

2016 wwPDB AC Meeting

John L. Markley, Stephen K. Burley,
Haruki Nakamura, and Sameer Velankar



wwpdb.org

Introduction and Overview of the wwPDB

John L. Markley



wwpdb.org

Welcome

Chair and ICMRBS Representative: R. Andrew Byrd

wwPDB Advisory Committee Members

- BMRB: Valérie Copié and Arthur Edison
- RCSB PDB: Paul Adams and Cynthia Wolberger
- PDBe: David Brown and Sarah Butcher
- PDBj: Tsuyoshi Inoue and Genji Kurisu

Welcome (contd.)

Associate Members

- China: Jianping Ding (sends regrets)
- India: Manju Bansal

IUCr Representative

- Edward Baker

Macromolecular EM Representative

- Wah Chiu

Developments post 2015 AC Meeting

- Addressing collaboration challenges
- PDBe Leadership
 - Sameer Velankar so designated
- Deployment of full OneDep system supporting Crystallography, NMR, and 3DEM
- Progress in developing/integrating the NMR Exchange Format (NEF)

OneDep Development On Track

- October 2015: wwPDB Software Development Team reorganized with Jasmine Young as Global Team Leader
- January 2016: Reformed Team successfully launched OneDep V2.0 supporting X-ray, NMR, and 3DEM
- January 14, 2016: PI and Leadership Team planning of Collaboration Reboot Meeting
- February 7, 2016: PIs issued “Guiding Principles” for future development of the OneDep system
- February 7, 2016: PIs issued a Charge to the OneDep Leadership Team (refined before Reboot Meeting)
- March 2-3, 2016: Collaboration Reboot Meeting

Collaboration Reboot Planning

- Reboot Meeting hosted at Rutgers March 2-3, 2016
- Before meeting, PIs shared written summaries of their experiences and frustrations with collaboration
- Professional Leadership Coach (Suzanne Matteson) engaged for Day One
 - AM: Worked with PIs and AC Chair
 - PM: Worked with OneDep Leadership Team
- Prior to meeting Suzanne surveyed and interviewed PIs, AC Chair, and OneDep Leadership Team

PI Charge to OneDep Leadership

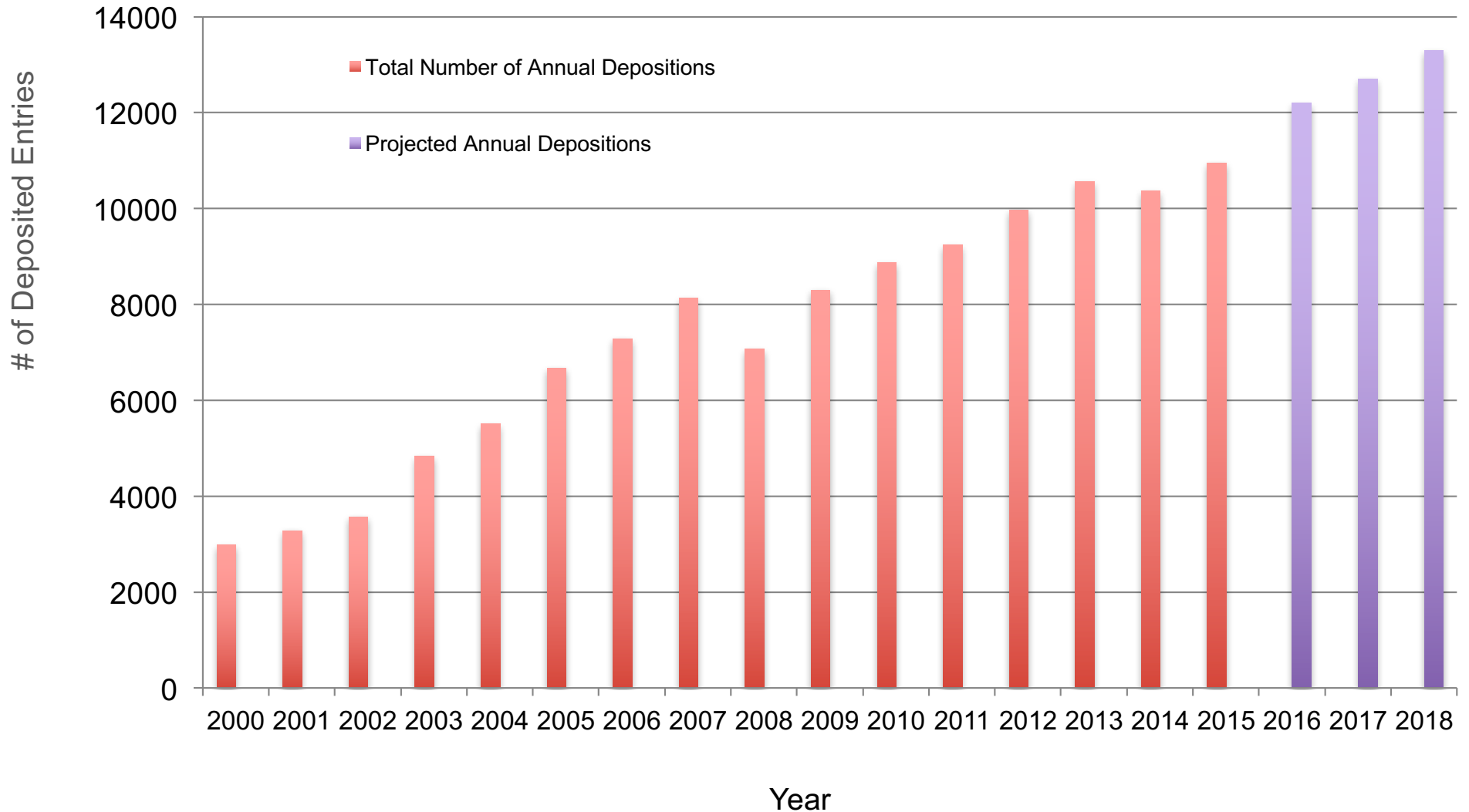
2016 wwPDB AC Meeting Deliverables

1. Capture Depositor ORCID IDs (initially voluntary)
2. Implement validation for X-ray, NMR and 3DEM
3. OneDep server operational at RCSB PDB West
4. OneDep servers operational at PDBe and PDBj
5. Transfer all in-process Asia and European depositions to PDBj and PDBe OneDep servers
6. Implement geographic redirection of OneDep Users
7. Establish a “warm failover” procedure to ensure continued operations when an individual OneDep site(s) go down
8. Develop plans for versioning of the PDB archive

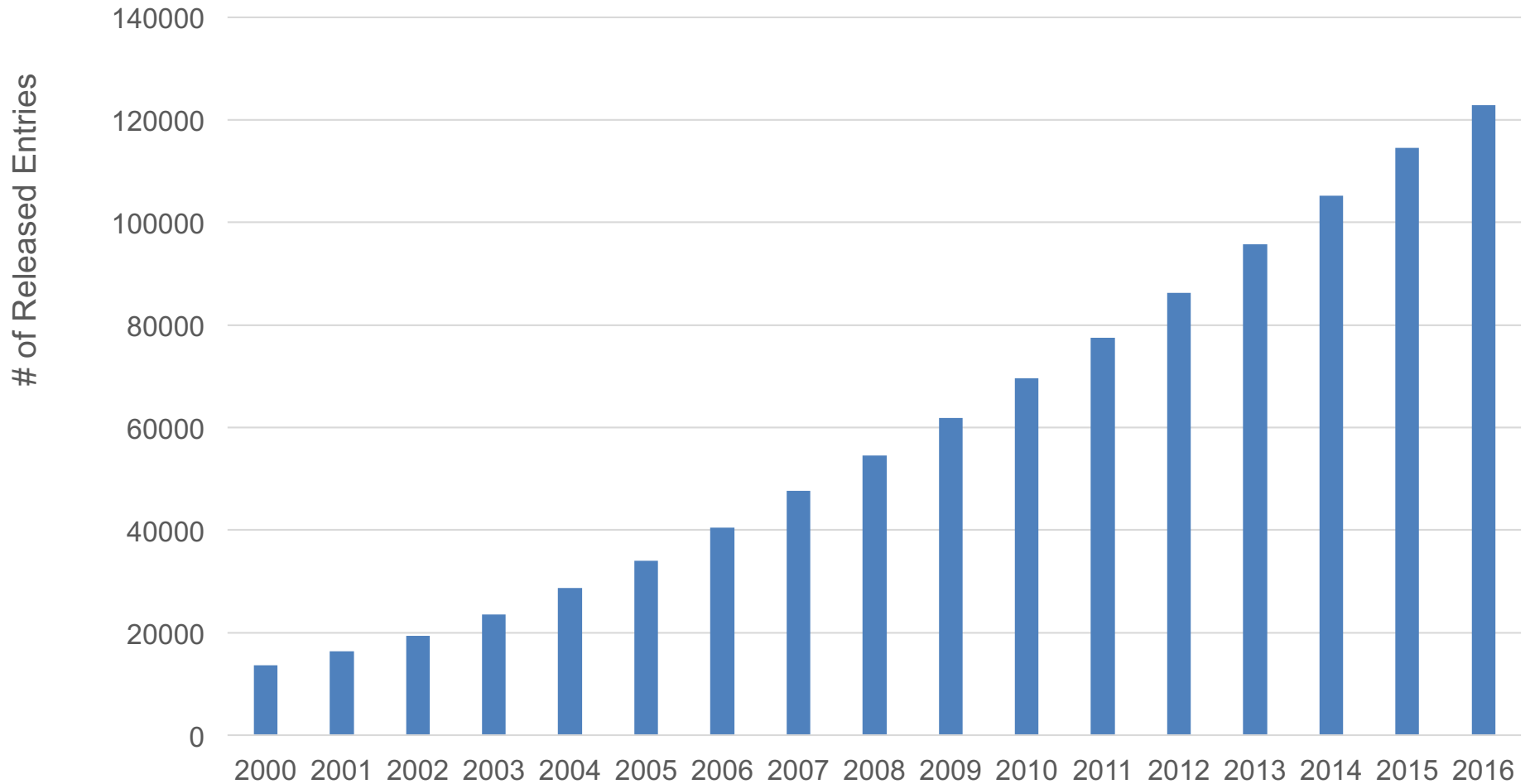
Reboot Meeting Outcome

- Unanimous commitment to build on successes since October 2015 wwPDB Meeting in Osaka
- PIs developed Guiding Principles of Openness, Transparency, 4-Way Communication
- Frustrations/conflict areas brought out into the open for discussion and resolution
- OneDep Leadership Team learned about communications styles and charted a path toward prioritizing and achieving the goals set out in the PI charge

Growing Number of Depositions



Growth of the PDB Archive



More than 1 billion atoms

Download Statistics

Year	Total	Total FTP Archive	Total Website	RCSB PDB FTP Archive	RCSB PDB Website	PDBe FTP Archive	PDBe Website	PDBj FTP Archive	PDBj Website
2009	328,362,536	271,116,934	57,245,602	222,984,760	53,507,785	30,141,339	1,475,116	17,990,835	2,262,701
2010	294,326,976	213,180,966	81,146,010	159,248,214	64,569,658	34,383,219	14,017,349	19,549,533	2,559,003
2011	383,131,048	276,952,286	106,178,762	204,939,406	81,560,098	40,960,368	18,515,245	31,052,512	6,103,419
2012	376,944,070	255,837,735	121,106,335	213,510,347	90,438,501	21,601,103	23,982,801	20,726,285	6,685,033
2013	441,262,210	296,176,290	145,085,920	215,331,908	97,549,580	43,684,850	37,762,496	37,159,532	9,773,844
2014	512,227,251	339,193,721	173,033,530	237,168,615	110,115,316	52,362,370	48,031,414	49,662,736	14,886,800
2015	534,339,871	368,244,766	166,095,105	255,346,630	111,802,897	48,544,330	41,127,219	64,353,806	13,164,989

More than 1.5 million / day



Geographic origins of FTP downloads, 2012-2015

wwPDB Policy Proposals

- PDB Data Release/Hold Policy for Pre-print Archives (Appendix 1)
- Atomic Coordinate Versioning (Appendix 2)
- Mandatory ORCID ID Capture (Appendix 3)

Meeting Reports

- Integrative/Hybrid Methods (I/HM) Task Force:
Federation Subgroup Planning Meeting
Nov 30, 2015
- wwPDB Software Engineering
Collaboration “Reboot” Meeting
Mar 2-4, 2016
- Joint wwPDB NMR VTF/NEF Workshop
Aug 26-27, 2016

Remaining Agenda Items (Lunch at 12:30pm)

- OneDep: Jasmine Young
- Outreach: Haruki Nakamura
- Crystallography: Stephen K. Burley
- 3DEM: Sameer Velankar
- NMR: John L. Markley
- Looking Ahead: John L. Markley
- Questions for the AC: Stephen K. Burley
- Executive Session

OneDep System

Jasmine Young



wwpdb.org

Agenda

- Building the OneDep Team
- Collaboration Reboot Meeting
- Deadlines/Deployment
- System Impact/Performance
- Recalculation of Validation Reports
- Extension of Validation Reports → NMR/3DEM
- ORCID ID Collection
- Rebranding of D&A → OneDep
- OneDep Publication Plan
- 2016/2017 Deliverables
- File Versioning Plan

Building the wwPDB OneDep Team

- Global Project Lead: Young
- Senior Leadership:
 - RCSB PDB: Westbrook, Feng, Lawson
 - PDBe: Gutmanas, Patwardhan
 - PDBj: Kobayashi
 - BMRB: Baskaran
- Meeting Frequency:
 - Operations – Weekly
 - Senior Leadership – Biweekly
 - Senior Leadership with wwPDB PIs – Biweekly
 - Lead Annotators – Weekly

Initial OneDep Milestones

2015

2016

Oct

Nov

Dec

Jan

Feb

Mar

V1.52

V2.0

V2.0

V2.0

V2.0

D&A Submit

Planning for V2.0 Release:

(1) Prioritized and addressed bug tracking tickets

(2) V2.0 Deployment Checklist

(3) Annotation guidelines for NMR and 3DEM

V2.0 in Beta Production
(in parallel with V1.52)

FAQ & Tutorials for Depositors

Planning for Full Production:

(1) Addressed Version compatibility

(2) Existing V1.52 session migration

(3) Data exchange with EMDB and BMRB

OneDep Full Production at RCSB and PDBj

Enabled FTP file upload

Regeneration of X-ray Validation Reports w/ 2015 stats

Validation FAQ

PDBe Server Setup Begun

Planning for Collaboration Reboot Meeting

Generation of NMR and 3DEM Validation Reports

Collaboration Reboot Meeting

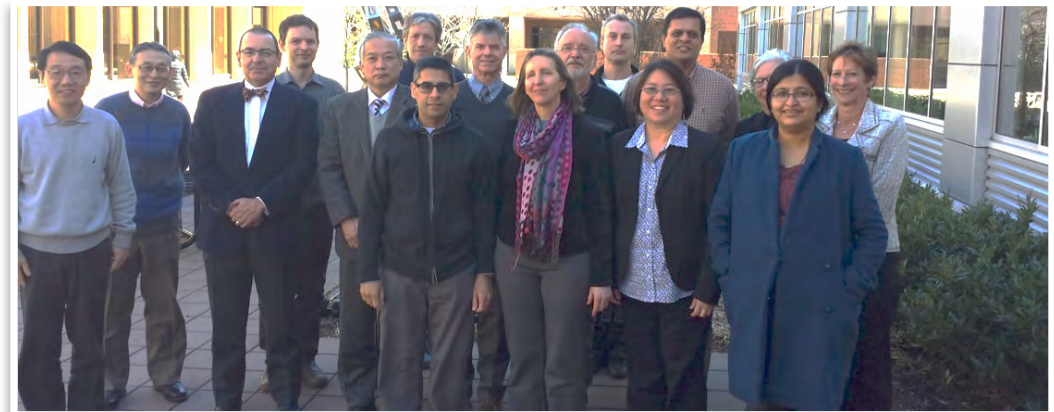
OneDep Project Planning

Update X-ray Validation Reports w/ 2015 stats

Biocuration Pipeline Improvement

Collaboration Reboot March 2016

- Outcomes
 - Communication Plan
 - Recommendations for Coding Standards
 - 2016 Deliverables
 - 2016 Deployment Checklist
 - wwPDB PI Resource Commitments



Defining Path Forward

- Strengthening the Team
- Functional Teams Utilize
 - Commitment/Passion
 - Shared Vision
 - Trust
 - Engagement
 - Transparency



Potential Team

Open minded,
shared leadership



Real Team








High productivity,
feelings of satisfaction
and pride

Planned 2016 OneDep Deliverables

March	Recalculation of Validation Reports across PDB Archive
April	PDBe Server Setup ORCID Collection
May	Suppression of Author Lists/Titles at Deposition Release of NMR/EM Validation Reports
June	Geographic Redirection to Regional Data Centers
July	Failover between Regional Data Centers Phase 1: Session Migration
August	Plan for Depositor of Record File Versioning
September	Implementation of Standalone Validation for all Methods (X-ray, NMR, and 3DEM) Failover between Regional Data Centers Phase 2: Active Session Replication

Delivered Milestones

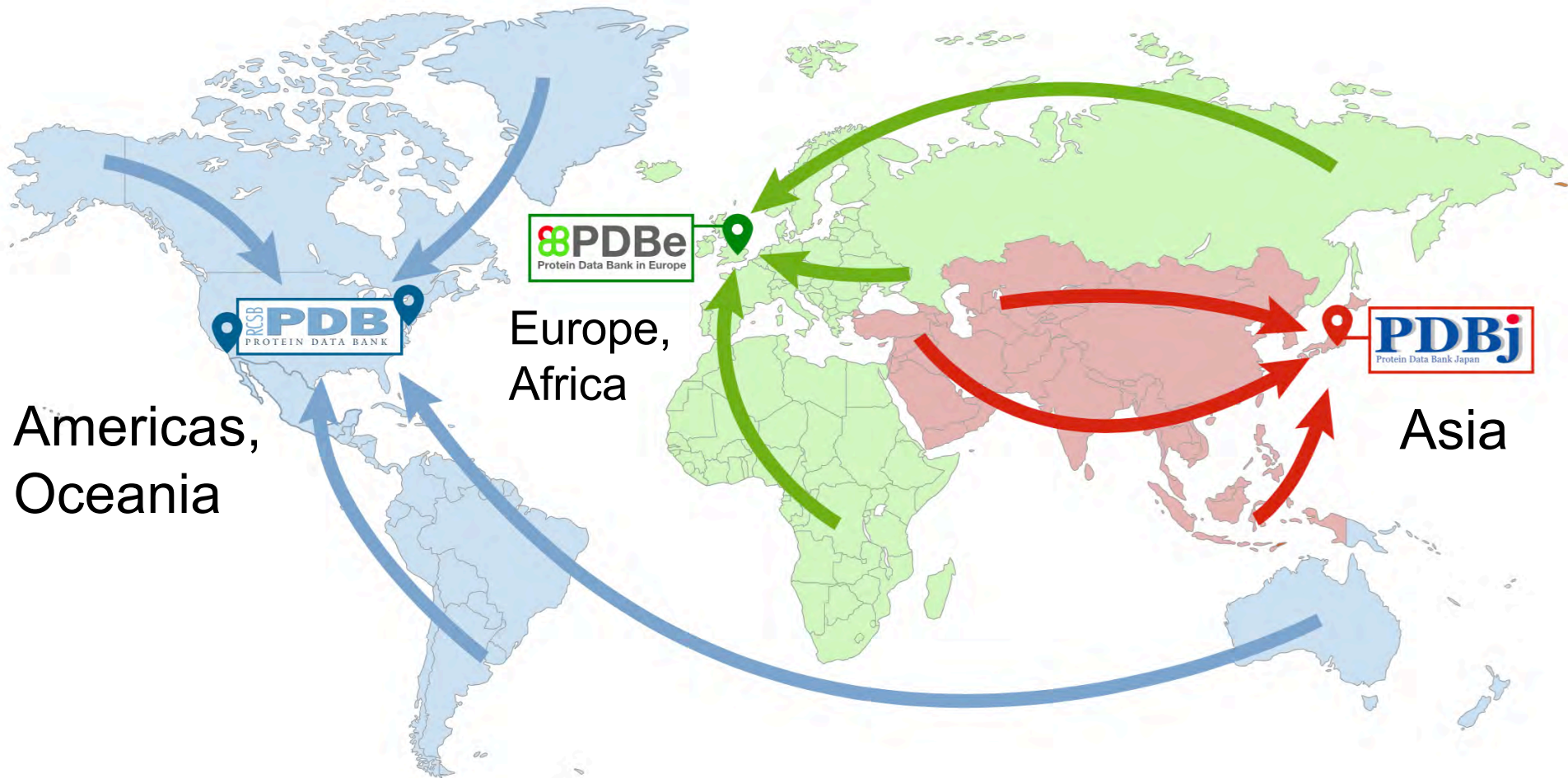
2016

Mar	Apr	May	Jun	Jul	Aug	Sep
						
Collaboration Reboot	PDBe Server in Production	Release of NMR/3DEM Validation Reports	Geographic Distribution	Session Migration	Standalone Validation Server Supporting X-ray, NMR & 3DEM	<i>Map Volume Mandatory for PDB deposition</i>
D&A Project Planning	ORCID Adoption	Suppression of Author List /Title			Active Server Failover	<i>Submit OneDep Paper</i>
Update X-ray Validation Reports w/ 2015 stats	<i>Support for CASP, D3R, and CAPRI Challenges</i>	Legacy Phase Out Begun			<i>3DEM remediation (V4 -> V5)</i>	File Versioning Planning
						Legacy Phase Out Complete

Biocuration Pipeline Improvement

New Workflow Manager

Geographic Direction

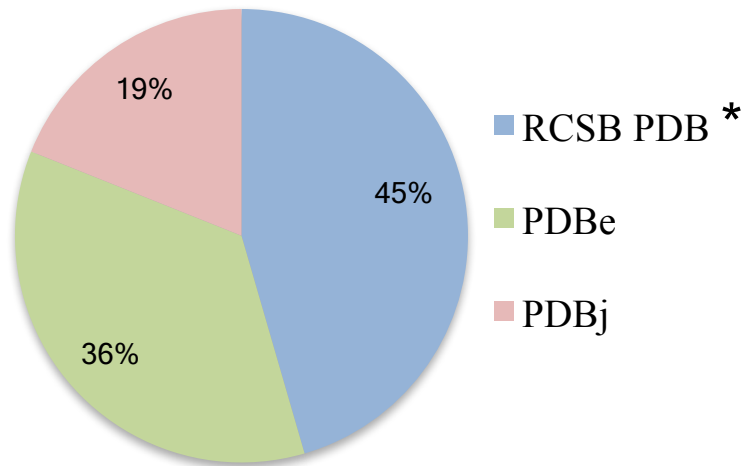


deposit.wwpdb.org

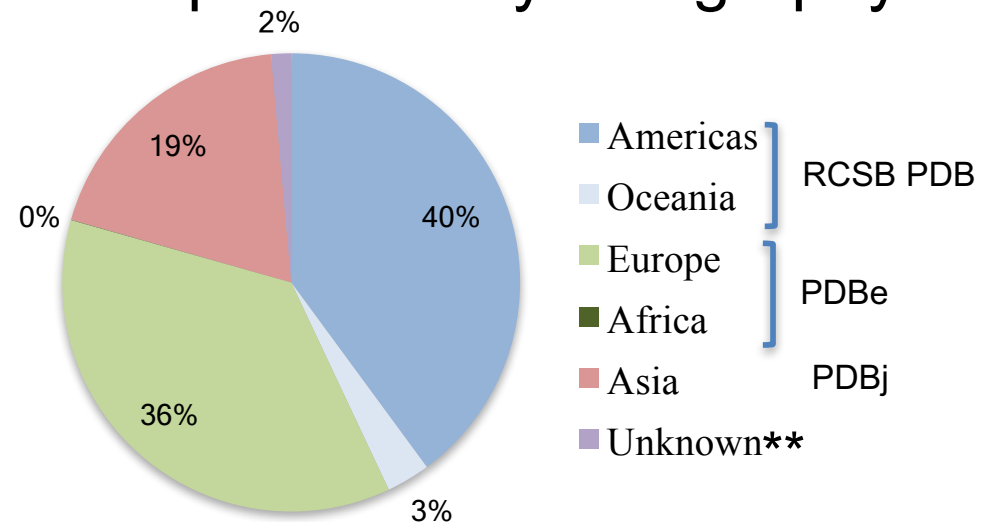
OneDep Impact: 09/2015→08/2016

Since mid-2016 Depositors have been directed to the appropriate Regional Data Center

Processing by Data Center



Depositions by Geography



* including Group depositions at RCSB PDB

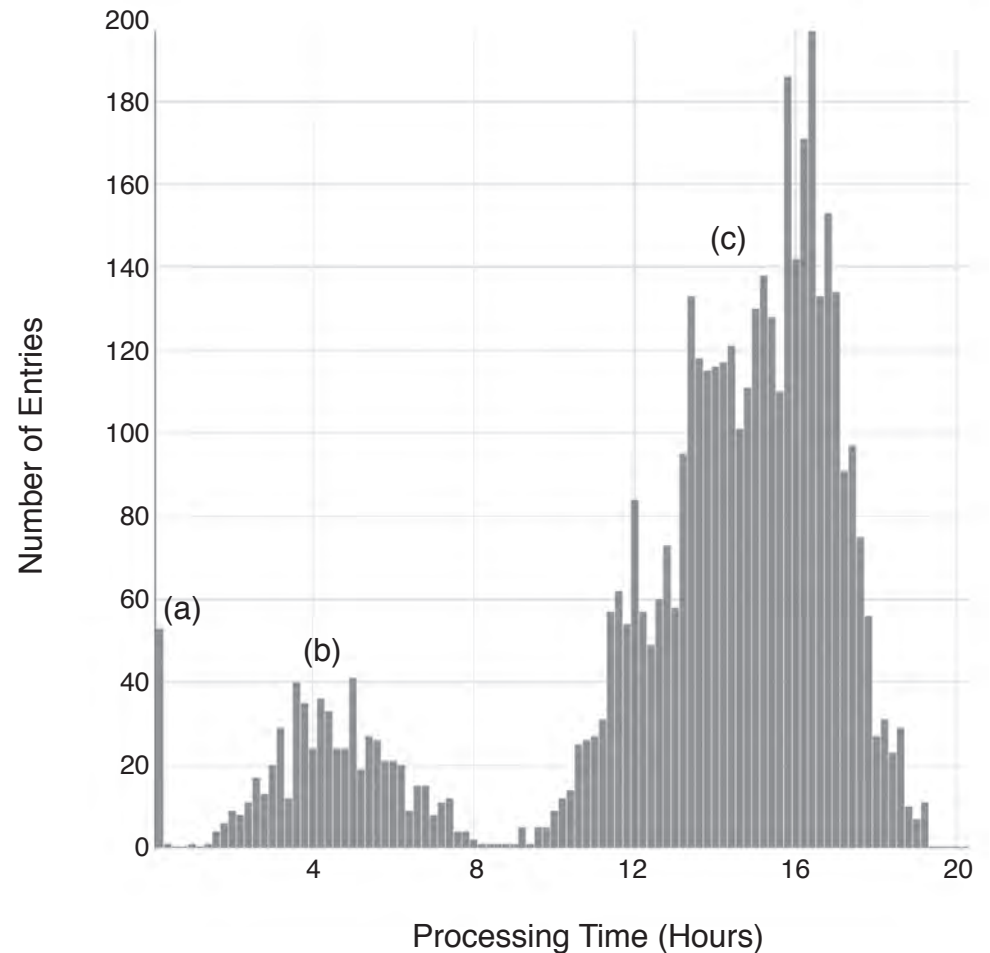
** Commercial depositions at legacy system

OneDep Processing Times

(a) ~1hr: Simple structures without issues

(b) ~4 hrs: More complex structures without issues

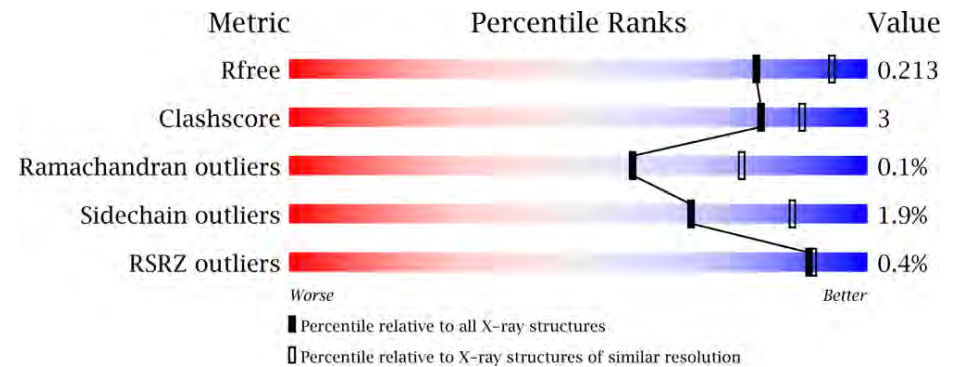
(c) ~15 hrs: Structures with issues, including Depositor response time



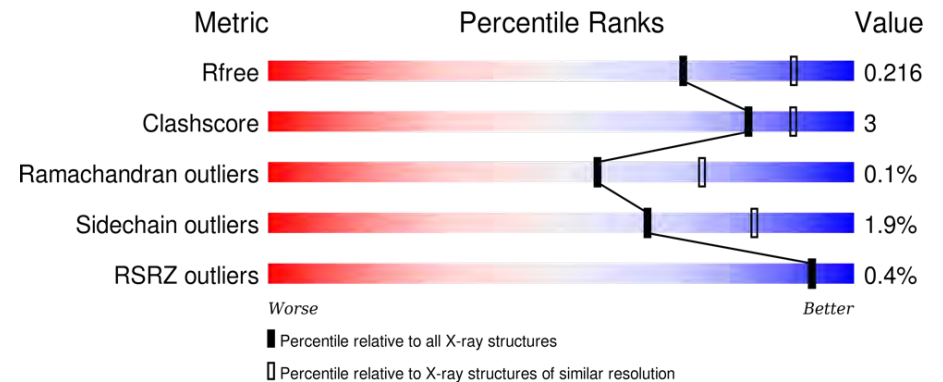
Recalculation of Validation Reports

- Archive Snapshot taken Dec 31, 2015
- Statistics recalculated
- March 2016: X-ray Validation Reports updated
- May 2016: NMR and EM Validation Reports released

PDB ID 4p1c



Statistics from Dec 31, 2013



Statistics from Dec 31, 2015
(~20% increase in Archive)

Validation Report → NMR and 3DEM

1.

Structure Determination

Pre-validate data independently before deposition

Aug 2016

2.

Deposition

Mandatory acknowledgement of report produced during deposition

Jan 2016

3.

Biocuration

wwPDB-recommended report for journal submission

Jan 2016

4.

Public Release

Report available for all released PDB entries

May 2016

Submission of Validation Report during manuscript review process is mandatory (*Nature*, *Acta D & F*, *FEBS*, *J Biol Chem*, *J Immunology*, *eLIFE*, *Angew Chem Int Ed Engl*) or recommended (*Cell*, *Molecular Cell*, *Structure*)

Finally!!

EDITORIAL

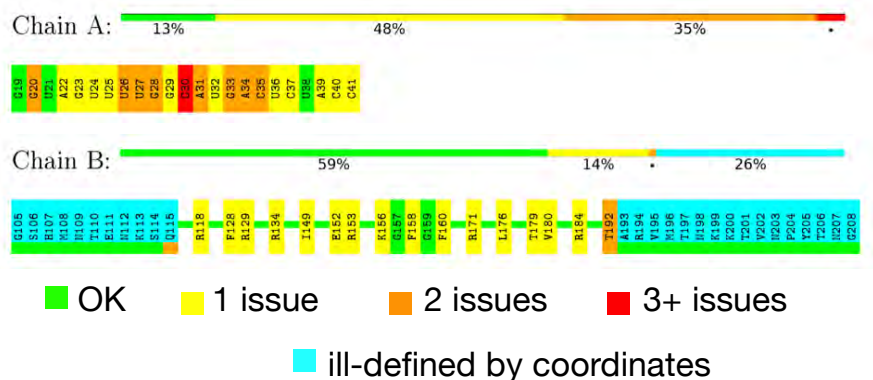
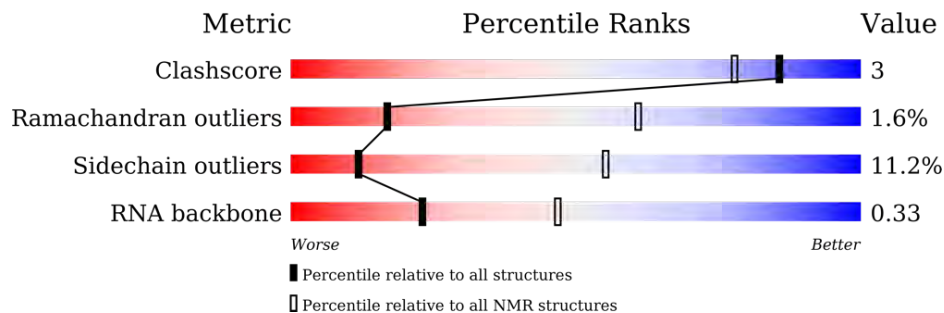
nature
structural &
molecular biology

Where are the data?

Here, we announce two policy changes across Nature journals: data-availability statements in all published papers and official Worldwide Protein Data Bank (wwPDB) validation reports for peer review.

As the research community embraces data sharing, academic journals can do their part to help. Starting this month, all research papers accepted for publication in *Nature* and an initial links to data in published articles is an effective approach to ensuring public data availability and policy compliance (T.H. Vines *et al.*, *FASEB J.* 27, 1304–1308, 2013).

Key Features of NMR Reports



Chemical Shifts:

- Referencing
- Assignment Completeness
- Statistical Outliers
- Random Coil Index

Nucleus	# values	Correction \pm precision, ppm	Suggested action
$^{13}\text{C}_\alpha$	88	-0.48 \pm 0.16	None needed (< 0.5 ppm)
$^{13}\text{C}_\beta$	86	-0.07 \pm 0.16	None needed (< 0.5 ppm)
$^{13}\text{C}'$	80	2.86 \pm 0.11	Should be applied
^{15}N	80	-0.01 \pm 0.34	None needed (< 0.5 ppm)

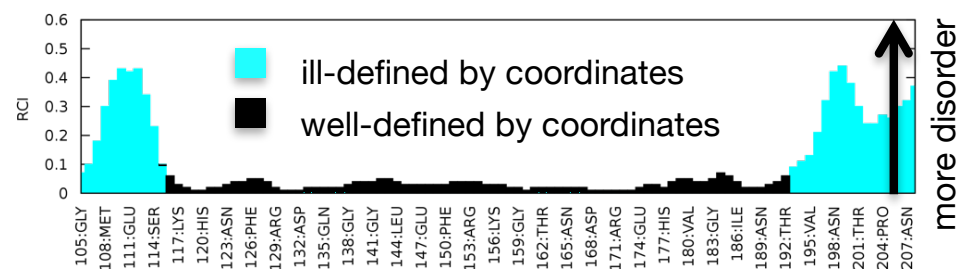
	Total	^1H	^{13}C	^{15}N
Backbone	499/512 (97%)	198/204 (97%)	204/208 (98%)	97/100 (97%)
Sidechain	563/713 (79%)	344/420 (82%)	201/245 (82%)	18/48 (38%)
Aromatic	70/101 (69%)	47/56 (84%)	23/42 (55%)	0/3 (0%)
Overall	1383/1753 (79%)	723/923 (78%)	545/650 (84%)	115/180 (64%)

Mol	Chain	Res	Type	Atom	Shift, ppm	Expected range, ppm	Z-score
1	A	56	LYS	NZ	109.70	49.86 – 18.16	23.9
1	A	42	ARG	NE	111.50	92.63 – 76.73	16.9
1	A	19	TRP	HD1	5.21	8.95 – 5.35	-5.4

Ensemble Analyses:

- Well-Defined vs. Ill-Defined
- Polymer Segments

Well-defined (core) protein residues			
Well-defined core	Residue range (total)	Backbone RMSD (Å)	Medoid model
1	A:5-A:84 (80)	0.48	22
2	A:96-A:158 (63)	0.52	17



Collection of ORCID IDs

- Successfully Implemented Apr 11, 2016
- Metrics (Apr 11 – Aug 31, 2016):
 - ~8% of Depositions have ORCID ID (374/4713)
 - 170 unique ORCID IDs (92 identified as PIs)
- Plans to Increase ORCID Adoption
 - Expand to all entry authors to provide ORCID (2017)
 - Distribute collected ORCID IDs at ftp archive (2017)
 - Mandatory going forward (2018)

Rebranding of D&A → OneDep

- Rebranding Process
 - Project Team nominated names/logos
 - Project Leadership recommended
 - wwPDB PIs made final selection
- OneDep logo



OneDep Publication Plan

Component	Focus Topics	Submission Timeline
<i>Full System</i>	<i>High-level overview of Deposition, Biocuration and Validation – Primary Reference</i>	Sep 2016
Validation	VTFs, supported methods, content, benefits to the depositors, annotators, and users, what have been implemented, limitations, and future improvements	Nov-Dec 2016
Full Biocuration Pipeline	Work Flow Manager, Ligand and Sequence annotation, checks, and communication	Jan-Feb 2017
Full Deposition Pipeline	More complete data, more checks, allow multiple file replacement, more efficient and better data quality	Mar-Apr 2017
3DEM	Changes/enhancements made, dictionary, extended data items, richer content for EMDB and PDB	Mar-Apr 2017
NMR Validation	VTF, CS, restraints validation, and NEF	TBD (dependency: VTF and NEF WG)
Enhanced Ligand Validation	Following implementation of Ligand Validation Workshop recommendations	TBD

Planned 2016/2017 Deliverables

Core Infrastructure Support:

Upgrade Security; Enable Use of External Computing Resources; Encrypt Traffic; Implement Depositor of Record Versioning; Management of User Credentials

New Content:

Migration of Legacy Entries→OneDep system; Begin Carbohydrate Remediation; Inclusion of NMR-SAXS Hybrid Method; Capture Experimental Assembly Data

Enhance Depositor Experience:

Data File Re-upload at Deposition; Conditional Controlled Vocabulary; Support Ligand Validation; Support NMR Exchange Format Files; Support Depositor Assembly with Experimental Evidence; Implement EM MAP Validation

Enhance Validation: Implement Ligand Validation Workshop Recommendations; Support Validation vs. NMR Restraints and CS in NMR Exchange Format

Enhance Biocurator Experience: Improve WorkFlow for Large Structures; Increase Reprocessing Automation; Improve CIF Editor Usability

File Versioning: Objectives

Current Issues:

- Loss of connection between PDB ID and Publication under current wwPDB Obsolete/Supersede Policy
- Current wwPDB Policy represents a non-trivial barrier to revisions by the Depositor of Record

Objectives:

- Introduce new procedure to manage revision of atomic coordinates by the Depositor of Record
- Establish a robust extensible framework for versioning of all archival data

File Versioning: Planning Process

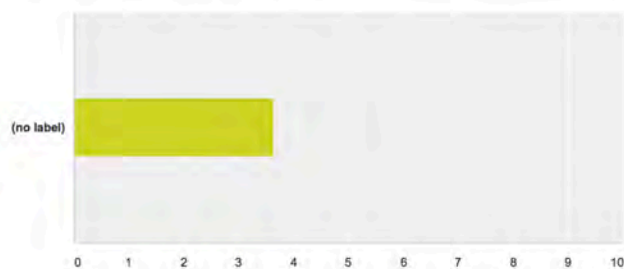
- User feedback solicited
- Enable revisions to entries updated by the Depositor of Record (e.g., Version 1-0 → 1-1; 1-0 → 2-0)
 - wwPDB will NOT assign a new PDB ID going forward (for Depositor of Record revision only)
- Introduce new PDB ID code format
 - Allow more informative and transparent delivery of revised data files
 - With PDB prefix and extension of 4 characters (e.g., from “1ABC” to “PDB_00001ABC”)
- Example: PDB_00001ABC_XYZ_V2-2.cif.gz

File Versioning: User Feedback

- Survey results from Depositors and Power users
 - 101 responses received
(42% access PDB data *via* ftp/rsync directly)
 - Overall positive feedback received (63%)
 - Changes in atomic coordinates, polymer sequences, or ligand chemistry are considered major revisions

Q4 wwPDB plans to assign version numbers to accession codes that will indicate how many times the atomic coordinate file has been revised. How valuable would this feature be for your research?

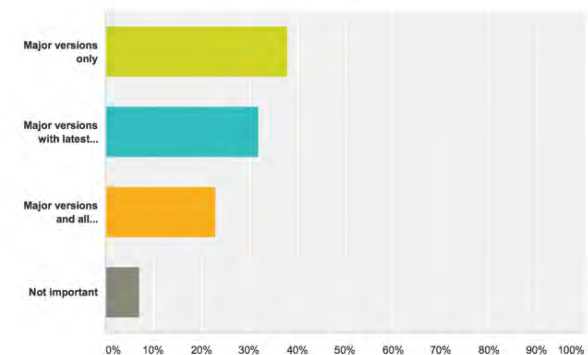
Answered: 98 Skipped: 4



	Not important	(no label)	(no label)	(no label)	Very important	Total	Weighted Average
(no label)	13.27%	7.14%	15.31%	31.63%	32.65%	98	3.63
	13	7	15	31	32		

Q6 What level of detail in versioning is important to you?

Answered: 100 Skipped: 2



Answer Choices	Responses
Major versions only	38.00% 38
Major versions with latest minor corrections	32.00% 32
Major versions and all versions including minor corrections	23.00% 23
Not important	7.00% 7
Total	100

File Versioning: Implementation

- Create new versioned ftp tree containing the latest minor revision to each major version
- Continue current ftp tree with current file naming convention
 - Files in this branch will serve latest version of each data file
- Communication plan
 - Public announcement of the plan (6 months ahead)
 - Public announcement of the implementation date (60 days in advance)
 - Public announcement of the roll out (on the roll out date)

OneDep Team

- Global Project Lead: Jasmine Y. Young
- RCSB PDB: Li Chen, Luigi Dicostanzo, Dimitris Dimitropoulos*, Zukang Feng, Sutapa Ghosh, Vladimir Guranovic, Brian Hudson, Cathy Lawson, Yuhe Liang, Ezra Peisach, Irina Persikova, Martha Quesada*, Raul Sala, Monica Sekharan, Raship Shah*, Chenghua Shao, Lihua Tan, John Westbrook, Huanwang Yang, Marina Zhuravleva, Helen M. Berman, Stephen K. Burley
- PDBe: David Armstrong, John M. Berrisford, Matthew J. Conroy, Dimitris Dimitropoulos*, Glen van Ginkel*, Swanand Gore*, Aleksandras Gutmanas, Pieter M.S. Hendrickx*, Lora Mak, Saqib Mir*, Abhik Mukhopadhyay, Thomas J. Oldfield*, Ardan Patwardhan, Luana Rinaldi*, Eduardo Sanchez-Garcia, Sanchayita Sen*, Oliver S. Smart, Ganesh J. Swaminathan*, Kim Henrick*, Gerard J. Kleywegt, Sameer Velankar
- PDBj: Minyu Chen, Reiko Igarashi, Yasuyo Ikegawa, Yumiko Kengaku, Junko Sato, Hirofumi Suzuki, Haruki Nakamura
- BMRB: Kumaran Baskaran, Dimitri Maziuk, Eldon L. Ulrich*, Hongyang Yao, John L. Markley
- BMRB at PDBj: Takeshi Iwata, Naohiro Kobayashi

Outreach

Haruki Nakamura



wwpdb.org

wwPDB Outreach

wwPDB Symposium: Integrative Structural Biology with Hybrid Methods

October 3, 2015 at Osaka University

9:00 Opening Remark **Haruki Nakamura** (PDBj, Osaka Univ)

9:10 **Helen M. Berman** (wwPDB, Rutgers Univ)

9:40 **Andrej Sali** (UCSF)

10:10 **Wah Chiu** (Baylor College of Medicine)

11:00 **Keiichi Namba** (Osaka Univ)

11:30 **Helen Saibil** (Birkbeck College)

12:00 **Kenji Iwasaki** (Osaka Univ)

14:00 **Angela Gronenborn** (Univ Pittsburgh)

14:30 **Florence Tama** (RIKEN)

15:00 **Takeshi Kawabata** (Osaka Univ)

15:50 **Mitsunori Ikeguchi** (Yokohama City Univ)

16:20 **Paul Adams** (Lawrence Berkeley Laboratory)

16:50 **R. Andrew Byrd** (NCI at Frederick)

17:20 Closing Remark **Stephen K Burley** (RCSB-PDB, Rutgers Univ)

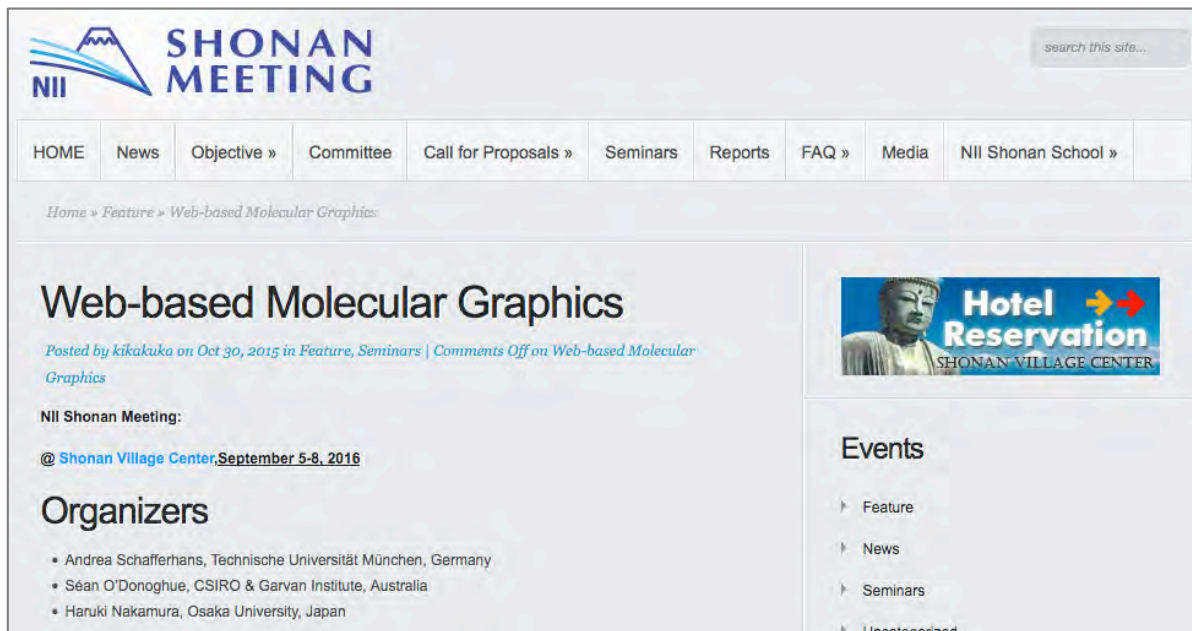


Helen M. Berman chaired by John L. Markley



Presentation by Andrej Sali

wwPDB Outreach



The screenshot shows the NII Shonan Meeting website. At the top left is the NII Shonan Meeting logo. A search bar is located at the top right. Below the logo is a navigation menu with links: HOME, News, Objective », Committee, Call for Proposals », Seminars, Reports, FAQ », Media, and NII Shonan School ». Below the menu is a breadcrumb trail: Home » Feature » Web-based Molecular Graphics. The main content area features a post titled 'Web-based Molecular Graphics' by kikakuka on Oct 30, 2015. To the right of the post is a banner for 'Hotel Reservation SHONAN VILLAGE CENTER' featuring a Buddha statue. Below the banner is an 'Events' section with a list of categories: Feature, News, Seminars, and Uncategorized.

