RCSB PDB
PROTEIN DATA BANK

RCSB Protein Data Bank Advisory Committee Meeting

November 1st 2016

rcsb.org

Welcome and Introductions

- RCSB Protein Data Bank Advisory Committee
- Funding Representatives
  - NSF
  - NIH
  - DoE
  - HHMI
- RCSB Protein Data Bank Leadership Team
  - Rutgers, The State University of New Jersey
  - University of California San Diego

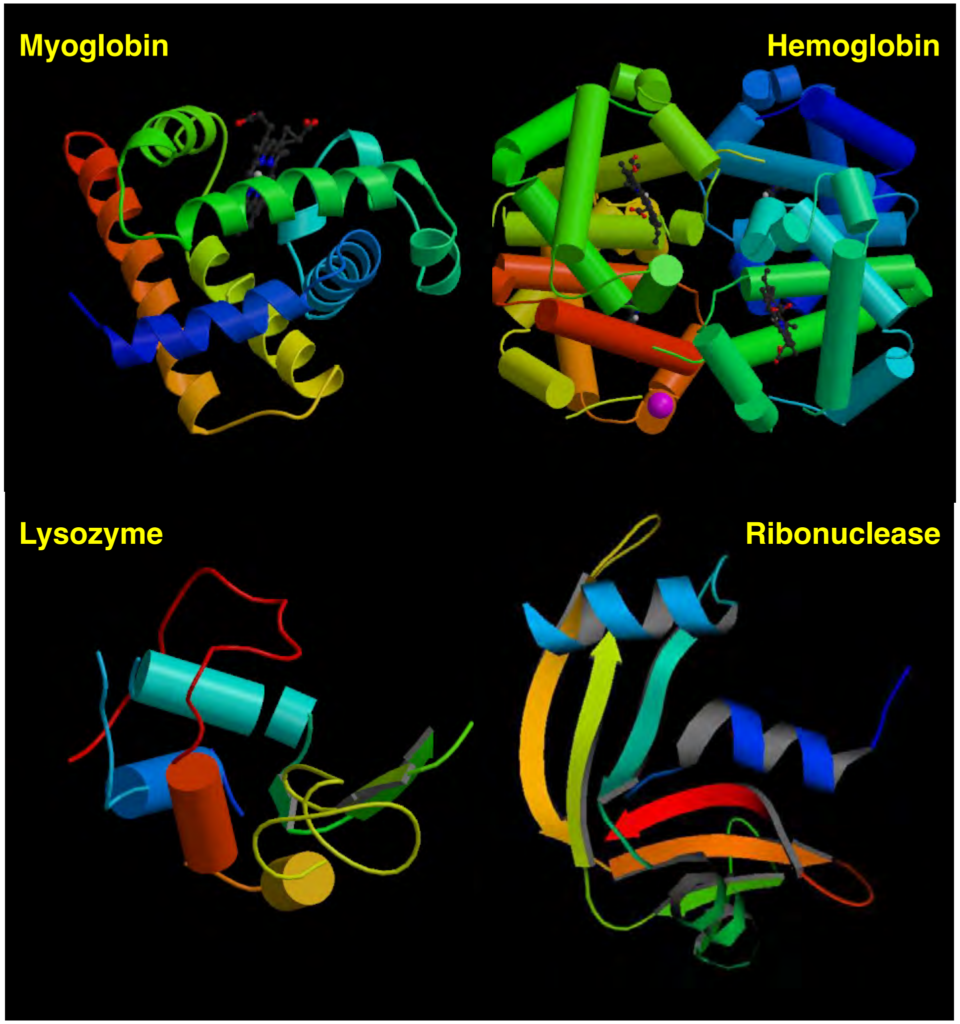
RCSB PDB
PROTEIN DATA BANK

Protein Data Bank Overview

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

Protein Data Bank Archive

- Single primary data archive for 3D structures of proteins, DNA, and RNA
- Established 1971 as 1st Global Open Access digital data resource in biology at Brookhaven (→RCSB PDB 1999)
- Since 2003, managed by an international partnership (wwPDB)

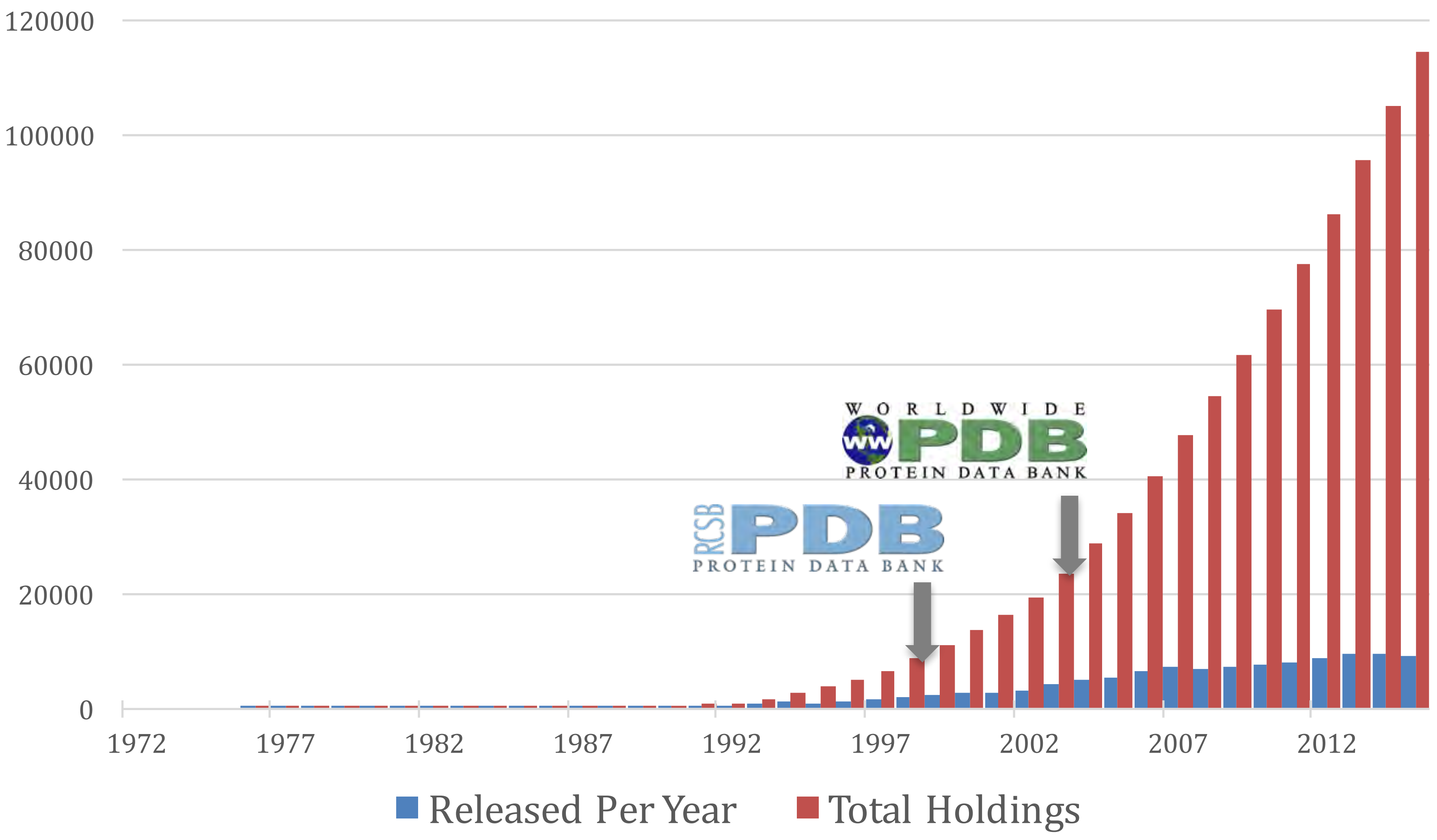


Some of the very first structures in the PDB

PDB Commitment to Scientific Data Reuse

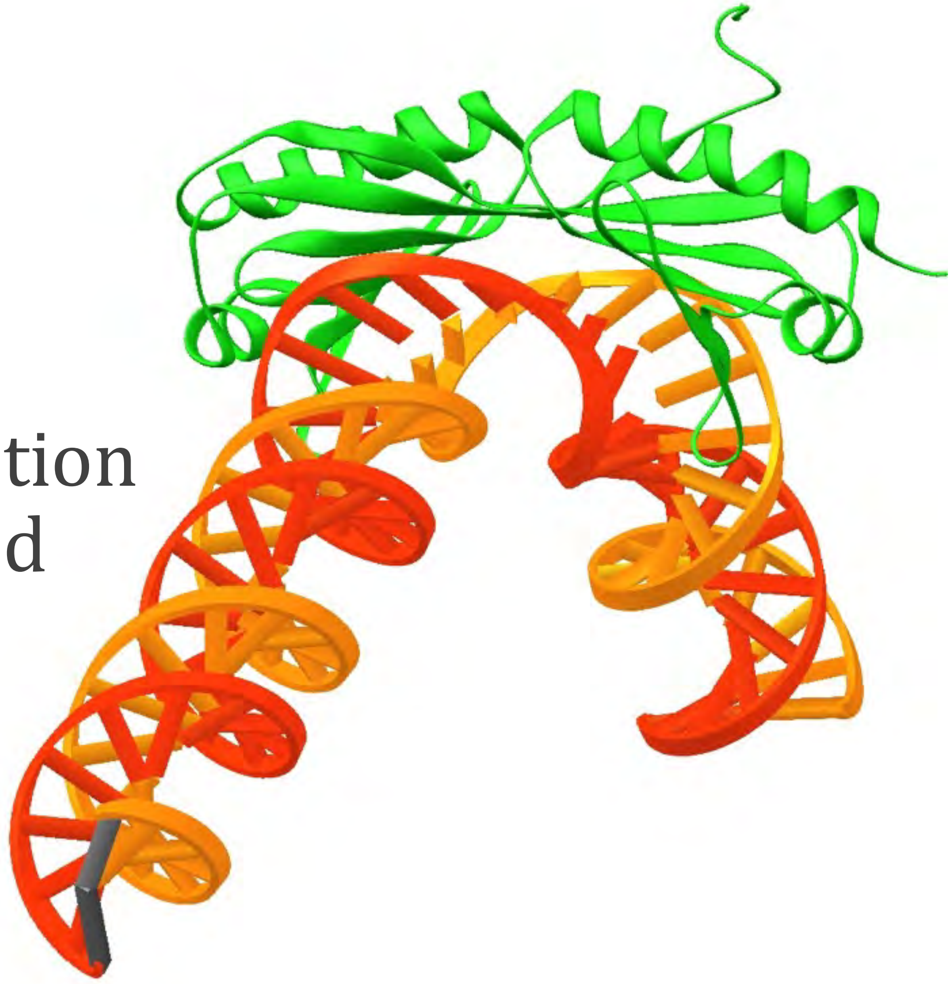
- Follow the *FAIR Guiding Principles for scientific data management and stewardship*
  - **F**indability
  - **A**ccessibility
  - **I**nteroperability
  - **R**eusability
- See Wilkinson *et al.* (2016) *Scientific Data* doi: 10.1038/sdata.2016.18

Released Entries 1971-2015



Function Follows Form in Biology

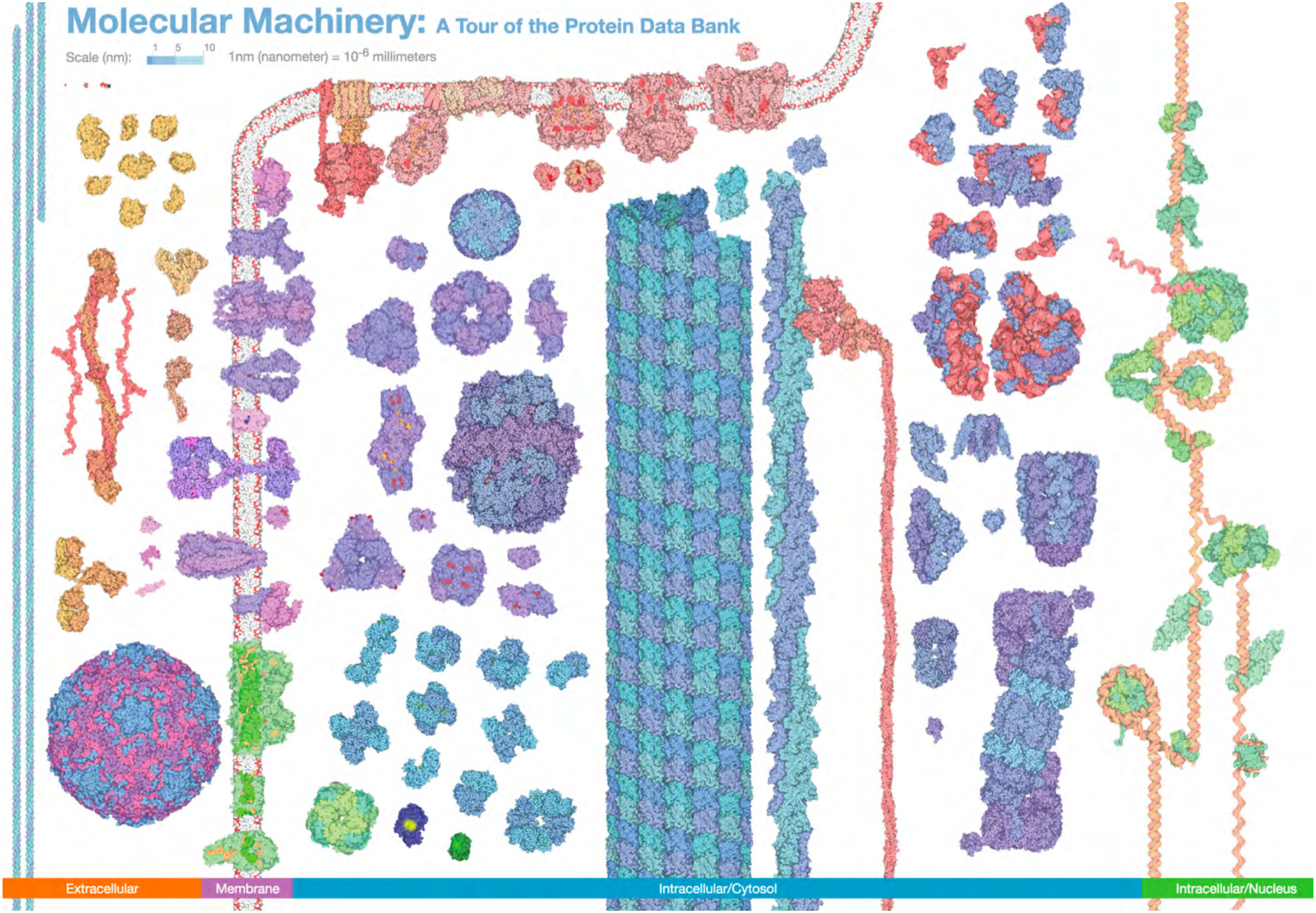
- 3D structure determines biological/biochemical function
- PDB data inform every area of research and education in biology, basic and applied
- PDB data are used every day to understand health and disease
- PDB data central to drug discovery



Arabidopsis thaliana
TATA-box Binding Protein + DNA (PDB 1VTL)

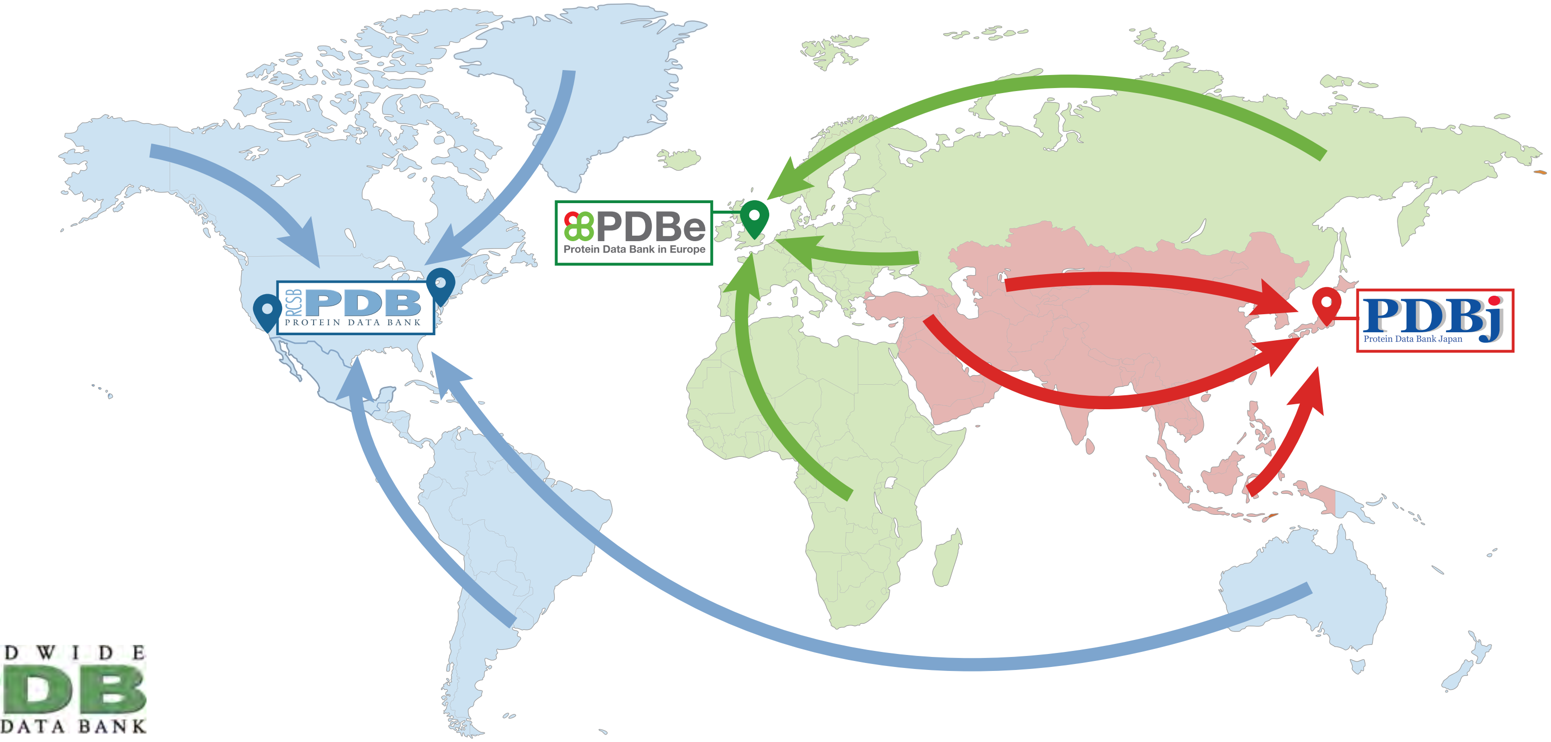
Molecular Machinery: A Tour of the Protein Data Bank

Scale (nm): 1 5 10 1nm (nanometer) = 10^{-6} millimeters



Worldwide Protein Data Bank (wwPDB)

- Structure data are globally produced/consumed
- Regional Data Centers: RCSB PDB (US), PDBj (Asia), PDBe (EU); BioMagResBank (US/Japan)
- Unfunded; Operated under formal MOU since 2003



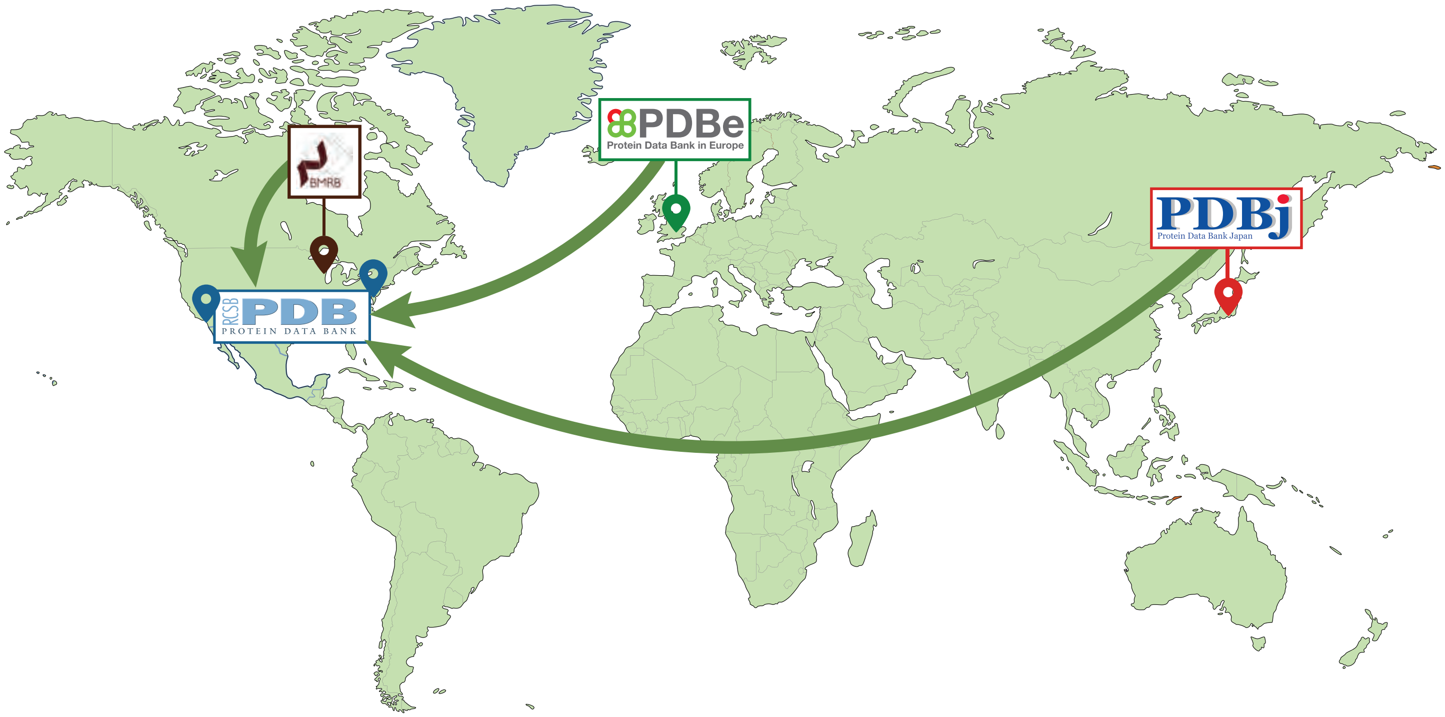
WORLDWIDE PDB PROTEIN DATA BANK

wwPDB Data Centers

- Ensure unrestricted PDB access worldwide
- Work with the scientific community to establish common data standards and best practices
- Collaborate on Global "Data In" Services: Deposition/Biocuration/Validation
- Operate identical FTP data distribution sites
- Develop/provide complementary Global Services for "Data Out"

RCSB PDB is the PDB Archive Keeper

- Support data security and global disaster recovery
- Ensure data uniformity and consistency
- Coordinate weekly updates and FTP distribution



Cost of Replicating the PDB Archive

- Data integrity and security are of paramount importance to the wwPDB partnership
- Estimated cost of replicating each PDB entry ranges from US$50,000 to > US$250,000
- Cost of replicating the PDB archive **US$12 billion** (assuming <unit cost>=US$100,000)

- Absent PDB data sharing, structural biology would never have reached current heights

What Has the PDB Archive Enabled?

- Reproducibility and Secure Storage
- Accelerated structure determination technologies
- Understanding evolution in 3D
  - Structure classification and prediction
- Structure-based drug discovery
- Functional understanding of Biology at molecular and atomic levels

Antigen Presenting Cell meets the T-cell
PDB 2CKB, Garcia *et al.* (1998)

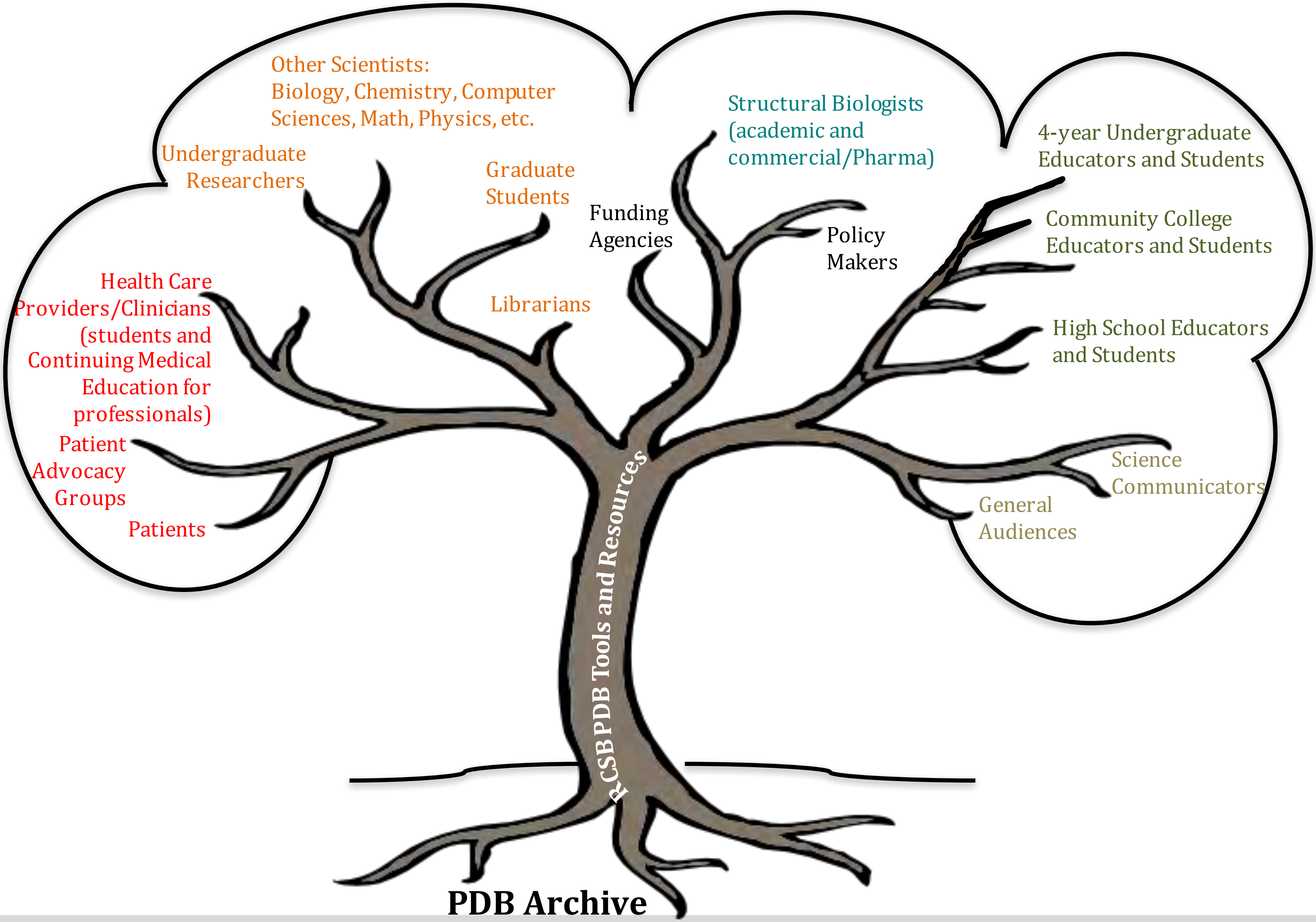
PDB Archive Facts and Figures

- Archival Contents
  - ~124,000 Structures Released since 1971
  - ~11,000 New Structures Deposited/Year
- Global User Base
  - ~30,000 Depositors Worldwide
  - >1 Million Unique Visitors/Year
    from 192/195 UN-recognized sovereign nations
- Impacts all of Biology and Medicine
  - >500 Million Data Files Downloaded/Year
  - ~1.5 Million Data Files Downloaded/Day
  - >200 derived data resources repackage PDB data

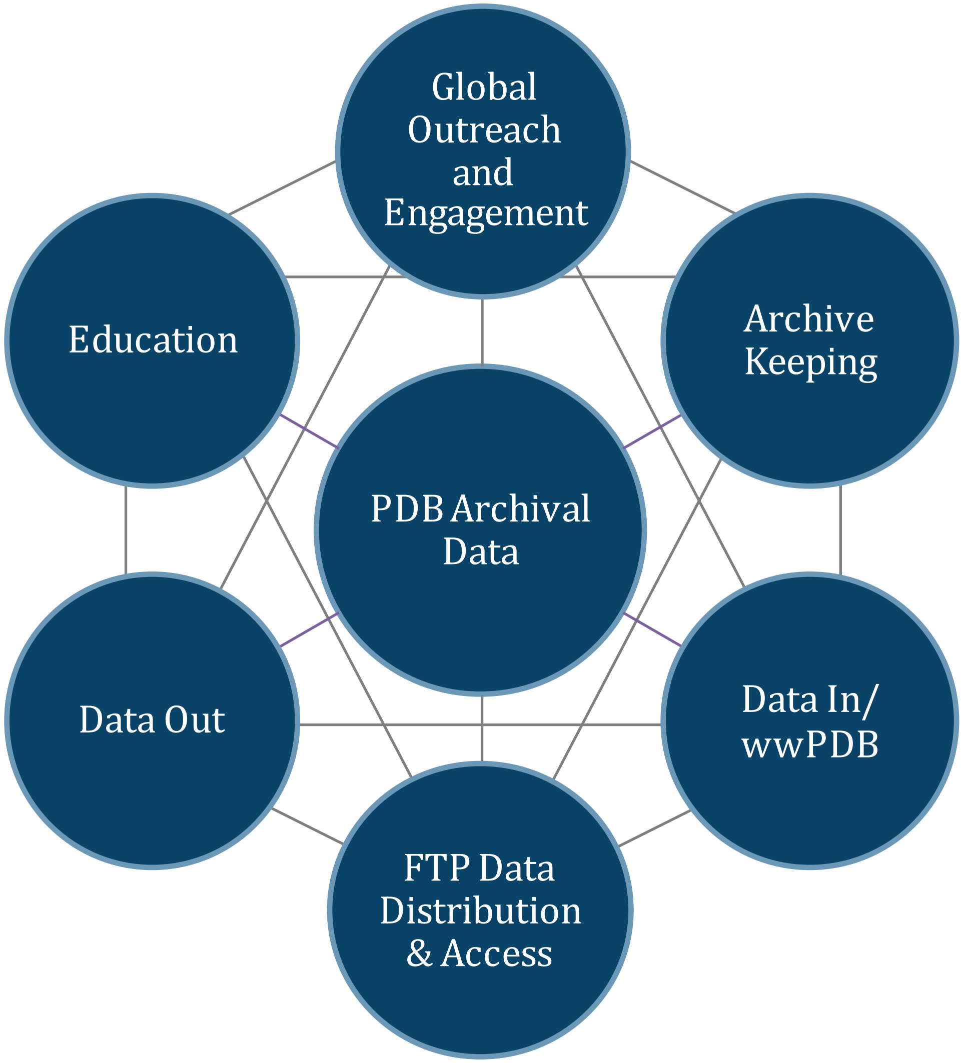
RCSB PDB: US wwPDB Data Center

- Established 1999 (RUTGERS | UC San Diego SDSC)
- Founded wwPDB in 2003 to support Data Producers
- Collaborates with international experts and resources to support Data Consumers
- Core activities funded by NSF (DBI-1338415), NIH, DOE
- Competes for additional funding for value-added activities

RCSB PDB Serves Diverse Stakeholders


Other Scientists:
Biology, Chemistry, Computer Sciences, Math, Physics, etc.
Undergraduate Researchers
Graduate Students
Funding Agencies
Structural Biologists (academic and commercial/Pharma)
4-year Undergraduate Educators and Students
Policy Makers
Community College Educators and Students
Health Care Providers/Clinicians (students and Continuing Medical Education for professionals)
Librarians
High School Educators and Students
Patient Advocacy Groups
Patients
Science Communicators
General Audiences
RCSB PDB Tools and Resources
PDB Archive

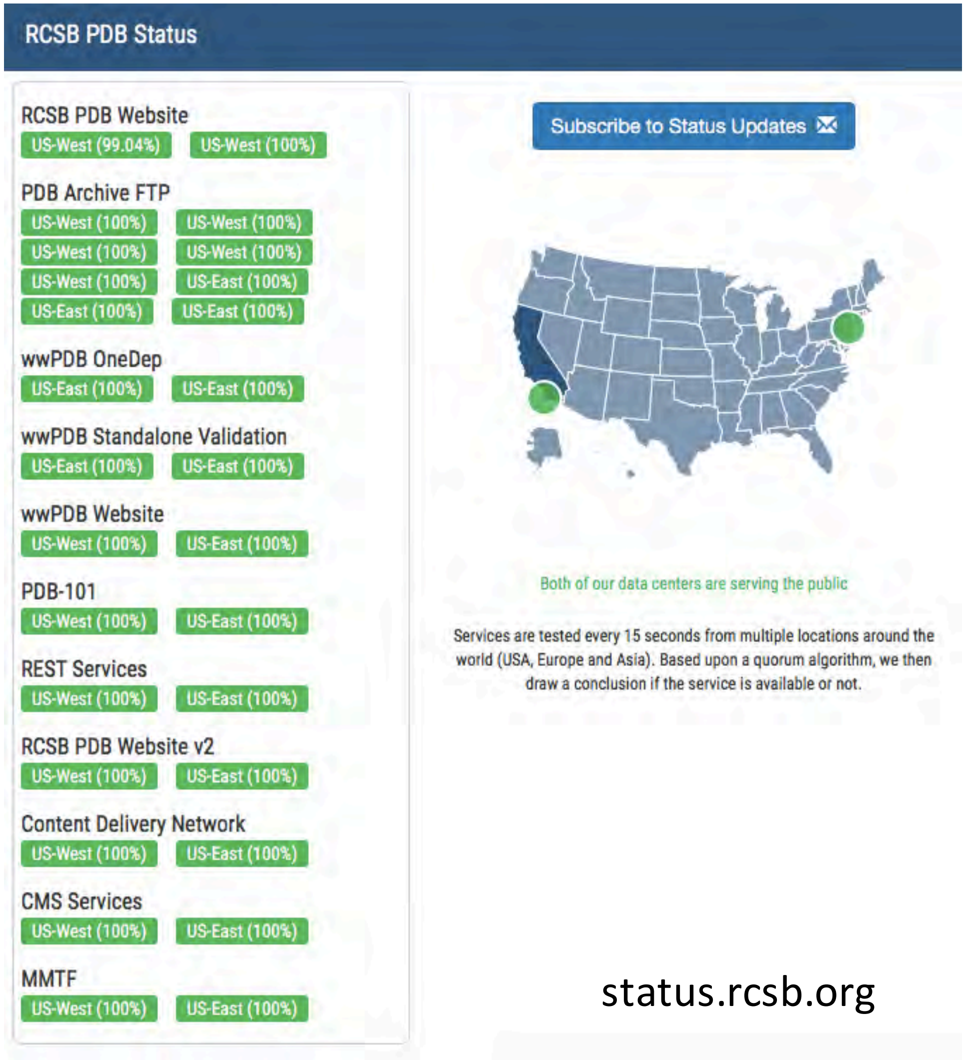
RCSB PDB Core Responsibilities



Funding for Core Activities: NSF [DBI-1338415], NIH, DOE

Archive Keeping

- PDB Archive FTP sites 24/7/366 availability with security and disaster recovery
- Failover and load balancing between Rutgers and UCSD
- Continuous global monitoring



RCSB PDB Status

RCSB PDB Website
US-West (99.04%) US-West (100%)

PDB Archive FTP
US-West (100%) US-West (100%)
US-West (100%) US-West (100%)
US-West (100%) US-East (100%)
US-East (100%) US-East (100%)

wwPDB OneDep
US-East (100%) US-East (100%)

wwPDB Standalone Validation
US-East (100%) US-East (100%)

wwPDB Website
US-West (100%) US-East (100%)

PDB-101
US-West (100%) US-East (100%)

REST Services
US-West (100%) US-East (100%)

RCSB PDB Website v2
US-West (100%) US-East (100%)

Content Delivery Network
US-West (100%) US-East (100%)

CMS Services
US-West (100%) US-East (100%)

MMTF
US-West (100%) US-East (100%)

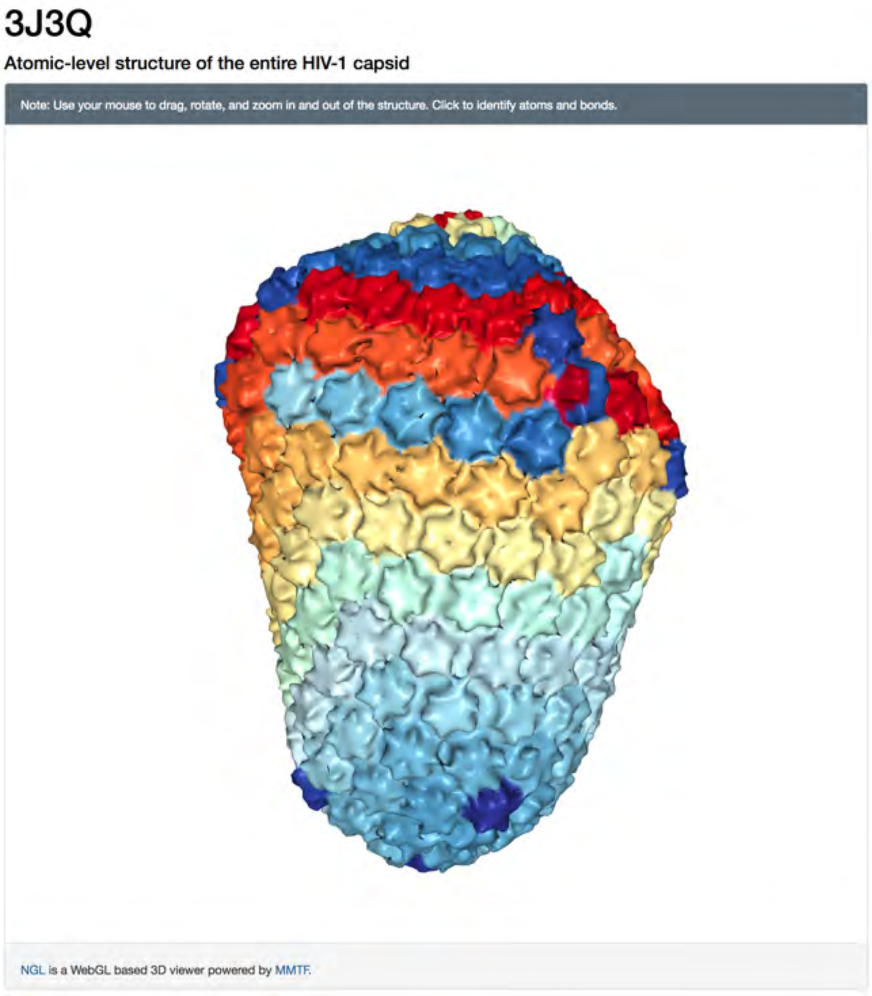
Both of our data centers are serving the public

Services are tested every 15 seconds from multiple locations around the world (USA, Europe and Asia). Based upon a quorum algorithm, we then draw a conclusion if the service is available or not.

status.rcsb.org

Data In/wwPDB

- Deployed unified global deposition system (OneDep)
- Validation for the entire PDB archive
- Facile management of very large structures (HIV capsid: 2.4 MM atoms)



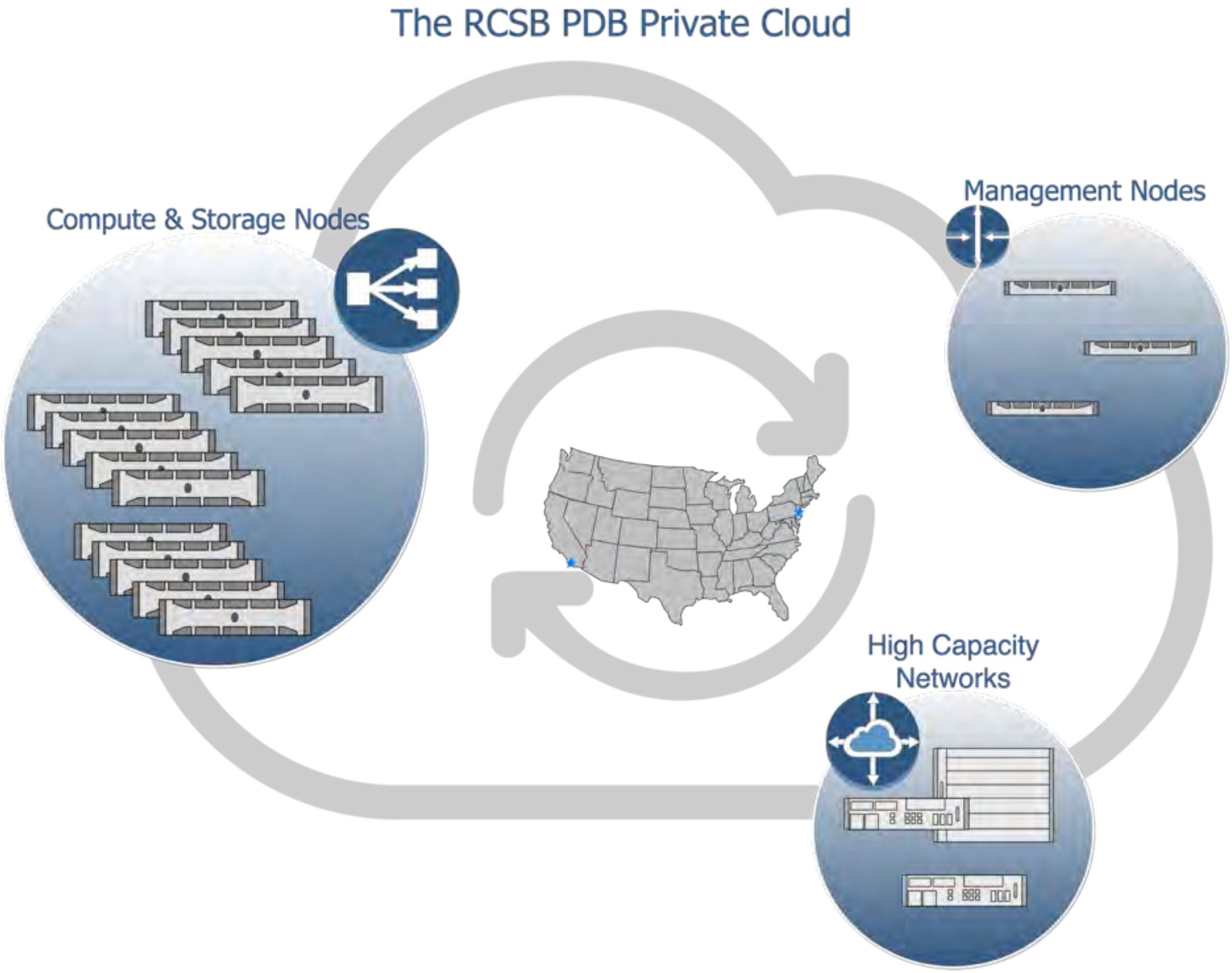
3J3Q

Atomic-level structure of the entire HIV-1 capsid

Note: Use your mouse to drag, rotate, and zoom in and out of the structure. Click to identify atoms and bonds.

