Protein Data Bank:

An open access resource enabling basic and applied research and education in biology and medicine

John Westbrook, Ph.D.

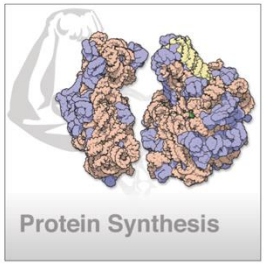
RCSB PDB Data & Software Architect Lead

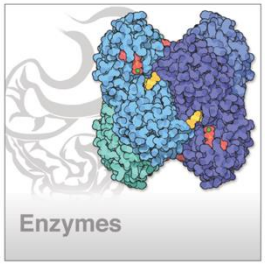
Overview

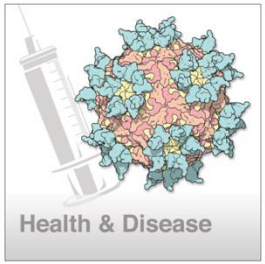
- A bit of background about the PDB
- PDB data content
- PDB data representation and data quality standards
- The PDB biocuration platform
- Some key features delivered by the RCSB PDB

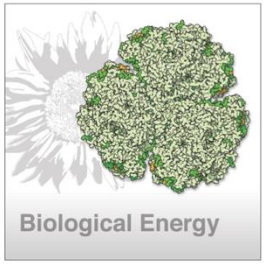
Protein Data Bank

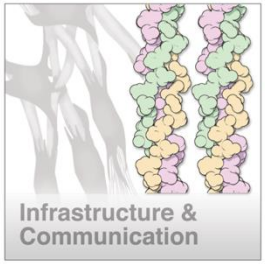
- First open access digital resource in biology (est. 1971 with 7 entries)
- Single global archive of 3-D macromolecular structures (contains >120,000 entries)
- Freely available to all at pdb.org
- US PDB headquartered at Rutgers/UCSD (NSF, NIH, DOE)
- US PDB part of Worldwide PDB with partners in EU and Japan

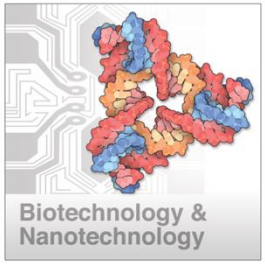










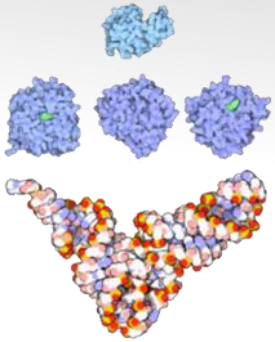
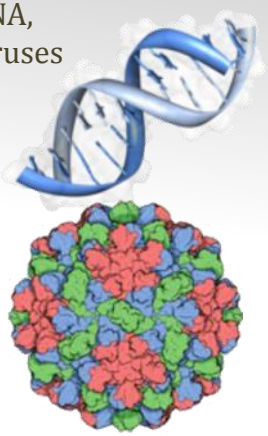
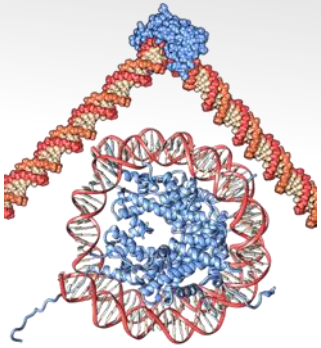
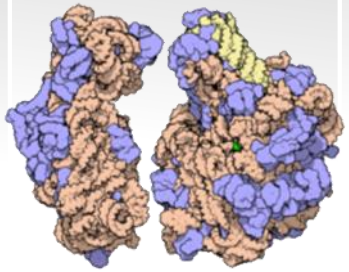
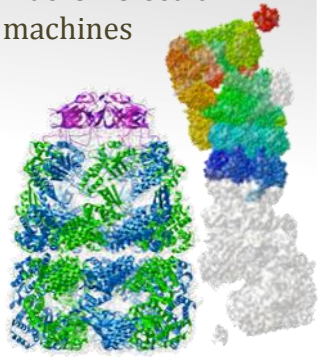
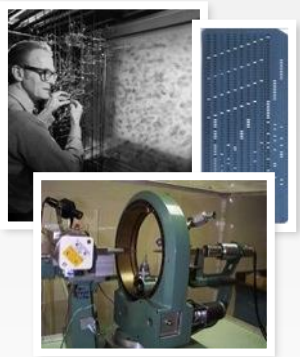
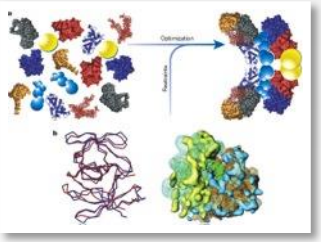
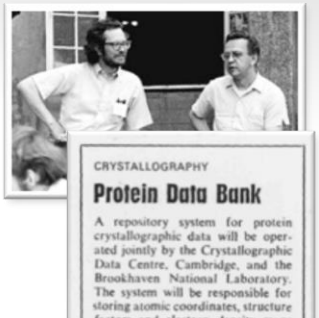
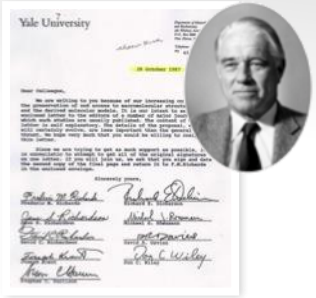
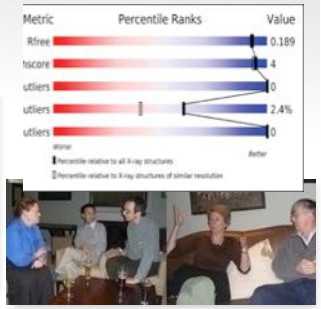