NII Shonan meeting on Web-based Molecular Graphics

Sep 5-8 2016 with 28 attendees



Contributions from all wwPDB partners

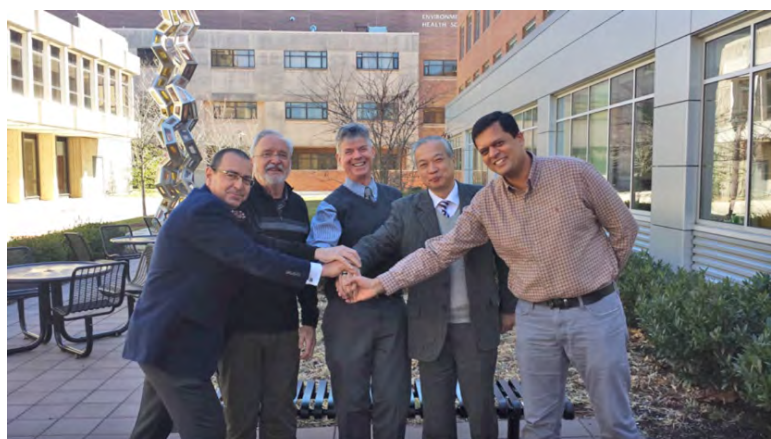


Peter Rose, Alex Rose
(UC San Diego, RCSB-PDB)
Sameer Velankar (PDBe)
Jon Wedell (BMRB)
Haruki Nakamura, Hirofumi Suzuki, Gert-jan Bekker
(PDBj)

wwPDB Leadership in Data Delivery

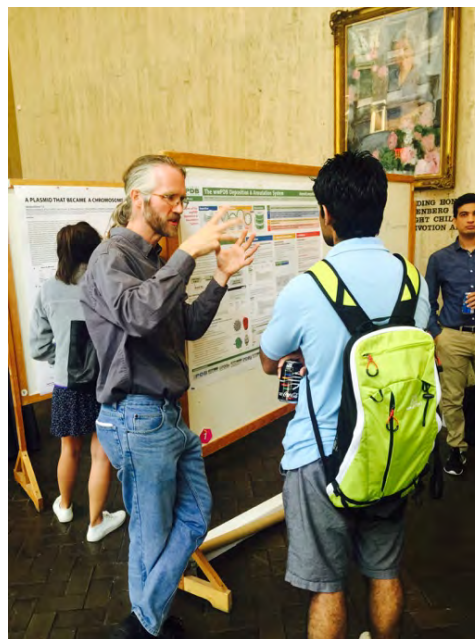
- Rapid Access to PDB Data
 - *MMTF* (Macromolecular Transmission Format) developed by RCSB PDB
 - Dynamic selection and compression approach – Atomic coordinate server developed by PDBe
- Sharing Core Technologies for Viewing Data
 - Molecular Graphics: *NGL Viewer* (RCSB PDB), *Molmil* (PDBj), *LiteMol* (PDBe)
 - Web-components for data presentation – *PDB component library* (PDBe) as a basic model
 - Database integration: *SIFTS* (EBI), *wwPDB/RDF* (PDBj)

wwPDB Outreach



D&A Summit, March, Rutgers

pdb.org now redirects to wwpdb.org

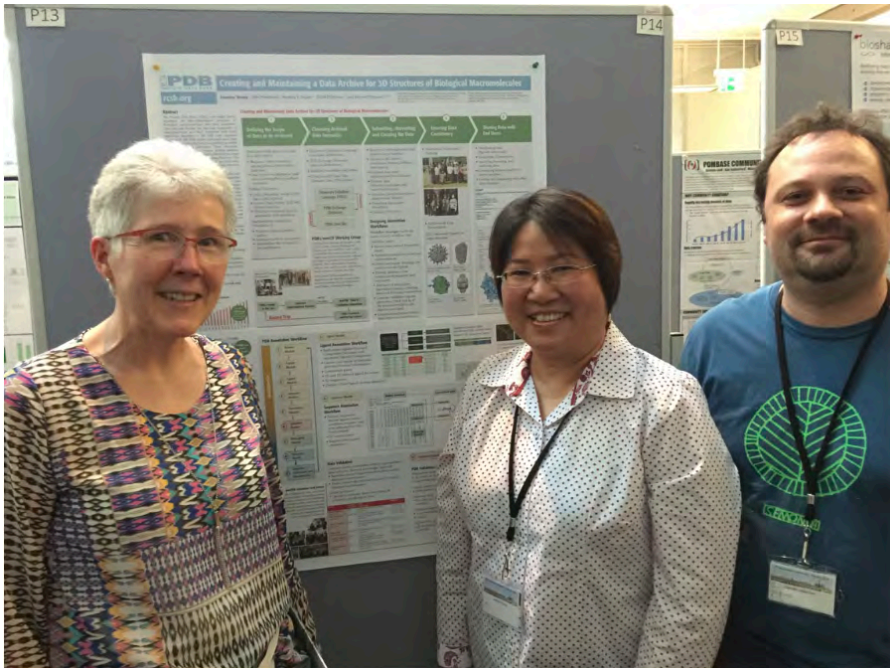


OneDep poster, NY Structural Biology Group Meeting, Aug 2016



Versioning Survey, ACA, Jul 2016

Outreach: RCSB PDB



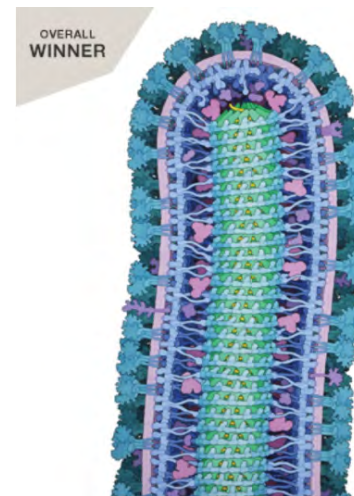
Creating and Maintaining a Data Archive at the PDB, Biocuration Meeting, Apr 2016



Science Olympiad, Jan 2016

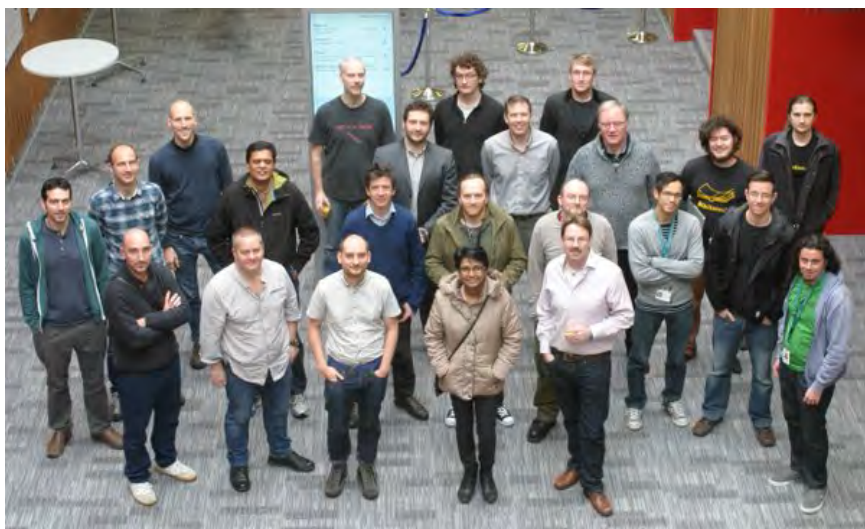


John Westbrook named 2017 Biocuration Career Award winner (for sustained contributions to the field of biocuration)

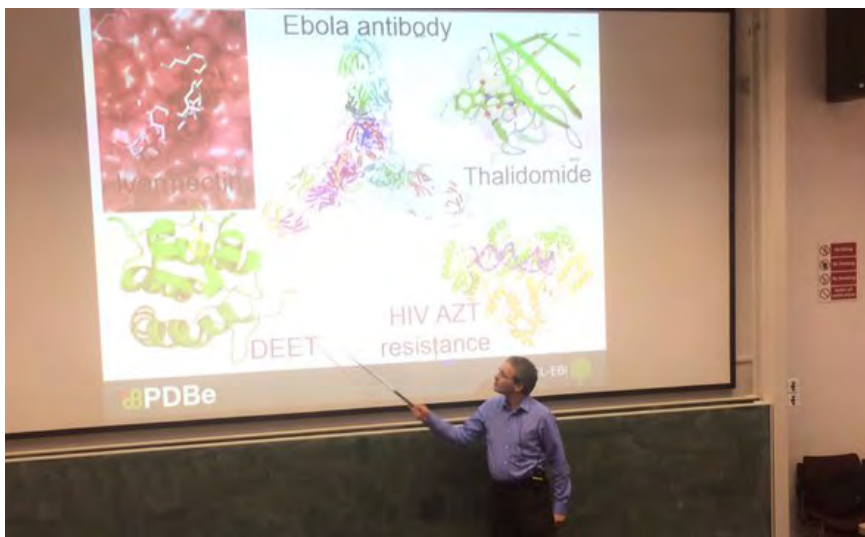


Wellcome Trust Image Awards 45

Outreach: PDBe



Open source search for bioinformatics workshop Feb 3-4, 2016



Protein Data Bank for Undergraduate Chemists
Nov 17, 2015 Lecture and practical session for Bioorganic
Chemistry at the University of Warwick, UK



British
Crystallographic
Association Spring
Meeting, Nottingham
UK, 4-7 Apr 7, 2016
Poster prize
presentation



Seminar from
Stephen Curry
Apr 19, 2016



30th European
Crystallography
Meeting, ECM-30,
Aug 28 –
Sep 1, 2016

Outreach: PDBj



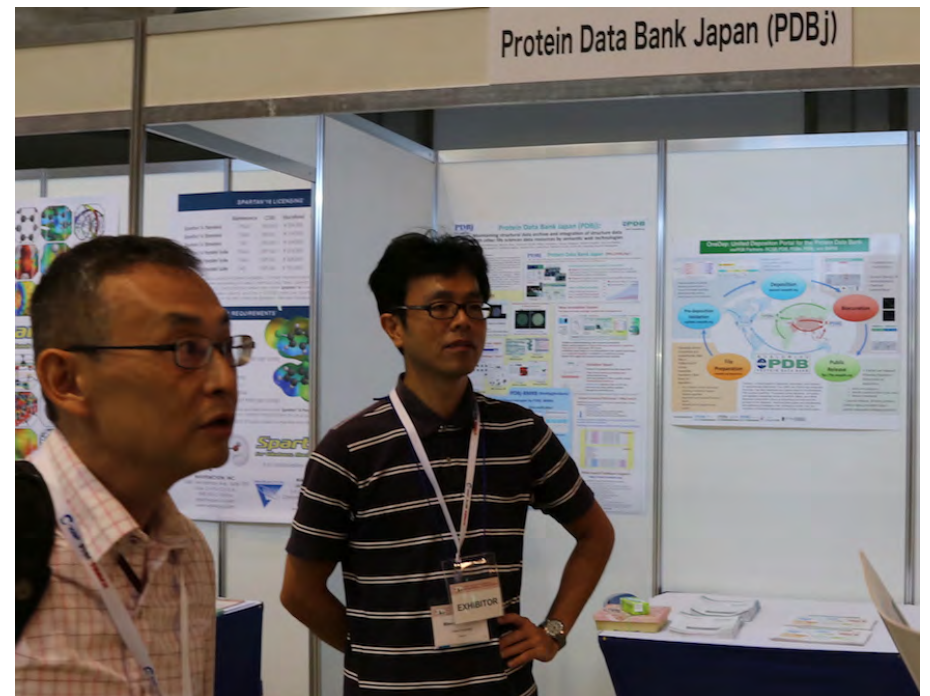
Luncheon Seminar at The Crystallographic Society of Japan (Oct 18, 2015, Osaka)



Science Agora 2015 - The Scientific Events Supported by JST. (November 13-15, 2015, Tokyo)



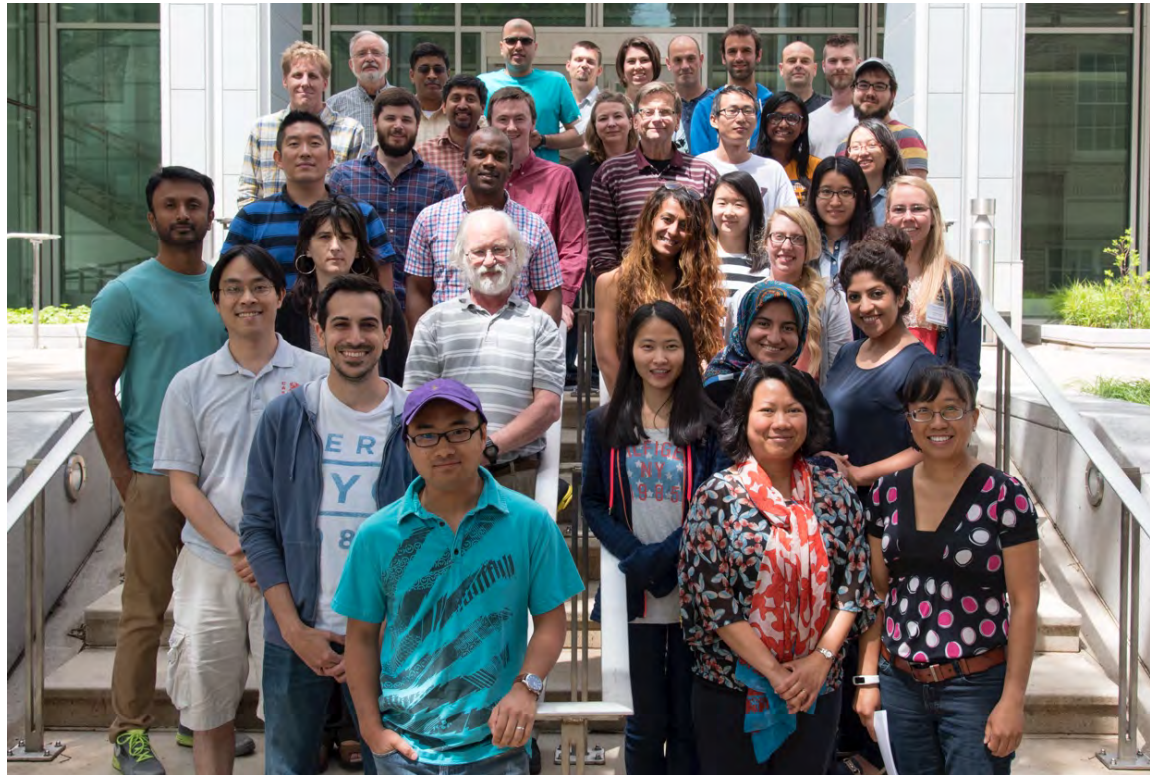
“All-in-One” joint workshop for the life science data bases (July 23, 2016, Osaka, Japan)



ICMRBS 2016 (August 21-26, 2016, Kyoto)

Outreach: BMRB

02/27-03/02/2016 Poster at the annual meeting of the Biophysical Society, Los Angeles, CA, USA



06/05-09/2016 Presentation at the NMRFAM Workshop on NMR Structure Determination, Madison, WI, USA

06/10/2016 Co-sponsor of the Workshop on “NMR-Based Metabolomics”, Morgridge Institute, Madison, WI, USA

Outreach: Book Publication Plan

- Title: "Integrative Structural Biology with Hybrid Methods"
- Publisher: Springer Japan
- Series: Advances in Experimental Medicine and Biology
- Publication Date: 2017

wwPDB Publications

Available online at www.sciencedirect.com

 ScienceDirect





The archiving and dissemination of biological structure data

Helen M Berman¹, Stephen K Burley^{1,2}, Gerard J Kleywegt³, John L Markley⁴, Haruki Nakamura⁵ and Sameer Velankar³



The global Protein Data Bank (PDB) was the first open-access digital archive in biology. The history and evolution of the PDB are described, together with the ways in which molecular structural biology data and information are collected, curated, validated, archived, and disseminated by the members of the Worldwide Protein Data Bank organization (wwPDB; <http://wwpdb.org>). Particular emphasis is placed on the role of community in establishing the standards and policies by which the PDB archive is managed day-to-day.

Addresses

¹Research Collaboratory for Structural Bioinformatics Protein Data Bank, Department of Chemistry and Chemical Biology, Center for Integrative Proteomics Research, Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, 174 Frelinghuysen Road, Piscataway, NJ 08854, USA

²Research Collaboratory for Structural Bioinformatics Protein Data Bank, Skaggs School of Pharmacy and Pharmaceutical Sciences and San Diego Supercomputer Center, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

³Protein Data Bank in Europe, European Molecular Biology Laboratory – European Bioinformatics Institute, Wellcome Genome Campus, Cambridge CB10 1SD, UK

⁴Biological Magnetic Resonance Bank, Department of Biochemistry, University of Wisconsin-Madison, Madison, WI 53706, USA

⁵Protein Data Bank Japan, Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka, 565-0871, Japan

Corresponding author: Berman, Helen M (berman@csb.rutgers.edu)

Current Opinion in Structural Biology 2016, 40:17–22

This review comes from a themed issue on **Biophysical and molecular biological methods**

Edited by **Petra Fromme** and **Andrej Sali**

<http://dx.doi.org/10.1016/j.sbi.2016.06.018>

0959-4401/© 2016 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

Historical background

Structural biology is a relatively young science that can trace its roots to the first X-ray diffraction studies of pepsin in 1935 by Dorothy Crowfoot (Hodgkin), who at the time was a student of J.D. Bernal [1]. Twenty years later, Kendrew determined the structure of myoglobin [2,3]; shortly thereafter, Perutz determined the

structure of hemoglobin [4,5]. Both won Nobel prizes for their achievements. Not long after these structures were published, the crystallographic community began discussions as to how to best archive these data and make them available. During this period, there were numerous grassroots meetings, one of which resulted in a petition, and many exchanges of handwritten documents. In 1971, the Cold Spring Harbor Laboratory hosted a symposium on protein crystallography, during which leaders in the field presented their seminal work [6]. Walter Hamilton, an attendee, offered to provide the first home for what is now known as the Protein Data Bank (PDB) [7]. The PDB was launched at Brookhaven National Laboratory, on the basis of the Protein Structure Library created by Edgar Meyer [8]. The initial PDB archive contained fewer than ten structures, all of which were determined by X-ray crystallography. In the 1980s, structures determined using NMR methods began to be deposited, and in 1990 the first structure determined by electron microscopy was deposited. In 1982 the PDB reached 100 entries, in 1993 1000 entries, in 1999 10,000, and in 2014 100,000 entries. At the time of writing, the PDB archive contains over 117,000 structures of proteins, nucleic acids, and their complexes with one another and with small molecule ligands.

The PDB as a community data resource

From its inception, the PDB has been a community effort that has evolved with changes in scientific culture. For example, when the PDB was first created, data submission was voluntary. However, in the 1980s, members of the community became outspoken about the need to enforce mandatory data deposition. Various committees were set up to define what data should be required and when to disseminate the data. These guidelines were published in 1989, and over time, adopted by virtually all of the scientific journals that now require PDB deposition(s) as a prerequisite for publication of structural studies [9]. In 2008, further shifts in community sentiment led to mandatory deposition of experimental data together with atomic coordinates. In the current decade, the importance of reproducibility has been highlighted. The PDB convened method-specific Validation Task Forces and Workshops [10^a, 11^b, 12^c, 13^d] to define what data should be collected and how best to validate the structural models, the experimental data, and the fit of the models to the data. Now every structure in the PDB comes with a publicly available validation report, and

www.sciencedirect.com Current Opinion in Structural Biology 2016, 40:17–22

 **Structure**

Meeting Report

Outcome of the First wwPDB/CCDC/D3R Ligand Validation Workshop

Paul D. Adams,¹ Kathleen Aertgeerts,² Cary Bauer,³ Jeffrey A. Bell,⁴ Helen M. Berman,^{5,6} Talapady N. Bhat,⁷ Jeff M. Blaney,⁸ Evan Bolton,⁹ Gerard Britoigne,¹⁰ David Brown,^{11,12} Stephen K. Burley,^{13,14} David A. Case,⁸ Kirk L. Clark,¹⁵ Tom Darden,¹⁵ Paul Emsley,¹⁶ Victoria A. Feher,¹⁷ Zukang Feng,¹⁸ Colin R. Groom,^{19,20} Seth F. Harris,²¹ Jorg Hendle,¹⁹ Thomas Holder,⁴ Andrzej Joachimiak,²⁰ Gerard J. Kleywegt,²¹

(Author list continued on next page)

¹Molecular Biophysics & Integrated Bioimaging Division, Lawrence Berkeley Laboratory, Department of Bioengineering, UC Berkeley, Berkeley, CA 94720-8235, USA

²DAPI NeuroScience, LLC, San Diego, CA 92131, USA

³Bruker AXS, Inc., Madison, WI 53711, USA

⁴Schrodinger, Inc., New York, NY 10036, USA

⁵Research Collaboratory for Structural Bioinformatics Protein Data Bank, Center for Integrative Proteomics Research, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

⁶Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

⁷BioSystems and Biomaterials Division, NIST, Gaithersburg, MD 20899, USA

⁸Genentech, Inc., South San Francisco, CA 94080, USA

⁹National Center for Biotechnology Information, U.S. National Library of Medicine, Bethesda, MD 20894, USA

¹⁰Global Phasing Ltd., Sanjour, CB3 0XK, UK

¹¹School of Biosciences, University of Kent, Canterbury CT2 7NH, UK

¹²Charles River Ltd., Structural Biology and Biophysics, Cambridge CB10 1XL, UK

¹³Skaggs School of Pharmacy and Pharmaceutical Sciences and San Diego Supercomputer Center, University of California, San Diego, La Jolla, CA 92093, USA

¹⁴Novartis Institutes for BioMedical Research, Cambridge, MA 02139, USA

¹⁵OpenEye Scientific, Cambridge, MA 02142, USA

(Affiliations continued on next page)

Crystallographic studies of ligands bound to biological macromolecules (proteins and nucleic acids) represent an important source of information concerning drug-target interactions, providing atomic level insights into the physical chemistry of complex formation between macromolecules and ligands. Of the more than 115,000 entries extant in the Protein Data Bank (PDB) archive, ~75% include at least one non-polymeric ligand. Ligand geometrical and stereochemical quality, the suitability of ligand models for in silico drug discovery and design, and the goodness-of-fit of ligand models to electron-density maps vary widely across the archive. We describe the proceedings and conclusions from the first Worldwide PDB/Cambridge Crystallographic Data Center/Drug Design Data Resource (wwPDB/CCDC/D3R) Ligand Validation Workshop held at the Research Collaboratory for Structural Bioinformatics at Rutgers University on July 30–31, 2015. Experts in protein crystallography from academe and industry came together with non-profit and for-profit software providers for crystallography and with experts in computational chemistry and data archiving to discuss and make recommendations on best practices, as framed by a series of questions central to structural studies of macromolecule-ligand complexes. What data concerning bound ligands should be archived in the PDB? How should the ligands be best represented? How should structural models of macromolecule-ligand complexes be validated? What supplementary information should accompany publications of structural studies of biological macromolecules? Consensus recommendations on best practices developed in response to each of these questions are provided, together with some details regarding implementation. Important issues addressed but not resolved at the workshop are also enumerated.

Background

The Worldwide PDB (wwPDB; wwpdb.org), the Cambridge Crystallographic Data Center (CCDC; www.ccdc.cam.ac.uk), and the Drug Design Data Resource (D3R; <https://www.drugdesigndata.org>) co-organized a Ligand Validation Workshop on July 30–31, 2015 at Rutgers University. The workshop brought together academic and industrial protein crystallographers, providers of software for crystallography, computational chemists, and experts in data archiving. More than 50 participants from more than 40 organizations discussed and made recommendations on best practices for structural studies of macromolecule-ligand complexes and archiving of the resulting information.

PDB and Historical Context for the Workshop

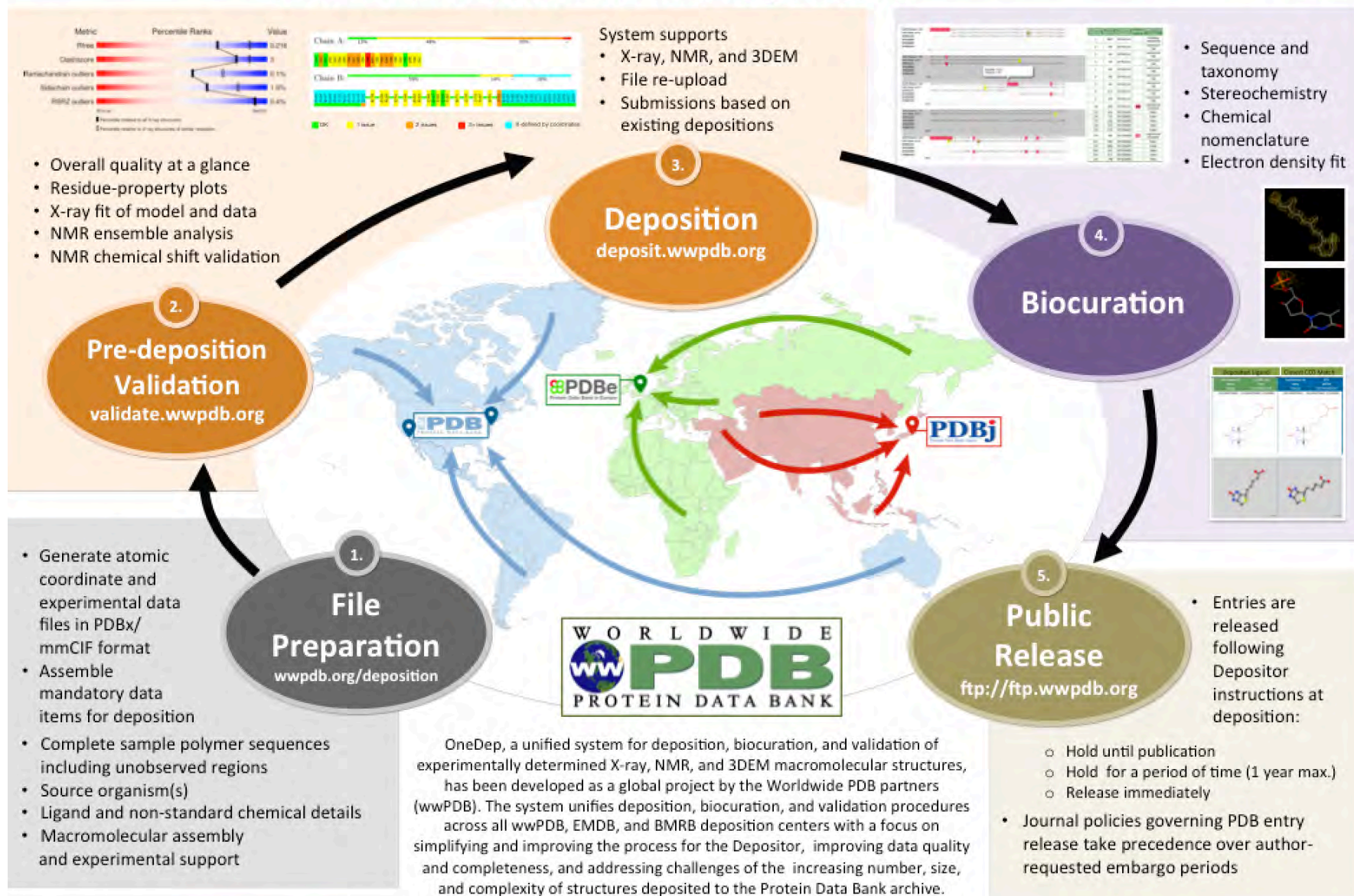
The PDB was established in 1971 with just seven X-ray crystallographic structures of proteins as the first open-access digital

502 Structure 24, April 5, 2016 ©2016 Elsevier Ltd All rights reserved

OneDep overview paper submitted Sep 30, 2016

OneDep: Unified Deposition Portal for the Protein Data Bank

wwPDB Partners - RCSB PDB, PDBe, PDBj, and BMRB



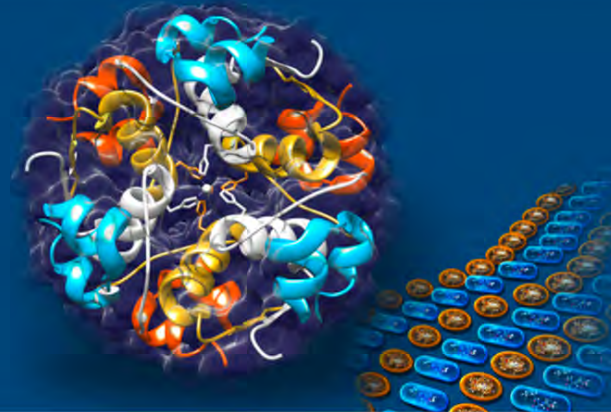
wwPDB Foundation Progress



HOME EVENTS SPONSORS AND DONATIONS BOARD

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.

SUPPORT US



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.

- Website released
- Fundraising on-going
- 2016 Events
 - OneDep Summit
 - Economics and Impact of the Protein Data Bank (PDB) Archive



SciDataCon 2016

<http://foundation.wwpdb.org/>

- HFSP Meeting on Sustainability 52

Crystallography

Stephen K. Burley

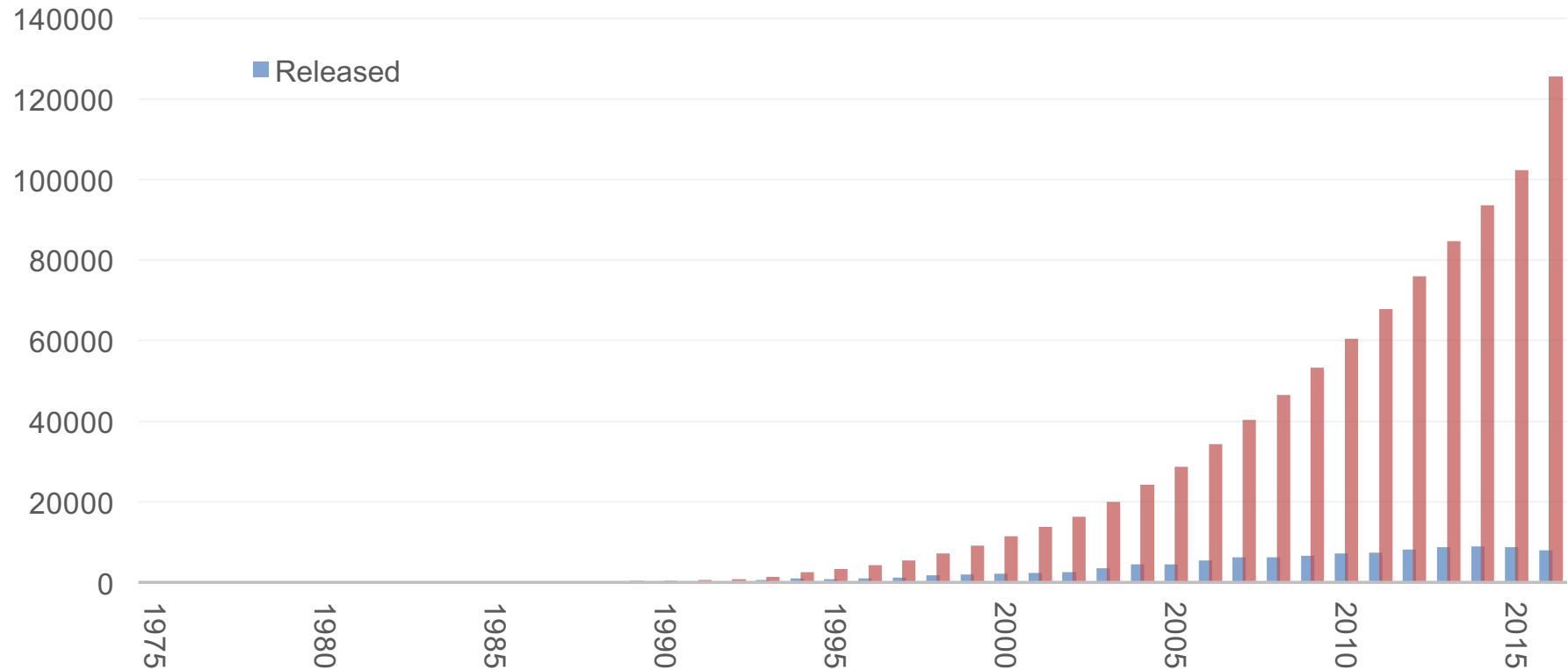


wwpdb.org

Agenda

- Crystallographic Data In Metrics
- wwPDB X-ray VTF 2.0 Meeting - Nov 2015
- Impact of Two-Stage PDB Data Release
- Enabling Depositions from Industry
- Plans for PDBx/mmCIF Working Group Meeting (2017)

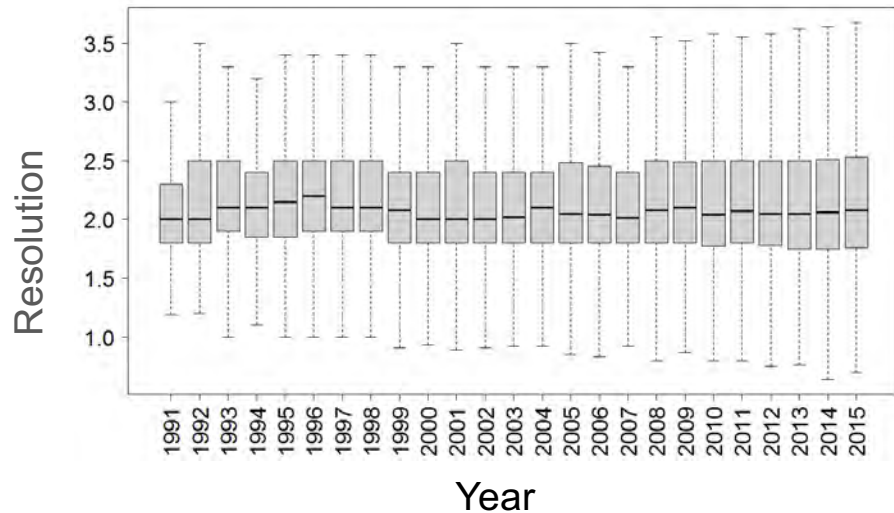
Growth of PX entries



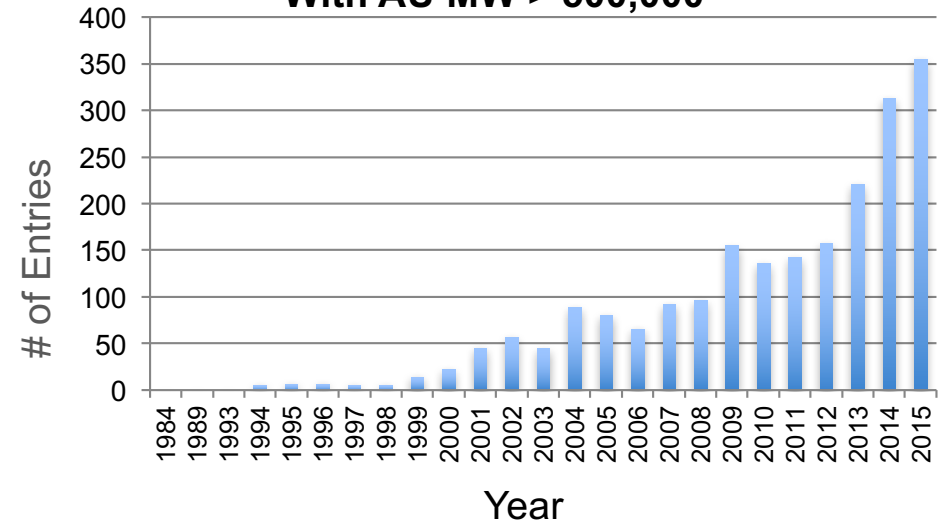
>9,000 New PX Entries Projected for Calendar 2016

Growing Complexity in PX Deposits

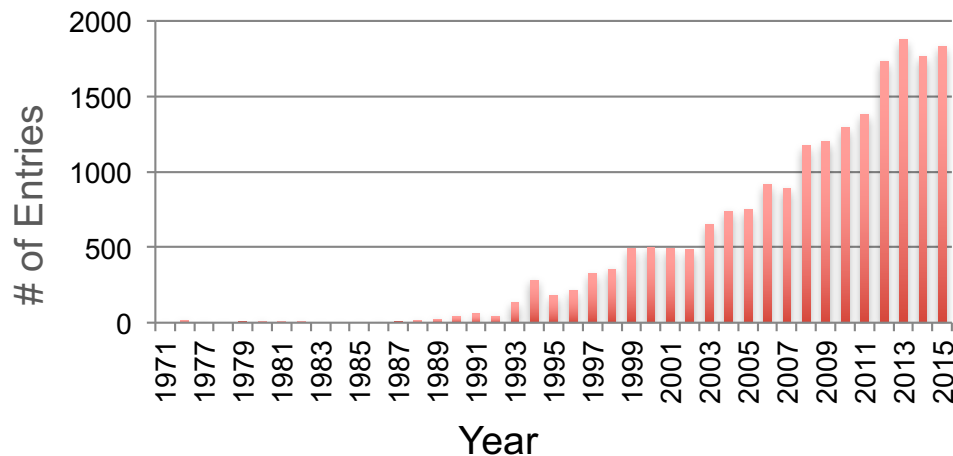
Annual Distribution for High Resolution Limit



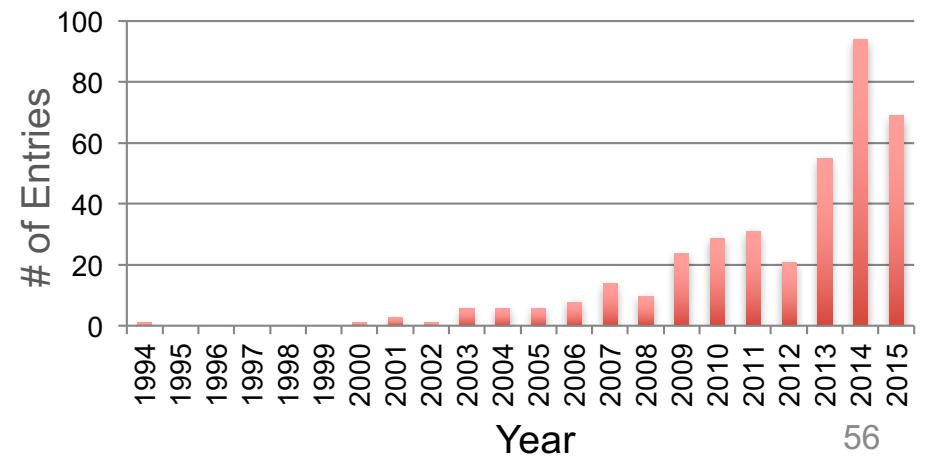
Annual Released Structures With AU MW > 500,000



Total Number of New CCD Entries



Annual Released Large Structures (chains > 62 & atoms > 99999)



wwPDB X-ray VTF 2.0 Meeting

Nov 16-17, 2015 at EMBL-EBI

- Ligands (wwPDB Ligand Validation Workshop, feedback, density-fit analysis and display, Buster Report and ligands, Radiation damage effects, Metal validation, Carbohydrate issues)
- Proteins (MolProbity, Cis-peptides, HNQ flips, Clashes and false positives)
- X-ray-specific (Xtrriage update, serial crystallography, NCS)
- wwPDB issues (pipeline, reports, metadata, annotation, prioritization)



VTF Members: Paul Adams, Gérard Bricogne, Dave Brown, Paul Emsley, Richard Henderson, Nobutoshi Ito, Robbie Joosten, Thomas Lütteke, Michael Nilges, Arwen Pearson, Tassos Perrakis, Randy Read (Chair), Jane Richardson, Janet Smith, Tom Terwilliger, Ian Tickle, Gert Vriend

wwPDB Attendees: Burley, Feng, Gutmanas, Velankar, Westbrook

Ligand Validation Workshop White Paper Published in 2016

- Adams *et al.* (2016) *Structure* 24, 502-508.
- 57 co-authors from 42 institutions/organizations
- Recommendations endorsed unanimously by wwPDB X-ray VTF 2.0

Cell^{press}

Structure
Meeting Report

Outcome of the First wwPDB/CCDC/D3R Ligand Validation Workshop

Paul D. Adams,¹ Kathleen Aertgeerts,² Cary Bauer,³ Jeffrey A. Bell,⁴ Helen M. Berman,^{5,6} Talapady N. Bhat,⁷ Jeff M. Blaney,⁸ Evan Bolton,⁹ Gerard Bricogne,¹⁰ David Brown,^{11,12} Stephen K. Burley,^{5,6,13,*} David A. Case,⁶ Kirk L. Clark,¹⁴ Tom Darden,¹⁵ Paul Emsley,¹⁶ Victoria A. Feher,^{17,*} Zukang Feng,^{5,6} Colin R. Groom,^{18,*} Seth F. Harris,⁸ Jorg Hendle,¹⁹ Thomas Holder,⁴ Andrzej Joachimiak,²⁰ Gerard J. Kleywegt,²¹

(Author list continued on next page)

