3DEM Data Archiving & Validation, Challenges

Cathy Lawson

NYSBC CryoEM Course

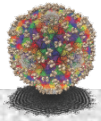
April 10, 2017

EMDataBank

Unified Data Resource for 3DEM

3DEM Data Archives for Structural Biology

- PDB – coordinates - managed by wwPDB
- EMDB – maps - managed by EMDataBank
- EMPIAR – raw images - managed by EBI



EMDataBank
Unified Data Resource for 3DEM

Protein Data Bank (PDB)

- Single global archive of 3-D macromolecular structures (>125,000 entries) est. 1971
- 1990s: First EM structures deposited
- 2017: ~1500 EM structures
- Managed by the world-wide PDB organization

EM Data Bank (EMDB)

- Global archive for maps produced using 3DEM methods, est. 2002 @ EBI
- Since 2007, managed by EMDataBank partners, with support from NIGMS
- We expect to reach 5000 entries by end of 2017

NCMI National Center for Macromolecular Imaging

RCSB PDB PROTEIN DATA BANK

PDBe Protein Data Bank in Europe

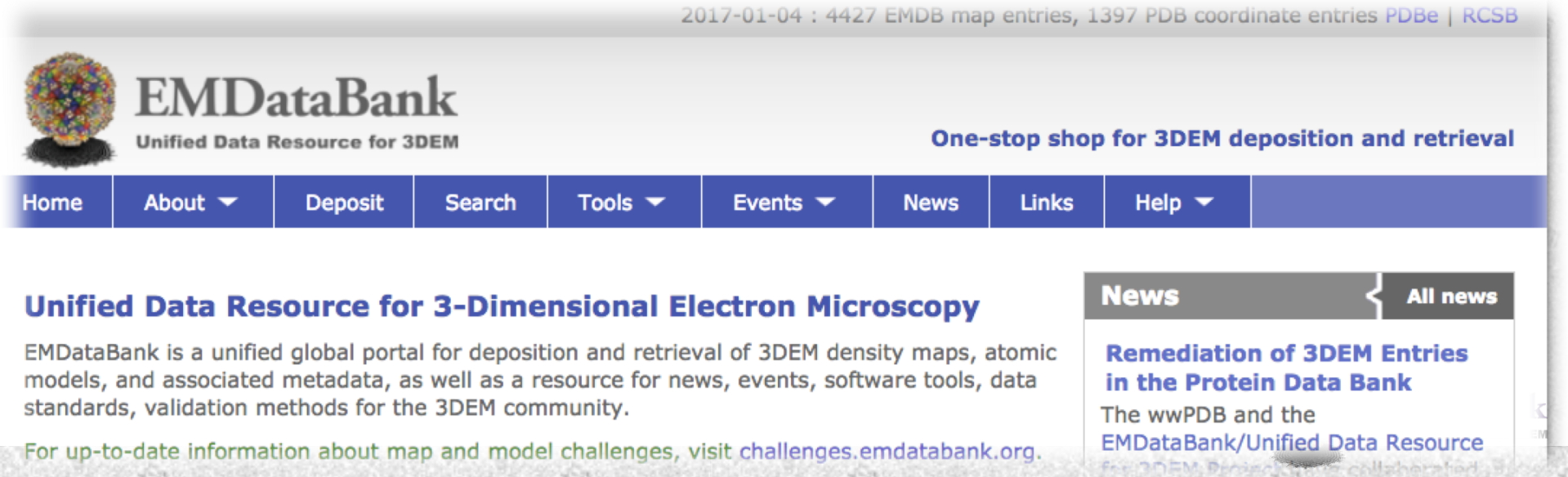
EMDataBank Unified Data Resource for 3DEM

EMPIAR (pronounced empire)

