EM data archiving

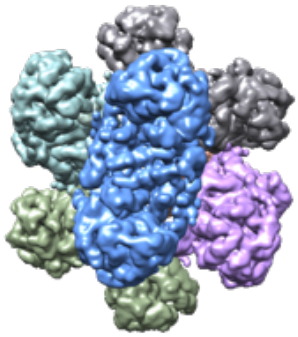
Gerard Kleywegt,
Ardan Patwardhan &
The EMDB & EMPIAR teams

EMBL-European Bioinformatics Institute
Hinxton, Cambridge, UK

EMBO Course, Birkbeck College
12 September 2019

EMBL-EBI

Outline



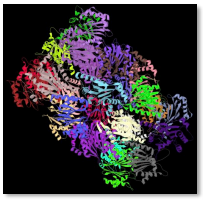
- Introduction to relevant archives
  - PDB
  - EMDB
  - EMPIAR
  - BioImage Archive
  - Other archives and resources
- What's in the pipeline?
- (Separate talk on searching/visualising/validating and depositing data to EMDB and EMPIAR by Osman Salih)

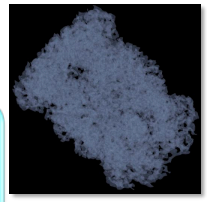
EMBL-EBI

Archives and other data resources at EMBL-EBI

EMBL-EBI

Structure and bioimage archives at EMBL-EBI



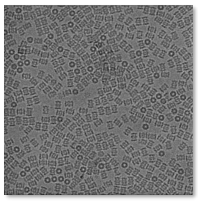


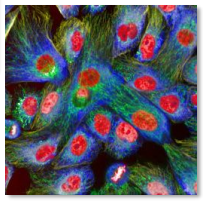
Protein Data Bank (PDB)

Electron Microscopy Data Bank (EMDB)

Electron Microscopy Public Image Archive (EMPIAR)

BioImage Archive

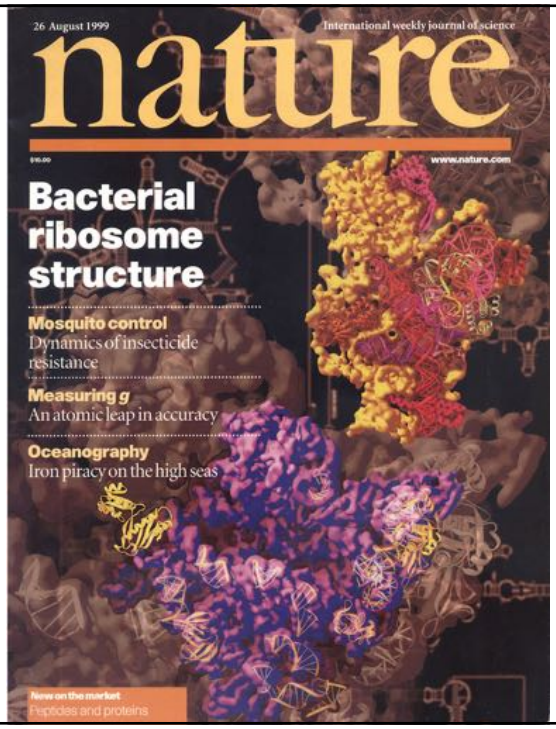




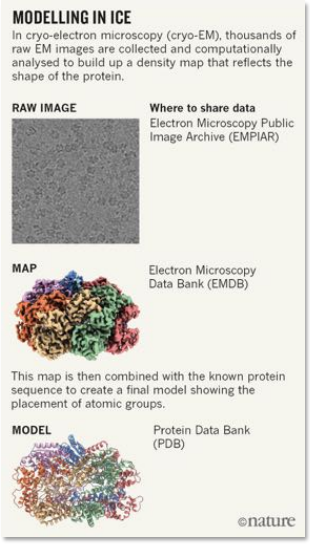
pdbe.org – emdb-empiar.org – empiar.org – www.ebi.ac.uk/bioimage-archive

EMBL-EBI

PDB –
Protein Data Bank



Molecular & Cellular Structure archives at EMBL-EBI



- PDB (est. 1971 at Brookhaven)
  - Atomistic models and some X-ray and NMR data
    - 155,800 entries and >0.5 TB data
  - EMBL-EBI involved since 1999
  - Managed by wwPDB partners since 2003
  - wwPDB: collaboration on all aspects of the PDB archive
    - Policies, procedures, formats, validation standards, ligands, journals, etc.
    - Exception: “data-out” (websites, APIs, value-added services, outreach, training, etc.)
    - Archive is identical at al sites!

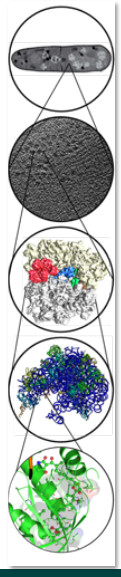
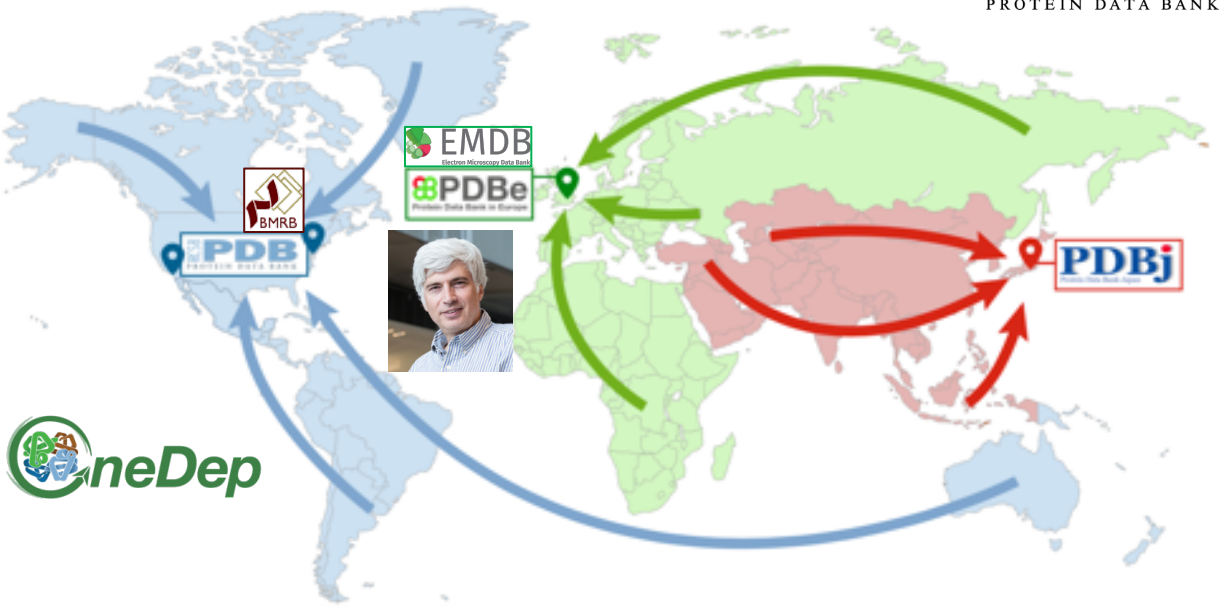


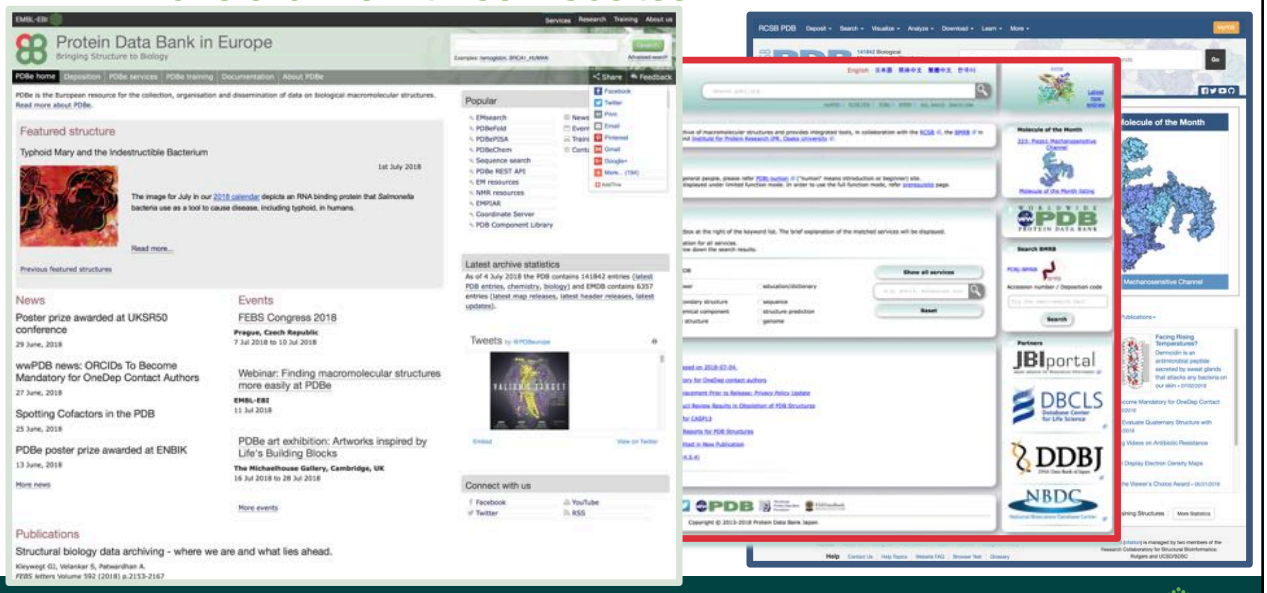
Image credit (left): *Nature* 561, 565-567 (2018)

