www.nature.com/scientificdata

scientific **data**

Check for updates

OPEN

ARTICLE

Unified access to up-to-date residue-level annotations from UniProtKB and other biological databases for PDB data

Preeti Choudhary[1] ✉, Stephen Anyango[1], John Berrisford[1,2], James Tolchard[1,3], Mihaly Varadi[1] & Sameer Velankar[1]

More than 61,000 proteins have up-to-date correspondence between their amino acid sequence (UniProtKB) and their 3D structures (PDB), enabled by the Structure Integration with Function, Taxonomy and Sequences (SIFTS) resource. SIFTS incorporates residue-level annotations from many other biological resources. SIFTS data is available in various formats like XML, CSV and TSV format or also accessible via the PDBe REST API but always maintained separately from the structure data (PDBx/mmCIF file) in the PDB archive. Here, we extended the wwPDB PDBx/mmCIF data dictionary with additional categories to accommodate SIFTS data and added the UniProtKB, Pfam, SCOP2, and CATH residue-level annotations directly into the PDBx/mmCIF files from the PDB archive. With the integrated UniProtKB annotations, these files now provide consistent numbering of residues in different PDB entries allowing easy comparison of structure models. The extended dictionary yields a more consistent, standardised metadata description without altering the core PDB information. This development enables up-to-date cross-reference information at the residue level resulting in better data interoperability, supporting improved data analysis and visualisation.

Introduction

As of March 2023, the Protein Data Bank (PDB)[1] contains over 200,000 entries representing over 61,000 unique entries in the Universal Protein Resource Knowledgebase (UniProtKB)[2]. Often, the PDB archive has the same protein in multiple entries under different experimental conditions or interacting with different macromolecules (proteins, DNA, RNA) or ligand molecules[3–5]. Multiple 3-dimensional coordinates of the same protein are invaluable for comparative structure-function studies[3,6,7]. Linking structure data with annotations available in other data resources such as UniProtKB[2] and to the structural and functional annotations is critical in order to understand biological function and processes at a molecular level. However, one of the barriers to comparative analysis or data integration is the independent, depositor-provided residue numbering in the coordinate files, which may not be the same as the protein sequence numbering[8]. While solving a protein 3D structure, many times the experiments are carried out only on a part of complete protein molecules (e.g. a domain) to make the sample amenable to experimental methods, especially in cases where there are highly flexible linker regions or intrinsically disordered regions[9,10]. Around 58% of the structures in the PDB contain smaller fragments (e.g. a domain) corresponding to different regions of a protein sequence. To determine where these fragments are located on the full-length protein sequence, these fragments need to be mapped to a common reference e.g. protein sequence numbering from a relevant entry in the UniProtKB database. The situation becomes complicated as often the flexible regions in the protein molecules are not modelled leading to unobserved residues i.e. residues without atomic coordinates in protein structures. The occurrence of missing residues makes structure-to-sequence mapping even more challenging. To address this fundamental problem of standardising residue numbering to make protein structure data more accessible to the broader scientific community, the

[1]Protein Data Bank in Europe, European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK. [2]AstraZeneca, Biomedical Campus, 1 Francis Crick Ave, Trumpington, Cambridge, CB2 0AA, UK. [3]Claude Bernard University, Villeurbanne, Lyon, 69100, France. ✉e-mail: cypreeti@ebi.ac.uk

PDBe[11] and UniProtKB[2] teams collaborated to establish the Structure Integration with Function, Taxonomy and Sequences (SIFTS) resource in 2002[12,13]. SIFTS provides up-to-date residue level mapping, with each weekly PDB release, between UniProtKB protein sequences and PDB protein structures allowing better integration of annotations based on protein sequence and structure.

In addition to mapping PDB structures to UniProtKB sequences, SIFTS also maps to other biological resources such as Pfam[14], InterPro[15], SCOP[16], CATH[17], IntEnz[18], GO[19,20], Ensembl[21], NCBI taxonomy database[22] and Homologene[23].

In the past 20 years, SIFTS has become an essential resource, and its data provides the foundation of many data services and web pages. SIFTS is fundamental to the PDBe and PDBe-KB data resources[24] and other databases, such as UniProtKB[2], Pfam[14], RCSB PDB[25], PDBj[26], SCOP2[27], InterPro[15] and MobiDB[28], rely on SIFTS to fetch cross-references between PDB structures and other biological databases. SIFTS data is distributed as summary flat files in CSV/TSV formats and also as a detailed per-entry XML files with residue-level information available from the EMBL-EBI FTP area (ftp://ftp.ebi.ac.uk/pub/databases/msd/sifts/). SIFTS data is also accessible via the PDBe API[29].

While SIFTS data has significantly improved the interoperability of PDB structure data with other key data resources, it still requires to be accessed separately from the 3D coordinates data in the PDB. The SIFTS output format is incompatible with 3D visualisation software that use the PDBx/mmCIF standard[30] and requires an additional step of parsing the data to display SIFTS annotations on protein 3D structure. To boost the FAIRness[31] (Findable, Accessible, Interoperable and Reusable) by further improving the findability and interoperability of PDB structures, the next logical step is to integrate SIFTS annotations alongside the 3D coordinates in the PDBx/mmCIF files. Moreover, with the availability of numerous high-quality, predicted protein structure models from resources like SWISS-MODEL[32] and AlphaFold DB[33,34], which generally follow the protein sequence numbering scheme, it was timely and essential to augment the protein sequence numbering for the experimentally determined 3D coordinates in the PDB. Using data from SIFTS resource, the PDBrenum[8] web server replaces author sequence numbering with UniProtKB numbering in PDB or PDBx/mmCIF format files but it has certain limitations while handling special cases. For instance, while renumbering if this web server does not find any mapping data in SIFTS, it simply adds a large number to the residue's sequence position number. These residues can be expression tags or insertions and need to be represented appropriately without losing the experimental context of the sample. Similarly, for chimeric proteins which are mapped to more than one protein sequence (UniProtKB accession), PDBrenum only renumbers according to the one protein sequence which has maximum coverage, losing information about remaining proteins in the chimeric construct. It does not integrate annotations to other data resources from SIFTS like Pfam, SCOP2 and CATH as well. Thus, there is a need to find a more consistent, sustainable and up-to-date solution while incorporating UniProtKB numbering and annotations from various other data resources in the 3D coordinate files.

Here, we describe incorporating SIFTS annotations in extended PDBx/mmCIF files to directly incorporate UniProtKB residue numbering next to the atomic coordinates. The PDBx/mmCIF is an extensible format that also provides a mechanism to maintain data integrity and is the master format for macromolecular structure data in the PDB[35]. We describe how this current work extends the PDBx/mmCIF dictionary by leveraging the extensibility of its structured framework, thereby providing a mechanism to enrich the biological context of a PDB structure.

Results

Extension to the core SIFTS pipeline.

The core SIFTS pipeline[13] includes (1) a semi-automated process to retrieve the manually curated UniProtKB cross-reference (or canonical UniProtKB accession) for each protein chain in the PDB and (2) an automated process that generates residue-level correspondences between structure (PDB) and the corresponding sequence (UniProtKB). Initial mapping of UniProtKB sequence to the PDB structure is manually curated during the wwPDB annotation process[36]. During the semi-automated process, these manually curated mappings are checked for obsoleted or secondary UniProtKB accessions and are updated accordingly. In the automatic process, the manually curated canonical accession is then expanded to include all its isoforms, and sequence alignment is computed for each PDB-UniProtKB pair. Taking only the PDB-UniProtKB pairs with the same source organism or atleast having a common ancestor within one or two levels up to species level in the taxonomy tree and having at least 90% sequence identity, the pair with the highest sequence identity is annotated as the best mapping. Once we have established the mapping between UniProtKB and PDB protein residues, the cross-references from other resources such as Pfam[14], InterPro[15], SCOP[16], CATH[17], IntEnz[18], GO[19,20], Ensembl[21] and Homologene[23] are added. The SIFTS annotations are stored in the SIFTS database, which is used to make the data accessible via the PDBe REST API. Individual XML files for each PDB entry with residue-level information are exported and the summary files are generated in CSV/TSV formats. An additional process was designed that reads the data from the SIFTS database and augments the PDB structure files with UniProtKB numbering and structure (SCOP2, and CATH resource) and sequence (Pfam resource) domain annotations. This update yields more consistent, standardised metadata. It is important to note that none of the core PDB information, such as atomic coordinates and experimental data, are altered in any way. Figure 1 shows the schematic overview of the data flow of the SIFTS process and highlights the additional process that was developed to export these data into PDBx/mmCIF files. The process helps researchers and data services access SIFTS data directly from the PDBx/mmCIF[37] files. To facilitate this update, additional "SIFTS-specific" mmCIF data categories were designed and integrated into the core PDBx/mmCIF data dictionary. These format specifications are discussed in detail below.

Extensions to the PDBx/mmCIF framework.

