# Introductions and Overview of the wwPDB

**Jeff Hoch**

WORLDWIDE PDB PROTEIN DATA BANK

wwpdb.org

# Welcome

**On behalf of the wwPDB/EMDB Principal Investigators**

- BMRB: Jeffrey C. Hoch
- RCSB PDB: Stephen K. Burley
- PDBe: Sameer Velankar
- PDBj : Genji Kurisu

- EMDB: Ardan Patwardhan (apology)

# Introductions

- Chair : Peter Rosenthal
- Co-Chair: Art Edison

wwPDB Advisory Committee Members

- RCSB PDB: Paul Adams and Kirk L. Clark
- PDBe: Arwen Pearson and Susan Lea
- PDBj: Daisuke Kohda and Masaki Yamamoto
- BMRB: Art Edison and Angela Gronenborn
- EMDB: Corinne Smith and Juha Huiskonen

# Introductions (cont.)

Associate Member candidates

- China: Wenqing Xu and Zhipu Luo
- India: Debasisa Mohanty

Institutional Representative

- Gerard Kleywegt (EMBL-EBI)

IUCr Representative

- Edward Baker

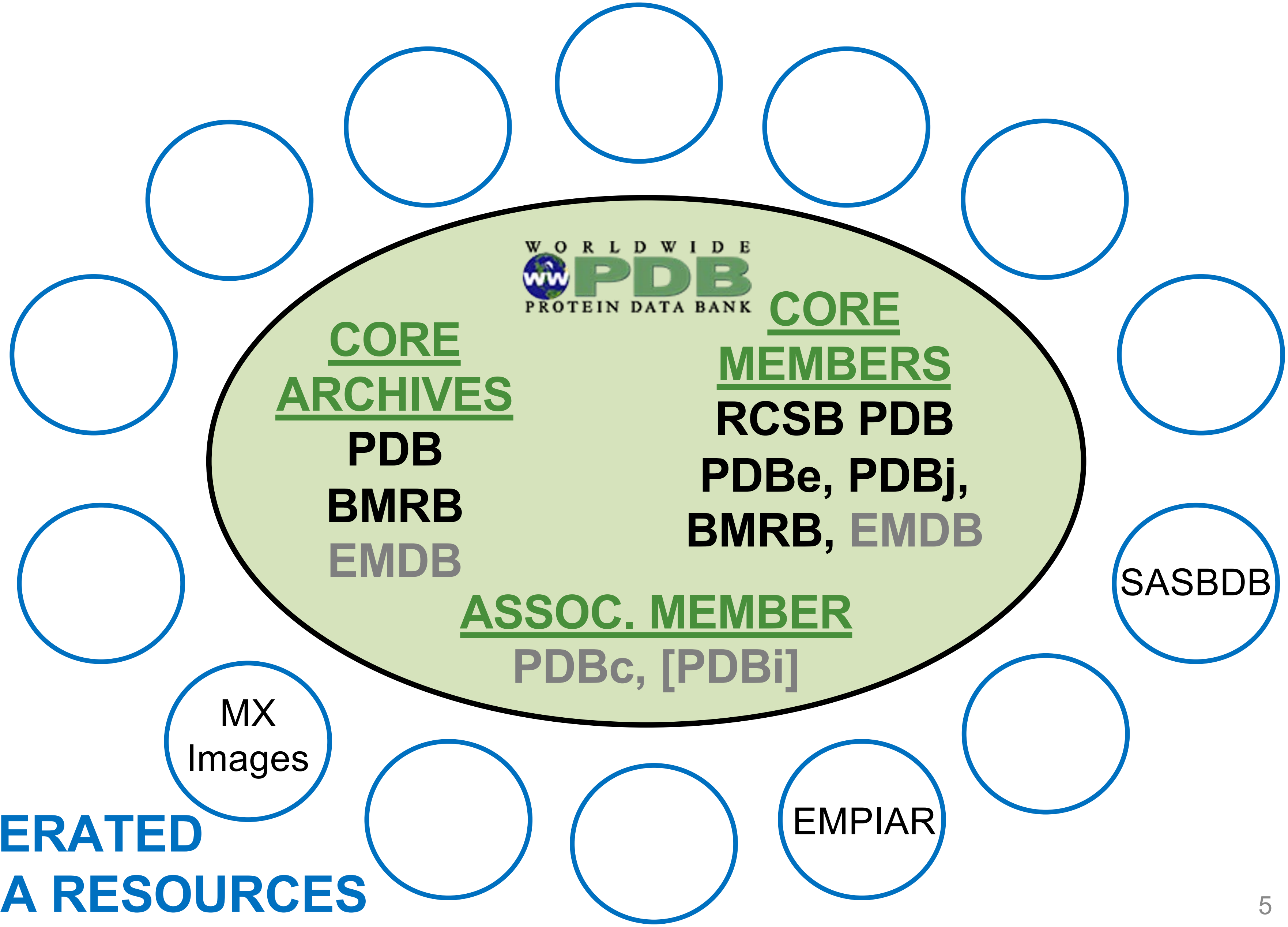
ISMAR Representative

- Andy Byrd

3DEM Representative

- Peter Rosenthal (concurrent)

# wwPDB Future Architecture



CORE ARCHIVES

PDB

BMRB

EMDB

CORE MEMBERS

RCSB PDB

PDBe, PDBj,

BMRB, EMDB

ASSOC. MEMBER

PDBc, [PDBi]

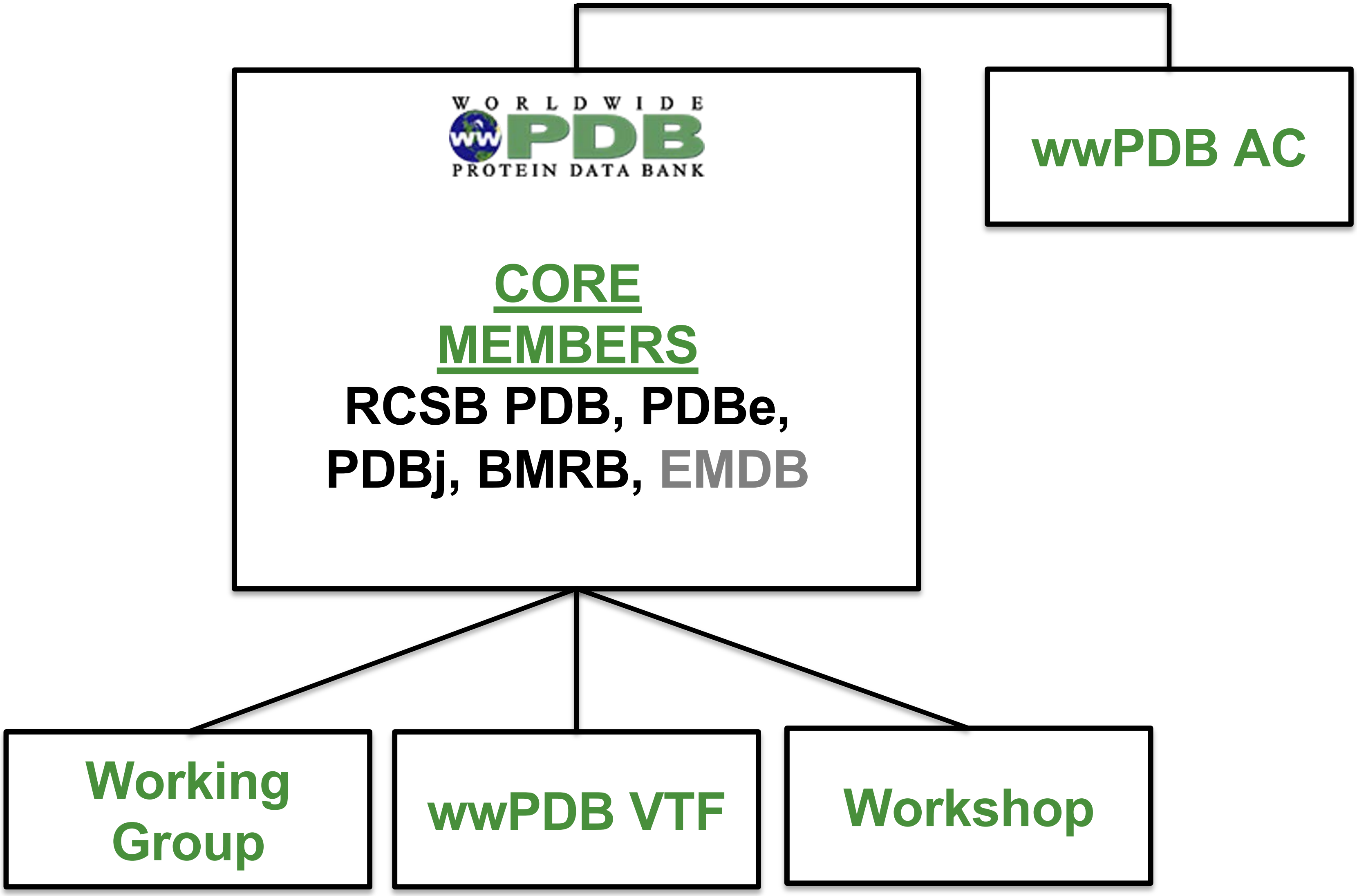
SASBDB

MX Images

EMPIAR

FEDERATED
DATA RESOURCES

# Current wwPDB Organization



WORLDWIDE wwPDB PROTEIN DATA BANK

CORE MEMBERS

**RCSB PDB, PDBe, PDBj, BMRB, EMDB**

wwPDB AC

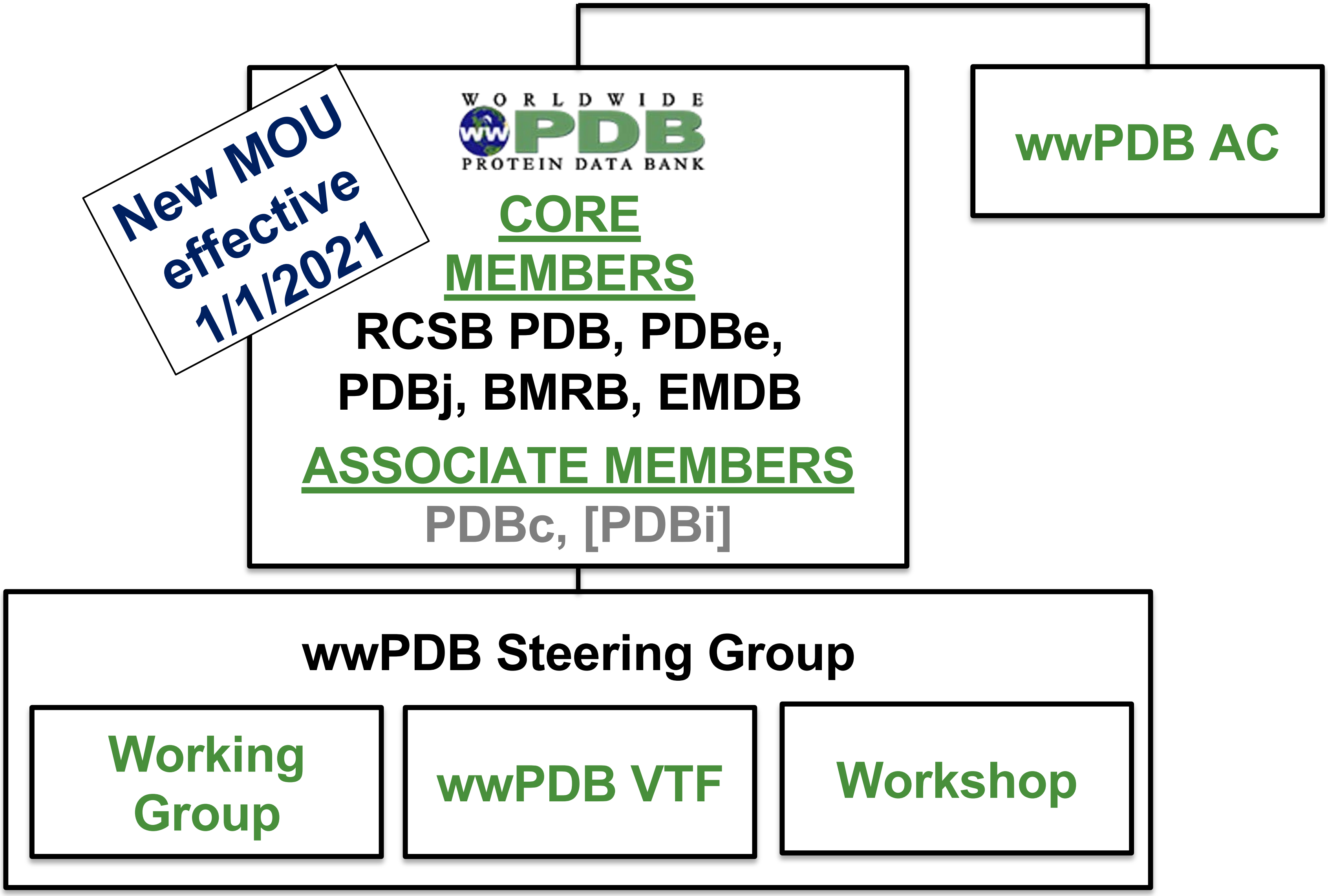
Working Group

wwPDB VTF

Workshop

# New wwPDB Organization

Ref. Appendix 1



New MOU effective 1/1/2021

WORLDWIDE PDB PROTEIN DATA BANK

CORE MEMBERS

RCSB PDB, PDBe, PDBj, BMRB, EMDB

ASSOCIATE MEMBERS

PDBc, [PDBi]

wwPDB AC

wwPDB Steering Group

Working Group

wwPDB VTF

Workshop

# Developments since 2019 Meeting I

**wwPDB**

- Continued enhancement of OneDep system for deposition/validation/biocuration of MX, NMR, and 3DEM
- Continued growth in 3DEM structure depositions and engagement with the 3DEM community
- Continued depositions to PDB-Dev for I/HM structures
- Presented at the Biophysical Society I/HM workshop (March 2019). Manuscript submitted
- Workshop on improving deposition and validation of single-particle EM data (January 2020)
- Finalizing the new MOU including EMDB

# Developments since 2019 Meeting II

**PDB Core Archive**

- OneDep upgraded to support remote operation
- Increased activity across the board resulting from Covid-19 pandemic
  - Projecting 15,224 depositions for calendar 2020 (13,377 depositions in 2019)
  - Increased communication from depositors
  - Near 100% compliance on voluntary immediate release of Covid-19 entries
  - 371 Covid-19 related entries as of Sept 8; 838 coronavirus-related entires

# Developments since 2019 Meeting III

## BMRB Core Archive I

- NMR-STAR dictionary enlarged with tags for unassigned coupling constants, updated enumerations for experiments including SSNMR
- Testing of pipeline to calculate structures using X-PLOR NIH with NMR-STAR as input file complete
- Majority of source code and NMR-STAR dictionary migrated to GitHub increases FAIRNESS
- New data visualizations added to entry summary pages
- BMRBdep now in production mode (449 depositions)
  - OneDep now employing PyNMR-Star to parse depositions
- ADIT-NMR decommissioned
- BMRbig conceived and beta deployed
- Graphic design for website redesign completed

# Developments since 2019 Meeting IV

## BMRB Core Archive II

- New API endpoints developed to support UNIPROT links
- Restraint validation package integrated into OneDep and testing underway
- Refactoring and containerization of multiple services improves efficiency and robustness
- NIH R01 grant migrated to UConn
- NIH U24 proposal submitted
- Visits to BMRB Eminent Community Champions:
  - Julie Forman-Kay, Lewis Kay, Mitsu Ikura, Cheryl Arrowsmith
  - Jane Dyson, Peter Wright

