Quarterly Reviews of Biophysics

cambridge.org/qrb

Anticipating innovations in structural biology

Helen M. Berman, Catherine L. Lawson, Brinda Vallat and Margaret J. Gabanyi

Department of Chemistry and Chemical Biology, Center for Integrative Proteomics Research, Institute for Quantitative Biomedicine, 174 Frelinghuysen Road, Piscataway, New Jersey 08854, USA

Review

Cite this article: Berman HM, Lawson CL, Vallat B, Gabanyi MJ (2018). Anticipating innovations in structural biology. *Quarterly Reviews of Biophysics* **51**, e8, 1–13. https://doi.org/10.1017/S0033583518000057

Received: 2 December 2017
Revised: 19 March 2018
Accepted: 20 March 2018

Key words:
Protein Data Bank; community data resources; structural biology; structural genomics; data standards

Author for correspondence:
Helen M. Berman, E-mail: berman@rcsb.rutgers.edu

Abstract

In this review, we describe how the interplay among science, technology and community interests contributed to the evolution of four structural biology data resources. We present the method by which data deposited by scientists are prepared for worldwide distribution, and argue that data archiving in a trusted repository must be an integral part of any scientific investigation.

Introduction

The structural biology community has been uniquely proactive in establishing data resources that archive the results of research and provide services to access and analyze those data. The Protein Data Bank (PDB) was established as a repository for biomacromolecular structural data more than 45 years ago (Protein Data Bank, 1971). It now contains more than 140 000 structures determined by X-ray crystallography, Nuclear magnetic resonance (NMR) spectroscopy, and three-dimensional electron microscopy (3DEM). A diverse community of researchers, students, educators and the general public downloads more than 1.9 million data sets every day. In this review, we demonstrate how the synergies among science, technology and community enabled the PDB to preserve the past while constantly evolving to reflect contemporary needs. We describe how and why two other structural biology data resources were created to supplement and collaborate with the PDB. We conclude by demonstrating how the experiences of the past inform how we are meeting the current challenges presented by the more recent determination of structural models of large macromolecular machines.

The synergies of science, technology and community in the development of the PDB

In 1957, the structure of myoglobin was determined (Kendrew *et al.*, 1958), followed shortly thereafter by hemoglobin (Perutz *et al.*, 1960). Thus, began the era of structural biology in which, one by one, structures of small proteins including enzymes such as lysozyme (Blake *et al.*, 1965), ribonuclease (Kartha *et al.*, 1967; Wyckoff *et al.*, 1967), and carboxypeptidase (Quiocho and Lipscomb, 1971) were determined using X-ray crystallography. By the late sixties, more than a dozen structures had been determined. In those days, X-ray crystallographic methods involved the use of calculators, newly emerging computers and manual model building relying on the Richards Box, an optical comparator that had to be housed in a large room (Richards, 1968). A single determination took years of painstaking work. The three-dimensional (3D) atomic coordinates obtained from these structure determinations contained a treasure trove of information that would eventually reveal new insights into biology, medicine, biophysics and biochemistry. Indeed, the award of the Nobel Prize to Kendrew and Perutz in 1962 (Nobelprize.org, 2017) recognized not just their achievements, but also the potential of X-ray crystallography. However, for others to help build on that knowledge, it would be necessary to have access to the 3D coordinates produced by all of these new structure determinations.

The coordinate data were stored on punched cards, paper tape and magnetic tape. Because the Internet was only beginning to be established, transfer of data between laboratories involved recording the data onto appropriate media and mailing it. Starting in 1966, a small community of scientists met periodically to discuss how best to archive and distribute these structures. In 1971, a seminal meeting was held in Cold Spring Harbor (Phillips, 1972) in which the practitioners and now pioneers of structural biology described their structures to a rapt and inspired audience. Among the attendees was Walter Hamilton, an energetic and highly respected chemical crystallographer from Brookhaven National Laboratory (BNL). Walter had been collaborating with Edgar Meyer who was creating a Protein Library (Meyer, 1997). When presented with the problem of needing an archive for biomacromolecular structures, Hamilton immediately offered to house one at BNL. He contacted Olga Kennard who was then head of the Cambridge Crystallographic Data Center (CCDC) in Cambridge, UK (Allen *et al.*, 1973) and they agreed to set up the PDB (Protein Data Bank, 1971) as collaboration between BNL and CCDC. After Hamilton's death in 1973, Tom Koetzle took over the

CAMBRIDGE
UNIVERSITY PRESS

https://doi.org/10.1017/S0033583518000057 Published online by Cambridge University Press

direction of the PDB. In 1979 there *were 53 structures in the PDB* (Fig. 1), some of which are shown in Fig. 2.

The 1980s saw a steady growth of structures in the PDB in large part because of the emergence of powerful new technologies. Genetic engineering made it possible to clone and express large quantities of protein without resorting to extraction from natural biological sources. Chemical synthesis could be used to obtain purified fragments of DNA. The advent of synchrotron sources allowed the collection of data with intense X-ray beams (Harmsen *et al.*, 1976). At the same time, development of the multiple anomalous diffraction phasing method (MAD) (Hendrickson *et al.*, 1985) leveraged the ability to tune the X-ray wavelength using synchrotron radiation. Flash freezing (Hope, 1988) to prevent crystal decay began to be more widely used. Multi-wire detectors made it possible to rapidly collect many diffraction reflections at once (Hamlin, 1985). Computing technology continued to improve. In particular, molecular graphics made it possible to fit structural models to electron density (Jones, 1978), replacing the need for the Richards Box. During this period, NMR spectroscopy began to be used for determining the structures of small proteins (Horst *et al.*, 2001), thus eliminating the requirement of crystallinity. During the 1980s, the first

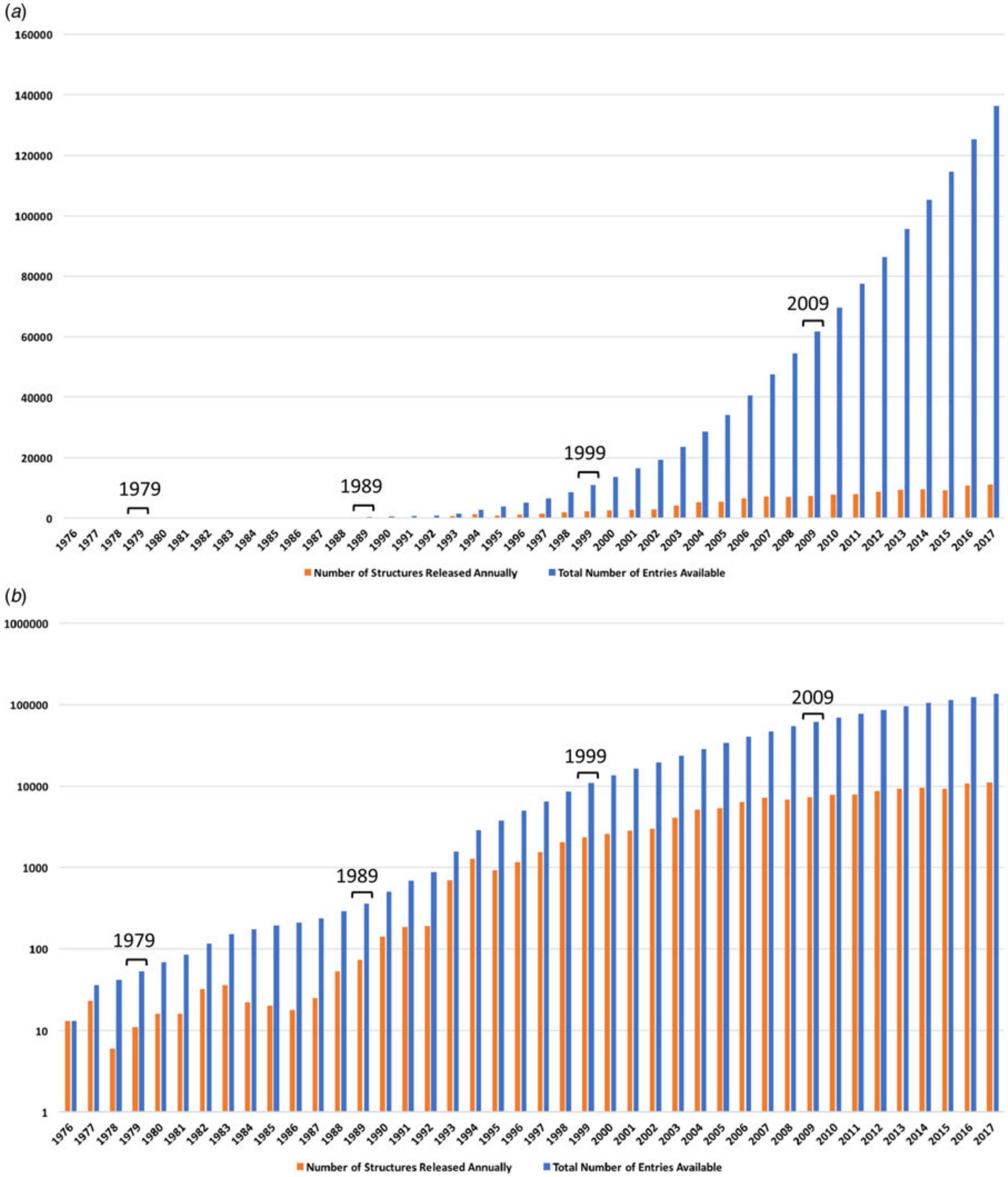

Fig. 1. Growth chart of structures in the PDB with indicators of each decade. (a) The number of structures released per year (blue) and the cumulative number of structures (orange). (b) The same information, using a log scale. The number of structures released at the end of each decade is indicated by black brackets.

