PDBe resources to help with starting model selection for molecular dynamics simulations

Sudakshina Ganguly

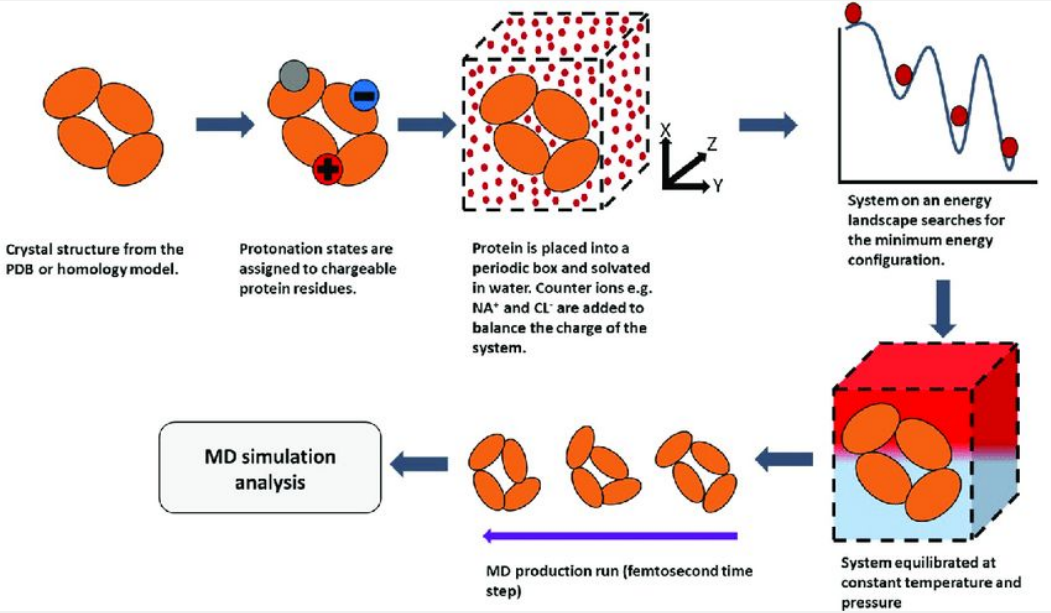
Scientific Database Curator

PDBe
Protein Data Bank in Europe

sganguly@ebi.ac.uk

EMBL-EBI

Molecular Dynamics Simulations



DOI: 10.3390/pharmaceutics10040165

EMBL-EBI

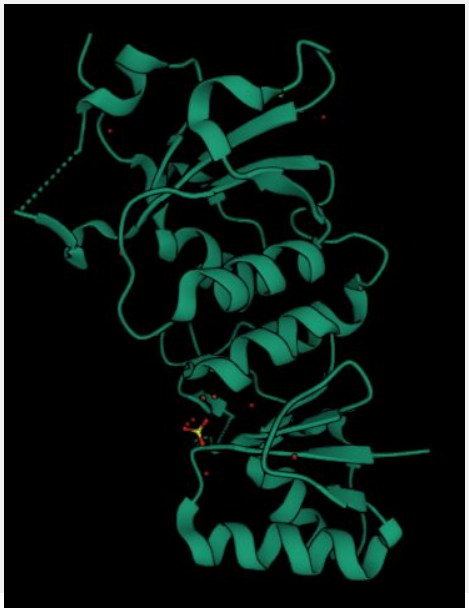
Starting model

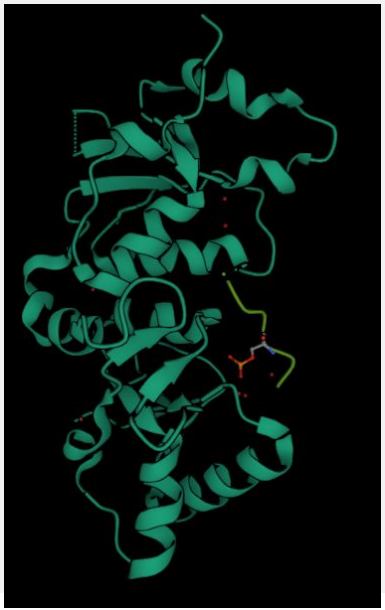
- Why is a starting model important?
- An example where different models led to different results?

PDB id 3pxc

BRCA BRCT protein

UniProt: P38398





PDB id 3pxe

BRCA BRCT protein

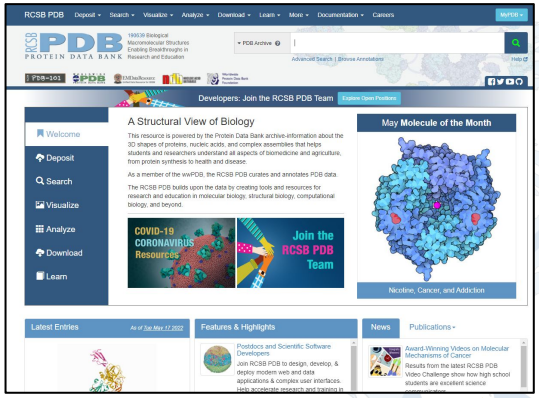
UniProt: P38398

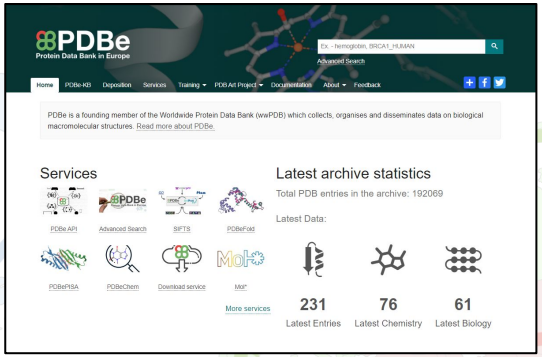
Relevant difference in clashes

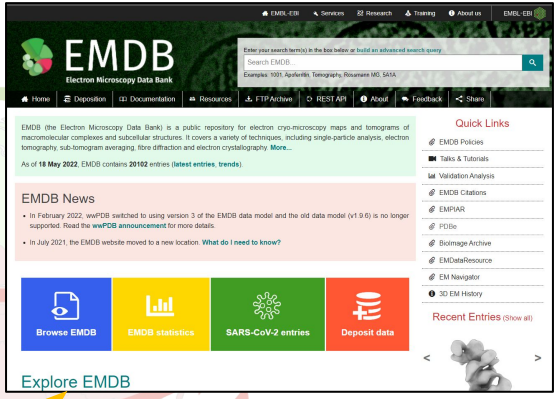
Objectives of this seminar

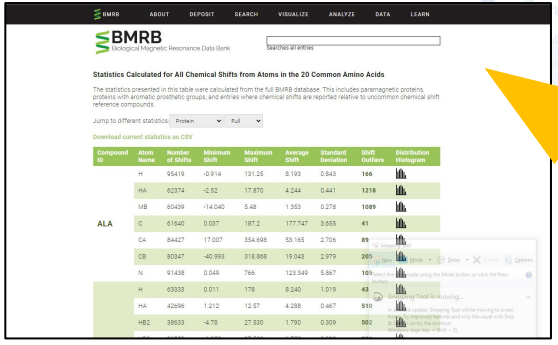
- Learn where and how to search for a structure
- Understanding the validation metrics
- Assimilate methods to visualise and compare structures
- Use an existing MD frame as an initial structure

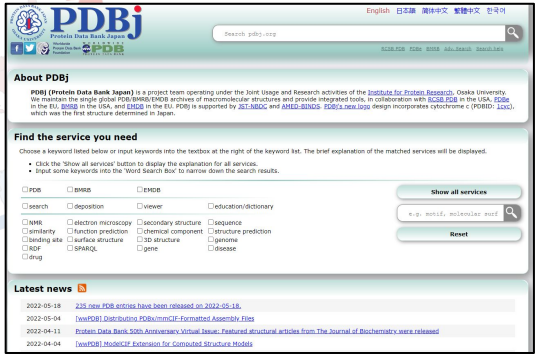
The Worldwide Protein Data Bank (wwPDB)

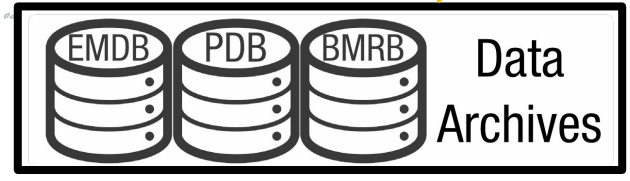






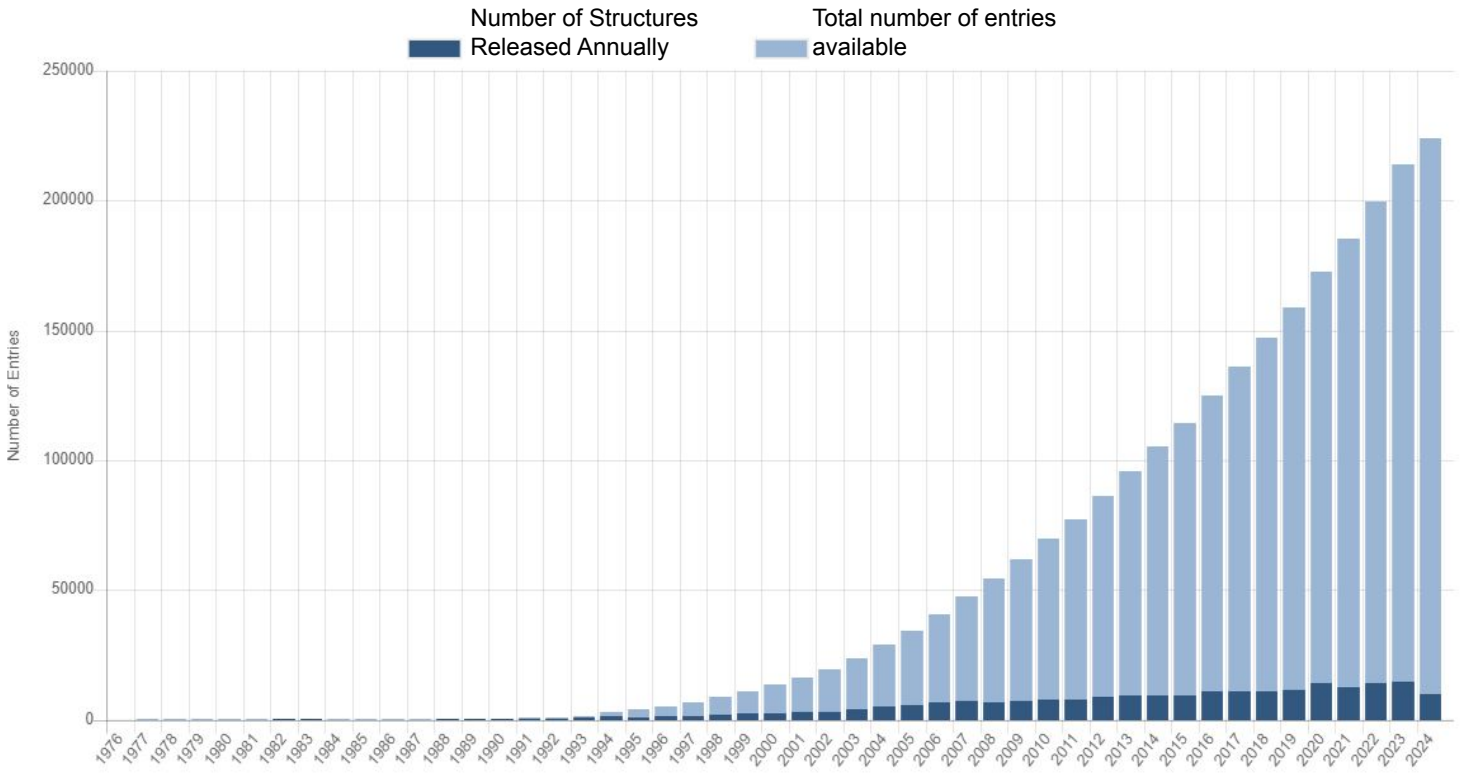






EMBL-EBI

How many structures in the PDB?



