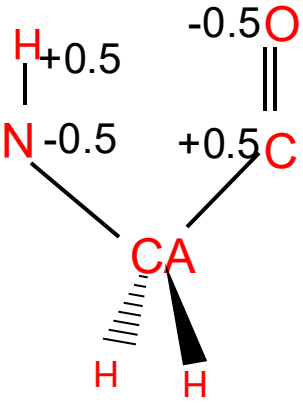
χ_1 rotameric
χ_2 non-rotameric

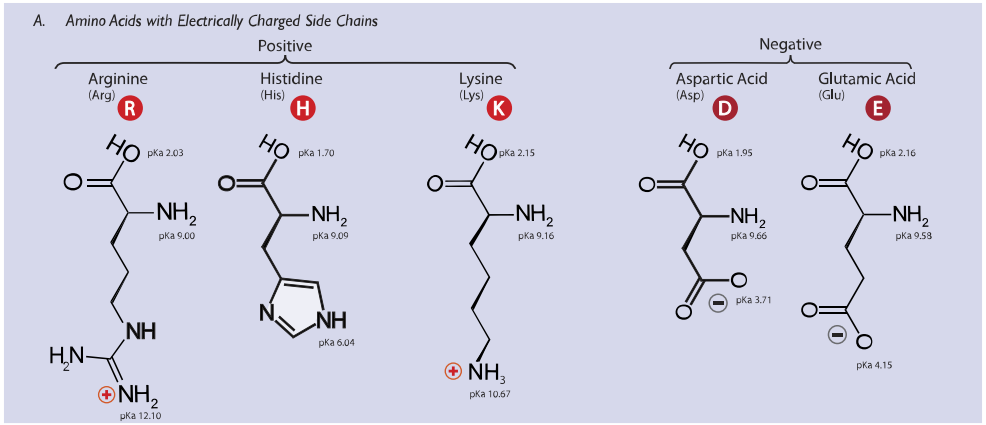


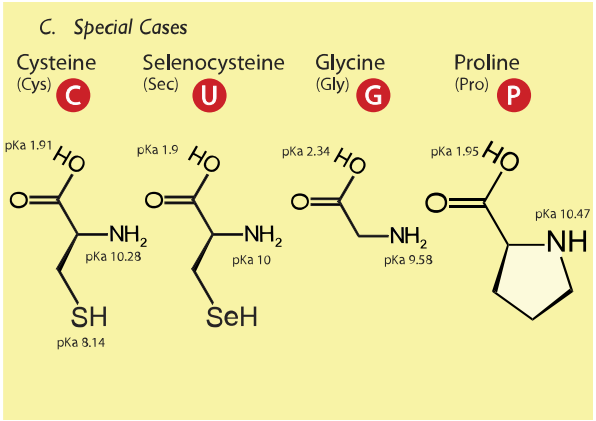
3. Amino acids contain hydrophobic, charged, and polar side groups, and the backbone is very polar.



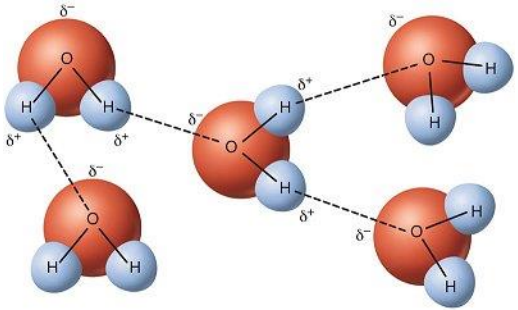






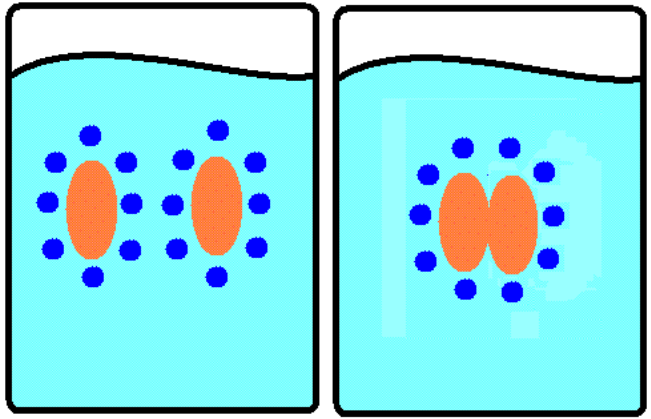


4. Hydrophobic side-chains need to be buried, mostly in contact with other hydrophobic groups, because otherwise water interacting with the hydrophobic group loses favorable water-water hydrogen bonds and entropy.



Water is very polar
Maintains ~3.8 h-bonds per molecule.

High entropy

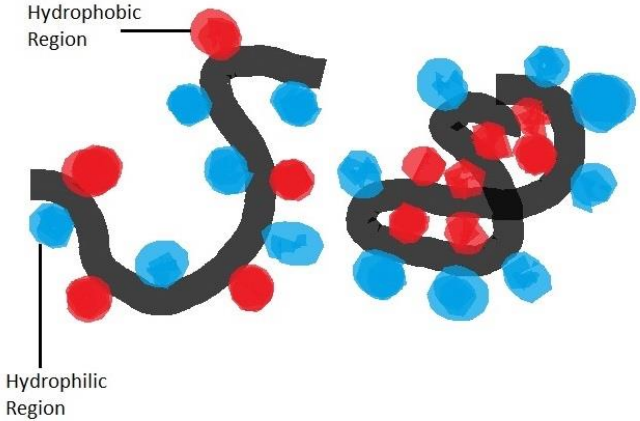


Exposed hydrophobic surface:

Losing h-bonds
Lose entropy

Buried hydrophobic surface

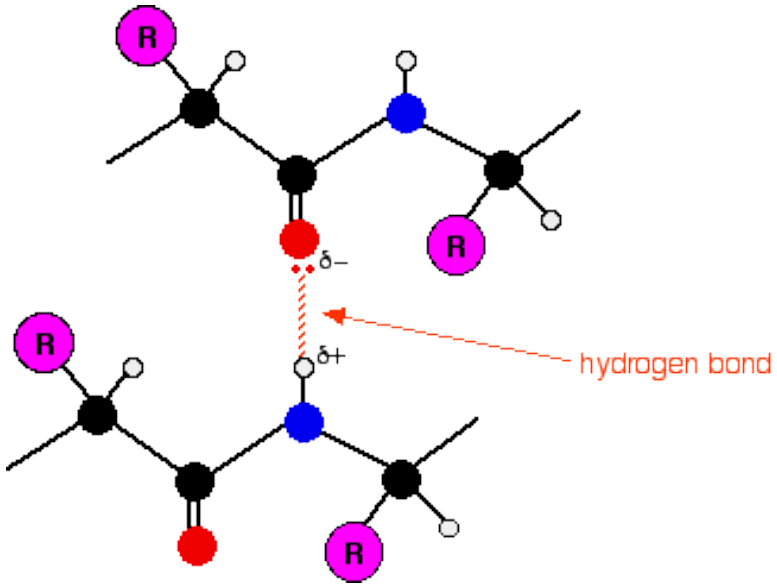
Losing fewer h-bonds
Lower less entropy

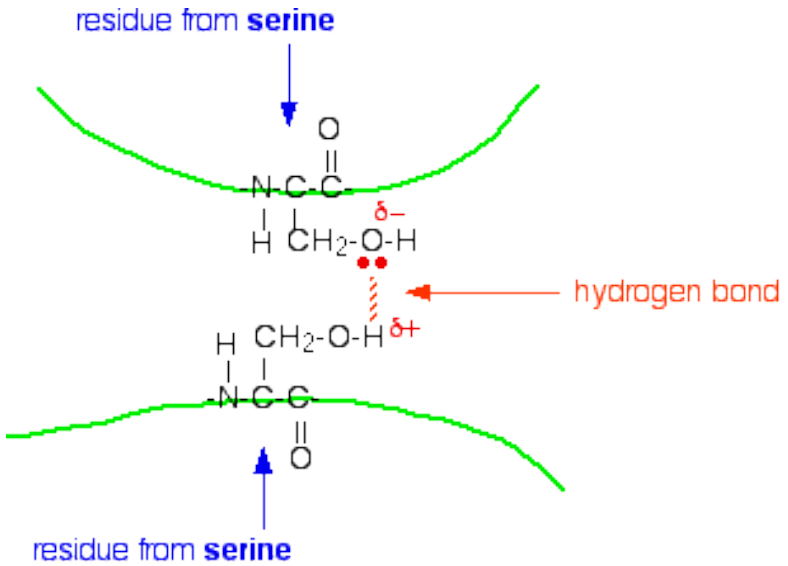


Unfolded protein

Folded protein

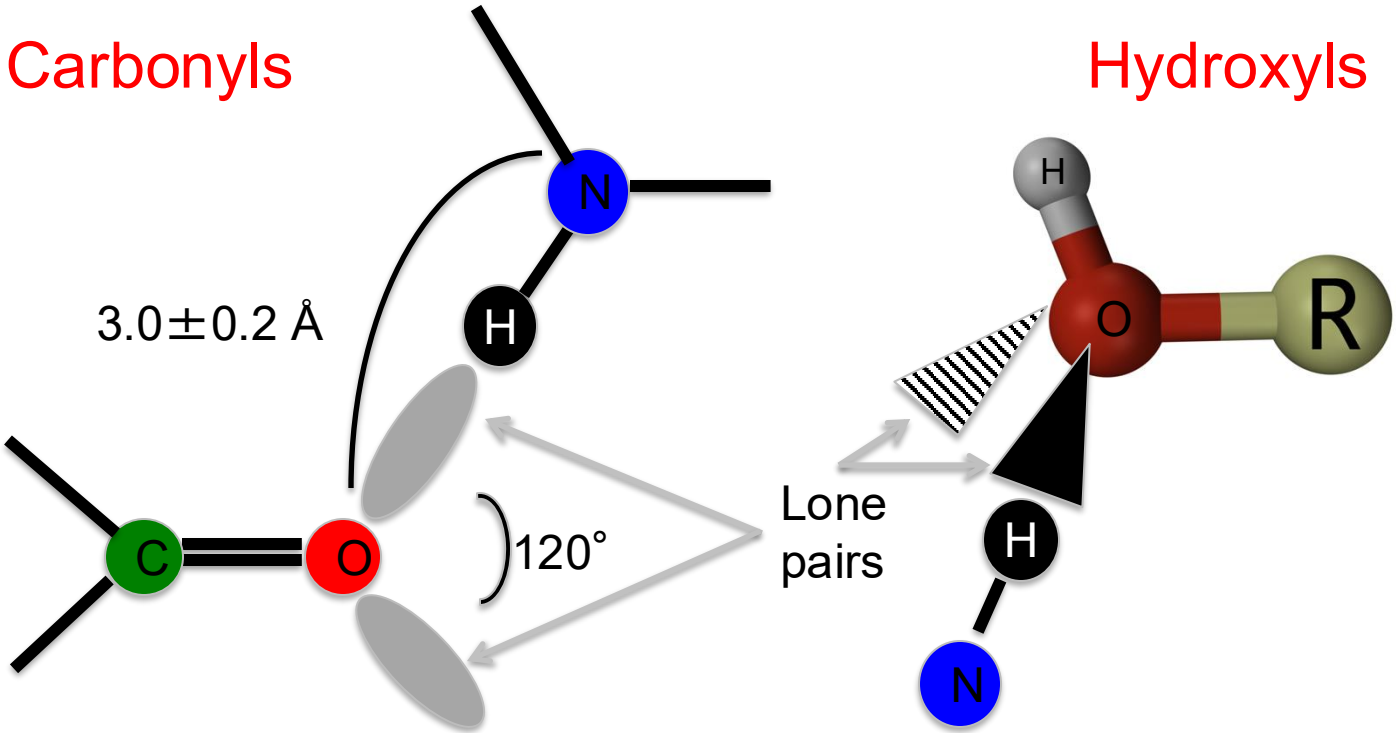
5. The polar groups need to be exposed to water (very polar) OR buried when forming intra or intermolecular hydrogen bonds.





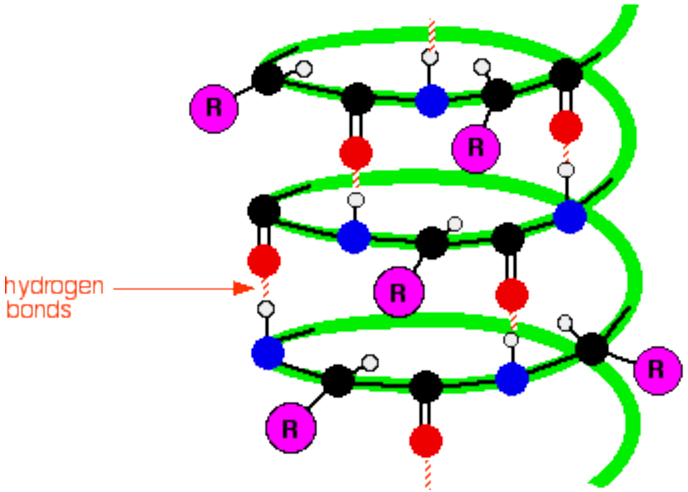
Hydrogen Bonds

Directionality Is Important



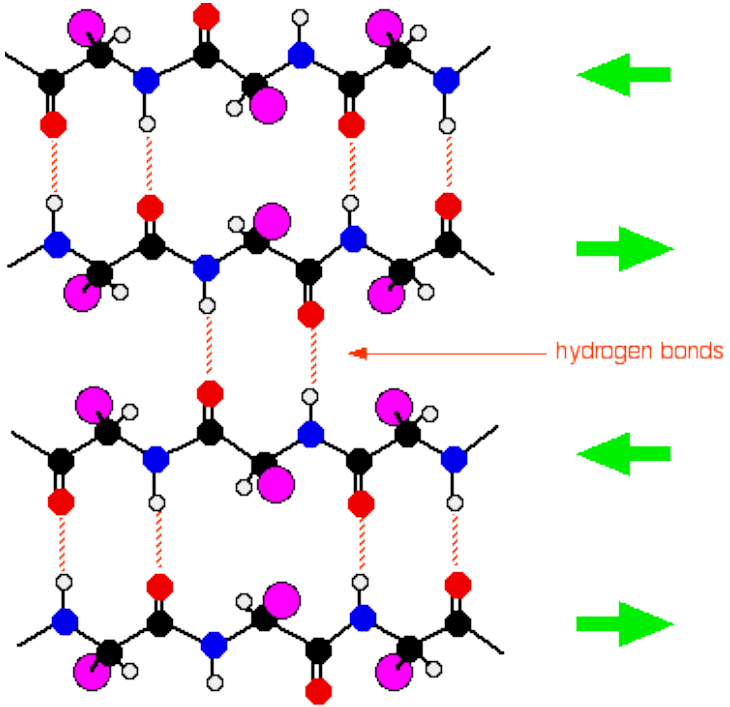
Burying Polar Groups by Means of Regular Secondary Structure

Alpha Helix



Hbond between C=O of residue *i* and HN of residue *i*+4

Beta Sheets

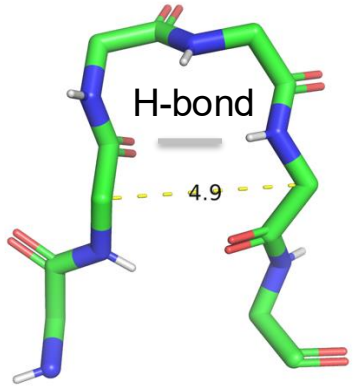


Beta Turns in Protein Loops

Protein loops are:
More like this …than this

Classical nomenclature ϕ,ψ

	Residue 2		Residue 3		
I	-60°	30°	-90°	0°	
I′	60°	-30°	90°	0°	
II	-60°	120°	80°	0°	
II′	60°	-120°	-80°	0°	
VIa1	-60°	120°	-90°	0°	(cisPro)
VIa2	-120°	120°	-90°	0°	(cisPro)
VIb	-135°	135°	-75°	160°	(cisPro)
VIII	-60°	-30°	-120°	120°	
IV	Everything else				