https://doi.org/10.1017/S0033583518000057 Published online by Cambridge University Press

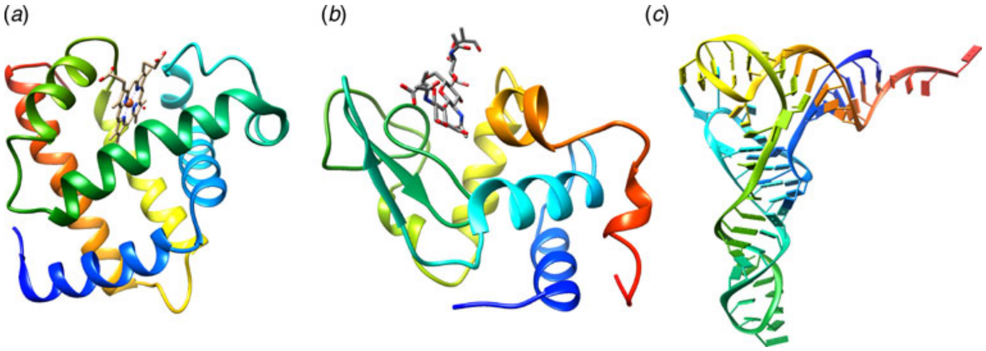

Fig. 2. Examples of structures determined in the 1970's. Ribbon representations were generated using UCSF Chimera (Pettersen *et al.*, 2004). (a) Myoglobin (Watson, 1969), PDB ID: 1MBN. First protein structure determined using X-ray crystallography. (b) Lysozyme (Blake *et al.*, 1965; Kelly *et al.*, 1979), PDB ID: 9LYZ. First enzyme structure determined using X-ray crystallography. (c) Yeast phenylalanine transfer RNA (Rich & Kim, 1978; Robertus *et al.*, 1974), PDB ID: 4TNA. First RNA structure determined using X-ray crystallography.

atomic structures of viruses were determined (Hopper *et al.*, 1984; Erickson *et al.*, 1985) as were those of DNA (Dickerson *et al.*, 1982) (Fig. 3).

With the potential of structural biology being realized at an increasing pace, members of the scientific community began to be concerned that valuable data would be lost if deposition of structures into the PDB were not mandatory (Barinaga, 1989). Starting in about 1982, committees were set up to determine exactly which data should be archived. Fred Richards created a petition signed by many of the leading structural biologists, urging deposition into the PDB (Hufton, 2014). In 1989, the International Union of Crystallography (IUCr) published guidelines for the deposition, archival and release of structural data (International Union of Crystallography, 1989). The National Institute of General Medical Sciences (NIGMS) then made a ruling that structure determinations funded by that institute had to be archived by the PDB. In time, virtually all journals required deposition of coordinates in the PDB as a mandatory condition of publication. Another important event in the 1980s was the inclusion of structural biology as a focus of research by Howard Hughes investigators (Howard Hughes Medical Institute, 2017). By 1989, there were 365 structures in the PDB (Fig. 1).

The rate of data deposition rapidly took off in the 1990s as even better methods for data collection, structure determination and refinement were developed and adopted. Computer performance continued to improve dramatically and structural biologists were more than eager to embrace the new capabilities. During this period, the very first atomic structure determined by electron microscopy methods was deposited into the PDB (Henderson *et al.*, 1990). The 1990s saw the deposition of many protein-nucleic acid complexes into the archive, including the structure of the nucleosome (Luger *et al.*, 1997) (Fig. 4). By 1999, there were 10,963 structures in the PDB10963 (Fig. 1).

When the PDB was first established, the focus was on the collection of the coordinate data as well as some other descriptive data. The PDB Format (Westbrook & Fitzgerald, 2009) was widely adopted because it was simple and 'human'-readable. However, it was lacking in many other ways: relationships among data items were implicit and not explicit, there was no controlled vocabulary, there were limitations on the number of atoms and residues,

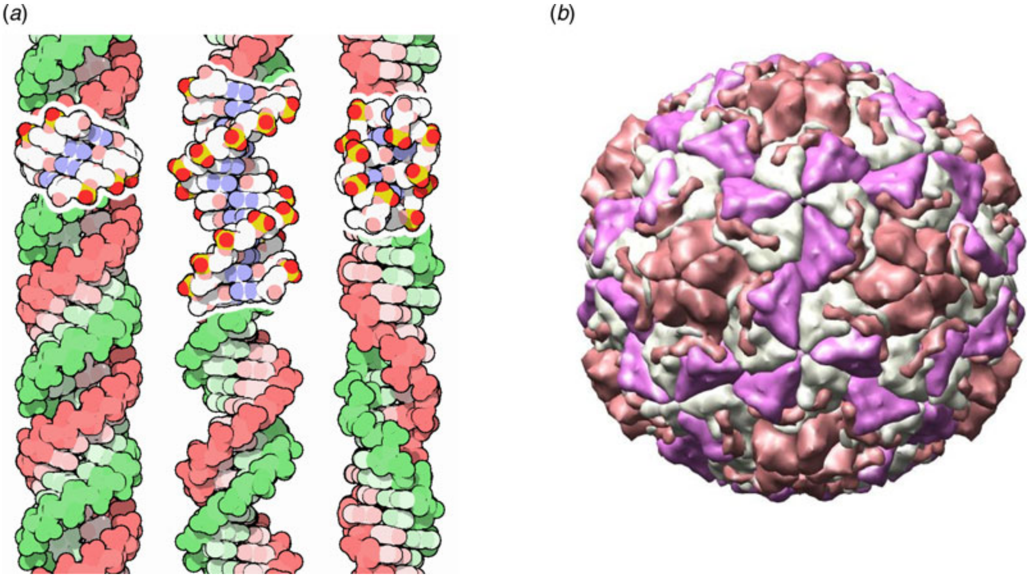

Fig. 3. Examples of structures determined in the 1980's. (a) A, B and Z DNA (Dickerson *et al.*, 1982). This representation of the three canonical forms of DNA is taken from the Molecule of the Month (Goodsell, 2001). (b) Rhinovirus (Arnold & Rossmann, 1988), PDB ID: 4RHV. This was one of the early virus structures determined using X-ray crystallography. Three unique chains (grey, pink, orange surfaces) are repeated 60-fold to create a virus capsid with icosahedral symmetry.

https://doi.org/10.1017/S0033583518000057 Published online by Cambridge University Press

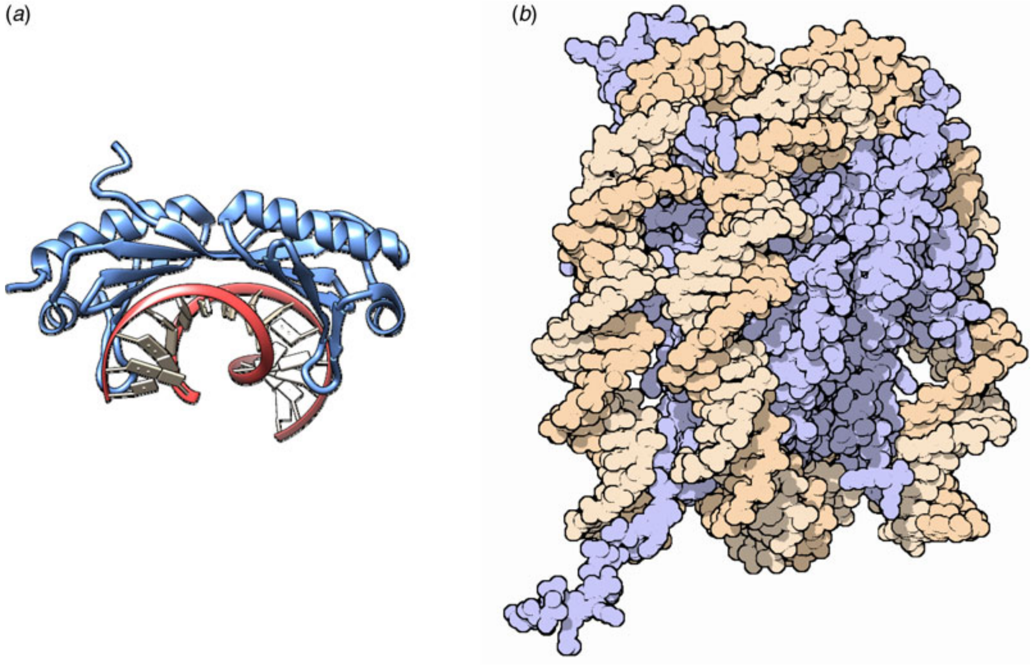

Fig. 4. Examples of structures determined in the 1990's. (a) The structure of a regulator of transcription called the TATA-binding protein bound to DNA. The binding of beta sheets into the minor groove of DNA causes a profound bend in the DNA (Patikoglou *et al.*, 1999), PDB ID: 1QN6. (b) Nucleosome (Luger *et al.*, 1997). The DNA is shown in orange wraps around the histone proteins shown in blue. Taken from the Molecule of the Month (Goodsell, 2000a).

and some of the definitions of data items were vague. In 1990, the IUCR set up a working group (WG) to create a Macromolecular Crystallographic Information File (mmCIF). It was originally supposed to be a variant of the Crystallographic Information File (CIF) that was already established for small molecules (Hall *et al.*, 1991). The mmCIF WG decided to use the opportunity to not only create richer data content with precise definitions for the macromolecular crystallographic experiment and its results but also to improve the data representation for PDB entries. A new data model was created that had data type definitions, explicit parent–child relationships among data items, enumerations for controlled vocabulary, and many other features. Workshops were held to obtain community feedback; by 1996, more than three thousand definitions were instantiated into a computer readable dictionary (Fitzgerald *et al.*, 2005). When the PDB moved from management by BNL to the Research Collaboratory for Structural Bioinformatics (RCSB) in 1998, mmCIF became the underlying data model that allowed for the creation of a relational database. However, uptake by the community was slow and it was not until 2011 that mmCIF became the Master Format for the PDB, allowing the PDB Format to be retired. As larger structures of macromolecular assemblies started to be deposited into the PDB, the limitations of the PDB format became more apparent, leading to wider acceptance of the mmCIF format.

