IHMCIF: An Extension of the PDBx/mmCIF Data Standard for Integrative Structure Determination Methods

Brinda Vallat[1,2], Benjamin M. Webb[3], John D. Westbrook[1,2,+], Thomas D. Goddard[4], Christian A. Hanke[5], Andrea Graziadei[6,7], Ezra Peisach[1], Arthur Zalevsky[3], Jared Sagendorf[3], Hongsuda Tangmunarunkit[8], Serban Voinea[8], Monica Sekharan[1], Jian Yu[9], Alexander A.M.J.J. Bonvin[10], Frank DiMaio[11], Gerhard Hummer[12], Jens Meiler[13], Emad Tajkhorshid[14], Thomas E. Ferrin[4], Catherine L. Lawson[1], Alexander Leitner[15], Juri Rappsilber[6,16], Claus A.M. Seidel[5], Cy M. Jeffries[17], Stephen K. Burley[1,2,18,19], Jeffrey C. Hoch[20], Genji Kurisu[9], Kyle Morris[21], Ardan Patwardhan[21], Sameer Velankar[22], Torsten Schwede[23], Jill Trewhella[24], Carl Kesselman[8], Helen M. Berman[1,19,25], 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]Cancer Institute of New Jersey, Rutgers, The State University of New Jersey, New Brunswick, NJ 08901, USA
[3]Department of Bioengineering and Therapeutic Sciences, Department of Pharmaceutical Chemistry, the Quantitative Biosciences Institute (QBI), and the Research Collaboratory for Structural Bioinformatics Protein Data Bank, University of California, San Francisco, San Francisco, CA 94157, USA
[4]Department of Pharmaceutical Chemistry, University of California, San Francisco, CA 94158, USA
[5]Molecular Physical Chemistry, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany
[6]Bioanalytics, Institute of Biotechnology, Technische Universität Berlin, 10623 Berlin, Germany
[7]Human Technopole, 20157 Milan, Italy
[8]Information Sciences Institute, Viterbi School of Engineering, University of Southern California, Los Angeles, California, USA
[9]Protein Data Bank Japan, Institute for Protein Research, Osaka University, Suita, Osaka 565-0871, Japan
[10]Bijvoet Centre for Biomolecular Research, Faculty of Science - Chemistry, Utrecht University, Padualaan 8, 3584 CH Utrecht, the Netherlands
[11]Department of Biochemistry and Institute for Protein Design, University of Washington, Seattle, WA 98195, USA
[12]Department of Theoretical Biophysics, Max Planck Institute of Biophysics, 60438 Frankfurt am Main, Germany; Institute for Biophysics, Goethe University Frankfurt, 60438 Frankfurt am Main, Germany

[13]Center for Structural Biology, Vanderbilt University, 465 21st Avenue South, Nashville, TN 37221, USA; Institute for Drug Discovery, Leipzig University Medical School, 04103 Leipzig, Germany
[14]NIH Resource for Macromolecular Modeling and Visualization, Beckman Institute for Advanced Science and Technology, Department of Biochemistry, and Center for Biophysics and Quantitative Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
[15]Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, 8093 Zurich, Switzerland
[16]Wellcome Centre for Cell Biology, University of Edinburgh, Max Born Crescent, Edinburgh EH9 3BF, UK
[17]European Molecular Biology Laboratory (EMBL), Hamburg Unit, c/o Deutsches Elektronen-Synchrotron (DESY), Notkestrasse 85, 22607 Hamburg, Germany
[18]Research Collaboratory for Structural Bioinformatics Protein Data Bank, San Diego Supercomputer Center, University of California, La Jolla, CA 92093, USA
[19]Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA
[20]Biological Magnetic Resonance Data Bank, Department of Molecular Biology and Biophysics, University of Connecticut, Farmington, CT 06030-3305, USA.
[21]Electron Microscopy Data Bank, European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, UK
[22]Protein Data Bank in Europe, European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD, UK
[23]Biozentrum, University of Basel, Basel, Switzerland; Computational Structural Biology & SIB Swiss Institute of Bioinformatics, Basel, Switzerland
[24]School of Life and Environmental Sciences, The University of Sydney, Sydney, NSW 2006, Australia; Department of Chemistry, University of Utah, Salt Lake City, UT 84112, USA.
[25]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles CA 90089, USA

