







# RCSB Protein Data bank: Tools for visualizing and understanding biological macromolecules in 3D

Stephen K. Burley<sup>1,2,3,4,5</sup>  | Charmi Bhikadiya<sup>1,2</sup> | Chunxiao Bi<sup>4</sup> | Sebastian Bittrich<sup>4</sup> | Henry Chao<sup>1,2</sup> | Li Chen<sup>1,2</sup> | Paul A. Craig<sup>6</sup> | Gregg V. Crichlow<sup>1,2</sup> | Kenneth Dalenberg<sup>1,2</sup> | Jose M. Duarte<sup>4</sup>  | Shuchismita Dutta<sup>1,2,3</sup> | Maryam Fayazi<sup>1,2</sup> | Zukang Feng<sup>1,2</sup> | Justin W. Flatt<sup>1,2</sup> | Sai J. Ganesan<sup>7,8</sup> | Sutapa Ghosh<sup>1,2</sup> | David S. Goodsell<sup>1,2,3,9</sup>  | Rachel Kramer Green<sup>1,2</sup> | Vladimir Guranovic<sup>1,2</sup> | Jeremy Henry<sup>4</sup> | Brian P. Hudson<sup>1,2</sup> | Igor Khokhriakov<sup>4</sup> | Catherine L. Lawson<sup>1,2</sup> | Yuhe Liang<sup>1,2</sup> | Robert Lowe<sup>1,2</sup> | Ezra Peisach<sup>1,2</sup> | Irina Persikova<sup>1,2</sup> | Dennis W. Piehl<sup>1,2</sup> | Yana Rose<sup>4</sup> | Andrej Sali<sup>7,8</sup> | Joan Segura<sup>4</sup> | Monica Sekharan<sup>1,2</sup> | Chenghua Shao<sup>1,2</sup> | Brinda Vallat<sup>1,2</sup> | Maria Voigt<sup>1,2</sup> | Benjamin Webb<sup>7,8</sup>  | John D. Westbrook<sup>1,2†</sup>  | Shamara Whetstone<sup>1,2</sup> | Jasmine Y. Young<sup>1,2</sup> | Arthur Zalevsky<sup>7,8</sup> | Christine Zardecki<sup>1,2</sup> 

<sup>1</sup>Research Collaboratory for Structural Bioinformatics Protein Data Bank, Rutgers, The State University of New Jersey, Piscataway, New Jersey, USA

<sup>2</sup>Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, New Jersey, USA

<sup>3</sup>Cancer Institute of New Jersey, Rutgers, The State University of New Jersey, New Brunswick, New Jersey, USA

<sup>4</sup>Research Collaboratory for Structural Bioinformatics Protein Data Bank, San Diego Supercomputer Center, University of California, La Jolla, California, USA

<sup>5</sup>Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, New Jersey, USA

<sup>6</sup>School of Chemistry and Materials Science, Rochester Institute of Technology, Rochester, New York, USA

<sup>7</sup>Research Collaboratory for Structural Bioinformatics Protein Data Bank, Department of Bioengineering and Therapeutic Sciences, Quantitative Biosciences Institute, University of California, San Francisco, California, USA

<sup>8</sup>Research Collaboratory for Structural Bioinformatics Protein Data Bank, Department of Pharmaceutical Chemistry, Quantitative Biosciences Institute, University of California, San Francisco, California, USA

<sup>9</sup>Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, California, USA

## Correspondence

Stephen K. Burley, 174 Frelinghuysen Road, Piscataway, NJ 08854, USA.  
Email: [stephen.burley@rcsb.org](mailto:stephen.burley@rcsb.org)

## Funding information

NIH-National Cancer Institute, Grant/Award Number: R01GM133198; NIH-National Institute of Allergy and Infectious Diseases, Grant/Award

## Abstract

Now in its 52nd year of continuous operations, the Protein Data Bank (PDB) is the premiere open-access global archive housing three-dimensional (3D) biomolecular structure data. It is jointly managed by the Worldwide Protein Data Bank (wwPDB) partnership. The Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) is funded by the National Science Foundation, National Institutes of Health, and US Department of

<sup>†</sup>Deceased

Number: R01GM133198; NIH-National Institute of General Medical Sciences, Grant/Award Number: R01GM133198; National Science Foundation, Grant/Award Number: DBI-1832184; U.S. Department of Energy, Grant/Award Number: DE-SC0019749; National Science Foundation, Grant/Award Number: DBI-2019297; National Science Foundation, Grant/Award Number: DBI-2129634; National Science Foundation, Grant/Award Numbers: DBI-2112967, DBI-1756250, DBI-2112966, DBI-1756248; NIH-NIGMS, Grant/Award Numbers: P41GM109824, R01GM083960; UK Biotechnology and Biological Research Council, Grant/Award Number: BB/V004247/1; UK Biotechnology and Biological Research Council, Grant/Award Number: BB/W017970/1

**Review Editor:** Nir Ben-Tal

Energy and serves as the US data center for the wwPDB. RCSB PDB is also responsible for the security of PDB data in its role as wwPDB-designated Archive Keeper. Every year, RCSB PDB serves tens of thousands of depositors of 3D macromolecular structure data (coming from macromolecular crystallography, nuclear magnetic resonance spectroscopy, electron microscopy, and micro-electron diffraction). The RCSB PDB research-focused web portal (RCSB.org) makes PDB data available at no charge and without usage restrictions to many millions of PDB data consumers around the world. The RCSB PDB training, outreach, and education web portal (PDB101.RCSB.org) serves nearly 700 K educators, students, and members of the public worldwide. This invited Tools Issue contribution describes how RCSB PDB (i) is organized; (ii) works with wwPDB partners to process new depositions; (iii) serves as the wwPDB-designated Archive Keeper; (iv) enables exploration and 3D visualization of PDB data via RCSB.org; and (v) supports training, outreach, and education via PDB101.RCSB.org. New tools and features at RCSB.org are presented using examples drawn from high-resolution structural studies of proteins relevant to treatment of human cancers by targeting immune checkpoints.

#### KEYWORDS

electron microscopy, macromolecular crystallography, micro-electron diffraction, Mol\*, nuclear magnetic resonance spectroscopy, open access, PDB, Protein Data Bank, RCSB Protein Data Bank, Worldwide Protein Data Bank

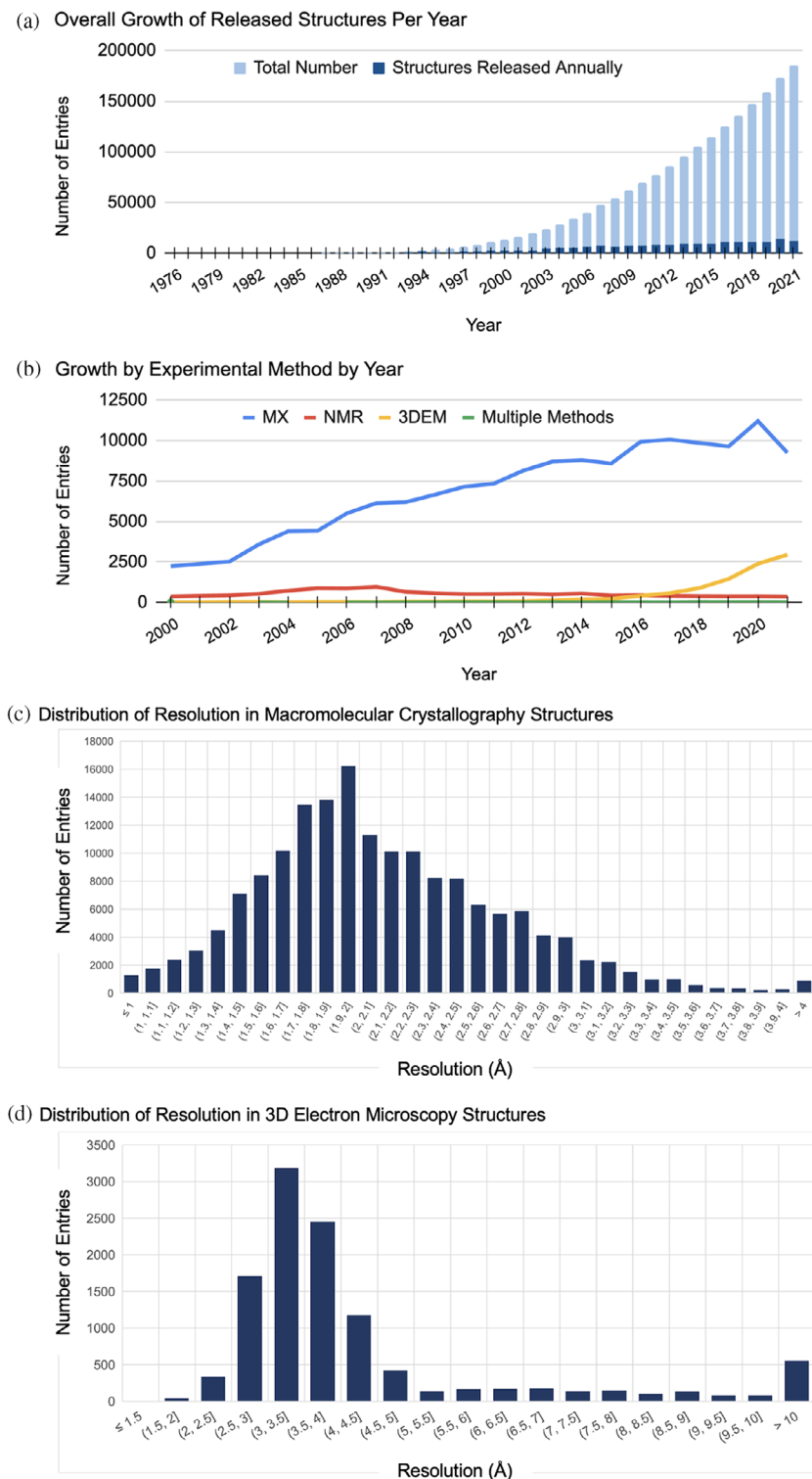
## 1 | INTRODUCTION

The Protein Data Bank (PDB) is in its 52nd year of continuous operation. Established in 1971 as the first open-access digital data resource in biology,<sup>1</sup> it currently houses ~200,000 experimentally determined 3D structures of proteins and nucleic acids (DNA and RNA) and their complexes with one another and with small-molecule ligands (e.g., enzyme cofactors, inhibitors, drugs). Since 2003, the PDB archive has been jointly managed by the Worldwide Protein Data Bank (wwPDB, [wwpdb.org](http://wwpdb.org)) partnership.<sup>2,3</sup> Full members of the wwPDB include three founding members—US-funded RCSB Protein Data Bank (RCSB PDB),<sup>4,5</sup> the Protein Data Bank in Europe (PDBe),<sup>6</sup> and Protein Data Bank Japan (PDBj)<sup>7</sup>—plus the Electron Microscopy Data Bank (EMDB)<sup>8</sup> and the Biological Magnetic Resonance Bank (BMRB).<sup>9</sup> Protein Data Bank China (PDBc) was recently admitted to the wwPDB as an Associate Member. The wwPDB is committed to managing PDB data for users around the world at no charge for data deposition or egress, with no limitations on data usage. All PDB data are made available by wwPDB partners under the most permissive Creative Commons CC0 license ([creativecommons.org/publicdomain/zero/1.0/](http://creativecommons.org/publicdomain/zero/1.0/)). It is no exaggeration to say that the PDB was “walking the walk” decades before scholars

began “talking the talk” regarding the principles emblematic of responsible data stewardship in the modern era: FAIR (Findability, Accessibility, Interoperability, and Reusability)<sup>10</sup> and FACT (Fairness, Accuracy, Confidentiality, and Transparency).<sup>11</sup>

Figure 1a documents the relentless growth of the PDB archive from its modest beginning with just seven X-ray crystal structures of proteins purified from natural sources. Technical advances in preparative biochemistry, exogenous protein production using bacterial and eukaryotic expression hosts, sample preparation and cryogenic preservation, structural biology instrumentation and facilities, methods of structure determination, and computational power have transformed structural biology (reviewed in Reference 12). Since publication of the first X-ray crystal structure of a protein (sperm whale myoglobin),<sup>13</sup> the discipline has transformed itself from a small number of dedicated teams laboring for years, sometimes decades, to determine a single structure to a core element of the biological sciences. At the time of writing, PDB data depositors numbered more than 58,000 working on all permanently inhabited continents. Figure 1b enumerates contributions from macromolecular crystallography, nuclear magnetic resonance (NMR) spectroscopy, 3D electron microscopy (3DEM), and multiple methods to the growth of the PDB archive. For

**FIGURE 1** PDB archival data metrics (as of July 2022). (a) Growth in archive (1976–2021). (b) Growth by structure experimental method annually 2000–2021 (macromolecular crystallography (MX), nuclear magnetic resonance spectroscopy (NMR), 3D electron microscopy (3DEM), and multiple methods). Resolution of all (c) macromolecular crystallography and (d) 3D electron microscopy structures



macromolecular crystallography and 3D electron microscopy structures (now ~93% of the PDB archive), resolution distributions are illustrated in Figure 1c, d. Although a majority of PDB structures determined at better than 2.5 Å resolution come from macromolecular crystallography (~68% of the archive), 3D electron microscopy is now capable of delivering structures in the most favorable

cases at nearly 1 Å resolution (e.g., 1.15 Å resolution structure of apoferritin, PDB ID 7a6a<sup>14</sup>).

The impact of PDB data spans fundamental biology, biomedicine, bioenergy, and bioengineering and biotechnology.<sup>15,16</sup> The PDB has even found its way into the field of economics.<sup>17</sup> Literature citations running from Agriculture to Zoology attest to the breadth of impact,

traversing all of the natural, mathematical, and engineering sciences.<sup>18</sup> It is no exaggeration to state that recent advances in de novo protein structure prediction by AlphaFold2<sup>19</sup> and RoseTTAFold<sup>20</sup> were only possible because of existence of the PDB as an enormous public archive of rigorously validated and expertly biocurated 3D biostructures.<sup>21</sup> These “gold-standard” data served as a reliable source of training and testing datasets for development of artificial intelligence/machine learning-based tools. Within biomedicine, PDB data have transformed our understanding of human health and disease,<sup>22–24</sup> and continue to provide valuable starting points for structure-guided discovery of small-molecule drugs and design of new biologics agents.<sup>23,25</sup>

United States federal funders have supported the PDB since its inception. As the US wwPDB data center, RCSB PDB is jointly funded by the National Science Foundation, the National Institutes of Health, and the US Department of Energy. Safeguarding and nurturing the PDB archive and providing open access to PDB data is the work of four coordinated RCSB PDB services, encompassing data deposition; archive management and access; data exploration; and training, outreach, and education.

This invited Tools Issue contribution describes how RCSB PDB (i) is organized to deliver services efficiently to many millions of users worldwide; (ii) works with wwPDB partners to process, validate, and biocurate ~16,000 new depositions annually; (iii) acts as the wwPDB-designated Archive Keeper; (iv) enables exploration and 3D visualization of PDB data integrated with

more than 50 external data resources through its research-focused web portal [RCSB.org](https://www.rcsb.org); and (v) supports training, outreach, and education through its introductory [PDB101.RCSB.org](https://www.rcsb.org/pdb101) web portal.