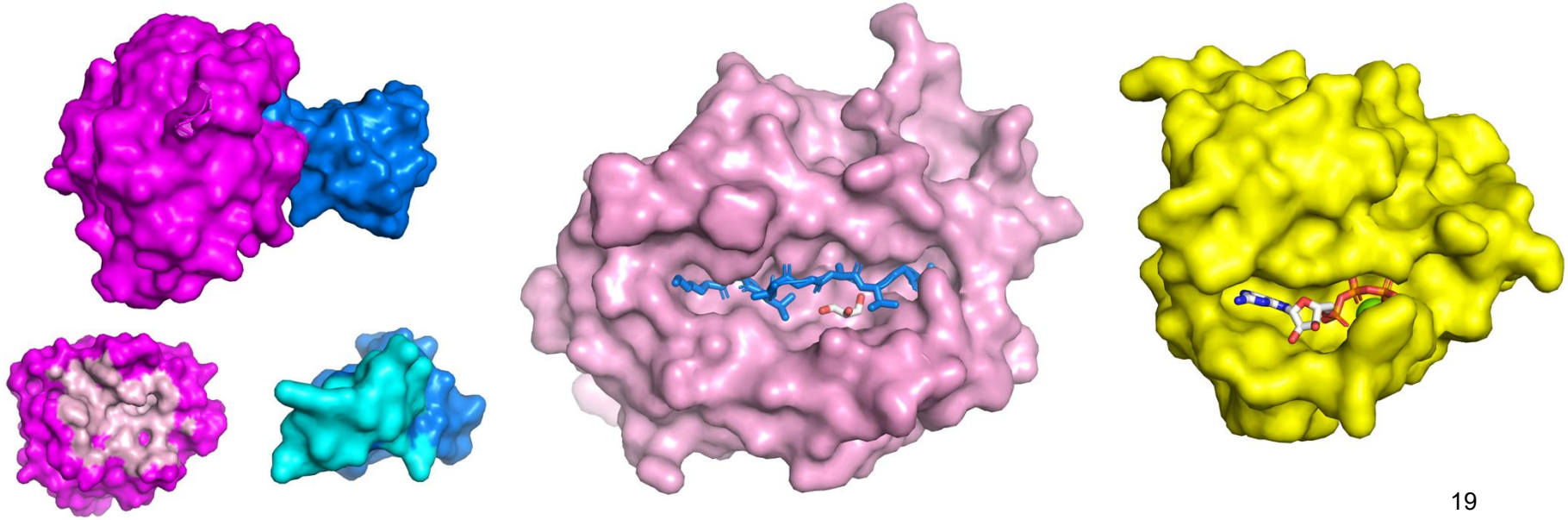


Definition:
(1) CA1-CA4 distance ≤ 7 Å
(2) not helix

https://github.com/DunbrackLab/Identify_18_BetaTurn_Types

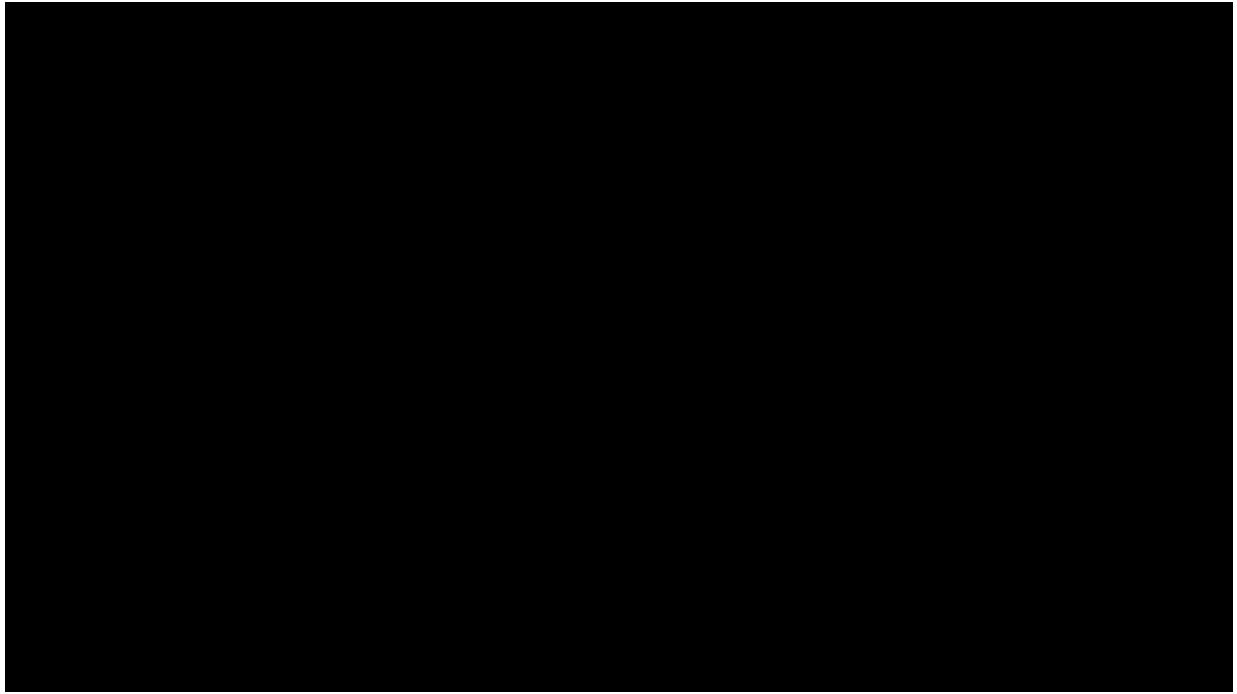
6. Proteins have evolved to fold into structures (“folds”, “domains”) that create surfaces, pockets & grooves to perform binding, catalysis, localization, transport.

Surfaces have complementary hydrophobicity, charges, and H-bond donors & acceptors for binding partners.

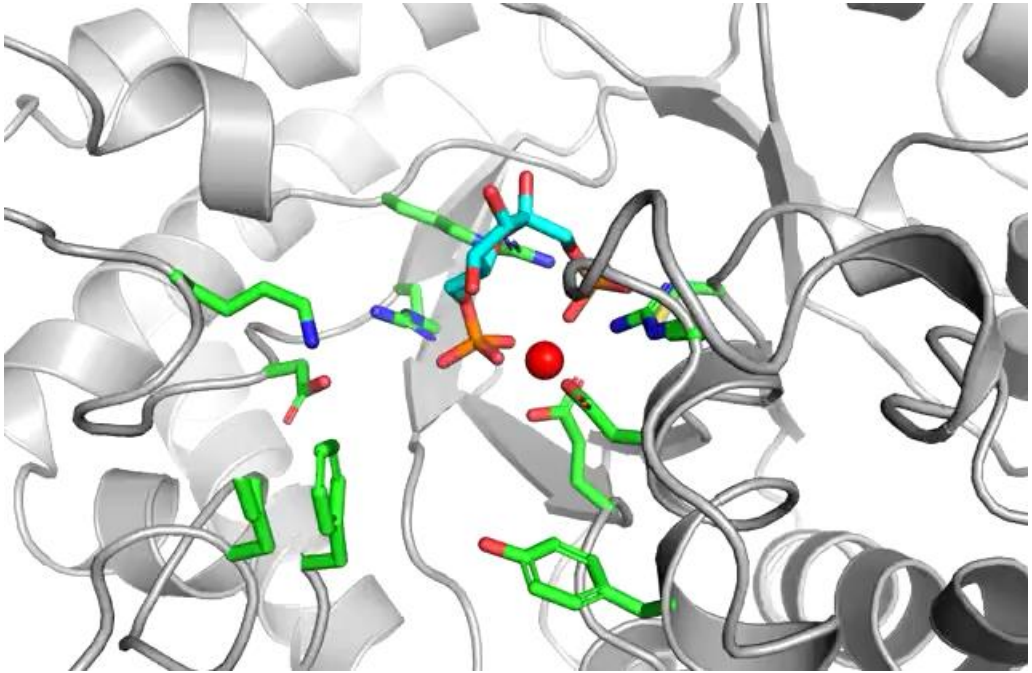


7. How? By burying (conserved) hydrophobic side-chains, forming backbone-backbone hydrogen bonds in alpha helices and beta sheets, forming side-chain.backbone and side-chain/side-chain hydrogen bonds. And exposing surface shapes and physical attributes for binding.

Hbond groups must be satisfied either internally or with solvent or binding partners

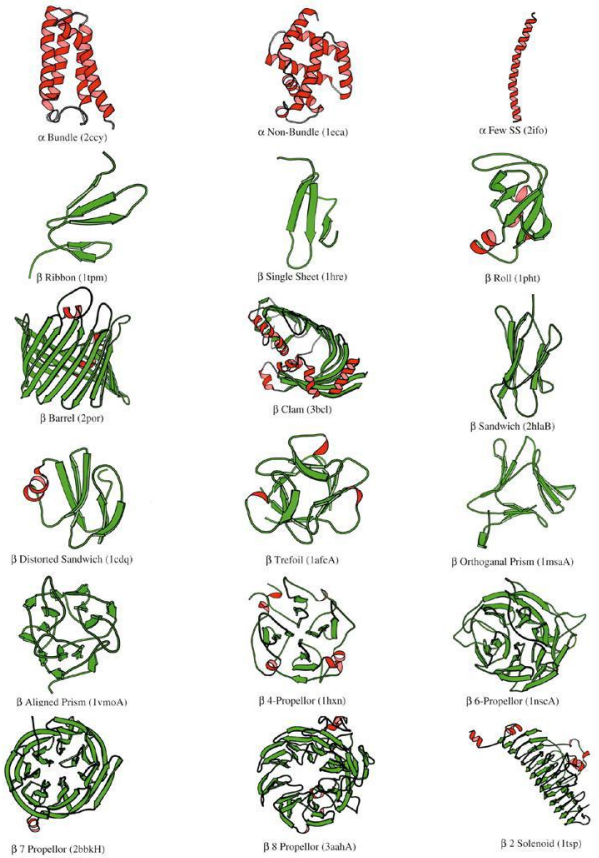


8. Proteins are dynamic, not rocks, as shown by experiments and “molecular dynamics” simulations, and that dynamics is *always* relevant to function.

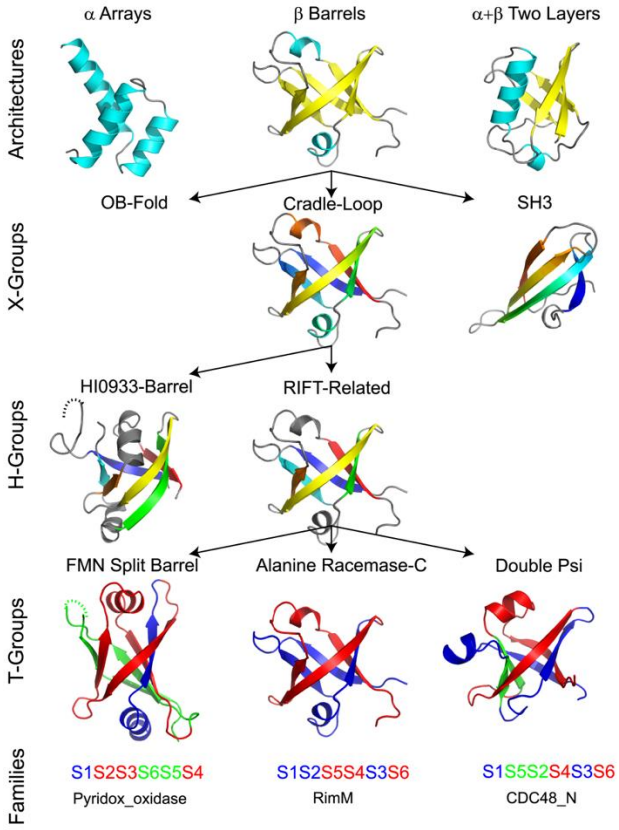


9. A set of ~4000 protein fold families (related by evolution) covers most proteins

Can be classified hierarchically (TED, ECOD)



ECOD: Evol. Classification of Domains



Architecture (all-α, all-β, mix)

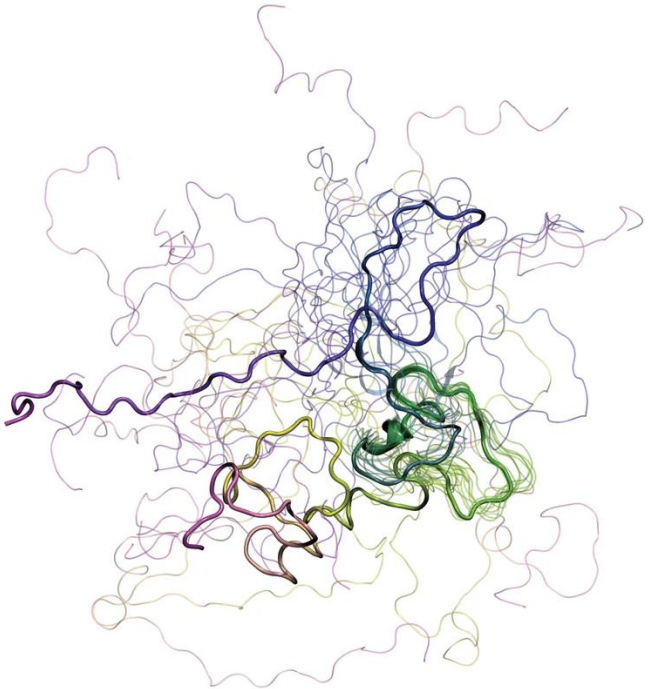
X-group (rough fold)

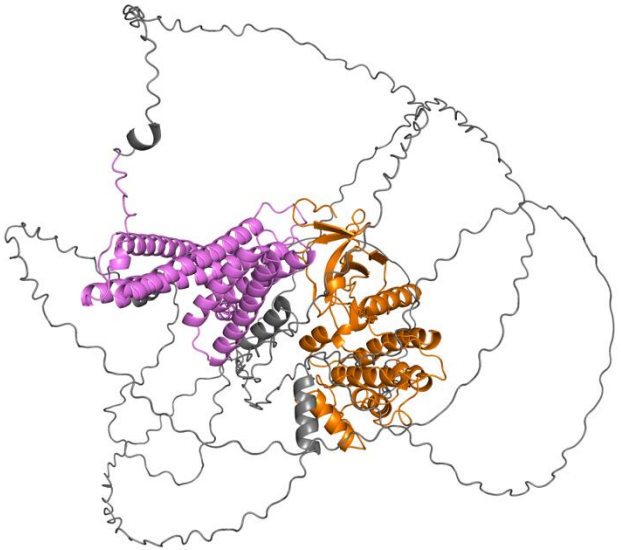
H-group (evol.-related)

T-group (Exact topology)

Protein families

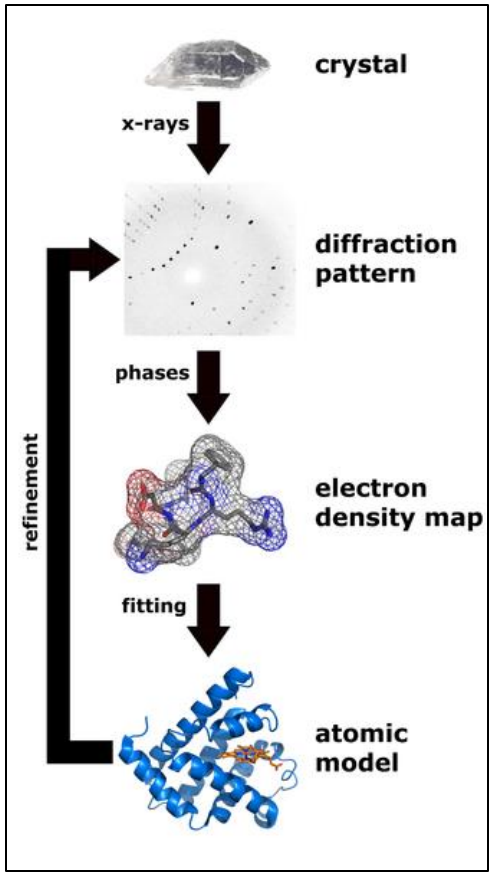
10. Some human proteins are entirely disordered (IDPs) and most human proteins have multiple folded domains connected by linkers, which can be disordered (IDRs) and which control domain-domain interactions, act as scaffolds for other proteins, and other functions

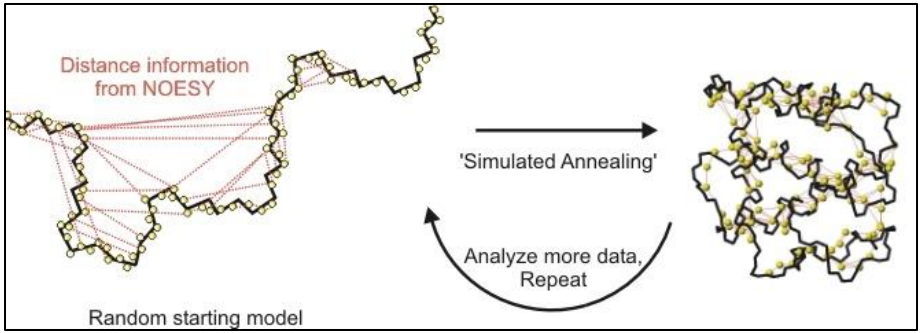


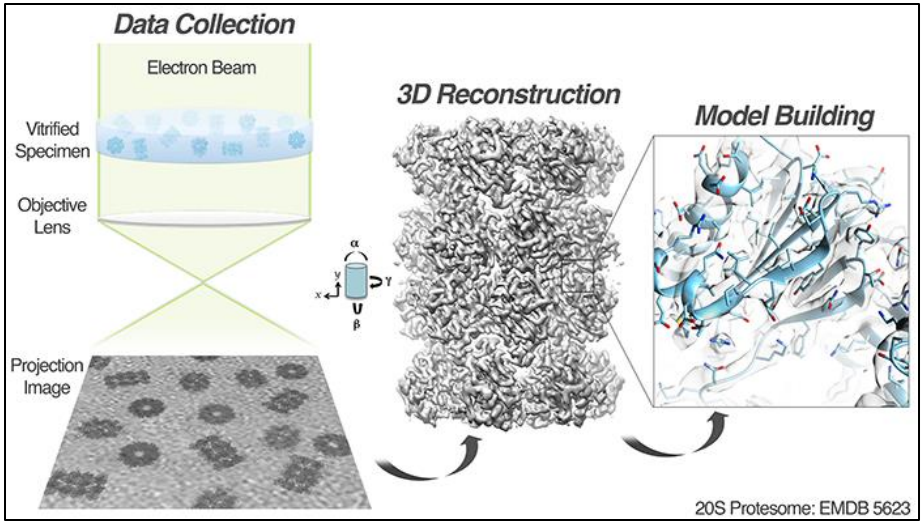


ULK1

11. Structures are determined by crystallography, NMR, and cryo-EM spectroscopy, which don't see IDRs/IDPs.

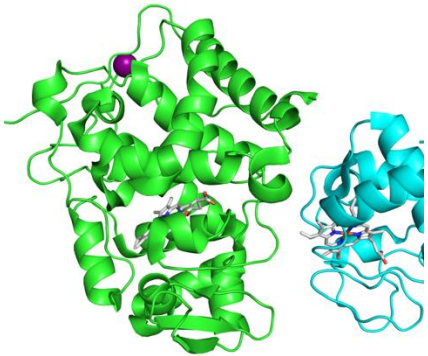




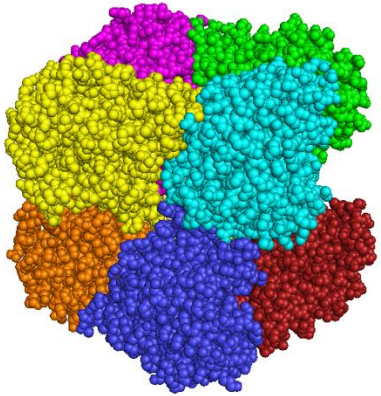


12. Structures show assemblies within crystals (more explicit in cryo-EM), including homooligomers which are usually symmetric assemblies and distinct from asymmetric units in crystals.