# Developments since 2019 Meeting V

**EMDB Core Archive**

- Development of EMDB Policies and Processing Procedures document
- Development of EMDB validation reports

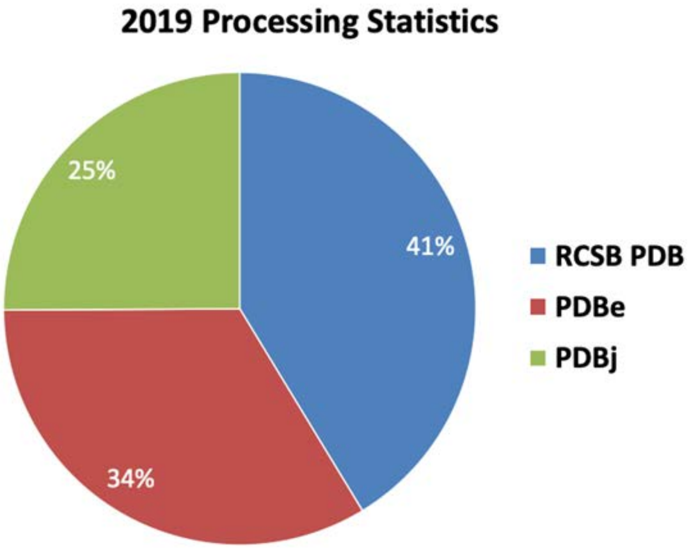
**Individual wwPDB partner sites**

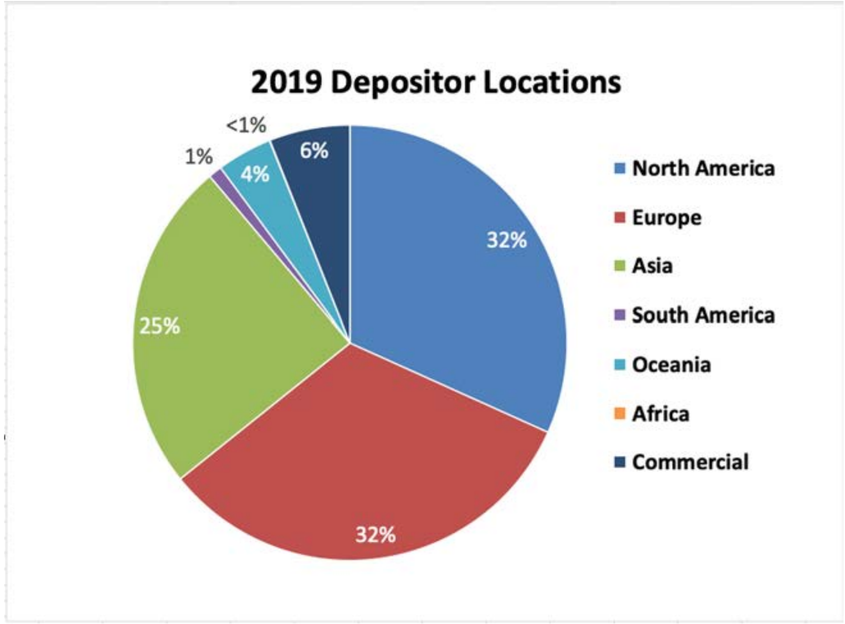
- RCSB PDB and PDBe received a joint NSF/BBSRC grant (3 years duration) to support development of the Next Generation PDB Archive (presented at 2019 wwPDB AC meeting)
- PDBe/RCSB PDB Mol* collaboration continues to go well

# PDB Core Archive Depositions

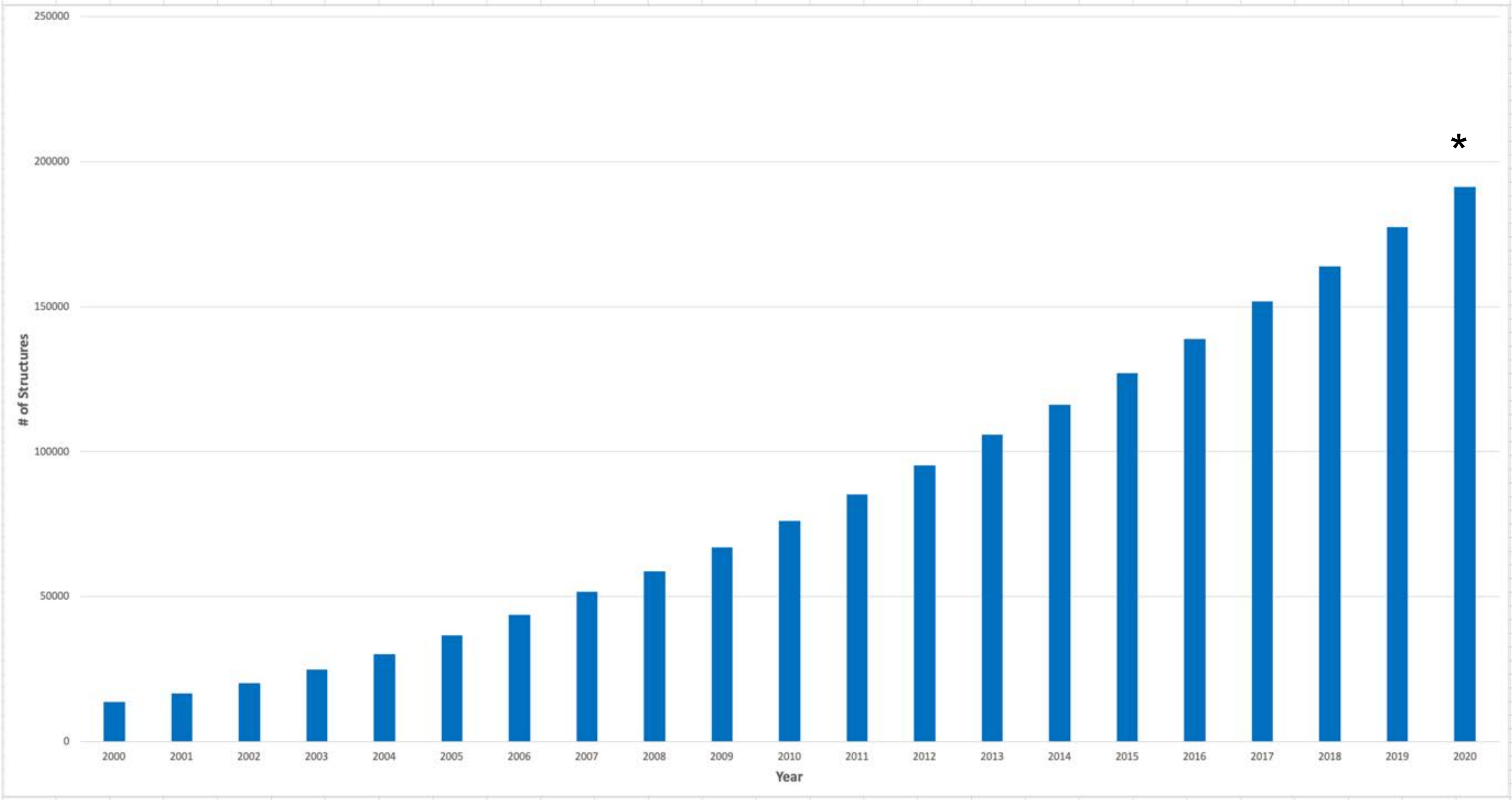
- 13,377 depositions in 2019 (~10% increase).
- Rapid growth in 3DEM.
  - Exceeded NMR depositions
  - Nearly doubled since 2018

| Method | 2019 Depositions | 2018 Depositions |
|---|---|---|
| MX | 10969 (81.9%) | 10594 (87.0%) |
| NMR | 403 (3.0%) | 418 (3.4%) |
| 3DEM | **1996 (14.9%)** | **1140 (9.4%)** |
| Other | 24 (0.2%) | 27 (0.2%) |



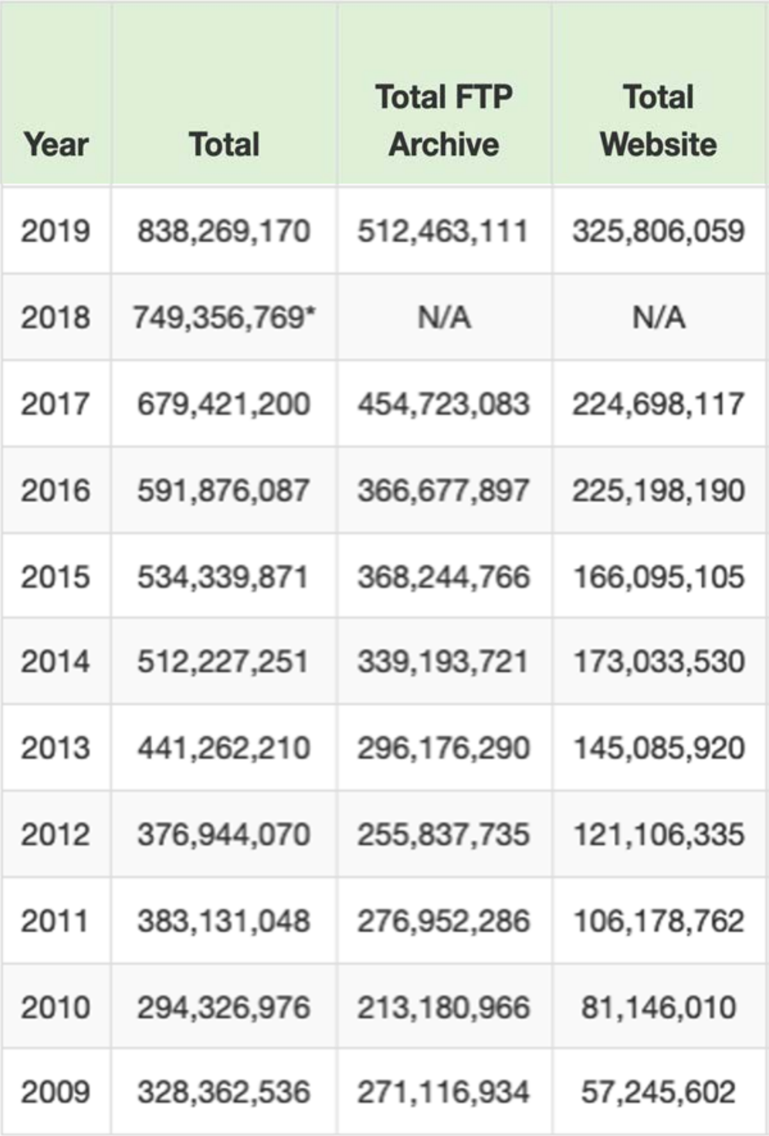


# PDB Core Archive Growth



* As of 1 Sep 2020

# PDB Core Archive Downloads



| Year | Total | Total FTP Archive | Total Website |
|---|---|---|---|
| 2019 | 838,269,170 | 512,463,111 | 325,806,059 |
| 2018 | 749,356,769* | N/A | N/A |
| 2017 | 679,421,200 | 454,723,083 | 224,698,117 |
| 2016 | 591,876,087 | 366,677,897 | 225,198,190 |
| 2015 | 534,339,871 | 368,244,766 | 166,095,105 |
| 2014 | 512,227,251 | 339,193,721 | 173,033,530 |
| 2013 | 441,262,210 | 296,176,290 | 145,085,920 |
| 2012 | 376,944,070 | 255,837,735 | 121,106,335 |
| 2011 | 383,131,048 | 276,952,286 | 106,178,762 |
| 2010 | 294,326,976 | 213,180,966 | 81,146,010 |
| 2009 | 328,362,536 | 271,116,934 | 57,245,602 |

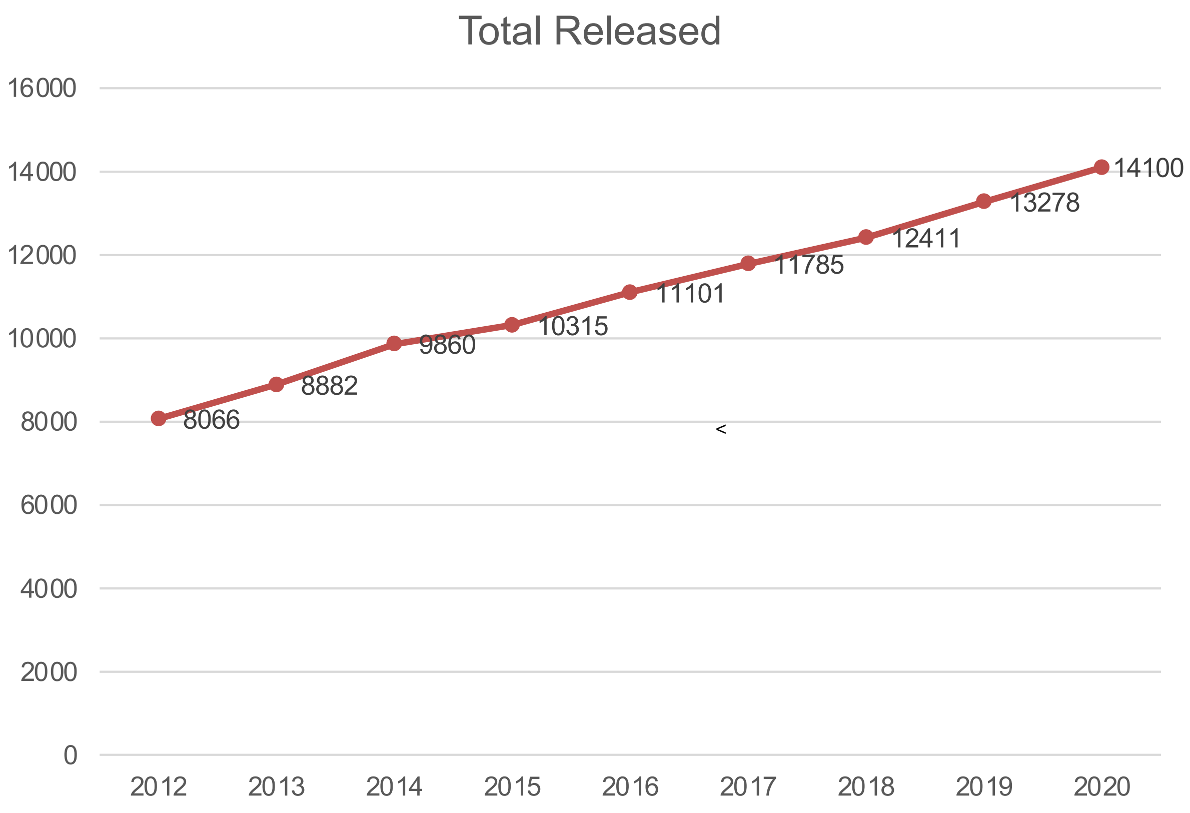
More than 2 million/day!

N.B.: Some 2018 data estimated due to GDPR.

**Geographic Origins of FTP downloads; 2012-2015**

# BMRB Core Archive Growth

- BMRB has released 595 new entries so far in 2020 (235 via OneDep)
- Total released entries estimated to reach ~14,100 by the end of 2020.



