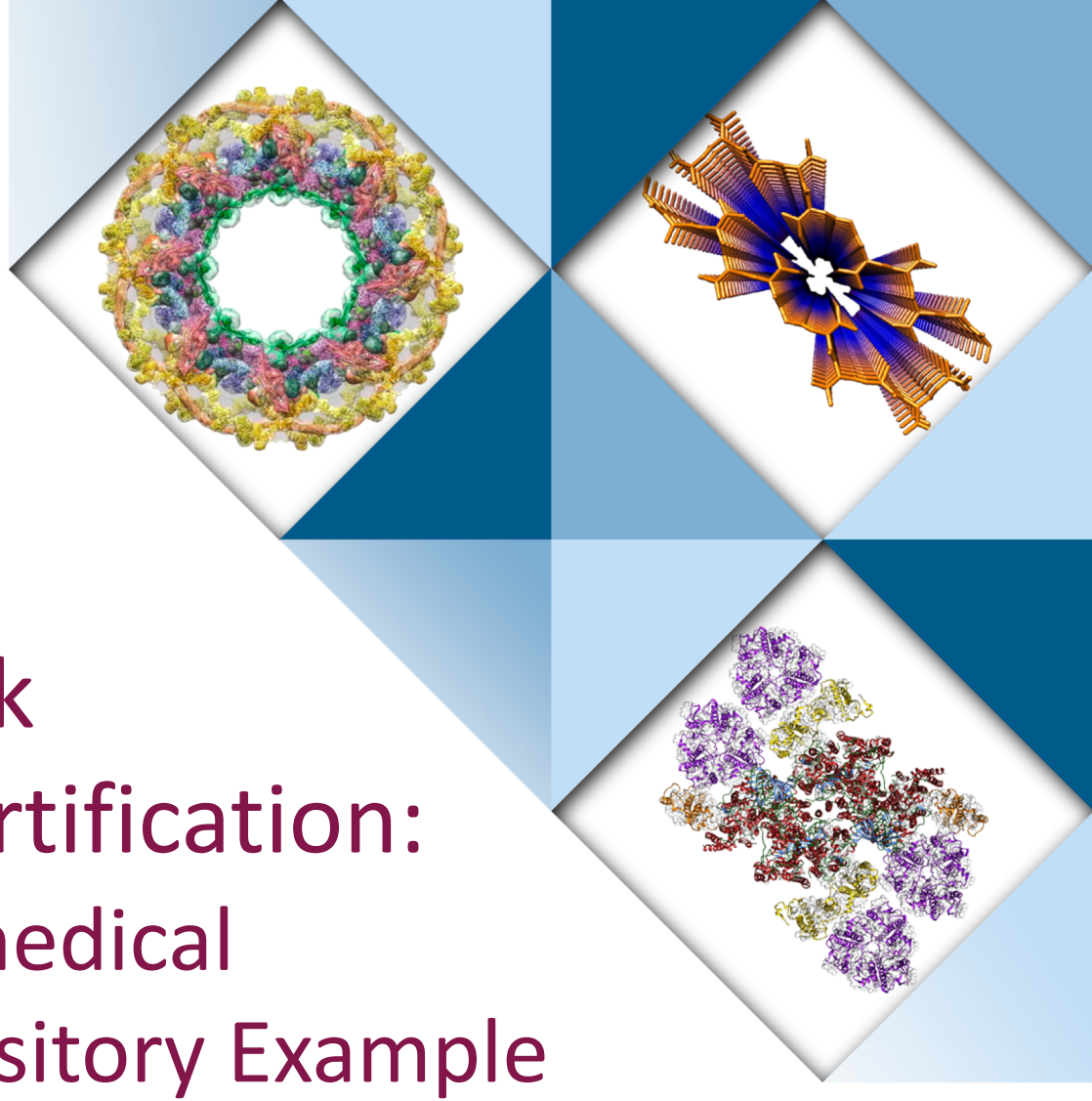