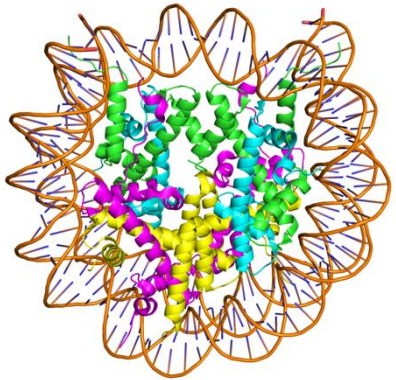
- Many proteins function as oligomers
- Observed from protein crystals
- Protein crystal structures are stored in Protein Data Bank (PDB).



Heterodimer of Peroxidase and Cytochrome C

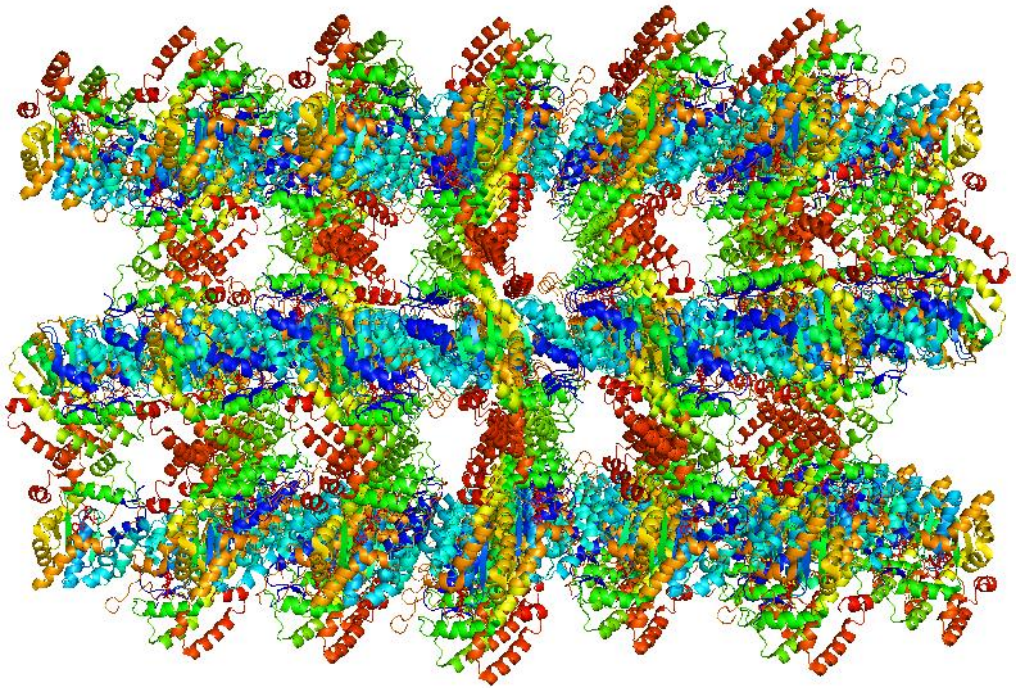


Guanine monophosphate reductase



Nucleosome with human Histone octamer and DNA

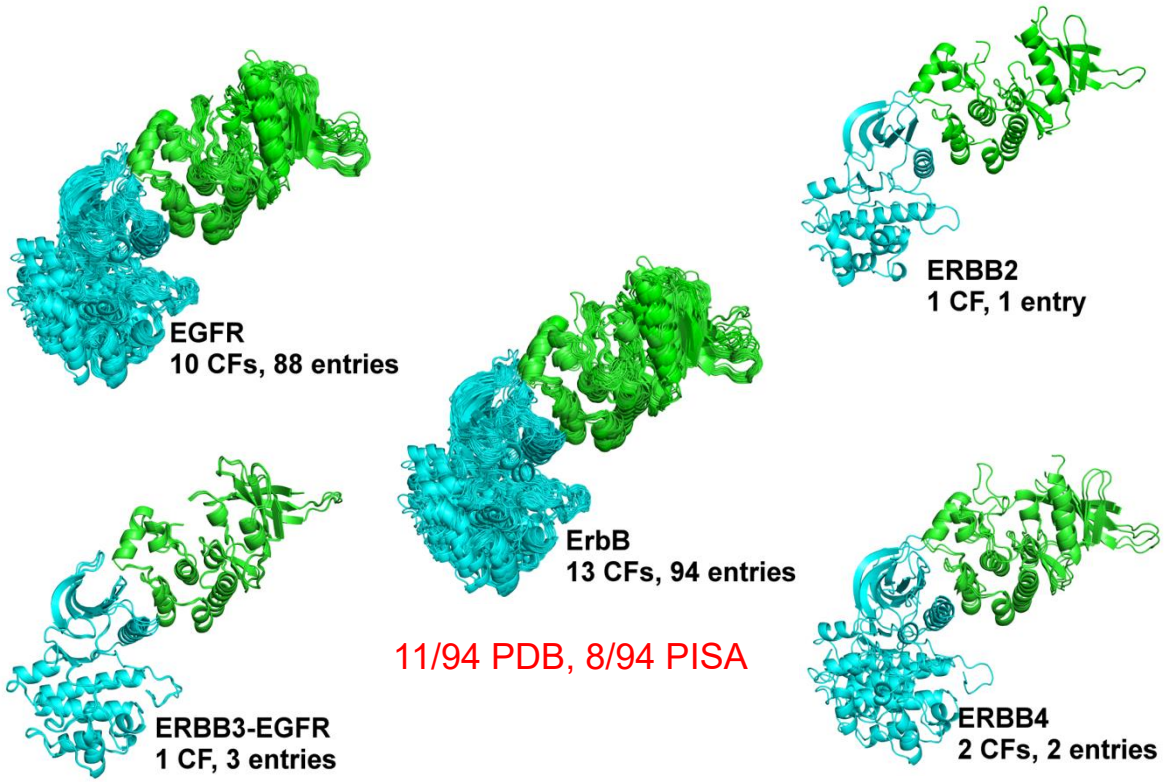
Biological assemblies in protein crystals



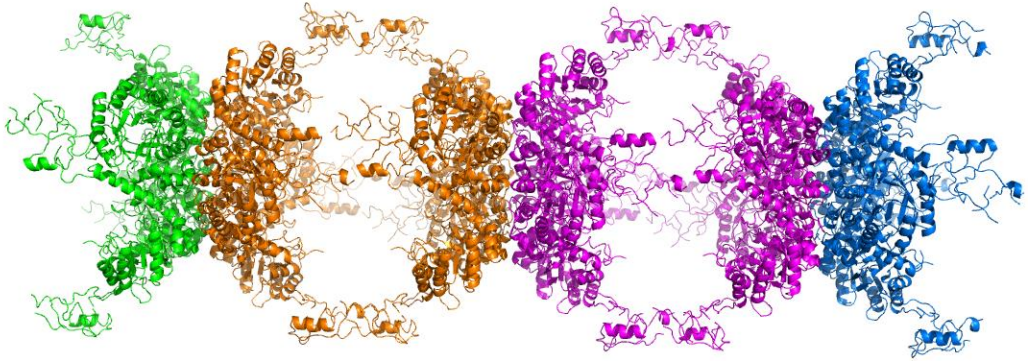
Human estrogen sulfotransferase is a dimer. Which dimer?

Seeing the same dimer or oligomer in different crystals can indicate “biological relevance”

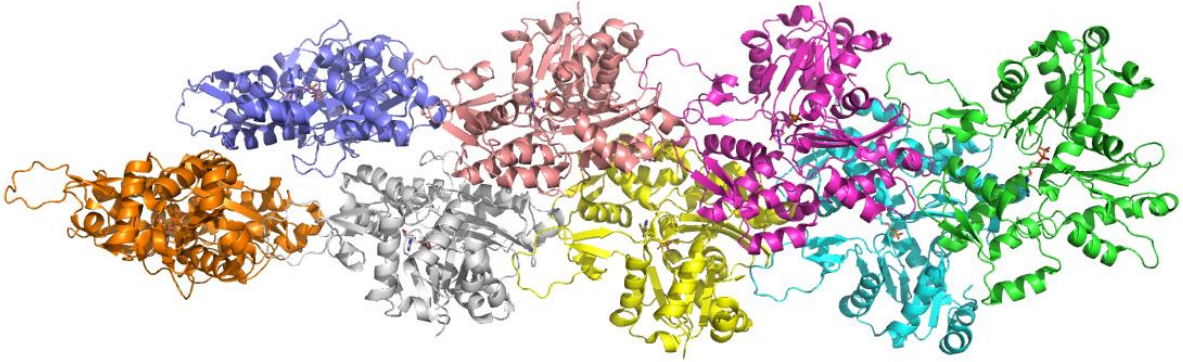
https://dunbrack2.fccc.edu/protcid
https://dunbrack2.fccc.edu/protcad



Filamentous proteins

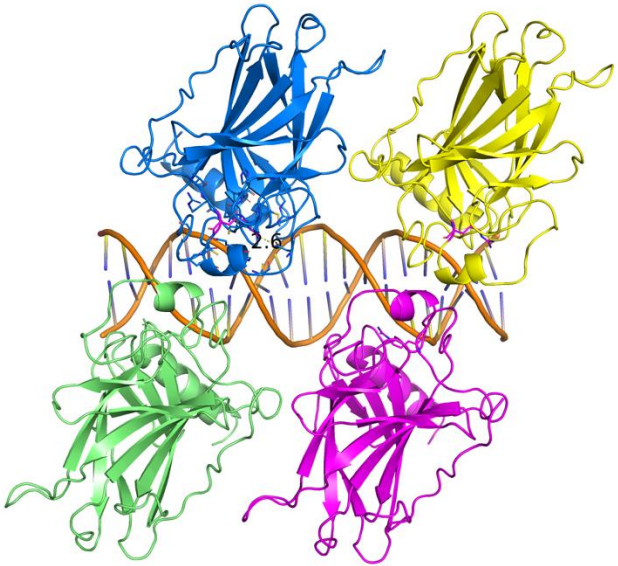


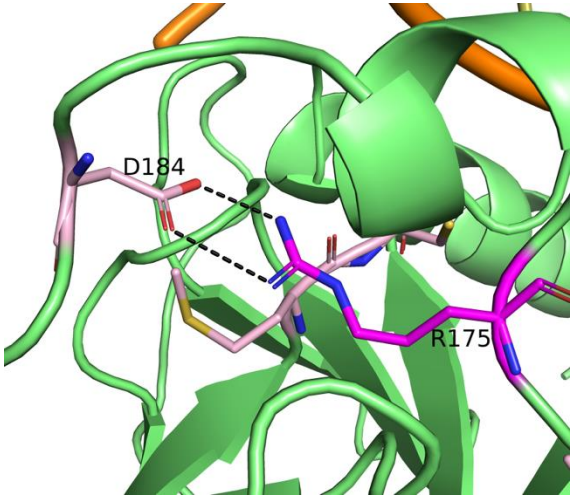
Inosine monophosphate dehydrogenase (D4 octamers)



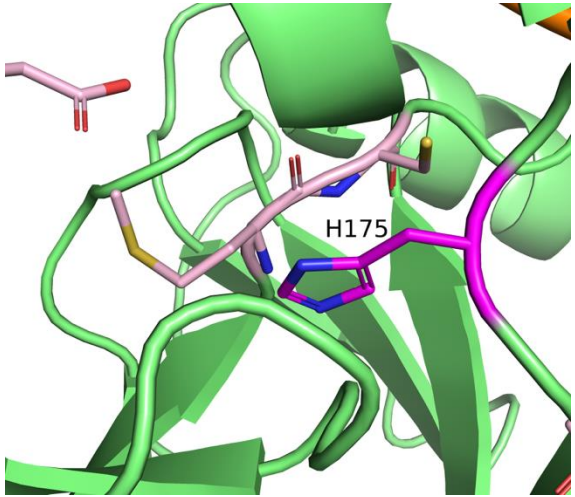
Actin filament (monomers)

13. Structure is important for interpreting/predicting function, conservation, inherited/somatic/experimental mutational effects.



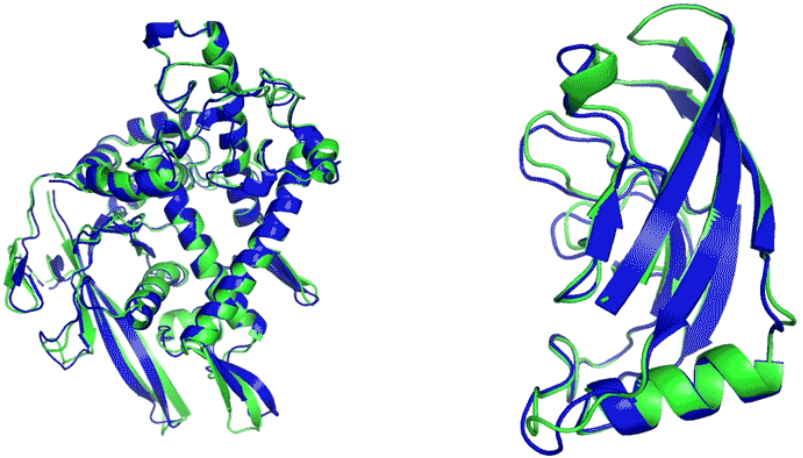


R175 with salt bridge



R175H, no salt bridge

14. But now we have Alphafold2 and Alphafold3 for predicting structure – fast and accurate structure prediction in the absence of experimental structures and sometimes capable of



T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

T1049 / 6y4f
93.3 GDT
(adhesin tip)

Experimental result
Computational prediction

CASP Experiments

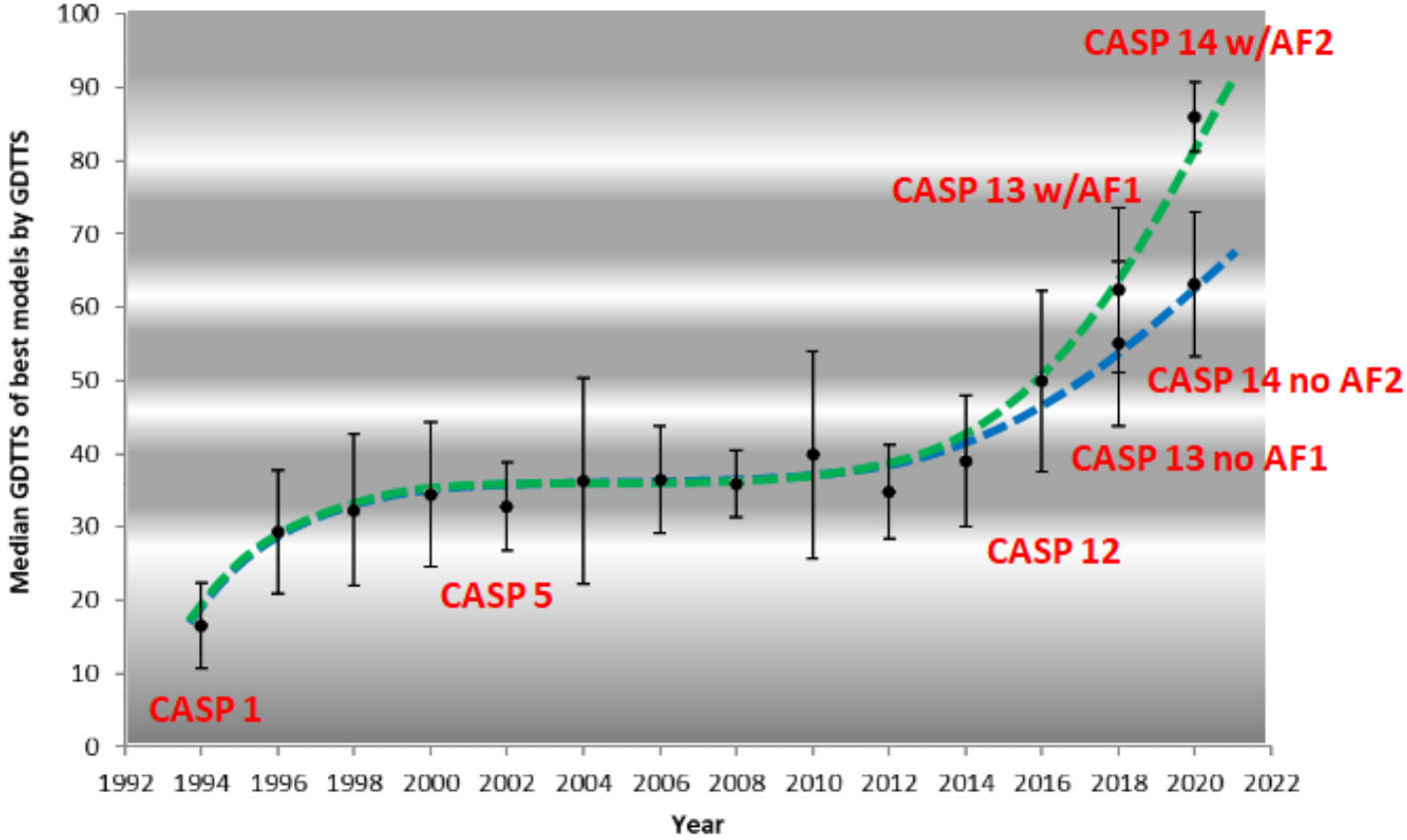
Almost indistinguishable from true structure

