EMDataResource: 3DEM Structure Data Archiving, Validation Challenges

Cathy Lawson

Rutgers University

NYSBC/Simons EM Center

Winter EM Course

April 16, 2018

EMDataResource

Unified Data Resource for 3DEM

Unified Data Resource for 3DEM

Wah Chiu, PI
Stanford University

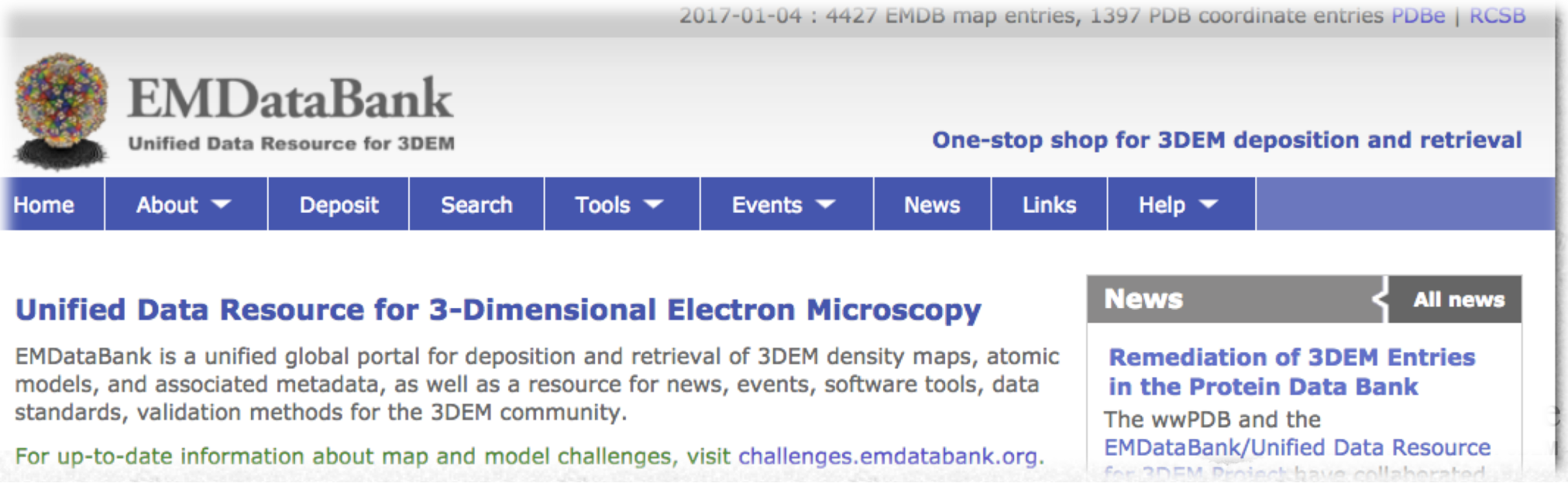
Helen Berman, CoPI
Cathy Lawson
Rutgers University

Ardan Patwardhan
European Bioinformatics Institute

- Established 2007 under NIGMS Support (R01GM079429) to:
- Develop Data Infrastructure/Archives for 3DEM
- Promote Community Development of Validation and Standards

Project Website

- Global portal for deposition and retrieval of 3DEM density maps, atomic models, and associated metadata (EMDB/PDB).
- Resource for news, events, software tools
- Outreach for data standards, validation methods



EM Standards / Validation Development

2004:
Dictionary Development Workshop

2010:
Validation Task Force Workshop

2010:
Model Challenge

2011, 2012, 2015:
Data Management Workshops

2015-2017:
Map and Model Challenges

EMDataResource
Unified Data Resource for 3DEM

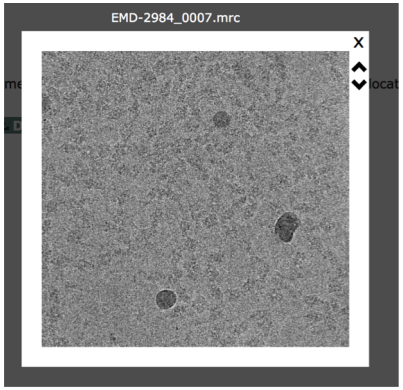
3DEM Data Archives

Empiar, EMDB, PDB

EMDataResource
Unified Data Resource for 3DEM

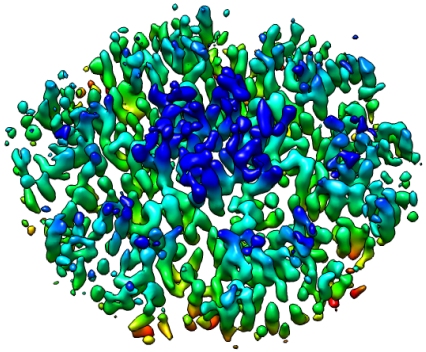
Data Archives: What data is found where...

2D Raw images



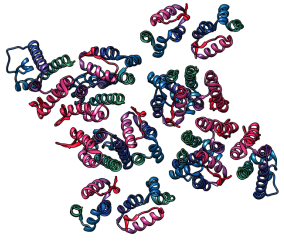
EMPIAR

3D Volumes



EMDB

Fitted models



PDB

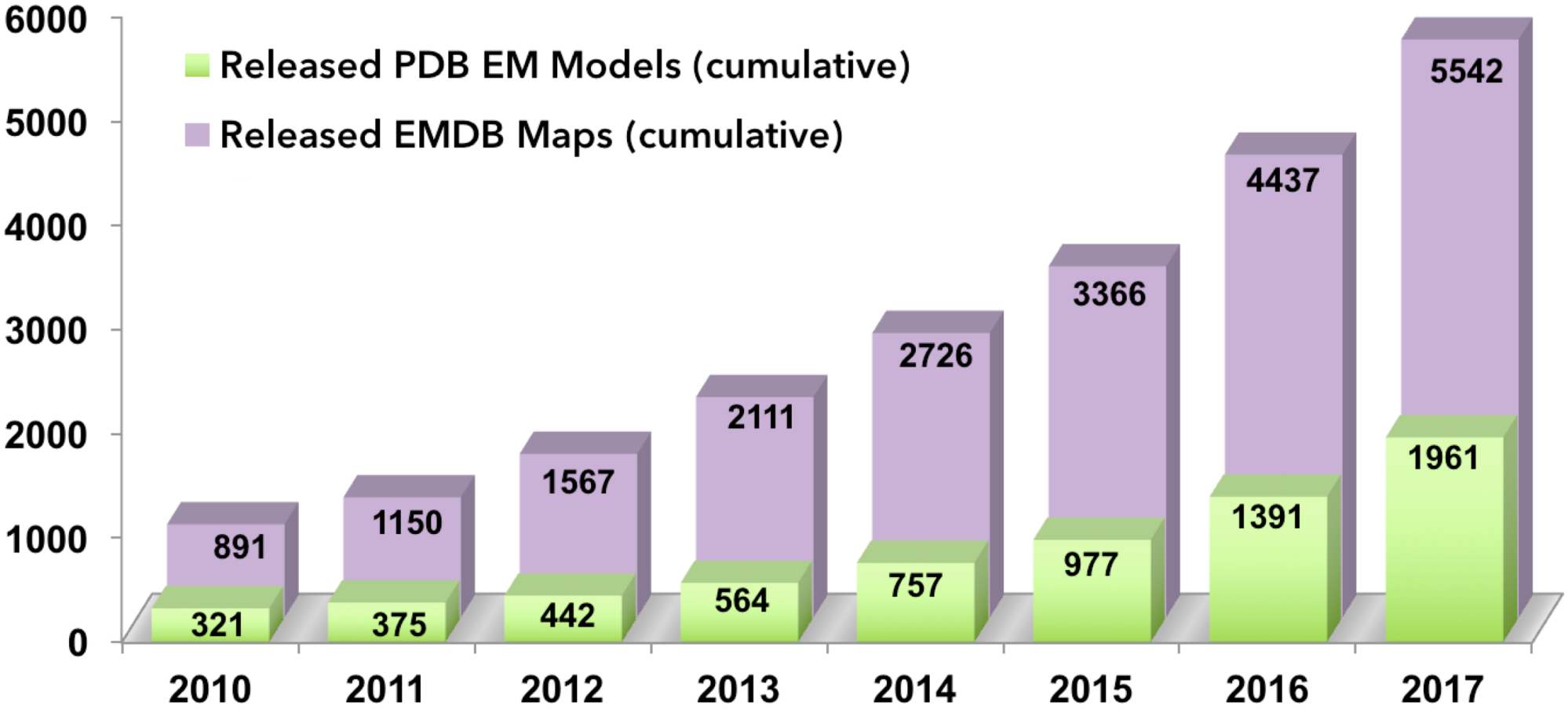
EMDataResource
Unified Data Resource for 3DEM

