EMDB data deposition

Image processing for cryo-electron microscopy, Birkbeck EMBO practical cryo-EM course
September 15, 2017

Ardan Patwardhan

EMBL-EBI

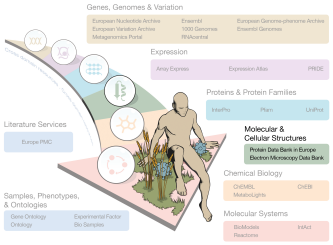
Plan

- Introduction to EMDB and EMPIAR
- Trends in the field
- Overview of web resources for searching visualisation and validation
- Planned initatives and developments
- How to deposit to EMDB and EMPIAR

EMBL-EBI

Molecular and Cellular Structure Cluster

- Maintain and manage archives
  - PDB for atomic coordinate models
  - EMDB for 3DEM reconstructions
  - EMPIAR for 3DEM raw data
- Develop and maintain web-services – searching, visualisation and validation
- Facilitate community-wide initiatives
- Key themes – integration (with other bioinformatics resources and imaging scales), validation, data-mining



EMBL-EBI

Cellular Structure and 3D Bioimaging Team


Sanja Abbott


Andrii Iudin


Paul Korir


Sriram Somasundharam

- PDB/EMDB annotation staff (Alice Clark…)
- Former members: E Sanz Garcia, C Lugo, J Salavert Torres (UPV), I Lagerstedt (Lilly)
- Former students: M Holmdahl, V Lysenkov, I Solanes Valero
- 1 positions available:
  - validation

EMBL-EBI

Collaborations & acknowledgements

- Birkbeck College (Elena Orlova, Maya Topf, Helen Saibil)
- Baylor (Wah Chiu)
- Chinese Academy of Sciences (Fei Sun)
- CNB Madrid (Jose Maria Carazo)
- Dundee (Jason Swedlow)
- EMBL-EBI (Helen Parkinson, Alvis Brazma, Ugis Sarkans)
- Francis Crick (Lucy Collinson, Raffa Carzaniga, Peter Rosenthal)
- Osaka University (Genji Kurisu)
- MRC-LMB (John Briggs, Paula da Fonseca, Wanda Kukulski, Garib Murshudov)
- RCSB (Cathy Lawson)
- STFC (Martyn Winn)
- University of Manchester (Alan Roseman)
- And many more...!

EMBL-EBI

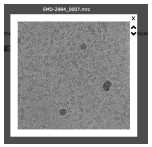
Websites

- emdb-empiar.org
- empiar.org
- pdbe.org
- emdatabank.org

EMBL-EBI

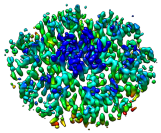
Data: What data is found where...

2D Raw image



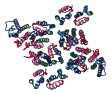
→ EMPIAR

3D Volume



→ EMDB

Fitted model



→ PDB

EMBL-EBI

Comparison of archives

	PDB	EMDB	EMPIAR
Inception	1971	2002	2014
Number of entries	133589 (2297 EM)	5163 (1068 in 2016)	84
Archive size	1 GB	489 GB	48 TB
Mandatory deposition	Coordinates (1989) Structure factors (2008)	Single particle and sub-tomogram averaged structures (2012) Representative tomogram strongly recommended	-

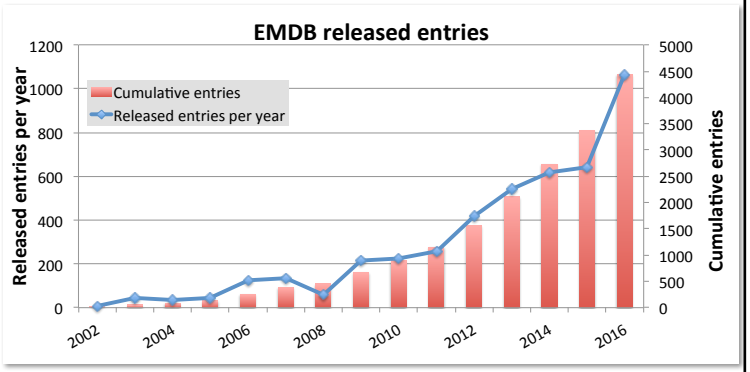
EMBL-EBI

EMDB Deposition Discussion at 3DEM GRC 2017

- Following changes were voted on with overwhelming majority (few abstentions and none against)
  - For all tomography-based studies, the deposition of representative tomographic reconstructions should be made mandatory. All tomographic reconstructions that are depicted in published figures (as orthoslices, renderings or similar), should be deposited
  - The deposition of un-modified half-maps should be mandatory in cases where half-maps have been used to assess resolution

EMBL-EBI

EMDB trends: released entries

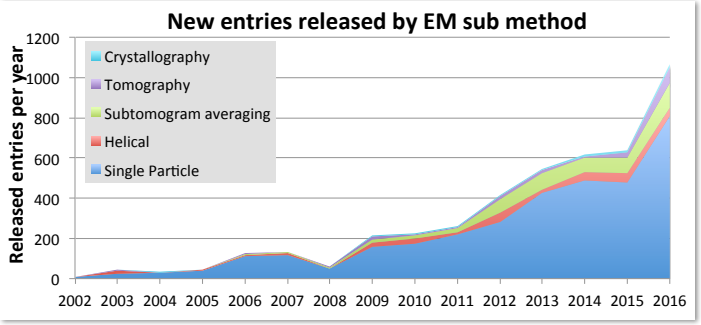


emdb-empiar.org/emstats

EMBL-EBI

EMDB trends: EM sub-methods

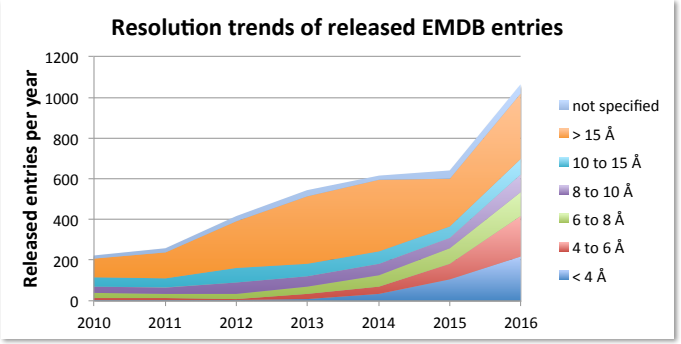
- ~20% from tomography in 2016 - will growth in tomography be sustained?



EMBL-EBI

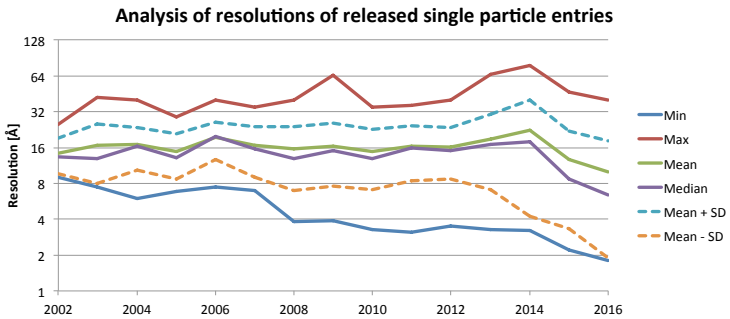
EMDB trends: resolution

- Strong growth of better than 4Å structures



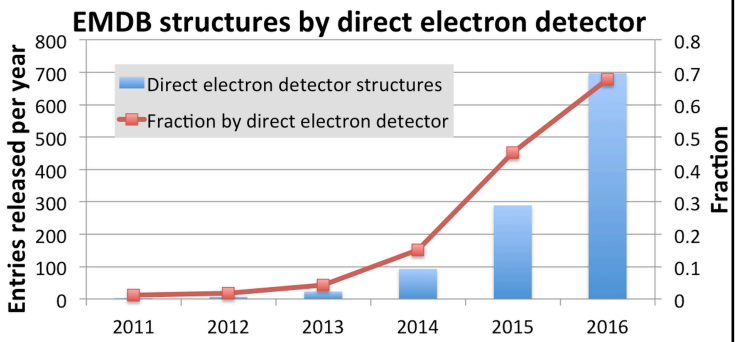
EMBL-EBI

EMDB trends: resolution of single-particle entries



EMBL-EBI

EMDB trends: direct electron detectors



EMBL-EBI

EMDB trends: geographic spread

- Analysis of author affiliations of EMDB-associated publications
- USA, Germany, UK continue to lead
- Strong growth in China
- Many more countries up and coming
- Patwardhan A., *Trends in the Electron Microscopy Data Bank (EMDB)*, Acta D **73**(6), 503-508

