Celebrating PROTEIN DATA BANK 50 Years

2020 ANNUAL REPORT

RCSB PDB
PROTEIN DATA BANK

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

TABLE OF CONTENTS

DIRECTOR'S MESSAGE

The year 2020 started off well for the RCSB Protein Data Bank (PDB). In January, team members from UCSD/SDSC and UCSF visited members at the Rutgers campus for an in-person retreat and a week-long Cloud Technologies Best Practices Boot Camp, co-organized with the NIH STRIDES Program, Google Cloud Platform, and Amazon Web Services.

On February 5th, the first SARS-CoV-2 coronavirus structure (6lu7) was made available in record time. Shortly thereafter as the COVID-19 pandemic worsened, RCSB PDB and our Worldwide Protein Data Bank (wwPDB) partners moved to remote work, prioritizing consistent, expeditious management and distribution of SARS-COV-2 structures. A total of 795 SARS-CoV-2 structures were deposited, validated, expertly biocurated, and made publicly available during the calendar year.

The RCSB PDB and PDB archive have thrived despite all of the challenges created by the new coronavirus. A record-breaking 15,436 new structures were deposited to the archive, of which 7,190 were processed by RCSB PDB – another record. And we broke yet another record in 2020. A total of 14,044 new structures were publicly released into the PDB, revealing more of the natural world in 3D at the atomic level.

As the 3D shape of the coronavirus reached iconic status, RCSB PDB developed various COVID-19 materials for the public. Described in this report, a series of articles, watercolor paintings, short videos, and other resources were released throughout the year, shifting in subject focus and scale as more became known about the virus and managing the pandemic.

During this same period, we launched a newly-designed search engine for discovering and understanding PDB data. Significantly revised and updated tools were made available, ranging from offering Boolean search options, to expanded display options for PDB structures, to powerful GraphQL and REST APIs for data access and searching. We expect that these new services will drive research and education across the sciences from agriculture to zoology for years to come.

In addition to developing the next generation of PDB users, 2021 marks a significant milestone in the history of the archive. Fifty years ago in 1971, the structural biology community established the single worldwide archive for

macromolecular structure data–the Protein Data Bank. From its inception, the PDB has embraced a culture of open access, leading to its widespread use by the research community. PDB data are freely accessed by millions of users exploring fundamental biology, energy, and biomedicine working in every sovereign country and territory recognized by the United Nations. These same data are also repackaged and redistributed by hundreds of external data resources.

Structural biology and structural bioinformatics have had an enormous impact on our understanding of the mechanism of action and biochemical function of biological macromolecules. The PDB archive safeguards these data, representing many of the achievements and milestones of structural biologists worldwide.

Together with our wwPDB partners, we will be celebrating the 50th anniversary of the PDB archive during the year with symposia, materials, and more (**rcsb.org/pdb50** and **wwpdb.org/pdb50**). All main symposia will be virtual, with additional events to be announced. We hope you can join one or more of our celebrations wherever you live and work.

Like all of you, the RCSB PDB team looks forward to a gradual return to work and a more normal life made possible by large-scale vaccination over the coming months. Please be sure to continue taking recommended public health precautions, so that when we do come together no one is missing.

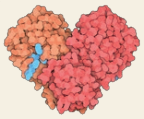
Sincerely,

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
Adjunct Professor, University of California San Diego

PDB HISTORY HIGHLIGHTS

2021 PDB50: Golden Anniversary of the PDB. *See page 6 for more details.*

2020 First SARS-CoV-2 structure in the PDB (6lu7). *See pages 4-5 for more details.*



2014 **100,000 Structures** in the archive

OneDep: New system for Deposition-Validation-Biocuration

2010 Worldwide Protein Data Bank Foundation established

2008 **50,000 Structures** in the archive

2003 wwPDB established by RCSB PDB, PDBe, PDBj. *See page 8 for more details.*

1999 **10,000 Structures** in the archive

1991 First 3DEM structure released in the PDB (1brd)

1989 First NMR structure released in the PDB (1bds)

1971 PDB established at Brookhaven National Laboratory and the Cambridge Crystallographic Data Center. *See page 6 for more details.*

To explore PDB History in more detail, visit **rcsb.org** *> More > PDB History*

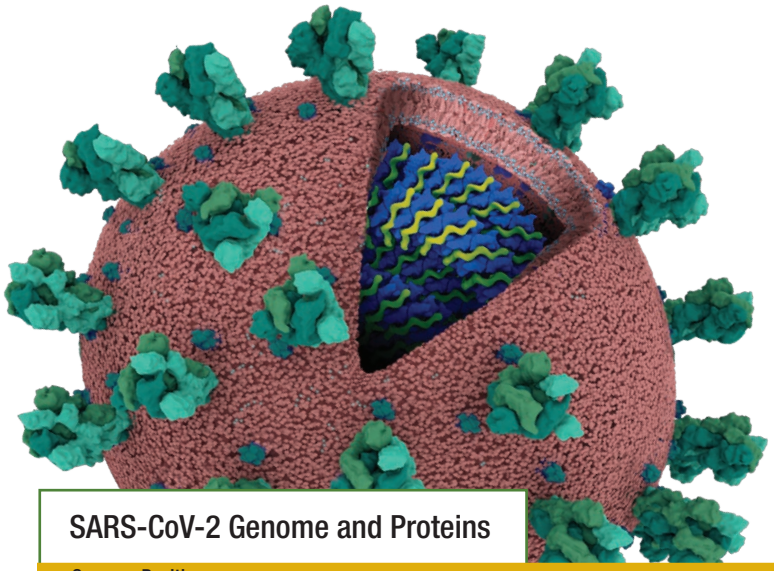
RCSB PDB team in September 2020

RCSB PDB IN THE TIME OF COVID-19

RCSB PDB is a central resource in the global fight against SARS-CoV-2, the agent of the COVID-19 pandemic. Open access to PDB data has facilitated design of effective vaccines, and provides the basic understanding and 3D structures needed to discover new therapies.