[+]Deceased

*Contact authors: brinda.vallat@rcsb.org; ihm-repval@salilab.org

ABSTRACT

IHMCIF (github.com/ihmwg/IHMCIF) is a data information framework that supports archiving and disseminating macromolecular structures determined by integrative or hybrid modeling (IHM), and making them Findable, Accessible, Interoperable, and Reusable (*FAIR*). IHMCIF is an extension of the Protein Data Bank Exchange/macromolecular Crystallographic Information Framework (PDBx/mmCIF) that serves as the framework for the Protein Data Bank (PDB) to archive experimentally determined atomic structures of biological macromolecules and their complexes with one another and small molecule ligands (e.g., enzyme cofactors and drugs). IHMCIF serves as the foundational data standard for the PDB-Dev prototype system, developed for archiving and disseminating integrative structures. It utilizes a flexible data representation to describe integrative structures that span multiple spatiotemporal scales and structural states with definitions for

restraints from a variety of experimental methods contributing to integrative structural biology. The IHMCIF extension was created with the benefit of considerable community input and recommendations gathered by the Worldwide Protein Data Bank (wwPDB) Task Force for Integrative or Hybrid Methods (wwpdb.org/task/hybrid). Herein, we describe the development of IHMCIF to support evolving methodologies and ongoing advancements in integrative structural biology. Ultimately, IHMCIF will facilitate the unification of PDB-Dev data and tools with the PDB archive so that integrative structures can be archived and disseminated through PDB.

Keywords: IHMCIF, PDBx/mmCIF, Data Standard, Open Access, Worldwide Protein Data Bank, wwPDB, Integrative Modeling, PDB-Dev

INTRODUCTION

Introduction to integrative modeling

Increasingly, structures of many complex biological systems are determined using integrative approaches that combine information from multiple experimental and computational methods [1, 2]. Such approaches are typically used for determining structures of complex macromolecular assemblies that cannot be solved using any one of the traditional methods, including macromolecular crystallography (MX), Nuclear Magnetic Resonance (NMR) spectroscopy, and three-dimensional electron microscopy (3DEM). Integrative modeling generally combines data from these traditional methods with information from complementary biophysical and proteomics methods, such as small angle scattering (SAS), chemical crosslinking mass spectrometry (crosslinking-MS), Förster resonance energy transfer (FRET) spectroscopy, electron paramagnetic resonance (EPR) spectroscopy, hydrogen-deuterium exchange mass spectrometry (HDX-MS), and atomic force microscopy (AFM) obtained from *in vitro*, *in situ* or even *in vivo* samples. In addition, experimental data can be combined with other information, such as structures of molecular components determined by experimental and computational methods as well as other types of bioinformatics analyses (e.g., predictions of binding sites and co-evolving residues); in particular, integrative modeling of large and/or dynamic biomolecular systems benefits from models of system components computed by emerging deep learning methods [3, 4]. The input information gathered is converted into an integrative model by: (i) defining molecular representation of the modeled system, (ii) constructing spatial restraints on the components, (iii) finding a model that satisfies these restraints by structural sampling, and (iv) validating the model (Figure 1).

