Received: 5 August 2022 | Revised: 2 September 2022 | Accepted: 5 September 2022

DOI: 10.1002/pro.4439

TOOLS FOR PROTEIN SCIENCE

PDBe and PDBe-KB: Providing high-quality, up-to-date and integrated resources of macromolecular structures to support basic and applied research and education

Mihaly Varadi[1] | **Stephen Anyango**[1] | **Sri Devan Appasamy**[1] | **David Armstrong**[1] | **Marcus Bage**[1] | **John Berrisford**[1] | **Preeti Choudhary**[1] | **Damian Bertoni**[1] | **Mandar Deshpande**[1] | **Grisell Diaz Leines**[1] | **Joseph Ellaway**[1] | **Genevieve Evans**[1] | **Romana Gaborova**[1] | **Deepti Gupta**[1] | **Aleksandras Gutmanas**[1] | **Deborah Harrus**[1] | **Gerard J. Kleywegt**[1] | **Weslley Morellato Bueno**[1] | **Nurul Nadzirin**[1] | **Sreenath Nair**[1] | **Lukas Pravda**[1] | **Marcelo Querino Lima Afonso**[1] | **David Sehnal**[1,2,3] | **Ahsan Tanweer**[1] | **James Tolchard**[1] | **Charlotte Abrams**[1] | **Roisin Dunlop**[1] | **Sameer Velankar**[1]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton

[2]CEITEC – Central European Institute of Technology, Masaryk University, Brno, Czech Republic

[3]National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Brno, Czech Republic

Correspondence
Sameer Velankar, European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK.
Email: sameer@ebi.ac.uk

Funding information
Biotechnology and Biological Sciences Research Council, Grant/Award Numbers: BB/M011674/1, BB/P024351/1, BB/P025846/1, BB/R015201/1, BB/S017135/1, BB/S020071/1, BB/T01959X/1, BB/V004247/1, BB/V016113/1, BB/V018779/1, BB/W017970/1; Wellcome Trust, Grant/Award Numbers: 104948/Z/14/Z, 218303/Z/19/Z, 221327/Z/20/Z, 223739/Z/21/Z; European Bioinformatics Institute; European Molecular Biology Laboratory

Review Editor: Nir Ben-Tal

Abstract

The archiving and dissemination of protein and nucleic acid structures as well as their structural, functional and biophysical annotations is an essential task that enables the broader scientific community to conduct impactful research in multiple fields of the life sciences. The Protein Data Bank in Europe (PDBe; pdbe.org) team develops and maintains several databases and web services to address this fundamental need. From data archiving as a member of the Worldwide PDB consortium (wwPDB; wwpdb.org), to the PDBe Knowledge Base (PDBe-KB; pdbekb.org), we provide data, data-access mechanisms, and visualizations that facilitate basic and applied research and education across the life sciences. Here, we provide an overview of the structural data and annotations that we integrate and make freely available. We describe the web services and data visualization tools we offer, and provide information on how to effectively use or even further develop them. Finally, we discuss the direction of our data services, and how we aim to tackle new challenges that arise from the recent, unprecedented advances in the field of structure determination and protein structure modeling.

KEYWORDS
bioinformatics, databases, protein data bank, structural biology

1 | INTRODUCTION

The Worldwide Protein Data Bank (wwPDB) is the consortium responsible for operating the PDB archive.[1] Its members are the Protein Data Bank in Europe (PDBe; pdbe.org),[2] RCSB Protein Data Bank,[3] Protein Data Bank Japan (PDBj),[4] the Biological Magnetic Resonance Bank (BMRB),[5] and the Electron Microscopy Data Bank (EMDB).[6] The archiving of experimentally determined macromolecule structures, painstakingly determined by researchers of the global scientific community, is a fundamental and valuable task.[1] It is estimated that reproducing the experimentally determined protein and nucleic acid structures currently in the PDB archive would cost around 19 billion dollars.[7] The wwPDB members collaborate on data archiving, but in addition each site provides its own data-access services, software, and enriched data sets.

Open access to the wealth of macromolecular structure data in the PDB archive enables research and software development in several scientific fields. Structure-based drug discovery, structural bioinformatics studies, experimental determination of new protein structures and scientific software development all benefit greatly from the data stored in the PDB archive. Indeed, the recent spectacular advances in protein structure prediction using Artificial Intelligence (AI) approaches, as demonstrated by AlphaFold[8] and RoseTTaFold,[9] would not have been possible without the training set of experimentally determined structures in the PDB archive.

While the three-dimensional coordinates of macromolecules are essential for answering a variety of scientific questions, these data are often insufficient if they lack a biological context.[10,11] A lot of data resources and scientific software specialize in providing parts of this biological context through functional, biophysical, and biochemical annotations.[12–15] Yet, a significant barrier to taking systematic advantage of these valuable annotations is that they are often lacking in FAIRness,[16] that is, they might not be sufficiently findable, accessible, interoperable or reusable. In particular, findability is hampered by the fragmented nature of these data resources, making it difficult for researchers to keep track of the location and existence of the annotations. Interoperability is also a significant challenge, as each data provider may use custom data formats, ontologies, and definitions.

PDBe has worked with research groups and scientific service teams to establish the PDBe Knowledge Base (PDBe-KB; pdbekb.org) consortium to tackle these challenges.[11] PDBe-KB is an open, collaborative consortium that provides FAIR access to the biological context of macromolecular structure data. It is one of the flagships of the ELIXIR 3D-BioInfo Community, a group of researchers and software developers working on improving all aspects of data management in the field of structural bioinformatics.[17] PDBe-KB has expanded significantly since its inception in 2018, and provided valuable data sets and tools for the scientific community.

In recent years, both PDBe and PDBe-KB have grown in terms of the amount of data and the number of services they provide. The PDB archive is growing steadily, with high-resolution structures determined using electron cryo-microscopy (cryo-EM) rising swiftly. The number of Cryo-EM depositions has now overtaken those of nuclear magnetic resonance (NMR) methods, and cryo-EM depositions are approaching those of X-ray crystallography, historically the most prevalent experimental method used to determine the structures in the PDB.

