EM Data Archives & EMDataBank Challenges

Cathy Lawson

NYSBC CryoEM Course

March 28, 2016

EMDataBank
Unified Data Resource for 3DEM

EM Archives for Structural Biology Data

PDB – managed by wwPDB

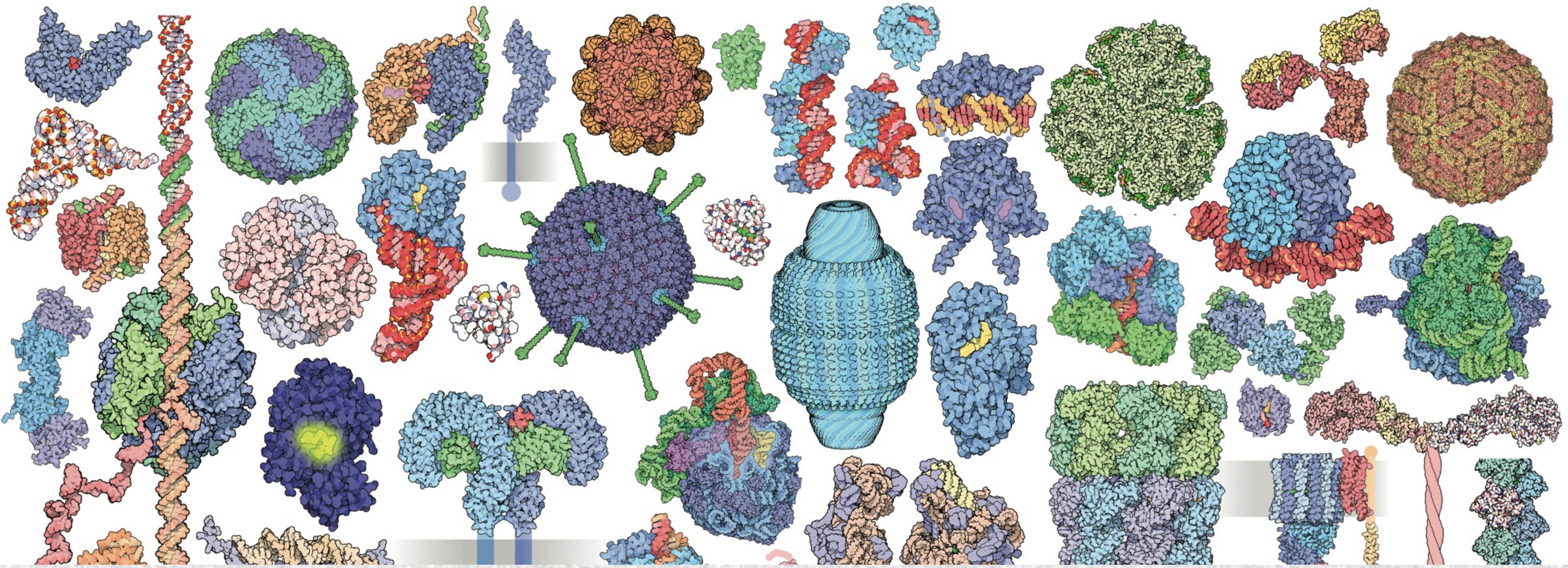
EMDB – managed by EMDataBank

EMPIAR – managed by PDBe

EMDataBank
Unified Data Resource for 3DEM

Protein Data Bank (PDB)

- Established in 1971 with 7 entries
  - Single global archive of 3-D macromolecular structures (>117,000 entries)
  - 1990s: First EM structures deposited



Worldwide Protein Data Bank

- Four Data Centers/Partners
  - RCSB PDB (Research Collaboratory for Structural Bioinformatics)
  - PDBj (Osaka University)
  - PDBe (EMBL-EBI)
  - BioMagResBank (University Wisconsin, Madison)
- Governing agreement
- Ensures data are freely available
- Formalized procedures for Data In: Deposition, Annotation, Representation, and Validation
- Each Data Center provides unique Data Out services

EM Data Bank (EMDB)

- 2002: **EM Data Bank (EMDB)** map archive est. at EBI
- 2004-5: Development workshops with EM community: call for "one-stop shop" for maps and models
- 2006: Proposal to NIH
- 2007: **EMDataBank Unified Data Resource** funded
- 2010: EM Validation Task Force and 1st Model Challenge
- 2013: NIH funding renewed
- 2015-6: New Map and Model Challenges

