EMDataResource: Structure Data Archiving, Validation Challenges

Cathy Lawson

EMDataResource & RCSB

Rutgers University

NYSBC NCCAT Single Particle
Short Course, March 5, 2020

EMDataResource
Unified Data Resource for 3DEM

Unified Data Resource for 3DEM

Stanford University/SLAC

RCSB PDB
PROTEIN DATA BANK

Rutgers University

EMDB
Electron Microscopy Data Bank

European Bioinformatics Institute

- Established 2007 under NIH Support (R01GM079429) to:
  - Develop Data Archives for 3DEM (EMDB + PDB)
  - Promote Community Development of Validation and Standards

Project Website

2019-07-10 : 8627 EMDB map entries, 3596 PDB coordinate entries RCSB PDB | PDBe

EMDataResource
Unified Data Resource for 3DEM

Home About Deposit Search Tools Events News Links Help

Search EMDB

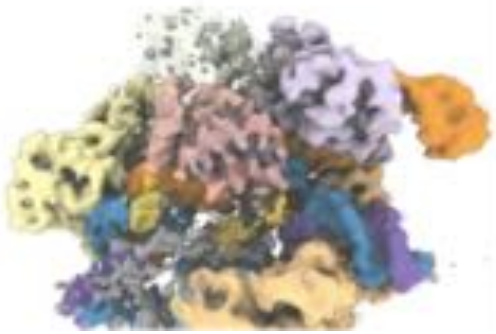
Unified Data Resource for 3DEM

Global resource for 3-Dimensional Electron Microscopy (3DEM) structure data archiving and retrieval, news, events, software tools, data standards, validation methods, and community challenges

Quick link: model metrics challenge

Recently released EMDB entries

All recent entries



EMD-0584
Deposited 2019-02-17
singleParticle 3.37 Å PDB:6o0y
Conformational states of Cas9-sgRNA-DNA ternary complex in the presence of magnesium
Zhu X, Clarke R, Puppala AK, Chittori S, Merk A, Merrill BJ, Simonovic M, Subramaniam S

News

All news

2019 Model Metrics Challenge

EMDataResource is trying something new and has put together a smaller, shorter-timeline EM model challenge. The deadline for submission of completed models fitted to target maps (2-3 Å resolution range) is ~~May 25~~ **May 28 (3 PM US Eastern).**
Read more...

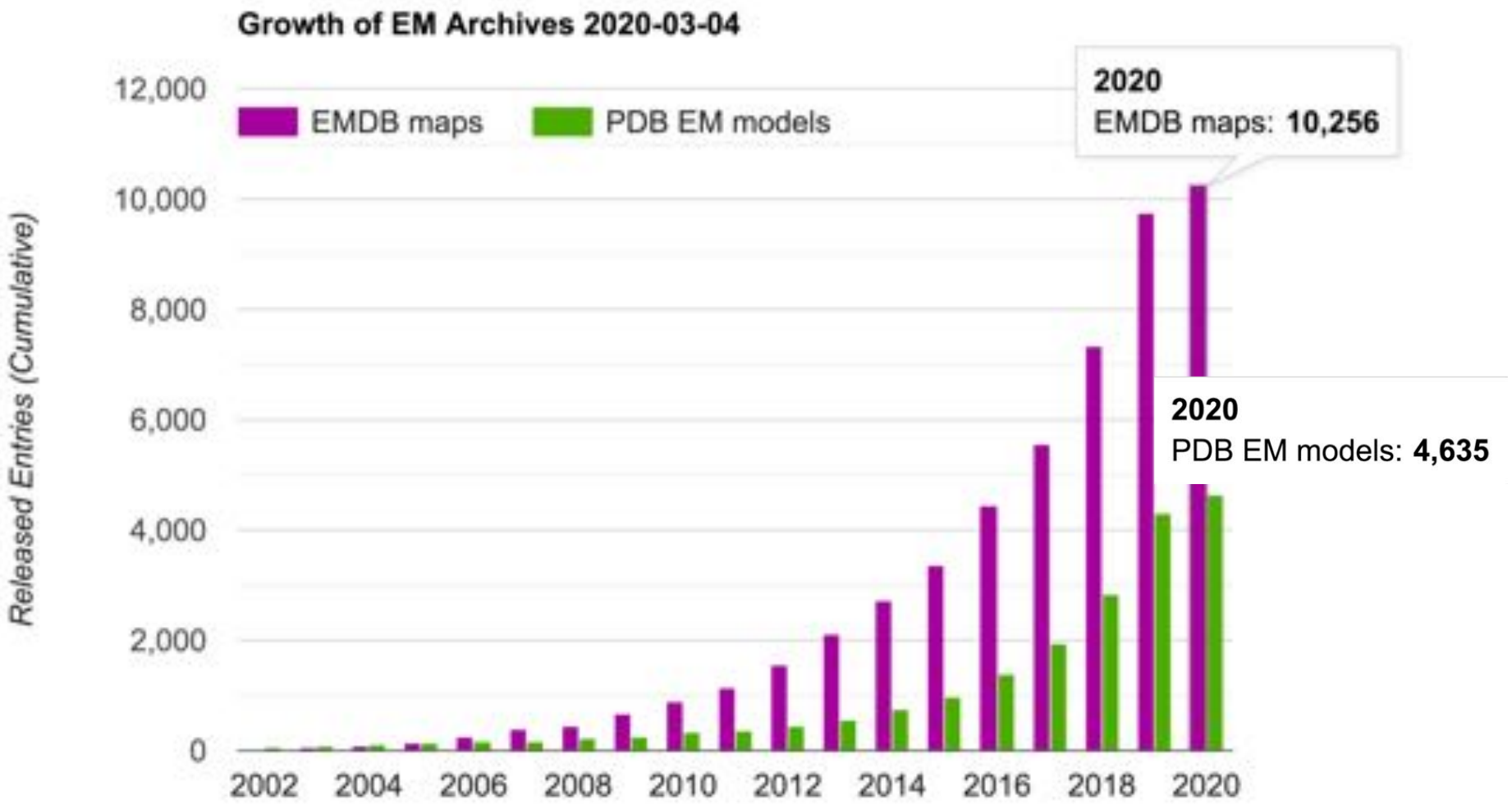
Reminder: EMDB Accession Codes are Changing Soon!

The allocation of 4 digits for EMDB accession codes will soon come to an end. New EMDB accession codes will include an additional digit and will