While the core PDB data is the same for every wwPDB consortium member, PDBe provides additional unique services. These have been driving an increase in usage in recent years, with almost half a million monthly unique users visiting PDBe and PDBe-KB entry pages and growing access of PDBe data through its public application programming interface (API).[18] The amount of data in PDBe-KB has also increased sharply, following integration of a growing number of structural, functional, biophysical, and biochemical annotations contributed by the PDBe-KB partner resources. These data allow us to provide an increasingly comprehensive context for PDB structures, as demonstrated by the so-called "aggregated views" pages of PDBe-KB, which collate all available structural data and their annotations for a specific protein of interest.[19]

Here, we give a detailed overview of the PDBe and its sister resource, the PDBe-KB.[2,11] We will discuss the infrastructure of these massive databases and provide descriptions of the services, data sets and data visualization tools developed under the umbrella of PDBe and PDBe-KB. We will subsequently highlight the availability of high-quality, up-to-date training materials and conclude by presenting an outlook for the future of these data resources.

2 | INFRASTRUCTURE

The PDBe team develops and maintains several public and open-access data resources and scientific tools. Some of these are under the umbrella of PDBe, while others are related to the PDBe-KB, the AlphaFold Protein Structure Database,[20] or the 3D-Beacons Network (Table 1). The following sections provide a detailed overview of these services.

The infrastructure can be divided into several well-defined, and stand-alone components: data-deposition systems, databases, programmatic access software, and data visualization (Figure 1). The wwPDB common data-deposition system, OneDep,[24] handles the processing and curation of new PDB entries, while a separate system processes and integrates functional and biophysical annotations from PDBe-KB partners. All the data are loaded into our internal databases, and we provide programmatic access to data both for external users and to PDBe services. The PDBe entry pages and the PDBe-KB aggregated views of proteins display the core PDB data, and enriched structural and functional annotations from PDBe-KB partners and our internal data pipelines.

2.1 | Deposition systems

PDBe is a core member of the wwPDB consortium.[1] All wwPDB members use OneDep, the common deposition, annotation, and validation system to support the deposition of macromolecular structure data and associated experimental data to the PDB and EMDB core archives. The OneDep system is collaboratively developed by PDBe, RCSB PDB, PDBj, and EMDB.[2–4,6]

Whereas the OneDep system serves the depositors of new PDB and EMDB entries, the PDBe-KB deposition system serves data resources that are members of the PDBe-KB consortium. This deposition system focuses on functional and biophysical annotations that can enrich core PDB data. The basis of the deposition system is a data-exchange format collaboratively developed by the PDBe-KB consortium and maintained by the PDBe team. This format is a JSON (JavaScript Object Notation) specification and aims to capture the minimal required

TABLE 1 Services of the PDBe team

Name of service	Brief description	(Example) URLs
3D-Beacons API	Programmatic access to models	https://3d-beacons.org/api
3D-Beacons network	Access to experimental and theoretical structures	https://3d-beacons.org
AFDB entry pages[20]	View single AlphaFold predictions	https://www.alphafold.ebi.ac.uk/entry/Q5VSL9
Aggregated views of proteins[11]	View aggregated structural data for proteins	https://pdbe-kb.org/proteins/Q14676
Density server	Access to volumetric density data	https://www.ebi.ac.uk/pdbe/densities/doc.html
FunPDBe annotations[19]	JSON files containing functional annotations	http://ftp.ebi.ac.uk/pub/databases/pdbe-kb/annotations/
Model server	Access to (sub-)structure models of molecules	https://molstar.org/docs/data-access-tools/model-server/
Neo4j database	Downloadable graph database of PDBe and PDBe-KB data	https://www.ebi.ac.uk/pdbe/pdbe-kb/graph-download
PDBe API[18]	Programmatic access to PDBe and PDBe-KB data	https://www.ebi.ac.uk/pdbe/pdbe-rest-api
PDBe component library	Reusable web components from PDBe	https://www.ebi.ac.uk/pdbe/pdb-component-library/
PDBe download service	Download PDB data for lists of PDB entries	https://www.ebi.ac.uk/pdbe/download/docs
PDBe entry pages[2]	View single PDB entries	https://pdbe.org/3bow
PDBe Mol*[21]	Interactive 3D molecular viewer	https://github.com/molstar/pdbe-molstar
PDBe PISA	Analyse molecular assemblies	https://www.ebi.ac.uk/msd-srv/prot_int/pistart.html
PDBe ProtVista[22]	Interactive 2D sequence feature viewer	https://github.com/PDBeurope/protvista-pdb
PDBe-KB component library	Reusable web components from PDBe-KB	https://github.com/PDBe-KB?q=component
PDBeChem	Search system for small molecules	https://www.ebi.ac.uk/pdbe-srv/pdbechem/
PDBeFold	Structure-based search	https://www.ebi.ac.uk/msd-srv/ssm/
SIFTS[23]	Mapping of PDB entries to UniProt and other databases	https://www.ebi.ac.uk/pdbe/docs/sifts/

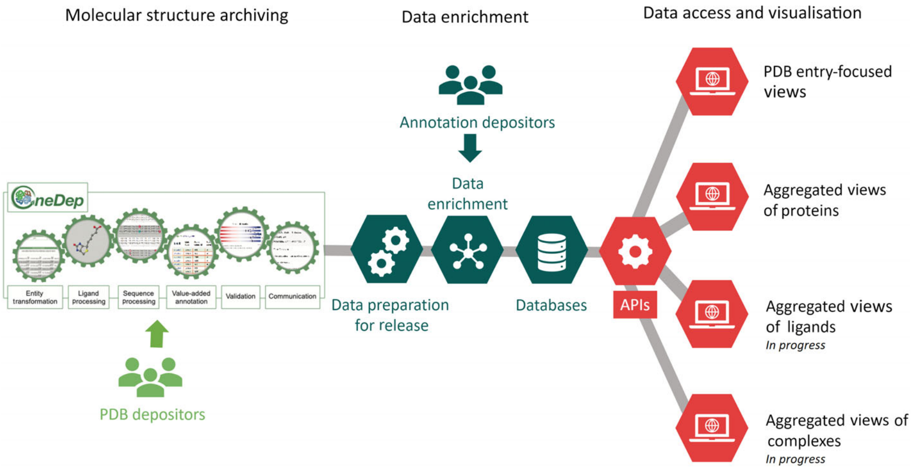


FIGURE 1 Overview of the infrastructure developed by the PDBe team The main wwPDB deposition system (OneDep) captures the core PDB data from depositors worldwide (light green). We enrich the core PDB data with annotations provided by PDBe-KB partner databases and archive the data in internal databases (dark green). Finally, we provide data-access mechanisms and visualizations for every aspect of the structural data we manage (red). PDBe, Protein Data Bank in Europe; PDBe-KB, PDBe Knowledge Base; wwPDB, Worldwide Protein Data Bank

metadata to describe residue-level annotations. For example, the schema supports archiving information on which PDB residues are predicted to form a druggable pocket or which residues have known variants that may have a deleterious effect on the stability of the protein chain.[12,25,26] The data-exchange format schema is available at https://github.com/PDBe-KB/funpdbe-schema.

