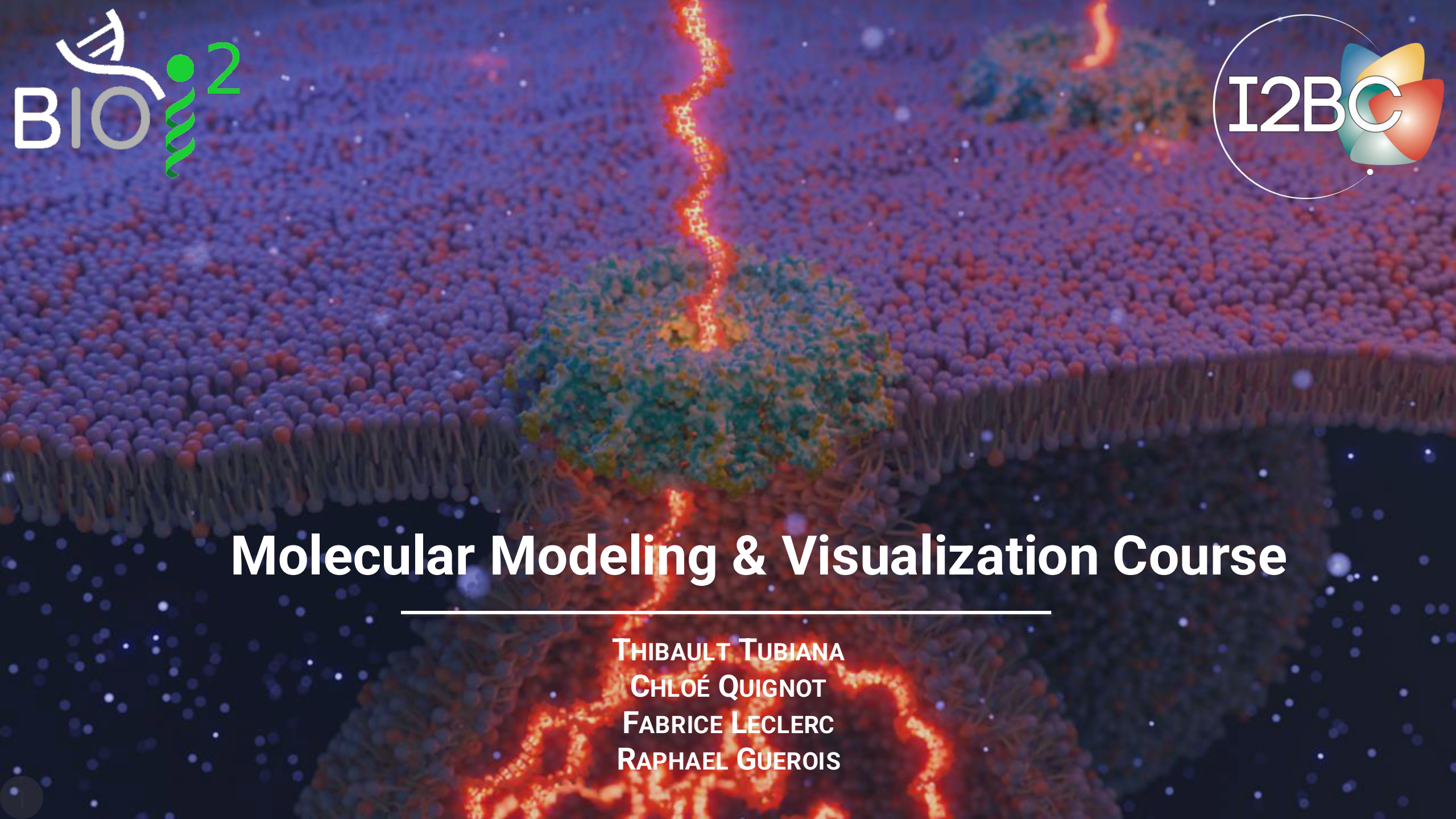


Molecular Modeling & Visualization Course

Thibault Tubiana
Chloé Quignot
Fabrice Leclerc
Raphael Guerois

Setup

- ***Do you all have internet? (Eduroam or Guest WiFi)***
- ***Do you all have ChimeraX installed?***

Instructions for Guest WiFi:

1) Connect to **I2BC_ENREGISTREMENT**,
   - It should open an internet window
   - if not, type the following in a browser: http://198.18.34.1

2) Enter your first name, last name and email address
   - "Personne visitée" : chloe.quignot@i2bc.paris-saclay.fr
   - **Note down the code that was given**

3) Wait for us to confirm your request, then connect to **I2BC_ACCESS** and type your code

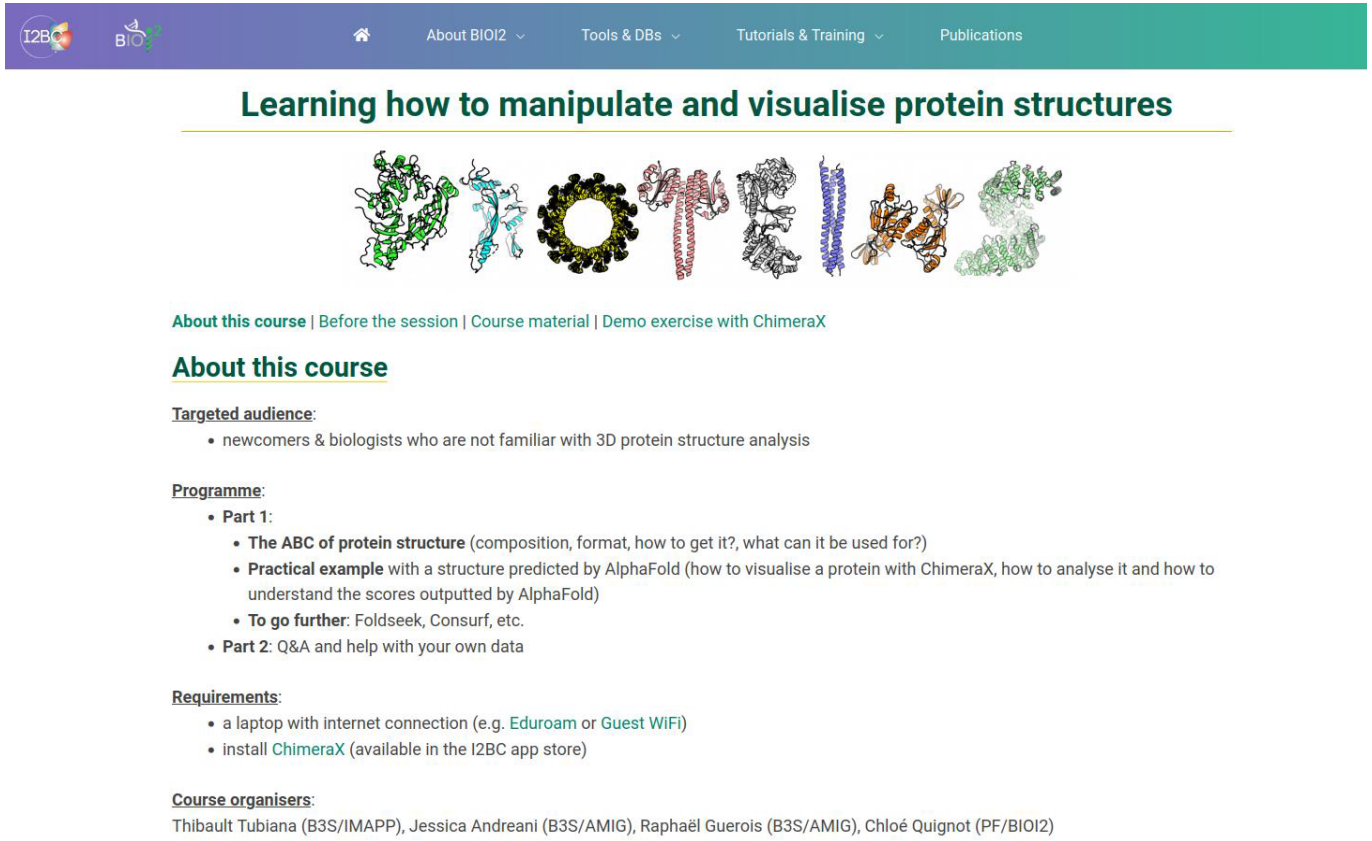
Who are we? Who are you?

Let's get to know each other!

- Thibault Tubiana (B3S/IMAPP, I2BC/CEA)
- Raphaël Guerois (B3S/AMIG, I2BC/CEA)
- Fabrice Leclerc (GENOMES/SSFA, I2BC)
- Chloé Quignot (BIOI2, I2BC/CEA)

Course material

https://bioi2.i2bc.paris-saclay.fr/training/visu-struct/

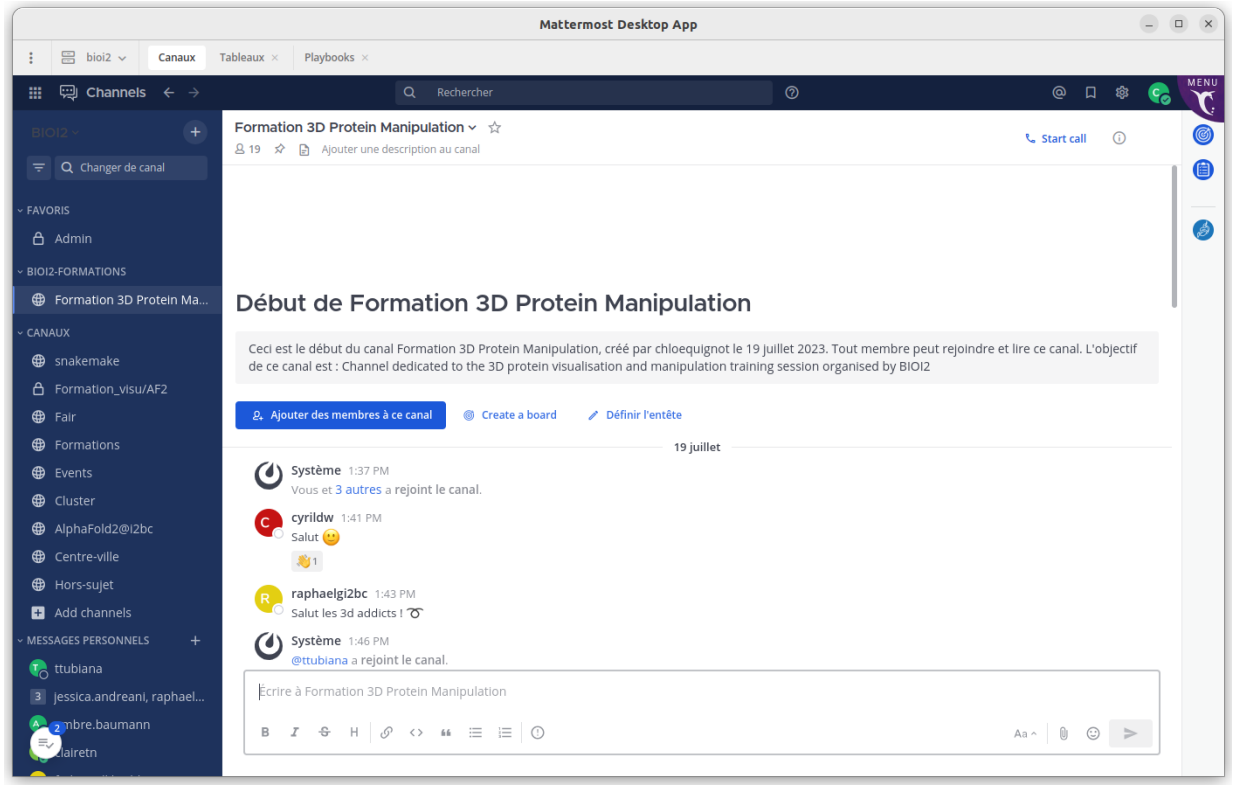


/!\ this presentation is password protected, type "alphafold" to unlock

BIOI2 discussion group

https://framateam.org/bioi2/channels/formation-3d-protein-manipulation
and https://framateam.org/bioi2/channels/alphafold2i2bc

To join, **create** yourself an account on FramaTeam using your I2BC account ("s'inscrire"): https://framateam.org/signup_user_complete/?id=y44u7h1x9jbyikhzmtbb6bw3hc



Summary

PART 1 – introduction to the 3D world of proteins

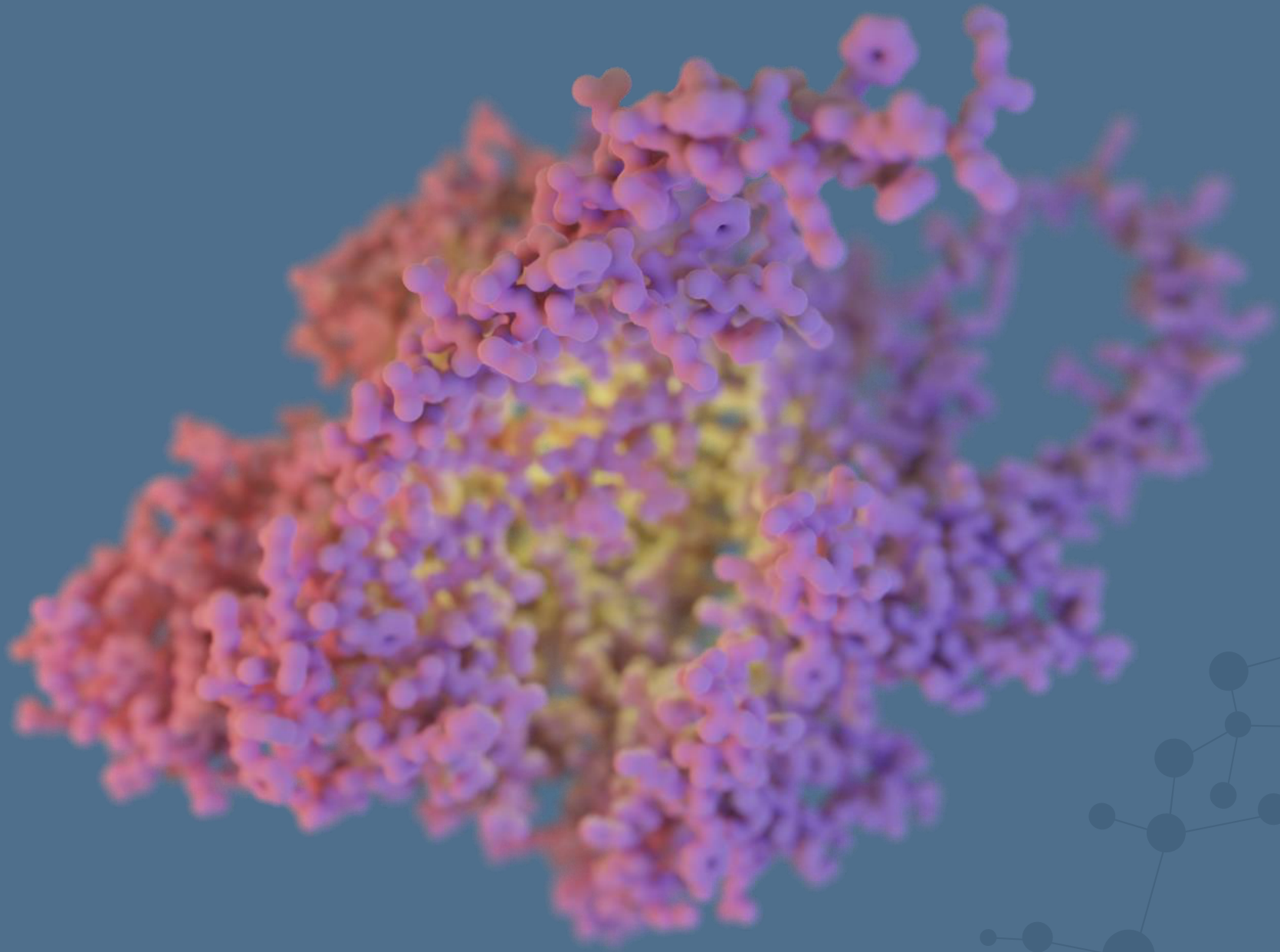
1. Protein structures, what are they?
2. Where to find them?
3. How to visualise them?

PART 2 – First steps with AlphaFold

1. What is AlphaFold?
2. How to use AlphaFold?
3. How to understand predictions?

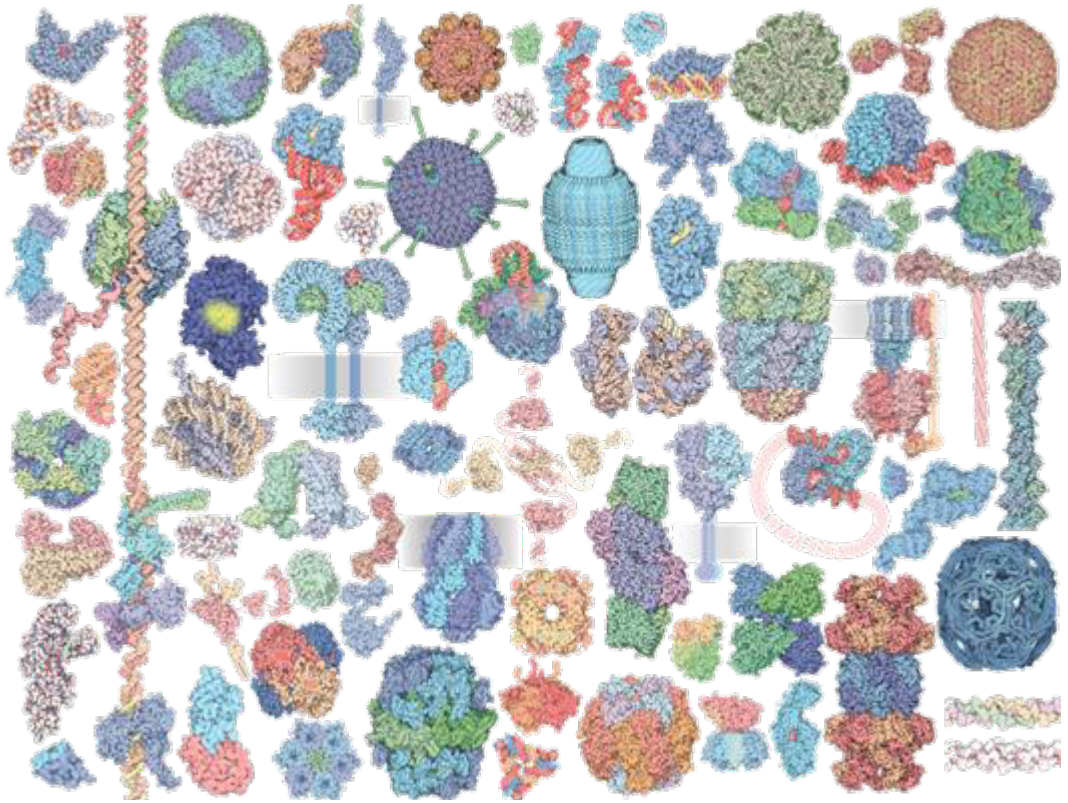
PART I

1. Protein structures: what are they?



Proteins : Why, What, How ?

- Proteins play diverse roles in the human body : Building, Repairing, Signaling, Defending….
- Experimental Methods or advances *in silico* methods can be used to obtain the atomic structure/model of proteins

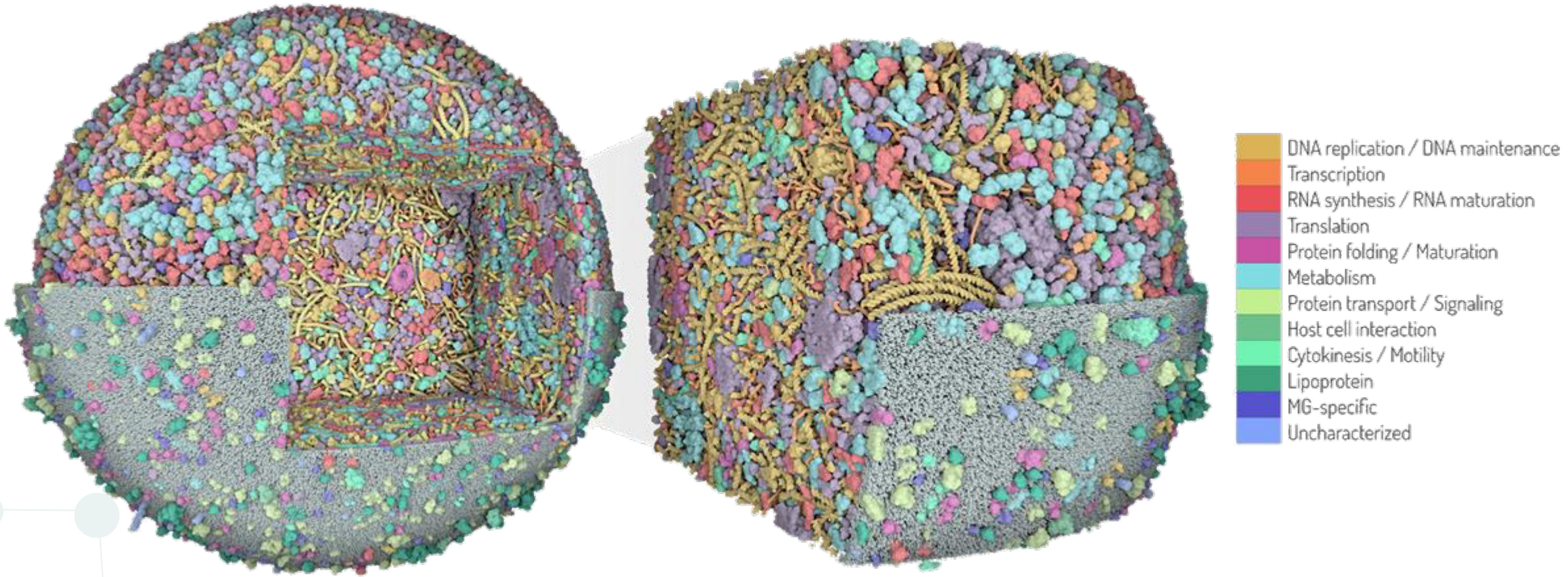


These are all representations of proteins or complexes of proteins

RCSB PDB molecules of the month by David S. Goodsell

Proteins : Why, What, How ?

Proteins are everywhere



3D Whole Cell Model of a Mycoplasma Bacterium

Made by Martina Maritan

Proteins : Why, What, How ?

Proteins are everywhere, mobile

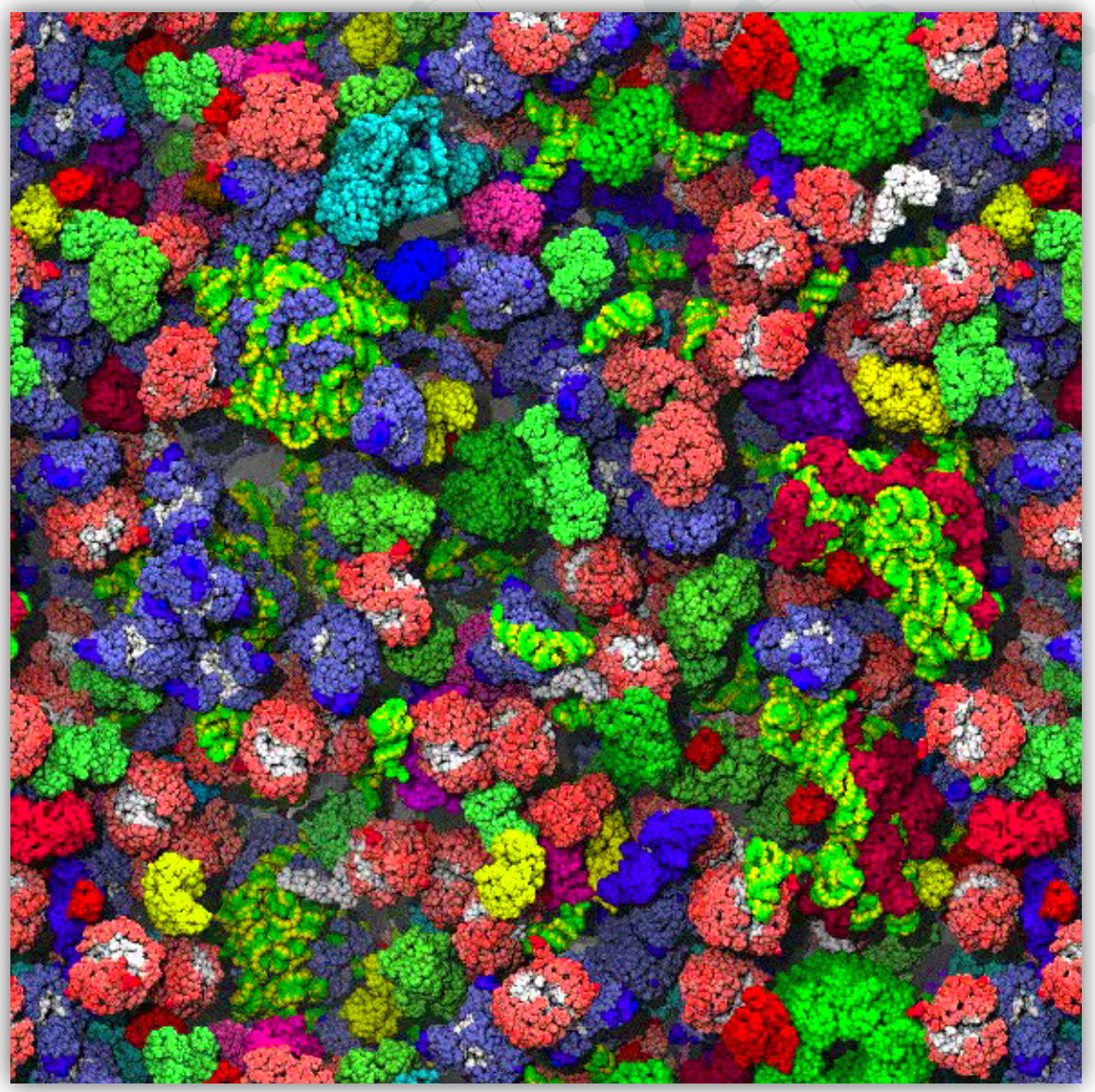
A

	Name	Mw	#
	Adk	24	14
	AhpC	187	7
	Asd	80	4
	Bcp	11	8
	CspC	7	72
	CysK	64	13
	DapA	125	2
	DnaK	41	11
	Efp	20	14
	Eno	91	18
	Fba	78	6
	Frr	21	7
	FusA	69	22

	Name	Mw	#
	GapA	142	10
	GlnA	621	1
	GltD	94	3
	GlyA	91	15
	GpmA	55	4
	Hns	5	7
	Hup	15	12
	IcdA	92	43
	IlvC	54	18
	Mdh	65	13
	MetE	84	213
	Mop	845	2

	Name	Mw	#
	PanB	140	2
	Pgk	41	26
	Pnp	190	3
	Ppa	116	9
	PpiB	18	7
	PurA	94	4
	PurC	42	7
	Pyr	308	3
	RpiA	46	3
	Rpo	260	4
	SerC	79	11
	SodA	46	13
	SodB	42	9

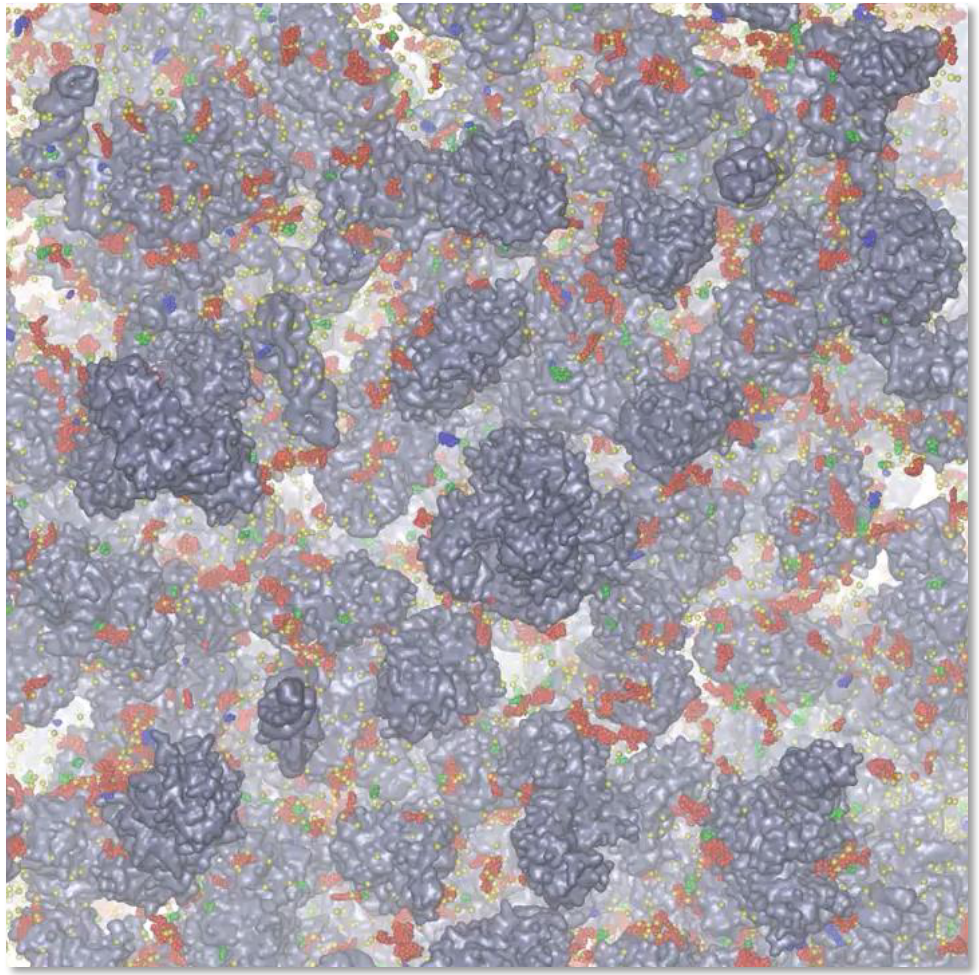
	Name	Mw	#
	Suc	142	4
	Tig	48	9
	TpiA	54	5
	Tsf	61	12
	TufA	84	181
	Upp	45	11
	UspA	31	7
	50S	1,355	10
	30S	788	10
	tRNA-C	24	37
	tRNA-Q	24	37
	tRNA-F	25	37
	GFP	26	8



McGuffee SR, Elcock AH, 2010. PLoS Comput Biol 6(3): e1000694.

Proteins : Why, What, How ?

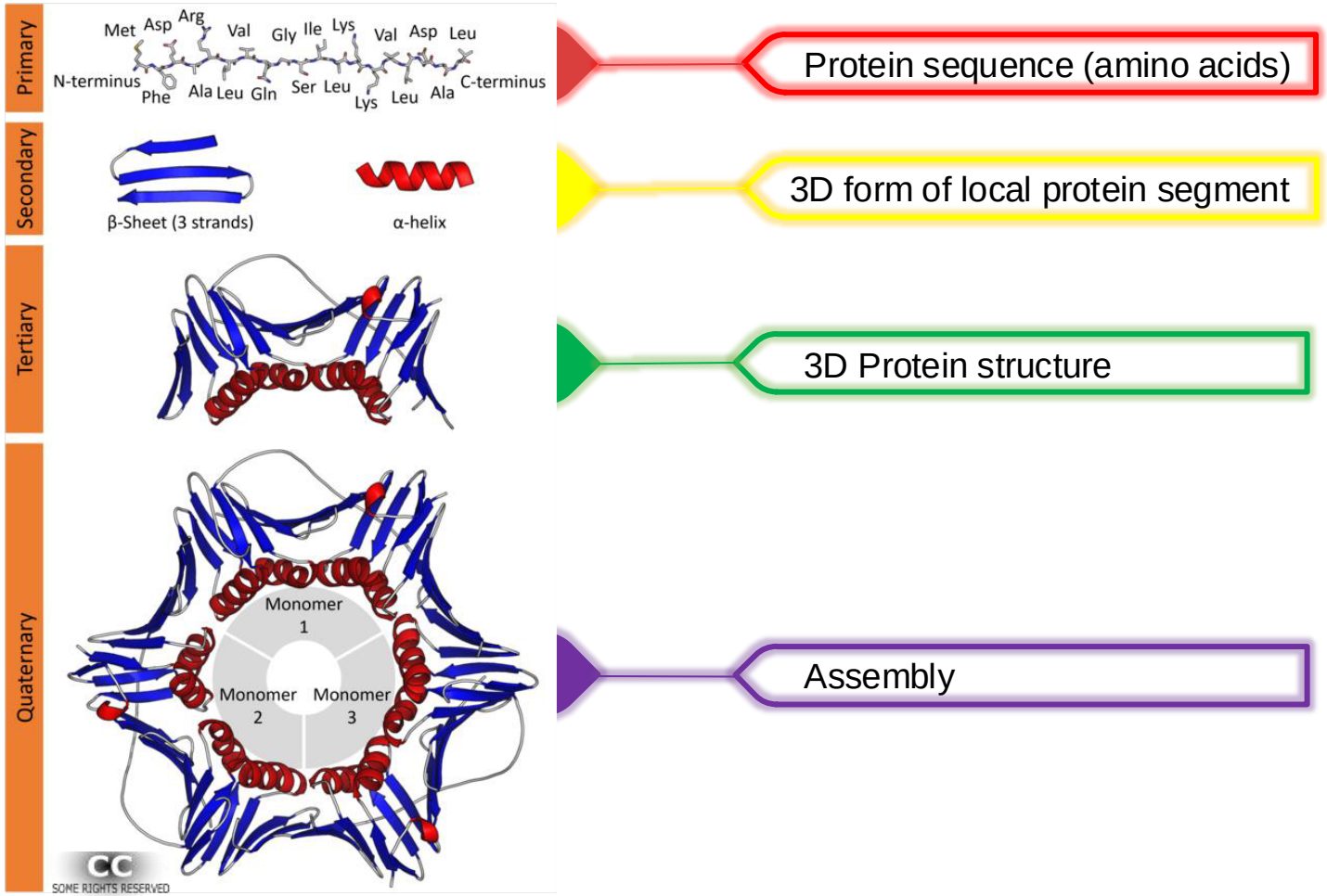
Proteins are everywhere, mobile, interacting



Yu, I., Mori, T., Ando, T., Harada, R., Jung, J., Sugita, Y., & Feig, M. (2016). Biomolecular interactions modulate macromolecular structure and dynamics in atomistic model of a bacterial cytoplasm. In eLife (Vol. 5). eLife Sciences Publications, Ltd. https://doi.org/10.7554/eLife.19274

Proteins : Why, What, How ?

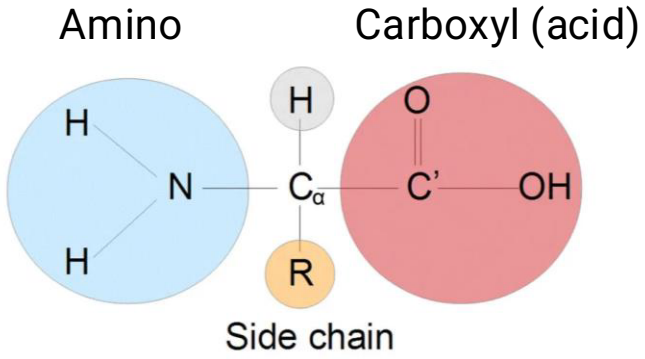
Proteins are made up of amino acids, which interact to form a 3D structure



https://en.wikipedia.org/wiki/Protein_secondary_structure

Proteins : Why, What, How ?

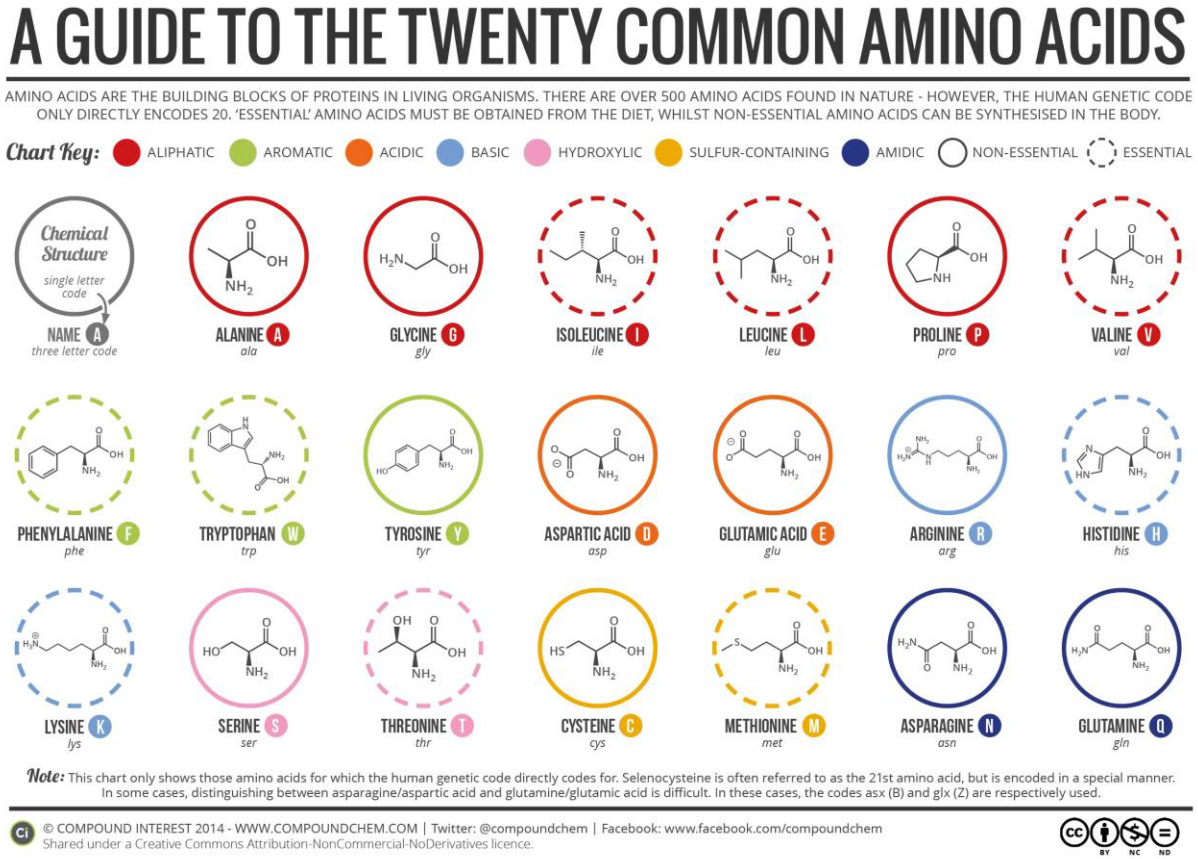
- Amino acids can be regrouped into families of similar properties (hydrophobic, hydrophilic, charged...)



