PDB-IHM: A System for Deposition, Curation, Validation, and Dissemination of Integrative Structures ☆

Brinda Vallat [1,2,*], Benjamin M. Webb [3], Arthur Zalevsky [3], Hongsuda Tangmunarunkit [4], Monica R. Sekharan [1], Serban Voinea [4], Aref Shafaeibejestan [4], Jared Sagendorf [3], Jeffrey C. Hoch [5], Genji Kurisu [6], Kyle L. Morris [7], Sameer Velankar [8], Carl Kesselman [4], Stephen K. Burley [1,2,9,10,11], Helen M. Berman [1,10,12], and Andrej Sali [3]

***1* - Research Collaboratory for Structural Bioinformatics Protein Data Bank and the Institute for Quantitative Biomedicine,** *Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA*

***2* - Rutgers Cancer Institute,** *Rutgers, The State University of New Jersey, New Brunswick, NJ 08901, USA*

***3* - Research Collaboratory for Structural Bioinformatics Protein Data Bank,** *Department of Bioengineering and Therapeutic Sciences, Quantitative Biosciences Institute (QBI), and Department of Pharmaceutical Chemistry, University of California, San Francisco., San Francisco, CA 94157, USA*

***4* - Information Sciences Institute,** *Viterbi School of Engineering, University of Southern California, Los Angeles, CA, USA*

***5* - Biological Magnetic Resonance Data Bank,** *Department of Molecular Biology and Biophysics, University of Connecticut, Farmington, CT 06030-3305, USA*

***6* - Protein Data Bank Japan,** *Institute for Protein Research, Osaka University, Suita, Osaka 565-0871, Japan*

***7* - Electron Microscopy Data Bank,** *European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD, UK*

***8* - Protein Data Bank in Europe,** *European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD, UK*

***9* - Research Collaboratory for Structural Bioinformatics Protein Data Bank,** *San Diego Supercomputer Center, University of California, La Jolla, CA 92093, USA*

***10* - Department of Chemistry and Chemical Biology,** *Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA*

***11* - Rutgers Artificial Intelligence and Data Science (RAD) Collaboratory,** *Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA*

***12* - Department of Quantitative and Computational Biology,** *University of Southern California, Los Angeles CA 90089, USA*

Correspondence to Brinda Vallat:Research Collaboratory for Structural Bioinformatics Protein Data Bank and the Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA.
brinda.vallat@rcsb.org (B. Vallat)

https://doi.org/10.1016/j.jmb.2025.168963
Editor: Michael Sternberg

Abstract

Structures of many large biomolecular assemblies are now being determined using integrative approaches. In these approaches, information derived from multiple experimental and computational methods is combined to compute three-dimensional structures of multi-protein complexes and other macromolecular machines. A standalone prototype data resource for integrative structures called PDB-Dev was built, based on recommendations of the Integrative and Hybrid Methods (IHM) Task Force of the Worldwide Protein Data Bank (wwPDB). This effort included developing data standards and

☆ This article is part of a special issue entitled: 'Computation Resources (2025)' published in Journal of Molecular Biology.

0022-2836/© 2025 The Author(s). Published by Elsevier Ltd.This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Please cite tts article as: B. Vallat, B.M. Webb, A. Zalevsky, et al., PDB-IHM: A System for Deposition, Curation, Validation, and Dissemination of Integrative StructuresJournal of Molecular Biology, https://doi.org/10.1016/j.jmb.2025.168963

software tools for collecting, curating, validating, visualizing, archiving, and disseminating integrative structures that span diverse spatiotemporal scales and conformational states. Mechanisms have been created to validate integrative structures based on the experimental data underpinning them. Building upon this foundational framework, PDB-Dev has been further expanded to handle large dynamic macromolecular systems and integrative structures that combine, for example, experimental restraints with atomic coordinates computed by machine learning algorithms. Data standards and supporting tools have also been extended to capture information about biomolecular dynamics, such as conformational transitions and related kinetic data derived from biophysical methods. Recently, PDB-Dev was unified with the PDB archive and rebranded as PDB-IHM (pdb-ihm.org), further promoting FAIR (Findable, Accessible, Interoperable, and Reusable) principles of data stewardship for integrative structural biology.

© 2025 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Introduction

The Protein Data Bank (PDB) is the single global repository for experimentally determined, three-dimensional (3D) structures of biomolecules and their complexes.[1–3] Established in 1971,[4] the PDB currently houses over 227,000 atomic-level structures of biomolecules, promoting the FAIR (Findability, Accessibility, Interoperability, Reusability) data principles[5] for structural biology. The PDB primarily archives biostructures determined using macromolecular X-ray crystallography (MX), nuclear magnetic resonance (NMR) spectroscopy, and three-dimensional electron microscopy (3DEM).