## 2 | RESULTS

### 2.1 | RCSB PDB organizational structure

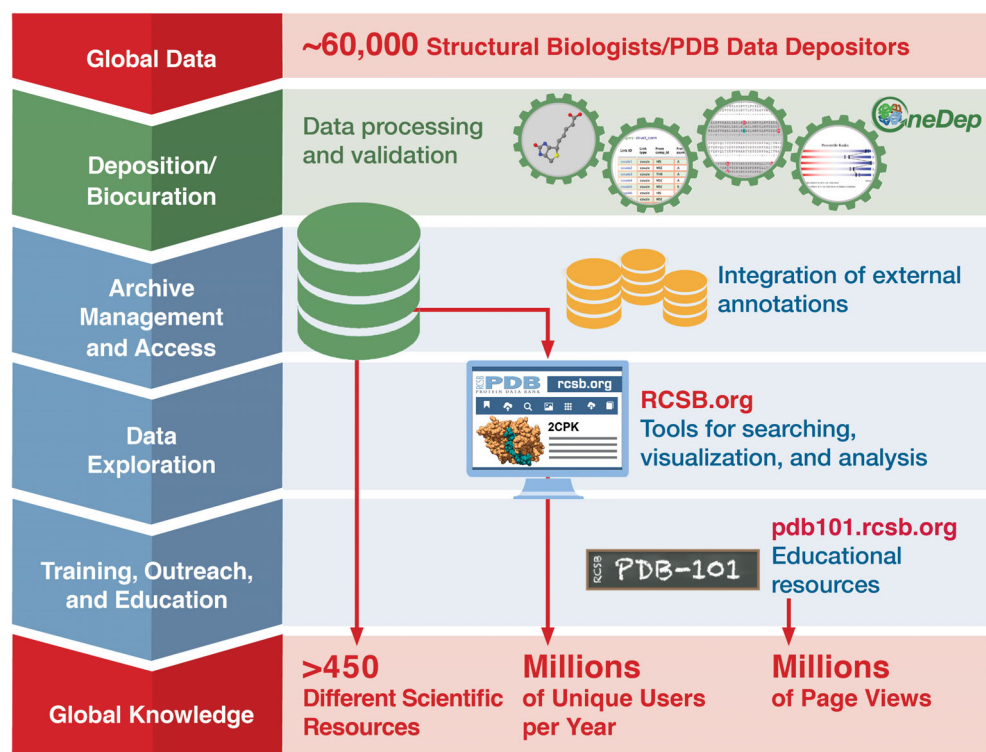
Four RCSB PDB services support transformation of global data into global knowledge (Figure 2).

#### 2.1.1 | Service 1: data deposition

The wwPDB OneDep software system<sup>26</sup> is used for global management of deposition, validation, expert biocuration, and remediation of macromolecular crystallography, 3D electron microscopy, nuclear magnetic resonance spectroscopy, and micro-electron diffraction structures, experimental data, and related metadata. RCSB PDB does not charge for data deposition.

#### 2.1.2 | Service 2: archive management and access

As the wwPDB-designated Archive Keeper,<sup>27</sup> RCSB PDB safeguards the PDB archive and maintains the PDBx/



**FIGURE 2** PDB data life cycle and RCSB PDB services. RCSB PDB hosts four integrated, interdependent cyberinfrastructure services, supported by a Customer Service Help Desk and an Infrastructure Team



mmCIF data dictionary (see below) that enables organization and searching of archived data. Programmatic access to PDB data is made available through multiple application programming interfaces (APIs).<sup>28</sup> Strict adherence to the PDBx/mmCIF data standard enables facile integration of 3D structure information with ~50 trusted external data resources (Table 1). RCSB PDB does not charge fees for data egress and there are no limitations on data usage.

### 2.1.3 | Service 3: data exploration and delivery

The RCSB PDB research-focused [RCSB.org](https://www.rcsb.org) web portal supports PDB data searching and download, browsing, 3D visualization, custom report generation, and analyses, again at no charge to users with no limitations on data usage.<sup>5,29</sup>

### 2.1.4 | Service 4: outreach and education

PDB-101 delivers training, outreach, and education resources at no charge via its introductory [PDB101.RCSB.org](https://www.pdb101.rcsb.org) web portal focused on structural biology and its impact across the sciences.<sup>30</sup>

Additional elements of RCSB PDB operations include a *Customer Service Help Desk*, supporting 3D structure depositors and PDB data consumers around the world, and an *Infrastructure Team*, working to ensure >99% 24 × 7 × 365 service availability uptime. Status of RCSB PDB servers, microservices, and APIs is monitored by NS1 ([ns1.com](https://ns1.com)) and made publicly available on a real time basis at [status.rcsb.org](https://status.rcsb.org).

RCSB PDB operations currently spans three performance sites, including its headquarters at Rutgers, The State University of New Jersey, plus smaller teams housed at the University of California San Diego-San Diego Supercomputer Center (UCSD-SDSC) and the University of California San Francisco (UCSF). RCSB PDB sub-teams have complementary roles within the organization. All *Service 1* team members are based at Rutgers. A bi-coastal team of data managers and software developers supports both *Services 2* and *3*. *Service 4* training, outreach, and education team members are based at Rutgers. The *Customer Service Help Desk* is coordinated from Rutgers. *Infrastructure Team* members are divided between Rutgers and UCSD-SDSC. The RCSB PDB team contains specialists in method-specific biocuration, software development, IT infrastructure, training, outreach, and education who work together to ensure that the RCSB PDB delivers high quality data and services.

Projects developed across all services receive guidance from the RCSB PDB Advisory Committee (recognized experts in fields, including but not limited to, structural biology, cell and molecular biology, computational biology, information technology, and education). The wwPDB Advisory Committee, an international team of experts in macromolecular crystallography, 3D electron microscopy, nuclear magnetic resonance spectroscopy, bioinformatics, and data science additionally provides guidance on matters surrounding the PDB archive.

All four RCSB PDB Services can be accessed from the [RCSB.org](https://www.rcsb.org) home page depicted in Figure 3. Navigation menus arrayed across the top of the page provide access to drop-down menus, with a selection echoed in the tabs in the center panel. Deposit provides access to all *Service 1* features. *Services 2* and *3* can be accessed from: Search, Visualize, Analyze, and Download. Learn provides access to all *Service 4* resources. Miscellaneous information regarding RCSB PDB is available from the More navigation menu. Documentation provides access to General Help and documentation for the entire web portal. Documentation and news updates are fully text searchable using the pull-down option from the search bar. Careers advertises current job openings across the organization.

## 2.2 | RCSB PDB Service 1: Data deposition

All PDB structures are deposited through the wwPDB OneDep software system ([deposit.wwpdb.org](https://deposit.wwpdb.org)), which supports structure deposition,<sup>26</sup> rigorous validation,<sup>31–33</sup> expert biocuration,<sup>34</sup> and occasional remediation.<sup>35–37</sup> Experimental methods supported by OneDep include macromolecular crystallography, 3D electron microscopy, nuclear magnetic resonance spectroscopy, and micro-electron diffraction. In addition to OneDep, RCSB PDB offers GroupDep<sup>38</sup> for deposition of large numbers (>10) of closely related macromolecular crystallography structures (e.g., results from crystallographic fragment screening<sup>39</sup>) via requests sent to [deposit@deposit.rcsb.org](mailto:deposit@deposit.rcsb.org). Development of OneDep enhanced the rate at which the international team of wwPDB biocurators can process incoming PDB data, permitting scaling of *Service 1* operations without the need of additional personnel.

During 2021, wwPDB partners processed 14,573 experimental structure depositions, a number slightly lower than the record of 15,436 set in 2020 during the COVID-19 lockdown but higher than in 2019 (13,377). During this same period, RCSB PDB processed ~39% of the global depositions (from the Americas and Oceania, and GroupDep<sup>38</sup> users globally), PDBe processed ~29% of global depositions (from Europe and Africa), and PDBj

**TABLE 1** Trusted external resources/data content integrated with PDB data. In response to community input, RCSB PDB continues to integrate new trusted external data resources as updated at [rcsb.org/docs/general-help/data-from-external-resources-integrated-into-rcsb-pdb](https://rcsb.org/docs/general-help/data-from-external-resources-integrated-into-rcsb-pdb)

Resource	Description
AlphaFold DB <sup>19,62</sup>	Computed structure models by AlphaFold2
ATC	Anatomical therapeutic chemical (ATC) classification System from World Health Organization
Binding MOAD <sup>63</sup>	Binding affinities
BindingDB <sup>64</sup>	Binding affinities
BMRB <sup>65</sup>	BMRB-to-PDB mappings
Catalytic site Atlas <sup>66</sup>	Active sites and catalytic residues in enzymes
CATH <sup>67</sup>	Protein structure classification
Cambridge structural Database <sup>68</sup>	Crystallographic small molecule data from the Cambridge crystallographic data Centre
ChEBI <sup>69</sup>	Chemical entities of biological interest
ChEMBL <sup>70</sup>	Manually curated database of bioactive molecules with drug-like properties
DrugBank <sup>71</sup>	Drug and drug target data
ECOD <sup>72</sup>	Evolutionary classification of protein domains
EMDB <sup>8</sup>	3D electron microscopy density maps and associated metadata
ExplorEnz <sup>73</sup>	IUBMB enzyme nomenclature and classification
Gencode <sup>74</sup>	Gene structure data
Gene Ontology <sup>75</sup>	Gene structure data
Genotype-tissue expression (GTEx) <sup>76</sup>	Tissue-specific gene expression data
GlyCosmos <sup>77</sup>	Web portal integrating the glycosciences with the life sciences
GlyGen <sup>78</sup>	Data integration and dissemination resource for carbohydrates and glycoconjugates
GlyTouCan <sup>79</sup>	Glycan structure repository
Human gene nomenclature committee ( <a href="https://www.genenames.org">genenames.org</a> )	Human gene name nomenclature and genomic information
IMGT <sup>80</sup>	International ImMunoGeneTics information system
Immune epitope Database <sup>81</sup>	Antibody and T cell epitopes
International mouse phenotyping consortium (IMPC, <a href="https://www.mousephenotype.org">mousephenotype.org</a> )	Mouse gene phenotype data
MemProtMD <sup>82</sup>	Database of membrane proteins embedded in lipid bilayers
ModelArchive ( <a href="https://modelarchive.org">modelarchive.org</a> )	Computed structure models (e.g., by RoseTTAFold)
Mpstruc ( <a href="https://blanco.biomol.uci.edu/mpstruc/">https://blanco.biomol.uci.edu/mpstruc/</a> )	Classification of transmembrane protein structures
NCBI Gene <sup>47</sup>	Gene info, reference sequences, et al.
NCBI Taxonomy <sup>47</sup>	Organism classification
NDB <sup>83</sup>	Experimentally determined nucleic acids and complex assemblies
OPM <sup>84</sup>	Orientations of proteins in membranes database; classification of transmembrane protein structures and membrane segments
PDBbind-CN <sup>85</sup>	Binding affinities
PDBflex <sup>86</sup>	Protein structure flexibility
PDBTM <sup>87</sup>	Protein Data Bank of Transmembrane Proteins
Pfam <sup>88</sup>	Protein families
Pharos <sup>89</sup>	Drug targets and diseases
<a href="https://protein-diffraction.org">ProteinDiffraction.org</a> ( <a href="https://proteindiffraction.org">proteindiffraction.org</a> )	Diffraction images
PubChem <sup>90</sup>	Chemical information

TABLE 1 (Continued)

Resource	Description
PubMed <sup>47</sup>	Citation information
PubMedCentral <sup>47</sup>	Open access literature
RECOORD <sup>91</sup>	Nuclear magnetic resonance spectroscopy structure ensembles
RESID <sup>92</sup>	Protein modifications
SAbDab <sup>93</sup>	The structural antibody database
SBGrid <sup>94</sup>	Structural biology data grid/diffraction images
SCOP <sup>95</sup>	Structural classification of proteins
SCOPe <sup>96</sup>	Structural classification of proteins—Extended
SIFTS <sup>97</sup>	Structure, function, taxonomy, sequence
Thera-SAbDab <sup>98</sup>	Therapeutic structural antibody database
UniProt <sup>46</sup>	Protein sequences and annotations

processed ~32% of global depositions (from Asia and the Middle East; reflecting significant year-on-year growth in the number of depositions coming from the People's Republic of China). A total of 12,593 new structures were publicly released into the PDB during 2021, which is lower than the record of 14,009 set in 2020 but again higher than in 2019 (11,497). Year-to-date deposition trends are on track (as of early October) for the number of new depositions in 2022 to exceed 16,000.

*Service 1* support for global PDB data deposition embodies four elements, including Prepare Data, Validate Data, Deposit Data, and related Help & Resources, all accessible from the Deposit menu and home page panel at [RCSB.org](https://www.rcsb.org).

- i. Prepare Data provides access to information about the PDBx/mmCIF data dictionary underpinning the entire PDB archive, plus software tools developed by RCSB PDB (e.g., `pdb_extract`, SF-Tool, Ligand Expo, Maxit)<sup>40–43</sup> that are used frequently by structural biologists for organizing information prior to data deposition.
- ii. Validate Data provides access to the stand-alone, anonymous wwPDB Validation Server and APIs, which should be used to identify possible issues with experimental data and/or atomic coordinates during structure determination. Depositors are strongly encouraged to resolve any issues identified by the Validation Server before proceeding to structure submission. Additional resources available under Validate Data include Information for Journals, pertaining to use of wwPDB Validation Reports during manuscript review, and Validation Task Forces, describing related community engagement. Multiple descriptions of structure validation have been published in peer-reviewed journals.<sup>31–33,35,44,45</sup>

- iii. Deposit Data provides access to the wwPDB One-Dep<sup>26</sup> global system for structure deposition, validation,<sup>31–33</sup> biocuration,<sup>34</sup> and remediation.<sup>35–37</sup> This software tool was developed jointly by wwPDB partners. It is used by all wwPDB data centers worldwide to ensure data uniformity, completeness, and consistency across the PDB archive. Data depositors now benefit from a streamlined deposition process, reducing the time required for data entry while ensuring complete data capture. Every few years (or when new features are added to validation software), wwPDB Validation Reports are recomputed for all structures in the archive to stay abreast of improvements in average structure quality and support statistics-based detection of outliers. wwPDB biocuration processes scrutinize biopolymers, small molecule ligands, and carbohydrates through different OneDep Modules. The Sequence Module cross-checks biopolymer sequences against UniProt<sup>46</sup> and NCBI<sup>47</sup> databases for taxonomy and sequence reference annotation. Small molecule ligands are extracted from the coordinates, searched for matched ligands in the wwPDB Chemical Component Dictionary (Chemical Component Dictionary),<sup>48</sup> and standardized with nomenclature in the Ligand Module. This module also allows biocurators to create Chemical Component Dictionary definitions for new ligands. Saccharide units covalently bound to a biopolymer are represented as branched carbohydrates based on connectivity (e.g., N-glycosylation).