224572 structures as of today, 10th Sep 2024

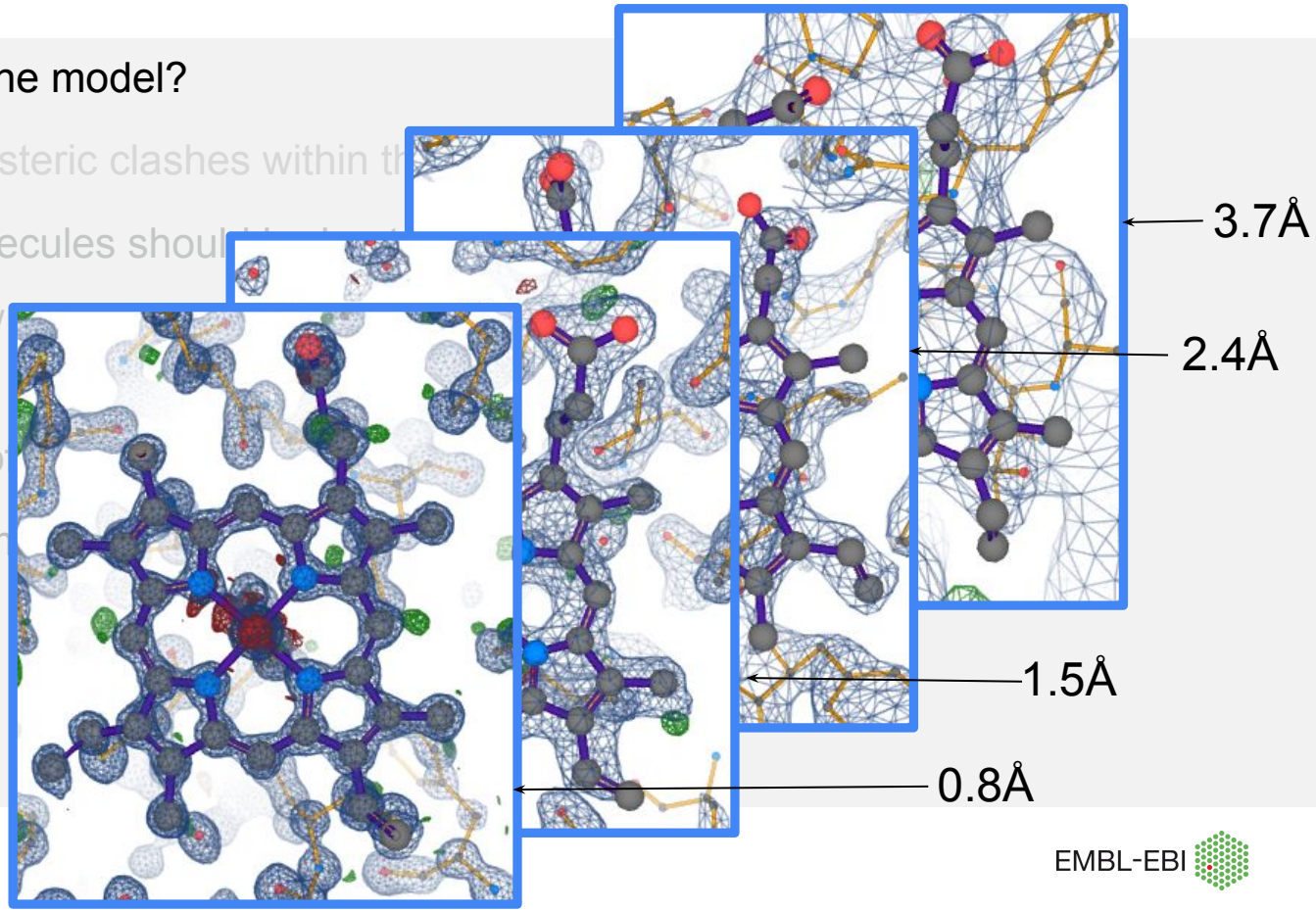
EMBL-EBI

Important points for selecting the starting model

- What is the resolution for the model?
- Are there any outstanding steric clashes within the structure?
- Assess whether water molecules should be kept in the structure or stripped.
- Are there missing regions / loops of the protein and do they have to be modelled using another tool?
- What is the conformation of the protein?
- What is the experimental method used to derive the structure?

Important points for selecting the starting model

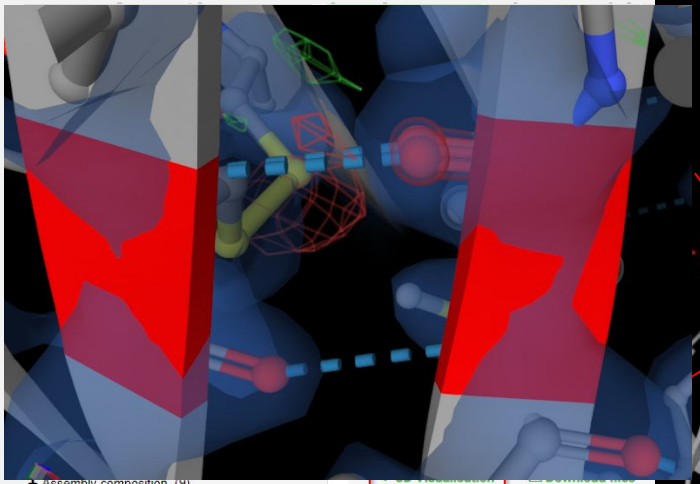
- What is the resolution for the model?
- Are there any outstanding steric clashes within th
- Assess whether water molecules shoul
- Are there missing regions / another tool?
- What is the conformation o
- What is the experimental m

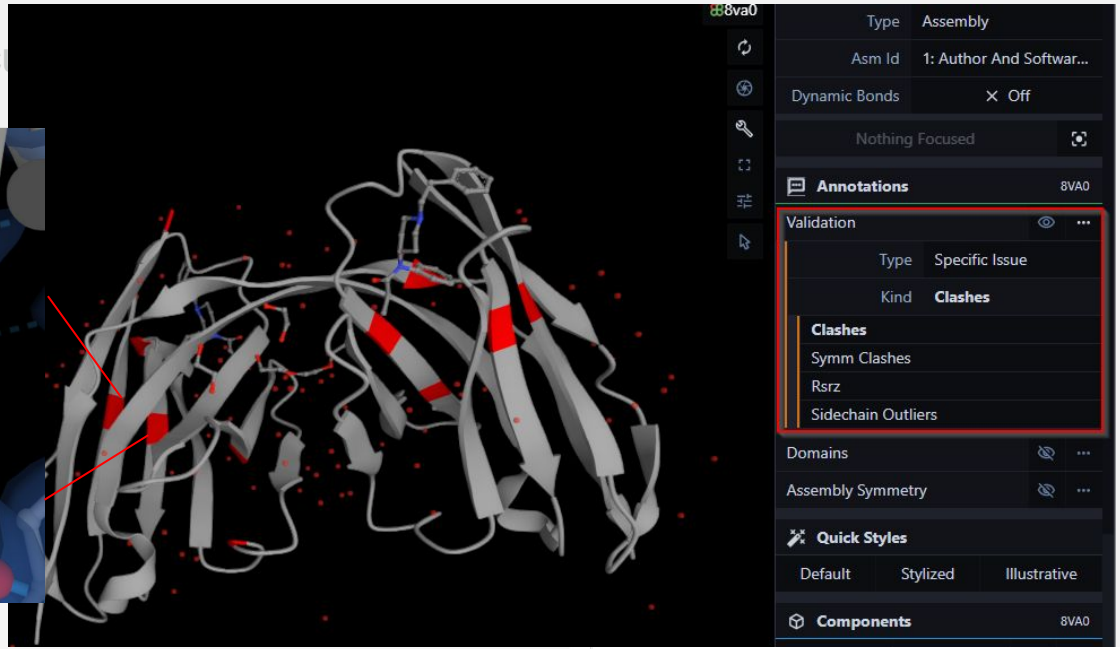


EMBL-EBI

Important points for selecting the starting model

- What is the resolution for the model?
- Are there any outstanding steric clashes within the structure?
- Assess whether water molec



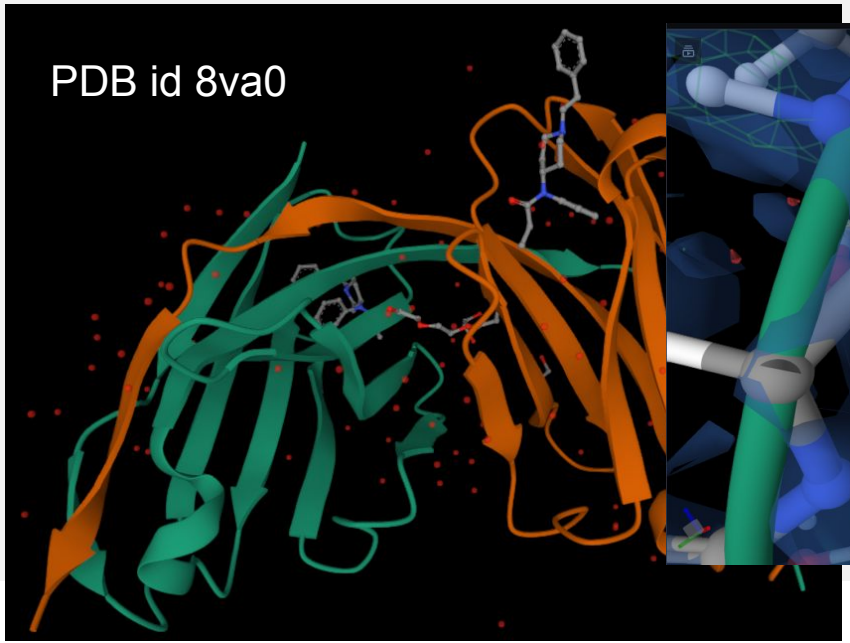


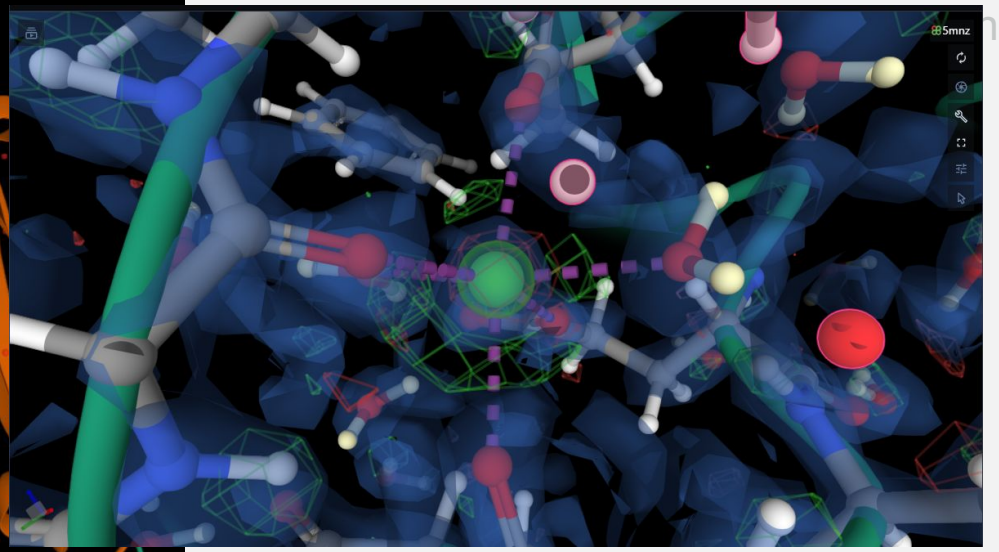
+ Assembly composition (9)
+ Assembly polymer count (42)

EMBL-EBI

Important points for selecting the starting model

- What is the resolution for the model?
- Are there any outstanding steric clashes within the structure?
- Assess whether water molecules should be kept in the structure or stripped.

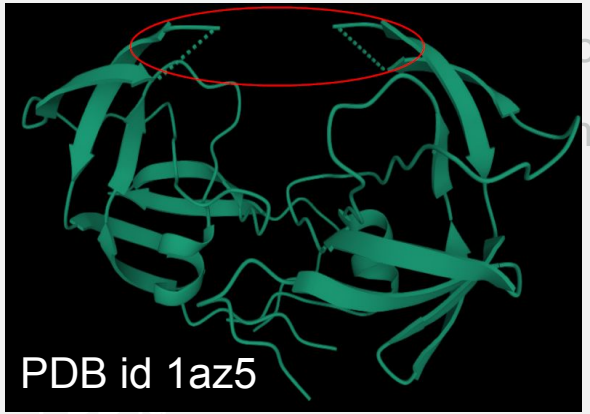




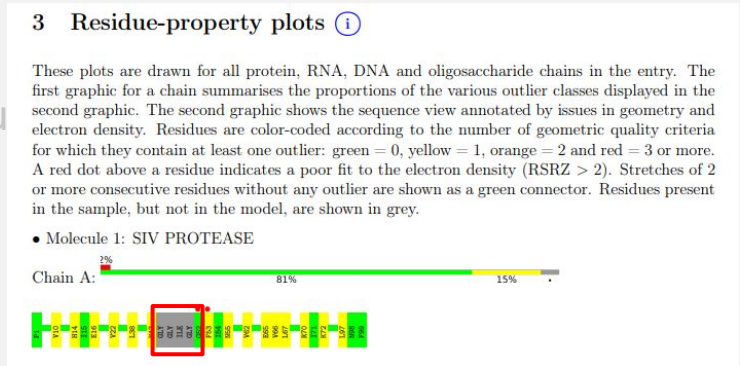
PDB id 5mnz

Important points for selecting the starting model

- What is the resolution for the model?
- Are there any outstanding steric clashes within the structure?
- Assess whether water molecules should be kept in the structure or stripped.
- Are there missing regions / loops of the protein and do they have to be modelled using another tool?




```
loop_
_pdbx_unobs_or_zero_occ_residues.id
_pdbx_unobs_or_zero_occ_residues.PDB_model_num
_pdbx_unobs_or_zero_occ_residues.polymer_flag
_pdbx_unobs_or_zero_occ_residues.occupancy_flag
_pdbx_unobs_or_zero_occ_residues.auth_asym_id
_pdbx_unobs_or_zero_occ_residues.auth_comp_id
_pdbx_unobs_or_zero_occ_residues.auth_seq_id
_pdbx_unobs_or_zero_occ_residues.PDB_ins_code
_pdbx_unobs_or_zero_occ_residues.label_asym_id
_pdbx_unobs_or_zero_occ_residues.label_comp_id
_pdbx_unobs_or_zero_occ_residues.label_seq_id
1 1 Y 1 A GLY 48 ? A GLY 48
2 1 Y 1 A GLY 49 ? A GLY 49
3 1 Y 1 A ILE 50 ? A ILE 50
4 1 Y 1 A GLY 51 ? A GLY 51
```



3 Residue-property plots

These plots are drawn for all protein, RNA, DNA and oligosaccharide chains in the entry. The first graphic for a chain summarises the proportions of the various outlier classes displayed in the second graphic. The second graphic shows the sequence view annotated by issues in geometry and electron density. Residues are color-coded according to the number of geometric quality criteria for which they contain at least one outlier: green = 0, yellow = 1, orange = 2 and red = 3 or more. A red dot above a residue indicates a poor fit to the electron density (RSRZ > 2). Stretches of 2 or more consecutive residues without any outlier are shown as a green connector. Residues present in the sample, but not in the model, are shown in grey.

