ANNUAL REPORT

2018

RCSB PDB
PROTEIN DATA BANK

rcsb.org • info@rcsb.org

A Living Digital Data Resource that Enables Scientific Breakthroughs

DIRECTOR'S MESSAGE

Knowing the 3D structure of biological macromolecules–nucleic acids, proteins, and large molecular complexes–is essential for understanding biology. These shapes and how they change over time play important roles in human and animal health and disease; have significant function in plants and food and energy production; and impact other challenges related to global prosperity and sustainability.

To enable open access to the accumulating knowledge of 3D structure, function, and evolution of biological macromolecules, the PDB Core Archive safeguards terabytes of X-ray, NMR, and 3DEM experimental data. Today, >150,000 validated and expertly-biocurated structures are freely available to all without restriction.

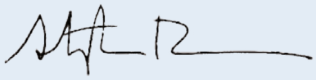
This powerful resource is jointly managed by the Worldwide Protein Data Bank organization, within which the RCSB PDB is responsible for US PDB operations and serves as Archive Keeper.

RCSB PDB provides expert stewardship of 3D bio-structure *Data for All*. Every day, researchers, educators, and students across different scientific fields and disciplines access PDB data as well as specialized tools and resources. These data enable researchers and educators to expand the frontiers of fundamental biology, biomedicine, and biotechnology.

As structural biologists continue transforming the discipline, RCSB PDB team members work every day to embrace and enhance these advances. Services are being scaled to support continued growth and complexity of structures deposited to the archive and to manage new data from evolving experimental methods (including X-ray Free Electron Lasers and 3D Electron Microscopy).

RCSB PDB is also focused on preparing for the onslaught of structural information coming from studies of larger and ever more complex molecular machines. Facilitating the archiving of data from studies that integrate both quantitative measurements and structural analyses enables the PDB data community to make wide-reaching breakthroughs in basic and applied research and outreach and education.

