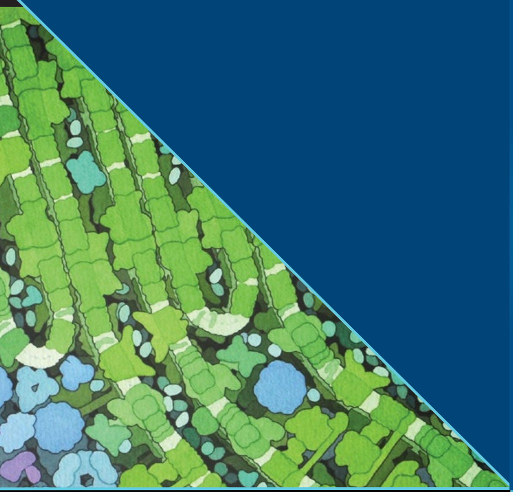
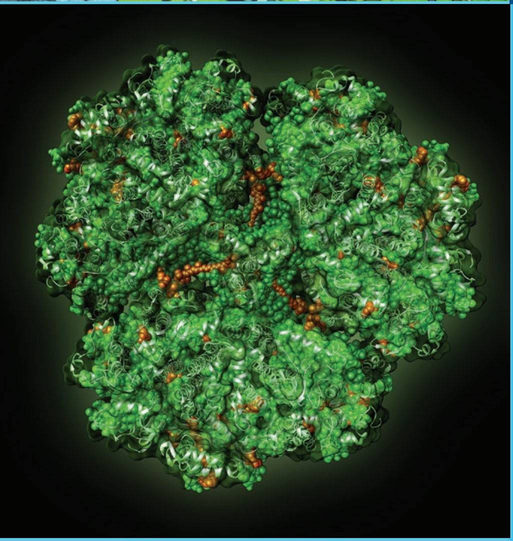
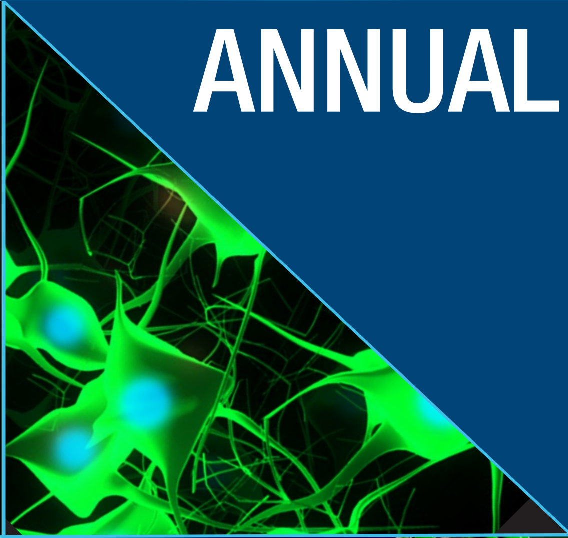
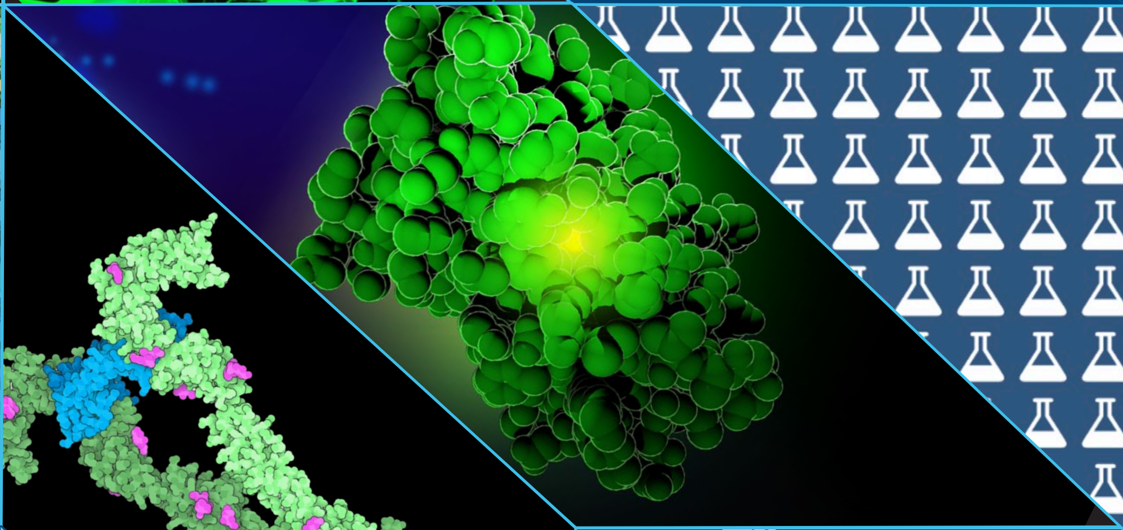