Country	2010	2011	2012	2013	2014	2015	2016	Sum
USA	55	50	86	116	139	135	162	743
Germany	19	25	20	22	42	53	53	234
UK	14	10	27	16	42	51	62	222
China	3	4	3	10	13	29	39	101
France	8	4	6	8	16	16	23	81
Japan	6	7	4	6	11	11	20	65
Spain	4	7	11	7	8	7	10	54
Canada	1	5	3	2	10	14	17	52
Switzerland	2	2	8	8	6	15	11	52
The Netherlands	0	1	1	3	9	15	16	45

EMBL-EBI

EMDB trends: <4A structures by organization

- Stats on author affiliations of EMDB related publications with high res structures (<4A)
- Organisations with 5 or more sructures in past 3 years shown

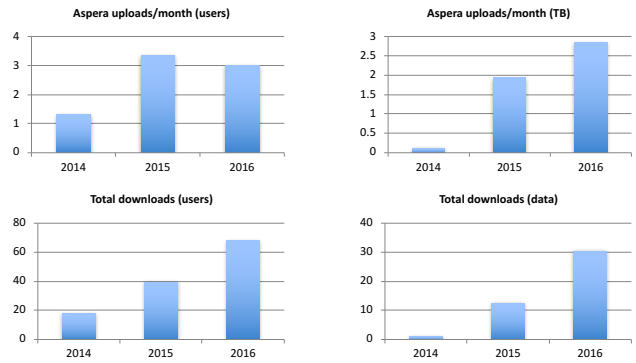
Institution	2014	2015	2016	Sum
MRC Laboratory of Molecular Biology	8	14	15	37
Tsinghua University	-	7	10	17
Howard Hughes Medical Institute	1	6	5	12
University of California San Francisco	-	5	4	9
Chinese Academy of Sciences	1	3	4	8
Harvard University/Medical School	2	4	2	8
National Cancer Institute	1	2	5	8
University of California Los Angeles	-	7	1	8
University of Munich	2	3	3	8
Purdue University	1	1	5	7
University of Texas	3	3	1	7
Columbia University	-	2	4	6
ETH	1	3	2	6
University of California Berkeley	1	1	4	6
University of Virginia	2	3	1	6
Max Planck Institute for Biochemistry	-	-	5	5
Max Planck Institute for Biophysical Chemistry	-	1	4	5
The Rockefeller University	-	-	5	5
University of Oxford	-	3	2	5
University of Washington	-	3	2	5

EMPIAR Raw image data archive for EM

- Started in 2014 – strong community support
- Control growth by requiring deposition of 3D to EMDB
- Usage: validation, development, testing, teaching and…
  - Safe storage of your data!
  - Was source for data in EMDatabank Map Validation Challenge
- Multi-frame micrographs, averaged micrographs, particle-stacks, tilt series
- Uses Aspera, Globus, ftp, http for data transfers

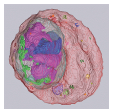
EMPIAR metrics

- Number of entries: 84 (~50TB; average size ~ 600GB)
- 12 datasets greater than 1 TB; one 10TB+ dataset
- Transfer speed: uploads 1-2 TB/24h (Europe, US, Australia)
- "empiar" cited 50+ times in full-text open-access papers
- Nature Methods publication (ludin et al., 2016)

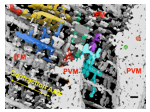


Beyond molecular cryo-EM

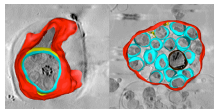
- We have already expanded EMPIAR beyond molecular cryo-EM
  - EMPIAR-10052/3/4/5: SBF-SEM
  - EMPIAR-10070: FIB-SEM
  - EMPIAR-10087: soft X-ray tomography
  - Discussion with integrative hybrid modelling community on deposition of related EM data
- EMBL-EBI Bioimaging Workhop (Jan 2017)
- Expert workshop on data models for EMPIAR for cellular imaging in December 2017



EMPIAR-10054

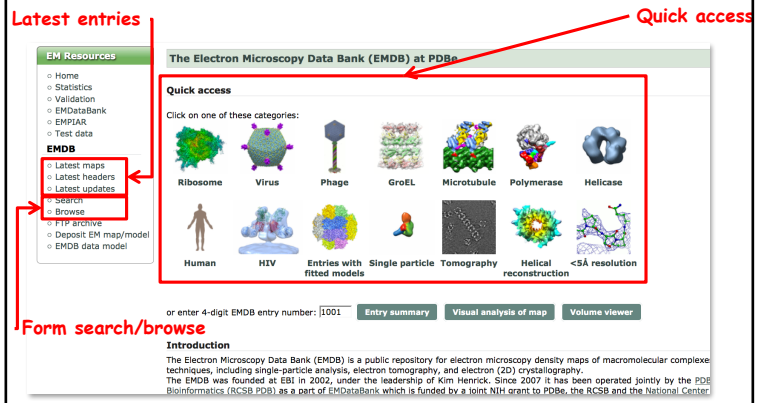


EMPIAR-10070



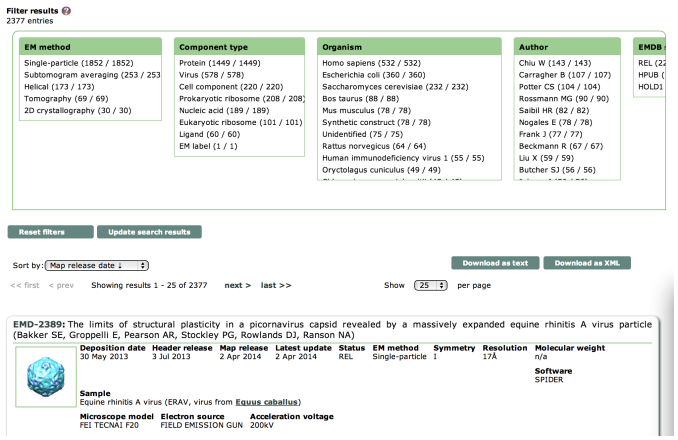
EMPIAR-10087

Searching EMDB



emdb-empiar.org EMBL-EBI

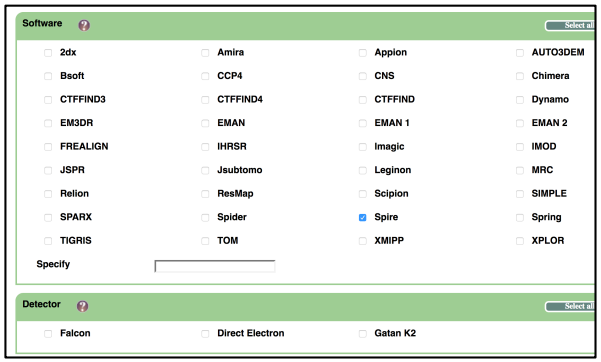
Searching EMDB – browse the archive



emdb-empiar.org/embrowse EMBL-EBI

Form based search

- Enumerations for software and microscopy recently updated
- Search by detector model



http://emdb-empiar.org/emsearch EMBL-EBI

EMStats – EMDB statistics

- Dynamically generated charts based on the current state of the archive

Charts

General
Map releases
Map size statistics for released maps
Distribution of maps released based on EM technique used
Trends and distribution of microscope usage for released maps
Trends and distribution of software package usage for released maps
Molecular weight statistics of single-particle released maps
Sample taxonomy statistics for released maps

