Quarterly Reviews of Biophysics

www.cambridge.org/qrb

Protein Data Bank (PDB): Fifty-three years young and having a transformative impact on science and society

Helen M. Berman[1,2,3] and Stephen K. Burley[1,2,4,5,6]

[1]Research Collaboratory for Structural Bioinformatics Protein Data Bank, Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, NJ, USA; [2]Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, NJ, USA; [3]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA, USA; [4]Rutgers Cancer Institute, Rutgers, The State University of New Jersey, New Brunswick, NJ, USA; [5]Rutgers Artificial Intelligence and Data Science (RAD) Collaboratory, Rutgers, The State University of New Jersey, Piscataway, NJ, USA and [6]Research Collaboratory for Structural Bioinformatics Protein Data Bank, San Diego Supercomputer Center, University of California San Diego, La Jolla, CA, USA

Review

Cite this article: Berman HM and Burley SK (2025). Protein Data Bank (PDB): Fifty-three years young and having a transformative impact on science and society. *Quarterly Reviews of Biophysics*, **58**, e9, 1–15
https://doi.org/10.1017/S0033583525000034

Received: 19 December 2024
Revised: 11 February 2025
Accepted: 12 February 2025

Keywords:
bioinformatics; computational biophysics; function; protein structure; structural biology

Corresponding authors:
Helen M. Berman and Stephen K. Burley;
Emails: berman@rcsb.rutgers.edu; stephen.burley@rcsb.org

Abstract

This review article describes the co-evolution of structural biology as a discipline and the Protein Data Bank (PDB), established in 1971 as the first open-access data resource in biology by like-minded structural scientists. As the PDB archive grew in size and scope to encompass macromolecular crystallography, NMR spectroscopy, and cryo-electron microscopy, new technologies were developed to ingest, validate, curate, store, and distribute the information. Community engagement ensured that the needs of structural biologists (data depositors) and data consumers were met. Today, the archive houses more than 230,000 experimentally determined structures of proteins, nucleic acids, and macromolecular machines and their complexes with one another and small-molecule ligands. Aggregate costs of PDB data preservation are ~1% of the cost of structure determination. The enormous impact of PDB data on basic and applied research and education across the natural and medical sciences is presented and highlighted with illustrative examples. Enablement of *de novo* protein structure prediction (AlphaFold2, RoseTTAfold, OpenFold, *etc.*) is the most widely appreciated benefit of having a corpus of rigorously validated, expertly curated 3D biostructure data.

Table of contents

Introduction

The Protein Data Bank (PDB) was the first open-access digital archive in biology; it was a vanguard in the open-access data movement (Berman, 2008; Protein Data Bank, 1971). Over the past fifty-three-plus years, it has co-evolved with the scientific research it supports and continually embraced new technologies for 3D biostructure data deposition, validation, biocuration, preservation, and dissemination.

In this review, we first trace how the contents of the PDB have grown in terms of the types of structures and the experimental methods used for determination. We show how cyberinfrastructure

has evolved in parallel to meet the needs of the ever-growing archive. We then describe how data standards and policies were established in collaboration with a growing cohort of stakeholders, thereby enabling the PDB to be a pioneer in embracing the FAIR (Findability, Accessibility, Interoperability, and Reusability (Wilkinson *et al.*, 2016)), FACT (FAIRness, Accuracy, Confidentiality, and Transparency (van der Aalst *et al.*, 2017)), and TRUST (Transparency, Responsibility, User focus, Sustainability, and Technology (Lin *et al.*, 2020)) principles emblematic of responsible data management. The costs of data capture, archiving, and delivery in accord with the FAIR and FACT principles are also discussed.

Thereafter, we describe the immense impact of the PDB on basic and applied research in virtually all areas of biology and medicine. The impact of the PDB on structure-guided drug discovery and vaccine development played central roles in helping to reduce the ravages of HIV and combat the COVID-19 pandemic. The existence of the PDB led to the creation of a new field of science–structural bioinformatics–that, in turn, yielded transformative advances in protein structure prediction and design. The impact of the PDB on the chemical, computational, mathematical, physical, and social sciences is also described.

Evolution and growth of the PDB

Content of the PDB

The types of structures archived in the PDB have evolved with the progress of structural biology. In the 1970s, atomic-level 3D biostructures were mostly smaller, single-domain globular proteins (Figure 1), such as myoglobin (Kendrew *et al.*, 1958; Kendrew et al., 1960), hemoglobin (Bolton and Perutz, 1970; Perutz *et al.*, 1960), lysozyme (Blake *et al.*, 1965), and ribonuclease (Kartha *et al.*, 1967; Wyckoff *et al.*, 1967). The first experimental structure of a large nucleic acid, yeast Phe tRNA, was determined in 1974 (Kim *et al.*, 1973; Robertus *et al.*, 1974). The 1980s saw the first atomic-resolution structure of a full turn of a B-DNA double helix (Drew *et al.*, 1981), icosahedral virus structures (Abad-Zapatero *et al.*, 1980; Harrison *et al.*, 1978), and an ever-increasing number of protein structures. In the late 1980s, the structure of the first protein-nucleic acid complex was determined (Anderson *et al.*, 1987), and then the first nucleosome structure was determined in the late 1990s (Luger *et al.*, 1997). The 2000s saw the determination of the first ribosome structures (Ban *et al.*, 2000; Carter *et al.*, 2000; Schluenzen *et al.*, 2000). By 1999, the PDB archive had reached 10,000 structures. By 2014, the archive housed 100,000 structures; now, there are more than 230,000 structures (Figure 2).

The size and complexity of structures deposited in the PDB reflect advances in technologies available for structure determination. In the early days, atomic-level structures were determined exclusively by macromolecular X-ray crystallography (MX). Although the steps needed to determine a structure remain the same (Figure 3a), methods for carrying out each step have evolved significantly over the years. In the 1950s, proteins were purified at a large scale from natural sources (*e.g.*, sperm whale muscle tissue) and crystallized using batch methods. Data were collected on photographic films using CuK? X-ray sources. The phases of each structure factor were determined using multiple isomorphous replacement in which additional diffraction data are measured from crystals soaked with heavy-atom labeling reagents. Resulting electron density maps were interpreted by building atomic models manually at 5 cm/1 Å scale with Kendrew's wire models inside a Richard's box (Martz and Francoeur, 2004; Richards, 1968). Until the late 1970s, the fit of atomic coordinates to electron density maps was not computationally optimized (refined); rubredoxin (a 54 amino acid protein, PDB ID 4rxn (Watenpaugh *et al.*, 1980)) was the first protein structure to be refined against experimental data (Watenpaugh *et al.*, 1972). New technologies for gene cloning

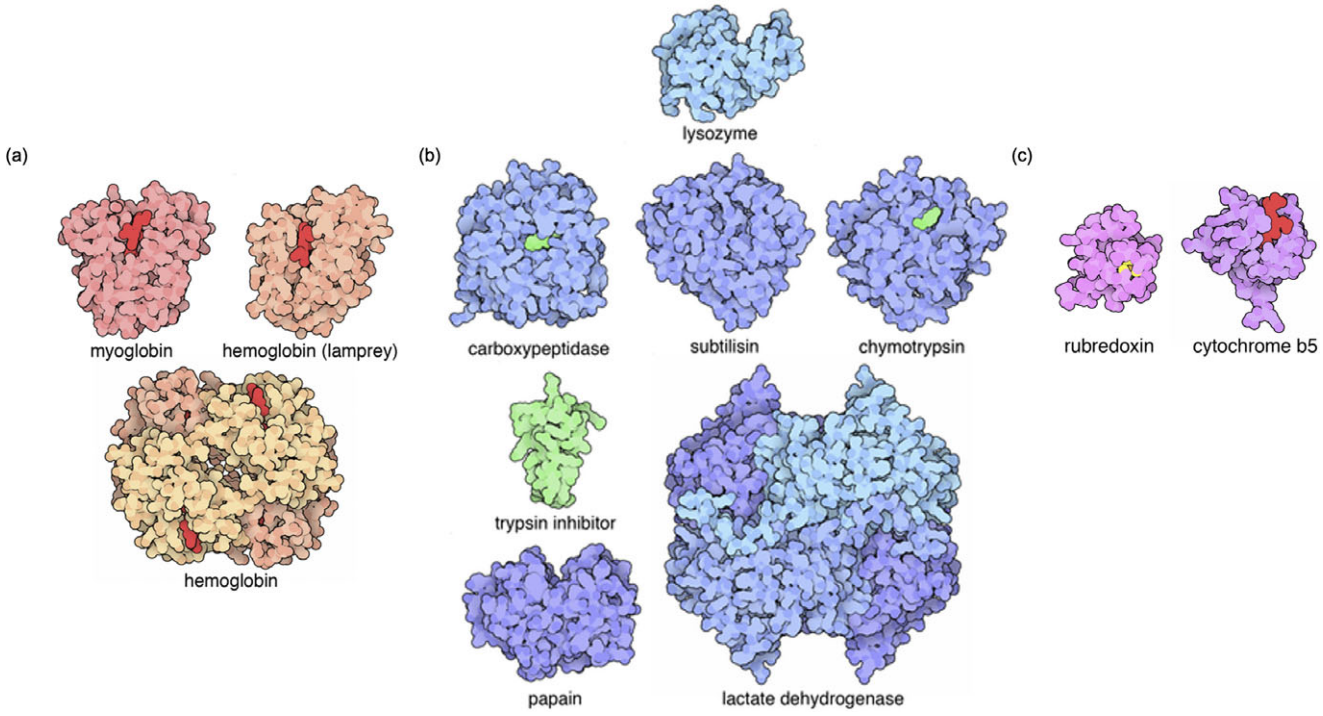

Figure 1. Early structures in the PDB: (a) Oxygen carrying; (b) enzymes. (c) Electron transport. Images from Molecule of the Month: PDB Pioneers (Goodsell 2011).

https://doi.org/10.1017/S0033583525000034 Published online by Cambridge University Press

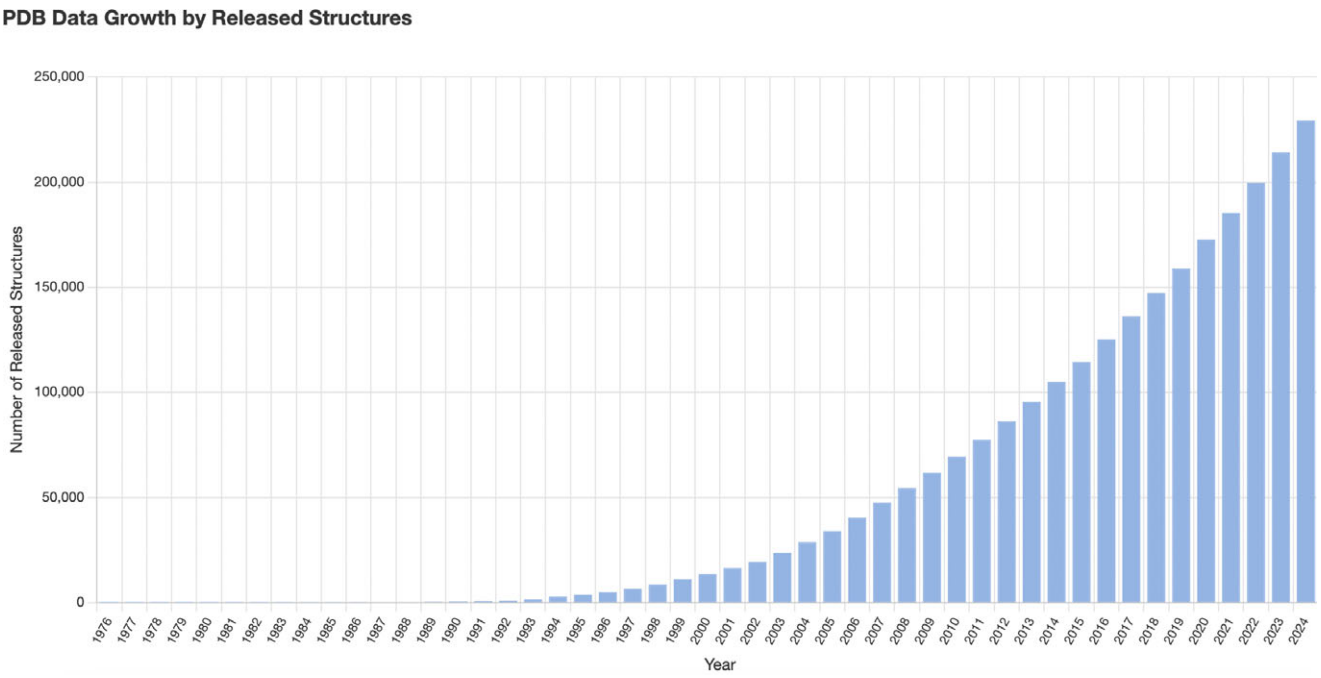

Figure 2. Overall growth of structures released in the PDB archive (https://www.rcsb.org/stats).