PDBx/mmCIF framework organises information in categories containing related data items[37]. The updated PDBx/mmCIF files contain the residue mappings between UniProtKB and PDB, and annotations from Pfam, SCOP2, and CATH. The SIFTS annotations are integrated

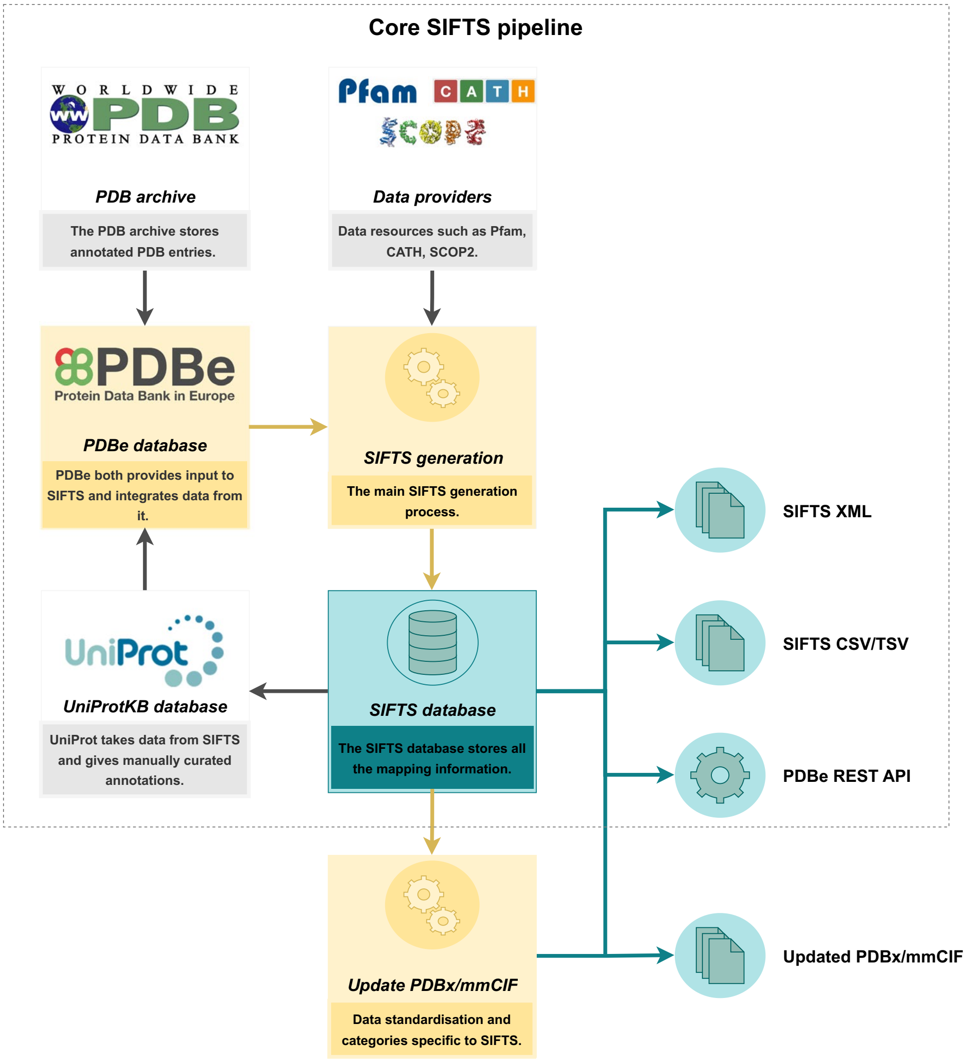

Core SIFTS pipeline
WORLDWIDE PDB PROTEIN DATA BANK
PDB archive
The PDB archive stores annotated PDB entries.
Pfam CATH SCOP2
Data providers
Data resources such as Pfam, CATH, SCOP2.
PDBe Protein Data Bank in Europe
PDBe database
PDBe both provides input to SIFTS and integrates data from it.
SIFTS generation
The main SIFTS generation process.
UniProt
UniProtKB database
UniProt takes data from SIFTS and gives manually curated annotations.
SIFTS database
The SIFTS database stores all the mapping information.
SIFTS XML
SIFTS CSV/TSV
PDBe REST API
Update PDBx/mmCIF
Data standardisation and categories specific to SIFTS.
Updated PDBx/mmCIF

Fig. 1 The schematic overview of the core SIFTS pipeline and an additional process for exporting data into PDBx/mmCIF Files. The figure illustrates the different components of the core SIFTS pipeline, represented in yellow, and the corresponding outputs, indicated in green. The core SIFTS process generates various output files, including the SIFTS database, XML, CSV, and TSV files. The additional process, represented in the figure, is responsible for augmenting SIFTS data in updated PDBx/mmCIF files. The grey components in the figure denote data resources that are external to the SIFTS pipeline.

in two ways: per-segment and per-residue. The per-segment annotations refer to a continuous segment in the protein sequence, where only the start and end positions for the annotations are provided. On the other hand, the per-residue annotations expand the segment boundaries to provide annotations for every residue that spans that region. The reason for having both types of annotation is that expanding segment annotations to the residue level can be complex due to factors such as missing residues, insertions, expression tags, and linker regions in the protein sequence. Moreover, the PDB residue numbers are not always uniquely defined and can have insert codes which together with the PDB residue number uniquely identify a particular residue. These factors can lead to gaps in the numbering between residues, which can make it challenging to expand segment annotations to the residue level. Therefore providing both per-segment and per-residue annotations affords the flexibility to visualise and analyse these data in a way that best suits the user needs. New data categories were added to represent these additional per-segment and per-residue mappings (Fig. 2). Two new categories "_pdbx_sifts_unp_segments" and "_pdbx_sifts_xref_db_segments" were added to represent per-segment mapping to UniProtKB and other data resources - Pfam, SCOP2, CATH. A third category, "_pdbx_sifts_xref_db", was added to provide per-residue mapping from all the external resources. The "_atom_site" category, which represents the coordinate information, was extended with additional data items to integrate UniProtKB residue numbering from the best mapping adjacent to the atomic coordinates.

A summary of the new and modified data categories necessary to encode the SIFTS annotations data is provided below:

1. **_pdbx_sifts_unp_segments**
   This new category describes residue range-based cross-references specific to the UniProtKB database. It shows segments/regions of PDB residues mapped to the canonical UniProtKB accession and all its isoforms. The residue mapping is established by aligning the PDB sequence to each UniProtKB accession (canonical and all the isoforms) and the sequence identity between the aligned PDB-UniProtKB pair is provided. This category also indicates the best mapped UniProtKB accession.
2. **_pdbx_sifts_xref_db_segments**
   This new category describes residue range-based cross-references to additional databases such as Pfam, SCOP2, and CATH.
3. **_pdbx_sifts_xref_db**
   PDB structures often have missing residues, expression tags or linker regions, making the expansion of mappings from segments (residue range) to individual residues cumbersome. An essential category, "_pdbx_sifts_xref_db", therefore describes residue level cross-references to external databases. This category provides annotations specific to the best mapped UniProtKB accession and can be used to identify all the mappings for each residue to external databases (Fig. 3).
4. **_atom_site**
   New data items were added to the "_atom_site" category to represent the best mapped UniProtKB accession, residue type and number. The new data item "_atom_site.pdbx_label_index" along with the "atom_site.label_asym_id" provide a unique identifier for all the polymer residues and individual non-polymer and solvent components.

There are two different numbering schemes followed to indicate each residue (amino-acid or nucleotide) in the PDBx/mmCIF file. Firstly, "auth_seq_id" which is the numbering provided by the author. An author can assign its value in any desired way and the values may be used to relate the given structure to a numbering scheme in a homologous structure, including sequence gaps or insertion codes, which are not necessarily numbers. Secondly, "label_seq_id " which is the wwPDB assigned numbering which starts from 1 and increments sequentially only for all the polymer residues. All the SIFTS-specific categories refer consistently to the wwPDB assigned numbering scheme defined by the "label_seq_id" data item in the atom_site category. The reference to labl_seq_id is provided by the data items ".seq_id", ".seq_id_start" and ".seq_id_end" in the relevant categories. Data on the author provided or the PDB numbering scheme can be retrieved using the appropriate relationships defined in the PDBx/mmCIF categories (Fig. 4).

Often in many proteins, several domains are tandemly repeated[38]. Additionally, researchers also synthesise structures where even the entire protein is repeated for specific research purposes[39,40]. Previously, there was no automated way to find corresponding UniProtKB mappings for multiple domains in a protein structure in the PDB. The data item ".instance_id" is designed to help identify multiple instances of the same protein segment. For example, in the single-chain dimeric Streptavidin structure (PDB 6s50), the two copies of Streptavidin[41] are easily identified by instance ids "1" and "2" for the UniProtKB accession P22629 (Fig. 5).

Similarly, users can rely on this data item to easily identify multiple copies of the same domains in a protein structure.

During evolution protein structures may evolve with an insertion of an additional domain which splits the original structural domain into a discontinuous range of residues in the sequence[42]. For example, the *E.coli* enzyme RNA 3′-terminal phosphate cyclase (PDB 1qmh) consists of two structural domains where a smaller insert domain (residues 186–276) splits the larger domain (residues 5–182 and 277–337)[43]. The identification of the split domain (residues 5–182, 277–337) is evident from the ".segment_id" data item (Fig. 6).

