

PDBe tools for an in-depth analysis of small molecules in the Protein Data Bank

Preeti Choudhary  | Ibrahim Roshan Kunnakkattu | Sreenath Nair |
Dare Kayode Lawal | Ivanna Pidruchna | Marcelo Querino Lima Afonso |
Jennifer R. Fleming | Sameer Velankar

Protein Data Bank in Europe, European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK

Correspondence

Sameer Velankar and Preeti Choudhary, Protein Data Bank in Europe, European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK.
Email: sameer@ebi.ac.uk; cypreeti@ebi.ac.uk

Funding information

UKRI-Biotechnology and Biological Sciences Research Council, Grant/Award Number: BB/T01959X/1

Review Editor: Nir Ben-Tal

Abstract

The Protein Data Bank (PDB) is the primary global repository for experimentally determined 3D structures of biological macromolecules and their complexes with ligands, proteins, and nucleic acids. PDB contains over 47,000 unique small molecules bound to the macromolecules. Despite the extensive data available, the complexity of small-molecule data in the PDB necessitates specialized tools for effective analysis and visualization. PDBe has developed a number of tools, including PDBe CCDUtils (<https://github.com/PDBEurope/ccdutils>) for accessing and enriching ligand data, PDBe Arpeggio (<https://github.com/PDBEurope/arpeggio>) for analyzing interactions between ligands and macromolecules, and PDBe RelLig (<https://github.com/PDBEurope/rellig>) for identifying the functional roles of ligands (such as reactants, cofactors, or drug-like molecules) within protein–ligand complexes. The enhanced ligand annotations and data generated by these tools are presented on the novel PDBe-KB ligand pages, offering a comprehensive overview of small molecules and providing valuable insights into their biological contexts (example page for Imatinib: <https://pdbe.org/chem/sti>). By improving the standardization of ligand identification, adding various annotations, and offering advanced visualization capabilities, these tools help researchers navigate the complexities of small molecules and their roles in biological systems, facilitating mechanistic understanding of biological functions. The ongoing enhancements to these resources are designed to support the scientific community in gaining valuable insights into ligands and their applications across various fields, including drug discovery, molecular biology, systems biology, structural biology, and pharmacology.

KEYWORDS

cofactors, drug discovery, drugs, ligand–protein interactions, ligands, PDB, PDBe-KB, python package, RDKit, reactants, similar ligands, small molecules, target validation

1 | INTRODUCTION

Protein Data Bank in Europe (PDBe) is one of the founding partners of the worldwide Protein Data Bank (wwPDB) consortium, dedicated to collecting, curating, and

Preeti Choudhary and Ibrahim Roshan Kunnakkattu contributed equally to this study.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2025 The Author(s). *Protein Science* published by Wiley Periodicals LLC on behalf of The Protein Society.

providing access to a global repository of macromolecular structure models, the Protein Data Bank (PDB) (wwPDB consortium 2019). The PDB contains over 225,000 experimentally determined macromolecular structures, with around 75% featuring at least one bound small molecule. These small molecules serve various purposes, including experimental necessities or acting as ligands with diverse biological functions, like cofactors, metabolites, substrates, and inhibitors. To ensure a standardized representation of small molecules, the wwPDB developed the Chemical Component Dictionary (CCD), a comprehensive reference resource that includes data for all unique chemical components found within PDB entries, including individual amino acids, nucleotides, and ligands (Westbrook et al. 2015). The CCD provides details such as chemical descriptors (e.g., chemical formula, molecular weight, SMILES, and InChI), systematic chemical names, chemical connectivities, stereochemical assignments, and idealized 3D coordinates generated using Molecular Networks' Corina (Schwab 2010) and OpenEye's OMEGA (Hawkins et al. 2010). The CCD is accessible at <https://www.wwpdb.org/data/ccd>. Each unique chemical component is assigned a CCD identifier, allowing for the identification of all instances of a specific small molecule in PDB structures. For instance, adenosine triphosphate has a CCD identifier: ATP, and is bound in over 3600 macromolecular structures.

Complex ligands are often fragmented into individual chemical components during the refinement and biocuration, creating challenges for their identification and mapping to other databases. To address this, the wwPDB introduced the Biologically Interesting Molecule Reference Dictionary (BIRD) in 2013 (Dutta et al. 2014). This manually curated reference dictionary assigns unique identifiers and detailed descriptions to peptide-like inhibitors, antibiotic molecules, and common oligosaccharides found in PDB entries. Each entry details the composition, connectivity, chemical structure, and functions of the reference molecule. The BIRD reference data is accessible at <https://www.wwpdb.org/data/bird>, with examples like Vancomycin, identified by the BIRD identifier: PRD_000204. In 2020, the wwPDB standardized the representation of carbohydrate polymers, which had been previously fragmented into individual monosaccharides, by introducing a new "branched" entity representation (Shao et al. 2021). This improved system offers a more accurate depiction of complex carbohydrate structures and is supported by consistent 2D representations based on the Symbol Nomenclature for Glycans (SNFG) (Neelamegham et al. 2019). While these wwPDB efforts have significantly improved the handling of peptide-like inhibitors, antibiotics, and carbohydrates, several multicomponent ligands remain fragmented into separate components.

To comprehensively tackle complex ligands fragmentation, PDBe introduced covalently linked components (CLCs) (Kunnakkattu et al. 2023). This novel class of reference small molecules identifies ligands composed of multiple covalently linked chemical

components (CCDs) across the entire PDB archive and is found in 5% of the PDB entries. CLCs provide a more complete and accurate representation of these complex ligands, filling gaps left by fragmented CCDs not included in the BIRD or carbohydrate remediation efforts. To streamline identification and analysis of CLCs, PDBe has implemented an automatic process to assign unique identifiers based on InChIKey. For instance, in PDB entry 1D83, there are two carbohydrates. Carbohydrate 1: 2,6-dideoxy-4-O-methyl- α -D-galactopyranose-(1-3)-(2R,3R,6R)-6-hydroxy-2-methyltetrahydro-2H-pyran-3-yl acetate, composed of subcomponent CCDs: CDR, CDR, and ERI. Carbohydrate 2: 3-C-methyl-4-O-acetyl- α -L-olivopyranose-(1-3)-(2R,5S,6R)-6-methyltetrahydro-2H-pyran-2,5-diol-(1-3)-(2R,5S,6R)-6-methyltetrahydro-2H-pyran-2,5-diol, consisting of subcomponent CCDs: ARI and 1GL. These two carbohydrates are covalently linked to the CCD: CPH to form a single small molecule, Chromomycin. This complex ligand, consisting of the covalently linked CCDs 1GL, ARI, CPH, CDR, CDR, and ERI, can now be identified as a single entity using the CLC identifier: CLC_000153 (Figure 1).

Notably, BIRD entries are manually curated, and some CLCs may eventually transition into BIRD entries. For example, in PDB entry 1AO4, Peplomycin—a glycopeptide antineoplastic antibiotic—was previously fragmented into the CCDs PMY, GUP, and 3FM. However, it can now be represented as a single entity with the CLC identifier CLC_000034. If Peplomycin is later classified as a BIRD, CLC_000034 will be automatically replaced by the corresponding BIRD identifier. Creating CLCs to represent chemically complete ligands allows for improved mapping to other chemical databases such as PubChem, ChEMBL, and KEGG, overcoming the limitations of fragmented representation using multiple CCDs. Together, the CCD, BIRD (PRD), and CLC reference dictionaries provide a comprehensive set of unique small molecules found within the PDB, enhancing the understanding and exploration of small molecules bound to biological macromolecules.

The complexity and diversity of small molecules in the PDB necessitate the development of specialized tools for their access and analysis. Although existing small-molecule reference dictionaries provide essential data, additional tools are required to enhance this information within a broader biological context. Furthermore, it is necessary to elucidate the functional roles of ligands within these complexes and distinguish them from experimental artifacts, such as buffers and cryoprotectants. Accurately identifying complete ligand-macromolecule complexes and key protein-ligand interactions is essential for understanding protein functions. Here, we discuss PDBe's strategies to address these challenges, explore PDBe tools, and provide tutorials (Table 1) on their utilization for ligand analysis and visualization. These tools empower researchers to navigate the complexities of ligand and ligand-macromolecule complex structures, assess their

