Introductions and Overview of the wwPDB

Genji Kurisu

WORLDWIDE PDB
PROTEIN DATA BANK

wwpdb.org

Welcome

On behalf of the wwPDB/EMDB Principal Investigators

- PDBj : Genji Kurisu

- RCSB: Stephen K. Burley

- PDBe: Sameer Velankar

- BMRB: John L. Markley and Jeffrey C. Hoch

- EMDB: Ardan Patwardhan (apology)

Introductions

- Chair : Peter Rosenthal
- Co-Chair: Tsuyoshi Inoue

wwPDB Advisory Committee Members

- RCSB PDB: Paul Adams and Kirk L. Clark
- PDBe: Dave Brown and Susan Lea
- PDBj: Tsuyoshi Inoue and Masaki Yamamoto
- BMRB: Art Edison and Masatsune Kainosho
- EMDB: Sarah Butcher 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)
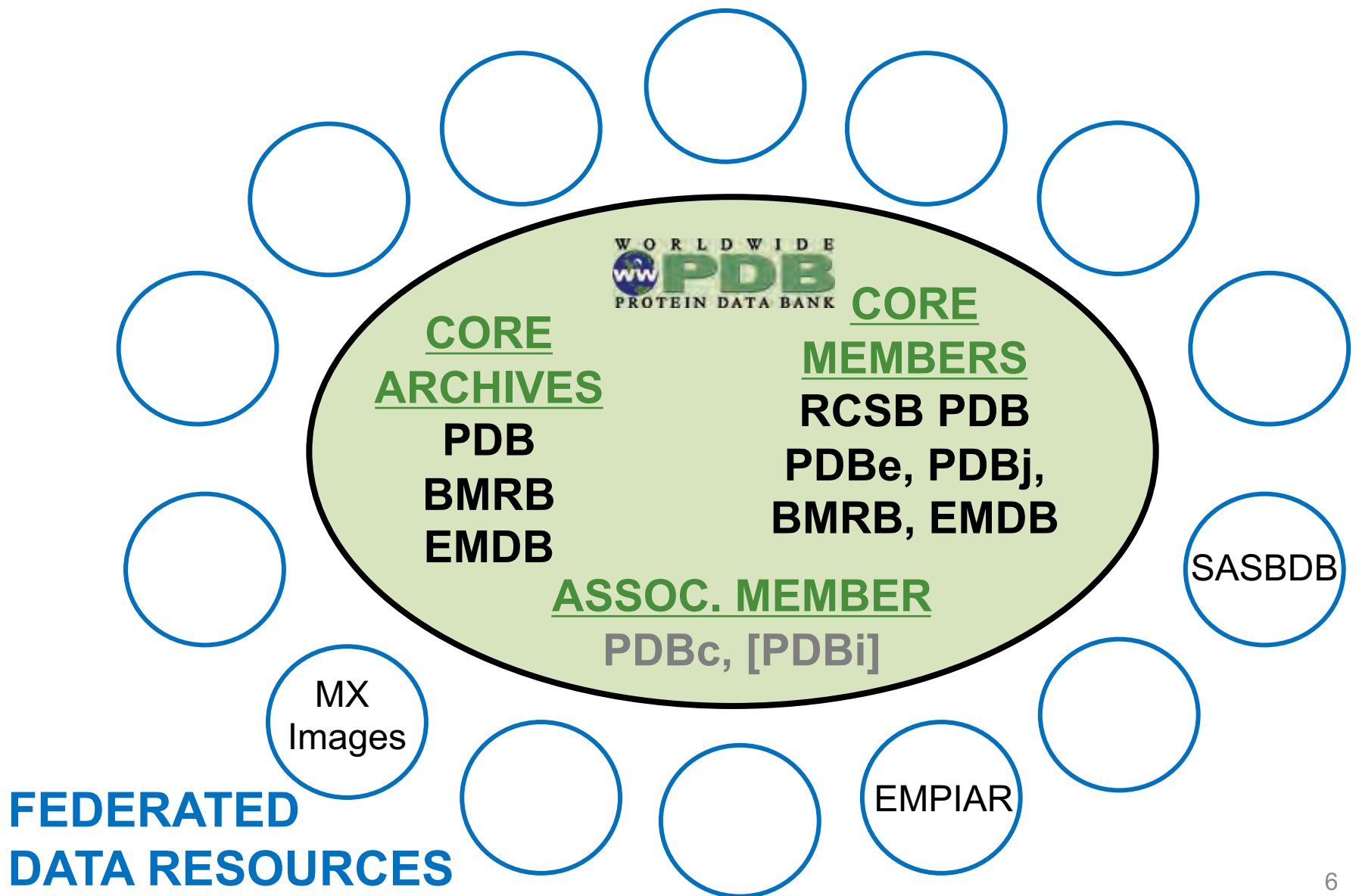
Introductions (cont.)
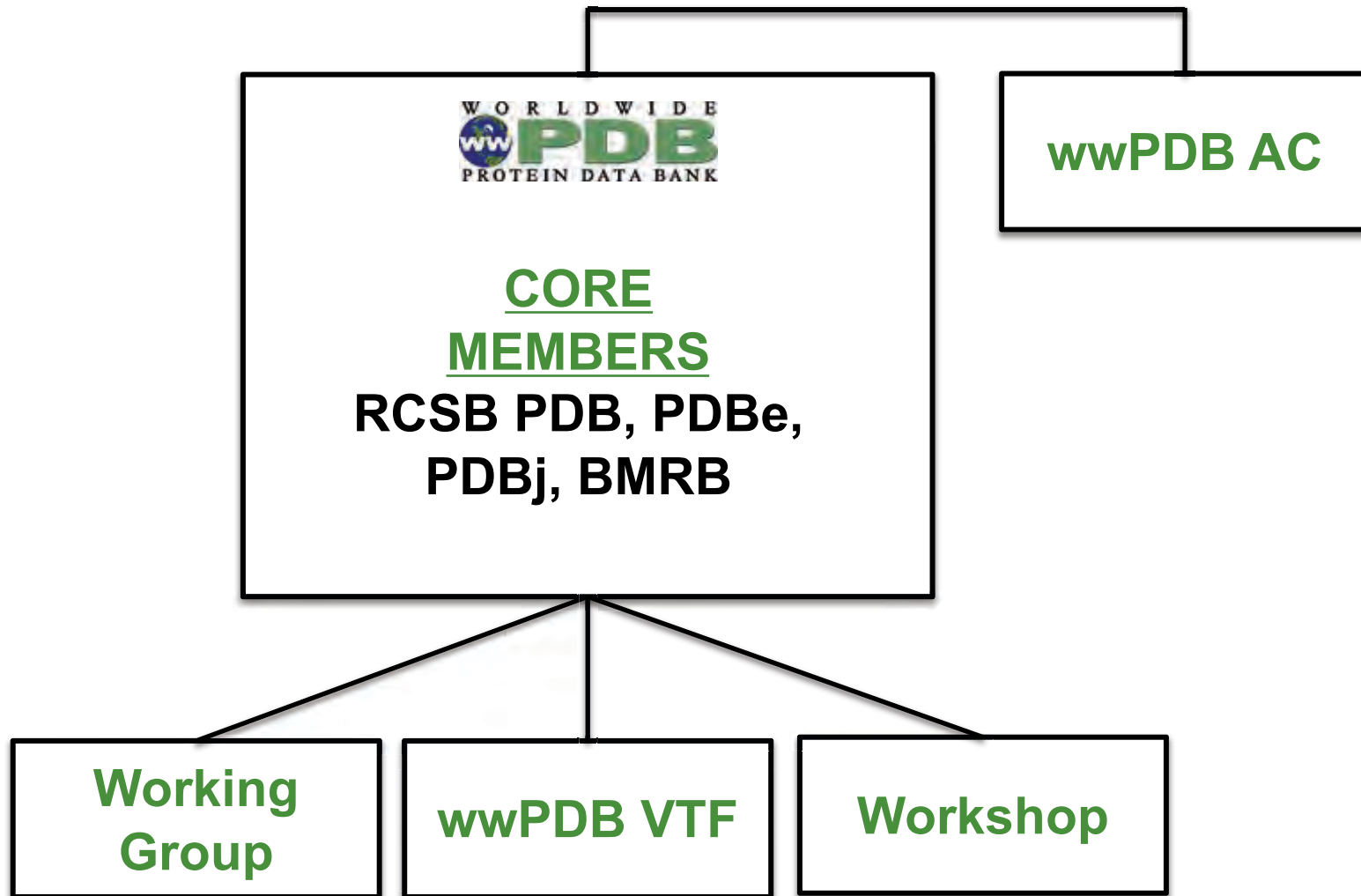
Observers

- Daisuke Kohda (PDBj-AC)

- Gert-Jan Bekker (PDBj)
- Yasuyo Ikegawa (PDBj)
- Kenji Mizuguchi (PDBj)
- Reiko Yamashita (PDBj)

- Takeshi Iwata (PDBj-BMRB)
- Yohei Miyanoiri (PDBj-BMRB)
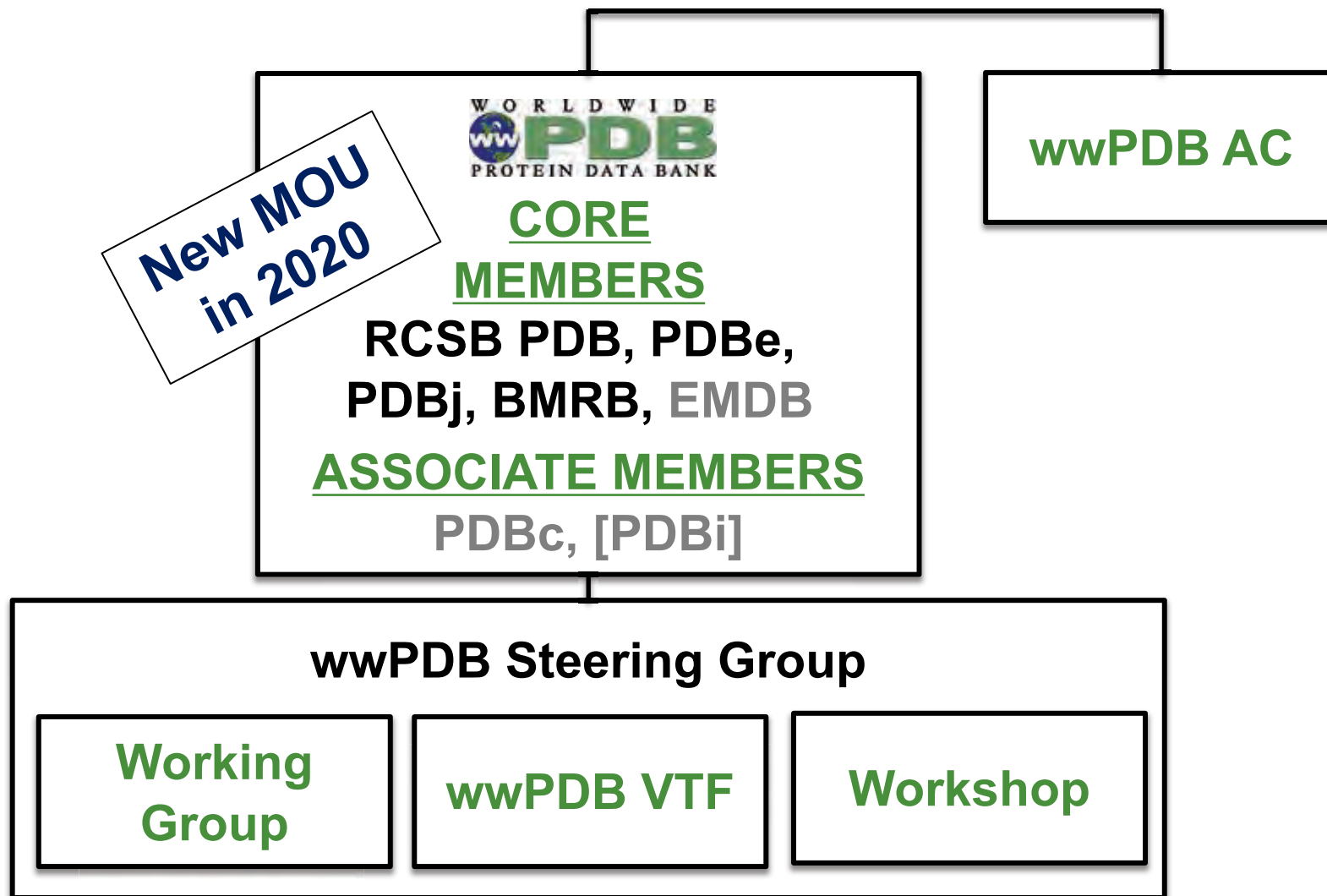
wwPDB Future Architecture