Comparison of Data Archives (Jan 2018)

	PDB	EMDB	EMPIAR
Inception Year	1971	2002	2013
# Entries	136594 (1963 EM)	5543	119
Archive size	1 GB	1/2 TB	50 TB
Community/Journal Deposition Policies	Coordinates (1989) Structure factors (2008)	Single particle, sub-tomogram avg. maps (2012) Representative tomogram recommended	-
Reference	Berman et al 2003 10.1038/nsb1203-980	Lawson et al 2016 10.1093/nar/gkv1126	Iudin et al 2016 10.1038/nmeth.3806

EMDataResource
Unified Data Resource for 3DEM

Growth of EM Structure Archives

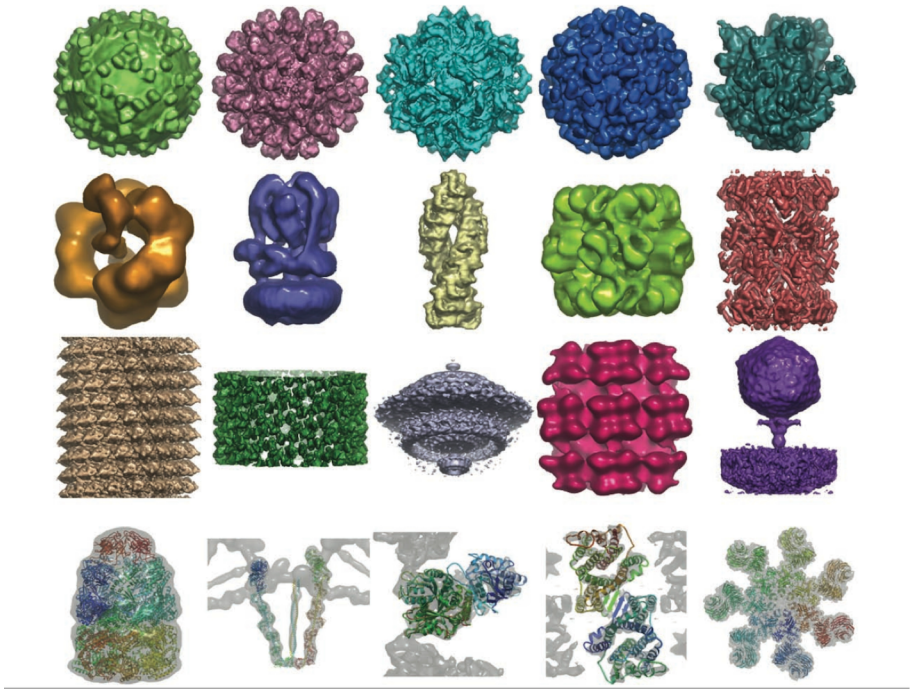


emdatabank.org/statistics.html

EMDataResource
Unified Data Resource for 3DEM

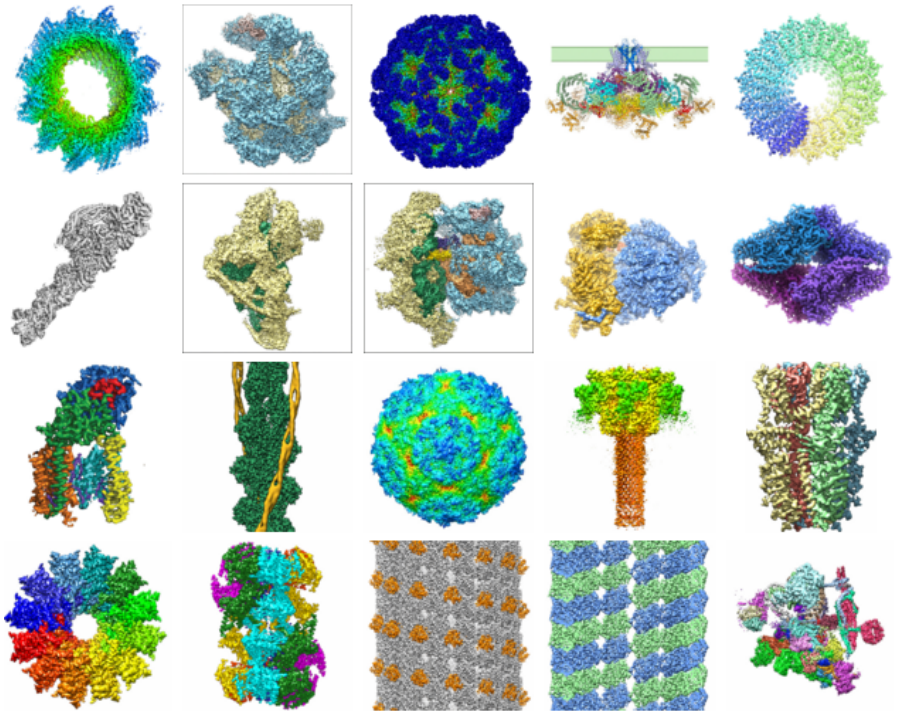
EM Structures 2010 vs 2015

2010: Molecular Shapes



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

2015: Traceable Densities

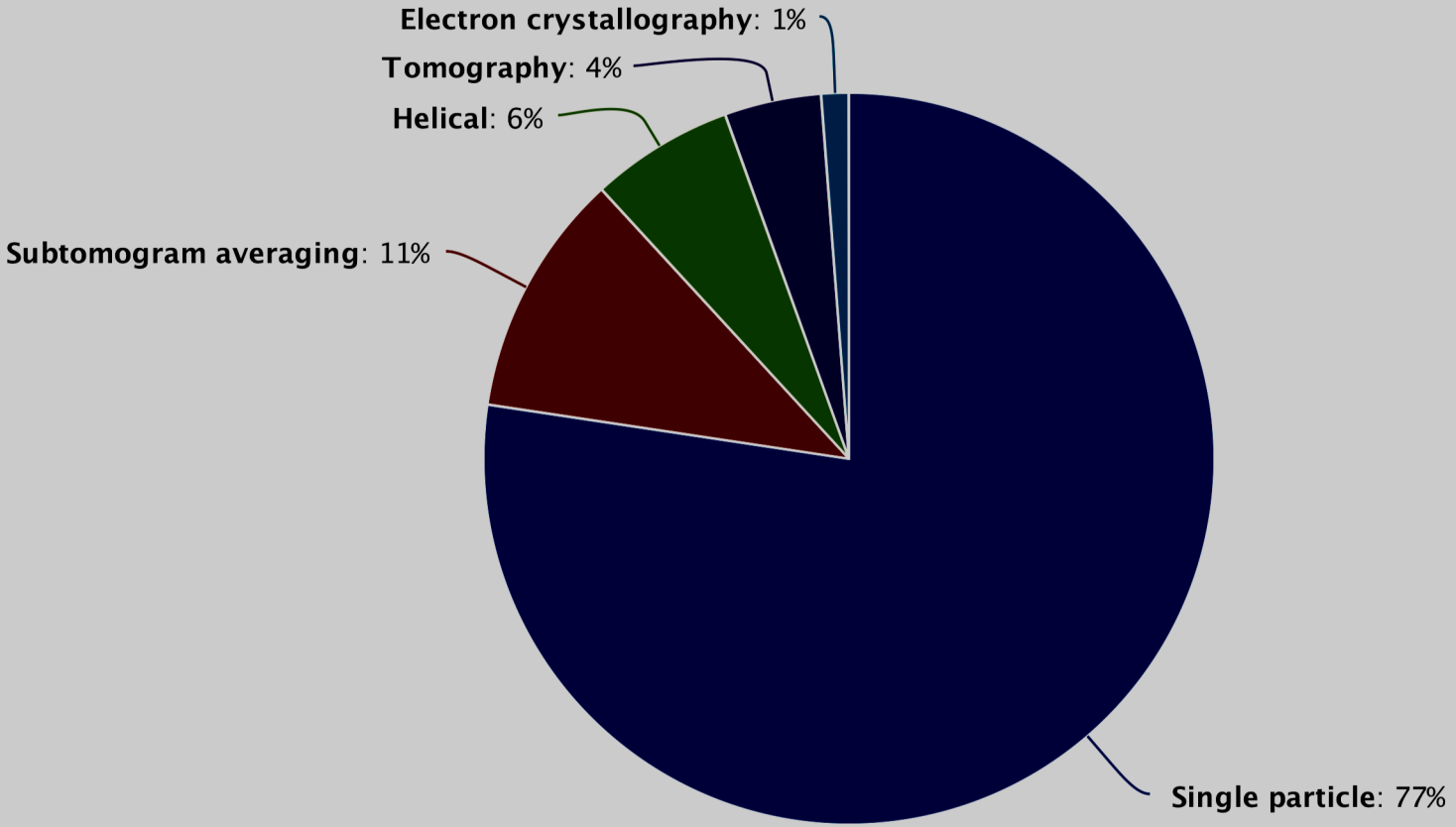


0.8% of all entries in PDB
(905 of 112400)

