ANNUAL REPORT 2017

RCSB PDB
PROTEIN DATA BANK

rcsb.org

A Living Digital Data Resource that Enables Scientific Breakthroughs

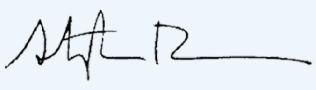
Director's Message

PDB structures are the molecules of life, and knowledge of their 3D structures (shapes), how they evolved with time, and their functions are essential for understanding critical areas of science.

The Protein Data Bank (PDB) is the single global archive of 3D macromolecular structure data, providing compelling insight into many of the diverse molecular machines found in living organisms and viruses. This powerful resource is jointly managed by the Worldwide Protein Data Bank organization, within which the RCSB PDB is responsible for US PDB operations.

The mission of the RCSB Protein Data Bank (RCSB PDB) is to sustain a unique living data resource of PDB structure information following the FAIR Guiding Principles for scientific data management and stewardship–structure data need to be Findable, Accessible, Interoperable, and Reusable.[1] By following these FAIR principles, usage of PDB data and RCSB PDB Services drive patent applications, drug discovery and development, publication of innovative research in scientific disciplines ranging from Agriculture to Zoology, and innovations leading to discovery and development of life-changing biopharmaceutical products.[2]

Efficient, ongoing fulfillment of the RCSB PDB mission depends critically on extensive collaborations and partnerships with PDB Data Depositors and Data Consumers, chemical and biological data resources, international leaders in biomolecular data curation (e.g., UniProt, Worldwide Protein Data Bank), scientific and professional societies, and stakeholder groups with interests in the life sciences, chemistry, computational sciences, medical sciences, drug discovery, biotechnology, and education. Together with this dynamic and evolving network of collaborators, RCSB PDB is committed to delivering data and related cyberinfrastructure services, and providing significant value to diverse user communities in the US and around the globe.



Stephen K. Burley, M.D., D.Phil.

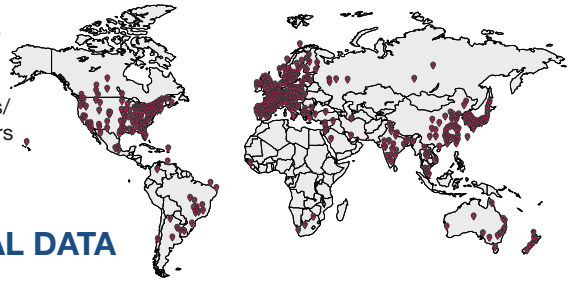
Director, RCSB PDB

University Professor and Henry Rutgers Chair
Rutgers, The State University of New Jersey

Adjunct Professor, University of California, San Diego

RCSB PDB SERVICES

Molecular machines studied by **>31,000** Structural Biologists/ PDB Data Depositors worldwide

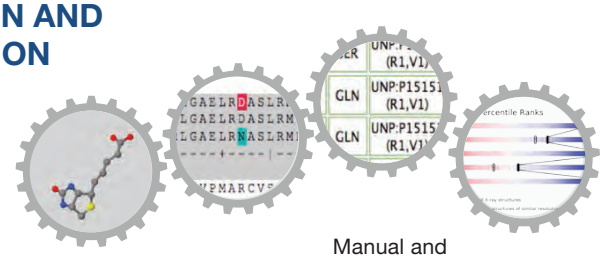


STRUCTURAL DATA

DEPOSITION AND BIOCURATION

WORLDWIDE PDB PROTEIN DATA BANK

Support Data Depositors worldwide with expert review and validation



Ligand Processing | Sequence Processing | Manual and Automatic Annotation | Validation

ARCHIVE MANAGEMENT AND ACCESS

ftp.wwpdb.org

Support PDB Data Consumers by maintaining the PDB archive and integrating PDB data with other available information

Data downloaded and utilized by **>400** different scientific resources

DATA EXPLORATION

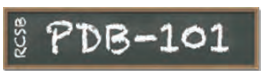
rcsb.org

Tools for query, analysis and visualization

>1 Million unique users per year

OUTREACH AND EDUCATION

pdb101.rcsb.org



For teachers, students, and the general public

>620,000 unique users per year

GLOBAL IMPACT

RCSB PDB: Enabling Breakthroughs in Scientific and Biomedical Research and Education

Protein Data Bank (PDB) was established as the 1st open access global resource for 3D macromolecular structure data.