Worldwide Protein Data Bank

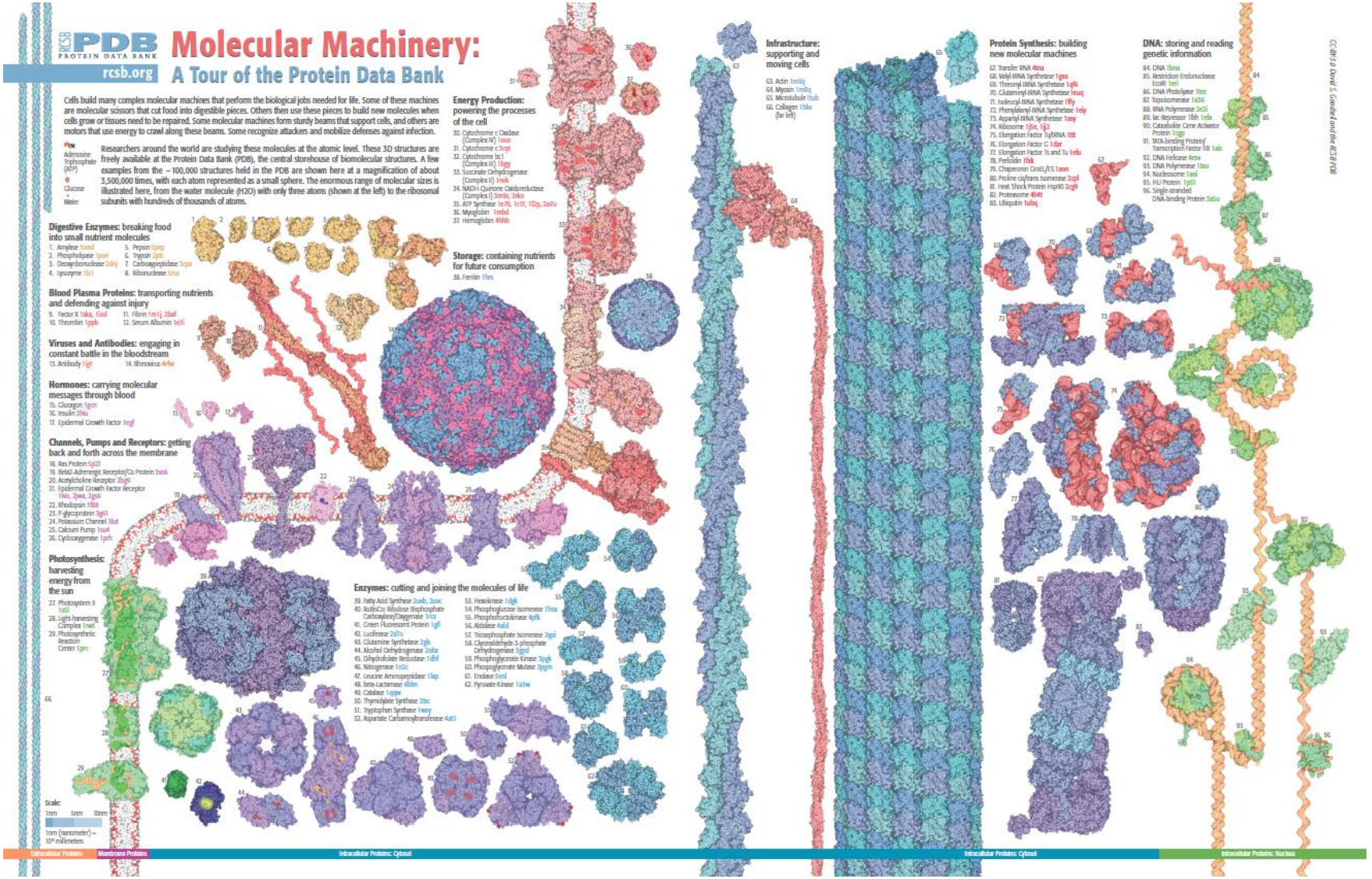
- Established in 2003, wwPDB ensures data are freely & globally available in a common repository
- Collaborate on data quality and representation standards, and tools and procedures for biocuration
- Each partner delivers different services and views of the common repository of data

wwPDB Advisory Committee Meeting, 2015

WORLDWIDE PDB PROTEIN DATA BANK | RCSB PDB PROTEIN DATA BANK | PDBe Protein Data Bank in Europe | PDBj Protein Data Bank Japan | BMRB

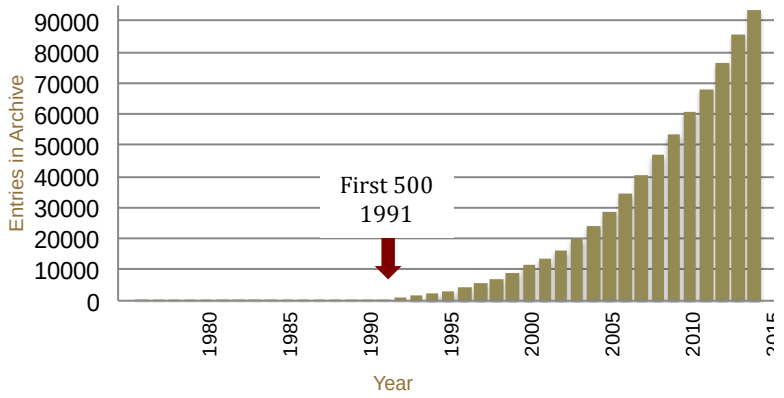
	1970s	1980s	1990s	2000s	2010s
Science	Small enzymes, RNA 	DNA, viruses 	Protein-DNA complexes 	Ribosomes 	Large macromolecular machines 
Technology	X-ray diffraction, diffractometers, punched cards 	Synchrotron radiation, computer graphics, NMR 	Electron microscopy, fast computers, fast detectors 	High throughput structural genomics, robots 	Hybrid methods 
Community	PDB archive established 	IUCr deposition guidelines 	Standardization, RCSB 	Experimental data required, wwPDB 	Validation standards 