WORLDWIDE PDB PROTEIN DATA BANK elixir EMBL-EBI

wwPDB in 2019: 3 core archives, 5 core members



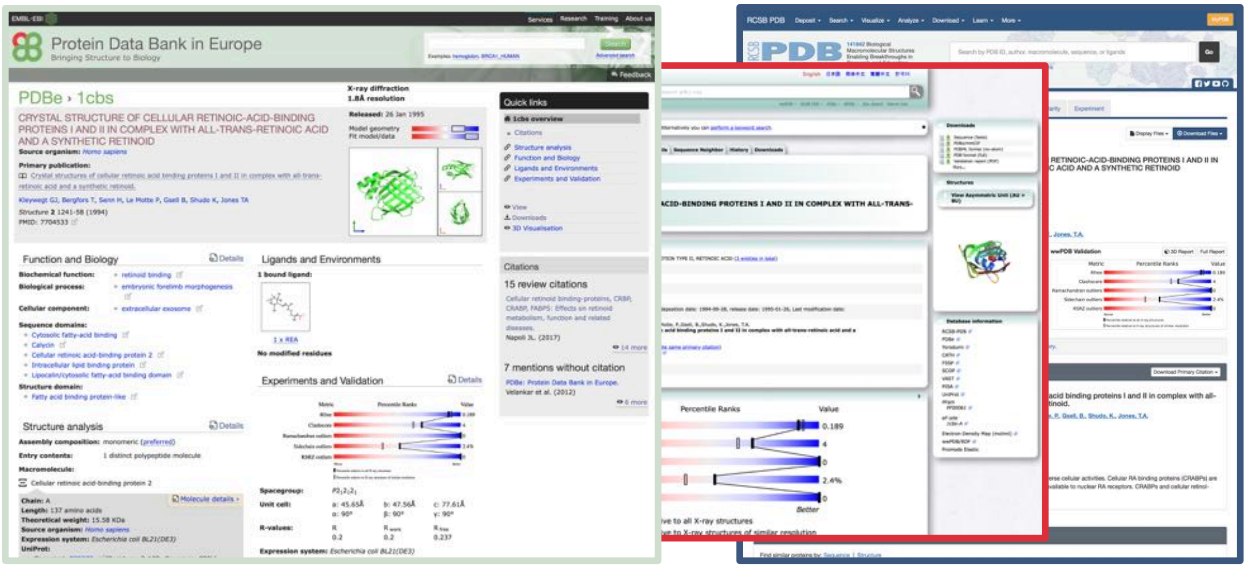
PDB: one archive – three websites



pdbe.org – pdbj.org – rcsb.org

EMBL-EBI

PDB: one archive – three websites

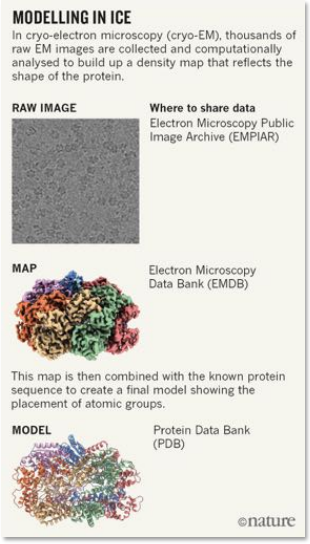


pdbe.org/1cbs – pdbj.org – rcsb.org

EMBL-EBI

EMDB – Electron Microscopy Data Bank

Molecular & Cellular Structure archives at EMBL-EBI



- EMDB (est. 2002 at EMBL-EBI)
  - Electron Microscopy Data Bank
  - Cryo-EM volume maps and tomograms
    - 9,016 entries and >1.2 TB data
  - Operated jointly by EMBL-EBI and RCSB PDB since 2007
    - PDBj involved since 2013
  - To become wwPDB Core Archive and Core Member in 2019

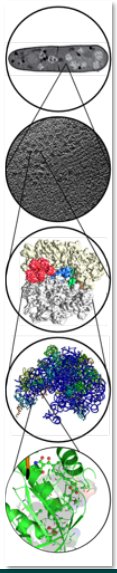


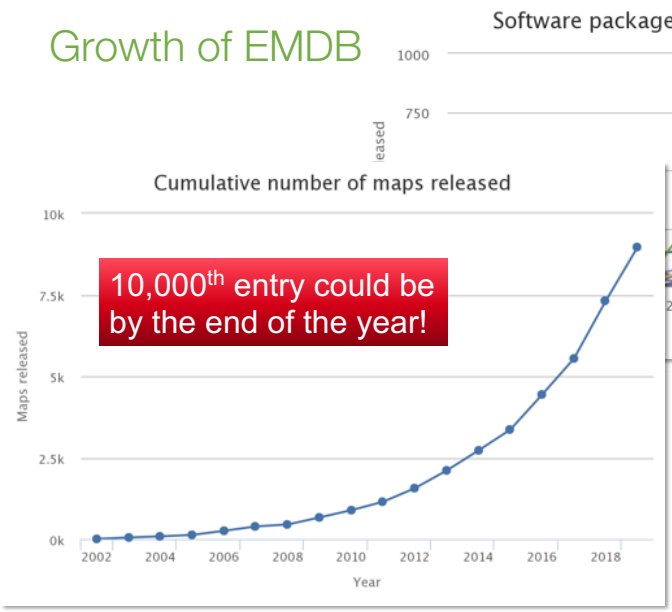
Image credit (left): *Nature* 561, 565-567 (2018)

WORLDWIDE PDB PROTEIN DATA BANK

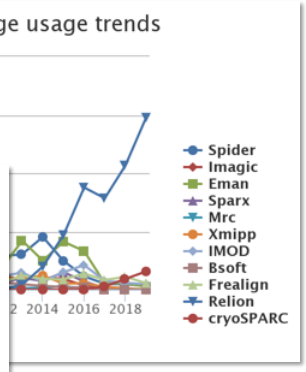
elixir

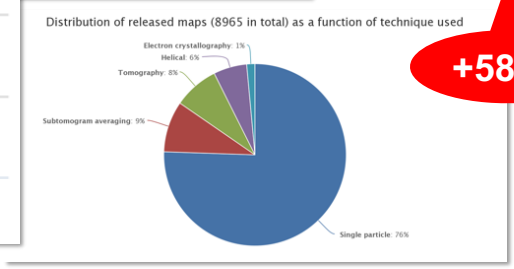
EMBL-EBI

Growth of EMDB



10,000th entry could be by the end of the year!





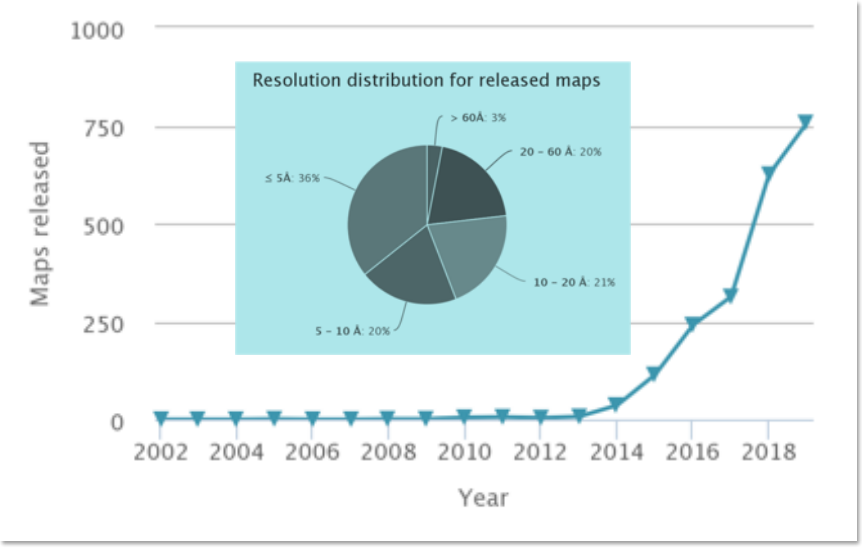
Year	EMDB depositions
2009	151
…	…
2015	780
2016	1074
2017	1390
2018	2197