# BMRB Core Archive Growth

## Total Released Entries

| Year | Total released | Yearly increase | Structures | Yearly increase | Non-structures | Yearly increase |
|---|---|---|---|---|---|---|
| 2012 | 8068 | 814 | 3953 | 536 | 4115 | 278 |
| 2013 | 8886 | 818 | 4524 | 571 | 4362 | 247 |
| 2014 | 9867 | 981 | 5182 | 658 | 4685 | 323 |
| 2015 | 10322 | 455 | 5481 | 299 | 4841 | 156 |
| 2016 | 11112 | 790 | 5977 | 496 | 5135 | 294 |
| 2017 | 11803 | 691 | 6395 | 418 | 5408 | 273 |
| 2018 | 12438 | 635 | 6666 | 271 | 5772 | 364 |
| 2019 | 13728 | 867 | 7147 | 491 | 6131 | 376 |

# BMRB Core Archive Growth

## Internet Server Traffic (Website) – All Mirrors*

| Year | Server requests | Page requests | File requests | Distinct hosts served | Total data transferred |
|---|---|---|---|---|---|
| 2012 | 34,371,708 | 9,147,444 | 3,204,767 | 310,043 | 23.4 TB |
| 2013 | 40,371,342 | 7,871,583 | 3,262,360 | 350,660 | 20.7 TB |
| 2014 | 33,015,619 | 7,762,480 | 2,296,483 | 391,574 | 27.8 TB |
| 2015 | 28,726,994 | 4,758,270 | 2,066,640 | 450,482 | 27.5 TB |
| 2016 | 36,418,752 | 6,637,758 | 3,301,130 | 458,671 | 29.3 TB |
| 2017 | 63,475,707 | 17,058,266 | 6,272,421 | 340,175 | 17.1 TB |
| 2018 | 75,233,603 | 15,444,841 | 11,508,248 | 440,728 | 15.5 TB |
| 2019 | 77,590,580 | 39,664,896 | 4,155,929 | 575,809 | 15.5 TB |

~300K/day server and page requests

- BMRB has mirror sites in Italy and Japan, and PDBj-BMRB branch for deposition
- Updates to accounting methods resulted in slight changes to historical data from previous reports

# BMRB Core Archive Growth

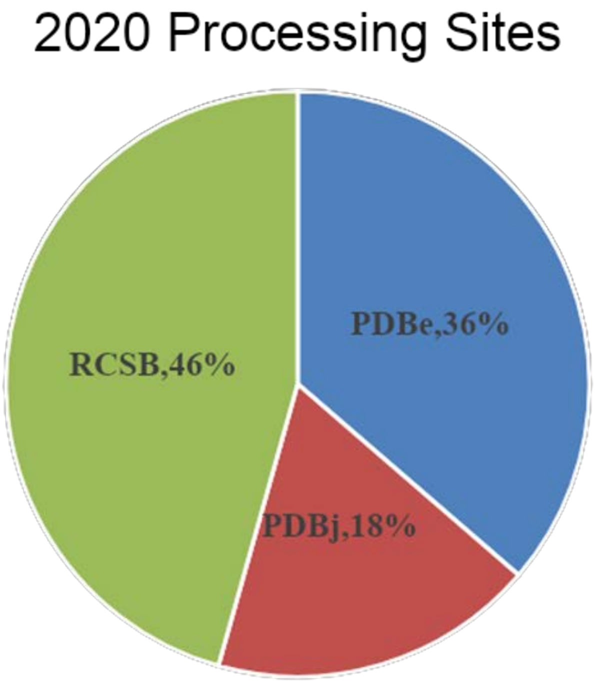
## Internet Server Traffic (FTP Servers) – All Mirrors*

| Year | Server requests | Distinct files requested | Distinct hosts served | Total data transferred |
|---|---|---|---|---|
| 2012 | 2,058,066 | 1,597,183 | 5,037 | 1.1 TB |
| 2013 | 2,018,662 | 1,503,932 | 5,494 | 1.4 TB |
| 2014 | 1,991,174 | 1,486,165 | 4,930 | 1.6 TB |
| 2015 | 2,185,255 | 1,655,143 | 3,915 | 0.9 TB |
| 2016 | 5,704,287 | 1,722,143 | 5,956 | 1.7 TB |
| 2017 | 4,862,305 | 2,335,675 | 4,226 | 4.6 TB |
| 2018 | 4,715,647 | 2,788,527 | 3,866 | 2.0 TB |
| 2019 | 4,845,421 | 2,423,941 | 3,908 | 5.5 TB |

*Updates to accounting methods resulted in changes to historical data from previous reports

# EMDB Core Archive Depositions

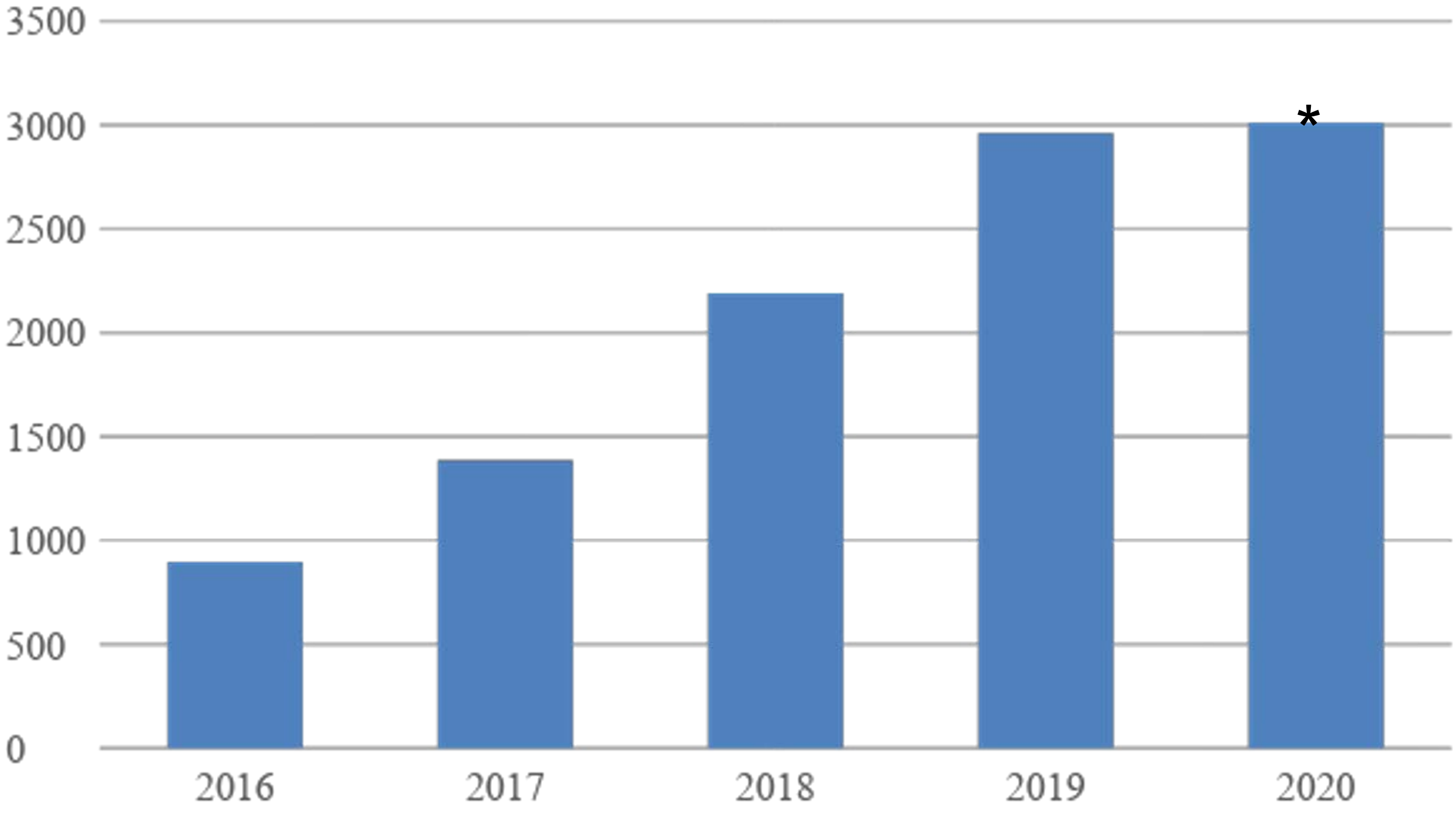
- Over 10,000 EMDB entries
- On track for ~4000 3DEM depositions in 2020.
- 1728 out of 3012 have PDB entries in 2020.



| Processing Site | 2019 Depositions | 2020 Depositions* |
|---|---|---|
| PDBj | 496 | 544 |
| PDBe | 1064 | 1094 |
| RCSB | 1400 | 1374 |
| **Total** | 2960 | 3012 |

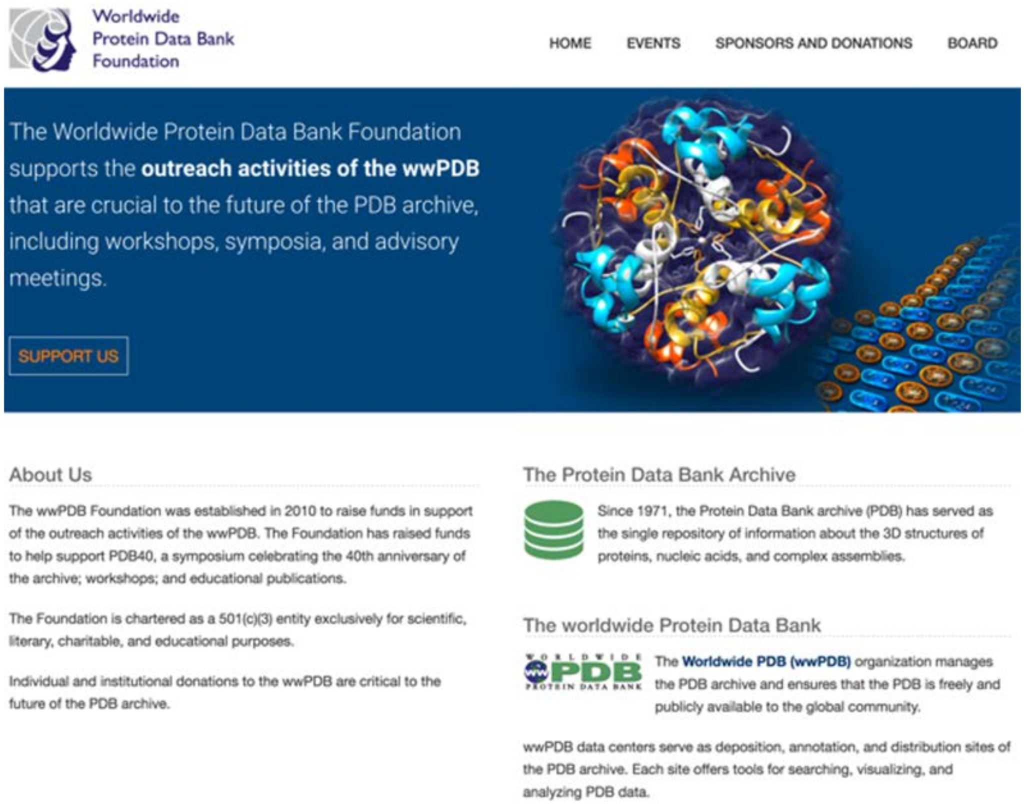
*2020 to 1st September

# EMDB Core Archive Growth



* Up to 1st September 2020

# wwPDB Foundation Progress



The Worldwide Protein Data Bank Foundation supports the **outreach activities of the wwPDB** that are crucial to the future of the PDB archive, including workshops, symposia, and advisory meetings.

About Us

The wwPDB Foundation was established in 2010 to raise funds in support of the outreach activities of the wwPDB. The Foundation has raised funds to help support PDB40, a symposium celebrating the 40th anniversary of the archive; workshops; and educational publications.

The Foundation is chartered as a 501(c)(3) entity exclusively for scientific, literary, charitable, and educational purposes.

Individual and institutional donations to the wwPDB are critical to the future of the PDB archive.

The Protein Data Bank Archive

Since 1971, the Protein Data Bank archive (PDB) has served as the single repository of information about the 3D structures of proteins, nucleic acids, and complex assemblies.

The worldwide Protein Data Bank

The **Worldwide PDB (wwPDB)** organization manages the PDB archive and ensures that the PDB is freely and publicly available to the global community.

wwPDB data centers serve as deposition, annotation, and distribution sites of the PDB archive. Each site offers tools for searching, visualizing, and analyzing PDB data.

http://foundation.wwpdb.org/

- Fundraising ongoing
- Planning for PDB50
  - May 5th @Online
  - July 24th @ACA
  - August 14th @IUCr
  - Oct. 20th-22nd @EMBL, Heidelberg
  - Dec. 6th @Kuala Lumpur Malaysia (Online?)

# wwPDB Collaboration Resource November 2019-October 2020

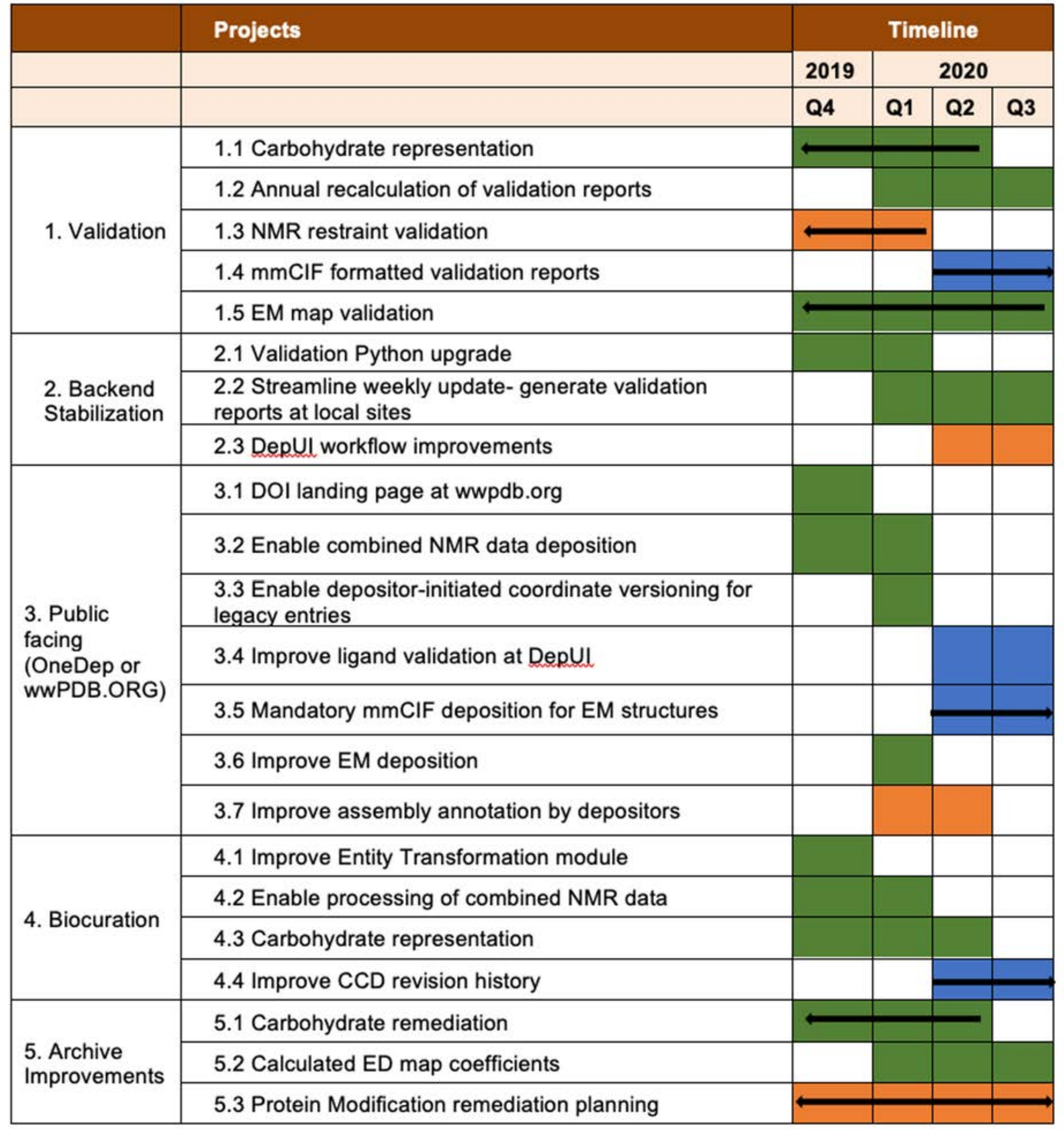
| wwPDB Partner | Software Development | Production Maintenance/ Management | Requirements Setting/ Testing | Core Archive Keeping* | Outreach | Biocuration/ Remediation | Total FTE Commitments |
|---|---|---|---|---|---|---|---|
| RCSB PDB | 2.0 | 1.6 | 0.35/0.35 | 2.0 | 0.3 | 6.0 | 12.6 |
| PDBe | 1.5 | 1.0 | 0.35/0.35 | - | 0.3 | 4.0 | 7.5 |
| PDBj | 0.4 | 0.4 | 0.2/0.2 | - | 0.1 | 4.5 | 5.8 |
| BMRB | 0.85 | - | 0.20 | 0.95 | - | 0.20 | 2.20 |
| EMDB | 0.9 | 0.35 | 0.1/0.2 | 0.3 | - | 0.5 | 2.35 |
| Total wwPDB | 5.65 | 3.35 | 2.3 | 3.25 | 0.7 | 15.2 | 30.45 |