Resolution
Trend of highest resolutions achieved annually for released maps
Number of released maps achieving given resolution levels
Single-particle released maps - resolution trends
Single-particle released maps - resolution versus acceleration voltage and source type
Single-particle released maps - resolution versus microscope type and software package used
Single-particle released maps - resolution versus number of projections
Tomography - resolution trends for released maps

Publications
Trend for publications associated with released maps
Journal distribution of articles associated with released maps
Publication trends and distribution for different microscope types
Publication trends and distribution for different software packages

FTP downloads
Monthly FTP downloads
Top 10 FTP downloads

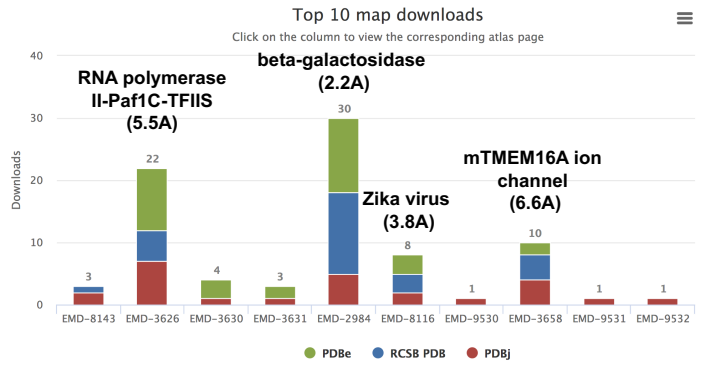
EMPIAR
Entry releases
Entry size statistics for released entries
EMPIAR transfer statistics

emdb-empiar.org/emstats

EMBL-EBI

EMStats – map downloads

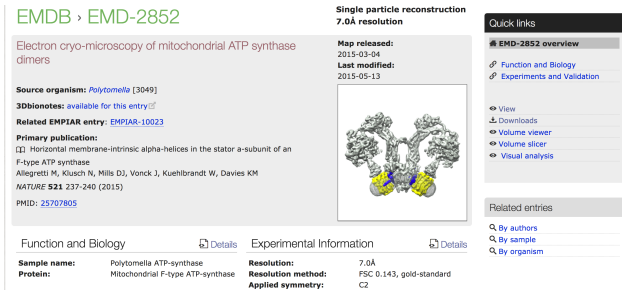
- Downloads stats for the past two years



EMBL-EBI

EMDB entry pages

- Displays summary info about entry
- Download header file (extensive meta-data), map, FSC etc
- Visual analysis pages and visualisation options

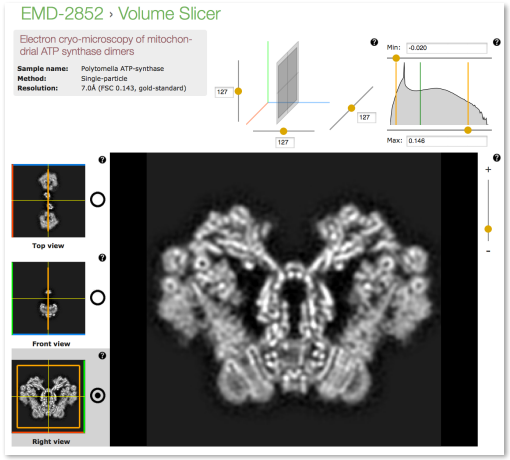


emdb-empiar.org/emd-2852

EMBL-EBI

Volume slicer

- Orthogonal slices in three directions
- Interactive 3D navigation

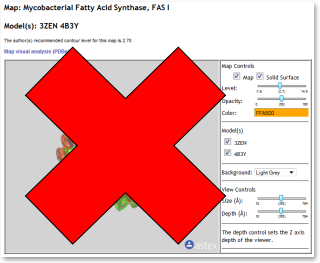


emdb-empiar.org/emd-2852/browse

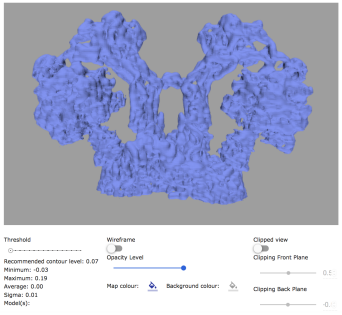
EMBL-EBI

3D interactive viewer

- Web GL based 3D viewer on the way…



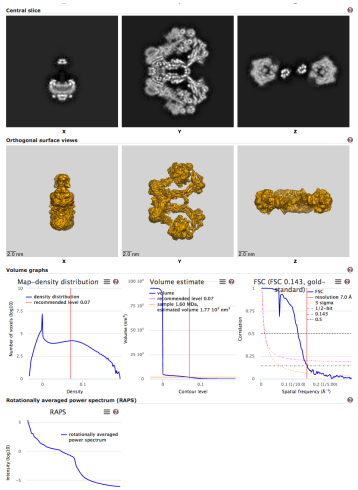
EMD-2852 › Volume viewer



EMBL-EBI

Visual analysis

- Basic sanity checking info
- FSC curves shown if available



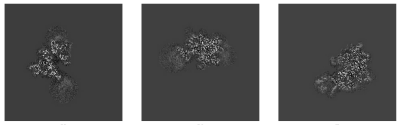
emdb-empiar.org/emd-2852/analysis

EMBL-EBI

Visual analysis: images

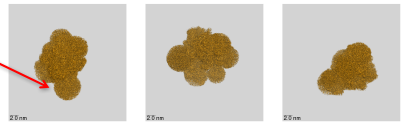
- Surface rendering shows that contour level is too low

5.3 Central slices



Contour level defines the mask!!

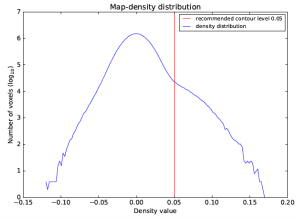
5.4 Orthogonal surface views



Visual analysis: map density distribution

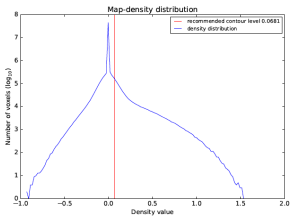
- Masking comes up as sharp spike ~0
- Check for artefacts

5.5 Map-density distribution



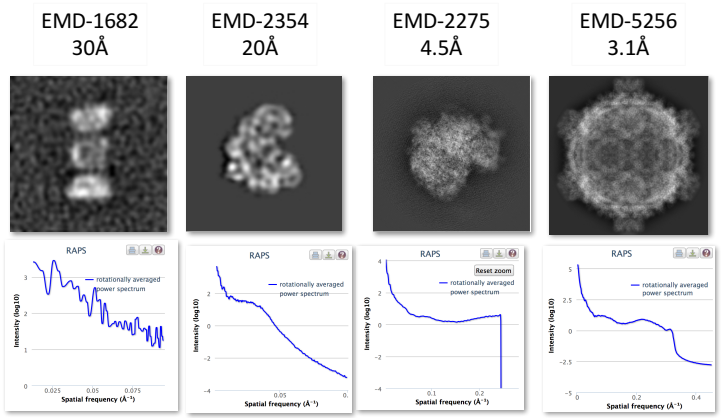
Unmasked map
beta galactosidase
EMD-2984

5.5 Map-density distribution



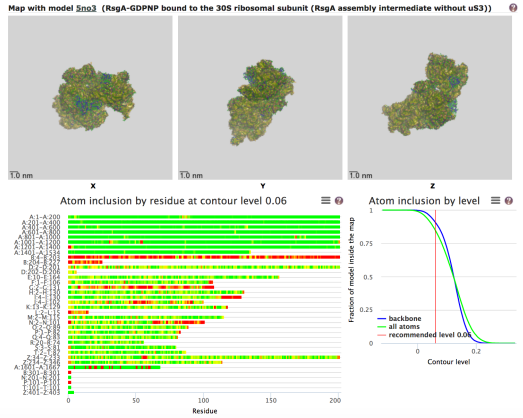
Masked map
ribosome
EMD-2566

Analysis: Rotationally averaged power spectra (RAPS)



Visual analysis: map-model fit

- How well does model fit into map density?