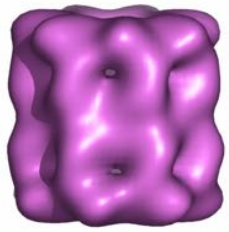
+58%!

emdb-empiar.org/emstats

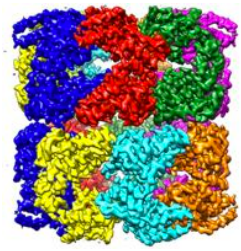
EMBL-EBI

The "resolution revolution" in cryo-EM





GroEL at 25Å in 2006 (EMDB:1291) and at 3.5Å in 2017 (EMDB:8750)

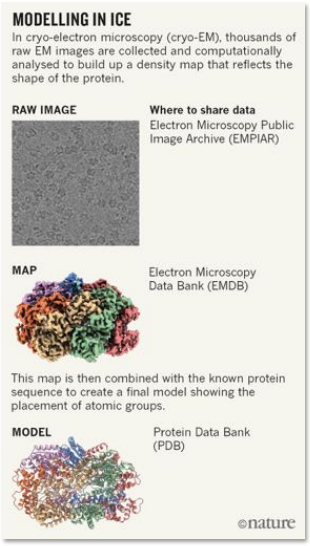


Number of maps at better than 4Å resolution released by EMDB per year

EMBL-EBI

EMPIAR – Electron Microscopy Public Image Archive

Molecular & Cellular Structure archives at EMBL-EBI



- EMPIAR (est. 2014 at EMBL-EBI)
  - Electron Microscopy Public Image Archive
  - Raw 2D image data (cryo-EM/ET), 3D volume maps (SXT, 3DSEM, …) and more
    - 240 entries and 200 TB data
  - Community-driven initiative
    - Established as a pilot archive in 2014
  - Used for validation, teaching/training, software/methods development, community challenges, …
  - Establishing new international collaborations
    - Mirror in Japan: empiar.pdbj.org
  - Part of EMBL-EBI BioImage Archive

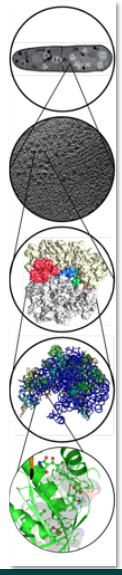
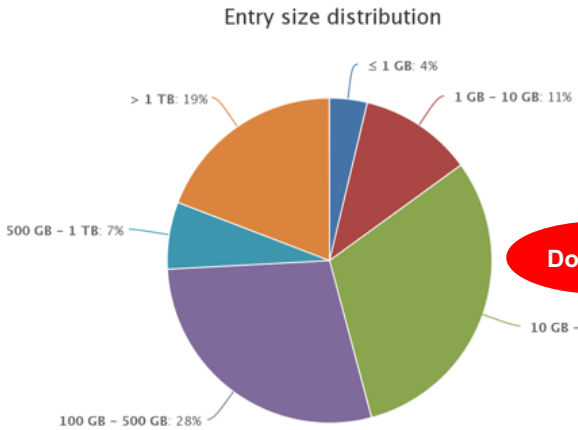


Image credit (left): *Nature* 561, 565-567 (2018)

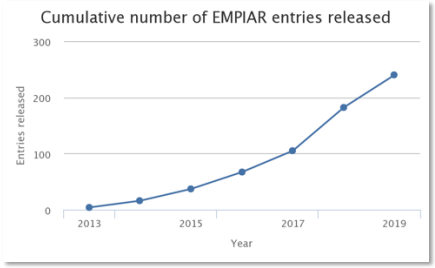
EMBL-EBI

Growth of EMPIAR

Year	EMPIAR releases	Average entry size (GB)	Size of largest entry (TB)
2013	4	1854	6.5
2014	12	258	1.7
2015	21	512	4.2
2016	30	764	12.4
2017	42	189	1.8
2018	79	777	11.7
(2019	58	1500	11.3)



Doubling!



emdb-empiar.org/emstats

EMBL-EBI

EMPIAR website

empiar.org

EMBL-EBI

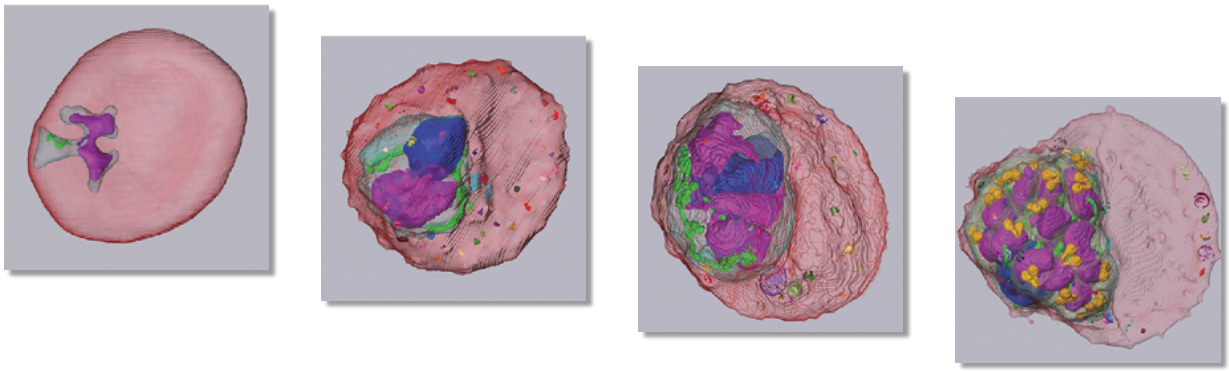
EMPIAR accepts more data types than you may know

- Raw data associated with an EMDB entry
- 2D/3D data from 3D imaging modalities not covered by EMDB (e.g., 3DSEM and SXT →)
- 2D EM data used in integrative/hybrid methods, associated with a structure deposited in the PDB or PDB-Dev archive
- Certain reference and benchmark datasets (to be decided on a case-by-case basis)
- Datasets used for certain community challenges (such as the 2015 Map Validation Challenge)
- Soon: EM/XM parts of CLEM/CLXM experiments

When in doubt, contact empiar-help@ebi.ac.uk

EMBL-EBI

First SBF-SEM datasets archived in 2016

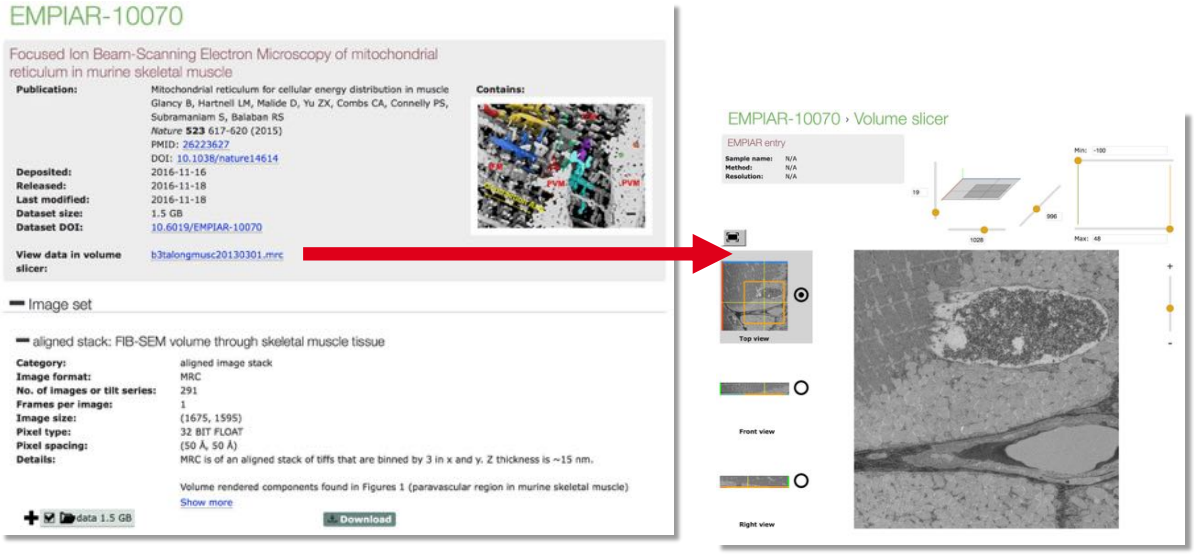


- Four different stages of malaria-parasite-infected red blood cell
- Sakaguchi *et al.*, *J Struct Biol* **193** (2016) 162-171
- EMPIAR entries 10052 to 10055

empiar.org/empiar-10052 (m.m.)

