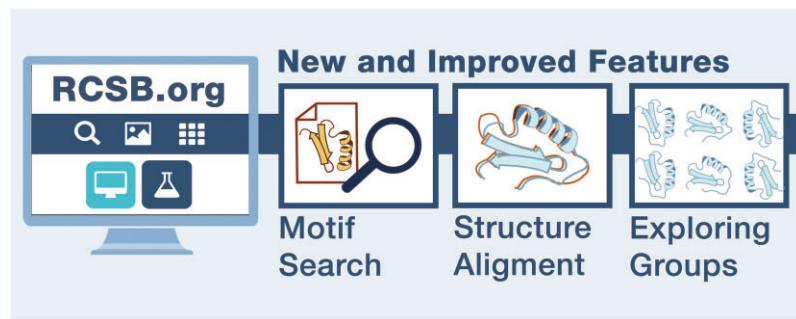


Updated resources for exploring experimentally-determined PDB structures and Computed Structure Models at the RCSB Protein Data Bank

Stephen K. Burley                                         <

Graphical abstract



Introduction

The Protein Data Bank (PDB) was established as the first open-access digital data resource in biology (1). It has been supported without interruption by the United States (U.S.) government since it was founded in 1971. At the time of writing in 2024, the RCSB Protein Data Bank (2) has been continuously funded by the U.S. National Science Foundation, National Institutes of Health and the U.S. Department of Energy for more than 25 years. With renewal of this joint funding for another five years in early 2024, the RCSB PDB will reach the milestone of 30 years of continuous U.S. federal funding in 2029. The RCSB PDB serves as the wwPDB data center in the U.S. for the global PDB archive of rigorously validated, expertly biocurated 3D structure data for large biological molecules (proteins, DNA, RNA, viruses, macromolecular machines and their complexes). PDB holdings at the time of writing include >227 000 experimentally determined atomic-level 3D structures. These data are essential for research and education across fundamental biology, health, energy and biotechnology (1,3). RCSB PDB collaborates with its wwPDB partners (including Protein Data Bank in Europe or PDBe, Protein Data Bank Japan or PDBj, Protein Data Bank China or PDBc, Electron Microscopy Data Bank or EMDB and Biological Magnetic Resonance Data Bank or BMRB) to maintain a single PDB archive and ensure that this information remains freely and publicly available to the global community (4,5).

The wwPDB adheres to the principles of fairness-accuracy-confidentiality-transparency (FACT (6)) and Findability-Accessibility-Interoperability-Reusability (FAIR (7)), ensuring equitable sharing and responsible management of the 3D biostructure data. Information stored in the PDB is made available under the most permissive Creative Commons CC0 1.0 Universal License (<https://creativecommons.org/licenses/by/4.0/>), enabling researchers around the world to access and utilize the information at no charge and with no restrictions on its usage. Recognizing its long-standing commitment to high standards of data preservation, management and open access, the PDB is accredited by CoreTrustSeal, an international organization that certifies data repositories (<https://amt.coretrustseal.org/certificates/>). More recently, the PDB was recognized by the Global Biodata Coalition (<https://globalbiodata.org>) as a Global Core Biodata Resource, of 'fundamental importance to the wider biological and life sciences community and the long-term preservation of biological data.' PDB remains a vanguard in the open access movement.

RCSB PDB activities are organized around four user-facing services (8): Service 1, deposition and biocuration; Service 2, archive management and access; Service 3, data exploration; Service 4, training, outreach and education. Recently, the organization decided to expand from the existing four integrated, interdependent services to five with the introduction of the new Service 0, IT Infrastructure. This change was designed to streamline cyberinfrastructure operations supporting the entire organization.

As reported previously (3,9–11), our research-focused RCSB.org web portal now provides access to >1M Computed Structure Models (CSMs) together with > 227 000 experimentally determined structures of biological macromolecules, obtained using macromolecular crystallography (MX), 3D electron microscopy (3DEM) or nuclear magnetic resonance spectroscopy. Development of the ModelCIF data format by the wwPDB partners in collaboration with community stakeholders (12) was key to accomplishing this integration effort, as it provided a way to seamlessly combine two kinds of 3D biostructure data, experimental and computational, while preserving the provenance and metadata that distinguish each type. Notably, the RCSB.org modular architecture dating from late 2020 (13) has proved to be highly extensible in its support of complex data and capable of accommodating a six-fold increase in data volume.

In this update prepared for publication in the 2025 *Nucleic Acids Research* Database Issue, we present new features and improvements developed and implemented since our last publication to our research-focused web portal RCSB.org (3).

Results

RCSB.org functionality supports four overarching activities comprising 3D biological macromolecule data exploration: Searching, Browsing, Visualizing and Comparing. Updates presented herein relate to these four topics, with the most prominent new features being searching with user-provided atomic coordinates and newly-introduced tools for visualizing and understanding 'Groups' of related structures (both experimentally determined and computationally predicted). New integrations with external information resources are also presented. They augment the wealth of structural and functional annotations incorporated from 50 public-domain biodata repositories. Beyond user interface (UI) improvement features, we also briefly present tools that enable easier access to RCSB PDB Application Programming Interfaces (APIs).

Structure and structure motif search for user-provided atomic coordinates

Our 3D structure similarity (14) and structure motif search (15) tools have been extended to support RCSB.org queries based on user-provided atomic coordinates, which can be uploaded in various standard formats. Previously, queries were restricted to structures available from RCSB.org (experimental PDB archive and/or select CSMs). Users can now work with atomic coordinates provided either by a publicly accessible URL (Uniform Resource Locator) or file upload. This new functionality provides a convenient and streamlined way to relate novel, previously uncharacterized 3D structures to the wealth of 3D biostructure information archived in the PDB, which includes expertly biocurated data and annotations from 50 trusted external resources, alongside >1 million CSMs coming from AlphaFold DB (16) and the ModelArchive (modelarchive.org) (17).

Search by URL or file upload is available in the Advanced Search Query Builder (Figure 1A) via the left-most dropdown menu under Structure Similarity or Structure Motif. The ‘File URL’ option can reference 3D structure data from external repositories (e.g. AlphaFold DB, the ModelArchive or the ESM Metagenomic Atlas (18)). Alternatively, the ‘File Upload’ option allows users to browse and select an atomic coordinate file locally stored on their computer, which is then automatically uploaded to the RCSB.org cyberinfrastructure and referenced using a bookmarkable and shareable URL. The Mol* 3D viewer (19) on RCSB.org (<https://www.rcsb.org/3d-view>) provides an alternative means of performing searches with user-provided atomic coordinates: after loading a user-provided structure via the Mol* interface, the structure gets automatically uploaded, and then custom Structure Motif searches can be performed (20). Regardless of the structure’s origin, users can easily export the global alignment generated by the Structure Motif Search tool as a ZIP archive. To do this, use the ‘Export Models’ option in Mol*, which becomes available after selecting the ‘Align in 3D’ link on the search results page.