Very good models

Topology ok

Overall shape ok

No better than a random model

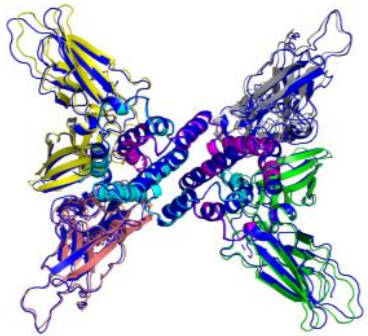


AlphaFold-Multimer

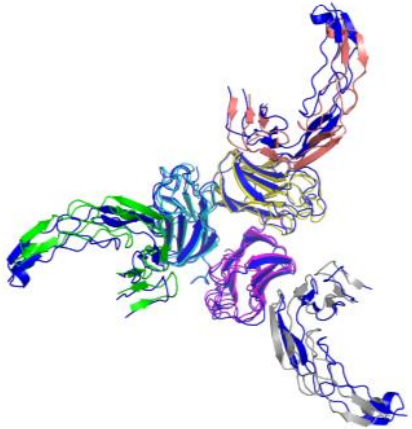
v2.1 in Nov 2021
Training data: 2018-4-30

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

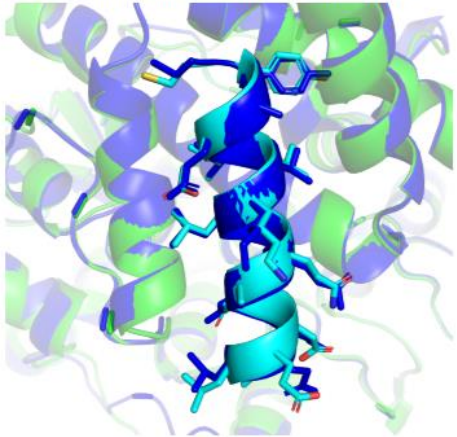
v2.3 in Dec 2022
Training data: 2021-09-30.



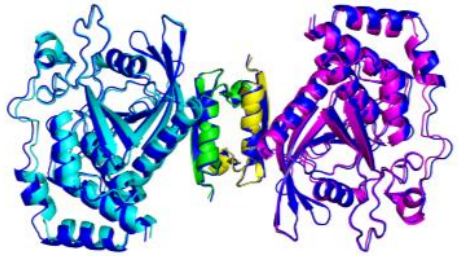
(a) A2B2C2 heteromer
TM-score = 98.0, N_{res} = 1,246, PDB ID = 6E3K



(b) A3B3 heteromer
TM-score = 89.3, N_{res} = 795, PDB ID = 7KHD



(c) Protein-peptide complex
TM-score = 96.0, DockQ = 0.948,
N_{res} = 385, PDB ID = 6JMT

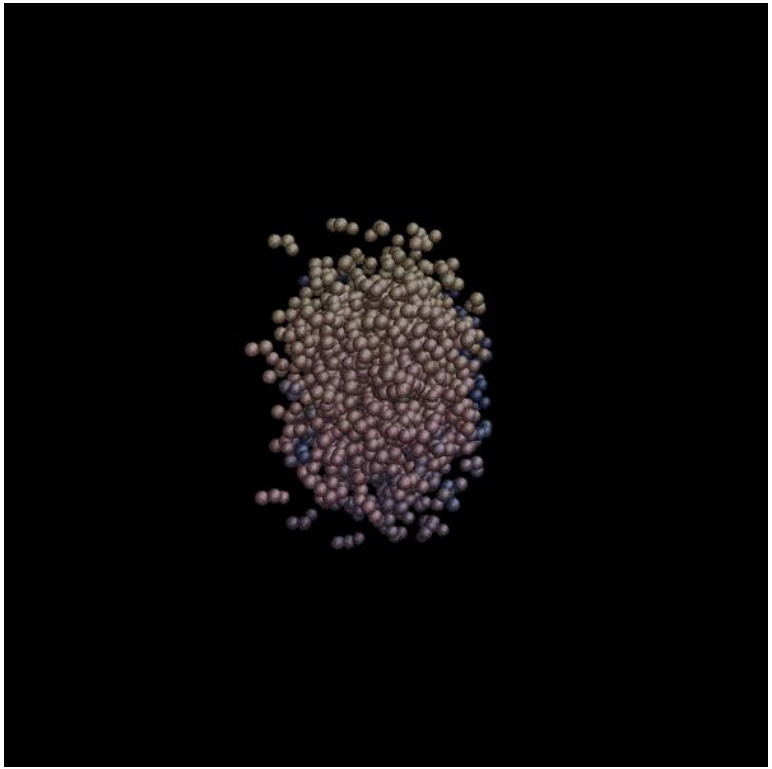


(d) A2B2 heteromer
TM-score = 98.3, N_{res} = 716, PDB ID = 6IWD

AlphaFold3 – Diffusion based structure prediction

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

Training data: 2021-09-30.

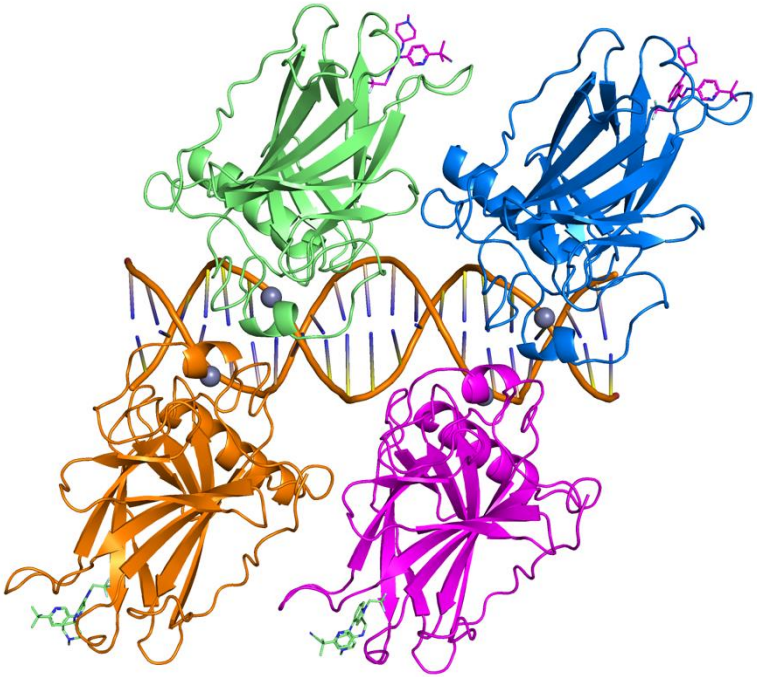


Random noise

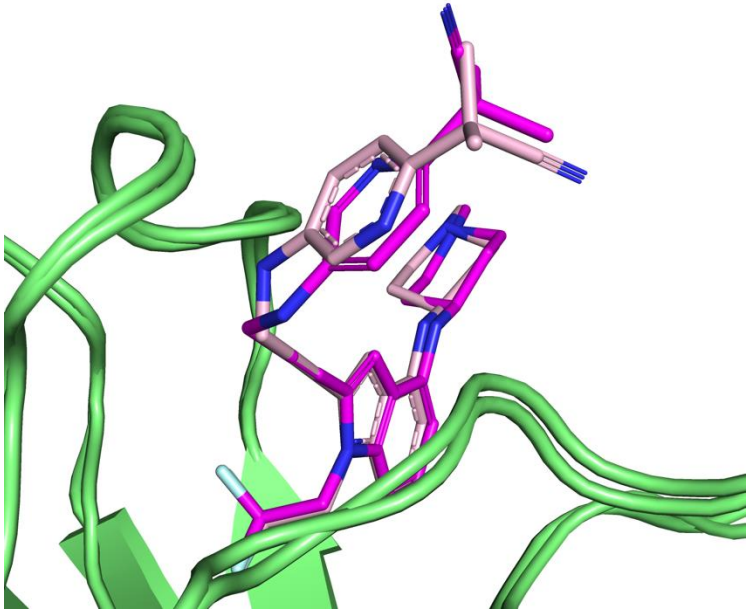
Denoising under informational constraints

AlphaFold3-like programs: AF3, Boltz2, Chai2

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

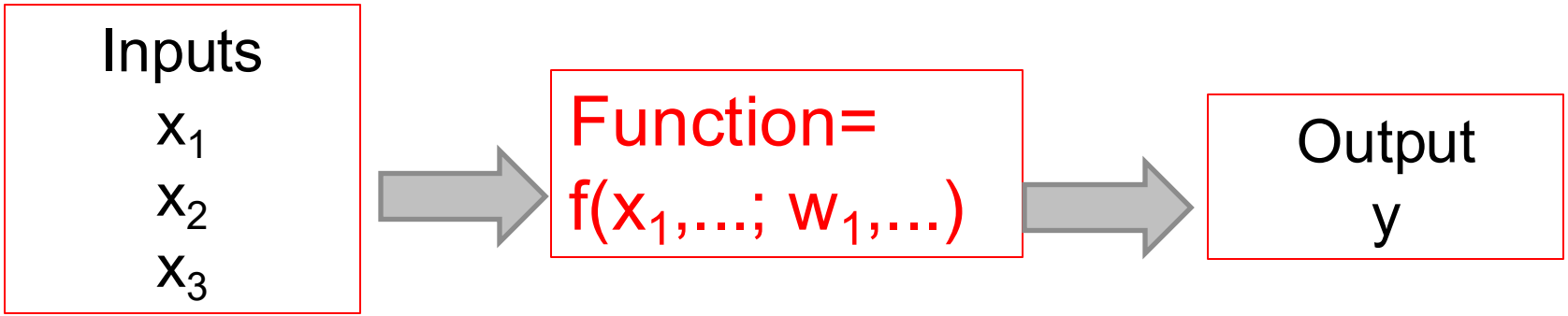
15. AF2 and AF3 are deep-learning neural networks which take inputs (sequences and sequence alignments) and output predicted structures via a small number of “tricks”.

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?



Output can be a class (or classes): cat, dog, tree = Classifier
Output can be a number (or numbers) = Regression
Function has to have 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

$$Z = w_1 x x_1 + w_2 x x_2 + w_3 x x_3$$

A dot product of vectors **x** and **w**

Z=w · x

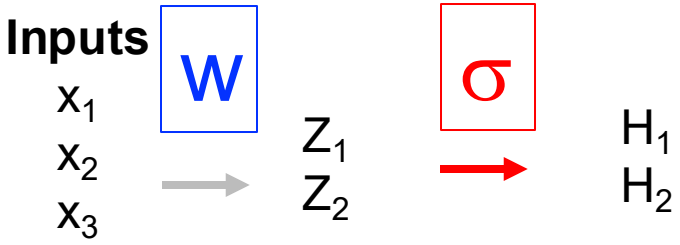
$$Z_1 = w_{11} x x_1 + w_{12} x x_2 + w_{13} x x_3$$

$$Z_2 = w_{21} x x_1 + w_{22} x x_2 + w_{23} x x_3$$

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

Z=wx

Trick #2. Apply an "Activation Function" (σ)

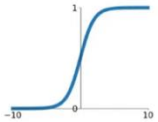


numbers between 0 and 1
or between -1 and 1

Activation Functions

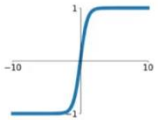
Sigmoid

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



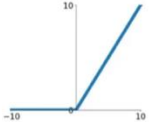
tanh

$\tanh(x)$



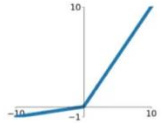
ReLU

$\max(0, x)$



Leaky ReLU

$\max(0.1x, x)$

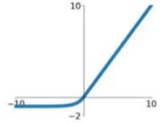


Maxout

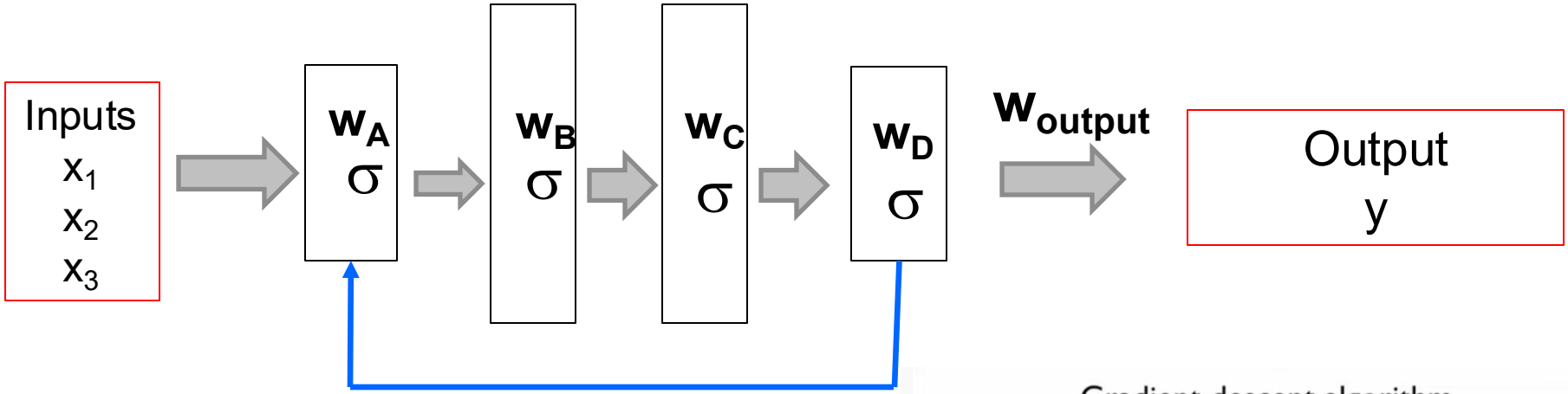
$\max(w_1^T x + b_1, w_2^T x + b_2)$

ELU

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



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



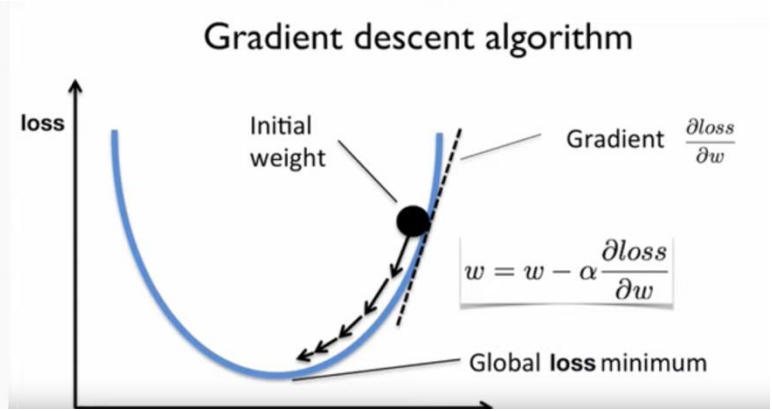
Learn the weights w_A, w_B, w_C, w_D, w_{output}

