AKADÉMIAI KIADÓ

Worldwide Protein Data Bank (wwPDB): A virtual treasure for research in biotechnology

PAYAM BEHZADI[1] and MÁRIÓ GAJDÁCS[2*]

[1] Department of Microbiology, College of Basic Sciences, Shahr-e-Qods Branch, Islamic Azad University, Tehran, 37541-374, Iran

[2] Department of Oral Biology and Experimental Dental Research, Faculty of Dentistry, University of Szeged, 6720, Szeged, Hungary

European Journal of Microbiology and Immunology

11 (2021) 4, 77–86

DOI:
10.1556/1886.2021.00020

Received: November 6, 2021 • Accepted: November 23, 2021
Published online: December 15, 2021

REVIEW PAPER

ABSTRACT

The Research Collaboratory for Structural Bioinformatics Protein Data Bank (RSCB PDB) provides a wide range of digital data regarding biology and biomedicine. This huge internet resource involves a wide range of important biological data, obtained from experiments around the globe by different scientists. The Worldwide Protein Data Bank (wwPDB) represents a brilliant collection of 3D structure data associated with important and vital biomolecules including nucleic acids (RNAs and DNAs) and proteins. Moreover, this database accumulates knowledge regarding function and evolution of biomacromolecules which supports different disciplines such as biotechnology. 3D structure, functional characteristics and phylogenetic properties of biomacromolecules give a deep understanding of the biomolecules' characteristics. An important advantage of the wwPDB database is the data updating time, which is done every week. This updating process helps users to have the newest data and information for their projects. The data and information in wwPDB can be a great support to have an accurate imagination and illustrations of the biomacromolecules in biotechnology. As demonstrated by the SARS-CoV-2 pandemic, rapidly reliable and accessible biological data for microbiology, immunology, vaccinology, and drug development are critical to address many healthcare-related challenges that are facing humanity. The aim of this paper is to introduce the readers to wwPDB, and to highlight the importance of this database in biotechnology, with the expectation that the number of scientists interested in the utilization of Protein Data Bank's resources will increase substantially in the coming years.

KEYWORDS

PDB, proteins, nucleic acids, RNA, DNA, drug design, vaccines, biotechnology, COVID-19

INTRODUCTION

The Protein Data Bank (PDB) is known as an international virtual data core, which serves as a fundamental information source in association with atomic structures, crystallography and three-dimensional (3D) structures of biomolecules, including nucleic acids and proteins (e.g., enzymes, immunoglycoproteins, adhesins) which are applicable for education and research. In this regard, biotechnology, biopharmaceutics, bioengineering, biomedicine, biology are disciplines that are directly dependent on the use of PDB [1–7]. Indeed, the data and information regarding crystallography and 3D structures of biomolecules released by PDB enable us to have an effective prognostication about the biochemical, biophysical and physicochemical properties comprising affinities and bonds of the related macromolecules and small biomolecules [2, 8–10]. Since 1971, the PDB as the first global open access recourse, which serves invaluable digital data for free. This international public good, supports vital data and information to visualize the biological structures and the related bindings between macro- and small biomolecules. Since 2013, the management of PDB is in accordance with the FAIR (the acronym depicts: Findable, Accessible, Interoperable, Reusable)

*Corresponding author. Tel.: +36-62-342-532.
E-mail: gajdacs.mario@stoma.szote.u-szeged.hu

guiding principles for scientific data [2, 11]. Figure 1 shows the timeline of PDB progression (https://www.rcsb.org/pages/about-us/history) [2, 12–18].

Interestingly, the open access "treasure" of PDB archives and represents several thousands of biomolecules to global users. Atomic and molecular structures of biological molecules together with their complexes (biomolecule-specific ligand(s)) are archived in PDB. Simultaneously, the PDB archive gets bigger and bigger every year. Up to now, the PDB is recognized as a high-managed resource for effective biodata. The FAIR principles are guaranteed via the application of OneDep software system. This software system controls the input structure data receiving by PDB data ecosystem for being validated, standard and biocurated. This process makes the data representing by PDB as findable, accessible, interoperable and reusable [11, 19–21]. Since the establishment of wwPDB [21] in 2003 (Fig. 1) up to now, several biocurators have been recruited by wwPDB centers in different continents such as Asia, Europe and the Americas. A collection of basic sciences and skills comprising enzymology, biophysics, computational chemistry, biochemistry, small molecule crystallography, electron microscopy, macromolecular crystallography and nuclear magnetic resonance (NMR) spectrometry supports the structural biology as the front line aim and goal of the PDB archive [19]. Even during the severe acute respiratory syndrome–related coronavirus (SARS-CoV-2) pandemic era, more than 2000 structures associated with the causative agent of the coronavirus disease (COVID-19) were released and have become accessible for global users for free. A brief collection of PDB deposits is available on SARS-CoV-2 related structures page (https://covid-19.bioreproducibility.org/) [7]. The structural properties of different organisms e.g., COVID-19 released by PDB archives give us this opportunity to find out the spatial conformation of ligands, ligand binding sites, protein-protein interactions and amino acid substitutions regarding different viral proteins. The related data may also be represented by other centers and websites rather than PDB (https://www.rcsb.org/news?year=2020&article=5e74d55d2d410731e9944f52&feature=true), including the COVID-19 Data Portal (https://www.covid19dataportal.org/) and PDBe-KB COVID-19 Data Portal (https://www.ebi.ac.uk/pdbe/covid-19) among others.

Moreover, chemical, functional and energetic characteristics are effective data, which may be gained from PDB to describe the potential capabilities for each individual molecule. These properties belonging to each structure and organisms may support us to determine the potential drug targets for drug design and vaccine preparation [22]. As an important documentary evidences, 210 new molecular entities (NMEs) were discovered and developed during a period of 2010–2016 and then were approved by the US Food and Drug Administration (FDA). The primary 3D structural data and information belonging to all of these NMEs compartments, were first produced and released via PDB archive. The representation of the related structures encouraged pharma companies to finance in drug discovery and development [2, 23]. Due to this fact, the aim of this review article is to show the vital importance of RCSB PDB as a virtual information "treasure" for research in biotechnology.

