Map Resolution   **Map Features**   **Model Geometry**   **Model Fit-to-Map**
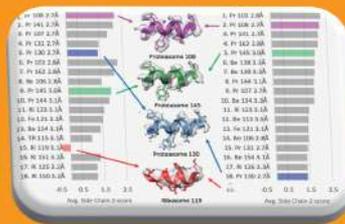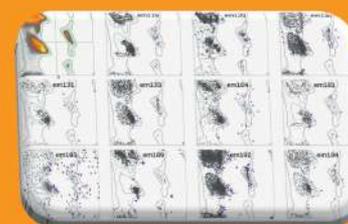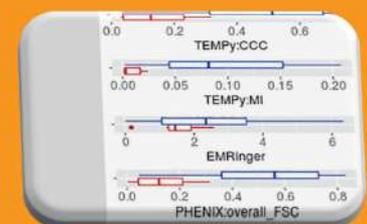
EMDataResource &
Model Challenges

Cathy Lawson

Rutgers University

NYSBC-NCCAT spring 2022 SPA short course –
3/17 lecture @4pm

EMDataResource
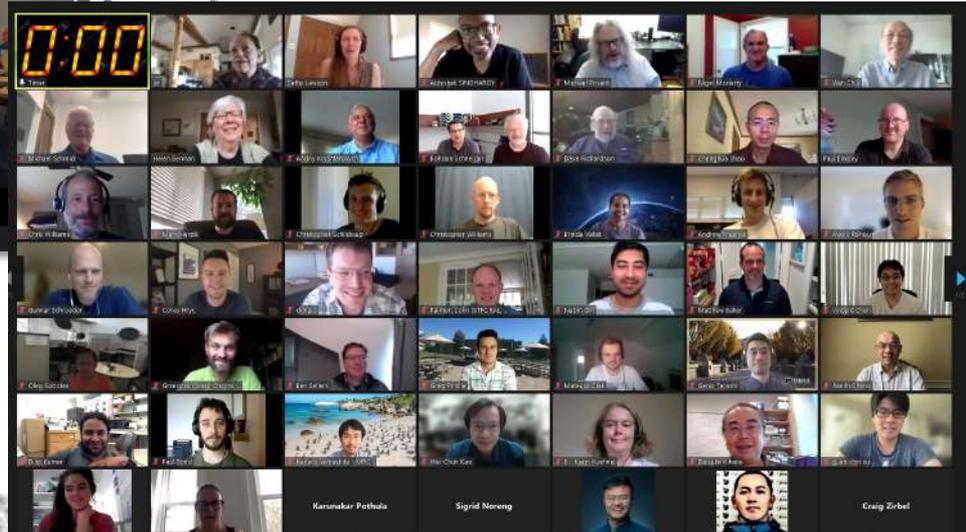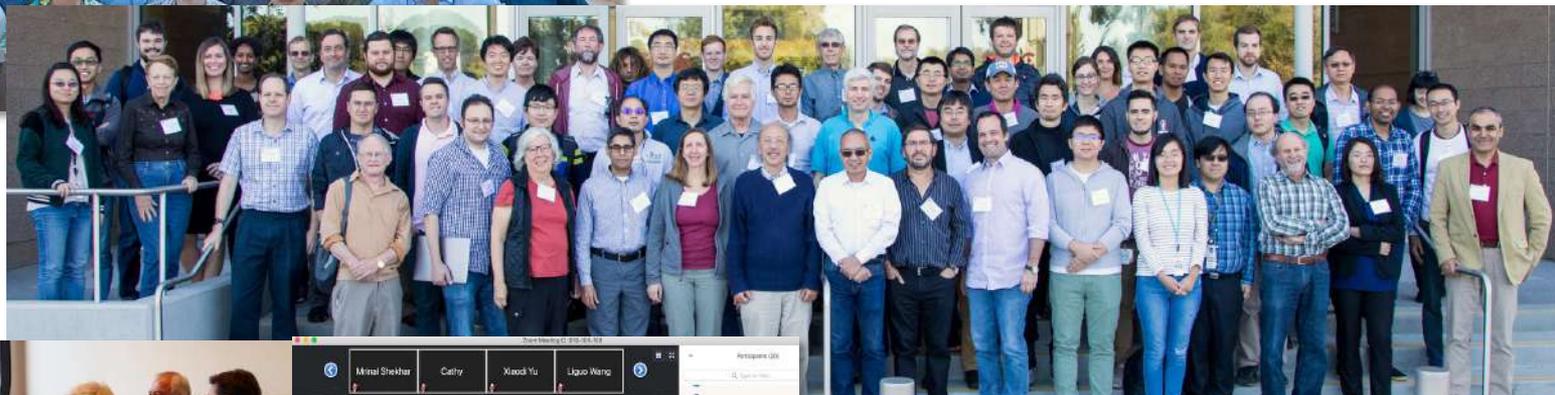Unified Data Resource for 3DEM

EMDataResource (.org)

- Established 2006 under NIH/NIGMS Support to:
  - Develop Data Archives for 3DEM
  - Promote Community Development of Validation and Standards via **Workshops** and **Challenges**

Lawson, Berman & Chiu (2020)
Evolving Data Standards for CryoEM
Struct. Dynamics 10.1063/1.5138589