2022 ANNUAL REPORT

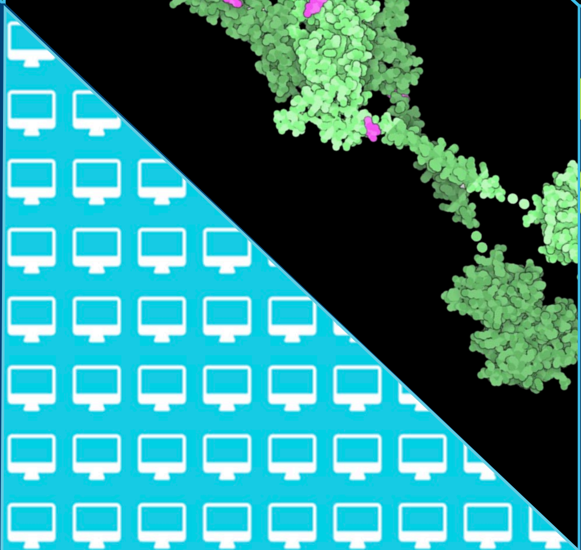


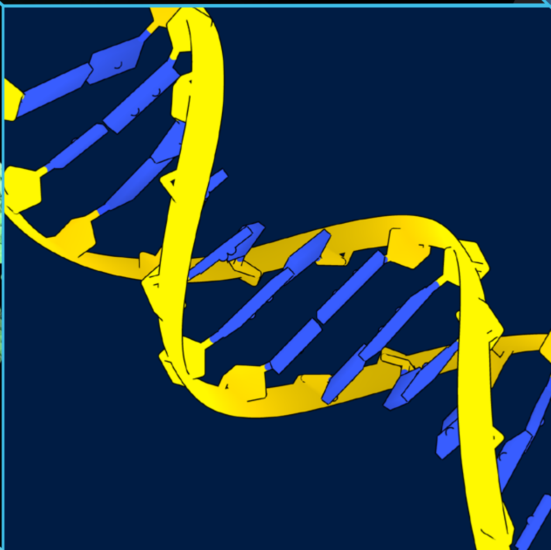




RCSB PDB
PROTEIN DATA BANK

RCSB.ORG A LIVING DIGITAL
DATA RESOURCE THAT ENABLES
SCIENTIFIC BREAKTHROUGHS





DIRECTOR'S MESSAGE

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

Director
RCSB Protein Data Bank

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

Research Scientist
San Diego Supercomputer Center

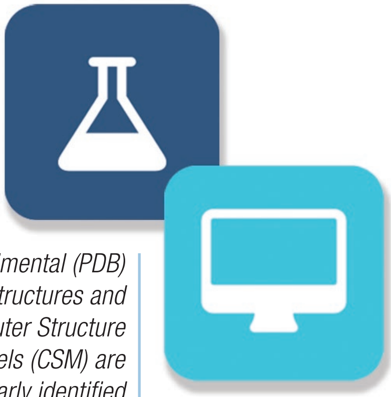
RCSB PDB (**RCSB.org**) is the US data center for the global Protein Data Bank (PDB) archive of three-dimensional (3D) structure data for large biological molecules (proteins, DNA, and RNA) essential for research and education in fundamental biology, biomedicine, energy sciences, and biotechnology. Team members at Rutgers, The State University of New Jersey, University of California San Diego, and University of California San Francisco support many millions of users around the world by delivering 3D biostructure data, tools for exploration and visualization, and resources for training, teaching, and understanding biology at the atomic level.

The PDB was established in 1971 as the first open-access digital data resource in all of biology, founded on bedrock values of open access and facile reuse. It is an exemplar of the FAIR (Findability, Accessibility, Interoperability, and Reusability) and FACT (FAIRness, Accuracy, Confidentiality, and Transparency) principles emblematic of responsible data stewardship for more than five decades. The PDB has been a vanguard in the open-access data movement. Beginning with only seven protein structures, the archive has grown to nearly 30 thousand-fold to become the authoritative data resource for atomic-level, experimentally-determined 3D biostructures. The PDB was recently designated by the Global Biodata Coalition as a Global Core Biodata Resource that is of fundamental importance to biological and life sciences communities.

International collaboration has been critical to the success of the archive. To recognize the archive as a global public good and ensure its sustainability, RCSB PDB led the formation of the Worldwide Protein Data Bank (wwPDB, wwpdb.org) in 2003. wwPDB members serve as regional data centers for 3D biostructure data curation and collaborate on software development. Current wwPDB Full Members include RCSB PDB, PDB in Europe (PDBe), and PDB Japan (PDBj), plus two newer Full Members–Electron Microscopy Data Bank (EMDB) and Biological Magnetic Resonance Bank (BMRB). Protein Data Bank China (PDBc) was admitted in 2022 as a wwPDB Associate Member.

The PDB has global reach. Approximately 60,000 structural biologists working on every inhabited continent have freely contributed data from their experiments to the archive. This information is today used in more than 200 United Nations-recognized sovereign countries by basic and applied researchers and their trainees, and educators and their students. The PDB is a community biodata resource central to understanding fundamental biology; explaining how human, animal, bacterial, fungal, and viral proteins contribute to health and cause disease; facilitating discovery of small-molecule drugs and design of biologics (*e.g.*, therapeutic antibodies) and vaccines; supporting basic and applied research in plant molecular biology and microbial physiology important for ensuring food security and addressing climate change; and enabling human and computer-based engineering of novel proteins to carry out biochemical tasks not previously seen in nature.

