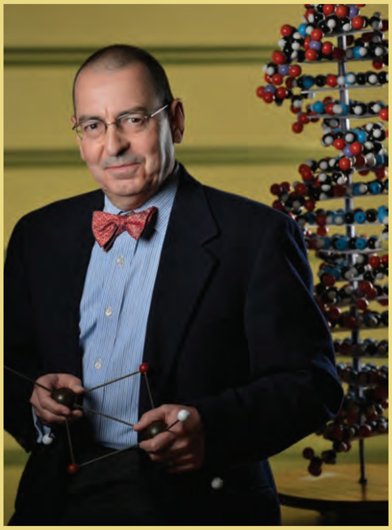
RCSB PDB ANNUAL REPORT 2016

RCSB PDB
PROTEIN DATA BANK

RCSB.ORG • INFO@RCSB.ORG

MOLECULAR EXPLORATIONS
OF BIOLOGY AND MEDICINE

DIRECTOR'S MESSAGE



The Protein Data Bank (PDB) is the single global archive of 3D macromolecular structure data, providing compelling atomic level views of many of the diverse molecular machines found in living organisms and viruses. PDB has been making data freely available since 1971, when it was established as the first open access digital data resource in biology with just 7 protein structures.

Today, the PDB safeguards structure data for more than 125,000 experimentally studied proteins and nucleic acids, and provides unrestricted access for more than 1 million researchers, educators, and students around the world. More than 1.5 million structure data files are downloaded over the Internet by our Users each and every day. The replacement value of the entire PDB archive is conservatively estimated to be more than US$13 billion.

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

The RCSB PDB team focuses on developing tools and best practices for biocuration and validation of PDB data. Additionally, RCSB PDB develops resources that offer rich structural views of biological systems. From tools for drug discovery to understanding the impact of genomic mutations on 3D protein structure and function, RCSB PDB scientists and computer programmers are working to enable research breakthroughs and leverage the power of the data in the PDB archive.

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.

Yours Faithfully,

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

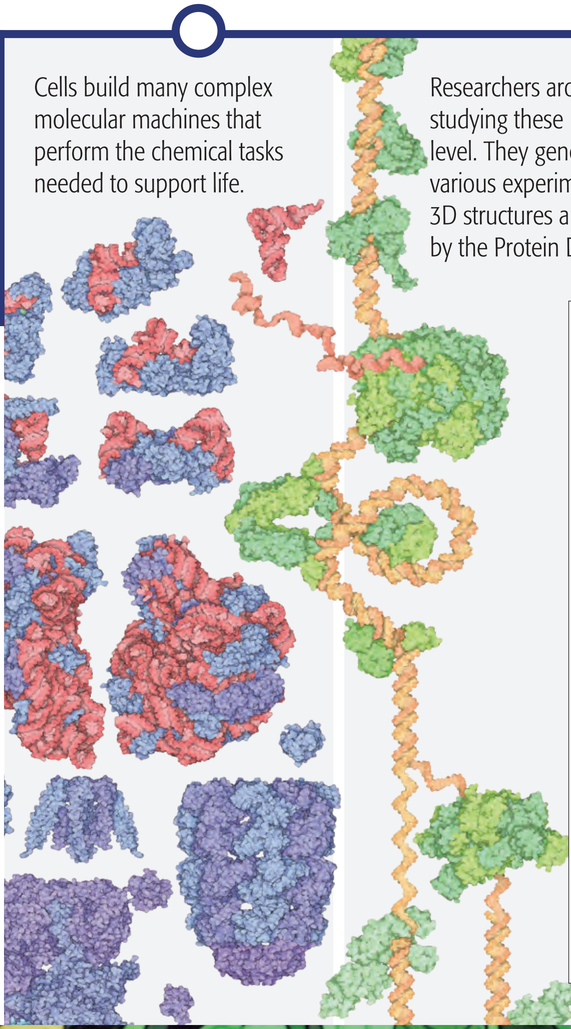
Director, RCSB PDB

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

Adjunct Professor
University of California, San Diego

Report Legend: ○ Independent 🟢 Managed by the wwPDB Staff 🔵 Managed by the RCSB PDB Staff

ABOUT THE PDB



○ Cells build many complex molecular machines that perform the chemical tasks needed to support life.

○ Researchers around the world are studying these molecules at the atomic level. They generate structural data using various experimental techniques. These 3D structures are made freely available by the Protein Data Bank (PDB).

