

Protein Data Bank in Europe

Depositing your data to the PDB

John Berrisford

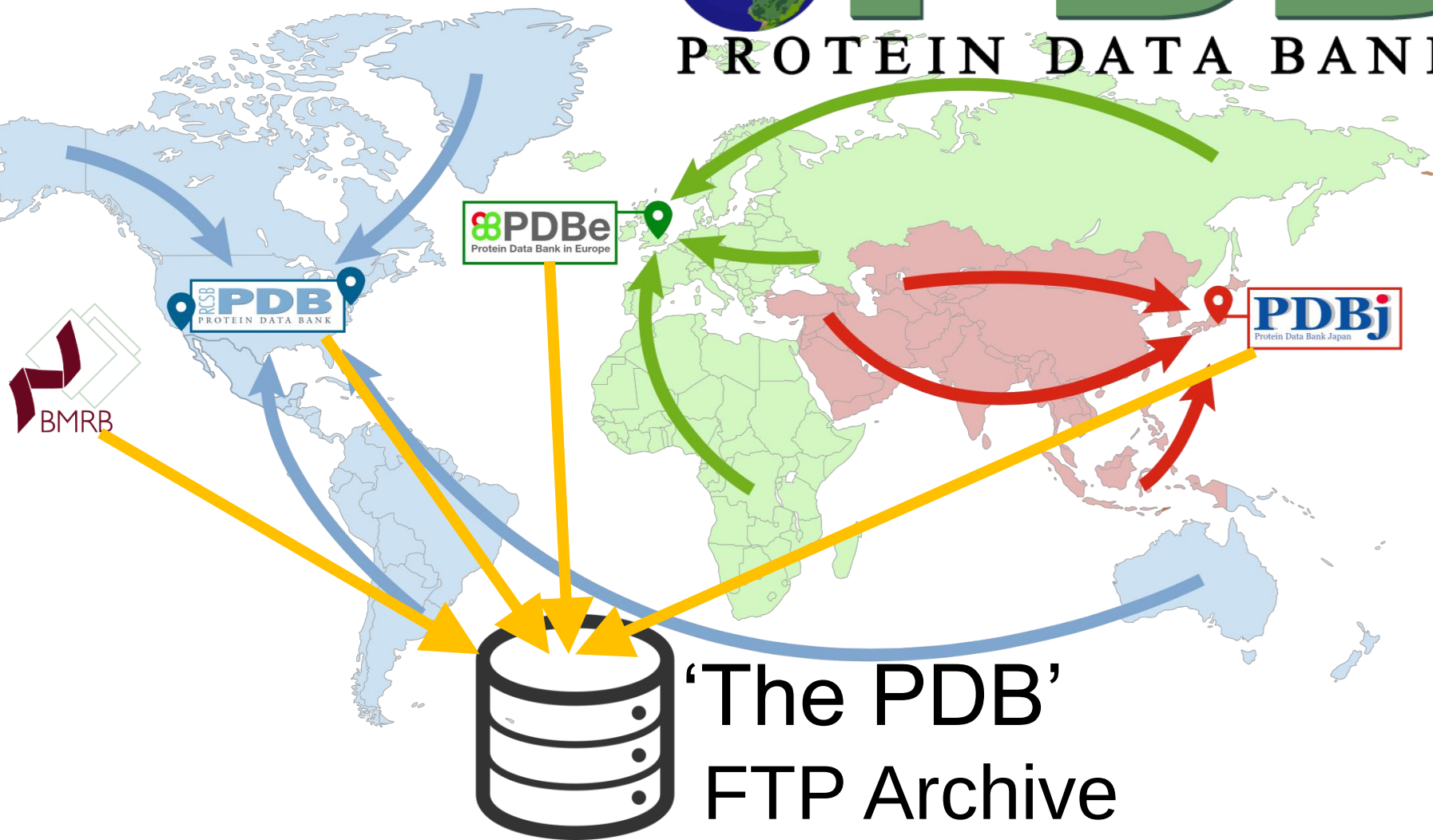
 pdhelp@ebi.ac.uk

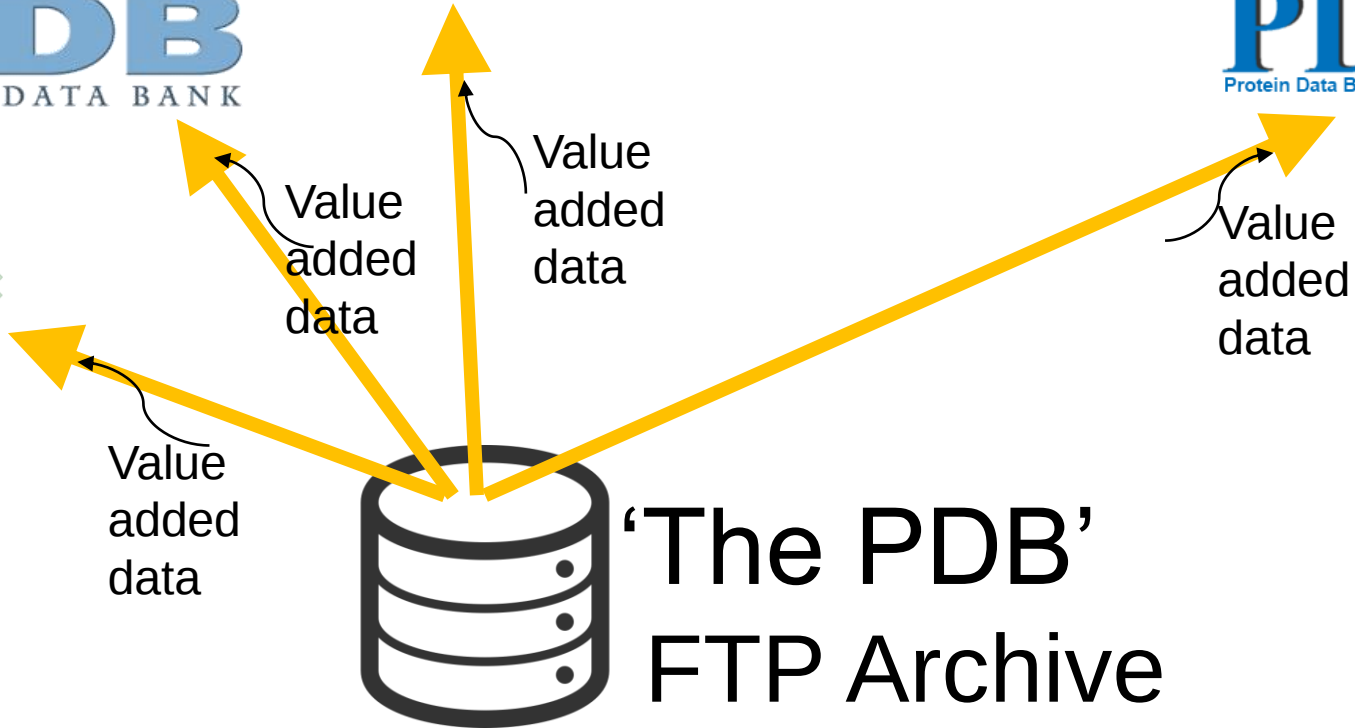
 [proteindatabank](https://www.facebook.com/proteindatabank)

 [@PDBEurope](https://twitter.com/PDBEurope)