Knowing the 3D structure of a biological macromolecule is essential for understanding its role in human and animal health and disease, its function in plants and food and energy production, and its importance to other topics related to global prosperity and sustainability.

THE PDB ARCHIVE:

136,472 Structures of Proteins, DNA, and RNA (December 31, 2017)

The cost to replicate the contents of the PDB archive is estimated at
$14 billion

PDB DATA AND RCSB DATA SERVICES IMPACT

Scientific Inquiry

Contributed data to nearly **1 MILLION** published research papers

Aid basic and applied research in subject areas from Agriculture to Zoology

Medicine

Lead to development of patent applications

Drug Discovery

Aid discovery of lifesaving drugs

Technology

Are a crucial part of innovative processes that lead to product development and company formation

Education

Support STEM education on all levels

>1,000,000 Data Consumers Served Each Year

Researchers, scientists, educators, students, curious public, medical professionals, patients, and patient advocates

Private sector, including pharmaceutical and biotechnology companies

Generating return on investment
~1500 times the federal funding[3]

Supporting US-Funded Research

- PDB connects US-funded research and scientists with worldwide structural biology data from public and private sector research
- PDB is 2nd most heavily cited online data resource after ClinicalTrials.gov by NIH-funded researchers
- PDB enables interdisciplinary collaborations and accelerate method development by highly visible collaborative networks
- PDB safeguards research and serves as the Data Management Plan
- PDB data are interoperable with current and future resources
- PDB ensures rigor/reproducibility across basic and applied research

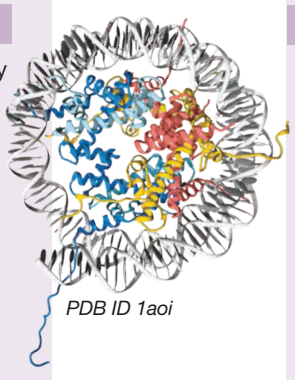
#1 CITED PDB STRUCTURE

NUCLEOSOME

>4,900 CITATIONS

Fundamental Biology and the National Science Foundation:

PDB data are used to understand DNA packing into nucleosomes and chromatin



PDB ID 1aoi

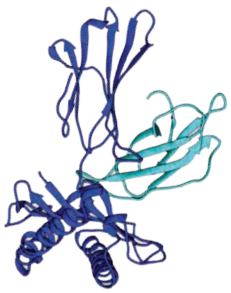
#4 CITED PDB STRUCTURE

MAJOR HISTOCOMPATIBILITY COMPLEX

>3,000 CITATIONS

Biomedicine and the National Institutes of Health:

MHC displays peptides on the surfaces of cells, allowing the immune system to sense the infection inside



PDB ID 1hla

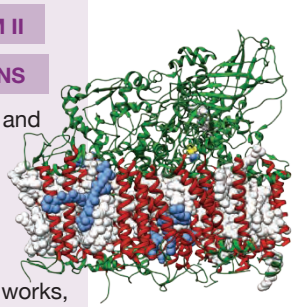
#14 CITED PDB STRUCTURE

PHOTOSYSTEM II

>2,200 CITATIONS

Photosynthesis and the Department of Energy:

PDB structures reveal how photosynthesis works, driving innovation in energy production



PDB ID 1s5l

DEPOSITION AND BIOCURATION

THE INTERNATIONAL COLLABORATION

The PDB archive is managed by the Worldwide PDB organization (wwPDB, **wwpdb.org**), an international collaboration involving regional data centers in the US, Europe, and Japan. wwPDB partners ensure that these valuable data are securely stored, expertly managed, and made freely available for the benefit of researchers in basic and applied science, educators, and students around the globe.