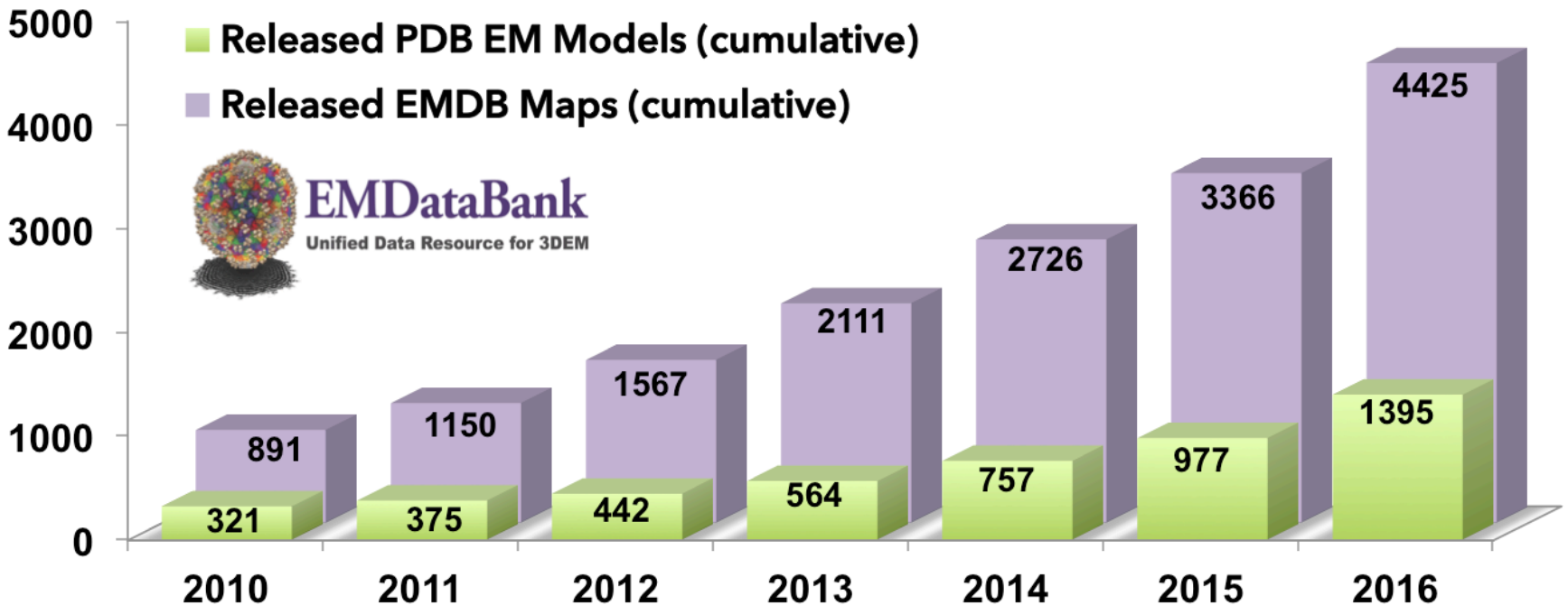
- Archive for raw electron microscopy image data est. @ EBI/PDBe in 2013
- ~100 entries
- Supported by MRC, BBSRC

EMDataBank.org

- Global portal for deposition and retrieval of 3DEM density maps, atomic models, and associated metadata
- Resource for news, events, software tools, data standards, validation methods for the 3DEM community

