Protein Data Bank and File Formats

Vinay Kakkar

December 5, 2022

Contents

Abstract

1 Protein Data Bank

The Protein Data Bank was established at Brookhaven National Laboratories [BKW+77] in 1971 as an archive for biological macromolecular crystal structures [BWF+00].

Definition 1.1 (Macromolecular) *Macromolecular is any very large molecule, usually with a diameter ranging from about 100 to 10,000 angstroms*

It is a information source for data retreved from with atomic structures, crystallography and three-dimensional structures of biomolecules, including nucleic acids and proteins [BG21].

At the time this was the first open-access digital dataresource in biology which started with just seven protein structures [BBB+22b].

Various groups such as the Protein Data Bank in Europe, Protein Data Bank Japan help manage the Protein Data Bank archive. Current wwPDB members also include the ElectronMicroscopy Data Bank and the Biological Magnetic Resonance Bank [BBB+22b].

Protein Data Bank China has recently joined the wwPDB as an Associate Member with its role as wwPDBdesignated PDB Archive Keeper. Where they are responsible for weekly updates of the archive and safeguarding both digital information and a physical archive of correspondence [BBB+22a].

The management of PDB must comply with FAIR (the acronym depicts: Findable, Accessible, Interoperable, Reusable) and FACT [vdABH17] guiding principles for scientific data [WDA+16] [WSHB20].

The FAIR Guiding Principles	
To be Findable:	F1. (meta)data are assigned a globally unique and persistent identifier F2. data are described with rich metadata (defined by R1 below) F3. metadata clearly and explicitly include the identifier of the data it describes F4. (meta)data are registered or indexed in a searchable resource
To be Accessible:	A1. (meta)data are retrievable by their identifier using a standardized communications protoco A1.1 the protocol is open, free, and universally implementable A1.2 the protocol allows for an authentication and authorization procedure, where necessary A2. metadata are accessible, even when the data are no longer available
To be Interoperable:	I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation. I2. (meta)data use vocabularies that follow FAIR principles I3. (meta)data include qualified references to other (meta)data
To be Reusable:	R1. meta(data) are richly described with a plurality of accurate and relevant attributes R1.1. (meta)data are released with a clear and accessible data usage license R1.2. (meta)data are associated with detailed provenance R1.3. (meta)data meet domain-relevant community standards

Table 1: The guidlines to what builds up the FAIR principles [WDA+16]

1.1 Aims and Objectives of PDB

Enzymology, electron microscopy, computational chemistry small molecule crystallography, biochemistry, biophysics, macromolecular crystallography and nuclear magnetic resonance spectrometry all help the aims and goals of the PDB archive.

supports the structural biology as the front line aim and goal of the PDB archive [19] [BG21]