Diverse Molecular Content of the PDB

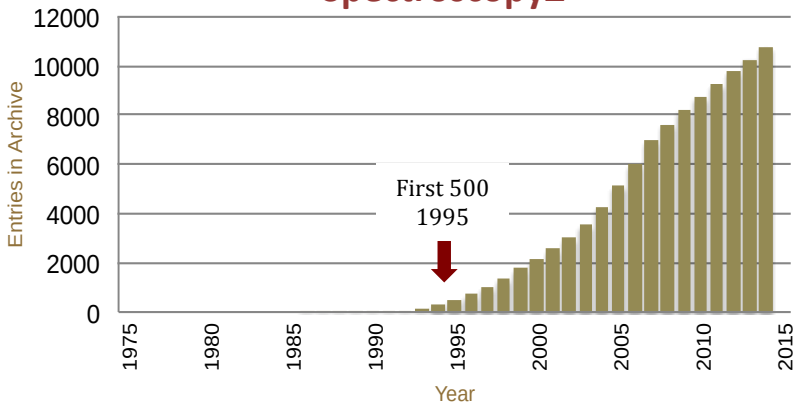


Supporting Rapidly Evolving Experimental Technologies

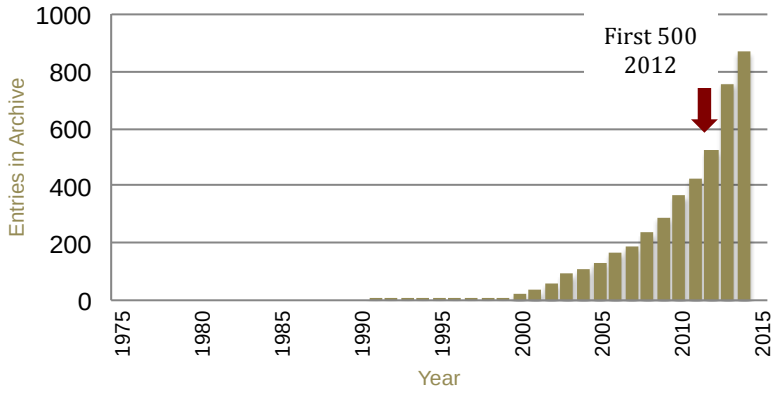
X-ray Crystallography



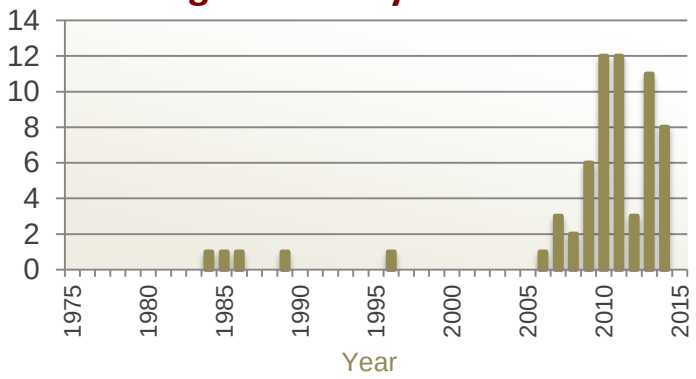
Nuclear Magnetic Resonance Spectroscopy



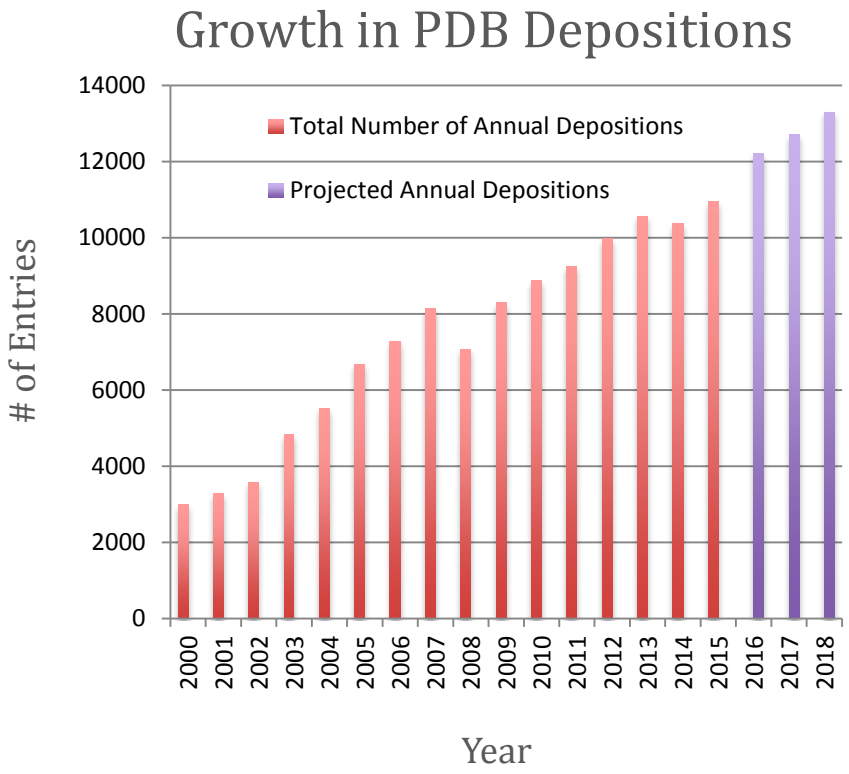
3D Electron Microscopy



Integrative & Hybrid Methods



PDB Growth and Data Usage



As of 2015, ~55% increase in the number of global depositions since 2008

PDB Depositors >800 new entries/month

PDB Users

FTP and RSYNC Download Traffic in 2015:
526 million downloads

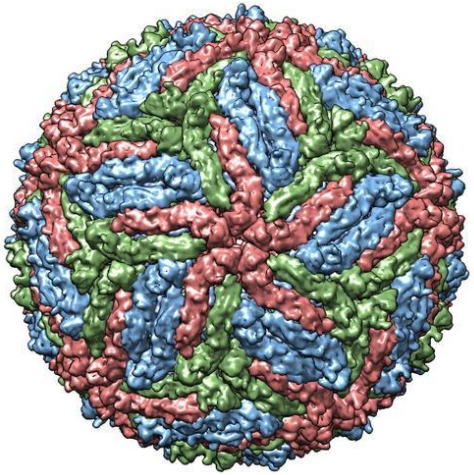
RCSB PDB
347 million

PDBe
100 million

PDBj
58 million

PDB Data Content

- Atomic coordinates and primary experimental data
- Sample composition and preparation details
- Protein and nucleic acid polymer sequences and taxonomy
- Small molecules (ligands)
- Experimental data collection, structure solution, and structure refinement
- Structure classification by sequence, function and other criteria
- Citation and references to related data resources
- PDB entries contain 250 – 1200+ unique items of data

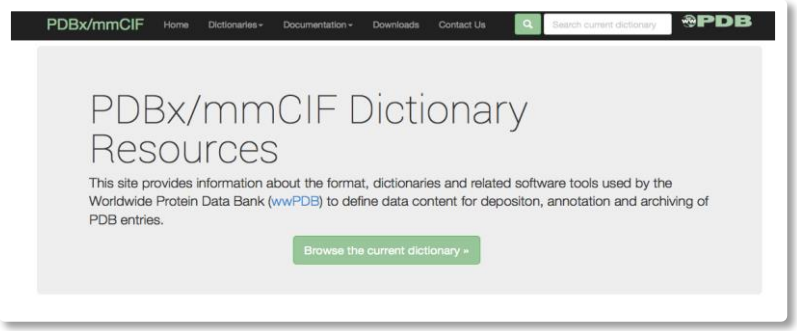


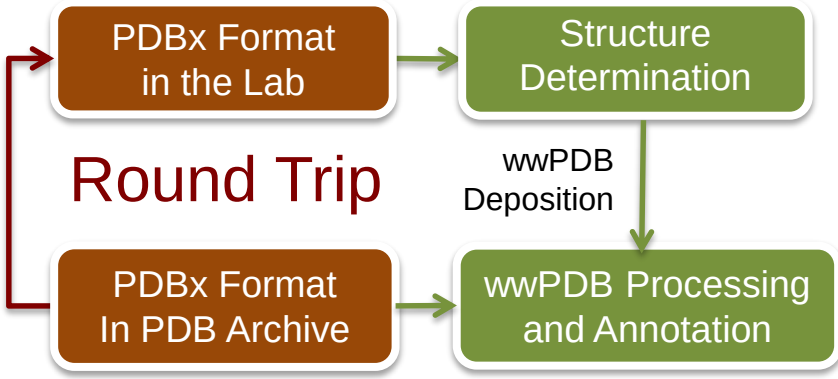
```
data_5IRE
#
_entry.id   5IRE
#
_em_entity_assembly.id            1
_em_entity_assembly.details
;Low passage Vero cells, derived from African gr
 were chosen for propagating the virus.
;
_em_entity_assembly.name          'Zika virus'
_em_entity_assembly.type          VIRUS
_em_entity_assembly.assembly_id   1
#
_em_imaging.id                                1
_em_imaging.entry_id                          5IRE
_em_imaging.accelerating_voltage              300
_em_imaging.electron_source                   'FIE
_em_imaging.illumination_mode                 'FLO
_em_imaging.microscope_model                  'FEI
_em_imaging.mode                              'BRI
_em_imaging.nominal_cs                        2.7
_em_imaging.nominal_defocus_max               2500
_em_imaging.nominal_defocus_min               1000
_em_imaging.nominal_magnification             14000
_em_imaging.recording_temperature_maximum     80
_em_imaging.recording_temperature_minimum     80
_em_imaging.specimen_holder_model             'FEI TITAN KRIOS AUTOGRID HOLDER'
_em_imaging.sample_support_id                 1

#
loop_
_entity_poly.entity_id
_entity_poly.type
_entity_poly.nstd_linkage
_entity_poly.nstd_monomer
_entity_poly.pdbx_seq_one_letter_code
_entity_poly.pdbx_seq_one_letter_code_can
_entity_poly.pdbx_strand_id
_entity_poly.pdbx_target_identifier
1 'polypeptide(L)' no no
;IRCIGVSNRDFVEGMSGGTWVDVVLEHGGCVTVMAQDKPTVDIELVTTTVSNMAEVRSYCYEASISDMASDSRCPTQGEA
YLDKQSDTQYVCKRTLVDRGWGNGCGLFGKGSLVTCAKFACSKKMTGKSIQPENLEYRIMLSVHGSQHSGMIVNDTGHET
DENRAKVEITPNSPRAEATLGGFGSLGLDCEPRTGLDFSDLYYLTMNNKHWLVHKEWFHDIPLPWHAGADTGTPHWNNKE
ALVEFKDAHAKRQTVVVLGSQEGAVHTALAGALEAEMDGAKGRLSSGHLKCRLKMDKLRLKGVSYSLCTAAFTFTKIPAE
TLHGTVTVEVQYAGTDGPCKVPAQMAVDMQTLTPVGRLITANPVITESTENSKMMLELDPPFGDSYIVIGVGEKKITHHW
HRSGSTIGKAFEATVRGAKRMAVLGDTAWDFGSVGGALNSLGKGIHQIFGAAFKSLFGGMSWFSQILIGTLLMWLGLNTK
NGSISLMCLALGGVLIFLSTAVSA
;
;IRCIGVSNRDFVEGMSGGTWVDVVLEHGGCVTVMAQDKPTVDIELVTTTVSNMAEVRSYCYEASISDMASDSRCPTQGEA
YLDKQSDTQYVCKRTLVDRGWGNGCGLFGKGSLVTCAKFACSKKMTGKSIQPENLEYRIMLSVHGSQHSGMIVNDTGHET
DENRAKVEITPNSPRAEATLGGFGSLGLDCEPRTGLDFSDLYYLTMNNKHWLVHKEWFHDIPLPWHAGADTGTPHWNNKE
```

