

Protein Data Bank (PDB) Archive: a new architecture (beta) for scalable, PDBx/mmCIF-based data distribution

Zukang Feng,^a Balakumaran Balasubramanian,^b Gert-Jan Bekker,^c Jose M. Duarte,^d Vladimir Guranovic,^a Jeremy Henry,^d Sreenath S. Nair,^b Ezra Peisach,^a Dennis W. Piehl,^a Aditya Pingale,^a James Smith,^a Brinda Vallat,^a Reiko Yamashita,^c Arthur Zalevsky,^e Kyle Morris,^f Jeff Hoch,^g Genji Kurisu,^{c,h} Sameer Velankar,^b Stephen K. Burley^{a,i,j,k,l} and Jasmine Y. Young^{a,*}

Received 15 April 2026

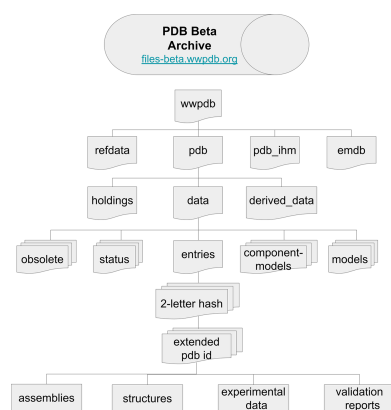
Accepted 12 June 2026

Edited by R. McKenna, University of Florida, USA

Keywords: Protein Data Bank; PDB; macromolecular structure; mmCIF; structural biology; data archive.

^aRCSB Protein Data Bank, Rutgers, The State University of New Jersey, Piscataway, New Jersey, USA, ^bPDBE, EMBL–European Bioinformatics Institute, Hinxton CB10 1SD, United Kingdom, ^cProtein Data Bank Japan, Institute for Protein Research, The University of Osaka, Osaka 565-0871, Japan, ^dRCSB Protein Data Bank, San Diego Supercomputer Center, University of California San Diego, San Diego, California, USA, ^eRCSB Protein Data Bank, University of California San Francisco, San Francisco, California, USA, ^fEMDB, EMBL–European Bioinformatics Institute, Hinxton CB10 1SD, United Kingdom, ^gBMRB, UConn Health, Farmington, Connecticut, USA, ^hProtein Data Bank Japan, Protein Research Foundation, Minoh, Osaka 562-8686, Japan, ⁱInstitute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, New Jersey, USA, ^jRutgers Cancer Institute, Rutgers, The State University of New Jersey, Piscataway, New Jersey, USA, ^kRutgers Artificial Intelligence and Data Science Collaboratory, Rutgers, The State University of New Jersey, Piscataway, New Jersey, USA, and ^lDepartment of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, New Jersey, USA. *Correspondence e-mail: jasmin@rcsb.rutgers.edu

With the continuous growth of the Protein Data Bank archive, the Worldwide Protein Data Bank (wwPDB) partnership anticipates that the entire complement of possible four-character PDB accession codes (for example 1ABC) will be exhausted by 2028. wwPDB is, therefore, revising the PDB accession code (PDB ID) to 12 characters by extending its length and prepending ‘pdb_’ (for example pdb_1000axyz) in lower case. This change will enable the robust detection of references to PDB entries in published literature. On or about July 21st 2027, the PDB will convert to releasing entries with extended PDB IDs only, which will not be compatible with the legacy PDB format. A beta version of the PDB Archive (PDB Beta Archive) is now available to help communities adapt to and embrace the extended PDB IDs and PDBx/mmCIF format during a transition phase. All files in the current PDB archive are reorganized in the Beta Archive with extended PDB IDs (including file naming and directories) on an entry-level basis, mirroring the data organization of the PDB Versioned Archive. wwPDB encourages scientific journals, PDB community members and users to transition to the PDBx/mmCIF format and adopt the new PDB ID format as early as possible. The PDB Beta Archive will replace the current public archive, and 12-character PDB IDs will be solely assigned to all newly deposited PDB entries.



1. Introduction

The Protein Data Bank (PDB) was established in 1971 (Protein Data Bank, 1971; wwPDB Consortium, 2019) as the first open-access digital data resource in biology, initially comprising just seven X-ray crystal structures of proteins. Today, it houses more than 250 000 experimentally determined three-dimensional (3D) structures of biological macromolecules that are freely used by many millions of PDB data consumers worldwide (Berman, 2008; Berman & Burley, 2025). This wealth of information serves as a cornerstone for basic and applied research and education endeavours across fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences.

The Worldwide Protein Data Bank partnership (wwPDB, <https://wwpdb.org>; Berman *et al.*, 2003) currently consists of five full members (RCSB Protein Data Bank or RCSB PDB, Protein Data Bank in Europe or PDBe, Protein Data Bank Japan or PDBj, Biological Magnetic Resonance Data Bank or BMRB, and Electron Microscopy Data Bank or EMDB) and one associate member (Protein Data Bank China or PDBc; Xu *et al.*, 2023). The wwPDB partners jointly manage three Core Archives (wwPDB Consortium, 2019), including PDB (archiving atomic coordinates, crystallographic data and NMR chemical shifts/restraints), EMDB (archiving 3DEM density maps) and BMRB (archiving NMR experimental data), all the while adhering to the FAIR (Findability, Accessibility, Interoperability, Reusability; Wilkinson *et al.*, 2016) and FACT (Fairness, Accuracy, Confidentiality, Transparency; van der Aalst *et al.*, 2017) principles emblematic of responsible data stewardship.

Stemming from advances in technology and methodologies, the number, complexity and size of biomolecular structural studies have grown rapidly, posing challenges for data archiving. Advances across structure-determination pipelines include improvements in sample preparation and delivery, the development and refinement of data-collection instruments and detectors, innovations in data processing and analysis algorithms, and increased computational power. Rapidly evolving experimental approaches, such as cryo-electron microscopy (cryo-EM), which has improved in average resolution from >14 Å to a best resolution of ~ 1 Å since 2013 (Kühlbrandt, 2014), X-ray free-electron lasers and serial crystallography for time-resolved studies, integrative/hybrid methods (IHM) that combine multiple experimental and computational techniques (Vallat *et al.*, 2025), and high-throughput screening of protein libraries with small molecules have further accelerated progress. Advances in experimental pipelines, coupled with computational and AI-driven modelling, and the structural biology community's commitment to open data sharing, have driven substantial growth of the PDB archive.