EMDataResource
Unified Data Resource for 3DEM

Types of Maps Archived in EMDB

Distribution of released maps (5543 in total) as a function of technique used



EMDataResource
Unified Data Resource for 3DEM

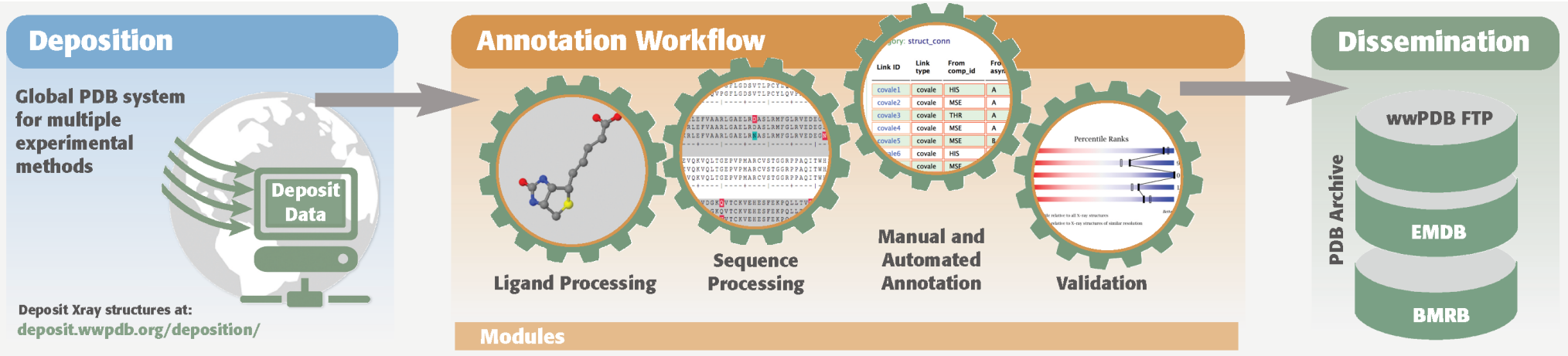
3DEM Structure Deposition

EMDB, PDB

EMDataResource
Unified Data Resource for 3DEM

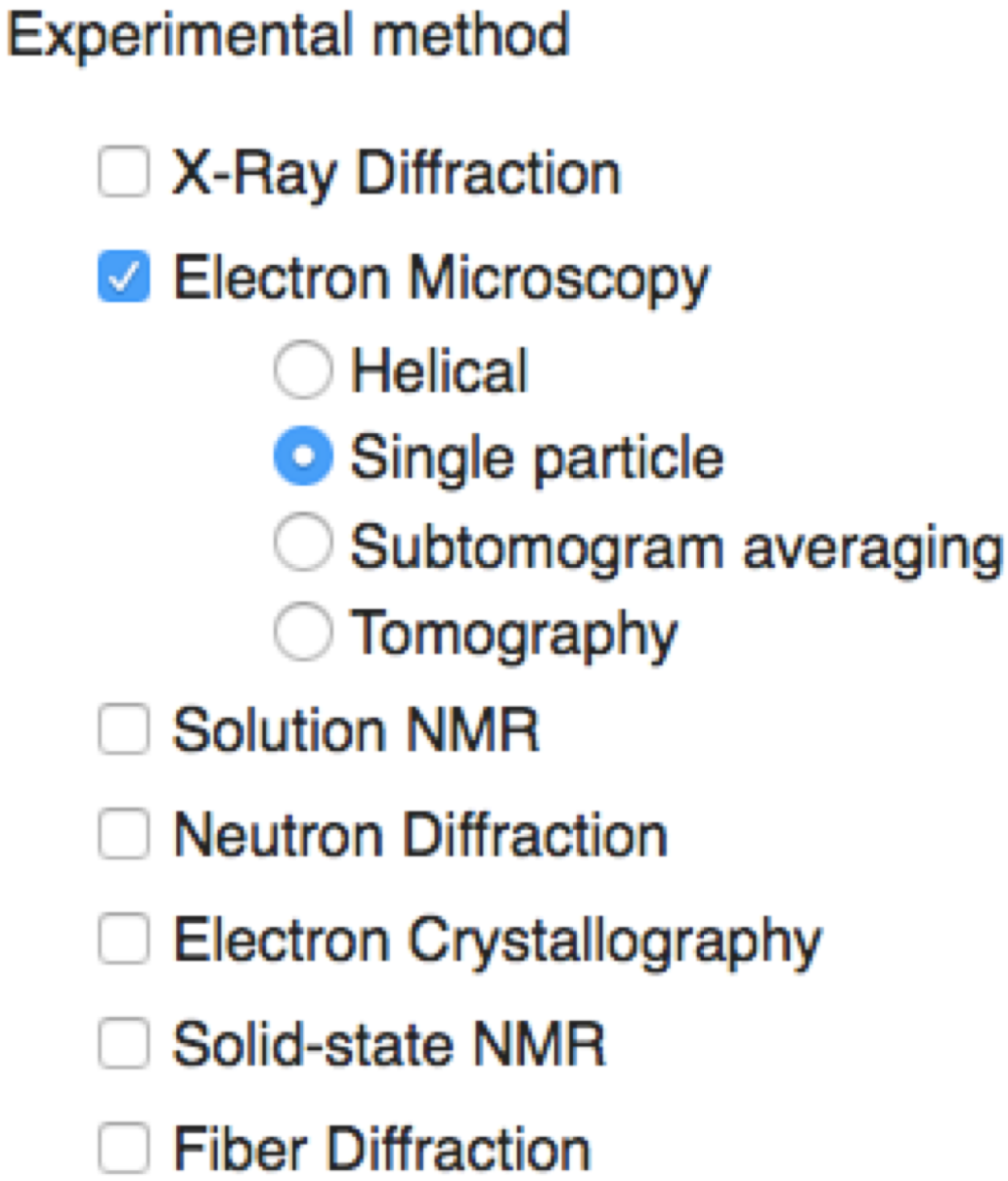
wwPDB OneDep System

- X-ray, NMR, and EM Methods (since 2016)
- EM Methods: Deposit map to EMDB with associated model to PDB
- Validation report produced



