HHS Public Access

Author manuscript

*Methods Mol Biol*. Author manuscript; available in PMC 2018 February 22.

Published in final edited form as:

*Methods Mol Biol*. 2017 ; 1607: 627–641. doi:10.1007/978-1-4939-7000-1_26.

# Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive

**Stephen K. Burley**, **Helen M. Berman**, **Gerard J. Kleywegt**, **John L. Markley**, **Haruki Nakamura**, and **Sameer Velankar**

## Abstract

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 houses ~130,000 entries (May 2017). 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.bmrb.wisc.edu). The four wwPDB partners operate a unified global software system that enforces community-agreed data standards and supports data Deposition, Biocuration, and Validation of ~11,000 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 complementary 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 and validation of 3D structural models and experimental data derived from integrative or hybrid methods.

### 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

## 1 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 data resource in the biological sciences [1]. Soon after X-ray structures of myoglobin [2, 3] and hemoglobin [4, 5] 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 [6]. Walter C. Hamilton, one of the attendees, offered to provide the first home for what is now the Protein Data Bank (PDB) [7]. Shortly thereafter, the PDB was launched from within the Department of Chemistry at Brookhaven National Laboratory (BNL), building on the Protein Structure Library framework [8]. The importance of scientific data archiving as 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 *via* the Cambridge Crystallographic Database Centre [1].

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 [9], and over time, adopted by virtually all of the scientific journals now requiring PDB deposition of atomic coordinates 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 include experimental data.

## 2 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 nonlinear (Fig. 1). By 1982, the PDB had reached only ~100 entries. Eleven years later, in 1993, there were 1000 entries. Before the end of the decade (1999), this number had grown to 10,000. Circa 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 extant PDB archival entries came 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 [9269 in 2012; 10,168 in 2015]. During that same period, 3DEM depositions have increased 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 1062/year, declining to 510/year in 2015. The PDB archive has also grown considerably in complexity since 1971. Some proxy measures of complexity are provided in Table 1.

## 3 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 [10]. Following an open re-competition for US federal funding of the PDB in 1998, responsibility for the archive was awarded 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 Institute of Standards and Technology [11]. Following a transition period that witnessed formalization of Protein Data Bank Japan (PDBj) [12] and the Macromolecular Structure Database (MSD) [13, 14], RCSB PDB, PDBj, and MSD came together in 2003 to establish the Worldwide Protein Data Bank (wwPDB; wwpdb.org) [15]. In 2006, a global NMR data repository BioMagResBank (BMRB), founded in 1989 [16], joined the wwPDB organization [17]. BMRB hosts deposition sites in both the US (BMRB; www.bmrb.wisc.edu) and Japan (PDBj-BMRB; bmrbdep.pdbj.org) [18]. (N.B.: MSD was rebranded in 2008 as the Protein Data Bank in Europe or PDBe [14, 19].)

The wwPDB organization is governed by a Memorandum of Understanding (wwpdb.org/about/agreement), which was renewed in 2013. Oversight of wwPDB partner activities is provided by an internationally recognized team of experts in structural biology and bioinformatics comprising the wwPDB Advisory Committee (http://wwpdb.org/about/advisory). As outlined in detail below, wwPDB partners collaborate on “Data In.” They are jointly responsible for standardizing, collecting, biocurating, validating, 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 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.

## 4 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 [11], and AutoDep, which was developed at BNL [20] and reengineered by MSD/PDBe [21]. NMR depositions were accepted *via* ADIT-NMR at BMRB and PDBj-BMRB, with coordinates and restraint data transferred to RCSB PDB or PDBj, respectively [17]. 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, Biocuration, 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 the Middle East and Asia (PDBj). The pipeline currently used by the wwPDB to process incoming structures is illustrated schematically in Fig. 2. Approximately 900 depositions are received monthly

from every inhabited continent (Fig. 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, biocurate, 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 [22]. Preliminary recommendations have also been made by VTFs for NMR [23] and 3DEM [24]. The work of these interoperating 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 [25], and, instead, an independent repository should be created to archive computed models elsewhere. The Protein Modeling Portal was established in 2007 [26]. 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 one or more SAS data repositories that should interoperate with the PDB archive [27]. Subsequently, some 49 PDB entries derived exclusively from SAS methods were transferred into the SAS Biological Data Bank (SASbDB; sasbdb.org) archive [28] 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) [29] 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 [30] were endorsed by the wwPDB X-ray VTF in late 2015. Implementation of these recommendations was underway at the time of writing.