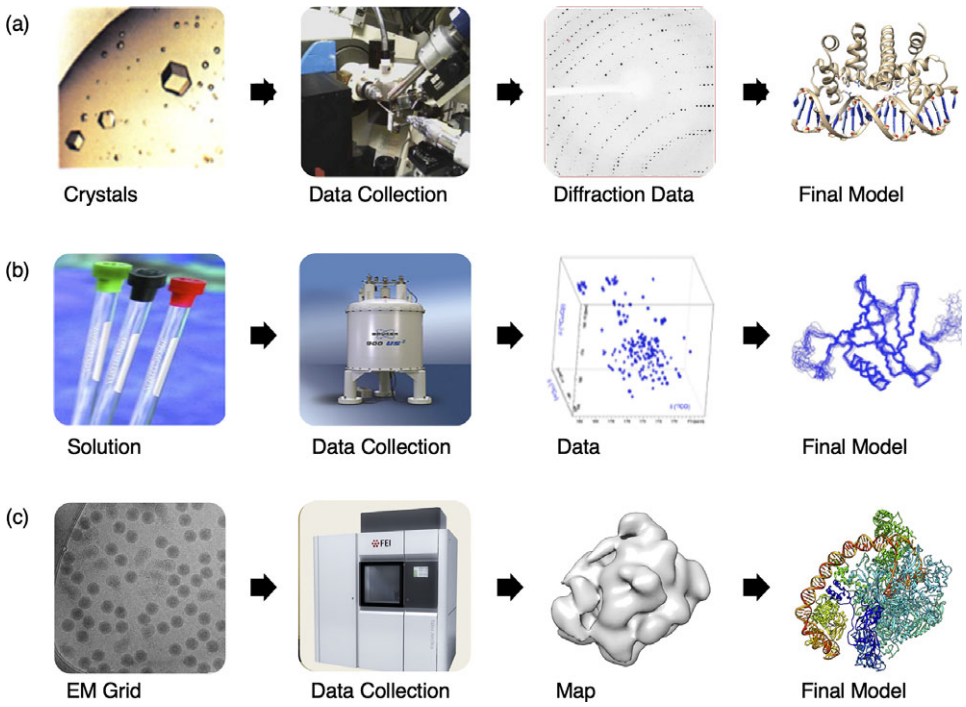

Figure 3. Structure determination pipelines for (a) MX, (b)NMR, and (c) 3DEM. Figure from https://pdb101.rcsb.org/learn/pdb-and-data-archiving-curriculum/about/ (Lawson *et al.* 2018).

and facile expression of exogenous proteins in *Escherichia coli* and so forth enabled rapid production of large quantities of proteins for structural analyses. Moreover, having control of which part of a protein to express enabled studies of individual protein domains when full-length proteins could not be crystallized. Multi-well hanging-drop/vapor diffusion crystallization plates began to be used for manual crystallization trials. Over time, crystallization reagent kits were designed, and robots did the job of setting up and screening for crystallinity (McPherson, 2017). The advent of bright synchrotron radiation sources made it possible to have more intense X-rays at tunable wavelengths, the latter supporting development of multiple-wavelength anomalous dispersion or MAD phase determination (Hendrickson, 1985). X-ray detectors have also improved dramatically in terms of efficiency and speed.

https://doi.org/10.1017/S0033583525000034 Published online by Cambridge University Press

Together, these myriad technical advances helped inspire the launch of the National Institute of General Medical Sciences Protein Structure Initiative (Norvell and Machalek, 2000) to determine the structures of all unique protein shapes, which, in addition to increasing the number and quality of PDB structures, resulted in major efficiency improvements in structure determination processes. These advances have made it possible to use extremely small samples and, for smaller proteins, produce structures in a matter of days to weeks rather than years. Today, efforts are being devoted to ever more challenging problems (e.g., integral membrane proteins, large multi-complex protein assemblies). Total archival holdings of MX structures as of January 2025 were ~191,000. Public release of new MX structures by the PDB averaged ~10,000/year for 2019–2024.

Nuclear magnetic resonance or NMR spectroscopy emerged as a structure determination method in the 1980s (Williamson *et al.*, 1985) (Figure 3b). Unlike MX, most NMR samples are dilute solutions (typically ~5 mM), which can make sample preparation easier. While relatively small protein structures are amenable to NMR structure determination methods, the technique is particularly well suited to measuring protein dynamics and exploring the behavior of intrinsically disordered proteins. Total archival holdings of NMR structures as of January 2025 were ~14,400, most of which are represented as ensembles of atomic-level structures. Public release of new NMR structures by the PDB has plateaued to a few hundred/year (311 in 2024).

The 1990s saw PDB deposition of the first electron microscopy or 3DEM structure (bacteriorhodopsin (Henderson and Schertler, 1990)) (Figure 3c). 3DEM offers three critical advantages *versus* MX: (i) crystals are not required, (ii) it is suitable for studying larger macromolecular systems, and (iii) it can be used for compositionally and conformationally heterogeneous samples. Over more than thirty years, significant advances in sample preparation and vitrification, electron optics, direct electron detection, motion correction, and cyberinfrastructure have made it possible to determine 3DEM structures at higher and higher resolution, leading to what has been termed the "Resolution Revolution (Kuhlbrandt, 2014)." As of late January 2025, 3DEM structure holdings in PDB exceeded those of NMR (24,379 *versus* 14,440), and public release of new 3DEM structures in 2024 by PDB was ~63% of new MX structures (5793 *versus* 9241). At the same time, the highest resolution 3DEM structure archived in the PDB was that of murine apo-ferritin at 1.09 Å resolution (PDB ID 8rqb (Kucukoglu *et al.*, 2024)).

In 2014, a structure of a Nup-84 sub-complex of the *Saccharomyces cerevisiae* nuclear pore complex was among the very first integrative structures to be determined by combining experimental information from multiple methods (PDB-Dev ID PDBDEV_00000001/PDB/PDB-IHM ID 8zz1 (Shi *et al.*, 2014)). Structures determined using information from various biophysical (e.g., MX, 3DEM, NMR, small-angle scattering, cross-linking mass spectrometry, Forster resonance energy transfer (FRET)) and computational (e.g., homology modeling and *de novo* structure prediction) methods are classified as integrative/hybrid methods (IHM) structures, which typically could not have been determined using a single method. Some of the early IHM structures were archived in a prototype data resource, PDB-Dev (pdb-dev.wwpdb.org (Burley *et al.*, 2017; Vallat *et al.*, 2021; Vallat *et al.*, 2018)). In late 2024, the contents of the PDB-Dev prototype resource were unified with PDB holdings and designated as PDB-IHM structures. Each of the original PDB-Dev structures now has both a PDB-Dev ID and a PDB ID. PDB-Dev has been rebranded as PDB-IHM (pdb-ihm.org (Vallat B *et al.*, in press)).

Policies

Policies for managing PDB data have evolved considerably since 1971. At the outset, deposition was purely voluntary. In the 1980s, it became clear that unless there were deposition guidelines, there was a high likelihood that valuable data would be lost. Fred Richards worked with colleagues on a petition demanding that deposition be a prerequisite for publication (Barinaga, 1989). The Biological Macromolecular Commission of the IUCr convened a committee of prominent structural biologists to establish data deposition guidelines. In 1989, these guidelines were published (International Union of Crystallography, 1989), and in time, most scientific journals began requiring the deposition of 3D biostructure to the PDB (as evidenced by the inclusion of a valid PDB ID) as a prerequisite for publication. Many funding organizations, both governmental and philanthropic, require PDB depositions by their awardees.

In 2003, the Worldwide Protein Data Bank (wwPDB) was established as a global consortium partnership to jointly manage the PDB archive (Berman *et al.*, 2003). PDB data centers in the US (RCSB Protein Data Bank or RCSB PDB), UK (Protein Data Bank in Europe or PDBe), and Japan (Protein Data Bank Japan or PDBj) were signatories to the first formal wwPDB Charter developed to ensure that all PDB data follow uniform standards and that the information remains freely available. The Charter is reviewed and renewed regularly. Current members include RCSB PDB (Berman *et al.*, 2000; Burley *et al.*, 2025), PDBe (Armstrong *et al.*, 2020), PDBj (Kinjo *et al.*, 2018), Biological Magnetic Resonance Bank (BMRB (Hoch *et al.*, 2023; Romero *et al.*, 2020; Ulrich *et al.*, 2008)), and Electron Microscopy Data Bank (EMDB (wwPDB Consortium, 2023)). Protein Data Bank China (PDBc (Xu *et al.*, 2023)) recently joined the organization as an Associate Member. Each wwPDB data center is responsible for ingesting structures deposited from within their assigned geographic catchment area (RCSB PDB: Americas and Oceania; PDBe: Europe, Africa, and Israel), PDBj (Asia and the Middle East), and PDBc (People's Republic of China). Leaders of each wwPDB partner organization meet frequently with one another and annually with the wwPDB Advisory Committee (https://www.wwpdb.org/about/advisory).

At present, wwPDB members jointly manage three Core Archives, including the Protein Data Bank, the Electron Microscopy Data Bank, and the Biological Magnetic Resonance Data Bank. Each wwPDB Core Archive is safeguarded and maintained by a wwPDB-designated Archive Keeper as follows: Protein Data Bank: RCSB PDB; Electron Microscopy Data Bank-EMDB; Biological Magnetic Resonance Data Bank-BMRB. The PDB Core Archive houses atomic coordinates of all PDB structures and related metadata and experimental data for all MX structures. The EMDB Core Archive houses electric Coulomb potential maps (hereafter 3DEM density maps) for all 3DEM structures stored in PDB and a sizeable number of additional density maps with no corresponding atomic coordinates in PDB (typically derived from lower-resolution 3DEM studies). The BMRB Core Archive houses NMR data for NMR structures stored in PDB and a considerable volume of additional biomolecule NMR data with no corresponding atomic coordinates in PDB. wwPDB members also jointly manage the NextGen PDB archive (files-nextgen.wwpdb.org), which is an enriched PDB archive that includes annotations from external database resources in the metadata that goes beyond content available in the PDB main archive (Choudhary *et al.*, 2024).

https://doi.org/10.1017/S0033583525000034 Published online by Cambridge University Press

Evolving infrastructure for ingesting, managing, and delivering PDB data

Key steps required for operating an open-access repository are data ingestion, validation, biocuration, archiving, query, and distribution (https://pdb101.rcsb.org/learn/pdb-and-data-archiving-curriculum/about/ (Lawson *et al.*, 2018)). In this section, we review how cyberinfrastructure supporting the PDB has evolved during more than 53 years of continuous PDB operations.

In 1971, data were transferred to magnetic tapes and mailed to the PDB, where they were processed on Control Data Corporation CDC 6600/7600 main-frame computers, which at the time were state-of-the-art machines. The CDC 6600 was networked *via* DARPANET (a precursor to the World Wide Web) to graphics workstations at two locations in the United States, allowing visualization of 3D biostructure data. The CRYSNET project, funded by the United States (US) National Science Foundation in 1973, was innovative in its time and provided some financial support for the PDB (Meyer *et al.*, 1974).