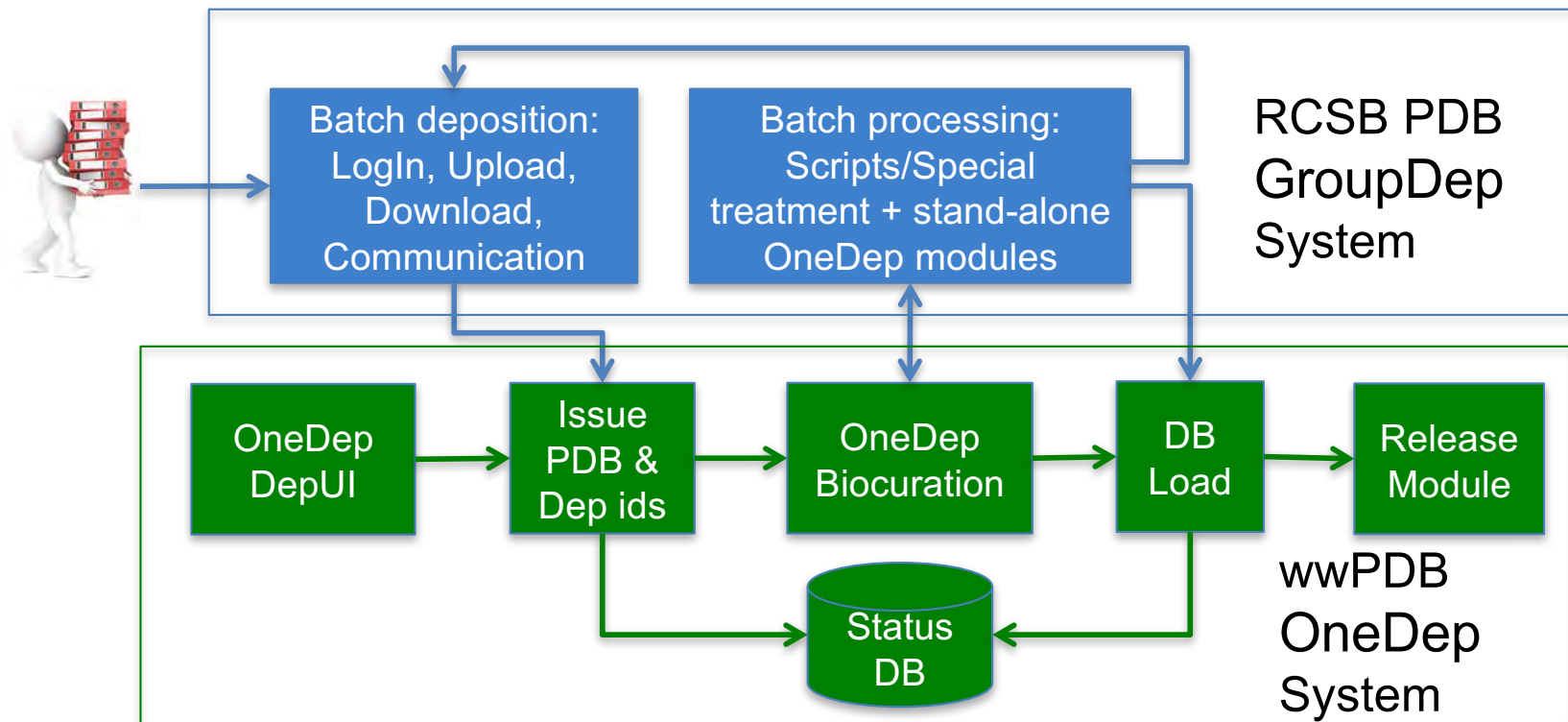
Impact of 2-Stage PDB Data Release

- Every Saturday by 3:00 UTC the wwPDB website provides the following for every new entry stage for Wednesday release:
 - Sequence/s (amino acid or nucleotide) for each distinct polymer
 - Where appropriate, InChI string(s) for each distinct ligand and crystallization pH value(s)
- Support/Statistics for CAMEO
 - 4066 targets: 26 predictors for protein structure prediction
 - 14200 targets: 5 predictors for ligand binding
- Support/Statistics for CAPRI, CASP, and D3R
 - CAPRI: 11 targets: 41 teams
 - CASP: 134 targets: 221 groups registered, 16099 models
 - D3R: Blinded Challenges predicting docking pose/binding affinity for 2 targets/211 compounds; Weekly CELPP challenge coming

Enabling Depositions from Industry

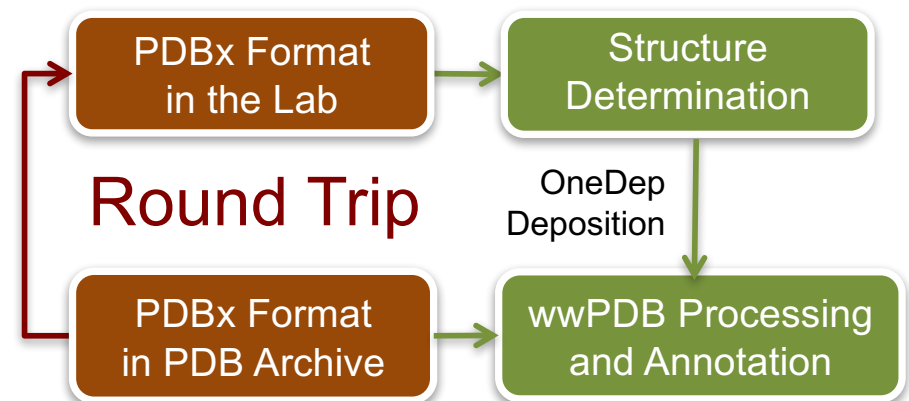
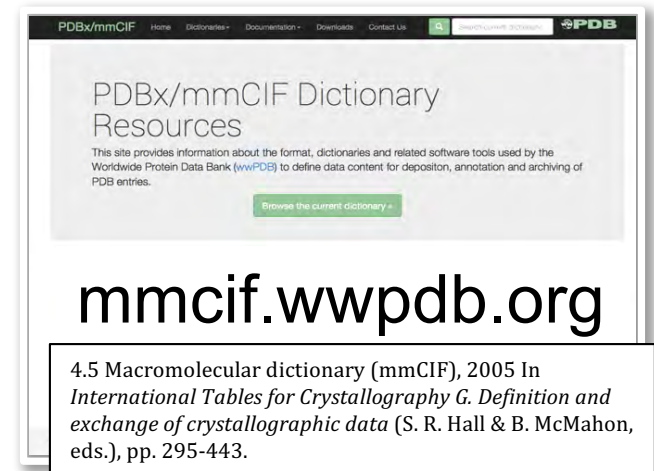
<https://deposit-group.rcsb.rutgers.edu/groupdeposit/>

- Group Deposition processing
 - Requirements set by wwPDB OneDep Team
 - Provided support for D3R Blind Challenges
 - Early Adopters: Roche, Merck Serono, U. Marburg, U. Essex



PDBx/mmCIF Working Group

- PDBx/mmCIF Dictionary
 - Essential for large/complex structures
 - Extensible to new and integrative methods
- PDBx/mmCIF Working Group
 - Community developers support dictionary development
 - PDBx/mmCIF files for OneDep available from CCP4 and Phenix
- Working Group to be reconvened for face-to-face meeting in 2017 at PDBe



Workshop
Participants,
October
2014



Electron Microscopy

Sameer Velankar

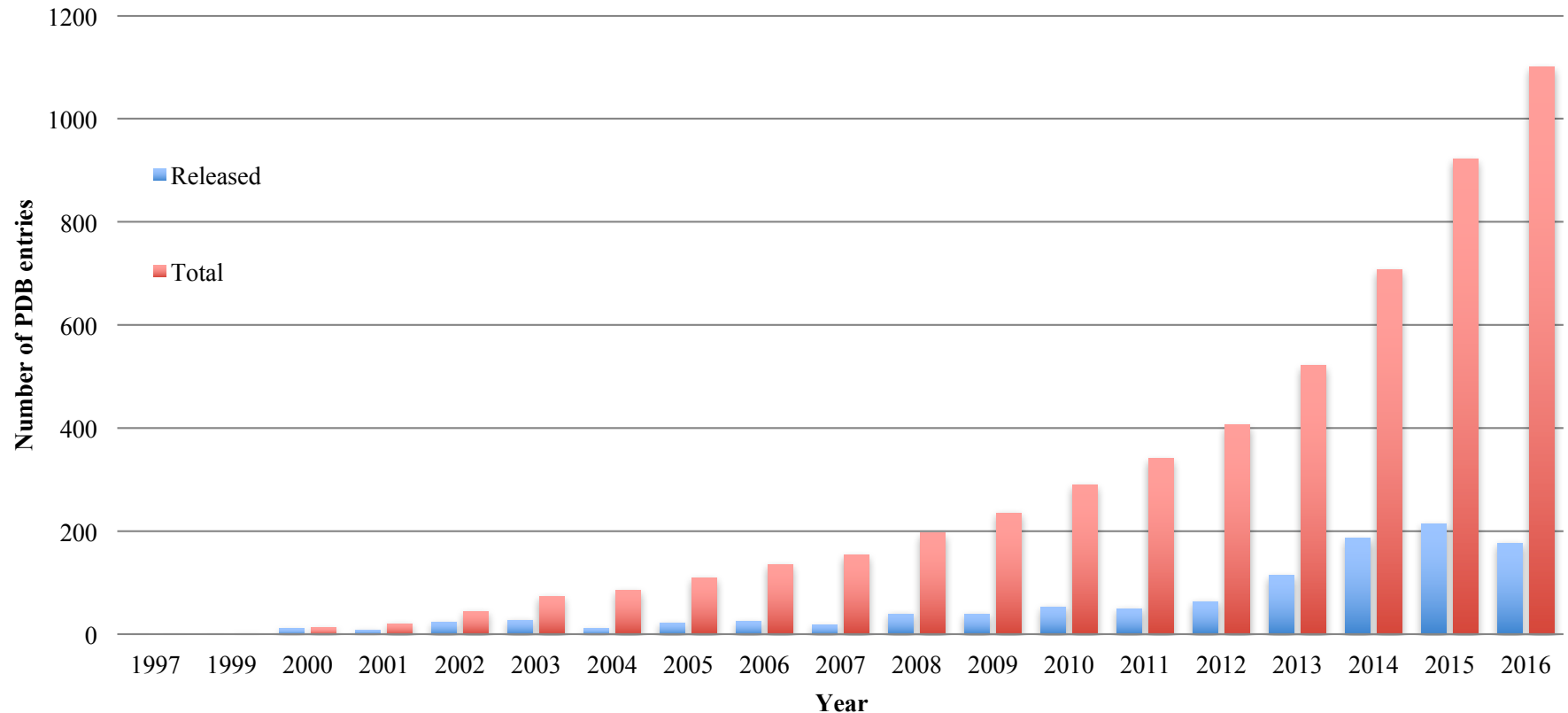


wwpdb.org

Agenda

- Electron Microscopy Data In Metrics
- EM Depositions with OneDep
- EM Validation Reports
- New PDB Policy for 3DEM Structures
- “Resolution Revolution”
- Recent Developments


Growth of PDB EM Entries



As of August 1, 2016, >1100 EM entries in the PDB archive


178 new entries released Jan 1 - Aug 1, 2016

EMDB Depositions with OneDep



WORLDWIDE
wwPDB
PROTEIN DATA BANK

wwPDB Deposition: D_8000200707 -- Requested ID: EMDB



Welcome to the Worldwide Protein Data Bank

[Submit deposition](#)

[All items](#)

[Mandatory items](#)

[Navigation](#)

- ✓ Welcome
- ✓ Communication
- ✓ Re-upload files
- ✓ **Upload summary**
- Admin
 - ✓ Contact information
 - ✓ Grant information
 - ✓ Release status
 - ✓ Entry title & author
 - ✓ Citation information
- Macromolecules
 - ✓ 1) Pleurotolysin
- EM sample
 - ✓ Sample description
- EM experiment
 - ✓ Specimen preparation
 - ✓ Microscopy
 - ✓ Image recording
 - ✓ Reconstruction
 - ✓ Fitting interpretation
- Summary & conditions
- Downloads & reports
 - All files

This page contains a summary of the uploaded data. Please check that the data content here is correct before proceeding. Data problems that require new data to be uploaded may result in the loss of information entered on subsequent pages.

▼ **Upload file summary**

You uploaded 1 files to the system.

Number	Used	File name	Size	File type	File header check
1	*	groel_stagg_map1.mrc	14896524	Main volume data	The file has correct format

▼ **Map conversion report -- groel_stagg_map1.mrc**

	Converted map	Original map
Map title:	::::EMDATABANK.org::::D_8000200707::::	bhead -Background ../maps/GroEL-NRAMM39_100.map ../maps/GroEL-NRAMM39_100.map
Map endianness:	Big endian	Big endian
Map mode:	Image stored as floating point number (4 bytes)	Image stored as floating point number (4 bytes)
Fast, medium and slow axes:	X, Y, Z	X, Y, Z
Grid sampling on x, y, and z:	155, 155, 155	155, 155, 155
Pixel sampling on x, y, and z:	1.2, 1.2, 1.2	1.64, 1.64, 1.64
Cell dimensions (x, y, and z, alpha, beta, gamma):	186.0, 186.0, 186.0, 90.0, 90.0, 90.0	254.2, 254.2, 254.2, 90.0, 90.0, 90.0
Number of columns, rows, and sections:	155, 155, 155	155, 155, 155
Start points on columns, rows, and sections:	-77, -77, -77	-77, -77, -77
Origin in MRC format:	0.0, 0.0, 0.0	0.0, 0.0, 0.0
Space group number:	1	0
Minimum density:	-0.10947169	-0.10947169
Maximum density:	0.08548745	0.08548745

EM Validation Reports

- Reports for all EM entries made public May 4, 2016
- Provides “Table 1” counterpart

4 Experimental information i

Property	Value	Source
Reconstruction method	SINGLE PARTICLE	Depositor
Imposed symmetry	POINT, Not provided	Depositor
Number of images	140155	Depositor
Resolution determination method	FSC 0.143 CUT-OFF	Depositor
CTF correction method	Not provided	Depositor
Microscope	FEI TITAN KRIOS	Depositor
Voltage (kV)	300	Depositor
Electron dose ($e^-/\text{\AA}^2$)	Not provided	Depositor
Minimum defocus (nm)	500	Depositor
Maximum defocus (nm)	3500	Depositor
Magnification	81000	Depositor
Image detector	Not provided	Depositor



wwPDB EM Map/Model Validation Report i

May 17, 2016 – 11:04 AM EDT

PDB ID : 5GAN
 EMD ID : EMD-8012
 Title : The overall structure of the yeast spliceosomal U4/U6.U5 tri-snRNP at 3.7 Angstrom
 Authors : Nguyen, T.H.D.; Galej, W.P.; Bai, X.C.; Oubridge, C.; Scheres, S.H.W.; Newman, A.J.; Nagai, K.
 Deposited on : 2015-12-15
 Resolution : 3.60 Å (reported)
 Based on PDB ID : ?

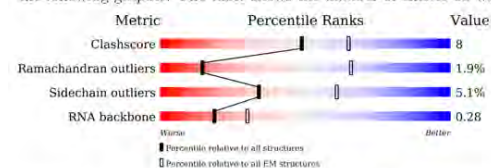
This is a wwPDB EM Map/Model Validation Report for a publicly released PDB/EMDB entry. For rigid body fitted models, validation errors reported here could stem from errors in the original structure(s) used in the fitting. We welcome your comments at validation@mail.wwpdb.org. A user guide is available at <http://wwpdb.org/validation/2016/EMValidationReportHelp>

1 Overall quality at a glance i

The following experimental techniques were used to determine the structure:
ELECTRON MICROSCOPY

The reported resolution of this entry is 3.60 Å.

Percentile scores (ranging between 0-100) for global validation metrics of the entry are shown in the following graphic. The table shows the number of entries on which the scores are based.



Metric	Whole archive (#Entries)	EM structures (#Entries)
Clashscore	114402	924
Ramachandran outliers	111179	726
Sidechain outliers	111093	686
RNA backbone	3027	244

The table below summarises the geometric issues observed across the polymeric chains. The red, orange, yellow and green segments on the bar indicate the fraction of residues that contain outliers for >=3, 2, 1 and 0 types of geometric quality criteria. A grey segment represents the fraction of residues that are not modelled. The numeric value for each fraction is indicated below the corresponding segment, with a dot representing fractions <=5%.

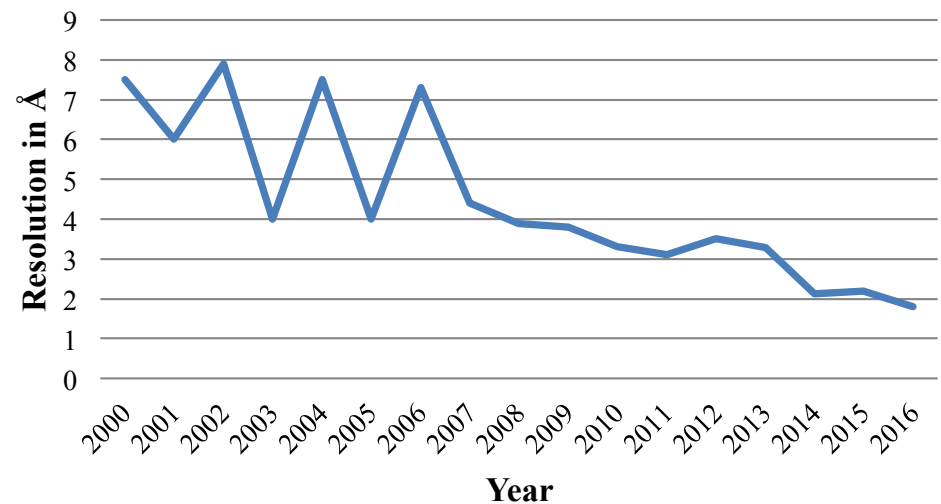
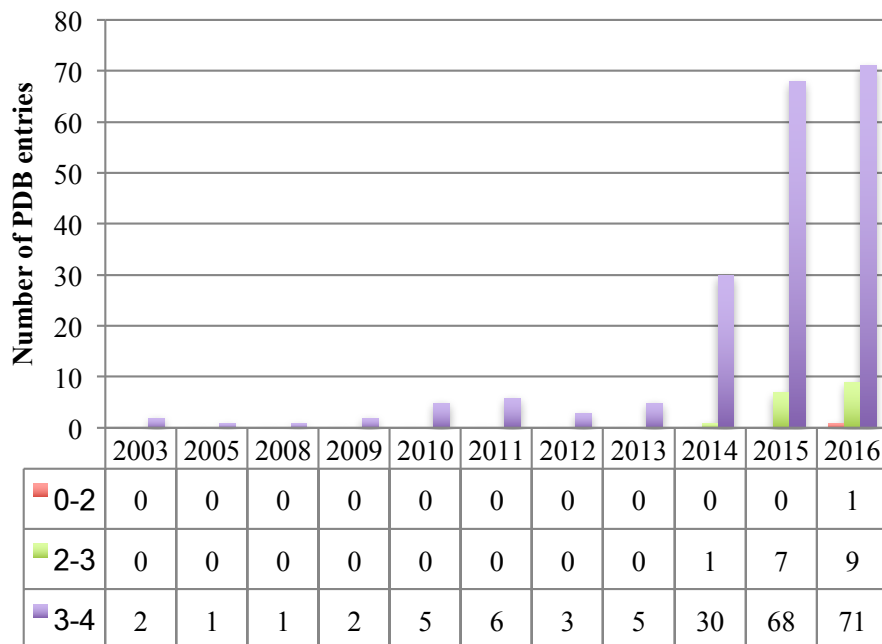
Mol	Chain	Length	Quality of chain
1	V	160	28% 38% 9% 23%
2	W	112	31% 23% 17% 29%

New wwPDB Policy for 3DEM Data

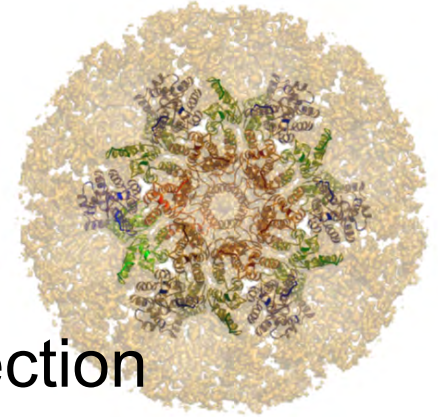
- Effective Sep 6, 2016, deposition of atomic models determined by 3DEM to the PDB requires prior or simultaneous deposition of the associated 3DEM mass density maps to EMDB
- For joint PDB/EMDB depositions, the hold period is the same for both map(s) and model(s)

“Resolution Revolution”

- 1.8Å structure in 2016 (PDB ID 5K12; EMD-8194)
- Increasing number of 3DEM structures at 2-4Å resolution (75 in calendar 2015 and 80 in first 7 months of 2016)



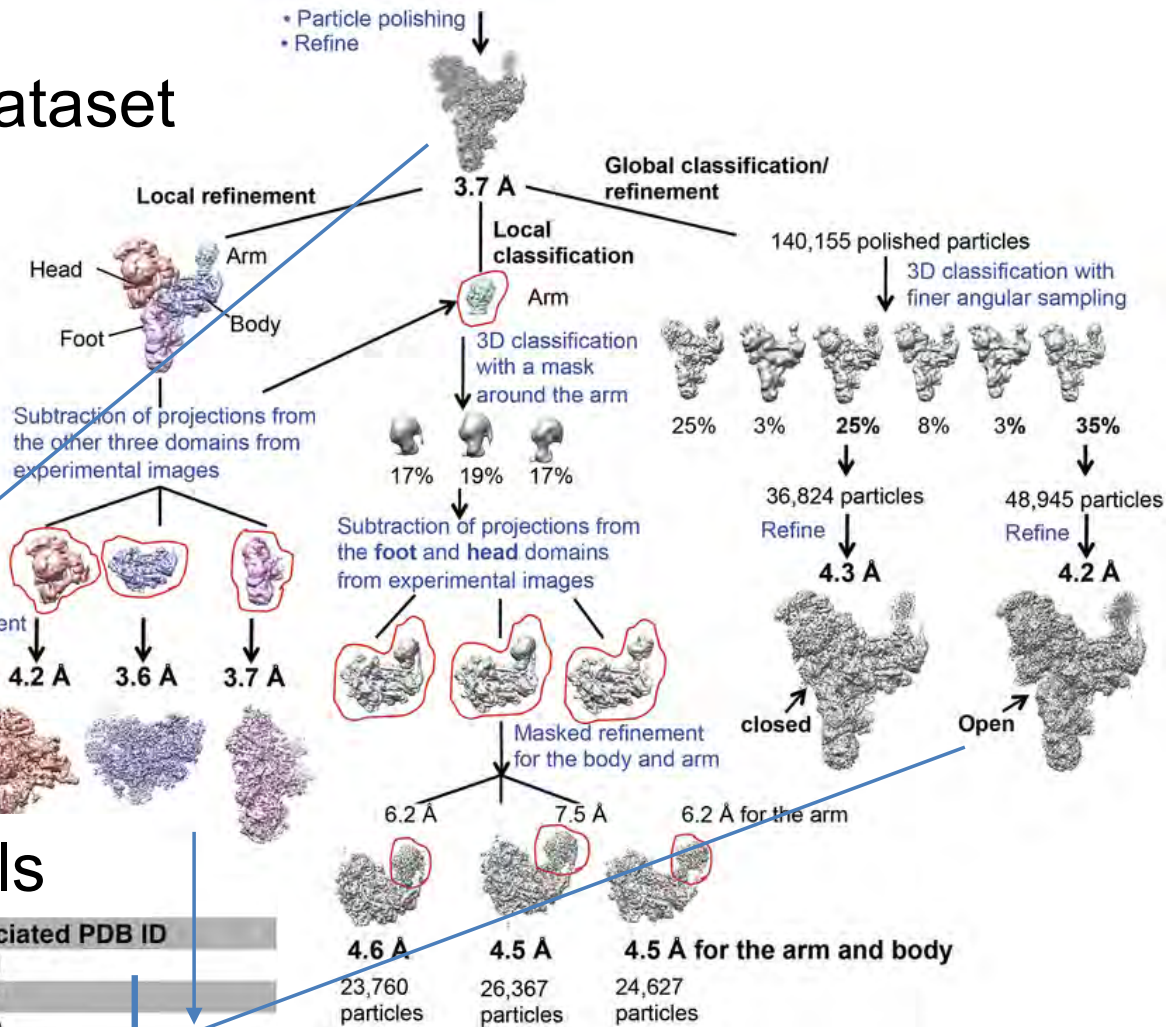
Recent Developments



- High Resolution Single-Particle Structures
 - Direct detectors; automated, fast data collection
 - Better sample preparation and handling (Automation)
 - Improved image processing (able to detect sample motion) and reconstruction software (3D classification)
 - 3.4Å resolution hemoglobin structure reported
- 2.2Å resolution β -galactosidase structure
PDB ID 5a1a - Bartesaghi *et al.* (2015) *Science* 348, 1147
- High-Resolution Structure from Cryo-EM Tomography
“We applied optimized cryo-electron tomography and subtomogram averaging to resolve this region within assembled immature HIV-1 particles at 3.9Å resolution and built an atomic model.”
PDB ID 5l93 - Briggs *et al.* (2016) *Science* 353, 506

1 Dataset → Multiple Coordinate Sets

a) One Dataset



b) *In silico* Classification

c) Multiple Maps and Models

Maps	EMDB code	Associated PDB ID
Overall map	EMD-8012	5GAN
Body map	EMD-8014	5GAP
Head map	EMD-8013	5GAO
Foot map	EMD-8011	5GAM
Global class 1 (closed state)	EMD-8007	
Global class 2 (open state)	EMD-8006	
Masked body/arm class 1	EMD-8008	
Masked body/arm class 2	EMD-8009	
Masked body/arm class 3	EMD-8010	

Nguyen et al., 2016,
Nature, **530**, 298-302

Archiving 3DEM Entries in the PDB

- Model Coordinates archived in PDB
- Mass Density Maps archived in EMDB
- Diverse model refinement strategies in use
 - e.g., Structure factors (“SF”s) derived from mass density maps used as “experimental data” in 3DEM atomic coordinate refinement (~10% of challenge)
 - e.g., Masked refinement strategies should consider B-factor distribution and variance within a given masked region *versus* across the entire map
 - How should wwPDB handle these cases?

Opportunities for Capturing 3DEM Refinement Workflows

- Make identification of model coordinates that have not been refined mandatory (i.e., docked as rigid bodies)?
- In masked refinement
 - Identify individually masked regions
 - Capture refinement statistics for each individually masked region
- Two of the commonly used EM model refinement programs (REFMAC and PHENIX) currently output mmCIF with detailed refinement statistics
 - Invite EM model refinement software developers to the the next face-to-face meeting of the PDBx/mmCIF Working Group at EBI

Engaging the EM Community

- EMDatabank is currently conducting two large challenges
 - EM Images → Mass density maps
 - EM Mass density maps → Atomic coordinate models
- Publication(s) from challenges expected in 2017
- wwPDB will work with EMDB and EMDatabank to reconvene the wwPDB EM Validation Task Force together with major 3DEM software providers to review the present status and determine a consensus path forward

NMR

John L. Markley



wwpdb.org

BMRB Staffing - Madison

- Director: Pedro Romero
- Other Staff members
 - Director Emeritus: Eldon Ulrich
 - Head Annotator: Hongyang Yao
 - Systems Manager (75%): Dmitri Maziuk
 - Programmer: Jonathan Wedell
 - Assistant Scientist: Kumaran Baskaran
 - Undergraduates
- Computer Science advisors
 - Miron Livny (Univ. of Wisconsin Madison)
 - Yannis Ioannidis (Univ. of Athens, Greece)

BMRB Staffing – PDBj Osaka

- Director: Toshimichi Fujiwara
- Other staff members
 - Naohiro Kobayashi
 - Takeshi Iwata
 - Masashi Yokochi

BMRB External Advisory Board

- Met Apr 9, 2016
- Membership
 - Art Edison - Athens, GA (2015-2019)
 - Valérie Copié - Bozeman, MT (2011-2016)
 - Peter Tompa - Brussels, Belgium (2015-2019)
 - Michael Summers - Baltimore, MD (2015-2019)
 - Mei Hong - Cambridge, MA (2011-2016)

PDBj-BMRB Representative - Naohiro Kobayashi

PDBe Representative - Aleks Gutmanas

- Advisory Board Report in Appendix 4

Advisory Board Recommendations

- Develop a new vision ASAP
 - Protein Structure Initiative (PSI) terminated
- BMRB expansion into new fields/communities
 - Intrinsically disordered Proteins (IDP)
R01 proposal on IDP NMR Toolset and Resources submitted (2016)
 - Metabolomics
 - NMR metabolomics workshop for UW community (6/10/2016)
 - Metabolomics Standards DB R01 proposal in preparation
 - Biological ssNMR
 - RNA structural biology
 - Outreach/Workshops
 - Multiple workshops through NMRbox project planned
 - Increase BMRB User base through tight NMRbox integration

NMRbox Project (Uconn/UW P41)

- Full NMR software toolset through Virtual Machine (VM)
 - Facilitates reproducibility
 - Eliminates software compatibility issues
 - Platform as a service (PAAS) option reduces infrastructure needs for users
- BMRB Integration
 - VM tightly integrated with BMRB through API (already available)
 - BMRB can be accessed from any step in the NMR workflow to provide stats and visualization
 - BMRB new graphical library (in R) is now available and complements our DEVis visualization system
 - Automated deposition of NMR data to BMRB
 - Workflow manager to be implemented
 - Workflows stored at BMRB

NMR Depositions (9/1/15 – 8/31/16)

Site	NMR structures (NMR data sets annotated)	Experimental NMR data without structures (annotated)	Total (number annotated)
BMRB	264 (273)*	231	496 (504)*
BMRB (OneDep)	208		208
PDBe (AutoDep)	9*		9*
PDBj-BMRB	26	3	29
Total	507	234	741

* Entries deposited *via* AutoDep were sent to BMRB for annotation

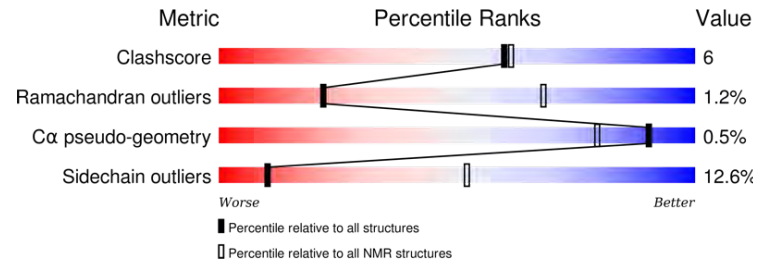
NMR Depositions (2013-2016)

Year	NMR depositions to PDB Archive	Experimental NMR data without structures	Total
2013	486	308	714
2014	506	240	746
2015	427	333	760
2016	507	234	741

OneDep-BMRB Integration

- All NMR structure depositions are routed to OneDep
 - ADIT-NMR no longer accepts structures
- Since early 2016, 208 NMR structures have been deposited *via* OneDep
- NMR validation is handled by OneDep
 - A basic NMR validation report has been implemented
 - BMRB will provide expertise and help in expanding the OneDep NMR validation protocol to include restraints
- Entries deposited *via* OneDep are sent to BMRB, where they are checked by BMRB software and biocurated as needed
- Biocuration protocol
 - E-mail consultations between BMRB biocurators and authors are copied to the OneDep system for archiving
- ADIT-NMR continues to be used by BMRB for NMR depositions that do not involve atomic coordinates

NMR Validation Pipeline

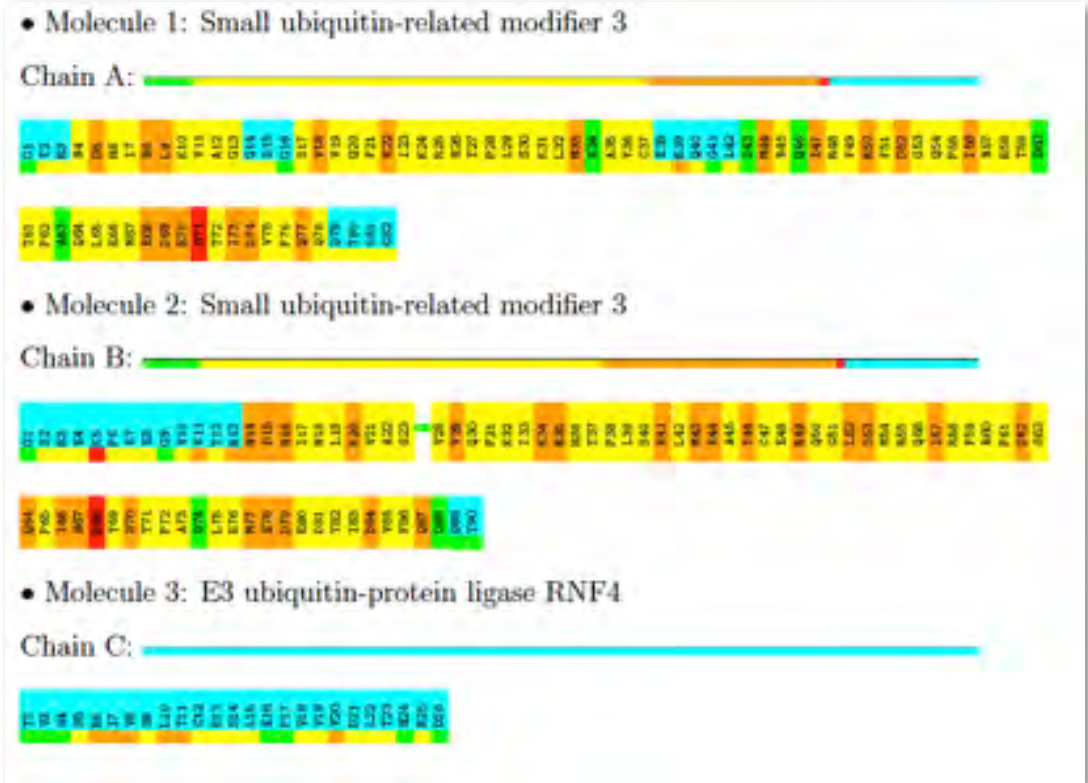


- Reports include
 - Model Validation – Updated following VTF feedback
 - Chemical Shift Validation
 - Annotation Information
- Incorporated into OneDep
- Applied to all NMR entries in PDB
- Reports released May 2016
- Outstanding Tasks:
 - Develop more complete NMR validation protocol: Incorporate restraint validation and eventually peak lists – BMRB to help formulate the definitive protocol

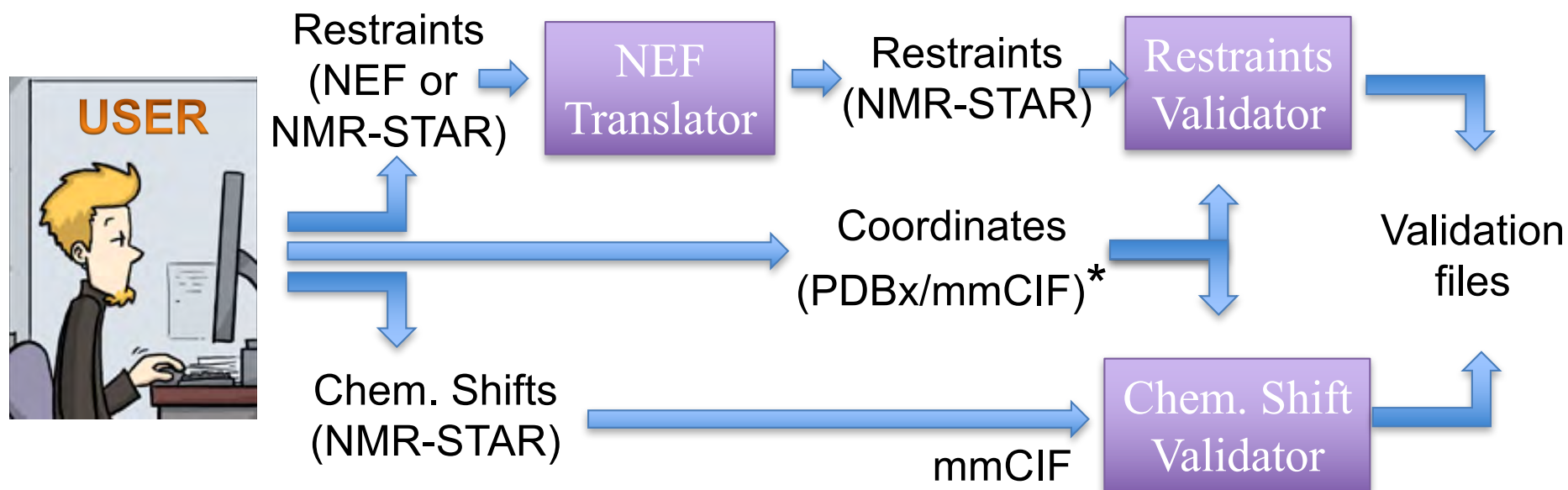
NMR Validation Report: New Features

Mol	Chain	Length	Quality of chain
1	A	82	
2	B	90	
3	C	25	

- Clearer distinction of well-defined and ill-defined (cyan) regions
- Residue plots for average scores and for representative (“medoid”) model
- Plots for all models in the full report



Proposed OneDep Workflow



* PDBx/mmCIF atom specification table now includes a column for NEF atom specifications for consistency

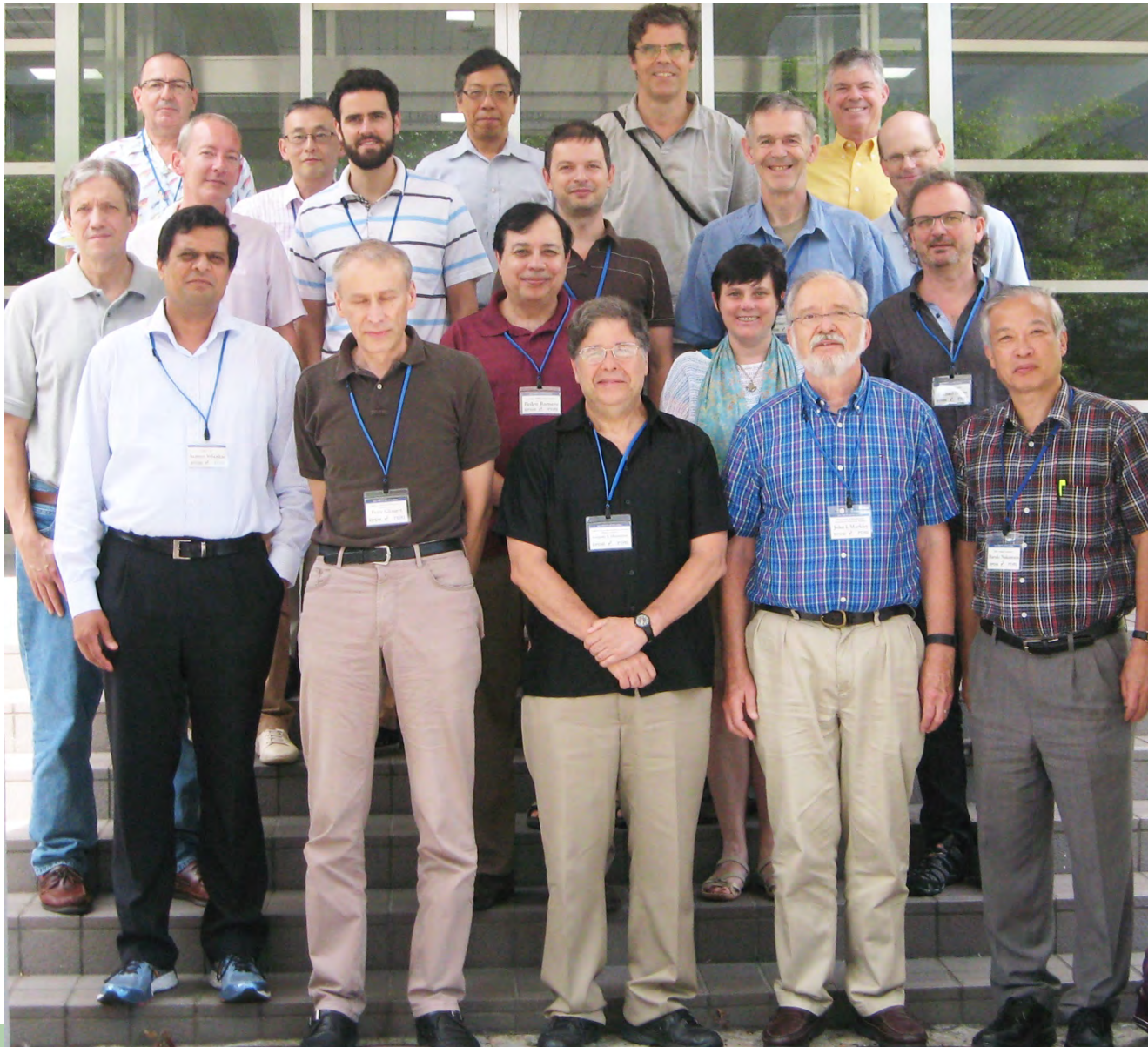
BMRB Involvement in NEF

- Participated in 2014 and 2015 NEF Workshops
- Reviewed format proposal documents (8/2015)
- Co-authored *Nature Struct. Mol. Biol.* 22, 433-434
- Transmitted whitepaper to wwPDB PIs and NEF team (outlining issues/recommendations, 9/15/2015)
- Designated as wwPDB liaison to NEF (3/2/2016)
- Received NEF response to whitepaper (3/10/2016)
- Surveyed NEF membership
 - Responses received 6/7/2016; Testing examples requested
- Co-organized Joint wwPDB NMR VTF-NEF Meeting in Osaka 8/26-27/2016)
- BMRB/NEF issue resolution meeting (9/8/2016)
- BMRB to participate in testing rounds (Fall 2016)

Joint wwPDB NMR VTF-NEF Meeting Hosted by PDBj Aug 26-27, 2016

- VTF-NEF Members (software tools) in attendance
 - A. Bax (TALOS), R.A. Byrd, P. Güntert (CYANA), T. Herrmann (UNIO), G.W. Montelione (Autostructure/PSVS), M. Nilges (CNS/ARIA), T. Polenova, C. Schwieters (Xplor-NIH), N. Sgourakis (CS-Rosetta), G. Vuister (CCPN/CING)
- Unable to attend
 - S. Butcher, D. Case (AMBER), J.S. Richardson (MolProbity), W. Vranken (CCPN), D. Wishart (PROSESS/PANAV/SHIFTX2)
- wwPDB Observers
 - (PDBj) H. Nakamura, T. Fujiwara, N. Kobayashi
 - (BMRB) J.L. Markley, P. Romero
 - (RCSB PDB) S.K. Burley, J. Westbrook
 - (PDBe) S. Velankar, A. Gutmanas

Joint wwPDB NMR VTF-NEF Meeting Hosted by PDBj Aug 26-27, 2016



NEF Implications for wwPDB

- OneDep system to start accepting NEF files
 - Nomenclature checks to be implemented
- NEF will be translated to NMR-STAR for NMR structure validation within OneDep (BMRB actively involved in this process)
 - Restraints-based structure validation in development at wwPDB
- NEF will include raw, software-specific data for use in cases not covered by the format (BMRB has legacy translators)
 - Software-specific NEF tags can be used
- Validation pipeline to process the translated NMR-STAR files
 - Translated from either NEF or raw data, depending on experiment and format coverage
- Eventually: Mandatory deposition of NEF or NMR-STAR formats

Looking Ahead

John L. Markley



wwpdb.org

Plans for the Coming Years I

- 2016/2017 (OneDep Team)
 - Implement Ligand Validation Workshop recommendations
 - Implement support for NMR/SAS Hybrid Structures
 - Collect experimental evidence from Depositors relating to Quaternary Structure (voluntary)

Plans for the Coming Years II

- 2016/2017
 - Remediation
 - Begin work on carbohydrates (RCSB PDB)
 - Begin work on post-translational modifications (PDBe)
 - wwPDB Partners
 - Extend PDBx/mmCIF dictionary mirroring and management
 - Audit weekly release process and assess
- wwPDB AC meeting at RCSB PDB Rutgers

Plans for the Coming Years III

- 2018
 - Begin process to extend franchise to appropriately qualified wwPDB partner sites in China and India
 - wwPDB AC meeting at PDBe EMBL-EBI
 - Begin planning process for PDB 50th Anniversary in 2021 with celebratory scientific meetings and outreach events by the wwPDB and each wwPDB partner

RCSB PDB to Host 2017 wwPDB AC

- Date: Friday Oct 6 or 13, 2017
- Preferences please?
- Location:
 - Center for Integrative Proteomics Research
 - Rutgers, The State University of New Jersey
 - Piscataway, NJ 08854
 - USA

Questions for the wwPDB AC

Stephen K. Burley



wwpdb.org

Questions for the wwPDB AC

1. Does the wwPDB AC concur with the recommendation by the wwPDB Partners that PDB Archival Entries associated with publicly-archived preprints be released as outlined in Appendix 1?
2. Does the wwPDB AC concur with adoption of the Implementation Plan for Versioning of PDB Archival Entries outlined in Appendix 2?

Questions for the wwPDB AC (cont.)

