Protein Data Bank in Europe

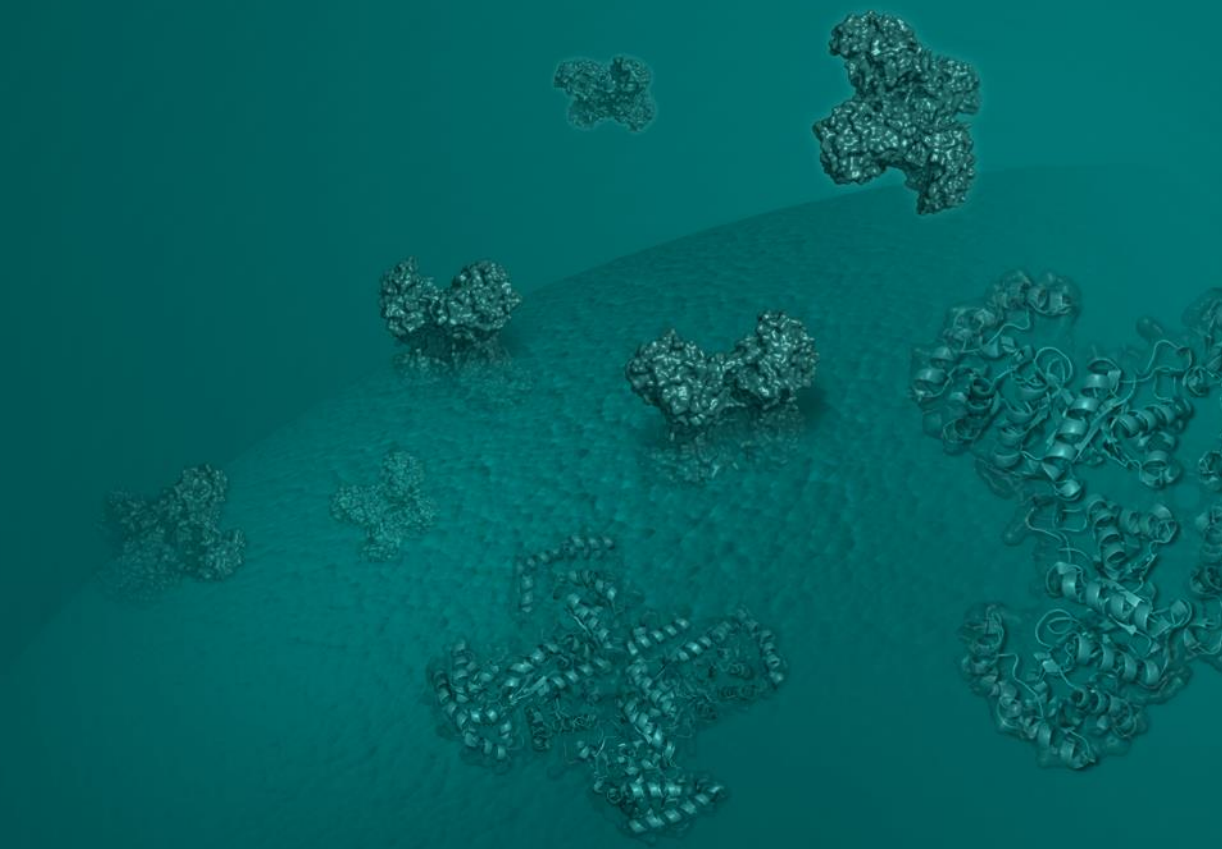
Depositing your data to the PDB

John Berrisford

pdbhelp@ebi.ac.uk

proteindatabank

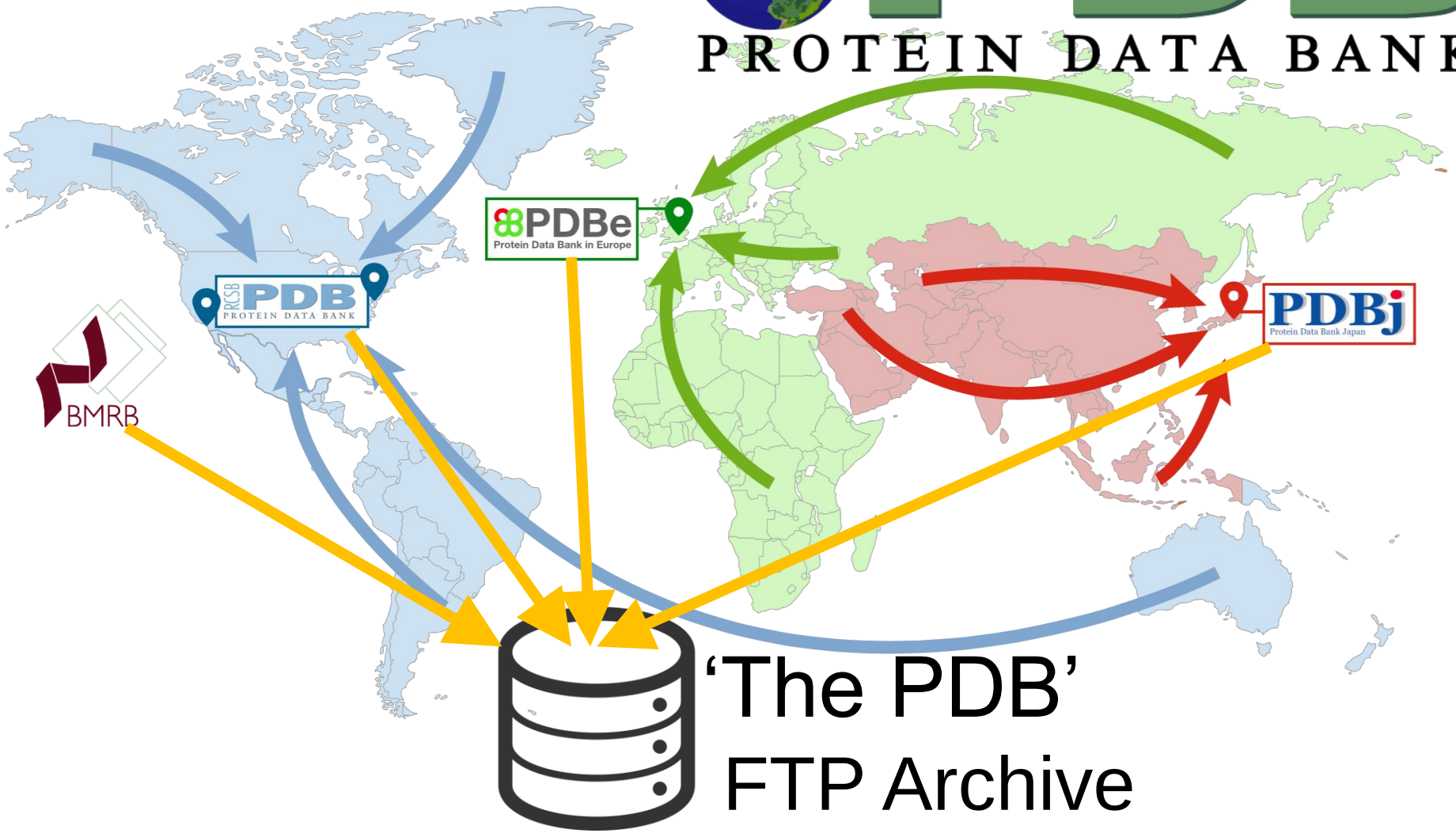
@PDBeurope



EMBL-EBI

Who are the PDB?