## 5 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 denoted 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 (IRRMC; www.proteindiffraction.org) and the Structural Biology Data Grid Consortium (SBGrid; sbgrid.org [31]) 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 determined by the facility. BMRB [32] 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 [33]. Voluntary archival deposition of raw 3DEM images is currently supported by EMPIAR [34].

The first data format used by the PDB archive was established in the early 1970s, and was based on the 80-column Hollerith format used for punched cards [35]. 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 (IUCr) charged a committee with creating a more informative and extensible data model for the PDB archive.

In response to the IUCR committee report, the Macromolecular Crystallographic Information File (mmCIF) was proposed [36]. 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 archived. 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 3000 data items pertaining to crystallography. Over time, data items specific to NMR and 3DEM were added, and the dictionary was subsequently rebranded PDBx/mmCIF [37]. 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 going forward all output from their programs would be available in PDBx/mmCIF.

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, the PDBx/mmCIF dictionary contained more than 4400 data items, including ~250 and ~1200 specific to NMR and 3DEM, respectively. PDBML, an XML format based on PDBx/mmCIF [38] 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 [39]. Recently, XML and RDF-formatted BMRB data have been provided as BMRB/XML and BMRB/RDF, respectively [40], 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 [41], which builds on PDBx/mmCIF. In addition to accelerating development of the SASbDB archive, creation of sasCIF will allow for facile inter-operation 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 [42]. As noted above, this format has been harmonized with PDBx/mmCIF and now serves as the preferred deposition format for NMR structures [43]. 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 addressed in part by the NMR Restraints Grid, which can process restraint files and convert them to the NMR-STAR or CCPN formats [44, 45]. 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 [46]. 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 precluded 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 (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.

## 6 Data Representation for Small Molecules

The PDB Chemical Component Dictionary (CCD) was originally developed [47] 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 [36]. Each new chemical component coming into the archive is identified by a unique three-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 (CSD) operated by CCDC [29] were identified in a collaborative effort, which revealed ~1400 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) [48], which contains information about oligopeptide-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 [49] or Norine [50] archived amino acid 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.

## 7 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 biocurated/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 or has been 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 entries 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 blinded prediction challenges (CAMEO: cameo3d.org [51]; and D3R CELPP: drugdesigndata.org/about/celpp). Stage 2: every Wednesday at 00:00 UTC, all new entries designated for release are made publicly available through four wwPDB FTP sites (wwPDB: ftp.wwpdb.org; RCSB PDB: ftp.rcsb.org; PDBe: ftp.ebi.ac.uk/pub/databases/pdb/; PDBj: ftp.pdbj.org). On average, ~200 structures are released every week, corresponding to ~111,000 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 complement 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: bmrb.wisc.edu). Collectively, wwPDB FTP sites and partner websites support in excess of 500 million downloads of data files annually. Simply put, more than one million data files are downloaded by PDB users distributed across all inhabited continents every day of the year. Our records show that FTP downloads of PDB data were made to all but four of the 195 recognized independent states worldwide during the period 2012–2015 (excluding Central African Republic, Cote d'Ivoire, Kosovo, and Swaziland). No PDB FTP download requests were recorded from the disputed territory of Western Sahara during the same period.

## 8 Future of Structural Biology and the Role of the wwPDB

At present, PDB archival entries come exclusively from measurements using crystallography, NMR, and 3DEM. These mainstay structure determination methods involve the same four basic steps: (1) making measurements from a physical sample of a biological macromolecule(s); (2) utilizing a representation of the measured data that allows encoding of these data for use by a computable scoring function encompassing spatial restraints that directly compares predicted and measured experimental results; (3) 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 (4) 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, structural biologists 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 [52]). 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 NMR data, particularly for dynamic systems [53, 54].

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 4 Å 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 cross-linking/mass spectrometry, Forster resonance energy transfer or FRET, and SAS (e.g., [55]). 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 bicuration/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 [56] 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 makeup 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 collaborate with the wwPDB partners in establishing the data resource federation. At the time of writing, the SASbDB resource [28] 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.

## 9 PDB Archive at 50 Years of Age

The PDB is just 5 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, Biocuration, 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.

## Acknowledgments

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). We thank Christine Zardecki for expert help with manuscript preparation.