The 2000s saw even more growth in the PDB. Ribosome structures, representing some of the very largest and most complex structures in the PDB, were deposited (Ban *et al.*, 2000; Carter *et al.*, 2000; Schluenzen *et al.*, 2000) (Fig. 5). Not surprisingly, the feat of determining these structures led to the award of a Nobel Prize in Chemistry in 2009, shared by three structural biologists. During the same period, the Protein Structure Initiative (PSI) began in which structures were determined on a genomic scale, resulting in nearly 7000 new structures in the PDB. *In 2009, there were 61,812 structures in the PDB* (Fig. 1).

When the PDB was first established, it was international in nature. Under BNL management, only one site curated the data, although there were multiple mirror or distribution sites. After RCSB was awarded the grant to manage the PDB, other sites were eager to become deposition sites. In 2003, three data centers, RCSB PDB in the USA, MSD (later PDBe) at the EMBL-EBI, and PDBj in Osaka, established the worldwide PDB (wwPDB) (Berman *et al.*, 2003). A formal agreement was created to ensure that all structures curated by the data centers follow the same rules for data processing and that there would be one archive with identical copies distributed by the wwPDB partners. At the time of this first agreement, compliance was difficult because there were two completely different processing pipelines. To ensure that the curated data files were in fact following the same rules, there were regular exchanges among the wwPDB partners to revalidate the data. The need for a single data processing pipeline became apparent. The project to create OneDep began in 2007; this new pipeline system was put into production in 2014 (Young *et al.*, 2017).

By establishing an international consortium whose goal was to develop and maintain a single, high-quality archive, it became possible to remediate existing data to meet more modern standards. One of the most important accomplishments was updating the PDB to use IUPAC nomenclature for standard amino acids and nucleotides (Henrick *et al.*, 2008). Other efforts resulted in an incrementally improved corpus of data. Structures that had been represented in multiple, inconsistent ways, for example, peptides and viruses, were corrected, and curation of data going forward was improved (Lawson *et al.*, 2008; Dutta *et al.*, 2014).

During this same era, the requirement for creating more stringent validation criteria emerged from the community. An important milestone was reached in 2008 when all crystallographic depositions were required to be accompanied by structure factors (Wlodawer *et al.*, 2008); in 2010, chemical shifts were required for

https://doi.org/10.1017/S0033583518000057 Published online by Cambridge University Press

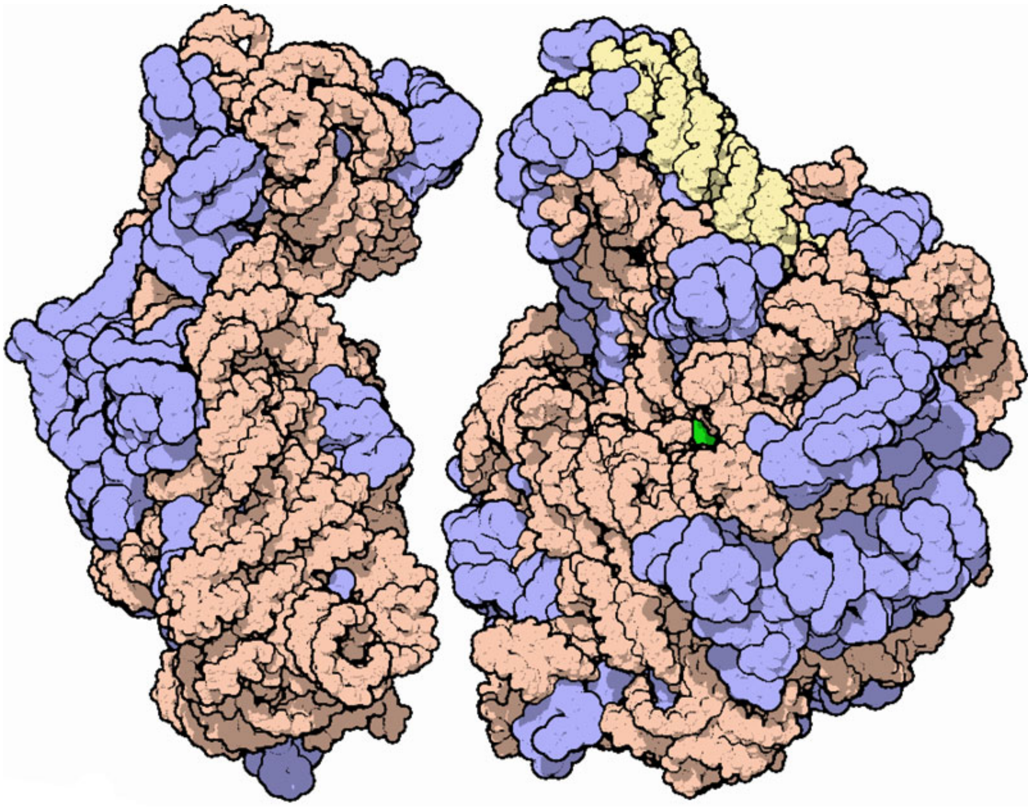

Fig. 5. Ribosome subunits. The small subunit is shown on the left and the large on the right (Ban *et al.*, 2000; Carter *et al.*, 2000; Schluenzen *et al.*, 2000). The protein is shown in blue and the RNA in orange and yellow. Taken from the Molecule of the Month (Goodsell, 2000b).

NMR structures. There was also increasing concern about the possibility that fraudulent structures had become a part of the archive. In 2008, the first of many method-specific wwPDB sponsored Validation Task Forces (VTFs) was set up. The charge to the X-ray VTF was to make recommendations to the wwPDB about validation of structures determined by that method. The X-ray VTF examined all available methods, tested them on the entire archive and reported their findings in a paper published in Structure (Read *et al.*, 2011). Their recommendations became the basis of the wwPDB OneDep Validation module (Gore *et al.*, 2017).

In this section, we have demonstrated how the PDB content and policies have evolved over the last 45 years and how the PDB has been agile in responding to rapid and unexpected scientific advances, technical improvements and strongly held beliefs of many stakeholders. Long before the introduction of the 'FAIR' guiding principles (Wilkinson *et al.*, 2016), the PDB archive has been making the results of structural biology investigations Findable, Accessible, Interoperable and Reusable.

Structural genomics and the Structural Biology Knowledgebase (SBKB)

The PDB contains many related structures, including homologs from different organisms, biomolecular complexes with different ligands, and even systematic small mutations of proteins introduced to investigate the effect on folding and activity; for example, PDB contains 566 structures of Bacteriophage T4 lysozyme variants (Matthews, 1996) and more than 250 structures of small molecule – HIV protease complexes (Wlodawer, 2002). The protein structure initiative (PSI) was launched to enable the determination of unique and diverse structures on a genomic scale (Norvell & Berg, 2007). The first phase focused on determining structures of proteins with extremely low sequence similarity to known structures, with the goal of finding new folds. The second phase focused on biology and linked the high throughput centers with projects on specific biological problems that would benefit from systematic structural approaches. For example, there were substantial gains made in determining structures of previously intractable membrane proteins (Pieper *et al.*, 2013). New high-throughput approaches were developed that allowed for advances in every part of the structure determination pipeline, including methods for producing pure protein samples, robotic crystallization, robotic crystal mounting and positioning and automated structure determination. Counter to some earlier concerns, the quality of the structures improved and the cost per structure determination decreased significantly (Grabowski *et al.*, 2016).

To meet the data management requirements of the PSI project, SBKB was created in 2008 (Berman *et al.*, 2009; Gabanyi *et al.*, 2011). The SBKB consisted of several modules that addressed the varying needs of the PSI project, described below.

TargetTrack provided information about the status of over 330 000 targets studied by the PSI Centers, including selection rationale, histories of protein production trials, and structure determination and deposition. It also collected and made public more than a thousand protocols routinely used by the centers, with variations noted on a trial-by-trial basis. Sequence-based annotations were also calculated and aggregated into each TargetTrack record. The data collected by TargetTrack were usually the first pieces of information available about a given sequence; to share

https://doi.org/10.1017/S0033583518000057 Published online by Cambridge University Press

it in the public domain, not only within the PSI Network, was unprecedented at that time.

A Technology Portal provided reports about the various technologies being developed to enable high-throughput protein production and structure determination (Gifford *et al.*, 2012). Summaries of over 450 novel technologies or protocols, along with their use cases, contact information, and references were collected. Categorization by experimental step enabled researchers to find new ideas for overcoming barriers that they could translate into their own laboratory.

Biosync (Kuller *et al.*, 2002; Flippen-Andersen *et al.*, 2010) became a module of the SBKB. This data resource collects synchrotron beamline parameters and experimental capabilities, and tracks the number of structures released per facility and beamline.

The Publication Portal tracked PSI publications along with their citations and journal impact factors. To date, 80% of the 2300+ articles published by the PSI have at least 5 citations.

The PSI Materials Repository, collected 90 000+ clones and 120 novel cloning and expression vectors created by the PSI centers and distributed them to researchers all over the world (Seiler *et al.*, 2014).