EMBL-EBI

First FIB-SEM dataset archived in 2016



empiar.org/empiar-10070

EMBL-EBI

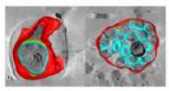
First SXT dataset archived in 2017

EMPIAR-10087

Soft X-ray tomography of Plasmodium falciparum infected human erythrocytes stalled in egress by the inhibitors Compound 2 and E64

Publication: Parasitophorous vacuole poration precedes its rupture and rapid host erythrocyte cytoskeleton collapse in Plasmodium falciparum egress
Hale VL, Watermeyer JM, Hackett F, Vizcay-Barrena G, van Ooij C, Thomas JA, Spink MC, Harkiolaki M, Duke E, Fleck RA, Blackman MJ, Saibil HR
Proc. Natl. Acad. Sci. U.S.A.
DOI: 10.1073/pnas.1619441114

Contains:



Related EMDB entries: EMD-3586, EMD-3587, EMD-3606, EMD-3610
Deposited: 2017-02-28
Released: 2017-03-15
Last modified: 2017-03-15
Dataset size: 280.6 MB
Dataset DOI: 10.6019/EMPIAR-10087

View data in volume slicer: c2_tomo02.mrc, e64_tomo03.mrc

Image set

Soft X-ray tomograms of Plasmodium falciparum infected human erythrocytes stalled in egress

Category: Soft X-ray tomograms
Image format: MRC
No. of images or tilt series: 2
Frames per image: 1
Image size: (variable, variable)
Pixel type: SIGNED BYTE
Pixel spacing: (160 Å, 160 Å)

data 280.6 MB — Download
- C2_tomo02.mrc 131.2 MB
- E64_tomo03.mrc 149.4 MB

EMPIAR-10087 › Volume slicer

EMPIAR entry
Sample name: N/A
Method: N/A
Resolution: N/A

Min: -128 Max: 127

Top view
Front view
Right view

empiar.org/empiar-10087

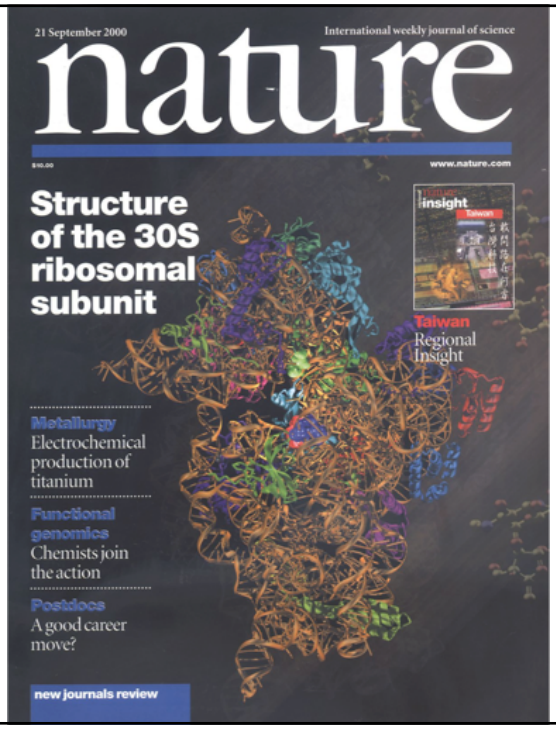
EMBL-EBI

EM in 2018

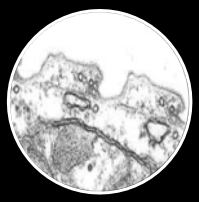
- 1779 new EMDB releases
- 869 (49%) had associated PDB entry
- 74 (4%) had raw datasets deposited in EMPIAR

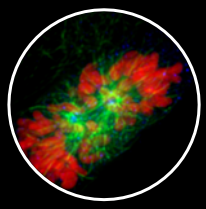
EMBL-EBI

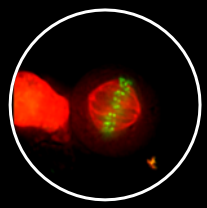
EMBL-EBI
BioImage
Archive



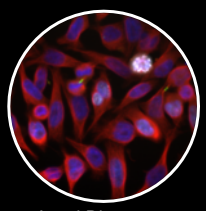
Bioimaging is ubiquitous

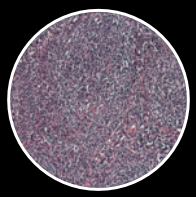

Organelles

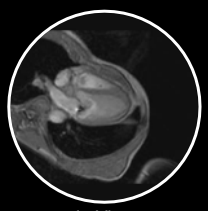

Cells


Dynamics


Physiology


Lead Discovery
Target Validation


Pathology


In Vivo

BioImage Archive (est. July 2019)

- Workshop at EMBL-EBI (January 2017) recommended establishment of a bioimage archive
  - White paper: *Nature Methods* **15**, 849–854 (2018)
- EMBL-EBI has now established BioImage Archive
- EMPIAR already in-house and relatively mature component of BioImage Archive
- Object store as back-end (petabyte scale)

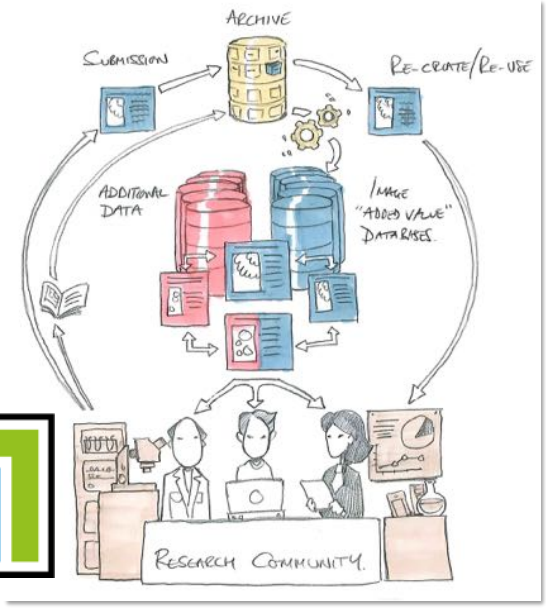
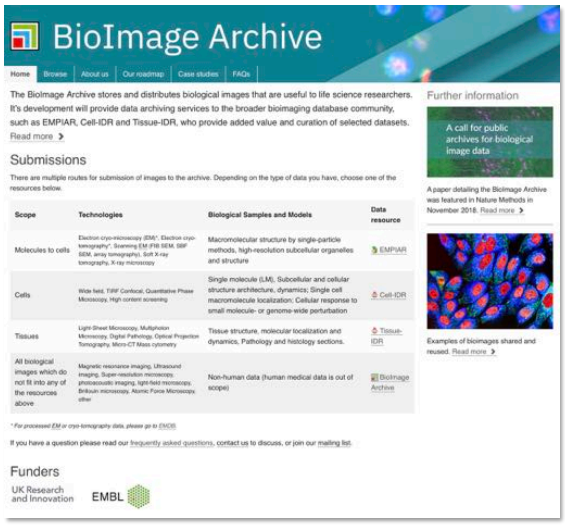
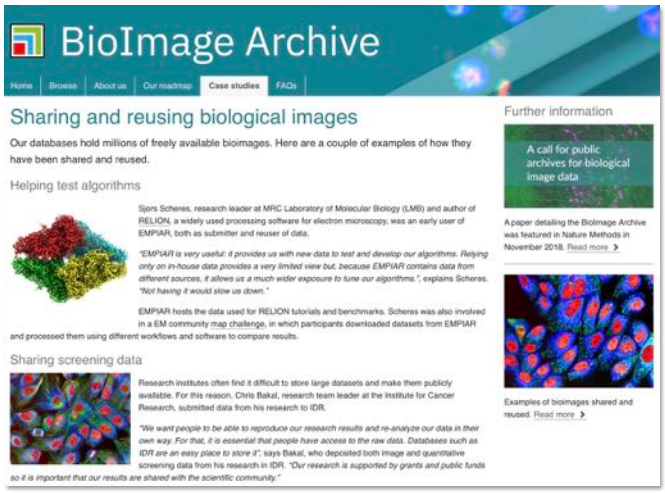


Image credit: Spencer Phillips (EMBL-EBI)