Integrative modeling has been applied to determine structures of macromolecular systems that participate in major cellular processes, such as replication, transcription, translation, regulation of gene expression, protein degradation, mitosis, muscle contraction, signal transduction, cellular communication, and immune response [1]. These structures greatly enhance our understanding of biological processes and pathways, regulatory interactions, antibody epitopes, and disease etiology. Therefore, efforts to make the results of integrative structure determinations publicly available are critical for advancing biological and biomedical research.

wwPDB IHM Task Force and Working Groups

Protein Data Bank (PDB) is the single global repository for atomic structures of macromolecules and their complexes determined using MX, NMR, and 3DEM [5, 6]. The archive is managed by the Worldwide Protein Data Bank (wwPDB) organization [7] that ensures open access to the structural data according to the *FAIR* (Findable, Accessible, Interoperable, and Reusable) principles [8]. Recognizing the growing application of integrative methods in structural biology, the wwPDB established an Integrative and Hybrid Methods (IHM) Task Force (hereafter Task Force, wwpdb.org/task/hybrid) to address and overcome challenges involved in archiving and disseminating integrative structures. The Task Force included members from different experimental method communities, as well as structural biologists, modelers, and data scientists. The inaugural workshop of the Task Force was held in 2014 at EMBL-EBI in Hinxton, UK, resulting in a whitepaper describing a series of recommendations for archiving integrative structures and associated experimental data and metadata [9]. In addition, two Working Groups were set up to address ongoing requirements for (a) developing model representation and validation methods for integrative structures and (b) creating a federated network of interoperating data resources contributing to integrative structural biology. A second meeting was organized by the Working Groups in Baltimore, Maryland, during the 2019 Biophysical Society annual meeting, resulting in another whitepaper [10] describing additional recommendations for developing data standards and methods for collecting, curating, validating, and disseminating integrative structures as well as recommendations for establishing mechanisms for interoperation among different experimental data and structural model repositories to build a federated network of resources. These workshops have fostered collaborative efforts across different scientific disciplines to create benchmarks, data standards, and other means of promoting open science and *FAIR* data practices [11-17]. We continue to work with a number of scientific communities contributing data for integrative structure determination, aiming to coordinate efforts of various data providers to develop data standards, supporting tools, and necessary infrastructure for archiving and disseminating data in the *FAIR* manner.

Significance of data standards and history of PDBx/mmCIF

Data standards are technical descriptions of data and metadata definitions, along with format specifications for encoding the data and metadata. They are the primary requirement for collecting, archiving, and disseminating data in a standard format, and ensuring that the data follow the *FAIR* principles. Scientific data standards provide definitions for representing the results of an investigation and additional metadata, such as authors, citations, samples, methods, software, etc. Using consistent, standard mechanisms to store this information enables better interoperation among resources and facilitates data search, retrieval, and reuse.

The legacy PDB format developed in the 1970s is one of the earliest archival formats in structural biology [18]. Due to its simplicity and popularity, the PDB format remained the standard archival format for PDB for over forty years. However, it posed serious limitations for archiving structures of large biomolecular assemblies due to its rigid requirements of fixed column positions and widths and limited metadata definitions. As structural biology evolved, a more general and flexible system for defining data standards was required to support larger structures and new experimental methods.

The Crystallographic Information Framework (CIF) was developed as the data and publication standard of the International Union of Crystallography (IUCr) for diffraction experiments on small molecules [19]. Subsequently, the macromolecular CIF (mmCIF) data representation was created to describe the structures of macromolecules and the results of MX structure determinations [20]. mmCIF takes into account the hierarchical representation of polymeric macromolecules and the relationship between sequence and three-dimensional (3D) structure. Over time, the original mmCIF data standard was extended by the wwPDB to create PDBx/mmCIF (Protein Data Bank Exchange/macromolecular Crystallographic Information Framework) [21, 22], which added support for archiving of structures determined using NMR and 3DEM experiments. It was officially adopted as the master format and archiving data standard for PDB in 2014. The underlying framework that supports PDBx/mmCIF [23] includes metadata definitions used for assessing and maintaining data consistency, such as primary data types (e.g., integers, real numbers, and text), controlled vocabularies, boundary conditions, and parent-child relationships among data items. Support for parent-child relationships within PDBx/mmCIF, which are necessary for archiving macromolecular structure data, represented a significant advance over the original CIF standard. PDBx/mmCIF was designed to be fully extensible and has been extended, for example, to represent small-angle solution scattering data [24, 25] and computed structure models [26]. In addition, a suite of software tools is available to support the PDBx/mmCIF format and its extensions (mmcif.wwpdb.org/docs/software-resources.html).