NGL is a WebGL based 3D viewer powered by MMTF.

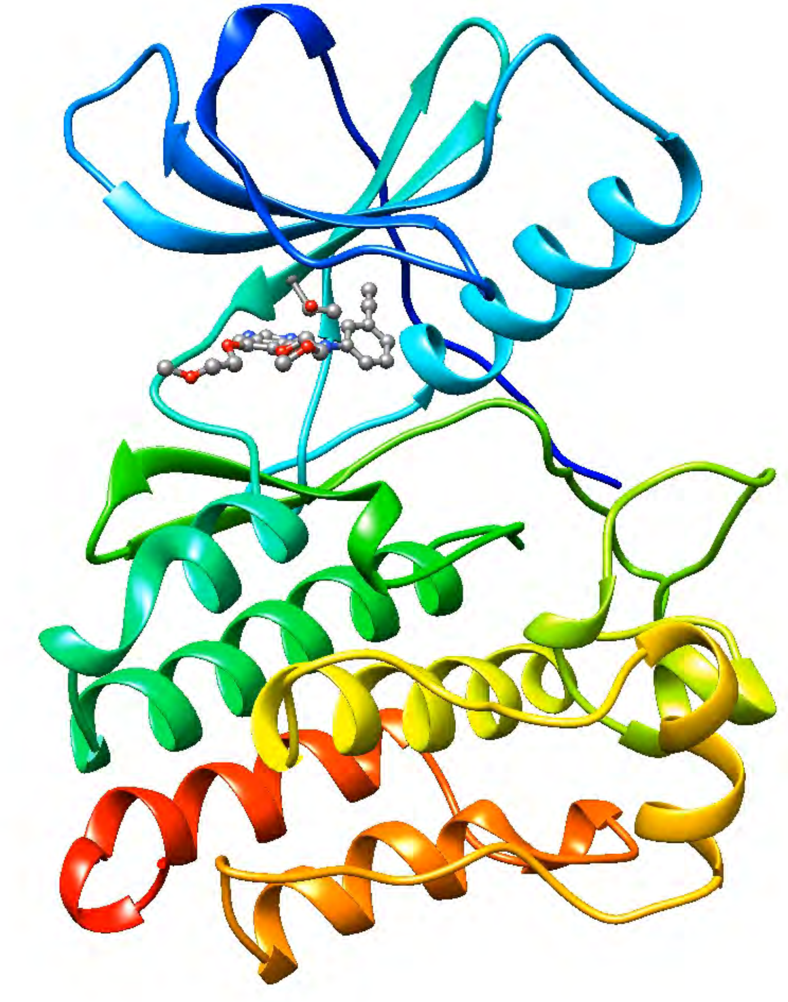
FTP Data Distribution/Access



RCSB PDB distributes data using a Private Cloud→Future Public Cloud

Data Out

- Optimization of rcsb.org website continues
- Integration of genome sequence →protein sequence →3D structure
- Visualization of pathways, ligands, very large structures, *etc*.

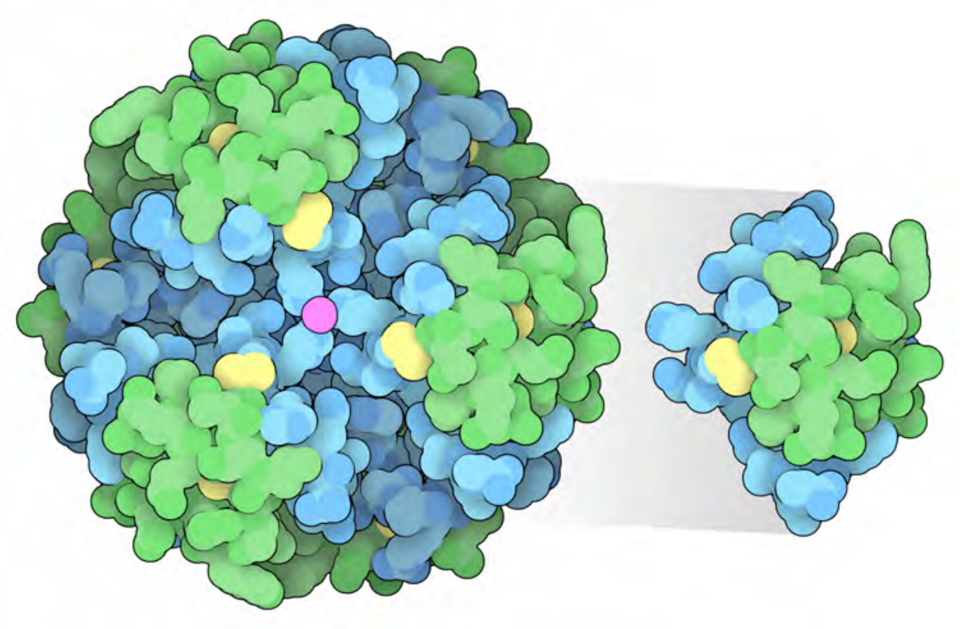


Erlotinib targeting EGFR for Lung Cancer (PDB 4HJO)

Education

- Developed modular curriculum on diabetes for use in high schools

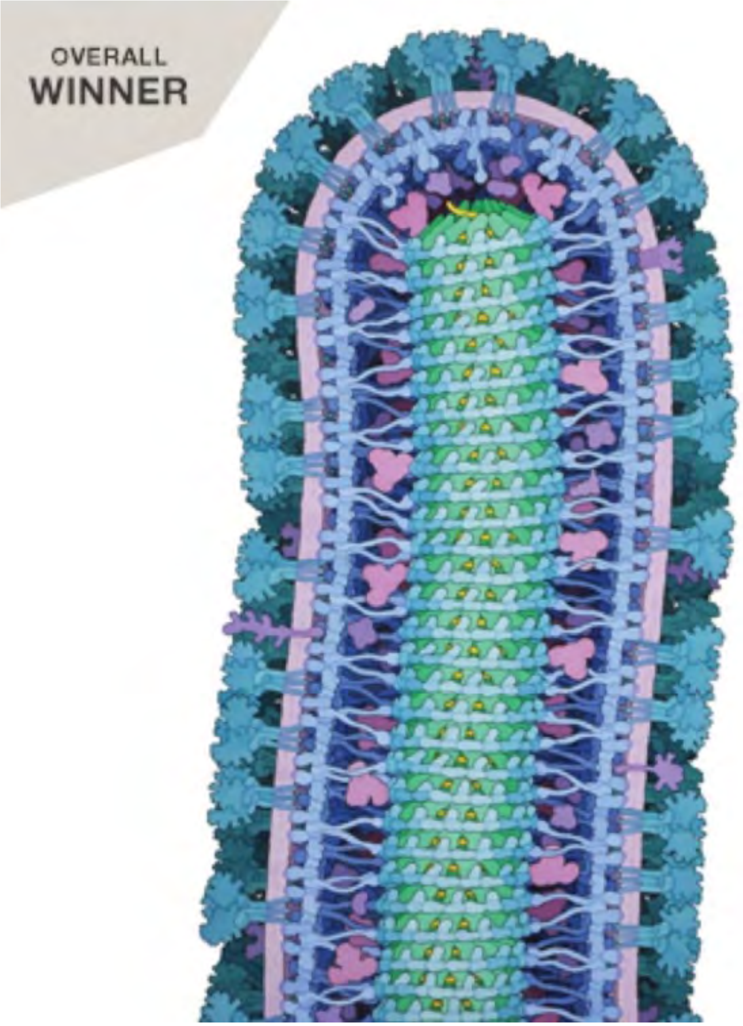
- Deployed diabetes materials (Molecule of the Month articles, poster) *via* PDB-101 website



Insulin Hexamer/Monomer from Molecule of the Month on *Designer Insulins* (February 2016)

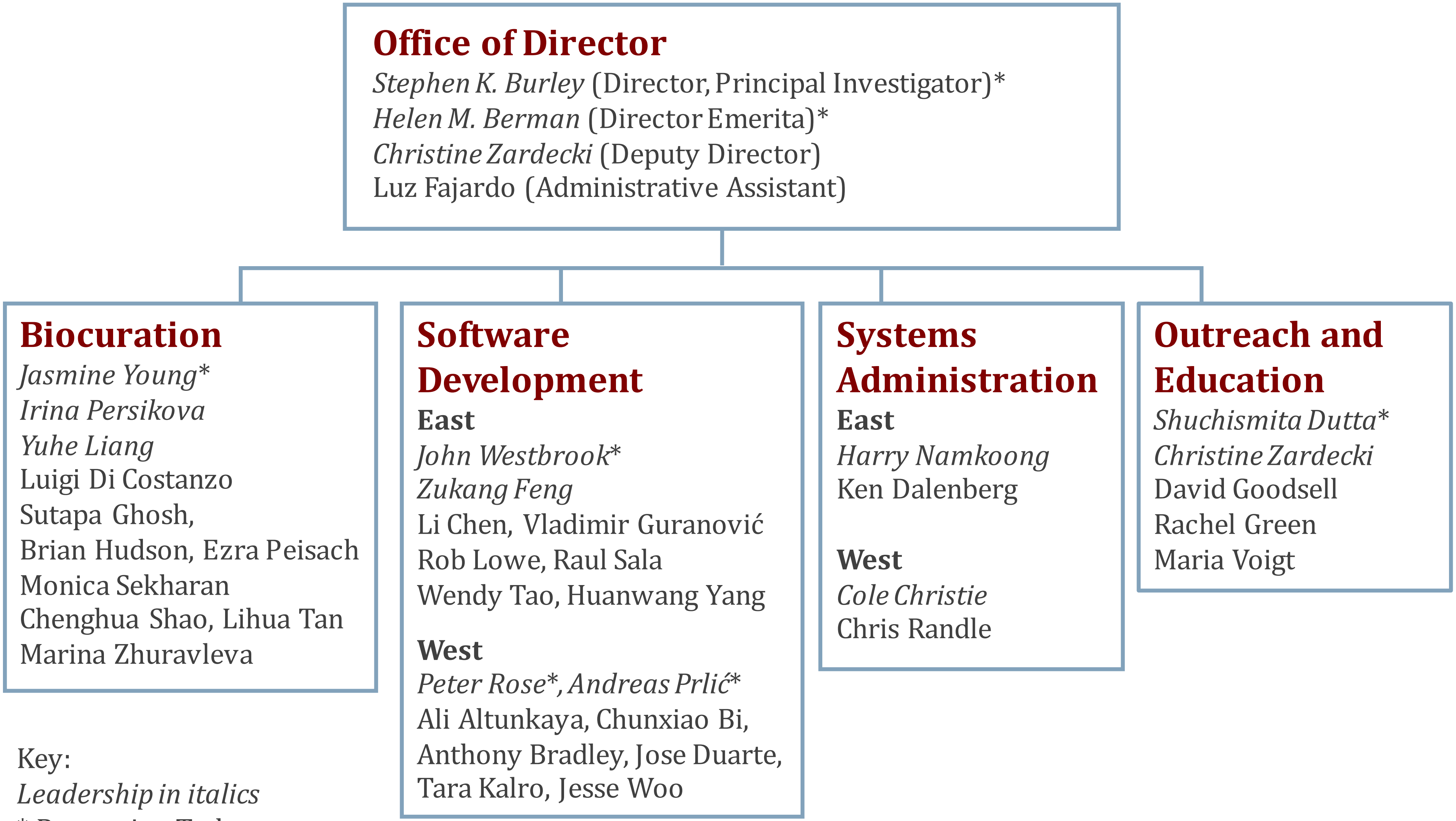
Global Outreach and Engagement

- Wellcome Image Award successes
- Deployment of Zika virus outreach materials with Purdue structure release
- Publication of Ligand Validation Workshop whitepaper in *Structure*



Ebola Virus painted for
Molecule of the Month by David Goodsell

RCSB PDB Organization



Office of Director

Stephen K. Burley (Director, Principal Investigator)*
Helen M. Berman (Director Emerita)*
Christine Zardecki (Deputy Director)
Luz Fajardo (Administrative Assistant)

Biocuration

*Jasmine Young**
Irina Persikova
Yuhe Liang
Luigi Di Costanzo
Sutapa Ghosh,
Brian Hudson, Ezra Peisach
Monica Sekharan
Chenghua Shao, Lihua Tan
Marina Zhuravleva

Software Development

East
*John Westbrook**
Zukang Feng
Li Chen, Vladimir Guranović
Rob Lowe, Raul Sala
Wendy Tao, Huanwang Yang

West
Peter Rose, Andreas Prlić**
Ali Altunkaya, Chunxiao Bi,
Anthony Bradley, Jose Duarte,
Tara Kalro, Jesse Woo

Systems Administration

East
Harry Namkoong
Ken Dalenberg

West
Cole Christie
Chris Randle

Outreach and Education

*Shuchismita Dutta**
Christine Zardecki
David Goodsell
Rachel Green
Maria Voigt

Key:
Leadership in italics
* Presenting Today

RCSB PDB Staff Outreach Commitment

NJ Science Olympiad

ABRCMS 2016

Undergraduate Research, Summer 2016, Rutgers

San Diego Science and Engineering Festival

High School and Undergraduate Research, Summer 2016, UCSD

Supporting Diversity

- Longtime commitment to supporting a diverse, inclusive, and family-friendly workplace

- Mentorship of students under-represented in sciences
  - External *via* Rutgers Office of Diversity and Inclusion Summer Program (RiSE, 3 students in 2016)
  - Rutgers Undergraduates (4 students in 2016)
  - UCSD Outreach Programs

- Joint participation with the Rutgers Center for Graduate Recruitment, Retention, and Diversity at ABRCMS and SACNAS 2016 national meetings

RCSB PDB Advisory Committee

- Provides independent advice to RCSB PDB Director and staff
  - Operates under formal Terms of Reference
  - Triennial rotation schedule (renewable)
  - Cynthia Wolberger agreed to chair through 2019
- Comments, advises, or makes recommendations for action on topical issues as they arise over the course of the time between meetings, and on any standing agenda items
  - Deposition Policies and Annotation Practices
  - Data Distribution, Query Policies, and Practices
  - Education and Outreach

Agenda

Overview	Stephen K. Burley
Data In *OneDep, Data Standards, Infrastructure, Plans Forward*	Jasmine Young and John Westbrook
Data Out *Access and Exploration*	Peter Rose and Andreas Prlić
Outreach	Helen M. Berman

Lunch

Education	Shuchismita Dutta
Funding and Sustainability **Response to 2015 Report**	Stephen K. Burley

Matters Arising & General Discussion

Data In: OneDep, Data Standards, Infrastructure, Plans Forward

Jasmine Young, Ph.D.

John Westbrook, Ph.D.

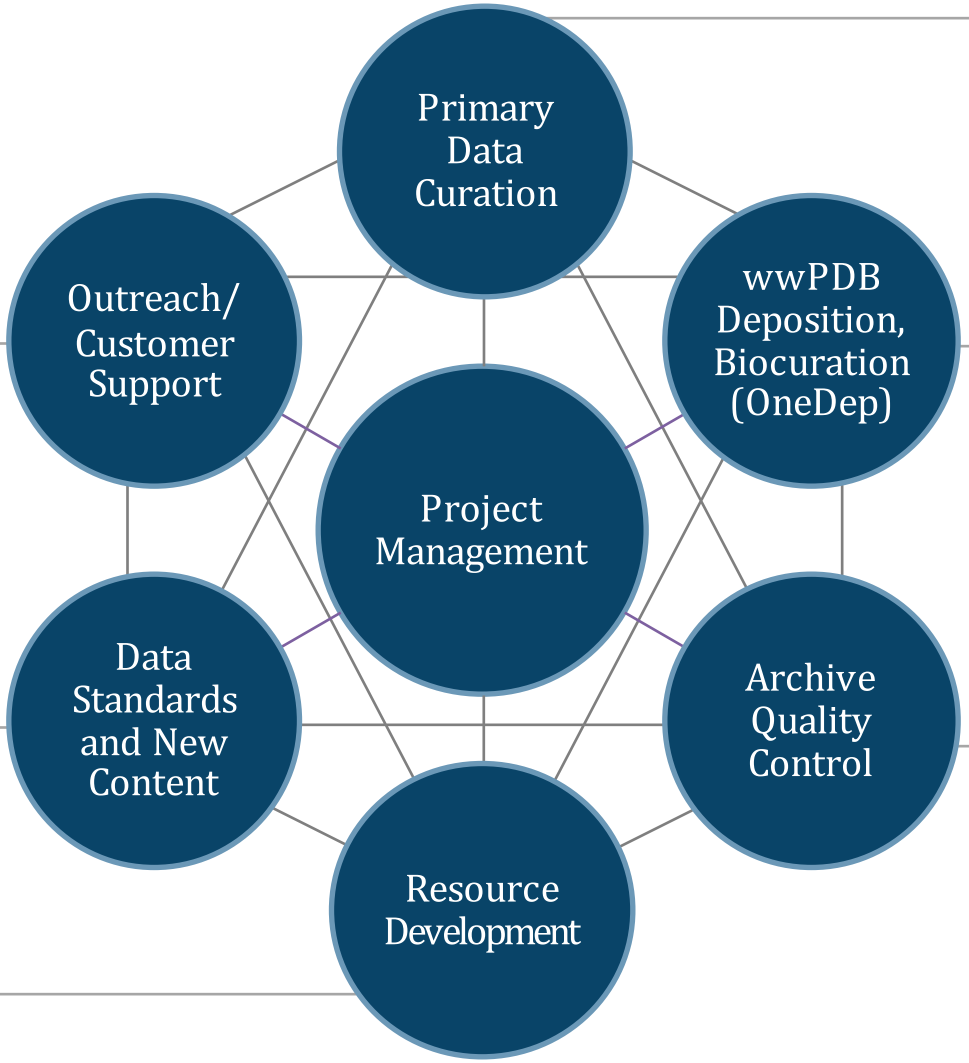
Outline

- Team and responsibilities
- Data life cycle
- wwPDB international partnership
- Importance of biocuration
- Engaging scientific communities
- Infrastructure
- Plans forward

Biocurators and Data In Developers

- 11 scientists,
  3 scientist programmers,
  4 software developers
- 14 Ph.D., 2 M.S., 2 B.S.
- 8 countries,
  3 continents
- Combined length of service 169 years
- Median length of service 9.4 years

Data In Responsibilities



Primary Data Curation

wwPDB Deposition, Biocuration (OneDep)

Archive Quality Control

Resource Development

Data Standards and New Content

Outreach/ Customer Support

Project Management

RCSB PDB curates >5,000 entries/year

Develop and maintain tools for Deposition, Biocuration, and Validation

Assure quality of weekly data release

Make larger archive-wide improvements (“remediation”)

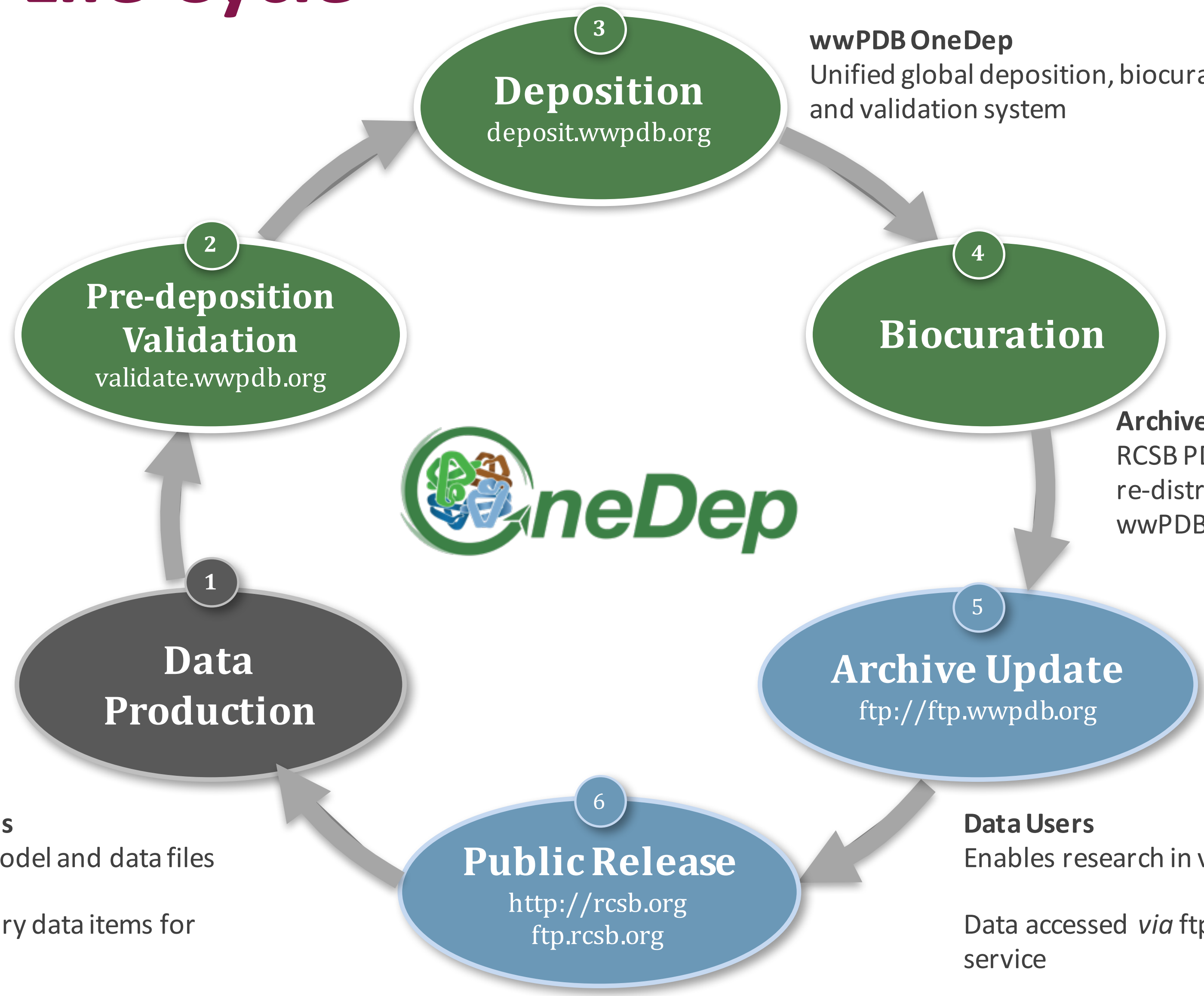
Resource integration across Data In pipeline

Create, maintain and refine data dictionaries

Communicate with Depositors and diverse User Communities

Funding for Core Activities: NSF [DBI-1338415], NIH, DOE

Data Life Cycle



3 Deposition
deposit.wwpdb.org

4 Biocuration

5 Archive Update
ftp://ftp.wwpdb.org

6 Public Release
http://rcsb.org
ftp.rcsb.org

1 Data Production

2 Pre-deposition Validation
validate.wwpdb.org

OneDep

wwPDB OneDep
Unified global deposition, biocuration, and validation system

Archive Keeper
RCSB PDB packages and re-distributes data to wwPDB partners

Structure Biologists
Generate atomic model and data files

Assemble mandatory data items for deposition

Data Users
Enables research in various fields

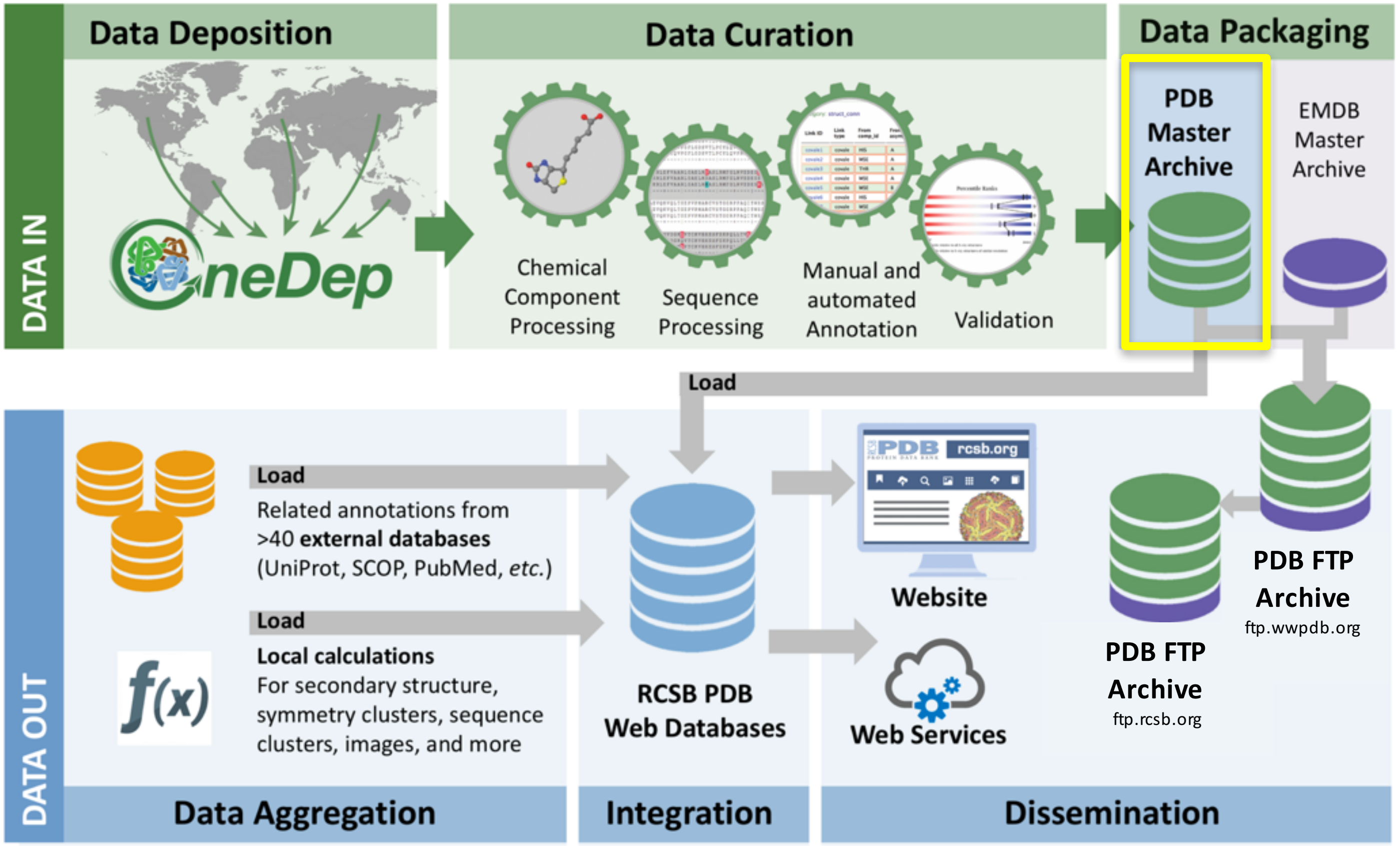
Data accessed *via* ftp download or web service

wwPDB International Collaboration (Data In)

- Developing unified global tools for deposition, validation, and biocuration
- Defining data standards and content: PDBx/mmCIF Dictionary
- Ensuring data uniformity in the PDB archive ("Remediation")
- Maintaining a single global archive

RCSB PDB is the Archive Keeper

- RCSB PDB is the Archive Keeper for world distribution of PDB data and leads the wwPDB collaboration in developing tools, setting data standards, and performing data remediation

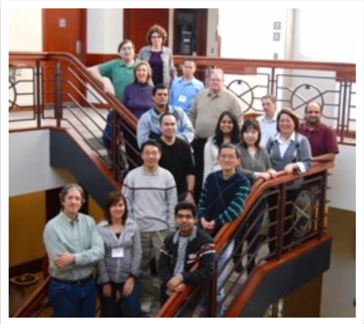


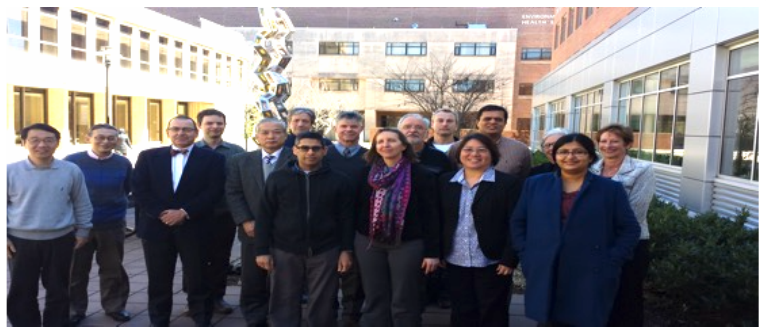
Unified Global OneDep Tool

- More complete data capture
- File format standardization
- Improved efficiency and consistency
  - Enables workload balancing
  - More automation
  - File replacement pre-submission
- Validation for all methods
  - Standalone Validation Server
  - Web Service API
- Support for larger and more complex structures

2007 Initial Discussions

2010 OneDep Team Meeting



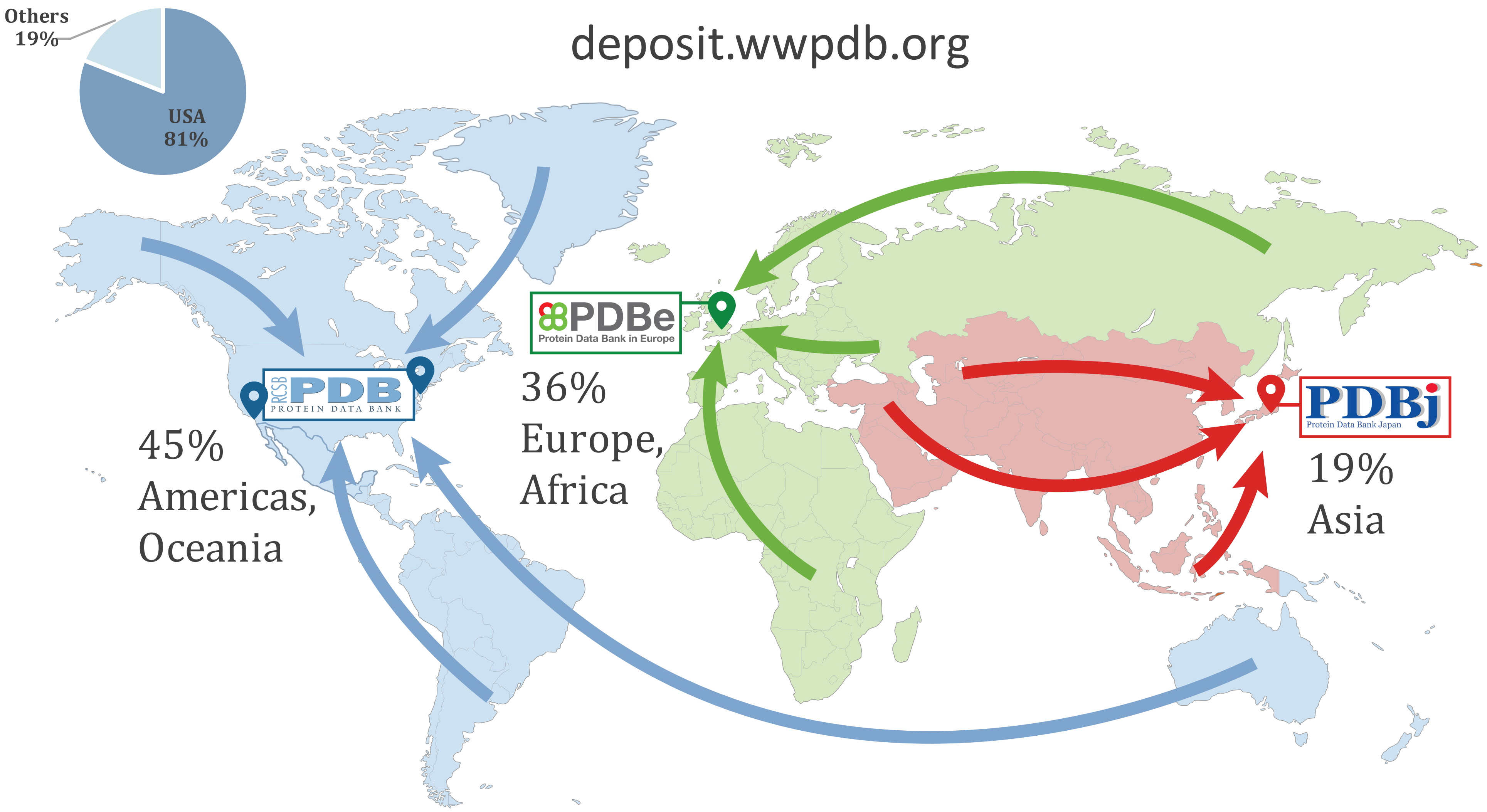


2016 OneDep Summit Meeting

RCSB PDB Effort on OneDep Project

- Provided technical and managerial leadership
- Responsibilities
  - Backend infrastructure and technology
  - Biocuration pipeline
  - Hosting wwPDB development servers
  - Technical support for partner server installation
- >50% of wwPDB-committed FTEs from RCSB PDB
  - Jasmine Young, Global Project Lead