## References

1. Protein Data Bank. Protein Data Bank. Nature New Biology. 1971; 233:223. [PubMed: 20480989]
2. Kendrew JC, Bodo G, Dintzis HM, et al. A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. Nature. 1958; 181:662–666. [PubMed: 13517261]
3. Kendrew JC, Dickerson RE, Strandberg BE, et al. Structure of myoglobin: a three-dimensional Fourier synthesis at 2 Å resolution. Nature. 1960; 185:422–427. [PubMed: 18990802]
4. Bolton W, Perutz MF. Three dimensional fourier synthesis of horse deoxyhaemoglobin at 2.8 Ångstrom units resolution. Nature. 1970; 228:551–552. [PubMed: 5472471]
5. Perutz MF, Rossmann MG, Cullis AF, et al. Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5 Å resolution, obtained by X-ray analysis. Nature. 1960; 185:416–422. [PubMed: 18990801]
6. Cold Spring Laboratory. Cold Spring Harbor Symposia on quantitative biology. Vol. 36. Cold Spring Laboratory Press; Cold Spring Harbor, NY: 1972.
7. Berman H. The Protein Data Bank: a historical perspective. Acta Crystallogr A. 2008; 64:88–95. [PubMed: 18156675]
8. Meyer EF. The first years of the Protein Data Bank. Protein Sci. 1997; 6:1591–1597. [PubMed: 9232661]
9. International Union of Crystallography. Policy on publication and the deposition of data from crystallographic studies of biological macromolecules. Acta Crystallogr A. 1989; 45:658.
10. Sussman JL, Lin D, Jiang J, et al. Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. Acta Crystallogr D Biol Crystallogr. 1998; 54:1078–1084. [PubMed: 10089483]
11. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. Nucleic Acids Res. 2000; 28:235–242. [PubMed: 10592235]
12. Standley DM, Kinjo AR, Kinoshita K, et al. Protein structure databases with new web services for structural biology and biomedical research. Brief Bioinform. 2008; 9:276–285. [PubMed: 18430752]
13. Keller PA, Henrick K, McNeil P, et al. Deposition of macromolecular structures. Acta Crystallogr D Biol Crystallogr. 1998; 54:1105–1108. [PubMed: 10089486]

14. Velankar S, van Ginkel G, Alhroub Y, et al. PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. Nucleic Acids Res. 2016; 44:D385–D395. [PubMed: 26476444]
15. Berman HM, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. Nat Struct Biol. 2003; 10:980. [PubMed: 14634627]
16. Ulrich EL, Markley JL, Kyogoku Y. Creation of a nuclear magnetic resonance data repository and literature database. Protein Seq Data Anal. 1989; 2:23–37. [PubMed: 2911559]
17. Markley JL, Ulrich EL, Berman HM, et al. 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. [PubMed: 18288446]
18. Ulrich EL, Akutsu H, Doreleijers JF, et al. BioMagResBank. Nucleic Acids Res. 2008; 36:D402–D408. [PubMed: 17984079]
19. Velankar S, Best C, Beuth B, et al. PDBe: Protein Data Bank in Europe. Nucleic Acids Res. 2010; 38:D308–D317. [PubMed: 19858099]
20. Lin D, Manning NO, Jiang J, et al. AutoDep: a web-based system for deposition and validation of macromolecular structural information. Acta Crystallogr D Biol Crystallogr. 2000; 56:828–841. [PubMed: 10930830]
21. Tagari M, Tate J, Swaminathan GJ, et al. E-MSD: improving data deposition and structure quality. Nucleic Acids Res. 2006; 34:D287–D290. [PubMed: 16381867]
22. Read RJ, Adams PD, Arendall WB, et al. A new generation of crystallographic validation tools for the Protein Data Bank. Structure. 2011; 19:1395–1412. [PubMed: 22000512]
23. Montelione GT, Nilges M, Bax A, et al. Recommendations of the wwPDB NMR Validation Task Force. Structure. 2013; 21:1563–1570. [PubMed: 24010715]
24. Henderson R, Sali A, Baker ML, et al. Outcome of the first electron microscopy validation task force meeting. Structure. 2012; 20:205–214. [PubMed: 22325770]
25. Berman HM, Burley SK, Chiu W, et al. Outcome of a workshop on archiving structural models of biological macromolecules. Structure. 2006; 14:1211–1217. [PubMed: 16955948]
26. Arnold K, Kiefer F, Kopp J, et al. The Protein Model Portal. J Struct Funct Genom. 2009; 10:1–8.
27. Trewhella J, Hendrickson WA, Kleywegt GJ, et al. Report of the wwPDB Small-Angle Scattering Task Force: data requirements for biomolecular modeling and the PDB. Structure. 2013; 21:875–881. [PubMed: 23747111]
28. Valentini E, Kikhney AG, Previtali G, et al. SASBDB, a repository for biological small-angle scattering data. Nucleic Acids Res. 2015; 43:D357–D363. [PubMed: 25352555]
29. Groom CR, Bruno IJ, Lightfoot MP, et al. The Cambridge Structural Database. Acta Crystallogr B. 2016; 72:171–179.
30. Adams PD, Aertgeerts K, Bauer C, et al. Outcome of the First wwPDB/CCDC/D3R Ligand Validation Workshop. Structure. 2016; 24:502–508. [PubMed: 27050687]
31. Meyer PA, Socias S, Key J, et al. Data publication with the structural biology data grid supports live analysis. Nature Commun. 2016; 7:10882. [PubMed: 26947396]
32. Markley, JL., Ulrich, EL., Westler, WM., et al. Macromolecular structure determination by NMR spectroscopy. In: Bourne, PE., Weissig, H., editors. Structural bioinformatics. John Wiley & Sons, Inc; Hoboken, NJ: 2003. p. 89-113.
33. Lawson CL, Patwardhan A, Baker ML, et al. EMDataBank unified data resource for 3DEM. Nucleic Acids Res. 2016; 44:D396–D403. [PubMed: 26578576]
34. Iudin A, Korir PK, Salavert-Torres J, et al. EMPIAR: a public archive for raw electron microscopy image data. Nat Methods. 2016; 13:387. [PubMed: 27067018]
35. Bernstein FC, Koetzle TF, Williams GJB, et al. Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol. 1977; 112:535–542. [PubMed: 875032]
36. Fitzgerald, PMD., Westbrook, JD., Bourne, PE., et al. 4.5 Macromolecular dictionary (mmCIF). In: Hall, SR., McMahon, B., editors. International Tables for Crystallography G. Definition and exchange of crystallographic data. Springer; Dordrecht, The Netherlands: 2005. p. 295-443.
37. Westbrook, JD., Henrick, K., Ulrich, EL., et al. Appendix 3.6.2. The Protein Data Bank Exchange Data Dictionary. In: Hall, SR., McMahon, B., editors. International Tables for Crystallography G.