WORLDWIDE PDB
PROTEIN DATA BANK



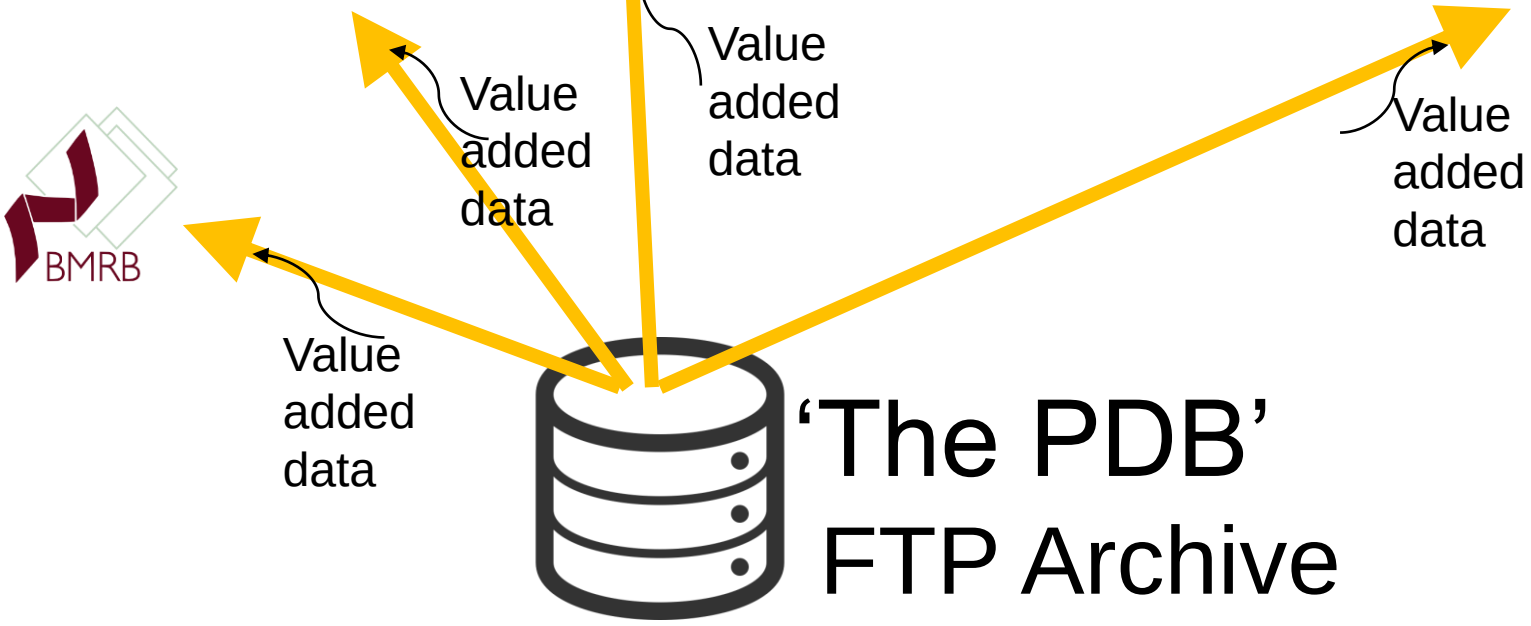
WORLDWIDE PDB
PROTEIN DATA BANK

PDBe
Protein Data Bank in Europe

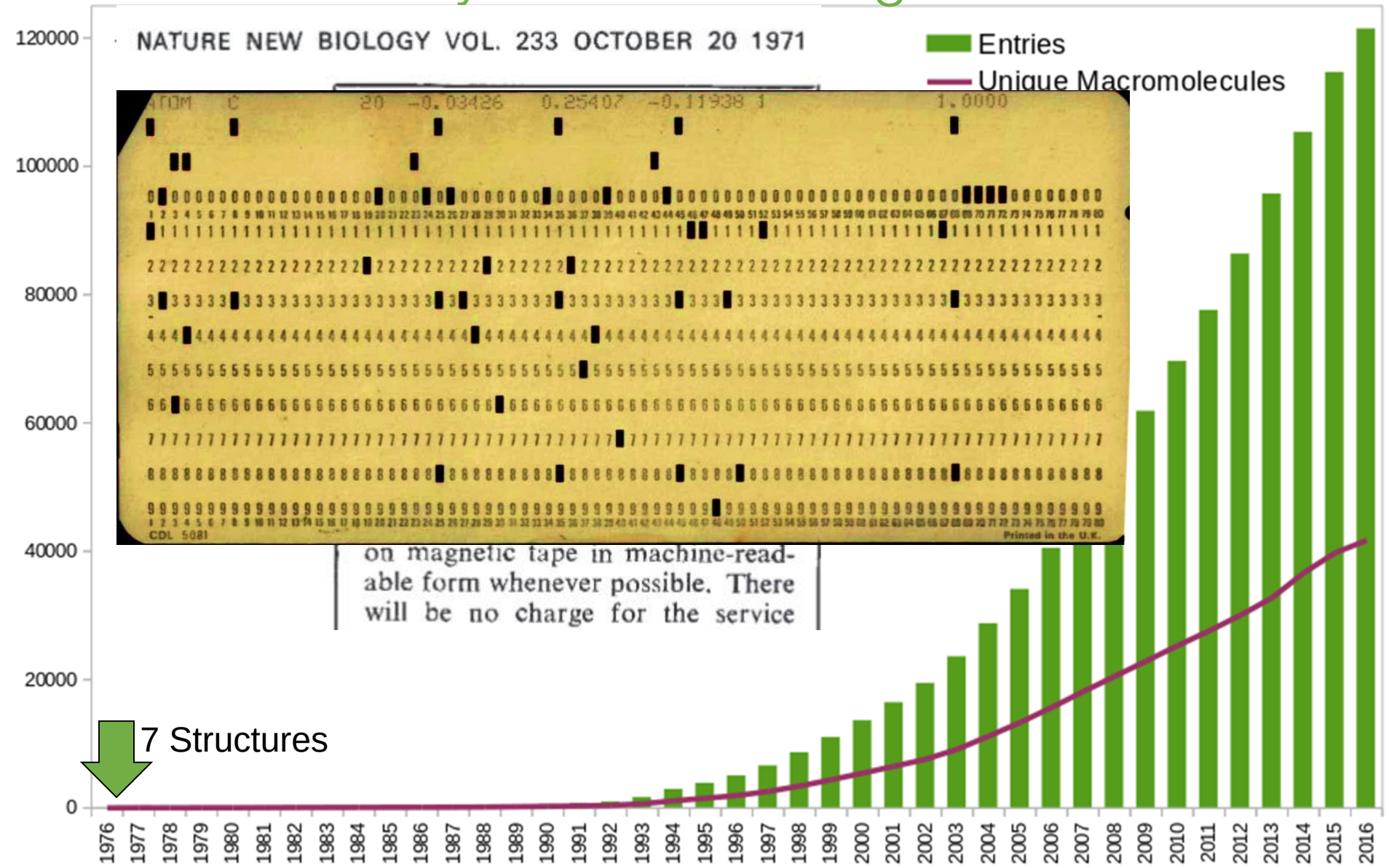
RCSB PDB
PROTEIN DATA BANK

PDBj
Protein Data Bank Japan

BMRB



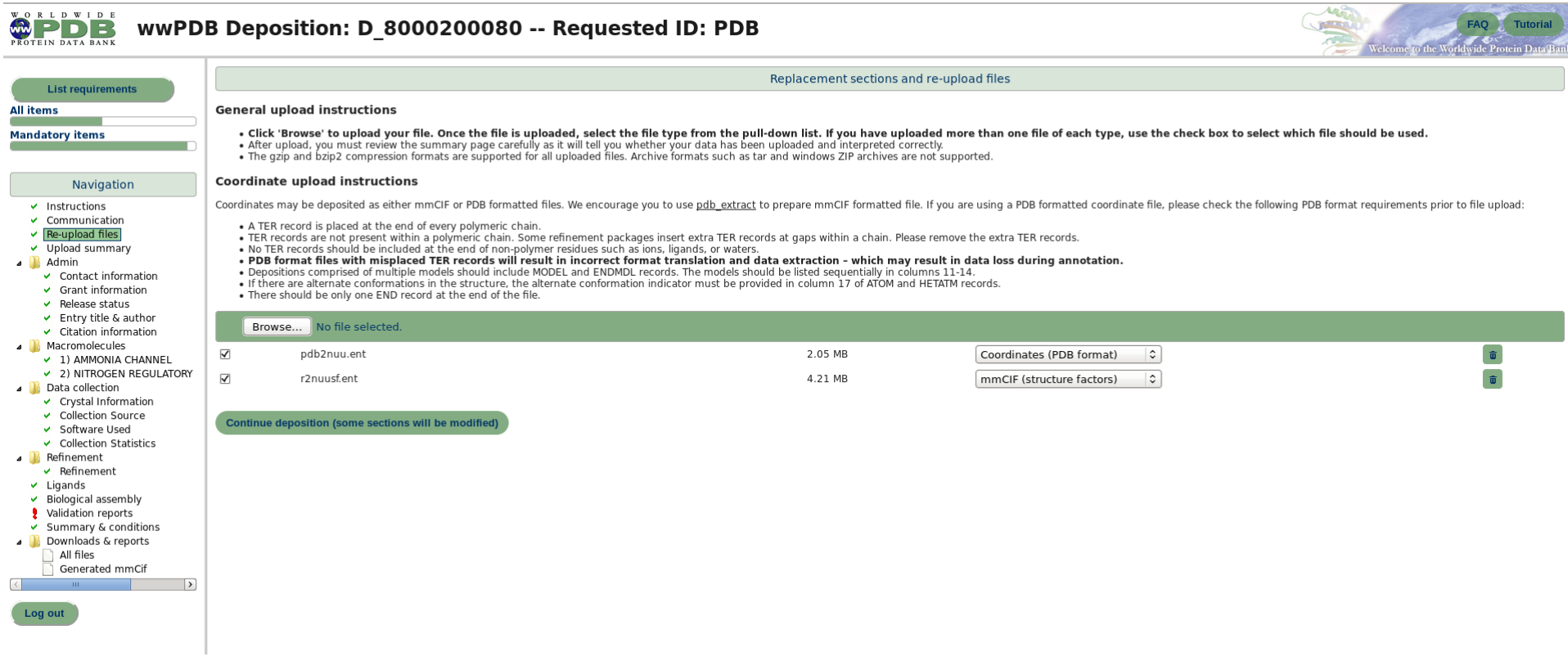
Oldest freely available biological database



Why deposit to the PDB?

- To make your data publicly available
- Required by journals

OneDep deposition system



What does deposition involve?

- File validation
- wwPDB validation
- Form filling
- Annotation – including more validation

File validation

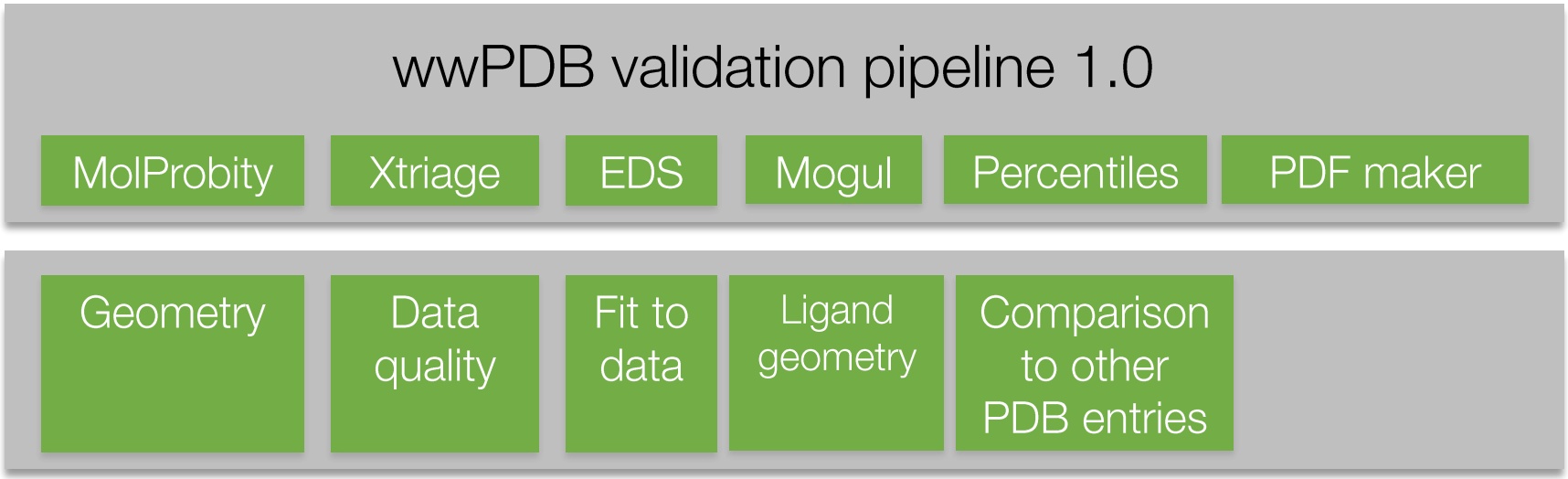
- Model file in mmCIF or PDB format
  - data harvesting (refinement stats) is better in mmCIF format
  - Refmac and phenix can output mmCIF format
    - http://www.wwpdb.org/deposition/PDBxDeposit