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

CORE MEMBERS
RCSB PDB, PDBe, PDBj, BMRB

wwPDB AC

Working Group

wwPDB VTF

Workshop

New wwPDB Organization

New MOU in 2020

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

Updated wwPDB Vision and Mission
Ref. Appendix 2

Developments since 2018 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
- Planned workshop on improving deposition and validation of single-particle EM data (January 2020)
- Finalized the new MOU including EMDB
- Supported PDBc Leadership appointment process

Developments since 2018 Meeting II

PDB Core Archive

- PDBx/mmCIF deposition for MX made mandatory
- As part of ligand validation development, ED map coefficients will be provided at the wwPDB FTP site
- Based on the collaboration with Global Phasing, the Buster Report code ported into OneDep
- PDB archive versioning in place
- Atomic coordinate replacement implemented

Developments since 2018 Meeting III

BMRB Core Archive I

- NMR-STAR dictionary enlarged with tags for non-uniform sampling and reconstruction, and structural restraint nomenclature

- Pipeline to calculate structures using X-Plor NIH with NMR-STAR as input file ready for testing

- Visualization tools in 'R' and Python for browsing and visualization of data in NMR-STAR format

- Visualization tools accessible through Jupyter notebooks and available through BMRB GitHub.

- BMRB R Library updated for new version of R.

- BMRBdep released as beta version (17 depositions)

Developments since 2018 Meeting IV

BMRB Core Archive II

- BMRB API migrated to new Python 3 Docker containerized version; GlusterFS stores API logs and shared data between cluster computers

- Delivered restraint validation package for use OneDep environment within wwPDB Validation Report

- NEF Translator along with support for NMR-STAR restraints file uploaded/implemented in OneDep System

- PLY (Python-Lex-Yacc) – based STAR parser developed and on GitHub

- BMRB APIs upgraded for use with BMRBdep

Developments since 2018 Meeting V

EMDB Core Archive

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

Individual wwPDB partner sites

- PDBj made a deal with EMPIAR to start EMPIAR-PDBj
- RCSB PDB renewed Funding 2019-2023, now includes Andrej Sali/UCSF
- PDBe renewed WT Funding 2020-2024, and launched PDBe Knowledge Base resource

PDB Core Archive Depositions

- 12,179 depositions in 2018.

- Rapid growth in 3DEM.
  - Exceeded NMR depositons

Method	2017 Depositions	2018 Depositions
MX	11,889 (91.1%)	10594 (87.0%)
NMR	460 (3.5%)	418 (3.4%)
3DEM	**674 (5.2%)**	**1140 (9.4%)**
Other	26 (0.2%)	27 (0.2%)

2018 Processing Sites

- RCSB PDB
- PDBe
- PDBj

42%
34%
24%
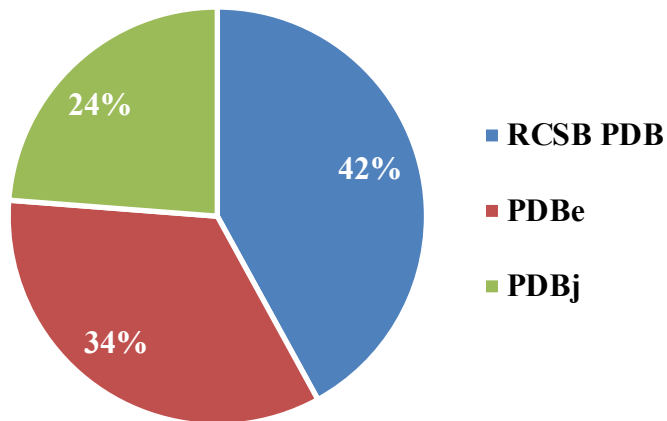
2018 Depositor Locations

- North America
- Europe
- Asia
- South America
- Oceania
- Africa
- Commercial

34%
33%
23%
6%
1%
3%
<1%
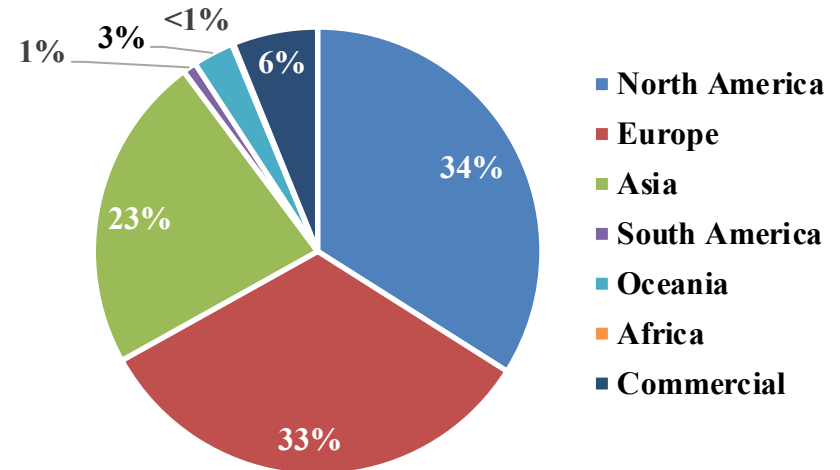
PDB Core Archive Growth
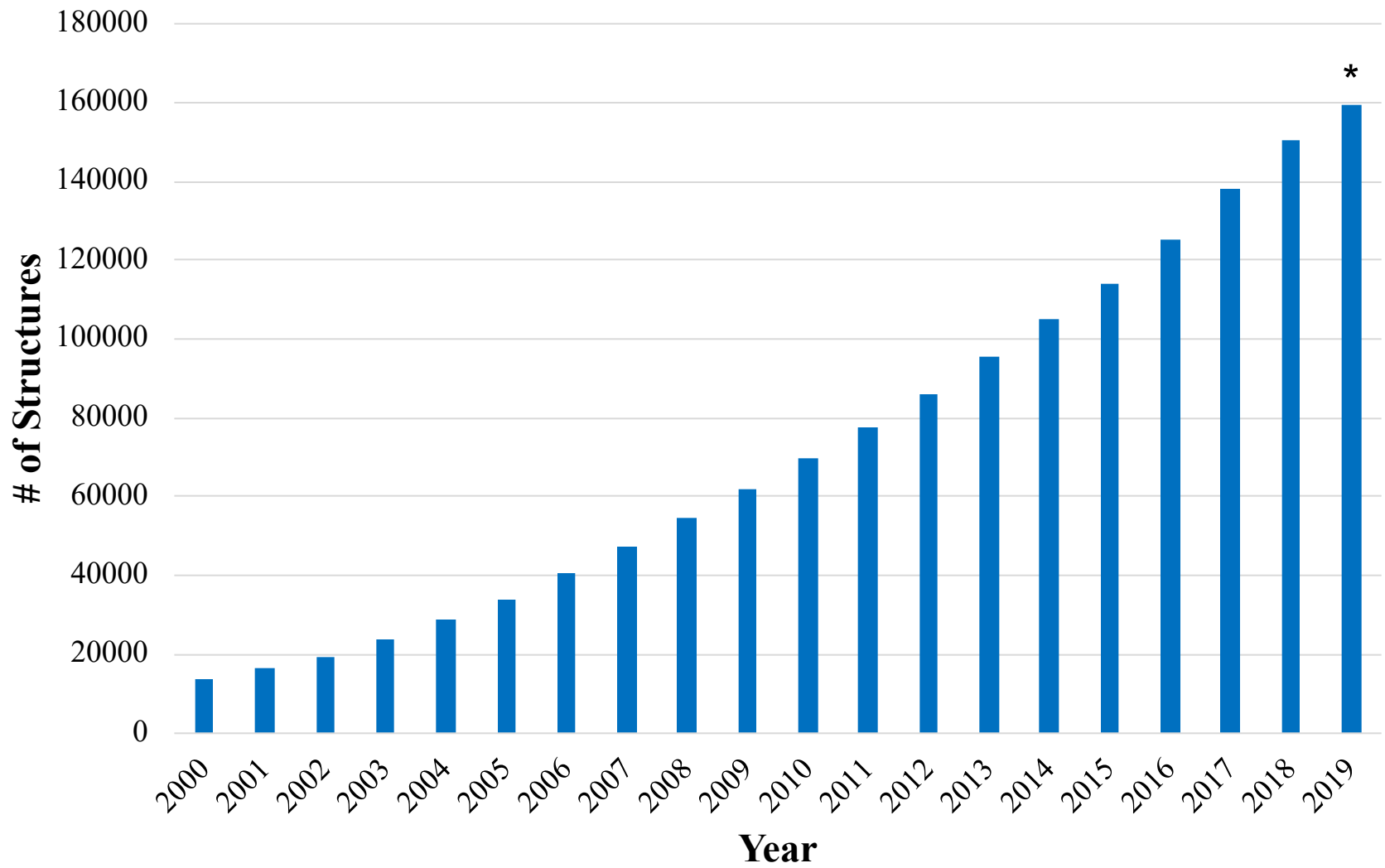
* As of 3 Sep 2019

PDB Core Archive Downloads

Year	Total	Total FTP Archive	Total Website
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. Hope to be on track for 2019→.

Geographic Origins of FTP downloads; 2012-2015

BMRB Core Archive Growth

- BMRB is on track for 773 new entries released so far in 2019.