Initially, atomic coordinate data were stored in the Diamond format (Diamond, 1971). In 1976, the (now legacy) PDB file format, based on the 80-column Hollerith punch cards, was created and became extremely popular and widely used by the structural biology community (Bernstein *et al.*, 1977). As the size and complexity of structures grew, it became clear that the 80-column format limited the number of atoms and/or polymer chains that could be contained in a single structure data file. During the 1990s, a Working Group convened by the International Union of Crystallography (IUCr) Commission on Biological Macromolecules developed a machine- readable format that was self-defining and gave explicit relationships (Fitzgerald *et al.*, 2005). The new "macromolecular Crystallographic Information File (mmCIF)" has no limitations on the number of atoms or residues. As the PDB archive contains structures determined by severa methods, the format is now called PDBx/mmCIF (Westbrook *et al.*, 2022). It became the Master PDB archival format in 2014 (Berman *et al.*, 2014).

At present, PDBx/mmCIF is jointly maintained by the wwPDB PDBx/mmCIF Working Group (Westbrook *et al.*, 2022) and the wwPDB partners. Data dictionary terms and definitions are continuously formulated, reviewed, and modified to support existing data remediation and inclusion of new and rapidly evolving methodologies. This fully extensible data standard also supports data items and metadata elements for newer experimental methods that could not be accommodated within the restrictive legacy PDB format.

Once the World Wide Web became available in the 1990s, a computer program named AutoDep made it possible to deposit 3D biostructure data to PDB electronically (Lin *et al.*, 2000). For more than two decades, Brookhaven National Laboratory (BNL) hosted the only PDB data center that accepted depositions and processed incoming atomic coordinate data. In the late 1990s, a data center at the European Bioinformatics Institute (originally called the Macromolecular Structure Database, later rebranded as PDBe) began a collaboration with the BNL PDB to process data (Boutselakis *et al.*, 2003). In 1998, the Research Collaboratory for Structural Bioinformatics (RCSB) formed as a collaboration between Rutgers, The State University of New Jersey, the National Institute for Standards and Technology, and the San Diego Supercomputer Center successfully competed for US federal agency funding to manage the United States (US) PDB data center (Berman *et al.*, 2000). Central to the RCSB PDB project was its development of an integrated 3D biostructure data deposition system (AutoDep Input Tool, ADIT), built atop the PDBx/mmCIF data dictionary. PDBj (Kinjo *et al.*, 2018) became the first Asian PDB data center in 2000. Initially, two different data deposition systems were used by the wwPDB: AutoDep by PDBe and ADIT by RCSB PDB and PDBj. To ensure that data were fully consistent, data were exchanged regularly between sites and reviewed. A major software development project created the global OneDep system (Young *et al.*, 2017) for comprehensive deposition, rigorous validation (Feng *et al.*, 2021; Gore *et al.*, 2017; Young *et al.*, 2017), and expert biocuration (Young *et al.*, 2018) of MX, 3DEM, NMR, and micro-electron diffraction structures, supporting experimental data and related metadata. Biocuration involves checking for self-consistencies, enforcing controlled vocabularies that are part of the PDBx/mmCIF dictionary, checking polymer sequences against the sequence databases, standardization of ligand atom naming, etc., and value-added annotations (i.e., disease-causing mutations and quaternary structure information).

In the early days of the PDB, validation was focused on the geometry and chemistry of both macromolecules and bound small-molecule ligands. In addition to polypeptide backbone Ramachandran checks, MolProbity evaluation (Williams *et al.*, 2018) became the part of wwPDB validation. Although it was always possible for MX structure factor data to be deposited into PDB, it was not until 2008 that they became mandatory (wwPDB, 2024). That important policy change made it possible for OneDep to validate atomic coordinates against experimental electron density map data. NMR chemical shift deposition became mandatory in 2010 (wwPDB, 2024). For 3DEM structures, deposition of 3DEM density maps became mandatory in 2016 (wwPDB, 2024).

In 2008, the wwPDB began establishing a series of Task Forces to define validation criteria for each structure determination method supported by the PDB (Henderson *et al.*, 2012; Montelione *et al.*, 2013; Read *et al.*, 2011; Sali *et al.*, 2015; Trewhella *et al.*, 2013; wwPDB, 2024). Method-specific Task Forces, consisting of subject matter experts, evaluated procedures for rigorous assessment of structures with reference data sets and made recommendations to the wwPDB for adoption within the OneDep software system. Their efforts gave rise to a rich suite of structure validation tools that are today used to generate a wwPDB Validation Report for every incoming structure (Gore *et al.*, 2017; Gore *et al.*, 2012; Smart *et al.*, 2018a, b; Young *et al.*, 2017). These validation reports are first used by depositors, then journal editors and manuscript reviewers, and finally by PDB data consumers.

Once all the structure data and related metadata are validated and reviewed by wwPDB biocurators and depositors, they are archived as flat PDBx/mmCIF formatted files. Other research communities have followed suit, and now there are multiple working groups setting data formatting, archiving, and validation standards for various biophysical methods (Hanke *et al.*, 2024; Leitner *et al.*, 2020; Trewhella, 2018).

An important attribute of the PDBx/mmCIF data dictionary/data standard is that all of the information is stored as tables, which lend themselves to creating a relational database that can be searched efficiently. In the 1990s, the Nucleic Acid Database (NDB) became a testbed for the utility of such a database built atop the PDBx/mmCIF data standard (Berman *et al.*, 1992). NDB proved to be fit for purpose; it supported many different kinds of queries of the data and presented results in various formats.

https://doi.org/10.1017/S0033583525000034 Published online by Cambridge University Press

Building on this experience, RCSB PDB used PDB data stored in PDBx/mmCIF format to build a core database and integrated features from external databases to provide rich contextual reports for every structure in the PDB (Berman *et al.*, 2000).

From 1977 to 1992, PDB data were distributed *via* magnetic tape. Just one tape was sufficient to store a copy of the entire archive. In 1977, a total of 14 tapes, each housing 77 structures, were publicly distributed; in 1992, 262 tapes (957 structures). Thereafter, distribution of PDB data utilized CDs, followed by DVDs first by BNL and then RCSB PDB. In the late 1990s, it became possible to distribute information *via* the internet (Stampf *et al.*, 1995), and now it is the only way PDB data are distributed. In 2023, more than 3.1 billion structure data files were downloaded from the PDB main archive and web portals operated by RCSB PDB, PDBe, and PDBj combined.

PDB stakeholders

When the PDB was launched, almost all of its users were data depositors – structural biologists. Before deposition became mandatory, motivations for deposition varied, including the assurance that the data would never be lost, the desire to have someone else check the data for serious errors, or the desire to share scientific information for the public good. As archival holdings grew, protein crystallographers increasingly used previously deposited structure data to determine new structures *via* the molecular replacement approach to diffraction data phasing. PDB structures are also used to interpret lower-resolution 3DEM density maps. Computational biologists began to use the resource to classify and compare structures, thus creating a whole new field of structural bioinformatics. Drug companies began to use the PDB to facilitate structure-guided drug discovery. Educators began to use the PDB to teach biology at all levels. Computer scientists, mathematicians, and statisticians used the large PDB data set for their analyses. Today, structural biologists probably represent <1% of the very large and diverse community of PDB data consumers numbering in the many millions worldwide.

The PDB has received funding from US government agencies since its inception. Current funders of wwPDB members are as follows: RCSB PDB: US National Science Foundation, National Institutes of Health, and US Department of Energy; PDBe: European Molecular Biology Laboratory-European Bioinformatics Institute, Wellcome Trust, Biotechnology and Biological Sciences Research Council, Medical Research Council, and European Union; PDBj: Japan Science and Technology Agency Department for Information Infrastructure and Japan Agency for Medical Research and Development; EMDB: European Molecular Biology Laboratory-European Bioinformatics Institute, and Wellcome Trust; BMRB: National Institute of General Medical Sciences; PDBc: Shanghai Advanced Research Institute, Chinese Academy of Sciences, and ShanghaiTech University.

Costs and benefits of 3D biostructure data archiving

Preservation and dissemination of research results have never been free. The purchase price of Charles Darwin's "On the Origin of the Species" was nearly US$100 in today's money when first published in 1859. That volume was but a summary of the vast amount of information that Darwin assembled and analyzed before presenting his ideas on natural selection to the world. Notwithstanding the cost of the book and the enormous body of research Darwin and others undertook to make it possible, there can be no serious debate as to the value proposition of preserving the observations and ideas that went on to stimulate generations of biologists in developing our current understanding of evolution. The second half of this review explores the cost of preserving and disseminating 3D biostructure data and enumerates tangible benefits therefrom.

How much does it cost to capture, archive, and distribute PDB data?

To explain how much it costs to ingest, safeguard, and distribute PDB data, we used key performance indicators and other metrics documented during 2023 RCSB PDB operations.

We first provide a summary of the results, followed by the full details.

1. Average one-time cost to archive each new PDB structure in 2023 is ~US$420 (<1% of the estimated cost of an inexpensive determination of the MX structure of a single domain globular protein from a prokaryote).
2. Average annual cost to preserve a PDB structure in 2023 was ~US$10 (<0.01% of the estimated average replacement cost of a PDB structure).
3. Average annual cost to serve a unique IP Address Client from the RCSB PDB research-focused web portal RCSB.org in 2023 was ~30 US Cents.

During 2023, wwPDB partners based in the US, Europe, and Asia received a record 17,063 new depositions from structural biologists working on every inhabited continent. RCSB PDB processed 6,698 (~40%) of the global depositions. US federal agency grant monies budgeted for data ingestion, rigorous validation, and expert biocuration at RCSB PDB in 2023 totaled ~US$2.8 million. The 2023 one-time, non-recurring cost to ensure that these 3D biostructure data are FAIR was ~US$420/structure received/processed.

In its role as wwPDB-designated PDB Archive Keeper during the same period, RCSB PDB safeguarded the data for a total of 214,070 PDB structures (~1.4 TB of total digital storage), with an estimated replacement cost of ~US$21 billion. US federal agency grant monies budgeted for data preservation by RCSB PDB in 2023 totaled ~US$2.2 million. The annual recurring cost to preserve and safeguard the entire PDB archive was ~US$10.30/structure/year (in 2023).

Under the wwPDB charter, RCSB PDB disseminates data at no charge and with no limitations on its usage from its RCSB.org research-focused web portal. In 2023, the web portal hosted visits from ~8.2 million unique IP addresses in nearly every country and territory recognized by the United Nations. US federal agency grant monies budgeted for data dissemination at RCSB PDB in 2023 totaled ~US$2.5 million. The average recurring cost of serving each RCSB.org client in 2023 was ~30 USCents/RCSB.org unique IP Address Client. These data dissemination metrics and cost estimates do not include any users the wwPDB reaches indirectly when PDB data are reused and redistributed by nearly 500 external data resources, serving many millions more users. They also do not capture usage statistics when copies of the PDB archive are held inside biopharmaceutical and biotechnology company firewalls or stored locally for the convenience of large bioinformatics research teams in academia.

What is the value in capturing, securely archiving, and freely distributing PDB data?

We now review the enormous impact of open-access PDB data on structural biology as a discipline; the natural, chemical, engineering,

https://doi.org/10.1017/S0033583525000034 Published online by Cambridge University Press

mathematical, and physical sciences; biomedicine; biotechnology innovation; protein structure prediction; the global biodata ecosystem; and regional, US, and global economies.

Structural biology as a scientific discipline

Open access to PDB data has both accelerated the development of structural biology as a scientific discipline and enabled its reproducibility. For MX, >90% of new structures are today determined by molecular replacement using previously deposited PDB structures or Computed Structure Models (CSMs, see the Protein Structure Prediction section) to overcome the crystallographic phase problem. The development of NMR and 3DEM as mainstream structural biology methods benefited significantly from open access to the PDB, BMRB, and EMDB Core Archives. wwPDB Biocurators and OneDep structure validation tools contribute to the reproducibility of experimental methods currently supported by the PDB. Inarguably, MX is among the most reproducible experimental techniques in the biological sciences (Liebschner *et al.*, 2013).

Open-access PDB data have also enabled the analyses of ensembles of structures to understand common principles in macromolecular anatomy and biological and biochemical function. Archival contents have been used to classify and understand protein domains, and there now exist knowledge bases providing such classifications and grouping them into superfamilies (Conte *et al.*, 2000; Orengo *et al.*, 1997). Protein–protein interactions (Jones and Thornton, 1996) and protein-nucleic acid interactions have also been analyzed across the archive (Jones *et al.*, 2001). More than 1,000 papers describing these types of analyses and resources have been published to date (Basner, 2017).