Every PDBe-KB consortium member converts their specific annotations to the agreed data-exchange format and transfers these JSON file sets to the PDBe team. Each week, we process, validate and merge the annotations from the JSON files with the core PDB data. We also make the JSON files publicly available at http://ftp.ebi.ac.uk/pub/databases/pdbe-kb/annotations/. These integrated annotations provide the biological context of proteins to users through our services, for example the aggregated views of proteins (e.g., https://www.ebi.ac.uk/pdbe/pdbe-kb/proteins/P0DTD1). As of July 2022, the PDBe-KB partners have provided over 80 million residue-level annotations for over 185,000 PDB entries. These annotations provide useful information about predicted and observed ligand-binding sites, potential druggable pockets, predicted post-translational modifications (PTMs), catalytic sites, backbone flexibility, interface classification, and topology classifications, to name a few.[11]

2.2 | Databases

We integrate the PDBe-KB annotations with the core PDB data using two approaches to data archival: a Neo4j graph database, that allows us to aggregate and perform sophisticated analysis and querying, and an Oracle relational database, which is more performant in case of simpler queries. The Oracle database currently powers the majority of the PDBe entry pages and PDBe search services. The Neo4j graph database powers the PDBe-KB aggregated views pages. Additionally, the graph database is a powerful tool that researchers can query using complex sub-graph patterns to answer specific scientific questions. For example, the graph database can easily help identify ligand molecules in the PDB that have the same Murcko scaffold and bind to the same binding site of a target protein (Figure 2). See Supplementary Material SS1 for the relevant Cipher query.

We make a weekly updated copy of the graph database available to the scientific community at https://www.ebi.ac.uk/pdbe/pdbe-kb/graph-download. Installing a local copy of this database allows researchers to use it as an in-house discovery tool. It is especially powerful if users integrate their own data into the graph, allowing them to map their data to that in the PDB seamlessly. This can be done by downloading and running a copy of the PDBe graph database, and modifying and loading new nodes and edges. We provide detailed documentation of the schema of the graph database at https://www.ebi.ac.uk/pdbe/pdbe-kb/schema.

2.3 | Programmatic access

We provide programmatic access to data from both the relational and the graph database through a rich set of API endpoints. More than 90 endpoints provide data in

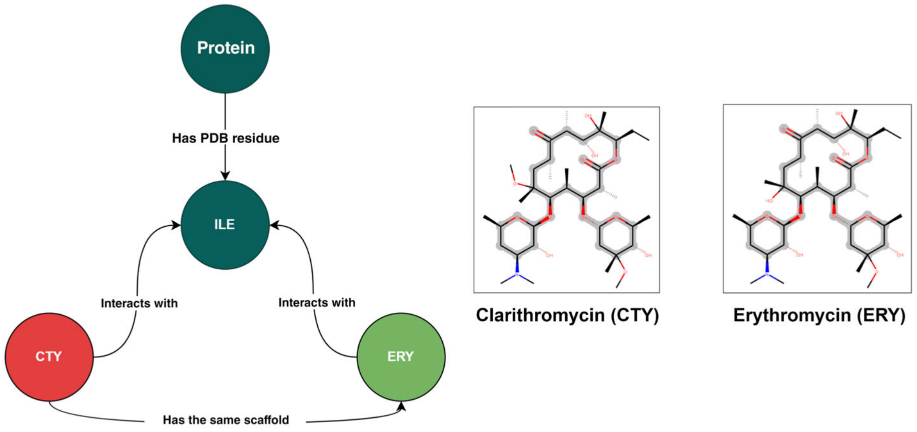


FIGURE 2 Example of using the PDBe graph database. The PDBe graph database is a powerful tool for research and discovery. In the example above, we queried the graph to identify ligand molecules in the PDB archive that bind to the same binding site and have the same molecular scaffolds. One of the ligand pairs matching this query is clarithromycin and erythromycin, a pair of antibiotics. Both molecules interact with the same amino acids of aminoglycoside phosphotransferase, for example with residue ILE105. These interactions can be viewed at https://www.ebi.ac.uk/pdbe/pdbe-kb/proteins/Q47396/ligands. PDBe, Protein Data Bank in Europe

JSON format that addresses specific use cases, such as mappings between PDB chains and UniProt accessions or lists all observed ligand-binding sites in a PDB structure.[18] To create these mappings, we use data from the Structure Integration with Function, Taxonomy, and Sequence (SIFTS)[23] resource. In addition to mappings between PDB residues and UniProt residues, SIFTS also provides annotations from the IntEnz, GO, InterPro, Pfam, CATH, SCOP, PubMed, Ensembl, and Homologene resources.[27–32]

Comprehensive documentation of both SIFTS-related and other API endpoints is available at https://www.ebi.ac.uk/pdbe/pdbe-rest-api. Users can search and test all API endpoints on these documentation pages, helping them find those relevant to their specific scientific analysis or bioinformatics pipeline needs.[18] The API endpoints are also used by the PDBe services and by external data resources and scientific software. For example, the API endpoints exposing data from the PDBe graph database power the PDBe-KB aggregated views of proteins and provide annotations to the 2D sequence-feature viewer, PDB ProtVista,[22] on the PDBe entry pages. Other data resources, such as the AlphaFold Protein Structure Database, UniProt, and GeneCards, depend on data accessed through our API endpoints.[20,33,34]