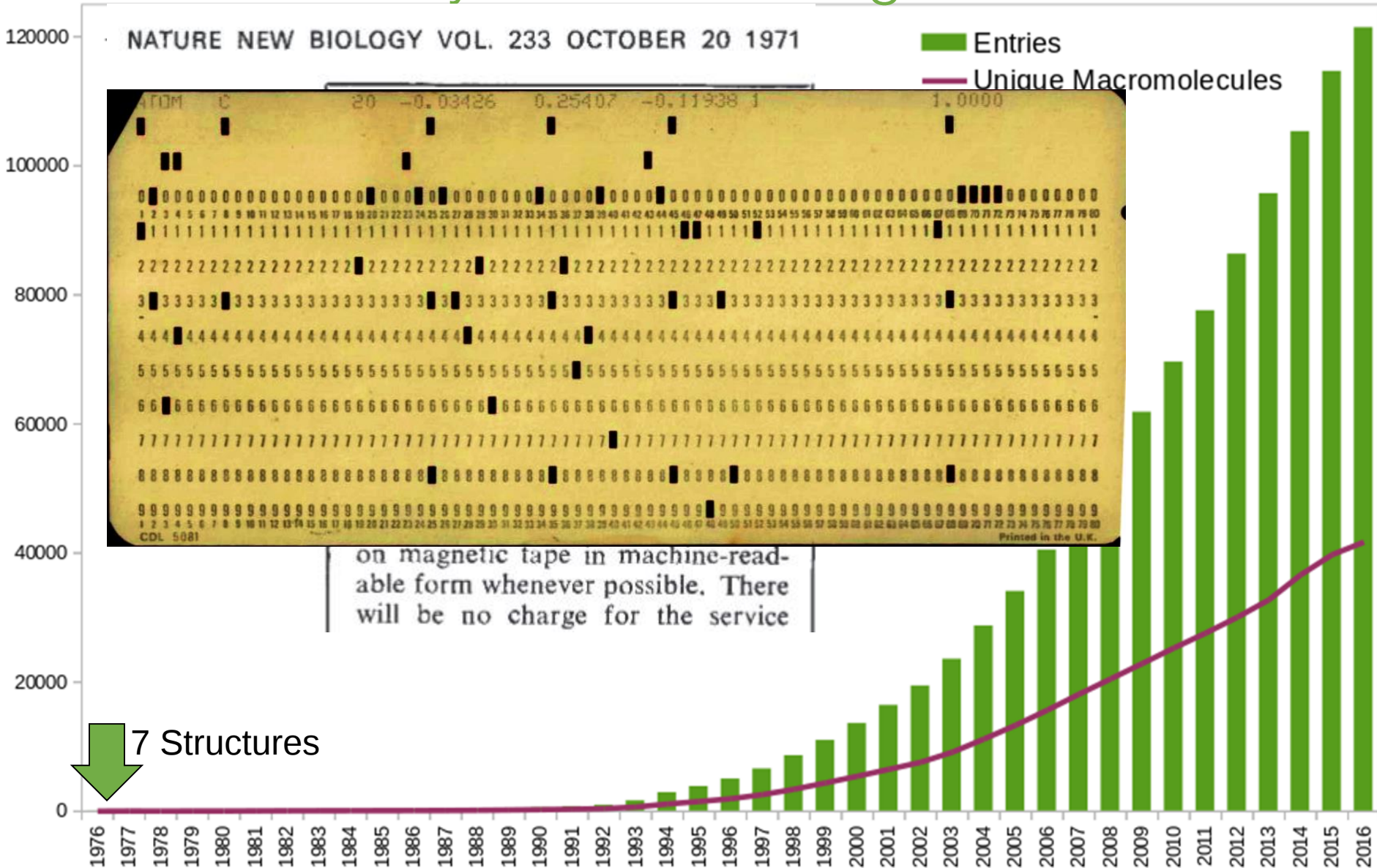
Who are the PDB?

W O R L D W I D E
wwPDB
P R O T E I N D A T A B A N K





Oldest freely available biological database



Why deposit to the PDB?

- To make your data publicly available
- Required by journals

OneDep deposition system



List requirements

All items

Mandatory items

Navigation

- ✓ Instructions
- ✓ Communication
- ✓ **Re-upload files**
- ✓ Upload summary
- ✚ Admin
 - ✓ Contact information
 - ✓ Grant information
 - ✓ Release status
 - ✓ Entry title & author
 - ✓ Citation information
- ✚ Macromolecules
 - ✓ 1) AMMONIA CHANNEL
 - ✓ 2) NITROGEN REGULATORY
- ✚ Data collection
 - ✓ Crystal Information
 - ✓ Collection Source
 - ✓ Software Used
 - ✓ Collection Statistics
- ✚ Refinement
 - ✓ Refinement
 - ✓ Ligands
 - ✓ Biological assembly
 - ✓ Validation reports
 - ✓ Summary & conditions
- ✚ Downloads & reports
 - ☐ All files
 - ☐ Generated mmCif

Log out

Replacement sections and re-upload files

General upload instructions

- Click 'Browse' to upload your file. Once the file is uploaded, select the file type from the pull-down list. If you have uploaded more than one file of each type, use the check box to select which file should be used.
- After upload, you must review the summary page carefully as it will tell you whether your data has been uploaded and interpreted correctly.
- The gzip and bzip2 compression formats are supported for all uploaded files. Archive formats such as tar and windows ZIP archives are not supported.

Coordinate upload instructions

Coordinates may be deposited as either mmCIF or PDB formatted files. We encourage you to use [pdb_extract](#) to prepare mmCIF formatted file. If you are using a PDB formatted coordinate file, please check the following PDB format requirements prior to file upload:

- A TER record is placed at the end of every polymeric chain.
- TER records are not present within a polymeric chain. Some refinement packages insert extra TER records at gaps within a chain. Please remove the extra TER records.
- No TER records should be included at the end of non-polymer residues such as ions, ligands, or waters.
- **PDB format files with misplaced TER records will result in incorrect format translation and data extraction - which may result in data loss during annotation.**
- Depositions comprised of multiple models should include MODEL and ENDMDL records. The models should be listed sequentially in columns 11-14.
- If there are alternate conformations in the structure, the alternate conformation indicator must be provided in column 17 of ATOM and HETATM records.
- There should be only one END record at the end of the file.

Browse... No file selected.

☒	pdb2nuu.ent	2.05 MB	Coordinates (PDB format)	
☒	r2nuusf.ent	4.21 MB	mmCIF (structure factors)	

Continue deposition (some sections will be modified)