Workload Balancing/Depositor Support



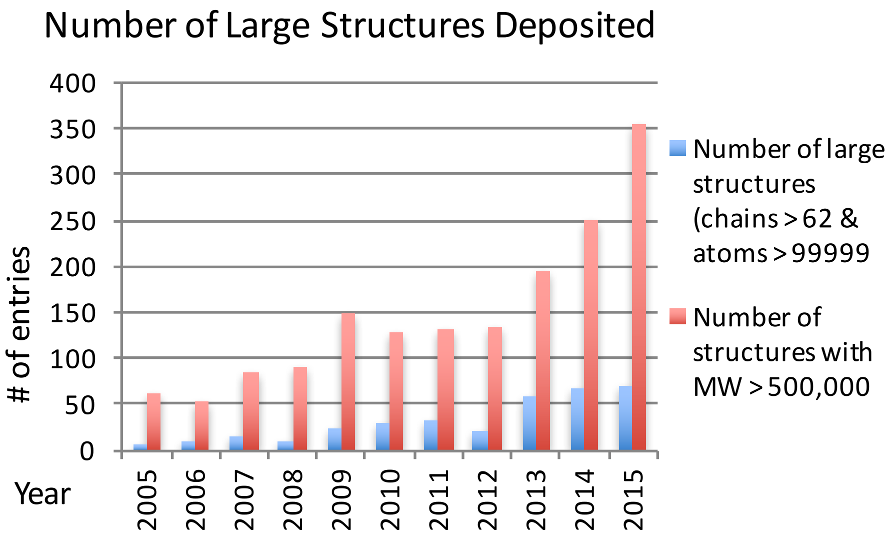
deposit.wwpdb.org

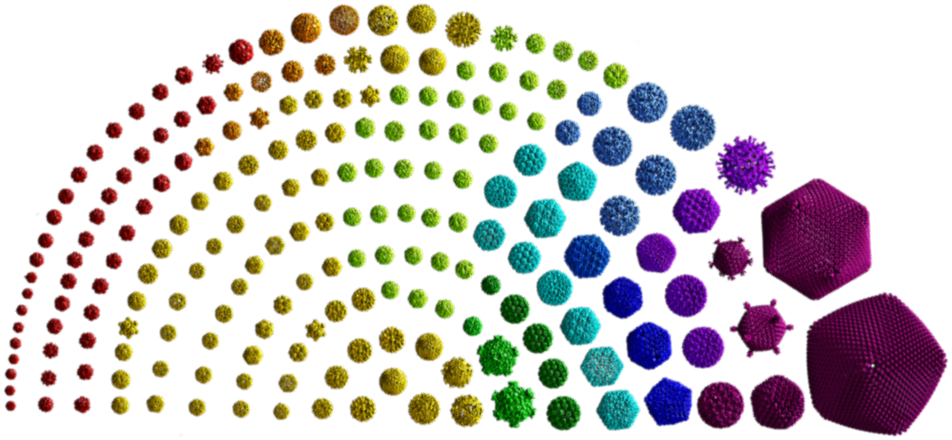
45%
Americas,
Oceania

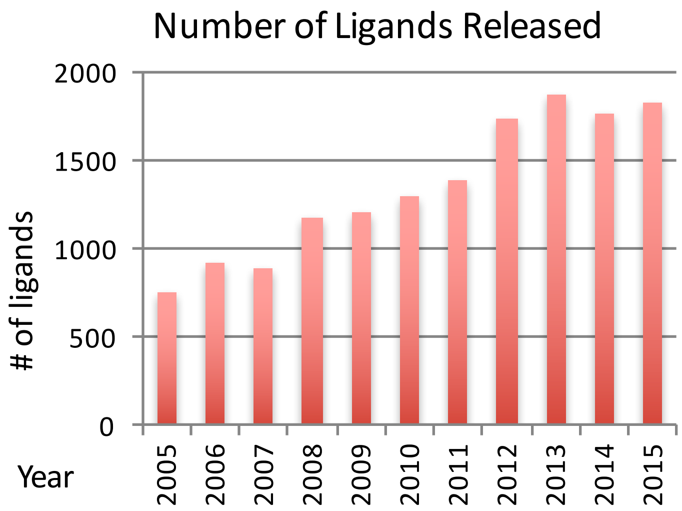
36%
Europe,
Africa

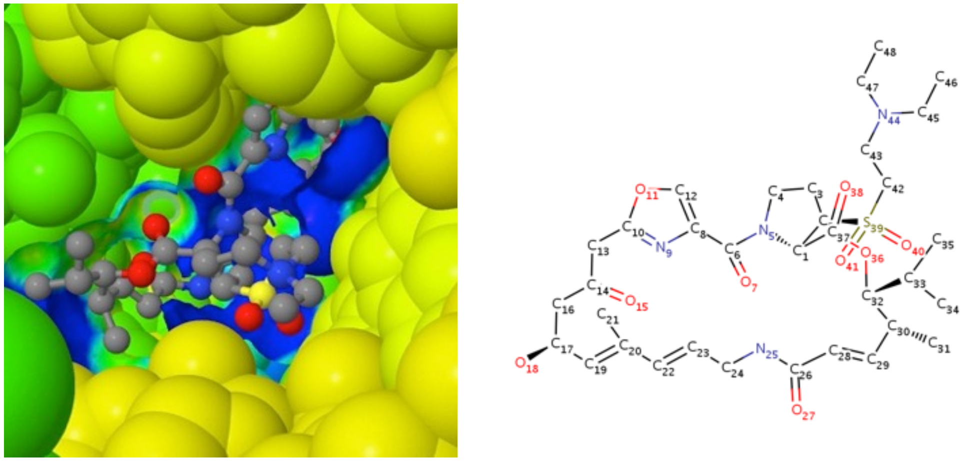
19%
Asia

Increasing Size and Complexity


Number of Large Structures Deposited

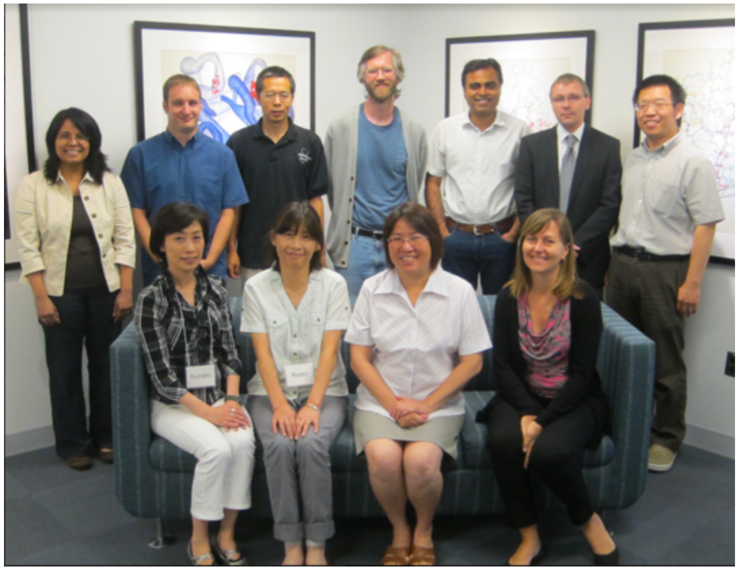

Faustovirus (PDB 5j7v, Klose et al., 2016) is the largest PDB structure


Number of Ligands Released


Antibiotic quinupristin/Dalfopristin bound to ribosome (PDB 4u26 Noeske et al., 2014)

Importance of Biocuration

- Enforces data standardization through policies and common biocuration practices
- Ensures data quality and provides value-added annotation
- Communicates possible errors to Depositors (wwPDB Validation Report)
- Maintains data uniformity and compliance in the PDB archive to enable data search and exploration
- Requires domain expertise
  - Cannot be replaced by purely computational means

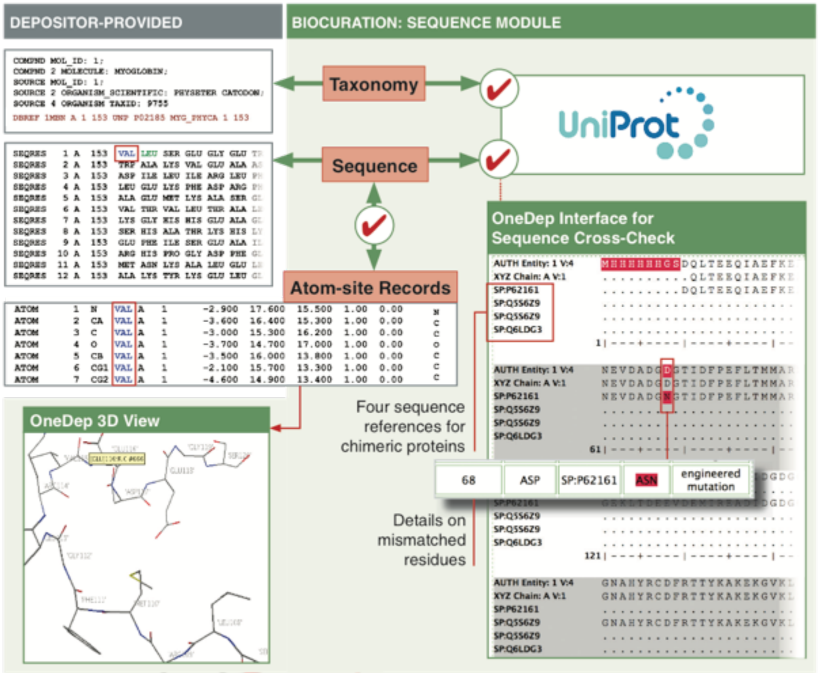


2014 wwPDB Biocurator Summit

2015 PDBj Biocuration Training

Data Quality and Value-Added Annotation

- Consistency checking
  - Polymer sequence and taxonomy
  - Ligand stereochemistry
  - Ligand density fit
- Integration with external data resources
- Overall quantitative and qualitative review of deposited data



DEPOSITOR-PROVIDED

```
COMPND MOL_ID: 1;
COMPND 2 MOLECULE: MYOGLOBIN;
SOURCE MOL_ID: 1;
SOURCE 2 ORGANISM_SCIENTIFIC: PHYSETER CATODON;
SOURCE 4 ORGANISM_TAXID: 9755
DBREF 1MBN A 1 153 UNP P02185 MYG_PHYCA 1 153
```

```
SEQRES   1 A  153  VAL LEU SER GLU GLY GLU
SEQRES   2 A  153  TRP ALA LYS VAL GLU ALA
SEQRES   3 A  153  ASP ILE LEU ILE ARG LEU
SEQRES   4 A  153  LEU GLU LYS PHE ASP ARG
SEQRES   5 A  153  ALA GLU MET LYS ALA SER
SEQRES   6 A  153  VAL THR VAL LEU THR ALA
SEQRES   7 A  153  LYS GLY HIS HIS GLU ALA
SEQRES   8 A  153  SER HIS ALA THR LYS HIS
SEQRES   9 A  153  GLU PHE ILE SER GLU ALA
SEQRES  10 A  153  ARG HIS PRO GLY ASP PHE
SEQRES  11 A  153  MET ASN LYS ALA LEU GLU
SEQRES  12 A  153  ALA LYS TYR LYS GLU LEU
```

```
ATOM      1  N   VAL A   1      -2.900  17.600  15.500  1.00  0.00     N
ATOM      2  CA  VAL A   1      -3.600  16.400  15.300  1.00  0.00     C
ATOM      3  C   VAL A   1      -3.000  15.300  16.200  1.00  0.00     C
ATOM      4  O   VAL A   1      -3.700  14.700  17.000  1.00  0.00     O
ATOM      5  CB  VAL A   1      -3.500  16.000  13.800  1.00  0.00     C
ATOM      6  CG1 VAL A   1      -2.100  15.700  13.300  1.00  0.00     C
ATOM      7  CG2 VAL A   1      -4.600  14.900  13.400  1.00  0.00     C
```

BIOCURATION: SEQUENCE MODULE

Taxonomy

Sequence

Atom-site Records

OneDep 3D View

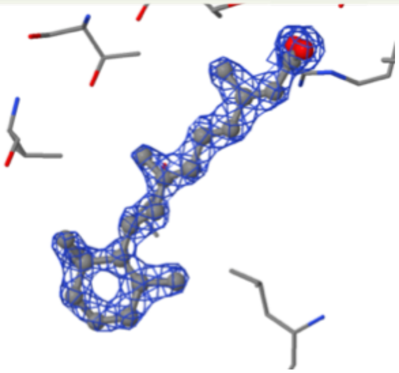
OneDep Interface for Sequence Cross-Check

```
AUTH Entity: 1 V:4   MHHHHHHGSDQLTEEQIAEFKE
XYZ Chain: A V:1     ...........QLTEEQIAEFKE
SP:P62161            ..........DQLTEEQIAEFKE
SP:Q5S6Z9            ......................
SP:Q5S6Z9            ......................
SP:Q6LDG3            ......................
                   1 |----+-----|----+-----|--

AUTH Entity: 1 V:4   NEVDADGDGTIDFPEFLTMMAR
XYZ Chain: A V:1     NEVDADGDGTIDFPEFLTMMAR
SP:P62161            NEVDADGNGTIDFPEFLTMMAR
SP:Q5S6Z9            ......................
SP:Q5S6Z9            ......................
SP:Q6LDG3            ......................
                  61 |----+-----|----+-----|--

SP:P62161            GEKLTDEEVDEMIREADIDGDG
SP:Q5S6Z9            ......................
SP:Q5S6Z9            ......................
SP:Q6LDG3            ......................
                 121 |----+-----|----+-----|--

AUTH Entity: 1 V:4   GNAHYRCDFRTTYKAKEKGVKL
XYZ Chain: A V:1     GNAHYRCDFRTTYKAKEKGVKL
SP:P62161            ......................
SP:Q5S6Z9            GNAHYRCDFRTTYKAKEKGVKL
SP:Q5S6Z9            ......................
SP:Q6LDG3            ......................
```

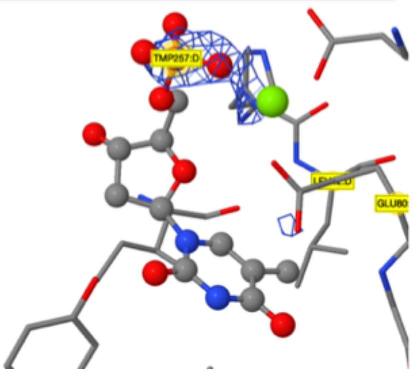
68	ASP	SP:P62161	ASN	engineered mutation

Four sequence references for chimeric proteins

Details on mismatched residues

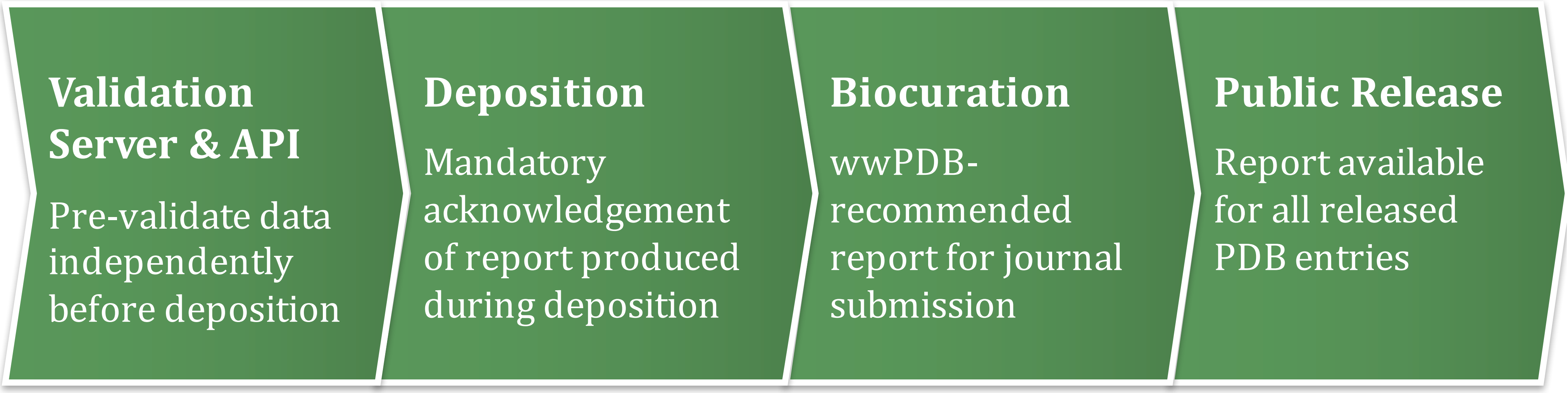


REA in PDB 1CBS
(Kleywegt et al., 1994)
RSR=0.10, RSCC=0.96



TMP in PDB 3HW4
(Kaushik et al., 2013)
RSR=0.41, RSCC=0.57

Improving Data Quality



Validation Server & API

Pre-validate data independently before deposition

Deposition

Mandatory acknowledgement of report produced during deposition

Biocuration

wwPDB-recommended report for journal submission

Public Release

Report available for all released PDB entries

Coordinates and data frequently replaced during Deposition and Biocuration

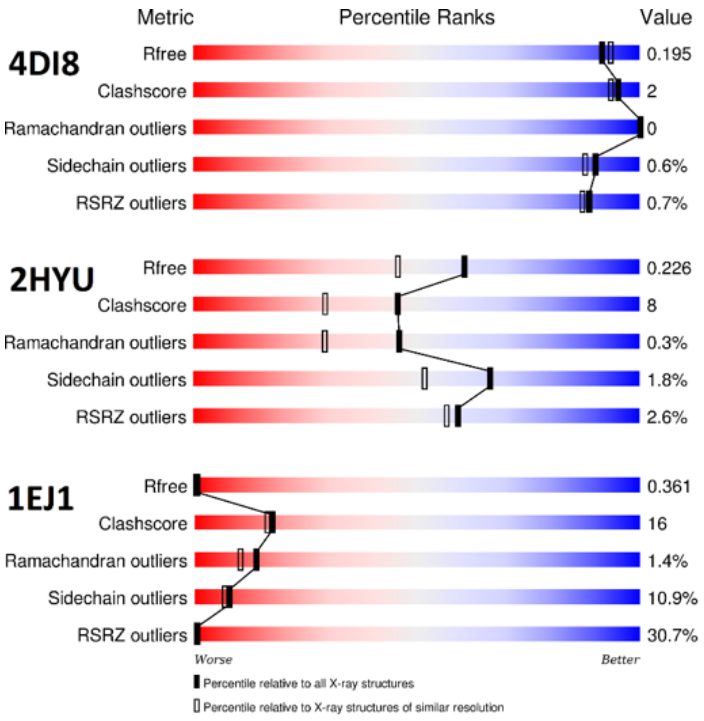
Validation Report submission during manuscript review process

- Mandatory: *Nature* journals, *Acta D & F, FEBS, JBC, J Immunology, eLIFE,* and *Angew Chem Int Ed Engl*
- Recommended: *Cell, Molecular Cell, Structure*

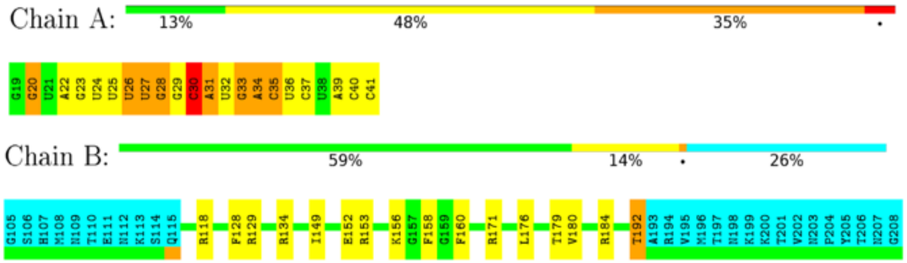
Key Features of wwPDB Validation Reports

- Graphical overview of data quality
- Residue plots
- Atomic model quality
- Experimental data quality

Overall Quality



Residue Plots



🟩 OK 🟨 1 issue 🟧 2 issues 🟥 3+ issues

🟦 ill-defined by coordinates

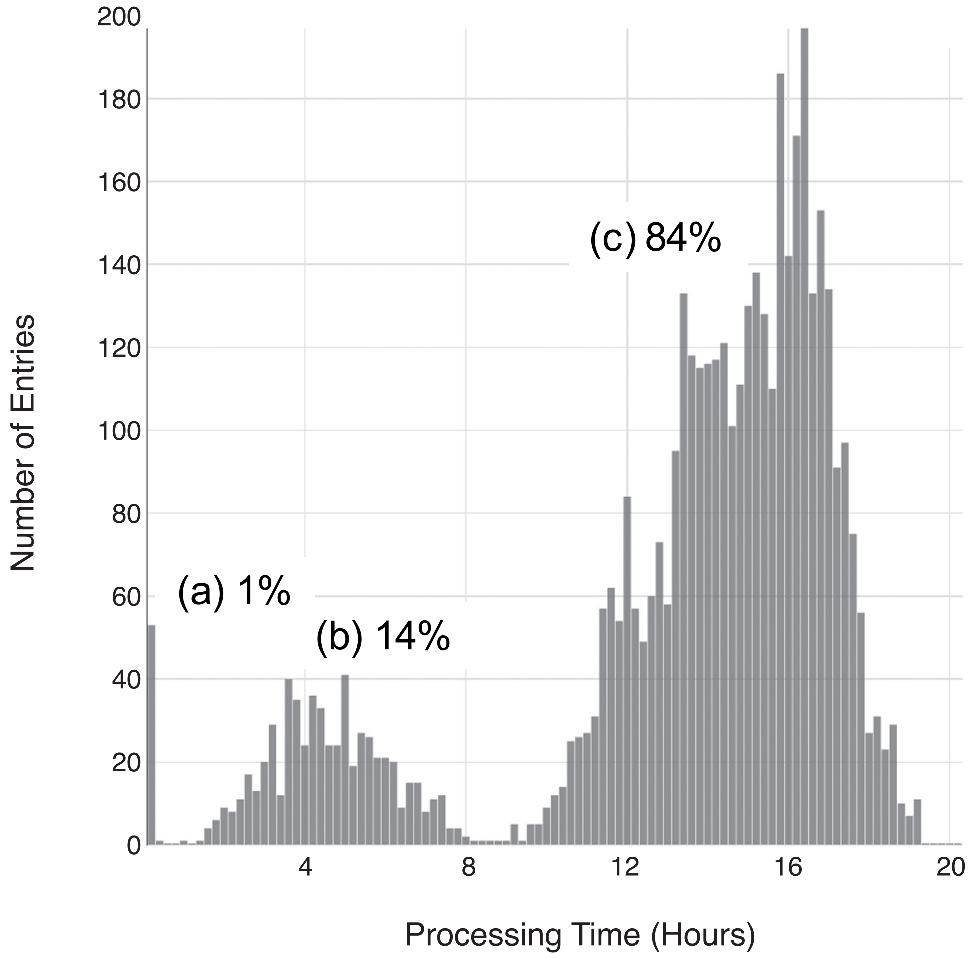
OneDep Biocuration Processing Time

(a) ~1hr: Simple structures without issues

(b) ~4 hrs: More complex structures without issues

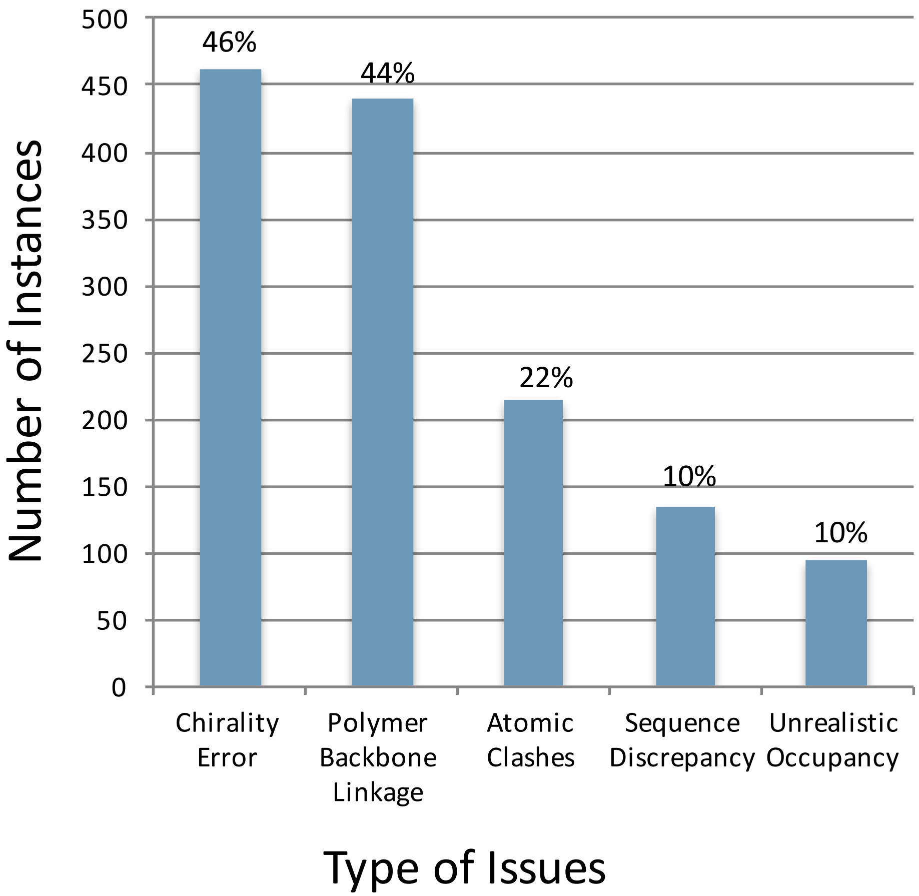
(c) ~15 hrs: Structures with issues, including Depositor response time

Legacy ADIT system:
4-5 days



OneDep Biocuration Impact

Top issues frequently raised during Biocuration



New data sets received in response to issues raised during Biocuration

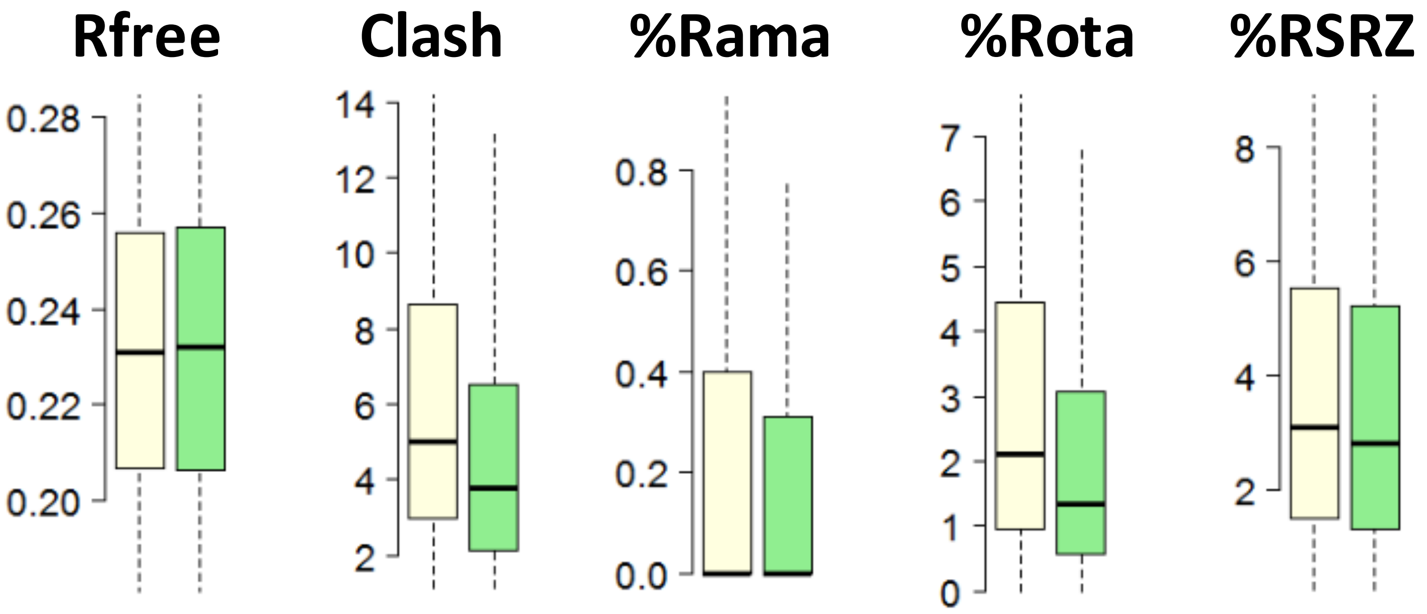
- ~29% of all 2015 entries
- ~25% of all 2016 entries

New tools to promote pre-deposition validation

- Standalone Validation Server (now supports NMR, 3DEM)
- Web Service API

Improving X-ray Structure Quality

Structure quality improvement since the advent of the wwPDB Validation Report



Yellow: Legacy System 2012-2013
Green: OneDep 2014-2015

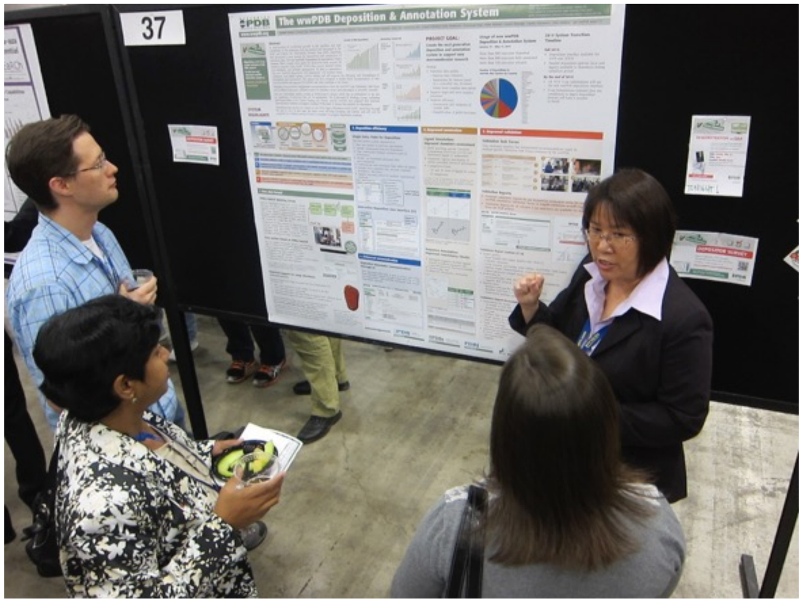
Enabling Data Search and Exploration

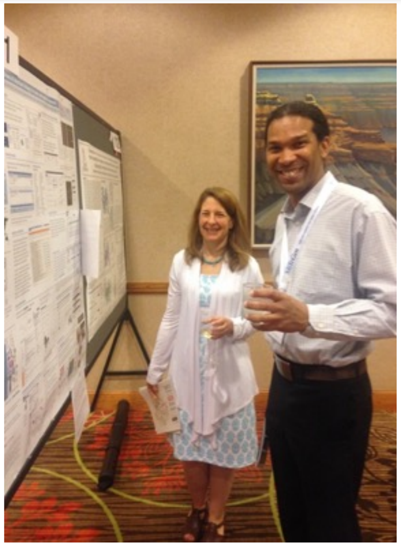
- Data standardization significantly impacts data query
- RCSB PDB has taken leadership
  - Developing and maintaining data standards
  - Data remediation
- 5 rounds of data remediation for entire PDB archive from 2007 to 2014
- OneDep system also supports data remediation
  - 3DEM (in collaboration with EMDataBank)
  - Carbohydrates

Henrick, *et al.*, 2008, *NAR*; Lawson, *et al.*, 2008, *Acta Cryst. D*; Dutta, *et al.*, 2014, *Biopolymers*

Depositor/User Feedback

- Daily communication between Biocurators and External Users (Depositors)
- ACA, IUCr meetings: demonstrations, posters and exhibit booth
- Internal Users (Biocurators)
- Continuous testing and improvement
- Weekly cross-site reviews of issues



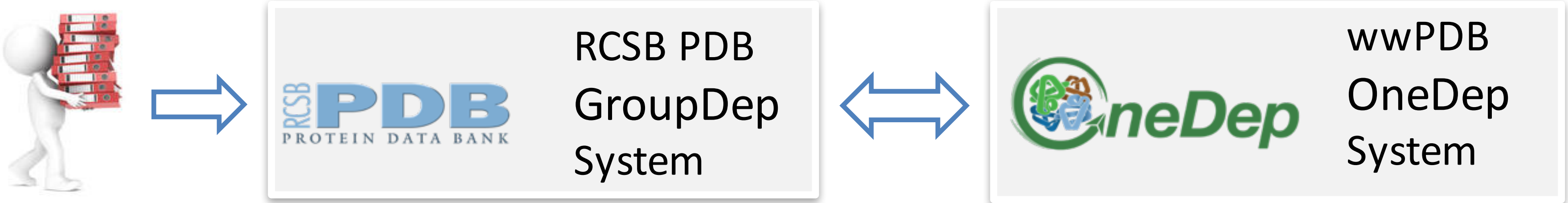


Enabling Bulk Depositions from Industry

"Group" Deposition developed to meet community need

- Requirements set by wwPDB OneDep Team
- Support for D3R Blind Challenges
- Depositors: Roche, EMD Serono, University of Marburg, University of Essex
- 364 depositions in single group processed in 5 days

deposit-group.rcsb.rutgers.edu/groupdeposit/



RCSB PDB GroupDep System

wwPDB OneDep System

External funding: NIH grant 1U01GM111528

Engaging Scientific Communities

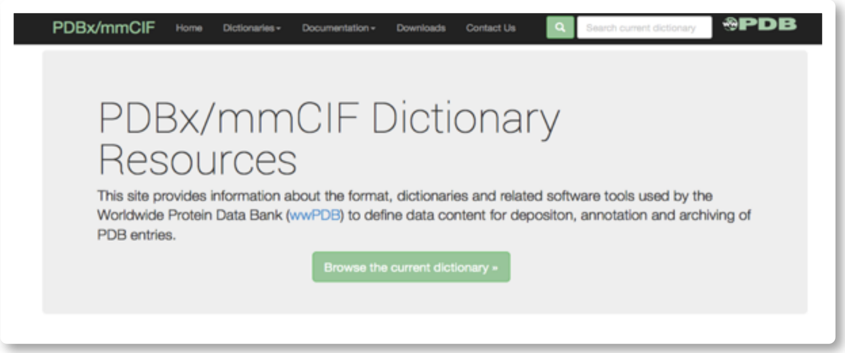
- Defining data content and quality standards
- Task Forces and Working Groups
  - Validation Task Forces (VTFs)
    - X-ray, NMR, 3DEM
  - Small Angle Scattering
  - Integrative/Hybrid Methods Task Force
  - PDBx/mmCIF Working Groups
  - NEF Working Group
- Ligand Validation Workshop

Defining Data Content & Quality Standards

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

Community Data Standards

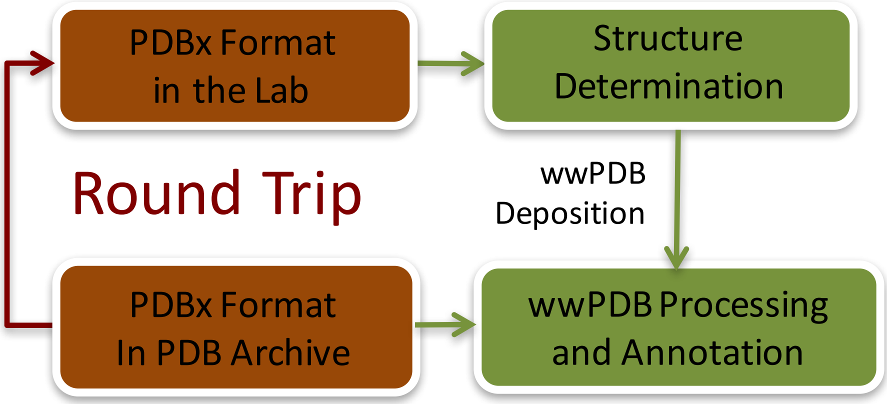
- Data managed using PDBx/mmCIF
  - Extends earlier IUCr data standard
  - PDBx/mmCIF dictionary has >4400 data terms
- Extensions now coordinated with wwPDB PDBx/mmCIF Working Group
  - Supports broader needs of both contributors and users of the archive
- Host community workshops
- mmCIF.wwpdb.org provides data dictionaries, schema, software tools



PDBx/mmCIF Dictionary Resources

This site provides information about the format, dictionaries and related software tools used by the Worldwide Protein Data Bank (wwPDB) to define data content for depositon, annotation and archiving of PDB entries.

mmcif.wwpdb.org

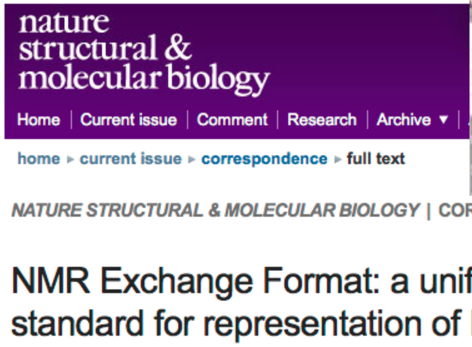


Data Standards Working Groups

- PDBx/mmCIF Working Group
  - Experts and methods developers
  - Ensures good support in key community software tools
- NMR Task Force Working Group
  - NMR Exchange Format
- SASCIF
  - PDBx-compatible extension dictionary supporting data exchange with SASBDB

PDBx Workshop, October 2014

NMR Workshop, August 2016



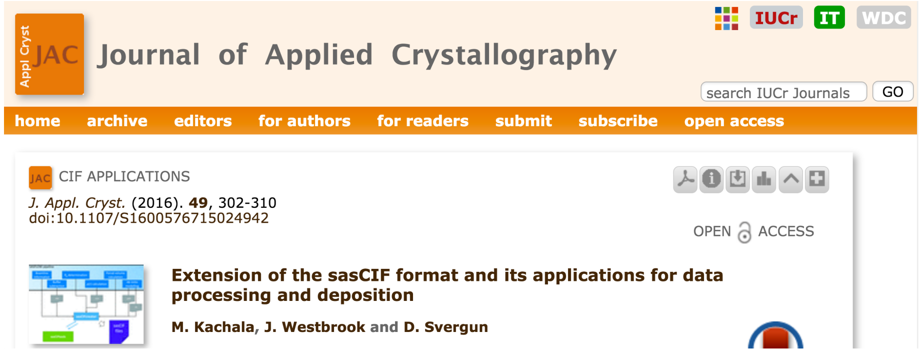
nature structural & molecular biology

Home | Current issue | Comment | Research | Archive ▼ |

home ▸ current issue ▸ correspondence ▸ full text

NATURE STRUCTURAL & MOLECULAR BIOLOGY | CORRESPONDENCE OPEN

NMR Exchange Format: a unified and open standard for representation of NMR restraint data



Journal of Applied Crystallography

search IUCr Journals GO

home archive editors for authors for readers submit subscribe open access

CIF APPLICATIONS

J. Appl. Cryst. (2016). **49**, 302-310
doi:10.1107/S1600576715024942

OPEN ACCESS

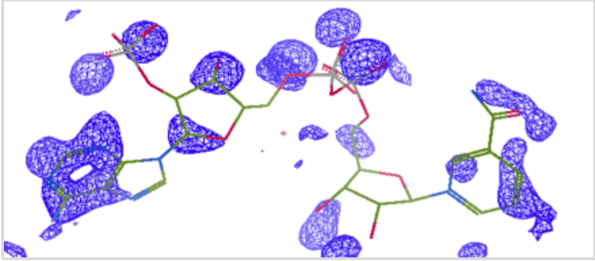
Extension of the sasCIF format and its applications for data processing and deposition

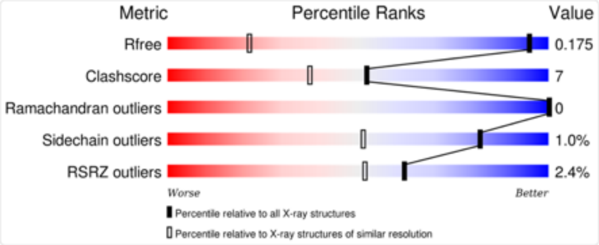
M. Kachala, J. Westbrook and D. Svergun

Focus on Ligand Quality

Current summary validation statistics may not identify poor electron density fit

NADP in PDB 1ZK4:
2|Fo|-|Fc| map at 1σ
Schlieben *et al.*, 2005

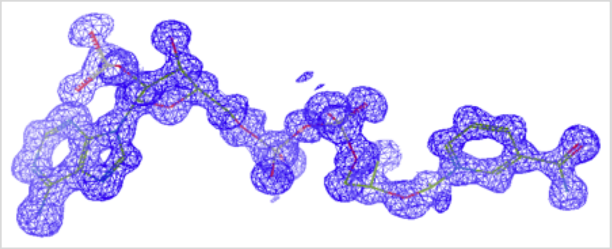


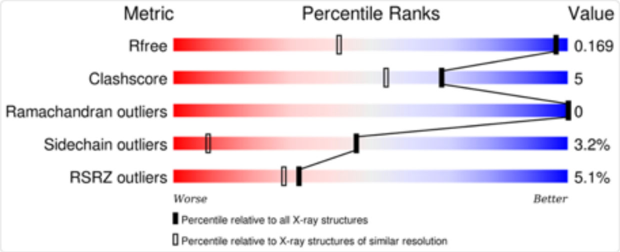


Metric	Percentile Ranks	Value
Rfree		0.175
Clashscore		7
Ramachandran outliers		0
Sidechain outliers		1.0%
RSRZ outliers		2.4%

▮ Percentile relative to all X-ray structures
▯ Percentile relative to X-ray structures of similar resolution

NADP in PDB 2FZD:
2|Fo|-|Fc| map at 1σ
Steuber *et al.*, 2006

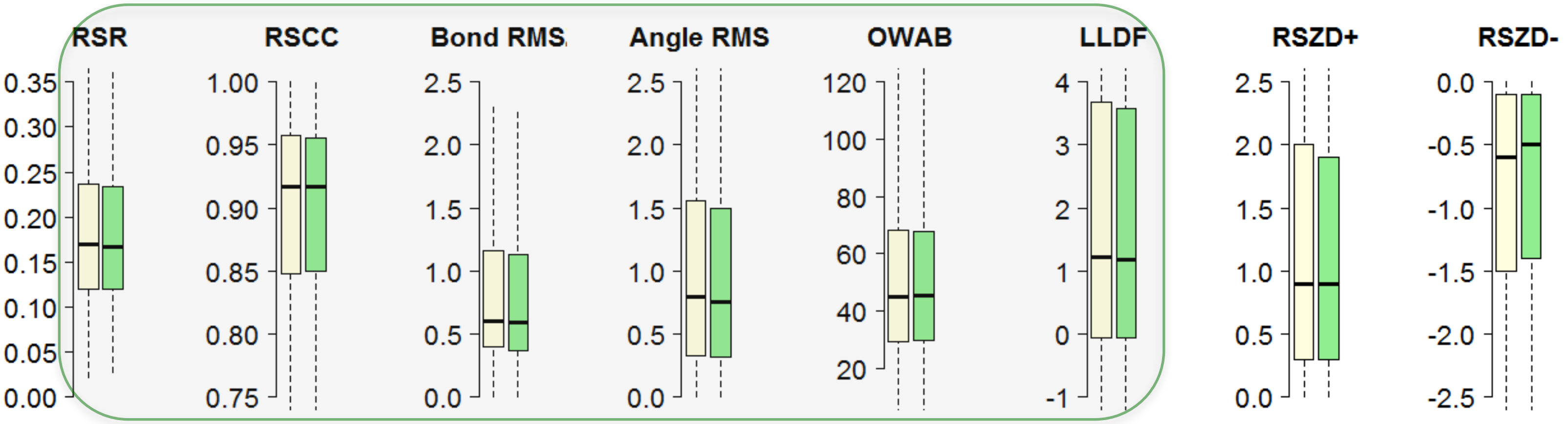




Metric	Percentile Ranks	Value
Rfree		0.169
Clashscore		5
Ramachandran outliers		0
Sidechain outliers		3.2%
RSRZ outliers		5.1%

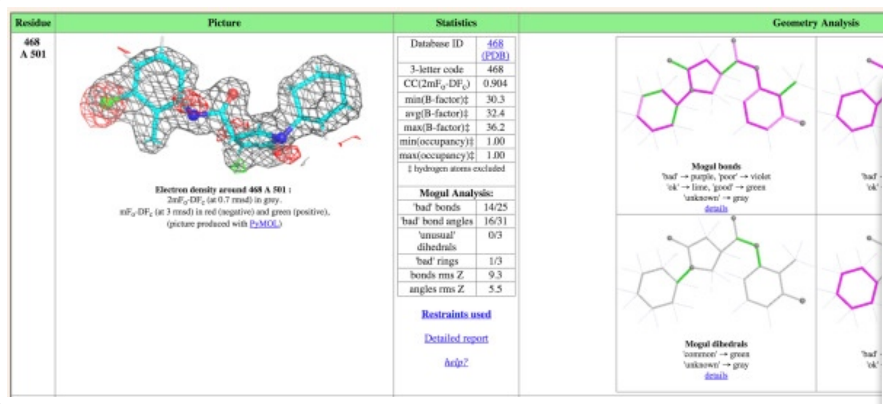
▮ Percentile relative to all X-ray structures
▯ Percentile relative to X-ray structures of similar resolution

Overall Ligand Model Quality: Legacy 2012-2013 (yellow) *vs*. OneDep 2014-2015 (green)



Ligand Validation Workshop

- Co-crystal structure determination experts (Academe and Industry) and Software Developers (X-ray Crystallography and Computational Chemistry) discussed, developed, and recommended:
  - Best practices for PDB archive deposition/validation of co-crystal structures
  - Editorial/Refereeing/Publication standards for co-crystal structures



Residue | Picture | Statistics | Geometry Analysis

468
A 501

Electron density around 468 A 501 :
$2mF_o$-DF_c (at 0.7 rmsd) in grey.
mF_o-DF_c (at 3 rmsd) in red (negative) and green (positive).
(picture produced with PyMOL)

Database ID	468 (PDB)
3-letter code	468
CC($2mF_o$-DF_c)	0.904
min(B-factor)‡	30.3
avg(B-factor)‡	32.4
max(B-factor)‡	36.2
min(occupancy)‡	1.00
max(occupancy)‡	1.00

‡ hydrogen atoms excluded

Mogul Analysis:	
'bad' bonds	14/25
'bad' bond angles	16/31
'unusual' dihedrals	0/3
'bad' rings	1/3
bonds rms Z	9.3
angles rms Z	5.5

Restraints used

Detailed report

help?