Sincerely,



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

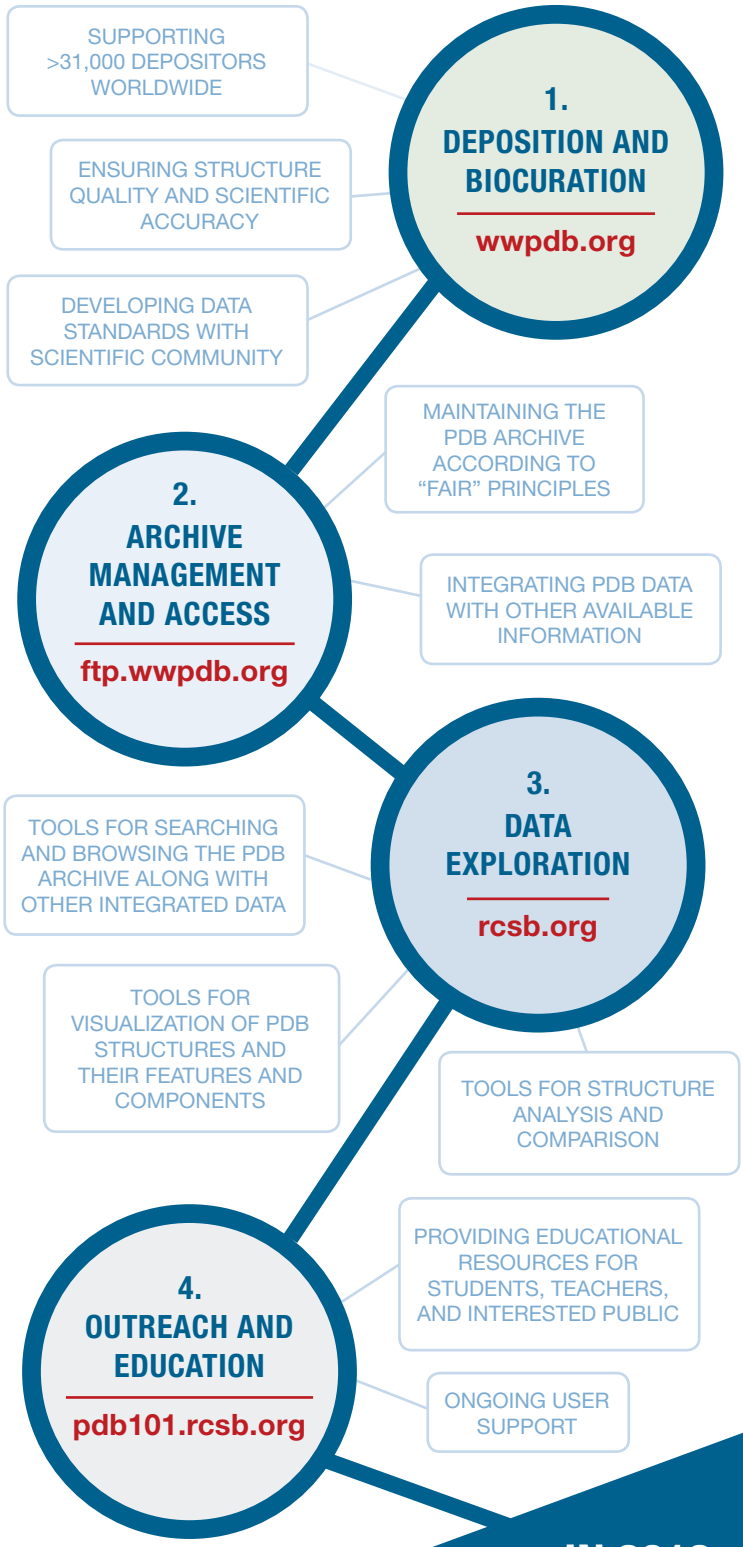
Director, RCSB PDB

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

Adjunct Professor, University of California, San Diego

RCSB PDB SERVICES AT A GLANCE

2018 achievements for each service are described in greater detail in this report.



IN 2018:

12,179 PDB structures were deposited by researchers from around the world

>400 scientific resources utilized the PDB Archive

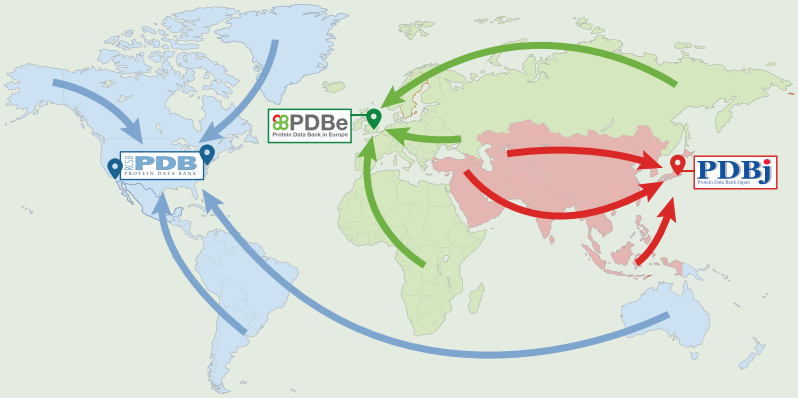
Millions of unique users visited **rcsb.org**

~600,000 unique users visited **pdb101.rcsb.org**

1. DEPOSITION AND BIOCURATION

The Worldwide Protein Data Bank (wwPDB) was established to manage a 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 2018, **12,179 PDB structures** were deposited by researchers from around the world and then prepared for public release by the wwPDB partners RCSB PDB, PDBe, and PDBj. Biocuration responsibilities are distributed geographically. As the US Data Center, RCSB PDB biocurates structures submitted by scientists working in the Americas and Oceania. During 2018, RCSB PDB processed 42% of all incoming structures.



Data Curation and Community-set Standards

PDB structures contain:

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

The CoreTrustSeal Board certified wwPDB as a Trusted Digital Repository in 2018.

CoreTrustSeal was launched by the ICSU World Data System and the Data Seal of Approval as a unified global organization for certification of data repositories. Accreditation requirements encompass the entire life cycle of PDB data management, project organization, and oversight. The wwPDB is dedicated to following CoreTrustSeal standards in order to sustain a freely accessible, single global PDB archive as an enduring public good.

2. 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**[1] 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.

PDB ARCHIVE DATA HOLDINGS DECEMBER 31, 2018

MOLECULE TYPE	Total Number	Number Added in 2018	2018 growth rate
Proteins, peptides, and viruses	136,849	10,308	7.5%
Nucleic acids	3,285	120	3.7%
Protein/nucleic acid complexes	7,367	787	10.7%
Non-polymer and other	30	0	0%
RELEASED ATOMIC COORDINATE ENTRIES	**147,531**	**11,215**	**7.6%**
EXPERIMENTAL TECHNIQUE			
Macromolecular Crystallography	131,928	9,901	7.5%
NMR Spectroscopy	12,477	396	3.2%
Electron Microscopy	2,726	849	31.1%
Multi Method	129	23	18%
Other	271	46	17%
RELATED EXPERIMENTAL DATA FILES			
Structure factors	121,872	9,903	8.1%
NMR restraints	9,818	393	4.0%
Chemical shifts	3,570	433	12.1%
3DEM map files	2,791	906	32.5%

As of April 16, 2019

As the wwPDB Archive Keeper, the RCSB PDB is responsible for safeguarding the PDB archive and maintaining the PDB FTP data access system (**ftp.wwpdb.org**). RCSB PDB coordinates weekly updates of the PDB archive with wwPDB Data Centers in Europe and Japan.

