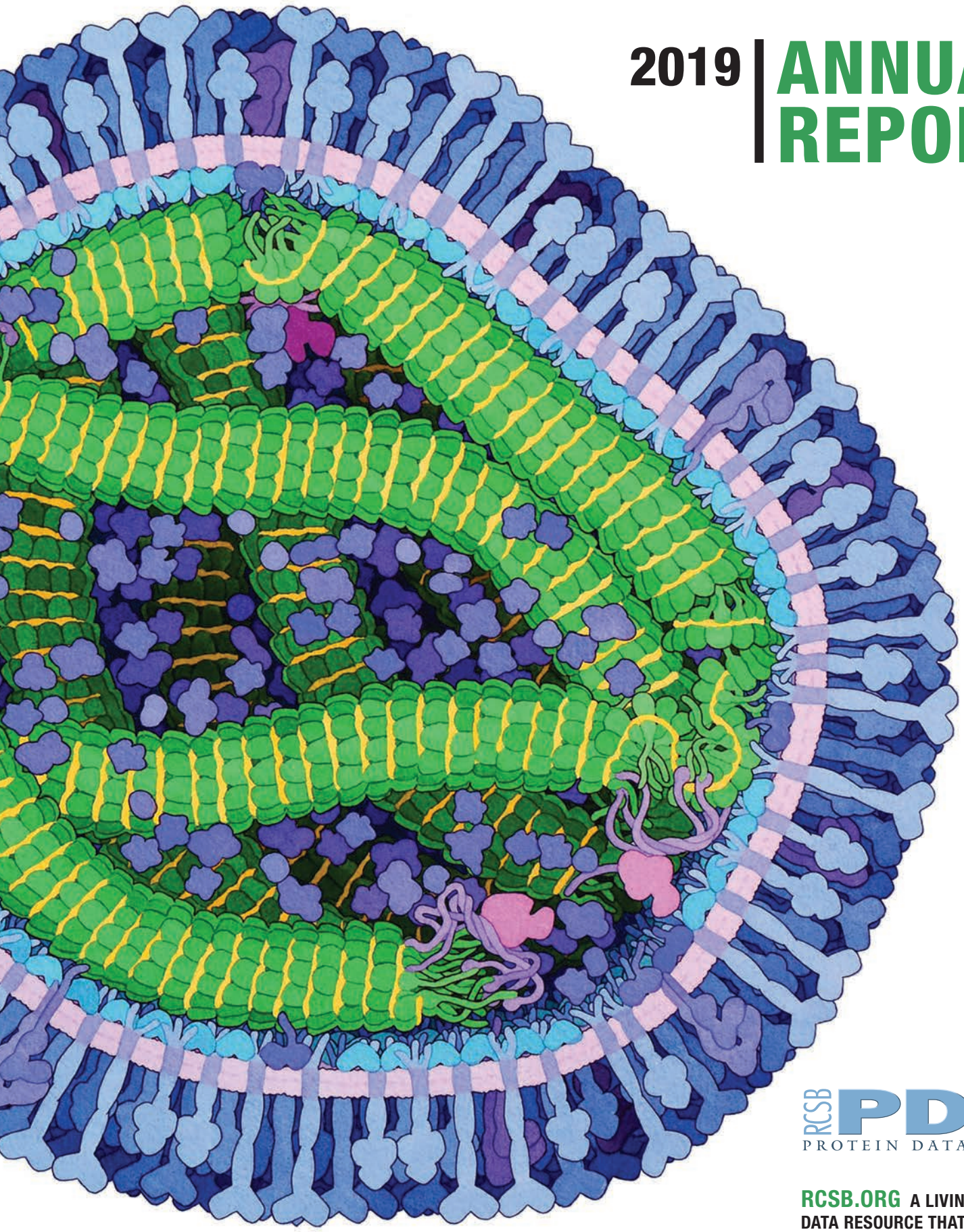


2019 ANNUAL REPORT

RCSB PDB
PROTEIN DATA BANK

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

DIRECTOR'S MESSAGE

As I write this in Spring 2020, the world is facing the COVID-19 pandemic. Our team may be working remotely, but the global research community is actively investigating all avenues for targeting this threat to humanity.