• Molecule 1: SIV PROTEASE

Chain A: 3% 81% 15%

EMBL-EBI

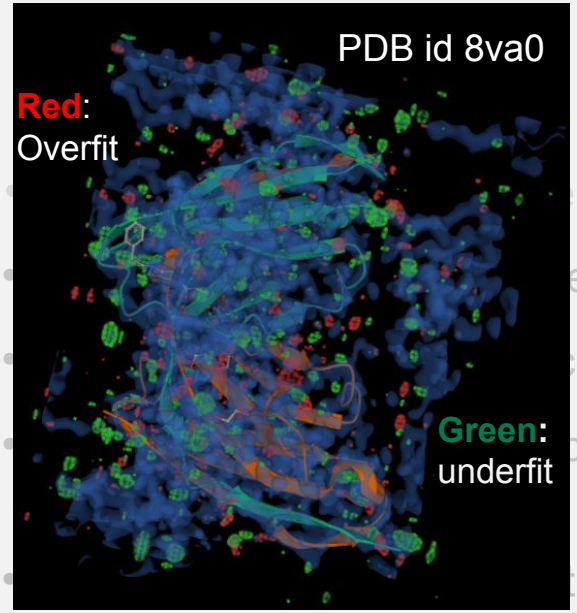
Important points for selecting the starting model

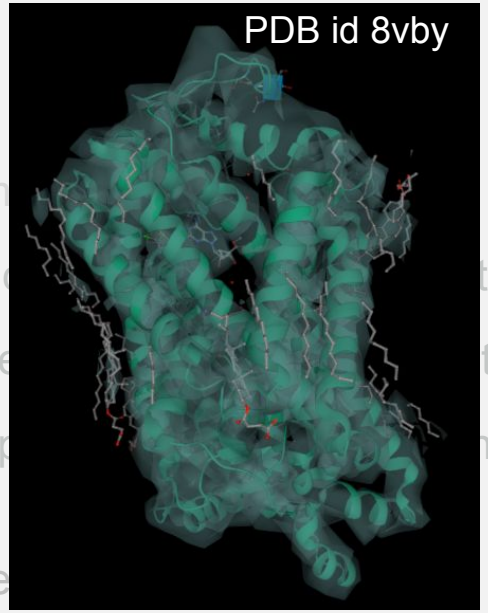
- What is the resolution for the model?
- Are there any outstanding steric clashes within the structure?
- Assess whether water molecules should be kept in the structure
- Are there missing regions / loops of the protein and do they have another tool?
- What is the conformation of the protein?
- What is the experimental method used to

Serine protease 1

UniProt: **P00760**

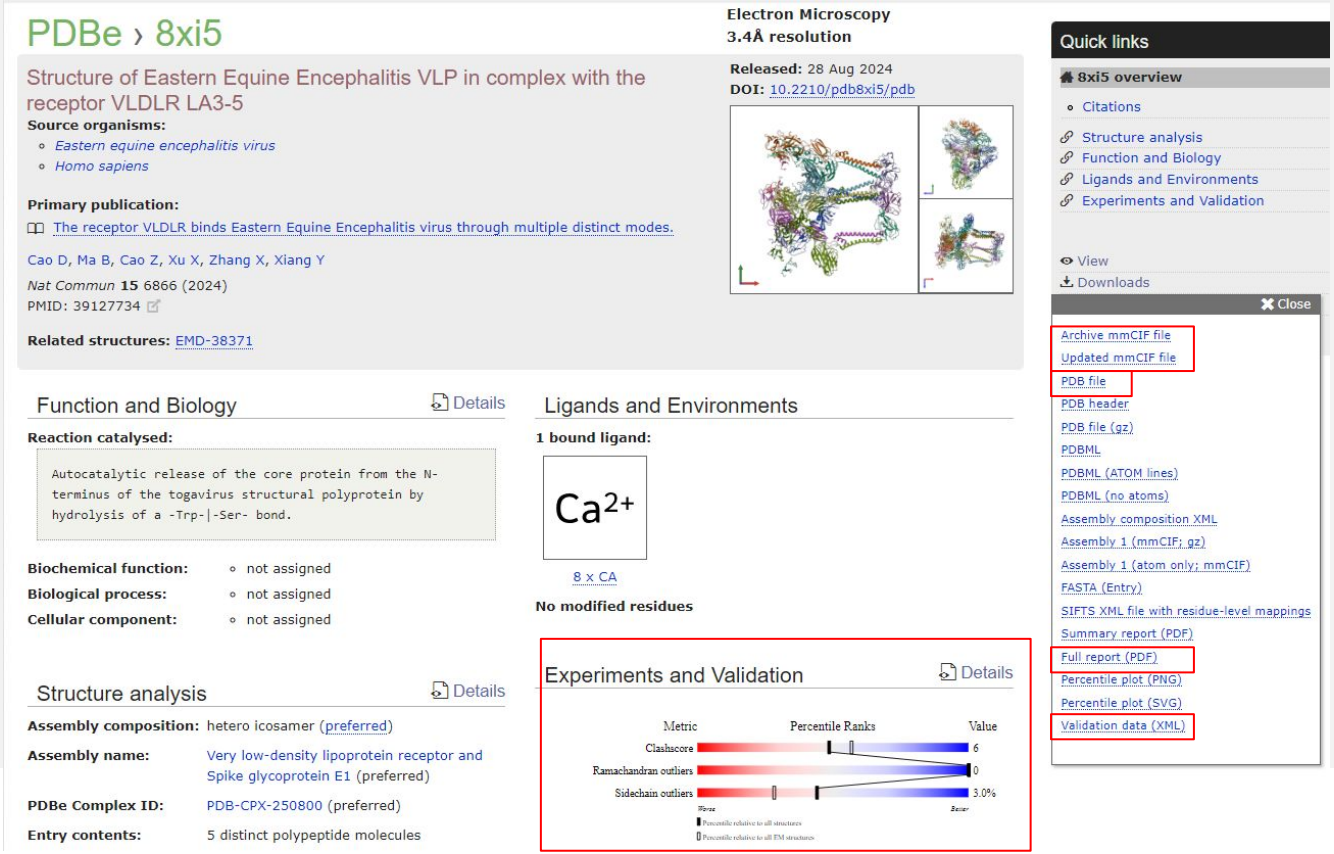
Important points for selecting the starting model





- What is the experimental method used to derive the structure?

Validation sliders: what you need to know



PDB format is not being supported any longer

mmCIF format is the new '**master**' format!

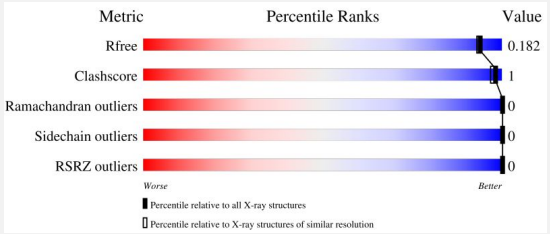
More info on the mmCIF format - https://mmcif.wwpdb.org/docs/user-guide/guide.html

EMBL-EBI

Validation sliders: what you need to know

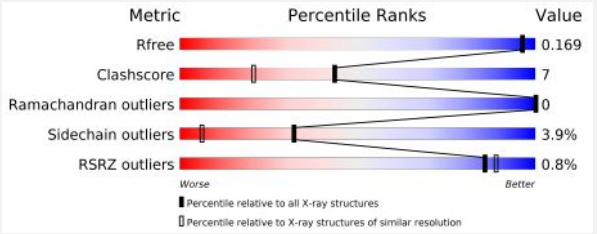
- Overall quality of a specific structure can be assessed
- Overall assessment of the quality of the modelling of the structure compared to a number of key validation metrics
- Includes both geometric validation and fit of the model to the experimental data

PDB id 2xhh



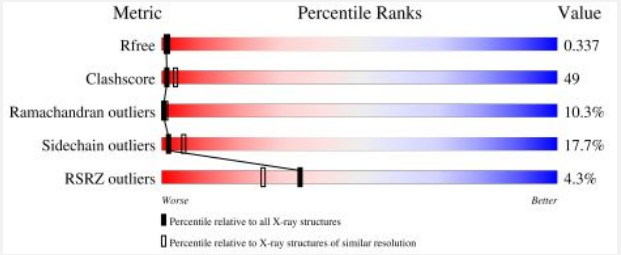
High quality

PDB id 5ovo



Medium quality

PDB id 3lnn



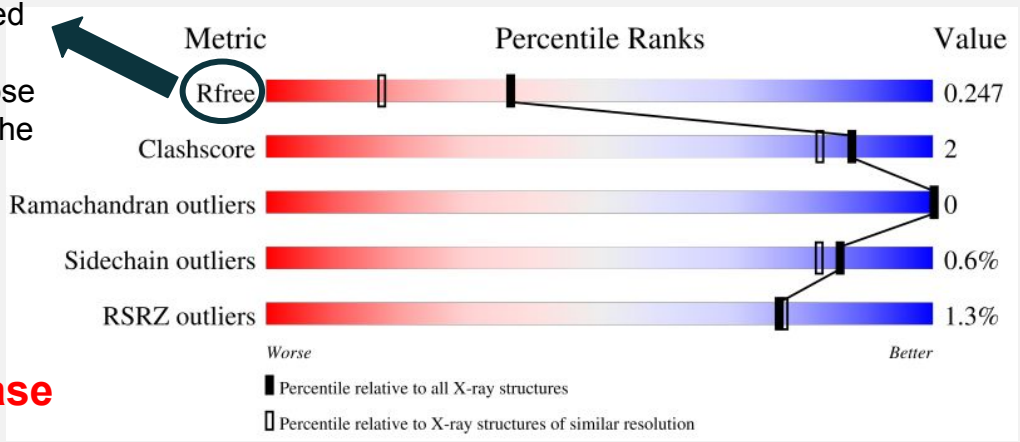
Low quality

EMBL-EBI

Validation sliders: what you need to know

measures correlation between observed structure factor amplitudes & those calculated from the model

