*Published online 6 November 2019* *Nucleic Acids Research, 2020, Vol. 48, Database issue* ***D335–D343***
*doi: 10.1093/nar/gkz990*

# PDBe: improved findability of macromolecular structure data in the PDB

**David R. Armstrong[1], John M. Berrisford[1], Matthew J. Conroy[1], Aleksandras Gutmanas[1], Stephen Anyango[1], Preeti Choudhary[1], Alice R. Clark[1], Jose M. Dana[1], Mandar Deshpande[1], Roisin Dunlop[1], Paul Gane[1], Romana Gáborová[2], Deepti Gupta[1], Pauline Haslam[1], Jaroslav Koča[2], Lora Mak[1], Saqib Mir[1], Abhik Mukhopadhyay[1], Nurul Nadzirin[1], Sreenath Nair[1], Typhaine Paysan-Lafosse[1,3], Lukas Pravda[1], David Sehnal[2], Osman Salih[4], Oliver Smart[1], James Tolchard[1], Mihaly Varadi[1], Radka Svobodova-Vařeková[2], Hossam Zaki[1], Gerard J. Kleywegt[1,4] and Sameer Velankar[1,*]**

[1]Protein Data Bank in Europe, European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK, [2]CEITEC - Central European Institute of Technology, Masaryk University Brno, Kamenice 5, 625 00 Brno-Bohunice, Czech Republic, [3]InterPro, European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK and [4]Electron Microscopy Data Bank (EMDB), European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received September 17, 2019; Revised October 11, 2019; Editorial Decision October 11, 2019; Accepted October 25, 2019

## ABSTRACT

**The Protein Data Bank in Europe (PDBe), a founding member of the Worldwide Protein Data Bank (wwPDB), actively participates in the deposition, curation, validation, archiving and dissemination of macromolecular structure data. PDBe supports diverse research communities in their use of macromolecular structures by enriching the PDB data and by providing advanced tools and services for effective data access, visualization and analysis. This paper details the enrichment of data at PDBe, including mapping of RNA structures to Rfam, and identification of molecules that act as cofactors. PDBe has developed an advanced search facility with ∼100 data categories and sequence searches. New features have been included in the LiteMol viewer at PDBe, with updated visualization of carbohydrates and nucleic acids. Small molecules are now mapped more extensively to external databases and their visual representation has been enhanced. These advances help users to more easily find and interpret macromolecular structure data in order to solve scientific problems.**

## INTRODUCTION

Protein Data Bank in Europe (PDBe; pdbe.org) is a founding member of the Worldwide Protein Data Bank (wwPDB; wwpdb.org) (1), the international consortium responsible for the management of the Protein Data Bank (PDB) (2), the single, global archive for experimentally determined three dimensional (3D) structures of biological macromolecules. The other wwPDB partners are Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB; rcsb.org) (3), Protein Data Bank Japan (PDBj; pdbj.org) (4) and Biological Magnetic Resonance Bank (BMRB; bmrb.wisc.edu) (5). Together, the four partners fully cooperate in the areas of deposition, curation, validation and dissemination of macromolecular structure data, guided by the FAIR principles of administering data resources, which ensure data is Findable, Accessible, Interoperable and Reusable (2,6).

Since 2014, the processing of all PDB depositions is managed through the OneDep system (7). PDBe is responsible for the processing of OneDep depositions from Euro-

[*]To whom correspondence should be addressed. Tel: +44 1223 49 4646; Email: sameer@ebi.ac.uk
Present addresses:
Alice R. Clark, Faculty of Science and Engineering, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY, UK.
Saqib Mir, GSK, Gunnels Wood Road, Stevenage, Herts SG1 2NY, UK.
Abhik Mukhopadhyay, The Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, UK.
Oliver Smart, School of Life Sciences, Anglia Ruskin University, East Road, Cambridge CB1 1PT, UK.
Hossam Zaki, Department of Molecular and Cellular Biology and Biochemistry, Brown University, Providence, RI, USA.

pean and African institutions, totalling over 4,000 PDB entries in 2018, which equates to 34% of the total PDB depositions that year (wwpdb.org/stats/deposition). PDBe and the wwPDB partners collaborate closely with the Electron Microscopy Data Bank (EMDB) (8), the archive for electron microscopy (EM) electric potential maps, which are also deposited via OneDep. While all wwPDB partners collaborate on processing deposited data in the PDB archive, thereby creating an authoritative source of macromolecular structure data, each partner has its own website and tools to support users in accessing the data.

In 2019, the PDB reached the milestone of 150 000 structures. This wealth of information is used by researchers in fields ranging from drug discovery to protein engineering. To support these diverse user communities, and to present macromolecular structure information in other biologically relevant contexts, PDBe enriches its data through collaborations with resources that specialize in other areas of bioinformatics. For example, the SIFTS project (9) brings together PDB structures with protein sequence data and annotations from UniProtKB (10), Pfam (11) and InterPro (12), structure domains from CATH (13) and SCOP (14), and, more recently, started incorporating genetic variation data from Ensembl (15) and putative homology groups from Homologene (16). PDBe provides online data access, analysis and visualization tools for PDB data through its website, FTP/RSYNC file download, Application Programming Interface (REST API) and PDB Coordinate/Electron-density servers. PDBe services were accessed from 2.2 million unique IP addresses in 2018, with 1.1 million users accessing the PDBe REST API in the same time period, based on statistics collected at the European Bioinformatics Institute (EMBL-EBI) (17).

A new resource, Protein Data Bank in Europe-Knowledge Base (PDBe-KB, pdbe-kb.org), has been developed by PDBe in collaboration with the structural bioinformatics community. The PDBe-KB resource aggregates macromolecular structure data in a series of dedicated views centred on commonly used biological objects (e.g. UniProtKB accessions) rather than on PDB entries. These views incorporate additional biological context and functional information and are discussed in a separate publication in this issue (18).

This paper describes data enrichment efforts and updates to the PDBe infrastructure, website and other services since our previous publication (19).