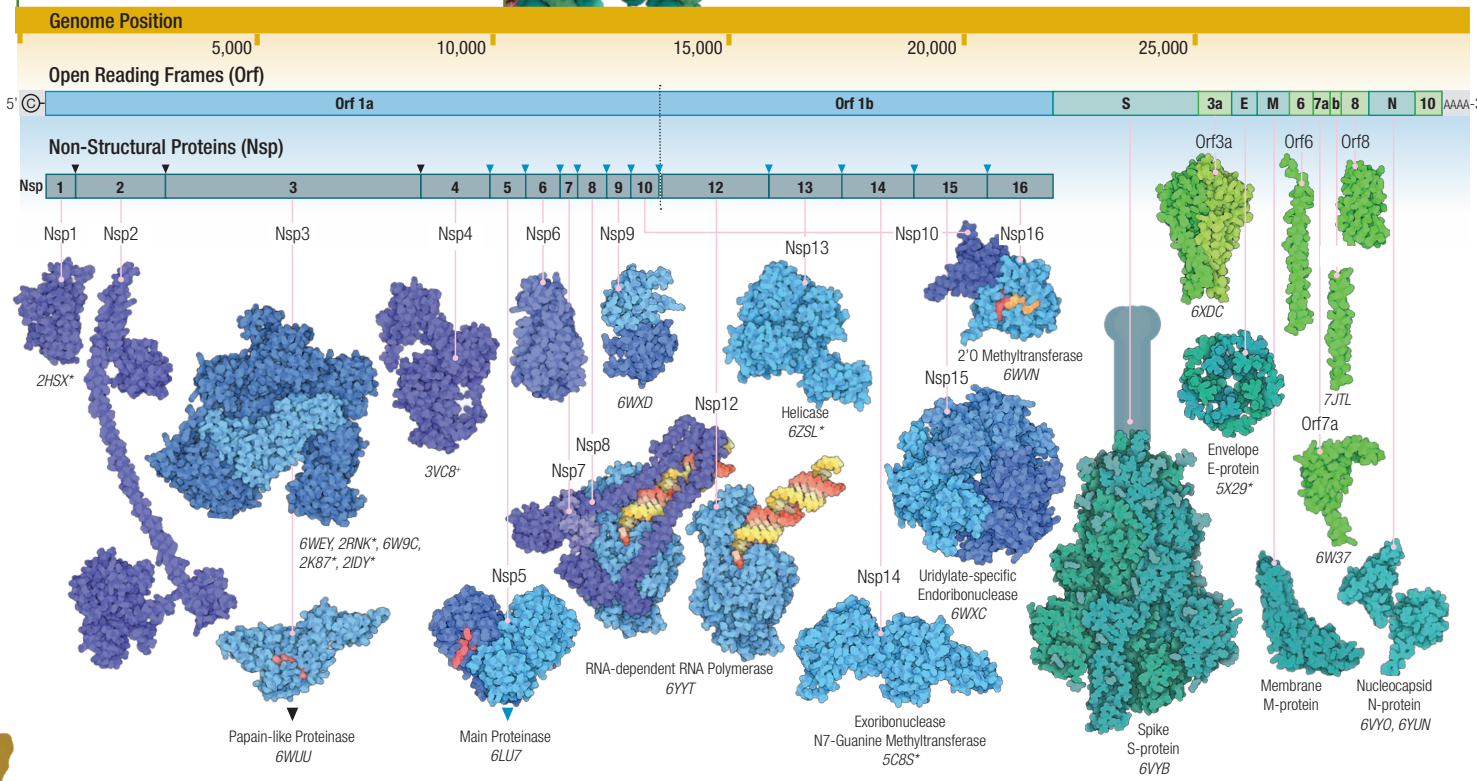
In addition, PDB structures are central to visualizing and understanding new variants of the virus emerging in different geographies. The impact of structural biologists on research related to COVID-19 is a testament to the power of the experimental tools they use and their commitment to making data open access for the public good.

In addition to expert biocuration of these structures, RCSB PDB offers many resources for access and analysis of these structures, and general resources to help understand the mechanisms of the viral life cycle and how they impact human health.



NJTV News spoke with biocurator Brian Hudson to learn more about fighting COVID-19.