3. Does the wwPDB AC concur with adoption of the Implementation Plan for broadening capture of ORCID identifiers outlined in Appendix 3?
4. Does the wwPDB AC have any questions or concerns regarding the individual RCSB PDB, PDBe, PDBj, or BMRB Advisory Committee reports provided in Appendix 4?

Acknowledgements and Closing Remarks

John L. Markley and R. Andrew Byrd



wwpdb.org

People who helped organize this meeting

- Lai Bergeman (Biochemistry)
- Allyson Miller (Morgridge Institute)
- Miron Livny (Morgridge Institute)
- BMRB Staff Members

Funding for the Meeting

- NIH grant that supports BMRB
- Biochemistry Department and University of Wisconsin-Madison
- wwPDB Foundation
- Grants supporting RCSB PDB, PDBj, and PDBe

Closing Remarks

Thank all of you for your support of the wwPDB and for taking the time to attend the meeting

Safe travels home

2016 wwPDB AC Meeting Appendix Materials

Appendix 1: Proposed policy regarding release of PDB archive depositions associated with publicly-archive preprints

Appendix 2: Proposed plan for Plan for Versioning of PDB Archival Entries

Appendix 3: ORCID Proposal

Appendix 4: wwPDB Partner Reports from their most recent advisory committee meetings

- BMRB
- PDBe
- RCSB PDB
- PDBj

Appendix 5: wwPDB Publications

1. Berman HM, Burley SK, Kleywegt GJ, Markley JL, Nakamura H, Velankar S. The archiving and dissemination of biological structure data. *Curr Opin Struct Biol.* 2016 Jul 21;40:17-22. doi: 10.1016/j.sbi.2016.06.018. [Epub ahead of print] Review. PubMed PMID: 27450113.
2. S. K. Burley, H. M. Berman, G. J. Kleywegt, J. L. Markley, H. Nakamura, and S. Velankar, "Protein Data Bank - the single global macromolecular structure archive managed by the Worldwide Protein Data Bank," In "Macromolecular Crystallography" Volume in *Methods in Molecular Biology* (Springer). Co-editors A. Wlodawer, Z. Dauter, and M. Jaskolski, submitted.
3. Sali A, Berman HM, Schwede T, Trewhella J, Kleywegt G, Burley SK, Markley J, Nakamura H, Adams P, Bonvin AM, Chiu W, Peraro MD, Di Maio F, Ferrin TE, Grünewald K, Gutmanas A, Henderson R, Hummer G, Iwasaki K, Johnson G, Lawson CL, Meiler J, Marti-Renom MA, Montelione GT, Nilges M, Nussinov R, Patwardhan A, Rappsilber J, Read RJ, Saibil H, Schröder GF, Schwieters CD, Seidel CA, Svergun D, Topf M, Ulrich EL, Velankar S, Westbrook JD. Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop. *Structure.* 2015 Jul 7;23(7):1156-67. doi: 10.1016/j.str.2015.05.013. Epub 2015 Jun 18. PubMed PMID: 26095030; PubMed Central PMCID: PMC4933300.
4. Adams PD, Aertgeerts K, Bauer C, Bell JA, Berman HM, Bhat TN, Blaney JM, Bolton E, Bricogne G, Brown D, Burley SK, Case DA, Clark KL, Darden T, Emsley P, Feher VA, Feng Z, Groom CR, Harris SF, Hendle J, Holder T, Joachimiak A, Kleywegt GJ, Krojer T, Marcotrigiano J, Mark AE, Markley JL, Miller M, Minor W, Montelione GT, Murshudov G, Nakagawa A, Nakamura H, Nicholls A, Nicklaus M, Nolte RT, Padyana AK, Peishoff CE, Pieniazek S, Read RJ, Shao C, Sheriff S, Smart O, Soisson S, Spurlino J, Stouch T, Svobodova R, Tempel W, Terwilliger TC, Tronrud D, Velankar S, Ward SC, Warren GL, Westbrook JD, Williams P, Yang H, Young J. Outcome of the First wwPDB/CCDC/D3R Ligand Validation Workshop. *Structure.* 2016 Apr 5;24(4):502-8. doi: 10.1016/j.str.2016.02.017. PubMed PMID: 27050687.

Appendix 1
Proposed wwPDB Policy for
Release of PDB Entries Described in Public Preprint Archives

Rationale:

wwPDB partners have received enquires relating to policies governing release of PDB Entries described in Public Preprint Archives (e.g., bioRxiv). Specifically, the wwPDB was asked if submission of a preprint to bioRxiv or equivalent resource constituted publication, which would normally trigger release of an on hold PDB Entry (HPUB).

Interim wwPDB Management:

In response to such enquires, the wwPDB PIs met and agreed to the following interim arrangement:

- PDB Depositors using Public Preprint Archives would be strongly encouraged to include the Official wwPDB Validation Report obtained when the PDB Entry is finalized with the preprint submission.
- PDB Depositors would be asked to provide the preprint DOI for use as the Primary Citation.
- wwPDB PIs would research the matter and present a definitive policy proposal to the Advisory Committee Meeting in October 2016.

Proposed wwPDB Policy:

wwPDB PIs conferred with key opinion leaders involved in Public Preprint Archives and developed the following policy proposal for presentation to the Advisory Committee:

- 1) The wwPDB regards submissions to Public Preprint Archives as *bona fide* publications.
- 2) PDB Depositors are required to provide the Public Preprint Archive DOI, which will be designated as the Preprint Citation in the PDBx/mmCIF file (new data item).
- 3) PDB Depositors are required to release any Entries described in the Preprint Citation similar to entries with HPUB status (i.e. release upon publication).
- 4) The wwPDB will not implement additional processes to identify when a preprint is made public, preferring, instead to rely on PDB Depositor integrity and peer pressure from users of Public Preprint Archives for ensure entry release. If the wwPDB becomes aware that a preprint associated with an HPUB entry has been made public, it will be released immediately.
- 5) When the journal publication appears, the journal DOI will appear as the Primary Citation in the PDBx/mmCIF file.

Roadmap for Implementation:

- PENDING: Draft appropriate wwPDB Policy statement for posting on wwpdb.org.

Appendix 2 Proposed wwPDB Plan for PDB Archive Versioning

As requested at the 2015 wwPDB AC Meeting, the wwPDB proposes the following plan to support versioning of the PDB archive as follows:

1. (A) Designate original Depositor provided experimental data as the Experimental Data of Record, which are NOT subject to versioning.

(B) Designate original Depositor provided atomic coordinates as the Structure of Record, identified as Version 1-0.

(C) Current wwPDB policies regarding deposition of new atomic coordinates refined against existing experimental data produced by another Depositor will remain unchanged (i.e., peer-reviewed publication required; new PDB ID Code issued).

(D) wwPDB will no longer assign new accession codes for atomic coordinate replacements by the original Depositor.

(E) All Major and the latest Minor Versions of a given archival entry will be maintained within the PDB ftp repository.

(F) Substantial revisions requiring assignment of a new Major Version include replacement of atom coordinates, or changes in polymer sequence(s) and ligand chemical description (e.g., Version 2-0 when the original Depositor, etc. replace the atomic coordinates for the first time).

(G) Other changes in archival data, such as updating citation information upon publication, will be treated as Minor Revisions (e.g., V1-1, V1-2, etc.).
2. Adopt an extended 12-character accession code of the form "PDB_00001ABC", consisting of the "PDB_" prefix to identify the archive and eight case-insensitive alphanumeric characters.

Existing 4-character PDB ID codes would be "grandfathered", and right justified within the new accession code style.
The wwPDB plans to continue generating codes compatible with the current 4-character PDB ID code for as long as possible.
3. Continue support for the existing wwPDB ftp tree, while providing a separate parallel file system in the wwPDB ftp repository containing data files with standardized file names containing explicit version numbers.

4. Standard template for data file names will include the new 12-character accession code, the data content type, the version identifier, the data file format, and compression type.

For example, the atomic coordinate data file for Version 2-3 of the PDB entry designated 1abc would be assigned the following file name:

PDB_00001abc_###_v2-3.cif.gz

PDB_00001abc = new accession code
= placeholder for data content type (e.g., xyz, nmr-cs)
v2-3 = Version identifier (Major Version 2; Minor Revision 3)
cif = data file format
gz = compression type

Roadmap for Implementation:

- PENDING: Formal requirements setting and project planning by the wwPDB OneDep Team.
- PENDING: Draft the requisite 60-day statement of notice for posting on wwpdb.org.

Appendix 3 Proposed wwPDB Plan for ORCID Identifiers

As agreed at the 2015 wwPDB AC Meeting, the wwPDB recently implemented ORCID identifier capture within the OneDep system. Initial adoption rates are not discouraging. The wwPDB now proposes to undertake a two-step process, culminating in mandatory capture of ORCID identifiers for all Depositors and public release of ORCID information with each new entry.

A summary of our proposal follows:

New ORCID-related OneDep Processes:

- 2017: Enable voluntary provision of ORCID identifiers for all Depositors contributing a new entry. [As opposed to our current practice of capturing ORCID identifiers volunteered by the primary Depositor(s) and the Principal Investigator.]
- 2017: Enable public release of captured ORCID identifiers in the PDBx/mmCIF files for new entries made available from wwPDB ftp download sites.
- 2018: Make capture of ORCID identifiers for all Depositors contributing to a new entry mandatory.

Benefits to Depositors, Journals, Users, PDB Archive, and wwPDB:

- Ensures definitive identification of our Depositors.
- Allows correlation of Depositor identities in released PDB entries with ORCID identifiers in journal articles and grant applications.
- Provides additional readiness for extending the PDB franchise to new wwPDB partners in Asia and South America.

Roadmap for Implementation:

- PENDING: Formal requirements setting and project planning by the wwPDB OneDep Team.
- PENDING: Drafting of the requisite 60-day statements of notice for posting on wwpdb.org in 2017 and 2018.

**Report from the BMRB Advisory Committee
Madison, WI, April 9, 2016**

Committee members present:

Art Edison, University of Georgia (chair)

Peter Tompa, Vrije Universiteit Brussel

Valérie Copié, Montana State University

Michael Summers, HHMI and University of Maryland, Baltimore County

Committee members absent:

Mei Hong, MIT

The advisory committee (AC) met with BMRB leadership and staff, as well as visitors from other wwPDB sites for a one-day progress report summary of activities in the BMRB. The NIH funding cycle is entering its second year (out of five). The BMRB has undergone several transitions, most notably a change in their leadership from Eldon Ulrich to Pedro Romero, who is the new BMRB Director. Another notable change is the recently funded P41 to Prof. Hoch on the NMRBox project, which includes the BMRB.

The AC recognizes that the BMRB is funded by an R01 mechanism, which is often used to fund smaller projects. However, we think that for the BMRB to be successful, the BMRB leadership needs to continue to think more broadly than a typical R01 grant. The BMRB has historically served the international NMR community, and this needs to continue with the additional focus on the growth in areas outlined below.

The field of biological NMR is undergoing some major transformations, and several of these have the potential to impact the BMRB. Some of the more notable recent changes are the elimination of the protein structure initiative (PSI) by the NIH and the company Agilent leaving the NMR field. The PSI is changing the landscape in biological protein NMR, and the statistics reported by the BMRB staff about data depositions are dropping because of the loss of several major projects in this area. Agilent's decision to disinvest from selling NMR instruments move will likely have less impact on the BMRB, but the commercial development in the field is now largely being led by only a single vendor (Bruker), which may become problematic to stimulate new technical developments in the field of biological NMR.

The primary concern of AC was that BMRB needs to develop a clear and bold vision for the next 5-10 years. We understand that there are many current challenges and a steep learning curve for the new leadership. However, the report to the AC lacked a big vision, and we feel that now is the time to define where the BMRB is going within the landscape of the biological NMR field. Some examples:

The BMRB staff clearly are aware of the declining protein data depositions, and it isn't difficult to forecast that this decline will continue or at least won't rebound quickly. Yet, most of the

focus of the BMRB staff is on maintaining and improving the (very valuable) tools that they have built for many years to support protein structural biology.

Very little was mentioned about the emerging field of **RNA structural biology**, despite the growing importance in biomedicine and biology in small RNAs. These are not easy to study by other techniques such as X-ray crystallography or cryo-EM, and this is a clear area of growth for NMR. The BMRB should be anticipating this change and responding with workshops, surveys of major labs doing RNA structural biology about their data needs, and discussions with software vendors who will be working to support the NMR community. The BMRB could do a better job of curating the existing RNA data. Bruce Johnson has developed his own tools for screening the RNA data and identifying outliers. These activities should be conducted within the BMRB, and the refined data made available to the public.

We received a report on intrinsically disordered proteins (**IDPs**), which is a good sign that the BMRB is anticipating one of the other areas of growth. However, the report sounded to the committee more like an investigator preparing for a research-focused R01 application, rather than an effort to begin to align the BMRB with this area. There is a very large body of literature and a large community of investigators who are working on IDPs, and it is a major thrust area of Bruker developments. Again, we think that the appropriate response of the BMRB is to take a leadership role in reaching out to the community and major stakeholders to learn about their needs and to better understand tools and resources that are missing now.

The BMRB has had a long and important role in **metabolomics**, most notably with the database of standards that they provide to the community. However, the staff seemed unaware of the use of these to the wider community. Who is using the resources? How are they using them? What do they need to improve their research? There seemed to be a lack of efforts to integrate with the resources and metabolomics centers that have been created via funding through the NIH common funds. This apparent “disconnect” is worrisome as it risks to endanger BMRB’s efforts to seek future funding to support the BMRB and to expand its role in the metabolomics research field. It would help to reach out to these people regularly to ask how these resources could be improved. The activities presented sounded much more like an interesting individual investigator project than a focus on building infrastructure for the community.

Advisory committee members did not hear anything about the emerging field of **biological ssNMR**, and its increasing role in solving significant biological problems. This is an area of exciting growth, and several groups are working on important biomedical applications. Like the other areas outlined above, the BMRB is encouraged to reach out to major stakeholders and conduct surveys and workshops to provide the data and infrastructure needs of this community. The BMRB has been focusing on building a better database infrastructure for **protein dynamics**; we encourage continued BMRB leadership in this important area.

To summarize, the AC encourages the BMRB to expand beyond its core of solution protein NMR, where it has made a major impact in the community. This was evident in the large outpouring of concern by that community in the previous funding crisis. This field is mature and

slowing in its scientific contributions, and the BMRB would benefit from playing a more active role in defining itself for the emerging areas outlined above. This requires involving the scientific community well beyond UW-Madison, and the BMRB staff would need to spend much more time collecting and paying attention to citations and user data on their website. These priorities should be presented front and center in future AC meetings and in preparation for new proposals for funding.

We recommend that the BMRB leadership consider organizing and/or participating in focused workshops. This is happening with metabolomics this summer, and we recommend that one of the other emerging areas discussed above be featured in a 1-day workshop each year. In fact, members of the board unanimously and enthusiastically like the idea of a one-day, or even half-day, workshops associated with each of the upcoming BMRB advisory board meetings. For example, a short workshop with participation by Bruce Johnson or other members of the RNA community could be helpful for bringing the BMRB staff up to speed on emerging needs in this area, and would add real value to the advisory committee meetings (perhaps preceding next year's AC meeting?).

There are other ways of reaching the user community besides publications, and the BMRB staff should consider a series of webinars, user surveys, short videos, and other mechanisms to connect with users. Social media would be a good way to reach younger scientists who may not be aware of the resources available at the BMRB.

The workshops suggested above will also help the new BMRB Director become more familiar with the experimental NMR field in general. Dr. Romero is clearly an outstanding choice with regards to the computational side of the BMRB, but he does not have as much experience with the experimental side as Dr. Ulrich. The workshops will help in this transition and will help Dr. Romero define a vision for the future of the BMRB, which needs to come into focus next year to prepare it for a successful renewal of funding. We also suggest that Dr. Romero appoint an NMR expert as an Associate Director or close advisor during this transition.

We are concerned about the NMR Exchange Format (NEF). Our sense is that a group of well-intentioned but incompletely informed individuals is trying to recreate the NMR-STAR format, which has taken decades of work to develop. We are particularly worried that this is unfunded and that the burden will fall entirely with the BMRB to fix. AC members were not convinced during the presentation by the BMRB staff that this had any significant value. We strongly recommend that the BMRB leadership take an active role in leading the NEF discussions and evaluating whether or not such endeavors should be pursued. A great deal of individual communications and "lobbying" should happen between BMRB and major stake holders, especially the software developers. Before another large meeting or workshop, the BMRB staff should talk to every group individually to better understand the problem and to educate them about the risks and difficulties that will come with a poorly executed change of format. After the BMRB has defined the path forward, they should approach the NIH for supplemental funding to do the job properly.

The BMRB is well poised to benefit from the major new push from the NIH on reproducibility in science. This will continue and will permeate all areas of biomedical research. However, the BMRB seems to be missing a major opportunity to lead the NMR community in this area through outreach and the development of tools that will help investigators achieve the required standards. This is a clear point of synergy with the NMRBox P41 project, which can really help push this area forward. The AC would love to see the BMRB undertake a strong leadership role in this area.

Two committee members are scheduled to rotate off after this year, Valérie Copié and Mei Hong. We recommend replacements that focus on some of the emerging areas that we outlined here. It is important for the BMRB that the members are generally able to attend meetings and to participate actively.

PDBe SAC report 2016:

Recommendations to the wwPDB

D&A

We are very pleased to see the D&A implemented and running, but now there is a need to look at the long term planning. Specifically, the pilots are in place, the vision looks promising, what are the long term goals and consequences? We consider that the management could be lighter which would be important in terms of resources. We were concerned by the resource overhead for difficult structure depositions. This is an area where we expect growth as larger complexes are tackled. We observed that there is still room for efficiency improvements by reducing the number of iterations of deposition and annotation: before pressing the submit button, the depositors apparently do not see all the validation information that will be available later and will lead to revisions being required.

Recommendation to wwPDB: Try to reduce resource overhead on annotators by reducing number of iterations of D&A required. Primarily this might be achieved by depositors seeing more validation information during deposition. Consider if the structure deposited can be modified on the fly rather than starting a new deposition each time.

EM

As the other databases in structural biology become established, it is important to clarify these relationships and their dependencies. The question of formal relationships between EMDB and EMdatabank arose, and also their status with respect to the wwPDB e.g. as federated archives.

Recommendation to the wwPDB clarification of the relationship of the role of EMDB with respect to the wwPDB Current wwPDB agreements do not cover EM, We recommend that the relationship should be formalised, including a decision as to whether EMDB and EMdatabase are federated archives or integral.

As part of the D&A process, there should be alignment of deposition policies.

Recommendation for the wwPDB: Options for release of EM data should be similar to those offered to wwPDB depositors.

Recommendation for the wwPDB: Adapt the policy of mandatory EM map deposition with the corresponding EM atomic models in the wwPDB.

NMR

Great progress was made on the validation reports- they look very good.

NEF is an important community effort involving all major software developers. NEF is a prerequisite for the PDB to validate NMR structures against data or restraints derived from data.

BMRB has become involved in the development of NEF, which is a highly positive development. BMRB has allocated resources to the project of mapping NEF onto NMR Star.

Recommendation to wwPDB: Align the efforts of the NEF group, the wwPDB and NMR VTF.

SAS

This is looking good - could this be the model for dealing with small federated databases?

Recommendation to wwPDB: To work with both SAS archives to bring them in to the D&A system smoothly.

X-ray

In general, given the maturity of the field, the X-ray component of the wwPDB is in good shape. The validation tools available for X-ray structures have been well-received. Going forward, it will be important to make these tools as intuitive as possible for the different user communities.

Linking to new synchrotron data policies.

The ESRF in France, is proposing to associate a DOI for each dataset collected there, and similar policies are expected to come into force in other large scale facilities such as the Diamond Light Source, UK, ISIS, ILL. It should be possible to link to this during structure deposition. Still under discussion are the metadata that ESRF should be archiving to go with diffraction/scattering data. A good start is the information currently supplied in ispyB, should the wwPDB be mining this to add to the structure information? Or is it sufficient to have a link to the DOI with the data being available on publication/release? This could be a strategic area where the wwPDB take a lead in helping to define the required metadata and formats to enable it to be easily extracted. As the types of data in the large scale facilities are expanding to hybrid methods, this could be one aspect of a possible hybrid methods funding application.

Recommendation to wwPDB: Liaise with ESRF and other large scale facilities on their open access policies for collected data to negotiate on metadata and good archiving practices. This should streamline the use of these data for the wwPDB. Work with the software producers to recover metadata as well.

Recommendation to wwPDB: ensure that D&A can accommodate links to federated databases, including those that contain raw data.

General Comments

Recommendation to wwPDB: The VTFs are really important in having community involvement and direction. All the VTF need chasing up to ensure that their work becomes publically available, and initiatives integrated.

**RCSB Protein Data Bank Advisory Committee
Report of November 3, 2015 Annual Meeting
Rutgers University, New Brunswick, New Jersey**

Chair: Cynthia Wolberger

Membership: Paul Adams, R. Andrew Byrd, Wah Chiu (absent), Kirk Clark, Paul Craig, Roland L. Dunbrack, Jr., Thomas E. Ferrin, Catherine E. Peishoff, Sue Rhee, Andrej Sali (absent), Torsten Schwede, Jill Trewhella and Cynthia Wolberger

US Government Representatives: Peter McCartney (NSF representative, present for Skype discussion)

RCSB Leadership: Stephen Burley, Helen Berman

RCSB PDB AC E-mail Addresses:

cwolberg@jhmi.edu, PDAdams@LBL.gov, byrda@mail.nih.gov, wah@bcm.edu, kirk.clark@novartis.com, paul.craig@rit.edu, roland.dunbrack@fcc.edu, tef@cgl.ucsf.edu, catherine.E.Peishoff@gsk.com, srhee@carnegiescience.edu, sali@salilab.org, torsten.schwede@unibas.ch, j.trewhella@mmb.usyd.edu.au

US Government Agency Representative E-mail Addresses:

pmccartney@nsf.org

RCSB Leadership E-mail Addresses:

sburley@proteomics.rutgers.edu, berman@rcsb.rutgers.edu

Executive Summary

The Advisory Committee to the Research Collaboratory for Structural Bioinformatics (RCSB) - met in New Brunswick, New Jersey on 3rd November 2015 to consider management and enhancement of the Protein Data Bank (PDB).

Agenda items included

- (1) Responses to 2014 RCSB PDB AC Recommendations;
- (2) State of the PDB;
- (3) Update on Integrative and Hybrid Methods
- (4) Data In: Deposition and Annotation;
- (5) Data Out: Access and Exploration;
- (6) Education plan
- (7) The PDB-101 website;
- (8) Management issues;
- (9) Discussion with Dr. Peter McCartney, NSF; and
- (10) Matters arising.

The meeting was held in the Rutgers University Center for Integrative Proteomics and opened by Dr. Stephen Burley, who gave an overview of the past year's activities and current state of the RCSB PDB. Dr. Burley welcomed the new members of the Advisory Committee and outlined the new policy of appointing Members to 3-year renewable terms. Burley outlined the responses to the 2014 RCSB PDB AC Recommendations and updated the committee on progress towards

completing the next version of the Deposition and Annotation (D&A) tool in partnership with PDBe. A summary of recent activities was subsequently provided by Berman, Young, Westbrook, Rose, Prlić, Dutta and Goodsell.

The Committee felt that the RCSB PDB has done a superlative job in addressing the issues raised in the 2014 PDB AC report. The Committee praises the leadership of Drs. Burley and Berman, whose well-managed team is effectively meeting new challenges and has been highly successful for obtaining additional funding for targeted initiatives, as recommended. The Committee was very enthusiastic about the restructuring of the Outreach and Education plan, one of the recommendations in last year's report. The new Education plan is focused and leverages successful components of previous education efforts to achieve maximal impact. The Committee encourages the RCSB PDB to continue to monitor the impact and effectiveness of the new education plan.

Together with impressive gains in efficiencies thanks to the automated D&A tool, the RCSB PDB is in an excellent position to deal with increasing numbers of depositions and to meet the challenges of handling more complex depositions of structures determined by hybrid methods. Accompanying the dramatic reduction in turnaround for coordinate deposition is a marked increase in coordinate replacement, which could be an indication of improvement in the quality of the model in light of validation information provided during the deposition process. The Committee recommends investigating the reasons that users replace coordinates after the initial deposition and identify mechanisms that would encourage researchers to validate coordinates and data prior to beginning the deposition. As part of this effort, the Committee recommends making the validation pipeline software through a web-accessible interface as well as a standalone downloadable version. The Committee emphasizes the **critical importance of completing version 2.0 of the D&A tool, which will be essential to meeting future demands across all four wwPDB sites**. It is thus a matter of deep concern to the Committee that completion of D&A 2.0 has been delayed by over a year. The Committee very much hopes that the new management agreement and new deadlines agreed upon by the wwPDB collaboration will result in release of D&A 2.0 in early 2016.

The Committee endorses several proposals by RCSB PDB to improve the quality of deposited data and enhance the ability of users to connect structural data to information on biological function. These include plans to remediate carbohydrates, residual B factors and crystal orientation, as previously discussed, as well as a proposal to include visualization of ligand electron density. The Committee also supports the proposal to map structures to protein families and to biological pathways, which will be highly useful to the general user community. It might be useful to assess which resources have the most intuitive representation of biological pathways for the bulk of PDB users.

The Committee emphasizes once again the importance of securing stable, long-term funding for RCSB PDB to serve the needs of the scientific, medical, industrial and education communities. The Committee is grateful to the NIH, NSF and DOE for their long-standing support of the RCSB PDB, which has served as a model for managing "big data" and making it accessible to a broad and diverse community of users. The Committee was thus particularly gratified by comments from the NSF representative regarding their increased recognition of the value of long-term support for databases like the RCSB PDB.

Responses to 2014 RCSB PDB AC Recommendations

- PDBAC: Pursue funding to develop approaches for supporting data from integrative/hybrid methods
Response: Proposals submitted.
- PDBAC: Terminate the legacy deposition system (ADIT)
Response: ADIT retired July 2015 for x-ray crystal structures
- PDBAC: Continue to provide mobile-friendly services
Response: Redesign of Structure Summary and PDB-101 pages to respond to display type.
- PDBAC: Develop a focused Education Plan
Response: Comprehensive redesign; described below.
- PDBAC: Make more information available on unpublished structures
Response: Requires further discussion with wwPDB and community stakeholders.

PDB Metrics

In aggregate, 10364 depositions were processed between January 1st and December 31st 2014 with a two-week average turnaround, a decrease from the 10566 entries deposited in 2013. Based upon the number of entries deposited in 2015 to date, it is estimated that 11000 entries will be deposited in 2015.

Breakdown of depositions by discipline in calendar 2014 was as follows:

X-ray:	9586 (93% of entries deposited, down from 9697 in 2013)
NMR:	515 (5%, down from 590 in 2013)
EM:	240 (2%, up from 234 in 2013)
Other:	23 (0.3%, down from 45 in 2013)

Breakdown of depositions by wwPDB processing site in calendar 2014 was as follows:

RCSB PDB:	6040 (58%)
PDBj:	1779 (17%)
PDBe-EBI:	2545 (25%)

Breakdown of depositors by location in calendar 2014 was as follows:

North America	37%
Europe	33%
Asia	19%
Industry	7%
South America	<1%
Australasia	4%
Africa	<1%

During 2014, RCSB PDB's website at <http://rcsb.org> was visited each month by an average of 283,358 unique visitors and 668,348 unique visits. A total of 25.033 GB of data were accessed.

During the same time period, more than 505 million data files were downloaded from the PDB archive *via* the wwPDB member FTP and websites (RCSB PDB: 347,283,931; PDBe: 100,393,784; PDBj: 57,683,377).

2015 RCSB PDB AC Discussion

Integrative/Hybrid Methods

Dr. Helen Berman presented an overview of how the RCSB PDB is meeting the new challenges presented by deposition of structures determined by multiple experimental methods. Berman summarized the discussions held at the Hybrid Methods Task Force meeting at EMBL-EBI in Hinxton, UK in October 2014. The resulting set of recommendations, which were published in *Structure* in July 2015, identified issues regarding model and data archiving, structure representation, validation, and publication standards to be dealt with by all the wwPDB partners. In addition, a federation of model and data archives, including the newly-formed Small-Angle Scattering Biological Data Bank (SASBDB), will be established to handle depositions and create a single hybrid model repository. A Working Group led by Berman and Advisory Committee members Trehwella, Sali and Schwede are leading a Task Force and subgroups that are grappling with these issues and confer monthly to discuss progress and coordinate efforts. The Committee was gratified to hear that an NSF EAGER grant has been obtained to support some of these new efforts and that a new proposal on hybrid model validation has been submitted to the NSF. The Committee fully supports the RCSB PDB efforts in this critically important new area in structural biology and hopes that the necessary additional funding will be forthcoming.

Data In: Deposition, Annotation, and Quality Assessment

Dr. John Westbrook gave an overview of the activities of the curators and developers who manage data deposition and annotation. The team does an impressive job of curating data and developing tools for submission and curation, thanks to their breadth of expertise in x-ray crystallography, NMR, EM, small molecules, software and statistics.

Dr. Jasmine Young provided an update on depositions, which during 2015 transitioned to exclusive use of the Common Deposition & Annotation System (D&A), with phase-out of the older ADIT system over the period January – June 2015. The new D&A system has made possible an impressive increase in throughput, with approximately 50 entries per month processed by each full-time employee (FTE). This enabled the RCSB PDB to handle over 6,000 entries over the past year. These entries are of increasing complexity and size, which can now be handled efficiently, thanks to the adoption of the PDBx format. Young also updated the Committee on numerous improvements to biocuration, including annotation of chimeric protein sequences, improved ligand annotation and better workflow management, which is improving both the user experience and increasing curation efficiency. The Committee views both of these as mission-critical to the long-term ability of the PDB to serve both depositors and users, the latter of which are increasingly non-experts. The remarkable decrease in processing time, which has decreased from a median of 16.5 days with ADIT to 1.6 days with the new D&A tool, has, however, had unintended consequences, namely a large increase (~150%) in the rate at which some users replace coordinates each month, presumably in response to the results of validation reports. The Committee felt that, while improvements to structures are to be welcomed by the community, it will be important to reduce the replacement rate to ensure long-term productivity

and throughput. The Committee recommends that the RCSB PDB investigate the reasons for coordinate replacements, and experiment with incentivizing researchers to validate their data prior to depositing coordinates and to correct coordinate errors prior to deposition. As part of this effort, the Committee recommends that the RCSB make the PDB validation software available to users and developers via a web-accessible interface as well as for download, for those who wish to install a local version.

Dr. John Westbrook informed the Committee on the deployment of the D&A tool, done in partnership with the wwPDB and implemented over the period January 2014 to September 2015. The Committee was gratified to hear that a secondary site for disaster recovery was set up in April 2015 in partnership with the wwPDB and feels this was a critically important measure. Dr. Westbrook also updated the committee on further developments in data standards, guided by recommendations of the PDBx/mmCIF Working Group chaired by PDB AC member, Dr. Paul Adams. Changes to be implemented include the NMR Exchange Format (NEF) for restraint data and external references files (ERFs) such as links to the Cambridge Structural Database. Looking ahead to 2016, the Committee endorses the plan to remediate carbohydrates, residual B factors and space group settings, as previously discussed. The Committee also looks forward to completion of version 2.0 of the D&A tool, which is being developed in partnership with PDBe. The Committee is deeply concerned that more than one year has passed since the original completion deadline, because the outstanding NMR and 3DEM deposition functionalities have not been completed. The timely completion and implementation of D&A 2.0 is of paramount importance to the long-term ability of the wwPDB to meet its obligations to the larger community. The Committee thanks Dr. Jasmine Young for her willingness to help manage the project and looks forward to a rollout in early 2016.

Data Out: Data Access and Exploration

Dr. Peter Rose introduced the newly designed Structure Summary page, which had become cluttered and difficult to navigate as features were incrementally added. The Committee found the new design, whose look and feel was based on last year's redesign of the main PDB page, to be clean and user-friendly, making recently added features such the Protein Feature View and structure visualization easier to access and use. The Committee was pleased to hear that this feature has also been made accessible on mobile devices, which last year constituted 10% of web traffic and are increasingly being used to access the PDB. Dr. Andreas Prlić showed the Committee how to access mutation information in Protein Feature View as well as graphical summaries of structure validation and the web-based 3D structure viewer. These features are thoughtfully designed and easy to use. The importance of providing multiple tools accessible to naïve users was driven home by the fact 75% of RCSB PDB users are non-specialists. The Committee commends the RCSB PDB for its ongoing comprehensive use of web analytics to measure usage and analyze user demographics. These data are important for the ongoing ability of the PDB to meet the needs of its user community and will be critical for making their case to funding agencies. At the same time, the RCSB has also made wise use of user communications with the Help Desk to obtain feedback and identify areas for improvement. Looking ahead, the Committee endorses the RCSB PDB plans for further improvements, including visualization of ligand electron density. The Committee also supports the proposal to map structures to protein families and to biological pathways, which will be highly useful to the general user community.

Education and Outreach

The RCSB PDB has long carried out an impressive array of outreach and education activities. Last year, the Committee expressed the concern that these efforts needed to be more focused in order to stay within the current budget constraints while maximizing impact. The Committee was highly impressed by the outstanding new education and outreach plan presented by Dr. Shuchismita Dutta. The new plan, a comprehensive and thoughtful restructuring of education efforts, builds upon successful elements of previous efforts. The partnerships with educators to develop teaching materials and use of HIV/AIDS and diabetes, as frameworks to educate students at various levels about biomolecular structure, are all highly attractive elements. The overall plan for developing curriculum modules and then field-testing and assessing their impact is well thought-out and focused. While currently aimed at high school and college students, the plan to extend the reach to healthcare professionals and continuing medical education could further broaden the impact. As additional materials are developed, the Committee expects that the RCSB PDB will determine how these can best be publicized and made readily visible on the web site.

One of the most popular education and outreach resources on the RCSB PDB web site is PDB-101, which provides a variety of educational materials that are utilized by students and faculty alike. Dr. David Goodsell provided the Committee with a preview of the redesigned PDB-101 web interface, which addresses some shortcomings of the current version while making the site more user-friendly and easier to update. Improvements include the ability to search the popular Molecule of the Month pages, which were previously accessible through a menu only, as well as clearer and more intuitive menus and organization. Given the popularity of PDB-101, which accounted for an impressive 12% of RCSB PDB web traffic in the past year, the Committee expects that these changes will maximize the utility of these features and looks forward to the planned rollout at the end of the year.

Management

Dr. Stephen Burley provided a summary of the RCSB PDB organization, current funding and responses to NSF requirement for the current funding period. The Committee once again commends Drs. Burley and Berman for ensuring a seamless leadership transition last year and for continuing to work as an effective team with the help of Deputy Director Christine Zardecki. Burley is currently also directing the UCSD site but hopes to recruit a replacement for Dr. Phil Bourne, who left for the NIH last year. The RCSB PDB has had impressive success in obtaining grants for specific outreach and technology development projects, in addition to its core support from the NSF/NIH/DOE. As stipulated in the NSF requirements for the current 2014-2018 funding period, the RCSB PDB has developed a business model, diversity plan, and assessment plan and has revised the guidelines for membership of the advisory committee. The plan to appoint members to 3-year renewable terms will ensure turnover and strengthen the ability of the RCSB PDB to appoint members in new areas, as witnessed by the addition this year of members with expertise in hybrid methods, cryoEM and visualization, as well as representatives from industry. The Committee Chair has agreed to stay on through the next major grant renewal and will be replaced in 2019. To ensure sustainability in future years, the RCSB has been able to dramatically increase efficiency and rebalance 'Data In' tasks, thanks to the automated D&A tool. Plans to extend the wwPDB franchise to China and India will enable the PDB to meet expected increased demands from investigators throughout Asia.

Plans for financial support

The RCSB is currently on solid financial footing, thanks to success in obtaining grants for targeted projects. Continued success in this arena, together with plans to seek private and corporate funding, will be important for implementing plans for additional activities.

There was a discussion with Dr. Peter McCartney of the NSF, who participated via telephone. Dr. McCartney told the committee that the NSF views the RCSB as a major resource that has been able to maintain support and a positive profile at the NSF because of its well-defined scope. Dr. McCartney praised the RCSB PDB for maintaining this focus and recognizing “what it is and what it isn’t.” The Committee was very gratified to hear from Dr. McCartney that the NSF recognizes the value of providing long-term financial support to databases like the PDB. This is a shift from the previous funding philosophy, which considered NSF support to be seed funding, and is enthusiastically welcomed by the Committee. In the discussion following the telephone conversation, the Committee and the RCSB leadership agreed that it would be beneficial to include representatives from the NIH and DOE in next year’s conversation with the NSF, with an eye towards planning for what is likely to be a competing renewal in 2018.

Matters arising

The Committee was asked to provide input on a number of matters confronting the RCSB PDB. The PDB leadership solicited advice on a proposal to enable three-dimensional visualization of the structural impact of coding SNPs (single nucleotide polymorphisms) and other genetic variations. While the Committee agreed that this could, in principal, be of great utility to the broader biological community, particularly those lacking expertise in structural biology, the issue generated a wide ranging discussion of how such a plan would be implemented, what the focus would be, who would carry out the project and how it could be ensured that meaningful models would be generated. The Committee recommends exploring the issue with a pilot project, perhaps focusing on disease-causing mutations and to measure interest in the research and educational communities before broadening the project’s scope. The Committee endorsed the RCSB PDB’s proposal to develop resources for exploring protein families and pathways and recommends also making available precomputed structure superpositions for families of related proteins.

The Committee discussed at length the question of what the RCSB PDB, or the wwPDB, should do about errors that are detected by validation software, biocurators or users, but are not corrected by the author. The Committee endorses the proposal to put a comment section on the RCSB structure summary page where comments from annotators could be posted and responses from depositors could be solicited. The Committee also supports the proposal to identify whether there are journals whose reported structures have particularly elevated rates of problematic structures and consider engaging editors in the effort to increase author responsiveness to queries from curators.

Protein Data Bank Japan (PDBj) Advisory Committee (PDBj-AC)
Report of 5th February 2016 Meeting
PDBj, Institute for Protein Research (IPR), Osaka University, Osaka, Japan

Chair: Prof. Haruki Nakamura (IPR)

PDBj member: Prof. Toshimichi Fujiwara (IPR)

Committee members (present): Prof. Genji Kurisu (IPR), Prof. Tsuyoshi Inoue (Faculty of Engineering, Osaka University), Prof. Daisuke Kohda (Medical Institute of Bioregulation, Kyushu University), Prof. Toshiya Senda (Photon Factory, KEK)

Committee members (absent): Dr. Yoshitsugu Shiro (SPring-8, RIKEN), Prof. Kei Yura (Graduate School of Humanities and Science, Ochanomizu University)

The advisory committee of PDBj (PDBj-AC) met with PDBj leadership and staff, at the Institute for Protein Research (IPR), Osaka University on 5th February 2016.

1. Current PDBj program:

- The PDBj has been funded as the Database Integration Coordination Program from JST (Japan Science and Technology Agency) – NBDC (National Bioscience Database Center) for three years from April 2014 – March 2017.
- The JST-NBDC budget in the FY 2015 (from April 2015 to March 2016) is 50 MY (direct) and 15 MY (indirect), and other additional costs are paid from the University budget for IPR as the Joint Usage and Research Center for Proteins
- The Data-in activity of PDBj was reported with the statistics in the Asian countries: Japan, China, India and others.
- The first wwPDB/CCDC/D3R Ligand Validation Workshop and the 12th wwPDB AC meeting were reported.
- The development of D&A version 2 was reported, and participation of Haruki Nakamura and Naohiro Kobayashi to the wwPDB D&A Software Engineering summit, which will be held on 2-3 March 2016 is addressed.
- The current situation of PDBj-BMRB was reported.
- About the major outreaches of the PDBj, News letter Vol. 17 and the wwPDB Symposium on Integrative Structural Biology with Hybrid Methods were reported.

2. Discussion for the future of PDBj

- The JST-NBDC budget in the FY 2016 (from April 2016 to March 2017) will also be 50 MY (direct) and 15 MY (indirect), and other additional costs will be paid from the University budget for IPR as the Joint Usage and Research Center for Proteins.
- A contract was concluded about the Collaboration for Life Science Databases among PDBj, DDBJ, and DBCLS (Database Center for Life Science) on 6th August 2015, and the contract will be renewed in the next year.
- It is still unclear for the new Database policy of JST, because the president and the organization system of JST changed in October 2015. Therefore, nobody knows the funding situation after March 2017. The members of PDBj-AC concern the budget situation of the PDBj, and they will continuously support the PDBj activities.
- Profs. Genji Kurisu and Tsuyoshi Inoue will attend the wwPDB AC meeting, which will be held at BMRB, University of Wisconsin-Madison on 7th October 2016.



The archiving and dissemination of biological structure data

Helen M Berman¹, Stephen K Burley^{1,2}, Gerard J Kleywegt³,
John L Markley⁴, Haruki Nakamura⁵ and Sameer Velankar³



The global Protein Data Bank (PDB) was the first open-access digital archive in biology. The history and evolution of the PDB are described, together with the ways in which molecular structural biology data and information are collected, curated, validated, archived, and disseminated by the members of the Worldwide Protein Data Bank organization (wwPDB; <http://wwpdb.org>). Particular emphasis is placed on the role of community in establishing the standards and policies by which the PDB archive is managed day-to-day.

Addresses

¹ Research Collaboratory for Structural Bioinformatics Protein Data Bank, Department of Chemistry and Chemical Biology, Center for Integrative Proteomics Research, Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, 174 Frelinghuysen Road, Piscataway, NJ 08854, USA

² Research Collaboratory for Structural Bioinformatics Protein Data Bank, Skaggs School of Pharmacy and Pharmaceutical Sciences and San Diego Supercomputer Center, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

³ Protein Data Bank in Europe, European Molecular Biology Laboratory – European Bioinformatics Institute, Wellcome Genome Campus, Cambridge CB10 1SD, UK

⁴ Biological Magnetic Resonance Bank, Department of Biochemistry, University of Wisconsin-Madison, Madison, WI 53706, USA

⁵ Protein Data Bank Japan, Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka, 565-0871, Japan

Corresponding author: Berman, Helen M (berman@rcsb.rutgers.edu)

structure of hemoglobin [4,5]. Both won Nobel prizes for their achievements. Not long after these structures were published, the crystallographic community began discussions as to how to best archive these data and make them available. During this period, there were numerous grass-roots meetings, one of which resulted in a petition, and many exchanges of handwritten documents. In 1971, the Cold Spring Harbor Laboratory hosted a symposium on protein crystallography, during which leaders in the field presented their seminal work [6]. Walter Hamilton, an attendee, offered to provide the first home for what is now known as the Protein Data Bank (PDB) [7]. The PDB was launched at Brookhaven National Laboratory, on the basis of the Protein Structure Library created by Edgar Meyer [8]. The initial PDB archive contained fewer than ten structures, all of which were determined by X-ray crystallography. In the 1980s, structures determined using NMR methods began to be deposited, and in 1990 the first structure determined by electron microscopy was deposited. In 1982 the PDB reached 100 entries, in 1993 1000 entries, in 1999 10 000, and in 2014 100 000 entries. At the time of writing, the PDB archive contains over 117 000 structures of proteins, nucleic acids, and their complexes with one another and with small molecule ligands.

Current Opinion in Structural Biology 2016, 40:17–22

This review comes from a themed issue on **Biophysical and molecular biological methods**

Edited by **Petra Fromme** and **Andrej Sali**

<http://dx.doi.org/10.1016/j.sbi.2016.06.018>

0959-440/© 2016 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

Historical background

Structural biology is a relatively young science that can trace its roots to the first X-ray diffraction studies of pepsin in 1935 by Dorothy Crowfoot (Hodgkin), who at the time was a student of J.D. Bernal [1]. Twenty years later, Kendrew determined the structure of myoglobin [2,3]; shortly thereafter, Perutz determined the

The PDB as a community data resource

From its inception, the PDB has been a community effort that has evolved with changes in scientific culture. For example, when the PDB was first created, data submission was voluntary. However, in the 1980s, members of the community became outspoken about the need to enforce mandatory data deposition. Various committees were set up to define what data should be required and when to disseminate the data. These guidelines were published in 1989, and over time, adopted by virtually all of the scientific journals that now require PDB deposition(s) as a prerequisite for publication of structural studies [9]. In 2008, further shifts in community sentiment led to mandatory deposition of experimental data together with atomic coordinates. In the current decade, the importance of reproducibility has been highlighted. The PDB convened method-specific Validation Task Forces and Workshops [10,11,12,13] to define what data should be collected and how best to validate the structural models, the experimental data, and the fit of the models to the data. Now every structure in the PDB comes with a publicly available validation report, and

authors are strongly encouraged to include these reports with their manuscript submissions to journals.