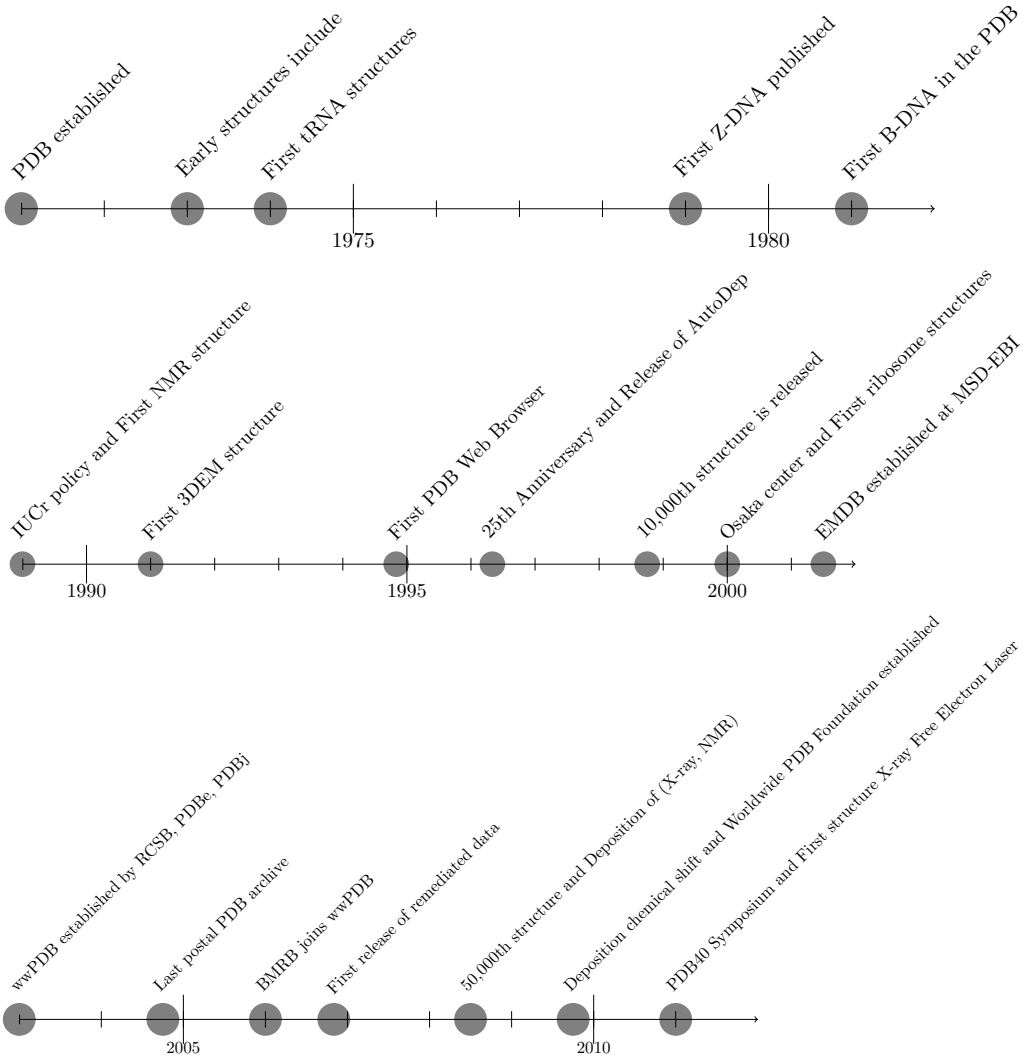
Definition 1.2 (Enzymology) *Enzymology is the branch of biochemistry aiming to understand how enzymes work*

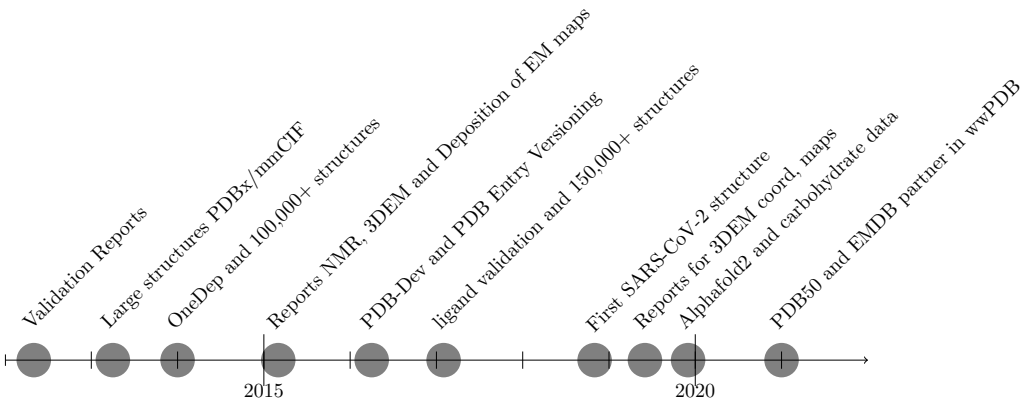
Definition 1.3 (Electron Microscopy) *Electron microscopy is a technique for obtaining high resolution images of biological and non-biological specimens.*

PDB's provides open access to nearly 200 000 archived, validated and biocurated experimentaly determined three-dimensional structures of biological macromolecules.3D structures archived in the PDB have enabled important scientific breakthroughs by basic and applied researchers [Bur21]. Open access to PDB data without restrictions on usage has also aided structural bioinformatics to areas such as computational biology.

1.2 Timeline of PDB

I have provided a timeline representing the milesstones achived within the protein data bank. Where PDB marked its 52st anniversary of continuous operations.