SARS-CoV-2 Genome and Proteins



© Denotes N7-methyl Guanine Cap
+ Denotes Mouse Hepatitis Virus Structures
* Denotes SARS-Cov PDB Structures
AAAA Denotes Poly-A Tail

Architecture of the SARS-CoV-2 genome and proteome from ***Evolution of the SARS-CoV-2 proteome in three dimensions (3D) during the first six months of the COVID-19 pandemic****,* bioRxiv 2020.12.01.406637*; doi: 10.1101/2020.12.01.406637*

BIOCURATION OF SARS-COV-2 STRUCTURES

The first SARS-CoV-2 structure, main protease (PDB ID 6lu7), was deposited on January 26, 2020 and immediately biocurated so that it could be released with the next weekly update of the PDB.

Since then, more than 1,000 COVID-19-related structures have been deposited to the PDB archive. Each new structure has been efficiently reviewed and expertly annotated by wwPDB biocurators following these "guiding principles":

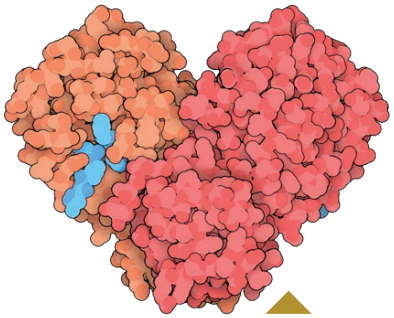
- Biocuration of COVID-19 structures is prioritized over that of other structures, including post-release revisions such as citation updates
- Authors are encouraged to release their structures immediately
- Consistent taxonomy name and ID (Severe acute respiratory syndrome coronavirus 2; 2697049) are applied to all COVID-19 structures
- Consistent UniProt referencing is incorporated: P0DTD1, P0DTC1, P0DTC2, P0DTC9

DATA ACCESS AND EXPLORATION

As a data archive, the PDB contains valuable clues in the form of protein structures from other coronaviruses. Insights from these structures could be vital in furthering our understanding of coronaviruses and in discovery and development of new treatments and vaccines to fight the current pandemic and to prepare for the next outbreak.

RCSB.org offers tools and services to search, analyze, and visualize PDB structures. These tools can be used to search and analyze SARS-CoV-2 structures in the context of other structures in the PDB, including those from SARS-CoV-1 and MERS.

All released SARS-CoV-2 structures and related resources are highlighted at RCSB.org/covid19.



OUTREACH AND EDUCATION

Early in the course of the pandemic, we aimed to develop outreach materials that explained the virus as quickly as new coronaviral structures were added to PDB. Coinciding with the release of the first SARS-CoV-2 structure (6lu7) on February 6, 2020 were publication of the February *Molecule of the Month* on the Coronavirus Main Protease and David S. Goodsell's watercolor *Coronavirus*.

All SARS-CoV-2 outreach materials can be accessed at the Coronavirus Collection at PDB-101 (pdb101.rcsb.org/browse/coronavirus). This body of work includes the video *Fighting Coronavirus with Soap*, which has been viewed nearly 500,000 times and was an Official Selection of the 2020 American Public Health Association Film Festival. Curricular materials were developed to help students explore COVID-19 in Molecular Detail and for a Boot Camp on *COVID-19 Evolution and Structural Biology*. This week-long summer experience for undergraduates kicked off a longer study of sequence and structure changes during the first six months of the pandemic.

The first COVID-19 Molecule of the Month *article was viewed a record-breaking 85,000 times in 2020. (For comparison, perennial favorites like hemoglobin and catalase were viewed more than 38,000 and 23,000 times, respectively, during the same time period). The popularity of this SARS-CoV-2 feature inspired future articles on the RNA-dependent polymerase and the spike protein. As more became known about how the biological macromolecules in the PDB were driving research and discovery, a new series of articles about* ***Resources to Fight the COVID-19 Pandemic*** *was launched.*

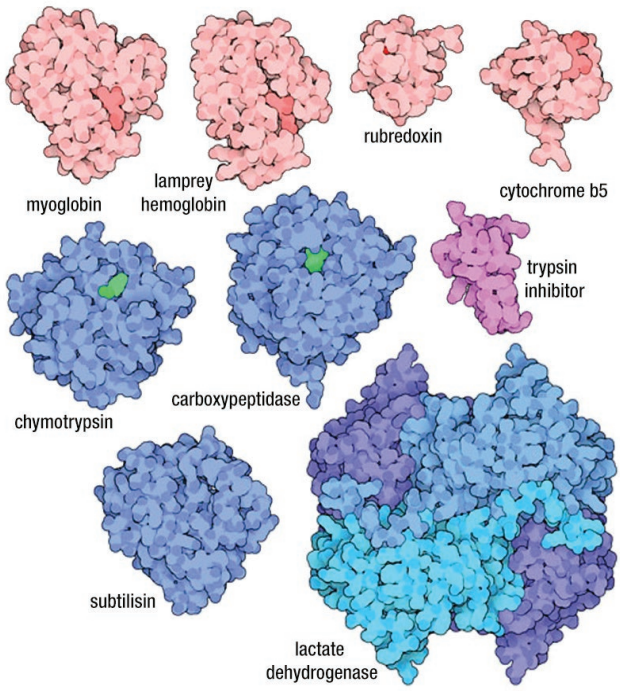
David Goodsell's first coronavirus painting depicts the virus just entering the lungs, surrounded by mucus secreted by respiratory cells, antibodies, and several small immune system proteins.