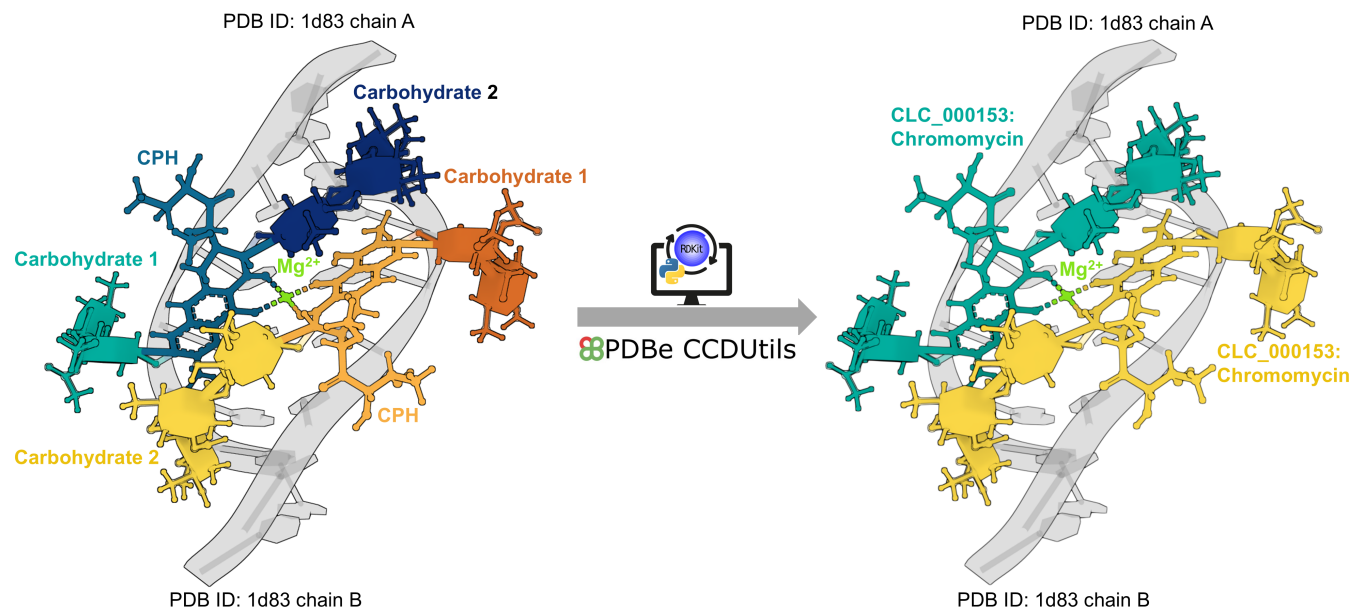


FIGURE 1 Structure of Chromomycin dimer/DNA oligomer complex (PDB ID 1D83). Originally annotated with four separate ligands per chain (two copies each of Carbohydrate 1, Carbohydrate 2, CPH, and Mg), the first three components are now unified as a single Chromomycin ligand (CLC_000153) per chain. The structure features a Chromomycin dimer coordinating an Mg.

TABLE 1 Tutorials demonstrating the use of PDBe tools for in-depth analysis and visualization of ligands in PDB.

Tutorial	Description	Access link
Tutorial 1	Guide on using various PDBe tools for processing and analyzing ligand structure data	https://github.com/PDBEurope/pdbe-notebooks/blob/main/pdbe_ligands_tutorials/PDBe_ligand_tools.ipynb
Tutorial 2	Using PDBe and PDBe-KB webpages to answer various scientific questions	https://github.com/PDBEurope/pdbe-notebooks/blob/main/pdbe_ligands_tutorials/PDBe_PDBe-KB_ligands_webpages_tutorials.pdf
Tutorial 3	Programmatically accessing ligand data using APIs	https://github.com/PDBEurope/pdbe-notebooks/blob/main/pdbe_ligands_tutorials/PDBe_KB_API_tutorials.ipynb

biological significance, and perform comparative analyses across various databases (Figure 2).

2 | PDBe TOOLS FOR PROCESSING LIGAND STRUCTURE DATA AND THEIR ANALYSIS

2.1 | PDBe CCDUtils: Accessing ligands and their enriched data in PDB

PDBe CCDUtils (Kunnakkattu et al. 2023) is an open-source chemistry toolkit designed to improve the

parsing, processing, and analysis of small molecules within the PDB's reference dictionaries, including CCD, PRD, and CLC. Built on RDKit (Landrum et al. 2024), it extends functionality by adding support for the PDBx/mmCIF file format (Westbrook et al. 2022) and provides access to a variety of metadata such as chemical descriptors (SMILES, InChI, InChIKey), 2D depictions, and 3D coordinates. It computes RDKit-derived physicochemical properties, generates 3D conformers, and identifies core chemical substructures, such as scaffolds. A scaffold is the core structural framework of a small molecule that forms the basis for its biological activity, to which functional groups or other modifications can be added. PDBe CCDUtils supports scaffold detection methods like Murcko Scaffolds (Bemis and Murcko 1996) and BRICS (Degen et al. 2008). PDBe CCDUtils can also be used to search ligand substructures against a curated library of over 2000 fragments from PDBe, ENAMINE, and the Diamond-SGC-iNext Poised Library (DSiP) (Cox et al. 2016). Users can additionally use their own external libraries for searches. It can identify similar PDB ligands for any given ligand using pairwise PARITY comparisons (Tyzack et al. 2018) and identify compounds related to a given PDB ligand from over 35 other small-molecule database identifiers, including ChEMBL, CCDC, PubChem, and DrugBank, through the UniChem cross-reference service (Chambers et al. 2013), facilitating the integration and exploration of ligand information across diverse data sources.

To ensure accurate representations of complex biological ligands, such as haem, PDBe CCDUtils incorporates an enhanced data sanitization process that iteratively identifies and rectifies unusual valency

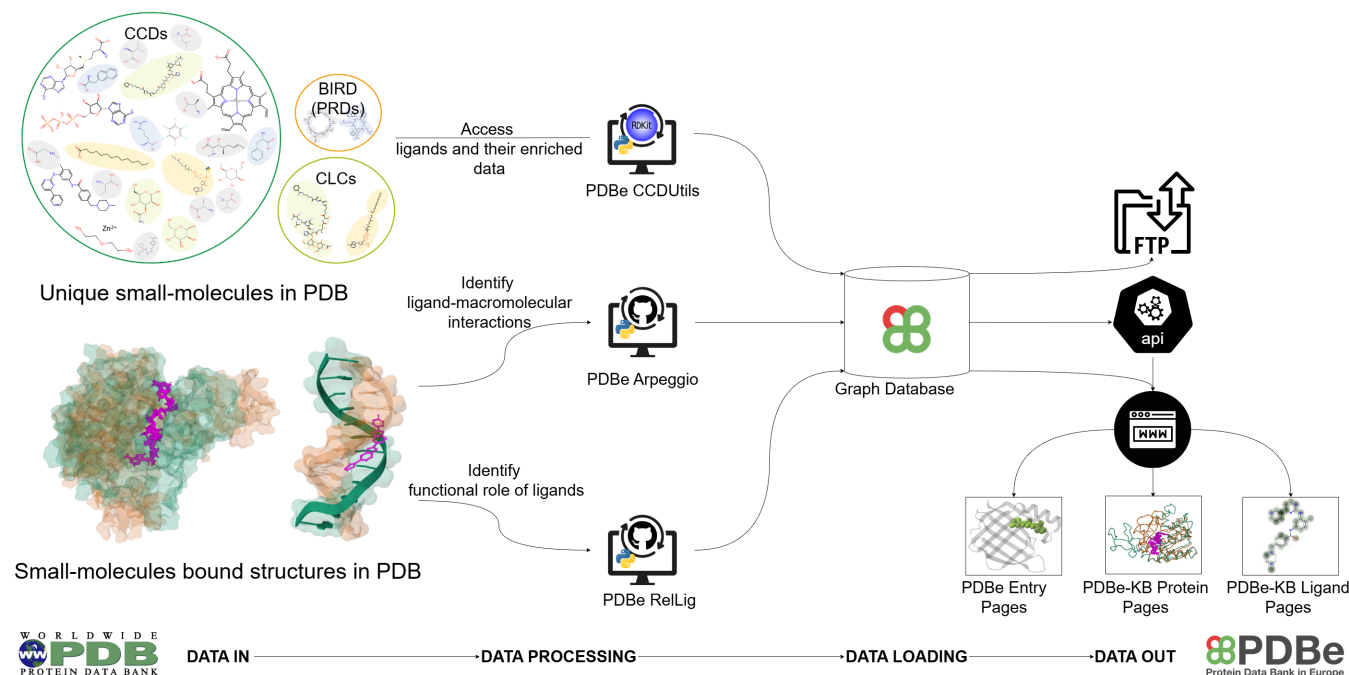


FIGURE 2 Schematic overview of the PDBe ligand tools for ligand analysis and visualization. The PDBe CCDUtils allows access to small molecules, while PDBe Arpeggio identifies ligand–macromolecular interactions, and PDBe RelLig determines ligand functional roles in ligand–protein complexes. PDB data is processed weekly through these pipelines and loaded into the PDBe Graph database, supporting up-to-date data files on FTP, API endpoints, and various web pages, including PDBe entry pages and PDBe-KB Protein and Ligand pages.

issues, adjusting bond types and formal charges as necessary. To generate high-quality 2D ligand depictions, the toolkit implements a Depiction Penalty Score (DPS) to evaluate depiction quality based on bond collisions and suboptimal atom positioning. It compares both template-based and connectivity-based methods to select the best image, ensuring high-quality visual representations of ligands.

Furthermore, PDBe CCDUtils identifies CLC compounds in PDB entries, providing a more complete and precise structural representation for multicomponent ligands represented using multiple CCDs with covalent links. Users can export small-molecule data in various formats, including SDF, CIF, PDB, JSON, XYZ, XML, and CML. All this enriched data generated by PDBe CCDUtils is pre-computed and also made available for download via FTP at https://ftp.ebi.ac.uk/pub/databases/msd/pdbechem_v2/ in the respective CCD, PRD, and CLC folders. The different file formats differ in structure and content, and further details about these files can be found in the readme file available at https://ftp.ebi.ac.uk/pub/databases/msd/pdbechem_v2/README.txt. PDBe CCDUtils is available at <https://github.com/PDBEurope/ccdutils>. With its comprehensive features, PDBe CCDUtils enables researchers to efficiently parse and enrich ligand data, significantly enhancing access to various ligand data in PDB. Tutorial 1.1 showcases how to read and generate enriched data for PDB's small-molecule reference dictionary. It also

demonstrates how to identify and access multi-component ligand systems in a given PDB entry.