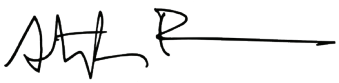
Key to the discovery and development of drugs and vaccines will be structural biology and the SARS-CoV2 structures freely available from the Protein Data Bank archive (PDB). The first COVID-19 coronavirus structure, the main protease determined by Zihe Rao and Haitao Yang's research team at ShanghaiTech University, was made available in record time on February 5, 2020. More than 100 related structures were released in the seven weeks since that initial structure, with many more on the horizon. Rapid public access to these detailed molecular portraits of the main and papain-like proteases, spike protein, RNA polymerase, and others help explain the biological and biochemical mechanisms central to the pathogenicity of the virus.

As a data archive, the PDB contains valuable clues in the structures of the corresponding protein from other coronaviruses. The 2003 outbreak of the closely-related Severe Acute Respiratory Syndrome-related coronavirus (SARS) led to the first 3D structures, and today there are more than 200 PDB structures of SARS proteins. Structural information from these related proteins could be vital in furthering our understanding of coronaviruses and in discovery and development of new treatments and vaccines to contain the current pandemic and manage the next coronavirus outbreak.

PDB-101, the educational arm of the RCSB PDB, has been producing materials to disseminate COVID-19 information beyond the research community. As noted by a follower on Twitter, SARS-CoV2 is not an invisible enemy but rather one that needs special tools to see. These PDB-101 videos, paintings, and other related features help students and educators better see the virus.

Facilitating the archiving of data from these studies that integrate both quantitative measurement and structural analysis enables the PDB data community to make wide-reaching breakthroughs in research and education.

Sincerely,



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

Director, RCSB PDB

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

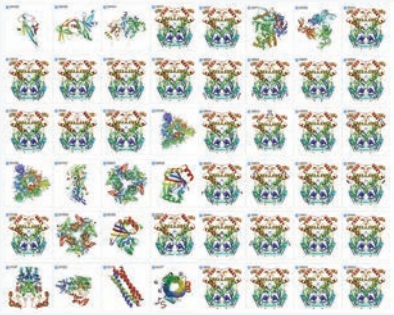
Adjunct Professor, University of California San Diego

RCSB PDB SERVICES & ENABLING THE STRUCTURAL EXPLORATION OF COVID-19

1 DEPOSITION AND BIOCURATION

RCSB PDB and other members of the Worldwide PDB support >40,000 depositors worldwide ensuring quality for the ever growing body of data.

Biocuration of incoming SARS-CoV-2 structures is prioritized by wwPDB annotators.



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.

Weekly data releases provide timely access to the current research on COVID-19, usually before it appears in a scientific journal (with author permission).

3 DATA EXPLORATION

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

Visitors to RCSB.org can instantly access all COVID-19 structures and resources along with exploration tools.

4 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.

Educational resources on COVID-19 facilitate our understanding of the coronavirus structure.

PDB-101 resources include *Coronavirus* painting and a video *Fighting Coronavirus with Soap*.

Individual RCSB PDB services and 2019 milestones are described in greater detail in this report.

1 RCSB PDB SERVICES
DEPOSITION AND BIOCURATION

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

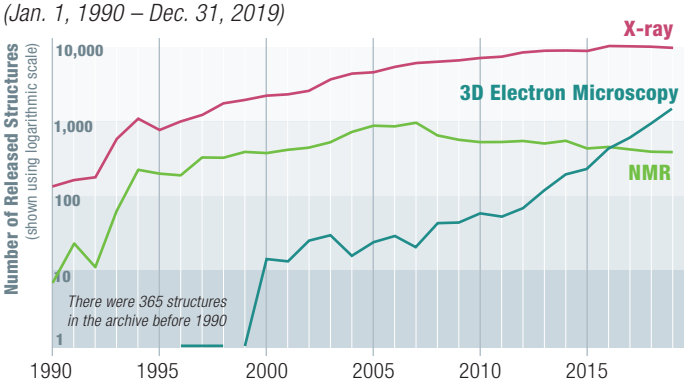
In 2019, 13,377 STRUCTURES WERE DEPOSITED by researchers from around the world (up from 12,179 in 2018).

Biocuration responsibilities are distributed geographically. As the US Data Center, RCSB PDB biocurates structures submitted by scientists working in the Americas and Oceania. During 2019, RCSB PDB processed 41% of all incoming structures (34% PDBe, 25% PDBj).