🟢 PDB is the central storehouse of biomolecular structures essential to basic and applied research and education in biology and medicine. PDB was established in 1971 with just 7 structures. On January 1st 2017, the archive contained 125,463 entries. During 2016, 10,881 new structures were added to the public archive.

PDB ARCHIVE SNAPSHOT: JANUARY 1, 2017

125,463 entries

Holdings by Molecule Type	
Proteins, peptides, and viruses	116,449
Nucleic acids	3,020
Protein/nucleic acid complexes	5,968
Other	26

Holdings by Experimental Techique	
X-ray	112,156
NMR	11,671
Electron Microscopy (3DEM)	1,329
Hybrid	103
Other	204

Related Experimental Data Files	
101,776	Structure factors
8,993	NMR restraints
2,762	Chemical shifts
1,314	3DEM map files

DATA DEPOSITION AND ANNOTATION

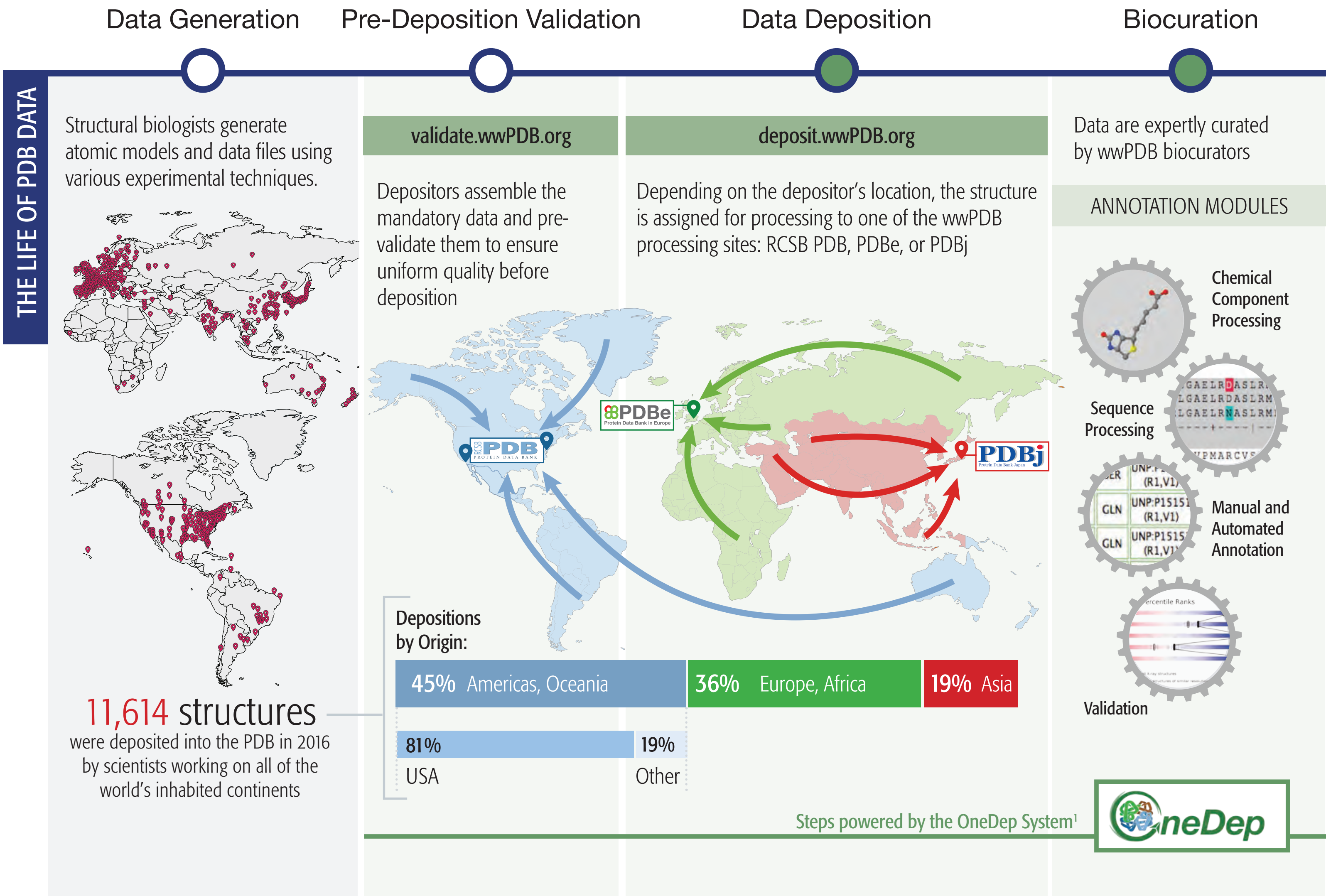
The PDB archive is managed by the Worldwide PDB organization (wwPDB, **wwPDB.org**), an international collaboration involving regional data centers in the US, Europe, and Japan.