EMBL-EBI

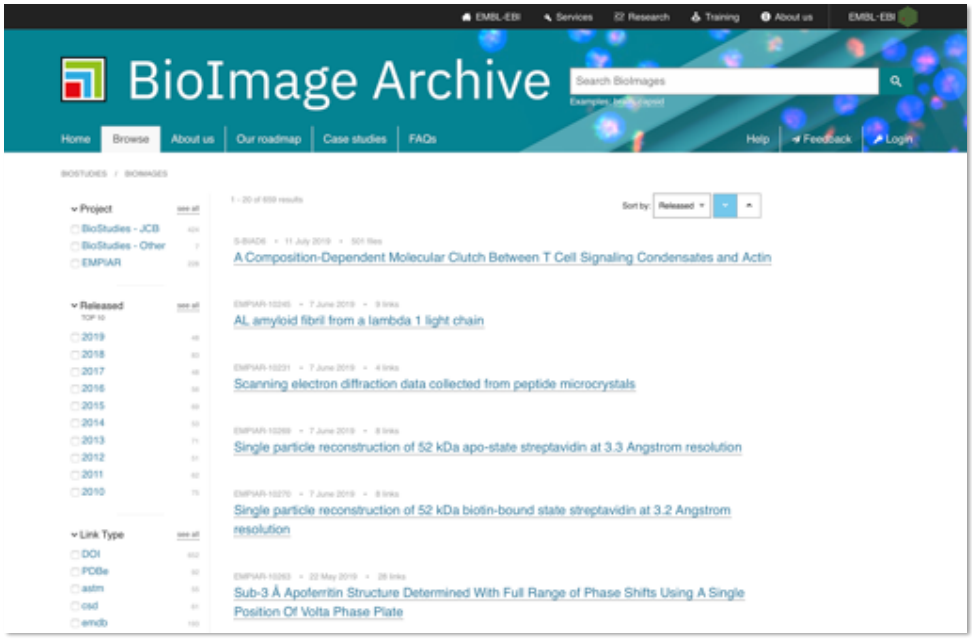
BioImage Archive (est. July 2019)





www.ebi.ac.uk/bioimage-archive

EMBL-EBI

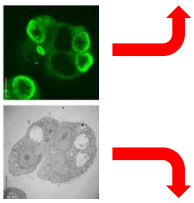


www.ebi.ac.uk/bioimage-archive

EMBL-EBI

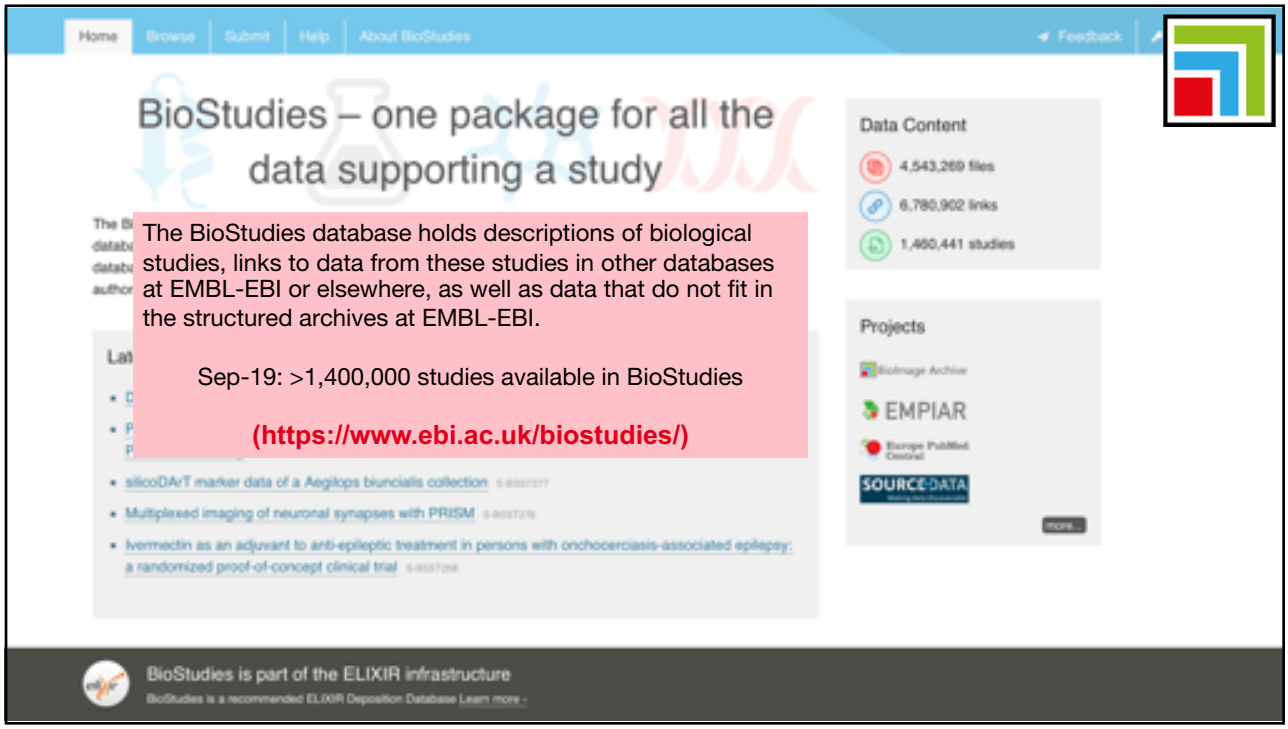
Archiving CLEM/CLXM data in the BioImage Archive

- BioStudies
  - Original 2D or 3D LM data
  - Description of image-registration process
  - ec-CLEM metadata XML file (or list of landmark pairs)
  - If warping was used, warped image(s) as well



- EMPIAR
  - 2D or 3D EM or SXT data
  - Raw data optional (e.g., tilt series)
  - BioStudies accession ID (for corresponding LM and registration data)

EMBL-EBI



The BioStudies database holds descriptions of biological studies, links to data from these studies in other databases at EMBL-EBI or elsewhere, as well as data that do not fit in the structured archives at EMBL-EBI.

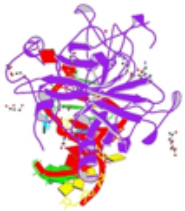
Sep-19: >1,400,000 studies available in BioStudies

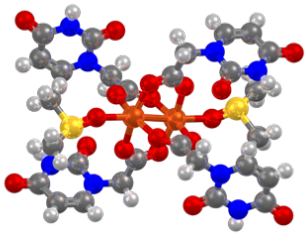
(https://www.ebi.ac.uk/biostudies/)

Other archives you may need or come across

Other experimental 3D model archives

- NDB - Nucleic acid DataBase
  - Operated by RCSB (1995)
  - >10,300 structures (Sep-19)
  - Most structures also in PDB
  - ndbserver.rutgers.edu
- CSD – Cambridge Structural Database
  - Operated by CCDC (1965)
  - Crystal structures of "small molecules"
  - >1,000,000 structures * (Sep-19)
  - www.ccdc.cam.ac.uk



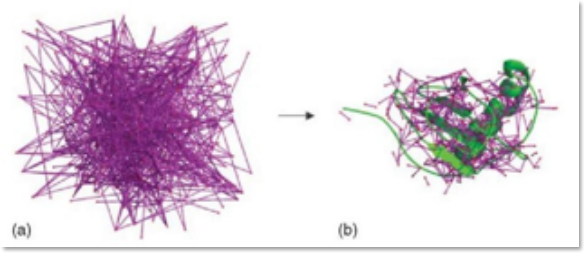


* ccdc.cam.ac.uk/csd-1-million/

EMBL-EBI

BMRB – Biological Magnetic Resonance Data Bank

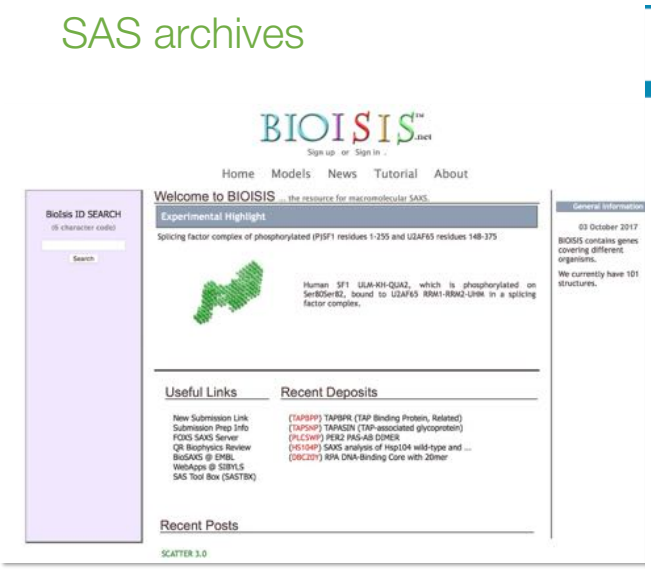
- Repository of experimental NMR data – mainly assigned chemical shifts (>13,000 entries, Sep-19)
- Other types of NMR data:
  - Experimental restraints
  - Relaxation parameters
  - Spectral peak lists
  - Metabolomics by NMR
  - Free-induction decay (FID) – raw spectral data
- Maintained at Univ Madison, Wisconsin (1996)
  - Also site in Japan (PDBj-BMRB)