Complete documentation for all the new and updated data categories and items is available at https://mmcif.wwpdb.org/dictionaries/ascii/mmcif_pdbx_v50.dic.

Applications. The SIFTS resource has been widely used in various research studies to retrieve residue correspondence between PDB structures and UniProtKB sequences[44–48]. However, in many cases, researchers have had to manually renumber the coordinate files to reflect UniProtKB numbering for subsequent comparative analysis across multiple PDB structures[49–51]. While SIFTS has been used in several functional studies, including mapping somatic mutations to protein structures to identify 3D clusters of mutations with functional significance[52] and mapping GPCR structures to their respective G protein structures to investigate the allosteric mechanism of GPCR activation[53], authors still had to manually validate missing positions in PDB structures to verify genuine cases of chimeric proteins, peptide tags, or point mutations. Unfortunately, this process was both time-consuming and error-prone. However, with the incorporation of SIFTS residue-level mapping to the best mapped UniProtKB sequence in the PDBx/mmCIF files, manual verification is no longer necessary, saving time and facilitating the analysis and interpretation of data.

Integration of UniProtKB sequence annotations and 3D-structures, can furnish the biological and functional context for the structural data. For instance, mapping variant annotations onto 3D-structure, can provide insights into the genetic basis of complex traits and diseases. SIFTS resource has also been used to fetch annotations like sequence domains and structural domains for various PDB structures[51,54]. Using the domain annotations mapped to a protein sequence in these PDBx/mmCIF files, researchers can easily identify the location, multiple copies and boundaries of different domains within a protein, which can help in understanding the overall structure and function of the protein. This also facilitates comparing proteins with similar domain structures and identifies potential functional relationships.

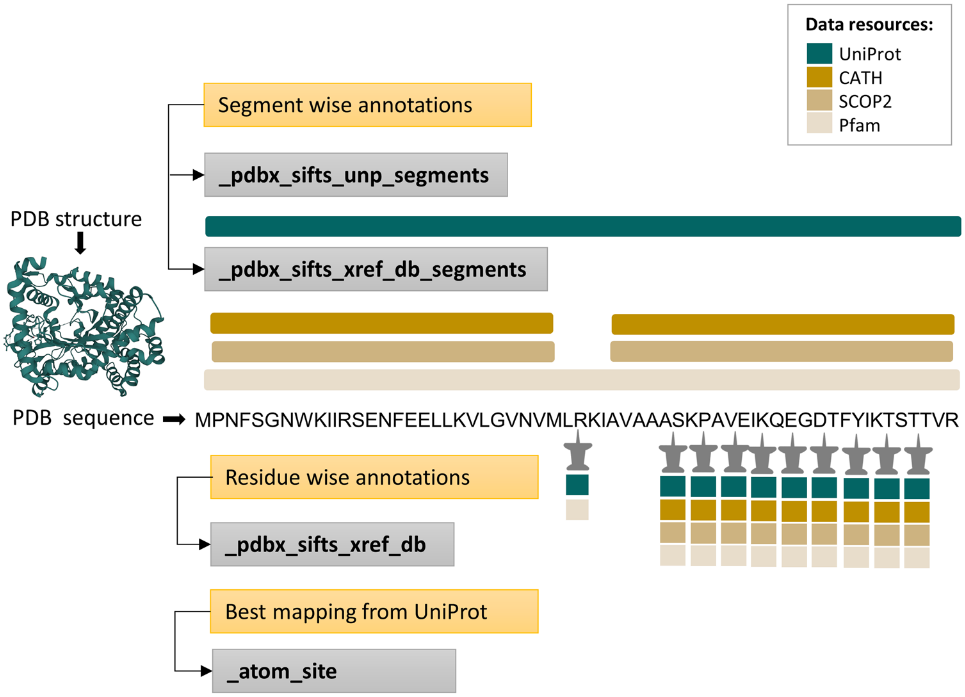
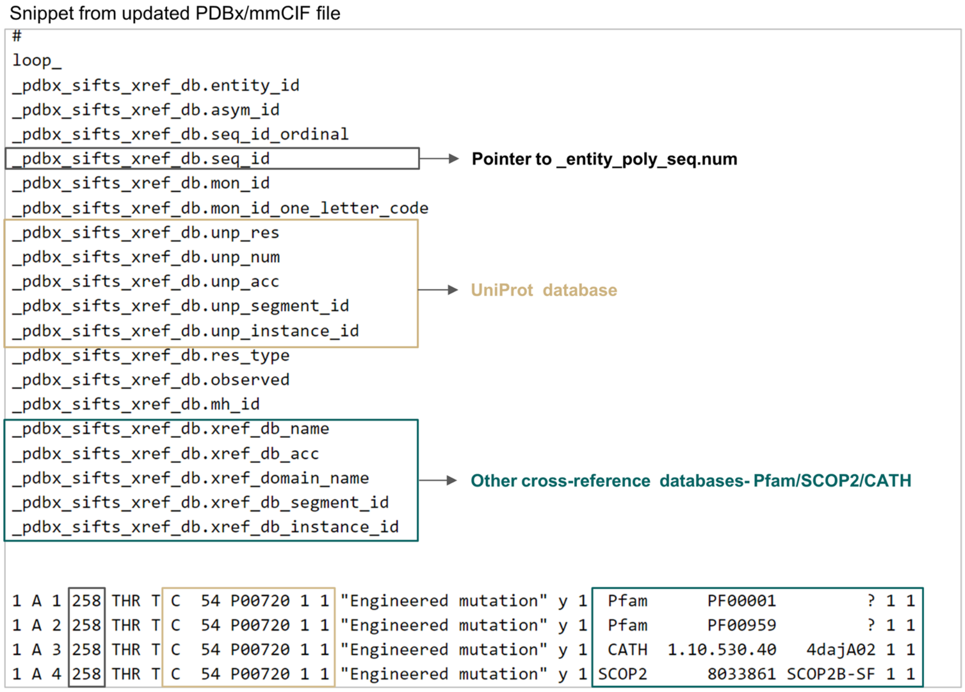


Fig. 2 The PDBx/mmCIF extension incorporates mappings from various data resources. SIFTS annotations mapping PDB residues to various data resources are shown both per-segment (top) and per-residue (bottom). All the new SIFTS-specific or modified PDBx/mmCIF categories are shown in grey boxes. The new SIFTS-specific PDBx/mmCIF categories introduced to show per-segment annotations from UniProtKB and all the other external data resources (Pfam, SCOP2, CATH) are "_pdbx_sifts_unp_segments" and "_pdbx_sifts_xref_db_segments" respectively. "_pdbx_sifts_xref_db" is another new SIFTS-specific PDBx/mmCIF category introduced to show per-residue annotations. We also modified the "_atom_site" category to indicate the best mapped UniProtKB sequence.


Snippet from updated PDBx/mmCIF file
Pointer to _entity_poly_seq.num
UniProt database
Other cross-reference databases- Pfam/SCOP2/CATH

```
#
loop_
_pdbx_sifts_xref_db.entity_id
_pdbx_sifts_xref_db.asym_id
_pdbx_sifts_xref_db.seq_id_ordinal
_pdbx_sifts_xref_db.seq_id
_pdbx_sifts_xref_db.mon_id
_pdbx_sifts_xref_db.mon_id_one_letter_code
_pdbx_sifts_xref_db.unp_res
_pdbx_sifts_xref_db.unp_num
_pdbx_sifts_xref_db.unp_acc
_pdbx_sifts_xref_db.unp_segment_id
_pdbx_sifts_xref_db.unp_instance_id
_pdbx_sifts_xref_db.res_type
_pdbx_sifts_xref_db.observed
_pdbx_sifts_xref_db.mh_id
_pdbx_sifts_xref_db.xref_db_name
_pdbx_sifts_xref_db.xref_db_acc
_pdbx_sifts_xref_db.xref_domain_name
_pdbx_sifts_xref_db.xref_db_segment_id
_pdbx_sifts_xref_db.xref_db_instance_id


1 A 1 258 THR T C  54 P00720 1 1 "Engineered mutation" y 1  Pfam      PF00001         ? 1 1
1 A 2 258 THR T C  54 P00720 1 1 "Engineered mutation" y 1  Pfam      PF00959         ? 1 1
1 A 3 258 THR T C  54 P00720 1 1 "Engineered mutation" y 1  CATH  1.10.530.40   4dajA02 1 1
1 A 4 258 THR T C  54 P00720 1 1 "Engineered mutation" y 1 SCOP2      8033861 SCOP2B-SF 1 1
```

Fig. 3 Single placeholder in PDBx/mmCIF files to find all the annotations associated with any residue from external databases. This figure shows the "_pdbx_sifts_xref_db" category for PDB 4daj. This critical new data category can describe residue-level cross-references to external databases. The items specific to the UniProtKB database and other cross-reference databases are marked in beige and green coloured boxes respectively.

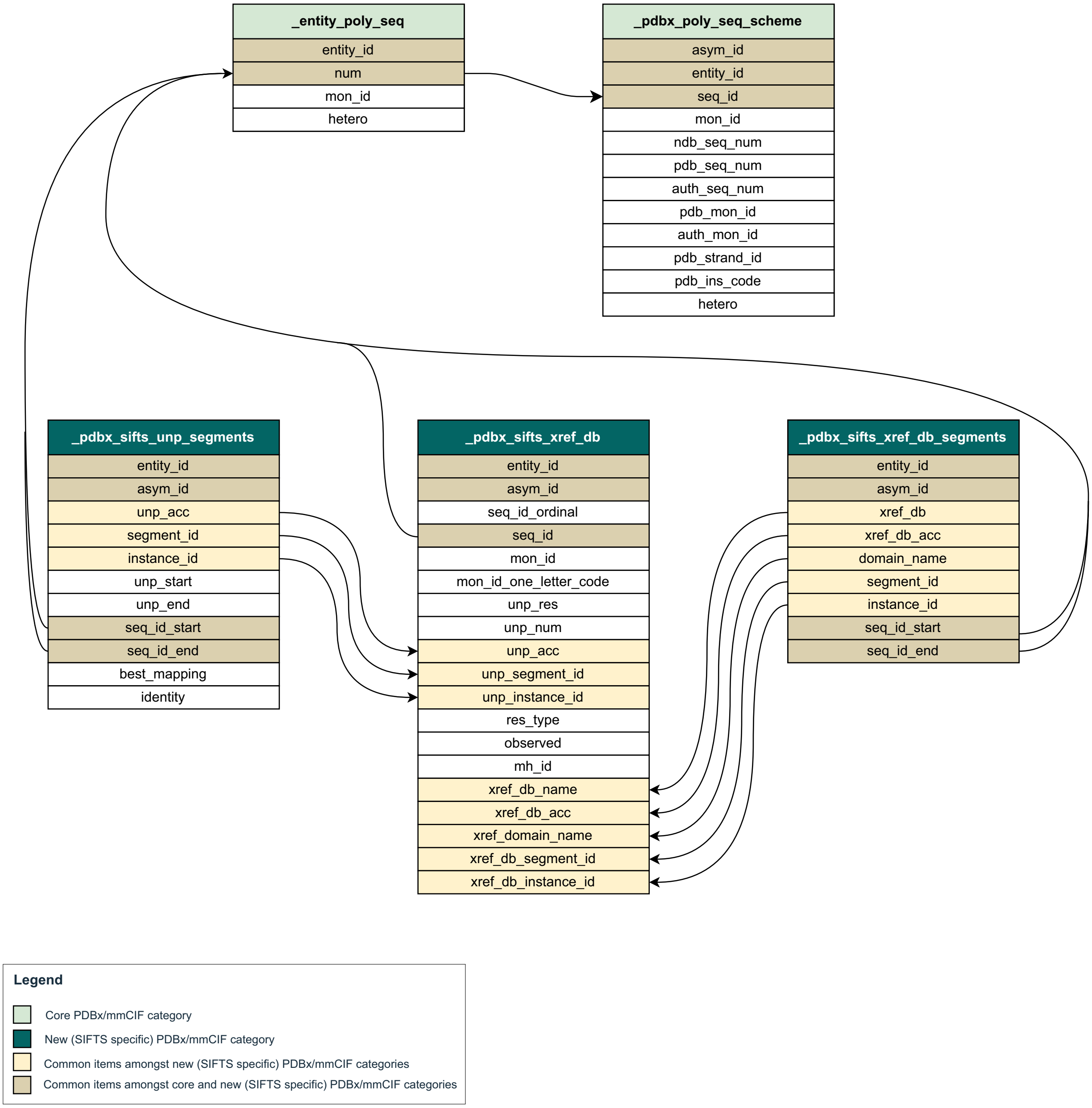
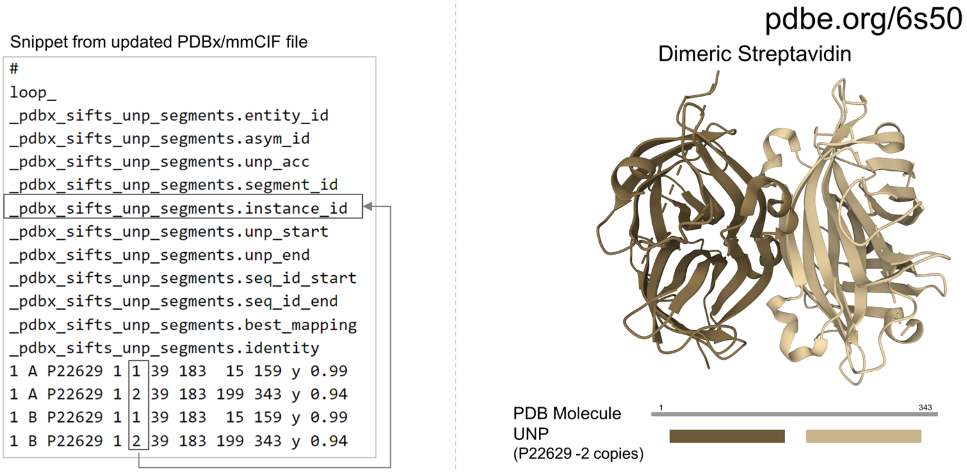


Fig. 4 Category relationship diagram including new SIFTS specific PDBx/mmCIF categories. New SIFTS specific PDBx/mmCIF data categories are shown along with their data items. All the common data items amongst these new data categories are highlighted and their relationship is shown. Further, the relationship of the data items representing PDB residue numbers - ".seq_id", ".seq_id_start" or ".seq_id_end" in these new data categories to existing data categories is shown.

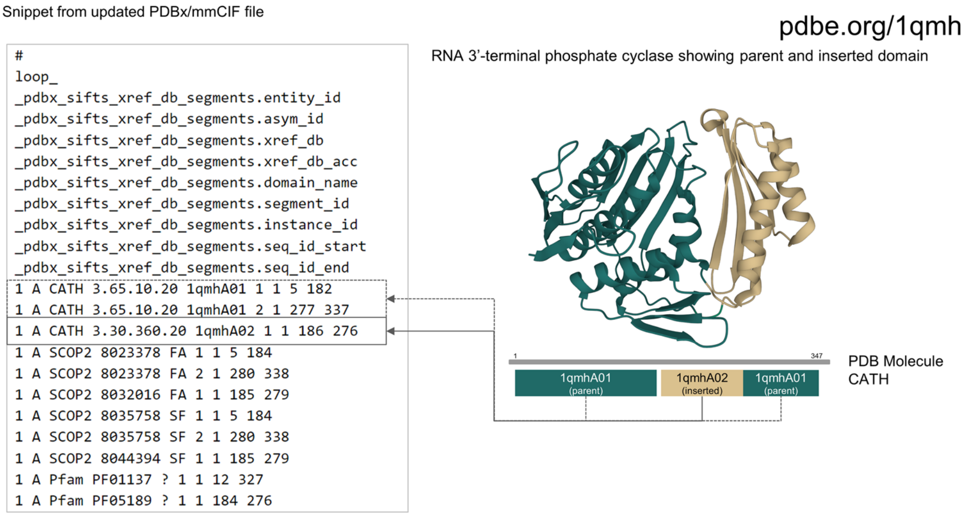
SIFTS is not only widely used in scientific research but also by several data resources[12]. For example, UniProtKB exploits SIFTS information to provide structure mapping in the UniProtKB database. SCOP[55] and Pfam[14,56] also use SIFTS to map protein domains and connect sequence domains with their corresponding structures. The web resource Kincore relies on SIFTS to map protein kinases to their respective structures, extract relevant information such as domain boundaries and ligand binding sites, and provide a structural classification of protein kinases and their inhibitors[50]. The PDBx/mmCIF files with SIFTS annotations address the fundamental need by combining data from various resources and providing coordinate files with a common reference frame, improving interoperability and reuse of these data. The availability of these files will streamline data extraction and promote consistent and efficient data sharing.

Adding UniProtKB, Pfam, SCOP2, and CATH annotations to PDB coordinate files can be very helpful for resources like Gene Integration with Function, Taxonomy, and Sequence (GIFTS, https://www.ebi.ac.uk/gifts/), Venus[57] or PhyreRisk[58]. These annotations provide valuable information to gain a deeper understanding of the relationships between protein structure and function[59], which can be used to link structural and functional data on a genome-wide scale[60]. By integrating these annotations in PDBx/mmCIF files, it becomes easier to



```
#
loop_
_pdbx_sifts_unp_segments.entity_id
_pdbx_sifts_unp_segments.asym_id
_pdbx_sifts_unp_segments.unp_acc
_pdbx_sifts_unp_segments.segment_id
_pdbx_sifts_unp_segments.instance_id
_pdbx_sifts_unp_segments.unp_start
_pdbx_sifts_unp_segments.unp_end
_pdbx_sifts_unp_segments.seq_id_start
_pdbx_sifts_unp_segments.seq_id_end
_pdbx_sifts_unp_segments.best_mapping
_pdbx_sifts_unp_segments.identity
1 A P22629 1 1 39 183  15 159 y 0.99
1 A P22629 1 2 39 183 199 343 y 0.94
1 B P22629 1 1 39 183  15 159 y 0.99
1 B P22629 1 2 39 183 199 343 y 0.94
```

Fig. 5 Distinguishing between multiple instances of the same protein in the PDBx/mmCIF file. The data item ".instance_id" enables users to identify the two copies of the same protein, Streptavidin (UniProtKB accession P22629), in the dimeric Streptavidin structure (PDB 6s50).