They all have the same "frame":

-> **1 amino** and **1 carboxyl group** with which they hold hands with others (=> they're part of the "**backbone**")

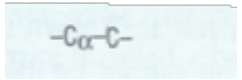
-> a carbon (Cα) which carries the **side chain**

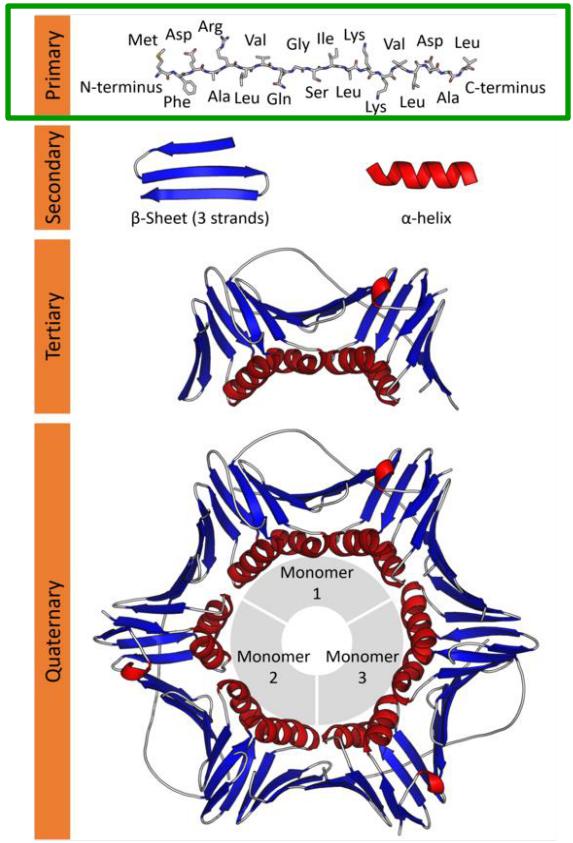


Protein world – From amino acid to protein

- Many chemical interactions stabilize proteins (polypeptides)
- Amino acids link together through peptide bonds => **Primary structure**

Interaction	Distance dependence	Typical distance	Free energy (bond dissociation enthalpies for the covalent bonds)
Covalent bond	-	1.5Å	356kJ/mole (610kJ/mole for a C=C bond)

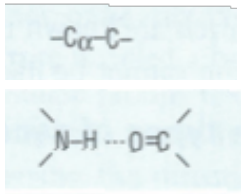




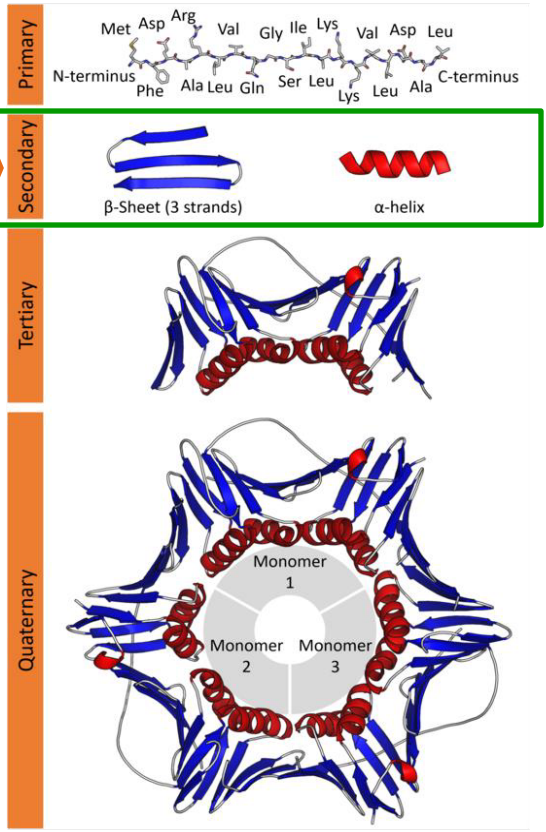
Protein world – From amino acid to protein

- Many chemical interactions stabilize proteins (polypeptides)
- **Secondary structures** are locally folded structures, mainly held by Hydrogen bonds (H-bonds)
  The most common secondary structures are α-helices and β-sheets... but there are others

Interaction	Distance dependence	Typical distance	Free energy (bond dissociation enthalpies for the covalent bonds)
Covalent bond	-	1.5Å	356kJ/mole (610kJ/mole for a C=C bond)
Hydrogen bond	Donor (here N), and acceptor (here O) atoms <3.5Å	3.0Å	2-6kJ/mole in water 12.5-21kJ/mole if either donor or acceptor is charged



hydrogen bonds
between the main-chain peptide groups



Protein world – From amino acid to protein

- Many chemical interactions stabilize proteins (polypeptides)
- **Secondary structures** are locally folded structures, mainly held by Hydrogen bonds (H-bonds)

α-helices are right-handed helices stabilized by H-bonds between the C=O group of residue "i" and the N-H group of residue "i+4"

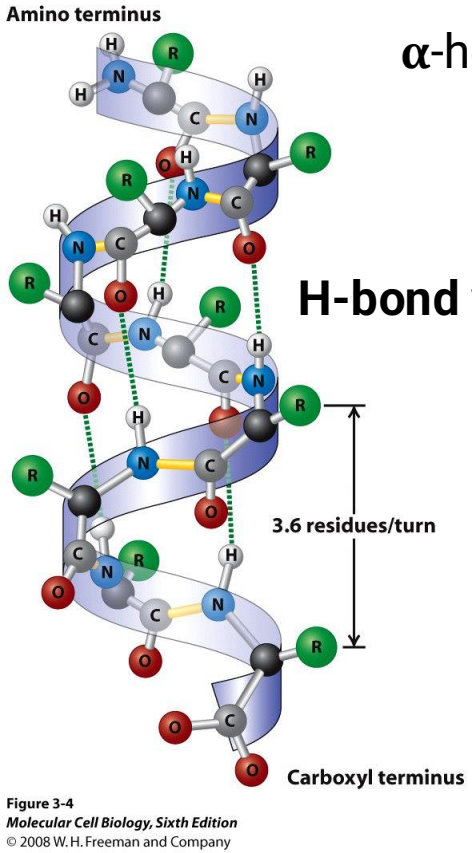
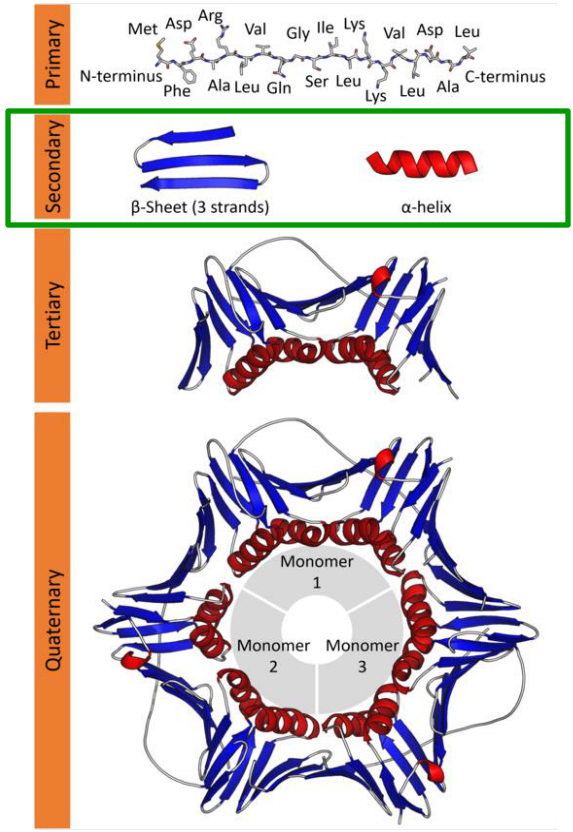


Figure 3-4
Molecular Cell Biology, Sixth Edition



📖 Want to know more ?
http://proteopedia.org/wiki/index.php/User:Stephen_Mills/Secondary_Structure:_Helices

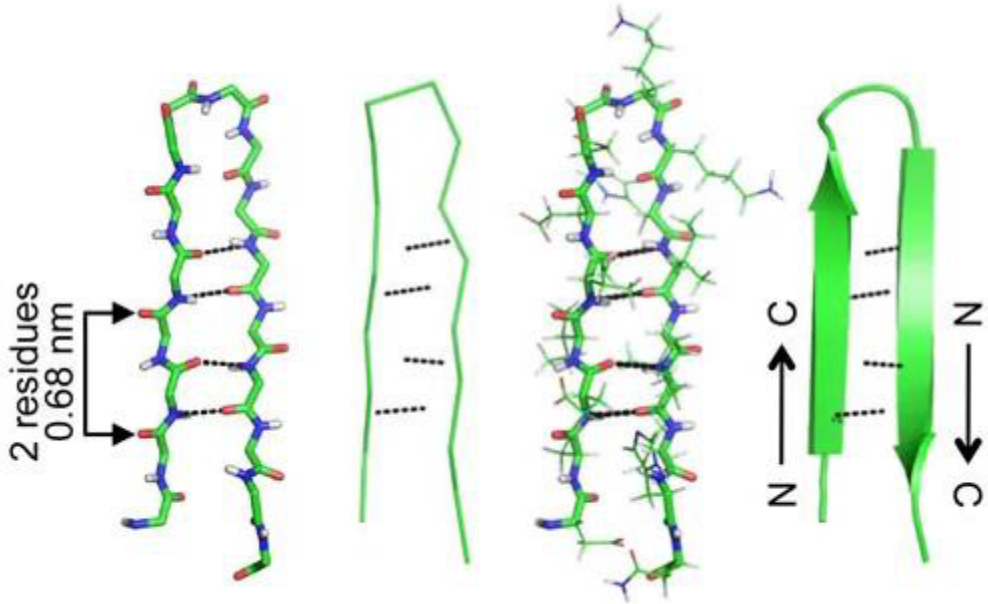
F. Rico, A. Rigato, L. Picas, and S. Scheuring, “Mechanics of proteins with a focus on atomic force microscopy,” Journal of Nanobiotechnology, vol. 11, no. Suppl 1, p. S3, 2013.

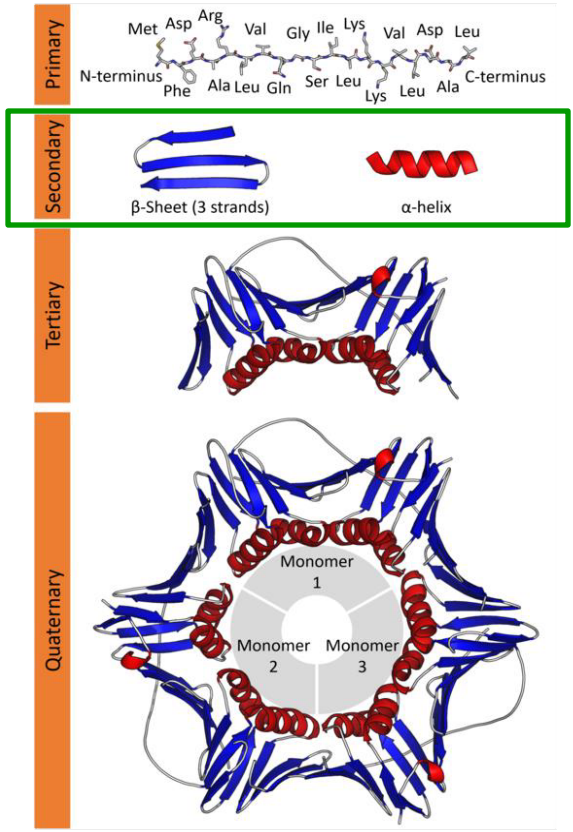
Protein world – From amino acid to protein

- Many chemical interactions stabilize proteins (polypeptides)
- **Secondary structures** are locally folded structures mainly held by Hydrogen bonds (H-bonds)

β-sheets *are made of laterally connected β-strands, stabilized by H-bonds, and forming a usually twisted, pleated sheet*

β-sheet





📖 Want to know more ?
http://proteopedia.org/wiki/index.php/User:Stephen_Mills/Secondary_Structure:_Helices

F. Rico, A. Rigato, L. Picas, and S. Scheuring, “Mechanics of proteins with a focus on atomic force microscopy,” Journal of Nanobiotechnology, vol. 11, no. Suppl 1, p. S3, 2013.

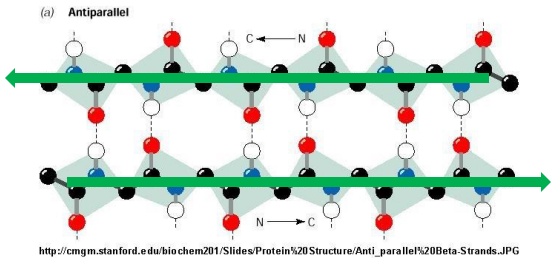
Protein world – From amino acid to protein

- Many chemical interactions stabilize proteins (polypeptides)
- **Secondary structures** are locally folded structures mainly held by Hydrogen bonds (H-bonds)

β-sheets *are made of laterally connected β-strands, stabilized by H-bonds, and forming a usually twisted, pleated sheet*

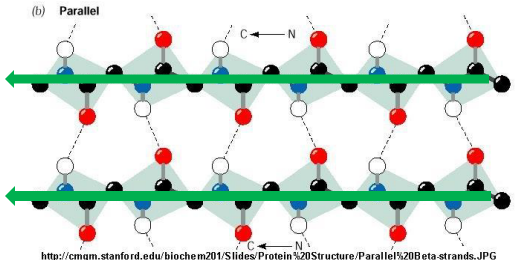
They can be parallel or antiparallel according to the relative orientation of the β-strands

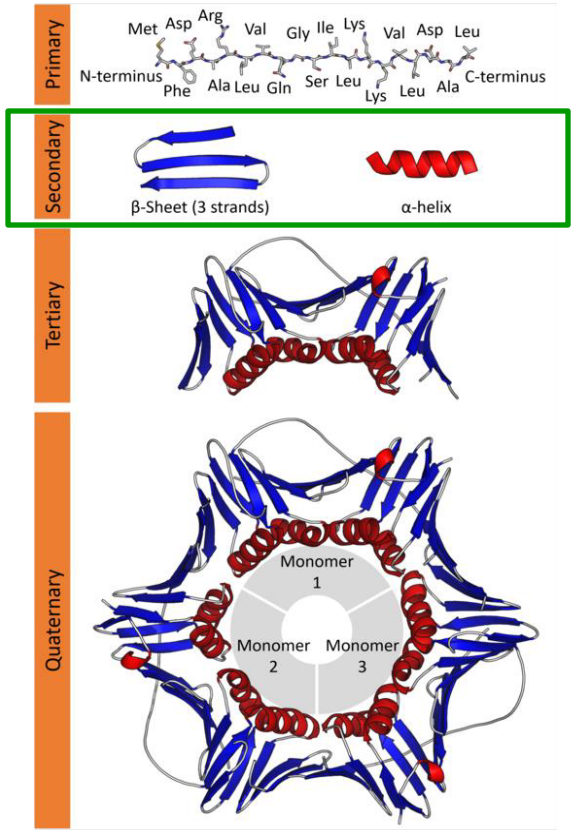
Antiparallel



<= MORE STABLE

Parallel



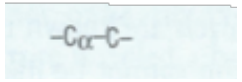
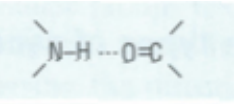
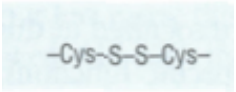
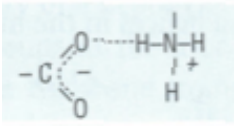
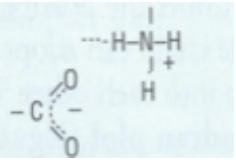
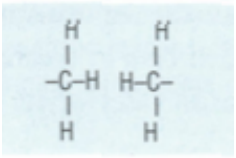


📖 Want to know more ?
http://proteopedia.org/wiki/index.php/User:Stephen_Mills/Secondary_Structure:_Helices

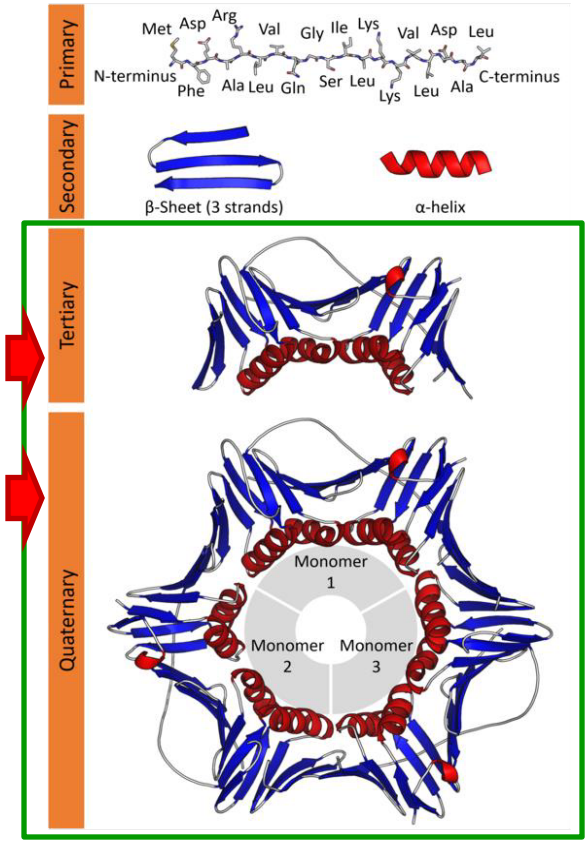
F. Rico, A. Rigato, L. Picas, and S. Scheuring, “Mechanics of proteins with a focus on atomic force microscopy,” Journal of Nanobiotechnology, vol. 11, no. Suppl 1, p. S3, 2013.

Protein world – From amino acid to protein

- Many chemical interactions stabilize proteins (polypeptides)
- **Tertiary structure** = the overall 3D structure of polypeptides
- Some proteins are made up of multiple polypeptide chains (subunits) = **Quaternary structure**

Interaction	Distance dependence	Typical distance	Free energy (bond dissociation enthalpies for the covalent bonds)	
Covalent bond	-	1.5Å	356kJ/mole (610kJ/mole for a C=C bond)	
Hydrogen bond	Donor (here N), and acceptor (here O) atoms <3.5Å	3.0Å	2-6kJ/mole in water 12.5-21kJ/mole if either donor or acceptor is charged	
Disulfide bond	-	2.2Å	167kJ/mole	
Salt bridge	Donor (here N), and acceptor (here O) atoms <3.5Å	2.8Å	12.5-17kJ/mole may be as high as 30kJ/mole for fully or partially buried salt bridges, less if the salt bridge is external	
Long-range electrostatic interaction	Depends on dielectric constant of medium. Screened by water. 1/r dependence	Variable	Depends on distance and environment. Can be very strong in nonpolar region but very weak in water	
Van der Waals interaction	Short range. Falls off rapidly beyond 4Å separation. 1/r^6 dependence	3.5Å	4kJ/mole (4-17 in protein interior) depending on the size of the group (for comparison, the average thermal energy of molecules at room temperature is 2.5 kJ/mole)	

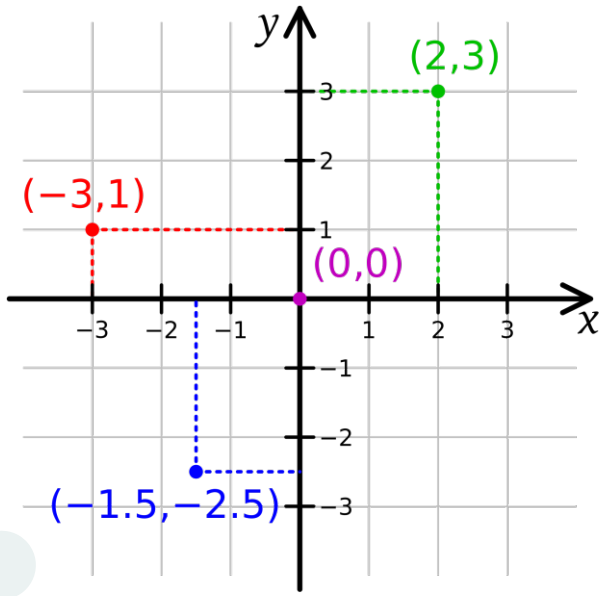
multiple bonds/ forces:
- hydrophobic interactions for protein folding,
- hydrogen & disulfide bonds for maintaining the stable structure



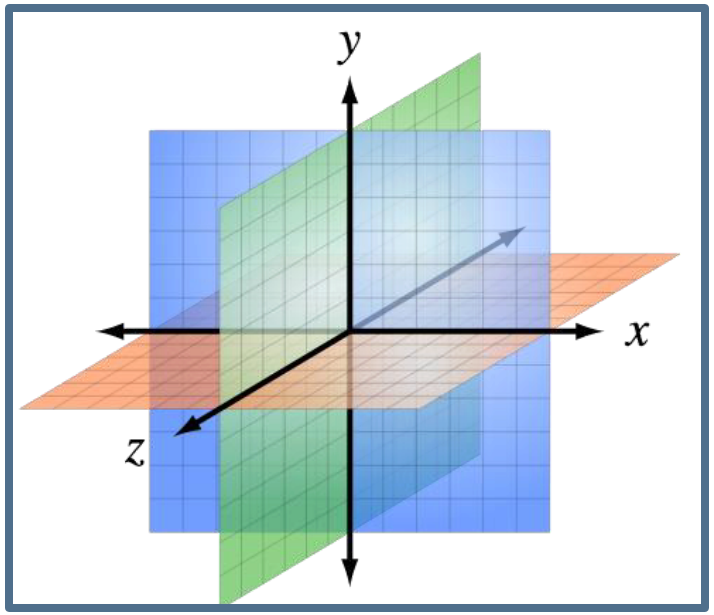
Protein world – coordinate system

- Proteins are 3D objects => described by 3D coordinates => (x, y, z) per atom
- Saved in a text file with a set syntax according to chosen format (.pdb, .cif, ...)

2D coordinate system



3D coordinate system



Protein world – data structure

- .pdb file structure, 1 line per atom (ids, types & coordinates), fixed length columns

```
column            1         2         3         4         5         6         7         8
number   12345678901234567890123456789012345678901234567890123456789012345678901234567890
```

ATOM	1	N	GLY	A	3	17.119	0.186	36.320	1.00	64.10	N	
ATOM	2	CA	GLY	A	3	16.944	-0.800	35.208	1.00	63.46	C	
ATOM	3	C	GLY	A	3	16.818	-0.087	33.851	1.00	61.22	C	
ATOM	4	O	GLY	A	3	15.721	0.337	33.463	1.00	62.81	O	
ATOM	5	N	PRO	A	4	17.944	0.077	33.129	1.00	57.39	N	
ATOM	6	CA	PRO	A	4	17.950	0.742	31.815	1.00	53.27	C	
ATOM	7	C	PRO	A	4	18.005	-0.247	30.629	1.00	49.78	C	
ATOM	8	O	PRO	A	4	19.086	-0.678	30.218	1.00	48.17	O	
ATOM	9	CB	PRO	A	4	19.191	1.613	31.898	1.00	54.33	C	
ATOM	10	CG	PRO	A	4	20.161	0.686	32.625	1.00	55.45	C	
ATOM	11	CD	PRO	A	4	19.305	0.019	33.701	1.00	55.83	C	
field id 1	2	3 4	5	6	7 8	9	10	11	12	13	14	15

field id	definition
1	"ATOM " or "HETATM"
2	atom serial number
3	atom name
4	alternate location indicator
5	residue name
6	chain identifier
7	residue sequence number
8	code for insertion of residues
9	orthogonal coordinates for X (in Angstroms)
10	orthogonal coordinates for Y (in Angstroms)
11	orthogonal coordinates for Z (in Angstroms)
12	occupancy
13	temperature factor
14	element symbol
15	charge on the atom

Pierre Poulain - http://cupnet.net/pdb-format/

PROBLEM!!! : 99999 atoms ID max (after all atoms are 99999 or start from 1 again)

Protein world – data structure

- .cif (PDBx/mmCIF) file structure – more recent – 1 line per atom but more flexible/less limitations, supporting data representing large structures, complex chemistry, and new & hybrid elements

```
loop_
_atom_site.group_PDB
_atom_site.id
_atom_site.type_symbol
_atom_site.label_atom_id
_atom_site.label_alt_id
_atom_site.label_comp_id
_atom_site.label_asym_id
_atom_site.label_entity_id
_atom_site.label_seq_id
_atom_site.pdbx_PDB_ins_code
_atom_site.Cartn_x
_atom_site.Cartn_y
_atom_site.Cartn_z
_atom_site.occupancy
_atom_site.B_iso_or_equiv
_atom_site.pdbx_formal_charge
_atom_site.auth_seq_id
_atom_site.auth_comp_id
_atom_site.auth_asym_id
_atom_site.auth_atom_id
_atom_site.pdbx_PDB_model_num
ATOM   1    N  N   . VAL A 1 1   ? 6.204  16.869 4.854  1.00 49.05 ? 1 VAL A N   1
ATOM   2    C  CA  . VAL A 1 1   ? 6.913  17.759 4.607  1.00 43.14 ? 1 VAL A CA  1
ATOM   3    C  C   . VAL A 1 1   ? 8.504  17.378 4.797  1.00 24.80 ? 1 VAL A C   1
ATOM   4    O  O   . VAL A 1 1   ? 8.805  17.011 5.943  1.00 37.68 ? 1 VAL A O   1
ATOM   5    C  CB  . VAL A 1 1   ? 6.369  19.044 5.810  1.00 72.12 ? 1 VAL A CB  1
ATOM   6    C  CG1 . VAL A 1 1   ? 7.009  20.127 5.418  1.00 61.79 ? 1 VAL A CG1 1
ATOM   7    C  CG2 . VAL A 1 1   ? 5.246  18.533 5.681  1.00 80.12 ? 1 VAL A CG2 1
```

Each block starts with `loop_`

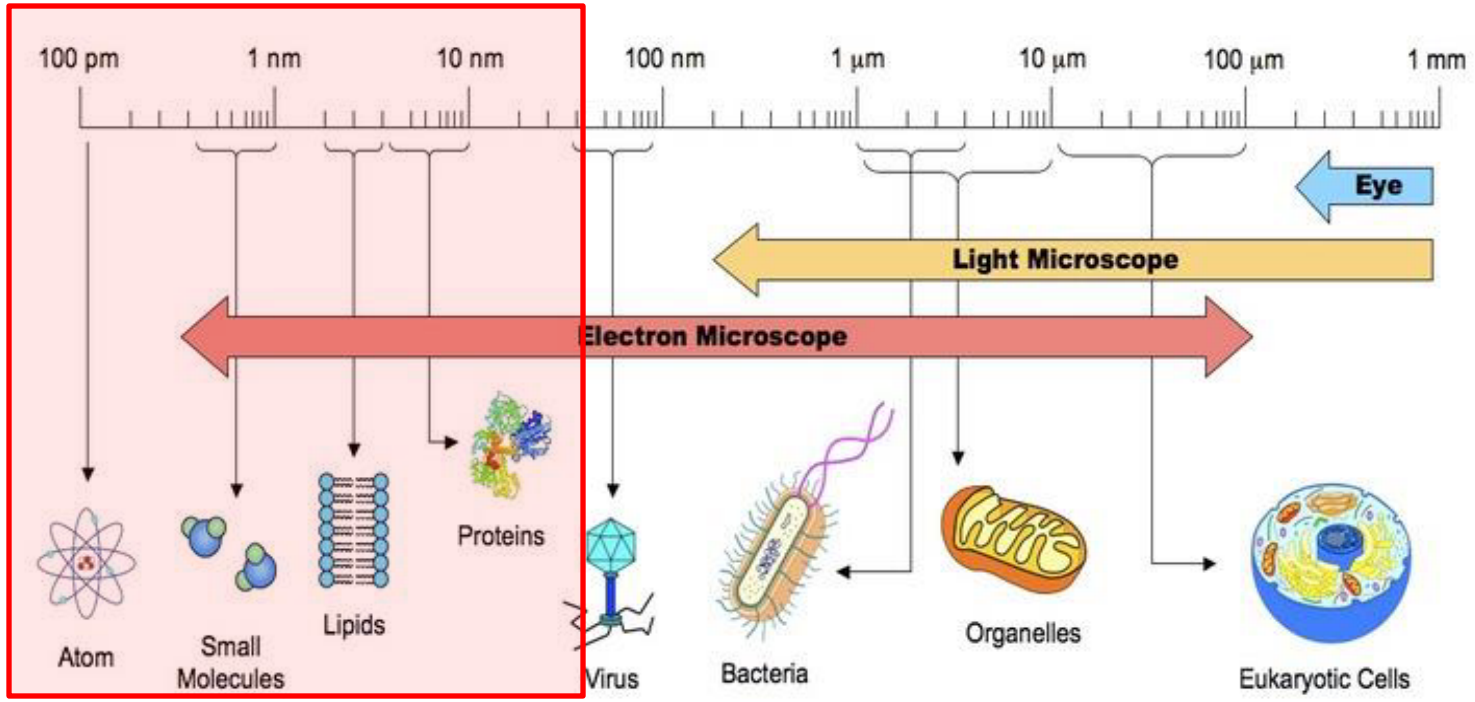
A first section to describe the columns below

1 atom per line, each column (labelled above) is separated by white space

Protein world – coordinate system

- **Be careful of the unit !!!**
  - **Ångström (Å) : (10^{-10} m)**
  - nanometer (nm) : (10^{-9} m)

Most used unit



Protein world – coordinate system

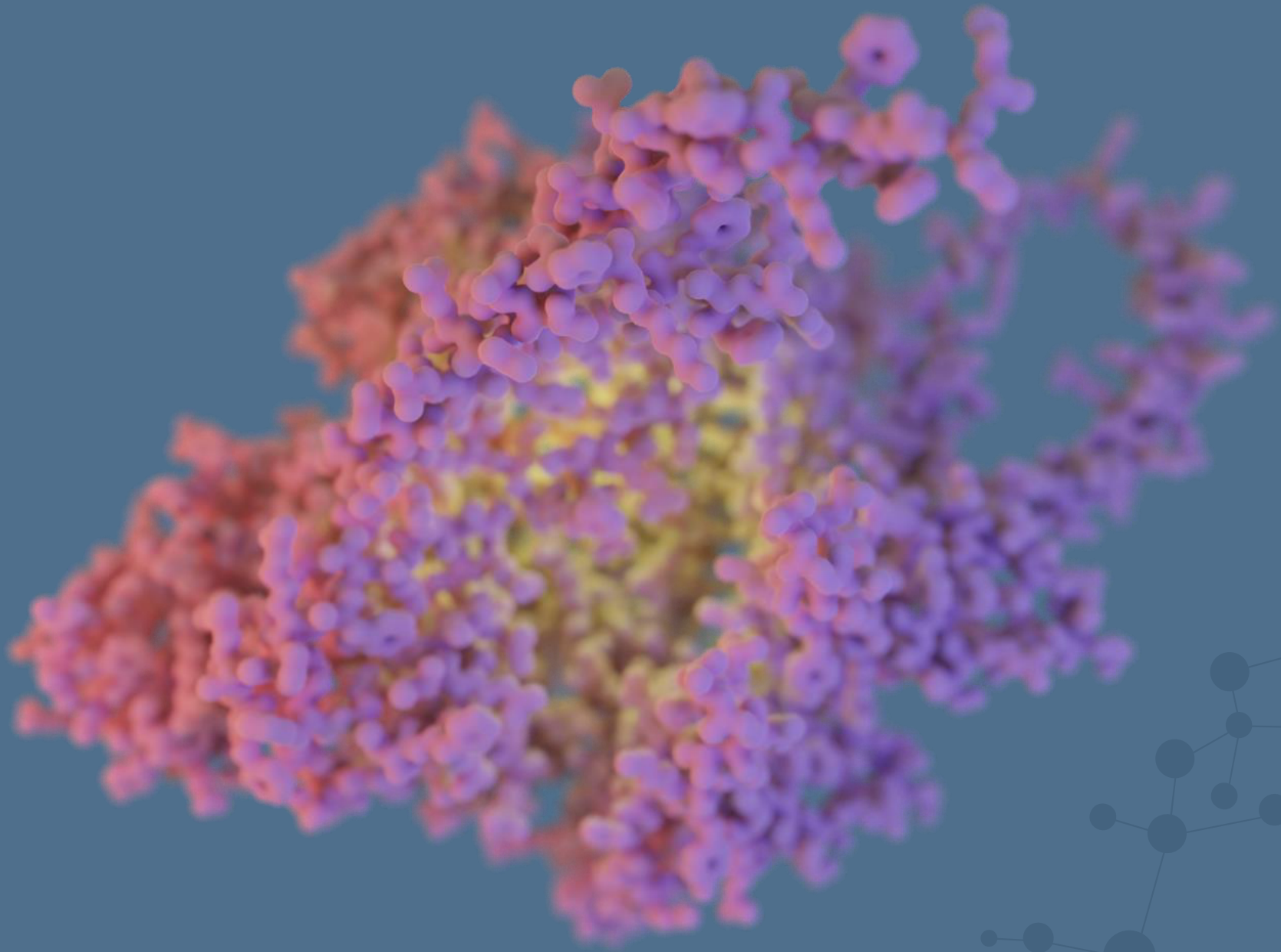
- **Be careful of the unit !!!**
  - **Ångström (Å) : (10^{-10} m)**
  - nanometer (nm) : (10^{-9} m)

Most used unit

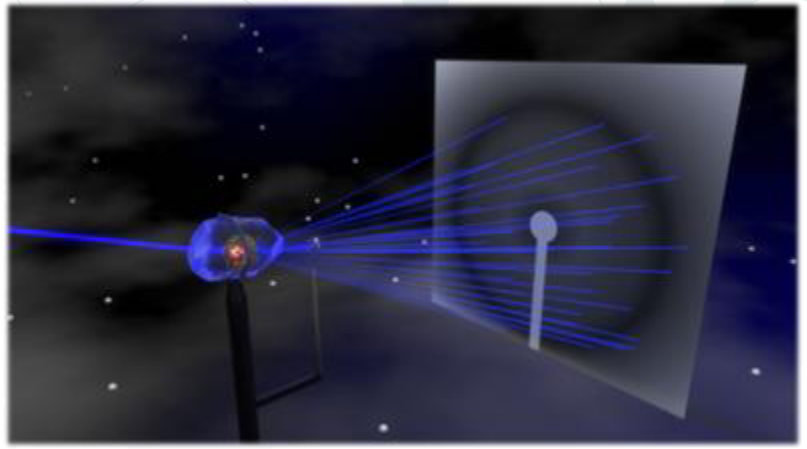
Remember the Mars Climate Orbiter incident from 1999?

PART I

2. Where to find them?

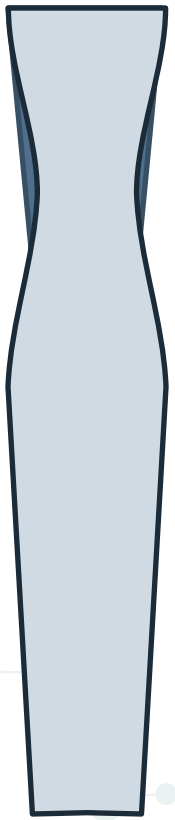


Protein world – Get the structures



For your own culture

How to determine 3D protein structures??

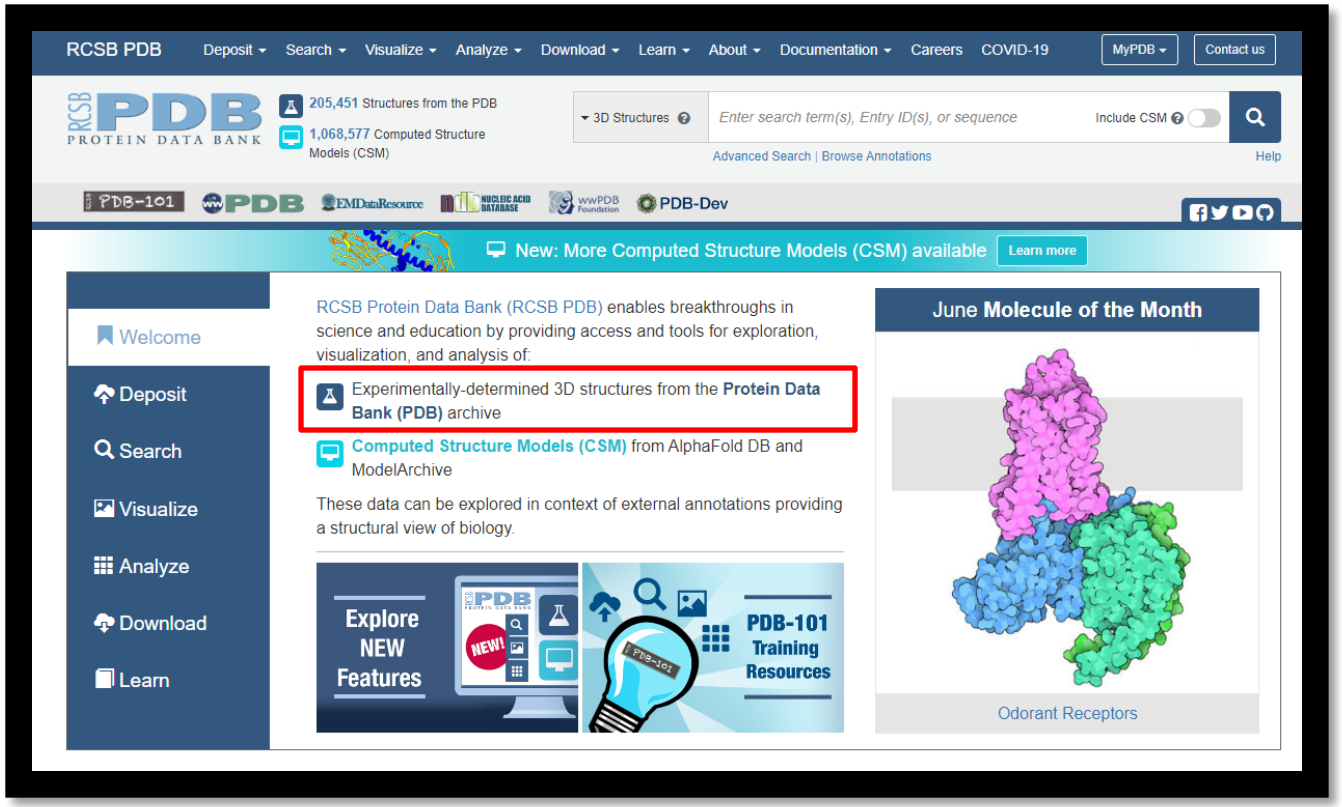


- Crystallography
  - *Compute electronic density from Xray diffraction*
- Nuclear Magnetic Resonance (NMR)
  - *Compute models from distance between amino acids*
- Cryo-EM
  - *Compute electronic density from electronic microscopy pictures*
- Modelling
  - *Compute a model 'from scratch' (de novo or based on an existing 3D structure)*

Protein world - Database

Where to find structures ?

→ Protein Data Bank (PDB) with >226K experimental structures *(october 2024)*



Experimental structures

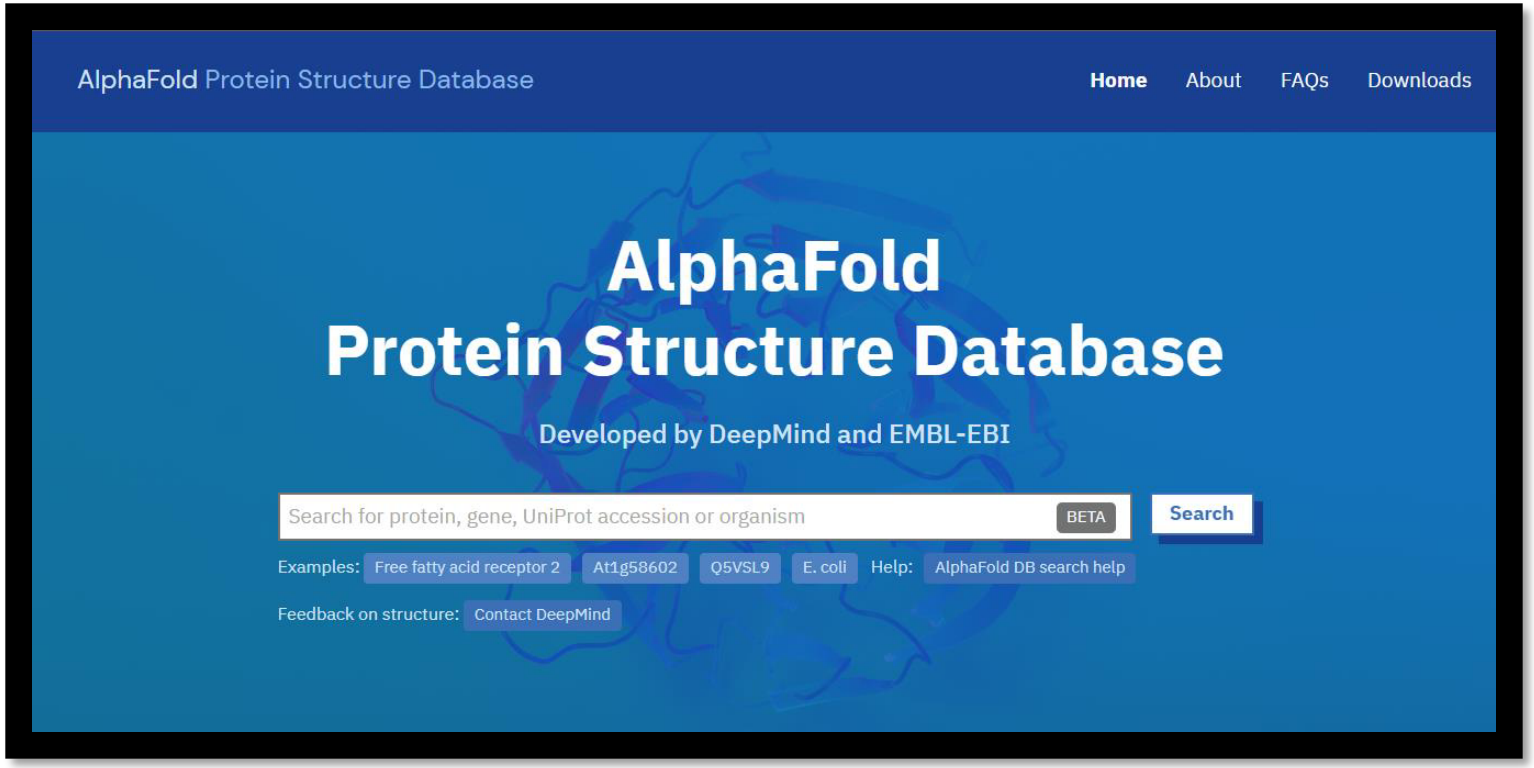
RCSB PDB web site also includes >1M models from AlphaFold DB and ModelArchive

Protein world - Database

Where to find structures ?