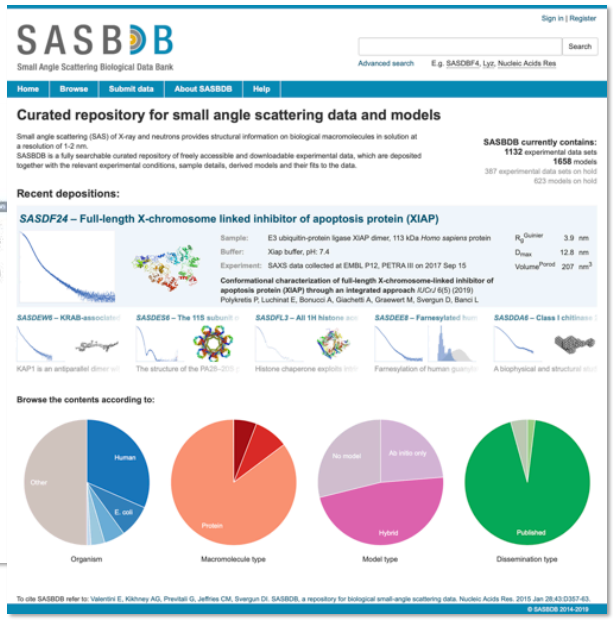


www.bmrb.wisc.edu

EMBL-EBI

SAS archives



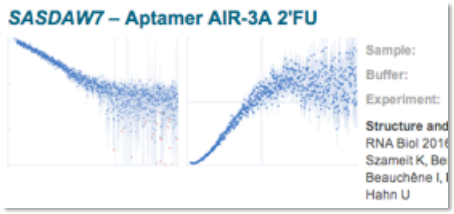


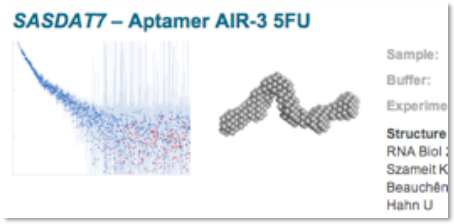
bioisis.net – sasbdb.org

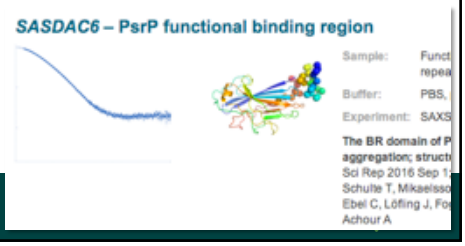
EMBL-EBI

SASBDB

- Maintained by EMBL-Hamburg (est. 2014)
- Archive for biological SAS
  - SAXS, SANS, WAXS data
  - Non-atomistic models (beads)
  - Atomistic models determined solely by SAS methods
  - SAS data that supports PDB structures

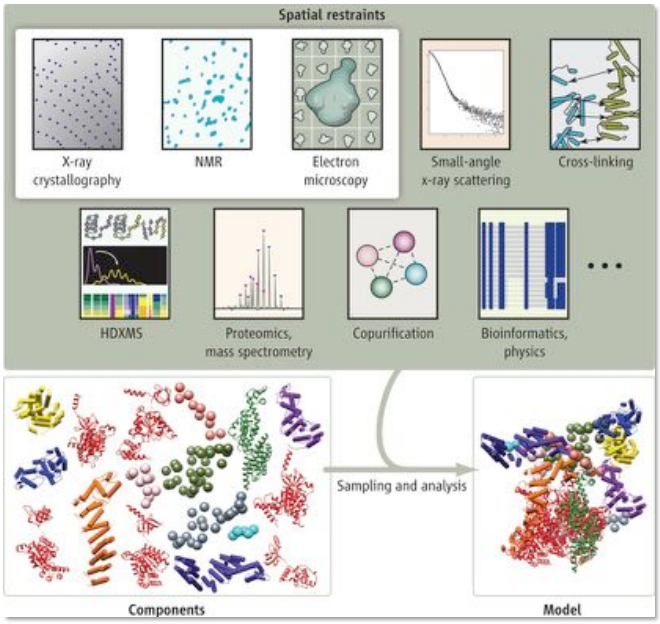






sasbdb.org

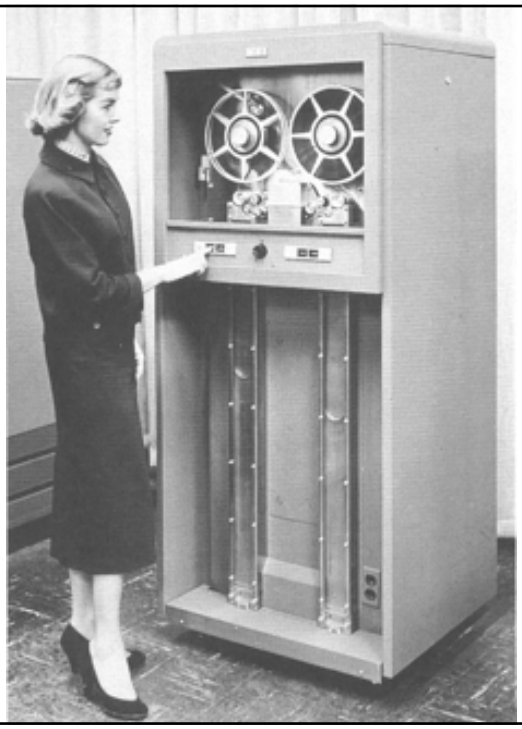
Future archives?

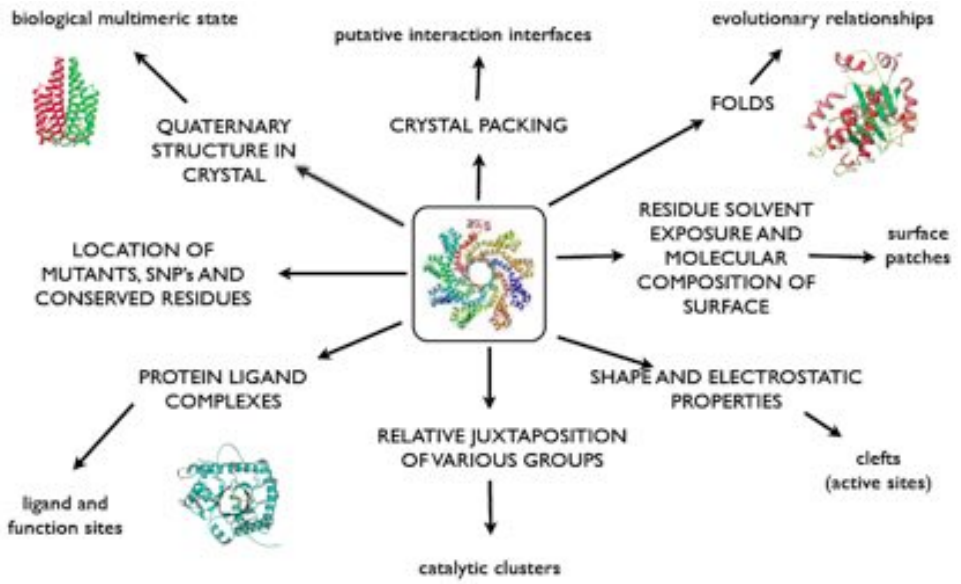


Integrative/Hybrid Modelling

EMBL-EBI

PDB-derived non-archival resources





Clues about function from structure

Resources for all or most PDB entries

- PDBSUM, OCA, Jena, MMDB – Summaries of PDB entries
- PDBREPORT - Validation reports
- EDS - Electron density, validation info
- CATH, SCOP - Classification
- PDB_REDO, RECOORD - Re-refined structures
- And many others
  - See January Database issues of "*Nucleic Acids Research*"

Electron Density Server

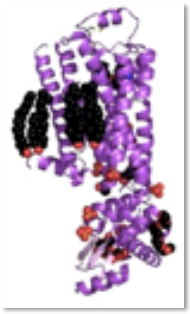
CATH PROTEIN STRUCTURE CLASSIFICATION

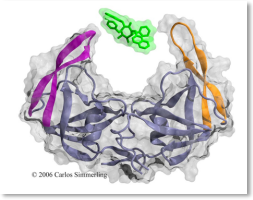
PDBsum

EMBL-EBI

Specialised structure databases

- Enzymes
  - Kinases, HIV proteases, (serine) proteases, carbohydrate-active enzymes, esterases, …
- Antibodies
- Allergenic proteins
- Nuclear receptors
- G-protein-coupled receptors
- Viruses
- Membrane proteins
- RNA structure
- Predicted structures

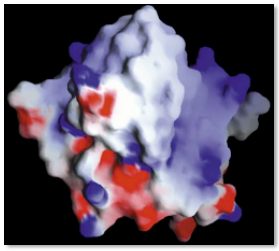


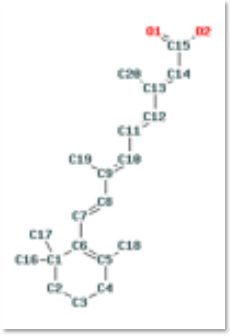


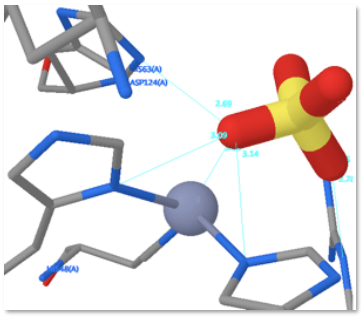
EMBL-EBI

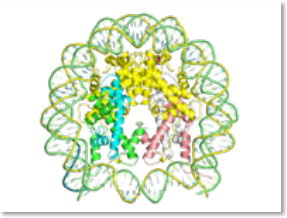
Specialised structure databases

- Ligands
- Domain definitions
- Quaternary structure
- Interactions
- Active sites
- Metal sites
- Surface properties
- Loops
- Torsion angles
- Dynamics
- Folding pathways
- …







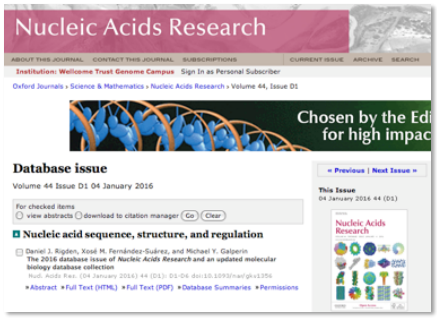


EMBL-EBI

Many PDB-derived databases/resources!