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.

1. M.D. Wilkinson *et al.* (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**: 1-9.

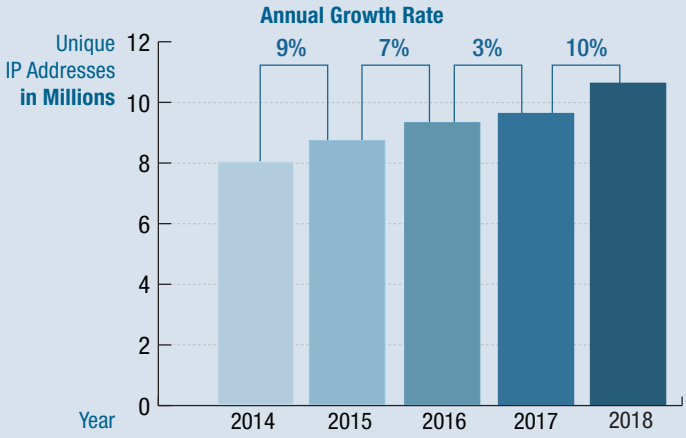
An estimated
~749 million structure data files
were downloaded from all wwPDB partners. More than 500 million structure data files were downloaded from RCSB PDB web and FTP sites, amounting to a total of **126 TB** of information.

3. 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 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 genetic variation.

RCSB.ORG: GROWTH 2014-2018

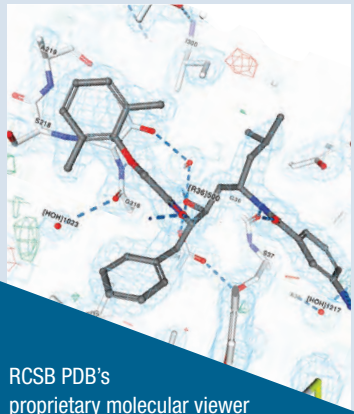


The number of unique IP address that access RCSB.org has increased on average 7.2% each year over the past five years.

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 and structure similarity clusters, and experimental data. These data are updated weekly, which means that while the original scientific publications reporting new structures remain static, RCSB PDB delivers contemporary views of all structures.

Structure Summary pages also offer fast, interactive 3D display of molecular complexes that in some cases contain millions of atoms. This web-native NGL (New Graphic Library) Viewer features also include rapid display of electron density maps, ligand-protein interactions, and validation data in 3D on desktop computers, tablets, and other mobile devices.

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



RCSB PDB's proprietary molecular viewer NGL now displays **electron density maps** which combine the structural model (coordinates) and the experimentally-collected data from an X-ray structure determination and serve to represent the fit of the model to the data.

Shown: Ligand R36 from PDB entry 1lee.

4. OUTREACH AND EDUCATION

PDB-101

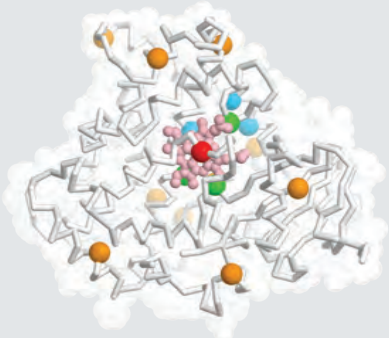
Today's students are tomorrow's PDB users. PDB-101 was created to provide support for teachers, students, and the curious public interested in exploring the world of proteins, DNA, and RNA.

A popular PDB-101 resource is the ***Molecule of the Month*** series, which presents short accounts describing selected molecules from the PDB archive that highlight stories surrounding Fundamental Biology, Biomedicine, and Energy.

