Impact of the Protein Data Bank Across Scientific Disciplines

Zukang Feng[1,2], Natalie Verdiguel[3], Luigi Di Costanzo[1,4], David S. Goodsell[1,5], John D. Westbrook[1,2], Stephen K. Burley,[1,2,6,7,8] Christine Zardecki[1,2 *],

[1] Research Collaboratory for Structural Bioinformatics Protein Data Bank,
Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA
[2] Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA
[3] University of Central Florida, Orlando, Florida, 32816 USA
[4] Current address: Department of Agricultural Sciences, University of Naples Federico II, Portici 80055, Italy
[5] Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA 92037, USA
[6] Research Collaboratory for Structural Bioinformatics Protein Data Bank, San Diego Supercomputer Center, University of California, San Diego, La Jolla, CA 92093, USA
[7] Rutgers Cancer Institute of New Jersey, Rutgers, The State University of New Jersey, New Brunswick, NJ 08903, USA
[8] Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, CA 92093, USA
*Corresponding author

AUTHORS' CONTRIBUTIONS:

A sentence or a short paragraph detailing the roles that each author held to contribute to the authorship of the submission. Individuals listed must fit within the definition of an author, as per our authorship guidelines.

Zukang Feng and Natalie Verdiguel contributed to the design and analysis of the data set and drafted the manuscript.

John Westbrook performed the initial citation analysis.

David Goodsell contributed to the analysis and interpretation of the data and revised the manuscript.

Christine Zardecki and Stephen K. Burley designed the study and finalized and approved the manuscript.

Abstract

The Protein Data Bank archive (PDB) was established in 1971 as the 1st open access digital data resource for biology and medicine. Today, the PDB contains >160,000 atomic-level, experimentally-determined 3D biomolecular structures. PDB data are freely and publicly available for download, without restrictions. Each entry contains summary information about the structure and experiment, atomic coordinates, and in most cases, a citation to a corresponding scientific publication. Individually and in bulk, PDB structures can be downloaded and/or analyzed and visualized online using tools at RCSB.org. As such, it is challenging to understand and monitor reuse of data. Citations of the scientific publications describing PDB structures provides one way of understanding which structures are being used, and in which research areas. Our analysis highlights frequently-cited structures and identifies milestone structures that have demonstrated impact across scientific fields.

Up to 6 keywords: Structural biology, open access, open science, citation patterns, Interdisciplinary

Introduction

Since 1971, the Protein Data Bank (PDB) has served the scientific community as the single, global repository for structural data of biomolecules (Protein Data Bank, 1971). Data archived at the PDB include atomic coordinates and related experimental data from macromolecular crystallography, nuclear magnetic resonance spectroscopy and 3D electron microscopy studies. Understanding these 3D structures of proteins, nucleic acids, and large molecular machines informs our understanding of fundamental biology, medicine and drug discovery, and energy.

The PDB was conceived as a resource for the crystallographic community, to archive their primary results. However, as the number of structures grew, it became apparent that this body of information would have much wider application. Communities of researchers emerged that focused on data mining, using the available structures to hypothesize and test overarching principles of biomolecular structure, folding, and function. Soon after, the archive showed growing application in the field of structure-guided drug design, and has since been instrumental in the discovery and development of dozens of blockbuster medical treatments (Westbrook and Burley, 2019). In addition, structures from the archive are used widely to provide structural understanding of biomolecular structure and function, promoting research in many fields of biology, but also in chemistry, physics, mathematics, computer science and beyond. The growing utility of the archive naturally lead to widespread policies of structure deposition, and today, most major journals require release of coordinates when publishing reports of structural studies (Young et al., 2018), ensuring that the results of structural research are available for these many derivative disciplines.

Today, the PDB is an established archive with >160,000 entries, which have been extensively curated for consistency and accuracy. In this report, we use an analysis of citations to reveal the impact of structural biology and the general availability of these atomic structures of biomolecules in diverse scientific communities.

The wwPDB archive

The PDB archive began in 1971 with seven structures, which were distributed by request on magnetic tape (Protein Data Bank, 1971). In subsequent years, the rapid growth of the archive and development of worldwide computational infrastructure required a more comprehensive approach to deposition, curation, and dissemination of the archive. Since 2003, the Worldwide PDB (wwPDB) organization has managed the PDB archive and ensured that PDB data are freely and publicly available (Berman et al., 2003, wwPDB consortium, 2019) following the FAIR principles (Wilkinson et al., 2016). Locally-funded, regional PDB Data Centers in the US (Berman et al., 2000), Europe (Velankar et al., 2016), and Japan (Kinjo et al., 2018) safeguard and disseminate PDB structures using a common data dictionary (Fitzgerald et al., 2005) and a unified global system for data deposition-validation-biocuration (Young et al., 2017).

The Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) (Berman et al., 2000, Burley et al., 2019) has served as the US PDB Data Center since 1999. RCSB PDB manages deposition and curation of roughly 42% of new structures, and free dissemination of the archive through the comprehensive RCSB PDB website. The PDB archive and RCSB.org are heavily used: during 2019, >838 million structure data files were downloaded from the archive. The RCSB PDB website also provides extensive value added through resources for visualization and analysis, integration with ~40 external resources, as well as a sister website, PDB-101, targeted for educational and outreach communities. Together, these resources provide rich structural views of fundamental biology, biomedicine, and energy sciences, which are accessed by millions of users from around the world and from a wide range of scientific disciplines.

Previous work on the impact of data reuse of the PDB archive

Evaluation of the impact of data archives is important for data depositors, data users, and for resource management, planning, and funding. For scientific databases, citations are often used as a tangible expression of reuse of the data. Since these publications are peer-reviewed and from reputable journals, it lends confidence that the derivative work is contributing to the growing body of scientific knowledge and validates the role of the data archive as a central resource for the community.

Previous citation analyses of the PDB archive have focused on the inaugural article describing the RCSB PDB resource, "The Protein Data Bank" (Berman et al., 2000) that appeared in *Nucleic Acids Research*. This inaugural article is regularly used to cite both the PDB data archive and RCSB PDB services. This reference is useful to study due to its high volume of citations. A 2014 analysis (Van Noorden et al., 2014) ranked the inaugural article 92nd among the top 100 most-cited research publications of all time and a 2017 study (Basner, 2017) placed it 5th among papers published since 2000. The 2017 analysis by Basner also found, using internal methods for normalizing across category, that articles citing the inaugural RCSB PDB publication had a citation-based impact exceeding the world-average in 16 scientific fields including Biology & Biochemistry, Computer Science, Plant & Animal Sciences, Physics, Environment/Ecology, Mathematics and Geosciences. Another study (Markosian et al., 2018) found the research areas for articles citing the inaugural RCSB PDB publication are changing over time, with more recent growth in disciplines such as Mathematical Computational Biology, Chemistry Medicinal, and Computer Science Interdisciplinary Applications.