METHODS (LITERATURE SEARCH)

The design of the present manuscript is a narrative review, with the aim of critically analyzing and contextualizing the

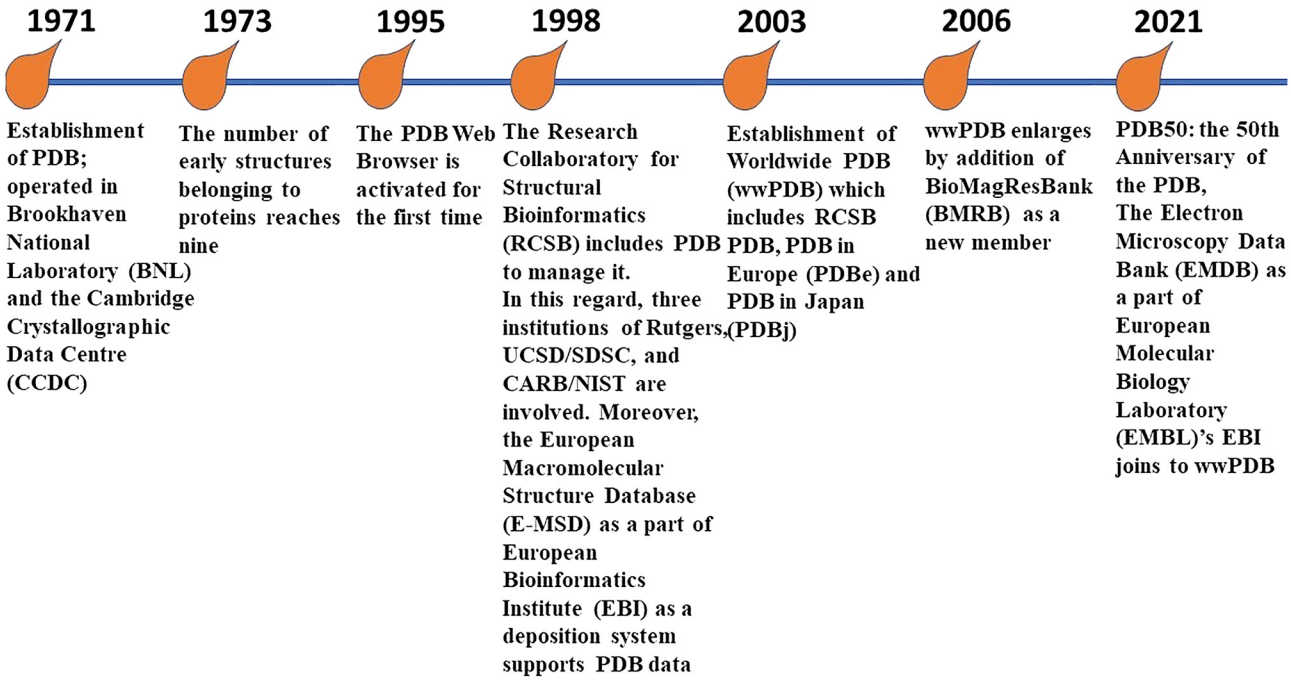

Fig. 1. Timeline of historical evolution of Protein Data Bank (PDB)

present knowledge and future perspectives on PDB. To formulate the present manuscript, a literature search was performed by the authors in the PubMed/MEDLINE, SCOPUS, EMBASE, and Web of Science databases up to 1st of September, 2021. No restrictions on article type, language or year of publication were set. The authors examined the primary search results and selected papers based on their suitability to be included in this review paper. After the selection of appropriate articles, the reference lists of these papers were also screened for relevant articles. Additionally, in case of some sub-topics of the review, authors also used references from their personal collection, totaling in $n = 106$ references.

PROTEIN DATA BANK (PDB)

The establishment of PDB in 1971 as an effective global open access resource for biological digital data was initiated by the introduction of only seven structures of proteins; and now at the time of writing this article PDB houses >182,600 biological macromolecule structures (https://www.rcsb.org/) pertaining to DNAs, proteins, RNAs, these biological molecules complexes with other molecules (e.g., drugs). The foundation of PDB as a unique feature was happened for the first time in the world's science history. Nowadays, PDB is identified as a remarkable gold standard and a great investment for archiving digital data regarding 3D structures of biological molecules. Therefore, PDB currently is known as an outstanding reference for researchers, trainers and students in the fields of applied and basic sciences associated with biology and biomedicine [23, 24].

For ensuring the highly validation and well-expertized biocurated of archived 3D macromolecular structures in PDB, the International consortium of wwPDB (RCSB PDB [25], PDB in Europe (PDBe) [26], PDB Japan (PDBj) [17] and Biological Magnetic Resonance Data Bank (BMRB) [27, 28]) (Fig. 1) has launched the OneDep software system which is known as a deposition-biocuration-validation tool [29]. These evaluations are achieved through professional expertized processes e.g., 3D cryo-electron microscopy (3DEM), X-ray crystallography and NMR [29]. Indeed, OneDep covers the wwPDB consortium through its unified software tool for deposition, biocuration and validation of the represented archived data associated with macromolecular structures [28]. To promote the validation and the quality of archived structures data in the wwPDB archive, availability of raw experimental data is enforced. OneDep system controls any ambiguity issues associated with experimental data and/or atomic models. This process facilitates the following handling processes for depositors to check and accomplishing any correction regarding a PDB deposition. Further doubtful issues will be rechecked by the manuscript reviewers or via wwPDB biocurators. To reduce the duration of validation process and to convene the validation task forces (VTFs) and effective validation metrics, the wwPDB has recruited a the OneDep software tool (https://deposit.wwpdb.org) for depositors server (https://validate.wwpdb.org/) [29] to check the experimental methodology containing electron microscopy [30], electron crystallography [31], solid-state- and solution NMR [31, 32], neutron diffraction [33], X-Ray diffraction [34, 35], fiber diffraction [24].