- Data is checked for:
  - Valid file
    - Can we read it!
    - NMR models are consistent
  - Residual B factors
  - “unusual” occupancies
    - Negative
    - Over 1

wwPDB validation

- Ligands checked by graph match to wwPDB Chemical Component Dictionary
  - pdbe.org/chem
- Pre-submission wwPDB validation report
  - Geometry issues
    - Polymer
    - Ligand
  - Chirality issues
  - Fit of model to data

wwPDB validation



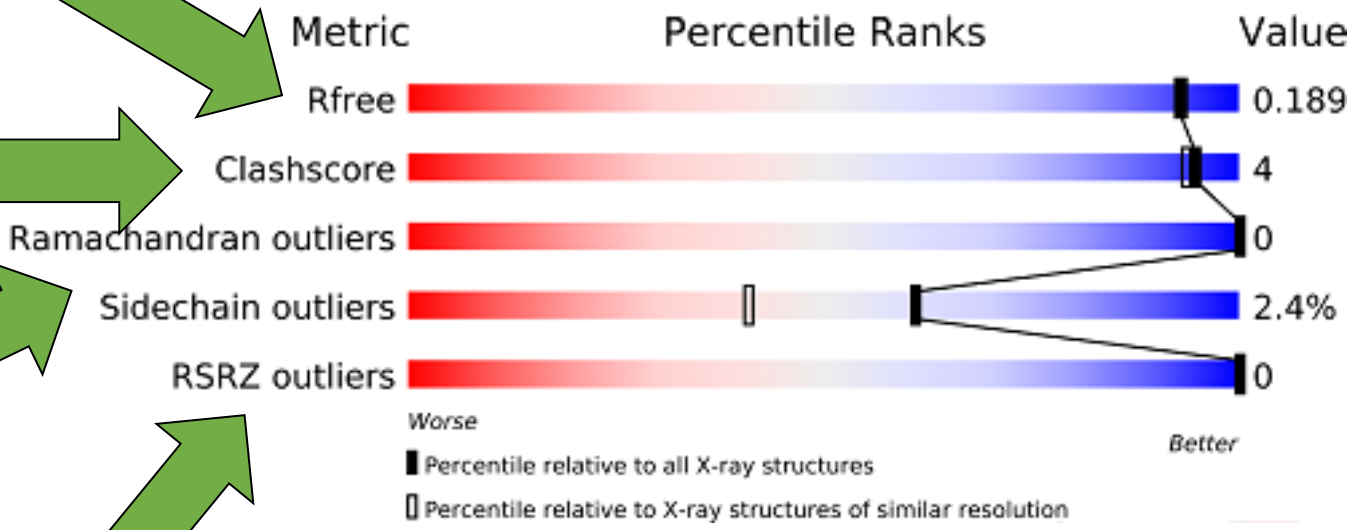
wwPDB validation

How well does the model back-predict the data?