→ AlphaFold Database (214M+ models, including 48 complete proteomes)

https://alphafold.ebi.ac.uk



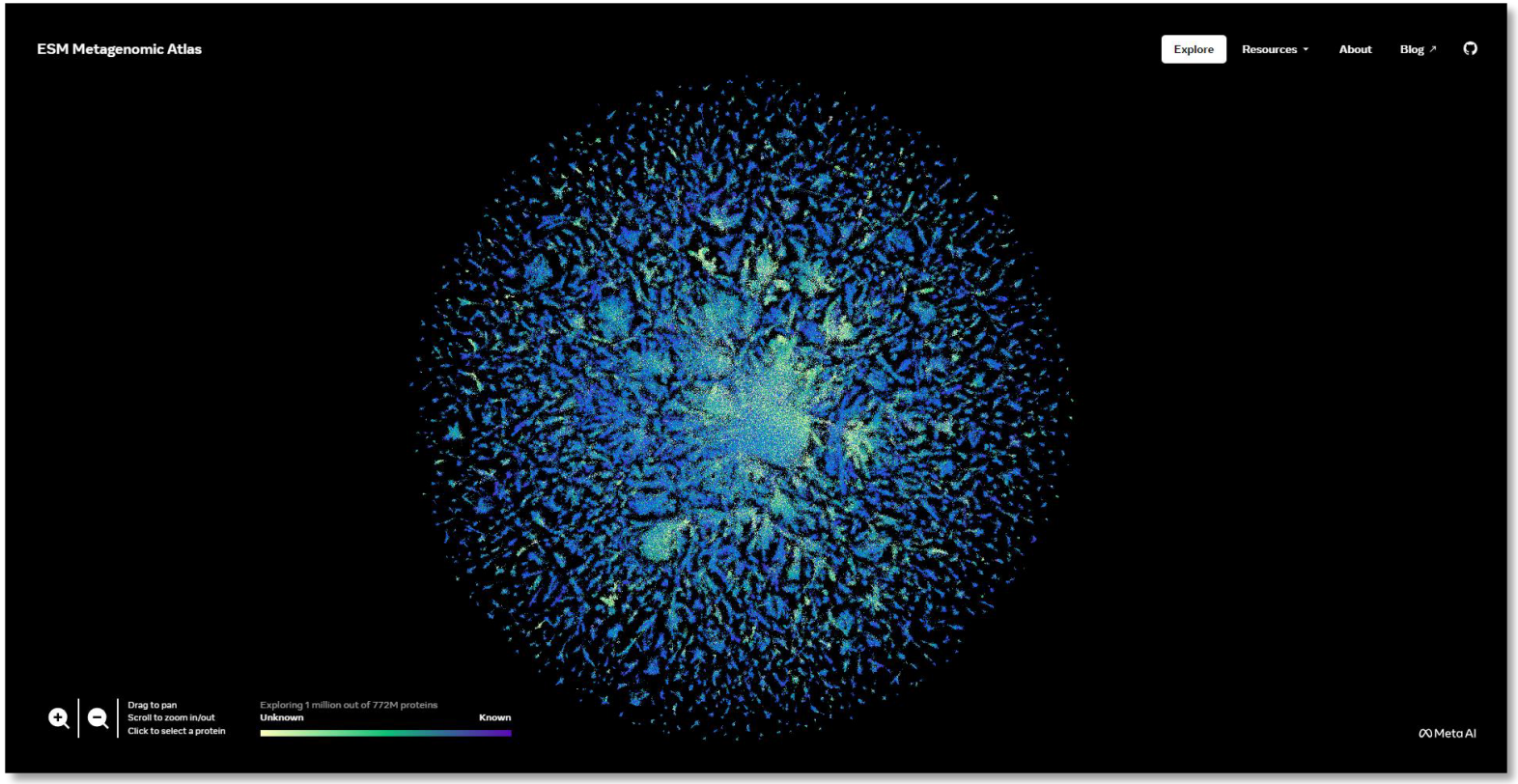
Models

Contains approx. all proteins from UniProt, excluding viral proteins and proteins exceeding their sequence length range (>16AA & <1280AA for UniProt / <2700AA for proteomes) - url

Protein world - Database

Where to find structures ?

→ ESM Metagenomic Atlas (772M+ models) (https://esmatlas.com/explore)



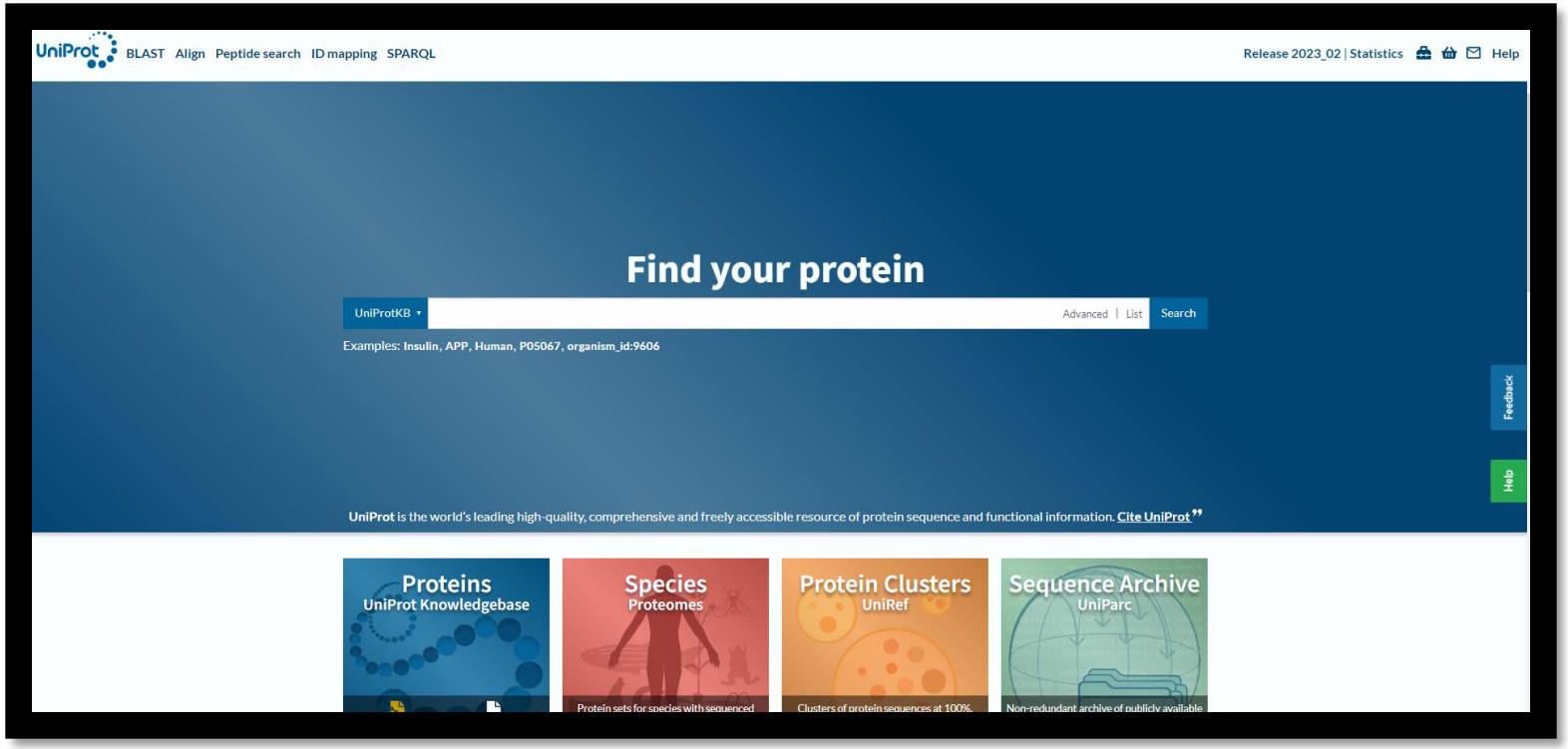
Models

Based on protein sequences in the MGnify database

Protein world - Database

Where to find structures ?

→ Or simply… Uniprot (https://www.uniprot.org)



Seq + Structure (database links)

Teacher Demo
From sequence to structure: querying DBs

What protein is this? Do we have a structure or model for it in the public databases?

```
MRIFAVFIFMTYWHLLNAFTVTVPKDLYVVEYGSNMTIECKFPVEKQLDLAALIVYWEME
DKNIIQFVHGEEDLKVQHSSYRQRARLLKDQLSLGNAALQITDVKLQDAGVYRCMISYGG
ADYKRITVKVNAPYNKINQRILVVDPVTSEHELTCQAEGYPKAEVIWTSSDHQVLSGKTT
TTNSKREEKLFNVTSTLRINTTTNEIFYCTFRRLDPEENHTAELVIPELPLAHPPNERTH
LVILGAILLCLGVALTFIFRLRKGRMMDVKKCGIQDTNSKKQSDTHLEET
```

Teacher Demo
From sequence to structure: querying DBs

What protein is this? Do we have a structure or model for it in the public databases?

```
MRIFAVFIFMTYWHLLNAFTVTVPKDLYVVEYGSNMTIECKFPVEKQLDLAALIVYWEME
DKNIIQFVHGEEDLKVQHSSYRQRARLLKDQLSLGNAALQITDVKLQDAGVYRCMISYGG
ADYKRITVKVNAPYNKINQRILVVDPVTSEHELTCQAEGYPKAEVIWTSSDHQVLSGKTT
TTNSKREEKLFNVTSTLRINTTTNEIFYCTFRRLDPEENHTAELVIPELPLAHPPNERTH
LVILGAILLCLGVALTFIFRLRKGRMMDVKKCGIQDTNSKKQSDTHLEET
```

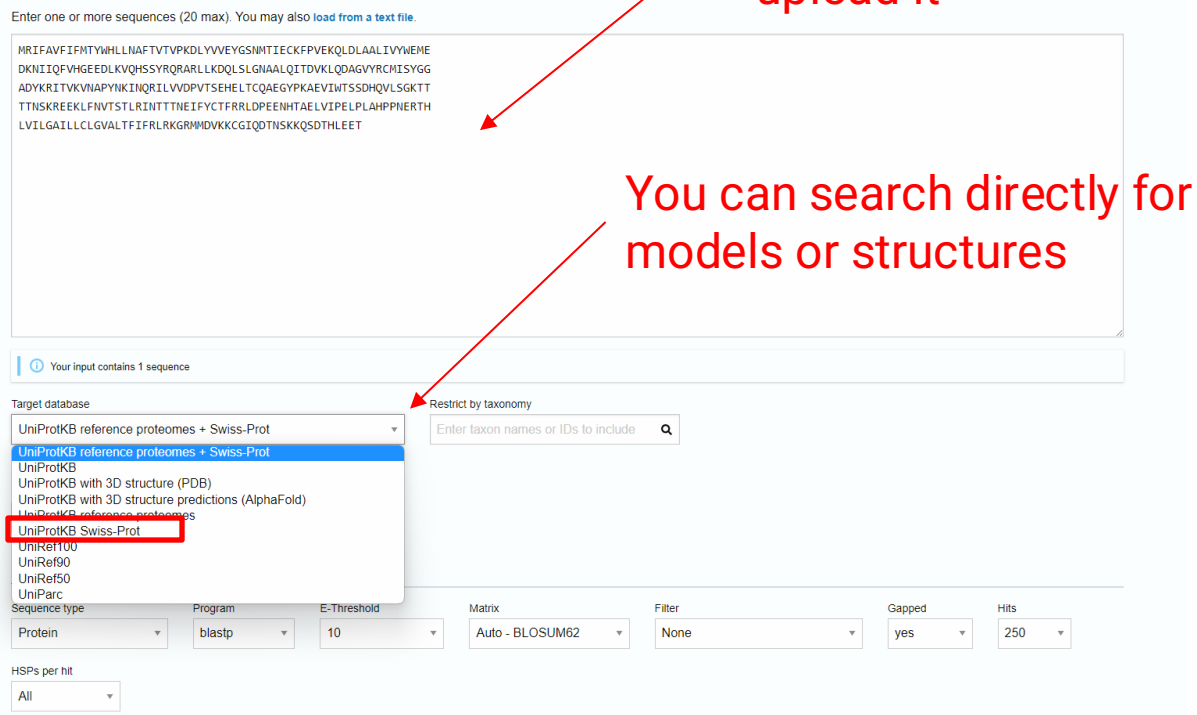
/!\ watch out for invisible characters when copy-pasting

> Protein Blast (blastp) - through UniProt:
> https://www.uniprot.org/blast

Swissprot : Only reviewed sequences

Copy/paste your fasta or upload it

You can search directly for models or structures



Teacher Demo
From sequence to structure: querying DBs

Blast output summary page:

BLAST 250 results found in UniProtKB

Overview | Taxonomy | Hit Distribution | Text Output | Input Parameters | API Request

BLAST | Align | Map IDs | Download | Add | Customize columns | Resubmit

Entry	Entry Name	Protein Names	Gene Names	Organism	Length	20 40 60 80 100 120 140 160 180 200 220 240 260 280
Q9NZQ7	PD1L1_HUMAN	Programmed cell death 1 ligand 1 [...]	**CD274**, B7H1, PDCD1L1, PDCD1LG1, PDL1	Homo sapiens (Human)	290 AA	100% 1511 0
A0A2R9B063	A0A2R9B063_PANPA	CD274 molecule	**CD274**	Pan paniscus (Pygmy chimpanzee) (Bonobo)	290 AA	99.7% 1503 0
H2QWZ8	H2QWZ8_PANTR	CD274 molecule	**CD274**	Pan troglodytes (Chimpanzee)	290 AA	99.7% 1503 0
G3QZN5	G3QZN5_GORGO	CD274 molecule	**CD274**	Gorilla gorilla gorilla (Western lowland gorilla)	290 AA	99% 1486 0
H2PS75	H2PS75_PONAB	CD274 isoform 2[...]	**CD274**, CR201_G0001691	Pongo abelii (Sumatran orangutan) (Pongo pygmaeus abelii)	290 AA	98.6% 1480 0

This protein is reviewed!

Protein Name

Origin (organism)

Protein ID (NCBI)

Sequence Identity

Teacher Demo
From sequence to structure: querying DBs

UniProt summary page for the best hit (Q9NZQ7):

All data about the protein

Let’s check if we have a structure

Teacher Demo
From sequence to structure: querying DBs

UniProt summary page for the best hit (Q9NZQ7):

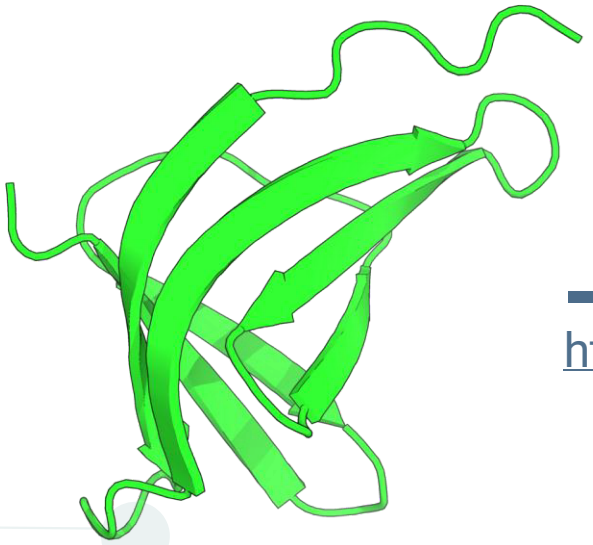
SOURCE	IDENTIFIER	METHOD	RESOLUTION	CHAIN	POSITIONS	LINKS	
-- Select --		-- Select --					
PDB	7C88	X-ray	2.00 Å	C/M	1-136	PDBe · RCSB-PDB · PDBj · PDBsum	Foldseek
PDB	7CZD	X-ray	1.64 Å	B/D	19-134	PDBe · RCSB-PDB · PDBj · PDBsum	Foldseek
PDB	7DCV	NMR		A	232-290	PDBe · RCSB-PDB · PDBj · PDBsum	Foldseek
PDB	7DY7	X-ray	2.42 Å	A/B	18-134	PDBe · RCSB-PDB · PDBj · PDBsum	Foldseek
PDB	7NLD	X-ray	2.30 Å	A/B/C/D/E/F	18-134	PDBe · RCSB-PDB · PDBj · PDBsum	Foldseek
PDB	7OUN	X-ray	1.90 Å	A	17-134	PDBe · RCSB-PDB · PDBj · PDBsum	Foldseek
PDB	7SJQ	X-ray	2.00 Å	A	18-134	PDBe · RCSB-PDB · PDBj · PDBsum	Foldseek
PDB	7TPS	X-ray	3.15 Å	B/D	19-227	PDBe · RCSB-PDB · PDBj · PDBsum	Foldseek
PDB	7VUN	X-ray	2.70 Å	A/B/C/D/E/F/G/H	18-134	PDBe · RCSB-PDB · PDBj · PDBsum	Foldseek
AlphaFold	AF-Q9NZQ7-F1	Predicted			1-290	AlphaFold	Foldseek

Download the coordinates
=> it's a text file

Yes! Plenty of structures covering different regions of our protein:
- 49 experimental structures (47 X-ray + 2 NMR) on segments
- 1 predicted structure (AlphaFold) of the whole sequence

Teacher Demo
From Structure to structure: FOLDSEEK

I got this fold in one model → What is this ?



https://search.foldseek.com/search

Domain Classification

Results for job: 69wZpG2nieU88Vp_BOMBZTn1YgTD6qmNtbVHjA

(252) AFDB50 (1000) BFMD (214) **CATH50 (274)** GMGCL_ID (197) MGNIFY_ESM30 (1(

CATH50 274 hits

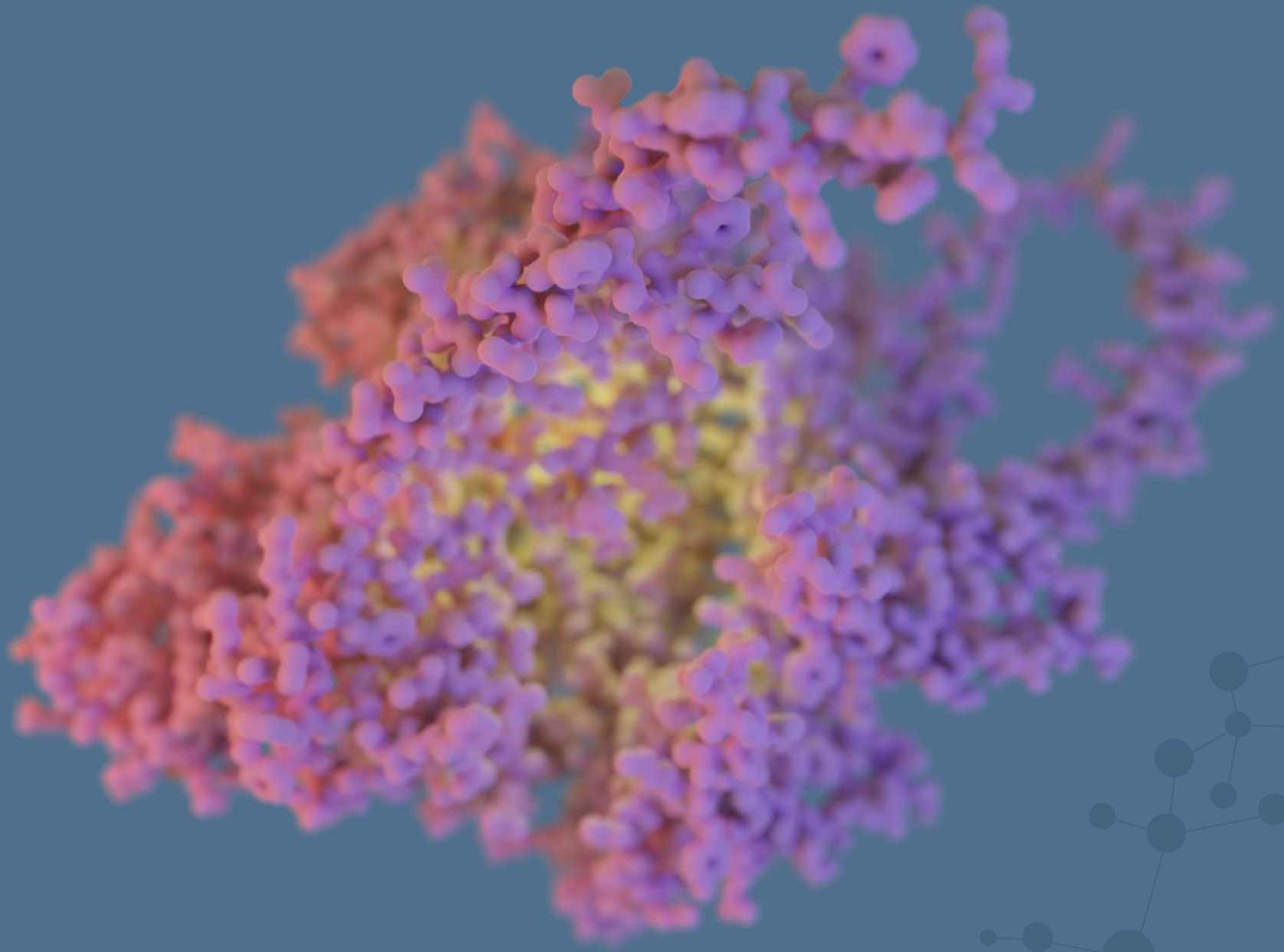
GRAPHICAL NUMERIC

Target	Scientific Name	Prob.	Seq. Id.	E-Value	Position in query	Alignment
2.30.29.30 O36027 2-131	Schizosaccharomyces ...	0.99	10	1.31e-1	5 72	
2.30.29.30 O08816 6-147	Rattus norvegicus	0.99	10.7	2.68e-1	7 72	
2.30.29.30 A0A1D8PU0...	Candida albicans	0.99	10	1.24e-1	5 72	
2.30.29.30 Q9VAT0 11-...	Drosophila melanogaster	0.98	12.5	1.93e-1	1 72	
2.30.29.30 Q9GSG9 2-...	Dictyostelium discoideum	0.97	8.8	2.15e-1	4 72	
1mkeA00	Rattus norvegicus	0.96	10.7	2.15e-1	6 72	
5nc7B00	Listeria monocytogenes...	0.95	14.2	1.73e-1	5 72	
2.30.29.30 P42768 15-...	Homo sapiens	0.90	9.3	8.99e-1	7 72	
2.30.29.30 Q8I5X1 250...	Plasmodium falciparum	0.89	6.5	1.46e-1	9 72	

→ 2.30.29.30 → PH DOMAIN

PART I

3. How to visualise them?



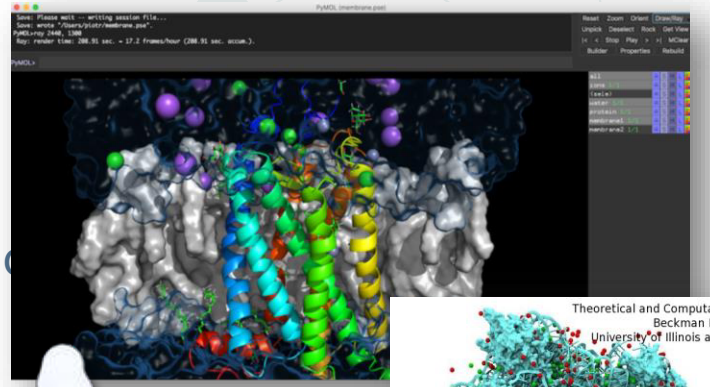
Protein viewers

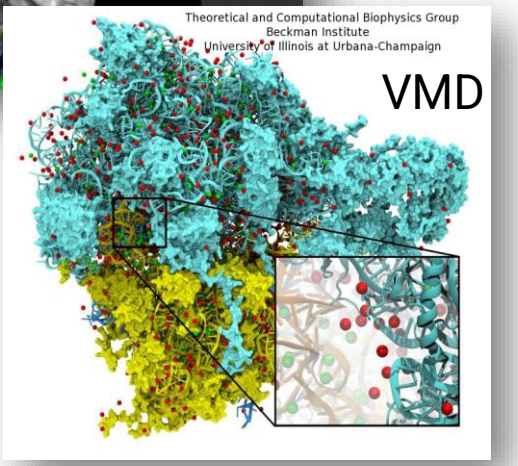
Free Protein viewers

- Standalone softwares
  - Pymol (educational version free, or open source versi
  - VMD
  - UnityMol
  - **ChimeraX**

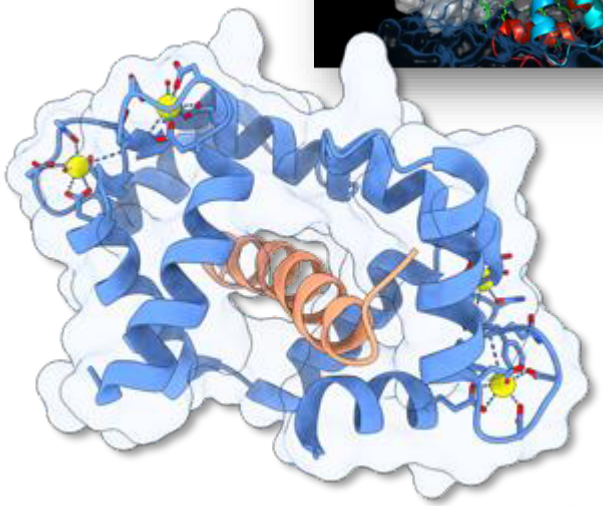
- Web based
  - 3DProteinImaging
  - Molstar
  - NGLViewer

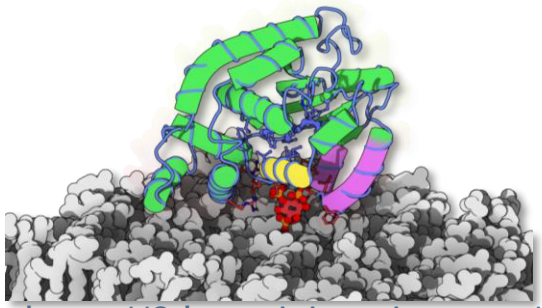
Pymol





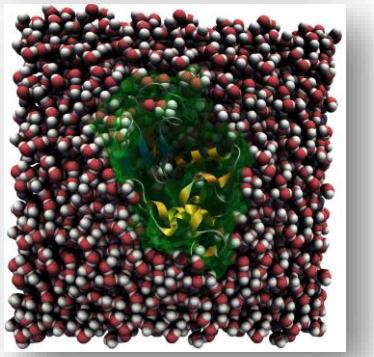
ChimeraX





https://3dproteinimaging.com/

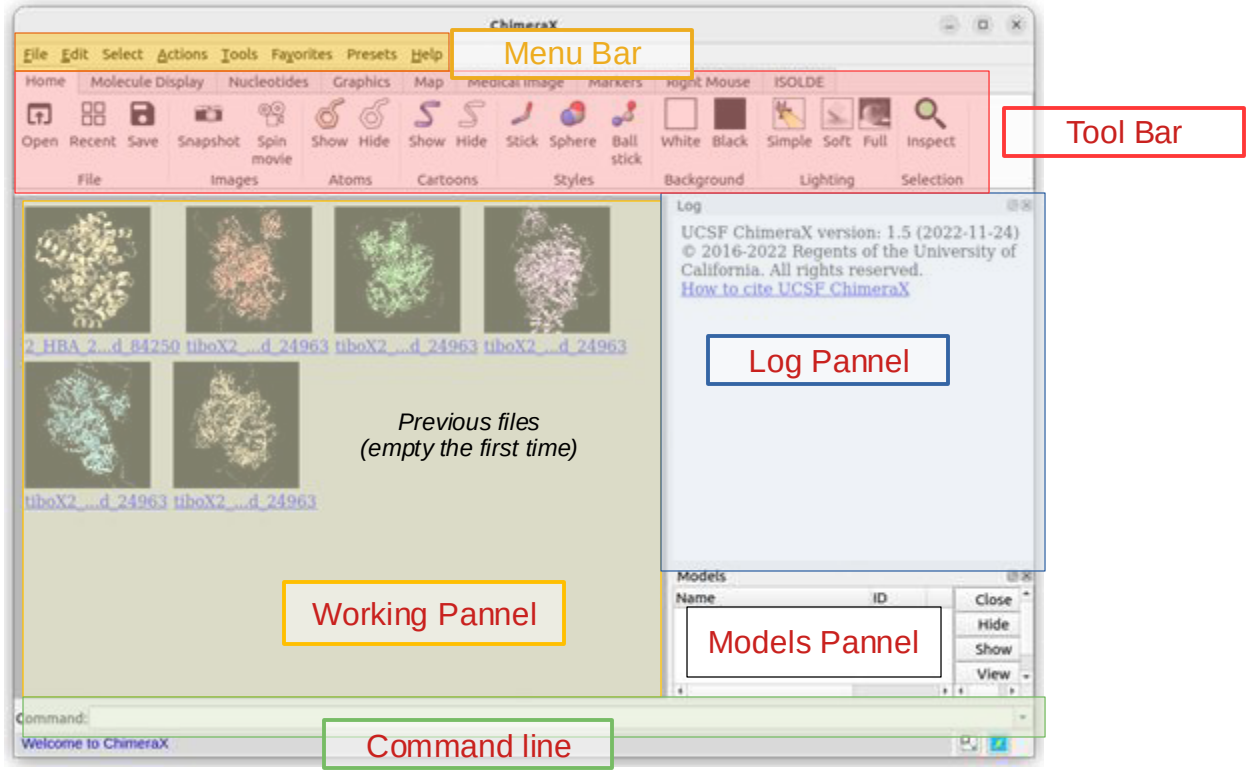
Unitymol



About ChimeraX

- ChimeraX is the new version of USCFChimera and is **free** (for academic) and available on every platform
- Optimized for CryoEM data and large systems as well
- No ray tracing engine, but can generate smooth images

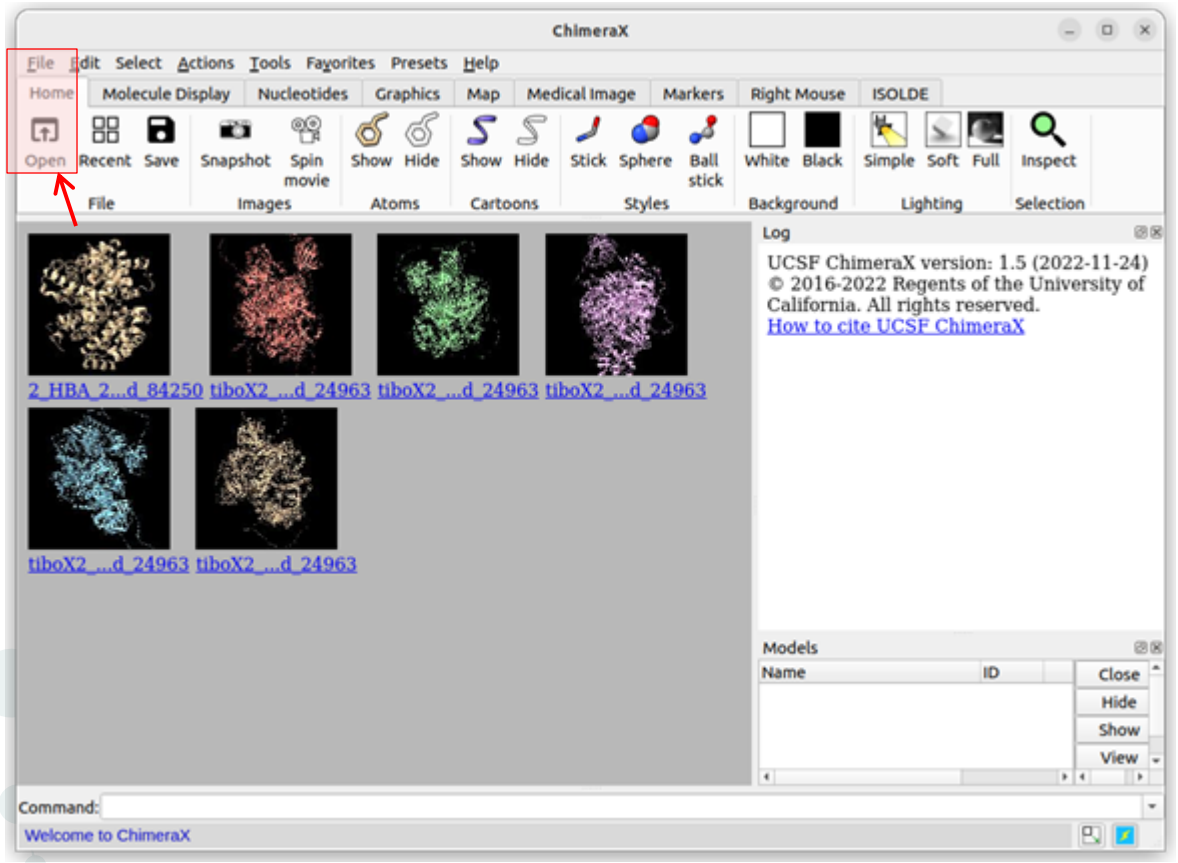
Main window when you open up ChimeraX



Step 1 – load a structure

Either directly from the databases: **File > Fetch by ID**

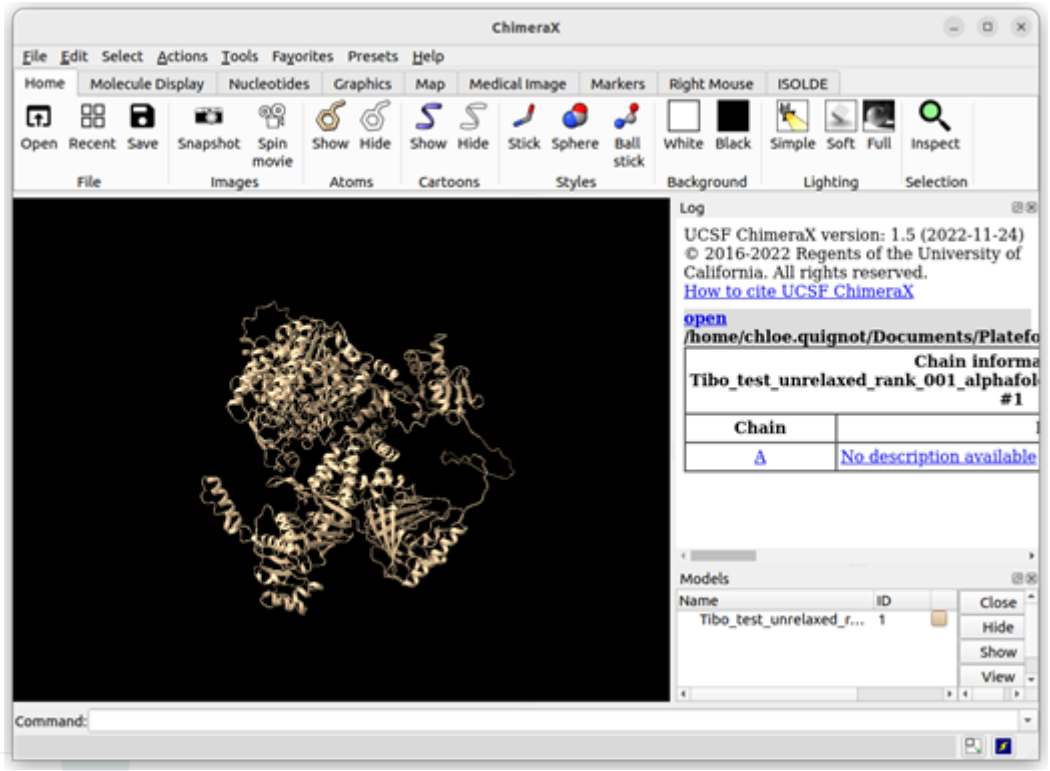
Or by loading a file on your computer: **File > Open** or **Home > Open**



Step 2 – move your protein around

The "Right Mouse" menu defines what action is associated to the right click of your mouse

Rotate and move



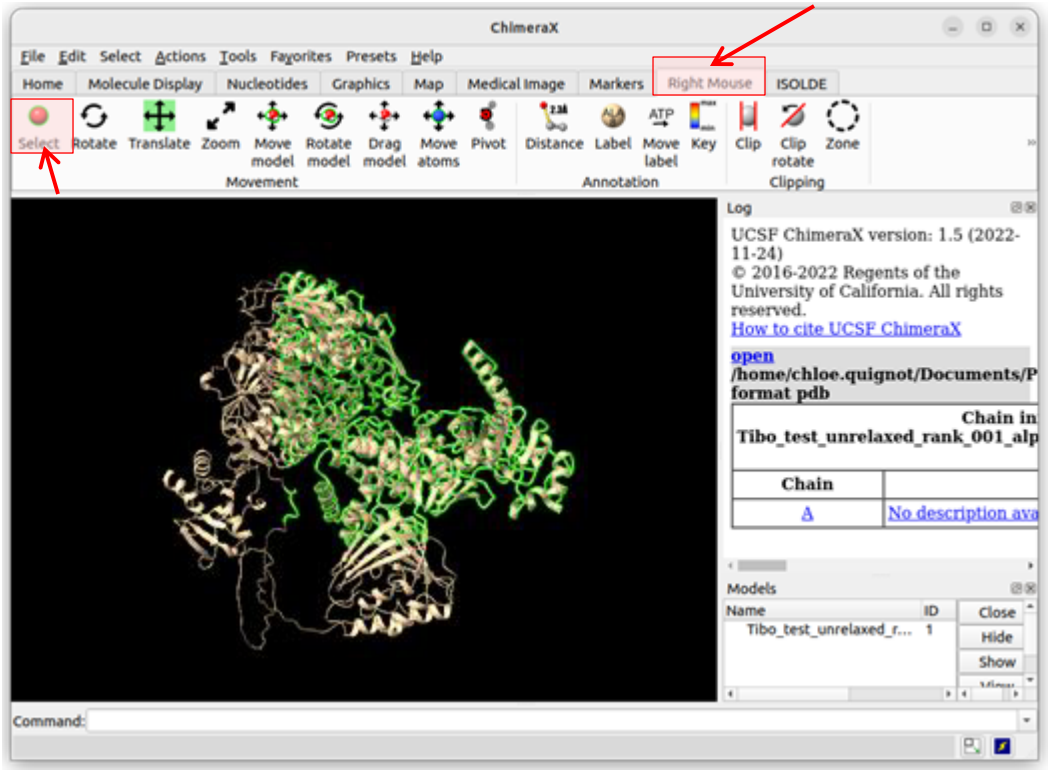

Left-click + hold **<u>on the protein</u>**:
=> move it about on vertical & horizontal axis


Left-click + hold **<u>on the background</u>**:
=> rotate it about clockwise or anticlockwise