Integrative modeling is a powerful method for obtaining structures of macromolecular assemblies that elude traditional structure determination approaches.[6] It involves combining information from multiple sources, including data from experimental methods and information from computational algorithms. Experimental data from MX, NMR spectroscopy, 3DEM, small angle solution scattering (SAS), chemical crosslinking mass spectrometry (Crosslinking-MS), hydrogen–deuterium exchange mass spectrometry (HDX-MS), Förster resonance energy transfer (FRET), electron paramagnetic spectroscopy (EPR), etc. are gathered and converted into spatial restraints. These restraints are combined with starting structures of components obtained from experimental or computational methods to determine structures of macromolecular assemblies. Integrative modeling has been particularly empowered by the recent successes of machine learning (ML) techniques, in two ways. First, prediction methods, such as AlphaFold2[7] and RoseTTAFold,[8] can often provide computed structure models (CSMs) for assembly components, both individual subunits or subcomplexes. Second, ML techniques can also be used to implement the entire integrative modeling workflow by incorporating experimental information into the predictions.[9]

In practice, all experimental methods involve a trade-off between resolution and scale (spatial and temporal). Methods that can determine structures at atomic resolution are limited principally by the scale of the system, and methods that can be applied at arbitrary scales are limited by resolution. Integrative modeling attempts to bridge the gap between scale and resolution by maximizing the input information that can be used for modeling.[6] To identify opportunities and challenges in archiving and disseminating integrative structures, the Worldwide Protein Data Bank (wwPDB)[2] assembled an Integrative and Hybrid Methods (IHM) Task Force (wwpdb.org/task/hybrid). This expert Task Force published guidelines,[10,11] based on which a prototype standalone system for archiving integrative structures (PDB-Dev[12–14] was developed. PDB-Dev supports archiving and disseminating integrative structures determined using experimental data from single or multiple methods that are not fully supported by the PDB. PDB-Dev infrastructure consists of (i) the IHMCIF data standard[15]; (ii) software applications that support IHMCIF; (iii) workflows and tools for data harvesting, deposition, curation, and archiving; (iv) methods for integrative structure validation and visualization; and (v) a website for data dissemination (Figure 1).

Integrative modeling is quickly becoming a mainstream method in structural biology. Over time, PDB-Dev has gone through feature extensions and improvements. It now robustly and comprehensively supports deposition, curation, validation, archiving, and dissemination of integrative structures. In August 2024, PDB-Dev was unified with the PDB to deliver integrative structures alongside archived experimental structures.[16] Unification was made possible because of the shared data standards between PDBx/mmCIF and IHMCIF. With unification, integrative structures are assigned PDB accession codes, annotated as IHM structures, and can be downloaded from the PDB archive. Now part of the PDB infrastructure, PDB-Dev has been rebranded as PDB-IHM (pdb-ihm.org), denoting structures from integrative and hybrid methods (IHM) archived in PDB. Herein, we present recent

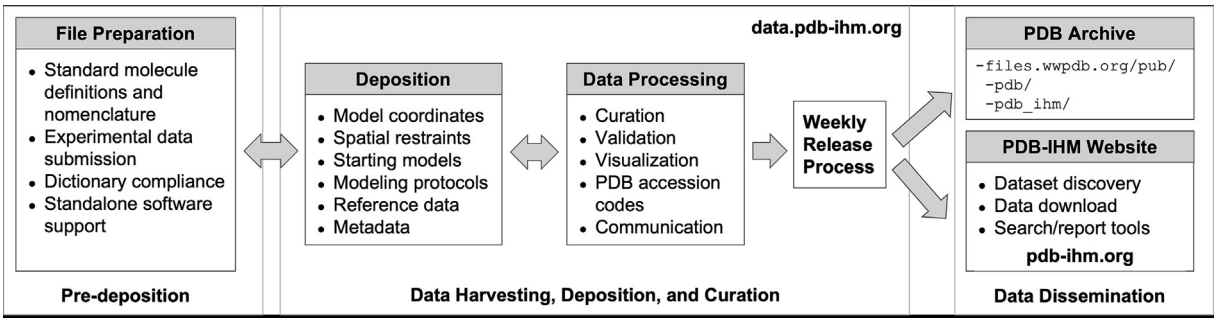

Figure 1. Components of the PDB-IHM data pipeline include methods for file preparation, data deposition, processing, release, and dissemination. Pre-deposition includes support from standalone software tools (e.g., python-ihm (github.com/ihmwg/python-ihm), MAXIT (sw-tools.rcsb.org/apps/MAXIT/)) and guidelines for preparing mmCIF files for submission. Data deposition handles submission of model coordinates, spatial restraints, and starting models along with related metadata. Data curation consists of checks for data consistency, completeness, and compliance with PDBx/mmCIF and IHMCIF dictionaries as well as verification of macromolecule and small molecule nomenclature. Additional processing steps include creation of complete mmCIF files from depositor provided data, generation of initial validation reports to help depositors identify and fix any errors, 3D visualization of structures, and issue of PDB accession codes. The file preparation, deposition, and curation processes are interrelated (shown by bi-directional arrows) and sometimes require active communication between curators and depositors to ensure that all required information is collected. Alternatively, depositors who use the deposition API or begin with a complete and compliant mmCIF file can skip many of the intermediate steps in the pipeline to expedite the issuing of an accession code. Finally, when the entry is set for release, it is added to the subsequent wwPDB weekly structure release process and delivered to the wwPDB archive portal and the PDB-IHM website.

developments in PDB-Dev (hereafter PDB-IHM) that enabled unification with PDB and provide continued support for evolving experimental and computational methods in integrative structural biology.