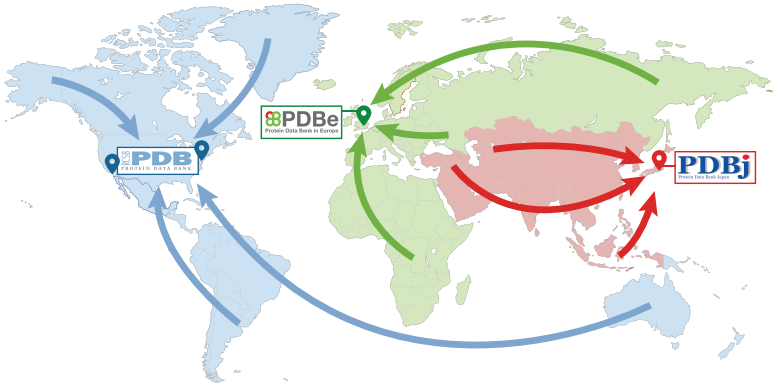
In 2017, 13,049 PDB structures were submitted by researchers from around the world and then biocurated, validated, and prepared for release by the wwPDB. This means that the RCSB PDB Deposition/Biocuration team received ~20 new structures each day. New depositions contain 3D atomic coordinates, experimental data, and related metadata that drive research across disciplines when made publicly available. Biocurators work closely with depositors to bring the most accurate data possible into the PDB archive. Specialized software is used to review and annotate sequences, chemistry, cross-references to other databases, experimental data, correspondence of atomic coordinates with primary experimental data, and more. This process helps to ensure that incoming data meet community-defined quality standards, thereby enabling meaningful analyses and comparisons across the entire archive.

DATA QUALITY STANDARDS

wwPDB data centers work closely with community experts to define data standards and resolve data representation challenges. Volunteer Task Forces make recommendations and contribute software tools used to generate wwPDB Validation Reports assessing the quality and accuracy of every structure stored in the PDB archive. Validation of structural data deposited to the PDB archive also helps to ensure the integrity of the peer-reviewed scientific literature. Access to validation reports helps referees and editors better evaluate the structure and improve publication quality. These same wwPDB Validation Reports are publicly available, helping to identify structures are of sufficient quality and accuracy for intended study.

2017

wwPDB data centers received **13,049** structures from researchers around the world



ARCHIVE MANAGEMENT AND ACCESS

MANAGING "BIG DATA" AS GLOBAL PUBLIC GOOD

RCSB PDB Archive Management/Access Services fulfill our responsibilities as the wwPDB "archive keeper", which include safeguarding the PDB archive and maintaining the PDB FTP data access system (**ftp.wwpdb.org**).

RCSB PDB coordinates weekly updates of the PDB archive with wwPDB data centers in Europe and Japan. Calculations are run as part of this process to generate clusters of similar sequences and 3D structures to support search and analysis applications. Data are also integrated with ~40 external data resources from across the Life Sciences ecosystem.

Additional responsibilities include developing the PDBx/mmCIF Data Dictionary used to support standardization across the archive and registration of Digital Object Identifiers (DOI) for all PDB structures. These DOIs can be used to access PDB data files directly and serve as references to PDB structures.

In addition to the website (**RCSB.org**), RCSB PDB provides numerous Web Services to access PDB structure data and data files. Several software libraries are made available as open source on GitHub.

2017

11,115 expertly annotated structures added to the PDB archive (**ftp.wwpdb.org**)

THE PDB ARCHIVE CONTENT, DECEMBER 31, 2017

136,472 Atomic structures of proteins, DNA, and RNA

Total PDB Archive traffic from all wwPDB partners:

Molecule Type	126,690	Proteins, peptides, and viruses
	3,174	Nucleic acids
	6,582	Protein/nucleic acid complexes
	26	Other
Experimental Technique	122,175	X-ray
	12,081	NMR
	1,881	Electron Microscopy
	111	Hybrid
	224	Other
Related Experimental Data Files	111,969	Structure factors
	9,425	NMR restraints
	3,137	Chemical shifts
	1,885	3DEM map files

> 679 million data file downloads

~75% downloaded from RCSB PDB web and FTP sites

DATA EXPLORATION

TOOLS FOR QUERY, ANALYSIS, AND VISUALIZATION

RCSB PDB Data Exploration Services utilize Web Services to deliver the information packaged and computed by the Archive Management/Access team to users around the globe.

The website at **RCSB.org** supports >1 million users representing a broad range of skills and interests. A variety of search and analysis tools provide access to structure data, comparative data, and external annotations such as information about point mutations and genetic variations.