- Total released entries estimated to reach ~13,600 by the end of 2019.
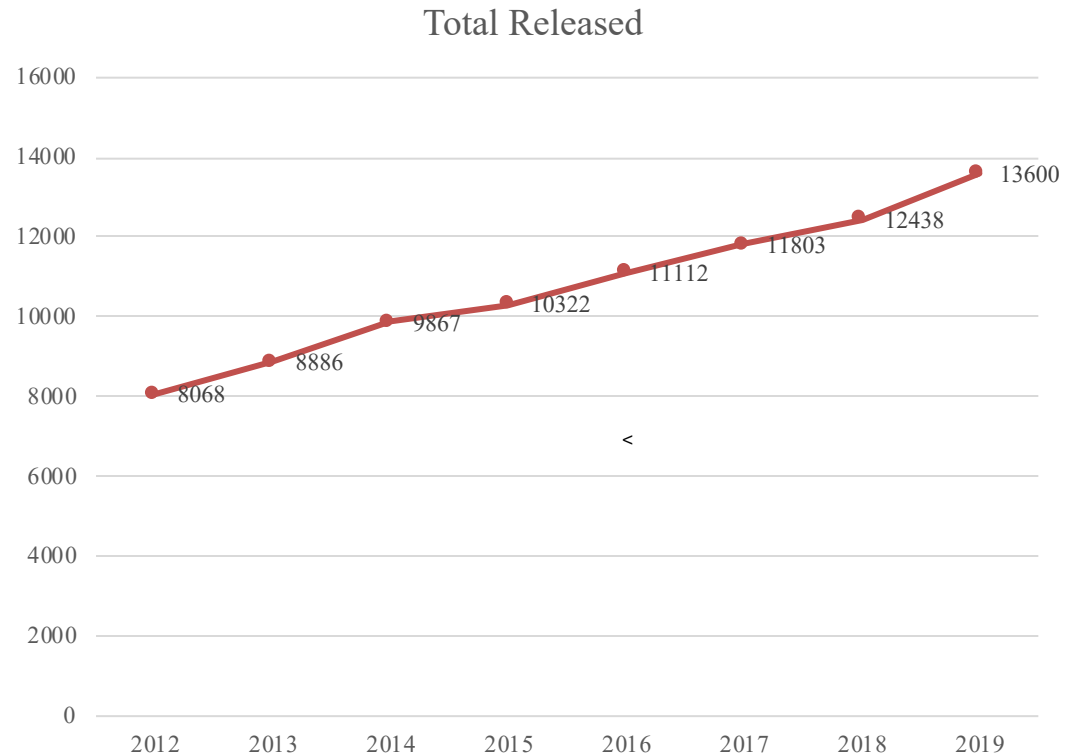
Total Released

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

BMRB Core Archive Growth

Internet Server Traffic (Website) – All Mirrors*

Year	Server requests	Page requests	File requests	Distinct hosts served	Total data transferred
2012	12,703,408	3,290,436	1,730,857	310,074	8.56 TB
2013	10,521,902	2,176,005	1,508,510	342,291	4.85 TB
2014	11,314,520	2,193,844	1,561,558	383,052	6.99 TB
2015	12,869,845	2,348,759	1,991,915	444,873	12.6 TB
2016	15,986,881	2,766,144	2,496,663	445,679	12.0 TB
2017	42,071,167	10,541,522	6,633,980	460,550	13.1 TB
2018	62,473,677	16,913,892	10,658,354	523,248	10.6 TB

~150K/day Server and page requests

* BMRB has mirror sites in Italy and Japan, and PDBj-BMRB branch for deposition

BMRB Core Archive Growth

Internet Server Traffic (FTP Servers) – All Mirrors*

Year	Server requests	Distinct files requested	Distinct hosts served	Total data transferred
2012	2,304,687	1,820,240	4,788	1.15 TB
2013	1,992,706	1,481,995	5,253	1.40 TB
2014	1,994,184	1,488,431	4,645	1.59 TB
2015	2,197,159	1,666,353	3,742	0.905 TB
2016	5,459,161	1,655,031	5,774	1.66 TB
2017	5,502,929	2,461,117	4,126	5.02 TB
2018	9,801,123	3,236,316	3,853	8.00 TB

EMDB Core Archive Depositions

- On track for ~2300 3DEM depositions in 2019.
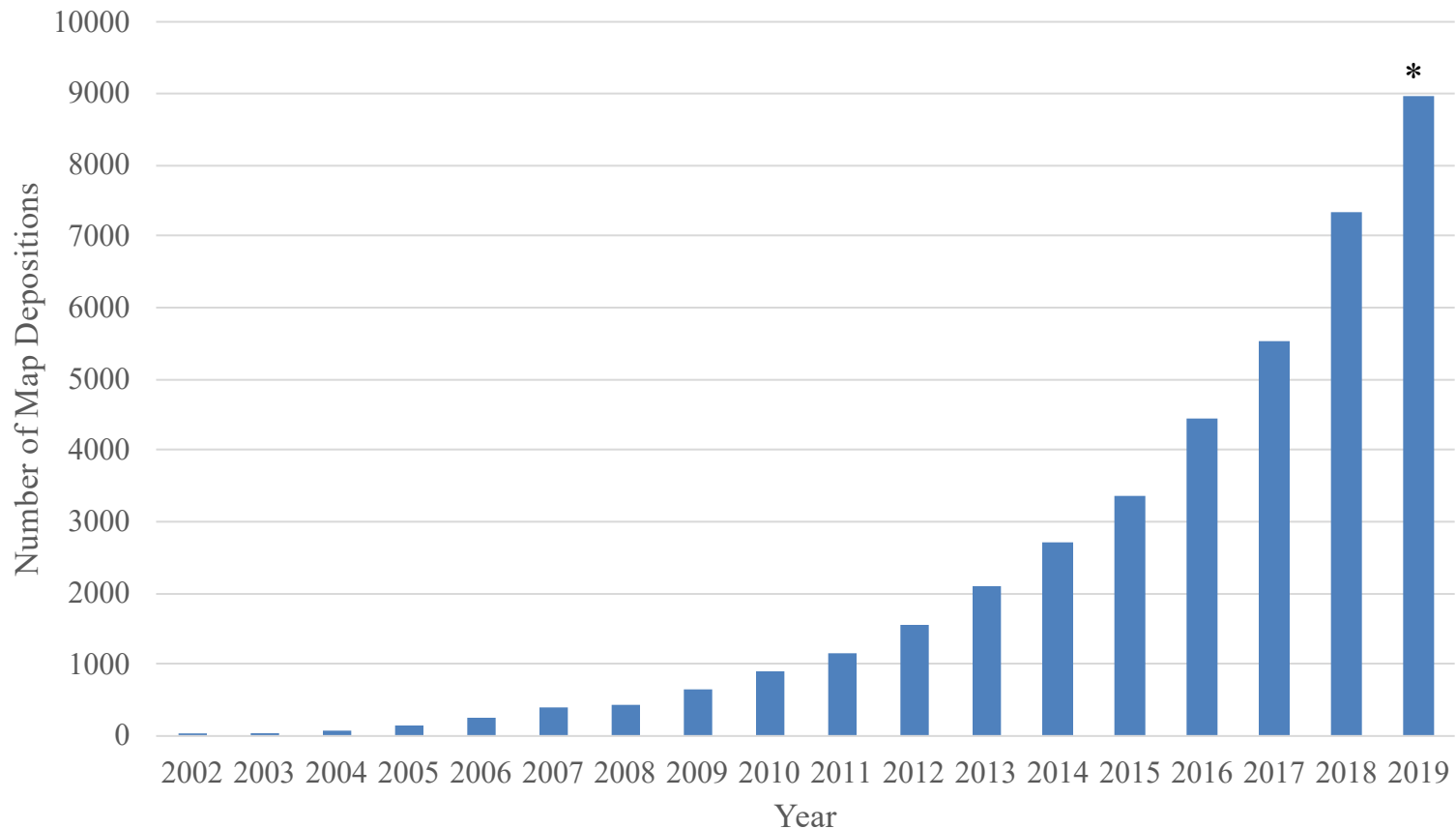- 936 out of 1692 have PDB entries in 2019.

Processing Site	2019 Depositions	2018 Depositions
PDBj	268	230
PDBe	599	523
RCSB	825	1026

2019 Processing Sites
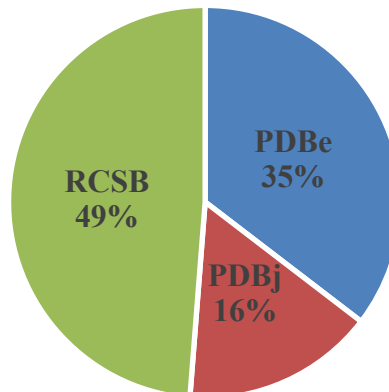
EMDB Core Archive Growth

* As of 4 September 2019

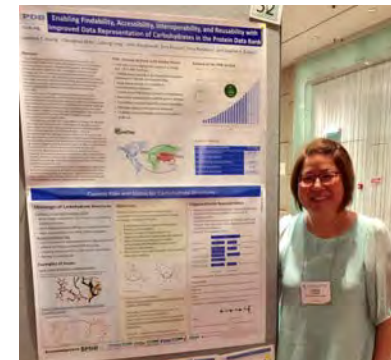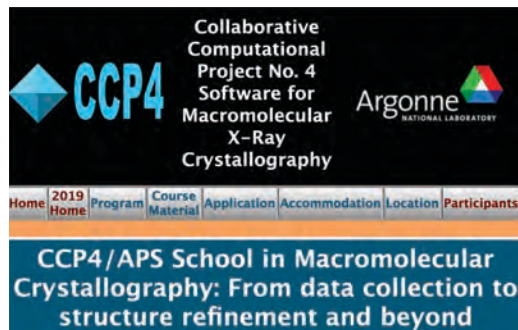# wwPDB Outreach

2019 OneDep Developer Summit

ACA

GRC Carbohydrates

Depositors Training

ECM32

AsCA

wwPDB 2018-2019 Publications

STRUCTURAL BIOLOGY
ISSN 2059-7983

Announcing mandatory submission of PDBx/mmCIF format files for crystallographic depositions to the Protein Data Bank (PDB)

Paul D. Adams,[a,b] Pavel V. Afonine,[a] Kumaran Baskaran,[c] Helen M. Berman,[d] John Berrisford,[e] Gerard Bricogne,[f] David G. Brown,[g] Stephen K. Burley,[d,h,i,a] Minyu Chen,[j] Zukang Feng,[d] Claus Flensburg,[f] Aleksandras Gutmanas,[e] Jeffrey C. Hoch,[c,*] Yasuyo Ikegawa,[j] Yumiko Kengaku,[j] Eugene Krissinel,[k] Genji Kurisu,[j,*] Yuhe Liang,[d] Dorothee Liebschner,[a] Lora Mak,[e] John L. Markley,[c,*] Nigel W. Moriarty,[a] Garib N. Murshudov,[m] Martin Noble,[n] Ezra Peisach,[d] Irina Persikova,[d] Billy K. Poon,[a] Oleg V. Sobolev,[a] Eldon L. Ulrich,[c] Sameer Velankar,[e,*] Clemens Vonrhein,[f] John Westbrook,[d] Marcin Wojdyr,[f,j] Masashi Yokochi[j] and Jasmine Y. Young[d]