Mogul bonds
'bad' → purple, 'poor' → violet
'ok' → lime, 'good' → green
'unknown' → gray
details

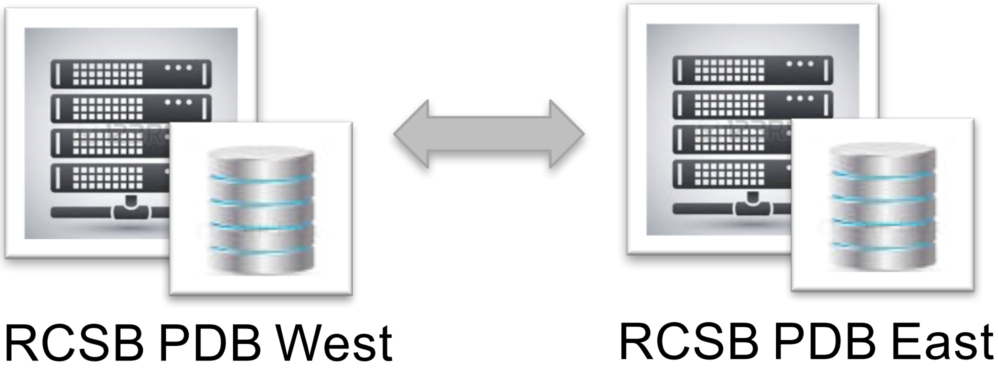
Mogul dihedrals
'common' → green
'unknown' → gray
details

Adams *et al.* (2016) *Structure* 24: 502-508

Rutgers July 30-31, 2015

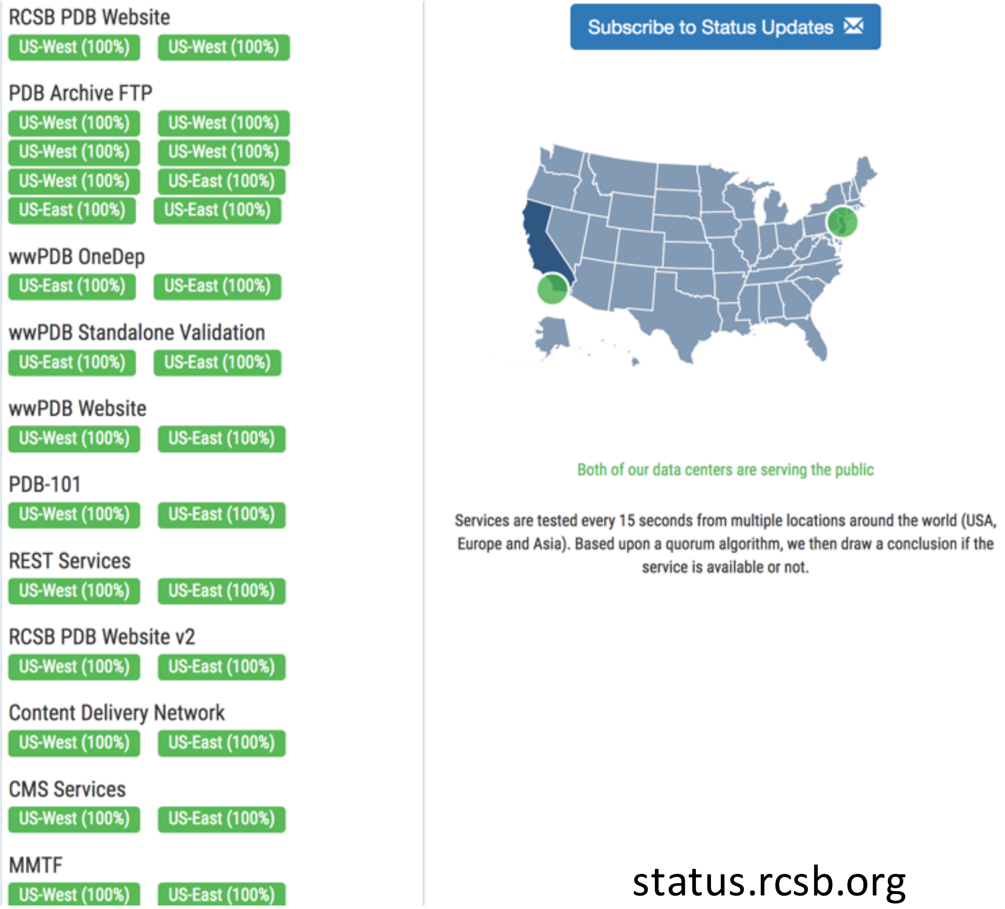
High Availability OneDep

- Redirection between global partners sites in the of event loss-of-service
- RCSB PDB hosts bi-coastal OneDep services
  - Warm failover: Independent East and West Coast OneDep installations
  - Active failover: In-progress East Coast deposition sessions mirrored to West Coast



Continuous Monitoring

- PDB archive FTP
- wwPDB OneDep systems
- wwPDB validation servers
- wwPDB website
- RCSB PDB website and related services



RCSB PDB Website
US-West (100%) US-West (100%)

PDB Archive FTP
US-West (100%) US-West (100%)
US-West (100%) US-West (100%)
US-West (100%) US-East (100%)
US-East (100%) US-East (100%)

wwPDB OneDep
US-East (100%) US-East (100%)

wwPDB Standalone Validation
US-East (100%) US-East (100%)

wwPDB Website
US-West (100%) US-East (100%)

PDB-101
US-West (100%) US-East (100%)

REST Services
US-West (100%) US-East (100%)

RCSB PDB Website v2
US-West (100%) US-East (100%)

Content Delivery Network
US-West (100%) US-East (100%)

CMS Services
US-West (100%) US-East (100%)

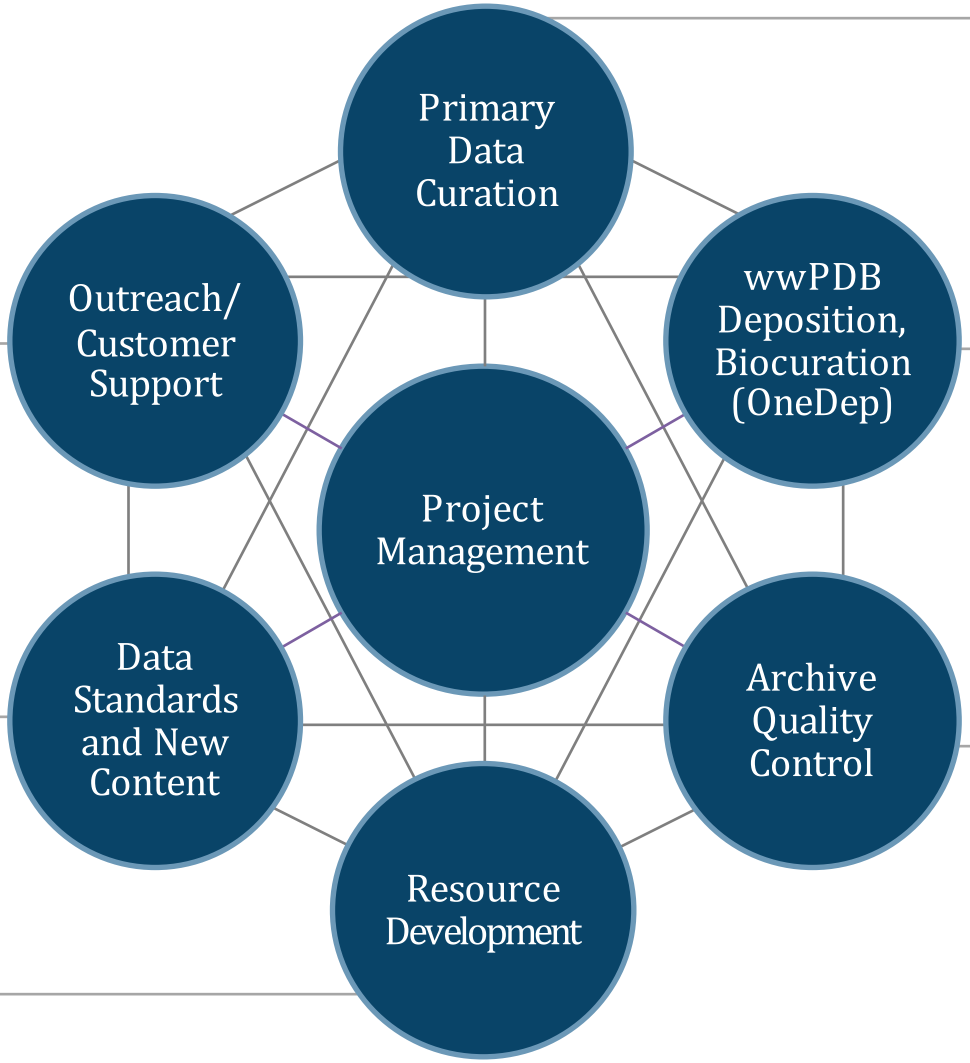
MMTF
US-West (100%) US-East (100%)

Both of our data centers are serving the public

Services are tested every 15 seconds from multiple locations around the world (USA, Europe and Asia). Based upon a quorum algorithm, we then draw a conclusion if the service is available or not.

status.rcsb.org

Data In Plans Forward



Primary Data Curation

Outreach/ Customer Support

wwPDB Deposition, Biocuration (OneDep)

Project Management

Data Standards and New Content

Archive Quality Control

Resource Development

Improve automation in Biocuration and data remediation

Continue to engage Depositors and diverse User communities

Harden OneDep infrastructure to support new wwPDB members (*e.g.*, China and India)

Continue data standardization efforts to support data remediation

Streamline weekly OneDep update operations

File versioning to improve archive management

Resource integration across Data In pipeline

Funding for Core Activities: NSF [DBI-1338415], NIH, DOE
External Funding for I/HM: NSF EAGER (DBI-1519158)

RCSB PDB
PROTEIN DATA BANK

Data Out: Access, Exploration, and Metrics

Peter Rose, Ph.D.

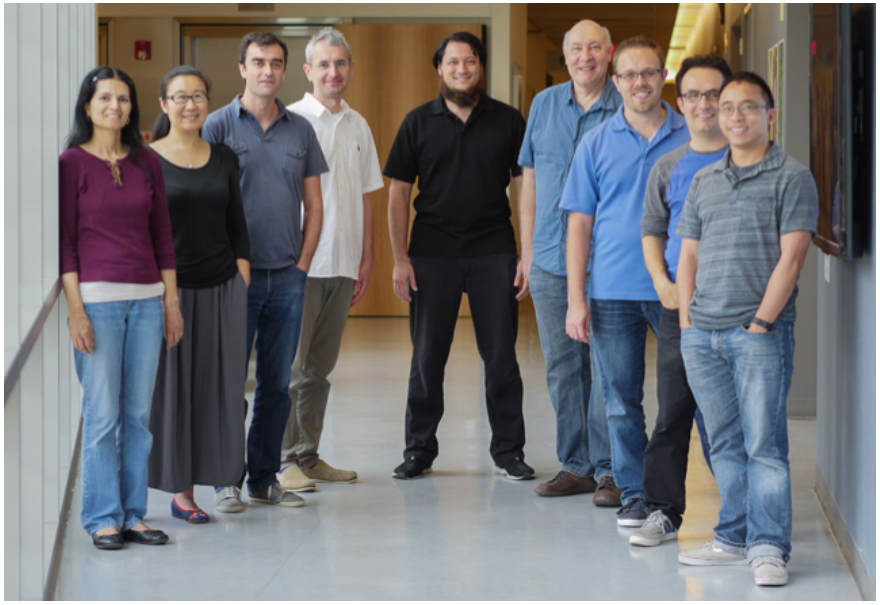
Andreas Prlić, Ph.D.

Outline

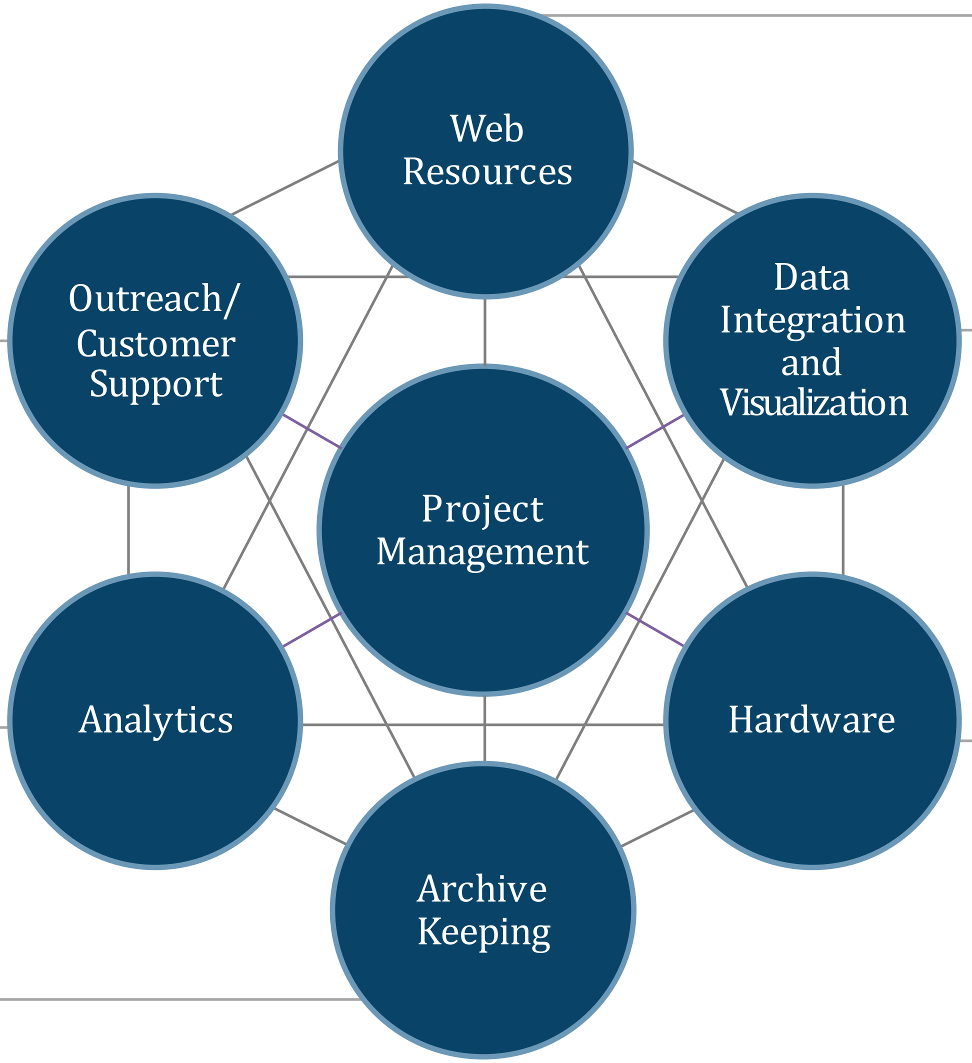
- Team and responsibilities
- Who uses RCSB PDB?
- What type of research do we enable?
- How broad is our impact?
- Plans forward

Bi-coastal Developer Team

- 3 scientists-software dev.,
  6 software developers,
  2 systems & infrastructure
- 4 Ph.D., 5 M.S., 2 B.S.
- 9 countries/3 continents
- 48 combined years of service
- 3 median years of service



Responsibilities



Web Resources

Outreach/ Customer Support

Data Integration and Visualization

Project Management

Analytics

Hardware

Archive Keeping

Website and resources utilized by >350,000 unique visitors/month
>1 million unique visitors/year

Communicate with diverse User communities

>40 External Resources

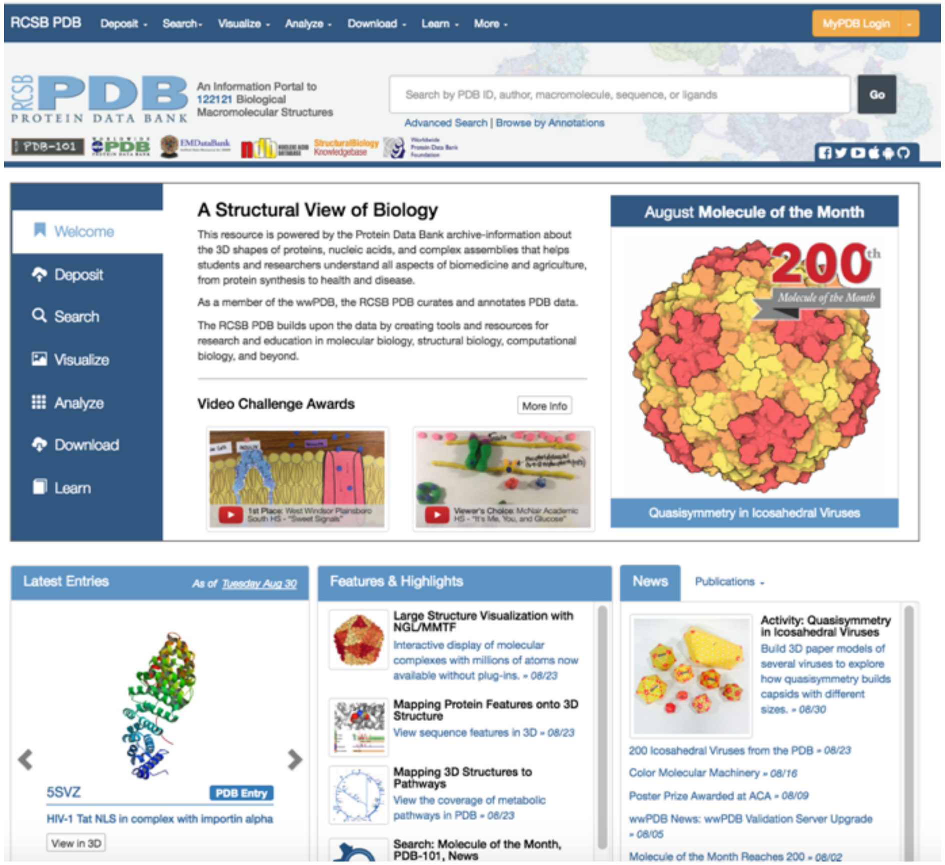
Ongoing usage assessment to inform development

System Design, Procurement, Maintenance

24/7/366 Support, Weekly Updates, Annual Snapshots

Funding for Core Activities: NSF [DBI-1338415], NIH, DOE

Who Are Our Users And How Do We Know?


RCSB PDB
Deposit
Search
Visualize
Analyze
Download
Learn
More
MyPDB Login
RCSB PDB PROTEIN DATA BANK
Search by PDB ID, author, macromolecule, sequence, or ligands
Go
Advanced Search | Browse by Annotations
PDB-101
EMDataBank
StructuralBiology Knowledgebase
Worldwide Protein Data Bank Foundation
Welcome
Deposit
Search
Visualize
Analyze
Download
Learn
200th
Molecule of the Month
More Info
View in 3D
PDB Entry

An Information Portal to 122121 Biological Macromolecular Structures

A Structural View of Biology

This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

As a member of the wwPDB, the RCSB PDB curates and annotates PDB data.

The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.

Video Challenge Awards

1st Place: West Windsor Plainsboro South HS - "Sweet Signals"

Viewer's Choice: McNair Academic HS - "It's Me, You, and Glucose"

August Molecule of the Month

Quasisymmetry in Icosahedral Viruses

Latest Entries As of Tuesday Aug 30

5SVZ

HIV-1 Tat NLS in complex with importin alpha

Features & Highlights

Large Structure Visualization with NGL/MMTF
Interactive display of molecular complexes with millions of atoms now available without plug-ins. » 08/23

Mapping Protein Features onto 3D Structure
View sequence features in 3D » 08/23

Mapping 3D Structures to Pathways
View the coverage of metabolic pathways in PDB » 08/23

Search: Molecule of the Month, PDB-101, News

News | Publications

Activity: Quasisymmetry in Icosahedral Viruses
Build 3D paper models of several viruses to explore how quasisymmetry builds capsids with different sizes. » 08/30

200 Icosahedral Viruses from the PDB » 08/23

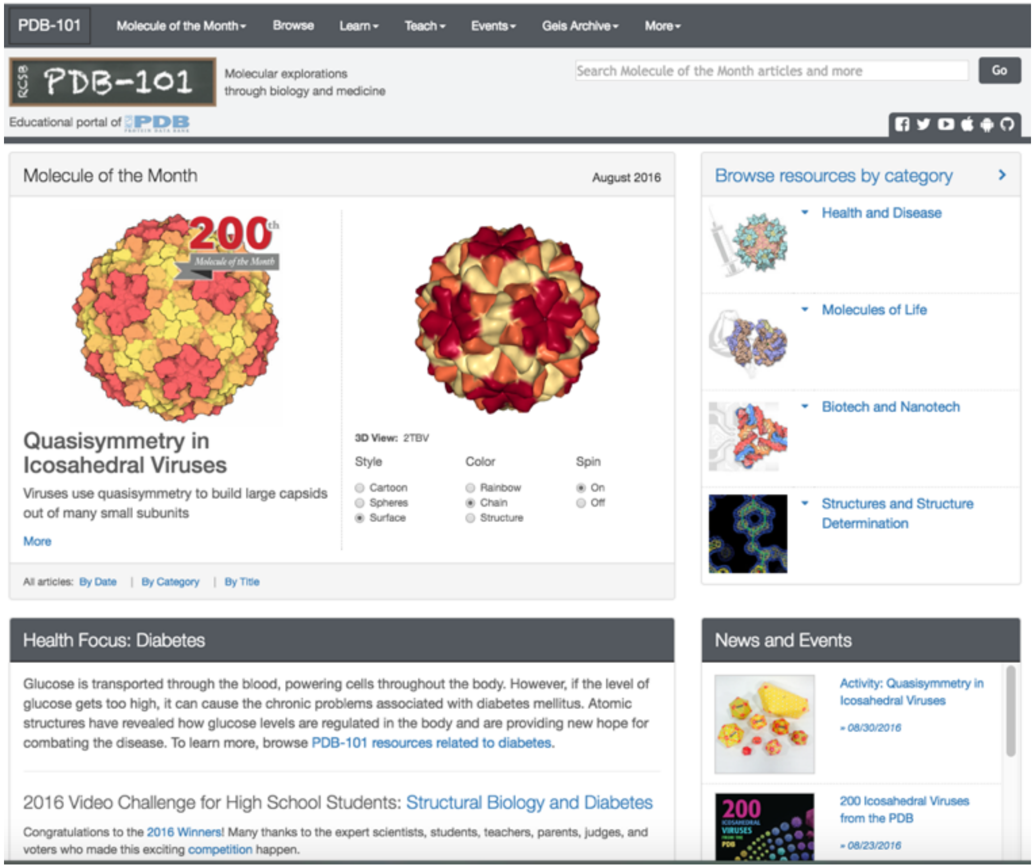
Color Molecular Machinery » 08/16

Poster Prize Awarded at ACA » 08/09

wwPDB News: wwPDB Validation Server Upgrade » 08/05

Molecule of the Month Reaches 200 » 08/02

rcsb.org



Molecular explorations through biology and medicine

Educational portal of PDB

Molecule of the Month August 2016

Quasisymmetry in Icosahedral Viruses

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

More

3D View: 2TBV

Style	Color	Spin
○ Cartoon	○ Rainbow	● On
○ Spheres	● Chain	○ Off
● Surface	○ Structure	

All articles: By Date | By Category | By Title

Browse resources by category

- Health and Disease
- Molecules of Life
- Biotech and Nanotech
- Structures and Structure Determination

Health Focus: Diabetes

Glucose is transported through the blood, powering cells throughout the body. However, if the level of glucose gets too high, it can cause the chronic problems associated with diabetes mellitus. Atomic structures have revealed how glucose levels are regulated in the body and are providing new hope for combating the disease. To learn more, browse PDB-101 resources related to diabetes.

2016 Video Challenge for High School Students: Structural Biology and Diabetes

Congratulations to the 2016 Winners! Many thanks to the expert scientists, students, teachers, parents, judges, and voters who made this exciting competition happen.

News and Events

Activity: Quasisymmetry in Icosahedral Viruses » 08/30/2016

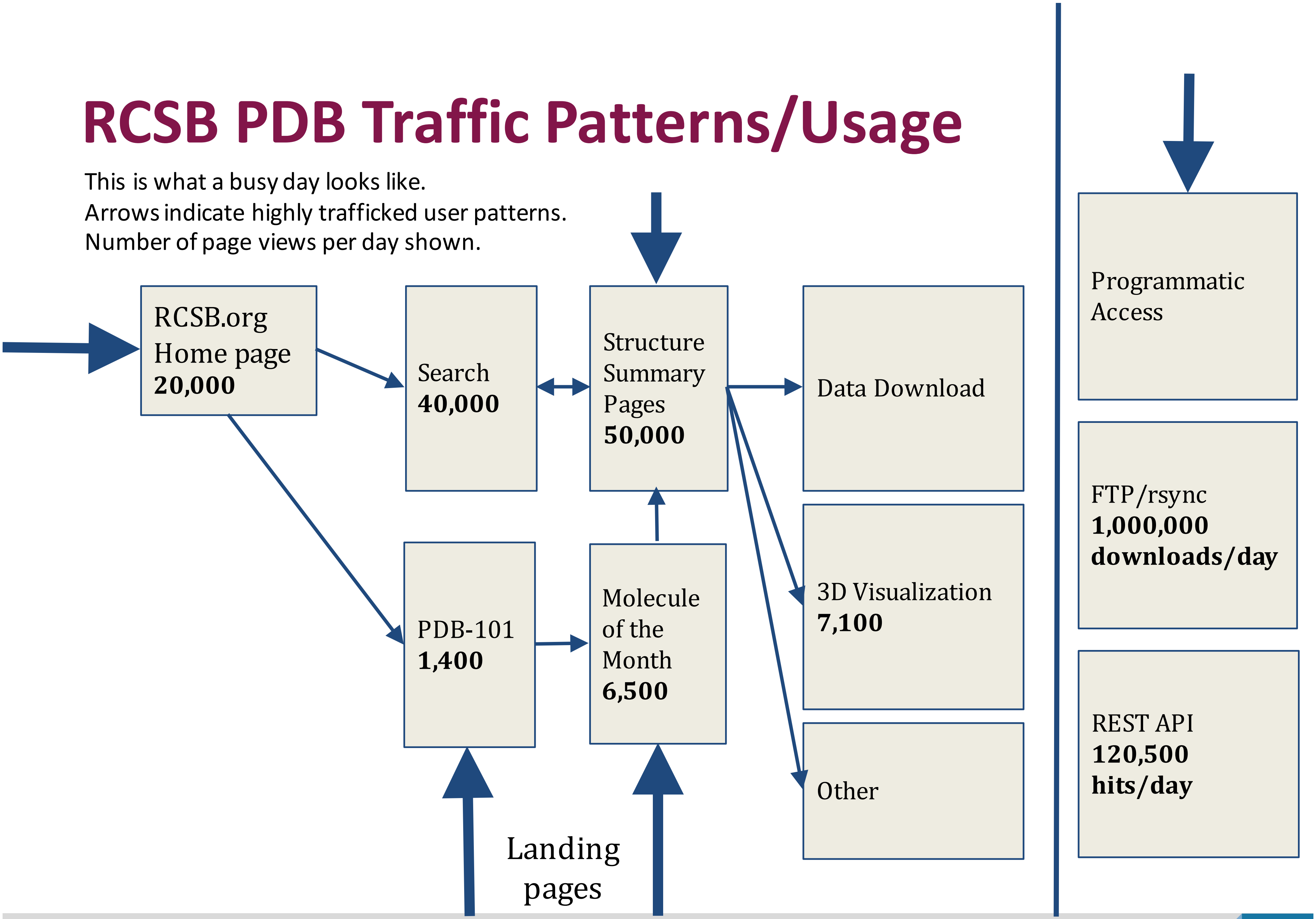
200 Icosahedral Viruses from the PDB » 08/23/2016

pdb101.rcsb.org

Google Analytics - Server Logs - External Metrics

RCSB PDB Traffic Patterns/Usage

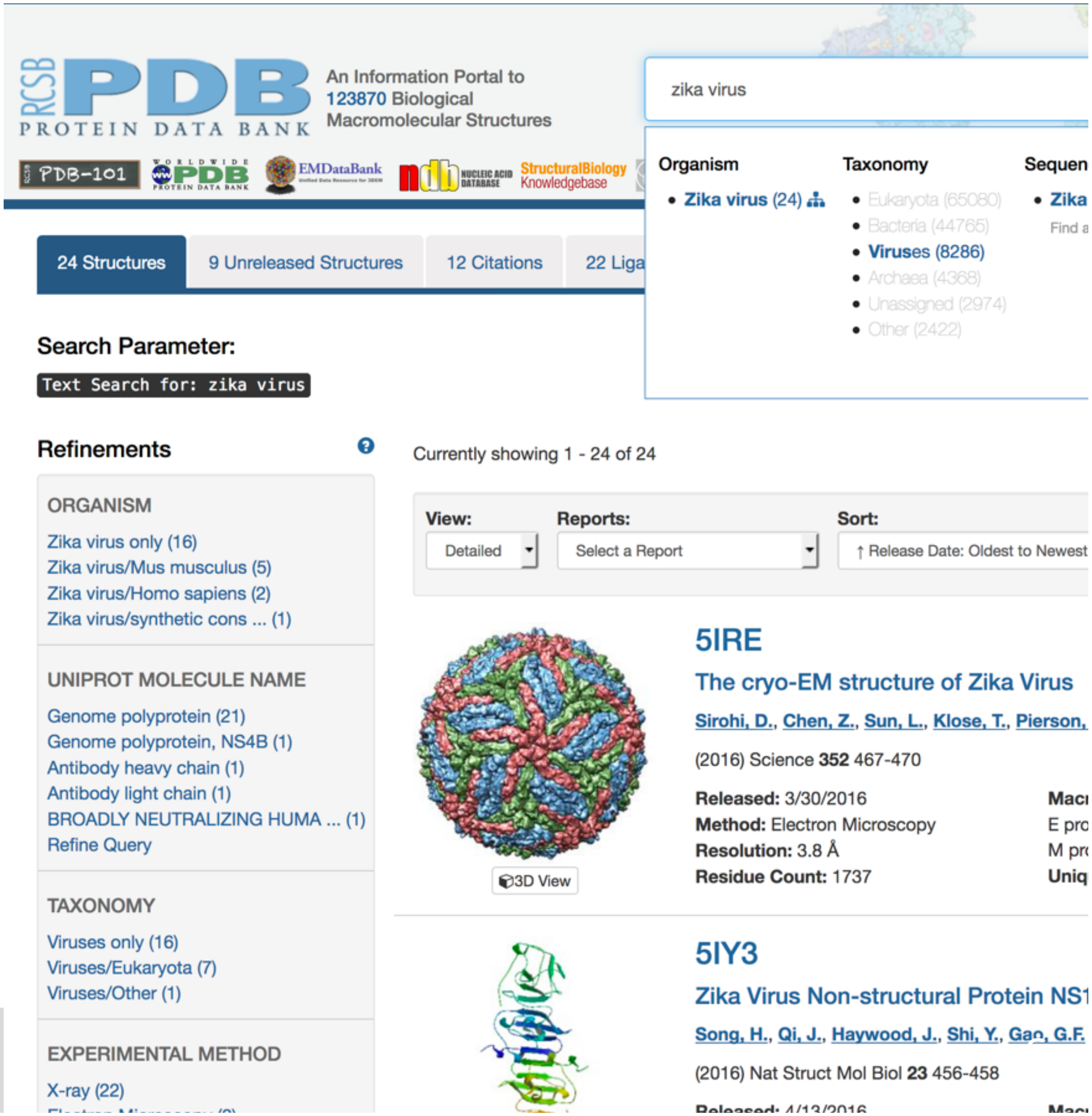
This is what a busy day looks like.
Arrows indicate highly trafficked user patterns.
Number of page views per day shown.



RCSB.org Home page **20,000**
Search **40,000**
Structure Summary Pages **50,000**
Data Download
3D Visualization **7,100**
Other
PDB-101 **1,400**
Molecule of the Month **6,500**
Landing pages
Programmatic Access
FTP/rsync **1,000,000 downloads/day**
REST API **120,500 hits/day**

Searching and Search Results

- Text search with autosuggestion
- Advanced Search
- New search features
  - Synonyms
  - Sequence clusters
  - PDB-101 content
- New Search Results page
  - Improved usability
  - Responsive layout
  - Fast page loads



An Information Portal to 123870 Biological Macromolecular Structures

zika virus

Organism
- Zika virus (24)

Taxonomy
- Eukaryota (65080)
- Bacteria (44765)
- Viruses (8286)
- Archaea (4368)
- Unassigned (2974)
- Other (2422)

Sequen
- Zika

Find a

24 Structures | 9 Unreleased Structures | 12 Citations | 22 Liga

Search Parameter:

`Text Search for: zika virus`

Refinements

ORGANISM

Zika virus only (16)
Zika virus/Mus musculus (5)
Zika virus/Homo sapiens (2)
Zika virus/synthetic cons ... (1)

UNIPROT MOLECULE NAME

Genome polyprotein (21)
Genome polyprotein, NS4B (1)
Antibody heavy chain (1)
Antibody light chain (1)
BROADLY NEUTRALIZING HUMA ... (1)
Refine Query

TAXONOMY

Viruses only (16)
Viruses/Eukaryota (7)
Viruses/Other (1)

EXPERIMENTAL METHOD

X-ray (22)
Electron Microscopy (2)

Currently showing 1 - 24 of 24

View: Detailed **Reports:** Select a Report **Sort:** ↑ Release Date: Oldest to Newest

3D View

5IRE

The cryo-EM structure of Zika Virus

Sirohi, D., Chen, Z., Sun, L., Klose, T., Pierson,

(2016) Science **352** 467-470

Released: 3/30/2016
Method: Electron Microscopy
Resolution: 3.8 Å
Residue Count: 1737

Macı
E pro
M pro
Uniq

5IY3

Zika Virus Non-structural Protein NS1

Song, H., Qi, J., Haywood, J., Shi, Y., Gao, G.F.

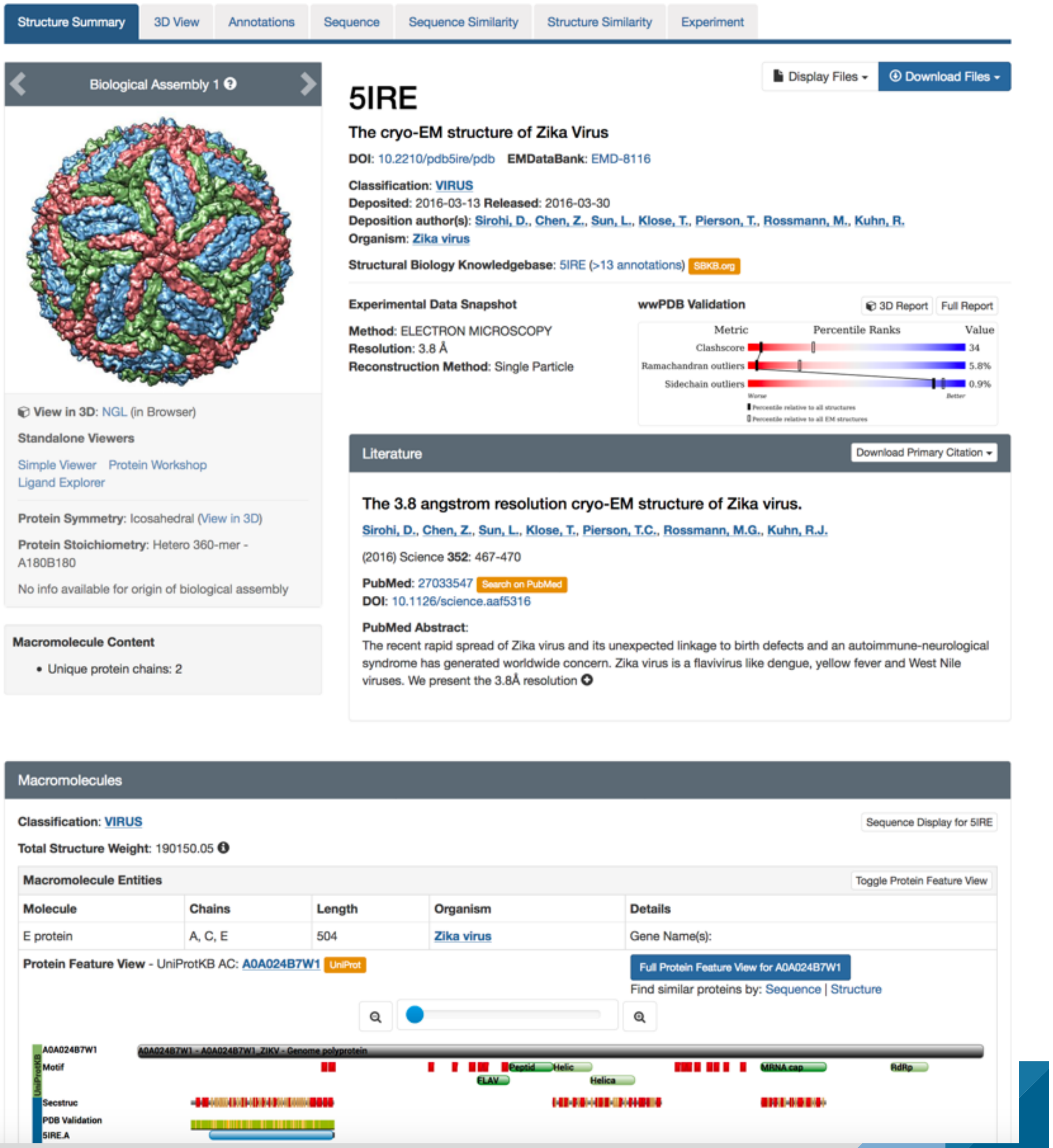
(2016) Nat Struct Mol Biol **23** 456-458

Released: 4/13/2016

Mac

Structure Summary Page

- Entry at a glance
- Detailed data organized in tabs
- New features
  - Reorganized content
  - Responsive layout
  - Integrated data views
    - Validation
    - Web-friendly NGL 3D viewer
    - Protein Feature View
    - Gene View
    - Pathway View



5IRE

The cryo-EM structure of Zika Virus

DOI: 10.2210/pdb5ire/pdb **EMDataBank:** EMD-8116

Classification: VIRUS
Deposited: 2016-03-13 **Released:** 2016-03-30
Deposition author(s): Sirohi, D., Chen, Z., Sun, L., Klose, T., Pierson, T., Rossmann, M., Kuhn, R.
Organism: Zika virus

Structural Biology Knowledgebase: 5IRE (>13 annotations)

View in 3D: NGL (in Browser)

Standalone Viewers

Simple Viewer Protein Workshop
Ligand Explorer

Protein Symmetry: Icosahedral (View in 3D)

Protein Stoichiometry: Hetero 360-mer - A180B180

No info available for origin of biological assembly

Macromolecule Content

- Unique protein chains: 2

Experimental Data Snapshot

Method: ELECTRON MICROSCOPY
Resolution: 3.8 Å
Reconstruction Method: Single Particle

wwPDB Validation

Literature

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

PubMed: 27033547
DOI: 10.1126/science.aaf5316

PubMed Abstract:
The recent rapid spread of Zika virus and its unexpected linkage to birth defects and an autoimmune-neurological syndrome has generated worldwide concern. Zika virus is a flavivirus like dengue, yellow fever and West Nile viruses. We present the 3.8Å resolution

Macromolecules

Classification: VIRUS

Total Structure Weight: 190150.05

Macromolecule Entities

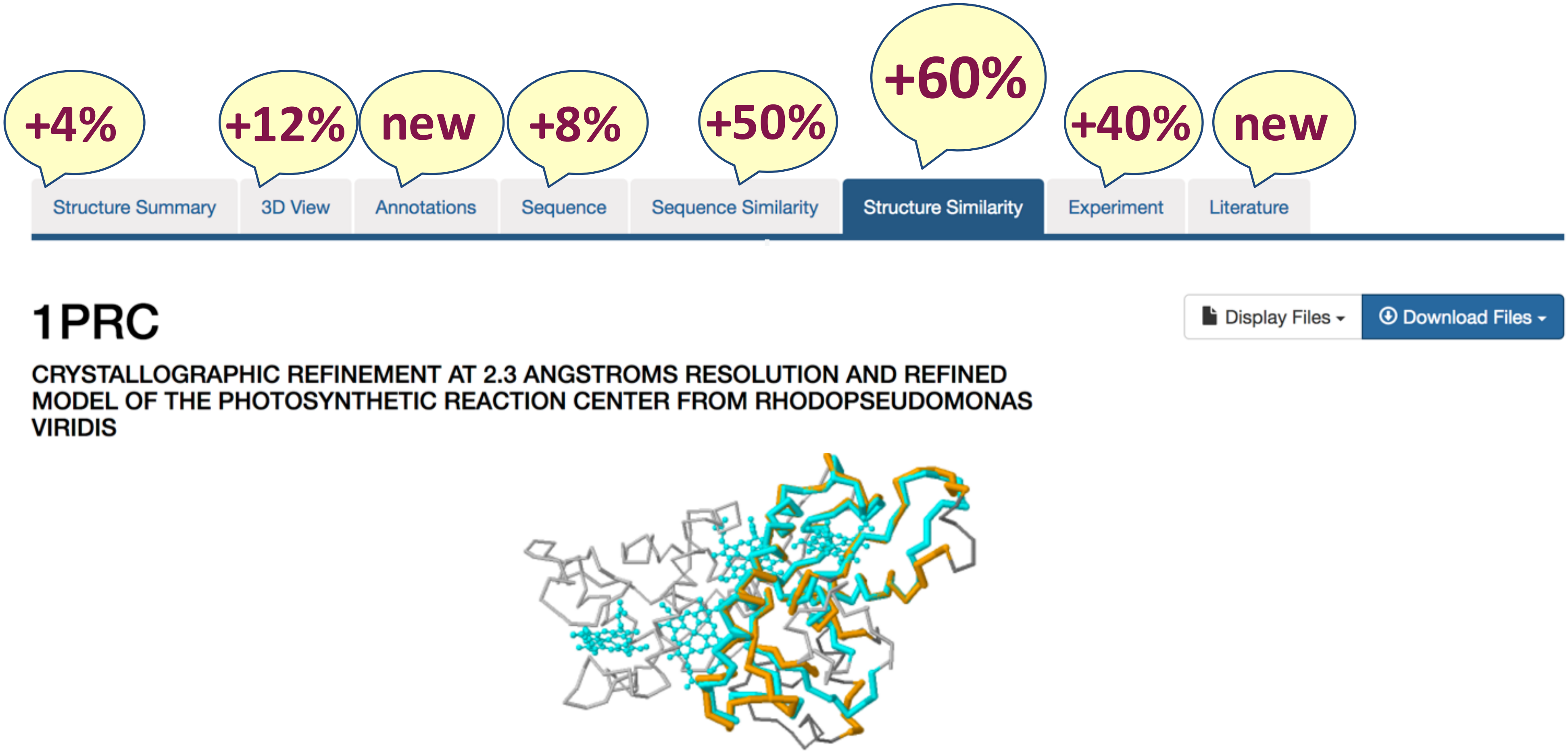
Molecule	Chains	Length	Organism	Details
E protein	A, C, E	504	Zika virus	Gene Name(s):

Protein Feature View - UniProtKB AC: A0A024B7W1

Find similar proteins by: Sequence | Structure

Impact of Design Changes

Growth in number of page views after October 2015 redesign

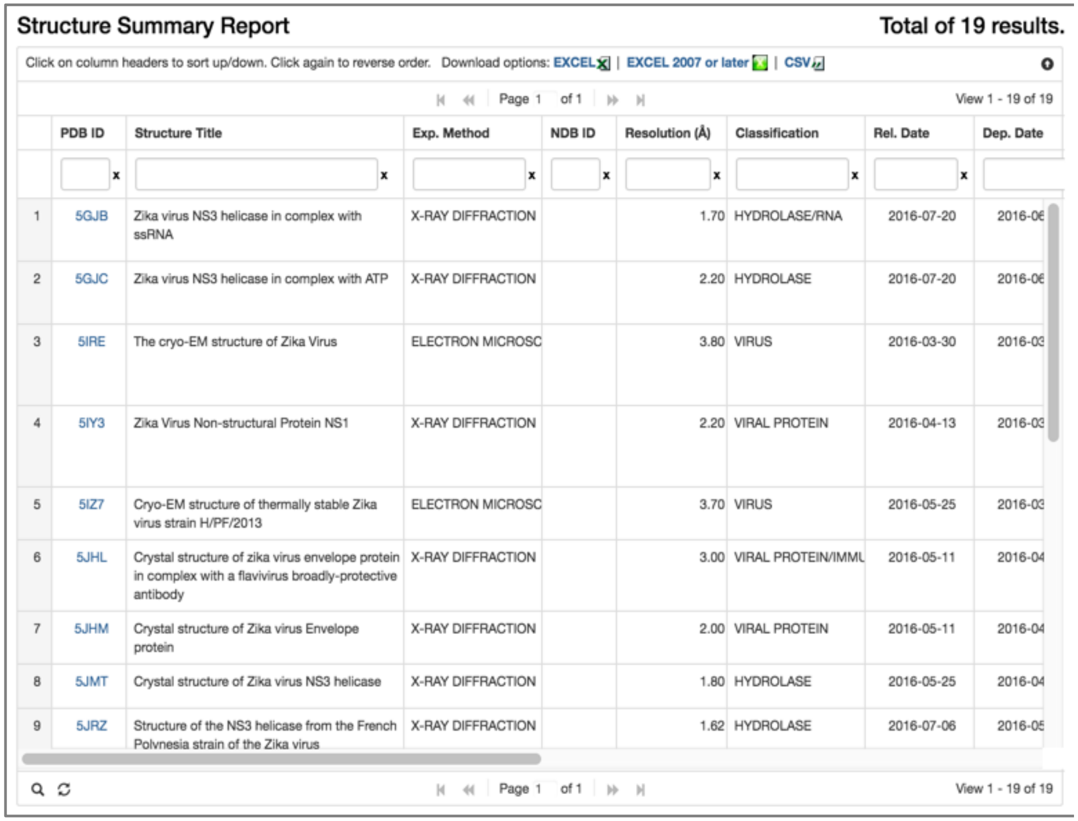


1PRC

CRYSTALLOGRAPHIC REFINEMENT AT 2.3 ANGSTROMS RESOLUTION AND REFINED MODEL OF THE PHOTOSYNTHETIC REACTION CENTER FROM RHODOPSEUDOMONAS VIRIDIS

Downloading and Reporting

- Tools used manually and programmatically
- Download structures, ligands, sequences, experimental data
- Report creation for search results
- New Features
  - User Interface overhaul
  - Sequence cluster information
  - Programmatic access improvements (RESTful web services)



Structure Summary Report **Total of 19 results.**

Click on column headers to sort up/down. Click again to reverse order. Download options: EXCEL | EXCEL 2007 or later | CSV

Page 1 of 1 View 1 - 19 of 19

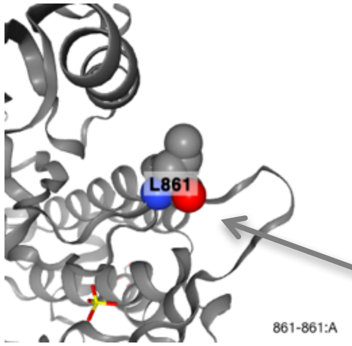
	PDB ID	Structure Title	Exp. Method	NDB ID	Resolution (Å)	Classification	Rel. Date	Dep. Date
1	5GJB	Zika virus NS3 helicase in complex with ssRNA	X-RAY DIFFRACTION		1.70	HYDROLASE/RNA	2016-07-20	2016-06
2	5GJC	Zika virus NS3 helicase in complex with ATP	X-RAY DIFFRACTION		2.20	HYDROLASE	2016-07-20	2016-06
3	5IRE	The cryo-EM structure of Zika Virus	ELECTRON MICROSC		3.80	VIRUS	2016-03-30	2016-03
4	5IY3	Zika Virus Non-structural Protein NS1	X-RAY DIFFRACTION		2.20	VIRAL PROTEIN	2016-04-13	2016-03
5	5IZ7	Cryo-EM structure of thermally stable Zika virus strain H/PF/2013	ELECTRON MICROSC		3.70	VIRUS	2016-05-25	2016-03
6	5JHL	Crystal structure of zika virus envelope protein in complex with a flavivirus broadly-protective antibody	X-RAY DIFFRACTION		3.00	VIRAL PROTEIN/IMMU	2016-05-11	2016-04
7	5JHM	Crystal structure of Zika virus Envelope protein	X-RAY DIFFRACTION		2.00	VIRAL PROTEIN	2016-05-11	2016-04
8	5JMT	Crystal structure of Zika virus NS3 helicase	X-RAY DIFFRACTION		1.80	HYDROLASE	2016-05-25	2016-04
9	5JRZ	Structure of the NS3 helicase from the French Polynesia strain of the Zika virus	X-RAY DIFFRACTION		1.62	HYDROLASE	2016-07-06	2016-05

Page 1 of 1 View 1 - 19 of 19

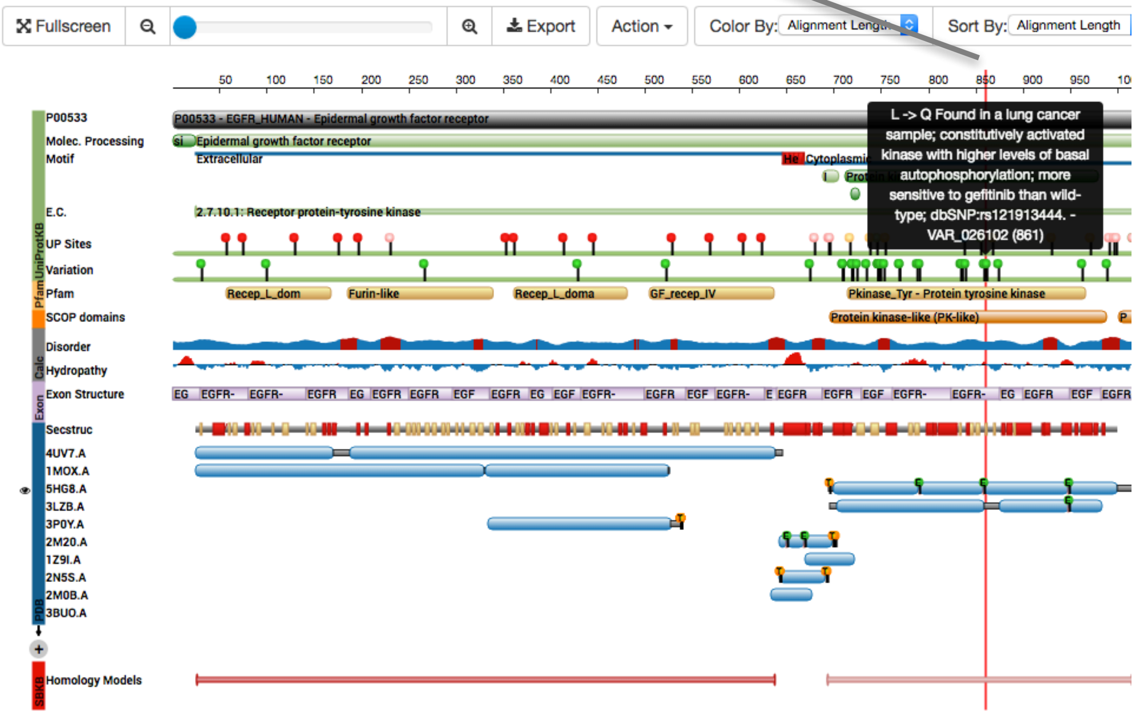
Data Integration Enabling Research I

Protein Feature View

- Sequence level view of protein features from UniProt, Pfam, PDB, Protein Model Portal, …
- New tracks
  - SNPs
  - Mutations in PDB
  - Protein modifications
  - Exon mapping
  - Link sequence position to 3D visualization



EGFR mutation
PDB 5HG8



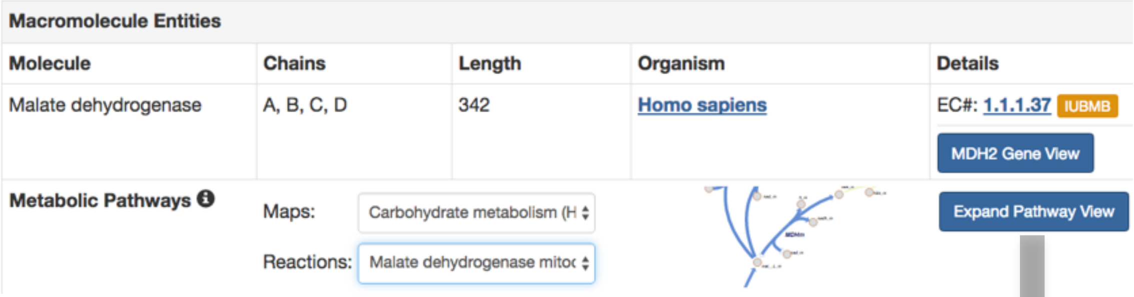
L -> Q Found in a lung cancer sample; constitutively activated kinase with higher levels of basal autophosphorylation; more sensitive to gefitinib than wild-type; dbSNP:rs121913444. - VAR_026102 (861)

Integrating Genomic Information with Protein Sequence and 3D Atomic Level Structure at the RCSB Protein Data Bank
Bioinformatics 2016 doi:10.1093/bioinformatics/btw547

Data Integration Enabling Research II

Pathway View

- Maps structures and metabolites to pathways
- Preview on Structure Summary page
- Interactive browsing
- Pathway name searching



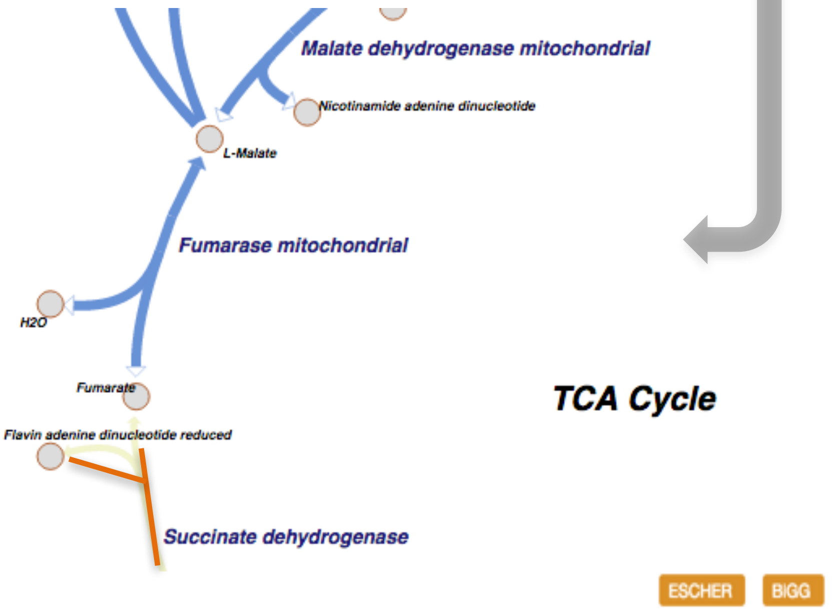
Macromolecule Entities

Molecule	Chains	Length	Organism	Details
Malate dehydrogenase	A, B, C, D	342	Homo sapiens	EC#: 1.1.1.37 IUBMB MDH2 Gene View

Metabolic Pathways

Maps: Carbohydrate metabolism (H

Reactions: Malate dehydrogenase mitoc



Malate dehydrogenase mitochondrial
Nicotinamide adenine dinucleotide
L-Malate
Fumarase mitochondrial
H2O
Fumarate
Flavin adenine dinucleotide reduced
Succinate dehydrogenase
TCA Cycle
ESCHER
BIGG

Blue: structures in PDB
Yellow: homology models

Geneticists, Genetic Counselors, Oncologists

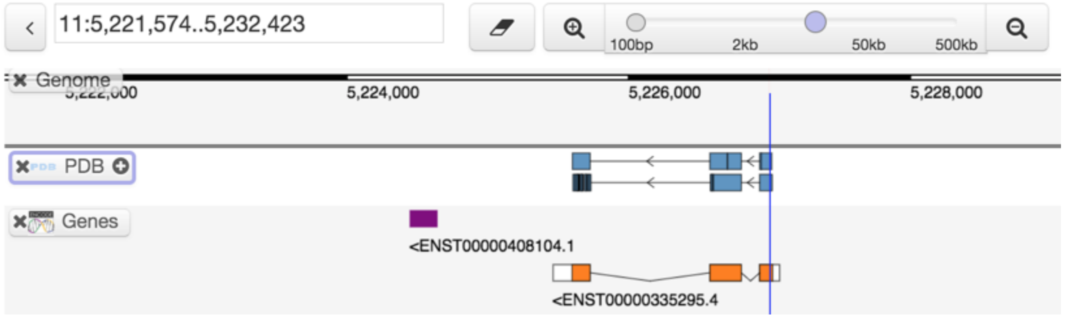
Data Integration Enabling Research III

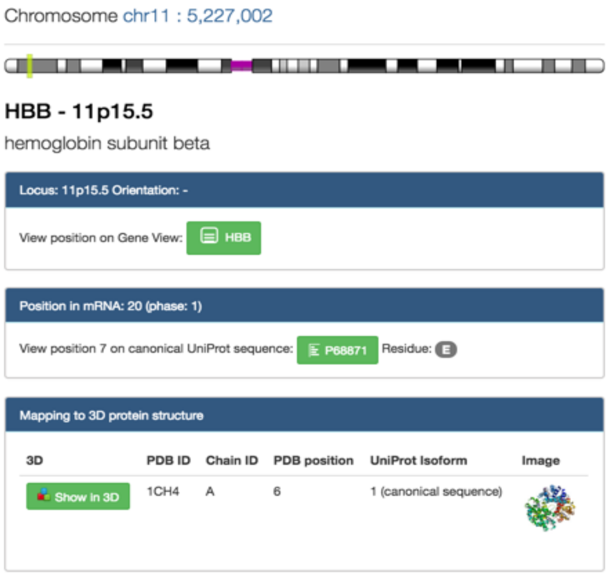
Gene View

- Mapping structural coverage onto human genome

Map Genomic Position

- To chromosome, protein position, and 3D structure





Chromosome chr11 : 5,227,002

HBB - 11p15.5

hemoglobin subunit beta

Locus: 11p15.5 Orientation: -

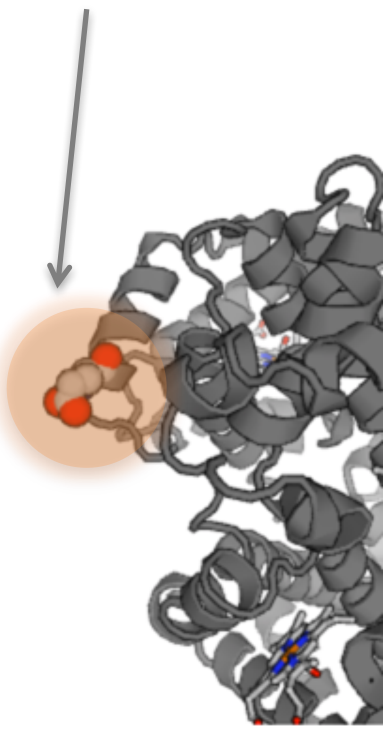
View position on Gene View: HBB

Position in mRNA: 20 (phase: 1)

View position 7 on canonical UniProt sequence: P68871 Residue: E

Mapping to 3D protein structure

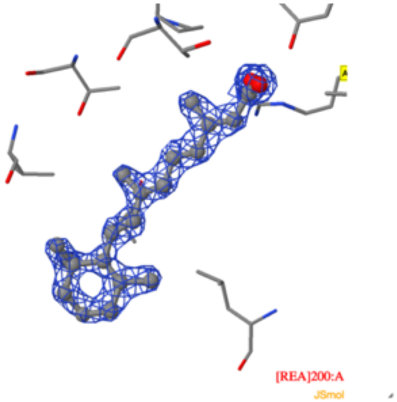
3D	PDB ID	Chain ID	PDB position	UniProt Isoform	Image
Show in 3D	1CH4	A	6	1 (canonical sequence)	



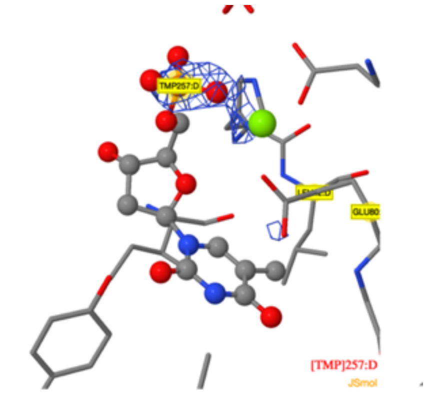
Data Integration Enabling Research IV

Electron Density

- Mini-maps to assess Ligand quality



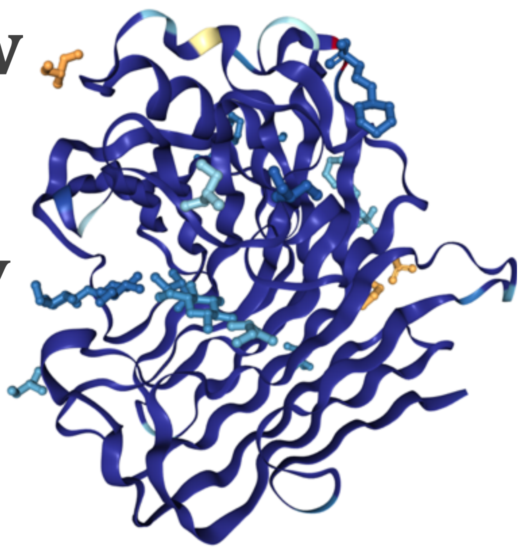
Good fit: REA in PDB 1CBS



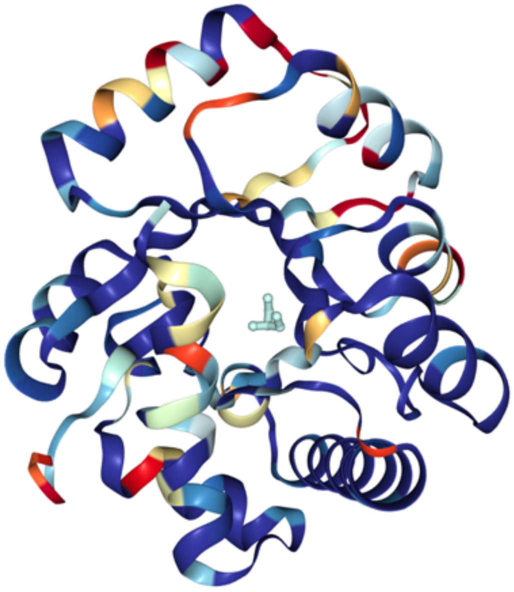
Bad fit: PDB 3HW4

NGL Validation 3D View

- Protein structure *versus* Electron density (Real Space R-factor Z score)

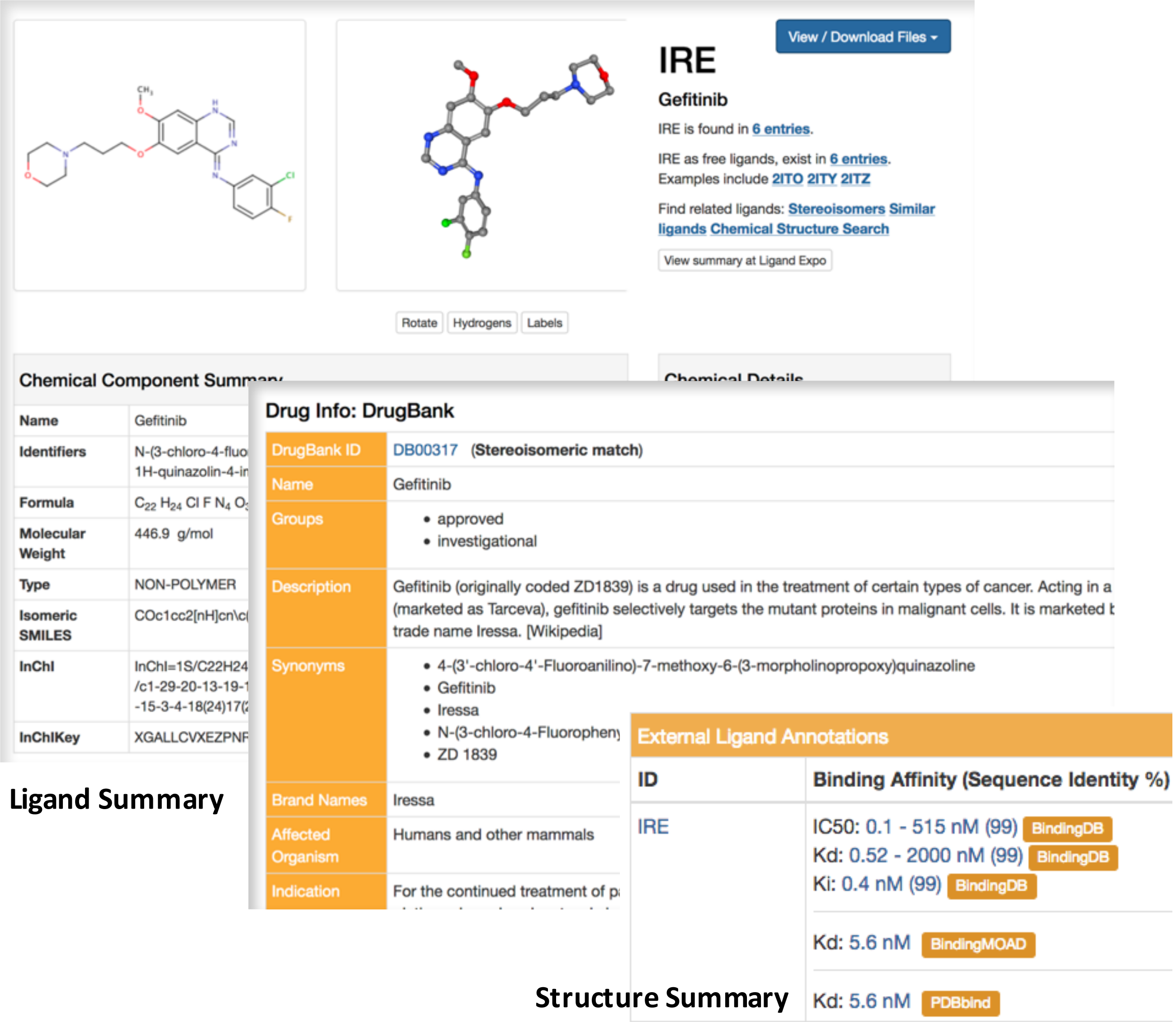


Good fit: PDB 3WY6



Bad fit: PDB 3WYZ

Enabling Drug Discovery in 3D



IRE

Gefitinib

IRE is found in 6 entries.