Natural, chemical, computational, engineering, mathematical, physical sciences, and beyond

Knowledge of 3D structures (shapes) of biomolecules helps to explain how they function in nature, accelerating discovery across the biosciences. PDB structures include proteins and nucleic acids coming from every living kingdom (see Figure 7 in (Burley *et al.*, 2022)) and increasing numbers of designed biopolymers. Among the latter are MX structures of a designed digoxigenin binding protein (PDB IDs 4j8t, 4j9a (Tinberg *et al.*, 2013)) and an engineered organophosphate hydrolase (PDB ID 3tig (Khare *et al.*, 2012)), and a designed DNA nanomaterial (PDB ID 3gbi (Zheng *et al.*, 2009)).

PDB data impact basic and applied research on health and diseases of humans, animals, and plants; production of food and energy; and other research about global prosperity, resilience, and environmental sustainability. There are many anecdotal accounts in the scientific literature attesting to the importance of the PDB. On the occasion of the 50th birthday of the PDB in 2021, for example, the *Journal of Biological Chemistry* published two PDB50-themed special issues (Berman and Gierasch, 2021; Gierasch and Berman, 2021).

Bibliometric analyses provide opportunities for quantitative assessment of the impact of PDB data. The inaugural RCSB PDB publication (Berman *et al.*, 2000) had been cited more than 30,000 times as of January 2025, according to the Clarivate *Web of Science* (Copyright Clarivate 2024. All rights reserved). Taking a broader view, a 2018 study (Burley *et al.*, 2018) demonstrated citations of PDB data spanning the biosciences from Agriculture to Zoology. Not surprisingly, nearly 90% of published PDB structures analyzed in 2018 were cited by Biochemistry & Molecular Biology journals. High impact was also documented in other areas of fundamental biology and biomedicine (Cell Biology, Pharmacology and Pharmacy, Microbiology, Genetics & Heredity). Related analyses highlighted PDB structure publications frequently cited in STEM-related journals focused on Materials Science, Physics, Computer Science, Chemistry, Engineering, and Mathematics (Feng *et al.*, 2020). PDB data are also being used in the Social Sciences to understand human behavior and incentives in academic research (Hill and Stein, 2019) and even by artists (Voss-Andreae, 2005).

Additional unpublished bibliometric analyses provide further evidence of PDB data impact. As of January 2023, 168,902 PDB structures (~84% of the archive) were reported in 78,334 unique primary publications, which were cited 5,601,496 times. At that time, individual publications of PDB structures had been cited ~38 times on average, and each of 148,874 (~75% of the archive) PDB structures had garnered at least one citation of their corresponding primary publication. Again, as of January 2023, the "most popular" PDB structure, PDB ID 1aoi (Luger *et al.*, 1997), that of the nucleosome core particle (Figure 4), had been cited >4,500 times. Additional highly cited PDB structure statistics are as follows: 85 PDB IDs had each been cited >1,000 times; nearly 600 PDB IDs had each been cited >500 times; and >11,300 PDB IDs had each been cited >100 times). The top 10% of published PDB structures had each been cited at least 79 times. These data provide compelling evidence of the enormous breadth of PDB data impact across the scientific literature. They also document that many PDB structures are reported in "citation classic" publications.

Biomedicine

3D structures of bacterial and viral proteins archived in the PDB are routinely used to discover and develop treatments and cures for infectious diseases. As of November 2024, the archive housed nearly 76,000 structures of bacterial proteins. The two National

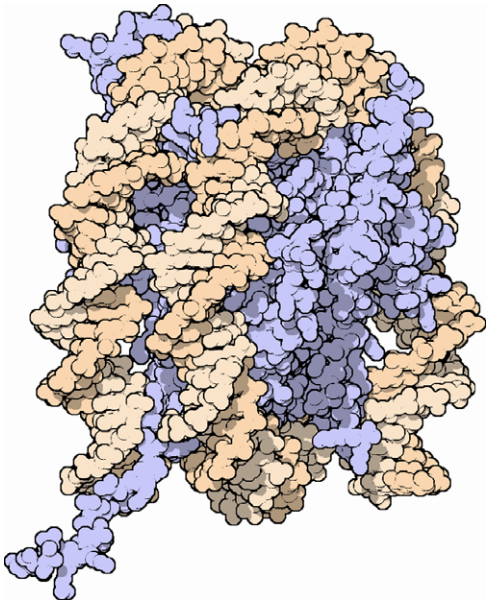

Figure 4. MX structure of the nucleosome core particle PDB ID 1aoi (Luger *et al.* 1997). Image from the Molecule of the Month (Goodsell, 2000).

https://doi.org/10.1017/S0033583525000034 Published online by Cambridge University Press

Institute for Allergy and Infectious Diseases Structural Genomics Centers for Infectious Diseases (Myler *et al.*, 2009; Stacy *et al.*, 2015) have together contributed >3,300 human pathogen protein structures to the archive. Collectively, bacterial protein structures in the PDB provide insights into microbial evolution (*e.g.*, (Koonin and Makarova, 2019)); metabolic pathways (*e.g.*, (Brunk *et al.*, 2018)); the human microbiome (*e.g.*, (Walker *et al.*, 2022)); potential targets for antimicrobial drug discovery (*e.g.*, (Shaikh *et al.*, 2021)); molecular mechanisms underpinning antibiotic resistance (*e.g.*, (Reeve *et al.*, 2015)); and structure-guided drug discovery (*e.g.*, (Simmons *et al.*, 2010)). Similarly, as of November 2024, the PDB archive housed ~21,400 structures of viral proteins. They provide valuable insights into virus evolution (*e.g.*, (Krupovic and Bamford, 2008)) and interactions with host cell proteins (*e.g.*, (Goodsell and Burley, 2020)). They also include information critical to combatting many of the viral pathogens already known to infect humans and others that may do so in the coming decades. For example, the PDB houses more than 2,600 human immunodeficiency virus-1 related structures, including more than 700 structures of the dimeric aspartyl protease (*e.g.*, PDB ID 3hpv (Wlodawer *et al.*, 1989)), many of which are co-crystal structures with bound to small-molecule inhibitors. These data played critical roles in structure-guided discovery of ten protease inhibitors approved for treating acquired immunodeficiency syndrome or AIDS, the first of which was saquinavir (PDB ID 1hxb (Krohn *et al.*, 1991)). More recently, PDB data (more than 4,600 experimentally determined SARS-CoV-2 protein structures) played central roles in the fight against COVID-19 (Figure 5, reviewed in (Burley, 2025)), contributing to both mRNA vaccine design (Corbett *et al.*, 2020) and discovery and development of nirmatrelvir, the active ingredient of Pfizer's Paxlovid (Owen *et al.*, 2021). Looking ahead to the possibility of a global pandemic caused by influenza A H5N1 virus (Kupferschmidt, 2023), there are currently >250 H5N1-related PDB structures and nearly 600 PDB structures of other influenza virus proteins (Bittrich *et al.*, 2025).

Published case studies (*e.g.*, (Hu *et al.*, 2018)) and anecdotal accounts presented at scientific meetings leave no doubt as to the important contributions to drug discovery made by structural biologists working within the biopharmaceutical industry. The first quantitative analysis of the impact of structural biologists and PDB structures on drug approvals across all therapeutic areas was published in 2019. PDB holdings were examined to identify 3D

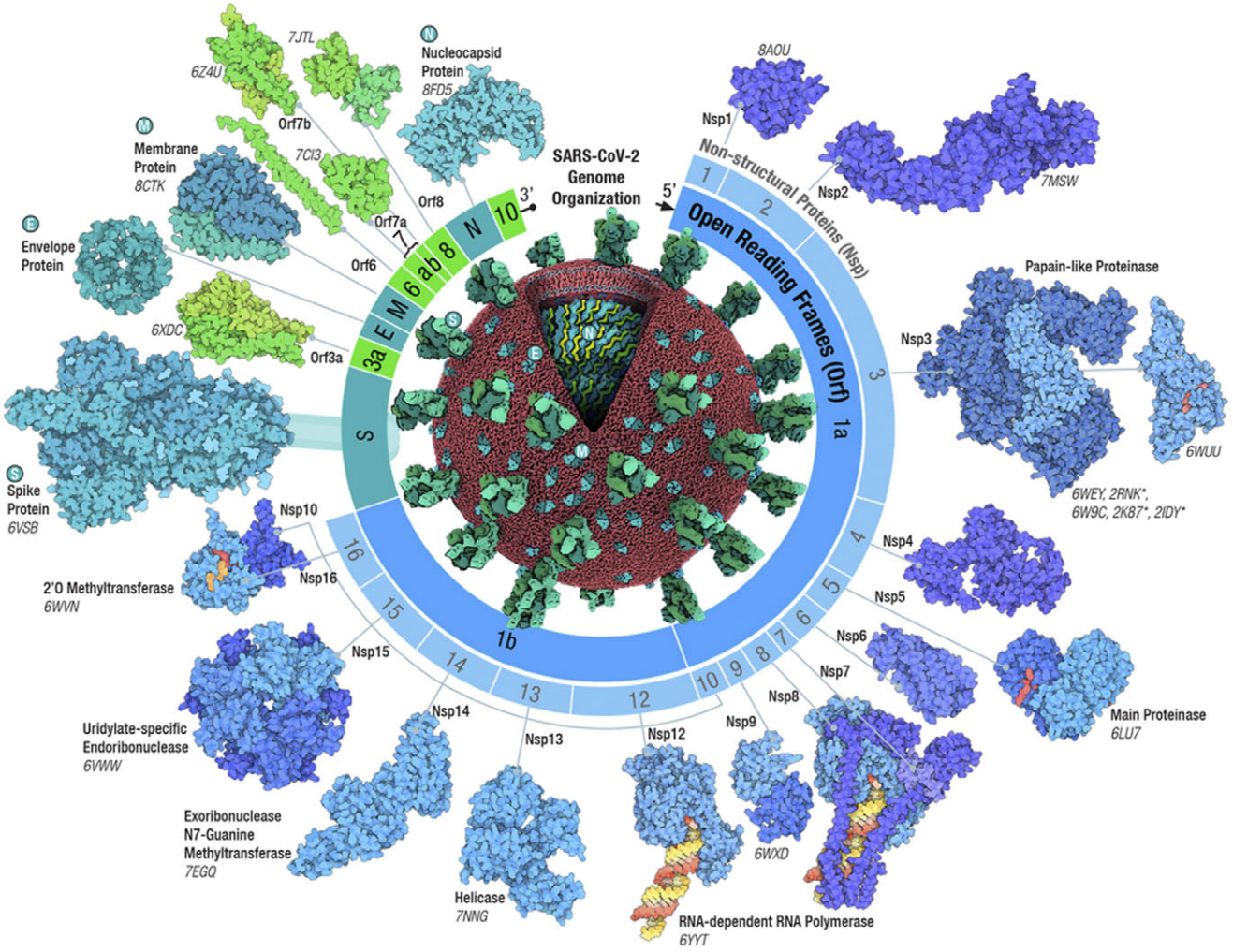

Figure 5. SARS-CoV-2 Genome and Proteome Organization. Near complete 3D knowledge of the SARS-CoV-2 proteome derives from >4,600 SARS-CoV-2 related PDB structures and CSM based on SARS-CoV-1 related structures archived in the PDB. Figure adapted from (Lubin *et al.*, 2022) and available from PDB-101 (https://pdb101.rcsb.org/learn/flyers-posters-and-calendars/flyer/sars-cov-2-genome-and-proteins). Color coding: shades of blue-non-structural proteins; shades of green: structural proteins and proteins encoded by various open-reading frames; yellow/orange/red-duplex RNA; orange-S-adenosyl methionine; and shades of red-enzyme inhibitors.

https://doi.org/10.1017/S0033583525000034 Published online by Cambridge University Press

biostructures relevant to the discovery and development of 171 new small-molecule drugs across all therapeutic areas approved by the US Food and Drug Administration (FDA) from 2010 to 2018 (Westbrook and Burley, 2019). The PDB archive housed 5,364 structures, providing atomic-level, 3D information for ~88% of the targets of these 171 small-molecule drugs. Structure-guided drug discovery approaches were used to generate >70% of these new drugs. In approximately 20% of cases, the number of PDB structures of the drug target exceeded 100.