Depositors are encouraged to submit macromolecular crystallography data in bulk using RCSB PDB Group Deposition capabilities,<sup>38</sup> where appropriate (e.g., crystallographic fragment screening). The efficiency of GroupDep, as it is informally known, encourages capture of valuable data that might

RCSB PDB

Deposit Search Visualize Analyze Download Learn More Documentation Careers

MyPDB Contact us

RCSB PDB PROTEIN DATA BANK

196,565 Structures from the PDB

1,000,361 Computed Structure Models (CSM)

3D Structures Enter search term(s), Entry ID(s), or sequence Include CSM

Advanced Search Browse Annotations Help

PDB-101 wwPDB EMDatabank NUCLEIC ACID DATABASE wwPDB Foundation

NEW! Computed Structure Models (CSM) Learn more

Welcome

Deposit

Search

Visualize

Analyze

Download

Learn

RCSB Protein Data Bank (RCSB PDB) enables breakthroughs in science and education by providing access and tools for exploration, visualization, and analysis of:

- Experimentally-determined 3D structures from the Protein Data Bank (PDB) archive
- Computed Structure Models (CSM) from AlphaFold DB and ModelArchive

These data can be explored in context of external annotations providing a structural view of biology.

COVID-19 CORONAVIRUS Resources

Join the RCSB PDB Team

October Molecule of the Month

Phytohormone Receptor DWARF14

Latest Entries As of Tue Oct 11 2022

8E9G

Mycobacterial respiratory complex I with both quinone positions modelled

Features & Highlights

Undergrads/Grads: Apply to the Molecule of the Month Boot Camp (January 2023)

Limited spaces available for the Science Communication in Biology and Medicine Virtual Boot Camp: January 9-13, 2023. Applications due October 31, 2022.

Improved EM validation with Q-score

wwPDB validation of EM structures for which there is both a model and an EM volume will include the Q-score metric

Register Now for Virtual Crash Course: Exploring Computed Structure Models from Artificial Intelligence/Machine Learning at RCSB.org

Learn how to search, visualize, and

News

Publications

Happy Birthday, Irving Geis

Celebrate Geis' birthday (October 18, 1908) with a tour of the Geis Digital Archive of his pioneering works of biomolecular art at PDB-101

10/16/2022

Education Corner: Inktober SciArt

Irina Bezsonova (UConn Health) describes how she created Inktober SciArt Celebrating PDB50 in 2021. Images of her amazing PDB-inspired drawings are available for download.

10/13/2022

wwPDB Charter: Full and Associate Members

wwPDB is committed to responsible, international stewardship of public

FIGURE 3 The RCSB PDB research-focused web portal home page (RCSB.org)

otherwise remain locked up within individual laboratories. Integrative or hybrid methods structures determined with more than one experimental technique (e.g., 3D electron microscopy combined with chemical cross linking) should be deposited to the wwPDB PDB-Dev prototype repository (pdb-dev.wwpdb.org).<sup>49–54</sup>

- iv. Deposition Help & Resources provides access to Deposit FAQs, Tutorials, Deposition Policies, Annotation Procedures, the PDBx/mmCIF Dictionary (see below), Chemical Component Dictionary, PDB File

Format Guides, BioSync Beamlines and Facilities information,<sup>55</sup> and Related Tools.

## 2.3 | RCSB PDB Service 2: Archive management and access

PDBx/mmCIF Data Standard: PDB data architecture is defined by the PDBx/mmCIF dictionary.<sup>56–58</sup> It is the macromolecular extension of an earlier community data standard, the Crystallization Information Framework<sup>59</sup>



([cif.iucr.org](http://cif.iucr.org)), developed by the International Union of Crystallography for small molecule X-ray diffraction. The macromolecular data standard is maintained by the wwPDB partners in collaboration with wwPDB PDBx/mmCIF Working Group domain experts ([wwpdb.org/task/mmcif](http://wwpdb.org/task/mmcif)).<sup>58</sup> wwPDB partners and the Working Group collaborate on developing terminologies for emerging and rapidly evolving methodologies (e.g., X-ray Free Electron Laser Serial Crystallography and 3D electron microscopy), and remediating (or enhancing) representations for existing data content (e.g., carbohydrates<sup>35</sup>).

Within the PDBx/mmCIF dictionary, PDB data follow a strict hierarchy. An individual PDB structure is referred to as an Entry, identified with a unique PDB ID (currently four alphanumeric characters, for example, 1q2w). Within each PDB Entry, every chemically unique molecule is defined as an Entity (including Polymer, Branched, or Non-polymer), denoted by a numeric Entity ID. Polymer Entities (including proteins or short polypeptide chains, DNA, and RNA) are made up of covalently linked chemical building blocks as described in the Chemical Component Dictionary and individually numbered according to position within the Polymer sequence. Branched Entities are either linear or branched carbohydrates, composed of saccharide units covalently linked via one or more glycosidic bonds. Non-polymer Entities are small molecules (enzyme cofactors, ligands, water molecules, etc.) and ions (e.g., Na<sup>+</sup>, Cl<sup>-</sup>, Zn<sup>++</sup>). Every Non-polymer Entity has a unique Chemical Component Dictionary ID (one to three or more character alphanumeric code). The Chemical Component Dictionary provides nomenclature standards and chemical descriptions for all Non-polymer Entities and every Polymer Entity component occurring within the PDB archive. Within a PDB Entry, there can be multiple Instances (or copies) of any particular Entity, each labeled with a unique Chain ID (one or more alphanumeric characters, for example, A, AA, ...). In nature, multiple proteins and/or nucleic acid chains frequently occur as supra-macromolecular complexes. Within the PDBx/mmCIF hierarchy, they are referred to as Assemblies, each with a unique numeric Assembly ID. The standardized nomenclature of the PDBx/mmCIF dictionary also provides the backbone of the Advanced Search interface, which is explained in greater detail below and in [RCSB.org](http://RCSB.org) Advanced Search documentation.

**Archive Keeping:** As wwPDB-designated PDB Archive Keeper, RCSB PDB is responsible for safeguarding and managing the PDB, and making its contents freely available. At the time of writing, the size of the archive was ~1 TB (excluding 3D electron microscopy experimental data stored in EMDB, which currently totals ~6 TB). The entire corpus of PDB data is safeguarded in multiple copies in secure institutional

computer machine rooms located on both US coasts, and backed up twice weekly (i.e., before and after each weekly update).

In its role as wwPDB-designated PDB Archive Keeper, RCSB PDB releases an updated version of the archive every week using a two-stage process. For each new PDB Entry slated for the upcoming weekly release, Stage One includes sequence(s) (amino acid or nucleotide) for each distinct Polymer Entity; InChI string(s) for each distinct ligand; and crystallization pH value(s), where appropriate. They are released at the wwPDB web portal (see [www.wwpdb.org/ftp/pdb-ftp-sites](http://www.wwpdb.org/ftp/pdb-ftp-sites)) every Saturday by 03:00 Universal Time Coordinated (UTC). Stage One supports weekly blind challenges for in silico prediction of protein structure (Continuous Automated Modeling Evaluation or CAMEO, [cameo3d.org](http://cameo3d.org))<sup>60</sup> and small-molecule docking (Continuous Evaluation of Ligand Protein Prediction or CELPP, [drugdesigndata.org/about/celpp](http://drugdesigndata.org/about/celpp))<sup>61</sup>. Stage Two completes the process every Wednesday at 00:00 UTC by releasing the updated PDB archive in full (currently adding ~300 new structures/week, updating previously released entries, and occasionally removing obsoleted structures).

PDB data are freely distributed via File Transfer Protocol (FTP) over Hypertext Transfer Protocol (HTTP) and remote sync (rsync), providing universal open access to the archival information in two forms (latest archive, [ftp.wwpdb.org/pub/pdb/data](http://ftp.wwpdb.org/pub/pdb/data); and latest and prior versions archive, [ftp-versioned.wwpdb.org](http://ftp-versioned.wwpdb.org)).

PDB data are also made available without fees or egress charges by Amazon Web Services (AWS) through its Open Data Sponsorship Program [registry.opendata.aws/pdb-3d-structural-biology-data/](http://registry.opendata.aws/pdb-3d-structural-biology-data/). Download tests conducted from Asia, the United Kingdom, and the United States documented significantly faster download speeds for larger PDB structure data files hosted by AWS versus delivery of the same information for the RCSB PDB web portal (and those maintained by both PDBe and PDBj). PDB data consumers wishing to carry out large-scale data downloads are strongly encouraged to do so using AWS ([s3.rcsb.org](http://s3.rcsb.org)).

During 2021, a total of ~2.36 billion PDB data files were downloaded from the archive, which is nearly double the previous record of 1.32 billion set in 2020. Historically, requests for data downloads have come from every sovereign country recognized by the United Nations. Approximately two-thirds of data downloads come from the FTP archive, with the remainder being downloaded from web portals maintained independently by RCSB PDB, PDBe, and PDBj. For the avoidance of doubt, all wwPDB partners distribute identical PDB data.

**External Data Integration:** *Service 2* also integrates PDB data with information from ~50 trusted external resources (Table 1). Integration of information from

external resources is a critical element of an exhaustive weekly process that loads the updated PDB archive into the RCSB PDB data warehouse. Thereafter, it is made available for our public data access APIs. At the time of writing, the entire corpus of data managed by RCSB PDB occupied >100 TB of digital storage.

To provide users with the most current information about a structure entry, information from trusted external resources is integrated on a weekly basis. [RCSB.org](https://www.rcsb.org) is a “living data resource,” while scholarly publications describing PDB Entries are static documents that reflect what was known about the biomolecule(s) at the time of study. Thereafter, it is not uncommon for new biological or biochemical functions of a macromolecule to come to light, or new disease-causing mutations to be identified. Such new findings are integrated with PDB data every week, thereby ensuring that [RCSB.org](https://www.rcsb.org) users have access to the most current information pertaining to every 3D biostructure in the public domain.

## 2.4 | RCSB PDB Service 3: Data exploration and delivery

Most RCSB PDB users access the archive through our [RCSB.org](https://www.rcsb.org) research-focused web portal. In 2021, more than 6.8 million unique internet protocol (IP) addresses from around the world were used to access [RCSB.org](https://www.rcsb.org), which exceeded our 2020 record of nearly 6.7 million (during the pandemic lockdown). We estimate that ~99% of these PDB data consumers are not experts in structural biology. Their research interests are very broad, encompassing fundamental biology, biomedicine, energy sciences, and bioengineering and biotechnology. *Service 3* is tasked with creation and maintenance of a public-facing web portal to serve the diverse needs of these multiple communities of users.

As described earlier, the [RCSB.org](https://www.rcsb.org) home page (Figure 3) displays navigation menus that provide access to Deposit, Search, Visualize, Analyze, Download, Learn, More, Documentation, and Careers resources. Immediately below these menus, users will find a top bar Search box that supports text searching (based on ElasticSearch, [elastic.co](https://www.elastic.co)) of the entire PDB archive or RCSB PDB Documentation, depending on the option selected. Immediately below the Search box, single click access is provided to Advanced Search and Browse Annotations features. The remainder of the home page is devoted to a description of the PDB archive and RCSB PDB, COVID-19 Coronavirus Resources, Careers, Molecule of the Month, Latest Entries, Features & Highlights, and News.

The flexibility, generality, and utility of RCSB PDB [RCSB.org](https://www.rcsb.org) web portal tools are exemplified below with a

case study focused on human programmed cell death protein 1 (PD-1), cancer immunotherapy, tumor neoantigens, monoclonal antibodies, and small-molecule drugs.

Therapeutic anti-PD-1 antibodies are used clinically to reduce the likelihood of “tumor sparing” by T-cells when they encounter malignant cells expressing program death-ligand proteins PD-L1 or PD-L2. Anti-cancer immune surveillance mechanisms, relying on detection of tumor neoepitopes (i.e., oligopeptide segments of tumor neoantigens) presented to the immune system by major histocompatibility complex (or MHC) classes 1 and 2, usually detect and kill nascent tumors before they begin to proliferate and cause disease. However, T-cell responses to tumor neoepitopes can be down regulated by a so-called “immune checkpoint” when PD-1 (normally displayed on T-cell surfaces) binds to PD-L1 (or PD-L2) present serendipitously on the surface cancer cells.<sup>99</sup> Targeting this immune checkpoint with monoclonal antibodies is revolutionizing treatment of solid tumor cancers.

President Jimmy Carter, for example, benefitted from intravenous infusions of pembrolizumab ([Keytruda.com](https://www.keytruda.com), trade name Keytruda), which put his late-stage melanoma into long-term remission and may well have cured him of his widely metastatic cancer. Pembrolizumab and other immune checkpoint therapies are thought to be most effective against high mutation burden cancers.<sup>100</sup> Both late-stage melanoma and tumors with defects in DNA mismatch repair (e.g., a subset of rectal cancers) carrying large numbers of tumor neoantigens can be treated successfully using antibodies that block interactions between PD-1 and PD-L1 (or PD-1 and PD-L2). Targeting PD-L1 with monoclonal antibodies has also proven effective in some individuals for some cancers (e.g., durvalumab, [Imfinzi.com](https://www.imfinzi.com), trade name Infizi). James Allison (MD Anderson Cancer Center) and Tasuku Honjo (Kyoto University) shared the 2018 Nobel Prize in Physiology or Medicine for their “discovery of cancer therapy by inhibition of negative immune regulation.” Honjo is credited with the discovery of PD-1, the target of pembrolizumab and nivolumab ([Opdivo.com](https://www.opdivo.com), trade name Opdivo).

