wwPDB biocuration: on the front line of structural biology

Biocurators, the backbone of the wwPDB, manage structural biology data deposition, quality, and integrity, and provide integral support to the research community worldwide.

Jasmine Y. Young, John Berrisford and Minyu Chen

Through the open-access Protein Data Bank (PDB) archive, structural biologists provide and gain access to atomic-level views of over 175,000 biological macromolecules. These structures augment our understanding of the connections between structure and function in biology. The archive continues to grow every year in both the number of structures and complexity––even during the COVID-19 pandemic. The PDB is widely regarded as one of the best-curated biodata resources[1], supporting the structural biology community and resourced by a team of expert biocurators vital to the integrity of the PDB data ecosystem. The biocurators collaborate with PDB depositors using the OneDep software system[2] to standardize, validate and biocurate incoming structure data, ensuring that they are findable, accessible, interoperable and reusable (FAIR). Now in its 50th year of operation, the PDB has exemplified the FAIR principles for responsible data management since long before they were widely known and adopted[3].

Key to the long-term success and sustainability of the PDB has been the Worldwide Protein Data Bank partnership[4,5], formalized in 2003. The wwPDB jointly manages the single global PDB archive. Today, more than 15 biocurators work at wwPDB data centers located in the United States, Europe and Asia. Using the OneDep system, we share responsibility for managing incoming PDB structures along geographic lines, bringing diverse expertise in biochemistry, biophysics, computational chemistry, enzymology and small-molecule crystallography to the global enterprise. With training in macromolecular crystallography, nuclear magnetic resonance spectrometry and electron microscopy, we operate on the front lines of structural biology, working closely with thousands of PDB depositors every year.

Since its launch in 1971, more than 100 biocurators have served the PDB, united by passion for science, thirst for knowledge, and interest in file formats, data dictionaries and ontologies. This is as true now as it was 47 years ago when PDB biocuration pioneer Frances C. Bernstein and others at Brookhaven National Laboratory curated data that were represented in the original PDB file format. Today, strict definitions of data types and file formats in the fully extensible PDBx/mmCIF data dictionary allow identification of errors and inconsistencies, highlight molecular properties, and connect data depositors, biocurators and data consumers. While structural biologists push the envelope by determining ever larger and more complex structures, developing novel experimental methods, or designing new validation tools, we work closely with software developers, structural biology facilities and scientific innovators to refine PDB data management practices. We also use OneDep periodically to look back and across the archive with 'remediation efforts' to bring previously released structure data up to modern standards.

The move for biocurators to work from home, away from each regional data center during the COVID-19 pandemic, was enabled by the global collaborative practices that support the wwPDB partnership. As the structural biology community stepped up to provide much-needed insights into SARS-CoV-2, wwPDB biocuration of SARS-CoV-2 protein structures was prioritized to ensure rapid public release of data while maintaining our enduring commitment to quality. More than 1,000 COVID-19-related structures are now freely available from the PDB, a year after release of the first SARS-CoV-2 structure. Combined with SARS-CoV-1 (over 200 structures) and MERS-CoV (70 structures) PDB structures from the two earlier epidemics, PDB data are facilitating structure-guided discovery and development of anti-coronavirus drugs, vaccines and neutralizing antibodies.

wwPDB Biocurator Summit held virtually in 2020 during the pandemic. Top row, from the left: Yuhe Liang (RCSB PDB), Jasmine Young (RCSB PDB), Irina Persikova (RCSB PDB). Second row, from the left: Deborah Harrus (PDBe), David Armstrong (PDBe), Brian Hudson (RCSB PDB). Bottom row, from the left: Ezra Peisach (RCSB PDB), John Berrisford (PDBe), Minyu Chen (PDBj).

Open access to PDB data has helped to expose the inner workings of the coronavirus beyond the scientific community. Throughout the COVID-19 pandemic, wwPDB biocuration staff have continued to support research, training and education worldwide. We are thrilled that our efforts have contributed to general awareness of the importance of scientific advances. As science evolves and new technologies and methods are adopted by the structural biology community, we will face increasing challenges in handling the growth in number, size and complexity of depositions to the PDB. In response to these challenges, we are working with the community to increase the accuracy of metadata in PDB entries through increased automatic data harvesting, increasing the level of automation in the biocuration pipeline to improve biocuration efficiency, and collaborating with the community task forces and mmCIF Working Group (http://wwpdb.org/task/mmcif) to extend the data model and improve validation of models and experimental data that will better support new technologies and methods. ❐

Jasmine Y. Young [1]✉, John Berrisford [2] and Minyu Chen[3]

[1]*Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB), Institute for Quantitative Biomedicine, Rutgers, The State*

University of New Jersey, Piscataway, NJ, USA. [2]*Protein Data Bank in Europe (PDBe), European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, UK.* [3]*Protein Data Bank Japan (PDBj), Institute for Protein Research, Osaka University, Osaka, Japan.*
✉*e-mail: jasmine.young@rcsb.org*

Published online: 7 May 2021
https://doi.org/10.1038/s41592-021-01137-z

References
1. Howe, D. et al. *Nature* **455**, 47–50 (2008).
2. Young, J. Y. et al. *Database* **2018**, bay002 (2018).
3. Wilkinson, M. D. et al. *Sci. Data* **3**, 160018 (2016).
4. Berman, H., Henrick, K. & Nakamura, H. *Nat. Struct. Biol.* **10**, 980 (2003).
5. wwPDB Consortium. *Nucleic Acids Res.* **47**, D520–D528 (2019).