3DEM Deposition: Method

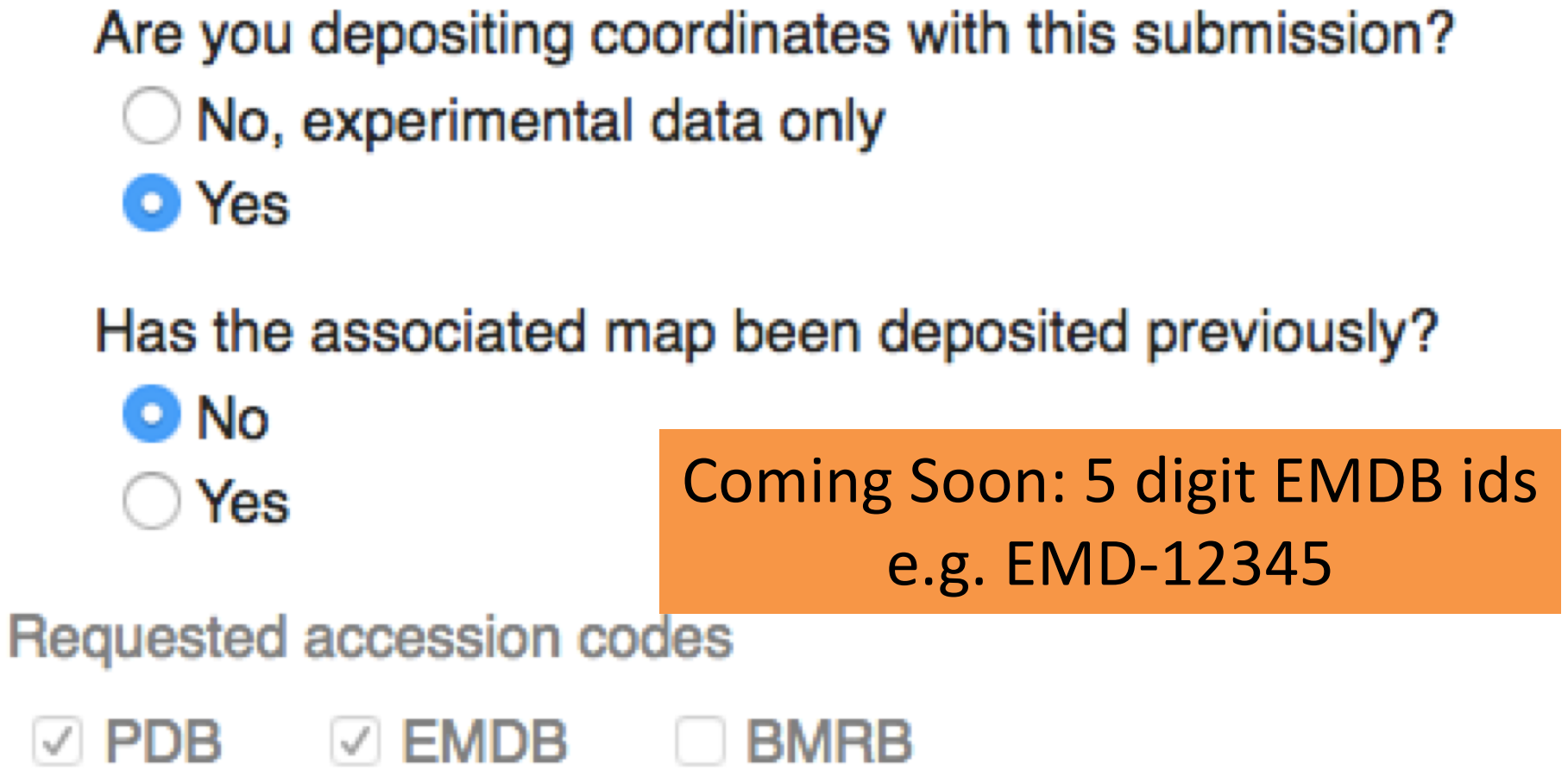
deposit.wwpdb.org



EMDataResource
Jnified Data Resource for 3DEM

3DEM Deposition: ID Assignment