emdataresource.org

EMDataResource
Unified Data Resource for 3DEM

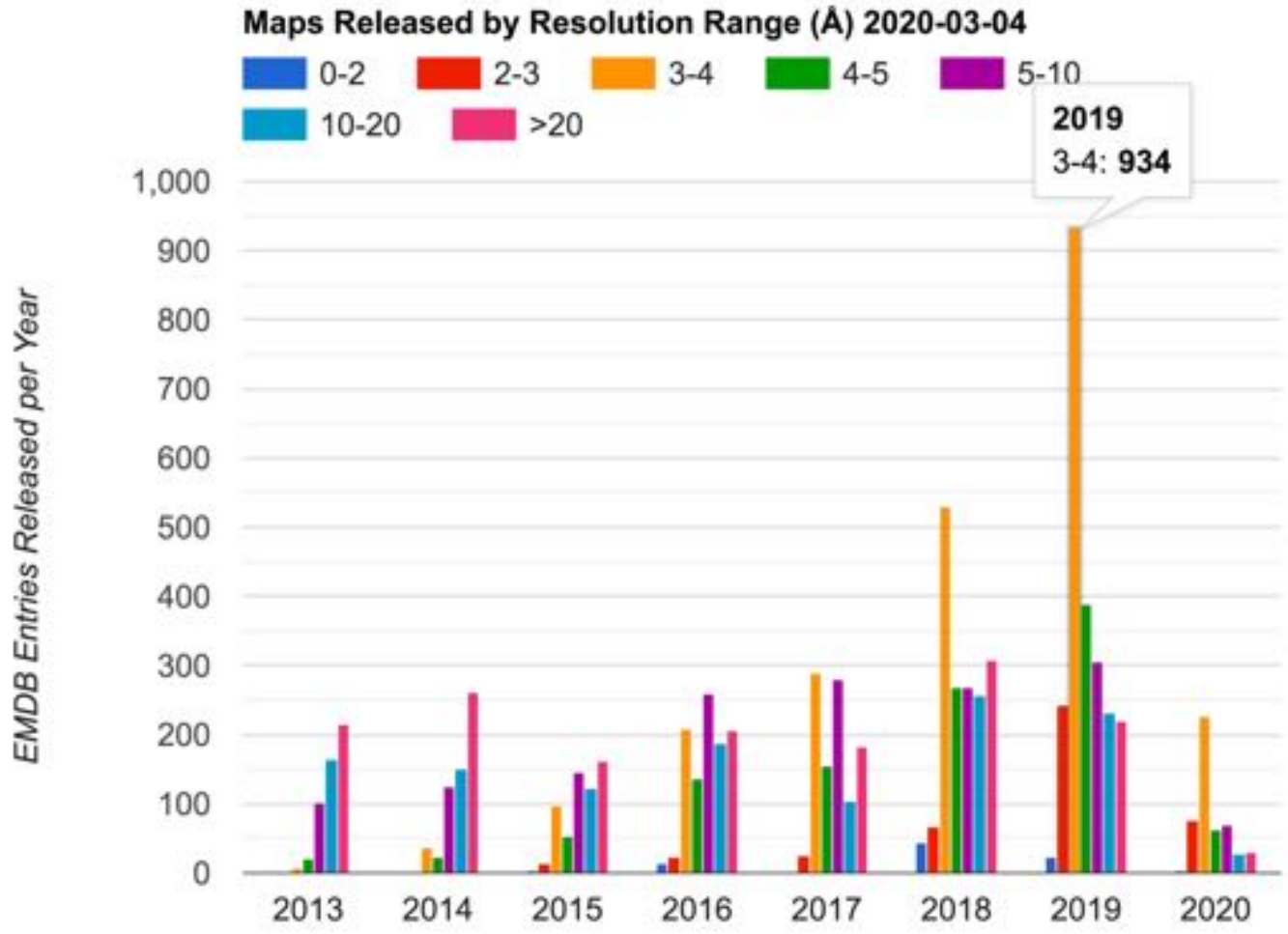
Growth of EM Archives



updated weekly: emdataresource.org/statistics.html

EMDataResource
Unified Data Resource for 3DEM

EMDB maps by year and resolution



updated weekly: emdataresource.org/statistics.html

EMDataResource
Unified Data Resource for 3DEM

Finding Cryo-EM Structures

emdataresource.org/search.html

EMDB maps

- EMDR Search
- Advanced Search @ EBI

PDB EM models

- RCSB-PDB
- PDBe

EMDB + PDB

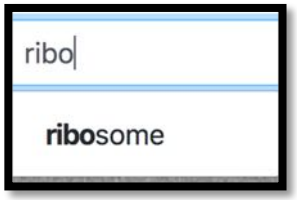
- EM Navigator @ PDBj
- 3DBioNotes @ CSIC

Raw image data

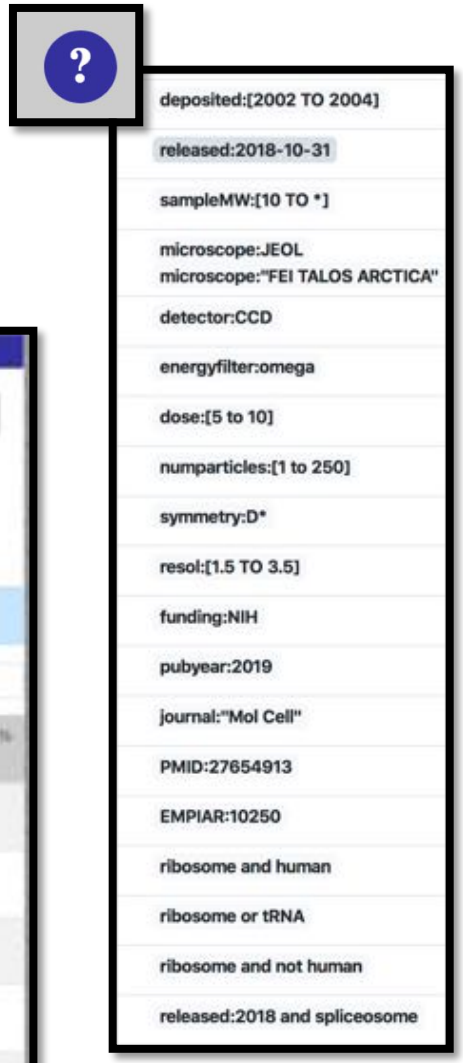
- EMPIAR @ EBI

EMDR Search: Features

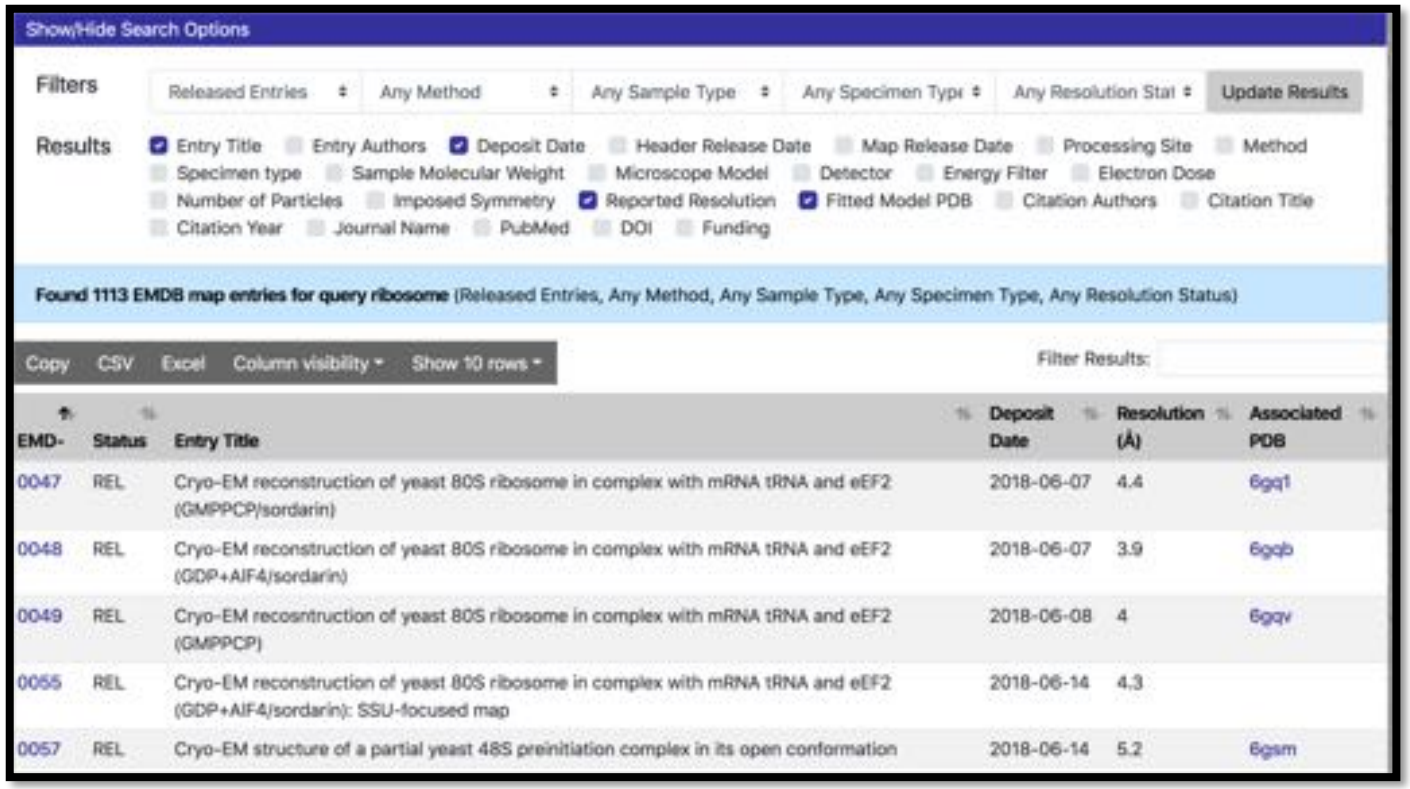
Autocomplete:



Interactive keyword help:



Sort, configure, filter, download search results:



EMDataResource
Unified Data Resource for 3DEM

EMDR Search: Demo

emdataresource.org/solrsearch.html

What EMDB map entries would you like to find?

EMDR Search: Programmatic Access

- https://www.emdataresource.org/node/solr/emd/select?q=5127&fl=id,PDB&rows=10 retrieves:

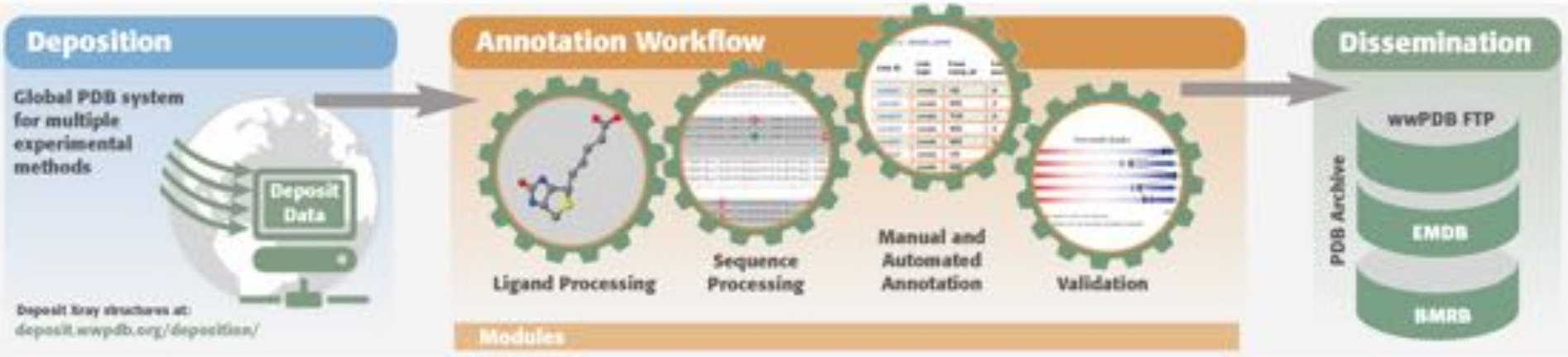
- {"responseHeader":{ "status":0, "QTime":3, "params":{ "q":"5127", "fl":"id,PDB", "rows":"10"}}, "response":{"numFound":1,"start":0,"docs":[{ "PDB":["3iyd"], "id":"5127"}] }}