(1) Random start

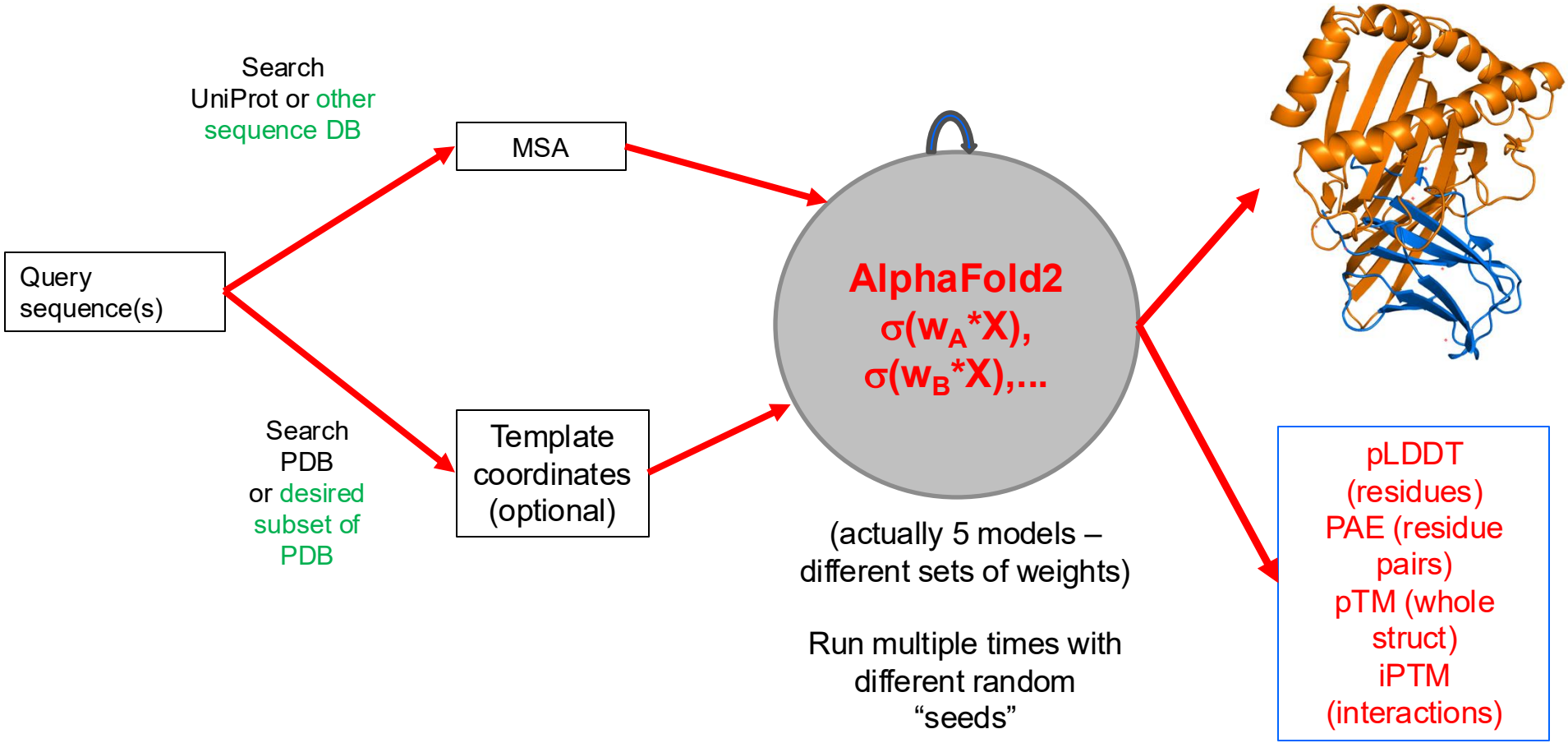
(2) Calculate y for training x;

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

(4) Adjust w to get better y



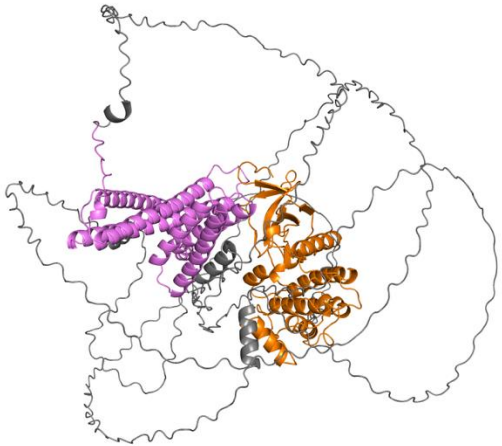
16. AF2 and AF3 output “scores: like pLDDT, PAE, ipTM.

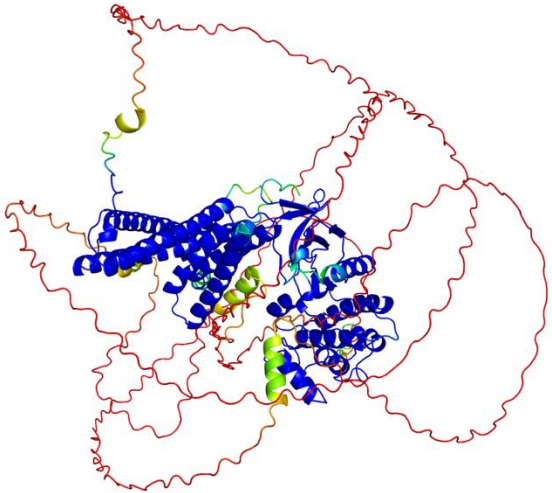


pLDDT – predicted local distance difference test

1. In experimental structure, identify all atoms within 15 Å of a residue and measure distances
2. In model, measure same distances
3. Calculate differences in distances between experimental structure and model
4. Calculate function of differences from 0 to 100
   → 100 = perfect match (all distances 0.0),
   → 0 = all differences very large

Colored by domains pLDDT<50 in gray

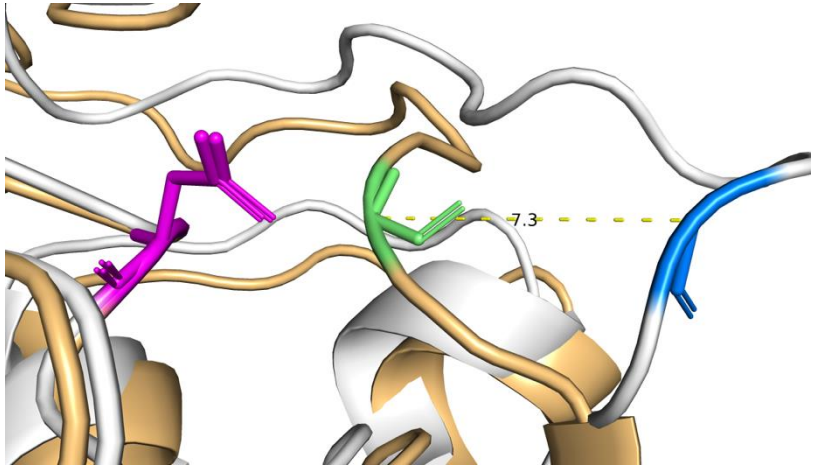


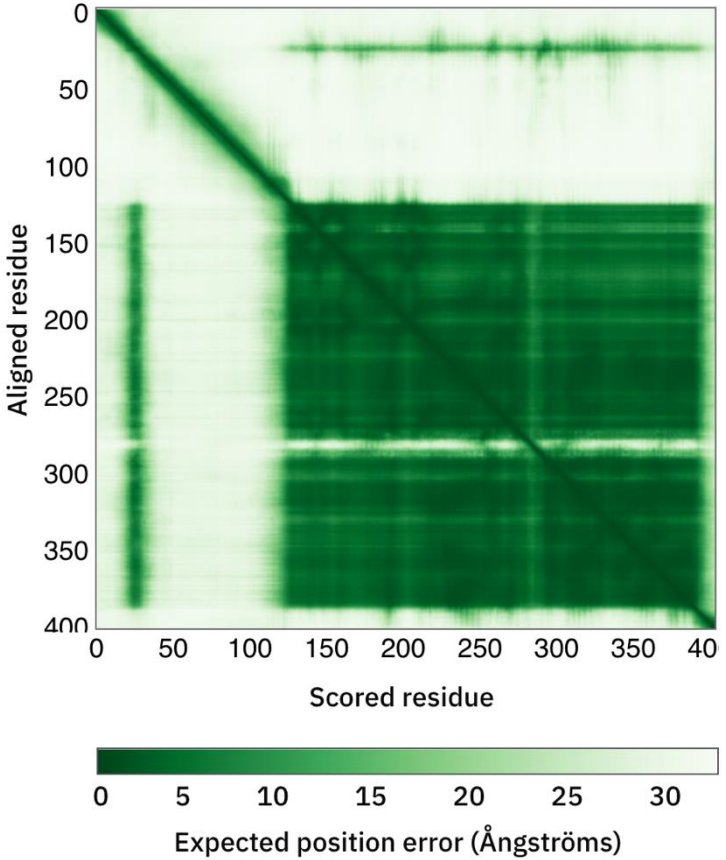


Colored by pLDDT (folded domains high pLDDT in blue)

PAE: Predicted Aligned Error

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

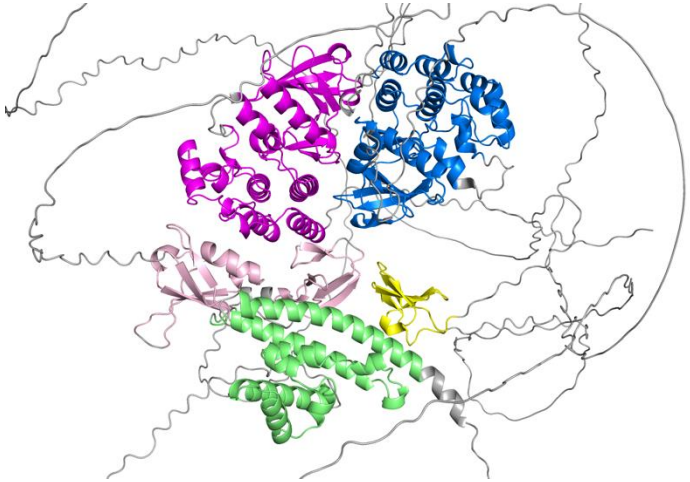




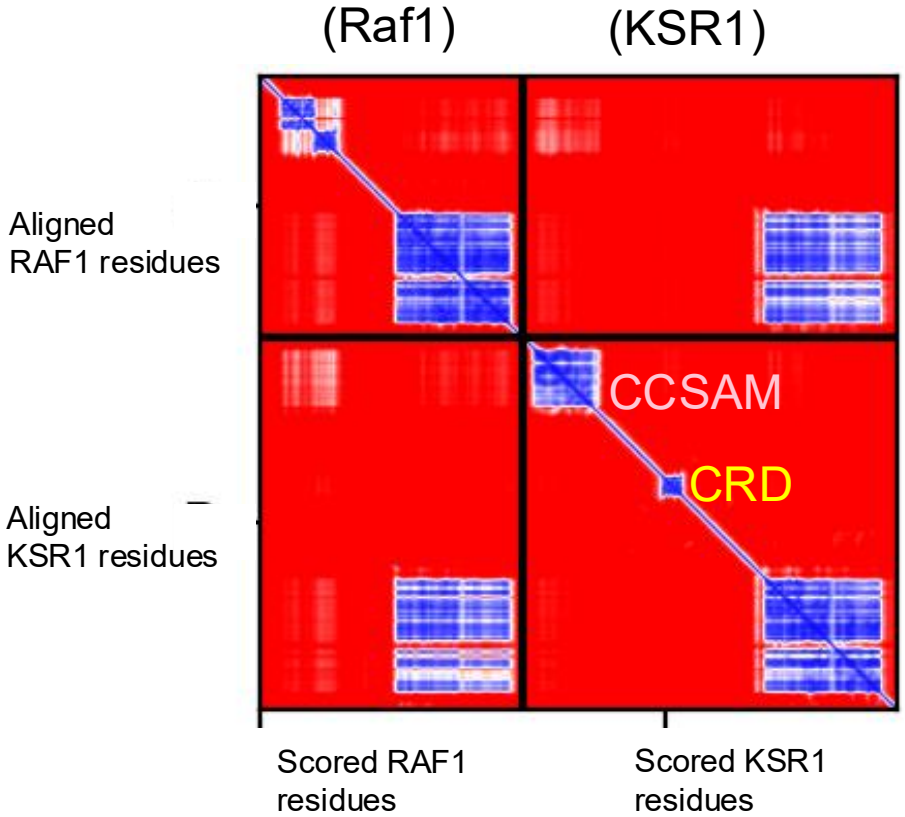
PAE is predicted error distance between Scored residue (Cα) in model and PDB

PAE matrix of a protein complex

ipTM=0.38



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



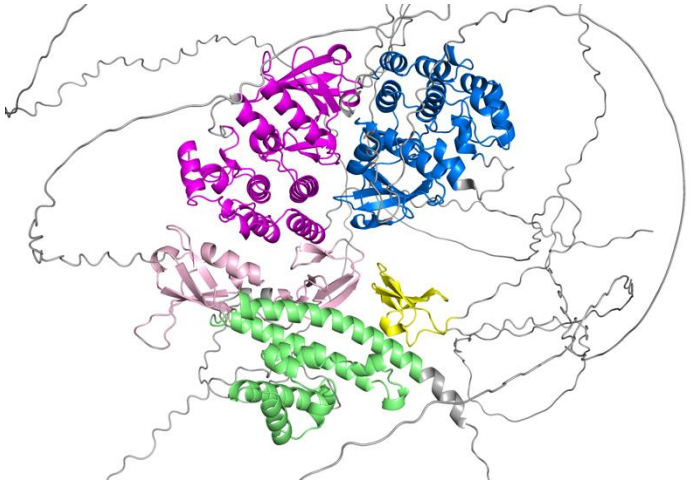
PAE
matrix:

blue<5 Å
red>25 Å

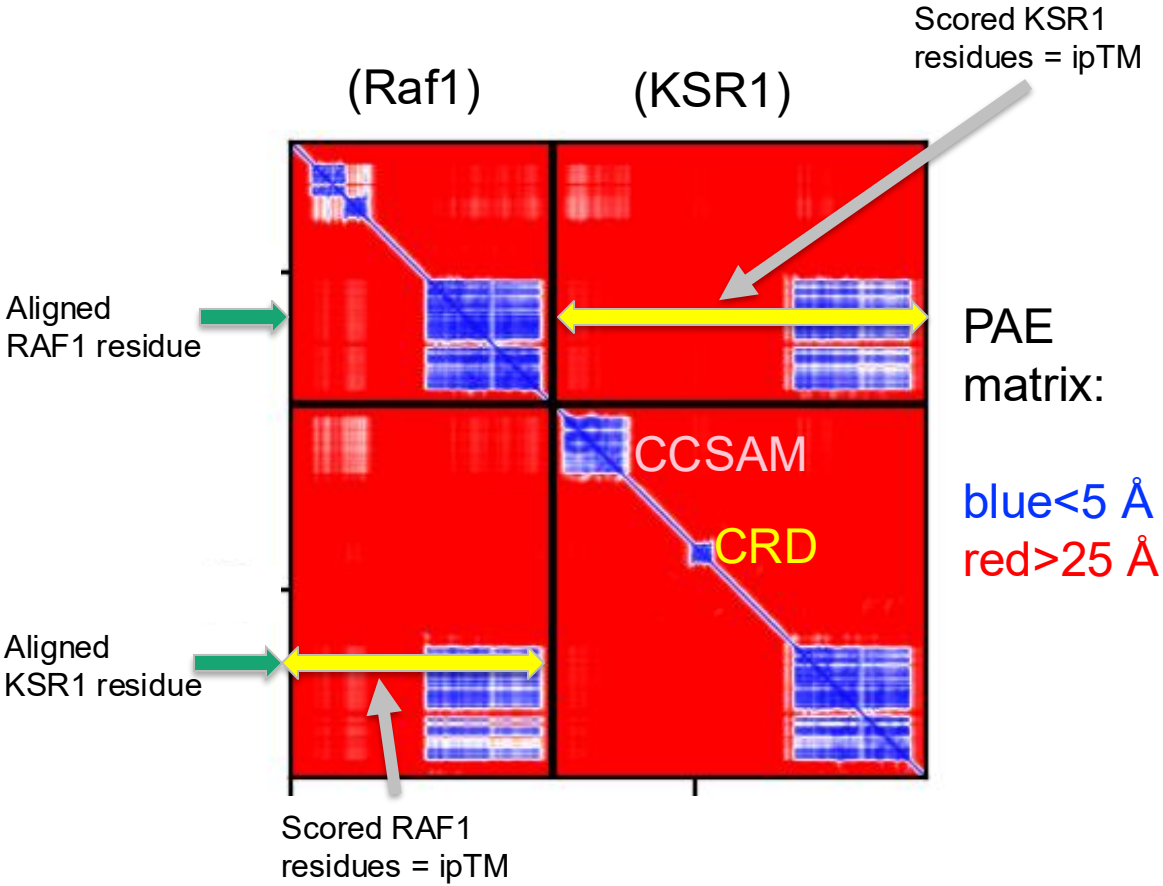
ipTM : interface predicted template modeling score

Calculated on the full structures, not interface. Between 0.0 and 1.0 (perfect model)