The importance of global participation in data archiving was understood early in the creation of the PDB. Indeed, the announcement of the PDB in 1971 described the collaboration with the Cambridge Crystallographic Database Centre [7]. In 2003, a Memorandum of Understanding (MOU) among partners in the US (RCSB Protein Data Bank; <http://www.rcsb.org>), Japan (Protein Data Bank Japan or PDBj; <http://www.pdbj.org>), and Europe (Protein Data Bank in Europe or PDBe; <http://pdbe.org>) established the Worldwide Protein Data Bank (wwPDB) partnership, which is responsible for formalizing the procedures involved in collecting, standardizing, annotating and disseminating the data [14^{*}]. Subsequently, a global NMR specialist data repository BioMagResBank, composed of deposition sites in the US (BMRB; <http://www.bmrb.wisc.edu>) and Japan (PDBj-BMRB; <http://bmrbdep.pdbj.org>), joined the wwPDB.

The X-ray crystallography community has led the biological sciences in the area of data sharing. While the sociological/anthropological underpinnings of this leadership role have not been fully explored, much of what has transpired in the creation and evolution of the PDB can be traced to J.D. Bernal, who, in addition to being a brilliant scientific innovator, was a prominent social activist, whose beliefs were consistent with the conduct of the PDB [15].

Content of the PDB archive

The PDB archive contains information about structural models that have been derived from experimental methods, including X-ray/neutron/electron crystallography, NMR spectroscopy, and 3D electron microscopy (3DEM). In addition to the 3D coordinates, the details of the chemistry of the polymers and small molecules are archived, as are metadata describing the experimental conditions, data-processing statistics and structural features such as the secondary and quaternary structure. The structure-factor amplitudes (or intensities) used to determine X-ray structures, and chemical shifts and restraints used in determining NMR structures are also archived. The electron density maps used to derive 3DEM models are archived in EMDB [16^{*}], and the experimental data underpinning them can be archived in EMPIAR [17]. In collaboration with community experts, pertinent data items are defined for each experimental field, with requirements evolving over time. The PDB data dictionary, originally developed to describe macromolecular crystallography, contains more than 4400 data items. The dictionary combines data items common to all methods as well as those that are method specific. For example, the current dictionary contains 250 NMR-specific data and 1200 3DEM-specific data definitions.

Over time, the holdings of the PDB have increased dramatically as has the complexity of the structures being archived (Figure 1).

A workshop held in 2005 led to the policy that purely *in silico* models should not be part of the PDB [18^{**}], and, instead, a modeling portal should be created for these models. The Protein Modeling Portal was established in 2007 [19].

Representation of PDB data

The first data format used by the PDB was established in the early 1970s and was on the basis of the 80-column Hollerith format used for punched cards. The atom records included atom name, residue name and sequence number. A 'header record' contained some metadata. This format was readily accepted because it was simple and both human- and machine-readable. However, it had many serious drawbacks in that the size of the structural models was limited to 99 999 atoms and that relationships among the data items were implicit. These inherent weaknesses meant that significant domain knowledge was necessary in order to write software using this format.

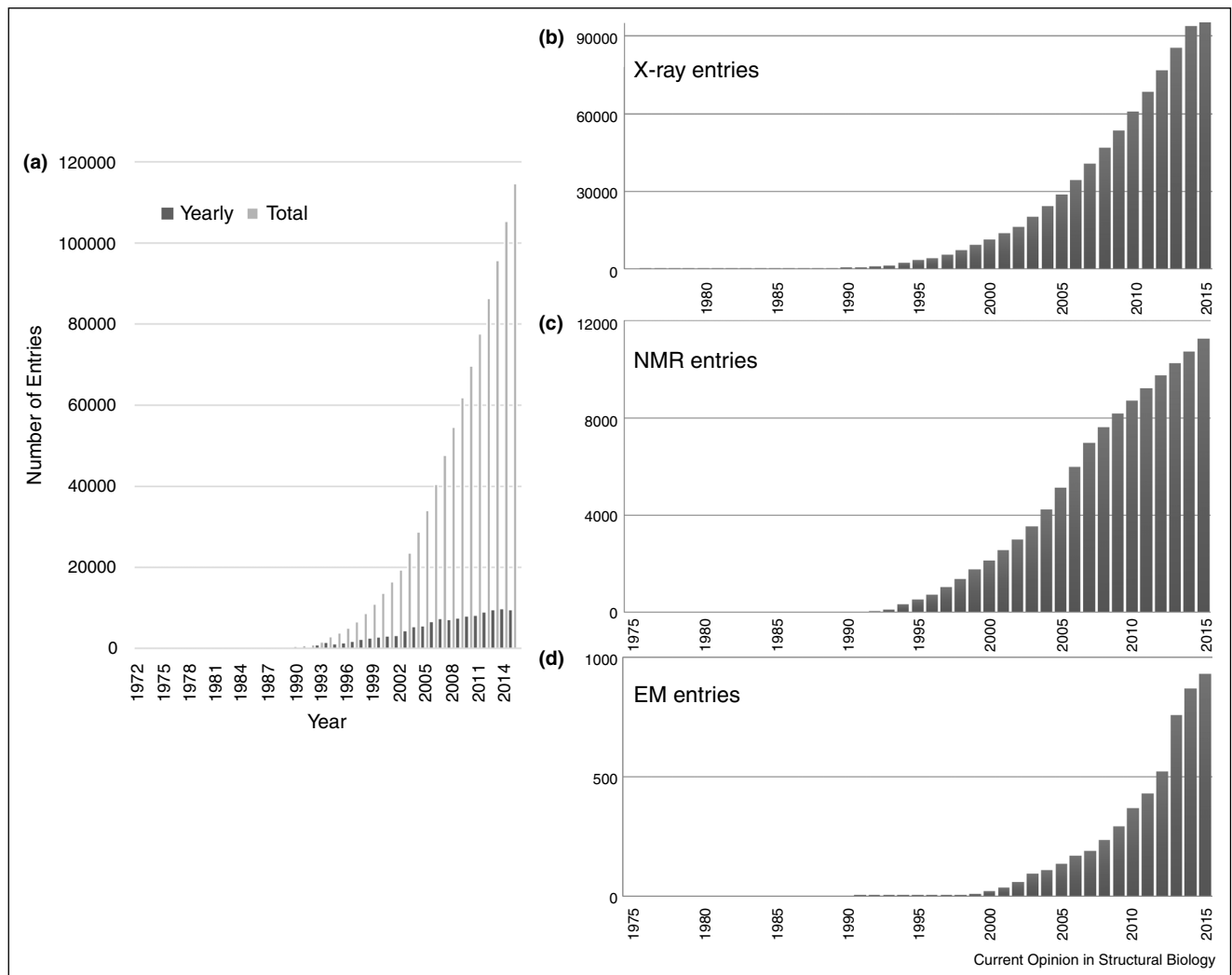
In the 1990s, the IUCr chartered a committee to create a more formal data model. This committee proposed the Macromolecular Crystallographic Information File (mmCIF) [20^{*}]. mmCIF is a self-defining format in which every data item has attributes describing its features including relationships to other data items. Most importantly, mmCIF has no limitations with respect to the size of the archived structural model. The dictionary and the data files are completely machine-readable, and no domain knowledge is required to read the files. The first dictionary contained over 3000 data items relevant to X-ray crystallography. Over time, terms specific to NMR and 3DEM were added, and the dictionary was renamed PDBx/mmCIF. In 2007, it was decided that PDBx would be the Master Format for data collected by the PDB. In 2011, major X-ray structure determination software developers agreed to adopt this data model so that all output from their programs would be in PDBx. In 2015, large structures archived in the PDB that had formerly been split into multiple entries were combined into single entries and mmCIF formatted files. Other structural biology communities are in the process of building on the PDBx/mmCIF framework to establish their own controlled vocabulary and specialist data items [19,21].

PDBML, an XML format on the basis of PDBx/mmCIF [22], and its RDF (Resource Description Framework) conversion were developed to facilitate the integration of structure data with other life sciences data resources could be facilitated [23^{*}].

The data pipeline

Every data resource has a set of procedures for deposition, curation, validation, archiving and dissemination of data.

Figure 1



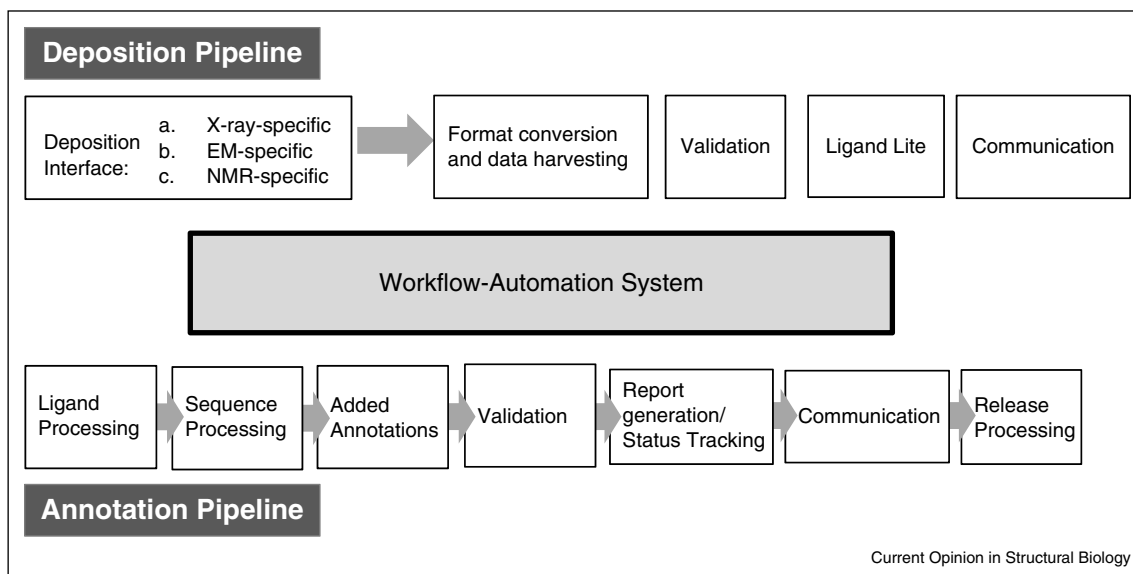
Growth of the PDB archive. **(a)** Number of entries deposited annually (dark gray) and available at the end of each year (light gray); **(b)** number of X-ray crystal structures; **(c)** NMR structures, and **(d)** 3DEM structures available each year.

The pipeline currently used by the wwPDB to populate the PDB archive is illustrated schematically in Figure 2.

In the very early days of the PDB, structures were deposited to BNL on magnetic tapes containing atomic coordinates with paper forms listing other data items, all sent first by mail and then via a web-based system, called AutoDep, was created in the 1990s [24]. This system was later modified and used by PDBe [25] until very recently. The RCSB PDB and PDBj collected data using a system on the basis of mmCIF called ADIT [26^{*}], and the BMRB in the US and its affiliate in Japan adopted a similar system called ADIT-NMR [27^{*}]. Although these systems were distinct, since 2003, the wwPDB partners have determined jointly what data should be collected and which procedures and algorithms should be used for data processing. In 2007, it was agreed within the wwPDB to

create a single deposition, Structures are made available to the public either immediately after they have been fully curated or -in most cases- when they are published in a journal. Usually, either the author or the journal informs wwPDB that the paper describing the structure is about to be published. PDB data are released in a two-stage process. Every Saturday at 03:00 UTC the polymer sequences, ligand SMILES strings, and crystallization pH for new structures designated for release are made public (<http://wwpdb.org/download/downloads>) as a courtesy to the protein structure modeling and computational chemistry communities to enable weekly blinded prediction challenge efforts (e.g., CAMEO [19] and D3R CELPP [28]). Every Wednesday at 00:00 UTC, all new structures designated for release are made publicly available through the wwPDB FTP sites. On average about 200 structures are released every week. As evidence

Figure 2



wwPDB Deposition, Annotation, and Validation pipeline. Each box represents a modular component of the data processing workflow.

for the importance of this archive, in 2015, more than 500 million sets of atomic coordinates were downloaded from the wwPDB FTP sites.

Value-added resources

The wwPDB FTP sites provide the core data for many databases, services, and websites, including those run by the individual wwPDB partners. In the original wwPDB MOU, it was agreed that to best serve science, wwPDB partner websites would compete with one another and would offer many different kinds of services and features. The RCSB PDB has extensive search and reporting capabilities as well as an education portal called PDB-101 [26*,29]. PDBe has multiple search and browse facilities as well as analysis and bioinformatics tools [30,31*]. PDBj provides a variety of services and viewers and supports browsing in multiple Asian languages [23*,32]. BMRB has many capabilities designed to serve the NMR community [33].

CATH [34] and SCOP [35,36] use the data in the PDB to classify the structural domains of proteins with an attempt to relate them to function. More recently, these two databases have agreed to work together and with other resources in the UK to provide predicted structural features under a unified system called Genome3D [37].

Additional specialty databases provide information on particular classes of macromolecules such as nucleic acids [38].

The Protein Structure Initiative (PSI) Structural Biology Knowledgebase (SBKB) [39] was an ambitious effort to

unify information about protein sequence, structure and function. Unfortunately, the decision to discontinue funding the PSI means that this resource will cease to exist.

Challenges going forward

A review of the holdings of the PDB shows a steady growth (~10,000 new structures annually). More significantly, the complexity of the structural models continues to increase with more and more large heterogeneous assemblies entering the archive. Fortunately, there are no longer technical restrictions to receiving, annotating, validating, and disseminating these very large structures.

Historically, most structures were determined exclusively with the aid of a single experimental method: X-ray crystallography, NMR or 3DEM. In recent years, these traditional techniques are being combined with other methods to yield improved models. For example, it is now common practice to add data from small-angle scattering measurements to NMR-derived restraints to determine solution structures [40,41]. Similarly, NMR or X-ray data can be combined with cryoEM data in integrative modeling approaches [42]. Such integrative methods make it possible to combine data from different biophysical techniques with computational methods to create models of very large macromolecular machines [43]. However, hybrid approaches also present a variety of challenges including how to validate these structures and then how to archive them. As in the past, with the help and advice of an expert Task Force [44**], this integrative challenge will be met by the wwPDB partners.

Acknowledgements

RCSB PDB is supported by NSF [DBI-1338415], NIH, DOE; PDBe by EMBL-EBI, Wellcome Trust [104948], BBSRC [BB/J007471/1, BB/K016970/1, BB/K020013/1, BB/M013146/1, BB/M011674/1, BB/M020347/1, BB/M020428/1], NIGMS [1RO1 GM079429-01A1], EU [284209,675858] and MRC [MR/L007835/1]; PDBj by JST-NBDC and BMRB by NIGMS [1R01 GM109046].

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest
- of outstanding interest

1. Bernal JD, Crowfoot DM: **X-ray photographs of crystalline pepsin.** *Nature* 1934, **133**:794-795.
2. Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC: **A three-dimensional model of the myoglobin molecule obtained by X-ray analysis.** *Nature* 1958, **181**:662-666.
3. Kendrew JC, Dickerson RE, Strandberg BE, Hart RG, Davies DR, Phillips DC, Shore VC: **Structure of myoglobin: a three-dimensional Fourier synthesis at 2 Å resolution.** *Nature* 1960, **185**:422-427.
4. Perutz MF, Rossmann MG, Cullis AF, Muirhead H, Will G, North ACT: **Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5 Å resolution, obtained by X-ray analysis.** *Nature* 1960, **185**:416-422.
5. Bolton W, Perutz MF: **Three dimensional fourier synthesis of horse deoxyhaemoglobin at 2.8 Ångstrom units resolution.** *Nature* 1970, **228**:551-552.
6. **Cold Spring Harbor Symposia on Quantitative Biology.** Cold Spring Laboratory Press; 1972.
This seminal meeting highlighted the key structures that had been determined and brought together the leading figures in structural biology. To quote David Phillips it was a 'Coming of age.'
7. Protein Data Bank: **Protein Data Bank.** *Nature New Biol.* 1971, **233**:223.
8. Meyer EF Jr, Morimoto CN, Villarreal J, Berman HM, Carrell HL, Stodola RK, Koetzle TF, Andrews LC, Bernstein FC, Bernstein HJ *et al.*: **CRYSTNET, a crystallographic computing network with interactive graphics display.** *Fed Proc* 1974, **33**:2402-2405.
9. International Union of Crystallography: **Policy on publication and the deposition of data from crystallographic studies of biological macromolecules.** *Acta Cryst* 1989, **A45**:658.
10. Read RJ, Adams PD, Arendall WB 3rd, Brunger AT, Emsley P, Joosten RP, Kleywegt GJ, Krissinel EB, Lutheke T, Otwinowski Z *et al.*: **A new generation of crystallographic validation tools for the protein data bank.** *Structure* 2011, **19**:1395-1412.
This paper contained a thorough analysis of methods that could be used for validation of structures determined by X-ray crystallography. The recommendations were used to create the validation tools used by the wwPDB.
11. Henderson R, Sali A, Baker ML, Carragher B, Devkota B, Downing KH, Egelman EH, Feng Z, Frank J, Grigorieff N *et al.*: **Outcome of the first electron microscopy validation task force meeting.** *Structure* 2012, **20**:205-214.
This paper was the first one to analyze what is needed to validate 3DEM maps and models. It is the basis for current research in this field.
12. Montelione GT, Nilges M, Bax A, Guntert P, Herrmann T, Richardson JS, Schwieters CD, Vranken WF, Vuister GW, Wishart DS *et al.*: **Recommendations of the wwPDB NMR validation task force.** *Structure* 2013, **21**:1563-1570.
This paper made recommendations for how to validate structures determined by NMR, and is the basis for current ongoing research.
13. Adams PD, Aertgeerts K, Bauer C, Bell JA, Berman HM, Bhat TN, Blaney JM, Bolton E, Bricogne G, Brown D *et al.*: **Outcome of the first wwPDB/CCDC/D3R ligand validation workshop.** *Structure* 2016, **24**:502-508.
The criteria for judging the quality of ligands in protein complexes are laid out and will form the basis for improved validation of these molecules.
14. Berman HM, Henrick K, Nakamura H: **Announcing the worldwide Protein Data Bank.** *Nat Struct Biol* 2003, **10**:980.
This is the formal announcement for how the Protein Data Bank will be managed by an international consortium.
15. Brown A, Bernal JD: *The Sage of Science.* Oxford: Oxford University Press; 2005.
16. Lawson CL, Patwardhan A, Baker ML, Hryc C, Garcia ES, Hudson BP, Lagerstedt I, Ludtke SJ, Pintilie G, Sala R *et al.*: **EMDataBank unified data resource for 3DEM.** *Nucleic Acids Res* 2016, **44**:D396-D403.
This paper describes the procedures for streamlining the deposition and distribution of 3DEM maps and models.
17. Iudin A, Korir PK, Salavert-Torres J, Kleywegt GJ, Patwardhan A: **EMPIAR: a public archive for raw electron microscopy image data.** *Nat Methods* 2016:13.
18. Berman HM, Burley SK, Chiu W, Sali A, Adzhubei A, Bourne PE, Bryant SH, Dunbrack RL Jr, Fidelis K, Frank J *et al.*: **Outcome of a workshop on archiving structural models of biological macromolecules.** *Structure* 2006, **14**:1211-1217.
The recommendation to remove purely *in silico* models from the PDB is contained in this paper.
19. Haas J, Roth S, Arnold K, Kiefer F, Schmidt T, Bordoli L, Schwede T: **The Protein Model Portal – a comprehensive resource for protein structure and model information.** *Database (Oxford)* 2013, **2013**:bat031.
20. Fitzgerald PMD, Westbrook JD, Bourne PE, McMahon B, Watenpaugh KD, Berman HM: **4.5 Macromolecular dictionary (mmCIF).** In *International Tables for Crystallography G. Definition and Exchange of Crystallographic Data.* Edited by Hall SR, McMahon B. Springer; 2005:295-443.
A complete description of the mmCIF data dictionary is contained here.
21. Malfois M, Svergun DI: **sasCIF: An Extension of Core Crystallographic Information File for SAS.** *Journal of Applied Crystallography* 2000, **33**:812-816.
22. Westbrook J, Ito N, Nakamura H, Henrick K, Berman HM: **PDBML: the representation of archival macromolecular structure data in XML.** *Bioinformatics* 2005, **21**:988-992.
23. Kinjo AR, Suzuki H, Yamashita R, Ikegawa Y, Kudou T, Igarashi R, Kengaku Y, Cho H, Standley DM, Nakagawa A *et al.*: **Protein Data Bank Japan (PDBj): maintaining a structural data archive and resource description framework format.** *Nucleic Acids Res* 2012, **40**:D453-D460.
Descriptions of PDBj services are given here as well as the RDF format.
24. Lin D, Manning NO, Jiang J, Abola EE, Stampf D, Prilusky J, Sussman JL: **AutoDep: a web-based system for deposition and validation of macromolecular structural information.** *Acta Cryst* 2000, **D56**:828-841.
25. Tagari M, Tate J, Swaminathan GJ, Newman R, Naim A, Vranken W, Kapopoulou A, Hussain A, Filion J, Henrick K *et al.*: **E-MSD: improving data deposition and structure quality.** *Nucleic Acids Res* 2006, **34**:D287-D290.
26. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28**:235-242.
This is the first complete description of the services provided by the RCSB PDB
27. Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z *et al.*: **BioMagResBank.** *Nucleic Acids Res* 2008, **36**:D402-D408.
A summary of the services provided by BMRB is given here.
28. Drug Design Data Resource Community: Continuous Evaluation of Ligand Pose Prediction. <https://drugdesigndata.org/about/celp> (accessed 18.04.16).
29. Rose PW, Prlic A, Bi C, Bluhm WF, Christie CH, Dutta S, Green RK, Goodsell DS, Westbrook JD, Woo J *et al.*: **The RCSB Protein Data Bank: views of structural biology for basic and applied research and education.** *Nucleic Acids Res* 2015, **43**:D345-D356.
30. Velankar S, van Ginkel G, Alhroub Y, Battle GM, Berrisford JM, Conroy MJ, Dana JM, Gore SP, Gutmanas A, Haslam P *et al.*: **PDBe: improved accessibility of macromolecular structure**

- data from PDB and EMDB.** *Nucleic Acids Res* 2016, **44**:D385-D395.
31. Gutmanas A, Alhroub Y, Battle GM, Berrisford JM, Bochet E, Conroy MJ, Dana JM, Fernandez Montecelo MA, van Ginkel G, Gore SP *et al.*: **PDBe: Protein Data Bank in Europe.** *Nucleic Acids Res* 2014, **42**:D285-D291.
The services offered by PDBe are described.
 32. Kinjo AR, Nakamura H: **Composite structural motifs of binding sites for delineating biological functions of proteins.** *PLoS One* 2012, **7**:e31437.
 33. Markley JL, Ulrich EL, Berman HM, Henrick K, Nakamura H, Akutsu H: **BioMagResBank (BMRB) as a partner in the Worldwide Protein Data Bank (wwPDB): new policies affecting biomolecular NMR depositions.** *J Biomol NMR* 2008, **40**:153-155.
 34. Sillitoe I, Lewis TE, Cuff A, Das S, Ashford P, Dawson NL, Furnham N, Laskowski RA, Lee D, Lees JG *et al.*: **CATH: comprehensive structural and functional annotations for genome sequences.** *Nucleic Acids Res* 2015, **43**:D376-D381.
 35. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG: **SCOP database in 2004: refinements integrate structure and sequence family data.** *Nucleic Acids Res* 2004, **32**:D226-D229.
 36. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, Murzin AG: **Data growth and its impact on the SCOP database: new developments.** *Nucleic Acids Res* 2008, **36**:D419-D425.
 37. Lewis TE, Sillitoe I, Andreeva A, Blundell TL, Buchan DW, Chothia C, Cozzetto D, Dana JM, Filippis I, Gough J *et al.*: **Genome3D: exploiting structure to help users understand their sequences.** *Nucleic Acids Res* 2015, **43**:D382-D386.
 38. Berman HM, Olson WK, Beveridge DL, Westbrook JD, Gelbin A, Demeny T, Hsieh S-h, Srinivasan AR, Schneider B: **The Nucleic Acid Database – a comprehensive relational database of three-dimensional structures of nucleic acids.** *Biophys. J.* 1992, **63**:751-759.
 39. Gabanyi MJ, Adams PD, Arnold K, Bordoli L, Carter LG, Flippen-Andersen J, Gifford L, Haas J, Kouranov A, McLaughlin WA *et al.*: **The Structural Biology Knowledgebase: a portal to protein structures, sequences, functions, and methods.** *J Struct Funct Genom* 2011, **12**:45-54.
 40. Madl T, Gabel F, Sattler M: **NMR and small-angle scattering-based structural analysis of protein complexes in solution.** *J Struct Biol* 2011, **173**:472-482.
 41. Wang Z, Chernyshev A, Koehn EM, Manuel TD, Lesley SA, Kohen A: **Oxidase activity of a flavin-dependent thymidylate synthase.** *FEBS J* 2009, **276**:2801-2810.
 42. Byeon IJ, Louis JM, Gronenborn AM: **A captured folding intermediate involved in dimerization and domain-swapping of GB1.** *J Mol Biol* 2004, **340**:615-625.
 43. Ward AB, Sali A, Wilson IA: **Biochemistry. Integrative structural biology.** *Science* 2013, **339**:913-915.
 44. Sali A, Berman HM, Schwede T, Trewhella J, Kleywegt G, Burley SK, Markley J, Nakamura H, Adams P, Bonvin AM *et al.*: **Outcome of the first wwPDB hybrid/integrative methods task force workshop.** *Structure* 2015, **23**:1156-1167.
- This paper summarized the steps necessary to establish an archive of structure data from hybrid/integrative methods.

**To be published in
“Macromolecular Crystallography”**

Volume in Methods in Molecular Biology (Springer).

Co-editors Alexander Wlodawer, Zbigniew Dauter, and Mariusz Jaskolski

**Protein Data Bank - the single global macromolecular structure archive managed
by the Worldwide Protein Data Bank**

Stephen K. Burley^{1,2,3}, Helen M. Berman¹, Gerard J. Kleywegt⁴, John L. Markley⁵,
Haruki Nakamura⁶, and Sameer Velankar⁴

Author Affiliations

¹ 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

² Rutgers Cancer Institute of New Jersey, Robert Wood Johnson Medical School, New Brunswick, NJ 08854, USA

³ Skaggs School of Pharmacy and Pharmaceutical Sciences and San Diego Supercomputer Center, University of California, San Diego, La Jolla, CA 92093, USA

⁴ Protein Data Bank in Europe, European Molecular Biology Laboratory–European Bioinformatics Institute, Wellcome Genome Campus, Cambridge CB10 1SD, UK

⁵ BioMagResBank, Department of Biochemistry, University of Wisconsin-Madison, Madison, WI 53706-1544, USA

⁶ Protein Data Bank Japan, Institute for Protein Research, Osaka University, Suita, Osaka 565-0871, Japan

Running Title: PDB Archive and the wwPDB

Keywords: Protein Data Bank, PDB, Worldwide Protein Data Bank, wwPDB, PDBx/mmCIF, Chemical Component Dictionary, Crystallography, NMR Spectroscopy, NMR-STAR, NMR Exchange Format, NEF, 3D Electron Microscopy, Integrative or Hybrid Methods

Summary

The Protein Data Bank (PDB)—the single global repository of experimentally determined 3D structures of biological macromolecules and their complexes—was established in 1971, becoming the first open-access digital resource in the biological sciences. The PDB archive currently includes ~119,000 entries (May 2016). It is managed by the Worldwide Protein Data Bank organization (wwPDB; wwpdb.org), which includes 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.bmrwisc.edu). The four wwPDB partners operate a unified global software system that enforces community-agreed data standards and supports data deposition, annotation, and validation of ~10,000 new PDB entries annually (deposit.wwpdb.org). The RCSB PDB currently acts as the archive keeper, ensuring disaster recovery of PDB data and coordinating weekly updates. wwPDB partners disseminate the same archival data from multiple FTP sites, while operating competing websites that provide their own views of PDB data with selected value-added information and links to related data resources. At present, the PDB archives experimental data, associated metadata, and 3D-atomic level structural models derived from three well-established methods: crystallography, nuclear magnetic resonance spectroscopy (NMR), and electron microscopy (3DEM). wwPDB partners are working closely with experts in related experimental areas (small-angle scattering, chemical cross linking/mass spectrometry, Forster energy resonance transfer or FRET, etc.) to establish a federation of data resources that will support sustainable archiving of 3D structural models and experimental data derived from integrative or hybrid methods.

Evolution of Data Sharing and Data Archiving in Structural Biology

The Protein Data Bank (PDB) was established in 1971 with fewer than ten X-ray crystallographic structures of proteins, becoming the first open access digital resource in the biological sciences (Protein Data Bank 1971). Soon after X-ray structures of myoglobin (Kendrew et al. 1958; Kendrew et al. 1960) and hemoglobin (Perutz et al. 1960; Bolton and Perutz 1970) were published, the structural biology community began discussions as to how best to archive protein crystallographic findings and make them broadly available. In 1971, the Cold Spring Harbor Laboratory hosted a symposium on protein crystallography, during which there was extensive discussion of data sharing (Cold Spring Harbor Symposia on Quantitative Biology 1972). Walter C. Hamilton, one of the attendees, offered to provide the first home for what is now the Protein Data Bank (PDB) (Berman 2008). Shortly thereafter, the PDB was launched from within the Department of Chemistry at Brookhaven National Laboratory (BNL), building on the Protein Structure Library framework (Meyer 1997). The importance of making scientific data archiving a global endeavor was understood at the outset, and public announcement of the PDB in 1971 explicitly mentioned collaboration with and the option of data submission to the Cambridge Crystallographic Database Centre (Protein Data Bank 1971).

When the PDB was launched, data submission was voluntary. In the 1980s, influential members of the structural biology community began to make the case for mandatory data deposition. Various committees were established to define what data should be required and when it should be disseminated. Guidelines were published in 1989 (International Union of Crystallography 1989), and over time, adopted by virtually all of the scientific journals now requiring PDB deposition(s) prior to publication of structural studies. In 2008, further evolution of community mores led to mandatory deposition of crystallographic structure factors and NMR restraints together with atomic coordinates. In 2010, deposition of NMR chemical shifts became mandatory. At the time of writing (May 2016), ~80% of PDB archival entries are accompanied by experimental data.

Growth of the Protein Data Bank Archive

The first 356 structures deposited to the PDB archive were determined by crystallography. In 1988, structures determined using NMR methods began to be deposited, and in 1996 the first structure determined by electron microscopy was deposited. Since 1971, growth of the archive has been

decidedly non-linear (Figure 1). By 1982, the PDB had reached only ~100 entries. Eleven years later, in 1993, there were 1,000 entries. Before the end of the decade (1999), this number had grown to 10,000. Fifteen years thereafter, archival contents exceeded 100,000 entries as of May 2014. At the time of writing (May 2016), the PDB archive contains more than 119,000 structures of proteins, nucleic acids, and their complexes with one another and with small molecule ligands. Calendar year depositions in 2015 numbered 10,956 (~900/month). The vast majority of the PDB archival entries come from X-ray, neutron, and combined X-ray/neutron crystallography (~90%), with the remainder produced by NMR (~9%) and 3DEM (~1%). Among the three experimental methods currently represented in the PDB archive, data deposition rates have varied markedly over time. From 2012 to 2015, annual crystallographic depositions have grown slowly year-on-year [9,269 in 2012; 10,168 in 2015] During that same period, 3DEM depositions continued to increase significantly year on year, rising from 103/year in 2012 to 254/year in 2015. NMR depositions, on the other hand, peaked in 2007 at 1,062/year, declining to 510/year in 2015. The PDB archive has also grown considerably in complexity since 1971. Proxy measures of complexity are provided in Table 1.

History and Role of the Worldwide Protein Data Bank

Prior to 1999, the PDB was headquartered at BNL, which acted as the sole global deposition site. Macromolecular structure data were then distributed internationally from BNL by authorized PDB mirror sites located in various countries, including Argentina, Australia, Brazil, China, France, Germany, India, Israel, Japan, Poland, and the United Kingdom (Sussman et al. 1998). Following an open re-competition for US federal funding of the PDB in 1998, responsibility for the archive was reassigned to the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB), which was headquartered at Rutgers, The State University of New Jersey with additional performance sites at the San Diego Supercomputer Center at UC San Diego and the National Institutes of Standards and Technology (Berman et al. 2000). Following a transition period that witnessed formalization of Protein Data Bank Japan (PDBj) (Standley et al. 2008) and the Macromolecular Structure Database (MSD) (Keller et al. 1998; Velankar et al. 2016), RCSB PDB, PDBj, and MSD came together in 2003 to establish the Worldwide Protein Data Bank (wwPDB; wwpdb.org) (Berman et al. 2003). In 2006, a global NMR data repository BioMagResBank (BMRB), founded in 1989 (Ulrich et al. 1989), joined the wwPDB organization (Markley et al. 2008). BMRB hosts deposition sites in both the US (BMRB; www.bmrwisc.edu) and Japan (PDBj-BMRB; bmrdep.pdbj.org) (Ulrich et al. 2008). [N.B.: MSD was rebranded in 2008 as the Protein Data Bank in Europe or PDBe (Velankar et al. 2010; Velankar et al. 2016).]

Current wwPDB activities are governed by a Memorandum of Understanding (wwpdb.org/about/agreement), which was renewed in 2013. As outlined in detail below, wwPDB partners collaborate on “Data In”. They are jointly responsible for standardizing, collecting, annotating, and disseminating macromolecular structure data as a single global archive. At present, RCSB PDB is formally designated as the Archive Keeper, responsible for ensuring disaster recovery of PDB data and coordinating weekly archival updates among the partner sites (or regional data centers).

Founding of the wwPDB organization helped to ensure that the PDB has continued to evolve as the single global archive of macromolecular structure data. In contrast, global archiving of nucleic acid sequences is accomplished by three independently operated regional archives comprising

the International Nucleotide Sequence Database Collaboration (INSDC), which exchange data nightly.

PDB Data Standardization, Deposition, Annotation, and Validation

Following launch of the wwPDB, crystallographic structure depositions to the PDB archive were accepted *via* two different portals; ADIT, which was operated jointly by RCSB PDB and PDBj (Berman et al. 2000), and AutoDep, which was developed at BNL (Lin et al. 2000) and reengineered by MSD/PDBe (Tagari et al. 2006). NMR depositions were accepted *via* ADIT-NMR at BMRB and PDBj-BMRB, with coordinates and restraint data transferred to RCSB PDB or PDBj, respectively (Markley et al. 2008). In addition, PDBe accepted NMR structures *via* AutoDep, with associated NMR data sent to BMRB for archiving. Early in 2016, the wwPDB partners launched a unified global system for deposition, annotation, and validation of incoming data supporting crystallography, NMR, and 3DEM (deposit.wwpdb.org). Working to a common set of standards, three wwPDB regional data centers take responsibility for depositions originating from the Americas and Oceania (RCSB PDB), Europe and Africa (PDBe), and Asia (PDBj). The pipeline currently used by the wwPDB to process incoming structures is illustrated schematically in Figure 2. Approximately 900 depositions are received monthly from every inhabited continent (Figure 3). RCSB PDB, PDBe, and PDBj refer depositors of NMR data unrelated to 3D structures to BMRB, and, conversely, BMRB refers depositors with atomic coordinate data to the three wwPDB regional data centers. NMR data archived in the PDB are also mirrored in the BMRB archive under a four-digit acquisition code, which in some cases contains additional data on the system supplied by depositors (e.g., NMR relaxation rates, order parameters, and files containing raw time-domain data). Deposited entries are then validated and annotated by wwPDB biocurators, with wwPDB Validation Reports (wwpdb.org/validation/validation-reports) returned to depositors for review before finalization and data release.

Considerable effort has gone into understanding how best to standardize, annotate, and validate incoming atomic coordinates and primary experimental data generated by crystallography, NMR, and 3DEM. Over the past decade, the wwPDB has convened a series of expert, method-specific Validation Task Forces (VTFs) to determine which experimental data and metadata from each method should be archived and how these data and the atomic level structural models derived therefrom should be validated. Initially, the wwPDB X-ray VTF made recommendations on how

best to validate crystallographic data (Read et al. 2011). Preliminary recommendations have also been made by VTFs for NMR (Montelione et al. 2013) and 3DEM (Henderson et al. 2012). The work of these VTFs has enabled a sea change in the way PDB entries are validated at the time of deposition/annotation. A wwPDB Validation Report is produced for every new entry, and more and more journals require authors of structure determination studies to submit these reports together with their manuscripts.

The wwPDB has also convened a number of workshops to address both policy and technical issues confronting the scientific community. A workshop held in 2005 led to adoption of the policy that purely *in silico* structural models do not belong in the PDB (Berman et al. 2006), and, instead, an independent repository should be created to archive computed models elsewhere. The Protein Modeling Portal was established in 2007 (Arnold et al. 2009). In 2012, to address the challenges posed by the presence of a number of non-atomistic structural models of proteins obtained *via* small-angle scattering (SAS), the wwPDB SAS Task Force was established. This group of community stakeholders met and recommended creation of a SAS data repository that should interoperate with the PDB archive (Trehwella et al. 2013). Subsequently, some 49 PDB entries derived exclusively from SAS methods were transferred into the SAS Biological Data Bank (SASBDB; sasbdb.org) archive (Valentini et al. 2015) and then obsoleted (retired) from the PDB archive. In 2015, the wwPDB partnered with the Cambridge Crystallographic Data Center (CCDC; www.ccdc.cam.ac.uk) (Groom et al. 2016) and the Drug Design Data Resource (D3R; drugdesigndata.org) to convene a Ligand Validation Workshop, focused on improving the quality and utility of co-crystal structures in the PDB archive. Published recommendations pertaining to representation of small-molecules and validation of co-crystal structures coming from this workshop (Adams et al. 2016) were endorsed by the wwPDB X-ray VTF in late 2015. Implementation thereof was underway at the time of writing (May 2016).

Data Representation for Biological Macromolecules, Metadata, and Experimental Methods and Results

The PDB archive contains comprehensive descriptions of structural models coming from crystallography, NMR, and 3DEM. Each archival entry is designated by a 4-character PDB identifier (e.g., 1VTL). In addition to atomic coordinates, details regarding the chemistry of biopolymers and any bound small molecules are archived, as are metadata describing biopolymer

sequence, sample composition and preparation, experimental procedures, data-processing methods/software/statistics, structure determination/refinement procedures and statistics, and certain structural features, such as the secondary and quaternary structure. Primary experimental data coming from crystallography (structure-factor amplitudes or intensities) and NMR (restraints and chemical shifts) must be archived in the PDB. Voluntary archiving of diffraction images is currently supported by two resources that operate independently of the PDB, including the Integrated Resource for Reproducibility in Macromolecular Crystallography (IRRM; www.proteindiffraction.org) and the Structural Biology Data Grid Consortium (SBGrid; sbgrid.org (Meyer et al. 2016)) both of which use digital object identifiers to make the data readily accessible. In addition, some synchrotron radiation facilities now store diffraction images in locally maintained repositories, with data retention and dissemination policies set by the facility. BMRB (Markley et al. 2003) has long served as a public repository for NMR experimental data that are not stored in the PDB. Mass density maps used to derive structural models from 3DEM can be archived in EMDB (Lawson et al. 2016). Voluntary archival deposition of raw 3DEM images is currently supported by EMPIAR (Iudin et al. 2016).

The first data format used by the PDB archive was established in the early 1970s, based on the 80-column Hollerith format used for punched cards (Bernstein et al. 1977). Atom records included atom name, residue name, polymer chain identifier, and polymer sequence number. A set of “header records” contained limited metadata. The community readily accepted this format, because it was simple and both human- and machine-readable. However, the format also had limitations that became serious liabilities as structural biologists took the field to new heights. Structural models were limited to 99,999 atoms and relationships among various data items were implicit. These and other weaknesses of the legacy PDB format meant that deep subject matter expertise was required to both create and use software relying on this format. In the 1990s, the International Union of Crystallography charged a committee with creating a more informative and extensible data model for the PDB archive.

In response to the report, the Macromolecular Crystallographic Information File (mmCIF) was proposed (Fitzgerald et al. 2005). mmCIF is a self-defining format in which every data item has attributes describing its features, including explicit definitions of relationships among data items. Most important, mmCIF has no limitations with respect to the size of the structural model to be described. In addition, the mmCIF dictionary and mmCIF format data files are fully machine-

readable, and no domain knowledge is required to read the files. At inception, the mmCIF dictionary contained over 3,000 data items pertaining to crystallography. Over time, data items specific to NMR and 3DEM were added, and the dictionary was subsequently rebranded PDBx/mmCIF (Westbrook et al. 2005b). In 2007, it was decided that PDBx would be the PDB Master Format for data collected by the wwPDB. In 2011, major crystallographic structure determination software developers agreed to adopt this data model so that all output from their programs would be available in PDBx/mmCIF going forward.

In collaboration with community stakeholders serving on the PDBx/mmCIF Working Group (wwpdb.org/task/mmcif), the wwPDB continues to extend and enhance archival data representations. As of December 2014, PDBx/mmCIF became the official format for distribution of PDB entries. At the time of writing (May 2016), the PDBx/mmCIF dictionary contains more than 4,400 data items, including ~250 and ~1200 specific to NMR and 3DEM, respectively. PDBML, an XML format based on PDBx/mmCIF (Westbrook et al. 2005a) and the requisite RDF (Resource Description Framework) conversion have also been developed to facilitate integration of structural biology data with other life sciences data resources (Kinjo et al. 2012). Recently, XML and RDF-formatted BMRB data have been provided as BMRB/XML and BMRB/RDF, respectively (Yokochi et al. 2016), by which a federated SPARQL query linking the BMRB is made available to other databases. Finally, other structural biology communities are building on the PDBx/mmCIF framework to establish their own controlled vocabulary and specialist data items. For example, SASBDB has been working in collaboration with wwPDB partners to develop sasCIF (Malfois and Svergun 2000), which builds on PDBx/mmCIF. In addition to accelerating development of SASBDB, creation of sasCIF will allow for its facile interoperation with the PDB archive using a common exchange protocol based on PDBx/mmCIF.

In 1996, BMRB adopted NMR-STAR (a version of mmCIF) as its archival format (Ulrich et al. 1996). As noted above, this format has been harmonized with PDBx/mmCIF and now serves as the preferred deposition format for NMR structures (Berman et al. 2009). Historically, most NMR experimental data have been deposited in “native” format provided by each software package and archived “as is” in the PDB. Format harmonization was partially addressed by the NMR Restraints Grid, which can process restraint files and convert them to NMR-STAR or CCPN format

(Doreleijers et al. 2009; Doreleijers et al. 2012). In 2013 and 2014, community stakeholders participating in a pair of NMR format meetings convened by the wwPDB NMR VTF, recommended that an NMR Exchange Format (NEF) be developed for facile data transfer among NMR software packages and faithful conversion to NMR-STAR (Gutmanas et al. 2015). BMRB-led efforts are now underway to complete harmonization of NEF with NMR-STAR/PDBx/mmCIF to support NMR data deposition, annotation, and validation using the wwPDB unified global system (deposit.wwpdb.org).

Prior to 2015, reliance on the original PDB format made it necessary for large structure depositions (e.g., ribosomes/ribosomal subunits) archived in the PDB to be “split” into multiple entries, each with its own 4-character PDB identifier and legacy PDB-format file. This stopgap arrangement was entirely suboptimal. Splitting depositions among multiple PDB entries effectively preclude routine visualization of some of the most interesting structural models in the PDB archive, owing to software limitations. With adoption of the PDBx/mmCIF standard, every PDB archival entry is now stored as a single PDBx/mmCIF file, including 277 large structures that had previously been “split”. At the time of writing (May 2016) and for the foreseeable future, archival entries are made available as a public service in “stripped down,” best-effort PDB legacy format files wherever possible. In time, visualization, computational chemistry, etc. software providers will need to adjust to the new format and use PDBx/mmCIF files directly.

Data Representation for Small-Molecules

The PDB Chemical Component Dictionary (CCD) was originally developed (Westbrook et al. 2015) to provide a more expressive alternative to the earliest PDB ligand descriptions, which were based purely on atom connectivity records. The CCD embraced data representations for chemical components developed for the PDBx/mmCIF data dictionary (Fitzgerald et al. 2005). Each new chemical component coming in to the archive is identified by a unique 3-character alphanumeric code assigned by the wwPDB. The dictionary contains detailed chemical descriptions for standard and modified amino acids/nucleotides, small molecule ligands, and solvent/solute molecules (e.g., chemical properties, such as stereo chemical assignments, chemical descriptors, and systematic chemical names). A set of atomic model coordinates from a selected PDB entry and a computed set of ideal atomic coordinates are provided for each CCD entry. Hydrogen atoms are

computationally added to the experimental coordinates and any unobserved heavy atoms, such as leaving groups, are included in the ideal coordinates. Exact matches between the PDB CCD and the Cambridge Structural Database operated by CCDC (Groom et al. 2016) were identified in a collaborative effort, which revealed >1,400 common entries. An External Reference File containing both CCD and CSD descriptors of such matches is available from the PDB Chemical Component Model file (wwpdb.org/data/ccd).

A related PDB chemical reference dictionary is the Biologically Interesting molecule Reference Dictionary (BIRD) (Dutta et al. 2014), which contains information about peptide-like molecules in the PDB archive. BIRD entries include molecular weight and chemical formula, polymer sequence and connectivity, descriptions of structural features and functional classification, natural source, and external references to corresponding UniProt (UniProt Consortium 2015) or Norine (Caboche et al. 2008) reference sequences. BIRD molecules may be represented as a polymer (with sequence information) or as a single compound (with chemical information). Preferred representations are specified in the BIRD file, with a representative PDB identifier. The BIRD resource provides both possible representations; sequence and chemical information are provided in parallel.

Distributed Data Dissemination and Value-Added wwPDB Partner Activities

PDB archival data are freely available to the public without limitations on use. Data are released either immediately after they have been fully annotated/validated or—in most cases—when they are published in a scientific journal. Typically, either the author or the journal informs the wwPDB that the paper describing a given structure is about to be published. At this stage, the primary literature reference for the entry is updated and all data are released together with the wwPDB Validation Report.

PDB data release occurs in two stages. Stage 1: every Saturday at 03:00 UTC the polymer sequences, ligand SMILES strings, and crystallization pH for new structures designated for release are made public (wwpdb.org/download/downloads). Two-stage release is performed as a courtesy to the protein structure modeling and computational chemistry communities to enable

two weekly blinded prediction challenges (CAMEO: cameo3d.org (Haas et al. 2013); D3R CELPP: drugdesigndata.org/about/celpp). Stage 2: every Wednesday at 00:00 UTC, all new structures designated for release are made publicly available through the wwPDB FTP sites (wwPDB: [ftp.wwpdb.org](ftp://wwpdb.org); RCSB PDB: [ftp.rcsb.org](ftp://rcsb.org); PDBe: [ftp.ebi.ac.uk/pub/databases/pdb/](ftp://ebi.ac.uk/pub/databases/pdb/); PDBj: [ftp.pdbj.org](ftp://pdbj.org)). On average, ~200 structures are released every week, corresponding to ~10,500 structures released/year. Annually, in late December, “snapshots” of the PDB archive are recorded and also made available for FTP download (RCSB PDB: <ftp://snapshots.wwpdb.org/>; PDBj: <ftp://snapshots.pdbj.org/>). The wwPDB FTP sites provide core data for many secondary data resources, services, and websites.

When the wwPDB was established in 2003, it was agreed that, to best serve science, wwPDB partner websites would compete with one another on “Data Out” and offer many different kinds of services and features (RCSB PDB: rcsb.org; PDBe: pdbe.org; PDBj: pdbj.org; BMRB: bmrwisc.edu). Collectively, wwPDB FTP sites and partner websites support in excess of 500 million downloads of atomic coordinate data sets annually. In other words, more than 1 million sets of atomic coordinate data are downloaded by PDB users distributed across all inhabited continents every day of the year (Figure 4).

Future of Structural Biology and the Role of the wwPDB

At the time of writing (May 2016), PDB archival entries come exclusively from measurements crystallography, NMR, and 3DEM. These mainstay structure determination methods involve the same four basic steps: i) making measurements from a physical sample; ii) utilizing a representation of the measured data that allows encoding of these data for use by a computable scoring function utilizing spatial restraints that allows direct comparison between predicted and measured experimental results; iii) construction of structural models of identical composition but differing spatial configurations, followed by identification of one or more models with superior scores from the scoring function; and iv) evaluation of structural models to quantify agreement between prediction and experiment and estimate the uncertainty of each structural model. Notwithstanding the enormous amounts of experimental data measured by structural biologists today, none of the three PDB-supported methods routinely produce sufficient data to serve as the sole source of spatial restraints with which to produce a high quality structural model of a biological

macromolecule. Instead, we are forced to combine available experimental data with molecular mechanics force field descriptions of atomic structure for both biopolymers and small molecule ligands. These descriptions represent an essential source of additional spatial restraints corresponding to familiar items such as bond lengths, bond angles, descriptions of chiral centers, aromaticity, etc., which together with experimental data help to ensure that a structural model of a protein or nucleic acid chain makes chemical “sense”.

Structural biologists today rely increasingly on complementary experimental measurements to improve research outcomes. For example, it is becoming commonplace to utilize, or “integrate”, the results of SAS measurements as an additional source of spatial restraints when computing ensembles of structural models derived primarily from NMR data (reviewed in (Prischi and Pastore 2016)). Specifically, SAS experimental data serve as a source of spatial restraints reflecting the overall dimensions and shape of the macromolecule, whereas NMR experimental data provide information regarding proximity of different parts of the biopolymer chain with respect to one another. Combined NMR-SAS structure determinations typically yield significant improvements in both accuracy and precision of structural models *versus* those computed solely with the NMR data, particularly for dynamic systems (Cornilescu et al. 2016; Venditti et al. 2016).

With the recent advent of direct electron detectors and improvements in sample preparation for electron microscopy under cryogenic conditions, 3DEM is poised to become the experimental method of choice for studying larger macromolecular systems, many of which are ill suited to either crystallography or NMR. While the number of 3DEM structural models determined at better than 4Å resolution and released in the PDB archive is on the rise (3 in 2012 *versus* 68 in 2015), many 3DEM data sets of biological macromolecules are unlikely to yield atomic level structural models absent integration of complementary experimental data with the mass density map coming from 3DEM. To this end, cryo-electron microscopy studies are increasingly being combined with measurements using one or more of the following methods: crystallography, NMR, chemical crosslinking/mass spectrometry, Forster resonance energy transfer or FRET, and SAS (e.g., (Erzberger et al. 2014)). Structural models produced with these integrative (or hybrid) methods have been deposited in the PDB archive, but there is currently no mechanism for PDB archiving of experimental data and associated metadata generated by methods other than crystallography,

NMR, and 3DEM. Moreover, there are no universally accepted procedures by which integrative structural models can be validated against experimental data combined from different methods.

In 2014, the wwPDB Integrative/Hybrid Methods Task Force was assembled to assess some of these challenges. Attendees included experts in relevant measurement techniques, integrative modeling, visualization, and experimental data/structural model archiving. The meeting culminated in a unanimous recommendation that the wwPDB work with subject matter experts from complementary experimental methods to ensure that integrative 3D structural models can be deposited to the PDB archive with appropriate annotation/validation, and that all of the supporting experimental data and associated metadata be made publicly available through a system of federated data resources. An account of this meeting (Sali et al. 2015) provides guidance as to what experimental data and metadata should be archived, how data should be exchanged among data resources, and how structural models should be validated. Meeting participants quite deliberately decided not to prescribe the make up of the federation. Instead, an Integrative/Hybrid Methods Working Group (led by Helen M. Berman, Andrej Sali, Torsten Schwede, and Jill Trewella) was established after the meeting to work with the wwPDB partners in establishing the data resource federation. At the time of writing (May 2016), the SASBDB resource (Valentini et al. 2015) is working closely with wwPDB partners to develop joint data exchange and validation protocols to allow for deposition, annotation, and validation of 3D atomic level structural models determined *via* crystallography, NMR, or 3DEM combined with SAS data.

PDB Archive at 50 Years of Age

The PDB is just five years short of its 50th birthday. Based on current deposition rates, archival contents in 2021 will number well in excess of 150,000 entries (i.e., >20,000-fold bigger than in 1971). wwPDB partners are working closely with one another and the global structural biology community to ensure that a federated data resource system is established to enable deposition, annotation, and validation of 3D integrative structural models of biological macromolecules together with supporting data from diverse experimental methods and associated metadata. By 2021, it is also likely that the wwPDB partnership will have grown to encompass one or more additional regional data centers to help meet the needs of growing structural biology communities in different parts of the world.

Table 1. Proxy measures of complexity for recent PDB archival entries (2012-2015).

Year	Number of new entries with number of polymer chains>62	Number of new entries with MW>500,000	Number of new protein-nucleic acid complexes	Number of new compounds added to the Chemical Component Dictionary
2012	14	133	~450	1733
2013	32	198	~440	1875
2014	49	164	~690	1767
2015	55	311	~580	1830

Figure 1. Growth of the PDB Archive since 1971.

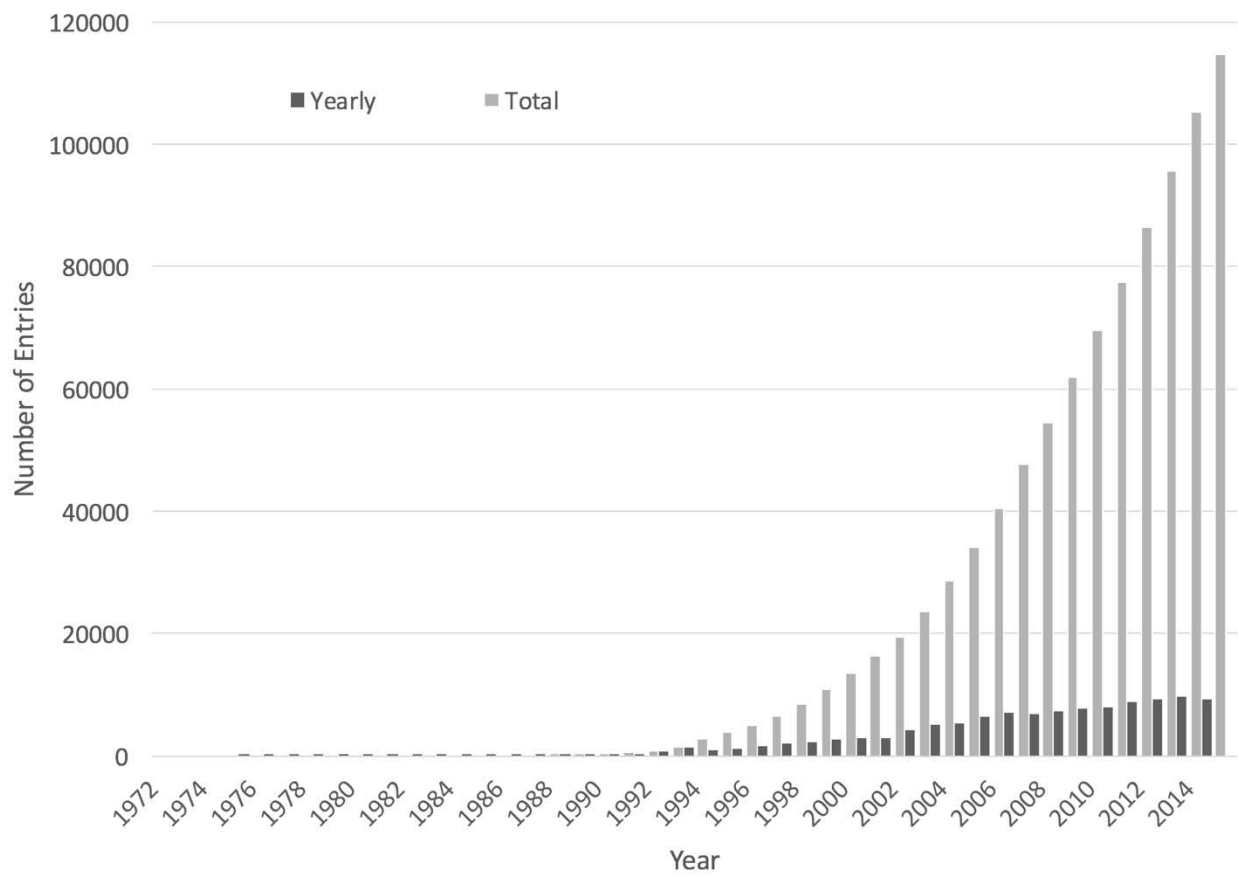


Figure 2. wwPDB Deposition, Annotation, and Validation Pipeline. Each box represents a modular component of the data processing workflow.

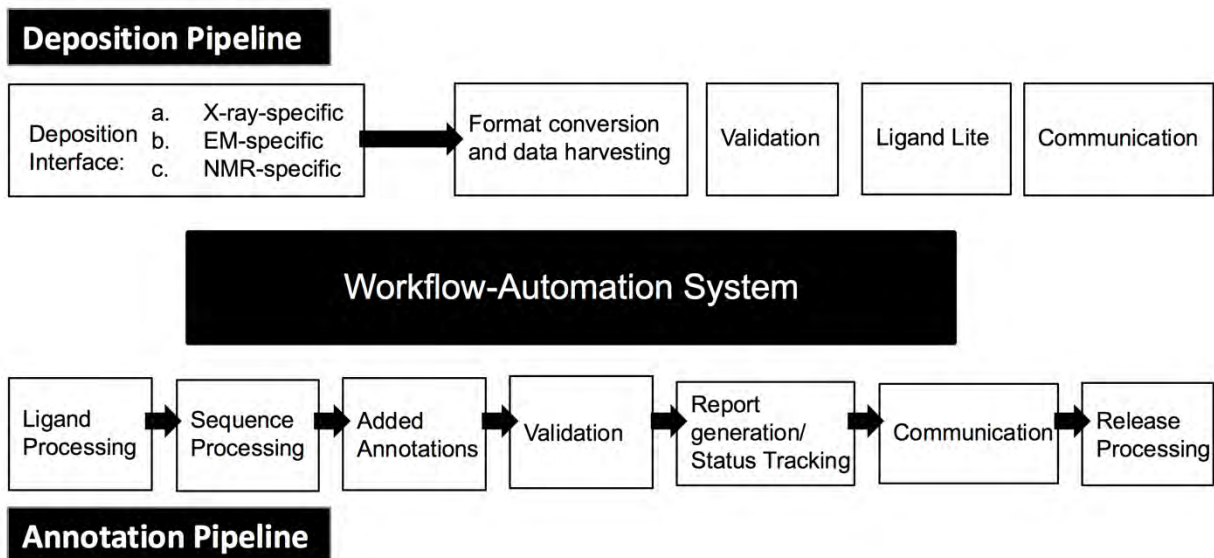


Figure 3. World map showing global distribution of PDB Depositors (2012-2015).



Figure 4. World map showing geographic distribution of PDB FTP download users (2012-2015).
[Note to Editors: An updated version of this figure is in preparation and will be provided ASAP.]



Acknowledgements

The RCSB PDB is supported by the National Science Foundation (DBI 1338415), National Institutes of Health, and the Department of Energy; PDBe by the Wellcome Trust, BBSRC, MRC, EU, CCP4, and EMBL-EBI; PDBj by JST-NBDC; and BMRB by the National Institute of General Medical Sciences (GM109046).

References

- Adams PD, Aertgeerts K, Bauer C, Bell JA, Berman HM, Bhat TN, Blaney JM, Bolton E, Bricogne G, Brown D, Burley SK, Case DA, Clark KL, Darden T, Emsley P, Feher VA, Feng Z, Groom CR, Harris SF, Hendle J, Holder T, Joachimiak A, Kleywegt GJ, Krojer T, Marcotrigiano J, Mark AE, Markley JL, Miller M, Minor W, Montelione GT, Murshudov G, Nakagawa A, Nakamura H, Nicholls A, Nicklaus M, Nolte RT, Padyana AK, Peishoff CE, Pieniazek S, Read RJ, Shao C, Sheriff S, Smart O, Soisson S, Spurlino J, Stouch T, Svobodova R, Tempel W, Terwilliger TC, Tronrud D, Velankar S, Ward SC, Warren GL, Westbrook JD, Williams P, Yang H, Young J (2016) Outcome of the First wwPDB/CCDC/D3R Ligand Validation Workshop. *Structure* 24 (4):502-508. doi:10.1016/j.str.2016.02.017
- Arnold K, Kiefer F, Kopp J, Battey JN, Podvinec M, Westbrook JD, Berman HM, Bordoli L, Schwede T (2009) The Protein Model Portal. *J Struct Funct Genomics* 10 (1):1-8. doi:10.1007/s10969-008-9048-5
- Berman H (2008) The Protein Data Bank: a historical perspective. *Acta Crystallogr A: Foundations of Crystallography* 64:88-95. doi:10.1107/S0108767307035623
- Berman HM, Burley SK, Chiu W, Sali A, Adzhubei A, Bourne PE, Bryant SH, Dunbrack RL, Jr., Fidelis K, Frank J, Godzik A, Henrick K, Joachimiak A, Heymann B, Jones D, Markley JL, Moulton J, Montelione GT, Orengo C, Rossmann MG, Rost B, Saibil H, Schwede T, Standley DM, Westbrook JD (2006) Outcome of a workshop on archiving structural models of biological macromolecules. *Structure* 14 (8):1211-1217. doi: 10.1016/j.str.2006.06.005
- Berman HM, Henrick K, Nakamura H (2003) Announcing the worldwide Protein Data Bank. *Nat Struct Biol* 10 (12):980. doi:10.1038/nsb1203-980
- Berman HM, Henrick K, Nakamura H, Markley JL (2009) Chapter 11 The Worldwide Protein Data Bank. In: Gu J, Bourne PE (eds) *Structural Bioinformatics, Second Edition*. John Wiley & Sons, Inc., Hoboken, NJ, pp 293-303.
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res* 28 (1):235-242. doi:10.1093/nar/28.1.235
- Bernstein FC, Koetzle TF, Williams GJB, Meyer Jr. EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M (1977) Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol* 112:535-542.
- Bolton W, Perutz MF (1970) Three dimensional fourier synthesis of horse deoxyhaemoglobin at 2.8 Ångstrom units resolution. *Nature* 228 (271):551-552. doi:10.1038/228551a0
- Caboche S, Pupin M, Leclere V, Fontaine A, Jacques P, Kucherov G (2008) NORINE: a database of nonribosomal peptides. *Nucleic Acids Res* 36 (Database issue):D326-331. doi:gkm792 [pii]
- Cold Spring Harbor Symposia on Quantitative Biology (1972), vol 36. Cold Spring Laboratory Press.

- Cornilescu G, Didychuk AL, Rodgers ML, Michael LA, Burke JE, Montemayor EJ, Hoskins AA, Butcher SE (2016) Structural Analysis of Multi-Helical RNAs by NMR-SAXS/WAXS: Application to the U4/U6 di-snRNA. *J Mol Biol* 428 (5 Pt A):777-789. doi:10.1016/j.jmb.2015.11.026
- Doreleijers JF, Vranken WF, Schulte C, Lin J, Wedell JR, Penkett CJ, Vuister GW, Vriend G, Markley JL, Ulrich EL (2009) The NMR restraints grid at BMRB for 5,266 protein and nucleic acid PDB entries. *J Biomol NMR* 45 (4):389-396. doi:10.1007/s10858-009-9378-z
- Doreleijers JF, Vranken WF, Schulte C, Markley JL, Ulrich EL, Vriend G, Vuister GW (2012) NRG-CING: integrated validation reports of remediated experimental biomolecular NMR data and coordinates in wwPDB. *Nucleic Acids Res* 40 (Database issue):D519-524. doi:10.1093/nar/gkr1134
- Dutta S, Dimitropoulos D, Feng Z, Persikova I, Sen S, Shao C, Westbrook J, Young J, Zhuravleva MA, Kleywegt GJ, Berman HM (2014) Improving the representation of peptide-like inhibitor and antibiotic molecules in the Protein Data Bank. *Biopolymers* 101 (6):659-668. doi:10.1002/bip.22434
- Erzberger JP, Stengel F, Pellarin R, Zhang S, Schaefer T, Aylett CH, Cimermancic P, Boehringer D, Sali A, Aebersold R, Ban N (2014) Molecular architecture of the 40S eIF3 translation initiation complex. *Cell* 158 (5):1123-1135. doi:10.1016/j.cell.2014.07.044
- Fitzgerald PMD, Westbrook JD, Bourne PE, McMahon B, Watenpaugh KD, Berman HM (2005) 4.5 Macromolecular dictionary (mmCIF). In: Hall SR, McMahon B (eds) *International Tables for Crystallography G. Definition and exchange of crystallographic data*. Springer, Dordrecht, The Netherlands, pp 295-443.
- Groom CR, Bruno IJ, Lightfoot MP, Ward SC (2016) The Cambridge Structural Database. *Acta Crystallogr B Struct Sci Cryst Eng Mater* 72 (Pt 2):171-179. doi:10.1107/S2052520616003954
- Gutmanas A, Adams PD, Bardiaux B, Berman HM, Case DA, Fogh RH, Guntert P, Hendrickx PM, Herrmann T, Kleywegt GJ, Kobayashi N, Lange OF, Markley JL, Montelione GT, Nilges M, Ragan TJ, Schwieters CD, Tejero R, Ulrich EL, Velankar S, Vranken WF, Wedell JR, Westbrook J, Wishart DS, Vuister GW (2015) NMR Exchange Format: a unified and open standard for representation of NMR restraint data. *Nat Struct Mol Biol* 22 (6):433-434. doi:10.1038/nsmb.3041
- Haas J, Roth S, Arnold K, Kiefer F, Schmidt T, Bordoli L, Schwede T (2013) The Protein Model Portal--a comprehensive resource for protein structure and model information. *Database : the journal of biological databases and curation* 2013:bat031. doi:10.1093/database/bat031
- Henderson R, Sali A, Baker ML, Carragher B, Devkota B, Downing KH, Egelman EH, Feng Z, Frank J, Grigorieff N, Jiang W, Ludtke SJ, Medalia O, Penczek PA, Rosenthal PB, Rossmann MG, Schmid MF, Schroder GF, Steven AC, Stokes DL, Westbrook JD, Wriggers W, Yang H, Young J, Berman HM, Chiu W, Kleywegt GJ, Lawson CL (2012) Outcome of the first electron microscopy validation task force meeting. *Structure* 20 (2):205-214. doi:10.1016/j.str.2011.12.014

- International Union of Crystallography (1989) Policy on publication and the deposition of data from crystallographic studies of biological macromolecules. *Acta Cryst* A45:658. doi:10.1107/S0108767389007695
- Iudin A, Korir PK, Salavert-Torres J, Kleywegt GJ, Patwardhan A (2016) EMPIAR: a public archive for raw electron microscopy image data. *Nature methods* 13. doi:10.1038/nmeth.3806
- Keller PA, Henrick K, McNeil P, Moodie S, Barton GJ (1998) Deposition of macromolecular structures. *Acta Crystallogr D Biol Crystallogr* 54 (Pt 6 Pt 1):1105-1108. doi:10.1107/S0907444998008464
- Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC (1958) A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* 181:662-666. doi:10.1038/181662a0
- Kendrew JC, Dickerson RE, Strandberg BE, Hart RG, Davies DR, Phillips DC, Shore VC (1960) Structure of myoglobin: A three-dimensional Fourier synthesis at 2 Å resolution. *Nature* 185 (4711):422-427. doi:10.1038/185422a0
- Kinjo AR, Suzuki H, Yamashita R, Ikegawa Y, Kudou T, Igarashi R, Kengaku Y, Cho H, Standley DM, Nakagawa A, Nakamura H (2012) Protein Data Bank Japan (PDBj): maintaining a structural data archive and resource description framework format. *Nucleic Acids Res* 40 (Database issue):D453-460. doi:10.1093/nar/gkr811
- Lawson CL, Patwardhan A, Baker ML, Hryc C, Garcia ES, Hudson BP, Lagerstedt I, Ludtke SJ, Pintilie G, Sala R, Westbrook JD, Berman HM, Kleywegt GJ, Chiu W (2016) EMDatabank unified data resource for 3DEM. *Nucleic Acids Res* 44 (D1):D396-403. doi:10.1093/nar/gkv1126
- Lin D, Manning NO, Jiang J, Abola EE, Stampf D, Prilusky J, Sussman JL (2000) AutoDep: a web-based system for deposition and validation of macromolecular structural information. *Acta Cryst D* D56:828-841. doi:10.1107/S0907444900005655
- Malfois M, Svergun DI (2000) sasCIF: an extension of core Crystallographic Information File for SAS. *J Appl Crystallogr* 33 (1):812-816. doi:10.1107/S0021889800001357
- Markley JL, Ulrich EL, Berman HM, Henrick K, Nakamura H, Akutsu H (2008) BioMagResBank (BMRB) as a partner in the Worldwide Protein Data Bank (wwPDB): new policies affecting biomolecular NMR depositions. *Journal of biomolecular NMR* 40 (3):153-155. DOI: 10.1007/s10858-008-9221-y
- Markley JL, Ulrich EL, Westler WM, Volkman BF (2003) Macromolecular structure determination by NMR spectroscopy. In: Bourne PE, Weissig H (eds) *Structural Bioinformatics*. John Wiley & Sons, Inc., Hoboken, NJ, pp 89-113.
- Meyer EF (1997) The first years of the Protein Data Bank. *Protein science : a publication of the Protein Society* 6:1591-1597.
- Meyer PA, Socias S, Key J, Ransey E, Tjon EC, Buschiazzi A, Lei M, Botka C, Withrow J, Neau D, Rajashankar K, Anderson KS, Baxter RH, Blacklow SC, Boggon TJ, Bonvin AM, Borek D, Brett TJ, Caflich A, Chang CI, Chazin WJ, Corbett KD, Cosgrove MS, Crosson S, Dhe-

- Paganon S, Di Cera E, Drennan CL, Eck MJ, Eichman BF, Fan QR, Ferre-D'Amare AR, Christopher Fromme J, Garcia KC, Gaudet R, Gong P, Harrison SC, Heldwein EE, Jia Z, Keenan RJ, Kruse AC, Kvensakul M, McLellan JS, Modis Y, Nam Y, Otwinowski Z, Pai EF, Pereira PJ, Petosa C, Raman CS, Rapoport TA, Roll-Mecak A, Rosen MK, Rudenko G, Schlessinger J, Schwartz TU, Shamoo Y, Sondermann H, Tao YJ, Tolia NH, Tsodikov OV, Westover KD, Wu H, Foster I, Fraser JS, Maia FR, Gonen T, Kirchhausen T, Diederichs K, Crosas M, Sliz P (2016) Data publication with the structural biology data grid supports live analysis. *Nature communications* 7:10882. doi:10.1038/ncomms10882
- Montelione GT, Nilges M, Bax A, Guntert P, Herrmann T, Richardson JS, Schwieters CD, Vranken WF, Vuister GW, Wishart DS, Berman HM, Kleywegt GJ, Markley JL (2013) Recommendations of the wwPDB NMR Validation Task Force. *Structure* 21 (9):1563-1570. doi:10.1016/j.str.2013.07.021
- Perutz MF, Rossmann MG, Cullis AF, Muirhead H, Will G, North ACT (1960) Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5 Å resolution, obtained by X-ray analysis. *Nature* 185:416-422. doi:10.1038/185416a0
- Prischi F, Pastore A (2016) Application of Nuclear Magnetic Resonance and Hybrid Methods to Structure Determination of Complex Systems. *Advances in experimental medicine and biology* 896:351-368. doi:10.1007/978-3-319-27216-0_22
- Protein Data Bank (1971) Protein Data Bank. *Nature New Biology* 233 (42):223. doi:10.1038/newbio233223b0
- Read RJ, Adams PD, Arendall WB, 3rd, Brunger AT, Emsley P, Joosten RP, Kleywegt GJ, Krissinel EB, Lutteke T, Otwinowski Z, Perrakis A, Richardson JS, Sheffler WH, Smith JL, Tickle IJ, Vriend G, Zwart PH (2011) A new generation of crystallographic validation tools for the protein data bank. *Structure* 19 (10):1395-1412. doi:10.1016/j.str.2011.08.006
- Sali A, Berman HM, Schwede T, Trewhella J, Kleywegt G, Burley SK, Markley J, Nakamura H, Adams P, Bonvin AM, Chiu W, Peraro MD, Di Maio F, Ferrin TE, Grunewald K, Gutmanas A, Henderson R, Hummer G, Iwasaki K, Johnson G, Lawson CL, Meiler J, Marti-Renom MA, Montelione GT, Nilges M, Nussinov R, Patwardhan A, Rappaport J, Read RJ, Saibil H, Schroder GF, Schwieters CD, Seidel CA, Svergun D, Topf M, Ulrich EL, Velankar S, Westbrook JD (2015) Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop. *Structure* 23 (7):1156-1167. doi:10.1016/j.str.2015.05.013
- Standley DM, Kinjo AR, Kinoshita K, Nakamura H (2008) Protein structure databases with new web services for structural biology and biomedical research. *Briefings in bioinformatics* 9 (4):276-285. doi: 10.1093/bib/bbn015
- Sussman JL, Lin D, Jiang J, Manning NO, Prilusky J, Ritter O, Abola EE (1998) Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta Crystallogr D Biol Crystallogr* 54 (Pt 6 Pt 1):1078-1084. doi:10.1107/S09074444998009378
- Tagari M, Tate J, Swaminathan GJ, Newman R, Naim A, Vranken W, Kapopoulou A, Hussain A, Fillon J, Henrick K, Velankar S (2006) E-MSD: improving data deposition and structure quality. *Nucleic Acids Res* 34 (Database issue):D287-290. doi:10.1093/nar/gkj163

- Trewhella J, Hendrickson WA, Kleywegt GJ, Sali A, Sato M, Schwede T, Svergun DI, Tainer JA, Westbrook J, Berman HM (2013) Report of the wwPDB Small-Angle Scattering Task Force: data requirements for biomolecular modeling and the PDB. *Structure* 21 (6):875-881. doi:10.1016/j.str.2013.04.020
- Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Kent Wenger R, Yao H, Markley JL (2008) BioMagResBank. *Nucleic Acids Res* 36 (Database issue):D402-408. doi:10.1093/nar/gkm957
- Ulrich EL, Argentar D, Klimowicz A, Markley JL (1996) STAR/CIF macromolecular NMR data dictionaries and data file formats. *Acta Cryst A* 52(a1):C577-C577.
- Ulrich EL, Markley JL, Kyogoku Y (1989) Creation of a Nuclear Magnetic Resonance Data Repository and Literature Database. *Protein Seq Data Anal* 2:23-37.
- UniProt Consortium (2015) UniProt: a hub for protein information. *Nucleic Acids Res* 43 (Database issue):D204-212. doi:10.1093/nar/gku989
- Valentini E, Kikhney AG, Previtali G, Jeffries CM, Svergun DI (2015) SASBDB, a repository for biological small-angle scattering data. *Nucleic Acids Res* 43 (Database issue):D357-363. doi:10.1093/nar/gku1047
- Velankar S, Best C, Beuth B, Boutselakis CH, Cobley N, Sousa Da Silva AW, Dimitropoulos D, Golovin A, Hirshberg M, John M, Krissinel EB, Newman R, Oldfield T, Pajon A, Penkett CJ, Pineda-Castillo J, Sahni G, Sen S, Slowley R, Suarez-Uruena A, Swaminathan J, van Ginkel G, Vranken WF, Henrick K, Kleywegt GJ (2010) PDBe: Protein Data Bank in Europe. *Nucleic Acids Res* 38 (Database issue):D308-317. doi:10.1093/nar/gkp916
- Velankar S, van Ginkel G, Alhroub Y, Battle GM, Berrisford JM, Conroy MJ, Dana JM, Gore SP, Gutmanas A, Haslam P, Hendrickx PM, Lagerstedt I, Mir S, Fernandez Montecelo MA, Mukhopadhyay A, Oldfield TJ, Patwardhan A, Sanz-Garcia E, Sen S, Slowley RA, Wainwright ME, Deshpande MS, Iudin A, Sahni G, Salavert Torres J, Hirshberg M, Mak L, Nadzirin N, Armstrong DR, Clark AR, Smart OS, Korir PK, Kleywegt GJ (2016) PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. *Nucleic Acids Res* 44 (D1):D385-395. doi:10.1093/nar/gkv1047
- Venditti V, Egner TK, Clore GM (2016) Hybrid Approaches to Structural Characterization of Conformational Ensembles of Complex Macromolecular Systems Combining NMR Residual Dipolar Couplings and Solution X-ray Scattering. *Chem Rev*: in the press. doi:10.1021/acs.chemrev.5b00592
- Westbrook J, Ito N, Nakamura H, Henrick K, Berman HM (2005a) PDBML: the representation of archival macromolecular structure data in XML. *Bioinformatics* 21 (7):988-992. doi:10.1093/bioinformatics/bti082
- Westbrook JD, Henrick K, Ulrich EL, Berman HM (2005b) Appendix 3.6.2 The Protein Data Bank Exchange Data Dictionary. In: Hall SR, McMahon B (eds) *International Tables for Crystallography G. Definition and exchange of crystallographic data*. Springer, Dordrecht, The Netherlands, pp 195-198.

- Westbrook JD, Shao C, Feng Z, Zhuravleva M, Velankar S, Young J (2015) The chemical component dictionary: complete descriptions of constituent molecules in experimentally determined 3D macromolecules in the Protein Data Bank. *Bioinformatics* 31 (8):1274-1278. doi:10.1093/bioinformatics/btu789
- Yokochi M, Kobayashi N, Ulrich EL, Kinjo AR, Iwata T, Ioannidis YE, Livny M, Markley JL, Nakamura H, Kojima C, Fujiwara T (2016) Publication of nuclear magnetic resonance experimental data with semantic web technology and the application thereof to biomedical research of proteins. *Journal of Biomedical Semantics* 7:16 doi: 10.1186/s13326-016-0057-1

Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop

Andrej Sali,^{1,*} Helen M. Berman,² Torsten Schwede,³ Jill Trehwella,⁴ Gerard Kleywegt,⁵ Stephen K. Burley,^{2,6} John Markley,⁷ Haruki Nakamura,⁸ Paul Adams,^{9,10} Alexandre M.J.J. Bonvin,¹¹ Wah Chiu,¹² Matteo Dal Peraro,¹³ Frank Di Maio,¹⁴ Thomas E. Ferrin,¹⁵ Kay Grunewald,¹⁶ Aleksandras Gutmanas,⁵ Richard Henderson,¹⁷ Gerhard Hummer,¹⁸ Kenji Iwasaki,¹⁹ Graham Johnson,²⁰ Catherine L. Lawson,² Jens Meiler,²¹ Marc A. Marti-Renom,²² Gaetano T. Montelione,^{23,24} Michael Nilges,^{25,26} Ruth Nussinov,^{27,28} Ardan Patwardhan,⁵ Juri Rappsilber,^{29,30} Randy J. Read,³¹ Helen Saibil,³² Gunnar F. Schröder,^{33,34} Charles D. Schwieters,³⁵ Claus A.M. Seidel,³⁶ Dmitri Svergun,³⁷ Maya Topf,³² Eldon L. Ulrich,⁷ Sameer Velankar,⁵ and John D. Westbrook²

Structures of biomolecular systems are increasingly computed by integrative modeling that relies on varied types of experimental data and theoretical information. We describe here the proceedings and conclusions from the first wwPDB Hybrid/Integrative Methods Task Force Workshop held at the European Bioinformatics Institute in Hinxton, UK, on October 6 and 7, 2014. At the workshop, experts in various experimental fields of structural biology, experts in integrative modeling and visualization, and experts in data archiving addressed a series of questions central to the future of structural biology. How should integrative models be represented? How should the data and integrative models be validated? What data should be archived? How should the data and models be archived? What information should accompany the publication of integrative models?

Background

Historical Rationale for the Workshop

The PDB (<http://wwpdb.org>) was founded in 1971 with seven protein structures as its first holdings (Protein Data Bank, 1971). The global PDB archive now holds more than 100,000 atomic structures of biological macromolecules and their complexes, all of which are freely accessible. Most structures in the PDB archive (~90%) have been determined by X-ray crystallography, with the remainder contributed by two newer 3D structure determination methods, nuclear magnetic resonance (NMR) spectroscopy and 3D electron microscopy (3DEM).

Considerable effort has gone into understanding how to best curate the structural models and experimental data produced with these methods. Over the past several years, the Worldwide PDB (wwPDB; the global organization responsible for maintaining the PDB archive) (Berman et al., 2003) has established expert, method-specific task forces to advise on which experimental data and metadata from each method should be archived and how these data and the resulting structure models should be validated. The wwPDB X-ray Validation Task Force (VTF) made detailed recommendations on how to best validate structures determined by X-ray crystallography (Read et al., 2011). These

¹Department of Bioengineering and Therapeutic Sciences, Department of Pharmaceutical Chemistry, California Institute for Quantitative Biosciences, Byers Hall Room 503B, University of California, San Francisco, 1700 4th Street, San Francisco, CA 94158-2330, USA

²Research Collaboratory for Structural Bioinformatics Protein Data Bank, Center for Integrative Proteomics Research, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

³Swiss Institute of Bioinformatics Biozentrum, University of Basel, Klingelbergstrasse 50-70, 4056 Basel, Switzerland

⁴School of Molecular Bioscience, The University of Sydney, NSW 2006, Australia

⁵Protein Data Bank in Europe, European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK

⁶Skaggs School of Pharmacy and Pharmaceutical Sciences and San Diego Supercomputer Center, University of California, San Diego, La Jolla, CA 92093, USA

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

⁸Protein Data Bank Japan, Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan

⁹Physical Biosciences Division, Lawrence Berkeley Laboratory, Berkeley, CA 94720-8235, USA

¹⁰Department of Bioengineering, UC Berkeley, Berkeley, CA 94720, USA

¹¹Bijvoet Center for Biomolecular Research, Faculty of Science - Chemistry, Utrecht University, Padualaan 8, Utrecht, 3584 CH, the Netherlands

¹²National Center for Macromolecular Imaging, Baylor College of Medicine, Houston, TX 77030, USA

¹³Institute of Bioengineering, School of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL) and Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

¹⁴Department of Biochemistry, University of Washington, Seattle, WA 98195-7370, USA

¹⁵Department of Pharmaceutical Chemistry and Department of Bioengineering and Therapeutic Sciences, California Institute for Quantitative Biosciences, University of California, San Francisco, 600 16th Street, San Francisco, CA 94158-2517, USA

¹⁶Division of Structural Biology, Wellcome Trust Centre of Human Genetics, University of Oxford, OX3 7BN Oxford, UK

(Affiliations continued on next page)

recommendations have been implemented as a software pipeline used within the wwPDB Deposition and Annotation (D&A) system. Initial recommendations of the wwPDB NMR (Montelione et al., 2013) and Electron Microscopy (Henderson et al., 2012) VTFs have also been implemented. In addition, the wwPDB and, in later years, the Structural Biology Knowledgebase (SBKB), spearheaded three workshops focused on validation, archiving, and dissemination of comparative protein structure models (Berman et al., 2006; Schwede et al., 2009). It is anticipated that as new validation methods are developed and as more experience is gained with existing ones, additional validation procedures will be implemented in the wwPDB D&A system.

Increasingly, structures of very large macromolecular machines are being determined by combining observations from complementary experimental methods, including X-ray crystallography, NMR spectroscopy, 3DEM, small-angle scattering (SAS), crosslinking, and many others (Figure 1; Table 1). Data from these complementary methods are used to compute integrative or hybrid models (Ward et al., 2013). Atomic models produced in this fashion have been deposited in the PDB, but there is currently no mechanism within the PDB framework for archiving the experimental data generated by methods other than X-ray crystallography, NMR spectroscopy, and 3DEM. The most recently established task force, the wwPDB SAS Task Force (Trehwella et al., 2013), recommended creation of a SAS data and model repository that would interoperate with the PDB. The SAS Task Force also recommended that an international meeting be held to consider how best to deal with the archiving of data and models derived from integrative structure determination approaches.

In response, a Hybrid/Integrative Methods Task Force was assembled by the wwPDB organization. Its inaugural meeting

was held at the EMBL European Bioinformatics Institute (EBI) on October 6 and 7, 2014 (<http://wwpdb.org/task/hybrid.php>). In all, 38 participants from 37 academic and government institutions worldwide attended the workshop, which was co-chaired by Andrej Sali (University of California, San Francisco, USA), Torsten Schwede (Swiss Institute of Bioinformatics [SIB] and University of Basel, Switzerland), and Jill Trehwella (University of Sydney, Australia). Attendees included experts in relevant experimental techniques, integrative modeling, visualization, and data and model archiving.

The workshop began with plenary talks followed by focused discussions. Gerard Kleywegt introduced the workshop objectives. Andrej Sali outlined the current state of integrative modeling. Helen Berman gave an overview of the history and status of the wwPDB organization. Jill Trehwella described the increasing role of SAS in integrative structural modeling, the need for the development of community standards and validation tools for biomolecular modeling using SAS data, and how SAS data and modeling resources could interoperate with the PDB. Claus Seidel outlined state-of-the-art single-molecule and ensemble Förster resonance energy transfer (FRET) spectroscopy (Kalinin et al., 2012) and live cell imaging, as well as related label-based spectroscopic methods for measuring select interatomic distances in macromolecular systems. Torsten Schwede presented the Protein Model Portal (Haas et al., 2013), including its linking of large databases of comparative models with experimental structure information in the PDB, and the Model Archive repository for all categories of *in silico* structural models.

Current Archives for Models and/or Supporting Data

In this section, we review the PDB and management of data derived from crystallography, NMR spectroscopy, 3DEM, and

¹⁷MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge CB2 0QH, UK

¹⁸Department of Theoretical Biophysics, Max Planck Institute of Biophysics, Max-von-Laue Straße 3, 60438 Frankfurt am Main, Germany

¹⁹Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan

²⁰Department of Bioengineering and Therapeutic Sciences, California Institute for Quantitative Biosciences, University of California, San Francisco, 600 16th Street, San Francisco, CA 94158-2330, USA

²¹Department of Chemistry, Center for Structural Biology, Vanderbilt University, Nashville, TN 37235, USA

²²Genome Biology Group, Centre Nacional d'Anàlisi Genòmica (CNAG), Gene Regulation, Stem Cells and Cancer Program, Center for Genomic Regulation (CRG) and Institutió Catalana de Recerca i Estudis Avançats (ICREA), 08028 Barcelona, Spain

²³Center for Advanced Biotechnology and Medicine, Department of Molecular Biology and Biochemistry, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

²⁴Department of Biochemistry, Robert Wood Johnson Medical School, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

²⁵Département de Biologie Structurale et Chimie, Unité de Bioinformatique Structurale, Institut Pasteur, F-75015 Paris, France

²⁶Unité Mixte de Recherche 3258, Centre National de la Recherche Scientifique, F-75015 Paris, France

²⁷Cancer and Inflammation Program, Leidos Biomedical Research Inc., Frederick National Laboratory, National Cancer Institute, Frederick, MD 21702, USA

²⁸Department of Human Molecular Genetics and Biochemistry, Sackler School of Medicine, Tel Aviv University, Tel Aviv 69978, Israel

²⁹Wellcome Trust Centre for Cell Biology, Institute of Cell Biology, University of Edinburgh, Edinburgh EH9 3BF, UK

³⁰Department of Bioanalytics, Institute of Biotechnology, Technische Universität Berlin, 13355 Berlin, Germany

³¹Department of Haematology, Cambridge Institute for Medical Research, University of Cambridge, Cambridge CB2 0XY, UK

³²Institute of Structural and Molecular Biology, Department of Biological Sciences, Birkbeck College, Malet Street, London WC1E 7HX, UK

³³Institute of Complex Systems (ICS-6), Forschungszentrum Jülich, 52425 Jülich, Germany

³⁴Physics Department, Heinrich-Heine University Düsseldorf, 40225 Düsseldorf, Germany

³⁵Division of Computational Bioscience, Center for Information Technology, National Institutes of Health, Bethesda, MD 20892-0520, USA

³⁶Chair for Molecular Physical Chemistry, Heinrich-Heine-Universität, Universitätsstraße 1, 40225 Düsseldorf, Germany

³⁷European Molecular Biology Laboratory, Hamburg Unit, Notkestrasse 85, 22607 Hamburg, Germany

*Correspondence: sali@salilab.org

<http://dx.doi.org/10.1016/j.str.2015.05.013>

All attendees of the Workshop are listed as authors.