## AUTOMATING THE PDBE RELEASE PROCESS

The wwPDB partners release PDB archive data through their respective websites and FTP sites at 00:00 UTC each Wednesday. PDBe have implemented a robust release process using Apache Airflow (airflow.apache.org), a free and open source software for workflow and dependency management. Airflow is used to manage job dependencies in the weekly release process, supporting automation and improving efficiency. The Airflow software enables thorough monitoring of the large number of processes that are involved in the release of data at PDBe, providing clear notification of any errors. The Airflow system includes a sophisticated graphical interface for review of the overall release process, highlighting any failed processes and allowing easy restart of individual tasks. This new infrastructure has significantly streamlined the complex process of releasing data at PDBe, ensuring timely availability of data and allowing easy integration of additional software into the PDBe release process, some of which is described below.

## IMPROVEMENTS TO SMALL-MOLECULE DEPICTION

Previous efforts to improve the PDBe website focused on macromolecules, and the area of small molecules was identified as a future improvement (19). To support clear and effective web-delivery of information pertaining to small molecules in the PDB, we have now redesigned the underlying process for handling small molecules present in the wwPDB chemical component dictionary (CCD) (20), deriving added-value information, improving two-dimensional (2D) depictions and adding cross-references to cheminformatics resources. The redesigned process is implemented as a freely available python package (pdbeccdutils; https://gitlab.ebi.ac.uk/pdbe/ccdutils) that builds on the RDKit software (http://rdkit.org) and its data model. The package provides functionality for reading PDBx/mmCIF-formatted files (compound definitions from the CCD) as well as utility functions to compute physicochemical properties, scaffolds, fragments and to draw ‘collision-free’ images using a series of templates (Figure 1). This process generates 2D coordinates and schematic depictions for the vast majority of small molecules in the PDB (e.g. only 447 out of 29,318 compounds do not have a collision-free image as of September 2019). The images have been added to the PDBe search interface and PDB entry pages (see below). Where possible, we have added cross-references to the database entries for the same compound in other resources, including ChEMBL (21), ChEBI (22), ZINC (23) and DrugBank (24).

## DATA ENRICHMENT

Macromolecular structures provide insight into the function of biological systems, but to capitalize fully on the value of the data inherently present in the PDB, it must be presented in the context of related biological and chemical information, particularly for scientists whose expertise may be outside of structural biology. PDBe works closely with other teams at EMBL-EBI and beyond, to enrich the data in the PDB archive with additional biological and chemical information. PDBe have been involved in a number of projects that have added value to PDB entries where this information was previously absent or incomplete (Table 1). Unlike the majority of proteins, for which UniProtKB is a universally accepted reference resource for sequence information, RNA molecules in PDB entries are not mapped to any sequence database that could provide consistent molecule naming using conventions adopted by the community. For example, it is difficult to find all transfer RNA structures in the PDB. To address this issue, PDBe, in collaboration with the Rfam team at EMBL-EBI, have mapped RNA sequences to RNA families provided by the Rfam resource (25). Functional annotation of small molecules in the PDB is lacking, and to start to address this issue, PDBe has collaborated with the Thornton group to identify cofac-

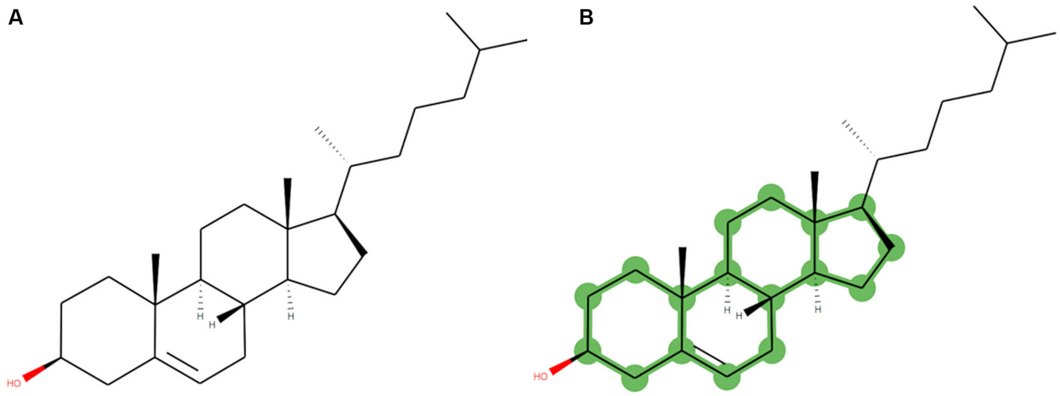


**Figure 1.** (**A**) Image of cholesterol (PDB three-letter code: CLR) generated from the PDB Chemical Component Dictionary, using the pdbeccdutils process. (**B**) Chemical scaffold (in green) of cholesterol, identified and highlighted using the pdbeccdutils process.

tor molecules in PDB entries (26). To address the discrepancy between the release cycles of the PDB and the Pfam and CATH resources, PDBe have integrated preliminary assignment of protein families (Pfam) and structural domains (CATHb (13)) such that these domains can be identified prior to formal assignment by Pfam and CATH (9). These data-enrichment projects are summarized in Table 1.

All the data from the above enrichment activities are made available via the PDBe REST API and are integrated into PDBe's search system.

## DATA QUALITY AND INTEGRITY

To ensure findability of data, users must be able to search with familiar terms and retrieve accurate and up-to-date results. One of the biggest barriers to findability is poor data quality and integrity. Therefore, to improve findability, PDBe strives to make the data as consistent as possible. As described previously (19), PDBe's 'clean-mmCIF' process updates PDBx/mmCIF files from the PDB archive with remapped enumerations and additional information, yielding more consistent, standardized metadata, without altering core PDB information, such as atomic coordinates and experimental data.

Since originally reported (19), a further 13 categories are now updated through the clean-mmCIF process, including information on synchrotron site and beamline for X-ray crystallography entries, and software information for all experimental techniques. A key aim of this process is to identify where data is inconsistent so that the core PDB archive can be improved. Many issues identified by this process have since been converted into checks or enumerations in the wwPDB OneDep system (7). Of the 25 categories updated in the clean-mmCIF process, nine have now been updated archive-wide, including experimental method and software classification, with a further five categories standardized for newly deposited entries, e.g. names of software packages used in EM and NMR.