The availability of rigorously validated, expertly biocurated PDB data laid the groundwork for protein structure prediction using Artificial Intelligence and Machine Learning methods. In 2021, nearly one million Computed Structure Models (CSMs) of predicted proteins with accuracies comparable to that of lower-resolution experimental methods were made freely available from AlphaFold DB and RoseTTAFold (from Model Archive). Within a year, the RCSB PDB built capacity and capabilities to support delivery of more than one million CSMs alongside all 200,000 plus experimental PDB structures integrated with trusted external data resources using the full complement of RCSB.org tools for search, visualization, and analysis. Full 3D



Experimental (PDB) structures and Computer Structure Models (CSM) are clearly identified throughout RCSB.org: a dark-blue flask icon is used for PDB structures and a cyan computer icon for CSMs.

structural coverage is now provided at RCSB PDB for the human proteome, and those of model organisms (*Arabidopsis*, nematode worm, *C. albicans*, zebrafish, *Dictyostelium*, fruit fly, *E. coli*, soybean, *M. jannaschii*, mouse, Asian rice, rat, budding yeast, fission yeast, and maize), select bacterial pathogens (*e.g.*, *M. tuberculosis*), and organisms important for understanding photosynthesis and CO2 sequestration (*e.g.*, peat moss). As RCSB PDB scales its data delivery capacity further to incorporate hundreds of millions of additional CSMs, RCSB.org users will have access to nearly every protein represented in UniProt (also a Global Core Biodata Resource).

The future of structural biology, structural biologists, and the PDB appears even brighter now that CSMs are available for essentially any amino acid sequence, accelerating structural studies of ever larger and more complex macromolecular machines that make life possible. Parallel delivery of millions of CSMs alongside experimentally-determined PDB structures will make 3D biostructure information easier to access and even more valuable for the many millions of RCSB.org users around the world. Data exchange partnerships with more than 475 external bioinformatics resources, including the US Department of Energy Knowledgebase (KBase), support repackaging, reuse, and redistribution of PDB data at no charge and with no limitations on utilization under the most permissive Creative Commons CC0 license. In the coming decades, 3D structures archived in the PDB will shine ever more brightly, helping to illuminate the way forward as we confront pressing global challenges, such as climate change, pandemics that cross the species barrier into humans, substance abuse, and the fight against cancer, cardiovascular, and metabolic diseases.

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

RCSB PDB SERVICES & IMPACT

SERVICE 1

DATA DEPOSITION AND BIOCURATION

RCSB PDB and other members of the Worldwide PDB support >60,000 individual data depositors around the globe, ensuring quality for the ever-growing body of experimental structure information.

SERVICE 2

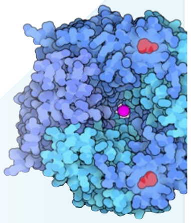
ARCHIVE MANAGEMENT AND ACCESS

RCSB PDB maintains the PDB archive according to FAIR principles, provides FTP access to the data, and integrates the structural information with other scientific resources.

SERVICE 3

DATA EXPLORATION

RCSB PDB develops tools for searching, visualization, and analysis of PDB structures and computed structure models (CSMs). These tools are freely available on **RCSB.org.**



SERVICE 4

TRAINING, OUTREACH, AND EDUCATION

RCSB PDB creates resources to support the broad user community in research and education.

PDB Data

- Enable research in subject areas from Agriculture to Zoology
- Contributed data to >1 million published research papers
- Used by 476 biological data resources

The PDB Archive

- Grows at the rate of nearly 10% per year
- Used to download >2 million structure data files per day
- Manages “Big Data” as a global Public Good
- Provides data critical to AI/ML development

PDB Data Impact

- Basic and applied research
- Patent applications
- Discovery of life-changing drugs
- Innovations that can lead to new product development and company formation
- PDB-101 materials illustrate how PDB data help explain fundamental biology, biomedicine, energy sciences, and biotechnology
- Training materials help our users make the most of the PDB data and tools in research and education

ECONOMIC IMPACT

The cost to replicate the contents of the PDB archive is estimated at **more than US $20 billion**

The PDB data and RCSB PDB services generate annual Return on Investment of **more than 1,500 times** federal funding

DEPOSITION AND BIOCURATION

Supporting Data Depositors who freely contribute the results of their structural studies of biological macromolecules to the PDB. All data deposited undergo expert review. Each structure is examined for self-consistency, standardized using controlled vocabularies, cross-referenced with other biological data resources, and validated for scientific/technical accuracy.

2022 STATISTICS

16,344
structures deposited and processed

3,327
new ligands created in the Chemical Component Dictionary

45
new Biologically Interesting Molecule Reference Dictionary items created

The Worldwide Protein Data Bank (wwPDB) was established to manage the PDB archive of macromolecular structural data, made freely and publicly available to the global community. It consists of organizations that act as deposition, data processing, and distribution centers for PDB data.

In 2022, a record 16,344 structures were deposited and processed (up from 14,573 in 2021). 3,327 new ligands and 45 new Biologically Interesting Molecule Reference dictionary items were created. RCSB PDB is currently responsible for managing depositions coming from the Americas and Oceania. Median processing time/new entry is 2.4 days (initial structure deposition, biocuration, and wwPDB validation report sent to depositors). Median time to finalize an entry is in 4 weeks (includes communication with Depositors and coordinate updates).

SARS-CoV-2 structures continued to be prioritized; 1,151 SARS-CoV-2 structures were deposited, rigorously validated, and expertly biocurated, and released in 2022; a total of 2,863 were available as of 31 December 2022.

PDB structures contain

- 3D atomic coordinates
- Experimental data
- Mandatory metadata
- Authors (*e.g.*, ORCID ID, PI contact information)
- Primary citation
- Sample preparation, data collection, and structure determination
- Polymer sequence(s) (proteins, DNA, RNA)
- Chemical ligand information

All deposited data undergo expert review by Ph.D.-trained biocurators. Each structure is examined for self-consistency, standardized using controlled vocabularies, cross-referenced with other biological data resources, and validated for scientific/technical accuracy.