Development of the PDB-Dev Prototype System and IHMCIF

Following recommendations of the wwPDB IHM Task Force, a prototype system called PDB-Dev was developed to archive and disseminate integrative structures and associated experimental data (pdb-dev.wwpdb.org) [27-29]. The PDB-Dev infrastructure consists of a deposition and data harvesting system, methods for data processing and curation, mechanisms for validation of experimental data and structures, tools for visualization of integrative structures, and a website for data distribution that supports search and retrieval, data access, dataset discovery, and download. The primary requirement for developing PDB-Dev was the development of data standards to represent the data and metadata involved in integrative structure modeling. The PDBx/mmCIF data representation was, therefore, extended to create the IHMCIF data standard [28]. IHMCIF incorporated community recommendations from the wwPDB IHM Task Force and contains specific definitions and attributes required for describing and archiving the results of integrative structure determination.

IHMCIF is developed and maintained as an open-source project (github.com/ihmwg/IHMCIF) by the Working Group on model representation and validation. This Working Group promotes adoption of IHMCIF in the integrative structural biology community, deposition of integrative structures to PDB-Dev, and development of software tools to support IHMCIF, such as the python-ihm library (github.com/ihmwg/python-ihm). The GitHub repository provides access to the IHMCIF extension dictionary as well as the consolidated dictionary, where IHMCIF is merged with the parent PDBx/mmCIF dictionary.

RESULTS AND DISCUSSION

Data definitions from PDBx/mmCIF

As an extension of PDBx/mmCIF, IHMCIF reuses many core definitions from PDBx/mmCIF (mmcif.wwpdb.org), including representation of polymeric macromolecules, small-molecule ligands, biomolecular complexes, and their atomic coordinates, as well as related metadata definitions pertaining to modeling software used, bibliographic citations, author names, and references for macromolecular sequences and small molecule nomenclature (Figure 2). These shared definitions facilitate interoperation of integrative structures and those determined experimentally using MX, NMR, and 3DEM.

IHMCIF data definitions

The IHMCIF extension was implemented based on recommendations made by the wwPDB IHM Task Force and representative integrative structures provided by Working Group members. IHMCIF extends PDBx/mmCIF definitions to address various requirements for archiving integrative structures (Figure 2).

(i) To accommodate the needs of integrative structural biology studies, IHMCIF allows for a flexible model representation that supports the following four features [9]:

First, a model can be multi-scale. Multi-scaling supports representing a model as a collection of particles at different resolutions corresponding to atoms, single or multi-residue spherical beads, and 3D Gaussian objects. For example, a protein complex can be simultaneously described as a low-resolution volume representation of protein subunits as well as a well-resolved atomic representation of individual residues. Multi-scale representation allows for optimally encoding the model such that spatial restraints from input data can be accurately applied while retaining sufficient information to make the resulting models useful for further research.

Second, a model can be multi-state. A set of multiple states can be used to describe a system that exists in a mixture of multiple structural and/or compositional states that collectively satisfy the input information. For example, a sample of enzyme molecules in solution is structurally heterogeneous when it exists in an equilibrium between open and closed states; it is compositionally heterogeneous when it contains enzyme molecules both with and without a ligand.

Third, the states in a multi-state model can be ordered in the form of a graph. This graph can be used to represent a model of a process such as an enzymatic reaction, a biochemical pathway, or a molecular dynamics trajectory.

Finally, IHMCIF also allows for specifying a collection of models, where each one is consistent with given input information within an acceptable threshold. The variability among the models in the collection helps in assessing the uncertainty of modeling and the completeness of input data.

(ii) IHMCIF captures the many different kinds of spatial restraints used for integrative modeling, including restraints derived from crosslinking-MS, HDX-MS, FRET spectroscopy, SAS, EPR spectroscopy, DNA footprinting, mutagenesis, and other biophysical techniques. To enable capture of a broad range of generic distance restraints (e.g., those from mutagenesis, DNA footprinting, and coevolution analysis), IHMCIF includes a general representation of distance restraints between features at various resolution scales (e.g., between individual atoms, single or multiple amino acid residues, and contiguous residue ranges) and the corresponding uncertainties. These definitions can be further extended to describe dihedral and orientational restraints at different granularities if specific requirements arise.