Cryo-EM Structure Deposition

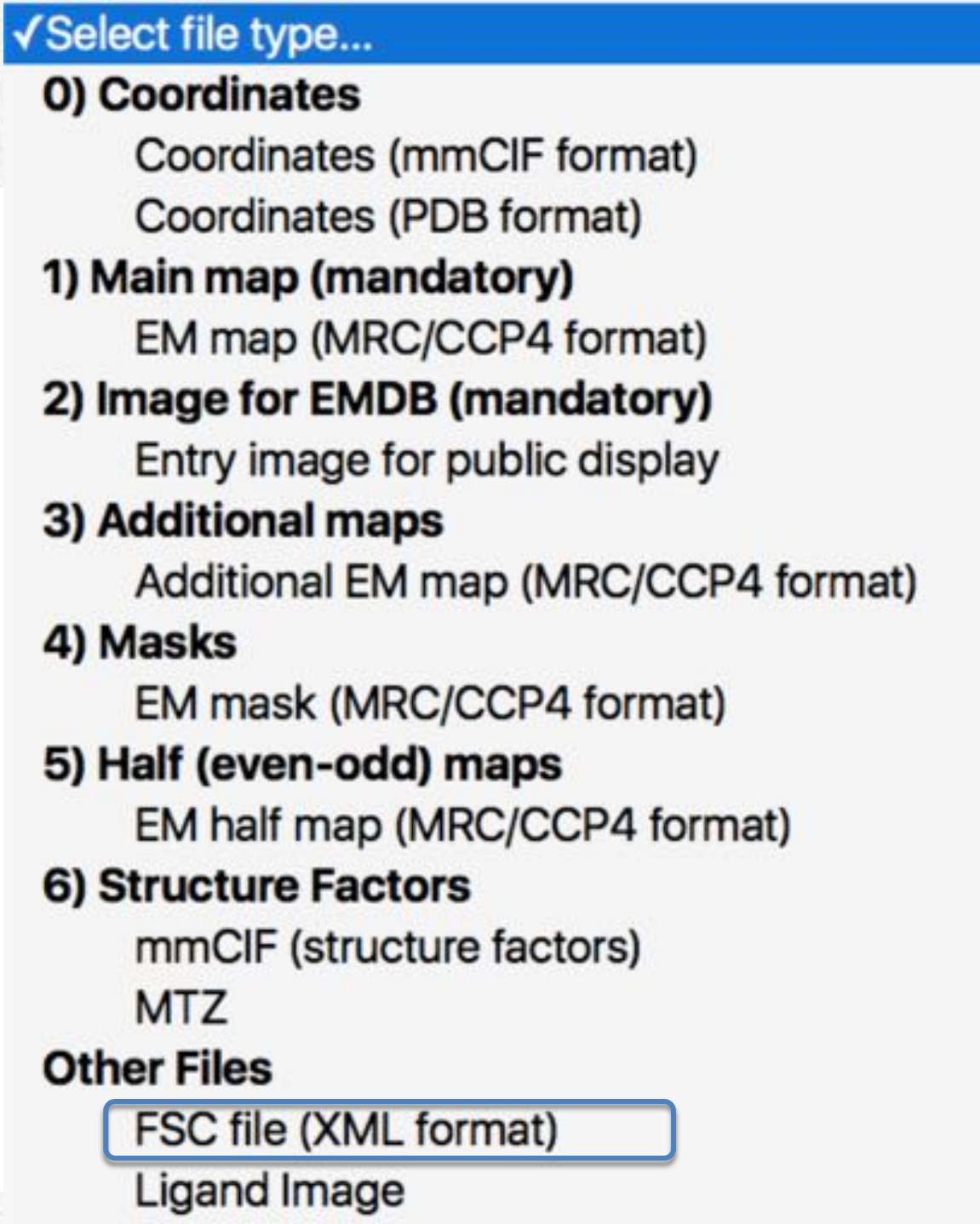
EMDB, PDB

wwPDB OneDep System

- Deposition system for X-ray, NMR, and EM Structures
- EM Depositions:
  - Map to EMDB
  - Coordinate model to PDB
- Validation report is produced

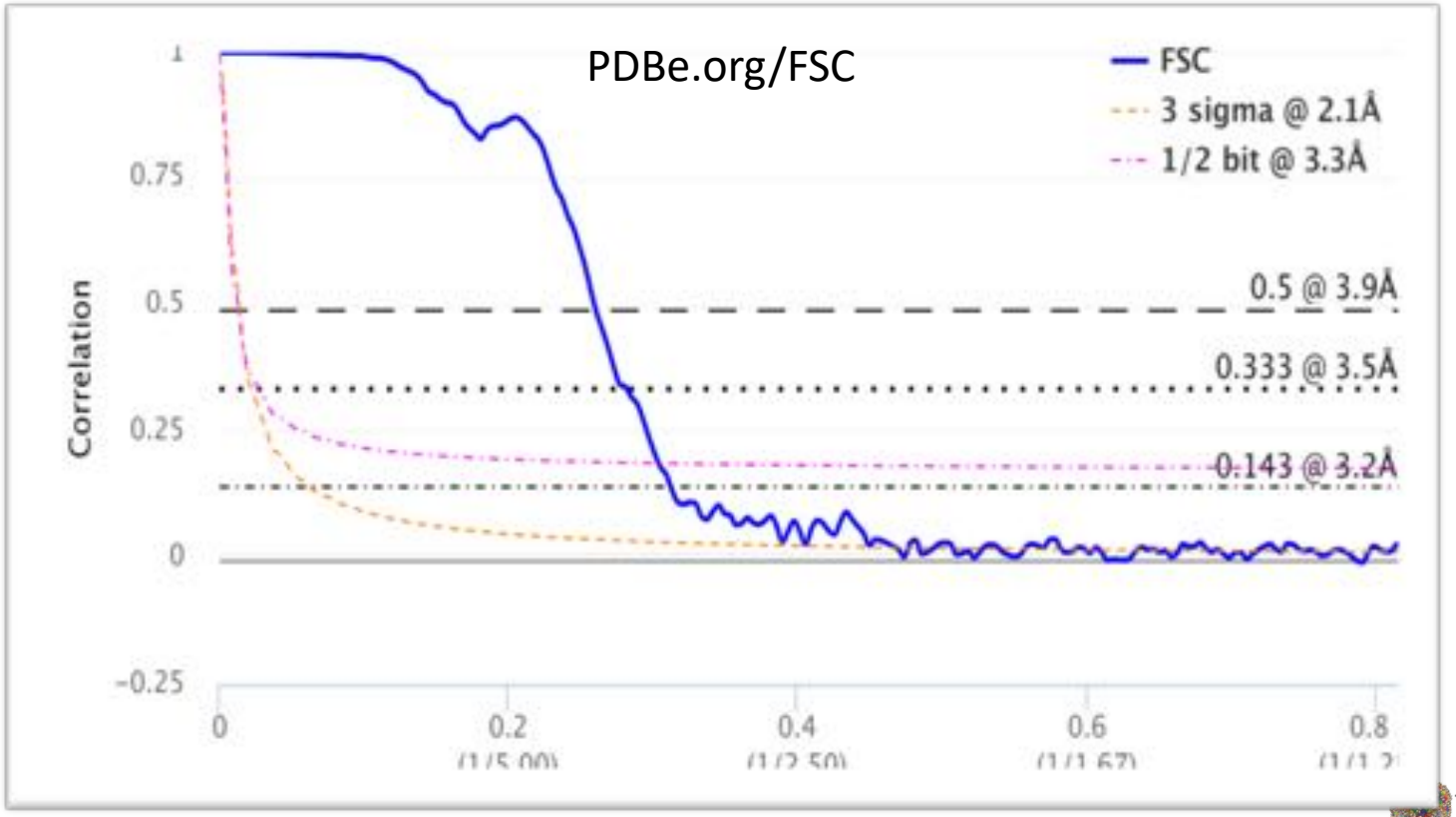


File uploads for EM deposition



FSC Curve

- Upload XML format file
- Create XML via a software package (e.g., Relion, EMAN, cisTEM), or use PDBe's [FSC server]



EMDataResource
Unified Data Resource for 3DEM

PDB Deposition Policies

- Map deposition to EMDB is mandatory for PDB depositions of 3DEM atomic coordinates
- MX atomic coordinates must be deposited in PDBx/mmCIF format
  - PDB format is still accepted for 3DEM and NMR (will change in future)
- Coordinates and experimental data share same release status (REL, HPUB, or HOLD).
  - Coordinate and experimental data are released simultaneously.
- ID's issued only after mandatory metadata are provided
- PI contact information and ORCiD ID are mandatory
- Author-initiated coordinate versioning is allowed post release

wwpdb.org/documentation/policy

3DEM Metadata Collection

- High-level classification of the EM experiment
- Software used for data collection, data processing, data analysis, structure calculations, and refinement
- Sample description (e.g., assembly, virus)
- Data collection (e.g., diffraction, imaging)
- Sample preparation (e.g., specimen, buffer, tomography, condition)
- Image processing and reconstruction
- Structure analysis for 3D fitting of atomic coordinates

mmCIF Data Dictionary for 3DEM

Top Level
em_experiment
em_software

Sample Description
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

Data Collection
em_diffraction
em_diffraction_shell
em_diffraction_stats
em_image_recording
em_image_scans
em_imaging
em_imaging_optics

Sample/Specimen Preparation
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*