*Resource from Archive Keeper: RCSB PDB; EMDB; BMRB

# OneDep 2019/2020 Progress *vs.* Goals

Delivered, To be delivered, Delayed

Ref. Appendix O



| | Projects | Timeline | | | |
|---|---|---|---|---|---|
| | | 2019 | 2020 | | |
| | | Q4 | Q1 | Q2 | Q3 |
| 1. Validation | 1.1 Carbohydrate representation | 🟩← | 🟩 | 🟩 | |
| | 1.2 Annual recalculation of validation reports | | 🟩 | 🟩 | 🟩 |
| | 1.3 NMR restraint validation | 🟧← | 🟧 | | |
| | 1.4 mmCIF formatted validation reports | | | 🟦 | 🟦→ |
| | 1.5 EM map validation | 🟩← | 🟩 | 🟩 | 🟩 |
| 2. Backend Stabilization | 2.1 Validation Python upgrade | 🟩 | 🟩 | | |
| | 2.2 Streamline weekly update- generate validation reports at local sites | | 🟩 | 🟩 | 🟩 |
| | 2.3 DepUI workflow improvements | | | 🟧 | 🟧 |
| 3. Public facing (OneDep or wwPDB.ORG) | 3.1 DOI landing page at wwpdb.org | 🟩 | | | |
| | 3.2 Enable combined NMR data deposition | 🟩 | 🟩 | | |
| | 3.3 Enable depositor-initiated coordinate versioning for legacy entries | | 🟩 | | |
| | 3.4 Improve ligand validation at DepUI | | | 🟦 | 🟦 |
| | 3.5 Mandatory mmCIF deposition for EM structures | | | 🟦 | 🟦→ |
| | 3.6 Improve EM deposition | | 🟩 | | |
| | 3.7 Improve assembly annotation by depositors | | 🟧 | 🟧 | |
| 4. Biocuration | 4.1 Improve Entity Transformation module | 🟩 | | | |
| | 4.2 Enable processing of combined NMR data | 🟩 | 🟩 | | |
| | 4.3 Carbohydrate representation | 🟩 | 🟩 | 🟩 | |
| | 4.4 Improve CCD revision history | | | 🟦 | 🟦→ |
| 5. Archive Improvements | 5.1 Carbohydrate remediation | 🟩← | 🟩 | 🟩 | |
| | 5.2 Calculated ED map coefficients | | 🟩 | 🟩 | 🟩 |
| | 5.3 Protein Modification remediation planning | 🟧← | 🟧 | 🟧 | 🟧→ |

# 2019/2020 Progress *vs.* Goals I

- Provided wwPDB DOI resolution
- Enabled author-initiated coordinate replacement (Legacy entries, phase II)
- Enabled single NMR data file deposition in NEF or NMR-STAR format
- Completed carbohydrate remediation
- Improved biocuration processes on entity transformation for BIRD molecules
- Streamlined weekly update- enabled per-site generation of validation reports
- Updated archive validation reports with enhancements for ligands and 3DEM maps and provided ED map coefficient files

# 2019/2020 Progress *vs.* Goals II

Re-forecasted

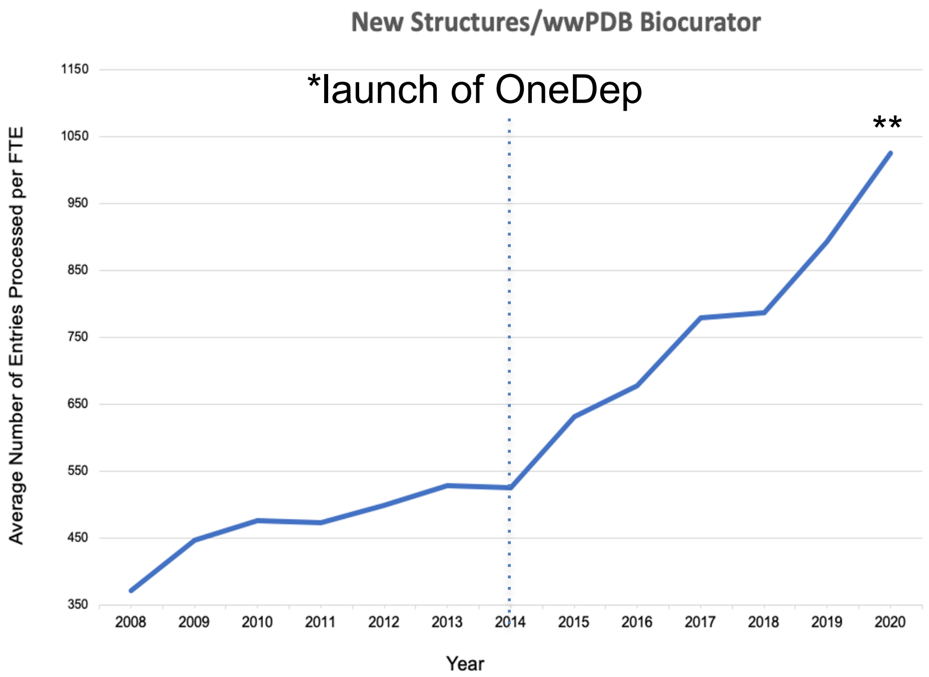
- Implement NMR restraint validation
- Depositor-annotated assembly
- Post-Translational Modification project planning

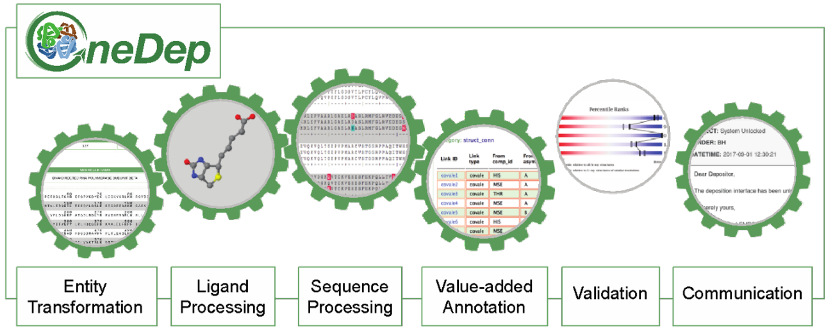
Mitigation

- Actively engaged NMR community in 2020
- Set clear requirements and phased plan for depositor-annotated assembly
- Follow carbohydrate remediation project template

# wwPDB Biocurator Productivity

- Continuing increased efficiency since 2009
- Significant increase with OneDep system
- Remote biocuration since Mar. 2020
- Pandemic doesn't impact biocuration productivity

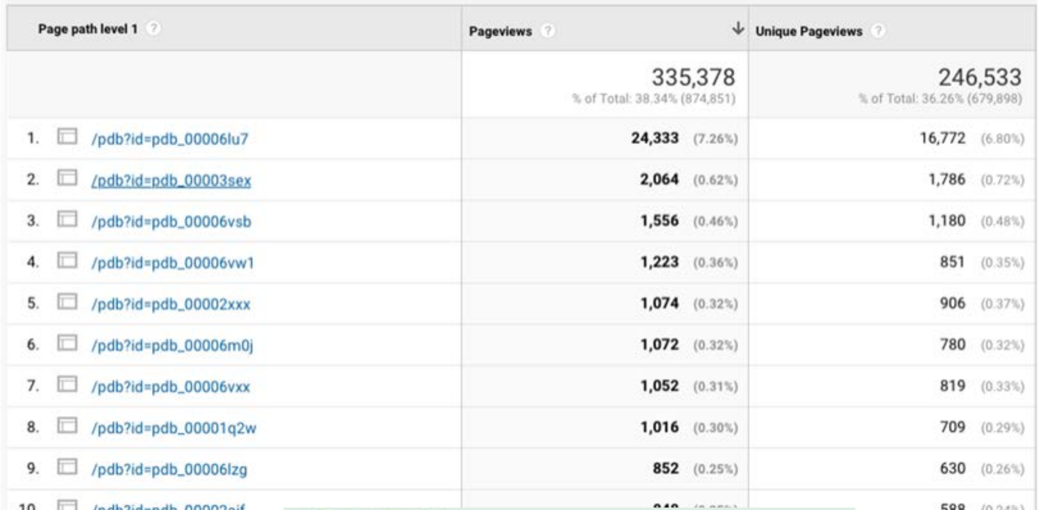




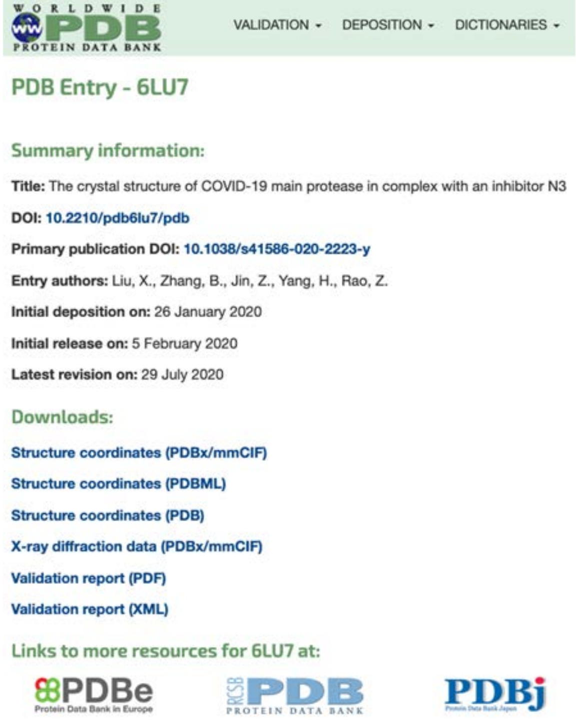
** As of September 1st 2020

# wwPDB DOI Resolution

- Provided for all onhold and released PDB entries
- Accessed > 335K times
- First coronavirus entry 6lu7 has top visit
- Some journals have adopted DOI links (Communication ongoing)
    - Acta Cryst. D & F
    - FEBS J.
    - JBC



| | Page path level 1 | Pageviews | Unique Pageviews |
|---|---|---|---|
| | | 335,378<br>% of Total: 38.34% (874,851) | 246,533<br>% of Total: 36.26% (679,898) |
| 1. | /pdb?id=pdb_00006lu7 | 24,333 (7.26%) | 16,772 (6.80%) |
| 2. | /pdb?id=pdb_00003sex | 2,064 (0.62%) | 1,786 (0.72%) |
| 3. | /pdb?id=pdb_00006vsb | 1,556 (0.46%) | 1,180 (0.48%) |
| 4. | /pdb?id=pdb_00006vw1 | 1,223 (0.36%) | 851 (0.35%) |
| 5. | /pdb?id=pdb_00002xxx | 1,074 (0.32%) | 906 (0.37%) |
| 6. | /pdb?id=pdb_00006m0j | 1,072 (0.32%) | 780 (0.32%) |
| 7. | /pdb?id=pdb_00006vxx | 1,052 (0.31%) | 819 (0.33%) |
| 8. | /pdb?id=pdb_00001q2w | 1,016 (0.30%) | 709 (0.29%) |
| 9. | /pdb?id=pdb_00006lzg | 852 (0.25%) | 630 (0.26%) |
| 10. | /pdb?id=pdb_00002ajf | [illegible] | 588 (0.24%) |



PDB Entry - 6LU7

Summary information:

**Title:** The crystal structure of COVID-19 main protease in complex with an inhibitor N3

**DOI:** 10.2210/pdb6lu7/pdb

**Primary publication DOI:** 10.1038/s41586-020-2223-y

**Entry authors:** Liu, X., Zhang, B., Jin, Z., Yang, H., Rao, Z.

**Initial deposition on:** 26 January 2020

**Initial release on:** 5 February 2020

**Latest revision on:** 29 July 2020

Downloads:

Structure coordinates (PDBx/mmCIF)

Structure coordinates (PDBML)

Structure coordinates (PDB)

X-ray diffraction data (PDBx/mmCIF)

Validation report (PDF)

Validation report (XML)

Links to more resources for 6LU7 at:

PDBe Protein Data Bank in Europe
RCSB PDB PROTEIN DATA BANK
PDBj Protein Data Bank Japan

# wwPDB Core Member Funding Status

- RCSB PDB: NSF/NIH/DOE funding renewed: 2019-2023

- BMRB: NIH NIGMS funding: 2019-2023
    - Inadequate budget: need to find additional support
        - NIH R01 transferred to UConn
        - NIH U24 submitted

- PDBe: EMBL-EBI, Wellcome Trust: 2021-2025

- PDBj: NBDC-JST and AMED funding: 2019-2022
    - Possible additional budget from S. Korea

- EMDB: EMBL-EBI, Wellcome Trust: 2019-2024

# wwPDB Collaboration Resources

## November 2020-October 2021

| wwPDB Partner | Software Development | Production Maintenance/ Management | Requirements Setting/ Testing | *Core Archive Keeping | Outreach | Biocuration/ Remediation | Total FTE Commitments |
|---|---|---|---|---|---|---|---|
| RCSB PDB | 2.0** | 1.3 | 0.35/0.35 | 2.0 | 0.3 | 6.3 | 12.6 |
| PDBe | 1.4** | 1.0 | 0.35/0.35 | - | 0.2 | 5.0 | 8.3 |
| PDBj | 0.4 | 0.4 | 0.2/0.2 | - | 0.1 | 4.5 | 5.8 |
| BMRB | 0.95 | - | 0.1/0/1 | 0.5 | - | 0.2 | 1.85 |
| EMDB | 0.9 | 0.35 | 0.1/0.2 | 0.3 | - | 0.5 | 2.35 |
| Total wwPDB | 5.65 | 3.05 | 2.3 | 2.8 | 0.6 | 16.5 | 30.9 |

