

**STATE**  
— *of the* —  
**RCSB PDB**

July 1<sup>st</sup> 2016

[rcsb.org](http://rcsb.org)

*Molecular Explorations of Biology and Medicine*

**RC****S****P****D****B**  
PROTEIN DATA BANK

## DIRECTOR'S MESSAGE



The Protein Data Bank (PDB) was established in 1971 as the first open access digital data resource in biology with just 7 X-ray crystal structures of proteins. Today, the PDB is a billion-atom primary experimental data repository. Three-dimensional (3D) structural data for more than 120,000 large biomolecules (proteins, DNA/RNA) are publicly accessible in the PDB archive.

The RCSB Protein Data Bank (RCSB PDB), the US data center for the global PDB archive, makes PDB data freely available to all users, from structural biologists to computational biologists and beyond.

The Worldwide Protein Data Bank partnership (wwPDB; including the RCSB PDB, Protein Data Bank in Europe, Protein Data Bank Japan, and BioMagResBank) exemplifies how an international collaboration can effectively manage 'big data' as a freely accessible public good that advances basic and applied research and education worldwide.

The RCSB PDB team, operating at Rutgers, The State University of New Jersey and the University of California San Diego/San Diego Supercomputer Center, serves as the PDB archive keeper, responsible for data integrity and disaster recovery. As the wwPDB regional data center for the Americas and Oceania, RCSB PDB manages and curates incoming structural data from these regions. Each week, RCSB PDB updates the archive by combining new data from the wwPDB partners and making the growing body of PDB data freely available to all via FTP mirror sites.

To serve researchers, educators, and students worldwide, RCSB PDB develops resources that offer rich structural views of biological systems. From tools focused on understanding drugs to the ability to map the impact of genomic variation on 3D protein structure, the team is working to enable breakthroughs in scientific inquiry across disciplines.

Together with Director Emerita Helen M. Berman, the entire RCSB PDB staff, and our Advisory Committee, I am committed to safeguarding the PDB and delivering the highest quality of service to our Users around the world. As a bicoastal team, we employ cutting edge software tools and state-of-the-art hardware to maintain exemplary standards of data integrity and security for the archive. We work at the forefront of biological big data integration and curation to connect the PDB archival content with more than 40 biological data resources, and provide our Customers with rich 3D views of biology and medicine.

Sincerely,

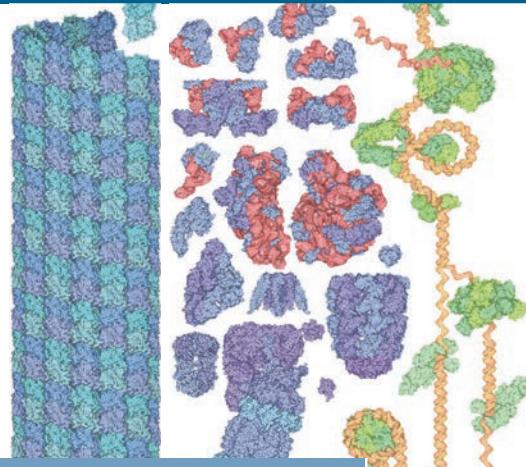
Stephen K. Burley, M.D., D.Phil.

Director, RCSB PDB  
Distinguished Professor, Department of Chemistry and Chemical Biology  
Rutgers, The State University of New Jersey  
Adjunct Professor, University of California, San Diego

## THE PDB ARCHIVE

Cells rely on many large molecular machines that carry out the complex biological and chemical tasks to sustain life.

3D structures of these machines are freely available at the Protein Data Bank, the global storehouse of biomolecular structures central to research and education.