The 3.8 angstrom resolution cryo-EM structure of Zika virus. Sirohi, D., Chen, Z., Sun, L., Klose, T., Pierson, T.C., Rossmann, M.G., Kuhn, R.J. (2016) Science 352: 467-470

Community Data Standards

- PDB manages data using the macromolecular extension to the Crystallographic Information Framework (mmCIF) originally developed as an IUCr data standard
- PDB coordinates the extension of the standard (PDBx/mmCIF) to support the broader needs of both contributors and users of the archive (> 4400 data terms)
- A PDBx/mmCIF Working Group of community experts and methods developers oversees the evolution of the standard and ensures that the standard is well supported by key community software tools.
- PDB hosts community workshops to support the data standard and maintains a web site serving PDBx/mmCIF data dictionaries, schema and software tools (mmcif.wwpdb.org)





Workshop Participants, September 2011

Workshop Participants, October 2014

PDBx/mmCIF Development Timeline

1991 1994 1997 2000 2003 2006 2009 2012

IUCr mmCIF Working Party

IUCr mmCIF Maintenance Group

Core CIF V1

mmCIF V1

mmCIF V2

mmCIF/Core sync'd

Workshops

York
Tarrytown
Brussels

Rutgers
St. Louis
Seattle

Glasgow

CARB
Rutgers

Honolulu
Orlando
EBI

DDL 1

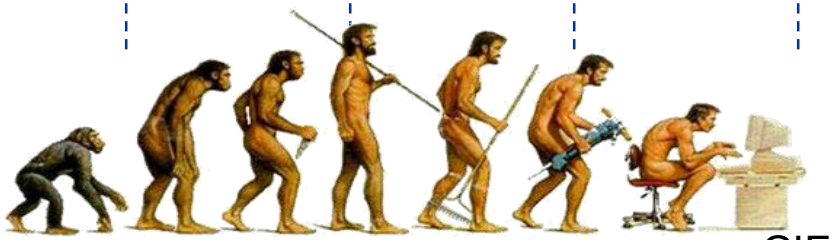
DDL 2

mmCIF +Extensions

PDB Exchange Dictionary

wwPDB
One Archive – One Dictionary

wwPDB Common Deposition & Annotation System



Data

mmCIF

Community Standards for Data Quality

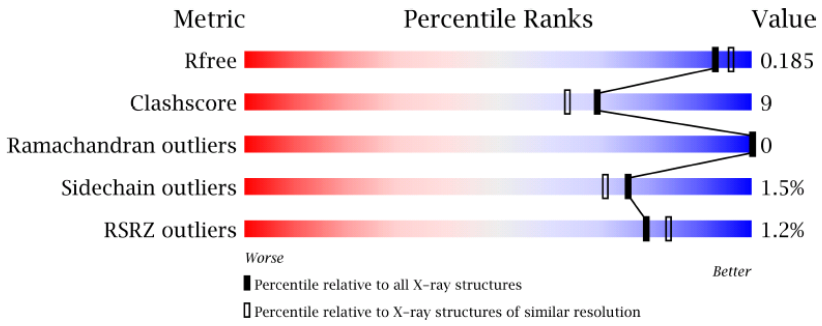
Method-specific Community Validation Task Forces have been convened to collect recommendations and develop consensus on data quality standards, identify software tools to perform required validation tasks, and to define related content requirements for archiving.

Task Force	Meeting/ Workshop	Chair(s)/Membership	Outcomes
X-ray Validation Task Force	2008 2015	Randy Read (Univ of Cambridge) 17 members	(2011) *Structure* 19: 1395-1412
NMR Validation Task Force	2009, 2011, 2013 (x2), 2015 2016	Gaetano Montelione (Rutgers) Michael Nilges (Institut Pasteur) 10 members	(2013) *Structure* 21: 1563-1570
3DEM Validation Task Force	2010	Richard Henderson (MRC-LMB) Andrej Sali (UCSF) 21 members	(2012) *Structure* 20: 205-214
Small-Angle Scattering Task Force	2011 2014	Jill Trewhella (Univ Sydney) 6 members	(2013) *Structure* 21: 875-881
Hybrid Methods Task Force	2014	Andrej Sali (UCSF), Torsten Schwede (Univ Basel), Jill Trewhella (Univ Sydney) 27 members	(2015) *Structure* 23: 1156-1167

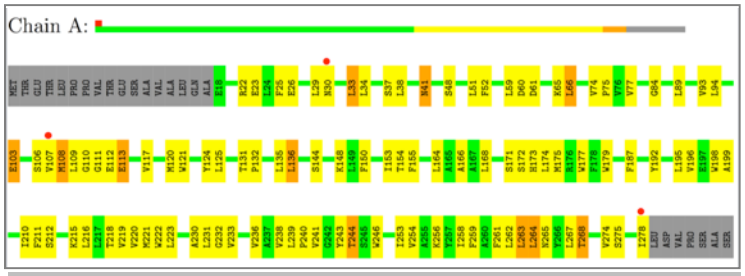
Presenting Data Quality to Diverse Audiences

- Provide relative and absolute quality metrics in graphical format
- Provide tabulations of key data, refinement statistics, and quality diagnostics
- Assess all macromolecular and ligand structural components
- PDF format reports can be uploaded with manuscript submission to a journal
- Diagnostics also delivered as an XML format data files

Overall Quality



Residue Plots

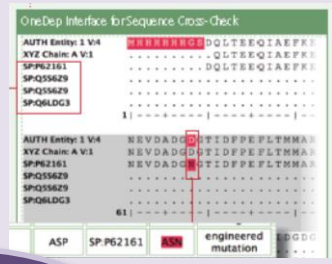


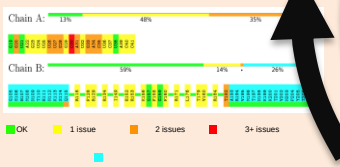
Grey – not modeled
Green, yellow, orange, red – 0,1,2, 3 or more issues
Red dot – poor fit to electron density