Other studies have looked at how individual structures are referenced in the literature to demonstrate data reuse. For example, Huang et al. found an increase in the number citations to PDB entries by URL rather than to publication (Huang et al., 2015), and Bousfield et al. cross-referenced open access literature with the PDB archive and found the average annual number of citations for a PDB structure is 6.7 (Bousfield et al., 2016). Scientific articles that are connected to open access data have been shown to be more highly cited than articles without data being made available (Colavizza et al., 2019). This is certainly reflected in the primary citations included in PDB entries. At the time the data for this study were collected (March 1, 2018), the PDB archive contained ~139,000 structures, and the Basner 2017 study found that the PDB archive from 2000-2016 had been cited by more than 1 million scientific publications in the Web of Science, giving an average number of ~40 citations per PDB structure publication (Burley et al., 2018). This number rises to ~80 citations per PDB structure publication for drug targets in all therapeutic areas (Westbrook and Burley, 2019).

This study explores citation patterns of individual PDB structures to identify PDB entries of most interest in specific fields and to examine trends in the application of structural biology. Each structure represents the results of an experiment determined by a laboratory and then deposited to the PDB data, biocurated by the wwPDB, and then made publicly available in the archive (Morris, 2018). The majority of PDB structures (~80%) have a corresponding primary citation that is the first paper to describe the molecule, its structure, and its function. Public release of most PDB data (87% in 2018) is coordinate with the time of publication of this primary citation. As identified by previous research, and due to the nature of the data, most papers citing PDB structures are in the field of biology. To identify strong examples of structures cited in other research categories, we identified the top cited structures within related disciplines.

Methods

For each entry in the PDB archive, we analyzed the set of articles that cite the primary citation of the PDB entry. First, the primary citations for each PDB entry were exported from the RCSB PDB database. A single primary citation may describe multiple PDB structures—in these cases, entries were treated and counted separately in the analysis. Then, publication data for articles that cite these PDB primary citations as of March 2018 were exported and organized by subject categories with Web of Science (Clarivate Analytics, 2019). Related subject categories were aggregated, for example, the Chemistry category reported here includes Chemistry Physical, Chemistry Organic, Chemistry Analytical, and others. Note that each publication can be assigned to more than one category. The ranking of citation impact across disciplines and longitudinally is an ongoing area of active bibliometric research (Abaci, 2017, Bronmann and Williams, 2020, Diamandis, 2017, Koelblinger et al., 2019, Pendlebury, 2009, Jesper W. Schneider et al., 2019). For this study, the citation data are not normalized in any way to provide a direct comparison of impact between categories. Exported data were analyzed in July 2018 and have been submitted to Dryad at https://doi.org/10.5061/dryad.4tmpg4f5v.

Results and Discussion

Overall citation of PDB structures

The top-cited PDB structures (Table 1) are landmark structures that signal achievements in fundamental biology and their application to biomedicine and biotechnology. A detailed description of the impact of the structures of the nucleosome (PDB 1aoi) and major histocompatibility complex 1 (1hla) has been published (Burley et al., 2018). The structure of bacteriorhodopsin was the first EM structure released in the PDB, and represents a ground-breaking use of electron crystallography of two-dimensional sheets to determine the membrane-bound structure of a protein (1brd). The structure of the F1 portion of ATP synthase (1bmf) revealed the atomic details of the rotary molecular motor, providing a structural explanation for decades of biochemical studies. Similarly, the structure of the potassium channel resolved a long-standing question about the nature of specificity, revealing the central role of hydration and dehydration of ions in controlling ion passage across the cell membrane. The structures of MHC I (1hla), MDM2 (1rv1), and serum albumin (1uor) are a testament to the utility of atomic structures in the understanding of biomedically-important biomolecules and in structure-based design of pharmaceuticals. Many additional milestone structures closely follow these top 10 entries, including photosystem II (1s5l) with 2286 citations and green fluorescent protein (1ema) with 1529 citations. In keeping with the importance of these molecules, all structures in the top-cited list (Figure 1) have been highlighted in the RCSB PDB's *Molecule of the Month* series at PDB101.rcsb.org (Goodsell et al., 2019).

The importance of these 10 entries is also supported by prominence of the journals where the primary citation was published. Of the nine articles describing the top 10 structures, four were published in *Nature*, four in *Science*, and one in the *Journal of Molecular Biology*. The oldest structure was published in 1987, and the most recent in 2007. Surprisingly, the initial 12 structures archived in the PDB since the 1970s and that represent the seminal achievements of structural biology are not included in this list. This may be due in part to inconsistent citation practices of the time, both for the citations that were included in the early structure deposition to the archive, and for how PDB structures were cited in the literature. For example, the citation included in the entry for Kendrew's landmark structure of myoglobin (1mbn) is not the initial structure solution (Kendrew et al., 1960), which currently shows >1100 citations, but rather a later report of the molecule (Watson, 1969) that is not included in the Web of Science. Similarly, multiple publications were presented over decades during the structure solution of hemoglobin (2dhb), including the primary citation associated with the entry (Bolton and Perutz, 1970) with 253 citations and a key *Nature* paper with 932 citations (Perutz et al., 1960).

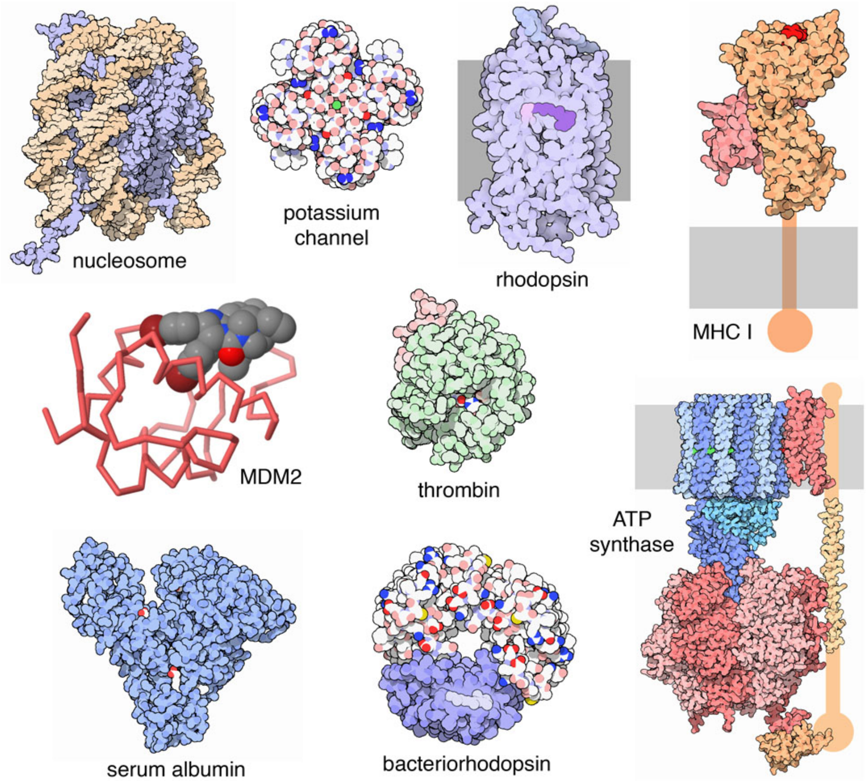

Figure 1. Images of the most highly-cited PDB structures, taken from PDB-101, the educational portal of the RCSB PDB. In several cases, a highly-cited structure entry was the example used in the *Molecule of the Month* article.