Scroll:
=> zoom in & out


Right-click + hold or
Middle-click + hold:
=> translate the protein

Select regions




Ctrl + Left-click + hold & drag:
=> select a zone on your protein

Ctrl + Left-click outside of the protein to unselect

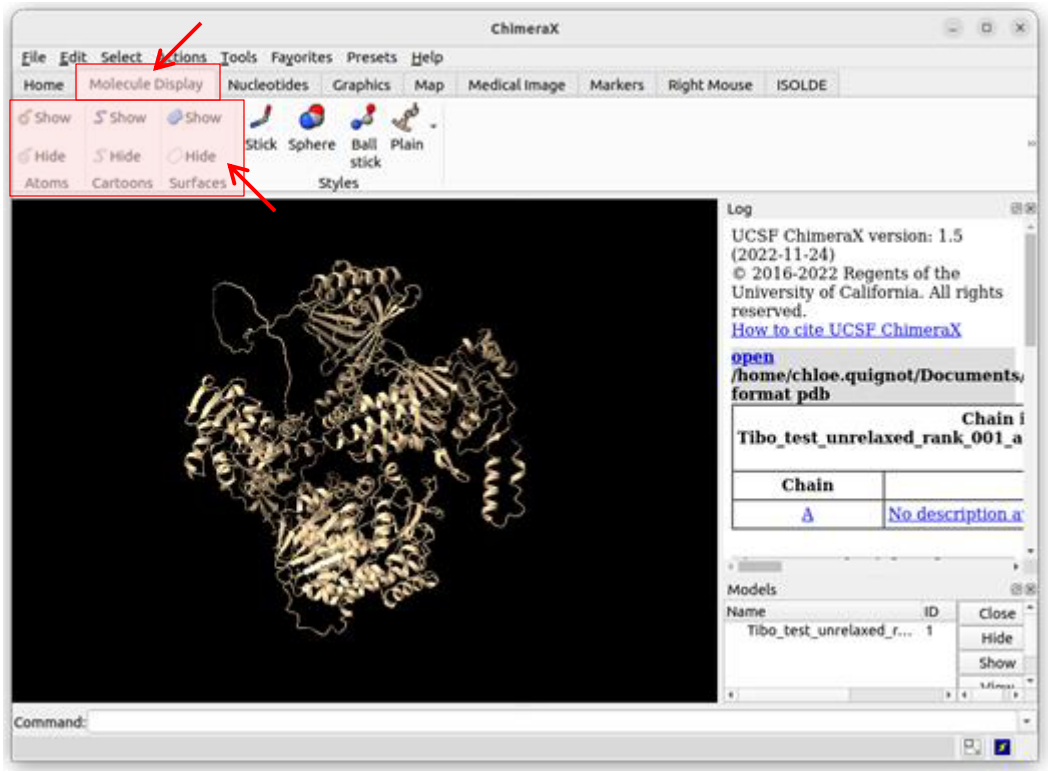
or

More functions in *Menu Bar > Right Mouse*

e.g. Menu Bar > Right Mouse > Select

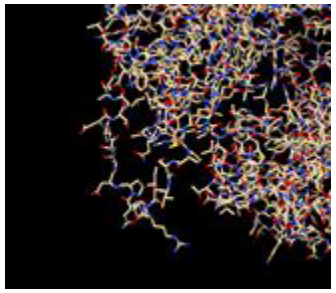

Right-click + hold & drag:
=> select a zone on your protein

Different outfits for different occasions

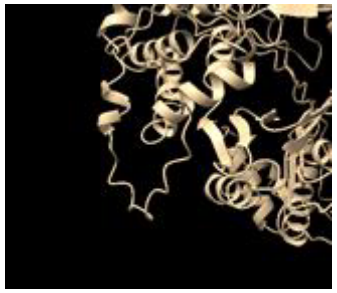


Show/hide
.“**Atoms**” view, several visualization options
-> in *Styles*, you can choose between:
“Sticks”, “Spheres” or “Ball stick” (for proteins) or “Plain” (for DNA)
.“**Cartoons**” view (default)
.“**Surfaces**” view

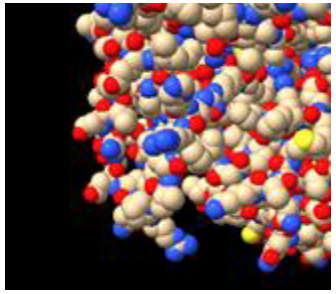
“**Atoms**” - Sticks



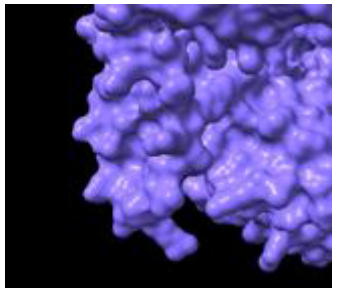
“**Cartoons**”



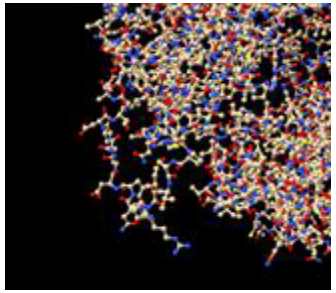
“**Atoms**” - Spheres



“**Surfaces**”

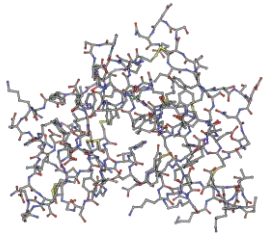


“**Atoms**” - Ball sticks



Visualisations & when to use them

Atoms + Ball sticks



Atoms = spheres, connecte by lines

Small systems, showing atomic connectivity & bonds

Big systems, simplistic figure

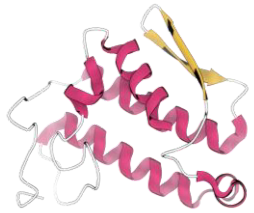
Atoms + Spheres

Atoms = spheres with van der waals radius

Small & big systems, simplistic figure

Bonding, interactions with small molecules

Cartoons

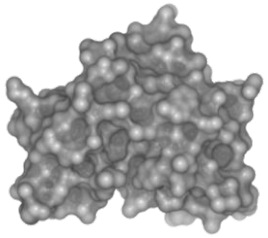


Standard 2^{ary} structure representation

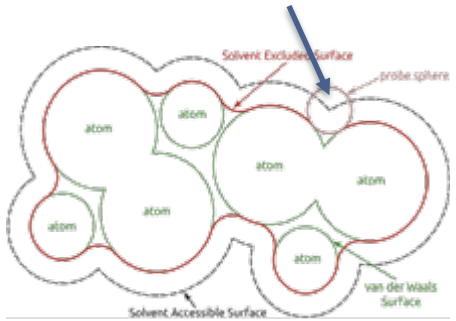
Big systems, see quaternary structure & protein shape, simplistic figure

Atomic interaction & pocket observation

Surface



Usual probe radius : 1.4Å (~ radius of a water molecule)



"real" shape of the protein

See quaternary structure & protein shape, pocket observation, simplistic figure

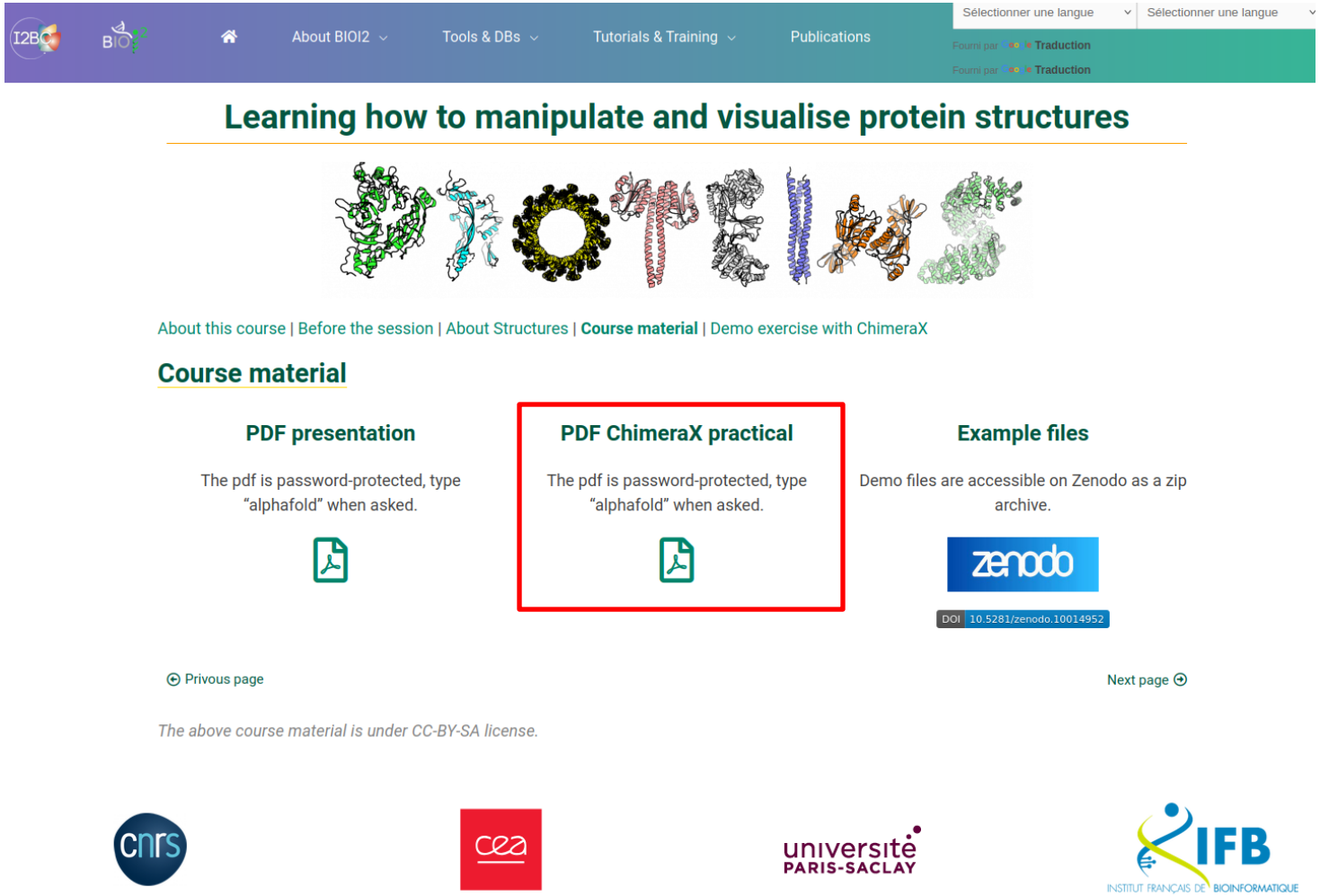
Atomic interaction, big systems (has to be calculated)

Sebastian Daberdaku

https://www.cgl.ucsf.edu/chimera/docs/UsersGuide/representation.html

Your turn!

https://bioi2.i2bc.paris-saclay.fr/training/visu-struct/course-material/



/!\ course material is password protected, type "alphafold" to unlock

Saving images

It's an art to itself!

Keys for a good vizualisation

What do I want to show ?

1

- Protein-Protein interaction ?
- Overall protein structure ?
- Protein-ligand interaction ?

What representation to use ?

2

- Ball and sticks, Licorice ?
- Surface
- Cartoon
- ....

Highly connected!!

What Software to use ?

3

- Pymol
- ChimeraX
- 3DProtein Imaging
- VMD
- Molstar
- Blender
- .....

Image format

JPEG

- Lossy compression
- No transparency
- Mostly use for photos
- Small file size

PNG

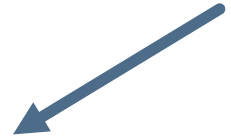
- Lossless compression
- Transparency (alpha Chanel)
- high quality illustration
- Medium file size

TIFF

- Can be compressed (lossless)
- Transparency + Multiple Layer
- High quality illustration
- Large File size

Image format

JPEG	PNG	TIFF
Lossy compression	Lossless compression	Can be compressed (lossless)
No transparency	Transparency (alpha Chanel)	Transparency + Multiple Layer
Mostly use for photos	high quality illustration	High quality illustration
Small file size	Medium file size	Large File size

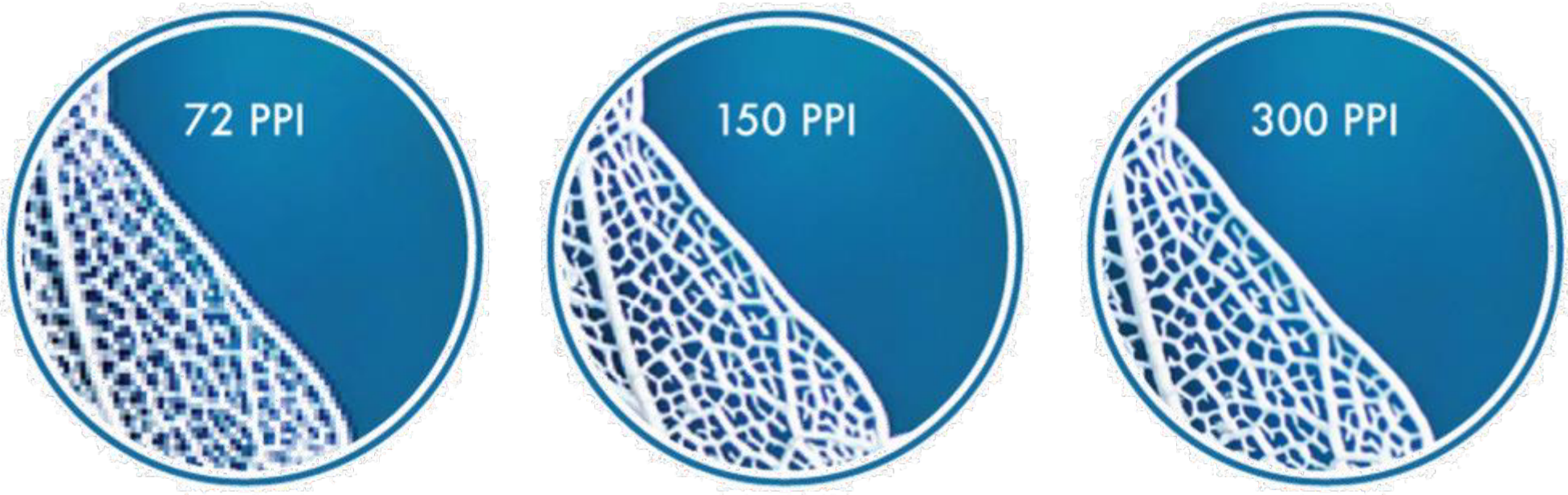


Before publication you should compress your TIFF with LZW Compression

- Online with Pace (which also check image problem for publication) https://pacev2.apexcovantage.com
- Command line : convert input.tiff -compress LZW output.tiff

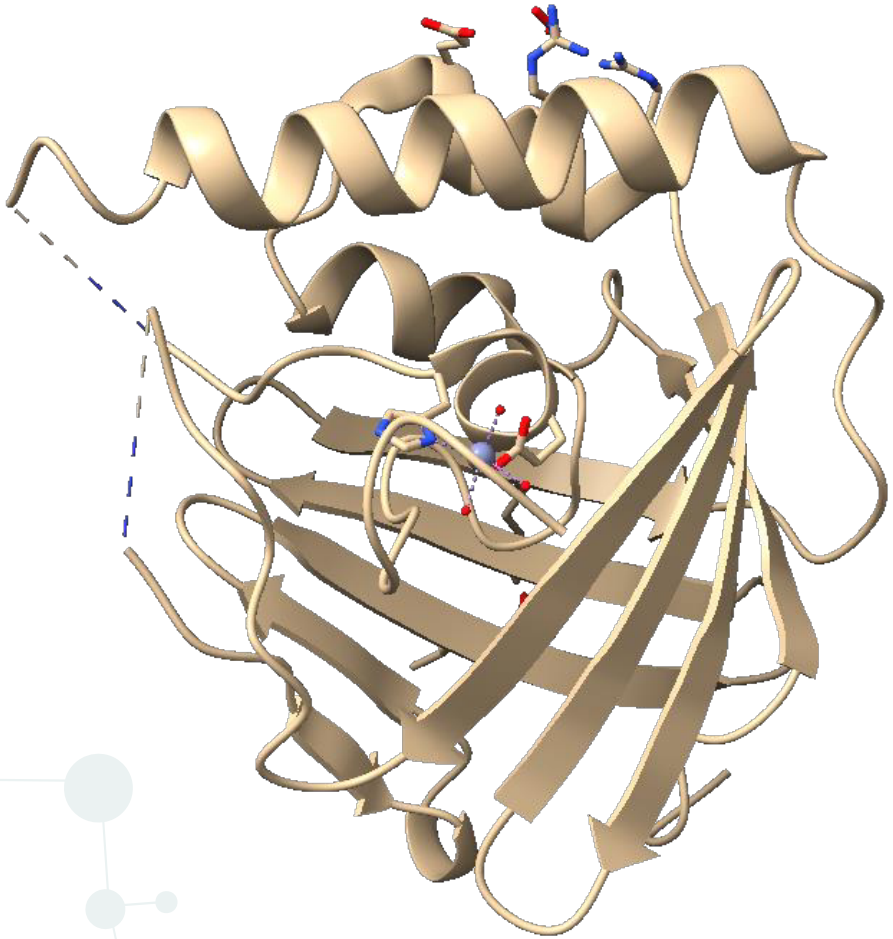
DPI / PPI

DPI (Dots Per Inch) or **PPI** (Pixels Per Inch) measure the **image resolution**
→ number of individual dots (or pixel) that can be display within a square inch.

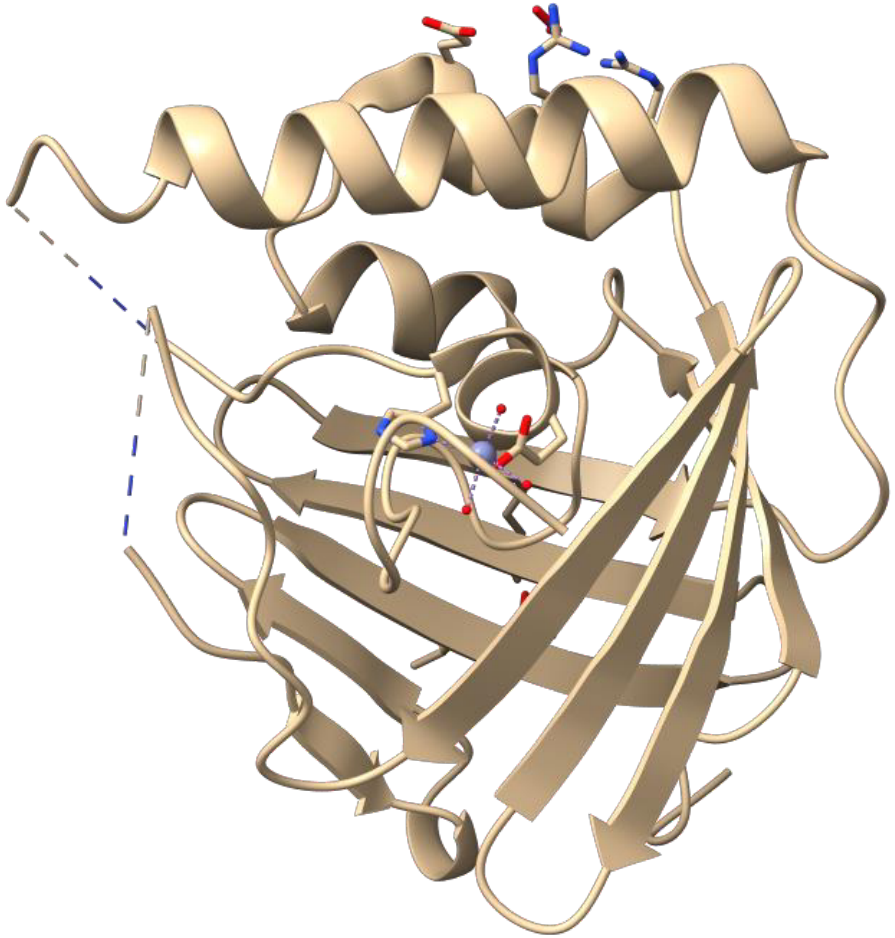


ChimeraX

SuperSampling

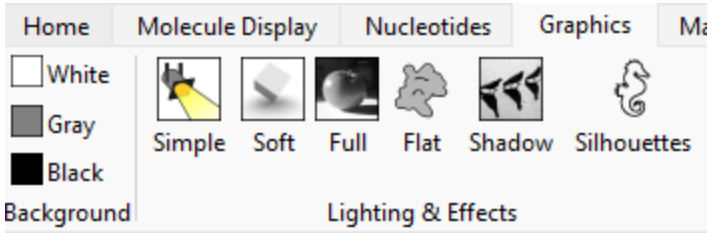


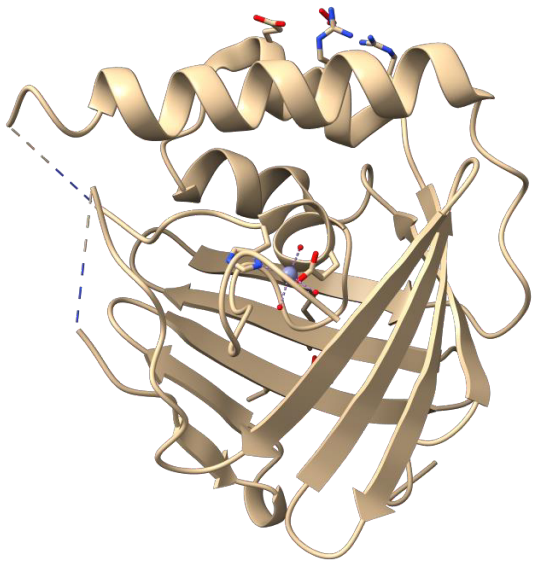
No Supersample (1)



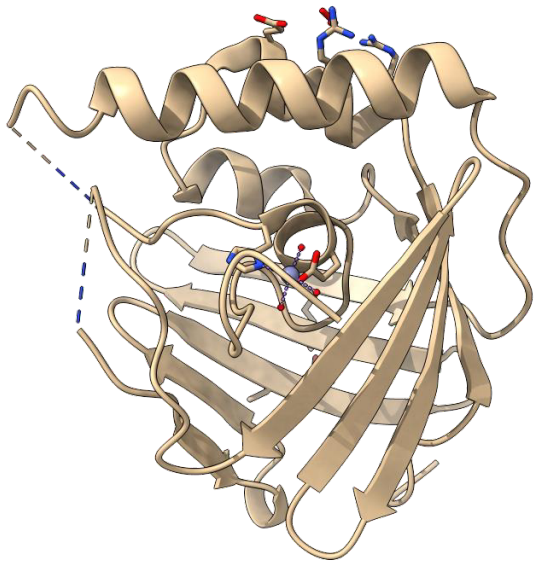
Supersample 8x

ChimeraX – Styles

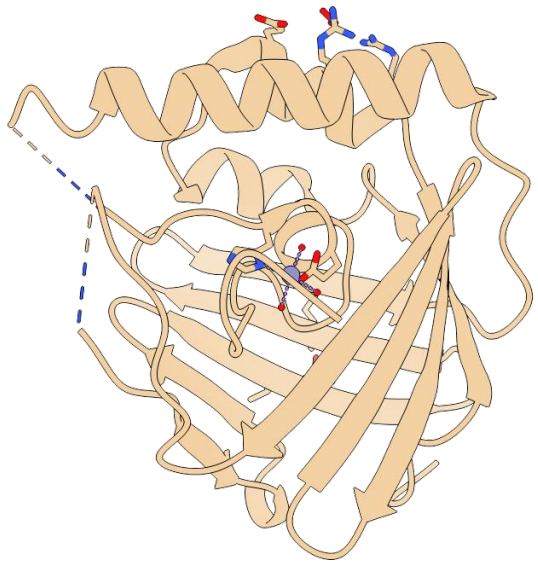




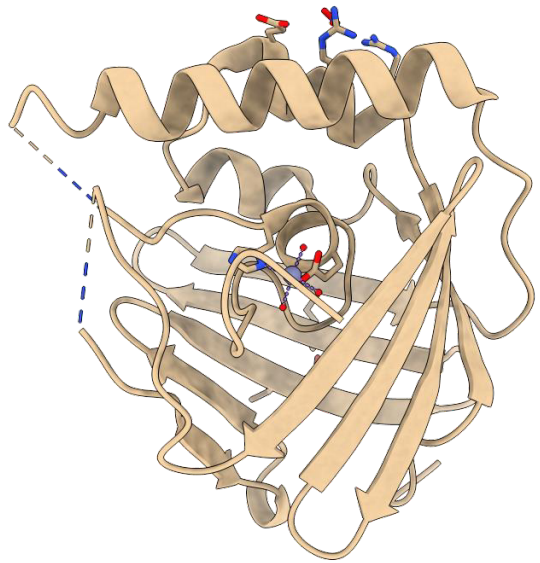
Simple



Full



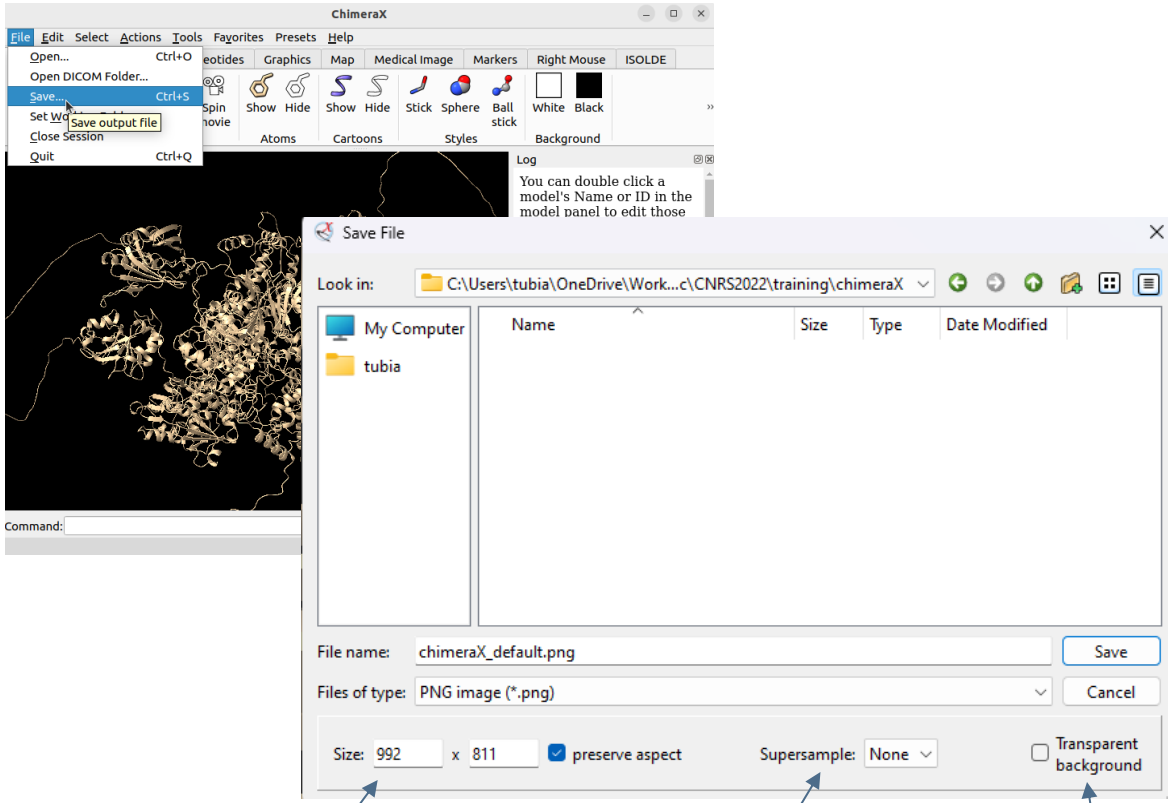
Flat



soft

Saving images in ChimeraX

Graphically



Resolution

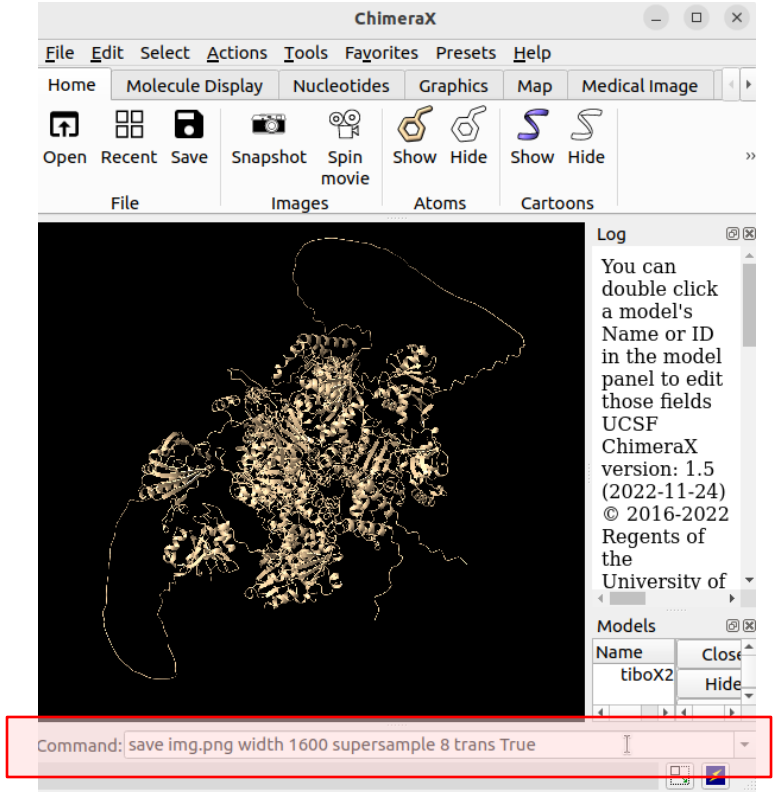
SuperSampling (better Quality)

Transparent Background

Or through command line:

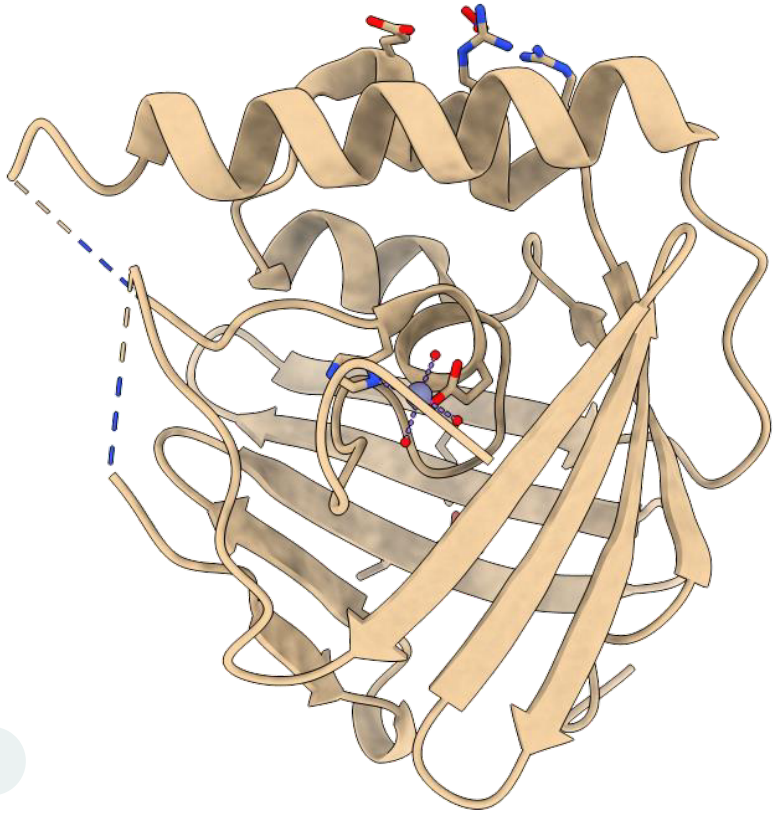
Type the following command in the command box:

```
save img.png width 1600 supersample 8 trans True
```

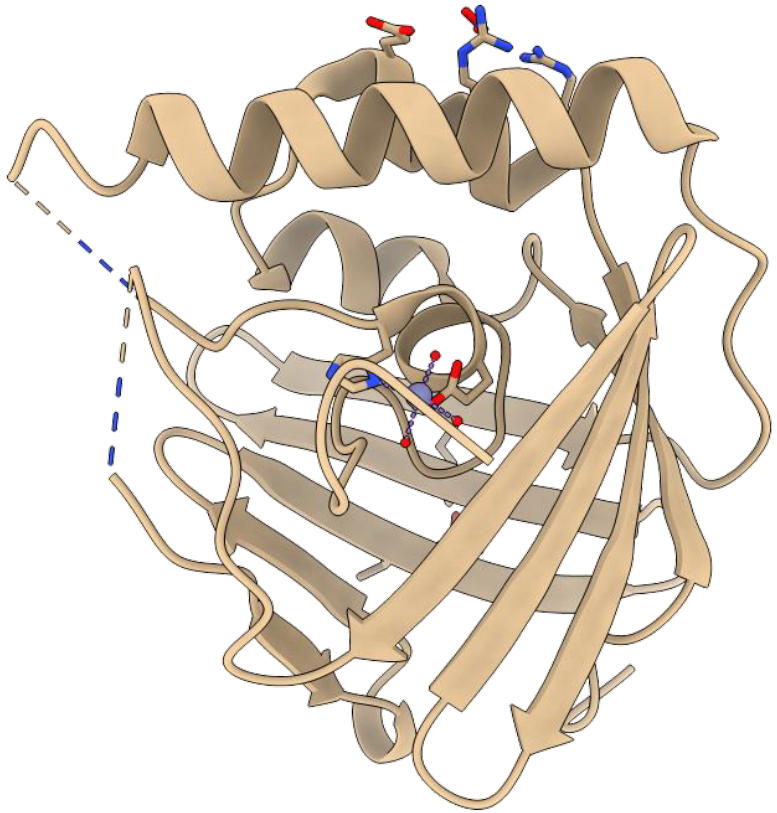


@Chloé Quignot, Biol2 platform, I2BC

ChimeraX – Styles



soft



soft + type in the command line:
`lighting multishadow 512`

FYI

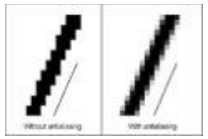
Render Style – Pymol

Example with *Pymol*

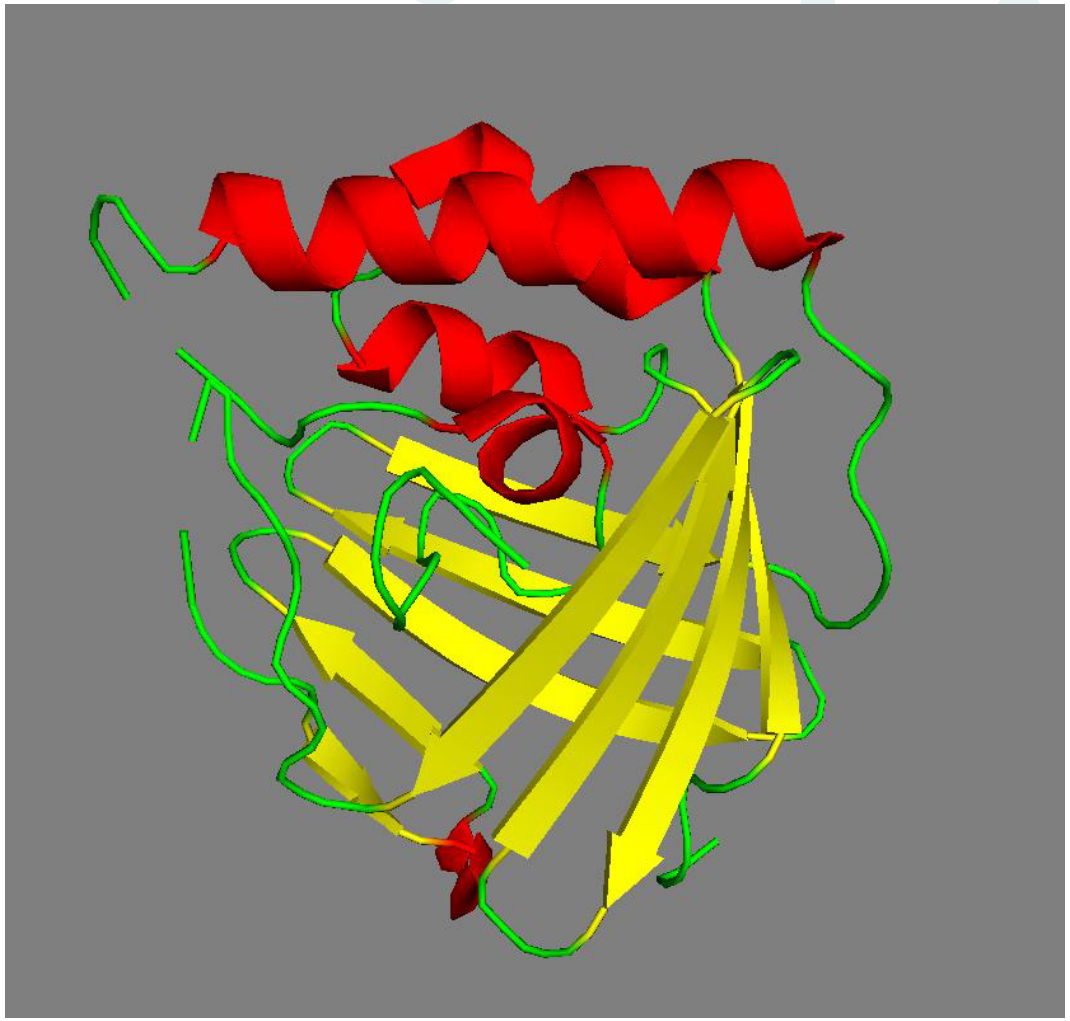
Screenshot:

+ Fast and dirty

– Bad Quality
– Bad resolution
– No background transparency
– No shadows
– No Antialiasing



Do not do that!



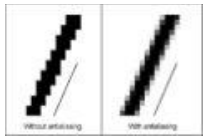
Render Style – Pymol

Example with *Pymol*

Screenshot:

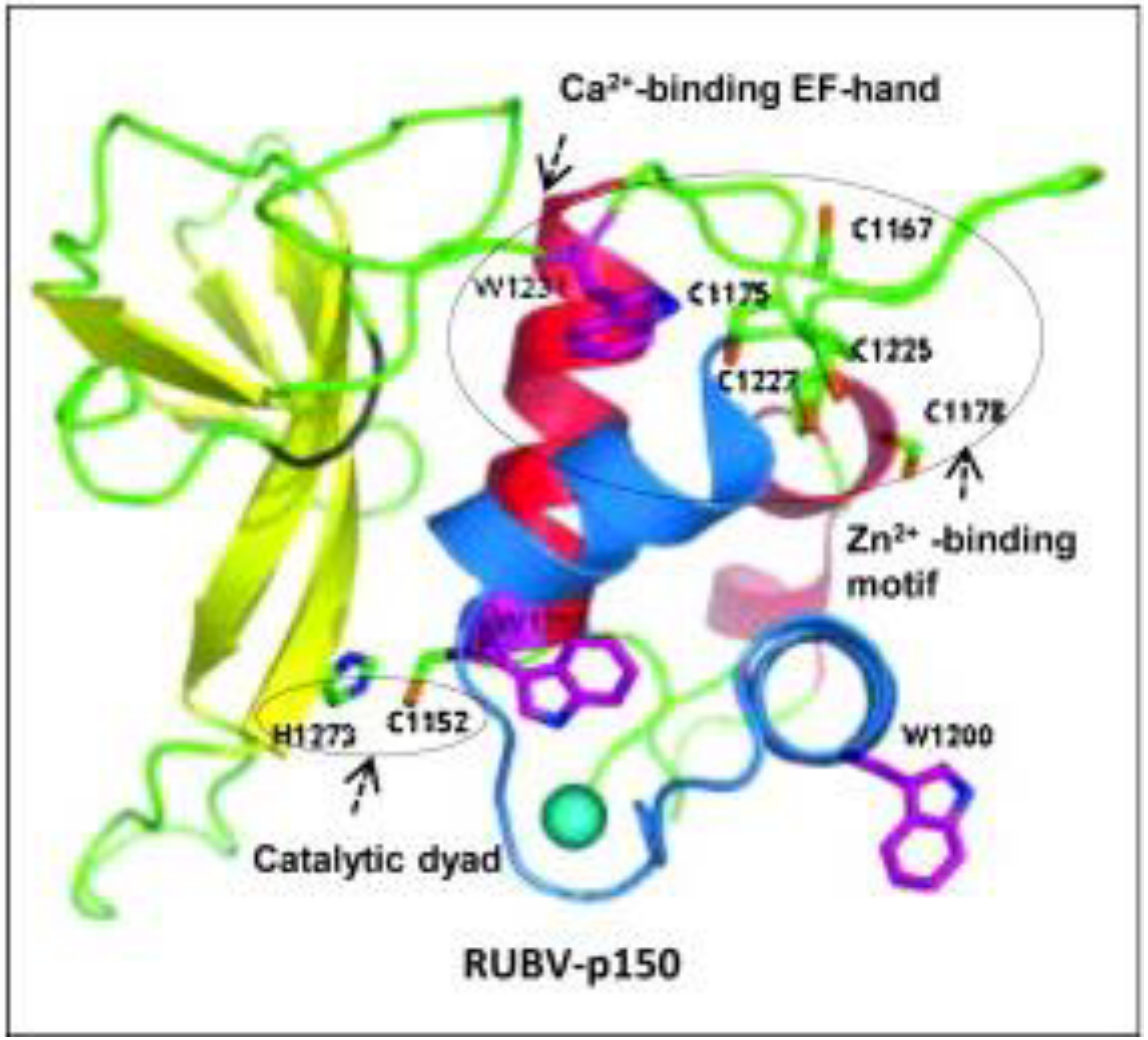
+ Fast and dirty

– Bad Quality
– Bad resolution
– No background transparency
– No shadows
– No Antialiasing



Do not do that!

+ Always think about the final resolution of your figure in the paper!