Cryo-EM Community

Key Early Workshop

2010 Validation Task Force

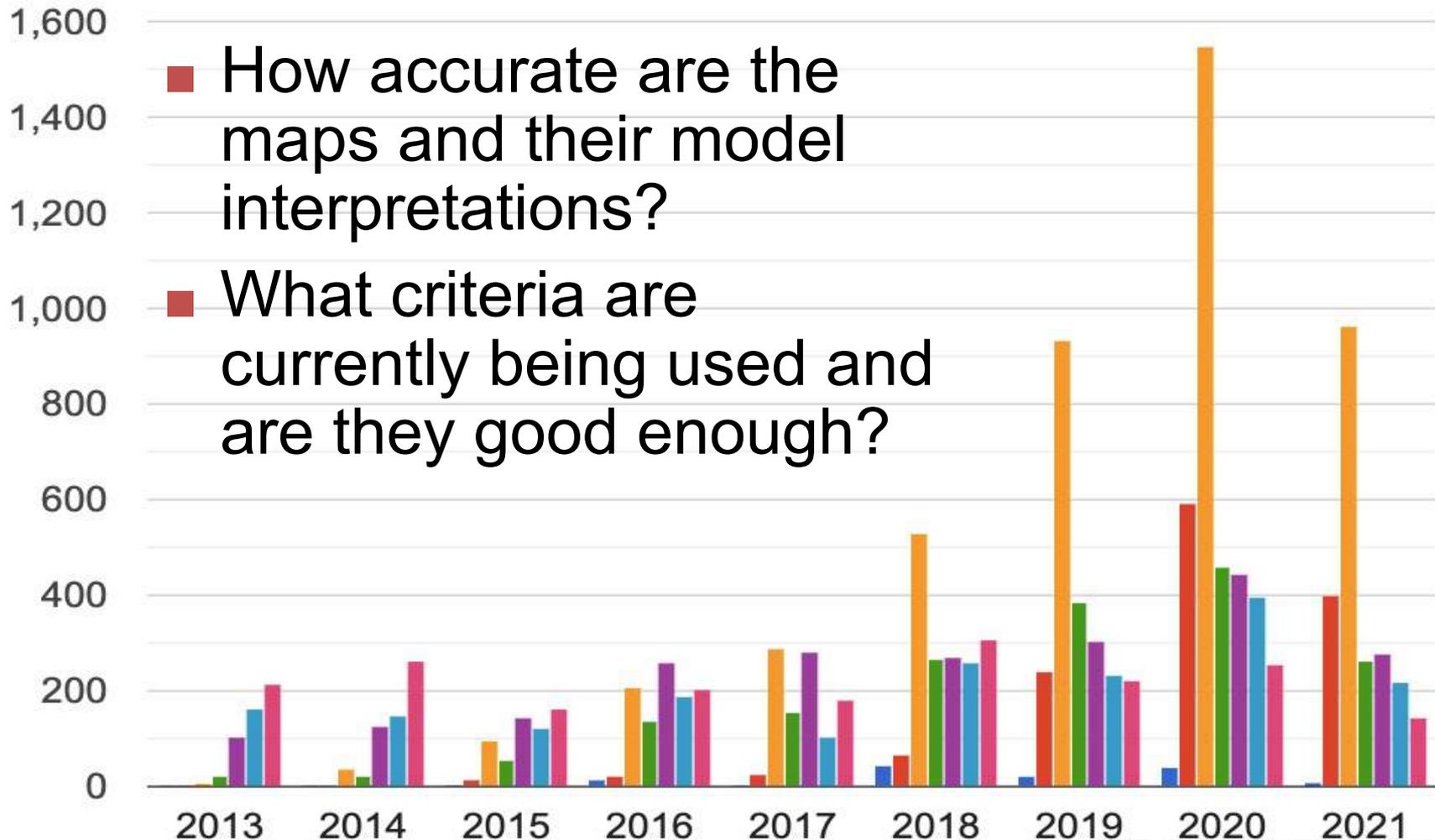❑ Map evaluation by independent half-map FSC
❑ Low-resolution map validation by tilt-pair
❑ Model validation same as for X-ray and NMR
❑ **Call for further development: Map, Model Fit-to-Map Validation**

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

Full workshop list: https://www.emdataresource.org/workshops.html

EMDataResource
Unified Data Resource for 3DEM

Validation in a Changing Landscape

Maps Released by Resolution Range (Å) 2021-07-21

Legend: 0-2, 2-3, 3-4, 4-5, 5-10, 10-20, >20

EMDB Entries Released per Year (y-axis: 0 to 1,600)

x-axis years: 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021

- How accurate are the maps and their model interpretations?
- What criteria are currently being used and are they good enough?

2010 (1ˢᵗ) Model Challenge

13 EMDB maps 2.5-24 Å

❑ Produce best models against selected maps
❑ Explore segmentation, secondary structure detection, rigid body, flexible fitting, *ab initio*

- Established modelling community around a common problem

- Identified critical standardization issues related to data deposition

- Identified key issues to explore in future challenges

130 Models submitted

Biopolymers

September 2012
9 Article Special Issue

The 2010 Cryo-EM Modeling Challenge
Guest Editor: Steven J. Ludtke

View this journal online at wileyonlinelibrary.com

WILEY-BLACKWELL

2016-2017 Map, Model Challenges

7 EMPIAR sets / 12 EMDB maps 2.5-5 Å

- ❑ Produce best maps from raw images
- ❑ Produce best models against maps
- ❑ Compare reconstruction, modeling practices
- ❑ Explore assessment strategies esp. model fit-to-map

- Innovative methods for map and model fit-to-map assessment introduced

- Map quality depended on participant level of experience

- Maps reported as same resolution looked different from each other

- Model quality was "all over the place"

66 maps, 63 models submitted

ELSEVIER

Volume 204, Number 3, December 2018   ISSN 1047-8477

Journal of

Structural Biology

December 2018
19 Article Special Issue

Map Resoluti

Model Geometry

Model Fit-to-Map

The CryoEM Structure Map and Model Challenges

2016-2017 Challenges Recommendations

7 EMPIAR sets / 12 EMDB maps 2.5-5 Å

- ❏ Produce best maps from raw images
- ❏ Produce best models against maps
- ❏ Compare reconstruction, modeling practices
- ❏ Explore assessment strategies esp. model fit-to-map

- Map Resolution by independent half-map FSC: uniform definition + software implementation needed

- Novel model-based methods may be useful for estimating map resolvability

- **Further review of global fit-to-map metrics needed**

66 maps, 63 models submitted

ELSEVIER

Volume 204, Number 3, December 2018   ISSN 1047-8477

Journal of **Structural Biology**

December 2018 19 Article Special Issue

Map Resolution

Model Geometry

TEMPy:CCC

TEMPy:MI

EMRinger

PHENIX:overall_FSC

Model Fit-to-Map

The CryoEM Structure Map and Model Challenges

2019 Model "Metrics" Challenge

4 EMDB maps 1.8-3.1 Å

❑ Produce best models against maps
❑ Explore Model metrics with focus on Fit-to-Map

- cryo-EM maps ≤3 Å with limited conformational flexibility have excellent information content