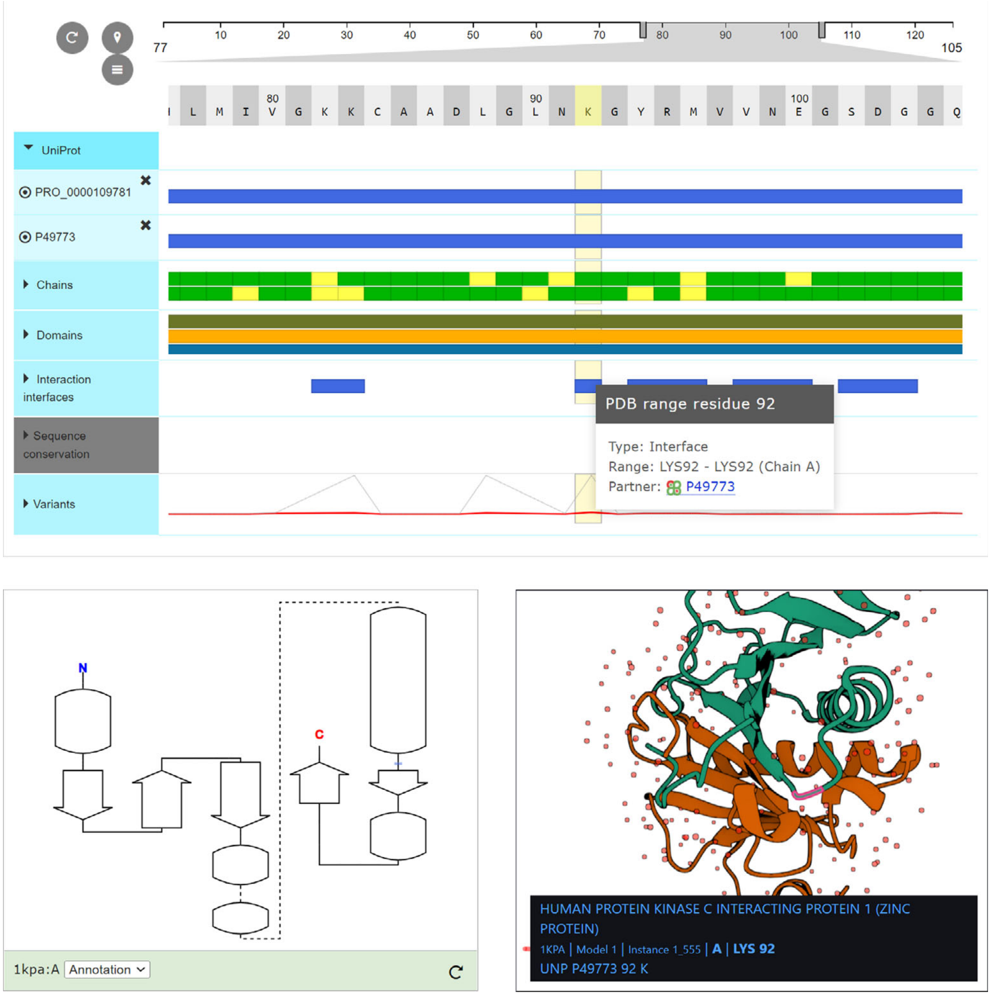
3 | PROTEIN DATA BANK IN EUROPE WEB SERVICES

PDBe offers web pages dedicated to describing individual PDB structures; for example, https://pdbe.org/3bow provides all the information for PDB entry 3bow. These entry pages offer direct access to files for download, display most of the textual metadata (such as biological function), experimental conditions, and validation information, and include interactive data-visualization tools to help users understand the molecular structure data and their annotations.

There are three primary, interactive data-visualization tools on the PDBe entry pages: the 3D molecular graphics viewer, PDBe Mol*,[21] the 2D sequence-feature viewer, PDB ProtVista,[22] and the 2D topology viewer for proteins and RNA molecules (Figure 3).

PDBe Mol* is a PDBe-specific implementation of the Mol* suite.[21] Its source code is available at https://github.com/molstar/pdbe-molstar. PDBe-Mol* includes only those functionalities of the Mol* suite, that are essential for the data visualizations on PDBe, PDBe-KB, and AlphaFold DB pages.[2,11,20] This interactive molecular graphics viewer can load, display, and save 3D visualizations of macromolecules, provide different rendering styles and support superposition views to allow users to compare various PDB structures of the same protein. PDBe Mol* is designed to have intuitive controls to ease basic usage, and documentation, with detailed interactive demos and tutorials on how to embed Mol* are provided at https://github.com/molstar/pdbe-molstar/wiki.

PDB ProtVista is another primary data-visualization tool we use on PDBe and PDBe-KB pages.[22] It is a 2D sequence-feature viewer that allows us to display the functional and biophysical annotations provided by the PDBe-KB partner resources. We make the source code of PDBe ProtVista available at https://github.com/PDBeurope/protvista-pdb and provide detailed demos and guides on integrating it with other data resources at https://github.com/PDBeurope/protvista-pdb/wiki. This viewer displays annotations in so-called "tracks." In its



PDB range residue 92

Type: Interface
Range: LYS92 - LYS92 (Chain A)
Partner: P49773

HUMAN PROTEIN KINASE C INTERACTING PROTEIN 1 (ZINC PROTEIN)
1KPA | Model 1 | Instance 1_555 | A | LYS 92
UNP P49773 92 K

FIGURE 3 PDBe Mol*, PDBe ProtVista and the protein topology viewer The three main data-visualization tools as displayed on the PDBe entry page of PDB 1KPA at https://www.ebi.ac.uk/pdbe/entry/pdb/1kpa/protein/1, corresponding to the human HINT1 protein. PDB ProtVista is at the top, the 2D topology viewer is at the bottom left, and PDBe Mol* is at the bottom right. Using these data-visualization tools, users can do visual analyses, for example, examining the potential effect of a known mutation at LYS92. It becomes apparent that this residue is in a loop region that is part of the dimeric interaction interface. Therefore, one can speculate that a mutation might disrupt or destabilize the dimerization of HINT1. PDBe, Protein Data Bank in Europe

current implementation, PDB ProtVista has four types of tracks: a segment-based track, a site/residue-based track, a variants track, and a sequence-conservation track (Supplementary Material SS2).