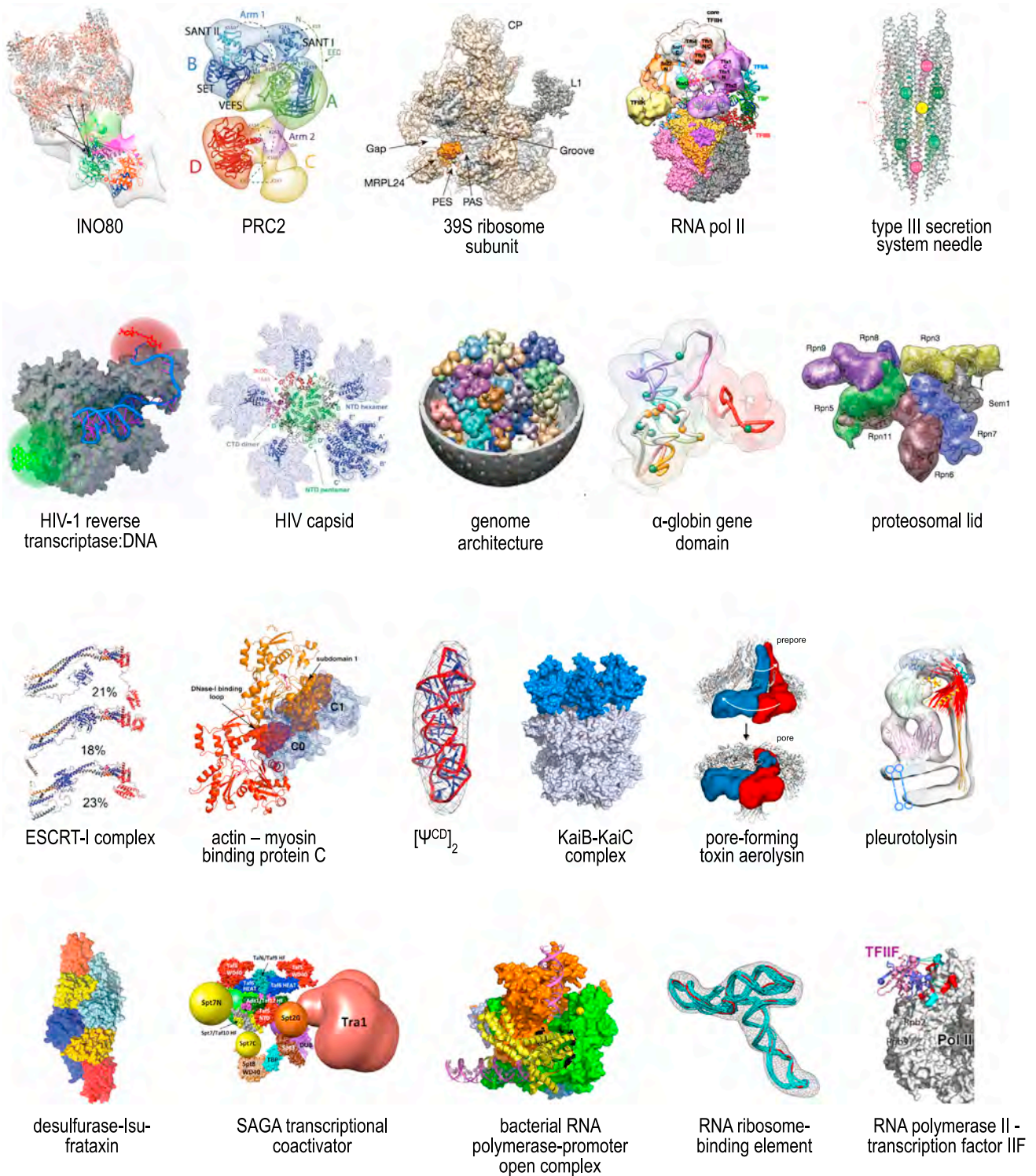


Figure 1. Examples of Recently Determined Integrative Structures

The molecular architecture of INO80 was determined with a 17-Å resolution cryo-electron microscopy (EM) map and 212 intra-protein and 116 inter-protein crosslinks (Russel et al., 2009). The molecular architecture of Polycomb Repressive Complex 2 (PRC2) was determined with a 21-Å resolution negative-stain EM map and ~60 intra-protein and inter-protein crosslinks (Shi et al., 2014). The molecular architecture of the large subunit of the mammalian mitochondrial ribosome (39S) was determined with a 4.9-Å resolution cryo-EM map and ~70 inter-protein crosslinks (Ward et al., 2013). The molecular architecture of the RNA polymerase II transcription pre-initiation complex was determined with a 16-Å resolution cryo-EM map plus 157 intra-protein and 109 inter-protein crosslinks (Alber et al., 2008). The atomic model of type III secretion system needle was determined with a 19.5-Å resolution cryo-EM map and solid-state nuclear magnetic resonance (NMR) data (Loquet et al., 2012). Molecular architecture of the productive HIV-1 reverse transcriptase:DNA primer-template complex in the open educt

(legend continued on next page)

Table 1. Types of Structural Data Used in Integrative Modeling

Structural Information	Method
Atomic structures of parts of the studied system	X-Ray and neutron crystallography, NMR spectroscopy, 3DEM, comparative modeling, and molecular docking
3D maps and 2D images	Electron microscopy and tomography
Atomic and protein distances	NMR, FRET, and other fluorescence techniques, DEER, EPR, and other spectroscopic techniques; chemical crosslinks detected by mass spectrometry, and disulfide bonds detected by gel electrophoresis
Binding site mapping	NMR spectroscopy, mutagenesis, FRET
Size, shape, and pairwise atomic distance distributions	SAS
Shape and size	Atomic force microscopy, ion mobility mass spectrometry, fluorescence correlation spectroscopy, and fluorescence anisotropy
Component positions	Super-resolution optical microscopy, FRET imaging
Physical proximity	Co-purification, native mass spectrometry, genetic methods, and gene/protein sequence covariance
Solvent accessibility	Footprinting methods, including H/D exchange assessed by mass spectrometry or NMR, and even functional consequences of point mutations
Proximity between different genome segments	Chromosome conformation capture and other data
Propensities for different interaction modes	Molecular mechanics force fields, potentials of mean force, statistical potentials, and sequence co-variation

Example methods that are informative about a variety of structural aspects of biomolecular systems are listed. 3DEM, 3D electron microscopy; DEER, double electron-electron resonance; EPR, electron paramagnetic resonance; FRET, Förster resonance energy transfer; H/D, hydrogen/deuterium; NMR, nuclear magnetic resonance; SAS, small-angle scattering.

SAS, plus archives for models derived exclusively on the basis on theoretical information.

PDB. For more than four decades, the PDB has served as the single global archive for atomic models of biological macromolecules; first for those derived from crystallography, and subsequently for models from NMR spectroscopy and 3DEM. The PDB also archives experimental data necessary to validate the structural models determined using these three methods. In addition, descriptions of the chemistry of polymers and ligands are collected, as are metadata describing sample preparation, experimental methods, model building, refinement statistics, literature references, and so forth. For all structural models in the PDB, geometric features are assessed with respect to standard valence geometry and intermolecular interactions, as recommended by the three wwPDB VTFs mentioned above.

Crystallography: Models and Data. For structures derived using X-ray, neutron, and combined X-ray/neutron crystallography, it has been mandatory to deposit structure factor amplitudes into the PDB since 2008 (<http://www.wwpdb.org/news/news?year=2007#29-November-2007>); until then, the submission of these primary data was optional. Additional validation against deposited structure factor amplitudes is carried out using procedures recommended by the X-ray VTF (Read et al., 2011). The resulting validation report includes graphical summaries of the quality of the overall model plus residue-specific features. Detailed assessments of various aspects of the model and its agreement with experimental and stereochemical data are also provided. In the near future, unmerged intensities will also be collected, enabling further validation activities.

state was determined by Förster resonance energy transfer (FRET) positioning and screening using a known HIV-1 reverse transcriptase structure (Kalinin et al., 2012). The structure of HIV-1 capsid protein was determined using residual dipolar couplings and small-angle X-ray scattering (SAXS) data (Deshmukh et al., 2013). The human genome architecture was determined based on tethered chromosome conformation capture and population-based modeling (Kalhor et al., 2012). The structural model of α -globin gene domain was determined based on Chromosome Conformation Capture Carbon Copy (5C) experiments (Bau et al., 2011). The molecular architecture of the proteasomal lid was determined using native mass spectrometry and 28 crosslinks (Politis et al., 2014). Structure models of the ESCRT-I complex were determined with SAXS, double electron-electron transfer, and FRET (Boura et al., 2011). Integrative model of actin and the cardiac myosin binding protein C was developed from a combination of crystallographic and NMR structures of subunits and domains, with positions and orientations optimized against SAXS and small-angle neutron scattering data to reveal information about the quaternary interactions (Whitten et al., 2008). The ensemble of $[\Psi^{CD}]_2$ NMR structures were fitted into the averaged cryo-electron tomography map (Miyazaki et al., 2010). Integrative model of the cyanobacterial circadian timing KaiB-KaiC complex was obtained based on hydrogen/deuterium exchange and collision cross-section data from mass spectrometry (Snijder et al., 2014). The pre-pore and pore conformations of the pore-forming toxin aerolysin were obtained combining cryo-EM data and molecular dynamics simulations (Degiacomi and Dal Peraro, 2013; Degiacomi et al., 2013). Segment of a pleurotolysin pore map (~11 Å resolution) with an ensemble of conformations shows the trajectory of β sheet opening during pore formation (Lukoyanova et al., 2015). A SAXS-based rigid-body model of a ternary complex of the iron-sulfur cluster assembly proteins desulfurase (orange) and scaffold protein Isu (blue) with bacterial ortholog of frataxin (yellow) was validated by NMR chemical shifts and mutagenesis (Prischi et al., 2010). The molecular architecture of the SAGA transcription coactivator complex was determined with 199 inter- and 240 intra-subunit crosslinks, several comparative models based on X-ray crystal structures, and a transcription factor IID core EM map at 31 Å resolution (Han et al., 2014). Structural organization of the bacterial (*Thermus aquaticus*) RNA polymerase-promoter open complex obtained by FRET (Mekler et al., 2002) was subsequently validated by a crystal structure (Zhang et al., 2012). The RNA ribosome-binding element from turnip crinkle virus genome was determined using NMR, SAXS, and EM data (Gong et al., 2015). The molecular architecture of the complex between RNA polymerase II and transcription factor IIF was determined using a deposited crystal structure of RNA polymerase II, homology models of crystal domains in transcription factor IIF, and 95 intra-protein and 129 inter-protein crosslinks (Chen et al., 2010).

NMR Spectroscopy: Models and Data. The Biological Magnetic Resonance DataBank (BioMagResBank or BMRB; <http://www.bmrwisc.edu>) is a repository for experimental and derived data gathered from NMR spectroscopic studies of biological molecules. The BMRB archive contains quantitative NMR spectral parameters, including assigned chemical shifts, coupling constants, and peak lists together with derived data, including relaxation parameters, residual dipolar couplings, hydrogen exchange rates, pK_a values, and so forth. Other data contained in the BMRB include: NMR restraints processed from original author depositions available from the PDB; time-domain spectral data from NMR experiments used to assign spectral resonances and determine structures of biological macromolecules; chemical shift and structure validation reports; and a database of 1D and 2D ^1H - and ^{13}C -NMR spectra for more than 1,200 metabolites. The BMRB website also provides tools for querying and retrieving data.

Since 2006, BMRB has been a member of the wwPDB organization (Markley et al., 2008). Chemical shift and restraint data that accompany model data are housed in both the BMRB and PDB archives. Deposited NMR data without model coordinates reside exclusively in the BMRB archive. The wwPDB D&A system provides for deposition, annotation, and validation of NMR models and related experimental data. Depositors of chemical shift and other data sets without accompanying models are automatically redirected to BMRB to deposit their data. Data exchange between the BMRB and PDB archives is facilitated by software tools utilizing correspondences maintained between the PDB Exchange Dictionary (PDBx) and the BMRB NMR-STAR Dictionary. Validation methods for NMR-derived models, measured chemical shifts, and restraint data are currently under development, in response to recommendations of the NMR VTF (Montelione et al., 2013). A working group composed of the major biomolecular NMR software developers has created a common NMR exchange format (NEF) for structural restraints, similar to NMR-STAR. The adoption of this NEF by NMR software developers will simplify data exchange and the archiving of NMR structural restraints by the wwPDB.

Electron Microscopy: Models and Maps. Atomistic structural models determined using 3DEM methods were first archived in the PDB in the 1990s. In 2002, the EM Data Bank (EMDB) was created by the Macromolecular Structure Database (now PDBe) at the EBI. In 2006, the EMDatabank (<http://www.EMDatabank.org>) was established as the unified global portal for one-stop deposition and retrieval of 3DEM density maps, atomic models, and associated metadata (Lawson et al., 2011). EMDatabank is a joint effort among PDBe, the Research Collaboratory for Structural Bioinformatics (RCSB) at Rutgers, and the National Center for Macromolecular Imaging (NCMI) at Baylor College of Medicine. EMDatabank also serves as a resource for news, events, software tools, data standards, raw data, and validation methods for the 3DEM community. 3DEM model and map data are now stored in separate branches of the wwPDB ftp archive site.

As for NMR-based models, the wwPDB D&A system supports processing of atomistic models and map data from 3DEM structure determinations. 3DEM map data deposited without atomistic models are stored exclusively in EMDb. Again, as for

NMR, a mapping is maintained between the PDBx data dictionary and the EMDb XML-based data model. Validation methods for 3DEM maps and atomistic models are currently under development in response to recommendations from the EM VTF (Henderson et al., 2012).

SAS: Data and Model Archiving. The report from the first meeting of the wwPDB SAS Task Force (Trehwella et al., 2013) made the case for establishing “a global repository that holds standard format X-ray and neutron SAS data that is searchable and freely accessible for download” and that “options should be provided for including in the repository SAS-derived shape and atomistic models based on rigid-body refinement against SAS data along with specific information regarding the uniqueness and uncertainty of the model, and the protocol used to obtain it.”

At present, there are two databases available for storing SAS data and models with associated metadata and analyses, both of which are freely accessible without limitations on data utilization via the Internet. As of March 2015, BIOISIS (<http://www.bioisis.net/>) contained 99 structures and is supported by teams at the Advanced Light Source and Diamond, while SASBDB (<http://www.sasbdb.org/>) (Valentini et al., 2015) contained 195 models and 114 experimental datasets and is supported by a team at EMBL-Hamburg.

Having evolved separately, these databases are distinctive in character. There was in principle agreement within the wwPDB SAS Task Force that BIOISIS and SASBDB will exchange data sets. Such exchange would be a step toward developing a federated approach to SAS data and model archiving, which in turn could ultimately be federated with the PDB, BMRB, and EMDb.

Further development of the sasCIF dictionary is required to permit full data exchange between the two SAS data repositories. sasCIF is a core crystallographic information file (CIF) developed to facilitate the SAS data exchange (Malfois and Svergun, 2000). As its name implies, sasCIF was implemented as an extension of the core CIF dictionary and has recently been extended to include new elements related to models, model fitting, validation tools, sample preparation, and experimental conditions (M.K., J.D.W., and D.S., unpublished data). sasCIFtools were developed as a documented set of publicly available programs for sasCIF data processing and format conversion; currently, SASBDB supports both import and export of sasCIF files.

Protein Model Portal. Comparative or homology modeling is routinely used to generate structural models of proteins for which experimentally determined structural models are not yet available (Marti-Renom et al., 2000; Schwede et al., 2009). Until 2006, such in silico models could be archived in the PDB, albeit in the absence of clear policies and procedures for their validation. Following recommendations from a stakeholder workshop convened in November 2005 (Berman et al., 2006), depositions to the PDB archive are limited to structural models substantially determined by experimental measurements from a defined physical sample (effective date October 15, 2006). The workshop also recommended that a central, publicly available archive or portal should be established for exclusively in silico models, and that methodology for estimating the accuracy of such computational models should be developed.

The Protein Model Portal (PMP) (Arnold et al., 2009; Haas et al., 2013) was developed at the SIB at the University of Basel

as a component of the SBKB (Berman et al., 2009; Gabanyi et al., 2011). Today, the SBKB integrates experimental information provided by the PDB with in silico models computed by automated modeling resources. In addition, the PMP provides access to several state-of-the-art model quality assessment services (Schwede et al., 2009). Since 2013, the Model Archive (<http://modelarchive.org>) resource has also served as a repository for individually generated in silico models of macromolecular structures, primarily those described in peer-reviewed publications. Finally, the Model Archive hosts all legacy models that were available from the PDB archive prior to 2006.

Each model in the PMP is assigned a stable, unique accession code (and digital object identifier or DOI) to ensure accurate cross-referencing in publications and other data repositories. Unlike experimentally determined structural models, in silico models are not the product of experimental measurements of a physical sample. They are generated computationally using various molecular modeling methods and underlying assumptions. Examples include comparative modeling, virtual docking of ligand molecules to protein targets, virtual docking of one protein to another, simulations of molecular dynamics and motions, and de novo (ab initio) protein modeling.

Effective archival storage of such models depends critically on capturing sufficient detail regarding underlying assumptions, parameters, methodology, and modeling constraints, to allow for assessment and faithful re-computation of the model. It is also essential that these models be accompanied by reliable estimates of uncertainty. In October 2013, a workshop on "Theoretical Model Archiving, Validation and PDBx/mmCIF Data Exchange Format" (<http://www.proteinmodelportal.org/workshop-2013/>) was hosted at Rutgers University to launch development of community standards for theoretical model archiving.

Integrative/Hybrid Structure Modeling

Motivation

Samples of many biological macromolecules prove recalcitrant to mainstream structural biology methods (i.e., crystallography, NMR, and 3DEM), because they are not crystallizable, are insoluble, are not of adequate purity, are conformationally heterogeneous, are too large or small, or do not remain intact during the course of the experiment. In such cases, integrative modeling is increasingly being used to compute structural models based on complementary experimental data and theoretical information (Figures 1 and 2; Table 1) (Alber et al., 2007, 2008; Robinson et al., 2007; Russel et al., 2012; Sali et al., 2003, 1990; Schneidman-Duhovny et al., 2014; Ward et al., 2013). Structural biology is no stranger to integrative models. Insights into the molecular details of the B-DNA double helix (Watson and Crick, 1953), the α helix, and the β sheet (Pauling et al., 1951) all depended on constructing structural models based on data derived from multiple sources (albeit without the benefit of digital computation). Integrative structure modeling of today has its origins in attempts to fit X-ray derived substructures into an EM density map of a larger assembly (Rayment et al., 1993). Other early examples include the model of the Gla-EGF domains from coagulation Factor X based on NMR and SAS data (Sunnerhagen et al., 1996), and the superhelical assembly of the bacteriophage fd gene 5 protein with single-stranded DNA based on neutron

and X-ray SAS data, EM data, and the crystal structure of G5P (Olah et al., 1995); the latter study was inspired in part by molecular dynamics simulations guided by contacts from an NMR structure of the G5P dimer and EM data (Folmer et al., 1994).

Beyond overcoming sample limitations, the integrative approach has several additional advantages (Alber et al., 2007). First, synergy among the input data minimizes the drawbacks of sparse, noisy, and ambiguous data obtained from compositionally and structurally heterogeneous samples. Each individual piece of data may contain relatively little structural information, but by simultaneously fitting a model to all data derived from independent experiments, the uncertainty of the structures that fit the data can be markedly reduced. Second, the integrative approach can be used to produce all structural models consistent with available data, instead of myopically focusing on just one model. Third, comparison of an ensemble of structural models permits estimation of precision and, sometimes, the accuracy of both the experimental data and the model. Fourth, the integrative approach can make structural biologists more efficient by identifying which additional measurements are likely to have the greatest impact on integrative model precision and accuracy. Finally, integrative modeling provides a framework for considering perturbations of the system that are often required to collect the data; for example, spin labels are required for electron paramagnetic resonance experiments, membrane proteins are often reconstituted in micelles for NMR spectroscopy, and point mutations or even entire domains are introduced to stabilize preferred conformations for crystallization. While such perturbations complicate structural analysis, integrative modeling may allow us to distinguish biologically relevant states from artifacts of any individual approach. In summary, integrative structure determination maximizes the accuracy, precision, completeness, and efficiency of the structural coverage of biomolecular systems.

Experimental and Computational Methods for Generating Structural Information

Input information for integrative modeling can come from various experimental methods, physical theories, and statistical analyses of databases of known structures, biopolymer sequences, and interactions. These methods probe different structural aspects of the system (Table 1). In addition to information about average structures, numerous methods provide insights into dynamics of the system, which can also be incorporated into integrative modeling procedures (Russel et al., 2009). For example, both NMR spectroscopy and X-ray crystallography provide access to various measures of conformational dynamics; FRET, time-dependent double electron-electron resonance (DEER) spectroscopies, and even quantitative crosslinking/mass spectrometry (Fischer et al., 2013) can map distance changes in time; small-angle X-ray scattering (SAXS) can provide time-resolved information on the structures and processes with the temporal resolution of a millisecond; molecular dynamics simulations can map the dynamics of an atomic structure up to the millisecond timescale; and high-speed atomic force microscopy imaging can provide the dynamic live images of single molecules (Ando, 2014).

Approach

All structural characterization approaches correspond to finding models that best fit input information, as judged by use of

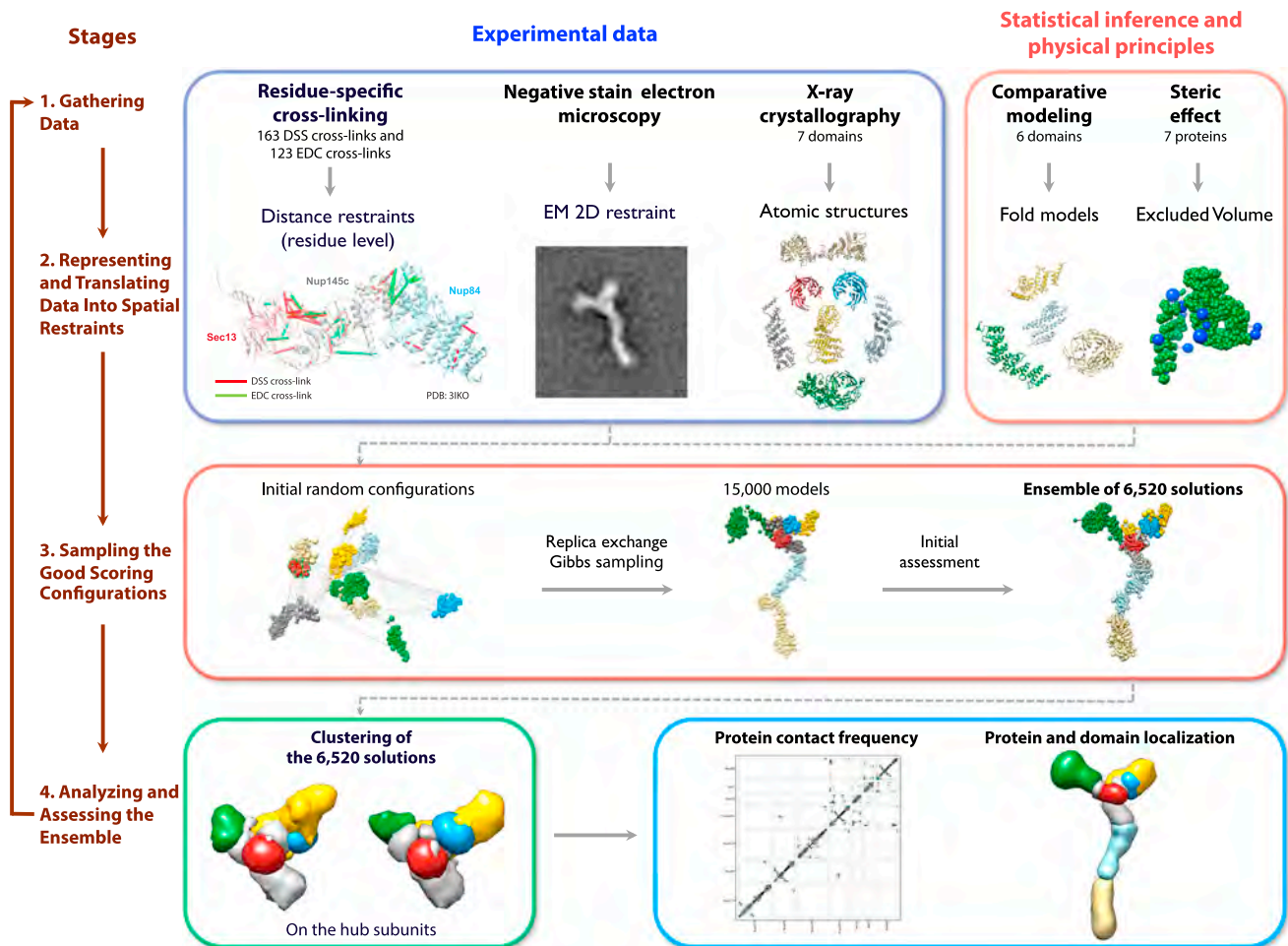


Figure 2. The Four Stages of Integrative Structure Determination

The approach is illustrated by its application to the heptameric Nup84 subcomplex of the nuclear pore complex (Shi et al., 2014).

a scoring function quantifying the difference between the observed data and the data computed from the model. Thus, any information about a structure determination target must always be converted to an explicit structural model through computation. Integrative approaches explicitly combine diverse experimental and theoretical information, with the goal of increasing accuracy, precision, coverage, and efficiency of structure determination. Input information can vary greatly in terms of resolution (i.e., precision, noise, uncertainty), accuracy, and quantity. All structure determination methods are integrative, albeit with differences in degree. At one end of the spectrum, even structure determination using predominantly crystallographic, NMR, or high-resolution single-particle EM data also generally requires a molecular mechanics force field description of atomic structure. At the other end of the spectrum, integrative methods rely more evenly on different types of information, often resulting in coarser models with higher uncertainty (Figure 1). Examples of such integrative methods include docking of comparative models of subunits into a 3DEM density map of the macromolecular assembly (Lasker et al., 2009); rigid-body fitting of multi-domain structures and complexes determined by crystallography or NMR to SAS data (Petoukhov and Svergun, 2005);

and use of conformational sampling methods with sparse NMR data (Lange et al., 2012; Mueller et al., 2000), chemical crosslinks (Young et al., 2000), or even chemical shift data alone (Shen et al., 2008). It is not difficult to appreciate how integrative methods blur distinctions between models based primarily on theoretical considerations and those based primarily on experimental measurements from a physical sample.

The practice of integrative structure determination is iterative, consisting of four stages (Figure 2): gathering of data; choosing the representation and encoding of all data within a numerical scoring function consisting of spatial restraints; configurational sampling to identify structural models with good scores; and analyzing the models, including quantifying agreement with input spatial restraints and estimating model uncertainty. Input information about the system can be used to (1) select the set of variables that best represent the system (system representation), (2) rank the different configurations (scoring function), (3) search for good-scoring solutions (sampling); and (4) further filter good-scoring solutions produced by sampling.

Types of Integrative Models

A structural model of a macromolecular assembly is defined by the relative positions and orientations of its components

(e.g., atoms, pseudo-atoms, residues, secondary structure elements, domains, subunits, and subcomplexes). While traditional structural biology methods usually produce a single atomistic model, integrative models tend to be more complex in at least four respects. First, a model can be multi-scale (Grime and Voth, 2014), representing different levels of structural detail by a collection of geometrical primitives (e.g., points, spheres, tubes, 3D Gaussians, or probability densities). Thus, the same part of a system can be described with multiple representations and different parts of a system can be represented differently. An optimal representation facilitates accurate formulation of spatial restraints together with efficient and complete sampling of good-scoring solutions, while retaining sufficient detail (without over fitting) such that the resulting models are maximally useful for subsequent biological analysis (Schneidman-Duhovny et al., 2014). Second, a model can be multi-state, specifying multiple discrete states of the system required to explain the input information (each state may differ in structure, composition, or both) (Molnar et al., 2014; Pelikan et al., 2009). Third, a model can also specify the order of states in time and/or transitions between the states. This feature allows representation of a multi-step biological process, a functional cycle (Diez et al., 2004), a kinetic network (Pirchi et al., 2011), time evolution of a system (e.g., a molecular dynamics trajectory) (Bock et al., 2013), or FRET trajectories; for a comprehensive description of biomolecular function, it is essential to register state lifetimes, characteristic relaxation times, and direct rate constants. Finally, an ensemble of models may be provided to underscore the uncertainty in the input information, with each individual model satisfying the input information within an acceptable threshold (e.g., NMR-derived ensembles currently available in the PDB [Clore and Gronenborn, 1991; Snyder et al., 2005, 2014] and the ensembles generated from SAXS [Tria et al., 2015]). This aspect of the representation allows us to describe model uncertainty and to assess the completeness of input information; such ensembles are distinct from multiple states that represent actual variations in the structure, as implied by experimental information that cannot be accounted for by a single representative structure (Schneidman-Duhovny et al., 2014; Schröder, 2015).

Task Force Deliberations and Recommendations Charge to the Task Force

A healthy debate is under way about how to classify structural models. A major motivation for this discussion is the lack of accurate general methods to assess the precision and accuracy of any model. As a result, models are often classified based on the predominant type of information used to compute them, which in turn tends to reflect the data-to-parameter ratio and thus model accuracy. However, as previously discussed, all structures are in fact integrative models that have been derived both from experimental measurements involving a physical sample of a biological macromolecule and prior knowledge of the underlying stereochemistry. It is therefore difficult, if not impossible, to draw definitive lines on the spectrum ranging from very well-determined ultra-high-resolution crystallographic structures (>40 experimental observations per non-hydrogen atom in the crystallographic asymmetric unit) and structural models based on a single or even no experimental observation.

Reflecting this debate about model classification, there are in principle several possibilities for archiving the models and associated data among distinct, publicly accessible model/data repositories, including: (1) a single mega-archive that serves as the repository for every type of structural model and data; (2) independent, free-standing repositories that house distinct types of models and data; and (3) a federated system of inter-operating repositories that archive models and data, with “spheres of influence” based on community consensus.

To address some of the challenges ahead and make recommendations about how best to proceed, the community stakeholders who assembled at the October 2014 meeting of the wwPDB Hybrid/Integrative Methods Task Force were divided into three discussion groups, each tasked with considering a series of related questions. What experimental data (beyond crystallography, NMR, and 3DEM) should be archived? Where and how should it be validated? What kinds of non-atomistic models can we expect and how should they be validated? What are the criteria for deciding where models should be archived? How should non-atomistic and mixed atomistic/non-atomistic models be archived? Should there be a separate archive for integrative (mixed) models (and data)? Should we establish a federated system of data and model archives to support integrative structural biology? The three breakout groups were asked to address these questions, report back with their findings, and make recommendations for the future. Each group independently approached the same set of questions. At the close of the meeting, the teams converged to compare notes, identify areas of commonality and diversity, and determine how best to move forward. The resulting consensus is reflected in this document.

Recommendations

Recommendation 1. In addition to archiving the models themselves, all relevant experimental data and metadata as well as experimental and computational protocols should be archived; inclusivity is key.

Ideally, structural models of any kind, derived by any method, should be archived.

Models are of greatest value when they are independently tested, potentially improved, and serve to further our understanding of how the function of a biological system is determined by its 3D structure(s). Therefore, models and necessary annotations must be freely available to the research community. The modeling process should be reproducible. Information concerning all aspects of a model should be deposited, including input data, corresponding spatial restraints, output models, and protocols used to convert input data into models. In addition to the input experimental data, the archival deposition should specify or include theoretically derived restraints used to compute the model (e.g., a statistical potential and a molecular mechanics force field). In practice, frequently used data types (e.g., distance information) should be prioritized for early complete implementation. Uncertainty in the input data needs to be well documented; some data uncertainty estimates may require modeling (e.g., Bayesian error estimates [Rieping et al., 2005]). Consistency between input data and the structural model should be documented as part of model validation.

Each expert community should drive decisions as to how much raw data, processed data, and metadata to deposit,

subject to the minimal requirement that the spatial restraints used for modeling must be derivable from the deposited information. Attention needs to be paid to annotating measurement conditions, such as temperature (Fenwick et al., 2014), sample concentration, environmental conditions (e.g., buffer), construct definition, and identification of all assembly components, all of which can significantly influence the experimental outcome. Cost-benefit analyses should be used to help guide which data should be archived. As much data as practical should be deposited, to facilitate model validation, future improvements of the model, and methods development (e.g., benchmarking sets). Of particular importance will be availability of some raw data to help drive improvement of data processing methods and for use by methods developers, who are often not generating the experimental data themselves.

Recommendation 2. A flexible model representation needs to be developed, allowing for multi-scale models, multi-state models, ensembles of models, and models related by time or other order.

Model representation should allow for as many types of “structural” models as possible, thereby encouraging collaboration among developers of integrative modeling software (Russel et al., 2012). At a minimum, the model representation should allow encoding of an ensemble of multi-scale, multi-state, time-ordered models (see the section on Types of Integrative Models). Uncertainty of the model coordinates should be tightly associated with the model coordinates in the model representation. Any model resident within an archive should be “self-contained” to facilitate utilization (e.g., for visualization). A common representation and format for models are useful for reasons of software interoperability. Particle-based representations/primitives need to be prioritized; non-particle-based model representations (e.g., continuum representations) merit further consideration by appropriate community stakeholders.

Recommendation 3. Procedures for estimating the uncertainty of integrative models should be developed, validated, and adopted.

Assessment of both an integrative model and the information on which it is based is of critical importance for guiding subsequent use of the model. For atomistic models, extant standard validation criteria from X-ray crystallography should be used. Beyond this test, validation of integrative models and data is a major research challenge that must be addressed and overcome. The following represent promising considerations (Alber et al., 2007; Schneidman-Duhovny et al., 2014): convergence of conformational sampling, fit of the model to the input information, test for clashes between geometrical primitives comprising the model, precision of the ensemble of solutions (visualized with, e.g., ribbon plots), cross-validation and statistical bootstrapping based on available data, tests based on data determined after the model was computed, and sensitivity analysis of the model to input data. Bayesian approaches may be particularly well suited to describe model uncertainty by computing posterior model densities from a forward model, noise model, and priors (Muschiello et al., 2008; Rieping et al., 2005). Tools for visualizing model validation should be developed.

Communities generating data used in integrative modeling should agree on the standard set of descriptors for data quality, as has been done for crystallography, NMR, and 3DEM.

Recommendation 4. A federated system of model and data archives should be created.

Integrative models can be based on a broad array of different experimental and computational techniques. While the specific spatial restraints implied by the data and used to construct an integrative model should be deposited with the model itself, the underlying experimental data often contain much richer information. This information should be captured in a federated system of domain-specific model and data archives. These individual member archives should be developed by community experts, based on method-specific standards for data archiving and validation. A federated system of model and data archives implies the need for a seamless exchange of information between independent archives. This seamless exchange requires a common dictionary of terms, agreed data formats, persistent and stable data object identifiers, and close synchronization of policies and procedures. Federated model and data archives need to develop efficient methods for data exchange to allow for transparent data access across the enterprise.

A single interface for the deposition of all data and models into the federated system is highly desirable. Such an interface would greatly facilitate the task of the depositor and, thereby, maximize compliance with deposition standards and requirements. In addition, reliance on a single entry point will help to ensure consistency across the federation at the time of deposition. Following successful deposition, individual datasets can be transferred to member databases for data curation and archiving if domain-specific databases exist. There should also be provision for collecting unstructured information in a “data commons,” as proposed by the data science initiative at the NIH (Margolis et al., 2014).

Access to the contents of the federated database through a single portal is also most desirable, to facilitate dissemination of data, models, and experimental/computational protocols.

Of particular importance for integrative modeling will be the option to modify or update any aspect of the modeling procedure, for example, by adding new data. The federated archive should allow versioning for each deposited model. Such capabilities will facilitate the cycle of experiment and modeling, and accelerate production of more accurate, precise, and complete models (Russel et al., 2012).

Recommendation 5. Publication standards for integrative models should be established.

Over the past decade, the wwPDB organization has worked with relevant scientific journals to help establish publication standards for structural models coming from crystallography, NMR spectroscopy, and 3DEM. Community standards now include requiring authors to make their validation reports available to reviewers and editors. Through the International Union of Crystallography Small Angle Scattering and Journals Commissions, the SAS community developed and agreed upon publication guidelines for structural modeling of biomolecules therefrom (Jacques et al., 2012). A set of standards for publishing integrative models should be developed along similar lines.

Implementation

Implementation of Recommendation 1 poses a host of cultural and technical challenges. Experimentalists and modelers need to provide the data, models, and protocols, thus at least partly addressing increasing concerns regarding reproducibility of

scientific results. From a technical perspective, inter-operating data dictionaries for all methods need to be created. In addition, potential storage bottlenecks need to be addressed.

Implementation of Recommendations 2 and 3 will require significant research as to how best to represent and validate the many different kinds of integrative models. In addition, the community will need to agree on a common set of standards that are sufficiently mutable to allow for future innovation. Efforts such as the “Cryo-EM Modeling Challenge” may facilitate this process (http://www.emdatabank.org/modeling_chllnge).

Implementation of Recommendation 4 will require agreement on a common data exchange system among member repositories. Based on past accomplishments, the wwPDB is well positioned to play a leadership role in establishing the proposed federated system, including provision of common deposition and access interfaces. The wwPDB should begin this process by providing training and advice on data archiving and curation to contributing domain-specific member repositories.

Implementation of Recommendation 5 will require continued work with the journals that publish structural models of biological macromolecules.

Significant resources will be required to implement these recommendations, including grants for research, infrastructure, and workshops. These efforts are international by their very nature and will require funding from multiple public and private sources, including in North America, Europe, and Asia.

ACKNOWLEDGMENTS

The workshop was supported by funding to PDBe by Wellcome Trust 088944; RCSB PDB by NSF DBI 1338415; PDBj by JST-NBDC; BMRB by NLM P41 LM05799; EMDatabank by NIH GM079429; and tax-deductible donations made to the wwPDB Foundation in support of wwPDB outreach activities.

REFERENCES

- Alber, F., Dokudovskaya, S., Veenhoff, L., Zhang, W., Kipper, J., Devos, D., Suprpto, A., Karni-Schmidt, O., Williams, R., Chait, B., et al. (2007). Determining the architectures of macromolecular assemblies. *Nature* *450*, 683–694.
- Alber, F., Förster, F., Korkin, D., Topf, M., and Sali, A. (2008). Integrating diverse data for structure determination of macromolecular assemblies. *Annu. Rev. Biochem.* *77*, 443–477.
- Ando, T. (2014). High-speed AFM imaging. *Curr. Opin. Struct. Biol.* *28*, 63–68.
- Arnold, K., Kiefer, F., Kopp, J., Battey, J.N., Podvinec, M., Westbrook, J.D., Berman, H.M., Bordoli, L., and Schwede, T. (2009). The Protein Model Portal. *J. Struct. Funct. Genomics* *10*, 1–8.
- Bau, D., Sanyal, A., Lajoie, B.R., Capriotti, E., Byron, M., Lawrence, J.B., Dekker, J., and Marti-Renom, M.A. (2011). The three-dimensional folding of the alpha-globin gene domain reveals formation of chromatin globules. *Nat. Struct. Mol. Biol.* *18*, 107–114.
- Berman, H.M., Henrick, K., and Nakamura, H. (2003). Announcing the world-wide Protein Data Bank. *Nat. Struct. Biol.* *10*, 980.
- Berman, H.M., Burley, S.K., Chiu, W., Sali, A., Adzhubei, A., Bourne, P.E., Bryant, S.H., Dunbrack, R.L., Jr., Fidelis, K., Frank, J., et al. (2006). Outcome of a workshop on archiving structural models of biological macromolecules. *Structure* *14*, 1211–1217.
- Berman, H.M., Westbrook, J.D., Gabanyi, M.J., Tao, W., Shah, R., Kouranov, A., Schwede, T., Arnold, K., Kiefer, F., Bordoli, L., et al. (2009). The protein structure initiative structural genomics knowledgebase. *Nucleic Acids Res.* *37*, D365–D368.
- Bock, L.V., Blau, C., Schröder, G.F., Davydov, I.I., Fischer, N., Stark, H., Rodnina, M.V., Vaiana, A.C., and Grubmüller, H. (2013). Energy barriers and driving forces in tRNA translocation through the ribosome. *Nat. Struct. Mol. Biol.* *20*, 1390–1396.
- Boura, E., Rozycki, B., Herrick, D.Z., Chung, H.S., Vecer, J., Eaton, W.A., Cafiso, D.S., Hummer, G., and Hurley, J.H. (2011). Solution structure of the ESCRT-I complex by small-angle X-ray scattering, EPR, and FRET spectroscopy. *Proc. Natl. Acad. Sci. USA* *108*, 9437–9442.
- Chen, Z.A., Jawhari, A., Fischer, L., Buchen, C., Tahir, S., Kamenski, T., Rasmussen, M., Larivière, L., Bukowski-Wills, J.C., Nilges, M., et al. (2010). Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry. *EMBO J.* *29*, 717–726.
- Clare, G.M., and Gronenborn, A.M. (1991). Structures of larger proteins in solution: three- and four-dimensional heteronuclear NMR spectroscopy. *Science* *252*, 1390–1399.
- Degiacomi, M.T., and Dal Peraro, M. (2013). Macromolecular symmetric assembly prediction using swarm intelligence dynamic modeling. *Structure* *21*, 1097–1106.
- Degiacomi, M.T., Iacovache, I., Pernot, L., Chami, M., Kudryashev, M., Stahlberg, H., van der Goot, F.G., and Dal Peraro, M. (2013). Molecular assembly of the aerolysin pore reveals a swirling membrane-insertion mechanism. *Nat. Chem. Biol.* *9*, 623–629.
- Deshmukh, L., Schwieters, C.D., Grishaev, A., Ghirlando, R., Baber, J.L., and Clare, G.M. (2013). Structure and dynamics of full-length HIV-1 capsid protein in solution. *J. Am. Chem. Soc.* *135*, 16133–16147.
- Diez, M., Zimmermann, B., Börsch, M., König, M., Schweinberger, E., Steigmüller, S., Reuter, R., Felekyan, S., Kudryavtsev, V., Seidel, C.A.M., and Gräber, P. (2004). Proton-powered subunit rotation in single membrane-bound F₀F₁-ATP synthase. *Nat. Struct. Mol. Biol.* *11*, 135–141.
- Fenwick, R.B., van den Bedem, H., Fraser, J.S., and Wright, P.E. (2014). Integrated description of protein dynamics from room-temperature X-ray crystallography and NMR. *Proc. Natl. Acad. Sci. USA* *111*, E445–E454.
- Fischer, L., Chen, Z.A., and Rappsilber, J. (2013). Quantitative cross-linking/mass spectrometry using isotope-labelled cross-linkers. *J. Proteomics* *88*, 120–128.
- Folmer, R.H., Nilges, M., Folkers, P.J., Konings, R.N., and Hilbers, C.W. (1994). A model of the complex between single-stranded DNA and the single-stranded DNA binding protein encoded by gene V of filamentous bacteriophage M13. *J. Mol. Biol.* *240*, 341–357.
- Gabanyi, M.J., Adams, P.D., Arnold, K., Bordoli, L., Carter, L.G., Flippen-Andersen, J., Gifford, L., Haas, J., Kouranov, A., McLaughlin, W.A., et al. (2011). The Structural Biology Knowledgebase: a portal to protein structures, sequences, functions, and methods. *J. Struct. Funct. Genomics* *12*, 45–54.
- Gong, Z., Schwieters, C.D., and Tang, C. (2015). Conjoined use of EM and NMR in RNA structure refinement. *PLoS One* *10*, e0120445.
- Grime, J.M.A., and Voth, G.A. (2014). Highly scalable and memory efficient ultra-coarse-grained molecular dynamics simulations. *J. Chem. Theory Comp.* *10*, 423–431.
- Haas, J., Roth, S., Arnold, K., Kiefer, F., Schmidt, T., Bordoli, L., and Schwede, T. (2013). The Protein Model Portal—a comprehensive resource for protein structure and model information. *Database* *2013*, bat031.
- Han, Y., Luo, J., Ranish, J., and Hahn, S. (2014). Architecture of the *Saccharomyces cerevisiae* SAGA transcription coactivator complex. *EMBO J.* *33*, 2534–2546.
- Henderson, R., Sali, A., Baker, M.L., Carragher, B., Devkota, B., Downing, K.H., Egelman, E.H., Feng, Z., Frank, J., Grigorieff, N., et al. (2012). Outcome of the first Electron Microscopy Validation Task Force meeting. *Structure* *20*, 205–214.
- Jacques, D.A., Guss, J.M., Svergun, D.I., and Trehwella, J. (2012). Publication guidelines for structural modelling of small-angle scattering data from biomolecules in solution. *Acta Crystallogr. D Biol. Crystallogr.* *68*, 620–626.
- Kalhor, R., Tjong, H., Jayatilaka, N., Alber, F., and Chen, L. (2012). Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.* *30*, 90–98.