- *ab initio* methods performed extremely well--though stumbled on ligands/ions

- Limited targets with a resolution series enabled observation of key metrics trends
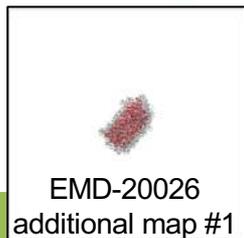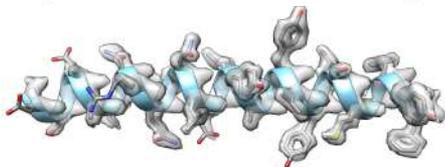
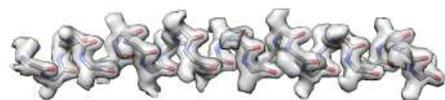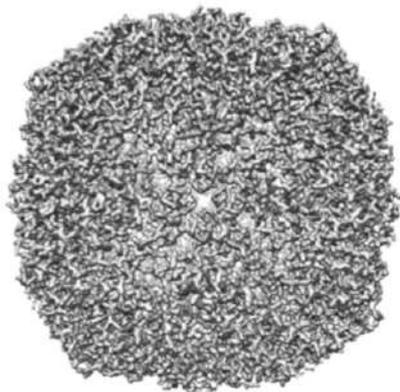**63 models submitted from 13 groups**

nature | methods

**Single Paper Summary
53 participant co-authors
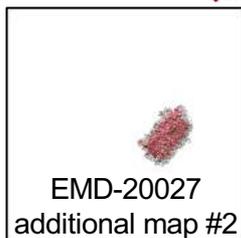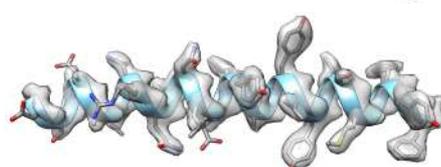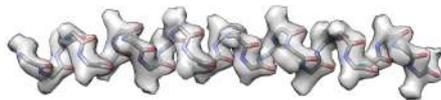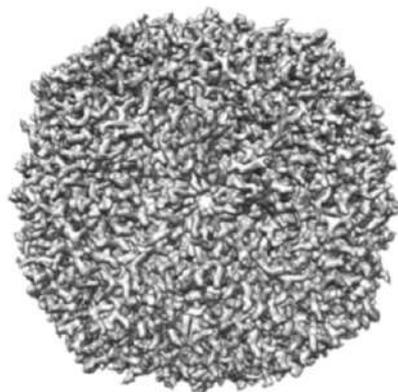Published Feb 2021**
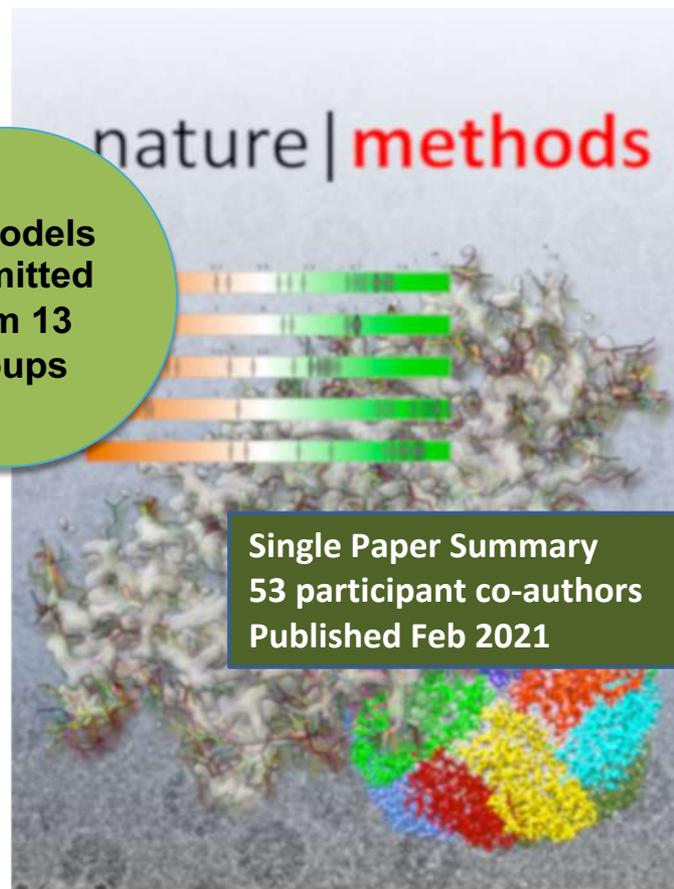
2019 Challenge: Targets

APOF 1.8 Å APOF 2.3 Å APOF 3.1 Å ADH 2.9 Å
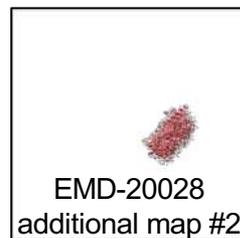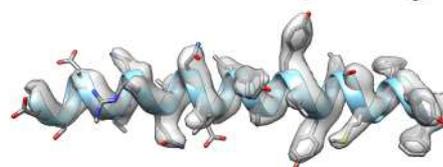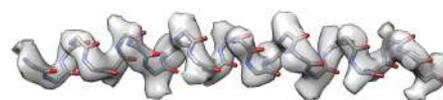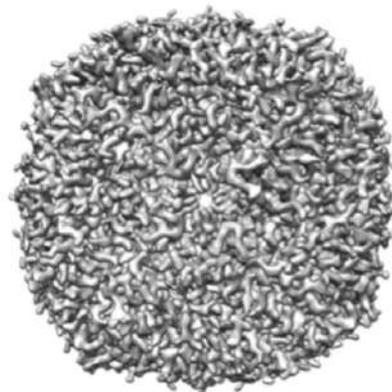
EMD-20026
additional map #1
masked