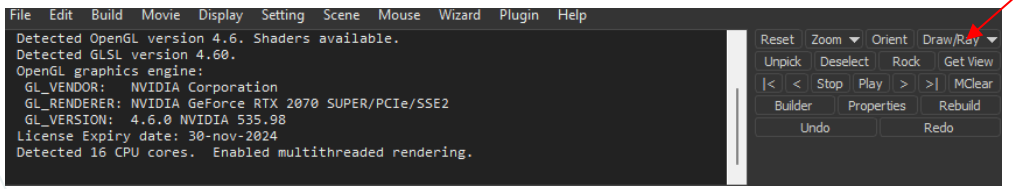
Render Style – Pymol

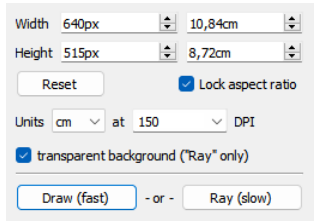
Example with *Pymol*

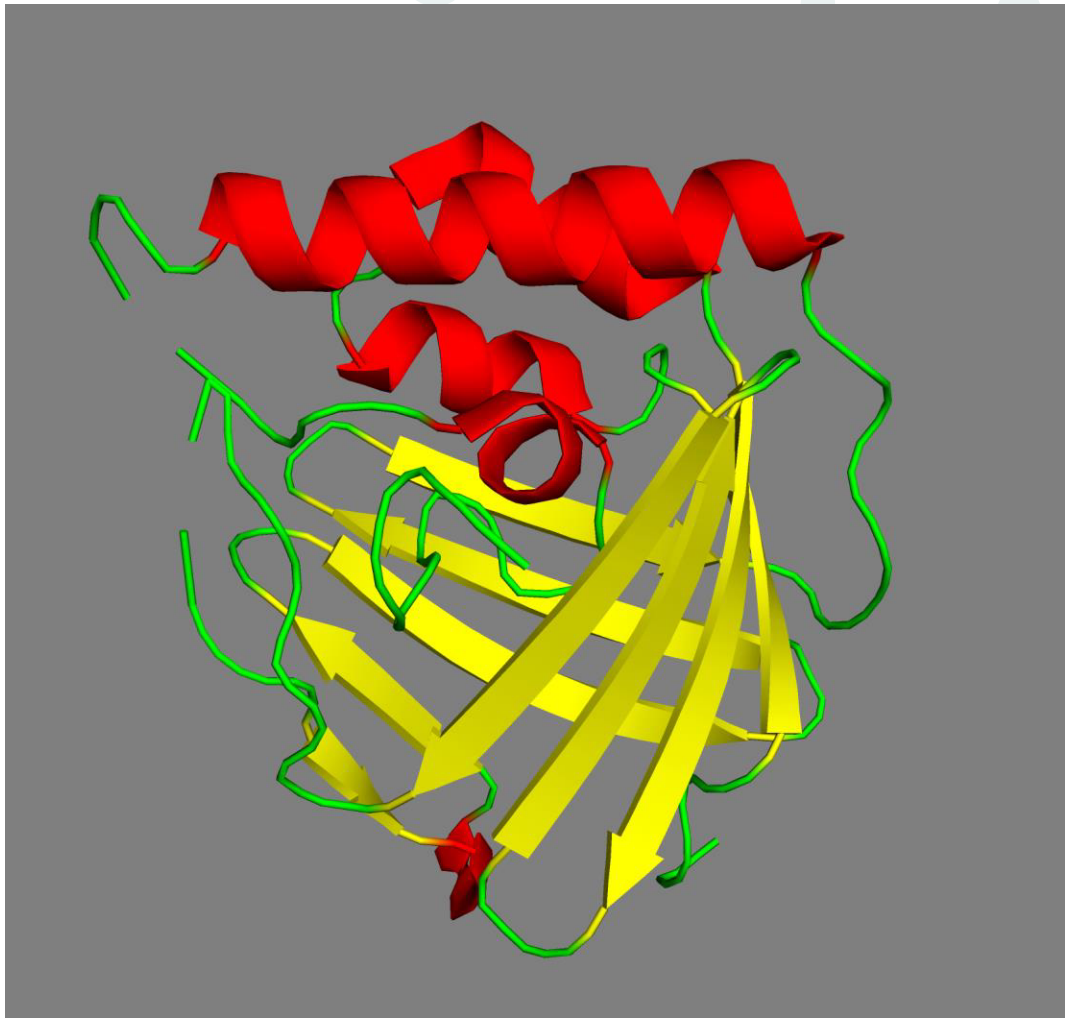
Draw (draw) :

+ Fast
+ Better Quality
+ Resolution can be set
+ Antialiasing

– No background transparency
– No shadows







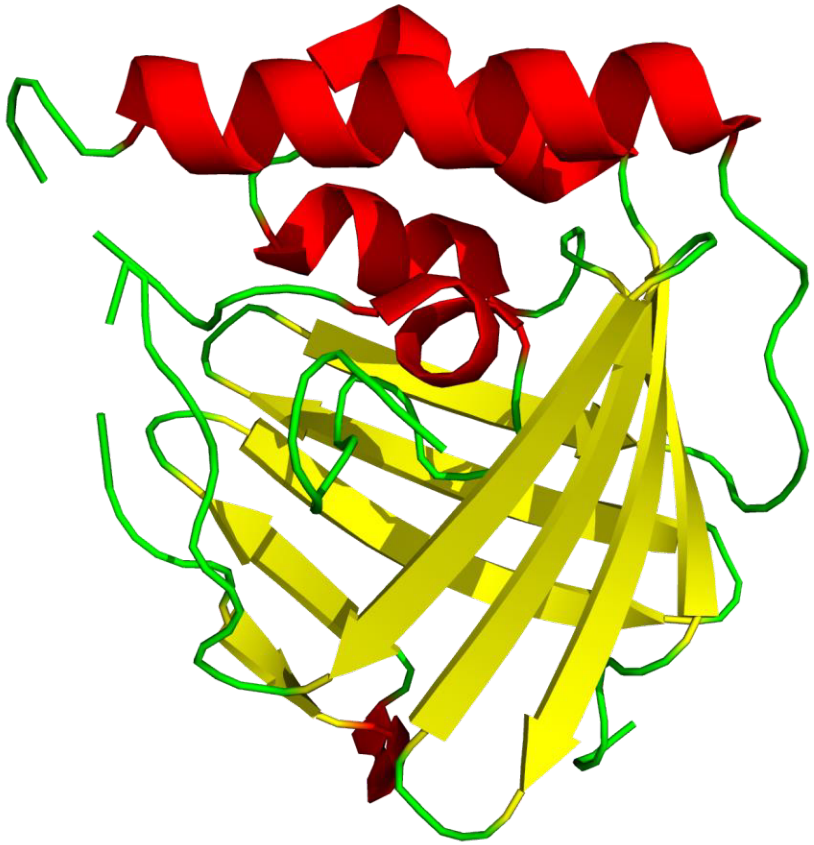
Render Style – Pymol

Example with *Pymol*

Ray Tracing (ray) :

+ Fast
+ Better Quality
+ Resolution can be set
+ Antialiasing
+ Background Transparency
+ Shadows

– Slower



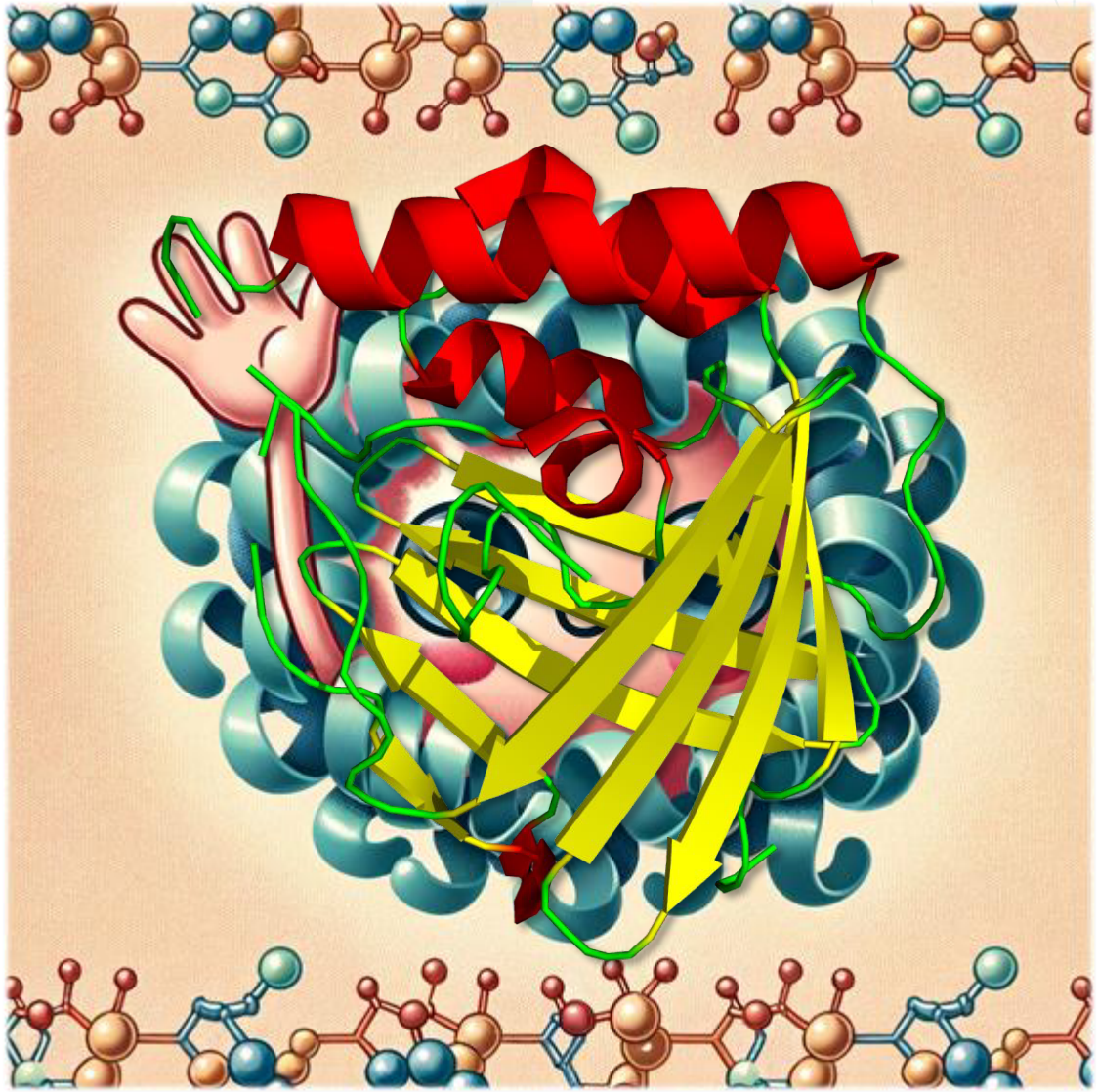
Render Style – Pymol

Example with *Pymol*

Ray Tracing (ray) :

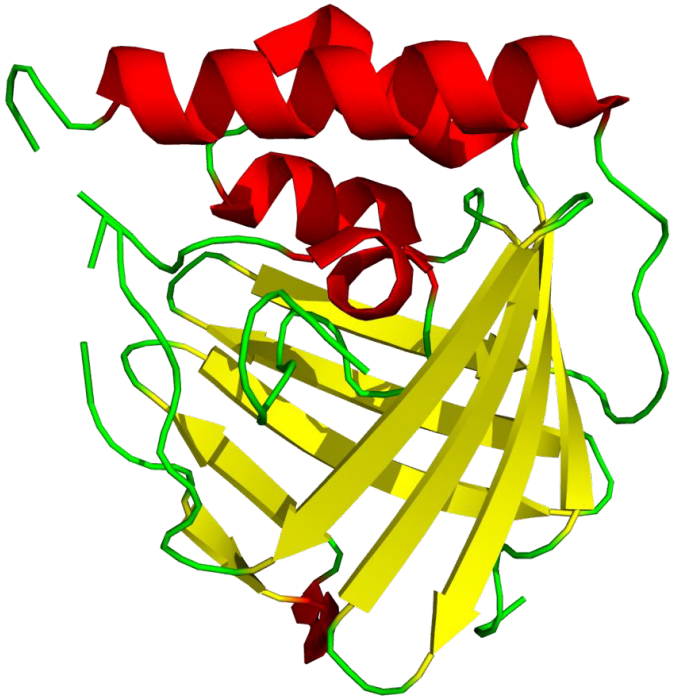
+ Fast
+ Better Quality
+ Resolution can be set
+ Antialiasing
+ Background Transparency
+ Shadows

– Slower

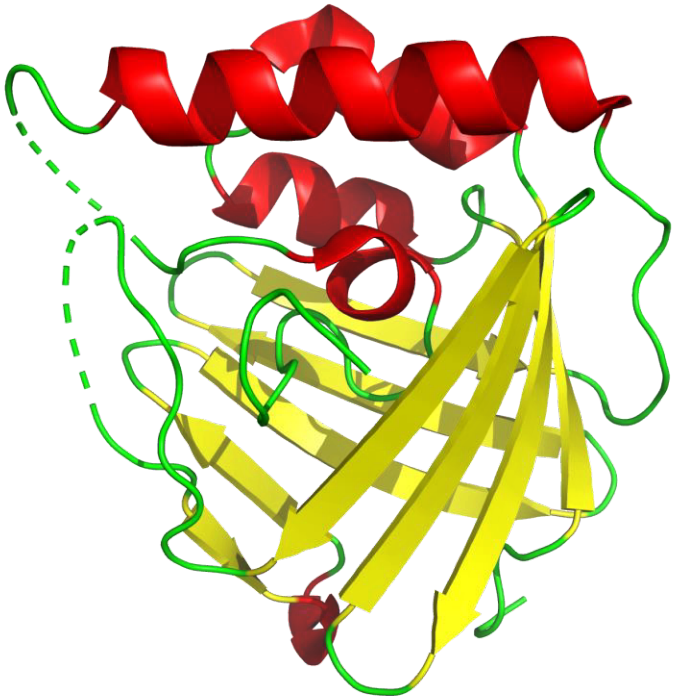


Render Style – Pymol

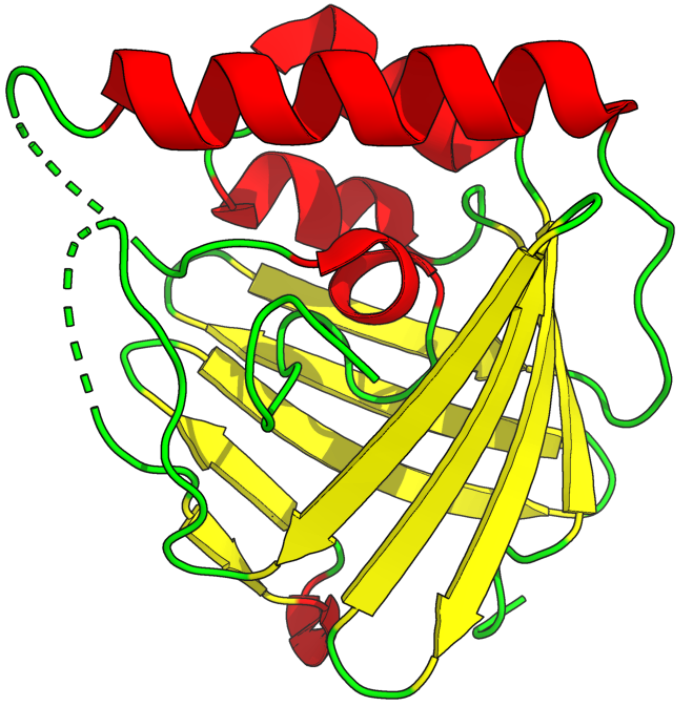
Learn to Play with Parameters!



Default



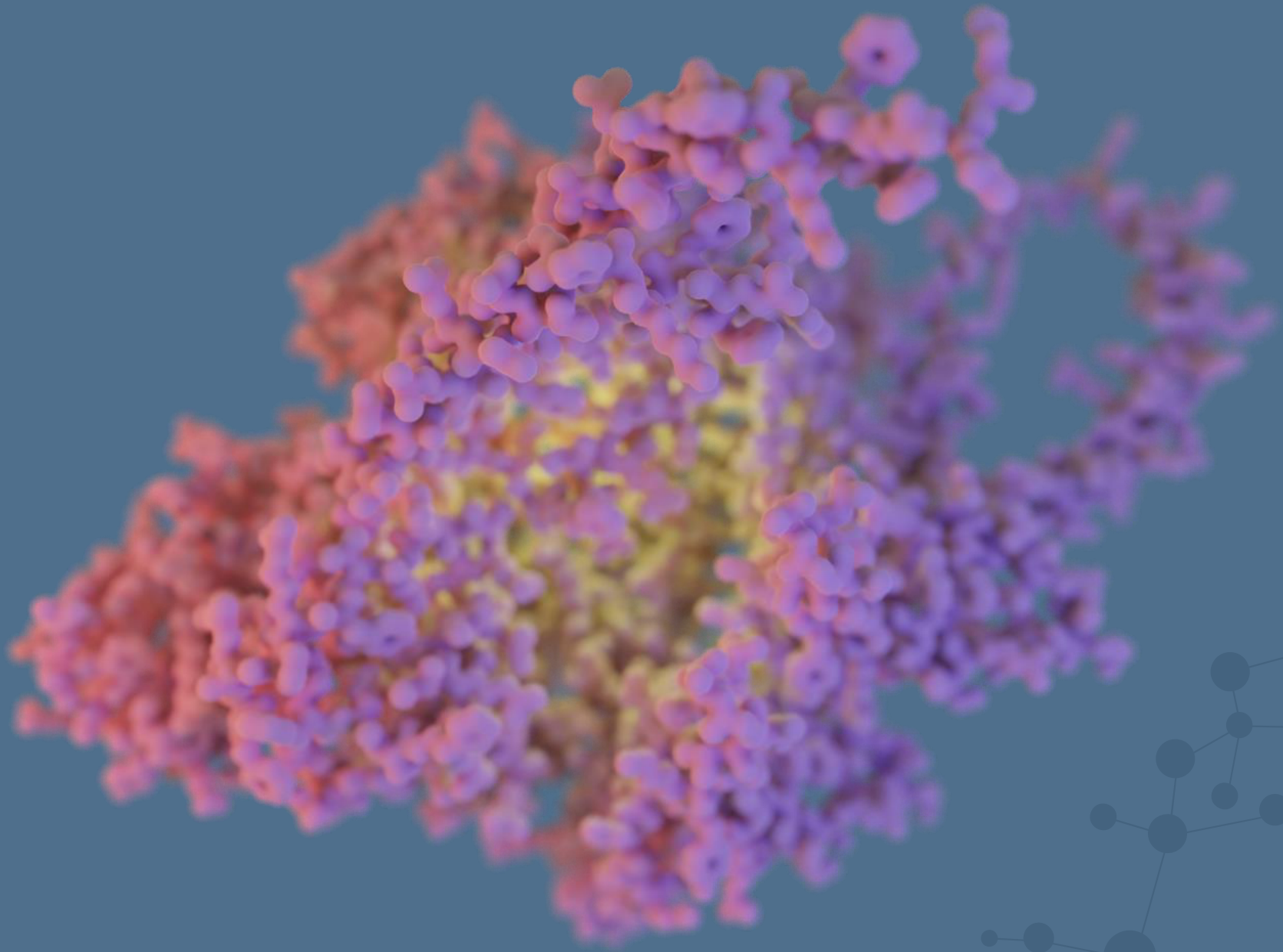
```
util.performance(0)
```



```
util.performance(0)
set ambient, 0.6
set specular, 0
set ray_trace_mode, 1
set ray_trace_color, black
set ray_trace_gain, 0.1
set antialias, 2
```

PART II

1. What is AlphaFold?



Protein modeling : sequence to structure

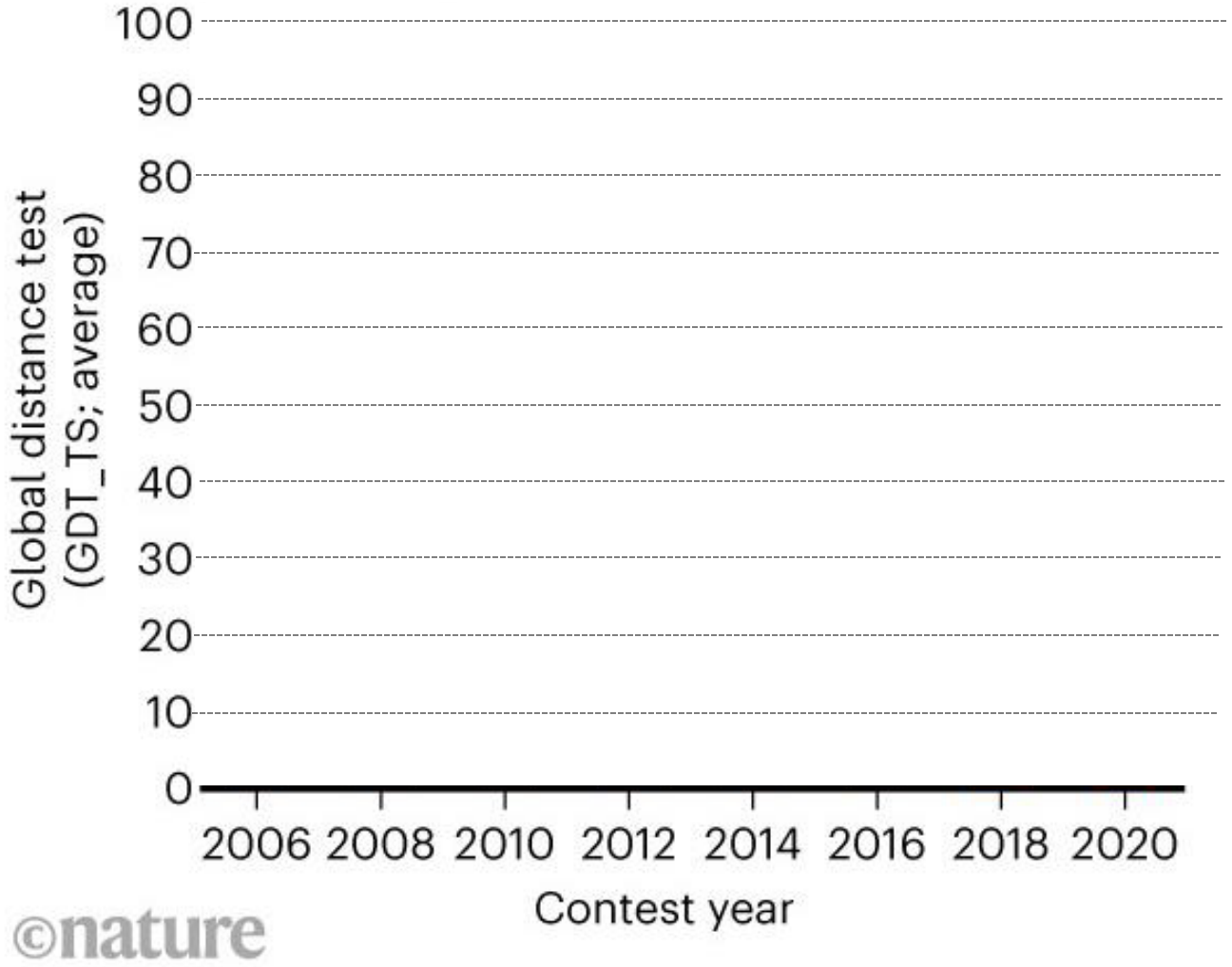
CASP (**C**ritical **A**ssessment of protein **S**tructure **P**rediction) :

Contest for ***de novo*** **protein structure prediction**

*"aim at **establishing** the **current state of the art** in **protein structure prediction**, identifying what progress has been made, and **highlighting where future effort** may be most productively focused." https://predictioncenter.org*

Olympic games for protein folding

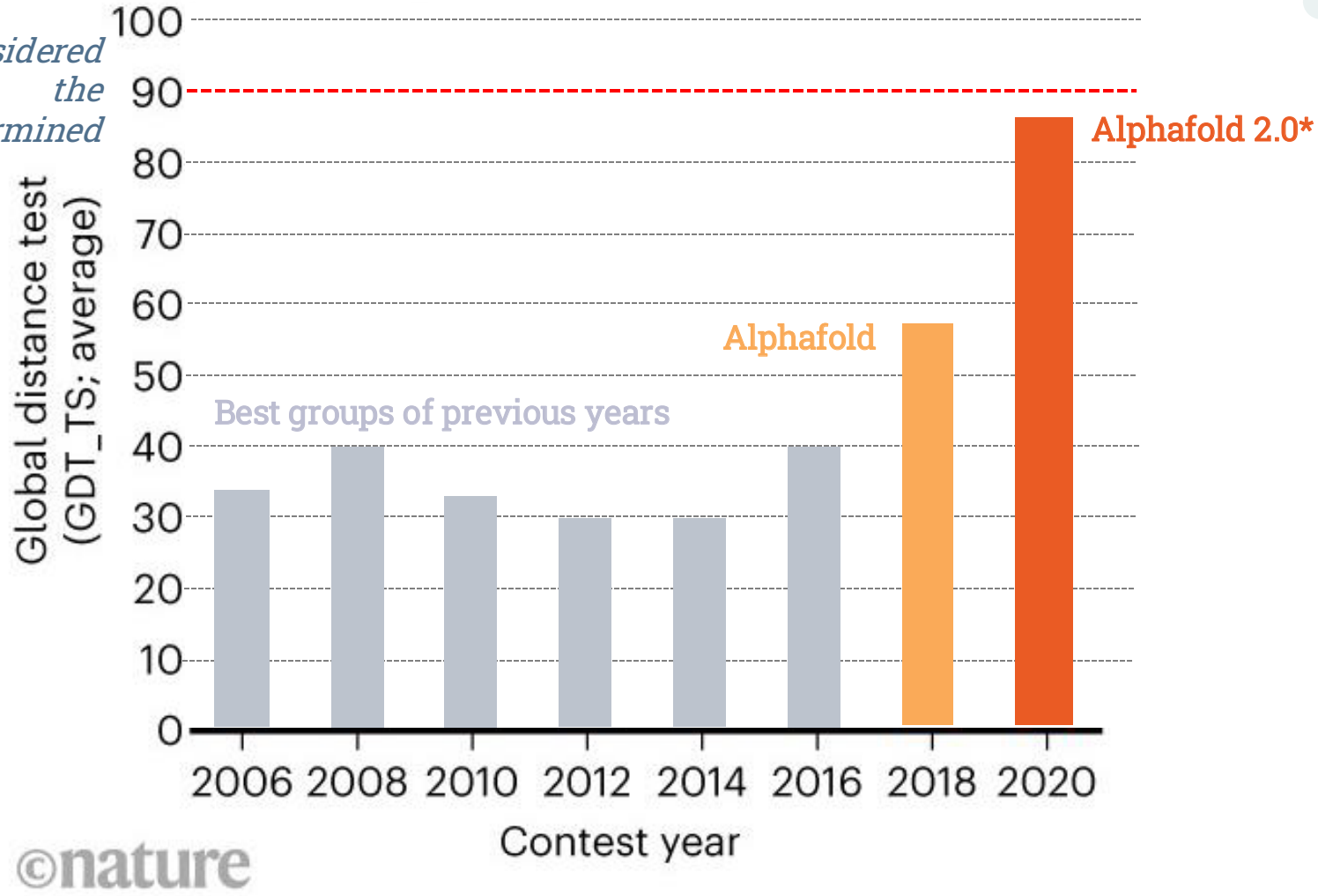
AlphaFold 2.0



https://www.nature.com/articles/d41586-020-03348-4

AlphaFold 2.0

*A score **above 90** is considered roughly **equivalent** to the **experimentally** determined **structure***



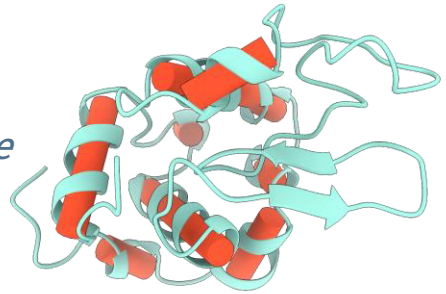
https://www.nature.com/articles/d41586-020-03348-4

* It is an average, AF2 got a GDT of 89.5, 91.4 and 96.5 on specific proteins.

Protein modeling : sequence to structure

Homology modeling

Template structure (reference)



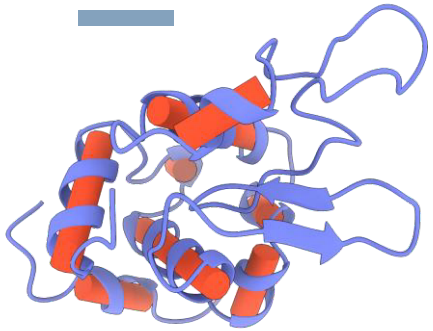
+

Target sequence

MDIGINSDPHPPHHHDHHGHGSGW[…]DHHHGDHHH

=

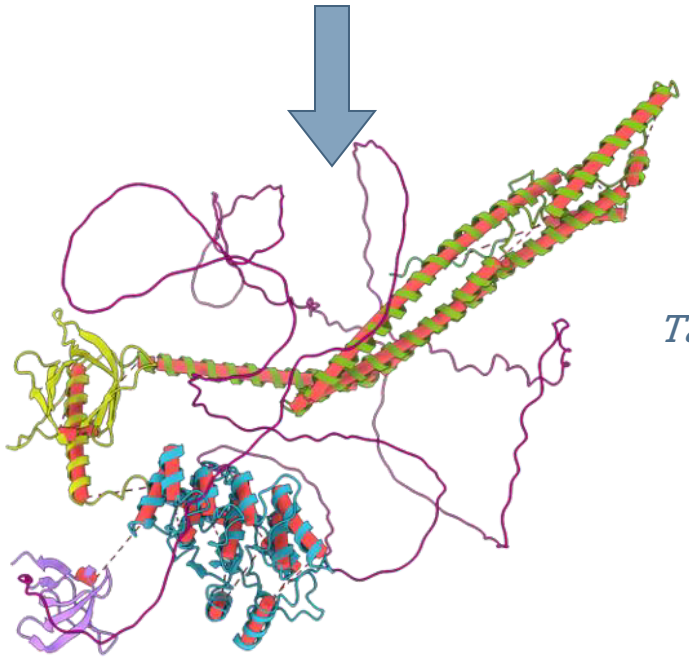
Target Model



De novo modeling

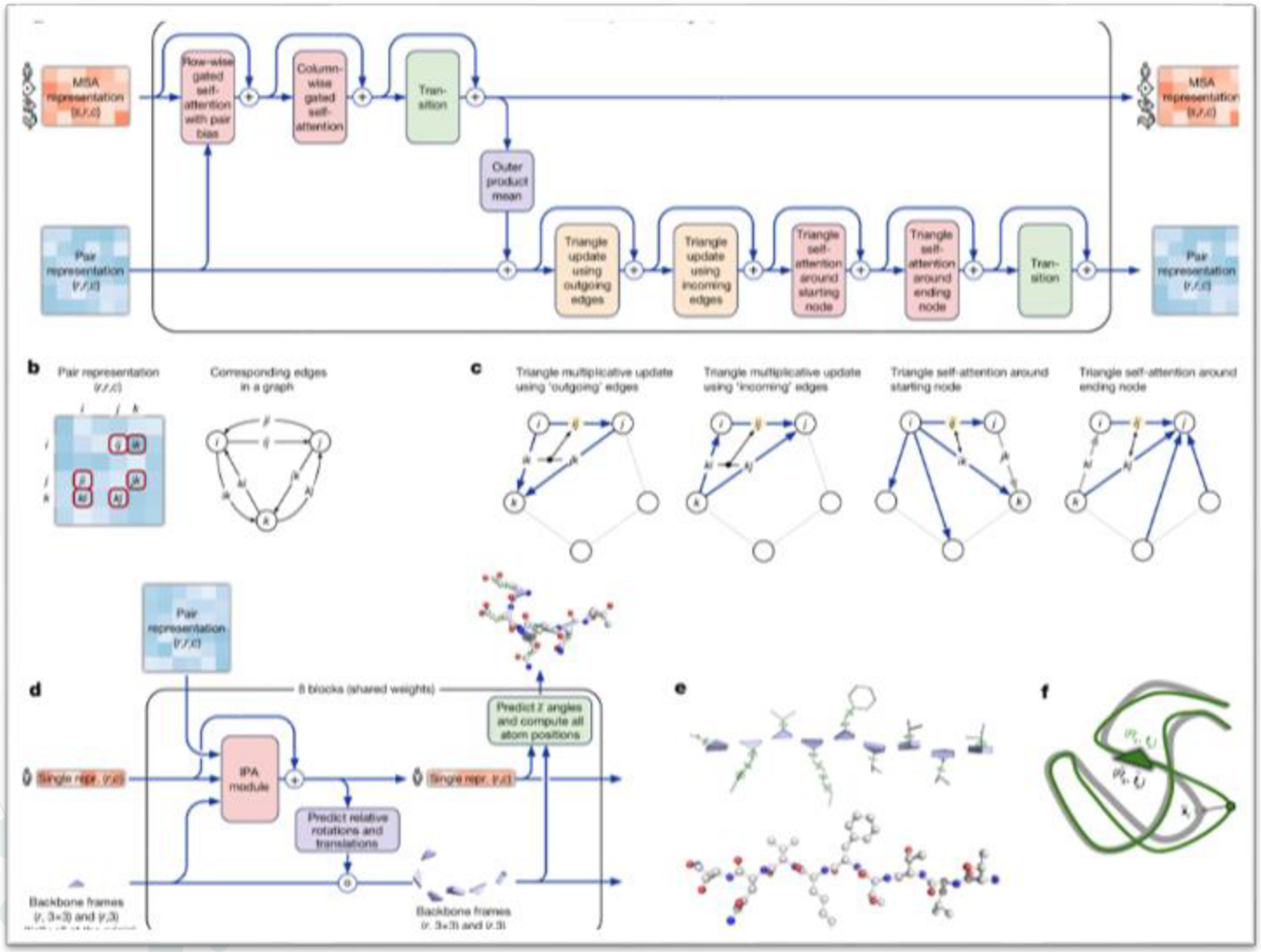
Target sequence

MRSSASRLSSFSSRDSLWNRMPDQI[…]SFVHILSD



Target Model

AlphaFold2 – How does it work?



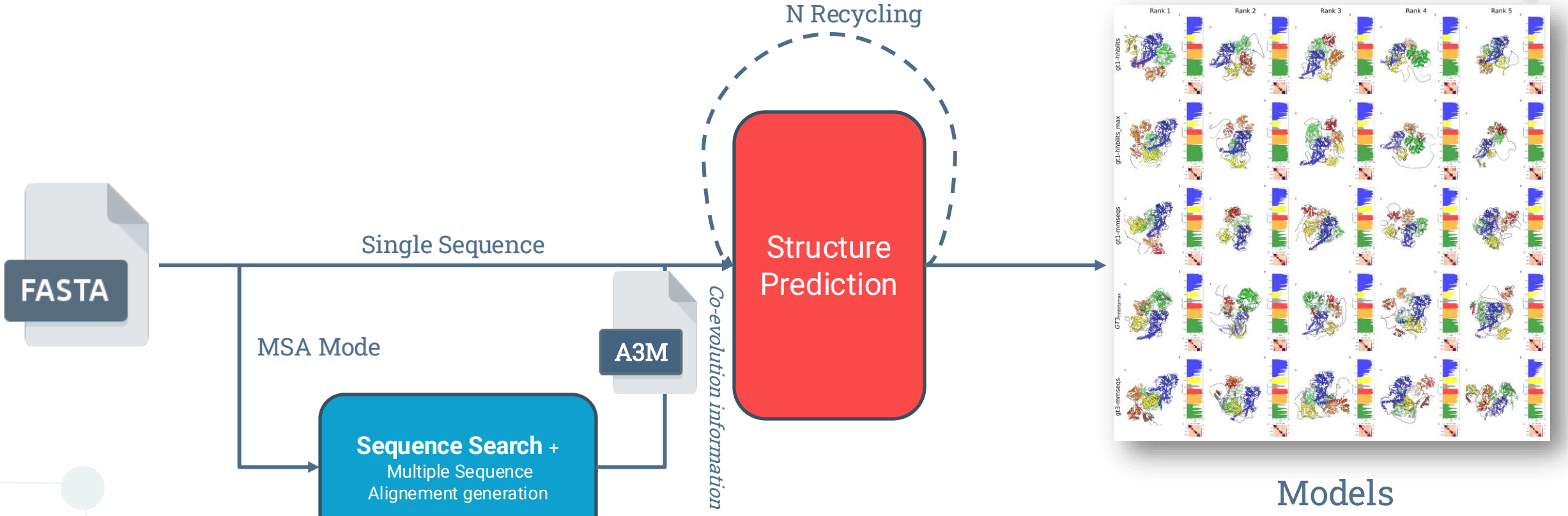
The full process may seem quite complex at first sight….

Jumper, J. *et al.* (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

AlphaFold2 – How does it work?

In practice….

All you need is the input sequence

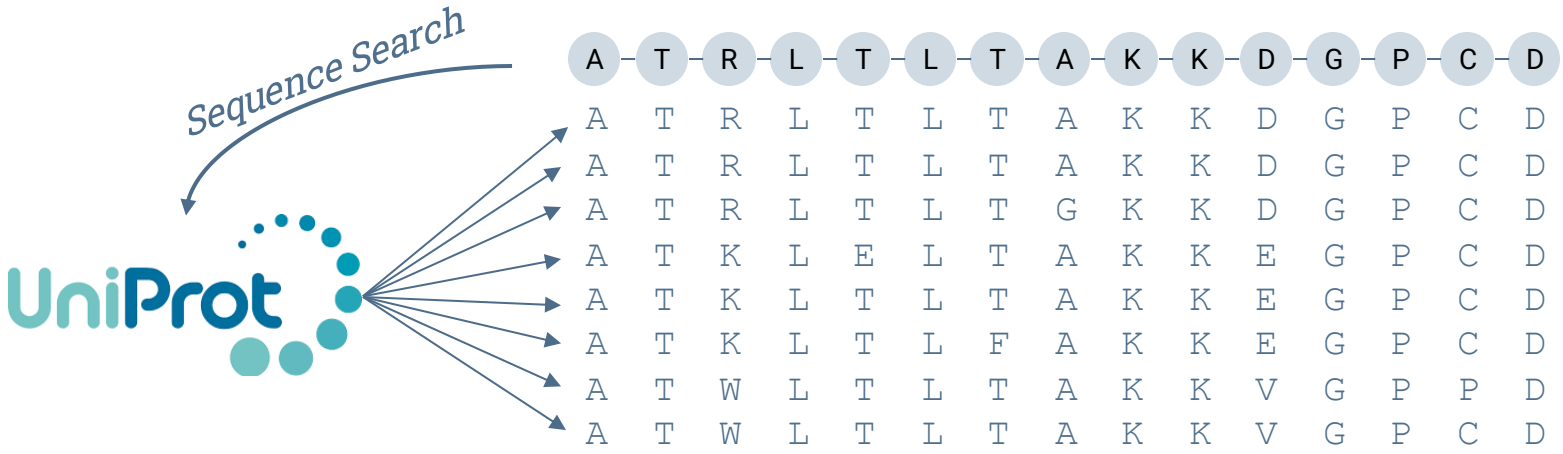


- Uniref 30
- Environmental database

AlphaFold2 – How does it work?