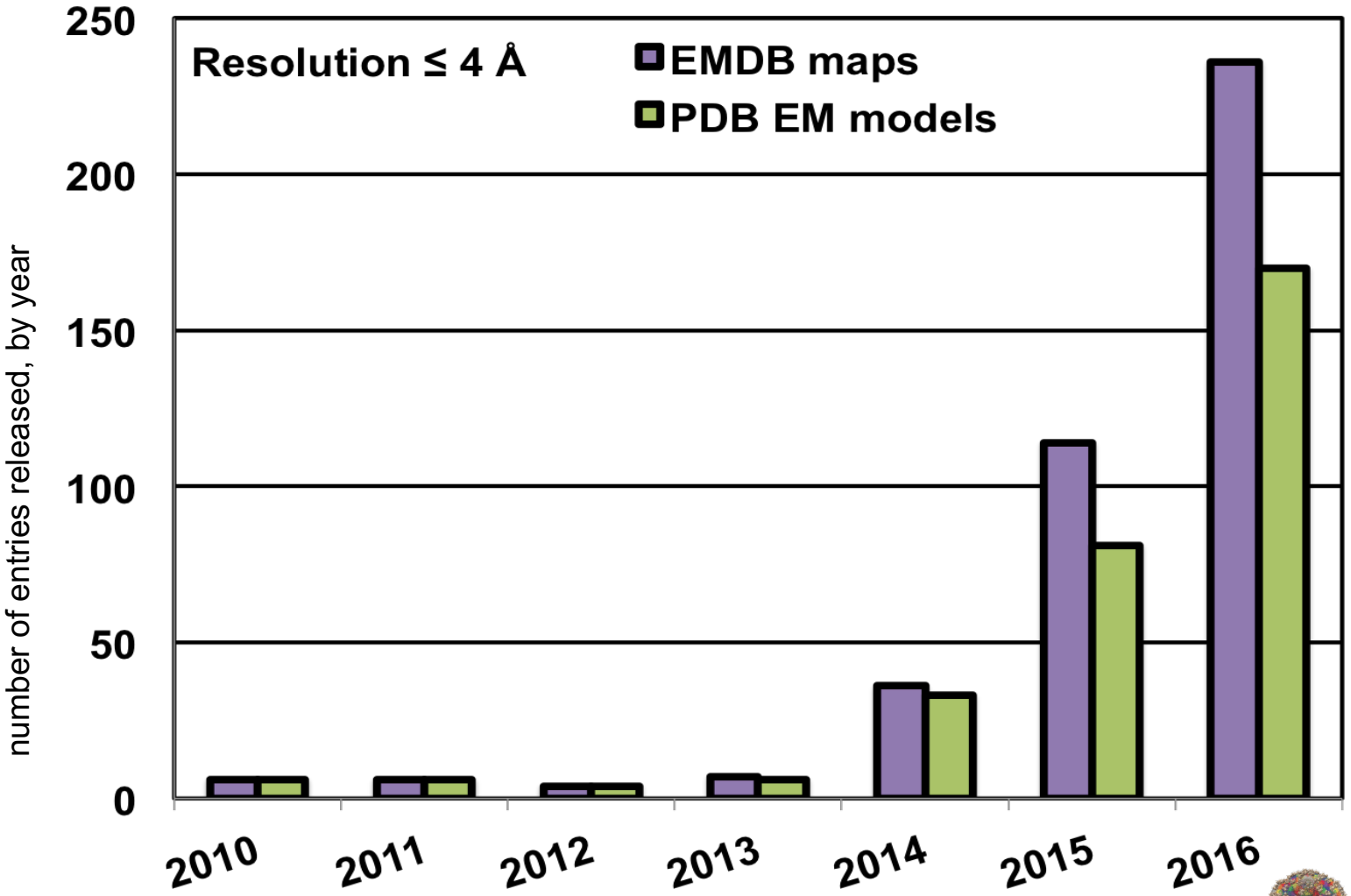


Growth of 3DEM Archives



EMDataBank
Unified Data Resource for 3DEM

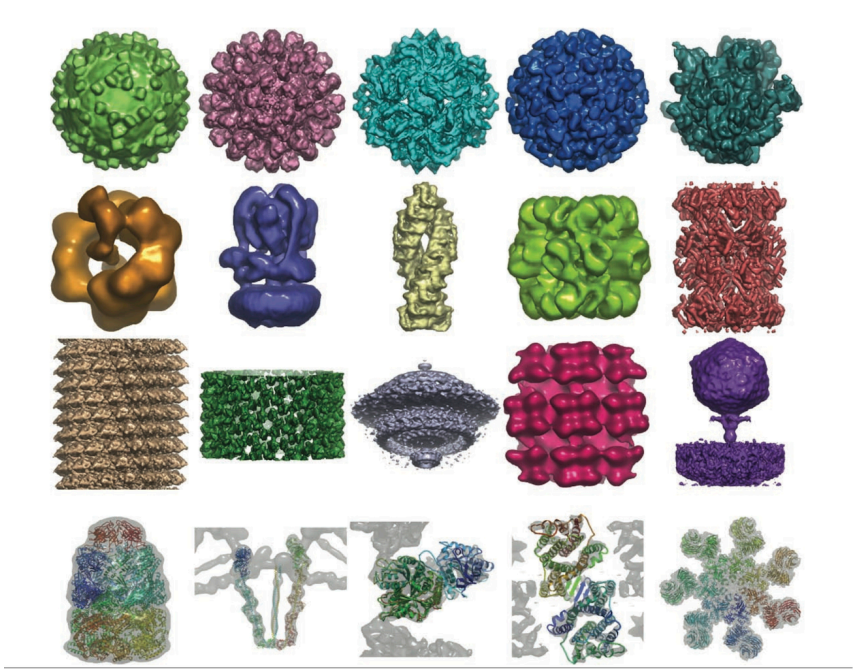
EM Structures @ 4 Å or better



EMDataBank
Unified Data Resource for 3DEM

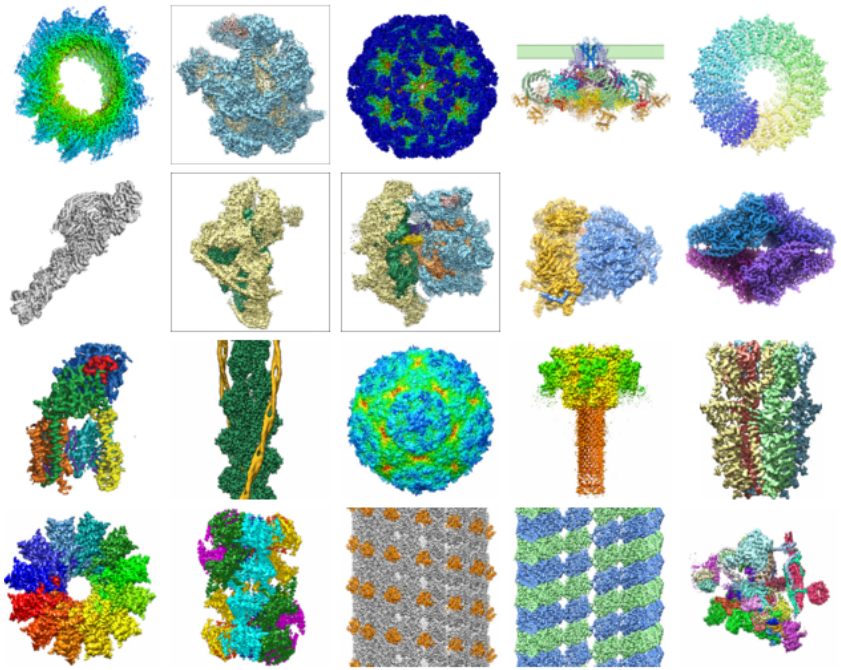
EM Structures 2010 vs 2015

2010: Molecular Shapes



0.5% of all entries in PDB
(332 of 67500)

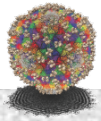
2015: Traceable Densities



0.8% of all entries in PDB
(905 of 112400)