Snippet from updated PDBx/mmCIF file
pdbe.org/1qmh
RNA 3'-terminal phosphate cyclase showing parent and inserted domain
1
347
PDB Molecule
CATH
1qmhA01
(parent)
1qmhA02
(inserted)
1qmhA01
(parent)

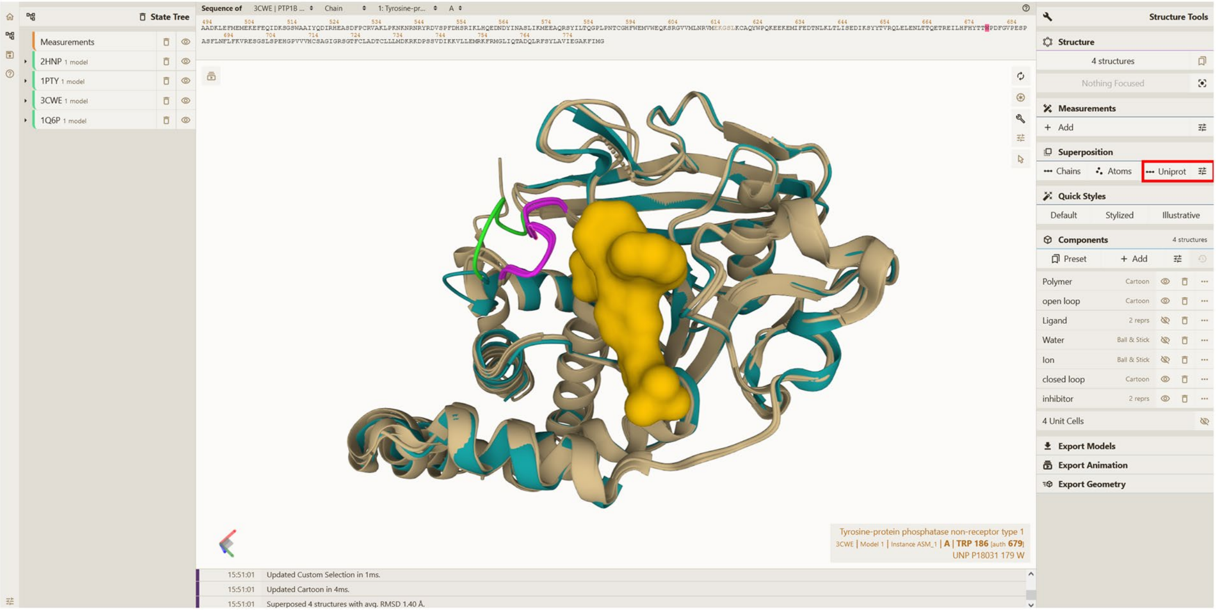
```
#
loop_
_pdbx_sifts_xref_db_segments.entity_id
_pdbx_sifts_xref_db_segments.asym_id
_pdbx_sifts_xref_db_segments.xref_db
_pdbx_sifts_xref_db_segments.xref_db_acc
_pdbx_sifts_xref_db_segments.domain_name
_pdbx_sifts_xref_db_segments.segment_id
_pdbx_sifts_xref_db_segments.instance_id
_pdbx_sifts_xref_db_segments.seq_id_start
_pdbx_sifts_xref_db_segments.seq_id_end
1 A CATH 3.65.10.20 1qmhA01 1 1 5 182
1 A CATH 3.65.10.20 1qmhA01 2 1 277 337
1 A CATH 3.30.360.20 1qmhA02 1 1 186 276
1 A SCOP2 8023378 FA 1 1 5 184
1 A SCOP2 8023378 FA 2 1 280 338
1 A SCOP2 8032016 FA 1 1 185 279
1 A SCOP2 8035758 SF 1 1 5 184
1 A SCOP2 8035758 SF 2 1 280 338
1 A SCOP2 8044394 SF 1 1 185 279
1 A Pfam PF01137 ? 1 1 12 327
1 A Pfam PF05189 ? 1 1 184 276
```

Fig. 6 Identification of split domains from PDBx/mmCIF file. The "_pdbx_sifts_xref_db_segments" category in the PDBx/mmCIF file of PDB 4daj helps to clearly identify discontinuous domains. The two halves of the M3 receptor domain are indicated by the same ".instance_id" but different ".segment_id".

map genetic variants to protein structures, which can greatly facilitate genome-wide studies. The use of SIFTS annotations in the COSMIC data resource is an excellent example of how this approach can be used to efficiently and accurately analyse the impact of genetic variants on protein function and stability[61]. This can be further expanded to support a wide range of computational approaches for analysing protein structure and function[62], including functional annotation[48], structural comparison[59], ligand binding analysis[63], identifying new protein-protein interactions[64], functional pathways, and potential drug targets[65] on a large scale.

Various data visualisation tools can directly use these PDBx/mmCIF files, making the mapping of 1D sequence data onto the 3D structure views straightforward. With our improvements, researchers from various scientific fields can easily map sequence feature data onto PDB structures. Users can directly retrieve all the SIFTS annotations like structural domains, sequence domains and conflicts between sequences and structures from the PDBx/mmCIF files.

These files also provide a basis for improved comparisons between experimentally determined and predicted protein models. UniProtKB numbering in the coordinate files allows direct residue correspondence making structural comparison and superposition easier. It also makes it easier to compare PDB structures with the predicted model structures from AlphaFold DB[33,34], SWISS-MODEL[32], RoseTTAFold[66], and many other resources, as these models follow a natural sequence numbering. These files are already being used by Mol*[67] (https://molstar.org/viewer/) to perform extremely fast superpositions using the SIFTS UniProtKB mapping. This superposition functionality in Mol*[67] is very powerful as it gives users the means to directly superimpose protein structures in their web browser without downloading any data or software. Mol* uses the SIFTS specific new



```
494       504       514       524       534       544       554       564       574       584       594       604       614       624       634       644       654       664       674       684
AADKLEFMEMEKEFEQIDKSGSWAAIYQDIRHEASDFPCRVAKLPKNKNRNRYRDVSPFDHSRIKLHQEDNDYINASLIKMEEAQRSYILTQGPLPNTCGHFWEMVWEQKSRGVVMLNRVMEKGSLKCAQYWPQKEEKEMIFEDTNLKLTLISEDIKSYYTVRQLELENLTTQETREILHFHYTTWPDFGVPESP
     694       704       714       724       734       744       754       764       774
ASFLNFLFKVRESGSLSPEHGPVVVHCSAGIGRSGTFCLADTCLLLMDKRKDPSSVDIKKVLLEMRKFRMGLIQTADQLRFSYLAVIEGAKFIMG
```

Polymer	Cartoon		
open loop	Cartoon		
Ligand	2 reprs		
Water	Ball & Stick		
Ion	Ball & Stick		
closed loop	Cartoon		
inhibitor	2 reprs		

Tyrosine-protein phosphatase non-receptor type 1
3CWE | Model 1 | Instance ASM_1 | A | TRP 186 [auth 679]
UNP P18031 179 W