Let's make it simple…

First, AF is based on the principle of **co-evolution**

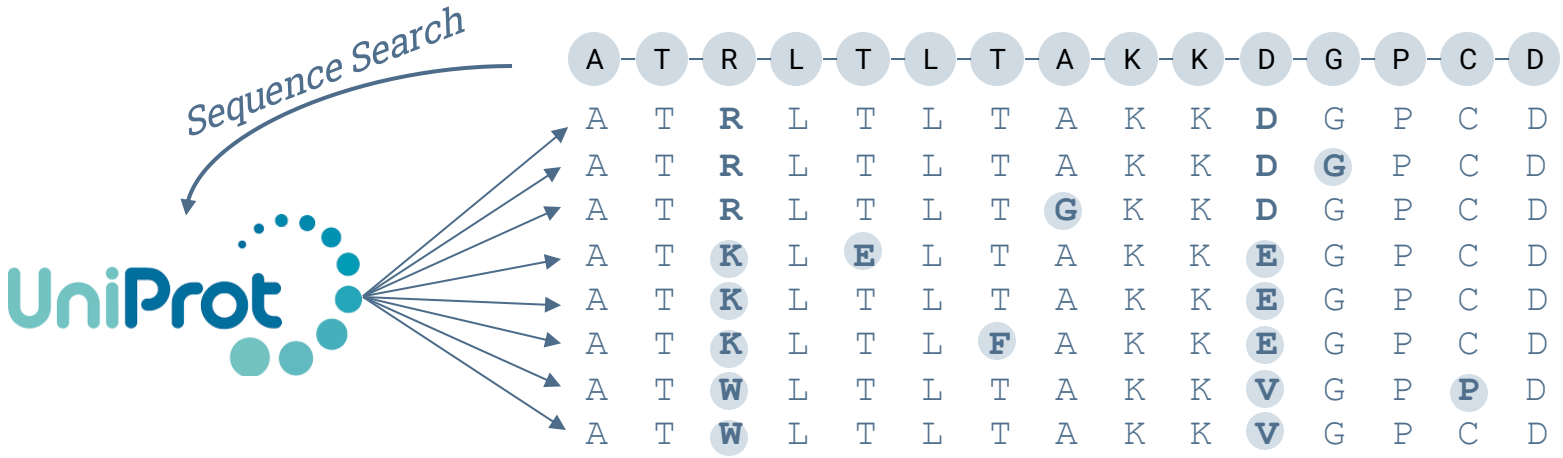


Inspired by https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-i

AlphaFold2 – How does it work?

Let's make it simple…

First, AF2 is based on the principle of **co-evolution**

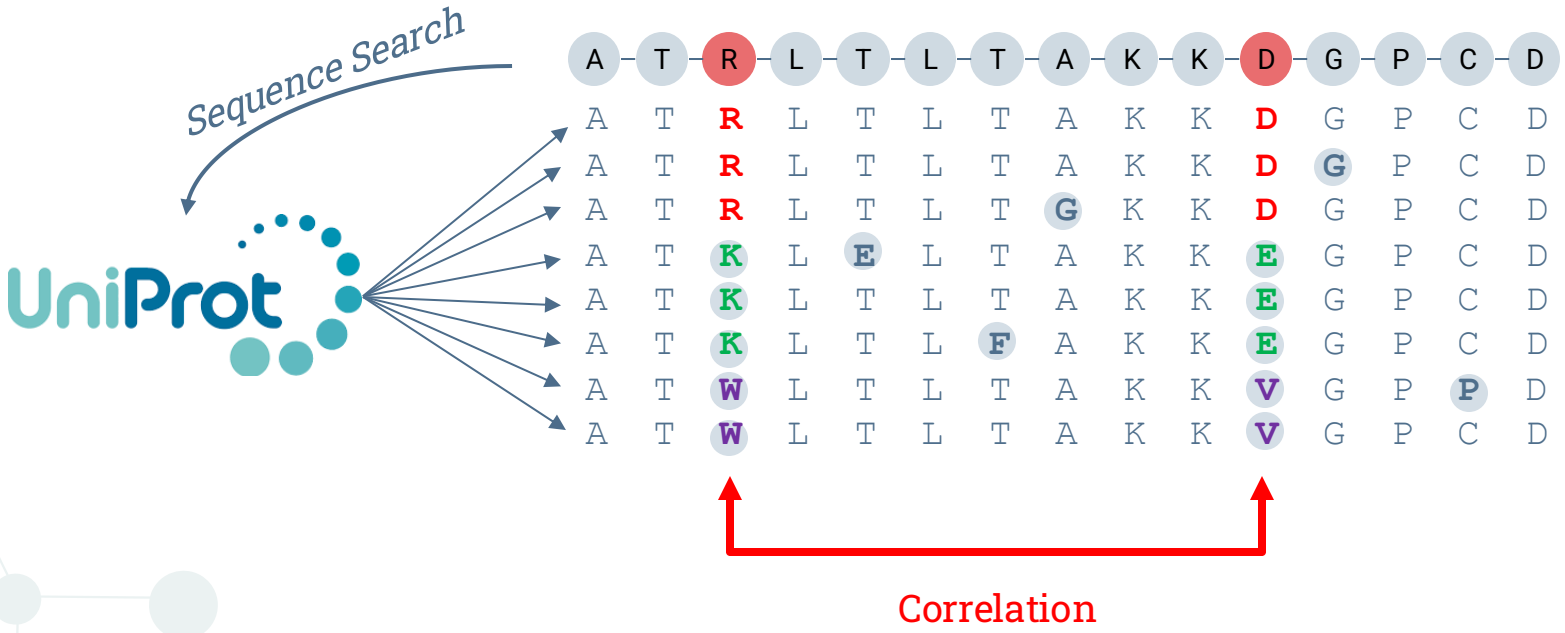


Inspired by https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-i

AlphaFold2 – How does it work?

Let's make it simple…

First, AF2 is based on the principle of **co-evolution**

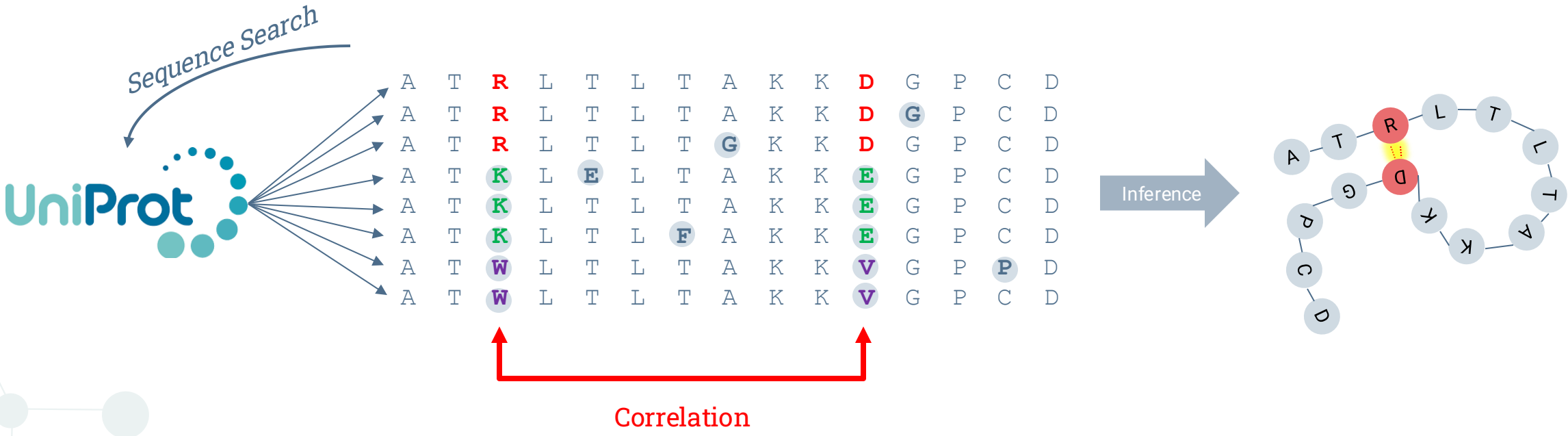


Inspired by https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-i

AlphaFold2 – How does it work?

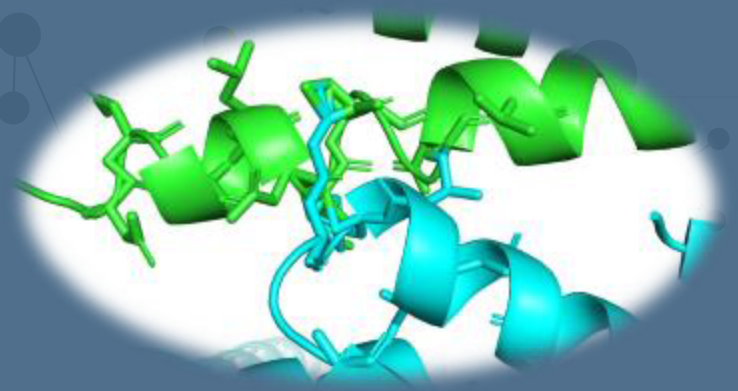
Let's make it simple…

First, AF2 is based on the principle of **co-evolution**

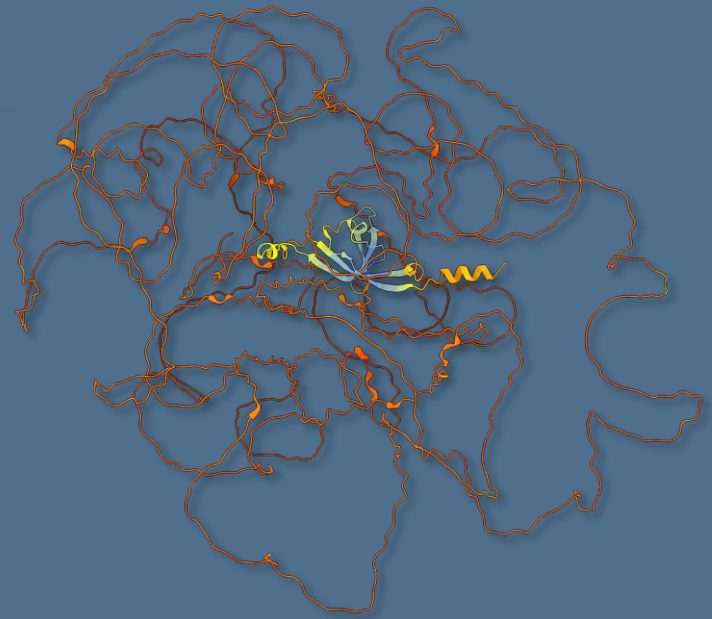


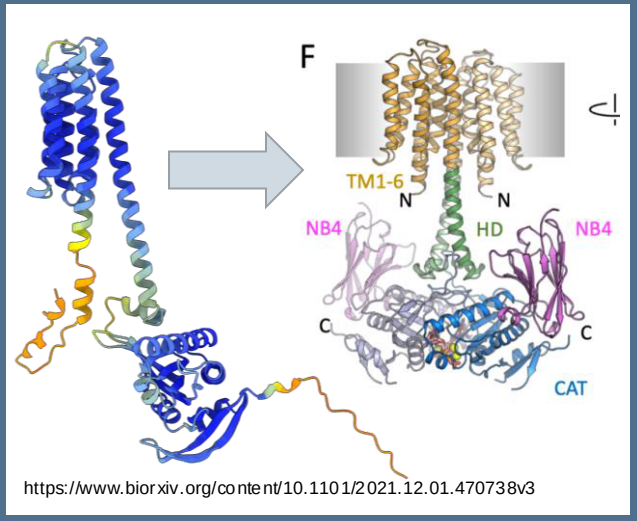
If 2 residues are **correlated**, they should be **close in space**.

Inspired by https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-i

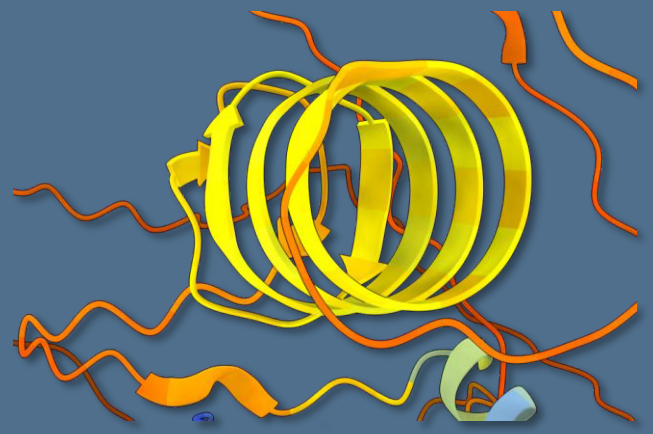


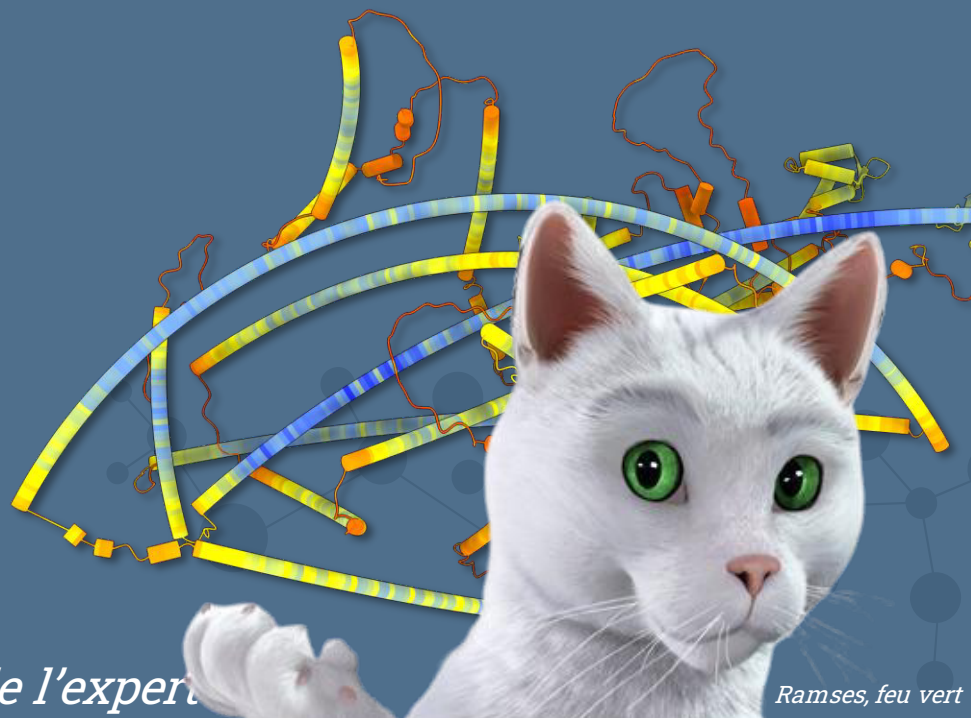
https://twitter.com/pedrobeltrao/status/1501214534681378818





https://www.biorxiv.org/content/10.1101/2021.12.01.470738v3





*AlphaFold is a **new** and **powerfull** tool, but **not perfect**.*
It cannot replace experimental data or the expert's eye (« la patte de l'exper

Ramses, feu vert

One more thing…

A few month ago, AF3 was release with « huge improvements »

- **Single protein**
- **Protein-Protein**
- **Antibody – Antigene**
- **Protein-DNA**
- **Protein-ION**
- **Protein-Ligand**
- **Post traductional modifications..**

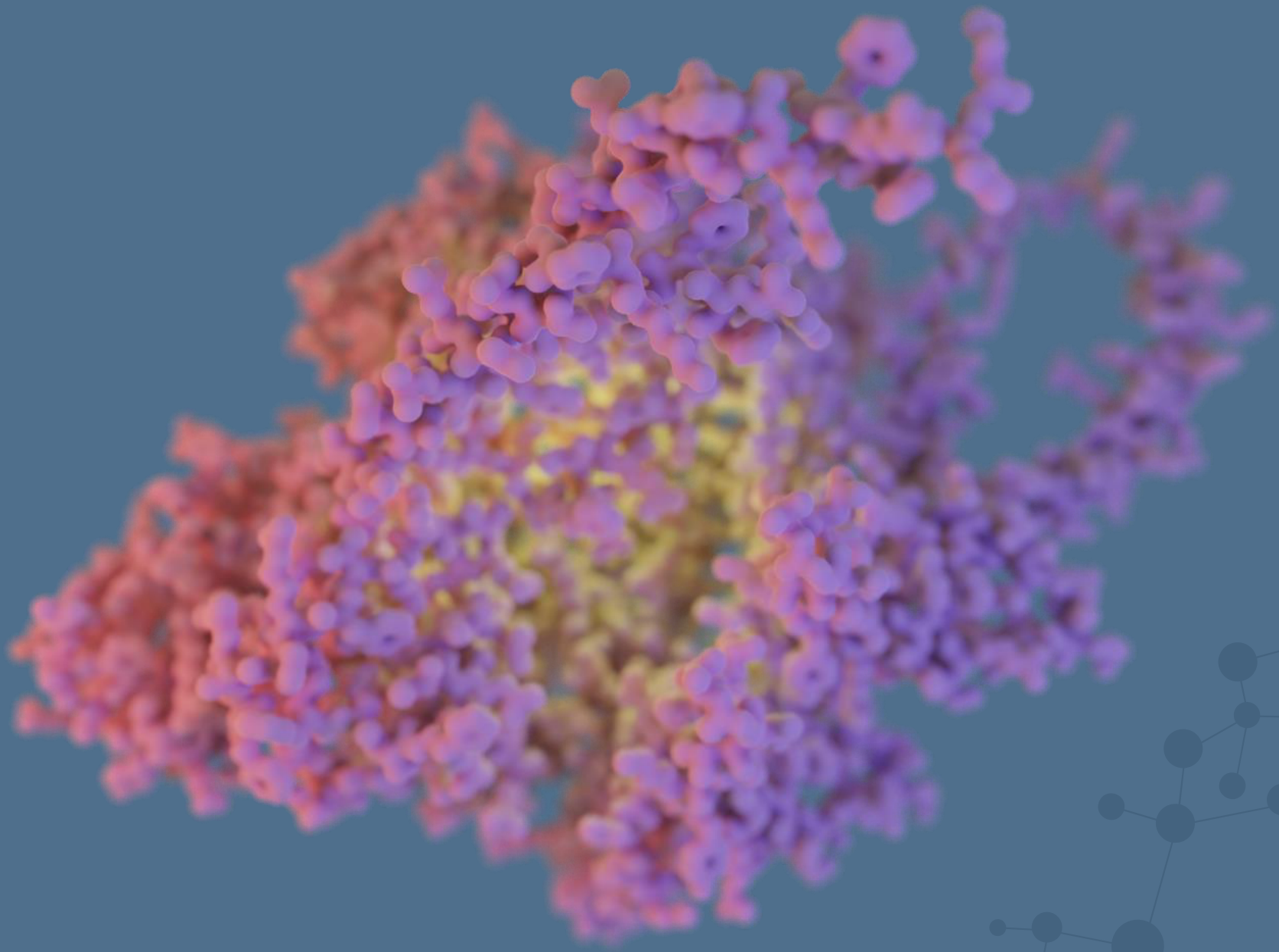
Code released only 2 week ago. Only limited to 20 daily jobs until now.

→ **Hard to have a extensive community validation!**

Instead of! the **attention** mechanism, they use a **diffusion** mechanism, a bit like Dall-E or Stable Diffusion

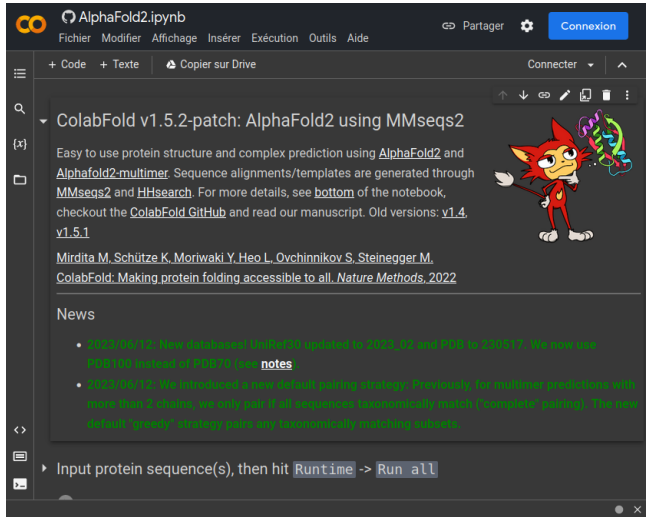
PART II

2. How to use AlphaFold?



AlphaFold2 – How to use it ?

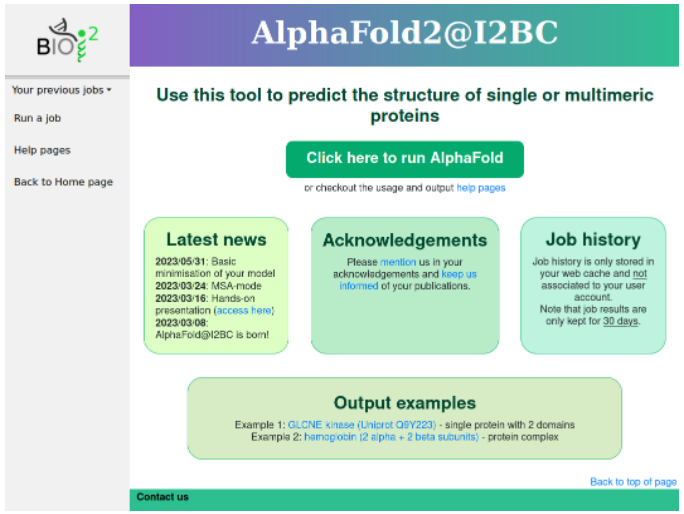
- ColabFold using Google Colab



freely accessible & uses Google's ressources (https://colab.research.google.com/github/sokrypton/ColabFold/)

ColabFold: Mirdita et al 2022

- AlphaFold@I2BC using ColabFold's pipeline



https://bioi2.i2bc.paris-saclay.fr/django/alphafold/

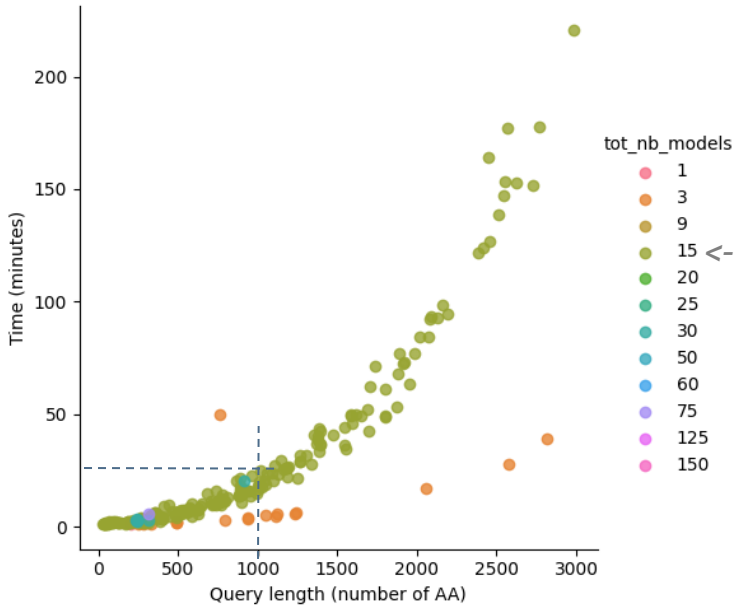
-> Its strength lies in the prediction of protein complexes (up to 3 000 AAs total)
-> User-friendly & fast
-> Within the I2BC (I2BC resources & storage, access restricted to users with an I2BC account)

⚠ **Multipass login & password** => use the Multipass website to get your credentials if not sure: https://multipass.i2bc.paris-saclay.fr/

AlphaFold2 @I2BC

- A few stats since march 2023: 86 users, 1570 jobs
  - Run time increases exponentially with input sequence length
  - Input < 1000 AA (~70% of jobs) with default options, run time should be less than 25 minutes

Time per model x recycles x repeats



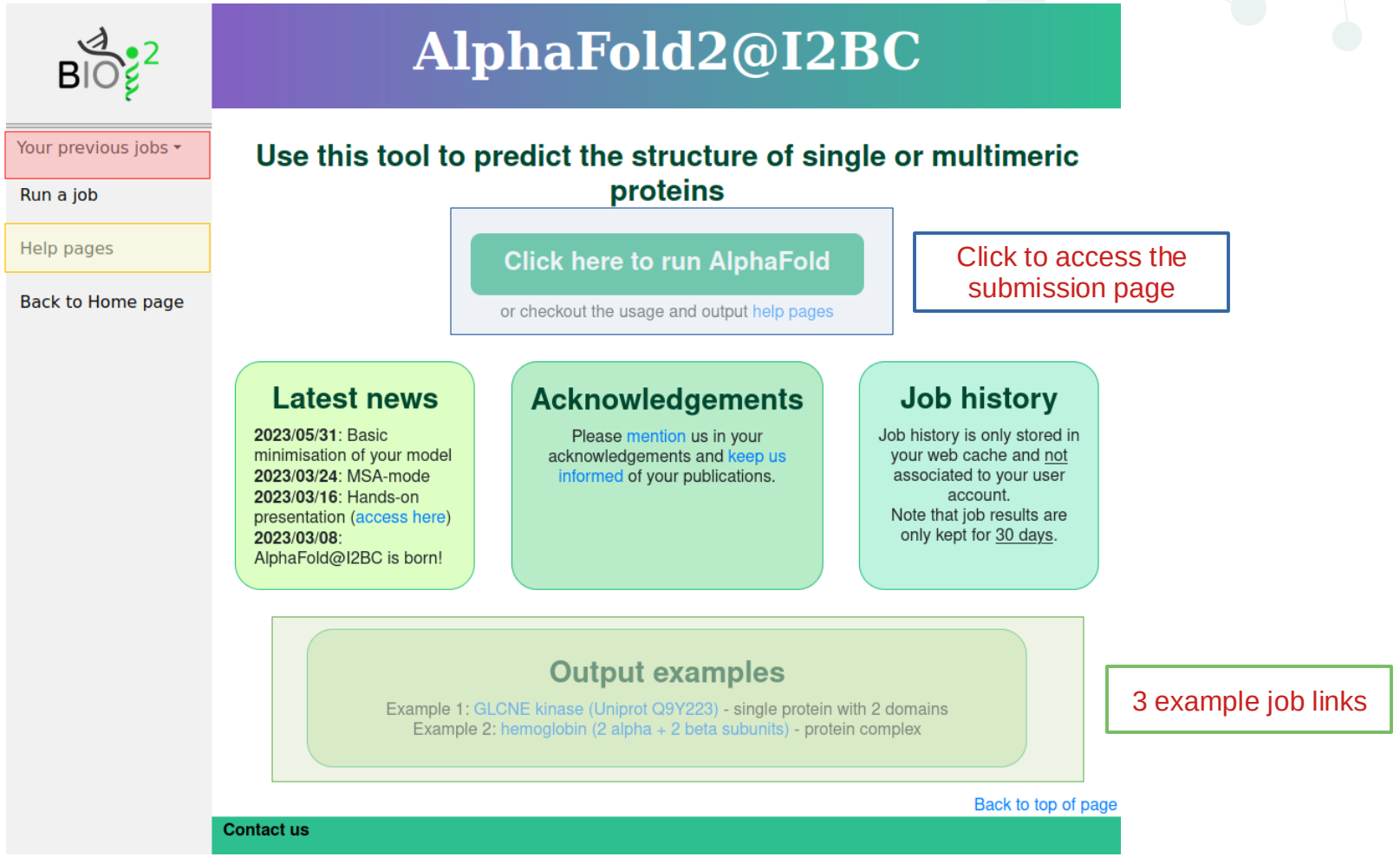
<- A run with default options (5 models x 3 recycles x 1 repeat)

Stats last updated on 7th of July 2023

AlphaFold2 @I2BC

List of previous jobs < 30 days old & only in web cache

Help pages on input format, output explanation and other details

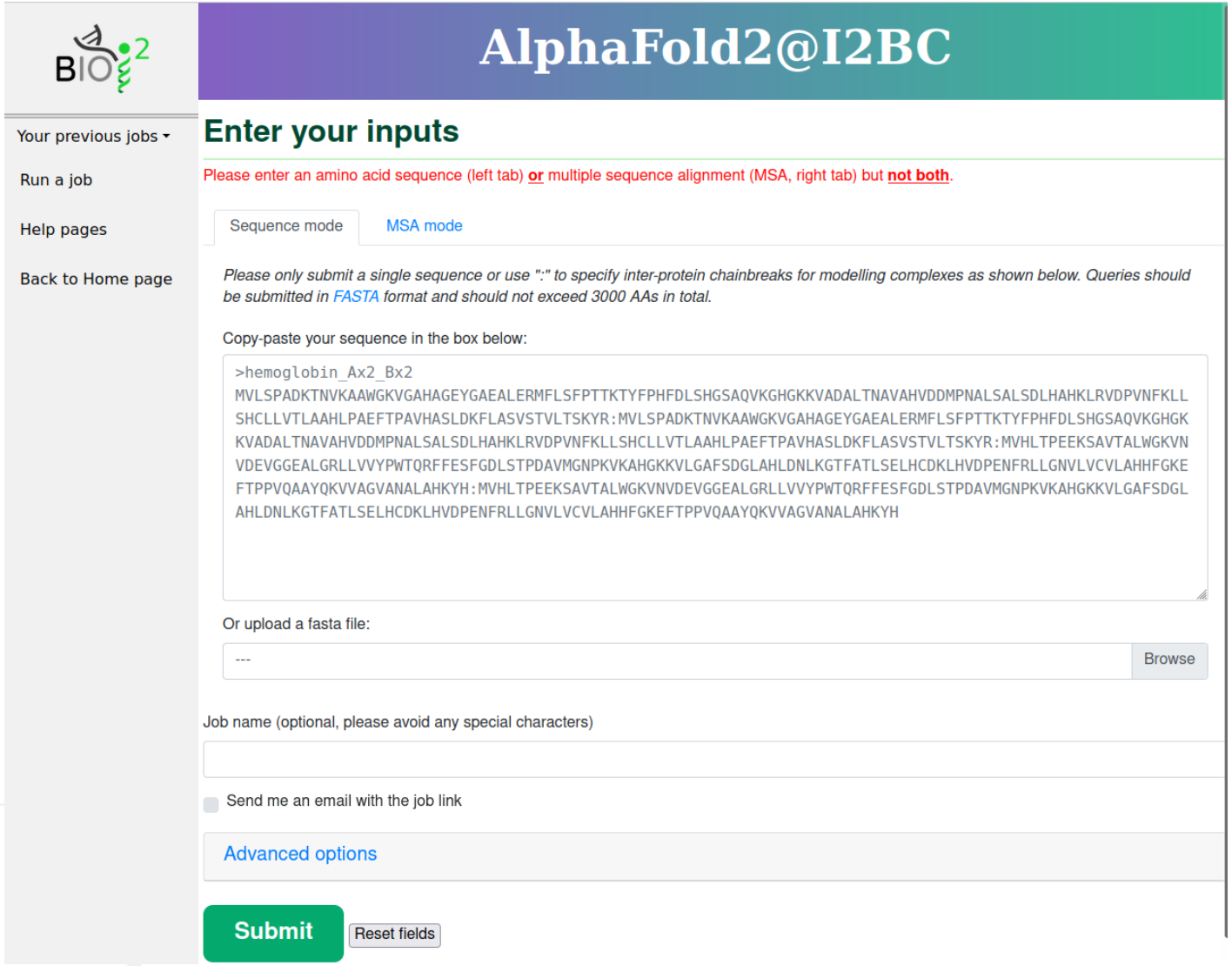


Click to access the submission page

3 example job links

https://bioi2.i2bc.paris-saclay.fr/alphafold/

AlphaFold2 @I2BC



*<- **2 input modes** available: sequence or MSA*

In both cases, you can either:

*<- **copy-paste** your input*

or

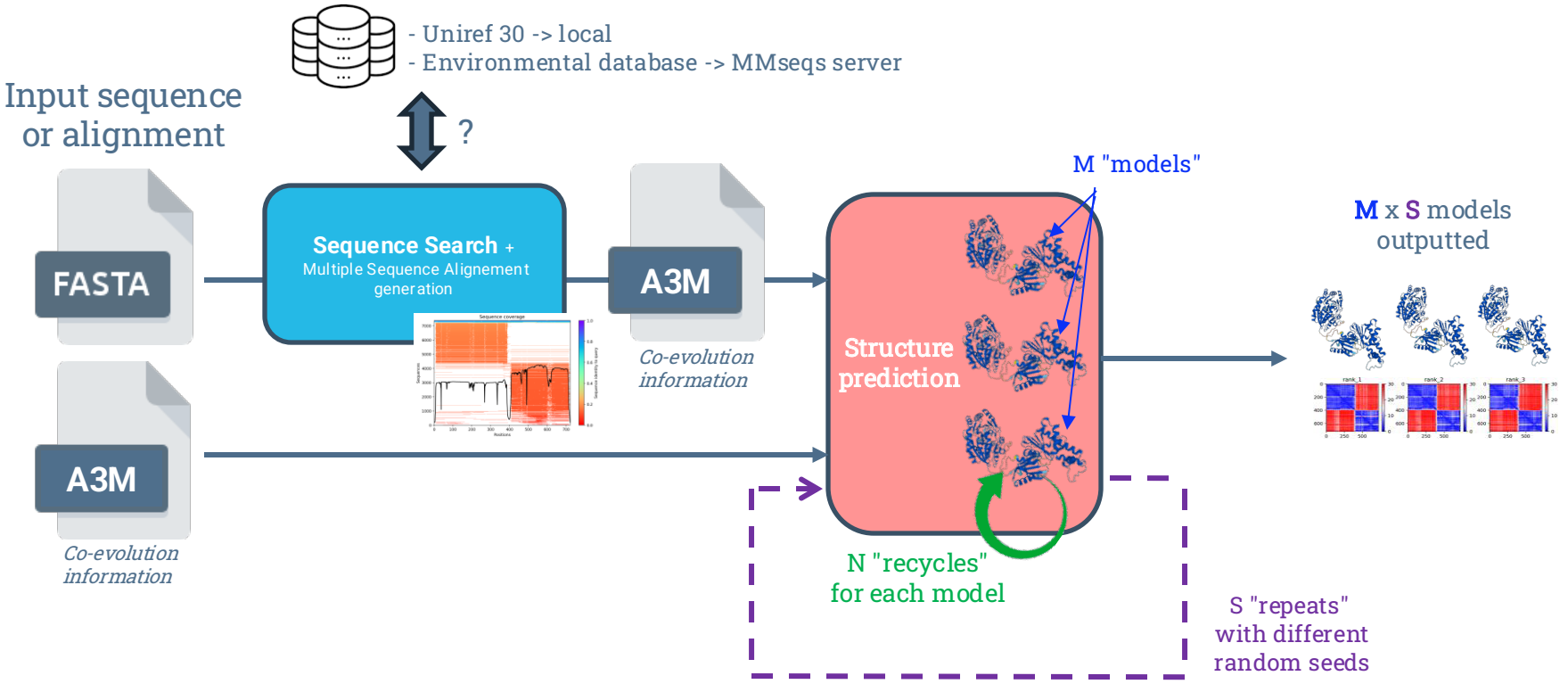
*<- **upload** it in the right format*

<- adding a job name is optional

<- advanced options

AlphaFold2 @I2BC

AlphaFold@I2BC's specific pipeline:



Resources available:

MSA generation

40 CPUs "big ram"
= 40 jobs at most
(very quick)

Structure prediction

6 GPUs
= 6 jobs at most
(bottleneck as longest step)

AlphaFold2 @I2BC – advanced options

-> (for sequence mode) Use metagenomic databases
Using these databases should be used only for rapidly diverging species which have few homolog sequences in the default ones

- Uniref 30
- Environmental database

Input sequence or alignment

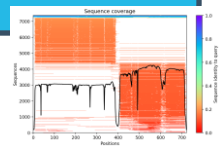
?

FASTA

A3M

Co-evolution information

Sequence Search +
Multiple Sequence Alignement generation



A3M

Co-evolution information

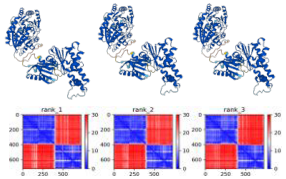
-> (for multimeric predictions) Pairing mode
To specify if you would like to use paired, unpaired or both sequence homologs found in the MSA generation step

-> Number of models
Different models correspond to different MSA sampling strategies during AlphaFold's training

M "models"

Structure prediction

M x **S** models outputted



N "recycles" for each model

S "repeats" with different random seeds

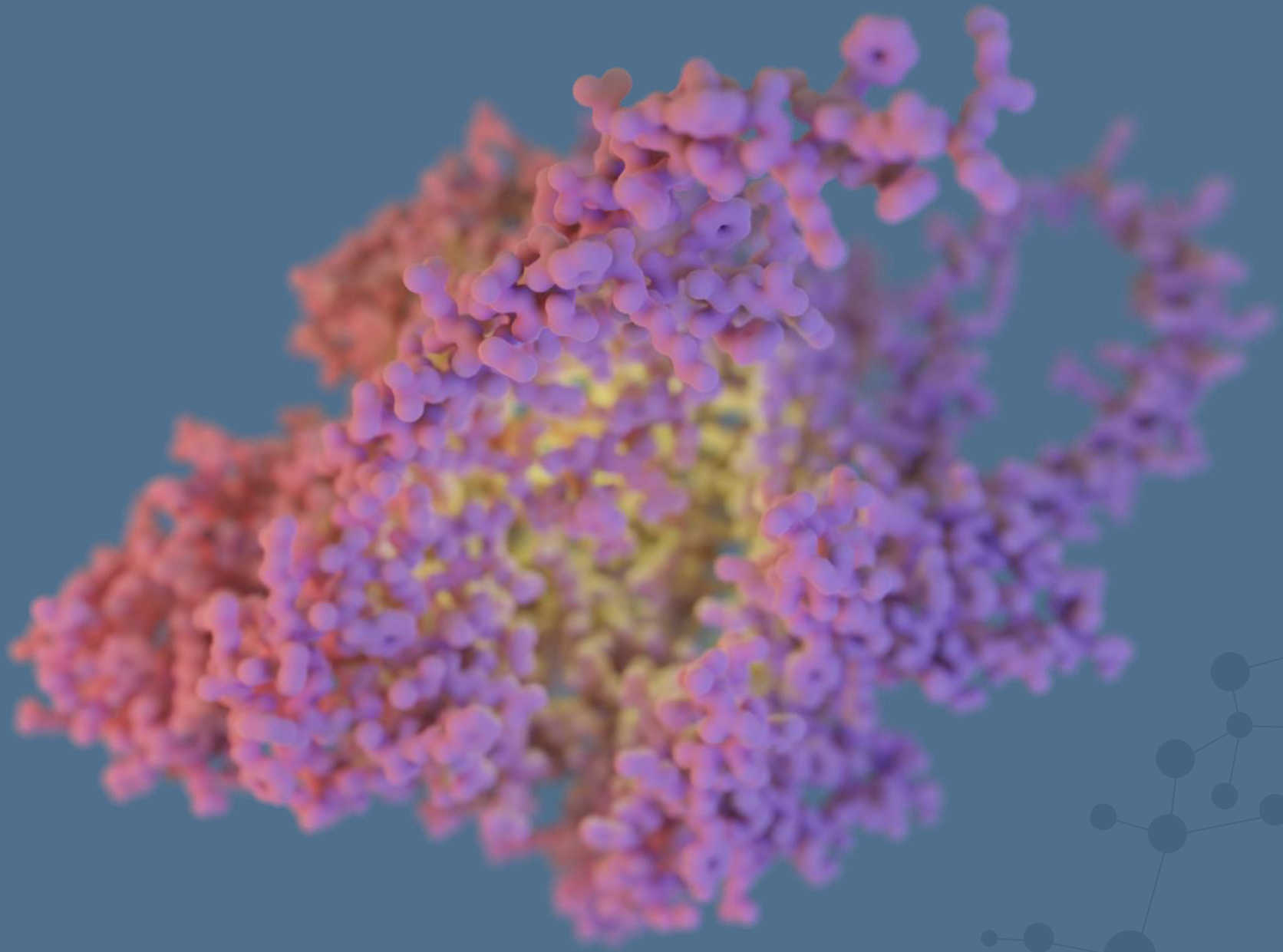
-> AlphaFold version: v2.2 or v2.3
v2.3 is the most recent and is the one that is advised. However, it can be useful to also use v2.2 in difficult cases

-> Number of recycles
The more recycles (=iterations), the more time your models have to converge but they also take longer to predict

-> Number of repetitions
For each repetition, AlphaFold will use a different random seed to begin with. This could change the output and is particularly useful for difficult models. As for recycles, repeating the whole prediction process will increase prediction time

PART II

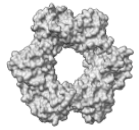
3. How to understand predictions?