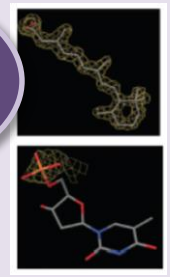
OneDep – The PDB Biocuration Platform

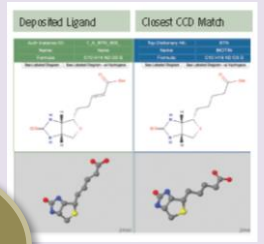
PDB OneDep: a unified global deposition, biocuration, and validation system

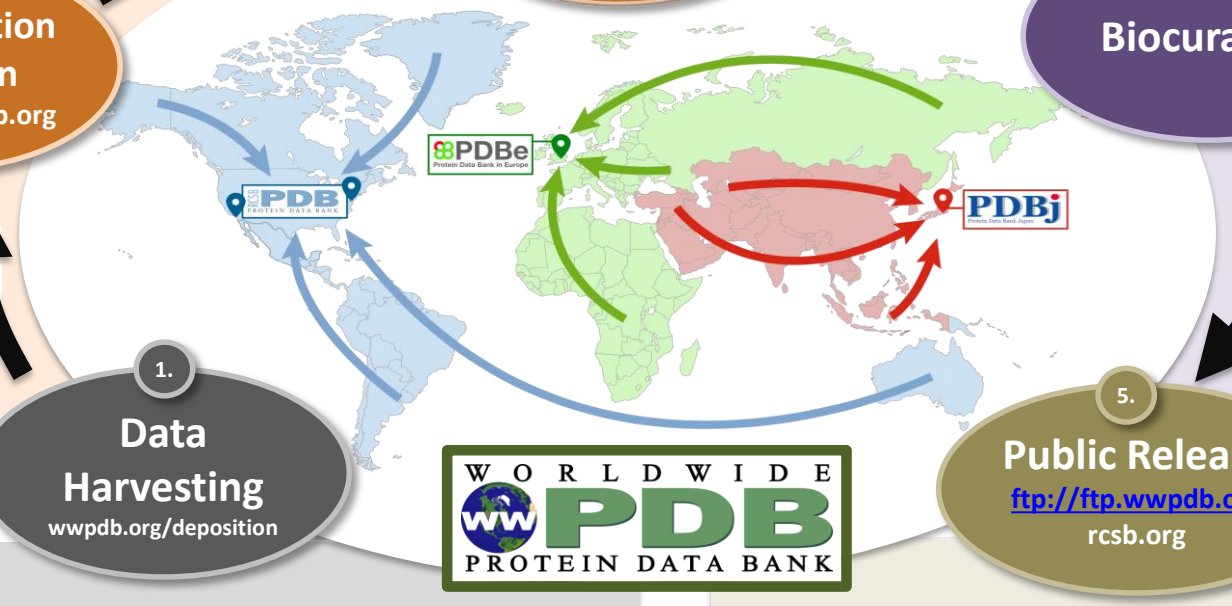
PDB
- Polymer check
- Ligand check
- Electron density fit











1. **Data Harvesting** wwpdb.org/deposition
2. **Pre-deposition Validation** validate.wwpdb.org
3. **Deposition** deposit.wwpdb.org
4. **Biocuration**
5. **Public Release** ftp://ftp.wwpdb.org rcsb.org

WORLDWIDE PDB PROTEIN DATA BANK

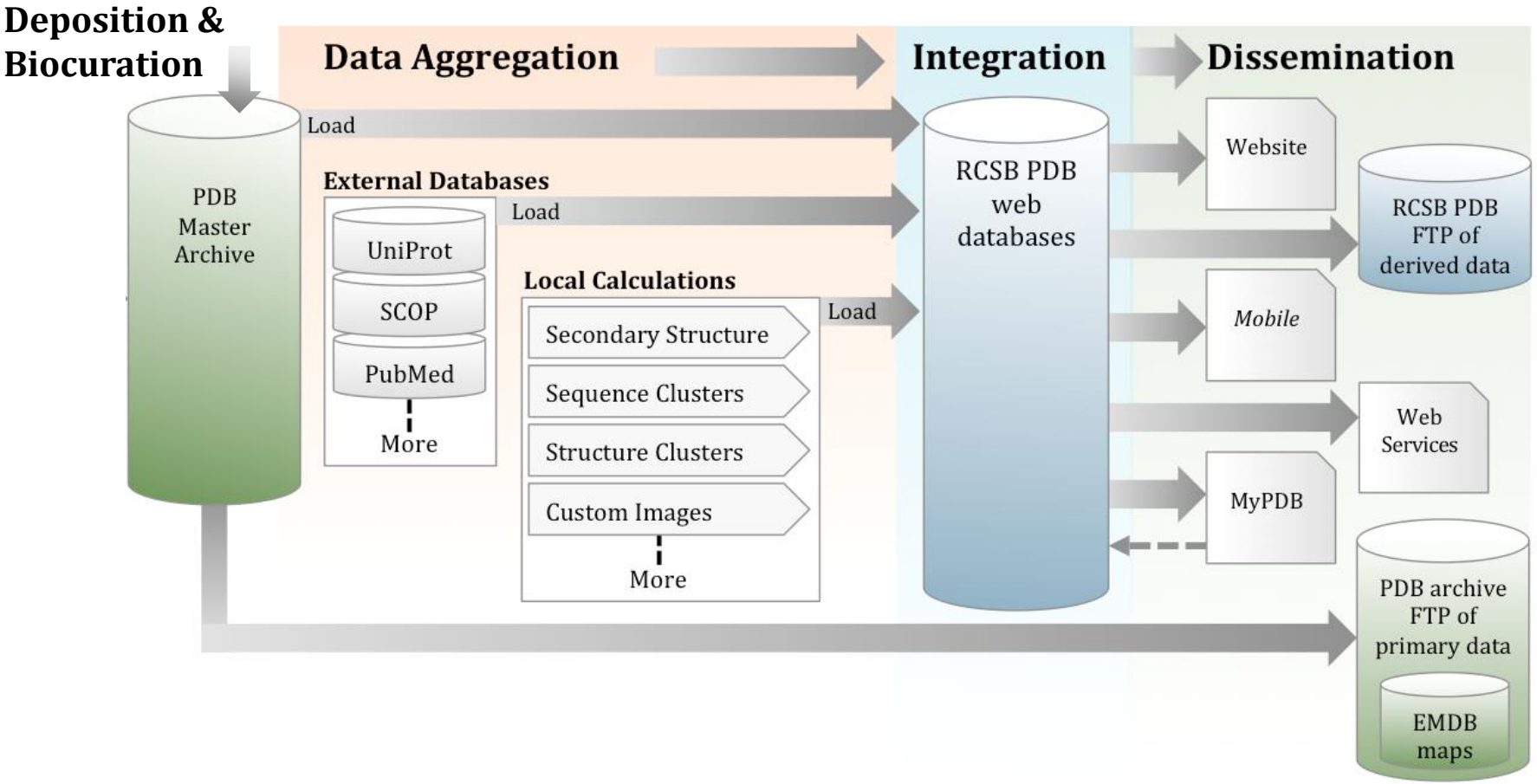
Data providers

- Generate atomic coordinate and experimental data files
- Assemble mandatory data items for deposition