THE ONEDEP SOFTWARE TOOL

The main goal of an open access digital data resource organization like wwPDB is to distribute high-quality data and information with no limitations to its global users. To provide this condition, the PDB archive is supported by strong system to enhance the quality of disseminated data. Today, the PDB archive as a progressive digital data resource encompasses numerous structures which are provided through 3DEM, crystallography and NMR spectroscopy [28]. These progressions are resulting from the successful efforts by the structural biology community. Simultaneously, the PDB archive is responsible for the validity of the released data. Due to this responsibility, since January 2014 the wwPDB employed the OneDep software system to support the atomic 3D structures obtained via crystallography (X-ray). Two years later in January 2016, the OneDep system was recruited for those structures obtained by 3DEM, crystallography (X-ray) and NMR [28]. Interestingly, the advanced OneDep software controls the repositories which are contained of a huge number of experimental data pertaining to crystallography (X-ray), 3DEM and NMR. These professional interoperations ensure the uniqueness of deposited data to assign PDB code. Subsequently, the deposited data get BMRB and Electron Microscopy Data Bank (EMDB) codes. In parallel with this, the employment of advanced OneDep system guarantees the uniformity, quality and accuracy of represented data and information through the wwPDB system [28].

The OneDep software tool is capable to support the most experimental approaches and tools as a single technique or combined ones. Moreover, the OneDep system recognizes and obstructs the defective deposited data; includes the new accepted data for different structures; controls the related data automatically in the process of deposition; checks the pre-validation reports before data deposition, supports the release of the molecular structures under deposition-biocuration-validation responsibilities in PDB archive and provides a quality service for global depositors in different geographical situation [15, 28, 29]. By conclusion of data deposition through the wwPDB OneDep validation pipeline, a pre-validation report is represented to depositor. The depositor reviews the deposited data to accept or reject pre-validation report. If accepted, the uploaded data undergo for biocuration. The biocurator analyses the accuracy of the obtained data. Accepted data by biocurators enters to the final step as the official validated data. The final validation report will be released by the wwPDB centers [29]. The official validation report issued by wwPDB involves entire

quality score for a PDB submission and certain issues. The wwPDB validation reports are accessible through the https://www.wwpdb.org/validation/validation-reports link [15, 28, 29]. The validation report issued by wwPDB is consisted of overall quality at a glance, entry composition, residue-property plot, data and refinement statistics, model quality, fit of model and data [15, 21, 29, 36].

The wwPDB data centers are able to serve their users around the world. The PDBe/UK (www.pdbe.org) supports Europe and Africa, the PDBj/Japan (www.pdbj.org) serves the Middle East and Asia and the RCSB PDB/US (www.rcsb.org) covers the Oceania and Americas [14, 17, 28, 37]. Due to this knowledge, each partner of PDB consortium e.g., PDBe is involved in processes data deposition. In addition, PDBe as a partner participates in archiving and releasing the related data pertaining to molecular structures. In parallel with these activities, the PDBe recruits advanced software tools and systems to serve their users by quality data availability, analyses and visualization. These facilities help the global users from drug discovery researchers to protein engineering scientists to find their target structure(s) much easier and have a fruitful interpretation from the target macromolecular structure(s). All in all, the partners of PDB consortium try to keep data resources in accordance with FAIR guiding principles [11, 15, 37].

PROTEIN DATA BANK IN EUROPE (PDBE): AN EFFECTIVE PARTNER OF WWPDB

As a partner of PDB consortium, PDBe collaborates with different resources of bioinformatics to enrich its data center. PDBe represents a collection of bioinformatic data through the project of Structure Integration with Function, Taxonomy and Sequence (SIFTS, http://pdbe.org/sifts/) [38]. The SIFTS project provides huge amounts of data pertaining to protein sequences and structures and annotations. This project bridges the core resources of PDBe and the Universal Protein Resource (UniProt) Knowledgebase (UniProtKB, http://uniprot.org) at the European Bioinformatics Institute (EMBL-EBI; http://www.ebi.ac.uk) [38, 39]. A portion of annotation resources which cover the SIFTS project data are consisted of CATH (https://www.cathdb.info) [40], Ensembl (www.ensembl.org) [41], Gene3D (http://gene3d.biochem.ucl.ac.uk/Gene3D/) [40, 42], Gene Ontology Annotation (GO/GOA) (http://www.ebi.ac.uk/GOA) [43], HomoloGene (https://www.ncbi.nlm.nih.gov/homologene) [44], Integrated relational Enzyme database (IntEnz) (http://www.ebi.ac.uk/intenz) [45], Integrative classification of Protein sequences (InterPro) (https://www.ebi.ac.uk/interpro/) [46], Protein families database (Pfam) (http://pfam.xfam.org/) [47], NCBI Taxonomy (https://www.ncbi.nlm.nih.gov/taxonomy/) [48], PubMed (http://www.ncbi.nlm.nih.gov/pubmed) [49] and Structural Classification of Proteins (SCOP) (http://scop.mrc-lmb.cam.ac.uk) [50].

In addition to SIFTS, FunPDBe is another project which supports Protein Data Bank in Europe-Knowledge Base (PDBe-KB) (https://pdbe-kb.org). In another word, the PDBe-KB contains all the data belongs to the projects of SIFTS and FunPDBe. The functional annotations and predictions associated with molecular structures data in the PDB archive are merged and compared through PDBe-KB [51]. Indeed, PDBe-KB supports the enhancement of annotations visibility disseminated by data resources and simultaneously decreases the splitting of annotations [51]. The structural data belonging to PDB are applied via a huge number of scientific software tools and data resources. In parallel with this feature, several numbers of these data resources promote the biological context of macromolecular structures through adding a wide range of effective annotations associated with biophysical and biochemical characteristics relating to data [51]. Due to this knowledge, biomacromolecular tunnels and pores, molecular pockets and channels [52], ligand binding sites [53–55], interactions between biomolecar complexes [56], structural and functional analyses of single nucleotide polymorphisms (SNPs) in biomolecules [57] and proteins catalytic sites [58, 59].

It is important that, several effective centers for bioinformatics e.g., InterPro [46], MobiDB (https://mobidb.org/) [60], PDBsum [61], PDBj [62], Pfam [47], RSCB PDB [63, 64], Reactome (https://reactome.org) [65], SCOP2 [50, 66] and UniProt [67] count on SIFTS as an active resource data to represent fruitful links between PDB consortium and the other biological bioinformatic digital data for serving their global users with up-to-date data and information [38]. The PDBe at the European Molecular Biology Laboratory (EMBL)-European Bioinformatics Institute (EBI) manages PDBe-KB; an activity which is covered by ELIXIR 3DBioInfo community [16, 68, 69]. Molecular recognition of inhibitors, signaling molecules and adaptors and substrates determine the strength of protein functions. Molecular dynamics and the dynamic characteristics of protein molecules are directly involved in spatial configuration and folding and unfolding activities of proteins. In this regard, a mass of software tools and systems has been designed and made [70–74].