RsgA-GDPNP bound to the 30S ribosomal subunit (RsgA assembly intermediate without uS3)

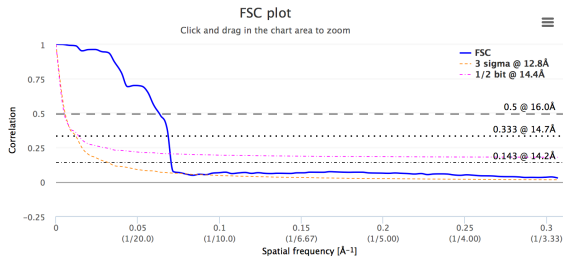
EMD-3662/5no3

Stand-alone validation servers

- Convenience
- Promote use and reporting
- Work with software developers to develop standards for capturing validation data

Fourier shell correlation (FSC) server

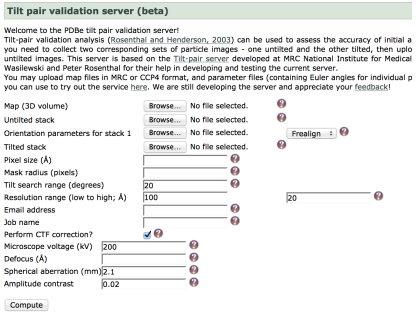
- User uploads two maps -> server calculates FSC curve
- FSC curve can be deposited to EMDB
- Automated calculation of thresholds
- Bsoft, Relion and EMAN2 generate FSC curves that can be deposited



emdb-empiar.org/fsc

Tilt-pair validation server

- Developed with Rosenthal lab
- Deposition of tilt-pair data to EMDB still to be developed
- Requested features – read Imagic files, Relion angles

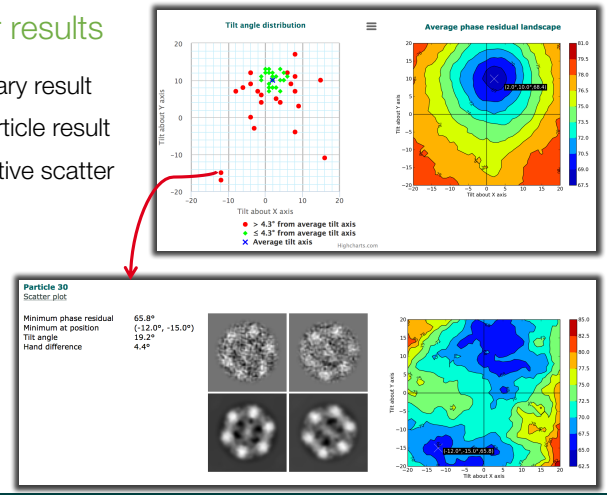


pdbe.org/tiltpair

EMBL-EBI

Tilt-pair results

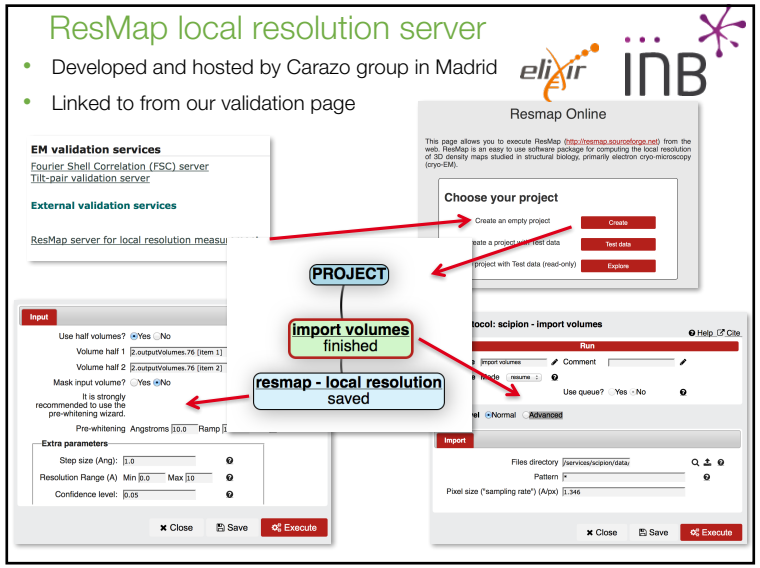
- Summary result
- Per particle result
- Interactive scatter plot



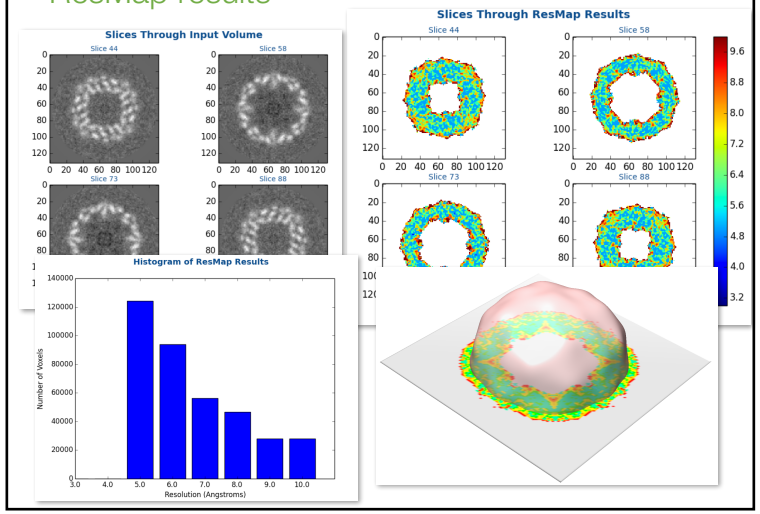
EMBL-EBI

ResMap local resolution server

- Developed and hosted by Carazo group in Madrid
- Linked to from our validation page



ResMap results



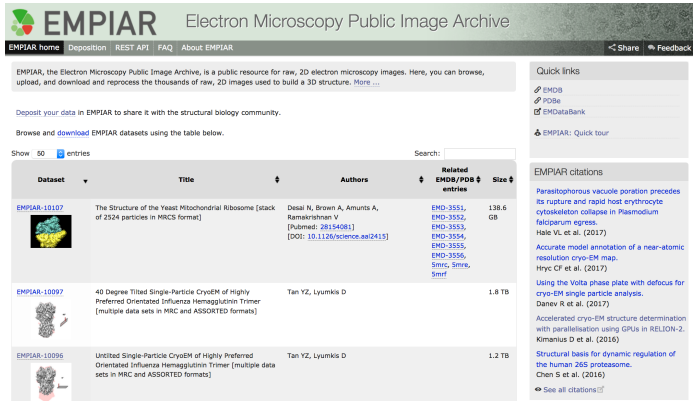
EMDB API

- Programmatically access EMDB meta-data
- Search API – will be discussed later in tutorial
- e.g., ebi.ac.uk/pdbe/api/emdb/entry/all/EMD-6007


```
    }
  ],
  "imaging": [
    {
      "electronSource": "FIELD EMISSION GUN",
      "imagingMode": "BRIGHT FIELD",
      "illuminationMode": "SPOT SCAN",
      "specimenHolder": "Eucentric",
      "microscope": "FEI/PHILIPS CM200FEG",
      "specimenHolderModel": "GATAN LIQUID NITROGEN",
      "acceleratingVoltage": {
        "value": 160.0,
        "units": "kV"
      }
    }
  ],
  "imageAcquisition": [
    {
```