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 pipeline 1.0

MolProbity

Xtrriage

EDS

Mogul

Percentiles

PDF maker

Geometry

Data
quality

Fit to
data

Ligand
geometry

Comparison
to other
PDB entries

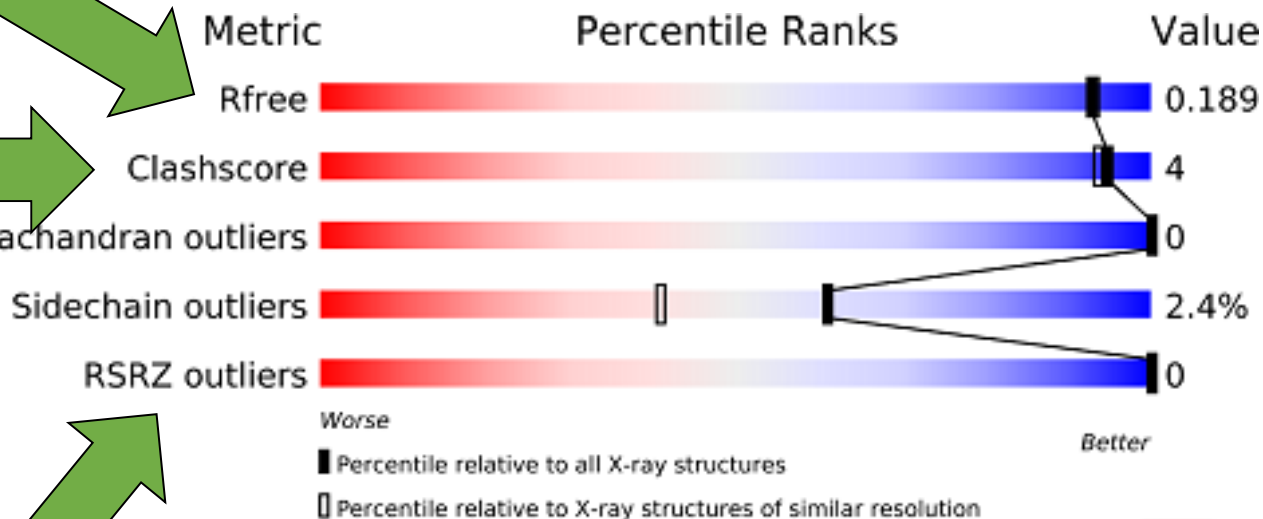
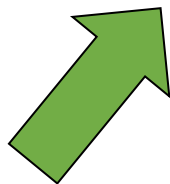
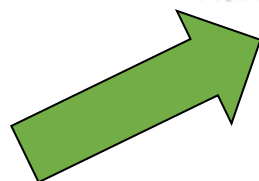
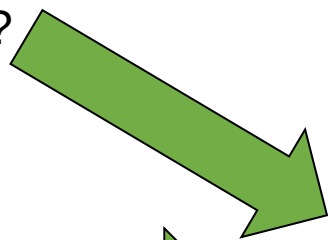
wwPDB validation

How well does the model
back-predict the data?

Atoms bumping into
each other

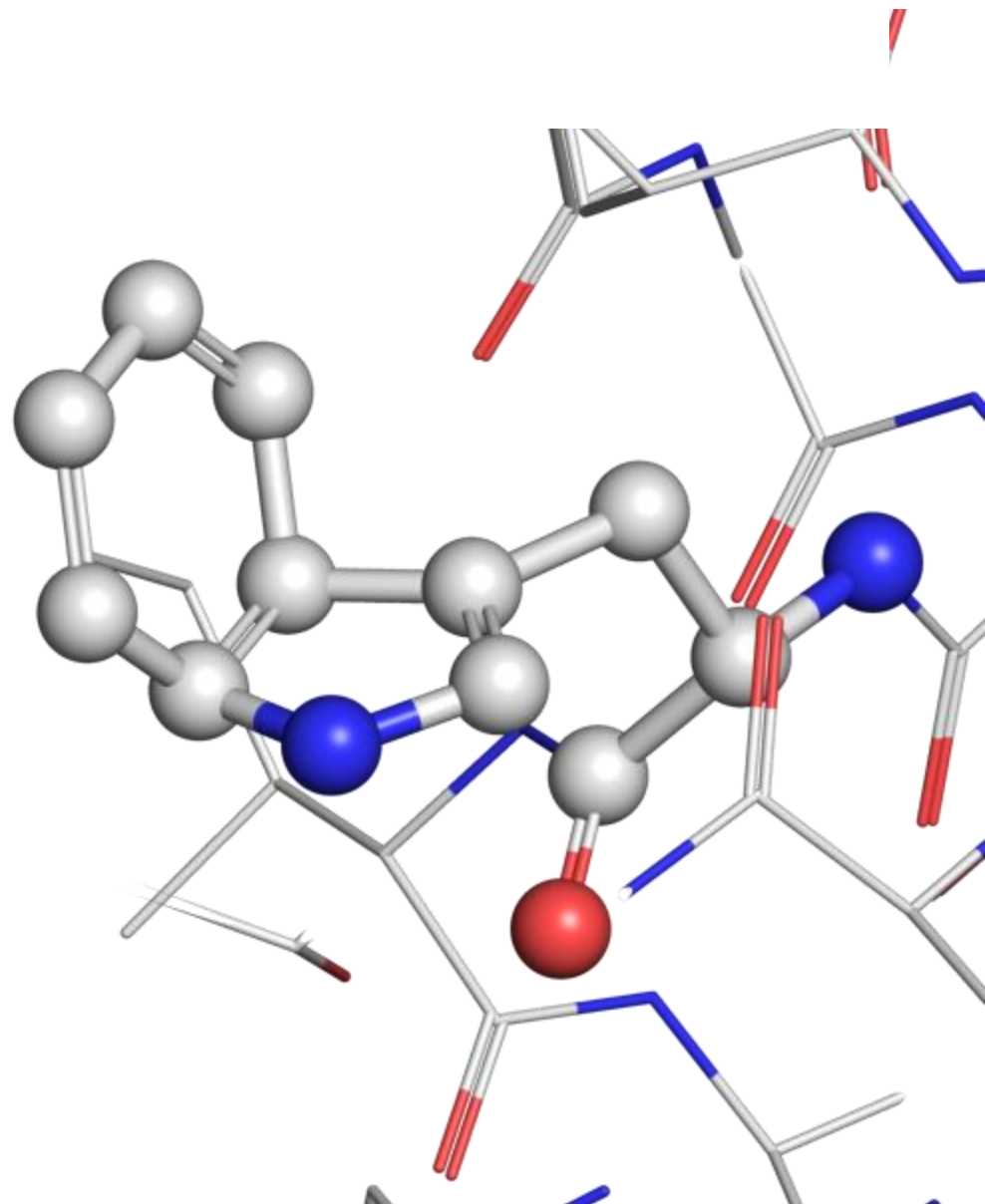
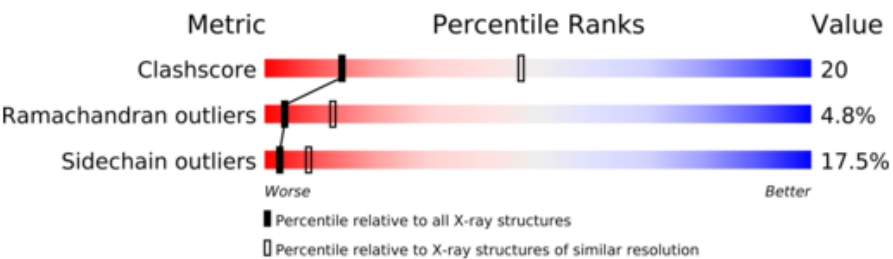
Surprising bond
angles