[a]Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720 USA, [b]Department of Bioengineering, University of California, Berkeley, CA 94720, USA, [c]BioMagResBank (BMRB), University of Wisconsin-Madison, Madison, WI 53706, USA, [d]Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB), Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA, [e]Protein Data Bank in Europe (PDBe), European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK, [f]Global Phasing Limited, Sheraton House, Castle Park, Cambridge, CB3 0AX, UK, [g]School of Biosciences, University of Kent, Canterbury, Kent CT2 7NJ, UK, [h]Rutgers Cancer Institute of New Jersey, Robert Wood Johnson Medical School, New Brunswick, NJ 08903, USA, [i]Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB), San Diego Supercomputer Center, University of California, San Diego, La Jolla, CA 92093, USA, [j]Protein Data Bank Japan (PDBj), Institute for Protein Research, Osaka University, Osaka, 565-0871, Japan, [k]BioMagResBank (BMRB), UConn Health, 263 Farmington Avenue, Farmington, CT 06030, USA, [l]CCP4, Research Complex at Harwell (RCaH), Rutherford Appleton Laboratory, Didcot, Oxon OX11 0FA, UK, [m]MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge Biomedical Campus, Cambridge, CB2 0QH, UK, and [n]Newcastle University, Framlington Place, Newcastle Upon Tyne, NE2 4HH, UK. *Correspondence e-mail: stephen.burley@rcsb.org, hoch@uchc.edu, gkurisu@protein.osaka-u.ac.jp, markley@biochem.wisc.edu, sameer@ebi.ac.uk

The Protein Data Bank (PDB) (wwPDB consortium, 2019) is the single global archive of experimentally determined three-dimensional (3D) structure data of biological macromolecules. The continuing growth in the numbers, size and complexity of macromolecular structures in the PDB archive, coupled with the rapid growth of evolving experimental methods such as 3D cryo-electron microscopy (3DEM) has made the traditional PDB format ('legacy PDB format') inadequate for fully representing these data. As described below, this format was based on a punched-card format that became obsolete long ago. In the following letter, we describe the changes necessary to address the challenges coming from the extraordinary success of structural biologists.

Since 2003, the PDB has been managed by the Worldwide Protein Data Bank (wwPDB; https://www.wwpdb.org/) (Berman et al., 2003), an international partnership that collaboratively oversees deposition, validation, biocuration and open-access dissemination of 3D macromolecular structure data, adhering to the FAIR principles of Findability, Accessibility, Interoperability and Reusability (Wilkinson et al., 2016). In 2007, the master file format for the archive was officially changed to PDB Exchange/Macromolecular Crystallographic Information File (PDBx/mmCIF), supported by the PDBx/mmCIF data dictionary, to address new challenges in structure archiving. Later, in 2012, the wwPDB terminated its support of the legacy PDB file format and froze its further development (https://wwpdb.org/documentation/file-formats-and-the-pdb).

We now announce that as of 1 July 2019, PDBx/mmCIF will be the only format allowed for deposition of the atomic coordinates for PDB structures resulting from macromolecular crystallography (MX), including X-ray, neutron, fiber and electron diffraction methods, via OneDep (Young et al., 2017). This requirement will be extended to PDB structures resulting from nuclear magnetic resonance (NMR) spectroscopy and 3DEM methods at a later date to be determined. Elimination of the legacy PDB format will improve the efficiency of the deposition process and enhance validation through capture

---

Protein Data Bank: the single global archive for 3D macromolecular structure data

wwPDB consortium

ABSTRACT

The Protein Data Bank (PDB) is the single global archive of experimentally determined three-dimensional (3D) structure data of biological macromolecules. Since 2003, the PDB has been managed by the Worldwide Protein Data Bank (wwPDB; wwpdb.org), an international consortium that collaboratively oversees deposition, validation, biocuration, and open access dissemination of 3D macromolecular structure data. The PDB Core Archive houses 3D atomic coordinates of more than 144 000 structural models of proteins, DNA/RNA, and their complexes with metals and small molecules and related experimental data and metadata. Structure and experimental data/metadata are also stored in the PDB Core Archive using the readily extensible ww-PDB PDBx/mmCIF master data format, which will continue to evolve as data/metadata from new experimental techniques and structure determination methods are incorporated by the wwPDB. Impacts of the recently developed universal wwPDB OneDep deposition/validation/biocuration system and various methods-specific wwPDB Validation Task Forces on improving the quality of structures and data housed in the PDB Core Archive are described together with current challenges and future plans.

INTRODUCTION

The Protein Data Bank (PDB, pdb.org) was established in 1971 as the first open-access, molecular data resource in biology (1). More than 47 years later, the PDB continues to serve as the single global repository for atomic-level, 3D structure data, making >144 000 experimentally-determined structures of proteins, DNA, and RNA, and their complexes with metal ions, drugs, and other small molecules freely available without restrictions on use. Since 2003, the PDB has been managed jointly by the Worldwide Protein Data Bank (wwPDB) consortium (2), in-

(6). The wwPDB partners are committed to ensuring adherence to the FAIR Principles of Findability-Accessibility-Interoperability-Reusability (7).

Today, the PDB is universally regarded as a core data resource essential for understanding the functional roles that macromolecules play in biology and medicine. Publication of new macromolecular structures in most scientific journals is contingent on mandatory deposition to the PDB of the 3D atomic coordinates comprising the structural model plus experimental data used to derive the structures and associated metadata. Many governmental and non-governmental research funders also require PDB deposition of unpublished macromolecular structure data. All of these 3D structural data are stored in one of two wwPDB Core Archives. The PDB Core Archive houses 3D atomic coordinates of >144 000 structural models of proteins, DNA/RNA, and their complexes with metals and small molecules. The PDB Core Archive also houses related experimental data/metadata from Macromolecular Crystallography (MX). The BioMagResBank (BMRB; www.bmrb.wisc.edu) Core Archive houses related experimental data/metadata from Nuclear Magnetic Resonance spectroscopy (NMR). The wwPDB partners work closely with the Electron Microscopy Data Bank (EMDB; emdb-empiar.org), which houses related experimental data/metadata from 3D Electron Microscopy (3DEM) and Electron Tomography (ET).

The PDB Core Archive has seen steady growth since its inception, with over 11,000 new structures plus experimental data/metadata released in 2017 (Figure 1A). In aggregate, most of the 3D structures (89.5%) in the PDB Core Archive were determined using macromolecular crystallography (MX), with the remainder determined by NMR (8.5%), 3DEM (1.6%), and other techniques (0.4%). These overall metrics mask recent trends, which show that in 2016 3DEM overtook NMR as the second most popular technique for determining atomic level structures (Figure 1B).
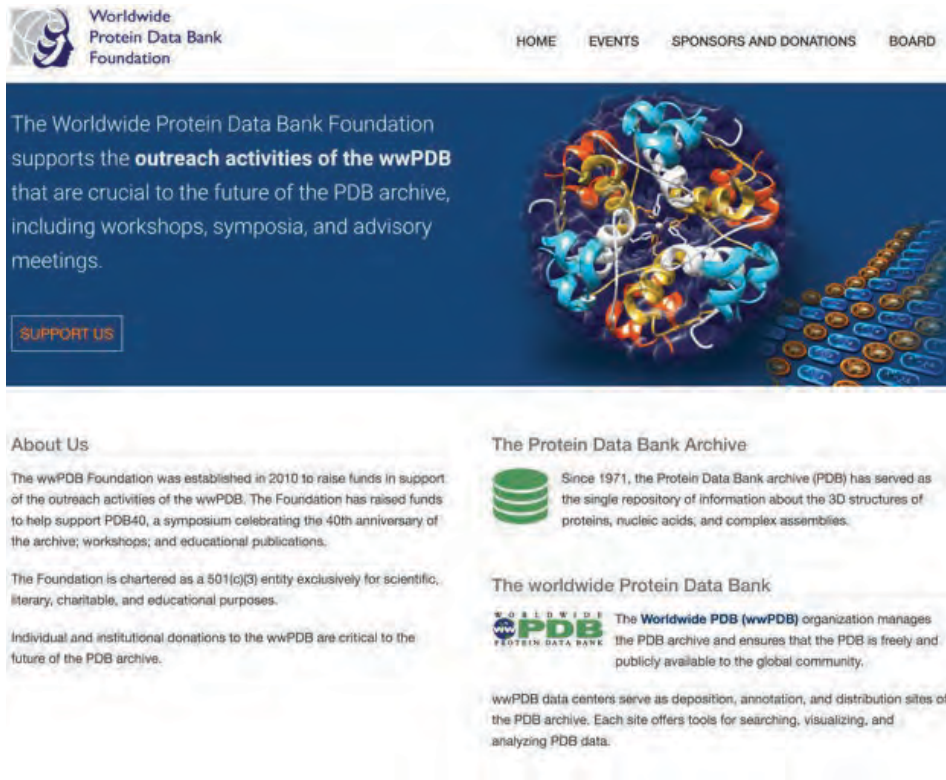
While the PDB Core Archive has grown enormously in scale and scope over the past 47 years and its management has evolved concurrently, adherence to the principle of open access and commitment to community engagement (1) continue to this day.

Mandatory mmCIF deposition announced on 3rd April 2019
https://doi.org/10.1107/S2059798319004522
I/HM manuscript was submitted.

wwPDB Foundation Progress

http://foundation.wwpdb.org/

- Fundraising ongoing.
- 2019 events:
  - New Chair: Celia A Schiffer, U. of Massachusetts Medical School
- Planning for PDB50
  - May 5th, 2021@Indianapolis
  - Oct.20th-22nd @EMBL, Heidelberg
  - Dec. @K.L. Malaysia

wwPDB Collaboration Resource November 2018- October 2019

wwPDB Partner	Software Development	Production Maintenance/ Management	Requirements Setting/ Testing	*Core Archive Keeping	Outreach	Biocuration/ Remediation	Total FTE Commitments
RCSB PDB	3.0	1.8	0.35/0.35	2.0	0.2	6.0	13.7
PDBe	1.6	1.0	0.35/0.35	-	0.2	4.2	7.7
PDBj	0.4	0.4	0.2/0.2	-	0.2	4.2	5.6
BMRB	1.5	-	-/-	1.5	0.2	0.2	3.9
EMDB	1	0.1	-/-	0.9	-	0.5	2.5
Total wwPDB	7.5	3.3	0.9/0.9	4.4	0.8	15.1	33.4

* Resource from Archive Keeper

Additional resources for software development made available by PDBj, EMDB and BMRB.