1.3 Recent Project

A project was undertaken to change the information management services for RCSB.org. The idea is was to have developed a primary place for studying 3D biostructures by extending RCSB.org web portal functionality to support parallel delivery of more than one million CSMs publiclyavailable from AlphaFold DB and ModelArchive together [BBB+22a].

1.4 Covid

During COVID-19 pandemic more than 2000 structures associated with the agent of the coronavirus disease were released and have become accessible for global users for free. The properties of these structures give us this opportunity to find out the ligand binding sites, spatial conformation of ligands, protein to protein interactions and amino acid substitutions regarding different viral proteins. Moreover, hemical, functional and energetic characteristics can also be gained from to describe the potential capabilities for each individual molecule. These properties might aid us to determine the potential drug targets for drug design and vaccine preparation [LZD+20].

Example case on why PDB is important

210 new molecular entities were discovered and developed during a period of 2010 to 2016 which were approved by the US Food and Drug Administration. The primary 3D structural data and information of these NMEs compartments, were first produced and released into the PDB archive. The representation of the related structures encouraged pharma companies to finance in drug discovery and development [WSHB20] [WB19] [BG21]

2 PDB Currently

As of 2022, the PDB has a vast number of 3D biostructures, eukaryotic protein structures exceeded 105 000. Bacterial protein structures were also numerous, totaling nearly 66 000. Archaeal protein structures were the least numerous totaling 5500. However the PDB coverage is decidedly limited, with mouse protein structures being most numerous at 8000 structures [BB21].

We have powerful tools developed by RCSB PDB for searching and analysis which include structure, sequence, sequence motif, structure motif, and visualization [BBB+22a].

Upon reaching the RCSB.org home page, users can query, organize, visualize, analyse, compare, and explore PDB structures and CSMs side-by-side. Searching 3D structure information can encompass PDB structures and CSMs or be limited to PDB structures only. Either PDB structures or CSMs can be excluded from the search results. The two types of structure information accessible via RCSB.org are clearly distinguished from each other. Top bar searching and data delivery for PDB structures and CSMs [BBB+22a].

The PDB Site currently

Figure 1: Search options at RCSB.org include Top Bar or Basic Search; Advanced Search; and Browse Annotations [BBB+22a].

(a) (b)

Figure 2: *Top Bar or Basic Search options available from every RCSB.org web page. Examples of searching for 3D structures using (A) simple text string insulin receptor; (B) drop down autosuggestions based on the text string insulin receptor; (C) Boolean operators to combine insulin + receptor (+ = AND); or (D) an amino acid sequence. (E) Searching RCSB.org documentation using a text string biological assembly [BBB+22a]*

The site has some key navigational features that provide users with access to Top Bar Search on the RCSB.org home page, Advanced Search, and Browse Annotations. The default option searches PDB structure data only. CSMs can be included with the toggle switch. Entering a keyword (e.g., molecule name, database entry ID (PDB, UniProt, AlphaFold DB, ModelArchive), author name (PDB structures only)) will launch autosuggestions organized by data category. Sequence searches can be run by entering single-letter code sequences for protein, DNA or RNA polymers and executing the query. Top Bar Search can also be used to search documentation and news announcements available

on both RCSB.org and the RCSB PDB outreach by changing the search type from 3D Structures to Documentation on the left of the search box [BBB+22a].

3 Recent Improvements

Recent RCSB PDB data architecture improvements

In 2020, RCSB PDB had a upgrade of its delivery architecture [RDL+21]at RCSB.org [BBB+21]. The legacy monolithic data delivery application was changed into a distributed deployment of individual microservices, each with a single responsibility. Data access services provide both Representational State Transfer and GraphQL API access to a data warehouse hosted in a MongoDB documentoriented database. Originaly, advanced Search QueryBuilder functionality encompassed text, PDB data attributes, 3D structure, sequence, biopolymer sequence motif, and chemical similarity. Every search function is implemented as an independent service. A separate search service is responsible for launching each search function, combining and delivering their integrated results to public programmatic search APIs. When each service has a single responsibility, we have a greater flexibility in scaling the deployment of services in response to changes in user load and significant reductions in the time required to develop, test, and deploy new features. The Sequence Motif search function has been extended with a new 3D Structure Motifsearch capability [BBR20].