2.2 | PDBe Arpeggio: Identifying ligand interactions in PDB

Accurate identification of macromolecule–ligand interactions depends on several key factors, including appropriate protonation, complete macromolecular structure in the form of its observed biological assembly, and complete small-molecule definitions. Determining the ligand protonation state is crucial for precisely defining and classifying its interactions, as protonation affects the ligand's charge and hydrogen bonding potential. Since most PDB structures lack the resolution necessary to model hydrogen atoms directly, a protonation step must be carried out before analyzing ligand interactions. Utilizing the biological assembly provides a functional state of a macromolecule, capturing the full context of interactions. Additionally, considering the entire small molecule, especially in the case of multi-component ligands from the PDB, ensures that all relevant structural details are included, offering a holistic system for interaction analysis.

The PDBe interactions pipeline addresses all these factors by preparing the ligand–macromolecule system before analysis. First, the preferred biological assembly is generated using Mol* Model Server (<https://molstar>).

[org/docs/data-access-tools/model-server/](https://www.ebi.ac.uk/pdbe/docs/data-access-tools/model-server/)). This assembly structure is then protonated using ChimeraX (Meng et al. 2023), representing the entire molecular complex rather than just the asymmetric unit for crystallographic structures. After this, complete bound molecules are inferred based on the connectivity of the non-protein chemical components (CCDs) within the assembly, and each molecule is assigned a unique identifier (bmlID). For each of these bound molecules, interactions are calculated utilizing PDBe Arpeggio. PDBe Arpeggio is a software tool based on the Arpeggio Python library (Jubb et al. 2017), designed to calculate interatomic contacts in macromolecules, including those between ligands and various biomolecules. PDBe Arpeggio extends the original library with features specifically tailored for analyzing structures in the PDB. Interatomic contacts are based on rules defined for CREDO, a protein–ligand interaction database (Schreyer and Blundell 2009), which formed the basis of the Arpeggio Python package. PDBe Arpeggio supports input files in PDBx/mmCIF format, the standard PDB archive format, and categorizes interactions into four types: atom–atom, atom–plane, plane–plane, and plane–group contacts. Detailed molecular interaction information is also available through the PDBe API (https://www.ebi.ac.uk/pdbe/graph-api/pdbe_doc/#api-PDB-GetBoundMoleculeInteractions) for individual atoms in ligands and polymer residues in a PDB entry. Additionally, protonated biological assemblies are accessible via PDBe Static file (https://www.ebi.ac.uk/pdbe/static/entry/download/1cbs_bio_h.cif.gz) and are used for accurate interaction computation. PDBe Arpeggio is an open-source python package available at <https://github.com/PDBEurope/arpeggio>. The tutorial 1.2 presents how the PDBe interactions pipeline prepares the structure and uses PDBe Arpeggio to calculate macromolecule–ligand interactions.

2.3 | PDBe RelLig: Identifying the functional role of ligands in PDB

Identifying ligand roles in protein structures is key to understanding macromolecular functions, enzyme mechanisms, and drug interactions (Burley et al. 2019; Credille et al. 2019; Moreira et al. 2019; Richard 2022; Vetting et al. 2015; Westbrook and Burley 2019). While cofactors and substrates reveal protein activity, others, like buffer components or cryoprotectants, may be introduced during experimental procedures (Caffrey and Cherezov 2009; Garman 2003; Jang et al. 2022; McPherson and Cudney 2014; Peat et al. 2005; Pflugrath 2015). Distinguishing functional ligands from non-functional ones is crucial for accurate biological interpretation. The PDBe RelLig (Relevant Ligands) pipeline addresses this by identifying biologically relevant ligands and categorizing them based on their

functional roles in the entries where they are bound—such as cofactor-like, reactant-like (similar to substrate or product), or drug-like—placing them within their biological context (Figure 3).

2.3.1 | Identifying cofactors-like ligands in PDB

The PDBe RelLig pipeline identifies cofactor-like ligands in the PDB using a semi-automated annotation process based on the 2D structural similarity of ligands to cofactor classes in the CoFactor database (Fischer et al. 2010). The CoFactor database contains 27 classes of manually curated organic enzyme cofactors and details of associated enzymes, including EC numbers. For each cofactor class, we identified a small molecule from the PDB closely matching the structure of the template molecule based on PARITY similarity and assigned it as a representative molecule (Mukhopadhyay et al. 2019). Additionally, we defined a minimum threshold for each cofactor and extended the list of enzyme EC numbers associated with the cofactor classes using the information from the BRENDA database—a comprehensive resource for functional, biochemical, and molecular biological information on enzymes, metabolites, and metabolic pathways (Chang et al. 2021). Using this information, for each newly released small molecule in the PDB, similarity is calculated against the template molecules of each cofactor class. If the similarity meets the minimum threshold of any cofactor class, the ligands are further evaluated against the representative molecule for the matched cofactor class. When the similarity score remains above the threshold and the ligand is present in a PDB entry with an approved EC number corresponding to the cofactor class, the ligand is classified as cofactor-like (Mukhopadhyay et al. 2019). Ligands that do not meet these criteria are flagged for manual annotation, ensuring accurate identification of biologically relevant ligands.

2.3.2 | Identifying reactant-like ligands in PDB

The PDBe RelLig pipeline identifies reactant-like ligands in the PDB by mapping to the Rhea database (Bansal et al. 2022), an expert-curated resource that uses the ChEBI ontology (Hastings et al. 2016) to describe reaction participants and their structures. For each reaction in Rhea, all associated PDB structures are mapped to the UniProt accession of the catalyzing protein. Using the PARITY method, the bound ligands in these PDB structures are compared to ChEBI compounds involved in the reaction. Ligands that achieve a minimum similarity score of 0.7 are annotated as reactant-like, ensuring accurate identification of biologically relevant reactants (substrate or product of the reaction).

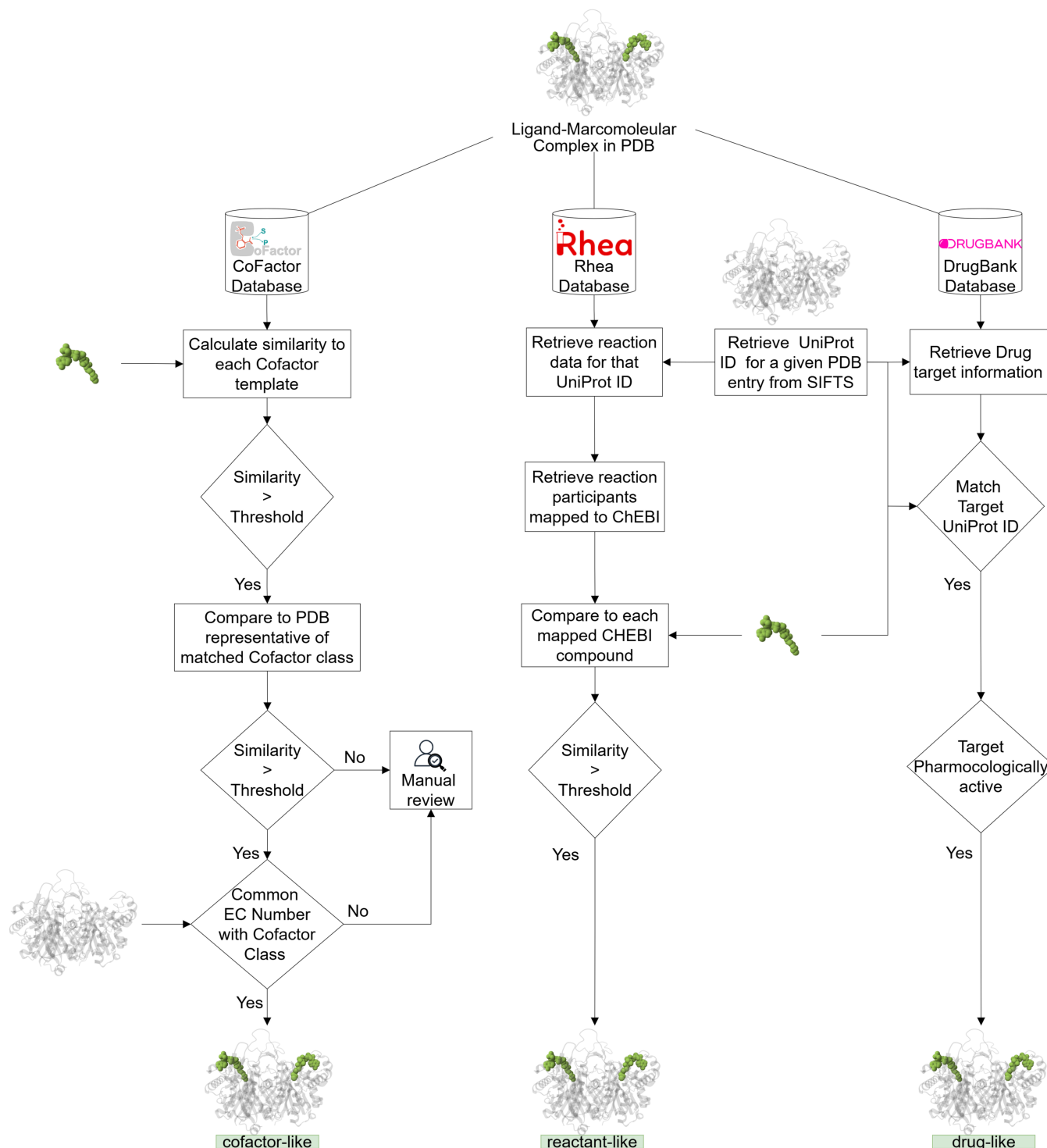


FIGURE 3 Schematic overview of the PDBe RelLig pipeline. This schematic outlines the identification of relevant ligands in the PDB and their categorization based on functional roles within protein structures. Ligands are classified as cofactor-like, reactant-like, or drug-like.