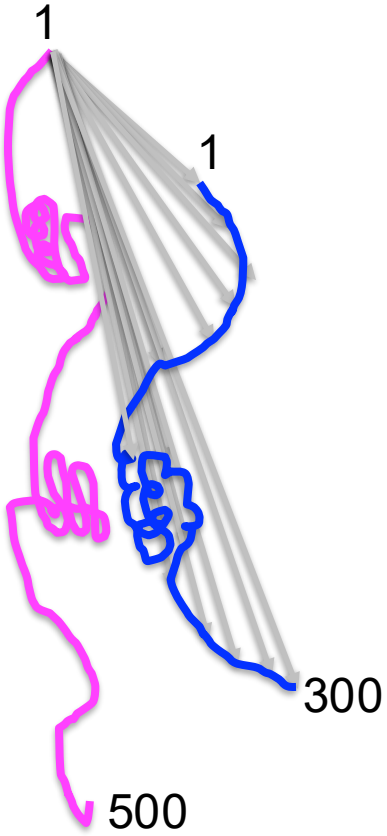
ipTM=0.38

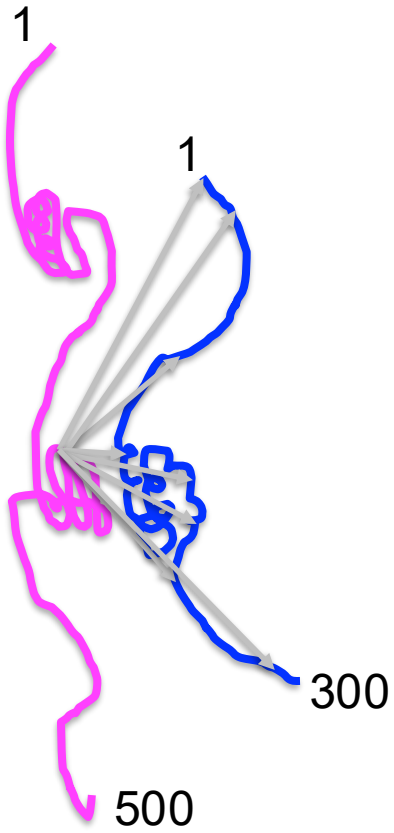


RAF1 (RBD-CRD, Kinase) and
KSR1 (CC-SAM, CRD, Kinase)



AlphaFold's ipTM Score





Align on one residue at a time in **chain A** Score ***ALL*** residues with ***TM(j)*** in **chain B** with TM formula given PAE values

Align on one residue at a time in **chain B** Score ***ALL*** residues with ***TM(j)*** in **chain A** with TM formula given PAE values

Each one is an alignment

ipTM is maximum score over all such alignments in both proteins

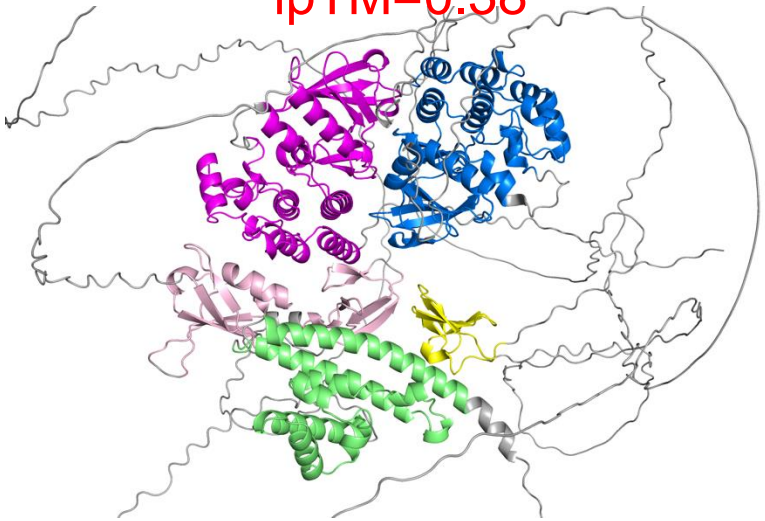
ipTM(res 1) = 0.15

ipTM(res 250) = 0.40

17. We developed ipSAE for protein-protein and domain-domain, domain-peptide interactions, which fixes problems in ipTM and works well for miniprotein-binder design.

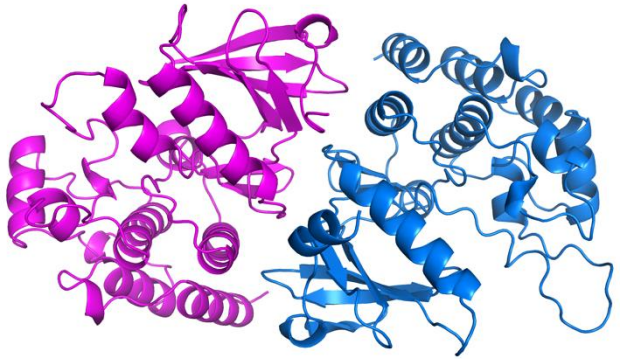
- Large constructs with disorder and lots of domains → low ipTM
- Identify interacting domains/segments in structure and in PAE plot
- Make new constructs based on structure/PAE and run AF → high ipTM
- Why?

ipTM=0.38



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

ipTM=0.82

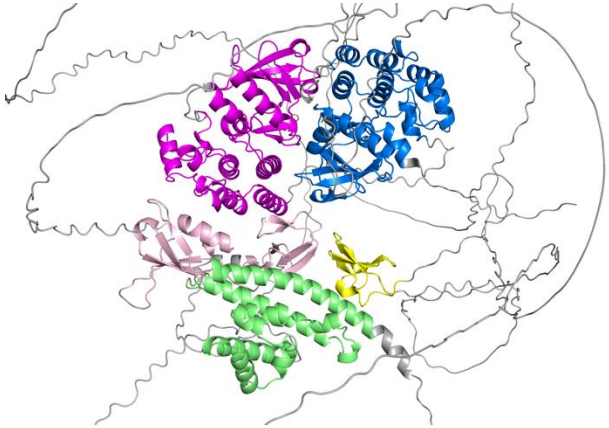


RAF1 (Kinase) and KSR1 (Kinase)

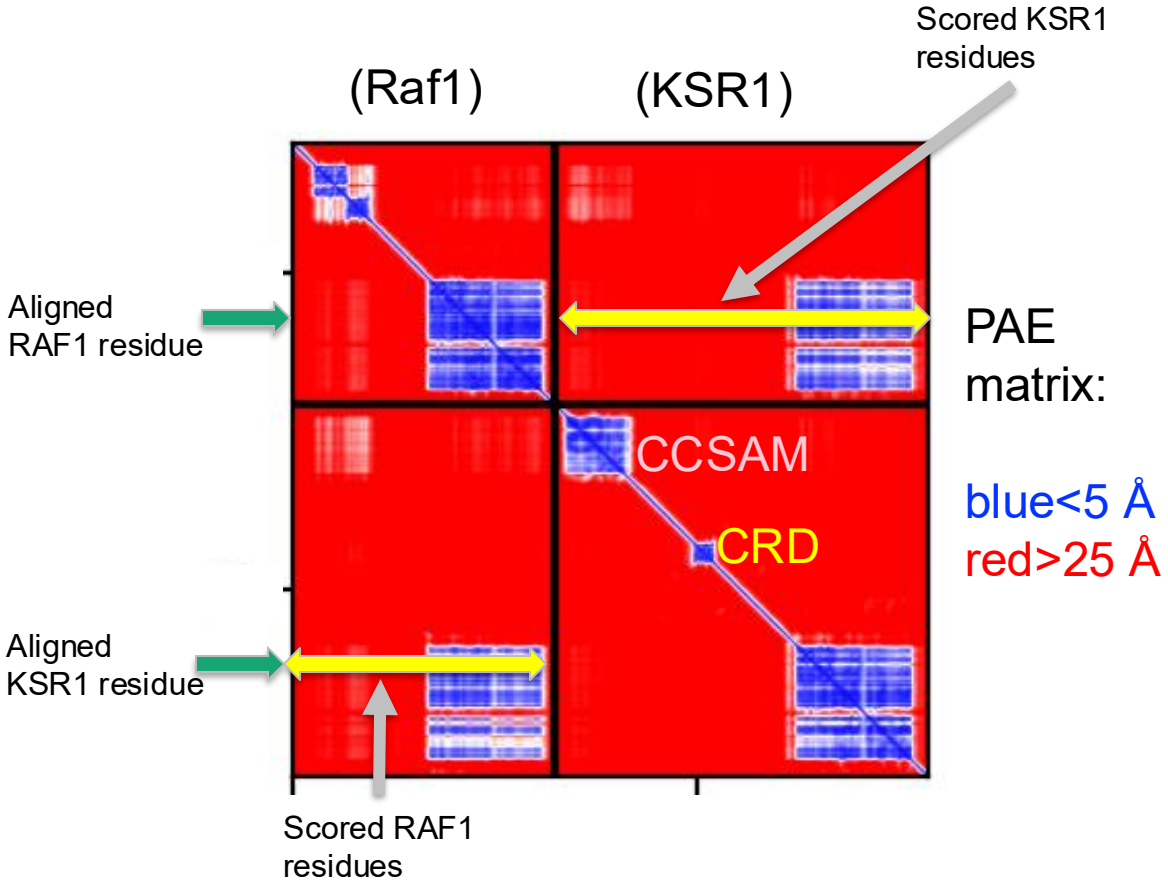
Disorder and extra domains lower *ipTM*

when present in both chains

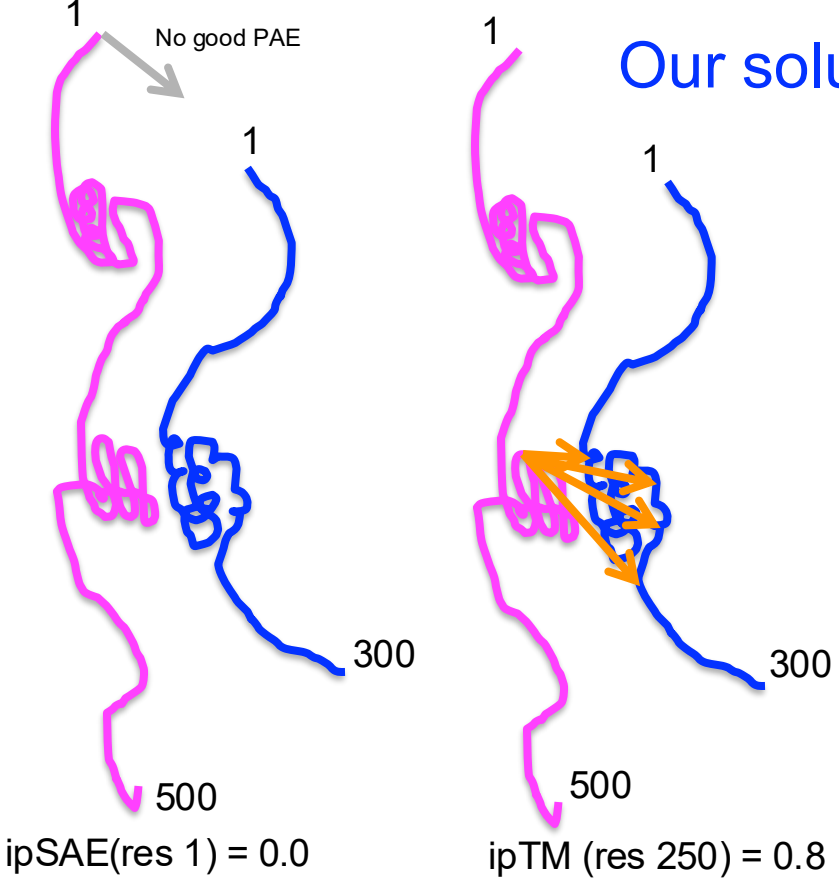
ipTM=0.38



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



Our solution (ipSAE)



ipSAE(res 1) = 0.0

ipTM (res 250) = 0.8

Align on one residue at a time in **chain A**
Average of ***TM(j)*** of residues in **chain B**
but only when there are PAE<cutoff (10 Å)

Align on one residue at a time in **chain B**
Average of ***TM(j)*** of residues in **chain A**
but only when there are PAE<cutoff (10 Å)

ipSAE is maximum score over all such alignments

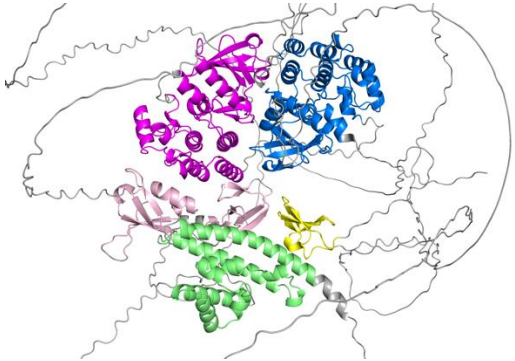
Disordered residues & 1st domain: no good PAE with chain B

Residue in 2nd domain of chain A: good PAE with all of folded domain in chain B

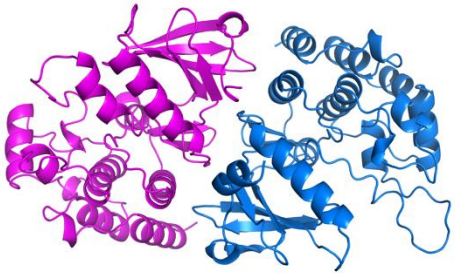
ipSAE -- Scoring full-length protein sequence models

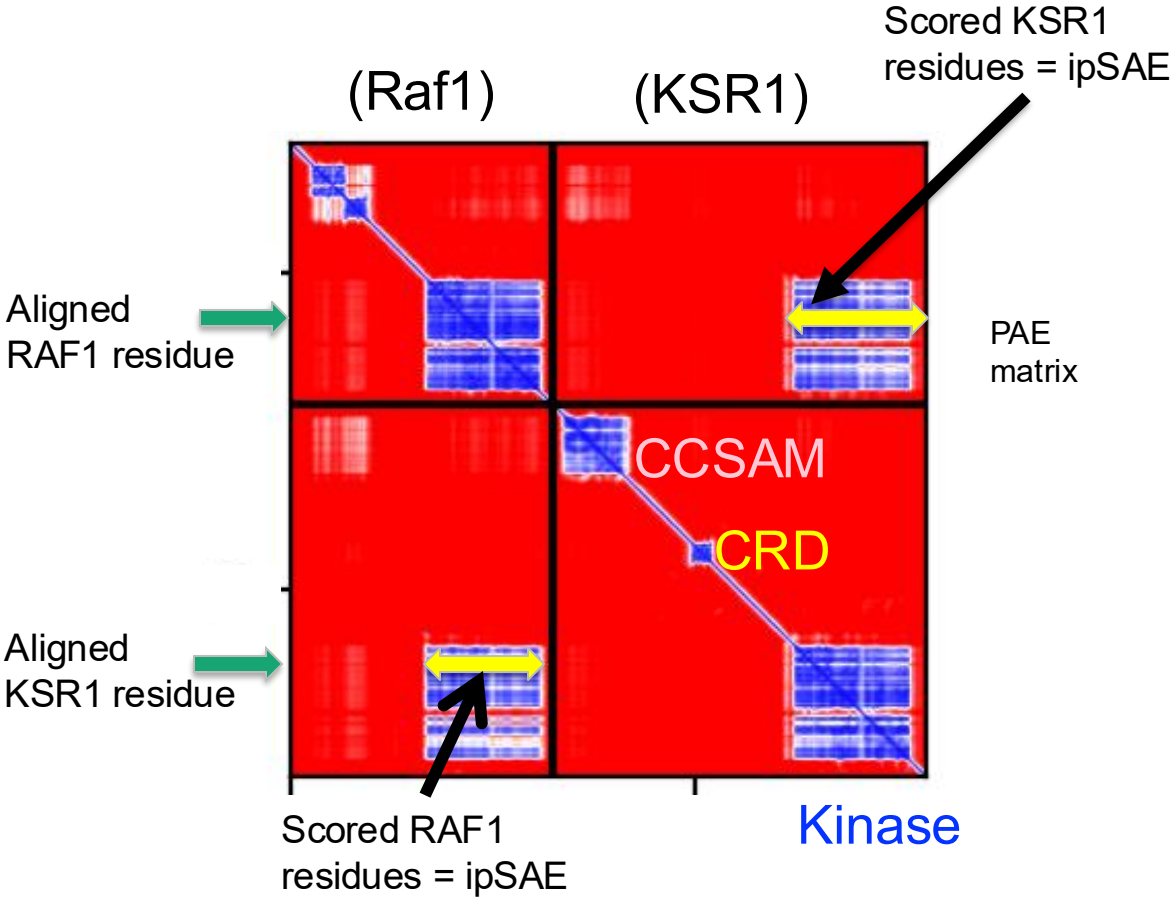
even when disorder and accessory domains are present in both chains