Atoms bumping into each other

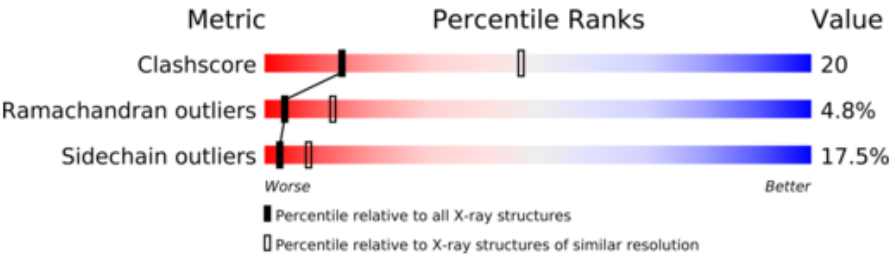
Surprising bond angles

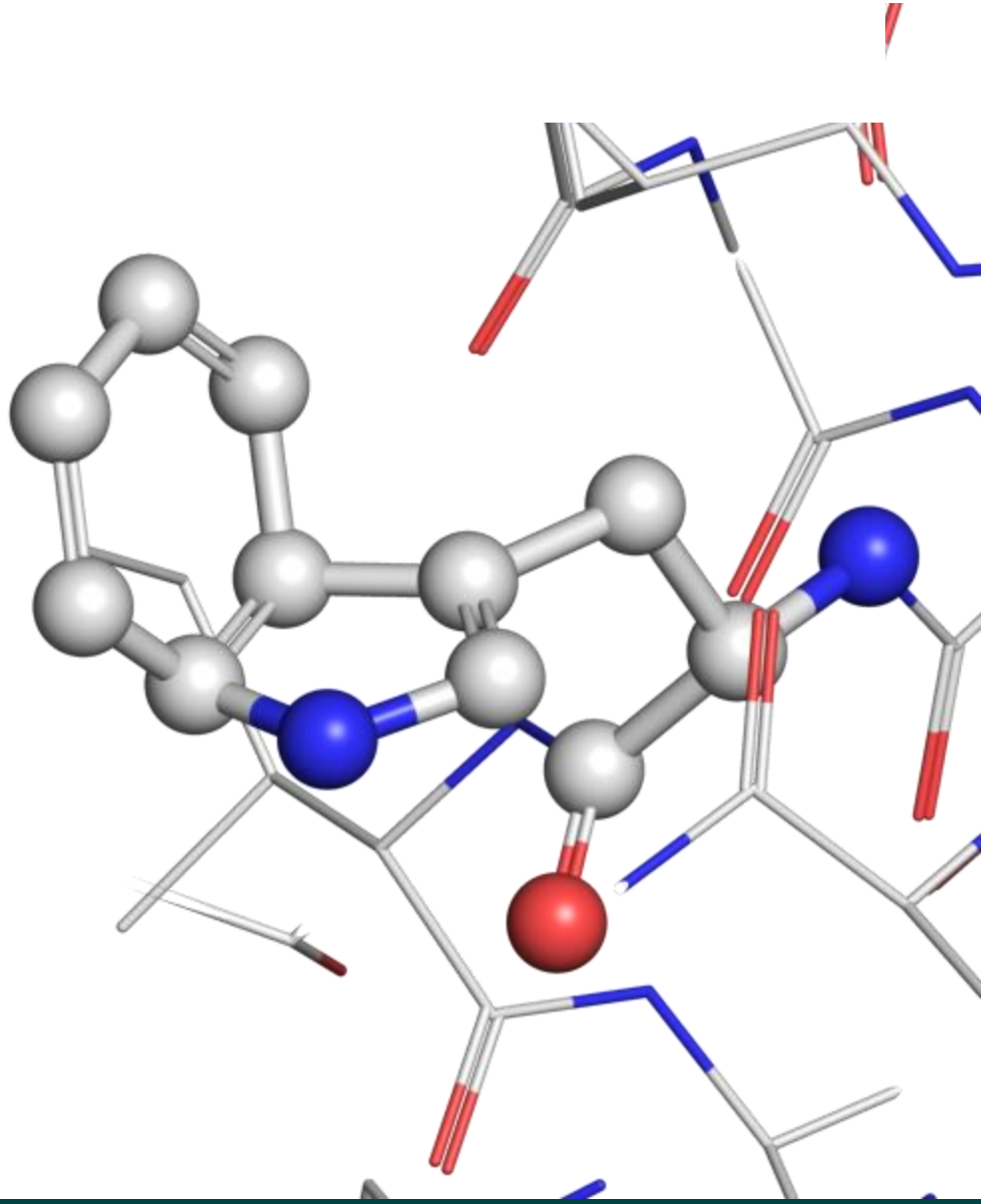
Atoms not in electron density



Interesting geometry

7GPB 2.9Å



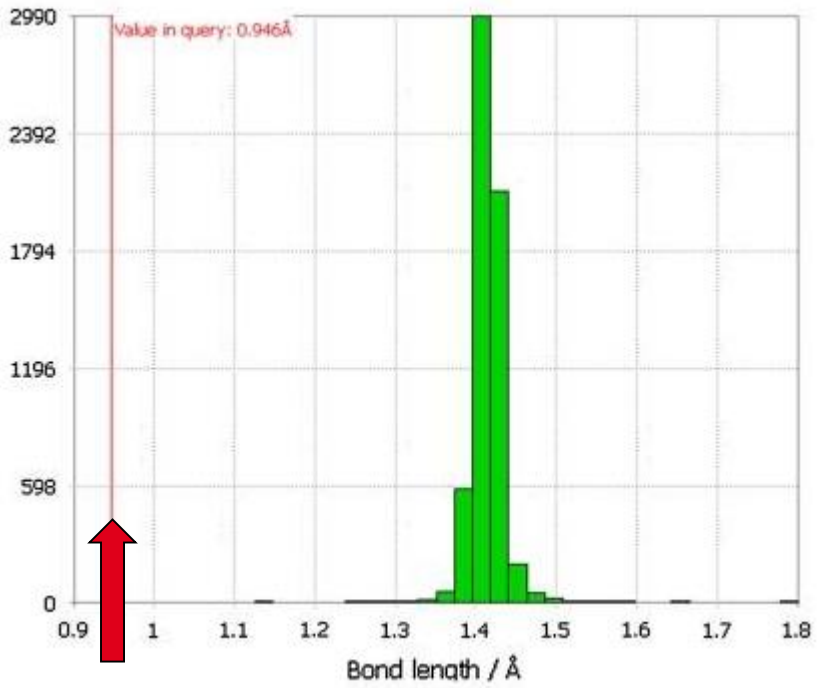


Validation for small molecules

Geometry

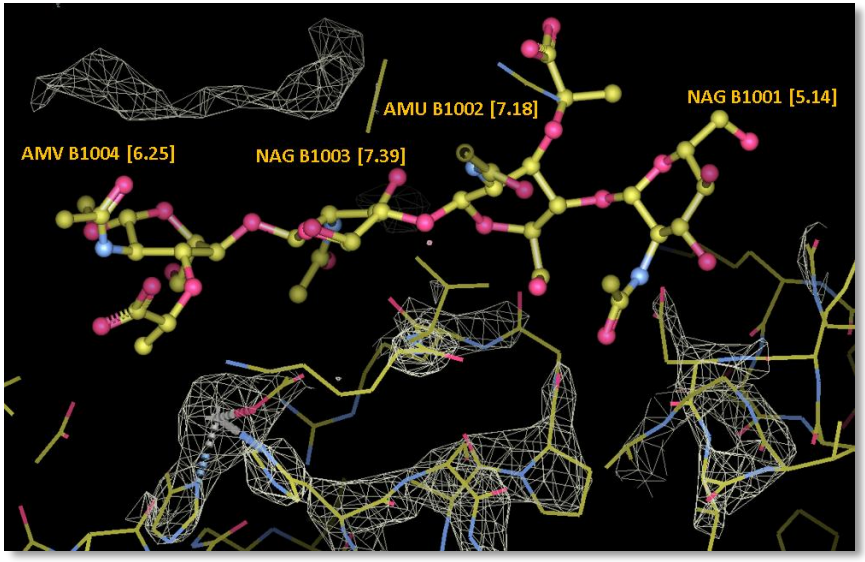
THE CAMBRIDGE CRYSTALLOGRAPHIC DATA CENTRE

Using Mogul

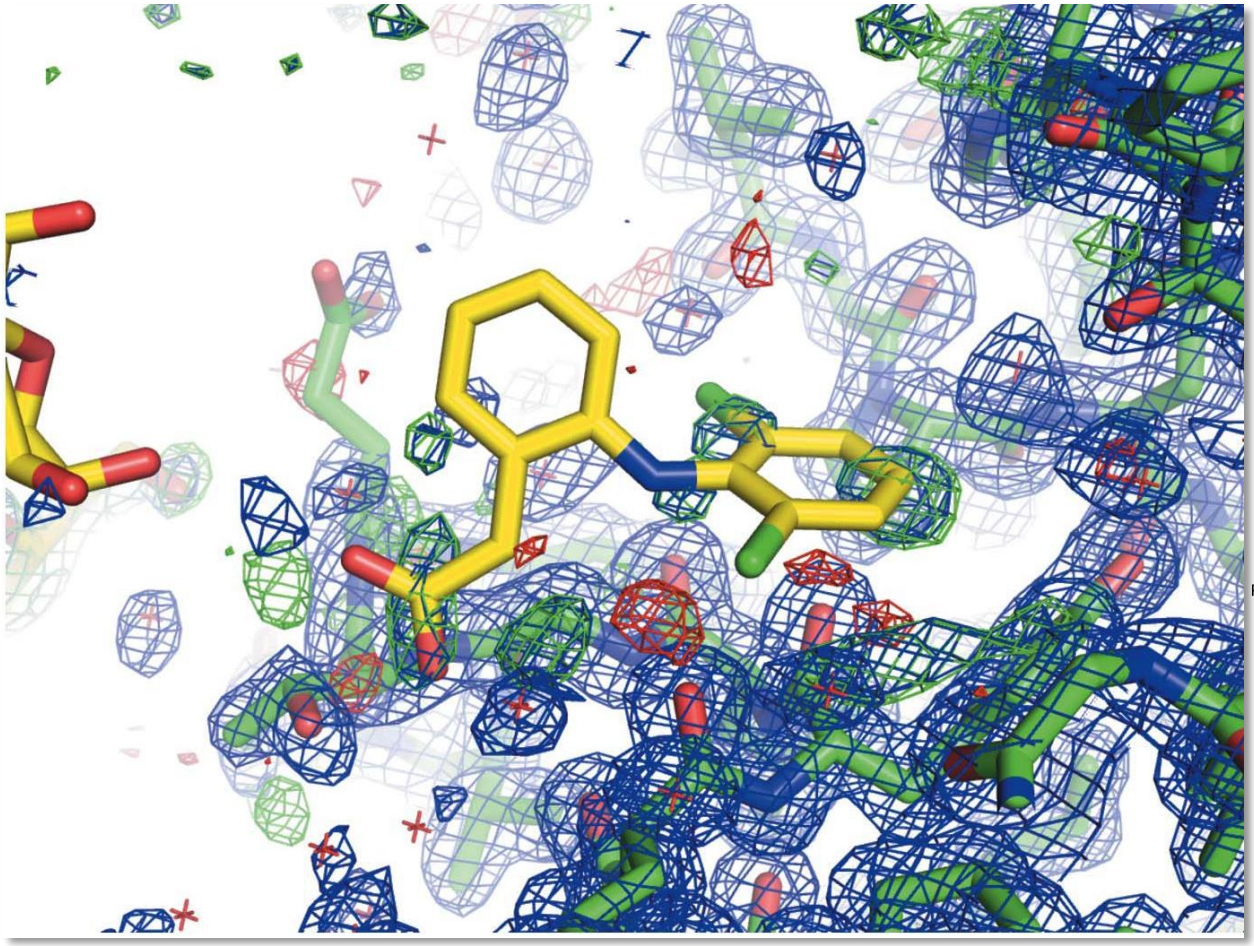


Fit to density

RSR: Measure of how well a residue fits its local density (Jones *et al.*, 1991)



Poor density for ligand

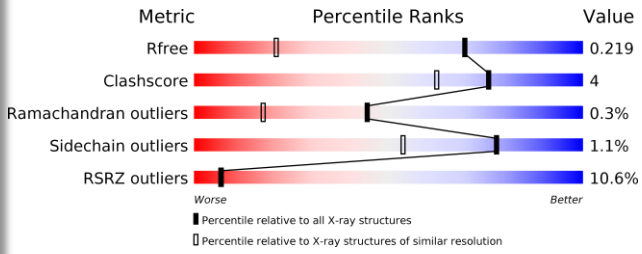


PDB: 3IB0 (1.4Å)

DIF A701

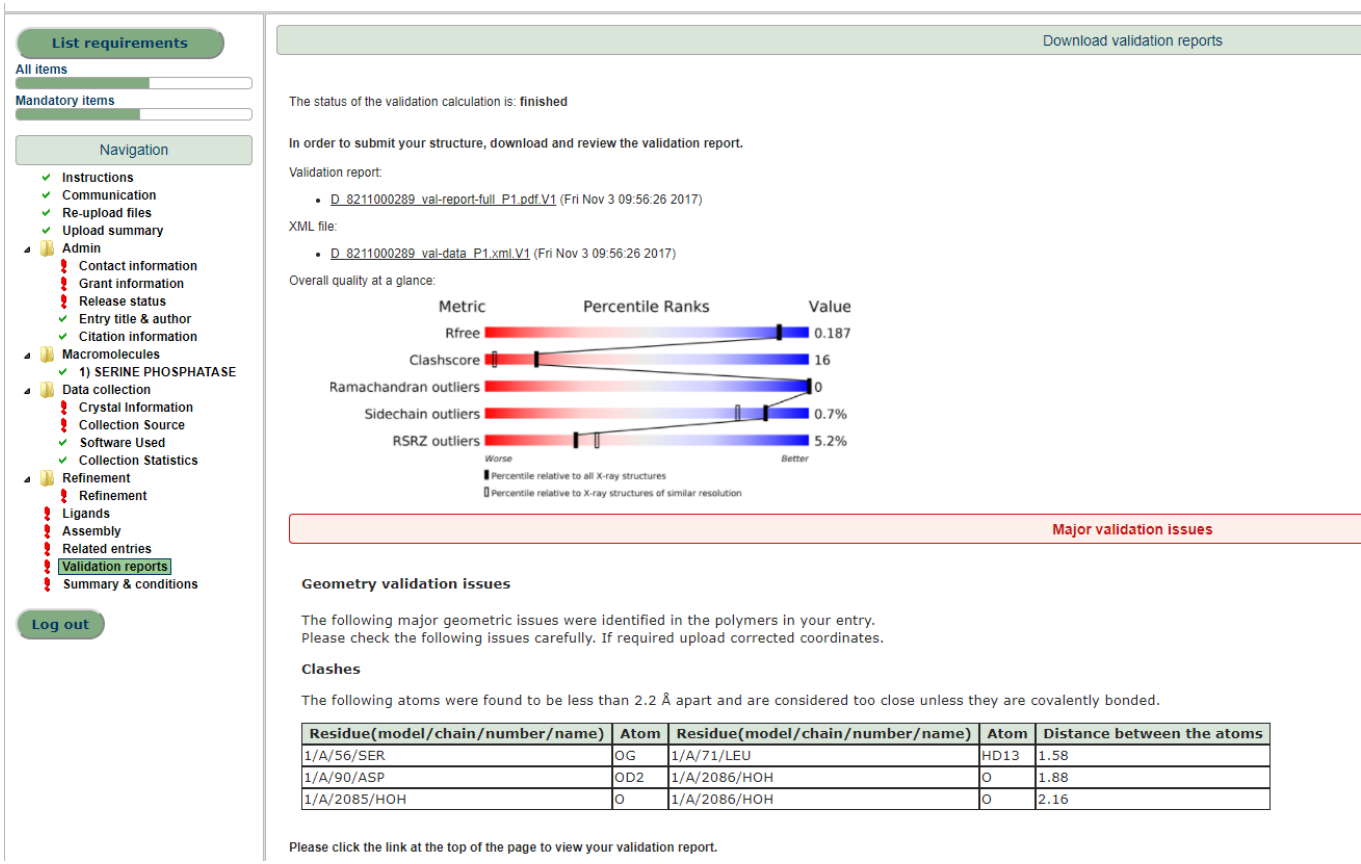
Ligand diclofenac not supported by electron density

https://www.ebi.ac.uk/pdbe/entry/pdb/3ib0/bound/DIF



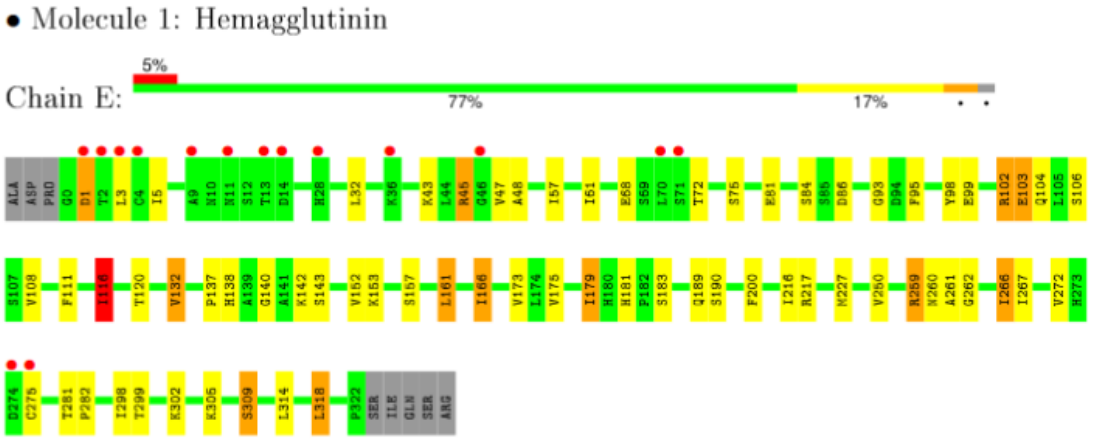
OneDep - highlighting outliers

- Geometry
  - Clashes
  - Bond lengths
  - Bond angles
  - Bonds between amino acids / nucleotides