Two follow-up studies focused on new anti-cancer drugs. One study documented that access to PDB structure information facilitated discovery and development of >90% of the 79 new anti-neoplastic agents approved by the US FDA from 2010 to 2018 (54 small-molecule drugs and 25 biologics) (Westbrook *et al.*, 2020). The other (Burley *et al.*, 2024) went on to review small-molecule anti-cancer drugs approved by US FDA from 2019 to 2023. During this latter period, open access to PDB structure information facilitated discovery and development of 100% of 34 newly approved anti-neoplastic agents. Approximately 80% of these new drugs were the products of structure-guided drug discovery. Figure 6, for example, illustrates PDB ID 6o8m (Canon *et al.*, 2019), showing the mechanism of action at the atomic level in 3D of sotorasib covalent targeting the G12C mutant form of KRAS (Lanman *et al.*, 2020). Before discovery and development of sotorasib, RAS oncoproteins were deemed undruggable. Structure-guided discovery, development, and regulatory approval of this first-in-class drug set the stage for targeting other mutant forms of RAS, which collectively occur in ~20% of all human cancers (Prior *et al.*, 2020).

Analyses of PDB holdings, the scientific literature, and related documents for each anti-cancer drug-protein target combination revealed that the impact of public-domain 3D structure data is broad and substantial, ranging from understanding target protein biology to identifying a given target protein as druggable to structure-guided drug discovery. There is every reason to believe that PDB structures of target proteins will continue to facilitate structure-guided discovery and subsequent development of new drugs benefiting patients and their families and society more broadly for decades to come.

Biotechnology innovation

Patent literature reviews conducted in August 2022 documented very broad impact of PDB data on innovation. As of June 2022, searching the US Patent and Trademark Office website (United States Patent and Trademark Office, 2022) identified ~10,000 issued US patents with PDB mentions (*vs.* ~6,500 issued patents in June 2017 (Burley *et al.*, 2018)). Analogous searches of global patent literature using PatSeer (Gridlogics, 2021) documented ~90,000 issued patents and in-process patent applications worldwide that include PDB mentions (*vs.* ~50,000 in mid-2017 (Burley *et al.*, 2018)). It should be noted that patents and patent applications mentioning PDB data do not involve attempts to patent protein structures *per se* (Committee on Intellectual Property Rights in Genomic And Protein Research and Innovation, 2006).

The top ten assignees of worldwide patents mentioning PDB in mid-2017 included four US research universities (Massachusetts Institute of Technology; New York University; University of California Regents; University of Texas), two biopharmaceutical companies (Genentech, Inc.; Amgen, Inc.), two biotechnology companies (Xencor, Inc.; Novozymes, Inc.), and two agribusiness companies (DuPont de Nemours, Inc.; Pioneer). These findings underscore the importance of open access to PDB data for basic and applied research carried out in universities and not-for-profit

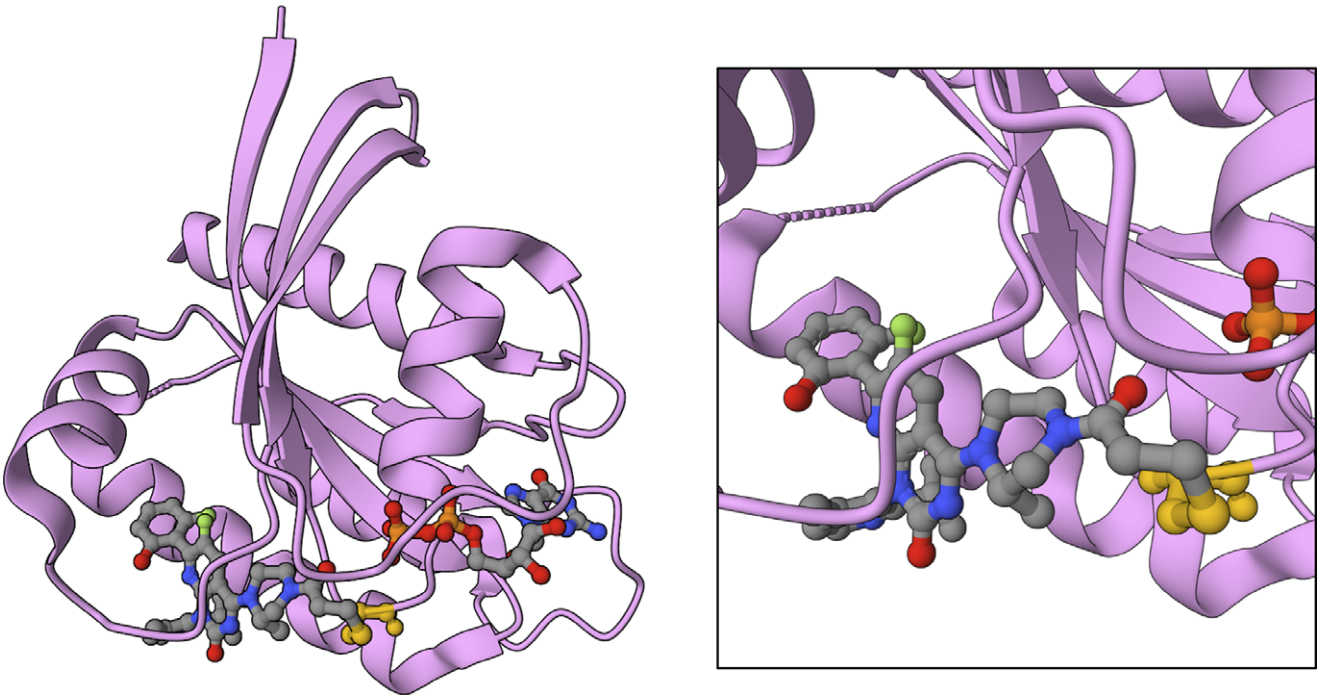

Figure 6. Ribbon representation of the co-crystal structure of sotorasib covalently bound to the G12C KRAS (pink)/GDP complex (PDB ID 6oim (Canon *et al.*, 2019)). Inset highlights a zoomed-in view of the sotorasib binding site, showing the covalent bond (half green/half yellow) between the drug and Cysteine 12 (yellow atomic ball-and-stick figure). Images generated using the Mol* Viewer (Sehnal *et al.*, 2021). Image adapted from (Burley *et al.*, 2024).

https://doi.org/10.1017/S0033583525000034 Published online by Cambridge University Press

institutes, and for-profit biopharmaceutical, biotechnology, and agribusiness companies.

Protein structure prediction

PDB data facilitated the development of structural bioinformatics as a vibrant subdiscipline of computational biology (Bourne and Weissig, 2003). Without an open-access repository of rigorously validated, expertly biocurated 3D structures of biological macromolecules, there would be no homology modeling, no computational docking of small-molecule ligands, and no *de novo* protein structure prediction. Inspired by the work of Anfinsen, who showed in the 1970s that the sequence of a polypeptide chain determines its shape or fold (Anfinsen, 1973); practitioners of this emerging field strove for decades to predict the 3D structures of proteins accurately. The 2020 Critical Assessment of Structure Prediction exercise (CASP (Alexander *et al.*, 2021)) witnessed a sea change in structural bioinformatics. Google DeepMind's AlphaFold2 (AF2) Artificial Intelligence/Machine Learning (AI/ML) software emerged as the top performer for *de novo* protein structure prediction, with accuracies often comparable to that of lower-resolution experimental methods (Jumper *et al.*, 2021; Shao *et al.*, 2022; Terwilliger *et al.*, 2024). Subsequently, the Rosetta team released RoseTTAFold (Baek *et al.*, 2021), which generates CSMs of proteins with accuracies comparable to AF2.

Today, Computed Structure Models for nearly every protein sequence represented in UniProt (UniProt Consortium, 2023) are publicly accessible from the AlphaFold Protein Structure Database (AlphaFold DB (Varadi *et al.*, 2022)). Some of the millions of CSMs generated independently of DeepMind (using RoseTTAFold, AF2 Colab, Meta, *etc.*) are available from the open-access ModelArchive (Protein Structure Bioinformatics Group, 2024). Both AlphaFold DB and the ModelArchive utilize the ModelCIF data standard (Vallat *et al.*, 2023), which interoperates seamlessly with the PDBx/mmCIF data dictionary described above. It is jointly managed by wwPDB partners and the wwPDB ModelCIF Working Group (www.wwpdb.org/task/modelcif). A 2021 *New England Journal of Medicine* publication described potential uses of CSMs in clinical research and practice (Burley *et al.*, 2021). Development of AF2 by John Jumper and Demis Hassabis and the pioneering protein design work of David Baker earned them shares of the 2024 Nobel Prize in Chemistry. All three newly minted Nobel Laureates explicitly acknowledged the key role that the Protein Data Bank played in providing highly curated, validated, machine-readable data (Callaway, 2024).

The RCSB PDB provides open access to more than one million CSMs alongside all PDB structures at RCSB.org (Figure 7). Access to both CSMs and PDB data benefits structural biologists who are using CSMs when initiating experimental studies (*e.g.*, for expression construct design) and during MX (Terwilliger *et al.*, 2022) and 3DEM (Subramaniam and Kleywegt, 2022) structure determination efforts. Making CSMs available to PDB data consumers working in areas such as plant sciences makes RCSB.org a much more valuable resource. The current experimental structure coverage of the *Arabidopsis thaliana* proteome in the PDB is ~4%. With combined delivery of PDB structures and CSMs at RCSB.org, plant molecular biologist users enjoy access to 3D structure information spanning the entire *A. thaliana* proteome.

Delivery of more than one million CSMs alongside PDB structures also provides full proteome coverage of 3D structural information for human, the major model organisms (mouse, rat, zebrafish, fruit fly, Dictyostelium, *Caenorhabditis elegans*, *A. thaliana*, *S. cerevisiae*, *Schizosaccharomyces pombe*, *C. albicans*, *E. coli*, and *Methanocaldococcus jannaschii*), 32 human pathogens, three important food crop plants (rice, maize, and soybean), and select organisms important for understanding the impact of climate change. Providing simultaneous access to experimentally determined structures and CSMs allows both types of information to be searched, visualized, and analyzed together. It also informs bioscience researchers and their trainees, and educators and their students as to some of the limitations of CSMs. They are comparable in accuracy to lower-resolution experimental structures and should not be relied on when a gold-standard, experimentally determined PDB structure(s) is available (Moore *et al.*, 2022; Shao *et al.*, 2022).

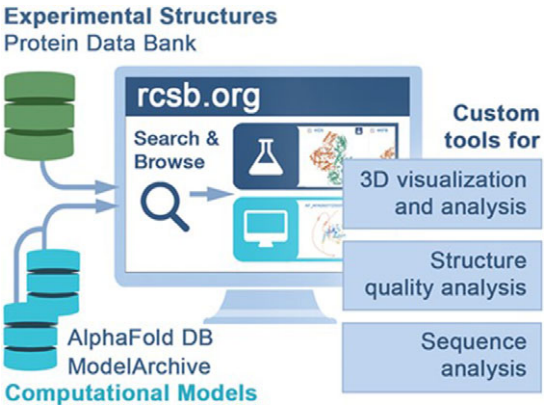

Figure 7. RCSB.org delivers PDB experimental structures (identified with an Erlenmeyer flask icon in dark blue) and CSMs (computer screen icon in cyan) from AI/ML that can be searched, analyzed, visualized, and explored using custom tools and features. Image from (Burley *et al.*, 2023).

Information stored in the PDB is made available under the most permissive Creative Commons CC0 1.0 Universal License (*https://creativecommons.org/licenses/by/4.0/*), enabling researchers around the world to access and utilize the information at no charge and with no restrictions on its usage. Recognizing its long-standing commitment to high standards of data preservation, management, and open access, the PDB is accredited by CoreTrustSeal, an international organization that certifies data repositories (*https://amt.coretrustseal.org/certificates/*). More recently, the PDB was recognized by the Global Biodata Coalition (*https://globalbiodata.org*) as a Global Core Biodata Resource of "fundamental importance to the wider biological and life sciences community and the long-term preservation of biological data." PDB remains a vanguard in the open-access movement.