This file upload feature is powered by a dedicated RCSB PDB API service (user-upload.rcsb.org). Files can be uploaded in PDBx/mmCIF (22), BinaryCIF (23) or legacy PDB file formats, which are then parsed and converted with the aid of the BioJava library (24). Each successfully uploaded file is assigned a shareable URL with a randomly generated 32-character identifier to limit access. Uploaded files are accessible *via* these ‘non-guessable’ URLs for 90 days. Structure Similarity and Structure Motif searches can be defined using URLs as input. Generated queries can be bookmarked and shared while the uploaded file URL (s) remain valid. The User Upload API is also utilized within the pairwise structure alignment application (described below). At the time of writing, this User Upload API was supporting ~20 000 external atomic coordinate file uploads monthly.

Overall, these enhancements to Structure Similarity and Structure Motif searches significantly expand the functionality and accessibility of RCSB.org cyberinfrastructure, providing researchers with powerful tools to explore, visualize, and analyze 3D biostructures with increased ease and flexibility.

Pairwise structure alignment tool enhancements

The RCSB PDB Pairwise Structure Alignment tool (Figure 2 (25)) was updated recently to provide more efficient op-

tions for constructing queries and additional interactive ways of exploring alignment results. We also introduced various improvements to streamline navigation through established alignment methods (e.g. TM-align (26) *versus* jCE-CP (27)). The tool now leverages the power of the RCSB PDB Sequence Annotations viewer (previously called 1D-3D viewer (28)) and Mol* (19) to facilitate bi-directional exploration between sequence alignments and structure superpositions. Key new features include synchronized highlighting of aligned sequences and superposed structures, the ability to show or hide aligned polymeric chains plus other polymers and non-polymers present in the PDB ID, and enhanced structure alignment visualization that distinguishes residues in spatially similar positions from those that are not.

To streamline access to the most recent predictions of protein 3D structures from AlphaFold DB (16) and ESM Metagenomic Atlas (18) for structure alignments, we introduced a feature that retrieves atomic coordinates directly via appropriate identifiers. Users can now input UniProt ID codes (34) to access AlphaFold DB and MGnify identifiers, access ESM Atlas and automatically locate and load the corresponding atomic coordinate files. Across RCSB.org, UniProt accession codes can be used to facilitate simultaneous selection of 3D biostructures of interest from among >227 000 experimentally determined PDB IDs and >1M incorporated CSMs (coming from both AlphaFold DB and ModelArchive). This feature facilitates comparison of ‘ground truth’ PDB structures with predicted atomic coordinates and enables fuller structural appreciation of many eukaryotic multi-domain proteins for which PDB coverage may not extend across the entire polypeptide chain length.

Other UI improvements were implemented to enhance navigation through available 3D structure alignment methods. Each option in the ‘Alignment Method’ dropdown now provides context-sensitive help that offers concise descriptions and guidance on usage. Available alignment methods are also grouped to clearly distinguish between rigid-body methods and methods designed to permit flexibility within proteins (e.g. FATCAT Flexible *versus* FATCAT Rigid (35)).

Importantly, RCSB.org users now can see the exact API request being issued to the RCSB PDB Alignment API service (alignment.rcsb.org) for calculating 3D structure alignments, after hitting the ‘Compare’ button on the Pairwise Structure Alignment page. The ‘Alignment API’ button opens a dedicated query editor page that provides a vehicle for further customizing API queries, enabling users with varying levels of technical expertise to utilize the Alignment API. Similar API buttons are available elsewhere across RCSB.org (described in the section entitled *API discoverability*).

Exploring Groups of similar proteins

The PDB archive has captured the outcomes of a multitude of structure determination experiments over more than five decades. Given funding emphases and challenges of trying to determine structures of some biological macromolecules, the archival contents do not provide uniform representations of protein and nucleic acid structure shapes. Many PDB IDs are structurally similar if not nearly identical (e.g. PDB holdings include >1450 independently determined X-ray crystal structures of the SARS-CoV-2 main protease, the target of nirmatrelvir, the active ingredient of Paxlovid (36), beginning with

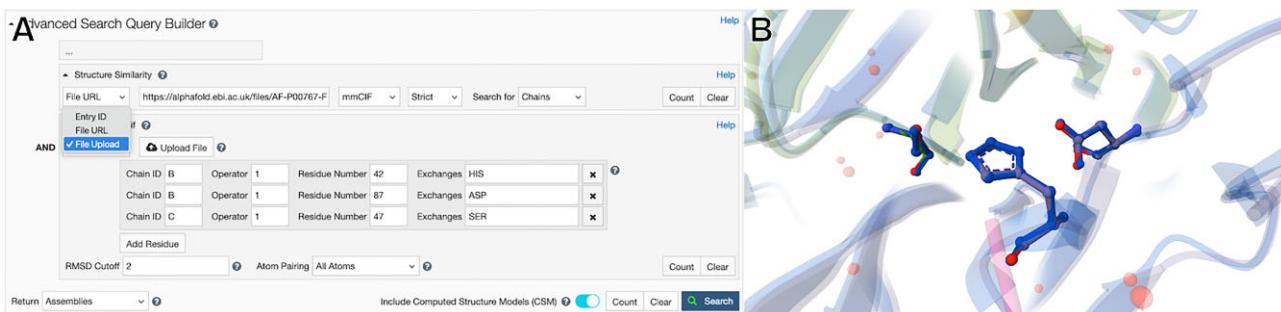


Figure 1. (A) Advanced Search Query Builder, showcasing support for searches defined by File URL and File Upload for Structure Similarity and Structure Motif searches. **(B)** Mol* 3D visualization example of a Structure Motif search match with the serine protease catalytic triad (search motif derived from PDB ID: 4cha (21)). Shown here is one of the returned matches, the AlphaFold CSM for bovine Chymotrypsinogen B (UniProt ID: P00767).

	Entry	Chain	RMSD	TM-score	Identity	Aligned Residues	Sequence Length	Modeled Residues
Glyceraldehyde-3-phosphate dehydrogenase	1CF2	A	-	-	-	-	337	336
	2YYY	A	0.79	0.97	60%	336	343	342
	2CZC	A	1.01	0.97	56%	332	334	334
Aspartate-semialdehyde dehydrogenase	3HSK	A	3.03	0.67	14%	280	381	358
	6C8W	A	3.04	0.68	14%	283	375	357

