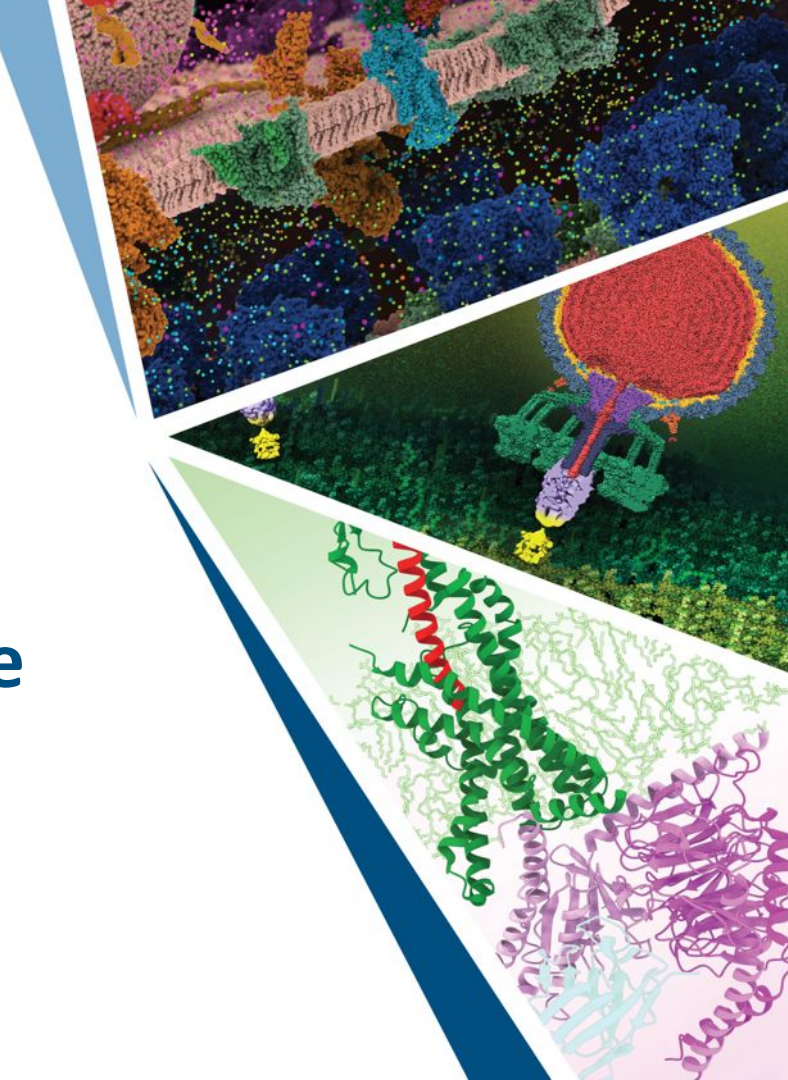
RCSB PDB
PROTEIN DATA BANK

RCSB.org
info@rcsb.org



RCSB PDB Advisory Committee

Virtual Meeting

Tuesday, March 15, 2022

Background Information Slides

Slides with a *green title* are provided as background information and can be presented and discussed at the meeting by request.

Main slides appear with blue titles

Underlined text indicates an active link

Introductions to Participants

Advisory Committee (full Roster)

- Returning: Paul Adams (Chair), Peter Andolfatto, Bridget Carragher, Wah Chiu*, Kirk Clark, Robert B. Darnell, Roland Dunbrack, Paul Falkowski, Thomas Ferrin, Mandë Holford, Cathy Peishoff, Sue Rhee, Torsten Schwede, Lance Stewart
- Welcome: Kevin H. Gardner, Takita F. Sumter
- Retiring: Judy Blake**, Paul Craig (observer; now on sabbatical with RCSB PDB), Jill Trewhella

RCSB PDB Participants

- Director: Stephen K. Burley; Director Emerita: Helen M. Berman; UCSF Site Head: Andrej Sali
- Operations Team Members
  - Rutgers: Zukang Feng, Robert Lowe, Jasmine Young, Christine Zardecki
  - UCSD: Jose Duarte (Site Manager), Yana Rose

Federal Funder Representatives

- Steven Ellis (NSF), Paula Flicker (NIGMS), Jerry Li (NCI), Amy Swain (DOE), Ramana Madupu (DOE)

*joining late, **unable to participate

Overview and 2021 Highlights

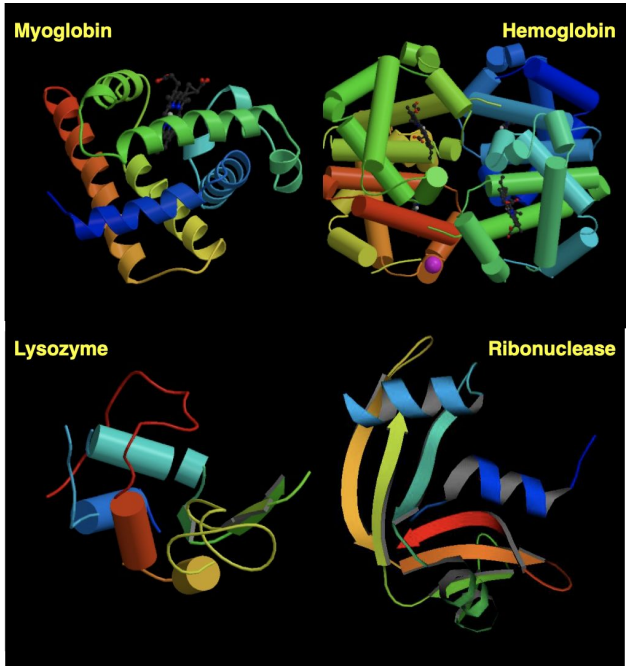
Stephen K. Burley

2021 Journals with RCSB PDB publications

Protein Data Bank History

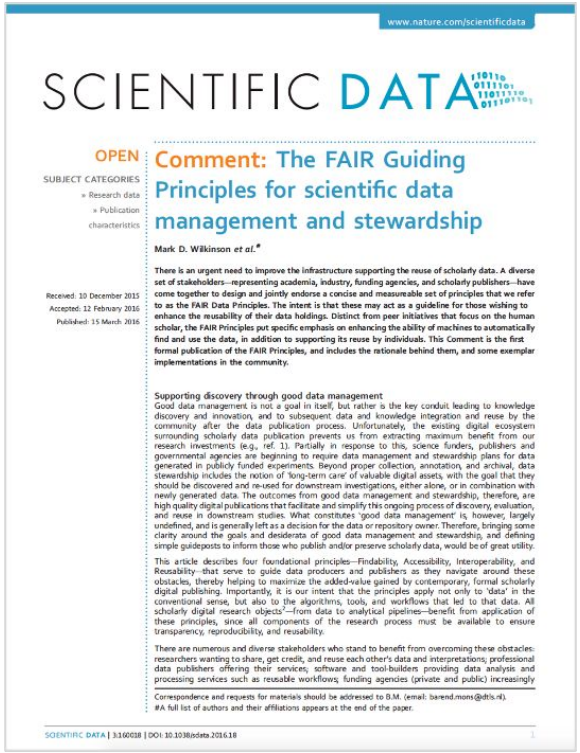
- PDB 1st Open Access digital data resource in biology
- Founded in 1971 with 7 X-ray structures of proteins
- Single global archive for protein and DNA/RNA experimental structures
- **Today, Open Access to >188,000 structures**

- wwPDB collaborates on management of the archive
  - Atomic Coordinates: US (RCSB PDB), EU (PDBe), and Japan (PDBj)
  - Related Experimental Data: PDB (MX), BMRB (NMR), and EMDB (3DEM maps)
- Honored the *FAIR*, *FACT*, and *TRUST* Principles since inception



Structures that Inspired Launch of the PDB

FAIR Principles



www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES
» Research data
» Publication characteristics

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson *et al.*#

Received: 10 December 2015
Accepted: 12 February 2016
Published: 15 March 2016

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

Supporting discovery through good data management

Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process. Unfortunately, the existing digital ecosystem surrounding scholarly data publication prevents us from extracting maximum benefit from our research investments (e.g., ref. 1). Partially in response to this, science funders, publishers and governmental agencies are beginning to require data management and stewardship plans for data generated in publicly funded experiments. Beyond proper collection, annotation, and archival, data stewardship includes the notion of 'long-term care' of valuable digital assets, with the goal that they should be discovered and re-used for downstream investigations, either alone, or in combination with newly generated data. The outcomes from good data management and stewardship, therefore, are high quality digital publications that facilitate and simplify this ongoing process of discovery, evaluation, and reuse in downstream studies. What constitutes 'good data management' is, however, largely undefined, and is generally left as a decision for the data or repository owner. Therefore, bringing some clarity around the goals and desiderata of good data management and stewardship, and defining simple guideposts to inform those who publish and/or preserve scholarly data, would be of great utility.

This article describes four foundational principles—Findability, Accessibility, Interoperability, and Reusability—that serve to guide data producers and publishers as they navigate around these obstacles, thereby helping to maximize the added-value gained by contemporary, formal scholarly digital publishing. Importantly, it is our intent that the principles apply not only to 'data' in the conventional sense, but also to the algorithms, tools, and workflows that led to that data. All scholarly digital research objects[2]—from data to analytical pipelines—benefit from application of these principles, since all components of the research process must be available to ensure transparency, reproducibility, and reusability.

There are numerous and diverse stakeholders who stand to benefit from overcoming these obstacles: researchers wanting to share, get credit, and reuse each other's data and interpretations; professional data publishers offering their services; software and tool-builders providing data analysis and processing services such as reusable workflows; funding agencies (private and public) increasingly

Correspondence and requests for materials should be addressed to B.M. (email: barend.mons@dtls.nl).
#A full list of authors and their affiliations appears at the end of the paper.

Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship.
Sci Data (2016) 3: 160018 doi: 10.1038/sdata.2016.18

FACT Principles

Bus Inf Syst Eng 59(5):311–313 (2017)
DOI 10.1007/s12599-017-0487-z

EDITORIAL

Responsible Data Science

Wil M. P. van der Aalst · Martin Bichler · Armin Heinzl

Published online: 26 June 2017

1 Introduction

An increasing fraction of research reported in BISE (Business & Information Systems Engineering) is data-driven. This is not surprising since torrents of data are vigorously changing the way we do business, socialize, conduct research, and govern society (Hilbert and Lopez 2011; Manyika et al. 2011; White House 2016). Data are collected on everything, at every time, and in every place. The Internet of Things (IoT) is rapidly expanding, with our homes, cars, and cities becoming "smart" by using the collected data in novel ways. These developments are also changing the way scientific research is performed. Model-driven approaches are supplemented with data-driven approaches. For example, genomics and evidence-based medicine are revolutionizing the understanding and treatment of diseases. From an epistemological point of view, data-driven approaches follow the logic of the new experimentalism (Mayo 1996; Chalmers 2013) in which knowledge is derived from experimental observations, not theory. Information systems which exploit the combination of data availability and powerful data science techniques dramatically improve our lives by enabling new services and products, while improving their efficiency and quality. *However, there are also great concerns about the use of data* (van der Aalst 2016a, b). Increasingly, customers, patients, and other stakeholders are concerned about irresponsible data use. Automated data decisions may be unfair or non-transparent. Confidential data may be shared unintentionally or abused by third parties. Each step in the "data science pipeline" (from raw data to insights and knowledge) may create inaccuracies, e.g., if the data used to learn a model reflects existing social biases, the algorithm is likely to incorporate these biases. These concerns could lead to resistance against the large-scale use of data and make it impossible to reap the benefits of data science. Rather than to avoid the use of data altogether, we strongly believe that data science techniques, infrastructures and approaches need be made responsible by design. Over the last year the first author has been leading a Dutch initiative called *Responsible Data Science* (RDS), cf. http://www.responsibledatascience.org/. In the context of RDS, there are research projects and regular meetings to discuss new ways to make data science more responsible. We believe that the insights obtained from these discussions are also relevant for the BISE community. The data-driven nature of today's (business) information systems makes it essential to incorporate safeguards against irresponsible data use already in the requirements and design phases.

Prof. dr. ir. W. M. P. van der Aalst (✉)
Department of Mathematics and Computer Science (MF 7.103),
Eindhoven University of Technology, PO Box 513,
5600 MB Eindhoven, The Netherlands
e-mail: w.m.p.v.d.aalst@tue.nl

Prof. Dr. M. Bichler
Department of Informatics, Decision Sciences & Systems,
Technical University of Munich (TUM), Boltzmannstr 3,
85748 Munich, Germany
e-mail: bichler@in.tum.de

Prof. Dr. A. Heinzl
Chair of General Management and Information Systems,
University of Mannheim, 68161 Mannheim, Germany
e-mail: heinzl@uni-mannheim.de

2 FACT: Fairness, Accuracy, Confidentiality, and Transparency

Responsible data science centers around four challenging questions (van der Aalst 2016a; Responsible Data Science Initiative 2016):

FACT: Fairness, Accuracy, Confidentiality, and Transparency

Responsible data science centers around four challenging questions (van der Aalst 2016a; Responsible Data Science Initiative 2016):

- Q1 fairness: data science without prejudice - how to avoid unfair conclusions even if they are true?
- Q2 accuracy: data science without guesswork - how to answer questions with a guaranteed level of accuracy?
- Q3 confidentiality: data science that ensures confidentiality - how to answer questions without revealing secrets?
- Q4 transparency: data science that provides transparency - how to clarify answers so that they become indisputable?

The terms fairness, accuracy, confidentiality, and transparency form the acronym FACT. This should not be confused with the well-known FAIR principles (Findable, Accessible, Interoperable, and Re-usable). Whereas FAIR looks at practical issues related to the sharing and distribution of data, FACT focuses more on the foundational scientific challenges.

van der Aalst W. M. P., Bichler M., and Heinzl A., "Responsible data science," *Bus Inf. Syst. Eng*. (2017) 59: 311–313 doi: 10.1007/s12599-017-0487-z

TRUST Principles

www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN COMMENT

The TRUST Principles for digital repositories

Dawei Lin[1,10], Jonathan Crabtree[2], Ingrid Dillo[3], Robert R. Downs[4], Rorie Edmunds[5], David Giaretta[6], Marisa De Giusti[7], Hervé L'Hours[8], Wim Hugo[9], Reyna Jenkyns[10], Varsha Khodiyar[11], Maryann E. Martone[12], Mustapha Mokrane[13], Vivek Navale[14], Jonathan Petters[15], Barbara Sierman[16], Dina V. Sokolova[16], Martina Stockhause[17] & John Westbrook[18]

As information and communication technology has become pervasive in our society, we are increasingly dependent on both digital data and repositories that provide access to and enable the use of such resources. Repositories must earn the trust of the communities they intend to serve and demonstrate that they are reliable and capable of appropriately managing the data they hold.

Following a year-long public discussion and building on existing community consensus[1], several stakeholders, representing various segments of the digital repository community, have collaboratively developed and endorsed a set of guiding principles to demonstrate digital repository trustworthiness. Transparency, Responsibility, User focus, Sustainability and Technology: the TRUST Principles provide a common framework to facilitate discussion and implementation of best practice in digital preservation by all stakeholders.

Context and History

For over sixty years, digital data stewardship and preservation have been central to the mission of academic institutions such as libraries, archives, and domain repositories[2] with many other stakeholders involved, including researchers, funders, infrastructure, and service providers. Scientific data management is receiving increasing attention inside and outside of the scientific community, particularly in the contemporary Open Science discourse. Consensus on 'good' data management practice is beginning to form, but there is still insufficient implementation in some scientific domains.

The FAIR Data Principles[3] highlight the need to embrace good practice by defining essential characteristics of data objects to ensure that data are reusable by humans and machines: they should be Findable, Accessible, Interoperable, and Reusable, i.e. FAIR. However, to make data FAIR whilst preserving them over time requires trustworthy digital repositories (TDRs) with sustainable governance and organizational frameworks, reliable infrastructure, and comprehensive policies supporting community-agreed practices. TDRs, with their clear remit to actively preserve data in response to changes in both technology and stakeholder requirements, play an important role in maintaining the value of data. They are held in a position of trust by their users as they accept the responsibilities of data stewardship. To fulfill this role, TDRs must demonstrate essential and enduring capabilities necessary to enable access and reuse of data over time for the communities they serve. TDRs support data

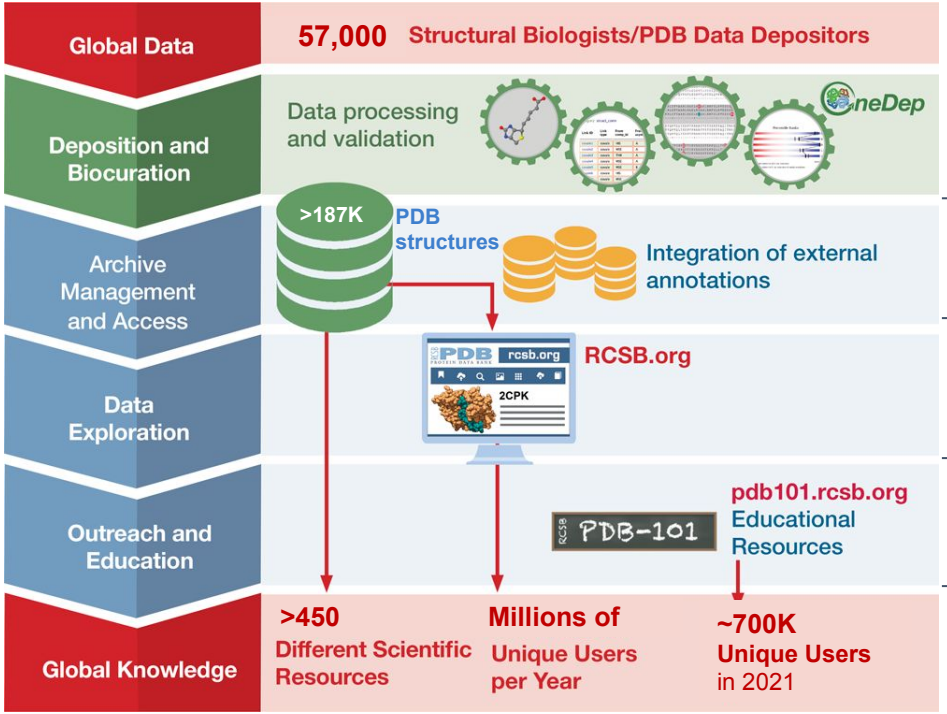
[1]Division of Allergy, Immunology, and Transplantation, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Maryland, USA. [2]HW Odum Institute for Research in Social Science, University of North Carolina at Chapel Hill, North Carolina, USA. [3]Data Archiving and Networked Services (DANS), The Hague, The Netherlands. [4]Center for International Earth Science Information Network (CIESIN), The Earth Institute, Columbia University, New York, USA. [5]World Data System of the International Science Council (WDS), WDS International Programme Office, Tokyo, Japan. [6]PTAB Ltd, Dorset, UK. [7]Universidad Nacional de La Plata, Comisión de Investigaciones Científicas de la Provincia de Buenos Aires, La Plata, Argentina. [8]UK Data Archive, UK Data Service, University of Essex, Colchester, UK. [9]South African Environmental Observation Network, Cape Town, South Africa. [10]Ocean Networks Canada, University of Victoria, Victoria, Canada. [11]Springer Nature, London, UK. [12]University of California, San Diego, California, USA and SciCrunch Inc., San Diego, USA. [13]Center for Information Technology, National Institutes of Health, Maryland, USA. [14]Data Services, University Libraries, Virginia Tech, Virginia, USA. [15]KB National Library of the Netherlands, The Hague, The Netherlands. [16]University Libraries, Columbia University, New York, USA. [17]German Climate Computing Center (DKRZ), Hamburg, Germany. [18]RCSB, Protein Data Bank, Rutgers, The State University of New Jersey, Institute for Quantitative Biomedicine at Rutgers, New Jersey, USA. [10]e-mail: dawei.lin@nih.gov

Box 1 The TRUST Principles

Principle	Guidance for repositories
Transparency	To be transparent about specific repository services and data holdings that are verifiable by publicly accessible evidence.
Responsibility	To be responsible for ensuring the authenticity and integrity of data holdings and for the reliability and persistence of its service.
User Focus	To ensure that the data management norms and expectations of target user communities are met.
Sustainability	To sustain services and preserve data holdings for the long-term.
Technology	To provide infrastructure and capabilities to support secure, persistent, and reliable services.