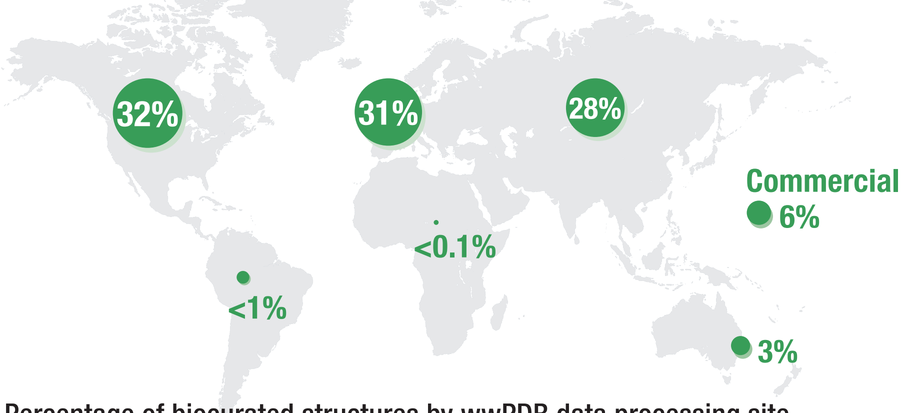
Validation is central to ensuring the highest quality data. wwPDB Working Groups and Task Forces, including more than 100 academic and industrial volunteers, make recommendations and contribute software tools used to generate wwPDB Validation Reports that assess the quality and accuracy of every structure stored in the PDB archive. Servers and APIs are provided for anonymous validation before data deposition, and reports are provided during biocuration.

Validation reports can be provided to journal editors and reviewers to help ensure the integrity of peer-reviewed scientific literature. Validation data are also provided publicly to enable meaningful analyses and comparisons across the entire archive.

2022 DEPOSITION STATISTICS

16,344 PDB structures were deposited by researchers from around the world

● Percentage of PDB Depositions by depositor location



Percentage of biocurated structures by wwPDB data processing site

42% RCSB PDB | **32% PDB Europe** | **26% PDB Japan**

ARCHIVE MANAGEMENT AND ACCESS

Supporting PDB Data Consumers by maintaining the PDB archive; data dictionary development and standardization, enabling global data delivery and DOI registration, and integrating PDB data with other available information.

A record 14,292 new PDB structures were released in 2022. They account for ~7% of the year-end total holdings of 199,755 available entries. The PDB archive went on to surpass the 200,000 structures mark on January 10, 2023.

In its role as wwPDB archive keeper, the RCSB PDB is responsible for safeguarding the PDB archive and maintaining the PDB FTP (ftp.wwpdb.org). RCSB PDB coordinates weekly updates of the PDB archive with wwPDB Data Centers in Europe and Japan.

PDB archival format and data standards are defined by the PDBx/mmCIF dictionary (mmCIF.wwpdb.org). At present, PDBx/mmCIF is jointly maintained by RCSB PDB, our wwPDB partners, and the wwPDB PDBx/mmCIF Working Group. Data dictionary terms and definitions are formulated, reviewed, and modified to support remediation of existing data and inclusion of new and rapidly evolving methodologies. This fully extensible data standard allows inclusion of data items and metadata elements for emerging technologies. For example, interoperation of CSMs with all tools and features at RCSB.org was enabled by the extension of the PDBx/mmCIF data standard with the related ModelCIF data standard developed for CSMs.

To support RCSB.org resources, calculations are run weekly to generate clusters of similar sequences and 3D structures to support search and analysis applications. Data are also integrated with 48 trusted external data resources from across the Life Sciences ecosystem.

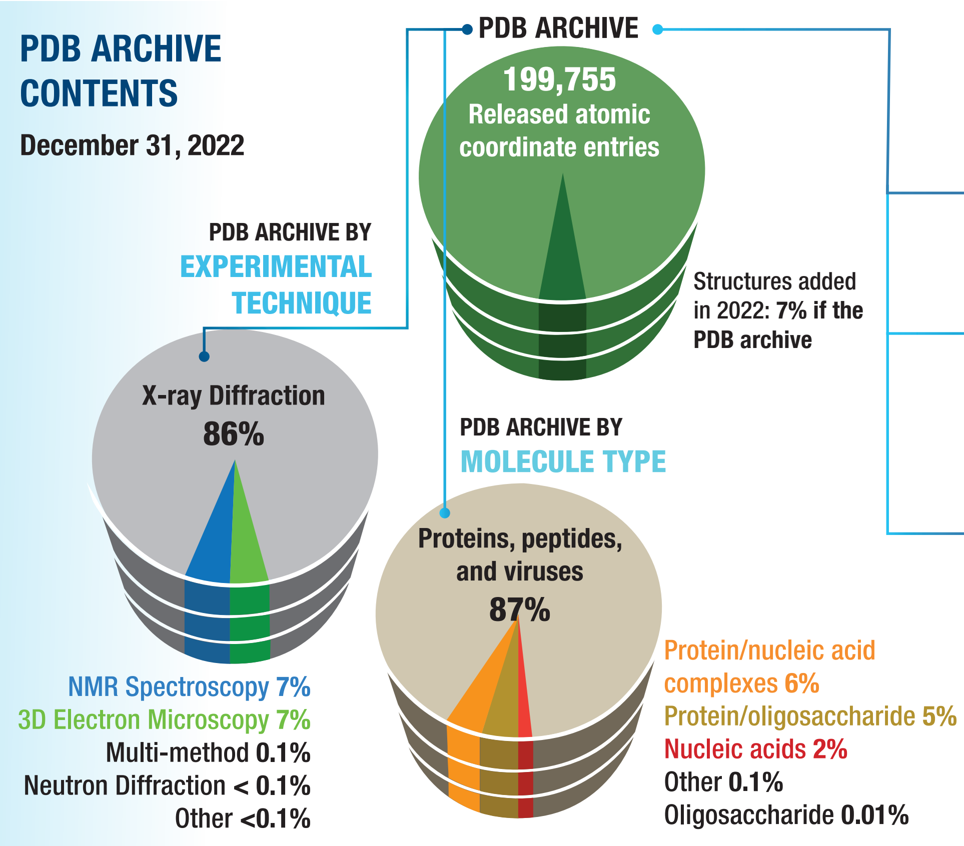
In 2022, 2.2 billion data files in various file formats, including structure files, experimental data files, chemical and molecular reference data files, and validation reports, were downloaded and/or viewed from RCSB PDB-hosted FTP and websites. Additional data were downloaded from wwPDB partners PDBe and PDBj, for a total of nearly **3 billion data file downloads**.