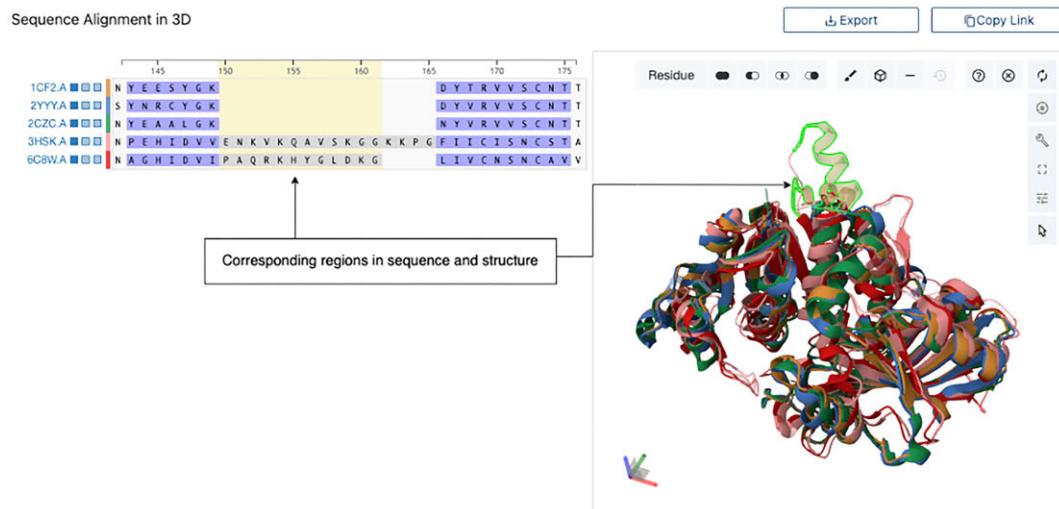


Figure 2. Alignment of structures to explore the NADP-binding sites of glyceraldehyde-3-phosphate dehydrogenase from *Methanothermus fervidus* (PDB ID: 1cf2 chain A, in orange, the reference structure (29)), *Methanocaldococcus jannaschii* DSM 2661 (PDB ID: 2yyy chain A (30), blue), *Pyrococcus horikoshii* OT3 (PDB ID: 2c2c chain A (31), green), aspartate-semialdehyde dehydrogenase from *Candida albicans* (PDB ID: 3hsk chain A (32), salmon) and *Blastomyces gilchristii* SLH14081 (PDB ID: 6c8w chain A (33), red).

PDB ID: 6lu7 ([37](#)). To help users make sense of this wealth of information, RCSB.org supports exploration of ‘Groups’ of structures ([10](#)).

The RCSB.org Advanced Search interface and Search API both provide tools that manage cases in which searches return PDB IDs that are similar in 3D structure. Different approaches are supported, depending on whether the search returns *Structures* or *Polymer Entities*. When the search return type is set to ‘Structure’ (meaning PDB ID), the results can be organized based on the deposition Group identifier denoting collections of structures deposited simultaneously *via* GroupDep (38). When set to ‘Polymer Entities’, results can be grouped based on amino acid sequence similarity clusters (39) or collections of structures that share the same UniProt ID. Once search results are grouped, results can be presented as a list of Groups

or as a list of representative members of each Group. The search configuration also offers different options for selecting Group representatives, including best experimental resolution (i.e. lowest number in Å for MX or 3DEM structures), largest total residue count, largest number of chains in the PDB ID, most extensive polypeptide chain coverage (when grouping by UniProt ID) or highest search relevance Score (as judged by our Search API system, based on the Elasticsearch relevance calculation). When RCSB.org Advanced Search presents results as Groups, each result item summarizes the Group, displaying relevant features such as Group name, the number of Group members satisfying the search conditions and the diversity of source organisms, domain families and Enzyme Classification (EC) numbers represented in the Group. Additionally, the title of the returned search result items also serves as a

link to the Group Summary Page (GSP) associated with the grouping strategy.

Examining collections of proteins with shared similarities can uncover patterns in both sequence and 3D structure that may be missed when proteins are analyzed individually. RCSB.org provides a robust set of tools for analyzing and visualizing Groups of proteins. These Groups are defined using the same strategies implemented for ‘redundancy’ reduction in search results:

- **Deposition Groups:** Sets of structures that researchers deposited simultaneously using the GroupDep tool (38).
- **UniProt Groups:** Encompass all proteins in the PDB and/or CSMs incorporated into RCSB.org with amino acid sequences sharing the same UniProt ID.
- **Sequence Identity Groups:** Defined by calculating sequence identity clusters of protein sequences across structures in the PDB archive and/or CSMs incorporated into RCSB.org.

Group tools now available at RCSB.org were developed to enable comprehensive understanding of structural and functional annotations across multiple members of a Group. They facilitate visualization of Group member property distributions, multiple sequence alignments (MSAs) and 3D structure superpositions. Group tools are accessible from GSPs, which serve as the main hub for exploring all Group-related features, including Group MSA visualization and Group 3D structure superpositions. GSPs can be accessed from structure summary pages (SSPs) and Advanced Search results displayed as Groups. SSPs link each of their protein sequences (PDB IDs, which are listed under the Macromolecules section) to their corresponding UniProt and Sequence Identity GSPs. When Advanced Search results are presented as Groups, the listed result items link to their corresponding GSP based on the selected grouping strategy (Deposition, UniProt or Sequence Identity).

GSPs (Figure 3) serve as the primary entry point for exploring Groups and Group-related resources, such as Group Sequence Pages and Group Sequence Alignments in 3D. GSPs contain all the data and metadata related to the Group and its members, including both experimentally determined PDB structures and CSMs. The GSP layout is similar to SSPs, featuring a page header, a carousel of images of each Group member (top-left), and various sections with key information about the Group member properties. Each header showcases the Group title, the methodology used for grouping members, the total count of elements within the Group and, when accessed from the Advanced Search, the number of Group elements matching the search query. The ‘carousel’ component enables exploration of individual Group members. Each slide in the carousel displays specific details about a Group member, including name, organism and key experimental information such as resolution and molecular weight. Group member properties are represented as histograms that plot the distribution of specific properties of the Group members (e.g. resolution, release date or source organism). These histograms are organized into different sections displaying distribution of related features. For example, the ‘Protein Domains’ section shows the distribution of protein domains within the Group members based on various protein domain classification databases, including CATH (40), SCOP/SCOPe (41,42), ECOD (43) and Pfam (44).

When GSPs are accessed from Advanced Search results, histograms display distributions using two different colors: blue represents the distribution of the Group members that matched the search query, while Group members that do not match the search are displayed in gray (Figure 3). In addition, histograms allow exploration of subsets of a distribution. Clicking on a histogram bar will update all plots to display the distribution for the Group members related to the selected property. For example, if the X-ray Crystallography bar is clicked in the Experimental Method histogram, all histogram plots will be updated to display the distribution of the Group members determined using X-ray Crystallography. A floating navigation menu ‘Query History’ allows going back to the previous unfiltered view.

Group Sequence Pages are available for UniProt- and sequence identity-based Groups. These pages use the Sequence Annotation Viewer (45) to provide detailed sequence information for Group members, divided into three main sections. The Sequence Alignment tab displays a graphical representation of the MSA of Group members, helping users visualize how sequences align across the Group and highlighting conserved and variable regions. The Structural Features tab shows the distribution of structural properties, such as secondary structure (e.g. α -helices and β -sheets) and protein domains by using a color gradient to represent the frequency of occurrence at a certain position. The Binding Sites section displays the distribution of protein-ligand binding sites. The content displayed on Group Sequence Pages is interconnected with information from GSPs. When RCSB.org users select a specific property to inspect a subset of the initial Group, the Group Sequence Page updates to display relevant data for that subset, allowing for more targeted analyses and visualization.

The Sequence Alignments in 3D tool is designed to integrate sequence and structural data for Group members (Figure 4). This tool is also available for both UniProt- and sequence identity-based Groups and is accessible from the GSP carousel and the Group Sequence Pages by clicking the ‘Sequence Alignments in 3D’ link. This tool utilizes the RCSB PDB 1D3D (45) to integrate the Sequence Annotation Viewer and the Mol* 3D viewer, presenting the MSA of Group members with their 3D structures. MSA visualization mirrors the information provided on GSPs, with the added functionality of clicking on track titles to load Group member 3D structures into the Mol* 3D viewer. Newly loaded structures are superimposed on the reference structure (first loaded), facilitating comparisons of Group members in 3D. The UI for this tool is identical to that used for the Pairwise Structure Alignment tool (see *Pairwise Structure Alignment tool enhancements* above).

Integration of data from new external resources

In addition to providing access to 3D structure data and related metadata, RCSB.org integrates information from numerous open-access bioinformatics resources and makes these data available for users to map onto 3D structures and visualize them. Depending on the type of annotation, integrated annotations may be presented for specific structures (entry-level annotations) or at the level of the biopolymers (entity- or instance-level annotations) present in the structure. These annotations can be used to learn about biomolecular structural attributes (e.g. CATH domains, membrane binding regions), functional properties (e.g. EC numbers, anatomical therapeutic classification of drugs) and locations of special significance

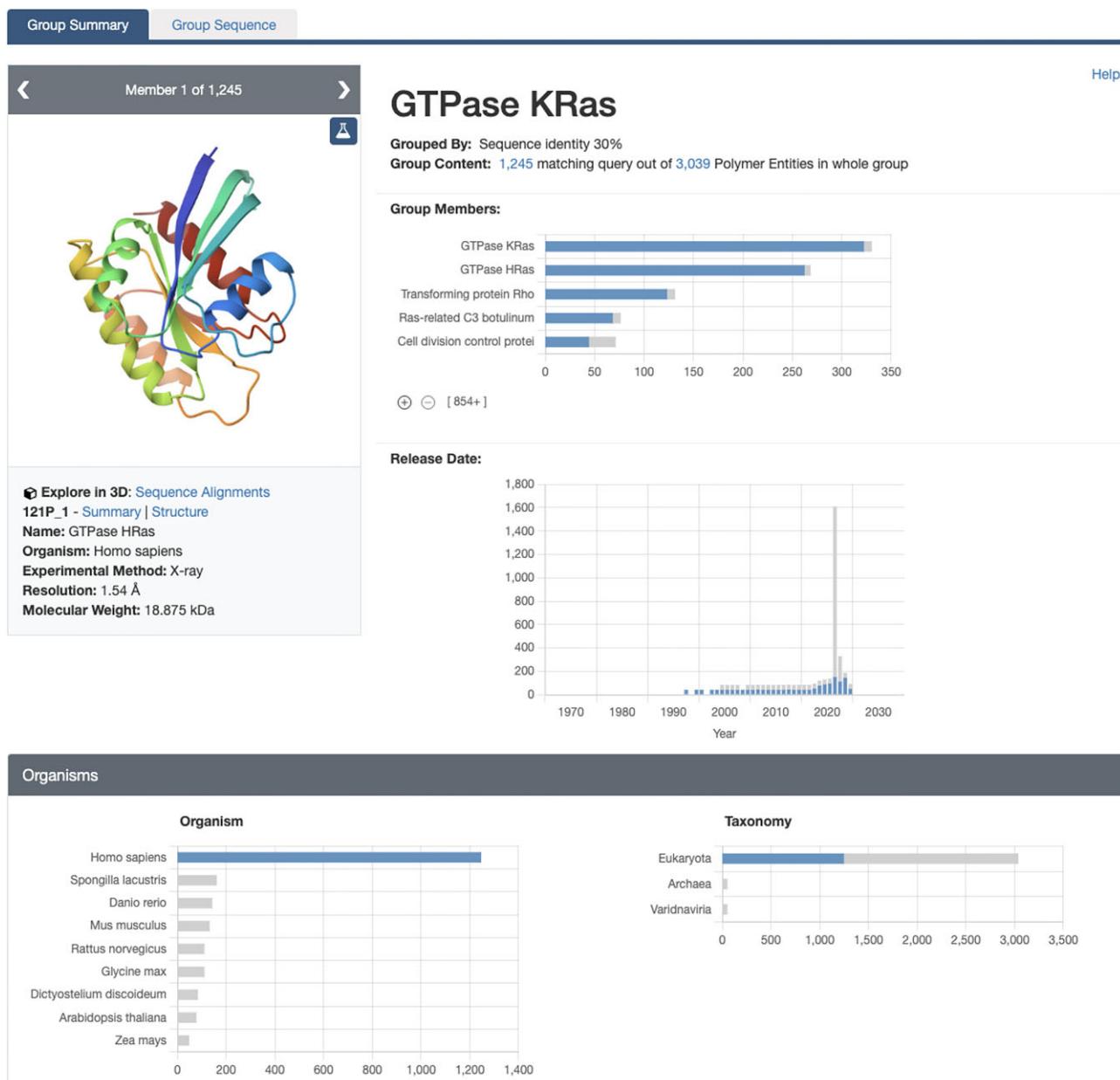


Figure 3. GSP for a 30% sequence identity cluster of Ras family proteins. Histograms display the distribution of organisms among the Group members, with the currently filtered *Homo sapiens* displayed in blue. Gray bars represent the distribution of Group members belonging to other organisms.

(e.g. sites of post-translational modification, commonly observed variants of the polymer sequence) (3). Newly integrated annotations enable users to explore antibiotic resistance and mechanisms of enzyme catalysis across the PDB archive. Including these annotations along with other criteria to construct complex queries can help identify structures relevant to diverse user research interests.

The Comprehensive Antibiotic Resistance Database (CARD (46)) is a highly curated database of antibiotics and antibiotic resistance genes, their protein products, and phenotypic information, organized by the CARD Antibiotic Resistance Ontology ('ARO'). Polymer sequences of proteins present in PDB IDs that perfectly or closely match (>95% amino acid sequence identity and >80% polypeptide chain coverage) the antibiotic resistance proteins referenced in CARD are identified and linked to specific annotations, such

as the matched protein gene name, ARO identifier, description, resistant antibiotic drug classes and the biochemical mechanism of resistance. When sequences of the protein in the PDB and the reference Antimicrobial Resistance Gene have <95% sequence identity, only gene family annotations are included at the Annotations page. Links at RCSB.org to the CARD database provide access to additional details about these annotations. Identifying related antibiotic resistance proteins can provide valuable insights into biochemical/biological function.