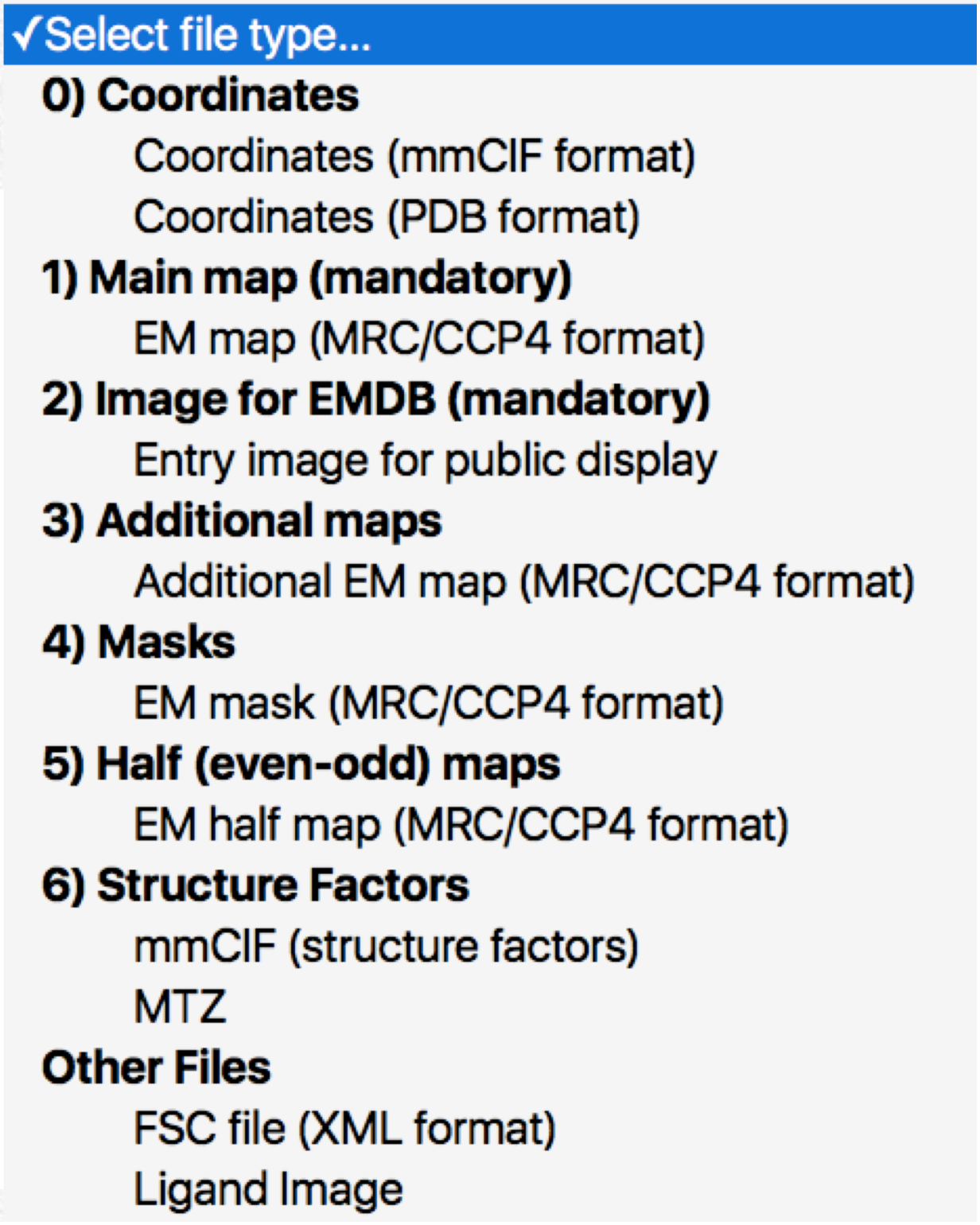
deposit.wwpdb.org



Coming Soon: 5 digit EMDB ids
e.g. EMD-12345

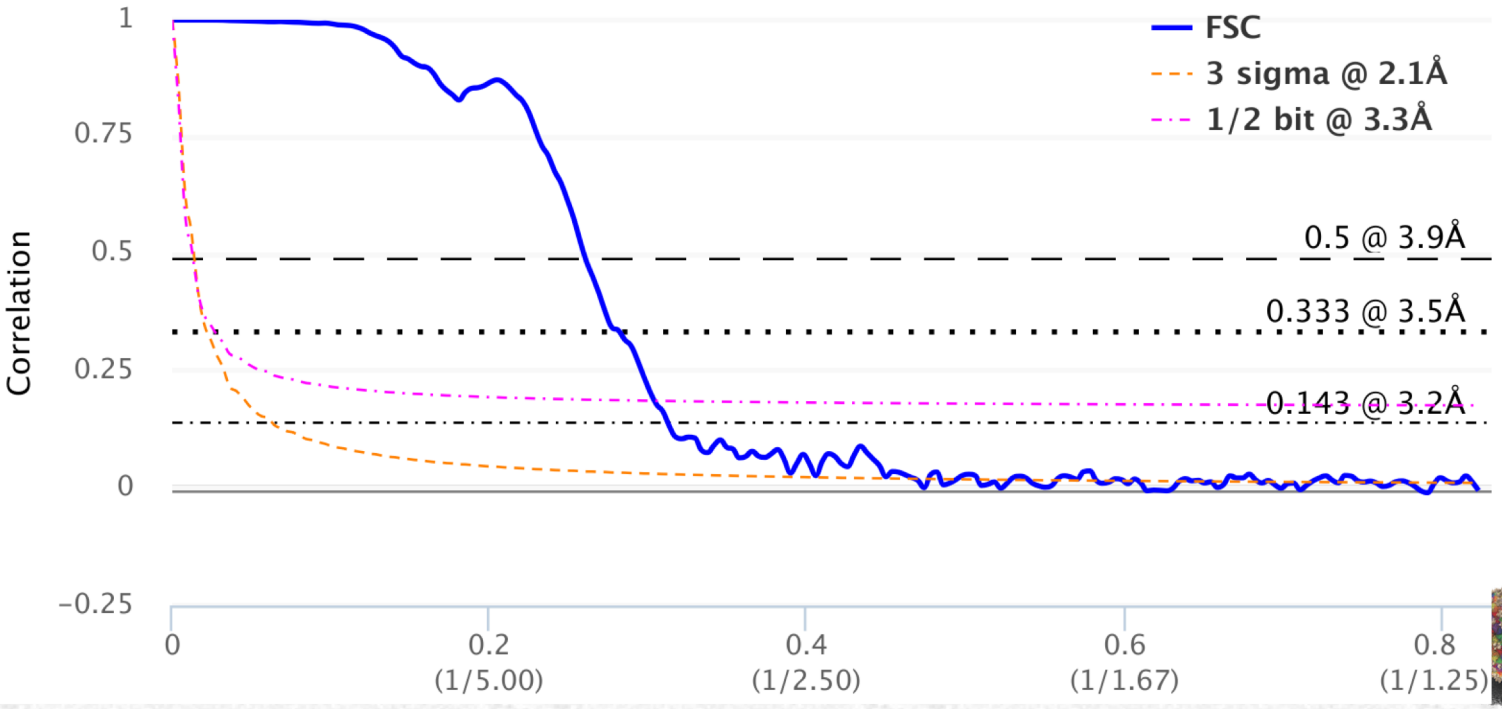
EMDataResource
Unified Data Resource for 3DEM

