2023 ANNUAL REPORT

RCSB PDB
PROTEIN DATA BANK

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

DIRECTOR'S MESSAGE

The Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) works with community stakeholders to preserve and deliver rigorously-validated, expertly-biocurated, three-dimensional (3D) biostructure information archived in the Protein Data Bank (hereafter PDB or the archive) to many millions of PDB Depositors and Data Consumers worldwide at no charge and with no limitations on usage.

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

Exploration of these experimentally-determined structures of proteins and nucleic acids (and their complexes with one another and small-molecule ligands) via RCSB PDB web portals is enabled by state-of-the-art data analysis, visualization, and download tools.

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. The archive is an exemplar of the FAIR (Findability, Accessibility, Interoperability, and Reusability) and FACT (FAIRness, Accuracy, Confidentiality, and Transparency) principles underpinning of responsible data stewardship. It is a vanguard in the open-access data movement. Beginning with only seven protein structures, PDB holdings have grown more than 30,000-fold.

Now in its 53rd year of continuous operations, the PDB has global reach. Approximately 60,000 structural biologists (Depositors) working on every inhabited continent have contributed data to the archive. This information is used by many millions of PDB Data Consumers (basic and applied researchers, trainees, educators, and students) based in >200 UN-recognized sovereign countries and territories. The archive has been designated by the Global Biodata Coalition as a Global Core Biodata Resource, and CoreTrustSeal-certified.

The PDB is managed by an international collaboration among data centers, called the Worldwide Protein Data Bank (wwPDB). RCSB PDB is the global archive keeper and the US data center in this collaboration. The estimated replacement cost of current archival contents exceeds US$20 billion.

PDB data are critical to public health. Structural biologists and PDB data contributed to design and rapid Emergency Use Authorization of two highly-effective mRNA vaccines against SARS-CoV-2, and to discovery and development of Pfizer's antiviral Paxlovid, saving many millions of lives worldwide. RCSB PDB developed many images, publications, videos and curricula to help explain how the virus works that were widely accessed during the pandemic.

It is undeniable that the recently-developed Artificial Intelligence (AI) and Machine Learning (ML)-based software tools for predicting protein structures from amino acid sequence information alone (computed structure models, CSMs) would not exist but for open access to PDB data. In 2022, RCSB PDB further solidified its position as the premier one-stop shop for studying 3D structures of biomolecules by enhancing its research-focused web portal (**RCSB.org**) to support parallel delivery of more than >1,000,000 CSMs of proteins alongside >200,000 PDB structures, enabling basic and applied research across the sciences.

Between 2019 and 2023, more than 67,000 atomic coordinate files and related experimental data files were publicly released. During the same five-year period, PDB data files were accessed more than 10.7 billion times from wwPDB data centers around the world. During 2023 alone, PDB data were accessed from RCSB PDB more than 2.6 billion times. Additionally, 476 trusted external information resources repackage and distribute PDB data for the global scientific community. PDB data are also maintained as standalone copies of the archive inside for-profit company firewalls.

RCSB.org supports an international community of users, including biologists (in fields such as structural biology, biochemistry, molecular and cellular

biology, genetics, pharmacology); other scientists (in fields such as bioinformatics, software developers for data analysis and visualization); students and educators (all levels); media writers, illustrators, textbook authors; and the general public.

Users access tools for searching, reporting, and visualization are integrated with a database that contains data from the PDB archive, data and links from external resources, and pre-calculated data. Searches range from quick queries (author name, ID) to asking more complicated biological questions. For 2023, RCSB PDB Usage Analytics recorded RCSB.org access by ~8.2 million unique IP addresses making ~3.5 billion requests/interactions (involving delivery of >622 TB of data).

RCSB PDB also supports US research by empowering a diverse STEM workforce through Training activities, advancing the frontiers of basic research and discovery through strategic initiatives (e.g., incorporating CSMs at RCSB.org), and enabling applied research in academe and industry with data and tools for exploring new technologies (e.g., protein engineering, de novo protein design, molecular nanotechnology). Training, Outreach, and Education materials are available on the dedicated website PDB-101 ("101", as in an entry-level course, **PDB101.RCSB.org**).

All feedback on RCSB PDB activities is welcome at RCSB.org and at **info@rcsb.org**.