AlphaFold – output

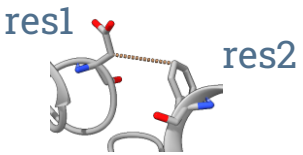
AlphaFold outputs quality estimates of your models at 3 levels:

- Global level



- Per residue level

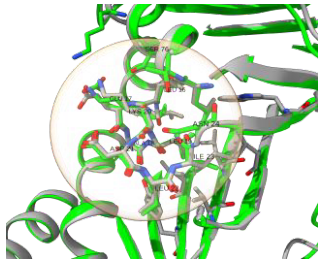
- Per residue pair level



Per-residue score

Based on an already-existing **similarity score (lDDT)**, used to compare 2 structures, usually a model *vs* a real structure:

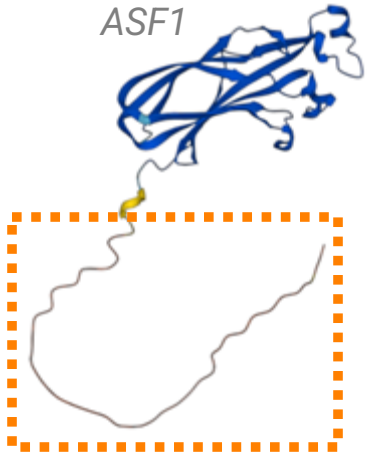
lDDT-Cα (Mariani et al., 2013)
local Distance Difference Test



- Assesses correctness of **local distances** without needing superposition of the structures
- **Per-residue** score

=> how well are local atomic interactions in the reference structure reproduced in the prediction?

predicted LDDT (pLDDT) *i.e.* theoretical true structure
= confidence in the local environment of a residue

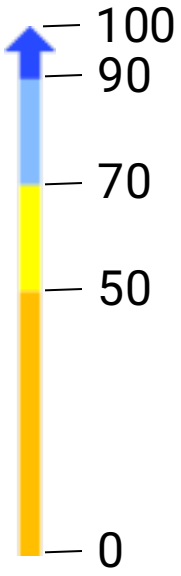


High confidence for BB + SC

Confident for BB

Low confidence

*Low confidence,
should not be interpreted,
reasonably strong predictor of disorder*

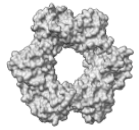


About pLDDT: https://alphafold.ebi.ac.uk/faq#faq-12

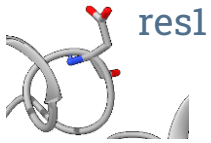
AlphaFold – output

AlphaFold outputs quality estimates of your models at 3 levels:

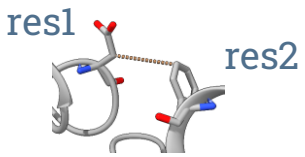
- Global level



- Per residue level = **pLDDT**



- Per residue pair level

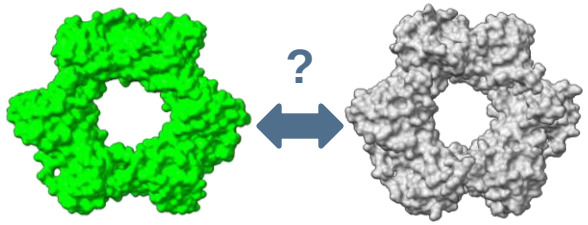


Global scores

- **Average pLDDT**: average over all residue scores in the model ([0-100]) => used to rank models in case of monomers

The other global scores are all based on another **similarity score**, also used to compare 2 structures:

TM-score (Zhang et al., 2004)
Template Modelling score



- Measures **topological similarity** between structures, with higher emphasis on small distance errors than larger ones
- **Structure-wide** metric

=> how similar is a prediction to a reference structure in terms of global shape?

- **predicted TM-score (pTM-score)**

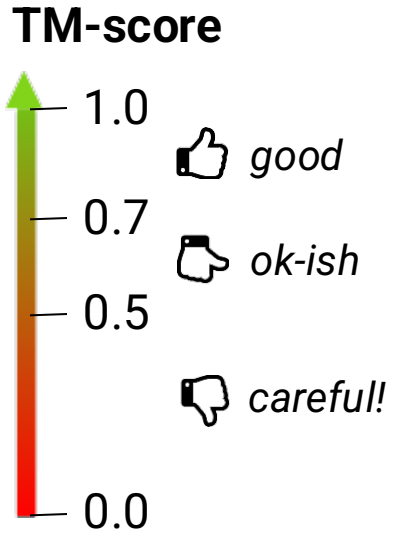
= confidence in global shape of the structure

- **interface pTM-score (ipTM-score)**

for multimers
= confidence in global structure of the interface

- **combined pTM-score**

for multimers
= 20% pTM-score + 80% ipTM-score

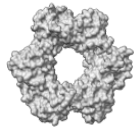


ipTM-score => used to rank multimeric models

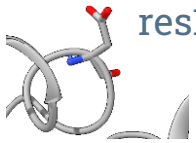
AlphaFold – output

AlphaFold outputs quality estimates of your models at 3 levels:

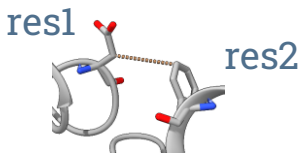
- Global level = **average pLDDT** + **pTM-score** + [**ipTM-score** + **combined pTM-score** (multimers)]



- Per residue level = **pLDDT**



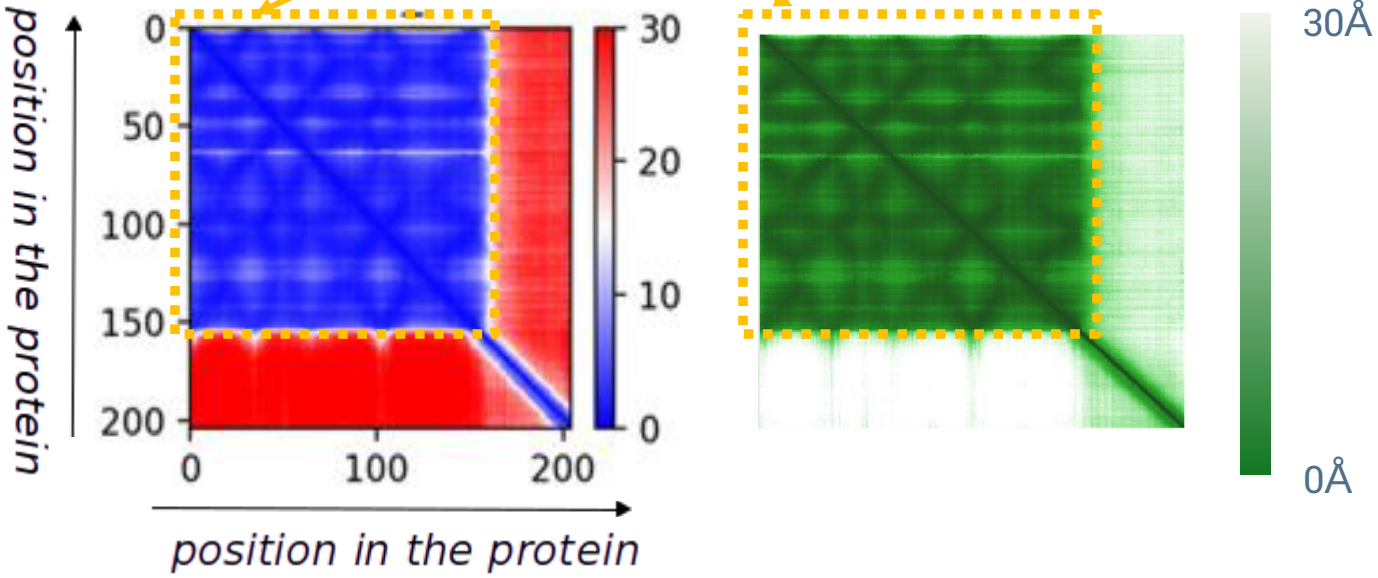
- Per residue pair level

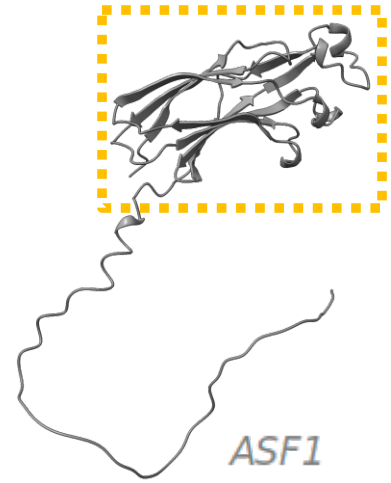


Pairwise "scores"

- **PAE** (predicted aligned error) = estimated **error on the distance** between 2 residues
- In Å, [0 - 32], lower = better

You can easily spot predicted domains
= uniform blocks





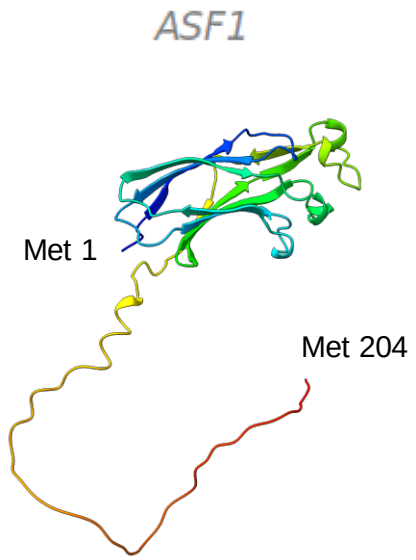
2 "parts":
-> well structured domain in *N-ter*
-> disordered tail in *C-ter*

- In short, measures confidence of:
  - Domain packing
  - Large-scale topology (*i.e. how accurate is the relative placement of the domains?*)

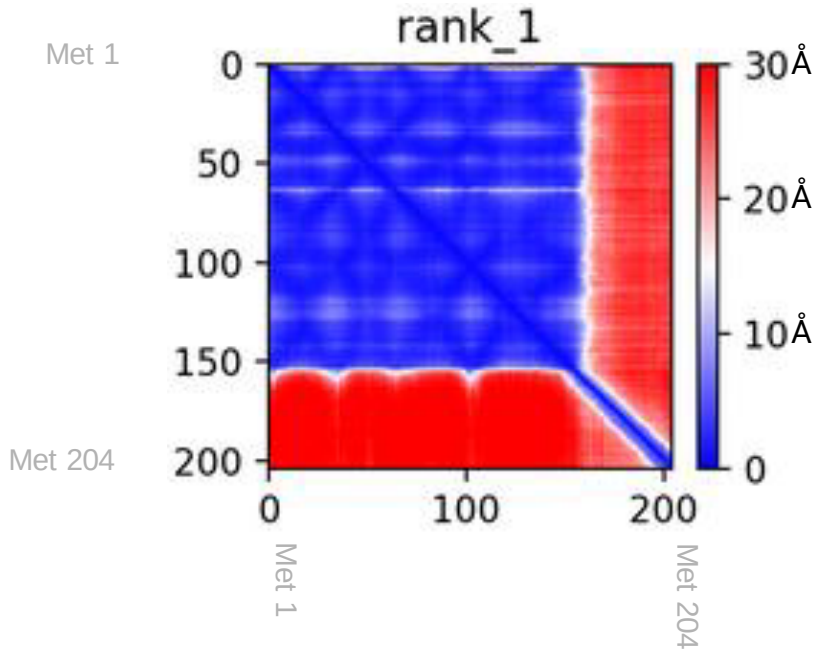
About PAE: https://alphafold.ebi.ac.uk/faq#faq-13

Pairwise "scores"

How to read the plot:



coloured by position



You can read it as: distance between residues x & y is equal to D +/- E

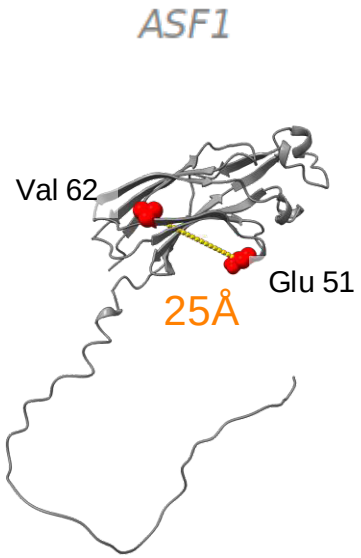
- E: error between x & y in the plot above
- D: distance between x & y in the model

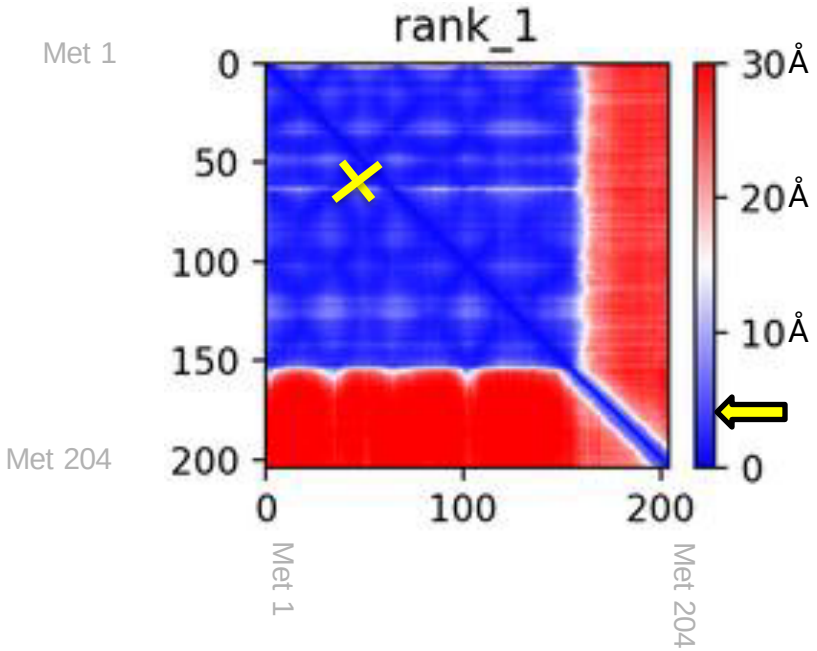
About PAE: https://alphafold.ebi.ac.uk/faq#faq-13

Pairwise "scores"

How to read the plot:

The distance between residues GLU-51 & VAL-62 is equal to 25 **+/- 4** Å
= ***high confidence*** *in relative position*
-> linked to GLU-51 and VAL-62 being in the same structural domain





You can read it as: distance between residues x & y is equal to D +/- E

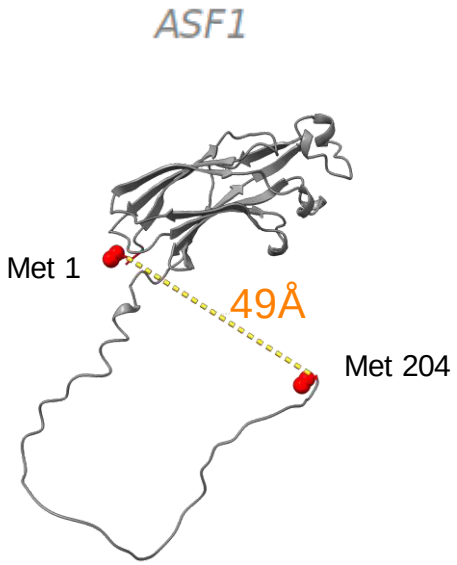
- E: error between x & y in the plot above
- D: distance between x & y in the model

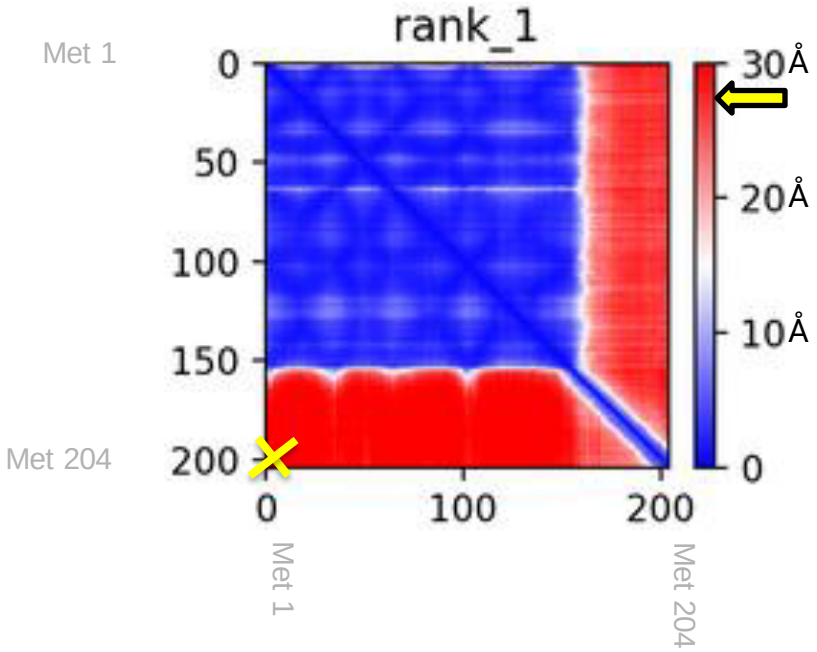
About PAE: https://alphafold.ebi.ac.uk/faq#faq-13

Pairwise "scores"

How to read the plot:

The distance between residues MET-1 & MET-204 is equal to 49 **+/- 27** Å
= ***low confidence*** *in relative position*





You can read it as: distance between residues x & y is equal to D +/- E

- E: error between x & y in the plot above
- D: distance between x & y in the model

About PAE: https://alphafold.ebi.ac.uk/faq#faq-13

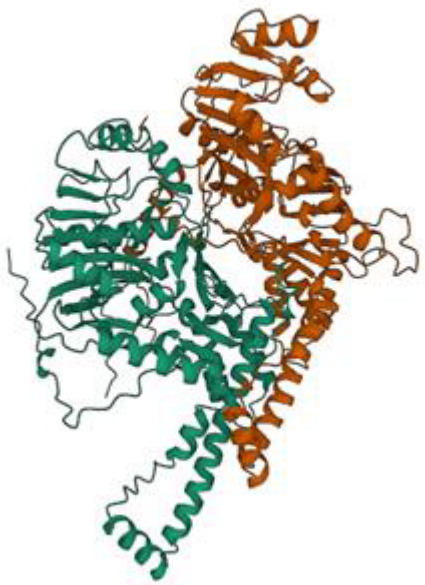
Pairwise "scores"

This plot is **mainly a general view**, it should not be read pixel by pixel!

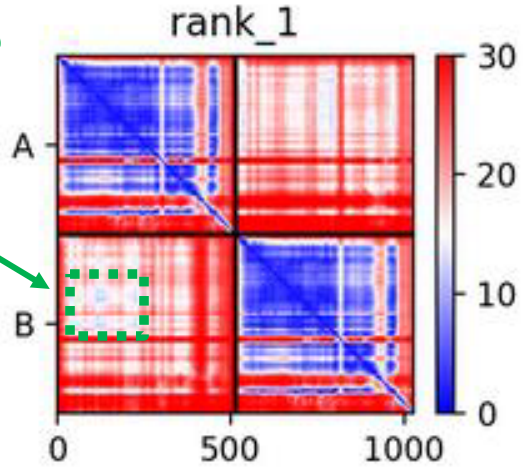
If you remember, it measures the confidence of:

- Domain packing
- Large-scale topology => *of which potential interfaces!*

e.g. N-ter dimer of pORF1:



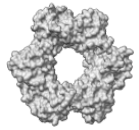
Interacting regions between sections (within a same protein or between different protein chains) usually show up as low-error patches that aren't along the diagonal



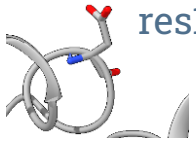
AlphaFold – output

AlphaFold outputs quality estimates of your models at 3 levels:

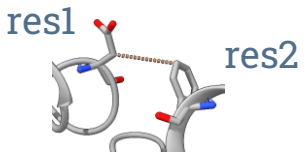
- Global level = **average pLDDT** + **pTM-score** + [**ipTM-score** + **combined pTM-score** (multimers)]



- Per residue level = **pLDDT**



- Per residue pair level = **PAE**

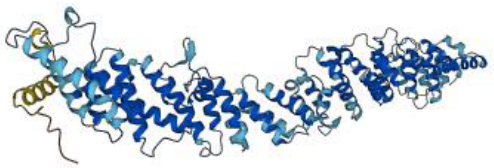


It also outputs summary plots for your MSA

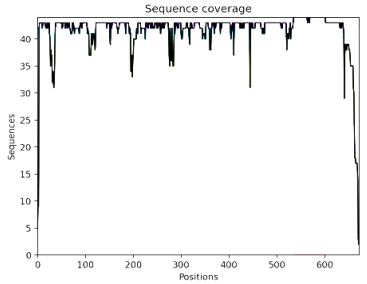
About the MSA plot

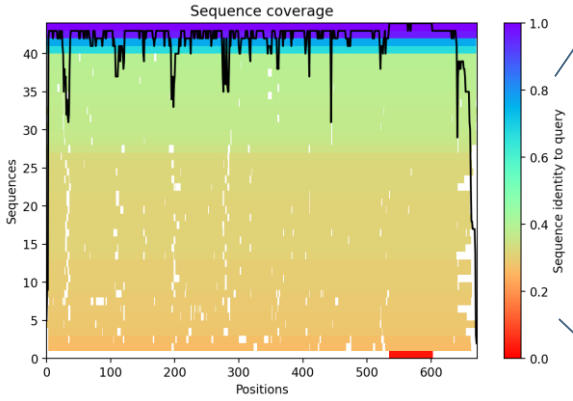
MSA plot = summary of your MSA (seq id + coverage)

SEC39 protein **monomer** (Q12745)



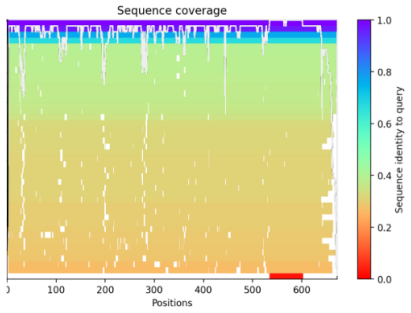
General coverage





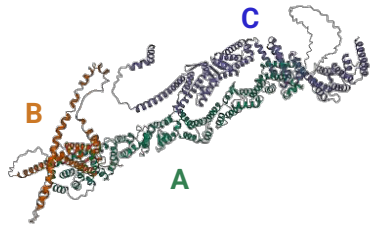
=> number of sequences covering each position along your input sequence

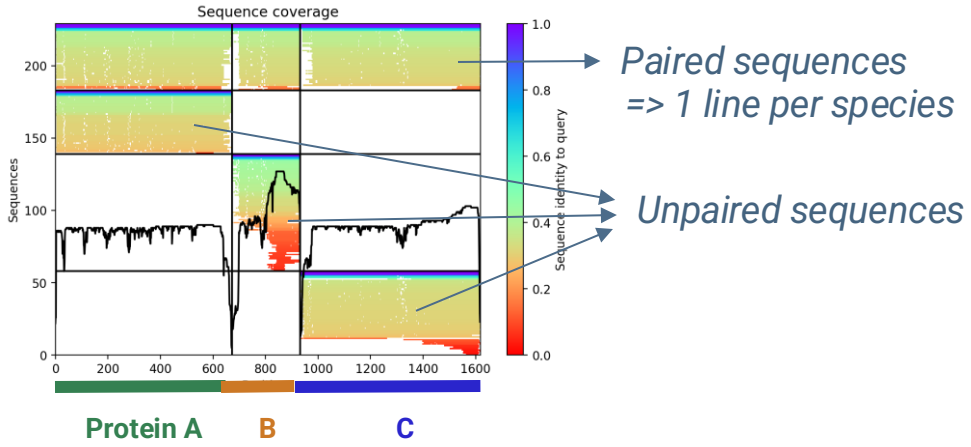
Sequence identities



=> 1 horizontal line per sequence in your MSA coloured by sequence identity with your input sequence

SEC39 (Q12745) **complex** with 2 other partners



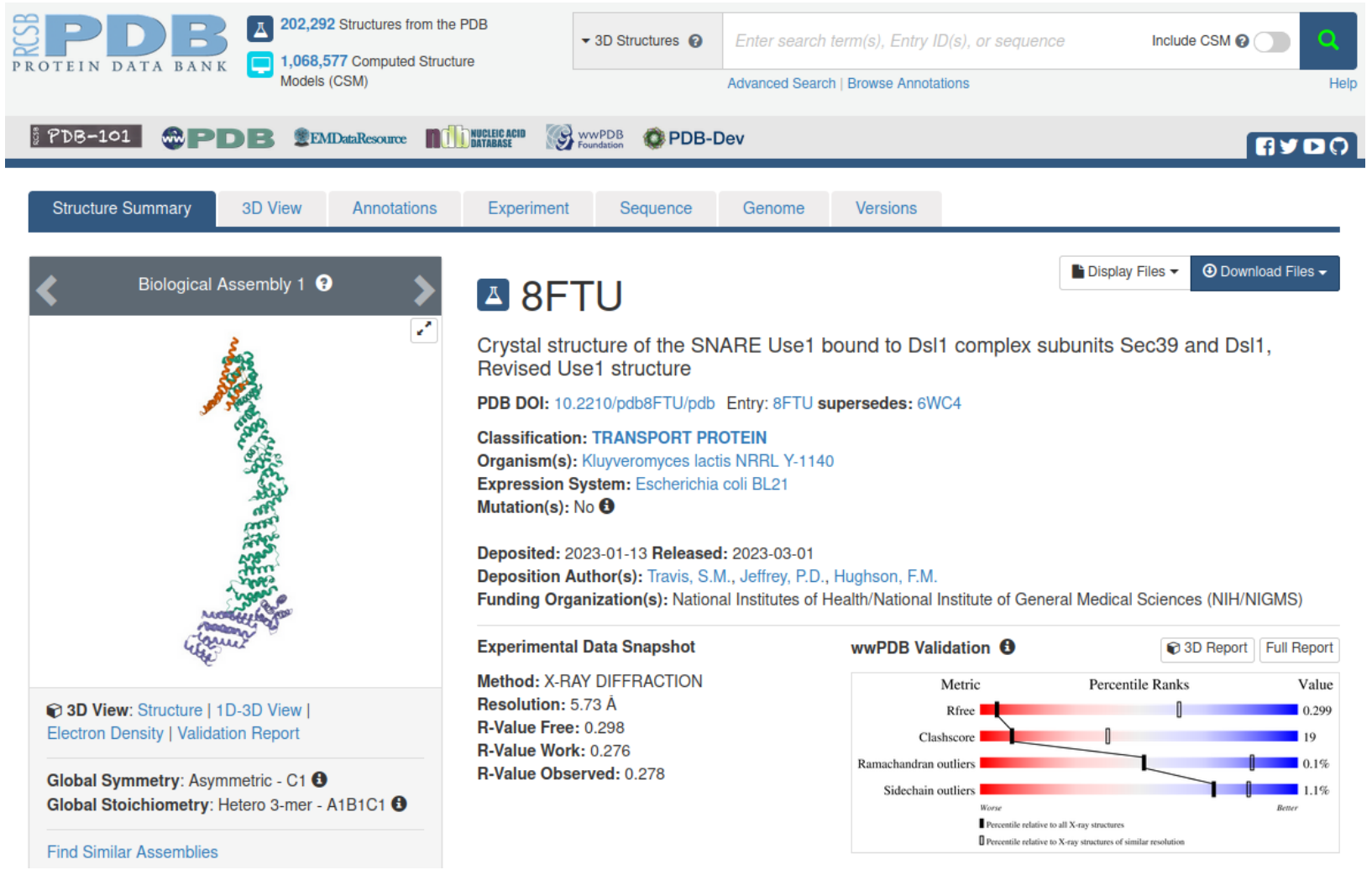


Paired sequences
=> 1 line per species

Unpaired sequences

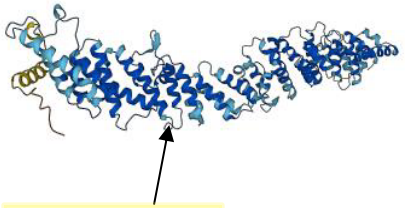
Example case - 8FTU

Recent PDB (8FTU), released after AF2 was trained, trimer between SEC39, DSL1 and USE1

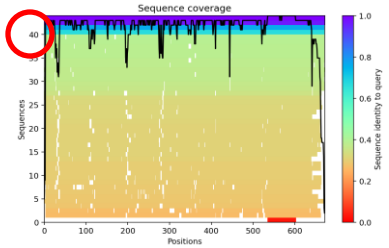


Example case - 8FTU

SEC39 alone https://bioi2.i2bc.paris-saclay.fr/django/alphafold/659a6d16-27b6-4c78-990a-e539e66c8d68/summary/

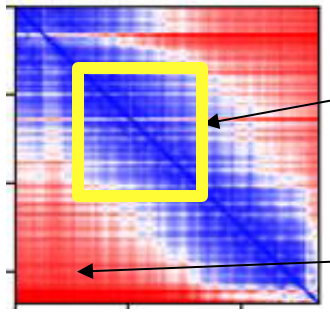


→ good pLDDT scores (mostly blue) even though the number of homologs is low (~45 sequences)



→ low pTM-scores => expected given the long shape of the protein (pTM-scores are calculated using the PAE matrix, i.e. the red in the matrix pulls the score down)

average pLDDT	pTM-score
87.58	0.67

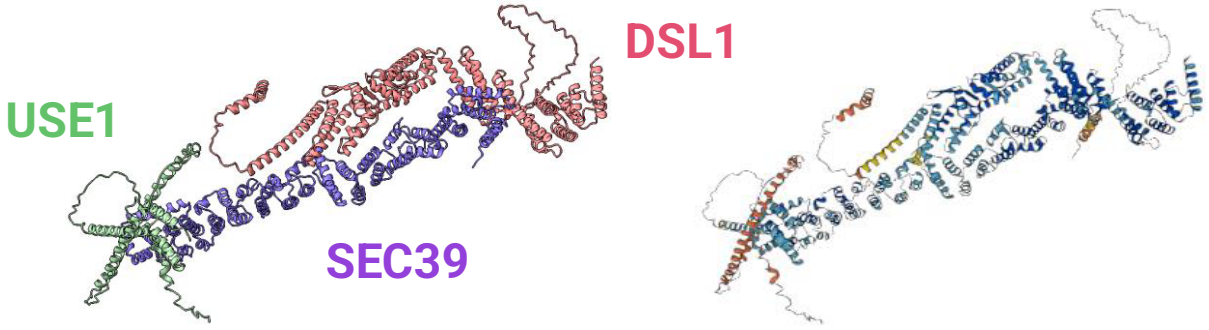


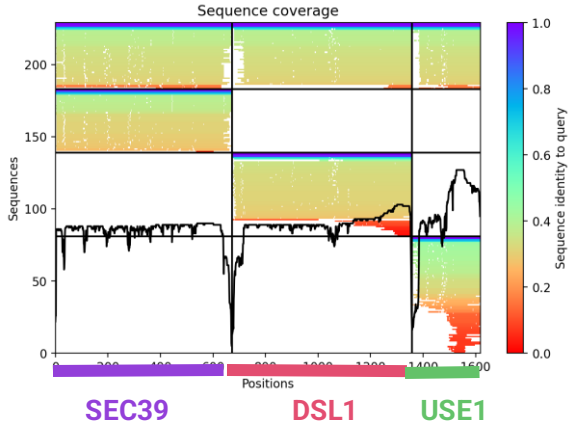
Sliding window along the diagonal of confident relative positions along the structure

...but relative position of Nter vs Cter is unsure

Example case - 8FTU

SEC39 + DSL1 + USE1 full lg https://bioi2.i2bc.paris-saclay.fr/django/alphafold/288fb6ea-bdd5-465f-8fa2-a336777e6e0e/summary

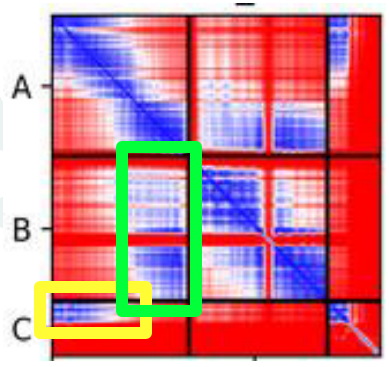




→ good pLDDT scores (mostly blue) except extremities and loops
→ uses all sequences found (paried+unpaired & redundancy, ~100)

average pLDDT	pTM-score	ipTM-score	Combined pTM-score
75.86	0.60	0.59	0.592

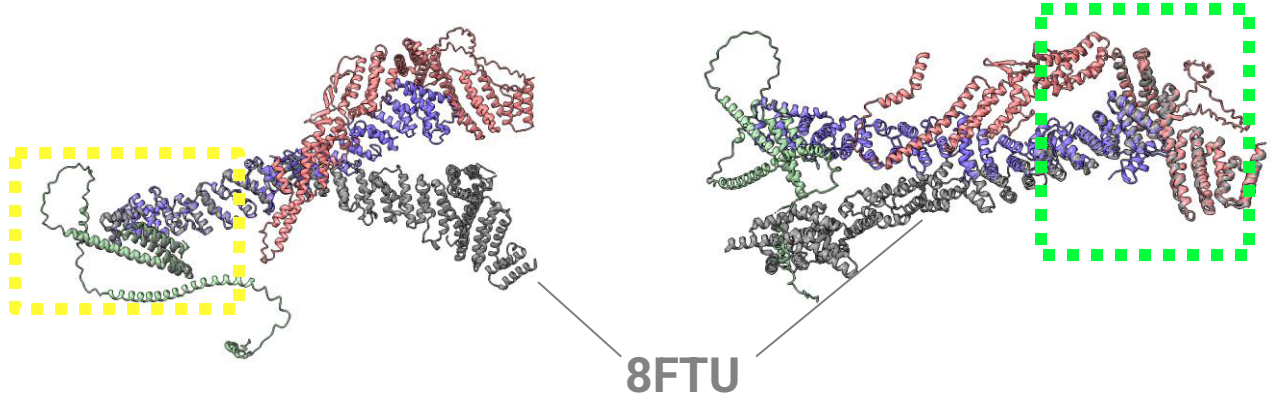
What do you think?



What now?

→ low pTM-score
→ low ipTM-score even though good interfaces vs true structure

Experimental structure vs prediction

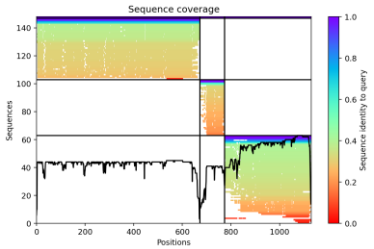


Example case - 8FTU

https://bioi2.i2bc.paris-saclay.fr/django/alphafold/ceb5190b-cc2e-4621-91a8-c0fb5921a628/summary/

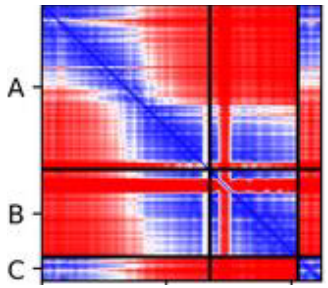
SEC39 + DSL1 + USE1 trimmed

→ we loose nearly all paired sequences (~45 unpaired / protein)

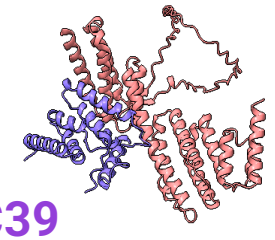


average pLDDT	pTM-score	ipTM-score	Combined pTM-score
82.62	0.60	0.68	0.664

→ low pTM-score
→ ipTM-score is slightly better but still low
→ we still see low-error regions @interfaces

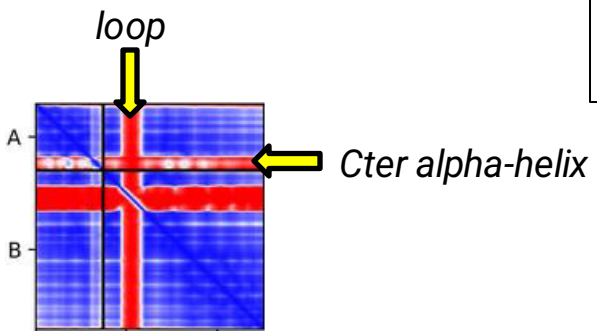


SEC39 Cter + DSL1 Nter



→ very globular, very good scores even though predicted interfaces
are very similar to the previous predictions

average pLDDT	pTM-score	ipTM-score	Combined pTM-score
83.29	0.81	0.85	0.842



Take home message:
Scores aren't absolute and depend on the context
-> playing on protein trimming or MSA

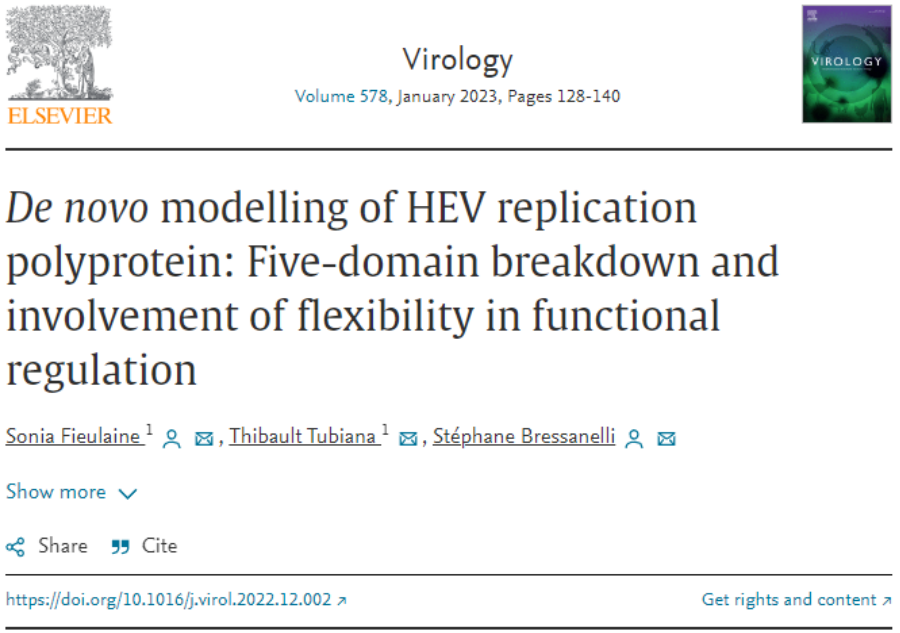
https://bioi2.i2bc.paris-saclay.fr/django/alphafold/6a55a9cd-0d5d-4163-a03b-6ea38cc56ec4/summary/

Real Life Demo

Goal: We want to study the structures of the pORF1 HEV polyprotein

About pORF1 HEV:
UniprotID : P33424
https://www.uniprot.org/uniprotkb/P33424/entry
Status : Reviewed

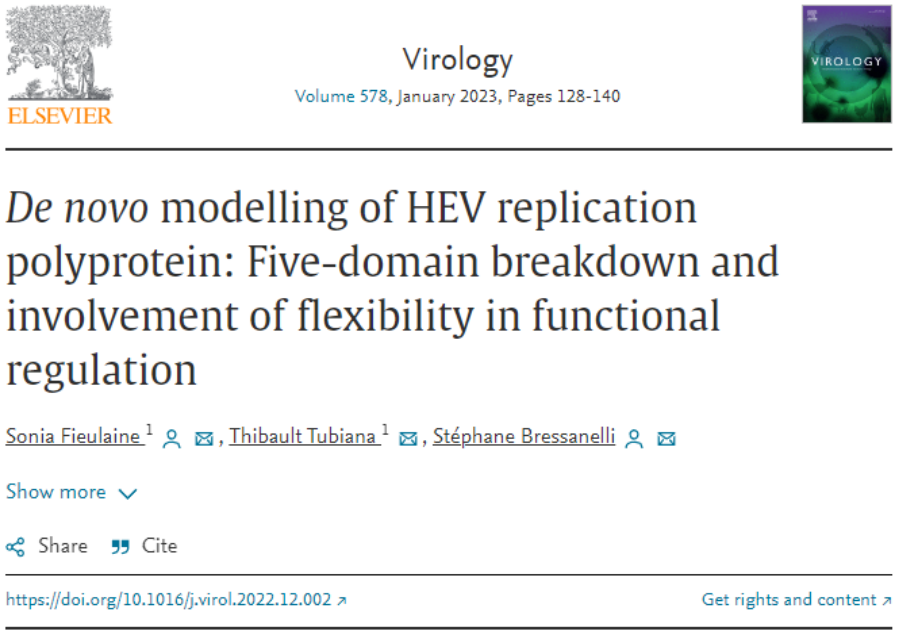
=> Have a look at the UniProt page for this protein, is there a structure available?



Real Life Demo

Goal: We want to study the structures of the pORF1 HEV polyprotein

About pORF1 HEV:
UniprotID : P33424
https://www.uniprot.org/uniprotkb/P33424/entry
Status : Reviewed
Structures : 1 X-ray (residues 510-691 only!!)
No AlphaFold models!! Why?
-> most of the viral proteins are excluded from automatic AlphaFold model generation…



Problem :
-> We don't have the full-length structure (no experimental data yet)
-> Protein is quite big (1700 aa)

"Solution": We can ***predict*** the structure ourselves

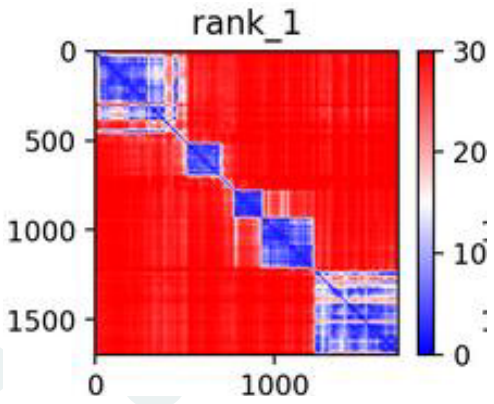
NB: Results of this study were made on local desktop with RTX A6000 + a lot of models + templates for the best one + minimisation with Isolde. But we'll be using the AlphaFold@I2BC server instead.