Results and Discussion

The PDB-IHM infrastructure consists of several software components that provide end-to-end support for collecting, curating, validating, visualizing, archiving, and disseminating integrative structures (Figure 1).

IHMCIF data dictionary

The data standard for PDB-IHM is captured in a dictionary of data definitions called IHMCIF.[15] IHMCIF is an extension of PDBx/mmCIF,[17,18] which serves as the data standard for the PDB archive. As an extension of PDBx/mmCIF, IHMCIF follows the same dictionary definition language[19] and reuses many core definitions from PDBx/mmCIF. Additional data and metadata descriptions specific to integrative modeling are captured in IHMCIF: (i) multi-scale models that simultaneously include atomic and/or coarse-grained models, (ii) multi-state models that only collectively satisfy the input data, (iii) ordered models that relate states across time, (iv) collections of models that individually satisfy the input data, (v) spatial restraints and starting structure models from various experimental and computational sources, (vi) conformational dynamics and kinetic information obtained from biophysical methods (e.g., FRET), (vii) modeling methodology, and (viii) provenance information and links to related experimental and reference data in trusted external repositories. IHMCIF is both human- and machine-readable and is extensible to support evolving methodologies. It is supported by a set of software tools that enable automated processing, reading, writing, and format validation. Unification with PDB was made possible by the seamless interoperability of IHMCIF and PDBx/mmCIF, in turn enabled by both dictionaries following the same fundamental data specifications.

System for data harvesting, deposition, and curation

Integrative structures are deposited to and processed by PDB-IHM through a system parallel to the wwPDB OneDep system[20] and are archived alongside experimental structures in the PDB archive. The PDB-IHM data harvesting, deposition, and curation system is hosted at data.pdb-ihm.org, which provides a web interface and workflows for depositors to assemble all the information required for archiving integrative structures in a format compliant with the PDBx/mmCIF and IHMCIF dictionaries. In addition, the system contains tools that support data consistency checks, data curation, generation of structure validation reports, 3D visualization, and public release of integrative structures to the PDB archive (Figure 1). In the following subsections, we describe the important features of the system, highlighting recent updates.

Implementation of the DERIVA platform. The PDB-IHM system has been built using the open source DERIVA scientific asset management platform.[21,22] The DERIVA ecosystem provides (i) a relational data catalog that can be used to describe the metadata associated with an entry, (ii) an object storage for managing user-uploaded and system generated files or assets, and (iii) a web interface to help depositors and curators navigate the system. Furthermore, the DERIVA platform assigns persistent unique identifiers and supports data versioning throughout the lifecycle of the digital object, thus allowing for efficiently capturing its history. DERIVA provides documented Representational State Transfer (REST) Application Programming Interfaces (APIs), high-level Python and R libraries, and command-line clients to programmatically access the stored data and metadata. The specific implementation of the DERIVA platform for PDB-IHM utilizes definitions in PDBx/mmCIF and IHMCIF to build a customized data catalog that supports the PDB-IHM data pipeline and uses the object storage to effectively track all files associated with an entry. The DERIVA-based PDB-IHM system can be deployed to a local on-premise private computing resource or on a public cloud infrastructure.

Improved workflows, curation processes, and access control policies. In addition to the DERIVA software stack, the PDB-IHM system is equipped with specialized workflows that have been implemented to carry out specific processes at different stages of the data pipeline. For example, after a depositor submits an entry, an automated workflow process collects all the information provided by the depositor, creates a complete mmCIF file, and checks the contents of the file to ensure that it is compliant with the PDBx/mmCIF and IHMCIF dictionaries. Similarly, when a curator triggers the release process for an entry, a complete mmCIF file with appropriate metadata about the release (e.g., release status and release date) is generated along with the validation reports. Each entry has a "Workflow_Status" property that is used to store the status of the entry at all times and communicate this information to all stakeholders (i.e., depositors, curators, and software agents). After specific milestones in the workflow, automated notifications are sent to curators with information about success or failure of a process along with any errors. The workflow processes rely on the PDBx/mmCIF and IHMCIF dictionaries and related software tools, such as python-ihm (github.com/ihmwg/python-ihm), py-mmcif (github.com/rcsb/py-mmcif), and the mmCIF dictionary suite (github.com/rcsb/cpp-dict-pack).

Beyond updates that increase the efficiency of workflows, PDB-IHM has been updated to allow curators to disable automated processing in favor of manual processing if required, thus giving them better control of the workflow. This feature is especially useful for handling entries from the legacy PDB-Dev deposition system that lack the full set of metadata required for processing as well as new types of integrative structures from evolving methods that are not immediately supported by the system. To improve performance of the system, we deployed the backend processing workflows as a separate service running from a different computing resource. In addition, we have incorporated fine-grained access control policies for depositors, curators, and system administrators that dictate who can create, modify, and view the metadata and assets related to an entry at different stages of the workflow until public release.