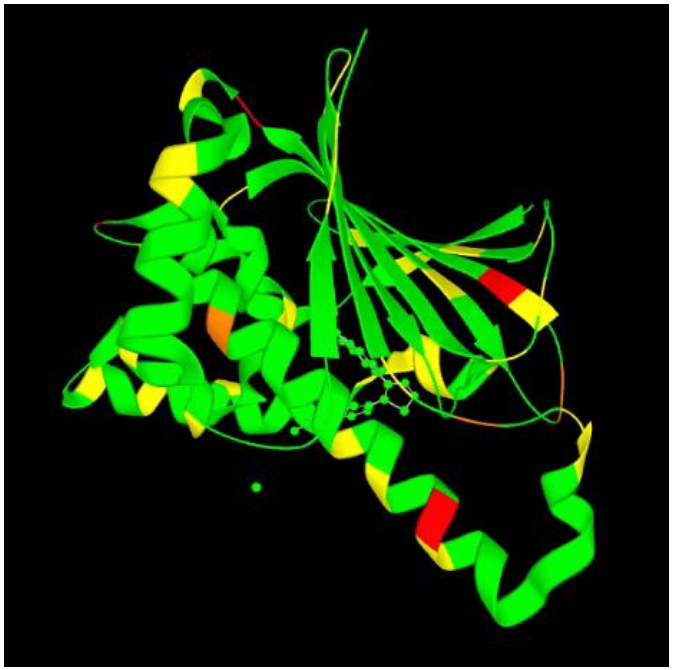


EMBL-EBI

Why give you all this validation upfront?



- PDF available on wwPDB websites
- XML available
  - Viewers use this to display validation
    - COOT
    - LiteMOL @PDBe
    - PyMOL (PDBe plugin)



What do we need?

- Mandatory items in deposition
  - Red in the deposition interface

- Admin information
  - Who are you?
  - When do you want your entry released?
- Sequence information for each macromolecule
  - Better to add this to your file in advance – ccp4i2 does this for you
- Data collection statistics
  - Where did you collect your data?
- Refinement statistics
  - Normally parsed directly from your uploaded file
- Ligand identity

Form filling

- Controlled vocabularies
  - Beamlines
  - Detectors
  - Software
- Sanity checks / Cross validation
  - Wavelength of greater than 2.0 is unusual
  - Resolution of data collected is higher than you refined!
  - Detector type is linked to detector

Diffraction ID*: 1

Data collection temperature (K)*: 10 ! Warning: Expected a value between 80.0 and 300.0

Data collection temperature details:

Diffraction Radiation

Protocol*: LAUE MAD SINGLE WAVELENGTH

Monochromator:

Scattering type*: electron neutron x-ray

Diffraction Source

Source type*: SYNCHROTRON

Source details*: DIAMOND BEAMLINE I02

Wavelength or list of wavelengths used for this experiment (Å)*: 15 ! Warning: Expected a value between 0.5 and 2.0

Diffraction Detector

Detector*: EIG ⚠ Value not in the enumeration list for this item.

Detector type*: ⚠ No value present for this mandatory item.

Date of collection*:

Optics:

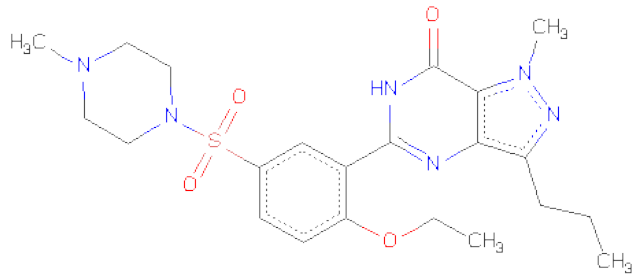
DECTRIS EIGER X 16M
DECTRIS EIGER X 500K
DECTRIS EIGER R 1M
DECTRIS EIGER R 4M
DECTRIS EIGER X 1M
DECTRIS EIGER X 9M
DECTRIS EIGER X 4M

Continue to next section

Sanity Check

Controlled vocabulary

Biocuration (annotation)



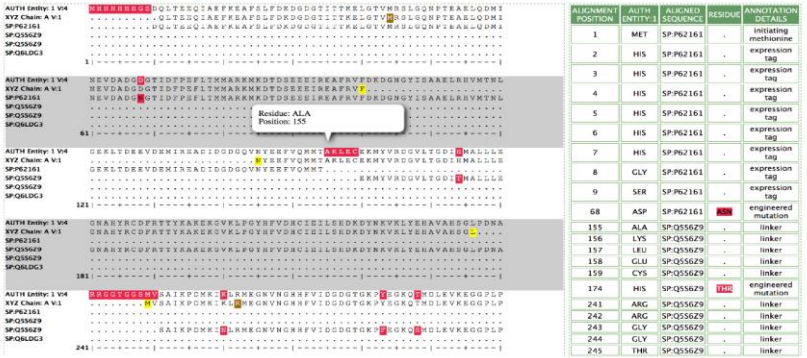
Ligand matching to dictionary

All identical chemicals have the same code

Taxonomy

Sequence matching to UniProt

Highlight mutations, conflicts, tags etc



Assemblies

Standardising of representations – helping make PDB searchable

Sanity - is that really purified from human hearts?

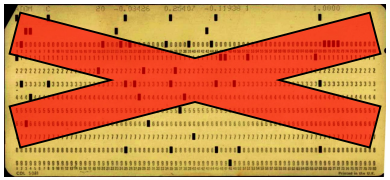
4.5 FTEs for all of Europe and Africa- >4000 structures a year

Can you try it out?

- YES
  - pdbe.org/deposition
  - gives a link to a validation server which can be used prior to deposition
  - https://validate-rcsb-1.wwpdb.org/

New wwPDB developments

PDBx/mmCIF V5

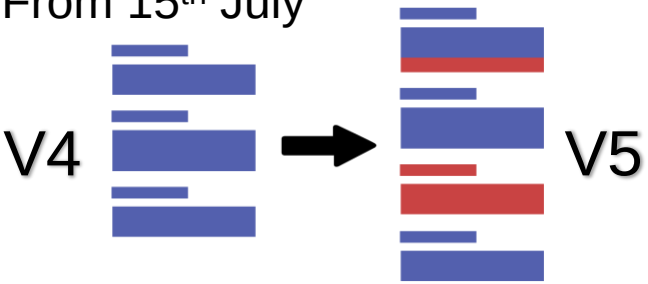


More metadata for EM derived models
all EM categories except em_map

Better audit information

Some data standardisation
work still on-going to clean up data

From 15th July

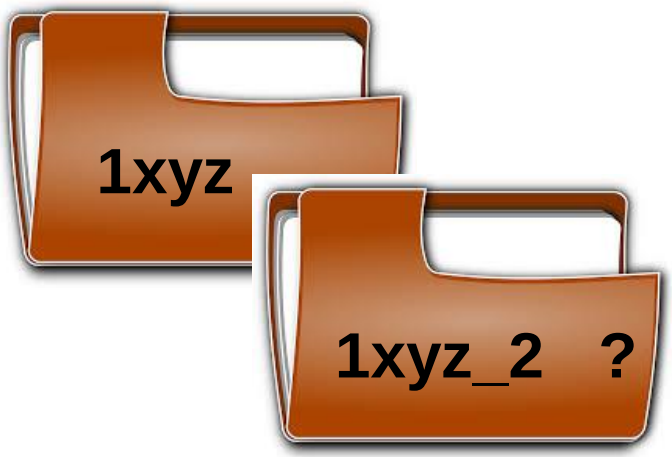


Coming soon

Versioning

Update released entries
And retain the ID code

Details yet to be finalised



Getting data out of the PDB

pdbhelp@ebi.ac.uk
proteindatabank
@PDBeurope

PDBe
Protein Data Bank in Europe

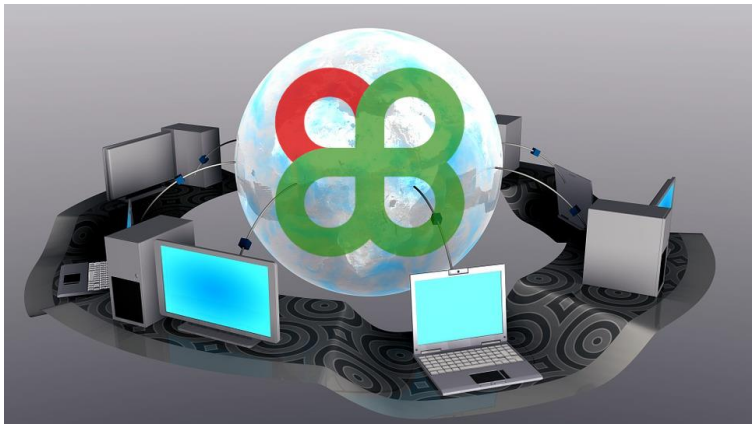
EMBL-EBI

Searching the PDB made simple

At PDBe we've implemented:

- Auto-complete search suggestions
- Facets to narrow down search results
- Quality presented on search results
- Four views of results
  - Moving away from entry-centricity

Challenges – from archive to resource
PDB is a historic archive



- "Archive" – i.e., what was described in the literature
- We have to accept and distribute everything
- Provider-centric
- Organised by entry not molecule/complex/function

- Integrate with other biological & chemical data
- Integrate structural data on scales from atoms to cells
- Deliver appropriate structure data to non-experts (in context of their work)

MetaData is vital

- *"Coordinates by themselves just specify shape and are not necessarily of intrinsic biological value, unless they can be related to other information."*

Gerstein, (2000)

Nature Structural Biology 7 960 - 963

- Sequence databases

- Taxonomy

- Publication

- Function

Turning an archive into a resource

- Easily searchable

- Cleaned up data
  - Standard vocabularies
- Value-added information

- Quality assessment

- Standard comparisons
- Flag irregularities

Lots of data clean-up locally

Historical data in the PDB are noisy

How many ways can you spell ‘engineered’?