File uploads: 3DEM map/model submission in OneDep



FSC Curve Upload

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

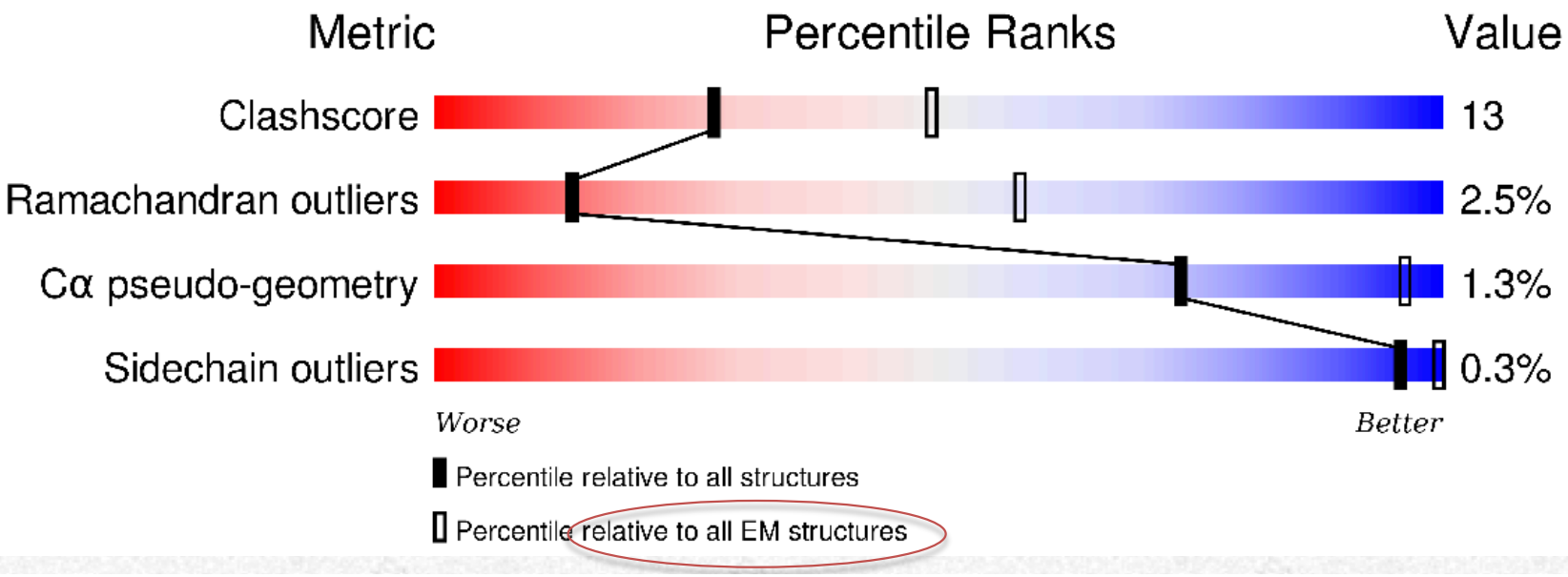


EMDataResource
Unified Data Resource for 3DEM

EM Validation Report

- "Table 1" + EM model metrics
- Comparative statistics updated annually
- Planned improvements: images/statistics

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 (		
Minimum defoc		
Maximum defoc		
Magnification		
Image detector		



Archive Files and Data Dictionaries

- EMDB produces EMDB/xml format files
- PDB produces PDBx/mmCIF files
- Underlying dictionaries are equivalent!

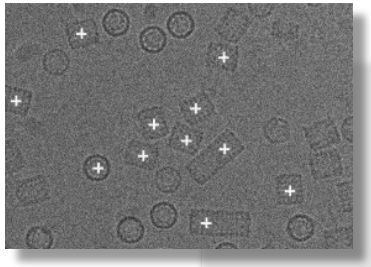
Example: Vitrification Instruments

```
<xs:simpleType name="vitrInstrType">
  <xs:restriction base="xs:string">
    <xs:enumeration value="BAL-TEC HPM 010"/>
    <xs:enumeration value="EMS-002 RAPID IMMERSION FREEZER"/>
    <xs:enumeration value="FEI VITROBOT"/>
    <xs:enumeration value="FEI VITROBOT MARK I"/>
    <xs:enumeration value="FEI VITROBOT MARK II"/>
    <xs:enumeration value="FEI VITROBOT MARK III"/>
    <xs:enumeration value="FEI VITROBOT MARK IV"/>
    <xs:enumeration value="GATAN CRYOPLUNGE 3"/>
    <xs:enumeration value="HOMEMADE PLUNGER"/>
    <xs:enumeration value="LEICA PLUNGER"/>
    <xs:enumeration value="LEICA EM GP"/>
    <xs:enumeration value="LEICA EM CPC"/>
    <xs:enumeration value="LEICA EM HPM100"/>
    <xs:enumeration value="LEICA EM PACT"/>
    <xs:enumeration value="LEICA EM PACT2"/>
    <xs:enumeration value="LEICA KF80"/>
    <xs:enumeration value="NONE"/>
    <xs:enumeration value="REICHERT-JUNG PLUNGER"/>
    <xs:enumeration value="ZEISS PLUNGE FREEZER CRYOBOX"/>
    <xs:enumeration value="OTHER"/>
    <xs:enumeration value="SPOTITON"/>
  </xs:restriction>
```