The Mechanism and Catalytic Site Atlas (M-CSA) resource is a database of enzyme reaction mechanisms (47). It also provides annotations concerning active site and catalytic residues plus enzyme reaction cofactor requirements. Each one of the ~1000 entries stored in this external resource is linked to at least one experimentally-determined structure archived in

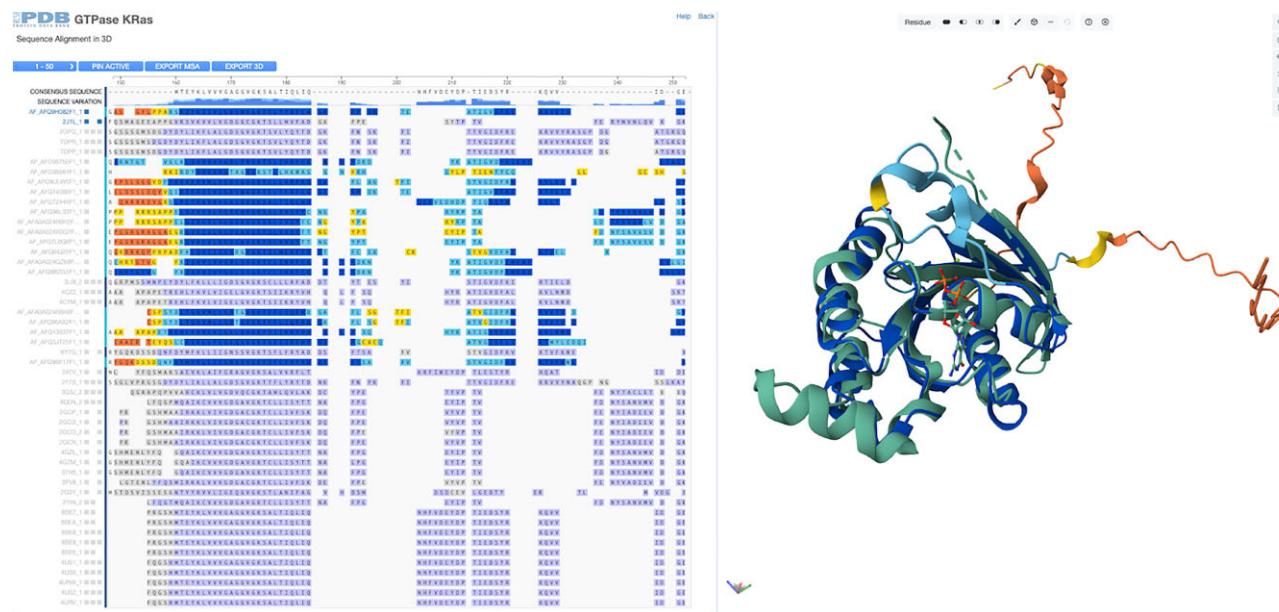


Figure 4. Group Sequence Alignments in 3D for Ras family proteins clustered at 30% amino acid sequence identity. On the left side, an MSA of the group members is displayed, including experimental structures and CSMs. The right panel shows the 3D structure of selected proteins from the alignment panel in Mol*.

the PDB. Linked PDB IDs are annotated with details from the M-CSA, which can in turn be used to query other structures (experimentally-determined from PDB or CSMs from AlphaFold DB or the ModelArchive) that have active-site 3D structure motifs similar to the enzyme of interest. Structure motif searches launched in this manner can help an RCSB.org user identify and learn about other PDB structures with comparable active site architecture. For example, in PDB ID: 1b73 (48), which includes an M-CSA annotation, the Catalytic Residues column can be used to search for other PDB structures or CSMs that have the same catalytic site residues and/or the same EC number, using options shown in Figure 5.

Recently, annotations of 3D structures available from RCSB.org were extended to CSMs. Thus polymer sequence-level annotations derived from the UniProt are mapped to CSM structures available from RCSB.org and made accessible from the SSP Annotations tab. Some of these annotations include Gene Ontology (49), InterPro (50) and Pharos (51), and can provide new perspectives about the CSMs being examined (e.g. Figure 6).

Support for extended chemical component identifiers

Several improvements for exploring PDB chemical components (small molecule chemicals) and their specific instances within individual PDB IDs have been added recently. The first, transparent to end-users but vitally important for software providers, concerned extension of PDB Chemical Component IDs from three to five characters. Three-character identifiers were predicted to be fully exhausted in late 2023, thus the wwPDB implemented necessary changes in PDB structure deposition, validation, and biocuration processes to extend the length of the identifier to five characters. With >60 million combinations, Chemical Component IDs are effectively future-proof. Our RCSB.org research-focused web portal and all APIs that underpin features therein were made fully com-

pliant with the five-character ID in late 2023 when the first such five-character ID was publicly released. As of June 2024, the PDB had issued 416 such Chemical Component IDs, represented at the time within 479 distinct PDB IDs.

Importantly, due to limitations in the legacy PDB format, PDB structures containing the new five-character Chemical Component IDs will only be provided in PDBx/mmCIF format. Transitioning to the PDBx/mmCIF format is highly encouraged. wwPDB resources (<https://mmcif.wwpdb.org/docs/user-guide/guide.html>) are available to assist data consumers, software providers, etc. in moving away from the legacy PDB format.

The Ligand Summary Page, the main landing page for exploration of Chemical Components at RCSB.org, now utilizes Mol* to unify the 3D visualization experience across our research-focused web portal. Users can also toggle views between ideal (from the Chemical Component Dictionary) and model atomic coordinates.

A further user experience improvement to ligand exploration was recently introduced at the SSPs 'Small Molecules' section, wherein the '3D Interactions' buttons are now dropdown menus that allow selection of specific ligand instances and two different ways of visualizing the atomic coordinates: with experimental electron density (now presented in a wireframe style by default), or atomic coordinates-only.

Incorporation of additional computed structure models

In previous publications (3,9–11), we described expansion of structural coverage at RCSB.org through the incorporation and parallel delivery of ~1 000 000 CSMs alongside a collection of the then-available ~195 000 experimentally-determined, rigorously validated, and expertly-biocurated PDB structures. The CSMs included in this release encompassed AlphaFold2 (52) predictions of entire proteomes for model organisms, common human pathogens, crop plants

Structure Motif Annotation: Mechanism and Catalytic Site Atlas

Chains	Enzyme Name	Description	Catalytic Residues
A	glutamate racemase M-CSA #1	Glutamate racemase is responsible for the synthesis of D-glutamate, an essential building block of peptidoglycan, found i... View More	Defined by 6 residues: ASP:A-7, SER:A-8, CYS:A-70, GLU:A_2-147, CYS:A-178, HIS:A-180 Explore in 3D: M-CSA 1 Motif Definition Search M-CSA Motif 2 EC: 5.1.1.3 (PDB Primary Data) 3 Search M-CSA Motif + EC 5.1.1.3 4

Figure 5. Tabular representation of M-CSA annotations for PDB ID: 1b73 (48). Links allow the user to (1) visualize the catalytic residues in Mol*; (2) launch a structure motif search based on the M-CSA motif definition; (3) launch a search for other structures in the PDB with the same EC number; (4) and launch a search with both the structure motif for similar arrangements of catalytic residues and PDB IDs or CSMs bearing the same EC number.

and nearly all of Swiss-Prot (coming from AlphaFold DB (16), <https://alphafold.ebi.ac.uk/>), plus a set of predicted yeast protein binary complexes generated using a combination of RoseTTAFold and AlphaFold2 (53,54) coming from the ModelArchive, thereby broadening structural coverage at RCSB.org substantially.

This expansion of coverage is critically important for RCSB.org users interested in model organisms with modest PDB coverage. Sole reliance on experimentally-determined PDB structures of *Arabidopsis thaliana* proteins, effectively limits 3D structural coverage of that proteome to ~4%. With incorporation of *A. thaliana* CSMs from AlphaFold DB, RCSB.org now provides plant molecular and cellular biologists with access to 3D structure information across the entire proteome. Similar coverage enhancements are now benefitting researchers focused on studying proteins from *C. elegans*, *D. melanogaster*, *M. musculus*, *S. pombe*, etc.