The Protein Model Portal (PMP) (Bordoli & Schwede, 2012) was created to help researchers locate homology models based on experimentally determined structures, thus further leveraging their impact. Users search the PMP by sequence or UniProt identifier, retrieving a list from among 22.8 million homology models pre-computed by Swiss-Model Repository (Kopp & Schwede, 2004), MODBASE (Pieper *et al.*, 2009), and the modeling groups within the PSI centers, as well as experimental structures from the PDB. A graphical map indicated how much of the sequence was covered by an experimental structure or derived from a model, and quality estimates were provided regarding the reliability of a model. If no model existed, new models could be requested and calculated by 6 public modeling servers. In 2013, the PMP group, with the support of the PSI and modeling community, created the Model Archive (Haas & Schwede, 2013). This new archive stores the computational model coordinates and details about assumptions, parameters and constraints applied in modeling. The Model Archive is open to all modelers and provides stable identifiers within publications as well as data storage and access in the public domain. To develop validation criteria for the modeling community, the PMP also constructed the Continuous Automated Model Evaluation (CAMEO) (Haas *et al.*, 2013) server that continuously evaluates the accuracy of predicted models, thus fostering the development of better modeling techniques.

The SBKB website integrated the results of the PSI with over 100 publicly available sequence, structure, function, proteomics and medicine databases. A search for any given protein sequence yielded all relevant annotations or products, presenting a view of what information was known, or still to be discovered. All structures, models, targets, and clones >40% identical in sequence were returned to allow for new connections to be found within the data. If a particular sequence yielded no annotations through the SBKB, users could nominate it for structure determination through the community-nomination portal, where users would be matched to collaborate with a PSI center. As the outreach arm of the PSI project, the SBKB also partnered with the Nature Publishing Group (now Macmillan Group) to write 320 research highlights on PSI advances for the SBKB portal. David Goodsell, the author of the RCSB PDB's Molecule of the Month series (Goodsell *et al.*, 2015), also created 90 illustrated essays of key PSI structures. PSI workshops were also archived on the SBKB.

By mid-2017, the PSI program produced 6920 structures, contributing over 5% of the current PDB archive (Table 1). Nearly 80% of these entries were distinct from each other and had less than 30% sequence identity to any structure pre-existing in the PDB (Dessailly *et al.*, 2009). A total of 600 structures were motivated by community requests. During PSI:Biology (2010–2015), the 9 membrane protein centers determined 160 structures and developed ∼40 novel technologies/methods for this difficult-to-determine class of proteins. Although the PSI program was terminated in 2015, the high throughput methods that enabled its productivity have endured. The SBKB is no longer operational following the end of the PSI program, but some of the modules continue to be available, including Protein Modeling Portal (Haas *et al.*, 2013) and Biosync (Flippen-Andersen *et al.*, 2010). The TargetTrack dataset has been archived (doi: 10.5281/zenodo.821654).

Electron Microscopy Data Bank

Bacterial rhodopsin was the first structure determined by electron microscopy deposited into the PDB (Henderson *et al.*, 1990). Because electron crystallography was used, it was possible for the PDB to curate the entry using a variation of the procedure for structures determined by X-ray crystallography. The determination of structures by cryo-electron microscopy (3DEM) became popular in the 2000s as software for reconstruction of 3D density maps from 2D single-particle images became available, even though the level of detail produced was typically limited (Chiu *et al.*, 2005). 3DEM scientists began to determine the overall shapes of large macromolecular complexes that could not be crystallized, opening up an important new avenue for structural biology investigations. The maps derived from 3DEM experiments could frequently be fitted with structures derived from X-ray crystallography, NMR spectroscopy or homology modeling, yielding 'pseudo-atomic' models that were able to provide useful insights and leads for further research (Rossmann *et al.*, 2005).

Table 1. Summary statistics of structures and other research products produced by the Protein Structure Initiative (PSI), 2000–2017

Products of the protein structure initiative	Total number
PSI structures	6920
Distinct structures	5472
Community-nominated structures	599
Membrane proteins	148
Homology models	22.8 M
Targets selected	335 714
Technology reports	458
Publications	2313
Publication with ⩾5 citations	1926
Citations of PSI publications	117 611
Research highlights from Nature Publishing	320
Illustrated featured molecules/systems	90

https://doi.org/10.1017/S0033583518000057 Published online by Cambridge University Press

In 2002, a new data archive called EM Data Bank (EMDB) containing maps and metadata was established at the EMBL-EBI (Editorial, 2003; Henrick *et al.*, 2003). Structures determined by 3DEM methods began to be deposited with maps archived in EMDB and models separately archived in PDB. An initial dictionary of data terms to describe 3DEM experiments was drafted jointly by the groups at EBI and RCSB, based on requirements provided by the 3DEM community in a series of international workshops. In 2006, the EBI and RCSB groups joined forces with Wah Chiu at the National Center for Macromolecular Imaging (NCMI) to create a 'one stop shop' for deposition and retrieval of maps and models at EMDataBank.org (Lawson *et al.*, 2011). Both groups launched 'serial' map + model deposition and annotation systems that directed users first to deposit their maps to EMDB using EmDep (Henrick *et al.*, 2003) and second to deposit their models to PDB with the transfer of relevant experimental metadata, as defined in the 3DEM data dictionary. The serial systems worked remarkably well, even though the underlying coordinate deposition and processing systems at the two sites were substantially different (see the section The synergies of science, technology and community in the development of the PDB). Over a 9-year period (2008–2015), nearly 4000 3DEM maps and 1000 3DEM models were processed in this manner. Truly joint map + model deposition for 3DEM structures was instantiated in 2016 with the OneDep system recently implemented by the wwPDB (Young *et al.*, 2017).

There has been substantial growth in 3DEM derived structures over the past few years (Fig. 6). Major technological advances, including the introduction of the direct electron detector and better data processing methods, have enabled the determination of structures derived from 2D single-particle images to near-atomic resolution, making it increasingly possible to visualize amino acid sidechains and nucleotide bases (Vinothkumar & Henderson, 2016). The award of the 2017 Nobel Prize in Chemistry to 3DEM pioneers Henderson, Frank and Dubochet recognized the potential of this rapidly evolving method to contribute to structural biology. Figure 7 provides several examples of maps deposited into EMDB just in the past year, each with a reported resolution of 4·5 Å or better.

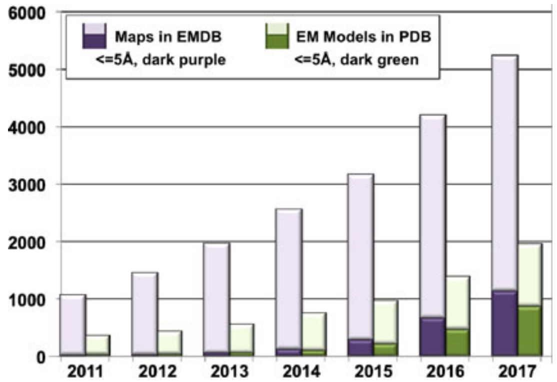

Fig. 6. Cumulative growth of 3DEM Structures. The number of structures available in EMDB for each recent year is indicated in purple (resolution better than 5·0 Å, dark purple); the number of EM-derived models available in PDB is indicated in green (resolution better than 5·0 Å, dark green).

The deluge of high-resolution 3DEM structures has made it a priority to establish robust validation methods for 3DEM derived

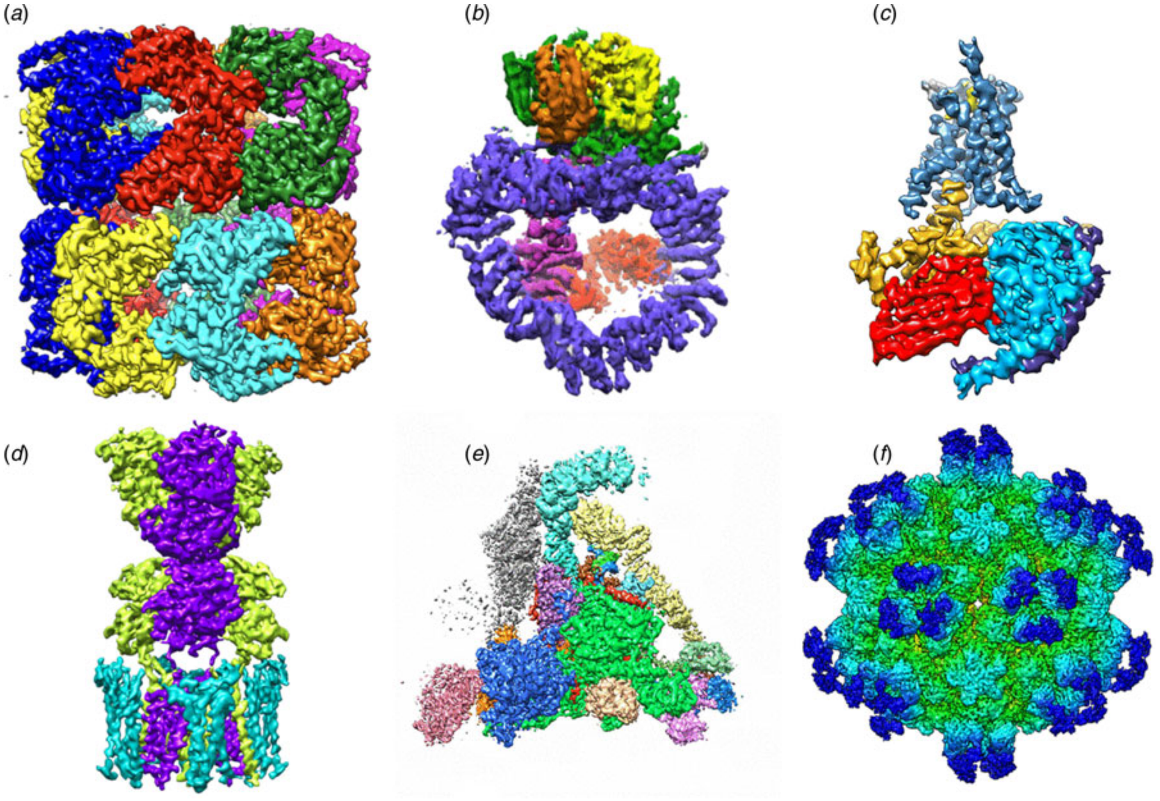

Fig. 7. Sampling of 3DEM structures recently released in EMDB: (a) GroEL (Roh *et al.*, 2017), EMD-8750 (b) DNA Protein Kinase (Sharif *et al.*, 2017), EMD-8751 (c) Heterotrimeric Gs protein complex (Liang *et al.*, 2017), EMD-8623 (d) Glutamate A2 receptor (Twomey *et al.*, 2017), EMD-8823 (e) Spliceosome (Wan *et al.*, 2016), EMD-9525 (f) Rhinovirus/Fab complex (Dong *et al.*, 2017), EMD-8763.

