Trends in macromolecular structure data across 50 years of the PDB

Irina Persikova[1], David Armstrong[2], John Berrisford[2], Minyu Chen[3], Gregg Crichlow[1], Genevieve Evans[2], Justin Flatt[1], Romana Gaborova[2], Sutapa Ghosh[1], Deepti Gupta[2], Deborah Harrus[2], Brian P. Hudson[1], Reiko Igarashi[3], Yumiko Kengaku[3], Ju Yaen Kim[3], Yuhe Liang[1], Ezra Peisach[1], Osman Salih[3], Junko Sato[3], Monica Sekharan[1], Chenghua Shao[1], James Tolchard[2], Jack Turner[2], Jasmine Young[1]

[1]RCSB Protein Data Bank, Rutgers, The State University of New Jersey, Piscataway, USA [2]Protein Data Bank in Europe, EMBL-European Bioinformatics Institute, Hinxton, UK [3]Protein Data Bank Japan, Institute for Protein Research, Osaka University, Japan

wwPDB.org

In 2021, we celebrate 50 years of the Protein Data Bank (PDB) archive - one of the longest-running open access scientific databases - managed collaboratively by the wwPDB (RCSB PDB, PDBe, PDBj, and BMRB). Throughout these 50 years, there has been significant evolution of the data in the PDB archive. This evolution has been driven by a number of factors including development in structural determination techniques, adaptation of biocuration practices, and increase in data capture via updated file formats. Here we present some of the key trends in data across the PDB archive, highlighting how structural biology data has changed over time and how wwPDB biocuration practices have adapted to handle these changes.

Fig.1 Number of Released PDB Structures

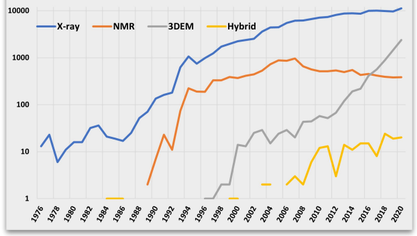


Fig.2 Median Structure Resolution (Å)

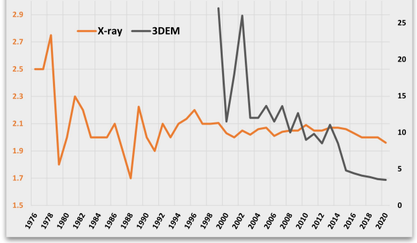


Fig.3 Number of X-ray Structures by Phasing Method

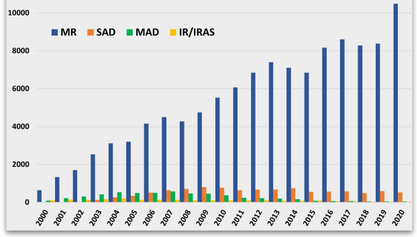
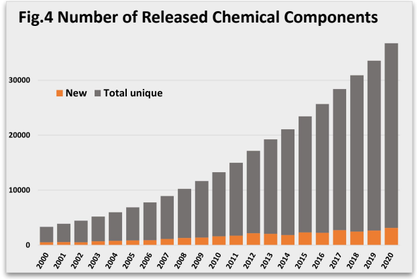
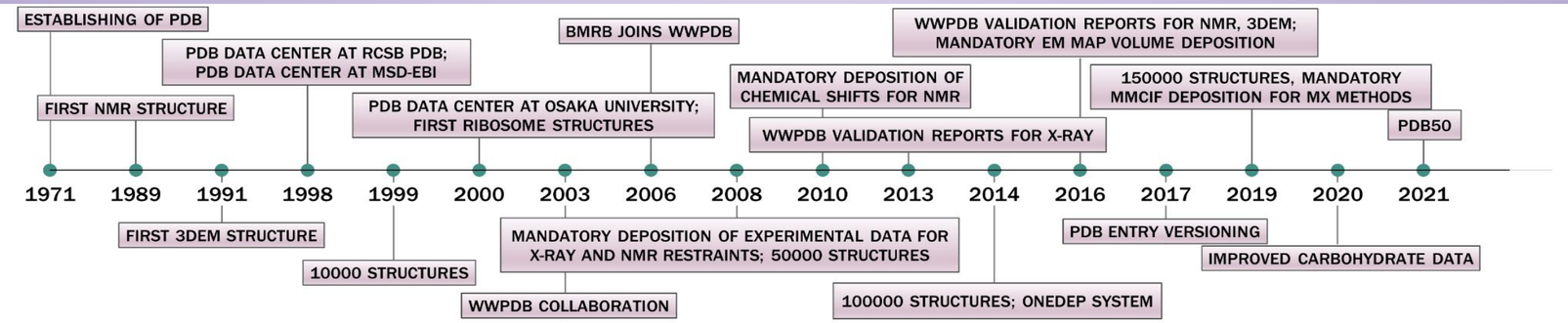