**Structure Summary Page:** For a topic like this, users often begin by viewing individual PDB entries, typically described in peer-reviewed publications that enumerate PDB structure IDs explicitly. When a single PDB ID (e.g., PDB structure 5jxe<sup>101</sup>) is entered into the [RCSB.org](https://www.rcsb.org) top bar Search box and users press the Return key or click the Magnifying Glass icon, they are taken directly to the relevant Structure Summary Page. Each Structure Summary Page provides summary information for that entry, with “tabs” linking to web portal pages displaying more detailed information specific to the entry. Figure 4a

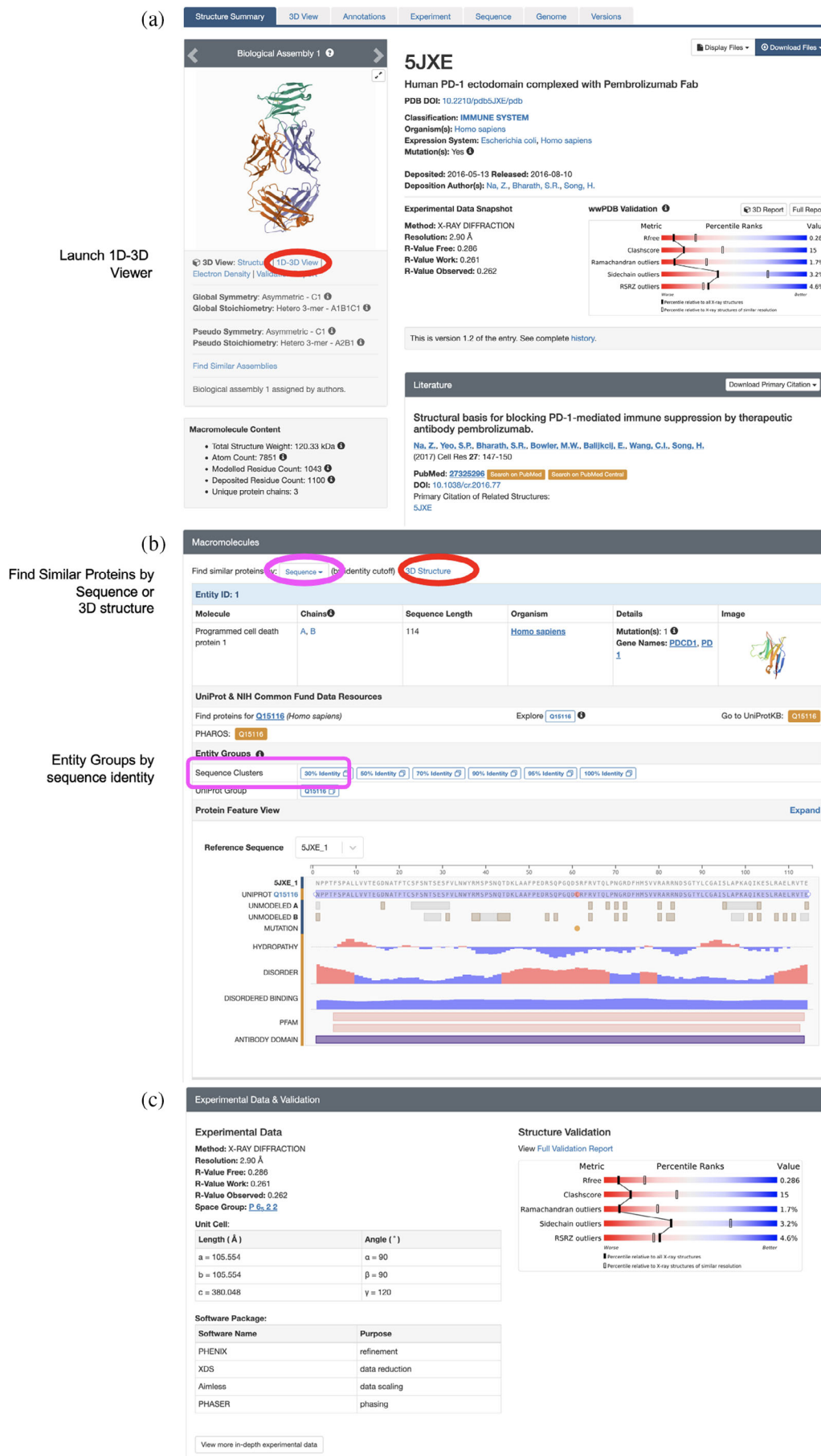
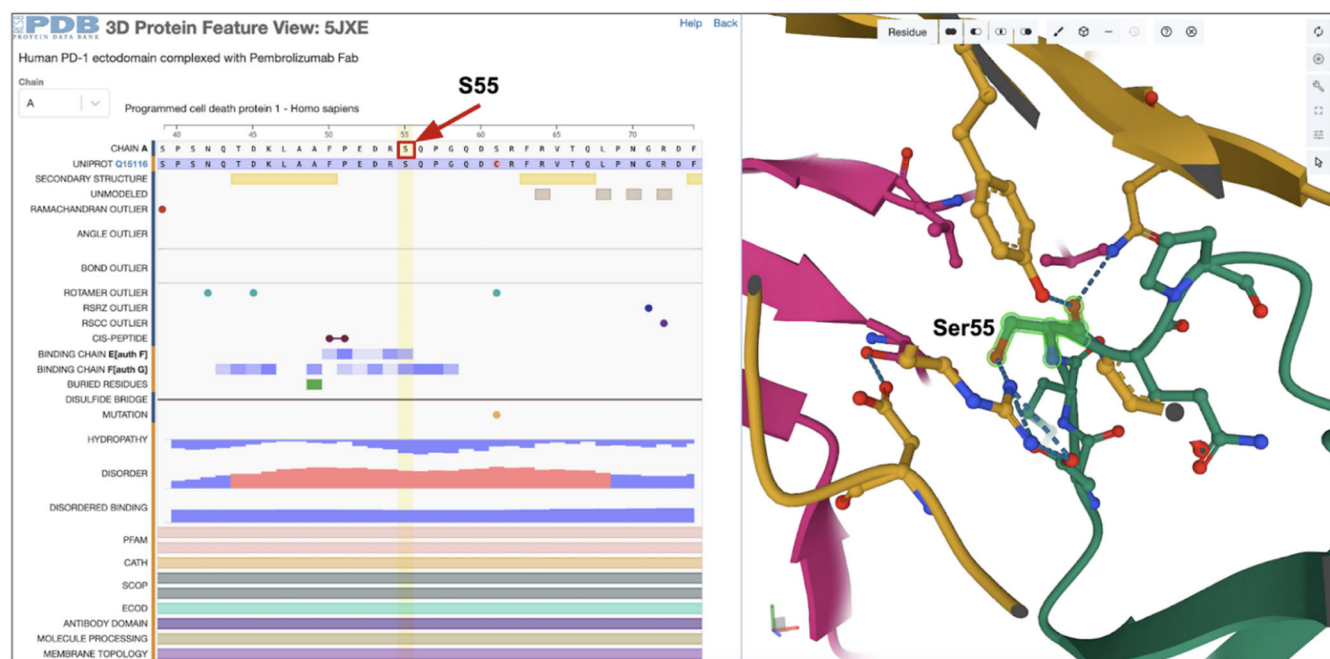


FIGURE 4 Legend on next page.

shows the upper portion of the PDB ID 5jxe Structure Summary Page, which provides structure images of pembrolizumab anti-PD1 monoclonal antibody Fab recognizing the ectodomain of PD-1, a summary of structure contents, wwPDB Validation quality assessment, and the primary literature citation. The Macromolecules section (Figure 4b) contains three subsections: the first providing a detailed summary for the two copies of human PD-1 present in the PDB Entry, plus two subsections each for two copies of the light and heavy chain of the Fab. The bottom portion of the Structure Summary Page (Figure 4c) includes sections for Experimental Data & Validation (providing a summary of the X-ray crystallographic experiment plus a button to View more in-depth experimental information) and an Entry History section with deposition and revision information.

1D-3D Data Visualization: PDB structure 5jxe reveals at the atomic level how pembrolizumab recognizes the

ectodomain of PD-1 on the surface of human T-cells. More details regarding the structure can be explored by clicking the 1D-3D View button in the upper section of the Structure Summary Page (Figure 4a, red ellipse), which takes the user to a simultaneous graphical display<sup>102</sup> of the one-dimensional (1D) amino acid sequences (Figure 5, left) and the 3D structure (Figure 5, right) with various annotations incorporated into the 1D view,<sup>103</sup> including identification of PD-1 residues involved in “Binding Chain F” (heavy chain of the pembrolizumab monoclonal antibody). The right-hand 3D view panel uses the Mol\* molecular graphics open-source software system,<sup>104</sup> first deployed on the RCSB.org web portal in 2020. Mol\* operates entirely within a web browser without the need to download additional software. It also inter-operates with the PDBx/mmCIF data standard, allowing 3D graphical delivery of value-added annotations at the level of individual amino acid



**FIGURE 5** 1D-3D View for PDB ID 5jxe. In the left panel, clicking on Ser55 (highlighted with a red arrow) on the track CHAIN A (at the top of the page) invokes 3D graphical display of the view on the right, showing a hydrogen bond between the sidechains of Ser55 of human PD-1 (labeled Ser55) and Arg99 of the pembrolizumab Heavy Chain (not labeled). Atom color coding: O-red; N-blue, C-green (for PD-1), C-yellow (for pembrolizumab Heavy Chain), and C-pink (for pembrolizumab Light Chain)

**FIGURE 4** Structure Summary Page for PDB ID 5jxe. (a) Upper section hosts structure images, summary of structure contents, wwPDB Validation quality assessment, and primary literature citation. The link to the 1D-3D viewer is highlighted with a red ellipse.

(b) Macromolecules Section: First of three subsections providing detailed summary for the two copies of human PD-1 present in the structure. (N.B.: Additional subsections for two copies each of the heavy and light chains of the pembrolizumab monoclonal antibody Fab are not shown.) “Find similar proteins by:” can be invoked for Sequence similarity searching (magenta ellipse) from a pull-down menu to select desired percentage of sequence similarity. “Find similar proteins by:” can be invoked for 3D Structure similarity searching (red ellipse) with a single mouse click. The Entity Group Sequence Clusters 30% Identity search button is denoted with a magenta rectangle.

(c) Experimental Data & Validation and Entry History Sections



residues. Extensive introductions to Mol\* can be found in a 2021 *Nucleic Acids Research* Web Server issue article,<sup>104</sup> the 2022 *Protein Science* Protein Tools special issue,<sup>29</sup> and [RCSB.org](https://www.rcsb.org) documentation.

Mousing over the interacting PD-1 residues in the 1D view highlights locations of these stabilizing interactions in the 3D view. Clicking on any one of the interacting PD-1 residues (e.g., Ser55, Figure 5, red arrow) within the 1D view causes the 3D view to center on the residue and display all non-H atoms within 5 Å, showing hydrogen bonds and other interactions as dashed lines (e.g., H-bond between PD-1 Ser55 and pembrolizumab heavy chain Arg99, Figure 5, right).

**Structure-guided Data Exploration:** The PDB archive often includes multiple entries related to a particular topic, which may be found using the “Find similar proteins by: Sequence” button on the Structure Summary Page. This feature is available for every Polymer Entity in the Entry (Figure 4b: magenta ellipse, with pull-down menu for 100% identity, 95%, ...). Starting from the Structure Summary Page for PDB structure 5jxe, invoking “Find similar proteins by: Sequence (100% identity)” returns a results page listing 10 Polymer Entities that are exact matches in amino acid sequence to human PD-1. Scrolling part way down this list reveals PDB structure 4zqk,<sup>105</sup> entitled “Structure of the complex of human programmed death-1 (PD-1) and its ligand PD-L1.” Clicking on 4zqk takes the user to the Structure Summary Page for this related structure. Selecting the 1D-3D View button then allows the user to identify the residues in PD-1 involved in molecular recognition and explore interactions between PD-1 and PD-L1 at the atomic level, as described above.

Alternatively, clicking on the “Find similar proteins by: 3D Structure” button on the structure summary page of PDB ID 5jxe (Figure 4b, red ellipse) returns a results page listing >100 polymer entities with similar 3D structure (root-mean-square-deviation or RMSD values ranging from 0.0 Å for the exact match to ~2.1 Å). The large number of structurally similar proteins represented in the PDB archive reflects the ubiquity of the immunoglobulin-like beta sandwich fold, as classified by SCOP<sup>95</sup>/SCOPe,<sup>96</sup> SCOP2,<sup>106</sup> ECOD,<sup>72</sup> CATH,<sup>67</sup> IMGT antibody Annotation,<sup>80</sup> and SABDab antibody Annotation,<sup>93</sup> all of which are available from the annotations tab on the structure summary page.

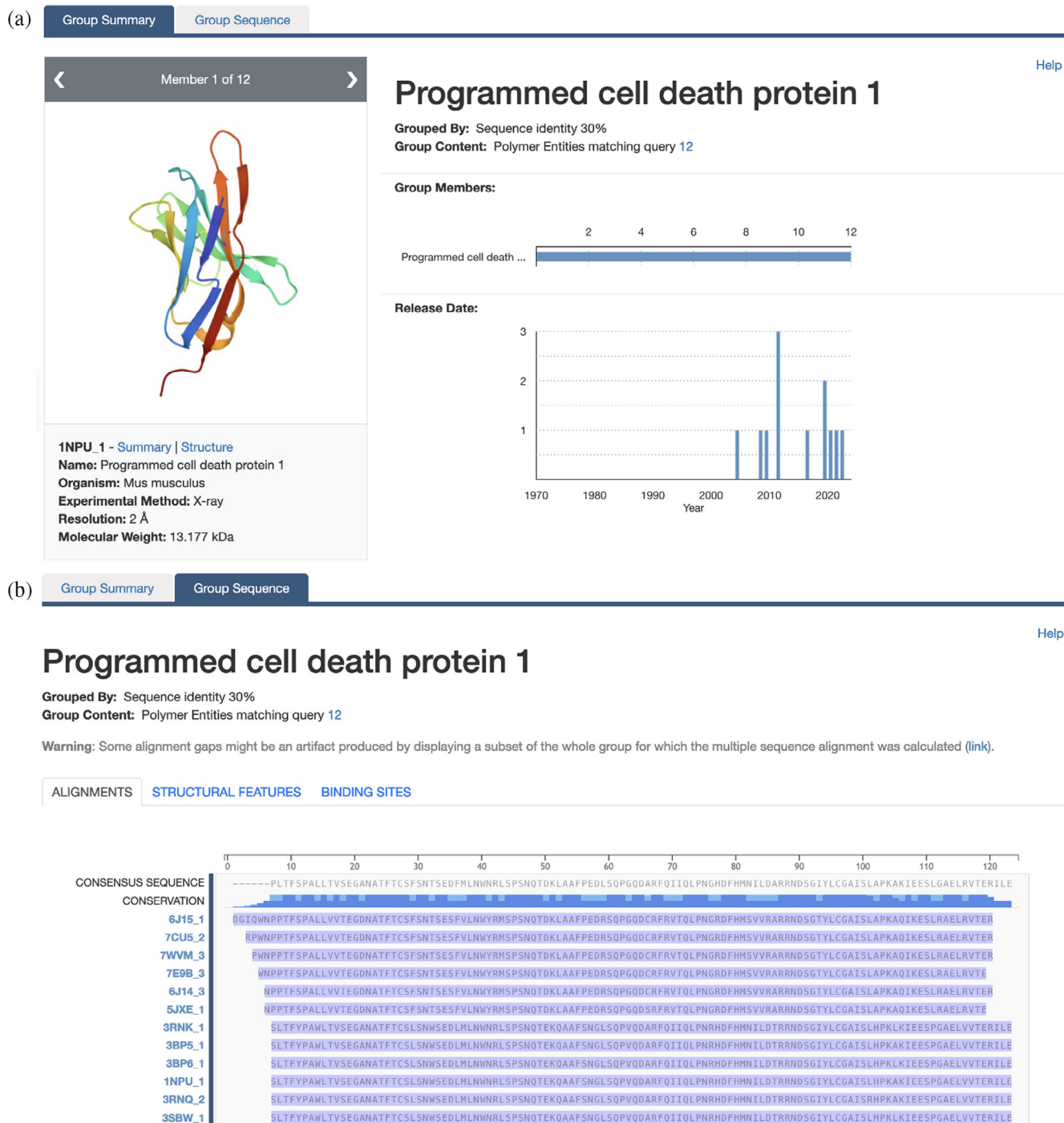
Another option for exploring related proteins is supported by the newly introduced RCSB PDB Group feature, located immediately below the description of each protein Polymer Entity occurring in a given PDB structure (Figure 4b). For the human PD-1 Polymer Entity in PDB structure 5jxe, clicking on the 30% identity Sequence Cluster button (Figure 4b, magenta rectangle) delivers

the user to the Group Summary page depicted in Figure 6a, which displays information for a cluster of 12 related proteins (encompassing all occurrences of human and murine PD-1 represented in the PDB archive). Clicking on the Group Sequence option reveals a multiple sequence alignment of all 12 related proteins (Figure 6b). Additional viewing options on the Group Sequence page include Structural Features (including a consensus sequence, secondary structural elements, and CATH and SCOP annotations) and Binding Sites (identifying amino acid residues in contact with various monoclonal antibody Fabs, PD-L1, and PD-L2).