(iii) IHMCIF includes definitions for the starting structural models of assembly components that are frequently used in integrative modeling. Starting models are mapped to molecular entities and their corresponding segments in the integrative structure, if applicable. Additionally, origins and provenance of starting models are specified, and existing structural templates and alignments used in building starting comparative models are defined. Representation of the spatial restraints and starting models enables validation of integrative structures based on all available information, including data used in the modeling and data reserved specifically for validation. Definitions are included to support preliminary model validation data, such as fit of models to input restraints (e.g., satisfied and violated crosslink restraints) and information regarding the precision and structural diversity of sampled models in each collection (e.g., localization densities [30]).

(iv) IHMCIF provides generic definitions for referencing related data from external resources via stable identifiers, such as accession codes or persistent digital object identifiers (DOIs) for data that do not have an established information repository. This approach facilitates inclusion of external annotations and provenance information regarding diverse sources of data and models used in integrative modeling, which is required for submission to PDB-Dev and is obtained during deposition. References to experimental data repositories such as BioMagResBank (BMRB [31]), Electron Microscopy Data Bank (EMDB [32]), Small Angle Scattering Biological Data Bank (SASBDB [33, 34]), and ProteomeXchange consortium resources [35], and 3D structural model repositories (PDB, ModelArchive (www.modelarchive.org), and AlphaFoldDB [36]) are supported. Dictionary support can be easily added for any new resources in the future.

(v) IHMCIF provides simplified definitions for describing the modeling workflow. It also includes mechanisms for linking modeling scripts and software program files, which are intended to promote reproducibility of modeling studies.

IHMCIF definitions are maintained and extended in an ongoing manner to support the evolving needs of integrative structural biology experiments. As spatial restraints from emerging methods are used in integrative modeling studies and innovative modeling algorithms are developed, new dictionary definitions are added to represent expanded data and metadata information. For example, recently IHMCIF was extended to describe (a) conformational dynamics and kinetic information of macromolecules obtained from FRET spectroscopy [37] or other biophysical methods, and (b) metadata regarding sets of entries belonging to an “investigation” or reported in a scholarly publication. The latter was implemented to archive the collection of structures resulting from the development of AlphaLink software [38], wherein machine learning algorithms are

combined with experimental restraints from crosslinking-MS to create new integrative modeling applications. Because the PDB-Dev infrastructure is built atop IHMCIF, the new definitions created in IHMCIF are automatically propagated to the tools supporting PDB-Dev, including the deposition and data harvesting system, curation and validation pipeline, and the search and data access services on the PDB-Dev website, to ensure comprehensive end-to-end support for the new definitions.

Software tools supporting IHMCIF

IHMCIF is supported by the open-source python-ihm software library (github.com/ihmwg/python-ihm), which enables reading, writing, and managing data files compliant with the IHMCIF dictionary [39]. Python-ihm represents an integrative model as a set of interrelated Python objects. It also provides mechanisms for converting these objects to or from IHMCIF or BinaryCIF [40] formats. Support for BinaryCIF provides improved parsing performance and efficient compression of IHMCIF files. Furthermore, python-ihm was designed to allow other developers to easily add support for IHMCIF in their software without needing to be fully aware of the underlying data model and the relationships between data items. For example, the *Integrative Modeling Platform* (IMP; [41]) and HADDOCK [42] modeling software packages currently use python-ihm to generate IHMCIF files for deposition to PDB-Dev. ChimeraX [43] uses this same library to visualize integrative structures archived in PDB-Dev. The python-ihm library can also be used standalone. Workflows in the PDB-Dev system for deposition, biocuration, and validation report generation use python-ihm to read and write IHMCIF files and validate the files against the IHMCIF dictionary.