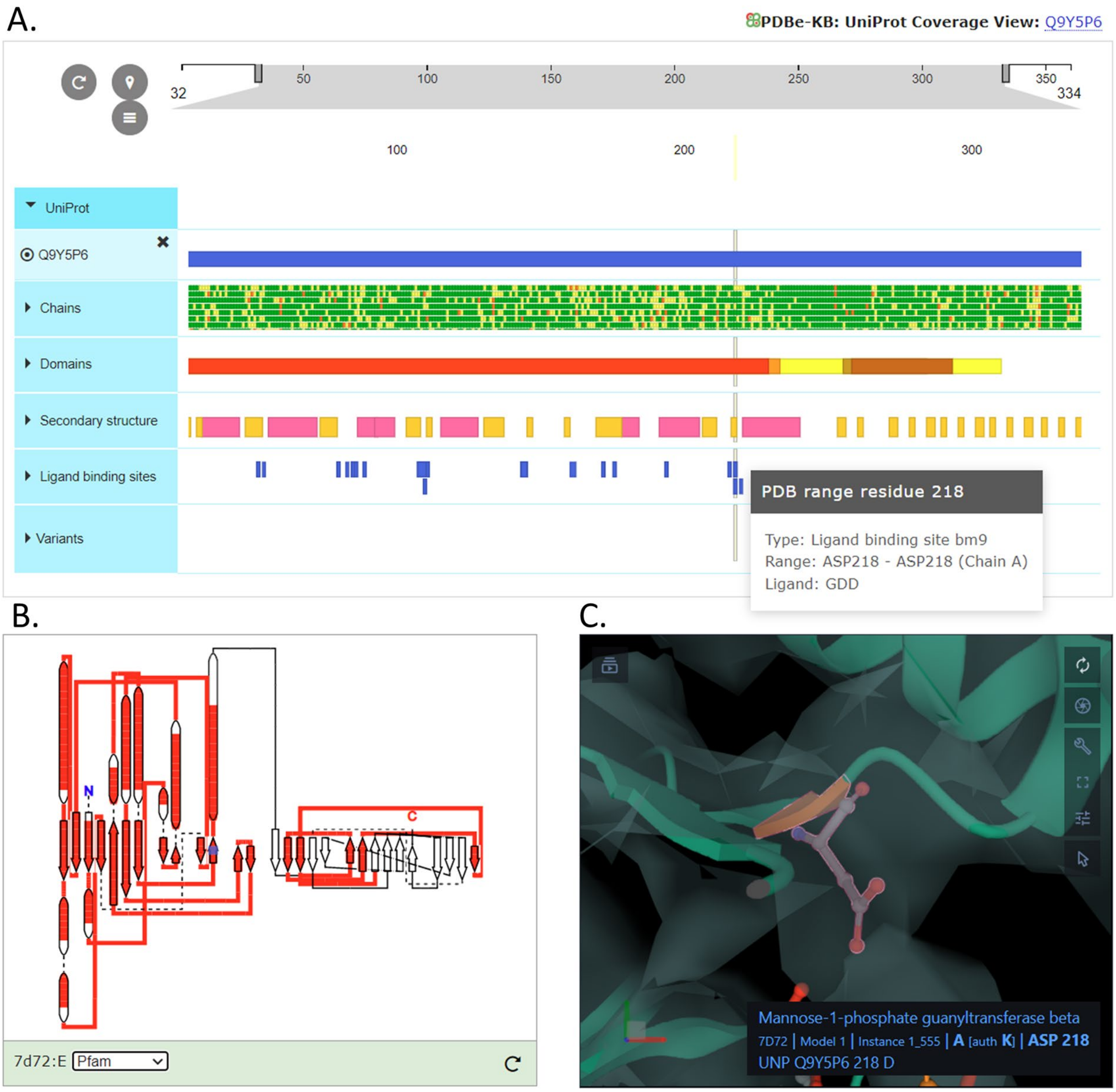
```
15:51:01    Updated Custom Selection in 1ms.
15:51:01    Updated Cartoon in 4ms.
15:51:01    Superposed 4 structures with avg. RMSD 1.40 Å.
```

Fig. 7 Superposition of protein structures using Mol*. The superposed apo and holo forms of human PTP1B protein are shown in green and beige colours, respectively, in Mol*. The WDP loop is in open (light green colour) conformation in the apo form (PDB 2HNP). Upon binding to various substrates/inhibitors this WDP loop attains closed (pink colour) conformation covering the catalytic site. The inhibitor bound in PDB 1Q6P is shown in the surface representation. The average RMSD between the four superposed structures as computed by Mol* is 1.40 Å. As seen in the tool-tip (bottom-right in the figure), residue W179 from PDB 3CWE and other residues in inhibitor bound PDBs 3CWE and 1Q6P have different author numbering compared to the unbound/substrate bound form (PDB 2HNP/1PTY). The UniProtKB numbering in the PDBx/mmCIF file provides a common reference frame for residue correspondence and supports superposition based on UniProtKB in Mol*.

data items added in the "_atom_site" category to establish the residue equivalence (UniProtKB residue number) from different PDB structures. Mol* superimposes the structures by calculating the optimal rotation and translation that align the corresponding atoms in each equivalent protein residue. Figure 7 shows the superposition of the unbound and bound forms of human Protein Tyrosine Phosphatase 1B protein (PTP1B, UniProtKB accession: P18031) performed using the "UniProt" button (highlighted in red box) in the Mol* Superposition panel[67]. This protein is known to be a signalling molecule regulating a variety of cellular processes including cell growth, differentiation and oncogenic transformation and is a potential therapeutic target for the treatment of type 2 diabetes and cancer[68]. Upon substrate/inhibitor binding, the WPD loop transitions from an open to a closed conformation[69–72] as shown in Fig. 7.

The new PDBx/mmCIF files also provided a basis for developing interactive visualisations. For example, the PDBe entry pages show the ProtVista component[73], a 2D visualisation for displaying the primary sequence features of proteins. ProtVista was developed in collaboration with UniProtKB and InterPro at EMBL-EBI. The PDBx/mmCIF files with PDB-UniprotKB residue mapping, enable interactivity between the 3D viewer (Fig. 8C), the ProtVista sequence viewer (Fig. 8A) and the 2D topology component (Fig. 8B). Consequently, Mol* can easily display all the annotations available in ProtVista and the 2D topology component on the 3D structure. As shown in Fig. 8, for Mannose-1-phosphate guanyltransferase, PDB 7d72 (https://www.ebi.ac.uk/pdbe/entry/pdb/7d72/protein/1), if users click on any residue annotation in the 2D viewer ProtVista, the residue or the residue segment is automatically highlighted in 3D,in the Mol* viewer. Similarly, users can highlight various structural or sequence domains, or other annotations in either the 2D topology component, 2D ProtVista component or Mol* viewer, and the three visualisations cross-talk with each other simultaneously, making visualisation and interpretation of data much easier. Mol* already uses these PDBx/mmCIF files to display various annotations on PDBe and PDBe-KB webpages. With SIFTS annotations directly available in the coordinate file, the 3D visualisation on PDBe and PDBe-KB webpages is more efficient and optimal.

It is important to note that adding additional data to a PDBx/mmCIF file, such as augmenting best mapped UniProtKB residue mapping in the "atom_site" category can come with a trade-off of an increase in the file size. While this may not be an issue for smaller PDB entries, it can become problematic for larger entries with significant file size. To address this issue, wwPDB provides binaryCIF[74] (bcif) files as an alternative to traditional PDBx/mmCIF files. The bcif format is a compressed binary version of the PDBx/mmCIF format that significantly reduces the file size, making it easier to handle and share large amounts of structural data. The Mol*, an open-source software for 3D molecular visualisation and analysis, also supports the bcif file format, allowing users to easily access and analyse structural data in this format.



PDB range residue 218

Type: Ligand binding site bm9
Range: ASP218 - ASP218 (Chain A)
Ligand: GDD

Mannose-1-phosphate guanyltransferase beta
7D72 | Model 1 | Instance 1_555 | A [auth K] | ASP 218
UNP Q9Y5P6 218 D

Fig. 8 The 2D visualisation components are interactively linked with 3D visualisation components on PDBe entry pages. Various 2D and 3D visualisation components seen on PDBe entry pages are interactively linked with each other. Here we show visualisation data for Mannose-1-phosphate guanyltransferase (PDB 7d72). (**A**) shows a 2D sequence feature viewer (ProtVista) and (**B**) shows a 2D topology viewer, along with (**C**) showing the 3D viewer, Mol*. As users select any residue (here ligand-binding residue ASP218 is selected) in ProtVista, it is automatically highlighted in Mol* and vice-versa. Users can also highlight a range of residues (e.g. domains) in any of these viewers. Here, we show the Pfam domain highlighted in red in the 2D topology viewer.

Discussion

Interoperability challenges between the protein structure data in the PDB and protein sequences in the UniProtKB presents a significant barrier to accessibility and reusability. The seemingly trivial task of mapping residue-level information proved to be a formidable task that necessitated the development of the SIFTS resource. While SIFTS has successfully provided up-to-date mappings between the PDB and other data resources for the past 20 years, using these mappings still required some level of technical expertise.

To remove a tedious but previously mandatory step in many structural data analyses, we worked on adding the SIFTS mapping data directly into the PDBx/mmCIF files, the master format for the PDB archive. We designed new data categories and extended existing ones to provide flexible support for residue-level annotations. This development will allow easy linking of structural and functional annotations derived using structure and sequence data. It will also streamline the vast majority of high-throughput bioinformatics analysis pipelines by allowing developers to remove a tedious and error-prone step from their processes. Including the SIFTS data in the PDBx/mmCIF will also improve the efficiency of data visualisation tools, both those that specialise in 3D molecular graphics and those that focus on the interactive mapping of annotations onto to the protein structure representations e.g. sequence or topology.

By extending the PDBx/mmCIF data format, this work has laid the foundation for the future integration of additional annotations, allowing the files to be more comprehensive and to provide the biological context for PDB structures.

Methods

PDBx/mmCIF file format and PDBx/mmCIF dictionary. The PDBx/mmCIF(Protein Data Bank exchange/macromolecular Crystallographic Information File) is a well-established data format utilised for storing and sharing information related to the three-dimensional structure of macromolecules, including proteins and nucleic acids. Widely considered as the master format for the PDB archive, it is extensively used for representing structural data. The format uses a text-based file format that encodes data and metadata utilising data items grouped into categories. The PDBx/mmCIF dictionary[30] defines a standardised set of categories and data items, along with controlled vocabularies and explicit relationships between different categories and data items. This format is extensible, allowing the incorporation of new data items and categories, as demonstrated by the IHM[75] and ModelCIF[76] extensions. The IHM extension enables the archiving of structural models of macromolecular assemblies obtained through integrative/hybrid methods, while the ModelCIF extension enables the consistent representation of molecular models obtained through computational methods. By facilitating such inclusion of new information and accommodating scientific advancements, the PDBx/mmCIF dictionary continues to remain relevant and valuable to the scientific community. The PDBx/mmCIF dictionary is maintained by the wwPDB consortium and is regularly updated with new data items to reflect changes in the field of structural biology. The mmCIF dictionary can be accessed and downloaded freely from https://mmcif.wwpdb.org/dictionaries/mmcif_pdbx_v50.dic/Index/.