- For 2011-2016, >25% of new databases described in annual *NAR* Database issues used PDB data (119 of 452)

- In total, >200 databases (of 1685 in Jan-2016 *NAR* Database collection) use PDB data, including:
  - 123 structure databases
  - 49 sequence databases
  - 22 metabolic and signalling pathways databases



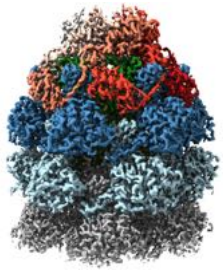
Data from Monica Sekharan, RCSB-PDB (2016)

EMBL-EBI

What's in the pipeline?

What's in the pipeline? A sneak peek!

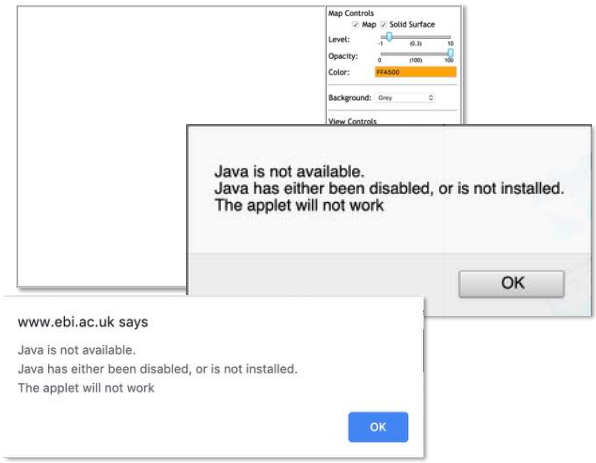
- EMDB
  - New 3D map viewer on entry pages (not requiring Java…)
  - Validation, validation, validation
    - New visual analysis pages for all entries
    - New validation methods (WT-funded UK EM Validation Network)
    - EM-specific components in wwPDB validation reports
  - Website redesign
- EMPIAR
  - More automatic deposition
- Both
  - Integrating structure and 3D bioimaging data across scales

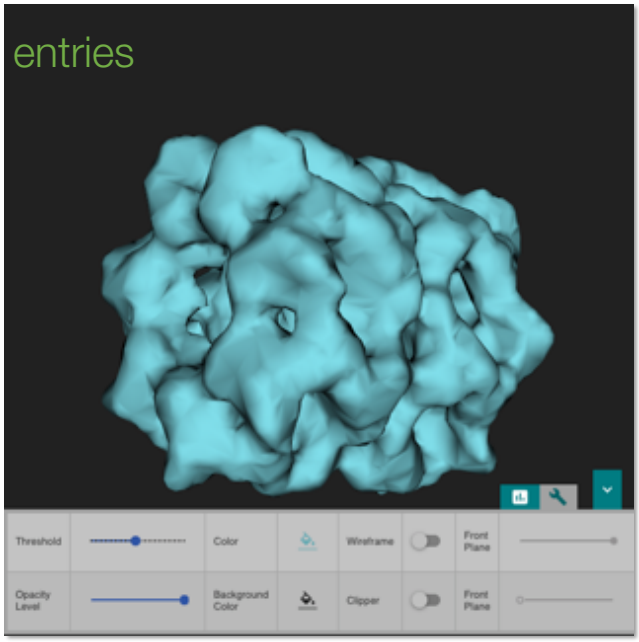


EMBL-EBI

New 3D map viewer for EMDB entries

- Current viewer is a Java applet

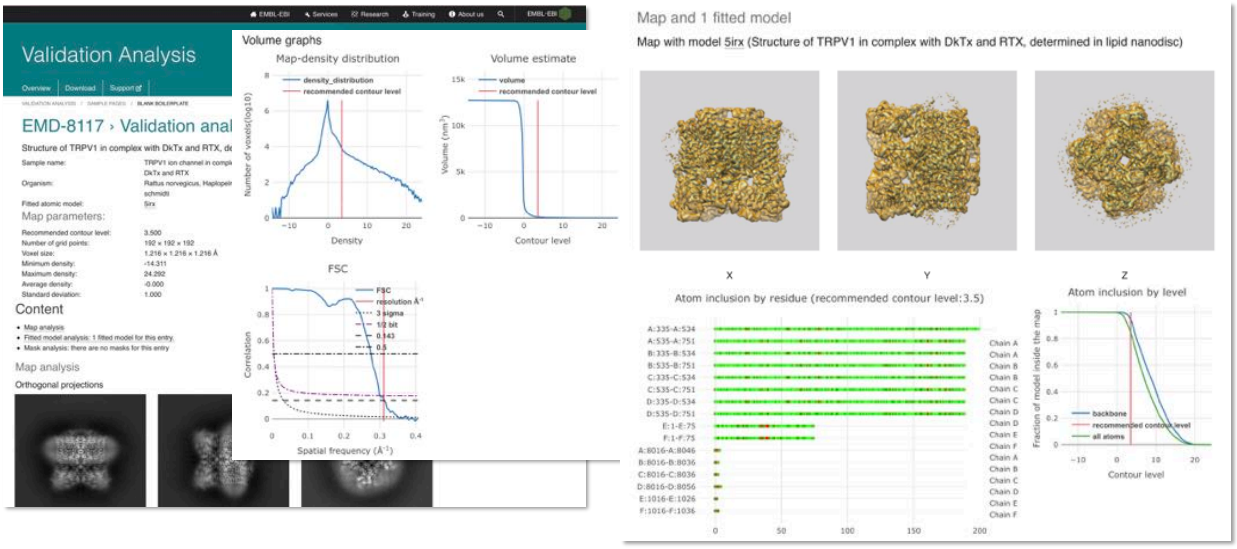




Coming soon! (Aim: Nov-2019)

EMBL-EBI

Refreshed visual analysis pages for EMDB entries



Coming soon! Also to the wwPDB validation pipeline! (Aim: Nov-2019)

EMBL-EBI

Integrating structure and 3D bioimaging data across scales

- We want to link structural information on scales from molecules to cells (and beyond…)
- Molecular structures are typically represented as atomistic models
- Lower resolution data usually does not allow construction of such models
- But if we capture the biologically meaningful objects and put “labels” on them, we can link them to both higher resolution data (e.g., PDB entries) and to other low-resolution objects
- This allows us to study the 3D structure of a system from the molecular detail to the cellular context – and *vice versa*

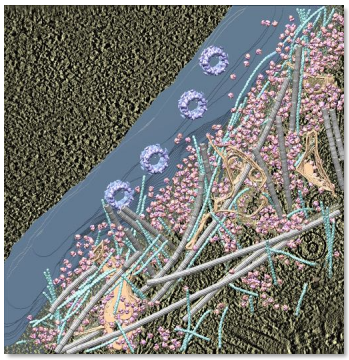
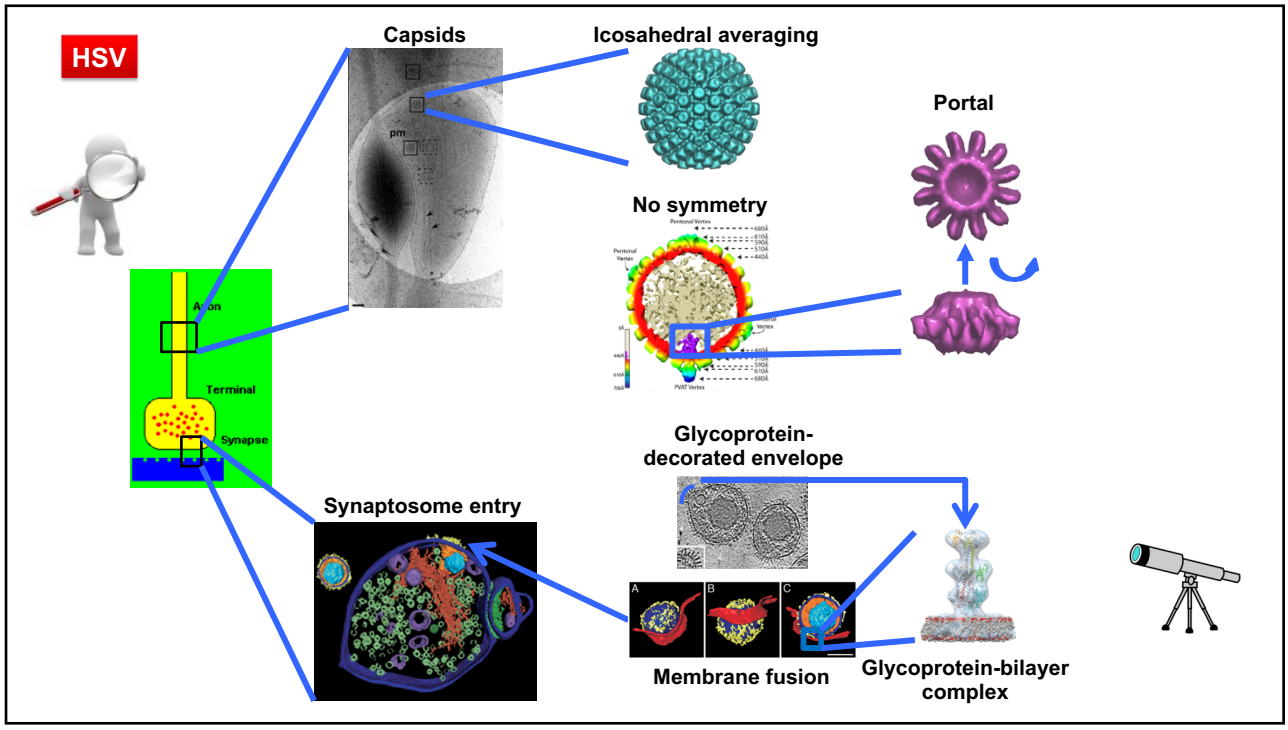