Support for PDB accession codes. As a prototype system, PDB-Dev provided accession codes with the "PDBDEV_" prefix for all deposited entries. To facilitate unification with the PDB archive, the wwPDB Archive Management team set aside a subset of 4-character PDB accession codes for integrative structures processed by PDB-IHM. We updated the PDB-IHM data catalog and workflows to handle PDB accession codes for new and existing entries. Integrative structures that are already released have been issued PDB accession codes, which are used as the primary identifier. Both PDB and PDB-Dev accession codes are supported for backward compatibility and the mmCIF files have been remediated to include both PDB and PDB-Dev accession codes. Updated mmCIF files and validation reports with PDB accession codes have been generated for all released entries. New depositions received after August 2024 are only issued PDB accession codes. In anticipation of the new format for the PDB accession codes (wwpdb.org/documentation/new-format-for-pdb-ids), PDB-IHM is also equipped to support both the current 4-character accession codes and the 12-character extended PDB accession codes to be adopted in the near future.

Deposition API and entry collections. Some integrative investigations may include a large collection of entries that share a common set of metadata (e.g., authors, publication, software) and modeling methodology. To enable depositing and processing such collections of entries, we modified PDB-IHM tools and workflows as follows: (i) IHMCIF and python-ihm were extended to support collection identifiers that can be assigned to entries in a collection; (ii) DERIVA data catalog was extended to allow for creation and update of collections and associated entries; (iii) a suite of DERIVA client tools were used to create a deposition API that facilitates bulk upload of dictionary-compliant data files; (iv) documentation

was created to help depositors with the bulk upload process using the deposition API; (v) mechanisms were created for curation, validation, and release of entries in a collection; and (vi) the PDB-IHM website was updated to support the new collection identifiers and facilitate search and retrieval of entries in a collection.

Structure visualization

3D visualization of integrative structures that follow the IHMCIF data standard is supported by the ChimeraX desktop viewer[23] and the Mol* web application.[24] Both software packages allow visualization of multi-scale and multi-state models in PDB-IHM. Furthermore, Mol* is available on the PDB-IHM website and has also been deployed within the PDB-IHM deposition system to aid deposition and biocuration.

Structure validation

Following the recommendations of the wwPDB IHM Task Force,[11] we organized information about entry, experimental and structural data used for structure generation, system representation, details of modeling protocol, and validation metrics into two key documents: a summary table and validation report. The summary table is a concise report with key information and quality metrics that can be used in supplementary information in publications. The validation report includes extended statistics and plots informative about the quality of the integrative structure and data on which it is based. Both data and models are assessed. The validation metrics are organized into 4 main categories: (i) data quality assessments, (ii) model quality assessments based on stereochemistry and physical principles, (iii) model fit to input data used for modeling, and (iv) model fit to data not used for modeling. The last category is currently under development. Data quality assessments and model fit to data used for modeling are dependent on the experimental method and input data used. Currently, these metrics are evaluated for models that are based on SAS data. They are described in the PDB-IHM validation documentation (pdb-ihm.org/about_validation.html). A similar implementation for models based on Crosslinking-MS is currently under testing and for models based on 3DEM is under development. Validation reports are available in human-readable HTML and PDF formats. The validation pipeline is fully integrated into the PDB-IHM deposition system, allowing depositors to review the validation report and, if necessary, update the entry before release. The validation software is also available for standalone use at github.com/salilab/IHMValidation.

Unification with PDB archive

In August 2024, PDB-Dev was unified with the PDB archive to serve integrative structures alongside experimental structures (and rebranded as PDB-IHM). Integrative structures and associated data can be accessed at and downloaded from files.wwpdb.org/pub/pdb_ihm/ (Figure 1). Currently, holdings files in JSON format, validation reports (summary and full reports) in PDF format, and model files in mmCIF format are provided. Holdings files include (i) current_file_holdings that lists all files associated with an entry including model files and validation reports, (ii) released_structures_last_modified_dates that provides the latest date of release/update for an entry, and (iii) unreleased_entries that lists all processed entries that are embargoed along with their deposition dates. The mmCIF files, validation reports, and holdings files use PDB accession codes and follow wwPDB conventions.

One of the prerequisites for unifying PDB-Dev with PDB was to establish a coordinated release process synchronized with the long-standing PDB weekly release pipeline. New releases in the PDB occur on Wednesdays at 00:00 Universal Time Coordinated (UTC). Entries to be released (or updated) each Wednesday are pre-packaged and staged for release on the previous Friday. Adopting the same release schedule, we created timed release workflows to identify new entries set for release/update by curators, create mmCIF files with proper release metadata, generate validation reports, create holdings files, package them together for archiving, and transfer them to the data exchange area set up by wwPDB on Fridays. Packaged PDB-IHM data are then taken up by the PDB archive weekly update process and released publicly on Wednesdays at 00:00 UTC by the wwPDB. In addition, digital object identifiers (DOI) are assigned to all released and processed entries in PDB-IHM based on their PDB accession codes. In the future, the wwPDB partners, including Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB, RCSB.org) in the United States, Protein Data Bank in Europe (PDBe, PDBe.org), and Protein Data Bank Japan (PDBj, PDBj.org), will disseminate integrative structures on their respective web portals. Electron Microscopy Data Bank (EMDB, ebi.ac.uk/emdb/) and Biological Magnetic Resonance Data Bank (BMRB, bmrb.io) will provide links to integrative structures based on 3DEM maps and NMR spectroscopy data respectively.