It was accessed more than 48,000 times from PDB-101, and circulated around the world. It was discussed in Forbes, Science, The Washington Post, The Paris Review, etc. *A black and white version was created for coloring, and that too generated much interest among young and old alike.*

Additional five paintings highlighting the molecular landscape of SARS-CoV-2 were released into the David Goodsell SciArt gallery, and enjoyed similar reuse worldwide.

CELEBRATING 50 YEARS OF THE PROTEIN DATA BANK ARCHIVE

In 1971, the structural biology community established the single global archive for macromolecular structure data–the Protein Data Bank (PDB). PDB was the 1st open access digital data resource in all of biology and medicine.



Protein Data Bank in 1973
To learn more about the historic structures that set the foundation for the PDB archive, read the October 2011 Molecule of the Month *column* PDB Pioneers

Growing from an initial seven structures to more than 175,000 today, the PDB archive is a leading global resource for experimental data central to scientific discovery. PDB data are repackaged and redistributed by hundreds of data resources and millions of users exploring fundamental biology, energy, and biomedicine.

NUMBER OF STRUCTURES RELEASED INTO THE PDB ARCHIVE BY DECADE

1971-80	1981-90	1991-2000	2001-10	2011-20
69 Structures	**438** Structures	**13,082** Structures	**55,917** Structures	**103,487** Structures

173, 005 Structures (as of Dec. 31, 2020)

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

In 2021, RCSB PDB and the wwPDB celebrate the 50th Anniversary of the PDB with symposia, materials, and more.

PDB50 CELEBRATORY SYMPOSIA

May 4-5, 2021

INAUGURAL (AND VIRTUAL) EVENT:
A special symposium celebrating the 50th anniversary of the Protein Data Bank

July 30-31, 2021

TRANSACTIONS SYMPOSIUM 2021:
Function Follows Form: Celebrating the 50th Anniversary of the Protein Data Bank.

This virtual event is part of the Annual Meeting of the American Crystallographic Association.

October 20-22, 2021

EMBL CONFERENCE:
Bringing Molecular Structure to Life: 50 Years of the PDB (virtual)

Additional meetings will be announced.

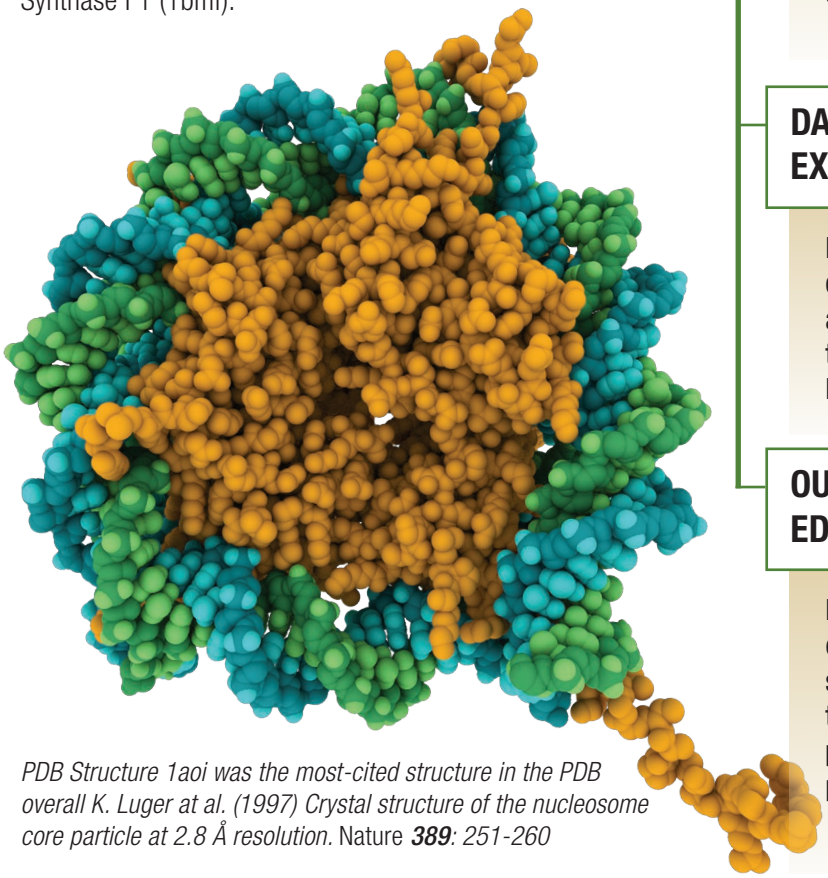
RCSB PDB AND 2020 AT A GLANCE

USERS AND IMPACT

RCSB PDB supports an international community of millions of users, including biologists (in fields such as structural biology, biochemistry, genetics, pharmacology); other research 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. The inaugural RCSB PDB publication (Berman *et al.*, *Nucleic Acids Research* 2000) is one of the top-cited scientific reports of all time.

RCSB PDB regularly explores different ways of measuring impact. A 2020 publication identified highly-cited PDB structures and analyzed those cited in Materials Science, Physics, Computer Science, Chemistry, Engineering, and Mathematics journals to explore how PDB data are used across scientific disciplines (Impact of the Protein Data Bank Across Scientific Disciplines (2020) *Data Science Journal* **19**: 25 doi: 10.5334/dsj-2020-025).