EMDataBank
Unified Data Resource for 3DEM

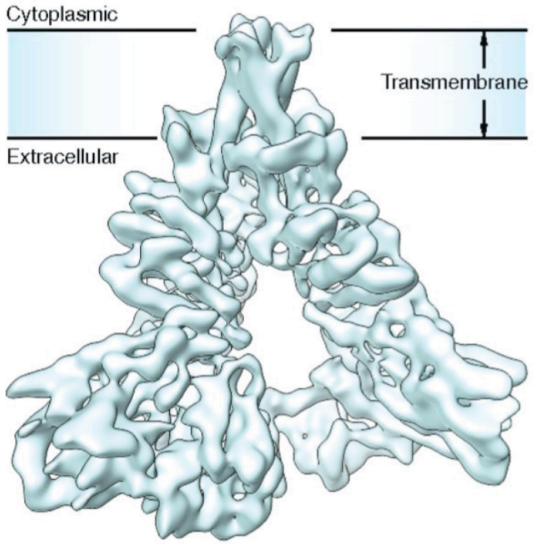
3DEM Structure Validation



EMDataBank
Unified Data Resource for 3DEM

Importance of Validation

- **J. Cohen, Is High-Tech View of HIV Too Good to Be True?** *Science* 341, 443-444 (2013)
- **R.M. Glaeser, Replication and validation of cryo-EM structures** *J. Struct. Biol.* 184, 379-380 (2013)
- **R. Henderson, Avoiding pitfalls of single particle cryo-electron microscopy: Einstein from noise,** *PNAS* 110, 18037-41 (2013)
- **M. van Heel , Finding trimeric HIV-1 envelope glycoproteins in random noise,** *PNAS* 110, E4175-7 (2013)
- **S. Subramaniam, Structure of trimeric HIV-1 envelope glycoproteins,** *PNAS* 110, E4172-4 (2013)



EMD-5418 Y Mao, JG Sodroski *et al*. Molecular architecture of the uncleaved HIV-1 envelope glycoprotein trimer *PNAS* 110, 12438-12443 (2013)

Community Input for Validation

Task Force	Meeting/ Workshop	Chair(s)/Membership	Outcome
X-ray Validation Task Force	2008 (2015)	Randy Read (Univ of Cambridge) 17 members	(2011) *Structure* 19: 1395-1412
NMR Validation Task Force	2009, 2011, 2013 (x2), 2015	Gaetano Montelione (Rutgers) Michael Nilges (Institut Pasteur) 10 members	(2013) *Structure*, 21: 1563-1570
3DEM Validation Task Force	**2010**	**Richard Henderson (MRC-LMB) Andrej Sali (UCSF) 21 members**	**(2012) *Structure* 20: 205-214**
Small-Angle Scattering Task Force	2012, 2014	Jill Trewhella (Univ Sydney) 6 members	(2013) *Structure* 21: 875-881
Hybrid Methods Workshop	2014	Andrej Sali (UCSF), Torsten Schwede (Univ Basel), Jill Trewhella (Univ Sydney) 27 members	(2014) *Structure* 23: 1156-1167

EM Validation Task Force

Henderson *et al*. (2012) ***Structure 20***, 205-214
http://www.ncbi.nlm.nih.gov/pubmed/22325770

EMDataBank
Unified Data Resource for 3DEM

EM VTF Recommendations

- Main recommendations: EM maps
  - Standards for assessing resolution and accuracy of a map need to be developed
  - Structural features in a map should be in accordance with the claimed resolution
- Main recommendations: models fitted into EM maps
  - Criteria for assessing models in context of map-model fit need to be developed
  - Capability to archive coarse-grained representations of models is needed
- More research and development needed!

EM VTC on Resolution

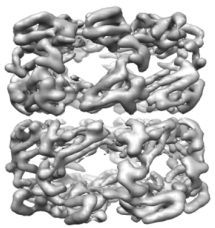
- Resolution via FSC: “Deposition of a published map should include its full FSC curve to the Nyquist frequency on a linear spatial-frequency scale.”
  - Masking: discouraged
  - Fully independent refinement: encouraged
- Low resolution structures (where secondary structure elements cannot be identified) can be validated using tilt-based methods

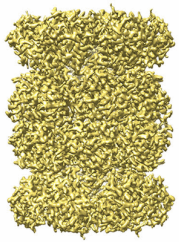
2015-Ongoing: Map, Model Challenges

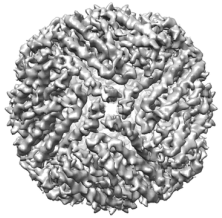
- Goals: Develop benchmarks, encourage development of best practices in 3DEM reconstruction and model fitting, **evolve criteria for validation**, compare and contrast different approaches
- Developed by expert committees
- Results discussion via Participant Workshops/Journal Special Issues
- http://challenges.emdatabank.org

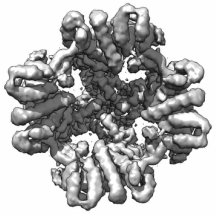
Benchmark Datasets

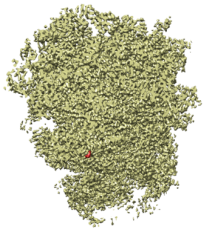
Map Challenge Targets: Raw Images @ EMPIAR