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.

PDB STRUCTURES AVAILABLE BY EXPERIMENTAL TECHNIQUE

(Jan. 1, 1990 – Dec. 31, 2019)



Ensuring Data Consistency and Quality

Processes and procedures follow data archiving best practices. PDB versioning, which enables depositors to update 3D atomic coordinates post release while retaining their original PDB ID, was enabled in 2019. Each PDB structure is assigned a DOI that resolves to a summary page with links to data files and resources at wwPDB partner sites.

wwPDB Working Groups and Task Forces include 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 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.

New 2019 validation features include 2D diagrams of ligands, highlighting geometric validation criteria and, for structures determined by macromolecular crystallography, 2D views of electron density fit for those designated as “Ligand of Interest”. Validation reports for depositors now incorporate an extensive EM map validation process, integrating validation methods for EM data provided by our EMDB collaborators. These changes should help depositors and users to identify potential errors and provide more clarity about potential limitations of the structural data.

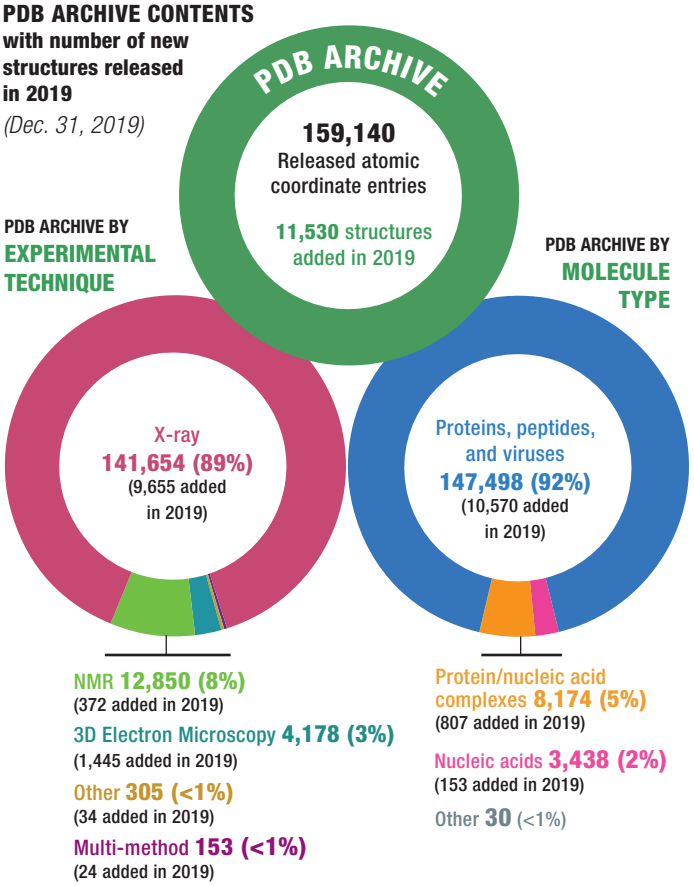
2 RCSB PDB SERVICES
ARCHIVE MANAGEMENT AND ACCESS

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 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.

In 2019, 11,530 NEW STRUCTURES WERE RELEASED to the PDB archive.

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 ARCHIVE CONTENTS
with number of new structures released in 2019
(Dec. 31, 2019)



RELATED EXPERIMENTAL DATA FILES

- **129,252** Structure factors (7,380 added in 2019)
- **10,125** NMR restraints (307 added in 2019)
- **3,877** Chemical shifts (307 added in 2019)
- **3,876** 3DEM map files (1,085 added in 2019)

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 information ecosystem.

More than 838 million structure data files were downloaded from all wwPDB partners, including more than 547 million from RCSB PDB-hosted FTP and websites (up from 500 million in 2018).