Following this initial CSM release, additional datasets (encompassing tens of thousands of new computationally predicted protein structures) have been incorporated into RCSB.org as part of our commitment to serving as a one-stop shop for the study of 3D structures of biological macromolecules. These datasets (all coming from the ModelArchive) include predicted protein structures from *Spongilla lacustris* (freshwater sponge) (55), African swine fever virus (56) and *Sphagnum divinum* (peat moss (57)), as well as a set of heterodimeric proteins from the cancer interactome in humans (58). Collectively, these additions have contributed ~68 000 CSMs, bringing the total number of CSMs available on RCSB.org to ~1 068 000. Importantly, owing to the development and adoption of the ModelCIF data standard (the extension to the PDB/mmCIF data standard developed for computationally predicted protein structures) (12), every one of these CSMs could be fully integrated at RCSB.org and made interoperable with the same set of powerful tools for search, browsing, analysis and visualization available for the extant collection of ~223 000 experimentally determined PDB structures.

API discoverability

SSPs and Ligand Summary Pages now contain a new button 'Data API' (located at the top-right), which opens the GraphQL query to RCSB PDB Data APIs in its query editor. This new feature improves discoverability of APIs for users who want to learn how certain data can be extracted using APIs and to understand how information is structured within

RCSB PDB data schemas. Beyond the main SSP tab, similar buttons are offered in other commonly accessed tabs: Annotations, Experiment, Ligands and Versions. Analogously, the search result page offers a 'Search API' button on top-right to open (in the Search API query editor) the search query actually used to obtain the search results displayed in the current view. A common cogwheel icon styling across all these API buttons provides a visual cue to users that serves to identify the functionality.

Another recent development that offers API usability improvements at RCSB.org is our release of the py-rscsbssearchapi Python library (<https://github.com/rscsbs/py-rscsbssearchapi>), a Python interface to the RCSB PDB Search API. Queries against the API can be constructed conveniently using either an operator interface or a fluent syntax. With access to this library, users can easily incorporate queries to the Search API in their own software or scripts.

Training opportunities

RCSB PDB hosts webinars, virtual and in-person crash courses, and virtual office hours to support users in their research and training across disciplines. A recent event demonstrated how to use RCSB.org features to navigate CSMs in the context of experimentally determined PDB structures all in 3D ('A Deep Dive into Computed Structure Model Exploration at RCSB.org'). Other events included an introduction to PDB structure quality metrics, including those presented in wwPDB Validation Report graphical sliders ('Understanding PDB Validation: Which experimental structures should I rely on?'); 'Visualize Biomolecular Structures with Mol*: From Atoms to Movies'; and 'Python Scripting for Molecular Docking'. Webinars are recorded and published on PDB-101 (<https://pdb101.rcsb.org/>) in the 'Train' section. PDB-101 hosts training, outreach and education resources developed by the RCSB PDB that are focused on structural biology and related topics (59). It aims to build confidence in current and future users to promote more effective utilization of RCSB.org tools and 3D biostructure data writ large. Materials are developed for users at various skill levels, targeting graduate students, postdoctoral fellows and established researchers in subject areas ranging from data deposition to data exploration.

Other PDB-101 training materials include a highly popular *Guide to Understanding PDB Data*, which supports RCSB.org users who do not have strong backgrounds in structural biology or data science. Available articles span from a 'Beginner's Guide to PDB/mmCIF' (the format that drives the PDB

Structure Summary
Structure
Annotations
Sequence
Genome

Display Files
Download Files
Data API

AF_AFP05108F1

Computed structure model of Cholesterol side-chain cleavage enzyme, mitochondrial

External Resource: Annotation

- Gene Ontology: Gene Product Annotation
- InterPro: Protein Family Classification

Gene Ontology Database Homepage

Gene Ontology: Gene Product Annotation

Chains	Polymer	Molecular Function	Biological Process	Cellular Component
A	Cholesterol side-chain cleavage enzyme, mitochondrial	<ul style="list-style-type: none"> • binding • organic cyclic compound binding • tetrapyrrole binding • heme binding • cation binding • transition metal ion binding • iron ion binding • ion binding • metal ion binding • small molecule binding • oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, reduced iron-sulfur protein as one donor, and incorporation of one atom of oxygen • oxidoreductase activity • monooxygenase activity • oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen 	<ul style="list-style-type: none"> • cellular biosynthetic process • regulation of hormone levels • steroid metabolic process • cellular metabolic process • biological regulation • C21-steroid hormone biosynthetic process • lipid biosynthetic process • primary metabolic process • C21-steroid hormone metabolic process • hormone metabolic process • biosynthetic process • metabolic process • cellular process • regulation of biological quality 	<ul style="list-style-type: none"> • membrane-bounded organelle • intracellular organelle • intracellular anatomical structure • mitochondrion • organelle • cytoplasm • cellular anatomical entity • intracellular membrane-bounded organelle • mitochondrial membrane • organelle inner membrane • membrane • mitochondrial inner membrane • organelle envelope • mitochondrial envelope

Gene Ontology Database Homepage

InterPro: Protein Family Classification

Chains	Accession	Name	Type
A	IPR002401	Cytochrome P450, E-class, group I	Family
A	IPR017972	Cytochrome P450, conserved site	Conserved Site
A	IPR001128	Cytochrome P450	Family
A	IPR036396	Cytochrome P450 superfamily	Homologous Superfamily

InterPro Database Homepage

Figure 6. Annotations available for a CSM of cholesterol side chain cleavage enzyme, mitochondrial (AlphaFold: AFP05108F1).

archive and RCSB.org) to a focus on biological assemblies. New articles are published online as major features are added to RCSB PDB, such as an ‘Introduction to RCSB PDB APIs’ and ‘Computed Structure Models’ (a.k.a. CSMs).

Future training events are announced at RCSB.org, PDB-101, and on social media; would-be attendees can also subscribe to our training events newsletter at <https://pdb101.rcsb.org/train/training-events> or info@rcsb.org to receive additional notices.

Data availability

No new data were generated or analyzed in support of this research. Resources described are freely available at RCSB.org.

Acknowledgements

The authors thank the tens of thousands of structural biologists working on all inhabited continents who have deposited structures to the PDB since 1971, and the many millions of researchers, educators and students around the world who consume PDB data. We thank the members of the RCSB PDB and wwPDB Advisory Committees for their valued advice. We also gratefully acknowledge contributions to the success of the PDB archive made by past members of RCSB PDB and our Worldwide Protein Data Bank partners (PDBe, PDBj, PDBc, EMDB and BMRB). The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Funding

U.S. National Science Foundation [DBI-2321666, PI: to S.K.B., RCSB PDB Core Operations]; U.S. Department of Energy [DE-SC0019749, PI to S.K.B.]; National Cancer Institute, the National Institute of Allergy and Infectious Diseases, and the National Institute of General Medical Sciences of the National Institutes of Health [R01GM157729, PI to S.K.B.]; NSF (to RCSB PDB); UK Biotechnology and Biological Sciences Research Council (to PDBe) jointly supported development of the Next Generation PDB archive [DBI-2019297, PI to S.K.B., BB/V004247/1, PI to S.V.]; new Mol* features [DBI-2129634, PI to S.K.B., BB/W017970/1, PI to S.V.]. NSF [DBI-1756248 and DBI-2112966, PI to B.V., DBI-1756250 and DBI-2112967, PI to A.S.]. NIH-NIGMS [R01GM083960, PI to A.S.; P41GM109824, PI to M.P.R.]. Funding for open access charge: NIH [R01GM157729].