To support robust record-keeping for the weekly update process, we have created mechanisms to systematically keep track of weekly new releases, updated re-releases, entries on hold, mmCIF files and validation reports associated with new and updated entries, holdings files generated each week, and entry-level history of updates over time. Furthermore, built-in DERIVA versioning tools help with recreating the files and metadata from previous weekly releases if required.

PDB-IHM data dissemination

In addition to disseminating integrative structures, validation reports, and holdings files through the "pdb_ihm" branch of the PDB archive, the PDB-IHM weekly release process also makes these data publicly available on the PDB-IHM website (pdb-ihm.org). The website was previously created to support the PDB-Dev infrastructure. It has been rebranded as PDB-IHM following unification with PDB. It provides mechanisms to search and retrieve archived data, view individual entry pages with structure highlights and validation metrics, access links to download mmCIF files and validation reports, and use the Mol* graphics viewer to visualize integrative structures. Integrative structures can also be downloaded in BinaryCIF format,[25] which is a serialization of the mmCIF format and provides improved parsing performance and compression. Moreover, all software components in the web application (e.g., backend database, frontend web pages, file download links, and search service) have been updated to support the new PDB accession codes (and existing PDB-Dev accession codes, if available). Entry pages also include the PDB DOIs that link to the wwPDB website (e.g., wwpdb.org/pdb?id=pdb_00009a8n).

Structures in PDB-IHM

PDB-IHM was created to archive macromolecular structures computed based on data from diverse sets of experimental and computational methods and/or structures that use new representations, such as multiple scales, multiple states, and ordered models. Figure 2 highlights examples of such structures now archived in PDB-IHM, including (i) structure of a fragment of plant secondary cell wall composed of complex branched polymers obtained from solid-state NMR data illustrating inter-molecular interactions and higher order architecture of the cell wall (9A3U),[26] (ii) multi-state structure of a GTPase, human Guanylate binding protein 1, determined using experimentally-obtained spatial restraints from FRET, SAS, and EPR that illustrates different conformers involved in oligomerization and reveals important mechanistic and kinetic information regarding conformational transition between the states (9A1G),[27] (iii) structure of Cullin4-RING ubiquitin ligase (CRL4) complex generated by combining AI protein structure prediction methods and *in vivo* Crosslinking-MS data using AlphaLink (9A40),[9] (iv) structure of human LINE-1 ORF2p protein based on 2D electron microscopy images (2DEM), 3DEM, Crosslinking-MS, AI structure predictions, and molecular dynamics simulations that revealed heterogeneous collection of states attributed to different stages of the LINE-1 integration cycle (9A3Q),[28] and (v) a model of the postmitotic assembly pathway of the human Nuclear Pore Complex obtained using restraints from time-dependent 3DEM and fluorescence correlation spectroscopy (FCS), revealing structures of assembly intermediates (9A25).[29] These structures utilize the full arsenal of new features of PDB-IHM.

PDB-IHM structures include those generated using input information derived from more than fifteen different types of experiments (Figure 3a) and a variety of modeling software (Figure 3b). Among the experimental methods, Crosslinking-MS, 3DEM, NMR, SAS, FRET, and HDX-MS are the most common and AlphaLink,[9] *Integrative Modeling Platform* (IMP),[30] HADDOCK,[31] and ROSETTA[32] are the most frequently used modeling software. Structures obtained with Crosslinking-MS, 3DEM, and/or SAS constitute ~94% of all entries, with Crosslinking-MS being the highest contributor. The sizable contribution of Crosslinking-MS can be explained by its high-throughput nature as well as its applicability to purified *in vitro* and *heterogeneous in situ* samples. Furthermore, many integrative structures that use 3DEM maps in combination with complementary methods, such as Crosslinking-MS, are currently

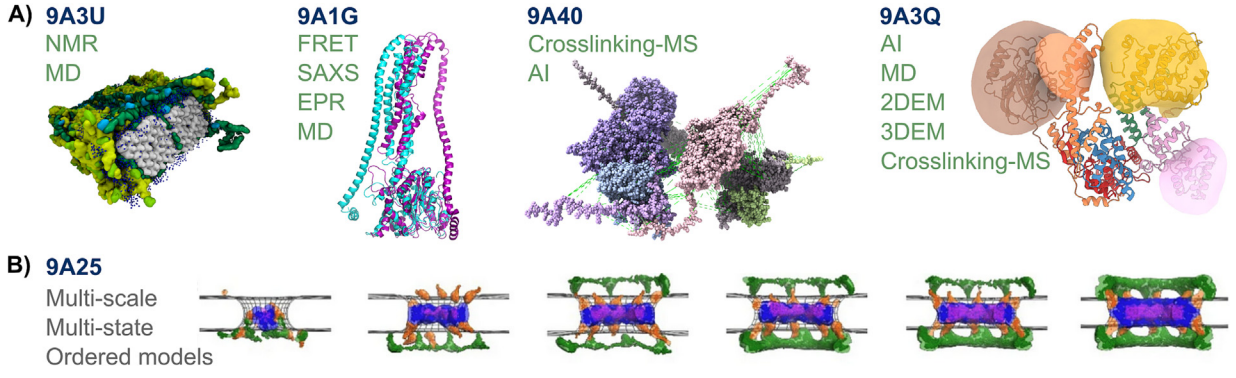

Figure 2. Selected PDB-IHM structures that illustrate the breadth of PDB-IHM capabilities. (A) Examples show different experimental methods used as sources of input information and methods used for modeling. 9A3U: Plant cell wall fragment; 9A1G: human Guanylate binding protein 1; 9A40: Cullin4-RING ubiquitin ligase; 9A3Q: LINE-1 ORF2p protein. (B) Assembly pathway of human Nuclear Pore Complex (9A25).

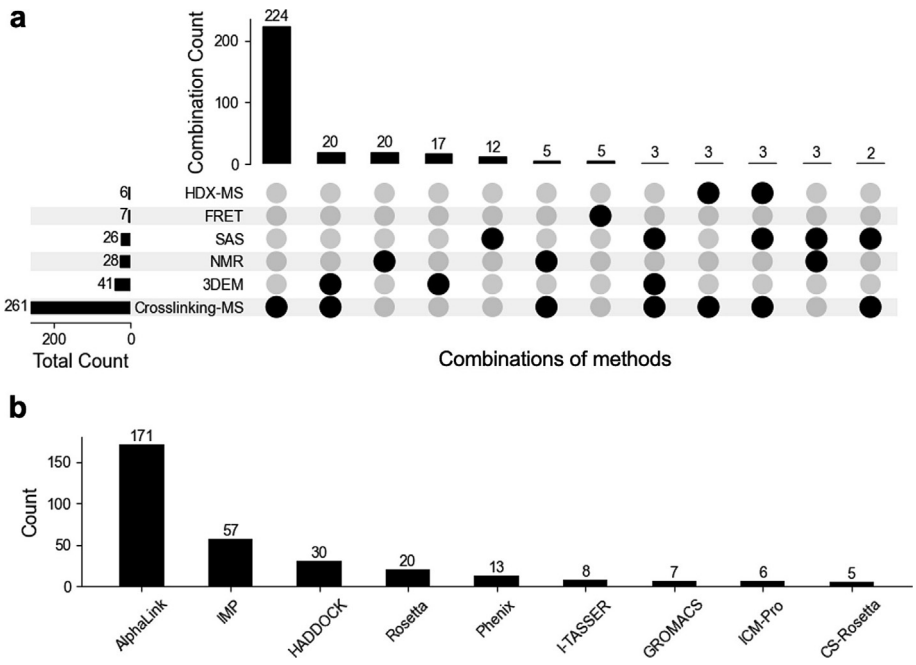

Figure 3. (a) Common experimental methods used for obtaining data for integrative structure determination in the PDB-IHM entries. The counts shown are the occurrences of various kinds of input information derived from experimental methods as reported in the PDB-IHM entries. Rows represent different experimental methods, and the counts correspond to the total number of entries that use a given source of input information (not a sum across the column counts shown). Columns represent particular combinations of input information, and the counts are the total number of times those combinations occur in PDB-IHM entries. Combinations that only occur once are not shown. Black circles indicate the methods included in each combination. **(b)** Common modeling software used for computing integrative structures in PDB-IHM. The counts indicate the number of entries that report using the given software in their modeling process.

deposited in the PDB as 3DEM structures without any information about other data used in structure determination, resulting in incomplete structure annotations and validations and loss of information. Following the unification of PDB-Dev with PDB, we aim for all new integrative structures to be deposited with the full set of information required for archiving. In addition, we will identify extant structures with likely missing annotations by processing the original publications and incorporating these annotations retroactively.

Conclusions

The PDB-IHM (formerly PDB-Dev) ecosystem provides a robust infrastructure for collecting, curating, validating, visualizing, archiving, and disseminating integrative structures of biomolecules and their complexes, thus addressing the recommendations made by the wwPDB IHM Task Force in 2015.[10] The recent unification of PDB-IHM with the PDB enables delivery of integrative structures alongside experimental atomic structures in the PDB archive and marks a significant milestone in the evolution of the PDB. To prevent data loss, depositors of integrative structures are encouraged to use PDB-IHM to allow deposition of all data involved in integrative modeling investigations (e.g., model coordinates, spatial restraints, starting models, modeling protocols, references to experimental data in other repositories). A complete deposition has at least four advantages. First, it makes the data more FAIR. Second, a complete deposition enables rigorous validation of the resulting structure based on the input data used. Third, it facilitates informed user interpretation of the structure. Finally, a large set of complete entries will facilitate the development of better integrative methods in the future.

