Validation And Quality Assessment for Small-Molecule Ligands In The Protein Data Bank

Chenghua Shao[1], Jasmine Y Young[2], Charmi Bhikadiya[3], Ezra Peisach[4], Jose M Duarte[5], Yana Rose[6], Zukang Feng[7], wwPDB Team[8], Stephen K Burley[9]

[1]RCSB Protein Data Bank Piscataway, NJ, [2]RCSB Protein Data Bank Piscataway, NJ, [3]RCSB Protein Data Bank, [4]RCSB Protein Data Bank, Piscataway, NJ [5]RCSB Protein Data Bank, La Jolla, CA, [6]RCSB Protein Data Bank, La Jolla, CA, [7]RCSB Protein Data Bank, Piscataway, NJ, [8]Worldwide Protein Data Bank, Piscataway, NJ, [9]RCSB Protein Data Bank Piscataway, NJ

chenghua.shao@rcsb.org

Validation and Quality Assessment for Small-Molecule Ligands in the Protein Data Bank

1.RCSB Protein Data Bank, Rutgers, The State University of New Jersey, New Jersey, USA
2.RCSB Protein Data Bank, San Diego Supercomputer Center, University of California San Diego, California, USA
3.PDBe, EMBL-European Bioinformatics Institute, Hinxton, United Kingdom
4.PDBj, Institute for Protein Research, Osaka University, Osaka, Japan
5.EMDB, EMBL-European Bioinformatics Institute, Hinxton, United Kingdom
6.BMRB, UConn Health, University of Wisconsin, Wisconsin, USA

The Protein Data Bank (PDB) was established in 1971 as the first open-access digital data resource in biology with just seven X-ray crystallographic structures of proteins. Today, the single global archive houses more than 200,000 experimentally-determined three- dimensional (3D) structures of biological macromolecules (and their complexes with one another and small-molecule ligands) that are made freely available to many millions of PDB Data Consumers worldwide with no limitations on usage. This information facilitates basic and applied research and education across the sciences, impacting fundamental biology, biomedicine, biotechnology, and energy sciences. The archive is jointly managed by the Worldwide Protein Data Bank (wwPDB, wwpdb.org) partnership (consisting of five Full Members: RCSB PDB, PDBe, PDBj, BMRB, and EMDB; plus one Associate Member: PDB China). The wwPDB is committed to making PDB data Findable, Accessible, Interoperable, and Reusable (FAIR). The archive has been designated as a Global Core Biodata Resource by the Global Biodata Coalition and accredited by CoreTrustSeal.
Experimentally-determined structures deposited into PDB via the wwPDB OneDep software system by tens of thousands of Depositors (structural biologists) working on every inhabited continent undergo rigorous validation and expert biocuration. wwPDB Validation Reports provide useful metrics for evaluating geometric quality of structures and fit between atomic coordinates and experimental data coming from macromolecular crystallography (MX), nuclear magnetic resonance spectroscopy, and 3D electron microscopy. Official wwPDB Validation Reports are made public upon weekly release of new structures (typically ~300 every Wednesday at 00:00 Universal Time Coordinated), providing both simplified and expert views of quality assessment to PDB Data Consumers based in >200 sovereign countries recognized by the United Nations. PDB Depositors are strongly encouraged to submit confidential pre-release versions of wwPDB Validation Reports to scientific journals alongside manuscripts reporting new 3D biostructures. Since comprehensive validation was implemented within OneDep, the quality of newly-deposited PDB structures has shown gradual improvement.
Small-molecule ligands are present in >70% of the PDB structures, and many of them are designated by Depositors as a focus of the structural study (termed Ligand(s) of Interest, LOI). All incoming ligand structures are validated by examining goodness-of-fit between atomic coordinates and experimental data; deviations in bond lengths, bond angles, and torsion angles from known ideal values; inappropriate interatomic clashes between the ligand and surrounding biomolecules; and more. For MX co-crystal structures, the fit between ligand atomic coordinates and the experimental electron density map is scrutinized, particularly for Depositor-designated LOIs. Geometric quality of every ligand structure represented in the PDB is independently validated against accurate small- molecule X-ray crystal structures stored in the Cambridge Structural Database.
We recently analyzed quality indicators of ~643,000 ligands present in >100,000 PDB MX structures, and established composite ligand quality scores based on principal component analyses and ranking. Composite ligand quality scores are reported at the RCSB PDB research-focused web portal RCSB.org, enabling quantitative comparisons of quality for chemically identical ligands represented in distinct PDB structures with easy-to-interpret two-dimensional ligand quality plots. This information allows PDB Data Consumers to quickly assess ligand structure quality and select the best exemplars of each small- molecule ligand.
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.