Conflict of interest statement

None declared.

References

1. Protein Data Bank (1971) Crystallography: Protein Data Bank. *Nat. New Biol.*, **233**, 223–223.
2. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
3. Burley,S.K., Bhikadiya,C., Bi,C., Bittrich,S., Chao,H., Chen,L., Craig,A.P., Crichlow,G.V., Dalenberg,K., Duarte,J.M., et al. (2023) RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million Computed Structure Models of proteins from Artificial Intelligence/Machine Learning. *Nucleic Acids Res.*, **51**, D488–D508.
4. Berman,H.M., Henrick,K. and Nakamura,H. (2003) Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.*, **10**, 980.
5. wwPDB consortium (2019) Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.*, **47**, D520–D528.
6. van der Aalst,W.M.P., Bichler,M. and Heinzl,A. (2017) Responsible data science. *Bus. Info. Syst. Eng.*, **59**, 311–313.
7. Wilkinson,M.D., Dumontier,M., Aalbersberg,I.J., Appleton,G., Axton,M., Baak,A., Blomberg,N., Boiten,J.W., da Silva Santos,L.B., Bourne,P.E., et al. (2016) The FAIR guiding principles for scientific data management and stewardship. *Sci. Data*, **3**, 160018.
8. Burley,S.K., Bhikadiya,C., Bi,C., Bittrich,S., Chen,L., Crichlow,G., Christie,C.H., Dalenberg,K., Costanzo,L.D., Duarte,J.M., et al. (2021) 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.*, **49**, D437–D451.
9. Bittrich,S., Bi,C., Bhikadiya,C., Chao,H., Duarte,J.M., Dutta,S., Fayazi,M., Henry,J., Khokhriakov,I., Lowe,R., et al. (2023) RCSB Protein Data Bank: efficient searching and simultaneous access to one million Computed Structure Models alongside the PDB structures enabled by architectural advances. *J. Mol. Biol.*, **435**, 167994.
10. Segura,J., Rose,Y., Bi,C., Duarte,J., Burley,S.K. and Bittrich,S. (2023) RCSB Protein Data Bank: visualizing groups of experimentally determined PDB structures alongside computed structure models of proteins. *Front. Bioinform.*, **3**, 1311287.
11. Burley,S.K., Piehl,D.W., Vallat,B. and Zardecki,C. (2024) RCSB Protein Data Bank: supporting research and education worldwide through explorations of experimentally determined and computationally predicted atomic level 3D biostructures. *IUCrJ*, **11**, 279–286.
12. Vallat,B., Tauriello,G., Bienert,S., Haas,J., Webb,B.M., Zidek,A., Zheng,W., Peisach,E., Piehl,D.W., Anischanka,I., et al. (2023) ModelCIF: an extension of PDBx/mmCIF data representation for computed structure models. *J. Mol. Biol.*, **435**, 168021.
13. Rose,Y., Duarte,J.M., Lowe,R., Segura,J., Bi,C., Bhikadiya,C., Chen,L., Rose,A.S., Bittrich,S., Burley,S.K., et al. (2021) RCSB Protein Data Bank: architectural advances towards integrated searching and efficient access to macromolecular structure data from the PDB archive. *J. Mol. Biol.*, **433**, 166704.
14. Guzenko,D., Burley,S.K. and Duarte,J.M. (2020) Real time structural search of the Protein Data Bank. *PLoS Comput. Biol.*, **16**, e1007970.
15. Bittrich,S., Burley,S.K. and Rose,A.S. (2020) Real-time structural motif searching in proteins using an inverted index strategy. *PLoS Comput. Biol.*, **16**, e1008502.
16. Varadi,M., Bertoni,D., Magana,P., Paramval,U., Pidruchna,I., Radhakrishnan,M., Tsenkov,M., Nair,S., Mirdita,M., Yeo,J., et al. (2024) AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res.*, **52**, D368–D375.
17. Schwede,T., Sali,A., Honig,B., Levitt,M., Berman,H.M., Jones,D., Brenner,S.E., Burley,S.K., Das,R., Dokholyan,N.V., et al. (2009) Outcome of a workshop on applications of protein models in biomedical research. *Structure*, **17**, 151–159.
18. Lin,Z., Akin,H., Rao,R., Hie,B., Zhu,Z., Lu,W., Smetanin,N., Verkuil,R., Kabeli,O., Shmueli,Y., et al. (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, **379**, 1123–1130.
19. Sehgal,D., Bittrich,S., Deshpande,M., Svobodova,R., Berka,K., Bazgier,V., Velankar,S., Burley,S.K., Koca,J. and Rose,A.S. (2021) Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Res.*, **49**, W431–W437.
20. Burley,S.K. and Berman,H.M. (2021) Open-access data: a cornerstone for artificial intelligence approaches to protein structure prediction. *Structure*, **29**, 515–520.
21. Tsukada,H. and Blow,D.M. (1985) Structure of alpha-chymotrypsin refined at 1.68 Å resolution. *J. Mol. Biol.*, **184**, 703–711.
22. Westbrook,J.D., Young,J.Y., Shao,C., Feng,Z., Guranovic,V., Lawson,C., Vallat,B., Adams,P.D., Berrisford,J.M., Bricogne,G., et al. (2022) PDBx/mmCIF Ecosystem: foundational semantic tools for structural biology. *J. Mol. Biol.*, **434**, 167599.
23. Sehgal,D., Bittrich,S., Velankar,S., Koca,J., Svobodova,R., Burley,S.K. and Rose,A.S. (2020) BinaryCIF and CIFTools-Lightweight, efficient and extensible macromolecular data management. *PLoS Comput. Biol.*, **16**, e1008247.
24. Lafita,A., Bliven,S., Prlic,A., Guzenko,D., Rose,P.W., Bradley,A., Pavan,P., Myers-Turnbull,D., Valasatava,Y., Heuer,M., et al. (2019) BioJava 5: a community driven open-source bioinformatics library. *PLoS Comput. Biol.*, **15**, e1006791.
25. Bittrich,S., Segura,J., Duarte,J.M., Burley,S.K. and Rose,Y. (2024) RCSB Protein Data Bank: exploring protein 3D similarities via comprehensive structural alignments. *Bioinformatics*, **40**, btae370.
26. Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.
27. Bliven,S.E., Bourne,P.E. and Prlic,A. (2015) Detection of circular permutations within protein structures using CE-CP. *Bioinformatics*, **31**, 1316–1318.
28. Segura,J., Rose,Y., Bittrich,S., Burley,S.K. and Duarte,J.M. (2022) RCSB Protein Data Bank 1D3D module: displaying positional features on macromolecular assemblies. *Bioinformatics*, **38**, 3304–3305.
29. Charron,C., Talfournier,F., Isupov,M.N., Branlant,G., Littlechild,J.A., Vitoux,B. and Aubry,A. (1999) Crystallization and preliminary X-ray diffraction studies of

D-glyceraldehyde-3-phosphate dehydrogenase from the hyperthermophilic archaeon *Methanothermus fervidus*. *Acta Crystallogr. Series D*, **55**, 1353–1355.