## SEQUENCE CLUSTERING

In many cases, the PDB contains more than one structure of the same protein or domain, with the current over 155 000 PDB entries representing only 48 000 unique UniProtKB entries (https://www.ebi.ac.uk/pdbe/docs/sifts/statistics.html). It is also common to find structures of homologous proteins. In many applications, this redundancy can be reduced, especially where users of the PDB archive wish to use a single representative structure for their protein of interest. Identification of redundant protein structures can be achieved by sequence comparison, as high sequence similarity correlates with structural similarity (29). PDBe have implemented a sequence-clustering process, using the MMseqs software suite (30), that groups structures based on sequence identity at various thresholds: 100%, 95%, 90%, 70%, 50%, 40% and 30%. PDB users looking for only individual representative structures at a certain level of sequence identity can now do so by applying a sequence-clustering filter in either the PDBe REST API or the PDBe search interface. This results in only individual representative structures being returned for each unique sequence cluster at the specified identity threshold, thus reducing the bulk of search results.

## IMPROVED QUERY SYSTEM

To ensure that data can be accessed effectively by a variety of user groups, it is vital that data of interest can be easily discovered in a format that is simple to interpret. In 2015, PDBe introduced a new query system (31), providing innovative and intuitive search options for accessing PDB data. More recent developments have focused on further improving the findability of macromolecular structures at PDBe, both by improving the data and by building on the existing search functionality.

### Improved autocomplete

The autocomplete functionality introduced previously at PDBe (19) has streamlined the querying of PDB data, while still presenting a large number of search terms. As data has been enriched with new mappings (Table 1), they have been included as new search items. The addition of more search items often expanded the autocomplete suggestions beyond a single page view, thereby reducing its usability. To mitigate this issue, the categories have been grouped into fewer sections. For example, rather than having separate groups for

**Table 1.** Summary of data enrichment collaborations

| Enriched data | Enrichment process | Main outcomes |
|---|---|---|
| Mapping to Rfam | • Data incorporated into the PDBe database and displayed on relevant PDBe entry pages.<br>• Data integrated into the PDBe REST API, search and in visualization, via web components.<br>• Links to Rfam resource on PDBe entry pages.<br>• Rfam data resource was updated where missing mapping information was identified. | • Assignment of over 5000 RNA chains to 98 Rfam families in >1500 structures in the PDB.<br>• Improved findability of RNA structures.<br>• Easier identification of RNA function in PDB entries. |
| Identification of cofactors | • PDBe worked with the Thornton group at EMBL-EBI to set up a process to identify cofactor and cofactor-like molecules in the PDB.<br>• Cofactors grouped by class.<br>• The process verifies that such molecules are bound to enzymes associated with the correct cofactor class.<br>• Data made available via PDBe REST API, on PDBe entry pages and search. | • Identified 417 unique cofactor-like small molecules.<br>• Annotation of over 78 000 PDB entries containing enzymes, 12 000 contain cofactor or cofactor-like molecules, representing over 1500 different Enzyme Commission numbers. |
| Preliminary Pfam domain assignments | • Collaboration with Pfam team to implement provisional domain-assignment process at PDBe.<br>• Pfam domains assigned in newly released PDB entries, in advance of Pfam database release.<br>• PDBe integrates this data, enabling Pfam domain assignments for PDB entries immediately after release. | • As of September 2019, over 16 000 PDB entries had Pfam domains assigned which are not in the official latest Pfam release (V32.0, September 2018).<br>• This includes over 3100 distinct Pfam domains. |
| Preliminary CATH domain assignments (CATHb) | • CATHb data released by CATH team provides preliminary CATH structural domain assignment for PDB structures on a weekly basis.<br>• PDBe integrates this data, enabling structure-domain assignments for PDB entries immediately after release. | • As of August 2019, around 30 000 new entries have CATH domains assigned which are not in the official full CATH release (V4.2, September 2017). |
| Standardized information on crystallographic cells dimensions (NIGGLI) | • Standardization of cell dimensions using Niggli reduction (27).<br>• Standardized cell dimensions made available through PDBe search API. | • Standardized cell dimensions in PDBe's search used by Phaser (28). |

sequence and structural domains, these have been grouped into a common domains group, including Pfam, InterPro, CATH, SCOP and Rfam, simplifying the display of the autocomplete suggestions.

### Advanced search

PDBe has increased the number of data categories available for searching to 122, enabling users to create richer and more customized queries. All search categories are now available in a new 'Advanced search' form (pdbe.org/search). They are grouped into sections containing related terms, with an autocomplete option also available to facilitate identification of relevant search fields (Figure 2). Furthermore, autocomplete has been implemented within each search category to simplify identification of relevant search terms and improve findability. Thirty new experimental data items have been added to the search related to EM and NMR structure determination methodologies. The 'resolution revolution' in EM (32) has significantly raised the interest in this method and these additional categories help users to identify structures determined with particular experimental procedures. The accuracy of the data in these additional data categories has improved by the development of OneDep and enumerations in the PDBx/mmCIF dictionary, in addition to clean-up efforts by wwPDB biocurators, ensuring consistent and accurate capture of data in PDB depositions (2,7,33).

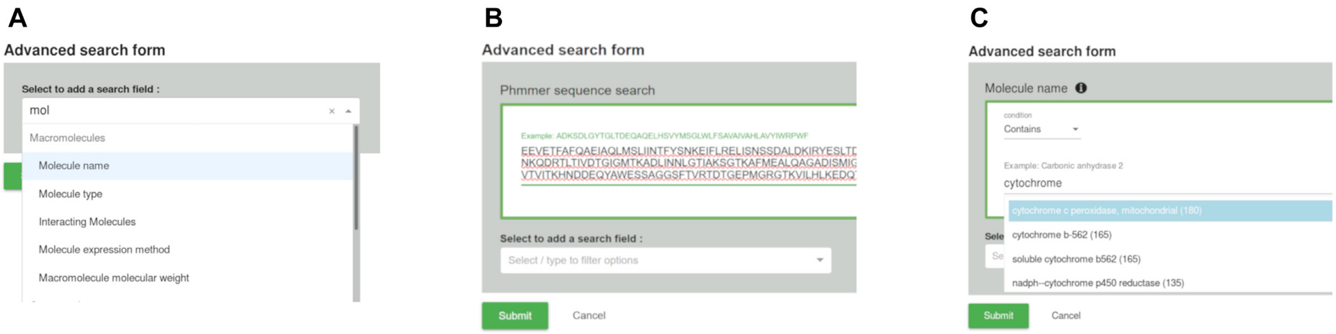
In search fields containing numeric and date terms, the form allows users to set specific ranges to fine-tune search queries. Each text-based term can be searched with options including 'contains' and 'equal to', and multiple search terms can be linked with Boolean operators ('AND', 'OR', and 'NOT').