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

RCSB PDB SERVICES OVERVIEW

SERVICE 0

IT INFRASTRUCTURE

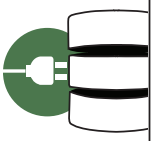
IT Infrastructure supports all RCSB PDB Services and systems by establishing policies and processes to ensure standardized systems configurations and management, redundancy, security, high availability, and disaster recovery.

SERVICE 1

DATA DEPOSITION AND BIOCURATION

RCSB PDB and other members of the Worldwide PDB support tens of thousands of individual data depositors around the globe, ensuring completeness and high 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, education, and training, made freely available on **PDB101.RCSB.org**.

IMPACT OF THE DATA AND SERVICES

PDB Data

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

The PDB Archive

- Grows at the rate of nearly 10% per year
- Used to download ~5 million 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 biotechnology company formation
- PDB-101 materials illustrate how PDB data help explain fundamental biology, biomedicine, energy sciences, and bioengineering/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 **many thousand 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.

SERVICE 1
2023 HIGHLIGHTS

17,063
structures deposited and processed

3,623
new ligands created in the Chemical Component Dictionary

30
new Biologically Interesting Molecule Reference Dictionary items created

The Worldwide Protein Data Bank (wwPDB) was established in 2003 to manage the single PDB archive of macromolecular structural data and make it freely and publicly available to the global community. It consists of organizations that act as deposition, data processing, and distribution centers for PDB data.

PDB structures contain

- 3D atomic coordinates
- Experimental data
- Mandatory metadata
- Authors (e.g., PI contact information)
- Primary citation
- Sample preparation, data collection, and structure determination details
- Polymer sequence(s) (proteins, DNA, RNA, oligosaccharides)
- Chemical 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 include more than 100 academic and industrial volunteers who develop data standards, 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 independent use by depositors prior to data submission, and reports are provided during biocuration.

wwPDB Validation Reports made available once a structure deposition is complete 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.

In 2023, 17,063 new structures were deposited and processed. 3,623 new small molecule ligands and 30 new Biologically Interesting Molecule Reference dictionary items were created. RCSB PDB is responsible for managing depositions from the Americas and Oceania. Of the 6,697 US structures released in 2023, ~69% reported NIH funding; ~11% reported NSF funding. ~89% of X-ray structures released in 2023 from US labs utilized DOE-supported synchrotron or XFEL facilities for data collection.

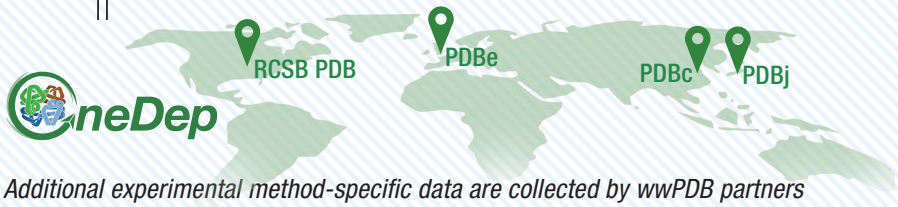
Open Access to SARS-CoV-2 structures remains a high priority; 1,053 SARS-CoV-2 structures were deposited, rigorously validated, expertly biocurated, and released in 2023; 3,913 were available at the start of 2024.

wwPDB regularly standardizes and updates PDB data across the archive. In 2023, updates included improved collection of starting models, standardized sequence annotations for expression tags, cross-referencing XFEL structures with diffraction images, translating atomic coordinates to the standard crystal frame, and standardizing restraint data formatting for NMR structures. Additionally, peptide residue chemical components were updated with standardized atom nomenclature and added backbone annotation.

OneDep

wwPDB develops and uses the global OneDep system for data deposition, validation, and biocuration. This single system helps to maintain high data quality and completeness within the PDB archive, while supporting growth in the number, size and complexity of the deposited structures.

OneDep software is regularly updated to facilitate data submission. In 2023, deposition contact authors became able to utilize their ORCiDs to access a summary table displaying all corresponding entries.



Additional experimental method-specific data are collected by wwPDB partners BioMagResBank (BMRB) and Electron Microscopy Data Bank (EMDB).

ARCHIVE MANAGEMENT AND ACCESS

Supporting PDB Data Consumers by maintaining the PDB archive, developing and standardizing the data dictionary, enabling global data delivery and DOI registration, and integrating PDB data with other trusted information.