### PDB ARCHIVE CONTENTS ON JULY 1<sup>ST</sup>, 2016

**120,057** Released atomic coordinate entries

MOLECULE TYPE	EXPERIMENTAL METHOD
<b>111,440</b> Proteins, peptides, and viruses	<b>107,264</b> X-ray crystallography
<b>5,658</b> Protein/nucleic acid complexes	<b>11,435</b> Nuclear Magnetic Resonance
<b>2,933</b> Nucleic acids	<b>1,065</b> 3D Electron Microscopy (3DEM)
<b>26</b> Other	<b>197</b> Other
	<b>96</b> Hybrid

#### RELATED EXPERIMENTAL DATA FILES

<b>91,864</b>	Structure factors
<b>8,539</b>	NMR restraints
<b>2,297</b>	Chemical shifts
<b>905</b>	3DEM maps

### ACCESS

Each month in 2015, [rcsb.org](http://rcsb.org) was visited **741,000** times on average by **315,000** unique visitors

A total of **35,260 GB** of data were accessed

PDB data are accessed from **192** of the **195** sovereign nations recognized by the United Nations worldwide (shown in blue)

More than **1.5 million PDB structures** are downloaded **EVERY day of the year**.

Significant PDB data usage also occurs behind the firewalls of pharmaceutical and biotechnology companies, where private copies of the PDB archive are maintained.

Shown: Structure of the E. coli ribosome inhibited by tetracycline (PDB ID 5j5b)

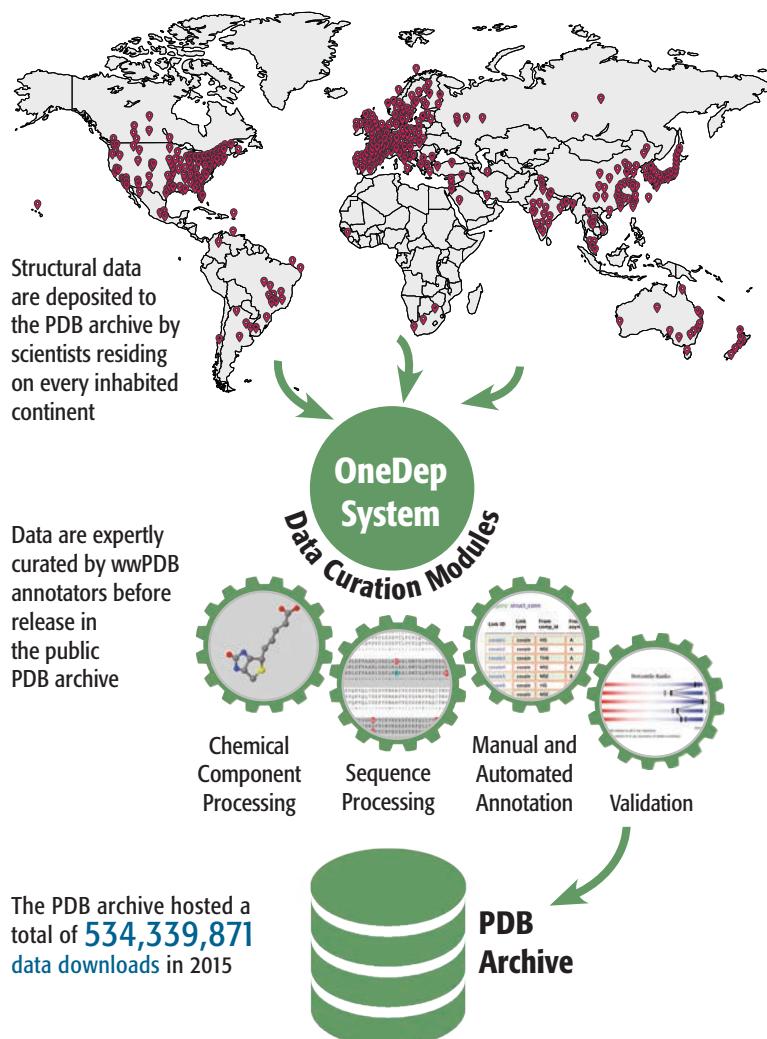


# DATA DEPOSITION AND ANNOTATION



The PDB archive is managed by the Worldwide PDB organization (wwPDB, [wwpdb.org](http://wwpdb.org)), a collaboration involving PDB data centers in the US, UK, and Japan. wwPDB ensures that these valuable data are securely stored, expertly managed, and made freely available for the benefit of scientists and educators around the globe.

wwPDB data centers work closely with community experts to define deposition and annotation policies, resolve data representation issues, and implement community validation standards.