Whether the query is submitted via the simple or advanced search interface, the user is offered further options to download the search results or filter them with facets (Figure 3). Direct links for 3D visualization and additional downloads for each search result allow faster access to the PDB structures and further information. There are links to the relevant protein-centric aggregated views served by PDBe-KB (pdbe-kb.org/proteins) for each protein, giving direct access to the wealth of additional information provided by that resource (18).

The 'Macromolecules' tab in the PDBe search results page lets users see their results grouped by individual molecule name, presenting a more succinct view of the available data. This feature was limited to grouping by protein name, but now includes grouping by DNA and RNA molecule names.

### Sequence search

There are numerous services available to search for proteins based on sequence similarity. Tools such as BLAST (34), FASTA (35), and HMMER (36) enable users to eas-



Advanced search form

Select to add a search field :
mol
Macromolecules
Molecule name
Molecule type
Interacting Molecules
Molecule expression method
Macromolecule molecular weight

Advanced search form

Phmmer sequence search

Example: ADKSDLGYTGLTDEQAQELHSVYMSGLWLFSAVAIVAHLAVYIWRPWF

EEVETFAFQAEIAQLMSLIINTFYSNKEIFLRELISNSSDALDKIRYESLTD
NKQDRTLTIVDTGIGMTKADLINNLGTIAKSGTKAFMEALQAGADISMIG
VTVITKHNDDEQYAWESSAGGSFTVRTDTGEPMGRGTKVILHLKEDQ

Select to add a search field :
Select / type to filter options

Submit Cancel

Advanced search form

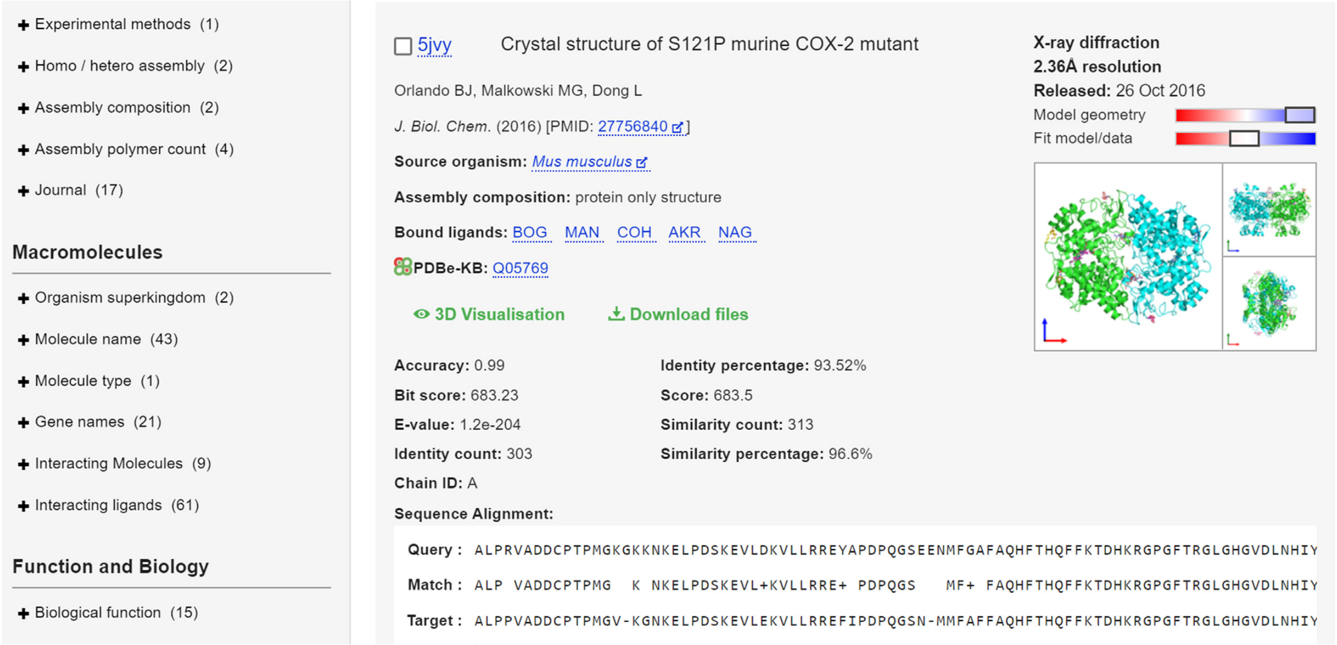
Molecule name

condition
Contains

Example: Carbonic anhydrase 2
cytochrome
cytochrome c peroxidase, mitochondrial (180)
cytochrome b-562 (165)
soluble cytochrome b562 (165)
nadph--cytochrome p450 reductase (135)

Submit Cancel

**Figure 2.** New features added to the PDBe search. (**A**) The advanced search form supports 122 fields in queries. (**B**) Sequence searching using HMMER. (**C**) Autocomplete option to find relevant search terms for each search field.



+ Experimental methods (1)
+ Homo / hetero assembly (2)
+ Assembly composition (2)
+ Assembly polymer count (4)
+ Journal (17)

Macromolecules

+ Organism superkingdom (2)
+ Molecule name (43)
+ Molecule type (1)
+ Gene names (21)
+ Interacting Molecules (9)
+ Interacting ligands (61)

Function and Biology

+ Biological function (15)

□ 5jvy Crystal structure of S121P murine COX-2 mutant

Orlando BJ, Malkowski MG, Dong L

*J. Biol. Chem.* (2016) [PMID: 27756840]

**Source organism:** *Mus musculus*

**Assembly composition:** protein only structure

**Bound ligands:** BOG MAN COH AKR NAG

**PDBe-KB:** Q05769

3D Visualisation Download files

**X-ray diffraction**
**2.36Å resolution**
**Released:** 26 Oct 2016
Model geometry
Fit model/data

**Accuracy:** 0.99
**Bit score:** 683.23
**E-value:** 1.2e-204
**Identity count:** 303
**Identity percentage:** 93.52%
**Score:** 683.5
**Similarity count:** 313
**Similarity percentage:** 96.6%
**Chain ID:** A

**Sequence Alignment:**

```
Query :  ALPRVADDCPTPMGKGKKNKELPDSKEVLDKVLLRREYAPDPQGSEENMFGAFAQHFTHQFFKTDHKRGPGFTRGLGHGVDLNHIY
Match :  ALP VADDCPTPMG  K NKELPDSKEVL+KVLLRRE+ PDPQGS   MF+ FAQHFTHQFFKTDHKRGPGFTRGLGHGVDLNHIY
Target : ALPPVADDCPTPMGV-KGNKELPDSKEVLEKVLLRREFIPDPQGSN-MMFAFFAQHFTHQFFKTDHKRGPGFTRGLGHGVDLNHIY
```

**Figure 3.** New features in presenting search results, including: number of facets (left) increased to 39 to improve filtering of results, access to 3D visualization and file downloads directly from the search results (green text); alignment and statistics provided for sequence searches, links to PDBe-KB aggregated views.

ily query very large databases of protein sequences and obtain meaningful results rapidly. Though these tools may provide links to PDB entries, they are usually presented as a list of accession codes and do not provide additional context and validation information for the search results. PDBe have integrated sequence searching with FASTA into the advanced search form, in addition to querying with the EMBL-EBI pHMMER service (www.ebi.ac.uk/Tools/hmmer/search/phmmer) which carries out fast and sensitive homology searches using protein sequence profiles. Results are initially sorted by similarity score and can be further refined with PDBe search facets or sorted by other criteria such as validation score or resolution.

## ADDITIONAL DATA FOR BULK ACCESS

As a wwPDB member, PDBe hosts copies of both the PDB FTP site (ftp://ftp.ebi.ac.uk/pub/databases/pdb) and the PDB versioned FTP site (http://ftp.ebi.ac.uk/pub/databases/pdb_versioned/), providing FTP/RSYNC access to PDB data. Enriched data unique to PDBe is provided by the PDBe REST API (pdbe.org/api), which powers both the PDBe search and its entry pages, and is distributed through an additional FTP site (ftp://ftp.ebi.ac.uk/pub/databases/msd). The data available from both the REST API and PDBe FTP site are updated to reflect changes and improvements to the PDBe website, with recent updates detailed in Table 2.

## DEVELOPMENTS TO VISUALIZATION IN LITEMOL

The LiteMol viewer (39), has enabled fast visualization of PDB structures and related functional data at PDBe. Several features that further improve the representation of PDB structures are described here. Visualization of carbohydrate molecules in LiteMol has been updated by integration of the 3D implementation of the Symbol Nomenclature for Glycans (SNFG) system of carbohydrate representation (40,41)

**Table 2.** Additions and updates to the PDBe REST API and FTP sites

| | |
|---|---|
| PDBe REST API (pdbe.org/api) | • New endpoints for protein information.<br>○ CATHb structure domains.<br>○ Pre-release Pfam sequence domains (via HMMER).<br>○ Ensembl identifiers.<br>• New endpoints for nucleic-acid information.<br>○ RNA molecules mapped to Rfam.<br>• New endpoints for information on small molecules in the Chemical Component Dictionary (CCD).<br>○ Cofactor information.<br>○ Cross-references to DrugBank (including UniProtKB accessions for drug targets), ChEBI (22), ChEMBL (21) and CSD (37). |
| PDBe enriched FTP (ftp://ftp.ebi.ac.uk/pub/databases/msd) | • mmCIF-format assembly files added (ftp://ftp.ebi.ac.uk/pub/databases/msd/assemblies).<br>○ These contain full assemblies for a given structure, along with rotation/translation operators required to generate the assemblies from the PDB entry file.<br>○ Includes XML files summarizing the composition of the assembly files, relating each chain in the assembly to an entity in the PDB entry.<br>• New FTP area for updated chemical-component process (ftp://ftp.ebi.ac.uk/pub/databases/msd/pdbechem_v2).<br>○ CCD mmCIF files with additional cross-references, mapped through the UniChem pipeline (38), to ChEMBL, DrugBank, KEGG, ChEBI, ZINC, Pubchem, NMRShiftDB, BindingDb, MetaboLights, BRENDA and Rhea.<br>○ Updated image files for molecules in CCD.<br>○ Coordinate data in various formats – including PDB and SDF. |

(Figure 4A). Fifty of the 71 monosaccharides defined by SNFG are present in the PDB in at least one anomeric form and are shown using 3D-SNFG symbols in LiteMol on the PDBe pages. This has simplified visualization and identification of carbohydrates, especially of complex and branched glycan chains.

From user testing it was clear that in the original implementation of LiteMol, the thin strands representing nucleic-acid backbones did not highlight these molecules sufficiently well. It is important that nucleic-acid chains be easily distinguishable from protein chains, and therefore a new visualization was developed for nucleic acids in LiteMol that addresses these problems and allows straightforward location of each base position (Figure 4B).

PDBe has worked with UniProt, Complex Portal (42), Reactome (43) and Ensembl (15) to support integration of LiteMol on their respective websites, placing the macromolecular structure within a context that is familiar to users of these resources. For example, Ensembl have implemented the display of variation data and exon boundaries directly on the macromolecular structure, helping researchers to more easily relate parts of the structure to genomic features.

## ACCESS TO EXPERIMENTAL DATA

The experimental data currently collected by wwPDB in OneDep (i.e., structure factors, chemical shifts and restraints, and electric potential maps) is a reduced (processed) representation of the raw experimental data. This enables computation of validation statistics, but to fully replicate the experiment, users require access to the raw experimental data which may have been deposited to specialized archives. PDBe previously introduced the Raw Experimental Data web component (19) to provide a summary and links to the raw experimental data from the Electron Microscopy Public Image Archive (EMPIAR) (44), SBGrid Data Bank (45) and Integrated Resource for Reproducibility in Macromolecular Crystallography (IRRMC) (46). Links to raw Free Induction Decay (FID) data held at BMRB for NMR entries have been added. The component relies on the respective archive to provide an API allowing queries based on PDB codes and serving basic information about, and a URL for, the raw experimental dataset. Independently, the OneDep system allows depositors to include a DOI to raw experimental data held in these or other archives. These DOIs have been standardized in OneDep and the component has been adapted to provide links to experimental datasets referenced by them.