IRE as free ligands, exist in 6 entries. Examples include 2ITO 2ITY 2ITZ

Find related ligands: Stereoisomers Similar ligands Chemical Structure Search

View summary at Ligand Expo

Chemical Component Summary

Name	Gefitinib
Identifiers	N-(3-chloro-4-fluo 1H-quinazolin-4-ir
Formula	C_{22} H_{24} Cl F N_4 O
Molecular Weight	446.9 g/mol
Type	NON-POLYMER
Isomeric SMILES	COc1cc2[nH]cn\c(
InChI	InChI=1S/C22H24 /c1-29-20-13-19-1 -15-3-4-18(24)17(
InChIKey	XGALLCVXEZPNF

Chemical Details

Drug Info: DrugBank

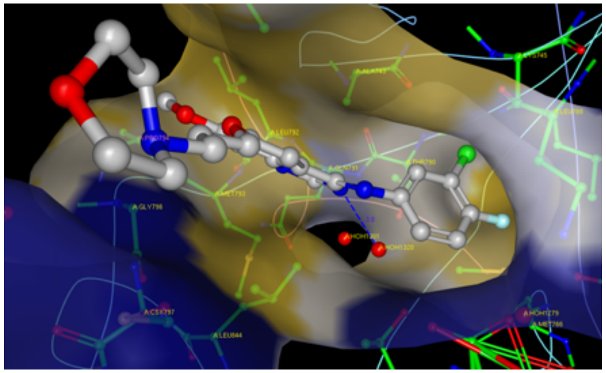
DrugBank ID	DB00317 **(Stereoisomeric match)**
Name	Gefitinib
Groups	• approved • investigational
Description	Gefitinib (originally coded ZD1839) is a drug used in the treatment of certain types of cancer. Acting in a (marketed as Tarceva), gefitinib selectively targets the mutant proteins in malignant cells. It is marketed t trade name Iressa. [Wikipedia]
Synonyms	• 4-(3'-chloro-4'-Fluoroanilino)-7-methoxy-6-(3-morpholinopropoxy)quinazoline • Gefitinib • Iressa • N-(3-chloro-4-Fluorophen • ZD 1839
Brand Names	Iressa
Affected Organism	Humans and other mammals
Indication	For the continued treatment of p

External Ligand Annotations

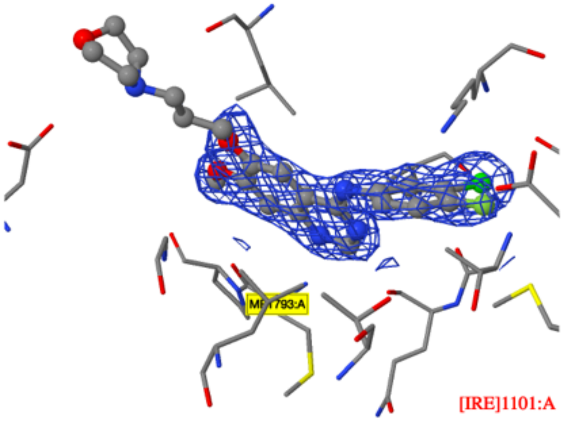
ID	Binding Affinity (Sequence Identity %)
IRE	IC50: 0.1 - 515 nM (99) BindingDB Kd: 0.52 - 2000 nM (99) BindingDB Ki: 0.4 nM (99) BindingDB
	Kd: 5.6 nM BindingMOAD
	Kd: 5.6 nM PDBbind

Ligand Summary

Structure Summary



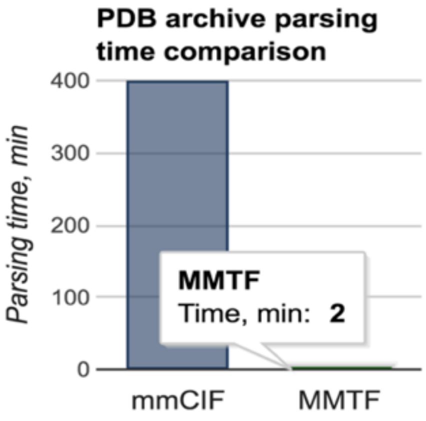
Gefitinib binding site view, PDB 4WKQ

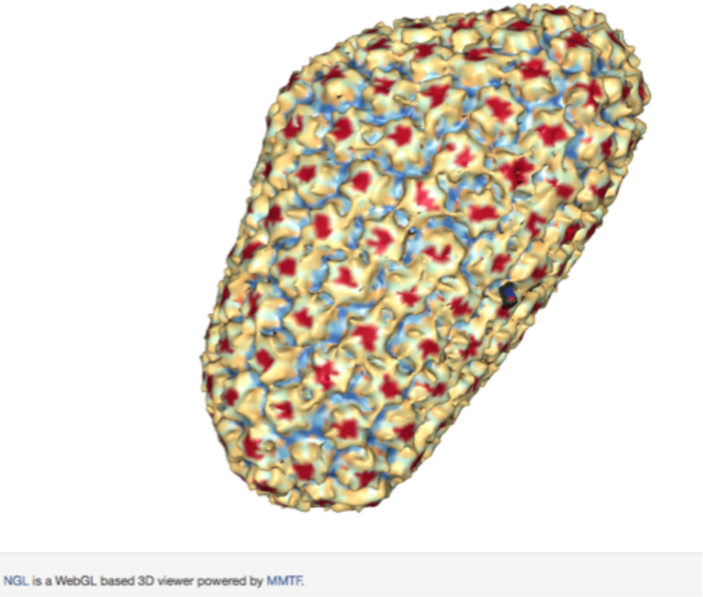


Gefitinib electron density map

Visualizing Large Structures in 3D

- 68 of the 100 largest PDB structures were deposited in the past three years
- Challenge: Data transmission and parsing
  - Developed MacroMolecular Transmission Format (MMTF)
  - Rapid adoption by community
    - Jmol, 3Dmol.js, iCn3D(NCBI), PyMol
    - BioJava, Biopython
- Challenge: Web visualization
  - NGL Viewer efficiently renders large complexes using MMTF on any device



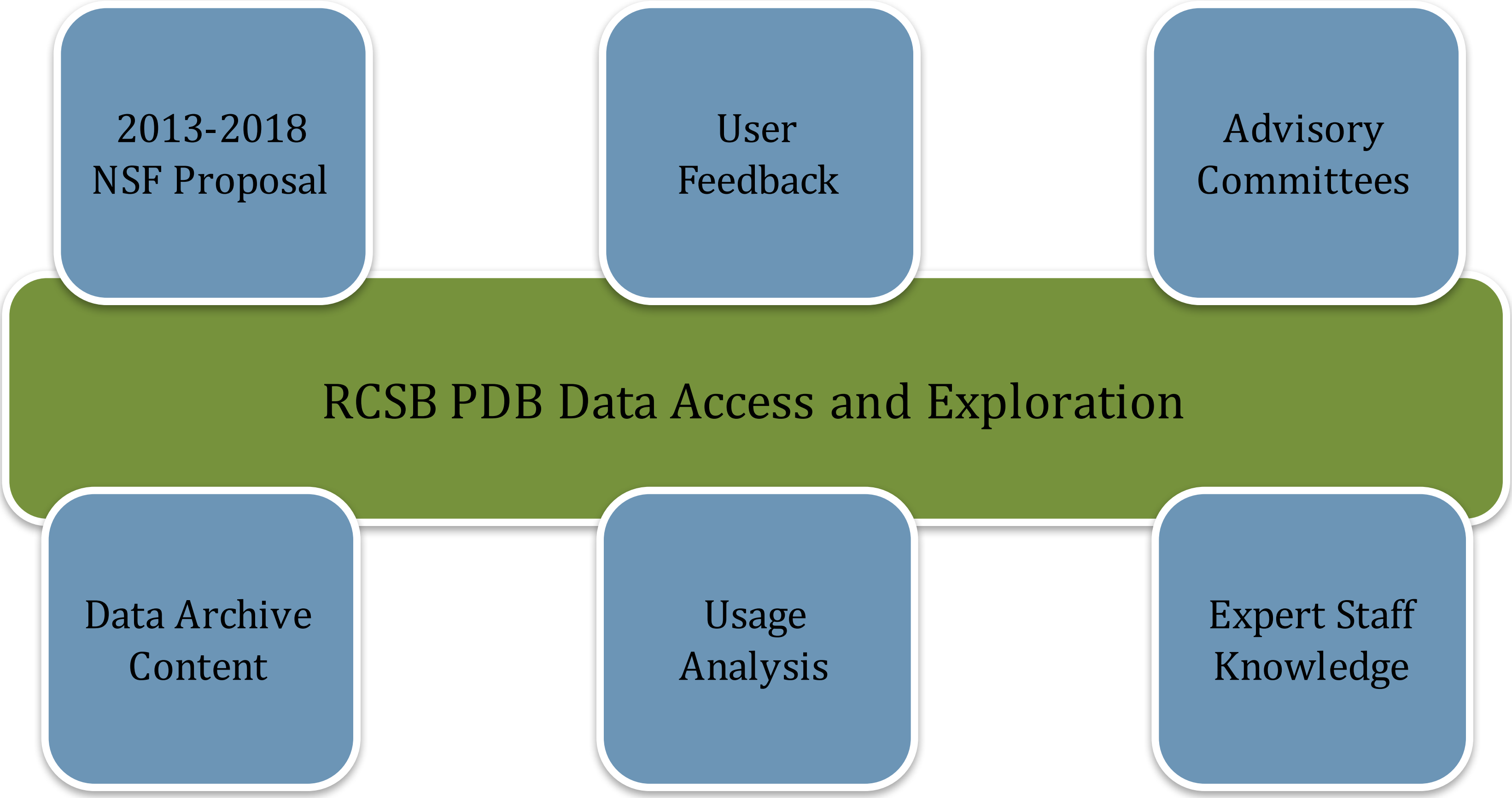


HIV Capsid PDB 3J3Q, ~2.4MM atoms

Web-based molecular graphics for large complexes
Web3D '16 Proceedings of the 21st International Conference on Web3D Technology
2016 doi: 10.1145/2945292.2945324

External Funding: NIH BD2K Award U01 CA198942

Factors Influencing Decision Making



2013-2018 NSF Proposal

User Feedback

Advisory Committees

RCSB PDB Data Access and Exploration

Data Archive Content

Usage Analysis

Expert Staff Knowledge

Website Usage

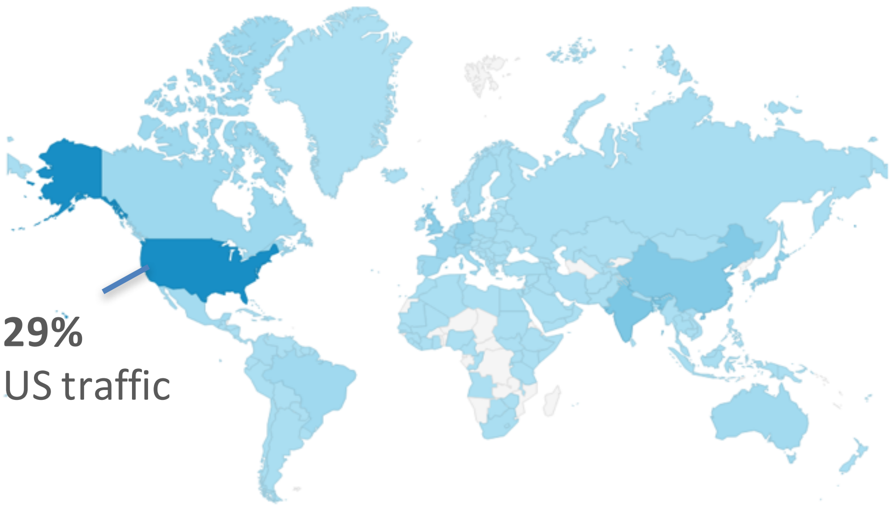
Users

- >350,000 monthly
  >1 million annually
- 2% annual growth in users (non-bounce sessions)

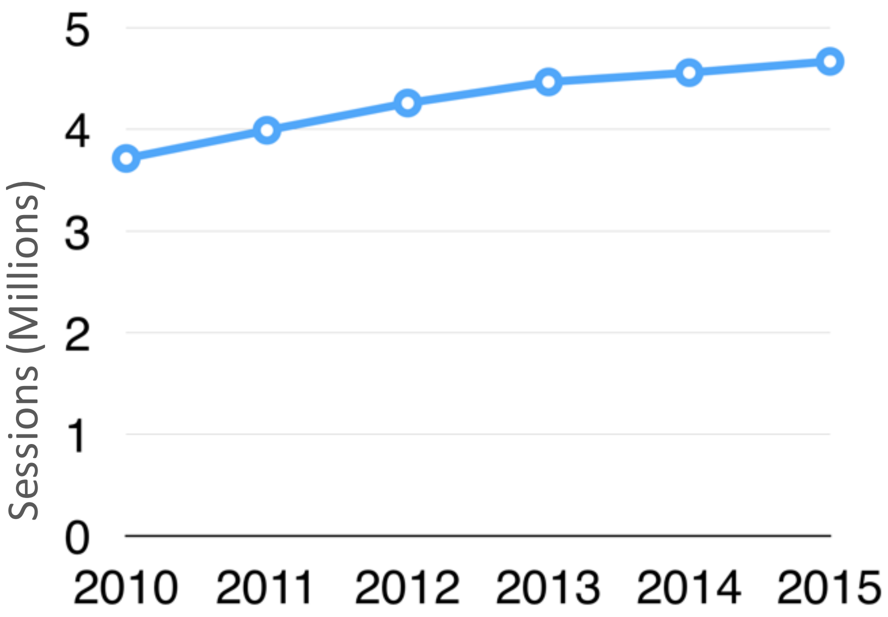
Sessions

- ~26% growth since 2010
- In 2015, ~1 million more sessions than in 2010
- High average session duration (~8 minutes)

Global Traffic


29%
US traffic

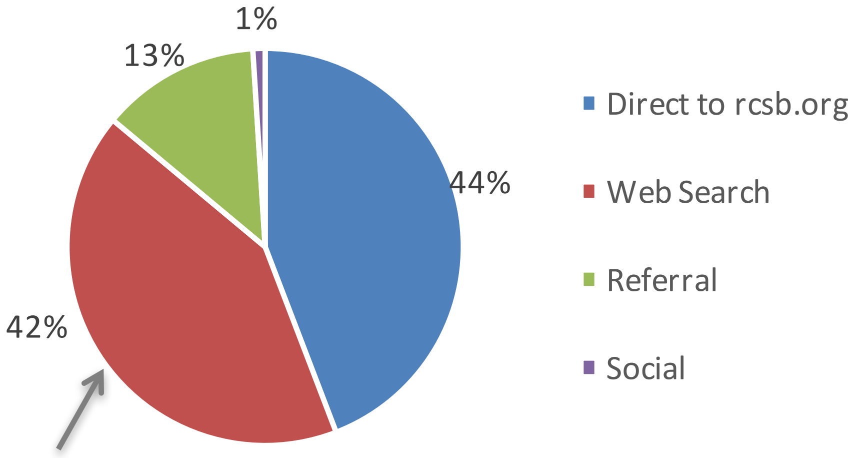
Non-bounce Sessions



Website Access

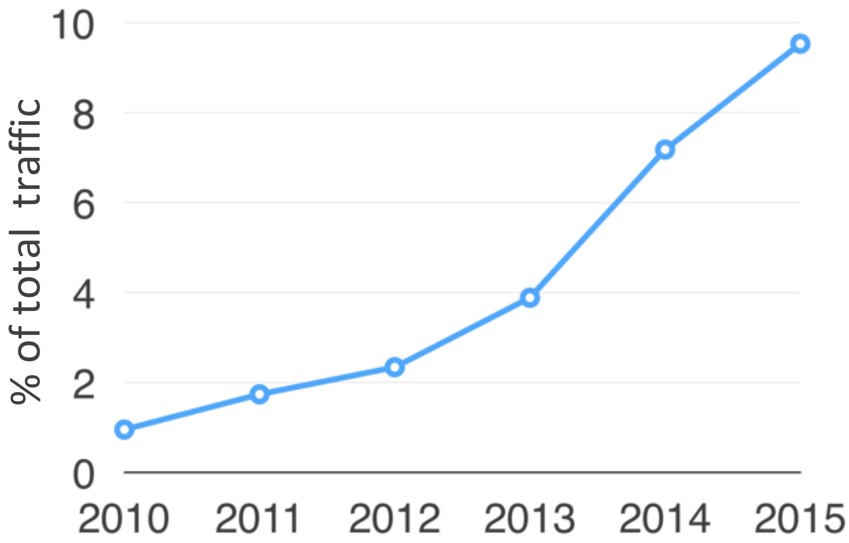
- Most traffic: direct access and web searches (*e.g.*, Google)
- Mobile usage growing rapidly
  - 10% to rcsb.org
  - 20% to PDB-101 (educational site)
- Supported by responsive layout and WebGL-based 3D visualization

Source of rcsb.org traffic



Search Engine Optimization (SEO) important

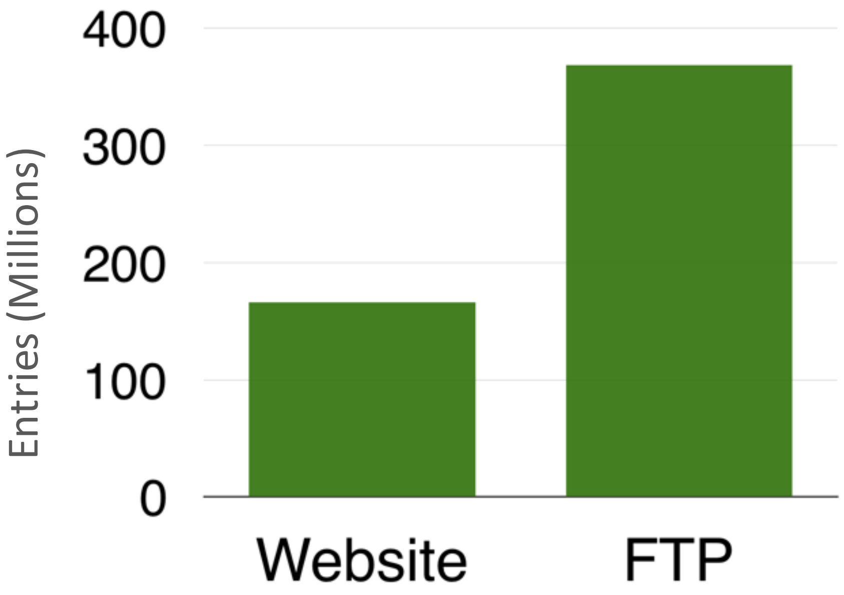
Mobile rcsb.org traffic (%)



Data Downloading/Programmatic Access

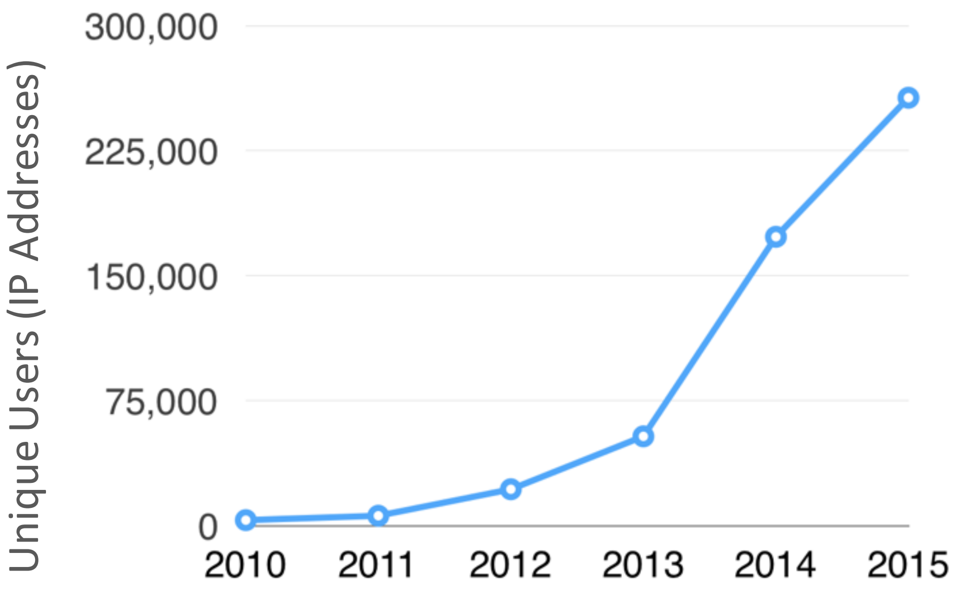
Access *via* Website and FTP

Entry Downloads and Views 2015

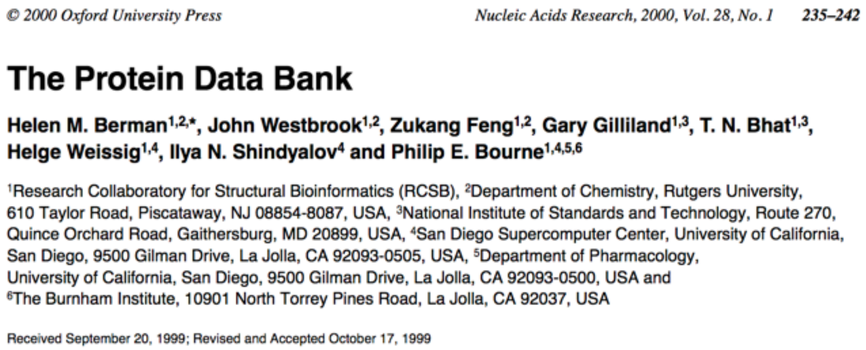


Programmatic Access (API)

RESTful Web Services:
fastest growing service



Impact: Primary RCSB PDB Publication



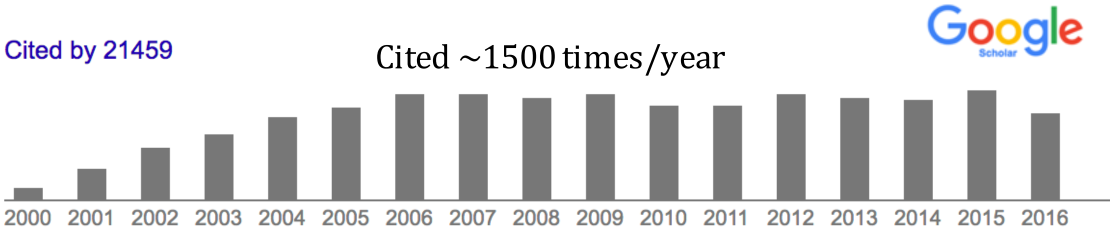
© 2000 Oxford University Press Nucleic Acids Research, 2000, Vol. 28, No. 1 235–242

The Protein Data Bank

Helen M. Berman[1,2,*], John Westbrook[1,2], Zukang Feng[1,2], Gary Gilliland[1,3], T. N. Bhat[1,3], Helge Weissig[1,4], Ilya N. Shindyalov[4] and Philip E. Bourne[1,4,5,6]

[1]Research Collaboratory for Structural Bioinformatics (RCSB), [2]Department of Chemistry, Rutgers University, 610 Taylor Road, Piscataway, NJ 08854-8087, USA, [3]National Institute of Standards and Technology, Route 270, Quince Orchard Road, Gaithersburg, MD 20899, USA, [4]San Diego Supercomputer Center, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0505, USA, [5]Department of Pharmacology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0500, USA and [6]The Burnham Institute, 10901 North Torrey Pines Road, La Jolla, CA 92037, USA

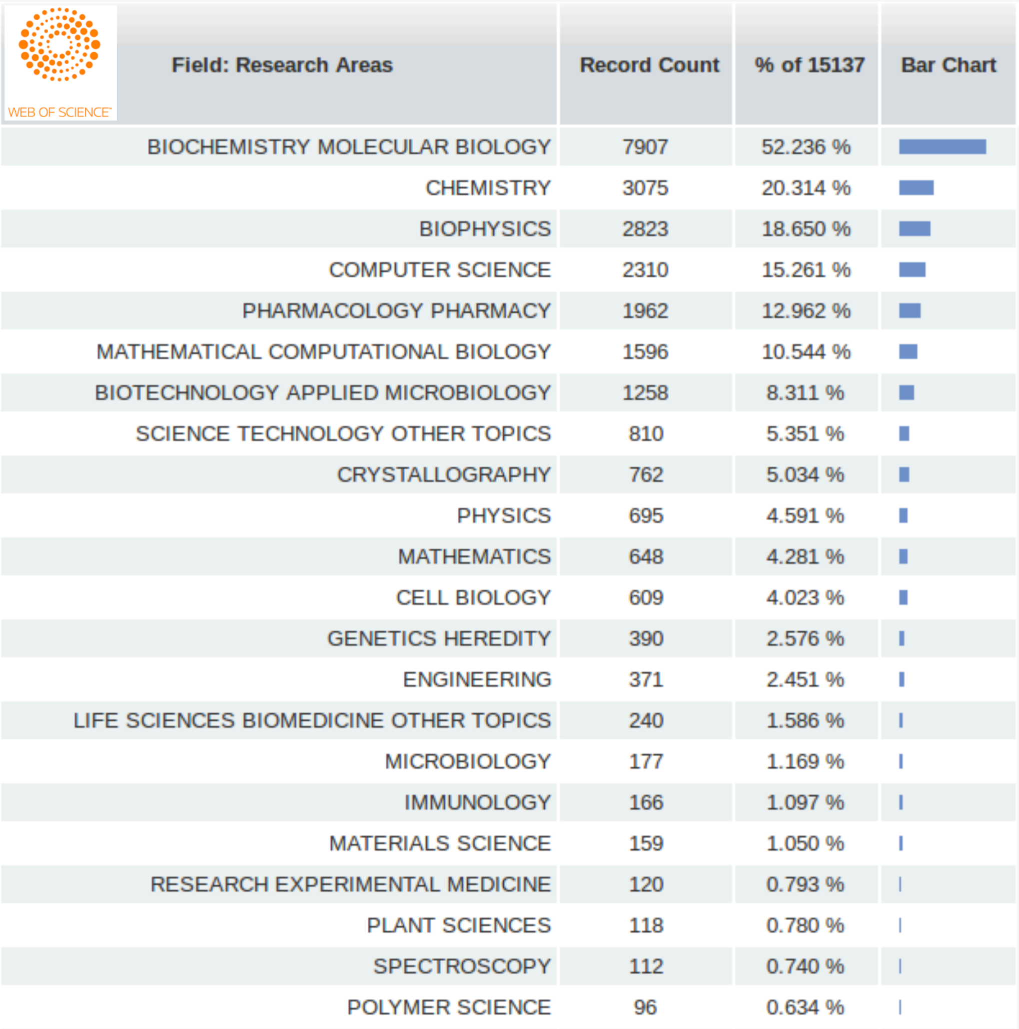
Received September 20, 1999; Revised and Accepted October 17, 1999



Cited by 21459

Cited ~1500 times/year

NEWS FEATURE
THE TOP 100 PAPERS
Nature explores the most-cited research of all time.
BY RICHARD VAN NOORDEN, BRENDAN MAHER AND REGINA NUZZO



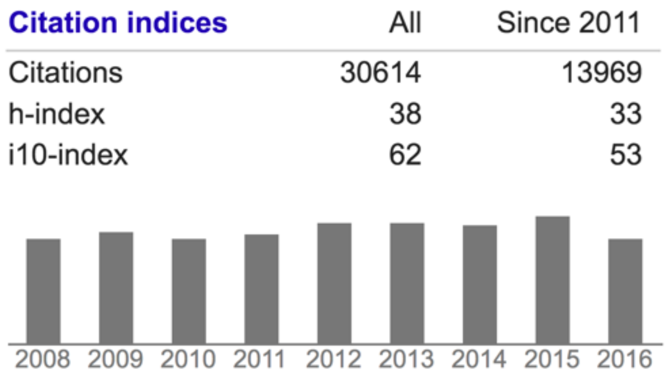
WEB OF SCIENCE

Field: Research Areas	Record Count	% of 15137	Bar Chart
BIOCHEMISTRY MOLECULAR BIOLOGY	7907	52.236 %	▬
CHEMISTRY	3075	20.314 %	▬
BIOPHYSICS	2823	18.650 %	▬
COMPUTER SCIENCE	2310	15.261 %	▬
PHARMACOLOGY PHARMACY	1962	12.962 %	▬
MATHEMATICAL COMPUTATIONAL BIOLOGY	1596	10.544 %	▬
BIOTECHNOLOGY APPLIED MICROBIOLOGY	1258	8.311 %	▬
SCIENCE TECHNOLOGY OTHER TOPICS	810	5.351 %	▬
CRYSTALLOGRAPHY	762	5.034 %	▬
PHYSICS	695	4.591 %	▬
MATHEMATICS	648	4.281 %	▬
CELL BIOLOGY	609	4.023 %	▬
GENETICS HEREDITY	390	2.576 %	|
ENGINEERING	371	2.451 %	|
LIFE SCIENCES BIOMEDICINE OTHER TOPICS	240	1.586 %	|
MICROBIOLOGY	177	1.169 %	|
IMMUNOLOGY	166	1.097 %	|
MATERIALS SCIENCE	159	1.050 %	|
RESEARCH EXPERIMENTAL MEDICINE	120	0.793 %	|
PLANT SCIENCES	118	0.780 %	|
SPECTROSCOPY	112	0.740 %	|
POLYMER SCIENCE	96	0.634 %	|

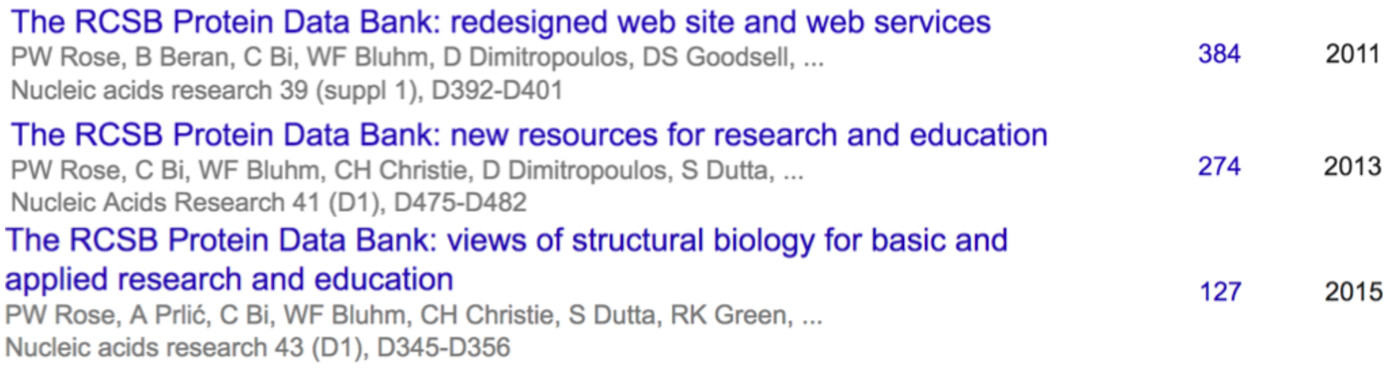
Impact: All RCSB PDB Publications

- 116 publications 2000→
- h-index: ~38
- i10-index: ~62
- Aggregate citations: ~2400/year

Google Scholar



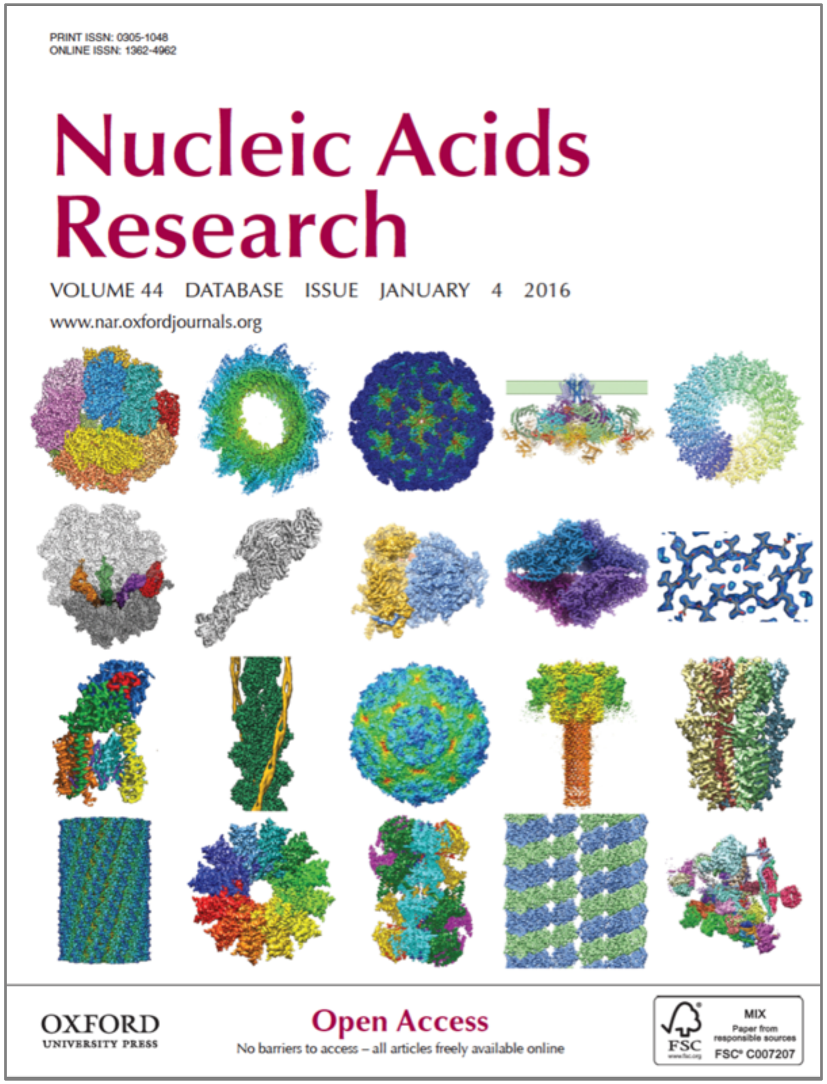
Citation indices	All	Since 2011
Citations	30614	13969
h-index	38	33
i10-index	62	53



The RCSB Protein Data Bank: redesigned web site and web services PW Rose, B Beran, C Bi, WF Bluhm, D Dimitropoulos, DS Goodsell, ... Nucleic acids research 39 (suppl 1), D392-D401	384	2011
The RCSB Protein Data Bank: new resources for research and education PW Rose, C Bi, WF Bluhm, CH Christie, D Dimitropoulos, S Dutta, ... Nucleic Acids Research 41 (D1), D475-D482	274	2013
The RCSB Protein Data Bank: views of structural biology for basic and applied research and education PW Rose, A Prlić, C Bi, WF Bluhm, CH Christie, S Dutta, RK Green, ... Nucleic acids research 43 (D1), D345-D356	127	2015

Impact: PDB Data Reuse

- PDB data used by >200 biological databases
  - Based on databases publishing in *NAR* 2011-2016
  - 11 Categories: Structure, Protein Sequence, Nucleotide Sequence, RNA Sequence, Genomics, Metabolic and Signaling, Human Genes and Diseases, Immunology, Proteomics, Plant, Other
- Since 2011, >25% of new databases utilize PDB data (119 out of 452 new databases)



PRINT ISSN: 0305-1048
ONLINE ISSN: 1362-4962

Nucleic Acids Research

VOLUME 44 DATABASE ISSUE JANUARY 4 2016

www.nar.oxfordjournals.org

OXFORD
UNIVERSITY PRESS

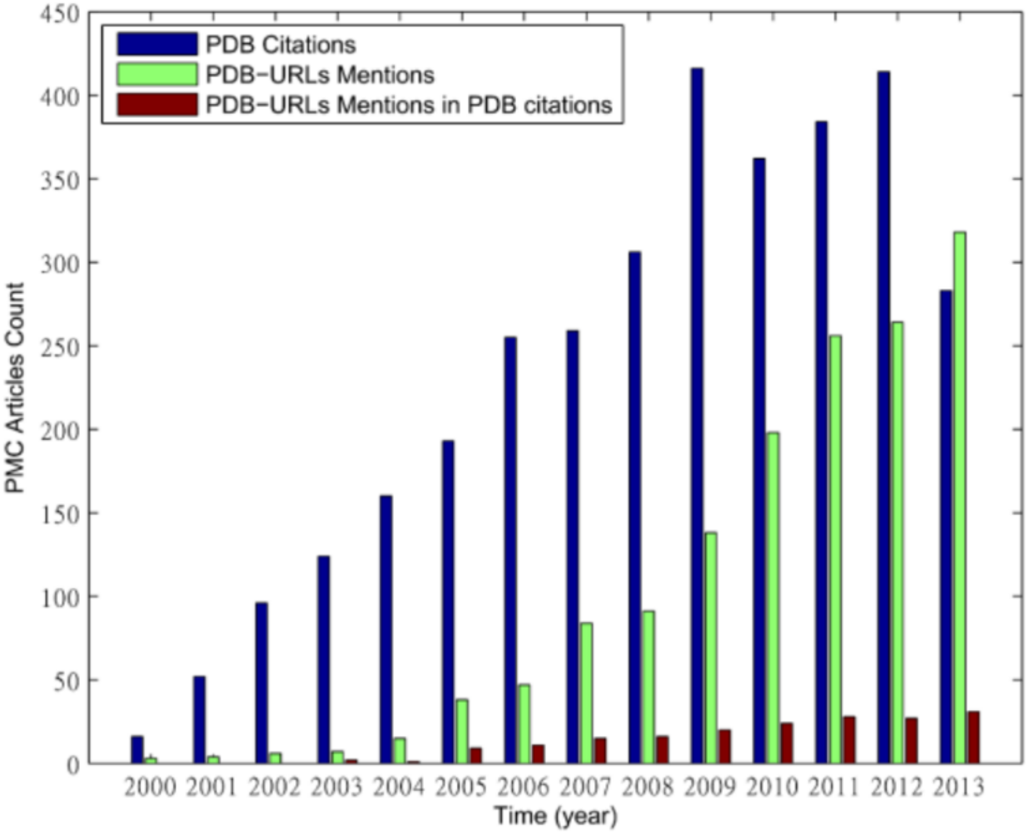
Open Access

No barriers to access – all articles freely available online

MIX
Paper from responsible sources
FSC
FSC® C007207

Citations in PMC Open Access Articles

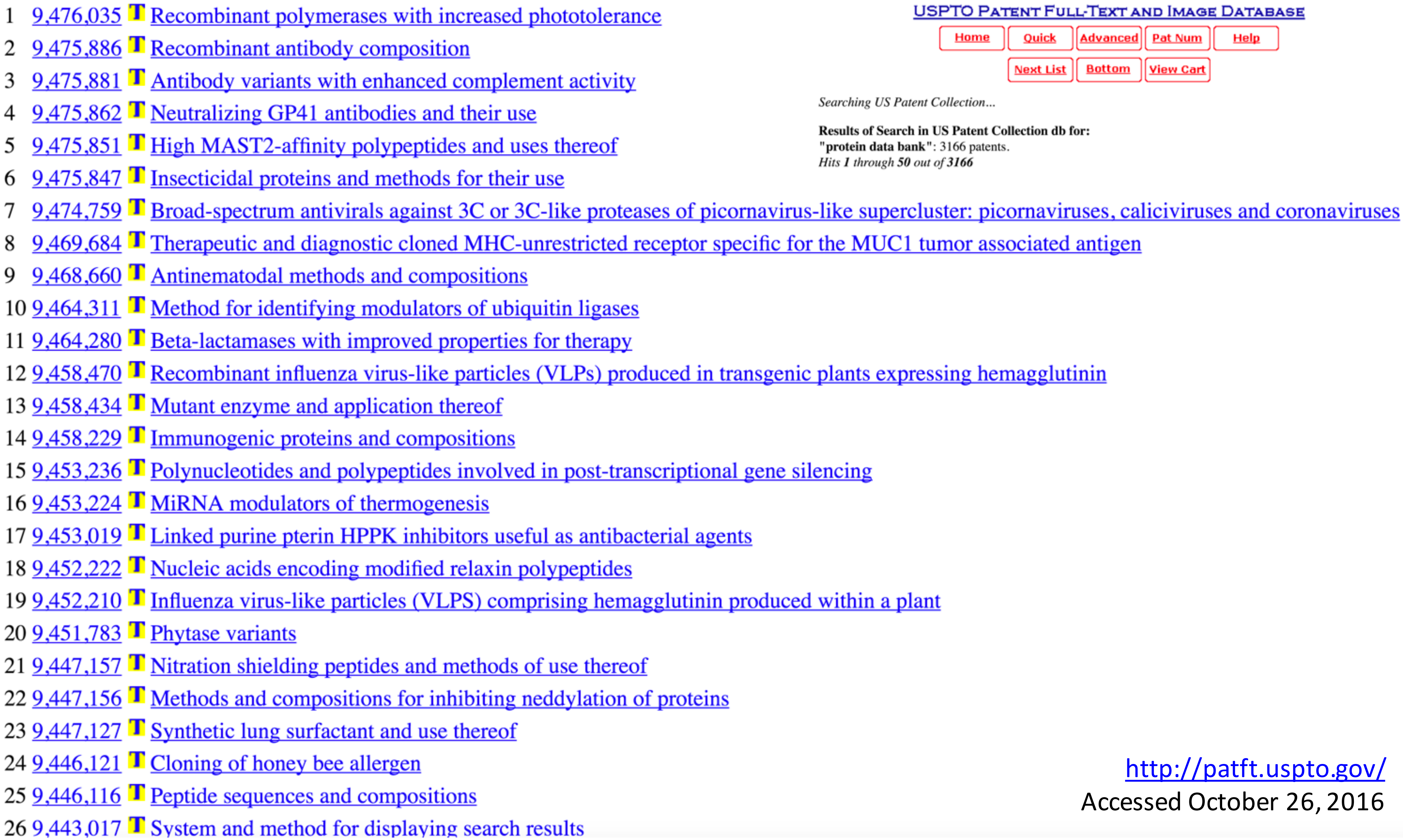
- Articles either cite the original PDB publication (Berman NAR 2000) or mention URL rcsb.org
  - Rarely are both referenced
- URL mentions are rising rapidly as data source references
- Citation statistics significantly underestimate the impact of the PDB data resource



Citing a Data Repository: A Case Study of the Protein Data Bank (2015) *PLoS ONE* 10(8): e0136631 doi:10.1371/journal.pone.0136631

External Funding: BioCADDIE Pilot Project 3.2, NIH U24AI117966

3166 Patents Mention "protein data bank"



Searching US Patent Collection...