The nucleosome structure PDB ID 1aoi was the most-cited structure in the PDB overall, followed by the potassium channel (1bl8), bacteriorhodopsin (1brd), rhodopsin (1f88), major histocompatibility class I (1hla), MDM2/imidazoline inhibitor (1rv1), thrombin (2v3o/2v3h), serum albumin (1uor), and ATP Synthase F1 (1bmf).



PDB Structure 1aoi was the most-cited structure in the PDB overall K. Luger at al. (1997) Crystal structure of the nucleosome core particle at 2.8 Å resolution. Nature ***389****: 251-260*

RCSB PDB SERVICES

DEPOSITION AND BIOCURATION

RCSB PDB and other members of the Worldwide PDB support >50,000 depositors worldwide ensuring quality for the ever growing body of data.

In 2020, **15,436** new structures were deposited to the archive

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.

In 2020, **14,044** new structures were added to the archive

DATA EXPLORATION

RCSB PDB develops tools for data searching, visualization and analysis, and makes them freely available on RCSB.org

In 2020, RCSB PDB released the **NEXT GENERATION** search tools

OUTREACH AND EDUCATION

RCSB PDB develops educational resources about structural biology and makes them freely available on pdb101.rcsb.org. It also provides ongoing user support.

In 2020, PDB-101 was visited **>2.6 million times**

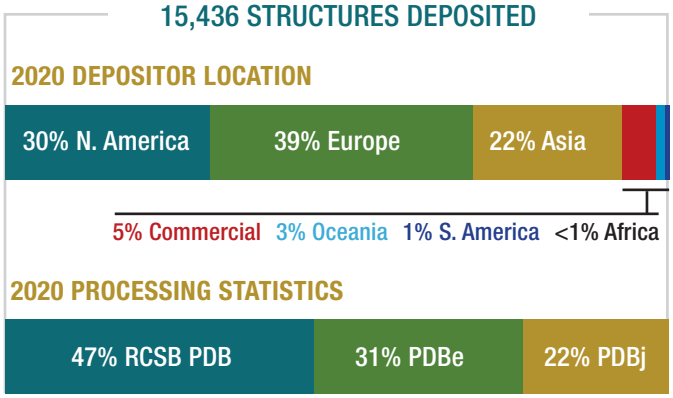
RCSB PDB SERVICES: DEPOSITION/BIOCURATION

The Worldwide Protein Data Bank (wwPDB) was established to manage the single PDB archive of macromolecular structural data that is 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 2020, 15,436 PDB structures were deposited by researchers from around the world and then prepared for public release by wwPDB partners RCSB PDB, PDBe, and PDBj. 3146 new ligands and 70 new Biologically Interesting Molecule Reference dictionary items were made available.

Biocuration responsibilities are distributed geographically. As the US Data Center, RCSB PDB biocurates structures contributed by scientists working in the Americas and Oceania. During 2020, RCSB PDB processed 47% of all incoming structures.

2020 DEPOSITION STATISTICS



PDB structures contain 3D atomic coordinates, experimental data, and mandatory metadata (authors, primary citation, sample preparation, data collection, and structure determination, polymer sequence(s) and chemical structures). All deposited data undergo expert review and curation. Each structure is examined for self-consistency, standardized using controlled vocabularies, cross-referenced with other biological data resources, and validated for scientific/technical accuracy.

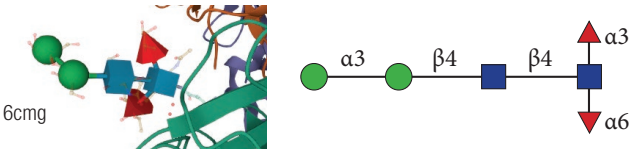
wwPDB Working Groups and Task Forces represent more than 100 academic and industrial volunteers who make recommendations and contribute software tools used to generate wwPDB Validation Reports that assess the quality and accuracy of every structure stored in the PDB archive. These reports can be provided to journal editors and manuscript 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.

IMPROVING CARBOHYDRATES IN THE PDB: The wwPDB regularly reviews how data are represented across the data and coordinates "remediation" efforts that help maintain consistency and accuracy across the PDB archive.

~10% of PDB structures contain carbohydrates. To improve the ability to search for these structures and their complex chemistry (e.g., stereo-isomers, anomeric configurations, branched chains), the wwPDB embarked on a remediation project that

- Standardized Chemical Component Dictionary definitions for monosaccharides
- Adopted carbohydrate nomenclature following IUPAC-IUBMB recommendations
- Provided a uniform representation for oligosaccharides
- Incorporated glycoscience community tools for naming conventions and linear descriptors into processing pipeline
- Annotated glycosylation sites in PDB structures

In July 2020, ~14,000 PDB structures were re-released to to conform to the this new representation. This update of the archive includes diverse oligosaccharides: glycosylation; metabolites such as maltose, sucrose, cellulose fragments; glycosaminoglycans, such as fragments of heparin and heparan sulfate; epitope patterns such as A/B blood group antigens and the H-type or Lewis-type stems; and many artificial carbohydrates mimicking natural products.