https://doi.org/10.1017/S0033583518000057 Published online by Cambridge University Press

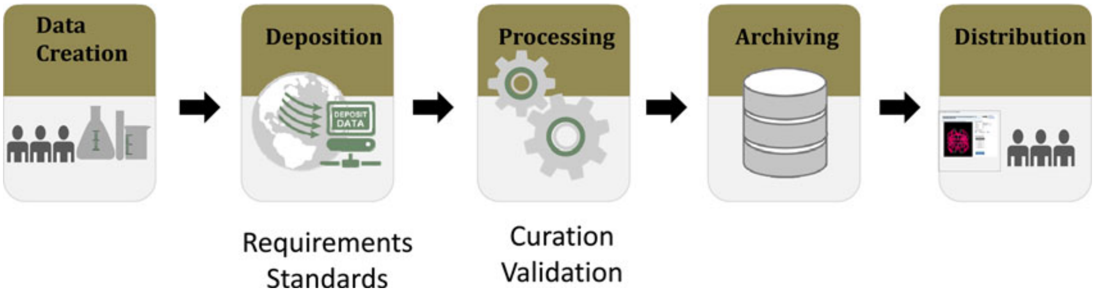

Fig. 8. The Data Processing Pipeline, from Data Creation through Distribution. Each component of the PDB pipeline is described in the section The current PDB pipeline.

maps and models. With OneDep now providing the facilities for 3DEM deposition, the current focus of EMDataBank.org is on enabling development of validation methods for 3DEM.

The current PDB pipeline

The PDB is responsible for collecting data entries from structural biologists and distributing curated data entries to users. To accomplish this goal, it is necessary to implement a data management pipeline with components for data deposition, curation, validation, archiving and distribution. Over time, data management has changed. Next, we describe current practices in the PDB data management pipeline (Fig. 8).

Requirements

In addition to the atomic coordinates, a considerable body of metadata is collected to describe how the coordinates were derived. The metadata are based on the details of each experimental method currently supported by the PDB: X-ray crystallography, NMR spectroscopy, and electron microscopy. Table 2 provides a summary of the various aspects of each method that need to be considered for data deposition.

Decisions about which data items must be collected are made in consultation with the community *via* the respective wwPDB Task Forces. Because the science, technology development and community sentiment change over time, the scope and level of granularity of the data to be collected also change over time. It is notable that protein production procedures are currently not collected. The PSI did, in fact, have procedures in place for collecting protein production protocols through TargetTrack (see the section Structural genomics and the structural biology knowledgebase (SBKB) above). However, compliance from the community was poor, which suggests that the time was not right for collecting and archiving protein production data.

Standards

To make the PDB archive computer searchable, it is essential that there are clear definitions for each data item collected. The PDBx/mmCIF format that is entirely computer readable is now the PDB Master Format. The data dictionary contains the definitions for all of the methods currently supported by the PDB (mmcif.wwpdb.org). The dictionary is extensible and allows for changes in existing methods and inclusion of new methods. A standing committee reviews the changing requirements and when necessary

Table 2. Experimental metadata requirements for the methods currently supported by the PDB

Method	X-ray crystallography	NMR spectroscopy	3D electron microscopy
Sample	• Buffer • Crystallization procedure	• Buffer • Isotope labeling	• Buffer • Sample support • Vitrification
Experiment	• Sample conditions • X-ray source • Detector • Collection protocol	• Sample conditions • Spectrometer • Acquisition parameters	• Sample conditions • Electron source • Detector • Imaging parameters
Measurements/data	• Diffraction images • Structure factors • Processing software • Statistics (resolution, Rsym)	• Resonance spectra • Resonance assignments • Chemical shifts • Contraints • Processing protocol	• Particle images • Final 3D map • Processing software • Processing protocol • Resolution (FSC)
Structure modeling	• Structure solution method • Refinement software, restraints • Fit-to-data statistics	• Structure calculation method • Refinement software	• Modeling method • Model source • Fitting software

https://doi.org/10.1017/S0033583518000057 Published online by Cambridge University Press

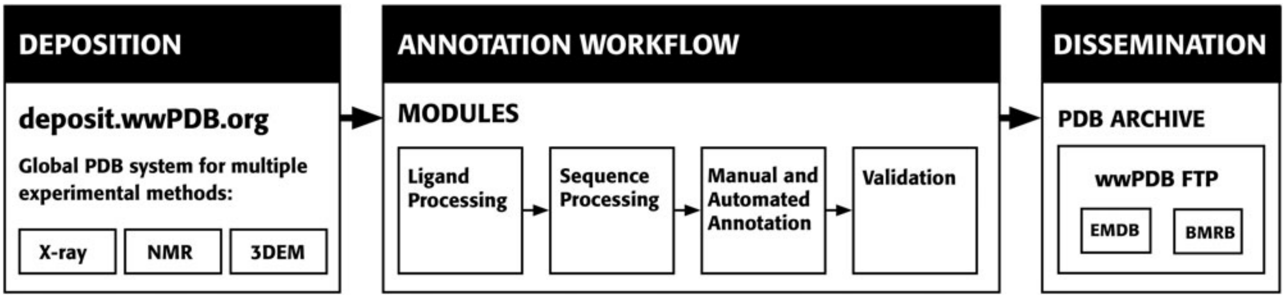

Fig. 9. OneDep System. Deposition is provided for X-ray, 3DEM and NMR. The annotation pipeline is made up of several modules that check the chemistry of the components, add new annotations and validate the structural model against standard geometries and the experimental data.

adds new definitions. In anticipation of changing needs, the dictionary also contains definitions for data items not currently in the PDB archive.

Data curation

All PDB entries are extensively curated. Many different aspects of the structure are carefully checked using a modular series of computational tools. For the polymer sequence, the following tasks are performed: cross-checks of author-provided sample sequence and coordinate sequence *versus* the sequence database, cross-checks of author-provided source organism *versus* the taxonomy database, assignments of database references and taxonomy identifiers to modeled protein polymer entities, and annotation of sequence discrepancies between sample sequence and database reference. For ligands, a search is performed to determine whether the ligand geometry is novel or equivalent to one of the ligands found in existing PDB entries. The ligand geometry is checked using a variety of 2D and 3D views. Derived data including the biological assembly are determined.

Data validation

Data in the PDB are validated according to recommendations made by Validation Task Forces that are convened by the wwPDB. Because X-ray crystallography is the oldest method supported by the PDB, its community has had the time and experience to develop the most extensive validation procedures (Read *et al*., 2011). The wwPDB has implemented the recommendations of the X-ray VTF directly into the data processing pipeline. Covalent geometry is checked against established standards. Intermolecular and intramolecular geometries of the polymer chains are checked for clashes using Molprobity (Chen *et al*., 2010). The geometry of ligands is checked against standards derived from small molecule structures archived in the CCDC (Bruno *et al*., 2004). The deposition of structure factors allows the checking of real space R factors for each residue and each ligand. A Validation Report is produced with the detailed analysis of the geometrical features of the model as well as the fit of the structure to the underlying experimental data. The graphical representation in the form of sliders gives a summary of the quality of the structure.

Validation of NMR derived structures follows the recommendation of the NMR VTF (Montelione *et al*., 2013). The model geometry is checked in the same way as for X-ray derived structures. Consistency checks across models are carried out for NMR structures along with an examination of outliers in NMR restraints. For 3DEM-derived structures, the 3DEM VTF recommended that the validation of model geometry follow the same criteria developed for X-ray derived structures and that new methods be developed for 3DEM map validation and map-to-model fit (Henderson *et al*., 2012). One of the ways to achieve this goal involves engaging the community in EM Challenges (Lawson *et al*., 2016), where participants attempt to fit models to benchmarked maps, followed by an assessment of the results. These exercises are likely to result in more robust methods for validating 3DEM structural models.

To enable efficient data deposition, curation and processing, a new tool called OneDep was developed by the wwPDB (Young *et al*., 2017) (Fig. 9). OneDep has a Deposition and Annotation Workflow system containing the modules required for making data curation as thorough and automatic as possible. Skilled wwPDB biocurators review all of the results of data processing and work with the depositors to ensure the best possible representation of the submitted data.