OneDep 2018/2019 Progress *vs*. Goals
Ref. Appendix 3

Delivered,
To be delivered,
Delayed

	Projects	2018 Q4	2019 Q1	2019 Q2	2019 Q3	2019 Q4
1. Validation	1.1 Ligand validation		Delivered	Delivered		
	1.2 NMR restraint validation			Delayed	Delayed	
	1.3 EM validation			Delayed		
	1.4 Update validation pipeline to handle new Carbohydrate representation				To be delivered	To be delivered
	1.5 Annual recalculation of validation reports					
2. Backend Stabilization	2.1 Github software release	Delivered	Delivered			
	2.2 Enable external computing	Delivered	Delivered			
	2.3 Deposition Django software upgrade					
	2.4 Python software upgrade				Delivered	Delivered
3. Public facing (OneDep or wwPDB.ORG)	3.1 Improve EM deposition		Delivered			
	3.2 Mandatory mmCIF deposition for MX structures			Delivered		
	3.3 Enable depositor-initiated coordinate versioning				Delivered	
	3.4 Enable NEF upload for NMR restraint validation				Delayed	Delayed
	3.5 ORCiD Authentication					To be delivered
	3.6 DOI resolution at wwpdb.org landing page				Delivered	Delivered
	3.7 Interfacing OneDep w/ BMRBdep				To be delivered	To be delivered
4. Biocuration	4.1 Provide calculated ED map coefficients to depositors			Delivered		
	4.2 NMR NEF processing				To be delivered	To be delivered
	4.3 Incorporate carbohydrate representation in new entries				To be delivered	To be delivered
	4.4 Incorporate better auditing for chemical components in CCD				To be delivered	To be delivered
5. Archive Improvements	5.1 Carbohydrate remediation	To be delivered	To be delivered	To be delivered	To be delivered	To be delivered
	5.2 Calculated ED map coefficients for new entries				To be delivered	To be delivered
	5.3 Protein Modification remediation requirements and planning		Delayed	Delayed	Delayed	Delayed
	5.4 CCD Versioning					To be delivered

28

2018/2019 Progress *vs.* Goals I

- Improved Biocuration processes
- Better software management *via* GitHub
- Enabled use of external computing resource
- Implemented Ligand Validation Workshop recommendations (2D depictions and ED maps)
- Made mmCIF deposition mandatory for MX
- Enabled Author-initiated Coordinate Replacement
- Provided wwPDB DOI landing page
- Upgraded Python software to support both V2 and V3
- Carbohydrate remediation (Ongoing)

2018/2019 Progress *vs.* Goals II
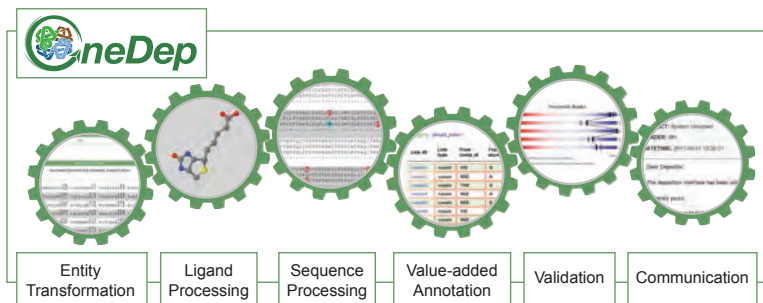
Re-forecasted

- Implement NMR restraint validation and NEF deposition
- Implement EM map validation
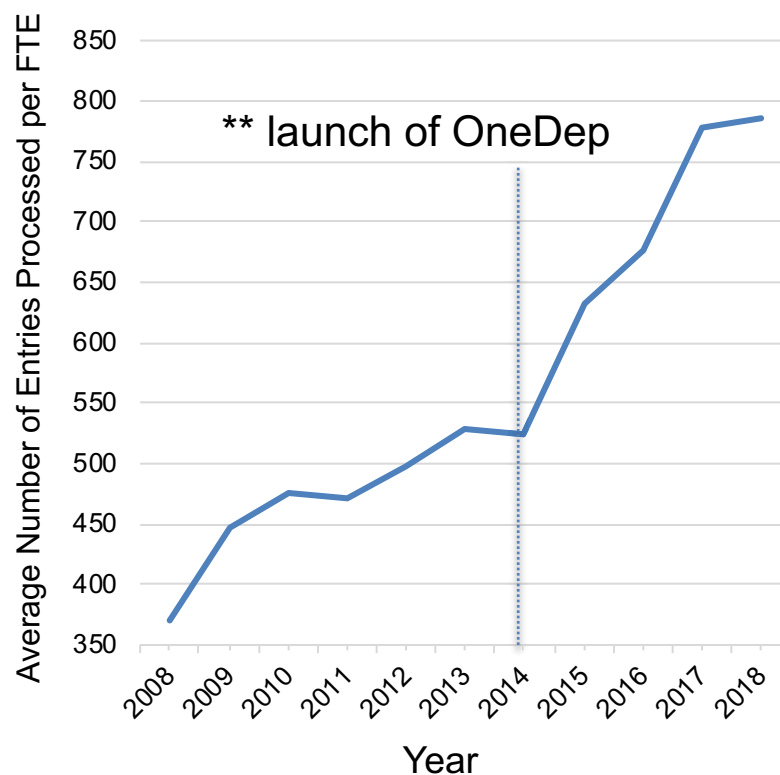- Post-Translational Modification project planning

Mitigation

- Actively engage NMR community in 2019-2020
- Set clear requirements and phased plan for EM validation
- Resource recruiting for PTM remediation

wwPDB Biocurator Productivity

- Continuing increased efficiency since 2009
- Significant increase with OneDep system
- Ongoing improvements in Biocuration processes
- Further "bending of the curve" is required

New Structures/wwPDB Biocurator

** launch of OneDep

| Entity Transformation | Ligand Processing | Sequence Processing | Value-added Annotation | Validation | Communication |

wwPDB DOI Resolution

- PDB ID with DOI 10.2210/pdb5AT1/pdb

- Redirect→wwPDB landing page via CrossRef (doi.org)

- Provides access to all data files and websites.

- Communication plan with journals under development

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

- PDBe: EMBL-EBI, Wellcome Trust: 2020→2024

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

- EMDB: EMBL-EBI, Wellcome Trust: 2019→2023

wwPDB Collaboration Resources

November 2018-October 2019

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

* RCSB PDB; EMDB; BMRB

OneDep 2019/2020 Goal Setting I

	Major Projects	Primary resource
Validation	**Implement EM map validation**	EMDB
	Implement NMR restraint validation	BMRB
	Provide carbohydrate presentation	PDBe
	Provide mmCIF formatted validation report	RCSB PDB
Public facing	Coordinate versioning for legacy entries	PDBe
	Improve ligand validation workflow at deposition	PDBe
Annotation	Implement new carbohydrate representation	RCSB PDB
	Use author's ligand definition for new ligand creation	RCSB PDB
	Automate audit revision when updating ligand definitions	RCSB PDB
Backend	Software upgrade (move to python 3)	RCSB PDB, PDBe
	mmCIF Dictionary maintenance	RCSB PDB
PDB Archive	Carbohydrate remediation (Review and production)	RCSB PDB, PDBe/j
	Recalculation of validation reports and ED map coefficients	PDBe
	PTM remediation (Requirement setting and planning)	PDBe
	Add ligand connectivity and extended PDB ID to model files	RCSB PDB, PDBe/j
	NMR restraint remediation	BMRB

Bold: re-forecasted from 2018/2019

OneDep 2019/2020 Goal Setting II

Timeline
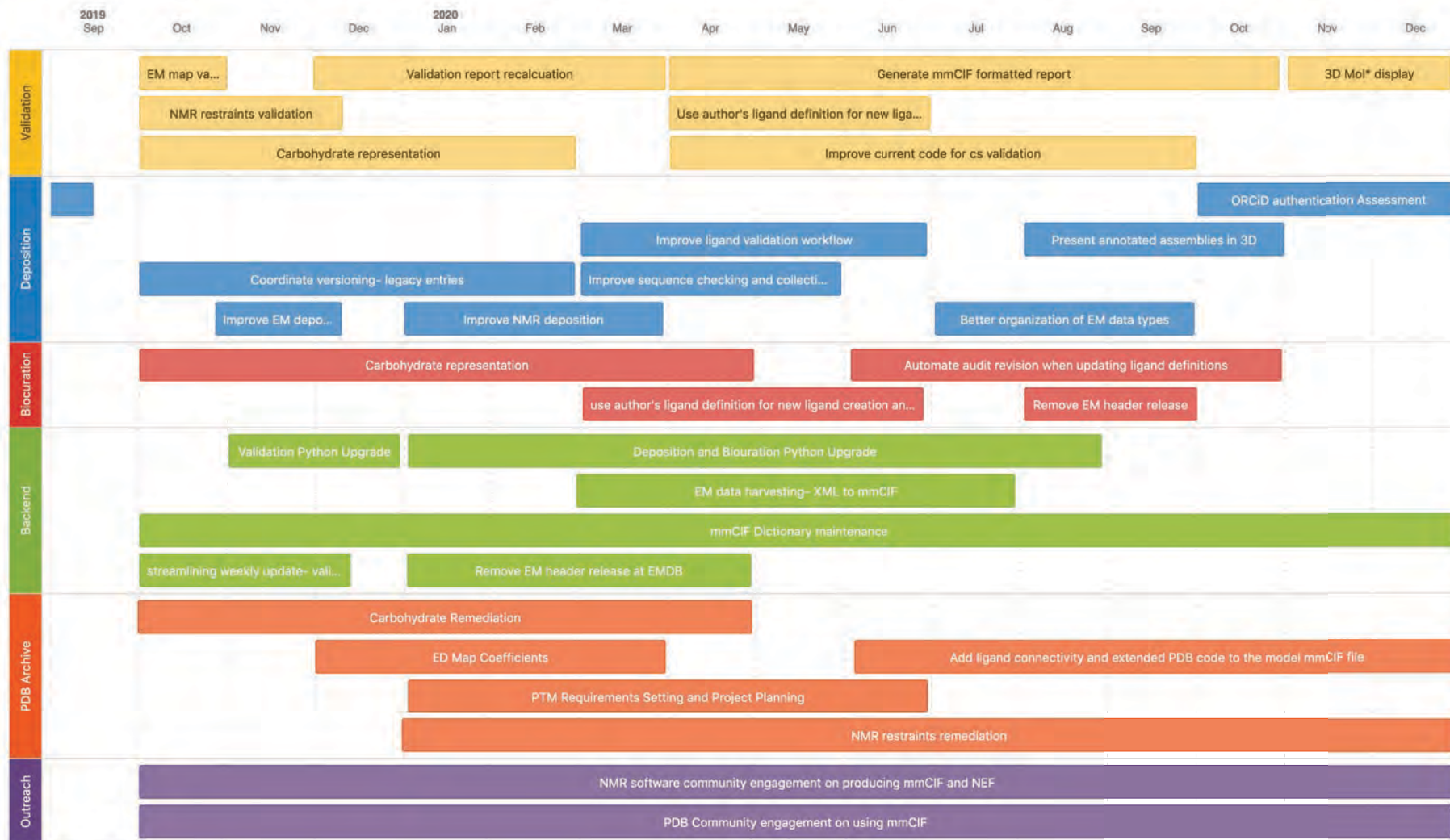
* Timeline will be further refined after requirement setting.