Atoms not in electron density



Interesting geometry

7GPB 2.9Å



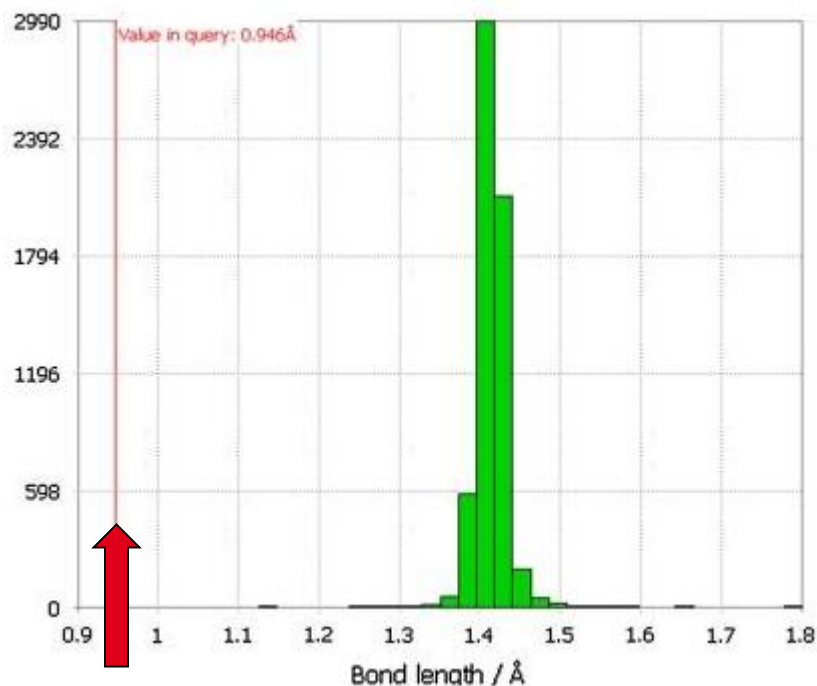
Validation for small molecules

Geometry



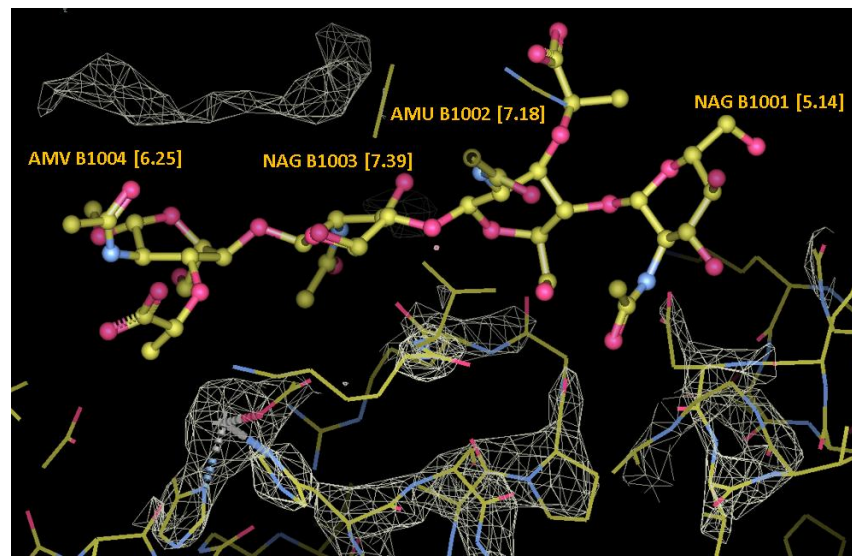
THE CAMBRIDGE
CRYSTALLOGRAPHIC
DATA CENTRE

Using Mogul

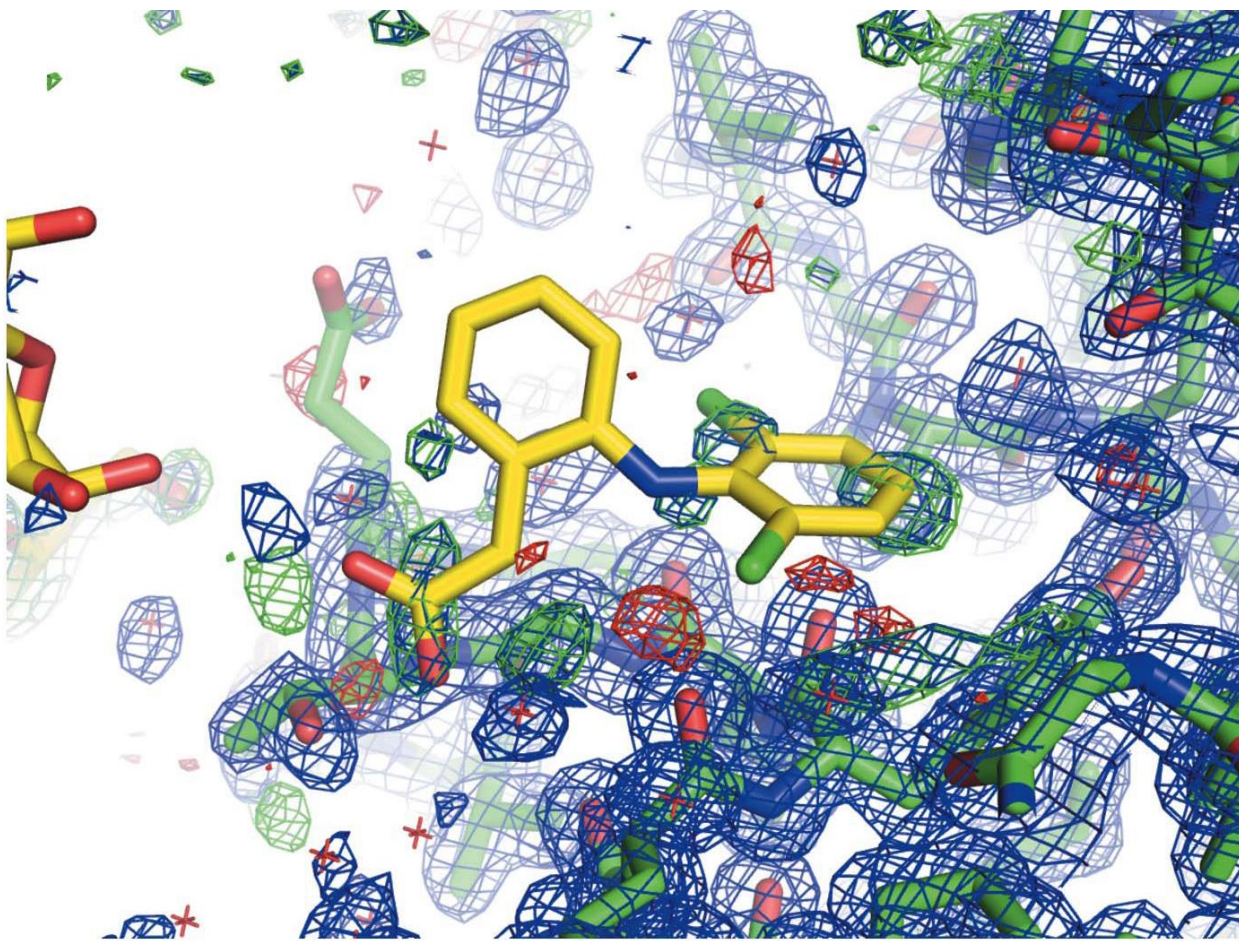


Fit to density

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



Poor density for ligand

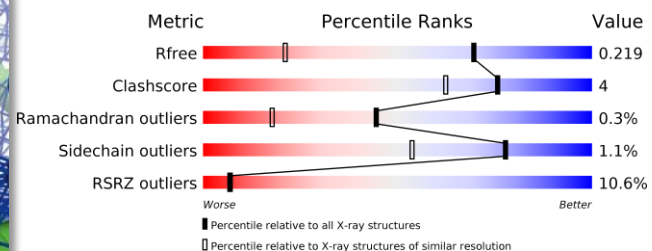


PDB: 3IB0 (1.4Å)

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

List requirements

All items

Mandatory items

Navigation

Instructions

Communication

Re-upload files

Upload summary

Admin

Contact information

Grant information

Release status

Entry title & author

Citation information

Macromolecules

1) SERINE PHOSPHATASE

Data collection

Crystal Information

Collection Source

Software Used

Collection Statistics

Refinement

Refinement

Ligands

Assembly

Related entries

Validation reports

Summary & conditions

Log out

Download validation reports

The status of the validation calculation is: finished

In order to submit your structure, download and review the validation report.