3 RCSB PDB SERVICES
DATA EXPLORATION

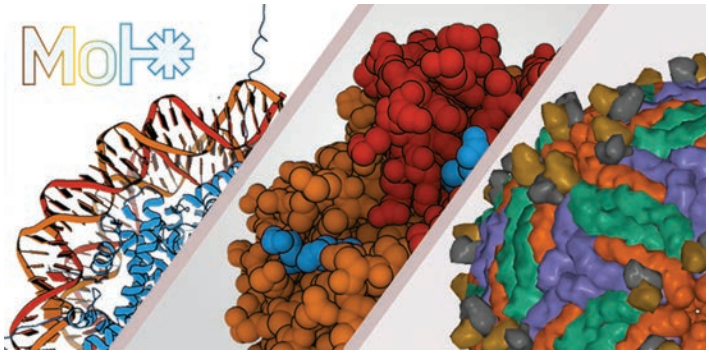
The open-access web portal RCSB.org supports PDB Data Consumers in the US and around the world with resources for PDB structure access, visualization, and analysis.

In 2019, RCSB.org hosted ~4 MILLION USERS.*

The website supports millions of users representing a broad range of skills and interests. In addition to retrieving structure data, PDB users access comparative data, and external annotations, such as information about point mutations and other genetic variations.

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 from 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.

A major 2019 milestone was the release of Mol*, a new web-native 3D molecular viewer that enables fast visualization of molecular structures and their corresponding data, along with high-quality rendering. New services have been deployed to deliver coordinate and map data to this new visualization tool. These services provide data in a binary encoding of PDBx/mmCIF that is compressed and organized to optimize interactive molecular visualization. These services will also enable delivery of combined map and model display essential for evaluating and comparing of data quality. Mol* is an open-source collaborative project between RCSB PDB, PDBe, and CETIEC.



Behind the scenes, a significant 2019 project has been the architectural redesign of the information management 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. Architectural redesign will yield operational efficiencies,improve maintainability and extensibility, enable more proactive monitoring and scaling of services, and ensure continued >99% uptime 24x7x365 service availability.

RCSB PDB tools and resources provide rich structural views of biological systems to enable breakthroughs in scientific inquiry, medicine, drug discovery, technology, and education.

**As reported by Google Analytics.*

4 RCSB PDB SERVICES
OUTREACH AND EDUCATION

Celebrating 20 Years of *Molecule of the Month*

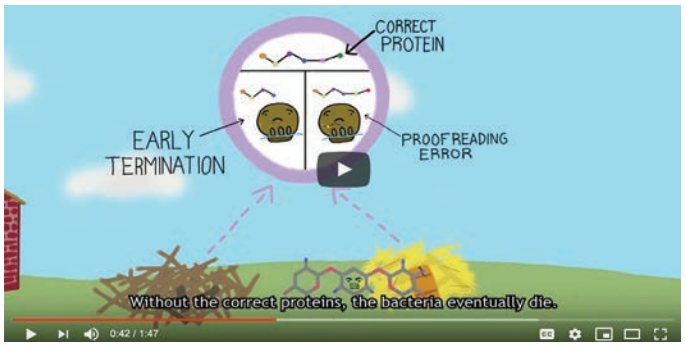
Since 2000, the RCSB PDB *Molecule of the Month* series has introduced millions of visitors to the shape and function of the 3D structures archived in the Protein Data Bank.

Created and illustrated by David S. Goodsell (RCSB PDB-Rutgers and The Scripps Research Institute), this feature tells stories about molecular structure and function, their diverse roles within living cells, and the growing connections between biology and nanotechnology. *Molecule of the Month* content from *Actin* to *Zika* has inspired readers around the world, and is a regular read for students and researchers alike. The series is so compelling that it was accessed nearly a million times in 2019. Each year, the highest-ranked articles are Hemoglobin and Catalase.

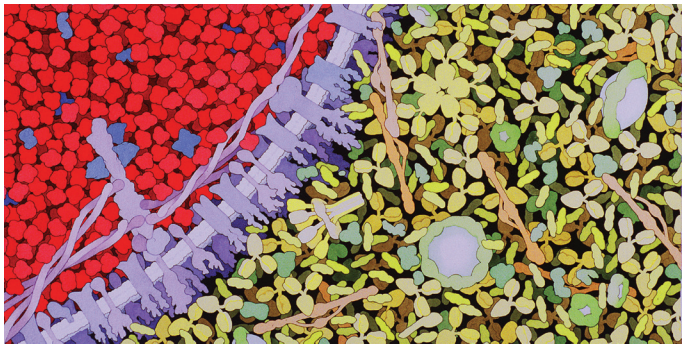
The growth and popularity of the *Molecule of the Month* column led to development of the PDB-101 educational website PDB-101 (**pdb101.rcsb.org**), which provides support for teachers, students, and the curious public interested in exploring the world of proteins, DNA, and RNA. PDB-101 aims to help train today's students who will be tomorrow's PDB users.

In 2019, more than 665,000 USERS viewed nearly 2 MILLION WEBSITE PAGES.*

Content can be searched or browsed using categories such as "Molecular Evolution" and "Nobel Prizes and PDB Structure." Along with *Molecule of the Month*, other features include molecular animations, foldable paper models, award-winning illustrations, and curricular materials.



The annual PDB-101 video challenge encourages high schools students to combine storytelling with molecular exploration.



PDB-101 also hosts an online gallery of David Goodsell's Molecular Landscapes. These watercolor paintings integrate information from structural biology, microscopy, and biophysics to simulate detailed views of the molecular structure of living cells. Shown is *Blood* (2000).

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.

Impact analyses performed within the RCSB PDB demonstrate the impact of PDB structures on approvals by the US Food and Drug Administration (FDA). In the first study (How Structural Biologists and the Protein Data Bank Contributed to Recent FDA New Drug Approvals (2018) *Structure* **27**: 211-217 doi: 10.1016/j.str.2018.11.007), discovery/development of 210 new molecular entities (NMEs; new drugs) approved by the US FDA 2010–2016 was facilitated by open access to 3D structures stored in the PDB. Nearly 6,000 relevant PDB structures contributed to approval of 88% of these NMEs across all therapeutic areas.

A more recent study (Westbrook, J.D. *et al.* Impact of the Protein Data Bank on antineoplastic approvals, *Drug Discov Today* (2020), doi.org/10.1016/j.drudis.2020.02.002) looked at PDB holdings, the scientific literature and related documents for each drug-target combination. It showed how open access to PDB structure data facilitated discovery and development of over 90% of the 79 new antineoplastic agents with known molecular targets approved by the US FDA during 2010–2018.

PDB ARCHIVE AND CANCER DRUG APPROVALS

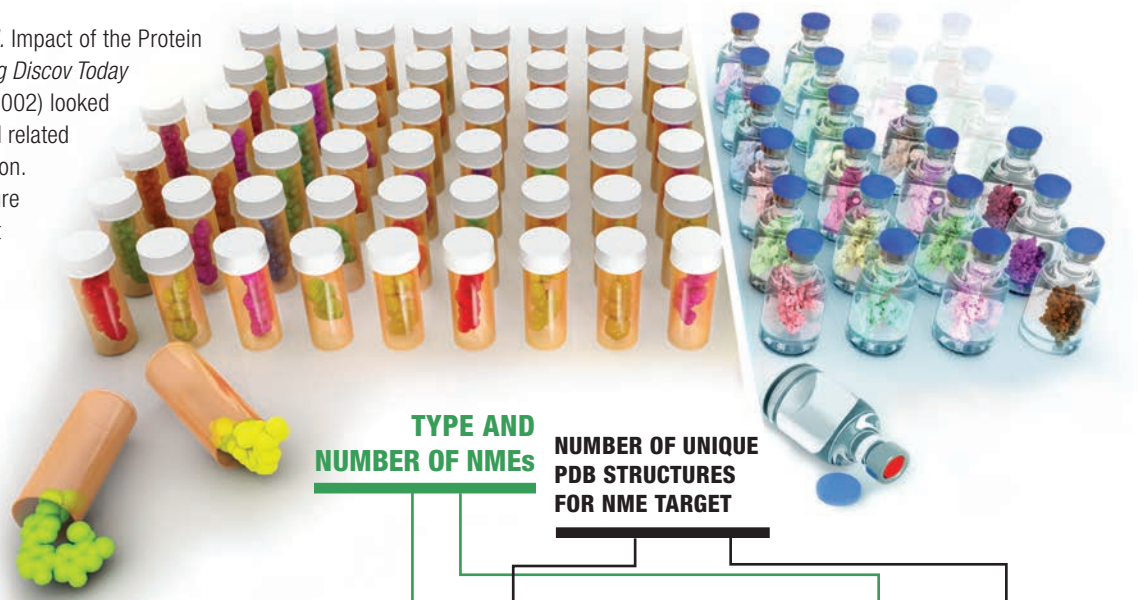
74 of the 79 NMEs approved during 2010–2018 had a total of **2412 unique structures in the PDB** exploring their biological targets in the pre-approval years. Impact ranges from understanding target biology through identifying a given target as likely druggable, to structure-guided optimization of potency and selectivity.