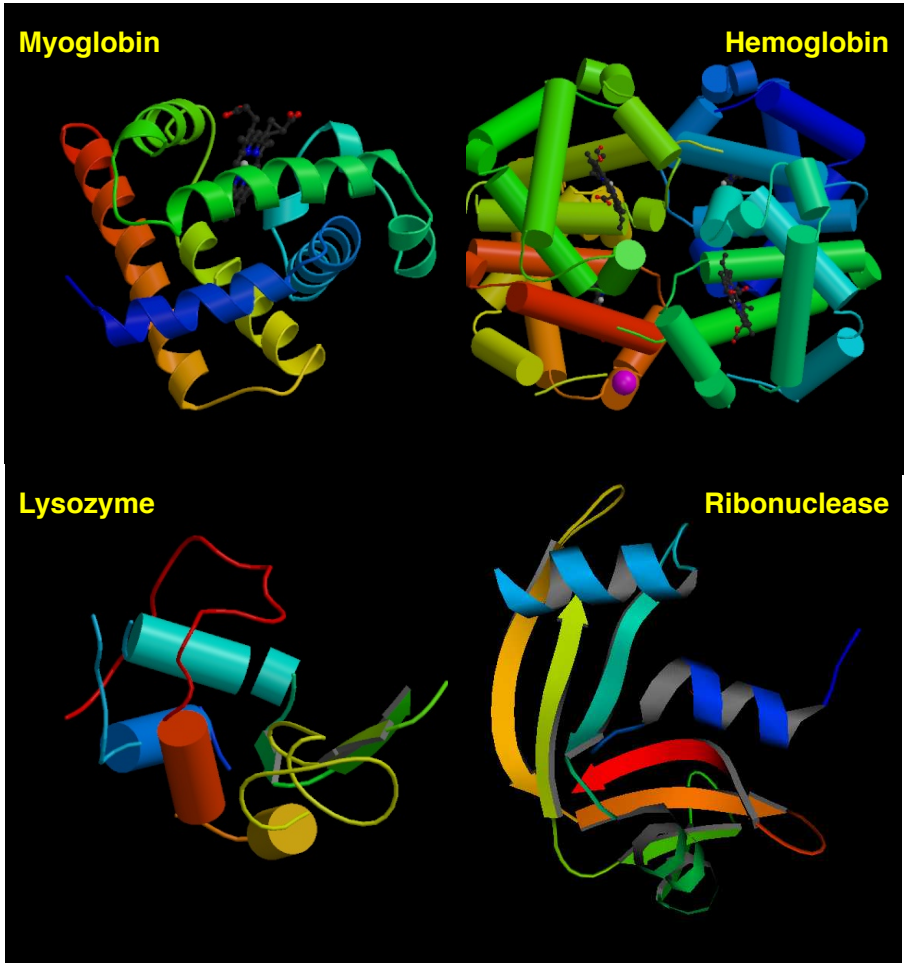


Protein Data Bank CoreTrustSeal Certification: A Community Biomedical Archival Data Repository Example

John Westbrook RCSB PDB

Protein Data Bank History

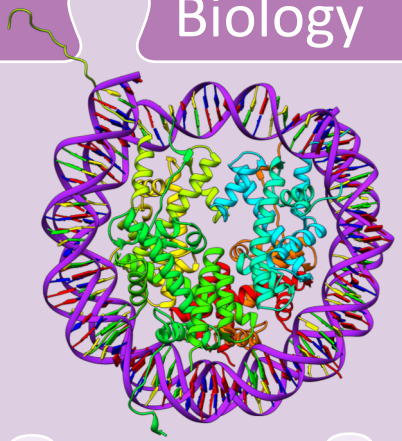
- PDB 1st Open Access archival digital data resource in all of biology
- Founded 1971with 7 X-ray structures of proteins
- Single global archive for protein and DNA/RNA 3D experimental structures
- Today, Open Access to >150,000 structures
- wwPDB collaboration US (RCSB PDB), EU (PDBe), Japan (PDBJ), and BMRB



Some of the earliest structures in the PDB

Structure Data Contributes to Fundamental Biology, Biomedicine, and Energy

Fundamental Biology



Nuclear Cell Biology

Molecular Evolution

Molecular Transport

Cellular Signaling

Enzymes

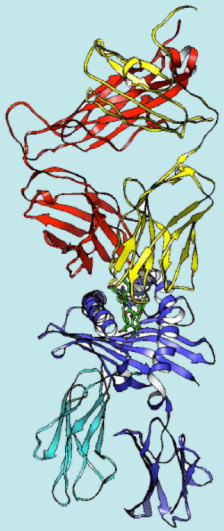
Nanotechnology

Protein Folding

Molecular Infrastructure

Biomedicine

Explaining T-cell of Immunology



Zika Virus

Cancer

Precision Medicine

HIV & AIDS

Anti-Microbial Resistance

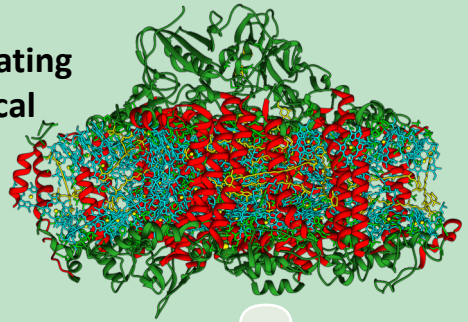
Vaccine Development

Ebola

Type II Diabetes

Energy

Illuminating Biological Energy



Biofuels

Renewable Energy

Cyanobacteria

Biotechnology

Crop Sustainability

Methane Production

Hydrogen Gas

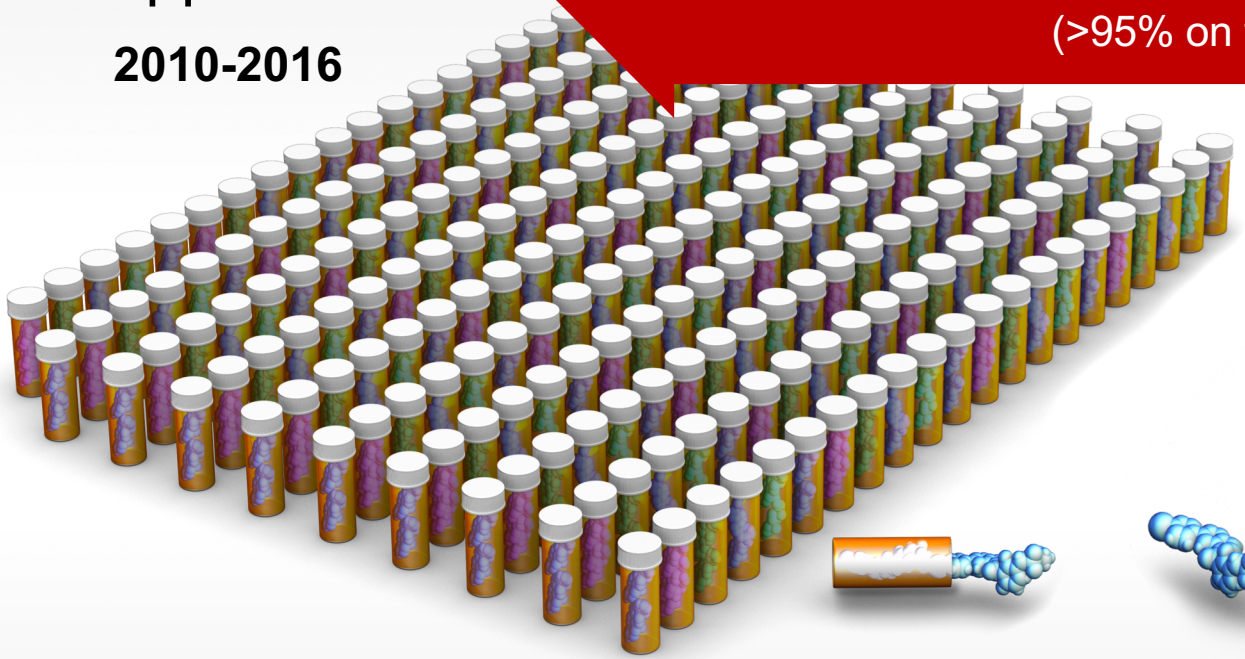
Impact of PDB Data on Drug Approvals

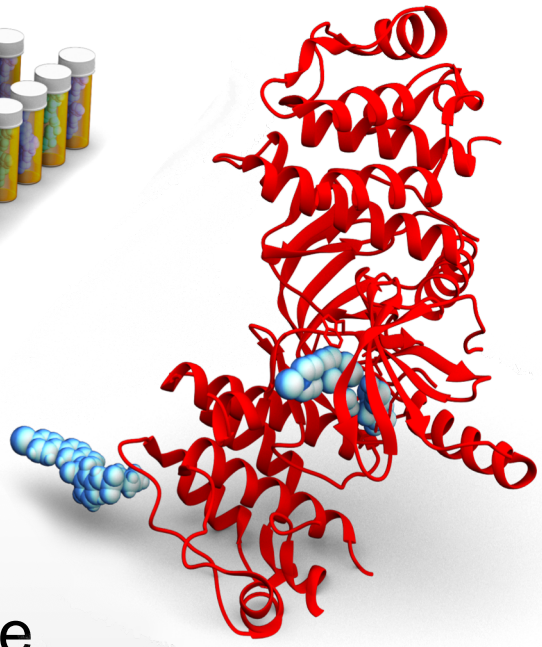
210 NEW DRUGS approved

2010-2016

2000-2016

>$100 BILLION **of NIH funding** contributed to these approvals (>95% on targets)[1]





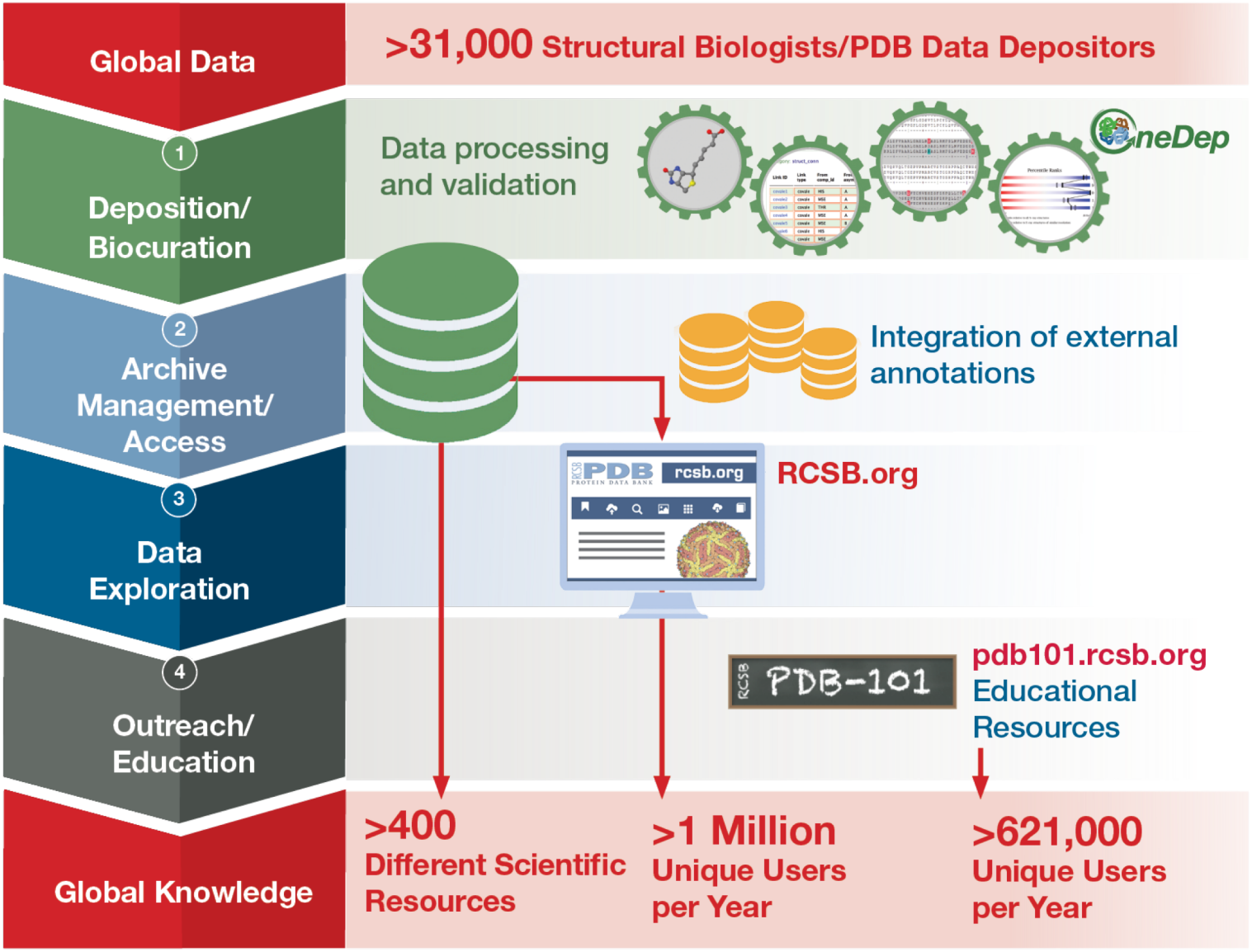
B-Raf Kinase complex with Vemurafenib ***PDB ID 3og7***