Finally, the topology viewer includes two distinct web components that display protein topology and RNA topology in a consistent style. Protein topology data is calculated using the PDBSum software package,[35] while for RNA topologies we generate the data using R2DT[36] and FR3D.[37] This 2D topology viewer communicates with PDBe Mol* and PDB ProtVista, allowing users to interactively identify and map residues between the three data visualization tools.

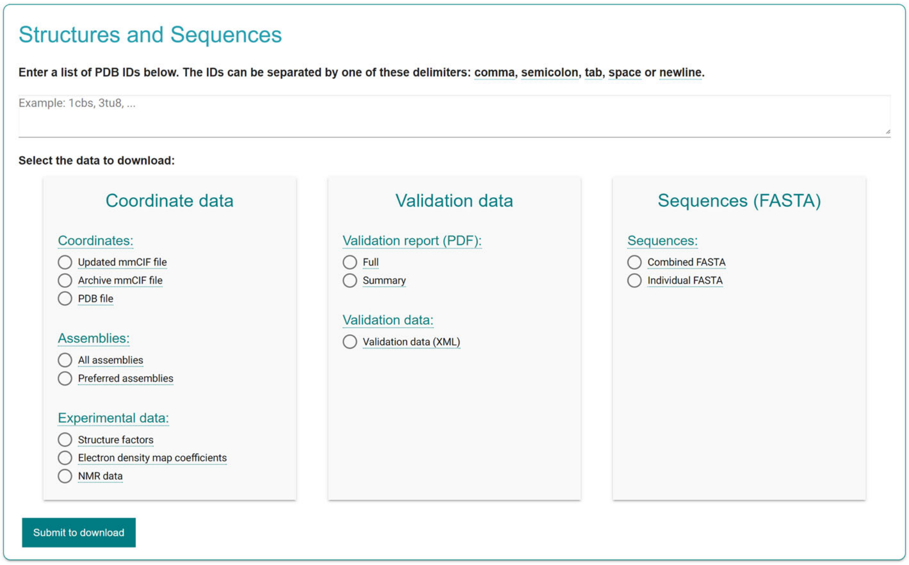
While Mol*, ProtVista, and the topology viewer are the main data-visualization web components of PDBe, we have several other reusable components for specific tasks, such as providing summary information about a PDB entry or displaying residue-interaction networks.[2] We make all these web components available to the scientific community in our PDBe web component library at https://www.ebi.ac.uk/pdbe/pdb-component-library/.

4 | PROTEIN DATA BANK IN EUROPE DOWNLOAD SERVICE

Researchers are often interested in a collection of PDB entries relevant to the specific scientific question they are investigating. While downloading data for individual PDB entries is straightforward, performing bulk downloads was often more complicated than necessary. Therefore, we have created a bulk-data download service with an intuitive user interface to help users select and download large volumes of PDB data (Figure 4). This service allows users to download coordinate data, validation data, sequences, data on ligands and other small molecules, and residue mappings between PDB and UniProt.

5 | PROTEIN DATA BANK KNOWLEDGE BASE WEB SERVICES

Currently, the main offering of PDBe-KB is the collection of pages that provide aggregated structural data and functional annotations on a per-protein basis, the so



Structures and Sequences

Enter a list of PDB IDs below. The IDs can be separated by one of these delimiters: comma, semicolon, tab, space or newline.

Example: 1cbs, 3tu8, ...

Select the data to download:

Coordinate data

Coordinates:
○ Updated mmCIF file
○ Archive mmCIF file
○ PDB file

Assemblies:
○ All assemblies
○ Preferred assemblies

Experimental data:
○ Structure factors
○ Electron density map coefficients
○ NMR data

Validation data

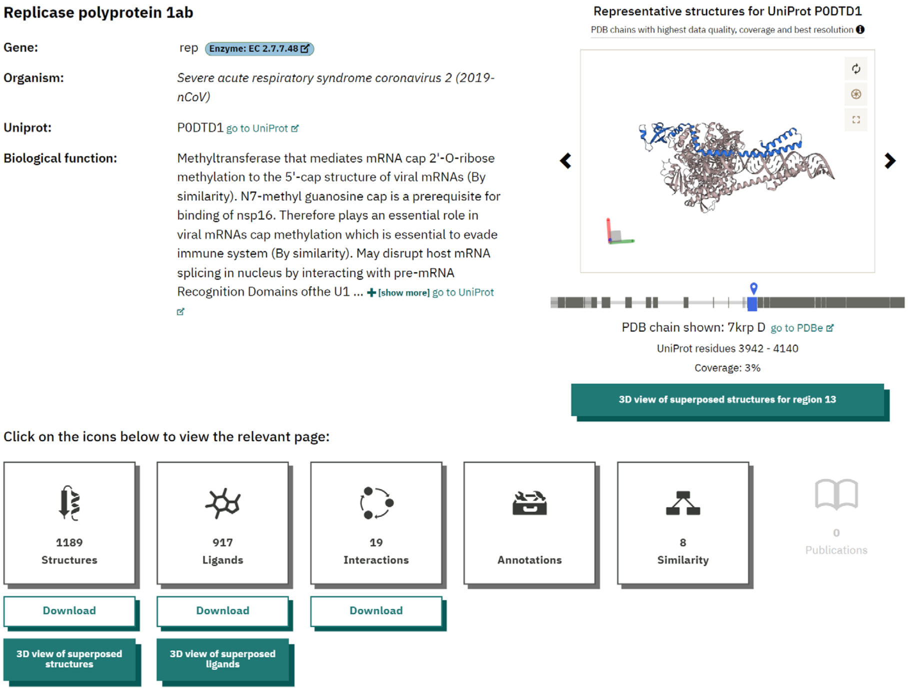
Validation report (PDF):
○ Full
○ Summary

Validation data:
○ Validation data (XML)

Sequences (FASTA)

Sequences:
○ Combined FASTA
○ Individual FASTA

FIGURE 4 Bulk-data download service We provide a bulk-data download service for all aspects of PDB data. Using a list of PDB entries or small molecule identifiers (HET codes), users can retrieve large volumes of data from coordinate files to validation data and more. PDB, Protein Data Bank