EM Data-Management Workshop

EMBL-EBI: January 23-24, 2020

- Developing a mechanism to improve EM data deposition to PDB and EMDB Core Archives
  - Formulate a list of mandatory and good-to-have data items for PDB and EMDB
  - Identify missing data categories and items
  - Update data dictionaries for PDB and EMDB
  - Discuss making PDBx/mmCIF format atomic coordinate deposition mandatory for EM (as for MX), and identify stakeholders and pain points and how wwPDB can support software developers

EM Data-Management Workshop

EMBL-EBI: January 23-24, 2020

- Review updated validation report and provide feedback
- Discuss additional model, map, and map-model validation metrics for single-particle EM depositions
- Discuss deposition of segmented maps with multiscale model representation

BMRB to Host 2020 wwPDB AC

- Next wwPDB AC meeting: Friday, Oct. 2nd 2020
  Host: BMRB
  Venue: University of Wisconsin, Madison, WI, USA

- 2021 wwPDB AC meeting: 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

Plan of PDB China

Wenqing Xu

WORLDWIDE PDB
PROTEIN DATA BANK

wwpdb.org

Getting ready for PDB-China

Joint effort of: **National Facility for Protein Science in Shanghai (NFPS)**
                                  iHuman Institute and SIAIS, ShanghaiTech University
Location: **Pudong, Shanghai, China**

Leadership for PDB-China (PDBc):

Prof. Zihe Rao: Academician of CAS; SAB Chair of NFPS
Prof. Ray Stevens: Founding director of iHuman Institute
Prof. Wenqing Xu: Director of NFPS
Prof. Zhijie Liu: Executive director of iHuman Institute
Prof. Ge Jiang: Executive director of SIAIS, vice-provost of ShanghaiTech Univ.

Jiang   Stevens   Rao   Xu   Liu

Key personnel:
Profs. Wenqing Xu, Zhijie Liu: Co-directors of PDBc; Prof. Zhipu Luo: deputy director of PDBc

Budget/committed funding:
NFPS has secured funding of 16M RMB¥ (~2.3M USD), from the Shanghai Science Commission, for PDB-China this year. The iHuman Institute will provide additional funding for 10 FTEs.

Promised long-term funding: NFPS (from the Zhangjiang National Lab), ShanghaiTech Univ., Chinese Academy of Science (CAS), and Shanghai municipal government.

Mission Statement: By serving as a wwPDB member, PDB-China will contribute to: (1) Manage the wwPDB Core Archives as a public good according to the FAIR Principles. (2) Provide expert deposition, validation, biocuration and remediation services at no charge to Data Depositors worldwide. (3) Ensure universal open access to public domain structural biology data with no limitations on usage. (4) Develop and promote community-endorsed data standards for archiving and exchange of global structural biology data.
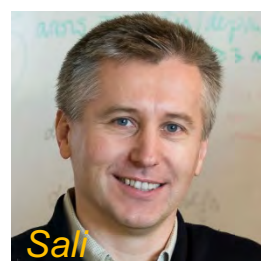
Getting ready for PDB-China

Data-in activity plan:

Key personnel:
Prof. Zhipu Luo (X-ray and general)
Dr. Bin Wu (NMR facility manager at NFPS)
Prof. Rongguang Zhang, lead of SSRF bio-stations (X-ray, part time)
(will increase to 5 FTE dedicated to data-in)

Key collaboration/consortium:
Prof. Quan Wang, Director of the Cryo-EM facility of ShanghaiTech University
Prof. Fei Sun, Director of the Cryo-EM facility at the Institute of Biophysics, CAS
Prof. Andrej Sali, Professor at iHuman Institute, a leader in the integrative structural biology field.
(Profs. Andrej Sali and Fei Sun will likely join the SAB of PDB-China in the near future).

Activity:
Upgrades for PDBc computation, backup and support systems are ongoing.

Plan:
To start on-site training by wwPDB in 2020;
To initiate data-in activity following completion of training.

Getting ready for PDB-China

<u>Data-out activity plan:</u>

<u>Key personnel:</u>
Prof. Zhipu Luo (X-ray and general)
Ms. Tingting Wang (IT and website design at NFPS)
Mr. Haiwei Fan (deputy director, IT Center, ShanghaiTech Univ.)
(will increase to 5 FTE dedicated to data-out)

<u>Key collaboration/consortium:</u>

EMDB-China (www.emdb-china.org.cn): will be merged to PDB-China.
(currently led by Prof. Fei Sun, at the Institute of Biophysics, CAS)

IT Center at the ShanghaiTech Univ.

Shanghai Center for Bioinformation Technology (eng.scbit.org/pages/index.do)

<u>Data-out Plan:</u>
To recruit additional members for the data-out team;
Upon data-out team buildup, merge EMDB-China into PDB-China (in 2020);
To optimize hardware and network setup;
To build PDBc user interface.

Plan of PDB India

Debasisa Mohanty

WORLDWIDE
wwPDB
PROTEIN DATA BANK

wwpdb.org

PDB-India

An Indian initiative on setting up a high fidelity structural data archival/retrieval system for Life Sciences – (PDBi)

Principal Investigator: **Prof. Manju Bansal**
Molecular Biophysics Unit
Indian Institute of Science (IISc)
Bangalore

Co-Investigators:
Prof. K. Sekar
Indian Institute of Science
Bangalore

Dr. Debasisa Mohanty
National Institute of Immunology
New Delhi

Location: **Indian Institute of Science, Bangalore**

Funding: National Supercomputing Mission (NSM), Govt. of India
under the Umbrella Program of
NSM Platform for Genomics & Drug Discovery (NPGDD)

Duration: 3 Years (October 2019 – September 2022)

45

PDB-India

Mission:

➢ Development and maintenance of a digital repository, The Protein Data Bank, India (PDBi) as a partner of wwPDB consortium for archiving, curation, validation and dissemination of experimental data on biomolecular structures.

➢ Development of user-friendly bioinformatics resources by integrating information from 3D structures with other forms of biological data for better understanding of biological pathways/processes.

Budget (3 Years):

➢ Salaries for staff (8 postdoctoral scientists or IT experts) & training/workshops: **INR 45 Million (0.6 Million USD)**

➢ Computer hardware (Servers/Storage/Networking): ***To be provided separately by another division of NSM**

* *Hardware installation could be delayed as a different wing of NSM is involved*

PDB-India

ACTION PLAN

Year 1 (Oct 2019 – Sept 2020)

- Hiring of Staff by December 2019
- Efforts towards accelerating hardware procurement
- Setting up prototype Data-Out System on minimal hardware available at IISc & NII
- Training of staff with help from other wwPDB partners on installation of softwares and scripts for mirroring & Data-Out

Year 2 (Oct 2020 – Sept 2021)

- Installation of required hardware
- Setting up of full fledged PDBi Data-out and making it available to public
- Testing of Data-in & training of staff for curation/validation using One-Dep with help from wwPDB partners

Year 3 (Oct 2021 – Sept 2022)

- Setting up of full fledged PDBi Data-in & accept data deposition from India
- Development of structural bioinformatics resources
- Hosting of bioinformatics tools developed by other research groups in India on PDBi portal
- Efforts to get funding for next phase

Remaining Agenda Items

- Discussion

- Lunch

- PDB Core Archive Plans (SV)
- BMRB Core Archive Plans (JLM)
- EMDB Core Archive Plans (GK for AP)
- Questions for the Advisory Committee (SKB)
- Executive Session

- Dinner at 19:00

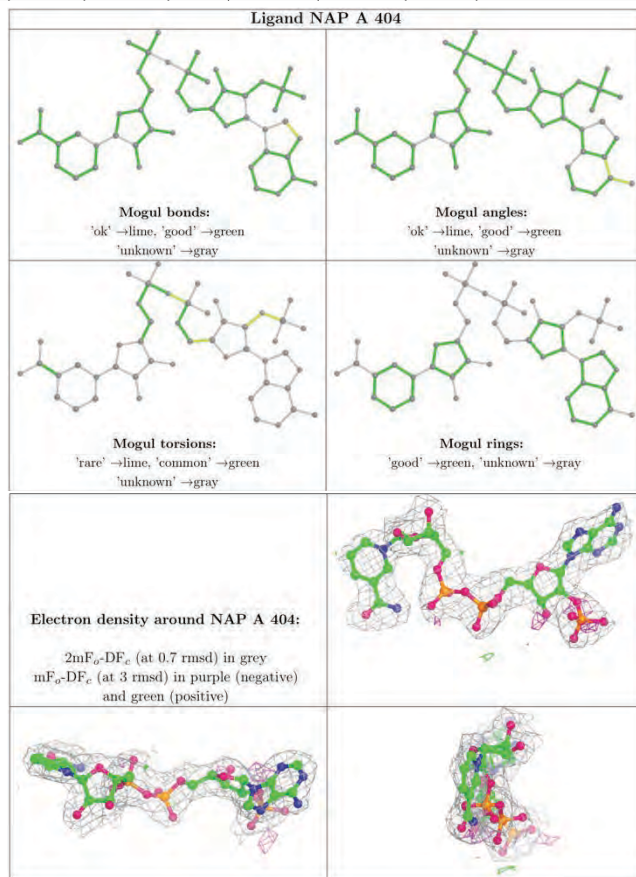# PDB Core Archive Update and Plans

Sameer Velankar

WORLDWIDE wwPDB

PROTEIN DATA BANK

wwpdb.org

PDB Core Archive Update I

Improved Ligand Validation

- Adapted software from Global Phasing Ltd. under a formal collaboration agreement with all wwPDB partners
- Benefits:
  - Provides geometrical quality in 2D depiction
  - Provides electron density fit for X-ray in 2D depiction
- Now mandatory at deposition: identification of Ligand(s) Of Interest (LOI, author's research focus)
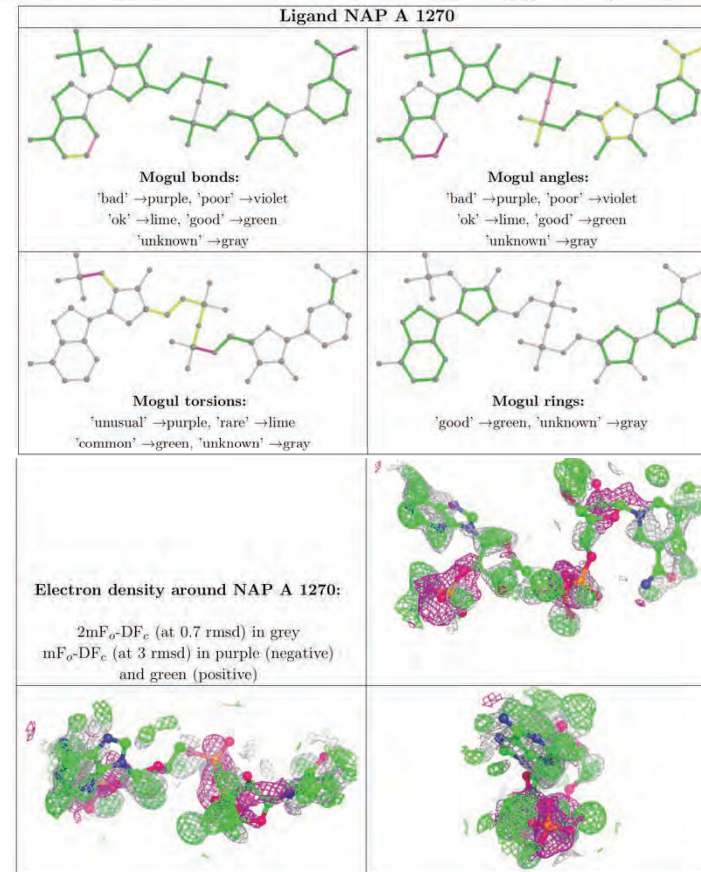 - 2D depictions provided in wwPDB Validation Report for all LOIs

Ligand Validation- NADP Examples

Mol	Type	Chain	Res	Atoms	RSCC	RSR	B-factors(Å²)	Q<0.9
3	NAP	A	404	48/48	0.96	0.14	31,43,66,70	0

Mol	Type	Chain	Res	Atoms	RSCC	RSR	B-factors(Å²)	Q<0.9
3	NAP	A	1270	48/48	-0.06	0.67	87,96,100,100	0

PDB entry 5zix (Better data quality)

PDB entry 1zk4 (Worse data quality)

3

PDB Core Archive Update II

<u>Coordinate Versioning to Improve Data Quality</u>

- Depositors can now make corrections to existing structures in the PDB Core Archive by updating the atomic coordinates preserving the original PDB identifier

- Deployed on July 27th 2019 for all structures deposited via OneDep (Phase one)

- 1st coordinate replacement (PDB ID 5T26) released on versioned at FTP August 7th 2019 following reviewers' comments based on the wwPDB validation reports

  - Moved two polymer chains to be in the same unit cell

  - Fixed water molecules on symmetry axis

ftp://ftp-versioned.wwpdb.org/pdb_versioned/data/entries/t2/pdb_00005t26/

PDB Core Archive Update III

<u>ORCiD Utilization to Improve Depositor Tracking</u>

- Enable better management of incoming data
- Credit PDB entries by author unambiguously
- 25% of all Unique Depositors have provided ORCiDs
- 3342 Unique PIs now have ORCiDs

ORC**iD**

Connecting Research and Researchers

PDB Core Archive Plans I

ED Map Coefficients

- ED map coefficients have been provided to depositors during biocuration for new entries

- As part of annual recalculation of validation reports, the ED map coefficients will be provided for existing entries at the wwPDB FTP site for better interpretation of the ligand validation by user community.