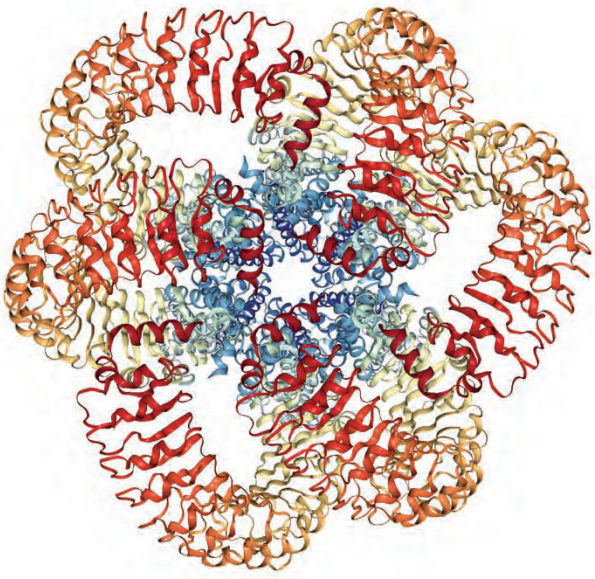
Each PDB structure is presented on a Structure Summary page that organizes access to important information, including a snapshot of the validation report and other high-level content, annotations, sequence information, sequence and structure similarity clusters, and experimental data. These data are updated weekly, which means that while the corresponding scientific publication stays static, RCSB PDB displays contemporary views of all structures.

Structure Summary pages also offer fast, interactive 3D display of molecular complexes containing millions of atoms–without plug-ins–on desktop computers and even smartphones using the NGL (New Graphic Library) Viewer. NGL Viewer uses a binary compressed format (Macromolecular Transmission Format) to massively reduce network transfer and parsing times.

These rich structural views of biological systems are provided to enable breakthroughs in scientific inquiry, medicine, drug discovery, technology, and education.

2017

RCSB.org supported **>1 Million** users



The NGL Viewer allows for rapid, interactive 3D display of molecular complexes across all platforms. Shown: PDB Structure 6g9l; Structure of a volume-regulated anion channel of the LRRC8 family (2018) D. Deneka et al. Nature **558***: 254-259.*

OUTREACH AND EDUCATION

PDB-101

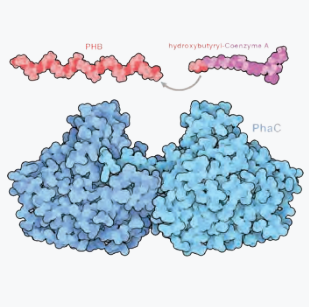
MOLECULAR EXPLORATIONS THROUGH BIOLOGY AND MEDICINE

The molecular stories of PDB structures reach an audience outside of the research community through RCSB PDB's Outreach and Education efforts.

PDB-101 is an online portal designed for teachers, students, and the curious public to promote exploration in the world of proteins, DNA, and RNA. Learning about the diverse shapes and functions of these biological macromolecules promotes understanding of all aspects of biomedicine and agriculture, from protein synthesis to health and disease to biological energy. Videos, posters, PDB data user guides, and other content support today's students who will become tomorrow's PDB users.

2017

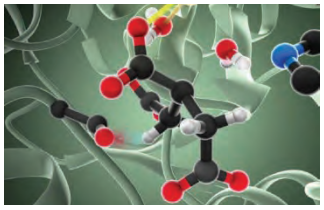
PDB-101 (**pdb101.rcsb.org**) provided training and support for **>620,000 visitors** interested in:



December 2017 feature on Biodegradable Plastic

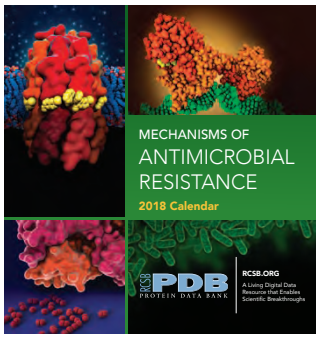
Molecule of the Month
This popular series presents short accounts describing selected molecules from the PDB archive that highlight stories surrounding Fundamental Biology, Biomedicine, and Energy.