Validation report:

D_8211000289_val-report-full_P1.pdf.V1 (Fri Nov 3 09:56:26 2017)

XML file:

D_8211000289_val-data_P1.xml.V1 (Fri Nov 3 09:56:26 2017)

Overall quality at a glance:

Metric	Percentile Ranks	Value
Rfree		0.187
Clashscore		16
Ramachandran outliers		0
Sidechain outliers		0.7%
RSRZ outliers		5.2%

Worse

Better

■ Percentile relative to all X-ray structures

▨ Percentile relative to X-ray structures of similar resolution

Major validation issues

Geometry validation issues

The following major geometric issues were identified in the polymers in your entry. Please check the following issues carefully. If required upload corrected coordinates.

Clashes

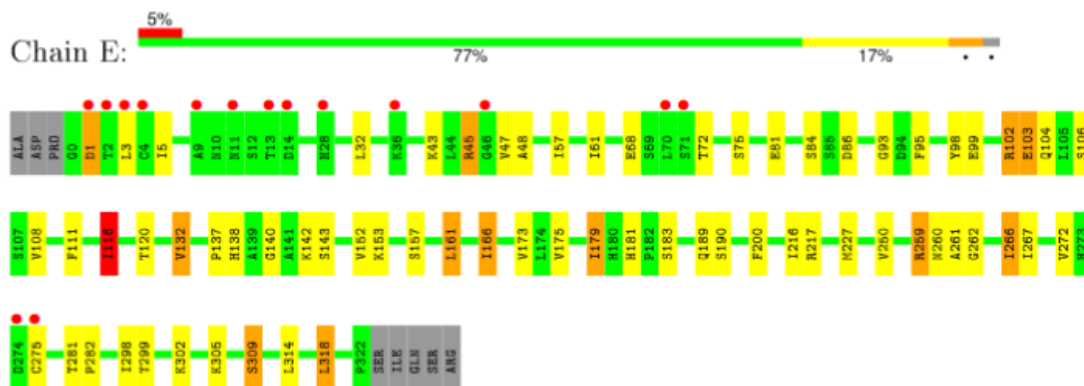
The following atoms were found to be less than 2.2 Å apart and are considered too close unless they are covalently bonded.

Residue(model/chain/number/name)	Atom	Residue(model/chain/number/name)	Atom	Distance between the atoms
1/A/56/SER	OG	1/A/71/LEU	HD13	1.58
1/A/90/ASP	OD2	1/A/2086/HOH	O	1.88
1/A/2085/HOH	O	1/A/2086/HOH	O	2.16

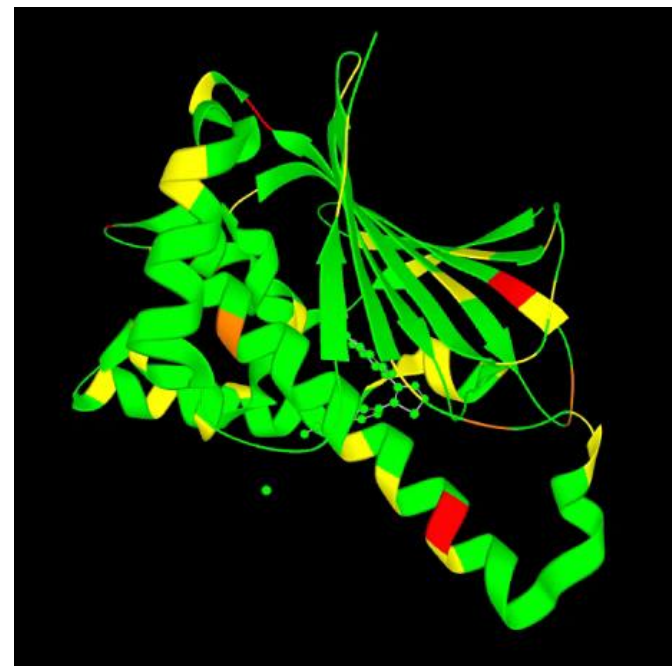
Please click the link at the top of the page to view your validation report.

Why give you all this validation upfront?

- Molecule 1: Hemagglutinin



- 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

Sanity Check

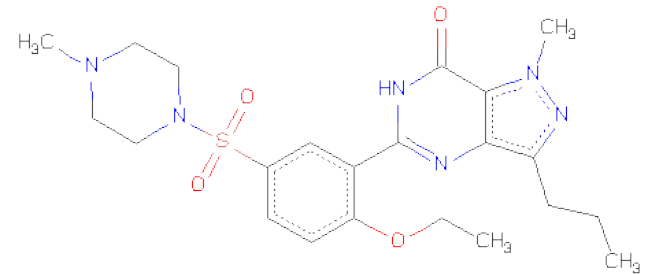
Controlled vocabulary

Continue to next section

Biocuration (annotation)

Ligand matching to dictionary

All identical chemicals have the same code



Taxonomy

Sequence matching to UniProt

Highlight mutations, conflicts, tags etc

AUTHORITY	ENTRY	SEQUENCE	RESIDUE	ANNOTATION
1	MET	SP-P62161	-	initiating methionine
2	HIS	SP-P62161	-	expression tag
3	HIS	SP-P62161	-	expression tag
4	HIS	SP-P62161	-	expression tag
5	HIS	SP-P62161	-	expression tag
6	HIS	SP-P62161	-	expression tag
7	HIS	SP-P62161	-	expression tag
8	GLY	SP-P62161	-	expression tag
9	SER	SP-P62161	-	expression tag
68	ASP	SP-P62161	-	engineered mutation
155	ALA	SP-Q55629	-	linker
156	LYS	SP-Q55629	-	linker
157	LEU	SP-Q55629	-	linker
158	GLU	SP-Q55629	-	linker
159	CYS	SP-Q55629	-	linker
174	HIS	SP-Q55629	-	engineered mutation
241	ARG	SP-Q55629	-	linker
242	ARG	SP-Q55629	-	linker
243	GLY	SP-Q55629	-	linker
244	GLY	SP-Q55629	-	linker
245	THR	SP-Q55629	-	linker

Assemblies

Standardising of representations – helping make PDB searchable

Sanity - is that really purified from human hearts?