PDB id 8va0



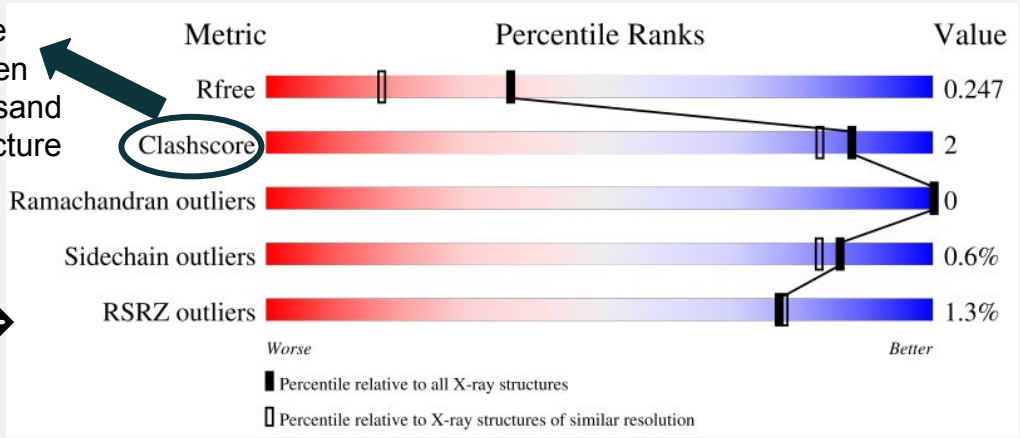
Rfree increase ⇒ quality decrease

Rfree is specific for X-ray diffraction entries

Validation sliders: what you need to know

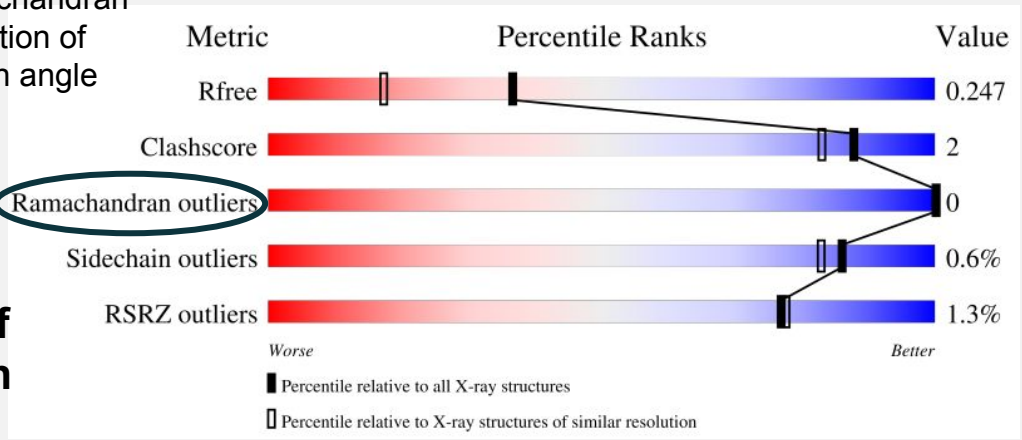
quantifies the number of close contacts between atoms per thousand atoms of a structure

Clashcore increases ⇒ quality decreases



Validation sliders: what you need to know

A standard residue is identified as a Ramachandran outlier if the combination of backbone φ-ψ torsion angle values is unusual

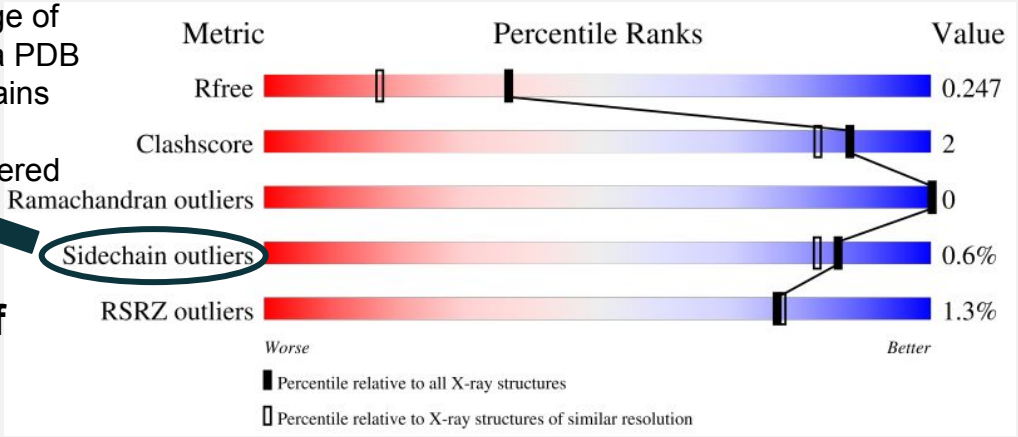


Low number of Ramachandran outliers ⇒ higher quality

EMBL-EBI

Validation sliders: what you need to know

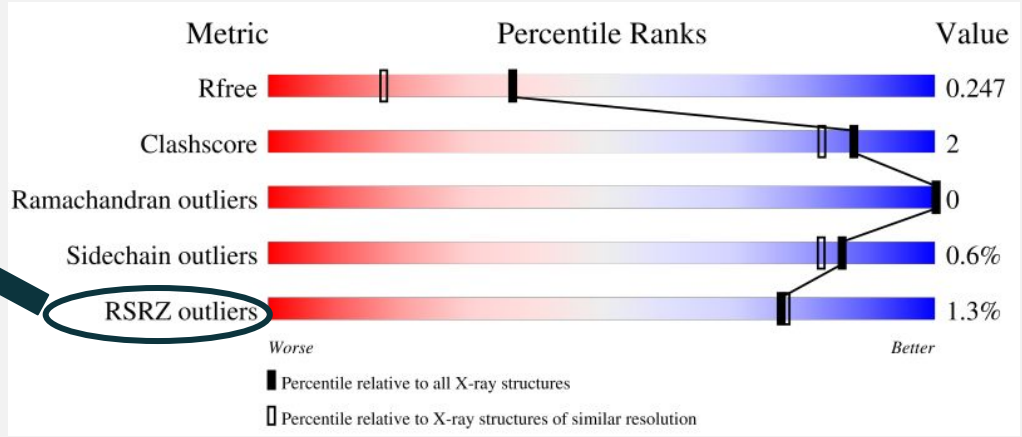
Represents percentage of standard residues in a PDB structure with side chains whose torsion angle combination is considered to be an outlier



Low number of sidechain outliers ⇒ higher quality

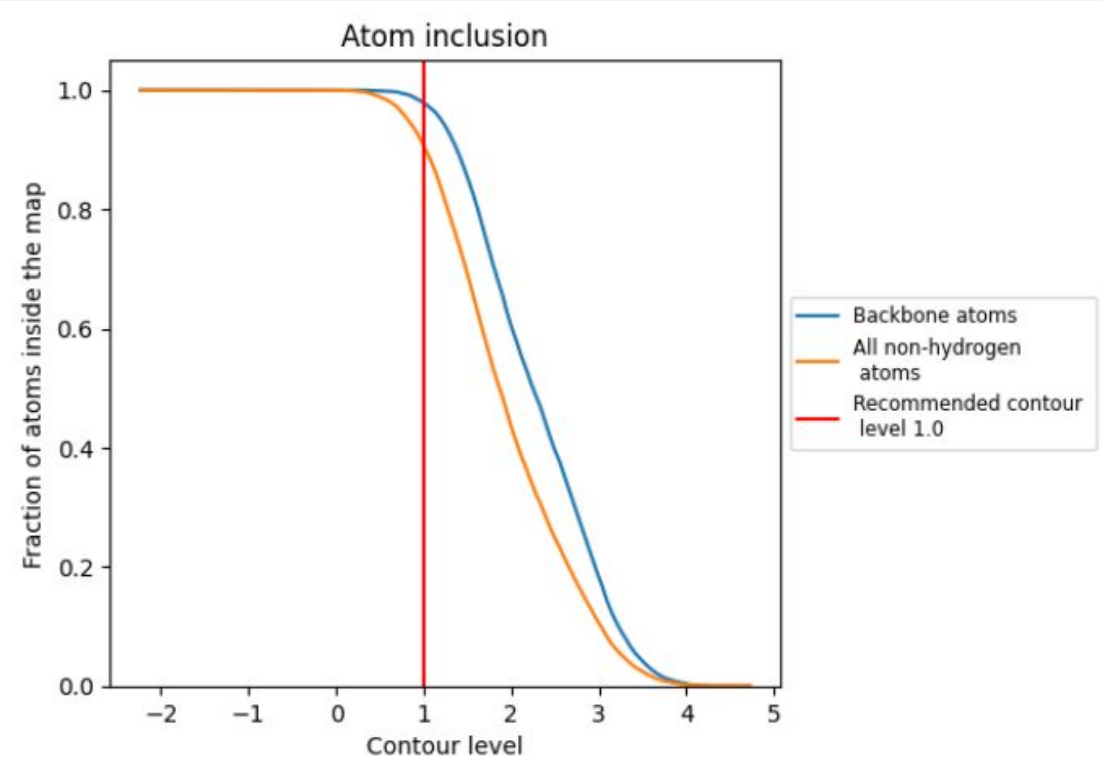
Validation sliders: what you need to know

Quantify deviation of bond angles and bond lengths in standard residues of a PDB structure



Low number of RSRZ outliers ⇒ higher quality

Data quality for EM structures



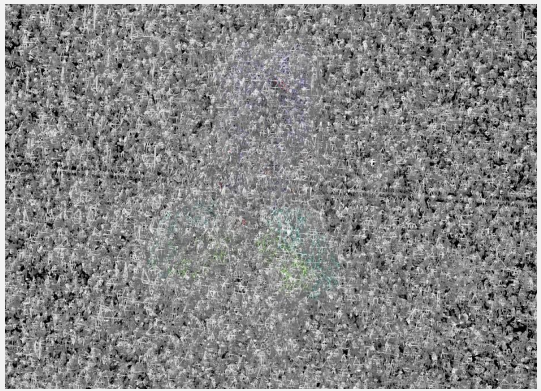
At the recommended contour level, 98% of all backbone atoms, 91% of all non-hydrogen atoms, are inside the map.

- Displays graphically the fraction of atoms that fit within the map at different map contour levels
- Also displays the recommended contour level for this map, as provided by the depositor at deposition.

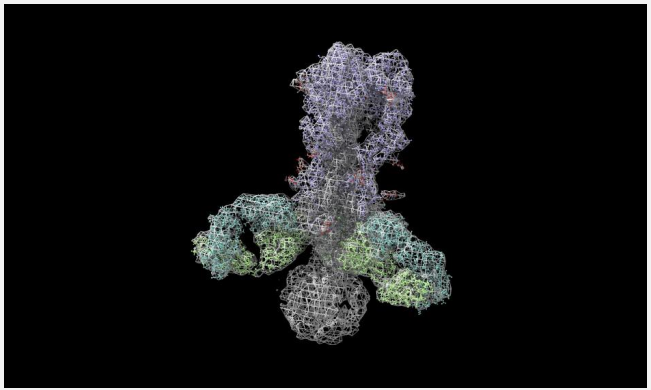
EMBL-EBI

Data quality for EM structures

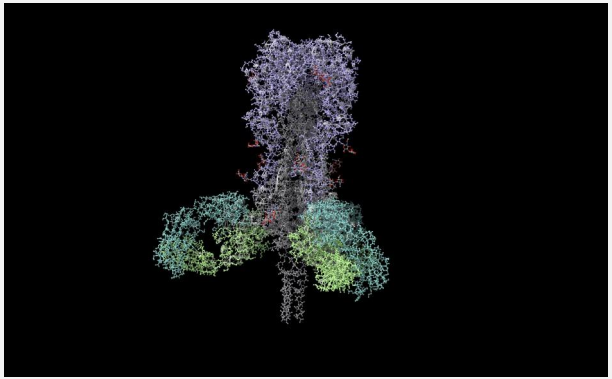
PDB id 6hjq



Contour level 0.2



Contour level 1

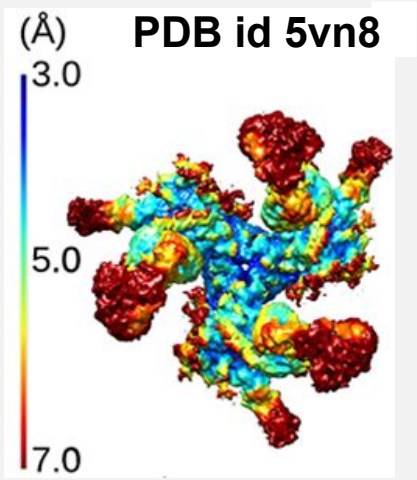


Contour level 2

Comparison of the map model fit at these different contour levels gives an indication of how changes in the contour level affect the apparent fit of atoms within the map.