Lin, D., Crabtree, J., Dillo, I. *et al.* The TRUST Principles for digital repositories. *Sci Data* (2020) 7: 144 doi: 10.1038/s41597-020-0486-7

RCSB PDB: Driving Breakthroughs in Research and Education



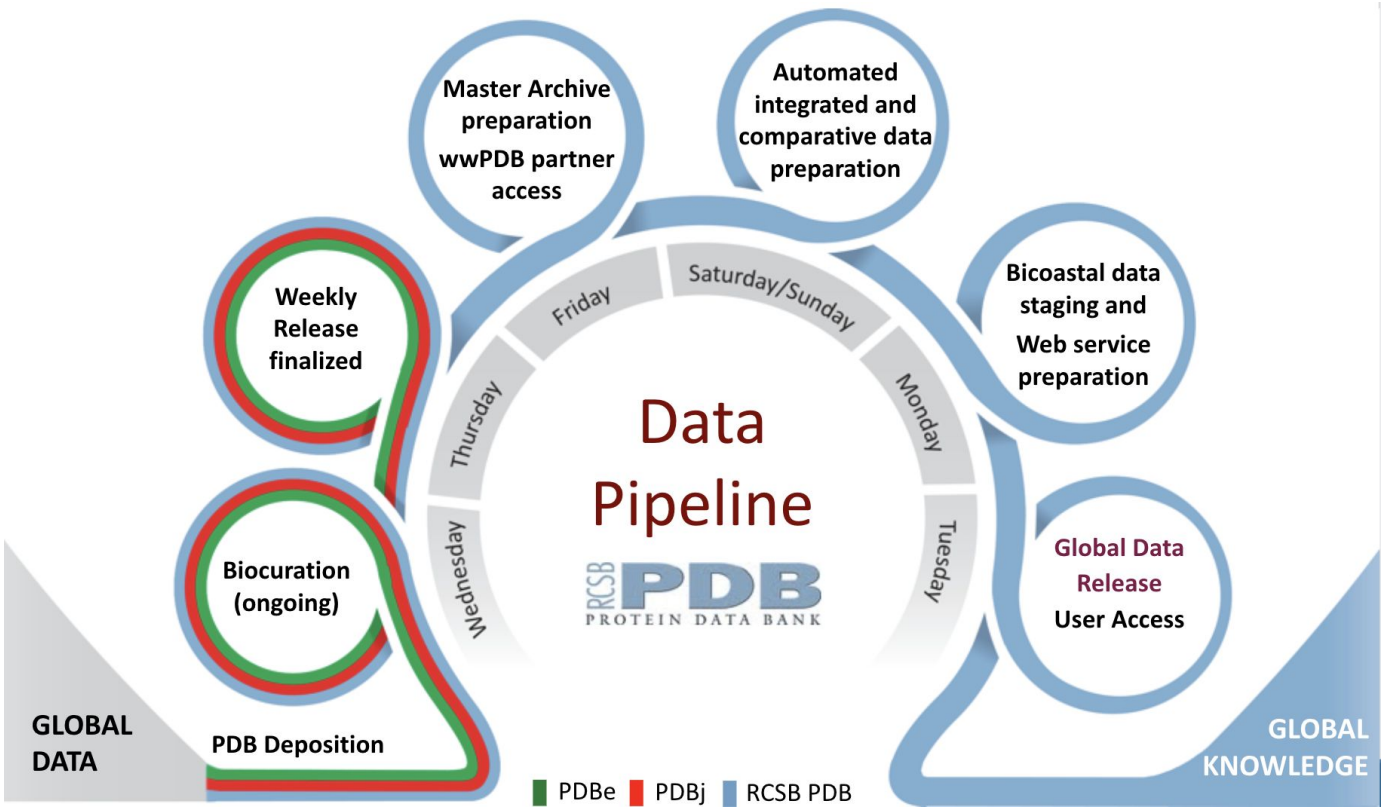
Deposition, Validation, Biocuration, Remediation

Data Standards, Integration, Storage, Access

Web Portal, Search/Analysis, 3D Visualization

Website and In-person Outreach

~250 Structures Released Per Week



Publicly Available PDB Data

PDB Archive contains >1 TB of Structure Data for Proteins, DNA, and RNA

The cost to replicate the contents of the PDB archive is estimated at **$18 billion (USD)** (Analysis)

The PDB Archive

- Grows at the rate of nearly 10% per year
- Used to download >2 million structure data files per day
- Managed by International collaboration US-Asia-Europe
- Manages "Big Data" as global Public Good

PDB Data

- Enable research in subject areas from Agriculture to Zoology (Analysis)
- Contributed data to nearly >1 million published research papers
- Used by >400 biological data resources

PDB Data Impact

- Basic and applied research
- Patent applications
- Discovery of lifesaving drugs
- Innovations that can lead to new product development and company formation
- STEAM education: PDB-101 provides curricula and online tools for teachers and students

Millions of Data Consumers worldwide served every year

Researchers, scientists, educators, students, curious public, medical professionals, patients, and patient advocates

Public and Private sectors, including pharmaceutical and biotechnology companies

Generates return on investment of **~1,500 times federal funding** (Analysis)

Remembering John Westbrook

For more than 25 years John supported Rutgers, RCSB PDB, and millions of data users worldwide with his vision and passion for innovative databases, ontologies, and other technologies for management of complex biological data.

As Data & Software Architect Lead of the RCSB PDB, John was central to the design and development of infrastructure and services to acquire, curate, archive, and deliver 3D macromolecular structure data to the broad community of PDB users.

His work helped establish the PDBx/mmCIF data dictionary and format as the foundation of the modern Protein Data Bank (PDB) archive.

John D. Westbrook Jr (1957–2021) Acta Cryst (2021) D77: 1475-1476 doi: 10.1107/S2059798321011402.

2021 Highlights

- 14,571 new PDB depositions
- 2,539 new ligands

- 185,541 released structures
- 2.3 billion data downloads from wwPDB sites combined
- Many millions of RCSB.org users

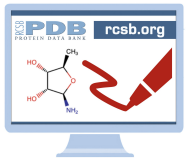
- ~700,000 PDB101.RCSB.org users
- 3 poster prizes for RCSB PDB undergraduate researchers

2021 Highlights

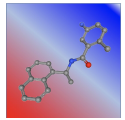
>14K depositions (>850 SARS-CoV-2)

BioArt Award

Chemical Sketch

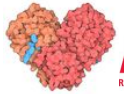


Ligand Validation



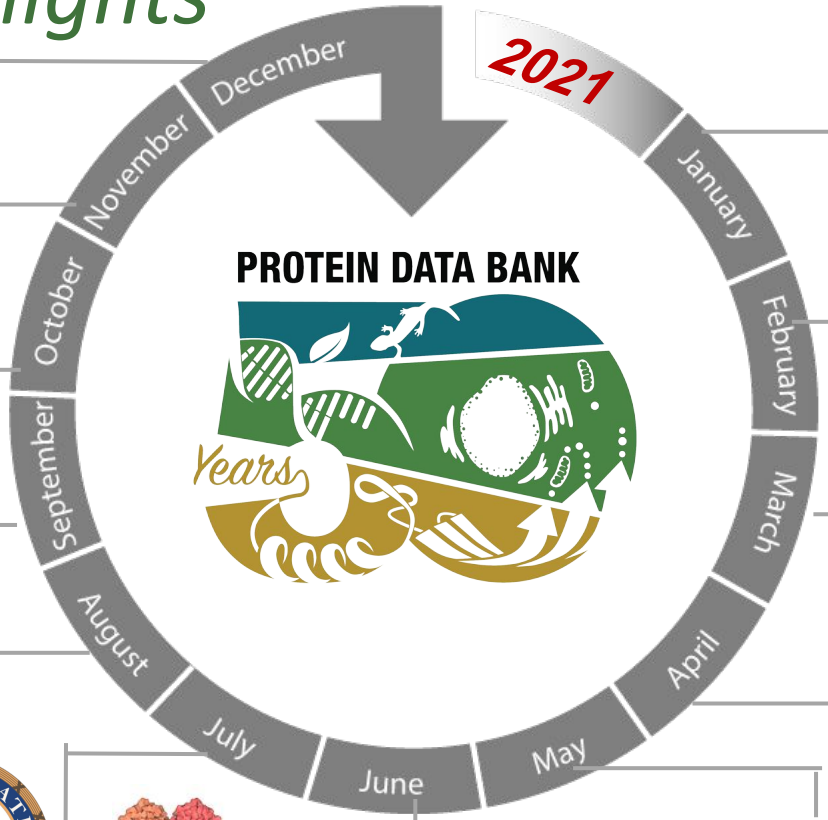
PDB50 Celebration

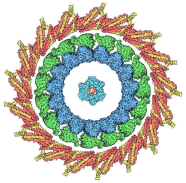
PDB50 read into Congressional Record



RISE at Rutgers
Research Intensive Summer Experience

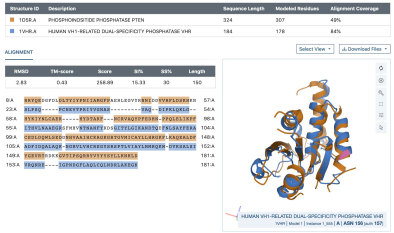
HS Video Challenge





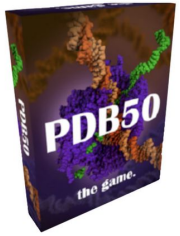
IQB Boot Camp
Scientific Communication

RU: Undergraduate Course
Drugs and the Brain



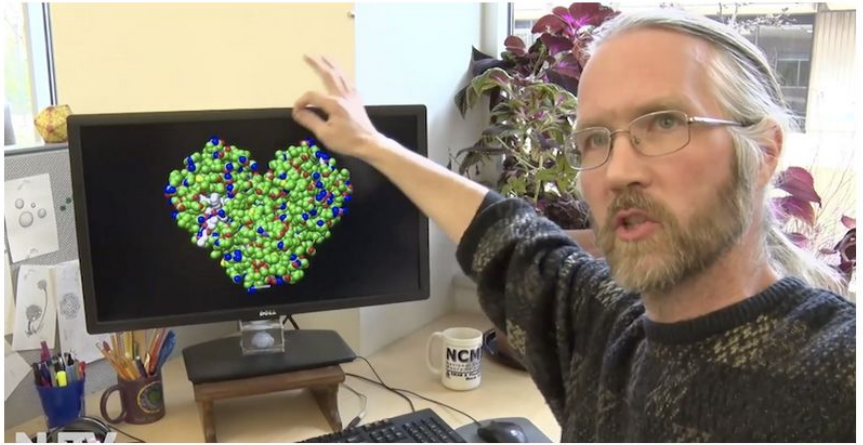
Pairwise Structure Alignment

1
4



Biocurator Milestone:
>10,000 Depositions Processed

RCSB PDB Biocurators Dr. Sutapa Ghosh and Dr. Monica Sekharan



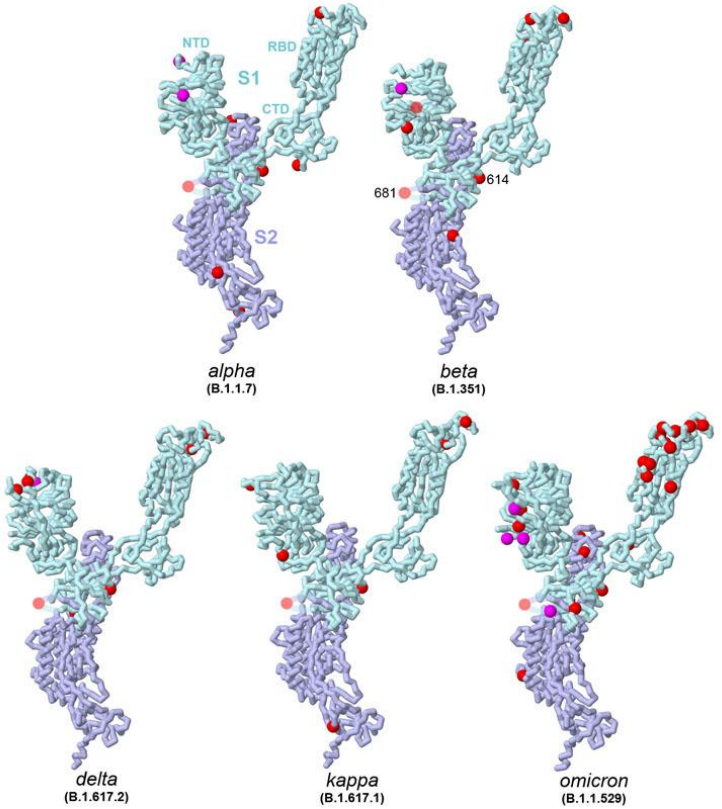
RCSB PDB Biocurator Dr. Brian P. Hudson talking about the SARS-CoV-2 main protease in February 2020.

Impact of COVID-19 on Operations

- Remote since March 2020
  - Regular All Hands and social hours in addition to standing meetings
- Biocuration of SARS-CoV-2 structures prioritized
  - >1,900 structures released as of March 9, 2022
  - 925 released in 2021
- Pandemic-related traffic to PDB-101 peaked in 2020
- Related resources updated weekly at RCSB.org/covid19

Return to Office

- Hybrid office at Rutgers, UCSD, UCSF
  - Will continue until institutions change policy and/or
  - Organization needs dictate otherwise
- Shared offices follow local social distancing guidelines
- Classes taking place in-person
- Current plans for in-person undergraduate summer research

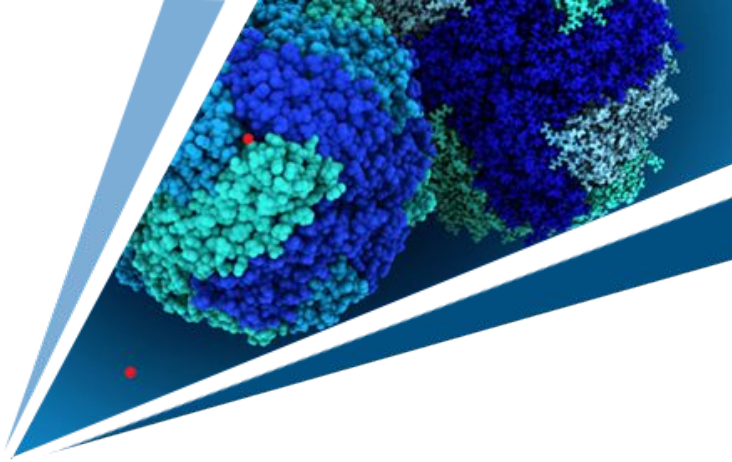


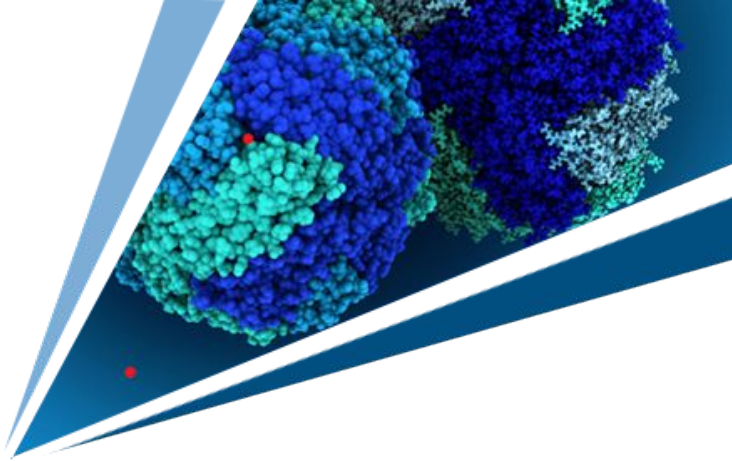
New Team Members

- Paul Craig: Visiting Scholar (UCSD + Rutgers) Sabbatical of collaboration on UXD, documentation, …
- Henry Chao: DevOps Lead (Rutgers)
- Connor Parker: Scientific Software Developer (Rutgers); working with RCSB PDB before graduate school in the fall

Recruitment, Interviews, and hiring are ongoing; More to come...