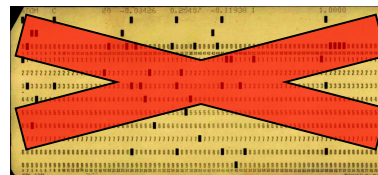
4.5 FTEs for all of Europe and Africa- >4000 structures a year

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

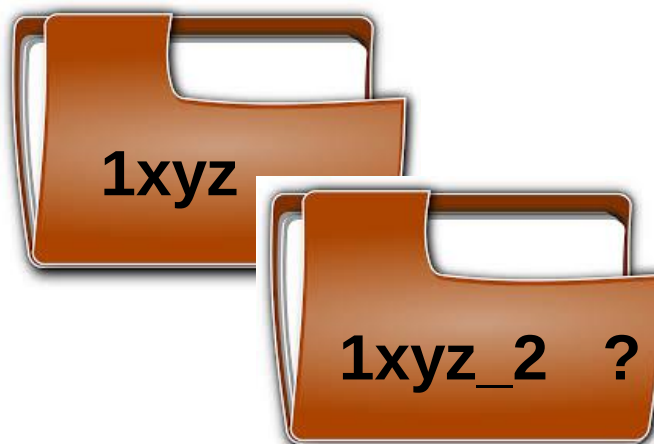


Coming soon

Versioning

Update released entries
And retain the ID code

Details yet to be finalised



Getting data out of the PDB



pdhelp@ebi.ac.uk



[proteindatabank](https://www.facebook.com/proteindatabank)



[@PDBEurope](https://twitter.com/PDBEurope)



Protein Data Bank in Europe

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

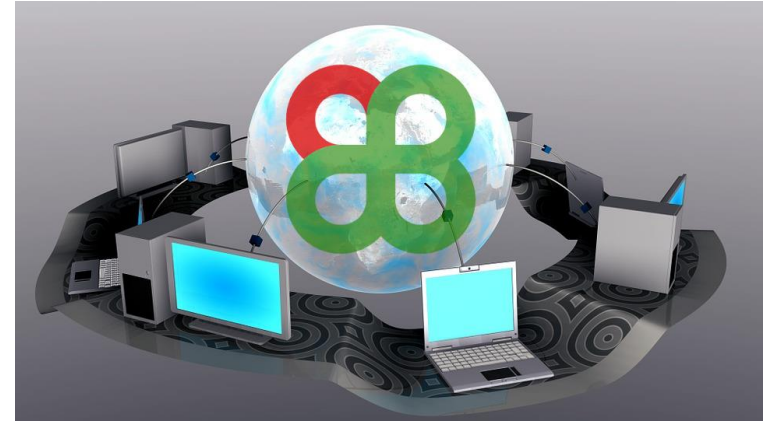


Challenges – from archive to resource

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?

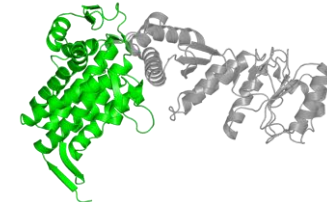
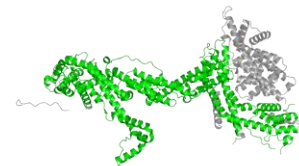
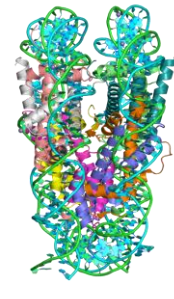
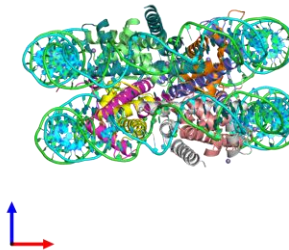
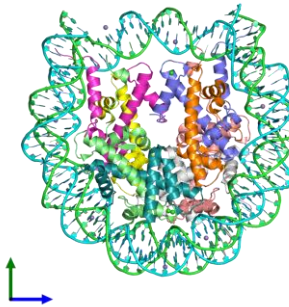
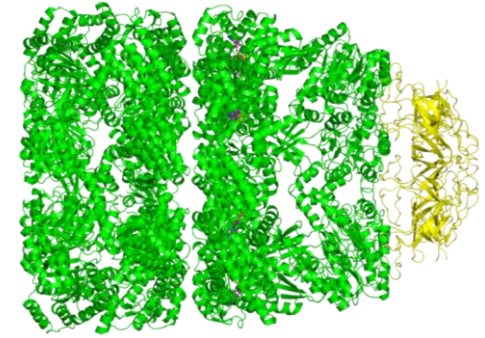
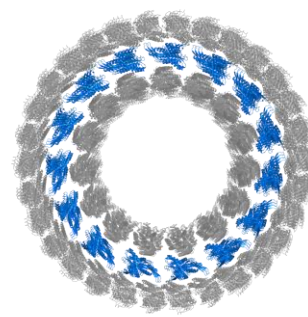
DATA



- 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





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)
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 (IHNV) (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

Latest archive statistics

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](#)).

[Previous featured structures](#)

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)

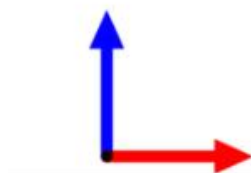
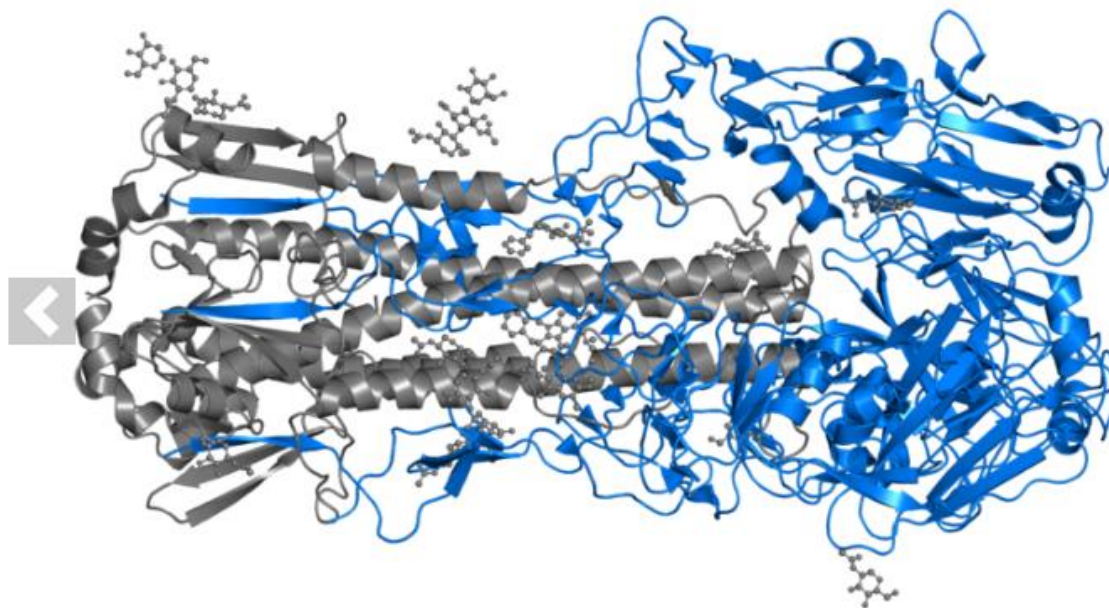
polysaccharide/protein complex
protein/protein complex
protein structure

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.