EMD-20027
additional map #2
masked

EMD-20028
additional map #2
masked
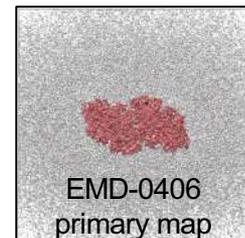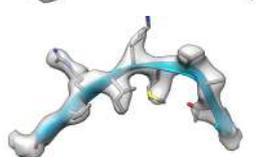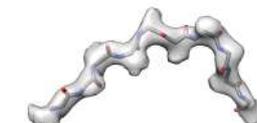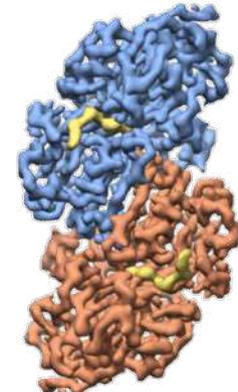
EMD-0406
primary map
unmasked

Fit-to-Map
Analyses ➜

EMDataResource
Unified Data Resource for 3DEM

2019 Challenge: Pipeline

Andriy Kryshtafovych
UC Davis

a — 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

b — SUBMISSIONS

63 models total
51 *ab initio*
12 optimization

13 participating teams from US and Europe

- ab initio, some manual steps, 39
- optimization, fully automated, 10
- ab initio, fully automated, 12
- optimization, some manual steps, 2

c — 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

d — SCORES COMPARISON

Map-Model FSC

T1 APOF TARGET
Submitted Models
(random colors)
vs Reference Model

1/ Resolution, 1/A

T4 ADH TARGET — Model #10_1 ▲ — Model #60_1 ▼ — All Submitted Models ◆

- Q-score
- CaBLAM Conf
- LDDT
- Davis-QA

model-compare.emdataresource.org

EMDataResource
Unified Data Resource for 3DEM

Team ID*, name	Team Members	# Models	Effort Type(s)	Software
10 yu	Xiaodi Yu	4	ab initio+manual	Phenix[20], Buccaneer[37], Chimera[38], Coot[28], Pymol
25 cdmd	Maxim Igaev, Andrea Vaiana, Helmut Grubmüller	4	optimization automated	CDMD[39]
27 kumar	Dilip Kumar	1	ab initio+manual	Phenix, Rosetta[40], Buccaneer, ARP/wARP[41], Coot
28 ccpem	Soon Wen Hoh, Kevin Cowtan, Agnel Praveen Joseph, Colin Palmer, Martyn Winn, Tom Burnley, Mateusz Olek, Paul Bond, Eleanor Dodson	4	ab initio+manual	CCP-EM[42], Refmac[12], Buccaneer, Coot, TEMPy[15-17]
35 phenix	Pavel Afonine, Tom Terwilliger, Li-Wei Hung	4	ab initio+manual	Phenix, Coot
38 fzjuelich	Gunnar Schroeder, Luisa Schaefer	3	optimization automated	Phenix, Chimera, DireX[43], MDFF[44], CNS, Gromacs
41 arpwarp	Grzegorz Chojnowski	8	ab initio automated, ab initio+manual	Refmac, ARP/wARP, Coot
54 kihara	Daisuke Kihara, Genki Terashi	8	ab initio+manual	Rosetta, Mainmast[45], MDFF, Chimera
60 deeptracer	Liguo Wang, Dong Si, Renzhi Cao, Jianlin Cheng, Spencer A. Moritz, Jonas Pfab, Tianqi Wu, Jie Hou	10	ab initio automated, ab initio+manual	Cascaded-CNN[46], Chimera
73 singharoy	Mrinal Shekhar, Genki Terashi, Sumit Mittal, Daipayan Sarkar, Daisuke Kihara, Ken Dill, Alberto Perez, Abishek Singharoy	5	ab initio+manual, optimization automated	reMDFF[47], MELD[48], VMD, Chimera, Mainmast
82 rosetta	Frank DiMaio, Dan Farrell	8	ab initio automated, ab initio+manual	Rosetta, Chimera
90 mbaker	Matt Baker	2	ab initio+manual	Pathwalker[49], Phenix, Chimera, Coot
91 chiu	Greg Pintilie, Wah Chiu	2	optimization+manual	Phenix, Chimera, Coot

Fit to Map	Correlation	Full map density	TEMPY CCC | PHENIX CCbox
		Density within a mask	TEMPy CCC_OV | Segment Mander's Overlap (SMOC) PHENIX CCpeaks | CCvol | CCmask
		Density-derived functions	TEMPY Mutual Information(MI) | MI_OV | Laplacian (LAP)
		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	REFMAC5 FSCavg
	Atom Inclusion		TEMPy Envelope | EMDB Atom Inclusion (AI_all)
	Rotamer		EMRinger Z-score protein C_γ atom paths around $\chi1$
Coordinates Only	Configuration		PHENIX Bond | Bond angle | Chirality | Planarity | Dihedral
	Conformation	Backbone	CaBLAM Cα-trace Cα-only virtual dihedrals CaBLAM Conformation Cα and CO-containing virtual dihedrals MOLPROBITY Ramachandran
		Sidechain	MOLPROBITY Rotamer
	Clashes		MOLPROBITY Clashscore
	Energy		PROQ3 energy and predicted features
Comparison with 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
	Distances	Per chain	LDDT Local difference distance test
		All chains	OPENSTRUCT oligomeric LDDT | weighted oligomeric LDDT
Comparison among Models*	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

Fit-to-Map: Score Correlations, all 63 models
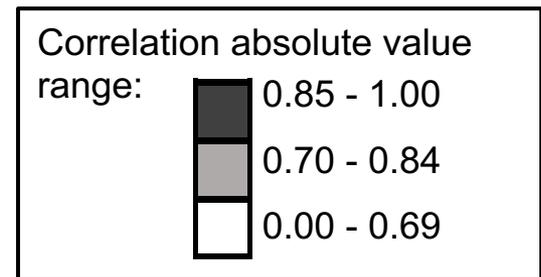
Fit-to-Map: Cluster 1 vs Cluster 2

Cluster 1
Correlation Score:
Map resolution input
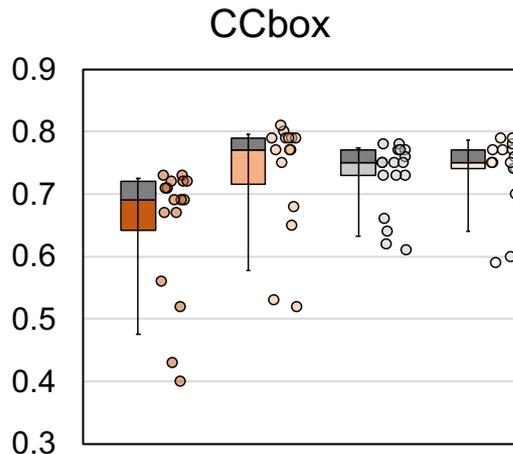
Cluster 2
Correlation Score:
Fixed Gaussian profile