The legacy PDB file format (Bernstein *et al.*, 1977), based on the 80-column Hollerith punch-card standard, has limitations on the number of columns and characters, including identifiers for molecules described in the Chemical Component Dictionary (CCD IDs; Westbrook *et al.*, 2015) and PDB accession codes (PDB IDs). Previously, wwPDB took steps to address the limitation in the PDB legacy format. Initially, historically large ribosome structures were split into several PDB files due to limits on the number of atoms and/or polymer chain IDs supported by the legacy PDB format, which hinder the ability for users to gain structural insight into the full structure and wwPDB realization of FAIR principles. In 2014, all large structures that were previously split into multiple PDB entries were combined into single files formatted according to the PDB Exchange/macromolecular Crystallographic Information Framework (PDBx/mmCIF; Fitzgerald *et al.*, 2005; Westbrook *et al.*, 2022). Another limitation of the legacy PDB format is the length restriction of residue names to three characters, as defined in the CCD.

Recently, all three-character CCD IDs were consumed. Newly deposited novel ligands are now assigned five-character CCD IDs, which are no longer compatible with the legacy PDB file format. Finally, with the continued growth of PDB depositions, wwPDB anticipates that four-character PDB IDs will be fully consumed by 2028. wwPDB has given careful consideration in defining the new extended PDB ID code to both facilitate text mining and support deposition growth.

For more than a decade, wwPDB has enhanced the content of the PDB archive by introducing innovative, new features: PDB Archive (<https://files.wwpdb.org>) provides the historical archive, PDB Versioned Archive (<https://files-versioned.wwpdb.org>) enables the distribution of different versions of PDB entries and PDB NextGen Archive (<https://files-nextgen.wwpdb.org>) distributes enhanced information available in the individual PDB entries, to support the diverse needs of our very broad user community. PDB Archive is the primary repository, which distributes the latest versions of all data files for each PDB entry.

Versioning of PDB entries (stored in the PDB Versioned Archive) was introduced in 2017 and distributes the latest versions of every major version of structure model files for each PDB entry. Necessary modifications made to a PDB entry after its initial release can be either 'major' or 'minor'. Updates to atomic coordinates, polymer sequence or chemical description in the coordinate file trigger a major version increment. Other changes to the metadata in the coordinate file are identified with a minor version increment. In both cases, the originally issued PDB accession code is retained. To enable tracking of changes between versions, new revision data categories were defined in the PDBx/mmCIF dictionary (https://mmcif.wwpdb.org/dictionaries/mmcif_pdbx_v50.dic/Groups/audit_group.html). The revision trail is provided within the PDBx/mmCIF-formatted coordinate files. This PDB Versioned Archive has adopted the new extended PDB ID in the directory and file naming in the archive layout.

The PDB NextGen Archive (Choudhary *et al.*, 2024) was created in 2023 as a prototype to provide enriched annotations for the latest version of each entry compared with the PDB Archive. This PDB NextGen Archive incorporates additional metadata annotations from external data resources within the structure model files in the PDB Archive. For example, up-to-date sequence annotations from the SIFTS resource (Dana *et al.*, 2019) to external resources such as UniProt, SCOP2 and Pfam are provided in this newer archive. The layout of the NextGen Archive follows the same layout as the Versioned Archive, using the extended PDB ID in the directory and file naming. To render the PDB Archive more sustainable while enhancing accessibility and scalability, wwPDB has restructured a beta version of the PDB Archive with data reorganization based on the new extended PDB ID at an entry-level basis.

Herein, we describe the beta version of the PDB Archive (henceforth referred to as the PDB Beta Archive), incorporating the new updates, with the goal of supporting PDB users as they transition to the extended PDB ID and PDBx/mmCIF

format. Eventually, the PDB Beta Archive will replace the current PDB Archive once 12-character PDB IDs are issued exclusively. Users are strongly encouraged to adapt to these changes in PDB data layout and content as soon as possible.

2. Methods

2.1. Extended PDB ID format

To support the ongoing growth of the PDB archive and make it more FAIR, wwPDB has extended the PDB accession code to be 12 characters in length with the prefix ‘pdb_’ (e.g. `pdb_1000axyz`) in lower case. The prefix ‘pdb_’ in the extended PDB IDs will facilitate text mining of PDB entries referenced in the published literature. PDB entries bearing only extended PDB IDs (*i.e.* those that do not have a corresponding four-character ID) are incompatible with legacy PDB format and will be distributed only in PDBx/mmCIF and PDBML (XML) formats.

Since the late 1990s, the PDB Archive has provided entries in PDBx/mmCIF format (Westbrook *et al.*, 2022; <https://mmcif.wwpdb.org>). The PDBx/mmCIF framework was built atop the self-validating Data Definition Language Version 2 (DDL2) dictionary (Westbrook *et al.*, 2005), which defines metadata (e.g. data types, allowed ranges, controlled vocabularies) and parent–child relationships. The PDBx/mmCIF dictionary explicitly defines all relationships between data items (e.g. atom and residue identifiers), which enables software applications to evaluate and validate referential integrity for any biostructure data represented in this form (e.g. within a PDB entry) and maps information between two related categories (e.g. the sample sequences and the observed sequences from the model coordinates). The PDBx/mmCIF dictionary is fully extensible and both human- and machine-readable, allows the accommodation of metadata from new and increasingly complex experimental methods such as serial crystallography, X-ray free-electron lasers and integrative/hybrid methods (Sali, 2021), and helps to ensure that new metadata items conform to data standards.

Since 2014, PDBx/mmCIF has become the master PDB archival format (Berman *et al.*, 2014) when the OneDep system (Young *et al.*, 2017) was launched for deposition, validation (Feng *et al.*, 2021; Gore *et al.*, 2017) and biocuration (Young *et al.*, 2018) of incoming PDB and EMDB entries. The PDBx/mmCIF format has no limitations. It can accommodate ribosomes and other large structures in a single file and supports better realization of FAIR principles. This format is extendable to support new methods development and the increasing complexity of structures. PDBx/mmCIF can be programmatically parsed, searched and edited, and is managed by the wwPDB and the mmCIF Working Group consisting of community experts (<https://www.wwpdb.org/task/mmcif>; Westbrook *et al.*, 2022).

All PDB entries, including IHM structures, issued with four-character PDB IDs have been assigned an extended PDB ID by prepending ‘pdb_0000’ to the four-character ID (e.g. PDB ID 1ABC will have the extended PDB ID `pdb_00001abc`). For

such entries, the PDB Digital Object Identifier (DOI) format with ‘pdb’ prefix, `10.2210/pdb<four-character PDB ID>/pdb`, remains unchanged, as necessary to respect FAIR data practices (e.g. `10.2210/pdb1abc/pdb`). Both the four-character and extended PDB ID are stored in the `database_2` category of the PDBx/mmCIF model file:

```
loop_
  _database_2.database_id
  _database_2.database_code
  _database_2.pdbx_database_accession
  _database_2.pdbx_DOI
PDB      1ABC      pdb_00001abc      10.2210/pdb1abc/pdb
```

PDB entries identified solely with extended PDB IDs issued (e.g. `pdb_1000axyz`) will have the same 12-character PDB ID in both the `_database_2.database_code` and `_database_2.pdbx_database_accession` data items. Such entries will have a new PDB DOI format based on the extended PDB ID: `10.2210/<12-character PDB ID>/pdb` (e.g. `10.2210/pdb_1000axyz/pdb`):

```
loop_
  _database_2.database_id
  _database_2.database_code
  _database_2.pdbx_database_accession
  _database_2.pdbx_DOI
PDB      pdb_1000axyz      pdb_1000axyz      10.2210/pdb_1000axyz/pdb
```

2.2. Creation of PDB Beta Archive with extended PDB IDs

A new PDB Beta Archive was created to allow better presentation of all the abovementioned changes and enriched annotations, and help users adapt to extended PDB IDs and the PDBx/mmCIF format. This PDB Beta Archive contains PDB entries in PDBx/mmCIF and PDBML formats, with extended PDB IDs in the datablock name. It is organized on an entry-level basis (e.g. all associated experimental and validation files for a particular entry are contained within the same directory) with extended PDB IDs used in the file and directory naming.

Entry data files (e.g. structure model and structure-factor files) with extended PDB IDs in PDBx/mmCIF-formatted files and file naming are generated by the OneDep release module alongside the data files with four-character PDB IDs in PDBx/mmCIF-formatted files and file naming, which will remain accessible in the current main archive when four-letter PDB ID code issuance is discontinued in July 2027. The wwPDB weekly update process was modified to create the PDB Beta Archive alongside the current PDB Archive. Both PDB Beta and the current main archives are updated in parallel each week and made available every Wednesday at 00:00 UTC.

The PDB-IHM data pipeline and workflows were extended to generate updated files for the Beta Archive during the weekly update process, alongside the PDB-IHM branch of the

current main archive. The model files and validation reports are organized at the entry level and follow the 12-character extended PDB ID naming convention, as for the PDB Beta Archive.

3. Results

A new PDB Beta Archive is available at <https://files-beta.wwpdb.org> to help communities adapt to and adopt extended PDB ID and PDBx/mmCIF format during the transition phase. A documentation page, <https://www.wwpdb.org/ftp/pdb-beta-ftp-sites>, with download instructions (including a script and URL short links) is provided to help users adapt to the new file organization and formats in the PDB Beta Archive. Existing legacy PDB-format files are retained and maintained wherever possible in this archive. Once four-character PDB IDs are consumed, newly deposited PDB entries, including the IHM structures, will only be issued extended PDB IDs and will only be distributed in PDBx/mmCIF and PDBML formats.

3.1. Data structure and file-naming conventions

A high-level subdirectory (`/wwpdb/`) under the `/pub` top directory has been introduced within the PDB Beta Archive (<https://files-beta.wwpdb.org/pub/wwpdb/>). Within this `/wwpdb` subdirectory are subdirectories for each data type, including small-molecule or CCD data (`/refdata/`), PDB data (`/pdb/`), IHM data (`/pdb_ihm/`) and EMDB data (`/emdb/`). Fig. 1 illustrates the tree structure data organization of the PDB Beta Archive. Within the `/pub/wwpdb/pdb` subdirectory are PDB data, derived data (*i.e.* wwPDB-computed data files based on PDB data *etc.*) and PDB holding files (*i.e.* inventory index files that offer a quick overview of archival contents).

All data files associated with a given entry, such as structure model, experimental data, assemblies and validation reports, are stored on an entry-level basis using the extended PDB ID as the folder name under a hash directory, <https://files-beta.wwpdb.org/pub/wwpdb/pdb/data/entries/>. The two-character hash is generated from the penultimate two characters of the PDB code (*i.e.* the second and third characters from the last character, https://files-beta.wwpdb.org/pub/wwpdb/pdb/data/entries/<two-letter-hash>/<pdb_accession_code>/<entry_data_file_names>). For example, PDB entry `pdb_1abc5678` will be stored under the `/67/` subdirectory. This measure will maintain consistency with the current PDB archive: PDB entry `1ABC` (with extended PDB ID `pdb_00001abc`) is stored under `/ab/`.

File naming has been standardized, such that the file type is used for the extension. For example, the file naming for structure-factor data was changed from `r116dsf.ent.gz` to `pdb_0000116d-sf.cif.gz` with a `.cif` extension as the file type and from `pdb318d.ent.gz` to `pdb_0000318d.pdb.gz` with a `.pdb` extension as the file type for the legacy PDB-formatted coordinate file.

3.2. File access and short links

The PDB Beta Archive is updated every Wednesday at 00:00 UTC. wwPDB has provided various download protocols, including `https` and `rsync`:

wwPDB: `https://files-beta.wwpdb.org`, `rsync://rsync-beta.wwpdb.org`.

RCSB PDB (US): `https://files-beta.rcsb.org`, `rsync://rsync-beta.rcsb.org`.