Moolecular Replacment?

Ribosomal RNA names

Lots more

Improving the Resource

The problem with the PDB is…

- I can't find what I want
- Too many false hits when I search
  - CaM, CaM-Kinase, CaM binding protein, Cam-like…
- Results are too complicated
  - Which lysozyme to use? 1600 structures from 50 species!
  - Which is best?/what is best?
- Redundancy
  - How many unique human protein structures?

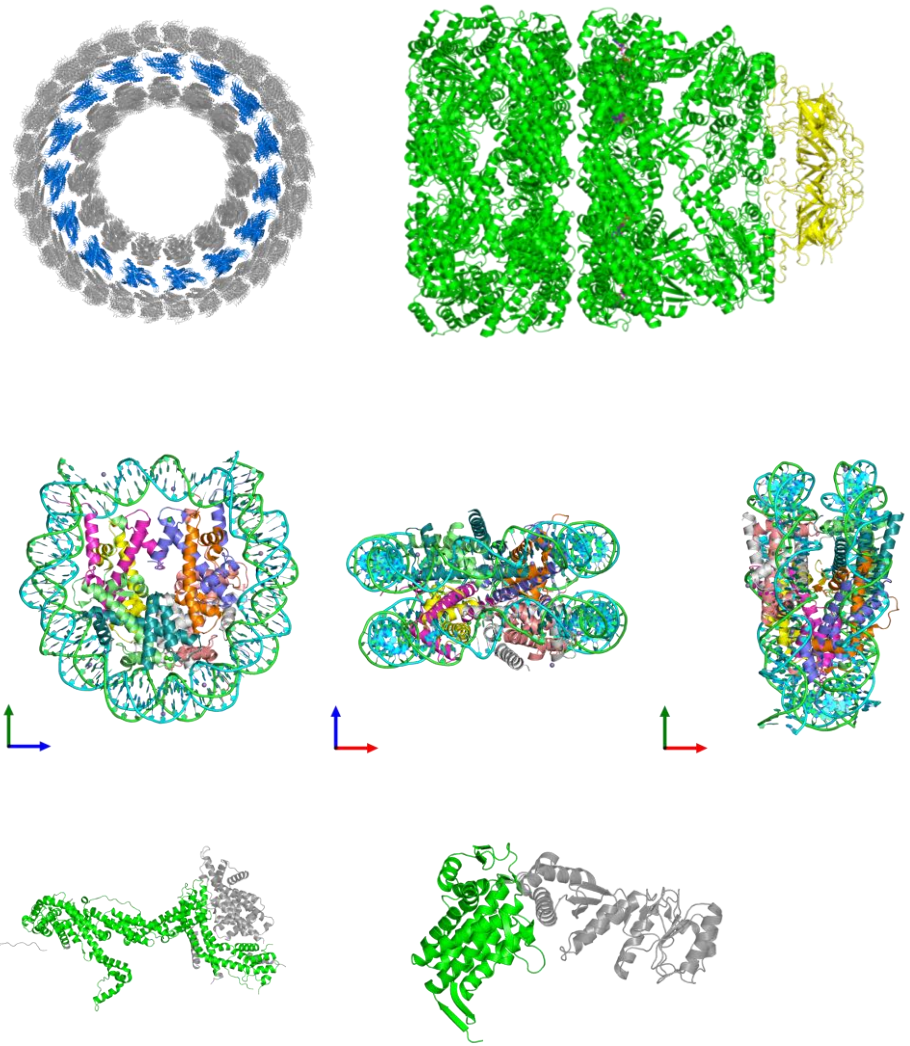
What is needed to answer such queries?

DATA

- Improved data quality
- Cleaned up data
  - Standard vocabularies
- Value-added information
- Intuitive ways to present 3D structure information
  - 1D-3D mapping
  - Images, images, images!

Using images to convey information

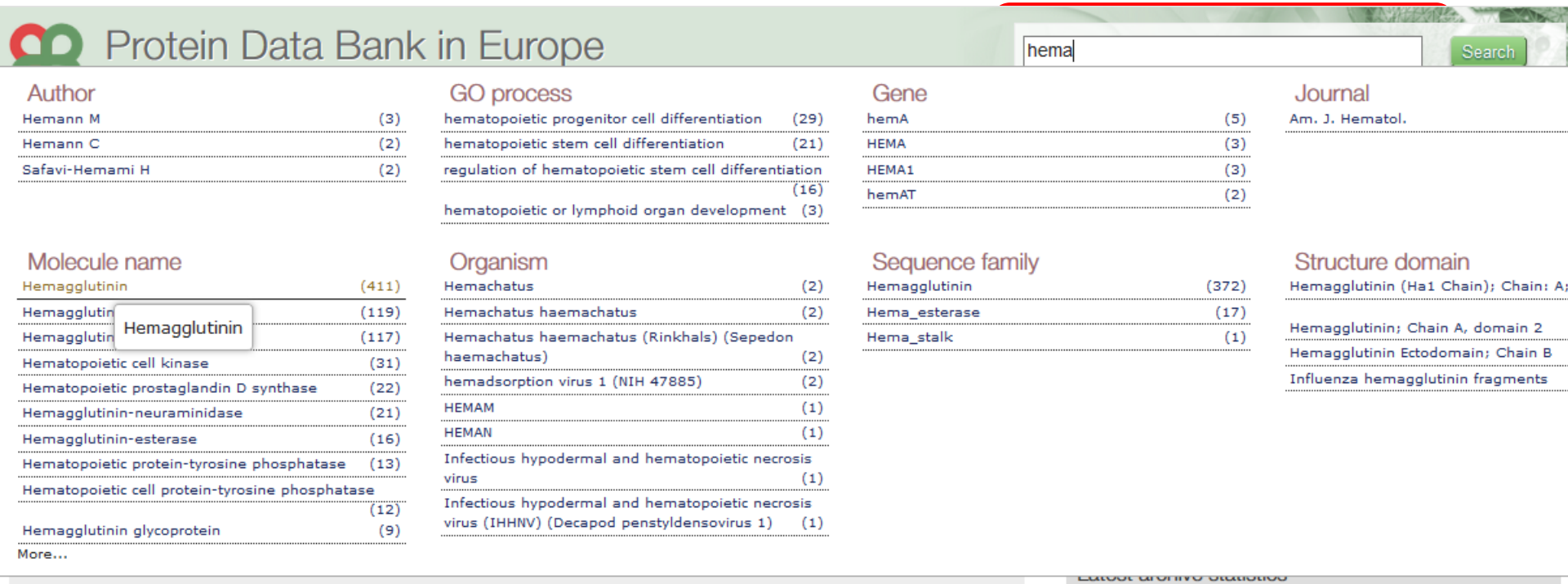
- Where is “my protein” in this assembly?
- How many unique macromolecules are there in this complex?
- What is the overall shape of the complex?
- Where is this sequence or structure domain in the 3D structure?



Searching the PDB made simple

At PDBe we've implemented:

- Auto-complete suggestions
- Facets to narrow down search results
- Quality presented on search results
- Four views of results
  - Moving away from entry-centricity


Protein Data Bank in Europe
hema
Search
Author
Hemann M (3)
Hemann C (2)
Safavi-Hemami H (2)
GO process
hematopoietic progenitor cell differentiation (29)
hematopoietic stem cell differentiation (21)
regulation of hematopoietic stem cell differentiation (16)
hematopoietic or lymphoid organ development (3)
Gene
hemA (5)
HEMA (3)
HEMA1 (3)
hemAT (2)
Journal
Am. J. Hematol.
Molecule name
Hemagglutinin (411)
Hemagglutin (119)
Hemagglutin (117)
Hemagglutinin
Hematopoietic cell kinase (31)
Hematopoietic prostaglandin D synthase (22)
Hemagglutinin-neuraminidase (21)
Hemagglutinin-esterase (16)
Hematopoietic protein-tyrosine phosphatase (13)
Hematopoietic cell protein-tyrosine phosphatase (12)
Hemagglutinin glycoprotein (9)
More...
Organism
Hemachatus (2)
Hemachatus haemachatus (2)
Hemachatus haemachatus (Rinkhals) (Sepedon haemachatus) (2)
hemadsorption virus 1 (NIH 47885) (2)
HEMAM (1)
HEMAN (1)
Infectious hypodermal and hematopoietic necrosis virus (1)
Infectious hypodermal and hematopoietic necrosis virus (IHHNV) (Decapod penstyldensovirus 1) (1)
Sequence family
Hemagglutinin (372)
Hema_esterase (17)
Hema_stalk (1)
Structure domain
Hemagglutinin (Ha1 Chain); Chain: A;
Hemagglutinin; Chain A, domain 2
Hemagglutinin Ectodomain; Chain B
Influenza hemagglutinin fragments
Previous featured structures
As of 28 December 2016 the PDB contains 125463 entries (latest PDB entries, chemistry, biology) and EMDB contains 4393 entries (latest map releases, latest header releases, latest updates).
News
Events

