Trust but verify: Woes of the Protein Data Bank

Pavel Afonine
Lawrence Berkeley National Laboratory, California, USA

The Protein Data Bank (PDB) is an open-access, global archive for experimentally determined 3D structures of biological macromolecules. Established in 1971 as the first open-access digital data resource in biology—starting with seven X-ray crystallographic protein structures—and now containing >250,000 structures. Managed by the wwPDB consortium, the PDB provides curated data under a CC0 (unrestricted free-to-use) license. Beyond serving structural biomedicine, PDB data have enabled major advances in structure modeling and AI/ML—supporting both homologous modeling and template-free approaches such as AlphaFold2—and are integrated into RCSB's research portal with advanced search and visualization tools. PDB's impact on drug discovery and its public-health role, including >6K SARS-CoV-2 structures, can't be overstated.

Efficient use of this wealth of information may seem trivial, but it is not; it requires understanding the files containing structural information—their format, content, and the meaning of the actual structural data—along with the associated metadata. Multiple traps exist that often novice, and sometimes even seasoned, users of the PDB regularly fall into. The presentation will discuss these caveats and share the speaker's decades of experience to help build a solid foundation for avoiding them.

Pavel Afonine, a research scientist at Lawrence Berkeley National Laboratory (LBNL), USA, has been a key developer of the Phenix software for over two decades, specializing in bio-macromolecular structure modeling through diffraction and cryo-EM methods. He co-founded the Quantum Refinement (Q|R) project during his professorship at Shanghai University (2014–2019) and was a Full Professor there from 2017 to 2019. Pavel holds degrees in physics, mathematics, biology, and biotechnology (Bs., Ms., MIPT, 2000) and a Ph.D. in computational structural biology from Henri Poincaré University, France (2003). His research focuses on three-dimensional macromolecular structure determination, with contributions to crystallographic and cryo-EM methods. He has published over 100 papers, including 14 in Nature, one in Science, and three in PNAS, accumulating over 35,000 citations with an H-index of 45. He has received the Bertaut Prize from the European Crystallographic and Neutron Scattering Associations (2012), Shanghai Eastern Scholar Awards (2015, 2017), and LBNL Outstanding Performance (2006) and Technology Transfer (2021) awards. Phenix, used in academia and industry—including GSK, Novartis, Merck, and Pfizer—has been cited over 25,000 times and contributed to approximately 40% of all Protein Data Bank (PDB) structures (~88,000). Beyond his scientific pursuits, he is passionate about education and enjoys backcountry hiking, endurance racing, world travel, food growing and experimenting with nutrition.