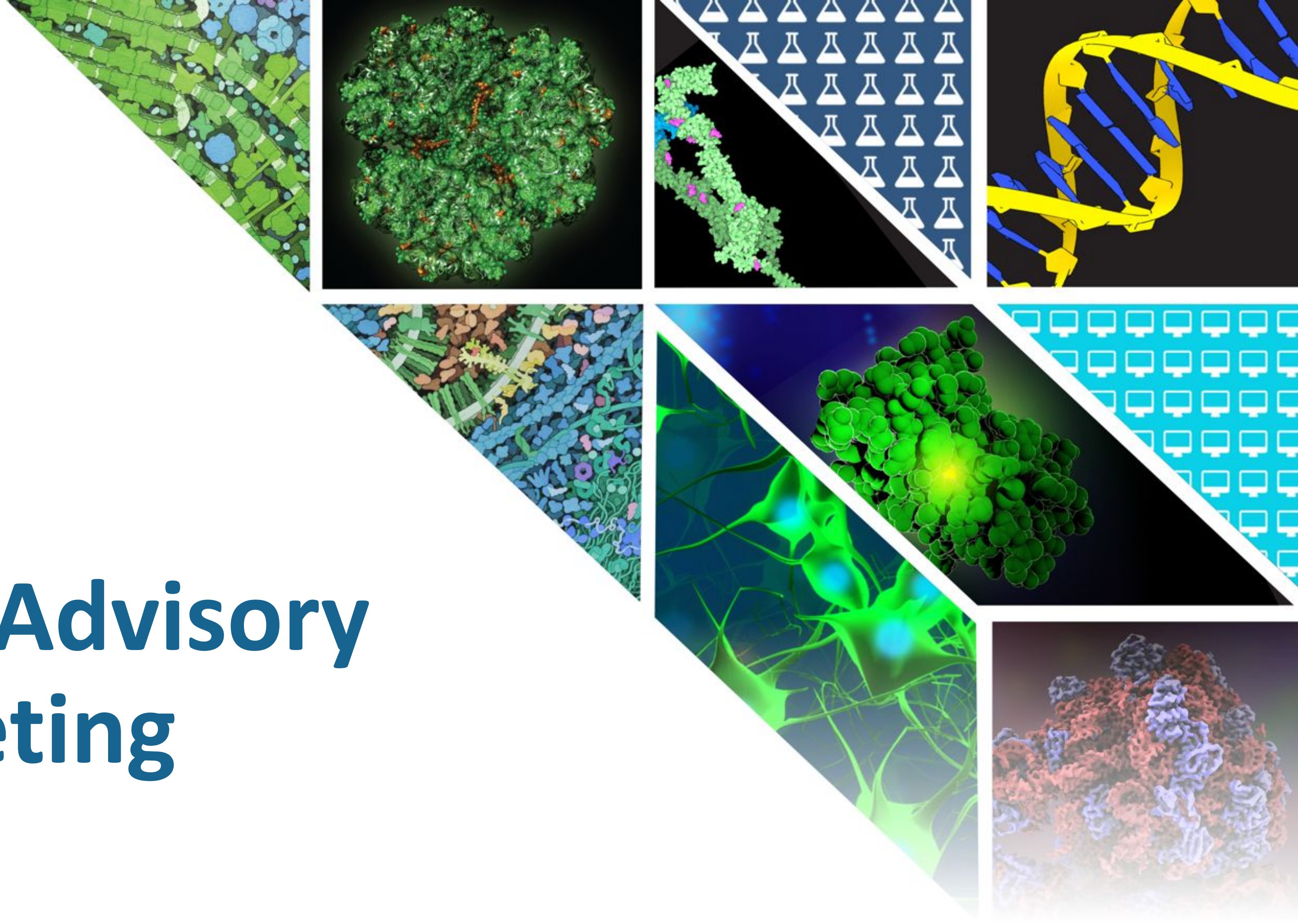
RCSB PDB
PROTEIN DATA BANK

rcsb.org

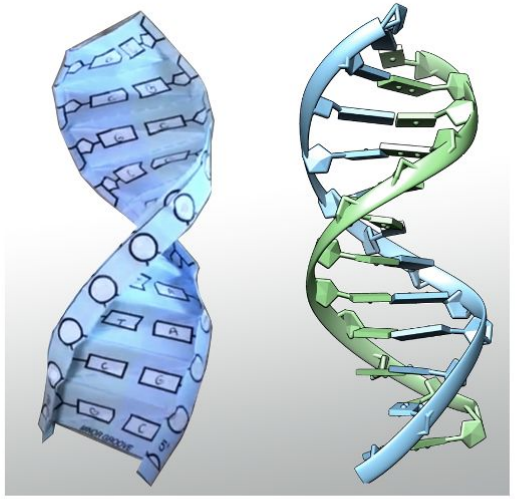


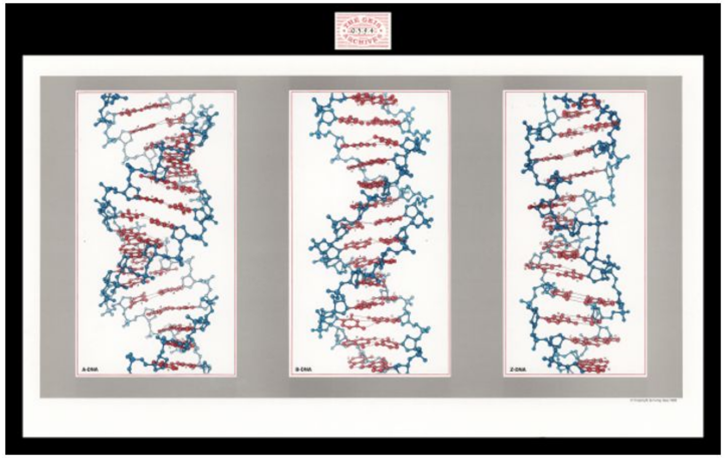
2024 RCSB PDB Advisory Committee Meeting

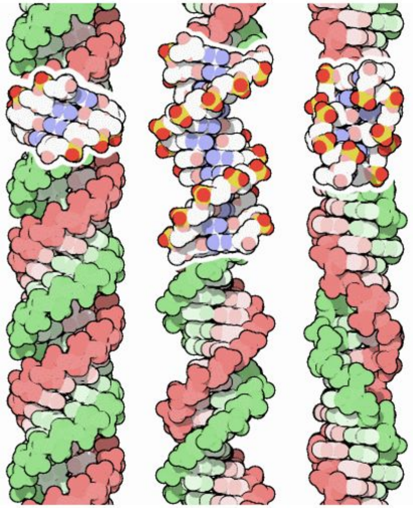
April 25, 2024

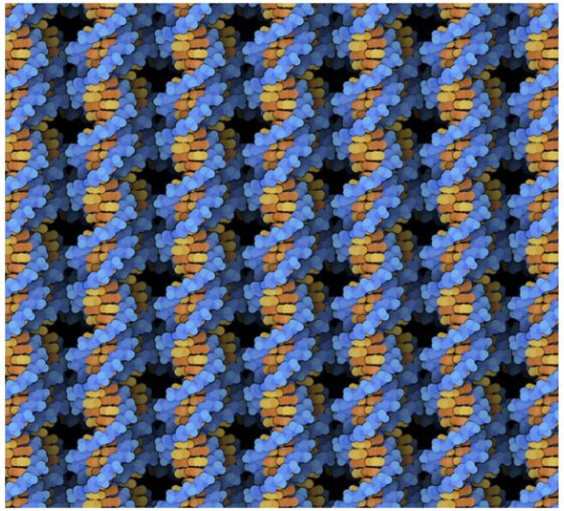
2:00pm-5:00pm EDT | 11:00am-2:00pm PDT

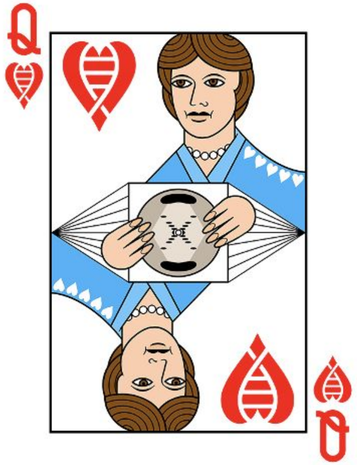
Happy DNA Day!











Agenda

Pacific	*Eastern*		
11:00am PT	**2:00pm ET**	**10’ Executive Session**	
11:10	2:10	Brief Welcome	Stephen Burley
11:15	2:15	Proposal Status and Updates	Stephen Burley, Henry Chao, Jasmine Young
11:30	2:30	PDBx/mmCIF Transition Update, S1-2 Roadmap Highlights	Jasmine Young
11:40	2:40	10’ Break	
11:50	2:50	Computed Structure Models (CSMs) at RCSB.org S3 Roadmap Highlights	Yana Rose
12:05	3:05	Recruiting Updates and Team Transitions	Stephen Burley
12:10	3:10	Questions for Committee	
		S4 Highlights	*if time allows*
12:30pm	**3:30**	**30’ Executive Session**	
1:00	4:00	60’ Discussion with Available Advisors and RCSB PDB	
2:00	5:00	Meeting ends	

Today’s Participants: Welcome

- **Advisory Committee**
  - *Confirmed*: Paul Adams, Wah Chiu, Kirk Clark, Roland Dunbrack, Paul Falkowski, Thomas Ferrin,
    Cathy Peishoff, Torsten Schwede, Lance Stewart, Kevin H. Gardner
  - *Unconfirmed*: Peter Andolfatto, Mandë Holford, Takita F. Sumter
  - *Absent*: Bridget Carragher, Robert B. Darnell, Sue Rhee
- **RCSB PDB Participants**
  - *Leadership*: Stephen K. Burley (Director/PI), Andrej Sali (UCSF Site Head)
  - *Operations Team Representatives*: Jose Duarte (Scientific Software Lead and UCSD Manager), Henry Chao (S0 Lead; IT Infrastructure), Zukang Feng (Principal Scientific Software Developer), Jasmine Young (S1-2 Lead; RCSB PDB Biocuration Team Lead & wwPDB Global Project Lead), Yana Rose (S3 Lead; Scientific Software Developer & Data Architect), Christine Zardecki (Associate Director; S4 Lead)

Background Information Slides

- Slides with an *italicized, light blue title* are provided as background information and can be presented and discussed at the meeting by request.

- Main slides appear with non-italicized blue titles

- Underlined text indicates an active link

2023 By The Numbers: Another Banner Year!

Scientific Support and User Engagement

- Maintained 99.9% availability of RCSB.org and APIs
- RCSB PDB help desk supported ~600 conversations with users
  - Additional OneDep and structure-related questions transferred to wwPDB Help Desk

S1: Deposition/Biocuration

- Record 17063 structures deposited and processed–PDB record (New record! Up from 16,344 in 2022)
  - 1,053 SARS-CoV-2 structures released (~4,000 available)
- 3623 new ligands and 30 new BIRD dictionary items

S2: Archive Management and Access

- PDB surpassed 200,000 structures on January 10, 2023
- Record ~3 billion data file downloads across the wwPDB
- PDB Certified as a Global Core Biodata Resource
- PDB chemical component IDs now issued in 5-character format

S3: Data Exploration

- Record 8.2 million unique RCSB.org clients (unique IP addresses, up from 7.2 million in 2022)
  - 63 million web page views
- 3.5 billion requests/interactions
  (e.g., data downloads, service usage, RCSB.org views)

S4: Training, Outreach, Education

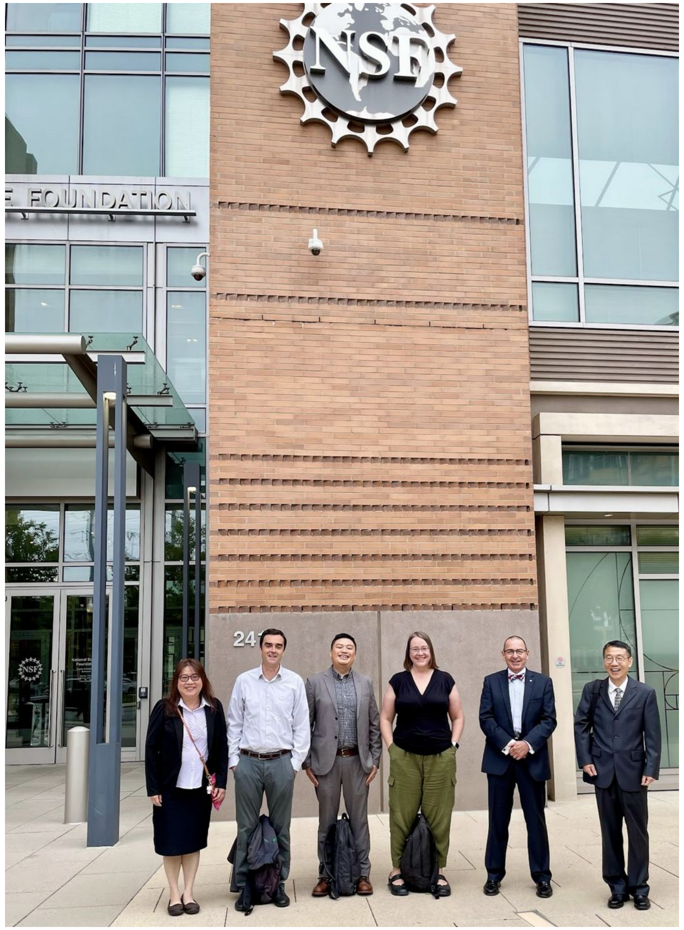
- ~548,000 PDB-101 users (down from ~663,000 in 2022)
- >1.8 million page views
- 850,000 YouTube Channel views
- Virtual “crash courses” and webinars
  - Understanding PDBx/mmCIF: ~450 participants
  - Python Scripting: ~160 participants
  - Leveraging RCSB PDB APIs: ~169 participants
  - Teaching Enzymology with the PDB: ~70 participants

Renewal Proposal Status and Updates

Stephen K. Burley

PDB Grant Renewal: Current Status

- Feb 28, 2023: Proposal submitted! (Thanks to RCSB PDB AC and others)
- Jun 22, 2023: Reverse Site Visit with Federal Funders and Review Team
- Sep 21, 2023: Follow-up Discussion with Federal Funders
- Nov 2, 2023: Review of RCSB PDB Response to Federal Funder feedback
- Jan-Feb 2024: Update to NSF, NIH, DOE
- April 10, 2024: NSF Notice of Award received



June 22, 2023 Review at NSF

Overview: Response to Review Critique

Cyberinfrastructure

- Designated as new Service 0 - IT Infrastructure (as Henry will highlight)
- Planned investments in owned hardware for S3 in Y1 and Y2 reduced significantly
- Will partner with an external high-performance computing provider (likely DOE-funded NERSC) to reduce reliance on owned-hardware for compute-intensive elements of prerelease data calculation process

PDB-Dev/PDB Unification

- Schedule accelerated
- PDB IDs will be allocated to extant PDB-Dev holdings and all newly deposited integrative/hybrid method structures starting in the second half of Y1 (versus Y3)

Computed Structure Models

- CSM caveats made more obvious for users who are not structural biologists (DONE; example)
- Existing documentation will be expanded and user training augmented to promote responsible CSM use

Data Volumes

- Refined hardware requirements to accommodate projected growth in data volumes

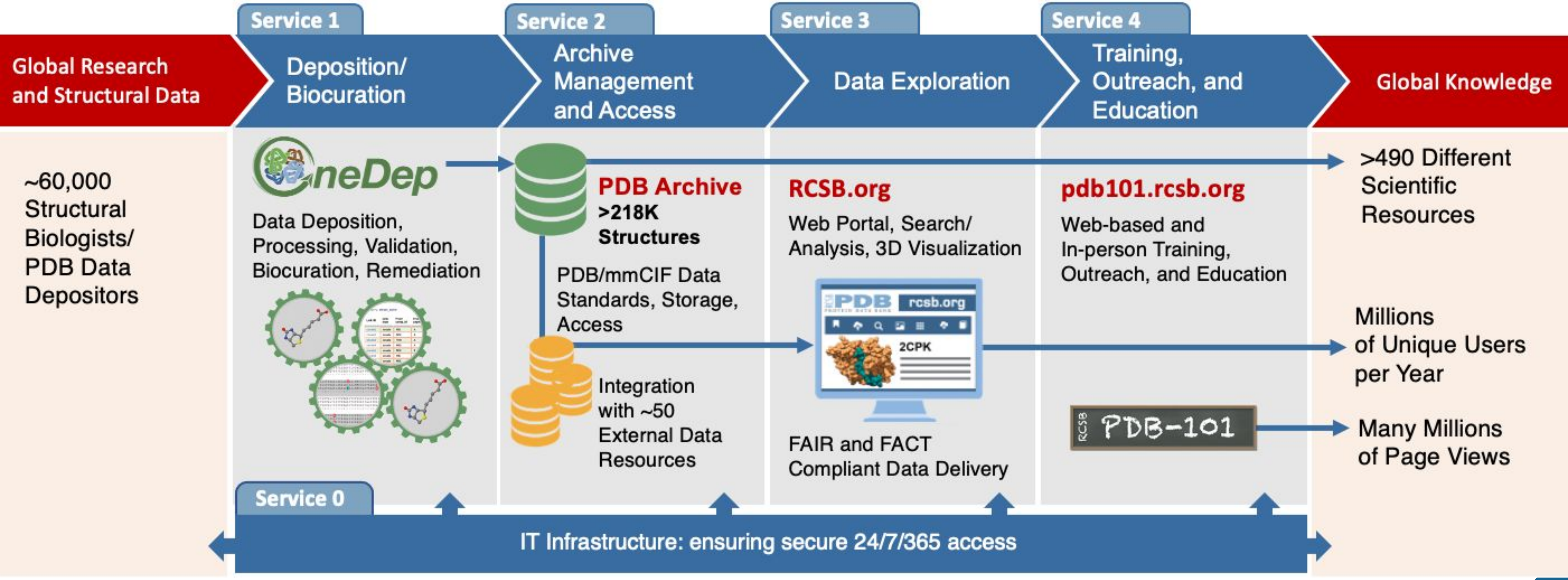
Agency Interactions

- RCSB PDB will continue to meet regularly with funding agencies to report progress and collect advice

Establishing Service 0: IT Infrastructure (Henry)

- Mission: Ensuring highly available, secure, and reliable access to IT resources for RCSB PDB by establishing and enforcing policies and processes around the management and operation of our IT infrastructure
- Team Members: Henry, Jeremy, Vladimir, and Aditya (starts June)
- Requirements/Activities
  - Ongoing participation and representation in Operations meetings
  - Roadmap for strategic and long term project planning
  - Ongoing KPI tracking/reviews
  - Strategic engagement with external Cyberinfrastructure organizations/resources/stakeholders

Introducing Service 0: IT Infrastructure



Global Research and Structural Data

Service 1

Deposition/ Biocuration

Service 2

Archive Management and Access

Service 3

Data Exploration

Service 4

Training, Outreach, and Education

Global Knowledge

~60,000 Structural Biologists/ PDB Data Depositors

OneDep

Data Deposition, Processing, Validation, Biocuration, Remediation

PDB Archive

>218K Structures

PDB/mmCIF Data Standards, Storage, Access

Integration with ~50 External Data Resources

RCSB.org

Web Portal, Search/ Analysis, 3D Visualization

FAIR and FACT Compliant Data Delivery

pdb101.rcsb.org

Web-based and In-person Training, Outreach, and Education

>490 Different Scientific Resources

Millions of Unique Users per Year

Many Millions of Page Views

Service 0

IT Infrastructure: ensuring secure 24/7/365 access

Service 0: Outcomes from the Reverse Site Visit

- Designation as an RCSB PDB Service to highlight importance of IT infrastructure work and increase visibility in reporting
- Cyberinfrastructure plans modified to utilize federal agency resources
  - Pivot from previously planned hardware purchases, to using no-cost high-performance computing resources from DOE National Energy Research Scientific Computing Center (NERSC) for our calculation intensive workloads
  - Limited purchase of new hardware to replace aging S1/S2 hardware and increase S3 computing resources to support future growth and user traffic
- Better positioning in the long run for
  - Anticipated scale of growth in data and user traffic
  - Planned development efforts
  - Access to more advanced resources and capabilities

Accelerating PDB-Dev Unification (Jasmine)

Goal: Accelerate PDB-Dev unification with the PDB archive to ensure capture and assessment of important IHM structure data

Strategy: Increase the originally proposed FTE effort in Y1-4 to enable

- Issuance of PDB IDs and DOIs for all existing and newly deposited IHM structures in Y1 (new)
- Creation of a parallel weekly release pipeline to pull IHM structures from PDB-Dev into PDB archive early in Y1 (originally planned for Y3)
- OneDep provides entry point to IHM structure deposition system by mid-Y1 (originally planned for Y5)
- Download of all IHM structures from the PDB archive (in parallel to PDB structures) by mid-Y1 (originally planned for Y5)

Unchanged

- Data Out data analysis and visualization on IHM structures (Y2-3)
- PDB-Dev website maintenance (Y1-5)
- Development of IHM validation tool, including Bayesian validation (Y1-5)

PDBx/mmCIF Transition Update Roadmap Highlights:

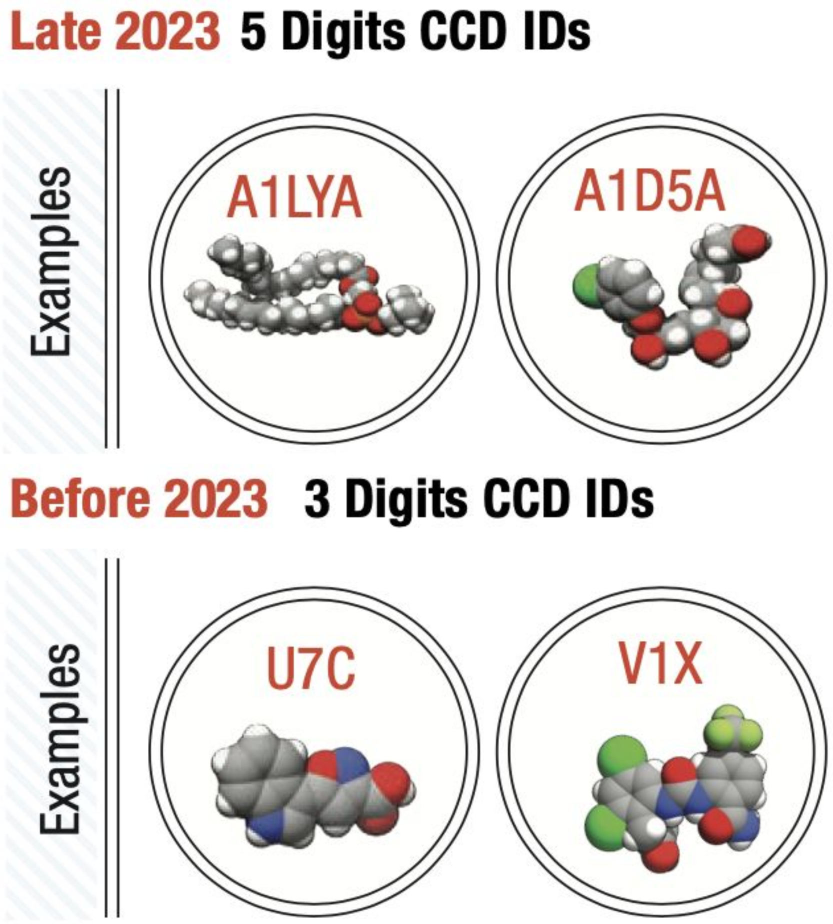
Service 1 Deposition/Biocuration
Service 2 Archive Management/Access

Jasmine Young

Chemical Component Dictionary (CCD) ID Extension

2-year project completed in 2023

- Regular ongoing community announcements (example) and presentations at meetings
- Users encouraged to utilize example files with 5-character IDs (provided via GitHub)
- Software upgrades to enable support (OneDep, PDB archive, partner websites)
- Three character IDs consumed December 2023
- 5-character CCD IDs in use in archive (example)
  - >700 new 5-character IDs already issued
  - N.B.: Files in PDB legacy format files cannot be provided for structures with extended CCD IDs



Transition to Extended PDB IDs and PDBx/mmCIF

Goals

- Help users prepare for full transition to PDBx/mmCIF format
- Increase community awareness of transition timeline and available resources

```
loop_
_database_2.database_id
_database_2.database_code
_database_2.pdbx_database_accession
_database_2.pdbx_DOI
PDB pdb_00001abc pdb_00001abc
10.2210/pdb_00001abc/pdb
```

5-year Plan for Transitioning to Extended PDB IDs and PDBx/mmCIF

- Create training materials for adoption of mmCIF and extended PDB ID (2024)
  - FAQs, software and documentation resources guide
- Register new PDB DOIs based on extended PDB IDs for the entire archive (2025)
- Establish “beta” PDB archive designed around extended PDB IDs (2026)
  - New PDB DOIs and extended PDB IDs available in the coordinate PDBx/mmCIF files
  - File directory organized at entry level (using same organization as the PDB Versioned Archive)
  - Directory and file naming use extended PDB ID
- OneDep and Data Out software re-tooling complete (early 2027)
- **“beta” PDB archive becomes PDB main archive (2027)**

Communication with Data Depositors, Data Consumers, and Scientific Journals/Editors throughout

S1-2 2024 Roadmap Highlights

- Deposition: Enhance EM deposition with more checks, improve file upload process with better tracking
- Validation: Upgrade 3rd party software (MolProbity, OpenBabel, Refmac), modularize and enable parallel calculations
- Biocuration: more automation, improve large structure processing performance
- Infrastructure: Data exchange among wwPDB partners via Globus service (replacing rsync protocol)
- Archive: PDB-Dev unification with PDB archive, PTM remediation, inclusion of extended PDB IDs

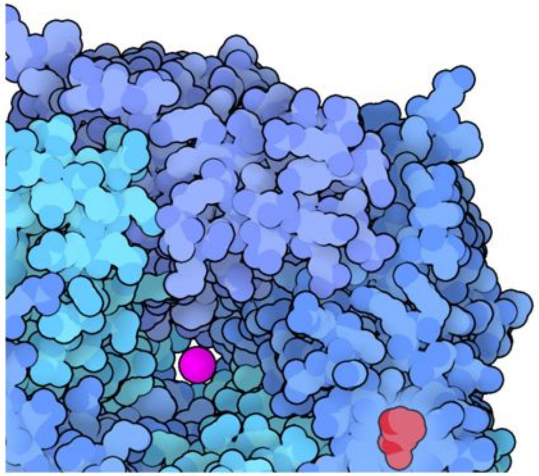
RCSB PDB PROTEIN DATA BANK

IQB

VIRTUAL WEBINAR

UNDERSTANDING PDB VALIDATION: WHICH EXPERIMENTAL STRUCTURES SHOULD I RELY ON?

Tuesday May 14th 2024
2-3pm Eastern | 11am-12pm Pacific



10’ Break

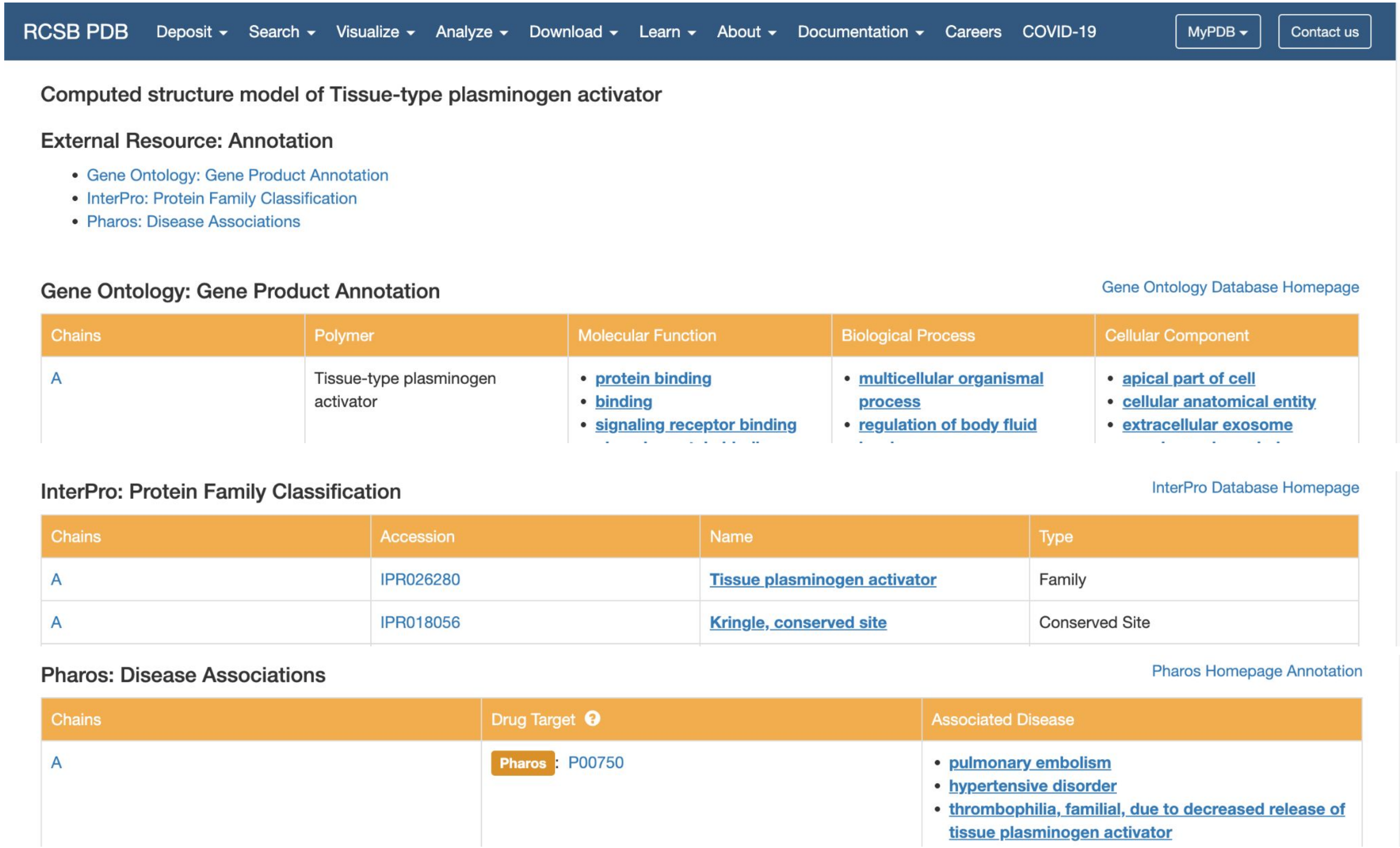
Computed Structure Models (CSMs) at RCSB.org; Roadmap Highlights: Service 3 Data Exploration

Yana Rose

Timeline of CSMs at RCSB.org

- **August 2022**: ~1 million CSMs from AlphaFold DB and ModelArchive for the entire human proteome and key organisms important in research and global health made accessible alongside experimental PDB Structures at RCSB.org
- **September 2022**: First Virtual Crash Course on CSMs (> 150 attendees)
- **February 2023**: updated AlphaFold DB models with latest release and ~68,000 ModelArchive CSMs added, providing coverage of model organisms important to funding agencies (*e.g.*, freshwater sponge, African swine fever virus, *Sphagnum divinum*, cancer interactome)
- **January 2024**: User survey on using CSMs at RCSB.org (results)
- **April 30, 2024**: Second Virtual Crash Course (~300 registered)
- **Spring 2024**: UXD review (in progress)

External Annotations Now Available for CSMs



Computed structure model of Tissue-type plasminogen activator

External Resource: Annotation

- Gene Ontology: Gene Product Annotation
- InterPro: Protein Family Classification
- Pharos: Disease Associations

Gene Ontology: Gene Product Annotation

Gene Ontology Database Homepage

Chains	Polymer	Molecular Function	Biological Process	Cellular Component
A	Tissue-type plasminogen activator	• protein binding • binding • signaling receptor binding	• multicellular organismal process • regulation of body fluid	• apical part of cell • cellular anatomical entity • extracellular exosome

InterPro: Protein Family Classification

InterPro Database Homepage

Chains	Accession	Name	Type
A	IPR026280	Tissue plasminogen activator	Family
A	IPR018056	Kringle, conserved site	Conserved Site

Pharos: Disease Associations

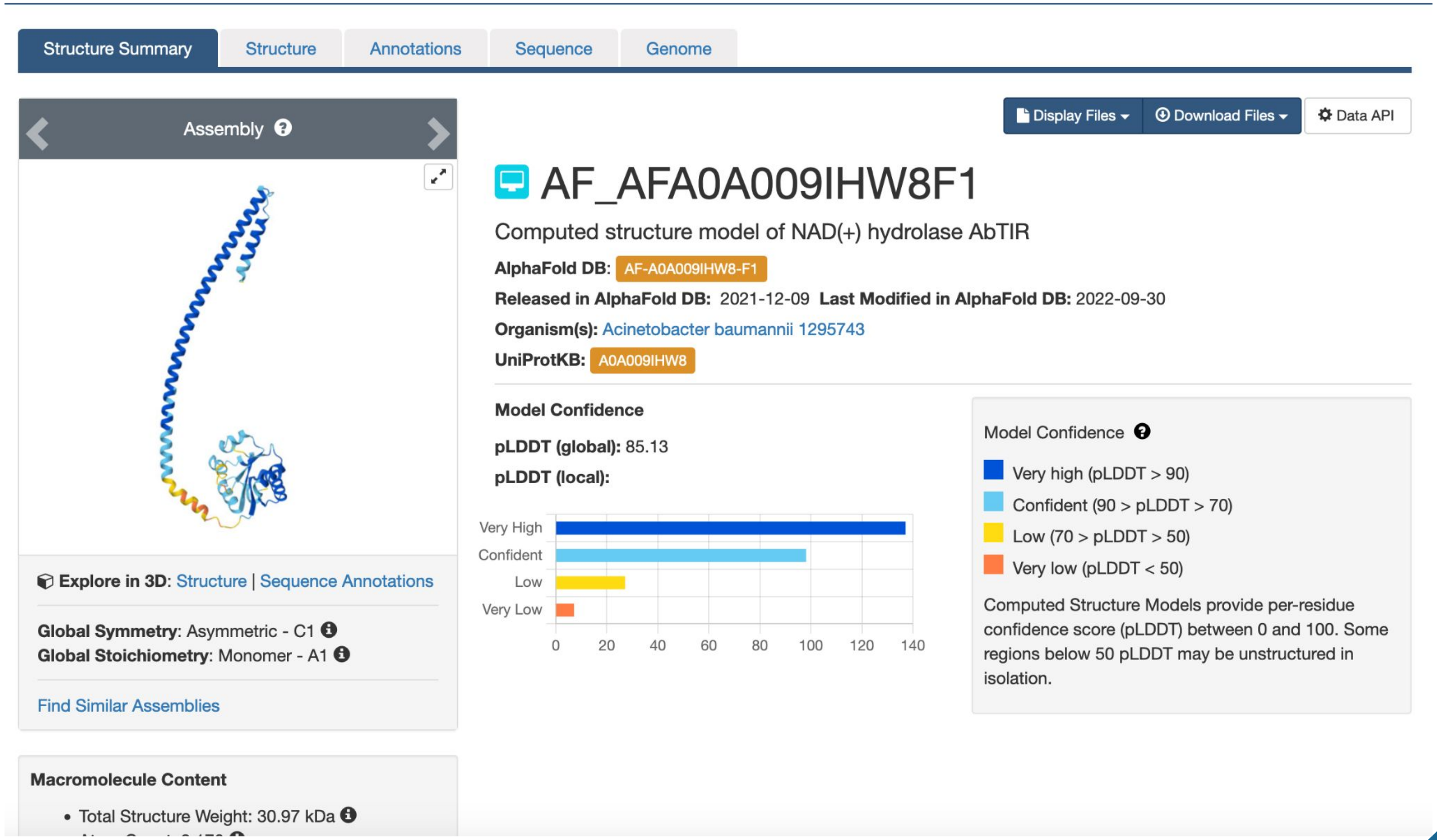
Pharos Homepage Annotation

Chains	Drug Target	Associated Disease
A	Pharos : P00750	• pulmonary embolism • hypertensive disorder • thrombophilia, familial, due to decreased release of tissue plasminogen activator

Example: AF_AFP00750F1

User Views of CSM Summary Pages at RCSB.org

Year	CSM Summary Page Views	Sequence Accesses
2023	1,088,013	134,890
2022	82,916	30,973



AF_AFA0A009IHW8F1

Computed structure model of NAD(+) hydrolase AbTIR

AlphaFold DB: AF-A0A009IHW8-F1

Released in AlphaFold DB: 2021-12-09 **Last Modified in AlphaFold DB:** 2022-09-30

Organism(s): Acinetobacter baumannii 1295743

UniProtKB: A0A009IHW8

Model Confidence

pLDDT (global): 85.13

pLDDT (local):

Model Confidence

- 🟦 Very high (pLDDT > 90)
- 🟦 Confident (90 > pLDDT > 70)
- 🟨 Low (70 > pLDDT > 50)
- 🟧 Very low (pLDDT < 50)

Computed Structure Models provide per-residue confidence score (pLDDT) between 0 and 100. Some regions below 50 pLDDT may be unstructured in isolation.

Explore in 3D: Structure | Sequence Annotations

Global Symmetry: Asymmetric - C1

Global Stoichiometry: Monomer - A1

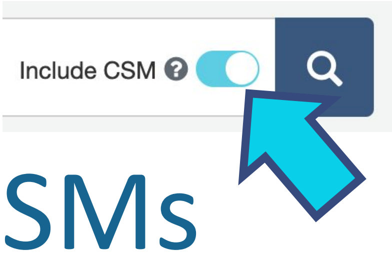
Find Similar Assemblies

Macromolecule Content

- Total Structure Weight: 30.97 kDa

Source: RCSB PDB Analytics

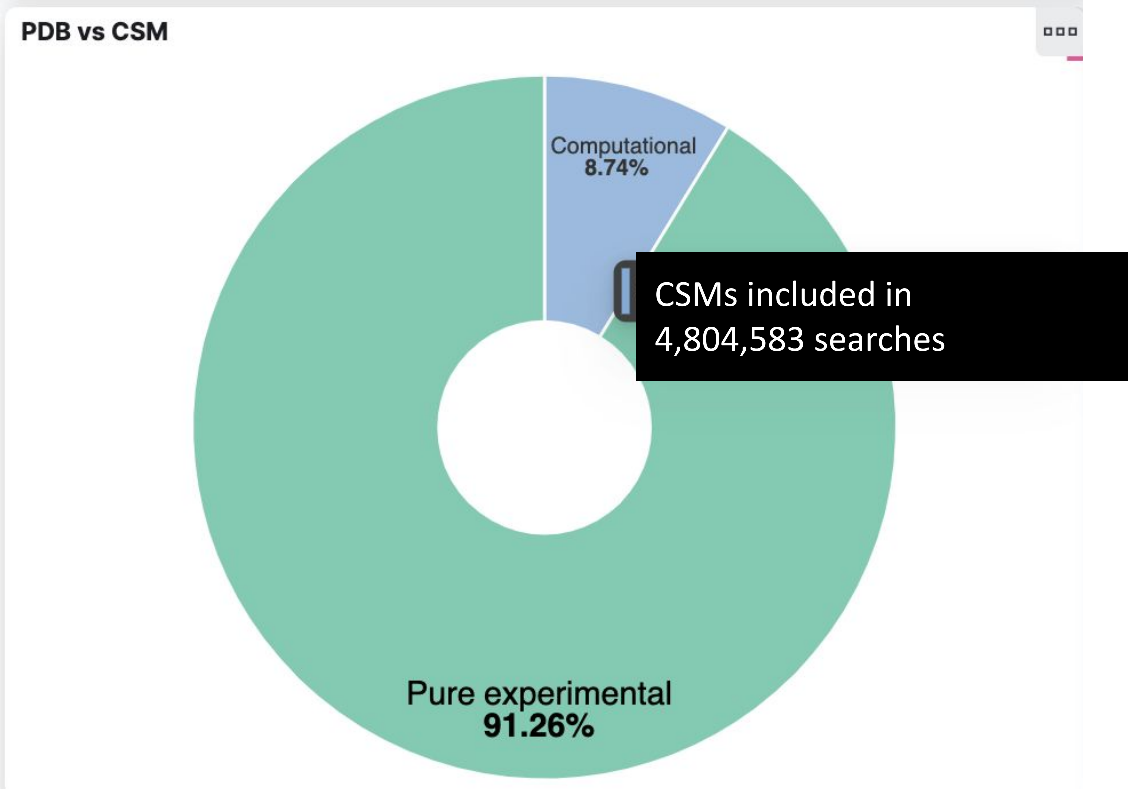
Example: AF_AFA0A009IHW8F1



Include CSM

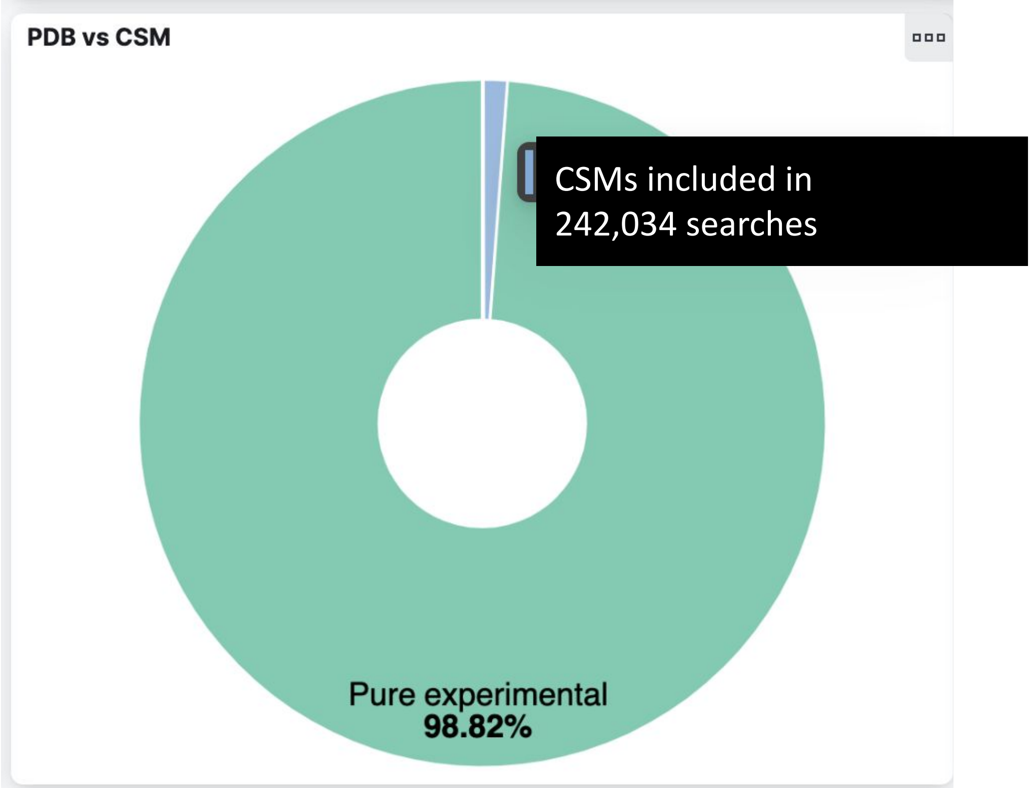
Programmatic Users "Opt-in" to Include CSMs

Programmatic (API) Searches (2023)



CSMs included in 4,804,583 searches

Manual Searches (2023)

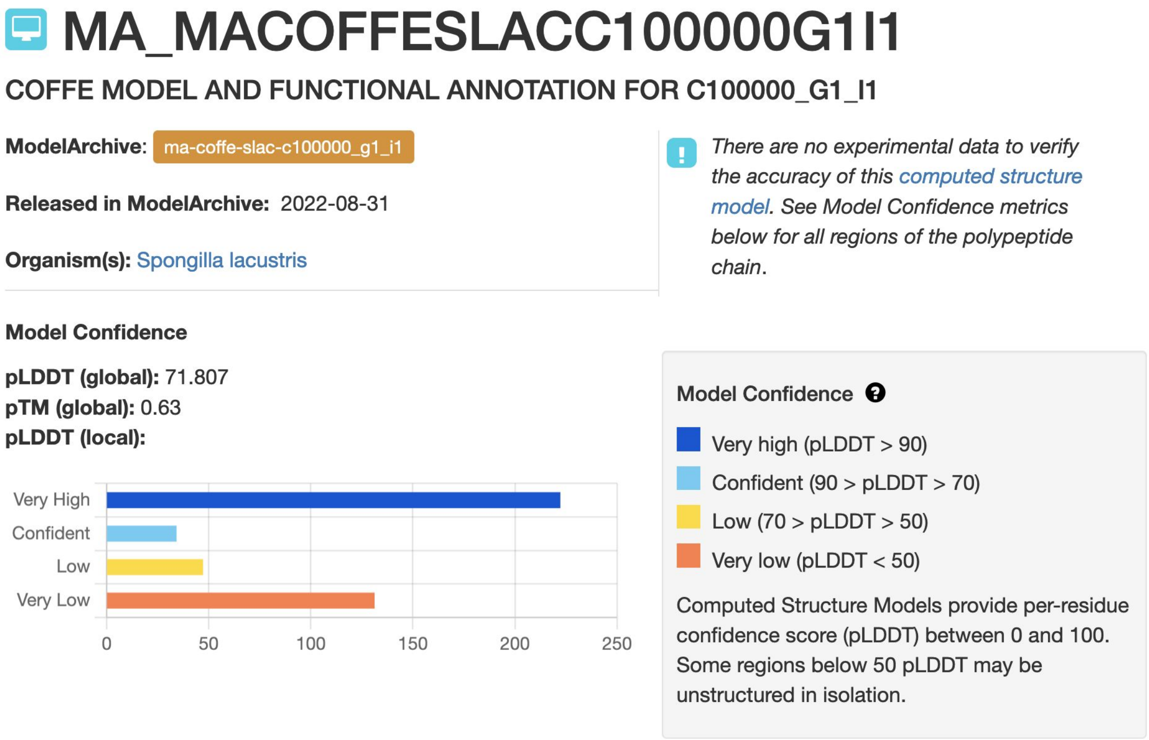


CSMs included in 242,034 searches

Source: RCSB PDB Analytics

Rutgers User Experience Design (UXD) CSM Review

- Spring Semester 2024 Course
  - Students in Master of Business and Science program
- Process
  - Design study to identify how users navigate CSMs at RCSB.org (done)
  - Collect and analyze user needs and pain points (in progress)
  - Deliver recommendations May 2024
- Surveying Encompassed
  - Are users able to include/exclude CSMs in searches?
  - Do users know if they are looking at a CSM or PDB structure?
  - Is it clear how to assess CSM quality?



MA_MACOFFESLACC100000G1I1

COFFE MODEL AND FUNCTIONAL ANNOTATION FOR C100000_G1_I1

ModelArchive: ma-coffe-slac-c100000_g1_i1

Released in ModelArchive: 2022-08-31

Organism(s): Spongilla lacustris

There are no experimental data to verify the accuracy of this computed structure model. See Model Confidence metrics below for all regions of the polypeptide chain.

Model Confidence

pLDDT (global): 71.807
pTM (global): 0.63
pLDDT (local):

Model Confidence

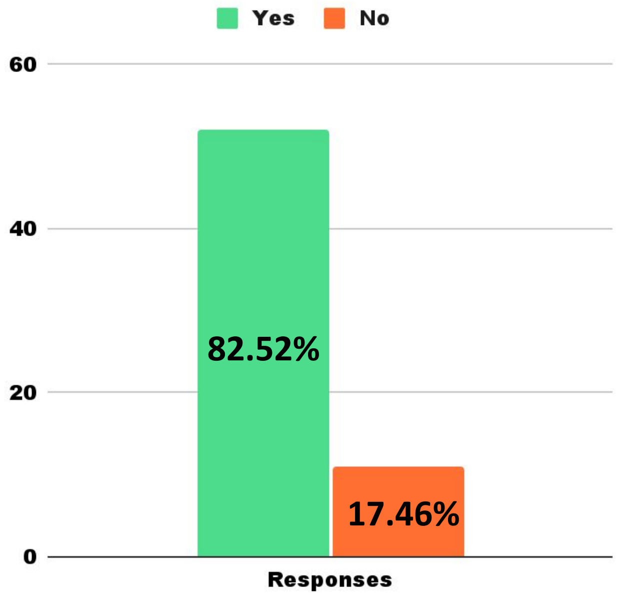
- 🟦 Very high (pLDDT > 90)
- 🟦 Confident (90 > pLDDT > 70)
- 🟨 Low (70 > pLDDT > 50)
- 🟧 Very low (pLDDT < 50)

Computed Structure Models provide per-residue confidence score (pLDDT) between 0 and 100. Some regions below 50 pLDDT may be unstructured in isolation.