PDBe (UK): `https://ftp.ebi.ac.uk/pub/databases/wwpdb/`.

PDBj (Japan): `ftp://ftp-beta.pdbj.org`, `https://files-beta.pdbj.org`, `rsync://rsync-beta.pdbj.org`.

To help PDB users readily access data files in the PDB Beta Archive, wwPDB has provided download short links that obviate the need for prior knowledge of file paths. Table 1 lists the file paths for each data type and provides download short links. To use these short links, simply replace the example ID, `pdb_00001abc`, with a real PDB ID.

To perform a batch download of all PDB entries of a file type or format from the PDB Beta Archive, wwPDB also provides a download script, <https://cdn.rcsb.org/wwpdb/docs/BetaArchiveBatchDownloader.py>. It uses an asynchronous `aiohttp` library to download multiple files in parallel. It requires Python 3.8 or higher, plus the `aiofiles` and `aiohttp`

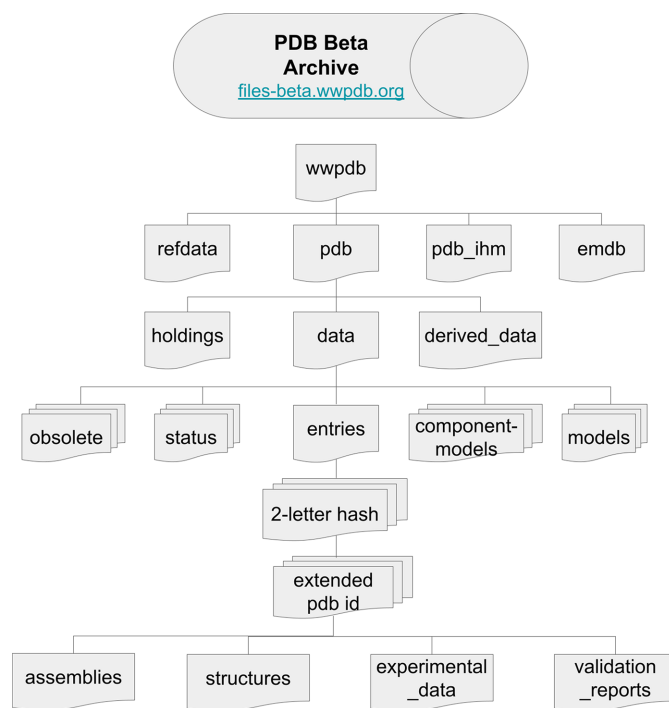


Figure 1

Organization of data within the PDB Beta Archive. A `/wwpdb` directory is introduced under the `/pub` top directory. Small-molecule data are stored in the `/refdata/` subdirectory. PDB data, derived data and PDB holding files, which are inventory data files that offer a quick overview of data in the archive, are stored under the `/pdb/` subfolder. 3D structure atomic coordinate model files, experimental data files, assembly files and validation reports associated with each entry are stored under its extended PDB ID subfolder under a hash directory. The `/pdb_ihm/` folder consists of `/data/` and `/holdings/` subfolders corresponding to IHM structures (not shown) (Johnson *et al.*, 2024; Baldwin *et al.*, 2024).

Table 1
PDB Beta Archive data organization and download short links.

Data types	File formats	Location	Short link
Atomic coordinates	PDBx/mmCIF, PDBML and PDB	https://files-beta.wwpdb.org/pub/wwpdb/pdb/data/entries/[two-letter-hash]/[extended-PDB-ID]/structures/	https://files-beta.wwpdb.org/download/[extended-PDB-ID].[extension].gz , e.g. https://files-beta.wwpdb.org/download/pdb_00001abc.cif.gz
X-ray data: structure factors	PDBx/mmCIF	https://files-beta.wwpdb.org/pub/wwpdb/pdb/data/entries/[two-letter-hash]/[extended-PDB-ID]/experimental_data/	https://files-beta.wwpdb.org/download/[extended-PDB-ID]-sf.cif.gz , e.g. https://files-beta.wwpdb.org/wodnload/pdb_00001abc-sf.cif.gz
NMR data: restraints and chemical shifts	NEF, NMR-STAR and native refinement program formats	https://files-beta.wwpdb.org/pub/wwpdb/pdb/data/entries/[two-letter-hash]/[extended-PDB-ID]/experimental_data/	https://files-beta.wwpdb.org/download/[extended-PDB-ID].[NMR data format].gz , e.g. https://files-beta.wwpdb.org/download/pdb_00001abc_nmr-data.str.gz , https://files-beta.wwpdb.org/download/pdb_00001abc_nmr-data.nef.gz
Small-molecule references			
CCD	PDBx/mmCIF	https://files-beta.wwpdb.org/pub/wwpdb/refdata/chem_comp/[last-character-hash]/[CCD ID]/	https://files-beta.wwpdb.org/ligands/download/[CCD ID].cif , e.g. https://files-beta.wwpdb.org/ligands/download/ATP.cif
BIRD		https://files-beta.wwpdb.org/pub/wwpdb/refdata/bird , e.g. https://files-beta.wwpdb.org/pub/wwpdb/refdata/bird/prd/[last-character-hash]†	https://files-beta.wwpdb.org/birds/download/[BIRD ID].cif , e.g. https://files-beta.wwpdb.org/birds/download/PRD_000006.cif
Other derived data		https://files-beta.wwpdb.org/pub/wwpdb/refdata/derived_data/	
CCD holdings	JSON	A list of released chemical reference entries, their content types (e.g. Chemical Component, BIRD) and the most recent modification date of the reference file, https://files-beta.wwpdb.org/pub/wwpdb/refdata/derived_data/refdata_id_list.json.gz .	
SMILES, InChI, InChIKey Variants	SDF, smi, inch		
Assemblies	PDBx/mmCIF	https://files-beta.wwpdb.org/pub/wwpdb/pdb/data/entries/ , https://files-beta.wwpdb.org/pub/wwpdb/pdb/data/entries/[two-letter-hash]/[extended-PDB-ID]/assemblies/	https://files-beta.wwpdb.org/download/[extended-PDB-ID]-assembly[number].cif.gz , e.g. https://files-beta.wwpdb.org/download/pdb_00001abc-assembly1.cif.gz
Archive holdings	JSON	https://files-beta.wwpdb.org/pub/wwpdb/pdb/holdings/	None, e.g. https://files-beta.wwpdb.org/pub/wwpdb/pdb/holdings/current_file_holdings.json.gz

† Note that paths are case-sensitive: use capital letters for hash and IDs.

packages. Detailed instructions can be found at <https://www.wwpdb.org/ftp/pdb-beta-ftp-sites#downloadPor>.