Coordinate Versioning (Phase 2 - Legacy Entries)

- Create OneDep deposition sessions

- Enable authors access and login

- Entries will be released at versioned FTP

PDB Core Archive Plans II

<u>Remediation to Improve Consistency and Searchability</u>
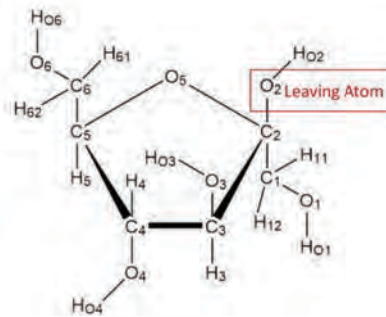
- Representation of carbohydrates (CHOs) led by RCSB PDB; post-translational modifications (PTM) led by PDBe will be the focus of PDB Core Archive plans in 2019 and 2020

  - 2019: Software development and testing for CHOs
  - 2020:
    - Q1: Data remediation and review of CHOs
    - Q2: Update CHO entries across PDB archive (~10%)
    - Re-forecasted: Requirement setting and project planning for PTM remediation

Details re Carbohydrate Remediation

- ~16,000 PDB Entries
- ~20,000 Oligosaccharides
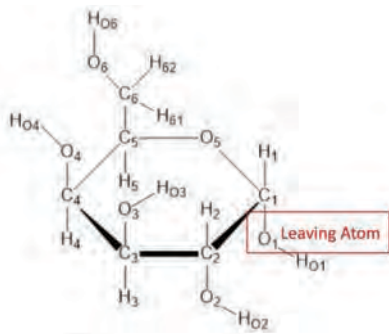
Alpha-D-Galactopyranose

Beta-D-Fructofuranose

Project summary:
https://www.wwpdb.org/documentation/carbohydrate-remediation

Examples of remediated data:
https://github.com/pdbxmmcifwg/carbohydrate-extension/tree/master/examples

DManpa1-3DManpb1-4DGlcpNAcb1-4[LFucpa1-3]DGlcpNAcb

- Project communicated with glycoscience, PDB and software developers communities via conferences/virtual meetings
- Standardized nomenclature following IUPAC/IUBMB
- Adopted glycoscience community software
- Provided uniform representation for oligosaccharides with community linear descriptor(s)
- To-do: Identify, validate, and biocurate glycosylation
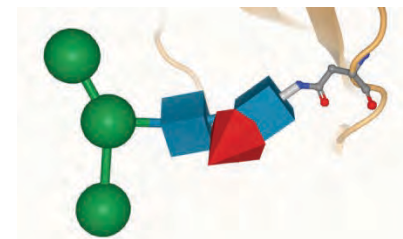
Current Data Distribution by Partners

- wwPDB partner sites dealing with PDB Core Archive (RCSB PDB, PDBj and PDBe) distribute the common PDB Core Archive FTP area

- RCSB PDB, PDBj and PDBe carry out individual processes to enrich PDB Core data for their websites

- RCSB PDB, PDBj and PDBe have site-specific FTP areas that provide additional site specific files

- PDBe has a weekly process to "clean up" PDB Core Archive information:

    - Jira tickets are added to ensure each site cleans up their entries – cleanup of PDB entries is ongoing

    - The "updated" files are distributed *via* PDBe website

Current Data Distribution by Partners
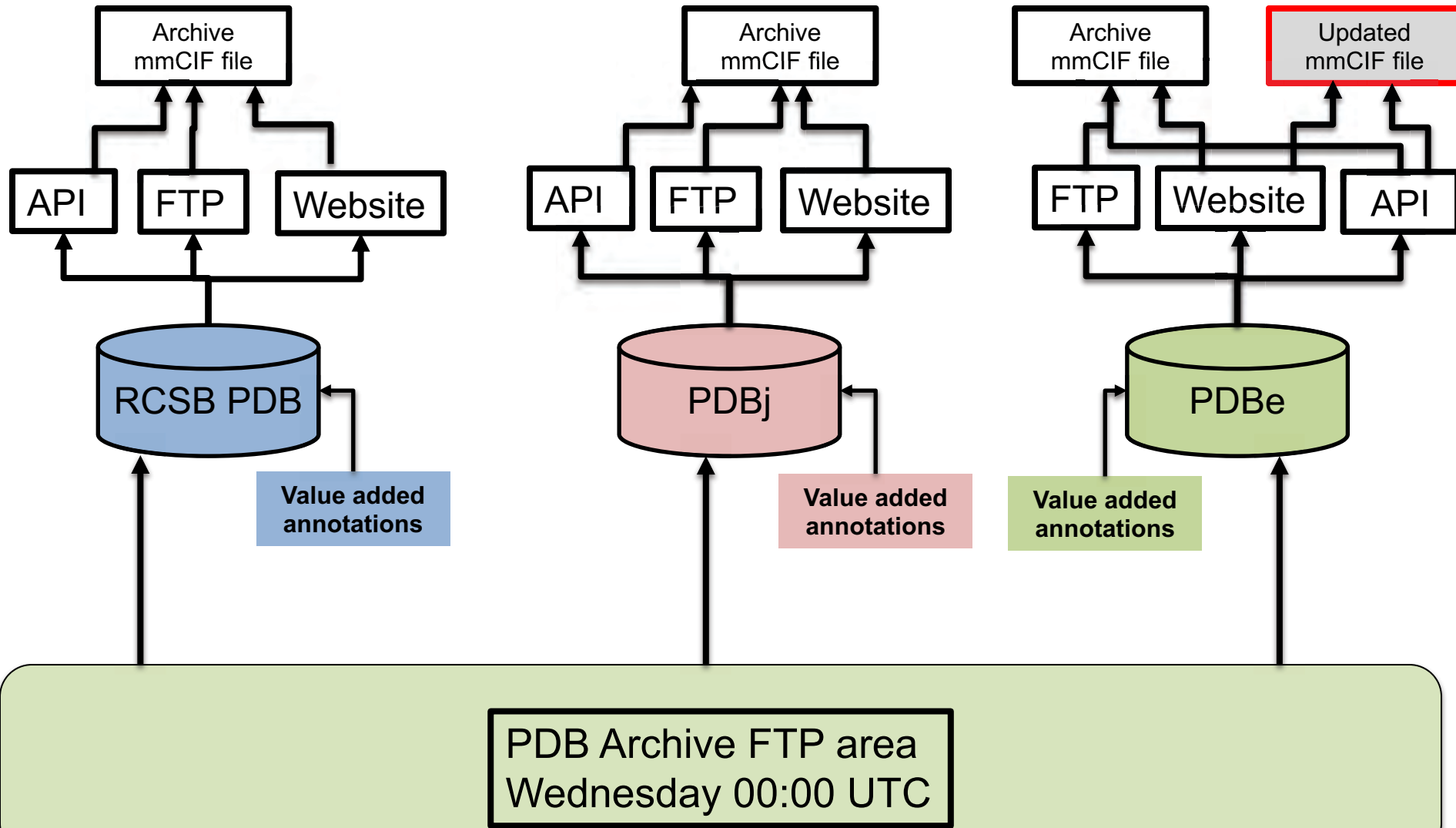
Current Data Distribution by Partners

- Pros
  - Each site is independent to develop processes to enrich "data-out" offerings
  - Breadth of site-specific focus and expertise benefit Users
- Cons
  - Some duplication of effort
  - Users may have trouble finding particular site-specific data
  - Coordination among wwPDB partners on "data-out" activities is not readily apparent to Users/Funders
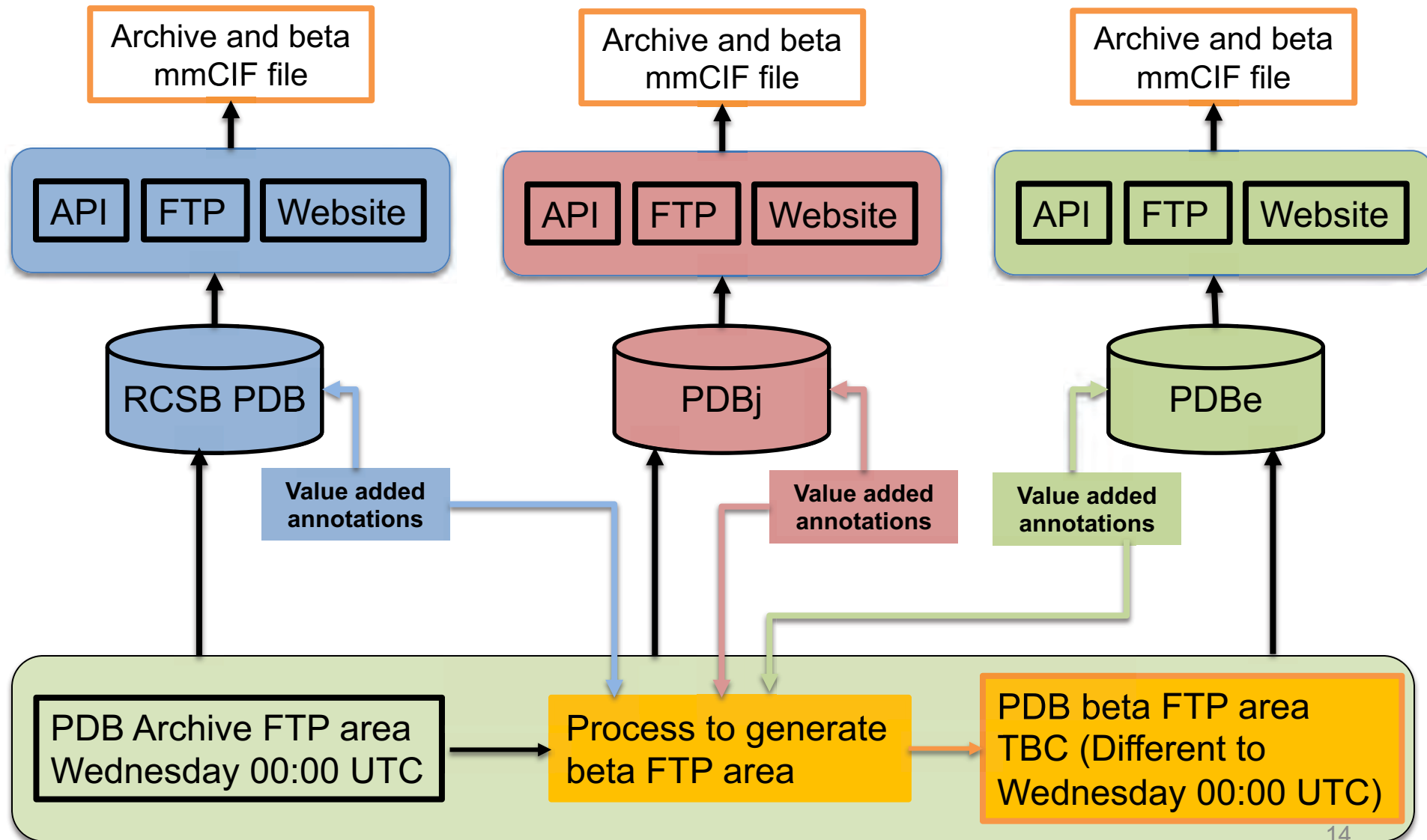
Future Plans I