2.3.3 | Identifying drug-like ligands in PDB

The PDBe RelLig pipeline identifies drug-like ligands in the PDB by mapping them to the DrugBank database (Knox et al. 2024). Ligands bound to PDB structures of pharmacologically active targets listed in DrugBank are classified

as drug-like, enabling the accurate annotation of ligands with potential therapeutic relevance. This approach ensures precise identification of biologically and pharmacologically significant drug-like ligands in the PDB.

The PDBe RelLig pipeline is open-source and available at <https://github.com/PDBeurope/rellig>. Tutorial 1.3

demonstrates how the PDBe RelLig pipeline utilizes the PARITY method to identify and classify ligands in PDB structures as reactant-like, cofactor-like, or drug-like.

3 | PDBe WEB PAGES FOR LIGAND VISUALIZATION AND ANALYSIS

For any given structure in the PDB, the “Ligands and Environments” section provides a detailed overview of bound ligands, cofactors, and modified residues, as seen in PDB ID 7bvp (Kim et al. 2020), which includes one cofactor (12 instances), one bound ligand (6 instances), and no modified residue. By clicking on the ligand image or ligand identifier (e.g., CCD ID: NAD), an interactive visualization is launched, showcasing

PDBe Arpeggio calculated residue-level interactions in 2D using the LigEnv web component and atomic-level interactions in 3D between ligands and their binding sites using Mol* (Figure 4). This dual-view visualization is intercommunicative—if a user clicks on a residue in the 2D schematic, it is highlighted in the 3D viewer and vice-versa. This integration facilitates easy analysis of ligand–protein interactions by visualizing spatial and molecular relationships intuitively. The LigEnv web component ensures that ligands are presented in a consistent orientation and layout across different PDB entries, facilitating comparison of the same ligands across various binding sites. Residues forming the ligand binding site are color-coded according to their properties, improving clarity and enhancing the user’s ability to interpret the interactions. Users can export the underlying interaction data in JSON format for computational

7bvp > NAD

NICOTINAMIDE-ADENINE-DINUCLEOTIDE

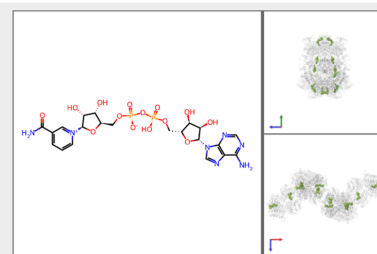
Formula: $C_{21} H_{27} N_7 O_{14} P_2$

Molecular weight: 663 Da

Putative function: Cofactor

Cofactor class: Nicotinamide-adenine dinucleotide

Similarity to cofactor representative (NAD): 1.0



Environment details NEW

NAD 901 bound to chain A ▾

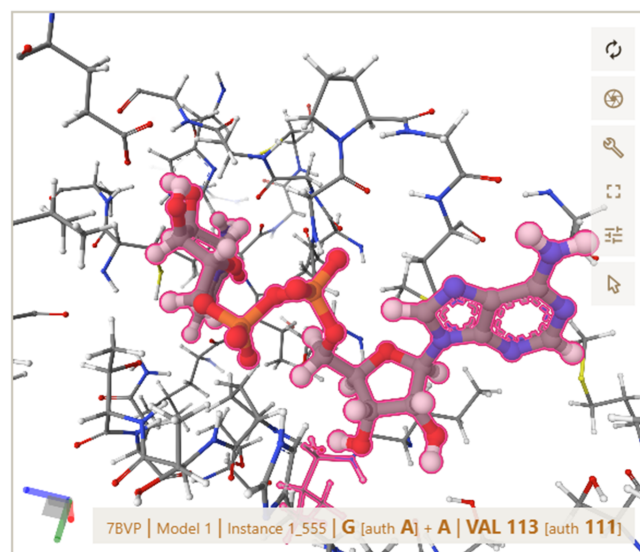
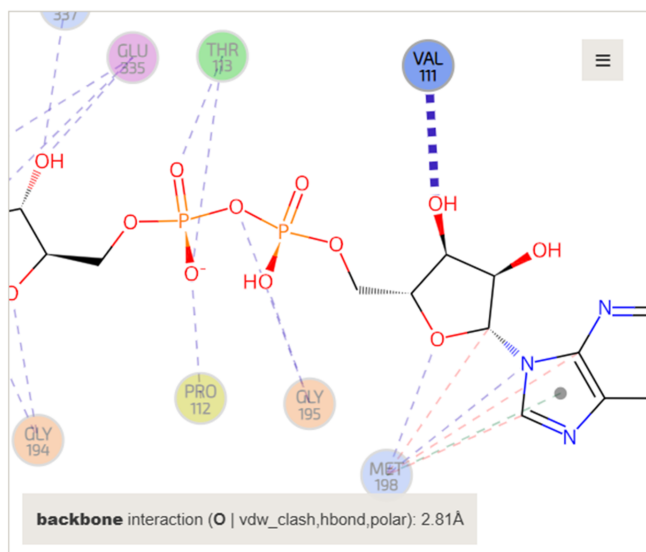


FIGURE 4 Cofactor binding site of aldehyde-alcohol dehydrogenase (PDB ID: 7BVP). The bound ligand NAD is annotated as a cofactor and required for the reduction of the carbonyl group. Hovering over the ligand’s interaction with residue VAL 111 in the LigEnv web component automatically highlights it in the Mol* viewer, enabling easy identification of key binding residues and interactions within the 3D structure. This view is interactive, with movable nodes and edges, allowing users to explore and customize the display as needed. Users can view the color scheme by accessing the help from the menu button in the LigEnv viewer.

analysis or export visual representations in SVG format for documentation. The LigEnv component is an open-source web component available at <https://github.com/PDBEurope/ligand-env/>.

Additionally, the wwPDB validation reports (Feng et al. 2021; Gore et al. 2017), accessible through PDB entry pages, provide essential information on the quality of ligands (Figure 5). Using the program Mogul (Bruno et al. 2004), these reports evaluate ligand geometry against the small-molecule structures in the Cambridge Structural Database (CSD) (Ferrence et al. 2023; Groom et al. 2016). Z-scores are calculated to quantify deviations in bond lengths, angles, torsion angles, and ring geometry, with values above 2.0 flagged as outliers. The root-mean-square value of the Z-scores (RMSZ score), which summarizes the overall ligand geometry, should ideally range between 0 and 1. For torsion angles, deviations are flagged if their local density measure is below 5%, while rings are flagged if their torsion angle RMSD exceeds 60°. Chirality and stereochemistry are also assessed for errors. Beyond geometric validation, ligand fit to the electron density is assessed using metrics like the RSR (real-space *R*-value) and RSCC (real-space correlation coefficient). Ligands with RSCC values below 0.8 or RSR values above 0.4 are highlighted for further scrutiny (Smart et al. 2018).

These validation tools help identify potential issues with ligand modeling, allowing users to compare ligand quality across models and select those with better fit and geometry for more accurate structural representation. Tutorial 2.1 demonstrates how to examine PDB–ligand interactions on PDB entry pages and assess the quality of ligands using wwPDB validation reports.

4 | PDBE-KB: ENABLING LIGAND DATA ANALYSIS IN ITS BIOLOGICAL CONTEXT

The PDBE-Knowledge Base (PDBE-KB), established in 2018 and maintained by the PDBE team at EMBL-EBI, is an open and collaborative data resource dedicated to placing macromolecular structure data within its biological context (PDBE-KB consortium 2022). PDBE-KB consortium partner resources contribute biochemical, biophysical, and functional annotations based on the analysis of the PDB structures. As of 2024, PDBE-KB has integrated contributions from 34 consortium partners, including numerous cheminformatics experts who contribute data on small molecules and their macromolecular binding sites, with collaborators such as canSAR (di Micco et al. 2023), 3DLigandSite (McGreig et al. 2022), P2Rank (Krivák and Hoksza 2018), and many others, as detailed at pdbe-kb.org/partners.