4. Discussion

4.1. Communication/community engagement

The development and maintenance of PDBx/mmCIF in collaboration with the macromolecular crystallography (MX) community led to the formation of the wwPDB mmCIF Working Group. The Working Group was initially established in 2011 to enable the use of PDBx/mmCIF-format files in major macromolecular crystallographic software tools, and to provide recommendations on format extensions required for deposition of larger structures to the PDB. This working group meets regularly with wwPDB leadership and select subject-matter experts to guide dictionary content and promotes the standard among structural biology methods developers.

PDBx/mmCIF has been adopted by the community broadly including major refinement-software packages [*Phenix* (Adams *et al.*, 2010), *REFMAC* (Murshudov *et al.*, 2011), *Global Phasing* (Blanc *et al.*, 2004; Bricogne *et al.*, 2009)], visualization programs [*PyMOL* (DeLano, 2002), *ChimeraX* (Meng *et al.*, 2023), *CCP4mg* (McNicholas *et al.*, 2011), *Mol** (Sehna *et al.*, 2021) *etc.*], *pdb_extract* tool (Yang *et al.*, 2004), *PDBj CIF Editor* and various mmCIF parsers (e.g. *py-mmCIF*,

Gemmi, *iotxb.cif*, *mmcif_pdbx*). The adoption of PDBx/mmCIF format was extended to other communities with the development of *ModelCIF* (Vallat *et al.*, 2023) for computed structure models and *IHMCIF* (Vallat *et al.*, 2024) for integrative/hybrid structures. In 2019, PDBx/mmCIF was made the mandatory format for the deposition of macromolecular crystallography structures to the PDB (Adams *et al.*, 2019).

wwPDB continues to engage with the editors of the top 25 scientific journals that publish most macromolecular structures deposited to the PDB to inform them of the extended PDB ID format (in text, tables and data-availability statements) and corresponding PDB DOI links in their pipelines. Early adoption will contribute to the long-term sustainability and interoperability of 3D biostructure data across the scientific community.

For all extant PDB entries with four-character PDB IDs, PDB DOIs are formatted as 10.2210/pdb[four-character-PDB-ID]/pdb and can be used to link to the corresponding wwPDB DOI landing page. For example, the PDB entry with ID 8Y9M (and extended PDB ID pdb_00008y9m) has the DOI <https://doi.org/10.2210/pdb8y9m/pdb>. Importantly, previously issued DOIs are persistent and will remain unchanged, following FAIR data best practices.

Once the issuance of four-character PDB IDs ceases in July 2027, all new PDB entries, including IHM structures in

Table 2

Resources and locations for extended PDB ID and PDBx/mmCIF.

Resource types	Resources	Locations
FAQ	Extended PDB ID	https://www.wwpdb.org/documentation/pdb-id-extension-faq
FAQ	PDBx/mmCIF	https://mmcif.wwpdb.org/docs/faqs/pdbx-mmcif-faq-general.html
Dictionary	PDBx/mmCIF	https://mmcif.wwpdb.org
Tutorials	PDBx/mmCIF	https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/beginner%E2%80%99s-guide-to-pdbx-mmcif
User guide	PDBx/mmCIF	https://mmcif.wwpdb.org/docs/user-guide/guide.html
Webinar	PDBx/mmCIF	https://pdb101.rcsb.org/train/training-events/mmcif
Webinar	PDB Beta Archive	https://pdb101.rcsb.org/train/training-events
Tools	<i>pdb_extract</i>	https://pdb-extract.wwpdb.org/
Tools	<i>CIF Editor</i>	https://pdbj.org/cif-editor
Tools	mmCIF parser and format conversion	https://mmcif.wwpdb.org/docs/software-resources.html
Tools	mmCIF <i>DB-Loader</i>	https://sw-tools.rcsb.org/apps/DB-LOADER/index.html
Tools	PostgreSQL database updater	https://gitlab.com/pdbjapan/mine2updater/
Download script and instruction	PDB Beta Archive	https://www.wwpdb.org/ftp/pdb-beta-ftp-sites
Archive	PDB Beta Archive	https://files-beta.wwpdb.org

PDB-IHM, will only be issued extended PDB IDs, and their corresponding PDB DOIs will be formatted as 10.2210/[extended_PDB_ID]/pdb, leading to the corresponding wwPDB DOI landing page. For example, a PDB entry with ID pdb_10001xyz will have the DOI https://doi.org/10.2210/pdb_10001xyz/pdb.

wwPDB has also reached out broadly to technically diverse research and technology communities, including structural biology, visualization tools, bioinformatics, cheminformatics, computational structure prediction and third-party database resources that utilize PDB data. We are strongly encouraging them to transition to extended PDB ID and PDBx/mmCIF formats through individual email communication, news announcements via various community mailing lists, poster and oral presentations at professional society meetings, and offering webinars and virtual office hours focused on these topics.

4.2. Training materials and tools for adoption

Many training materials have been created to help users learn about extended PDB IDs and PDBx/mmCIF, including an mmCIF user guide, FAQ pages, recorded webinars and tutorials. Resource tools for mmCIF parsers and file-format conversion are available at <https://mmcif.wwpdb.org/docs/software-resources.html> for users to extract information from PDBx/mmCIF files or convert PDBx/mmCIF to PDB format, if needed, until all users have adopted PDBx/mmCIF. For example, the wwPDB software tool *Maxit* can be used to convert PDBx/mmCIF to PDB format for PDB entries with four-character PDB IDs and without five-character CCD IDs. Alternatively, a community software tool, such as *Gemmi*, can convert PDBx/mmCIF to PDB format with minimal support for metadata and accession codes. For database users, a *DB-Loader* tool for loading data from PDBx/mmCIF files into a database is available for user adoption at <https://sw-tools.rcsb.org/apps/DB-LOADER/index.html>. *Mine2Updater* can be used to update a PostgreSQL database with weekly PDB releases: <https://gitlab.com/pdbjapan/mine2updater/>.

wwPDB also provides a CIF editor (see Table 2) that allows users to prepare and modify their own PDBx/mmCIF files. In addition, the *pdb_extract* tool allows PDB depositors to

generate a PDBx/mmCIF file with common metadata that can be used to prepare a complete PDBx/mmCIF file for depositions via OneDep. wwPDB partners, and the software developers (CCP4, Phenix, CCP-EM, Global Phasing and others), are actively working to support the generation of PDBx/mmCIF files for data deposition. Table 2 provides information regarding available resources, and where users can access them to learn and utilize these tools to work with PDBx/mmCIF format files.

5. Conclusion and future perspectives

PDBx/mmCIF is the file format and archival data standard of the PDB. Its flexible and infinitely extensible format allows the accurate representation of large structures, while supporting new/evolving structure-determination methodologies. It is both human- and machine-readable. PDBx/mmCIF-format file contents can be parsed programmatically, or using three-dimensional visualization tools. As the PDB archive grows, PDBx/mmCIF- and PDBML-format files will become the exclusive means of distributing atomic coordinate information and related metadata.

Many resources have been made available to assist users in transitioning to extended PDB ID and PDBx/mmCIF. wwPDB is committed to providing additional resources, including webinars and virtual office hours, as needed for a smooth transition.

5.1. Timeline for user adoption

The PDB Beta Archive will become the PDB main core archive in July 2027. The <https://files-beta.wwpdb.org> URL will point to the main archive (<https://files.wwpdb.org>) for some time. Thereafter, the PDB Beta Archive URL (not the PDB Beta Archive) will be retired. PDB entries with extended PDB IDs only will not be compatible with the legacy PDB format. wwPDB strongly encourages scientific journals, PDB depositors and the PDB user community writ large to transition to the PDBx/mmCIF format and adopt extended PDB IDs as soon as possible.

When providing extended PDB IDs to scientific journals and citing extended PDB IDs in submitted manuscripts, all 12

characters, including the prefix 'pdb_', should be included (in lower case). After mid-2027, scientific journals should be routinely accepting 12-character extended PDB IDs with their new DOI format (e.g. PDB ID `pdb_10001xyz` will have DOI https://doi.org/10.2210/pdb_10001xyz/pdb), while retaining the existing DOI format for PDB entries with four-character PDB IDs or those starting with `pdb_0000` (e.g. PDB ID `8Y9M` or `pdb_00008y9m` will have the DOI <https://doi.org/10.2210/pdb8y9m/pdb>) for accurate cross-referencing in electronic manuscripts. (N.B. DOIs for existing PDB entries with legacy four-character PDB IDs have not been redefined, because DOIs are intended to be persistent as per the best FAIR data practices.)

5.2. Future plans for PDB archive management

Maintaining multiple PDB archives (PDB, Versioned and NextGen) represents a substantial challenge for the wwPDB. wwPDB addresses this challenge by creating a more sustainable PDB Beta Archive, which is structured to be scalable and extensible. This beta archive will replace the current PDB archive. Looking ahead, wwPDB envisions enhanced support for fragment-screening and time-resolved studies, which require grouping multiple PDB structures and additional metadata to describe overall characteristics of the experimental study. After the PDB Beta Archive becomes the PDB main core archive, this scalable archive will be augmented with content from the NextGen archive, thereby providing users with access to additional enriched annotations and the possibility of grouping multiple related entries. This PDB main core archive has been designed such that it can also be extended to house the versioned archival files, currently held in the Versioned archive. wwPDB partners welcome constructive feedback on potential improvements to the PDB Beta Archive for appropriate consideration.

Acknowledgements

The authors thank the >60 000 depositors from around the world for freely depositing macromolecular structures to the PDB. We thank wwPDB biocurators for testing the software tools used in creating data files for the PDB Beta Archive through the OneDep system. We also gratefully acknowledge the wwPDB outreach team for promoting extended PDB ID and PDBx/mmCIF formats to user communities. Author contributions are as follows. ZF developed software to generate data files, created beta archive release processes and wrote a batch download script; VG and JH set up the DNS domain and PDB Beta Archive at RCSB PDB, and provided the rsync protocol; JMD created download short links; DWP and AP developed file-download loggings and usage analytics; EP validated PDBx/mmCIF data files; BV and AZ generated IHM data files and modified PDB-IHM release processes; JS created the wwPDB file-download instruction page; BB and SSN set up the PDB Beta Archive at PDBe; G-JB and RY set up the PDB Beta Archive at PDBj; JYY created the beta archive concept, managed the development of this work and

wrote the original manuscript; KM, JH, GK, SV and SKB acquired funding to support this work and provided overall project direction. All authors read and contributed to the text.

Conflict of interest

The authors declare that there are no conflicts of interest.

Data availability

The PDB Beta Archive is accessible at <https://files-beta.wwpdb.org>. The PDBx/mmCIF dictionary is available at <https://mmcif.wwpdb.org/>.

Funding information

RCSB PDB core operations are jointly funded by the National Science Foundation (NSF) (DBI-2321666, PI: S. K. Burley), the US Department of Energy (DE-SC0019749, PI: S. K. Burley) and the National Cancer Institute, the National Institute of Allergy and Infectious Diseases and the National Institute of General Medical Sciences of the NIH (R01GM157729, PI: S. K. Burley). PDBj is supported by grants from the Database Integration Coordination Program from the Office of NBDC Program, Japan Science and Technology Agency (JPMJND2205, PI: G. Kurisu), the Basis for Supporting Innovative Drug Discovery and Life Science Research (BINDS) from the Japan Agency for Medical Research and Development (25ama121001, PI: G. Kurisu) and the Joint Usage/Research Center program for Protein Science from the Ministry of Education, Culture, Sports, Science and Technology, Japan. PDBe core operations are funded by the European Molecular Biology Laboratory–European Bioinformatics Institute (EMBL–EBI) and Wellcome Trust (218303/Z/19/Z, PI: S. Velankar). EMDB core operations are funded by the European Molecular Biology Laboratory–European Bioinformatics Institute (EMBL–EBI) and the Wellcome Trust (310300/Z/24/Z). BMRB is supported by the National Institute of General Medical Sciences of the NIH (R24GM150793, PI: J. C. Hoch) and the University of Connecticut Office of the Vice President for Research.