Definition and exchange of crystallographic data. Springer; Dordrecht, The Netherlands: 2005. p. 195-198.

38. Westbrook J, Ito N, Nakamura H, et al. PDBML: the representation of archival macromolecular structure data in XML. Bioinformatics. 2005; 21:988–992. [PubMed: 15509603]
39. Kinjo AR, Suzuki H, Yamashita R, et al. Protein Data Bank Japan (PDBj): maintaining a structural data archive and resource description framework format. Nucleic Acids Res. 2012; 40:D453–D460. [PubMed: 21976737]
40. Yokochi M, Kobayashi N, Ulrich EL, et al. Publication of nuclear magnetic resonance experimental data with semantic web technology and the application thereof to biomedical research of proteins. J Biomed Semantics. 2016; 7:16. [PubMed: 27927232]
41. Malfois M, Svergun DI. sasCIF: an extension of core Crystallographic Information File for SAS. J Appl Crystallogr. 2000; 33:812–816.
42. Ulrich EL, Argentar D, Klimowicz A, et al. STAR/CIF macromolecular NMR data dictionaries and data file formats. Acta Crystallogr A. 1996; 52:C577–C577.
43. Berman, HM., Henrick, K., Nakamura, H., et al. The Worldwide Protein Data Bank. In: Gu, J., Bourne, PE., editors. Structural bioinformatics. 2. Wiley; Hoboken, NJ: 2009. p. 293-303.
44. Doreleijers JF, Vranken WF, Schulte C, et al. NRG-CING: integrated validation reports of remediated experimental biomolecular NMR data and coordinates in wwPDB. Nucleic Acids Res. 2012; 40:D519–D524. [PubMed: 22139937]
45. Doreleijers JF, Vranken WF, Schulte C, et al. The NMR restraints grid at BMRB for 5,266 protein and nucleic acid PDB entries. J Biomol NMR. 2009; 45:389–396. [PubMed: 19809795]
46. Gutmanas A, Adams PD, Bardiaux B, et al. NMR Exchange Format: a unified and open standard for representation of NMR restraint data. Nat Struct Mol Biol. 2015; 22:433–434. [PubMed: 26036565]
47. Westbrook JD, Shao C, Feng Z, et al. The chemical component dictionary: complete descriptions of constituent molecules in experimentally determined 3D macromolecules in the Protein Data Bank. Bioinformatics. 2015; 31:1274–1278. [PubMed: 25540181]
48. Dutta S, Dimitropoulos D, Feng Z, et al. Improving the representation of peptide-like inhibitor and antibiotic molecules in the Protein Data Bank. Biopolymers. 2014; 101:659–668. [PubMed: 24173824]
49. UniProt Consortium. UniProt: a hub for protein information. Nucleic Acids Res. 2015; 43:D204–D212. [PubMed: 25348405]
50. Caboche S, Pupin M, Leclere V, et al. NORINE: a database of nonribosomal peptides. Nucleic Acids Res. 2008; 36:D326–D331. [PubMed: 17913739]
51. Haas J, Roth S, Arnold K, et al. The Protein Model Portal—a comprehensive resource for protein structure and model information. Database. 2013; 2013:bat031. [PubMed: 23624946]
52. Prischi F, Pastore A. Application of nuclear magnetic resonance and hybrid methods to structure determination of complex systems. Adv Exper Med Biol. 2016; 896:351–368. [PubMed: 27165336]
53. Cornilescu G, Didychuk AL, Rodgers ML, et al. Structural analysis of multi-helical RNAs by NMR-SAXS/WAXS: application to the U4/U6 di-snRNA. J Mol Biol. 2016; 428:777–789. [PubMed: 26655855]
54. Venditti V, Egner TK, Clore GM. Hybrid approaches to structural characterization of conformational ensembles of complex macromolecular systems combining NMR residual dipolar couplings and solution X-ray scattering. Chem Rev. 2016; 116:6305–6322. [PubMed: 26739383]
55. Erzberger JP, Stengel F, Pellarin R, et al. Molecular architecture of the 40SeIF1eIF3 translation initiation complex. Cell. 2014; 158:1123–1135. [PubMed: 25171412]
56. Sali A, Berman HM, Schwede T, et al. Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop. Structure. 2015; 23:1156–1167. [PubMed: 26095030]