Results of Search in US Patent Collection db for:
"protein data bank": 3166 patents.
*Hits **1** through **50** out of **3166***

1 9,476,035 T Recombinant polymerases with increased phototolerance
2 9,475,886 T Recombinant antibody composition
3 9,475,881 T Antibody variants with enhanced complement activity
4 9,475,862 T Neutralizing GP41 antibodies and their use
5 9,475,851 T High MAST2-affinity polypeptides and uses thereof
6 9,475,847 T Insecticidal proteins and methods for their use
7 9,474,759 T Broad-spectrum antivirals against 3C or 3C-like proteases of picornavirus-like supercluster: picornaviruses, caliciviruses and coronaviruses
8 9,469,684 T Therapeutic and diagnostic cloned MHC-unrestricted receptor specific for the MUC1 tumor associated antigen
9 9,468,660 T Antinematodal methods and compositions
10 9,464,311 T Method for identifying modulators of ubiquitin ligases
11 9,464,280 T Beta-lactamases with improved properties for therapy
12 9,458,470 T Recombinant influenza virus-like particles (VLPs) produced in transgenic plants expressing hemagglutinin
13 9,458,434 T Mutant enzyme and application thereof
14 9,458,229 T Immunogenic proteins and compositions
15 9,453,236 T Polynucleotides and polypeptides involved in post-transcriptional gene silencing
16 9,453,224 T MiRNA modulators of thermogenesis
17 9,453,019 T Linked purine pterin HPPK inhibitors useful as antibacterial agents
18 9,452,222 T Nucleic acids encoding modified relaxin polypeptides
19 9,452,210 T Influenza virus-like particles (VLPS) comprising hemagglutinin produced within a plant
20 9,451,783 T Phytase variants
21 9,447,157 T Nitration shielding peptides and methods of use thereof
22 9,447,156 T Methods and compositions for inhibiting neddylation of proteins
23 9,447,127 T Synthetic lung surfactant and use thereof
24 9,446,121 T Cloning of honey bee allergen
25 9,446,116 T Peptide sequences and compositions
26 9,443,017 T System and method for displaying search results

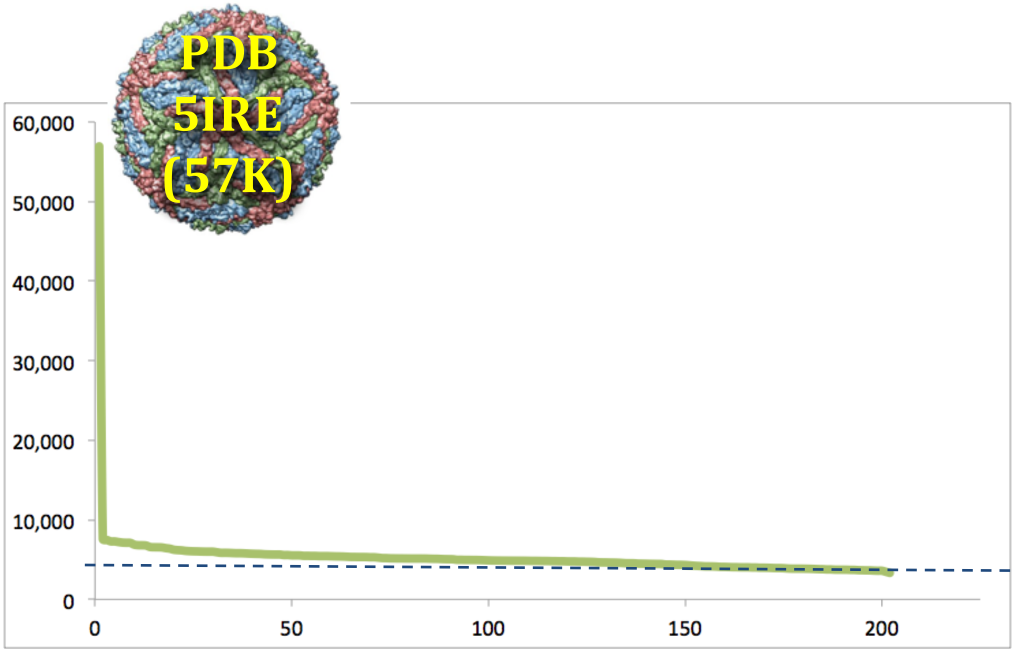
http://patft.uspto.gov/
Accessed October 26, 2016

Case Study: Zika Virus Data Release

- Zika virus structure PDB 5IRE released March 30, 2016 (Sirohi et al., 2016)
- Downloaded (8K) and viewed on website (49K) times
- >10x usage *versus* 201 other entries released same week
  - Average ~5K

Data Downloads & Web Views

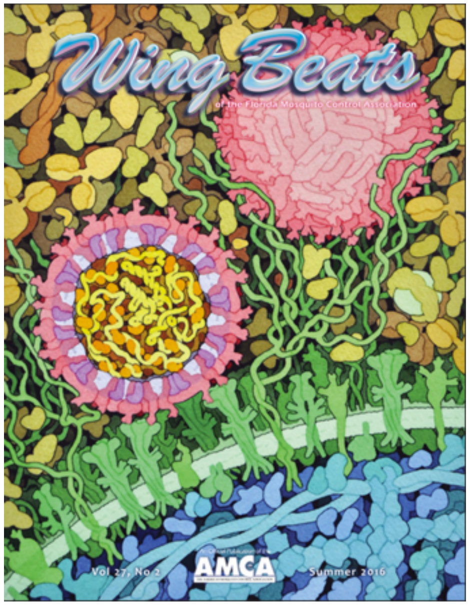
March 30-October 4, 2016

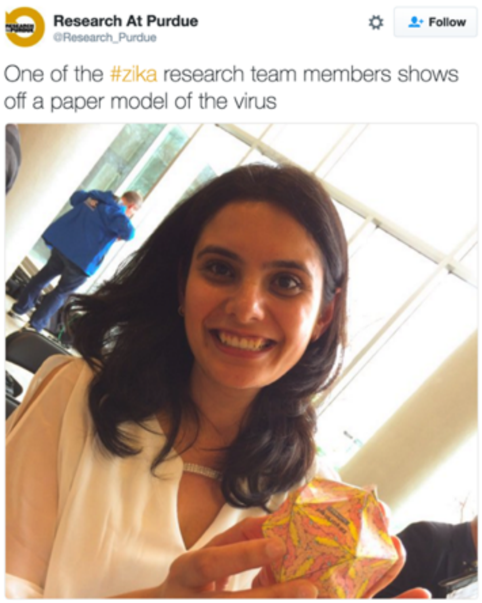


202 structures released March 30, 2016, sorted by usage frequency

Related Outreach

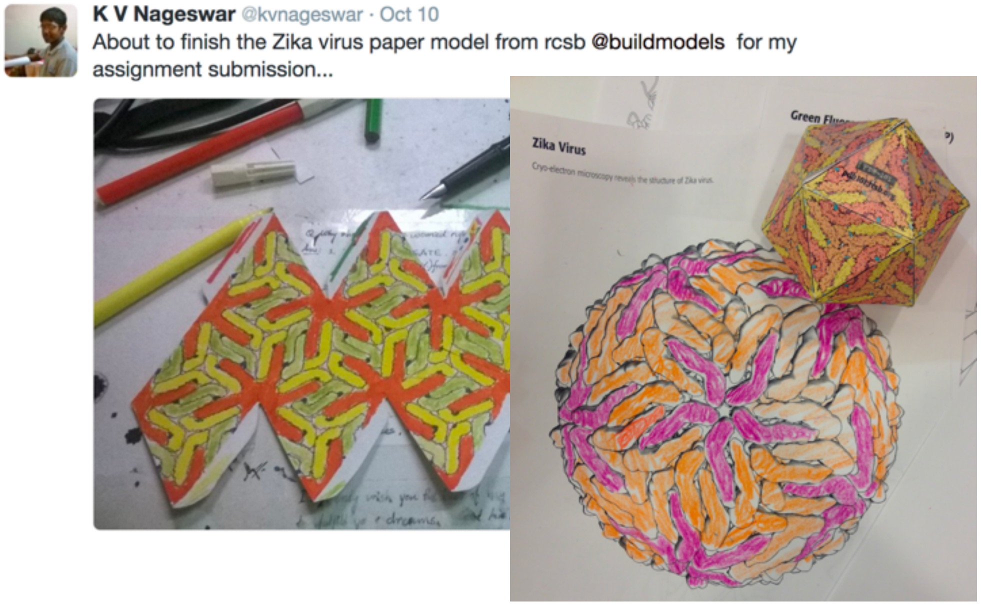
- Molecule of the Month May 2016
  - Zika watercolor highlighted on Cover of *Cell Host & Microbe* and many blogs (NPR, NIH Director, Smithsonian, *Anthropology News*, ...)
  - ~12,000 page views
- Molecular Origami PDF paper models
- RCSB PDB Coloring Book





Research At Purdue @Research_Purdue

One of the #zika research team members shows off a paper model of the virus



K V Nageswar @kvnageswar · Oct 10

About to finish the Zika virus paper model from rcsb @buildmodels for my assignment submission...

Zika Virus
Cryo-electron microscopy reveals the structure of Zika virus

Infrastructure

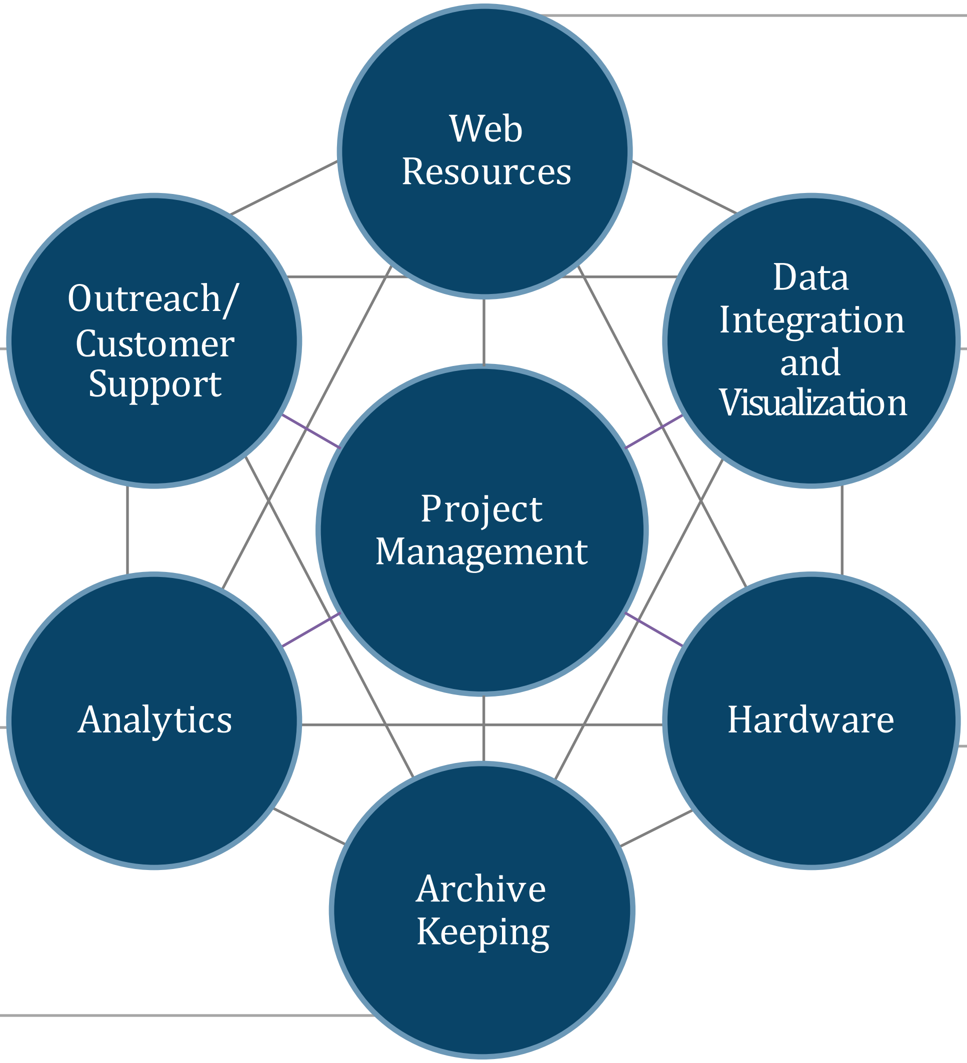
- Hosted at SDSC/UCSD and Rutgers
- Disaster preparedness
- Geographic load balancing
- Private cloud
  - Expanded capacity
  - More flexibility → on demand resourcing
  - Better analytics
- High accessibility

Future Needs

Continue to transform PDB Data → Knowledge for growing and diverse User base:

- Data Consumers → expand breadth
  - Easy access and understanding of data
  - Customized views
- Power Users → expand depth
  - 3D queries and mining of PDB
  - Web services (API)

Data Out Plans Forward



Protein family views, 3D structural queries, APIs

Communicate with diverse User communities, develop reusable open source components

Map genomic (variation) data to structure, link to homology model resources

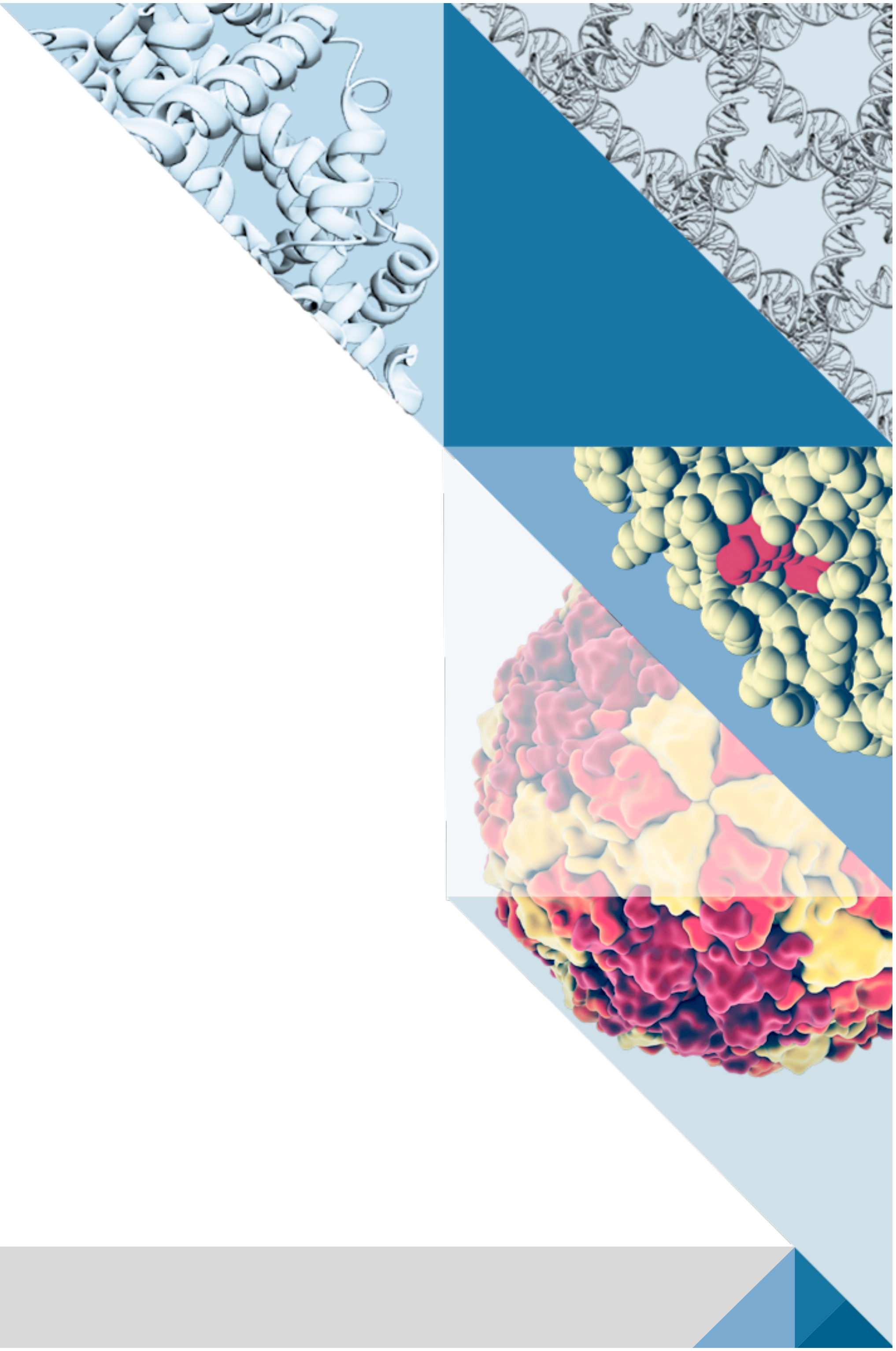
Assess usage to inform development (e.g., A/B Testing)

Expand cloud infrastructure to accommodate growing user needs and new features

Support 24/7/366 Operations, Weekly Updates, Annual Snapshots

Funding for Core Activities: NSF [DBI-1338415], NIH, DOE
External Funding: BD2K , BioCADDIE

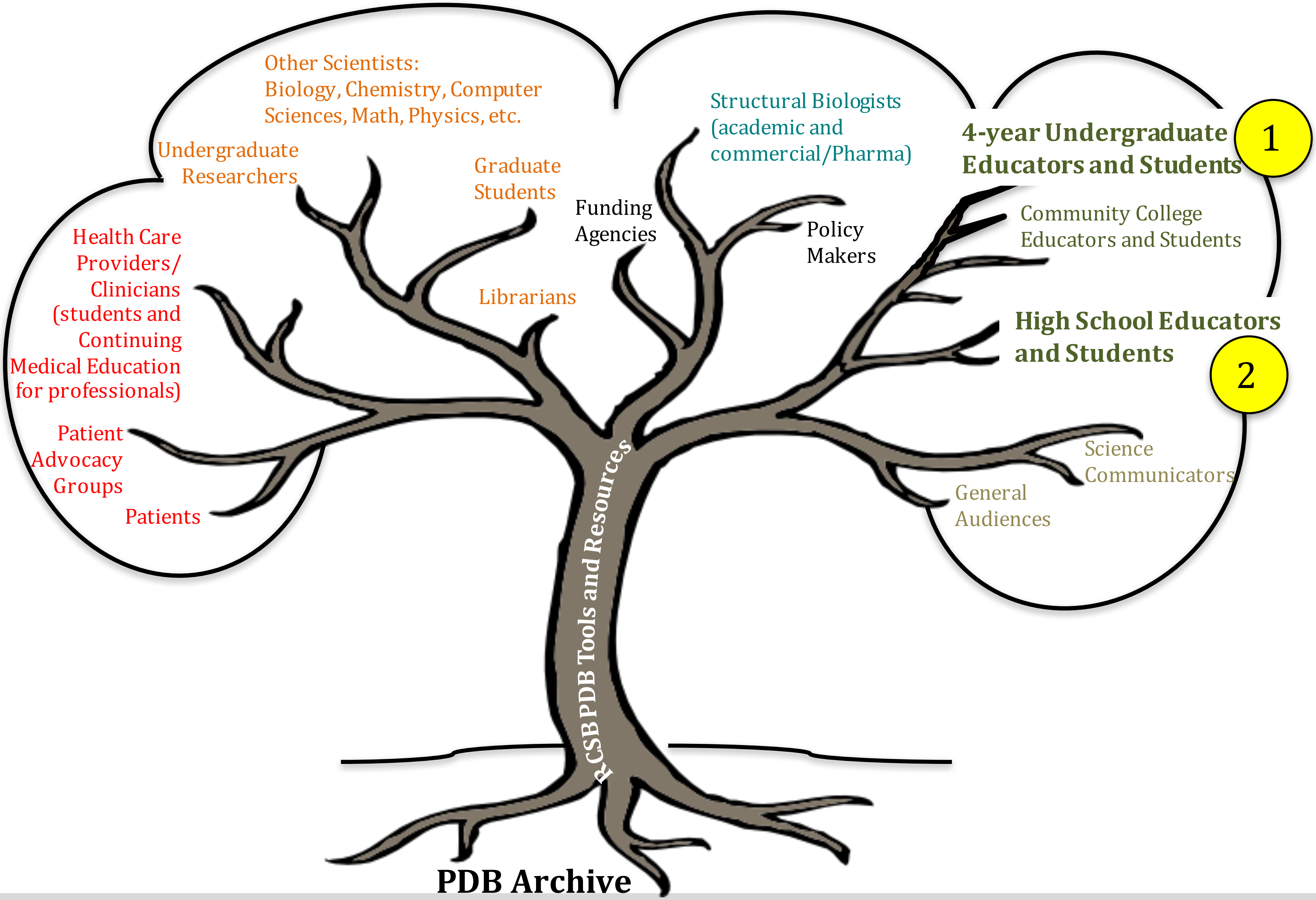
RCSB PDB
PROTEIN DATA BANK



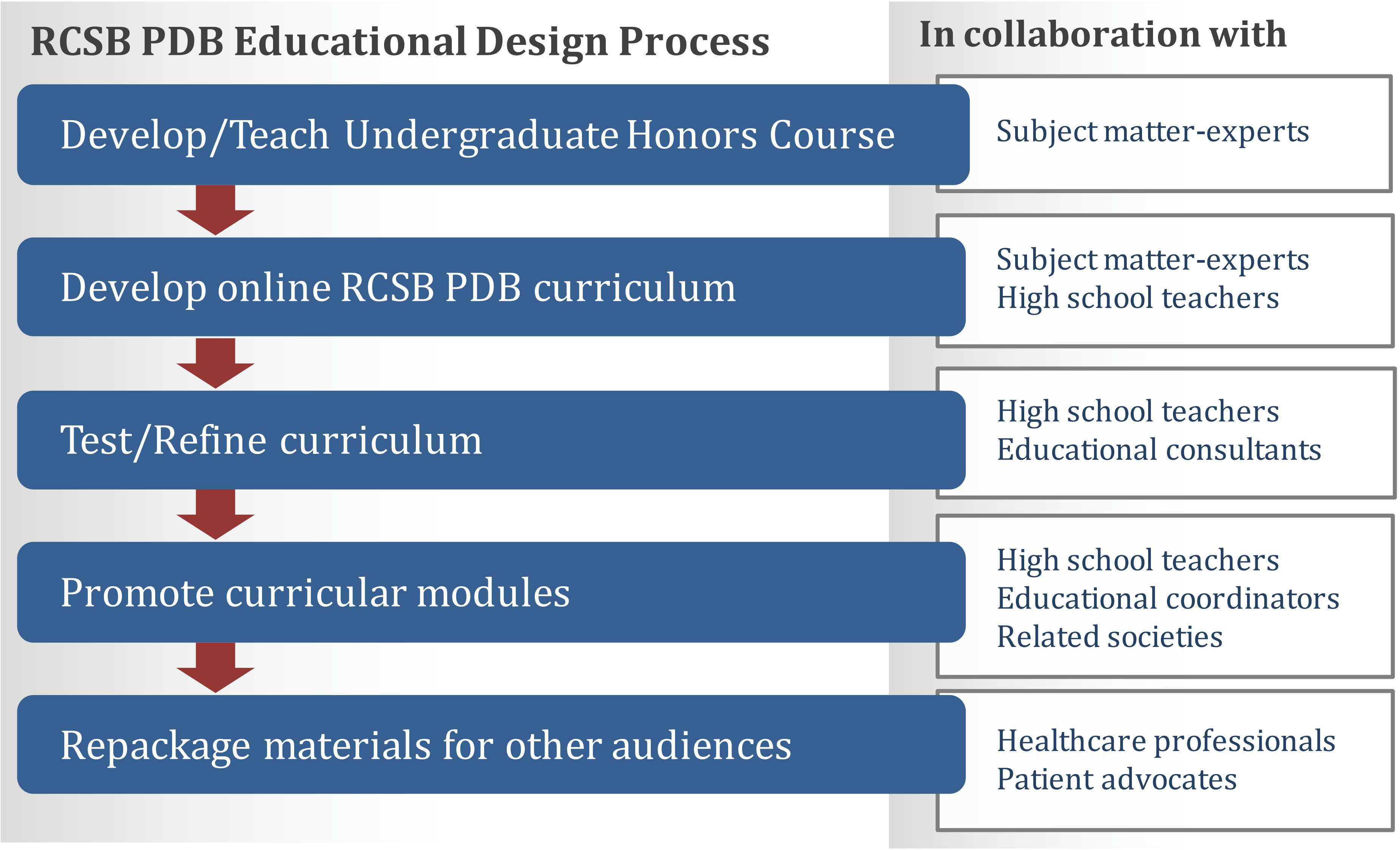
Education

Shuchismita Dutta, Ph.D.

RCSB PDB User Communities

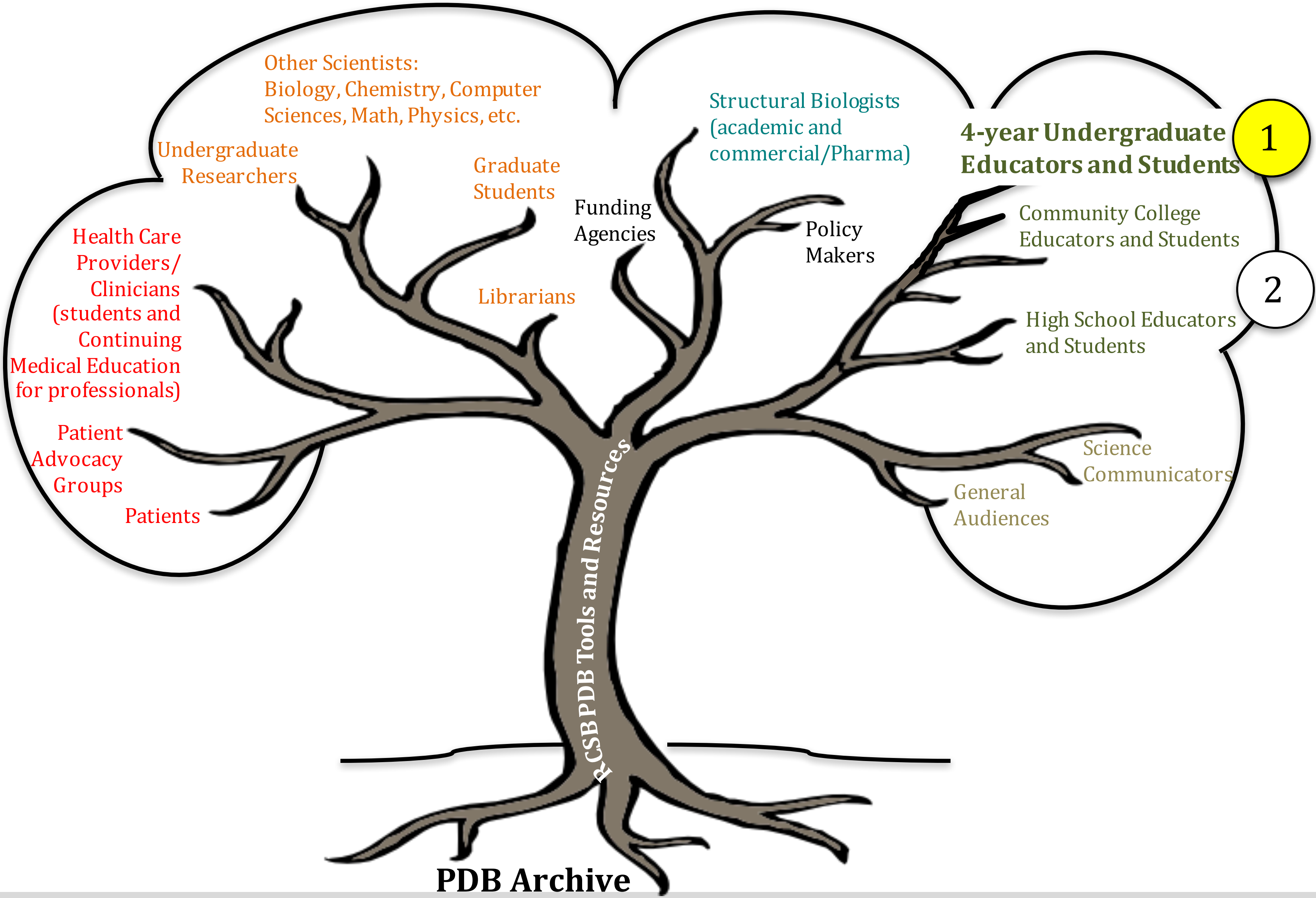

Other Scientists:
Biology, Chemistry, Computer Sciences, Math, Physics, etc.
Undergraduate Researchers
Graduate Students
Funding Agencies
Structural Biologists (academic and commercial/Pharma)
Policy Makers
4-year Undergraduate Educators and Students
1
Community College Educators and Students
High School Educators and Students
2
Librarians
Health Care Providers/ Clinicians (students and Continuing Medical Education for professionals)
Patient Advocacy Groups
Patients
Science Communicators
General Audiences
RCSB PDB Tools and Resources
PDB Archive

Theme-Based Education Strategy



RCSB PDB Educational Design Process	In collaboration with
Develop/Teach Undergraduate Honors Course	Subject matter-experts
Develop online RCSB PDB curriculum	Subject matter-experts High school teachers
Test/Refine curriculum	High school teachers Educational consultants
Promote curricular modules	High school teachers Educational coordinators Related societies
Repackage materials for other audiences	Healthcare professionals Patient advocates

Offering Courses and Developing Curricula



Other Scientists:
Biology, Chemistry, Computer
Sciences, Math, Physics, etc.

Undergraduate
Researchers

Graduate
Students

Funding
Agencies

Librarians

Structural Biologists
(academic and
commercial/Pharma)

Policy
Makers

**4-year Undergraduate
Educators and Students**

1

Community College
Educators and Students

2

High School Educators
and Students

Health Care
Providers/
Clinicians
(students and
Continuing
Medical Education
for professionals)

Patient
Advocacy
Groups

Patients

Science
Communicators

General
Audiences

RCSB PDB Tools and Resources

PDB Archive

Syllabus of UG Honors Course on Diabetes

Syllabus (Spring 2015-7)

- Introduction to Insulin and Diabetes
- Understanding the subject matter in 3D
- Clinical aspects of Diabetes and its treatments – Expert lectures
- Approaches to managing Diabetes
  - Non-pharmacological
  - Pharmacological

Student Assessment Projects

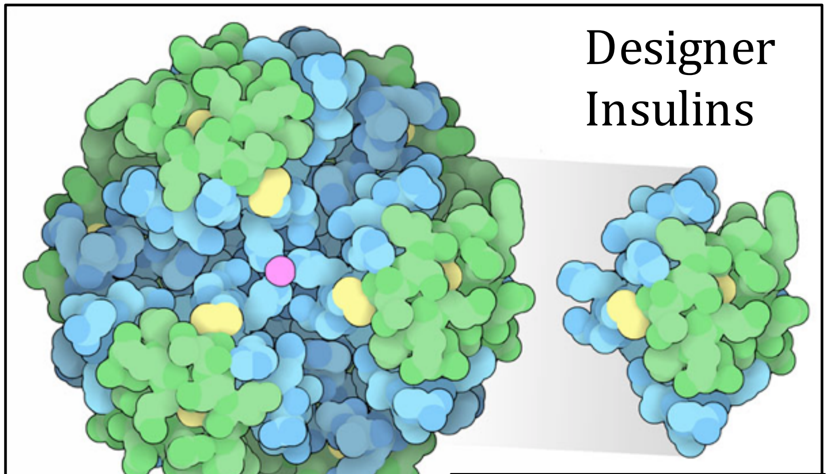
- Molecules involved in glucose homeostasis and causes of Diabetes
- Current pharmacological approaches for treatment of Type 2 Diabetes

Dr. L. Amorosa
RWJMS, Endocrinology

Cynthia Seidman
RD, CDN, Formerly at
Rockefeller University

Dr. T. Schneider
RWJMS, Endocrinology

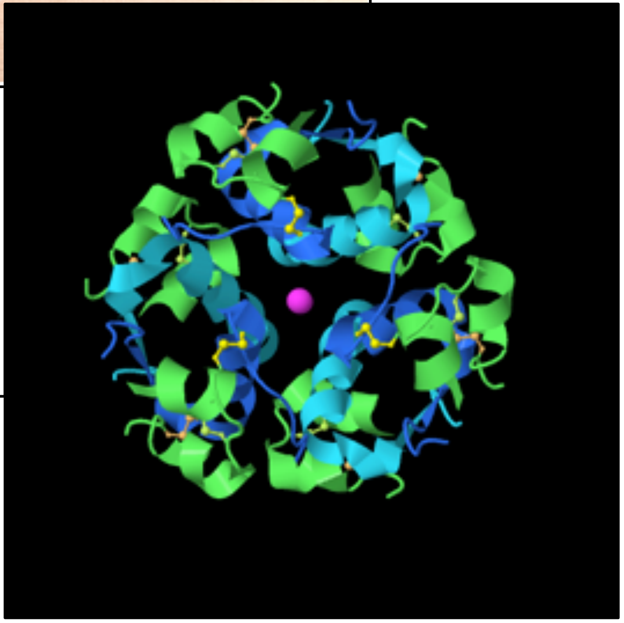
PDB-101 Resources/Activities



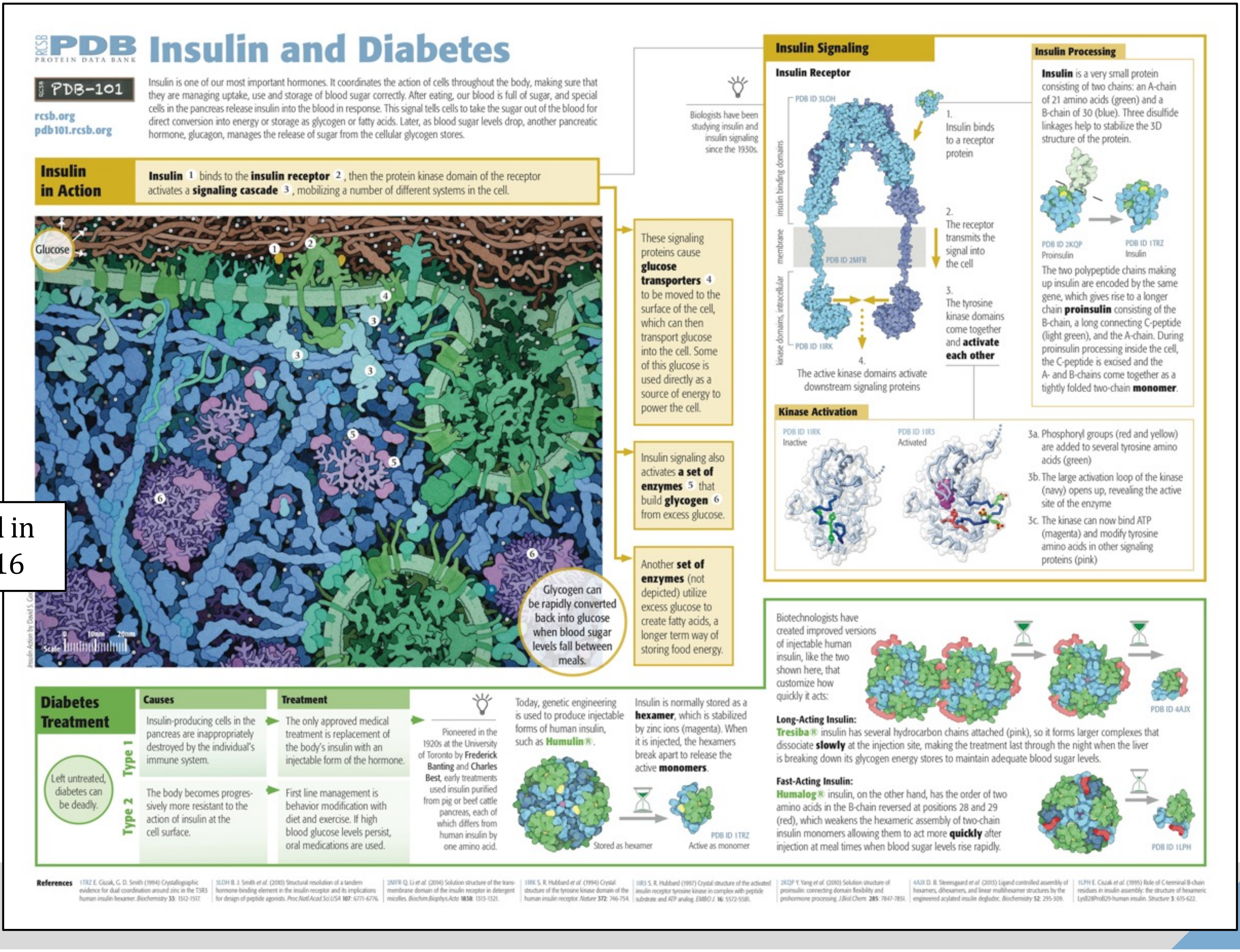
MotM, Spring 2016

Insulin	Feb. 2001
Insulin Receptor	Feb. 2015
Glucagon	Apr. 2015
Receptor for Advanced Glycation End Products	Jun. 2015
Designer Insulins	Feb. 2016
Dipeptidyl Peptidase 4	Oct. 2016

Molecular Origami: Insulin Paper Model and Visualization Activity
Fall 2015



Diabetes Poster



Insulin and Diabetes

rcsb.org
pdb101.rcsb.org

Insulin is one of our most important hormones. It coordinates the action of cells throughout the body, making sure that they are managing uptake, use and storage of blood sugar correctly. After eating, our blood is full of sugar, and special cells in the pancreas release insulin into the blood in response. This signal tells cells to take the sugar out of the blood for direct conversion into energy or storage as glycogen or fatty acids. Later, as blood sugar levels drop, another pancreatic hormone, glucagon, manages the release of sugar from the cellular glycogen stores.

Biologists have been studying insulin and insulin signaling since the 1930s.

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.

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.

Insulin Signaling

Insulin Receptor

1. Insulin binds to a receptor protein

2. The receptor transmits the signal into the cell

3. The tyrosine kinase domains come together and **activate each other**

4. The active kinase domains activate downstream signaling proteins

Insulin Processing

Insulin is a very small protein consisting of two chains: an A-chain of 21 amino acids (green) and a B-chain of 30 (blue). Three disulfide linkages help to stabilize the 3D structure of the protein.

The two polypeptide chains making up insulin are encoded by the same gene, which gives rise to a longer chain **proinsulin** consisting of the B-chain, a long connecting C-peptide (light green), and the A-chain. During proinsulin processing inside the cell, the C-peptide is excised and the A- and B-chains come together as a tightly folded two-chain **monomer**.

Kinase Activation

3a. Phosphoryl groups (red and yellow) are added to several tyrosine amino acids (green)

3b. The large activation loop of the kinase (navy) opens up, revealing the active site of the enzyme

3c. The kinase can now bind ATP (magenta) and modify tyrosine amino acids in other signaling proteins (pink)

Diabetes Treatment

Left untreated, diabetes can be deadly.

	Causes	Treatment
Type 1	Insulin-producing cells in the pancreas are inappropriately destroyed by the individual's immune system.	The only approved medical treatment is replacement of the body's insulin with an injectable form of the hormone.
Type 2	The body becomes progressively more resistant to the action of insulin at the cell surface.	First line management is behavior modification with diet and exercise. If high blood glucose levels persist, oral medications are used.

Pioneered in the 1920s at the University of Toronto by **Frederick Banting** and **Charles Best**, early treatments used insulin purified from pig or beef cattle pancreas, each of which differs from human insulin by one amino acid.

Today, genetic engineering is used to produce injectable forms of human insulin, such as **Humulin®**.

Insulin is normally stored as a **hexamer**, which is stabilized by zinc ions (magenta). When it is injected, the hexamers break apart to release the active **monomers**.

Biotechnologists have created improved versions of injectable human insulin, like the two shown here, that customize how quickly it acts:

Long-Acting Insulin:
Tresiba® insulin has several hydrocarbon chains attached (pink), so it forms larger complexes that dissociate **slowly** at the injection site, making the treatment last through the night when the liver is breaking down its glycogen energy stores to maintain adequate blood sugar levels.

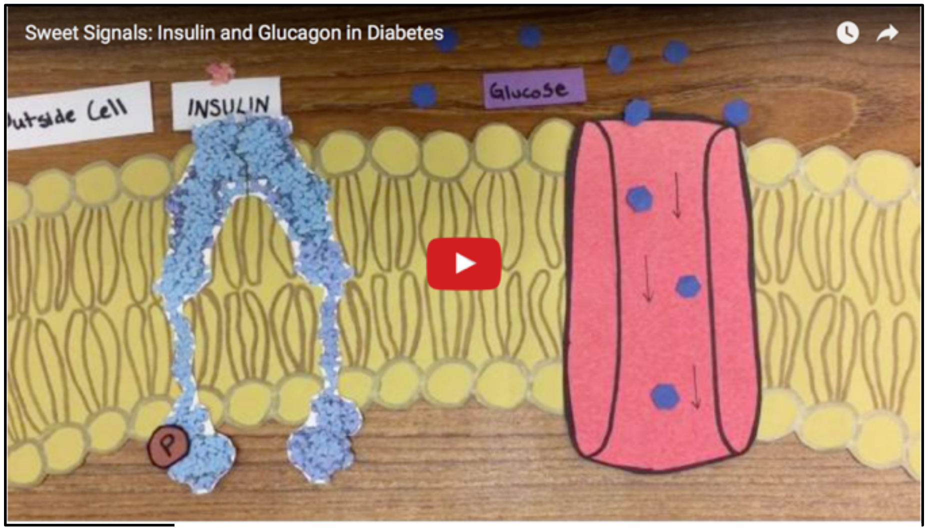
Fast-Acting Insulin:
Humalog® insulin, on the other hand, has the order of two amino acids in the B-chain reversed at positions 28 and 29 (red), which weakens the hexameric assembly of two-chain insulin monomers allowing them to act more **quickly** after injection at meal times when blood sugar levels rise rapidly.