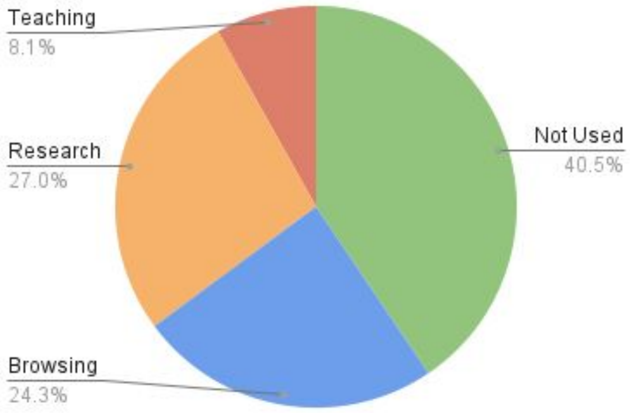
Example: MA_MACOFFESLACC100000G1I1

Q1 User Survey on Using CSMs at RCSB.org

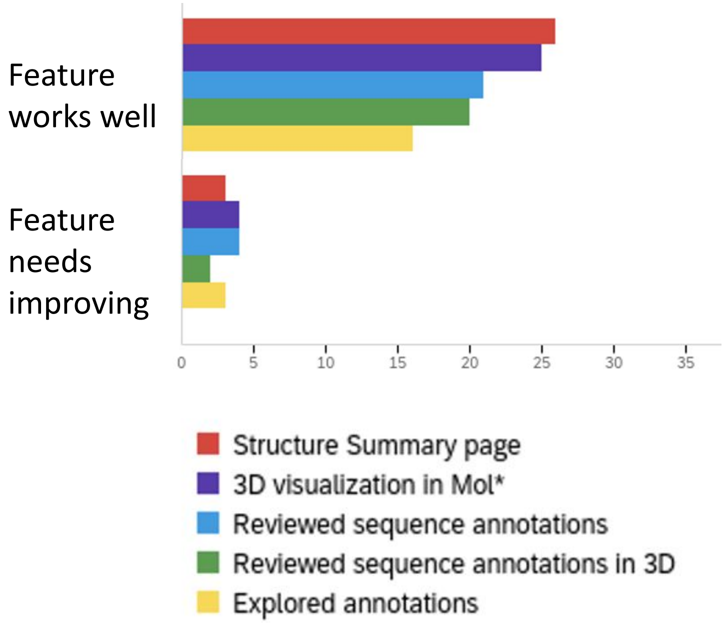
Do you understand the icons used for experimentally-determined structures (🧪) and CSMs (🖥)?



What have you used CSMs at RCSB.org for?

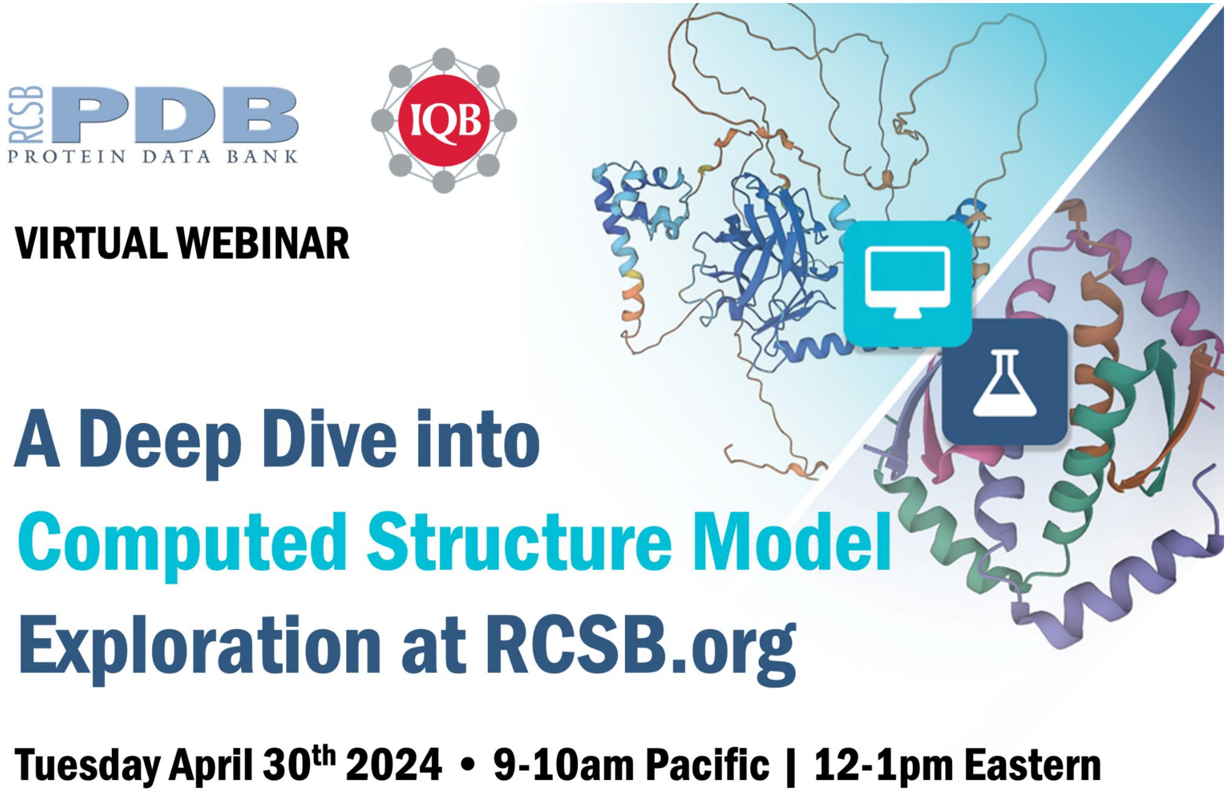


Which CSM features have you used/liked?



63 Responses

Upcoming Training Webinar on CSMs



VIRTUAL WEBINAR

A Deep Dive into Computed Structure Model Exploration at RCSB.org

Tuesday April 30th 2024 • 9-10am Pacific | 12-1pm Eastern

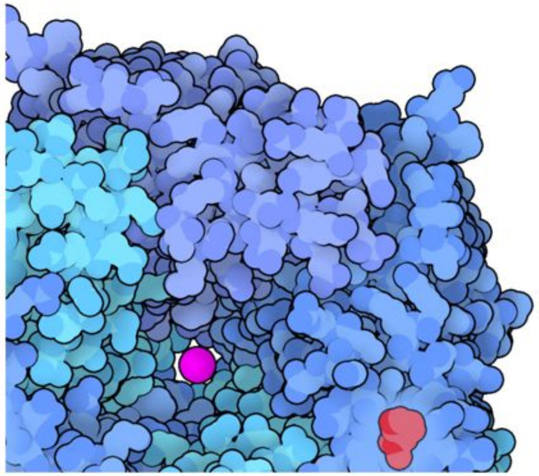
Join us as we demonstrate how RCSB.org serves as your gateway to structural data exploration on Tuesday April 30, 2024 from 9-10am Pacific, noon-1pm Eastern.

This event will equip you with the knowledge of how to use RCSB.org features to navigate 3D predicted protein structures in the context of experimentally-determined PDB structures.

Registration: https://go.rutgers.edu/1ztidbcw

S3 2024 Roadmap Highlights

- Research/Prototyping: Exploring application of AI/ML methods in scientific search applications
- Advanced Search Redesign: Improve user experience and increase user engagement with the Advanced Search tool
- Homepage Redesign: Improve navigation and increase traffic towards advanced features
- Display Metadata for Evolving Methods: SX/XFEL
- Documentation Homepage: Enhance user experience for individuals seeking information pertaining to the features and data available on RCSB.org



Recruiting Updates and Team Transitions

Stephen Burley

Other Team Member Transitions (April 2023-present)

Recent Hires

- Senior Front-End Web Developer (Rutgers): Ronald Brown starts April 29
- Junior DevOps Engineer (Rutgers): offer accepted, background check passed, starting soon
- Scientific Software Developers:
  - Jared Sagendorf (UCSF)
  - Douglas Myers-Turnbull (UCSD)
- Gap Year Science Communication Amy Wu-Wu (Rutgers)
- Jason Kaebler (IQB) serving as new 3DEM advisor

Departures

- Scientific Software Developers: Li Chen (Rutgers), Alicia Evans (Rutgers), Maryam Fayazi (Rutgers), Igor Khokhriakov (UCSD), Zintis May-Krumins (Rutgers)

Li Chen retirement after 22 years;
Shamara Whetstone recruiting at Rutgers Job Fair

OPPORTUNITIES for SCIENTIFIC SOFTWARE DEVELOPERS, GRADUATES, and UNDERGRADUATES

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.