Image Processing & Reconstruction
em_3d_reconstruction
em_image_processing
em_particle_selection
em_volume_selection
em_ctf_correction

em_2d_crystal_entity
em_3d_crystal_entity
em_helical_entity
em_single_particle_entity

em_euler_angle_assignment*
em_final_classification*
em_start_model*

Structure Analysis
em_3d_fitting
em_3d_fitting_list
em_fsc_curve*

Experimental Data
em_map*
em_structure_factors*
em_layer_lines*

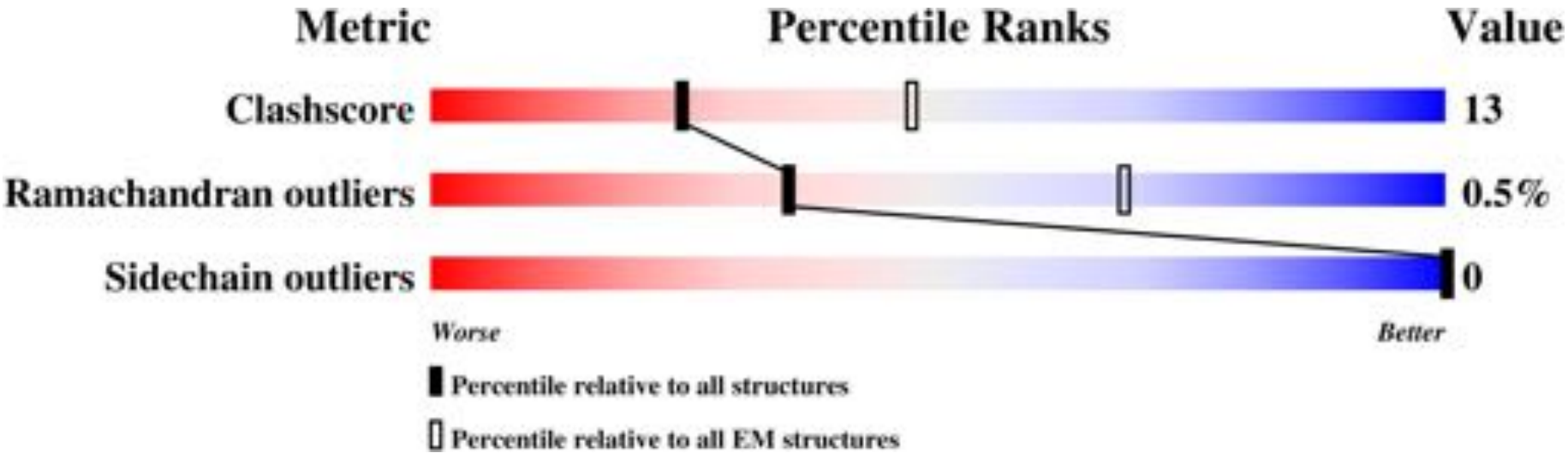
*All categories are collected by the OneDep system. Data from most categories are archived in both PDB and EMDB; asterisked categories are archived only in EMDB.

EMDataResource
Unified Data Resource for 3DEM

Cryo-EM Structure Validation

Validation Report for EM Structures v1 (2016-2019)

- Map resolution reported by depositor
- Model geometry statistics
- No fit-to-map validation



EMDataResource
Unified Data Resource for 3DEM

Validation Report for EM Structures v2 (2020-)

- New:
  - Map, Map+Model Images
  - FSC curve(s)
  - Rotationally averaged power spectrum
  - Fit-to-Map: Atom inclusion at recommended contour level