Ligand Validation Workshop Participants



## OneDep System

A major focus of the wwPDB collaboration has been to develop a single software system that could be used to validate, deposit, and annotate PDB data from all accepted methods. The resulting powerhouse program OneDep has been successfully used at all wwPDB sites for X-ray, NMR, and 3DEM data since early 2016.

### OneDep:

- ✓ produces more uniform data, as it uses the PDBx/mmCIF data format and dictionaries
- ✓ supports replacement of data files pre- and post-deposition
- ✓ enhances communication with Depositors
- ✓ enables improved annotation
- ✓ provides detailed validation reports based on recommendations from wwPDB expert task forces.

Global adoption of the OneDep system has already resulted in the improved representation of data in the archive.

In 2015, the wwPDB curated **10,956** deposited structures. The PDB archive is on track to receive **~12,000** new structures in 2016

## Best Practices in Data Representation

The wwPDB brings together community experts to provide guidance on best practices. In 2015, the **wwPDB/CCDC/D3R Ligand Validation Workshop** brought co-crystal structure determination experts from academe and industry together with X-ray Crystallography and Computational Chemistry software developers to discuss and develop best practices for validation of co-crystal structures; editorial/refereeing standards for publishing co-crystal structures; and recommendations for ligand representation across the archive.

These recommendations have been published in *Structure*<sup>1</sup>, providing guidance for data archiving, annotation, and validation practices of co-crystal structural data.

<sup>1</sup>(2016) *Structure* 24: 502-508 doi: [10.1016/j.str.2016.02.017](https://doi.org/10.1016/j.str.2016.02.017)

# DATA ACCESS, QUERY, AND REPORTING

As the wwPDB archive keeper, the RCSB PDB ensures the integrity and disaster recovery of PDB data, and coordinates weekly updates of the PDB archive at [ftp.wwpdb.org](http://ftp.wwpdb.org).

Data and RCSB PDB services can be accessed freely online *via* our website, FTP, Web Services, and a mobile app. Our software libraries are open source on GitHub. The RCSB PDB *Mobile* app supports easy on-the-go searching and visualization.

**RCSB.ORG** provides rich structural views of biological systems to enable breakthroughs in scientific inquiry, medicine, drug discovery, technology, and education

To enable data exploration, multiple tools are offered for structure query, analysis, and molecular visualization.

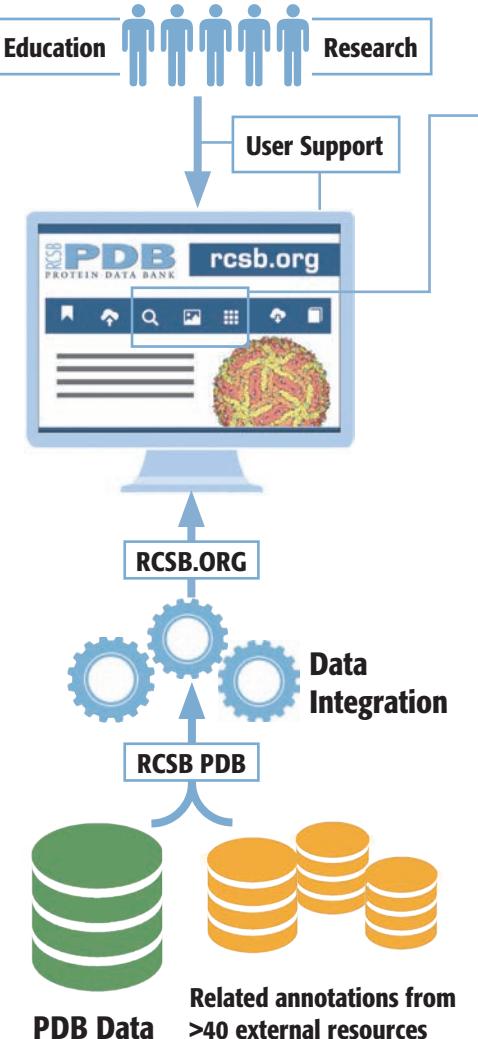


The PDB contains many large macromolecules bound to smaller chemical components, including many US FDA-approved drugs.

## DATA ACCESS AT RCSB.ORG

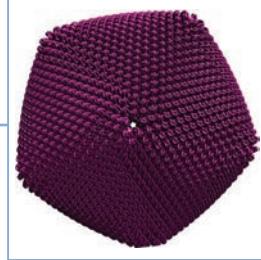
## DATA QUERY, EXPLORATION, ANALYSIS, AND UTILIZATION

### RCSB PDB Users



### SEARCH

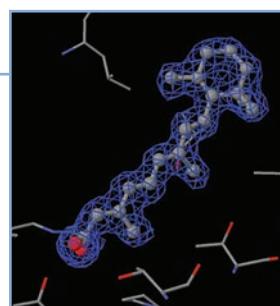
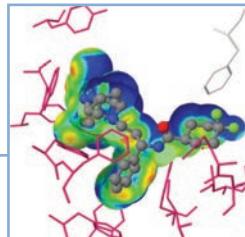
Users can perform simple searches (e.g., ID, name, sequence, ligand) or build complex search combinations of parameters and criteria. Other classification systems are used to organize PDB data into hierarchical trees for browsing and searching (e.g., Membrane Proteins, Gene Ontology, Enzyme Classification).



RCSB PDB tools support visualization and access to extremely large molecular machines. 68 of the 100 largest protein structures in the PDB archive have been released in the past three years, including this faustovirus, the largest known virus (PDB ID 5j7v).

### VISUALIZE

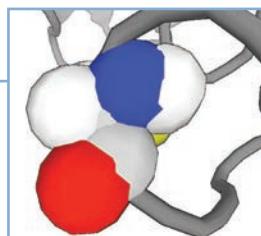
Visualization features include **Protein Feature View**, a graphic comparison of a PDB sequence with UniProt and other annotations; and **Gene View**, a tool that illustrates the correspondences between the human genome and 3D structure.



To help users visualize electron density maps for ligands, Sigma-weighted  $2m|F_O - F_C|$  electron density "mini-maps" can be viewed in 3D.

### ANALYZE

Users can analyze various features of 3D structures as well as inspect the sites of genetic mutations using the **Map Genomic Position to Protein Sequence and 3D Structure** tool.



Mutations of specific amino acids can lead to disease. The molecular locations and impact of these mutations can be viewed in 3D at [rscsb.org](http://rscsb.org).

## OUTREACH AND EDUCATION

The PDB-101 website ([pdb101.rcsb.org](http://pdb101.rcsb.org)) was recently redesigned to help teachers, students, and the general public explore the 3D world of proteins and nucleic acids

### RESOURCES FOR LEARNING

PDB-101 hosts videos, posters, and even paper models for hands-on exploration, including the Zika virus shown here.



### SEARCHING AND BROWSING

All resources are both searchable and organized by topics, including Health and Disease, Molecules of Life, Biotech and Nanotech, and Structures and Structure Determination.

### MOLECULE OF THE MONTH

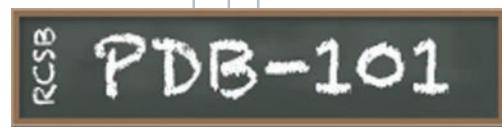
This web publication series presents short accounts describing selected molecules from the PDB that highlight the structure and function of the molecule and its relevance to health and welfare.