Acknowledgements
RCSB PDB is jointly funded by the US 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 US National Institutes of Health (R01GM133198). The Protein Data Bank in Europe is supported by the European Molecular Biology Laboratory–European Bioinformatics Institute and Wellcome Trust (104948). Protein Data Bank Japan is supported by the Database Integration Coordination Program from the National Bioscience Database Center (NBDC)–JST (Japan Science and Technology Agency), the Platform Project for Supporting in Drug Discovery and Life Science Research from AMED, and the joint usage program of Institute for Protein Research, Osaka University.

Competing interests
The authors declare no competing interests.

Check for updates

A new era of synchrotron-enabled macromolecular crystallography

The future of macromolecular crystallography includes new X-ray sources, enhanced remote-accessible capabilities and time-resolved methods to capture intermediate structures along reaction pathways.

Aina E. Cohen

In 1974, three years after the official start of the Protein Data Bank, the first beamlines dedicated to structural studies began operation at multi-GeV storage rings. During early macromolecular crystallography experiments, a crystal exposed for only a few minutes would produce a higher resolution diffraction pattern than could be measured after hours with a rotating anode source, first demonstrating the value of synchrotron radiation to structural biology[1]. At state-of-the-art synchrotron beamlines today, complete datasets are measured in seconds with crystal rotation speeds exceeding 90° s^{-1} and diffraction image frame rates exceeding 100 Hz. The resulting images are monitored for diffraction quality in real time and transferred to automated processing pipelines to simplify data analysis. As technologies advance, structural investigations are transitioning beyond solving single static structures. Sequential series of structural snapshots are being applied to provide details of the atomic positions and motions that define the relationships involved in molecular recognition, transition state stabilization, allostery and other aspects of the biocatalytic process.

Many proteins retain biological function in the crystalline state, and macromolecular crystallography offers opportunities to study these in different environments and temperatures. Brighter microbeam sources have enabled the use of smaller samples that have improved diffusion and optical transmission characteristics, supporting burgeoning methods for rapid microcrystal freeze-trapping, whereby a process of interest is initiated using light or chemical mixing to capture intermediate structures along the reaction pathway. However, it is extremely difficult to freeze-capture events faster than ~10 ms. To observe faster reactions, time-resolved methods

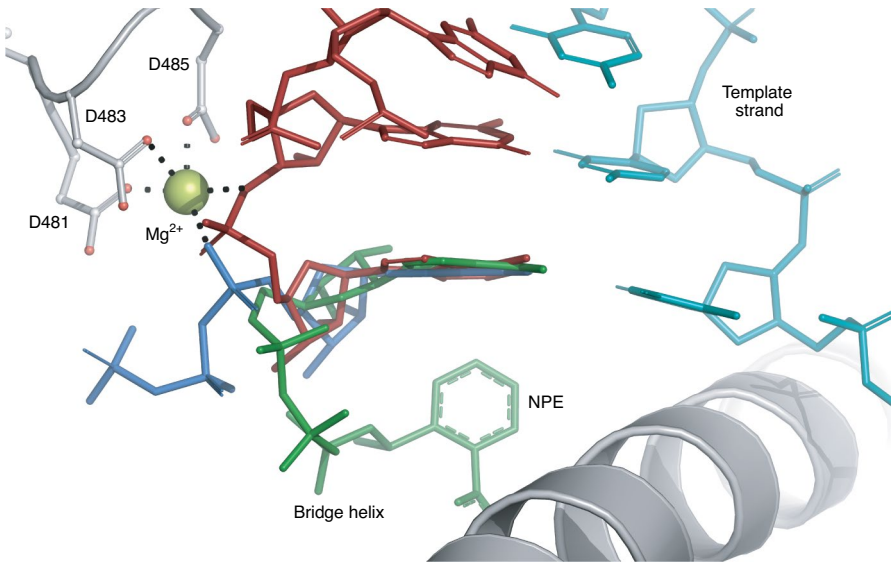

Fig. 1 | Time-resolved studies of transcription. The Calero lab (University of Pittsburgh) has demonstrated that photoactive caged ATP (green) can bind to RNA polymerase II and that ultraviolet illumination can break the nitrobenzyl group (NPE) in crystallo, allowing ATP release, metal coordination (blue) and phosphodiester bond formation (red). These experiments were performed remotely using SSRL beam line 12-1, where crystals were exposed to UV light to break the cage, followed by a temperature increase from 100 K to 170 K (above the glass transition temperature of water), followed by rapid helical-rotation data collection (2 s per dataset). Credit:Guillermo Calero.