30. Malay,A.D., Bessho,Y., Ellis,M.J., Antonyuk,S.V., Strange,R.W., Hasnain,S.S., Shinkai,A., Padmanabhan,B. and Yokoyama,S. (2009) Structure of glyceraldehyde-3-phosphate dehydrogenase from the archaeal hyperthermophile *Methanocaldococcus jannaschii*. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.*, **65**, 1227–1233.

31. Ito,K., Arai,R., Kamo-Uchikubo,T., Shirouzu,M., Yokoyama,S. and IKEN Structural Genomics/Proteomics Initiative (RSGI)IKEN Structural Genomics/Proteomics Initiative (RSGI) (2005) Crystal structure of glyceraldehyde-3-phosphate dehydrogenase from *Pyrococcus horikoshii* OT3. <https://doi.org/10.2210/pdb2CZC/pdb>.

32. Arachea,B.T., Liu,X., Pavlovsky,A.G. and Viola,R.E. (2010) Expansion of the aspartate beta-semialdehyde dehydrogenase family: the first structure of a fungal ortholog. *Acta Crystallogr. Series D*, **66**, 205–212.

33. Dahal,G.P. and Viola,R.E. (2018) Structural insights into inhibitor binding to a fungal ortholog of aspartate semialdehyde dehydrogenase. *Biochem. Biophys. Res. Commun.*, **503**, 2848–2854.

34. UniProt Consortium (2023) UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.*, **51**, D523–D531.

35. Ye,Y. and Godzik,A. (2003) Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics*, **19**, ii246–ii255.

36. Burley,S.K. (2024) Protein Data Bank: from two epidemics to the global pandemic to mRNA vaccines and Paxlovid. *Curr. Opin. Struct. Biol.*, <https://doi.org/10.1016/j.sbi.2024.102954>.

37. Jin,Z., Du,X., Xu,Y., Deng,Y., Liu,M., Zhao,Y., Zhang,B., Li,X., Zhang,L., Peng,C., et al. (2020) Structure of M (pro) from SARS-CoV-2 and discovery of its inhibitors. *Nature*, **582**, 289–293.

38. 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.

39. Steinegger,M. and Soding,J. (2018) Clustering huge protein sequence sets in linear time. *Nat. Commun.*, **9**, 2542.

40. Sillitoe,I., Bordin,N., Dawson,N., Waman,V.P., Ashford,P., Scholes,H.M., Pang,C.S.M., Woodridge,L., Rauer,C., Sen,N., et al. (2021) CATH: increased structural coverage of functional space. *Nucleic Acids Res.*, **49**, D266–D273.

41. Andreeva,A., Kulesha,E., Gough,J. and Murzin,A.G. (2020) The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Res.*, **48**, D376–D382.

42. Chandonia,J.M., Fox,N.K. and Brenner,S.E. (2019) SCOPe: classification of large macromolecular structures in the structural classification of proteins-extended database. *Nucleic Acids Res.*, **47**, D475–D481.

43. Cheng,H., Liao,Y., Schaeffer,R.D. and Grishin,N.V. (2015) Manual classification strategies in the ECOD database. *Proteins Struct. Funct. Genet.*, **83**, 1238–1251.

44. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M.,

Sangrador-Vegas,A., et al. (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.

45. Segura,J., Rose,Y., Westbrook,J., Burley,S.K. and Duarte,J.M. (2020) RCSB Protein Data Bank 1D tools and services. *Bioinformatics*, **36**, 5526–5527.

46. Alcock,B.P., Huynh,W., Chalil,R., Smith,K.W., Raphenya,A.R., Włodarski,M.A., Edalatmand,A., Petkau,A., Syed,S.A., Tsang,K.K., et al. (2023) CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Res.*, **51**, D690–D699.

47. Ribeiro,A.J.M., Holliday,G.L., Furnham,N., Tyzack,J.D., Ferris,K. and Thornton,J.M. (2018) Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Res.*, **46**, D618–D623.

48. Hwang,K.Y., Cho,C.S., Kim,S.S., Sung,H.C., Yu,Y.G. and Cho,Y. (1999) Structure and mechanism of glutamate racemase from *Aquifex pyrophilus*. *Nat. Struct. Biol.*, **6**, 422–426.

49. Gene Ontology Consortium (2021) The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.*, **49**, D325–D334.

50. Blum,M., Chang,H.Y., Chuguransky,S., Grego,T., Kandasamy,S., Mitchell,A., Nuka,G., Paysan-Lafosse,T., Qureshi,M., Raj,S., et al. (2021) The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.*, **49**, D344–D354.

51. Nguyen,D.T., Mathias,S., Bologa,C., Brunak,S., Fernandez,N., Gaulton,A., Hersey,A., Holmes,J., Jensen,L.J., Karlsson,A., et al. (2017) Pharos: collating protein information to shed light on the druggable genome. *Nucleic Acids Res.*, **45**, D995–D1002.

52. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Zidek,A., Potapenko,A., et al. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.

53. Humphreys,J.R., Pei,J., Baek,M., Krishnakumar,A., Anishchenko,I., Ovchinnikov,S., Zhang,J., Ness,T.J., Banjade,S., Bagde,S.R., et al. (2021) Computed structures of core eukaryotic protein complexes. *Science*, **374**, eabm4805.

54. Baek,M., DiMaio,F., Anishchenko,I., Dauparas,J., Ovchinnikov,S., Lee,G.R., Wang,J., Cong,Q., Kinch,L.N., Schaeffer,R.D., et al. (2021) Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, **373**, 871–876.

55. Ruperti,F., Papadopoulos,N., Musser,J.M., Mirdita,M., Steinegger,M. and Arendt,D. (2023) Cross-phyla protein annotation by structural prediction and alignment. *Genome Biol.*, **24**, 113.

56. Spinard,E., Azzinaro,P., Rai,A., Espinoza,N., Ramirez-Medina,E., Valladares,A., Borca,M.V. and Gladue,D.P. (2022) Complete structural predictions of the proteome of African Swine Fever virus strain georgia 2007. *Microbiol. Resour. Announc.*, **11**, e0088122.

57. Gao,M., Coletti,M., Davison,R.B., Prout,R., Abraham,S., Hernandez,B. and Sedova,A. (2022) IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW). *IEEE*, **2022**, 206–215.

58. Zhang,J., Pei,J., Durham,J., Bos,T. and Cong,Q. (2022) Computed cancer interactome explains the effects of somatic mutations in cancers. *Protein Sci.*, **31**, e4479.

59. Zardecki,C., Dutta,S., Goodsell,D.S., Lowe,R., Voigt,M. and Burley,S.K. (2022) PDB-101: educational resources supporting molecular explorations through biology and medicine. *Protein Sci.*, **31**, 129–140.