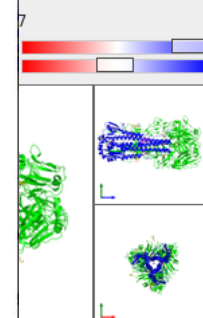
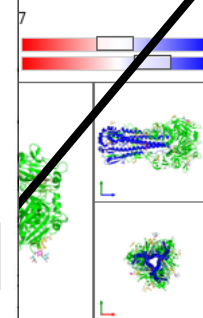
Search

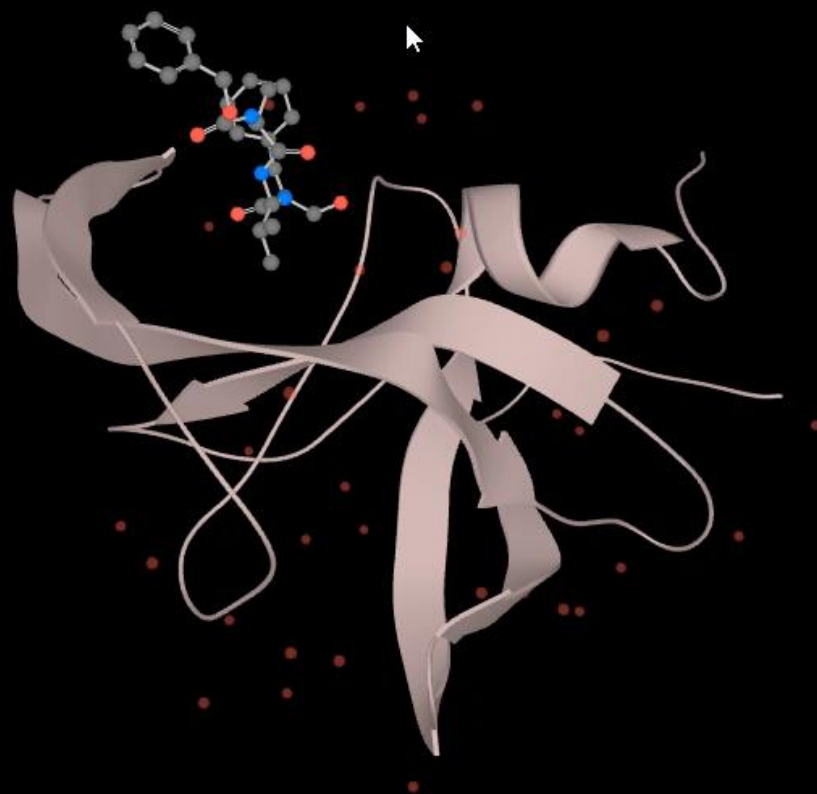
Search EMBL

Feedback

View basket (0)

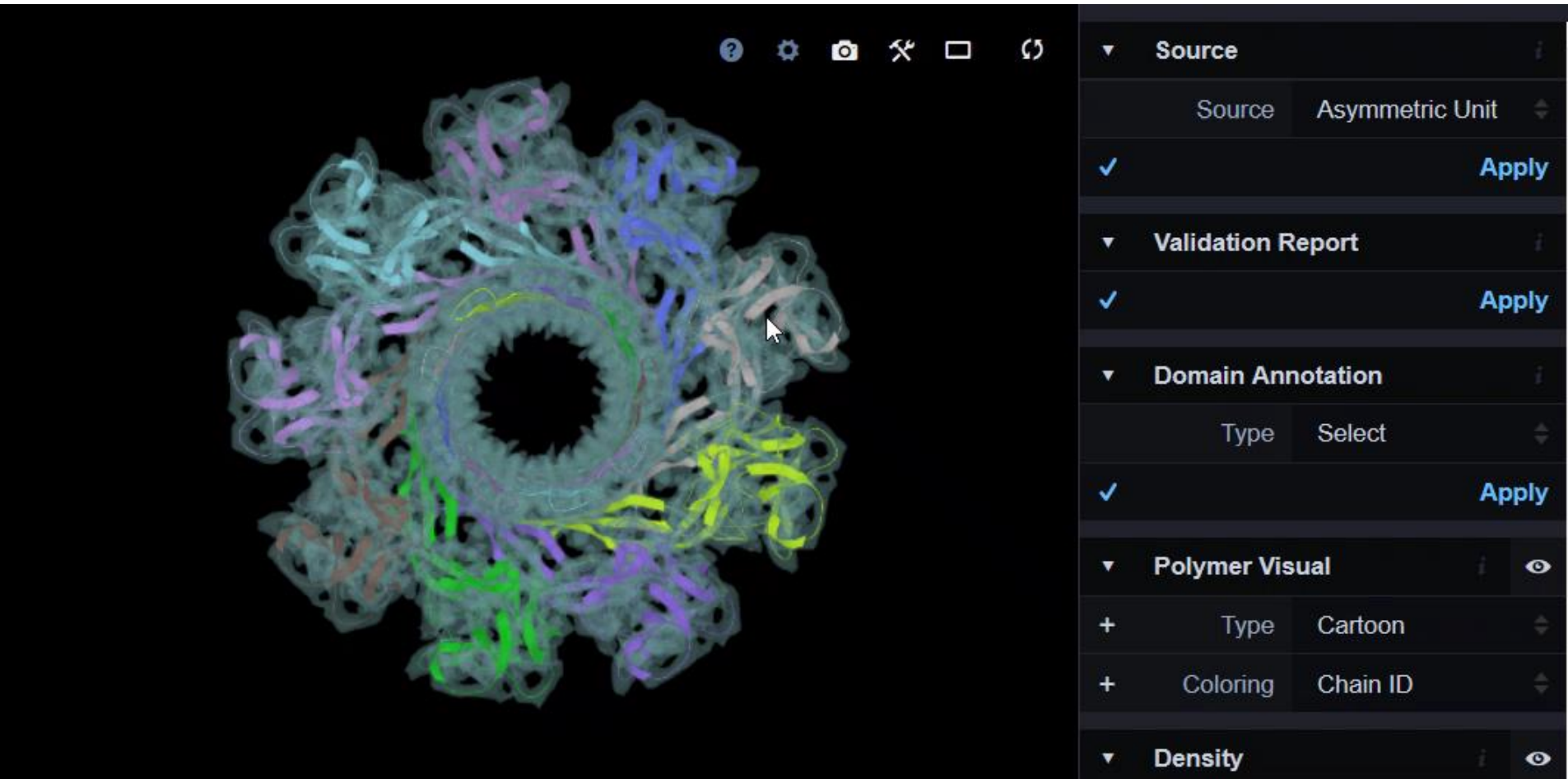
Release date (desc)