PDB GOLDEN ANNIVERSARY

CELEBRATING 50 YEARS OF OPEN ACCESS TO BIOMOLECULAR STRUCTURES

The Protein Data Bank archive was established as the 1st open access digital data resource in all of biology and medicine. Growing from the initial seven structures to more than 162,000 as of April 2019, the PDB archive is a leading global resource for experimental data central to scientific discovery.

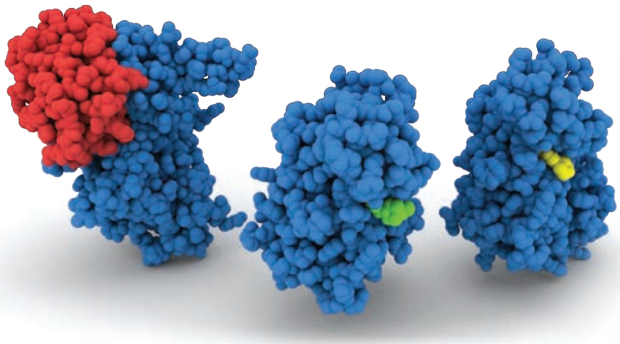
In 2021, a milestone 50th Anniversary will be celebrated by wwPDB Partners RCSB PDB, PDBe, PDBj, and BMRB. Information about symposia, events, and educational materials celebrating the scientific breakthroughs of the PDB will be posted at **wwPDB.org** throughout our golden year.



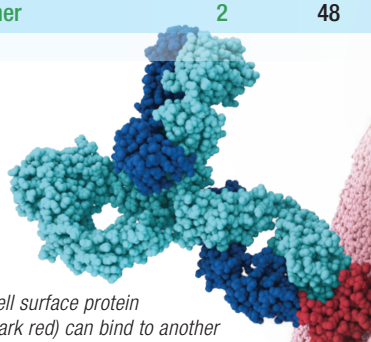
TYPE AND NUMBER OF NMEs | **NUMBER OF UNIQUE PDB STRUCTURES FOR NME TARGET**

Target classes	Low Molecular Weight NMEs	54	2007
	Protein Kinase	33	1,136
	Tubulin	2	249
	Ribosome A Site	1	134
	Androgen Receptor	2	96
	HDAC	2	92
	PARPs	4	89
	Proteasome	2	62
	E3 Ubiquitin Ligase	1	50
	IDH1	1	40
	BCL-2	1	26
	CYP17A1	1	13
	Smoothened	3	11
	IDH2	1	9

	Biologic NMEs	20	405
	Antibody	14	395
	Antibody-Drug Complex	4	30
	Other	2	48



*Abnormal activity of cyclin-dependent kinases (blue) and their regulatory proteins, the cyclins (red), can lead to accelerated cell division, a hallmark of cancer. The new LMW NMEs, abemaciclib (green) and ribociclib (yellow) inhibit the kinase, thereby stopping the malignant cells from dividing. PDB entries shown: **1bi7**, **5l2s**, and **5l2t**.*



*The T-cell surface protein PD-1 (dark red) can bind to another protein on the surface of cancer cell suppressing the body's immune response. The recently approved antibody pembrolizumab (blue), blocks PD-1, thereby enabling the body to kill cancer. Illustration created from PDB entries **5ggs** and **5dk3**.*