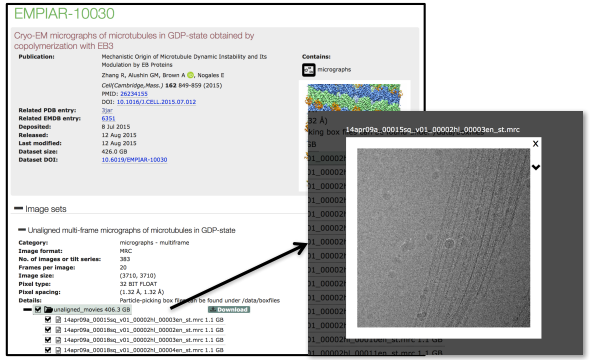

EMPIAR website - empiar.org

- Search, browse EMPIAR entries and download and deposit data



EMPIAR entry pages

- Browse more info on entry
- View thumbnail images before downloading
- Select images to download



empiar.org/empiar-10030

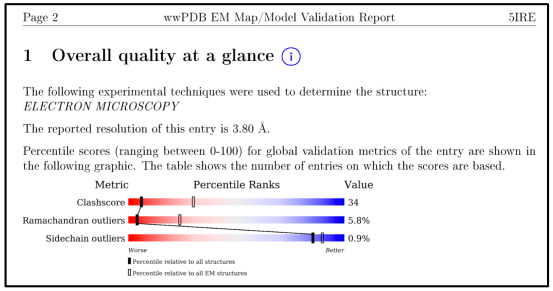
EMBL-EBI

Planned developments

EMBL-EBI

EM validation reports

- In production: mainly info relating to EM models
- Validation report for PDB EM models calculated and available from ftp sites



Page 2 wwPDB EM Map/Model Validation Report 5IRE

1 Overall quality at a glance

The following experimental techniques were used to determine the structure:
ELECTRON MICROSCOPY

The reported resolution of this entry is 3.80 Å.

Percentile scores (ranging between 0-100) for global validation metrics of the entry are shown in the following graphic. The table shows the number of entries on which the scores are based.

Metric	Percentile Ranks	Value
Clashscore		34
Ramachandran outliers		5.8%
Sidechain outliers		0.9%

Worse — Better

Percentile relative to all structures

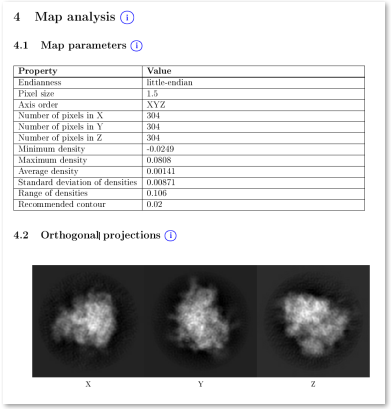
Percentile relative to all EM structures

pdbe.org/5ire

EMBL-EBI

EM validation reports – map validation

- Based on EMDB Visual analysis pages (available for released entries)
- Limited by data available – mostly sanity checking
- In testing



4 Map analysis

4.1 Map parameters

Property	Value
Endianness	little-endian
Pixel size	1.5
Axis order	XYZ
Number of pixels in X	304
Number of pixels in Y	304
Number of pixels in Z	304
Minimum density	-0.0249
Maximum density	0.0808
Average density	0.00141
Standard deviation of densities	0.00871
Range of densities	0.106
Recommended contour	0.02

4.2 Orthogonal projections

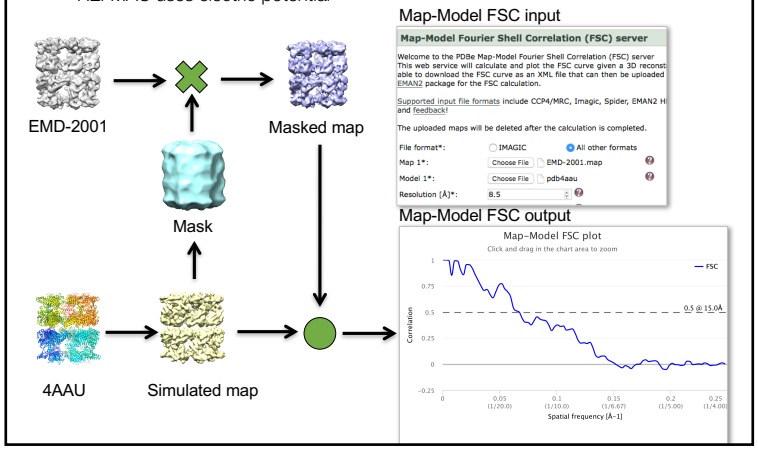
X Y Z

emd-5801

EMBL-EBI

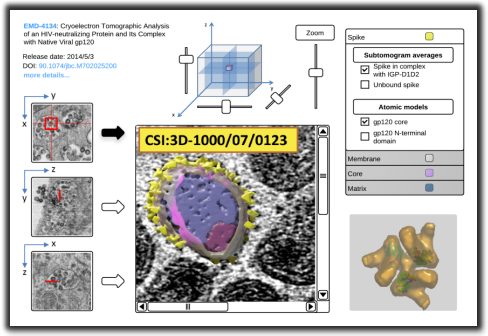
Map/Model FSC server using EMAN2/REFMAC

- EMAN2 generates map from model using Gaussian blurring function
- REFMAC uses electric potential

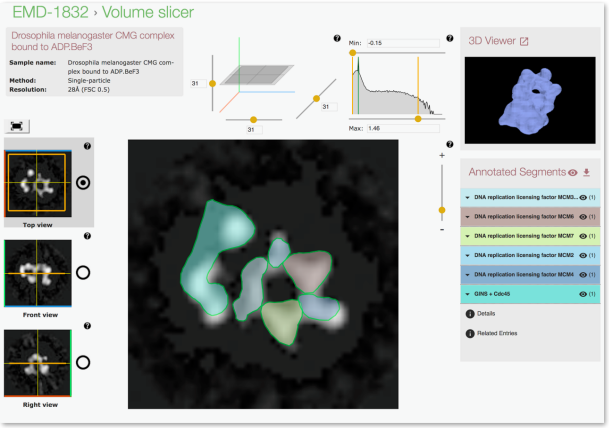


Volume browser – the idea

- Integrated visualisation of structural data
- Spanning scales from cells to molecules



Volume browser – the reality (almost there!)