Data quality for EM structures

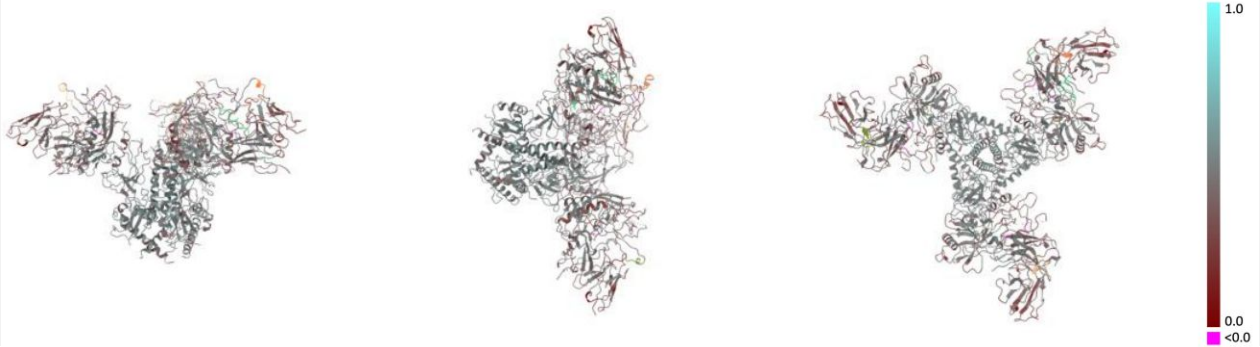
It is important to understand the concept of **global** vs **local** resolution for cryo-EM structures



Vilas, J.L. et al. Local resolution estimates of cryoEM reconstructions. *Curr Opin Struct Biol.* 2020, 64:74-78

Q-score: calculated directly from map values around an atom's position

PDB id 5vn8

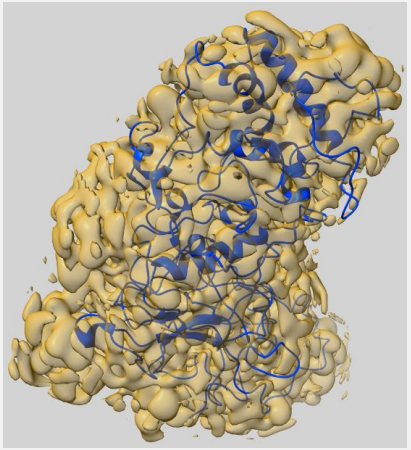


EMBL-EBI

Data quality for EM structures

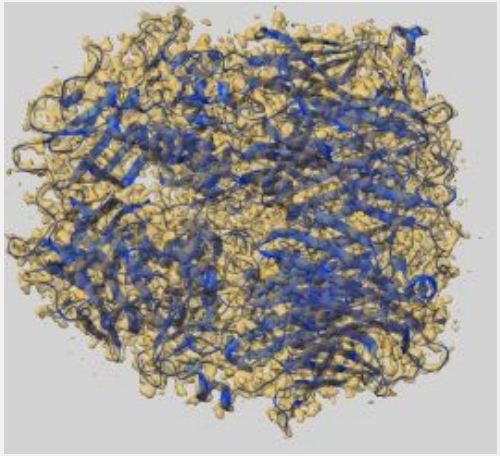
Higher Q-score values reflect better resolvability

PDB id 8fd5



Model	Metric	Evaluation bar	Average value
8fd5	Q-score	Worse — Better	0.006

◆ : Percentile relative to EM structures of ± 1 Å ◇ : Percentile relative to all EM structures



PDB id 9ikc

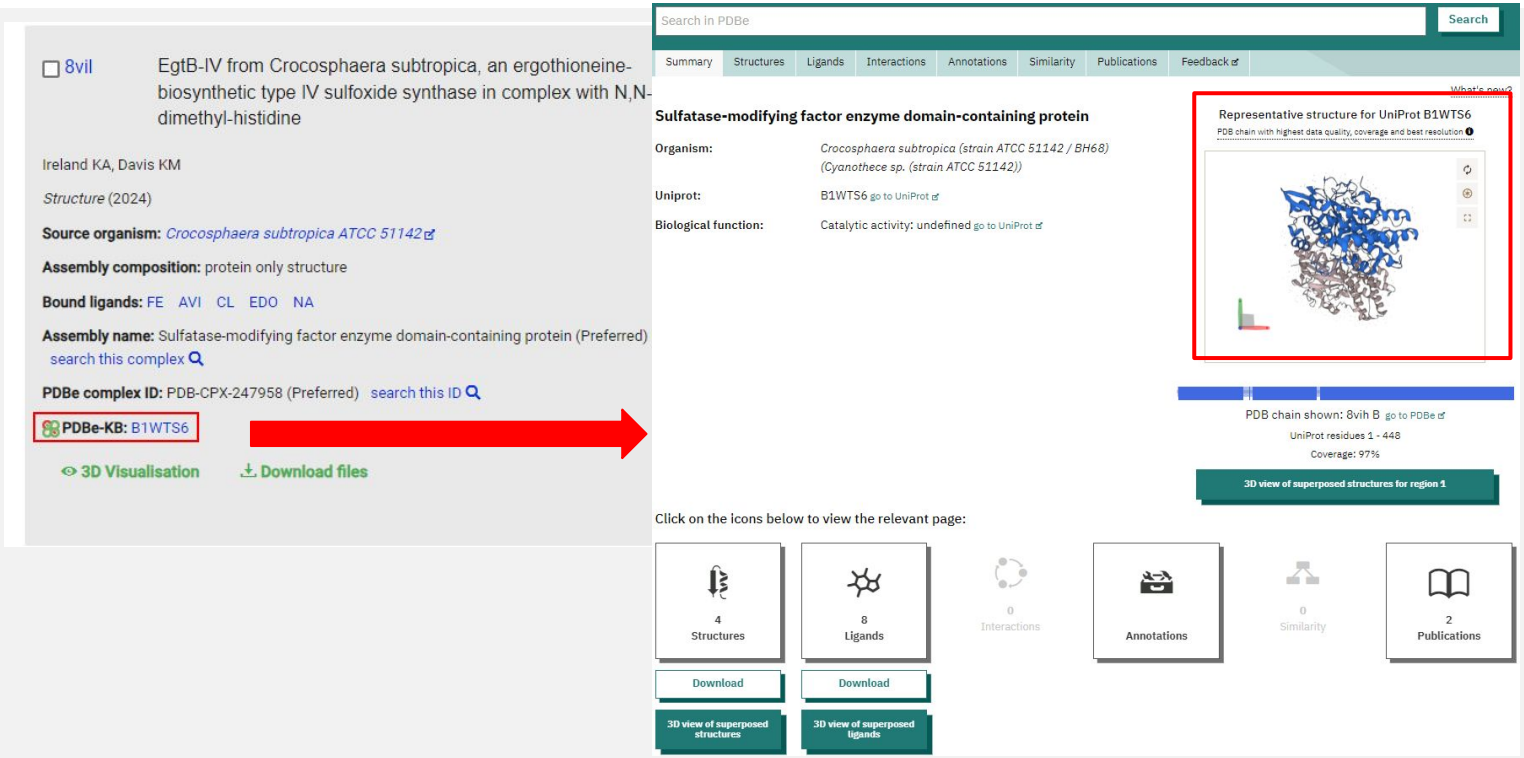
Model	Metric	Evaluation bar	Average value
9ikc	Q-score	Worse — Better	0.763

◆ : Percentile relative to EM structures of ± 1 Å ◇ : Percentile relative to all EM structures

https://www.ebi.ac.uk/emdb/

EMBL-EBI

PDBe-KB: Comparing structures in the PDB

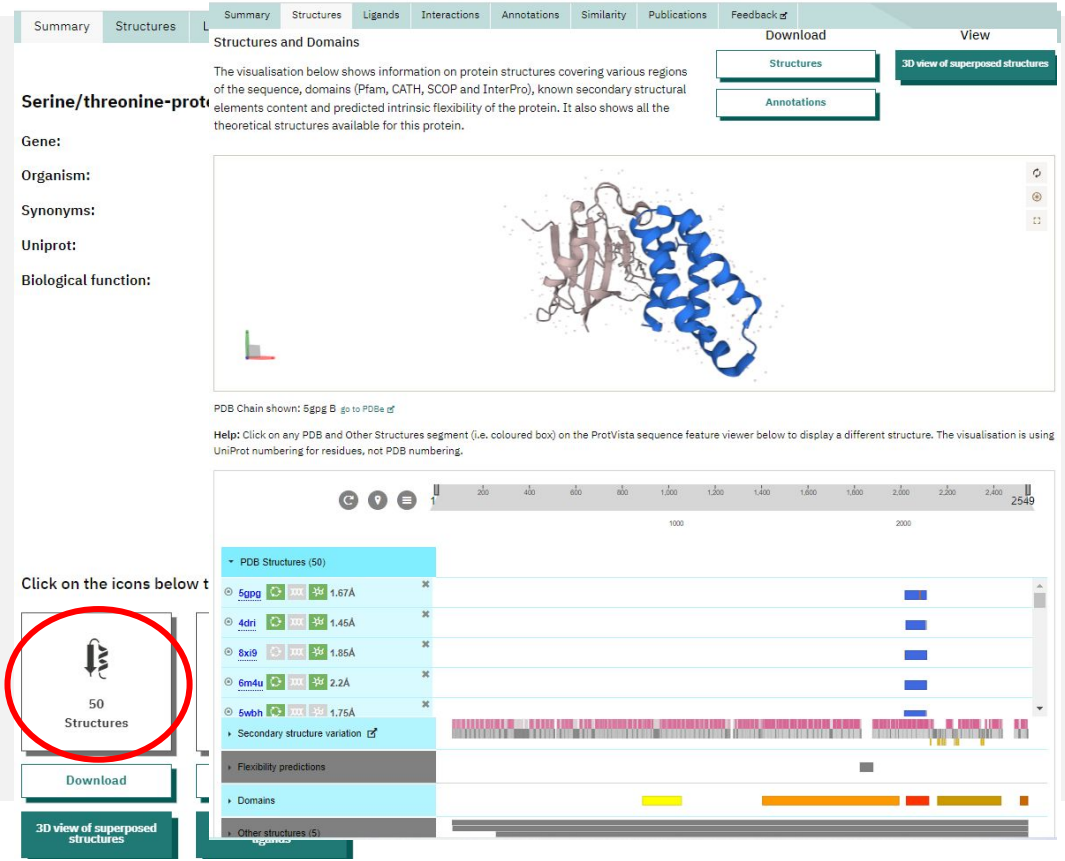


PDBe-KB selects the best representative structure based on the PDB chains with highest data quality, coverage and resolution