Questions and Comments?





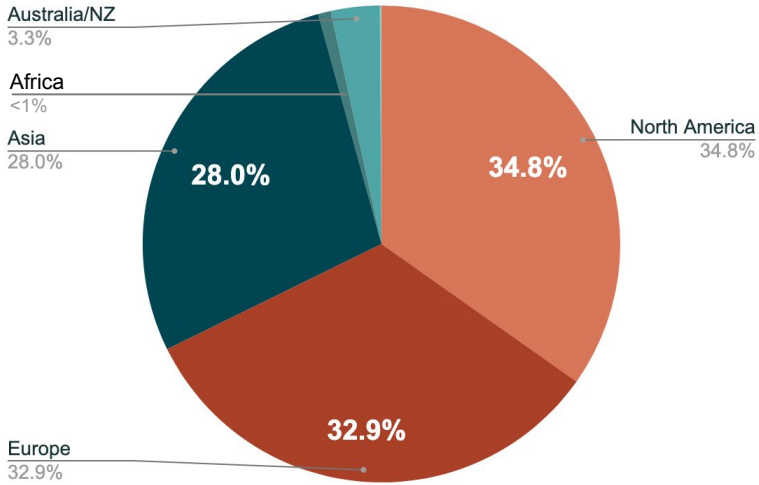
Data Deposition, Validation, Biocuration, Archive Management

Jasmine Young

2021 wwPDB Deposition/Biocuration Statistics

- 14,571 new depositions
  (15,436 in 2020; 13,377 in 2019)
- 5,687 RCSB PDB-biocurated
  (7,190 in 2020; 5,523 in 2019)
- Biocuration workload is geographically-balanced
  - 38.1% Americas, Oceania
  - 33.0% Europe, Africa
  - 28.0% Asia

Depositor Location



Deposition/Validation/Biocuration in 2021

- OneDep System for Deposition, Validation, and Biocuration:
  12 software upgrades; improvements included Mol* 3D assembly views for depositors, inclusion of extended PDB IDs and DOIs in PDBx/mmCIF data files, and validation and other enhancements
- Improved automation for biocurating incoming depositions
  - 76% of new entries now pass automatically through Entity Transformer module standardizes polymer and non-polymer entity representation
  - 25% pass without biocurator assistance through sequence processing
  - 57% pass automatically through ligand processing

Recent Validation Report Enhancements

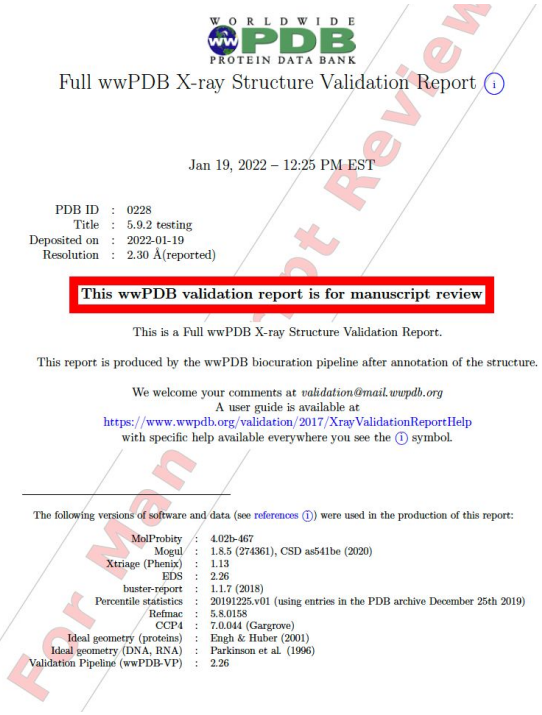
- Available for all released 3DEM PDB structures and EMDB maps
  - Include images for deposited masks and improved map-model overlay
- Provided in PDBx/mmCIF to make validation data more interoperable with the archival format and more "database-friendly"
  - Easier to interpret (more "readable" by humans)
  - Contains a high-level summary and easier access to residue-level information
- More prominent labeling of the types of validation reports
  - *e.g.*, For Manuscript Review *vs.* Not For Manuscript Review

Communicating Data Quality to Depositors and Journals

- OneDep submission blocked for certain data errors (e.g., physically impossible values)
- Outliers highlighted in the wwPDB validation reports
  - Biocurators highlight major issues for depositor correction
  - Depositors must download and accept data quality assessed in the validation reports for submission
- Many journals require wwPDB Validation Report with manuscript submission
  - Validation reports with *For Manuscript Review* watermark provided by biocurators

Types of wwPDB Validation Reports

Official Report: For Manuscript Review



WORLDWIDE PDB
PROTEIN DATA BANK

Full wwPDB X-ray Structure Validation Report (i)

Jan 19, 2022 – 12:25 PM EST

PDB ID : 0228
Title : 5.9.2 testing
Deposited on : 2022-01-19
Resolution : 2.30 Å(reported)

This wwPDB validation report is for manuscript review

This is a Full wwPDB X-ray Structure Validation Report.

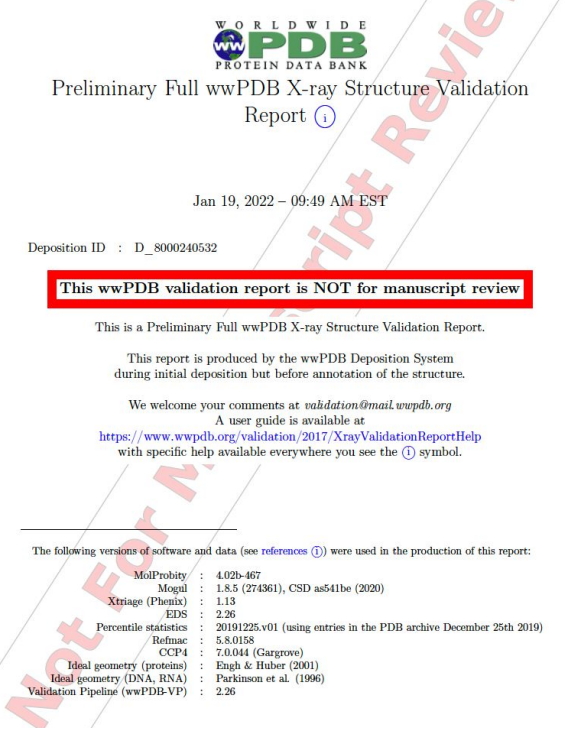
This report is produced by the wwPDB biocuration pipeline after annotation of the structure.

We welcome your comments at *validation@mail.wwpdb.org*
A user guide is available at
https://www.wwpdb.org/validation/2017/XrayValidationReportHelp
with specific help available everywhere you see the (i) symbol.

The following versions of software and data (see references (i)) were used in the production of this report:

MolProbity	:	4.02b-467
Mogul	:	1.8.5 (274361), CSD as541be (2020)
Xtriage (Phenix)	:	1.13
EDS	:	2.26
buster-report	:	1.1.7 (2018)
Percentile statistics	:	20191225.v01 (using entries in the PDB archive December 25th 2019)
Refmac	:	5.8.0158
CCP4	:	7.0.044 (Gargrove)
Ideal geometry (proteins)	:	Engh & Huber (2001)
Ideal geometry (DNA, RNA)	:	Parkinson et al. (1996)
Validation Pipeline (wwPDB-VP)	:	2.26

Preliminary Report: Not For Manuscript Review



WORLDWIDE PDB
PROTEIN DATA BANK

Preliminary Full wwPDB X-ray Structure Validation Report (i)

Jan 19, 2022 – 09:49 AM EST

Deposition ID : D_8000240532

This wwPDB validation report is NOT for manuscript review

This is a Preliminary Full wwPDB X-ray Structure Validation Report.

This report is produced by the wwPDB Deposition System
during initial deposition but before annotation of the structure.

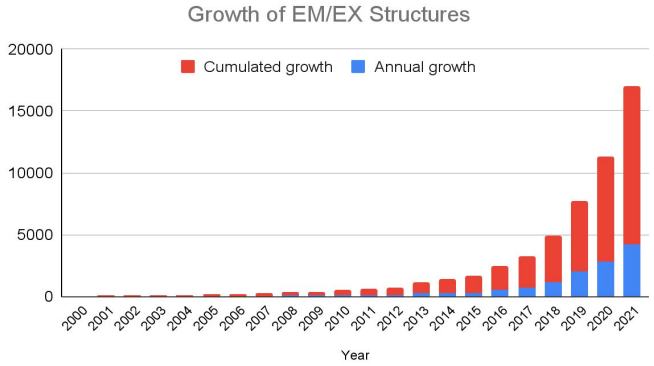
We welcome your comments at *validation@mail.wwpdb.org*
A user guide is available at
https://www.wwpdb.org/validation/2017/XrayValidationReportHelp
with specific help available everywhere you see the (i) symbol.

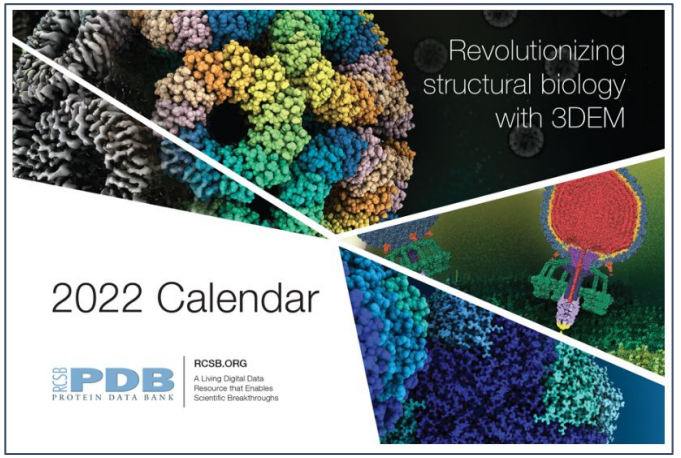
The following versions of software and data (see references (i)) were used in the production of this report:

MolProbity	:	4.02b-467
Mogul	:	1.8.5 (274361), CSD as541be (2020)
Xtriage (Phenix)	:	1.13
EDS	:	2.26
Percentile statistics	:	20191225.v01 (using entries in the PDB archive December 25th 2019)
Refmac	:	5.8.0158
CCP4	:	7.0.044 (Gargrove)
Ideal geometry (proteins)	:	Engh & Huber (2001)
Ideal geometry (DNA, RNA)	:	Parkinson et al. (1996)
Validation Pipeline (wwPDB-VP)	:	2.26

3DEM Growth: Higher Resolution/Complexity

- Continue rapid growth in 3DEM deposition
  - 4254 new structures in 2021 (53% increase *versus* 2020)
- Some 3DEM structures reaching resolution limits comparable to MX
- Complexity continues to grow (see 2022 RCSB PDB Calendar)
- 2022 should see total 3DEM holdings exceed NMR





Community Development of 3DEM Validation

2020 EM Data Management Workshop

- White paper to be submitted soon
- Recommendations include
  - Show calculated and author-provided FSC curves in the same plot (Implemented)
  - Include which types of model restraints were used
  - Include a map-model FSC plot to indicate correlation between the experimental map and one computed from the model as a function of resolution

EMDR and Validation Challenges

- EMDR funded by NIGMS R01GM079429-12 (Chiu, PI)
- Promotes community development of validation and standards
- Produces best models against selected maps
- Explores model metrics with focus on fit-to-map
- Burley now Rutgers PI on EMDR R01 submission to NIGMS

Plans for 3DEM-related Improvements

Completed/Ongoing

- Enhanced validation reports
- Better streamlined release between PDB and EMDB
- Made deposition of half-maps mandatory
- More mandatory metadata items
- Increasing controlled vocabularies

Future Plan

- Distribute status information of unreleased entries
- Improve metadata checking with more rigorous validation
- Streamline deposition
- Provide per-amino acid fitting to map

PDB Archive Growth in 2021

Year-end holdings > 185K structures

- 12,602 new structures released (14,044 in 2020; 11,501 in 2019)

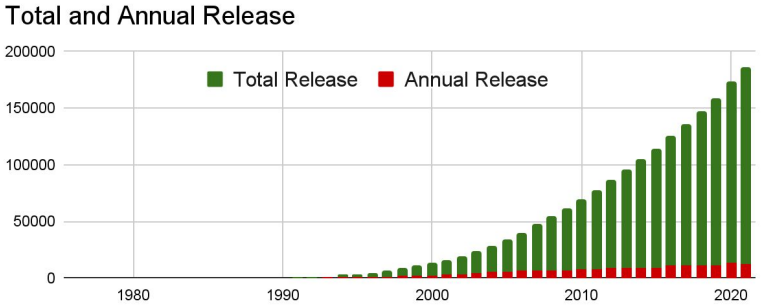
Data In Storage

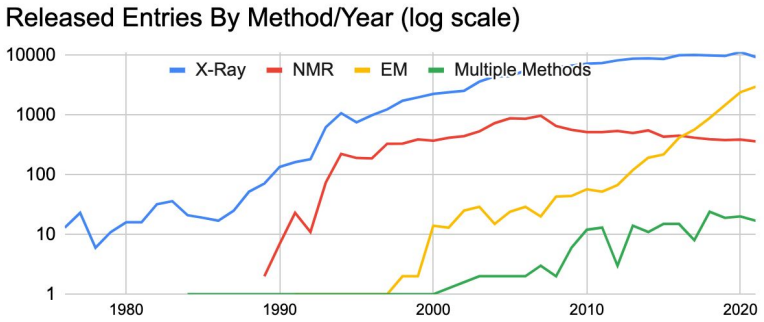
- OneDep data: 31.3 TB
- PDB ftp (legacy + versioned): 923 GB
- EMDB ftp: 4.5 TB

Storage for all Deployed Core Archive Services: 109 TB

- Redundant copies of all ftp servers, map and model services, images, and annual snapshots

CoreTrustSeal certification (through April 2024) (CoreTrustSeal.org)





All PDB Statistics

Archive Management and PDB Exchange Dictionary Enhancements

PDB Core Archive

- Distribution of PDBx/mmCIF-formatted validation reports
- Distribution of individual Chemical Component definition files
- Better-organized inventory lists
- Standardized SARS-CoV-2 annotation
- Standardized Chemical Component Dictionary (CCD) synonyms

PDBx/mmCIF Dictionary

- Better support for X-ray data: anisotropic diffraction limits, unmerged reflection data, and anomalous diffraction data
- Support for extended PDB ID and DOI
- Support enhanced assembly files
- Better collection of starting model
- Support better-organized CCD synonyms
- Support 3DEM improvements
- Repackaging of dictionaries on GitHub

http://www.wwpdb.org/news/news?year=2021#613b93b3ef055f03d1f222cf

Chemical Reference Data Growth

Improved Access: Individual Chemical Component entries are available for download as individual files from the PDB archive

>36K Chemical Component Definitions

- 2538 new in 2021 (2683 in 2019, 3143 in 2020)
- 3137 updated (2708 in 2019, 4334 in 2020)

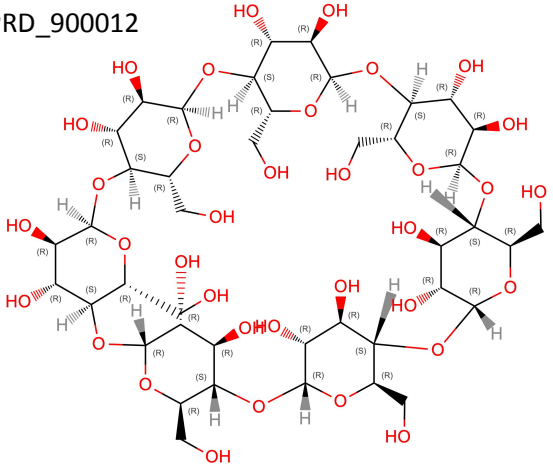
~1K Biologically Interesting molecule Reference Dictionary (BIRD) Definitions

- 29 new in 2021 (9 in 2019, 170 in 2020)
- 3 updated (4 in 2019, 144 in 2020)

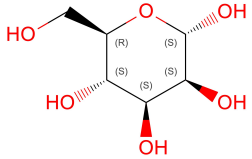
Definitions Enhanced by External Resources

CCDC, PubChem, ChEMBL, ChEBI, DrugBank, Pharos, CAS

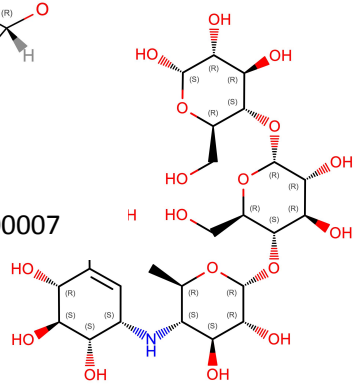
PRD_900012



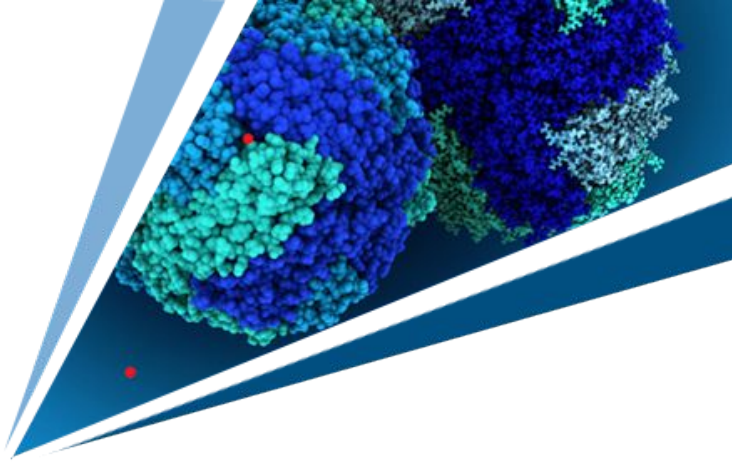
MAN

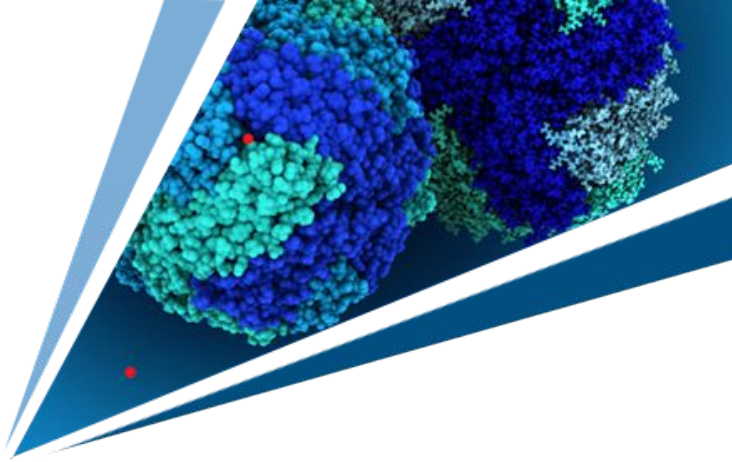


PRD_900007



Questions and Comments?



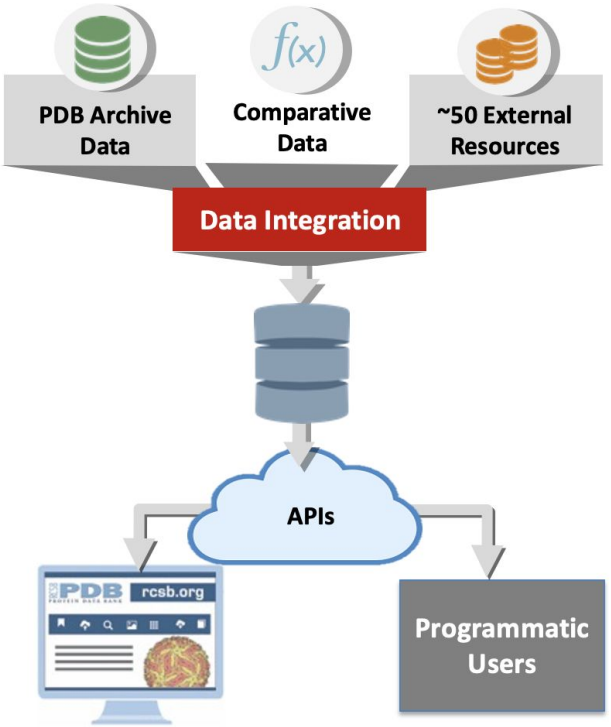


Data Exploration (RCSB.org)

Yana Rose

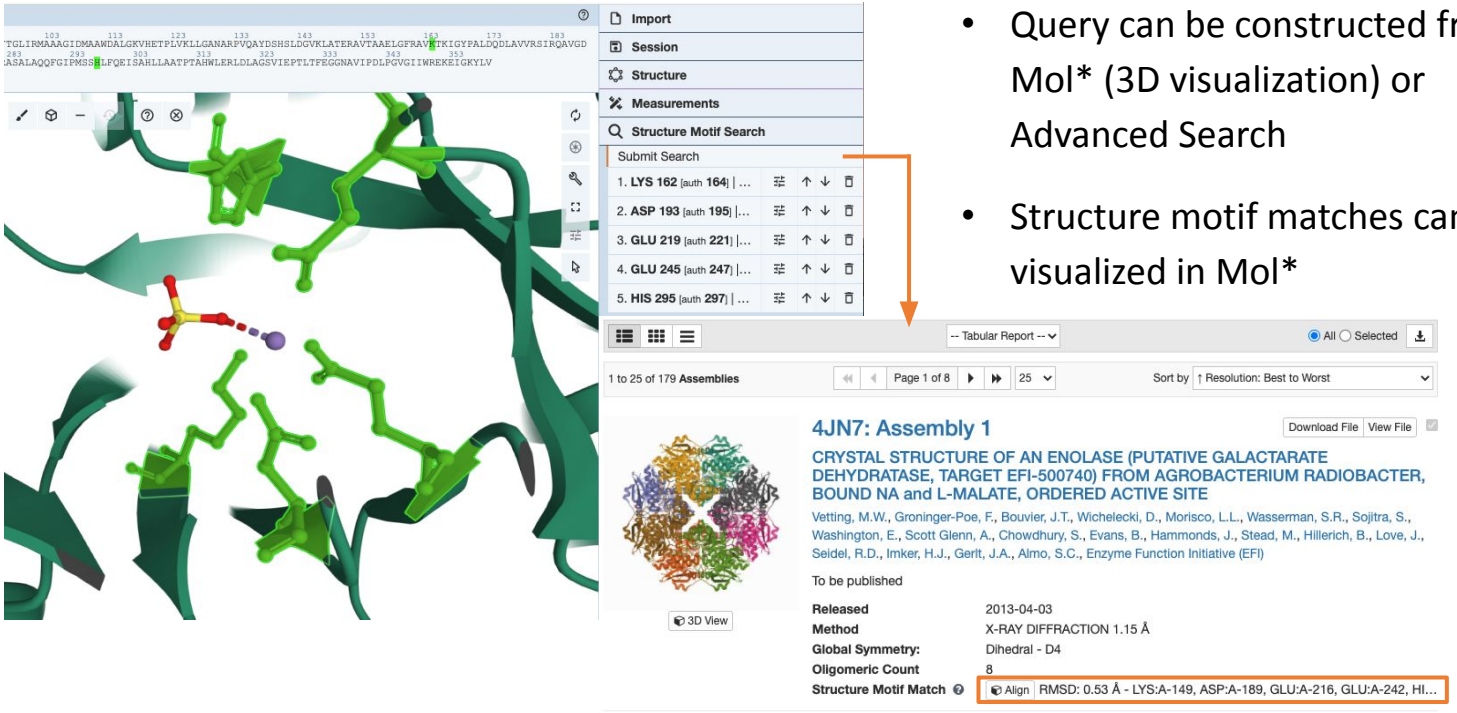
RCSB.org Services Overview

- Data Integration
  - Loading PDB archival data
  - Generating data derived from archival contents
  - Integrating data from external resources
- Data Access
  - Search services
  - Data delivery services
- Data Exploration
  - Resources that help users explore, visualize, and analyze PDB data
  - MyPDB: stores searches, sends alerts when related data are released
  - Related content: documentation, news, newsletters, etc.



New: Structure Motif Search

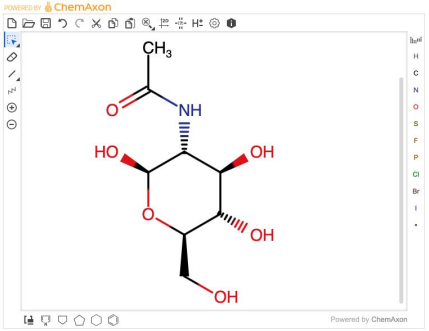
A new service finds structures containing a small number of specific amino acids in proximity



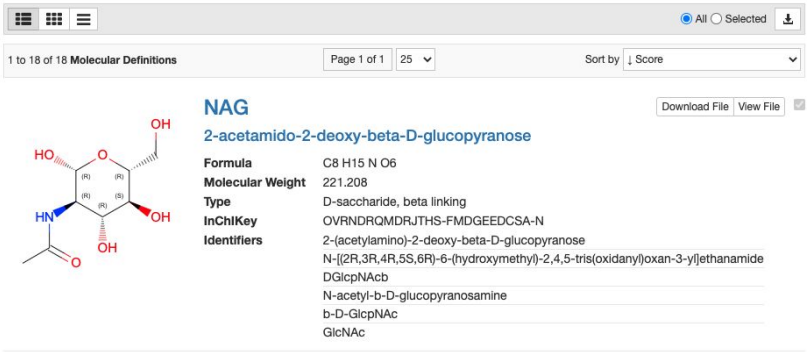
- Query can be constructed from Mol* (3D visualization) or Advanced Search
- Structure motif matches can be visualized in Mol*

Small Molecule Search Options

- Chemical sketch tool: web-based tool to build chemical drawings to search for small molecules in the PDB, based on chemical drawing tool Marvin JS from ChemAxon

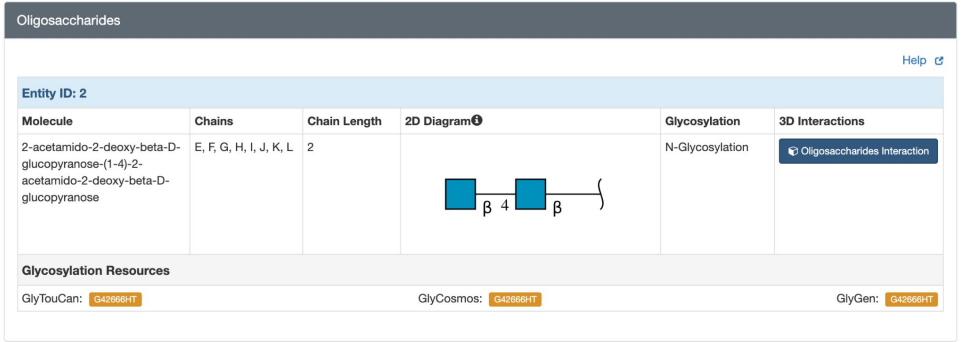


- Molecular Definitions search results option: capability to perform integrated searches and return components as defined in the Chemical Component Dictionary (CCD) and Biologically Interesting molecule Reference Dictionary (BIRD)



New: Integration with External Resources

- Glycosylation resources: GlyTouCan, GlyCosmos, and GlyGen

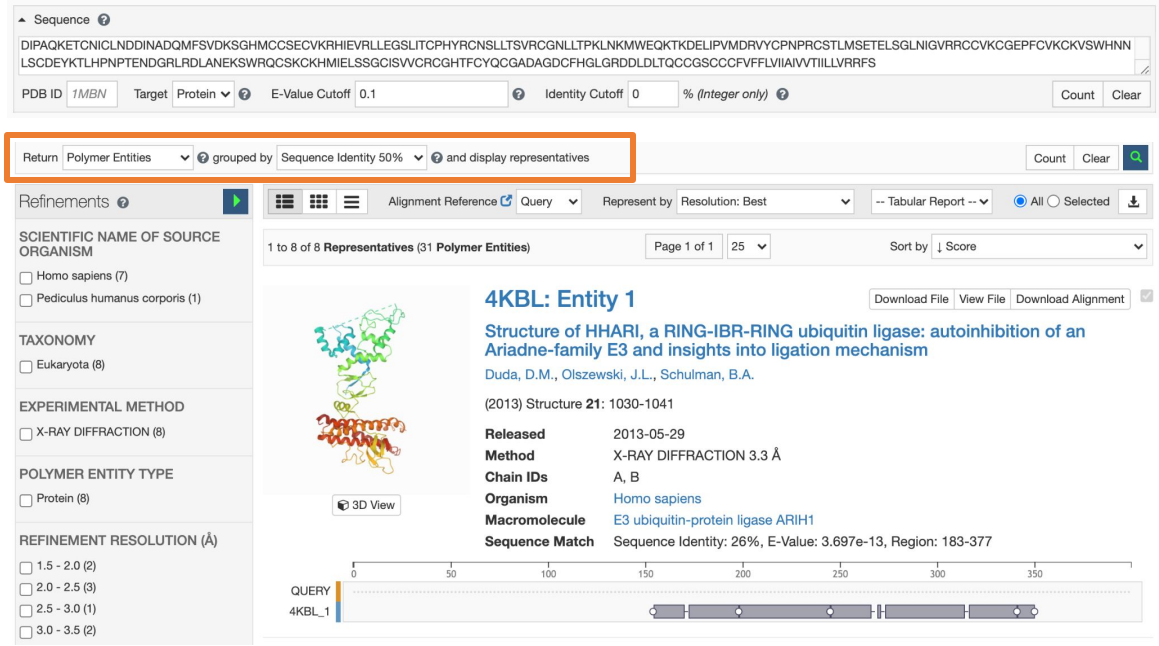


- New membrane protein annotations from OPM, PDBTM, and MemProtMD (in addition to mpstruct)

New: Streamlining Search Results

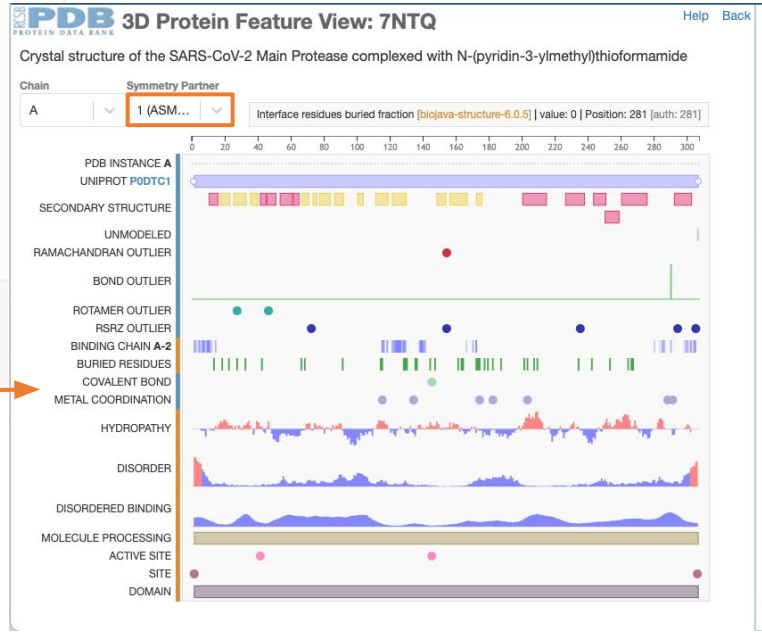
Filtering of highly-similar Search Results based on

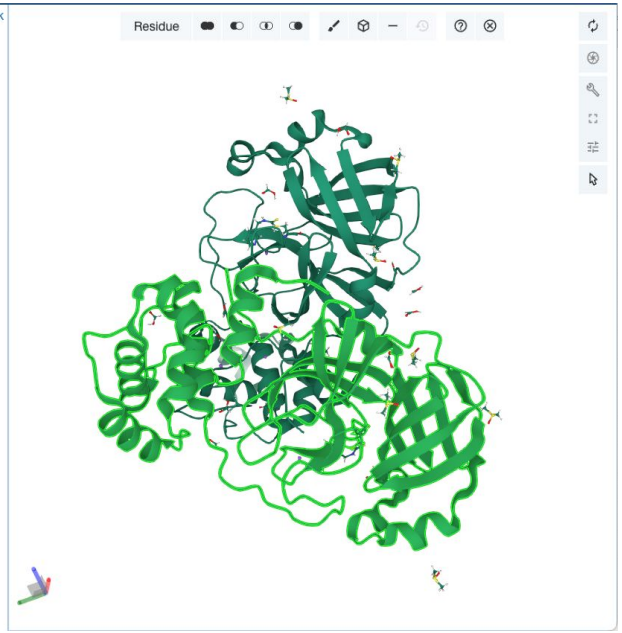
- Sequence Identity
- UniProt ID
- Group Deposition ID



3D Protein Feature View

Users can view and select sequence features in specific symmetry partners.

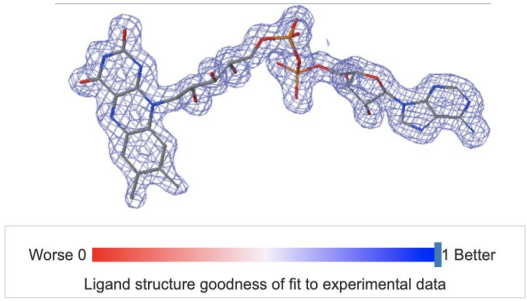




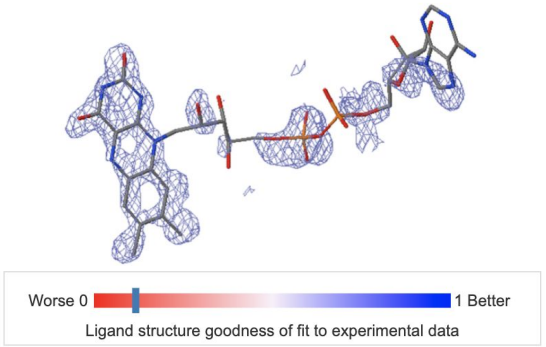
3D Protein Feature View can be easily accessed from SSP

New: Ligand Quality Assessment at RCSB.org

- New graphics summarize the fitting quality of the Ligand of Interest on Structure Summary pages
- Clicking on the bar will take the user to the *Ligands* tab for more information about this ligand in this particular entry and across the archive
- Response to recommendation made at 2019 RCSB PDB AC Meeting



FAD in PDB structure 5NAK



FAD in PDB structure 2CZ8

Shao et al. (2022) *Structure* 252-262.e4 doi: 10.1016/j.str.2021.10.003

FAD in PDB structure 5NAK

Structure Summary | 3D View | Annotations | Experiment | Sequence | Genome | Ligands | Versions

Display Files | Download Files

5NAK

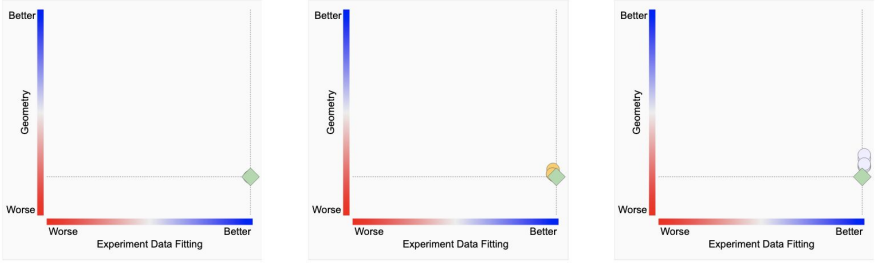
FAD | CL | GOL | KYN

FAD: FLAVIN-ADENINE DINUCLEOTIDE

Ligand Definition and Summary of FAD

FAD is a Ligand Of Interest in 5NAK designated by the RCSB

Help



Best-fitted instance in this entry
Other instances in this entry
Best-fitted PDB instances with same target (top 5)
Best-fitted PDB instances (top 5)