EMD-0273

Map Images

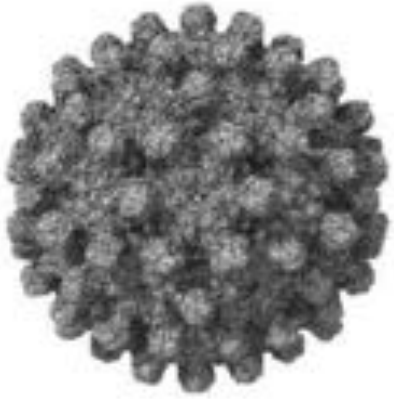
6.1 Orthogonal projections

X

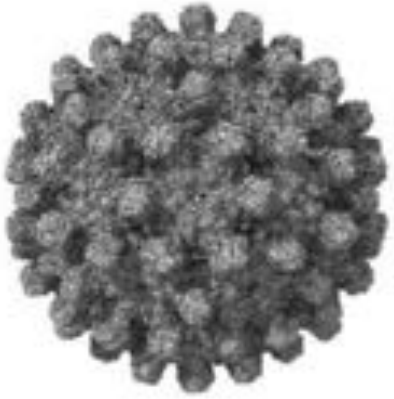
Y

Z

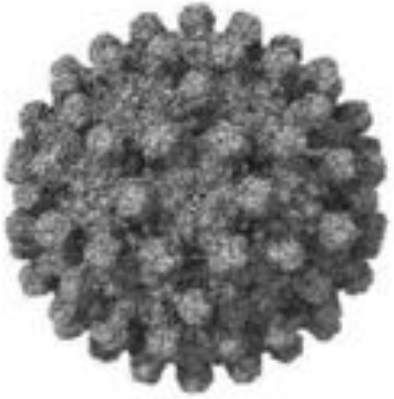
6.4 Orthogonal surface views



X



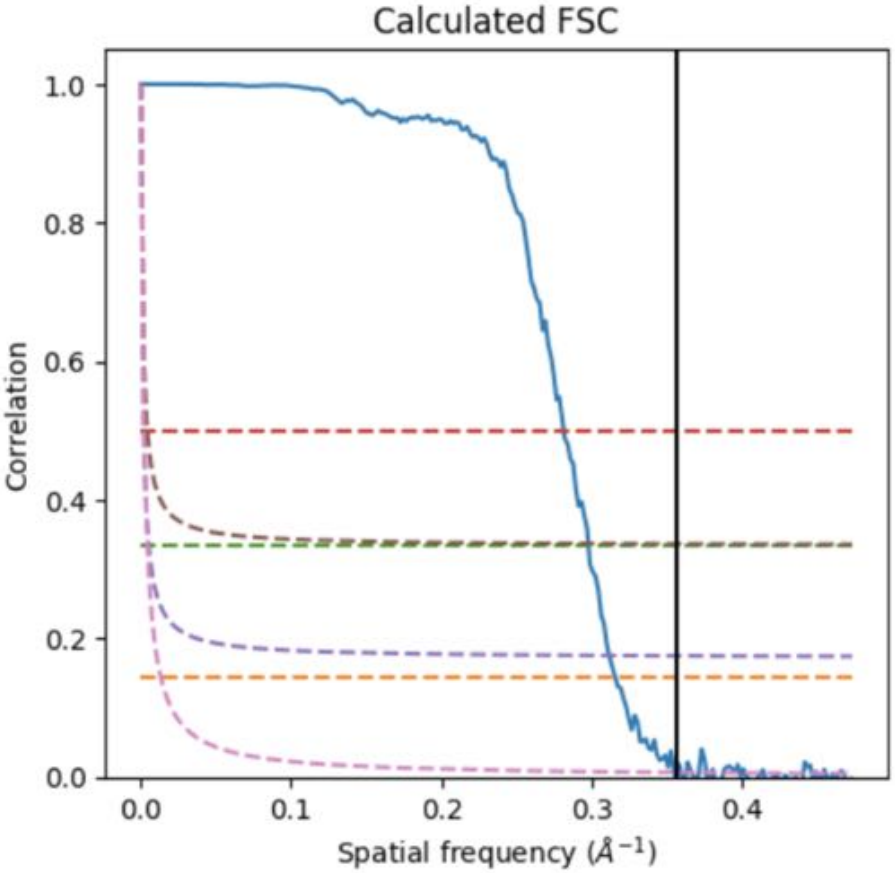
Y



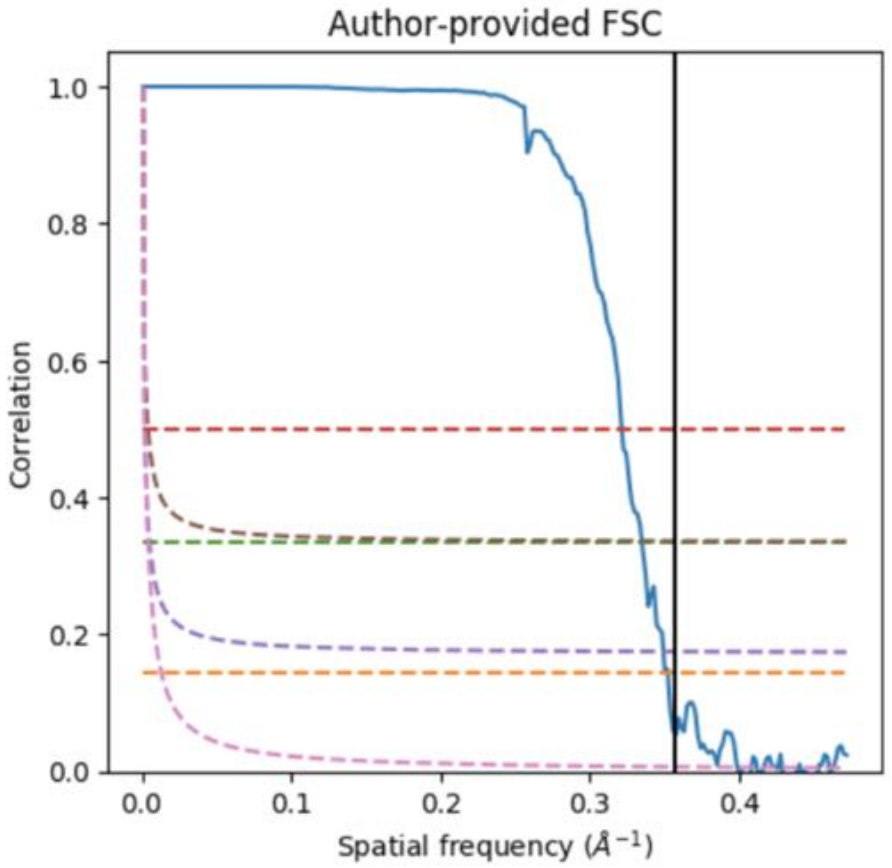
Z

Resolution Estimate by FSC

Calculated by archive from deposited half-maps



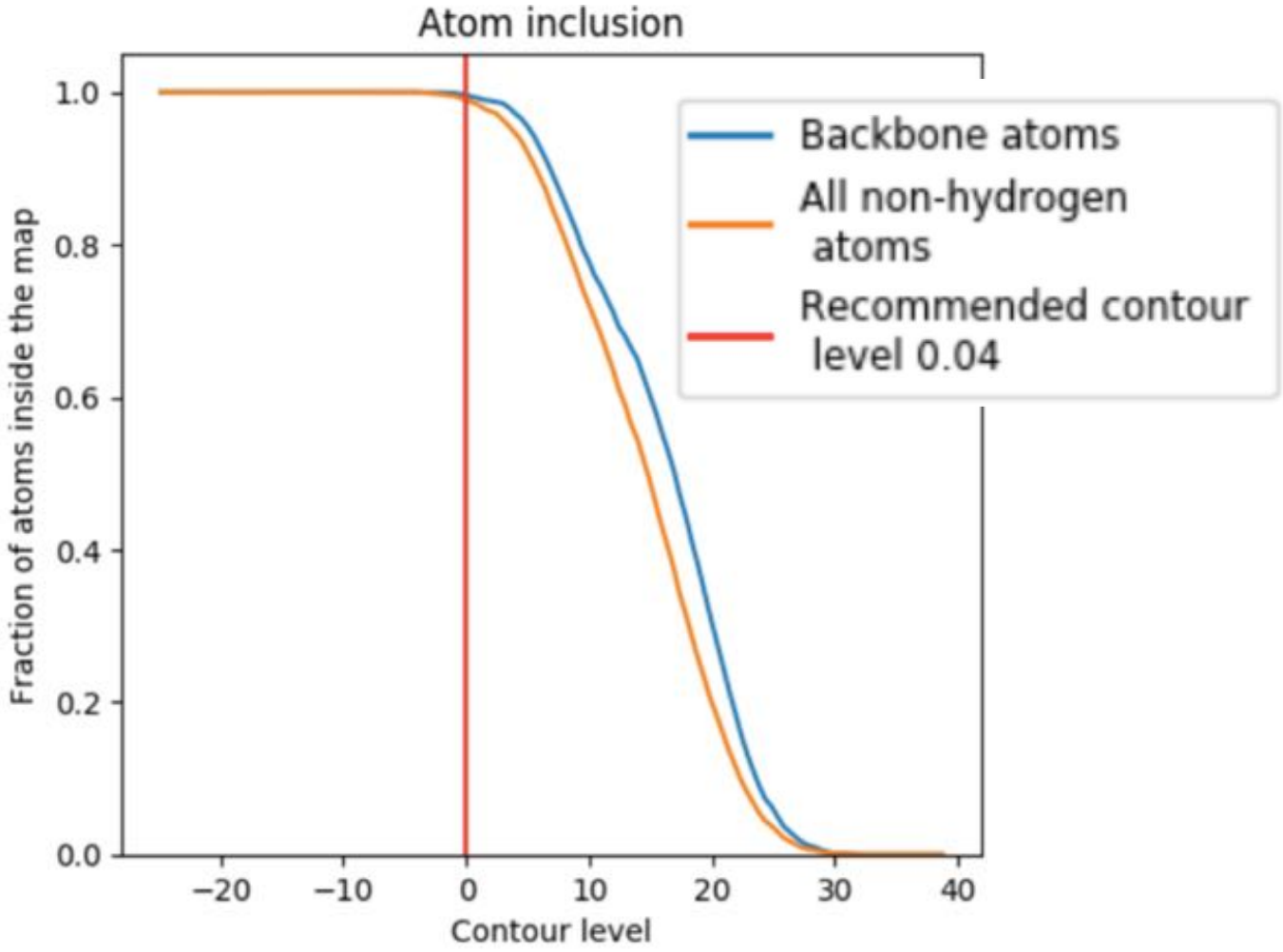
Calculated by depositor

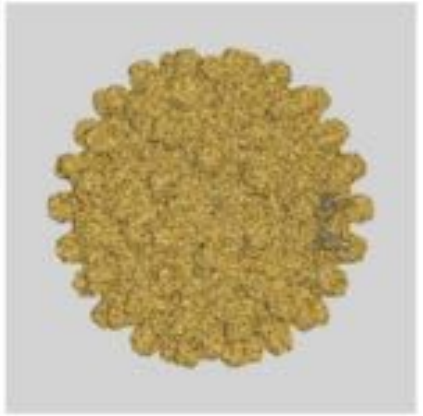


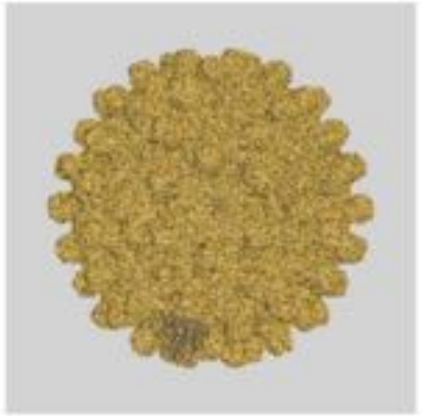
Blue: FSC curve; Vertical Black Line: reported resolution

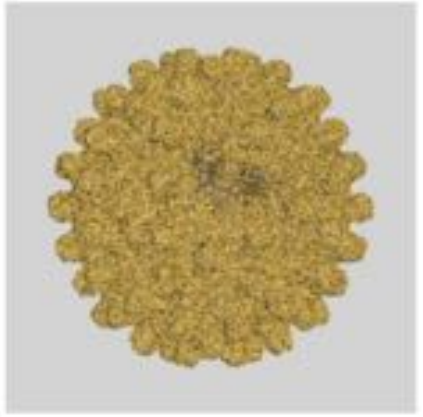
EMD-0273 PDB 6UH7

Atom Inclusion









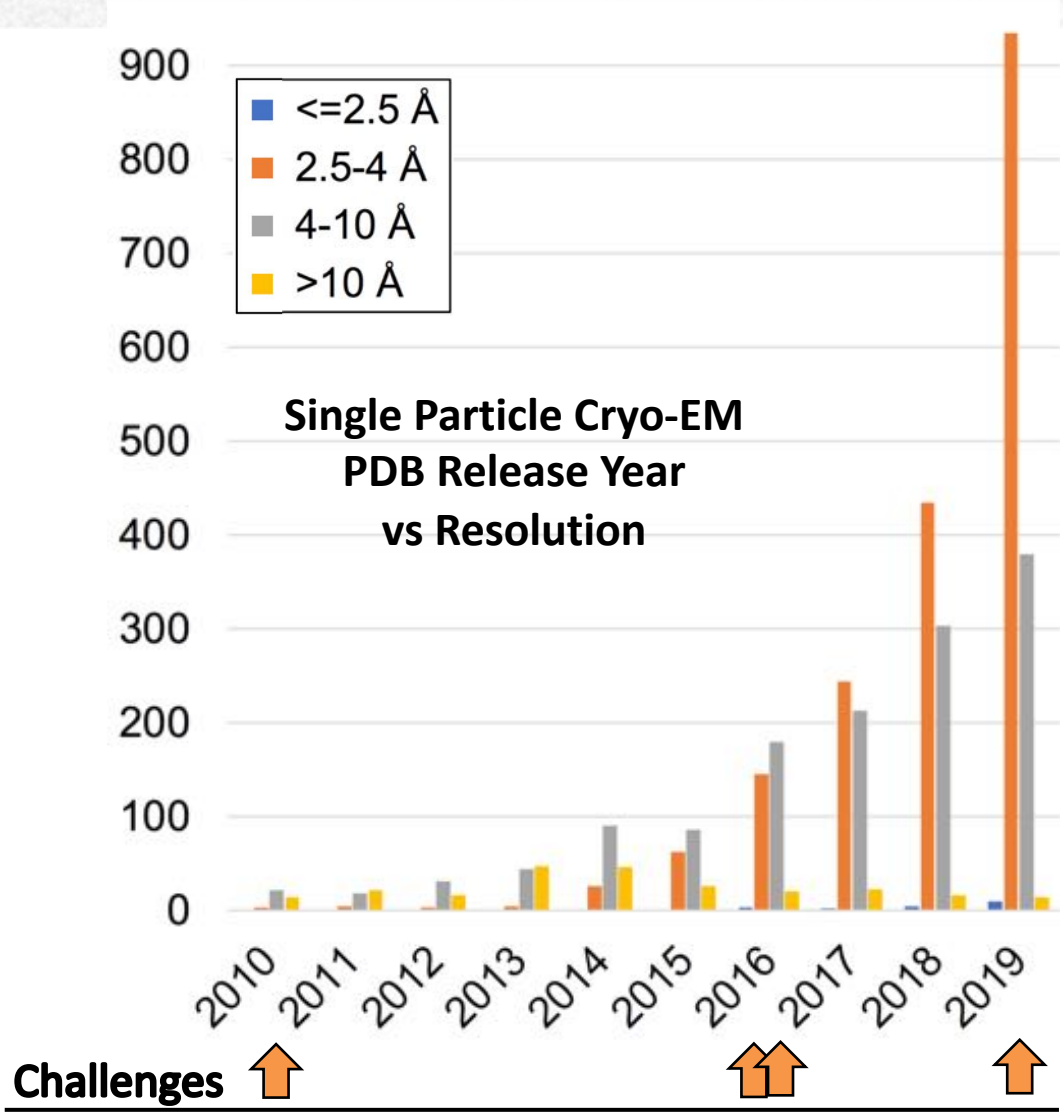
Validation Challenges

EM Validation Task Force 2010 Recommendations

- Full FSC curve from independent half-maps
- Model Stereochemistry same as X-ray / NMR
- Other Metrics: **More Research Needed**