EMDataBank
Unified Data Resource for 3DEM

EMDataBank
Unified Data Resource

- Unified 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

NCMI National Center for Macromolecular Imaging

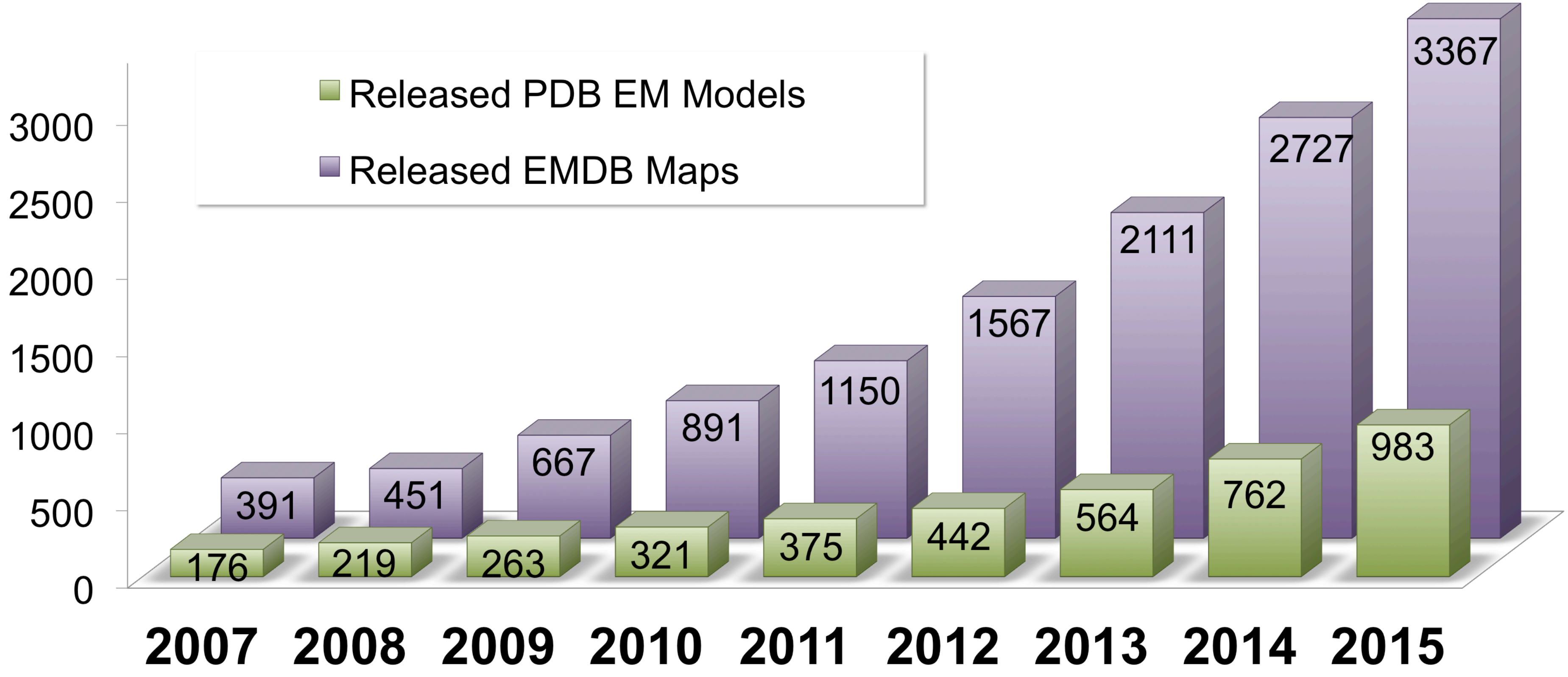
RCSB PDB PROTEIN DATA BANK

PDBe Protein Data Bank in Europe

Supported by NIH National Institute of General Medical Sciences

EMDataBank Unified Data Resource for 3DEM

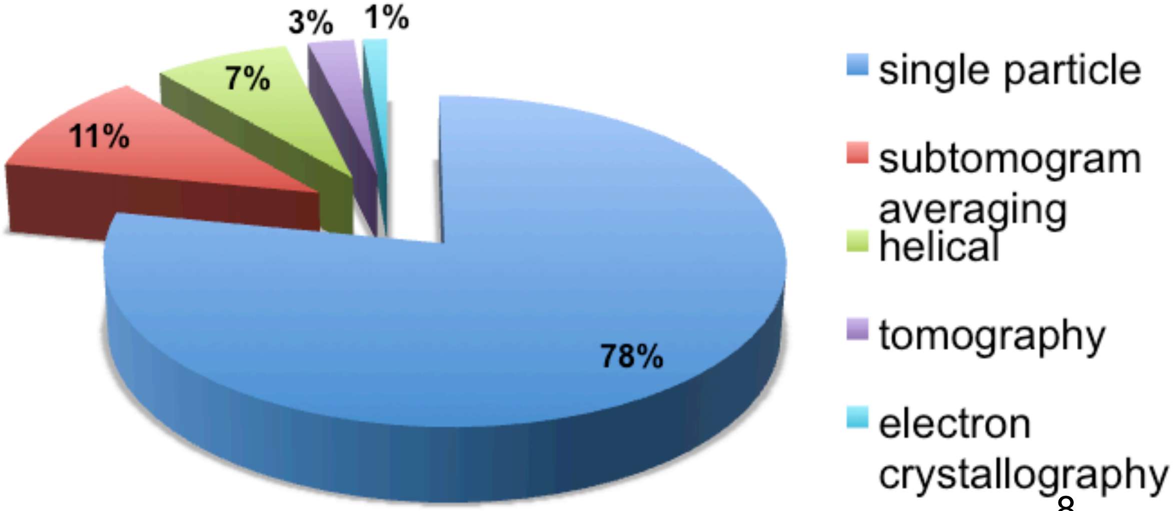
Growth of EM Archives



EMDataBank
Unified Data Resource for 3DEM

EMDB Content

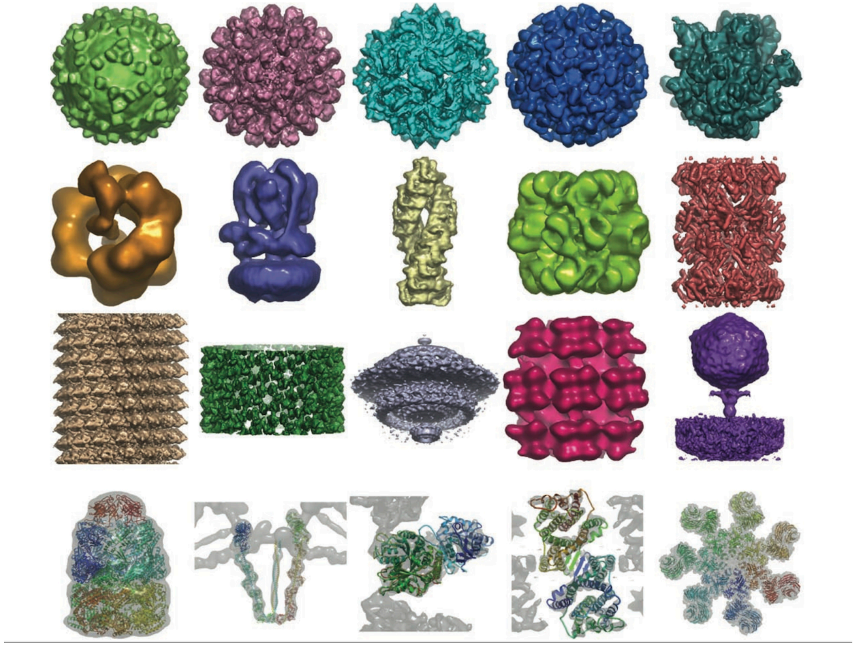
- Archived maps range from macromolecular complexes to cellular tomograms
- Broad resolution range (100-2Å)
- ~1/3 of maps have coordinates



EMDataBank
Unified Data Resource for 3DEM

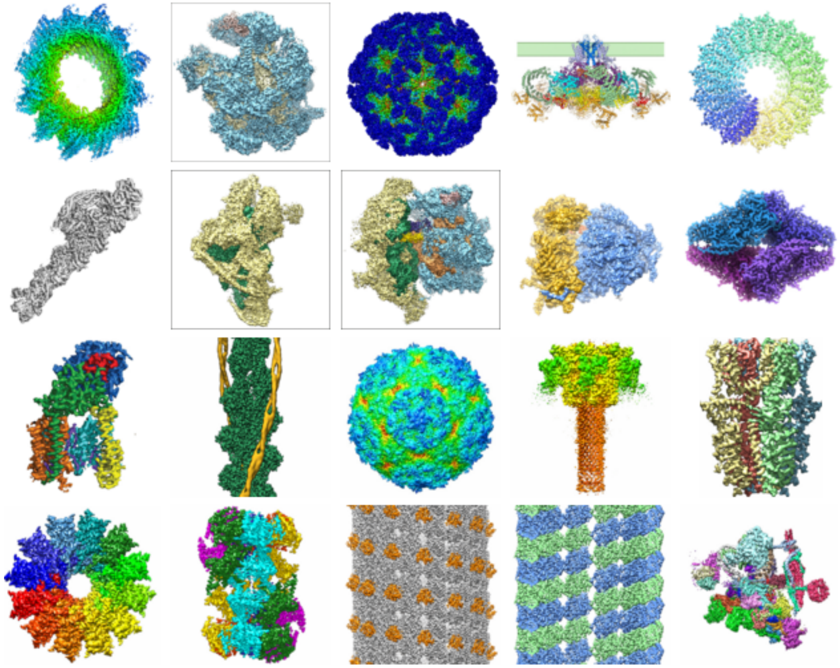
EM Structures 2010 vs 2015

2010: Molecular Shapes



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

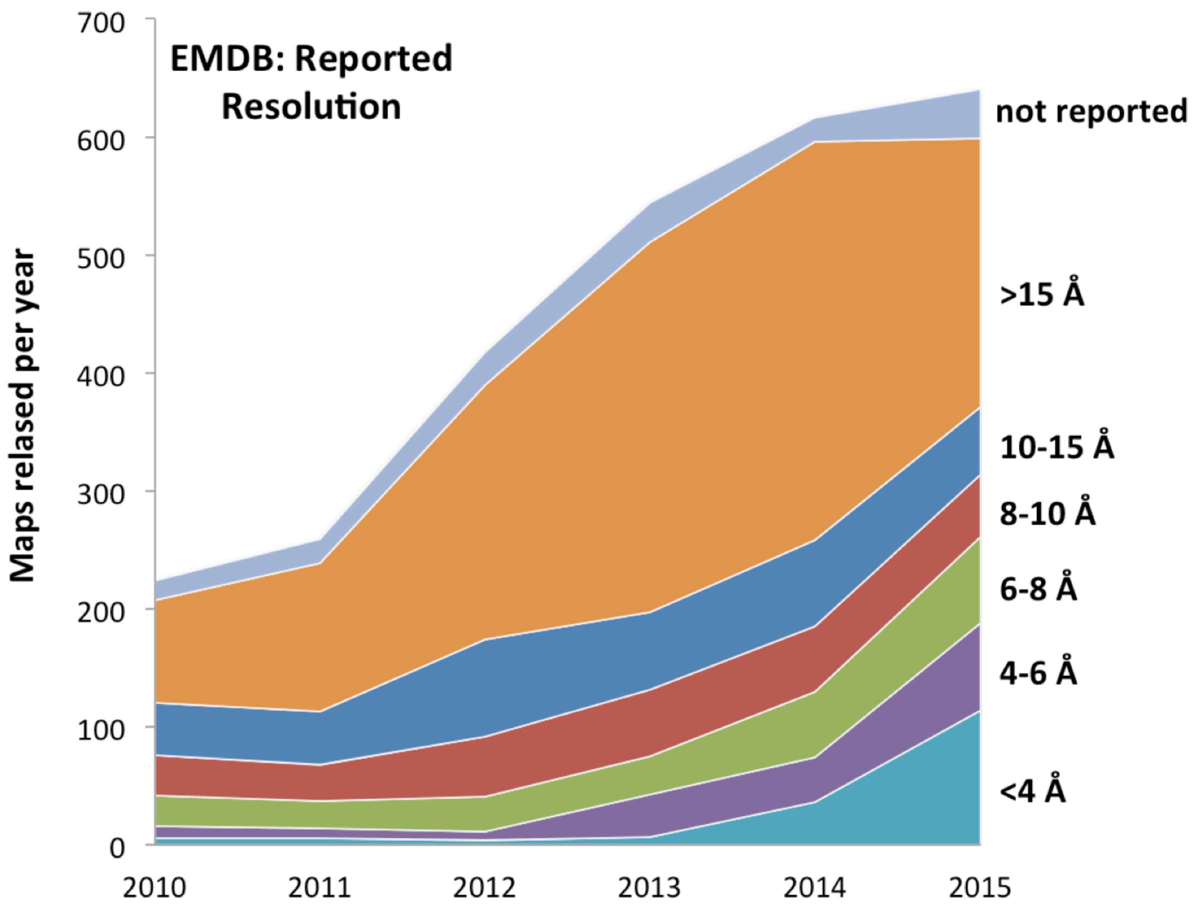
2015: Traceable Densities



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

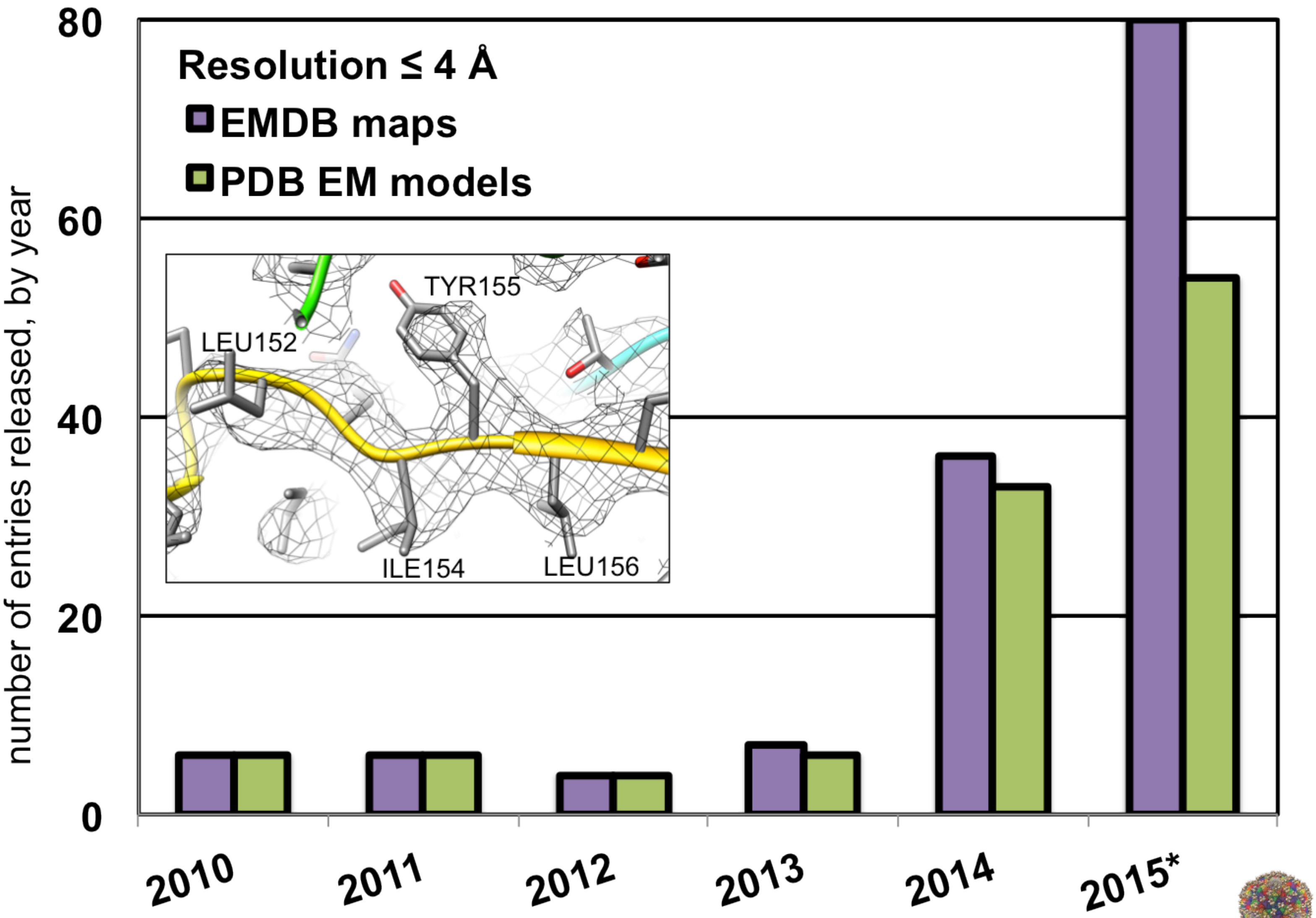
EMDataBank
Unified Data Resource for 3DEM

EMDB Map entries vs Resolution



EMDataBank
Unified Data Resource for 3DEM

EM Structures @ 4 Å or better



EMDataBank

Unified Data Resource for 3DEM

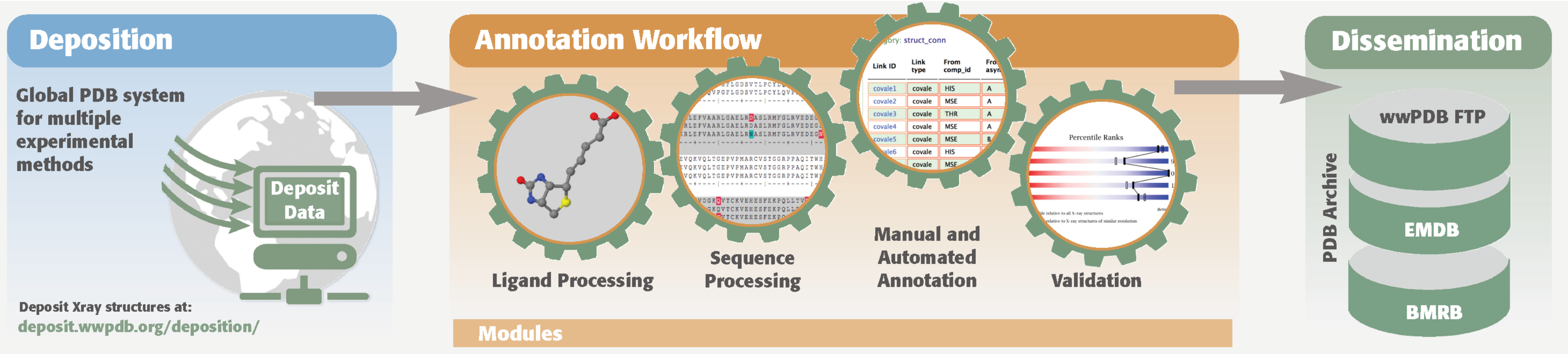
3DEM Structure Deposition

Development: 2004 CryoEM Workshop

- 30 Attendees including cryo-EM, programming and database experts, funding agency and journal representatives
- Recommendations:
  - EM data dictionary
  - One-stop-shop

wwPDB Deposition & Annotation System

- New Version for depositing structures from X-ray, NMR, and EM Launched January 2016



wwPDB Deposition & Annotation System

- Standard file format based on a controlled data dictionary (mmCIF/PDBx)
- Provides support for larger and more complex structures
- Improves efficiency for data capture through automation and validation
- Balances workload internationally based on resource capacity and location

wwPDB D&A: EM Deposition

- “one stop shop” realized: deposit Model and associated Map in the same session
  - Map is assigned an EMDB id
  - Model is assigned a PDB id
- Expanded data dictionary for EM
- Provision for uploading half-maps, FSC curves
- 3DEM validation reports for models (in future to include map-model fit)

- Old systems for maps and models (EMDEP, EM-ADIT, AUTODEP) will be retired this year

EM Dictionary Categories 2016

High Level
em_experiment
em_software

Sample
em_entity_assembly
em_entity_assembly_molwt
em_entity_assembly_naturalsource
em_entity_assembly_recombinant
em_virus_entity
em_virus_natural_host
em_virus_shell

Imaging
em_diffraction
em_diffraction_shell
em_diffraction_stats
em_image_recording
em_image_scans em_imaging
em_imaging_optics

Specimen
em_buffer
em_buffer_component
em_crystal_formation
em_embedding
em_sample_support
em_specimen
em_staining
em_vitrification
em_fiducial_markers
em_focused_ion_beam
em_grid_pretreatment
em_high_pressure_freezing
em_shadowing
em_support_film
em_tomography
em_tomography_specimen
em_ultramicrotomy

Reconstruction
em_3d_reconstruction
em_image_processing
em_particle_selection
em_volume_selection
em_ctf_correction
em_euler_angle_assignment
em_final_classifiation
em_start_model

Symmetry
em_2d_crystal_entity
em_3d_crystal_entity
em_helical_entity
em_single_particle_entity

Fitting
em_3d_fitting
em_3d_fitting_list

EMDataBank
Unified Data Resource for 3DEM

Experiment: EM Imaging

Parent/Child Relationships

entry_id*
specimen_id*

Equipment & Basic Settings (enumerations)

cryogen
electron_source*
illumination_mode*
microscope_model*
mode*
specimen_holder_model

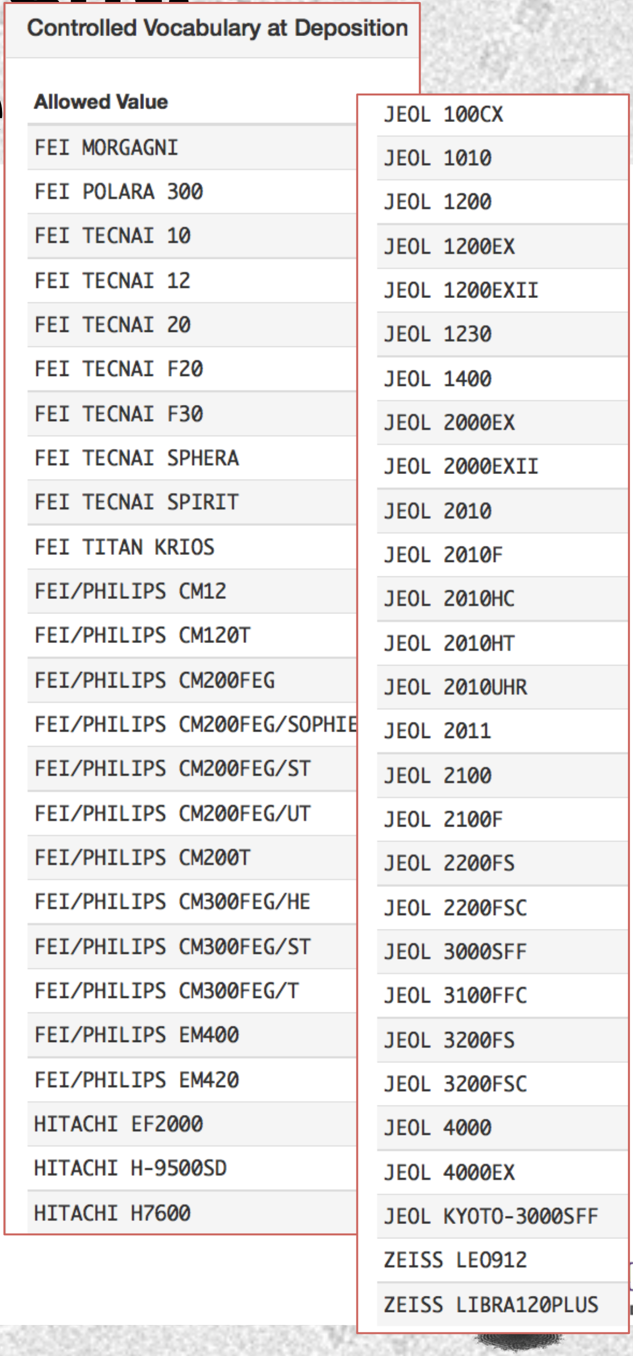
Parameters (Units, with value limits)

accelerating_voltage (kV)*
c2_aperture_diameter (mm)
nominal_cs (mm)
nominal_defocus_max (nm)
nominal_defocus_min (nm)
nominal_magnification (fold x)
recording_temperature_maximum (^{o}K)
recording_temperature_minimum (^{o}K)
tilt_angle_max (^{o})
tilt_angle_min (^{o})

*mandatory data item

Experiment: Microscope

- mandatory data item
- controlled vocabulary
- input from
  - 3DEM experts
  - microscope manufacturers

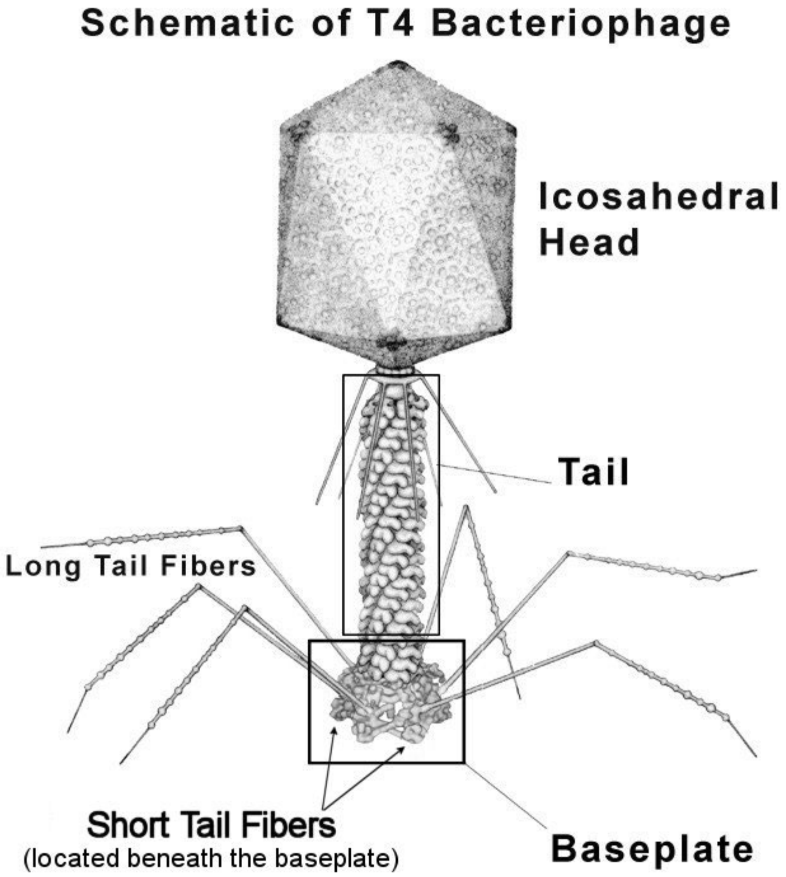


Controlled Vocabulary at Deposition

Allowed Value
FEI MORGAGNI
FEI POLARA 300
FEI TECNAI 10
FEI TECNAI 12
FEI TECNAI 20
FEI TECNAI F20
FEI TECNAI F30
FEI TECNAI SPHERA
FEI TECNAI SPIRIT
FEI TITAN KRIOS
FEI/PHILIPS CM12
FEI/PHILIPS CM120T
FEI/PHILIPS CM200FEG
FEI/PHILIPS CM200FEG/SOPHIE
FEI/PHILIPS CM200FEG/ST
FEI/PHILIPS CM200FEG/UT
FEI/PHILIPS CM200T
FEI/PHILIPS CM300FEG/HE
FEI/PHILIPS CM300FEG/ST
FEI/PHILIPS CM300FEG/T
FEI/PHILIPS EM400
FEI/PHILIPS EM420
HITACHI EF2000
HITACHI H-9500SD
HITACHI H7600

JEOL 100CX
JEOL 1010
JEOL 1200
JEOL 1200EX
JEOL 1200EXII
JEOL 1230
JEOL 1400
JEOL 2000EX
JEOL 2000EXII
JEOL 2010
JEOL 2010F
JEOL 2010HC
JEOL 2010HT
JEOL 2010UHR
JEOL 2011
JEOL 2100
JEOL 2100F
JEOL 2200FS
JEOL 2200FSC
JEOL 3000SFF
JEOL 3100FFC
JEOL 3200FS
JEOL 3200FSC
JEOL 4000
JEOL 4000EX
JEOL KYOTO-3000SFF
ZEISS LEO912
ZEISS LIBRA120PLUS

Hierarchical Description



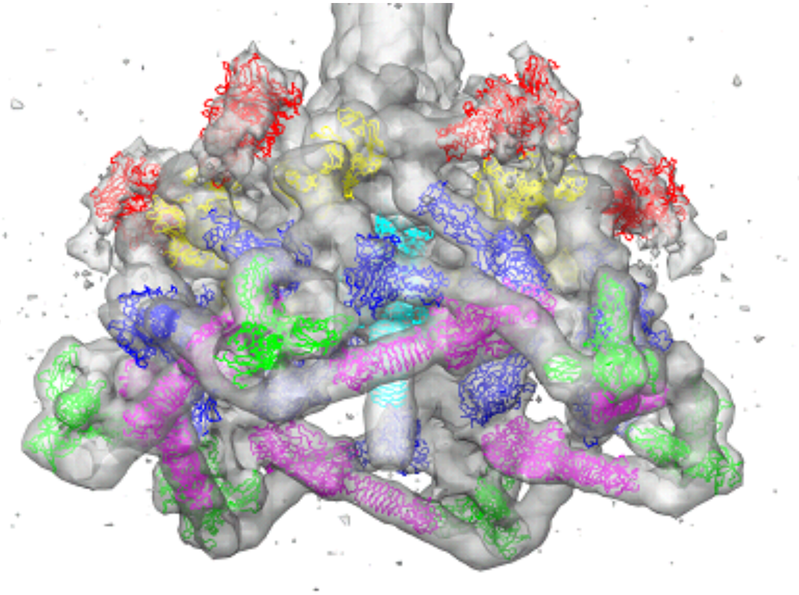
- head
  - capsid
  - genomic DNA
  - portal assembly
- neck
  - neck base
  - collar fibers
  - whisker fibers
- tail
  - sheath
  - tail tube
  - tail terminator
  - cell puncturing device
- baseplate
  - baseplate base
  - short tail fibers
  - long tail fibers

--Yap & Rossmann (2014) Future Microbiol 12, 1319-27

EMDataBank
Unified Data Resource for 3DEM

Assembly-Polymer Linkage

- T4 Baseplate:
  - gp11, gp10, gp8, gp6, gp25, gp9, gp5, gp27

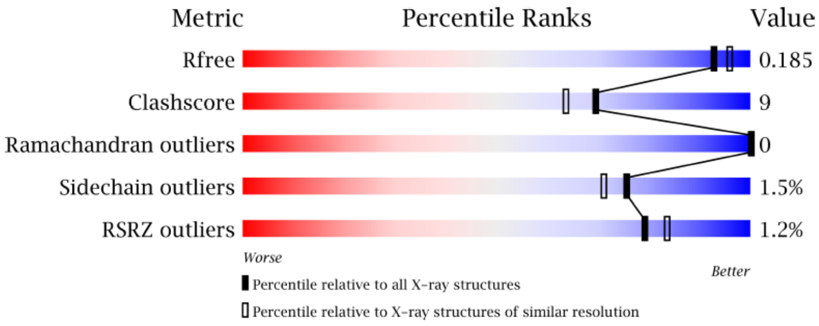


EMDataBank
Unified Data Resource for 3DEM

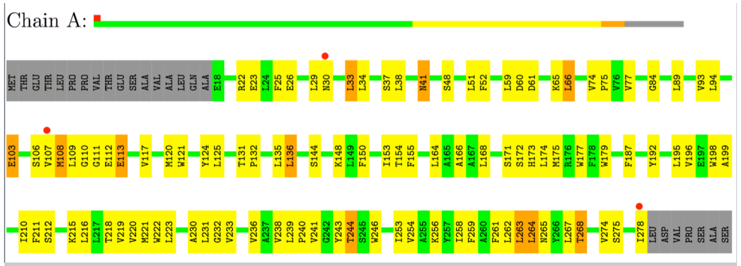
Validation Report: X-ray

- Overall quality at-a-glance
- "Table 1" with key data & refinement statistics
- Component diagnostics for all macromolecules & ligands
- Depositor also receives detailed XML report
- PDF can be uploaded with manuscript submission to a journal

Overall Quality



Residue Plots



Grey – not modeled

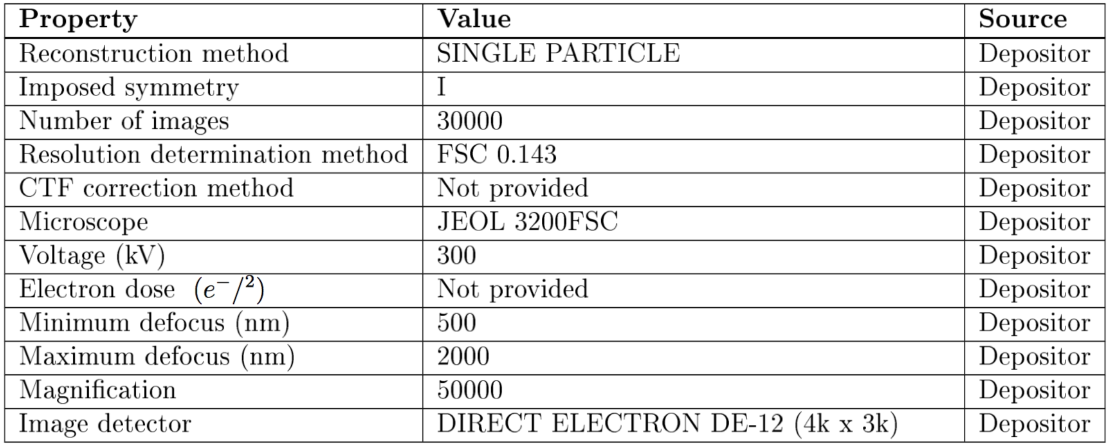
Green, yellow, orange, red – 0,1,2, 3 or more issues

Red dot – poor fit to electron density

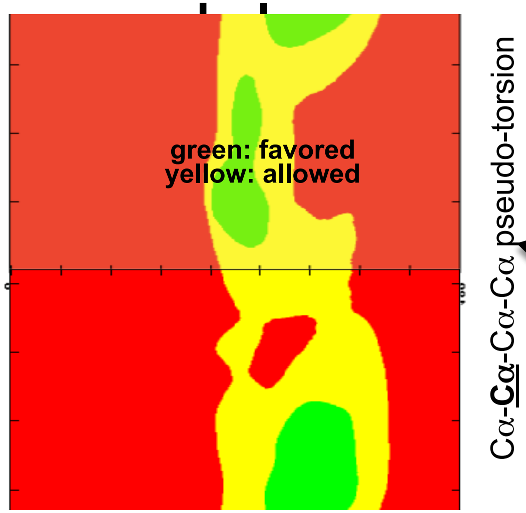
EMDataBank
Unified Data Resource for 3DEM

EM Validation Reports

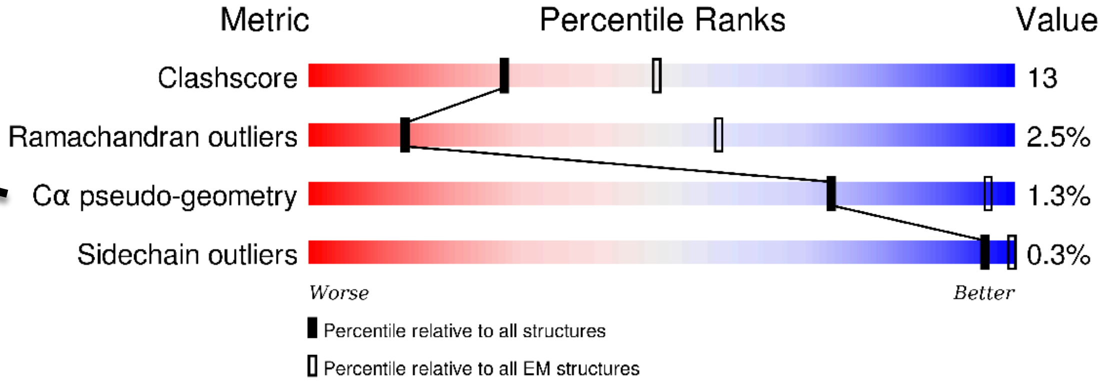
- "Table 1" for EM
- Metrics relevant for EM



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



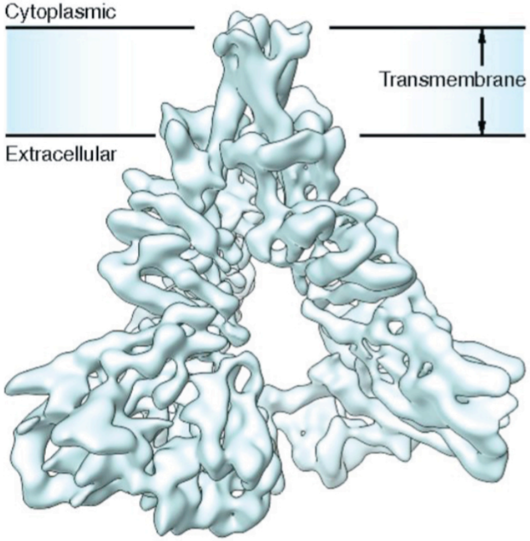
Cα-**Cα**-Cα pseudo-angle



Validation: Map and Model Challenges

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)

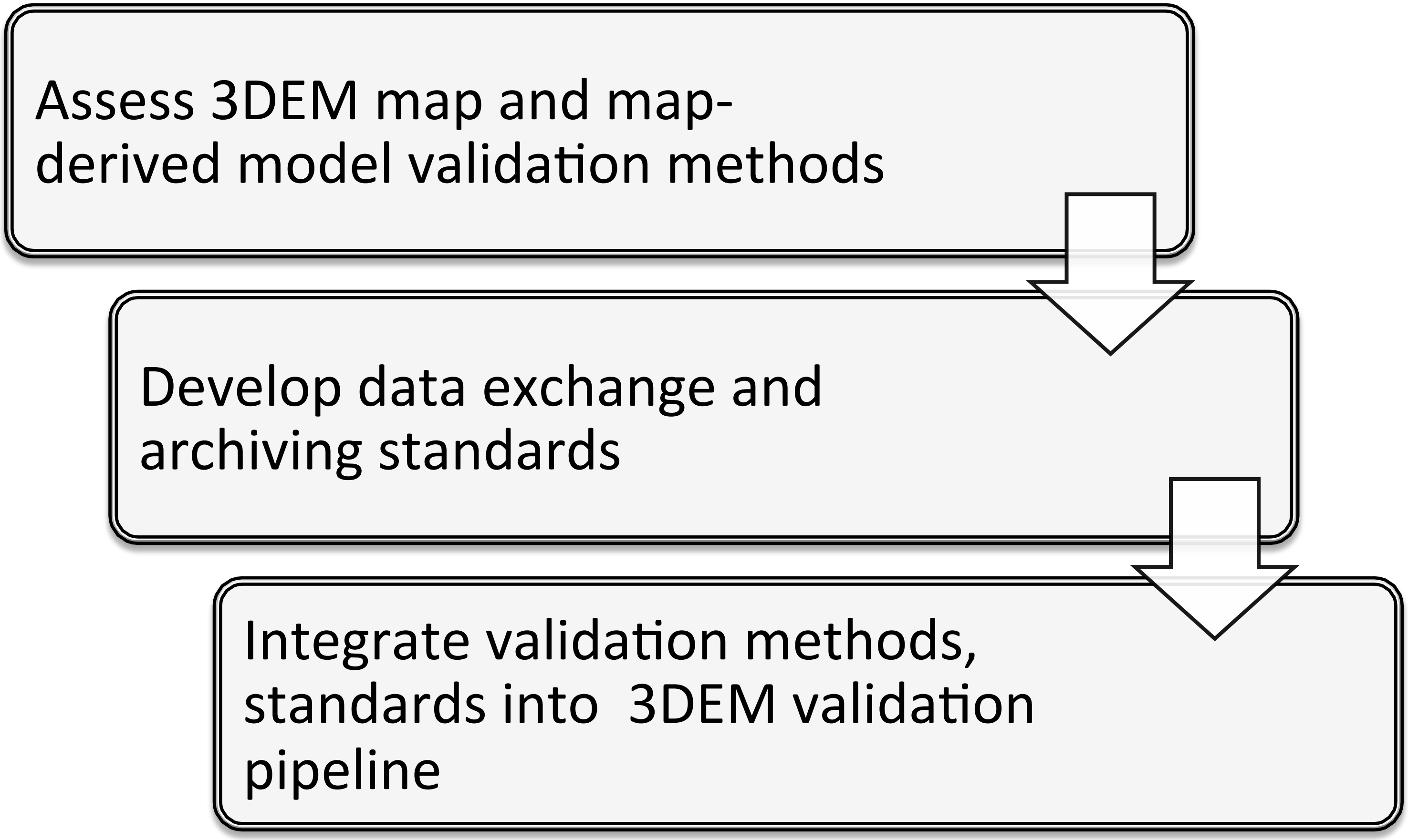


EMD-5418 Y Mao, JG Sodroski *et al*. Molecular architecture of the uncleaved HIV-1 envelope glycoprotein trimer *PNAS* 110, 12438-12443 (2013)

EMDataBank
Unified Data Resource for 3DEM

Validation Development

through research collaborations
with the community, workshops, challenges:



Assess 3DEM map and map-derived model validation methods

Develop data exchange and archiving standards

Integrate validation methods, standards into 3DEM validation pipeline

EMDataBank
Unified Data Resource for 3DEM

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

Unified Data Resource for 3DEM

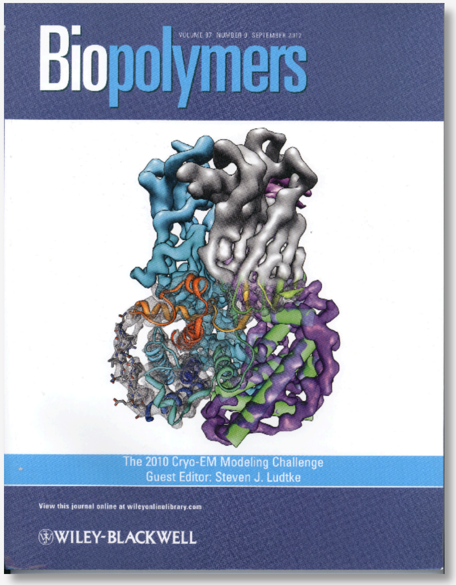
Validation for 3DEM

2010
EM-VTF

EM Validation Task Force
Henderson *et al*. (2012) ***Structure 20***, 205-214

Maps: Standards for assessing resolution and accuracy need to be developed

Models: Criteria needed for model only, fit to map, and fit to additional structural data


Biopolymers
VOLUME 97 NUMBER 9 SEPTEMBER 2012
The 2010 Cryo-EM Modeling Challenge
Guest Editor: Steven J. Ludtke
View this journal online at wileyonlinelibrary.com
WILEY-BLACKWELL

2010 CryoEM Modeling Challenge
Collected papers in a special issue of **Biopolymers** September 2012

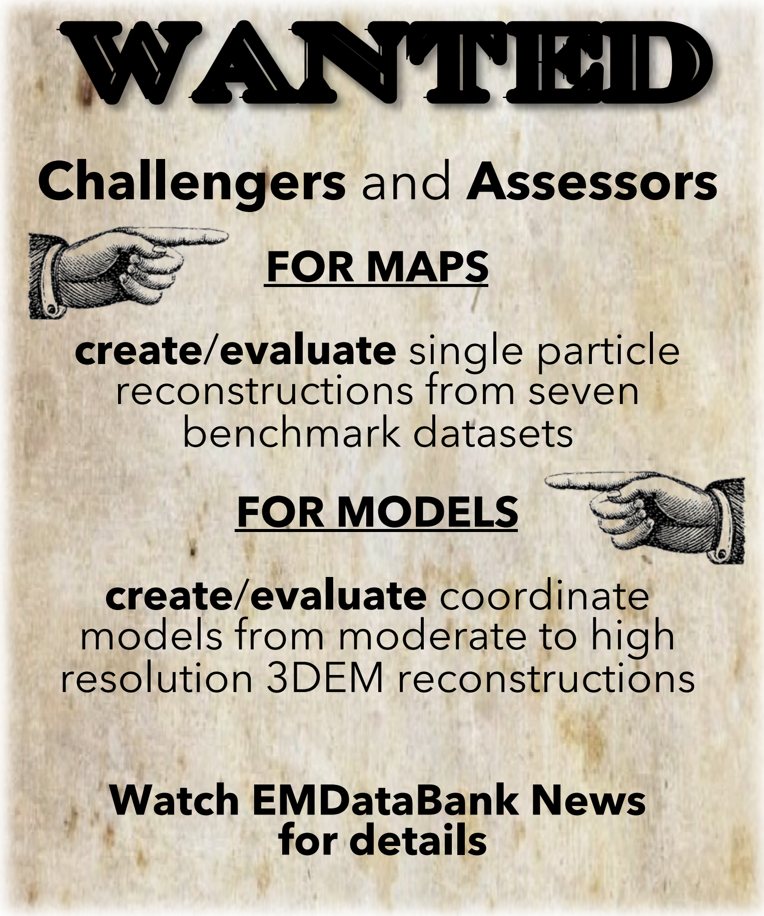
13 target maps
58 participants
10 research groups
136 submitted models
13 software packages

EM VTF Recommendations

- Main recommendations for 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 for models fitted into EM maps
  - Criteria for assessing models need to be developed
  - Capability to archive coarse-grained representations of models is needed
- More research and development needed!

EMDataBank
Unified Data Resource for 3DEM

2015/2016 Map, Model Challenges



WANTED

Challengers and **Assessors**

FOR MAPS

create/evaluate single particle reconstructions from seven benchmark datasets

FOR MODELS

create/evaluate coordinate models from moderate to high resolution 3DEM reconstructions

Watch EMDataBank News for details

- Each challenge formulated by a community-based committee
- Targets selected from recently deposited maps, models, 2.2-4.5 Å resolution

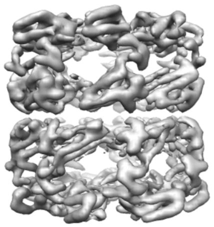
EMDataBank
Unified Data Resource for 3DEM

2015/2016 Map, Model Challenges

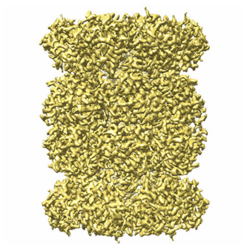
- Goals: Develop benchmarks, encourage development of best practices in 3DEM reconstruction and model fitting, evolve criteria for validation, compare and contrast different approaches
- Results Discussion via Participant Workshops/Journal Special Issues

Benchmark Datasets

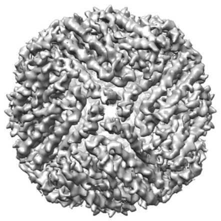
Map Challenge Targets: Raw Images @ EMPIAR



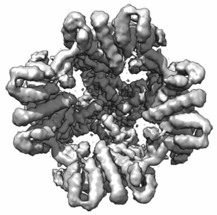
GroEL



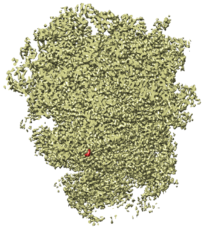
T20S Proteasome



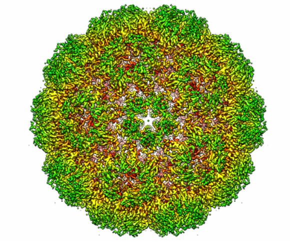
Apo-Ferritin



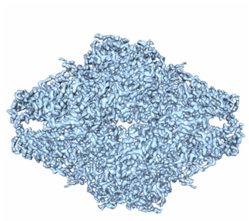
TrpV1 channel



80S Ribosome

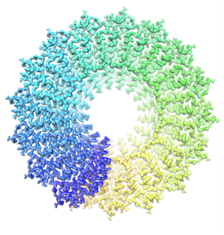


Brome Mosaic Virus

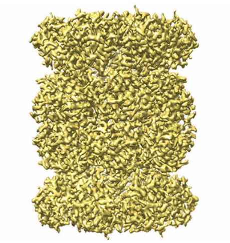


ß-galacto-sidase

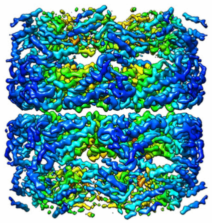
Model Challenge Targets: Maps @ EMDB



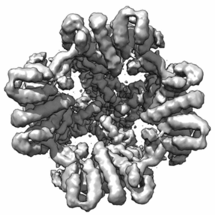
Tobacco Mosaic Virus



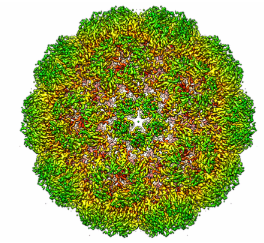
T20S Proteasome



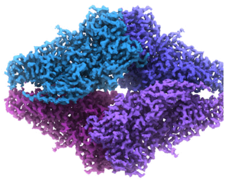
GroEL



TrpV1 channel

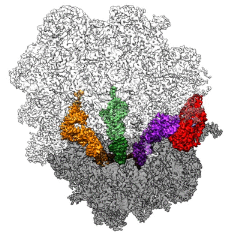


Brome Mosaic Virus



ß-galacto-sidase

γ-Secretase



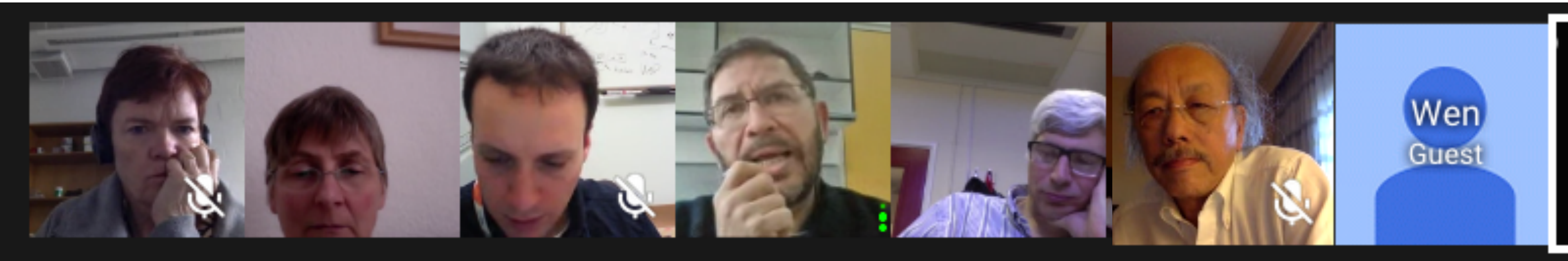
70S Ribosome

EMDataBank
Unified Data Resource for 3DEM

Map Challenge

- **Timing**: **Registration NOW OPEN**
  - Challengers: reconstruction submissions open **August** thru **March 31**
  - Assessors: open data assessment period commences **May 2016**
  - Results Workshop Fall 2016

- **Committee**: Bridget Carragher (Chair), Jose-Maria Carazo, Wen Jiang, John Rubinstein, Peter Rosenthal, Fei Sun, Janet Vonck



Model Challenge

- **Timing**: **Registration NOW OPEN**
  - Challengers: model submissions open **November 2015** thru **April 2016**
  - Assessors: open data assessment period **Summer 2016**
  - Results Workshop **Fall 2016**

- **Committee**: Paul Adams (Chair), Axel Brunger, Randy Read, Torsten Schwede, Maya Topf, Gerard Kleywegt

EMDataBank
Unified Data Resource for 3DEM

wwPDB Hybrid Methods Task Force

EMBL-EBI, Hinxton, UK 6-7 October 2014

Task Force Recommendations

1. Archive Structures, Models, Data/MetaData, and Work Flows
2. Adopt Flexible Structure Representation
3. Assess Structure Uncertainty
4. Federate Structure, Model, and Data /MetaData, and Work Flow Archives
5. Establish Publication Standards

Now in print Sali *et al*. (2015) *Structure 23*, 1156-1167.

Center for Integrative Proteomics Research

- Physical home for structural biology on the Rutgers campus
  - Protein Data Bank
  - Experimental methods: x-ray, NMR, 3DEM....
  - Computational methods

EMDataBank Project Team

NCMI
National Center for Macromolecular Imaging

Baylor College of Medicine

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

RCSB PDB
PROTEIN DATA BANK

Rutgers University

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

PDBe
Protein Data Bank in Europe

EMBL-European Bioinformatics Institute

Gerard Kleywegt, co-PI
Ardan Patwardhan
Eduardo Sanz Garcia
Ingvar Lagerstedt
Matthew Conroy

EMDataBank Advisory Committee
Paul Adams (Chair), Richard Henderson, Bram Koster, Maryanne Martone, Andrej Sali