wwPDB partners ensure that valuable structure data are securely stored, expertly managed, and made freely available for the benefit of scientists, educators, and students around the globe. wwPDB data centers work closely with community experts to define deposition and biocuration policies, and resolve data representation challenges. Structure validation standards are developed with the help of volunteer wwPDB Validation Task Forces that make recommendations and contribute software tools used to prepare wwPDB validation reports assessing the quality and accuracy of every structure stored in the PDB archive.

In 2016, 11614 structures were deposited into the PDB by scientists working on all of the world's inhabited continents. After deposition, wwPDB biocurators carefully review and annotate each structure, and perform numerous validation checks for the depositor before the new entry is finalized for public release. Open communication is of paramount importance; wwPDB biocurators are in regular contact with depositors, working closely with them to bring the most accurate data possible into the PDB archive. Expert biocuration ensures that each PDB entry is a faithful representation of both the structure and the experiment. wwPDB biocurators review polymer sequences, small molecule ligand chemistry, cross-references to other databases, experimental methods, correspondence of atomic coordinates with primary experimental data, protein and nucleic acid polymer geometry, biological assemblies, and crystal packing following well-established wwPDB protocols. This exacting process helps to ensure that all incoming data meet community defined quality standards, thereby enabling meaningful analyses and comparisons across the entire archive.

Validation of structural data deposited to the PDB archive also helps to ensure the integrity of the peer-reviewed scientific literature. Most journals publishing newly determined 3D structures of biomolecules now require (or strongly encourage) authors to provide wwPDB validation reports along with their submitted manuscripts. Access to the wwPDB validation reports helps both referees and editors better evaluate the science and improve publication quality. After structure release, these same wwPDB validation reports are publicly available, helping consumers of PDB data ensure that the structures they are studying are of sufficient quality and accuracy for their intended purpose.

wwPDB develops specialized software to streamline the functions of data deposition, validation, and biocuration while maintaining a high-quality archive. In 2016, the wwPDB OneDep system was deployed to more efficiently support structure data depositions coming from X-ray crystallography, NMR, and 3DEM experiments[1]. OneDep provides detailed validation reports to depositors. In a recently published study[2], data in the PDB showed improved quality of structures deposited using the wwPDB OneDep system versus structures deposited with legacy systems, demonstrating that introduction of the new wwPDB validation reporting system is having the intended impact.



THE LIFE OF PDB DATA

Data Generation

Structural biologists generate atomic models and data files using various experimental techniques.

11,614 structures were deposited into the PDB in 2016 by scientists working on all of the world's inhabited continents

Pre-Deposition Validation

validate.wwPDB.org

Depositors assemble the mandatory data and pre-validate them to ensure uniform quality before deposition

Data Deposition

deposit.wwPDB.org

Depending on the depositor's location, the structure is assigned for processing to one of the wwPDB processing sites: RCSB PDB, PDBe, or PDBj

Depositions by Origin:

45% Americas, Oceania	36% Europe, Africa	19% Asia

81% USA	19% Other

Biocuration

Data are expertly curated by wwPDB biocurators

ANNOTATION MODULES

Chemical Component Processing

Sequence Processing

Manual and Automated Annotation

Validation

Steps powered by the OneDep System[1]

RCSB PDB PROTEIN DATA BANK
PDBe Protein Data Bank in Europe
PDBj Protein Data Bank Japan
OneDep

DATA DISTRIBUTUTION AND ACCESS

RCSB PDB
PROTEIN DATA BANK

The RCSB PDB website (**RCSB.org**) provides rich structural views of biological systems to enable breakthroughs in scientific inquiry, medicine, drug discovery, biotechnology, and education.

In 2016, RCSB.org supported more than 1 million Users from around the world by providing access to a diverse array of tools and services for structure searching, analysis, and molecular visualization.