## FURTHER VISUALIZATION IMPROVEMENTS

Improving the usability of PDBe services helps users in interpreting and exploiting the information provided, and a number of recent changes at PDBe have focused on improving the visualization experience. A specifically tailored mobile view has been developed for the PDBe search pages, which maintains the information provided on the desktop version but improves the usability on mobile devices. The use of web components across all PDBe pages gives an inherently more adaptable visualization, which has been utilized in the implementation of mobile-specific views.

The ProtVista component (47), developed in collaboration with UniProt and InterPro, enables a graphical visualization of molecular features corresponding to a protein's primary sequence. PDBe will in the near future include the ProtVista component on the PDB entry pages, where it will be interactively linked to a 2D topology component and the LiteMol viewer. The ProtVista implementation on these pages will display information related to the macromolecule in the PDB structure, including related functional data to provide further biological context.

## TRAINING AND OUTREACH

PDBe continues to run a broad programme of training and outreach activities to ensure that users can more easily utilize the tools and services to understand macromolecular structure data. These activities include webinars, workshops and seminars for researchers in the fields of chemistry, biology and bioinformatics. Since the beginning of 2018, PDBe have delivered 28 workshops in 13 different

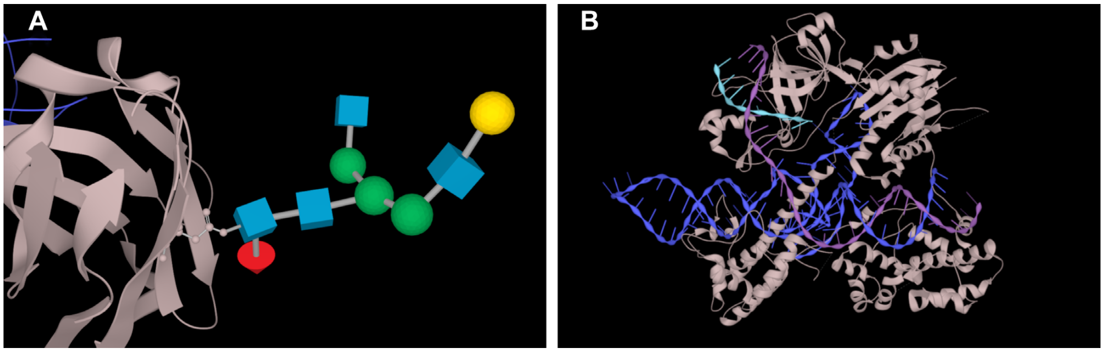


**Figure 4.** Improved visualization of nucleic acids and carbohydrates in LiteMol. (**A**) PDB entry 5ezi (48), the structure of a malaria antigen in complex with an antibody, focused on the branched glycan in the structure. The glycan is shown using the 3D-SNFG representation. (**B**) PDB entry 5x2g (49), structure of a CRISPR-Cas9 complex of protein with nucleic acids, using the new visualization for RNA and DNA.

countries on 3 continents. The use of web-based components on the PDBe website enables easy delivery of workshops to a range of users without additional software requirements for the host or attendees. In addition, where users cannot attend PDBe workshops, training materials are accessible through the EMBL-EBI Train online portal (www.ebi.ac.uk/training/online) and the PDBe YouTube channel (youtube.com/proteindatabank). There have been 10 new videos added to the PDBe YouTube channel since the beginning of 2018, with videos on the PDBe YouTube channel viewed more than 15,000 times over the same time period.

The PDB Art project (pdbe.org/art) is a public engagement effort in collaboration with art societies and local schools. It engages students and the general public with macromolecular structure data through the medium of art. Students learn about protein structures in the PDB and the biology behind them, create artworks inspired by structures of their choice, and exhibit these works in a public art exhibition in Cambridge, UK. This project is now in its fifth year and has engaged hundreds of students and reached thousands of members of the general public. This project has been extremely valuable in communicating the importance of macromolecular structures with the general public and we encourage more scientists to use this approach, and to contact us for more information.

## FUTURE DEVELOPMENTS

PDBe will continue to improve the quality of data, aiming to improve findability through both the PDBe search facility and its REST API. This will include the integration of more data from the PDBe-KB resource into the search and entry pages, enriching the data further and improving navigation to relevant PDB entry-specific pages and the aggregated views offered by PDBe-KB. There will also be a continued effort to better classify and enrich data related to nucleic acids, as well as improved naming and classification of antibodies to help users more easily locate data for these molecules. In collaboration with Complex Portal (42), PDBe is working to classify known macromolecular complexes, which will enable standardization of assembly information available in the PDB, improving findability and further enriching data related to these assemblies.

The PDB archive continues to be a vital resource to drive the development of scientific research, with the tools provided by PDBe supporting a wide variety of users in making the most of these data. PDBe will continue to develop these tools and services for its continuously growing user community.

## ACKNOWLEDGEMENTS

We would like to thank all collaborators and partners at the EMBL-EBI, EMBL, wwPDB, EMDB, CCP4, CCPEM, CCPN and other collaborative efforts, as well as the structural biology and bioinformatics communities. The PDB Art project is a collaboration with Cambridge art societies CANTAB and GRANTA.

## FUNDING

The Protein Data Bank in Europe is supported by European Molecular Biology Laboratory-European Bioinformatics Institute; Wellcome Trust [104948]; Biotechnology and Biological Sciences Research Council [BB/N019172/1, BB/G022577/1, BB/J007471/1, BB/K016970/1, BB/K020013/1, BB/M013146/1, BB/M011674/1, BB/M020347/1, BB/M020428/1, BB/P024351/1]; European Union [284209]; ELIXIR and Open Targets. Funding for open access charge: EMBL
*Conflict of interest statement.* None declared.

This paper is linked to: https://doi.org/10.1093/nar/gkz853.