The annotations pertaining to structural and functional data associated with proteins represent an effective activity in the field of protein engineering (e.g., antibodies and enzymes). Due to this fact, the canonical structures were identified in spatial configurations of antibodies' 3D structures within their hypervariable domains. Indeed, the pivotal role of biocomputational methods in determination of canonical structures in 3D structures belonging to immunoglobulin molecules led to influential progression in predictive procedures through the bioinformatic and computational tools and techniques to obtain effective and accurate structural data in antibodies and other proteins. The effective and strong employment of bioinformatic and biocomputational procedures and methodologies in protein engineering resulted in development and progression in biotechnology through the establishment of a significant number of biotechnological companies to represent influent clinical procedures, tools and methodologies for advanced research fields [68, 75, 76].

ELIXIR encompasses a wide range of platforms which is able to support different digital data centers around Europe. The PDBe and InterPro – as the core digital resources of ELIXIR – are linked to other important annotation and structure prediction resources including CATH-Gene3D [42], FUGUE [77], GenTHREADER [78], PHYRE [79], SUPERFAMILY [80] and SWISS-MODEL [81]. Moreover, since 2018 BRENDA enzyme data base (https://www.brenda-enzymes.org) is known as the ELIXIR core data resource (https://elixir-europe.org/platforms/data/core-data-resources), too [82, 83]. BRENDA as a continuous curated system releases effective and reliable data, updated categorization of enzymes and simultaneously involves new identified enzymes. BRENDA shares new and high-quality data to support the needs of global users in the fields of biotechnology, systems biology, pharmaceutics, and medicine [82]. The core data resource of BRENDA belongs to German Network for Bioinformatics Infrastructure (de.NBI (https://www.denbi.de/)) which is covered by the German Node of ELIXIR [82, 84].

The availability, 3D visualization and structural analyses of macromalecules constitute the core of structural biology and structural bioinformatics. Hence, the recruitment of Mol*Viewer as a part of the Mol* open-source project supports the development of a common library and tools for web-based molecular visualization, graphics and analyses. This software tool covers services for the structural biology and structural bioinformatics to feed international PDB consortium [68, 73, 85].

THE RESEARCH COLLABORATORY FOR STRUCTURAL BIOINFORMATICS PROTEIN DATA BANK (RCSB PDB)

The RCSB PDB – as the US Data Center of wwPDB – serves several thousands of American and Oceanian depositors in Americas and Oceania continents. The US Data Center of serves its millions of global users with a huge number of structural data relating to macromolecules for free, all the disseminated data via wwPDB and in particular RCSB PDB are unlimited and free of charge. It is estimated that more than 660 *k* of RCSB PDB users are students, researchers and educators (from different fields involving bioengineering, biomedicine, biotechnology and fundamental biology) who utilize PDB101 center service (www.PDB101.RCSB.org). Since 2019, the portal of RCSB PDB web has been equipped with modern software tools a systems for an easy search and availability through a full Boolean operator logic [64].

Because of the importance of 3D biostructure data in research and investigation, software tools are developed to manage the related services in the field of bioengineering, biomedicine, biotechnology and fundamental biology [14, 64]. The facilities including search of protein and nucleic acid sequences [86, 87], short sequence motifs in protein and amino acid sequences, protein structure similarities [88], recognition of amino acids constituting binding or catalytic sites and ligands [64]. Due to this information, the 3D biostructure digital data belonging to wwPDB consortium such as RCSB PDB has had pivotal role associated with drug designing, drug discovery targes and vaccines against the COVID-19 pandemic era [2, 23, 89]. At the time of writing this article, by searching the keywords of "'COVID-19' drug targets" in RCSB PDB search box you may find 178,740 viral structures (e.g., the SARS-CoV-2 Spike ectodomain, PDB ID 7CN9 [90] (Fig. 2)); SARS-CoV-2 Main Protease, PDB ID 7AQE [91] (Fig. 2); the SARS-CoV-2 spike receptor-binding domain (RBD), PDB ID 7JVB (Fig. 3) [92]; SARS-CoV-2 3CL protease, PDB ID 7DPP [93] (Fig. 3).

RCSB PDB weekly supports PDB structure data through integrating more than 40 external digital biodata resources to refresh and enrich structural views for its global users, many of them are mentioned in the PDBe section [64, 89]. As the RCSB PDB covers US PDB operations, this center receives financial supports from some important institutes including Department of Energy, the National Cancer Institute, the National Institute of Allergy and Infectious

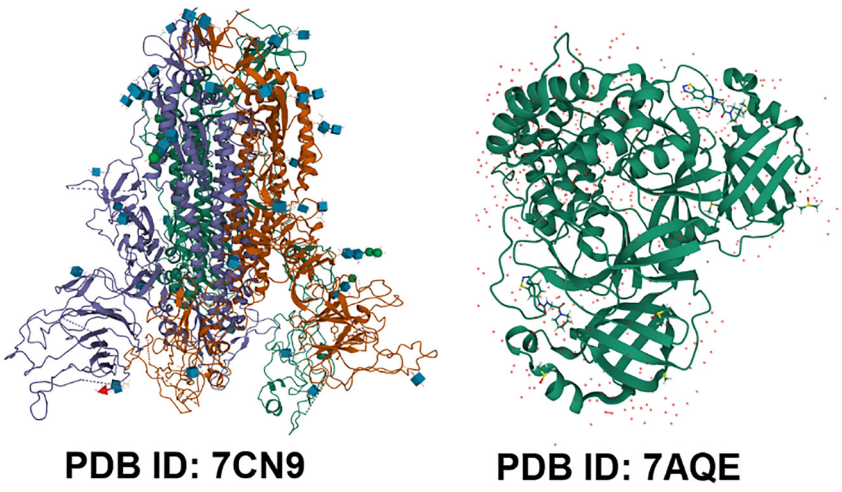

Fig. 2. SARS-CoV-2 Spike ectodomain, PDB ID 7CN9; SARS-CoV-2 Main Protease, PDB ID 7AQE