SIFTS-specific data categories and items in PDBx/mmCIF Dictionary. The PDBx/mmCIF dictionary was extended with, three new data categories to provide the necessary semantic organisation to represent SIFTS annotations: "_pdbx_sifts_unp_segments", "_pdbx_sifts_xref_db_segments", and "_pdbx_sifts_xref_db".

The "_pdbx_sifts_unp_segments" category displays the UniProtKB sequence segments that correspond to the PDB structure. The "_pdbx_sifts_xref_db_segments" category provides information about the cross-references between the PDB structure and other databases, such as Pfam, CATH, and SCOP2. Finally, the "_pdbx_sifts_xref_db" category displays per-residue annotations between the PDB structure, UniProtKB, and other data resources.

Additionally, the "_atom_site" category was modified to integrate residue-level cross-reference data to the best mapped UniProtKB sequence. The updated PDBx/mmCIF dictionary, including all the new and updated data categories and items, is publicly available at https://mmcif.wwpdb.org/dictionaries/mmcif_pdbx_v50.dic/Index/.

Augmenting the core SIFTS process. SIFTS (Structure Integration with Function, Taxonomy and Sequences) is a collaborative resource between the PDBe (Protein Data Bank in Europe) and UniProtKB teams at EMBL-EBI. It is designed to map the protein structures available in PDB to the protein sequences in UniProtKB at the individual residue level. The SIFTS mapping can facilitate transfer annotations from a variety of biological resources including the NCBI taxonomy database, IntEnz, GO, Pfam, InterPro, SCOP, CATH, PubMed, Ensembl, Homologene, and automatic Pfam domain assignments based on HMM profiles. The pipeline is run weekly by PDBe as part of the PDB release process.

The mapping between PDB protein structures and UniProtKB protein sequences is manually curated by PDB and UniProtKB annotators. SIFTS performs automatic sequence alignment and generates a residue-level mapping between aligned protein structures and sequences. The pipeline downloads and parses data from various biological resources, which is then loaded into the SIFTS database (Fig. 1). SIFTS database is queried to derive residue-level annotations for all these biological resources. The SIFTS process generates per-entry XML files, summary CSV and TSV files to distribute all the SIFTS annotations. The SIFTS database also powers all the SIFTS related PDBe API[29].

To update PDBx/mmCIF files with residue-level annotations from SIFTS resources, a new process was added to the existing SIFTS pipeline. For a given PDB entry, the new process reads all the relevant data from the SIFTS database and integrates it into the PDBx/mmCIF file. The integration of SIFTS data uses the extended PDBx/mmCIF dictionary discussed earlier. The new process is implemented in Python and uses gemmi[77] to parse the PDBx/mmCIF file and write the SIFTS annotations in the corresponding categories. The process is executed as part of the PDBe weekly release pipeline, ensuring up-to-date SIFTS data in the PDBx/mmCIF files every Wednesday to coincide with the weekly PDB release. Currently, residue-level SIFTS annotations for UniProtKB, Pfam, SCOP2, and CATH databases are integrated in the PDBx/mmCIF files.

Data availability

We expanded the PDBe release pipeline with a process that adds SIFTS annotations to the PDBx/mmCIF files for individual structures in the PDB archive. The scientific community can download these PDBx/mmCIF files from the PDBe entry pages (https://pdbe.org/7dr0) and through direct URLs (https://www.ebi.ac.uk/pdbe/static/entry/7o9f_updated.cif), using the PDBe download service (https://www.ebi.ac.uk/pdbe/download/api) or from the EMBL-EBI FTP area (https://ftp.ebi.ac.uk/pub/databases/msd/updated_mmcif/).

Code availability

To assist users in utilising the updated PDBx/mmCIF files and SIFTS annotations, a Google Colab notebook is available at https://colab.research.google.com/github/PDBe-KB/sifts_data_analysis/blob/main/sifts.ipynb or via GitHub at https://github.com/PDBe-KB/sifts_data_analysis. This notebook provides information on how to parse, extract and filter SIFTS annotations from the updated PDBx/mmCIF files. Additionally, the notebook demonstrates how users can compare various numbering schemes of a given residue across different PDB structures of the same protein.

Received: 6 December 2022; Accepted: 23 March 2023;
Published online: 12 April 2023

References

1. wwPDB consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* **47**, D520–D528 (2019).
2. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).
3. Brylinski, M. & Skolnick, J. What is the relationship between the global structures of apo and holo proteins? *Proteins* **70**, 363–377 (2008).
4. Burra, P. V., Zhang, Y., Godzik, A. & Stec, B. Global distribution of conformational states derived from redundant models in the PDB points to non-uniqueness of the protein structure. *Proc. Natl. Acad. Sci.* **106**, 10505 (2009).
5. Lobanov, M. Y. *et al.* ComSin: database of protein structures in bound (complex) and unbound (single) states in relation to their intrinsic disorder. *Nucleic Acids Res.* **38**, D283–D287 (2010).
6. Gutteridge, A. & Thornton, J. Conformational changes observed in enzyme crystal structures upon substrate binding. *J. Mol. Biol.* **346**, 21–28 (2005).
7. Vishwanath, S., de Brevern, A. G. & Srinivasan, N. Same but not alike: Structure, flexibility and energetics of domains in multi-domain proteins are influenced by the presence of other domains. *PLOS Comput. Biol.* **14**, e1006008 (2018).
8. Faezov, B. & Dunbrack, R. L. Jr. PDBrenum: A webserver and program providing Protein Data Bank files renumbered according to their UniProt sequences. *PLOS ONE* **16**, e0253411 (2021).
9. Oldfield, C. J. *et al.* Utilization of protein intrinsic disorder knowledge in structural proteomics. *Biochim. Biophys. Acta* **1834**, 487–498 (2013).
10. Seffernick, J. T. & Lindert, S. Hybrid methods for combined experimental and computational determination of protein structure. *J. Chem. Phys.* **153**, 240901 (2020).
11. Armstrong, D. R. *et al.* PDBe: improved findability of macromolecular structure data in the PDB. *Nucleic Acids Res.* **48**, D335–D343 (2020).
12. Dana, J. M. *et al.* SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic Acids Res.* **47**, D482–D489 (2019).
13. Velankar, S. *et al.* SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Res.* **41**, D483–D489 (2013).
14. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
15. Blum, M. *et al.* The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res* **49**, D344–D354 (2021).
16. Andreeva, A. *et al.* Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.* **36**, D419–D425 (2008).
17. Sillitoe, I. *et al.* CATH: increased structural coverage of functional space. *Nucleic Acids Res.* **49**, D266–D273 (2021).
18. Fleischmann, A. *et al.* IntEnz, the integrated relational enzyme database. *Nucleic Acids Res.* **32**, D434–437 (2004).
19. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
20. The Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.* **49**, D325–D334 (2021).
21. Howe, K. L. *et al.* Ensembl 2021. *Nucleic Acids Res.* **49**, D884–D891 (2021).
22. Schoch, C. L. *et al.* NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database J. Biol. Databases Curation* **2020**, (2020).
23. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **49**, D10–D17 (2021).
24. PDBe-KB consortium. PDBe-KB: collaboratively defining the biological context of structural data. *Nucleic Acids Res.* **50**, D534–D542 (2022).
25. Bittrich, S. *et al.* RCSB Protein Data Bank: improved annotation, search and visualization of membrane protein structures archived in the PDB. *Bioinformatics* **38**, 1452–1454 (2022).
26. Bekker, G.-J. *et al.* Protein Data Bank Japan: Celebrating our 20th anniversary during a global pandemic as the Asian hub of three dimensional macromolecular structural data. *Protein Sci.* **31**, 173–186 (2022).
27. Andreeva, A., Howorth, D., Chothia, C., Kulesha, E. & Murzin, A. G. Investigating Protein Structure and Evolution with SCOP2. *Curr. Protoc. Bioinforma.* **49**, 1.26.1–1.26.21 (2015).
28. Piovesan, D. *et al.* MobiDB: intrinsically disordered proteins in 2021. *Nucleic Acids Res.* **49**, D361–D367 (2021).
29. Nair, S. *et al.* PDBe aggregated API: programmatic access to an integrative knowledge graph of molecular structure data. *Bioinformatics* **37**, 3950–3952 (2021).
30. Westbrook, J. D. *et al.* PDBx/mmCIF Ecosystem: Foundational Semantic Tools for Structural Biology. *J. Mol. Biol.* **434**, 167599 (2022).
31. FAIR principles for data stewardship. *Nat. Genet.* **48**, 343–343 (2016).
32. Waterhouse, A. *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018).
33. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).
34. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
35. Bourne, P. E. *et al.* [30] Macromolecular crystallographic information file. in *Methods in Enzymology* vol. 277 571–590 (Academic Press, 1997).
36. Young, J. Y. *et al.* Worldwide Protein Data Bank biocuration supporting open access to high-quality 3D structural biology data. *Database* **2018**, bay002 (2018).
37. Bourne, P. *et al.* The Macromolecular Crystallographic Information File (mmCIF). (2001).
38. Björklund, A. K., Ekman, D. & Elofsson, A. Expansion of protein domain repeats. *PLoS Comput. Biol.* **2**, e114 (2006).
39. Aslan, F. M., Yu, Y., Mohr, S. C. & Cantor, C. R. Engineered single-chain dimeric streptavidins with an unexpected strong preference for biotin-4-fluorescein. *Proc. Natl. Acad. Sci.* **102**, 8507–8512 (2005).
40. Mikel, P., Vasickova, P. & Kralik, P. One-plasmid double-expression His-tag system for rapid production and easy purification of MS2 phage-like particles. *Sci. Rep.* **7**, 17501 (2017).
41. Wu, S. *et al.* Breaking Symmetry: Engineering Single-Chain Dimeric Streptavidin as Host for Artificial Metalloenzymes. *J. Am. Chem. Soc.* **141**, 15869–15878 (2019).