RCSB.org users can explore carbohydrates in 3D with Mol. By default they are displayed using the symbolic representation.*

IMPROVING WWPDB VALIDATION REPORTS: Validation is a key part of PDB data deposition. Validation Reports are generated to help depositors identify and evaluate discrepancies in their data. During 2020, branched representation and 2D geometric quality images for ligands and carbohydrates, and 3D electron density fits (determined by macromolecular crystallography) for ligands and carbohydrates, restraints validation for NMR, and map volume analysis for 3DEM were added to these reports.

RCSB PDB SERVICES: ARCHIVE MANAGEMENT/ACCESS

The mission of the 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.

By following the 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.

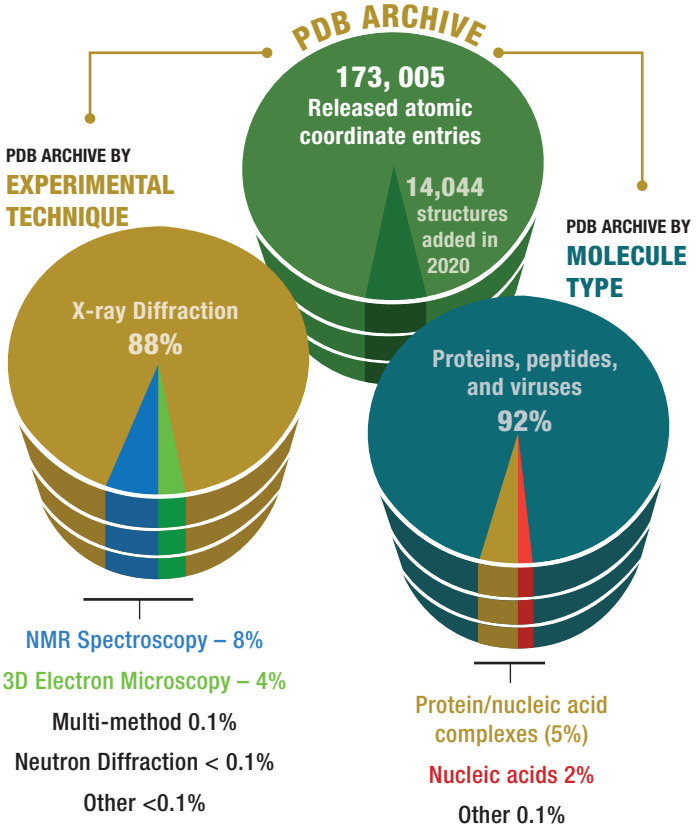
In 2020, 14,044 new PDB structures were added to the archive, for a total of 173,005 available entries as of January 1, 2021. As wwPDB archive keeper, the RCSB PDB is responsible for safeguarding the PDB archive and maintaining the PDB FTP (ftp.wwpdb.org). RCSB PDB coordinates weekly updates of the PDB archive with wwPDB Data Centers in the United Kingdom and Japan.

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 ~40 external data resources from across the Life Sciences ecosystem.

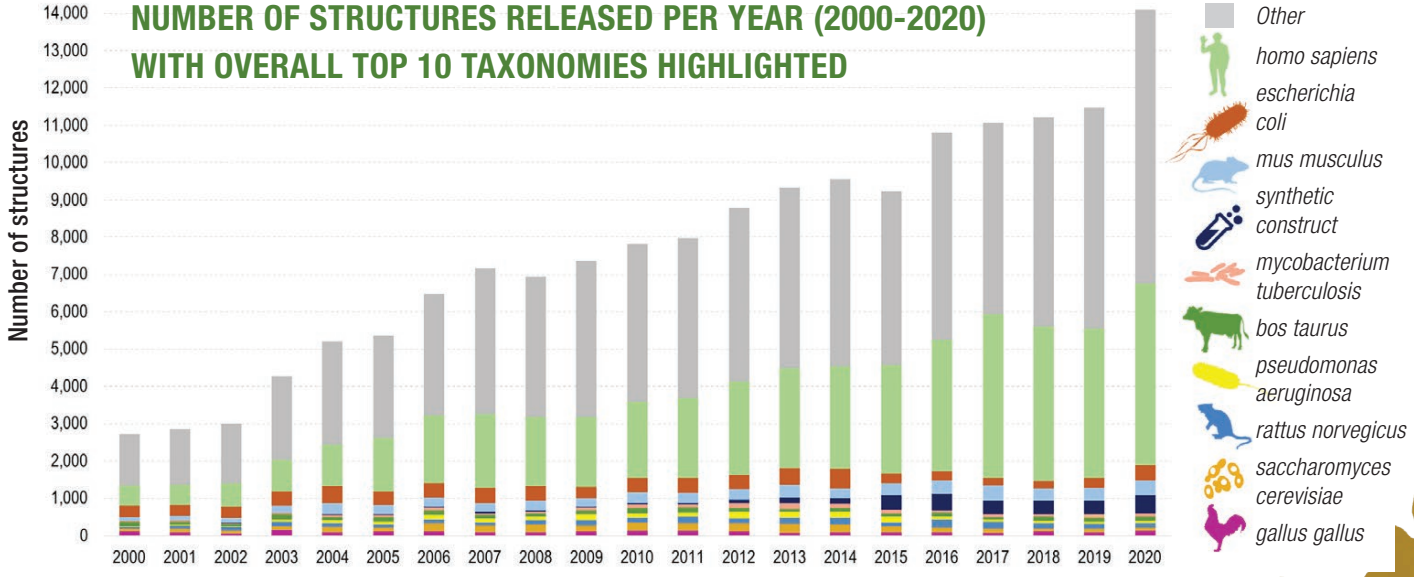
New data access API services have been deployed to support the newly-released RCSB.org website architecture. These services provide efficient access to primary PDB data and annotations, PDB atomic coordinate data, electron density maps, 3DEM maps.