## REFERENCES

1. Berman,H., Henrick,K., Nakamura,H. and Markley,J.L. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, **35**, D301–D302.
2. wwPDB consortium (2019) Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.*, **47**, D520–D528.
3. Burley,S.K., Berman,H.M., Bhikadiya,C., Bi,C., Chen,L., Di Costanzo,L., Christie,C., Dalenberg,K., Duarte,J.M., Dutta,S. *et al.* (2019) RCSB Protein Data Bank: Biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.*, **47**, D464–D474.

4. Kinjo,A.R., Bekker,G.-J., Suzuki,H., Tsuchiya,Y., Kawabata,T., Ikegawa,Y. and Nakamura,H. (2017) Protein Data Bank Japan (PDBj): updated user interfaces, resource description framework, analysis tools for large structures. *Nucleic Acids Res.*, **45**, D282–D288.
5. Ulrich,E.L., Akutsu,H., Doreleijers,J.F., Harano,Y., Ioannidis,Y.E., Lin,J., Livny,M., Mading,S., Maziuk,D., Miller,Z. *et al.* (2008) BioMagResBank. *Nucleic Acids Res.*, **36**, D402–D408.
6. Wilkinson,M.D., Dumontier,M., Aalbersberg,Ij.J., Appleton,G., Axton,M., Baak,A., Blomberg,N., Boiten,J.-W., da Silva Santos,L.B., Bourne,P.E. *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, **3**, 160018.
7. Young,J.Y.J.Y., Westbrook,J.D.J.D., Feng,Z., Sala,R., Peisach,E., Oldfield,T.J.T.J., Sen,S., Gutmanas,A., Armstrong,D.R.D.R., Berrisford,J.M.J.M. *et al.* (2017) OneDep: Unified wwPDB System for Deposition, Biocuration, and Validation of Macromolecular Structures in the PDB Archive. *Structure*, **25**, 536–545.
8. Abbott,S., Iudin,A., Korir,P.K., Somasundharam,S. and Patwardhan,A. (2018) EMDB Web Resources. *Curr. Protoc. Bioinforma*, **61**, 5.10.1–5.10.12.
9. Dana,J.M., Gutmanas,A., Tyagi,N., Qi,G., O'Donovan,C., Martin,M. and Velankar,S. (2019) 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.*, **47**, D482–D489.
10. Bateman,A. (2019) UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
11. El-Gebali,S., Mistry,J., Bateman,A., Eddy,S.R., Luciani,A., Potter,S.C., Qureshi,M., Richardson,L.J., Salazar,G.A., Smart,A. *et al.* (2019) The Pfam protein families database in 2019. *Nucleic Acids Res.*, **47**, D427–D432.
12. Mitchell,A.L., Attwood,T.K., Babbitt,P.C., Blum,M., Bork,P., Bridge,A., Brown,S.D., Chang,H.Y., El-Gebali,S., Fraser,M.I. *et al.* (2019) InterPro in 2019: Improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.*, **47**, D351–D360.
13. Dawson,N.L., Lewis,T.E., Das,S., Lees,J.G., Lee,D., Ashford,P., Orengo,C.A. and Sillitoe,I. (2017) CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Res.*, **45**, D289–D295.
14. Lo Conte,L., Ailey,B., Hubbard,T.J., Brenner,S.E., Murzin,A.G. and Chothia,C. (2000) SCOP: a structural classification of proteins database. *Nucleic Acids Res.*, **28**, 257–259.
15. Hunt,S.E., McLaren,W., Gil,L., Thormann,A., Schuilenburg,H., Sheppard,D., Parton,A., Armean,I.M., Trevanion,S.J., Flicek,P. *et al.* (2018) Ensembl variation resources. *Database (Oxford).*, **2018**, bay119.
16. Agarwala,R., Barrett,T., Beck,J., Benson,D.A., Bollin,C., Bolton,E., Bourexis,D., Brister,J.R., Bryant,S.H., Canese,K. *et al.* (2018) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **46**, D8–D13.
17. Cook,C.E., Lopez,R., Stroe,O., Cochrane,G., Brooksbank,C., Birney,E. and Apweiler,R. (2019) The European Bioinformatics Institute in 2018: Tools, infrastructure and training. *Nucleic Acids Res.*, **47**, D15–D22.
18. PDBe-KB consortium (2020) PDBe-KB: a community-driven resource for structural and functional annotations. *Nucleic Acids Res.*, doi:10.1093/nar/gkz853.
19. Mir,S., Alhroub,Y., Anyango,S., Armstrong,D.R., Berrisford,J.M., Clark,A.R., Conroy,M.J., Dana,J.M., Deshpande,M., Gupta,D. *et al.* (2018) PDBe: towards reusable data delivery infrastructure at protein data bank in Europe. *Nucleic Acids Res.*, **46**, D486–D492.
20. Westbrook,J.D., Shao,C., Feng,Z., Zhuravleva,M., Velankar,S. and Young,J. (2015) The chemical component dictionary: complete descriptions of constituent molecules in experimentally determined 3D macromolecules in the Protein Data Bank. *Bioinformatics*, **31**, 1274–1278.
21. Gaulton,A., Hersey,A., Nowotka,M.L., Patricia Bento,A., Chambers,J., Mendez,D., Mutowo,P., Atkinson,F., Bellis,L.J., Cibrian-Uhalte,E. *et al.* (2017) The ChEMBL database in 2017. *Nucleic Acids Res.*, **45**, D945–D954.
22. Hastings,J., Owen,G., Dekker,A., Ennis,M., Kale,N., Muthukrishnan,V., Turner,S., Swainston,N., Mendes,P. and Steinbeck,C. (2016) ChEBI in 2016: improved services and an expanding collection of metabolites. *Nucleic Acids Res.*, **44**, D1214–D1219.
23. Sterling,T. and Irwin,J.J. (2015) ZINC 15 - ligand discovery for everyone. *J. Chem. Inf. Model.*, **55**, 2324–2337.
24. Wishart,D.S., Feunang,Y.D., Guo,A.C., Lo,E.J., Marcu,A., Grant,J.R., Sajed,T., Johnson,D., Li,C., Sayeeda,Z. *et al.* (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.*, **46**, D1074–D1082.
25. Kalvari,I., Argasinska,J., Quinones-Olvera,N., Nawrocki,E.P., Rivas,E., Eddy,S.R., Bateman,A., Finn,R.D. and Petrov,A.I. (2018) Rfam 13.0: Shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.*, **46**, D335–D342.
26. Mukhopadhyay,A., Borkakoti,N., Pravda,L., Tyzack,J.D., Thornton,J.M. and Velankar,S. (2019) Finding enzyme cofactors in Protein Data Bank. *Bioinformatics*, **35**, 3510–3511.
27. Niggli,P. (1928) Krystallographische und Strukturtheoretische Grundbegriffe. *Handb. der Exp.*, **7**, 108–176.
28. McCoy,A.J., Grosse-Kunstleve,R.W., Adams,P.D., Winn,M.D., Storoni,L.C. and Read,R.J. (2007) Phaser crystallographic software. *J. Appl. Crystallogr.*, **40**, 658–674.
29. Krissinel,E. (2007) On the relationship between sequence and structure similarities in proteomics. *Bioinformatics*, **23**, 717–723.
30. Mirdita,M., Steinegger,M. and Söding,J. (2019) MMseqs2 desktop and local web server app for fast, interactive sequence searches. *Bioinformatics*, **35**, 2856–2858.
31. Velankar,S., Van Ginkel,G., Alhroub,Y., Battle,G.M.G.M., Berrisford,J.M.J.M., Conroy,M.J.M.J., Dana,J.M.J.M., Gore,S.P.S.P., Gutmanas,A., Haslam,P. *et al.* (2016) PDBe: Improved accessibility of macromolecular structure data from PDB and EMDB. *Nucleic Acids Res.*, **44**, D385–D395.
32. Callaway,E. (2015) The revolution will not be crystallized: A new method sweeps through structural biology. *Nature*, **525**, 172–174.
33. Young,J.Y., Westbrook,J.D., Feng,Z., Peisach,E., Persikova,I., Sala,R., Sen,S., Berrisford,J.M., Swaminathan,G.J., Oldfield,T.J. *et al.* (2018) Worldwide Protein Data Bank biocuration supporting open access to high-quality 3D structural biology data. *Database (Oxford).*, **2018**, D520–D528.
34. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
35. Lipman,D.J. and Pearson,W.R. (1985) Rapid and sensitive protein similarity searches published by: american association for the advancement of science stable URL, *Science*, **227**, 1435–1441.
36. Finn,R.D., Clements,J. and Eddy,S.R. (2011) HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res.*, **39**, W29–W37.
37. Groom,C.R., Bruno,I.J., Lightfoot,M.P. and Ward,S.C. (2016) The Cambridge structural database. *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.*, **B72**, 171–179.
38. Chambers,J., Davies,M., Gaulton,A., Hersey,A., Velankar,S., Petryszak,R., Hastings,J., Bellis,L., McGlinchey,S. and Overington,J.P. (2013) UniChem: A unified chemical structure cross-referencing and identifier tracking system. *J. Cheminform.*, **5**, 3.
39. Sehnal,D., Deshpande,M., Vařeková,R.S., Mir,S., Berka,K., Midlik,A., Pravda,L., Velankar,S. and Koča,J. (2017) LiteMol suite: Interactive web-based visualization of large-scale macromolecular structure data. *Nat. Methods*, **14**, 1121–1122.
40. Thieker,D.F., Hadden,J.A., Schulten,K. and Woods,R.J. (2016) 3D implementation of the symbol nomenclature for graphical representation of glycans. *Glycobiology*, **26**, 786–787.
41. Sehnal,D. and Grant,O.C. (2019) Rapidly Display Glycan Symbols in 3D Structures: 3D-SNFG in LiteMol. *J. Proteome Res.*, **18**, 770–774.
42. Meldal,B.H.M., Forner-Martinez,O., Costanzo,M.C., Dana,J., Demeter,J., Dumousseau,M., Dwight,S.S., Gaulton,A., Licata,L., Melidoni,A.N. *et al.* (2015) The complex portal - An encyclopaedia of macromolecular complexes. *Nucleic Acids Res.*, **43**, D479–D484.
43. Fabregat,A., Jupe,S., Matthews,L., Sidiropoulos,K., Gillespie,M., Garapati,P., Haw,R., Jassal,B., Korninger,F., May,B. *et al.* (2018) The Reactome Pathway Knowledgebase. *Nucleic Acids Res.*, **46**, D649–D655.
44. Iudin,A., Korir,P.K., Salavert-Torres,J., Kleywegt,G.J. and Patwardhan,A. (2016) EMPIAR: a public archive for raw electron microscopy image data. *Nat. Methods*, **13**, 387–388.
45. Morin,A., Eisenbraun,B., Key,J., Sanschagrin,P.C., Timony,M.A., Ottaviano,M. and Sliz,P. (2013) Collaboration gets the most out of software. *Elife*, **2**, e01456.