References 1TRZ E. Ciszak, G. D. Smith (1994) Crystallographic evidence for dual coordination around zinc in the T3R3 human insulin hexamer. *Biochemistry* **33**: 1512-1517. | 3LOH B. J. Smith *et al.* (2010) Structural resolution of a tandem hormone-binding element in the insulin receptor and its implications for design of peptide agonists. *Proc.Natl.Acad.Sci.USA* **107**: 6771-6776. | 2MFR Q. Li *et al.* (2014) Solution structure of the trans-membrane domain of the insulin receptor in detergent micelles. *Biochim.Biophys.Acta* **1838**: 1313-1321. | 1IRK S. R. Hubbard *et al.* (1994) Crystal structure of the tyrosine kinase domain of the human insulin receptor. *Nature* **372**: 746-754. | 1IR3 S. R. Hubbard (1997) Crystal structure of the activated insulin receptor tyrosine kinase in complex with peptide substrate and ATP analog. *EMBO J.* **16**: 5572-5581. | 2KQP Y. Yang *et al.* (2010) Solution structure of proinsulin: connecting domain flexibility and prohormone processing. *J.Biol.Chem.* **285**: 7847-7851. | 4AJX D. B. Steensgaard *et al.* (2013) Ligand controlled assembly of hexamers, dihexamers, and linear multihexamer structures by the engineered acylated insulin degludec. *Biochemistry* **52**: 295-309. | 1LPH E. Ciszak *et al.* (1995) Role of C-terminal B-chain residues in insulin assembly: the structure of hexameric LysB28ProB29-human insulin. *Structure* **3**: 615-622.

Developed in Spring 2016

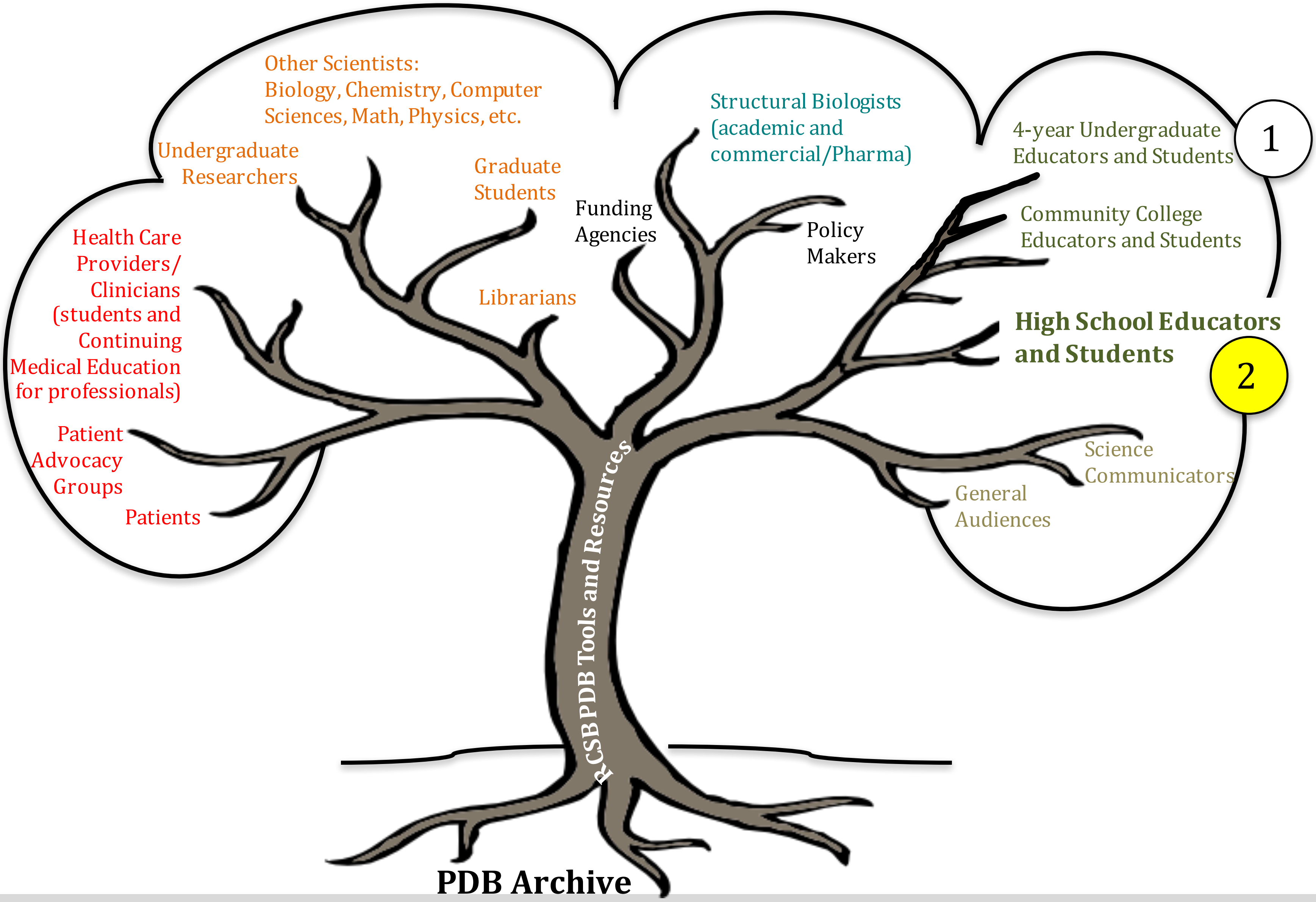
Video Challenge for HS Students

- 2016
  - Topic: Structural Biology & Diabetes
  - Participation: 82 entries (up from 38 in 2015)
  - Judges:
    - Endocrinology Chief
    - Science Animator
    - Drug Discovery Scientist
    - Scientist/Educator
- 2017
  - Topic: Treating Diabetes



2016 Judge's Awards First Place

Offering Courses and Developing Curricula


Other Scientists:
Biology, Chemistry, Computer Sciences, Math, Physics, etc.
Undergraduate Researchers
Graduate Students
Funding Agencies
Structural Biologists (academic and commercial/Pharma)
Policy Makers
4-year Undergraduate Educators and Students
1
Community College Educators and Students
High School Educators and Students
2
Librarians
Health Care Providers/ Clinicians (students and Continuing Medical Education for professionals)
Patient Advocacy Groups
Patients
Science Communicators
General Audiences
RCSB PDB Tools and Resources
PDB Archive

Develop/Test Diabetes HS Curriculum

- Development workshop
  (July 2016)
  - Draft Modular Curriculum
  - Meet Next Generation Science Standards (NGSS)
- Recruitment workshop
  (September 2016)
  - 45 NJ HS Teachers and Science Supervisors attended
  - 28 teachers from 17 schools committed to Pilot Testing
- NJ Science Convention
  (October 2016)
  - More recruitment of Pilot Testers
- Pilot Testing
  - 2016-17 Academic year

Dr. L. Amorosa
RWJMS, Endocrinology

Dr. A. Ohri
RWJMS, Endocrinology

Dr. M. Kamienski
Rutgers School of Nursing

K. Shah

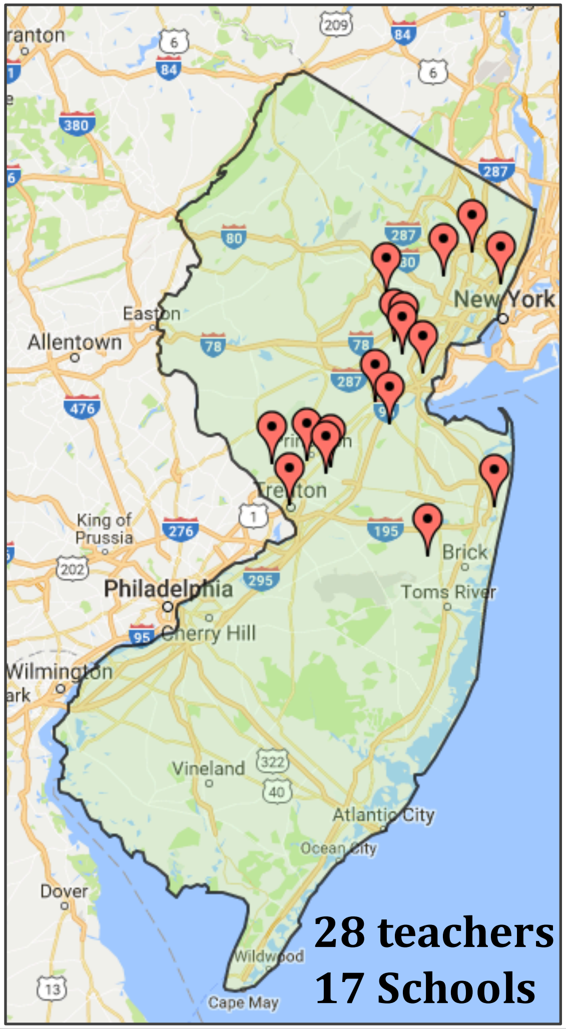
Standing (L to R): Mr. R. Tempsick, Mr. B. Buck, Ms. S. Coletta, Mrs. A. Sanelli, Ms. J. Jiang, Mrs. H. Sharif, Mrs. S. Eswaran
Sitting (L to R): Dr. B. Ameer, Ms. M. Dominguez, Dr. S. Dutta, Dr. M. Battacharya

Pilot Testing Diabetes HS Curriculum

Curriculum at a Glance

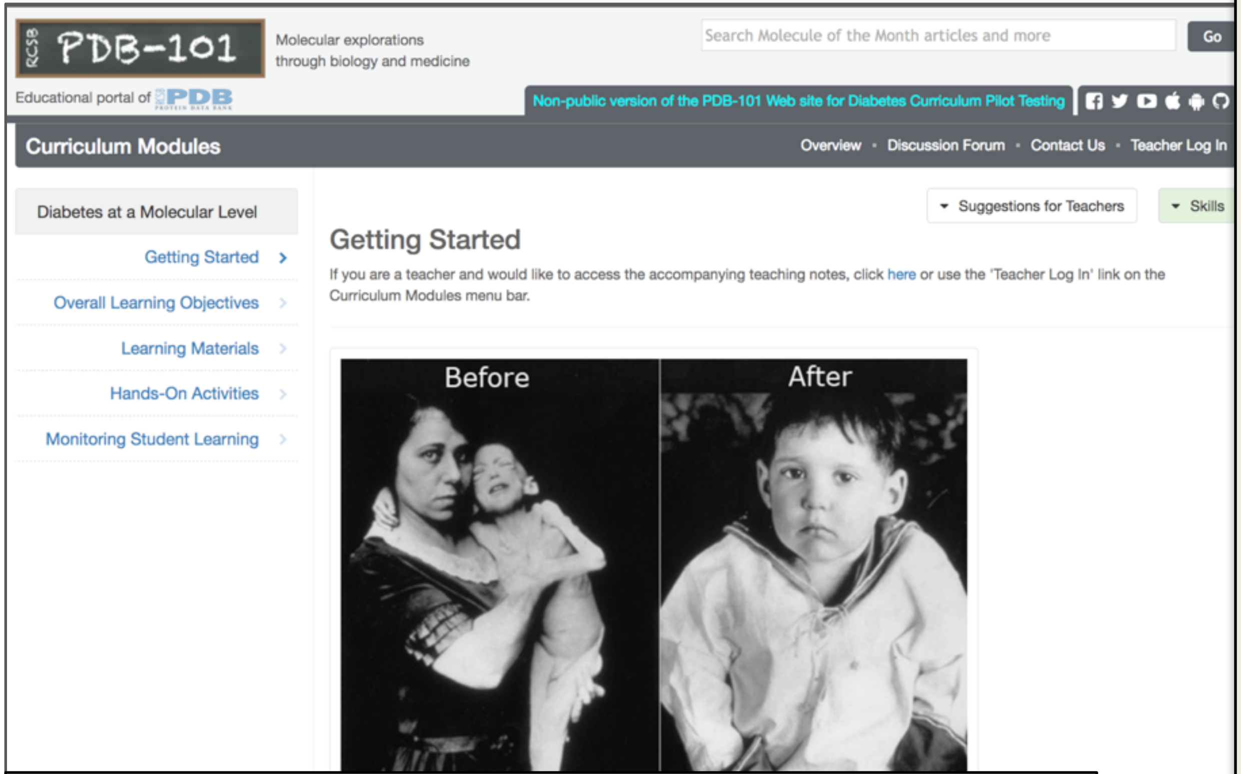
- Modular Components
  - Introduction to Proteins
  - Learning to use RCSB PDB data, tools, and resources
  - Enzymes
  - Protein Synthesis
  - Endocrine System
  - Cell Signaling
  - Genetics
  - Evolution
  - Managing Diabetes
- Module content
  - Pre- and Post-Tests
  - Learning Materials with Notes
  - Activities with Teacher Notes

Where are our Testers?



28 teachers
17 Schools

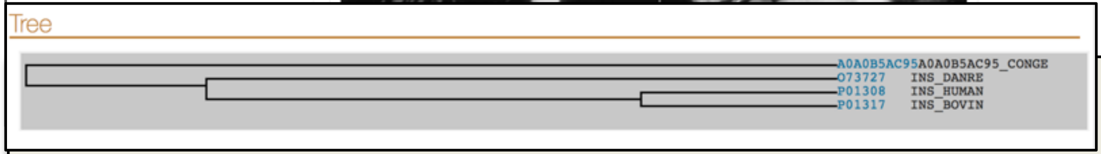
Selected HS Activities



RCSB PDB-101
Molecular explorations through biology and medicine
Search Molecule of the Month articles and more
Go
Educational portal of PDB PROTEIN DATA BANK
Non-public version of the PDB-101 Web site for Diabetes Curriculum Pilot Testing
Curriculum Modules
Overview · Discussion Forum · Contact Us · Teacher Log In
Diabetes at a Molecular Level
Getting Started
Overall Learning Objectives
Learning Materials
Hands-On Activities
Monitoring Student Learning
Suggestions for Teachers
Skills
Before
After

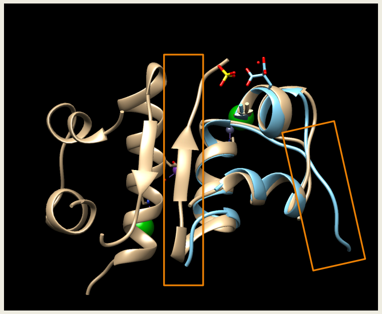
Getting Started

If you are a teacher and would like to access the accompanying teaching notes, click here or use the 'Teacher Log In' link on the Curriculum Modules menu bar.



Tree

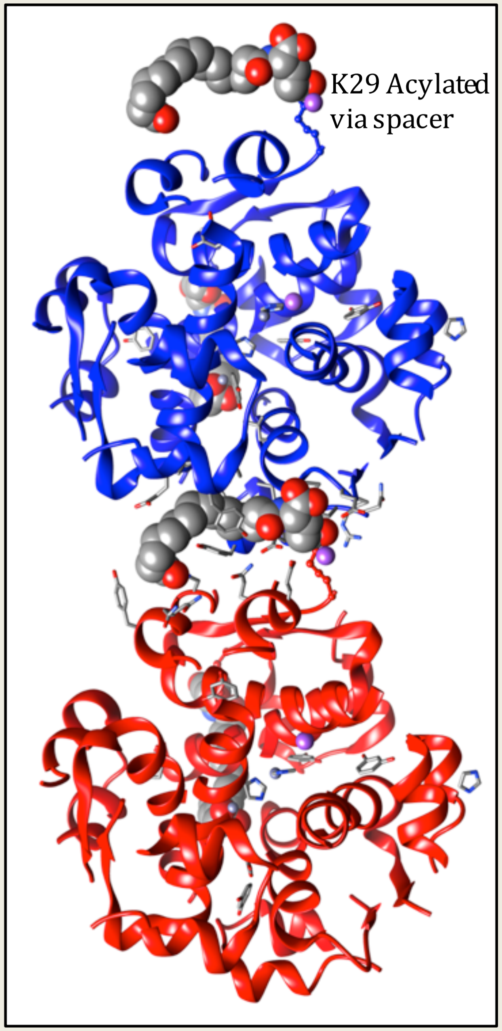
```
A0A0B5AC95A0A0B5AC95_CONGE
O73727    INS_DANRE
P01308    INS_HUMAN
P01317    INS_BOVIN
```



Weaponized Insulin

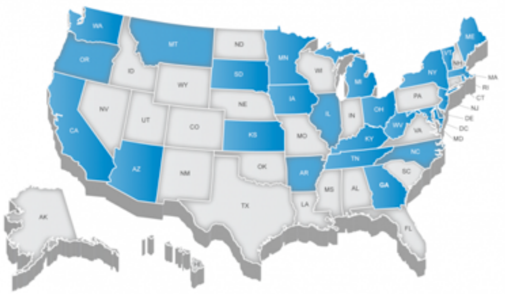
PDB 5JYQ
PDB 1TRZ

Engineered Insulin Molecules



Degludec (Acylated Insulin), PDB 4AJX

Evaluation and Expansion



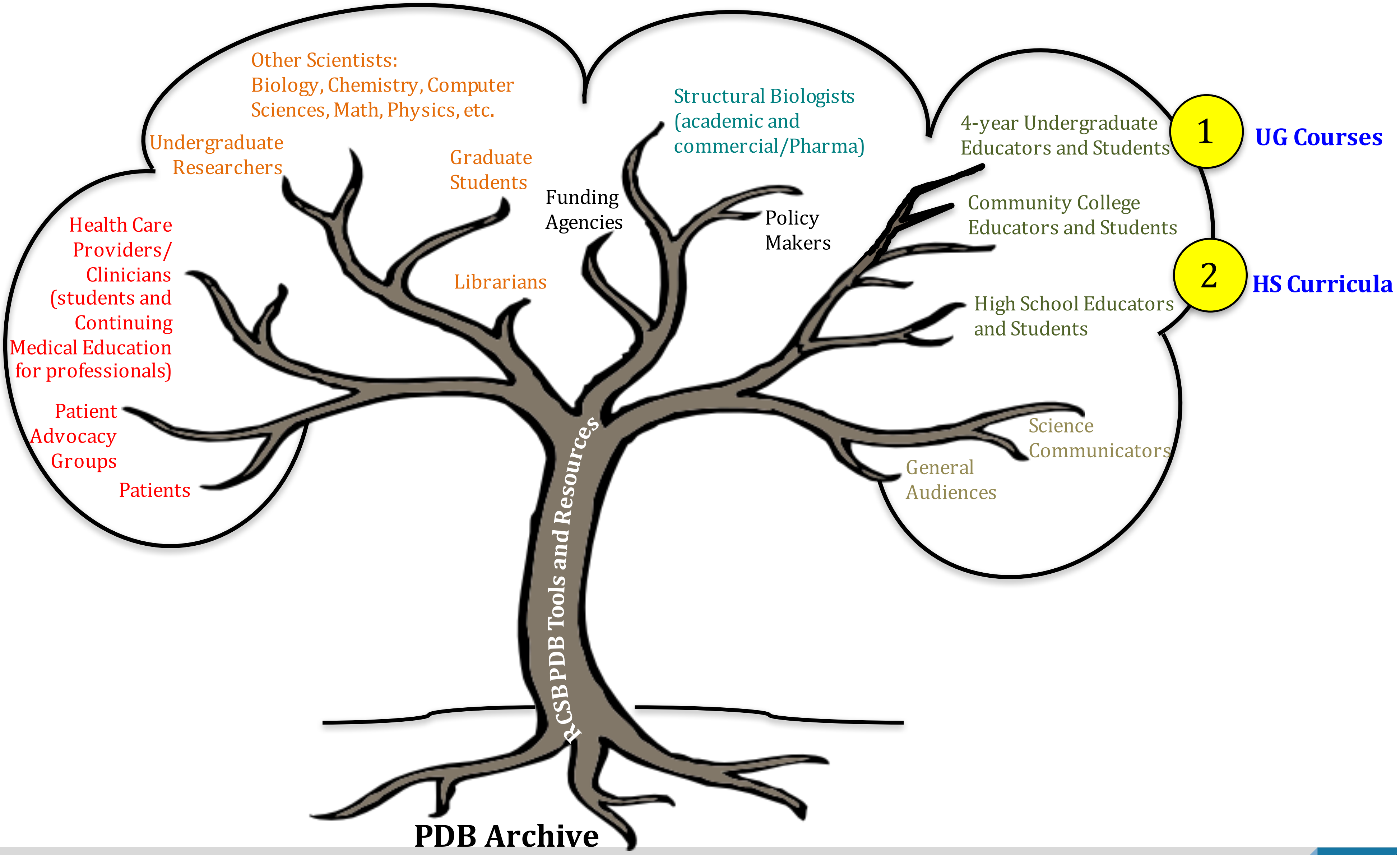
Expert Advisors/Plans

- Karen Collias
  - Founder
- Jennifer Childress
  - Director Instructional Support
- Sue Coletta
  - WSSP/Colleague
- Planned evaluation
  - EQuIP applied to NGSS
    - Alignment to the NGSS
    - Instructional Supports
    - Monitoring Student Progress

Evaluation Process

- Gather Data: Pilot Testing of Diabetes curriculum
  - Pre- and Post-tests
    - – Student Progress
  - Teacher Surveys
    - – Alignment to NGSS
    - – Instructional Support
  - Student Artifacts
    - – Student Progress
- Contract for Professional Evaluation of Curriculum (using above data)

Summary: Education Efforts



Other Scientists:
Biology, Chemistry, Computer Sciences, Math, Physics, etc.

Undergraduate Researchers

Graduate Students

Funding Agencies

Structural Biologists (academic and commercial/Pharma)

Policy Makers

Health Care Providers/ Clinicians (students and Continuing Medical Education for professionals)

Librarians

Patient Advocacy Groups

Patients

4-year Undergraduate Educators and Students

1 UG Courses

Community College Educators and Students

2 HS Curricula

High School Educators and Students

Science Communicators

General Audiences

RCSB PDB Tools and Resources

PDB Archive

RCSB PDB
PROTEIN DATA BANK

Funding and Sustainability

Response to 2015 Advisory Committee Report

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

Current Funding

- Core mission support DBI-1338415 for 2014-2018
  (NSF, NIH, DOE; competing renewal likely in 2018)

- Non-core activity support
  - NIGMS Drug Design Data Resource (Amaro/Burley, UCSD)
  - NCI BD2K-Structural Biology Data Compression (Rose, UCSD)
  - NIH BD2K-BioCaddie (Rose, UCSD)
  - NLM BD2K Data Science Course (Lawson, Rutgers)
  - NSF-Integrative/Hybrid Methods EAGER (Berman, Rutgers)
  - NSF-Data Management EAGER (Berman, Rutgers)
  - NSF Big Data Spoke Planning Grant (Prlić, UCSD)
  - NSF REU Minority Summer Students (Burley, Rutgers)
  - NIDA Science Olympiad (Herman/Dutta, Milwaukee School of Engineering)

- RCSB collaborative projects:
  EMDataBank, BioSync, NDB, SBKB

- Private support for Outreach projects
  - HIV film (Viiv, IBM, Rutgers, *et al.*)
  - Symposium on Aesthetics and the Life Sciences
    (Wellcome Trust, Princeton, Rutgers)

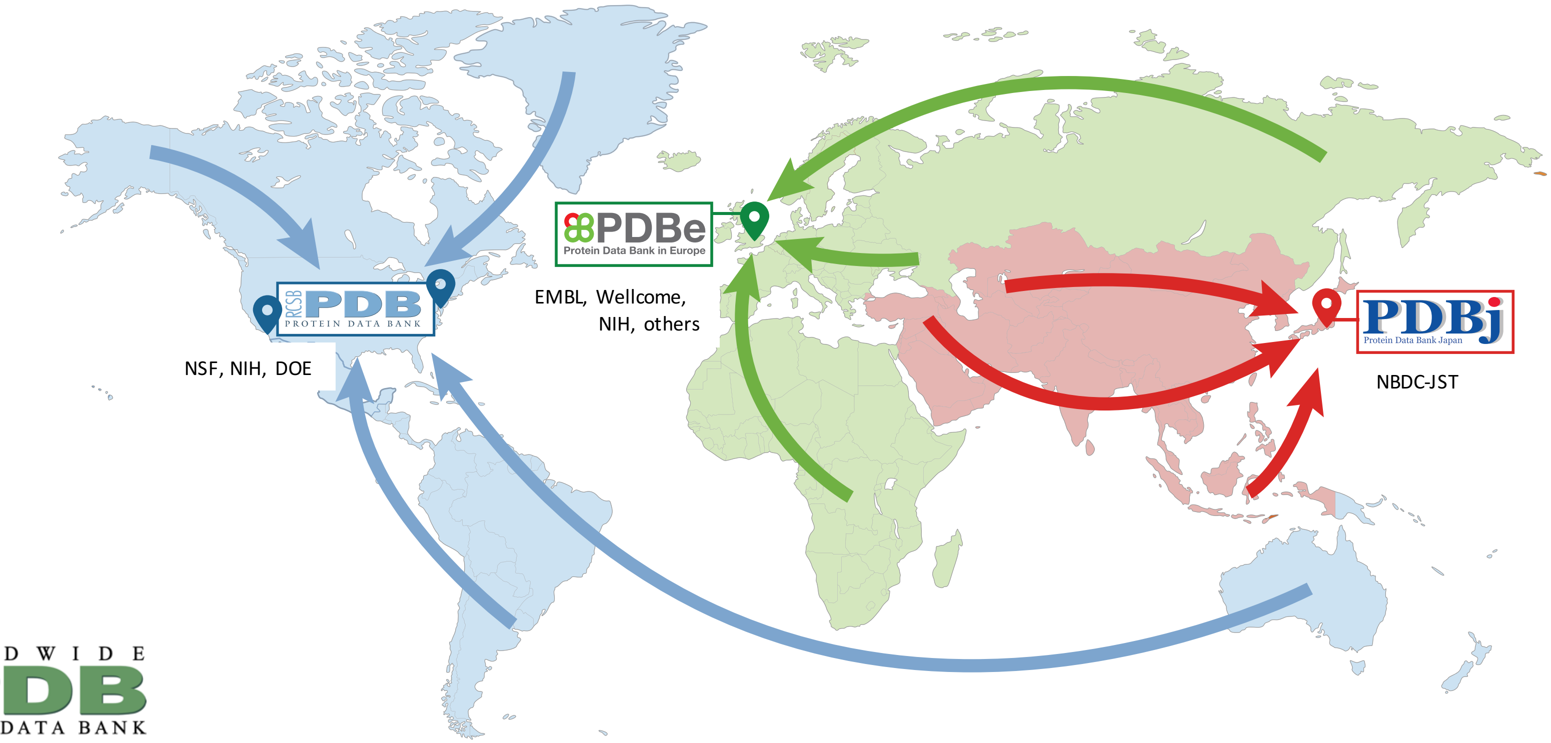
RCSB PDB Funding

- >10 years of nearly flat funding has resulted in a substantial decline in our purchasing power
  - 2004 funding was $5,926,617
    - Equivalent to ~$7,574,649 in 2016 (inflation)
  - 2016 funding is $6,455,369
    - Purchasing power down by ~$1,119K (⬇~14.8%)
- To add “Insult to Injury”
  - 2013 funding was $6,688,486
  - 2016 funding is $6,455,369 (⬇3.5%)

Data In Sustainability

Worldwide Protein Data Bank (wwPDB)

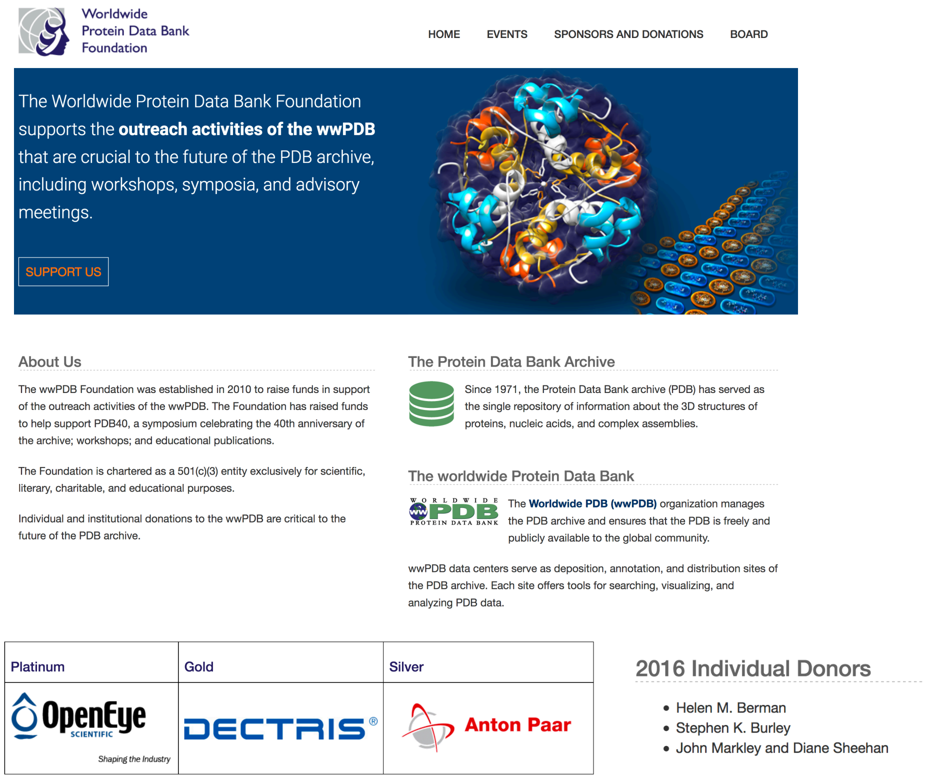
- Data In shared among Regional Data Centers: RCSB PDB (US), PDBj (Asia), PDBe (EU); BioMagResBank (US/Japan)


RCSB PDB PROTEIN DATA BANK
NSF, NIH, DOE
PDBe Protein Data Bank in Europe
EMBL, Wellcome, NIH, others
PDBj Protein Data Bank Japan
NBDC-JST

WORLDWIDE PDB PROTEIN DATA BANK

wwPDB Foundation: Outreach Fundraising

- Established to support specific wwPDB activities
  - Outreach and education activities, including seminars and workshops
  - Partial support for Advisory Committee meetings
  - Inaugural event: PDB 40th anniversary in 2011
  - Next major milestone: PDB50
- 501(c)3 organization
  - American, tax-exempt association dedicated to scientific, literary, charitable, and educational purposes
- Fundraising on-going



Worldwide
Protein Data Bank
Foundation

The Worldwide Protein Data Bank Foundation supports the **outreach activities of the wwPDB** that are crucial to the future of the PDB archive, including workshops, symposia, and advisory meetings.

About Us

The wwPDB Foundation was established in 2010 to raise funds in support of the outreach activities of the wwPDB. The Foundation has raised funds to help support PDB40, a symposium celebrating the 40th anniversary of the archive; workshops; and educational publications.

The Foundation is chartered as a 501(c)(3) entity exclusively for scientific, literary, charitable, and educational purposes.

Individual and institutional donations to the wwPDB are critical to the future of the PDB archive.

The Protein Data Bank Archive

Since 1971, the Protein Data Bank archive (PDB) has served as the single repository of information about the 3D structures of proteins, nucleic acids, and complex assemblies.

The worldwide Protein Data Bank

The **Worldwide PDB (wwPDB)** organization manages the PDB archive and ensures that the PDB is freely and publicly available to the global community.

wwPDB data centers serve as deposition, annotation, and distribution sites of the PDB archive. Each site offers tools for searching, visualizing, and analyzing PDB data.

Platinum	Gold	Silver
OpenEye SCIENTIFIC *Shaping the Industry*	DECTRIS®	Anton Paar

2016 Individual Donors

- Helen M. Berman
- Stephen K. Burley
- John Markley and Diane Sheehan

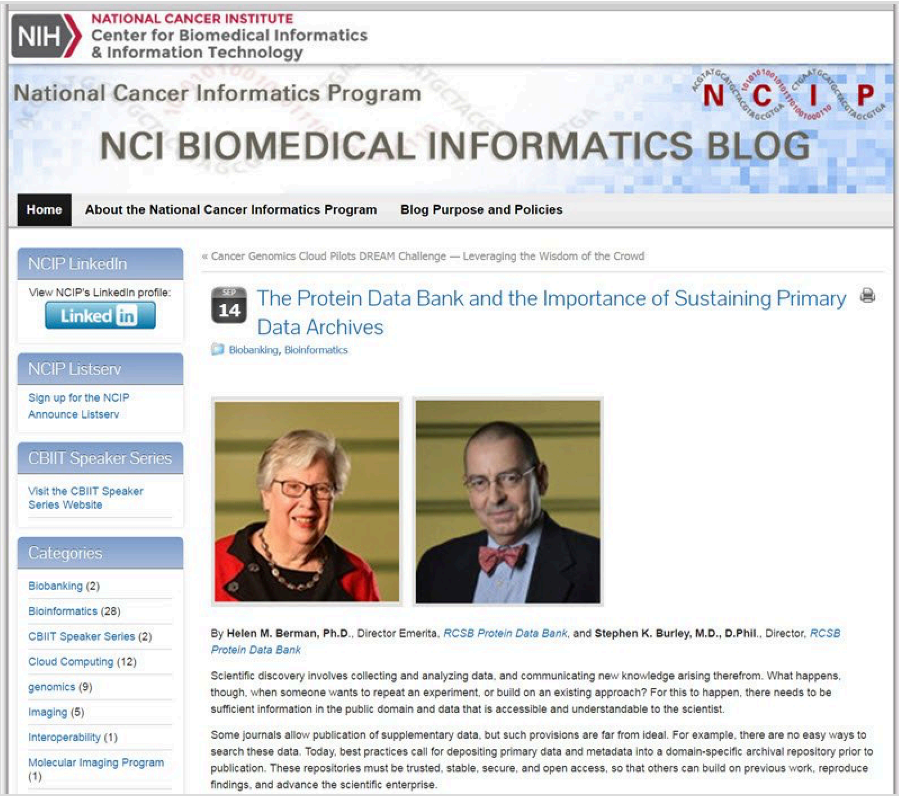
Members of the PDB, past and present, in attendance at PDB40

Data Out Sustainability

- RCSB.org one of most heavily-used primary biology data resources worldwide
  - Users: >1 million unique visitors/year
  - Global Reach: ~30% US, ~70% non-US
  - Most Data Consumers are not Data Producers
- Core activities enhanced *via* peer-reviewed grant applications for discrete technology development
- Joint wwPDB proposals planned for developing features common to both Data In and Data Out

Contributions to Sustainability Dialogue

- Sustaining Domain Repositories for Digital Data Working Group (Helen M. Berman)
- Sustaining Biological Infrastructure Advisory Board (Helen M. Berman)
- CODATA/SciDataCon (R. Andrew Byrd, Economics and Impact of the Protein Data Bank (PDB) Archive)
- International Human Frontier Science Program Organization (HFSPO) Life sciences data resources and the future
- NSF Advisory Committee for Cyberinfrastructure (Helen M. Berman)
- Gateways 2016: 11th Gateway Computing Environments Conference


NIH
NATIONAL CANCER INSTITUTE
Center for Biomedical Informatics
& Information Technology
National Cancer Informatics Program
N C I P
NCI BIOMEDICAL INFORMATICS BLOG
Home
About the National Cancer Informatics Program
Blog Purpose and Policies
NCIP LinkedIn
View NCIP's LinkedIn profile:
Linked in
NCIP Listserv
Sign up for the NCIP Announce Listserv
CBIIT Speaker Series
Visit the CBIIT Speaker Series Website
Categories
Biobanking (2)
Bioinformatics (28)
CBIIT Speaker Series (2)
Cloud Computing (12)
genomics (9)
Imaging (5)
Interoperability (1)
Molecular Imaging Program (1)
« Cancer Genomics Cloud Pilots DREAM Challenge — Leveraging the Wisdom of the Crowd
SEP
14

The Protein Data Bank and the Importance of Sustaining Primary Data Archives

Biobanking, Bioinformatics

By **Helen M. Berman, Ph.D.**, Director Emerita, *RCSB Protein Data Bank*, and **Stephen K. Burley, M.D., D.Phil.**, Director, *RCSB Protein Data Bank*

Scientific discovery involves collecting and analyzing data, and communicating new knowledge arising therefrom. What happens, though, when someone wants to repeat an experiment, or build on an existing approach? For this to happen, there needs to be sufficient information in the public domain and data that is accessible and understandable to the scientist.

Some journals allow publication of supplementary data, but such provisions are far from ideal. For example, there are no easy ways to search these data. Today, best practices call for depositing primary data and metadata into a domain-specific archival repository prior to publication. These repositories must be trusted, stable, secure, and open access, so that others can build on previous work, reproduce findings, and advance the scientific enterprise.

Longstanding participation in formal and informal sustainability discussions

Sustaining Domain Repositories for Digital Data Working Group *Principles*

1. Research data are a Public Good
2. Science requires a durable and permanent record
3. Repositories provide essential domain expertise
4. Data should be prepared for curation prior to publication (not after the fact or never!)
5. Sufficient and long-term financial support is critical
6. Global partnerships, both public and private, should be encouraged
7. Fiscal transparency is essential

Sustaining Domain Repositories for Digital Data: A White Paper (2013)
http://datacommunity.icpsr.umich.edu/sites/default/files/WhitePaper_ICPSR_SDRDD_121113.pdf

Sustaining Domain Repositories for Digital Data Working Group *Funding Requirements*

- Economic Stability/Long-term Sustainability
- Global Open Access
- Equity for Data Depositors
- Equity for Research/Teaching Institutions

8 different funding models examined
Only one meets all requirements

The Infrastructure Funding Model

- Funding agencies commit to direct payment of the costs of archiving experimental data/metadata generated with the research support they provide
- Data Resource funding comes in the form of strategic, long-term infrastructure investments (divorced from typical 3-5 year grant cycles)
- Ensures Economic Stability/Sustainability for an Open Access Data Resource Ecosystem with Equity for Data Depositors and Consumers

- PDB replacement cost: $US 12 billion
- Estimated archiving cost/year: ~2% of structural biology funding

Questions for the Committee

- Sustainability
  - Are there other funding sources we should be exploring?
  - Are we making convincing arguments?
  - Are there other data you want to see?
  - Are we reaching the right people?