Chemical Search function now has the ability to perform exhaustive substructure searching across the small molecules represented in the PDB archive, managed by the search Aggregator service and available through our public search APIs. To monitor the new service architecture, there is a system for processing service logs and indexing the time of request all whilst respecting PDB data consumer privacy. These statistics can be used to aid in monitoring the health of RCSB PDB production services [BBB+22b].

Recent advances in RCSB PDB data integration

RCSB PDB integrates the content of each expertly biocurated Entry with information from more than 50 external data resources.

Intergrated external data needs to follow a data schema that defines the organization of the RCSB PDB data warehouse. Finally it is available to RCSB PDB front end services, public data access APIs, and our text search indexing service [BBB+22b].

Recent PDBx/mmCIF data standard improvements

The PDBx/mmCIF data standard is maintained by the wwPDB organization in collaboration with wwPDBPDBx/mmCIF Working Group domain experts recruited from the scientific community. The PDBx/mmCIF web resource supports browse and search access to standard terminology. The Working Group includes developers for many of the widely used structure determination software systems, who ensure that data produced by these programs comply with the PDBx/mmCIF data standard, generating complete and correct data files for PDB deposition. The wwPDB and the Working Group collaborate on developing terminologies for new and rapidly evolving methodologies such as Free Electron Laser, 3DEM, Serial Crystallography, and X-ray, whilst improving representations for existing data content. Most recently, the Working Group has focused on modernizing content descriptions for processed X-ray diffraction data, including extensions describing anisotropic diffraction limits, unmerged reflection data, and new quality metrics of anomalous diffraction data. Deposition and delivery improve our ability to assess experimental data quality, and every PDB data consumer's ability to Find and Reuse relevant PDB Entries [BBB+22b].

Mol* molecular graphics visualization

PDB data provides 3D shapes and interactions of biological macromolecules where we ultimatly would like to understand the biological function.

Mol* was deployed in 2020 [SBD+21] and set as the default RCSB.org molecular graphics tool for visualizing. This software system was developed as a community project, co-led by RCSB PDB and the Protein Data Bank in Europe.

External Resources	
AlphaFold DB	Computed Structure Models by AlphaFold2
ATC	Anatomical Therapeutic Chemical (ATC) Classification System from World Health Organization
Binding MOAD	Binding affinities
Binding DB	Binding affinities
BMRB	BMRB-to-PDB mappings
Cambridge structural Database	Crystallographic small molecule data from the Cambridge Crystallographic Data Centre
CATH	Protein structure classification- Class, Architecture, Topology/fold, and Homologous superfamily
ChEMBL	Manually curated database of bioactive molecules with drug-like properties
CSD	Cambridge Structural Database: Validated and curated small-molecule organic and metal-organic crystal structures from the Cambridge Crystallographic Data Centre
DrugBank	Drug and drug target data
ECOD	Evolutionary Classification of Protein Domains
EMDB	3DEM density maps and associated metadata
ExplorEnz	IUBMB Enzyme nomenclature and classification
Gencode	Human and Mouse Gene annotations

Table 2: Some of the External Resources Integrated Into RCSB PDB

Mol* graphics tool is used for interrogating 3D macromolecular structure data from the PDB or computed structure models. It works entirely within the PDB site. Mol* also supports integration of information from other bioinformatics resources. These insights can help develop new hypotheses for research and facilitate analysis. Overall it allows to easily visualize entire polypeptide or nucleic acid chains, whole biological Assemblies, or specific atoms or groups of atoms ina particular biological macromolecule it can also display molecular surfaces, and noncovalent interactions with bound ligands, ions, drugs, and inhibitors [BBB+22b].