EMDataResource
Unified Data Resource for 3DEM

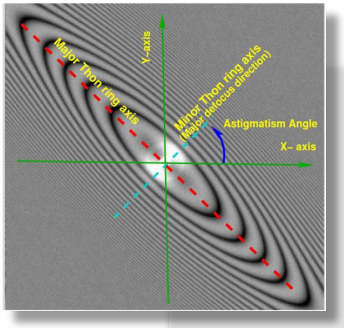
3DEM Validation Challenges

Community Challenges in 3DEM



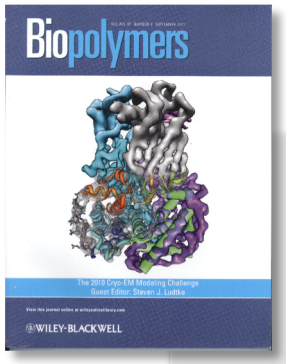
Particle Picking Bakeoff

Zhu et al 2004



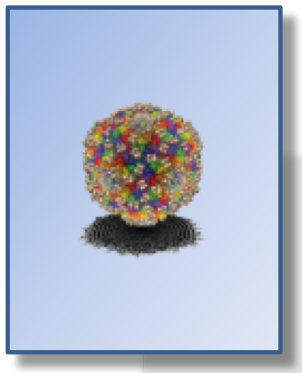
CTF Challenge

Marabini et al 2015



EMDataBank Model Challenge 2010

Biopolymers special issue 2012



EMDataBank Map and Model Challenges 2016

Journal special issue 2018

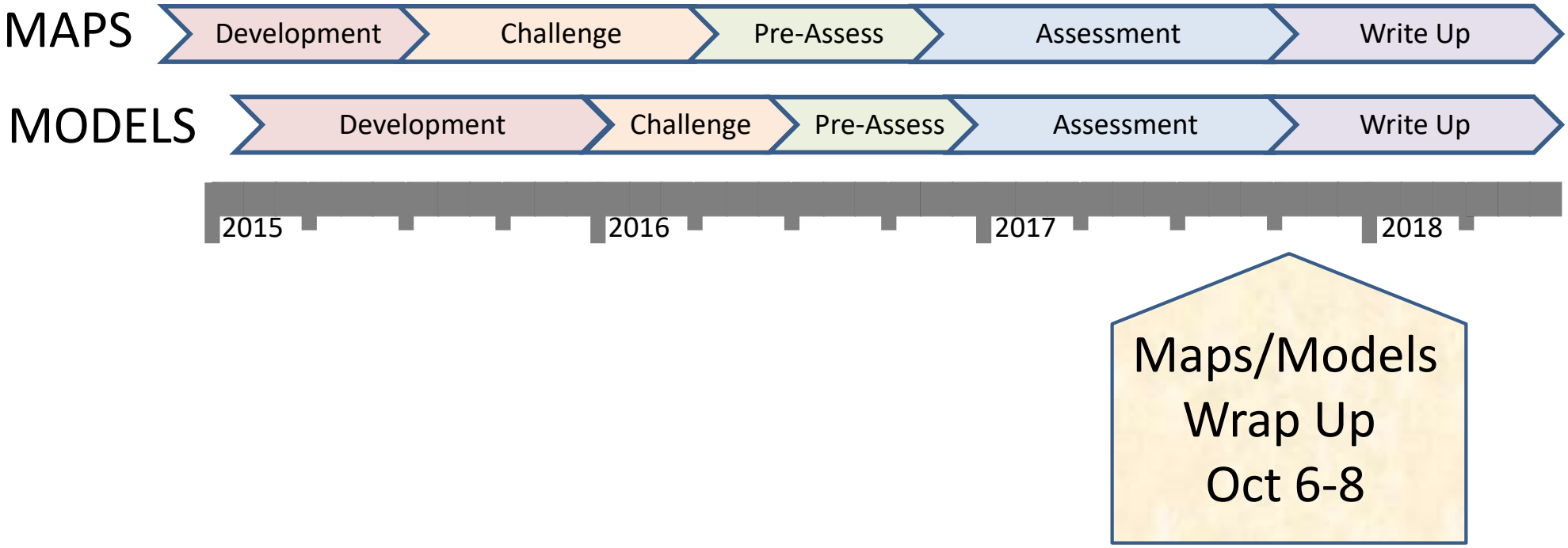
EMDataResource
Unified Data Resource for 3DEM

2015/2016 Map, Model Challenges

- Goals: Develop benchmarks, encourage development of best practices in reconstruction and model fitting, **evolve criteria for validation**, compare and contrast different approaches
- Based on data archived in EMPIAR, EMDB, PDB
- Results discussion via Participant Workshops/Journal Special Issue
- http://challenges.emdatabank.org

Joint Challenges Workshop Oct 6-8, 2017, Stanford/SLAC

The Process



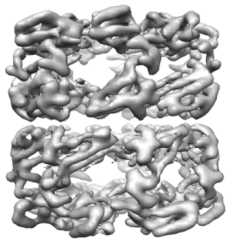
EMDataResource
Unified Data Resource for 3DEM

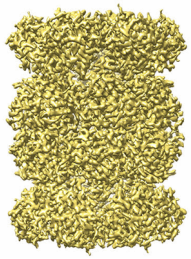
Committee Meetings

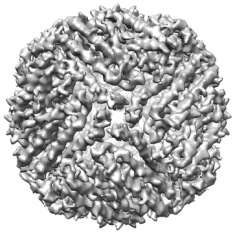
EMDataResource
Unified Data Resource for 3DEM