Millions of structure data files were downloaded from all wwPDB partner web and FTP sites.

PDB ARCHIVE CONTENTS (DEC. 31, 2020)



NUMBER OF STRUCTURES RELEASED PER YEAR (2000-2020) WITH OVERALL TOP 10 TAXONOMIES HIGHLIGHTED



RCSB PDB SERVICES: DATA EXPLORATION

The open-access web portal **RCSB.org** supports PDB Data Consumers in the US and around the world with resources for PDB structure access, visualization, and analysis.

The open-access web portal RCSB.org supports PDB Data Consumers in the US and around the world with resources for PDB structure access, visualization, and analysis.

The website supports millions of users representing a broad range of skills 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 provide information going 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 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 structures and related information.

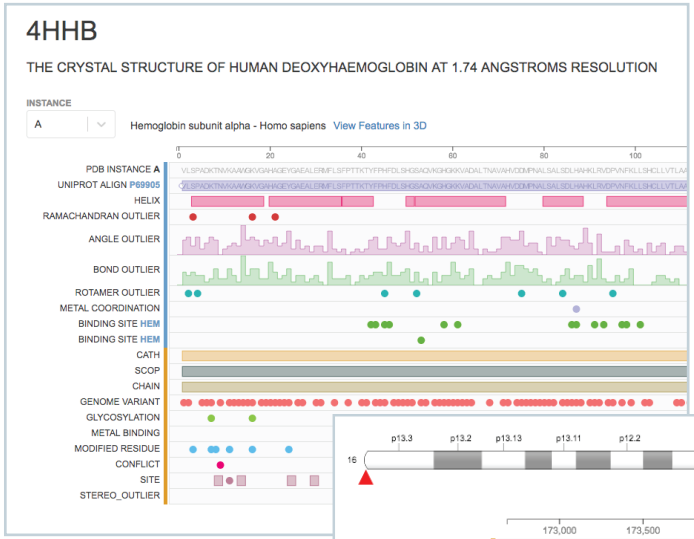
A significant project has been the architectural redesign of the services supporting RCSB.org. This effort includes re-structuring and simplifying website searching and reporting (to leverage existing and new APIs) and adopting a more modern and extensible front-end web framework.

The next generation of RCSB PDB tools was released in 2020, including search functionality that combines text, sequence, and coordinate queries. Other new and improved features include

- Advanced Search Boolean options
- Display of Advanced Search results for PDB structures, assemblies, or unique polymeric molecular entities (unique sequences)
- Left sidebar menus for refining Search Results
- Stored History of searches made during a session
- Improved MyPDB functionality, including registration using Google, Facebook, or ORCID and email notifications of stored searches
- Enhanced GraphQL and REST APIs for accessing and searching PDB data
- Sequence motif search to identify residue patterns that recur throughout the PDB
- Integration of multiple search types in Advanced Search: Structure (3D shape), Sequence Motif, Sequence Similarity, Chemical Search and Annotation Browsing
- Improved Download Tool for batches of data files
- Protein Feature View/Genome View: graphical summaries of correspondences between PDB entity sequences and annotations/reference genomes

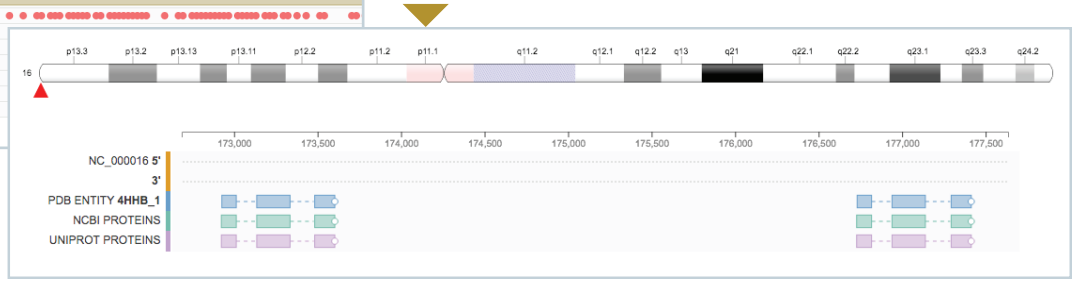
Documentation is provided for all new features. These RCSB PDB tools and resources provide rich structural views of biological systems to enable breakthroughs in scientific inquiry, medicine, drug discovery, technology, and education.

New versions of the graphic browsers Protein Feature View and Genome View were released in 2020.