More than 220 articles are available, ranging from Actin to Zika. In 2017, the most heavily-accessed PDB-101 articles included catalase, hemoglobin, lysozyme, green fluorescent protein, and insulin.



The video "How enzymes work?" explains the process of catalysis

Materials for Learning
A variety of materials to support exploration are available, including videos, posters, and printable templates for building molecular origami. PDB-101 can be searched by keyword or browsed by popular topic categories.



The 2018 calendar features PDB structures related to antibiotic action and resistance

Focus on Public Health: Antimicrobial Resistance
PDB-101 resource development is heavily focused on global health concerns. The release of the 2017 calendar kicked off a 2-year focus on Antimicrobial Resistance that will lead to the development of curricular modules and other educational materials. Curricular modules and other educational materials help explore other health topics, including HIV/AIDS and insulin/diabetes.

REFERENCES

1. MD Wilkinson *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**: 160018.
2. SK Burley *et al.* (2018) RCSB Protein Data Bank: Sustaining a living digital data resource that enables breakthroughs in scientific research and biomedical education. *Protein Sci* **27**: 316–330.
3. KP Sullivan *et al.* Economic Impacts of the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank. Rutgers Office of Research Analytics (2017). doi: 10.2210/rcsb_pdb/pdb-econ-imp-2017

CITE THE RCSB PDB

The Protein Data Bank (2000) *Nucleic Acids Res* **28**: 235-242. doi: 10.1093/nar/28.1.235

The RCSB Protein Data Bank: Views of structural biology for basic and applied research and education (2015) *Nucleic Acids Res* **43**: D345-D356. doi:10.1093/nar/gku1214

RCSB PDB
PROTEIN DATA BANK

rcsb.org • info@rcsb.org

RCSB PDB is funded by a grant (DBI-1338415) from the National Science Foundation, the National Institutes of Health, and the US Department of Energy.

The RCSB PDB is a member of the wwPDB organization (wwPDB.org).

The RCSB PDB is managed by the members of the Research Collaboratory for Structural Bioinformatics: Rutgers and UCSD/SDSC

RUTGERS

UC San Diego

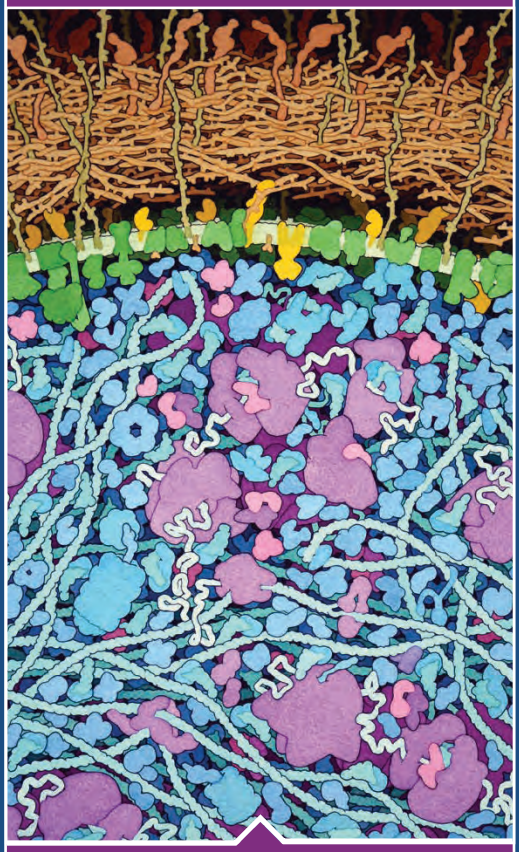
SDSC SAN DIEGO SUPERCOMPUTER CENTER

/RCSBPDB

/buildmodels

/RCSBProteinDataBank

/rcsb



ABOUT THE COVER

STAPHYLOCOCCUS AUREUS
DAVID S. GOODSELL

This illustration shows the surface of a *Staphylococcus aureus* bacterium, highlighting the many ways this pathogen evades antibiotics. These include membrane-bound proteins (yellow) that break down beta-lactam antibiotics or expel antibiotics out of the cell, and many diverse cytoplasmic proteins (magenta) that destroy antibiotics or block binding to their cellular targets.