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

Visualization features include Protein Feature View (a graphical comparison of a PDB sequence with UniProt and other annotations) and Gene View (an additional graphical tool that illustrates the correspondences between the human genome and 3D structure).

As the official wwPDB archive keeper, the RCSB PDB has the added responsibility of safeguarding the PDB archive and maintaining the PDB FTP data access system (**ftp.wwpdb.org**). Weekly updates of the PDB archive and the RCSB PDB website are coordinated by the RCSB PDB with the wwPDB data centers in Europe and Japan.

In addition to the website (**RCSB.org**), RCSB PDB provides PDB data and services via the FTP server (**ftp.RCSB.org**) and numerous Web Services. Several software libraries are made available as open source on GitHub.

Archive Update

ftp.wwPDB.org

Newly curated PDB data are added to the master PDB archive. **As archive keeper, RCSB PDB** packages and re-distributes the data to wwPDB partners. The archive is updated weekly.

10,881 structures were released into the public archive in 2016

In 2016, the PDB archive hosted a total of

591,876,087 data downloads

Public Release and Access

RCSB.org • ftp.RCSB.org

RCSB PDB provides open access to PDB data *via* three avenues:

1. **RCSB.ORG** WEBSITE

>1 million unique users visited RCSB.org in 2016

PDB DATA + INTEGRATED RELATED ANNOTATIONS FROM ~40 EXTERNAL RESOURCES

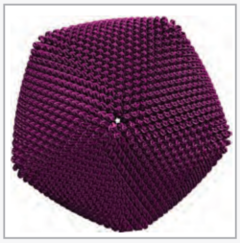
2. **FTP.RCSB.ORG** SERVER

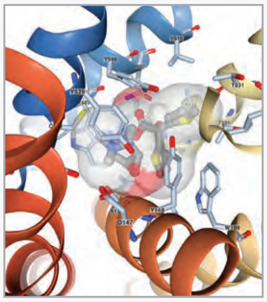
3. VARIOUS WEB SERVICES

~455 million data files were downloaded from RCSB PDB web and FTP sites

VISUALIZING AND ANALYZING PDB DATA ON RCSB.ORG (SELECTED FEATURES)

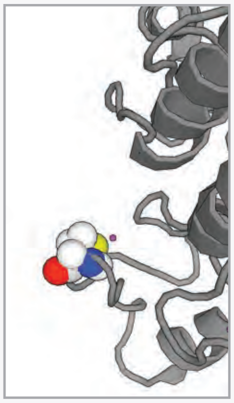
RCSB PDB tools can support visualization and access to extremely large molecular machines, such as the Faustovirus, the largest known virus (PDB ID 5j7v).

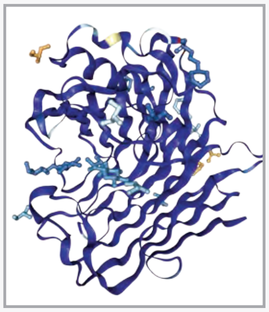


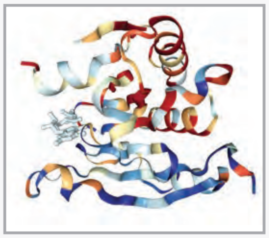


RCSB PDB tools can search for drugs and drug targets. Visualization tools allow for complex analysis of the drug binding site. Shown COPD drug tiotropium bound to the Muscarinic acetylcholine receptor M3 (PDB ID 4u15).

Mutations in a gene can have profound effects on the function of a protein. RCSB tools allow Users to map the locations of the SNPs onto proteins. Shown is the PDB structure 1jm7 with highlighted cysteine 64 with zinc ion (purple) aligned to it, forming a part of the zinc finger domain. A mutation on the BRCA1 gene leads to replacement of the cysteine by glycine, which has been linked to breast cancer formation.







wwPDB Validation Reports provide an assessment of the quality of a structure. Concerns are highlighted by considering the atomic coordinates, experimental data and fit between the two. RCSB PDB visualization tools allow Users to map validation report information onto the 3D structure.

The PDB structure 3wy6 shown on the left top has a good density fit; The PDB structure 1ej1 shown below has areas with problematic fit.