PDB data are also delivered through 476 external resources that repackage and redistribute PDB information, and through copies of the archival data stored inside for-profit company firewalls.

2022 STATISTICS

14,292
new structures released into the PDB archive

2.2 billion
data file downloads from RCSB PDB-hosted FTP and websites



PDB ARCHIVE CONTENTS

December 31, 2022

PDB ARCHIVE

199,755
Released atomic coordinate entries

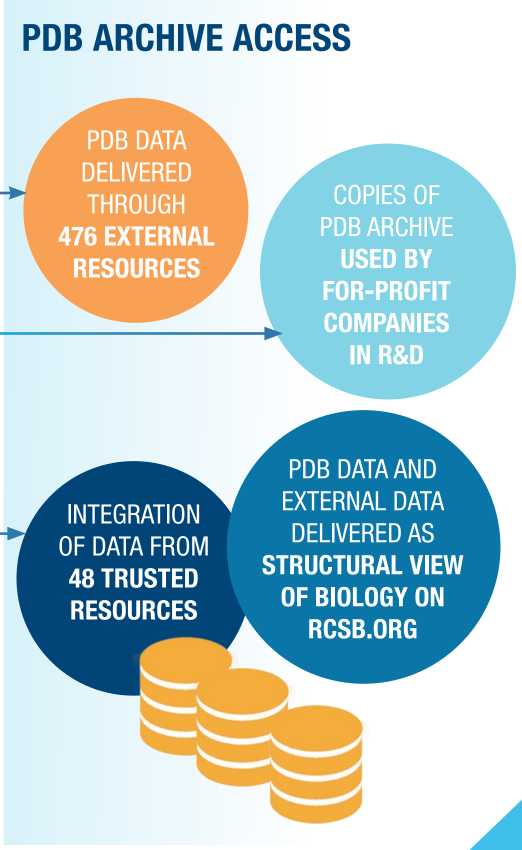
Structures added in 2022: 7% if the PDB archive

PDB ARCHIVE BY EXPERIMENTAL TECHNIQUE

X-ray Diffraction 86%
NMR Spectroscopy **7%**
3D Electron Microscopy **7%**
Multi-method **0.1%**
Neutron Diffraction **< 0.1%**
Other **<0.1%**

PDB ARCHIVE BY MOLECULE TYPE

Proteins, peptides, and viruses 87%
Protein/nucleic acid complexes **6%**
Protein/oligosaccharide **5%**
Nucleic acids **2%**
Other **0.1%**
Oligosaccharide **0.01%**



PDB ARCHIVE ACCESS

PDB DATA DELIVERED THROUGH **476 EXTERNAL RESOURCES**

COPIES OF PDB ARCHIVE **USED BY FOR-PROFIT COMPANIES IN R&D**

INTEGRATION OF DATA FROM **48 TRUSTED RESOURCES**

PDB DATA AND EXTERNAL DATA DELIVERED AS **STRUCTURAL VIEW OF BIOLOGY ON RCSB.ORG**

DATA EXPLORATION

Supporting PDB Data Consumers in the US and around the world through our open-access web portal RCSB.org that provides tools for structure visualization and analysis.

2022
STATISTICS

~7.2 million
clients (unique IP addresses)

~61 million/~37 million
total sessions/human user sessions

~2.8 billion
requests/interactions (downloads, service usage, web page content views)

RCSB PDB is one of the most heavily used biological data resources worldwide. For 2022, internal analytics reported ~7.2 million clients (unique IP addresses, up from ~6.8 million in 2021), 61 million sessions (~37 million of which are estimated to be human users), and ~2.8 billion requests/interactions (*e.g.*, data downloads, service usage, web page content views).

The website supports a broad range of skill levels and interests. In addition to retrieving 3D structure data, PDB users access comparative data, and external annotations, such as information about point mutations and genetic variations.

RCSB PDB services go well beyond the original structure and scientific publication. 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 remains static, RCSB PDB delivers contemporary views of all archived structures.

Our research-focused RCSB.org web portal offers a rich collection of software tools and features that can be used to search, browse, analyze, and visualize PDB data. These include powerful search and API services plus interactive analytical and visualization tools such as the Mol* molecular graphics system, 1D-3D protein feature view, and specialized tools that provide redundancy-reduced "groups" view of similar data.