4 Summary

4.1 Future and struggles of PDB

Future

As the PDB archive has started its 52nd year, it gives open acess to analyses of strucutres and much more to: basic and applied researchers, educators, and students spanning fundamental biology, biomedicine, bioenergy, bioengineering and biotechnology, with key points that help many communities

that use this ficlity. Firstly It delivers Data In and Data Out services efficiently to a user base that is now numbering many millions worldwide. Secondly it has wwPDB partners that process, validate, and biocurate the growing number of increasingly complex PDB depositions received. Manages and safeguards the growing PDB archive in its role as wwPDB designated Archive Keeper. Thirdly it enables searching, visualization, exploration and analysis of experimentally-determined PDB structures integrated with more than one million CSMs through its web portal. Lastly it supports user training, education, and outreach through PDB101.RCSB.org. [BBB+22a].

Struggles

Even after all the advancments PDB has gone through there are still additional challenges lying ahead which include:

- Rapid growth in public-domain CSMs of individual polypeptide chains, already numbering ¿200 million at the time of writing.
- Anticipated advances in AI/ML-based prediction of structures of multi-protein complexes.
- Continued development of biomolecular structure determination methods using X-ray Free Electron Lasers, revealing the microscopic details of chemical reactions in real time.
- Growth in the number and complexity of atomic-level cryoelectrontomography structures of macromolecular machines.
- Integration of PDB structures and CSMs with complementary information coming from correlative light microscopy and related imaging methods across length scales ranging from atoms to small molecules to individual biomolecules to macromolecular assemblies to organelles to cells and ultimately tissues
- Merging of the PDB-Dev prototype archiving system for integrative methods structures with the PDB archive
- Federating other biodata resources, such as the SmallAngle Scattering Database and the Proteomics Identification Database, with the PDB, EMDB and BMRB core archives jointly managed by the wwPDB partnership

[BBB+22a].

5 File Formats

The PDB archive holds a few different types of file types which hold data such as atomic coordinates and other information describing proteins and other biological macromolecules. Depending on what the data is created from it can fall into a different category.

5.1 PDB Data

The main information in the PDB archive is coordinate files for biological molecules. These files list the atoms in each protein, and their 3D coordinates.

These files are available in several formats:

- PDB
- mmCIF
- XML

The header section of text summarizes the protein, citation information, and the details of the structure solution, which is then followed by the sequence and a long list of the atoms and their coordinates. It also contains the experimental observations used to deterimne atomic coordinates [Goo].

.pdb Files

The PB format consists of a collection of fixed format records that describe the atomic coordinates, chemical and biochemical features, experimental details of the structure determination. and some structural features such as secondary structure assignments. hydrogen bonding, and biological assemblies and active sites [WF03].

Each item of data in the PDB format is assigned to a range of character positions in one of many PDB record types (HEADER. SOURCE. REMARK, etc.). The ATOM records encode the atomic coordinate data. ATOM records are among the more than 45 named data records in the PDB format. These named has records have suit column-formatting rules [WF03].

In addressing changes in experimental methodology, the PDB format has been extended with new REMARK records. For example, the organization and information content of REMARK 3 that encodes refinement information has been modified and extended for each new refinement program and program version Although extending REMARK records in this way captures information in a manner that is easy for a human to read, the diversity of organization of this data makes it very difficult to design software that can automatically and reliably extract information from these records Data m these records is also defined only in terms of the program that computed the information [WF03].

The PB format uses fixed-width fields to represent data, and this restriction places absolute limits on the size of certain items of data. For instance, the maximum number of atom records that can be represented in a single structure model is limited to 99.999 and the field width of the identifier for polymer chains is limited to a single character.Although these restrictions were certainly reasonable when the format was first defined, this is no longer the case. Many large molecular systems, such as the ribosomal subunit structures, cannot be represented in a single PDB entry. These entries must be divided into multiple PB files, seriously complicating their use [WF03].

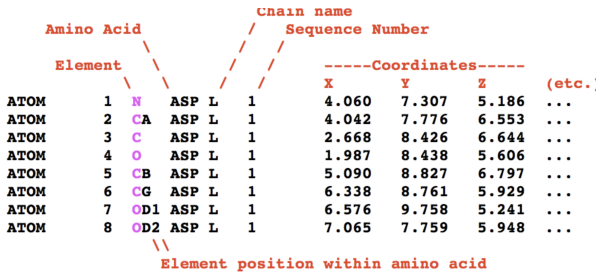


Figure 3: Showing contents of a PDB file for the Atom values [AAB+19].