Protein Feature View *provides graphical summaries of PDB protein biological and structural features and their relationships with UniProtKB entries. Structural features are extracted from the PDB while structural domains are annotated from CATH and SCOPe databases. Biochemical and biomedical features are collected from UniProtKB.*

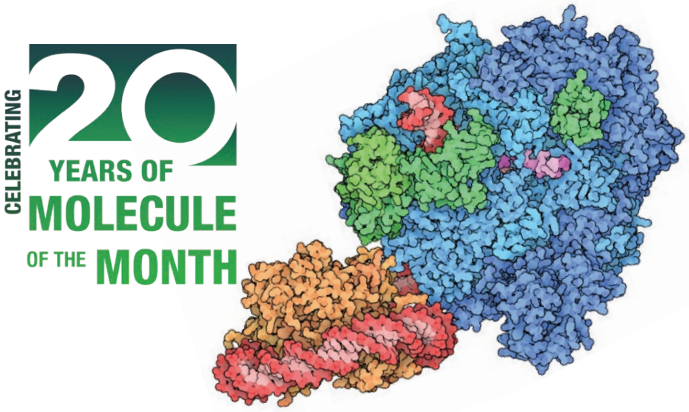
Genome View *provides graphical summaries of the correspondences between PDB entity sequences and genomes. This resource integrates alignments from multiple resources and databases including UniProtKB, RefSeq and SIFTS.*



RCSB PDB SERVICES: OUTREACH AND EDUCATION

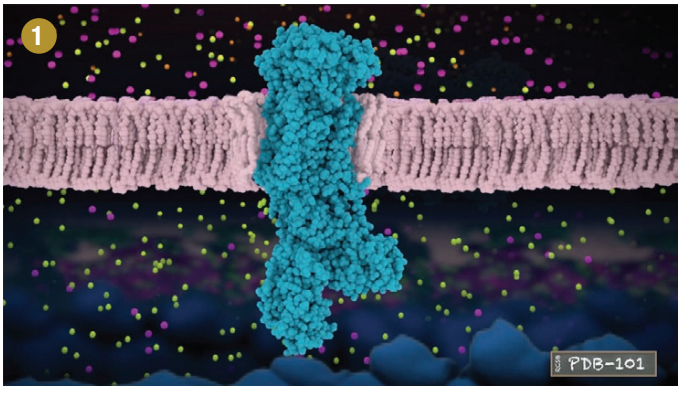
PDB-101 (**pdb101.rcsb.org**) hosts educational materials for a broad audience. In 2020, PDB-101 more than 850,000 users accounted for >2.6 million page views.

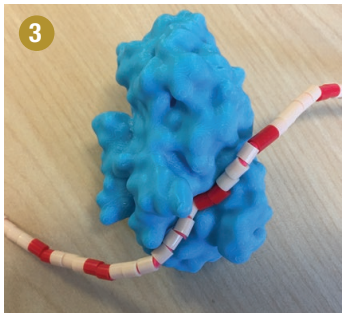
The popular *Molecule of the Month* series tells stories about molecular structure and function, their diverse roles within living cells, and the growing connections between biology and nanotechnology. This content has inspired readers around the world, and is a regular read for students and researchers alike.



2020 marked the 20th anniversary of Molecule of the Month. *The year's inaugural article* Twenty Years of Molecules *explored the structural biology revolution and included the insights from the column creator David Goodsell*

Selected PDB-101 additions in 2020 included a video on *Neuronal Signaling and Sodium-Potassium Pump* (1), a template to create a 3D paper model of Red Fluorescent Protein (2); and a collection of files for 3D printing of key molecules (3).

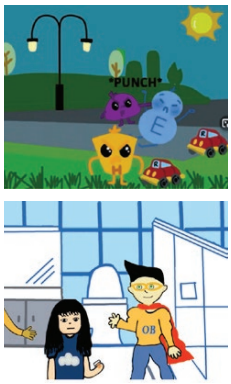




HEALTH FOCUS: DRUGS AND THE BRAIN

PDB-101 adopts a bi-annual public health theme to focus development and student engagement. The theme for 2020-2021 is Drugs and the Brain. The theme is used at Rutgers University for an Undergraduate Honors Seminar Course exploring the foundations of structural biology and for an annual RCSB PDB-sponsored video contest. For this challenge, high school students focused on the molecular mechanisms behind modulation of pain signals by opioids.

Browse PDB-101 materials on Drugs and the Brain at pdb101.rcsb.org/browse/drugs-and-the-brain



The 2020 video challenge entries feature many original story-telling approaches

CUSTOMER SERVICE

Over the past 50 years, correspondence with users has kept pace with the times. Early on, interactions with depositors and users took place by letter and phone. Beginning in the 1990s, correspondence blossomed as email became available and website technology made PDB data more accessible. Now in the 21st century, customer service has evolved to help users worldwide access the many tools available to explore the rich depth of structural information stored in the PDB.

In 2020, the RCSB.org customer service department engaged in over 1,700 email conversations with about 1,400 unique users, who range from students to power users, and those new to the field to subject experts. User input helps drive development through remarks about existing features and suggestions for new ones.

wwPDB depositors correspond with the biocuration team using a dedicated communication module in the OneDep deposition, validation, and biocuration software system and through a dedicated email help desk.

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.

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: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences (2021) *Nucleic Acids Research* **49**: D437–D451
doi: **10.1093/nar/gkaa1038**

FOLLOW US

/RCSBPDB
/RCSBProteinDataBank
/buildmodels
/rcsb

FUNDING

RCSB PDB is 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.