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

CITE RCSB PDB

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

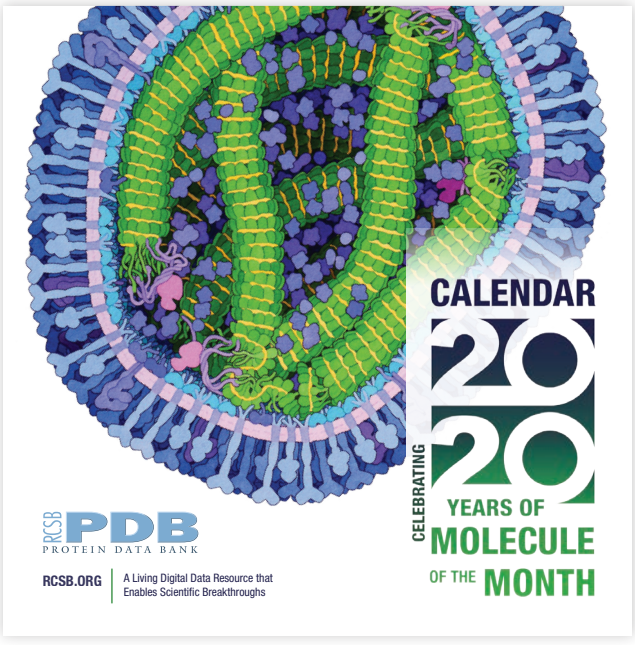
RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy (2019)
Nucleic Acids Research **47**: D464–D474.
doi:**10.1093/nar/gky1004**

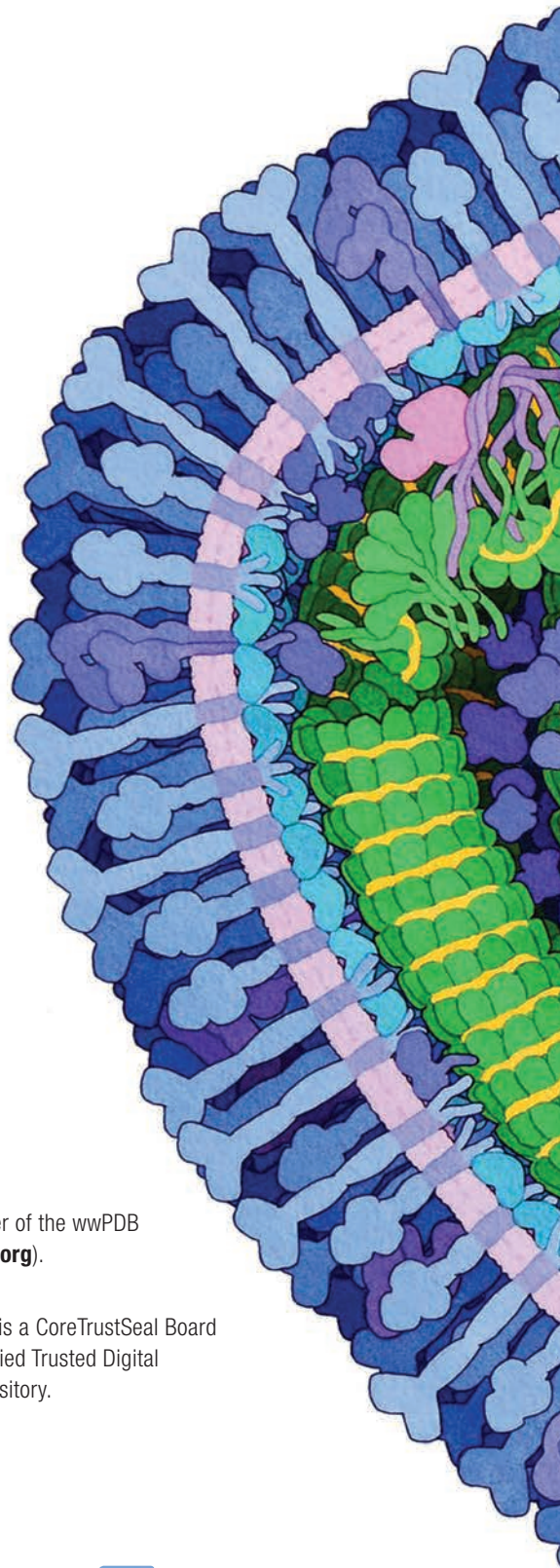
FUNDING

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.

ON THE COVER

To celebrate the 20th anniversary of the *Molecule of the Month* (see more in the *Outreach and Education* section), RCSB PDB produced a 2020 calendar featuring a compilation of the most popular features of the series. The painting shown on the cover by David Goodsell is part of the 2019 article on *Measles Virus Proteins* and illustrates how the six viral proteins work together to infect cells. The calendar can be downloaded on PDB-101 **pdb101.rcsb.org** > *Learn* > *Flyers, Posters & Other Resources*





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.

FOLLOW US

/RCSBPDB
/buildmodels
/RCSBProteinDataBank
/rcsb