Table 1. Top-cited PDB Structure Primary Citations as of March 1, 2018

Structure	PDB ID	Journal	Structure Primary Citation	Times Cited
Nucleosome	1aoi	Nature	(Luger et al., 1997)	4,927
Potassium channel	1bl8	Science	(Doyle et al., 1998)	4,695
Bacteriorhodopsin	1brd	Journal of Molecular Biology	(Henderson et al., 1990)	4,298
Rhodopsin	1f88	Science	(Palczewski et al., 2000)	4,237
Major histocompatibility class I	1hla	Nature	(Bjorkman et al., 1987)	3,081
MDM2/imidazoline inhibitor	1rv1	Science	(Vassilev et al., 2004)	2,649
Thrombin	2v3o/2v3h	Science	(Muller et al., 2007)	2,596
Serum albumin	1uor	Nature	(He and Carter, 1992)	2,552
ATP Synthase F1	1bmf	Nature	(Abrahams et al., 1994)	2,453

Top cited structures by category

We also analyzed subject categories for the journals where citing articles were published, to assess the range of disciplines where data from the PDB archive is having impact. Not surprisingly, PDB structures are most cited by publications with the subject category Biochemistry Molecular Biology, including 101,921 unique structures (72% of archive) at the time of this study. Six non-biological categories were chosen to show utility of the archive in related disciplines, including Materials Science, Physics, Computer Science, Chemistry, Engineering, and Mathematics.

We also identified the most highly-cited article that cited a PDB structure in each category. For the physics-related categories these highly-cited papers included reviews related to biotechnology and nanotechnology: Materials Science (Nel et al., 2009), Physics (Zweib et al., 1989), and Engineering (Hersel et al., 2003). For Computer Science, Chemistry, and Mathematics, the papers were primary citations for the widely-used molecular visualization program VMD (Humphrey et al., 1996), small and macromolecular structure determination program SHELX (Sheldrick, 2008), and database of theoretical models SWISS-MODEL (Arnold et al., 2006), respectively. These citations highlight how available data in the repository support cross-disciplinary use across the physical sciences.

Figure 2 reveals that individual PDB entries have impact on a wide range of disciplines. The three most highly-cited structures are included, along with the number of citations falling into the top 5 categories. Not surprisingly, all have Biochemistry & Molecular Biology as the top category. The following categories are quite different for these three entries, reflecting the different uses that are made of these structures: the nucleosome (1aoi) in basic biology and understanding of genetic mechanisms, the potassium channel (1bl8) as a central structure used for understanding and engineering specific ion channels, and rhodopsin (1f88) which was used for many years as a template for understanding and modeling the pharmacology of G-protein coupled receptors (GPCRs). The citations also fall into numerous other categories: for example, citations for the nucleosome structure fall into over 100 separate categories.

In the sections below, we identify the top-cited structures in each of these subject categories, and describe how these structures have impacted the fields.

Figure 2. Top Subject Categories of Articles Citing the Most-Referenced PDB Structures

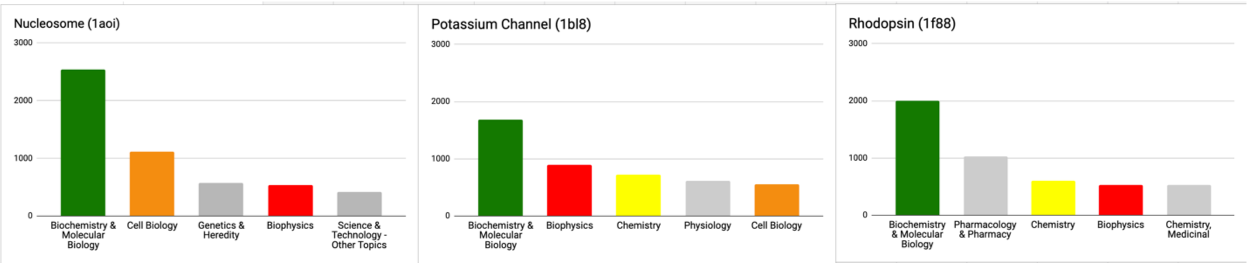

Figure 2. Top subject categories for publications citing the top-cited PDB structures. Common categories include Biochemistry & Molecular Biology (green), Cell Biology (orange), Biophysics (red), and Chemistry (yellow).

Materials Science

Articles in this category cited 18,495 unique structures, or roughly 12% of the archive. Two of the top cited PDB entries here, serum albumin (1uor) and potassium channel (1bl8) also appear in the overall top cited list (Table 2). Surprisingly, two additional structures of serum albumin (1ao6/1bm0) also appear in this top cited list. The reasons for citation of these structures in materials science journals are reflected in the most frequently used keywords in these citing articles: in vivo, mechanism, drug delivery, adsorption, in vitro, crystal structure, protein/s, nanoparticles, and binding.

These structures provide information that is useful in a variety of current bioengineering and nanotechnology goals. Serum albumin (1uor, 1ao6, 1bm0) plays essential roles in delivery of a wide variety of small molecules in the blood, thus it is often a key for assessing the ADME (Absorption, Distribution, Metabolism, Excretion) properties of engineered molecules. Many of these cited papers explore design of nanoparticles for delivery of molecules in the blood, building on knowledge of the structure. Similarly, alpha-hemolysin (7ahl) and potassium channels (1bl8) are worked examples of selective channels and have been used in bioengineering efforts. In particular, structural understanding of alpha-hemolysin has been instrumental in the engineering of nanopores for DNA sequencing. Designed DNA structures (3gbi) are some of the first successful examples of *de novo* design in bionanotechnology, and the strong streptavidin-biotin interaction (1stp) is often used to connect modular components in designed nanostructures. Materials Science publications reference the *in situ* structure of collagen microfibrils (3hqv and 3hr2), including reports exploring the properties of connective tissue and biomineralization. Photosystem II appears in Materials Science (3wu2), as well as in nearly all of the other categories below, since these structures revealed the water-splitting details of the oxygen-evolving center.

Table 2. Top-Cited PDB Structures in Materials Science

Structure	**PDB ID**	**Journal**	**Structure Primary Citation**	**Times Cited**
Serum albumin	1uor*	Nature	(He and Carter, 1992)	215
Alpha-hemolysin	7ahl	Science	(Song et al., 1996)	140
Designed DNA	3gbi	Nature	(Zheng et al., 2009)	138
Biotin/streptavidin	1stp	Science	(Weber et al., 1989)	136
Collagen	3hqv/3hr2	Proceedings of National Academy of Sciences USA	(Orgel et al., 2006)	116
Photosystem II	3wu2	Nature	(Umena et al., 2011)	100
Potassium channel	1bl8*	Science	(Doyle et al., 1998)	95
Serum albumin	1ao6/1bm0	Protein Engineering	(Sugio et al., 1999)	93

*also appears in top cited overall list

Physics

Articles in this category cited 50,819 unique structures. Two of the top cited structures, potassium channel (1bl8) and the nucleosome (1aoi) also appear in the overall top cited list (Table 3). The most frequently used keywords in these citing articles include: model, mechanism, protein/s, binding, molecular dynamics/simulations, spectroscopy, and crystal structure.