CCbox

Qscore

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

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

CCbox

Qscore

EMDataResource
Unified Data Resource for 3DEM
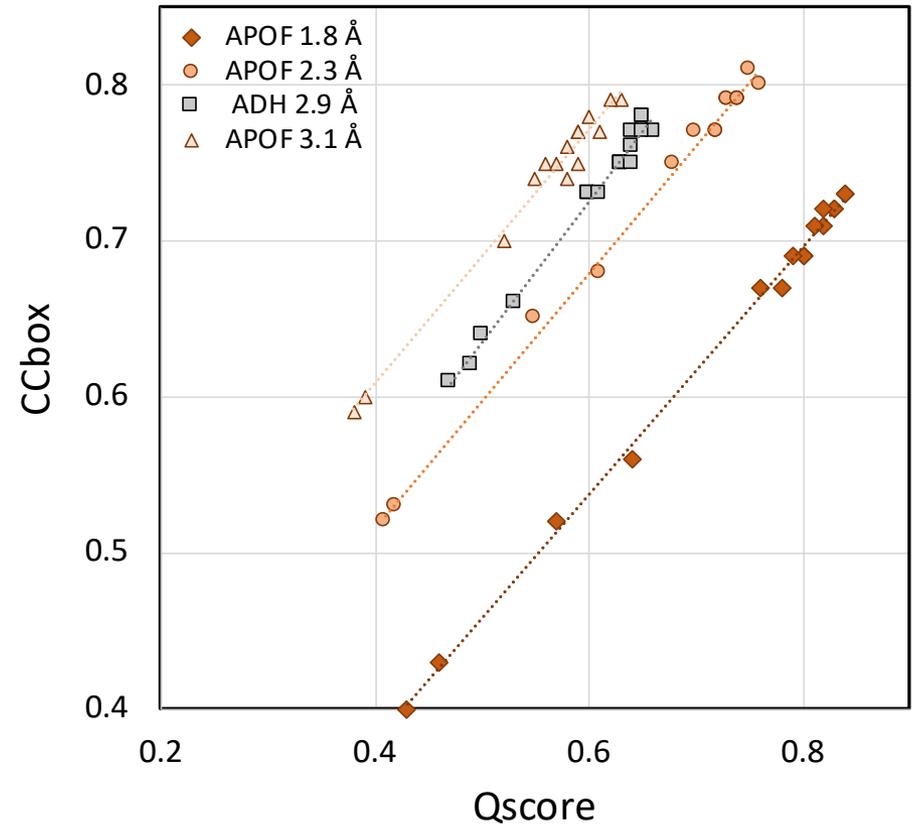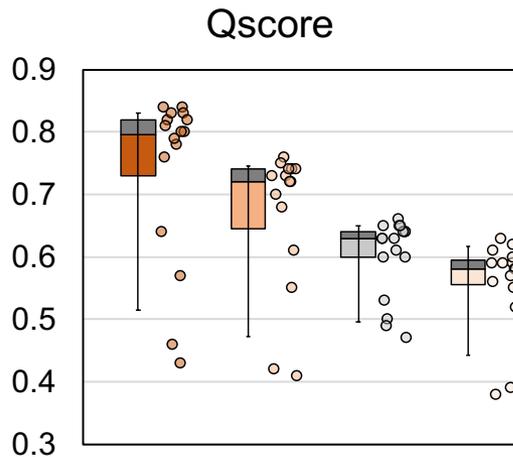
2019 Challenge: Recommendations

❑ **For researchers optimizing a model against a single map at near atomic resolution (2-3 Å)**, nearly all evaluated Fit-to-Map metrics behaved equivalently (exception: TEMPy ENV is more appropriate for lower resolution)

❑ **For archive-wide ranking**, the ideal Fit-to-Map metric will be insensitive to background noise, will not require estimated parameters that affect score value (resolution limit, threshold), and will yield overall better scores for maps with higher-resolution features. (Cluster 2!)

❑ **CaBLAM and Molprobity cis-peptide detection** are valuable tools for evaluating protein backbone conformation, especially when backbone carbonyl oxygens are not resolved (2.5-4 Å).

# 2021 Ligand Model Challenge

3 EMDB maps 1.9-2.5 Å

❑ Focus on ligands/waters/ions built into cryoEM maps
❑ Is manual intervention needed? How to validate?

- wide spectrum of modeling workflows represented

- novel assessments to evaluate ligand pose, as well as traditional fit-to-map scores

61 models submitted from 17 groups

Single Summary Paper Planned

EMDataResource

Unique Modeled Ligands in PDB EM Structures
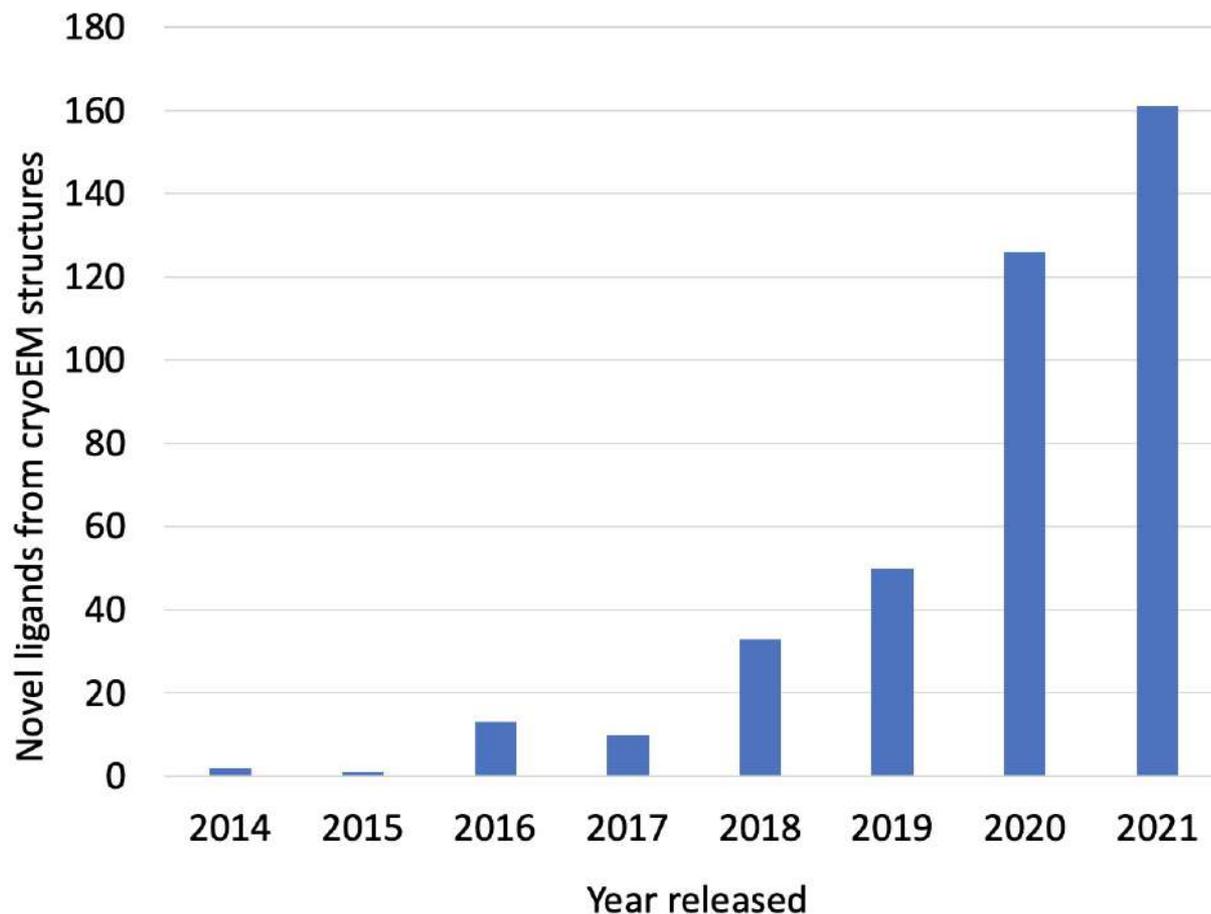
Targets

| Specimen/Reference	Target Map	EMDB	Symmetry	Reference Model	Model Components	Resol (Å)
E. coli β-Galactosidase (**EMPIAR 10061** reprocessed) **doi**		**EMD-7770**	D2	**6cvm**	Protein (1 unique chain, A,B,C,D), Water, Mg, Na ions, PETG: 2-phenylethyl 1-thio-beta-D-galactopyranoside (**PTQ**)	1.9
SARS-CoV-2 RNA-dependent RNA polymerase **doi**		**EMD-30210**	P1	**7bv2**	Protein (3 unique chains A|B|C), RNA (2 unique chains D|E), Water, Zn, Mg ions, pyrophosphate, **F86 (remdesivir**, covalent inhibitor)	2.5
SARS-CoV-2 ORF3a putative ion channel in nanodisc **doi**		**EMD-22898**	C2	**7kjr**	Protein (2 unique chains A,B|C,D), Water, 1,2-Dioleoyl-sn-glycero-3-phosphoethanolamine (**PEE**)	2.1

Assessment Methods Presented

- Model-Compare Pipeline (multiple metrics) – Andriy Kryshtafovych

- Mogul (wwPDB Validation Report) – Chenghua Shao

- NtC nucleic acid geometry, solvation– Bohdan Schneider and Jiri Cerny

- Probescore, Undowser, "eyes" -- Jane Richardson

- Hot spot Chem Environment -- Chris Williams

- Q-scores -- Greg Pintilie

- Ligand Strain Analysis --Ben Sellers

EMDataResource
Unified Data Resource for 3DEM

bGal PETG Ligand-Only Q-score

Best (0.69)

6cvm

Worst (0.21)
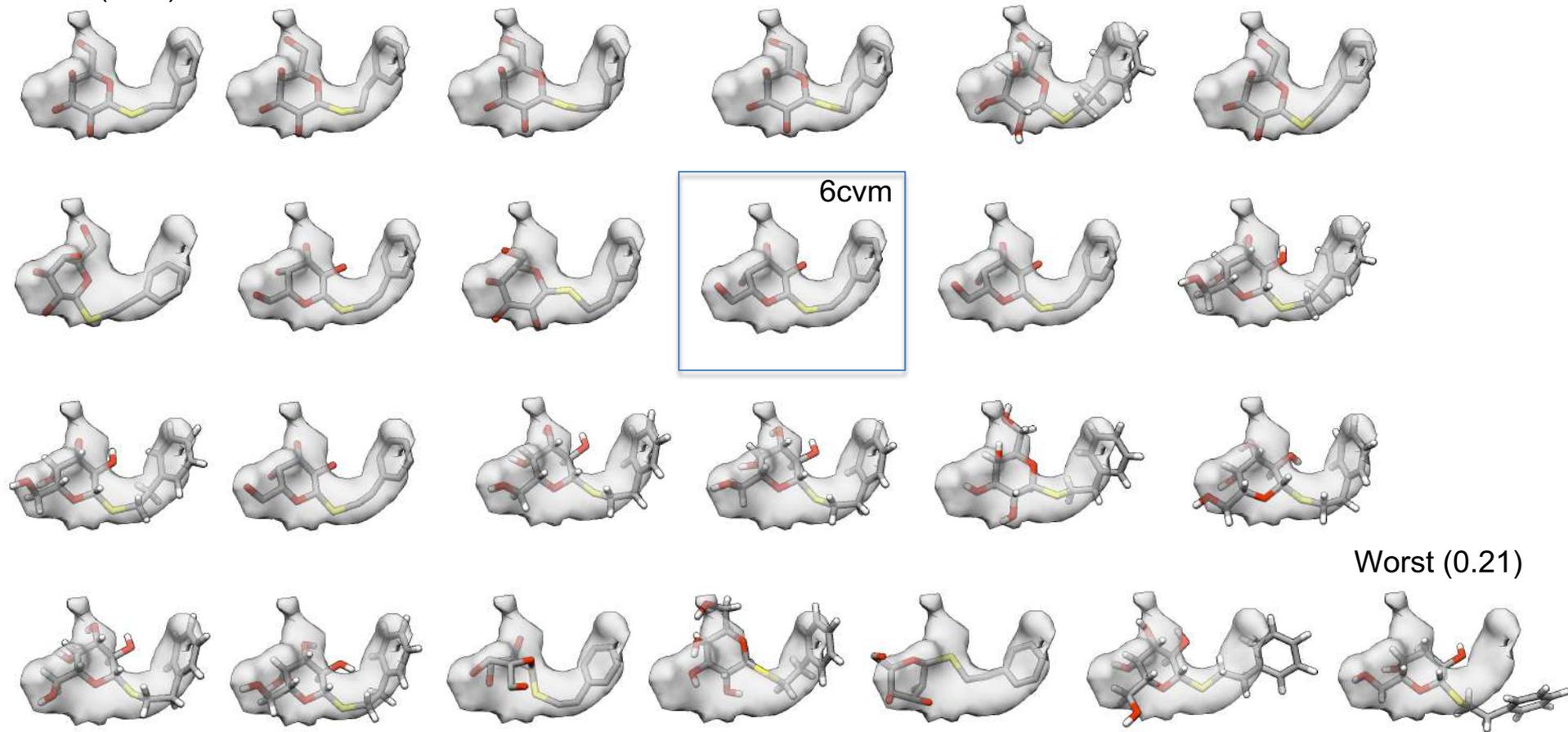
Source: Greg Pintilie

Some of the observations/issues raised…

- Local map resolution may not match overall resolution

- Incomplete occupancy

- Multiple conformations/poses

- Ligand may be flexible/disordered in regions

- Radiation damage

- Difference maps useful for interpretation

- B-factor type parameter needed beyond x,y,z to optimize model-based map

EMDataResource

Future Challenges

- Ligands/Ions
- Lower Resolution
- Membrane Proteins
- Nucleic Acids
- Models derived from Tomograms

Please Join Us!

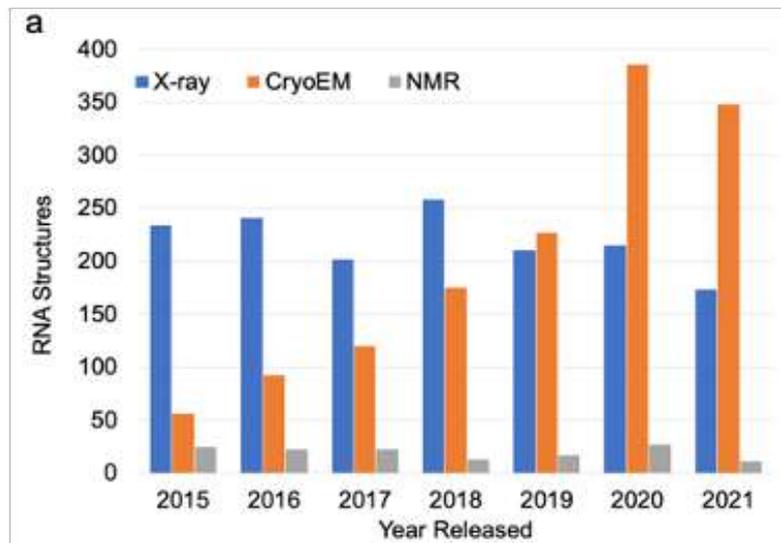Use Previous Challenge maps/models to test your software:
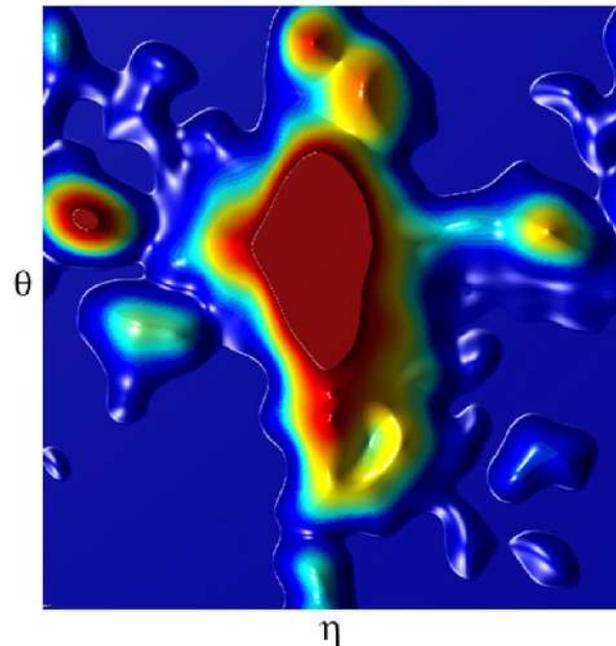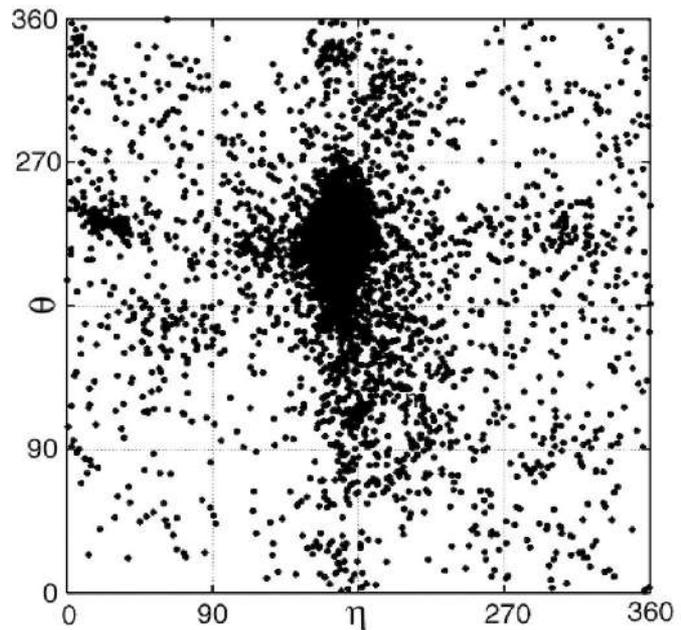https://zenodo.org/communities/3dcryoemchallenges