Data Users

- Access data via web and ftp download and web services
- Enable other research for user community

RCSB PDB Data Delivery Pipeline



RCSB PDB Web Portal

- Launchpad for a wide range of functionalities
  - Deposit
  - Search
  - Analyze
  - Visualize
  - Tabulate
  - Download

http://rcsb.org

RCSB PDB Mobile App

- Provides convenient access to PDB data *on the go* with a minimal feature set
- Provides a browser, simple search, and an interactive 3D viewer
- Supports iPhone, iPad, and Android

http://www.rcsb.org/pdb/static.do?p=mobile/RCSBapp.html

Web Services (RESTful APIs)

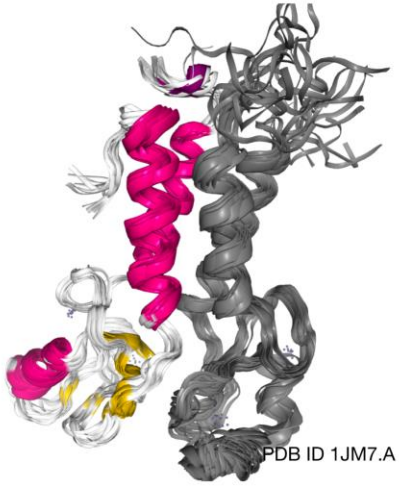
- Programmatic access to data: application-to-application communication
- Provides external workflows and analysis tools with direct access to a wide range of PDB data and services
- Enables integration of PDB data and services with programs and scripts in a variety of computer languages and computing environments

http://www.rcsb.org/pdb/software/rest.do

Enabling Data Access Through Integration and Visualization

- Protein Feature View: mapping protein sequence to 3D structure
- Gene View: mapping genome location to 3D structure
- Visualization
  - Browser native visualization tools using an efficient data compression protocol
  - Small molecule electron density and binding interactions
- Data integration resource files provided for download:
  - Correspondences with CCDC ligand structures
  - Sequence cluster data files
  - Phased release of data to support blinded molecular docking tests

Protein Sequence Integrated View



Organism: **Human** Length:1863 UniProt

Isoforms: **8**, currently showing only the 'canonical' sequence.

Gene View for BRCA1

Chromosome Location

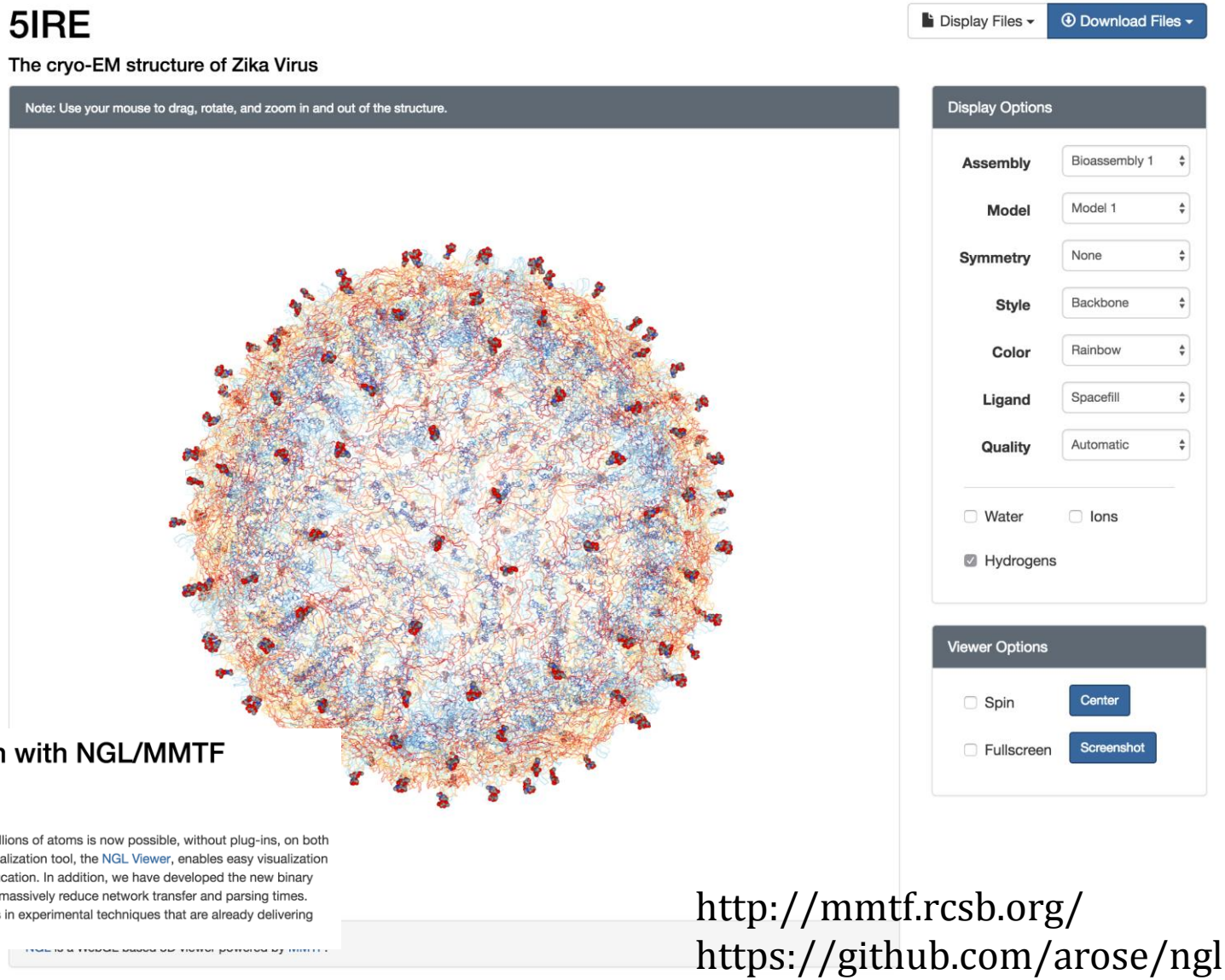
Other Gene names: RNF53

This protein in other organisms (by gene name):

P38398 - Homo sapiens 27

Genome Sequence Integrated View

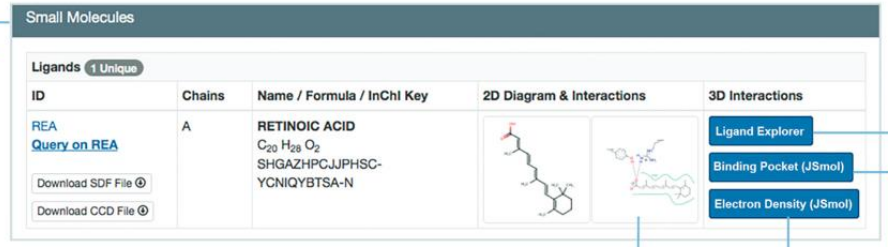
Molecular Visualization



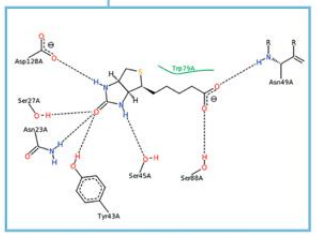
http://mmtf.rcsb.org/
https://github.com/arose/ngl

Visualizing Small Molecule Interactions

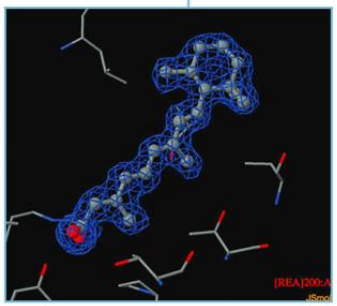
The **SMALL MOLECULES** section on entry's Structure Summary page offers tools for ligand visualization in 2D and 3D.