**Additional Functionality:** Additional Structure Summary Page functionality can be accessed using the other top tabs shown in Figure 4a, which include 3D View, Experiment, Sequence, Genome, and Versions. Clicking on the 3D View tab takes the user to a stand-alone Mol\* 3D graphics window, equivalent to the right-hand portion of the 1D-3D View. The Experiment tab (typically used only by structural biologists) displays a detailed summary of the X-ray crystallography study yielding the structure. The Sequence tab takes the user to the equivalent of the left-hand portion of the 1D-3D View. The Genome tab presents an alignment of the human PD-1 polypeptide chain with human chromosome 2. (N.B.: At the time of writing, this functionality was only supported for 1,614 organisms.) Clicking on the Versions tab displays information regarding any revisions to the PDB Entry following its initial release. Going beyond necessary minor revisions of a clerical nature made to individual PDB entries by wwPDB biocuration team members from time to time, depositors-of-record (i.e., principal investigators, research team leaders) are encouraged to “correct” their PDB entries (wherever appropriate) by providing updated atomic coordinate files. Since 2018, atomic coordinate versioning without triggering issuance of a new PDB ID has been supported by the wwPDB in response to depositor feedback.

## 2.5 | Searching the PDB archive

In cases where users do not know the PDB ID of a relevant Entry, multiple options for searching the archive are available on the [RCSB.org](https://www.rcsb.org) web portal. The top bar Search box located in the upper section of the homepage is the workhorse of RCSB PDB search capabilities. When non-PDB ID text is entered into the Search box, a drop-down list appears with possible search refinements identifying the input text as being found within Additional Structure Keywords, Structure Title, Structure Author, Polymer Entity Description, Gene Name Source Organism, Taxonomy Name, etc. Clicking on one of the options in the



**FIGURE 6** Entity Group Sequence Clusters 30% Identity for human PD-1. (a) Group Summary page, consisting of three components: ribbon representation drawings for individual Group Members with buttons provide access to the Structure Summary Pages (Summary) and a Mol\* display (Structure); a graphical table showing Group Members; and bar-graph summarizing public release dates for individual Group Members. (b) Group Sequence page, providing a multiple sequence alignment of Group Members. Additional tabs on the Group Sequence page provide access to information about Structural Features and Binding Sites

drop-down list takes the user to a combined results page. In most cases, scrolling down the results page will allow identification of relevant PDB structures for exploration via their Structure Summary Pages by clicking on a PDB ID. Another option supports free text searches, which are carried out most expeditiously when the phrase of

interest is enclosed within double quotes. Otherwise, structures containing any of the text words in the query will be returned and may include false positives.

Directed searches using full Boolean logic are supported by the Advanced Search page, which can be reached from the top of any [RCSB.org](https://www.rcsb.org) web portal page by

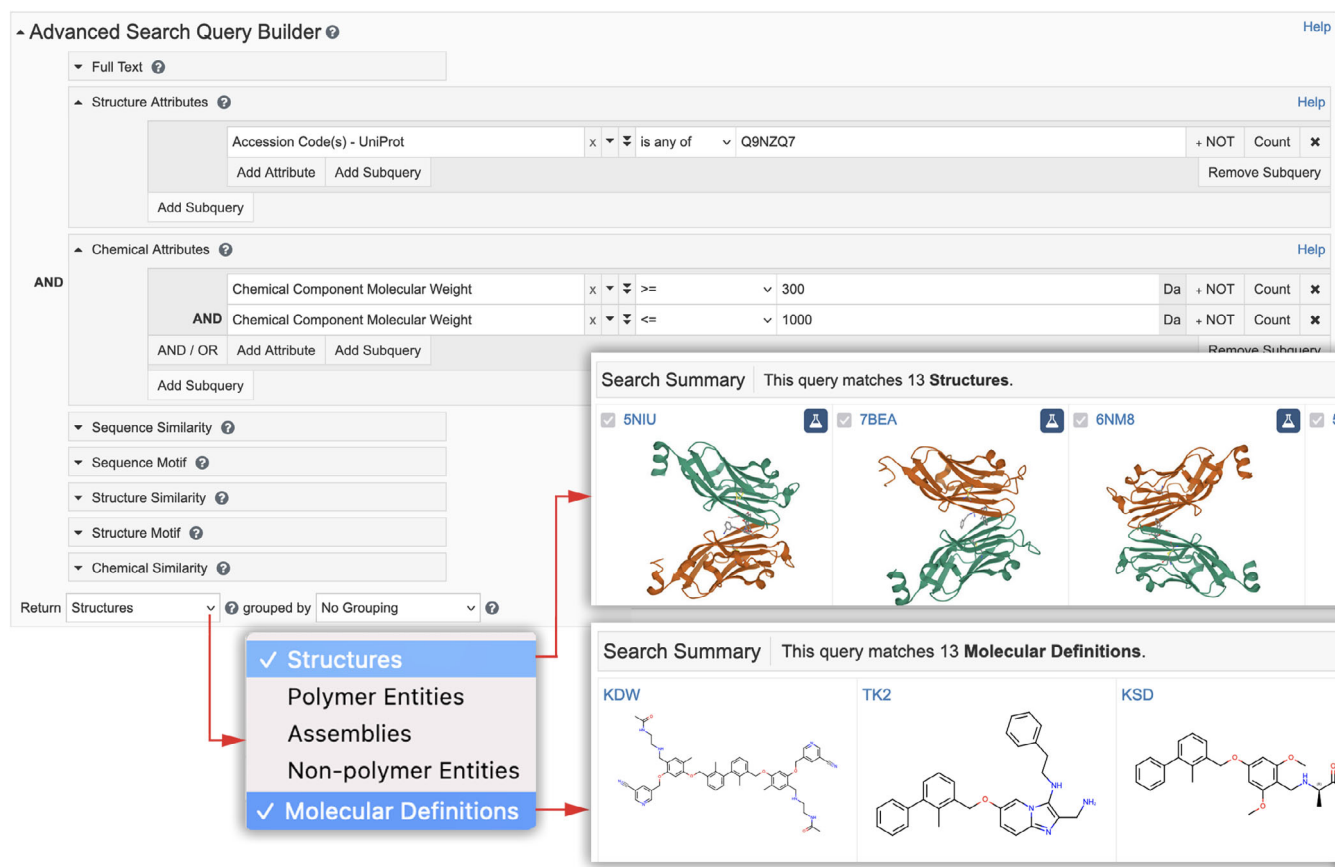
clicking on the Advanced Search button. The example depicted in Figure 7 was used to identify PDB structures of human PD-L1 (UniProt Accession Code Q9NZQ7) that also contain small-molecule ligands with molecular weight  $\geq 300$  and  $\leq 1000$  Da. At the time of writing, invoking this Advanced Search yielded 13 PDB structures containing human PD-L1 bound to a distinct small-molecule ligand. Toggling between Return Structures and Return Molecular Definitions (highlighted in figure) changes the output from individual PDB structures to individual small-molecule ligands.

These 13 PDB structures were described in six peer-reviewed journal publications by a number of research groups focused on discovering and developing orally bio-available, small-molecule drugs as alternatives to monoclonal antibody immune checkpoint anti-cancer therapies.<sup>107–113</sup> Such ligands are thought to reduce tumor immune evasion via small molecule-induced dimerization PD-L1 and subsequent internalization. This effect on the immune checkpoint is mechanistically distinct from that of anti-PD-1 and anti-PD-L1 monoclonal antibodies, which interdict PD-1/PD-L1 mediated

interactions between T-cells and tumor cells. Disappearance of PD-L1 from the surface of the tumor cell prevents them from engaging with T-cells in a manner that could down regulate the immune response to the tumor. A comparable Advanced Search intended to identify 3D structures of human PD-1 (UniProt Accession Code Q15116) bound to small-molecule ligands in PDB detected zero matches at the time of writing. Effective small-molecule drugs targeting immune checkpoints would provide access to cancer immunotherapies for millions of patients around the world unable to afford monoclonal antibody treatments (for reference, the current list price for a full 2-year course of pembrolizumab is ~US \$360,000, plus ancillary charges for intravenous infusions occurring every 3 or 6 weeks).

## 2.6 | Pairwise structure comparison

The Pairwise Structure Comparison tool is one of the most popular [RCSB.org](https://www.rcsb.org) web portal features and was described in detail in the 2022 Protein Science



**Advanced Search Query Builder**

**Full Text**

**Structure Attributes**

Accession Code(s) - UniProt  is any of

**Chemical Attributes**

Chemical Component Molecular Weight   $\geq$

Chemical Component Molecular Weight   $\leq$

**Search Summary** This query matches 13 Structures.

5NIU 7BEA 6NM8

**Search Summary** This query matches 13 Molecular Definitions.

KDW TK2 KSD

Return Structures

☒ Structures

☐ Polymer Entities

☐ Assemblies

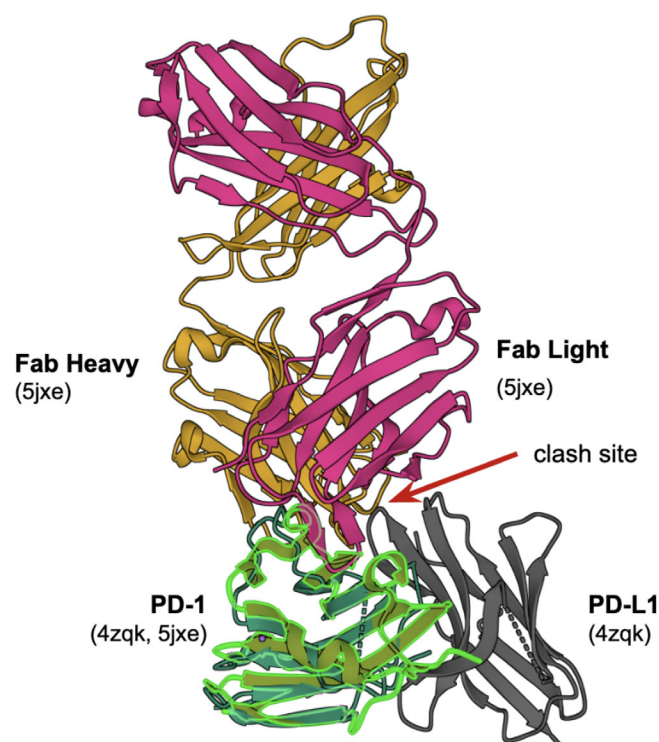
☐ Non-polymer Entities

☒ Molecular Definitions

**FIGURE 7** Advanced Search Query Builder used to combine Structure Attribute search for PDB structures containing human PD-L1 (UniProt ID Q9NZQ7) with Chemical Attribute search for ligands with molecular weight between 300 and 1000 Da occurring is shown. The Options for returning Structure data and Molecular Definition data are highlighted

Tools Issue.<sup>29</sup> Available from the Analyze navigation menu, Pairwise Structure Comparison supports superposition and quantitative comparison of 3D structures using Java implementations of well-established tools (e.g., Combinatorial Extension or CE,<sup>114</sup> CE with Circular Permutations,<sup>115</sup> FATCAT,<sup>116</sup> TM-align,<sup>117</sup> and superposition guided by Smith-Waterman local sequence alignment<sup>118</sup>) provided by the BioJava project.<sup>119</sup> This feature has also been implemented within the stand-alone version of the Mol\* 3D Viewer, which can be accessed from the Visualize navigation menu. Users can open atomic coordinate files and import them into the Mol\* session in either legacy PDB or PDBx/mmCIF formats, or simply download them directly from the PDB archive. Once the two PDB structures are visible on the Mol\* 3D-Canvas, clicking on the Superposition Panel allows the user to select displayed objects (i.e., by chains or by atoms) and superpose them.

Figure 8, downloaded by activating the Screenshot feature (camera iris icon) available on the Mol\* 3D-Canvas and clicking the Download button, illustrates the result of superposing Chain A of PDB ID 5jxe (human



**FIGURE 8** Superposition of PD-1 proteins in the PDB IDs 5jxe<sup>10</sup> and 4zqk<sup>105</sup> created using Mol\*. Ribbon representations of PD-1 (bright green outline), PD-L1 in PDB ID 4zqk (dark gray), and pembrolizumab Fab Heavy (gold) and Light (magenta) Chains. PD-L1 and the Fab Heavy Chain cannot bind PD-1 simultaneously because the two proteins would clash sterically (see labeled red arrow)

PD-1 bound to pembrolizumab) with Chain B of PDB ID 4zqk (human PD-1 bound to human PD-L1), which gave RMSD ~1.0 Å for common Cα atom pairs. The image of the superposed structures reveals the mechanism of action of pembrolizumab at the atomic level. Binding of the Fab would sterically interfere with PD-L1 binding to PD-1, thereby preventing down regulation of the T-cell when the T-cell receptor detects a neoepitope displayed on the surface of the malignant cell by major histocompatibility complex classes I and II.

A detailed guide to the superposition process can be accessed by using the path documentation navigation menu → Mol\* → FAQs/scenarios leading to the FAQ/scenario, entitled “How do I compare/superpose multiple structures?” an accompanying FAQ/scenario, entitled “How can I save a Mol\* session or state and return to it at a later time?” explains how to generate and download files for saving a Mol\* session or Mol\* state for later use.

## 2.7 | RCSB PDB Service 4: Training, outreach, and education

RCSB PDB has a long-standing commitment to training and community outreach and education. The dedicated web portal PDB-101 (denoting an introductory course; [PDB101.RCSB.org](http://PDB101.RCSB.org)) was launched in 2011 to support PDB archive exploration by teachers, students, and the general public. A detailed description of PDB-101 activities was provided most recently in the 2022 Protein Science Tools Issue.<sup>30</sup> Service 4 resources help train the next generation of PDB users and promote the importance of structural biology and protein science to nonexperts. Regularly published features include the highly popular Molecule of the Month (MotM) series,<sup>120</sup> 3D biostructure-related activities, molecular animations and videos, and educational curricula. Materials are organized into various categories (Health and Disease, Molecules of Life, Biotech and Nanotech, and Structures and Structure Determination) and searchable by keyword (e.g., cancer, checkpoint therapy, and antibody).