This category includes an interesting mix of structures related to photosynthesis (3pcq, 1s5l, 3wu2, 1jb0, 1lgh) and structures related to development of experimental methods (also 3pcq, 1m8m, 1l2y). The structures of photosystems revealed the detailed arrangements of chromophores in the protein complexes, and thus provide concrete information on the types of geometries and distances that are relevant for excitation and electron transfer. These structures also include several seminal developments in structural science with strong ties to physics, including determination of photosystem I by femtosecond X-ray protein nanocrystallography (3pcq) and determination of the spectrin domain by solid-state magic-angle-spinning NMR spectroscopy (1m8m). In addition, two structures related to nanotechnology appear in the list: a very small *de novo* designed protein (1l2y) and alpha-hemolysin (7ahl), mentioned in the section above. Inclusion of the nucleosome (1aoi) in this list may seem like a bit of a puzzle, until we understand that much effort has been expended with trying to understand and model the physics of DNA bending as it relates to nucleosome positioning and higher-order chromatin structure.

Table 3. Top-Cited PDB Structures in Physics

Structure	PDB ID	Journal	Structure Primary Citation	Times Cited
Photosystem I	3pcq	Nature	(Chapman et al., 2011)	319
Potassium channel	1bl8*	Science	(Doyle et al., 1998)	311
Nucleosome	1aoi*	Nature	(Luger et al., 1997)	186
Photosystem II	1s5l	Science	(Ferreira et al., 2004)	186
Photosystem II	3wu2	Nature	(Umena et al., 2011)	177
Alpha-spectrin sh3 domain	1m8m	Nature	(Castellani et al., 2002)	174
Photosystem I	1jb0	Nature	(Jordan et al., 2001)	148
Designed TRP-cage miniprotein	1l2y	Nature Structural Biology	(Neidigh et al., 2002)	147
Light-harvesting complex	1lgh	Structure	(Koepke et al., 1996)	131
Alpha-hemolysin	7ahl	Science	(Song et al., 1996)	127

* also appears in top cited overall list

Computer Science

Articles in this category cited 28,122 unique structures, and rhodopsin (1f88) also appears in the overall top cited list (Table 4). The most frequently used keywords in these citing articles included: docking, identification, Inhibitors, prediction, molecular dynamics, forcefield, design, binding, and crystal structure.

These structures represent important targets for drug development, and thus are often used to test new structure-based drug design methodology, including a shape-based 3-D scaffold hopping method (1y2f, 1y2g) and novel use of cyclic ureas to mimic substrate binding and displace a key water molecule in HIV protease (1hvr). Half of the list are GPCRs (2rh1, 3eml, 2vt4, 2r4s), along with the landmark structure of rhodopsin (1f88), which was used for many years to model GPCRs and ligand binding thereto.

Table 4. Top-Cited PDB Structures in Computer Science

Structure	**PDB ID**	**Journal**	**Structure Primary Citation**	**Times Cited**
Rhodopsin	1f88*	Science	(Palczewski et al., 2000)	185
Beta2 adrenergic receptor	2rh1	Science	(Cherezov et al., 2007)	134
ZipA/inhibitor	1y2g/1y2f	Journal of Medicinal Chemistry	(Rush et al., 2005)	129
Adenosine receptor	3eml	Science	(Jaakola et al., 2008)	94
HIV protease/cyclic ureas	1hvr	Science	(Lam et al., 1994)	77
Beta1 adrenergic receptor	2vt4	Nature	(Warne et al., 2008)	69
Beta2 adrenergic receptor	2r4r/2r4s	Nature	(Rasmussen et al., 2007)	67
Dihydrofolate reductase	3dfr/4dfr	Journal of Biological Chemistry	(Bolin et al., 1982)	64

* also appears in top cited overall list

Chemistry

Articles in this category cited 87,073 unique structures, and given the strong connections between biology and chemistry, half of the top cited structures (2v3h/2v3o, 1uor, 1f88, 1bl8) appear in the overall top cited list (Table 5). The most frequently used keywords in these citing articles were: complexes, protein/s, *E. coli*, inhibitors, mechanism, design, derivatives, binding, and crystal structure.

The top two entries are thrombin with an inhibitor (2v3h) and with a fluorinated version of the inhibitor (2v3o), and the primary reference is a review that is cited by studies of the effects of fluorination in inhibitor design. Serum albumin (1uor) showed up in “Materials Science” in relation to bioengineering efforts, but many of the citations in “Chemistry” are directly related to binding of molecules to the protein and characterizing its functional properties in blood. Two hydrogenase enzymes (1feh, 1hfe) and photosystem II (1s5l, 3wu2) perform interesting chemistry catalyzed by unusual metal clusters. The structure of a human telomeric quadruplex (1k8p) is cited by all manner of studies looking at its chemical properties and interactions with ions, small molecules, and proteins.

Table 5. Top-Cited PDB Structures in Chemistry

Structure	PDB ID	Journal	Structure Primary Citation	Times Cited
Thrombin	2v3o/2v3h*	Science	(Muller et al., 2007)	2,425
Serum albumin	1uor*	Nature	(He and Carter, 1992)	1,334
Photosystem II	1s5l	Science	(Ferreira et al., 2004)	1,043
Rhodopsin	1f88*	Science	(Palczewski et al., 2000)	1,003
Photosystem II	3wu2	Nature	(Umena et al., 2011)	936
Fe-only hydrogenase	1feh	Science	(Peters et al., 1998)	925
Potassium channel	1bl8*	Science	(Doyle et al., 1998)	803
Hydrogenase	1hfe	Structure	(Nicolet et al., 1999)	743
DNA quadruplex	1k8p	Nature	(Parkinson et al., 2002)	671

* also appears in top cited overall list

Engineering

Articles in this category cited 16,190 unique structures, and serum albumin (1uor) and potassium channel (1bl8) again appear in the ten list (Table 6). The most frequently used keywords in these citing articles were: mechanism, in vivo, protein/s, purification, binding, in vitro, *E. coli*, expression.

Several of these structures are related to bioengineering projects. The citations for serum albumin (1uor) include studies about the interaction with a wide variety of dyes, nanoparticles and other engineered molecules. Structures of collagen (3hqv, 3hr2, 1cag), fibronectin (1fnf), and osteocalcin (1q8h) played key roles in the understanding of cell adhesion, connective tissues and bone. The potassium channel structure (1bl8) is cited by studies looking at nanopores and biosensors. The lipase (5tgl) and laccase (1gyc) structures were cited in studies of engineered and immobilized versions of the enzymes.