2015 marked the 15<sup>th</sup> year of the *Molecule of the Month* (MotM) series. Since its launch in January 2000, these articles have explored the structure and function of PDB macromolecules from AAA+ proteases to Zika virus. A very popular RCSB PDB feature, MotM articles are utilized in classrooms around the world.



### CURRICULUM MATERIALS

With a focus on Biomolecular Structures, Immunology, and HIV/AIDS, these modules include authentic data from the PDB and other public resources, hands-on activities, teaching materials, individual and group activities, and assessment suggestions. They were collaboratively developed by researchers, curriculum designers, educators, clinicians, and teachers from local schools.



### EVENTS

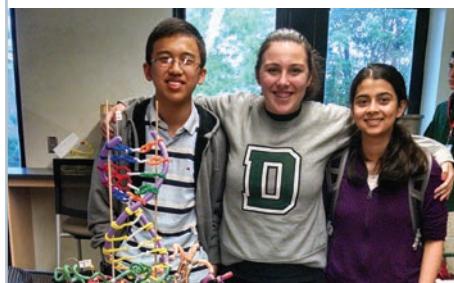


A team from Academy of Allied Health and Science (Neptune City, NJ) received 1<sup>st</sup> place in the 2015 Video Challenge for High School Students.

The *Art of Science* traveling exhibition highlights the fascinating 3D shapes and symmetries of biological macromolecules found in the PDB.

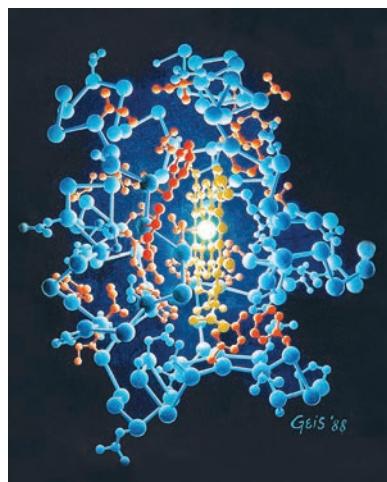


A team from Harriton High School (Rosemont, PA) took first place in the 2015 National Science Olympiad Protein Modeling Event co-developed by RCSB PDB and MSOE Center for BioMolecular Modeling.



### GEIS DIGITAL ARCHIVE

Through a collaboration with the Howard Hughes Medical Institute (HHMI), which owns the Geis Archives, RCSB PDB has established a digital archive of Geis' molecular art. This resource displays many of Geis' illustrations in the context of the corresponding PDB structures and related molecular information.



Irving Geis, Cytochrome c (1988)

## On the Front Cover:

*Insulin Action* by David S. Goodsell

The painting depicts a 190 nanometer section of a human cell. The hormone insulin (yellow) is shown binding to the insulin receptor (green) thus activating a signaling cascade leading to glucose (white) being transported inside the cells to be used as energy, or to be stored as glycogen (purple). A new poster *Insulin and Diabetes* built around the painting can be downloaded at [pdb101.rcsb.org](http://pdb101.rcsb.org).

## Funding:

Core RCSB PDB operations are funded by a grant (DBI-1338415) from the National Science Foundation, the National Institutes of Health, and the US Department of Energy. RCSB PDB technology development activities are funded by grants to individual team members from the National Science Foundation and the National Institutes of Health.

## Cite RCSB PDB:

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

The RCSB Protein Data Bank: views of structural biology for basic and applied research and education (2015)  
*Nucleic Acids Res* **43**: D345-D356.  
doi:10.1093/nar/gku1214



[rcsb.org](http://rcsb.org) • [info@rcsb.org](mailto:info@rcsb.org)

The RCSB PDB is managed by the members of the Research Collaboratory for Structural Bioinformatics:

RUTGERS

UC San Diego [SDSC](#)  
SAN DIEGO SUPERCOMPUTER CENTER