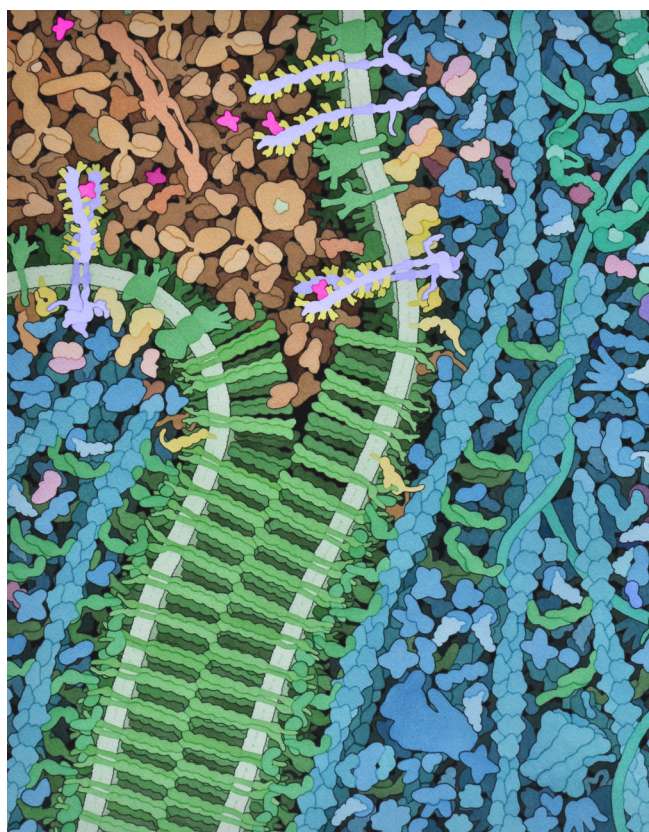
“Cancer Biology and Therapeutics” is the biennial health focus of PDB-101 for 2022 and 2023. Recently completed and ongoing activities related to cancer are as follows:

1. Publication of six new MotM articles that were coauthored by under-represented minority undergraduate students participating in the Rutgers University Institute for Quantitative Biomedicine 2022 Winter Boot Camp ([iqb.rutgers.edu](http://iqb.rutgers.edu)), entitled “Science Communication in Biology and Medicine.” Article topics include Vascular Endothelial Growth Factor (VegF)



and Angiogenesis (Figure 9; March 2022)<sup>121</sup>; Nicotine, Cancer, and Addiction (May 2022)<sup>122</sup>; Pyruvate Kinase (June 2022)<sup>123</sup>; HER2/neu and Trastuzumab (April 2022)<sup>124</sup>; Non-Homologous End Joining Supercomplexes (July 2022)<sup>125</sup>; and Secretory Antibodies (August 2022).<sup>126</sup>

2. RCSB PDB hosted the 2022 High School Video Challenge ([pdb101.rcsb.org/events/video-challenge/the-challenge](https://pdb101.rcsb.org/events/video-challenge/the-challenge)). Participating students were asked to explore two cancer-related molecular signaling systems: the p53/p21 pathway and the Epidermal Growth Factor Receptor (EGFR)/Ras pathway. In all, 38 teams



**FIGURE 9** Artistic conception of VegF signaling from the Molecule of the Month article *Vascular Endothelial Growth Factor (VegF) and Angiogenesis*.<sup>121</sup> Two neighboring cells are shown with cell membrane in green and cytoplasm in blue, and blood plasma at the top left in tan. VegF (magenta) arrives through the blood to the potential site of a new blood vessel and brings together two copies of VegF receptor (top center, lavender with glycosylation in yellow) to form an active dimer. Activated VegF receptor then initiates a signal cascade of kinases (yellow molecules attached to the membrane and pink molecules in the cytoplasm), leading to phosphorylation of many proteins, including cadherin (green). The phosphorylated cadherins separate making room for new blood vessels. Cytoplasmic proteins (blue) in the two cells include long actin filaments, L-shaped tRNA, a small ribosomal subunit, and many metabolic enzymes. The full image is available at PDB-101 (doi: [10.2210/rcsb\\_pdb/goodsell-gallery-041](https://doi.org/10.2210/rcsb_pdb/goodsell-gallery-041))

from around the US created short videos, each telling a coherent story explaining underlying scientific concepts and touching on public health aspects of cancer, such as screening, prevention, and awareness. External expert reviewers selected the winning videos based on scientific content and communication effectiveness, while the public was invited to vote on a “viewer’s choice” video.

Previously published PDB-101 activities related to cancer included MotM articles on p53 Tumor Suppressor (July 2002)<sup>127</sup>; Estrogen Receptor (September 2003)<sup>128</sup>; Epidermal Growth Factor (June 2010)<sup>129</sup>; RAS Protein (April 2012)<sup>130</sup>; RAF Protein Kinases (March 2016)<sup>131</sup>; Programmed Cell Death Protein 1 (December 2016)<sup>132</sup>; Chimeric Antigen Receptors (October 2017)<sup>133</sup>; MDM2 (June 2019)<sup>134</sup>; Cyclin and Cyclin Dependent Kinase (August 2019)<sup>135</sup>; and Cisplatin and DNA (March 2021).<sup>136</sup> Given the public health importance of human papilloma viruses (HPV, a cause of genital warts and cervical, genital, anal, and oropharyngeal cancers), PDB-101 developed related content in the form of (i) a molecular origami paper-folding model of the reconstituted HPV capsid that has been used successfully for vaccine design, (ii) a Human Papillomavirus and Vaccines MotM article (May 2018)<sup>137</sup>; and (iii) a *Nature Oncogene* review article focused on the biology of HPV viewed through the lens of the PDB.<sup>138</sup>

Going beyond our biennial theme of Cancer Biology and Therapeutics, we use our [PDB101.RCSB.org](https://PDB101.RCSB.org) web portal and *Service 4* social media channels (Twitter: @build-models; Facebook: RCSBPDB) to empower educators, students, and the general public to explore and understand how 3D biostructures are being used in research to address global health challenges of immediate concern. Compelling SARS-CoV-2 images created by *Service 4* team member Professor David S. Goodsell ([pdb101.rcsb.org/sci-art/goodsell-gallery](https://pdb101.rcsb.org/sci-art/goodsell-gallery)) and introductory materials related to coronaviruses were provided in the early days of the COVID-19 pandemic.<sup>139</sup> Up-to-date information on PDB structures and resources related to COVID-19 is available at [RCSB.org/covid19](https://RCSB.org/covid19), and new MotM articles, illustrations, animations, and flyers are created and presented on the [PDB101.RCSB.org](https://PDB101.RCSB.org) web portal as important new findings emerge from COVID-19 researchers. A recently published RCSB PDB article described PDB-related resources related to structure-facilitated messenger RNA vaccine design.<sup>140</sup> Similarly, new PDB-101 resources are being used to explain what we know about the structural biology and vaccinology of poxviruses in response to concerning reports of outbreaks of monkeypox in non-endemic areas (accessible by searching [PDB101.RCSB.org](https://PDB101.RCSB.org) with the keyword “poxviruses”). PDB-

101 features on Ebola Virus published in 2014, including a MotM article,<sup>141</sup> an award-winning Ebola Virus Illustration, and an Ebola Virus Protein Video, are likely to attract renewed interest in the wake of recent outbreaks in Uganda.

### 3 | DISCUSSION

RCSB PDB is currently in its 24th year of continuous operations, making it nearly half as mature (as opposed to old) as the PDB itself. The 50th anniversary of the PDB was celebrated in 2021 with numerous peer-reviewed publications and journal special issues,<sup>21,25,29,142–145</sup> and virtual scientific meetings hosted by the wwPDB Foundation ([foundation.wwpdb.org](http://foundation.wwpdb.org)) and wwPDB members ([wwpdb.org/pdb50](http://wwpdb.org/pdb50)). Throughout its golden anniversary celebrations, there was universal acclaim for the pioneering efforts of the PDB as a vanguard of the open-access data movement and global sharing of research findings.

Today, there is broad appreciation that advances in basic and applied research depend critically on unfettered access to the research findings of the broader scientific community. Given that the vast majority of 3D biostructures currently in the public domain were generated with governmental or private philanthropic support, it is only right that this information be made freely available with no limitations on usage. Promulgation of the FAIR and FACT principles by government funders in many nations and the efforts of non-governmental organizations such as the CoreTrustSeal ([coretrustseal.org](http://coretrustseal.org)) and the Global Biodata Coalition ([globalbiodata.org](http://globalbiodata.org)) are playing critical roles in raising awareness of the value of open data sharing.<sup>146</sup> Equally important, going forward, will be development of sustainable funding mechanisms to support global core biodata resources such as the PDB.<sup>147,148</sup>

Looking ahead, perspectives on the intertwined futures of structural biologists, the PDB archive, the RCSB PDB, and the Worldwide PDB partnership are perforce informed by the COVID-19 experience. Less than 1 month after release of the viral genome sequence in early 2020, the first experimentally determined structure of a SARS-CoV-2 protein—that of the main protease<sup>149</sup>—was deposited to the PDB and made publicly available. Thereafter, structural biologists and wwPDB data centers worked in concert to make 3D structures of viral proteins freely available to basic and applied researchers, vaccine designers, and drug hunters. At the time of writing, there were ~ 2700 structures of SARS-CoV-2 proteins and hundreds of SARS-CoV protein structures archived in the PDB. Together, they helped fuel an immense effort in basic and applied virology and immunology research; supported design of two widely available, life-saving

messenger RNA vaccines; and facilitated structure-guided discovery and development of a highly effective main protease-targeting drug (i.e., Pfizer's nirmatrelvir,<sup>150</sup> co-administered with ritonavir as Paxlovid) and multiple monoclonal antibodies approved for passive immunization (e.g., Evusheld, a fixed-dosed combination of tixagevimab and cilgavimab).

The words “what's past is prologue” spoken by Antonio in Act 2 Scene 1 of Shakespeare's *The Tempest* ring as true today as they did when the play was first performed on November first 1611. Evolution of structural biology as a scientific discipline and growth PDB as the first open access digital data resource in biology served us remarkably well as preparation for the pandemic. With the advent of artificial intelligence/machine learning-based de novo protein structure prediction, there is every reason to believe that structural biologists will be even more prolific, particularly when it really matters while SARS-CoV-2 and its variants of concern continue to threaten global public health. Their generous contributions of new information will ensure that the PDB continues to grow in importance as the single global open access data resource for experimentally determined 3D biostructures. As the RCSB PDB enters its 25th year of continuous operations, it remains committed to meeting ongoing challenges and delivering the ever-increasing corpus of structure information in myriad ways that will meet the needs of its diverse community of users around the world.

### AUTHOR CONTRIBUTIONS

**Stephen K. Burley:** Funding acquisition (lead); project administration (lead); supervision (lead); writing – original draft (lead). **Charmi Bhikadiya:** Software (supporting). **Chunxiao Bi:** Software (supporting). **Sebastian Bittrich:** Software (supporting). **Henry Chao:** Software (supporting). **Li Chen:** Software (supporting). **Paul A. Craig:** Writing – review and editing (supporting). **Gregg V. Crichlow:** Data curation (supporting); validation (supporting). **Kenneth Dalenberg:** Software (supporting). **Jose M. Duarte:** Project administration (supporting); software (lead); supervision (supporting). **Shuchismita Dutta:** Visualization (equal); writing – review and editing (supporting). **Maryam Fayazi:** Software (supporting). **Zukang Feng:** Project administration (lead); software (supporting); supervision (supporting); validation (supporting). **Justin W. Flatt:** Data curation (supporting); validation (supporting). **Sai J. Ganesan:** Validation (supporting). **Sutapa Ghosh:** Data curation (supporting); validation (supporting). **David S. Goodsell:** Visualization (equal); writing – review and editing (supporting). **Rachel Kramer Green:** Project administration (supporting). **Vladimir Guranovic:** Software



(supporting). **Jeremy Henry:** Software (supporting). **Brian P. Hudson:** Data curation (supporting); validation (supporting). **Igor Khokhriakov:** Software (supporting). **Catherine L. Lawson:** Data curation (supporting); validation (supporting). **YuHe Liang:** Data curation (lead); validation (lead). **Robert Lowe:** Software (lead); supervision (supporting). **Ezra Peisach:** Data curation (supporting); software (supporting); validation (supporting). **Irina Persikova:** Data curation (lead); validation (lead). **Den- nis W. Piehl:** Software (supporting). **Yana Rose:** Project administration (supporting); software (lead). **Andrej Sali:** Project administration (lead). **Joan Segura:** Software (supporting). **Monica Sekharan:** Data curation (supporting); validation (supporting). **Chenghua Shao:** Data curation (supporting); software (supporting); validation (supporting). **Brinda Vallat:** Software (supporting). **Maria Voigt:** Software (supporting); visualization (equal). **Benjamin Webb:** Software (supporting). **John D. Westbrook:** Software (lead); supervision (supporting). **Shamara Whetstone:** Project administration (supporting). **Jasmine Y. Young:** Data curation (lead); project administration (supporting); supervision (supporting); validation (lead). **Arthur Zalevsky:** Software (supporting). **Christine Zardecki:** Project administration (lead); supervision (supporting); visualization (equal); writing – review and editing (supporting).

## ACKNOWLEDGEMENTS

The authors thank the tens of thousands of structural biologists who deposited structures to the PDB since 1971 and the many millions of researchers, educators, and students around the world who utilize PDB data. We thank the members of the RCSB PDB and wwPDB Advisory Committees for their valued advice. We gratefully acknowledge contributions to the growth and maintenance of the PDB archive made by past members of RCSB PDB and our Worldwide Protein Data Bank partners (PDBe, PDBj, EMDB, and BMRB).

## FUNDING INFORMATION

RCSB PDB is jointly funded by the National Science Foundation (DBI-1832184, PI: S.K. Burley), the US Department of Energy (DE-SC0019749, PI: S.K. Burley), and the National Cancer Institute, the National Institute of Allergy and Infectious Diseases, and the National Institute of General Medical Sciences of the National Institutes of Health (R01GM133198, PI: S.K. Burley). Additional funding from NSF is supporting development of a next generation PDB archive (DBI-2019297, PI: S.K. Burley) and new Mol\* features (DBI-2129634, PI: S.K. Burley). Other funding awards to RCSB PDB by the NSF and to PDBe by the UK Biotechnology and Biological Research Council are jointly supporting development

of a Next Generation PDB archive (DBI-2019297, PI: S.K. Burley; BB/V004247/1, PI: Sameer Velankar) and new Mol\* features (DBI-2129634, PI: S.K. Burley; BB/W017970/1, PI: Sameer Velankar). PDB-Dev is supported by NSF (DBI-1756248 and DBI-2112966, PI: B. Vallat; DBI-1756250 and DBI-2112967, PI: A. Sali). Sali acknowledges additional support from NIH-NIGMS (R01GM083960, PI: A. Sali; P41GM109824, PI: M.P. Rout).

## CONFLICT OF INTEREST

The authors declare no conflict of interest with this publication. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.