Table 6. Top-Cited PDB Structures in Engineering

Structure	PDB ID	Journal	Structure Primary Citation	Times Cited

Serum albumin	1uor*	Nature	(He and Carter, 1992)	84
Collagen	3hqv/3hr2	Proceedings of the National Academy of Sciences USA	(Orgel et al., 2006)	68
Potassium channel	1bl8*	Science	(Doyle et al., 1998)	51
Fibronectin	1fnf	Cell	(Leahy et al., 1996)	51
Photosystem II	1s5l	Science	(Ferreira et al., 2004)	44
Lipase/inhibitor	5tgl	Nature	(Brzozowski et al., 1991)	43
Laccase	1gyc	Journal of Biological Chemistry	(Piontek et al., 2002)	43
Osteocalcin	1q8h	Nature	(Hoang et al., 2003)	42
Collagen	1cag	Science	(Bella et al., 1994)	40

* also appears in top cited overall list

Mathematics

Articles in this category cited 7,306 unique structures. Four of the top cited structures (1bl8, 1f88, 1aoi, 1brd) appear in the overall top cited list (Table 7). The most frequently used keywords in these citing articles were: molecular dynamics, protein/s, binding, identification, recognition, sequence, prediction, database, crystal structure.

Many of these papers are involved in modeling and analysis of protein and nucleic acid structures, with “Mathematical Computational Biology” being the major mathematics-related category. For example, the potassium channel (1bl8) citations include many computation studies exploring the dynamics of channel gating and permeation, as well as methods for predicting structure and function of other channels based on this structure, and the rhodopsin structure (1f88) includes several citations for methods that model the structure of GPCRs. The nucleosome (1aoi) citations include studies about nucleosome positioning and modeling of DNA bending or higher chromatin structure. The designed protein structures (1qys, 1fsv, 1fsd, 1fsm) are cited by methods papers that explore prediction of protein folding and design, and the ribosome structure (1ffk) is cited by methods exploring RNA structure and interaction of RNA and protein. The first atomic structure of a B-DNA helix (1bna) is cited in modeling studies of DNA conformation and interaction.

Table 7. Top-Cited PDB Structures in Mathematics

Structure	PDB ID	Journal	Structure Primary Citation	Times Cited
Potassium channel	1bl8	Science	(Doyle et al., 1998)	29
Designed protein	1qys	Science	(Kuhlman et al., 2003)	24
Rhodopsin	1f88	Science	(Palczewski et al., 2000)	22
Nucleosome	1aoi	Nature	(Luger et al., 1997)	20
Designed protein	1fsv/1fsd	Science	(Dahiyat and Mayo, 1997)	20
Ribosomal subunit	1ffk	Science	(Ban et al., 2000)	13
Bacteriorhodopsin	1brd*	Journal of Molecular Biology	(Henderson et al., 1990)	11
Photosystem II	2axt	Nature	(Loll et al., 2005)	11
B-DNA dodecamer	1bna	Proceedings of National Academy of Sciences USA	(Drew et al., 1981)	10

* also appears in top cited overall list

Conclusions

This analysis has shown a large impact of PDB archive within the discipline of molecular biology and in many related disciplines. Of course, this analysis uses only one metric for assessing impact—the record of citations. Additional information could be obtained through analysis of instances of PDB structure IDs in publications, or linkage of specific PDB entries in digital resources. We are also interested in assessing the impact of the PDB archive in education and public understanding, which may potentially be approached through analysis of usage and citation of entries in textbooks and popular publications. That effort may be more difficult, however, given the citation practices in those publications are not a tightly codified as in professional scientific publications.

The PDB archive was originally established to serve the structural biology community. The extensive usage of PDB structures across a variety of disciplines demonstrates the importance of structural studies and how data archives support interdisciplinary research.

Acknowledgements

We thank the more than 40,000 structural biologists who have deposited structures to the PDB and the researchers who have cited them. RCSB PDB is a member of the Worldwide Protein Data Bank (wwPDB.org).

Funding Information

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. Natalie Verdiguel was supported by an NSF REU to the RCSB PDB (DBI-1338415).

References

Abaci, A. (2017). Scientific competition, impact factor, and Altmetrics. Anatol J Cardiol, 18(5), 313. doi:10.14744/AnatolJCardiol.2017.11

Abrahams, J. P., Leslie, A. G., Lutter, R. & Walker, J. E. 1994. Structure at 2.8 A resolution of F1-ATPase from bovine heart mitochondria. *Nature,* 370**,** 621-8. 10.1038/370621a0

Arnold, K., Bordoli, L., Kopp, J. & Schwede, T. 2006. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics,* 22**,** 195-201. 10.1093/bioinformatics/bti770

Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. 2000. The complete atomic structure of the large ribosomal subunit at a 2.4 Å resolution. *Science,* 289**,** 905-920. 10.1126/science.289.5481.905

Basner, J. 2017. Impact Analysis of “Berman HM et al., (2000), The Protein Data Bank”. doi: 10.2210/rcsb_pdb/pdb-cit-anal-2017. doi: 10.2210/rcsb_pdb/pdb-cit-anal-2017

Bella, J., Eaton, M., Brodsky, B. & Berman, H. M. 1994. Crystal and molecular structure of a collagen-like peptide at 1.9 A resolution. *Science,* 266**,** 75-81. 10.1126/science.7695699

Berman, H. M., Henrick, K. & Nakamura, H. 2003. Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.,* 10**,** 980. 10.1038/nsb1203-980

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. 2000. The Protein Data Bank. *Nucleic Acids Res,* 28**,** 235-42. 10.1093/nar/28.1.235

Bjorkman, P. J., Saper, M. A., Samraoui, B., Bennett, W. S., Strominger, J. L. & Wiley, D. C. 1987. Structure of the human class I histocompatibility antigen, HLA-A2. *Nature,* 329**,** 506-12. 10.1038/329506a0

Bolin, J. T., Filman, D. J., Matthews, D. A., Hamlin, R. C. & Kraut, J. 1982. Crystal structures of Escherichia coli and Lactobacillus casei dihydrofolate reductase refined at 1.7 A resolution. I. General features and binding of methotrexate. *J Biol Chem,* 257**,** 13650-62.

Bolton, W. & Perutz, M. F. 1970. Three dimensional fourier synthesis of horse deoxyhaemoglobin at 2.8 Ångstrom units resolution. *Nature,* 228**,** 551-2. 10.1038/228551a0

Bousfield, D., McEntyre, J., Velankar, S., Papadatos, G., Bateman, A., Cochrane, G., Kim, J. H., Graef, F., Vartak, V., Alako, B. & Blomberg, N. 2016. Patterns of database citation in articles and patents indicate long-term scientific and industry value of biological data resources. *F1000Res,* 5. 10.12688/f1000research.7911.1

Bronmann, L., & Williams, R. (2020). An Evaluation of Percentile Measures of Citation Impact, and a Proposal for Making Them Better. arXiv, arXiv:2001.04290.

Brzozowski, A. M., Derewenda, U., Derewenda, Z. S., Dodson, G. G., Lawson, D. M., Turkenburg, J. P., Bjorkling, F., Huge-Jensen, B., Patkar, S. A. & Thim, L. 1991. A model for interfacial activation in lipases from the structure of a fungal lipase-inhibitor complex. *Nature,* 351**,** 491-4. 10.1038/351491a0

Burley, S. K., Berman, H. M., Bhikadiya, C., Bi, C., Chen, L., Di Costanzo, L., Christie, C., Dalenberg, K., Duarte, J. M., Dutta, S., Feng, Z., Ghosh, S., Goodsell, D. S., Green, R. K., Guranovic, V., Guzenko, D., Hudson, B. P., Kalro, T., Liang, Y., Lowe, R., Namkoong, H., Peisach, E., Periskova, I., Prlic, A., Randle, C., Rose, A., Rose, P., Sala, R., Sekharan, M., Shao, C., Tan, L., Tao, Y. P., Valasatava, Y., Voigt, M., Westbrook, J., Woo, J., Yang, H., Young, J., Zhuravleva, M. & Zardecki, C. 2019. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res,* 47**,** D464-D474. 10.1093/nar/gky1004

Burley, S. K., Berman, H. M., Christie, C., Duarte, J. M., Feng, Z., Westbrook, J., Young, J. & Zardecki, C. 2018. RCSB Protein Data Bank: Sustaining a living digital data resource that enables breakthroughs in scientific research and biomedical education. *Protein Sci,* 27**,** 316-330. 10.1002/pro.3331

Castellani, F., van Rossum, B., Diehl, A., Schubert, M., Rehbein, K. & Oschkinat, H. 2002. Structure of a protein determined by solid-state magic-angle-spinning NMR spectroscopy. *Nature,* 420**,** 98-102. 10.1038/nature01070

Chapman, H. N., Fromme, P., Barty, A., White, T. A., Kirian, R. A., Aquila, A., Hunter, M. S., Schulz, J., DePonte, D. P., Weierstall, U., Doak, R. B., Maia, F. R., Martin, A. V., Schlichting, I., Lomb, L., Coppola, N., Shoeman, R. L., Epp, S. W., Hartmann, R., Rolles, D., Rudenko, A., Foucar, L., Kimmel, N., Weidenspointner, G., Holl, P., Liang, M., Barthelmess, M., Caleman, C., Boutet, S., Bogan, M. J., Krzywinski, J., Bostedt, C., Bajt, S., Gumprecht, L., Rudek, B., Erk, B., Schmidt, C., Homke, A., Reich, C., Pietschner, D., Struder, L., Hauser, G., Gorke, H., Ullrich, J., Herrmann, S., Schaller, G., Schopper, F., Soltau, H., Kuhnel, K. U., Messerschmidt, M., Bozek, J. D., Hau-Riege, S. P., Frank, M., Hampton, C. Y., Sierra, R. G., Starodub, D., Williams, G. J., Hajdu, J., Timneanu, N., Seibert, M. M., Andreasson, J., Rocker, A., Jonsson, O., Svenda, M., Stern, S., Nass, K., Andritschke, R., Schroter, C. D., Krasniqi, F., Bott, M., Schmidt, K. E., Wang, X., Grotjohann, I., Holton, J. M., Barends, T. R., Neutze, R., Marchesini, S., Fromme, R., Schorb, S., Rupp, D., Adolph, M., Gorkhover, T., Andersson, I., Hirsemann, H., Potdevin, G., Graafsma, H., Nilsson, B. & Spence, J. C. 2011. Femtosecond X-ray protein nanocrystallography. *Nature,* 470**,** 73-7. 10.1038/nature09750

Cherezov, V., Rosenbaum, D. M., Hanson, M. A., Rasmussen, S. G., Thian, F. S., Kobilka, T. S., Choi, H. J., Kuhn, P., Weis, W. I., Kobilka, B. K. & Stevens, R. C. 2007. High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. *Science,* 318**,** 1258-65. 10.1126/science.1150577

Clarivate Analytics Web of ScienceTM. © Copyright Clarivate Analytics 2019, All rights reserved.

Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K. & McGillivray, B. 2019. The citation advantage of linking publications to research data. *arXiv***,** arXiv:1907.02565 [cs.DL].

Dahiyat, B. I. & Mayo, S. L. 1997. De novo protein design: fully automated sequence selection. *Science,* 278**,** 82-7. 10.1126/science.278.5335.82

Diamandis, E. P. (2017). The Journal Impact Factor is under attack - use the CAPCI factor instead. BMC Med, 15(1), 9. doi:10.1186/s12916-016-0773-5

Doyle, D. A., Morais Cabral, J., Pfuetzner, R. A., Kuo, A., Gulbis, J. M., Cohen, S. L., Chait, B. T. & MacKinnon, R. 1998. The structure of the potassium channel: molecular basis of K+ conduction and selectivity. *Science,* 280**,** 69-77. 10.1126/science.280.5360.69

Drew, H. R., Wing, R. M., Takano, T., Broka, C., Tanaka, S., Itakura, K. & Dickerson, R. E. 1981. Structure of a B-DNA dodecamer: conformation and dynamics. *Proc. Natl. Acad. Sci. U.S.A.,* 78**,** 2179-2183. 10.1073/pnas.78.4.2179

Ferreira, K. N., Iverson, T. M., Maghlaoui, K., Barber, J. & Iwata, S. 2004. Architecture of the photosynthetic oxygen-evolving center. *Science,* 303**,** 1831-8. 10.1126/science.1093087

Fitzgerald, P. M. D., Westbrook, J. D., Bourne, P. E., McMahon, B., Watenpaugh, K. D. & Berman, H. M. 2005. 4.5 Macromolecular dictionary (mmCIF). *In:* HALL, S. R. & MCMAHON, B. (eds.) *International Tables for Crystallography G. Definition and exchange of crystallographic data.* Dordrecht, The Netherlands: Springer.

Goodsell, D. S., Zardecki, C., Berman, H. M. & Burley, S. K. 2019. Insights from 20 Years of the Molecule of the Month. *Biochemistry and Molecular Biology Education***,** submitted.

He, X. M. & Carter, D. C. 1992. Atomic structure and chemistry of human serum albumin. *Nature,* 358**,** 209-15. 10.1038/358209a0

Henderson, R., Baldwin, J. M., Ceska, T. A., Zemlin, F., Beckmann, E. & Downing, K. H. 1990. Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *J Mol Biol,* 213**,** 899-929. 10.1016/S0022-2836(05)80271-2

Hersel, U., Dahmen, C. & Kessler, H. 2003. RGD modified polymers: biomaterials for stimulated cell adhesion and beyond. *Biomaterials,* 24**,** 4385-415. 10.1016/s0142-9612(03)00343-0

Hoang, Q. Q., Sicheri, F., Howard, A. J. & Yang, D. S. 2003. Bone recognition mechanism of porcine osteocalcin from crystal structure. *Nature,* 425**,** 977-80. 10.1038/nature02079

Huang, Y. H., Rose, P. W. & Hsu, C. N. 2015. Citing a Data Repository: A Case Study of the Protein Data Bank. *PLoS One,* 10**,** e0136631. 10.1371/journal.pone.0136631

Humphrey, W., Dalke, A. & Schulten, K. 1996. VMD: visual molecular dynamics. *J Mol Graph,* 14**,** 33-38.