EMDataResource
Unified Data Resource for 3DEM

RNA Structures by Method 2015-2021
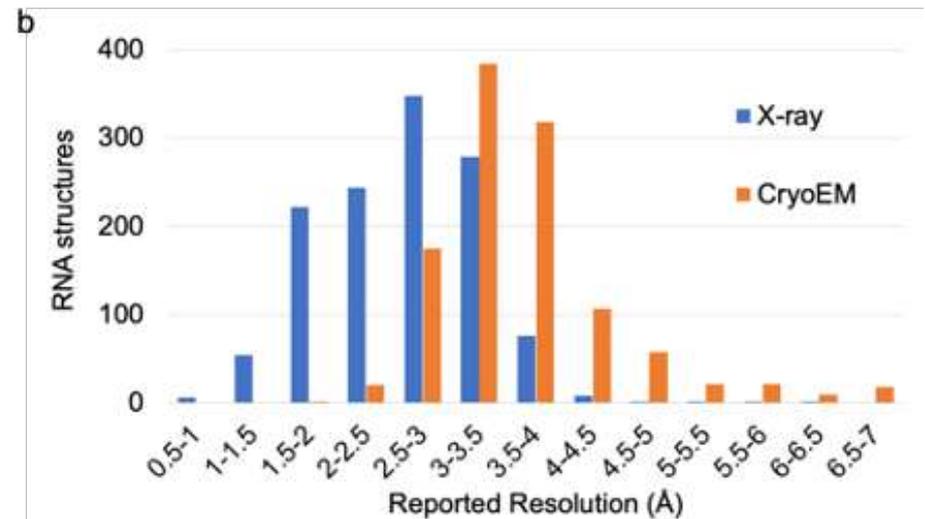
η-θ Plots
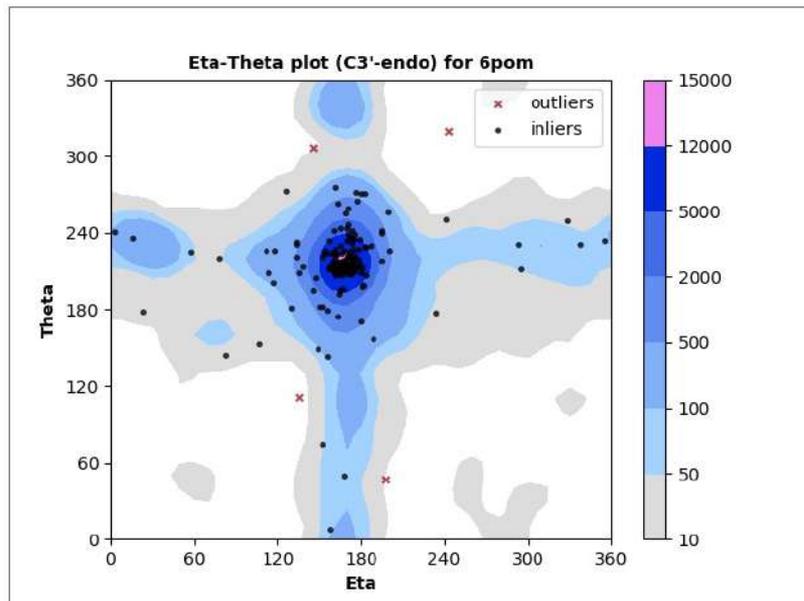

Data windowing (kernel smoothing or kernel density estimation) was employed using Blackman window function. It transforms the η–θ scatter plot into a density plot.
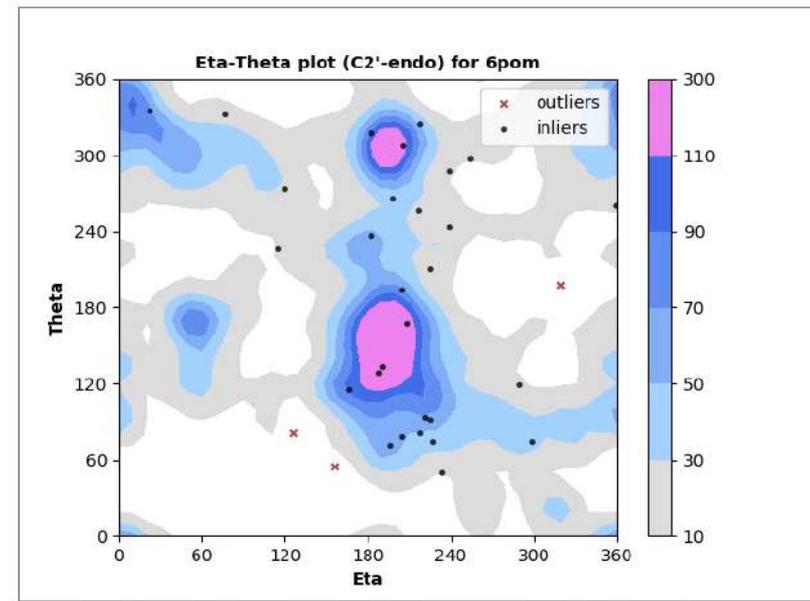
Wadley LM et al., 2007. J. Mol. Biol. 372, 942-957

EMDataResource
Unified Data Resource for 3DEM

η-θ Plot: Riboswitch with tRNA (6POM)

186 C3'-endo RNA residues with 4 outliers (2.15 %) from Eta-Theta plot

32 C2'-endo RNA residues with 3 outliers (9.38 %) from Eta-Theta plot

https://ptp.emdataresource.org

Unified Data Resource for 3DEM

Stanford University

Wah Chiu (PI)
Greg Pintilie
Mike Schmid
UC Davis:
Andriy Kryshtafovych

(Baylor Coll. Med.)
Steven Ludtke
Matt Baker
Corey Hryc
Ian Rees

Rutgers University

Cathy Lawson
Helen Berman
Brinda Vallat

John Westbrook
Brian Hudson
Batsal Devkota
Raul Sala
Chunxiao Bi

European Bioinformatics Institute

Ardan Patwardhan
Gerard Kleywegt
Sanja Abbott
Zhe Wang
Kim Henrick
Richard Newman
Christoph Best
Glen van Ginkel
Eduardo Sanz-Garcia
Ingvar Lagerstedt