Replicase polyprotein 1ab

Gene: rep Enzyme: EC 2.7.7.48

Organism: *Severe acute respiratory syndrome coronavirus 2 (2019-nCoV)*

Uniprot: P0DTD1 go to UniProt

Biological function: Methyltransferase that mediates mRNA cap 2'-O-ribose methylation to the 5'-cap structure of viral mRNAs (By similarity). N7-methyl guanosine cap is a prerequisite for binding of nsp16. Therefore plays an essential role in viral mRNAs cap methylation which is essential to evade immune system (By similarity). May disrupt host mRNA splicing in nucleus by interacting with pre-mRNA Recognition Domains ofthe U1 ... + [show more] go to UniProt

Representative structures for UniProt P0DTD1

PDB chains with highest data quality, coverage and best resolution

PDB chain shown: 7krp D go to PDBe

UniProt residues 3942 - 4140

Coverage: 3%

Click on the icons below to view the relevant page:

3D view of superposed structures for region 13
1189
Structures
917
Ligands
19
Interactions
Annotations
8
Similarity
0
Publications
Download
Download
Download
3D view of superposed structures
3D view of superposed ligands

FIGURE 5 PDBe-KB aggregated views of proteins The latest update of the PDBe-KB aggregated views of proteins offers a streamlined user experience with better separation and description of the various data we provide access to. Each primary section, such as "structures," "ligands," etc. has its own subpage. This allowed us to fine-tune the performance of the subpages while also displaying more data before. The example shown here is the Replicase polyprotein 1ab of SARS-CoV-2, https://pdbe-kb.org/proteins/P0DTD1. PDB, Protein Data Bank; PDBe-KB, PDBe Knowledge Base

called aggregated views of proteins.[18] We have recently redesigned these pages to provide a more intuitive user interface, while also making the pages more performant to account for the massive increase in annotations provided by PDBe-KB partner resources (Figure 5).

The latest version of the aggregated views of proteins display functional annotations for ligands, such as whether an observed ligand is a reactant-like (similar to product or substrate), a cofactor-like, or a drug-like molecule.[38] We use internal data pipelines to create these annotations weekly, using the latest PDB data. These pipelines are available as part of our open-source PDBe CCDutils package (https://github.com/PDBeurope/ccdutils). The pages also include biophysical descriptions such as per-residue solvent-accessible surface area from POPS[39] and flexibility propensities from webNMA.[40] Together, the annotations from the PDBe-KB consortium place proteins in a more comprehensive biological and functional context, allowing researchers to better understand the implications of the structural data in the PDB archive.

In addition to displaying annotations and PDB structures, we also make the underlying data available to researchers and software developers. We have worked with the ELIXIR 3D-BioInfo community to make specific benchmark datasets accessible for a number of well-defined aspects of structural bioinformatics.[17] Some of these datasets focus on distinct conformational states and their biological role, while others focus on training prediction methods for ligand-binding sites, PTM sites or variants. We provide access to these datasets at http://ftp.ebi.ac.uk/pub/databases/pdbe-kb/benchmarking/.

Finally, we open-sourced all the web components we used to build the PDBe-KB aggregated views so that other data services can reuse these data visualizations: https://github.com/PDBe-KB?q=component.

6 | TRAINING MATERIALS

Keeping up with the amount and types of data can at times be challenging. Therefore, it is very important that we provide up-to-date and user-friendly training materials, tutorials, demos, and webinars for all our services that help users take advantage of the wealth of structure data and their functional annotations. These training materials are available from our training portal at https://www.ebi.ac.uk/pdbe/pdbe-training.

As part of the EMBL-EBI training initiatives, we are regularly offering webinars and in-person workshops that cover every aspect of our services from the core PDB data to programmatic access of PDBe-KB annotations. Within the on-demand training portfolio of EMBL-EBI, we provide an introductory course to PDBe at https://www.ebi.ac.uk/training/online/courses/exploring-pdb-entry/. We regularly host webinars and we make the recordings available at our YouTube channel: https://www.youtube.com/user/ProteinDataBank. These webinars cover topics such as the PDBe API, the 3D molecular viewer Mol*, and how to take advantage of the PDBe-KB aggregated views of proteins. We also have interactive tutorials covering PDBe search, API, and other PDBe tools at https://pdbeurope.github.io/api-webinars/index.html.

We follow a policy of open-sourcing the majority of our data pipelines and data visualization tools, generally in public repositories on the GitHub platform at PDBe (https://github.com/PDBeurope) and PDBe-KB (https://github.com/PDBe-KB), under Apache 2.0 license. The code repositories provide technical descriptions and instructions on how to install and use the tools. We provide technical documentation which includes demos, such as the ones described in previous sections related to PDBe Mol* and PDB ProtVista.

Feedback from the user community is very valuable, and drives all our work including improving the training materials, to reflect the most common use cases and address the most complicated aspects of our services.

7 | DISCUSSION

The archival of experimentally determined structures is a fundamental service both to structural biologists and the broader scientific community. The importance of protein structure archiving was highlighted by the recent, unprecedented advances in AI-based protein-structure prediction, as demonstrated by AlphaFold[8] and RoseTTaFold.[9] Neither these nor any other prediction methods would have been possible without the painstaking effort, ingenuity, and determination of generations of structural biologists or without the open, and transparent access to protein structure data from the Protein Data Bank.[1]

The advent of these advanced computationally predicted models disrupted several fields in the life sciences, from structure-based drug discovery and bioinformatics analysis to software development and structure determination efforts.[41–48] Researchers regularly use AlphaFold models to solve protein structures using their (sometimes old) experimental data.

Going forward, it becomes increasingly important to provide seamless integration of all structural data; experimental, predicted, and hybrid. Initiatives, such as the 3D-Beacons network, are increasingly important in this aspect, allowing researchers to find and access protein structures from several different data resources.

To address the changes in the field of structural biology, we are now focusing on improving the integration between PDBe and PDBe-KB web services, and investigating new approaches to aggregating structural data and their corresponding functional and biophysical annotations. We are exploring ways to identify and collate data around single ligand molecules observed in the PDB archive, as well as for macromolecular assemblies. The latter are often the de facto functional units in many biological systems, therefore they are at the crux of answering scientific questions even more so than individual protein structures.