.mmCIF Files

Mmcif is a dictionary based approach to data about crystallographic experiments and results and is the format in which all structures described in articles sent to Act Crystallographic C are submitted [WF03].

Implicitly this mandate included the need to describe all the data items included in a PDB entry. Also the need to provide sufficient data names so that the experimental section of a structure paper could be written automatically and to facilitate the development of tools so that computer programs could easily access and validate mmCIF data files [WF03].

In January 1997, the mmCIF dictionary containing 1700 definitions was completed and submitted to the IUCr committee that oversees dictionary development (COMCIFS) for review and in June 1997. Version 1.0 was released [WF03].

.xml

The representation of PDB data in XML builds from the content of the PDB Exchange dictionary, both for assignment of data item names and for defining data organization. Although presented in very different syntaxes, the PDB Exchange and XML representations use the same logical data organization.

The dictionary data block is mapped to the standard top-level XML schema element, and the data file data block is mapped to a datablock element. The schema and datablock elements provide namespace definitions, linkages to the supporting XML schema definition documents and linkages

to the location of the current supporting schema. Category or table definitions in the Exchange dictionary are described as XML complexTypes. The category definition and examples are mapped to XML annotation and documentation elements. The data items within the category are defined as an unordered sequence of XML elements named according to the attribute portion of their dictionary equivalents. The special data items that form the primary key for the category are defined as XML attributes.

PDB Exchange data dictionary attributes	XML schema mapping
Data block	Root level *schema element*
Category groups	
Categories	*complexTypes*
Definition	*annotation* and *documentation* elements
Examples	*annotation* and *documentation* elements
Primary keys	*attributes* of the data category
Items	*elements* of the data category
Definition	*annotation* and *documentation* elements
Examples	annotation and documentation elements
Data types	*simpleTypes*
Range restrictions and allowed values	*restrictions* within *simpleTypes* or *unions* of *simpleTypes*
Mandatory data code	Element attributes *minOccurs* and *nillable*
Parent-child relationships	*key/keyref* elements
Interdependency/exclusivity	
Units of measurement	Additional *fixed attributes*
Subcategory membership	

Figure 4: Summary of the correspondences between PDB Exchange data dictionary and XML schema metadata [WIN+05].

5.2 Visualizing Structures

While you can view PDB files directly using a text editor, Pdb files can be viewed from text editors but we can also use a browsing or visualization program. RCSB PDB allow you to search and explore the information, including information on experimental methods and the chemistry and biology of the protein. Visualization programs allow to read in the PDB file and, display the protein structure generating custom pictures of it. These programs can contain analysis tools that allow you to measure distances and bond angles, and identify interesting structural features [Goo].

5.3 Reading Coordinate Files

Before exploring structures in the PDB archive we need some prior understanding of the coordinate files. For example we can find a diverse mixture of biological molecules, small molecules, ions, and water which can get confusing we can use he names and chain IDs to help sort these out. In structures determined from crystallography, atoms are annotated with temperature factors that describe their vibration and occupancies that show if they are seen in several conformations. NMR structures often include several different models of the molecule [Goo].

5.4 Potential Challanges

There are some things to note as you could fall into some challanges when browsing through the PDB archive. Many structures, particular those determined by crystallography, only include information about part of the functional biological assembly. One thing to note is that the PDB can aid with this. Another note is many PDB entries are missing portions of the molecule that were not observed in the experiment. These include structures that include only alpha carbon positions, structures with missing loops, structures of individual domains, or subunits from a larger molecule. In addition, most of the crystallographic structure entries do not have information on hydrogen atoms [Goo].

Bibliography

[AAB+19] Paul D. Adams, Pavel V. Afonine, Kumaran Baskaran, Helen M. Berman, John Berrisford, Gerard Bricogne, David G. Brown, Stephen K. Burley, Minyu Chen, Zukang Feng, Claus Flensburg, Aleksandras Gutmanas, Jeffrey C. Hoch, Yasuyo Ikegawa, Yumiko Kengaku, Eugene Krissinel, Genji Kurisu, Yuhe Liang, Dorothee Liebschner, Lora Mak, John L. Markley, Nigel W. Moriarty, Garib N. Murshudov, Martin Noble, Ezra Peisach, Irina Persikova, Billy K. Poon, Oleg V. Sobolev, Eldon L. Ulrich, Sameer Velankar, Clemens Vonrhein, John Westbrook, Marcin Wojdyr, Masashi Yokochi, and Jasmine Y. Young. Announcing mandatory submission of PDBx/mmCIF format files for crystallographic depositions to the Protein Data Bank (PDB). *Acta Crystallographica Section D Structural Biology*, 75(4):451–454, April 2019.