SERVICE 2
2023 HIGHLIGHTS

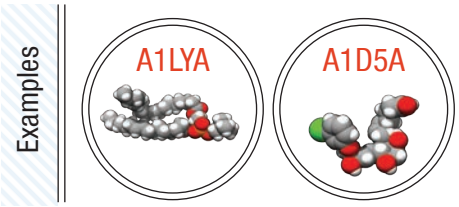
14,517
new structures released into the PDB archive

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

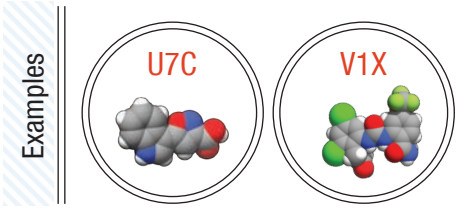
On January 10, 2023, the PDB archive surpassed the milestone of 200,000 structures. Throughout the year, a record 14,517 new PDB structures were released, leading to year-end total holdings of 214,121 available PDB entries.

Another milestone was reached this year. Historically, identifiers for small Chemical Components like ligands have utilized three characters. However, all available three-character Chemical Component IDs have been exhausted, leading wwPDB to introduce five-character alphanumeric accession codes for new components.

Late 2023 5 Digits CCD IDs



Before 2023 3 Digits CCD IDs



In its role as wwPDB-designated archive keeper, the RCSB PDB is responsible for safeguarding and maintaining the archive. RCSB PDB coordinates weekly updates of the PDB archive with other wwPDB Data Centers in Europe and Asia.

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, from new experimental techniques to Computed Structure Models.

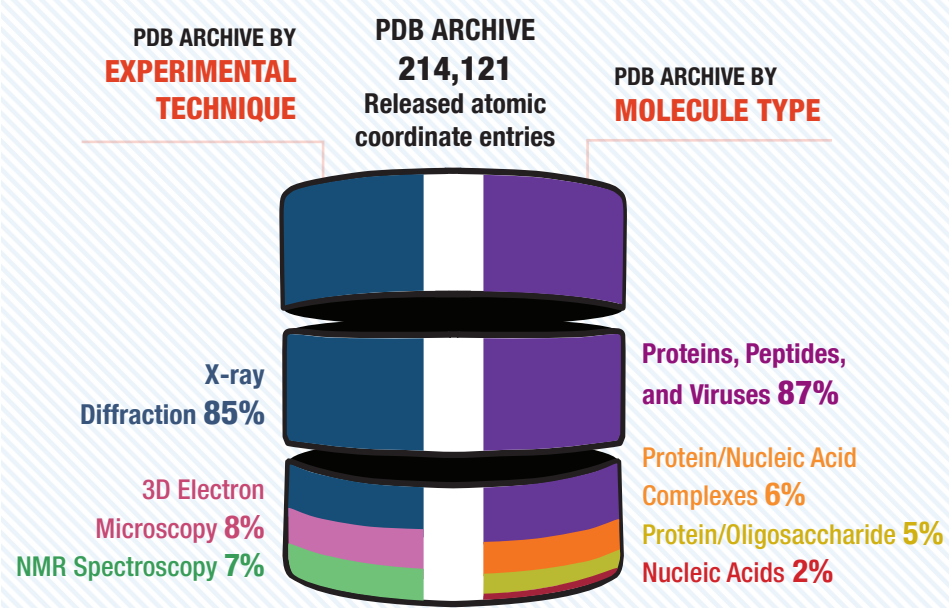
PDBx/mmCIF also enabled development of a "Next Generation" PDB Archive (NextGen, **files-nextgen.wwpdb.org**) that offers enriched annotation from external database resources such as UniProt, SCOP2, and Pfam at atom, residue, and chain levels; and intra-molecular connectivity for each residue present in an entry.

To support RCSB.org data delivery, calculations are run weekly to generate clusters of similar sequences and 3D structures to support search and analysis applications. Data are also integrated with 50 trusted external data resources from across the Life Sciences information ecosystem to provide the most complete and up-to-date representation of a PDB structure.

In 2023, 2.6 billion data files in various file formats, including structure files, experimental data files, chemical and molecular reference data files, FASTA sequence 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, giving a wwPDB-wide total of >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 biopharmaceutical company firewalls.

PDB ARCHIVE CONTENTS | December 31, 2023



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, analysis, and download.