42. Aroul-Selvam, R., Hubbard, T. & Sasidharan, R. Domain insertions in protein structures. *J. Mol. Biol.* **338**, 633–641 (2004).
43. Palm, G. J., Billy, E., Filipowicz, W. & Wlodawer, A. Crystal structure of RNA 3′-terminal phosphate cyclase, a ubiquitous enzyme with unusual topology. *Structure* **8**, 13–23 (2000).
44. MacGowan, S. A. & Barton, G. J. Missense variants in ACE2 are predicted to encourage and inhibit interaction with SARS-CoV-2 Spike and contribute to genetic risk in COVID-19. *bioRxiv* 2020.05.03.074781, https://doi.org/10.1101/2020.05.03.074781 (2020).
45. Hall, M. W. J., Shorthouse, D., Jones, P. H. & Hall, B. A. Investigating structure function relationships in the NOTCH family through large-scale somatic DNA sequencing studies. *bioRxiv* 2020.03.31.018325, https://doi.org/10.1101/2020.03.31.018325 (2020).
46. Utgés, J. S., Tsenkov, M. I., Dietrich, N. J. M., MacGowan, S. A. & Barton, G. J. Ankyrin repeats in context with human population variation. *PLoS Comput. Biol.* **17**, e1009335 (2021).
47. Betts, M. J. *et al*. Systematic identification of phosphorylation-mediated protein interaction switches. *PLoS Comput. Biol.* **13**, e1005462 (2017).
48. Li, B., Roden, D. M. & Capra, J. A. The 3D mutational constraint on amino acid sites in the human proteome. *Nat. Commun.* **13**, 3273 (2022).
49. Xu, Q. *et al*. Identifying three-dimensional structures of autophosphorylation complexes in crystals of protein kinases. *Sci. Signal.* **8**, rs13 (2015).
50. Modi, V. & Dunbrack, R. L. Jr. Kincore: a web resource for structural classification of protein kinases and their inhibitors. *Nucleic Acids Res.* **50**, D654–D664 (2022).
51. Frappier, V., Duran, M. & Keating, A. E. PixelDB: Protein–peptide complexes annotated with structural conservation of the peptide binding mode. *Protein Sci.* **27**, 276–285 (2018).
52. Gao, J. *et al*. 3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets. *Genome Med.* **9**, 4 (2017).
53. Flock, T. *et al*. Universal allosteric mechanism for Gα activation by GPCRs. *Nature* **524**, 173–179 (2015).
54. Hashemi, S., Nowzari Dalini, A., Jalali, A., Banaei-Moghaddam, A. M. & Razaghi-Moghadam, Z. Cancerouspdomains: comprehensive analysis of cancer type-specific recurrent somatic mutations in proteins and domains. *BMC Bioinformatics* **18**, 370 (2017).
55. Andreeva, A., Kulesha, E., Gough, J. & Murzin, A. G. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Res.* **48**, D376–D382 (2020).
56. Finn, R. D. *et al*. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–285 (2016).
57. Ferla, M. P., Pagnamenta, A. T., Koukouflis, L., Taylor, J. C. & Marsden, B. D. Venus: Elucidating the Impact of Amino Acid Variants on Protein Function Beyond Structure Destabilisation. *Comput. Resour. Mol. Biol.* **434**, 167567 (2022).
58. Ofoegbu, T. C. *et al*. PhyreRisk: A Dynamic Web Application to Bridge Genomics, Proteomics and 3D Structural Data to Guide Interpretation of Human Genetic Variants. *Comput. Resour. Mol. Biol.* **431**, 2460–2466 (2019).
59. Slodkowicz, G. & Goldman, N. Integrated structural and evolutionary analysis reveals common mechanisms underlying adaptive evolution in mammals. *Proc. Natl. Acad. Sci.* **117**, 5977–5986 (2020).
60. Zerbino, D. R., Frankish, A. & Flicek, P. Progress, Challenges, and Surprises in Annotating the Human Genome. *Annu. Rev. Genomics Hum. Genet.* **21**, 55–79 (2020).
61. Tate, J. G. *et al*. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
62. Gligorijević, V. *et al*. Structure-based protein function prediction using graph convolutional networks. *Nat. Commun.* **12**, 3168 (2021).
63. Coudert, E. *et al*. Annotation of biologically relevant ligands in UniProtKB using ChEBI. *Bioinformatics* **39**, btac793 (2023).
64. Huttlin, E. L. *et al*. Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *Cell* **184**, 3022–3040.e28 (2021).
65. Sargsyan, K., Mazmanian, K. & Lim, C. A strategy for evaluating potential antiviral resistance to small molecule drugs and application to SARS-CoV-2. *Sci. Rep.* **13**, 502 (2023).
66. Baek, M. *et al*. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
67. Sehnal, D. *et al*. Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Res.* **49**, W431–W437 (2021).
68. Combs, A. P. Recent Advances in the Discovery of Competitive Protein Tyrosine Phosphatase 1B Inhibitors for the Treatment of Diabetes, Obesity, and Cancer. *J. Med. Chem.* **53**, 2333–2344 (2010).
69. Han, Y. *et al*. Discovery of [(3-bromo-7-cyano-2-naphthyl)(difluoro)methyl]phosphonic acid, a potent and orally active small molecule PTP1B inhibitor. *Bioorg Med Chem Lett* **18**, 3200–3205 (2008).
70. Scapin, G. *et al*. The Structural Basis for the Selectivity of Benzotriazole Inhibitors of PTP1B. *Biochemistry* **42**, 11451–11459 (2003).
71. Barford, D., Flint, A. J. & Tonks, N. K. Crystal Structure of Human Protein Tyrosine Phosphatase 1B. *Science* **263**, 1397–1404 (1994).
72. Puius, Y. A. *et al*. Identification of a second aryl phosphate-binding site in protein-tyrosine phosphatase 1B: A paradigm for inhibitor design. *Proc. Natl. Acad. Sci.* **94**, 13420–13425 (1997).
73. Deshpande, M. *et al*. *PDB ProtVista: A reusable and open-source sequence feature viewer* https://doi.org/10.1101/2022.07.22.500790 (2022).
74. Sehnal, D. *et al*. BinaryCIF and CIFTools—Lightweight, efficient and extensible macromolecular data management. *PLOS Comput. Biol.* **16**, e1008247 (2020).
75. Vallat, B. *et al*. New system for archiving integrative structures. *Acta Crystallogr. Sect. D* **77**, 1486–1496 (2021).
76. Vallat, B. *et al*. ModelCIF: An extension of PDBx/mmCIF data representation for computed structure models. *J. Mol. Biol.* 168021, https://doi.org/10.1016/j.jmb.2023.168021 (2023).
77. Wojdyr, M. GEMMI: A library for structural biology. *J. Open Source Softw.* **7**, 4200 (2022).

Acknowledgements

This paper is dedicated to the fond memory of our dear collaborator and wwPDB member John Westbrook. John critically reviewed the new “SIFTS-specific” categories in PDBx/mmCIF data dictionary and provided valuable feedback. We also thank Ezra Peisach for his valuable comments while updating the PDBx/mmCIF data dictionary. Both John Westbrook and Ezra Peisach are members of the RCSB Protein Data Bank, a co-founder of the wwPDB along with PDBj and PDBe. We also thank EMBL and EMBL-EBI for providing the necessary infrastructure to run this process weekly. This work was supported by funding from EMBL and funding awarded to PDBe by the UK Biotechnology and Biological Research Council (BB/V004247/1, PI:Sameer Velankar) and RCSB PDB by the NSF (DBI-2019297, PI: S.K. Burley) supporting development of a Next Generation PDB archive.

Author contributions

Conceptualization: P.C., J.B., S.V.; Methodology: P.C., J.T., J.B.; Software: P.C., S.A.; Investigation: P.C., J.B., S.V.; Funding: S.V.; Writing - Original Draft: P.C.; Writing - Review and Editing: P.C., M.V., S.V.; All authors read and approved the final version of the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to P.C.