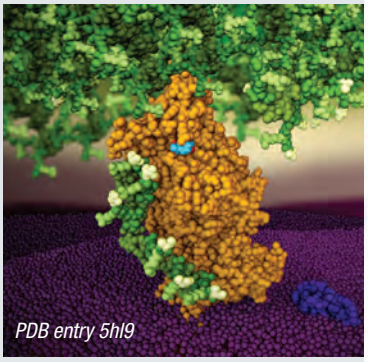
Articles from *Actin* to *Zika* are used in classrooms around the world. In 2018, the most heavily-accessed features included *Hemoglobin*, *Catalase*, and *Green Fluorescent Protein*. Topical articles included *Directed Evolution of Enzymes*, which was inspired by the 2018 Nobel Prize in Chemistry.

Curricular modules and other educational materials help explore other health topics, including HIV & AIDS, and insulin & diabetes.

PDB-101 content can be searched or browsed using categories such as *Molecular Evolution* and *Nobel Prizes and PDB Structure*.

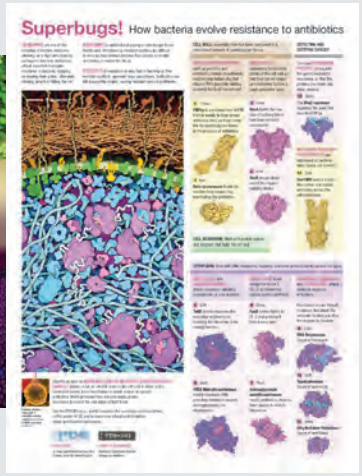


An evolved P411 enzyme (PDB entry 5ucw), with sites of mutation shown with colored spheres as featured in the December 2018 Molecule of the Month article.



PDB entry 5hl9

PDB-101 resource development is also concerned with public health, with the **2018-2019 Health Focus** zooming in on Antibiotic Resistance. A highlight is the annual Video Challenge where high school students submit brief stories that combine medicine and structural biology.



Educational resources developed as part of the 2018-19 Health Focus use PDB structures to explain the molecular mechanisms of bacterial resistance to antibiotics.

RCSB PDB Customer Service

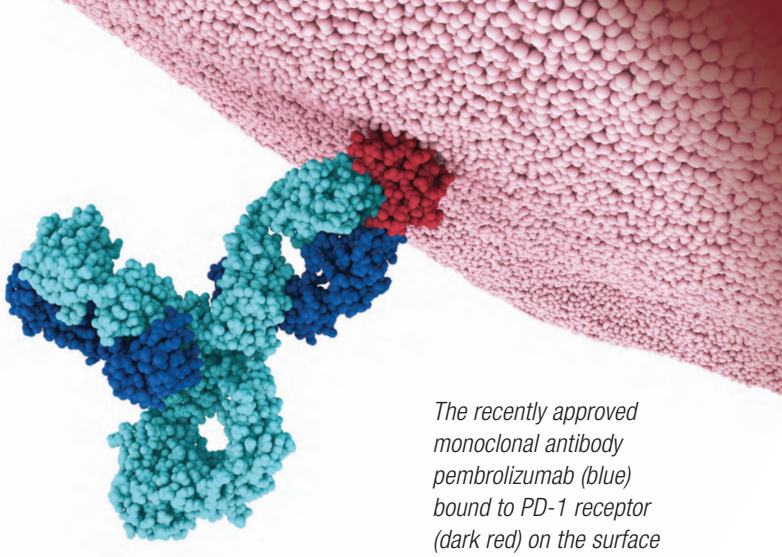
collects and answers questions about the website, PDB data, and structural biology. Nearly **1,000 unique users** initiate new electronic conversations each year. Questions and comments come from students new to structural biology, users involved in the general study of science, and domain experts from the various disciplines that utilize PDB data.

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.

RCSB PDB services broadly impact research and education. The inaugural RCSB PDB publication (Berman *et al.*, *Nucleic Acids Research* 2000) is one of the top-cited scientific reports of all time. A 2017 bibliometric analysis performed independently by Clarivate Analytics shows that the PDB motivates high-quality basic and applied research throughout the world. Papers citing the inaugural publication had a citation-based impact exceeding the world-average in 16 scientific fields including Biology & Biochemistry, Computer Science, Plant & Animal Sciences, Physics, Environment/Ecology, Mathematics, and Geosciences.

An independent economic analysis performed in 2017 by the Rutgers University Office of Research Analytics noted that a reasonable estimate to replicate the data held in the PDB Core Archive at the time exceeded $12 billion. The same study documented a ~1,500-fold return on investment for federal funding of the RCSB PDB, excluding the impact of PDB data on the biopharmaceutical industry.



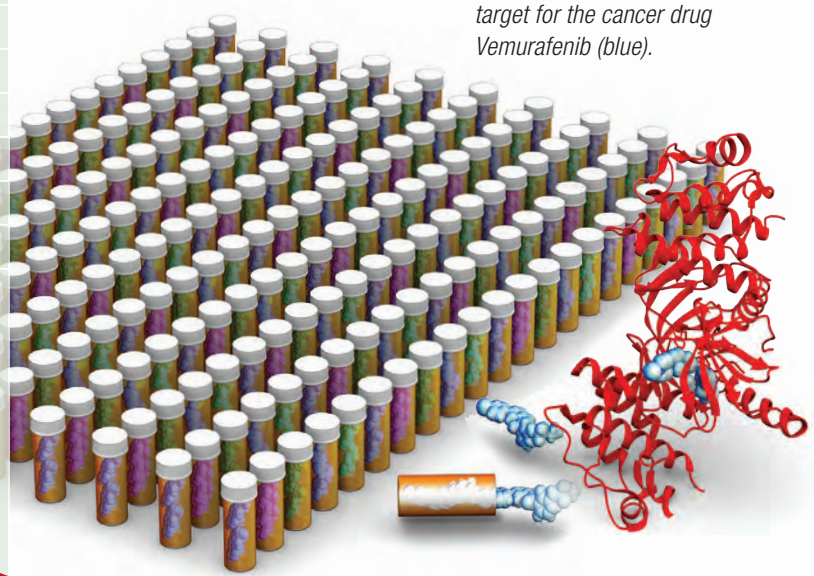
The recently approved monoclonal antibody pembrolizumab (blue) bound to PD-1 receptor (dark red) on the surface of T-cell. This illustration was created based on PDB entries 5ggs and 5dk3, both available in the PDB Archive before this life-saving drug was approved.

A more recent impact analysis[1] performed within the RCSB PDB showed that structural biologists and the PDB Core Archive contributed to the pre-competitive and proprietary research efforts leading to US FDA regulatory approval of nearly 90% of the 210 new drugs approved for patients between 2010 and 2016. Representative examples of new biopharmaceutical agents include the monoclonal antibodies nivolumab and pembrolizumab, which both reactivate the immune system to kill malignant cells in a wide range of cancers; linagliptin, which potentiates the effect of the body's natural system for regulating blood sugar to combat type 2 diabetes mellitus; and tofacitinib, which blocks the action of the Janus kinase enzyme for treatment of rheumatoid arthritis and other autoimmune diseases.

Almost 90% of the 210 FDA-approved drugs (2010-2016), had the total of 5,913 related atomic structures in the PDB Archive in the pre-approval years, supporting pre-competitive research. Many of these structures are direct targets of the drug, like the B-Raf Kinase shown below in red from the PDB entry 3og7, a target for the cancer drug Vemurafenib (blue).

PDB ARCHIVE AND FDA-DRUG APPROVALS

2010-2016 Type and Number of FDA-approved New Drugs	**210**	Share of FDA-Approved Drugs with PDB Structures **184**	Number of PDB Target and Target Biology Structures **5,913**
Anti-neoplastic	59	55	1,325
Anti-infective	31	27	1,354
Cardiovascular	21	16	1,261
Central Nervous System	21	18	486
Endocrine	7	7	627
Gastrointestinal	6	5	45
Immunologic	21	19	861
Metabolic	26	23	315
Respiratory	13	11	663
Miscellaneous	4	3	35



The value of the data held in the PDB Core Archive:
> $12 billion (as of 2017)

Return on investment for federal funding of the RCSB PDB:
~1,500-fold (excluding impact of PDB data on the biopharma industry)

Data Source: Economic Impacts of the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank. Rutgers Office of Research Analytics, May 2017. doi: **10.2210/rcsb_pdb/pdb-econ-imp-2017**

1. 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**

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 and UCSD/SDSC

RUTGERS

UC San Diego

SDSC SAN DIEGO SUPERCOMPUTER CENTER

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**

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

FUNDING

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

The cover illustrations in this report are sourced from RCSB PDB's 2019 calendar and educational video *What is a Protein?* These resources focus on the intimate relationship between protein structure and function, and are available from **pdb101.rcsb.org**.

PDB entries shown on the cover:
Front: 1i6h and 1bna (top), 1igt and 4rhv (middle), 1cag (bottom)
Back: 6dde

/RCSBPDB

/buildmodels

/RCSBProteinDataBank

/rcsb