SERVICE 3
2023 HIGHLIGHTS

~8.2 million
clients (unique IP addresses)

~63 million
web page views

~3.5 billion
requests/interactions (downloads, service usage, web page content views)

RCSB.org is one of the most heavily used biological data resources worldwide. For 2023, internal analytics reported ~8.2 million unique IP addresses, making ~3.5 billion requests, downloads, and views. During the same period, Google Analytics estimated that RCSB.org hosted ~5.2 million unique users (~20% from USA) viewing ~63 million web pages.

The website supports a broad range of skill levels and interests. In addition to retrieving 3D structure data, RCSB.org users access comparative data, and external annotations from resources such as Comprehensive Antibiotic Resistance Database (CARD) and Pharos, a comprehensive knowledge base for the Druggable Genome. Gene Ontology, InterPro, and Pharos annotations were also made available for the computed structure models (CSMs) in addition to experimentally-determined PDB structures.

RCSB PDB services go well beyond the original structure and scientific publication. Each PDB structure is represented by 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 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 experimentally-determined PDB structures and CSMs.

The 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 services as well as interactive analytical and visualization tools such as the Mol* molecular graphics system, sequence annotations view, and specialized tools that provide redundancy-reduced "groups" view of similar data. The Pairwise Structure Alignment Tool can align one or more protein chains to a reference structure in a pairwise manner, for simultaneous analysis and visualization of 3D structure alignments and structure-based 1D sequence alignments.

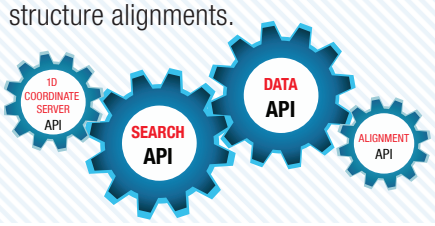
3D structure similarity and motif searches can be executed based on 3D atomic coordinates uploaded by the user. This allows identification of similarities between uploaded coordinates to experimentally-determined PDB structures and CSMs available from RCSB.org.

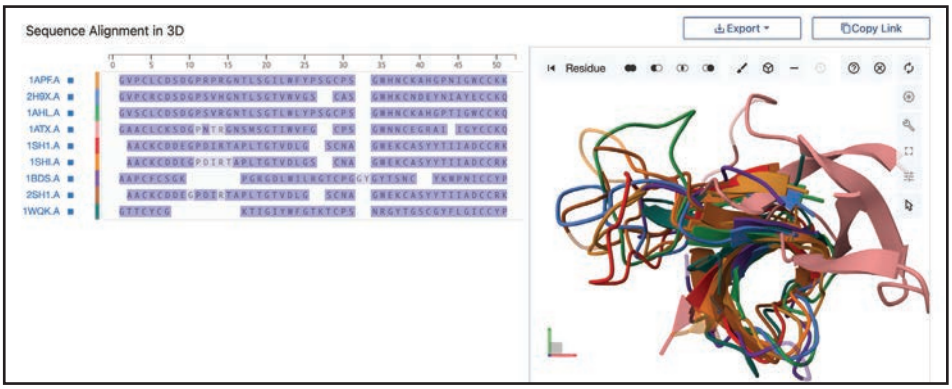
RCSB.org will continue to expand its one-stop shop for studying the 3D structures of biomolecules by providing PDB data consumers with access to CMSs of model organisms, select pathogens, agriculturally-important plants, and organisms crucial in the battle against climate change.

RCSB PDB APIs

RCSB PDB services are accessed via our website (**RCSB.org**) and public API endpoints for programmatic access, with two main APIs: Data (**data.rcsb.org**) and Search (**search.rcsb.org**).

Supporting API services include the 1D Coordinate Server API (**1d-coordinates.rcsb.org**) that serves alignments between structural and sequence databases and integrates protein positional features from multiple resources and Alignment API (**alignment.rcsb.org**) that serves as a comprehensive platform for the seamless computation of structure alignments.





Pairwise alignment tool on RCSB.org showing structural comparison of PDB structures of toxins produced by sea anemones

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 the world of proteins and nucleic acids. The diverse shapes and functions of biological macromolecules help explain aspects of fundamental biology, biomedicine, and bioenergy, from protein synthesis to health and disease to biofuels.

In 2023, ~548,000 users visited PDB-101, viewing >1.8 million pages.

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 courses in 2023 included Leveraging *RCSB PDB APIs for Bioinformatics Analyses and Machine Learning*; *Use PDB Data To Their Full Extent: Understanding PDBx/mmCIF*; and *Python Scripting for Biochemistry & Molecular Biology II*.