A major project undertaken in 2022 involved extending the RCSB.org web portal and its functionalities to support parallel delivery of more than one million computed structure models (CSMs) from AlphaFold DB (alphafold.ebi.ac.uk) and the ModelArchive (modelarchive.org) alongside all of the experimental structures in the PDB archive. The infrastructure and software stack behind RCSB.org was updated to accommodate these new data, which represented a 6-fold increase in data volume. User guides and help documentation are provided for all new features.

Expansion of the purview of RCSB.org allows us to continue serving as a one-stop shop for studying 3D structures of biomolecules by providing PDB data consumers with access to CSMs covering the entire human proteome plus those of many model organisms, selected pathogens, organisms relevant to bioenergy research, and protein-protein complexes from select studies.

Cladogram showing PDB holdings. Within each of the three branches, PDB structure totals are provided for selected organisms. The PDB also houses 3D structures that solely contain nucleic acids (DNA, RNA, DNA-RNA hybrids, etc.) and/or viral proteins or human-designed proteins, which collectively accounted for ~8% of archival holdings as of end of 2022.

199,755 Total Number of Structures in the PDB as of December 31, 2022

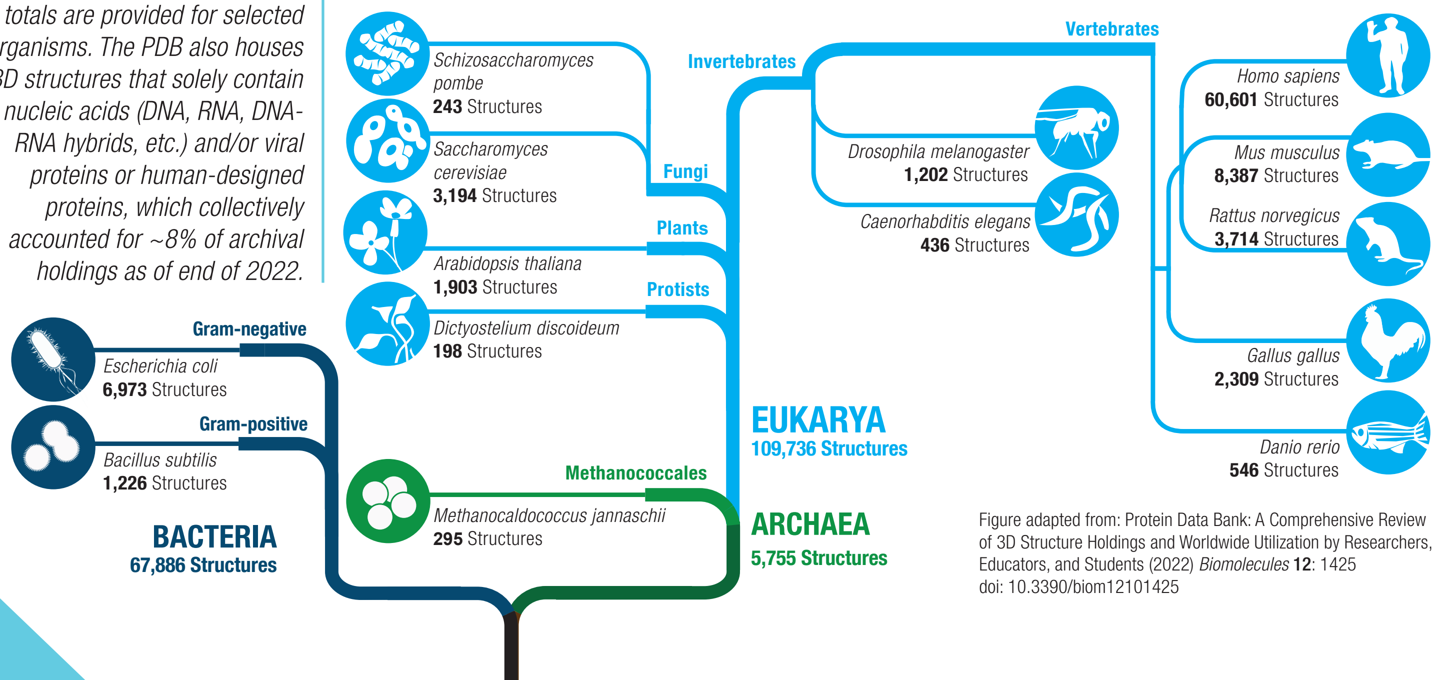


Figure adapted from: Protein Data Bank: A Comprehensive Review of 3D Structure Holdings and Worldwide Utilization by Researchers, Educators, and Students (2022) *Biomolecules* **12**: 1425
doi: 10.3390/biom12101425

TRAINING, OUTREACH, AND EDUCATION

Building and supporting the broad PDB user community with a wide range of resources for understanding 3D biostructures.

PDB-101 (**pdb101.rcsb.org**) is an online portal for exploring proteins and nucleic acids. The diverse shapes and functions of biological macromolecules help explain many aspects of biomedicine and agriculture, from protein synthesis to health and disease to biological energy.