Archiving

Once the data are processed, the files are put into a temporary archive until they are ready for release, usually upon publication of the structure. The released structures reside in the PDB Archive, which can be accessed using methods such as the File Transport Protocol (FTP) and rsync. The PDB Archive consists of flat files that contain several types of data, including atomic coordinates, a molecular description of macromolecules and ligands, metadata describing the experimental method, and experimental data including structure factors, chemical shifts and restraints. 3DEM map data are curated by EMDB partner sites and archived under a separate, parallel branch of the archive. The PDB Archive is mirrored by all three wwPDB partners.

Data distribution

The PDB is distributed in several ways. Data can be downloaded *via* rsync or ftp protocols following the directions provided on the wwPDB website (https://www.wwpdb.org/download/downloads). In addition, each of the wwPDB data centers has websites that provide a multitude of services including downloading, searching and browsing (Berman *et al*., 2000; Ulrich *et al*., 2008; Velankar *et al*., 2016; Kinjo *et al*., 2017; Rose *et al*., 2017). Coordinate sets are currently downloaded from the wwPDB FTP and websites more than 550 000 000 times per year.

The future: integrative hybrid (I/H) methods

Traditionally, each PDB entry contains an atomic structural model derived from a single structure determination method,

https://doi.org/10.1017/S0033583518000057 Published online by Cambridge University Press

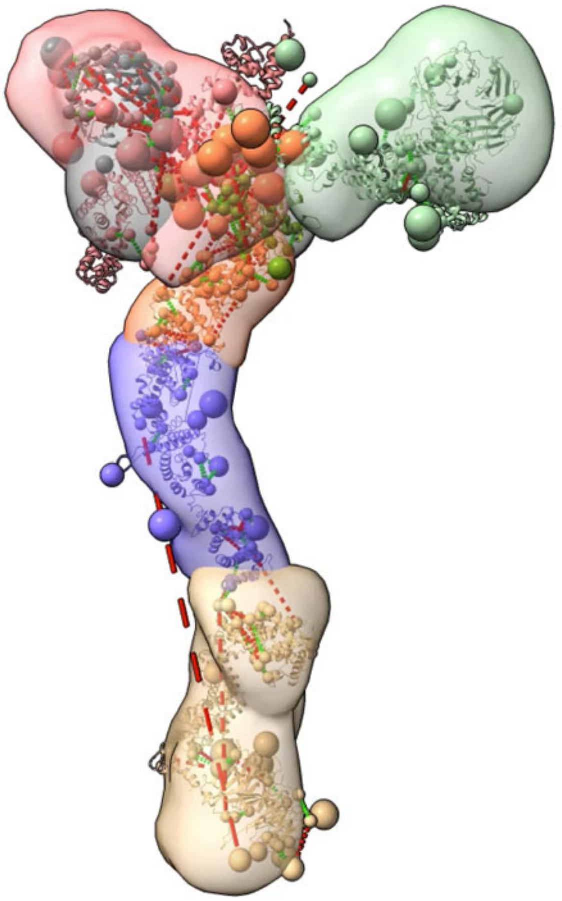

Fig. 10. I/H model of the Nup84 sub-complex from the Nuclear Pore Complex (Shi *et al.*, 2014) available from PDB-Dev (Burley *et al.*, 2017; Vallat *et al.*, 2018), PDB-Dev ID: PDBDEV_00000001. Multi-scale structural model of the heptameric Nup84 sub-complex is shown (colored ribbons and spheres) along with the localization densities of the sampled structures (colored contoured surfaces). The model is obtained using the Integrative Modeling Platform (IMP) software (Russel *et al.*, 2012) and visualized using ChimeraX software (Goddard *et al.*, 2018).

including X-ray crystallography, NMR spectroscopy and 3D electron microscopy. Recently, I/H methods have been developed that simultaneously use data from multiple experimental techniques to compute structures of single macromolecules or macromolecular complexes (Ward *et al.*, 2013). In some cases, data from a primary method such as NMR are combined with additional information obtained from a secondary method such as small-angle solution scattering (SAS). In other cases, information from multiple experimental sources, such as Fluorescence resonance energy transfer (FRET), SAS, chemical crosslinking (CX) and mass spectrometry (MS) are pooled together to derive a set of spatial restraints that enable computation of a structural model. Combining multiple complementary experimental methods makes it possible to determine structures of large macromolecular machines that have previously eluded traditional structure determination methods. I/H methods have led to the elucidation of structures of macromolecular assemblies such as the nuclear pore complex (Alber *et al.*, 2007*a*, 2007*b*) and its sub-complexes (Fig. 10, (Kim *et al.*, 2014; Shi *et al.*, 2014)), the type III secretion system needle (Loquet *et al.*, 2012), the proteasomal lid complex (Politis *et al.*, 2014), the exosome complex (Shi *et al.*, 2015) and the mediator complex (Robinson *et al.*, 2015). Although many important structures have been determined using I/H methods, there are no standard mechanisms to archive these structures and make them available to the public. An important distinction between structural models obtained through I/H methods and the atomistic models currently archived in the PDB is that I/H models are often coarse-grained. The existing PDB data pipeline expects fully atomistic models and hence cannot process coarse-grained I/H models.

In 2014, 38 experimental and computational scientists assembled at the EMBL-EBI to discuss how best to archive the results of I/H structure determinations. The wwPDB I/H methods Task Force (I/HTF) made the following series of recommendations that would enable the wwPDB to address this problem (Sali *et al.*, 2015): (1) a flexible model representation should be developed, allowing for multi-scale models (with atomistic and non-atomistic coarse-grained representations), multi-state models (existing in various conformations), ensembles of models, and models related by time or other order; (2) procedures for estimating the uncertainty of integrative models should be developed, validated, and adopted; (3) all relevant experimental data and metadata as well as experimental and computational protocols should be archived; (4) a Federation of model and data archives should be created; and (5) publication standards for integrative models should be established.

To address these recommendations, two subgroups of the I/HTF have been established: the Model Validation Subgroup and the Federation Subgroup. The concept of a Federation of model and data repositories would allow individual disciplines to create appropriate repositories for their experimental data based on the requirements of their communities. Mechanisms for data exchange would promote seamless interoperation among the federated repositories (Fig. 11).

Following the recommendations of the I/HTF, a preliminary dictionary has been created to address the flexible data representation required to describe I/H results (Berman *et al.*, 2016; Vallat *et al.*, 2017, 2018). This dictionary is a modular extension of the PDBx/mmCIF dictionary (Fitzgerald *et al.*, 2005) used by the PDB archive and contains data definitions necessary to describe the details of I/H models, associated spatial restraints and modeling protocols. The newly developed I/H methods extension dictionary provides the fundamental data specifications required for archiving I/H models. Based on this dictionary extension, a prototype pipeline called PDB-Development (PDB-Dev; pdb-dev.wwpdb.org) has been built to enable testing and development of deposition and archiving for I/H structural models (Burley *et al.*, 2017; Vallat *et al.*, 2018). Fifteen I/H models obtained using different modeling software such as the Integrative Modeling Platform (IMP) (Russel *et al.*, 2012), Rosetta (Leaver-Fay *et al.*, 2011), HADDOCK (Dominguez *et al.*, 2003), TADbit (Serra *et al.*, 2017) and XPLOR-NIH (Schwieters *et al.*, 2018) have been deposited into PDB-Dev in a format compliant with the I/H methods dictionary. These include the Nup84 sub-complex of the nuclear pore complex (Shi *et al.*, 2014), the exosome complex (Shi *et al.*, 2015), the mediator complex (Robinson *et al.*, 2015), lysine-linked Diubiquitin complex (Liu *et al.*, 2018), structures of the human serum albumin domains in their native environment (Belsom *et al.*, 2016), the chromatin model of the first 4·5Mb of chromosome 2L from *Drosophila Melanogaster* (Trussart *et al.*, 2015) and the ribosomal RNA small subunit methyltransferase A complexed with 16S ribosomal RNA (van Zundert *et al.*, 2015). These structures are now publicly available from the PDB-Dev website (Vallat *et al.*, 2018;

https://doi.org/10.1017/S0033583518000057 Published online by Cambridge University Press

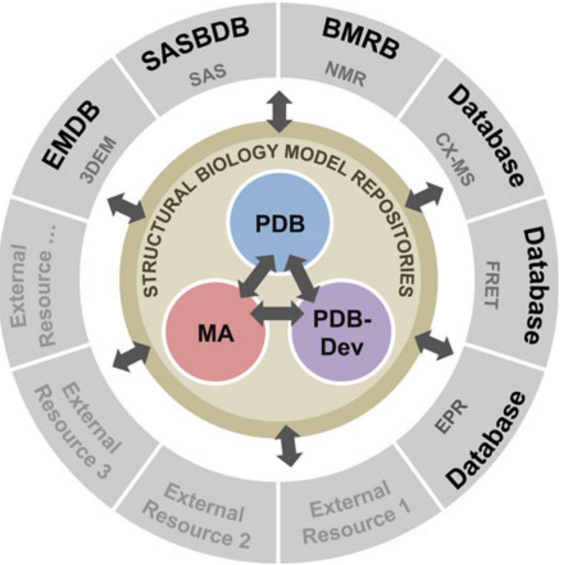

Fig. 11. Conceptual diagram of the I/H Methods Federation. At the center are the three structural biology model repositories: the PDB archives experimentally determined structures of macromolecules (Berman *et al.*, 2000); the Model Archive (MA), part of the Protein Model Portal (PMP), archives *in silico* structural models (Bordoli and Schwede, 2012; Haas and Schwede, 2013; Haas *et al.*, 2013); and PDB-development (PDB-Dev) is the prototype system for archiving I/H models (Vallat *et al.*, 2018; Burley *et al.*, 2017). The outer circle consists of experimental data repositories that contribute to structural biology. Only a limited set of experimental data archives have been identified at present and many others may be included as the field evolves and the respective research communities build their own repositories.