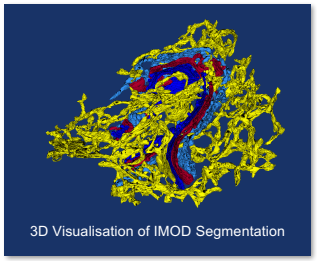


EMDB-SFF (Segmentation File Format)

- Expert workshop on "3D segmentations and transformations - building bridges between cellular ar molecular structural biology", Dec 2015
  - Manuscript submitted to eLife
- Working group to advise on development
- EMDB-SFF supports structured biological annotation linking segments to ontologies and other bioinformatics resources
- Implementations in XML, HDF5 and JSON
- Also capable of supporting transforms between subtomogram averages and tomograms

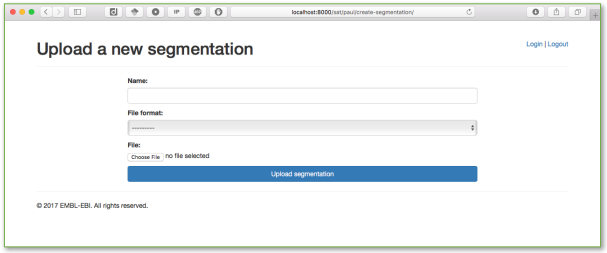
EMDB-SFFTK (toolkit)

- Available from CCP-EM SVN: http://bit.ly/2oQnvxd
- Contains code for manipulating EMDB-SFF
  - Translators support external formats
    - IMOD
    - Segger
    - Amira (AmiraMesh and Amira HxSurface)
    - STL
    - EMDB MAP masks
  - Annotation utilities
    - Search ontologies (OLS hosted at EBI)
    - View, add, edit, delete annotations
- Output formats: XML, HDF5, JSON
- Create and load ROIs into OMERO

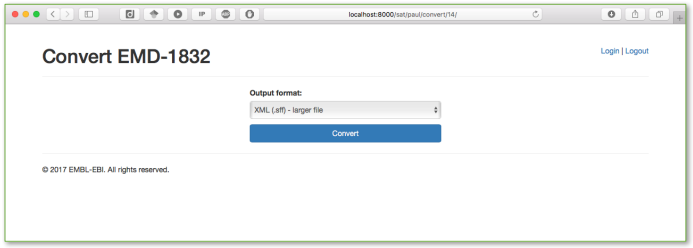


Segmentation annotation tool (SAT)

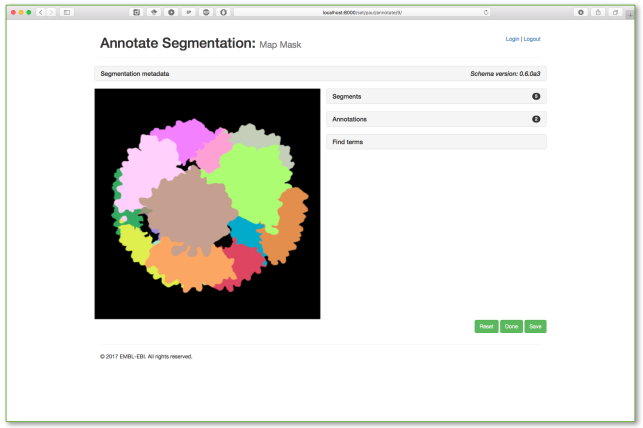
- Web based service
- Desktop tool



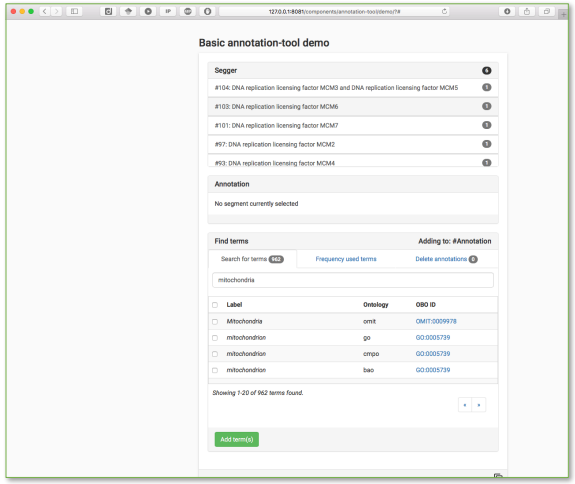
SAT – conversion to EMDB-SFF



SAT



SAT – annotation via OLS (Ontology Lookup Service)



Future outlook

- Crowd sourced annotation of segmentation
- Archiving for related imaging modalities including – 3DSEM, SXT and CLEM
- Data-harvesting and submission pipelines to EMDB and EMPIAR
- Visualisation of and standards for correlative imaging and for relating sub-tomogram averages to tomograms
- Validation
- Automated segmentation – deep learning
- Fast archive-wide sub-structure volumetric (or shape-based) searches

References

1. Patwardhan A., *et al.* "Data management challenges in three-dimensional EM". *Nat Struct Mol Biol*, **19**, 1203-1207 (2012).

2. Gutmanas A., Oldfield T., Patwardhan A., Sen S., Velankar S. and Kleywegt G. "The role of structural bioinformatics resources in the era of integrative structural biology". *Acta Cryst*, **D69**, 710-721 (2013).

3. Lagerstedt I., Moore W.J., Patwardhan A., Sanz-Garcia E., Best C., Swedlow J.R. and Kleywegt G.J. "Web-based visualisation and analysis of 3D electron-microscopy data from EMDB and PDB". *J Struct Biol*, **184**, 173-181 (2013).

4. Patwardhan A., *et al.* "A 3D cellular context for the macromolecular world". *Nat Struct Mol Biol*, **21**, 841-845 (2014).

5. Wood C., Burnley T., Patwardhan A., Scheres S., Topf M., Roseman A. and Winn M. "Collaborative computational project for electron cryo-microscopy". *Acta Crystallogr D Biol Crystallogr*, **71**, 123-126 (2015).

6. Iudin A., Korir P.K., Salavert-Torres J., Kleywegt G.J. and Patwardhan A. "EMPIAR: A public archive for raw electron microscopy image data". *Nature Methods*, **13**, 387-388 (2016).

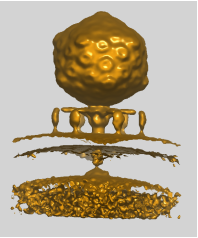
7. Patwardhan A. and Lawson C.L. "Databases and Archiving for CryoEM". *Methods Enzymol*, **579**, 393-412 (2016).

8. Salavert-Torres J., Iudin A., Lagerstedt I., Sanz-Garcia E., Kleywegt G.J. and Patwardhan A. "Web-based volume slicer for 3D electron-microscopy data from EMDB". *J Struct Biol*, **194**, 164-170 (2016).

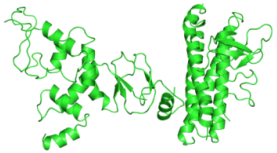
9. Patwardhan A. "Trends in the Electron Microscopy Data Bank (EMDB)". *Acta Crystallographica Section D*, **73**, (2017).

How to deposit?

EMDBspeak 1



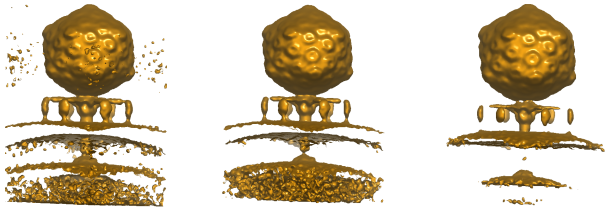
Map: 3D EM reconstruction (EM volume)



Model: description of atomic coordinates

EMDBspeak 2

- **Contour level**: Density level at which the surface of molecule is defined
- Important for visual interpretation
- Important for many calculations, e.g., weight, volume, atom inclusion



0.03 0.25 1.0

EMDBspeak 3

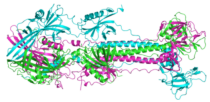
- **EMDB header file**: XML file with metadata describing how map was obtained


```
- <imaging>
    <astigmatism>objective lens astigmatism was corrected at 120,000 times magnification</astigmatism>
    <electronSource>FIELD EMISSION GUN</electronSource>
    <electronDose units="e/A**2">16</electronDose>
    <imagingMode>BRIGHT FIELD</imagingMode>
    <nominalDefocusMin units="nm">1000</nominalDefocusMin>
    <nominalDefocusMax units="nm">3000</nominalDefocusMax>
    <illuminationMode>FLOOD BEAM</illuminationMode>
    <detector>Direct Electron DE-12</detector>
    <nominalCs units="mm">2</nominalCs>
    <tiltAngleMin units="degrees">0</tiltAngleMin>
    <calibratedMagnification>42134</calibratedMagnification>
    <microscope>FEI TECNAI F20</microscope>
```