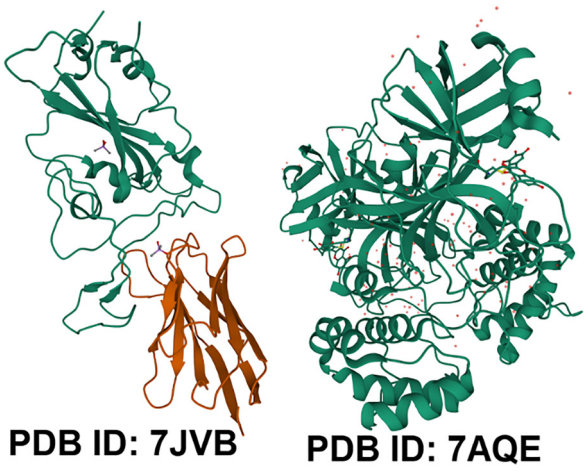

Fig. 3. SARS-CoV-2 spike receptor-binding domain (RBD), PDB ID 7JVB; SARS-CoV-2 3CL protease, PDB ID 7DPP

Diseases, the National Institute of General Medical Sciences and the National Science Foundation. Moreover, the University of California San Fransisco (UCSF), the State University of New Jersey, Rutgers and the San Diego Supercomputer Center at the University of California San Diego support the human resources and specialists of RCSB PDB [89].

The RCSB PDB as a super-professional data center controls, supports and coordinates the updating process archival data in PDBe and PDBj as the wwPDB international consortium in Europe and Asia, respectively [89]. The RCSB PDB is continuously in progression; the growth of macromolecular structures, small molecule ligands, integral membrane protein structures serves users to apply for biotechnology and the related sciences [89]. Since 2014, the National Institutes of Health (NIH) has started the project of Illuminating the Druggable Genome (IDG); the aim of this project is to detect unknown proteins and to enhance our knowledge regarding those proteins that interact with small molecules. The Target Central Resource Database (TCRD) (http://juniper.health.unm.edu/tcrd/) and Pharos (https://pharos.nih.gov/) are resulted from the IDG project. Both of TCRD and Pharaos as the IDG resources cover the related facilities to have better understanding of undiscovered regions pertaining to human genome [94]. The National Institutes of Health (NIH) Common Fund Data Resources are Pharos [95], Genotype-Tissue Expression (GTEx (https://gtexportal.org)) [96] and the International Mouse Phenotyping Consortium (IMPC (https://www.mousephenotype.org) [97]. The characterized chemical compounds supports a portion of PDB data resource and now are accessible through the wwPDB chemical component dictionary (wwPDB CCD) [98]. Moreover, the DrugBank database (https://www.drugbank.ca) [99], which collaborates with RCSB PDB, disseminates the molecular data and information associated with antibiotics and drugs, drug metabolism, drug pharmacokinetics, drug pharmacodynamics and the mechanism of their activities and the related target molecules. These facilities served by DrugBank provide the researchers to design a wide range of drugs and predict drug metabolites *in silico* [99, 100].

PROTEIN DATA BANK JAPAN (PDBJ)

The PDBj is the Japanese member of the wwPDB international consortium contributes to biological structures of macromolecules acceptance and annotation together with its other partners such as BMRB, RCSB PDB and PDBe [17, 62]. The PDBj covers the processing and annotation of those depositions received from the Middle East and Asia. All of the partners involving in wwPDB international consortium like PDBj release their updated digital structural data at midnight of Wednesday, every week. The PDBj represents updated databases and remarkable service tools for different research fields of bioinformatics and structural biology [17, 62]. The specific recruited tools in PDBj services consist of PDB mine 2 (which supports the users to search 3D structures with different resolutions and residues and clarifies the PDB metadata) [62], Molmil (a web-based molecular reviewer and graphics program (http://gjbekker.github.io/molmil/)) [62, 101], ProMode-Elastic a normal mode analysis-based database of PDB which is achieved via the program of **E**lastic-network-model based normal mode analysis (PDBETA) and computes the structures of proteins, DNAs, RNAs and ligands (https://pdbj.org/promode-elastic) [62, 102–104], electrostatic surface of functional-site (eF-site) with virtual reality (VR) technology (a database provides the electrostatic surfaces in association protein functional site (http://www.pdbj.org/eF-site/) [62, 105] and Omakage search (a web-based service to find out the global shape similarities in association with 3DEM or atomic model of biological macromolecules and the related assemblies in EMDB and PDB (https://pdbj.org/omokage) and Gaussian mixture model fitting (Gmfit) program [62, 106].

CONCLUSIONS

Even since the advent of molecular biology technologies and crystallography, it has been widely recognized that knowledge pertaining to the structures of biologically-relevant macromolecules hold valuable and critical information for chemistry, biology and various branches of medicine. However, since the beginning of the 21st century, the interest in atomic structures, three-dimensional (3D) structures of biomolecules and various molecular interaction studies have received substantial interest, both from researchers in basic science, from pharmaceutical and/or biotechnology companies, and people involved in clinical medicine. Although substantial information in this field is scattered in the literature (both in freely-available and subscription-only sources), there are few relevant, comprehensive and freely available global sources in this field. The Worldwide Protein Data Bank (wwPDB) – and its affiliates – is one of these sources, providing reliable, curated and easily accessible data and tools to visualize biological structures and the

interaction between biomolecules on the micro- and macromolecular scale, which may be relevant to all users of the biomedical sciences. The present paper aimed to surmise the main aspects, branches and advantages of using the wwPDB during research and the development for novel pharmaceutical and biotechnological products. As demonstrated by the SARS-CoV-2 pandemic, rapidly reliable and accessible biological data for microbiology, immunology, vaccinology, and drug development are critical to address many healthcare-related challenges that are facing humanity. As a consequence, the importance of databases such as wwPDB has been further validated in recent times, with the expectation that the number of scientists interested in the utilization of Protein Data Bank's resources will increase substantially in the coming years.

CRediT authorship contribution statement: **Payam Behzadi:** Conceptualization, Methodology, Formal analysis, Investigation, Writing-original draft, Writing-review&editing. **Márió Gajdács:** Supervision, Writing-original draft, Writing-review & editing, Project administration.

Declaration of competing interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding: M.G. was supported by the János Bolyai Research Scholarship (BO/00144/20/5) of the Hungarian Academy of Sciences. The research was supported by the ÚNKP-21-5-540-SZTE New National Excellence Program of the Ministry for Innovation and Technology from the source of the National Research, Development and Innovation Fund. M.G. would also like to acknowledge the support of ESCMID's "30 under 30" Award.

Ethics statement: Not applicable (review paper).

ACKNOWLEDGEMENTS

None.

REFERENCES

1. Burley SK. Impact of structural biologists and the Protein Data Bank on small-molecule drug discovery and development. J Biol Chem 2021;296.
2. Westbrook JD, Soskind R, Hudson BP, Burley SK. Impact of the protein Data Bank on antineoplastic approvals. Drug Discov Today 2020;25(5):837–50.
3. Behzadi P, García-Perdomo HA, Karpiński TM, Issakhanian L. Metallo-ß-lactamases: a review. Mol Biol Rep 2020:1–14.
4. Issakhanian L, Behzadi P. Antimicrobial agents and urinary tract infections. Curr Pharm Des 2019;25(12):1409–23.
5. Behzadi P, García-Perdomo HA, Karpiński TM. Toll-like receptors: general molecular and structural biology. J Immunol Res 2021;2021:e9914854.
6. Behzadi P, Gajdács M. Writing a strong scientific paper in medicine and the biomedical sciences: a checklist and recommendations for early career researchers. Biologia Futura 2021:1–13.
7. Wlodawer A, Dauter Z, Shabalin IG, Gilski M, Brzezinski D, Kowiel M, et al. Ligand-centered assessment of SARS-CoV-2 drug target models in the Protein Data Bank. FEBS J 2020;287(17):3703–18.
8. Blundell TL. Protein crystallography and drug discovery: recollections of knowledge exchange between academia and industry. IUCrJ 2017;4(4):308–21.
9. Burley SK, Berman HM, Christie C, Duarte JM, Feng Z, Westbrook J, 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):316–30.
10. Brown KK, Hann MM, Lakdawala AS, Santos R, Thomas PJ, Todd K. Approaches to target tractability assessment–a practical perspective. MedChemComm 2018;9(4):606–13.
11. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data 2016;3(1):1–9.
12. Berman HM, Kleywegt GJ, Nakamura H, Markley JL. The Protein Data Bank at 40: reflecting on the past to prepare for the future. Structure 2012;20(3):391–6.
13. Boutselakis H, Dimitropoulos D, Fillon J, Golovin A, Henrick K, Hussain A, et al. E-MSD: The European bioinformatics institute macromolecular structure database. Nucleic Acids Res 2003;31(1):458–62.
14. Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, Di Costanzo L, 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):D464–74.
15. Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res 2019;47(D1):D520–8.
16. Mir S, Alhroub Y, Anyango S, Armstrong DR, Berrisford JM, Clark AR, et al. PDBe: towards reusable data delivery infrastructure at protein data bank in Europe. Nucleic Acids Res 2018;46(D1):D486–92.
17. Kinjo AR, Bekker G-J, Suzuki H, Tsuchiya Y, Kawabata T, Ikegawa Y, et al. Protein Data Bank Japan (PDBj): updated user interfaces, resource description framework, analysis tools for large structures. Nucleic Acids Res 2016:gkw962.
18. Lawson CL, Patwardhan A, Baker ML, Hryc C, Garcia ES, Hudson BP, et al. EMDataBank unified data resource for 3DEM. Nucleic Acids Res 2016;44(D1):D396–403.
19. Young JY, Berrisford J, Chen M. wwPDB biocuration: on the front line of structural biology. Nat Methods 2021;18(5):431–2.
20. Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, et al. The future of biocuration. Nature 2008;455(7209):47–50.
21. Berman H, Henrick K, Nakamura H. Announcing the worldwide protein data bank. Nat Struct Mol Biol 2003;10(12):980.
22. Lubin JH, Zardecki C, Dolan EM, Lu C, Shen Z, Dutta S, et al. Evolution of the SARS-CoV-2 proteome in three dimensions (3D) during the first six months of the COVID-19 pandemic: bioRxiv; 2020.