Search results

Refine query:

Macromolecules (11)

Hemagglutinin
Hemagglutinin HA1 chain
Hemagglutinin HA2 chain
Outer capsid protein VP8*
Outer capsid protein sigma-1
HA-17
HA33
HA70
Outer capsid protein VP5*

Molecule type (1)

Interacting macromolecules (100+)

Interacting compounds (46)

Species name (100+)

Experimental methods (2)

Assembly composition (3)

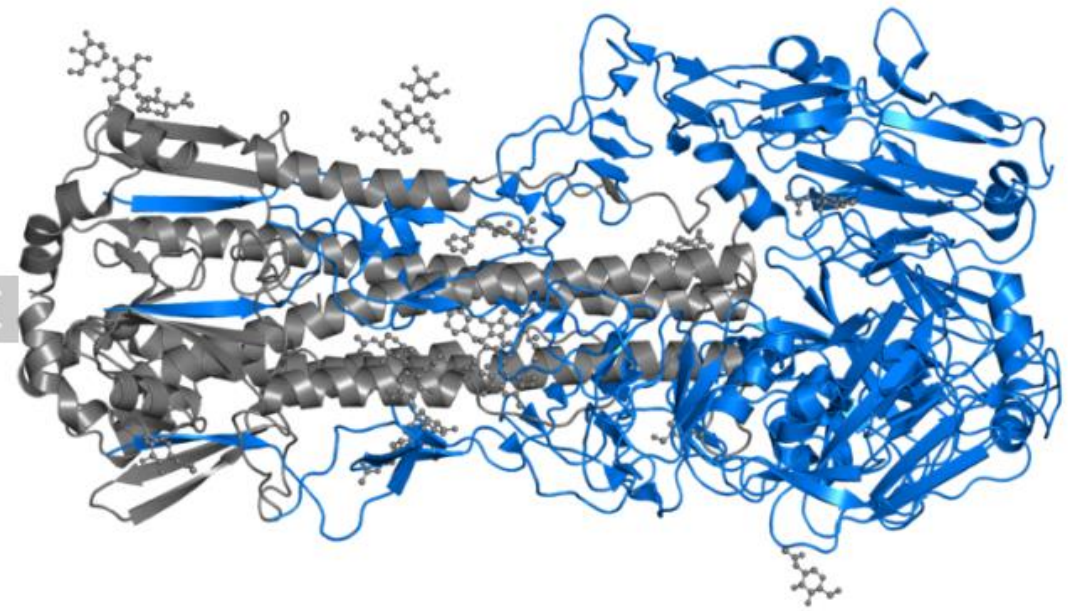
polysaccharide/protein complex
protein/protein complex
protein structure

Assembly polymer count (13)

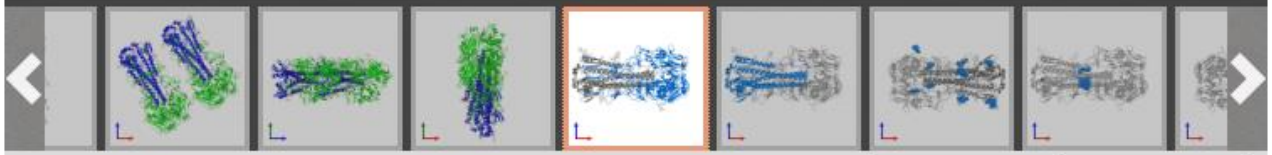
Homo / hetero assembly (2)

Journal (35)

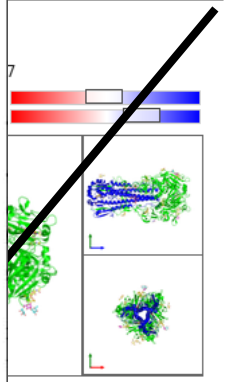
Resolution distribution

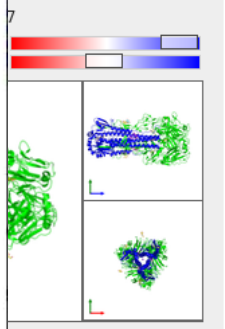


PDB 5t6s contains 3 copies of Hemagglutinin in assembly 2. This protein is highlighted and viewed from the front.



Release date (desc)