Identifier	Ranking for goodness of fit	Ranking for geometry	Real space R factor	Real space correlation coefficient	RMSZ-bond-length	RMSZ-bond-angle	Outliers of bond length	Outliers of bond angle	Atomic clashes	Stereochemical errors	Model completeness	Average occupancy
5NAK_FAD_A_501	99%	19%	0.056	0.986	1.52	2.13	5	7	0	0	100%	1
5NAK_FAD_B_501	99%	19%	0.057	0.983	1.51	2.14	5	6	0	0	100%	1
5N7T_FAD_B_501	98%	19%	0.067	0.987	1.48	2.12	5	6	0	0	100%	1
5NAE_FAD_B_501	98%	20%	0.067	0.986	1.4	2.16	5	6	0	0	100%	1
5NAH_FAD_B_501	98%	20%	0.061	0.978	1.35	2.17	5	6	0	0	100%	1
6FOY_FAD_B_501	97%	22%	0.073	0.989	1.33	2.07	6	7	0	0	100%	1
5NAB_FAD_A_501	97%	20%	0.071	0.985	1.41	2.14	4	7	0	0	100%	1
4KGD_FAD_B_701	100%	28%	0.031	0.997	1.29	1.79	6	5	0	0	100%	1
4FEE_FAD_B_701	100%	24%	0.034	0.996	1.45	1.87	4	6	0	0	100%	1
4FEG_FAD_B_710	100%	24%	0.033	0.997	1.44	1.84	5	9	1	0	100%	1
2PGO_FAD_B_613	100%	29%	0.036	0.993	1.12	1.85	4	5	0	0	100%	1
6RK0_FAD_B_403	100%	25%	0.036	0.993	1.12	2.1	5	7	0	0	100%	1

FAD in PDB structure 2CZ8

Structure Summary | 3D View | Annotations | Experiment | Sequence | Genome | Ligands | Versions

Display Files | Download Files

2CZ8

PO4 | K | FAD

FAD: FLAVIN-ADENINE DINUCLEOTIDE

Ligand Definition and Summary of FAD

FAD is a Ligand Of Interest in 2CZ8 designated by the RCSB

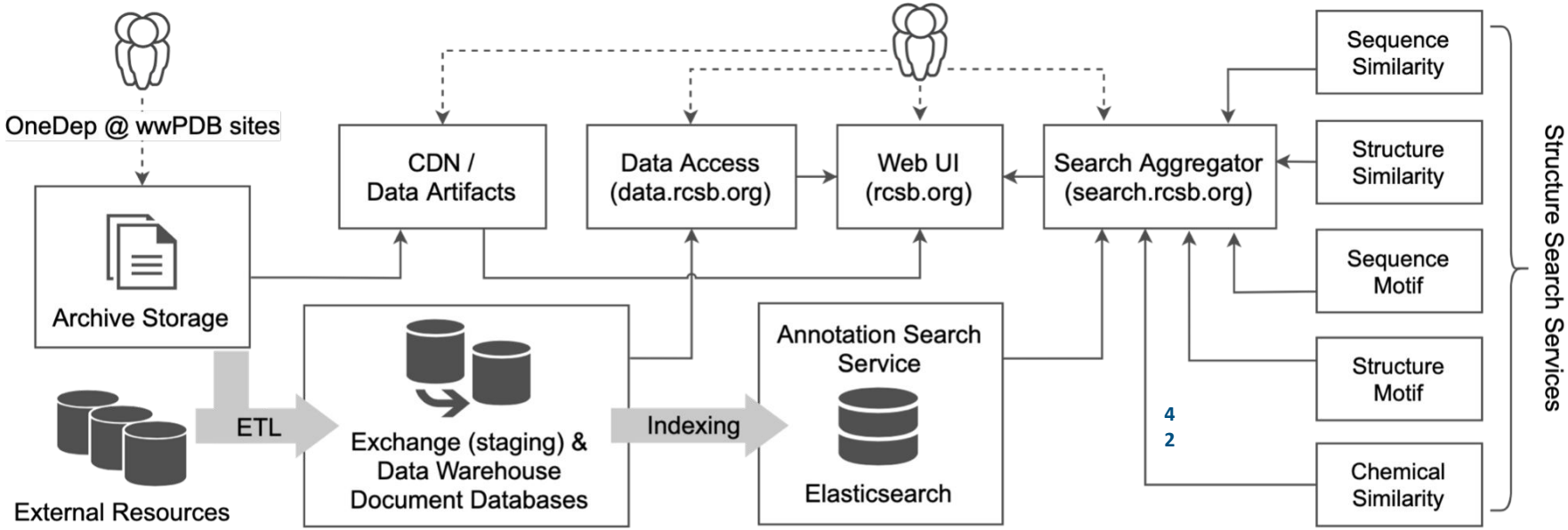
Help

Best-fitted instance in this entry
Other instances in this entry
Best-fitted instance in this entry
Best-fitted PDB instances (top 5)

Identifier	Ranking for goodness of fit	Ranking for geometry	Real space R factor	Real space correlation coefficient	RMSZ-bond-length	RMSZ-bond-angle	Outliers of bond length	Outliers of bond angle	Atomic clashes	Stereochemical errors	Model completeness	Average occupancy
2CZ8_FAD_D_1204	12%	14%	0.23	0.748	2.04	1.97	9	3	2	0	100%	1
2CZ8_FAD_H_1208	11%	15%	0.241	0.746	1.99	1.96	9	3	2	0	100%	1
2CZ8_FAD_C_1203	8%	15%	0.249	0.72	2.01	1.94	9	4	5	0	100%	1
2CZ8_FAD_G_1207	6%	15%	0.284	0.707	2.02	1.94	9	4	5	0	100%	1
2CZ8_FAD_E_1205	5%	15%	0.296	0.698	2.01	1.96	9	4	3	0	100%	1
2CZ8_FAD_F_1206	5%	15%	0.25	0.647	2.02	1.97	9	5	1	0	100%	1
2CZ8_FAD_A_1201	4%	15%	0.318	0.693	2.02	1.98	9	4	4	0	100%	1
2CZ8_FAD_B_1202	2%	15%	0.3	0.589	2.04	1.96	9	4	3	0	100%	1
4KGD_FAD_B_701	100%	28%	0.031	0.997	1.29	1.79	6	5	0	0	100%	1
4FEE_FAD_B_701	100%	24%	0.034	0.996	1.45	1.87	4	6	0	0	100%	1
4FEG_FAD_B_710	100%	24%	0.033	0.997	1.44	1.84	5	9	1	0	100%	1
2PGO_FAD_B_613	100%	29%	0.036	0.993	1.12	1.85	4	5	0	0	100%	1
6RK0_FAD_B_403	100%	25%	0.036	0.993	1.12	2.1	5	7	0	0	100%	1

Search Architecture Driving RCSB.org

Background Information; *RCSB Protein Data Bank: Architectural Advances Towards Integrated Searching and Efficient Access to Macromolecular Structure Data from the PDB Archive,* https://doi.org/10.1016/j.jmb.2020.11.003



Rose et al. (2021) *Journal of Molecular Biology 433,* 166704 doi: doi.org/10.1016/j.jmb.2020.11.003

2021 RCSB.org Usage Summary

RCSB.org

- Google Analytics: ~4.7 million unique users (~3.9 M in 2019)
  - 12 million sessions
  - 46 million page views
  - ~257.71 TB of data accessed
- Internal tracking: ~7 million unique IP addresses

Data File Downloads

- RCSB PDB: 1,828,603,104 downloads;
  604,544,570 (web), 1,224,058,534 (FTP)
- All wwPDB: 2,364,150,827 downloads;
  719,560,727 (web), 1,644,590,100 (FTP)

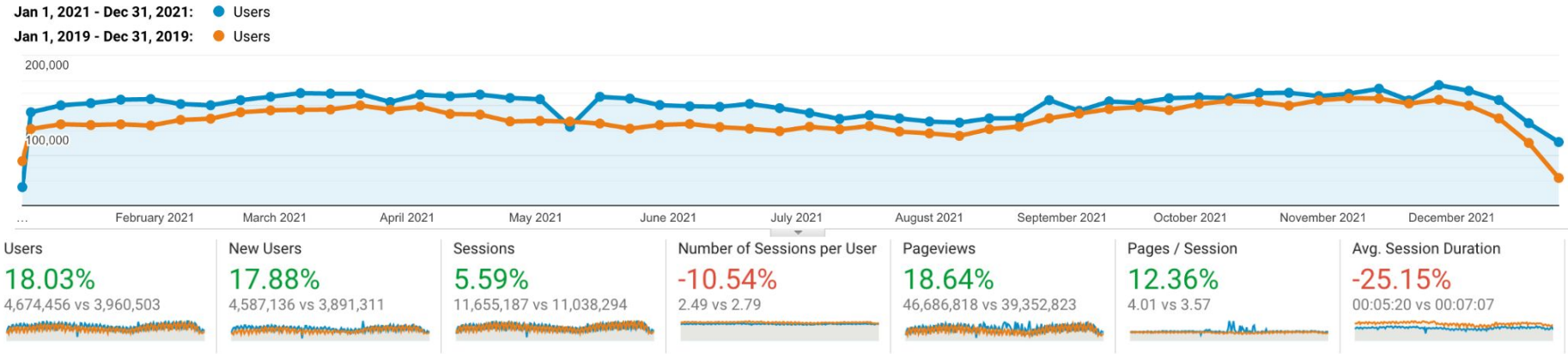
External Utilization

- >450 external resources repackage and distribute data
- Biopharm/biotech companies use PDB data internally
- Protein structure prediction accelerated by artificial intelligence approaches depend upon open access to PDB data

2.1M sessions from US (GA)

Tracked using Google Analytics and internal logging services

RCSB.org Traffic: Increase *versus* 2019 Visits

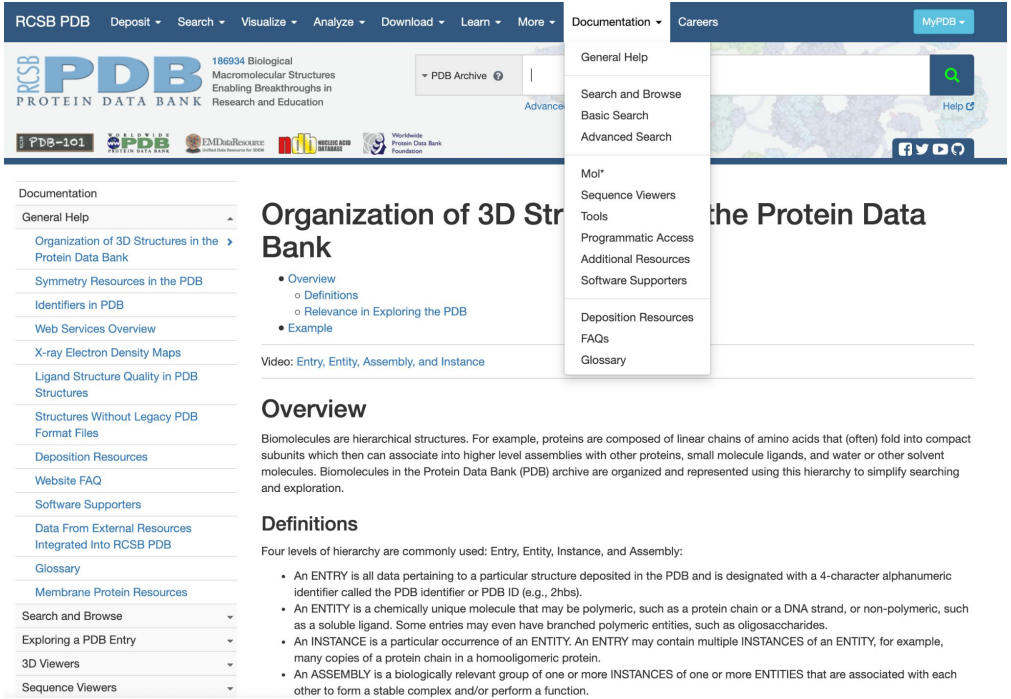


	Google Analytics: Unique Users	Internal Tracking: Unique IP Addresses	Google Analytics: Page Views	Internal Tracking: Hits (all pages, images, files)
2021	4,674,456	6,845,233	46,686,818	2,154,402,236
2020	3,632,445 (incomplete data)	6,677,853	42,113,933*	1,692,857,853
2019	3,960,503	6,415,390	39,352,823	1,230,667,138
2018	3,595,930	4,784,921	36,484,191	1,286,305,223

We assume that Google Analytics filters out some robotic access; *Estimated page views; Google Analytics data for 2020 is incomplete.

Customer Service and Improved Documentation

- Electronic Help Desk supports ~1,000 conversations/year
  - Feature requests
  - Request for guidance on usage
  - Bug reports
- New, detailed RCSB.org documentation
  - Generated in tandem with RCSB.org development
  - Linked from RCSB.org features and top menu
  - Searchable from header search box



Rutgers User Experience Design (UXD) Review

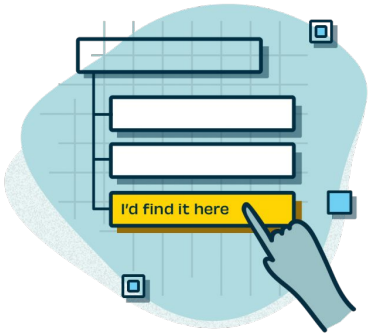
Spring Semester 2022: Two students in Master of Business and Science (MBS) Degree program

Goals

- Identify how undergraduate professors use RCSB.org in classroom
- Improve the usability and experience of the tools they frequently use

Process

- Research
  - Paul Craig organizing interviews with BioMolViz community
- Analyze
- Design wireframes
- Test
- Deliver recommendations



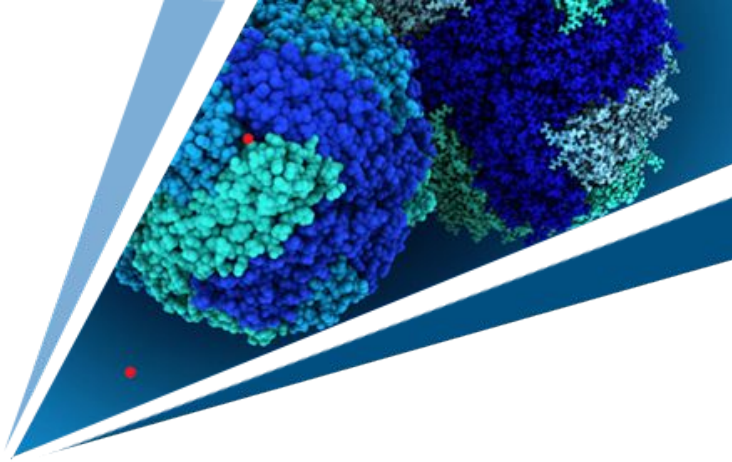
Treejack Testing

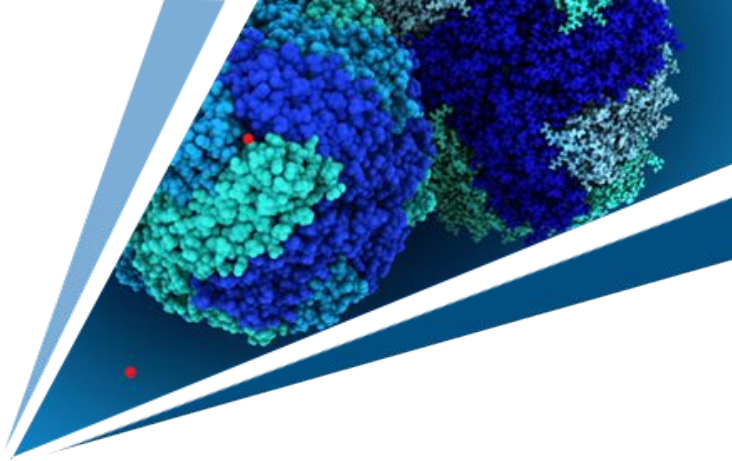
To find out how easily users can find information on the website

First-Click Testing

To verify that the first click a user makes on an interface to carry out a given task is clear and easy

Questions and Comments?





Outreach/Education

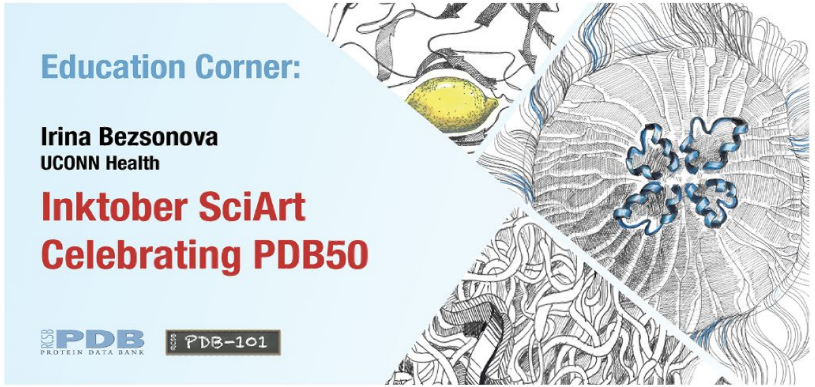
Christine Zardecki

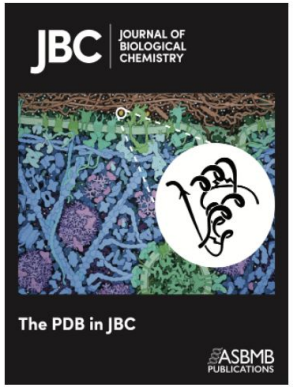
2021: Celebrating 50 Years of the PDB

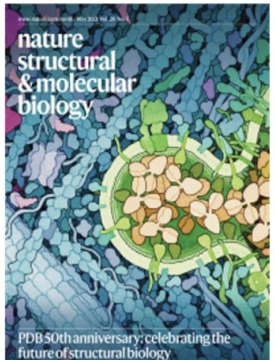
- 8 virtual symposia and webinars, each hosting hundreds of participants
  - Co-hosted with ASBMB, ACA, EMBL, Rutgers, ACS, Biophysical Society, Royal Society of Chemistry, Biophysical Society of Japan
  - Video recordings are being published on YouTube
- Special journal collections and publications
  - *Nature*
  - *Journal of Biological Chemistry*
- Related materials hosted at RCSB.org/pdb50

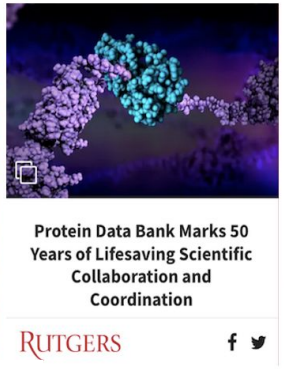
The inaugural event was sponsored by wwPDB Foundation and hosted 916 participants and 275 posters; 209 student/postdocs were eligible for 4 prize awards

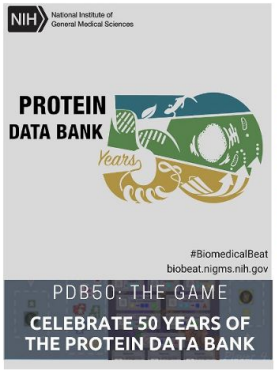
Journal Collections and Articles

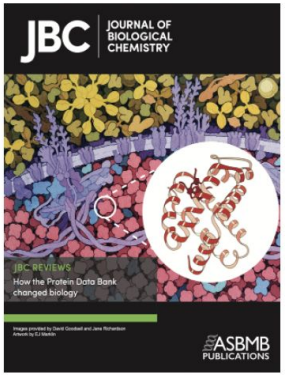






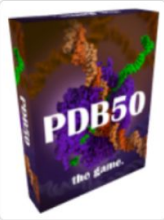






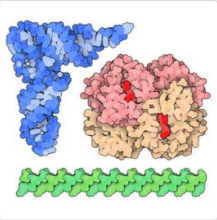
RCSB.org/pdb50

Special Projects Celebrating PDB50

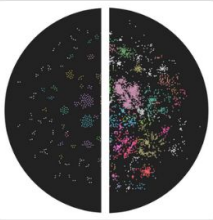


PDB50 Game: This gives players the opportunity to explore the process of structure discovery.