ipSAE=0.75



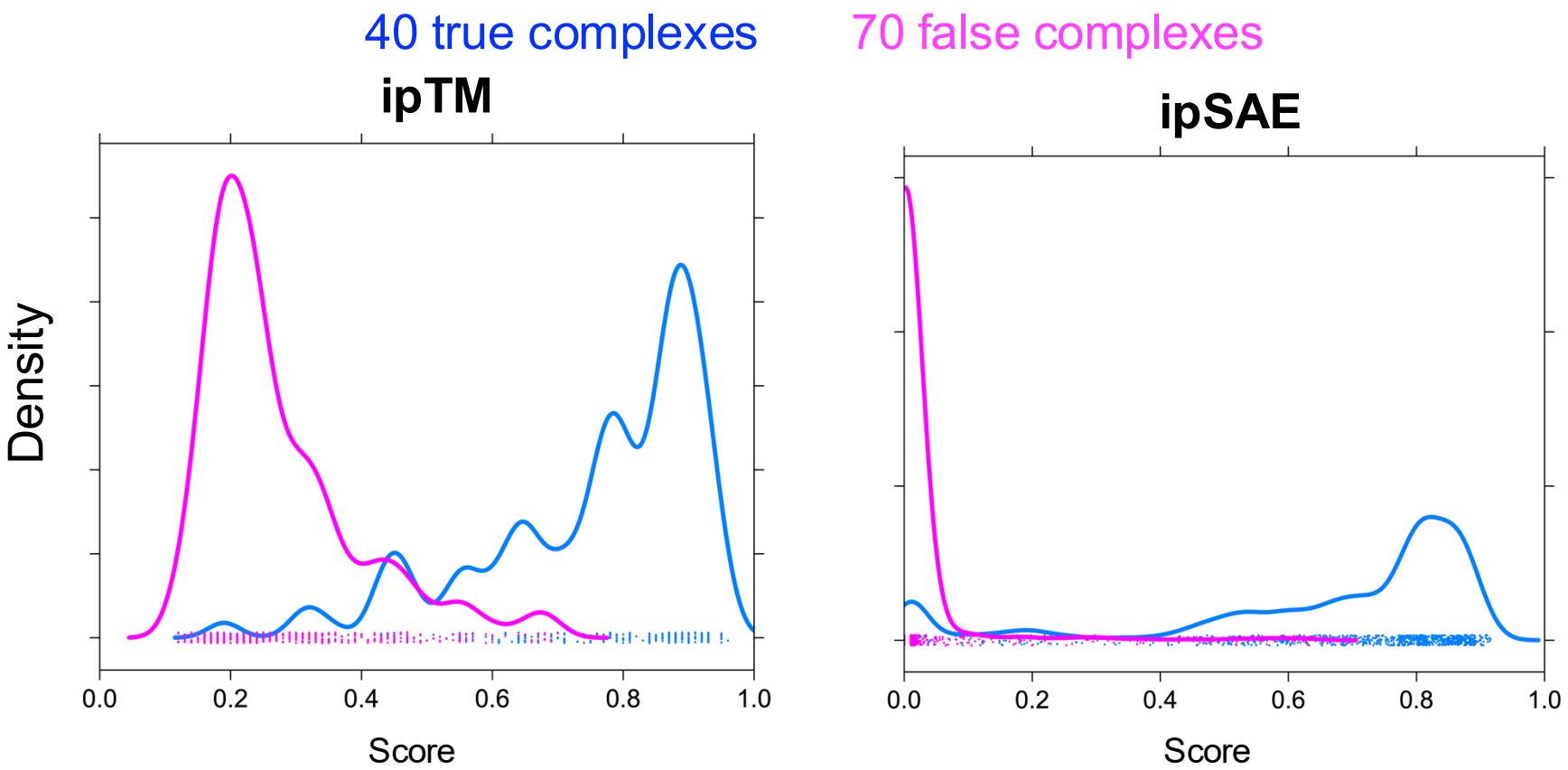
ipSAE = 0.75





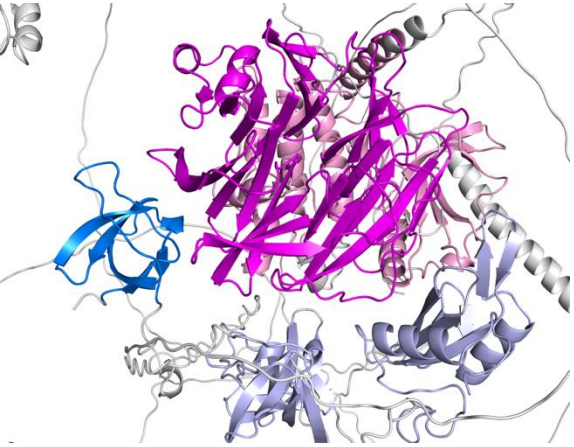
ipSAE distinguishes interactors from non-interactors better than ipTM

Full Uniprot sequences given to AlphaFold2

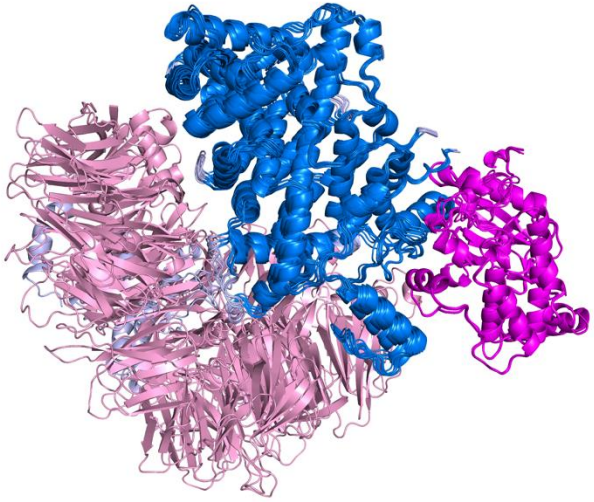


ipSAE enables interactome studies: TNIK

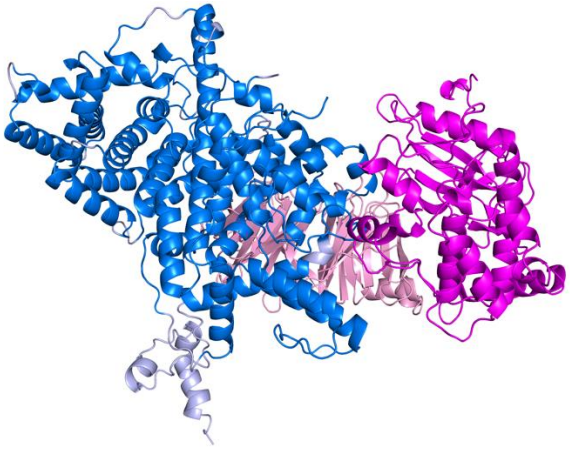
92 interactors from Biogrid (out of 161)
Full Uniprot sequences of TNIK and partner
11 score better than ipSAE=0.50
22 score better than ipSAE=0.40



TNIK beta-propeller domain with NCK2 SH3 domain
ipSAE=0.57, ipTM=0.34
pDockQ2=0.02



TNIK kinase domain with STRIP1
Top6:
ipSAE=0.51-0.65, ipTM=0.64-0.70
pDockQ2=0.01-0.04



TNIK kinase domain with STRIP1
Rank011
ipSAE=0.09, ipTM=0.43
pDockQ2=0.01

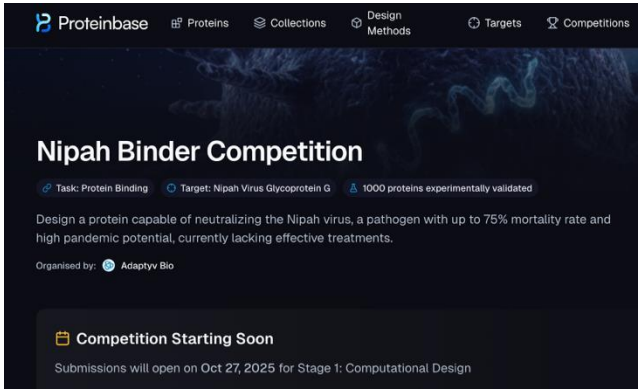
ipSAE for protein binder design

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

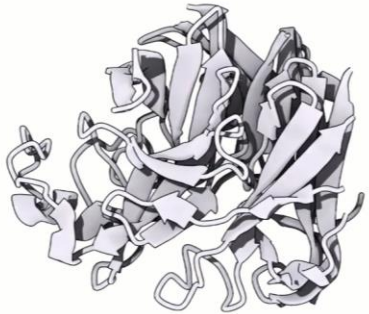
Max D. Overath[†1,*], Andreas Rygaard[†1,2,*], Christian P. Jacobsen[1], Valentas Brasas[1], Oliver Morell[1], Pietro Sormanni[2], and Timothy P. Jenkins[1,*]

[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

We show that interface-focused metrics, most notably the AF3-derived interaction prediction Score from Aligned Errors (ipSAE) outperform commonly used scores such as ipAE and ipTM, with a significant 1.4-fold increase in average precision compared to ipAE.



Using ipSAE to select 600 designs to synthesize



18. AF is very good for fixing understanding of domain architecture of large human proteins as demonstrated by our analysis of full-length structure prediction of human proteins that contain protein kinase domains.

Full-length models of UniProt sequences now available (or computable)

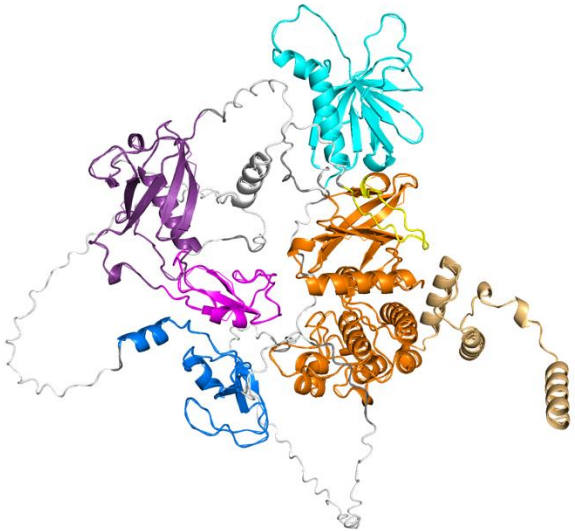
Approximately 1700 non-kinase domains in 480 kinase genes

Understudied dark kinome

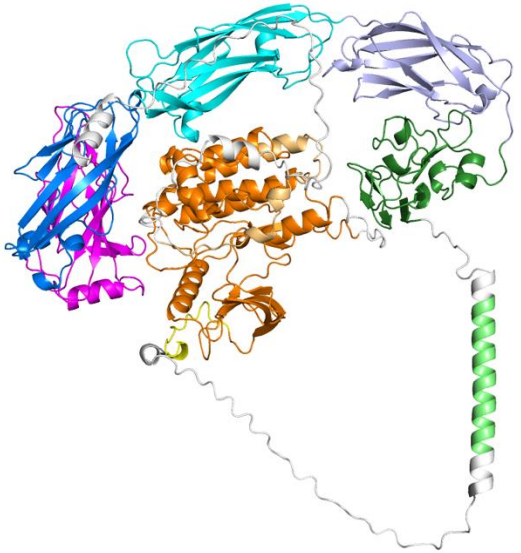
Mutational effects in cancer and other diseases

NT and CT extensions of kinase domain for crystallization studies

Functional domains can be:
Activating
Inhibiting
Membrane-binding
Substrate-association
Catalytic activity
Scaffolding



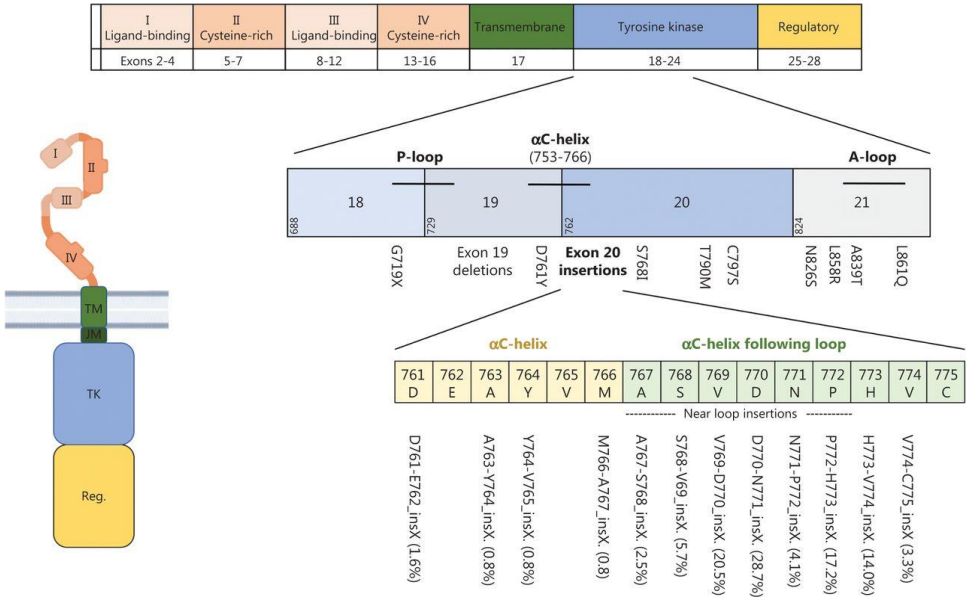
CAMK_PRKD1



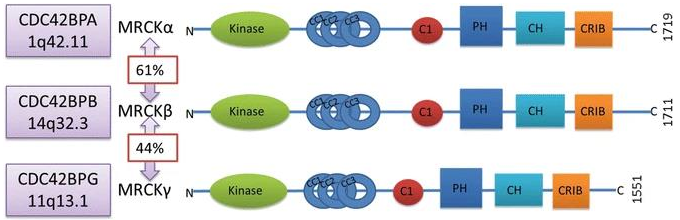
TYR_RET

Domain diagrams are missing a lot or not to scale

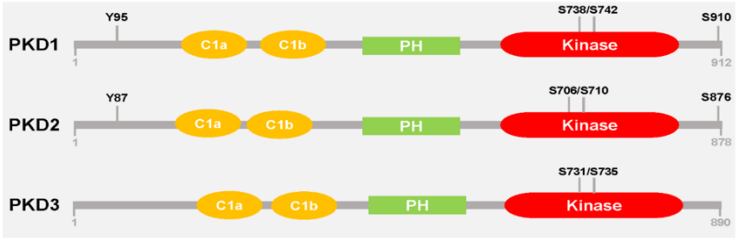
EGFR not to scale



MRCK kinases – missing domain boundaries and helice



PRKD1,2,3: Missing N-terminal Ubq domains



Information Sources

EBI

AFDB models – one model from Model 1 of AF2

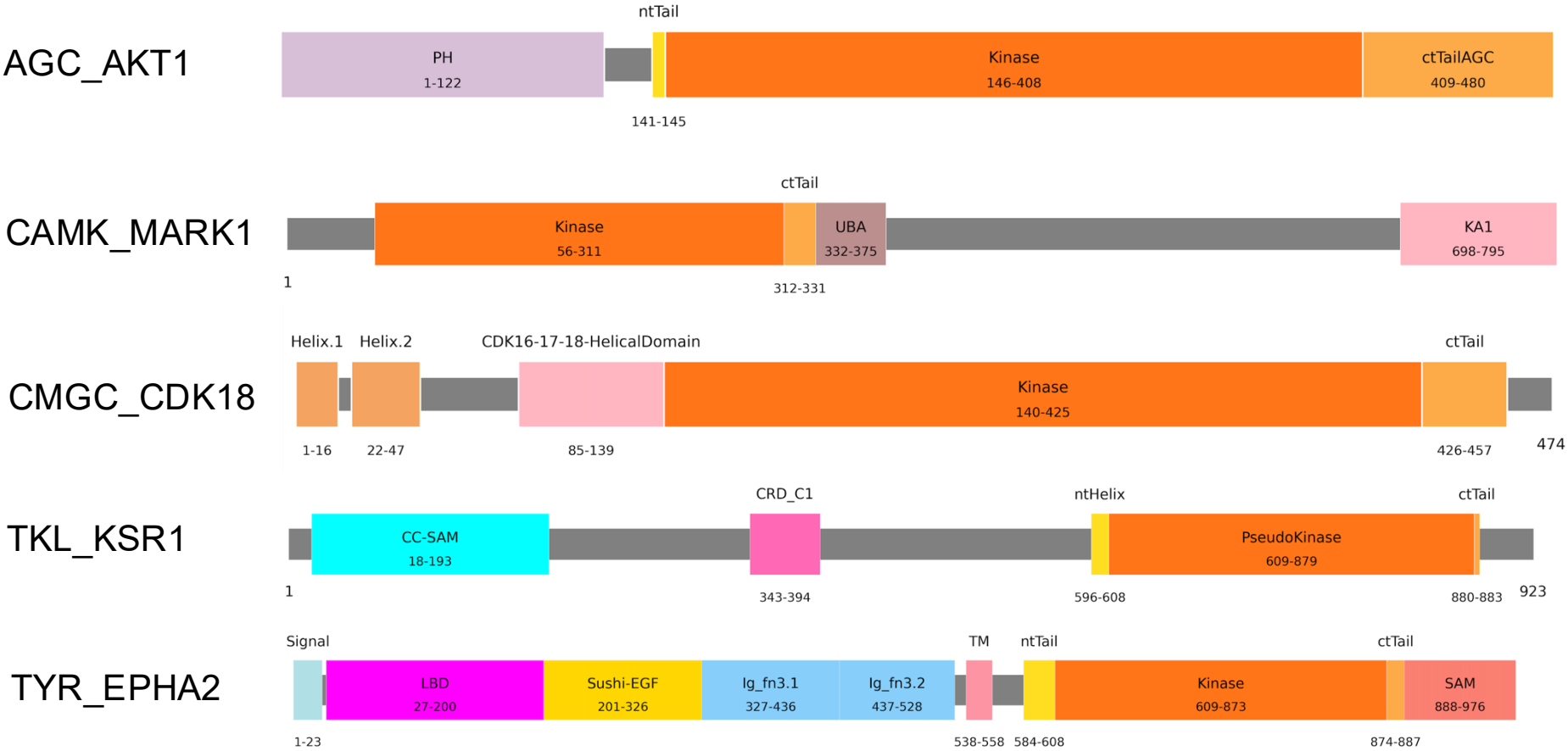
Public domain assignments

TED -- The Encyclopedia of Domains (CATH)
ECOD – Evolutionary Classification of Domains
UniProt (from HMMs from Pfam/Interpro)
PDB structures (constructs, resolved coordinates)