Worldwide distribution of PDB data is not limited to wwPDB partner web portals. Review of the *Nucleic Acids Research* Online Molecular Biology Database Collection (Rigden and Fernandez, 2023), which comprises databases from the journal's annual Database Issues, identified ~500 external data resources that distribute repackaged PDB data to individuals who may not routinely visit RCSB.org or one of the wwPDB partner web portals (Resources as of 2022 (Rigden and Fernandez, 2022) listed in able S1 of (Burley *et al.*, 2022)). Beyond utilization of PDB from open-access knowledgebases, *etc.*, there is substantial reuse of public domain 3D biostructures within global biopharmaceutical companies (*e.g.*, Pfizer, Novartis, Eli Lilly, and Company), most if not all of which maintain copies of the archive inside company firewalls. Therein, PDB data are used daily alongside proprietary MX, NMR, and

https://doi.org/10.1017/S0033583525000034 Published online by Cambridge University Press

3DEM structures determined by the company to better understand target protein biology, identify target proteins as likely to be druggable, and support structure-guided drug discovery and preclinical development of drug candidates.

Regional, US, and global economies

Although it has not been possible to carry out comprehensive analyses of the economic impact of wwPDB Core Archives and wwPDB partner activities, some data about RCSB PDB operations are available. A 2017 Rutgers University Office of Research Analytics (ORA) study documented the substantial contributions of PDB data and RCSB PDB to public sector economies (Sullivan *et al.*, 2017). The corpus of scientific data (>227,000 3D biostructures) has an estimated replacement cost of nearly US$23 billion. The Rutgers ORA analyses of 2017 public sector usage of PDB data delivered *via* RCSB.org estimated an aggregate economic value of ~US$9.2 billion (>1,500 times ~US$6.1 million federal funding of RCSB PDB at that time). Since 2017, the PDB archive has grown by ~67%, and the number of unique IP clients visiting RCSB.org annually has grown by >80%, suggesting that the public sector economic impacts of PDB data and RCSB PDB operations have increased substantially (as a multiple of ~US$10 million in 2024 federal funding of RCSB PDB Core Operations).

The Rutgers ORA analysis did not attempt to estimate quantitatively the economic impact of PDB data accruing from societal benefits generated by pharmaceutical and biotechnology companies. However, some sense of the magnitude of impact on for-profit companies and the global economy can be gleaned from the metrics presented above under Biotechnology Innovation and Biomedicine. We are also unable to quantify the impact of open access to PDB data on education and STEM workforce training. Introductory RCSB PDB training materials and documentation delivered at PDB101.RCSB.org help researchers and their trainees, and educators and their students learn how to connect 3D biostructures to knowledge. PDBe and PDBj also provide training resources to users of their web portals (pdbe.org and pdbj.org, respectively), as do our other two wwPDB partners EMDB (ebi.ac.uk/emdb/) and BMRB (bmrb.io).

Perspectives and future directions

The advent of AF2, RoseTTAfold, *etc.*, caused some scientists to suggest that structural biology as a discipline and those who determine structures using experimental methods would no longer be necessary. They failed to recognize that structural biologists have never shied away from embracing new biophysical and computational methods to achieve their ultimate goals of visualizing and understanding biomolecules in 3D at the atomic level. They also failed to appreciate how useful the results of *de novo* structure prediction would be for structural scientists, particularly for those research teams relying on 3DEM methods. Individual CSMs can be fitted into 3DEM density maps coming from both single-particle and tomographic 3DEM measurements. At present, there are no computational methods capable of delivering predicted structures of large macromolecular complexes with accuracies comparable to lower-resolution experimental methods. Even for individual proteins, experimentally determined structures are more accurate than CSMs. Moreover, they are often more informative because they provide atomic-level insights into binding of small-molecule ligands (*e.g.*, enzyme co-factors, inhibitors, US FDA-approved drugs).

It is important to note, however, that the promise of new AI/ML tools for structural biology depends critically on open access to ever more experimental structures in the PDB; well-determined, rigorously validated, and expertly biocurated structures are essential for improving the quality of the training sets on which AI methods development depends. Thus, the importance of continued focus on validation by the wwPDB has never been greater. In addition, biochemical analyses performed by PDB structure depositors are essential for us to understand complex relationships between structure and function. The central importance of PDB data to the development of *de novo* protein structure prediction tools reliant on AI/ML approaches raises important questions regarding the "dark matter" of structural biology – the hundred thousand or more X-ray co-crystal structures of protein-ligand complexes preserved inside biopharmaceutical company firewalls as trade secrets. Successful development of AI/ML tools that support structure-guided drug discovery could well hinge on public access to some of this information, much of which is post- competitive (meaning that its release will in no way diminish company shareholder value). The lesson from AF2, RoseTTAfold, *etc.*, is clear. Open access to tens of thousands of entirely new co-crystal structures "donated" to the PDB by biopharmaceutical companies will accelerate structure-guided drug discovery for the benefit of patients, their families, and all of humanity. Finally, the ambitious goals of providing structural descriptions of organelles and even whole cells (*e.g.*, the pancreatic beta cell (Singla *et al.*, 2018)) can and will be realized if structural biologists continue to develop new structure determination methods such as integrative and hybrid methods, and PDB, EMDB, and BMRB continue to ingest, validate, biocurate, and archive reliable atomic-level 3D structure information for proteins and nucleic acids, and their complexes with one another and small-molecule ligands.

The overarching mission of the wwPDB is to make highly curated and therefore trustworthy atomic-level 3D macromolecular structure information freely available to anyone working and learning anywhere in the world, with no limitations on data usage. The wwPDB was founded by three Core Members representing three continents to ensure the success of the PDB Core Archive; today, it has five Core Members and an Associate Member that jointly manage three Core Archives. Various Task Forces and Working groups have developed rigorous structure validation criteria implemented by the wwPDB. Active involvement of subject matter experts from the global scientific community ensures that wwPDB data will remain well-curated and reliable. On a weekly basis, each wwPDB Core Archive is updated by its respective Archive Keeper and released to the public. Thereafter, web portals maintained independently by RCSB PDB, PDBe, PDBj, EMDB, and BMRB distribute identical 3D biostructure information, together with unique services and value-added information. With PDB, EMDB, and BMRB serving as singular wwPDB Core Archive data resources, fragmentation and balkanization of the world's 3D biostructure data have been avoided. The wwPDB is a highly effective consortium, one that is laterally aligned. That is, it allows member organizations to be independent as appropriate while collaborating to achieve common goals (The Stakeholder Alignment Collaborative, 2025). We believe that the model provided by the wwPDB for international collaboration to preserve and disseminate high-quality information can be adopted by other scientific disciplines, thereby enabling exciting new technical and scientific breakthroughs, particularly those using data-reliant AI/ML approaches.

Acknowledgments. The authors thank the tens of thousands of structural biologists working on all inhabited continents who have deposited structures to

https://doi.org/10.1017/S0033583525000034 Published online by Cambridge University Press

the PDB since 1971 and the many millions of researchers, educators, and students around the world who consume PDB data. We thank the members of the RCSB PDB and wwPDB Advisory Committees for their valued advice. We also gratefully acknowledge contributions to the success of the PDB archive made by past members of RCSB PDB and our Worldwide Protein Data Bank partners (PDBe, PDBj, PDBc, EMDB, and BMRB) and thank Dr. Brinda Vallat and Christine Zardecki for help with manuscript preparation.

Financial support. RCSB PDB Core Operations are jointly funded by the U.S. National Science Foundation [DBI- 2321666, PI: S.K. Burley], the U.S. Department of Energy [DE-SC0019749, PI: S.K. Burley], and the National Cancer Institute, the National Institute of Allergy and Infectious Diseases, and the National Institute of General Medical Sciences of the National Institutes of Health [R01GM157729, PI: S.K. Burley]. The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Competing interest. H.M. Berman and S.K. Burley declare none.

References

Abad-Zapatero C *et al* (1980) Structure of southern bean mosaic virus at 2.8 A resolution. *Nature* **286**(5768), 33–39. https://doi.org/10.1038/286033a0.

Alexander LT *et al* (2021) Target highlights in CASP14: Analysis of models by structure providers. *Proteins* **89**(12), 1647–1672. https://doi.org/10.1002/prot.26247.

Anderson JE, Ptashne M and Harrison SC (1987) Structure of the repressor-operator complex at bacteriophage 434. *Nature* **326**, 846–852. https://doi.org/10.1038/326846a0

Anfinsen CR (1973) Principles that govern the folding of protein chains. *Science* **181**, 223–230. https://doi.org/10.1126/science.181.4096.223.

Armstrong DR *et al* (2020) PDBe: Improved findability of macromolecular structure data in the PDB. *Nucleic Acids Research* **48**(D1), D335–D343. https://doi.org/10.1093/nar/gkz990.

Baek M *et al* (2021) Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**(6557), 871–876. https://doi.org/10.1126/science.abj8754.

Ban N *et al* (2000) The complete atomic structure of the large ribosomal subunit at a 2.4 Å resolution. *Science* **289**, 905–920. https://doi.org/10.1126/science.289.5481.905.

Barinaga M (1989) The missing crystallography data. *Science* **245**(4923), 1179–1181. https://doi.org/10.1126/science.2781276.

Basner J (2017) Impact Analysis of "Berman HM et al., (2000), The Protein Data Bank. doi: 10.2210/rcsb_pdb/pdb-econ-imp-2017.

Berman HM (2008) The Protein Data Bank: A historical perspective. *Acta Crystallographica Section A* **64**(1), 88–95. https://doi.org/10.1107/S0108767307035623.

Berman HM and Gierasch LM (2021) How the Protein Data Bank changed biology: An introduction to the JBC Reviews thematic series, part 1. *Journal of Biological Chemistry* **296**, 100608. https://doi.org/10.1016/j.jbc.2021.100608.

Berman HM, Henrick K and Nakamura H (2003) Announcing the worldwide Protein Data Bank. *Nature Structural Biology* **10**(12), 980. https://doi.org/10.1038/nsb1203-980.

Berman HM *et al* (2014) The Protein Data Bank archive as an open data resource. *Journal of Computer-Aided Molecular Design* **28**, 1009–1014. https://doi.org/10.1007/s10822-014-9770-y.

Berman HM *et al* (1992) The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophysical Journal* **63**(3), 751–759. https://doi.org/10.1016/S0006-3495(92)81649-1.

Berman HM *et al* (2000) The protein data bank. *Nucleic Acids Research* **28**(1), 235–242. https://doi.org/10.1093/nar/28.1.235.

Bernstein FC *et al* (1977) The Protein Data Bank. A computer-based archival file for macromolecular structures. *European Journal of Biochemistry* **80**(2), 319–324. https://doi.org/10.1016/s0022-2836(77)80200-3

Bittrich S *et al* (2025) Visualizing and Analyzing 3D Biomolecular Structures using Mol* at RCSB.org: Influenza A H5N1 virus proteome case study. *Protein Science*. https://doi.org/10.1002/pro.70093

Blake CCF *et al* (1965) Structure of hen egg-white lysozyme. A three dimensional Fourier synthesis at 2 Å resolution. *Nature* **206**, 757–761. https://doi.org/10.1038/206757a0.

Bolton W and Perutz MF (1970) Three dimensional fourier synthesis of horse deoxyhaemoglobin at 2.8 Ångstrom units resolution. *Nature* **228**(271), 551–552. https://doi.org/10.1038/228551a0.

Bourne PE and Weissig H (eds) (2003) *Structural Bioinformatics*. Hoboken, NJ: John Wiley & Sons, Inc.

Boutselakis H, (2003) E-MSD: The European Bioinformatics Institute Macromolecular Structure Database. *Nucleic Acids Research* **31**(1), 458–462. https://doi.org/10.1093/nar/gkg065.

Brunk E *et al* (2018) Recon3D enables a three-dimensional view of gene variation in human metabolism. *Nature Biotechnology* **36**(3), 272–281. https://doi.org/10.1038/nbt.4072.

Burley SK (2025) Protein Data Bank: From two epidemics to the global pandemic to mRNA vaccines and Paxlovid. *Current Opinion in Structural Biology* **90**, 102954. https://doi.org/10.1016/j.sbi.2024.102954.

Burley SK, Arap W and Pasqualini R (2021) Predicting proteome-scale protein structure with artificial intelligence. *New England Journal of Medicine* **385**(23), 2191–2194. https://doi.org/10.1056/NEJMcibr2113027.