https://www.ebi.ac.uk/pdbe/pdbe-kb/

EMBL-EBI

PDBe-KB: Comparing structures in the PDB

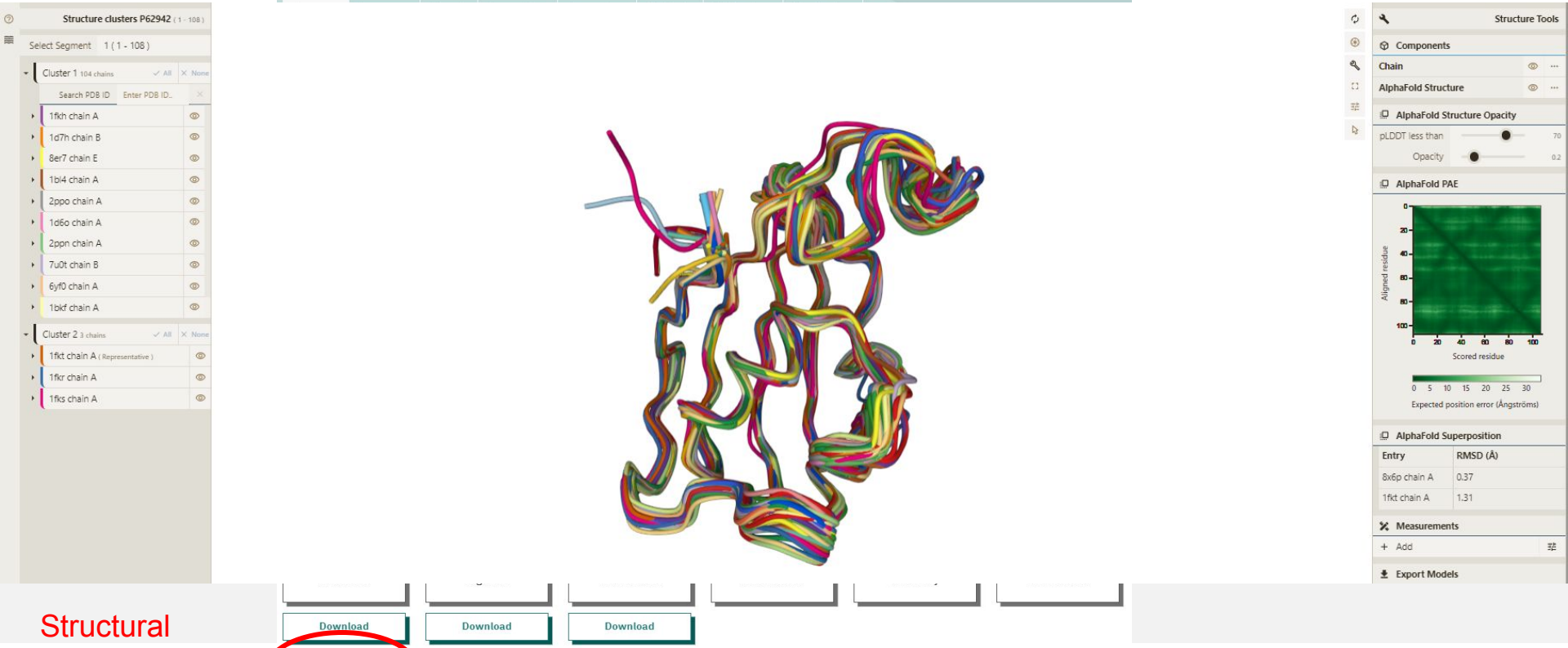


All the PDB entries with this protein

https://www.ebi.ac.uk/pdbe/pdbe-kb/

EMBL-EBI

PDBe-KB: Comparing structures in the PDB

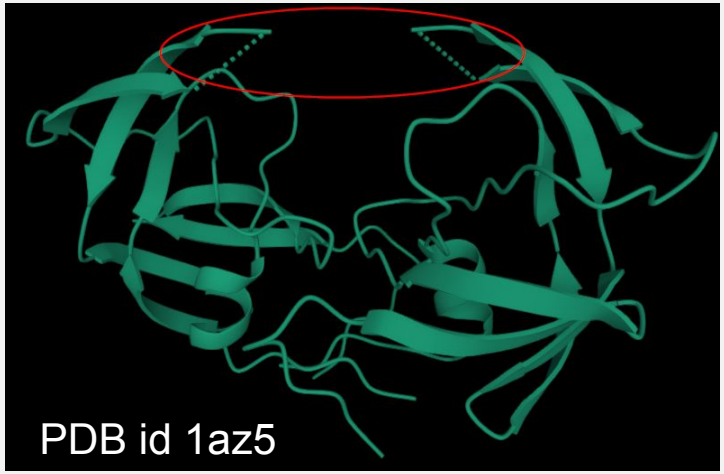


Structural clusters of the protein

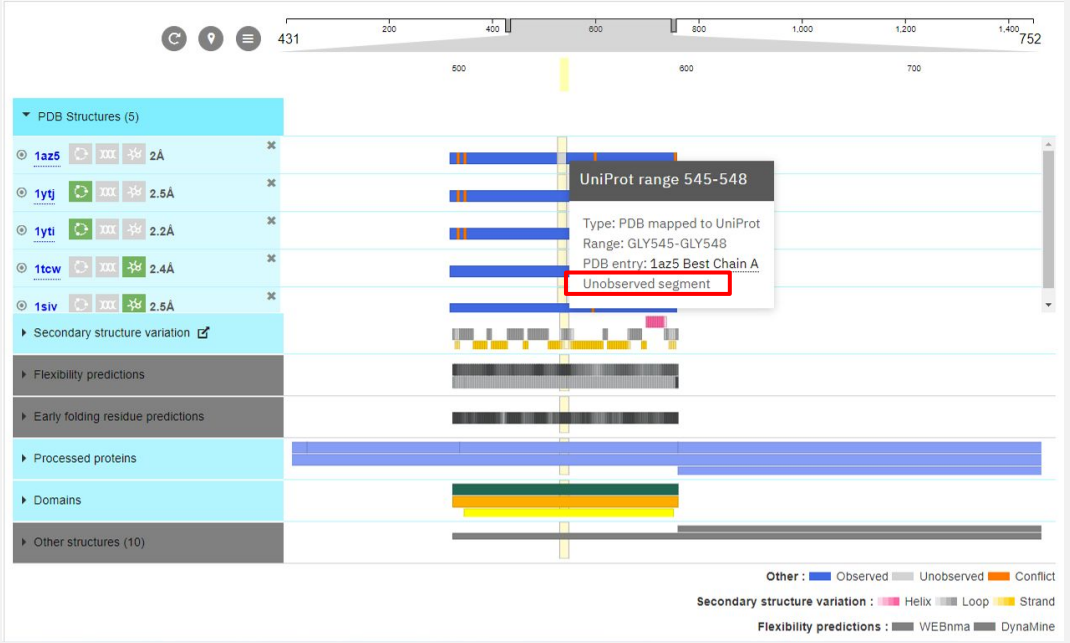
https://www.ebi.ac.uk/pdbe/pdbe-kb/

EMBL-EBI

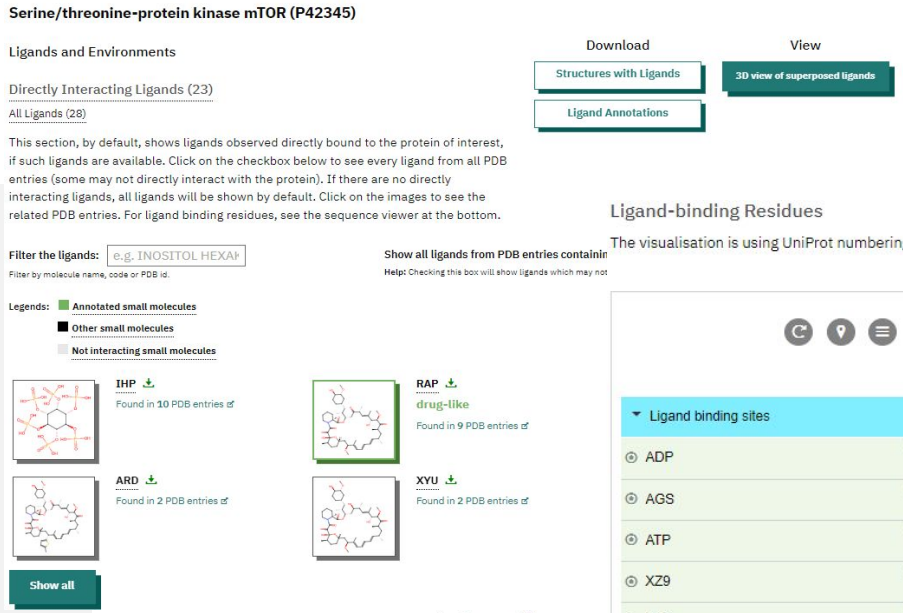
PDBe-KB: Comparing structures in the PDB



1az5

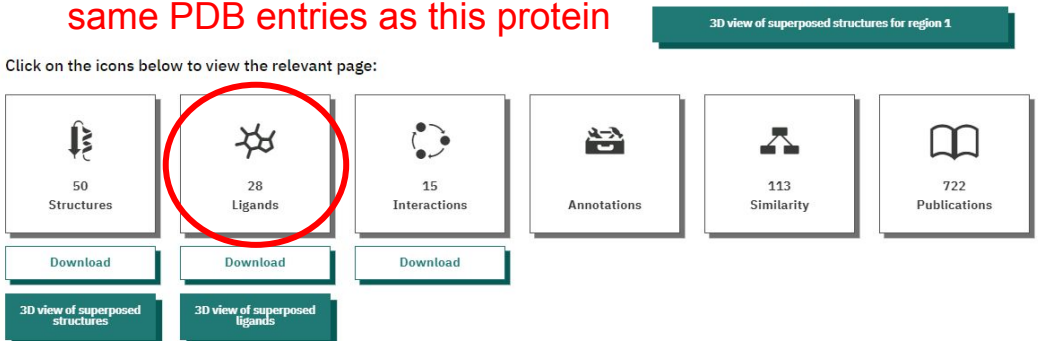


EMBL-EBI



structures in the PDB

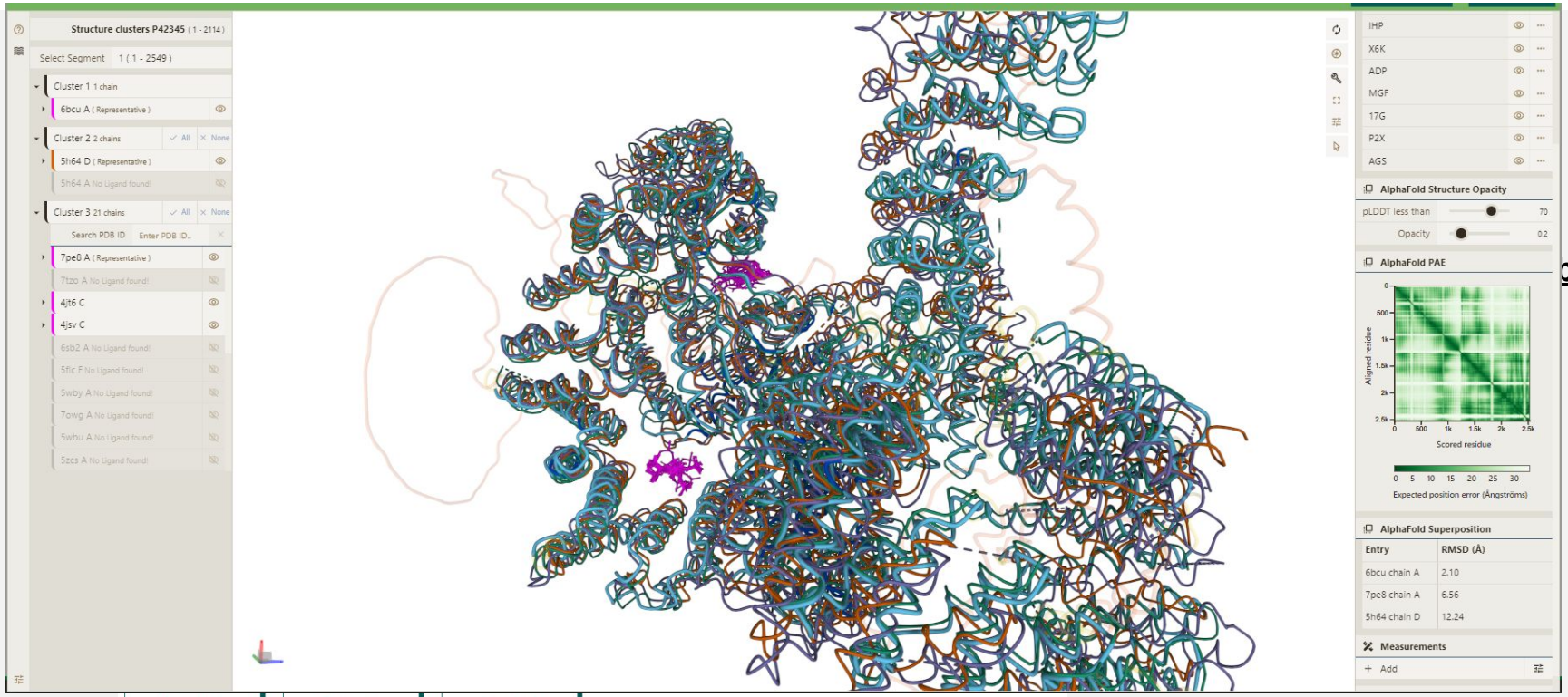
all the ligands observed in the same PDB entries as this protein



