The Protein Data Bank (in Europe)

An introduction to protein structure and the PDB

Genevieve L. Evans

Structural Biologist
& Scientific Database
Curator

PDBe
Protein Data Bank in Europe

pdbhelp@ebi.ac.uk
proteindatabank
@PDBeurope
proteindatabank
pdbeurope
pdbart

EMBL-EBI

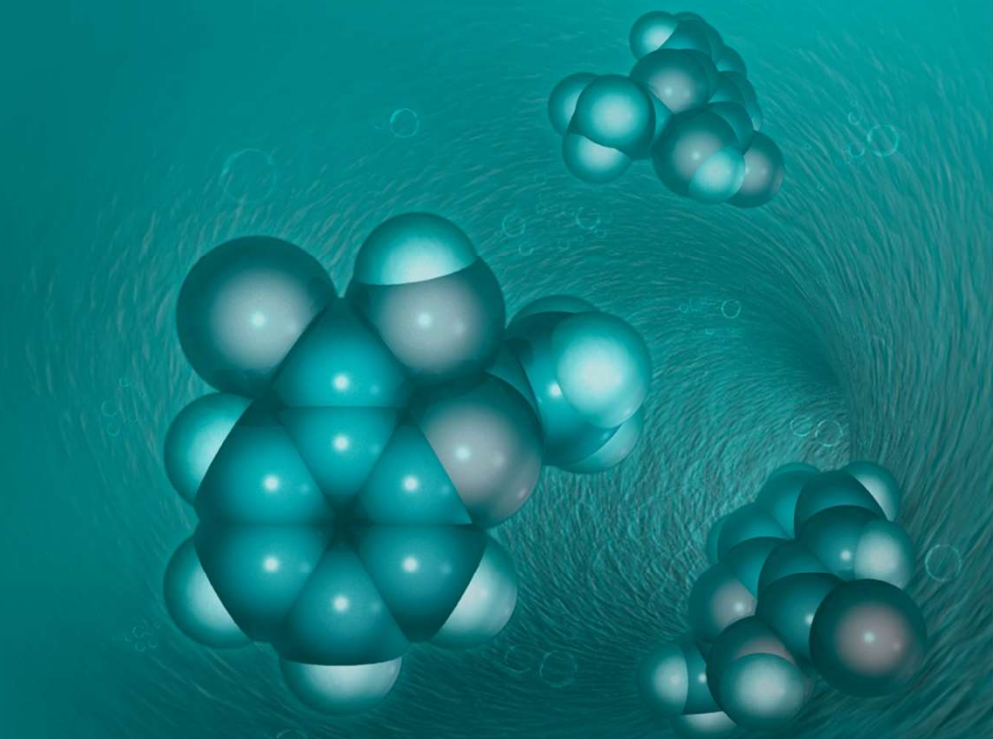
What is the PDB?

https://www.menti.com/alpvu5wen5e4

Join code: 8419 3777

PDBe

Protein Data Bank in Europe



EMBL-EBI

More about PDBe tools and services

https://www.menti.com/alpvu5wen5e4

Join code: 8419 3777

Go to PDBe.org/training > tutorials

- Links to EBI train online tutorials for PDBe services
  - From basic introductory tutorials to advanced tools like PDBeFold
  - https://www.ebi.ac.uk/pdbe/online-tutorials

- Webinars available on our YouTube channel
  - Walk-throughs of specific features on PDBe and PDBe-KB (KnowledgeBase) webpages
  - https://www.youtube.com/@ProteinDataBank

API - Application Programming Interface

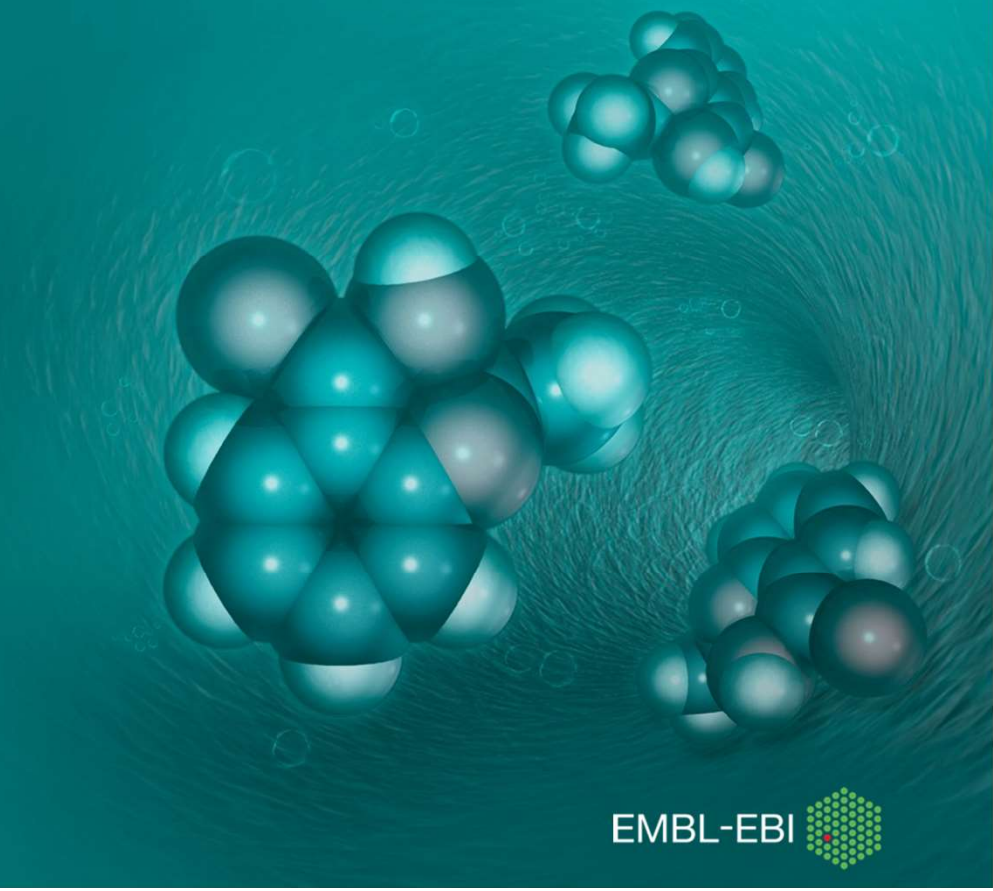
https://www.menti.com/alpvu5wen5e4

Join code: 8419 3777

- Programmatic access webinar series
  - Six-part series ranging from basics in PDBe programmatic access to advanced visualisation
  - Visit pdbeurope.github.io/api-webinars for videos and tutorials

- API 'start page'
  - Visit: https://www.ebi.ac.uk/pdbe/pdbe-rest-api

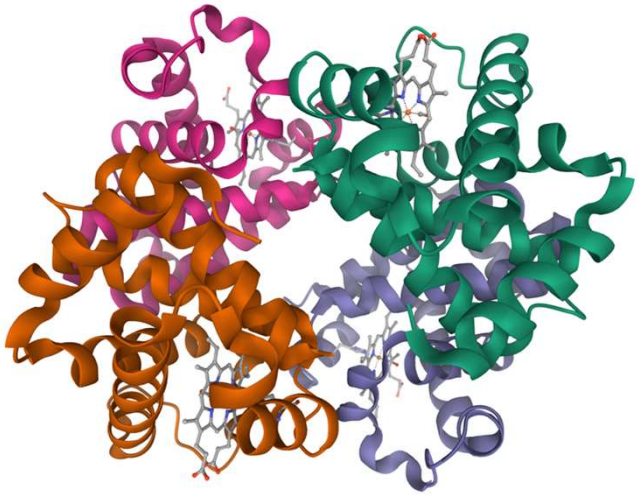
What is the PDB?



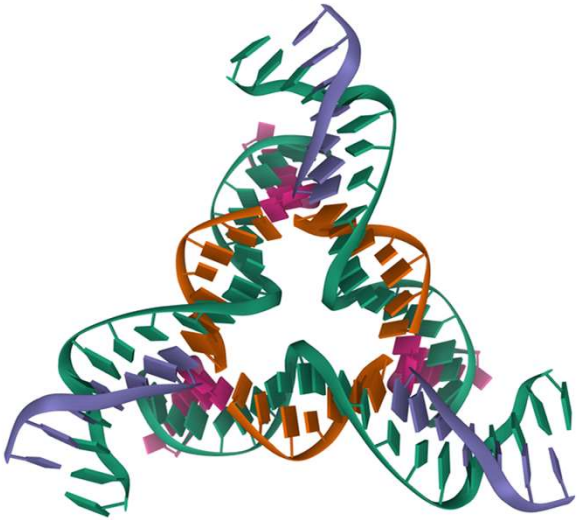
PDBe
Protein Data Bank in Europe

EMBL-EBI

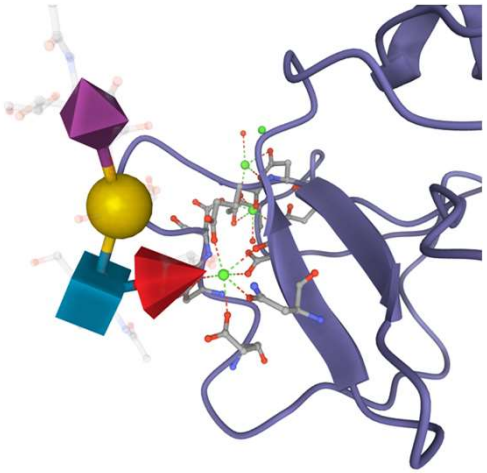
The Protein Data Bank (PDB) is an archive of experimentally determined 3-dimensional structures of biological macromolecules



Proteins

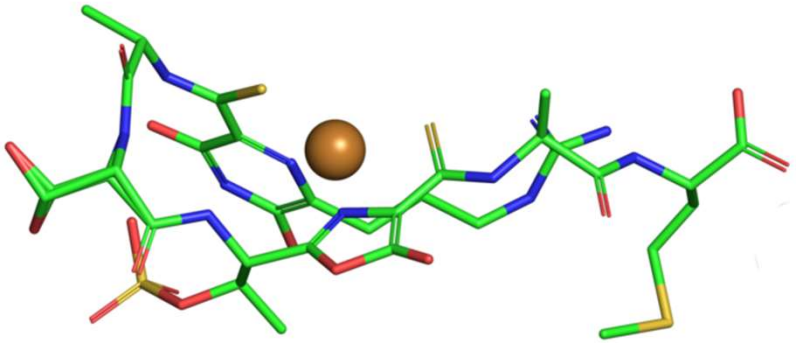


DNA/RNA

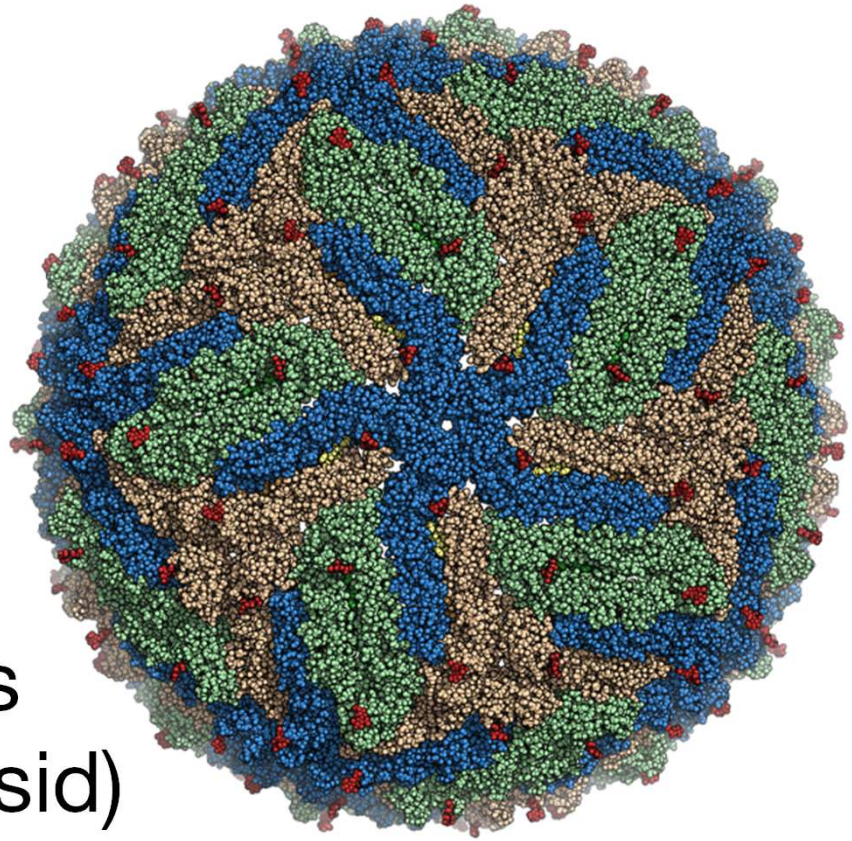


Bound ligands

Varying sizes of structure - can be very large!

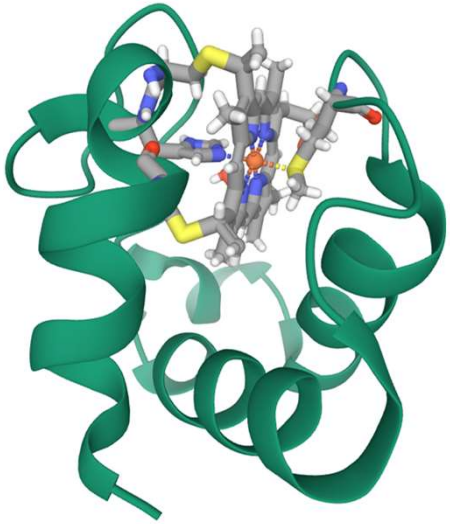


Copper scavenger
<1kDa



Zika virus
4 MDa (capsid)

A “PDB code” refers to a structure



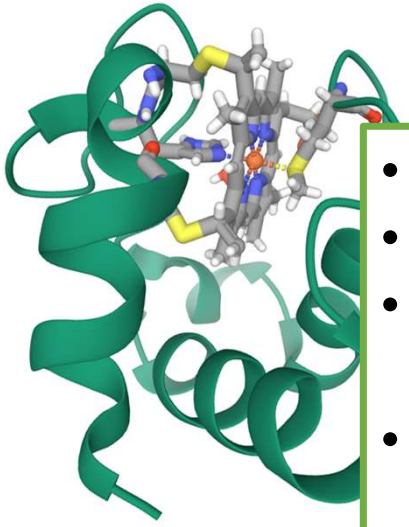
Deposit in the PDB

PDB code

Described in a paper (or maybe not published)

Referenced in the paper

A “PDB code” refers to a structure



Deposit in the PDB

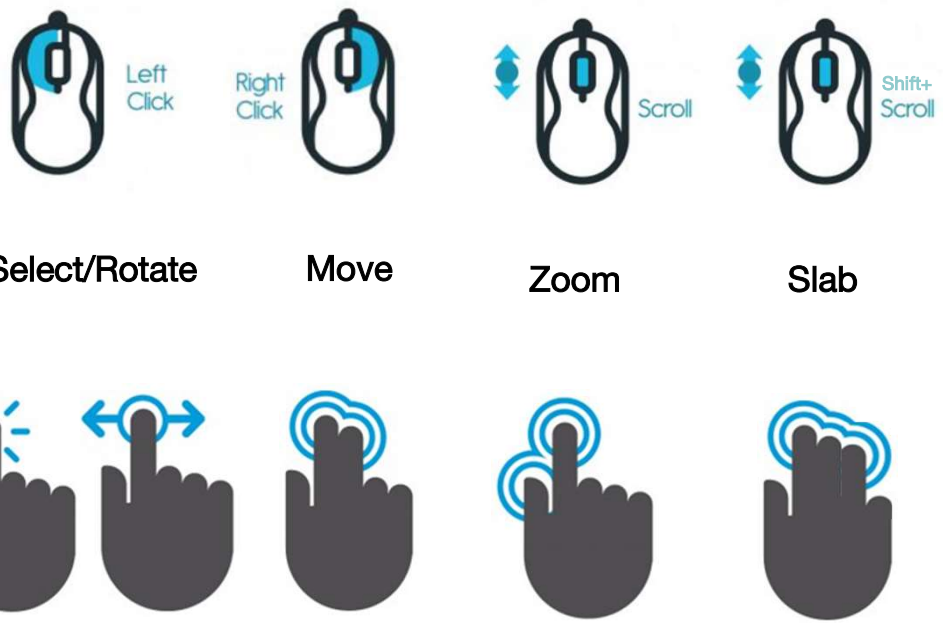
PDB cod

- Unique code, currently 4 characters
- Identifies the data within the PDB archive
- Always starts with a number

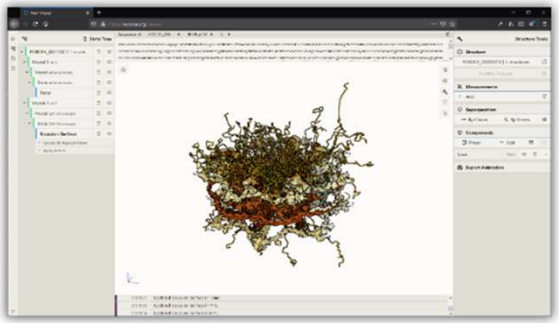
- *e.g.* 2ins, 4xyz, 2f48
  → pdb_00002ins, pdb_00004xyz, pdb_00002f48

  More information:
  https://www.wwpdb.org/documentation/new-format-for-pdb-ids

Mol* website - molstar.org



Mol* (*/'molstar/*) is a modern web-based open-source toolkit for visualisation and analysis of large-scale molecular data



Open Mol* Viewer

Viewer Documentation

Issues & Feedback

High-performance graphics and data handling of the Mol* Viewer allow users to simultaneously visualise up to hundreds of (superimposed) protein structures, play molecular dynamics trajectories, render cell-level models at atomic detail with tens of millions of atoms, or display huge models obtained by I/HM such as the Nuclear Pore Complex.

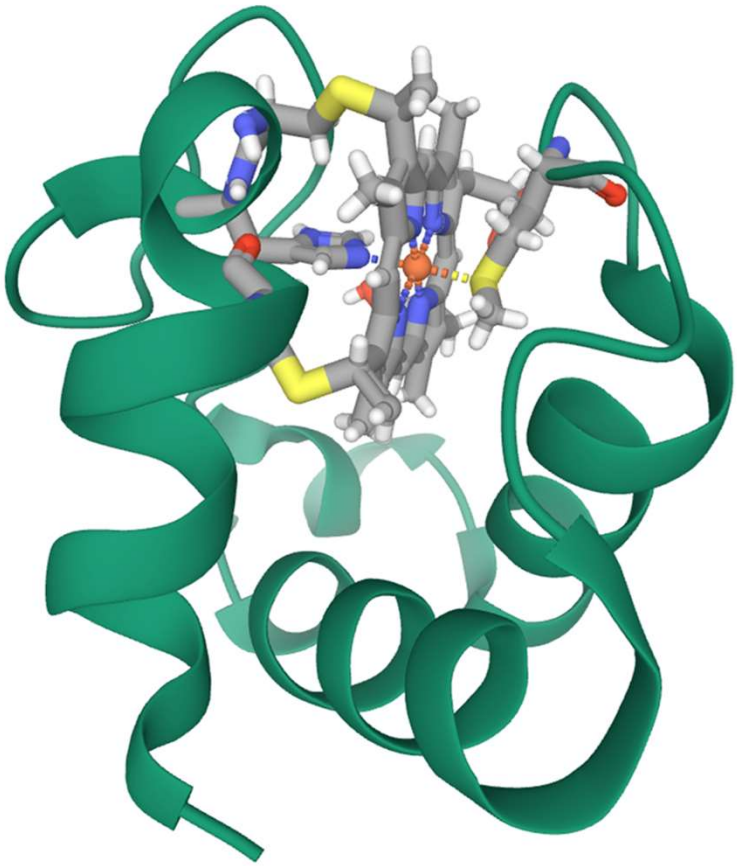
Ligands

- In the PDB:
  drugs, metal ions, cofactors, etc
  = ligands
  = small molecules binding to protein

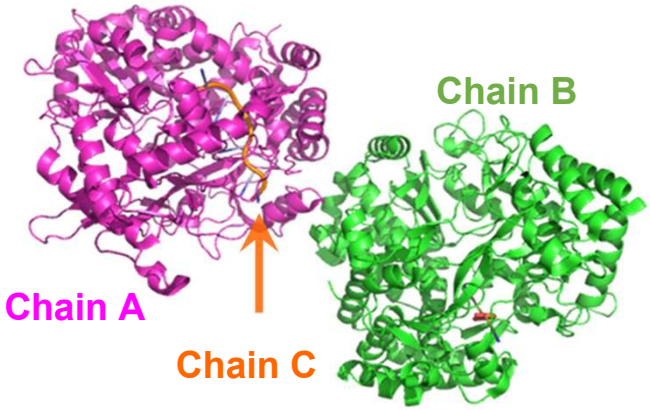
- Each unique ligand has its own ID code
  - Currently 3 characters, *e.g.* AIN (aspirin)

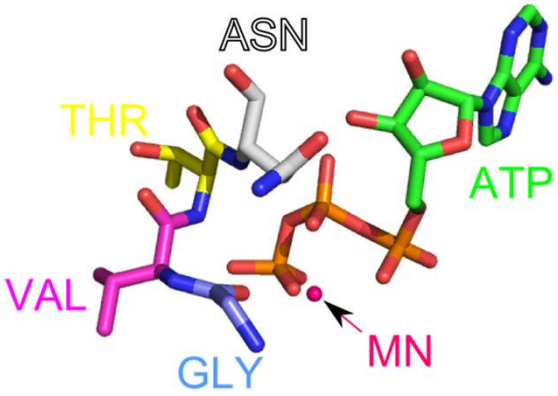
    → IDs with 5 characters will exist in future
  - also called: Chemical component (CC) id
  - Helps find it in the archive

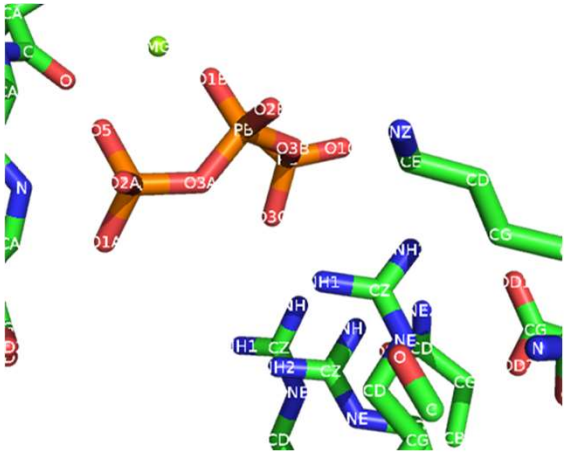
- **PDBeChem** web pages contain ligand-specific info:
  https://www.ebi.ac.uk/pdbe-srv/pdbechem/



