EMDataResource: Structure Data Archiving, Validation Challenges

Cathy Lawson
EMDataResource & RCSB PDB
Rutgers University

S2C2 cryoEM Image Processing Workshop
June 12, 2020
Unified Data Resource for 3DEM

Established 2007 under NIH Support to:

- Develop Data Archives for 3DEM (EMDB + PDB)
- Promote Community Development of Validation and Standards
Growth of EM Archives

Growth of EM Archives 2020-03-04

- EMDB maps
- PDB EM models

2020
EMDB maps: 10,256
PDB EM models: 4,635

Released Entries (Cumulative)

updated weekly: emdataresource.org/statistics.html
EMDB maps by year and resolution

Maps Released by Resolution Range (Å) 2020-06-10

- Blue: 0-2 Å
- Red: 2-3 Å
- Orange: 3-4 Å
- Green: 4-5 Å
- Purple: 5-10 Å
- Cyan: 10-20 Å
- Pink: >20 Å

EMDB Entries Released per Year

- 2020: 594 entries

updated weekly: emdataresource.org/statistics.html
# Finding Cryo-EM Structures

[emdataresource.org/search.html](http://emdataresource.org/search.html)

<table>
<thead>
<tr>
<th>EMDB maps</th>
<th>PDB EM models</th>
<th>EMDB + PDB</th>
<th>Raw image data</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMDR Search</td>
<td>RCSB-PDB</td>
<td>EM Navigator @ PDBj</td>
<td>EMPIAR @ EBI</td>
</tr>
<tr>
<td>Advanced Search @ EBI</td>
<td>PDBe</td>
<td>3DBioNotes @ CSIC</td>
<td></td>
</tr>
</tbody>
</table>
EMDR Search: Features

Autocomplete: ribo

Interactive keyword help:

Sort, configure, filter, download search results:

Found 1113 EMDB map entries for query ribosome (Released Entries, Any Method, Any Sample Type, Any Specimen Type, Any Resolution Status)

<table>
<thead>
<tr>
<th>EMD-</th>
<th>Status</th>
<th>Entry Title</th>
<th>% Deposit Date</th>
<th>% Resolution (Å)</th>
<th>% Associated PDB</th>
</tr>
</thead>
<tbody>
<tr>
<td>0047</td>
<td>REL</td>
<td>Cryo-EM reconstruction of yeast 80S ribosome in complex with mRNA tRNA and eEF2 (GMPPCP/sordarin)</td>
<td>2018-06-07</td>
<td>4.4</td>
<td>6gq1</td>
</tr>
<tr>
<td>0048</td>
<td>REL</td>
<td>Cryo-EM reconstruction of yeast 80S ribosome in complex with mRNA tRNA and eEF2 (GDP+AIF4/sordarin)</td>
<td>2018-06-07</td>
<td>3.9</td>
<td>6ggb</td>
</tr>
<tr>
<td>0049</td>
<td>REL</td>
<td>Cryo-EM reconstruction of yeast 80S ribosome in complex with mRNA tRNA and eEF2 (GMPPCP)</td>
<td>2018-06-08</td>
<td>4</td>
<td>6gqv</td>
</tr>
<tr>
<td>0055</td>
<td>REL</td>
<td>Cryo-EM reconstruction of yeast 80S ribosome in complex with mRNA tRNA and eEF2 (GDP+AIF4/sordarin): SSU-focused map</td>
<td>2018-06-14</td>
<td>4.3</td>
<td>6gsm</td>
</tr>
<tr>
<td>0057</td>
<td>REL</td>
<td>Cryo-EM structure of a partial yeast 48S preinitiation complex in its open conformation</td>
<td>2018-06-14</td>
<td>5.2</td>
<td></td>
</tr>
</tbody>
</table>
New: Simple Search

EMDB maps: ribosome

Click here for EMDR's enhanced coronavirus search

- all
- single particle
- electron crystallography
- tomography
- subtomogram averaging
- helical
- virus
- ribosome
- PDB coordinates
- no PDB coordinates
- EMPIAR raw data

Enter keyword

Sort by Resolution

Show 20 Entries

Found 1138 EMDB map entries with resolution=any value. View as EMDR Search List.

20 of 1138 shown

EMD-20353  PDB:6pj6  Released 2020-01-22 Deposited
2019-06-27 singleParticle 2.2 Å High resolution cryo-EM structure of E.coli 50S Stojevik V, Myasnikov A, Frost A, Fujiwara DG PubMed DOI

EMD-10077  PDB:6s0z  Released 2019-08-21 Deposited

EMD-7025  PDB:8az3  Released 2017-12-06 Deposited

EMD-8361  PDB:81sh  Released 2016-10-12 Deposited

EMD-10377  PDB:814q  Released 2019-12-25 Deposited
New: Coronavirus Search

EMDB holdings for Coronavirus-Related Structures

List of additional CoV resources
- all coronavirus
- SARS-CoV-2
- SARS-CoV
- MERS-CoV
- H-CoV
- MHV
- AIVB
- FIPV
- PD-CoV
- PEDV
- Spike
- Non-Spike
- ACE2 receptor
- Fab
- with PDB
- no PDB
- EMPIAR raw data

Enter keyword

Sort by
- Release Date

Show 20 Entries

Found 30 EMDB map entries. View as EMDR Search List.

20 of 30 shown

EMD-22078  PDB:6x6p  Released 2020-06-10  Deposited

EMD-11068  PDB:6z43  Released 2020-06-03  Deposited

EMD-30276  PDB:7c2k  Released 2020-06-03  Deposited

EMD-30247  PDB:7byr  Released 2020-06-10  Deposited

EMD-30252  PDB:7bzf  Released 2020-06-03  Deposited

EMD-30283  Released 2020-06-03  Deposited 2020-05-10
2020-06-07 singleParticle 3.03 Å 0.15 MÅ COVID-19 RNA-dependent RNA polymerase pre-translocated catalytic complex in conformation I Wang Q, Gao Y, Ji W, Mu A, Rao Z DOI
Cryo-EM Structure Deposition

EMDB, PDB
wwPDB OneDep System

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

---

- Deposition
- Annotation Workflow
- Dissemination

- Deposit Data
- Ligand Processing
- Sequence Processing
- Manual and Automated Annotation
- Validation

- Global PDB system for multiple experimental methods
- wwPDB FTP
- EMDB
- BMRB

Deposit X-ray structures at: deposit.wwpdb.org/deposition/
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

![FSC Curve Image](PDBe.org/FSC)

- FSC
- 3 sigma @ 2.1Å
- 1/2 bit @ 3.3Å
- 0.5 @ 3.9Å
- 0.333 @ 3.5Å
- 0.143 @ 3.2Å
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

<table>
<thead>
<tr>
<th>Top Level</th>
<th>Sample/Specimen Preparation</th>
<th>Image Processing &amp; Reconstruction</th>
<th>Experimental Data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>em_experiment</strong></td>
<td><strong>em_buffer</strong></td>
<td><strong>em_3d_recognition</strong></td>
<td><strong>em_map</strong></td>
</tr>
<tr>
<td><strong>em_software</strong></td>
<td><strong>em_buffer_component</strong></td>
<td><strong>em_image_processing</strong></td>
<td><strong>em_structure_factors</strong></td>
</tr>
<tr>
<td><strong>Sample Description</strong></td>
<td><strong>em_crystal_formation</strong></td>
<td><strong>em_particle_selection</strong></td>
<td><strong>em_layer_lines</strong></td>
</tr>
<tr>
<td><strong>em_entity_assembly</strong></td>
<td><strong>em_embedding</strong></td>
<td><strong>em_volume_selection</strong></td>
<td></td>
</tr>
<tr>
<td><strong>em_entity_assembly_molwt</strong></td>
<td><strong>em_sample_support</strong></td>
<td><strong>em_ctf_correction</strong></td>
<td></td>
</tr>
<tr>
<td><strong>em_entity_assembly_naturalsource</strong></td>
<td><strong>em_specimen</strong></td>
<td><strong>em_2d_crystal_entity</strong></td>
<td></td>
</tr>
<tr>
<td><strong>em_entity_assembly_recombinant</strong></td>
<td><strong>em_vitrification</strong></td>
<td><strong>em_3d_crystal_entity</strong></td>
<td></td>
</tr>
<tr>
<td><strong>em_virus_entity</strong></td>
<td><strong>em_3d_crystal_entity</strong></td>
<td><strong>em_helical_entity</strong></td>
<td></td>
</tr>
<tr>
<td><strong>em_virus_natural_host</strong></td>
<td><strong>em_vitrification</strong></td>
<td><strong>em_volume_selection</strong></td>
<td></td>
</tr>
<tr>
<td><strong>em_virus_shell</strong></td>
<td></td>
<td><strong>em_final_classification</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Data Collection</strong></td>
<td><strong>em_2d_crystal_entity</strong></td>
<td><strong>em_start_model</strong></td>
<td></td>
</tr>
<tr>
<td><strong>em_diffraction</strong></td>
<td><strong>em_3d_fitting</strong></td>
<td><strong>em_3d_fitting_list</strong></td>
<td></td>
</tr>
<tr>
<td><strong>em_diffraction_shell</strong></td>
<td><strong>em_high_pressure_freezing</strong></td>
<td><strong>em_euler_angle_assignment</strong></td>
<td></td>
</tr>
<tr>
<td><strong>em_diffraction_stats</strong></td>
<td><strong>em_shadowing</strong></td>
<td><strong>em_final_classification</strong></td>
<td></td>
</tr>
<tr>
<td><strong>em_image_recording</strong></td>
<td><strong>em_support_film</strong></td>
<td><strong>em_start_model</strong></td>
<td></td>
</tr>
<tr>
<td><strong>em_image_scans</strong></td>
<td><strong>em_tomography</strong></td>
<td><strong>em_euler_angle_assignment</strong></td>
<td></td>
</tr>
<tr>
<td><strong>em_imaging</strong></td>
<td><strong>em_tomography</strong></td>
<td><strong>em_final_classification</strong></td>
<td></td>
</tr>
<tr>
<td><strong>em_imaging_optics</strong></td>
<td><strong>em_vitrification</strong></td>
<td><strong>em_start_model</strong></td>
<td></td>
</tr>
</tbody>
</table>

*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.*
Cryo-EM Structure Validation

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

### Metrics and Percentile Ranks

<table>
<thead>
<tr>
<th>Metric</th>
<th>Percentile Ranks</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clashscore</td>
<td></td>
<td>13</td>
</tr>
<tr>
<td>Ramachandran outliers</td>
<td></td>
<td>0.5%</td>
</tr>
<tr>
<td>Sidechain outliers</td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

- **Worse**
  - Percentile relative to all structures
- **Better**
  - Percentile relative to all EM structures
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
  - Reports for both EMDB-only and EMDB+PDB

- **Right now:**
  - V2 reports are only available to depositors

- **Very Soon:**
  - Will be available for download from EMDB and PDB archives
Ongoing Cryo-EM Validation Discussion

- Frontiers in Cryo-EM Validation Jan 2019

- wwPDB Single-Particle Data Management Jan 2020
Preview of Some Outcomes/Recommendations

■ Map Resolution:
  ■ (unmasked, minimally filtered) half-maps should be deposited
  ■ Data archive should calculate FSC from these half-maps

■ Global Map-Model Fit:
  ■ By Map-Model FSC (full curve, read at threshold=0.5)
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
# Promoting Standards Development: Challenges

<table>
<thead>
<tr>
<th>Challenge Workshop(s)</th>
<th>Resol (Å)</th>
<th>Goals for Participants and Assessors</th>
</tr>
</thead>
</table>
| **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 | 2.5-5 | - Produce best models against selected maps  
- Explore Model metrics with focus on Fit-to-Map |
| **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 |
Challenge Workshops

2011

2017

2019
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: 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 Challenge Targets

APOF 1.8 Å

APOF 2.3 Å

APOF 3.1 Å

ADH 2.9 Å

EMD-20026 additional map #1

EMD-20027 additional map #2

EMD-20028 additional map #2

EMD-0406 primary map
Challenge Pipeline

**SETUP**
- **Map Targets:**
  - T1 APOF 1.8 Å
  - T2 APOF 2.3 Å
  - T3 APOF 3.1 Å
  - T4 ADH 2.9 Å

**Reference Models:**
- T1-T3 APOF: 3ajo
- T4 ADH: 6nbb

**SUBMISSIONS**
- 63 models total
  - 51 ab initio
  - 12 optimization

13 participating teams from US and Europe

**EVALUATION**
- **Fit to Map**
  - Correlation
  - Map-Model FSC
  - Atom Inclusion
  - Rotamer

- **Coordinates Only**
  - Configuration
  - Conformation
  - Clashes
  - Energy

- **Comparison to Reference Model**
  - Superposition
  - Distances
  - Contacts

- **Comparison among Models**
  - Superposition

**SCORES COMPARISON**

model-compare.emdataresource.org
<table>
<thead>
<tr>
<th>Coordinates Only</th>
<th>Conformation</th>
<th>Backbone</th>
<th>CaBLAM Co-trace Cα-only virtual dihedrals</th>
<th>CaBLAM Conformation Cα and CO-containing virtual dihedrals</th>
<th>MOLPROBITY Ramachandran</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sidechain</td>
<td>MOLPROBITY</td>
<td>MOLPROBITY Rotamer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Valence Geometry</td>
<td>Density within a mask</td>
<td>TEMPy CCC_overlap</td>
<td>Segment Mander’s Overlap</td>
<td>PHENIX CC_peaks</td>
<td>CC_volume</td>
</tr>
<tr>
<td></td>
<td>Density-derived functions</td>
<td>TEMPy Mutual Information(MI)</td>
<td>MI_overlap</td>
<td>Laplacian Filtered</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Density at atom positions</td>
<td>MAPQ Q-score: vs Reference Gaussians (r=0-2 Å)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clashes</td>
<td>FSC curve</td>
<td>Single point</td>
<td>PHENIX Resolution Map-Model FSC = 0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Integration</td>
<td>CCPEM REFMAC5 FSCavg curve area to defined resolution limit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Energy</td>
<td>Atom Inclusion</td>
<td>TEMPy Envelope</td>
<td>EMDB Atom Inclusion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rotamer</td>
<td></td>
<td>EMRinger Z-score protein Cα, atom paths around χ1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>vs Reference Model</td>
<td>Superposition</td>
<td>Cα Superposition</td>
<td>OPENSTRUCT RMSD-Cα</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Distance cutoffs</td>
<td>OPENSTRUCT Global Distance Calculation (GDC) all</td>
<td>Global Distance Test (GDT) total score</td>
<td>high accuracy</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sequence assignment</td>
<td>PHENIX seq match</td>
<td>Cα atom position match</td>
<td>overall score</td>
<td></td>
</tr>
<tr>
<td></td>
<td>*Multiple references</td>
<td>DAVIS-QA average of pairwise GDT_TS scores</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>vs Models Consensus*</td>
<td>Distances</td>
<td>Per chain</td>
<td>LDDT Local difference distance test</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>All chains</td>
<td>OPENSTRUCT oligomeric LDDT</td>
<td>weighted oligomeric LDDT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contacts</td>
<td>Contact area</td>
<td>CAD Contact Area Difference</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shared contacts</td>
<td>OPENSTRUCT Quaternary Structure (QS) best, global</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hydrogen bonds</td>
<td>HBPLUS H-bond Precision all</td>
<td>nonlocal</td>
<td>Similarity all</td>
<td>nonlocal</td>
</tr>
</tbody>
</table>
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]

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 Cγ atom paths around χ1 [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
- COVID-19
Unified Data Resource for 3DEM

Stanford University (Baylor Coll. Med.)
- Wah Chiu
- Greg Pintilie
- Mike Schmid
- Steven Ludtke
- Matt Baker
- Corey Hryc
- Ian Rees

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

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

- Manuscript preprint will be available shortly at [BioRxiv: Outcomes of the 2019 EMDataResource model challenge: validation of cryo-EM models at near-atomic resolution](https://www.biorxiv.org/)