- Preserve existing independence, site specific focus and expertise to benefit Users (e.g., tools/services)
- Address drawbacks for Users and wwPDB partners in existing arrangements, by –
  - Minimizing duplication of effort
  - Minimizing redundancies in delivering data from external sources (data resources or calculations)
  - Ensuring that Users can benefit from all data enrichment activities whichever wwPDB partner website they access

Future Plans II

- Continuous review, correction, of PDB Core Archive data and coordinated, timely re-release
- Continue archive wide remediation
- Establish a beta version of PDB Core Archive containing:
    - All data items from the PDB Core Archive without modification
    - Additional annotations and associated provenance information described in a common dictionary integrated into the PDBx/mmCIF format files
    - Where annotations cannot be integrated in the PDBx/mmCIF format files, data with provenance information will be presented in separate directories

Future Plans III

Future Plans IV

- Beta version of the PDB Core Archive will be updated on a weekly basis, using an automated process that operates separately from the update schedule and release processes for the PDB Core Archive

- Roadmap/Timeline

  - Publish plan for full release of beta version with supporting update/release processes in 2020

  - Make beta version publicly available in 2021 (PDB50)

  - wwPDB partners to seek additional joint funding to support requirement setting and software development of automated parallel, update and release processes tailored to each of the PDB Core Archive and the beta version

Mol* Core Library

- Parsing, primary focus on (Binary)CIF (backwards compatible with CIF)

- Efficient in-memory representation of molecular data (particularly assemblies)
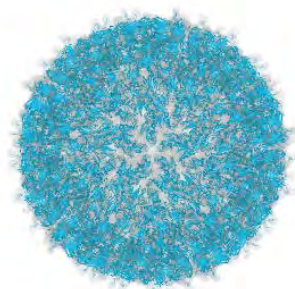
- I/HM support

- Powerful query language MolQL
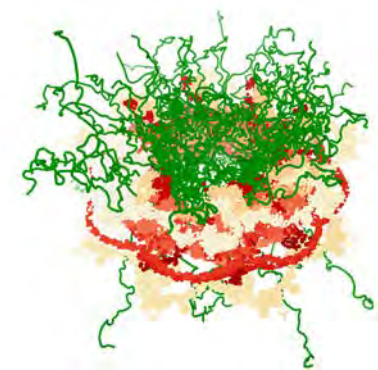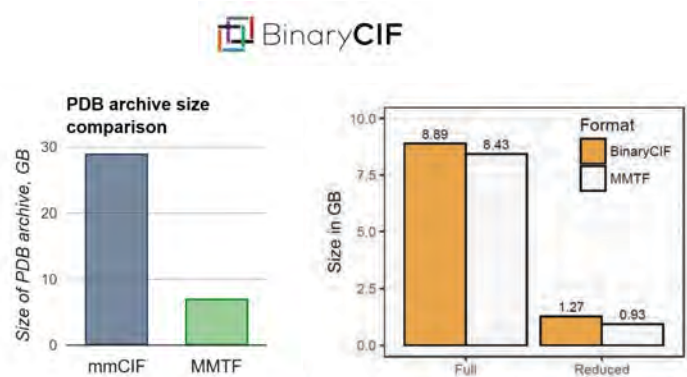
IH/M: Structure of the 552-protein Nuclear Pore Complex (NPC) from yeast
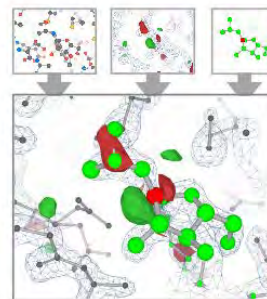
Mol* Data Delivery

- Coordinate/trajectory data
- Experimental data (X-ray, CryoEM)
- Deliver annotations

Cryo-EM structure of Zika Virus

Original data: **1.6 GB**
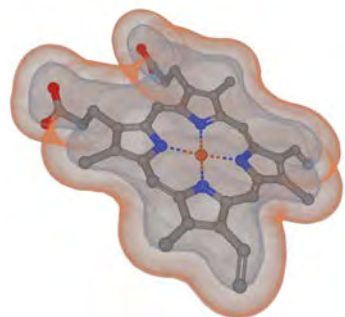
Reduced: **1.03 MB**   1572× less

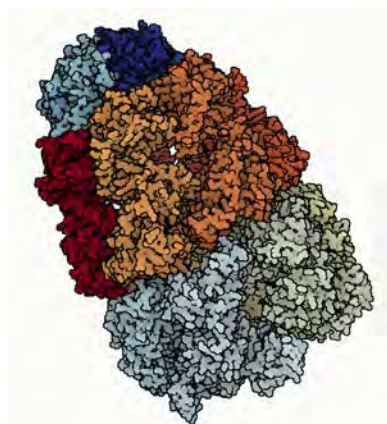**Ligand Binding Site with Validation**

Original data: **20 MB**

Reduced: **40 KB**   508× less

Mol* Viewer

- Modular architecture
- Rich visuals
- State saving and sharing
- Animations

Faustovirus (40M atoms)

60M animated atoms in a single scene
https://molstar.org/viewer/?snapshot-id=capsing-capsids&hide-controls=1

3D-SNFG (carbohydrates)

Volumetric rendering

Cartoon/illustrative shading
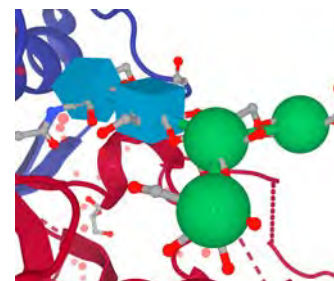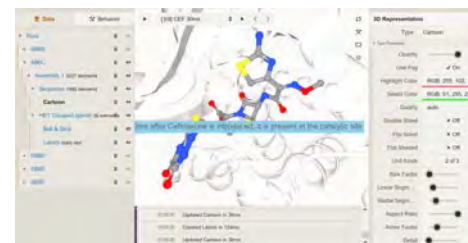
Mol* UI

BMRB Core Archive Plans

John Markley and Jeff Hoch

wwpdb.org

BMRB Core Archive Plans I

- Rework BMRB website: design, graphics, content
- Complete testing of BMRBdep and then close down legacy deposition system
- Continue development of NMR-STAR dictionary, deposition, and data out for solid-state NMR data
- Work with NMRFAM to develop spin system matrices for metabolites that contain corrections for pH-dependent chemical shifts
- Remediate and enlarge the small molecule database to include additional spin system matrices, carbohydrates and important missing metabolites
- Release the beta version of the small-molecule version of BMRBdep

BMRB Core Archive Plans II

- Work with BMRB Curators to expand the NMR-STAR data dictionary and archive content to data types that are under-represented in the BMRB archive: dynamics, integrative/hybrid, intrinsically disordered proteins, membrane proteins, metabolomics, molecular interactions, natural products, RNA, solid-state NMR

- As part of the OneDep Team, continue development of software suite (external packages and our own software) for 3D structure validation vs. NMR data of various kinds

- As part of the PDB-Dev team, develop protocols for validation and annotation of NMR-specific data involved in an integrative/hybrid structure deposition

BMRB Core Archive Plans III

Ref. Appendix 4

- As part of OneDep Team, explore the development of a deposition system for NMR structures that leverages BMRBdep and meets the following criteria:
    - Eliminates/reduces redundant requests for metadata
    - Harvests full range of experimental NMR data pertinent to biomolecular structure
    - Maintains provenance of atom naming and numbering to ensure accurate validation of structural models against deposited NMR data
    - Supports software that validates structural models against the full range of NMR experimental data
    - Supports BMRB as the sole point of contact with Depositors
    - Maximizes interoperability, portability, and maintainability of code

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

EMDB Core Archive Plans

Ardan Patwardhan

EMDB Archive→wwPDB Core Archive

- Draft policies and procedures document under review by wwPDB PIs

- Progress on EM Validation

  - EM map, map/model, and model validation components for validation reports have now been developed based on the "Visual Analysis" pages for entries at EMDB

  - Map and map/model components will be made available in wwPDB Validation Report.

  - New validation reports will be made available as part of OneDep in November 2019.

Map Segmentation Data in EMDB

- Currently there is only one way to represent segmentations – as masks

  - Lack of structured annotation make these almost unusable

  - Tedious for users to convert segmentations into masks for deposition – option is rarely used

- Medium-Term Plan: To add support for map segmentation data in EMDB data model and OneDep

Questions for the wwPDB AC

Stephen K. Burley

wwpdb.org

Questions for the wwPDB AC

1. Does the wwPDB AC concur with the updated Terms of Reference for the wwPDB Advisory Committee (Appendix 5, see Section 3.8)?

2. What is the wwPDB AC guidance regarding participant representation for the 3DEM community (Appendix 6)?

3. What is the wwPDB AC guidance regarding the duration of the comment period for the planned policy change making PDBx/mmCIF atomic coordinate files mandatory for deposition of 3DEM and microED structures (Appendix 7)?

Questions for the wwPDB AC (cont.)

4. What is the wwPDB AC guidance regarding the duration of the comment period for the planned policy change making PDBx/mmCIF atomic coordinate files mandatory for deposition of NMR structures (Appendix 8)?

5. Does the wwPDB AC concur with the BMRB proposal to seek federal funding that will support work with NMR software developers to adopt NMR-STAR (Appendix 9)?

Questions for the wwPDB AC (cont.)

6. Does the wwPDB AC have any questions or concerns regarding the individual RCSB PDB, PDBe/EMDB, PDBj, or BMRB Advisory Committee reports (Appendix 10)?

Group Photo