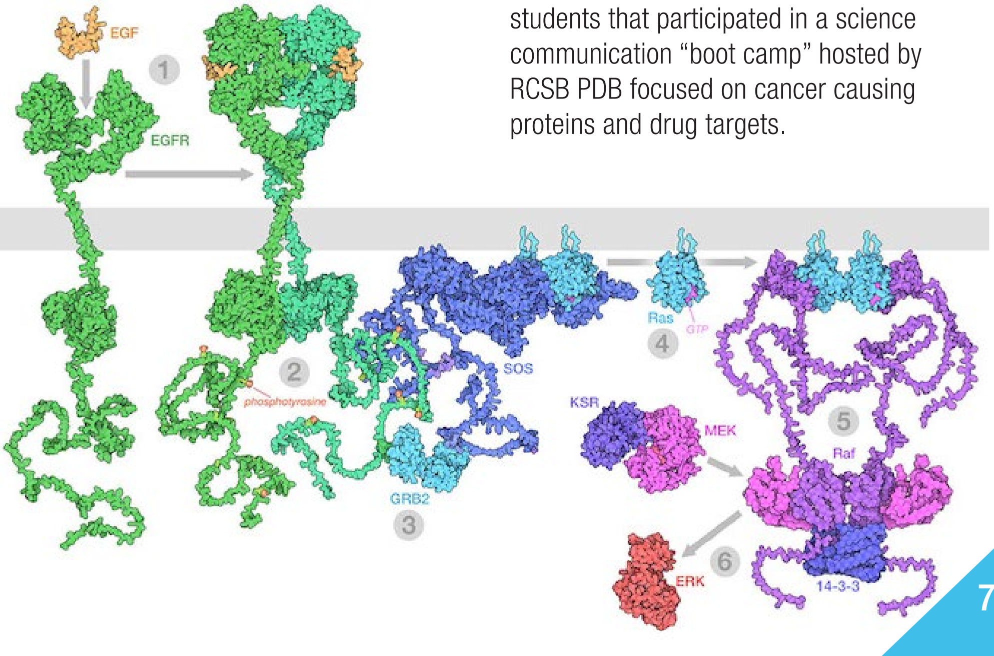
In 2022, **~663,000 users visited PDB-101** for

- Training materials, such as the *Guide to Understanding PDB Data* and webinars are available to help graduate students, postdoctoral scholars, and researchers use PDB data and RCSB PDB tools.
- Popular training course materials in 2022 highlighted how to explore CSMs at RCSB.org; how to access PDB data and CSMs from KBase; and Python scripting in biochemistry and molecular biology.
- Outreach content, including the *Molecule of the Month* series and molecular animations, demonstrate how PDB data can be used to understand fundamental biology, biomedicine, bioengineering/biotechnology, and energy sciences in 3D by a diverse and multidisciplinary user community.
- Education Materials, such as Curriculum Modules, provide lessons and activities for teaching and learning.

PDB-101 team members were recognized for their contributions in 2022. *Molecule of the Month* creator David S. Goodsell was honored with the Carl Brändén Award from The Protein Society, which recognizes an outstanding protein scientist, who has also made exceptional contributions in the areas of education and/or service, and the I. Fankuchen Award from the American Crystallographic Association, which recognizes contributions to crystallographic research by one who is known to be an effective teacher of crystallography. Scientific Training, Education, and Documentation Lead Shuchismita Dutta was named a 2022 Fellow of the American Association for the Advancement of Science in the section of Education.

Other PDB-101 content released in 2022 included the guide *Exploring Structural Biology of Cancer*, a video *How Neurons Communicate? An Introduction to Neurotransmission and Action Potential*, a video on Photosynthesis, tutorial videos in English and Spanish on folding a paper DNA model, and new Molecular Landscape illustrations in the David Goodsell SciArt gallery.

This illustration from the PDB-101 resource Exploring the Structural Biology of Cancer shows the Raf/MEK/ERK (MAPK) pathway. In six consecutive steps, it presents the key proteins that disseminate the growth signal inside the cell: the epidermal growth factor receptor, GRB2, SOS, Ras, Raf, MEK, KSR, and ERK. This illustration was created from decades of structural studies in PDB IDs 1egf, 1nql, 1m17, 2jwa, 3njp, 2gs6, 1gri, 1xd2, 3ksy, 5p21, 6xi7, 6q0j, 2y4i, 1pme, and unstructured chains from AlphaFold2.



SCIENTIFIC SUPPORT AND USER ENGAGEMENT

RCSB PDB regularly participates in meetings and scientific societies, hosts Working Groups, and convenes method-specific Task Forces. Recognized experts in fields, including but not limited to, structural biology, cell and molecular biology, computational biology, information technology, and education serve as advisors to the RCSB PDB.

Depositors and PDB Data Consumers are supported by responsive Help Desks covering all RCSB PDB and wwPDB services. User feedback helps inform prioritization and resource development to meet the needs of diverse research and education communities.