POSEVIEW 2D diagrams show a ligand and interacting residues. Black dashed lines indicate hydrogen bonds, salt bridges, and metal interactions. Green solid line show hydrophobic interactions and green dashed lines show π-π and π-cation interactions.



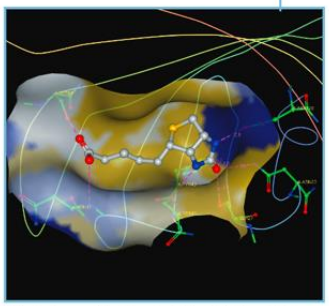
*Visualization of Biotin bound to Streptavidin PDB ID: 1STP with PoseView (**poseview.zbh.uni-hamburg.de**)*



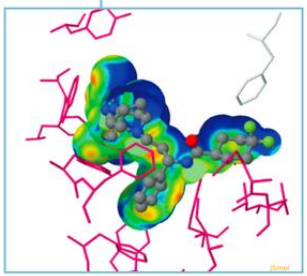
Electron density mini-map for Retinoic Acid in PDB ID 1CBS at 1sigma contour level. In this example, the ligand fits well into the density.

View sigma-weighted 2m|Fo|-d|Fc| **ELECTRON DENSITY "MINI-MAPS"** for bound ligands using the JSmol 3D View. This feature is available for ligands with more than one atom (ions excluded) in PDB entries with structure factor data. Different sigma values can be selected.

BINDING POCKET (JSmol) offers a 3D view of residues contact with the selected ligand. A cropped ligand van der Waals surface will appear that is color coded by the proximity to the van der Waals surface of the binding site (red: close contact, blue: far contact).



Biotin bound to Streptavidin PDB ID: 1STP



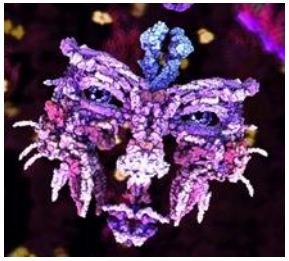
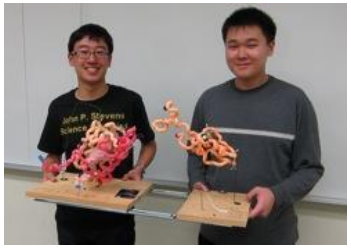
Maraviroc with its receptor C-C chemokine receptor type 5 PDB ID: 4MBS

LIGAND EXPLORER is a Java Web Start application with options to display hydrogen bonds, hydrophobic contacts, water mediated hydrogen bonds, metal interactions, and surfaces that can be color-coded using various parameters.

Reaching Diverse User Communities

Who are our users?	What are they using?
Biologists: structural biology, biophysics, biochemistry, genetics, Immunology, pharmacology, cell and molecular biology …	RCSB PDB website, deposition tools, data
Other scientists: bioinformatics, software developers, …	Web Services, search engines, data
Students & teachers	PDB-101
Media: Writers, textbook authors, patient advocacy groups, …	Images, data, information, outreach material, e.g., posters
General public: Curious/interested individuals, artists, sculptors, …	Images, *Molecule of the Month*, information from external media

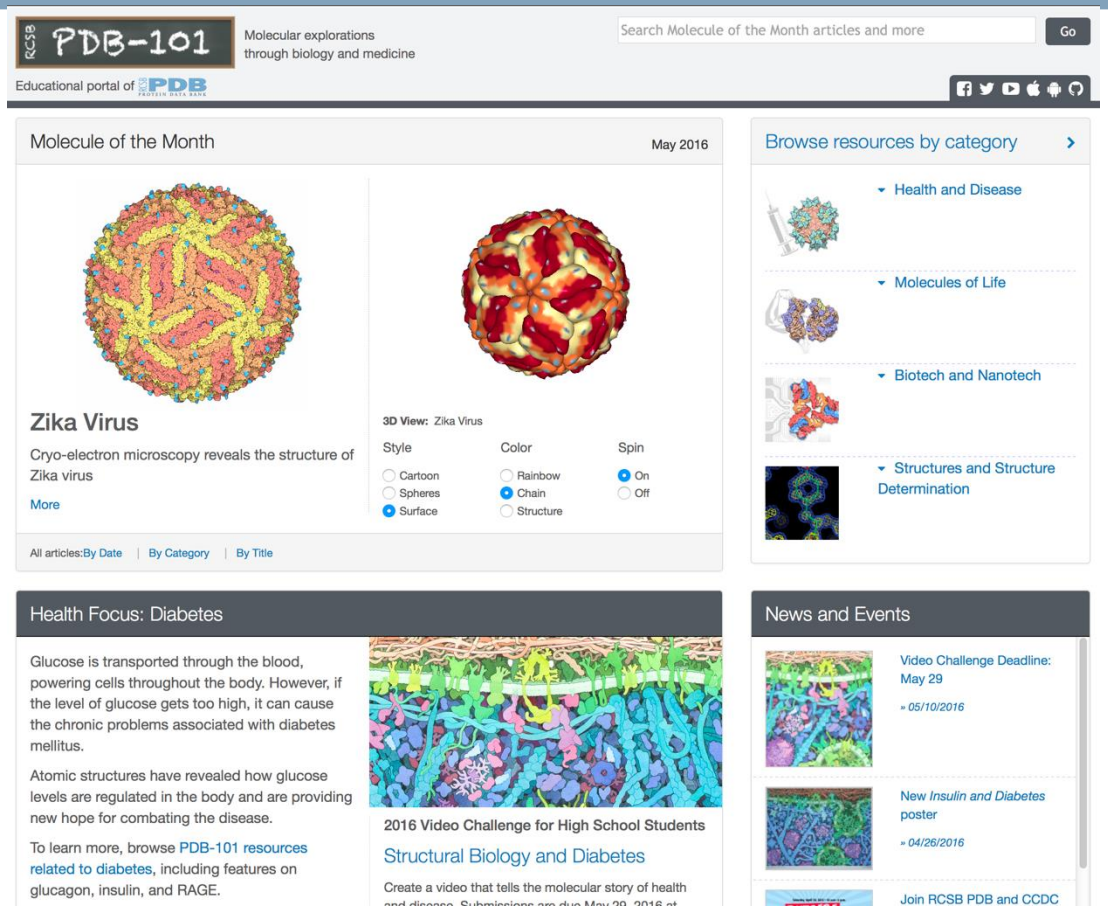


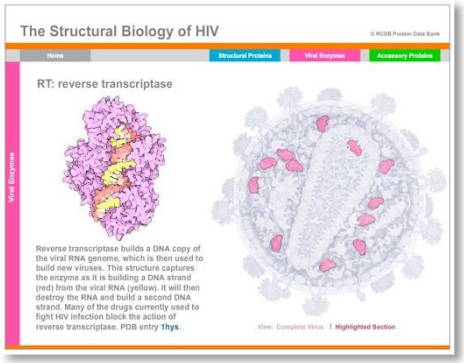
Online Educational Resources

Resources to help understand biology at the molecular level

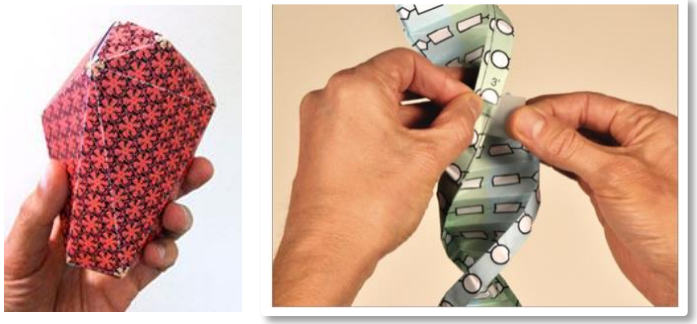
http://pdb101.rcsb.org/



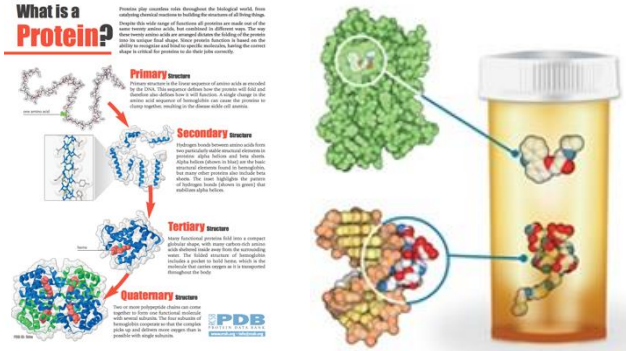
Animations



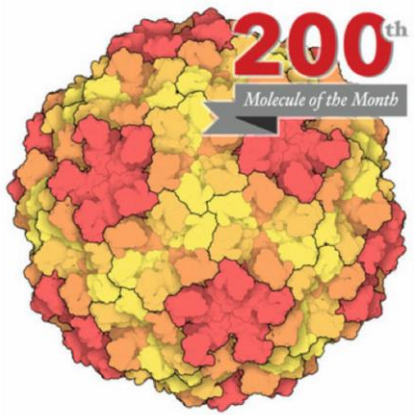
Paper Models



Posters



Molecule of the Month



PDB-101 | Molecule of the Month | Browse | Learn | Teach | Events | Geis Archive | More

RCSB PDB-101 Molecular explorations through biology and medicine

Search Molecule of the Month articles and more | Go

Educational portal of PDB

Molecule of the Month — By Category · By Date · By Title

Quasisymmetry in Icosahedral Viruses

Viruses use quasisymmetry to build large capsids out of many small subunits

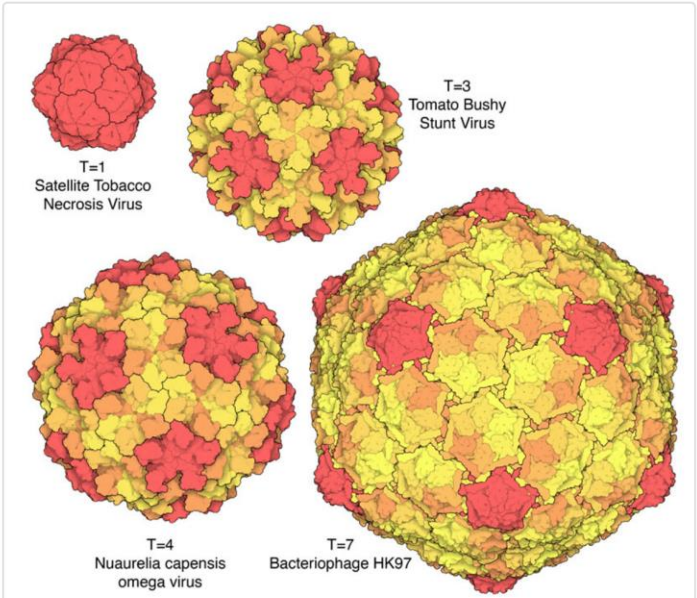
In the 1950s, before any virus structures had been determined, Watson and Crick predicted that viruses are symmetrical assemblies of many small subunits. They based this proposal on a puzzling observation: only a small amount of RNA (or DNA) can fit inside a tiny virus, but this RNA needs to encode the sequences for the proteins comprising the virus. This problem is particularly severe for the smallest viruses, such as satellite tobacco necrosis virus (shown here from PDB entry 2buk). Its capsid is just big enough to hold a single strand of RNA that encodes only one protein: the capsid subunit. It's called a "satellite" virus because it can't infect a cell on its own. Instead, it relies on other viruses, which infect the cell at the same time, to provide the necessary viral machinery for its reproduction.

Quasisymmetry

Viruses typically need to be larger than these tiny satellite viruses, so that they can encode additional viral proteins. One way they do this is by using *quasisymmetry*. In perfectly icosahedral viruses, the subunits form pentamers that assemble at the vertices of a small icosahedron. In quasisymmetrical viruses, a single type of subunit plays several different roles. They form pentamers that assemble at the vertices, but they also assemble in a slightly different way to form hexamers that fill out the faces of larger icosahedrons. In the early 1960s, Caspar and Klug discovered that there are a few systematic ways to arrange these pentamers and hexamers, based on a *triangulation number* that quantifies the number of slightly different shapes that the subunit needs to adopt in the final capsid.

Triangulation Numbers

Perfectly symmetrical viruses have a triangulation number T=1, because all of the subunits are identical. In satellite tobacco necrosis virus, this is just big enough to hold an RNA strand 1239 nucleotides long. Tomato bushy stunt virus (PDB entry 2tbv) has the next larger triangulation number, T=3, with pentamers on the vertices of the icosahedron and a hexamer centered on each face. It can hold a larger genome with 4776 nucleotides that encodes five proteins, including a viral replicase. The T=4 capsid of Nudaurelia capensis omega virus (PDB entry 1ohf) has hexamers centered on each edge of the icosahedron, and holds two strands of RNA that encode replicase and capsid proteins. Bacteriophage HK97 (PDB entry 1ohg) has T=7 quasisymmetry, and is large enough to enclose a DNA strand with 39,732 nucleotides that encodes 61 proteins.



Icosahedral virus capsids. In each virus, all of the subunits are chemically identical, but they adopt a few different quasisymmetrical shapes, each colored differently here. Pentamers of subunits are colored red, and hexamers of subunits are colored in shades of yellow and orange.

Download high quality TIFF image

PDB Management

PDB members past and present at the PDB40 Anniversary Symposium, 2011

The Protein Data Bank Archive is managed by:

WORLDWIDE PDB PROTEIN DATA BANK

wwpdb.org

Worldwide Protein Data Bank

Members

RCSB PDB PROTEIN DATA BANK

rcsb.org

RCSB Protein Data Bank

@buildmodels

Funding:
NSF, NIH, DOE

PDBe Protein Data Bank in Europe

pdbe.org

proteindatabank

@PDBEurope

Funding:
EMBL-EBI, Wellcome Trust, BBSRC, NIGMS, EU

PDBj Protein Data Bank Japan

pdbj.org

ja-jp.facebook.com/PDBjapan

@PDB_ja

Funding:
NBDC-JST

BMRB

bmrb.wisc.edu

Funding:
NLM