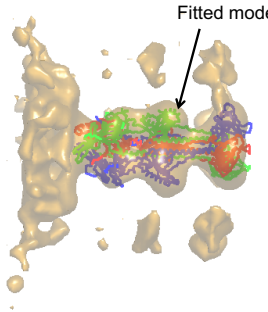

- Each EMDB entry consists (at least) of 1 header file + 1 map file
- Map files in **EMDB map format** (EMDB interpretation of CCP4 format)
- Formats: emdb-empiar.org/emschema

EMDBspeak 4

- **Fitting model**: model that will be shifted, rotated and possibly flexed to fit into EM map
- **Fitted model**: model that lies in frame with the map

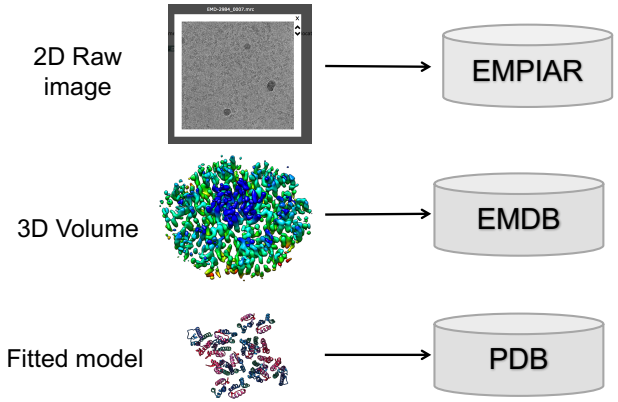


Fitting model



Data: What data goes where...

- Final single-particle and sub-tomogram average maps must go to EMDB (tomograms strongly recommended)
- Fitted/built models must go to PDB



From deposition to release

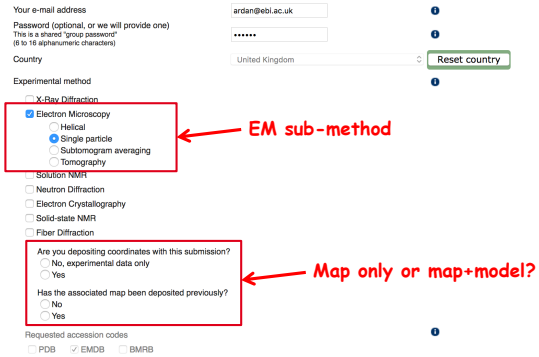
- You start by filling out a form and uploading the map and other data
- When you submit your deposition, it is locked and forwarded to the annotation team; you are assigned a deposition ID and an accession code
- The annotation team will check your deposition and will communicate with you regarding changes and clarifications
- Once annotation is completed, and you agree -> EMDB header file is released
- Map can be released immediately, kept on-hold until publication, or on-hold for 1 year
- "Immediate" is not immediate – weekly release cycle!

How to deposit?

- Some preparations
  - which EM maps? you can upload compressed maps
  - which fitted models? check that they overlap with the maps
    - note: each deposition can only accommodate as a maximum 1 map and 1 model
  - contour level
  - voxel size
  - what is the sample composition?
  - dig out your lab notebooks so that experimental info is readily available
- Check out the 12 golden rules and other help documents on emdatabank.org -> deposit
- Go to the OneDep system (https://deposit.wwpdb.org/deposition/)
  - When you select your country you will be directed to the correct site (PDBe, RCSB PDB or PDBj)

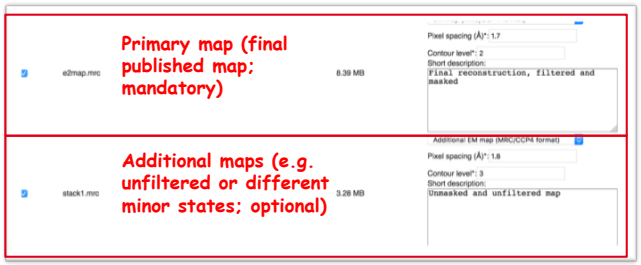
Deposition: submission options

- Chose EM sub-method carefully as this will determine options presented in deposition
- Tell the annotator if this is a part of a multipart deposition



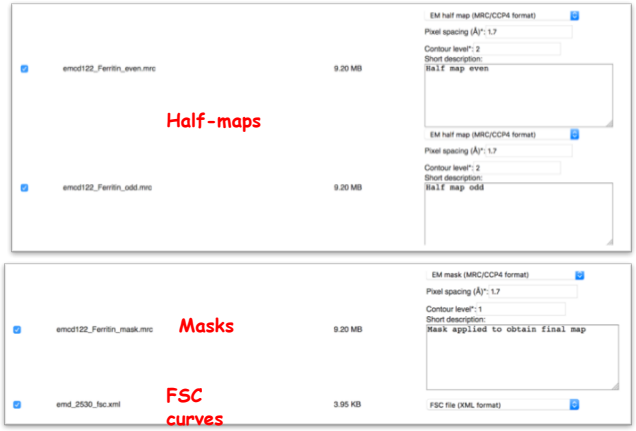
Deposition: file upload I

- Maps may be compressed: '.zip', '.gzip' or '.bz2'
- Maps larger than 1.5Gb -> ftp
- All maps converted to EMDB map format -> check that converted map and fitted models overlap!!
- Allowed formats: CCP4/MRC
- You are also asked to upload representative image (mandatory) that will be used on the website etc



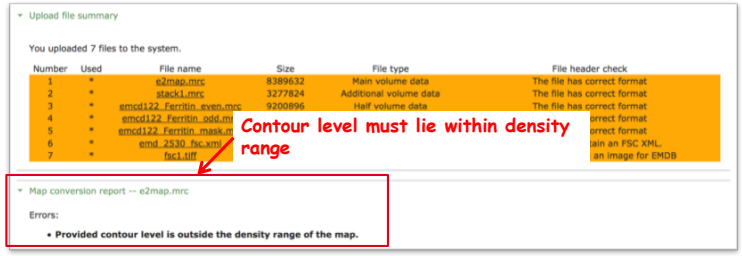
Deposition: file upload II

- Half maps, masks, fsc curves can be uploaded as specific file types (optional but strongly recommended)



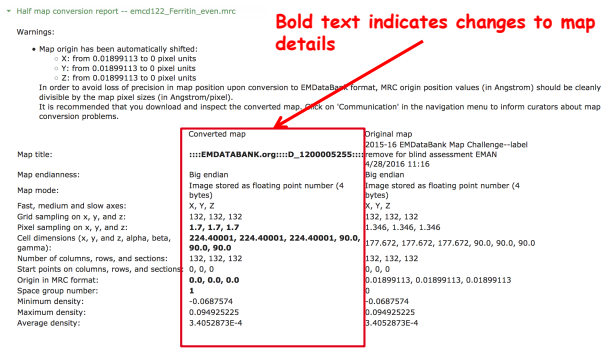
Check output of file upload I

- Download converted maps/models and check overlays etc



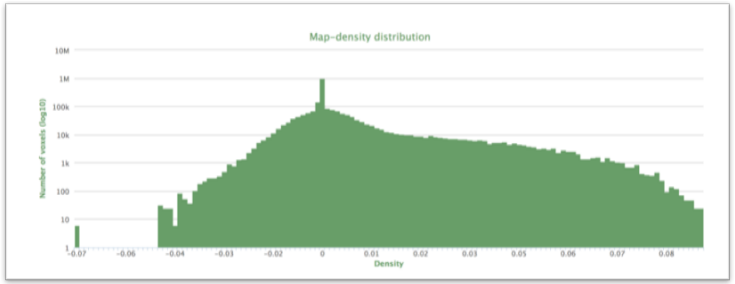
Check output of file upload II

- Again conversion may shift map etc. so download converted maps/models and check overlays etc