46. Grabowski,M., Langner,K.M., Cymborowski,M., Porebski,P.J., Sroka,P., Zheng,H., Cooper,D.R., Zimmerman,M.D., Elsliger,M.A., Burley,S.K. *et al.* (2016) A public database of macromolecular diffraction experiments. *Acta Crystallogr. Sect. D Struct. Biol.*, **72**, 1181–1193.
47. Watkins,X., Garcia,L.J., Pundir,S. and Martin,M.J. (2017) ProtVista: Visualization of protein sequence annotations. *Bioinformatics*, **33**, 2040–2041.
48. Favuzza,P., Guffart,E., Tamborrini,M., Scherer,B., Dreyer,A.M., Rufer,A.C., Erny,J., Hoernschemeyer,J., Thoma,R., Schmid,G. *et al.* (2017) Structure of the malaria vaccine candidate antigen CyRPA and its complex with a parasite invasion inhibitory antibody. *Elife*, **6**, e20383.
49. Yamada,M., Watanabe,Y., Gootenberg,J.S., Hirano,H., Ran,F.A., Nakane,T., Ishitani,R., Zhang,F., Nishimasu,H. and Nureki,O. (2017) Crystal structure of the minimal cas9 from campylobacter jejuni reveals the molecular diversity in the crispr-cas9 systems. *Mol. Cell*, **65**, 1109–1121.