23. Westbrook JD, Burley SK. How structural biologists and the Protein Data Bank contributed to recent FDA new drug approvals. Structure 2019;27(2):211–7.
24. Gabanyi MJ, Berman HM. Structural databases of biological macromolecules: eLS John Wiley & Sons; 2012.
25. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. Nucleic Acids Res 2000;28(1): 235–42.
26. Velankar S, van Ginkel G, Alhroub Y, Battle GM, Berrisford JM, Conroy MJ, et al. PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. Nucleic Acids Res 2016; 44(D1):D385–95.
27. Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, et al. BioMagResBank. Nucleic Acids Research 2007; 36(suppl_1):D402–8.
28. Young JY, Westbrook JD, Feng Z, Sala R, Peisach E, Oldfield TJ, et al. OneDep: unified wwPDB system for deposition, biocuration, and validation of macromolecular structures in the PDB archive. Structure 2017;25(3):536–45.
29. Gore S, García ES, Hendrickx PM, Gutmanas A, Westbrook JD, Yang H, et al. Validation of structures in the protein Data Bank. Structure 2017;25(12):1916–27.
30. Henderson R, Sali A, Baker ML, Carragher B, Devkota B, Downing KH, et al. Outcome of the first electron microscopy validation task force meeting. Structure 2012;20(2):205–14.
31. Read RJ, Adams PD, Arendall III WB, Brunger AT, Emsley P, Joosten RP, et al. A new generation of crystallographic validation tools for the protein data bank. Structure 2011;19(10):1395–412.
32. Montelione GT, Nilges M, Bax A, Güntert P, Herrmann T, Richardson JS, et al. Recommendations of the wwPDB NMR validation task force. Structure 2013;21(9):1563–70.
33. Liebschner D, Afonine PV, Moriarty NW, Langan P, Adams PD. Evaluation of models determined by neutron diffraction and proposed improvements to their validation and deposition. Acta Crystallogr Section D: Struct Biol 2018;74(8):800–13.
34. Meyer PA, Socias S, Key J, Ransey E, Tjon EC, Buschiazzo A, et al. Data publication with the structural biology data grid supports live analysis. Nat Commun 2016;7(1):1–12.
35. Grabowski M, Langner KM, Cymborowski M, Porebski PJ, Sroka P, Zheng H, et al. A public database of macromolecular diffraction experiments. Acta Crystallogr Section D: Struct Biol 2016;72(11): 1181–93.
36. Berman H, Henrick K, Nakamura H, Markley JL. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res 2007;35(suppl_1):D301–3.
37. Armstrong DR, Berrisford JM, Conroy MJ, Gutmanas A, Anyango S, Choudhary P, et al. PDBe: improved findability of macromolecular structure data in the PDB. Nucleic Acids Res 2020;48(D1): D335–43.
38. Dana JM, Gutmanas A, Tyagi N, Qi G, O'Donovan C, Martin M, 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):D482–9.
39. Cook CE, Bergman MT, Cochrane G, Apweiler R, Birney E. The European Bioinformatics Institute in 2017: data coordination and integration. Nucleic Acids Res 2018;46(D1):D21–9.
40. Sillitoe I, Bordin N, Dawson N, Waman VP, Ashford P, Scholes HM, et al. CATH: increased structural coverage of functional space. Nucleic Acids Res 2021;49(D1):D266–73.
41. Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, et al. Ensembl 2020. Nucleic Acids Res 2020;48(D1):D682–8.
42. Lewis TE, Sillitoe I, Dawson N, Lam SD, Clarke T, Lee D, et al. Gene3D: extensive prediction of globular domains in proteins. Nucleic Acids Res 2018;46(D1):D435–9.
43. Huntley RP, Sawford T, Mutowo-Meullenet P, Shypitsyna A, Bonilla C, Martin MJ, et al. The Goa database: gene ontology annotation updates for 2015. Nucleic Acids Res 2015;43(D1): D1057–63.
44. Sayers EW, Beck J, Bolton EE, Bourexis D, Brister JR, Canese K, et al. Database resources of the national center for biotechnology information. Nucleic Acids Res 2021;49(D1):D10.
45. Fleischmann A, Darsow M, Degtyarenko K, Fleischmann W, Boyce S, Axelsen KB, et al. IntEnz, the integrated relational enzyme database. Nucleic Acids Res 2004;32(suppl_1):D434–7.
46. Blum M, Chang H-Y, Chuguransky S, Grego T, Kandasaamy S, Mitchell A, et al. The InterPro protein families and domains database: 20 years on. Nucleic Acids Res 2021;49(D1):D344–54.
47. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer EL, et al. Pfam: the protein families database in 2021. Nucleic Acids Res 2021;49(D1):D412–9.
48. Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database 2020;2020.
49. Fiorini N, Lipman DJ, Lu Z. Cutting edge: towards PubMed 2.0. Elife 2017;6:e28801.
50. 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):D376–82.
51. Varadi M, Berrisford J, Deshpande M, Nair SS, Gutmanas A, Armstrong D, et al. PDBe-KB: a community-driven resource for structural and functional annotations. Nucleic Acids Res 2020; 48(D1):D344–53.
52. Pravda L, Sehnal D, Svobodová Vařeková R, Navrátilová V, Toušek D, Berka K, et al. ChannelsDB: database of biomacromolecular tunnels and pores. Nucleic Acids Res 2018; 46(D1):D399–405.
53. Krivák R, Hoksza D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. J Cheminformatics 2018;10(1):1–12.
54. Tym JE, Mitsopoulos C, Coker EA, Razaz P, Schierz AC, Antolin AA, et al. canSAR: an updated cancer research and drug discovery knowledgebase. Nucleic Acids Res 2016;44(D1):D938–43.
55. Wass MN, Kelley LA, Sternberg MJ. 3DLigandSite: predicting ligand-binding sites using similar structures. Nucleic Acids Res 2010;38(suppl_2):W469–73.
56. Vangone A, Spinelli R, Scarano V, Cavallo L, Oliva R. COCOMAPS: a web application to analyze and visualize contacts at the interface of biomolecular complexes. Bioinformatics 2011;27(20): 2915–6.
57. Lu H-C, Herrera Braga J, Fraternali F. PinSnps: structural and functional analysis of SNPs in the context of protein interaction networks. Bioinformatics 2016;32(16):2534–6.