[BB21] Stephen K. Burley and Helen M. Berman. Open-access data: A cornerstone for artificial intelligence approaches to protein structure prediction. *Structure (London, England: 1993)*, 29(6):515–520, June 2021.

[BBB+21] Stephen K. Burley, Charmi Bhikadiya, Chunxiao Bi, Sebastian Bittrich, Li Chen, Gregg V. Crichlow, Cole H. Christie, Kenneth Dalenberg, Luigi Di Costanzo, Jose M. Duarte, Shuchismita Dutta, Zukang Feng, Sai Ganesan, David S. Goodsell, Sutapa Ghosh, Rachel Kramer Green, Vladimir Guranović, Dmytro Guzenko, Brian P. Hudson, Catherine L. Lawson, Yuhe Liang, Robert Lowe, Harry Namkoong, Ezra Peisach, Irina Persikova, Chris Randle, Alexander Rose, Yana Rose, Andrej Sali, Joan Segura, Monica Sekharan, Chenghua Shao, Yi-Ping Tao, Maria Voigt, John D. Westbrook, Jasmine Y. Young, Christine Zardecki, and Marina Zhuravleva. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Research*, 49(D1):D437–D451, January 2021.

[BBB+22a] Stephen Burley, Charmi Bhikadiya, Chunxiao Bi, Sebastian Bittrich, Henry Chao, Li Chen, Paul Craig, Gregg Crichlow, Kenneth Dalenberg, Jose Duarte, Shuchismita Dutta, Maryam Fayazi, Zukang Feng, Justin Flatt, Sai Ganesan, Sutapa Ghosh, David Goodsell, Rachel Kramer, Vladimir Guranovic, and Christine Zardecki. RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Research*, November 2022.

[BBB+22b] Stephen K. Burley, Charmi Bhikadiya, Chunxiao Bi, Sebastian Bittrich, Li Chen, Gregg V. Crichlow, Jose M. Duarte, Shuchismita Dutta, Maryam Fayazi, Zukang Feng, Justin W. Flatt, Sai J. Ganesan, David S. Goodsell, Sutapa Ghosh, Rachel Kramer Green, Vladimir Guranovic, Jeremy Henry, Brian P. Hudson, Catherine L. Lawson, Yuhe Liang, Robert Lowe, Ezra Peisach, Irina Persikova, Dennis W. Piehl, Yana Rose, Andrej Sali, Joan Segura, Monica Sekharan, Chenghua Shao, Brinda Vallat, Maria Voigt, John D. Westbrook, Shamara Whetstone, Jasmine Y. Young, and Christine Zardecki. RCSB Protein Data Bank: Celebrating 50 years of the PDB with new tools for understanding and visualizing biological macromolecules in 3D. *Protein Science*, 31(1):187–208, 2022. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.4213.

[BBR20] Sebastian Bittrich, Stephen K. Burley, and Alexander S. Rose. Real-time structural motif searching in proteins using an inverted index strategy. *PLOS Computational Biology*, 16(12):e1008502, December 2020. Publisher: Public Library of Science.

[BG21] Payam Behzadi and Márió Gajdács. Worldwide Protein Data Bank (wwPDB): A virtual treasure for research in biotechnology. *European Journal of Microbiology and Immunology*, 11(4):77–86, December 2021. Publisher: Akadémiai Kiadó Section: European Journal of Microbiology and Immunology.

[BKW+77] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The Protein Data Bank: a computer-based archival file for macromolecular structures. *Journal of Molecular Biology*, 112(3):535–542, May 1977.

[Bur21] Stephen K. Burley. Impact of structural biologists and the Protein Data Bank on small-molecule drug discovery and development. *The Journal of Biological Chemistry*, 296:100559, 2021.