In addition to the above tools, the Mol* [44] web application supports visualization of integrative structures described using IHMCIF. Other modeling and visualization applications such as ROSETTA [45], *Bayesian Inference of ENsembles* (BioEn [46]), *BioChemical Library* (BCL [47]), *FRET Positioning and Screening* (FPS [48]), and *Visual Molecular Dynamics* (VMD [49]) are in the process of adding support for IHMCIF.

Advantages of IHMCIF

IHMCIF serves as the foundational data standard for archiving integrative structures. In addition, IHMCIF enables creation of automated mechanisms for data collection, processing, validation, and open access dissemination of integrative structures. As an extension of PDBx/mmCIF, IHMCIF provides a number of advantages. First, existing definitions in PDBx/mmCIF for representing the atomic structures of polymeric macromolecules, small-molecules, and macromolecular assemblies are reused. Second, software tools developed to support PDBx/mmCIF have been extended to support IHMCIF; for example, IHMCIF files can be validated against the dictionary and converted to BinaryCIF files using software applications developed for PDBx/mmCIF (e.g., github.com/rcsb/py-mmcif and sw-tools.rcsb.org/apps/MMCIF-DICT-SUITE/). Third, IHMCIF can be readily extended to support ongoing and future methodological developments. Finally, IHMCIF enables interoperation with other structural biology data resources (e.g., PDB, ModelArchive, AlphaFoldDB, and SASBDB).

CONCLUSION AND PERSPECTIVES

Development of IHMCIF enabled creation of the PDB-Dev prototype system for archiving and disseminating integrative structures, thereby promoting *FAIR* data principles, and providing free and open access to the results of integrative structure determinations. PDB-Dev was implemented separately from PDB to facilitate agile development, with the eventual goal of unifying PDB-Dev with PDB. Work is currently in progress to integrate the structures and tools in PDB-Dev with PDB. As a result, integrative structures can be collected, curated, validated, archived, and disseminated through PDB. This unification is made possible by the IHMCIF extension and will expand the capabilities of the PDB to support emerging structural biology methods and archive spatiotemporal and dynamic biostructures spanning diverse scales. As structural biology expands its scope from macromolecular machines to entire cells [2, 50] and beyond, the application of integrative modeling to address future challenges will be essential. Analysis of recent depositions in both PDB and PDB-Dev revealed increasing use of 3DEM in combination with complementary methods such as crosslinking-MS in integrative modeling studies. Furthermore, the much-heralded successes of machine learning algorithms, such as AlphaFold2 [3] and RoseTTAFold [4], in predicting the structures of proteins from amino acid sequence alone provide an enormous pool of starting component models for integrative modeling studies of larger systems across size scales ranging from macromolecular assemblies to whole cells. Integration of experimental technologies with machine learning-driven structure prediction approaches will lead to novel integrative modeling methods that will shape structural biology discovery in the next decade.

ACKNOWLEDGEMENTS

The authors thank all members of the wwPDB IHM Task Force and Working Groups for their continued support and recommendations. We thank all the researchers worldwide who have deposited structures to PDB and PDB-Dev. We also gratefully acknowledge contributions to the PDBx/mmCIF data standard made by past members of the Worldwide Protein Data Bank partner organizations (Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB), Protein Data Bank in Europe (PDBe), Protein Data Bank Japan (PDBj), EMDB, and BMRB) and members of the structural biology community. We thank the developers of Molstar for providing support for visualizing integrative structures.

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). H. Berman acknowledges funding from NSF (DBI-1519158). 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-1832184, 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 (R01GM133198, PI: S.K. Burley). Other funding awards to RCSB PDB by NSF and to PDBe by the UK Biotechnology and Biological

Research Council are jointly supporting development of a Next Generation PDB archive (DBI-2019297, PI: S.K. Burley; BB/V004247/1, PI: Sameer Velankar) and new Mol* features (DBI-2129634, PI: S.K. Burley; BB/W017970/1, PI: Sameer Velankar). G. Hummer acknowledges support by the Max Planck Society. E. