EM Structure Archives, Validation and Challenges

Cathy Lawson
EMDataResource & RCSB
Rutgers University

S2C2 Modeling Course
January 17, 2020
Unified Data Resource for 3DEM

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

https://www.emdataresource.org
Growth of EM Archives

- **EMDB maps**: 9,901
- **PDB EM models**: 4,386

*Graph showing the growth of EM Archives from 2002 to 2020.*
EMDB maps by year and resolution

Maps Released by Resolution Range (Å) 2020-01-15
- 0-2
- 2-3
- 3-4
- 4-5
- 5-10
- 10-20
- >20

EMDB Entries Released per Year

- 2013
- 2014
- 2015
- 2016
- 2017
- 2018
- 2019
- 2020

2019
3-4: 934
Finding Cryo-EM Structures

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

Found 947 EMDB map entries for query ribosome (Released Entries, Any Method, Any Specimen Type, Any Resolution Status)

<table>
<thead>
<tr>
<th>EMDB id</th>
<th>Status</th>
<th>Entry Title</th>
<th>Entry Authors</th>
<th>Deposit Date</th>
<th>Resolution (Å)</th>
<th>Associated PDB</th>
</tr>
</thead>
<tbody>
<tr>
<td>MD-0048</td>
<td>REL</td>
<td>Cryo-EM reconstruction of yeast 80S ribosome in complex with mRNA tRNA and eEF2 (GDP+AlF4/sordarin)</td>
<td>Pellegrino S, Yusupov M, Yusupova G, Hashem Y</td>
<td>2018-06-07</td>
<td>3.9</td>
<td>6gqb</td>
</tr>
<tr>
<td>MD-0049</td>
<td>REL</td>
<td>Cryo-EM reconstruction of yeast 80S ribosome in complex with mRNA tRNA and eEF2 (GMPPCP)</td>
<td>Pellegrino S, Yusupov M, Yusupova G, Hashem Y</td>
<td>2018-06-08</td>
<td>4</td>
<td>6gqv</td>
</tr>
<tr>
<td>MD-0055</td>
<td>REL</td>
<td>Cryo-EM reconstruction of yeast 80S ribosome in complex with mRNA tRNA and eEF2 (GDP+AlF4/sordarin): SSU-focused mao</td>
<td>Pellegrino S, Yusupov M, Yusupova G, Hashem Y</td>
<td>2018-06-14</td>
<td>4.3</td>
<td></td>
</tr>
</tbody>
</table>
What EMDB entries do you want to search for?
Cryo-EM Structure Deposition

EMDB, PDB
wwPDB "OneDep" System

- Deposition system for X-ray, NMR, and EM Structures
- EM: Deposit map to EMDB, coordinate model to PDB
- Validation report is produced
File uploads for EM deposition

- Coordinates (mmCIF format)
- Coordinates (PDB format)
- Main map (mandatory)
  - EM map (MRC/CCP4 format)
- Image for EMDB (mandatory)
  - Entry image for public display
- Additional maps
  - Additional EM map (MRC/CCP4 format)
- Masks
  - EM mask (MRC/CCP4 format)
- Half (even-odd) maps
  - EM half map (MRC/CCP4 format)
- Structure Factors
  - mmCIF (structure factors)
  - MTZ
- Other Files
  - FSC file (XML format)
  - Ligand Image
FSC Curve

- Upload XML format file
- Create via software package (e.g., Relion, EMAN, cisTEM), or FSC server

PDBe.org/FSC
# mmCIF Data Dictionary for cryo-EM

<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>em_experiment</td>
<td>em_buffer</td>
<td>em_3d_reconstruction</td>
<td>em_map*</td>
</tr>
<tr>
<td>em_software</td>
<td>em_buffer_component</td>
<td>em_image_processing</td>
<td>em_structure_factors*</td>
</tr>
<tr>
<td></td>
<td>em_crystalFormation</td>
<td>em_particle_selection</td>
<td>em_layer_lines*</td>
</tr>
<tr>
<td></td>
<td>em_embedding</td>
<td>em_volume_selection</td>
<td></td>
</tr>
<tr>
<td></td>
<td>em_sample_support</td>
<td>em_ctf_correction</td>
<td></td>
</tr>
<tr>
<td></td>
<td>em_specimen</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>em_staining</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>em_vitrification</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>em_fiducial_markers*</td>
<td></td>
</tr>
<tr>
<td>Sample Description</td>
<td>em_buffer_component</td>
<td>em_grid_pretreatment</td>
<td></td>
</tr>
<tr>
<td>em_entity_assembly</td>
<td>em_crystal Formation</td>
<td>em_high_pressure_freezing</td>
<td></td>
</tr>
<tr>
<td>em_entity_assembly_molwt</td>
<td>em_embedding</td>
<td>em_shadowing</td>
<td></td>
</tr>
<tr>
<td>em_entity_assembly_naturalsource</td>
<td>em_sample_support</td>
<td>em_support_film</td>
<td></td>
</tr>
<tr>
<td>em_entity_assembly_recombinant</td>
<td>em_staining</td>
<td>em_tomography</td>
<td></td>
</tr>
<tr>
<td>em_virus_entity</td>
<td>em_vitrification</td>
<td>em_euler_angle_assignment</td>
<td></td>
</tr>
<tr>
<td>em_virus_natural_host</td>
<td></td>
<td>em_final_classification</td>
<td></td>
</tr>
<tr>
<td>em_virus_shell</td>
<td></td>
<td>em_start_model</td>
<td></td>
</tr>
<tr>
<td>Data Collection</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>em_diffraction</td>
<td>Sample/Specimen Preparation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>em_diffraction_shell</td>
<td>Image Processing &amp; Reconstruction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>em_diffraction_stats</td>
<td>Experimental Data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>em_image_recording</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>em_image_scans</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>em_imaging</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>em_imaging_optics</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*All categories are collected by the OneDep system. Most categories are archived in both PDB and EMDB; asterisked categories are archived only in EMDB.
Cryo-EM Structure Validation

- Map resolution reported by depositor
- Model geometry statistics
- No fit-to-map validation
New:
- Map and Map+Model Images
- FSC curve plot(s)
- Rotationally averaged power spectrum plot
- Fit-to-Map: Atom inclusion at recommended contour level plot
6.1 Orthogonal projections

6.4 Orthogonal surface views
FSC Curves

Calculated by archive from deposited half-maps

Calculated by depositor

Blue: FSC curve; Vertical Black Line: reported resolution
Map + Model Images

9.1 Map-model overlay

X Y Z
Atom Inclusion

EMD-0273 PDB 6UH7
Ribosome Example

EMD-2914  PDB 5AJ4
Ribosome Example

Recommended contour level

rcsb.org/3d-view/molstar/5aj4

Lower contour level
CryoEM 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

vtf.emdataresource.org
Validation in a Changing Landscape

- How accurate are the maps and their model interpretations?
- What criteria are currently being used and are they good enough?
2016 Map & Model Challenges

- Reconstruction and Modelling tasks at 2.5-5 Å

- Major outcomes:
  - Estimation of map resolution needs to be better standardized across the community
  - Novel model-based methods may be useful for estimating map resolvability
  - Further review of global fit-to-map metrics is needed
Goal: Identify metrics most suitable for evaluating and comparing fit of atomic coordinate models into cryo-EM maps for specimens in the 1.5-4.0 Å reported overall resolution range.

We received 63 models from 16 modelling teams
- 51 of 63 were modelled using *ab initio* methods
Model Compare Pipeline

- “Laboratory” for evaluating assessments

- [http://model-compare.emdataresource.org](http://model-compare.emdataresource.org)

Andriy Kryshtafovych
UC Davis
Model Challenge Meeting @ Stanford/SLAC June 2019

- **External Advisors/Assessors:** Peter Rosenthal, Paul Emsley, Jane Richardson, Paul Adams, James Fraser, Frank DiMaio, Pavel Afonine, Tom Terwilliger, Mark Herzik

- **Challengers:** Soon Wen Hoh, Gunnar Schroeder, Andrea Vaiana, Grzegorz Chojnowski, Daisuke Kihara, Pavel Afonine, Abishek Singharoy, Xiaodi Yu, Liguo Wang, Frank DiMaio, Matt Baker

- **EMDataResource:** Andriy Kryshtafovych, Cathy Lawson, Wah Chiu, Greg Pintilie, Helen Berman
<table>
<thead>
<tr>
<th>Coordinates Only</th>
<th>Conformation</th>
<th>Backbone</th>
<th>CaBLAM Cα-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>PHENIX CCC</td>
<td>PHENIX box_CC</td>
<td>TEMPy CCC_overlap</td>
<td>Segment Mander’s Overlap</td>
</tr>
<tr>
<td>Clashes</td>
<td>Density-derived functions</td>
<td>PHENIX</td>
<td>MANDER’s Overlap</td>
<td>TEMPy Mutual Information(MI)</td>
<td>MI_overlap</td>
</tr>
<tr>
<td>Energy</td>
<td>Density at atom positions</td>
<td>MAPQ Q-score: vs Reference Gaussians (r=0-2 Å)</td>
<td></td>
<td>MAPQ Q-score: vs Reference Gaussians (r=0-2 Å)</td>
<td></td>
</tr>
<tr>
<td>Correlation</td>
<td>Single point</td>
<td>PHENIX Resolution Map-Model FSC = 0.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FSC curve</td>
<td>Integration</td>
<td>CCPM REFMAC5 FSCavg curve area to defined resolution limit</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Atom Inclusion</td>
<td></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>OPENSTRUCT RMSD-Cα</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Distance cutoffs</td>
<td>OPENSTRUCT Global Distance Calculation (GDC) all</td>
<td>OPENSTRUCT Global Distance Calculation (GDC) all</td>
<td>sidechain Global Distance Test (GDT) total score</td>
<td>high accuracy</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></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></td>
<td>Contacts</td>
<td>Contact area</td>
<td>CAD Contact Area Difference</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>
Global Fit-to-Map Metrics Comparison

Average correlation per target map

Correlation across all four target maps
3 Fit-to-Map Metrics yielded Resolution-Sensitive Model Ranking

- **Within single map targets**, all fit-to-map metrics were largely equivalent (gave similar model rankings)

- **Across all map targets**, only 3 fit-to-map metrics yielded resolution-sensitive (human-intuitive) model rankings:
  - Q-score
  - EMRinger
  - Map-Model FSC @ 0.5
Resolution-Sensitive Ranking

APOF 1.8 Å

APOF 2.3 Å

APOF 3.1 Å
Resolution-Sensitive Ranking: Q-Score

APOF 1.8 Å
APOF 2.3 Å
APOF 3.1 Å
ADH 2.9 Å

Reference model score
Submitted models scores

APOF: 3ajo
ADH: 6nbb
Q-score: novel way to estimate map resolution

Maps from EMDB and Models from PDB

Pintilie et al *Nature Methods* in press
CaBLAM: proxy for a reference model

- CaBLAM compares virtual dihedrals (based on backbone carbonyl-O, $C_{\alpha}$) to PDB statistics
- Especially valuable when carbonyl-O’s are not obvious (>3 Å)
- CaBLAM performed similarly to metrics commonly used in CASP competitions

Cross-correlations between CaBLAM and “vs. Reference Model” metrics scores for all models in the 2019 Model Challenge
2019 Model Challenge Major Outcomes

- Most *ab initio* methods represented in the challenge performed extremely well
- Most fit-to-map metrics are fine for optimization against a single experimental map
- Resolution-sensitive metrics are best for ranking diverse structures in an archive
- CaBLAM is a valuable new tool for evaluating protein backbone conformation issues
## EM Structure Validation Servers

### Map:

<table>
<thead>
<tr>
<th>Overall Shape &amp; Hand</th>
<th>Tilt-Pair</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>pdbe.org/tiltpair</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Resolution by FSC</th>
<th>FSC</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>pdbe.org/FSC</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Local Resolution</th>
<th>3DFSC</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>3dfsc.salk.edu</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Local Resolution</th>
<th>Scipion</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>scipion.cnbcsic.es/m/myresmap#</td>
</tr>
</tbody>
</table>

### Model:

<table>
<thead>
<tr>
<th>Stereochemistry, compare with all PDB structures</th>
<th>wwPDB</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>validate.wwpdb.org</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stereochemistry</th>
<th>Molprobity</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>molprobity.biochem.duke.edu</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nucleic Acid conformation</th>
<th>DNATCO</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>dnatco.org</td>
</tr>
</tbody>
</table>

### Map/Model Fit:

<table>
<thead>
<tr>
<th>“backbone bumpiness”</th>
<th>EMRinger</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>emringer.com  (@UCSF)</td>
</tr>
</tbody>
</table>

See also: [www.emdataresource.org/validation.html](http://www.emdataresource.org/validation.html)
Unified Data Resource for 3DEM

Stanford University/SLAC

Wah Chiu
Greg Pintilie
Andriy Kryshtafovych (UC Davis)

Rutgers University

Catherine Lawson
Helen Berman
Brinda Vallat
Brian Hudson
John Westbrook

European Bioinformatics Institute

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

www.emdataresource.org
References