Structural Biology Playing Cards



Molecule of the Month: Fifty Years of Open Access to PDB Structures



PDB Citation MeSH Network Explorer

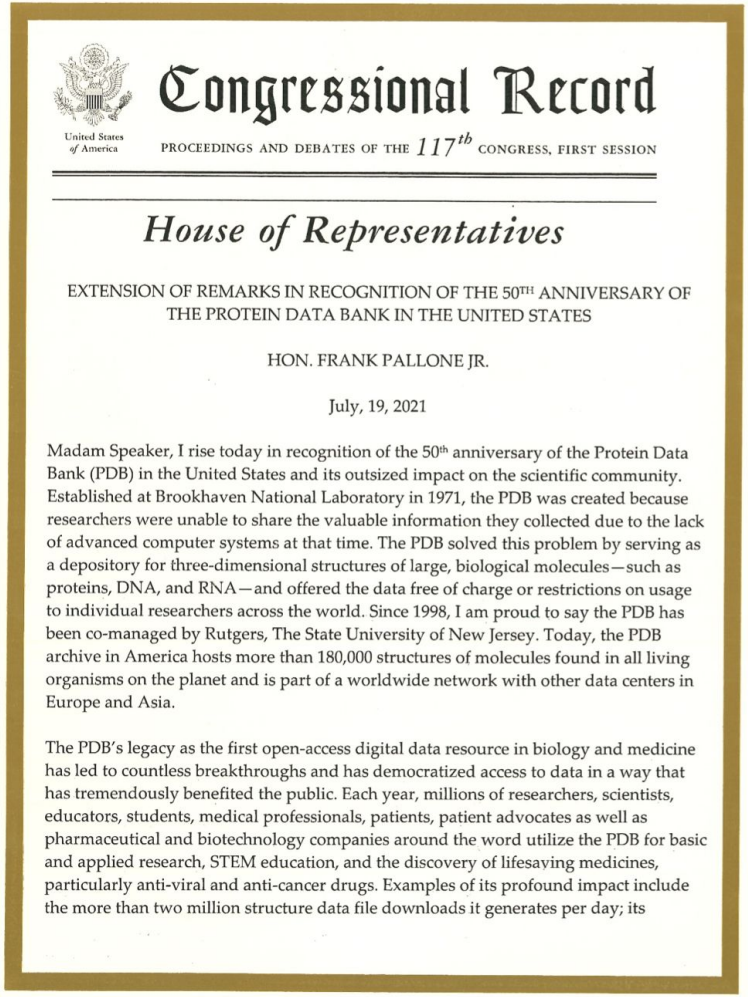
Video: Celebrating in Virtual Reality
Exploration of groundbreaking PDB structures for structure-based drug design.

Education Corner: Inktober SciArt Celebrating PDB50

PDB50 was noted in the Congressional Record (HTML | PDF)

Cookies for team members

RCSB.org/pdb50



Congressional Record

United States of America

PROCEEDINGS AND DEBATES OF THE 117th CONGRESS, FIRST SESSION

House of Representatives

EXTENSION OF REMARKS IN RECOGNITION OF THE 50TH ANNIVERSARY OF THE PROTEIN DATA BANK IN THE UNITED STATES

HON. FRANK PALLONE JR.

July, 19, 2021

Madam Speaker, I rise today in recognition of the 50th anniversary of the Protein Data Bank (PDB) in the United States and its outsized impact on the scientific community. Established at Brookhaven National Laboratory in 1971, the PDB was created because researchers were unable to share the valuable information they collected due to the lack of advanced computer systems at that time. The PDB solved this problem by serving as a depository for three-dimensional structures of large, biological molecules—such as proteins, DNA, and RNA—and offered the data free of charge or restrictions on usage to individual researchers across the world. Since 1998, I am proud to say the PDB has been co-managed by Rutgers, The State University of New Jersey. Today, the PDB archive in America hosts more than 180,000 structures of molecules found in all living organisms on the planet and is part of a worldwide network with other data centers in Europe and Asia.

The PDB's legacy as the first open-access digital data resource in biology and medicine has led to countless breakthroughs and has democratized access to data in a way that has tremendously benefited the public. Each year, millions of researchers, scientists, educators, students, medical professionals, patients, patient advocates as well as pharmaceutical and biotechnology companies around the word utilize the PDB for basic and applied research, STEM education, and the discovery of lifesaving medicines, particularly anti-viral and anti-cancer drugs. Examples of its profound impact include the more than two million structure data file downloads it generates per day; its contribution to over one million published research papers; and its enabling of research in many subject areas, ranging from Agriculture to Zoology.

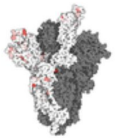
But this ambitious project would not have been possible without significant federal investment in scientific research. The National Science Foundation, the Department of Energy, and the National Institutes of Health all contribute funding to the PDB, which generates a return on investment 1,500 times more than federal dollars it receives. This kind of investment clearly demonstrates the importance of the government-university research partnership and why it is critical that we maintain and strengthen that partnership to bolster American innovation and competitiveness for years to come.

In conclusion, Madam Speaker, I offer my congratulations to the PDB in the United States as it celebrates its first 50 years as a leading global resource for experimental data central to scientific discovery, and I look forward to its contributions to science and our society over the next 50 years.

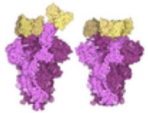
Congressman Frank Pallone, Jr., who represents New Jersey's 6th district that includes Rutgers University, noted the 50th anniversary of the PDB in the Congressional Record (HTML | PDF)

Recent SARS-CoV-2 Efforts

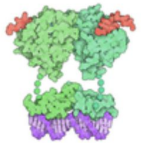
Molecules of the Month



December 2021
SARS-CoV-2 Spike Variants

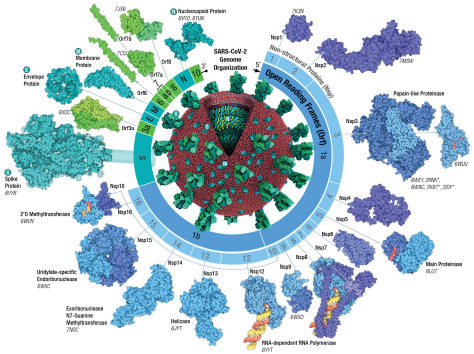


April 2021
SARS-CoV-2 Spike and Antibodies

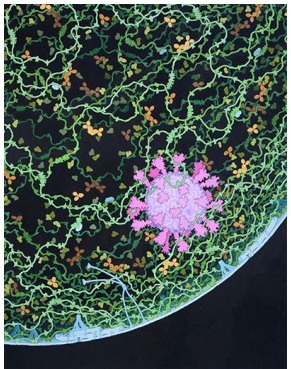


June 2021
Glucocorticoid Receptor and Dexamethasone

Genome and Proteins Flyer



2021 BioArt Award



Research Intensive Summer Experience at Rutgers

Clockwise, starting with upper left: Mary Agnes Balogun (Morgan State), Mickayla Bacorn (University of Maryland, Baltimore County), Cassandra Olivas (California State University Stanislaus), and Amy Wu Wu (University of Puerto Rico-Mayaguez)

Studied protease evolution across known *Coronaviridae*; 2 awarded poster prizes at American Crystallographic Association Meeting; 1 at ABRCMS; manuscripts in

Recent and Related Publications

Design and proof of concept for targeted phage-based COVID-19 vaccination strategies with a streamlined cold-free supply chain

Daniela I. Staquicini[a,b,1], Fenny H. F. Tang[a,b,1], Christopher Markosian[a,b,1], Virginia J. Yao[a,b,2], Fernanda I. Staquicini[a,b,3], Esteban Dodero-Rojas[c], Vinicius G. Contessoto[c,d], Deodate Davis[a,b], Paul O'Brien[a,b], Nazia Habib[a,b], Tracey L. Smith[a,b], Natalie Bruiners[a,b], Richard L. Sidman[f], Maria L. Gennaro[g], Edmund C. Lattime[a,h], Steven K. Libutti[a,b], Paul C. Whitford[i], Stephen K. Burley[a,j,k], José N. Onuchic[c,m,n,o,4,5], Wadih Arap[a,b,4,5], and Renata Pasqualini[a,b,4,5]

[a]Rutgers Cancer Institute of New Jersey, Newark, NJ 07101; [b]Division of Cancer Biology, Department of Radiation Oncology, Rutgers New Jersey Medical School, Newark, NJ 07103; [c]Center for Theoretical Biological Physics, Rice University, Houston, TX 77005; [d]Department of Physics, Institute of Biosciences, Humanities and Exact Sciences, São Paulo State University, São José do Rio Preto, SP 15054, Brazil; [e]Public Health Research Institute, Rutgers New Jersey Medical School, Newark, NJ 07103; [f]Department of Neurology, Harvard Medical School, Boston, MA 02115; [g]Rutgers Cancer Institute of New Jersey, New Brunswick, NJ 08901; [h]Department of Surgery, Rutgers Robert Wood Johnson Medical School, New Brunswick, NJ 08901; [i]Department of Physics and Center for Theoretical Biological Physics, Northeastern University, Boston, MA 02115; [j]RCSB Protein Data Bank and Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, NJ 08854; [k]Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, NJ 08854; [l]RCSB Protein Data Bank, San Diego Supercomputer Center and Skaggs School of Pharmacy & Pharmaceutical Sciences, University of California San Diego, La Jolla, CA 92067; [m]Department of Biosciences, Rice University, Houston, TX 77005; [n]Department of Chemistry, Rice University, Houston, TX 77005; [o]Department of Physics and Astronomy, Rice University, Houston, TX 77005; and [p]Division of Hematology/Oncology, Department of Medicine, Rutgers New Jersey Medical School, Newark, NJ 07103

Contributed by José N. Onuchic, May 3, 2021 (sent for review March 29, 2021); reviewed by Luiz V. Rizzo and Angel E. García

Development of effective vaccines against coronavirus disease 2019 (COVID-19) is a global imperative. Rapid immunization of the entire human population against a widespread, continually evolving, and highly pathogenic virus is an unprecedented challenge, and different vaccine approaches are being pursued. Engineered filamentous bacteriophage (phage) particles have unique potential in vaccine development due to their inherent immunogenicity, genetic plasticity, stability, cost-effectiveness for large-scale production, and proven safety profile in humans. Herein we report the development and initial evaluation of two targeted phage-based vaccination approaches against SARS-CoV-2: dual ligand peptide-targeted phage and adeno-associated virus/phage (AAVP) particles. For peptide-targeted phage, we performed structure-guided antigen design to select six solvent-exposed epitopes of the SARS-CoV-2 spike (S) protein. One of these epitopes displayed on the major capsid protein pVIII of phage induced a specific and sustained humoral response when injected in mice. These phage were further engineered to simultaneously display the peptide CAKSMGDIVC on the minor capsid protein pIII to enable their transport from the lung epithelium into the systemic circulation. Aerosolization of these "dual-display" phage into the lungs of mice generated a systemic and specific antibody response. In the second approach, targeted AAVP particles were engineered to deliver the entire S protein gene under the control of a constitutive CMV promoter. This induced tissue-specific transgene expression, stimulating a systemic S protein-specific antibody response in mice. With these proof-of-concept preclinical experiments, we show that both targeted phage- and AAVP-based particles serve as robust yet versatile platforms that can promptly yield COVID-19 vaccine prototypes for translational development.

AAVP | COVID-19 | gene delivery | phage display | SARS-CoV-2

Since early 2021, the World Health Organization (WHO) has estimated that nearly 3 million deaths in 223 countries/territories have been caused by complications of coronavirus disease 2019 (COVID-19). This unprecedented pandemic has prompted a worldwide collaborative effort to develop vaccines and antiviral therapies to control global spread. Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) is the third zoonotic coronavirus to infect humans in less than 20 y (1, 2). Previous coronavirus epidemics, such as severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS), foreshadowed the risk of emerging disease outbreaks and the imminent need for novel and versatile technologies for rapid manufacturing and large-scale distribution of vaccines and therapies as emergency countermeasures.

SARS-CoV-2 is a single-stranded enveloped RNA virus with four main structural proteins. The spike (S) protein mediates both host cell recognition and membrane fusion and is pivotal for viral entry. The S protein, composed of the S1 and S2 subunits, is displayed as a trimer on the surface of the viral particle (3). Within

Author contributions: D.I.S., F.H.F.T., C.M., V.J.Y., F.I.S., E.D.-R., V.G.C., D.D., P.O., N.H., T.L.S., N.B., R.L.S., M.L.G., E.C.L., S.K.L., P.C.W., S.K.B., J.N.O., W.A., and R.P. designed research; D.I.S., F.H.F.T., C.M., V.J.Y., F.I.S., E.D.-R., V.G.C., D.D., P.O., N.H., T.L.S., N.B., M.L.G., E.C.L., P.C.W., W.A., and R.P. performed research; S.K.L., S.K.B., and J.N.O. contributed new reagents/analytic tools; D.I.S., F.H.F.T., C.M., V.J.Y., F.I.S., E.D.-R., V.G.C., D.D., P.O., N.H., T.L.S., N.B., R.L.S., M.L.G., E.C.L., P.C.W., J.N.O., W.A., and R.P. analyzed data; D.I.S., F.H.F.T., C.M., V.J.Y., F.I.S., E.D.-R., V.G.C., P.C.W., J.N.O., W.A., and R.P. wrote the paper.

Reviewers: L.V.R., Hospital Israelita Albert Einstein; and A.E.G., Los Alamos National Laboratory.

Competing interest statement: D.I.S., W.A., and R.P. are listed as inventors on a patent application related to this technology (International Patent Application no. PCT/US2020/053758, entitled "Targeted Pulmonary Delivery Compositions and Methods Using Same"). Provisional patent application nos. 63/048, 279, and 63/161,136, entitled "Enhancing Immune Responses Through Targeted Antigen Expression," have also been filed on the technology and intellectual property reported here. PhageNova Bio has licensed these intellectual properties and D.I.S., F.H.F.T., C.M., V.J.Y., T.L.S., S.K.L., W.A., and R.P. may be entitled to standard royalties. R.P., S.K.L., and W.A. are founders and equity stockholders of PhageNova Bio. S.K.L. is a board member and R.P. is chief scientific officer and a paid consultant of PhageNova Bio. V.J.Y. is currently a full-time employee of PhageNova Bio. R.P. and W.A. are founders and equity shareholders of MBrace Therapeutics; R.P. serves as the chief scientific officer and W.A. is a member of the scientific advisory board at MBrace Therapeutics. F.I.S. is currently a full-time employee of MBrace Therapeutics. These arrangements are managed in accordance with the established institutional conflict-of-interest policies of Rutgers, The State University of New Jersey.

[1]D.I.S., F.H.F.T., and C.M. contributed equally to this work.

[2]Present address: PhageNova Bio, T. O. Daniel Research Incubator & Collaboration Center, Summit, NJ 07901.

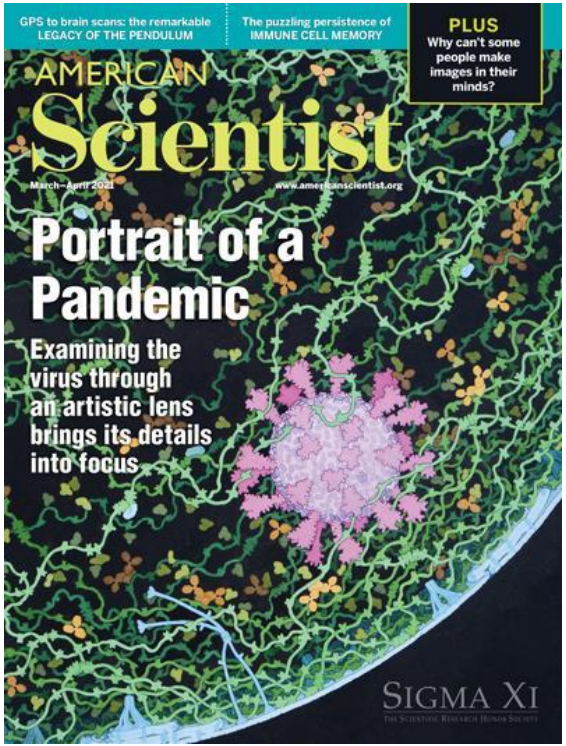
[3]Present address: MBrace Therapeutics, T. O. Daniel Research Incubator & Collaboration Center, Summit, NJ 07901.

[4]J.N.O., W.A., and R.P. contributed equally to this work.