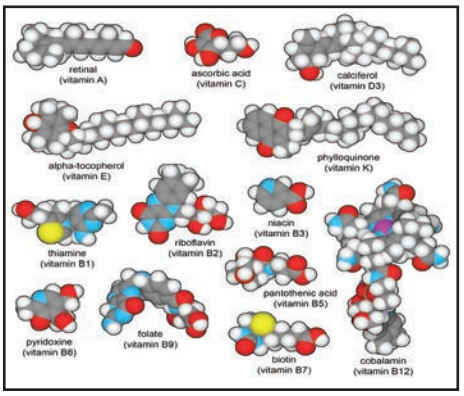
Outreach content, including the *Molecule of the Month* series and molecular animations, demonstrate how PDB data can be used to understand fundamental biology, biomedicine, bioengineering/ biotechnology, and energy sciences in 3D by a diverse and multidisciplinary user community.

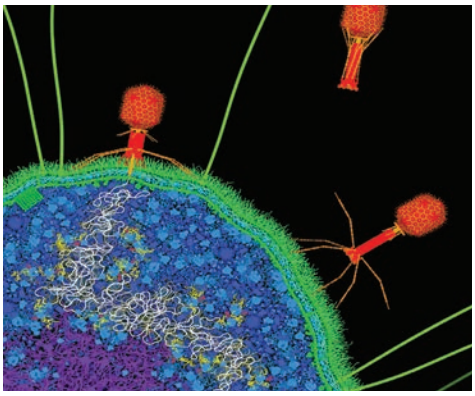
Education Materials, such as Curriculum Modules, provide lessons and activities for teaching and learning.

Other PDB-101 content released this year included a 2024 calendar highlighting the *Structural Biology of Peak Performance*; new features *Exploring the Structural Biology of Bioenergy*; *Exploring the Structural Biology of Viruses*; and *Exploring the Structural Biology of Health and Nutrition*; an introduction to APIs; and Global Health articles highlighting diabetes drugs used for blood glucose management, weight loss, and other topics.

PDB-101 molecular animation videos, webinar recordings, and tutorials are hosted on YouTube. The channel (RCSBProteinDataBank) has ~85,000 subscribers; videos were viewed 858,703 times in 2023.

All training resources are accessible through the PDB-101 website from the new "Train" section in the top menu.





Illustrations from "Exploring the Structural Biology of Health and Nutrition" and "Exploring the Structural Biology of Viruses".

SERVICE 4
2023 HIGHLIGHTS

~548,000
unique users

>1.8 million
web page views

>850,000
video views including molecular animations, training materials, and tutorials

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.

ABOUT THE COVER

Acknowledgement: Illustration by David S. Goodsell, RCSB Protein Data Bank and Scripps Research.
doi: 10.2210/rcsb_pdb/goodsell-gallery-048

Snapshots from the life cycle of bacteriophage T4.

At left, a bacteriophage (red) is injecting its DNA genome (white) into an *Escherichia coli* cell. At center, the bacteriophage has taken over the cell, destroying the cellular DNA (purple) and forcing the cell to make many new copies of bacterial virus. At right, the bacteriophage produces a channel-forming protein that pierces the inner cell membrane, allowing lysozyme enzymes to break down the peptidoglycan sheath that supports the cell. The cell bursts, releasing several hundreds of new bacteriophages.

Many structures of parts of bacteriophage T4 are available in the PDB archive, including the head that holds the DNA in PDB ID 7vs5 (mature form) and 7vrt (immature form), and the baseplate with the injection machinery in PDB ID 5iv5.

FOLLOW US

/RCSBPDB /RCSBProteinDataBank /buildmodels /rcsb

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

FUNDING

RCSB PDB 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.

CITE RCSB PDB

The Protein Data Bank (2000) *Nucleic Acids Research* **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**

RCSB Protein Data Bank: Tools for visualizing and understanding biological macromolecules in 3D (2022) *Protein Science* **31**: e4482
doi: **10.1002/pro.4482**

RCSB Protein Data Bank: Efficient Searching and Simultaneous Access to One Million Computed Structure Models Alongside the PDB Structures Enabled by Architectural Advances (2023)
Journal of Molecular Biology **435**: 167994
doi: **10.1016/j.jmb.2023.167994**

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.

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.