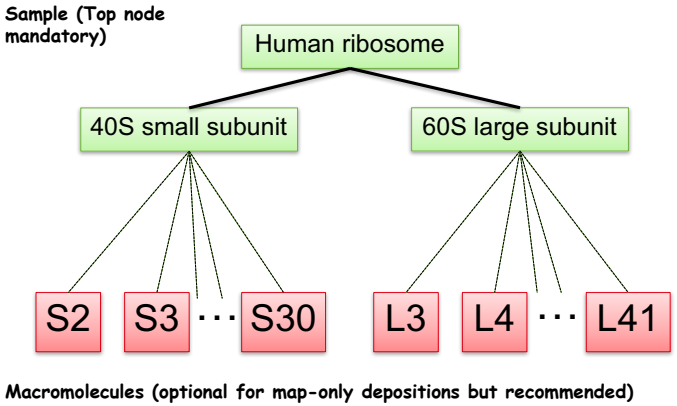
Check output of file upload III

- Check that histogram looks like similar to that e.g., in Chimera
- Weird histograms indicate byte-swapping issues etc



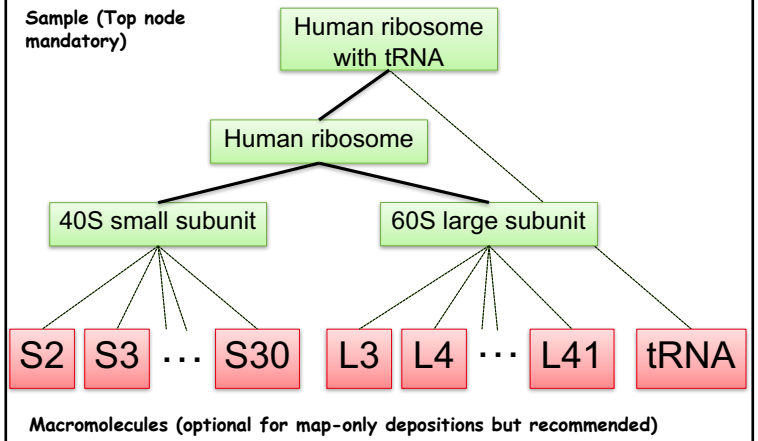
Sample description

- Two tiered description to accommodate molecular and cellular EM



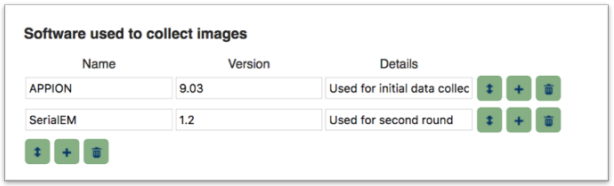
Sample description II

- Links can be from any sample node to macromolecules



Software usage

- During the deposition there are several places where you can provide info about software used
- You can specify more than one package
- This information is very important for software developers

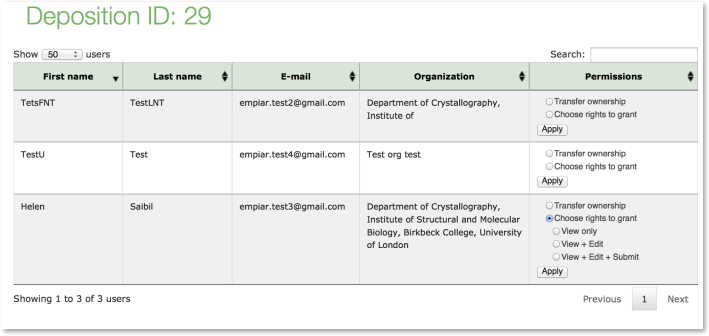


EMPIAR deposition

- Go to empiar.org/deposition
- Read the deposition manual
- User based login – one user can have multiple depositions and users can share depositions
- Creat login account or login using facebook/google/orcid

EMPIAR deposition – user roles

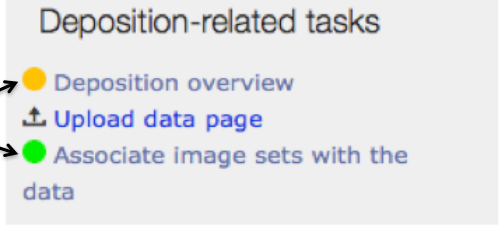
- PI can, for example, grant access to student to fill out form but not submit



EMPIAR deposition workflow

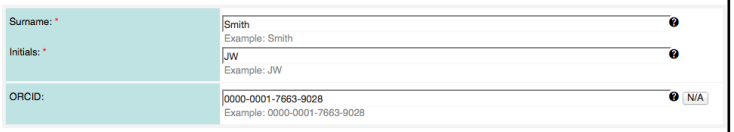
- providing the general metadata about the deposition — EMDB accession code, title authors, etc.;
- uploading the data — the transfer can take some time, so the next step most likely would not be undertaken immediately;
- associating the uploaded data with the corresponding image sets — that is, identifying the image sets present and describing them.

Traffic lights indicate if mandatory items on page have been completed



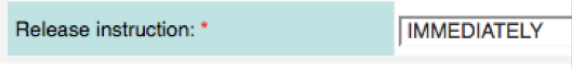
Deposition overview

- For single particle EM and electron tomography, a link to EMDB entry is mandatory
  - This also means that we can extract some info from the related EMDB entry, e.g., authors
- Red star(*) indicates mandatory items
- For fields with an N/A button, you have to explicitly press the button if you have no value for the field



Release instruction

- IMMEDIATELY: As soon as the processing has finished the release procedure will be initiated. For datasets in the TB range this could take several days
- EMDBPUB: Wait till the associated EMDB entry has been released before releasing the EMPIAR entry. This would require the public release by EMDB of both the header and the map
- HPUB: Wait until the primary citation for the associated EMDB entry has been published or max 1 year (whichever comes first)
- HOLD1: Wait for 1 year

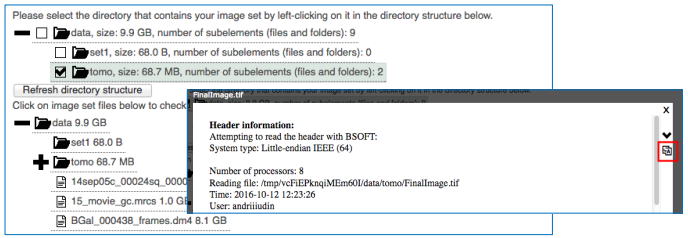


File upload page

- Three options — Aspera via command line, Aspera via web-client or Globus
- The Aspera web client allows you to upload max ~ 300 files at a time and best for < 400GB datasets
- Data-transfers commonly proceed at 50 - 200 GB per hour so expect TB+ sized datasets to take days in some cases
- Structure your data so that each raw image set is in a separate directory
- Multiple image sets are fine
- Each image set should be consistent (images of the same format/conditions/sizes etc)
- Use of common file formats is encourages, e.g., MRC
- You can include auxilliary data such as particle picking coordinates
  - Put them in directories but do not mark them as “image sets”
  - Use free text ”details” field to describe them for now

Associate image sets with data page

- You need to tell us how your data is organized
- Checking the uploaded data
  - System will highlight zero sized files
  - there is a downloadable script for md5 checksum verification
- Associate datasets with directories
- File parameters can be harvested from files



Invite reviewers

- Give anonymous access to reviewers to browse and download your data prior to submission
- Generate link and pass the link on to the journal
- The link can only be used once