* Resource from Archive Keeper: RCSB PDB; EMDB; BMRB

**excluding additional resource from BBSRC/NSF joint grant, 1.0 FTE at PDBe and 1.3 FTE at RCSB PDB

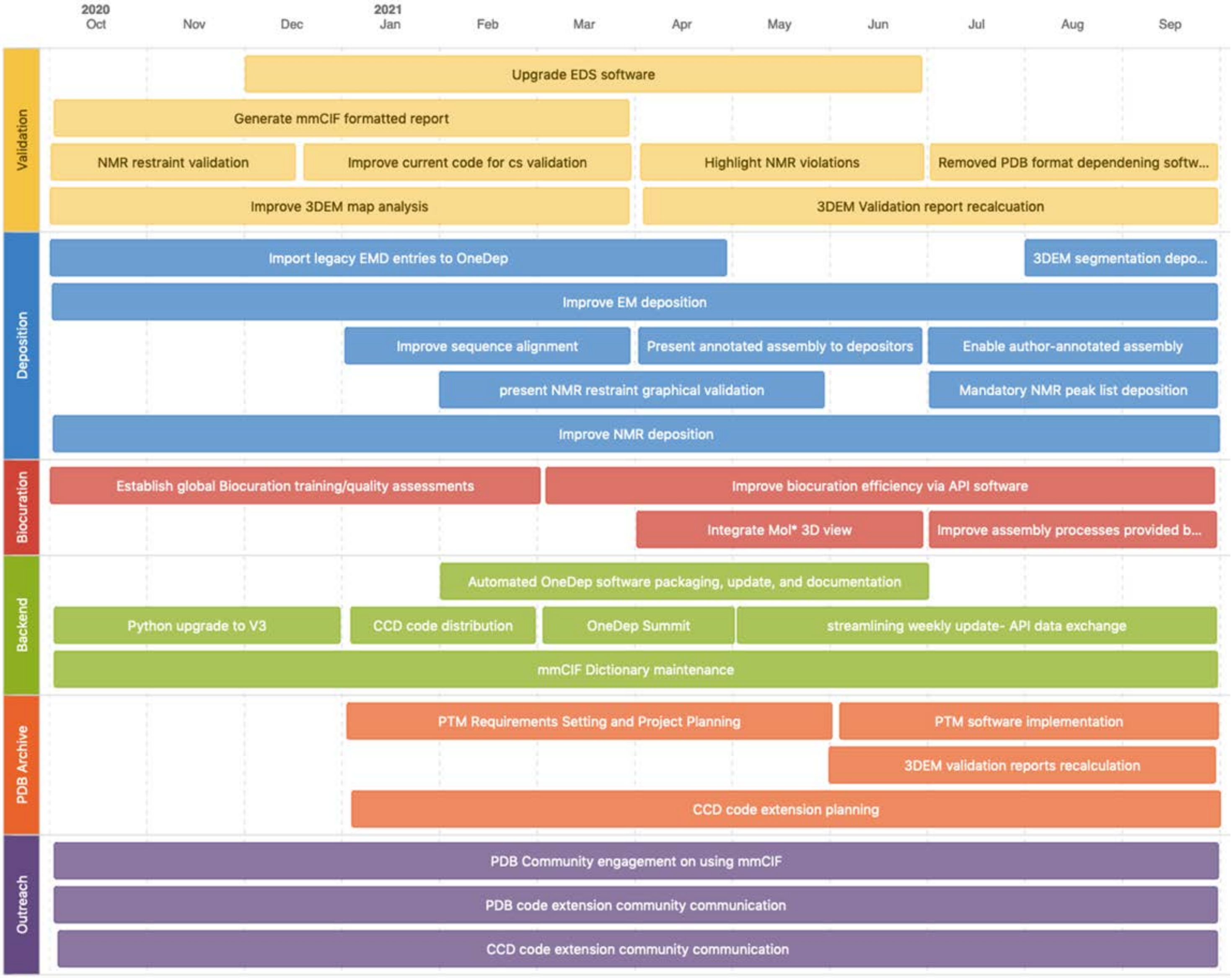
# OneDep 2020/2021 Goal Setting I

| | Major Projects | Primary resource |
|---|---|---|
| Validation | **Implement NMR restraint validation** | **BMRB** |
| | Improve EM map validation | EMDB |
| | Provide mmCIF formatted validation report | RCSB PDB |
| | Upgrade 3rd party EDS software | PDBe |
| | Refactor NMR chemical shifts validation | BMRB |
| Public facing | Improve sequence alignment at DepUI | PDBe |
| | Improve NMR and EM depositions | PDBj/EMDB |
| | **Enable author-annotated assembly** | **PDBe** |
| Annotation | Establish global Biocuration training/quality assessments | RCSB PDB/PDBe/PDBj |
| | Improve assembly processes provided by authors | RCSB PDB/PDBe |
| | Improve biocuration efficiency via API software | RCSB PDB/PDBe |
| Backend | Automated OneDep software packaging and update | RCSB PDB/PDBe |
| | Ligand ID extension planning | RCSB PDB |
| | mmCIF Dictionary maintenance | RCSB PDB |
| PDB Archive | **PTM remediation** | **PDBe** |
| | 3DEM validation reports recalculation | EMDB/PDBe |

**Bold**: re-forecasted from 2019/2020

# OneDep 2020/2021 Goal Setting II

## Timeline



* Timeline will be further refined after requirement setting.

# EM Data-Management Workshop

**EMBL-EBI: January 23-24, 2020**

- Details in EMDB Presentation

# EMDB to Host 2021 wwPDB AC

- Next wwPDB AC Meeting
  Date: Tuesday, Oct. 19th 2021
  Host: EMDB
  Venue: EMBL-Heidelberg, Boxberg, Germany

- PDB50 Celebration (Europe) to follow immediately thereafter (Oct. 20th-22nd 2021) at EMBL-Heidelberg

- 2022 wwPDB AC Meeting Scheduling
  Date Options: Friday Oct. 14th or Friday Oct. 21nd
  Host: RCSB PDB
  Venue: Rutgers University, Piscataway, NJ, USA

# Remaining Agenda Items

- Discussion

- MX Update  (SKB)
- NMR Update (JCH)
- EMDB Update (SV for AP)
- Outreach and Training Update (GK)
- Questions for the Advisory Committee (SKB)
- Executive Session

# Update on Macromolecular Crystallography

Stephen K. Burley

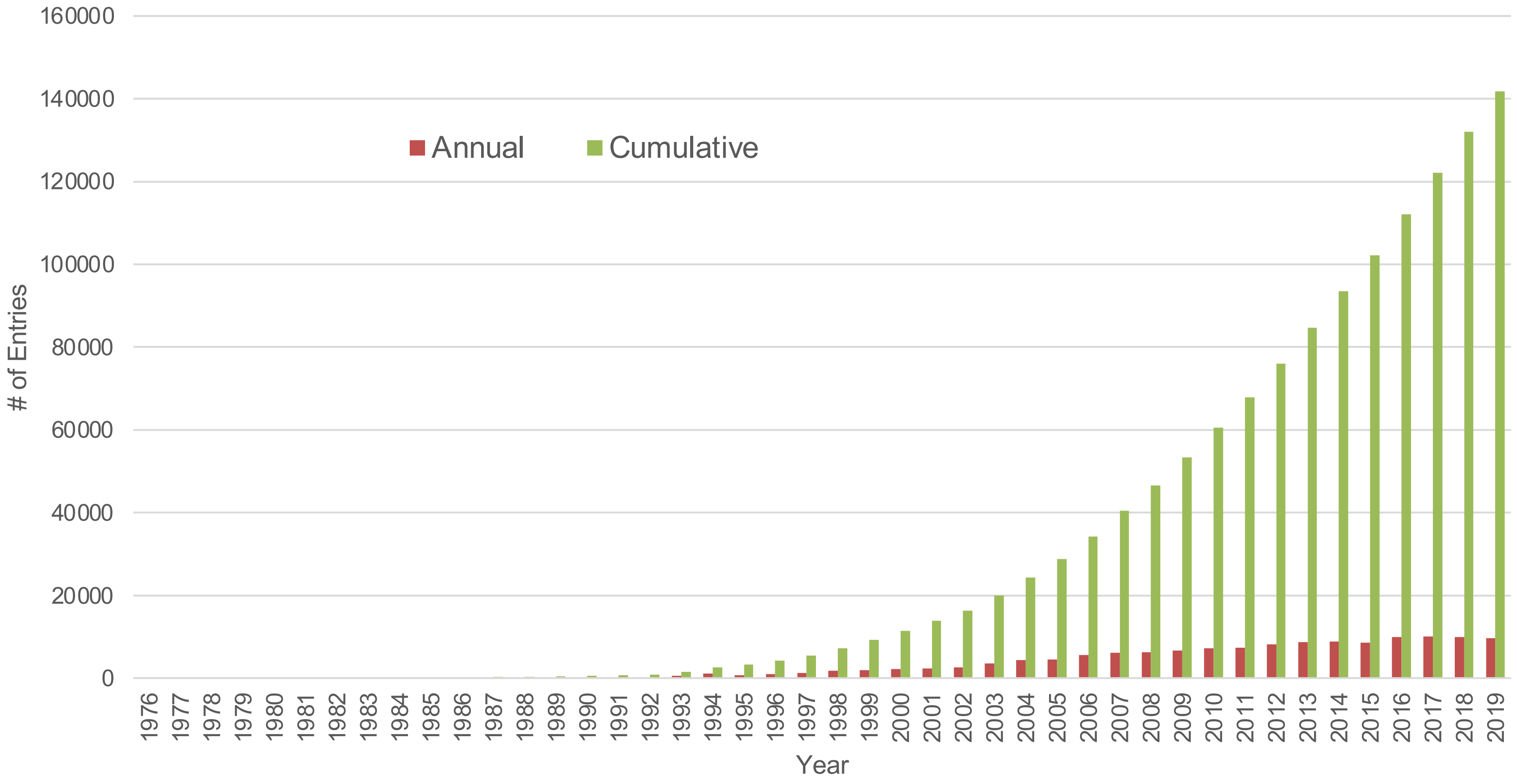
WORLDWIDE PDB PROTEIN DATA BANK

wwpdb.org

# Agenda

- MX Data Deposition Metrics
- Update on Structure Versioning
- Update on PDBx/mmCIF Mandatory Deposition
- Update on PDBx/mmCIF Working Group Activities

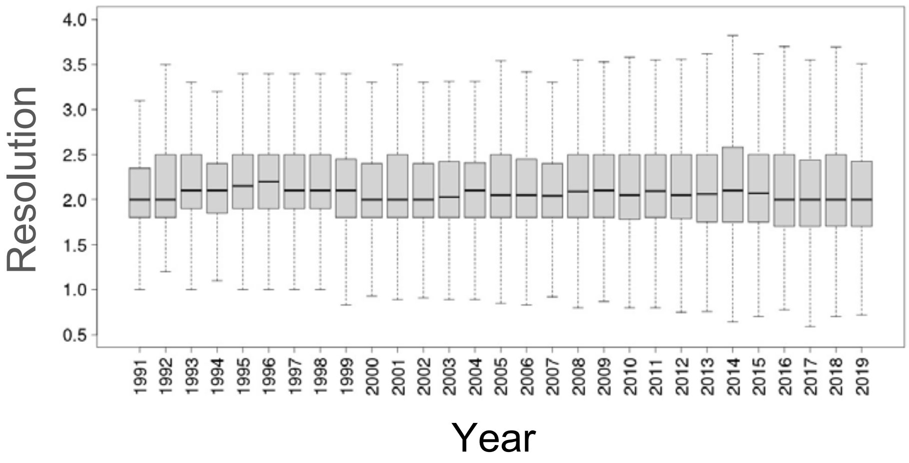
# Growth of Released MX Entries



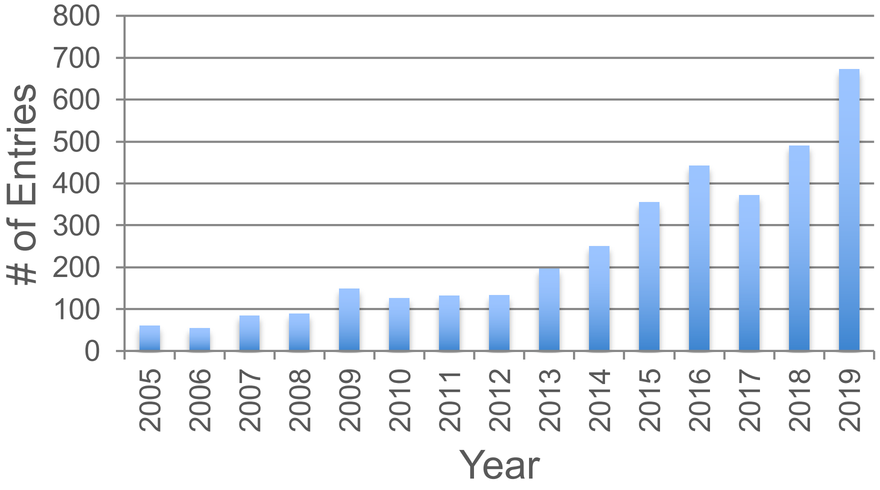
>151,000 Total Released MX Entries Projected for End 2020

# MX Deposition Size and Complexity

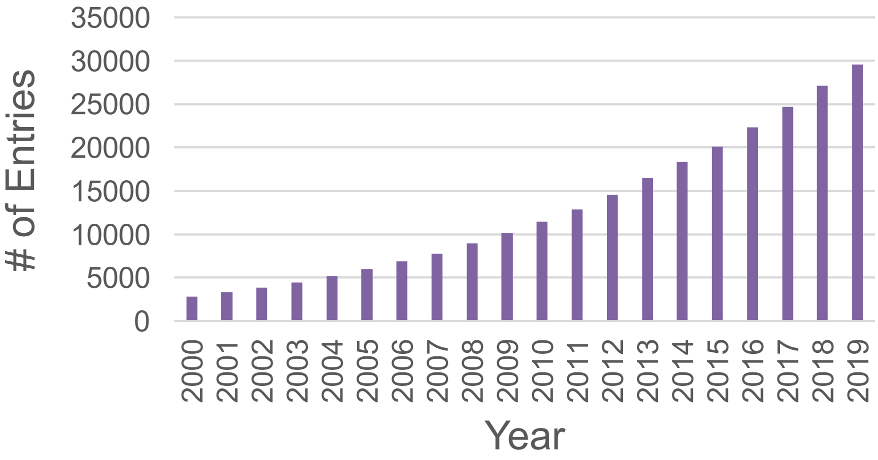
### Annual Distribution for High Resolution Limit



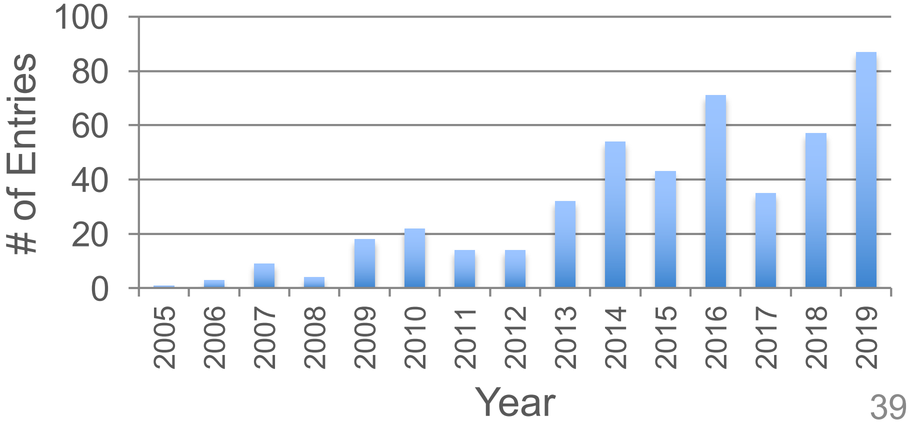
### Annual Released Structures With AU MW > 500,000