Burley SK *et al* (2018) RCSB Protein Data Bank: Sustaining a living digital data resource that enables breakthroughs in scientific research and biomedical education. *Protein Science* **27**(1), 316–330. https://doi.org/10.1002/pro.3331.

Burley SK *et al* (2022) Protein Data Bank: A comprehensive review of 3D structure holdings and worldwide utilization by researchers, educators, and students. *Biomolecules* **12**, 1425. https://doi.org/10.3390/biom12101425.

Burley SK *et al* (2025) Updated Resources for exploring experimental PDB structures and computed structure models at the RCSB protein data bank. *Nucleic Acids Research* **53**, D564–D574. https://doi.org/10.1093/nar/gkae1091.

Burley SK *et al* (2023) RCSB Protein Data Bank (RCSB.org): Delivery of experimentally-determined PDB structures alongside one million Computed Structure Models of proteins from Artificial Intelligence/Machine Learning. *Nucleic Acids Research* **51**, D488–D508. https://doi.org/10.1093/nar/gkac1077.

Burley SK *et al* (2017) PDB-Dev: A Prototype System for Depositing Integrative/Hybrid Structural Models. *Structure* **25**(9), 1317–1318. https://doi.org/10.1016/j.str.2017.08.001.

Burley SK *et al* (2024) Impact of structural biology and the Protein Data Bank on US FDA new drug approvals of low molecular weight antineoplastic agents 2019-2023. *Oncogene* **43**(29), 2229–2243. https://doi.org/10.1038/s41388-024-03077-2.

Callaway E (2024) The huge protein database that spawned AlphaFold and biology's AI revolution. *Nature* **634**(8036), 1028–1029. https://doi.org/10.1038/d41586-024-03423-0.

Canon J *et al* (2019) The clinical KRAS(G12C) inhibitor AMG 510 drives anti-tumour immunity. *Nature* **575**(7781), 217–223. https://doi.org/10.1038/s41586-019-1694-1.

Carter AP *et al* (2000) Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. *Nature* **407**, 340–348. https://doi.org/10.1038/35030019.

Choudhary P *et al* (2024) PDB NextGen Archive: centralizing access to integrated annotations and enriched structural information by the Worldwide Protein Data Bank. *Database (Oxford)* **2024**, baae041. https://doi.org/10.1093/database/baae041.

Committee on Intellectual Property Rights in Genomic and Protein Research and Innovation (2006) In Merrill SA and Mazza AM (eds), *Reaping the Benefits of Genomic and Proteomic Research: Intellectual Property Rights, Innovation, and Public Health*. Washington (DC).

Conte L *et al* (2000) SCOP: A structural classification of proteins database. *Nucleic Acids Research* **28**(1), 257–259. https://doi.org/10.1093/nar/28.1.257.

Corbett KS *et al* (2020) SARS-CoV-2 mRNA vaccine design enabled by prototype pathogen preparedness. *Nature* **586**(7830), 567–571. https://doi.org/10.1038/s41586-020-2622-0.

Diamond R (1971) A real-space refinement procedure for proteins. *Acta Crystallographica Section* **A27**, 436–452. https://doi.org/10.1107/S0567739471000986.

Drew HR *et al* (1981) Structure of a B-DNA dodecamer: Conformation and dynamics. *Proceedings of the National Academy of Sciences U.S.A.* **78**(4), 2179–2183. https://doi.org/10.1073/pnas.78.4.2179.

Feng Z *et al* (2020) Impact of the protein data bank across scientific disciplines. *Data Science Journal* **19**, 1–14. https://doi.org/10.5334/dsj-2020-025.

Feng Z *et al* (2021) Enhanced validation of small-molecule ligands and carbohydrates in the protein databank. *Structure* **29**, 393–400.e391. https://doi.org/10.1016/j.str.2021.02.004.

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.

Gierasch LM and Berman HM (2021) How the Protein Data Bank changed biology: An introduction to the JBC Reviews thematic series, part 2. *Journal of Biological Chemistry* **296**, 100748. https://doi.org/10.1016/j.jbc.2021.100748.

Goodsell DS (2000) Nucleosome. *RCSB PDB Molecule of the Month*. https://doi.org/10.2210/rcsb_pdb/mom_2000_7.

Goodsell DS (2011) PDB Pioneers. *RCSB PDB Molecule of the Month*. https://doi.org/10.2210/rcsb_pdb/mom_2011_10.

Goodsell DS and Burley SK (2020) RCSB Protein Data Bank tools for 3D structure-guided cancer research: Human papillomavirus (HPV) case study. *Oncogene* **39**(43), 6623–6632. https://doi.org/10.1038/s41388-020-01461-2.

Gore S *et al* (2017) Validation of structures in the Protein Data Bank. *Structure* **25**(12), 1916–1927. https://doi.org/10.1016/j.str.2017.10.009.

Gore S, Velankar S and Kleywegt GJ (2012) Implementing an X-ray validation pipeline for the Protein Data Bank. *Acta Crystallographica D* **68**(Pt 4), 478–483. https://doi.org/10.1107/S0907444911050359.

Gridlogics (2021) PatSeer™. Available at patseer.com (accessed 2021).

Hanke CA *et al* (2024) Making fluorescence-based integrative structures and associated kinetic information accessible. *Nature Methods*. https://doi.org/10.1038/s41592-024-02428-x.

Harrison SC *et al* (1978) Tomato bushy stunt virus at 2.9 A resolution. *Nature* **276**(5686), 368–373. https://doi.org/10.1038/276368a0.

Henderson R *et al* (2012) Outcome of the first electron microscopy validation task force meeting. *Structure* **20**(2), 205–214. https://doi.org/10.1016/j.str.2011.12.014.

Henderson R and Schertler GF (1990) The structure of bacteriorhodopsin and its relevance to the visual opsins and other seven-helix G-protein coupled receptors. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences* **326**(1236), 379–389. https://doi.org/10.1098/rstb.1990.0019.

Hendrickson WA (1985) Analysis of protein structure from diffraction measurement at multiple wavelengths. *Transactions of the American Crystallographic Association* **21**, 11–21.

Hill R and Stein C (2019) Scooped! estimating rewards for priority in science. Available at https://carolynstein.github.io/files/scooped.pdf (accessed 2019).

Hoch JC *et al* (2023) Biological magnetic resonance data bank. *Nucleic Acids Research* **51**, D368–D376. https://doi.org/10.1093/nar/gkac1050.

Hu T *et al* (2018) The impact of structural biology in medicine illustrated with four case studies. *Journal of Molecular Medicine* **96**(1), 9–19. https://doi.org/10.1007/s00109-017-1565-x.

International Union of Crystallography (1989) Policy on publication and the deposition of data from crystallographic studies of biological macromolecules. *Acta Crystallographica* **A45**, 658. https://doi.org/10.1107/S0108767389007695.

Jones S *et al* (2001) Protein-RNA interactions: a structural analysis. *Nucleic Acids Research* **29**(4), 943–954. https://doi.org/10.1093/nar/29.4.943.

Jones S and Thornton JM (1996) Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences U.S.A.* **93**, 13–20. https://doi.org/10.1073/pnas.93.1.13.

Jumper J *et al* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* **596**(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.

Kartha G, Bello J and Harker D (1967) Tertiary structure of ribonuclease. *Nature* **213**, 862–865. https://doi.org/10.1038/213862a0.

Kendrew JC *et al* (1958) A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* **181**, 662–666. https://doi.org/10.1038/181662a0.

Kendrew JC *et al* (1960) Structure of myoglobin: A three-dimensional Fourier synthesis at 2 A. resolution. *Nature* **185**(4711), 422–427. https://doi.org/10.1038/185422a0.

Khare SD *et al* (2012) Computational redesign of a mononuclear zinc metalloenzyme for organophosphate hydrolysis. *Nature Chemical Biology* **8**(3), 294–300. https://doi.org/10.1038/nchembio.777.

Kim SH *et al* (1973) Three-dimensional structure of yeast phenylalanine transfer RNA: Folding of the polynucleotide chain. *Science* **179**(4070), 285–288. https://doi.org/10.1126/science.179.4070.285.

Kinjo AR, *et al* (2018) New tools and functions in data-out activities at Protein Data Bank Japan (PDBj). *Protein Science* **27**(1), 95–102. https://doi.org/10.1002/pro.3273.

Koonin EV and Makarova KS (2019) Origins and evolution of CRISPR-Cas systems. *Philos Transactions of the Royal Society of London Series B, Biological Sciences* **374**(1772), 20180087. https://doi.org/10.1098/rstb.2018.0087.

Krohn A *et al* (1991) Novel binding mode of highly potent HIV-proteinase inhibitors incorporating the (R)-hydroxyethylamine isostere. *Journal of Medicinal Chemistry* **34**(11), 3340–3342. https://doi.org/10.1021/jm00115a028.

Krupovic M and Bamford DH (2008) Virus evolution: how far does the double beta-barrel viral lineage extend? *Nature Reviews Microbiology* **6**(12), 941–948. https://doi.org/10.1038/nrmicro2033.

Kucukoglu B *et al* (2024) Low-dose cryo-electron ptychography of proteins at sub-nanometer resolution. *Nature Communications* **15**(1), 8062. https://doi.org/10.1038/s41467-024-52403-5.

Kuhlbrandt W (2014) Biochemistry. The resolution revolution. *Science* **343**(6178), 1443–1444. https://doi.org/10.1126/science.1251652.

Kupferschmidt K (2023) Bird flu spread between mink is a 'warning bell'. *Science* **379**(6630), 316–317. https://doi.org/10.1126/science.adg8342.

Lanman BA *et al* (2020) Discovery of a covalent inhibitor of KRAS(G12C) (AMG 510) for the treatment of solid tumors. *Journal of Medicinal Chemistry* **63**(1), 52–65. https://doi.org/10.1021/acs.jmedchem.9b01180.

Lawson CL *et al* (2018) New online curriculum: the PDB pipeline and data archiving. *Acta Crystallographica Section A* **74**(a1), a243. https://doi.org/10.1107/S0108767318097568.

Leitner A *et al* (2020) Toward increased reliability, transparency, and accessibility in cross-linking mass spectrometry. *Structure* **28**(11), 1259–1268. https://doi.org/10.1016/j.str.2020.09.011.

Liebschner D *et al* (2013) On the reproducibility of protein crystal structures: five atomic resolution structures of trypsin. *Acta Crystallographica Section D* **69**(Pt 8), 1447–1462. https://doi.org/10.1107/S0907444913009050.

Lin D *et al* (2020) The TRUST principles for digital repositories. *Scientific Data* **7**(1), 144. https://doi.org/10.1038/s41597-020-0486-7.

Lin D *et al* (2000) AutoDep: A web-based system for deposition and validation of macromolecular structural information. *Acta Crystallographica Section D* **56**, 828–841. https://doi.org/10.1107/S0907444900005655.

Lubin JH *et al* (2022) Evolution of the SARS-CoV-2 proteome in three dimensions (3D) during the first 6 months of the COVID-19 pandemic. *Proteins* **90**, 1054–1080. https://doi.org/10.1002/prot.26250.

Luger K *et al* (1997) Crystal structure of the nucleosome core particle at 2.8Å resolution. *Nature* **389**, 251–260. https://doi.org/10.1038/38444.

Martz E and Francoeur E (2004) History of Visualization of Biological Macromolecules. Available at https://www.umass.edu/microbio/rasmol/history.htm#physical (accessed 1 December, 2024).

McPherson A (2017) Protein crystallization. *Methods in Molecular Biology* **1607**, 17–50. https://doi.org/10.1007/978-1-4939-7000-1_2.

Meyer EF, Jr. *et al* (1974) CRYSNET, a crystallographic computing network with interactive graphics display. *Federation Procedures* **33**(12), 2402–2405.

Montelione GT *et al* (2013) Recommendations of the wwPDB NMR Validation Task Force. *Structure* **21**(9), 1563–1570. https://doi.org/10.1016/j.str.2013.07.021.

Moore PB *et al* (2022) The protein-folding problem: Not yet solved. *Science* **375**(6580), 507. https://doi.org/10.1126/science.abn9422.

Myler PJ *et al* (2009) The seattle structural genomics center for infectious disease (SSGCID). *Infectious Disorders – Drug Targets* **9**(5), 493–506. https://doi.org/10.2174/187152609789105687.