Dunbrack lab

50 models with AlphaFold2 (ColabFold) – no templates
25 models with AlphaFold2 (ColabFold) – with templates

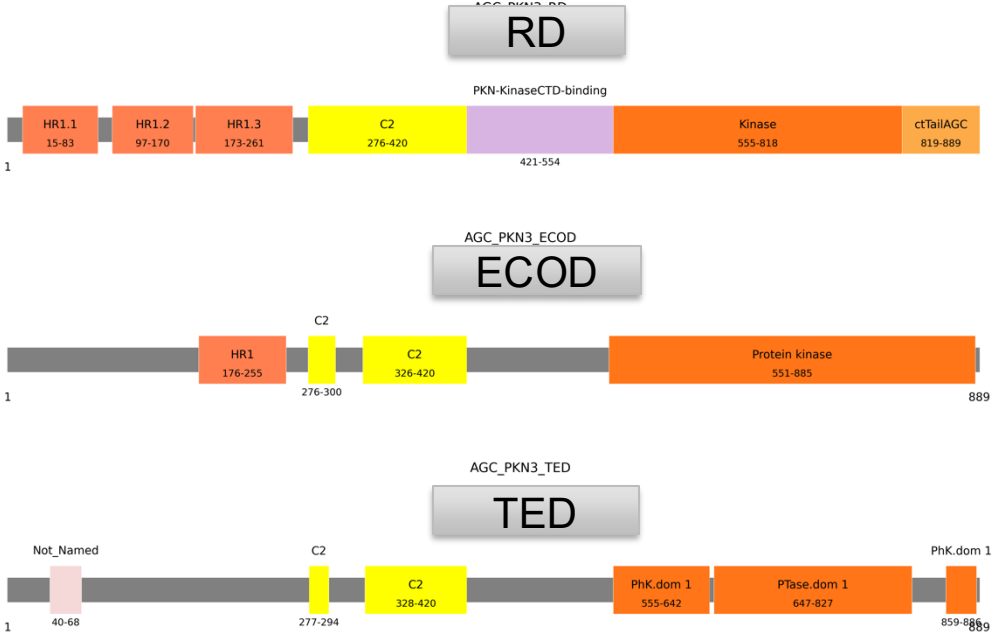
Domain diagrams for all human kinases

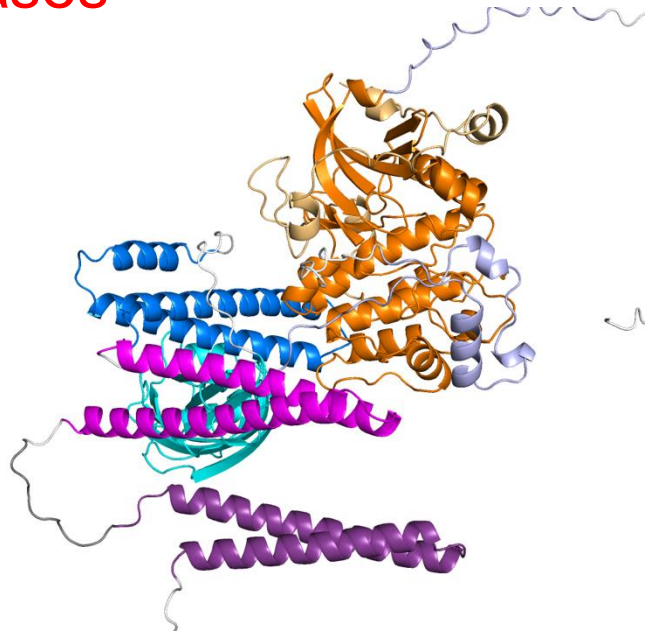


Our 1673 domain assignments vs other sources

Source	#MissingDomains	% 1673 domains	% 480 kinases
ECOD	53	3.2%	8.1%
TED	110	6.6%	16.2%
Uniprot/Interpro	284	17.0%	40.0%

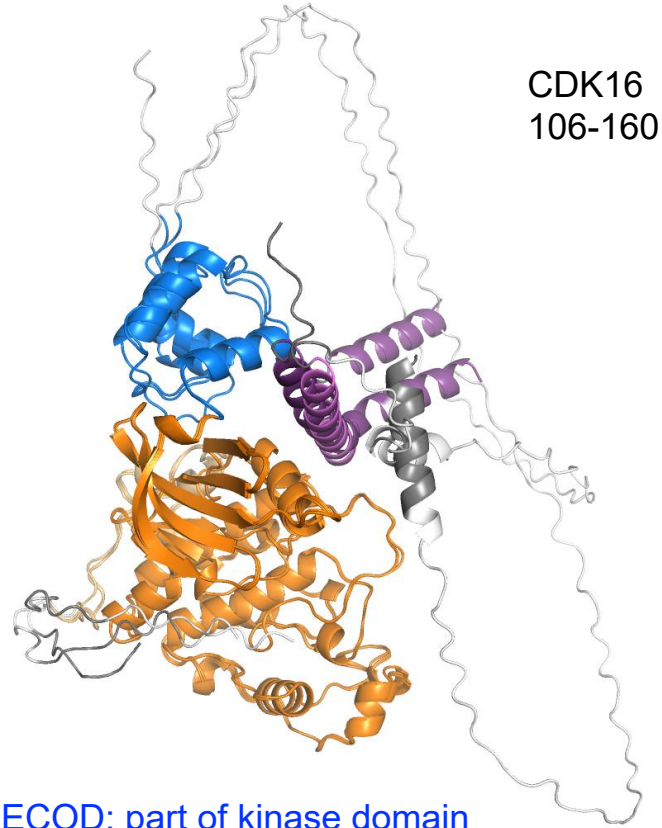
Example: HR1 repeats in PKN kinases



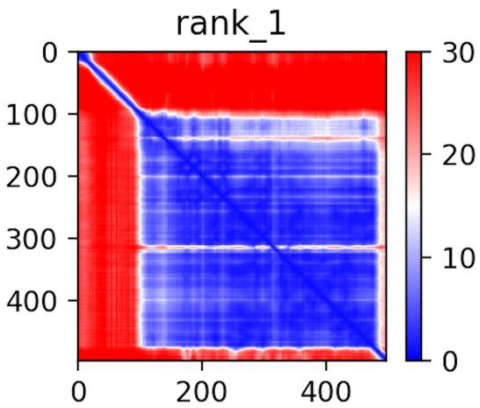


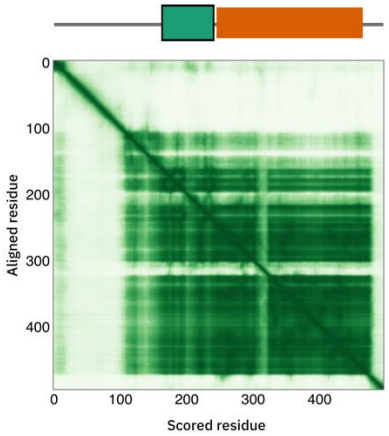
17 Novel Domains in Protein Kinases

Example: CDK16-CDK17-CDK18 Nterm



CDK16
106-160





AFDB-SWISSPROT 22 hits

SHOW TAXONOMY | GRAPHICAL

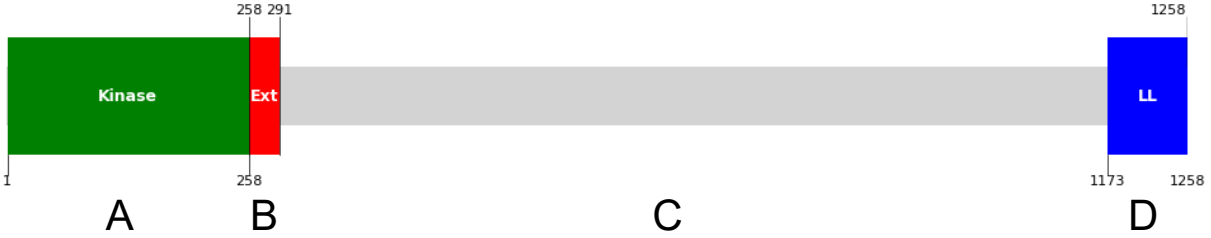
Target	Description	Scientific Name	Prob.	Seq. Id.	E-Value	Position in query
AF-Q00537-F1-model_v4	Cyclin-dependent kinase 17	Homo sapiens	1.00	77.7	4.20e-6	
AF-Q04735-F1-model_v4	Cyclin-dependent kinase 16	Mus musculus	1.00	100	7.31e-7	
AF-O35831-F1-model_v4	Cyclin-dependent kinase 17	Rattus norvegicus	1.00	74.5	1.01e-5	
AF-Q8K0D0-F1-model_v4	Cyclin-dependent kinase 17	Mus musculus	1.00	74.5	1.08e-5	
AF-Q00536-F1-model_v4	Cyclin-dependent kinase 16	Homo sapiens	1.00	100	3.92e-6	
AF-Q07002-F1-model_v4	Cyclin-dependent kinase 18	Homo sapiens	1.00	62.9	7.05e-5	
AF-Q04899-F1-model_v4	Cyclin-dependent kinase 18	Mus musculus	1.00	62.9	2.07e-4	
AF-O35832-F1-model_v4	Cyclin-dependent kinase 18	Rattus norvegicus	1.00	60	2.07e-4	
AF-Q5RD01-F1-model_v4	Cyclin-dependent kinase 18	Pongo abelii	1.00	62.9	1.45e-3	
AF-Q8I7M8-F1-model_v4	Cyclin-dependent kinase 17	Caenorhabditis elegans	1.00	61.1	2.84e-3	
AF-P52730-F1-model_v4	Homeobox protein engrailed-2-B	Xenopus laevis	0.08	18.1	4.93e+0	
AF-Q05917-F1-model_v4	Homeobox protein engrailed-2	Gallus gallus	0.08	20.3	4.61e+0	
AF-P19622-F1-model_v4	Homeobox protein engrailed-2	Homo sapiens	0.08	16.6	4.61e+0	
AF-C0ZYA5-F1-model_v4	Translation initiation factor IF-2	Rhodococcus erythropolis PR4	0.07	10.3	7.37e+0	

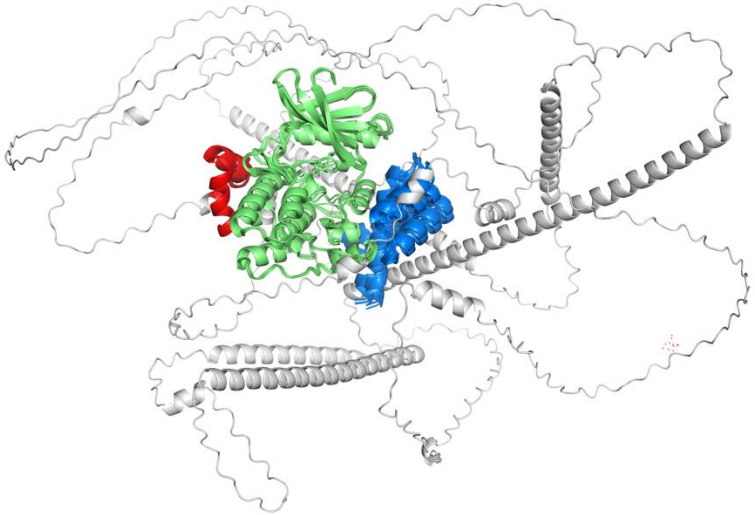
ECOD: part of kinase domain
TED: not covered

Intramolecular domain-domain interactions with ipSAE score

Domain assignments for all 480 human protein kinase proteins from full-length AF2 models

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				

Workshop Tasks

1. Learn basics of Uniprot Pages for proteins
2. Learn basics of PDB pages for protein structures
3. Learn how to use Colabfold (AlphaFold2)
4. Learn how to use AlphaFold3 server
5. Look at ECOD, TED, Foldseek

https://dunbrack.fccc.edu/bioinfo

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

colabfold_batch

```
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

nohup ./run_smo
```

```
>SMO_PKA_mouse_F577A
RGAASSGNATGPGPRSAGGSARRSAAVTGPPPPLSHCGRAAPCEPLRYNVCLGSVLPYGATSTLLAGDSDSQEEAHGKLVLWSGLRNAPRCWAVIQPLLCAVYMPKCENDR
VELPSRTLCQATRGPCAIVERERGWPDFLRCTPDSDF:
GNAAAAKKGSEQESVKEFLAKAKEDFLKKWESPAQNTAHLDQFERIKTLGTGSFGRVMLVKHKETGNHYAMKILDKQKVVKLKQIEHTLNEKRILQAVNFPFLVKLEFSFK
DNSNLYMVMEYVPGGEMFSHLRRIGRFSEPHARFYAAQIVLTFEYLHSLDLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWTLCGTPEYL

>SMO_PKA_mouse_R569E
VILHPNETIFNDFCKKSTTCEVLKYNTCLGSPLPYTHTSLILAEDSETQEEAFEKLAMWSGLRNAPRCWAVIQPLLCAVYMPKCENGKVELPSQHLCQATRNPCSIVERER
GWPNFLKCENKEQFPKGCQNEVQKLKFNTSGQCEAPLVKTDTLLCTFFTLATFLASMMDVGRTAVVPRADGRRGVQIHSRTNLMDAELLDADSDF:
GNAPTAKNKGNEMESVKEFLAKAKEDFLKKWENPAQNTASLDHFERLKTLGTGSFGRVMLVKHKESGQHFAMKILDKQKVVKLKQIEHTLNEKRILQAVSFPFLVRLEHSF
KDNTNLYMVMEYVPGGEMFSHLRRIGRFSEPHARFYAAQIVLTFEYLHSLDLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWTLCGTPEY
```

colabfold_batch output file (log.txt or SMO.out)

```
2024-07-01 00:11:20,812 alphafold2_multimer_v3_model_3_seed_001 recycle=0 pLDDT=69.4 pTM=0.489 ipTM=0.224
2024-07-01 00:11:53,920 alphafold2_multimer_v3_model_3_seed_001 recycle=1 pLDDT=70.1 pTM=0.502 ipTM=0.275 tol=7.77
2024-07-01 00:12:26,991 alphafold2_multimer_v3_model_3_seed_001 recycle=2 pLDDT=71.6 pTM=0.521 ipTM=0.443 tol=5.99
2024-07-01 00:13:00,084 alphafold2_multimer_v3_model_3_seed_001 recycle=3 pLDDT=71.1 pTM=0.51 ipTM=0.281 tol=4.53
2024-07-01 00:13:33,153 alphafold2_multimer_v3_model_3_seed_001 recycle=4 pLDDT=71.2 pTM=0.512 ipTM=0.338 tol=5.06
2024-07-01 00:14:06,239 alphafold2_multimer_v3_model_3_seed_001 recycle=5 pLDDT=71.2 pTM=0.509 ipTM=0.262 tol=5.19
2024-07-01 00:14:39,304 alphafold2_multimer_v3_model_3_seed_001 recycle=6 pLDDT=71.3 pTM=0.518 ipTM=0.291 tol=3.88
2024-07-01 00:15:12,388 alphafold2_multimer_v3_model_3_seed_001 recycle=7 pLDDT=71.5 pTM=0.511 ipTM=0.283 tol=2.41
2024-07-01 00:15:45,451 alphafold2_multimer_v3_model_3_seed_001 recycle=8 pLDDT=71.6 pTM=0.522 ipTM=0.738 tol=9.98
2024-07-01 00:16:18,596 alphafold2_multimer_v3_model_3_seed_001 recycle=9 pLDDT=70.8 pTM=0.495 ipTM=0.447 tol=8.3
2024-07-01 00:16:51,674 alphafold2_multimer_v3_model_3_seed_001 recycle=10 pLDDT=71.8 pTM=0.51 ipTM=0.61 tol=5.34
2024-07-01 00:16:51,675 alphafold2_multimer_v3_model_3_seed_001 took 364.0s (10 recycles)
```

colabfold_batch output (zip file)

```
SMO_PKA_mouse_F577A_coverage.png
SMO_PKA_mouse_F577A_pae.png
SMO_PKA_mouse_F577A_plddt.png
SMO_PKA_mouse_F577A_relaxed_rank_001_alphafold2_multimer_v3_model_5_seed_002.pdb
SMO_PKA_mouse_F577A_relaxed_rank_002_alphafold2_multimer_v3_model_5_seed_005.pdb
SMO_PKA_mouse_F577A_relaxed_rank_003_alphafold2_multimer_v3_model_5_seed_003.pdb
SMO_PKA_mouse_F577A_scores_rank_001_alphafold2_multimer_v3_model_5_seed_002.json
SMO_PKA_mouse_F577A_scores_rank_002_alphafold2_multimer_v3_model_5_seed_005.json
SMO_PKA_mouse_F577A_scores_rank_003_alphafold2_multimer_v3_model_5_seed_003.json
SMO_PKA_mouse_F577A_unrelaxed_rank_001_alphafold2_multimer_v3_model_5_seed_002.pdb
SMO_PKA_mouse_F577A_unrelaxed_rank_002_alphafold2_multimer_v3_model_5_seed_005.pdb
SMO_PKA_mouse_F577A_unrelaxed_rank_003_alphafold2_multimer_v3_model_5_seed_003.pdb
```

```
python3 ipsae.py file_rank_001.json file_rank_001.pdb 10 10
```