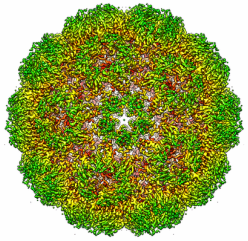

GroEL

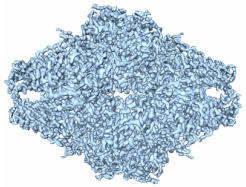

T20S Proteasome


Apo-Ferritin

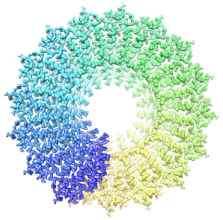

TrpV1 channel

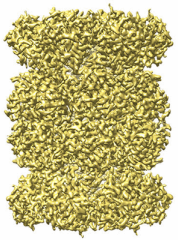

80S Ribosome

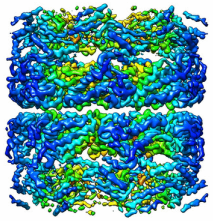

Brome Mosaic Virus

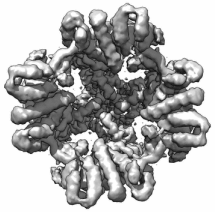

ß-galacto-sidase

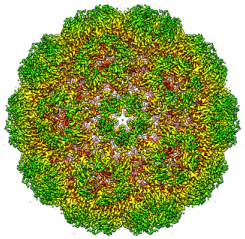
Model Challenge Targets: Maps @ EMDB

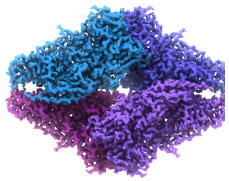

Tobacco Mosaic Virus

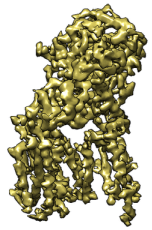

T20S Proteasome

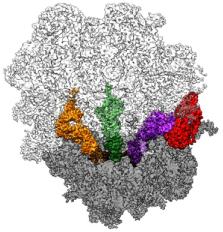

GroEL


TrpV1 channel


Brome Mosaic Virus


ß-galacto-sidase


γ-Secretase


70S Ribosome

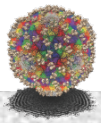
EMDataBank
Unified Data Resource for 3DEM

Challenges: Progress

- Map Challenge:
  - 27 participants submitted 66 maps
- Model Challenge:
  - 16 participants submitted 106 models