Jaakola, V. P., Griffith, M. T., Hanson, M. A., Cherezov, V., Chien, E. Y., Lane, J. R., Ijzerman, A. P. & Stevens, R. C. 2008. The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist. *Science,* 322**,** 1211-7. 10.1126/science.1164772

Jordan, P., Fromme, P., Witt, H. T., Klukas, O., Saenger, W. & Krauss, N. 2001. Three-dimensional structure of cyanobacterial photosystem I at 2.5 A resolution. *Nature,* 411**,** 909-17. 10.1038/35082000

Kendrew, J. C., Dickerson, R. E., Strandberg, B. E., Hart, R. G., Davies, D. R., Phillips, D. C. & Shore, V. C. 1960. Structure of myoglobin: A three-dimensional Fourier synthesis at 2 A. resolution. *Nature,* 185**,** 422-7. 10.1038/185422a0

Kinjo, A. R., Bekker, G. J., Wako, H., Endo, S., Tsuchiya, Y., Sato, H., Nishi, H., Kinoshita, K., Suzuki, H., Kawabata, T., Yokochi, M., Iwata, T., Kobayashi, N., Fujiwara, T., Kurisu, G. & Nakamura, H. 2018. New tools and functions in data-out activities at Protein Data Bank Japan (PDBj). *Protein Sci.,* 27**,** 95-102. 10.1002/pro.3273

Koelblinger, D., Zimmermann, G., Weineck, S. B., & Kiesslich, T. (2019). Size matters! Association between journal size and longitudinal variability of the Journal Impact Factor. PLoS ONE, 14(11), e0225360. doi:10.1371/journal.pone.0225360

Koepke, J., Hu, X., Muenke, C., Schulten, K. & Michel, H. 1996. The crystal structure of the light-harvesting complex II (B800-850) from Rhodospirillum molischianum. *Structure,* 4**,** 581-97. 10.1016/s0969-2126(96)00063-9

Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L. & Baker, D. 2003. Design of a novel globular protein fold with atomic-level accuracy. *Science,* 302**,** 1364-8. 10.1126/science.1089427

Lam, P. Y., Jadhav, P. K., Eyermann, C. J., Hodge, C. N., Ru, Y., Bacheler, L. T., Meek, J. L., Otto, M. J., Rayner, M. M., Wong, Y. N. & et al. 1994. Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors. *Science,* 263**,** 380-4. 10.1126/science.8278812

Leahy, D. J., Aukhil, I. & Erickson, H. P. 1996. 2.0 A crystal structure of a four-domain segment of human fibronectin encompassing the RGD loop and synergy region. *Cell,* 84**,** 155-64. 10.1016/s0092-8674(00)81002-8

Loll, B., Kern, J., Saenger, W., Zouni, A. & Biesiadka, J. 2005. Towards complete cofactor arrangement in the 3.0 A resolution structure of photosystem II. *Nature,* 438**,** 1040-4. 10.1038/nature04224

Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. 1997. Crystal structure of the nucleosome core particle at 2.8Å resolution. *Nature,* 389**,** 251-260.

Markosian, C., Di Costanzo, L., Sekharan, M., Shao, C., Burley, S. K. & Zardecki, C. 2018. Analysis of impact metrics for the Protein Data Bank. *Sci Data,* 5**,** 180212. 10.1038/sdata.2018.212

Morris, C. 2018. The Life Cycle of Structural Biology Data. *Data Science Journal,* 17**,** 26. 10.5334/dsj-2018-026/

Muller, K., Faeh, C. & Diederich, F. 2007. Fluorine in pharmaceuticals: looking beyond intuition. *Science,* 317**,** 1881-6. 10.1126/science.1131943

Neidigh, J. W., Fesinmeyer, R. M. & Andersen, N. H. 2002. Designing a 20-residue protein. *Nat Struct Biol,* 9**,** 425-30. 10.1038/nsb798

Nel, A. E., Madler, L., Velegol, D., Xia, T., Hoek, E. M., Somasundaran, P., Klaessig, F., Castranova, V. & Thompson, M. 2009. Understanding biophysicochemical interactions at the nano-bio interface. *Nat Mater,* 8**,** 543-57. 10.1038/nmat2442

Nicolet, Y., Piras, C., Legrand, P., Hatchikian, C. E. & Fontecilla-Camps, J. C. 1999. Desulfovibrio desulfuricans iron hydrogenase: the structure shows unusual coordination to an active site Fe binuclear center. *Structure,* 7**,** 13-23. 10.1016/s0969-2126(99)80005-7

Orgel, J. P., Irving, T. C., Miller, A. & Wess, T. J. 2006. Microfibrillar structure of type I collagen *in situ*. *Proc Natl Acad Sci U S A,* 103**,** 9001-5. 10.1073/pnas.0502718103

Palczewski, K., Kumasaka, T., Hori, T., Behnke, C. A., Motoshima, H., Fox, B. A., Le Trong, I., Teller, D. C., Okada, T., Stenkamp, R. E., Yamamoto, M. & Miyano, M. 2000. Crystal structure of rhodopsin: A G protein-coupled receptor. *Science,* 289**,** 739-45. 10.1126/science.289.5480.739

Parkinson, G. N., Lee, M. P. & Neidle, S. 2002. Crystal structure of parallel quadruplexes from human telomeric DNA. *Nature,* 417**,** 876-80. 10.1038/nature755

Pendlebury, D. A. (2009). The use and misuse of journal metrics and other citation indicators. Arch Immunol Ther Exp (Warsz), 57(1), 1-11. doi:10.1007/s00005-009-0008-y

Perutz, M. F., Rossmann, M. G., Cullis, A. F., Muirhead, H., Will, G. & North, A. C. T. 1960. Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5 Å resolution, obtained by X-ray analysis. *Nature,* 185**,** 416-422. 10.1038/185416a0

Peters, J. W., Lanzilotta, W. N., Lemon, B. J. & Seefeldt, L. C. 1998. X-ray crystal structure of the Fe-only hydrogenase (CpI) from Clostridium pasteurianum to 1.8 angstrom resolution. *Science,* 282**,** 1853-8. 10.1126/science.282.5395.1853

Piontek, K., Antorini, M. & Choinowski, T. 2002. Crystal structure of a laccase from the fungus Trametes versicolor at 1.90-A resolution containing a full complement of coppers. *J Biol Chem,* 277**,** 37663-9. 10.1074/jbc.M204571200

Protein Data Bank 1971. Crystallography: Protein Data Bank. *Nature (London), New Biol.,* 233**,** 223-223. 10.1038/newbio233223b0

Rasmussen, S. G., Choi, H. J., Rosenbaum, D. M., Kobilka, T. S., Thian, F. S., Edwards, P. C., Burghammer, M., Ratnala, V. R., Sanishvili, R., Fischetti, R. F., Schertler, G. F., Weis,

W. I. & Kobilka, B. K. 2007. Crystal structure of the human beta2 adrenergic G-protein-coupled receptor. *Nature,* 450**,** 383-7. 10.1038/nature06325