References

- Adams, P. D., Afonine, P. V., Baskaran, K., Berman, H. M., Berrisford, J., Bricogne, G., Brown, D. G., Burley, S. K., Chen, M., Feng, Z., Flensburg, C., Gutmanas, A., Hoch, J. C., Ikegawa, Y., Kengaku, Y., Krissinel, E., Kurisu, G., Liang, Y., Liebschner, D., Mak, L., Markley, J. L., Moriarty, N. W., Murshudov, G. N., Noble, M., Peisach, E., Persikova, I., Poon, B. K., Sobolev, O. V., Ulrich, E. L., Velankar, S., Vonnrhein, C., Westbrook, J., Wojdyr, M., Yokochi, M. & Young, J. Y. (2019). *Acta Cryst.* **D75**, 451–454.
- Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L.-W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C. & Zwart, P. H. (2010). *Acta Cryst.* **D66**, 213–221.
- Baldwin, E. T., van Eeuwen, T., Hoyos, D., Zalevsky, A., Tchesnokov, E. P., Sánchez, R., Miller, B. D., Di Stefano, L. H., Ruiz, F. X., Hancock, M., Işik, E., Mendez-Dorantes, C., Walpole, T., Nichols, C., Wan, P., Riento, K., Halls-Kass, R., Augustin, M., Lammens, A., Jestel, A., Upla, P., Xibinaku, K., Congreve, S., Hennink, M., Rogala, K. B., Schneider, A. M., Fairman, J. E., Christensen, S. M.,