Lots of Data Received! – assessments now in progress

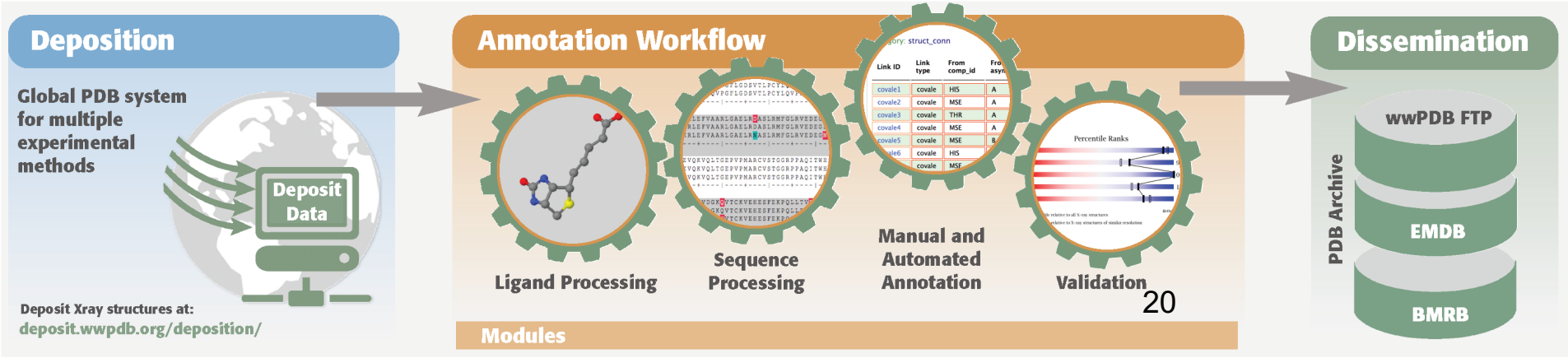
3DEM Structure Deposition



EMDataBank
Unified Data Resource for 3DEM

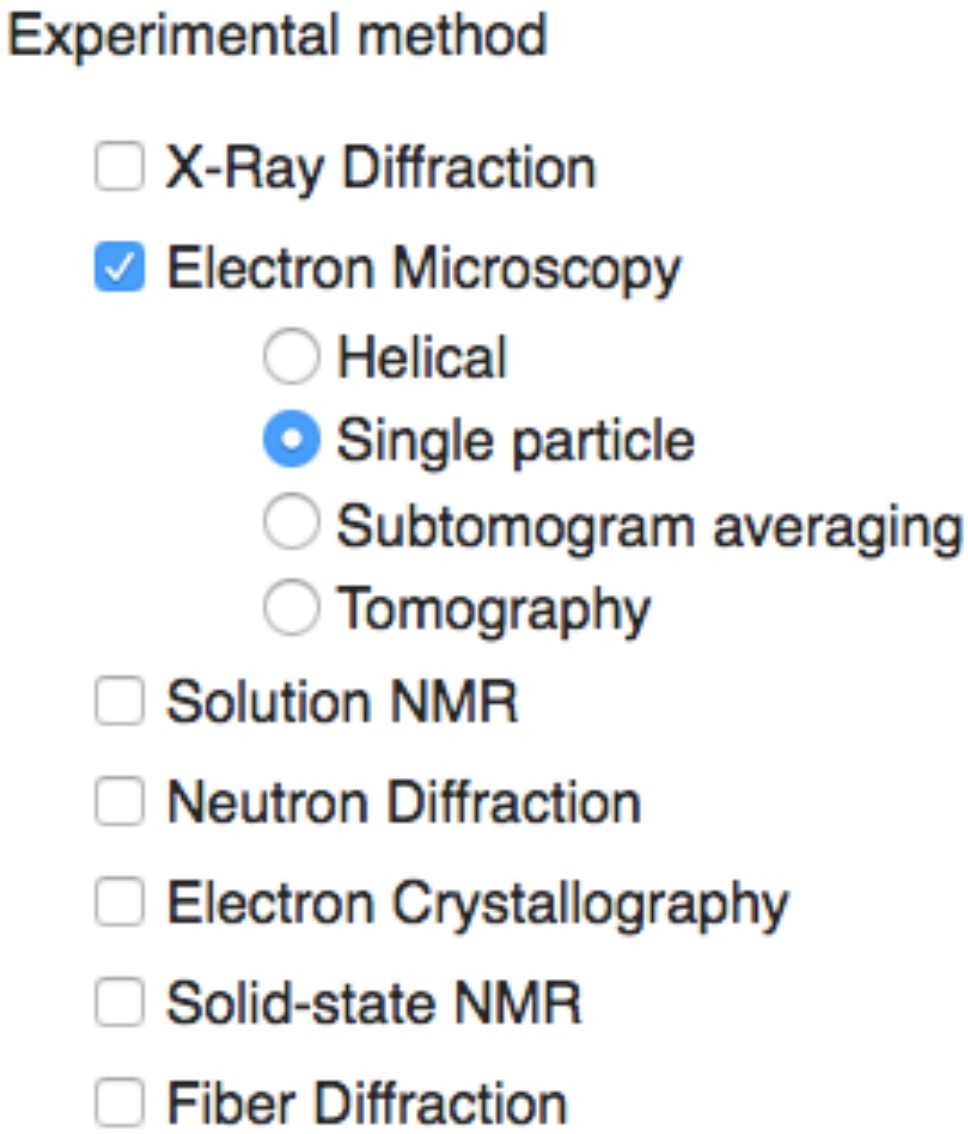
wwPDB OneDep System

- New Version for depositing structures from X-ray, NMR, and EM Launched January 2016
- Deposit map to EMDB with associated model to PDB in single session
- Validation report provided