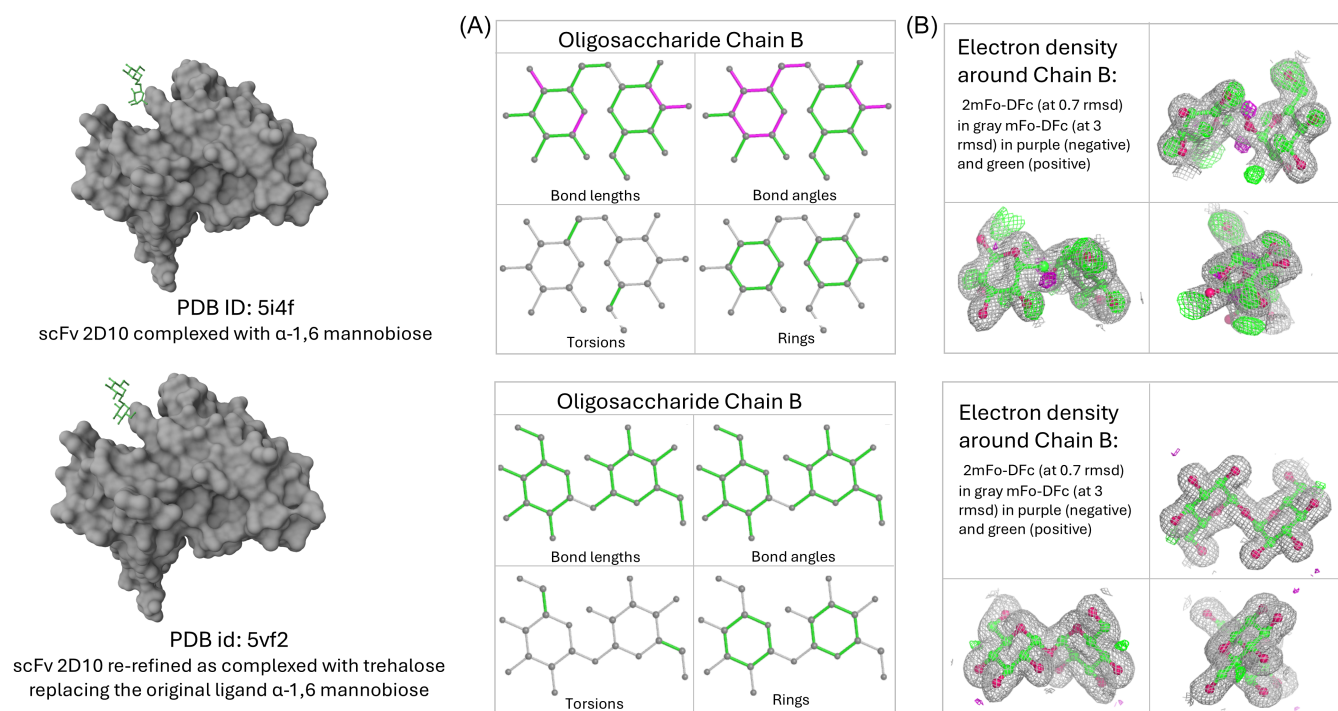


FIGURE 5 Ligand validation assessment of scFv 2D10 structures. Comparison of the original structure bound to α -1,6-mannobiose (PDB ID: 5i4F) and the re-refined structure bound to trehalose (PDB ID: 5VF2) at the same binding site. (a) 2D image of ligand geometric quality. Commonly observed values are shown in green, unusual values in magenta, and features with insufficient data in gray. (b) The 3D view for the ligand atomic model fit the experimental electron density map (shown in gray) with positive and negative difference density maps shown in green and magenta, respectively. The re-refined structure exhibits no geometric outliers and a significantly better fit, indicating improved model quality and confirming the misidentification of the principal ligand in the original structure model.

TABLE 2 Small-molecule-related resources and annotations in PDBe-KB.

Resource name	Annotation type	Number of annotated PDB entries (proteins)	Example PDB entry (PDBe-KB aggregated view)
MetalPDB	Metal binding sites	14,208 (3748)	PDB Id 5czz (https://www.ebi.ac.uk/pdbe/pdbe-kb/proteins/Q9K169/annotations)
M-CSA (mechanism and catalytic site atlas)	Catalytic residues	1004 (918)	PDB id 1bg0 (https://www.ebi.ac.uk/pdbe/pdbe-kb/proteins/P51541/annotations)
canSAR	Predicted druggable pockets	22,917 (22,278)	PDB id 3nbu (https://www.ebi.ac.uk/pdbe/pdbe-kb/proteins/P0A6T1/annotations)
ChannelsDB	Molecular channels	41,314 (7348)	PDB id 246L (https://www.ebi.ac.uk/pdbe/pdbe-kb/proteins/P00720/annotations)
P2rank	Predicted ligand binding sites	192,703 (57,364)	PDB id 246L (https://www.ebi.ac.uk/pdbe/pdbe-kb/proteins/P00720/annotations)
3DLigandSite	Predicted ligand binding sites	916 (478)	PDB id 246L (https://www.ebi.ac.uk/pdbe/pdbe-kb/proteins/P00720/annotations)
CATH-FunSites	Predicted functional sites	20,673 (4785)	PDB id 246L (https://www.ebi.ac.uk/pdbe/pdbe-kb/proteins/P00720/annotations)
Covalentizer	Covalent analogs of bound ligands	1693 (510)	PDB id 5bnr (https://www.ebi.ac.uk/pdbe/pdbe-kb/proteins/P0A6R0/annotations)

These annotations, encompassing curated and predicted data, are crucial for understanding the roles and mechanisms of small molecules interacting with proteins and nucleic acids. Examples of these annotations include catalytic sites, biologically relevant metal ions, druggable pockets, and predicted ligand binding sites. Table 2 provides more details on various ligand resources available in PDBe-KB.

PDBe-KB aims to consolidate the wealth of information and enable efficient use of structural data by aggregating all the data in a knowledge graph for a given biological entity (i.e., complexes, proteins, domains, ligands). The aggregated protein view, the first PDBe-KB webpage, was launched in March 2019 and is being actively developed with new annotations and features.

4.1 | Ligand annotations on PDBe-KB protein pages

The PDBe-KB aggregated views of proteins integrate data for a given protein (based on common UniProt accession) from multiple PDB entries to present a comprehensive web-based visualization of all the relevant structural data. These web pages feature a sub-page showcasing an integrated view, “Ligands and Environments,” of all small molecules from all relevant entries from the PDB archive that interact with the given protein, along with annotations regarding their functional roles, such as drugs, cofactors, or reactants identified using the PDBe RelLig pipeline (Figure 6a). The ligands are organized based on their chemical scaffolds, enhancing visual comparison and accessibility. Each entry in the gallery displays the chemical structure, its CCD ID, recommended name, and additional information. Users can explore detailed 3D representations of

ligand interactions through the Mol* 3D viewer by clicking on the ligand image and by downloading relevant PDB entries in CSV or JSON formats using the download button.

The “Ligands-binding Residues” section provides a comprehensive view of ligand binding residues from all relevant PDB entries containing the given protein, using a 2D sequence viewer, PDBe ProtVista (Figure 6b). This viewer uses darker shades of blue to indicate frequently interacting residues, enabling researchers to gain insights into key binding residues and sites. A collapsed view provides a summary that can help in identifying binding sites that bind a diverse set of small molecules.

Furthermore, users can access a superposed view of small molecules from various PDB entries for specific protein residue ranges (“segment”), highlighting all binding pockets. By clicking the “3D view of superposed ligands” button, users can visualize representative conformations for a given protein (Ellaway et al. 2024; PDBe-KB consortium 2022) alongside all ligands associated with that segment (Figure 6c), which is particularly beneficial for analyzing proteins identified in fragment screening experiments. Collectively, these features enhance the understanding of ligand–protein interactions and support more accurate structural representation and analysis. Tutorial 2.2 illustrates how to identify key binding residues on PDBe-KB aggregated views of proteins and locate important ligand binding sites using a 3D view of superimposed ligands.

4.2 | PDBe-KB ligand pages

In the next phase of PDBe-KB development, an aggregated view for individual ligands has been created to

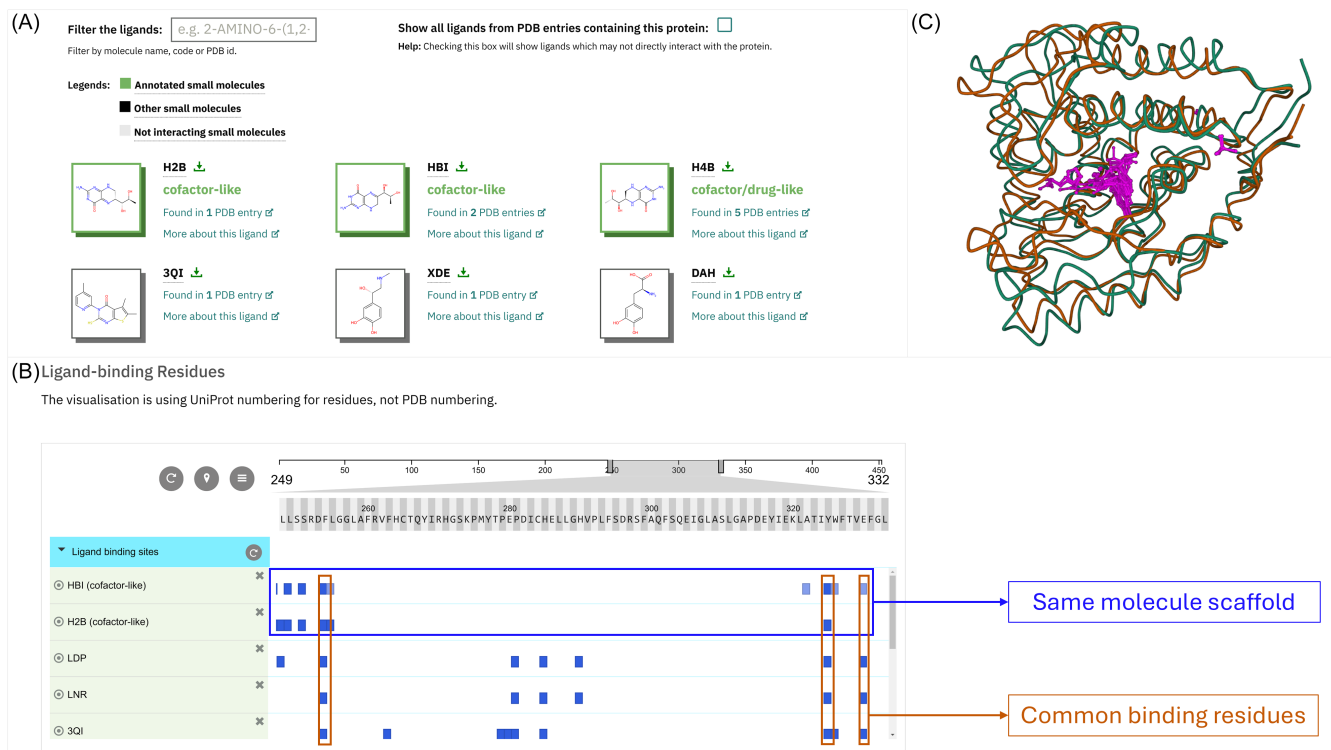


FIGURE 6 Ligand data for Phenylalanine-4-hydroxylase (P00439) on PDBE-KB Protein Page. (a) Ligand gallery displaying all ligands bound to Phenylalanine-4-hydroxylase, with cofactors highlighted in green. (b) PDBE ProtVista viewer displaying ligand-binding residues, facilitating easy identification of common binding residues and those specific to cofactors H2B, HBI, and H4B, which share a common molecular scaffold. (c) 3D superposition of ligands, highlighting the most prevalent binding pockets.

consolidate data on small molecules from the Protein Data Bank (PDB) and provide enhanced biological context. By leveraging the integrated PDBE-KB knowledge graph, this view simplifies the traditionally labor-intensive process of linking and analyzing macromolecular interactions with small molecules. These ligand-centric pages, keyed on PDB ligand identifiers such as CCD, PRD, and CLC IDs, present a comprehensive overview of each ligand's chemical, structural, interaction, and functional information in a user-friendly and intuitive manner.

The essential ligand features, including molecular name, synonyms, molecular formula, and chemical descriptors such as IUPAC InChI, InChIKey, and SMILES, are displayed in the “Description” tab. Structural representations are provided in 2D and 3D formats, with distinct views highlighting the overall ligand structure, individual atoms, the Murcko scaffold, and specific ligand fragments. Physicochemical properties, categorized into molecular, ring, conformational, surface, functional group, and stereochemical properties, are also shown. Additionally, if the CCD is part of a larger BIRD (PRD) or CLC entry, this information is also shown here. All data in this tab is generated using PDBE CCDUtils.

The “Structures” tab provides access to all PDB entries, where ligand–protein interactions for a given

ligand are observed in a table format. The table includes key details such as the protein name, UniProt accession (with link to PDBE-KB aggregated view for that protein), total number of PDB structures, source organism, EC number, and the functional role of the ligand (e.g., drug-like, cofactor-like, or reactant-like) as determined by the PDBE RelLig pipeline. Since the same protein may have multiple PDB structures resolved under different experimental conditions or with mutations, the data is grouped by default in the “Proteins” view, showing the total number of PDB structures. Users can switch to the “Structures” view to see all the individual PDB entries and easily find relevant protein–ligand complexes.

The “Interaction statistics” tab aggregates atom-level interaction data for all instances of the ligand bound to proteins in the PDB. A heatmap displays the percentage of relative interaction frequency for each ligand atom with respect to the total number of interactions by the ligand, with darker colors indicating more frequent interactions (Figure 7). This data is also mapped onto a 2D ligand image, providing an overview of interaction hotspots. The heatmap and ligand image are interactive, allowing users to highlight corresponding atoms across both views simultaneously. Additionally, the heatmap shows the percentage of relative interaction frequency of each ligand atom–amino acid

Interaction statistics

Interaction statistics shows the summary of aggregated protein-ligand interaction data of STI from 32 ligand instances in 16 PDB structures and 28 PDB chains. The protein-ligand interactions are computed using [PDBe Arpeggio](#)

[Download all interactions](#) [Documentation](#)

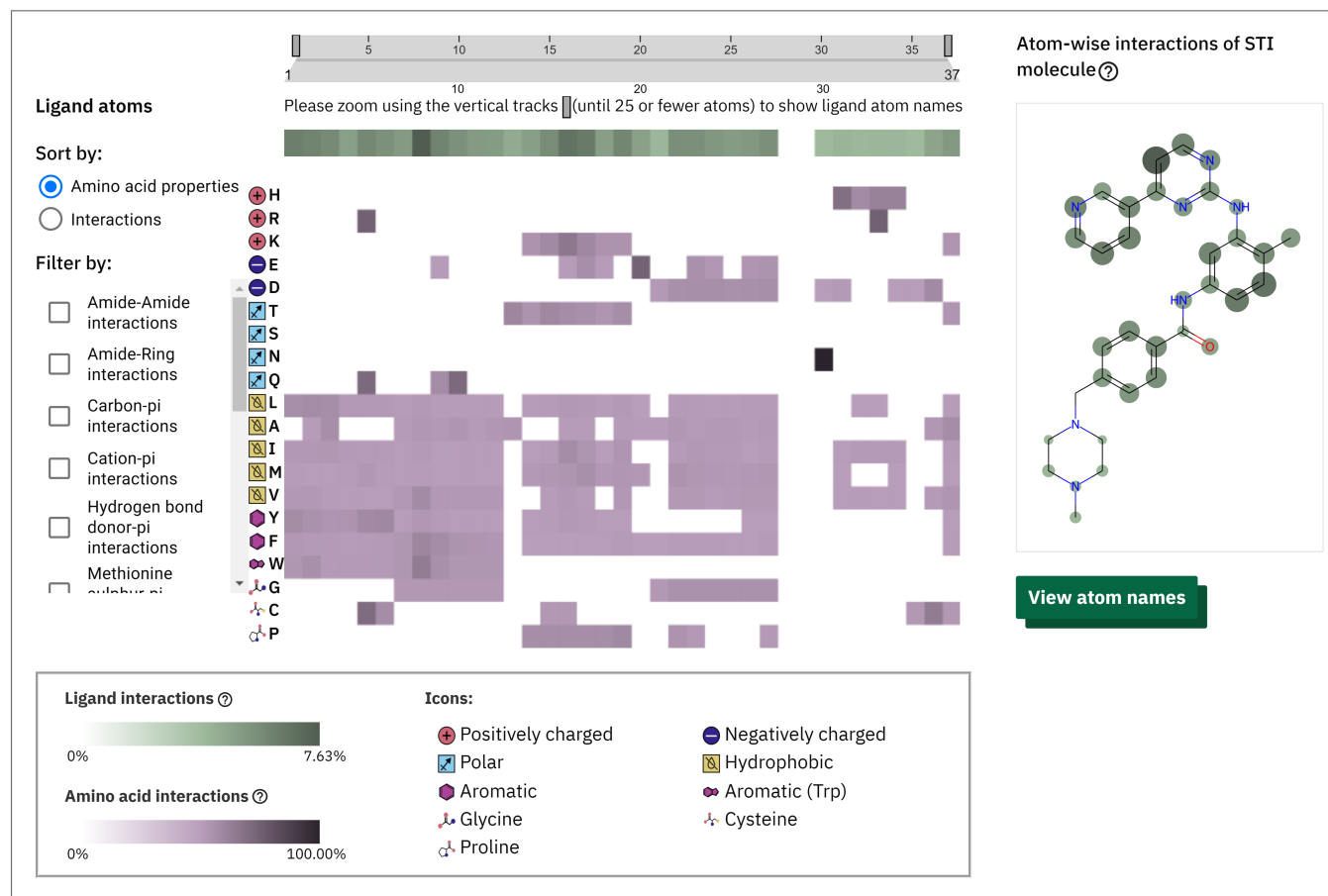


FIGURE 7 Aggregated interaction patterns for Imatinib (CCD ID: STI) on PDBe-KB ligand pages. Imatinib primarily interacts with hydrophobic residues, along with a few aromatic residues as illustrated on the heatmap. Key interaction hotspots for Imatinib are highlighted on its 2D image, showing that interactions primarily occur through its phenyl, pyridine, and pyrimidine rings. Source: <https://pdbe.org/chem/sti>

pairs with respect to the total number of interactions of the amino acid, indicating the likelihood of a given ligand atom to interact with a specific amino acid. The relative interaction frequencies can be sorted by either amino acid types or their values, offering insights into whether a ligand preferentially interacts with certain types of residues and allowing for quick identification of the most frequently interacting amino acids. Interaction data can also be filtered by interaction types, such as hydrogen bonds, Van der Waals, polar, or hydrophobic interactions.

The “Related ligands” tab provides access to structurally similar ligands, identified using the PDBe RelLig pipeline. The similar ligands are organized into three subsections based on the type of similarity: ligands with the same scaffold, those with $\geq 60\%$ similarity (similarity score ≥ 0.6 by PARITY method), and any stereoisomers. Each subsection features an image gallery displaying

the ligand structures, with a 3D view of the ligand upon clicking an image. Additionally, details such as the ligand name, CCD ID, percentage similarity, and the number of PDB structures in which the similar ligands are found are provided, along with a link to the PDBe search page. This search page lists all relevant PDB structures and includes various filtering options on the left menu, such as experimental method, source organism, sequence, structural classification, and so on.

Finally, the “Ligand-specific databases” offer cross-references to other small-molecule data resources generated using UniChem integration in PDBe CCDUtils. A data download functionality is integrated within each section for easy access to data for further analysis. By integrating chemical, structural, interaction, and functional data, the PDBe-KB Ligand Pages provide an aggregated view of small molecules, enhancing the accessibility and biological relevance of ligand data

within the PDB. Tutorial 2.3 showcases how to answer scientific questions using PDBe-KB ligand pages. Users can access PDBe-KB ligand pages in the following ways: 1) by searching for a ligand using the PDBe-Chem landing pages (<https://pdbe.org/chem>), 2) by exploring ligand details from the ligand gallery shown on PDBe-KB protein page for a given protein and clicking “More about this ligand” (e.g: <https://www.ebi.ac.uk/pdbe/pdbe-kb/proteins/P00720/ligands>), and 3) by directly accessing the PDBe-KB ligand page if they know the PDB ligand ID, using the URL format: https://pdbe.org/chem/{ligand_id} (replacing {ligand_id} with the actual ligand ID which can be CCD, PRD or CLC ID).

4.3 | PDBe-KB programmatic data access

All the data displayed on PDBe and PDBe-KB web pages are served by REST API endpoints, enabling users to access this information programmatically (Nair et al. 2021). These API endpoints are organized into categories such as compounds, proteins (UniProt), residues, PDB entries, and validation for efficient navigation. The compound-related endpoints provide information on all PDB entries containing specific ligands, details about their atoms and bonds, similar ligands, ligand substructures, and functional annotations. Additionally, they offer summary data, including descriptors such as InChI, InChIKey, SMILES, and physicochemical properties. A complete list of API endpoints is available at https://www.ebi.ac.uk/pdbe/graph-api/pdbe_doc. These API endpoints query the PDBe graph database to retrieve the necessary information. The PDBe graph database is freely available for download at <https://www.ebi.ac.uk/pdbe/pdbe-kb/graph>, enabling users to conduct advanced queries and custom analyses. Tutorial 3 illustrates how to programmatically access ligand data in PDBe and PDBe-KB through various API endpoints.

4.4 | Summary files for comprehensive analysis

PDBe offers summary files to facilitate extensive analyses of ligand interactions in the PDB archive. Two essential files are `Interacting_chains_with_ligand_functions.tsv` and `pdbe_bound_molecules.tsv`, are available at https://ftp.ebi.ac.uk/pub/databases/msd/pdbechem/v2/additional_data/pdb_ligand_interactions/. The first file summarizes ligand interactions, detailing interacting macromolecule chains, their UniProt accessions, functional annotations, and identifiers such as InChIKey, bmlID, and LigandType. The second file contains information on CLC molecules and their composition.

The ChEMBL and PDBe teams have mapped over 17,000 PDB–ligand complexes to approximately 39,000 bioactivity records, encompassing various experimental data such as binding affinities and functional assays. These results are available at https://ftp.ebi.ac.uk/pub/databases/msd/pdbechem_v2/additional_data/bioactivity_reports/, including a comprehensive report and a simplified version with aggregated values for each target–ligand complex and direct links for further exploration of ChEMBL data.

5 | DISCUSSION

Identifying and accurately representing ligands is crucial for understanding protein–ligand interactions, which have wide-ranging applications in understanding protein function, target validation, drug development, and repurposing. With the introduction of CLCs, PDBe enables chemically complete and precise representations of multicomponent ligands. The accurate identification and data standardization of small molecules via PDBe CCDUtils have enabled the mapping of many previously overlooked small molecules from the PDB to other databases. This process has significantly enhanced the completeness and accuracy of ligand definitions and facilitated the integration of ligand data across these resources and, as a result, facilitates, for example, the identification of relevant ligand–target complexes. The PDBe RelLig pipeline provides annotations that clarify the functional roles of ligands in their targets, helping researchers distinguish biologically significant molecules from experimental artifacts and streamlining the analysis process.

Moreover, the integrated information on PDBe ligand pages enables users to view small-molecule data within a comprehensive structural context. This holistic view encompasses ligand descriptions, physicochemical properties, ligand–protein complexes, and the functional roles of ligands within those complexes. Additionally, it provides overall binding interaction statistics, related ligands, and cross-links to other small-molecule databases. This wealth of information aids researchers in identifying critical patterns in ligand interactions, understanding these molecules’ biological significance, and exploring potential applications in various fields, such as drug discovery, enzyme design, and biomolecular research.

6 | CONCLUSIONS

The PDB contains a wealth of information on small molecules and their interactions with macromolecules, but specialized tools are essential to fully leverage this data. PDBe has developed several open-access tools—PDBe CCDUtils, PDBe Arpeggio, and PDBe

RelLig—that enrich ligand data, enable detailed interaction analysis, and classify ligands based on their biological roles. The PDBe-KB ligand pages also provide an integrated platform for accessing and visualizing this enriched data. Together, these resources empower researchers to efficiently explore the properties of small molecules, analyze their interactions with proteins, and distinguish between biologically relevant ligands and experimental artifacts. By integrating these tools and resources into the PDBe ecosystem, we provide a comprehensive platform for analyzing and visualizing small molecules within the PDB, greatly enhancing understanding of their biological context and functional significance to facilitate basic and translational research.

AUTHOR CONTRIBUTIONS

Preeti Choudhary: Conceptualization; project administration; supervision; writing – original draft; software; formal analysis; visualization. **Ibrahim Roshan Kunakkattu:** Software; methodology; formal analysis; visualization; writing – review and editing. **Sreenath Nair:** Software; supervision. **Dare Kayode Lawal:** Software; visualization. **Ivanna Pidruchna:** Visualization. **Marcelo Querino Lima Afonso:** Software; visualization. **Jennifer R. Fleming:** Supervision; conceptualization; writing – review and editing. **Sameer Velankar:** Conceptualization; funding acquisition; investigation; resources; supervision; writing – review and editing.


ACKNOWLEDGMENTS

The authors would like to express their gratitude to the UKRI-Biotechnology and Biological Sciences Research Council for funding provided under the BioChemGraph project (BB/T01959X/1) and to the European Molecular Biology Laboratory-European Bioinformatics Institute for their support. Special thanks are extended to collaborators from ChEMBL: Melissa F. Adasme, James Blackshaw, and Andrew Leach, as well as from CCDC: David Lowe and Ian Bruno. Open Access funding enabled and organized by Projekt DEAL.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in PDBe small-molecules FTP area at https://ftp.ebi.ac.uk/pub/databases/msd/pdbechem_v2/.

ORCID

Preeti Choudhary  <https://orcid.org/0000-0003-2340-3278>

REFERENCES

- Bansal P, Morgat A, Axelsen KB, Muthukrishnan V, Coudert E, Aimo L, et al. Rhea, the reaction knowledgebase in 2022. *Nucleic Acids Res.* 2022;50(D1):D693–700. <https://doi.org/10.1093/nar/gkab1016>
- Bemis GW, Murcko MA. The properties of known drugs. 1. Molecular frameworks. *J Med Chem.* 1996;39(15):2887–93. <https://doi.org/10.1021/jm9602928>
- Bruno IJ, Cole JC, Kessler M, Jie Luo WD, Motherwell S, Purkis LH, et al. Retrieval of crystallographically-derived molecular geometry information. *J Chem Inf Comput Sci.* 2004;44(6):2133–44. <https://doi.org/10.1021/ci049780b>
- Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, Di Costanzo L, et al. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.* 2019;47(D1):D464–74. <https://doi.org/10.1093/nar/gky1004>
- Caffrey M, Cherezov V. Crystallizing membrane proteins using lipidic mesophases. *Nat Protoc.* 2009;4(5):706–31. <https://doi.org/10.1038/nprot.2009.31>
- Chambers J, Davies M, Gaulton A, Hersey A, Velankar S, Petryszak R, et al. UniChem: a unified chemical structure cross-referencing and identifier tracking system. *J Chem.* 2013;5(1):3. <https://doi.org/10.1186/1758-2946-5-3>
- Chang A, Jeske L, Ulbrich S, Hofmann J, Koblitz J, Schomburg I, et al. BRENDA, the ELIXIR Core Data Resource in 2021: new developments and updates. *Nucleic Acids Res.* 2021;49(D1):D498–508. <https://doi.org/10.1093/nar/gkaa1025>
- Cox OB, Krojer T, Collins P, Monteiro O, Talon R, Bradley A, et al. A poised fragment library enables rapid synthetic expansion yielding the first reported inhibitors of PHIP(2), an atypical bromodomain. *Chem Sci.* 2016;7(3):2322–30. <https://doi.org/10.1039/C5SC03115J>
- Credille CV, Morrison CN, Stokes RW, Dick BL, Feng Y, Sun J, et al. SAR exploration of tight-binding inhibitors of influenza virus PA endonuclease. *J Med Chem.* 2019;62(21):9438–49. <https://doi.org/10.1021/acs.jmedchem.9b00747>
- Degen J, Wegscheid-Gerlach C, Zaliani A, Rarey M. On the art of compiling and using “drug-like” chemical fragment spaces. *Chem Med Chem.* 2008;3(10):1503–7. <https://doi.org/10.1002/cmdc.200800178>
- di Micco P, Antolin AA, Mitsopoulos C, Villasclaras-Fernandez E, Sanfelice D, Dolciemi D, et al. canSAR: update to the cancer translational research and drug discovery knowledgebase. *Nucleic Acids Res.* 2023;51(D1):D1212–9. <https://doi.org/10.1093/nar/gkac1004>
- Dutta S, Dimitropoulos D, Feng Z, Persikova I, Sen S, Shao C, et al. Improving the representation of peptide-like inhibitor and antibiotic molecules in the Protein Data Bank. *Biopolymers.* 2014; 101(6):659–68. <https://doi.org/10.1002/bip.22434>
- Ellaway JIJ, Anyango S, Nair S, Zaki HA, Nadzirin N, Powell HR, et al. Identifying protein conformational states in the Protein Data Bank: toward unlocking the potential of integrative dynamics studies. *Struct Dyn.* 2024;11(3):034701. <https://doi.org/10.1063/4.0000251>
- Feng Z, Westbrook JD, Sala R, Smart OS, Bricogne G, Matsubara M, et al. Enhanced validation of small-molecule ligands and carbohydrates in the Protein Data Bank. *Structure.* 2021;29(4):393–400.e1. <https://doi.org/10.1016/j.str.2021.02.004>
- Ferrence GM, Tovee CA, Holgate SJW, Johnson NT, Lightfoot MP, Nowakowska-Orzechowska KL, et al. CSD communications of the cambridge structural database. *IUCrJ.* 2023;10:6–15. <https://doi.org/10.1107/S2052252522010545>
- Fischer JD, Holliday GL, Thornton JM. The CoFactor database: organic cofactors in enzyme catalysis. *Bioinformatics.* 2010; 26(19):2496–7. <https://doi.org/10.1093/bioinformatics/btq442>
- Garman E. “Cool” crystals: macromolecular cryocrystallography and radiation damage. *Curr Opin Struct Biol.* 2003;13(5):545–51. <https://doi.org/10.1016/j.sbi.2003.09.013>
- Gore S, García ES, Hendrickx PMS, Gutmanas A, Westbrook JD, Yang H, et al. Validation of structures in the Protein Data Bank. *Structure.* 2017;25(12):1916–27. <https://doi.org/10.1016/j.str.2017.10.009>

- Groom CR, Bruno IJ, Lightfoot MP, Ward SC. The Cambridge structural database. *Acta Crystallogr Sect B Struct Sci, Cryst Eng Mater.* 2016;72(Pt 2):171–9. <https://doi.org/10.1107/S2052520616003954>
- Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, et al. ChEBI in 2016: improved services and an expanding collection of metabolites. *Nucleic Acids Res.* 2016;44(D1):D1214–9. <https://doi.org/10.1093/nar/gkv1031>
- Hawkins PCD, Geoffrey Skillman A, Warren GL, Ellingson BA, Stahl MT. Conformer generation with OMEGA: algorithm and validation using high quality structures from the protein databank and Cambridge structural database. *J Chem Inf Model.* 2010;50(4):572–84. <https://doi.org/10.1021/ci100031x>
- Jang K, Kim HG, Hlaing SHS, Kang MS, Choe H-W, Kim YJ. A short review on cryoprotectants for 3D protein structure analysis. *Crystals.* 2022;12(2):138. <https://doi.org/10.3390/cryst12020138>
- Jubb HC, Higuero AP, Ochoa-Montano B, Pitt WR, Ascher DB, Blundell TL. Arpeggio: a web server for calculating and Visualising interatomic interactions in protein structures. *Comput Resour Mol Biol.* 2017;429(3):365–71. <https://doi.org/10.1016/j.jmb.2016.12.004>
- Kim G, Yang J, Jang J, Choi J-S, Roe AJ, Byron O, et al. Aldehyde-alcohol dehydrogenase undergoes structural transition to form extended Spirosomes for substrate channeling. *Commun Biol.* 2020;3(1):298. <https://doi.org/10.1038/s42003-020-1030-1>
- Knox C, Wilson M, Klinger CM, Franklin M, Oler E, Wilson A, et al. DrugBank 6.0: the DrugBank knowledgebase for 2024. *Nucleic Acids Res.* 2024;52(D1):D1265–75. <https://doi.org/10.1093/nar/gkad976>
- Krivák R, Hoksza D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *J Chem.* 2018;10(1):39. <https://doi.org/10.1186/s13321-018-0285-8>
- Kunnakkattu IR, Choudhary P, Pravda L, Nadzirin N, Smart OS, Yuan Q, et al. PDBe CCDUtils: an RDKit-based toolkit for handling and analysing small molecules in the Protein Data Bank. *J Chem.* 2023;15(1):117. <https://doi.org/10.1186/s13321-023-00786-w>
- Landrum G, Tosco P, Kelley B, Rodriguez R, Cosgrove D, Vianello R, et al. Rdkit/Rdkit: 2024_09_2 (Q3 2024) Release. Zenodo. 2024 <https://doi.org/10.5281/zenodo.13990314>
- McGreig JE, Uri H, Antczak M, Sternberg MJE, Michaelis M, Wass MN. 3DLigandSite: structure-based prediction of protein-ligand binding sites. *Nucleic Acids Res.* 2022;50(W1):W13–20. <https://doi.org/10.1093/nar/gkac250>
- McPherson A, Cudney B. Optimization of crystallization conditions for biological macromolecules. *Acta Crystallogr Sect F Struct Biol Commun.* 2014;70:1445–67. <https://doi.org/10.1107/S2053230X14019670>
- Meng EC, Goddard TD, Pettersen EF, Couch GS, Pearson ZJ, Morris JH, et al. UCSF ChimeraX: tools for structure building and analysis. *Protein Sci.* 2023;32(11):e4792. <https://doi.org/10.1002/pro.4792>
- Moreira C, Calixto AR, Richard JP, Kamerlin SCL. The role of ligand-gated conformational changes in enzyme catalysis. *Biochem Soc Trans.* 2019;47(5):1449–60. <https://doi.org/10.1042/BST20190298>
- Mukhopadhyay A, Borkakoti N, Pravda L, Tyzack JD, Thornton JM, Velankar S. Finding enzyme cofactors in Protein Data Bank. *Bioinformatics (Oxford, England).* 2019;35(18):3510–1. <https://doi.org/10.1093/bioinformatics/btz115>
- Nair S, Váradi M, Nadzirin N, Pravda L, Anyango S, Mir S, et al. PDBe aggregated API: programmatic access to an integrative knowledge graph of molecular structure data. *Bioinformatics.* 2021;37(21):3950–2. <https://doi.org/10.1093/bioinformatics/btab424>
- Neelamegham S, Aoki-Kinoshita K, Bolton E, Frank M, Lisacek F, Lütke T, et al. Updates to the symbol nomenclature for glycans guidelines. *Glycobiology.* 2019;29(9):620–4. <https://doi.org/10.1093/glycob/cwz045>
- PDBe-KB consortium. PDBe-KB: collaboratively defining the biological context of structural data. *Nucleic Acids Res.* 2022;50(D1):D534–42. <https://doi.org/10.1093/nar/gkab988>
- Peat TS, Christopher JA, Newman J. Tapping the Protein Data Bank for crystallization information. *Acta Crystallogr Sect D.* 2005;61(12):1662–9. <https://doi.org/10.1107/S0907444905033202>
- Pflugrath JW. Practical macromolecular cryocrystallography. *Acta Crystallogr Sect F Struct Biol Commun.* 2015;71:622–42. <https://doi.org/10.1107/S2053230X15008304>
- Richard JP. Enabling role of ligand-driven conformational changes in enzyme evolution. *Biochemistry.* 2022;61(15):1533–42. <https://doi.org/10.1021/acs.biochem.2c00178>
- Schreyer A, Blundell T. CREDO: a protein-ligand interaction database for drug discovery. *Chem Biol Drug des.* 2009;73(2):157–67. <https://doi.org/10.1111/j.1747-0285.2008.00762.x>
- Schwab CH. Conformations and 3D pharmacophore searching. *Drug Discovery Today Technol.* 2010;7(4):e245–53. <https://doi.org/10.1016/j.ddtec.2010.10.003>
- Shao C, Feng Z, Westbrook JD, Peisach E, Berrisford J, Ikegawa Y, et al. Modernized uniform representation of carbohydrate molecules in the Protein Data Bank. *Glycobiology.* 2021;31(9):1204–18. <https://doi.org/10.1093/glycob/cwab039>
- Smart OS, Horský V, Gore S, Vařeková RS, Bendová V, Kleywegt GJ, et al. Validation of ligands in macromolecular structures determined by X-ray crystallography. *Acta Crystallogr Sect D.* 2018;74(3):228–36. <https://doi.org/10.1107/S2059798318002541>
- Tyzack JD, Fernando L, Ribeiro AJM, Borkakoti N, Thornton JM. Ranking enzyme structures in the PDB by bound ligand similarity to biological substrates. *Structure.* 2018;26(4):565–571.e3. <https://doi.org/10.1016/j.str.2018.02.009>
- Vetting MW, Al-Obaidi N, Zhao S, Francisco BS, Kim J, Wichelecki DJ, et al. Experimental strategies for functional annotation and metabolism discovery: targeted screening of solute binding proteins and unbiased panning of metabolomes. *Biochemistry.* 2015;54(3):909–31. <https://doi.org/10.1021/bi501388y>
- Westbrook JD, Burley SK. How structural biologists and the Protein Data Bank contributed to recent FDA new drug approvals. *Structure.* 2019;27(2):211–7. <https://doi.org/10.1016/j.str.2018.11.007>
- Westbrook JD, Shao C, Feng Z, Zhuravleva M, Velankar S, Young J. The chemical component dictionary: complete descriptions of constituent molecules in experimentally determined 3D macromolecules in the Protein Data Bank. *Bioinformatics.* 2015;31(8):1274–8. <https://doi.org/10.1093/bioinformatics/btu789>
- Westbrook JD, Young JY, Shao C, Feng Z, Guranovic V, Lawson CL, et al. PDBx/mmCIF ecosystem: foundational semantic tools for structural biology. *J Mol Biol.* 2022;434(11):167599. <https://doi.org/10.1016/j.jmb.2022.167599>
- wwPDB consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* 2019;47(D1):D520–8. <https://doi.org/10.1093/nar/gky949>

How to cite this article: Choudhary P, Kunnakkattu IR, Nair S, Lawal DK, Pidruchna I, Afonso MQL, et al. PDBe tools for an in-depth analysis of small molecules in the Protein Data Bank. *Protein Science.* 2025;34(4):e70084. <https://doi.org/10.1002/pro.70084>