Henderson *et al*. (2012) ***Structure 20***, 205-214
10.1016/j.str.2011.12.014

Validation in a Changing Landscape



- How accurate are the maps?
- Do the models conform to good valence geometry and stereochemistry?
- How well do the models fit the maps?
- What are the metrics for evaluation and are they good enough?

To answer these questions, we conducted a series of Challenges

EMDataResource
Unified Data Resource for 3DEM

Promoting Standards Development: Challenges

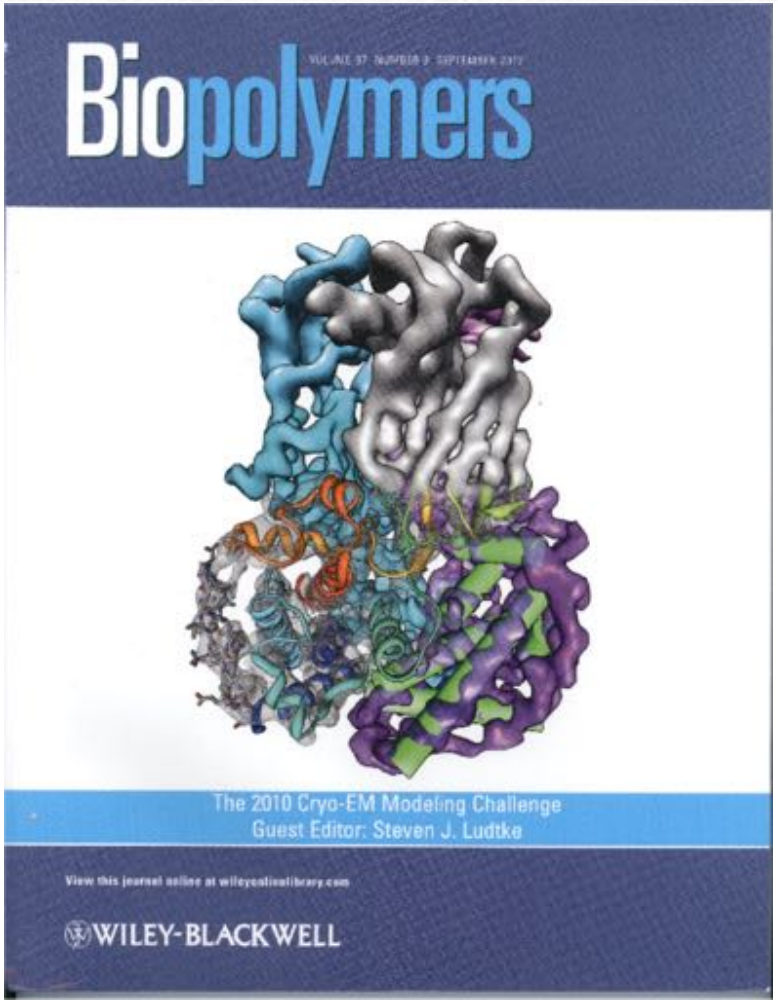
Challenge Workshop(s)	Resol (Å)	Goals for Participants and Assessors
2010 **Model Challenge** 2011 Hawaii 2012 Houston	2.5-24	❑ Produce best models against selected maps ❑ Explore segmentation, secondary structure detection, rigid body, flexible fitting, *ab initio*
2016-2017 **Map Challenge** 2017 Lake Tahoe 2017 Stanford/SLAC	2.5-5	❑ Produce best maps from selected raw images ❑ Produce best models against selected maps ❑ Compare reconstruction, modeling practices ❑ Explore assessment strategies esp. map and model fit-to-map
2016-2017 **Model Challenge** 2015 Boston 2017 New Orleans 2017 Stanford/SLAC		
2019 **Model "Metrics"** **Challenge** 2019 Stanford/SLAC	1.8-3.1	❑ Produce best models against selected maps ❑ Explore Model metrics with focus on Fit-to-Map

2010 Model Challenge: Participants

Chandrajit Bajaj, David Baker, Mariah Baker, Matthew Baker, Helen Berman, Radhakrishna Bettadapura, Virginia Burger, Kwok-Yan Chan, Chakra Chennubhotla, Wah Chiu, Frank DiMaio, Joachim Frank, Yaser Hashem, Tommy Hofmann, Shuiwang Ji, Tao Ju, Mert Karakaş, Steffen Lindert, Steven Ludtke, Cathy Lawson, Gerard Kleywegt, Jing He, Corey F. Hryc, Jens Meiler, Kamal Al Nasr, Grigore Pintilie, Ian Rees, Eduard Schreiner, Gunnar F. Schröder, Klaus Schulten, Dong Si, Phoebe L. Stewart, Leonardo G. Trabuco, Zhe Wang, Nils Wötzel, Qin Zhang

2010 Model Challenge: Observations

- Established community around a common problem
- Identified critical standardization issues related to data deposition
- Identified issues to explore in future challenges
- Identified Challenges as mechanism to establish modeling benchmarks

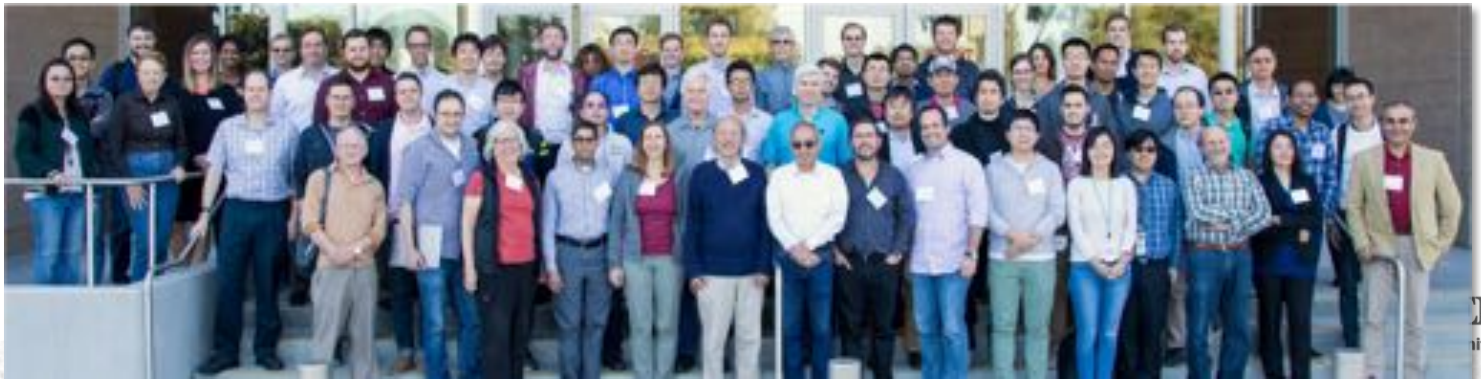


10.1002/bip.22081

2016-2017 Map & Model Challenges: Participants

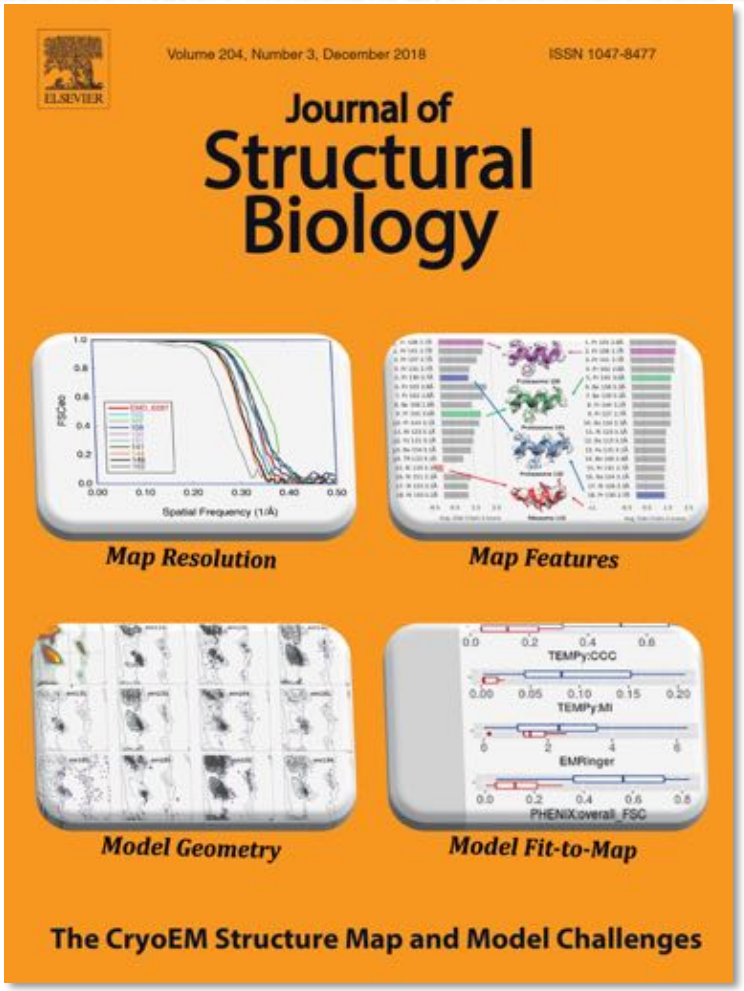
Maps: **Committee:** B Carragher (Chair), J-M Carazo, W Jiang, J Rubinstein, P Rosenthal, F Sun, J Vonck **Data Contributors**: Y Deng, F Sun, M Campbell, B Carragher, C Russo, L Passmore, J-P Armache, M Liao, Y Cheng, X Bai, S Scheres, Z Wang, W Chiu, A Bartesaghi, S Subramaniam **Challenge Participants:** A Punjani, A Nans, A Leith, A Chakraborty, JB Heymann, CO Sorzano, C Gati, D Tegunov, D-H Chen, F Li, G Yu, JM Bell, J Chen, JG Montoya, J Gomez-Blanco, K Yang, L Donati, L Estrozi, M Nilchian, N Caputo, N Grigorieff, S Stagg, S Shakeel, S Scheres, S Ludtke, X Bai, Y Lu **Assessors:** M Holmdahl, A Patwardhan, R Marabini, J-M Carazo, JB Heymann, J Mendez, S Stagg, G Pintilie,W Chiu, S Jonic, E Palovcak, J-P Armache, J Zhao, Y Cheng

Models: **Committee:** P Adams (Chair), A Brunger, R Read, T Schwede, M Topf, G Kleywegt **Data Contributors**: S Fromm, C Sachse, J-P Armache, Y Cheng, M Campbell, B Carragher, S-H Roh, C Hryc, W Chiu, Z Wang, A Bartesaghi, S Subramaniam, Xi Bai, S Scheres, N Fischer, H Stark, W Li, Z Liu, J Frank **Challenge Participants:** A Joseph, M Topf, B Frenz, F DiMaio, B Mao, C Hryc, W Chiu, D Kihara, G Terashi, D Matthies, G Schroeder, T Braun, K Wang, I Yu, H Zhou, M Baker, P Afonine, R McGreevy, A Singharoy, S Chittori, S Grudinin, A Hoffmann, T Kawabata, H Nakamura, T Terwilliger, T Croll, W Cao **Assessors:** A Krystafovych, A Joseph, M Topf, J Richardson, T Terwilliger, B Barad, J Fraser, A Jakobi, C Sachse



2016-2017 Map & Model Challenges: Observations

- Innovative methods for map and model fit-to-map assessment were introduced
- Map quality depended on participant level of experience
- Maps reported as the same resolution looked different from each other
- Models were “all over the place”



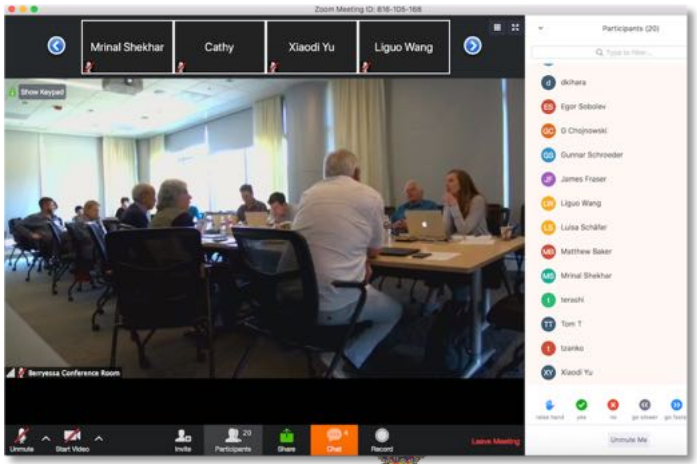
10.1016/j.jsb.2018.10.004

2016 Map & Model Challenges: Recommendations

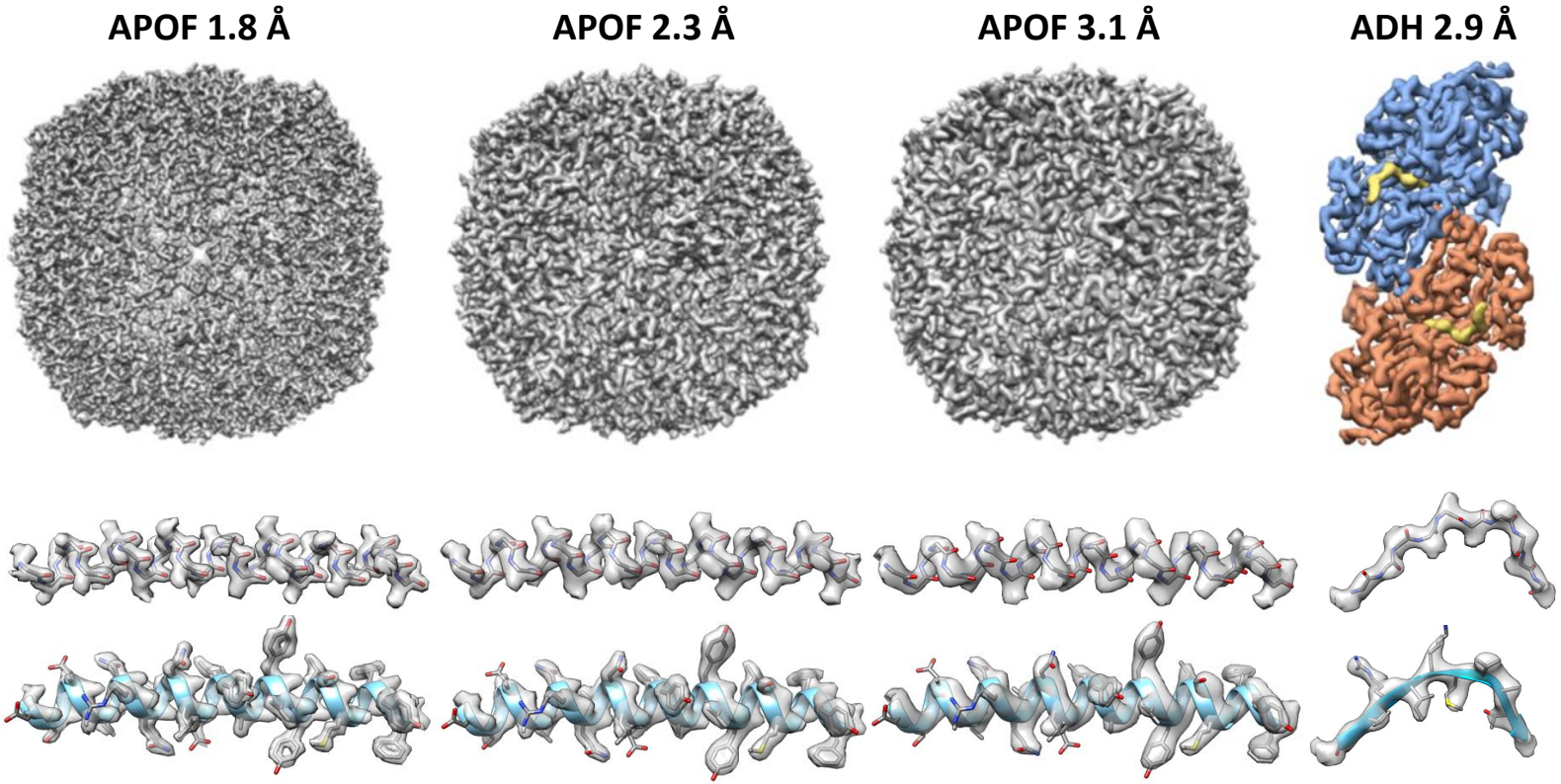
- Map Resolution by independent half-map FSC: uniform definition + software implementation needed
- Novel *model-based* methods may be useful for estimating map resolvability
- Needed: further review of fit-to-map metrics

2019 Model “Metrics” Challenge: Participants

Paul Adams, Pavel Afonine, Matt Baker, Helen Berman, Paul Bond, Tom Burnley, Renzhi Cao, Jianlin Cheng, Wah Chiu, Grzegorz Chojnowski, Kevin Cowtan, Frank DiMaio, Dan Farrell, James Fraser, Mark Herzik, Soon Wen Hoh, Maxim Igaev, Agnel Joseph, Daisuke Kihara, Andriy Kryshtafovych, Dilip Kumar, Cathy Lawson, Shanshan Li, Sumit Mittal, Bohdan Monastyrskyy, James Murray, Mateusz Olek, Colin Palmer, Ardan Patwardhan, Greg Pintilie, Alberto Perez, Jane Richardson, Peter Rosenthal, Daipayan Sarkar, Luisa Schaefer, Mike Schmid, Gunnar Schröder, Mrinal Shekhar, Dong Si, Abishek Singharoy, Genki Terashi, Tom Terwilliger, Maya Topf, Andrea Vaiana, Liguo Wang, Christopher Williams, Martyn Winn, Xiaodi Yu, Kaiming Zhang



2019 Challenge Targets



Challenge Pipeline

SETUP

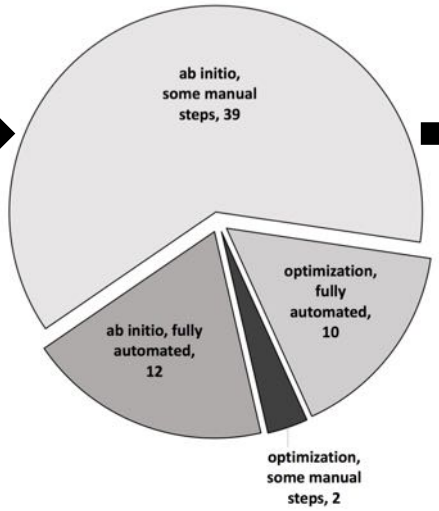
Map Targets:

Reference Models:
APOF: 3ajo
ADH: 6nbb

SUBMISSIONS

63 models total
51 ab initio
12 optimization

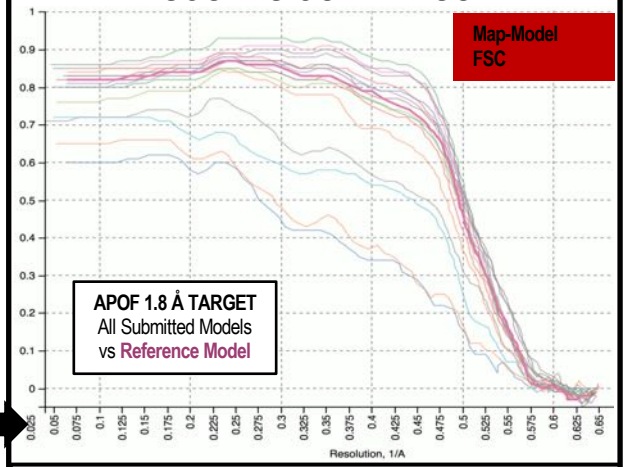
13 participating teams
from US and Europe

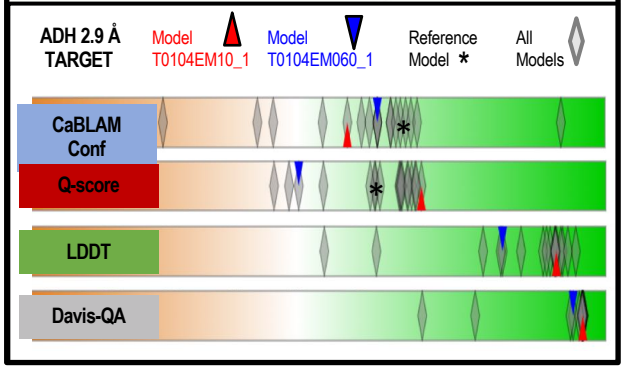


EVALUATION

Coordinates Only	•Configuration •Conformation •Clashes •Energy
Fit to Map	•Correlation •FSC-curve •Atom Inclusion •Rotamer
Comparison to Reference Model	•Superposition •Distances •Contacts
Comparison among Models	•Superposition

SCORES COMPARISON





model-compare.emdataresource.org

EMDataResource
Unified Data Resource for 3DEM

Coordinates Only	**Conformation**	Backbone	CaBLAM Cα-trace Cα-only virtual dihedrals CaBLAM Conformation Cα and CO-containing virtual dihedrals MOLPROBITY Ramachandran
		Sidechain	MOLPROBITY Rotamer
	Valence Geometry		PHENIX Bond | Bond angle | Chirality | Planarity | Dihedral
	Clashes		MOLPROBITY Clashscore
	Energy		PROQ3 energy and predicted features
Fit to Map	**Correlation**	Full map density	TEMPY CCC | PHENIX box_CC
		Density within a mask	TEMPy CCC_overlap | Segment Mander's Overlap PHENIX CC_peaks | CC_volume | CC_mask
		Density-derived functions	TEMPY Mutual Information(MI) | MI_overlap | Laplacian Filtered
		Density at atom positions	MAPQ Q-score: vs Reference Gaussians (r=0-2 Å)
	FSC curve	Single point	PHENIX Resolution Map-Model FSC = 0.5
		Integration	CCPEM REFMAC5 FSCavg curve area to defined resolution limit
	Atom Inclusion		TEMPy Envelope | EMDB Atom Inclusion
	Rotamer		EMRinger Z-score protein C_γ atom paths around $\chi 1$
vs Reference Model	**Superposition**	Cα Superposition	OPENSTRUCT RMSD-Cα
		Distance cutoffs	OPENSTRUCT Global Distance Calculation (GDC) all | sidechain Global Distance Test (GDT) total score | high accuracy
		Sequence assignment	PHENIX seq match | Cα atom position match | overall score
		*Multiple references	DAVIS-QA average of pairwise GDT_TS scores
vs Models Consensus*	**Distances**	Per chain	LDDT Local difference distance test
		All chains	OPENSTRUCT oligomeric LDDT | weighted oligomeric LDDT
	Contacts	Contact area	CAD Contact Area Difference
		Shared contacts	OPENSTRUCT Quaternary Structure (QS) best, global
		Hydrogen bonds	HBPLUS H-bond Precision all | nonlocal | Similarity all | nonlocal

2019 Challenge: 4 Metrics Stood Out

Coordinates alone:

CaBLAM: virtual dihedrals based on Cα, C=O compared to statistics from high quality PDB models [Williams 2018]

Fit-to-Map:

MAPQ Q-score: Per-atom Correlation vs Reference Gaussian (r=0-2 Å) [Pintilie in press]

PHENIX Resolution Map-Model FSC = 0.5 [Afonine 2018]

EMRinger Z-score "rotamericity" of map for protein C_γ atom paths around $\chi 1$ [Barad 2015]

2019 Challenge: Observations

- *ab initio* methods represented in the challenge performed extremely well for near-atomic maps (1.8 - 3.1 Å)
- For evaluating conformation: CaBLAM was a useful "orthogonal" metric to Ramachandran statistics
- *Within single map targets*, all fit-to-map metrics were equivalent (similar model rankings)
- only Q-score, EMRinger, and Map-Model FSC @ 0.5 provided useful comparisons *across map targets*

2019 Challenge: Recommendations

- Most fit-to-map metrics are fine for optimization against a single experimental map
- Resolution-sensitive metrics are preferred for ranking diverse structures in an archive
- CaBLAM is a valuable new tool for evaluating protein backbone conformation

Topics for Future Challenges

- Lower Resolution
- Membrane Proteins
- Nucleic Acids
- Models derived from Tomograms

Unified Data Resource for 3DEM

Stanford University (Baylor Coll. Med.)

Wah Chiu
Greg Pintilie
Mike Schmid

Steven Ludtke
Matt Baker
Corey Hryc
Ian Rees

UC Davis:
Andriy Kryshtafovych

RCSB PDB
PROTEIN DATA BANK

Rutgers University

Cathy Lawson
Helen Berman
Brinda Vallat
Brian Hudson

John Westbrook
Batsal Devkota
Raul Sala
Chunxiao Bi

European Bioinformatics Institute

Ardan Patwardhan
Gerard Kelywegt
Sanja Abbott
Ryan Pye
Osman Salih
Zhe Wang

Kim Henrick
Richard Newman
Christoph Best
Glen van Ginkel
Eduardo Sanz-Garcia
Ingvar Lagerstedt

NIH
National Institute of General Medical Sciences

EMDataResource is funded by the US National Institutes of Health/National Institute of General Medical Science, ***R01GM079429-12***

EMDataResource
Unified Data Resource for 3DEM

EM Structure Validation Servers

Map:	Service/Name	Link
Overall Shape & Hand	Tilt-Pair	pdbe.org/tiltpair
Resolution by FSC	FSC	pdbe.org/FSC
Local Resolution	3DFSC	3dfsc.salk.edu
Local Resolution	Scipion	scipion.cnb.csic.es/m/myresmap#

Model:	Service/Name	Link
Stereochemistry, compare with all PDB structures	wwPDB	validate.wwpdb.org
Stereochemistry	Molprobity	molprobity.biochem.duke.edu
Nucleic Acid conformation	DNATCO	dnatco.org

Map/Model Fit:	Service/Name	Link
“backbone bumpiness”	EMRinger	emringer.com (@UCSF)

See also: www.emdataresource.org/validation.html

EMDataResource
Unified Data Resource for 3DEM

References

- Lawson CL, Chiu W (2018) Comparing cryo-EM structures (Editorial). **J Struct Biol. 204**, 523-526. 10.1016/j.jsb.2018.10.004
- Patwardhan A & Lawson CL (2016). *Databases and Archiving for CryoEM*. **Methods Enzymol 579**, 393-412. 10.1016/bs.mie.2016.04.015
- Lagerstedt I, et al (2013). *Web-based visualisation and analysis of 3D electron-microscopy data from EMDB and PDB.* **J Struct Biol. 184,** 173-81. 10.1016/j.jsb.2013.09.021
- Henderson R, et al (2012) *Outcome of the first electron microscopy validation task force meeting.* **Structure 20**, 205-214. 10.1016/j.str.2011.12.014