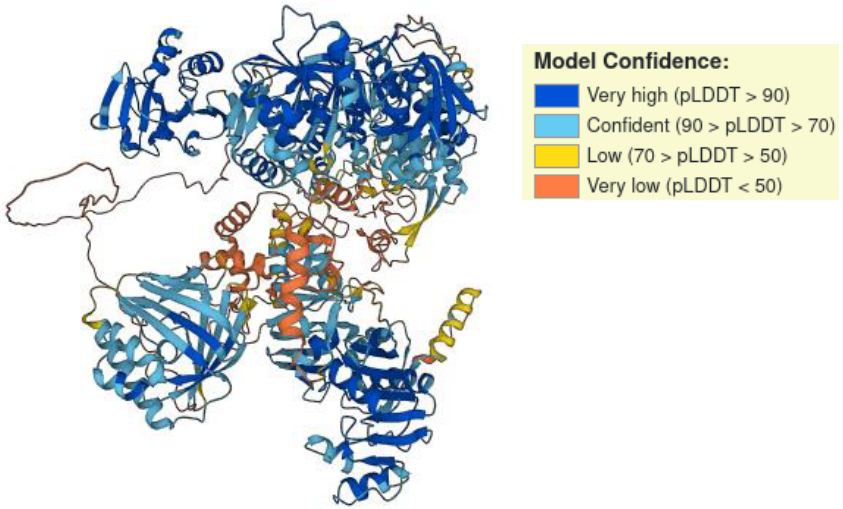
Real Life Demo

Goal: We want to study the structures of the pORF1 HEV polyprotein

=> no existing full-length structures in the known databases

=> We managed to predict a structure with AlphaFold, now what?





ELSEVIER

Virology

Volume 578, January 2023, Pages 128-140

De novo modelling of HEV replication polyprotein: Five-domain breakdown and involvement of flexibility in functional regulation

Sonia Fieulaine [1], Thibault Tubiana [1], Stéphane Bressanelli

Show more

Share Cite

https://doi.org/10.1016/j.virol.2022.12.002 ↗

Real Life Demo

Goal: We want to study the structures of the pORF1 HEV polyprotein

=> no existing full-length structures in the known databases

=> We managed to predict a structure with AlphaFold, now what?

… can we find any similar structures?

-> Let's search for structural homologs based on our models!

Foldseek (quick and dirty): https://search.foldseek.com/search

Dali (slow but more accurate): http://ekhidna2.biocenter.helsinki.fi/dali/

ELSEVIER

Virology

Volume 578, January 2023, Pages 128-140

De novo modelling of HEV replication polyprotein: Five-domain breakdown and involvement of flexibility in functional regulation

Sonia Fieulaine [1], Thibault Tubiana [1], Stéphane Bressanelli

Show more

Share Cite

https://doi.org/10.1016/j.virol.2022.12.002 Get rights and content

Results: job.pdb

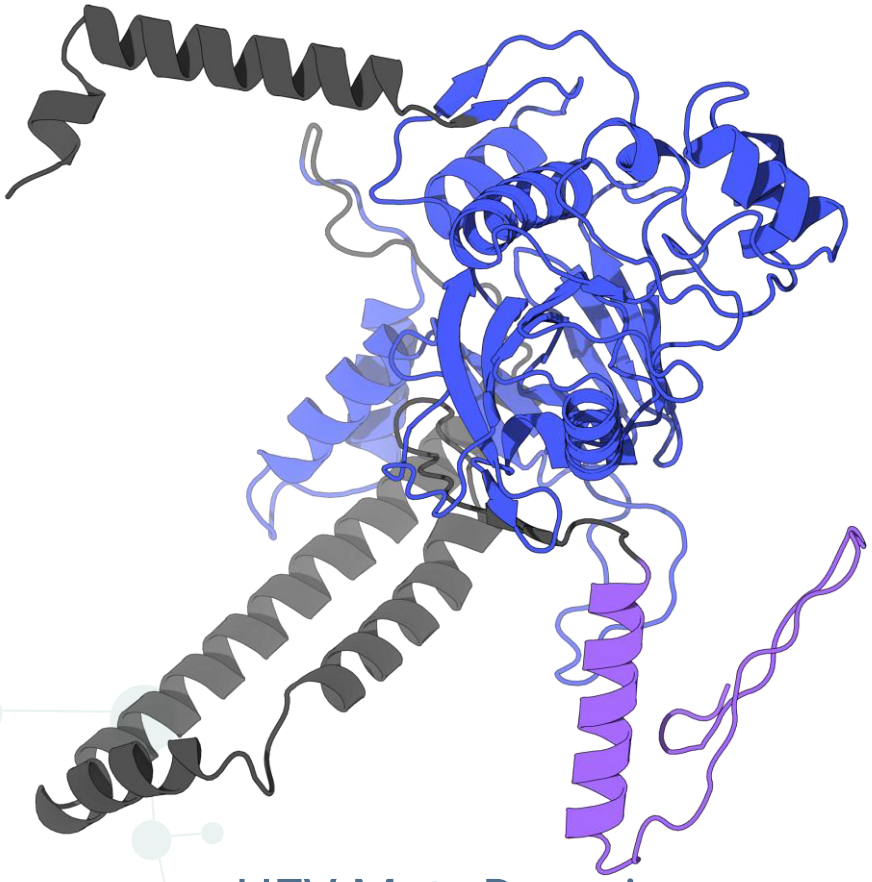
ALL DATABASES | AFDB-PROTEOME (2) | AFDB-SWISSPROT (1) | AFDB50 (77) | GMGCL_ID (0) | MGNIFY_ESM30 (1000) | PDB100 (35)

PDB100 35 hits — GRAPHICAL | NUMERIC

Target	Description	Scientific Name	Prob.	Seq. Id.	E-Value	Position in query	Alignment
6nu9_A	Crystal Structure of a Zinc-Binding No…	Hepatitis E virus (strain Pakistan)	1.00	95.4	8.37e-27	515 689	
4n0o_C	Complex structure of Arterivirus nons…	Equine arteritis virus Bucyrus	1.00	20.5	9.93e-9	959 1191	
3ew5_B	Structure of the tetragonal crystal for…	Feline infectious peritonitis virus (strai…	1.00	16.6	2.97e-5	779 925	
3jzt_E	Structure of a cubic crystal form of X …	Feline infectious peritonitis virus (strai…	1.00	19.8	3.12e-5	778 925	
3eti_C	Structure of a cubic crystal form of X …	Feline infectious peritonitis virus (strai…	1.00	18.7	5.83e-5	778 925	
5ziu_A	Crystal structure of human Entervirus …	enterovirus D68	1.00	13.5	2.40e-8	1267 1661	
5y6z_A	Crystal structure of the coxsackieviru…	Coxsackievirus A16	1.00	12.2	2.81e-8	1246 1661	

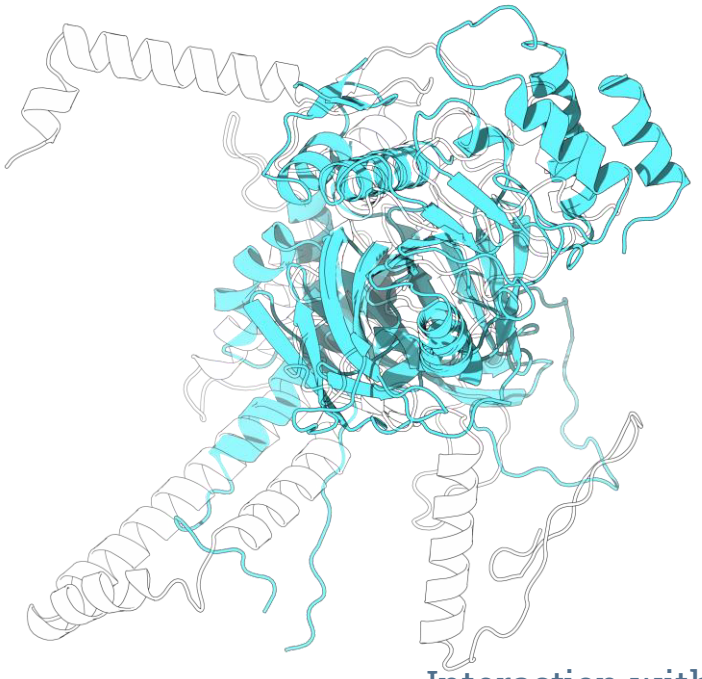
Real Life Demo

We have a Match!

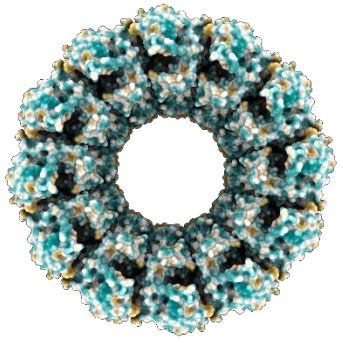


HEV Mety Domain

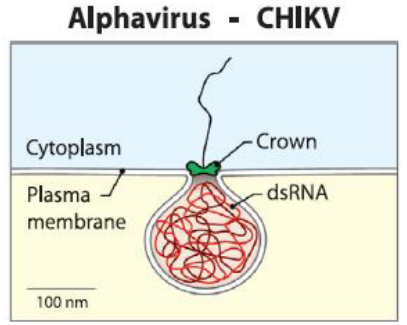
Structural homolog: NSP1 CHIKV



Natively dodecameric

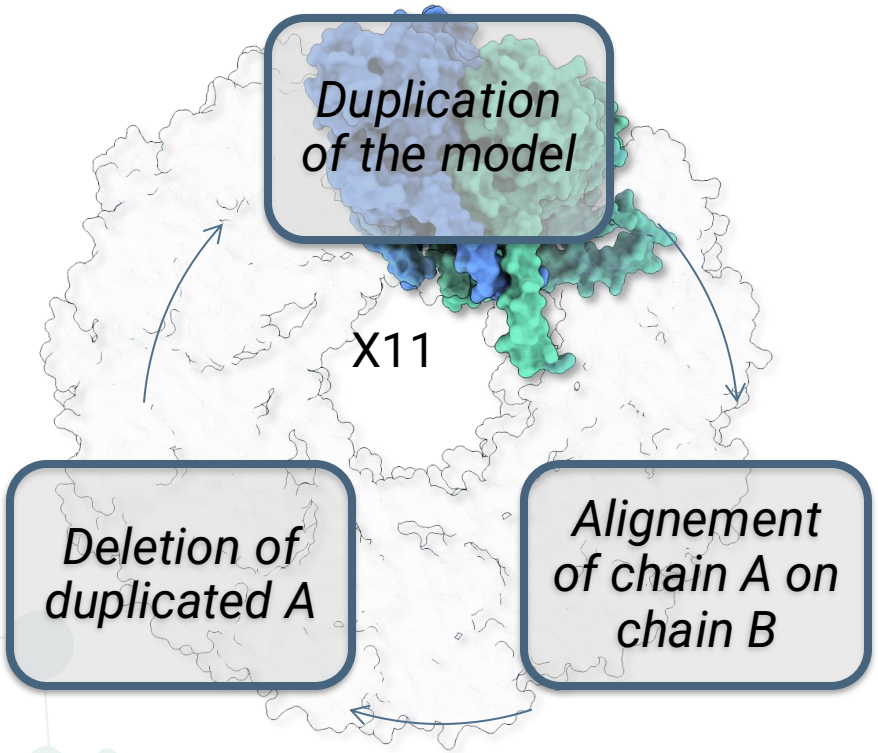


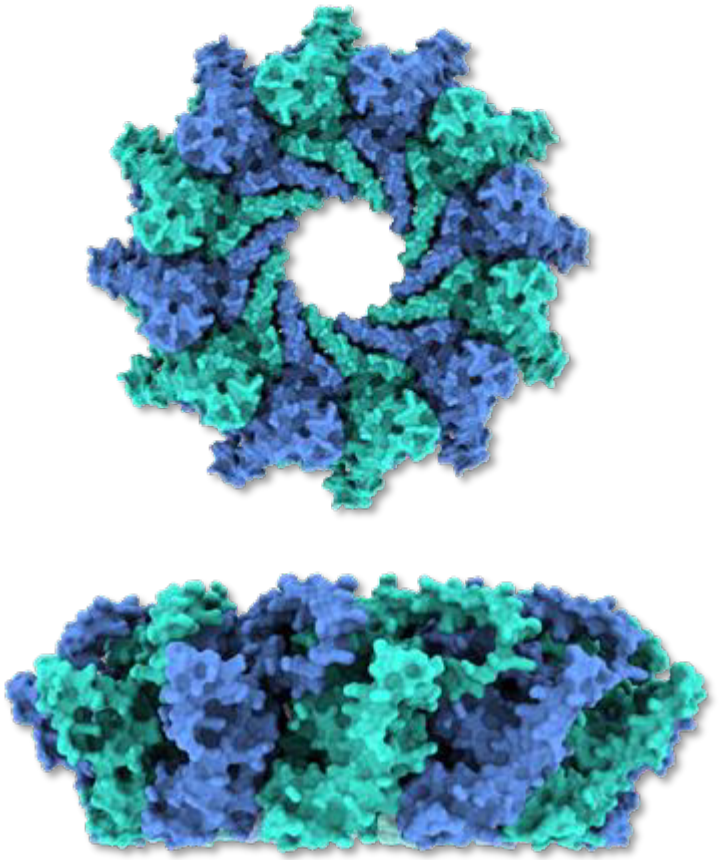
Interaction with the membrane ?



Real Life Demo

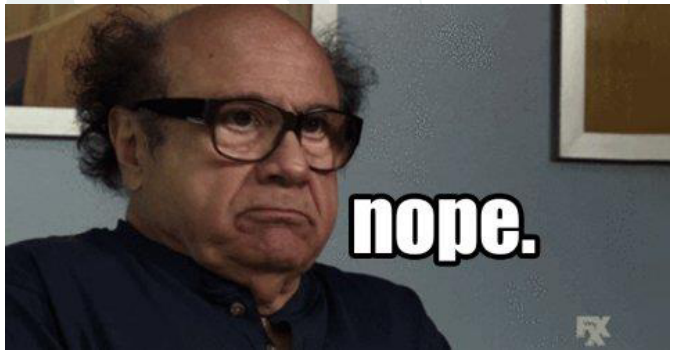
→ Let's try with something… smaller!
After all, a dodecamer is 6 dimers right? And a dimer is easier to predict!



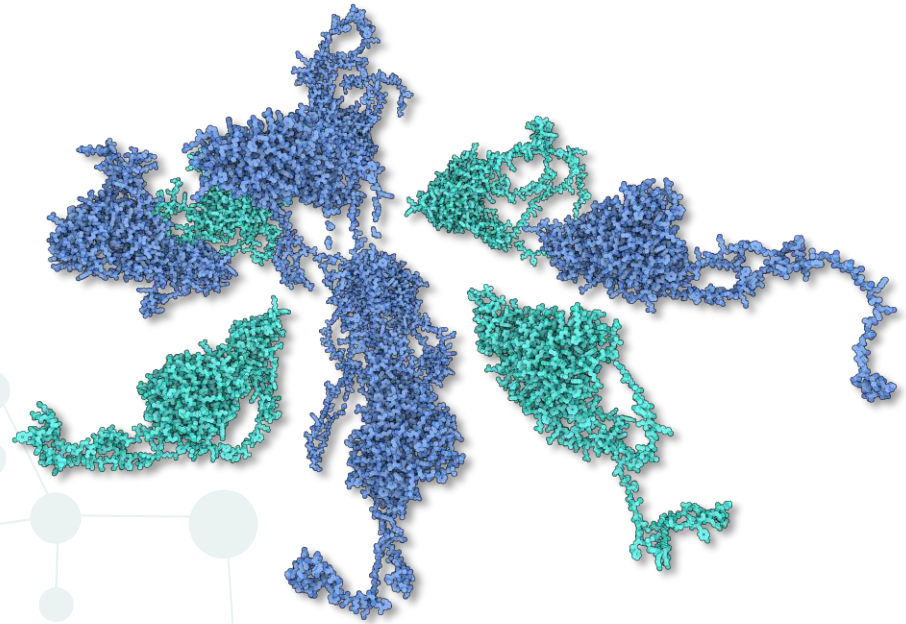


Real Life Demo

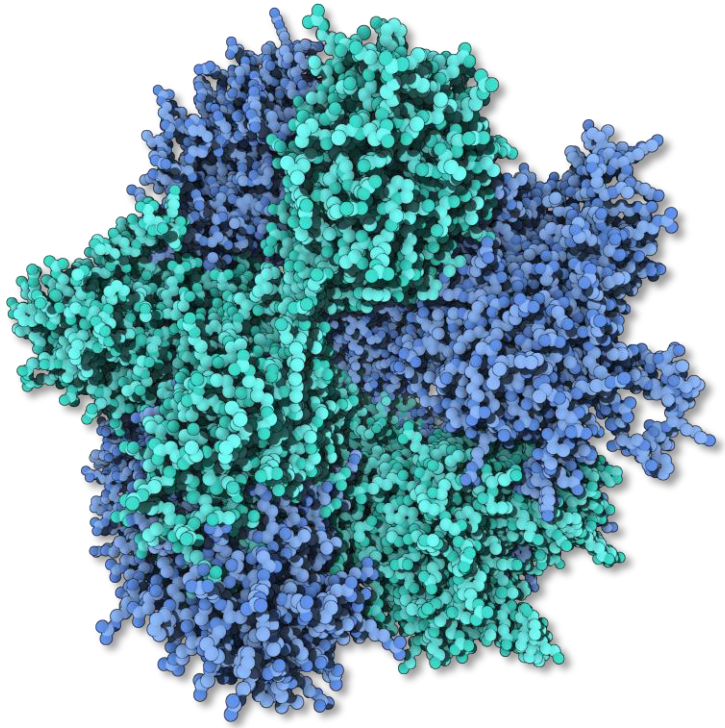
Could our protein be, in fact, dodecameric?
Let's try with AlphaFold!



*AlphaFold 2.2 - 10mer**



AlphaFold 2.3 - 12mer



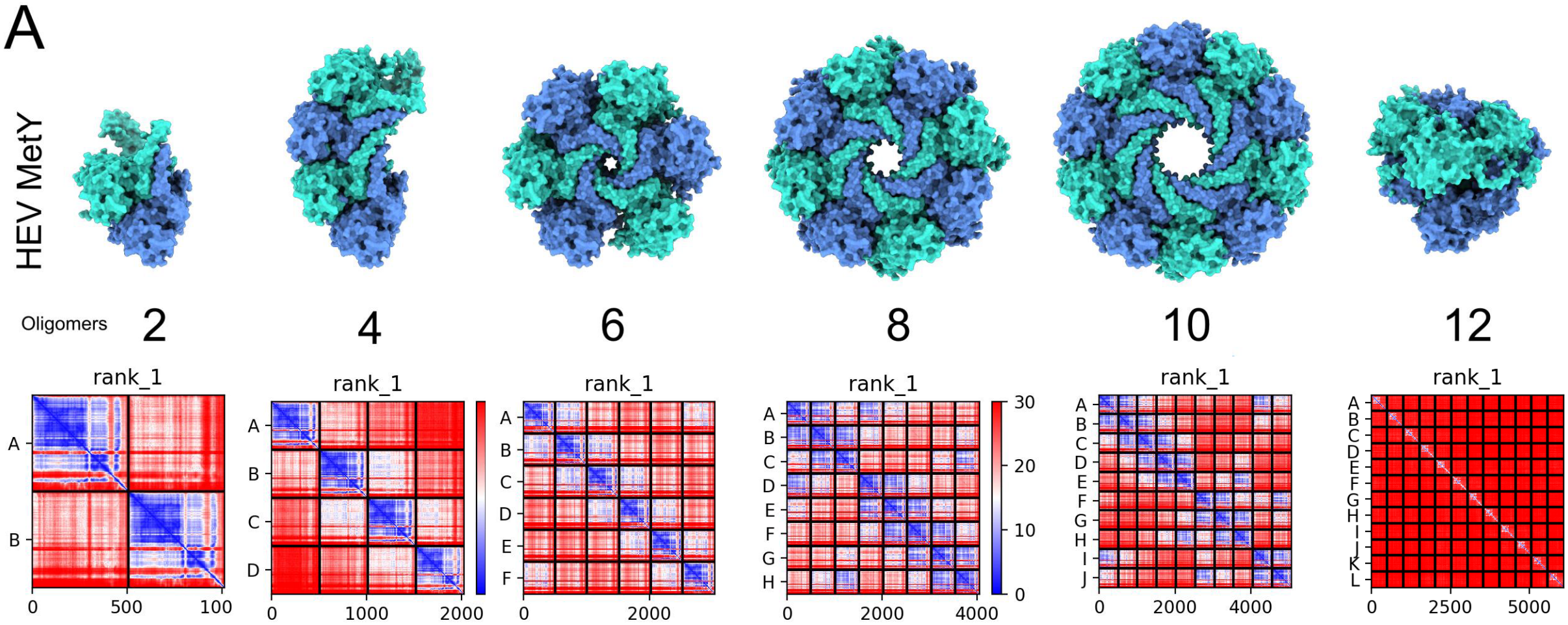
With large systems, the prediction decreases in accuracy leading to bad contact prediction as well as bad local folding...

- Need a tremendous amount of memory
- Very long computing time

Real Life Demo

Could our protein be, in fact, dodecameric?

Let’s try with AlphaFold!

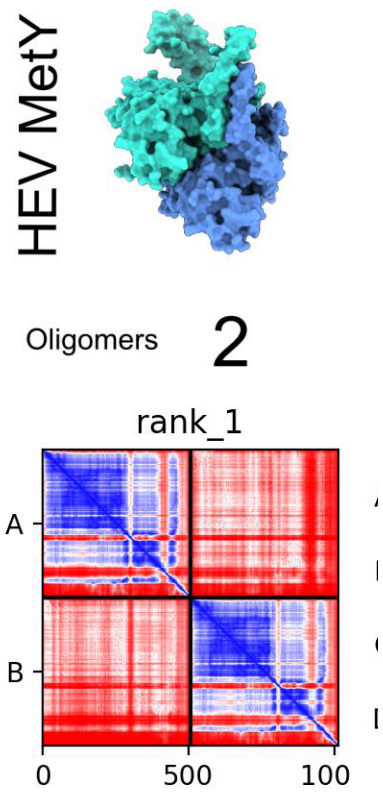


Real Life Demo

Could our protein be, in fact, dodecameric?

Let’s try with AlphaFold!

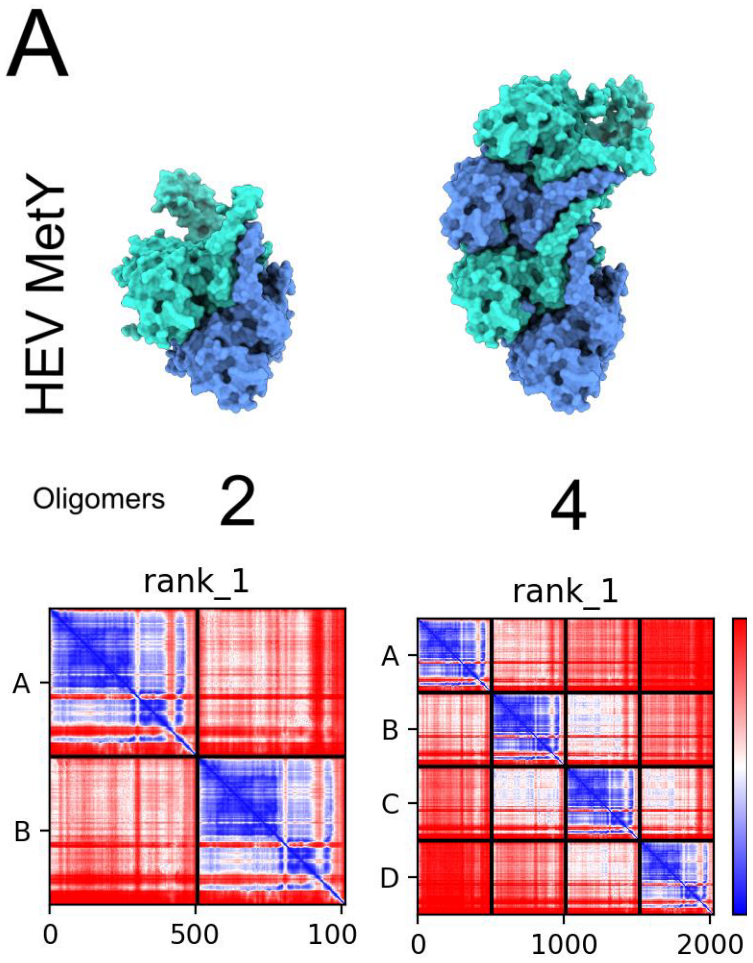
A



Real Life Demo

Could our protein be, in fact, dodecameric?

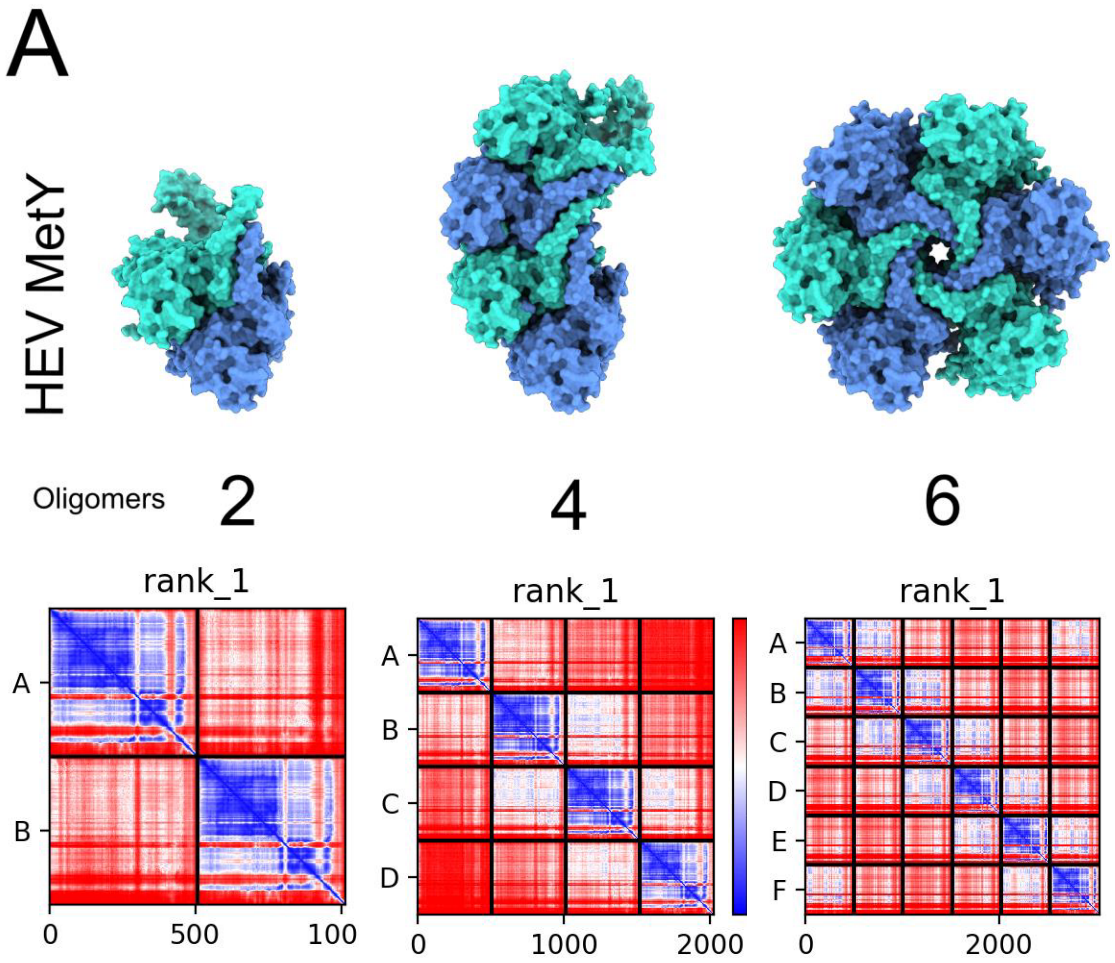
Let’s try with AlphaFold!



Real Life Demo

Could our protein be, in fact, dodecameric?

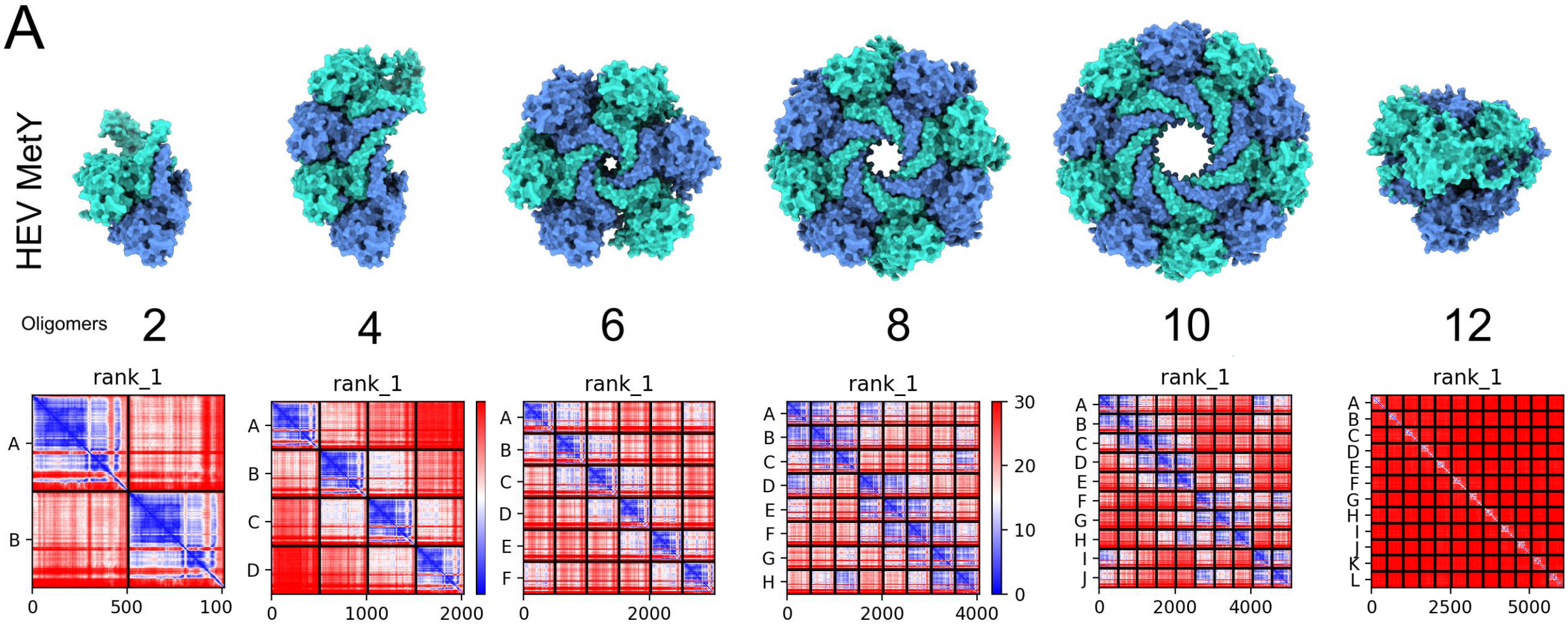
Let’s try with AlphaFold!



Real Life Demo

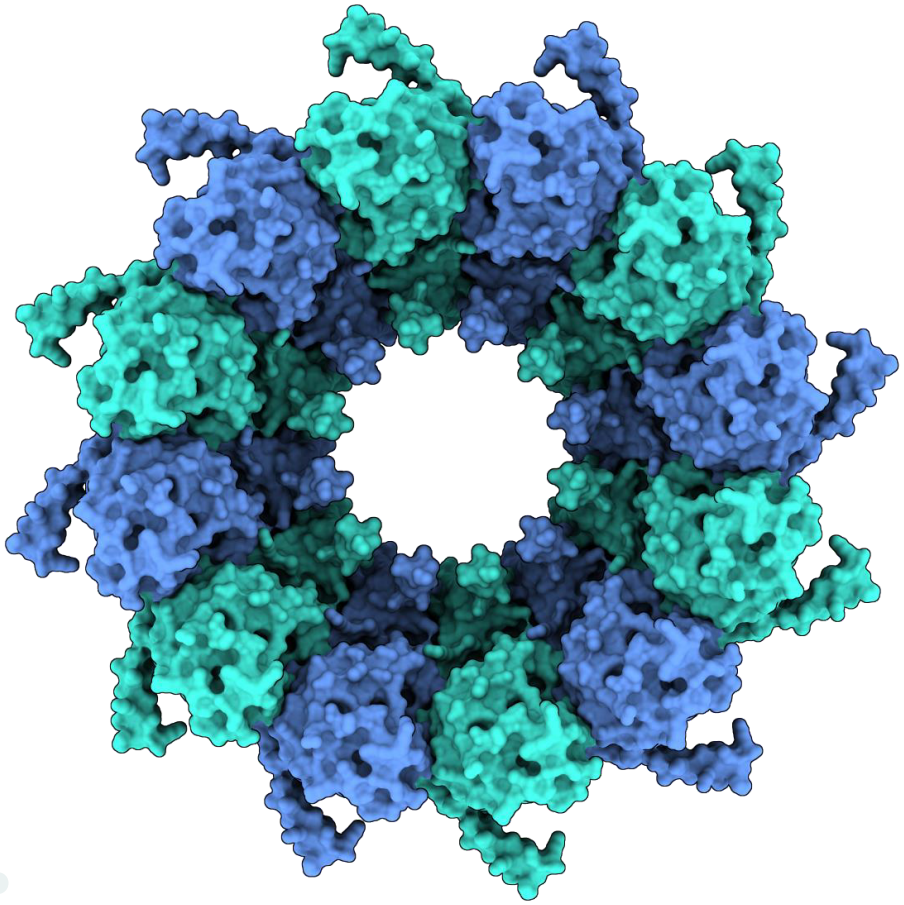
Could our protein be, in fact, dodecameric?

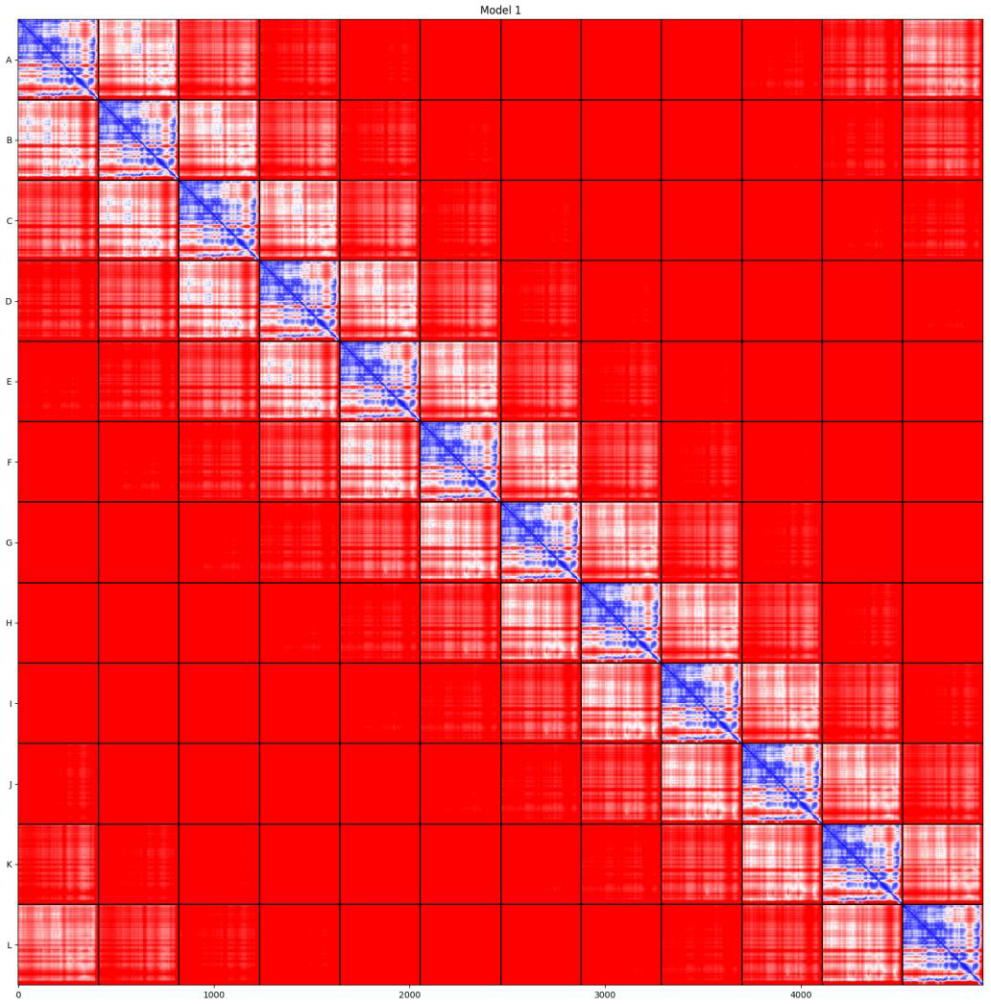
Let’s try with AlphaFold!



Real Life Demo

And AlphaFold 3???

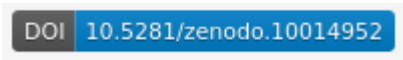




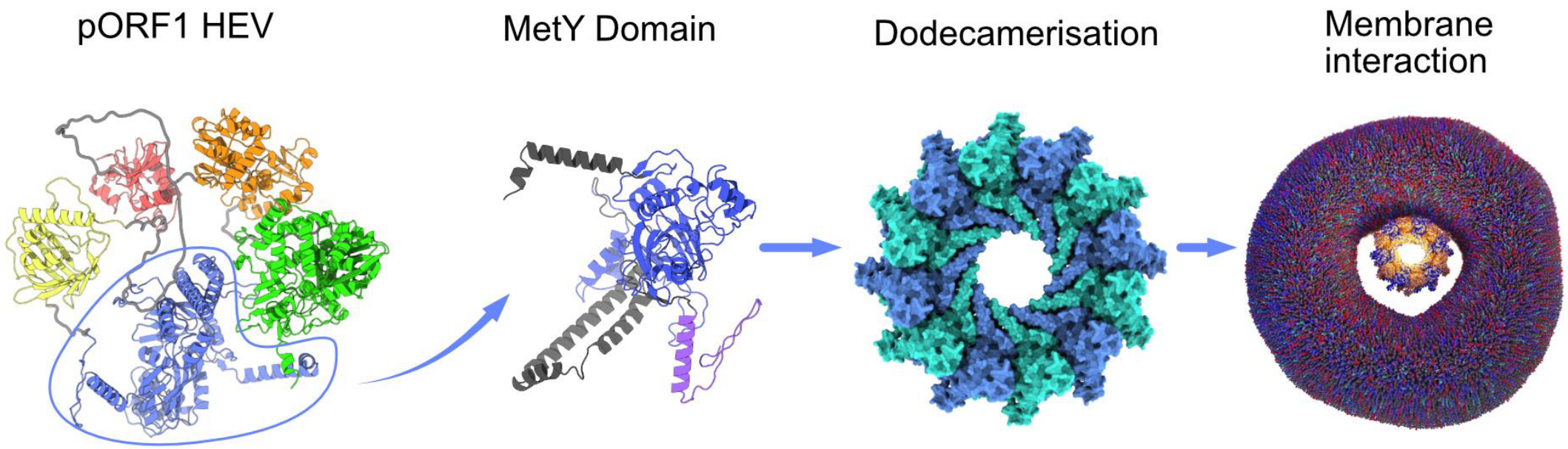
Real Life Demo

→ We already ran AlphaFold2 for you on the Nter dimer of pORF1 => that's what we're going to be working on today

AlphaFold@I2BC page: https://bioi2.i2bc.paris-saclay.fr/django/alphafold/ORF1_Nter_dimer/summary/

We also deposited the results on Zenodo: https://doi.org/10.5281/zenodo.10014952 

End of the story…



In our case, we

1. used Alphafold to generate a model
2. databases to find homologues
3. made a oligomerisation hypothesis base on structural homologues
4. We study now its membrane interaction.