[BWF+00] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, January 2000.

[Goo] David S. Goodsell. PDB101: Learn: Guide to Understanding PDB Data: Introduction.

[LZD+20] Joseph H. Lubin, Christine Zardecki, Elliott M. Dolan, Changpeng Lu, Zhuofan Shen, Shuchismita Dutta, John D. Westbrook, Brian P. Hudson, David S. Goodsell, Jonathan K. Williams, Maria Voigt, Vidur Sarma, Lingjun Xie, Thejasvi Venkatachalam, Steven Arnold, Luz Helena Alfaro Alvarado, Kevin Catalfano, Aaliyah Khan, Erika McCarthy, Sophia Staggers, Brea Tinsley, Alan Trudeau, Jitendra Singh, Lindsey Whitmore, Helen Zheng, Matthew Benedek, Jenna Currier, Mark Dresel, Ashish Duvvuru, Britney Dyszel, Emily Fingar, Elizabeth M. Hennen, Michael Kirsch, Ali A. Khan, Charlotte Labrie-Cleary, Stephanie Laporte, Evan Lenkeit, Kailey Martin, Marilyn Orellana, Melanie Ortiz-Alvarez de la Campa, Isaac Paredes, Baleigh Wheeler, Allison Rupert, Andrew Sam, Katherine See, Santiago Soto Zapata, Paul A. Craig, Bonnie L. Hall, Jennifer Jiang, Julia R. Koeppe, Stephen A. Mills, Michael J. Pikaart, Rebecca Roberts, Yana Bromberg, J. Steen Hoyer, Siobain Duffy, Jay Tischfield, Francesc X. Ruiz, Eddy Arnold, Jean Baum, Jesse Sandberg, Grace Brannigan, Sagar D. Khare, and Stephen K. Burley. Evolution of the SARS-CoV-2 proteome in three dimensions (3D) during the first six months of the COVID-19 pandemic. *bioRxiv: The Preprint Server for Biology*, page 2020.12.01.406637, December 2020.

[RDL+21] Yana Rose, Jose M. Duarte, Robert Lowe, Joan Segura, Chunxiao Bi, Charmi Bhikadiya, Li Chen, Alexander S. Rose, Sebastian Bittrich, Stephen K. Burley, and John D. Westbrook. RCSB Protein Data Bank: Architectural Advances Towards Integrated Searching and Efficient Access to Macromolecular Structure Data from the PDB Archive. *Journal of Molecular Biology*, 433(11):166704, May 2021.

[SBD+21] David Sehnal, Sebastian Bittrich, Mandar Deshpande, Radka Svobodová, Karel Berka, Václav Bazgier, Sameer Velankar, Stephen K Burley, Jaroslav Koča, and Alexander S Rose. Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Research*, 49(W1):W431–W437, July 2021.

[vdABH17] Wil M. P. van der Aalst, Martin Bichler, and Armin Heinzl. Responsible Data Science. *Business & Information Systems Engineering*, 59(5):311–313, October 2017.

[WB19] John D. Westbrook and Stephen K. Burley. How Structural Biologists and the Protein Data Bank Contributed to Recent FDA New Drug Approvals. *Structure (London, England: 1993)*, 27(2):211–217, February 2019.

[WDA+16] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018, March 2016. Number: 1 Publisher: Nature Publishing Group.

[WF03] John D. Westbrook and Paula M. D. Fitzgerald. The PDB Format, mmCIF Formats, and Other Data Formats. In *Structural Bioinformatics*, pages 159–179. John Wiley & Sons, Ltd, 2003. Section: 8 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/0471721204.ch8.

[WIN+05] John Westbrook, Nobutoshi Ito, Haruki Nakamura, Kim Henrick, and Helen M. Berman. PDBML: the representation of archival macromolecular structure data in XML. *Bioinformatics*, 21(7):988–992, April 2005.

[WSHB20] John D. Westbrook, Rose Soskind, Brian P. Hudson, and Stephen K. Burley. Impact of the Protein Data Bank on antineoplastic approvals. *Drug Discovery Today*, 25(5):837–850, May 2020.