https://www.ebi.ac.uk/pdbe/pdbe-kb/

EMBL-EBI

PDBe-KB: Comparing structures in the PDB

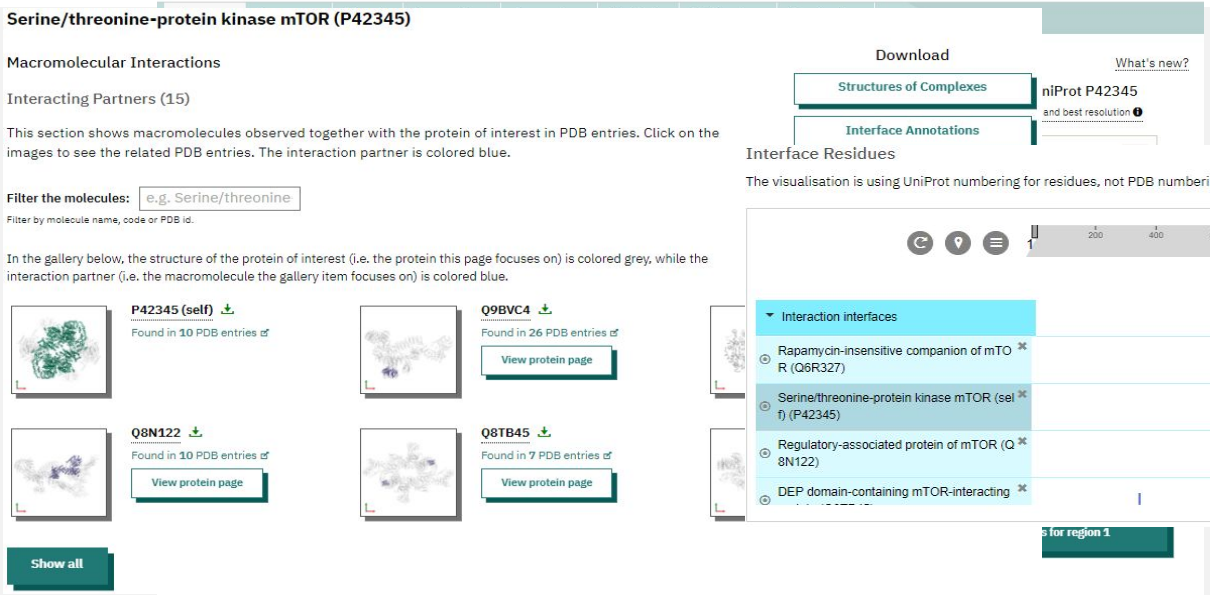


pe-kb/

all the ligands superposed on this protein

EMBL-EBI

PDBe-KB: Comparing structures in the PDB

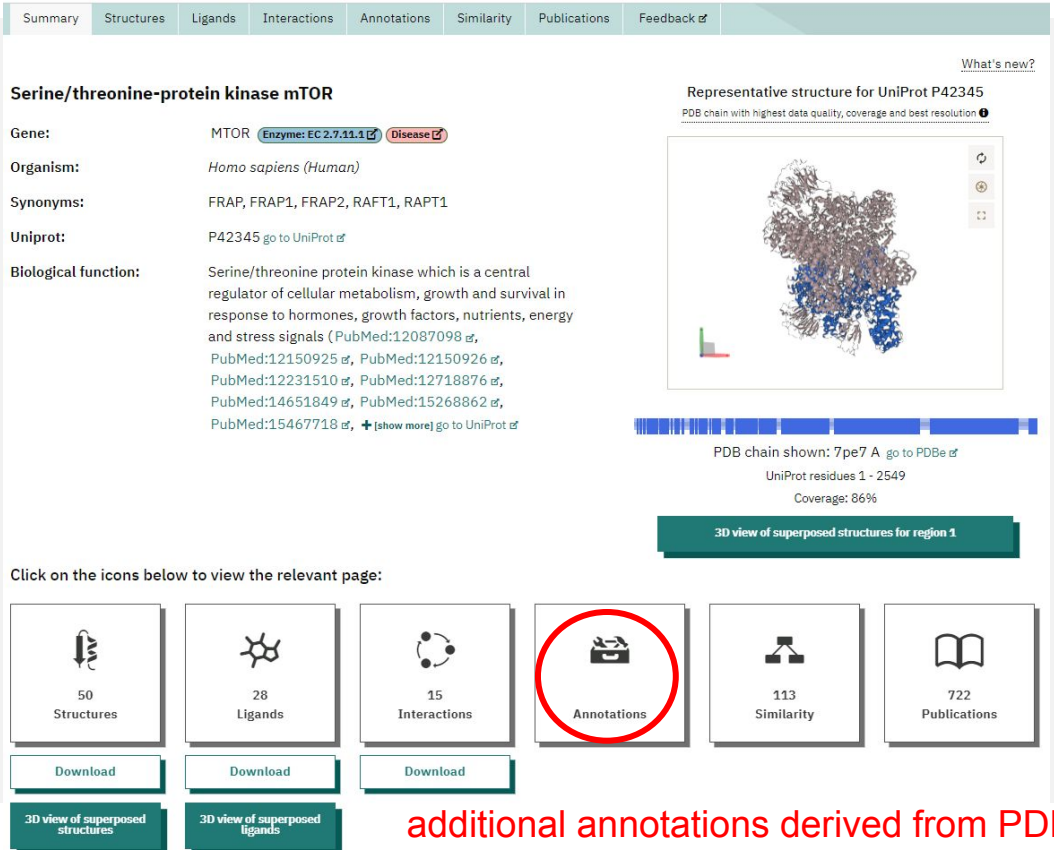


https://www.ebi.ac.uk/pdbe/pdbe-kb/

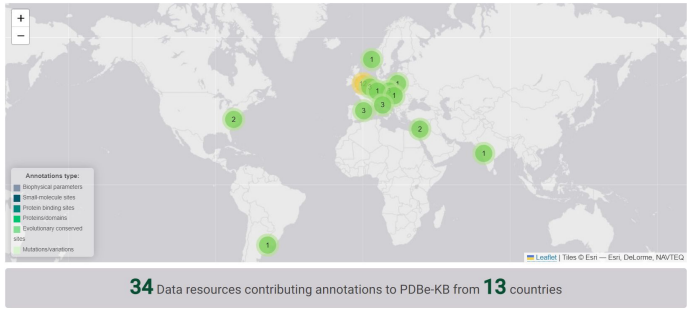
all the macromolecular interaction partners observed in the same PDB entries as this protein

EMBL-EBI

PDBe-KB: Comparing structures in the PDB



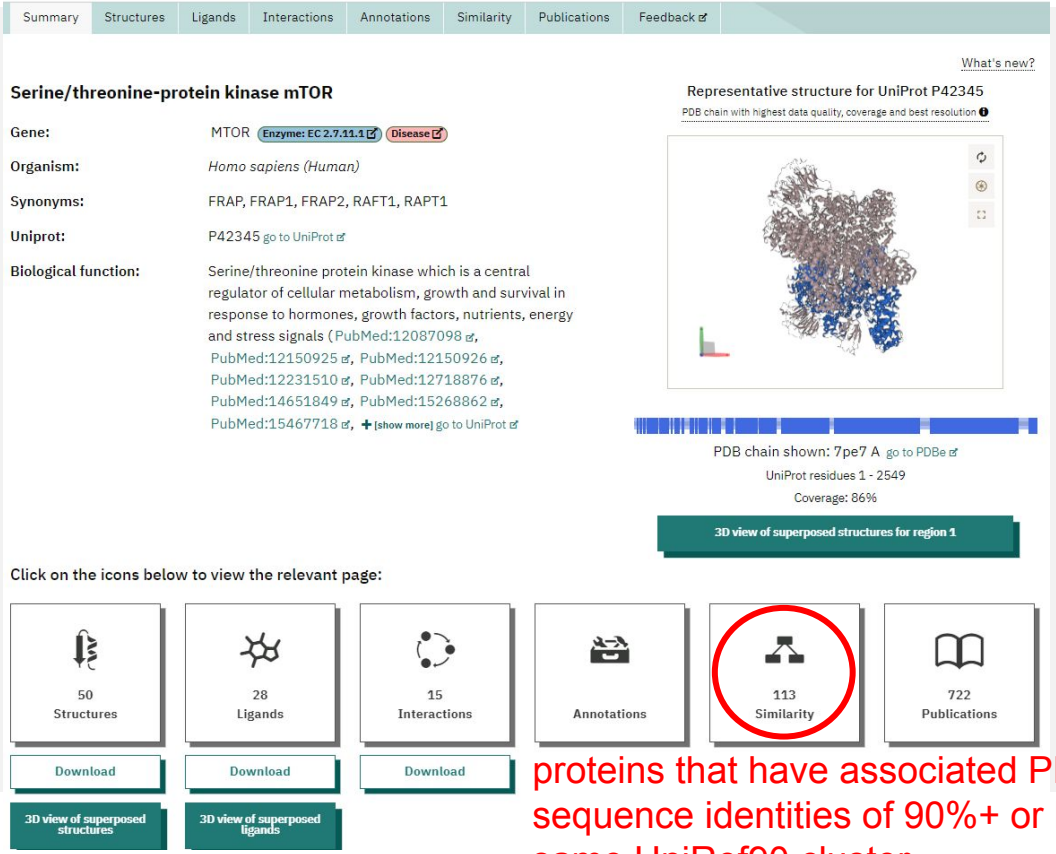
https://www.ebi.ac.uk/pdbe/pdbe-kb/partners



additional annotations derived from PDBe-KB partners (pdbekb.org/partners)

EMBL-EBI

PDBe-KB: Comparing structures in the PDB

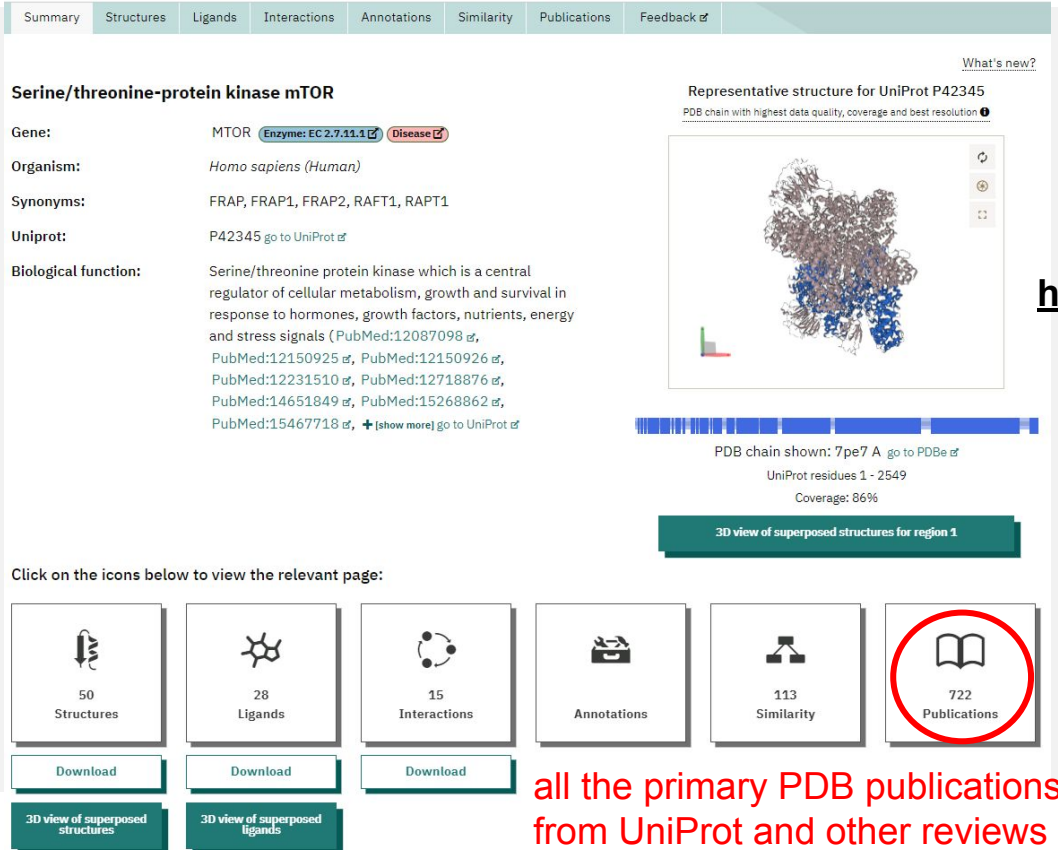


https://www.ebi.ac.uk/pdbe/pdbe-kb/

proteins that have associated PDBs with sequence identities of 90%+ or belong to the same UniRef90 cluster

EMBL-EBI

PDBe-KB: Comparing structures in the PDB



https://www.ebi.ac.uk/pdbe/pdbe-kb/

all the primary PDB publications, publications from UniProt and other reviews

EMBL-EBI

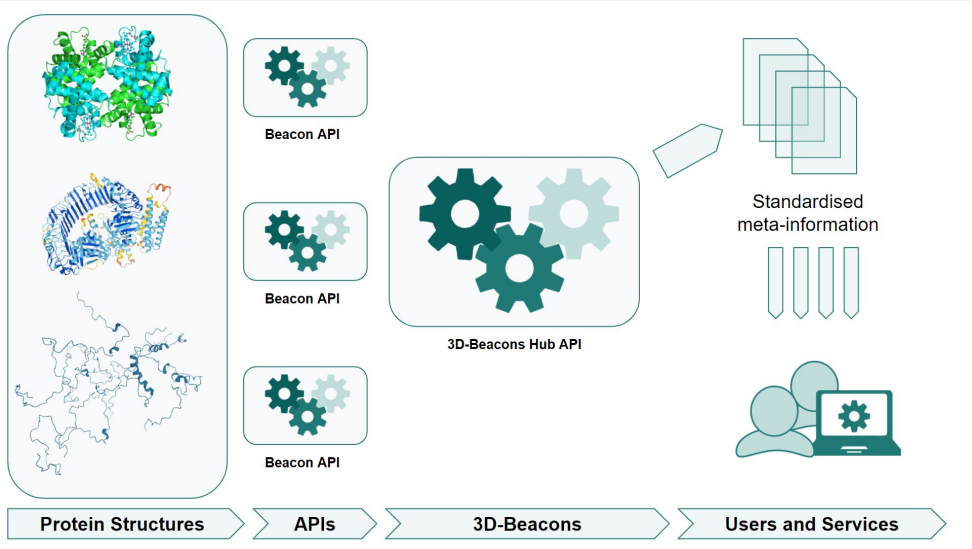
3D-Beacons: Finding computational models

- The 3D-Beacons Network[5] provides standardised access to both experimentally determined and predicted protein structures
- It is an open collaboration between many model providers:
  - PDBe
  - SWISS-MODEL
  - AlphaFold DB
  - Genome3D
  - SASBDB
  - AlphaFill
  - ModelArchive
  - Protein Ensemble Database

https://3d-beacons.org

3D-Beacons: Finding computational models

- You can access data and their descriptions in a standardised format
  - Where are the models (i.e. URLs)?
  - What is the overall quality of a model?
  - What is the context (i.e. metadata)?
    - Species?
    - Gene?
    - Sequence identifier?