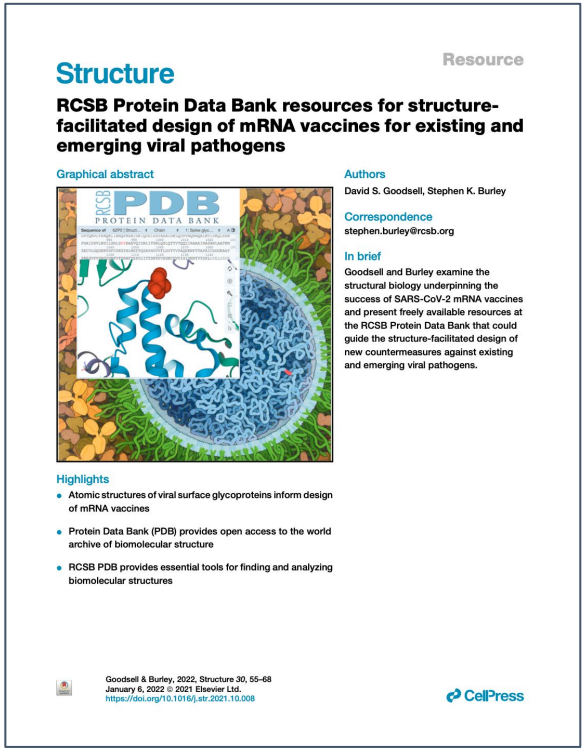
[5]To whom correspondence may be addressed. Email: jonuchic@rice.edu, or renata.pasqualini@rutgers.edu, or wadih.arap@rutgers.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2105739118/-/DCSupplemental.

Published July 7, 2021.



(2021) *American Scientist* 109: 88-93



Resource

Structure

RCSB Protein Data Bank resources for structure-facilitated design of mRNA vaccines for existing and emerging viral pathogens

Graphical abstract

Authors

David S. Goodsell, Stephen K. Burley

Correspondence

stephen.burley@rcsb.org

In brief

Goodsell and Burley examine the structural biology underpinning the success of SARS-CoV-2 mRNA vaccines and present freely available resources at the RCSB Protein Data Bank that could guide the structure-facilitated design of new countermeasures against existing and emerging viral pathogens.

Highlights

- Atomic structures of viral surface glycoproteins inform design of mRNA vaccines
- Protein Data Bank (PDB) provides open access to the world archive of biomolecular structure
- RCSB PDB provides essential tools for finding and analyzing biomolecular structures

Goodsell & Burley, 2022, Structure 30, 55–68
January 6, 2022 © 2021 Elsevier Ltd.
https://doi.org/10.1016/j.str.2021.10.008

CellPress

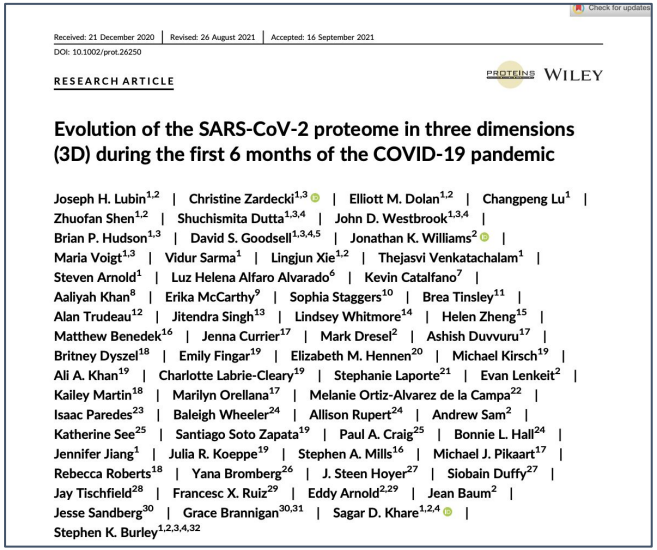
(2021) *Proceedings of the National Academy of Sciences* 118 (30) e21057391181 doi: 10.1073/pnas.2105739118

(2022) *Structure* 30: P55-68.e2 doi: 10.1016/j.str.2021.10.008

2020 Summer of the "Coronaverse" Published

One-week Boot Camp: 35 student researchers studied sequence-structure evolution of the main protease

12 continued for 5 weeks to explore the other structures (3 supported by NSF REU)



Received: 21 December 2020 | Revised: 26 August 2021 | Accepted: 16 September 2021
DOI: 10.1002/prot.26250

RESEARCH ARTICLE

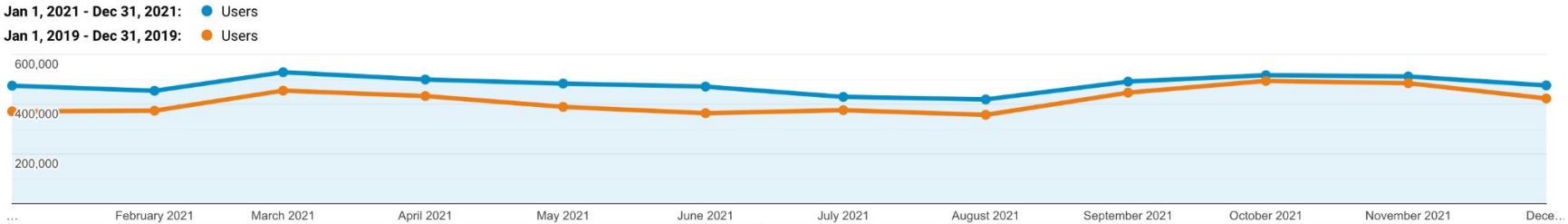
PROTEINS WILEY

Evolution of the SARS-CoV-2 proteome in three dimensions (3D) during the first 6 months of the COVID-19 pandemic

Joseph H. Lubin[1,2] | Christine Zardecki[1,3] | Elliott M. Dolan[1,2] | Changpeng Lu[1] | Zhuofan Shen[1,2] | Shuchismita Dutta[1,3,4] | John D. Westbrook[1,3,4] | Brian P. Hudson[1,3] | David S. Goodsell[1,3,4,5] | Jonathan K. Williams[2] | Maria Voigt[1,3] | Vidur Sarma[1] | Lingjun Xie[1,2] | Thejasvi Venkatachalam[1] | Steven Arnold[1] | Luz Helena Alfaro Alvarado[6] | Kevin Catalfano[7] | Aaliyah Khan[8] | Erika McCarthy[9] | Sophia Staggers[10] | Brea Tinsley[11] | Alan Trudeau[12] | Jitendra Singh[13] | Lindsey Whitmore[14] | Helen Zheng[15] | Matthew Benedek[16] | Jenna Currier[17] | Mark Dresel[2] | Ashish Duvvuru[17] | Britney Dyszel[18] | Emily Fingar[19] | Elizabeth M. Hennen[20] | Michael Kirsch[19] | Ali A. Khan[19] | Charlotte Labrie-Cleary[19] | Stephanie Laporte[21] | Evan Lenkeit[2] | Kailey Martin[18] | Marilyn Orellana[17] | Melanie Ortiz-Alvarez de la Campa[22] | Isaac Paredes[23] | Baleigh Wheeler[24] | Allison Rupert[24] | Andrew Sam[2] | Katherine See[25] | Santiago Soto Zapata[19] | Paul A. Craig[25] | Bonnie L. Hall[24] | Jennifer Jiang[1] | Julia R. Koeppe[19] | Stephen A. Mills[16] | Michael J. Pikaart[17] | Rebecca Roberts[18] | Yana Bromberg[26] | J. Steen Hoyer[27] | Siobain Duffy[27] | Jay Tischfield[28] | Francesc X. Ruiz[29] | Eddy Arnold[2,29] | Jean Baum[2] | Jesse Sandberg[30] | Grace Brannigan[30,31] | Sagar D. Khare[1,2,4] | Stephen K. Burley[1,2,3,4,32]

Proteins: Structure, Function, and Bioinformatics (2021) doi: 10.1002/prot.26250

PDB-101 Traffic: Increase over 2019 Visits



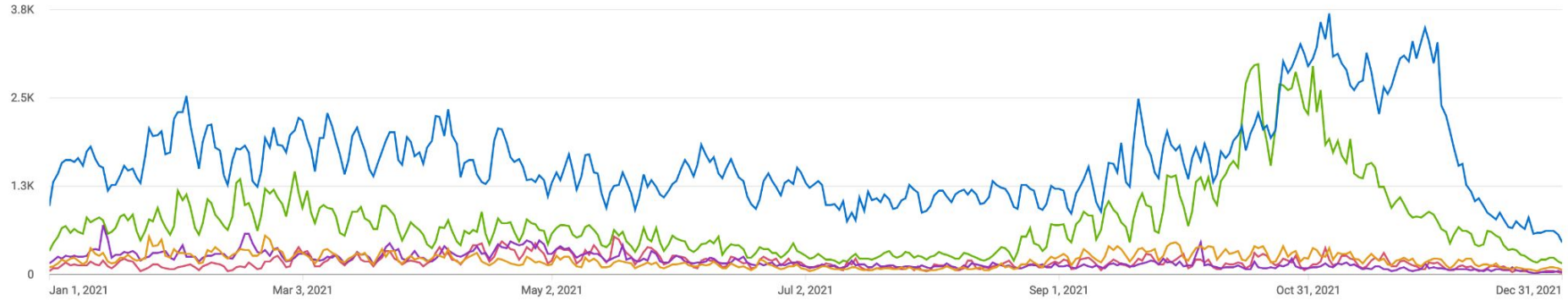
Users	New Users	Sessions	Number of Sessions per User	Pageviews	Pages / Session	Avg. Session Duration
4.08%	4.72%	13.08%	8.65%	16.24%	2.80%	5.90%
692,526 vs 665,404	690,283 vs 659,158	1,050,599 vs 929,057	1.52 vs 1.40	2,310,472 vs 1,987,604	2.20 vs 2.14	00:02:08 vs 00:02:01

	Google Analytics: Unique Users	Internal Tracking: Unique IP Addresses	Google Analytics: Page Views	Internal: Page Views
2021	**692,526**	**906,103**	**2,310,472**	**4,104,583**
2020	853,734	1,001,957	2,608,213	4,140,229
2019	665,958	749,005	1,989,074	3,632,911
2018	594,073	760,797	1,816,972	3,496,896

We assume that Google Analytics filters out some robotic access

PDB-101 Traffic 2021 vs 2020

RCSB PDB YouTube: 1.3M views in 2021



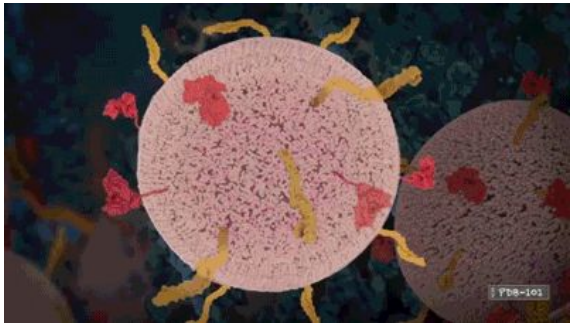
Video	Views ↓	Watch time (hours)	Subscribers	Impressions	Impressions click-through rate
Total	**1,379,974**	**52,101.2**	**15,502**	**24,276,781**	**2.5%**
What is a Protein? (from PDB-101)	591,039 42.8%	25,403.9 48.8%	9,230 59.5%	13,685,178	2.1%
How Enzymes Work (from PDB-101)	278,466 20.2%	9,833.1 18.9%	3,025 19.5%	5,023,291	2.2%
What is a Protein? Learn about the 3D shape and function of macro...	74,181 5.4%	2,281.8 4.4%	284 1.8%	315,442	4.8%
Fighting Coronavirus with Soap (from PDB-101)	70,006 5.1%	1,432.5 2.8%	158 1.0%	836,378	3.4%
¿Qué es una proteína? (Spanish)	62,999 4.6%	3,904.4 7.5%	477 3.1%	365,287	6.5%
Penicillin and Antibiotic Resistance (from PDB-101)	33,408 2.4%	1,566.3 3.0%	202 1.3%	513,871	3.4%
SARS-CoV-2 Life Cycle (Summer 2020)	23,608 1.7%	643.4 1.2%	59 0.4%	287,913	5.2%

PDB-101 in 2021: Beyond Covid-19 and PDB50

- 2022 3DEM calendar
- 9 Goodsell Molecular Landscapes
- Videos
  - Opioids and Pain Signaling
  - Experimental methods
- Health Focus: Drugs & Brain
  - HS Video Challenge
  - Undergraduate course
- Guide to Exploring Carbohydrates
- 3D Printed Model and Lesson Plan: Serum Albumin
- *Molecule of the Month* articles

Casein Micelle and Fat Globule in Milk

Cellulose Synthase

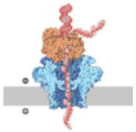


Opioids and Pain Signaling; >7K YouTube views

2021 Boot Camp: Scientific Communication

RCSB PDB and the Rutgers Institute for Quantitative Biomedicine hosted virtual boot camps for 20 undergraduates/ graduates with a focus on *Science Communication in Biology and Medicine*.

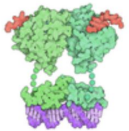
Students developed writing and 3D molecular visualization skills by collaborating on articles for the *Molecule of the Month* series at PDB-101.



September 2021
DNA-Sequencing Nanopores

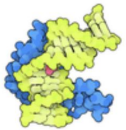
August 2021
Ribonuclease P

July 2021
Designed Proteins and Citizen Science



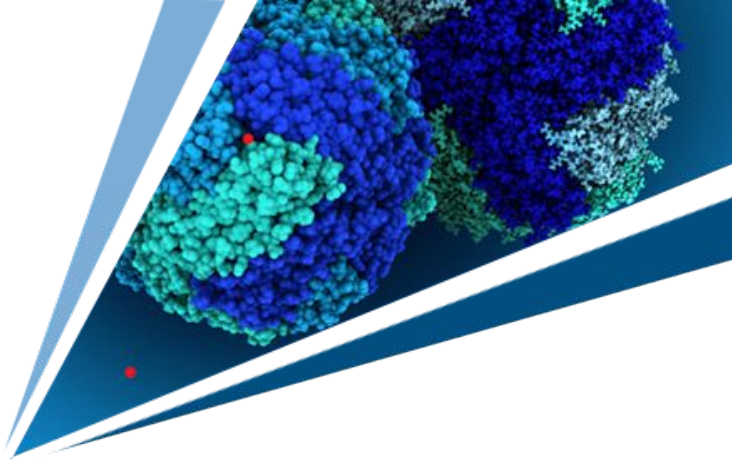
June 2021
Glucocorticoid Receptor and Dexamethasone

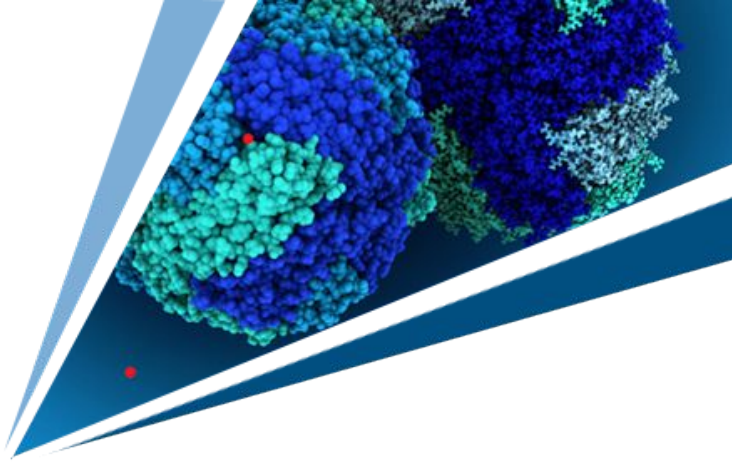
May 2021
Fetal Hemoglobin



March 2021
Cisplatin and DNA

Questions and Comments?