Burley *et al.*, 2017) and can be downloaded and visualized using the ChimeraX software (Goddard *et al.*, 2018) (Fig. 10).

The lessons learned from creating and maintaining the PDB archive are informing the process of developing the PDB-Dev system for archiving I/H structures. To adapt to the evolving needs of the scientific community, many important tasks have been accomplished: consulting with the community to determine requirements, carefully creating standard dictionary definitions and making sure that those dictionary standards are extensible. Once the PDB-Dev system is fully developed, it will be straightforward to include structures derived from I/H methods in the PDB archive, thus making the rich content from structures of complex macromolecular machines available to PDB users.

Conclusion

In this review, we describe the interplay among science, technology and community in creating data resources. The way in which the PDB developed in many ways follows the principles set forth by Elinor Ostrom for the management of natural resources (Ostrom, 1990). Those principles emphasize that bottom-up collective action can work better than top-down enforcement. Although building a community resource in this way can take much longer, the involvement of the various stakeholders in meaningful ways can better ensure its sustainability.

Domain repositories such as the PDB are key to the conduct of science and development of scientific knowledge. Preserving the data and making it freely available enables reproducibility and the ability to build on previous work to carry out new research. Structural biologists were early adopters of the concept of archiving as being an integral part of the research and publication life cycle. Not only has the availability of data helped enable further discoveries in the field, but it also has allowed computational biologists to analyze the entire corpus of data to understand the underlying principles that govern protein folding and interactions; it is impossible to imagine structural bioinformatics without the PDB. The PDB thus provides a compelling roadmap that could be applied to all of science.

Acknowledgements. We thank the wwPDB data center members, the EMDataBank group and the SBKB partners, with special thanks to Stephen Burley and Wah Chiu for their leadership of the PDB and EM projects and John Westbrook for his vision with respect to the mmCIF effort. This work has been supported by grants to the RCSB PDB from NSF, NIH and DOE (DBI-1338415), SBKB (U01 GM093324), EMDataBank (R01 GM079429), I/H methods (NSF EAGER award DBI-1519158), and the Enabling Data Science in Biology BD2K curriculum development project (R25 LM012286).

References

Alber F *et al.* (2007*a*) Determining the architectures of macromolecular assemblies. *Nature* **450**, 683–694.

Alber F *et al.* (2007*b*) The molecular architecture of the nuclear pore complex. *Nature* **450**, 695–701.

Allen FH *et al.* (1973) Cambridge crystallographic data centre. II. Structural data file. *Journal of Chemical Documentation* **13**, 119–123.

Arnold E and Rossmann MG (1988) The use of molecular-replacement phases for the refinement of the human rhinovirus 14 structure. *Acta Crystallographica Section A* **44**, 270–282.

Ban N *et al.* (2000) The complete atomic structure of the large ribosomal subunit at a 2·4 Å resolution. *Science* **289**, 905–920.

Barinaga M (1989) The missing crystallography data. *Science* **245**, 1179–1181.

Belsom A *et al.* (2016) Serum albumin domain structures in human blood serum by mass spectrometry and computational biology. *Molecular and Cellular Proteomics* **15**, 1105–1116.

Berman HM, Henrick K and Nakamura H (2003) Announcing the worldwide Protein Data Bank. *Nature Structural Biology* **10**, 980.

Berman HM *et al.* (2000) The Protein Data Bank. *Nucleic Acids Research* **28**, 235–242.

Berman HM *et al.* (2009) The protein structure initiative structural genomics knowledgebase. *Nucleic Acids Research* **37**, D365–D368.

Berman HM *et al.* (2016) A data dictionary for archiving integrative/hybrid models. In *66th Annual Meeting of the American Crystallographic Association*, Denver, CO, USA, pp. 85-SA.

Blake CCF *et al.* (1965) Structure of hen egg-white lysozyme. A three dimensional Fourier synthesis at 2 Å resolution. *Nature* **206**, 757–761.

Bordoli L and Schwede T (2012) Automated protein structure modeling with SWISS-MODEL workspace and the Protein Model Portal. *Methods in Molecular Biology* **857**, 107–136.

Bruno IJ *et al.* (2004) Retrieval of crystallographically-derived molecular geometry information. *Journal of Chemical Information and Computer Sciences* **44**, 2133–2144.

Burley SK *et al.* (2017) PDB-Dev: a prototype system for depositing integrative/hybrid structural models. *Structure* **25**, 1317–1318.

Carter AP *et al.* (2000) Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. *Nature* **407**, 340–348.

Chen VB *et al.* (2010) Molprobity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography* **66**, 12–21.

Chiu W *et al.* (2005) Electron cryomicroscopy of biological machines at subnanometer resolution. *Structure* **13**, 363–372.

Dessailly BH *et al.* (2009) PSI-2: structural genomics to cover protein domain family space. *Structure* **17**, 869–881.

Dickerson RE *et al.* (1982) The anatomy of a-DNA, B-DNA, and Z-DNA. *Science* **216**, 475–485.

Dominguez C, Boelens R and Bonvin AM (2003) HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society* **125**, 1731–1737.

https://doi.org/10.1017/S0033583518000057 Published online by Cambridge University Press

Dong Y *et al.* (2017) Antibody-induced uncoating of human rhinovirus B14. *Proceedings of the National Academy of Sciences of the United States of America* **114**, 8017–8022.

Dutta S *et al.* (2014) Improving the representation of peptide-like inhibitor and antibiotic molecules in the Protein Data Bank. *Biopolymers* **101**, 659–668.

Editorial (2003) A database for 'em. *Nature Structural Biology* **10**(5), 313.

Erickson JW *et al.* (1985) The structure of a $T = 1$ icosahedral empty particle from southern bean mosaic virus. *Science* **229**, 625–629.

Fitzgerald PMD *et al.* (2005) 4·5 macromolecular dictionary (mmCIF). In Hall SR and McMahon B (eds), *International Tables for Crystallography G. Definition and Exchange of Crystallographic Data*. Dordrecht, The Netherlands: Springer, pp. 295–443.

Flippen-Andersen J, Gabanyi MJ, Chen L, Sala R, Westbrook JD and Berman HM (2010) BioSync: a structural biologist's guide to high energy data collection facilities, http://biosync.rcsb.org.

Gabanyi MJ *et al.* (2011) The Structural Biology Knowledgebase: a portal to protein structures, sequences, functions, and methods. *Journal of Structural and Functional Genomics* **12**, 45–54.

Gifford LK *et al.* (2012) The protein structure initiative structural biology knowledgebase technology portal: a structural biology web resource. *Journal of Structural and Functional Genomics* **13**, 57–62.

Goddard TD *et al.* (2018) UCSF chimerax: meeting modern challenges in visualization and analysis. *Protein Science* **27**, 14–25.

Goodsell D (2000a) Nucleosome. PDB-101 Molecule of the Month series. doi:10.2210/rcsb_pdb/mom_2000_7.

Goodsell D (2000b) Ribosomal subunits. PDB-101 Molecule of the Month series. doi:10.2210/rcsb_pdb/mom_2000_10.

Goodsell D (2001) DNA. PDB-101 Molecule of the Month series. doi:10.2210/rcsb_pdb/mom_2001_11.

Goodsell DS *et al.* (2015) The RCSB PDB "molecule of the month": inspiring a molecular view of biology. *PLoS Biology* **13**, e1002140.

Gore S *et al.* (2017) Validation of the structures in the Protein Data Bank. *Structure* **25**, 1916–1927.

Grabowski M *et al.* (2016) The impact of structural genomics: the first quindecennial. *Journal of Structural and Functional Genomics* **17**, 1–16.

Haas J and Schwede T (2013) Model Archive. http://www.modelarchive.org/.

Haas J *et al.* (2013) The Protein Model Portal – a comprehensive resource for protein structure and model information. *Database (Oxford)* **2013**, bat031.

Hall SR, Allen FH and Brown ID (1991) The Crystallographic Information File (Cif) – a new standard archive file for crystallography. *Acta Crystallographica Section A* **47**, 655–685.

Hamlin RC (1985) Multiwire area X-ray diffractometers. *Methods in Enzymology* **114**, 416–452.

Harmsen A, Leberman R and Schulz GE (1976) Comparison of protein crystal diffraction patterns and absolute intensities from synchrotron and conventional x-ray sources. *Journal of Molecular Biology* **104**, 311–314.

Henderson R *et al.* (1990) Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *Journal of Molecular Biology* **213**, 899–929.

Henderson R *et al.* (2012) Outcome of the first electron microscopy validation task force meeting. *Structure* **20**, 205–214.

Hendrickson WA, Smith JL and Sheriff S (1985) Direct phase determination based on anomalous scattering. *Methods in Enzymology* **115**, 41–55.

Henrick K *et al.* (2003) EMDep: a web-based system for the deposition and validation of high-resolution electron microscopy macromolecular structural information. *Journal of Structural Biology* **144**, 228–237.

Henrick K *et al.* (2008) Remediation of the Protein Data Bank archive. *Nucleic Acids Research* **36**(Database issue), D426–D433.

Hope H (1988) Cryocrystallography of biological macromolecules: a generally applicable method. *Acta Crystallographica Section B: Structural Science* **44**, 22–26.

Hopper P, Harrison SC and Sauer RT (1984) Structure of tomato bushy stunt virus. V. Coat protein sequence determination and its structural implications. *Journal of Molecular Biology* **177**, 701–713.

Horst R *et al.* (2001) NMR structure reveals intramolecular regulation mechanism for pheromone binding and release. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 14374–14379.

Howard Hughes Medical Institute (2017). http://www.hhmi.org/about/history.