- Kalinin, S., Peulen, T., Sindbert, S., Rothwell, P.J., Berger, S., Restle, T., Goody, R.S., Gohlke, H., and Seidel, C.A.M. (2012). A toolkit and benchmark study for FRET-restrained high-precision structural modeling. *Nat. Methods* **9**, 1218–1225.
- Lange, O.F., Rossi, P., Sgourakis, N.G., Song, Y., Lee, H.W., Aramini, J.M., Ertekin, A., Xiao, R., Acton, T.B., Montelione, G.T., and Baker, D. (2012). Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. *Proc. Natl. Acad. Sci. USA* **109**, 10873–10878.
- Lasker, K., Topf, M., Sali, A., and Wolfson, H.J. (2009). Inferential optimization for simultaneous fitting of multiple components into a CryoEM map of their assembly. *J. Mol. Biol.* **388**, 180–194.
- Lawson, C.L., Baker, M.L., Best, C., Bi, C., Dougherty, M., Feng, P., van Ginkel, G., Devkota, B., Lagerstedt, I., Ludtke, S.J., et al. (2011). EMDatabank.org: unified data resource for CryoEM. *Nucleic Acids Res.* **39**, D456–D464.
- Loquet, A., Sgourakis, N.G., Gupta, R., Giller, K., Riedel, D., Goosmann, C., Griesinger, C., Kolbe, M., Baker, D., Becker, S., and Lange, A. (2012). Atomic model of the type III secretion system needle. *Nature* **486**, 276–279.
- Lukoyanova, N., Kondos, S.C., Farabella, I., Law, R.H., Reboul, C.F., Caradoc-Davies, T.T., Spicer, B.A., Kleifeld, O., Traore, D.A., Ekkel, S.M., et al. (2015). Conformational changes during pore formation by the perforin-related protein pleurotolysin. *PLoS Biol.* **13**, e1002049.
- Malfois, M., and Svergun, D. (2000). sasCIF: an extension of core crystallographic information file for SAS. *J. App. Cryst.* **33**, 812–816.
- Margolis, R., Derr, L., Dunn, M., Huerta, M., Larkin, J., Sheehan, J., Guyer, M., and Green, E.D. (2014). The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. *J. Am. Med. Inform. Assoc.* **21**, 957–958.
- Markley, J.L., Ulrich, E.L., Berman, H.M., Henrick, K., Nakamura, H., and Akutsu, H. (2008). BioMagResBank (BMRB) as a partner in the Worldwide Protein Data Bank (wwPDB): new policies affecting biomolecular NMR depositions. *J. Biomol. NMR* **40**, 153–155.
- Marti-Renom, M.A., Stuart, A.C., Fiser, A., Sanchez, R., Melo, F., and Sali, A. (2000). Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 291–325.
- Mekler, V., Kortkhonjia, E., Mukhopadhyay, J., Knight, J., Revyakin, A., Kapandis, A.N., Niu, W., Ebright, Y.W., Levy, R., and Ebright, R.H. (2002). Structural organization of bacterial RNA polymerase holoenzyme and the RNA polymerase-promoter open complex. *Cell* **108**, 599–614.
- Miyazaki, Y., Irobalieva, R.N., Tolbert, B.S., Smalls-Mantey, A., Iyalla, K., Loeliger, K., D'Souza, V., Khant, H., Schmid, M.F., Garcia, E.L., et al. (2010). Structure of a conserved retroviral RNA packaging element by NMR spectroscopy and cryo-electron tomography. *J. Mol. Biol.* **404**, 751–772.
- Molnar, K.S., Bonomi, M., Pellarin, R., Clinthorne, G.D., Gonzalez, G., Goldberg, S.D., Goulian, M., Sali, A., and DeGrado, W.F. (2014). Cys-scanning disulfide crosslinking and the bayesian modeling probe the transmembrane signaling mechanism of the histidine kinase, PhoQ. *Structure* **22**, 1239–1251.
- Montelione, G.T., Nilges, M., Bax, A., Güntert, P., Herrmann, T., Markley, J.L., Richardson, J., Schwieters, C., Vuister, G.W., Vranken, W., and Wishart, D. (2013). Recommendations of the wwPDB NMR Structure Validation Task Force. *Structure* **21**, 1563–1570.
- Mueller, G.A., Choy, W.Y., Yang, D., Forman-Kay, J.D., Venters, R.A., and Kay, L.E. (2000). Global folds of proteins with low densities of NOEs using residual dipolar couplings: application to the 370-residue maltodextrin-binding protein. *J. Mol. Biol.* **300**, 197–212.
- Muschielok, A., Andrecka, J., Jawhari, A., Bruckner, F., Cramer, P., and Michaelis, J. (2008). A nano-positioning system for macromolecular structural analysis. *Nat. Methods* **5**, 965–971.
- Olah, G.A., Gray, D.M., Gray, C.W., Kergil, D.L., Sosnick, T.R., Mark, B.L., Vaughan, M.R., and Trewthella, J. (1995). Structures of fd gene 5 protein-nucleic acid complexes: a combined solution scattering and electron microscopy study. *J. Mol. Biol.* **249**, 576–594.
- Pauling, L., Corey, R.B., and Branson, H.R. (1951). The structures of proteins. *Proc. Natl. Acad. Sci. USA* **37**, 205.
- Pelikan, M., Hura, G.L., and Hammel, M. (2009). Structure and flexibility within proteins as identified through small angle X-ray scattering. *Gen. Physiol. Biophys.* **28**, 174–189.
- Petoukhov, M.V., and Svergun, D.I. (2005). Global rigid body modeling of macromolecular complexes against small-angle scattering data. *Biophys. J.* **89**, 1237–1250.
- Pirchi, M., Ziv, G., Riven, I., Cohen, S.S., Zohar, N., Barak, Y., and Haran, G. (2011). Single-molecule fluorescence spectroscopy maps the folding landscape of a large protein. *Nat. Commun.* **2**, 493.
- Politis, A., Stengel, F., Hall, Z., Hernandez, H., Leitner, A., Walzthoeni, T., Robinson, C.V., and Aebersold, R. (2014). A mass spectrometry-based hybrid method for structural modeling of protein complexes. *Nat. Methods* **11**, 403–406.
- Prischi, F., Konarev, P.V., Iannuzzi, C., Pastore, C., Adinolfi, S., Martin, S.R., Svergun, D.I., and Pastore, A. (2010). Structural bases for the interaction of frataxin with the central components of iron-sulphur cluster assembly. *Nat. Commun.* **1**, 95.
- Protein Data Bank. (1971). Protein Data Bank. *Nat. New Biol.* **233**, 223.
- Rayment, I., Holden, H.M., Whittaker, M., Yohn, C.B., Lorenz, M., Holmes, K.C., and Milligan, R.A. (1993). Structure of the actin-myosin complex and its implications for muscle contraction. *Science* **261**, 58–65.
- Read, R.J., Adams, P.D., Arendall, W.B., III, Brunger, A.T., Emsley, P., Joosten, R.P., Kleywegt, G.J., Krissinel, E.B., Lutteke, T., Otwinowski, Z., et al. (2011). A new generation of crystallographic validation tools for the Protein Data Bank. *Structure* **19**, 1395–1412.
- Rieping, W., Habeck, M., and Nilges, M. (2005). Inferential structure determination. *Science* **309**, 303–306.
- Robinson, C.V., Sali, A., and Baumeister, W. (2007). The molecular sociology of the cell. *Nature* **450**, 973–982.
- Russel, D., Lasker, K., Phillips, J., Schneidman-Duhovny, D., Velazquez-Muriel, J., and Sali, A. (2009). The structural dynamics of macromolecular processes. *Curr. Opin. Cell Biol.* **21**, 97–108.
- Russel, D., Lasker, K., Webb, B., Velazquez-Muriel, J., Tjioe, E., Schneidman-Duhovny, D., Peterson, B., and Sali, A. (2012). Putting the pieces together: integrative structure determination of macromolecular assemblies. *PLoS Biol.* **10**, e1001244.
- Sali, A., Overington, J.P., Johnson, M.S., and Blundell, T.L. (1990). From comparisons of protein sequences and structures to protein modelling and design. *Trends Biochem. Sci.* **15**, 235–240.
- Sali, A., Glaeser, R., Earnest, T., and Baumeister, W. (2003). From words to literature in structural proteomics. *Nature* **422**, 216–225.
- Schneidman-Duhovny, D., Pellarin, R., and Sali, A. (2014). Uncertainty in integrative structural modeling. *Curr. Opin. Struct. Biol.* **28**, 96–104.
- Schröder, G.F. (2015). Hybrid methods for macromolecular structure determination: experiment with expectations. *Curr. Opin. Struct. Biol.* **31**, 20–27.
- Schwede, T., Sali, A., Honig, B., Levitt, M., Berman, H.M., Jones, D., Brenner, S.E., Burley, S.K., Das, R., Dokholyan, N.V., et al. (2009). Outcome of a workshop on applications of protein models in biomedical research. *Structure* **17**, 151–159.
- Shen, Y., Lange, O., Delaglio, F., Rossi, P., Aramini, J.M., Liu, G., Eletsky, A., Wu, Y., Singarapu, K.K., Lemak, A., et al. (2008). Consistent blind protein structure generation from NMR chemical shift data. *Proc. Natl. Acad. Sci. USA* **105**, 4685–4690.
- Shi, Y., Fernandez-Martinez, J., Tjioe, E., Pellarin, R., Kim, S.J., Williams, R., Schneidman, D., Sali, A., Rout, M.P., and Chait, B.T. (2014). Structural characterization by cross-linking reveals the detailed architecture of a coatomer-related heptameric module from the nuclear pore complex. *Mol. Cell. Proteomics* **13**, 2927–2943.
- Snijder, J., Burnley, R.J., Wiegand, A., Melquiond, A.S.J., Bonvin, A.M.J.J., Axmann, I.M., and Heck, A.J.R. (2014). Insight into cyanobacterial circadian timing from structural details of the KaiB-KaiC interaction. *Proc. Natl. Acad. Sci. USA* **111**, 1379–1383.

Snyder, D.A., Bhattacharya, A., Huang, Y.J., and Montelione, G.T. (2005). Assessing precision and accuracy of protein structures derived from NMR data. *Proteins* 59, 655–661.

Snyder, D.A., Grullon, J., Huang, Y.J., Tejero, R., and Montelione, G.T. (2014). The expanded FindCore method for identification of a core atom set for assessment of protein structure prediction. *Proteins* 82 (Suppl 2), 219–230.

Sunnerhagen, M., Olah, G.A., Stenflo, J., Forsen, S., Drakenberg, T., and Trehwella, J. (1996). The relative orientation of Gla and EGF domains in coagulation factor X is altered by Ca²⁺ binding to the first EGF domain. A combined NMR-small angle X-ray scattering study. *Biochemistry* 35, 11547–11559.

Trehwella, J., Hendrickson, W.A., Kleywegt, G.J., Sali, A., Sato, M., Schwede, T., Svergun, D.I., Tainer, J.A., Westbrook, J., and Berman, H.M. (2013). Report of the wwPDB Small-Angle Scattering Task Force: data requirements for biomolecular modeling and the PDB. *Structure* 21, 875–881.

Tria, G., Mertens, H.D.T., Kachala, M., and Svergun, D.I. (2015). Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering. *IUCr* 2, 207–217.

Valentini, E., Kikhney, A.G., Previtali, G., Jeffries, C.M., and Svergun, D.I. (2015). SASBDB, a repository for biological small-angle scattering data. *Nucleic Acids Res.* 43, D357–D363.

Ward, A., Sali, A., and Wilson, I. (2013). Integrative structural biology. *Science* 339, 913–915.

Watson, J.D., and Crick, F.H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 171, 737–738.

Whitten, A.E., Jeffries, C.M., Harris, S.P., and Trehwella, J. (2008). Cardiac myosin-binding protein C decorates F-actin: implications for cardiac function. *Proc. Natl. Acad. Sci. USA* 105, 18360–18365.

Young, M.M., Tang, N., Hempel, J.C., Oshiro, C.M., Taylor, E.W., Kuntz, I.D., Gibson, B.W., and Dollinger, G. (2000). High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. *Proc. Natl. Acad. Sci. USA* 97, 5802–5806.

Zhang, Y., Feng, Y., Chatterjee, S., Tuske, S., Ho, M.X., Arnold, E., and Ebright, R.H. (2012). Structural basis of transcription initiation. *Science* 338, 1076–1080.

Outcome of the First wwPDB/CCDC/D3R Ligand Validation Workshop

Paul D. Adams,¹ Kathleen Aertgeerts,² Cary Bauer,³ Jeffrey A. Bell,⁴ Helen M. Berman,^{5,6} Talapady N. Bhat,⁷ Jeff M. Blaney,⁸ Evan Bolton,⁹ Gerard Bricogne,¹⁰ David Brown,^{11,12} Stephen K. Burley,^{5,6,13,*} David A. Case,⁶ Kirk L. Clark,¹⁴ Tom Darden,¹⁵ Paul Emsley,¹⁶ Victoria A. Feher,^{17,*} Zukang Feng,^{5,6} Colin R. Groom,^{18,*} Seth F. Harris,⁸ Jorg Hendle,¹⁹ Thomas Holder,⁴ Andrzej Joachimiak,²⁰ Gerard J. Kleywegt,²¹

(Author list continued on next page)

¹Molecular Biophysics & Integrated Bioimaging Division, Lawrence Berkeley Laboratory, Department of Bioengineering, UC Berkeley, Berkeley, CA 94720-8235, USA

²DART NeuroScience, LLC, San Diego, CA 92131, USA

³Bruker AXS, Inc., Madison, WI 53711, USA

⁴Schrödinger, Inc., New York, NY 10036, USA

⁵Research Collaboratory for Structural Bioinformatics Protein Data Bank, Center for Integrative Proteomics Research, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

⁶Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

⁷Biosystems and Biomaterials Division, NIST, Gaithersburg, MD 20899, USA

⁸Genentech, Inc., South San Francisco, CA 94080, USA

⁹National Center for Biotechnology Information, U.S. National Library of Medicine, Bethesda, MD 20894, USA

¹⁰Global Phasing Ltd., Cambridge CB3 0AX, UK

¹¹School of Biosciences, University of Kent, Canterbury CT2 7NH, UK

¹²Charles River Ltd., Structural Biology and Biophysics, Cambridge CB10 1XL, UK

¹³Skaggs School of Pharmacy and Pharmaceutical Sciences and San Diego Supercomputer Center, University of California, San Diego, La Jolla, CA 92093, USA

¹⁴Novartis Institutes for BioMedical Research, Cambridge, MA 02139, USA

¹⁵OpenEye Scientific, Cambridge, MA 02142, USA

(Affiliations continued on next page)

Crystallographic studies of ligands bound to biological macromolecules (proteins and nucleic acids) represent an important source of information concerning drug-target interactions, providing atomic level insights into the physical chemistry of complex formation between macromolecules and ligands. Of the more than 115,000 entries extant in the Protein Data Bank (PDB) archive, ~75% include at least one non-polymeric ligand. Ligand geometrical and stereochemical quality, the suitability of ligand models for in silico drug discovery and design, and the goodness-of-fit of ligand models to electron-density maps vary widely across the archive. We describe the proceedings and conclusions from the first Worldwide PDB/Cambridge Crystallographic Data Center/Drug Design Data Resource (wwPDB/CCDC/D3R) Ligand Validation Workshop held at the Research Collaboratory for Structural Bioinformatics at Rutgers University on July 30–31, 2015. Experts in protein crystallography from academe and industry came together with non-profit and for-profit software providers for crystallography and with experts in computational chemistry and data archiving to discuss and make recommendations on best practices, as framed by a series of questions central to structural studies of macromolecule-ligand complexes. What data concerning bound ligands should be archived in the PDB? How should the ligands be best represented? How should structural models of macromolecule-ligand complexes be validated? What supplementary information should accompany publications of structural studies of biological macromolecules? Consensus recommendations on best practices developed in response to each of these questions are provided, together with some details regarding implementation. Important issues addressed but not resolved at the workshop are also enumerated.

Background

The Worldwide PDB (wwPDB; wwpdb.org), the Cambridge Crystallographic Data Center (CCDC; www.ccdc.cam.ac.uk), and the Drug Design Data Resource (D3R; <https://www.drugdesigndata.org>) co-organized a Ligand Validation Workshop on July 30–31 2015 at Rutgers University. The workshop brought together academic and industrial protein crystallographers, providers of software for crystallography, computational

chemists, and experts in data archiving. More than 50 participants from more than 40 organizations discussed and made recommendations on best practices for structural studies of macromolecule-ligand complexes and archiving of the resulting information.

PDB and Historical Context for the Workshop

The PDB was established in 1971 with just seven X-ray crystallographic structures of proteins as the first open-access digital

Tobias Krojer,²² Joseph Marcotrigiano,^{6,23} Alan E. Mark,²⁴ John L. Markley,²⁵ Matthew Miller,²³ Wladek Minor,²⁶ Gaetano T. Montelione,^{23,27} Garib Murshudov,¹⁶ Atsushi Nakagawa,²⁸ Haruki Nakamura,²⁸ Anthony Nicholls,¹⁵ Marc Nicklaus,²⁹ Robert T. Nolte,³⁰ Anil K. Padyana,³¹ Catherine E. Peishoff,³⁰ Susan Pieniazek,³² Randy J. Read,³³ Chenghua Shao,⁵ Steven Sheriff,³⁴ Oliver Smart,²¹ Stephen Soisson,³⁵ John Spurlino,³⁶ Terry Stouch,³⁷ Radka Svobodova,³⁸ Wolfram Tempel,³⁹ Thomas C. Terwilliger,⁴⁰ Dale Tronrud,⁴¹ Sameer Velankar,²¹ Suzanna C. Ward,¹⁸ Gregory L. Warren,¹⁵ John D. Westbrook,^{5,6} Pamela Williams,⁴² and Huanwang Yang,^{5,6} and Jasmine Young^{5,6}

¹⁶MRC Laboratory of Molecular Biology, Cambridge CB2 0QH, UK

¹⁷Drug Design Data Resource and Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, CA 92093, USA

¹⁸Cambridge Crystallographic Data Centre, Cambridge CB2 1EZ, UK

¹⁹Structural Biology, Lilly Biotechnology Center, San Diego, CA 92121, USA

²⁰Structural Biology Center, Biosciences, Argonne National Laboratory, Argonne, IL 60439, USA

²¹Protein Data Bank in Europe, European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK

²²Structural Genomics Consortium, University of Oxford, Oxford OX3 7DQ, UK

²³Center for Advanced Biotechnology and Medicine, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

²⁴School of Chemistry & Molecular Biosciences, University of Queensland, St Lucia, QLD 4072, Australia

²⁵BioMagResBank, Department of Biochemistry, University of Wisconsin-Madison, Madison, WI 53706-1544, USA

²⁶Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA 22908, USA

²⁷Department of Molecular Biology and Biochemistry, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

²⁸Protein Data Bank Japan, Institute for Protein Research, Osaka University, Osaka 565-0871, Japan

²⁹Computer-Aided Drug Design Group, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Frederick, MD 21702, USA

³⁰GlaxoSmithKline, Collegeville, PA 19426, USA

³¹Agios Pharmaceuticals, Inc., Cambridge, MA 02139, USA

³²Bristol-Myers Squibb Research and Development, Pennington, NJ 08534, USA

³³Department of Haematology, Cambridge Institute for Medical Research, University of Cambridge, Cambridge CB2 0XY, UK

³⁴Bristol-Myers Squibb Research and Development, Princeton, NJ 08543, USA

³⁵Merck Research Laboratories, West Point, PA 19486, USA

³⁶Janssen Pharmaceuticals, Inc., Spring House, PA 19002, USA

³⁷Science For Solutions, LLC, West Windsor, NJ 08550, USA

³⁸CEITEC-Central European Institute of Technology and National Centre for Biomolecular Research, Masaryk University Brno, 625 00 Brno, Czech Republic

³⁹Structural Genomics Consortium, University of Toronto, Toronto, ON M5G 1L7, Canada

⁴⁰Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

⁴¹Department of Biochemistry and Biophysics, Oregon State University, Corvallis, OR 97331, USA

⁴²Astex Pharmaceuticals, Cambridge CB4 0QA, UK

*Correspondence: stephen.burley@rcsb.org (S.K.B.), vfeher@ucsd.edu (V.A.F.), groom@ccdc.cam.ac.uk (C.R.G.)

<http://dx.doi.org/10.1016/j.str.2016.02.017>

resource in the biological sciences (Protein Data Bank, 1971 in [Supplemental References](#)). In February 2016, some 44 years later, this sui generis global archive holds more than 115,000 experimentally determined 3D structural models of biological macromolecules and their complexes with a wide variety of ligands. In addition, descriptions of the chemistry of biopolymers and ligands are collected, as are metadata describing sample preparation, experimental methodology, structural model building and refinement statistics, literature references, and so forth. PDB data are made freely available without restrictions on usage. The vast majority of data in the PDB (~90%) come from X-ray, neutron, and combined X-ray/neutron crystallography, with the remainder contributed by two newer 3D structure determination methods: nuclear magnetic resonance (NMR) spectroscopy and electron microscopy (3DEM).

Considerable effort has gone into understanding how best to curate structural models and primary experimental data from X-ray, NMR, and 3DEM. Over the past decade, the wwPDB, the global organization responsible for managing the PDB archive ([Berman et al., 2003](#)), has formed expert, method-specific Validation Task Forces (VTFs) to identify which experimental data and metadata from each structure determination method should be archived and how these data and the atomic level structural models therefrom should be validated. Initially, the

wwPDB X-ray VTF made recommendations on how to best validate crystallographic data (Read et al., 2011 in [Supplemental References](#)). These initial recommendations have been implemented as a validation pipeline used within the wwPDB Deposition and Annotation (D&A) system. A wwPDB Validation Report accompanies every PDB deposition (ftp://ftp.wwpdb.org/pub/pdb/validation_reports/). Preliminary recommendations have also been made by wwPDB VTFs for NMR (Montelione et al., 2013 in [Supplemental References](#)) and 3DEM (Henderson et al., 2012 in [Supplemental References](#)). Implementation of NMR and 3DEM VTF recommendations within the wwPDB D&A validation pipeline is currently underway. It is anticipated that additional validation measures will be implemented within the wwPDB D&A system as new methods are developed and more experience is gained with existing procedures.

Crystallographic Data in the PDB

For structural models determined via X-ray, neutron, and combined X-ray/neutron crystallography methods, together with those determined using electron diffraction from 2D crystals, deposition of experimental data (i.e., diffracted intensities or structure factor amplitudes) into the PDB has been mandatory since 2008 (<http://www.wwpdb.org/news/news?year=2007#29-November-2007>). Validation against deposited structure factor amplitudes is carried out using procedures

recommended by the wwPDB X-ray VTF (Read et al., 2011 in [Supplemental References](#)). wwPDB Validation Reports include graphical summaries of the quality of the overall structural model and residue-specific features. Detailed assessments of various aspects of the structural model, such as agreement with experimental data and chemical expectations, are also provided. In the near future, unmerged intensities will also be collected during PDB deposition, thereby enabling additional validation.

Chemical Component Dictionary

The Chemical Component Dictionary (CCD) was originally developed (Feng et al., 2004) to provide a more expressive alternative to the early PDB ligand descriptions, which were based purely on atom connectivity records. The CCD embraced the data representation for chemical components developed for the Macromolecular Crystallographic Information Framework or mmCIF data dictionary (Fitzgerald et al., 2005). Following a major wwPDB undertaking to standardize nomenclature concluded in 2007 (Henrick et al., 2008 in [Supplemental References](#)), the global organization adopted a common dictionary of chemical definitions. The current CCD (Westbrook et al., 2015 in [Supplemental References](#)) is an extended reference file describing all polymer components and small molecules found in PDB archival entries. This dictionary contains detailed chemical descriptions for standard and modified amino acids/nucleotides, small-molecule ligands, and solvent/solute molecules. Each chemical definition includes descriptions of chemical properties, such as stereochemical assignments, chemical descriptors (SMILES, Weininger, 1988 in [Supplemental References](#); InChI; and InChIKeys, Heller et al., 2013 in [Supplemental References](#)), and systematic chemical names. A set of atomic model coordinates from a selected experimental entry and a computed set of ideal atomic coordinates are provided for each entry in the CCD. Hydrogen atoms are computationally added to the experimental coordinates and unobserved heavy atoms, such as leaving groups specified by depositors, are added to the ideal coordinates if they are not explicitly modeled in the experimental entry. Computed ideal coordinates are obtained from the software tools Corina (Gasteiger et al., 1990) or OpenEye/Omega (Hawkins et al., 2010). Cahn-Ingold-Prelog stereochemical assignments (Cahn et al., 1966) and aromatic annotations are documented for each atom present in each CCD entry. The dictionary is organized by the three-character alphanumeric code that the wwPDB assigns to each chemical component, and updated with each weekly release of the PDB archive (Sen et al., 2014 in [Supplemental References](#)).

A related PDB archive chemical reference dictionary is the Biologically Interesting molecule Reference Dictionary (BIRD) (Dutta et al., 2014; Young et al., 2013 in [Supplemental References](#)), which contains information about peptide-like antibiotic and inhibitor molecules present in the PDB archive. BIRD entries include molecular weight and chemical formula, polymer sequence and connectivity, descriptions of structural features and functional classification, natural source, and external references to corresponding UniProt (UniProt Consortium, 2015 in [Supplemental References](#)) or Norine (Caboche et al., 2008) reference sequences.

A BIRD molecule may be represented in a PDB archival entry as a polymer with sequence information or as a single ligand with chemical information. The preferred representation is specified in the BIRD file, with a representative PDB code. All PDB entries containing the same BIRD molecule or its analog(s) are represented uniformly. An important feature of BIRD is to provide dual representation, both sequence and chemical information is provided, regardless of whether the molecule is represented as a polymer or a ligand in the PDB archive.

Current Validation of Macromolecule-Ligand Complexes

The initial recommendations of the wwPDB X-ray VTF (Read et al., 2011 in [Supplemental References](#)) have been implemented in a software pipeline (Gore et al., 2012) embedded within the wwPDB D&A system. Officially watermarked wwPDB Validation Reports are provided to PDB contributors at the time of deposition. An increasing number of journals require that these reports accompany manuscripts reporting structural studies of biological macromolecules. Structural biologists can obtain a similar report using the wwPDB Validation Server (<http://wwpdb-validation.wwpdb.org/>) prior to deposition. For ligands, the wwPDB Validation Report includes both geometrical and model fit diagnostic information. Bond lengths and angles, acyclic torsion angles, and ring systems are assessed (Bruno et al., 2004) by comparison with preferred molecular geometries derived from high-quality, small-molecule structures in the Cambridge Structural Database (Groom and Allen, 2014).

A Z score is calculated for every bond length and bond angle in each ligand. Individual bond lengths or bond angles with a Z score magnitude >2 are highlighted. The root-mean-square value of the Z scores (RMSZ) of bond lengths (or angles) is calculated for the entire molecule. The EDS software (Kleywegt et al., 2004 in [Supplemental References](#)) is used to calculate density maps from deposited atomic coordinates and experimental data, which are compared with idealized map density with the difference reported as a real space R value (RSR). This analysis is performed on an individual ligand basis. A local ligand density fit (for a description of this calculation see <http://www.wwpdb.org/validation/ValidationPDFNotes.html>) then compares the RSR of a molecule with the mean and SD of RSR for the neighboring polymeric standard amino acids and/or nucleotides. Minimum, median, 95th percentile, and maximum atomic displacement parameters (isotropic B values) for all atoms in the molecule are presented along with the number of atoms in the ligand molecule with occupancies of less than 0.9.

Quality of Macromolecule-Ligand Complexes in the PDB

Of the more than 115,000 entries in the PDB today, ~75% include at least one non-polymeric small-molecule ligand. While some of these ligands are almost certainly crystallization solutes, many were intentionally included in the experimental sample or co-purified with the structure determination target and are of considerable biological, biochemical, or medical interest. Recently published review articles assessing the quality of macromolecule-ligand complexes in the PDB can be usefully broken down into three categories:

1. Assessments of geometrical and stereochemical quality

(Affiliations continued on next page)

(Liesbeschuetz et al., 2012; Sehnal et al., 2015; Zheng et al., 2014 in [Supplemental References](#))

2. The suitability of ligand models for in silico drug discovery/design (Davis et al., 2008; Smart and Bricogne, 2015; Warren et al., 2012 in [Supplemental References](#))
3. General issues with ligand atomic model fit to the electron-density map (Malde and Mark, 2011; Pozharski et al., 2013; Sitzmann et al., 2012; Weichenberger et al., 2013 in [Supplemental References](#))

It has been emphasized by some that a non-negligible number of structural biologists err by interpreting weak density map features as indicating the presence of a bound small molecule that has been included in the crystallization process or soaked into a pre-formed crystal (e.g., Rupp, 2010 in [Supplemental References](#)). Current validation and journal refereeing policies and practices do not always prevent such cases from entering either the PDB archive or the scientific literature. Other explanations of problems with macromolecule-ligand complexes in the PDB include the following:

1. Some ligands undergo chemical transformation upon binding, which may not be reflected in the atomic model used for refinement
2. The ligand may be present, but was modeled incorrectly or refinement was performed with incorrect restraint targets
3. The ligand does bind, but the experimentalist does not provide an accurate chemical descriptor

Workshop Format and Charge to Participants

Catherine E. Peishoff (GlaxoSmithKline) gave the keynote address emphasizing the value of atomic level structural information for pharmaceutical discovery research and the growing opportunities for pre-competitive engagement and data sharing. She stressed the importance of data and structural model quality and the need for data archived in the PDB to be fit for purpose. Finally, she suggested a move away from the historical view of the PDB as an archival database toward an increased emphasis on data provisioning, which would shift the focus from any single structure to the structures as a collective. Increased attention to data standards, governance, and quality, together with improving tools to analyze the collective data, will significantly help researchers derive insight from this valuable scientific resource.

Stephen K. Burley (Research Collaboratory for Structural Bioinformatics [RCSB] PDB) and Gerard J. Kleywegt (PDB in Europe) then introduced the workshop rationale/objectives, and charged the participants with dividing among smaller breakout groups and addressing five questions regarding best practices for macromolecule-ligand complex data deposition and validation and journal editorial, refereeing, and publication practices. Breakout group members were selected on the bases of interest and expertise as follows: Group A, Academic and Industrial Crystallographers; Group B, Crystallographic Software Specialists; Group C, Computational Chemistry Software Specialists; and Group D, Academic and Industrial Crystallographers. After lengthy and lively discussions, the four breakout groups reconvened to report their findings and develop consensus recommendations. Each group independently approached the same set of questions.

Workshop Deliberations and Recommendations Charge to the Workshop

To address some of the myriad challenges facing PDB depositors and users and editors and referees of scientific journals that publish the results of structural studies of macromolecule-ligand complexes, the community stakeholders assembled at Rutgers considered the following five questions:

1. What are current best practices for selecting an initial ligand atomic model(s) for co-crystal structure refinement against diffraction data?
2. What are current best practices for validating the ligand(s) coming from such a co-crystal structure refinement?
3. What new data pertaining to co-crystal structures should be required for PDB depositions going forward?
4. What information should accompany journal submissions reporting co-crystal structure determinations? What supplementary materials should accompany publication of co-crystal structure determinations?
5. What do you recommend be done with existing co-crystal structures in the PDB archive?

Toward the close of the meeting, the groups reconvened to compare findings, identify areas of commonality and divergence, and determine how best to move forward. This document reflects the resulting consensus.

Workshop Recommendations

Recommended best practices for PDB archive deposition of co-crystal structure data:

Depositors should

1. Provide unambiguous chemical definitions for ligands present in the crystal mother liquor and in the refined structural model, including hydrogen atoms and covalent modifications
2. Provide the geometry of the starting model of the refined ligand(s), ligand-related refinement restraints, and their provenance
3. Use the PDBx/mmCIF dictionary `_atom_site.calc_flag` to identify non-experimentally modeled atoms. Non-experimentally modeled atoms, for the purposes of this recommendation, are defined as those atoms whose positions are not adequately localized by experimental data (e.g., electron-density map) to be assigned (x, y, z) positional coordinates, but whose presence is deduced by chemical knowledge of the crystal content and other information. This flag will usually be applicable to the hydrogen atom records for ligands. It is intended for use as an alternative to zero occupancy, which would be a less accurate indicator of the status of these atoms
4. Provide the Fourier coefficients of the density map(s) used for ligand(s) structure interpretation
5. Identify
 - a. Any ligand that is a focus of the study, where appropriate;
 - b. Any other biologically important ligand(s);

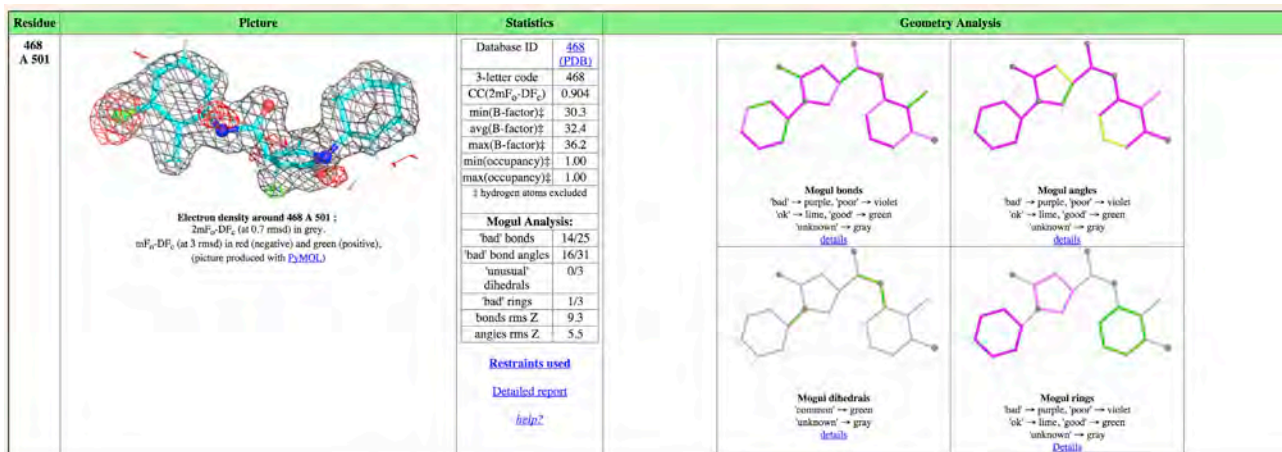


Figure 1. Representative Views of Ligand Chemical Structure and Electron Density

Example highlighting the value of presenting ligand electron-density model fit and geometrical analysis from CCDC Mogul from the Global Phasing Buster Report (PDB: 2H7P, later superseded by PDB: 4T2T [He et al., 2006]; CCD: 468).

- c. Adventitiously bound ligand(s) (i.e., co-purified) and ligands added for experimental convenience (e.g., crystallization additives or cryo-protectants); and
- d. The experimental method (crystal soaking versus co-crystallization) for (a) and (b).
6. As applicable, communicate other experimental findings, judgment calls, and perceived ambiguities regarding tautomers and protonation states of ligands not determined conclusively from the crystallographic data and chemical environment of the ligand by either (a) using the existing alternative conformation mechanism with partial occupancies or (b) providing the chemical descriptions recommended in Item 1 above.
7. Where appropriate, include comments explaining outliers, etc. identified in the wwPDB Validation Report
5. Identification of ligands capable of tautomerism or alternative protonation states within the pH range typical of protein crystals, nominally 4–10
6. The wwPDB D&A Validation pipeline should be described in full in peer-reviewed publications and continue to be publicly available for use in improving models prior to PDB archive deposition. The reference data used to calculate quality metrics/percentile scores should also continue to be publicly available. All the details describing the wwPDB Validation pipeline should be made available so that it can be implemented in an external environment. Specifically, details related to wwPDB Validation pipeline script(s), versions of the publicly available and commercial programs used therein, and input parameters and any other details necessary for reproducibility should be made public as soon as possible

Recommended best practices for wwPDB validation of co-crystal data:

Building on the framework of the current wwPDB Validation Report, the following new items should be included.

1. Informative images of ligand pose(s) plus nearby density map features using Fourier coefficients endorsed by the wwPDB X-ray VTF (e.g., $2m|F_o| - D|F_c|$, $m|F_o| - D|F_c|$, and omit map [Bhat, 1988; Bhat and Cohen, 1984]) and those provided by the depositor. The presentation style in the Buster Report tool (Smart and Bricogne, 2015 in Supplemental References) exemplifies the diagnostic utility of such representations (Figure 1)
2. Stick-figure representations of ligand(s) with non-hydrogen atom labels annotated with geometric validation findings
3. Identification of atoms modeled but not interpreted from density maps
4. Quality assessment metrics for each study compound and biologically important ligand(s)

Recommendations regarding editorial/refereeing/publication standards for co-crystal structure publications:
Journals should

1. Require submission of officially watermarked PDF wwPDB Validation Reports as Supplementary Materials accompanying manuscripts describing macromolecular structure determinations
2. Ensure that at least one of the referees selected for manuscript review has the technical expertise to evaluate in full the content of the wwPDB Validation Report

Response of the wwPDB X-Ray Validation Task Force

Following the conclusion of the Workshop, the recommendations outlined herein were presented to the membership of the wwPDB X-ray VTF (Chair: Randy Read, Cambridge University) when the group reconvened at the European Bioinformatics Institute in November 2015. The recommendations received strong support from the VTF. The wwPDB partner organizations (RCSB PDB, PDBe, PDBj, and Biological Magnetic Resonance Bank [BMRB]) are currently developing an implementation plan for recommendations relating to data requirements and updates

of the PDBx/mmCIF dictionary, and will finalize the plan in due course with the benefit of further advice from the VTF and the PDBx/mmCIF Working Group (Chair: Paul Adams, Lawrence Berkeley Laboratory).

Implementation Details

Implementation of the recommendations regarding additional archival content will require extension of the PDBx/mmCIF dictionary to capture details of the starting ligand model and the depositors' identification of the role of the ligand in each study. The PDBx/mmCIF Working Group (Chair: Paul Adams, Lawrence Berkeley Laboratory) is currently working on developing deposition standards for ligand refinement restraints and delivery of additional supporting data in the form of density map Fourier coefficients and unmerged intensities. Further extensions of the PDBx/mmCIF dictionary can be made as needed. With the requisite PDBx/mmCIF dictionary items in place, the wwPDB D&A system can be modified to ensure efficient capture of these new data during deposition.

An enhanced version of the wwPDB Validation Report will furnish the recommended depictions for ligand fits to map density and the annotated stick-figure models, with geometrical, stereochemical, and absence annotations. Development of a summary indicator of ligand quality for inclusion within the wwPDB Validation Report summary graphic requires additional research.

The wwPDB Validation Report also provides a convenient vehicle for delivering the recommended depictions of ligand density map to improve publication practices. Some scientific journals already require that wwPDB Validation Reports accompany structure manuscripts. Further community lobbying of editors is needed to expand the number of journals requiring submission of the wwPDB Validation Report. Finally, it is incumbent on the scientific community that experts continue to undertake rigorous review of manuscripts describing structural studies of macromolecule-ligand complexes.

Strong sentiments expressed both in the literature (e.g., Terwilliger and Bricogne, 2014 in [Supplemental References](#)) and during the workshop favored revision of the current wwPDB policy requiring issuance of new PDB codes following update of deposited atomic coordinates. Indeed, some depositors report being reluctant to update atomic coordinates because issuance of the new PDB code is thought to weaken the connection between the revised PDB archival entry and prior publications describing the structure. It was agreed that the wwPDB leadership, in consultation with the wwPDB Advisory Committee, should come to closure on the matter of versioning of atomic coordinates and other archival data as soon as possible.

Binding of ligands to macromolecules can also be studied using NMR spectroscopy. Members of the wwPDB NMR VTF present at the workshop volunteered the services of their task force to develop recommendations regarding data deposition and validation standards for structural models of macromolecule-ligand complexes determined by NMR.

Issues Addressed but Not Resolved at the Workshop

Workshop participants discussed three additional topics without reaching consensus.

First, some participants strongly advocated mandatory journal submission of processed diffraction data and atomic coordi-

nates to accompany manuscripts describing crystallographic studies of biological macromolecules. This practice is the norm for small-molecule crystallography publications. With the benefit of full and frank discussion, it was recognized that author sensitivities regarding providing primary data and atomic coordinates in advance of publication to reviewers, who may also be competitors, precluded consensus on this matter. The wwPDB leadership in consultation with the wwPDB Advisory Committee will revisit this issue.

Second, some participants strongly advocated mandatory PDB deposition of all-atom structural models, including computed positions of hydrogen atoms (properly identified with the `_atom_site.calc_flag`). As inclusion of explicit hydrogen atoms will affect the entire PDB archive, it was agreed that (1) technical recommendations on this front should be made by the wwPDB X-ray VTF, and (2) the wwPDB leadership, in consultation with the wwPDB Advisory Committee, should make further policy recommendations as necessary.

Finally, workshop participants identified a number of challenges that will come to the fore once enhanced validation of macromolecule-ligand complexes already archived in the PDB is concluded and updated wwPDB Validation Reports are made publicly available for every entry. Simply put, what should be done with existing PDB entries found wanting by the validation procedures recommended herein?

Workshop participants believe that the majority of depositors would be motivated to correct entries identified as not meeting minimal standards for enhanced ligand validation. However, it was also recognized that, over time, increasing numbers of depositors would not be in a position to make corrections. To ensure the integrity of the database, workshop participants propose that, following a reasonable interval for self-correction, community experts could be mobilized to apply targeted corrections to any remaining PDB archival entries with poor validation outcomes, particularly for bound ligands of significant biological and/or medical interest.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental References and can be found with this article online at <http://dx.doi.org/10.1016/j.str.2016.02.017>.

ACKNOWLEDGMENTS

The workshop was supported by funding to RCSB PDB by the National Science Foundation (DBI 1338415); PDBe by the Wellcome Trust (104948); PDBj by JST-NBDC; BMRB by the National Institute of General Medical Sciences (GM109046); D3R by the National Institute of General Medical Sciences (GM111528); registration fees from industrial participants; and a tax-deductible donation to the wwPDB Foundation by the Bristol-Myers Squibb Foundation.

REFERENCES

- Berman, H.M., Henrick, K., and Nakamura, H. (2003). Announcing the Worldwide Protein Data Bank. *Nat. Struct. Biol.* 10, 980.
- Bhat, T.N. (1988). Calculation of an OMIT map. *J. Appl. Cryst.* 21, 279–281.
- Bhat, T.N., and Cohen, G.H. (1984). OMITMAP—an electron-density map suitable for the examination of errors in a macromolecular model. *J. Appl. Cryst.* 17, 244–248.

- Bruno, I.J., Cole, J.C., Kessler, M., Luo, J., Motherwell, W.D., Purkis, L.H., Smith, B.R., Taylor, R., Cooper, R.I., Harris, S.E., et al. (2004). Retrieval of crystallographically-derived molecular geometry information. *J. Chem. Inf. Comput. Sci.* **44**, 2133–2144.
- Caboche, S., Pupin, M., Leclere, V., Fontaine, A., Jacques, P., and Kucherov, G. (2008). NORINE: a database of nonribosomal peptides. *Nucleic Acids Res.* **36**, D326–D331.
- Cahn, R.S., Ingold, C.K., and Prelog, V. (1966). Specification of molecular chirality. *Angew. Chem. Int. Ed. Engl.* **5**, 385–415.
- Davis, A.M., St-Gallay, S.A., and Kleywegt, G.J. (2008). Limitations and lessons in the use of X-ray structural information in drug design. *Drug Discov. Today* **13**, 831–841.
- Dutta, S., Dimitropoulos, D., Feng, Z., Persikova, I., Sen, S., Shao, C., Westbrook, J., Young, J., Zhuravleva, M.A., Kleywegt, G.J., et al. (2014). Improving the representation of peptide-like inhibitor and antibiotic molecules in the Protein Data Bank. *Biopolymers* **101**, 659–668.
- Feng, Z., Chen, L., Maddula, H., Akcan, O., Oughtred, R., Berman, H.M., and Westbrook, J. (2004). Ligand Depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics* **20**, 2153–2155.
- Fitzgerald, P.M.D., Westbrook, J.D., Bourne, P.E., McMahon, B., Watenpugh, K.D., and Berman, H.M. (2005). 4.5 macromolecular dictionary (mmCIF). In *International Tables for Crystallography G. Definition and Exchange of Crystallographic Data*, S.R. Hall and B. McMahon, eds. (Springer), pp. 295–443.
- Gasteiger, J., Rudolph, C., and Sadowski, J. (1990). Automatic generation of 3D-atomic coordinates for organic molecules. *Tetrahedron Comput. Methodol.* **3**, 537–547.
- Gore, S., Velankar, S., and Kleywegt, G.J. (2012). Implementing an x-ray validation pipeline for the protein data bank. *Acta Crystallogr. D Biol. Crystallogr.* **68**, 478–483.
- Groom, C.R., and Allen, F.H. (2014). The Cambridge Structural Database in retrospect and prospect. *Angew. Chem. Int. Ed. Engl.* **53**, 662–671.
- Hawkins, P.C., Skillman, A.G., Warren, G.L., Ellingson, B.A., and Stahl, M.T. (2010). Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model* **50**, 572–584.
- He, X., Alian, A., Stroud, R., and Ortiz de Montellano, P.R. (2006). Pyrrolidine carboxamides as a novel class of inhibitors of enoyl acyl carrier protein reductase from *Mycobacterium tuberculosis*. *J. Med. Chem.* **49**, 6308–6323.