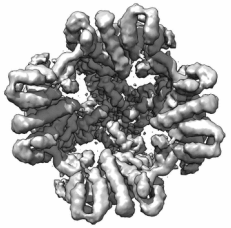
Benchmark Targets

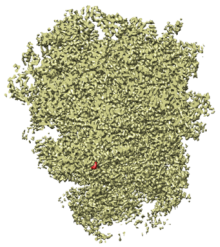
Map Challenge: Raw Images @ EMPIAR

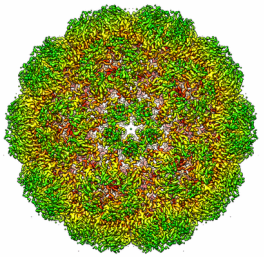

GroEL

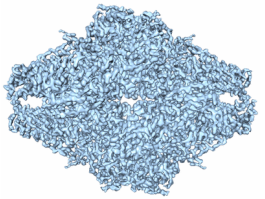

T20S Proteasome


Apo-Ferritin

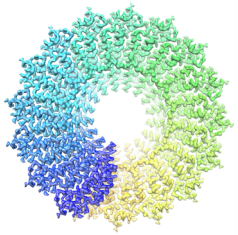

TrpV1 channel

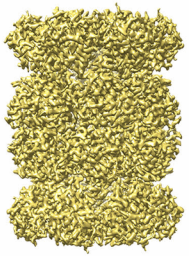

80S Ribosome

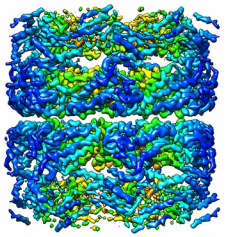

Brome Mosaic Virus

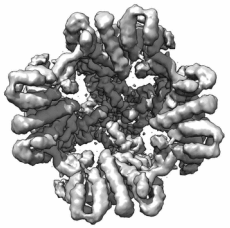

β-galacto-sidase

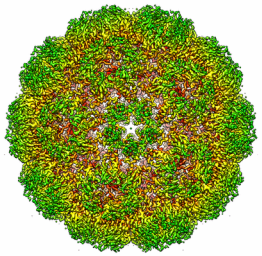
Model Challenge: Maps @ EMDB

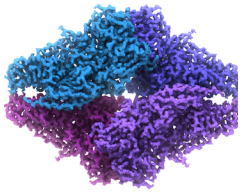

Tobacco Mosaic Virus

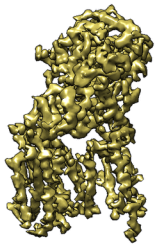

T20S Proteasome

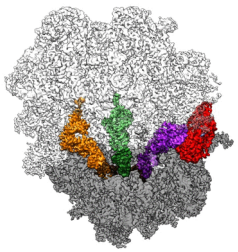

GroEL


TrpV1 channel


Brome Mosaic Virus


β-galacto-sidase


γ-Secretase


70S Ribosome

EMDataResource
Unified Data Resource for 3DEM

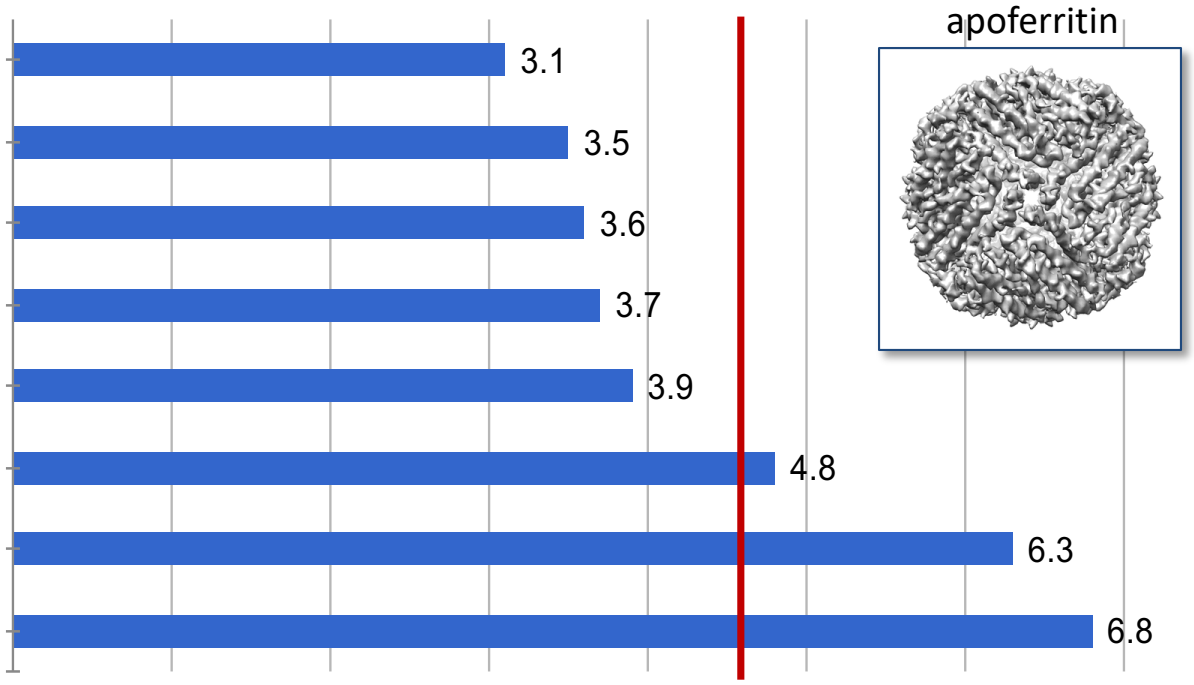
Challenger Locations

In total: more than 80 participants from 3DEM and modelling communities

EMDataResource
Unified Data Resource for 3DEM

Map Challenge: Apoferritin Target

Reported resolution distribution of submitted maps
Red line: resolution reported in original study

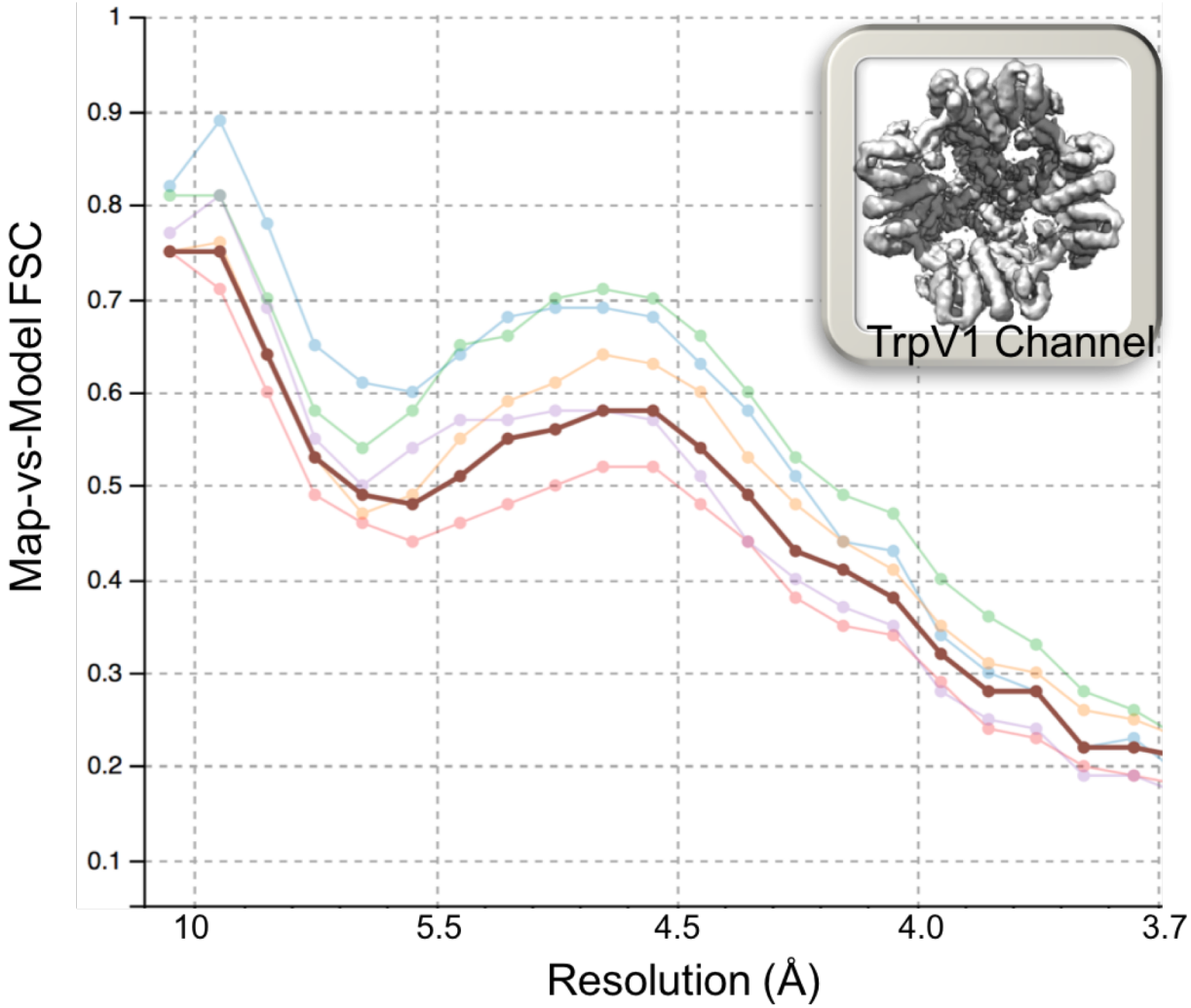


EMDataResource
Unified Data Resource for 3DEM

Challenges Wrap-Up: Maps

- Results
  - All major reconstruction packages produced maps of equivalent quality.
  - However .... quality could vary considerably between different practitioners.
  - Reported resolution was not a reliable indicator of resolvability.
- Conclusions
  - Current (FSC) practices are inconsistent.
  - Bullet-proof reconstruction workflows, best-practice standards for post-reconstruction processing, and FSC-based resolution evaluation are needed.

Model Challenge: TrpV1 Target



http://model-compare.emdatabank.org

EMDataResource
Unified Data Resource for 3DEM

Challenges Wrap-Up: Models

- Results
  - Challengers were able to correctly trace significant portions of the benchmarks, in some cases making substantive improvements.
- Conclusions
  - Further review of global fit metrics (e.g., Map-Model FSC, correlation coefficients) is needed to determine which combinations are most useful.
  - Residue-level metrics that properly account for electron scattering properties of charged residues are needed.
  - Model-based metrics may be useful to analyse map resolvability.

Questions/Comments:
help@emdataresource.org