A hierarchy in each PDB entry

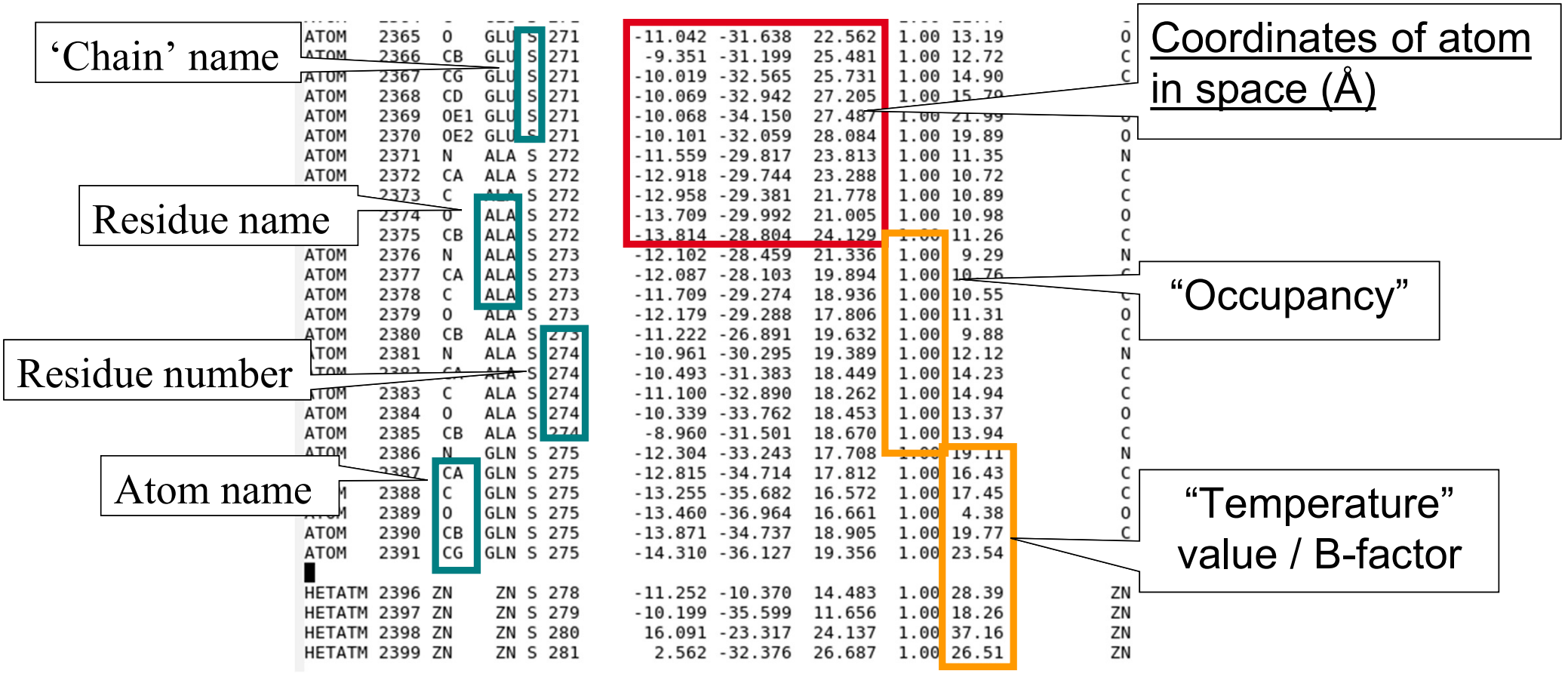






Chains	Residues	Atoms
Macromolecules are referred to as '**chains**'	Chains have **residues** with three letter code for chemical type (but some are not 3 letters!) and unique number for where it is in the chain	Residues have **atoms** Each has defined position in 3D space

What does a PDB file look like?
A text file with fixed column width - Card legacy



mmCIF is the ‘master format’
It’s still a text file!
More (modern) computer readable

The fastest open-source mmCIF parser:
GEMMI (https://gemmi.readthedocs.io/en/latest/)

```
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.Cartn_x_esd
_atom_site.Cartn_y_esd
_atom_site.Cartn_z_esd
_atom_site.occupancy_esd
_atom_site.B_iso_or_equiv_esd
_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    . MET A 1 1 ? 4.940   -11.534 16.748 1.00 0.00 ? ? ? ? ? ? 1 MET A N    1
ATOM 2    C CA   . MET A 1 1 ? 4.627   -10.082 16.703 1.00 0.00 ? ? ? ? ? ? 1 MET A CA   1
ATOM 3    C C    . MET A 1 1 ? 4.375   -9.616  15.273 1.00 0.00 ? ? ? ? ? ? 1 MET A C    1
ATOM 4    O O    . MET A 1 1 ? 4.604   -10.360 14.318 1.00 0.00 ? ? ? ? ? ? 1 MET A O    1
ATOM 5    C CB   . MET A 1 1 ? 5.800   -9.312  17.310 1.00 0.00 ? ? ? ? ? ? 1 MET A CB   1
ATOM 6    C CG   . MET A 1 1 ? 7.097   -9.468  16.535 1.00 0.00 ? ? ? ? ? ? 1 MET A CG   1
ATOM 7    S SD   . MET A 1 1 ? 8.437   -8.471  17.215 1.00 0.00 ? ? ? ? ? ? 1 MET A SD   1
ATOM 8    C CE   . MET A 1 1 ? 8.819   -9.388  18.706 1.00 0.00 ? ? ? ? ? ? 1 MET A CE   1
ATOM 9    H H1   . MET A 1 1 ? 5.323   -11.746 17.691 1.00 0.00 ? ? ? ? ? ? 1 MET A H1   1
ATOM 10   H H2   . MET A 1 1 ? 5.641   -11.730 16.006 1.00 0.00 ? ? ? ? ? ? 1 MET A H2   1
ATOM 11   H H3   . MET A 1 1 ? 4.056   -12.055 16.578 1.00 0.00 ? ? ? ? ? ? 1 MET A H3   1
ATOM 12   H HA   . MET A 1 1 ? 3.740   -9.904  17.292 1.00 0.00 ? ? ? ? ? ? 1 MET A HA   1
ATOM 13   H HB2  . MET A 1 1 ? 5.548   -8.261  17.340 1.00 0.00 ? ? ? ? ? ? 1 MET A HB2  1
```

Residue name

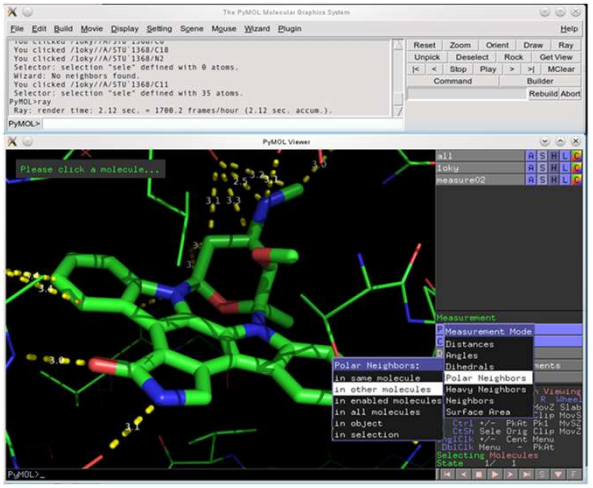
Coordinates of atom in space (Å)

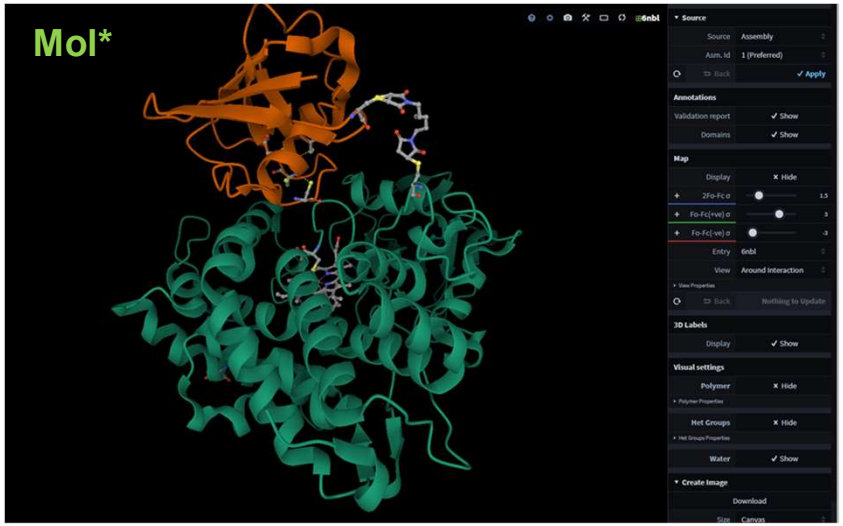
“Occupancy”

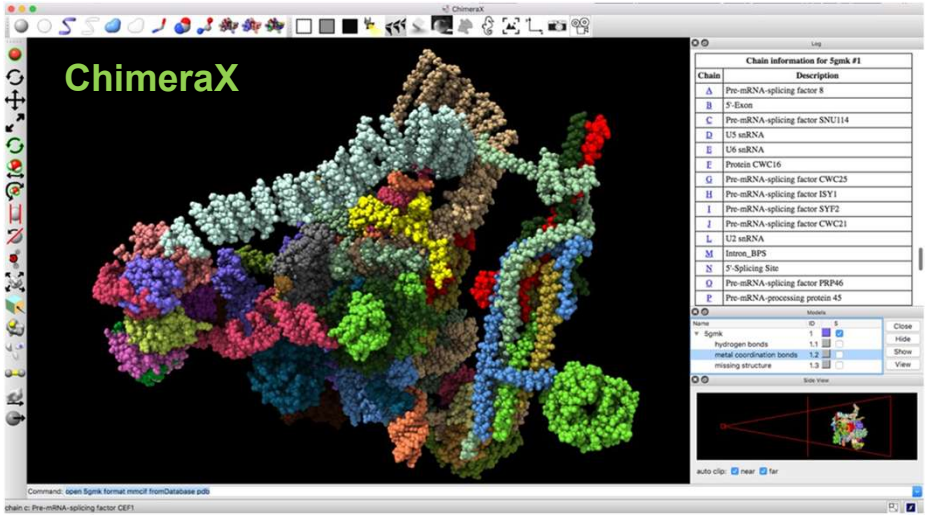
How to view mmCIF or PDB file in 3D

- Different molecular graphics programmes
- Display the atoms in 3D space and interpret the bonds.
- PyMol, ChimeraX, Mol*

Pymol



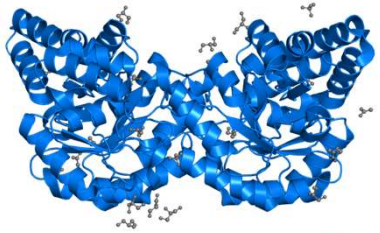




Why have a central archive of structures?

1. Accessibility
   - One-stop shop
   - Searchable

2.1Å
PDB: 3L21

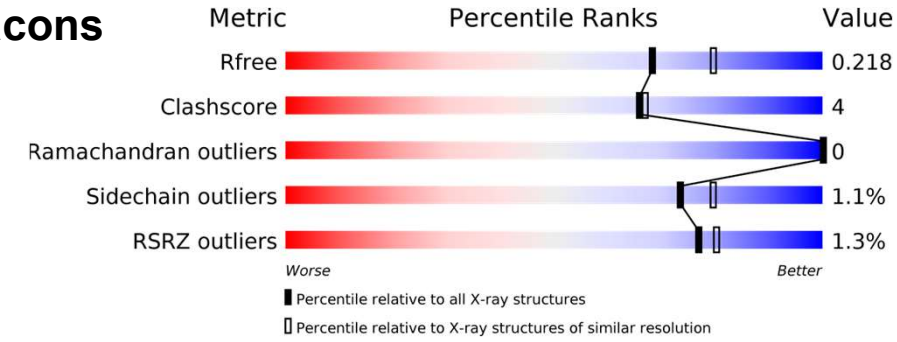


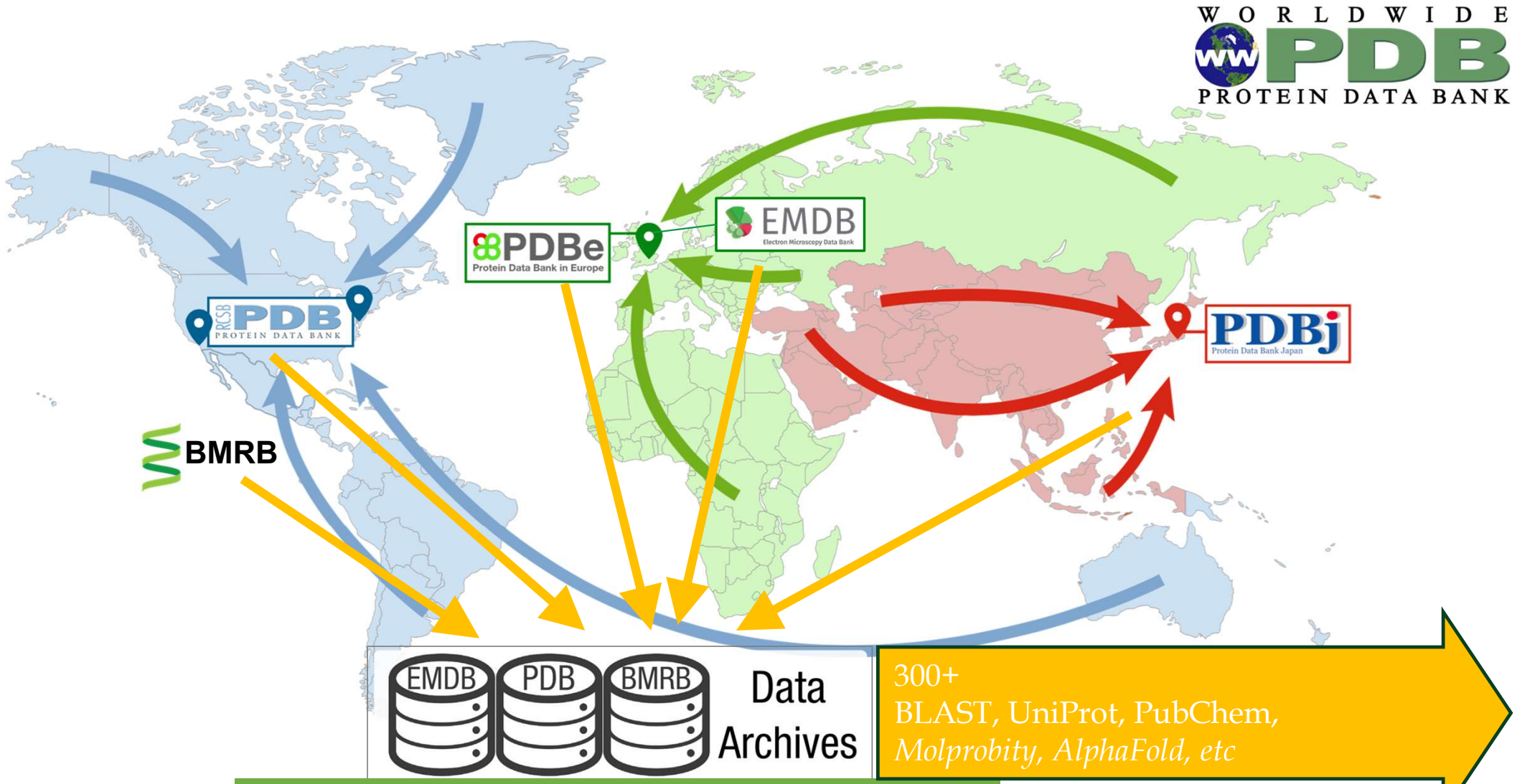
2. Context & connection
   - ‘Structural bioinformatics’
   - Integration with other resources

PDBe-KB
Protein Data Bank in Europe - Knowledge Base

3D Beacons

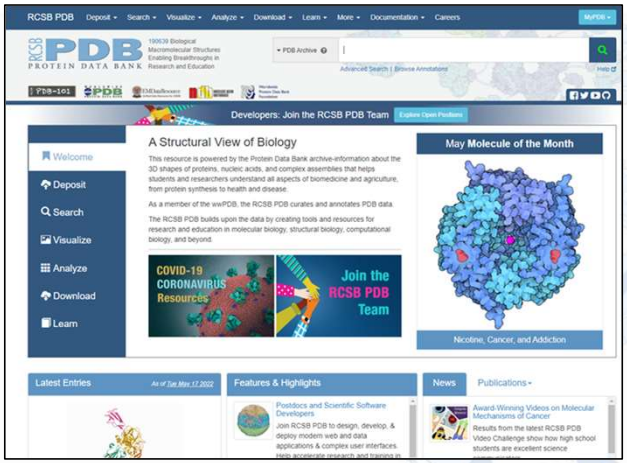
3. Consistency
   - Comparison against all other entries
   - → Validation

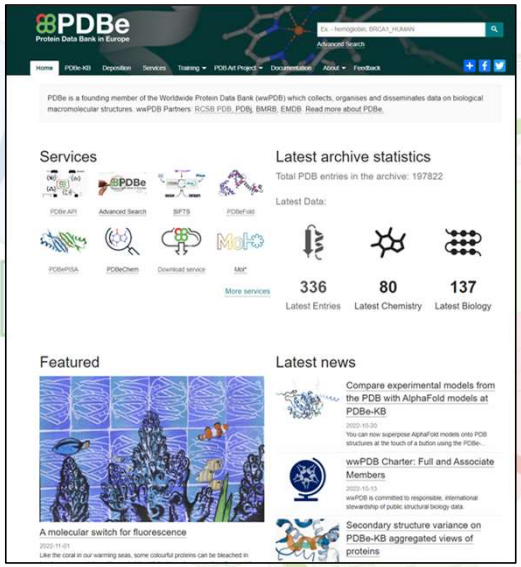


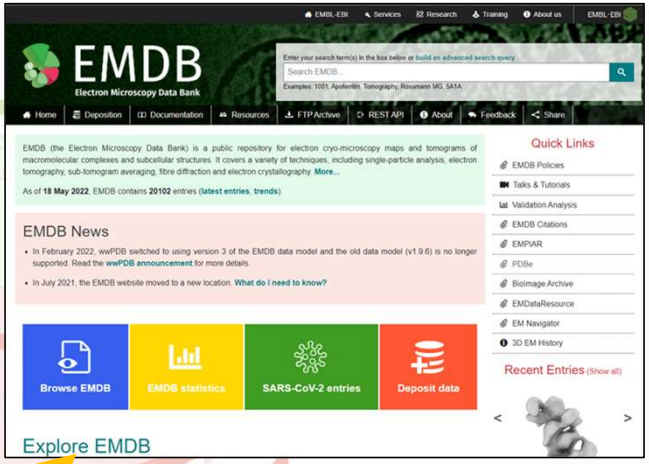


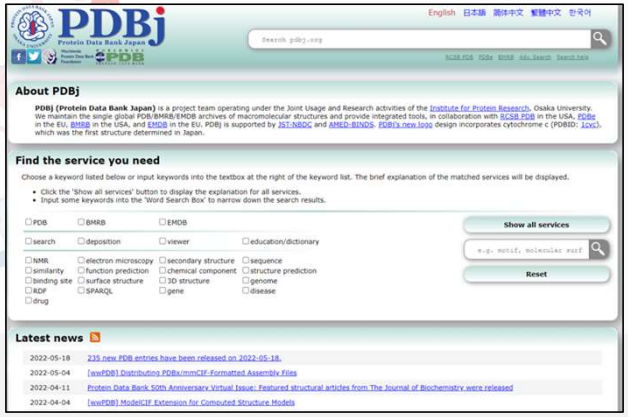
Collect, curate, store and distribute experimentally determined 3D structures of biological macromolecules.

The Worldwide Protein Data Bank (wwPDB)







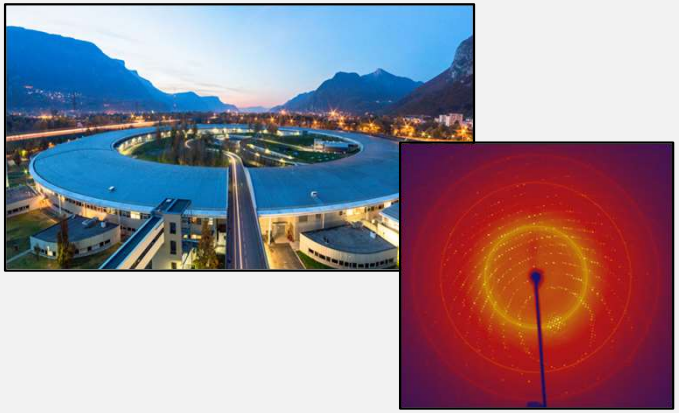


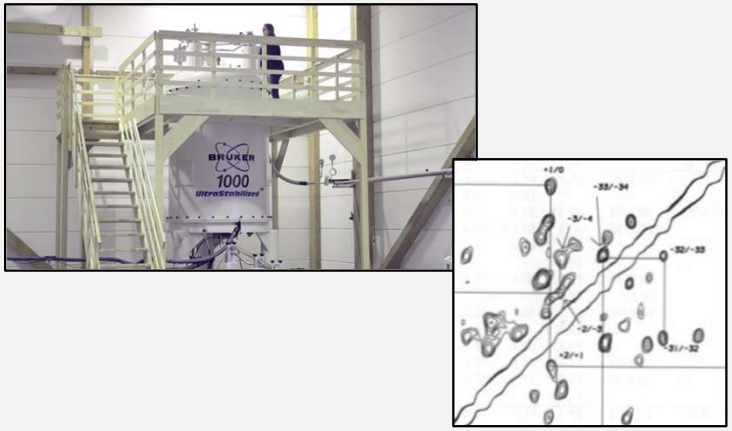
EMBL-EBI

What experimental methods are accepted?