OUTREACH AND EDUCATION

PDB-101 (**pdb101.rcsb.org**) is an online portal designed for teachers, students, and the curious public to promote exploration of the world of proteins, DNA, and RNA. Learning about the diverse shapes and functions of these biological macromolecules promotes understanding of all aspects of fundamental biology, biomedicine, and energy.

SELECTED PDB-101 FEATURES

Molecule of the Month

This popular series presents short accounts describing selected molecules from the PDB archive that highlight the structure and function of the molecule and its relevance to health and welfare. More than 200 articles are available, ranging from Actin to Zika.

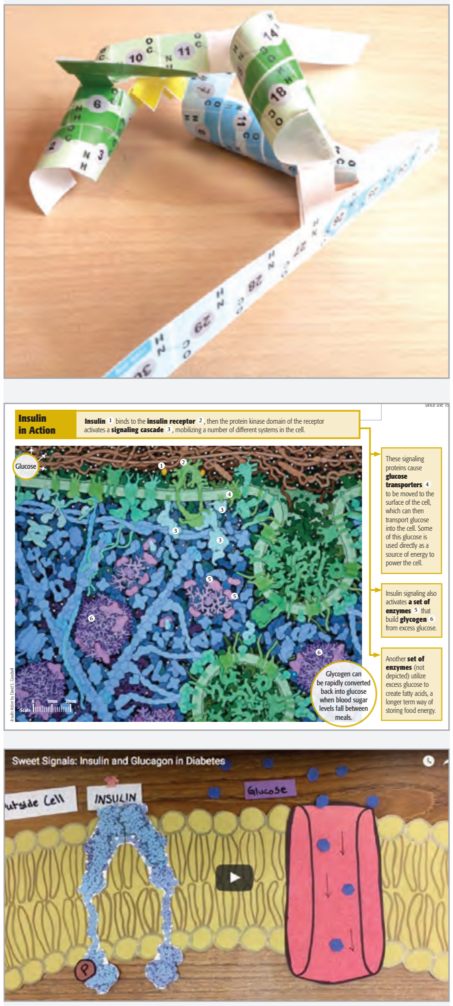
In 2016, the most heavily-accessed PDB-101 articles included catalase, hemoglobin, lysozyme, green fluorescent protein, alcohol dehydrogenase, penicillin-binding proteins, ribosomal subunits, ferritin, insulin, and the acetylcholine receptor.

200th
Molecule of the Month
200
ICOSAHEDRAL VIRUSES
FROM THE PDB
NUMBER OF PROTEIN CHAINS in each capsid
SCALE
PDB

August's 2016 feature on Quasisymmetry in Icosahedral Viruses marked the 200th installment of *Molecule of the Month*. Several resources accompanied this article to celebrate the milestone. They included the Quasisymmetry in Icosahedral Viruses activity for building several paper models of viruses to explore how quasisymmetry is used to build capsids with different sizes. In the poster *200 Icosahedral Viruses from the PDB*, two hundred virus structures are arranged by size and colored by the number of protein chains in each capsid.

Focus on Public Health: Diabetes

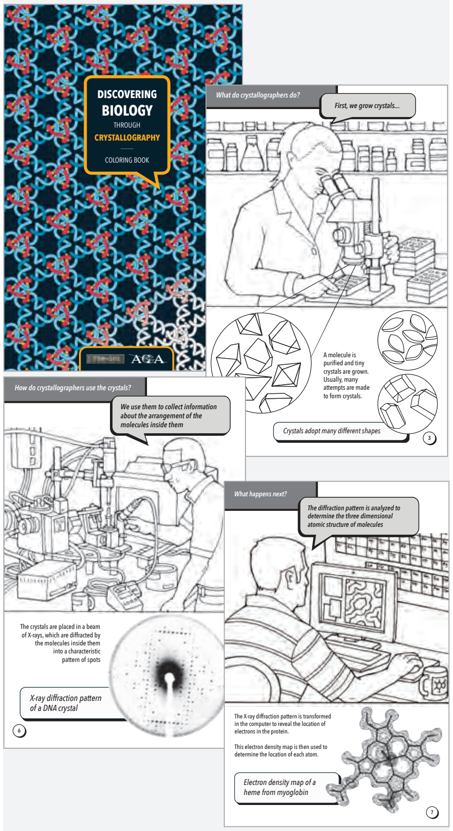
PDB-101 resource development is heavily focused on global health concerns. In 2016, curricular modules and other educational materials were made available to promote better understanding of insulin and diabetes at the molecular level.