Norvell JC and Machalek AZ (2000) Structural genomics programs at the US National Institute of General Medical Sciences. *Nature Structural Biologyl* **7 Suppl,** 931. https://doi.org/10.1038/80694.

Orengo CA *et al* (1997) CATH–a hierarchic classification of protein domain structures. *Structure* **5**, 1093–1108. https://doi.org/10.1016/s0969-2126(97)00260-8.

https://doi.org/10.1017/S0033583525000034 Published online by Cambridge University Press

Owen DR *et al* (2021) An oral SARS-CoV-2 M(pro) inhibitor clinical candidate for the treatment of COVID-19. *Science* **374**(6575), 1586–1593. https://doi.org/10.1126/science.abl4784.

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. https://doi.org/10.1038/185416a0.

Prior IA, **Hood FE and Hartley JL** (2020) The frequency of ras mutations in cancer. *Cancer Research* **80**(14), 2969–2974. https://doi.org/10.1158/0008-5472.CAN-19-3682.

Protein Data Bank (1971) Crystallography: Protein Data bank. *Nature New Biology* **233**(42), 223–223. https://doi.org/10.1038/newbio233223b0.

Protein Structure Bioinformatics Group (2024) ModelArchive. Available at modelarchive.org (accessed 2024).

Read RJ *et al* (2011) A new generation of crystallographic validation tools for the protein data bank. *Structure* **19**(10), 1395–1412. https://doi.org/10.1016/j.str.2011.08.006.

Reeve SM, **Lombardo MN and Anderson AC** (2015) Understanding the structural mechanisms of antibiotic resistance sets the platform for new discovery. *Future Microbiology* **10**(11), 1727–1733. https://doi.org/10.2217/fmb.15.78.

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. https://doi.org/10.1016/0022-2836(68)90085-5.

Rigden DJ and Fernandez XM (2022) The 2022 Nucleic Acids Research database issue and the online molecular biology database collection. *Nucleic Acids Research* **50**(D1), D1–D10. https://doi.org/10.1093/nar/gkab1195.

Rigden DJ and Fernandez XM (2023) The 2023 Nucleic Acids Research Database Issue and the online molecular biology database collection. *Nucleic Acids Research* **51**(D1), D1–D8. https://doi.org/10.1093/nar/gkac1186.

Robertus JD *et al* (1974) Structure of yeast phenylalanine tRNA at 3 Å resolution. *Nature* **250**, 546–551. https://doi.org/10.1038/250546a0.

Romero PR *et al* (2020) BioMagResBank (BMRB) as a resource for structural biology. *Methods in Molecular Biology* **2112**, 187–218. https://doi.org/10.1007/978-1-0716-0270-6_14.

Sali A *et al* (2015) Outcome of the first wwPDB hybrid/integrative methods task force workshop. *Structure* **23**(7), 1156–1167. https://doi.org/10.1016/j.str.2015.05.013.

Schluenzen F *et al* (2000) Structure of functionally activated small ribosomal subunit at 3.3 Å resolution. *Cell* **102**, 615–623. https://doi.org/10.1016/s0092-8674(00)00084-2.

Sehnal D *et al* (2021) Mol* Viewer: Modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Research* **49**, W431–W437. https://doi.org/10.1093/nar/gkab314.

Shaikh F *et al* (2021) LigTMap: Ligand and structure-based target identification and activity prediction for small molecular compounds. *Journal of Cheminformatics* **13**(1), 44. https://doi.org/10.1186/s13321-021-00523-1.

Shao C *et al* (2022) Assessing PDB macromolecular crystal structure confidence at the individual amino acid residue level. *Structure* **30**, 1385–1394. https://doi.org/10.1016/j.str.2022.08.004.

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 & Cellular Proteomics* **13**(11), 2927–2943. https://doi.org/10.1074/mcp.M114.041673.

Simmons KJ, **Chopra I and Fishwick CW** (2010) Structure-based discovery of antibacterial drugs. *Nature Reviews Microbiology* **8**(7), 501–510. https://doi.org/10.1038/nrmicro2349.

Singla J *et al* (2018) Opportunities and Challenges in Building a Spatiotemporal Multi-scale Model of the Human Pancreatic beta Cell. *Cell* **173**(1), 11–19. https://doi.org/10.1016/j.cell.2018.03.014.

Smart OS *et al* (2018a) Validation of ligands in macromolecular structures determined by X-ray crystallography. *Acta Crystallographica Section D* **74**(Pt 3), 228–236. https://doi.org/10.1107/S2059798318002541.

Smart OS *et al* (2018b) Worldwide Protein Data Bank validation information: usage and trends. *Acta Crystallographica Section D* **74**(Pt 3), 237–244. https://doi.org/10.1107/S2059798318003303.

Stacy R, **Anderson WF and Myler PJ** (2015) Structural genomics support for infectious disease drug design. *ACS Infectious Diseases* **1**(3), 127–129. https://doi.org/10.1021/id500048p.

Stampf D, **Felder C and Sussman J** (1995) PDBBrowse – a graphics interface to the Brookhaven Protein Data Bank. *Nature* **374**, 572–574. https://doi.org/10.1038/374572a0.

Subramaniam S and Kleywegt GJ (2022) A paradigm shift in structural biology. *Nature Methods* **19**(1), 20–23. https://doi.org/10.1038/s41592-021-01361-7.

Sullivan KP, **Brennan-Tonetta P and Marxen LJ** (2017) *Economic Impacts of the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank*. https://doi.org/10.2210/rcsb_pdb/pdb-econ-imp-2017.

Terwilliger TC *et al* (2024) AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. *Nature Methods* **21**, 110–116. https://doi.org/10.1038/s41592-023-02087-4.

Terwilliger TC *et al* (2022) Improved AlphaFold modeling with implicit experimental information. *Nature Methods* **19**(11), 1376–1382. https://doi.org/10.1038/s41592-022-01645-6.

The Stakeholder Alignment Collaborative (2025) *The Consortia Century Aligning for Impact*. New York, NY: Oxford University Press.

Tinberg CE *et al* (2013) Computational design of ligand-binding proteins with high affinity and selectivity. *Nature* **501**(7466), 212–216. https://doi.org/10.1038/nature12443.

Trewhella J (2018) Small angle scattering and structural biology: Data quality and model validation. *Advances in Experimental Medicine and Biology* **1105**, 77–100. https://doi.org/10.1007/978-981-13-2200-6_7.

Trewhella J, ***et al*** (2013) Report of the wwPDB Small-Angle Scattering Task Force: Data requirements for biomolecular modeling and the PDB. *Structure* **21**(6), 875–881. https://doi.org/10.1016/j.str.2013.04.020.

Ulrich EL *et al* (2008) BioMagResBank. *Nucleic Acids Research* **36**(Database issue), D402–408. https://doi.org/10.1093/nar/gkm957.

UniProt Consortium (2023) UniProt: The Universal protein knowledgebase in 2023. *Nucleic Acids Research* **51**, D523–D531. https://doi.org/10.1093/nar/gkac1052.

United States Patent and Trademark Office (2022) Available at www.uspto.gov (accessed 2022).

Vallat B *et al* (in press) PDB-IHM: A system for deposition, curation, validation, and dissemination of integrative structures. *Journal of Molecular Biology*

Vallat B *et al* (2023) ModelCIF: An extension of PDBx/mmCIF data representation for computed structure models. *Journal of Molecular Biology* https://doi.org/10.1016/j.jmb.2023.168021.

Vallat B, ***et al*** (2021) New system for archiving integrative structures. *Acta Crystallographica Section D* **77**(Pt 12), 1486–1496. https://doi.org/10.1107/S2059798321010871.

Vallat B *et al* (2018) Development of a prototype system for archiving integrative/hybrid structure models of biological macromolecules. *Structure* **26**(6), 894–904 e892. https://doi.org/10.1016/j.str.2018.03.011.

van der Aalst WMP, **Bichler M and Heinzl A** (2017) Responsible data science. *Business & Information Systems Engineering* **59**(5), 311–313. https://doi.org/10.1007/s12599-017-0487-z.

Varadi M *et al* (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research* **50**(D1), D439–D444. https://doi.org/10.1093/nar/gkab1061.

Voss-Andreae J (2005) Protein sculptures: Life's building blocks inspire art. *Leonardo* **38**, 41–45. https://doi.org/10.1162/leon.2005.38.1.41.

Walker ME, **Simpson JB and Redinbo MR** (2022) A structural metagenomics pipeline for examining the gut microbiome. *Current Opinion in Structural Biology* **75**, 102416. https://doi.org/10.1016/j.sbi.2022.102416.

Watenpaugh KD *et al* (1972) The structure of a non-heme iron protein: Rubredoxin at 1.5 Angstrom resolution. *Cold Spring Harbor Symposia on Quantitative Biology* **36**, 359–367. https://doi.org/10.1101/sqb.1972.036.01.047.

Watenpaugh KD, **Sieker LC and Jensen LH** (1980) Crystallographic refinement of rubredoxin at 1 x 2 A degrees resolution. *Journal of Molecular Biology* **138**(3), 615–633. https://doi.org/10.1016/s0022-2836(80)80020-9.

Westbrook JD and Burley SK (2019) How structural biologists and the protein data bank contributed to recent FDA new drug approvals. *Structure* **27**, 211–217. https://doi.org/10.1016/j.str.2018.11.007.

Westbrook JD *et al* (2020) Impact of protein data bank on antineoplastic approvals. *Drug Discovery Today* **25**, 837–850. https://doi.org/10.1016/j.drudis.2020.02.002.

https://doi.org/10.1017/S0033583525000034 Published online by Cambridge University Press

Westbrook JD *et al* (2022) PDBx/mmCIF Ecosystem: Foundational semantic tools for structural biology. *Journal of Molecular Biology* **434**, 167599. https://doi.org/10.1016/j.jmb.2022.167599.

Wilkinson MD *et al* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**(160018), 1–9. https://doi.org/10.1038/sdata.2016.18.

Williams CJ *et al* (2018) MolProbity: More and better reference data for improved all-atom structure validation. *Protein Science* **27**(1), 293–315. https://doi.org/10.1002/pro.3330.

Williamson MP, **Havel TF and Wuthrich K** (1985) Solution conformation of proteinase inhibitor IIA from bull seminal plasma by 1H nuclear magnetic resonance and distance geometry. *Journal of Molecular Biology* **182**(2), 295–315. https://doi.org/10.1016/0022-2836(85)90347-x.

Wlodawer A *et al* (1989) Conserved folding in retroviral proteases: Crystal structure of a synthetic HIV-1 protease. *Science* **245**(4918), 616–621. https://doi.org/10.1126/science.2548279.

wwPDB (2024) wwPDB Deposition Policies and wwPDB Biocuration Procedures version 5.4 https://www.wwpdb.org/documentation/policy. (accessed 1 December, 2024).

wwPDB Consortium (2023) EMDB-the electron microscopy data bank. *Nucleic Acids Research* **52**, D456–D465. https://doi.org/10.1093/nar/gkad1019.

Wyckoff HW *et al* (1967) The structure of ribonuclease-S at 6 Å resolution. *Journal of Biological Chemistry* **242**, 3749–3753. https://doi.org/10.1016/S0021-9258(18)95874-6.

Xu W *et al* (2023) Announcing the launch of Protein Data Bank China as an Associate Member of the Worldwide Protein Data Bank Partnership. *Acta Crystallographica Section D***79**, 792–795. https://doi.org/10.1107/S2059798323006381.

Young JY *et al* (2018) Worldwide Protein Data Bank biocuration supporting open access to high-quality 3D structural biology data. *Database* **2018**, bay002. https://doi.org/10.1093/database/bay002.

Young JY *et al* (2017) OneDep: Unified wwPDB system for deposition, biocuration, and validation of macromolecular structures in the pdb archive. *Structure* **25**(3), 536–545. https://doi.org/10.1016/j.str.2017.01.004.

Zheng J *et al* (2009) From molecular to macroscopic via the rational design of a self-assembled 3D DNA crystal. *Nature* **461**(7260), 74–77. https://doi.org/10.1038/nature08274.

https://doi.org/10.1017/S0033583525000034 Published online by Cambridge University Press