### Total Number of New CCD Entries



### Annual Released Large Structures (chains > 62)



# Update on Structure Versioning

- Atomic coordinate replacement Phase 1 (OneDep) began July 31 2019
- Phase 2 (Legacy) Feb 18 2020
- Initial uptake has been modest
  - OneDep: 93
  - Legacy: 7

- Motivations for replacement include:
  - Incomplete structural model
  - Ligand geometry
  - Sequence discrepancy
  - Ligand identity
  - Polymer geometry

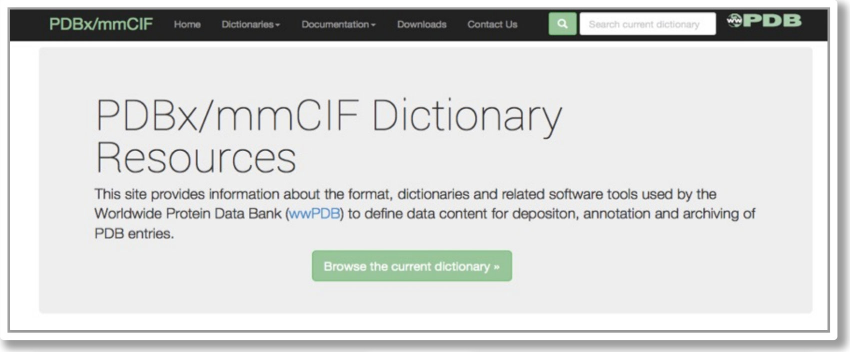
# Update on PDBx/mmCIF Mandatory Deposition

- PDBx/mmCIF atomic coordinate deposition made mandatory July 1 2019
- Announced: Apr 2019
  - doi:10.1107/S2059798319004522
- Compliance: 100%
- Depositor Feedback
  - Depositors do not upgrade software as frequently as they

- Lessons Learned:
  - Need broader set of examples and test cases for developers
  - Need more accessible documentation for depositors to access native mmCIF package features
  - Need testing development versions of software

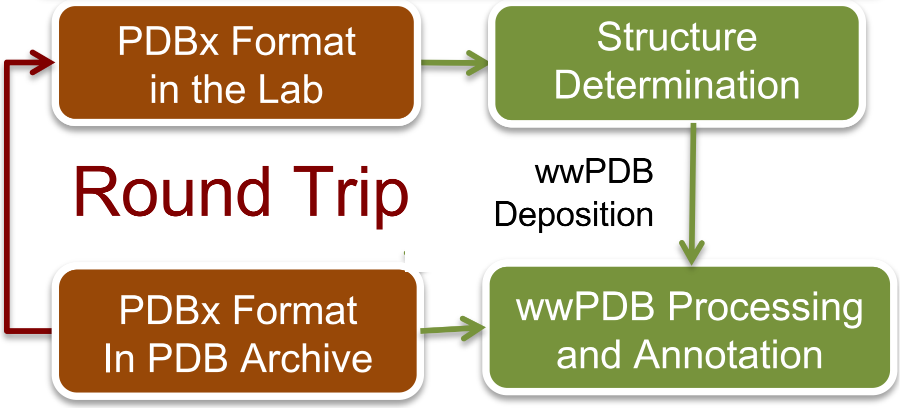
# Update on PDBx/mmCIF Working Group

- PDBx/mmCIF is the deposition and archiving data standard for the repository
- wwPDB together with the PDBx/mmCIF Working Group of community experts and methods developers oversee the evolution of the standard
- Working Group ensures that the standard is well supported by key community software tools.
- 2019-2020 PDBx/mmCIF Working Group focus areas:
    - Mandatory mmCIF deposition
    - Incorporate ligand and modified monomer chemical definitions with deposition input
    - Finalizing improvements in processed diffraction data organization



PDBx/mmCIF Dictionary Resources

This site provides information about the format, dictionaries and related software tools used by the Worldwide Protein Data Bank (wwPDB) to define data content for depositon, annotation and archiving of PDB entries.



PDBx Format in the Lab

Structure Determination

Round Trip

wwPDB Deposition

PDBx Format In PDB Archive

wwPDB Processing and Annotation

PDBx/mmCIF Workshop Participants, July 2017

# BMRB Core Archive: Transition and Plans

**Jeff Hoch**

# BMRB Leadership Transition

- As of April 1, 2020, John Markley will withdraw as Co-Head of BMRB and be replaced by Chad Rienstra; Jeff Hoch will become sole BMRB representative to wwPDB, sole PI of the BMRB NIH grant
  - Chad, an expert in biological solid-state NMR, was recently recruited to UW-Madison from University of Illinois

- John Markley will continue to be associated with BMRB as an Emeritus Professor on a voluntary basis and will provide advice and assistance as needed

Canceled

# Plan “B” Successfully Launched

- NIH hurdles resulted in funding hiatus from 4/1 to 7/21
- NIH grant awarded in entirety to UConn 7/21 with Hoch as sole PI
- UConn Vice President for Research (Radenka Maric) commits resources to aid transition :
  - $80K in hardware capital costs to recapitulate UW infrastructure for BMRB operations
  - 20% effort of IT/Bioinformatics Analyst
  - 20% effort of IT Project Manager to assist with WBS for U24 grant proposal
  - Challenge commitment: 50% of PM if BMRB can raise 50% through grant(s)

# BMRB Plan "B" Status

- 12 40-core Dell servers installed
- Kumaran Baskaran moves to CT
- Jon Weddell, Dmitri Maziuk, Kumaran Baskaran, Hongyang Yao transition to interim contracts via temp agency service provider
- Michael Wilson, Mark Maciejewski liaise with Dmitri Maziuk, Jon Weddell to transition BMRB services to UConn
- Services migrated as of 8/31/20:
  - ETS completely moved
  - Database, web site, API move imminent
  - Deposition system ready to switch upon annotation workflow move

# BMRB Core Archive Plans I

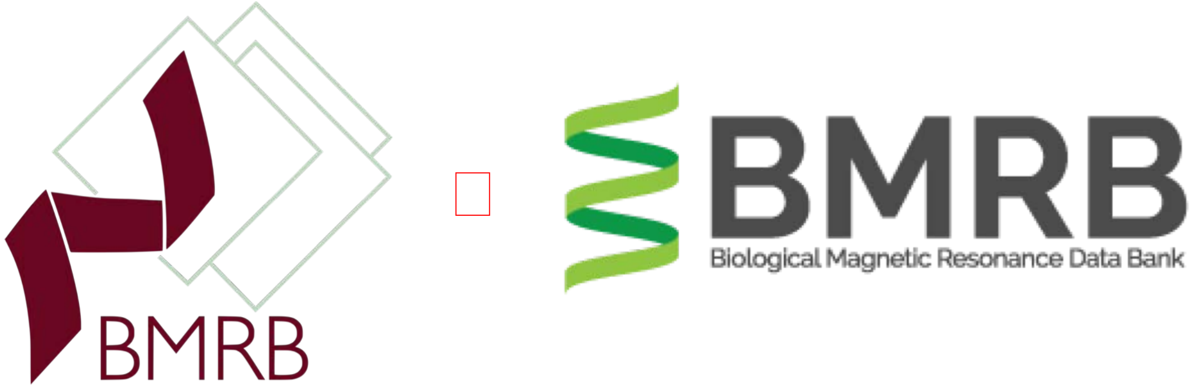
- Policy statements on OneDep/BMRBdep, NMR-STAR/NEF:
  - As an essential partner in the OneDep Team, BMRB commits to ensuring that BMRBdep is fully integrated in OneDep
  - While NMR-STAR remains the archive format for biomolecular NMR data hosted by BMRB, BMRB is fully committed to supporting NEF as an exchange and deposition format

- Seek additional funding for (1) quotidian operations and (2) R & D on additional value-added services for biomolecular NMR

# BMRB Core Archive Plans II

- Explore expansion of small molecule data sharing with PDB (aligning with CCD)

- Complete overhaul of web site
  - Logo



  - bmrb.wisc.edu □ bmrb.io

- Work on documenting, strengthening, and streamlining internal systems and SOPs

# BMRB Core Archive Plans III

- Finalize work on curated/normalized NMR-STAR schema

- Continue expansion of curated NMR data types

- Continue expansion of curated collections pertinent to specialized areas
  - SSNMR
  - Disordered systems
  - Metabolomics

- Continue rollout and testing of BMRbig
  - Develop tools to facilitate populating BMRB and PDB depositions from BMRbig uploads

# 3DEM Plans

Ardan Patwardhan/Sameer Velankar

WORLDWIDE PDB PROTEIN DATA BANK

wwpdb.org

# EMDB Core Archive

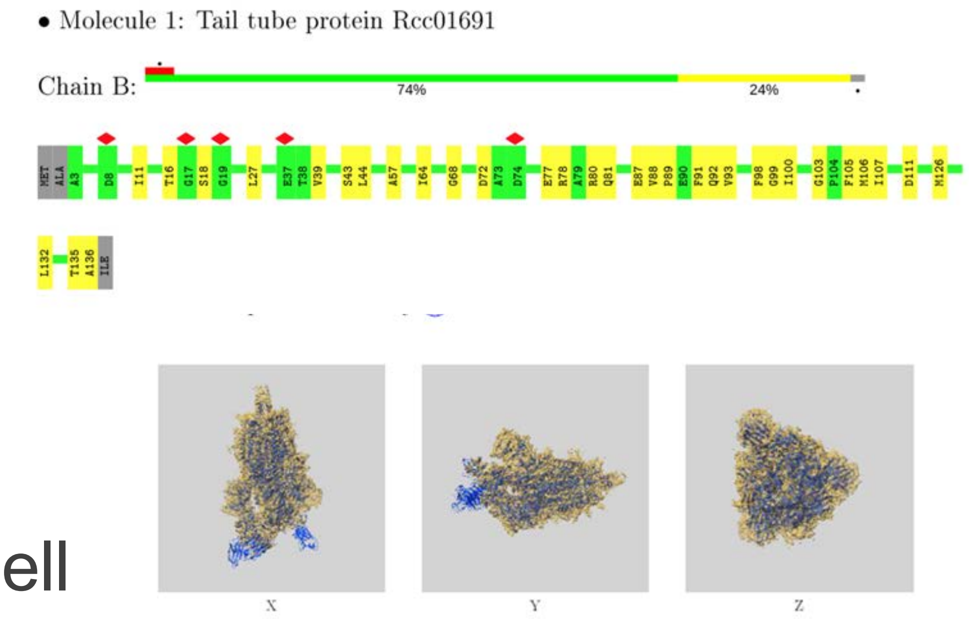
- EMDB policy created and publicly released
  - https://www.ebi.ac.uk/pdbe/emdb/policies.html
- EMDB release policies are consistent with PDB policies
- EMDB XML header no longer released upon entry approval

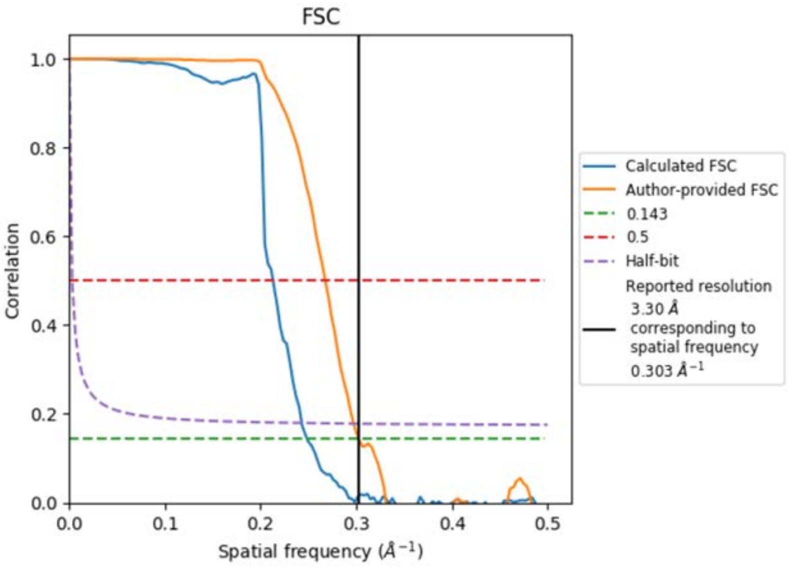
# EM validation reports – an overview

- Validation reports for EM entries: map-only, map-model and map with multiple models
- EM validation report types
  - Map-only validation report (created for all the above EM entries)
  - Map-model validation report (created for EM entries with one or more models)
- Currently only depositors are provided with the preliminary reports

# wwPDB validation report improvements - EM

- Map + model quality
- Map-model fit
  - Per-residue quality plots
  - Image of map + model overlayed
- Improvements to Fourier-Shell Correlation (FSC) plot:
  - Limited cut-offs to '0.143', '0.5' and 'halfbit' unless another criteria used
  - FSC curves provided by the authors and recalculated from the deposited half-maps shown in one plot





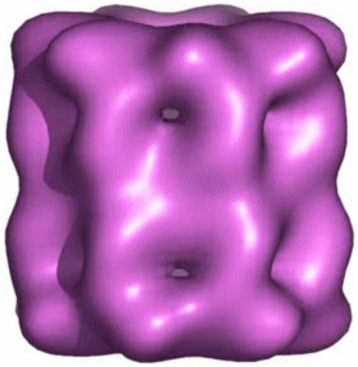
# wwPDB validation report improvements - EM

- Addition of the following images of the raw map calculated from the deposited half-maps:
  - Orthogonal projections
  - Central slices
  - Largest variance slices
  - Orthogonal surface views

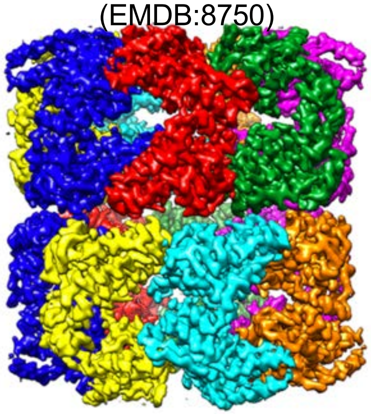
- The images are presented below those of the primary map

- Rotationally-Averaged Power Spectrum (RAPS) plot shown for both primary and raw map

# EM data management workshop (Jan 2020)



GroEL at 25Å in 2006 (EMDB:1291) and at 3.5Å in 2017 (EMDB:8750)



# Outcomes of the EM workshop

- Recommendations about improved data capture by wwPDB/EMDB, e.g.
  - Mandatory model deposition in PDBx/mmCIF format & support for software developers
    - Recommend to make PDBx/mmCIf mandatory for model deposition from 1st July 2021
  - Capture particle-picking metadata
  - Deposit half-maps if used
  - Add an “investigation/project” level to group related entries
- Comments and recommendations on validation reports
- Model validation recommendations, e.g.
  - Track model restraints used
  - Additional coordinate-based metric not biased by torsion-angle restraints

# Outcomes of the EM workshop

- Data and map validation recommendations, e.g.
  - Evaluate various local resolution metrics (qualitative rather than quantitative) (ResMap, BlocRes, MonoRes, …)
  - Evaluate measures of map anisotropy and angular coverage (3DFSC, CryoEF, MonoDIR, EMDA, SCF, …)
  - Add symmetry analysis (ProShade)
  - Deposition of particle stacks would be extremely useful for developing new validation metrics, and also allow re-processing and potential map and model improvement

# Outcomes of the EM workshop

- Validation of map-model fit recommendations, e.g.
  - Currently lacking good metric (atom inclusion subjective due to contour-level choice)
  - Add map/model FSC plots, also per-chain
  - Evaluate measures of real-space fit (RSCC, SMOC, EMringer, Q-score, …)
  - Evaluate measures for difference map calculation
  - Consider visual illustration of map-model fit in both a relatively good and a relatively poor part of the map
- EMDB to implement many methods where recommendations cannot yet be made to enable archive-wide analysis and expert assessment of performance in individual cases

# Outcomes of the EM workshop

- Challenges for methods and software developers
  - Identify criteria that could go in the “slider” plots for both map-model fit and data/map validation! (Must be hard to “fudge”)
  - Can 2D raw data (particle stacks) be used to validate the 3D map? How?
  - Develop model-validation criteria for reduced/coarse-grain models (e.g., Cα-only)
  - Develop a method to define an unbiased contour level (global and local)
  - Develop methods to assess if structural features observed at a given resolution are commensurate with expectations/experience (using machine learning?)