Insulin in Action
Insulin 1 binds to the **insulin receptor** 2, then the protein kinase domain of the receptor activates a **signaling cascade** 3, mobilizing a number of different systems in the cell.
Glucose
1
2
3
4
5
6
These signaling proteins cause **glucose transporters** 4 to be moved to the surface of the cell, which can then transport glucose into the cell. Some of this glucose is used directly as a source of energy to power the cell.
Insulin signaling also activates **a set of enzymes** 5 that build **glycogen** 6 from excess glucose.
Another **set of enzymes** (not depicted) utilize excess glucose to create fatty acids, a longer term way of storing food energy.
Glycogen can be rapidly converted back into glucose when blood sugar levels fall between meals.
Scale 0 10nm 20nm
Insulin Action by David S. Goodsell
Sweet Signals: Insulin and Glucagon in Diabetes
utside Cell
INSULIN
Glucose

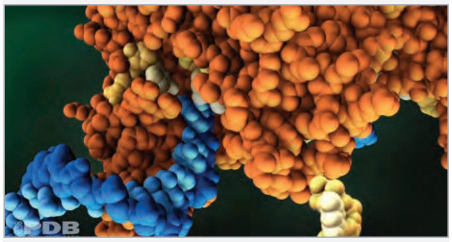
The insulin paper model (top) and the *Insulin and Diabetes* poster (fragment, middle) were some of the resources available to high school students participating in the 4th edition of the national video challenge *Structural Biology and Diabetes*. The winning video (bottom) from West Windsor-Plainsboro High School in NJ describes how insulin and glucagon regulate blood sugar levels.

Materials for Learning

A variety of materials to support exploration are available, including videos, posters, animations, and even a crossword puzzle.


DISCOVERING BIOLOGY THROUGH CRYSTALLOGRAPHY
COLORING BOOK
ACA
What do crystallographers do?
First, we grow crystals...
A molecule is purified and tiny crystals are grown. Usually, many attempts are made to form crystals.
Crystals adopt many different shapes
3
How do crystallographers use the crystals?
We use them to collect information about the arrangement of the molecules inside them
The crystals are placed in a beam of X-rays, which are diffracted by the molecules inside them into a characteristic pattern of spots
X-ray diffraction pattern of a DNA crystal
6
What happens next?
The diffraction pattern is analyzed to determine the three dimensional atomic structure of molecules
The X-ray diffraction pattern is transformed in the computer to reveal the location of electrons in the protein.
This electron density map is then used to determine the location of each atom.
Electron density map of a heme from myoglobin
7

A unique coloring book *Discovering Biology Through Crystallography* focused on the diverse 3D shapes and functions of biological molecules was created with grant support from the American Crystallographic Association and distributed to schools and science festivals. Bulk copies may be requested for use in outreach and education at **bit.ly/2m18avx**



New animation *Molecular Views of HIV Therapy* uses PDB structures to show how antiretroviral drugs inhibit essential viral enzymes.

ABOUT THE COVER

ZIKA VIRUS
DAVID S. GOODSELL

Zika virus is shown in cross section at center left. Visible on the periphery are envelope proteins (pink) and membrane proteins (magenta) embedded in a lipid membrane (light purple). Within the interior of the virus, the RNA genome (yellow) is associated with capsid proteins (orange).

Two viruses are shown interacting with cell surface receptors (green) and are surrounded by blood plasma proteins outside the cell.

This painting was recognized by the National Science Foundation (NSF) and *Popular Science* as one of the best science images of the year and selected as the "People's Choice" in the illustration category for the 2017 Vizzies.

REFERENCES

REFERENCES IN THIS REPORT

1. OneDep: Unified wwPDB System for Deposition, Biocuration, and Validation of Macromolecular Structures in the PDB Archive (2017)
   Structure **25**: 536–545.
   doi: 10.1016/j.str.2017.01.004

2. Multivariate Analyses of Quality Metrics for Crystal Structures in the Protein Data Bank Archive (2017)
   Structure **25**: 458-468.
   doi: 10.1016/j.str.2017.01.013

CITE THE PDB

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

2. 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

ABOUT RCSB PDB

RCSB PDB
PROTEIN DATA BANK

RCSB.ORG • INFO@RCSB.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.

MANAGEMENT

The 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

/RCSBPDB

/buildmodels

/RCSBProteinDataBank

/rcsb