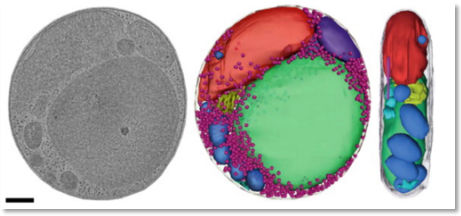
Image credit: Julia Mahamid (EMBL-Heidelberg)

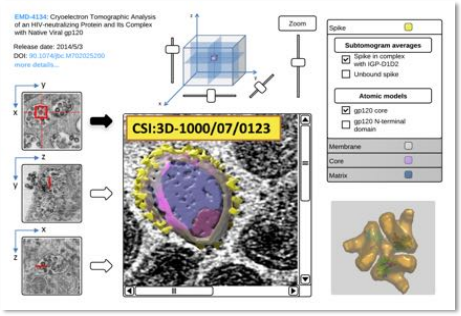
EMBL-EBI



Integrating structure and 3D bioimaging data across scales

- Requirements
  1) Data accessible in public archives
     - PDB, EMDB, EMPIAR, BioImage Archive, …
  2) Semantic segmentation of bioimaging datasets
     - Capture using SAT (Segmentation Annotation Tool) and EMDB-SFF (Segmentation File Format) and toolkit (sfftk)
  3) Linking archive entries through annotation
     - Using established ontologies and accession IDs
  4) Dissemination and visualisation
     - Developing Volume Browser



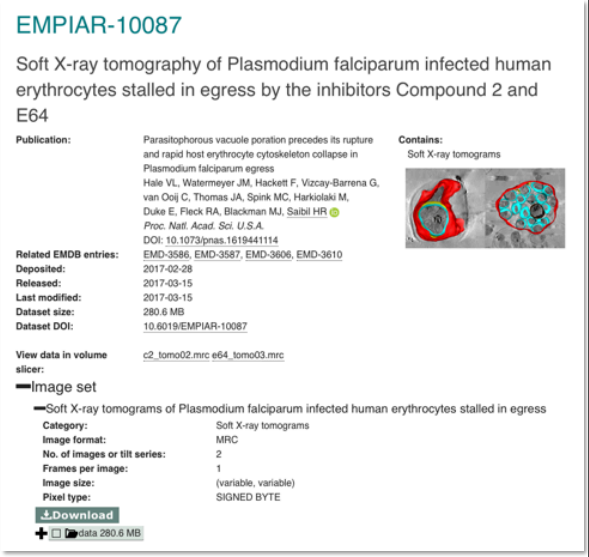


Try the Volume Browser prototype: https://bit.ly/2Wblo6U

EMBL-EBI

Integrating structure and 3D bioimaging data across scales

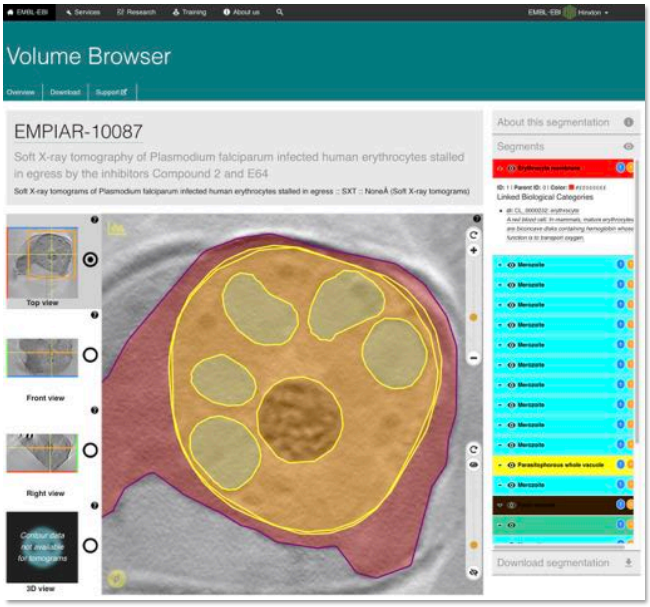
- Example: SXT dataset – EMPIAR-10087
  - Two tomograms with different inhibitors
- Worked with Vickie Hale to get segmentations and annotate them with our SAT (Segmentation Annotation Tool)
- Expose through our Volume Browser
  - Interactive inspection
  - Download



Try the Volume Browser prototype: https://bit.ly/2Wblo6U

EMBL-EBI

Volume Browser

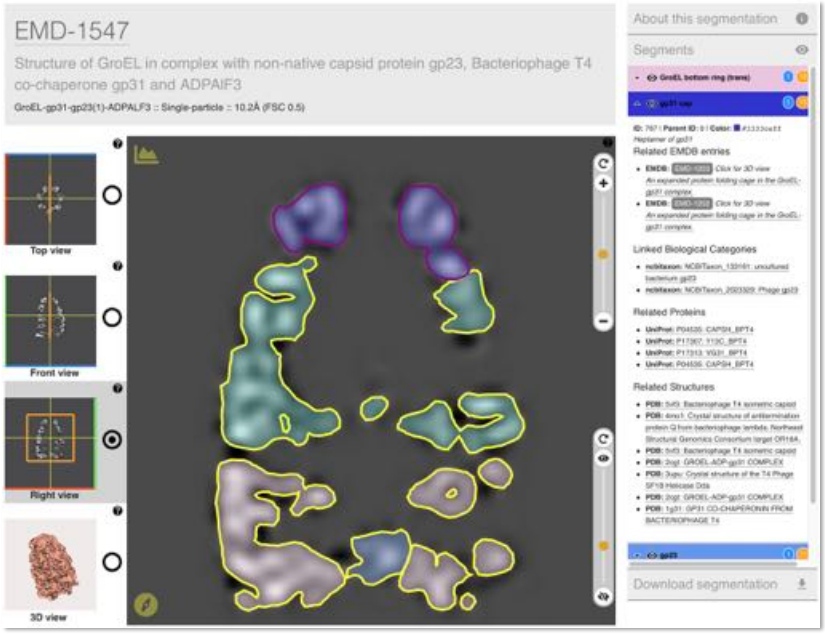


Try the Volume Browser prototype: https://bit.ly/2Wblo6U

EMBL-EBI

Volume Browser

- Can also be used for molecular EMDB maps to segment and identify individual components and link them to other structures of the same component in PDB and EMDB



Try the Volume Browser prototype: https://bit.ly/2Wblo6U

EMBL-EBI

Integrating structure and 3D bioimaging data across scales

- Everything depends on having high-quality annotated segmentations
- We plan to acquire these by
  - Working with selected depositors to segment and annotate some of their datasets to showcase what this makes possible
  - Encourage depositors to segment their data and use SAT to annotate it, and then deposit it in EMPIAR and EMDB
  - Organise “segmentathons” where specialists spend a few days doing semantic segmentation of existing EMDB and EMPIAR datasets
  - Considering crowd-sourcing, *e.g.* using Zooniverse
    - Used successfully, *e.g.* by Lucy Collinson *et al.* and by Michele Darrow *et al.*

Try the Volume Browser prototype: https://bit.ly/2Wblo6U

EMBL-EBI

Thank you!

https://emdb-empiar.org/ https://empiar.org/

https://pdbe.org/ https://pdbekb.org/

https://twitter.com/EMDB_EMPIAR

https://twitter.com/PDBeurope

EMBL-EBI