Fig.4 Number of Released Chemical Components

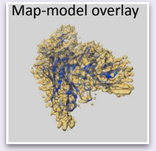


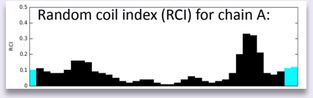


Supporting multiple experimental techniques

Experimental method specific:

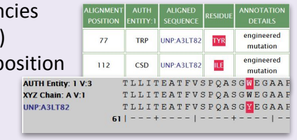
- Deposition interfaces
- Dictionaries (NMR: 244; EM: 613; X-FEL: 49 dictionary definitions)
- Extended data - anisotropic, unmerged
- Enumerations
- Validation (X-ray, NMR, EM)





Diverse sequences/sources

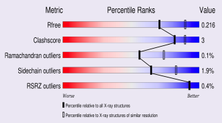
- Sequence cross-reference with UniProt /GB
- Curation of sequence discrepancies
- Taxonomy ids since 2007 (NCBI)
- Taxonomy enumerations at deposition



Data quality and reproducibility

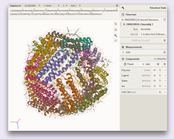
PDF validation reports since 2013:

- Required by the journals
- Overall quality at a glance
- Geometry and chemistry reports
- Model vs experimental data validation



Large structures and assembly annotation

- Large structures are handled as single files at deposition, curation and distribution since 2014
- Biological assembly is defined for each entry
- PISA is used to assess stable interfaces
- Annotated assembly verification in 3D
- Assembly files are available for download

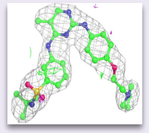


Archival data consistency and remediation

2007 - data uniformity across the archive
2008 - data enhancement, chemistry, assemblies, binding sites
2011 - biological assemblies, residual B factors, peptide inhibitors, nonstandard crystal frames
2013 - remediated structure factor files
2014 - integration of large structures
2016 - remediation of 3DEM entries
2017 - PDBx/mmCIF files to support OneDep
2020 - Carbohydrate remediation

Small molecules

- Chemical component definitions: stereochemical assignments, chemical descriptors (SMILES & InChI), systematic chemical names, idealized coordinates
- Validation reports: geometrical quality assessment, 2D graphical views of electron density fit



Supporting data uniformity, biocuration efficiency

- OneDep common deposition portal since 2014
- Curation efficiency and consistency (15436 structures curated in 2020)
- Provides enumerations and value limits
- Sequence, taxonomy, chemistry, model vs data, assembly annotation and validation

Looking towards the future

- Group depositions: multiple structures in one session
- PDB and CCD code extension
- Deposition APIs: Data harvesting pipeline from structure determination software packages
- PTM remediation
- Assembly curation improvements
- Validation improvements based on VTF recommendations

Fig.5 Unique Sequences (Cluster of 95% Identity)

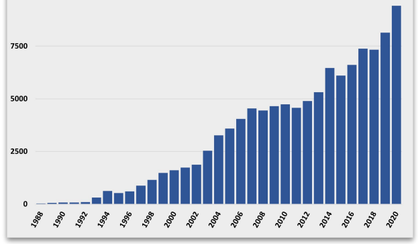


Fig.6 New and Cumulative Unique Sequences

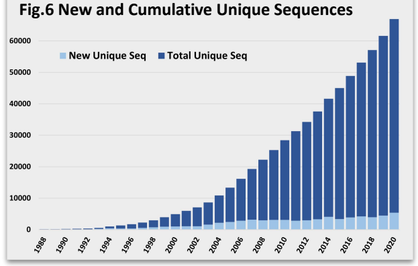


Fig.7 Mean Number of Atoms and Molecular Weight

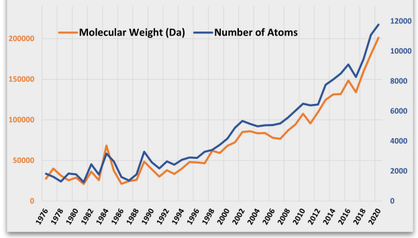
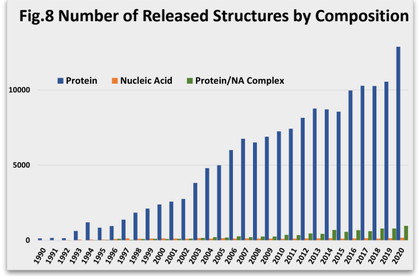


Fig.8 Number of Released Structures by Composition



Presenter contact: irina.persikova@rcsb.org