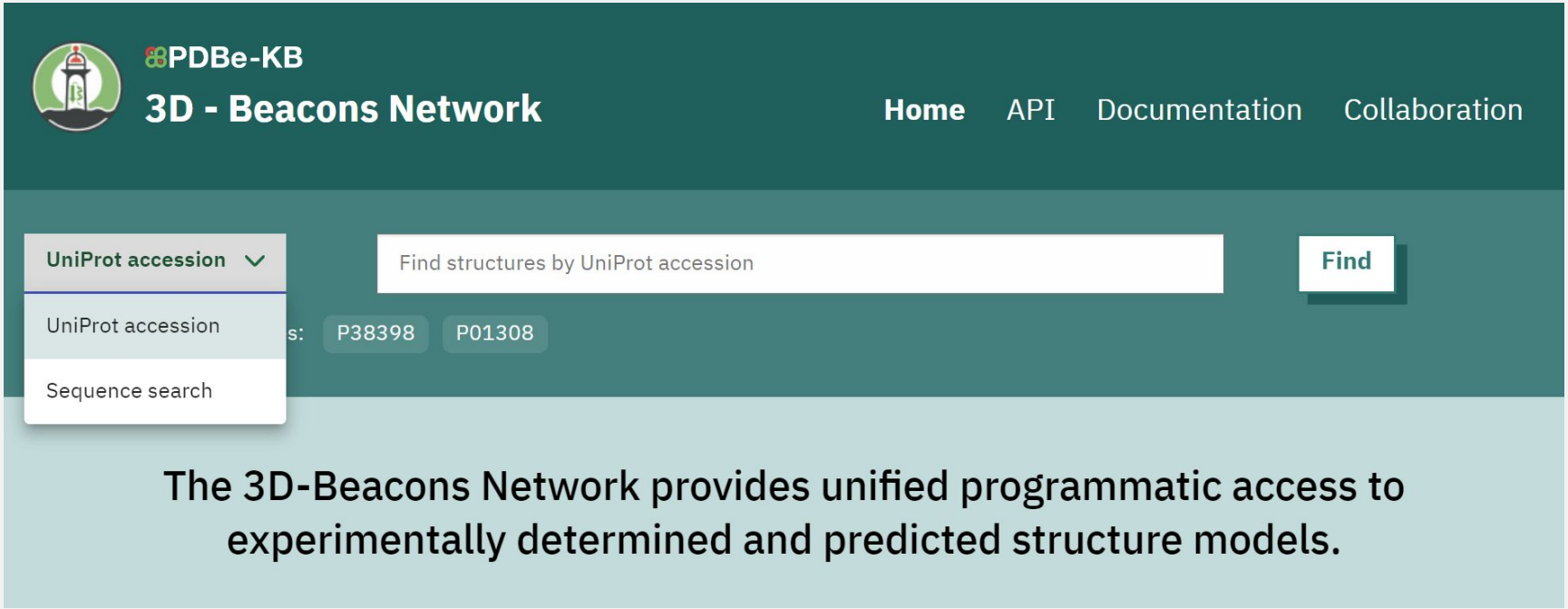


https://3d-beacons.org

EMBL-EBI

3D-Beacons: Finding computational models

Search by UniProt or sequence



https://3d-beacons.org

EMBL-EBI

3D-Beacons: Finding computational models

P38398 (BRCA1_HUMAN) - 64 Structures available

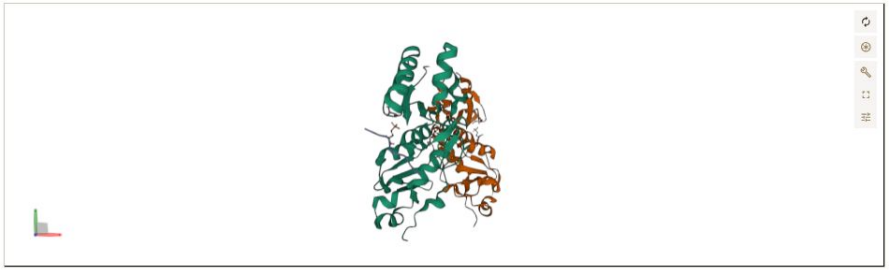
Information

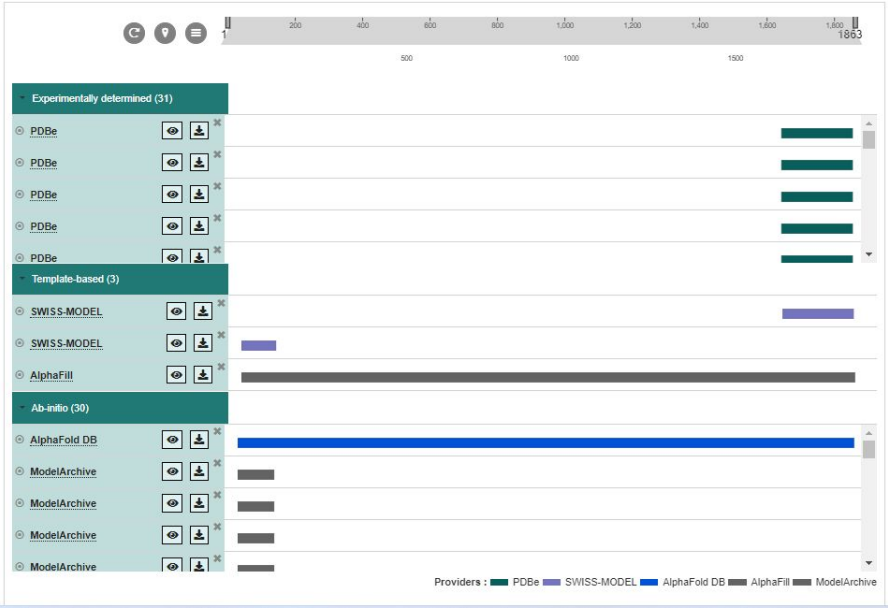
Protein	Breast cancer type 1 susceptibility protein Go to UniProt
Gene	BRCA1
Source organism	*Homo sapiens*
Biological function	E3 ubiquitin-protein ligase that specifically mediates the formation of polyubiquitin chains and plays a central role in DNA repair by facilitating responses to DNA ... Show more

31
Experimentally Determined Structures

0
Conformational Ensembles

Tem





https://3d-beacons.org

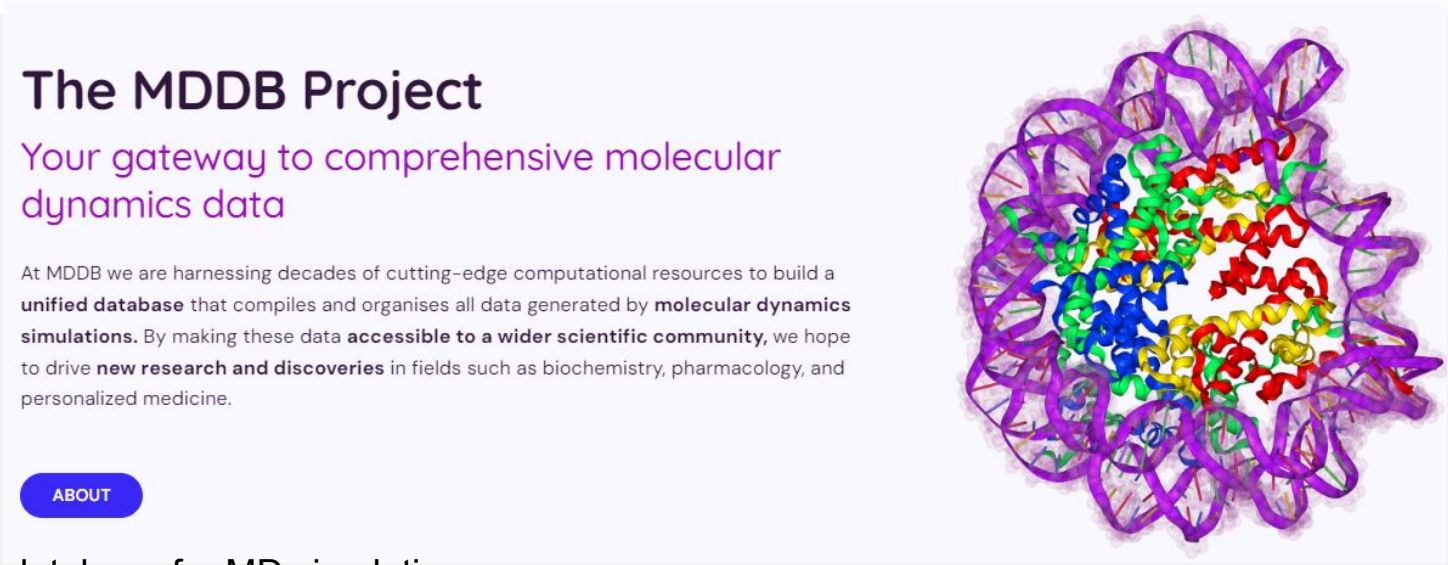
EMBL-EBI

Starting model from a MD simulation: MDposit

- open platform designed to provide web access to atomistic molecular dynamics (MD) simulations
- Possible to search simulations by title, authors or group names
- Download trajectory files and extract structures where the protein is solvated, energy minimized and ions added & continue the MD.

https://mdposit-dev.mddbr.eu/

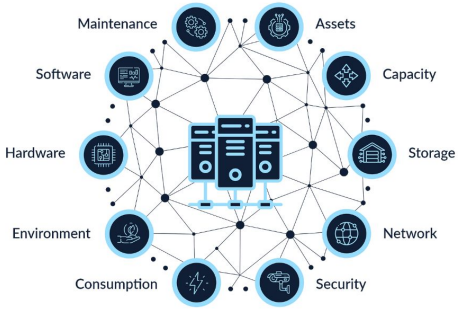
Starting model from a MD simulation: MDDB



- unified database for MD simulations
- multidisciplinary consortium for validated MD simulations
- EU horizon project including 7 partners and lead by IRB Barcelona

https://mddbr.eu/

MDDB Objectives



Establish FAIR principles for MD data

Define and promote good practices for the generation and analysis of trajectories

Design the technical infrastructure for a general repository for simulation data

Design a sustainability and governance model for MDDB

Funded by the European Union

MDDB received funding from EU's Horizon Europe programme under grant agreement 101094651

EMBL-EBI

Conclusions

- Selection of the correct starting model is extremely important for a successful MD simulation
- You should be careful of several factors, such as the protein conformation, missing loops, presence of ligands, etc.
- It is also important to be aware of the resolution and experimental technique used to derive the structure
- PDBe-KB & 3D Beacons are useful resources to compare protein structures across experimental and computational models.

Stay in touch

https://www.ebi.ac.uk/pdbe/

@PDBeurope

Protein Data Bank in Europe (PDBe)

@ProteinDataBank

proteindatabank

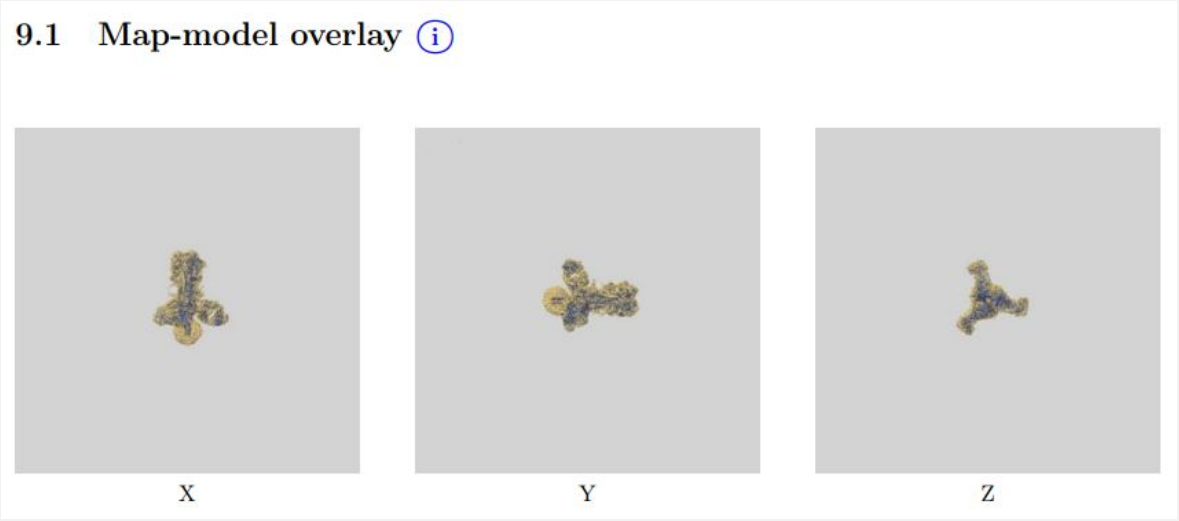
EMBL-EBI

Feedback

https://tinyurl.com/PDBe-bioexcel

EMBL-EBI

Data quality for EM structures



- Displays graphically the fraction of atoms that fit within the map at different map contour levels
- Also displays the recommended contour level for this map, as provided by the depositor at deposition.