>6,000 PDB Structures contributed to **183** of these drug approvals

Westbrook and Burley (2019) *Structure 27,* 211-2117.

Galkina Cleary *et al.* (2018) *PNAS 115,* 2329-2334.

RCSB PDB Services Support the Full Structural Biology Data Life Cycle



PDB Incentives for CTS Certification

- Strong commitment and tradition within our scientific community for support of data and process standards
- Expectations of both our repository contributors and users to adopt and maintain best practices in archiving and data management
- Increasing focus of funders on supporting FAIR data management practices
- Certification documents the resource investment required to responsibly manage the full life cycle of archival data
- Relatively low barrier and modest effort certification process
- Good balance between rigor and certification effort

Benefits of the Certification Process

- Requires an audit of the full life cycle of the repository data pipeline
- Uncovers implicit knowledge of processes that may lack proper documentation
- Useful exercise to identify systematic weaknesses and potential areas for improvement
- Provides an opportunity to explore how other disciplines are addressing similar data management challenges
- Provides a useful benchmark for resource and capacity planning
- Provides an excellent learning experience

The CTS Certification Process

- Straight forward application process with a variety of examples to frame your input
- The majority of the required information was already in public view or in existing project documents
- Leveraged our efforts in supporting the OAIS Archive Reference Model
- Required/expected level of detail is a bit ambiguous

Reference Model for an Open Archival Information System (OAIS). Magenta Book CCSDS 6500-M-2. Washington: Consultative Committee for Space Data Systems, NASA; 2012.

Some Certification Outcomes for PDB

- Harmonized practices and documentation across our regional data centers
- Improved alignment of our documentation with FAIR/FACT objectives
- Introspection helped focus our long-term plans to improve availability and disaster preparedness
- Explored some new approaches for schema registration, exchange and data discovery
- Certification beneficially contributed to our funding reapplication

Some CTS and Certification Challenges

- Supporting CTS requires diverse expertise in data science and engineering as well as in the target domain
- The long time horizon of some CTS objectives are difficult to support with typical 3-5 year competitive funding cycles
- Addressing long term objectives is similarly complicated for leased or cloud deployed infrastructure
- The resource burdens for robust CTS support may not be:
  - fully accounted in the scope of current program offerings
  - fully appreciated by grant reviewers
- Meetings and workshops like this will be important in providing the broader education to address some of these challenges

RCSB PDB Team

RCSB PDB PROTEIN DATA BANK

RCSB.ORG

info@rcsb.org

Funding

RCSB PDB is funded by a grant (DBI-1338415)
from the National Science Foundation,
the National Cancer Institute,
the National Institute of General Medical Sciences,
and the US Department of Energy

Management

RCSB PDB is hosted by:

RUTGERS

UC San Diego

SDSC SAN DIEGO SUPERCOMPUTER CENTER

UCSF University of California San Francisco

WORLDWIDE PDB PROTEIN DATA BANK

RCSB PDB is a member of the Worldwide Protein Data Bank partnership (wwPDB; **wwpdb.org**)

Follow us