The landscape of structural biology continues to change rapidly, and our services need to evolve to facilitate basic and translational research by supporting macromolecular structure depositors, specialist data resources, researchers and software developers alike.

AUTHOR CONTRIBUTIONS

Mihaly Varadi: Conceptualization (equal); formal analysis (equal); investigation (equal); project administration (lead); resources (equal); software (equal); supervision (lead); visualization (supporting); writing – original draft (lead). **Stephen Anyango:** Software (equal). **Sri Devan Appasamy:** Formal analysis (equal); investigation (equal); software (equal); visualization (equal). **David Armstrong:** Data curation (equal); writing – review and editing (supporting). **Marcus Bage:** Data curation (equal); formal analysis (equal); software (equal); visualization (equal); writing – review and editing (supporting). **John Berrisford:** Conceptualization (equal); data curation (equal); project administration (equal); software

(equal); supervision (equal); validation (equal); writing – review and editing (equal). **Preeti Choudhary:** Conceptualization (equal); data curation (equal); formal analysis (equal); software (equal); visualization (equal). **Damian Bertoni:** Software (equal); visualization (equal). **Mandar Deshpande:** Software (equal); visualization (equal). **Grisell Diaz Leines:** Software (equal); writing – review and editing (equal). **Joseph Ellaway:** Software (equal); visualization (equal). **Genevieve Evans:** Data curation (equal). **Romana Gaborova:** Data curation (equal); formal analysis (equal). **Deepti Gupta:** Data curation (equal); writing – review and editing (supporting). **Aleksandras Gutmanas:** Project administration (supporting); supervision (supporting). **Deborah Harrus:** Data curation (equal); project administration (supporting). **Gerard J. Kleywegt:** Writing – review and editing (equal). **Weslley Morellato Bueno:** Software (equal). **Nurul Nadzirin:** Software (equal). **Sreenath S Nair:** Software (lead); visualization (equal). **Lukas Pravda:** Formal analysis (equal); software (equal). **Marcelo Querino Lima Afonso:** Software (equal). **David Sehnal:** Software (equal); visualization (equal). **Ahsan Tanweer:** Software (equal). **James Tolchard:** Data curation (equal); software (supporting). **Charlotte Abrams:** Writing – review and editing (supporting). **Roisin Dunlop:** Writing – review and editing (supporting). **Sameer Velankar:** Conceptualization (equal); funding acquisition (lead); investigation (equal); resources (lead); software (equal); supervision (equal).

ACKNOWLEDGMENTS

We would like to acknowledge the work of all the researchers depositing their structures in the PDB archive, all the wwPDB partner sites and their staff, and all the specialist data resources and scientific software developers who make their valuable functional annotations available through the PDBe-KB consortium. Open Access funding enabled and organized by Projekt DEAL.

FUNDING INFORMATION

This work was supported by funding from the European Molecular Biology Laboratory-European Bioinformatics Institute, by the Wellcome Trust via grants PDBe-3D [104948/Z/14/Z], PDBe [218303/Z/19/Z], PDBe-KB [223739/Z/21/Z], and FunSites [221327/Z/20/Z]; by the Biotechnology and Biological Sciences Research Council (BBSRC) through grants SIFTS [BB/M011674/1], PDBe-India [BB/P025846/1], FunPDBe [BB/P024351/1], CATH-BBR [BB/R015201/1], Structure Coverage [BB/S017135/1], 3D-Gateway [BB/S020071/1], BioChemGRAPH [BB/T01959X/1], PDB-FACT [BB/V004247/1], FunCLAN [BB/V016113/1], 3D-Proteomics [BB/V018779/1], and PDBe-RCSB [BB/W017970/1].

DATA AVAILABILITY STATEMENT

All of the data provided is freely available for both academic and commercial use under Creative Commons Attribution 4.0 (CC-BY 4.0) license terms.

ORCID

Mihaly Varadi https://orcid.org/0000-0002-3687-0839

REFERENCES

1. wwPDB consortium, Burley SK, Berman HM, et al. Protein Data Bank: The single global archive for 3D macromolecular structure data. Nucleic Acids Res. 2019;47:D520–D528.
2. Armstrong DR, Berrisford JM, Conroy MJ, et al. PDBe: Improved findability of macromolecular structure data in the PDB. Nucleic Acids Res. 2019;48(D1):D335–D343.
3. 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;433:166704.
4. Kinjo AR, Bekker G-J, Suzuki H, et al. Protein Data Bank Japan (PDBj): Updated user interfaces, resource description framework, analysis tools for large structures. Nucleic Acids Res. 2017;45:D282–D288.
5. Romero PR, Kobayashi N, Wedell JR, et al. BioMagResBank (BMRB) as a resource for structural biology. Methods Mol Biol. 2020;2112:187–218.
6. Abbott S, Iudin A, Korir PK, Somasundharam S, Patwardhan A. EMDB web resources. Curr Protoc Bioinformatics. 2018;61:5.10.1–5.10.12.
7. 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 Publ Protein Soc. 2018;27:316–330.
8. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–589.
9. Baek M, DiMaio F, Anishchenko I, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science. 2021;373:871–876.
10. Gerstein M. Integrative database analysis in structural genomics. Nat Struct Biol. 2000;7(Suppl):960–963.
11. PDBe-KB consortium. PDBe-KB: Collaboratively defining the biological context of structural data. Nucleic Acids Res. 2022;50:D534–D542.
12. Mitsopoulos C, Di Micco P, Fernandez EV, et al. canSAR: Update to the cancer translational research and drug discovery knowledgebase. Nucleic Acids Res. 2021;49:D1074–D1082.
13. Cilia E, Pancsa R, Tompa P, Lenaerts T, Vranken WF. The DynaMine webserver: Predicting protein dynamics from sequence. Nucleic Acids Res. 2014;42:W264–W270.
14. Wass MN, Kelley LA, Sternberg MJE. 3DLigandSite: Predicting ligand-binding sites using similar structures. Nucleic Acids Res. 2010;38:W469–W473.
15. Pravda L, Sehnal D, Svobodová Vařeková R, et al. ChannelsDB: Database of biomacromolecular tunnels and pores. Nucleic Acids Res. 2018;46:D399–D405.
16. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data. 2016;3:160018.

