

RCSB Protein Data Bank: Making connections from genes to ecosystems

John D. Westbrook^{1*} (john.westbrook@rcsb.org), Jasmine Young,¹ Robert Lowe,¹ Christine Zardecki,¹ Jose Duarte,² and **Stephen K. Burley**^{1,2}

¹Rutgers, The State University of New Jersey, Piscataway, NJ; ²University of California San Diego, La Jolla, CA

<http://www.rcsb.org> and <http://pdb101.rcsb.org>

Project Goals: The Vision of the RCSB PDB is to enable open access to the accumulating knowledge of 3D structure, function, and evolution of biological macromolecules, expanding the frontiers of fundamental biology, biomedicine, and biotechnology.

Protein Data Bank (PDB) was established as the 1st open access digital data resource in biology and medicine. Today, the PDB is one of two global resources for experimental data central to science as a public good (the other key Primary Data Archive being the International Nucleotide Sequence Database Collaboration). PDB currently houses >160,000 atomic level biomolecular structures determined by crystallography, NMR spectroscopy, and 3D electron microscopy. It is managed by the Worldwide Protein Data Bank partnership (wwPDB; wwpdb.org) according to the FAIR principles (*i.e.*, Findability, Accessibility, Interoperability, and Reusability).

Through an internet information portal and downloadable data archive, researchers and educators can access 3D structure data for large biological molecules, such as proteins and DNA. These are the molecules of life, found in all organisms on the planet. Knowing the 3D structure or shape of a biological macromolecule is essential for understanding the role the molecule plays in health and disease of humans, animals, and plants, food and energy production, and other topics of concern to global prosperity and sustainability.

The RCSB PDB operates the US data center for PDB, serves as Archive Keeper for the global PDB archive, and makes PDB data available at no charge to all Data Consumers without limitations on usage. Studies of website usage, bibliometrics, and economics demonstrate the powerful impact of the PDB data on basic and applied research, clinical medicine, education, and the economy.

During calendar 2019, >800 million structure data files were downloaded from the RCSB PDB by Data Consumers working worldwide. During this same period, the RCSB PDB processed >5,530 new atomic level biomolecular structures plus experimental data and metadata coming into the archive from Data Depositors working in the Americas and Oceania. In addition, RCSB PDB served millions RCSB.org users worldwide with PDB data integrated with ~40 external data resources providing rich structural views of fundamental biology, biomedicine, and energy sciences, and supported >650,000 PDB101.rcsb.org educational website users around the globe.

RCSB PDB serves a rich collection application programmable interfaces (APIs) that enable searching of and access to PDB data content. These APIs enable search by key citation, biological and structural features, and retrieval of individual PDB structure entries, reports, and chemical and molecular reference data. Building on this functionality in 2020, RCSB PDB will deploy new API services that integrate structure and sequence comparison in combination with more granular access to the corpus of PDB data.

Access to PDB data and services contribute to patent applications, drug discovery and development, publication of scientific studies, innovations that can lead to new product development and company formation, and STEM education.

Publications

1. The Protein Data Bank (2000) *Nucleic Acids Research* 28: 235-242. doi: 10.1093/nar/28.1.235.
2. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy (2019) *Nucleic Acids Research* 47: D464–D474. doi: 10.1093/nar/gky1004.
3. RCSB Protein Data Bank: Enabling biomedical research and drug discovery (2019) *Protein Science* 29: 52-65 doi: 10.1002/pro.3730.

RCSB PDB is funded by the National Science Foundation (DBI-1832184), the US Department of Energy (DE-SC0019749), and the National Cancer Institute, National Institute of Allergy and Infectious Diseases, and National Institute of General Medical Sciences of the National Institutes of Health under grant R01GM133198.