## DATA AVAILABILITY STATEMENT

wwPDB policy states that data files contained in the PDB archive are available under the CC0 1.0 Universal (CC0 1.0) Public Domain Dedication.

## ORCID


Stephen K. Burley  <https://orcid.org/0000-0002-2487-9713>

Jose M. Duarte  <https://orcid.org/0000-0002-9544-5621>

David S. Goodsell  <https://orcid.org/0000-0002-5932-2130>

Benjamin Webb  <https://orcid.org/0000-0003-3360-4540>

John D. Westbrook  <https://orcid.org/0000-0002-6686-5475>

Christine Zardecki  <https://orcid.org/0000-0002-4149-1745>

## REFERENCES

1. Protein Data Bank. Crystallography: Protein data bank. *Nature* (London). New Biol. 1971;233(42):223.
2. Berman HM, Henrick K, Nakamura H. Announcing the worldwide protein data bank. *Nat Struct Biol*. 2003; 10(12):980.
3. wwPDB consortium. Protein data bank: The single global archive for 3d macromolecular structure data. *Nucleic Acids Res*. 2019;47(D1):D520–D528.
4. Berman HM, Westbrook J, Feng Z, et al. The protein data bank. *Nucleic Acids Res*. 2000;28(1):235–242.
5. Burley SK, Bhikadiya C, Bi C, et al. 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 Acid Res*. 2021;49: D437–D451.
6. Armstrong DR, Berrisford JM, Conroy MJ, et al. PdBe: Improved findability of macromolecular structure data in the pdb. *Nucleic Acids Res*. 2020;48(D1):D335–D343.
7. Bekker GJ, Yokochi M, Suzuki H, et al. Protein data bank Japan: Celebrating our 20th anniversary during a global

- pandemic as the Asian hub of three dimensional macromolecular structural data. *Protein Sci.* 2022;31(1):173–186.
8. Lawson CL, Patwardhan A, Baker ML, et al. Emdatabank unified data resource for 3dem. *Nucleic Acids Res.* 2016;44(D1):D396–D403.
  9. Ulrich EL, Akutsu H, Doreleijers JF, et al. Biomagresbank. *Nucleic Acids Res.* 2008;36:D402–D408.
  10. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The fair guiding principles for scientific data management and stewardship. *Sci Data.* 2016;3(160018):1–9.
  11. van der Aalst WMP, Bichler M, Heinzl A. Responsible data science. *Bus Inf Syst Eng.* 2017;59(5):311–313.
  12. Burley SK, Berman HM, Duarte JM, et al. Protein data bank: A comprehensive review of 3D structure holdings and worldwide utilization by researchers, educators, and students. *Biomolecules.* 2022;12:1425.
  13. Kendrew JC, Dickerson RE, Strandberg BE, et al. Structure of myoglobin: A three-dimensional fourier synthesis at 2 Å resolution. *Nature.* 1960;185(4711):422–427.
  14. Yip KM, Fischer N, Paknia E, Chari A, Stark H. Atomic-resolution protein structure determination by cryo-em. *Nature.* 2020;587(7832):157–161.
  15. Burley SK, Berman HM, Christie C, et al. RCSB protein data bank: Sustaining a living digital data resource that enables breakthroughs in scientific research and biomedical education. *Protein Sci.* 2018;27(1):316–330.
  16. Markosian C, Di Costanzo L, Sekharan M, Shao C, Burley SK, Zardecki C. Analysis of impact metrics for the protein data bank. *Sci Data.* 2018;5:180212.
  17. Hill R, Stein C. 2019. Scooped! Estimating rewards for priority in science. Working Paper. Massachusetts Institute of Technology.
  18. Feng Z, Verdigué N, Di Costanzo L, et al. Impact of the protein data bank across scientific disciplines. *Data Sci J.* 2020;19:1–14.
  19. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with alphafold. *Nature.* 2021;596(7873):583–589.
  20. Baek M, DiMaio F, Anishchenko I, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science.* 2021;373(6557):871–876.
  21. Burley SK, Berman HM. Open-access data: A cornerstone for artificial intelligence approaches to protein structure prediction. *Structure.* 2021;29(6):515–520.
  22. Westbrook JD, Burley SK. How structural biologists and the protein data bank contributed to recent FDA new drug approvals. *Structure.* 2019;27:211–217.
  23. Westbrook JD, Soskind R, Hudson BP, Burley SK. Impact of protein data bank on anti-neoplastic approvals. *Drug Discov Today.* 2020;25:837–850.
  24. Goodsell DS, Zardecki C, Di Costanzo L, et al. RCSB protein data bank: Enabling biomedical research and drug discovery. *Protein Sci.* 2020;29:52–65.
  25. Burley SK. Impact of structural biologists and the protein data bank on small-molecule drug discovery and development. *J Biol Chem.* 2021;296:100559.
  26. Young JY, Westbrook JD, Feng Z, et al. Onedep: Unified wwPDB system for deposition, biocuration, and validation of macromolecular structures in the PDB archive. *Structure.* 2017;25(3):536–545.
  27. Burley SK, Berman HM, Kleywegt GJ, Markley JL, Nakamura H, Velankar S. Protein data bank (pdb): The single global macromolecular structure archive. In: Wlodawer A, Jaskolski M, editors. *Methods in molecular biology: Protein crystallography methods and protocols.* New York: Springer, 2017; p. 627–641.
  28. Rose Y, Duarte JM, Lowe R, et al. RCSB protein data bank: Architectural advances towards integrated searching and efficient access to macromolecular structure data from the pdb archive. *J Mol Biol.* 2021;443:166704.
  29. Burley SK, Bhikadiya C, Bi C, et al. RCSB protein data bank: Celebrating 50 years of the pdb with new tools for understanding and visualizing biological macromolecules in 3d. *Protein Sci.* 2022;31(1):187–208.
  30. Zardecki C, Dutta S, Goodsell DS, Lowe R, Voigt M, Burley SK. Pdb-101: Educational resources supporting molecular explorations through biology and medicine. *Protein Sci.* 2022;31(1):129–140.
  31. Gore S, Sanz Garcia E, Hendrickx PMS, et al. Validation of structures in the protein data bank. *Structure.* 2017;25(12):1916–1927.
  32. Feng Z, Westbrook JD, Sala R, et al. Enhanced validation of small-molecule ligands and carbohydrates in the protein data bank. *Structure.* 2021;29:393–400.e391.
  33. Shao C, Liu Z, Yang H, Wang S, Burley SK. Outlier analyses of the protein data bank archive using a probability-density-ranking approach. *Sci Data.* 2018;5:180293.
  34. Young JY, Westbrook JD, Feng Z, et al. Worldwide protein data bank biocuration supporting open access to high-quality 3d structural biology data. *Database.* 2018;2018:bay002.
  35. Shao C, Feng Z, Westbrook JD, et al. Modernized uniform representation of carbohydrate molecules in the protein data bank. *Glycobiology.* 2021;31:1204–1218.
  36. Henrick K, Feng Z, Bluhm WF, et al. Remediation of the protein data bank archive. *Nucleic Acids Res.* 2008;36:D426–D433.
  37. Lawson CL, Dutta S, Westbrook JD, Henrick K, Berman HM. Representation of viruses in the remediated pdb archive. *Acta Crystallogr D.* 2008;D64(Pt 8):874–882.
  38. Burley SK, Berman HM, Bhikadiya C, et al. RCSB protein data bank: Biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.* 2019;47(D1):D464–D474.
  39. Pearce NM, Krojer T, Bradley AR, et al. A multi-crystal method for extracting obscured crystallographic states from conventionally uninterpretable electron density. *Nat Commun.* 2017;8:15123.
  40. Yang H, Guranovic V, Dutta S, Feng Z, Berman HM, Westbrook JD. Automated and accurate deposition of structures solved by x-ray diffraction to the protein data bank. *Acta Crystallogr D.* 2004;60(Pt 10):1833–1839.
  41. Yang H, Peisach E, Westbrook JD, Young J, Berman HM, Burley SK. DCC: A swiss army knife for structure factor analysis and validation. *J Appl Cryst.* 2016;49:1081–1084.
  42. Kramer RZ, Venugopal MG, Bella J, Mayville P, Brodsky B, Berman HM. Staggered molecular packing in crystals of a collagen-like peptide with a single charged pair. *J Mol Biol.* 2000;301(5):1191–1205.



43. Feng Z, Chen L, Maddula H, et al. Ligand depot: A data warehouse for ligands bound to macromolecules. *Bioinformatics*. 2004;20(13):2153–2155.
44. Shao C, Westbrook JD, Lu C, et al. Simplified quality assessment for small-molecule ligands in the pdb archive. *Structure*. 2022;30:252–262.e4.
45. Shao C, Bittrich S, Wang W, Burley SK. Assessing pdb macromolecular crystal structure confidence at the individual amino acid residue level. *Structure*. 2022;30:1385–1394.e3.
46. UniProt Consortium. Uniprot: The universal protein knowledgebase in 2021. *Nucleic Acids Res*. 2021;49(D1):D480–D489.
47. Sayers EW, Bolton EE, Brister JR, et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res*. 2022;50(D1):D20–D26.
48. Westbrook JD, Shao C, Feng Z, Zhuravleva M, Velankar S, Young J. The chemical component dictionary: Complete descriptions of constituent molecules in experimentally determined 3D macromolecules in the protein data bank. *Bioinformatics*. 2015;31(8):1274–1278.
49. Berman HM, Trewhealla J, Vallat B, Westbrook JD. Archiving of integrative structural models. *Adv Exp Med Biol*. 2018;1105:261–272.
50. Vallat B, Webb B, Westbrook JD, Sali A, Berman HM. Development of a prototype system for archiving integrative/hybrid structure models of biological macromolecules. *Structure*. 2018;26:894–904.
51. Berman HM, Adams PD, Bonvin AA, et al. Federating structural models and data: Outcomes from a workshop on archiving integrative structures. *Structure*. 2019;27(12):1745–1759.
52. Vallat B, Webb B, Westbrook J, Sali A, Berman HM. Archiving and disseminating integrative structure models. *J Biomol NMR*. 2019;73:385–398.
53. Vallat B, Webb B, Fayazi M, et al. New system for archiving integrative structures. *Acta Crystallogr D Struct Biol*. 2021;77(Pt 12):1486–1496.
54. Burley SK, Kurisu G, Markley JL, et al. Pdb-dev: A prototype system for depositing integrative/hybrid structural models. *Structure*. 2017;25(9):1317–1318.
55. Kuller A, Fleri W, Bluhm WF, Smith JL, Westbrook J, Bourne PE. A biologist's guide to synchrotron facilities: The biosync web resource. *TIBS*. 2002;27:213–215.
56. Westbrook J, Bourne PE. Star/mmCIF: An extensive ontology for macromolecular structure and beyond. *Bioinformatics*. 2000;16:159–168.
57. Fitzgerald PMD, Westbrook JD, Bourne PE, McMahon B, Watenpaugh KD, Berman HM. 4.5 macromolecular dictionary (mmCIF). In: Hall SR, McMahon B, editors. *International tables for crystallography g definition and exchange of crystallographic data*. Dordrecht, The Netherlands: Springer; 2005; p. 295–443.
58. Westbrook JD, Young JY, Shao C, et al. Pdbx/mmCIF ecosystem: Foundational semantic tools for structural biology. *J Mol Biol*. 2022;434:167599.
59. Hall SR, Allen FH, Brown ID. The crystallographic information file (CIF): A new standard archive file for crystallography. *Acta Crystallogr A Found Crystallogr*. 1991;47(6):655–685.
60. Haas J, Barbato A, Behringer D, et al. Continuous automated model evaluation (cameo) complementing the critical assessment of structure prediction in CASP12. *Proteins*. 2018;86-(Suppl 1):387–398.
61. Wagner JR, Churas CP, Liu S, et al. Continuous evaluation of ligand protein predictions: A weekly community challenge for drug docking. *Structure*. 2019;27(8):1326–1335.
62. Varadi M, Anyango S, Deshpande M, et al. AlphaFold protein structure database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res*. 2022;50(D1):D439–D444.
63. Ahmed A, Smith RD, Clark JJ, Dunbar JB Jr, Carlson HA. Recent improvements to binding MOAD: A resource for protein-ligand binding affinities and structures. *Nucleic Acids Res*. 2015;43:D465–D469.
64. Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong J. Binding DB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res*. 2016;44(D1):D1045–D1053.
65. Romero PR, Kobayashi N, Wedell JR, et al. Biomagresbank (BMRB) as a resource for structural biology. *Methods Mol Biol*. 2020;2112:187–218.
66. Ribeiro AJM, Holliday GL, Furnham N, Tyzack JD, Ferris K, Thornton JM. Mechanism and catalytic site atlas (M-CSA): A database of enzyme reaction mechanisms and active sites. *Nucleic Acids Res*. 2018;46(D1):D618–D623.
67. Sillitoe I, Bordin N, Dawson N, et al. Cath: Increased structural coverage of functional space. *Nucleic Acids Res*. 2021;49(D1):D266–D273.
68. Groom CR, Bruno IJ, Lightfoot MP, Ward SC. The Cambridge structural database. *Acta Crystallogr B Struct Sci Cryst Eng Mater*. 2016;72(Pt 2):171–179.
69. Hastings J, Owen G, Dekker A, et al. ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res*. 2016;44(D1):D1214–D1219.
70. Gaulton A, Hersey A, Nowotka M, et al. The ChEMBL database in 2017. *Nucleic Acids Res*. 2017;45(D1):D945–D954.
71. Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res*. 2018;46(D1):D1074–D1082.
72. Cheng H, Liao Y, Schaeffer RD, Grishin NV. Manual classification strategies in the ECD database. *Proteins*. 2015;83(7):1238–1251.
73. McDonald AG, Boyce S, Tipton KF. Explorenz: The primary source of the IUBMB enzyme list. *Nucleic Acids Res*. 2009;37:D593–D597.
74. Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: The reference human genome annotation for the ENCODE project. *Genome Res*. 2012;22(9):1760–1774.
75. Gene Ontology Consortium. The gene ontology resource: Enriching a gold mine. *Nucleic Acids Res*. 2021;49(D1):D325–D334.
76. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*. 2020;369(6509):1318–1330.
77. Yamada I, Shiota M, Shinmachi D, et al. The Glycosmos portal: A unified and comprehensive web resource for the glycosciences. *Nat Methods*. 2020;17(7):649–650.
78. York WS, Mazumder R, Ranzinger R, et al. GlyGen: Computational and informatics resources for glycoscience. *Glycobiology*. 2020;30(2):72–73.
79. Tiemeyer M, Aoki K, Paulson J, et al. GlyToucan: An accessible glycan structure repository. *Glycobiology*. 2017;27(10):915–919.