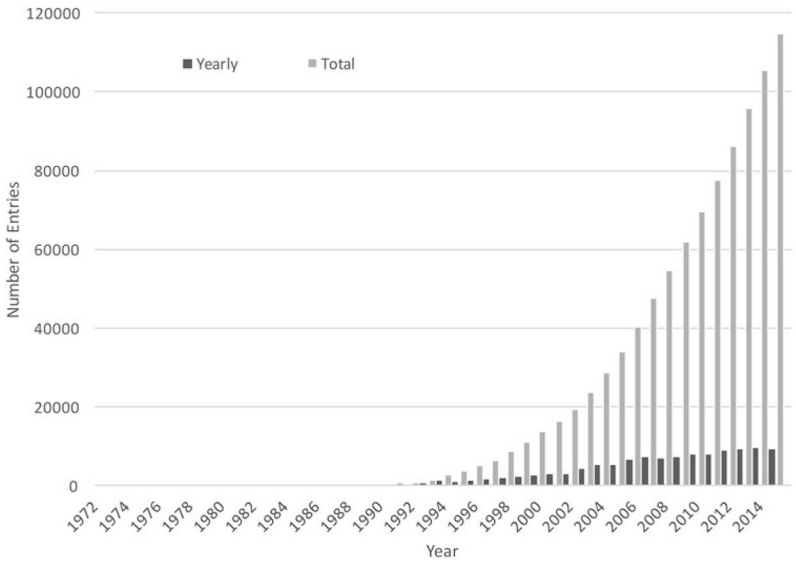


**Fig. 1.**
Growth of the PDB Archive since 1971

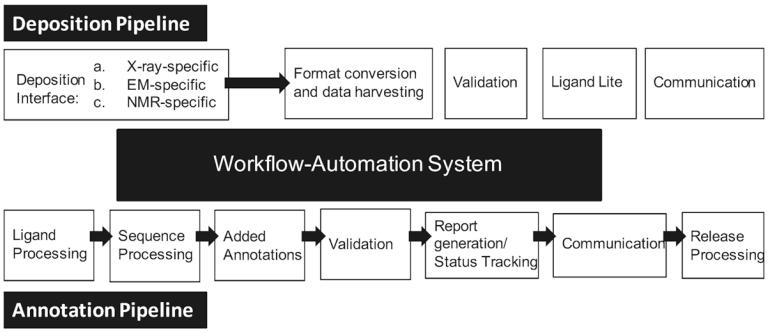


**Fig. 2.**

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

**Fig. 3.**
World map showing global distribution of PDB Depositors (2012–2015)

**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 (CCD) |
|---|---|---|---|---|
| 2012 | 14 | 133 | ~450 | 1733 |
| 2013 | 32 | 198 | ~440 | 1875 |
| 2014 | 49 | 164 | ~690 | 1767 |
| 2015 | 55 | 311 | ~580 | 1830 |