Visit www.rcsb.org/pages/jobs for more information

- Back End Software Engineer (Rutgers)
- High Performance Computing Workflows Architect (Rutgers)
- Postdoctoral Researcher in Bioinformatics (UCSD)
- Gap Year Opportunities (Rutgers)
- Undergraduate Summer Research (RISE at Rutgers)

RCSB PDB members with RISE 2023 student developers

RCSB PDB Team

RCSB.ORG

info@rcsb.org

Core Operations 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
THE STATE UNIVERSITY
OF NEW JERSEY

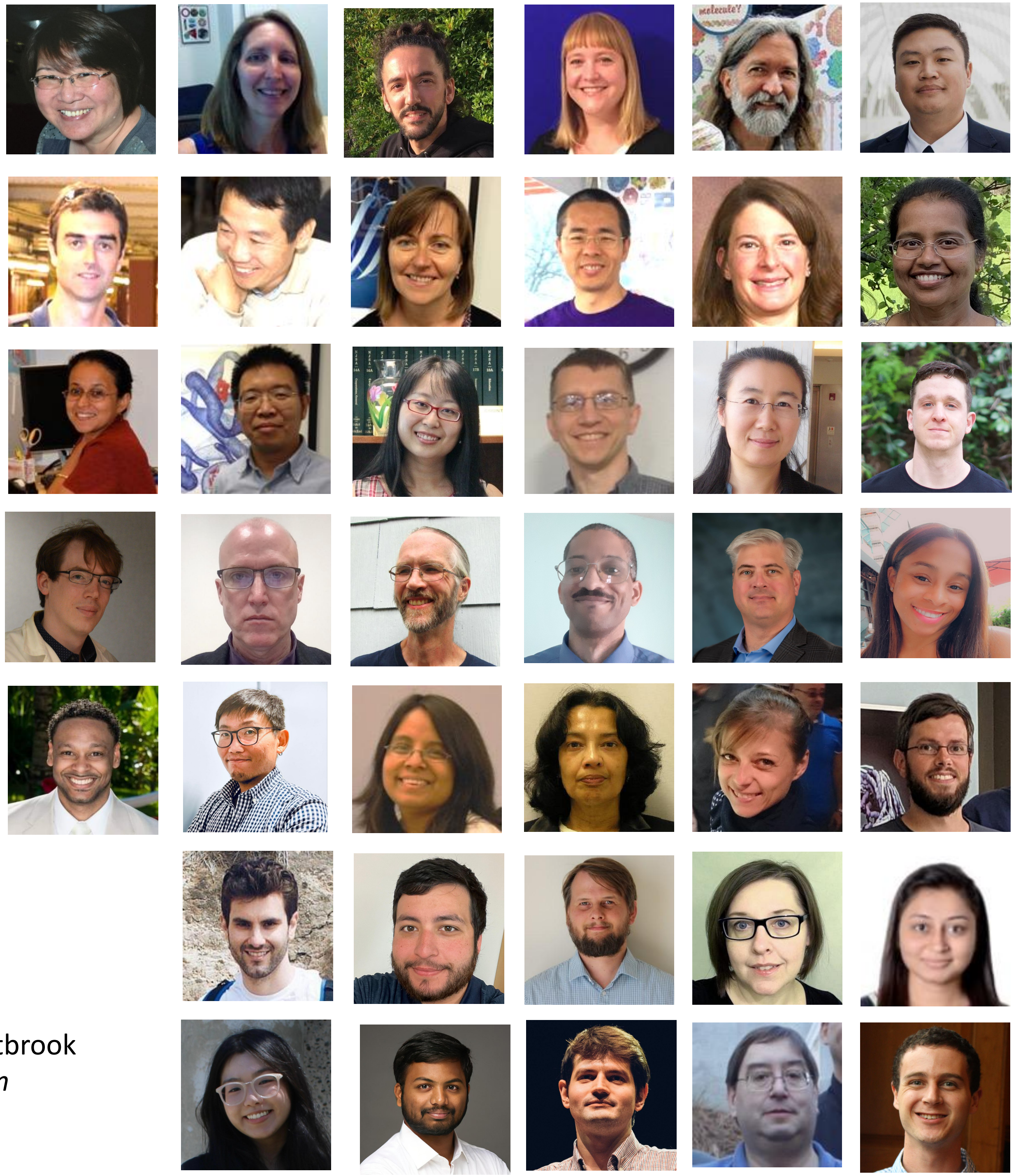
UC San Diego

SDSC SAN DIEGO SUPERCOMPUTER CENTER

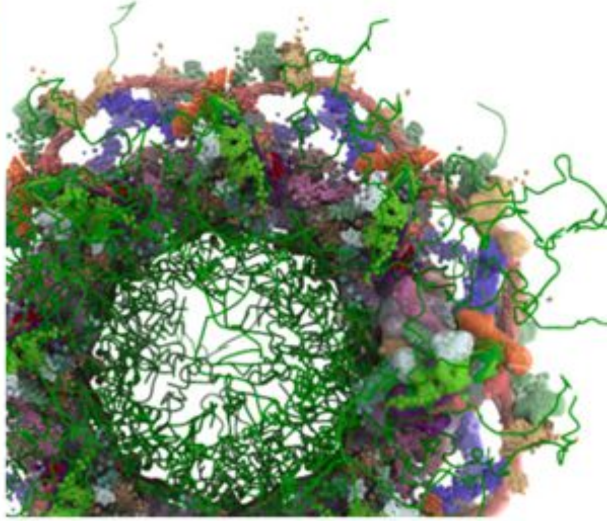
UCSF
University of California
San Francisco

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

Follow us



John D. Westbrook
In memoriam
1957-2021



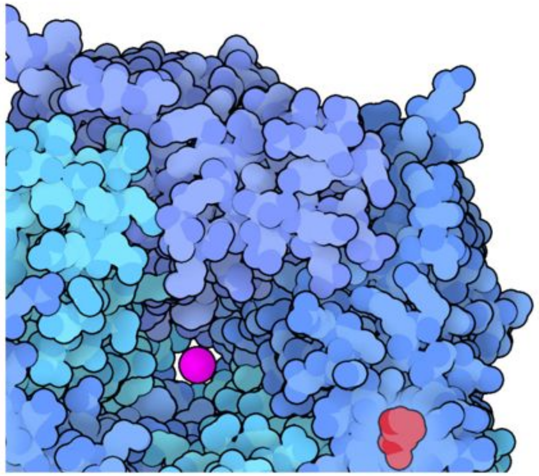
Responses to 2023 Report Questions for Committee

Stephen Burley

Responses to 2023 Recommendations

Increase the adoption of community-based feedback to refine services and user experience	RCSB PDB will continue to collect feedback (meetings, help desk, user surveys, and UXD reviews). In 2024, we are testing virtual "office hours" as another mechanism
Engage the User Experience Design group on CSMs	CSMs at RCSB.org will be reviewed as part of the Spring 2024 UXD Review
More user engagement to inform about CSMs and their strengths and weaknesses.	We are highlighting CSMs at meetings and will host an training event in April; materials will be published with other webinars at PDB-101
Create a web-based short video guides to inform users about new features on YouTube	RCSB PDB plans to explore options and best practices for video guides in 2024, with a goal of publishing videos in 2025 RCSB PDB will continue to collaborate with our wwPDB partners on depositor-focused videos published at wwPDB.org and YouTube
Consider a redesign of RCSB.org	RCSB PDB plans to start this process by improving the home page at RCSB.org as well as the Advanced Search interface later in 2024
Crash Courses could expand to advanced searches, APIs, CSMs, and Mol*	We began to offer more courses in 2023, including a focus on PDBx/mmCIF; APIs; and using RCSB.org. 2024 training will similarly include crash courses/webinars that target data depositors and data consumers. Advanced Search training events will be scheduled post-redesign.

2023 RCSB PDB AC: Report (PDF) and RCSB PDB Responses



Roadmap Highlights: Service 4 Training, Outreach, Education

Will be presented if time allows

S4 2023 Selected Roadmap Achievements

- Webinars
  - *Leveraging RCSB PDB APIs for Bioinformatics Analyses and Machine Learning*
  - *Teaching enzymology with the Protein Data Bank: from pandemic to Paxlovid*
  - *Use PDB data to their full extent: Understanding PDBx/mmCIF*
- New features: Exploring the Structural Biology of
  - *Health and Nutrition*
  - *Viruses*
  - *Bioenergy*

Brinda Vallat, Santiago Blaumann, Rusham Bhatt, Dennis Piehl developed a Python package (rcsbsearchapi) that can be used for accessing the RCSB PDB Search API as part of the 2023 Research Intensive Experience at Rutgers

S4 2024 *Roadmap Planned Highlights*

- Virtual Training Events: Mol*, CSM Exploration, PDB Validation, Teaching Enzymology (recordings at PDB-101)
- Virtual Office Hours: RCSB.org, “Ask a Biocurator”, Mol*, APIs
- New feature: *Exploring the Structural Biology of Evolution*
- New feature: protein folding poster and activity
- DEIA
  - Undergraduate training: API development (through Rutgers Research Intensive Summer Experience program for outstanding students from diverse backgrounds)
  - Annual Biomedical Research Conference For Minoritized Scientists
  - National Diversity in STEM Conference