Rush, T. S., 3rd, Grant, J. A., Mosyak, L. & Nicholls, A. 2005. A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J Med Chem,* 48**,** 1489-95. 10.1021/jm040163o

Schneider, J. W., Van Leeuwen, T. Visser, M., Aagaard, K. 2019. Examining national citation impact by comparing developments in a fixed and a dynamic journal set. Scientometrics, 973-985. 10.1007/s11192-019-03082-3

Sheldrick, G. M. 2008. A short history of SHELX. *Acta Crystallogr A,* 64**,** 112-22. 10.1107/S0108767307043930

Song, L., Hobaugh, M. R., Shustak, C., Cheley, S., Bayley, H. & Gouaux, J. E. 1996. Structure of staphylococcal alpha-hemolysin, a heptameric transmembrane pore. *Science,* 274**,** 1859-66. 10.1126/science.274.5294.1859

Sugio, S., Kashima, A., Mochizuki, S., Noda, M. & Kobayashi, K. 1999. Crystal structure of human serum albumin at 2.5 A resolution. *Protein Eng,* 12**,** 439-46. 10.1093/protein/12.6.439

Umena, Y., Kawakami, K., Shen, J. R. & Kamiya, N. 2011. Crystal structure of oxygen-evolving photosystem II at a resolution of 1.9 A. *Nature,* 473**,** 55-60. 10.1038/nature09913

Van Noorden, R., Maher, B. & Nuzzo, R. 2014. The top 100 papers. *Nature,* 514**,** 550-3. 10.1038/514550a

Vassilev, L. T., Vu, B. T., Graves, B., Carvajal, D., Podlaski, F., Filipovic, Z., Kong, N., Kammlott, U., Lukacs, C., Klein, C., Fotouhi, N. & Liu, E. A. 2004. In vivo activation of the p53 pathway by small-molecule antagonists of MDM2. *Science,* 303**,** 844-8. 10.1126/science.1092472

Velankar, S., van Ginkel, G., Alhroub, Y., Battle, G. M., Berrisford, J. M., Conroy, M. J., Dana, J. M., Gore, S. P., Gutmanas, A., Haslam, P., Hendrickx, P. M., Lagerstedt, I., Mir, S., Fernandez Montecelo, M. A., Mukhopadhyay, A., Oldfield, T. J., Patwardhan, A., Sanz-Garcia, E., Sen, S., Slowley, R. A., Wainwright, M. E., Deshpande, M. S., Iudin, A., Sahni, G., Salavert Torres, J., Hirshberg, M., Mak, L., Nadzirin, N., Armstrong, D. R., Clark, A. R., Smart, O. S., Korir, P. K. & Kleywegt, G. J. 2016. PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. *Nucleic Acids Res,* 44**,** D385-95. 10.1093/nar/gkv1047

Warne, T., Serrano-Vega, M. J., Baker, J. G., Moukhametzianov, R., Edwards, P. C., Henderson, R., Leslie, A. G., Tate, C. G. & Schertler, G. F. 2008. Structure of a beta1-adrenergic G-protein-coupled Receptor. *Nature,* 454**,** 486-91. 10.1038/nature07101

Watson, H. C. 1969. The stereochemistry of the protein myoglobin. *Prog. Stereochem.,* 4**,** 299.

Weber, P. C., Ohlendorf, D. H., Wendoloski, J. J. & Salemme, F. R. 1989. Structural origins of high-affinity biotin binding to streptavidin. *Science,* 243**,** 85-8. 10.1126/science.2911722

Westbrook, J. D. & Burley, S. K. 2019. How Structural Biologists and the Protein Data Bank Contributed to Recent FDA New Drug Approvals. *Structure,* 27**,** 211-217. 10.1016/j.str.2018.11.007

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J., Groth, P., Goble, C., Grethe, J. S., Heringa, J., t Hoen, P. A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S. A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. & Mons, B. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data,* 3**,** 1-9. 10.1038/sdata.2016.18

wwPDB consortium 2019. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res,* 47**,** D520-D528. 10.1093/nar/gky949

Young, J. Y., Westbrook, J. D., Feng, Z., Peisach, E., Persikova, I., Sala, R., Sen, S., Berrisford, J. M., Swaminathan, G. J., Oldfield, T. J., Gutmanas, A., Igarashi, R., Armstrong, D. R., Baskaran, K., Chen, L., Chen, M., Clark, A. R., Costanzo, L. D., Dimitropoulos, D., Gao, G., Ghosh, S., Gore, S., Guranovic, V., Hendrickx, P. M. S., Hudson, B. P., Ikegawa, Y., Kengaku, Y., Lawson, C. L., Liang, Y., Mak, L., Mukhopadhyay, A., Narayanan, B., Nishiyama, K., Patwardhan, A., Sahni, G., Sanz-García, E., Sato, J., Sekharan, M. R., Shao, C., Smart, O. S., Tan, L., Ginkel, G. v., Yang, H., Zhuravleva, M. A., Markley, J. L., Nakamura, H., Kurisu, G., Kleywegt, G. J., Velankar, S., Berman, H. M. & Burley, S. K. 2018. Worldwide Protein Data Bank biocuration supporting open access to high-quality 3D structural biology data. *Database,* 2018**,** bay002.

Young, J. Y., Westbrook, J. D., Feng, Z., Sala, R., Peisach, E., Oldfield, T. J., Sen, S., Gutmanas, A., Armstrong, D. R., Berrisford, J. M., Chen, L., Chen, M., Di Costanzo, L., Dimitropoulos, D., Gao, G., Ghosh, S., Gore, S., Guranovic, V., Hendrickx, P. M., Hudson, B. P., Igarashi, R., Ikegawa, Y., Kobayashi, N., Lawson, C. L., Liang, Y., Mading, S., Mak, L., Mir, M. S., Mukhopadhyay, A., Patwardhan, A., Persikova, I., Rinaldi, L., Sanz-Garcia, E., Sekharan, M. R., Shao, C., Swaminathan, G. J., Tan, L., Ulrich, E. L., van Ginkel, G., Yamashita, R., Yang, H., Zhuravleva, M. A., Quesada, M., Kleywegt, G. J., Berman, H. M., Markley, J. L., Nakamura, H., Velankar, S. & Burley, S. K. 2017. OneDep: Unified wwPDB System for Deposition, Biocuration, and Validation of Macromolecular Structures in the PDB Archive. *Structure,* 25**,** 536-545. 10.1016/j.str.2017.01.004

Zheng, J., Birktoft, J. J., Chen, Y., Wang, T., Sha, R., Constantinou, P. E., Ginell, S. L., Mao, C. & Seeman, N. C. 2009. From molecular to macroscopic via the rational design of a self-assembled 3D DNA crystal. *Nature,* 461**,** 74-7. 10.1038/nature08274

nature08274 [pii]

Zweib, C., Kim, J. & Adhya, S. 1989. DNA bending by negative regulatory proteins: *Gal* and *Lac* repressor. *Gene Develop.,* 3**,** 606-611. 10.1101/gad.3.5.606