58. 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):D618–23.
59. Sharir-Ivry A, Xia Y. Quantifying evolutionary importance of protein sites: a Tale of two measures. PLoS Genet 2021;17(4): e1009476.
60. Piovesan D, Necci M, Escobedo N, Monzon AM, Hatos A, Mičetić I, et al. MobiDB: intrinsically disordered proteins in 2021. Nucleic Acids Res 2021;49(D1):D361–7.
61. Laskowski RA, Jabłońska J, Pravda L, Vařeková RS, Thornton JM. PDBsum: structural summaries of PDB entries. Protein Sci 2018; 27(1):129–34.
62. Kinjo AR, Bekker GJ, Wako H, Endo S, Tsuchiya Y, Sato H, et al. New tools and functions in data-out activities at Protein Data Bank Japan (PDBj). Protein Sci 2018;27(1):95–102.
63. Rose Y, Duarte JM, Lowe R, Segura J, Bi C, Bhikadiya C, 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;433(11):166704.
64. Burley SK, Bhikadiya C, Bi C, Bittrich S, Chen L, Crichlow GV, 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 Acids Res 2021;49(D1):D437–51.
65. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. Nucleic Acids Res 2020;48(D1):D498–503.
66. Andreeva A, Howorth D, Chothia C, Kulesha E, Murzin AG. Investigating protein structure and evolution with SCOP2. Curr Protoc Bioinformatics 2015;49(1):1.26. 1-1. 1.
67. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res 2021;49(D1):D480–9.
68. Orengo C, Velankar S, Wodak S, Zoete V, Bonvin AM, Elofsson A, et al. A community proposal to integrate structural bioinformatics activities in ELIXIR (3D-Bioinfo Community). F1000Research 2020;9.
69. Orengo C, Schneider B, Schwede T, Sussman JL, Thornton JM, Velankar S, et al. Coordination of structural bioinformatics activities across Europe. F1000Research 2018;7.
70. Śledź P, Caflisch A. Protein structure-based drug design: from docking to molecular dynamics. Curr Opin Struct Biol 2018;48: 93–102.
71. Gioia D, Bertazzo M, Recanatini M, Masetti M, Cavalli A. Dynamic docking: a paradigm shift in computational drug discovery. Molecules 2017;22(11):2029.
72. Rachman MM, Barril X, Hubbard RE. Predicting how drug molecules bind to their protein targets. Curr Opin Pharmacol 2018;42:34–9.
73. Chodera JD, Noé F. Markov state models of biomolecular conformational dynamics. Curr Opin Struct Biol 2014;25:135–44.
74. Vreede J, Juraszek J, Bolhuis PG. Predicting the reaction coordinates of millisecond light-induced conformational changes in photoactive yellow protein. Proc Natl Acad Sci 2010;107(6): 2397–402.
75. Chothia C, Lesk AM. Canonical structures for the hypervariable regions of immunoglobulins. J Mol Biol 1987;196(4):901–17.
76. Chothia C, Lesk AM, Levitt M, Amit AG, Mariuzza RA, Phillips S, et al. The predicted structure of immunoglobulin D1. 3 and its comparison with the crystal structure. Science 1986;233(4765): 755–8.
77. Shi J, Blundell TL, Mizuguchi K. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 2001; 310(1):243–57.
78. McGuffin LJ, Street SA, Bryson K, Sørensen SA, Jones DT. The Genomic Threading Database: a comprehensive resource for structural annotations of the genomes from key organisms. Nucleic Acids Res 2004;32(suppl_1):D196–9.
79. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 2015;10(6):845–58.
80. Pandurangan AP, Stahlhacke J, Oates ME, Smithers B, Gough J. The SUPERFAMILY 2.0 database: a significant proteome update and a new webserver. Nucleic Acids Res 2019;47(D1):D490–4.
81. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res 2018; 46(W1):W296–303.
82. Chang A, Jeske L, Ulbrich S, Hofmann J, Koblitz J, Schomburg I, et al. BRENDA, the ELIXIR core data resource in 2021: new developments and updates. Nucleic Acids Res 2021;49(D1):D498–508.
83. Jeske L, Placzek S, Schomburg I, Chang A, Schomburg D. BRENDA in 2019: a European ELIXIR core data resource. Nucleic Acids Res 2019;47(D1):D542–9.
84. Drysdale R, Cook CE, Petryszak R, Baillie-Gerritsen V, Barlow M, Gasteiger E, et al. The ELIXIR Core Data Resources: fundamental infrastructure for the life sciences. Bioinformatics 2020.
85. Sehnal D, Bittrich S, Deshpande M, Svobodová R, Berka K, Bazgier V, et al. Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. Nucleic Acids Res 2021.
86. Mirdita M, Steinegger M, Söding J. MMseqs2 desktop and local web server app for fast, interactive sequence searches. Bioinformatics 2019;35(16):2856–8.
87. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol 2017;35(11):1026–8.
88. Guzenko D, Burley SK, Duarte JM. Real time structural search of the protein Data Bank. PLoS Comput Biol 2020;16(7): e1007970.
89. Goodsell DS, Zardecki C, Di Costanzo L, Duarte JM, Hudson BP, Persikova I, et al. RCSB Protein Data Bank: enabling biomedical research and drug discovery. Protein Sci 2020;29(1):52–65.
90. Liu Y-M, Shahed-Al-Mahmud M, Chen X, Chen T-H, Liao K-S, Lo JM, et al. A carbohydrate-binding protein from the edible Lablab beans effectively blocks the infections of influenza viruses and SARS-CoV-2. Cel Rep 2020;32(6):108016.
91. Günther S, Reinke PY, Fernández-García Y, Lieske J, Lane TJ, Ginn HM, et al. X-ray screening identifies active site and allosteric inhibitors of SARS-CoV-2 main protease. Science 2021;372(6542): 642–6.
92. Xiang Y, Nambulli S, Xiao Z, Liu H, Sang Z, Duprex WP, et al. Versatile and multivalent nanobodies efficiently neutralize SARS-CoV-2. Science 2020;370(6523):1479–84.

93. Su H, Yao S, Zhao W, Zhang Y, Liu J, Shao Q, et al. Identification of pyrogallol as a warhead in design of covalent inhibitors for the SARS-CoV-2 3CL protease. Nat Commun 2021;12(1):1–12.
94. Sheils TK, Mathias SL, Kelleher KJ, Siramshetty VB, Nguyen D-T, Bologa CG, et al. TCRD and Pharos 2021: mining the human proteome for disease biology. Nucleic Acids Res 2021;49(D1): D1334–46.
95. Nguyen D-T, Mathias S, Bologa C, Brunak S, Fernandez N, Gaulton A, et al. Pharos: collating protein information to shed light on the druggable genome. Nucleic Acids Res 2017;45(D1): D995–1002.
96. Consortium G. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science 2020;369(6509): 1318–30.
97. Skarnes WC, Rosen B, West AP, Koutsourakis M, Bushell W, Iyer V, et al. A conditional knockout resource for the genome-wide study of mouse gene function. Nature 2011;474(7351):337–42.
98. 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): 1274–8.
99. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 2018;46(D1):D1074–82.
100. Wishart DS, Wu A. Using DrugBank for in silico drug exploration and discovery. Curr Protoc Bioinformatics 2016;54(1):14.4. 1-.4. 31.
101. Bekker G-J, Nakamura H, Kinjo AR. Molmil: a molecular viewer for the PDB and beyond. J Cheminformatics 2016;8(1):1–5.
102. Wako H, Endo S. Normal mode analysis as a method to derive protein dynamics information from the Protein Data Bank. Biophysical Rev 2017;9(6):877–93.
103. Wako H, Endo S. Normal mode analysis based on an elastic network model for biomolecules in the Protein Data Bank, which uses dihedral angles as independent variables. Comput Biol Chem 2013;44:22–30.
104. Wako H, Endo S. Ligand-induced conformational change of a protein reproduced by a linear combination of displacement vectors obtained from normal mode analysis. Biophysical Chem 2011;159(2–3):257–66.
105. Kinoshita K, Nakamura H. eF-site and PDBjViewer: database and viewer for protein functional sites. Bioinformatics 2004;20(8):1329–30.
106. Suzuki H, Kawabata T, Nakamura H. Omokage search: shape similarity search service for biomolecular structures in both the PDB and EMDB. Bioinformatics 2016;32(4):619–20.