17. Orengo C, Velankar S, Wodak S, et al. A community proposal to integrate structural bioinformatics activities in ELIXIR (3D-Bioinfo Community). F1000Research. 2020;9:278.
18. Nair S, Váradi M, Nadzirin N, et al. PDBe aggregated API: Programmatic access to an integrative knowledge graph of molecular structure data Gorodkin J, editor. Bioinformatics. 2021;37(21):3950–3952.
19. PDBe-KB consortium, Varadi M, Berrisford J, et al. PDBe-KB: A community-driven resource for structural and functional annotations. Nucleic Acids Res. 2020;48:D344–D353.
20. 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:D439–D444.
21. Sehnal 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.
22. Deshpande M, Varadi M, Paysan-Lafosse T, et al. PDB ProtVista: A reusable and open-source sequence feature viewer. BioRxiv. 2022.
23. 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:D482–D489.
24. Young JY, Westbrook JD, Feng Z, et al. (2017) OneDep: Unified wwPDB system for deposition, biocuration, and validation of macromolecular structures in the PDB archive. Struct Lond Engl. 1993;25:536–545.
25. Khanna T, Hanna G, Sternberg MJE, David A. Missense3D-DB web catalogue: An atom-based analysis and repository of 4M human protein-coding genetic variants. Hum Genet. 2021;140:805–812.
26. Delgado J, Radusky LG, Cianferoni D, Serrano L. FoldX 5.0: Working with RNA, small molecules and a new graphical interface Valencia A, editor. Bioinformatics. 2019;35:4168–4169.
27. Blum M, Chang H-Y, Chuguransky S, et al. The InterPro protein families and domains database: 20 years on. Nucleic Acids Res. 2021;49:D344–D354.
28. Sillitoe I, Dawson N, Lewis TE, et al. CATH: Expanding the horizons of structure-based functional annotations for genome sequences. Nucleic Acids Res. 2019;47:D280–D284.
29. Mistry J, Chuguransky S, Williams L, et al. Pfam: The protein families database in 2021. Nucleic Acids Res. 2021;49:D412–D419.
30. Cunningham F, Allen JE, Allen J, et al. Ensembl 2022. Nucleic Acids Res. 2022;50:D988–D995.
31. 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:D376–D382.
32. Feolo M, Helmberg W, Sherry S, Maglott DR. NCBI genetic resources supporting immunogenetic research. Rev Immunogenet. 2000;2:461–467.
33. UniProt Consortium. UniProt: The universal protein knowledgebase in 2021. Nucleic Acids Res. 2021;49:D480–D489.
34. Stelzer G, Rosen N, Plaschkes I, et al. The GeneCards suite: From gene data mining to disease genome sequence analyses. Curr Protoc Bioinforma. 2016;54:1.30.1–1.30.33.
35. Laskowski RA, Jabłońska J, Pravda L, Vařeková RS, Thornton JM. PDBsum: Structural summaries of PDB entries. Protein Sci Publ Protein Soc. 2018;27:129–134.
36. Sweeney BA, Hoksza D, Nawrocki EP, et al. R2DT is a framework for predicting and visualising RNA secondary structure using templates. Nat Commun. 2021;12:3494.
37. Sarver M, Zirbel CL, Stombaugh J, Mokdad A, Leontis NB. FR3D: Finding local and composite recurrent structural motifs in RNA 3D structures. J Math Biol. 2008;56:215–252.
38. Mukhopadhyay A, Borkakoti N, Pravda L, Tyzack JD, Thornton JM, Velankar S. Finding enzyme cofactors in Protein Data Bank Wren J, editor. Bioinformatics. 2019;35:3510–3511.
39. Kleinjung J, Fraternali F. POPSCOMP: An automated interaction analysis of biomolecular complexes. Nucleic Acids Res. 2005;33:W342–W346.
40. Tiwari SP, Fuglebakk E, Hollup SM, et al. WEBnm@ v2.0: Web server and services for comparing protein flexibility. BMC Bioinformatics. 2014;15(427):427.
41. Fontana P, Dong Y, Pi X, et al. Structure of cytoplasmic ring of nuclear pore complex by integrative cryo-EM and AlphaFold. Science. 2022;376:eabm9326.
42. Mosalaganti S, Obarska-Kosinska A, Siggel M, et al. AI-based structure prediction empowers integrative structural analysis of human nuclear pores. Science. 2022;376:eabm9506.
43. Tian R, Li Y, Wang X, et al. A pharmacoinformatics analysis of artemisinin targets and de novo design of hits for treating ulcerative colitis. Front Pharmacol. 2022;13:843043.
44. Binder JL, Berendzen J, Stevens AO, et al. AlphaFold illuminates half of the dark human proteins. Curr Opin Struct Biol. 2022;74:102372.
45. Cai SW, Zinder JC, Svetlov V, et al. Cryo-EM structure of the human CST-Polα/primase complex in a recruitment state. Nat Struct Mol Biol. 2022;29(8):813–819.
46. Collar AL, Linville AC, Core SB, Frietze KM. Epitope-based vaccines against the chlamydia trachomatis major outer membrane protein variable domain 4 elicit protection in mice. Vaccine. 2022;10:875.
47. Wehrspan ZJ, McDonnell RT, Elcock AH. Identification of iron-sulfur (Fe-S) cluster and zinc (Zn) binding sites within proteomes predicted by DeepMind's AlphaFold2 program dramatically expands the metalloproteome. J Mol Biol. 2022;434:167377.
48. Bludau I, Willems S, Zeng W-F, et al. The structural context of posttranslational modifications at a proteome-wide scale. PLoS Biol. 2022;20:e3001636.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Varadi M, Anyango S, Appasamy SD, Armstrong D, Bage M, Berrisford J, et al. PDBe and PDBe-KB: Providing high-quality, up-to-date and integrated resources of macromolecular structures to support basic and applied research and education. Protein Science. 2022;31(10):e4439. https://doi.org/10.1002/pro.4439