Operations

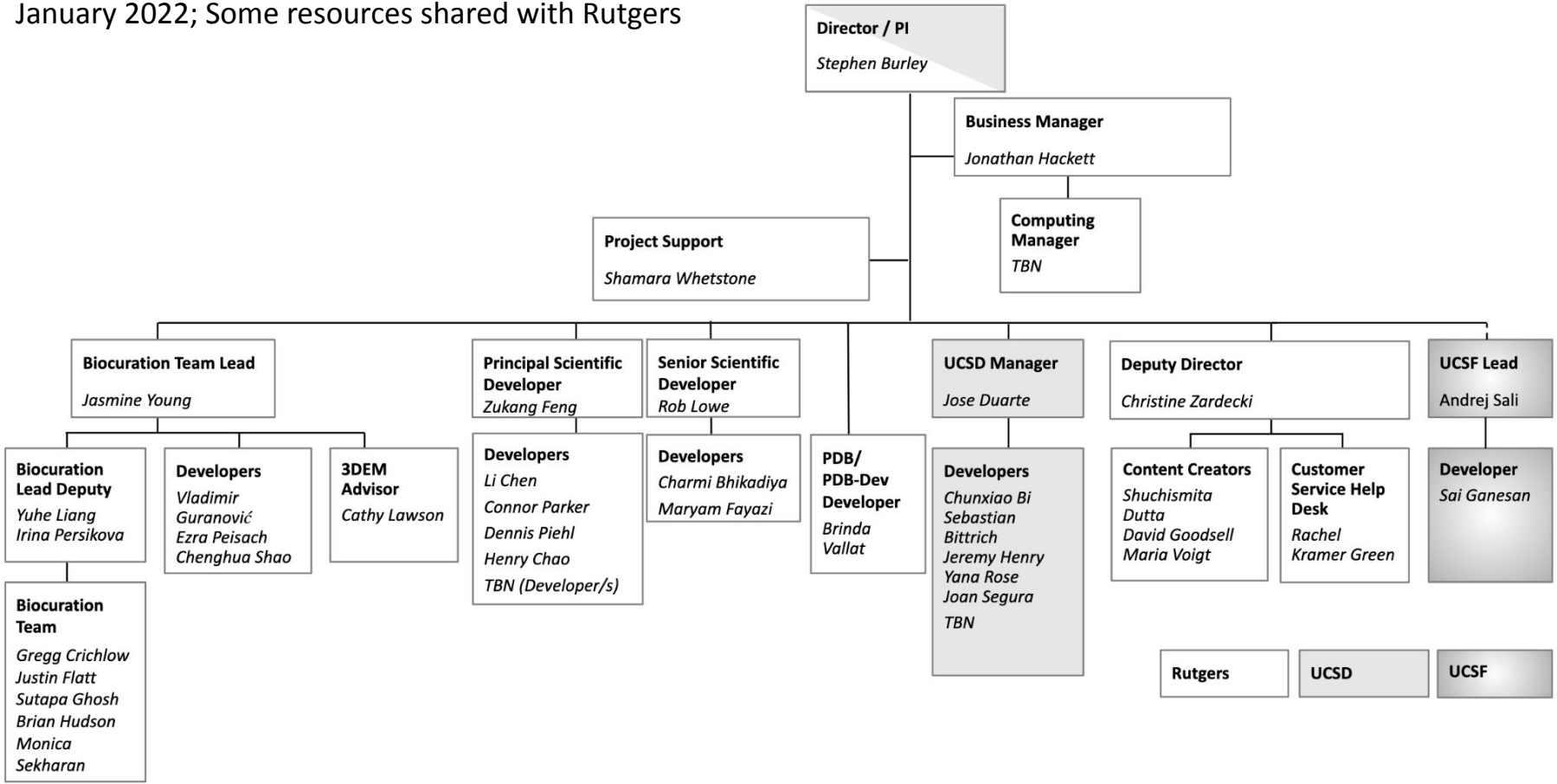
Stephen K. Burley

John Westbrook (1957-2021)

- Outpouring of support from the global community
- Data & Software Architect Lead responsibilities have been divided among team members and services
  - Main projects have been split across Deposition/Biocuration/ Archiving and RCSB.org
  - Zukang now managing direct reports
  - Yana now project manager for RCSB.org
- Will re-examine the organization in July 2022 to inform recruiting

RCSB PDB Line Management

January 2022; Some resources shared with Rutgers

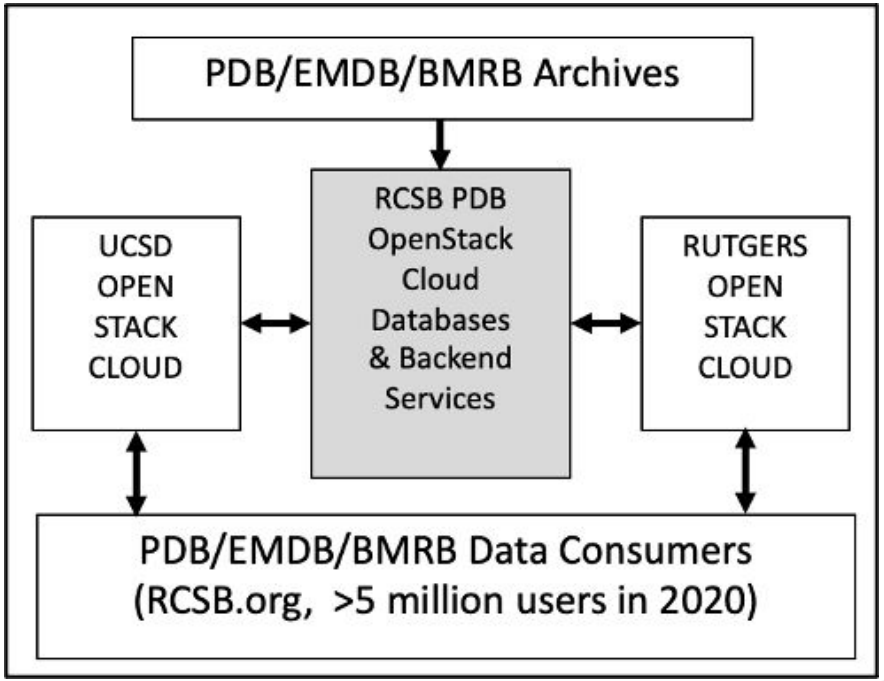


NSF Equipment Supplement Impact

- Urgently required expansion of server and storage infrastructure now installed on both coasts
  - 50% increase in Server CPUs
  - 150% increase in Server Memory
  - 200% increase in Solid State Disk
  - 100% increase in Hard Disk

- Improved overall RCSB.org user experience (reduced latency, *etc*.)

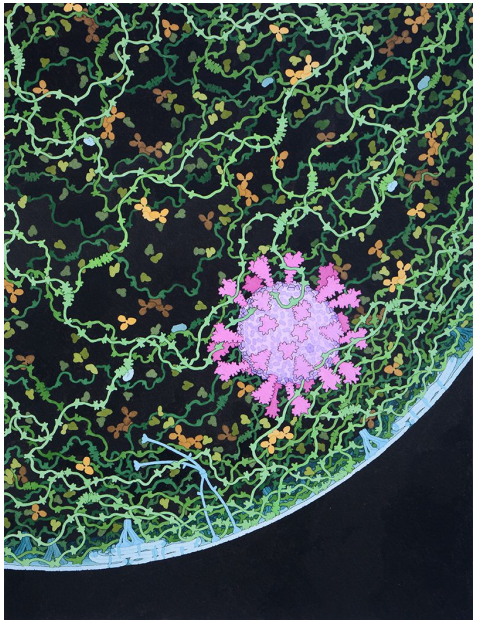
- Expected growth can be accommodated through 2024-2025 for current RCSB.org services



CAREER OPPORTUNITIES for
SCIENTIFIC SOFTWARE DEVELOPERS

Join the multidisciplinary RCSB Protein Data Bank Team in Piscataway, NJ; San Diego, CA; or San Francisco, CA

Develop innovative analysis, integration, query, and visualization tools for 3D biomolecular structures at **RCSB.org** to help accelerate research and training in biology, medicine, and related disciplines.



Open Positions:
- Scientific Web Application Developer (Rutgers)
- Scientific/Bioinformatics Software Developer (Rutgers)
- Scientific Software Developer (UCSD)
- Postdoctoral Researchers (UCSD/UCSF)

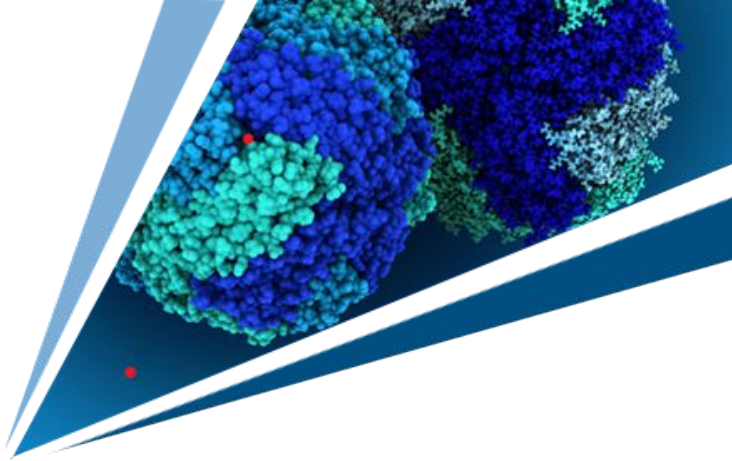
www.rcsb.org/pages/jobs

Recruitment in 2021

- Professional societies: American Crystallographic Association (ACA), American Society for Biochemistry and Molecular Biology (ASBMB), BioXFEL, International Society for Computational Biology (ISCB), Nature, Society for Advancement of Chicanos/Hispanics and Native Americans in Science (SACNAS)
- Informal postings: Black Women in Computational Biology Network, 500 Queer Scientists, Facebook, Twitter, LinkedIn, CCP4 & pdbl bulletin boards

Diversity, Equity, and Inclusion

- Working closely with Rutgers in new University-wide initiative to
  - Recruit, Retain, and Develop a Diverse Community
  - Promote Inclusive Scholarship and Teaching
  - Define Sustainable and Substantive Community Engagement
  - Build the Capacity of Leaders to Create Inclusive Climates
  - Develop an Institutional Infrastructure to Drive Change
- Working with UCSD Equity, Diversity, and Inclusion to
  - Address barriers to success for our underrepresented faculty, staff, and students
  - Further our efforts toward inclusive excellence
  - Foster a more welcoming and supportive campus climate
- Working with UCSF Office of Diversity and Outreach to offer all campus community members an equitable, inclusive, welcoming, secure, responsive, and affirming environment that fosters mutual respect, empathy and trust
- Partnering with Rutgers School of Graduate Studies in program for underrepresented minority undergraduate training and recruitment to Ph.D. programs (2015-present)
  - Research Intensive Summer Experience at Rutgers
  - Nationally-acclaimed summer research program for outstanding undergraduates from diverse backgrounds
  - RCSB PDB scholars (21) funded in part for past 7 years *via* NSF supplement



Strategic Initiatives

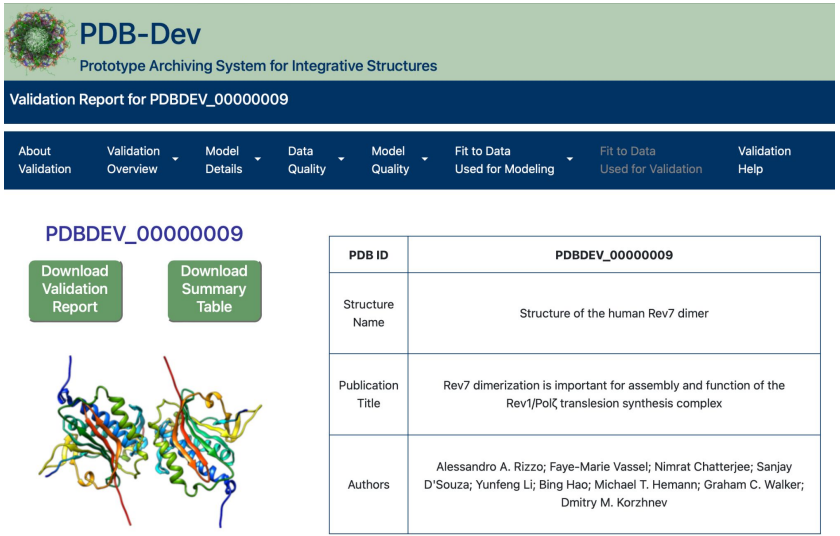
Stephen K. Burley
Andrej Sali

Cloud Migration Plans for 2022

- Set up on-premises server hardware for testing Docker Container deployment of RCSB.org backend services
- Convert RCSB.org backend services from current OpenStack configuration for deployment within Docker Containers
- Develop new orchestration tools for RCSB.org services, to interface with Kubernetes instead of OpenStack
- Deliver annual PDB core archive snapshots from AWS storage
- Deliver current PDB archive *via* ftp from AWS storage (updated weekly to accommodate newly released structures)

PDB-Dev and Integrative Structure Validation

- PDB-Dev deposition system has been improved and streamlined
- Designed to interoperate with existing wwPDB OneDep system
- Version 1.0 Integrative Methods Structure Validation Report, including a Summary Table
- Stepping-stone to Bayesian validation of structures of all types (w/ Jeff Hoch *et al*)
- mmCIF development (*cf*, Model Archive)



I/HM Validation project funded by RCSB PDB.

PDB-Dev Validation Report Example

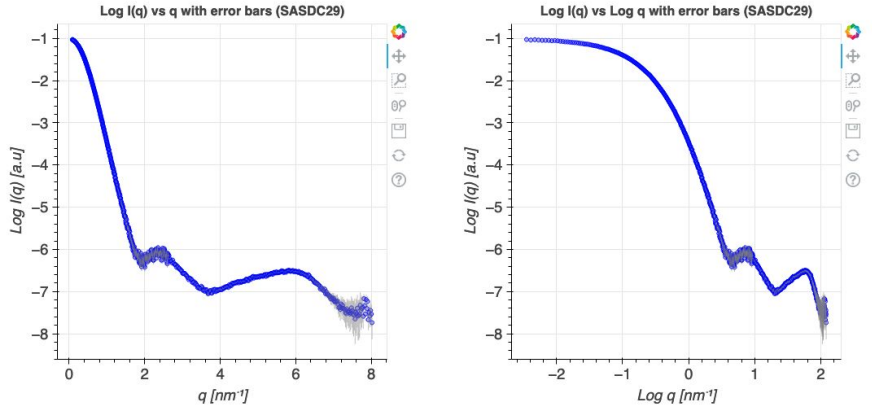
Data Quality

Data quality

SAS:Scattering profile

SAS data used in this integrative model was obtained from 1 deposited SASBDB entry (entries).

Scattering profile for SASDC29: data from solutions of biological macromolecules are presented as both log I(q) vs q and log I(q) vs log (q) based on SAS validation task force (SASvtf) recommendations. I(q) is the intensity (in arbitrary units) and q is the modulus of the scattering vector.



Fit of Model to Data

Fit of model to data used for modeling

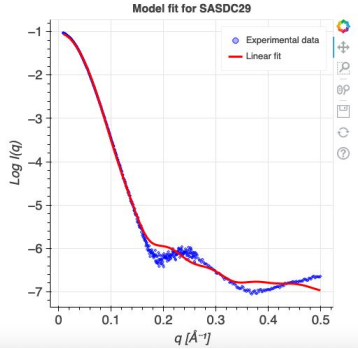
Fit of model(s) to SAS data

χ² goodness of fit analysis

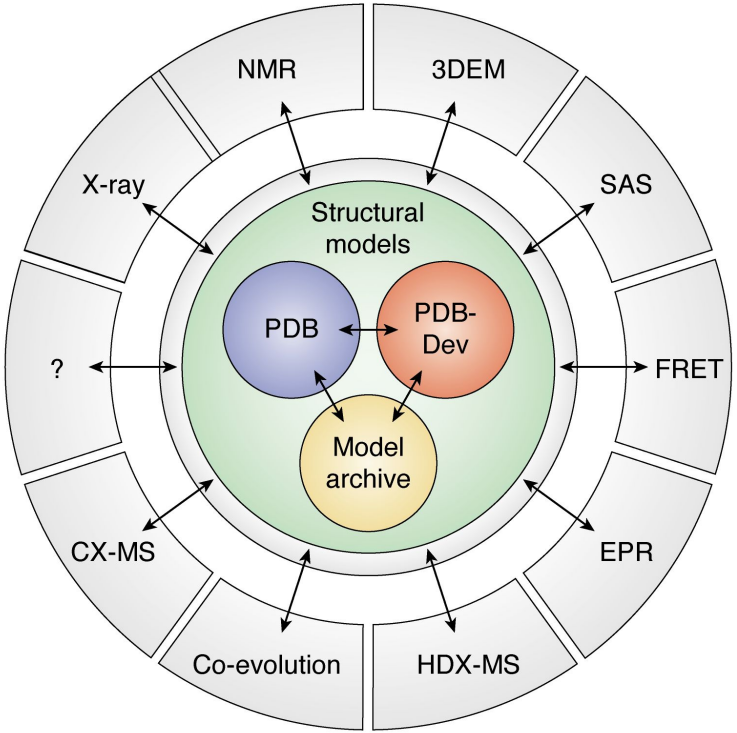
Model and fits displayed below were obtained from SASBDB. χ^2 values are a measure of fit of the model to data. A perfect fit has a χ^2 value of zero.

SASDB ID	Model	χ^2
SASDC29	1	25.13

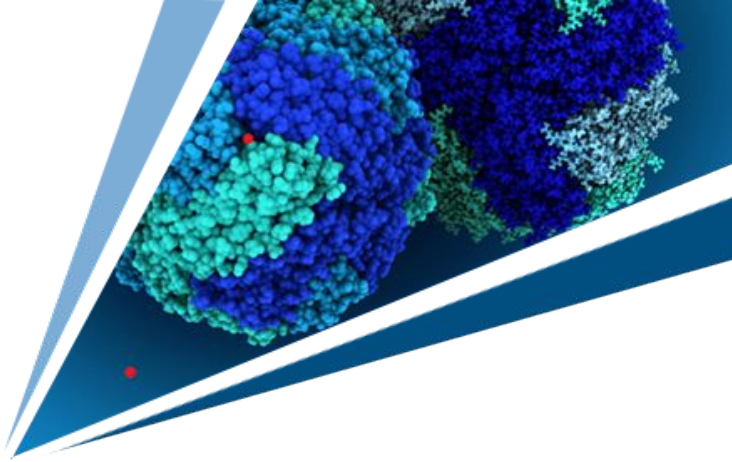
Model fit for SASDC29 (fit/model number 1): The experimental scattering curve (in blue) can be compared with the theoretical curve calculated from a model (in red).

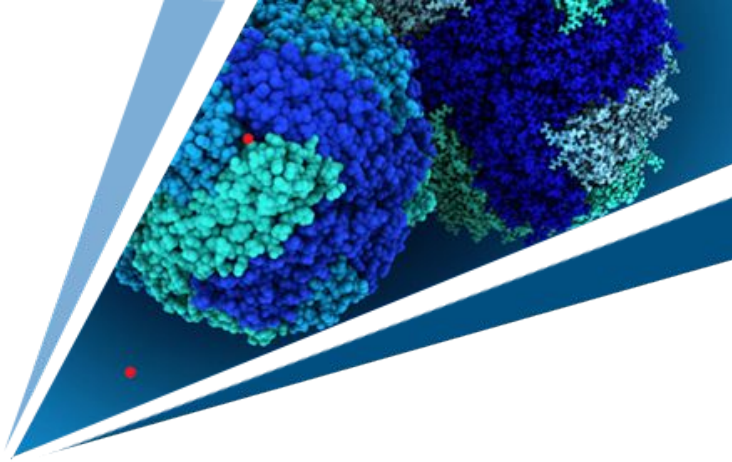


Vision for Global Federation of 3D Structure Models and Experimental Data



Questions and Comments?





Discussion

RCSB PDB Team

RCSB PDB PROTEIN DATA BANK

RCSB.ORG

info@rcsb.org

Funding

National Science Foundation (DBI-1832184), National Institute of General Medical Sciences, National Institute of Allergy and Infectious Disease, and National Cancer Institute (NIH R01GM133198), and the US Department of Energy (DE-SC0019749)

Management

RUTGERS

UC San Diego

SDSC SAN DIEGO SUPERCOMPUTER CENTER

UCSF University of California San Francisco

WORLDWIDE PDB PROTEIN DATA BANK

Member of the Worldwide Protein Data Bank (wwPDB; **wwpdb.org**)

Follow us

John D. Westbrook
In memoriam
1957-2021