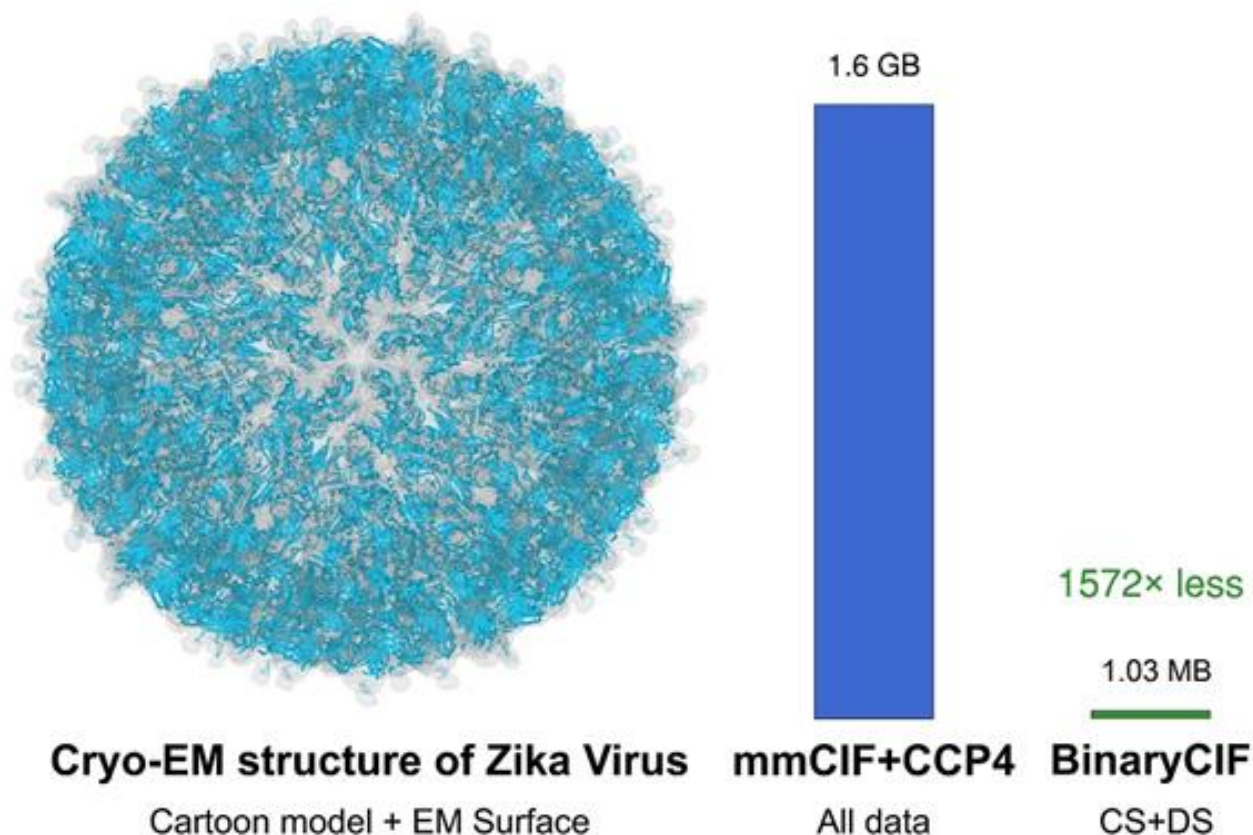


Source		
Source	Asymmetric Unit	
✓		Apply
Validation Report		
✓		Apply
Domain Annotation		
Type	Select	
✓		Apply
Polymer Visual		
+	Type	Cartoon
+	Coloring	Chain ID
HET Groups Visual		
+	Type	Balls and Sticks
+	Coloring	Element Symbol
Water Visual		
+	Type	Balls and Sticks
+	Coloring	Element Symbol
Density : 2Fo-Fc		
+	Iso Value (σ)	<input type="range"/> 1.5
+	Color	<input type="color"/>
Show	Around Selection	

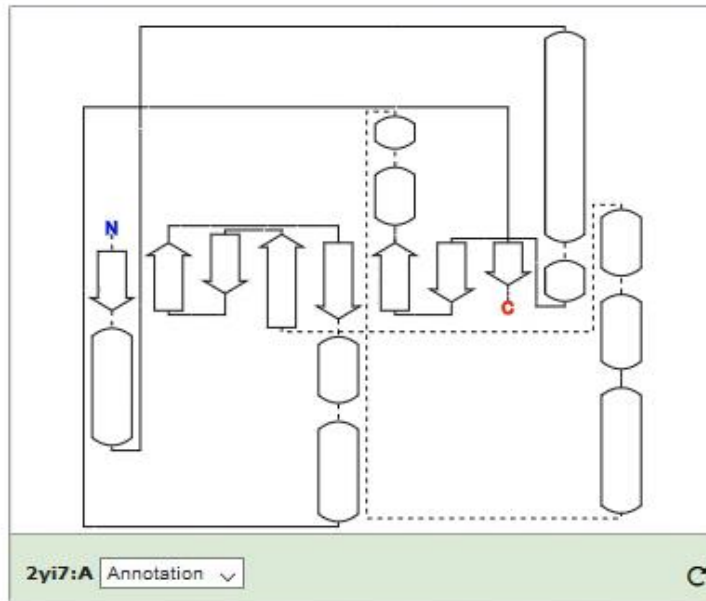
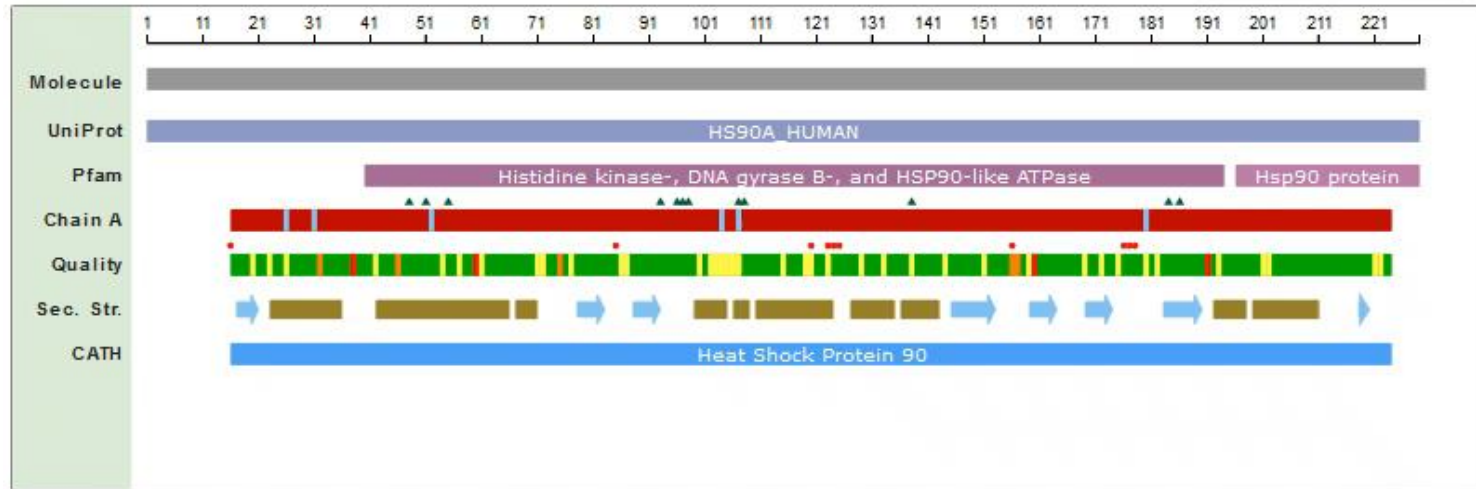
Also shows maps for EM entries



Improvements for viewing entries



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>

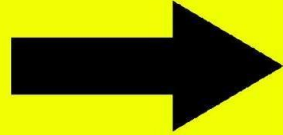
The screenshot displays the PDBE Advanced Search interface. A 'Search Form' modal window is open, showing two search criteria:

- Chimera**
 - condition: Contains
 - Chimera: y
- pfam name**
 - condition: AND Contains
 - pfam name: Flavoprotein

Below the criteria, there is a 'Select / Add Search Field' dropdown menu with the option 'Select search field'. At the bottom of the modal are 'Submit' and 'Cancel' buttons.

The background shows the PDBE search results page with filters on the left (Author Names, Macromolecules, Molecule Type, Interacting Compounds, Species Name, Experimental Methods, Assembly Composition, Assembly Polymer Count, Homo / Hetero Assembly, Journal) and search results on the right. The search results include a list of authors (Han, G.W., Hanson, M.A., Stevens, R.C., Wu, Y., Zhao, S., Ding, K.) and a list of macromolecules (6). The search results also show a 3D structure of a protein complex (X-ray diffraction 2.4Å resolution) and a 2D structure (X-ray diffraction 2.9Å resolution).

USEFUL STUFF



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



Thanks!



You, for providing the data!



pdbhelp@ebi.ac.uk



[proteindatabank](https://www.facebook.com/proteindatabank)



[@PDBEurope](https://twitter.com/PDBEurope)



PDBe.org