3DEM Deposition: Method

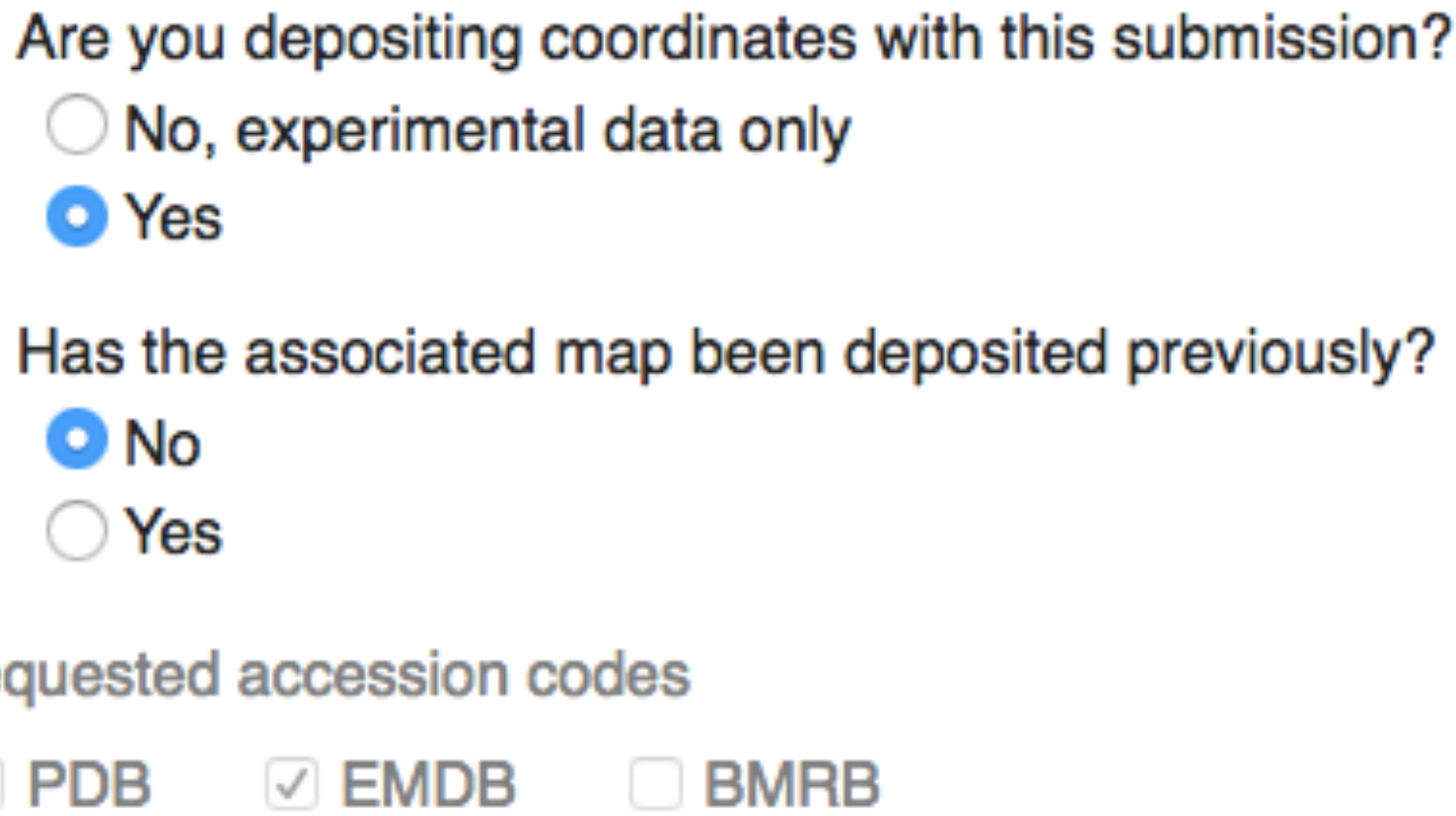
http://deposit.wwpdb.org/deposition/



EMDataBank
Unified Data Resource for 3DEM

3DEM Deposition: IDs

http://deposit.wwpdb.org/deposition/



EMDataBank
Unified Data Resource for 3DEM

File uploads: 3DEM map/model submission in OneDep

✓Select file type...

0) Coordinates

- Coordinates (mmCIF format)
- Coordinates (PDB format)

1) Main map (mandatory)

- EM map (MRC/CCP4 format)

2) Image for EMDB (mandatory)

- Entry image for public display

3) Additional maps

- Additional EM map (MRC/CCP4 format)

4) Masks

- EM mask (MRC/CCP4 format)

5) Half (even-odd) maps

- EM half map (MRC/CCP4 format)