Hufton AL (2014) Sharing the structures. In Nature Milestones: Crystallography. (1970s) Open software and crystallographic databases. Nature, Scientific Data.

International Union of Crystallography (1989) Commission on Biological Macromolecules. *Acta Crystallographica Section A* **45**, 658.

Jones TA (1978) FRODO: a graphic model building and refinement system for macromolecules. *Journal of Applied Crystallography* **11**, 268–272.

Kartha G, Bello J and Harker D (1967) Tertiary structure of ribonuclease. *Nature* **213**, 862–865.

Kelly JA *et al.* (1979) X-ray crystallography of the binding of the bacterial cell wall trisaccharide NAM-NAG-NAM to lysozyme. *Nature* **282**, 875–878.

Kendrew JC *et al.* (1958) A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature* **181**, 662–666.

Kim SJ *et al.* (2014) Integrative structure-function mapping of the nucleoporin nup133 suggests a conserved mechanism for membrane anchoring of the nuclear pore complex. *Molecular and Cellular Proteomics* **13**, 2911–2926.

Kinjo AR *et al.* (2017) Protein Data Bank Japan (PDBj): updated user interfaces, resource description framework, analysis tools for large structures. *Nucleic Acids Research* **45**, D282–D288.

Kopp J and Schwede T (2004) The SWISS-MODEL Repository of annotated three-dimensional protein structure homology models. *Nucleic Acids Research* **32**, D230–D234.

Kuller A *et al.* (2002) A biologist's guide to synchrotron facilities: the BioSync web resource. *TIBS* **27**, 213–215.

Lawson CL *et al.* (2008) Representation of viruses in the remediated PDB archive. *Acta Crystallographica Section D: Biological Crystallography* **64**, 874–882.

Lawson CL *et al.* (2011) EMDatabank.org: unified data resource for CryoEM. *Nucleic Acids Research.* **39**, D456–D464.

Lawson CL *et al.* (2016) EMDatabank unified data resource for 3DEM. *Nucleic Acids Research* **44**, D396–D403.

Leaver-Fay A *et al.* (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods in Enzymology* **487**, 545–574.

Liang YL *et al.* (2017) Phase-plate cryo-EM structure of a class B GPCR-G-protein complex. *Nature* **546**, 118–123.

Liu Z *et al.* (2018) Characterizing protein dynamics with integrative use of bulk and single-molecule techniques. *Biochemistry* **57**, 305–313.

Loquet A *et al.* (2012) Atomic model of the type III secretion system needle. *Nature* **486**, 276–279.

Luger K *et al.* (1997) Crystal structure of the nucleosome core particle at 2·8 A resolution. *Nature* **389**, 251–260.

Matthews BW (1996) Structural and genetic analysis of the folding and function of T4 lysozyme. *FASEB Journal* **10**, 35–41.

Meyer EF (1997) The first years of the Protein Data Bank. *Protein Science* **6**(7), 1591–1597.

Montelione GT *et al.* (2013) Recommendations of the wwPDB NMR validation task force. *Structure* **21**, 1563–1570.

Norvell JC and Berg JM (2007) Update on the protein structure initiative. *Structure* **15**, 1519–1522.

Ostrom E (1990) *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, Cambridge, United Kingdom.

Patikoglou GA *et al.* (1999) TATA element recognition by the TATA box-binding protein has been conserved throughout evolution. *Genes and Development* **13**, 3217–3230.

Perutz MF *et al.* (1960) Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5 Å resolution, obtained by X-ray analysis. *Nature* **185**, 416–422.

Pettersen EF *et al.* (2004) UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of Computational Chemistry* **25**, 1605–1612.

Phillips DC (1972) Protein crystallography 1971: coming of age. In *Cold Spring Harbor Symposia on Quantitative Biology*, Cold Spring Harbor: Cold Spring Harbor Laboratory Press **36**, 589–592.

https://doi.org/10.1017/S0033583518000057 Published online by Cambridge University Press

Pieper U *et al.* (2009) MODBASE, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Research* **37**, D347–D354.

Pieper U *et al.* (2013) Coordinating the impact of structural genomics on the human alpha-helical transmembrane proteome. *Nature Structural and Molecular Biology* **20**, 135–138.

Politis A *et al.* (2014) A mass spectrometry-based hybrid method for structural modeling of protein complexes. *Nature Methods* **11**, 403–406.

Protein Data Bank. (1971) Crystallography: Protein Data Bank. Nature: New Biology **233**, 223–223.

Quiocho FA and Lipscomb WN (1971) Carboxypeptidase A: a protein and an enzyme. *Advances in Protein Chemistry* **25**, 1–78.

Read RJ *et al.* (2011) A new generation of crystallographic validation tools for the Protein Data Bank. *Structure* **19**, 1395–1412.

Rich A and Kim S-H (1978) The three-dimensional structure of transfer RNA. *Scientific American* **238**, 52–62.

Richards FM (1968) The matching of physical models to three-dimensional electron-density maps: a simple optical device. *Journal of Molecular Biology* **37**, 225–230.

Robertus JD *et al.* (1974) Structure of yeast phenylalanine tRNA at 3 Å resolution. *Nature* **250**, 546–551.

Robinson PJ *et al.* (2015) Molecular architecture of the yeast Mediator complex. *Elife* **4**, pii: e08719.

Roh SH *et al.* (2017) Subunit conformational variation within individual GroEL oligomers resolved by Cryo-EM. *Proceedings of the National Academy of Sciences of the United States of America* **114**, 8259–8264.

Rose PW *et al.* (2017) The RCSB Protein Data Bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Research* **45**, D271–D281.

Rossmann MG *et al.* (2005) Combining X-ray crystallography and electron microscopy. *Structure* **13**, 355–362.

Russel D *et al.* (2012) Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biology* **10**, e1001244.

Sali A *et al.* (2015) Outcome of the first wwPDB hybrid/integrative methods task force workshop. *Structure* **23**, 1156–1167.

Schluenzen F *et al.* (2000) Structure of functionally activated small ribosomal subunit at 3·3 Å resolution. *Cell* **102**, 615–623.

Schwieters CD, Bermejo GA and Clore GM (2018) Xplor-NIH for molecular structure determination from NMR and other data sources. *Protein Science* **27**, 26–40.

Seiler CY *et al.* (2014) DNASU plasmid and PSI:Biology-Materials repositories: resources to accelerate biological research. *Nucleic Acids Research* **42**, D1253–D1260.

Serra F *et al.* (2017) Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. *PLoS Computational Biology* **13**, e1005665.

Sharif H *et al.* (2017) Cryo-EM structure of the DNA-PK holoenzyme. *Proceedings of the National Academy of Sciences of the United States of America* **114**, 7367–7372.

Shi Y *et al.* (2014) Structural characterization by cross-linking reveals the detailed architecture of a coatomer-related heptameric module from the nuclear pore complex. *Molecular and Cellular Proteomics* **13**, 2927–2943.

Shi Y *et al.* (2015) A strategy for dissecting the architectures of native macromolecular assemblies. *Nature Methods* **12**, 1135–1138.

The Nobel Prize in Chemistry. (1962) http://www.nobelprize.org/nobel_prizes/chemistry/laureates/1962/.

Trussart M *et al.* (2015) Assessing the limits of restraint-based 3D modeling of genomes and genomic domains. *Nucleic Acids Research* **43**, 3465–3477.

Twomey EC *et al.* (2017) Channel opening and gating mechanism in AMPA-subtype glutamate receptors. *Nature* **549**, 60–65.

Ulrich EL *et al.* (2008) Biomagresbank. *Nucleic Acids Research* **36**, D402–D408.

Vallat B *et al.* (2017) A Data Dictionary For Archiving Integrative/Hybrid Models. In 24th IUCr Congress and General Assembly. International Union of Crystallography, Hyderabad, India.

Vallat B, Webb B, Westbrook JD, Sali A and Berman HM (2018) Development of a Prototype System for Archiving Integrative/Hybrid Structure Models of Biological Macromolecules. *Structure* **26** 894–904.

Van Zundert GCP, Melquiond ASJ and Bonvin A (2015) Integrative modeling of biomolecular complexes: HADDOCKing with cryo-electron microscopy data. *Structure* **23**, 949–960.

Velankar S *et al.* (2016) PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. *Nucleic Acids Research* **44**, D385–D395.

Vinothkumar KR and Henderson R (2016). Single particle electron cryomicroscopy: trends, issues and future perspective. *Quarterly Reviews of Biophysics* **49**, e13.

Wan R *et al.* (2016). Structure of a yeast catalytic step I spliceosome at 3·4 A resolution. *Science* **353**, 895–904.

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

Watson HC (1969). The stereochemistry of the protein myoglobin. *Progress in Stereochemistry* **4**, 299.

Westbrook JD and Fitzgerald PM D. (2009). Chapter 10 The PDB format, mmCIF formats, and other data formats. In Bourne PE and Gu J (eds), *Structural Bioinformatics*, 2nd Edn. Hoboken, NJ: John Wiley & Sons, Inc., pp. 271–291.

Wilkinson MD *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**, 160018.

Wlodawer A (2002) Rational approach to AIDS drug design through structural biology. *Annual Review of Medicine* **53**, 595–614.

Wlodawer A *et al.* (2008) Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *FEBS Journal* **275**, 1–21.

Wyckoff HW *et al.* (1967) The structure of ribonuclease-S at 6 Å resolution. *Journal of Biological Chemistry* **242**, 3749–3753.

Young JY *et al.* (2017) Onedep: unified wwPDB system for deposition, biocuration, and validation of macromolecular structures in the PDB archive. *Structure* **25**, 536–545.

https://doi.org/10.1017/S0033583518000057 Published online by Cambridge University Press