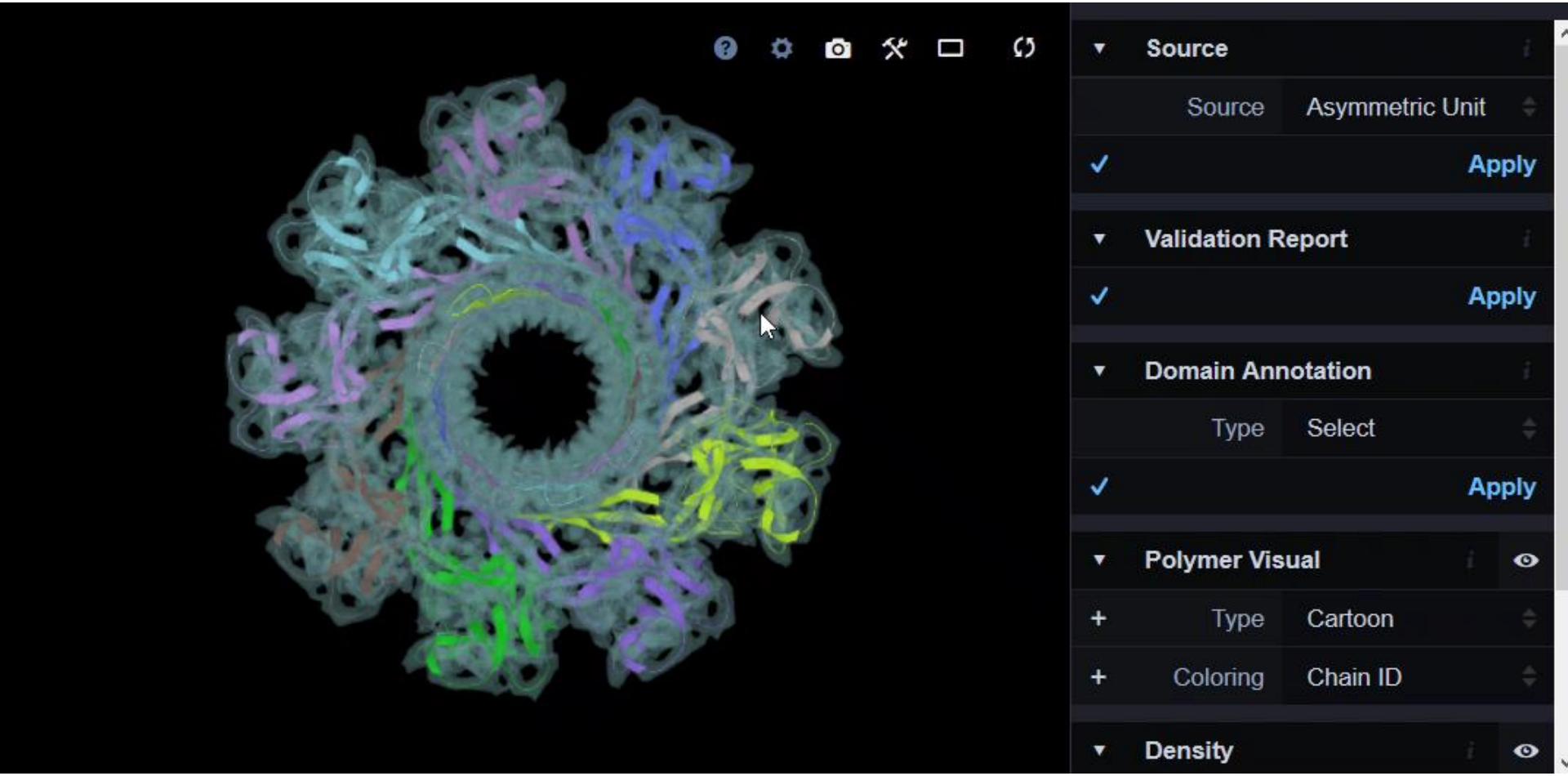


EBI

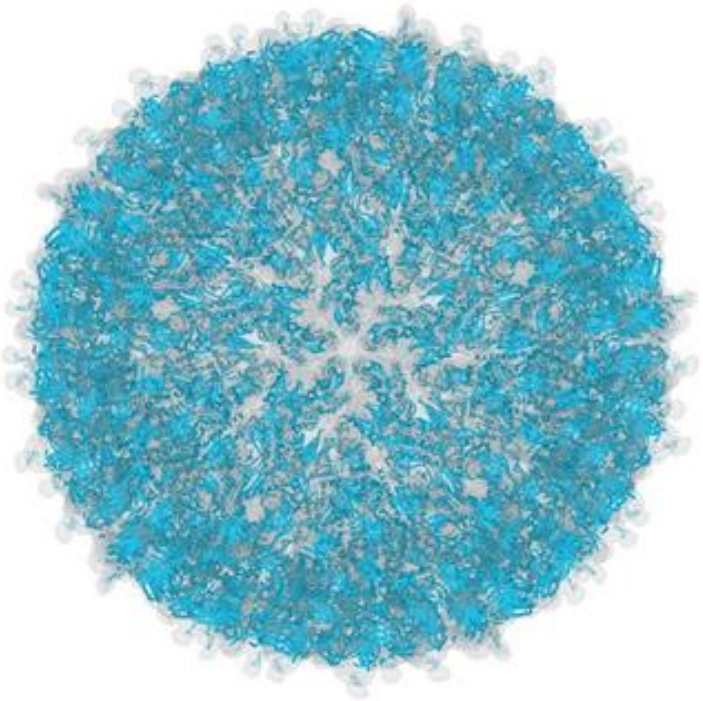
pdbe.org/2az9

EMBL-EBI

Also shows maps for EM entries



Improvements for viewing entries



Cryo-EM structure of Zika Virus

Cartoon model + EM Surface

mmCIF+CCP4

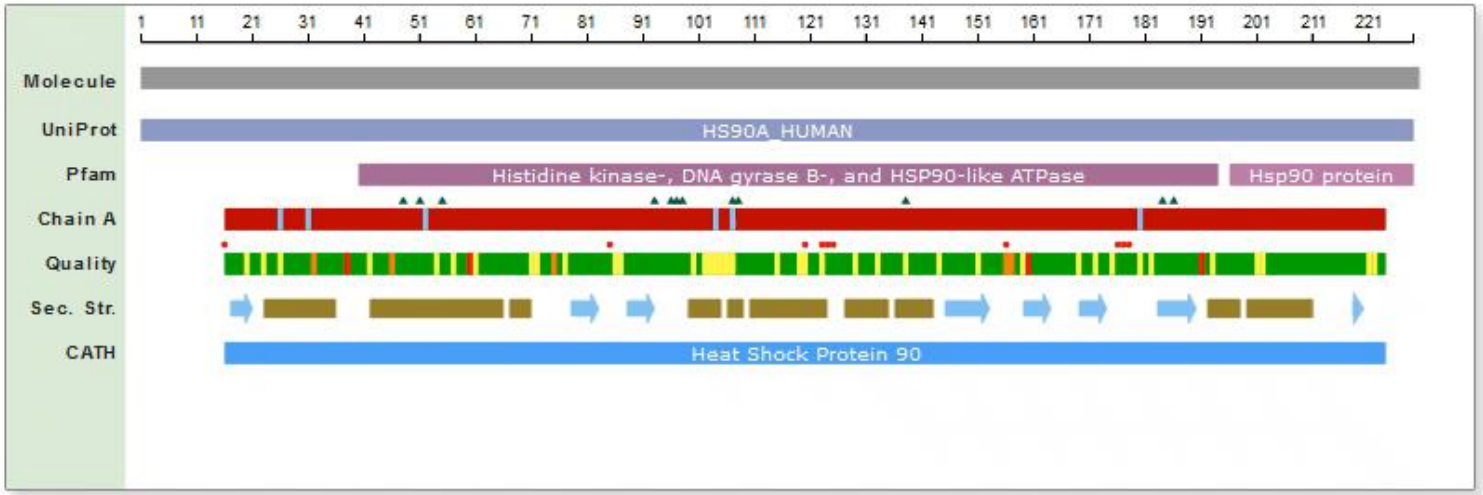
All data

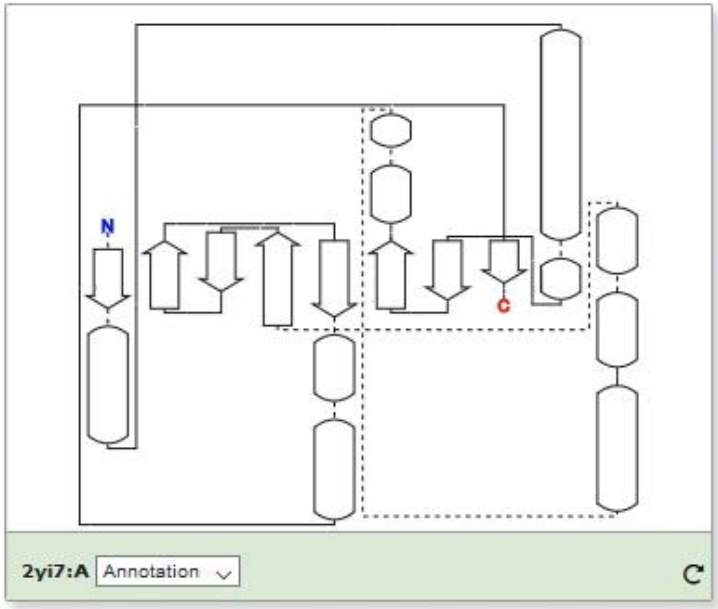
1572× less

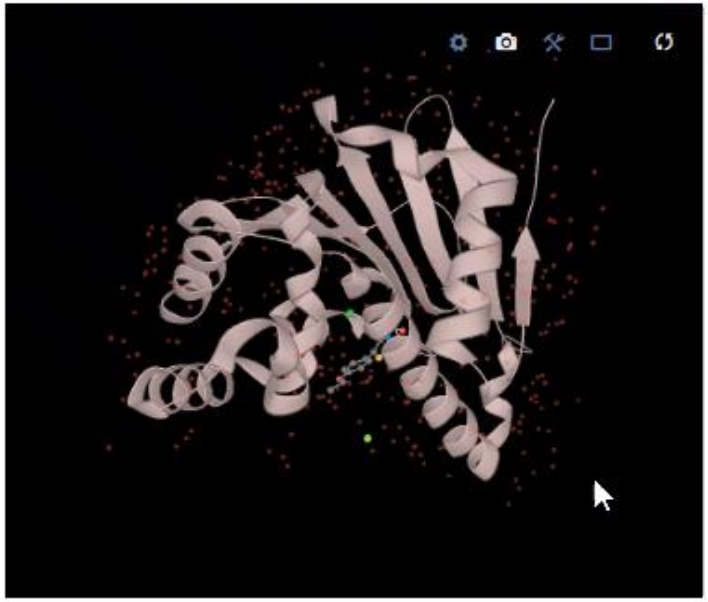
BinaryCIF

CS+DS

Linking sequence, topology and structure

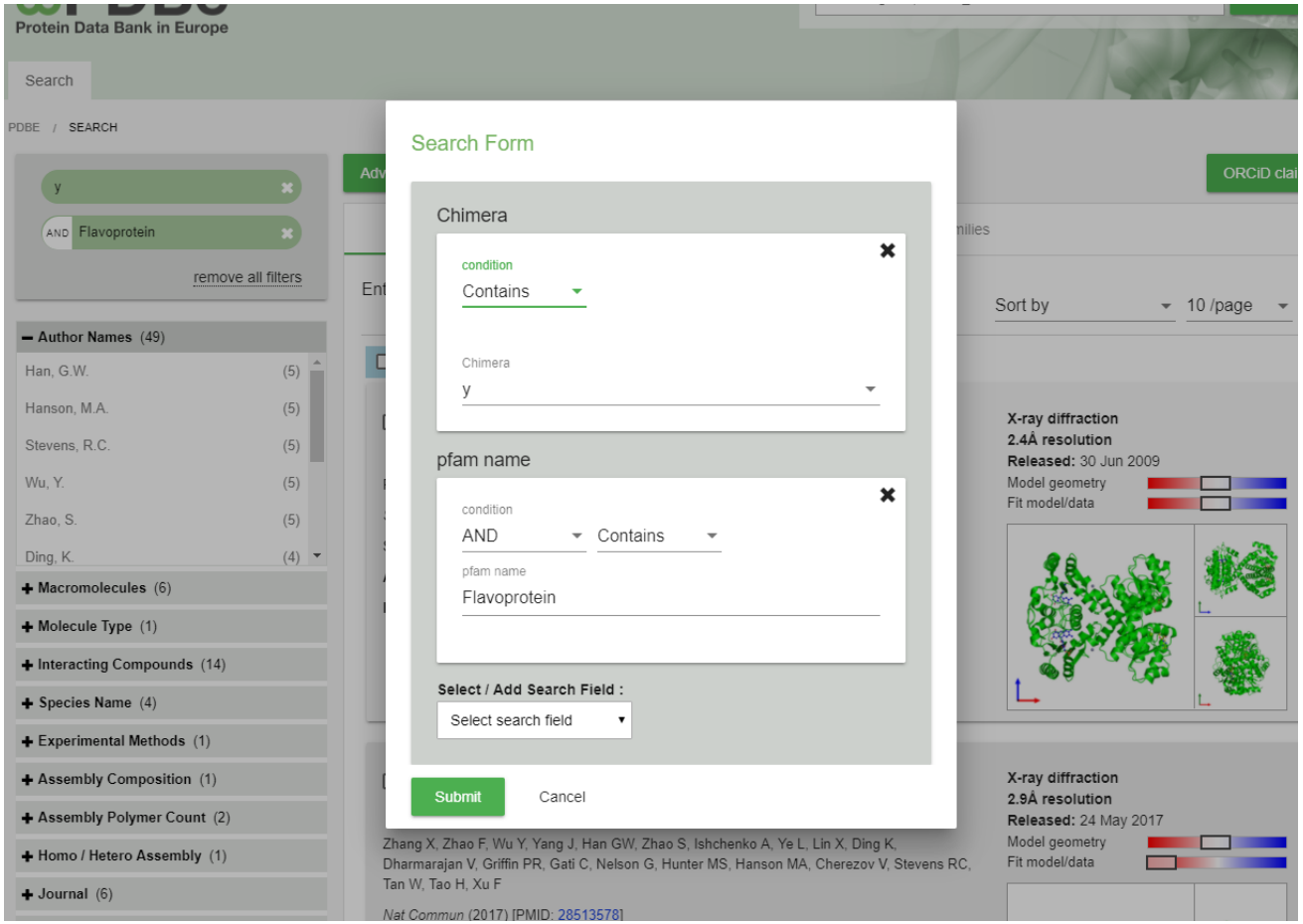




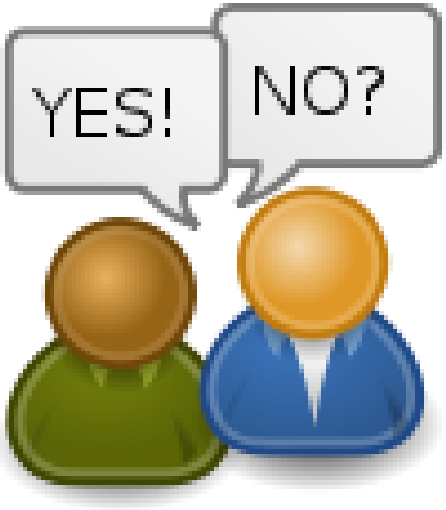


New features – available for testing and feedback

- Advanced search
- https://wwwdev.ebi.ac.uk/pdbe/widgets/advance-search-test/search



USEFUL STUFF



- Use PDBe because it's useful
- If it isn't, tell us how to improve

Thanks!

WORLDWIDE PDB
PROTEIN DATA BANK

CEITEC | Central European Institute of Technology
BRNO | CZECH REPUBLIC

EMBL

wellcome trust

West-Life

You, for providing the data!

pdbhelp@ebi.ac.uk

proteindatabank

@PDBeurope

PDBe.org

EMBL-EBI