6) Structure Factors

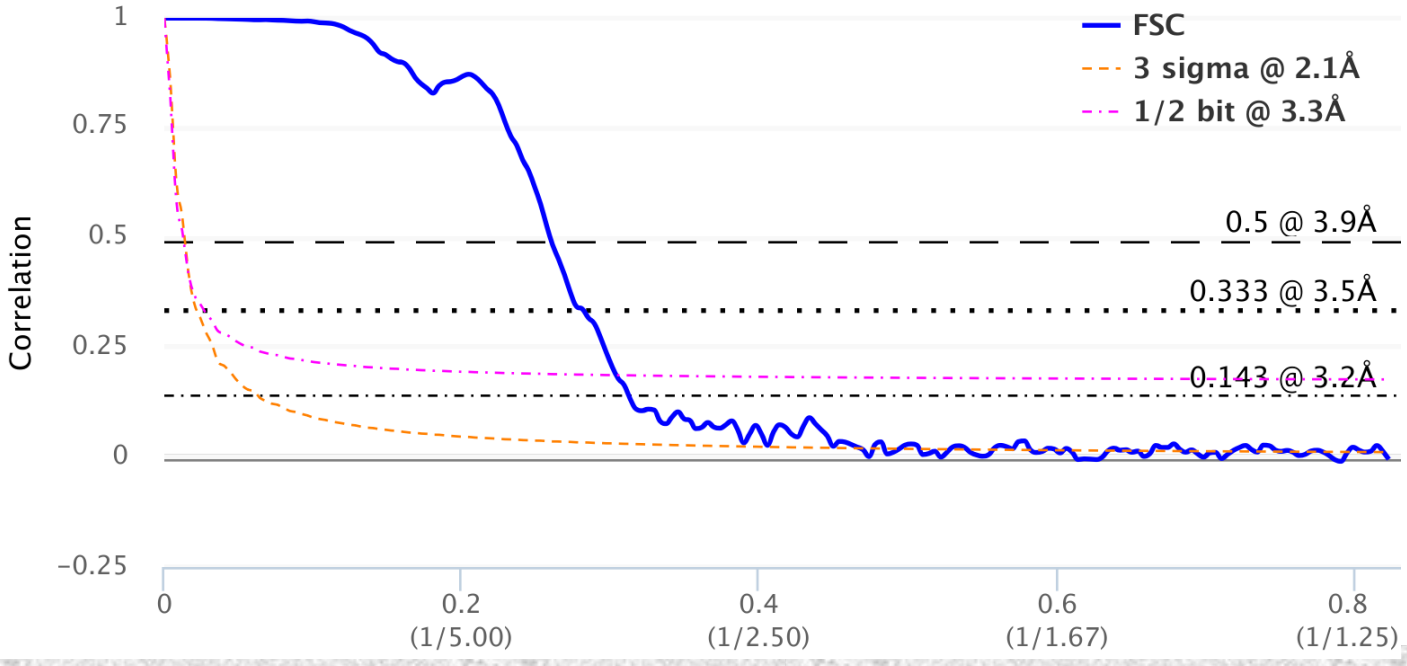
- mmCIF (structure factors)
- MTZ

Other Files

- FSC file (XML format)
- Ligand Image

FSC Curve Upload

- Create xml format file using a software package (e.g., Relion, EMAN)
- Use PDBe's Server: PDBe.org/FSC

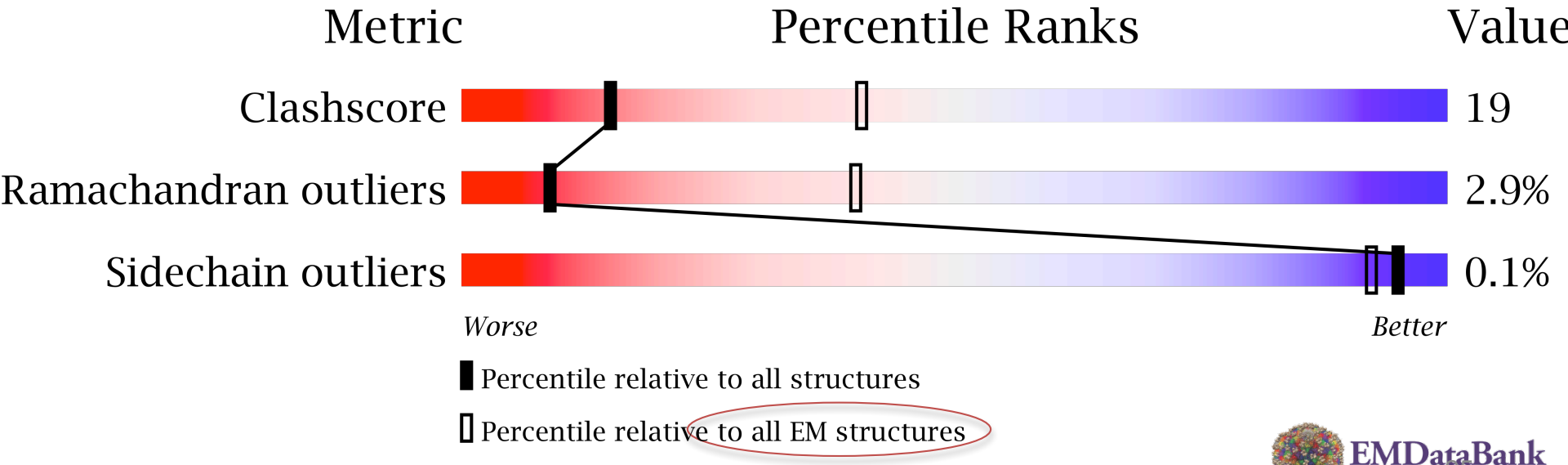


EMDataBank
Unified Data Resource for 3DEM

EM Validation Report

- "Table 1"
- Metrics of the EM model

Property	Value	Source
Reconstruction method	SINGLE PARTICLE	Depositor
Imposed symmetry	I	Depositor
Number of images	30000	Depositor
Resolution determination method	FSC 0.143	Depositor
CTF correction method	Not provided	Depositor
Microscope	JEOL 3200FSC	Depositor
Voltage (kV)	300	Depositor
Electron dose ($e^-/^2$)	Not provided	Depositor
Minimum defocus (nm)	500	Depositor
Maximum defocus (nm)	2000	Depositor
Magnification	50000	Depositor
Image detector	DIRECT ELECTRON DE-12 (4k x 3k)	Depositor



EM Validation Report: Proposed additions

- Table of Map Parameters
- Map images
- Plots
- Model to map fit
- Example: http://www.ebi.ac.uk/pdbe/entry/emdb/EMD-2984/analysis

Project Team

Baylor College of Medicine

Wah Chiu, PI
Steven Ludtke
Corey Hryc
Grigore Pintilie
Matthew Baker
Matthew Dougherty

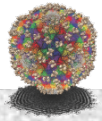
Rutgers University

Helen Berman, co-PI
Catherine Lawson
Raul Sala
Brian Hudson
John Westbrook

European Bioinformatics Institute

Gerard Kleywegt, co-PI
Ardan Patwardhan
Sanja Abbott
Eduardo Sanz Garcia
Ingvar Lagerstedt

Supported by NIH National Institute of General Medical Sciences



EMDataBank
Unified Data Resource for 3DEM