80. Lefranc MP, Giudicelli V, Duroux P, et al. Imgt(r), the international immunogenetics information system(r) 25 years on. *Nucleic Acids Res.* 2015;43:D413–D422.
81. Vita R, Overton JA, Greenbaum JA, et al. The immune epitope database (iedb) 3.0. *Nucleic Acids Res.* 2015;43:D405–D412.
82. Newport TD, Sansom MSP, Stansfeld PJ. The memprotmd database: A resource for membrane-embedded protein structures and their lipid interactions. *Nucleic Acids Res.* 2019;47(D1):D390–D397.
83. Berman HM, Olson WK, Beveridge DL, et al. The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys J.* 1992;63(3):751–759.
84. Lomize MA, Lomize AL, Pogozheva ID, Mosberg HI. OPM: Orientations of proteins in membranes database. *Bioinformatics.* 2006;22(5):623–625.
85. Su M, Yang Q, Du Y, et al. Comparative assessment of scoring functions: The casf-2016 update. *J Chem Inf Model.* 2019;59(2):895–913.
86. Hrabe T, Li Z, Sedova M, Rotkiewicz P, Jaroszewski L, Godzik A. Pdbflex: Exploring flexibility in protein structures. *Nucleic Acids Res.* 2016;44(D1):D423–D428.
87. Tusnady GE, Dosztanyi Z, Simon I. Transmembrane proteins in the protein data bank: Identification and classification. *Bioinformatics.* 2004;20(17):2964–2972.
88. Finn RD, Coghill P, Eberhardt RY, et al. The pfam protein families database: Towards a more sustainable future. *Nucleic Acids Res.* 2016;44(D1):D279–D285.
89. Nguyen DT, Mathias S, Bologa C, et al. Pharos: Collating protein information to shed light on the druggable genome. *Nucleic Acids Res.* 2017;45(D1):D995–D1002.
90. Kim S, Chen J, Cheng T, et al. Pubchem in 2021: New data content and improved web interfaces. *Nucleic Acids Res.* 2021;49(D1):D1388–D1395.
91. Nederveen AJ, Doreleijers JF, Vranken W, et al. Recoord: A recalculated coordinate database of 500+ proteins from the pdb using restraints from the biomagresbank. *Proteins.* 2005;59(4):662–672.
92. Garavelli JS. The resid database of protein modifications as a resource and annotation tool. *Proteomics.* 2004;4(6):1527–1533.
93. Dunbar J, Krawczyk K, Leem J, et al. Sabdab: The structural antibody database. *Nucleic Acids Res.* 2014;42:D1140–D1146.
94. Morin A, Eisenbraun B, Key J, et al. Collaboration gets the most out of software. *Elife.* 2013;2:e01456.
95. Andreeva A, Kulesha E, Gough J, Murzin AG. The scop database in 2020: Expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Res.* 2020;48(D1):D376–D382.
96. Chandonia JM, Fox NK, Brenner SE. Scope: Classification of large macromolecular structures in the structural classification of proteins-extended database. *Nucleic Acids Res.* 2019;47(D1):D475–D481.
97. Dana JM, Gutmanas A, Tyagi N, et al. Sifts: Updated structure integration with function, taxonomy and sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic Acids Res.* 2019;47(D1):D482–D489.
98. Raybould MIJ, Marks C, Lewis AP, et al. Thera-sabdab: The therapeutic structural antibody database. *Nucleic Acids Res.* 2020;48(D1):D383–D388.
99. Sharma P, Allison JP. Dissecting the mechanisms of immune checkpoint therapy. *Nat Rev Immunol.* 2020;20(2):75–76.
100. Klempner SJ, Fabrizio D, Bane S, et al. Tumor mutational burden as a predictive biomarker for response to immune checkpoint inhibitors: A review of current evidence. *Oncologist.* 2020;25(1):e147–e159.
101. Na Z, Yeo SP, Bharath SR, et al. Structural basis for blocking pd-1-mediated immune suppression by therapeutic antibody pembrolizumab. *Cell Res.* 2017;27(1):147–150.
102. Segura J, Rose Y, Bittrich S, Burley SK, Duarte JM. RCSB protein data bank 1d3d module: Displaying positional features on macromolecular assemblies. *Bioinformatics.* 2022;38:3304–3305.
103. Segura J, Rose Y, Westbrook J, Burley SK, Duarte JM. RCSB protein data bank 1d tools and services. *Bioinformatics.* 2020;36(22–23):5526–5527.
104. Sehna D, Bittrich S, Deshpande M, et al. Mol\* viewer: Modern web app for 3d visualization and analysis of large biomolecular structures. *Nucleic Acids Res.* 2021;49:W431–W437.
105. Zak KM, Kite R, Przetoicka S, et al. Structure of the complex of human programmed death 1, pd-1, and its ligand pd-l1. *Structure.* 2015;23(12):2341–2348.
106. Andreeva A, Howorth D, Chothia C, Kulesha E, Murzin AG. Scop2 prototype: A new approach to protein structure mining. *Nucleic Acids Res.* 2014;42:D310–D314.
107. Guzik K, Zak KM, Grudnik P, et al. Small-molecule inhibitors of the programmed cell death-1/programmed death-ligand 1 (pd-1/pd-l1) interaction via transiently induced protein states and dimerization of pd-l1. *J Med Chem.* 2017;60(13):5857–5867.
108. Skalniak L, Zak KM, Guzik K, et al. Small-molecule inhibitors of pd-1/pd-l1 immune checkpoint alleviate the pd-l1-induced exhaustion of t-cells. *Oncotarget.* 2017;8(42):72167–72181.
109. Perry E, Mills JJ, Zhao B, et al. Fragment-based screening of programmed death ligand 1 (pd-l1). *Bioorg Med Chem Lett.* 2019;29(6):786–790.
110. Butera R, Wazynska M, Magiera-Mularz K, et al. Design, synthesis, and biological evaluation of imidazopyridines as pd-1/pd-l1 antagonists. *ACS Med Chem Lett.* 2021;12(5):768–773.
111. Park JJ, Thi EP, Carpio VH, et al. Checkpoint inhibition through small molecule-induced internalization of programmed death-ligand 1. *Nat Commun.* 2021;12(1):1222.
112. Wang T, Cai S, Cheng Y, et al. Discovery of small-molecule inhibitors of the pd-1/pd-l1 axis that promote pd-l1 internalization and degradation. *J Med Chem.* 2022;65(5):3879–3893.
113. Sun C, Cheng Y, Liu X, et al. Novel phthalimides regulating pd-1/pd-l1 interaction as potential immunotherapy agents. *Acta Pharmaceut Sin B.* 2022.
114. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension of the optimum path. *Protein Eng.* 1998;11:739–747.
115. Bliven SE, Bourne PE, Prlic A. Detection of circular permutations within protein structures using ce-cp. *Bioinformatics.* 2015;31(8):1316–1318.
116. Ye Y, Godzik A. Fatcat: A web server for flexible structure comparison and structure similarity searching. *Nucleic Acids Res.* 2004;32:W582–W585.



117. Zhang Y, Skolnick J. Tm-align: A protein structure alignment algorithm based on the tm-score. *Nucleic Acids Res.* 2005; 33(7):2302–2309.
118. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol.* 1981;147(1):195–197.
119. Lafita A, Bliven S, Prlic A, et al. Biojava 5: A community driven open-source bioinformatics library. *PLoS Comput Biol.* 2019;15(2):e1006791.
120. Goodsell DS, Zardecki C, Berman HM, Burley SK. Insights from 20 years of the molecule of the month. *Biochem Mol Biol Educ.* 2020;48:350–355.
121. Cartagena E, Gelashvili M, Keyes J, Rosenzweig E, Goodsell DS, Burley SK. 2022. Vascular endothelial growth factor (vegf) and angiogenesis. *RCSB PDB Molecule of the Month.* [https://doi.org/10.2210/rcsb\\_pdb/mom\\_2022\\_3](https://doi.org/10.2210/rcsb_pdb/mom_2022_3).
122. Choudhry K, Muse D, Maio DPD, Soto Acevedo AA, Goodsell DS, Dutta S. 2022. Nicotine, cancer, and addiction. *RCSB PDB Molecule of the Month.* [https://doi.org/10.2210/rcsb\\_pdb/mom\\_2022\\_5](https://doi.org/10.2210/rcsb_pdb/mom_2022_5).
123. Ahmed F, Ash J, Patel T, Sanders A, Goodsell DS, Dutta S. 2022. Pyruvate kinase m2. *RCSB PDB Molecule of the Month.* [https://doi.org/10.2210/rcsb\\_pdb/mom\\_2022\\_6](https://doi.org/10.2210/rcsb_pdb/mom_2022_6).
124. de Leon Cruz S, Herrod A, Park KH, Wu A, Goodsell DS, Burley SK. 2022. Her2/neu and trastuzumab *RCSB PDB Molecule of the Month.* [https://doi.org/10.2210/rcsb\\_pdb/mom\\_2022\\_4](https://doi.org/10.2210/rcsb_pdb/mom_2022_4).
125. Diaz-Figueroa G, Egozi M, Jannath S, Maddy J, Goodsell DS. 2022. Non-homologous end joining super complexes. *RCSB PDB Molecule of the Month.* [https://doi.org/10.2210/rcsb\\_pdb/mom\\_2022\\_7](https://doi.org/10.2210/rcsb_pdb/mom_2022_7).
126. Brooks L, Colón-Colón C, Patel A, Polat A, Goodsell DS. 2022. Secretory antibodies. *RCSB PDB Molecule of the Month.* [https://doi.org/10.2210/rcsb\\_pdb/mom\\_2022\\_8](https://doi.org/10.2210/rcsb_pdb/mom_2022_8).
127. Goodsell DS. 2002. P53 tumor suppressor. *RCSB PDB Molecule of the Month.* [https://doi.org/10.2210/rcsb\\_pdb/mom\\_2002\\_7](https://doi.org/10.2210/rcsb_pdb/mom_2002_7).
128. Goodsell DS. 2003. Estrogen receptor. *RCSB PDB Molecule of the Month.* [https://doi.org/10.2210/rcsb\\_pdb/mom\\_2003\\_9](https://doi.org/10.2210/rcsb_pdb/mom_2003_9).
129. Goodsell DS. 2010. Epidermal growth factor. *RCSB PDB Molecule of the Month.* [https://doi.org/10.2210/rcsb\\_pdb/mom\\_2010\\_6](https://doi.org/10.2210/rcsb_pdb/mom_2010_6).
130. Goodsell DS. 2012. Ras protein. *RCSB PDB Molecule of the Month.* [https://doi.org/10.2210/rcsb\\_pdb/mom\\_2012\\_4](https://doi.org/10.2210/rcsb_pdb/mom_2012_4).
131. Goodsell DS. 2016. Raf protein kinases. *RCSB PDB Molecule of the Month.* [https://doi.org/10.2210/rcsb\\_pdb/mom\\_2016\\_3](https://doi.org/10.2210/rcsb_pdb/mom_2016_3).
132. Goodsell DS. 2016. Pd-1 (programmed cell death protein 1). *RCSB PDB Molecule of the Month.* [https://doi.org/10.2210/rcsb\\_pdb/mom\\_2016\\_2](https://doi.org/10.2210/rcsb_pdb/mom_2016_2).
133. Goodsell DS. 2017. Chimeric antigen receptors. *RCSB PDB Molecule of the Month.* [https://doi.org/10.2210/rcsb\\_pdb/mom\\_2017\\_10](https://doi.org/10.2210/rcsb_pdb/mom_2017_10).
134. Goodsell DS. 2019. MDM 2 and cancer. *RCSB PDB Molecule of the Month.* [https://doi.org/10.2210/rcsb\\_pdb/mom\\_2019\\_26](https://doi.org/10.2210/rcsb_pdb/mom_2019_26).
135. Goodsell DS. 2019. Cyclin and cyclin-dependent kinase. *RCSB PDB Molecule of the Month.* [https://doi.org/10.2210/rcsb\\_pdb/mom\\_2019\\_8](https://doi.org/10.2210/rcsb_pdb/mom_2019_8).
136. Gao H, Shrem SG, Suryanarayanan S, Goodsell DS. 2021. Cis-platin and DNA. *RCSB PDB Molecule of the Month.* [https://doi.org/10.2210/rcsb\\_pdb/mom\\_2021\\_3](https://doi.org/10.2210/rcsb_pdb/mom_2021_3).
137. Goodsell DS. 2018. Human papillomavirus and vaccines. *RCSB PDB Molecule of the Month.* [https://doi.org/10.2210/rcsb\\_pdb/mom\\_2018\\_5](https://doi.org/10.2210/rcsb_pdb/mom_2018_5).
138. Goodsell DS, Burley SK. RCSB protein data bank tools for 3d structure-guided cancer research: Human papillomavirus (hpv) case study. *Oncogene.* 2020;39(43):6623–6632.
139. Goodsell DS, Voigt M, Zardecki C, Burley SK. Integrative illustration for coronavirus outreach. *PLoS Biol.* 2020;18(8):e3000815.
140. Goodsell DS, Burley SK. RCSB protein data Bank resources for structure-facilitated design of mRNA vaccines for existing and emerging viral pathogens. *Structure.* 2022;30:252–262. e254.
141. Goodsell DS. 2014. Ebola virus proteins. *RCSB PDB Molecule of the Month.* [http://doi.org/10.2210/rcsb\\_pdb/mom\\_2014\\_10](http://doi.org/10.2210/rcsb_pdb/mom_2014_10)
142. Gierasch L, Berman HM. 2021. How the Protein Data Bank changed biology: A thematic series. <https://www.jbc.org/thematic-how-the-protein-data-bank-changed-biology>
143. Gierasch L, Berman HM. 2021. Virtual issue: The PDB in JBC. *Journal of Biological Chemistry.* <https://www.jbc.org/the-pdb-in-jbc>
144. PDB 50th Anniversary: Celebrating the future of structural biology. 2021. *Nature Methods.* 18: <https://www.nature.com/nmeth/volumes/18/issues/15>.
145. PDB 50th anniversary: Celebrating the future of structural biology. 2021. *Nature Structural & Molecular Biology.* 28: <https://www.nature.com/nsmb/volumes/28/issues/25>.
146. Beierlein JM, McNamee LM, Walsh MJ, Kaitin KI, DiMasi JA, Ledley FD. Landscape of innovation for cardiovascular pharmaceuticals: From basic science to new molecular entities. *Clin Ther.* 2017;39(7):1409–1425.
147. Anderson WP. Data management: A global coalition to sustain core data. *Nature.* 2017;543(7644):179.
148. Anderson W, Apweiler R, Bateman A, et al. Towards coordinated international support of core data resources for the life sciences. *bioRxiv.* 2017:1–7. <https://doi.org/10.1101/110825>.
149. Jin Z, Du X, Xu Y, et al. Structure of m(pro) from sars-cov-2 and discovery of its inhibitors. *Nature.* 2020;582(7811):289–293.
150. Owen DR, Allerton CMN, Anderson AS, et al. An oral sars-cov-2 m(pro) inhibitor clinical candidate for the treatment of covid-19. *Science.* 2021;374(6575):1586–1593.

**How to cite this article:** Burley SK, Bhikadiya C, Bi C, Bittrich S, Chao H, Chen L, et al. RCSB Protein Data bank: Tools for visualizing and understanding biological macromolecules in 3D. *Protein Science.* 2022;31(12):e4482. <https://doi.org/10.1002/pro.4482>