Looking forward, we anticipate continued development of integrative modeling methods, including those relying on machine learning algorithms. We also expect that new types of experimental data and theoretical information will be used for computing integrative structures. For example, we project an increase in cryo-electron tomography-derived depositions due to the development of new automated data collection and modeling techniques.[33] Similarly, we foresee a surge in application of Crosslinking-MS and other proteomics methods for structural characterization of protein–protein interactions at the cellular level.[34] Furthermore, emerging capabilities in NMR relaxation methods enable detection of transient

conformational states in proteins, which can be represented as multi-state integrative models.[35] Finally, to facilitate modeling based on new types of data and addressing new biological questions based on the resulting models, we also anticipate a concomitant increase in the diversity of the types of integrative structures. For instance, the scope of integrative modeling will continue to expand beyond static structures of biomolecular assemblies to provide insights into biomolecular interactions, dynamics, and function at increasingly higher levels of the hierarchy of biological organization, including genomes, organelles, and cells.[36,37,6] The PDB-IHM infrastructure is designed to be extensible and is positioned to support these advances in integrative structural biology.

CRediT authorship contribution statement

Brinda Vallat: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Benjamin M. Webb:** Writing – review & editing, Software, Methodology. **Arthur Zalevsky:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation. **Hongsuda Tangmunarunkit:** Writing – review & editing, Visualization, Validation, Software, Methodology. **Monica R. Sekharan:** Writing – review & editing, Visualization, Validation, Data curation. **Serban Voinea:** Software. **Aref Shafaeibejestan:** Software. **Jared Sagendorf:** Writing – review & editing, Visualization. **Jeffrey C. Hoch:** Writing – review & editing. **Genji Kurisu:** Writing – review & editing. **Kyle L. Morris:** Writing – review & editing. **Sameer Velankar:** Writing – review & editing. **Carl Kesselman:** Writing – review & editing, Supervision, Funding acquisition. **Stephen K. Burley:** Writing – review & editing, Supervision, Funding acquisition. **Helen M. Berman:** Writing – review & editing, Supervision, Funding acquisition. **Andrej Sali:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

Funding

B. Vallat acknowledges funding from the United States National Science Foundation (NSF) awards DBI-2112966 (PI: B. Vallat) and DBI-1756248 (PI: B. Vallat). A. Sali acknowledges funding from NSF and the United States National Institutes of Health (NIH) (NSF DBI-2112967, PI: A. Sali; NSF DBI-1756250, PI: A. Sali; NIH R01GM083960, PI: A. Sali; NIH P41GM109824, PI: M.P. Rout). C. Kesselman acknowledges funding from NSF (DBI-2112968). RCSB PDB core operations are jointly funded by 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). BMRB is supported by the National Institute of General Medical Sciences of the NIH (R24GM150793, PI: J.C. Hoch).

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors thank RCSB PDB team members Jasmine Young, Ezra Peisach, Zukang Feng, Vladimir Guranovic, Henry Chao, Jeremy Henry, Aditya Pingale, Catherine Lawson, Dennis Piehl, and James Smith for enabling unification of PDB-Dev with PDB. We thank Jian Yu from PDBj and all the wwPDB OneDep team members for their support. We are grateful to members of the wwPDB IHM Task Force for their continued support and recommendations. Finally, we thank the more than 60,000 researchers worldwide who have deposited structures to PDB and PDB-IHM, and the many millions of public 3D biostructure data consumers working and learning in nearly every country and territory.

Received 26 November 2024;
Accepted 22 January 2025;
Available online xxxx

Keywords:
PDB;
PDBx/mmCIF;
integrative modeling;
IHMCIF;
structure validation

References

1. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., et al., (2000). The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242.
2. Berman, H.M., Henrick, K., Nakamura, H., (2003). Announcing the worldwide Protein Data Bank. *Nature Struct. Biol.* **10**, 980.
3. wwPDB consortium, (2019). Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* **47**, D520–D528.
4. Protein Data Bank, (1971). Crystallography: Protein Data Bank. *Nature New Biol.* **233**, 223.

5. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., et al., (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 1–9.
6. Sali, A., (2021). From integrative structural biology to cell biology. *J. Biol. Chem.* **296**, 100743
7. Jumper, J., Evans, R., Pritzel, A., Green, T., et al., (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589.
8. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., et al., (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876.
9. Stahl, K., Warneke, R., Demann, L., Bremenkamp, R., et al., (2024). Modelling protein complexes with crosslinking mass spectrometry and deep learning. *Nature Commun.* **15**, 7866.
10. Sali, A., Berman, H.M., Schwede, T., Trewhella, J., et al., (2015). Outcome of the first wwPDB hybrid/integrative methods task force workshop. *Structure* **23**, 1156–1167.
11. Berman, H.M., Adams, P.D., Bonvin, A.A., Burley, S.K., et al., (2019). Federating structural models and data: outcomes from a workshop on archiving integrative structures. *Structure* **27**, 1745–1759.
12. Vallat, B., Webb, B., Fayazi, M., Voinea, S., et al., (2021). New system for archiving integrative structures. *Acta Crystallogr. Sect. D Struct. Biol.* **77**, 1486–1496.
13. Vallat, B., Webb, B., Westbrook, J.D., Sali, A., et al., (2018). Development of a prototype system for archiving integrative/hybrid structure models of biological macromolecules. *Structure* **26**, 894–904.e892.
14. Burley, S.K., Kurisu, G., Markley, J.L., Nakamura, H., et al., (2017). PDB-Dev: a prototype system for depositing integrative/hybrid structural models. *Structure* **25**, 1317–1318.
15. Vallat, B., Webb, B.M., Westbrook, J.D., Goddard, T.D., et al., (2024). IHMCIF: an extension of the PDBx/mmCIF data standard for integrative structure determination methods. *J. Mol. Biol.* **436**, 168546
16. Vallat B, Young J, Feng Z, Peisach E, et al. (unpublished results). Centralizing access to integrative structures alongside experimental structures in the Protein Data Bank archive.
17. Fitzgerald, P.M.D., Westbrook, J.D., Bourne, P.E., McMahon, B., et al., (2005). 4.5 Macromolecular dictionary (mmCIF). In: Hall, S.R., McMahon, B. (Eds.), *International tables for crystallography G definition and exchange of crystallographic data*. Springer, Dordrecht, The Netherlands, pp. 295–443.
18. Westbrook, J.D., Young, J.Y., Shao, C., Feng, Z., et al., (2022). PDBx/mmCIF ecosystem: foundational semantic tools for structural biology. *J. Mol. Biol.* **434**, 167599
19. Westbrook, J.D., Berman, H.M., Hall, S.R., (2005). 2.6 Specification of a relational Dictionary Definition Language (DDL2). In: Hall, S.R., McMahon, B. (Eds.), *International tables for crystallography*. Springer, Dordrecht, The Netherlands, pp. 61–72.
20. Young, J.Y., Westbrook, J.D., Feng, Z., Sala, R., et al., (2017). OneDep: unified wwPDB system for deposition, biocuration, and validation of macromolecular structures in the PDB archive. *Structure* **25**, 536–545.
21. Schuler, R.E., Kesselman, C., Czajkowski, K., (2016). Accelerating data-driven discovery with scientific asset management. *Second. Accelerat. Data-Driven Discov. Sci. Asset Manage.*, 31–40. https://doi.org/10.1109/eScience.2016.7870883.
22. Bugacov, A., Czajkowski, K., Kesselman, C., Kumar, A., et al., (2017). Experiences with deriva: an asset management platform for accelerating. *In: Proc IEEE Int Conf Escience*, pp. 79–88.
23. Meng, E.C., Goddard, T.D., Pettersen, E.F., Couch, G.S., et al., (2023). UCSF ChimeraX: tools for structure building and analysis. *Protein Sci.* **32**, e4792.
24. Sehnal, D., Bittrich, S., Deshpande, M., Svobodova, R., et al., (2021). Mol* viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Res.* **49**, W431–W437.
25. Sehnal, D., Bittrich, S., Velankar, S., Koca, J., et al., (2020). BinaryCIF and CIFTools-Lightweight, efficient and extensible macromolecular data management. *PLoS Comput. Biol.* **16**, e1008247
26. Addison, B., Bu, L., Bharadwaj, V., Crowley, M.F., et al., (2024). Atomistic, macromolecular model of the Populus secondary cell wall informed by solid-state NMR. *Sci. Adv.* **10**, eadi7965
27. Peulen, T.O., Hengstenberg, C.S., Biehl, R., Dimura, M., et al., (2023). Integrative dynamic structural biology unveils conformers essential for the oligomerization of a large GTPase. *eLife* **12**
28. Baldwin, E.T., van Eeuwen, T., Hoyos, D., Zalevsky, A., et al., (2024). Structures, functions and adaptations of the human LINE-1 ORF2 protein. *Nature* **626**, 194–206.
29. Otsuka, S., Tempkin, J.O.B., Zhang, W., Politi, A.Z., et al., (2023). A quantitative map of nuclear pore assembly reveals two distinct mechanisms. *Nature* **613**, 575–581.
30. Russel, D., Lasker, K., Webb, B., Velazquez-Muriel, J., et al., (2012). Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol.* **10**, e1001244
31. Dominguez, C., Boelens, R., Bonvin, A.M., (2003). HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* **125**, 1731–1737.
32. Leaver-Fay, A., Tyka, M., Lewis, S.M., Lange, O.F., et al., (2011). ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* **487**, 545–574.
33. Klumpe, S., Fung, H.K., Goetz, S.K., Zagoriy, I., et al., (2021). A modular platform for automated cryo-FIB workflows. *eLife* **10**
34. Graziadei, A., Rappsilber, J., (2022). Leveraging crosslinking mass spectrometry in structural and cell biology. *Structure* **30**, 37–54.
35. Fenwick, R.B., Oyen, D., van den Bedem, H., Dyson, H.J., et al., (2021). Modeling of hidden structures using sparse chemical shift data from NMR relaxation dispersion. *Biophys. J.* **120**, 296–305.
36. Singla, J., McClary, K.M., White, K.L., Alber, F., et al., (2018). Opportunities and challenges in building a spatiotemporal multi-scale model of the human pancreatic beta cell. *Cell* **173**, 11–19.
37. Hua, N., Tjong, H., Shin, H., Gong, K., et al., (2018). Producing genome structure populations with the dynamic and automated PGS software. *Nature Protoc.* **13**, 915–926.