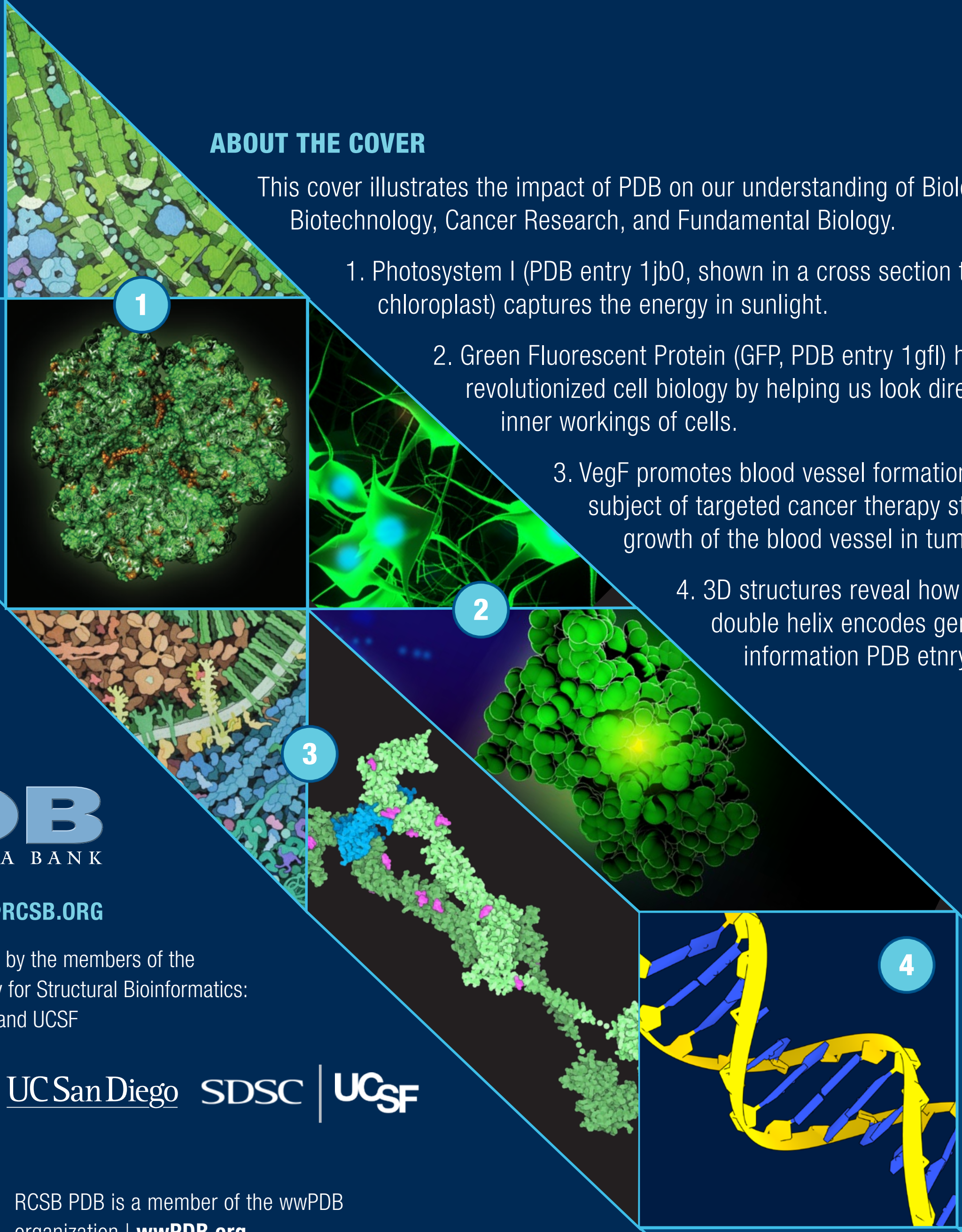
HEALTH FOCUS: MOLECULAR MECHANISMS OF CANCER

PDB-101 adopts a biennial public health theme to focus development and student engagement. The theme for 2022-2022 is *Molecular Mechanisms of Cancer*. These themes have been used at Rutgers for an Undergraduate Honors Seminar Course exploring the foundations of structural biology and for an annual RCSB PDB-sponsored video contest for high school students. Several *Molecule of the Month* features published in 2022 were co-written by undergraduate and graduate students that participated in a science communication “boot camp” hosted by RCSB PDB focused on cancer causing proteins and drug targets.

ABOUT THE COVER

This cover illustrates the impact of PDB on our understanding of Biological Energy, Biotechnology, Cancer Research, and Fundamental Biology.

1. Photosystem I (PDB entry 1jb0, shown in a cross section through a chloroplast) captures the energy in sunlight.
2. Green Fluorescent Protein (GFP, PDB entry 1gfl) has revolutionized cell biology by helping us look directly into the inner workings of cells.
3. VegF promotes blood vessel formation and is a subject of targeted cancer therapy stopping the growth of the blood vessel in tumors.
4. 3D structures reveal how the DNA double helix encodes genomic information PDB etnry 1bna).



RCSB PDB PROTEIN DATA BANK

RCSB.ORG • INFO@RCSB.ORG

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

RUTGERS | UC San Diego SDSC | UCSF

WORLDWIDE PDB PROTEIN DATA BANK

RCSB PDB is a member of the wwPDB organization | **wwPDB.org**

CORE TRUST SEAL

PDB is a CoreTrustSeal Board certified Trusted Digital Repository.

GBC GLOBAL CORE BIODATA RESOURCE

PDB is a Global Core Biodata Resource whose long term funding and sustainability is of fundamental importance to biological and life science communities.

CITE RCSB PDB

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

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 (2023) *Nucleic Acids Research* **51**: D488–D508 doi: **10.1093/nar/gkac1077**

FUNDING

RCSB PDB's core operations are funded by National Science Foundation (DBI-1832184), the US Department of Energy (DE-SC0019749), and the National Cancer Institute, National Institute of Allergy and Infectious Diseases, and National Institute of General Medical Sciences of the National Institutes of Health under grant R01GM133198.

FOLLOW US

/RCSBPDB

/RCSBProteinDataBank

/buildmodels

/rcsb