# What's next?

- 2020:
  - Publishing recommendations, incl. as preprint (white paper in progress)
  - Implement easy-to-do recommendations (on-going); then:
    - Update OneDep and validation server
    - Calculate and release validation reports for all current EM entries in EMDB and PDB
- Later:
  - Add many new methods to EMDB Validation Analysis (VA) web pages to enable evaluation of robustness, reliability, information, usefulness, etc. (on-going)
  - Implement additional recommendations in validation pipeline/reports
  - Wait for the field to do additional work and review in a few years' time
  - Get recommendations for EM modalities other than single particle analysis

# wwPDB Outreach & Expansion of the Franchise

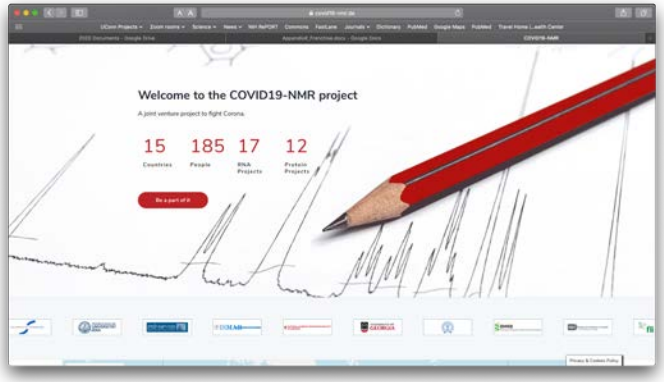
**Genji Kurisu**

WORLDWIDE PDB PROTEIN DATA BANK

wwpdb.org

# wwPDB Outreach

2020 OneDep Developer Summit
@ Zoom



Welcome to the COVID19-NMR project

A joint venture project to fight Corona

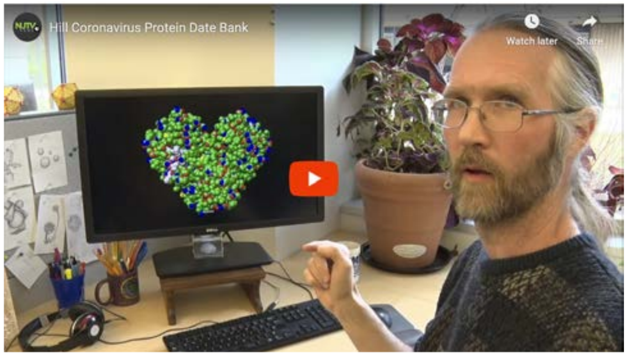
15 Countries

185 People

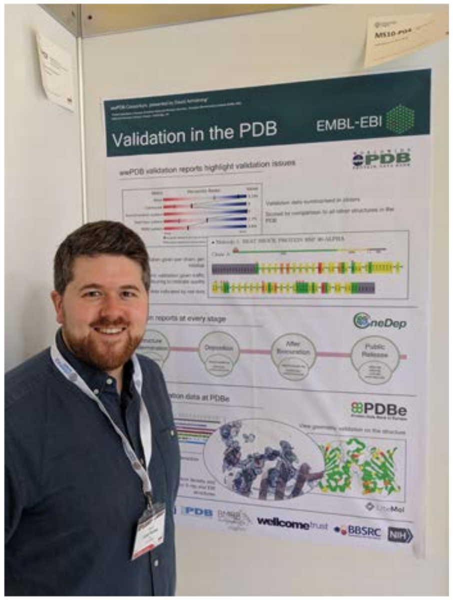
17 RNA Projects

12 Protein Projects

COVID-19 related
activity by BMRB



COVID-19 related activity by
RCSB PDB

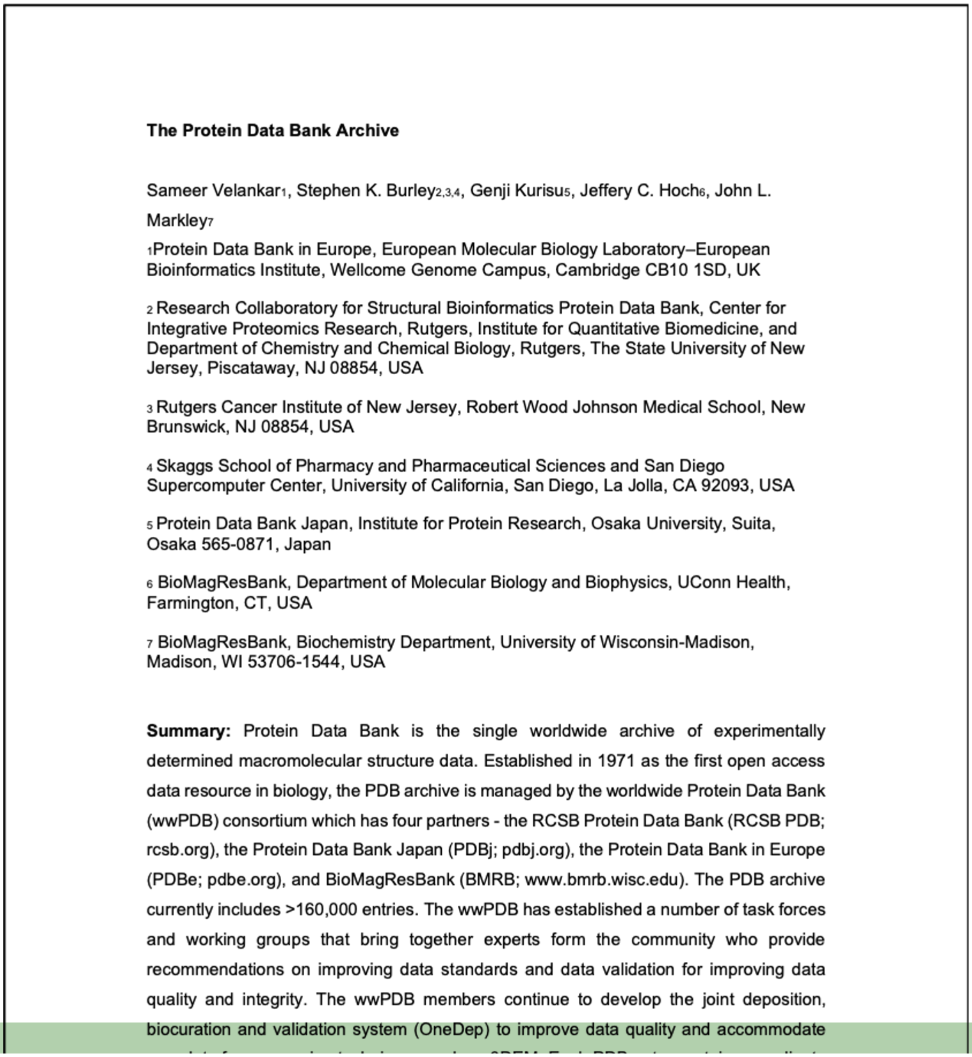


ECM 2019

AsCA@Singapore

# wwPDB 2019-2020 Publication

Title of Book: **Structural Proteomics: Methods and Protocols, Second Edition**
Editor name: **Raymond J. Owens. PhD.**
Title of Chapter: **The Protein Data Bank Archive**



**The Protein Data Bank Archive**

Sameer Velankar[1], Stephen K. Burley[2,3,4], Genji Kurisu[5], Jeffery C. Hoch[6], John L. Markley[7]

[1]Protein Data Bank in Europe, European Molecular Biology Laboratory–European Bioinformatics Institute, Wellcome Genome Campus, Cambridge CB10 1SD, UK

[2] Research Collaboratory for Structural Bioinformatics Protein Data Bank, Center for Integrative Proteomics Research, Rutgers, Institute for Quantitative Biomedicine, and Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

[3] Rutgers Cancer Institute of New Jersey, Robert Wood Johnson Medical School, New Brunswick, NJ 08854, USA

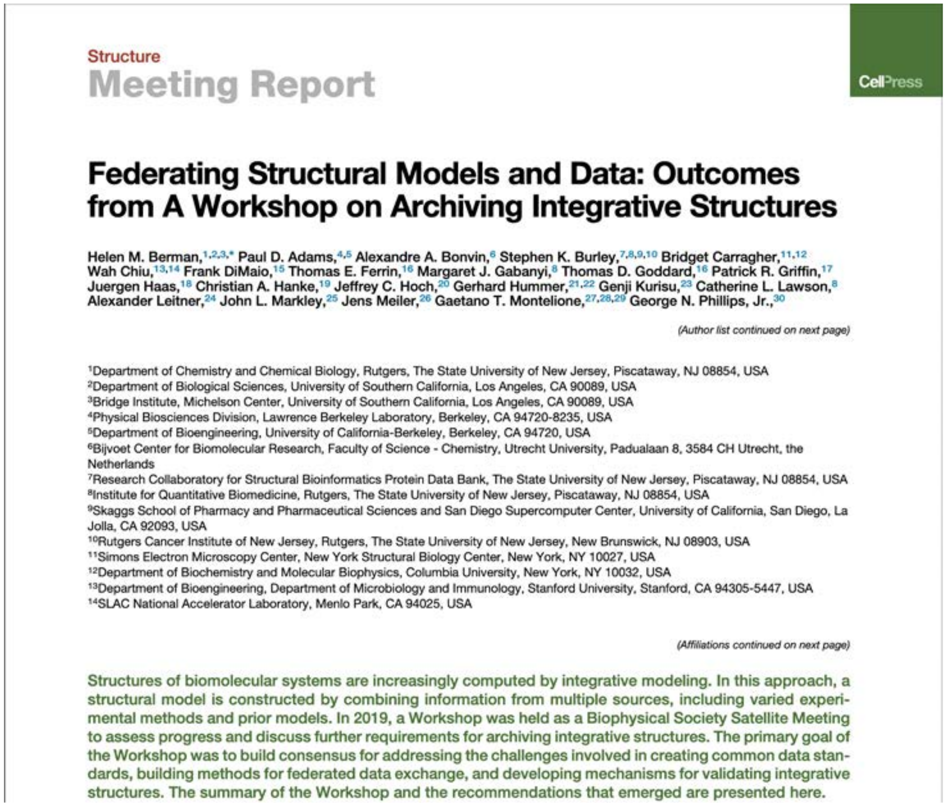
[4] Skaggs School of Pharmacy and Pharmaceutical Sciences and San Diego Supercomputer Center, University of California, San Diego, La Jolla, CA 92093, USA

[5] Protein Data Bank Japan, Institute for Protein Research, Osaka University, Suita, Osaka 565-0871, Japan

[6] BioMagResBank, Department of Molecular Biology and Biophysics, UConn Health, Farmington, CT, USA

[7] BioMagResBank, Biochemistry Department, University of Wisconsin-Madison, Madison, WI 53706-1544, USA

**Summary:** Protein Data Bank is the single worldwide archive of experimentally determined macromolecular structure data. Established in 1971 as the first open access data resource in biology, the PDB archive is managed by the worldwide Protein Data Bank (wwPDB) consortium which has four partners - the RCSB Protein Data Bank (RCSB PDB; rcsb.org), the Protein Data Bank Japan (PDBj; pdbj.org), the Protein Data Bank in Europe (PDBe; pdbe.org), and BioMagResBank (BMRB; www.bmrb.wisc.edu). The PDB archive currently includes >160,000 entries. The wwPDB has established a number of task forces and working groups that bring together experts form the community who provide recommendations on improving data standards and data validation for improving data quality and integrity. The wwPDB members continue to develop the joint deposition, biocuration and validation system (OneDep) to improve data quality and accommodate



Structure
Meeting Report

CellPress

## Federating Structural Models and Data: Outcomes from A Workshop on Archiving Integrative Structures

Helen M. Berman,[1,2,3,*] Paul D. Adams,[4,5] Alexandre A. Bonvin,[6] Stephen K. Burley,[7,8,9,10] Bridget Carragher,[11,12] Wah Chiu,[13,14] Frank DiMaio,[15] Thomas E. Ferrin,[16] Margaret J. Gabanyi,[8] Thomas D. Goddard,[16] Patrick R. Griffin,[17] Juergen Haas,[18] Christian A. Hanke,[19] Jeffrey C. Hoch,[20] Gerhard Hummer,[21,22] Genji Kurisu,[23] Catherine L. Lawson,[8] Alexander Leitner,[24] John L. Markley,[25] Jens Meiler,[26] Gaetano T. Montelione,[27,28,29] George N. Phillips, Jr.,[30]

*(Author list continued on next page)*

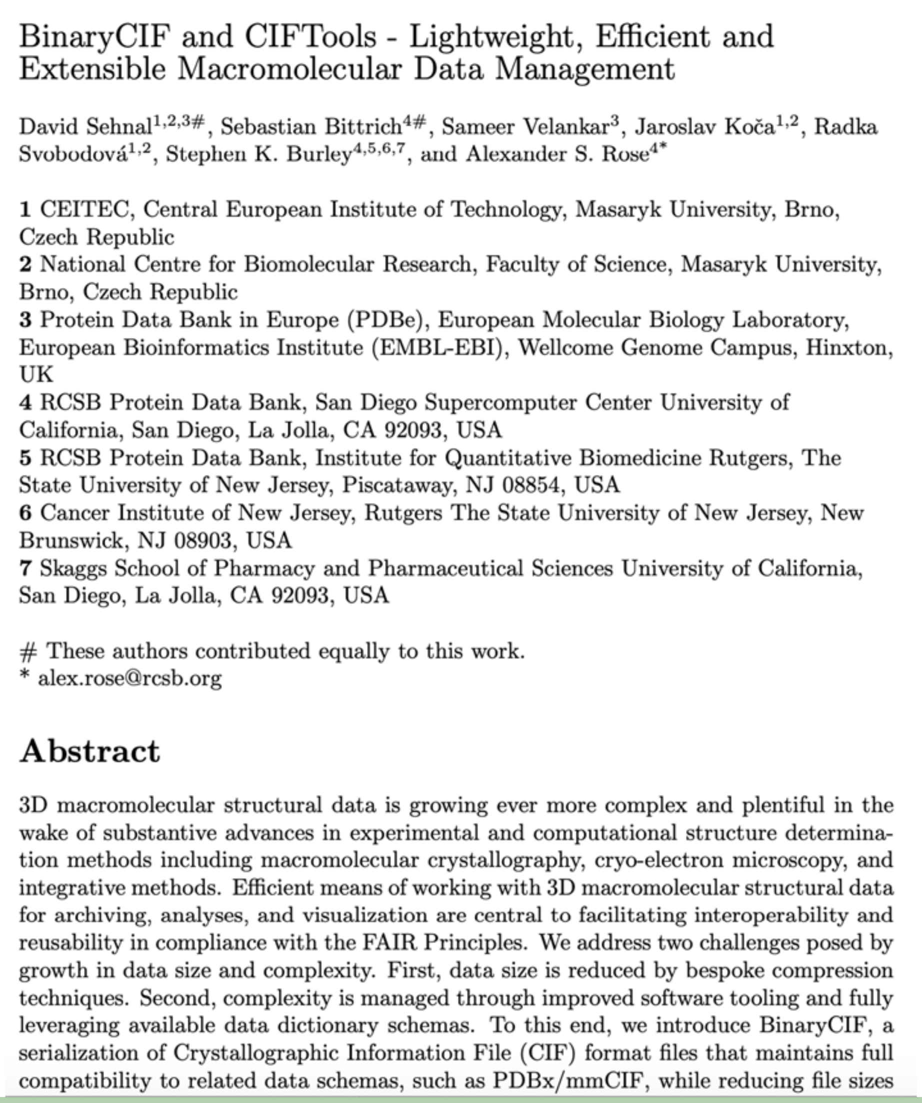
[1]Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA
[2]Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089, USA
[3]Bridge Institute, Michelson Center, University of Southern California, Los Angeles, CA 90089, USA
[4]Physical Biosciences Division, Lawrence Berkeley Laboratory, Berkeley, CA 94720-8235, USA
[5]Department of Bioengineering, University of California-Berkeley, Berkeley, CA 94720, USA
[6]Bijvoet Center for Biomolecular Research, Faculty of Science - Chemistry, Utrecht University, Padualaan 8, 3584 CH Utrecht, the Netherlands
[7]Research Collaboratory for Structural Bioinformatics Protein Data Bank, The State University of New Jersey, Piscataway, NJ 08854, USA
[8]Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA
[9]Skaggs School of Pharmacy and Pharmaceutical Sciences and San Diego Supercomputer Center, University of California, San Diego, La Jolla, CA 92093, USA
[10]Rutgers Cancer Institute of New Jersey, Rutgers, The State University of New Jersey, New Brunswick, NJ 08903, USA
[11]Simons Electron Microscopy Center, New York Structural Biology Center, New York, NY 10027, USA
[12]Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032, USA
[13]Department of Bioengineering, Department of Microbiology and Immunology, Stanford University, Stanford, CA 94305-5447, USA
[14]SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA

*(Affiliations continued on next page)*

Structures of biomolecular systems are increasingly computed by integrative modeling. In this approach, a structural model is constructed by combining information from multiple sources, including varied experimental methods and prior models. In 2019, a Workshop was held as a Biophysical Society Satellite Meeting to assess progress and discuss further requirements for archiving integrative structures. The primary goal of the Workshop was to build consensus for addressing the challenges involved in creating common data standards, building methods for federated data exchange, and developing mechanisms for validating integrative structures. The summary of the Workshop and the recommendations that emerged are presented here.

# wwPDB 2019-2020 Publication

Manuscript accepted in PLOS Computational Biology



## BinaryCIF and CIFTools - Lightweight, Efficient and Extensible Macromolecular Data Management

David Sehnal[1,2,3#], Sebastian Bittrich[4#], Sameer Velankar[3], Jaroslav Koča[1,2], Radka Svobodová[1,2], Stephen K. Burley[4,5,6,7], and Alexander S. Rose[4*]

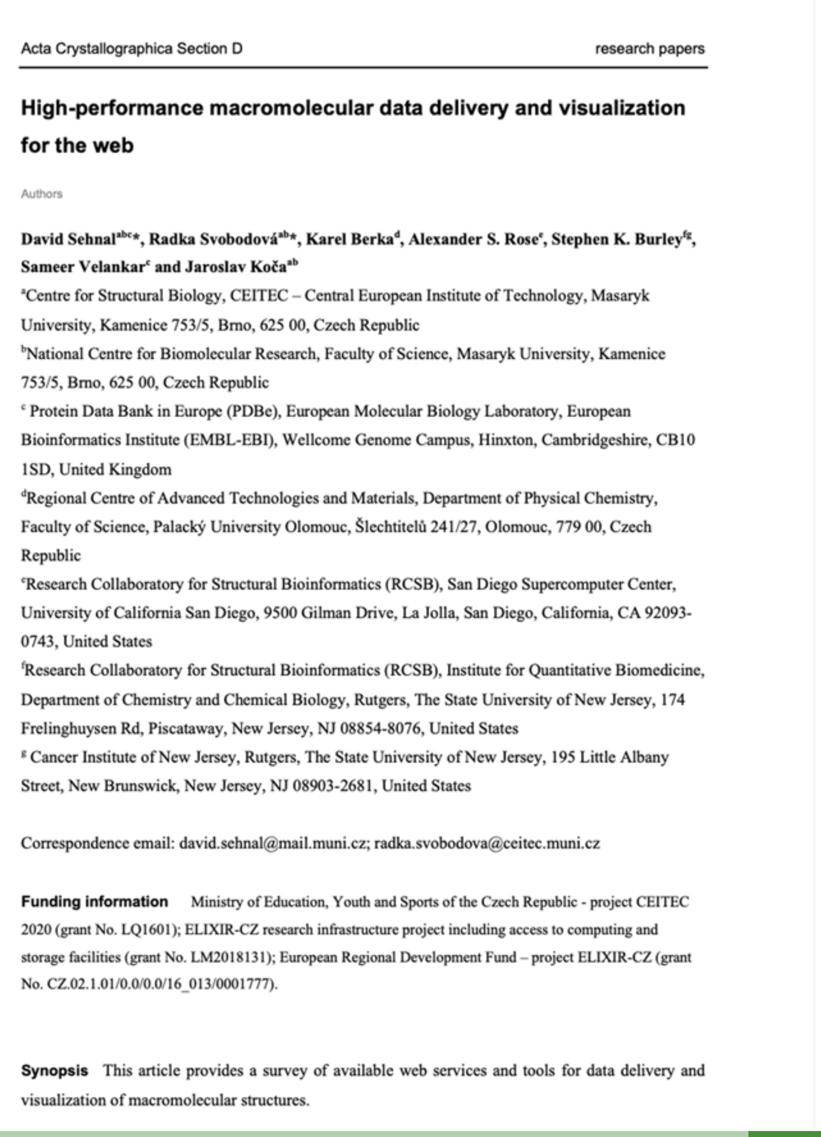
**1** CEITEC, Central European Institute of Technology, Masaryk University, Brno, Czech Republic
**2** National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Brno, Czech Republic
**3** Protein Data Bank in Europe (PDBe), European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, UK
**4** RCSB Protein Data Bank, San Diego Supercomputer Center University of California, San Diego, La Jolla, CA 92093, USA
**5** RCSB Protein Data Bank, Institute for Quantitative Biomedicine Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA
**6** Cancer Institute of New Jersey, Rutgers The State University of New Jersey, New Brunswick, NJ 08903, USA
**7** Skaggs School of Pharmacy and Pharmaceutical Sciences University of California, San Diego, La Jolla, CA 92093, USA

# These authors contributed equally to this work.
* alex.rose@rcsb.org

## Abstract

3D macromolecular structural data is growing ever more complex and plentiful in the wake of substantive advances in experimental and computational structure determination methods including macromolecular crystallography, cryo-electron microscopy, and integrative methods. Efficient means of working with 3D macromolecular structural data for archiving, analyses, and visualization are central to facilitating interoperability and reusability in compliance with the FAIR Principles. We address two challenges posed by growth in data size and complexity. First, data size is reduced by bespoke compression techniques. Second, complexity is managed through improved software tooling and fully leveraging available data dictionary schemas. To this end, we introduce BinaryCIF, a serialization of Crystallographic Information File (CIF) format files that maintains full compatibility to related data schemas, such as PDBx/mmCIF, while reducing file sizes

Manuscript submitted to Acta D - CCP4 study weekend



Acta Crystallographica Section D research papers

## High-performance macromolecular data delivery and visualization for the web

Authors

**David Sehnal[abc]*, Radka Svobodová[ab]*, Karel Berka[d], Alexander S. Rose[e], Stephen K. Burley[fg], Sameer Velankar[c] and Jaroslav Koča[ab]**

[a]Centre for Structural Biology, CEITEC – Central European Institute of Technology, Masaryk University, Kamenice 753/5, Brno, 625 00, Czech Republic

[b]National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Kamenice 753/5, Brno, 625 00, Czech Republic

[c] Protein Data Bank in Europe (PDBe), European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, United Kingdom

[d]Regional Centre of Advanced Technologies and Materials, Department of Physical Chemistry, Faculty of Science, Palacký University Olomouc, Šlechtitelů 241/27, Olomouc, 779 00, Czech Republic

[e]Research Collaboratory for Structural Bioinformatics (RCSB), San Diego Supercomputer Center, University of California San Diego, 9500 Gilman Drive, La Jolla, San Diego, California, CA 92093-0743, United States

[f]Research Collaboratory for Structural Bioinformatics (RCSB), Institute for Quantitative Biomedicine, Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, 174 Frelinghuysen Rd, Piscataway, New Jersey, NJ 08854-8076, United States

[g] Cancer Institute of New Jersey, Rutgers, The State University of New Jersey, 195 Little Albany Street, New Brunswick, New Jersey, NJ 08903-2681, United States

Correspondence email: david.sehnal@mail.muni.cz; radka.svobodova@ceitec.muni.cz

**Funding information** Ministry of Education, Youth and Sports of the Czech Republic - project CEITEC 2020 (grant No. LQ1601); ELIXIR-CZ research infrastructure project including access to computing and storage facilities (grant No. LM2018131); European Regional Development Fund – project ELIXIR-CZ (grant No. CZ.02.1.01/0.0/0.0/16_013/0001777).

**Synopsis** This article provides a survey of available web services and tools for data delivery and visualization of macromolecular structures.

# wwPDB Associate Members

**PDB China**

National Facility for Protein Science in Shanghai (NFPS) and iHuman Institute and SIAIS, Shanghai Tech University, Pudong, Shanghai, China

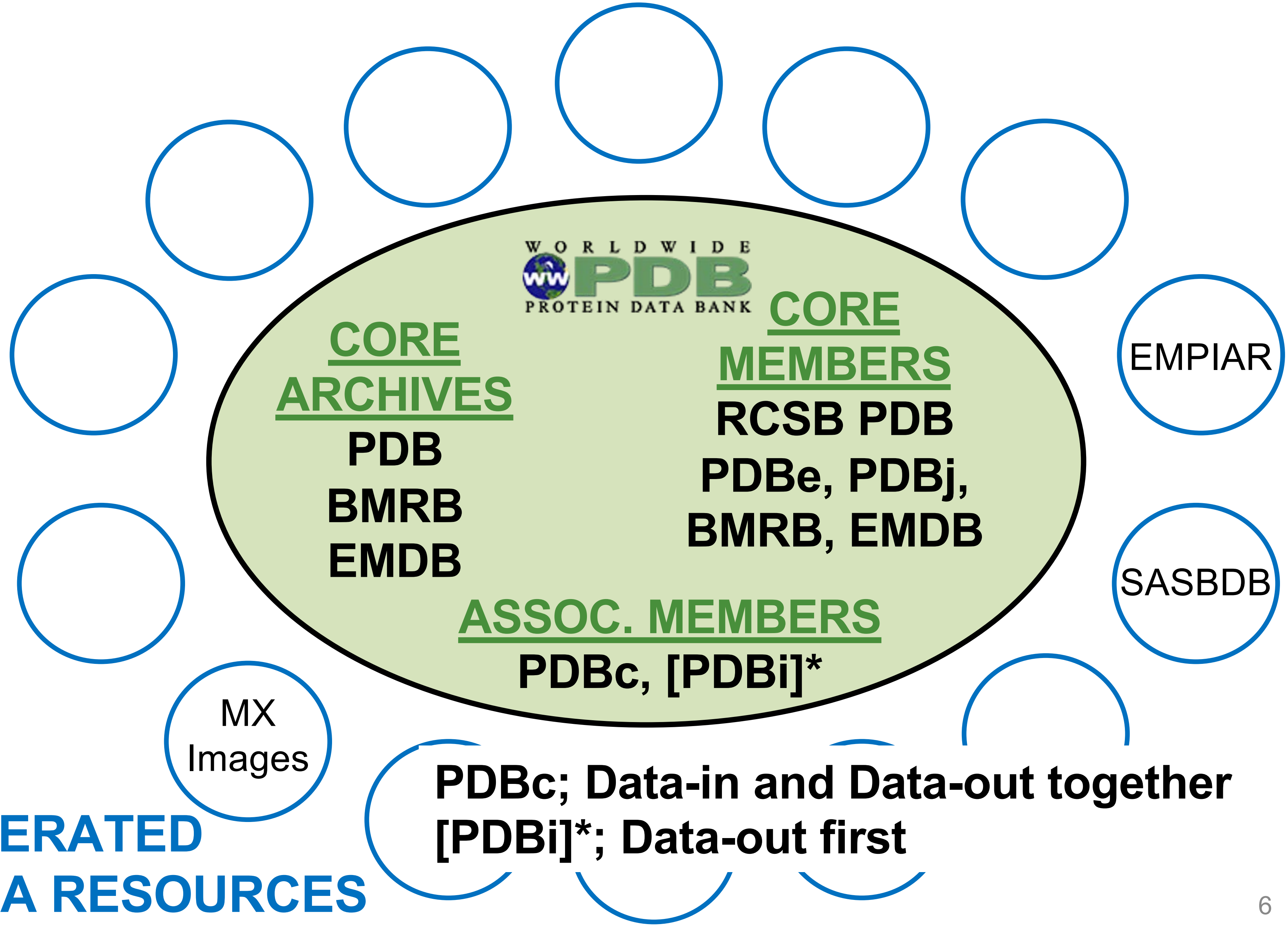
- Director, Wenqing Xu

**PDB India**

Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India

- PI, Manju Bansal
- Co-Investigator, Debasisa Mohanty and K. Sekar

# wwPDB Future Architecture



**CORE ARCHIVES**

**PDB**

**BMRB**

**EMDB**

**CORE MEMBERS**

**RCSB PDB**

**PDBe, PDBj,**

**BMRB, EMDB**

**ASSOC. MEMBERS**

**PDBc, [PDBi]***

**PDBc; Data-in and Data-out together**
**[PDBi]*; Data-out first**

**FEDERATED DATA RESOURCES**

# Implementation Plan for Data-out activities at PDBc and PDBi

- Background Training (remote)
  - Setting up the original Data-out services

- Hardware setup (local with remote support)
  - wwPDB authorized ftp service
  - Setting up the original pdbc.org or pdbi.org web sites.

# Implementation Plan for Data-in activity at PDBc (remote/onsite)

Remote Training of PDBc biocurators
(by RCSB/PDBe/PDBj)

- Scientific Training
- OneDep system Education
- OneDep system training

Onsite Training at PDBj (by PDBj)

- Invitation of PDBj members (postponed)
- Onsite OneDep system training (postponed)

# Implementation Plan of OneDep system for PDBc

**A similar scenario when PDBj started processing in 2000**

- Setting up PDBc’s OneDep system at Osaka (at  their own cost)
- Adding PDBc biocurators to the workflow manager
- PDBj will assign the depositions with PROC status in PDBc’s OneDep@Osaka
- Obviously, new PDBc biocurators cannot handle all depositions from China. PDBj biocurators should process the rest of data from China.
- Start with X-ray entry only

# Items for discussion

- PDBc should invite all wwPDB PIs to Shanghai to check their status. After approval by the wwPDB PIs, an official announcement that the data processing at Shanghai starting gradually will be announced to Chinese depositors.

- PDBj-BMRB will keep covering all BMRB deposition mainly from Asia.