- Desrosiers, B., Bisacchi, G. S., Saunders, O. L., Hafeez, N., Miao, W., Kapeller, R., Zaller, D. M., Sali, A., Weichenrieder, O., Burns, K. H., Götze, M., Rout, M. P., Arnold, E., Greenbaum, B. D., Romero, D. L., LaCava, J. & Taylor, M. S. (2024). *Nature*, **626**, 194–206.
- Berman, H. M. (2008). *Acta Cryst.* **A64**, 88–95.
- Berman, H. M. & Burley, S. K. (2025). *Q. Rev. Biophys.* **58**, e9.
- Berman, H. M., Henrick, K. & Nakamura, H. (2003). *Nat. Struct. Mol. Biol.* **10**, 980.
- Berman, H. M., Kleywegt, G. J., Nakamura, H. & Markley, J. L. (2014). *J. Comput. Aided Mol. Des.* **28**, 1009–1014.
- Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.
- Blanc, E., Roversi, P., Vonnrhein, C., Flensburg, C., Lea, S. M. & Bricogne, G. (2004). *Acta Cryst.* **D60**, 2210–2221.
- Bricogne, G., Blanc, E., Brandl, M., Flensburg, C., Keller, P., Paciorek, W., Roversi, P., Sharff, A., Smart, O. S., Vonnrhein, C. & Womack, T. O. (2009). *Global Phasing*. Cambridge: Global Phasing Ltd.
- Choudhary, P., Feng, Z., Berrisford, J., Chao, H., Ikegawa, Y., Peisach, E., Piehl, D. W., Smith, J., Tanweer, A., Varadi, M., Westbrook, J. D., Young, J. Y., Patwardhan, A., Morris, K. L., Hoch, J. C., Kurisu, G., Velankar, S. & Burley, S. K. (2024). *Database*, **2024**, baee041.
- Dana, J. M., Gutmanas, A., Tyagi, N., Qi, G., O'Donovan, C., Martin, M. & Velankar, S. (2019). *Nucleic Acids Res.* **47**, D482–D489.
- DeLano, W. L. (2002). *The PyMOL Molecular Graphics System*. <https://www.pymol.org>.
- Feng, Z., Westbrook, J. D., Sala, R., Smart, O. S., Bricogne, G., Matsubara, M., Yamada, I., Tsuchiya, S., Aoki-Kinoshita, K. F., Hoch, J. C., Kurisu, G., Velankar, S., Burley, S. K. & Young, J. Y. (2021). *Structure*, **29**, 393–400.
- Fitzgerald, P. M. D., Westbrook, J. D., Bourne, P. E., McMahon, B., Watenpaugh, K. D. & Berman, H. M. (2005). *International Tables for Crystallography*, Vol. G, edited by S. R. Hall & B. McMahon, pp. 295–443. Dordrecht: Springer.
- Gore, S., Sanz García, E., Hendrickx, P. M. S., Gutmanas, A., Westbrook, J. D., Yang, H., Feng, Z., Baskaran, K., Berrisford, J. M., Hudson, B. P., Ikegawa, Y., Kobayashi, N., Lawson, C. L., Mading, S., Mak, L., Mukhopadhyay, A., Oldfield, T. J., Patwardhan, A., Peisach, E., Sahni, G., Sekharan, M. R., Sen, S., Shao, C., Smart, O. S., Ulrich, E. L., Yamashita, R., Quesada, M., Young, J. Y., Nakamura, H., Markley, J. L., Berman, H. M., Burley, S. K., Velankar, S. & Kleywegt, G. J. (2017). *Structure*, **25**, 1916–1927.
- Johnson, A. G., Mayer, M. L., Schaefer, S. L., McNamara-Bordewick, N. K., Hummer, G. & Kranszus, P. J. (2024). *Nature*, **628**, 657–663.
- Kühlbrandt, W. (2014). *Science*, **343**, 1443–1444.
- McNicholas, S., Potterton, E., Wilson, K. S. & Noble, M. E. M. (2011). *Acta Cryst.* **D67**, 386–394.
- Meng, E. C., Goddard, T. D., Pettersen, E. F., Couch, G. S., Pearson, Z. J., Morris, J. H. & Ferrin, T. E. (2023). *Protein Sci.* **32**, e4792.
- Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* **D67**, 355–367.
- Protein Data Bank (1971). *Nature New Biol.* **233**, 223.
- Sali, A. (2021). *J. Biol. Chem.* **296**, 100743.
- Sehnal, D., Bittrich, S., Deshpande, M., Svobodová, R., Berka, K., Bazzig, V., Velankar, S., Burley, S. K., Koča, J. & Rose, A. S. (2021). *Nucleic Acids Res.* **49**, W431–W437.
- Vallat, B., Tauriello, G., Bienert, S., Haas, J., Webb, B. M., Židek, A., Zheng, W., Peisach, E., Piehl, D. W., Anischanka, I., Sillitoe, I., Tolchard, J., Varadi, M., Baker, D., Orengo, C., Zhang, Y., Hoch, J. C., Kurisu, G., Patwardhan, A., Velankar, S., Burley, S. K., Sali, A., Schwede, T., Berman, H. M. & Westbrook, J. D. (2023). *J. Mol. Biol.* **435**, 168021.
- Vallat, B., Webb, B. M., Westbrook, J. D., Goddard, T. D., Hanke, C. A., Graziadei, A., Peisach, E., Zalevsky, A., Sagendorf, J., Tangmunarunkit, H., Voinea, S., Sekharan, M., Yu, J., Bonvin, A., DiMaio, F., Hummer, G., Meiler, J., Tajkhorshid, E., Ferrin, T. E., Lawson, C. L., Leitner, A., Rappsilber, J., Seidel, C. A. M., Jeffries, C. M., Burley, S. K., Hoch, J. C., Kurisu, G., Morris, K., Patwardhan, A., Velankar, S., Schwede, T., Trewella, J., Kesselman, C., Berman, H. M. & Sali, A. (2024). *J. Mol. Biol.* **436**, 168546.
- Vallat, B., Webb, B. M., Zalevsky, A., Tangmunarunkit, H., Sekharan, M. R., Voinea, S., Shafaeibejestan, A., Sagendorf, J., Hoch, J. C., Kurisu, G., Morris, K. L., Velankar, S., Kesselman, C., Burley, S. K., Berman, H. M. & Sali, A. (2025). *J. Mol. Biol.* **437**, 168963.
- van der Aalst, W. M. P., Bichler, M. & Heinzl, A. (2017). *Bus. Inf. Syst. Eng.* **59**, 311–313.
- Westbrook, J., Henrick, K., Ulrich, E. L. & Berman, H. M. (2005). *International Tables for Crystallography*, Vol. G, edited by S. R. Hall & B. McMahon, pp. 195–198. Dordrecht: Springer.
- Westbrook, J. D., Shao, C., Feng, Z., Zhuravleva, M., Velankar, S. & Young, J. (2015). *Bioinformatics*, **31**, 1274–1278.
- Westbrook, J. D., Young, J. Y., Shao, C., Feng, Z., Guranovic, V., Lawson, C., Vallat, B., Adams, P. D., Berrisford, J. M., Bricogne, G., Diederichs, K., Joosten, R. P., Keller, P., Moriarty, N. W., Sobolev, O. V., Velankar, S., Vonnrhein, C., Waterman, D. G., Kurisu, G., Berman, H. M., Burley, S. K. & Peisach, E. (2022). *J. Mol. Biol.* **434**, 167599.
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S. A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. & Mons, B. (2016). *Sci. Data*, **3**, 160018.
- wwPDB Consortium (2019). *Nucleic Acids Res.* **47**, D520–D528.
- Xu, W., Velankar, S., Patwardhan, A., Hoch, J. C., Burley, S. K. & Kurisu, G. (2023). *Acta Cryst.* **D79**, 792–795.
- Yang, H., Guranovic, V., Dutta, S., Feng, Z., Berman, H. M. & Westbrook, J. D. (2004). *Acta Cryst.* **D60**, 1833–1839.
- Young, J. Y., Westbrook, J. D., Feng, Z., Peisach, E., Persikova, I., Sala, R., Sen, S., Berrisford, J. M., Swaminathan, G. J., Oldfield, T. J., Gutmanas, A., Igarashi, R., Armstrong, D. R., Baskaran, K., Chen, L., Chen, M., Clark, A. R., Costanzo, L. D., Dimitropoulos, D., Gao, G., Ghosh, S., Gore, S., Guranovic, V., Hendrickx, P. M. S., Hudson, B. P., Ikegawa, Y., Kengaku, Y., Lawson, C. L., Liang, Y., Mak, L., Mukhopadhyay, A., Narayanan, B., Nishiyama, K., Patwardhan, A., Sahni, G., Sanz-García, E., Sato, J., Sekharan, M. R., Shao, C., Smart, O. S., Tan, L., van Ginkel, G., Yang, H., Zhuravleva, M. A., Markley, J. L., Nakamura, H., Kurisu, G., Kleywegt, G. J., Velankar, S., Berman, H. M. & Burley, S. K. (2018). *Database*, **2018**, bay002.
- Young, J. Y., Westbrook, J. D., Feng, Z., Sala, R., Peisach, E., Oldfield, T. J., Sen, S., Gutmanas, A., Armstrong, D. R., Berrisford, J. M., Chen, L., Chen, M., Di Costanzo, L., Dimitropoulos, D., Gao, G., Ghosh, S., Gore, S., Guranovic, V., Hendrickx, P. M., Hudson, B. P., Igarashi, R., Ikegawa, Y., Kobayashi, N., Lawson, C. L., Liang, Y., Mading, S., Mak, L., Mir, M. S., Mukhopadhyay, A., Patwardhan, A., Persikova, I., Rinaldi, L., Sanz-García, E., Sekharan, M. R., Shao, C., Swaminathan, G. J., Tan, L., Ulrich, E. L., van Ginkel, G., Yamashita, R., Yang, H., Zhuravleva, M. A., Quesada, M., Kleywegt, G. J., Berman, H. M., Markley, J. L., Nakamura, H., Velankar, S. & Burley, S. K. (2017). *Structure*, **25**, 536–545.