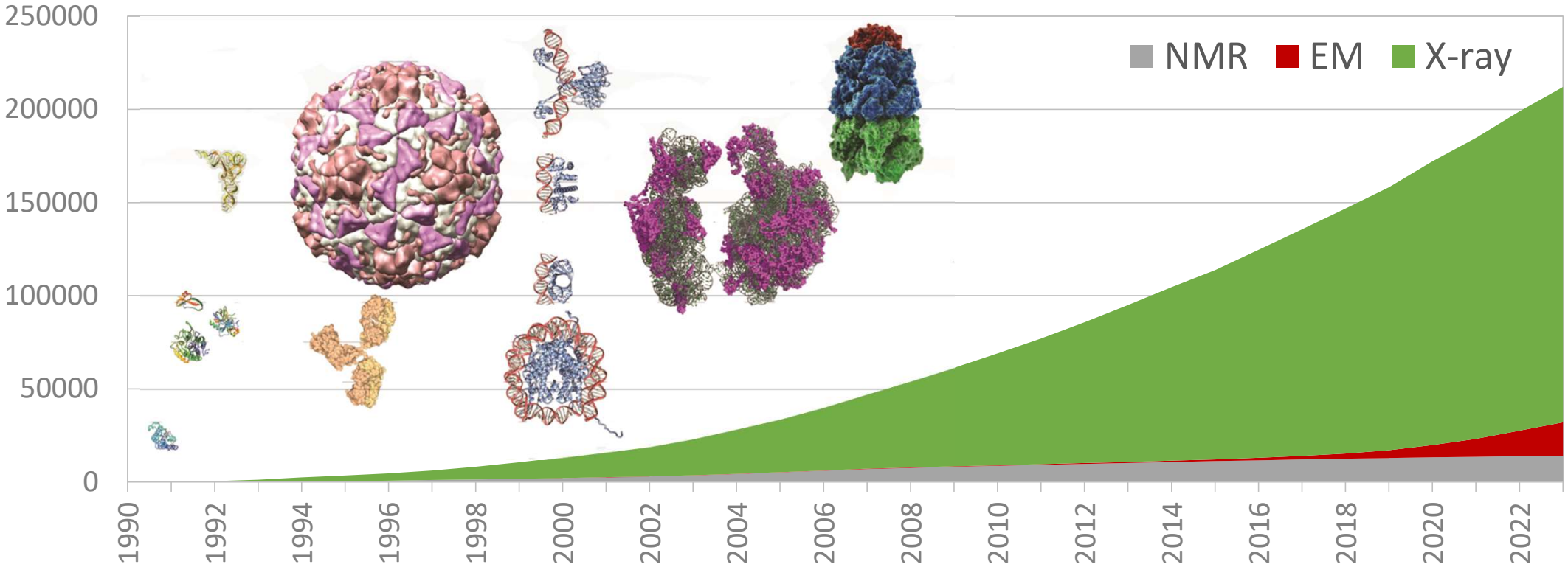
Method	Total	2019	2020	2021	2022	2023
X-ray Crystallography	85%	84%	80%	74%	69%	66%
Solution NMR Spectroscopy	6.7%	3.3%	2.7%	2.9%	2.1%	2.0%
Electron Microscopy	8.4%	13%	17%	24%	29%	32%

~1%: Fibre Diffraction, Neutron Diffraction, Electron Crystallography, Solid state NMR…



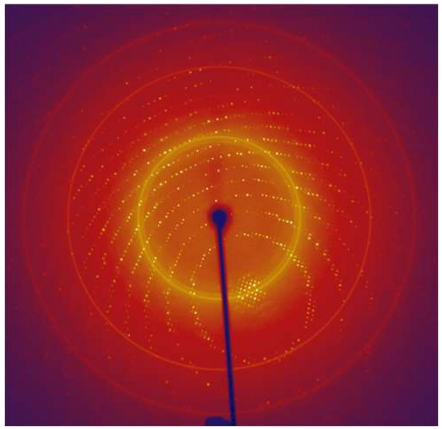


Growth of data in the PDB



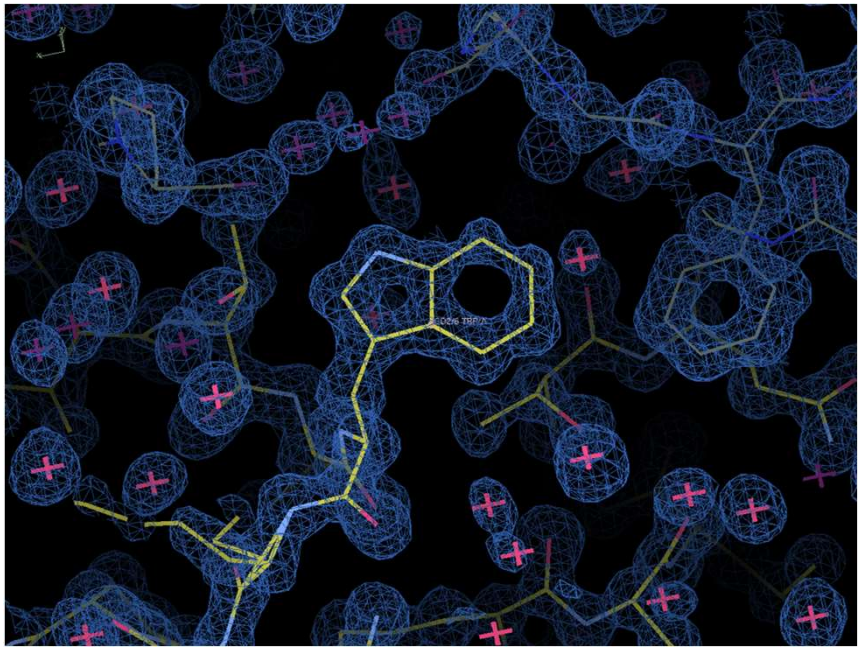
In January 2023, we passed the 200,000 milestone. As of today, ~215,000 structures

X-ray crystallography data



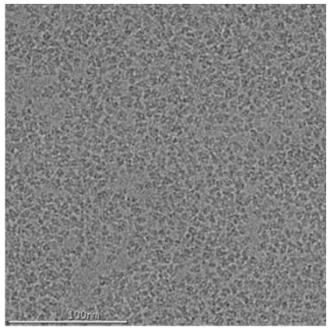
From the diffraction pattern, a map can be calculated.

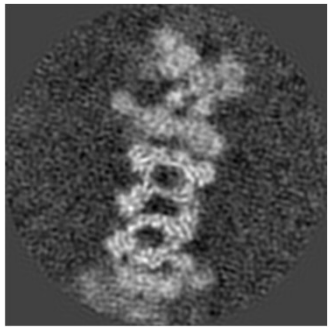
This indicates the location of electrons (therefore atoms) in the crystal
Hence - ‘electron density map’



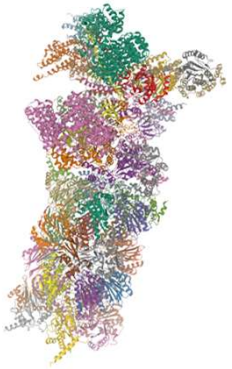
Model is built into this map in an iterative process

Cryo-electron Microscopy (cryoEM)

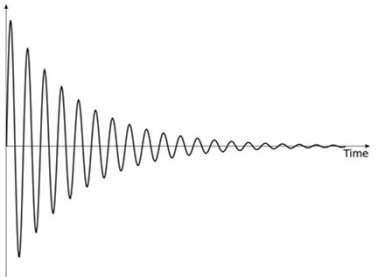


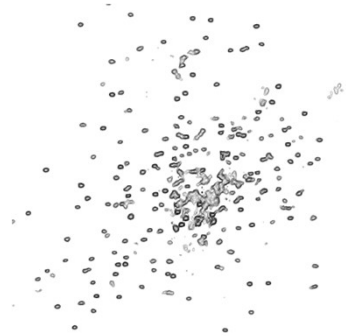


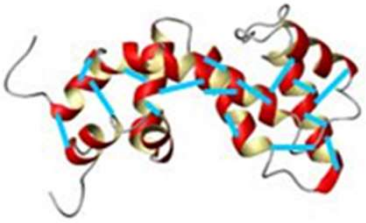
EM volume map

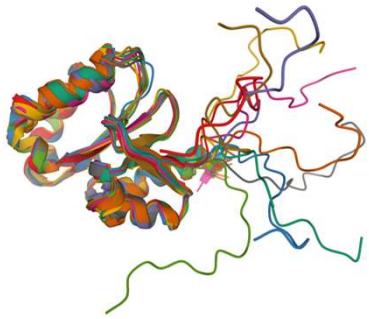


Nuclear Magnetic Resonance (NMR)

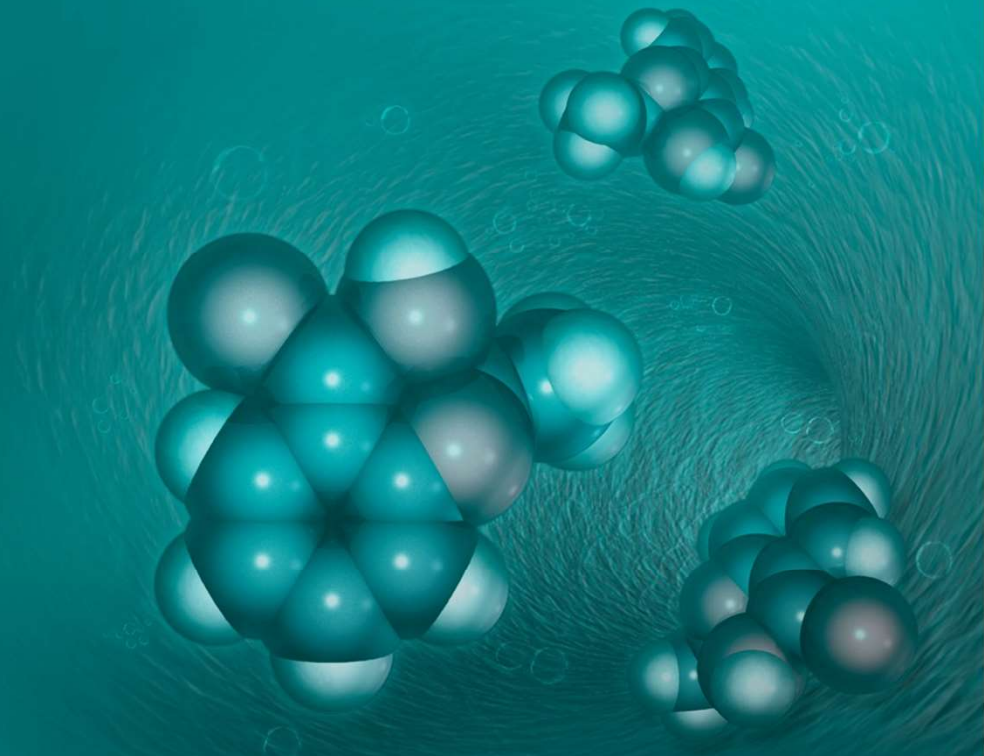








Macromolecules and ligands

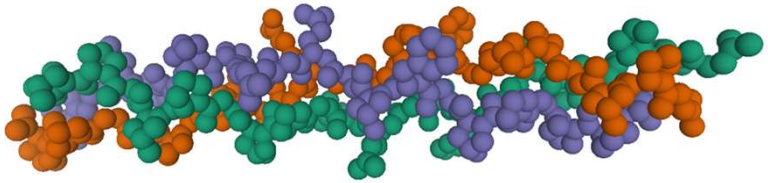


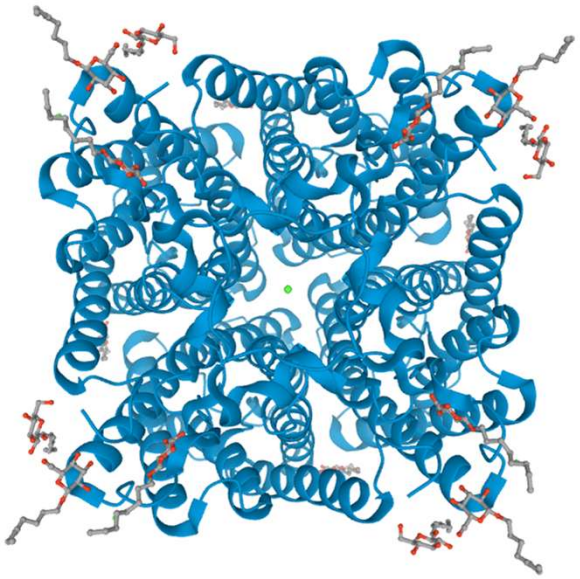
PDBe
Protein Data Bank in Europe

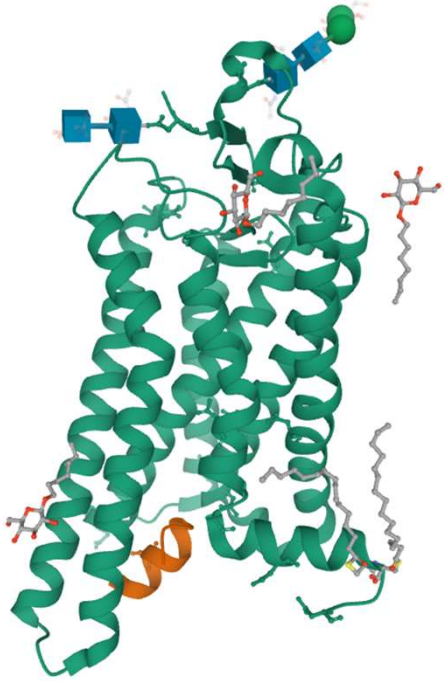
EMBL-EBI

Functions of proteins

- Structural
- Catalysis (i.e. enzymes)
- Receptors
- Channels
- Transport



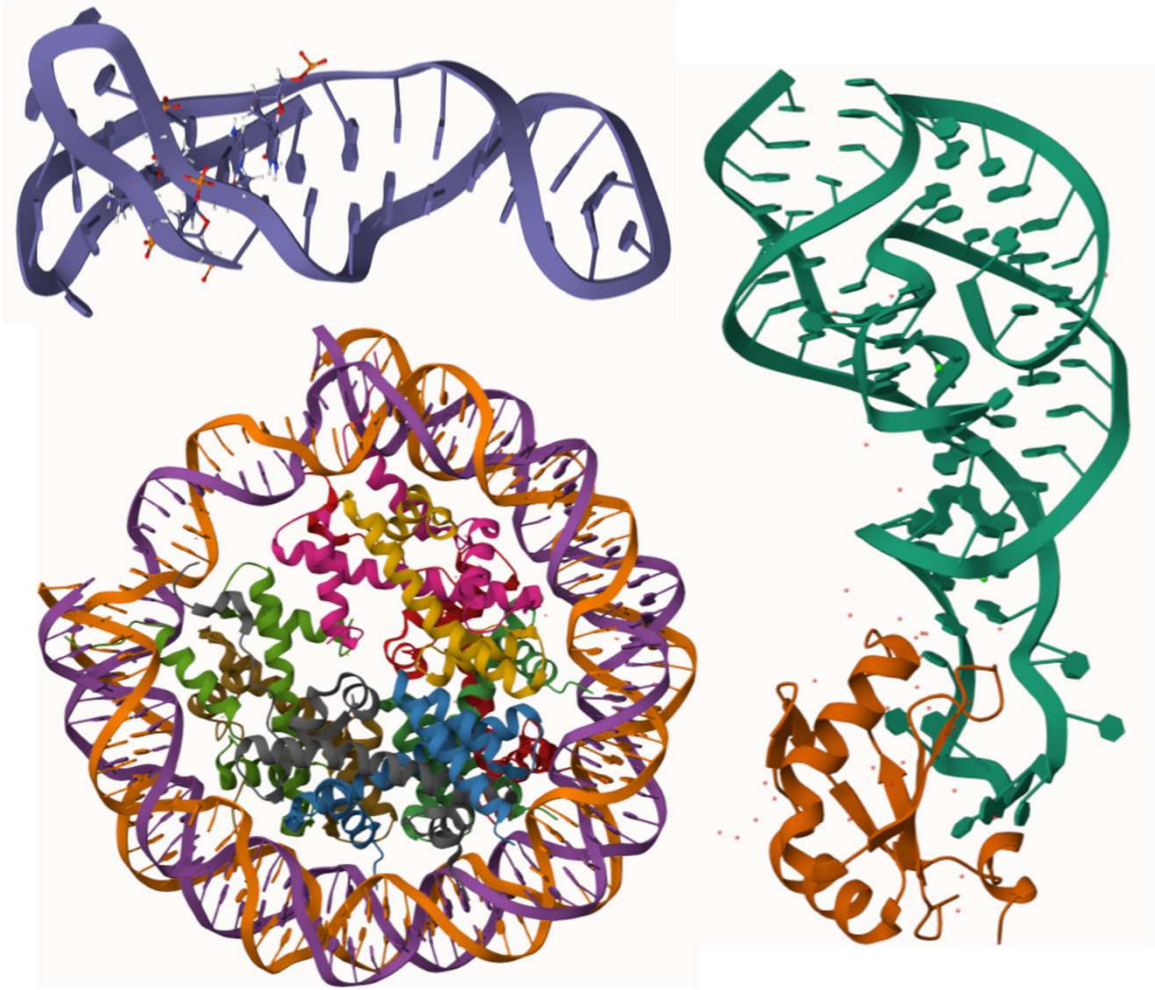




1Q7D, 3DQB, 2W1P

Functions of nucleic acids

- Storage of genetic information
- Catalysis (i.e. ribozymes)
- Protein synthesis
- Regulation of gene expression
- Epigenetic regulation

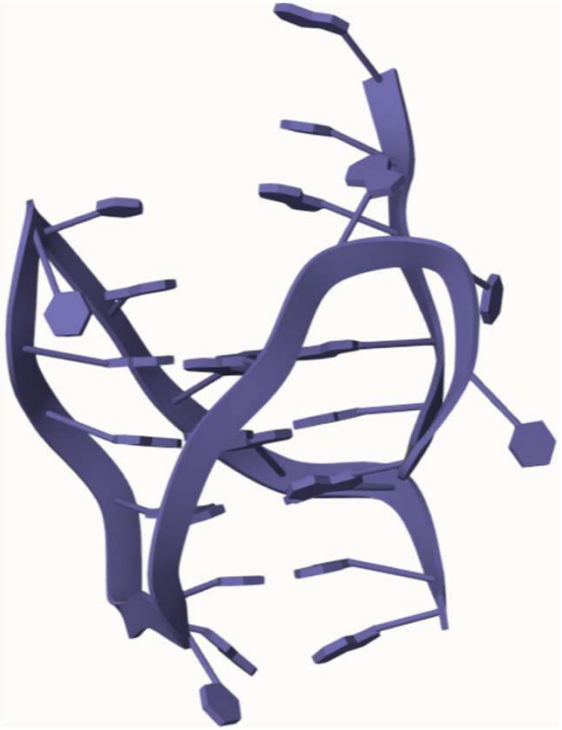


3KWQ, 4PRF, 7O5E

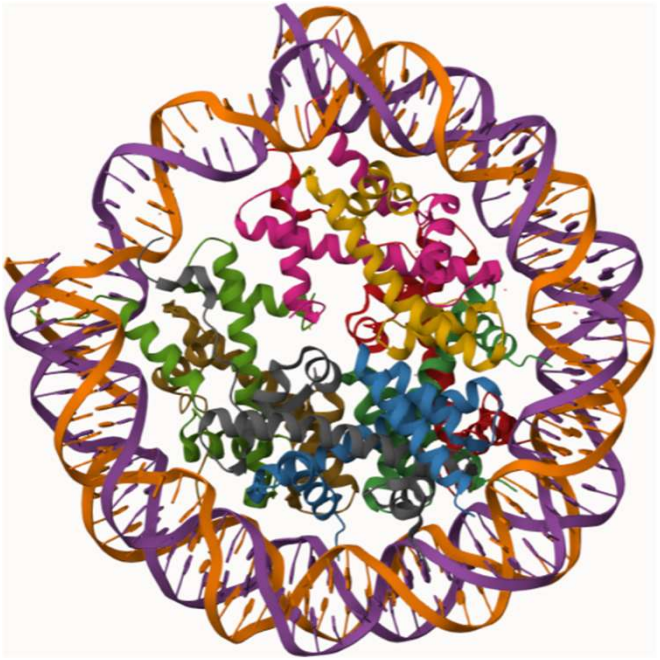
Secondary structure - nucleotides

Single stranded DNA

Double stranded DNA



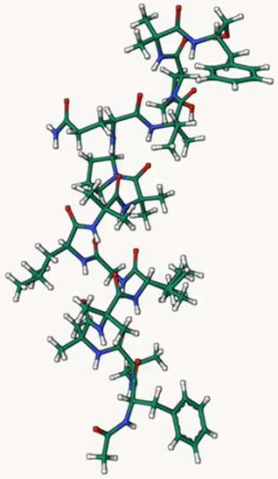
Quadruplex DNA



Nucleosome

Secondary structure - proteins

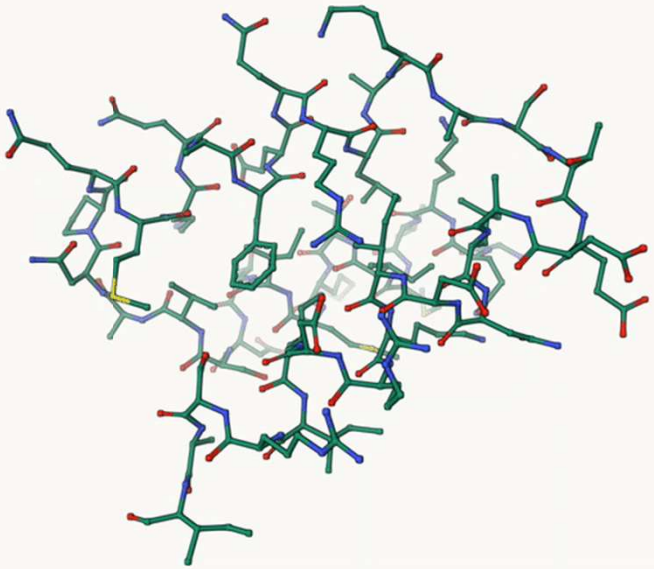
Peptide
(Alpha helix)



Secondary structure - proteins

Peptide
(Alpha helix)

Alpha helices

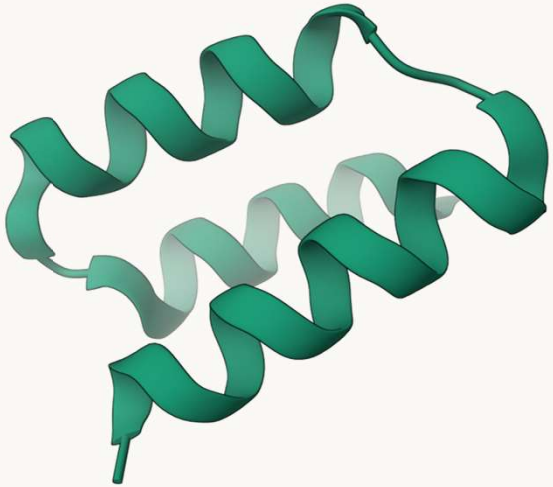


4G13, 3BBZ, 3E6Z

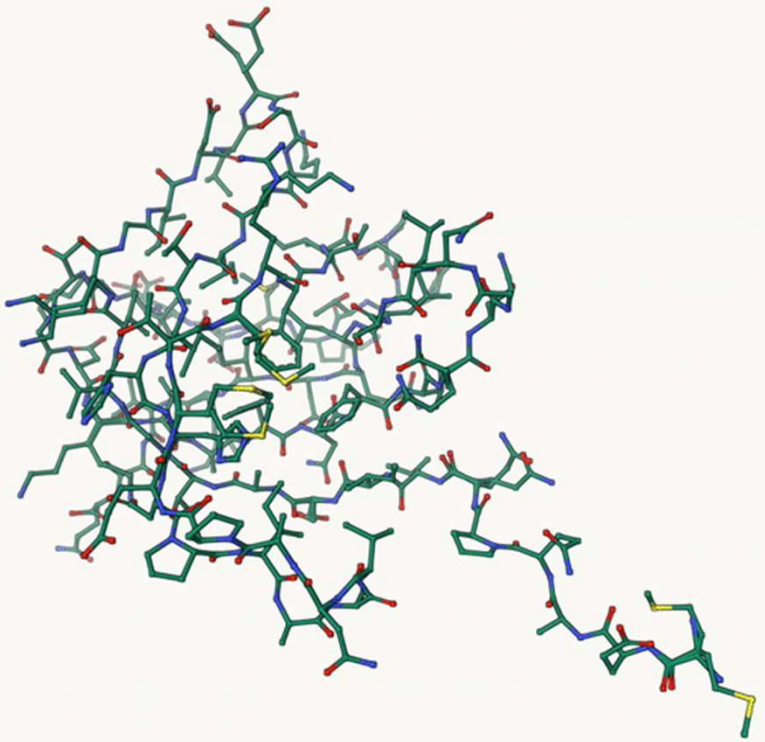
Secondary structure - proteins

Peptide
(Alpha helix)

Alpha helices



Beta strands

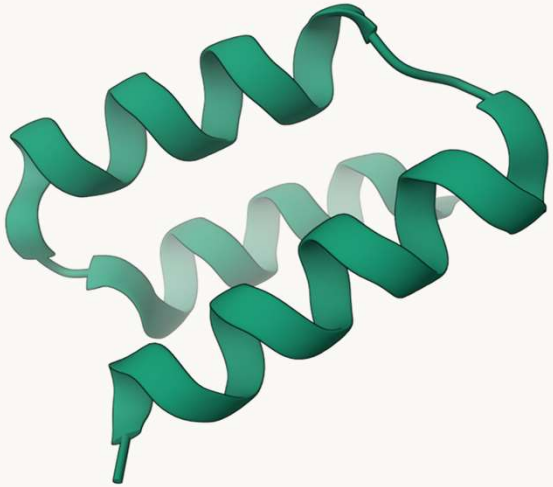


4G13, 3BBZ, 3E6Z

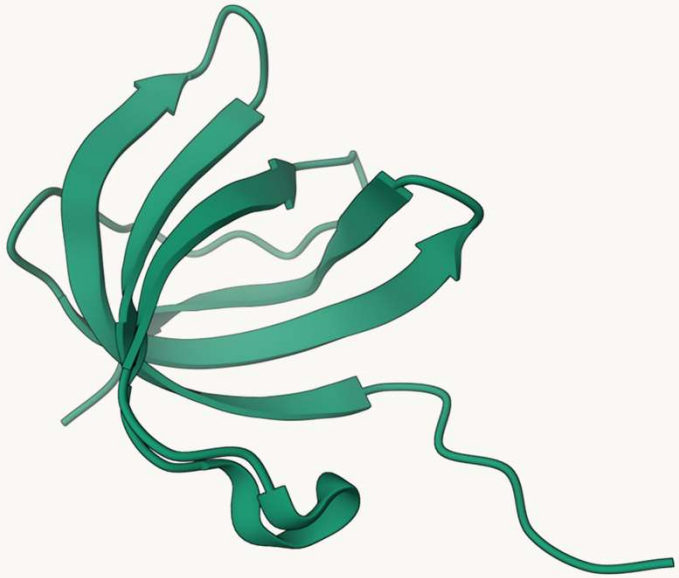
Secondary structure - proteins

Peptide
(Alpha helix)

Alpha helices



Beta strands



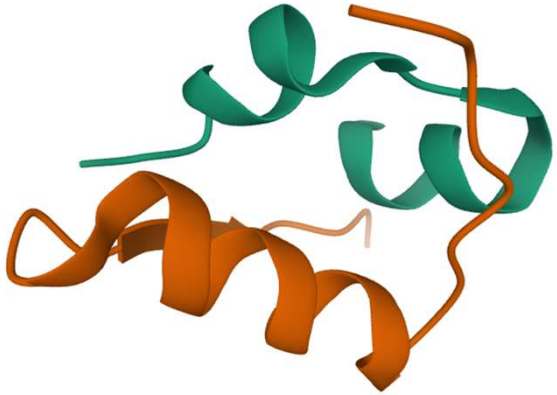
4G13, 3BBZ, 3E6Z

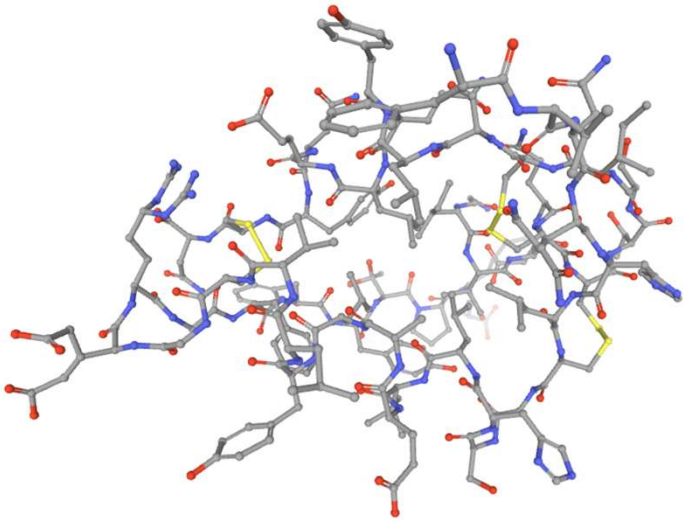
Structural folds - PDBeFold

- Tertiary folds up into a specific shape
  - Driven by hydrophobic interactions
  - Also, more specific interactions:
    e.g. disulphide bonds, salt bridges
  - Puts amino acids in the right place for function

→ **PDBeFold** – search for structural similarity (https://www.ebi.ac.uk/msd-srv/ssm/)

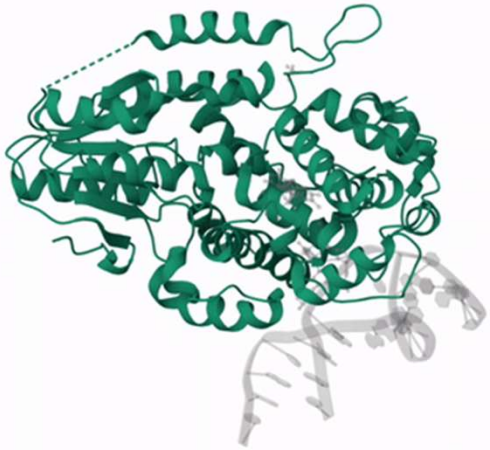
→ Protein domains / protein families (InterPro)





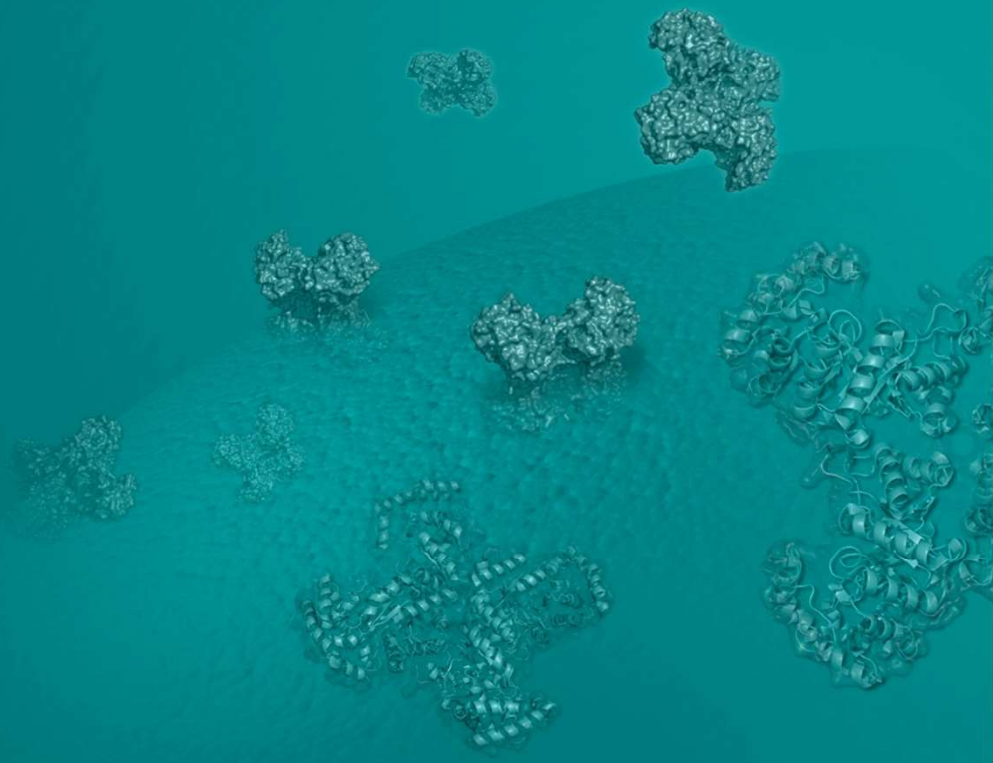
Assemblies - PDBePISA

- Multiple macromolecules interacting together
- Can be simple dimers or huge complexes
- Amino acids mediate interactions
  - Usually, hydrophobic patches
  - Can also be disulphide bridges



→**PDBePISA** to analyse assemblies and interfaces
(https://www.ebi.ac.uk/pdbe/pisa/)

Searching the PDBe

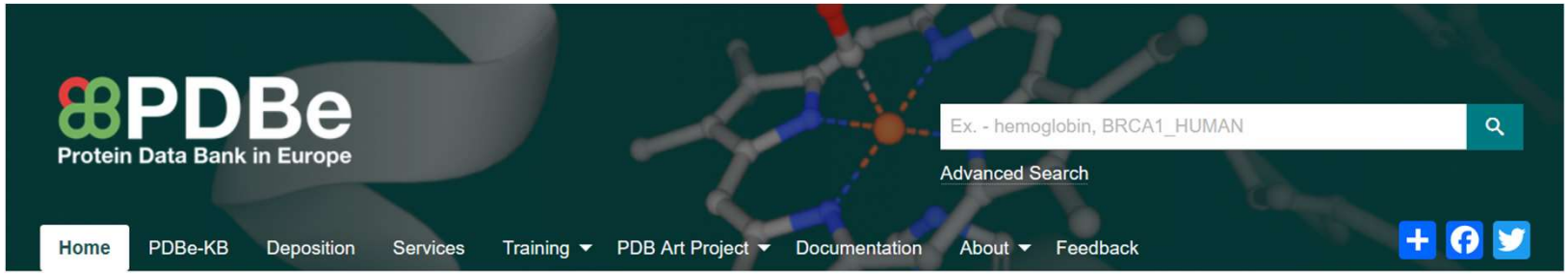


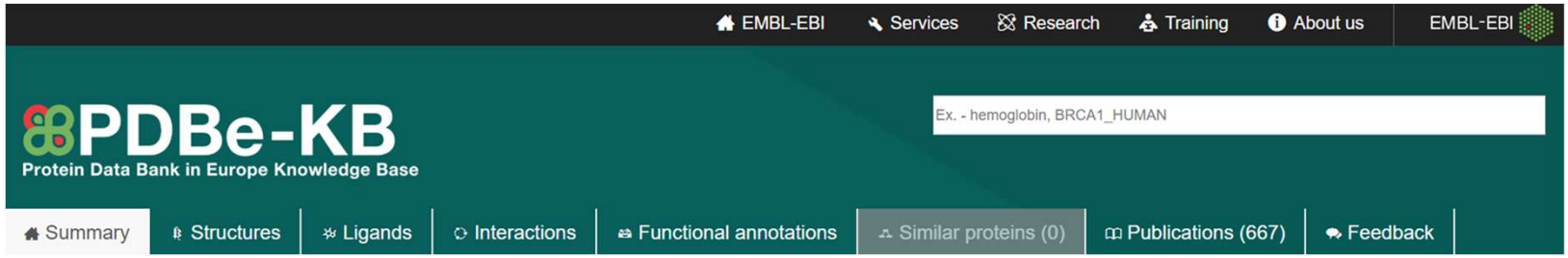
PDBe

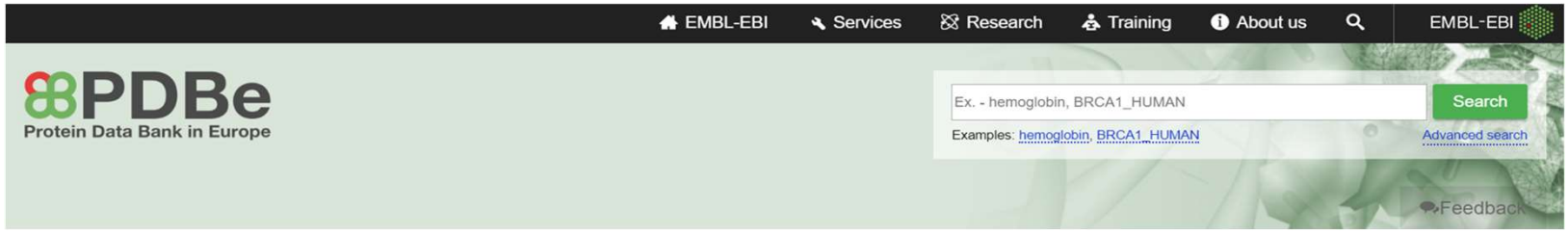
Protein Data Bank in Europe

EMBL-EBI

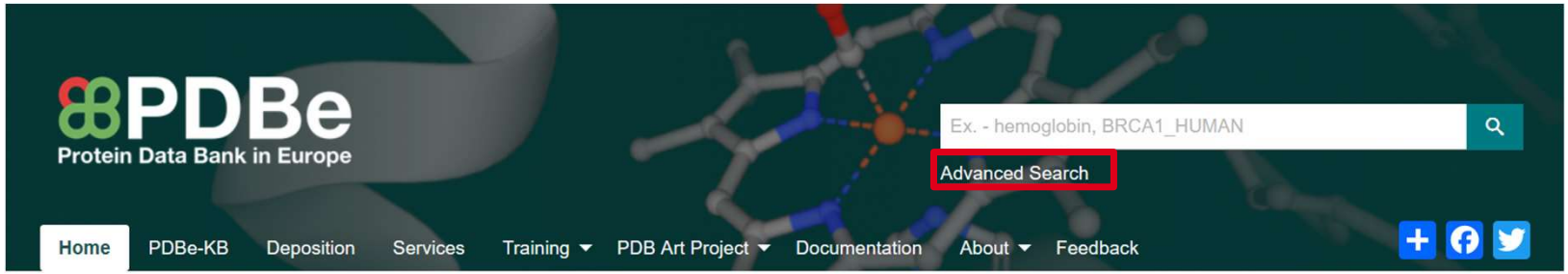
At PDBe: look for the search bars

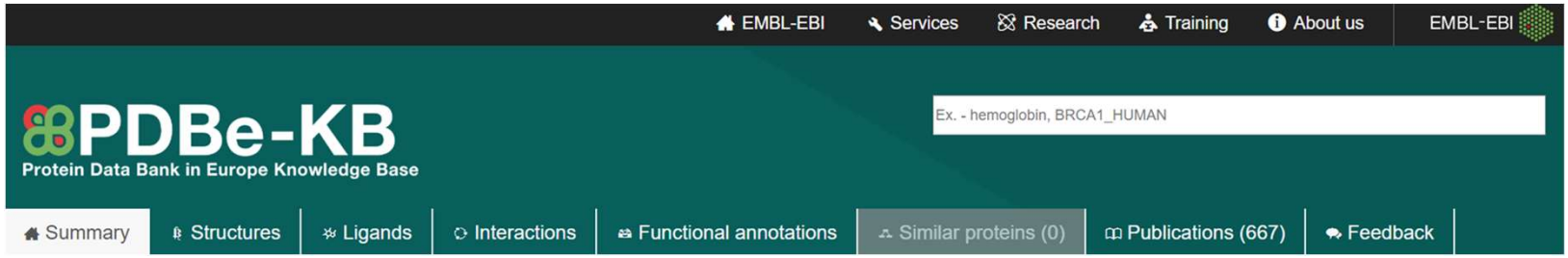


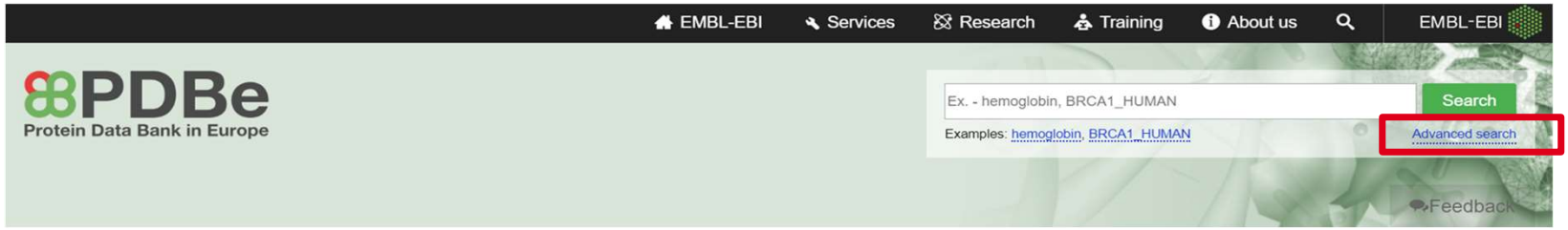




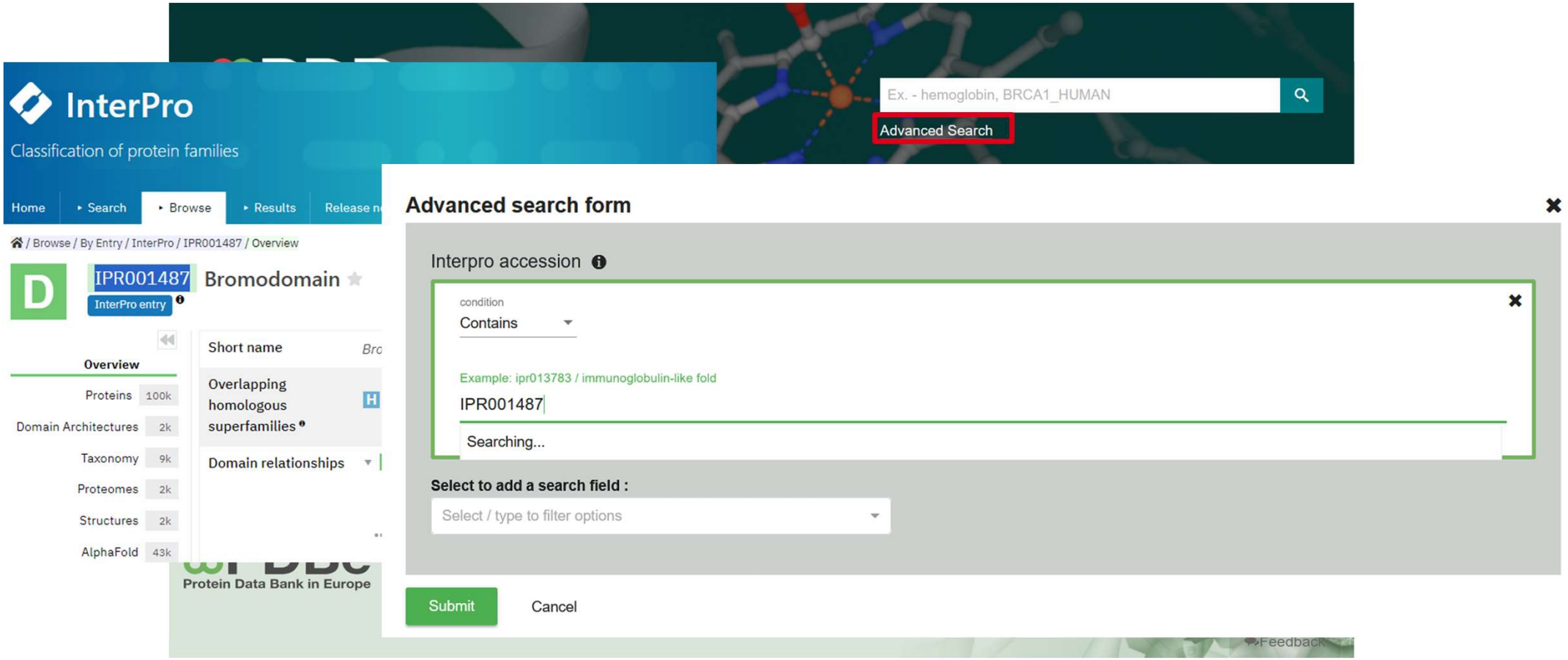
Search bars → multi-purpose tool → ‘Advanced search’



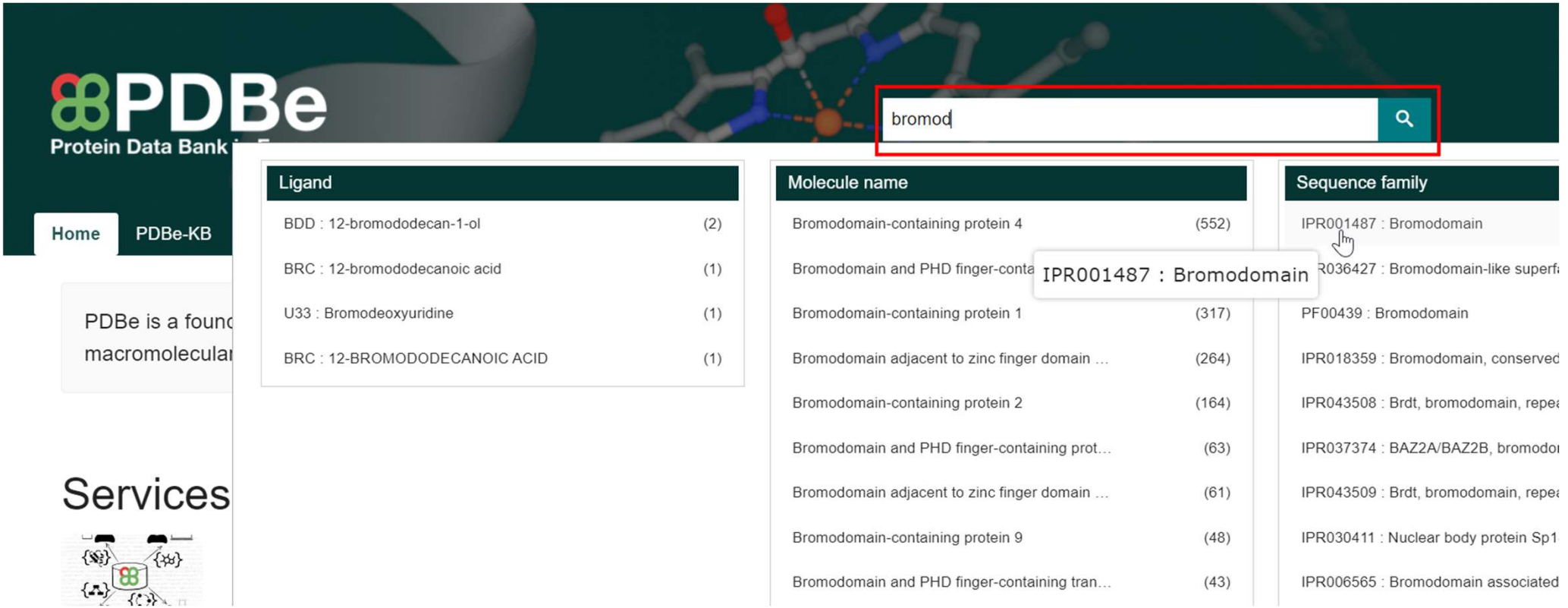




‘Advanced search’ - Protein domains



Search bars → multi-purpose tool → ‘autocomplete options’



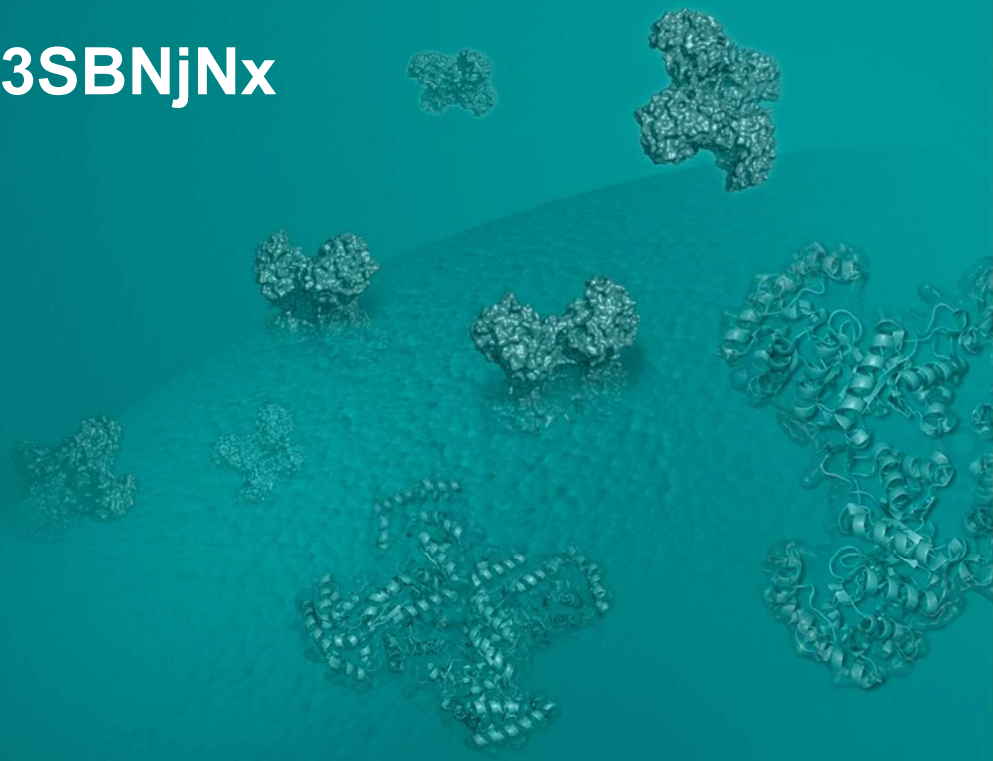
API tutorials

https://github.com/glevans/7ADD-workshop-2024

1_API_basics.ipynb

2_GET_vs_POST.ipynb

https://bit.ly/3SBNjNx



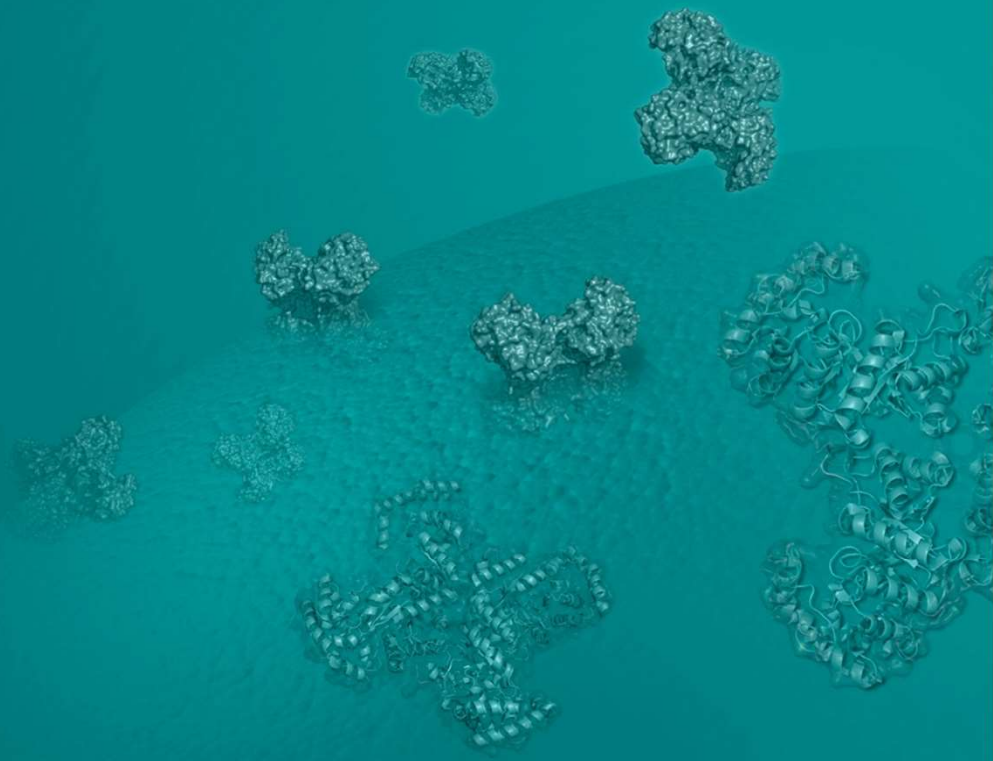
More advanced:

https://github.com/PDBeurope/pdbe-api-training/
→ subfolder api_tutorials

PDBe
Protein Data Bank in Europe

EMBL-EBI

How do I pick the best structure for me? (part1)



PDBe
Protein Data Bank in Europe

EMBL-EBI

How good is a structure?

- Maps*

  * - *X-ray crystallography & cryo electron microscopy*

- Resolution*

- Geometry

- Validation

How good is a structure?

- Maps* are the real data

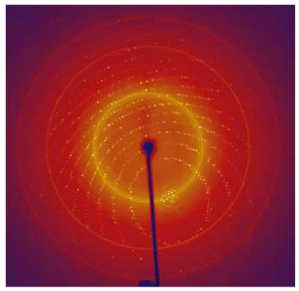
  * - *X-ray crystallography & cryo electron microscopy*

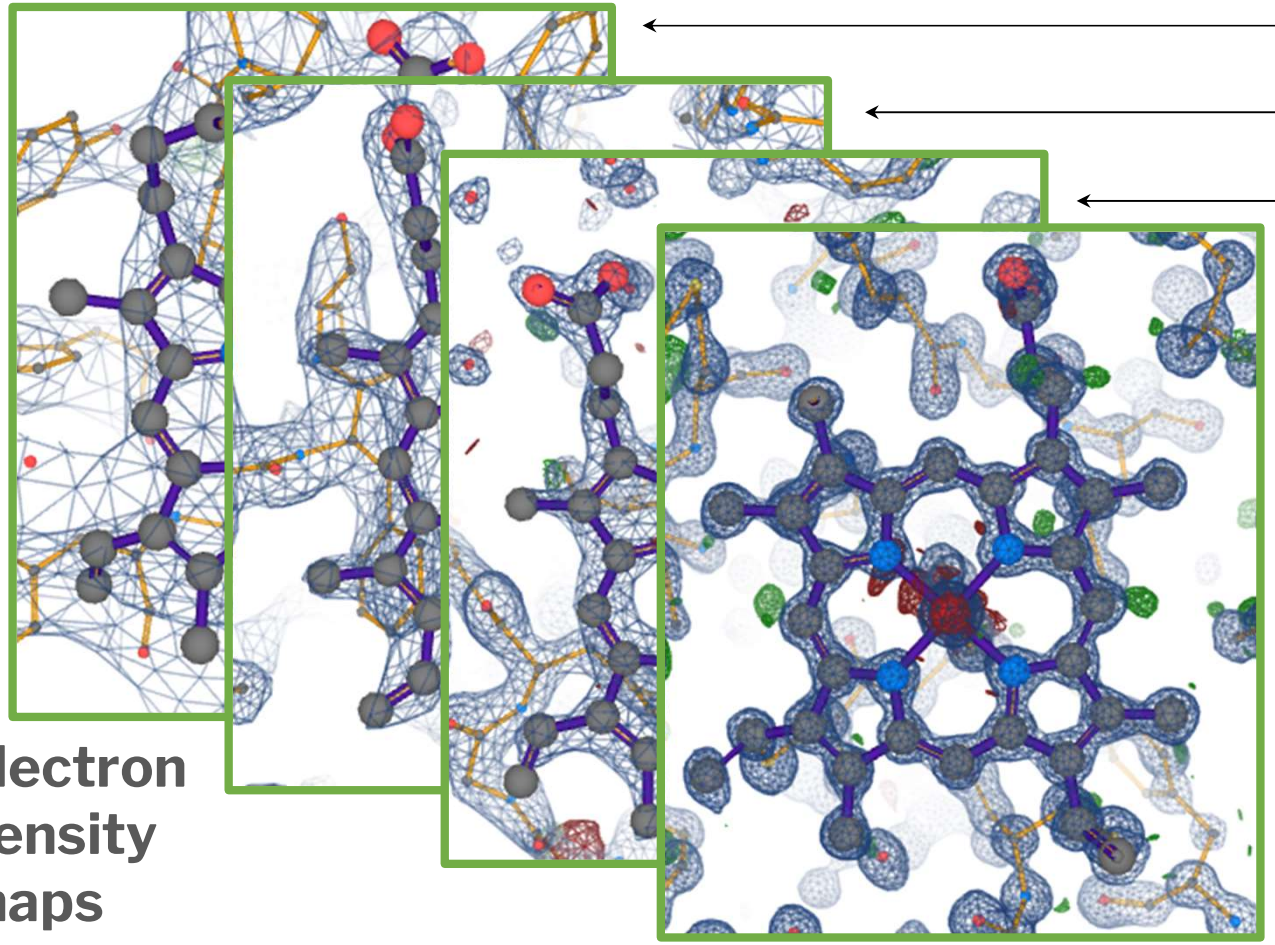
- Resolution* matters, but isn't everything

- Geometry

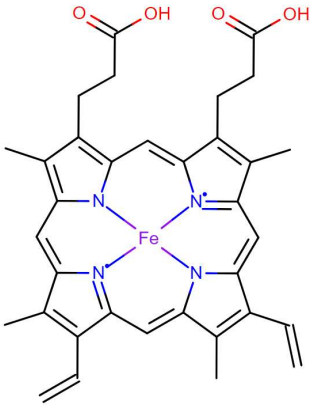
- Validation

Maps are the real data - X-ray crystallography density maps





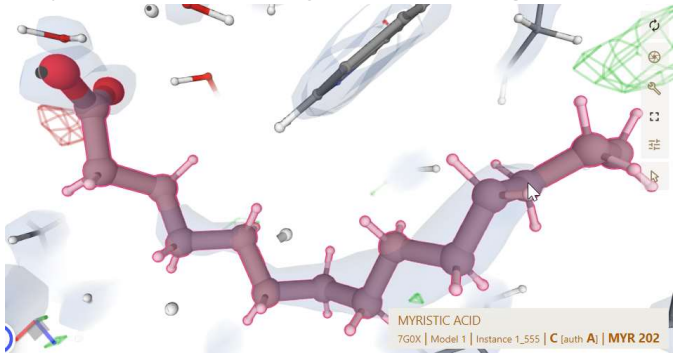
Electron density maps



Viewing electron density in Mol*

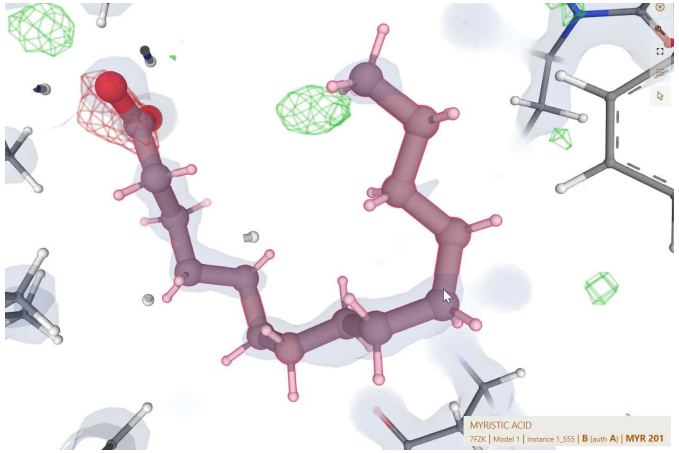
- To initiate electron maps display:
  clicking on ligand or protein amino acid
- Regular map (**blue**) '2Fo-Fc' electron density map
  → *should surround atoms*
- **Negative** and **positive** density
  → *highlights extra and missing atoms, respectively*

Myristic acid in **Fatty acid-binding protein 4**

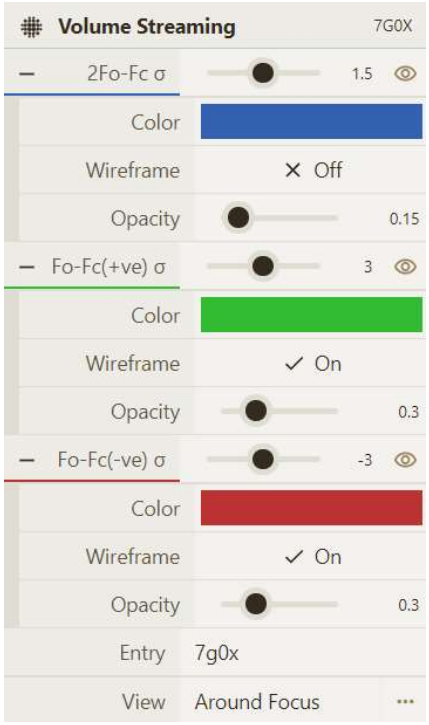


PDB ID: 7G0X

Myristic acid in **Fatty acid-binding protein 4**



PDB ID: 7FZK

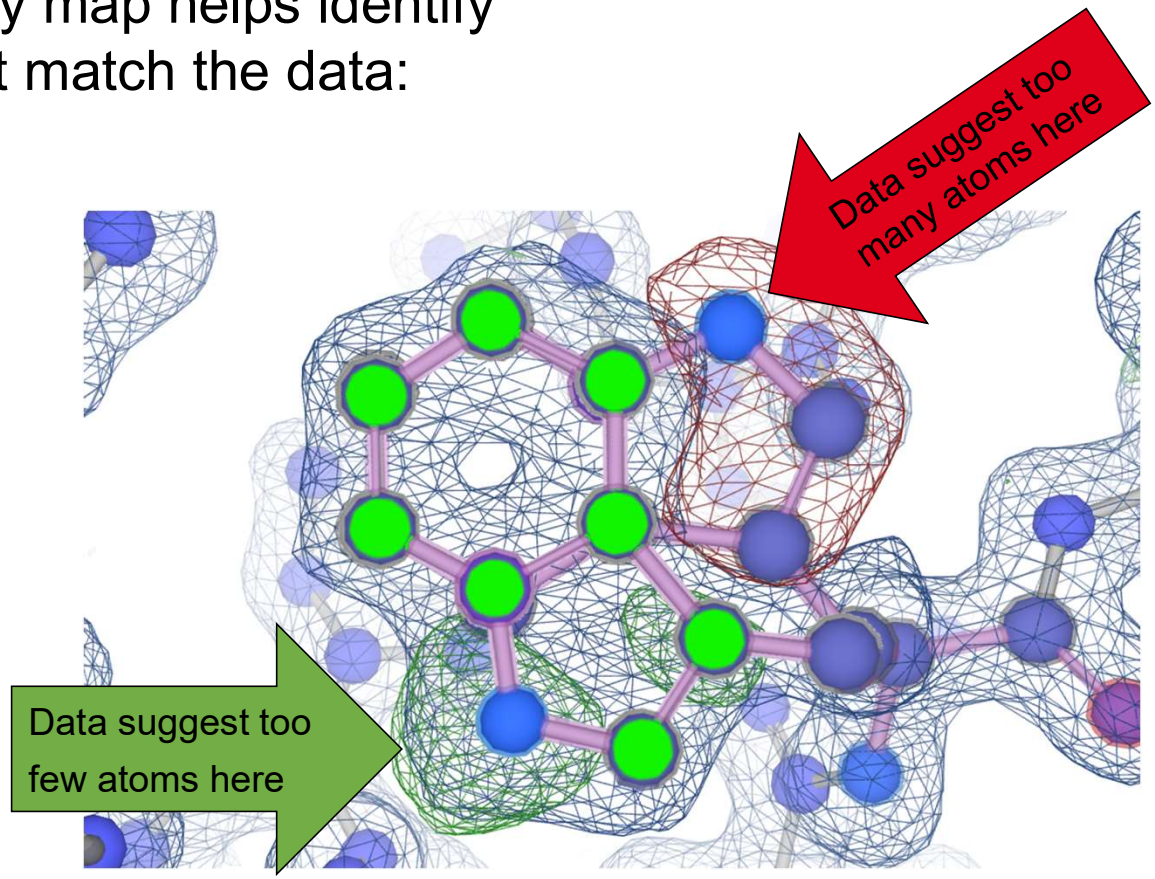


Maps are the real data - X-ray crystallography density maps

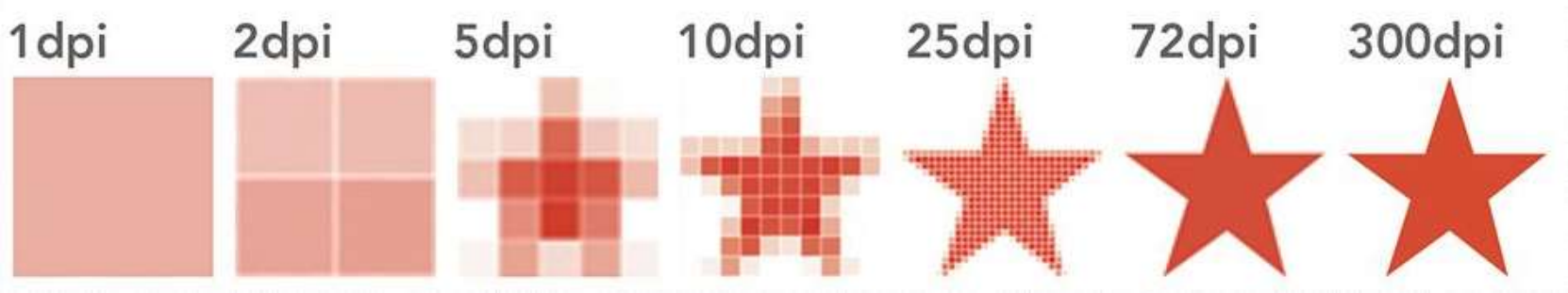
- Difference electron density map helps identify where the model does not match the data:

Areas where the model has too many atoms for the data

Areas where the model has too few atoms for the data

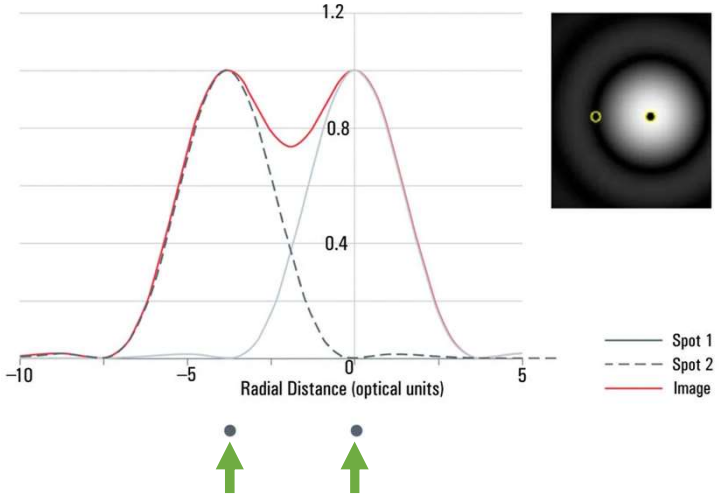


Resolution



Digital resolution

Larger number
= better resolution

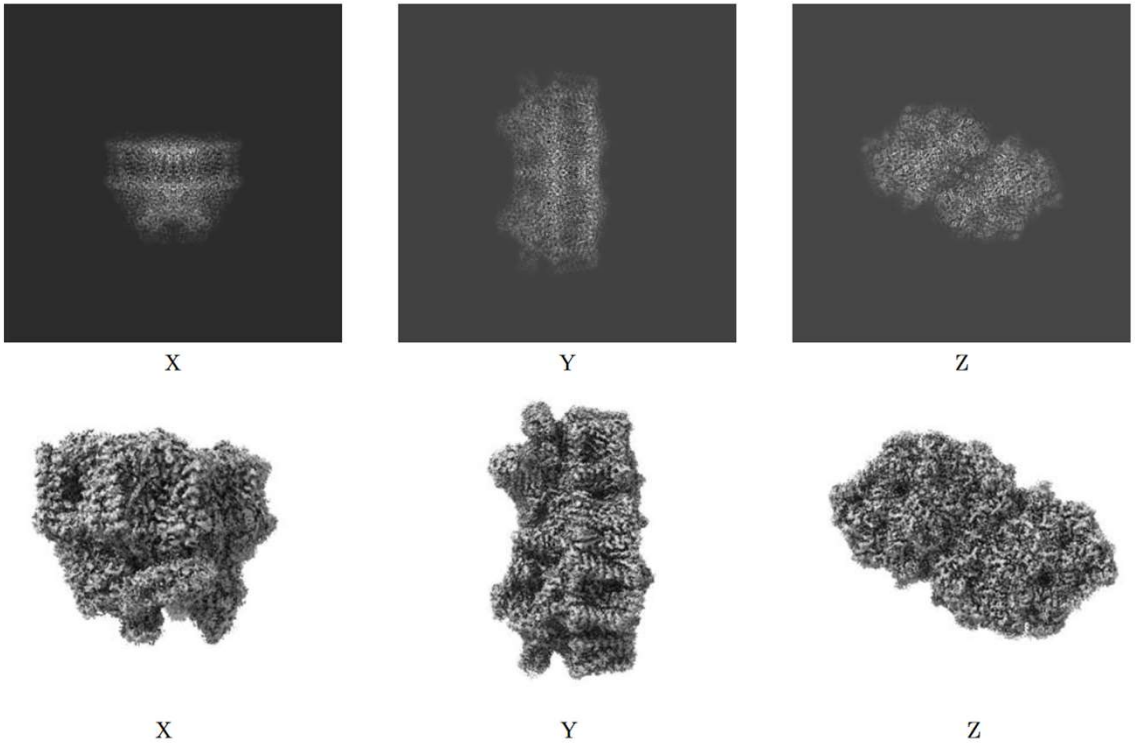


Microscope resolution

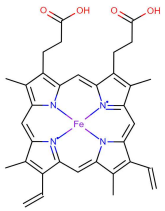
Smaller number
= better resolution

Maps are the real data - Cryo-electron microscopy volumes

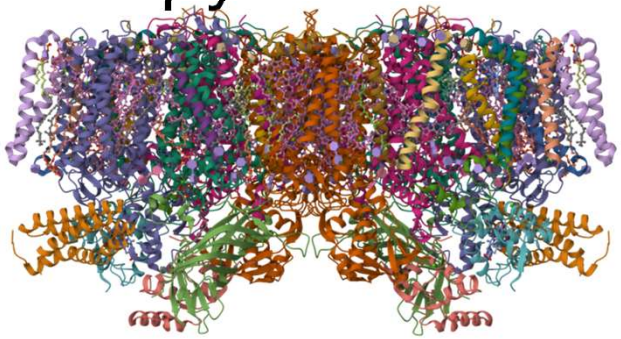
Photosystem II from cyanobacterium
(Resolution:1.9 Å)



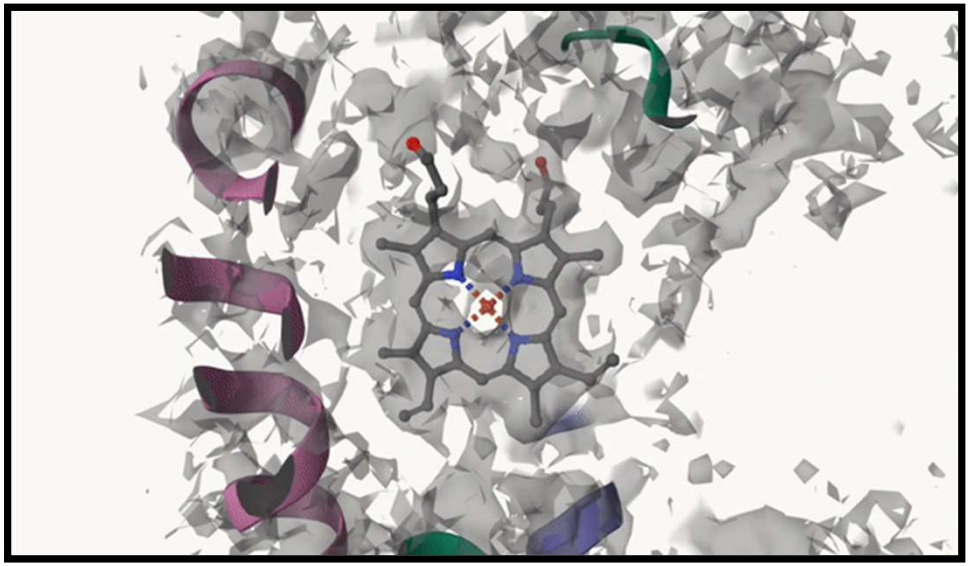
EM volume map



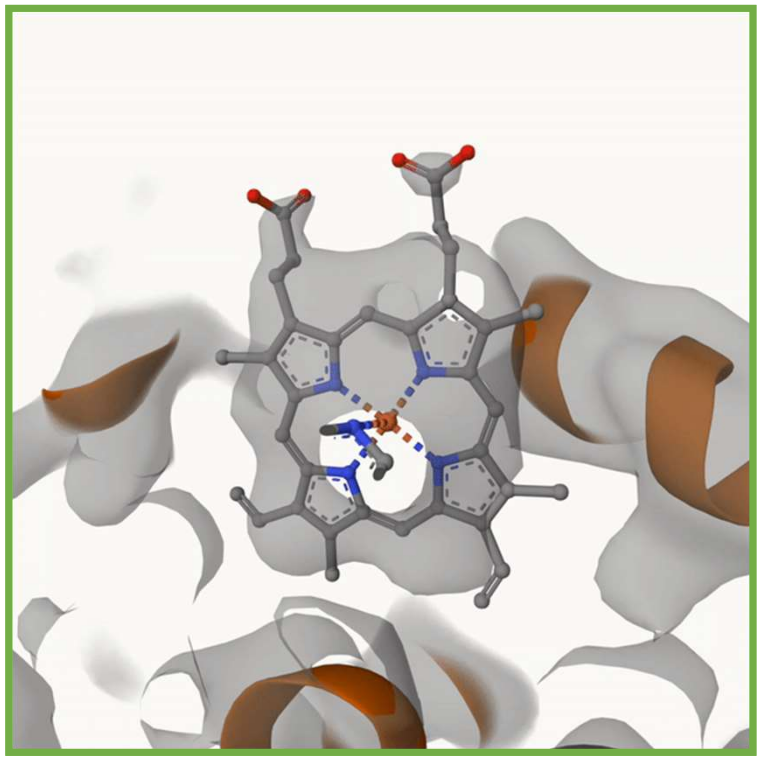
HEME



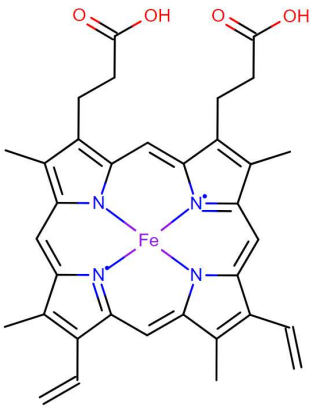
21 distinct protein chains



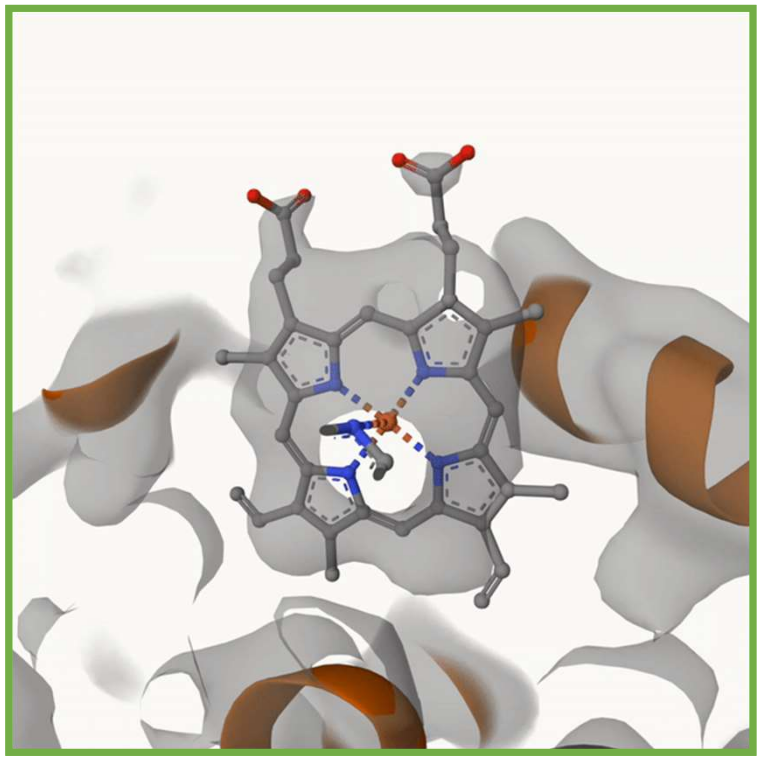
Maps are the real data - Cryo-electron microscopy volumes



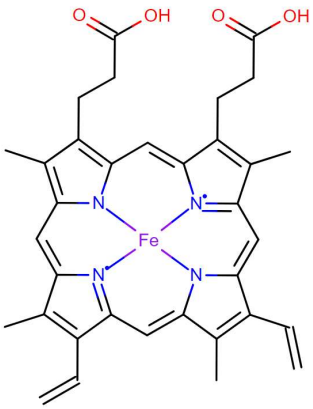
3.5Å



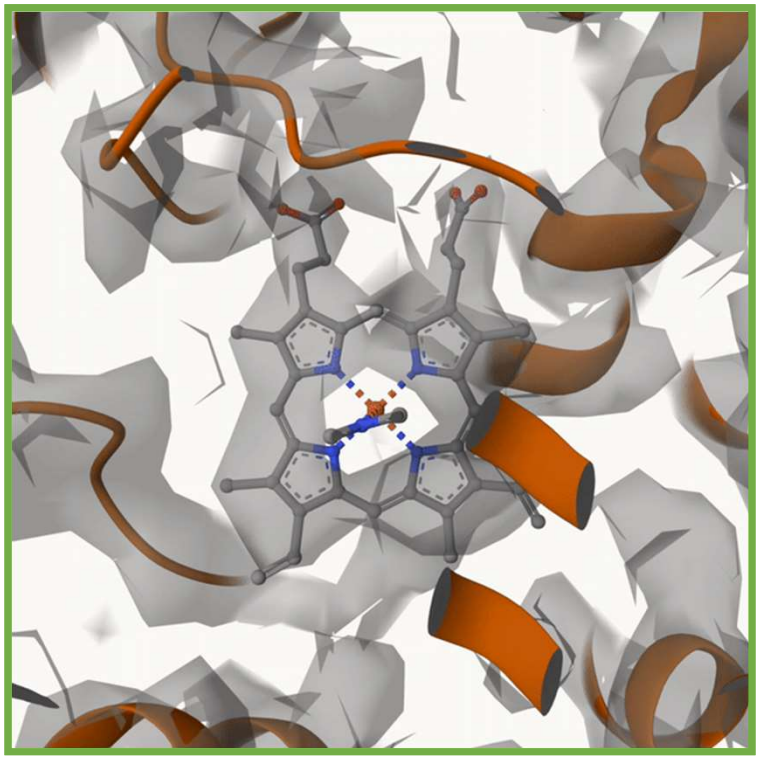
Maps are the real data - Cryo-electron microscopy volumes



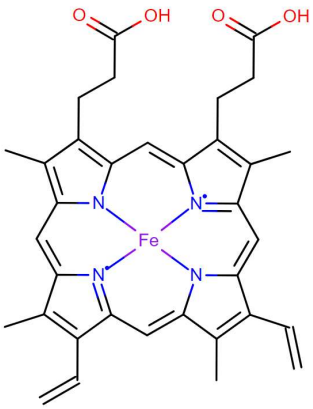
3.5Å



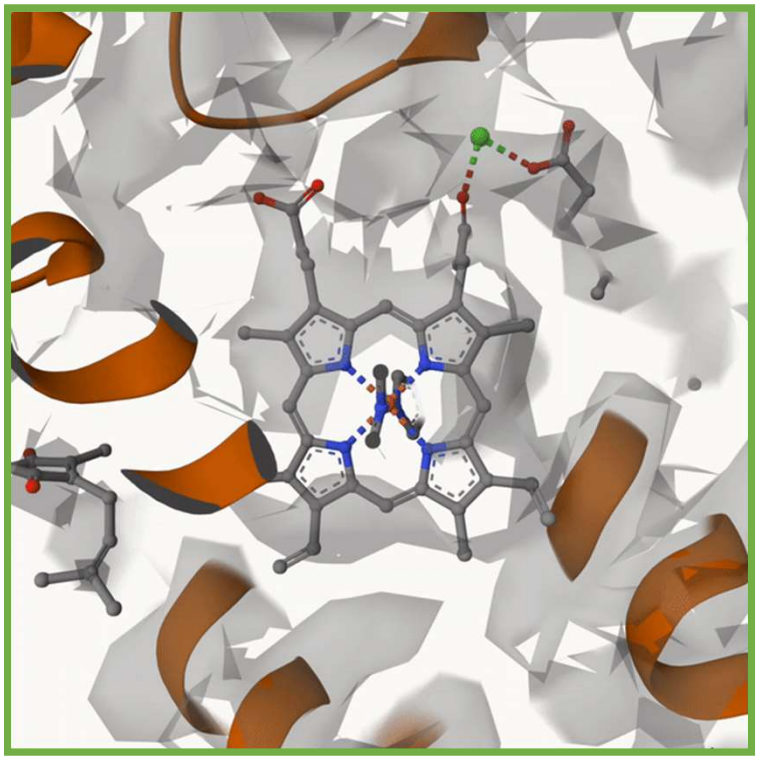
Maps are the real data - Cryo-electron microscopy volumes



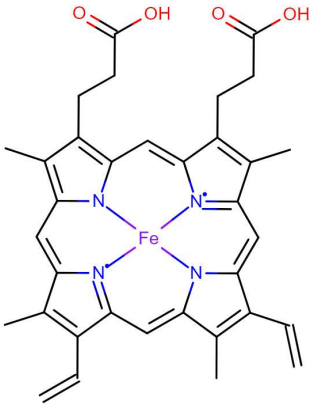
2.7Å



Maps are the real data - Cryo-electron microscopy volumes

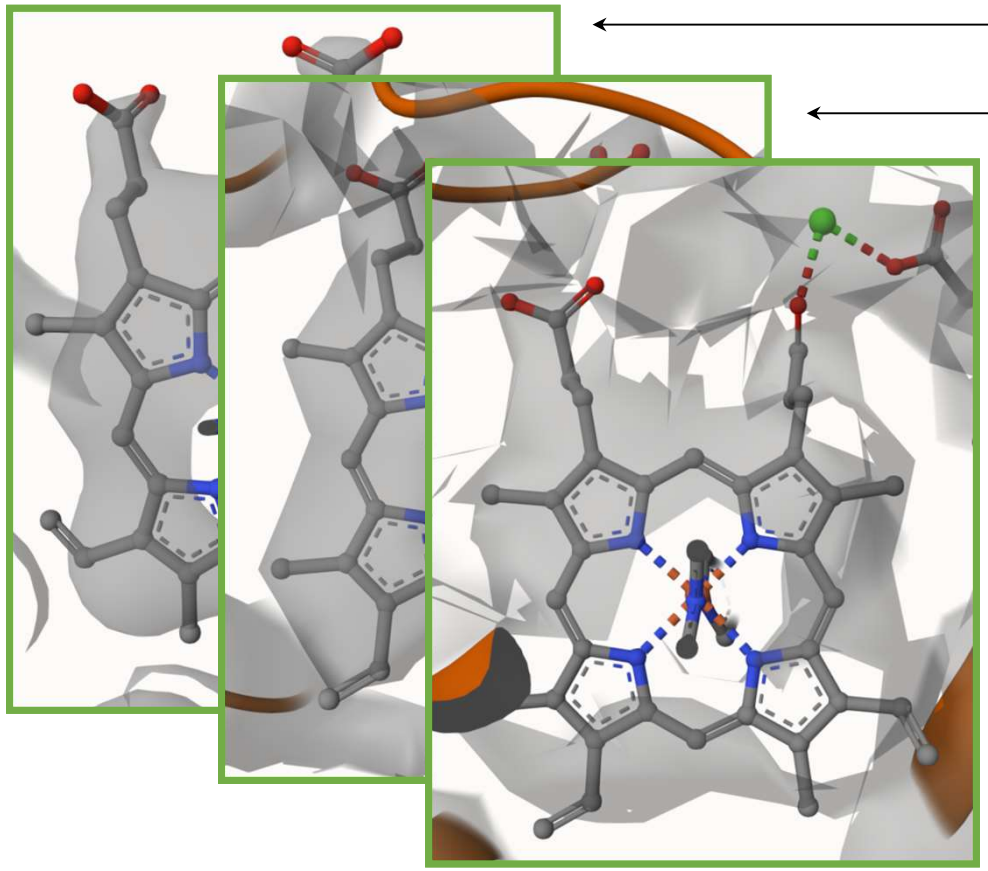


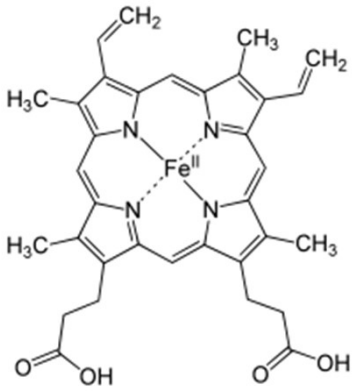
← 2.2Å



Maps are the real data - Cryo-electron microscopy volumes

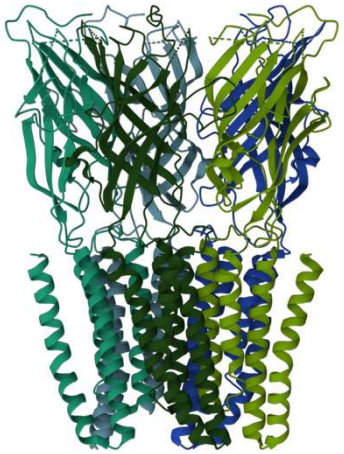
Global resolution

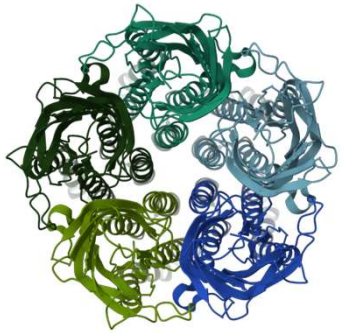




Local resolution vs. Global resolution

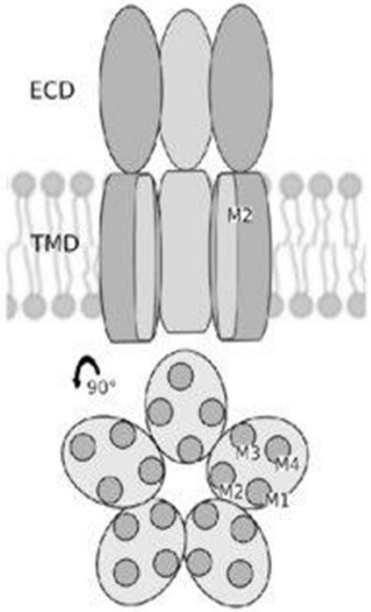
Example: Ion channel (ligand-gated)

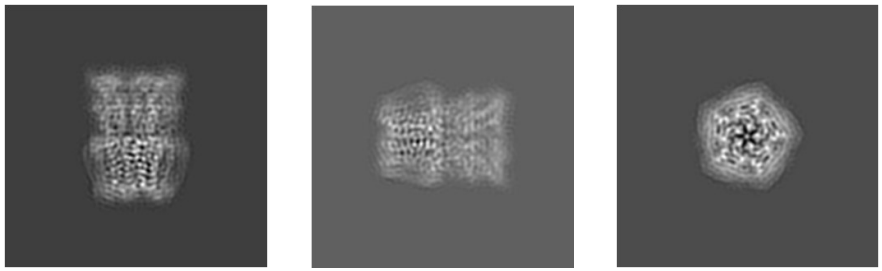




ECD
= extra-cellular domain

TMD
= transmembrane domain





X Y Z

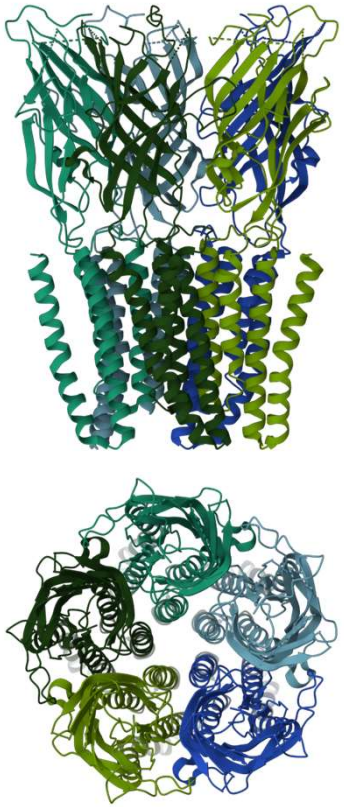
EM volume map

6ZGD, 6GZJ, 6ZGK

EMBL-EBI

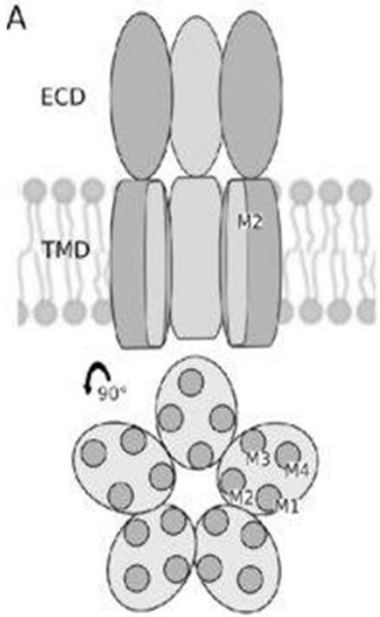
Local resolution vs. Global resolution

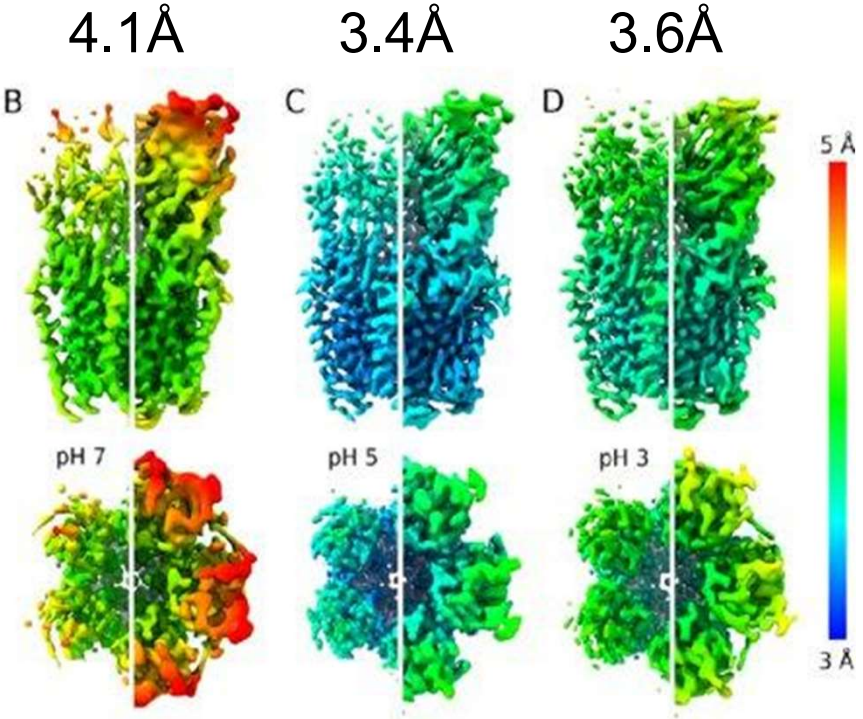
Example: Ion channel (ligand-gated)



ECD
= extra-cellular domain

TMD
= transmembrane domain





EM volume maps

6ZGD, 6GZJ, 6ZGK

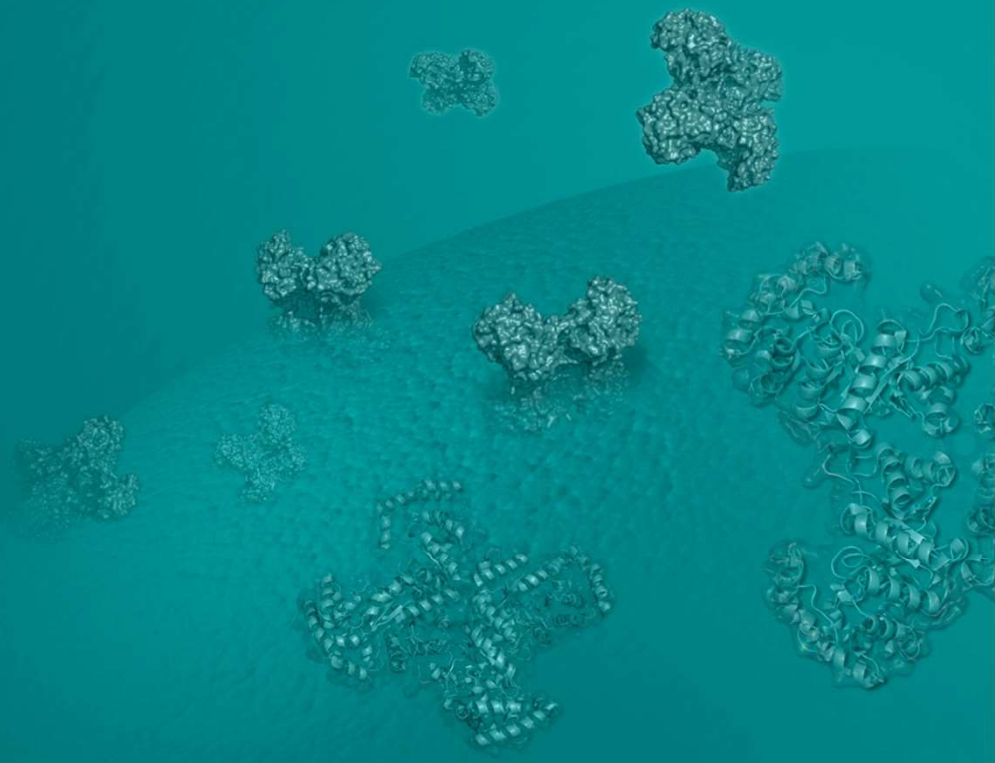
Molstar & ligand site tutorials

https://github.com/glevans/7ADD-workshop-2024

3_PDBe_PDBe-KB_ligand_view_tutorial.docx

3_PDBe_PDBe-KB_ligand_view_tutorial.pdf

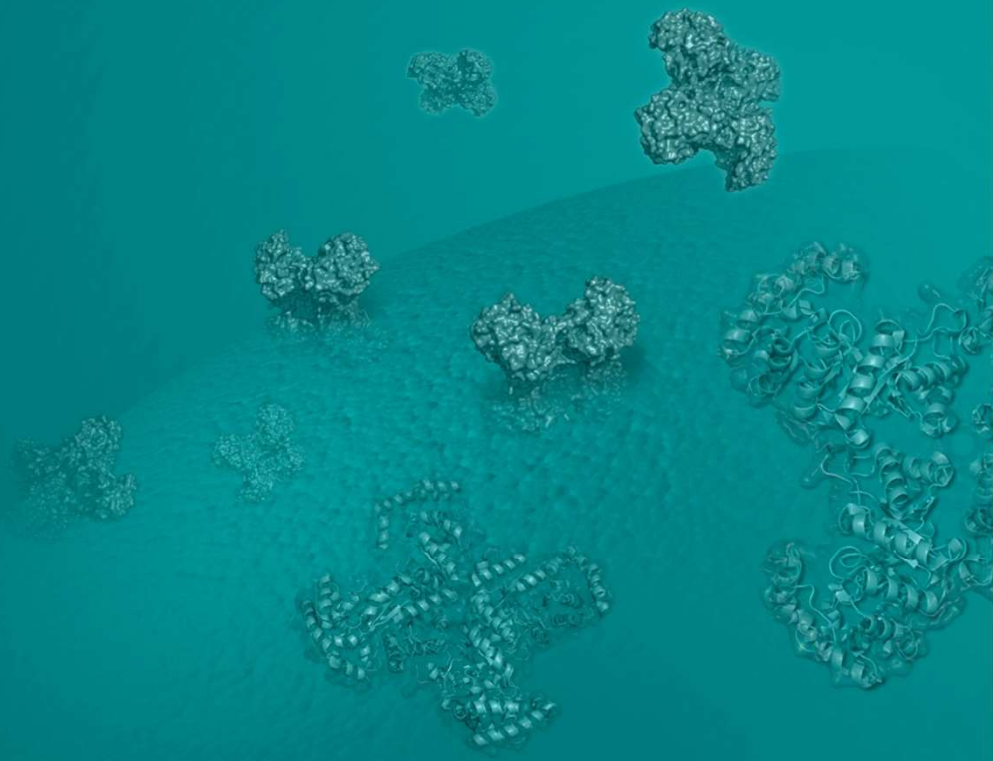
https://bit.ly/3SkNRpo



PDBe
Protein Data Bank in Europe

EMBL-EBI

How do I pick the best structure for me? (part2)



PDBe
Protein Data Bank in Europe

EMBL-EBI

How good is a structure?

- Maps are the real data*

  * - *X-ray crystallography & cryo electron microscopy*

- Resolution matters, but isn't everything

- Geometry can be evaluated

- Validation reports with 'sliders'

View Validation information

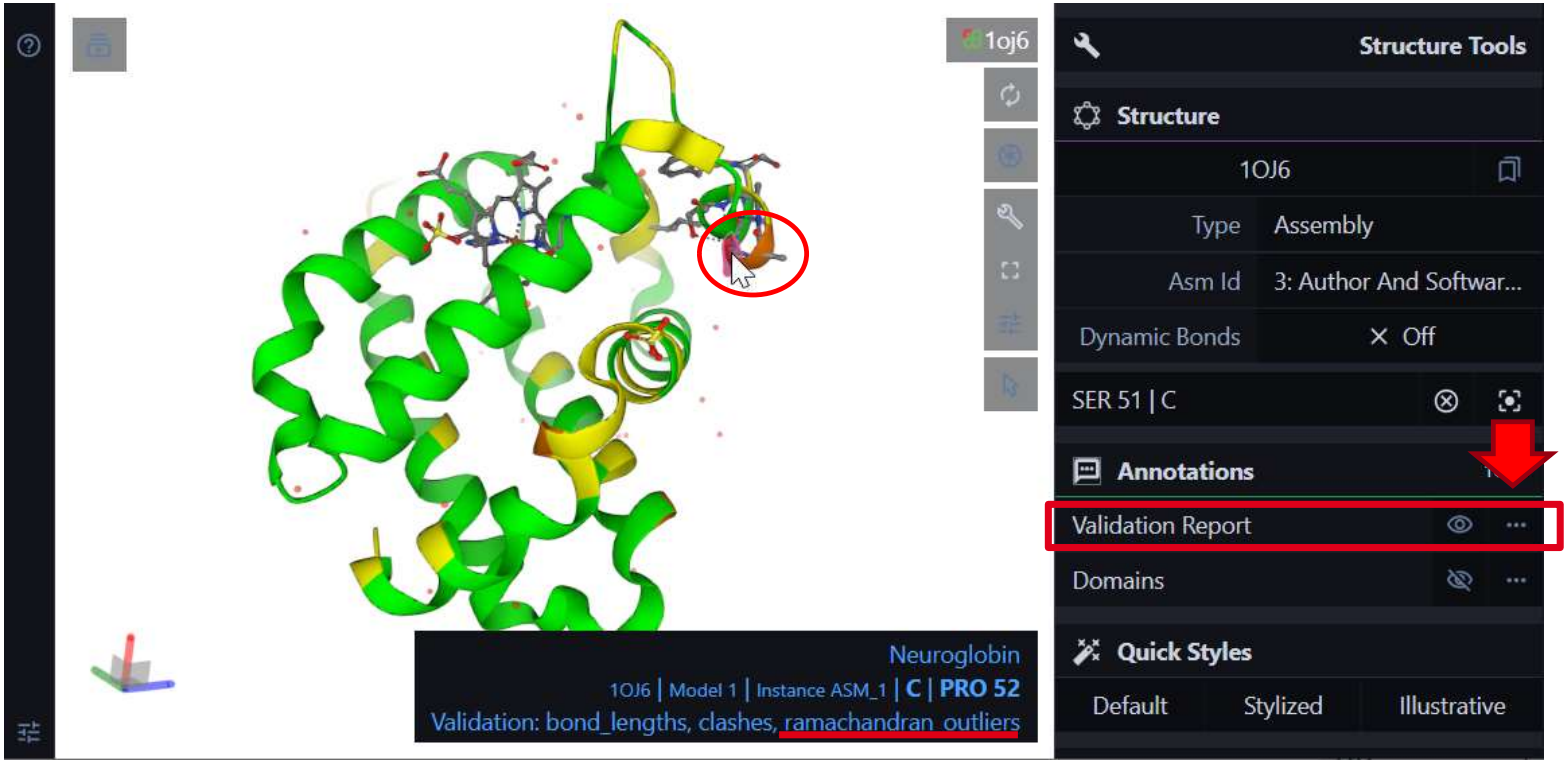
- Validation information (geometry information etc) can be displayed on the structure

View Validation information

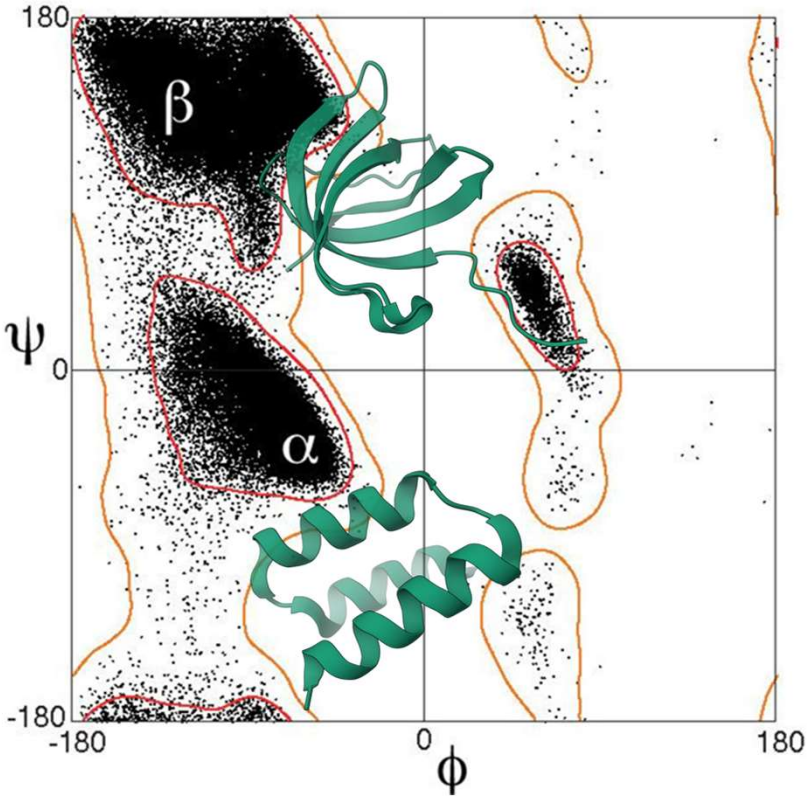
- Validation information (geometry information etc) can be displayed on the structure

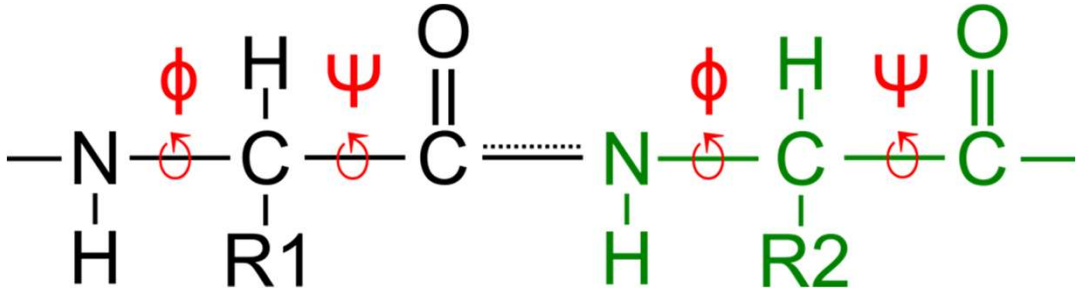
View Validation information

- Validation information (geometry information etc) can be displayed on the structure



How we judge geometry:
Ramachandran Plot / Ramachandran Outliers

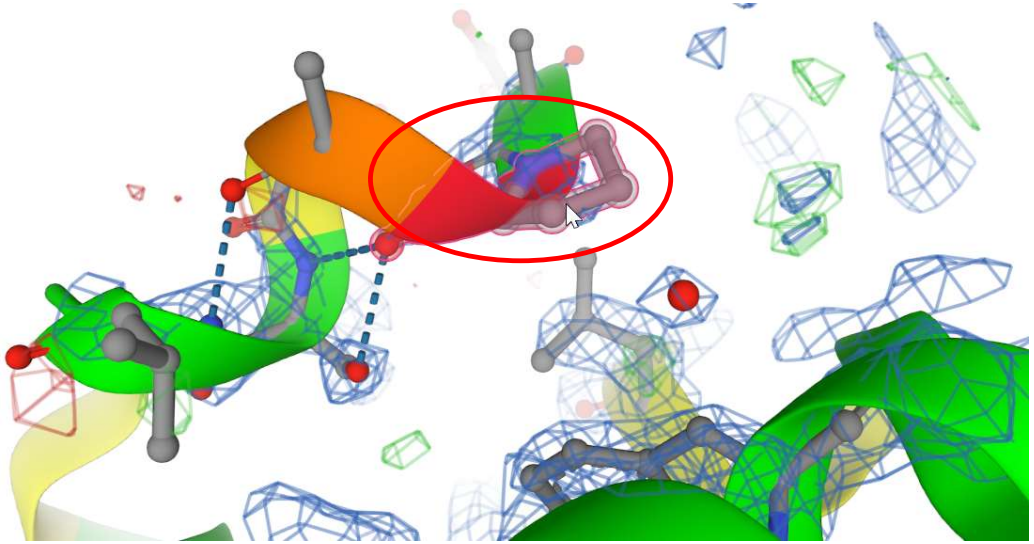




Geometry and fit to data can go hand-in-hand

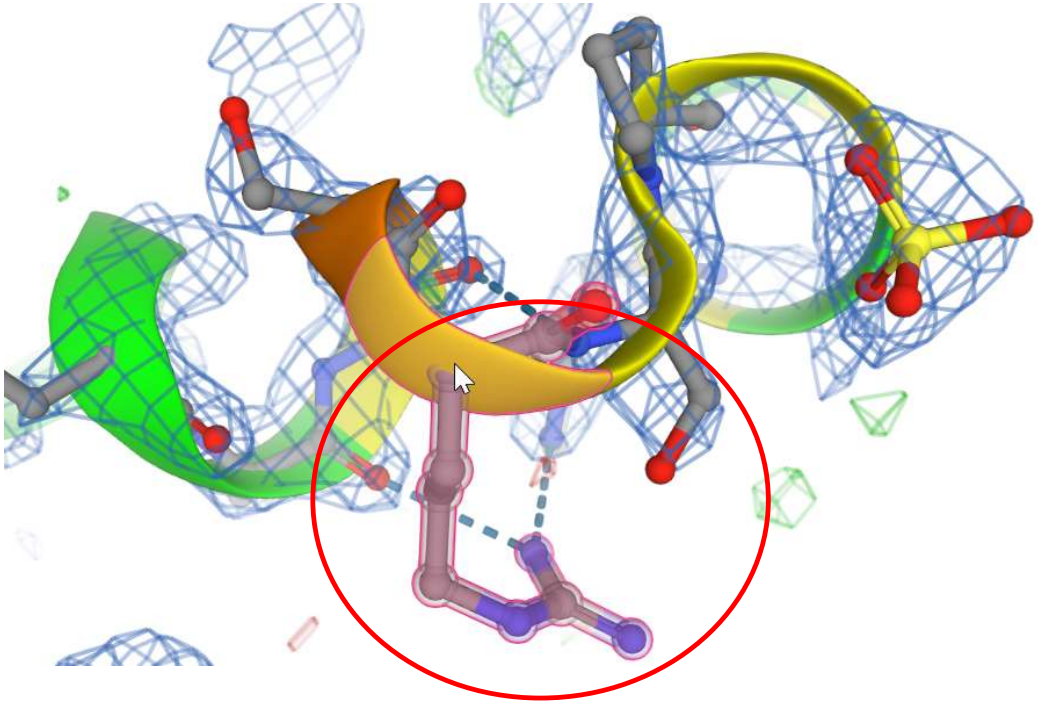
Pro 52 in entry 1oj6 is a Ramachandran outlier
(and has unusual bond lengths etc)

The strained conformation is NOT supported by data



Arg 18 in entry 1oj6 is a Sidechain outlier

The unusual conformation is NOT supported by data



Not all structures are created equal!

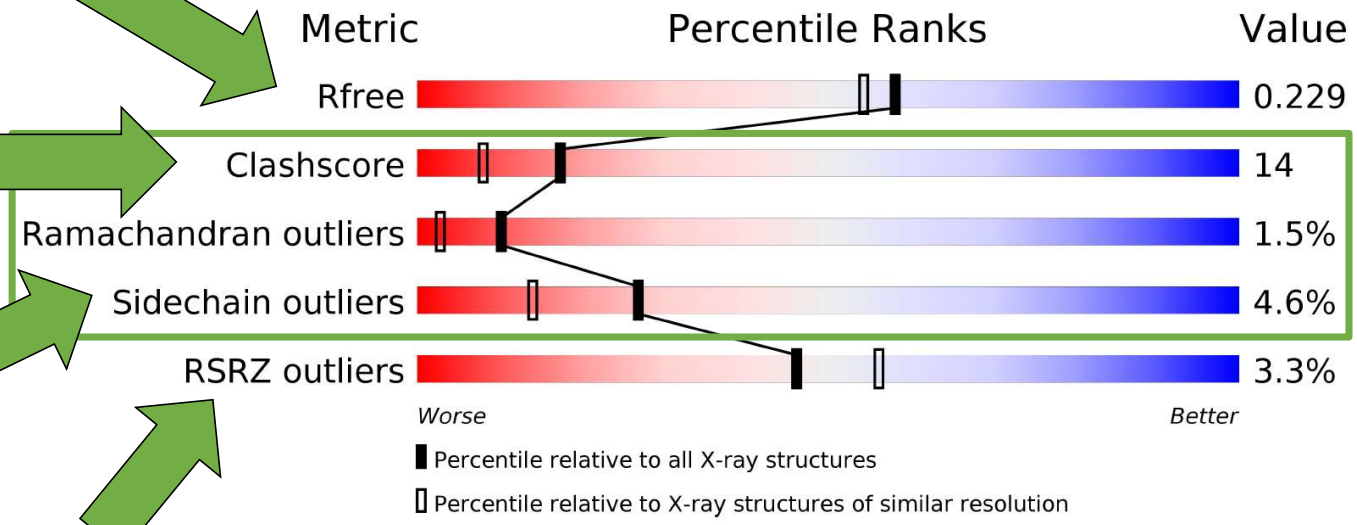
Validation sliders are here to help

How well does the model back-predict the data?

Atoms bumping into each other

Surprising bond angles

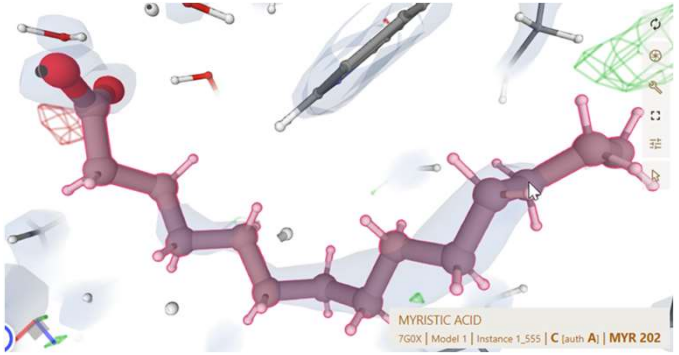
Atoms not in electron density



Viewing maps / electron density in Mol*

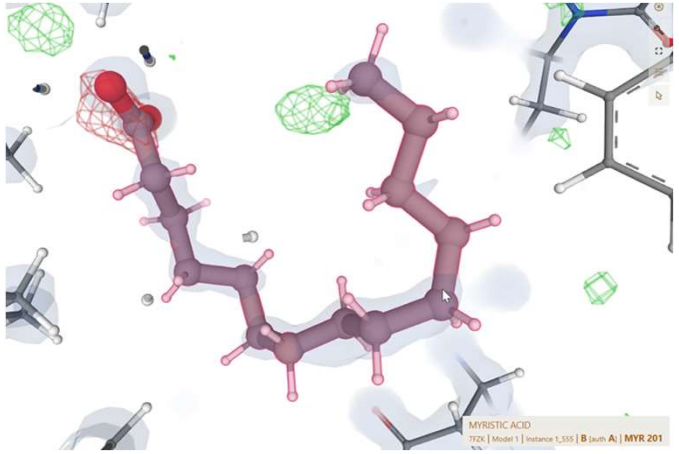
- To initiate electron maps display:
  clicking on ligand or protein amino acid
- Regular map (blue) '2Fo-Fc' electron density map
  should surround atoms
- Negative and positive density
  highlights extra and missing atoms, respectively

Myristic acid in **Fatty acid-binding protein**



PDB ID: 7G0X

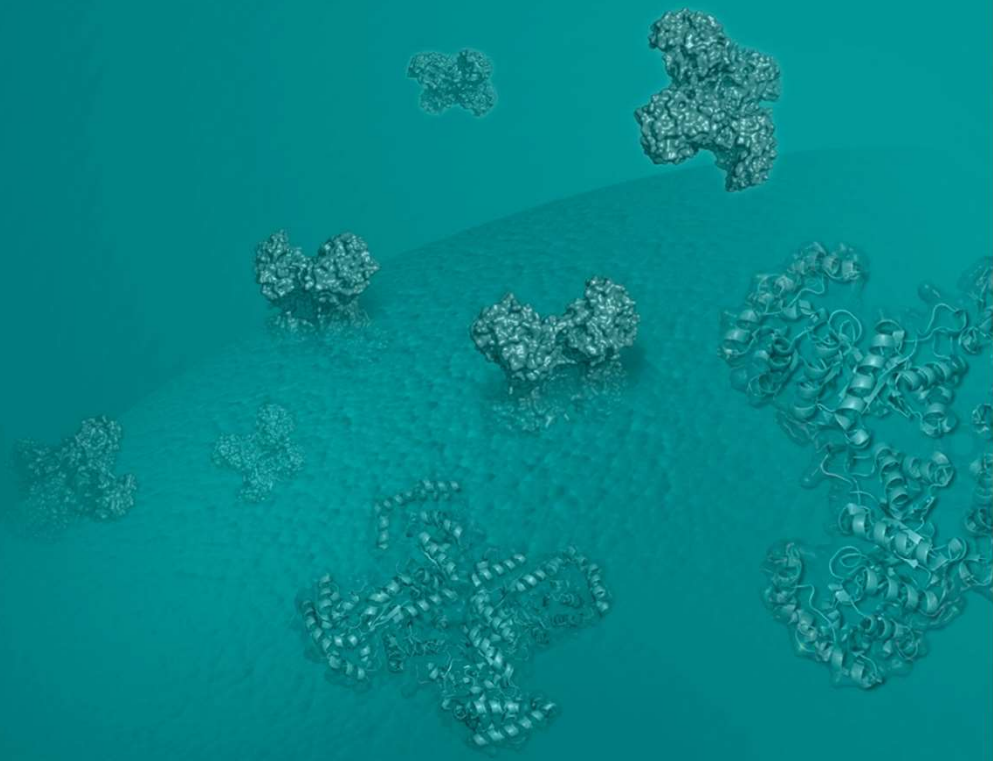
Myristic acid in **Fatty acid-binding protein**



PDB ID: 7FZK

PDBe-KB aggregated views of proteins

Genevieve Evans



PDBe-KB

Protein Data Bank in Europe - Knowledge Base

EMBL-EBI

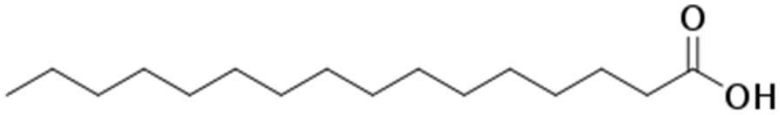
In January 2023, we passed the 200,000 milestone. As of today, ~214,000 structures

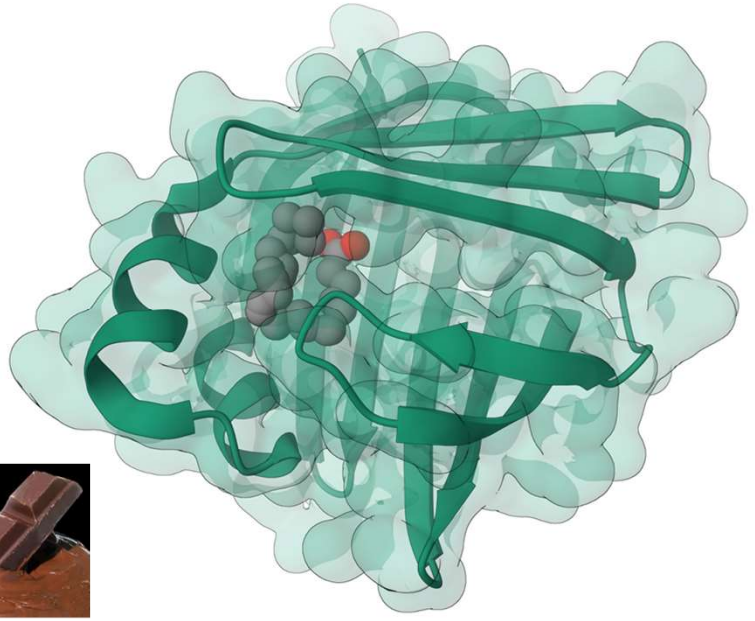
Not all ~215,000 structures are unique molecules

There are ~60k structures of human proteins

But only ~15k different human proteins
(~4K different protein families)

Why?





Solved by *different experimental methods*

With *different compounds* bound

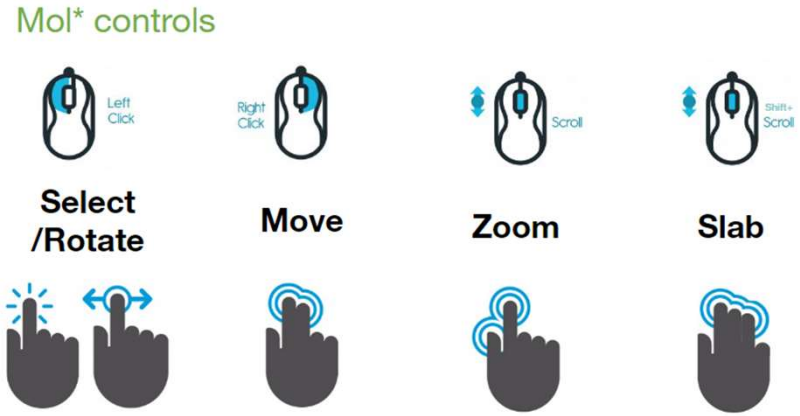
By *different people*

In *different conditions*

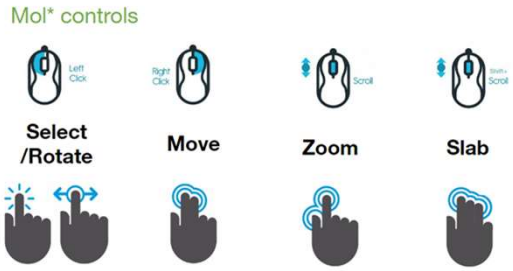
Fatty acid-binding protein: 242 structures,
227 different small molecules

Visualization in 3D with Mol*

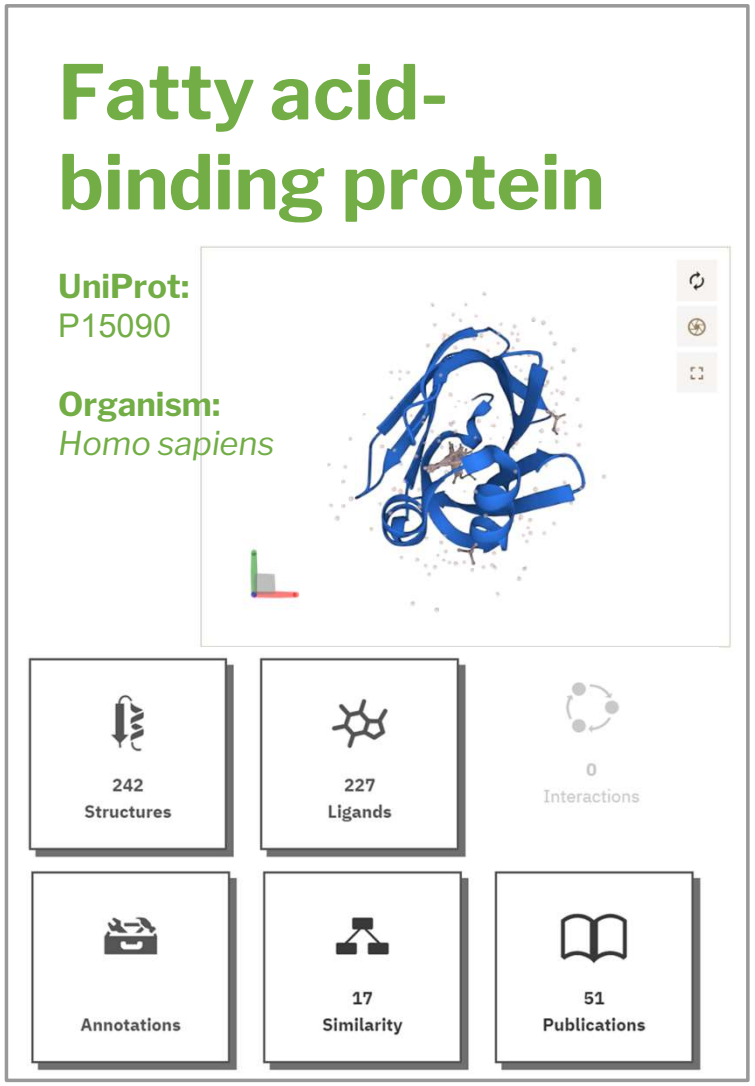
- Browser-based
  - Interactive & many features
  - Generate images / animations in browser window
  - Relatively easy to navigate



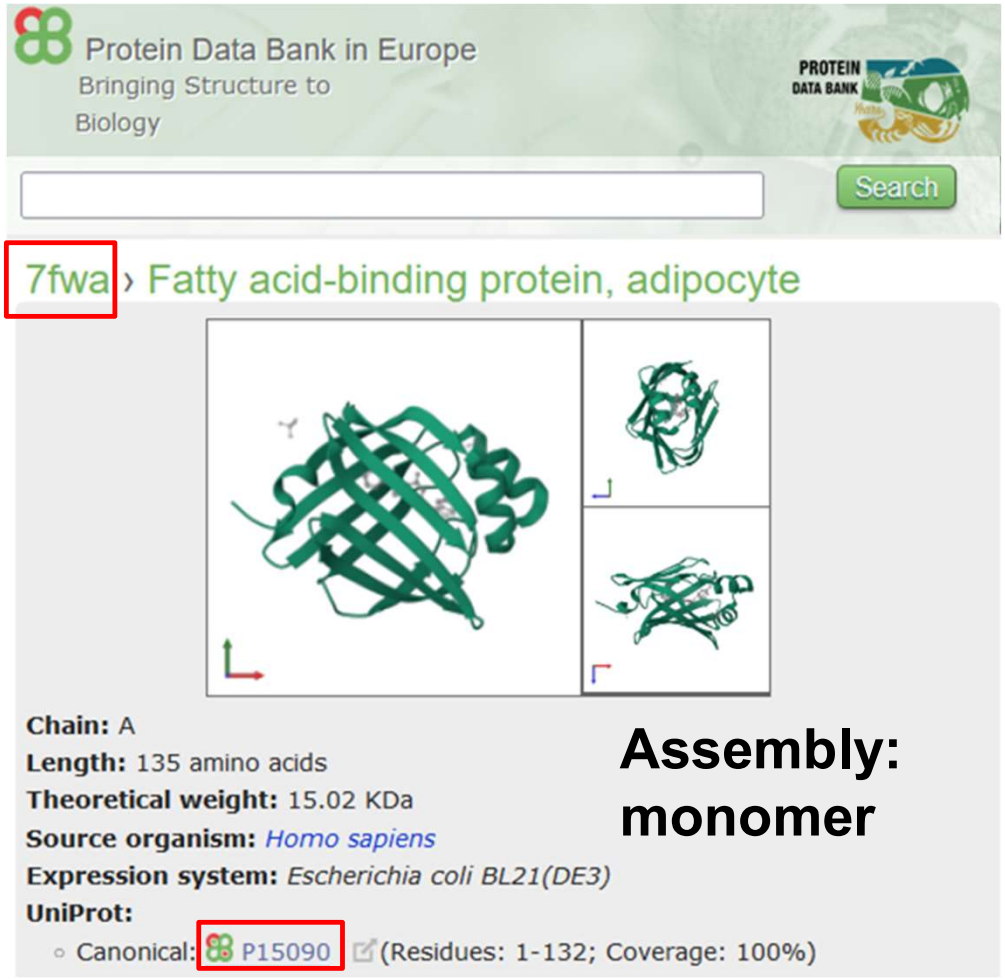
Protein KB pages



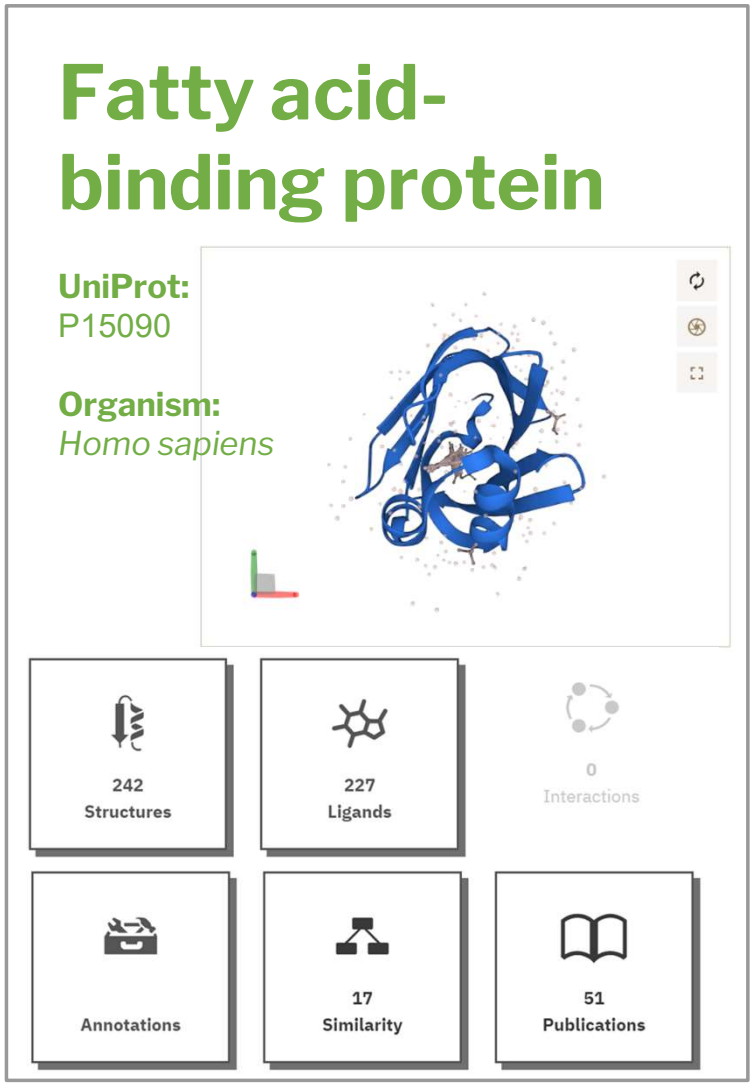
SOURCE: https://www.ebi.ac.uk/pdbe/entry/pdb/7fwa/protein/1



Protein Entry pages



Assembly: monomer

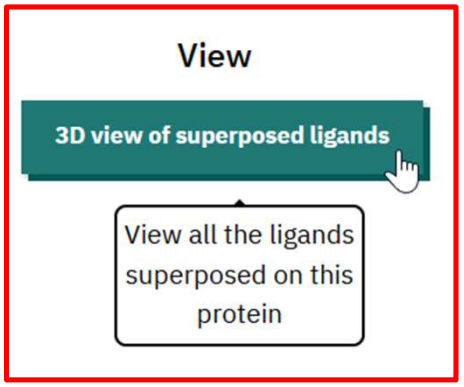


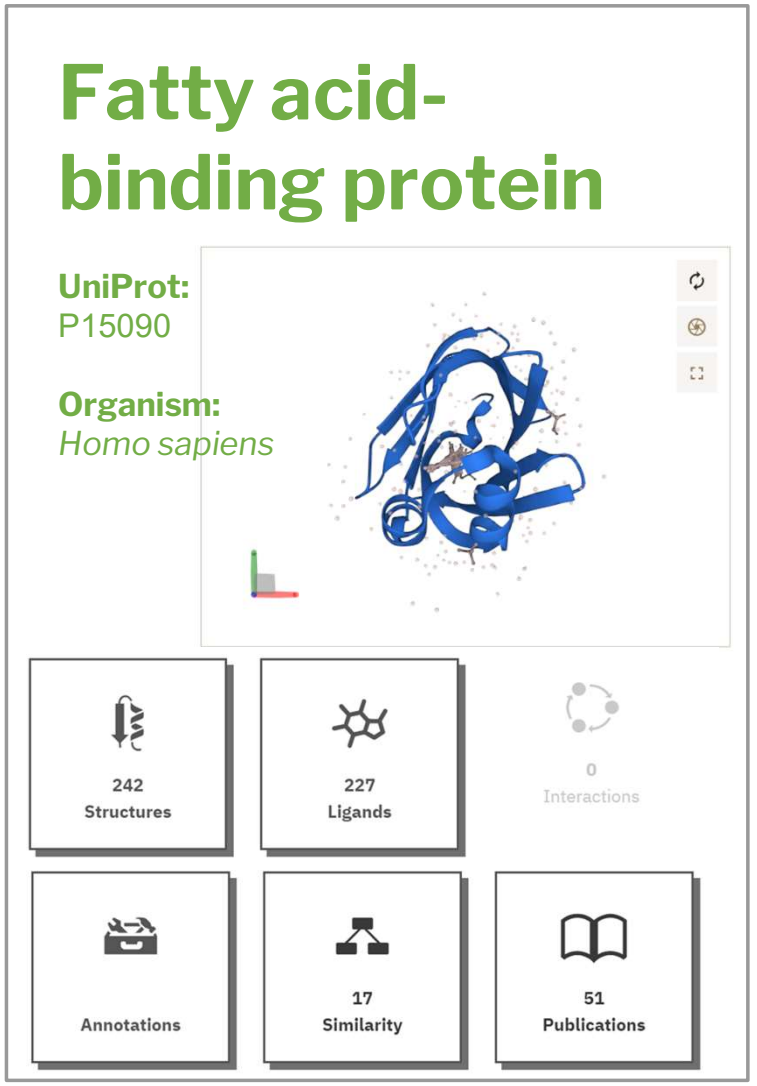
Protein KB pages

- Structural context for a given protein

UniProt

P15090





SOURCE: https://www.ebi.ac.uk/pdbe/pdbe-kb/proteins/P15090/ligands

Protein KB pages -- superposition

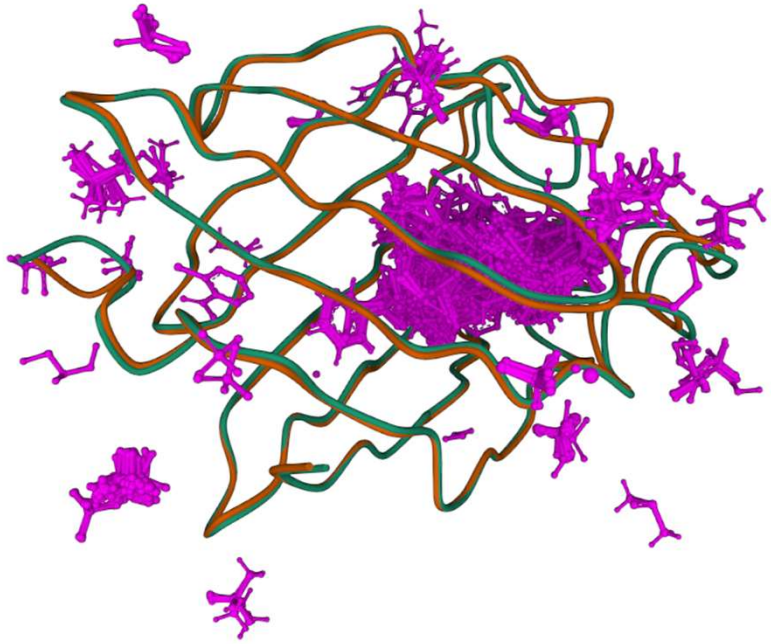
- We have a process that calculates superposition for individual proteins in the whole PDB archive

  → YouTube tutorial:
  https://www.youtube.com/watch?v=fqP50iFZHtY&t=17s

- These data are displayed in:
  'superposed clusters' (orange vs green)

- But:
  The clusters do not necessarily correspond to biological functions

- Allows you to view common interaction sites and important regions of the protein



Fatty acid-binding protein: 242 structures, 227 different small molecules

SOURCE: https://www.ebi.ac.uk/pdbe/pdbe-kb/proteins/P15090

Protein KB pages -- superposition

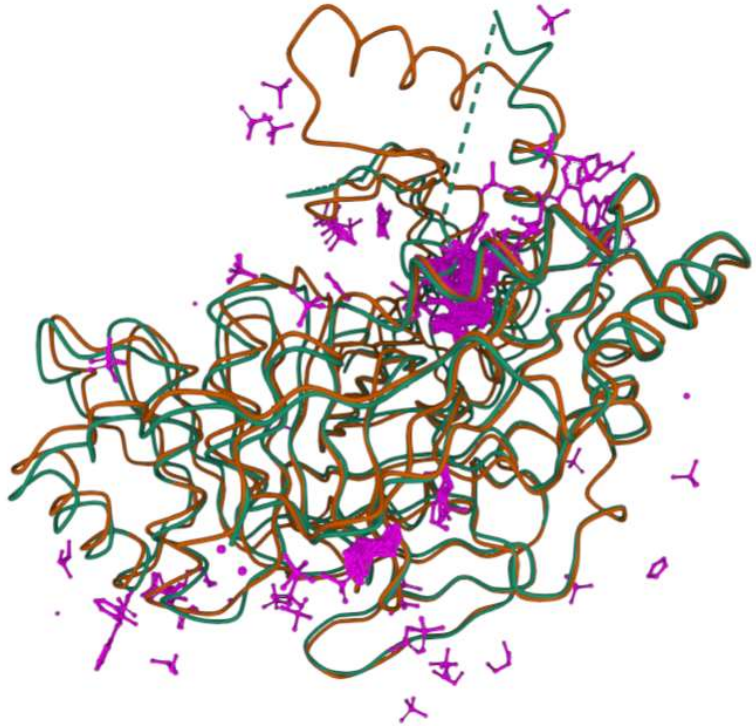
- We have a process that calculates superposition for individual proteins in the whole PDB archive

  → YouTube tutorial:
  https://www.youtube.com/watch?v=fqP50iFZHtY&t=17s

- These data are displayed in:
  'superposed clusters' (orange vs green)

- But:
  The clusters do not necessarily correspond to biological functions

- Allows you to view common interaction sites and important regions of the protein



Salicylate Synthase: 14 structures,
28 different small molecules

SOURCE: https://www.ebi.ac.uk/pdbe/pdbe-kb/proteins/P15090

More information: https://pubs.acs.org/doi/10.1021/bi3002067

EMBL-EBI

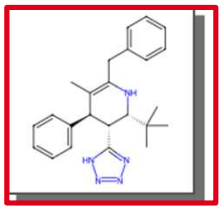
Protein KB pages -- ligands

Overview of ligands and binding residues

Legends: Annotated small molecules

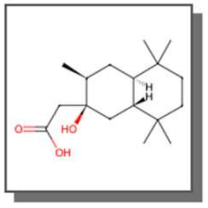
Other small molecules

Not interacting small molecules



UUX

Found in 2 PDB entries



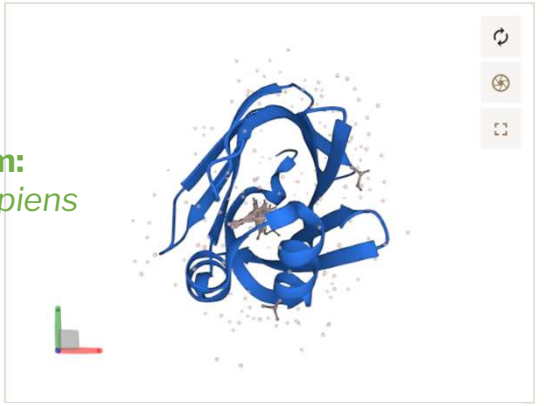
WOO

Found in 1 PDB entry

Fatty acid-binding protein

UniProt:
P15090

Organism:
Homo sapiens



242 Structures

227 Ligands

0 Interactions

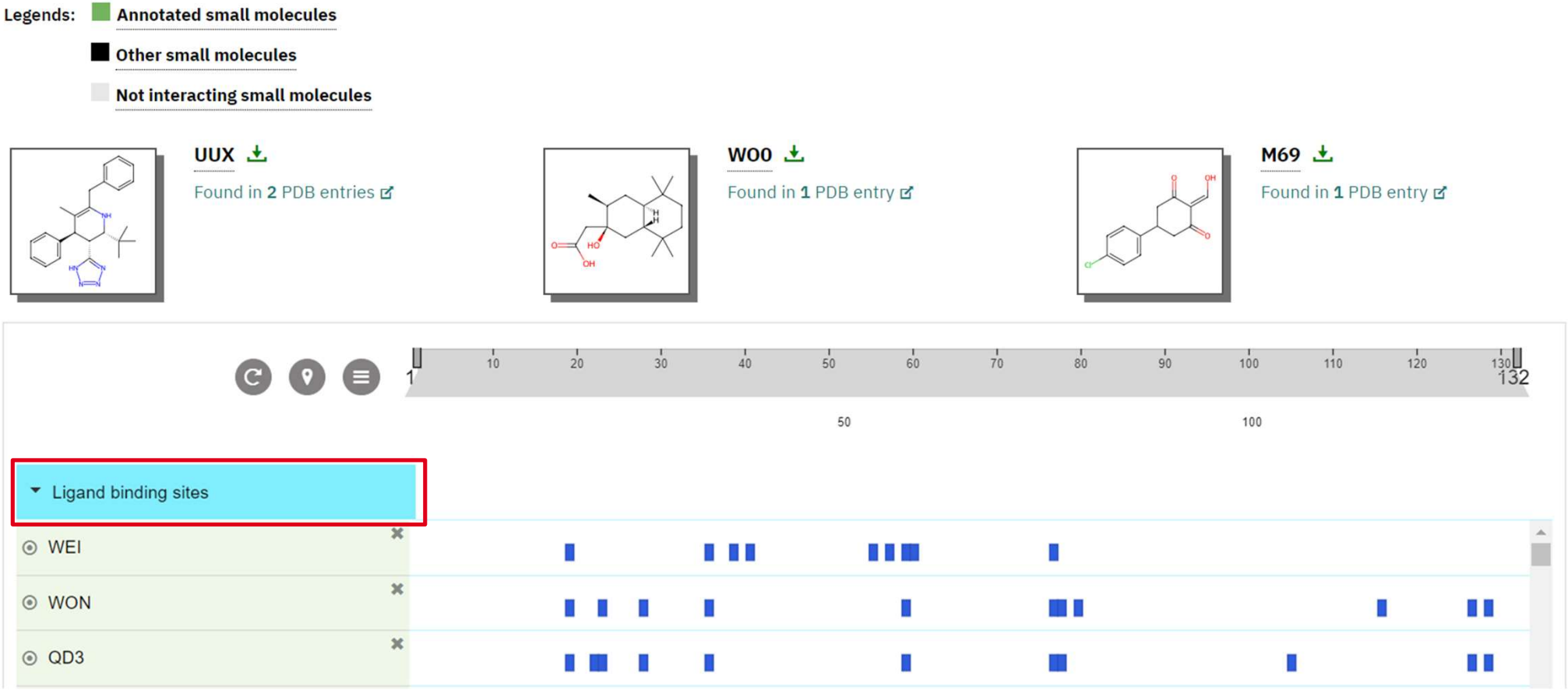
Annotations

17 Similarity

51 Publications

Protein KB pages -- ligands

Overview of ligands and binding residues



Protein KB pages

- Structural context

  UniProt ID

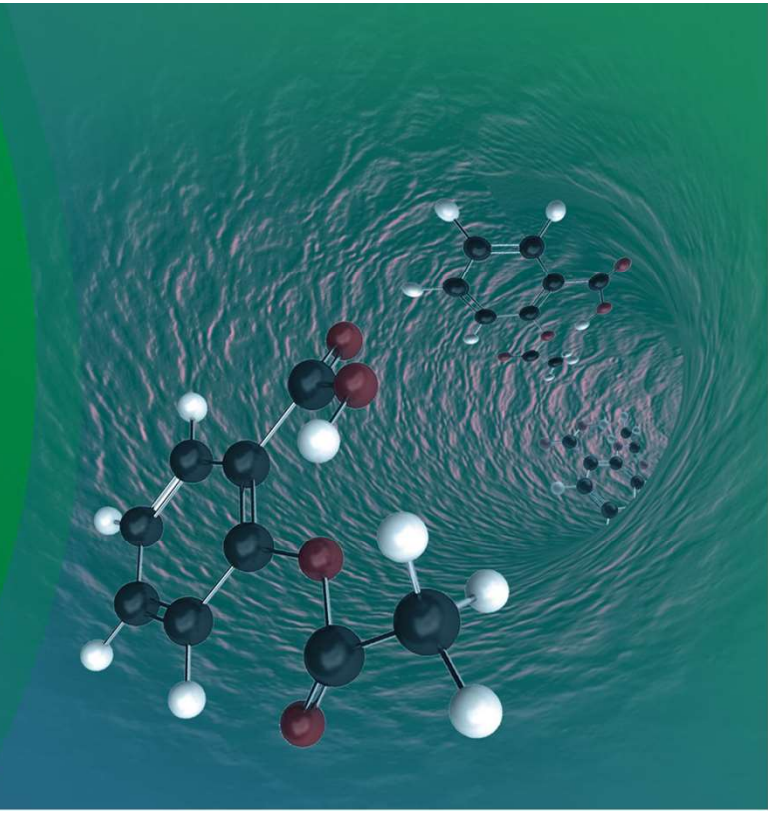
  = specific protein / protein component

  from a specific organism

- Information clustered in different sections
  - Structure coverage
  - Ligands and binding sites
  - Protein interactions
  - Additional annotations
  - Proteins with similar sequence
  - Publications

We've used APIs and our analysis pipeline to pre-cluster & analysis multiple structures for you ☺

Acknowledgements



Genevieve Evans

PDBe

pdbhelp@ebi.ac.uk
proteindatabank
@PDBeurope
proteindatabank
pdbeurope
pdbart

PDBe
Protein Data Bank in Europe

EMBL-EBI

Feedback Welcome
Thank you!

PDBe.org


pdbhelp@ebi.ac.uk


proteindatabank


@PDBeurope


proteindatabank


pdbeurope


pdbart

PDBe
Protein Data Bank in Europe

EMBL-EBI


EMBL-EBI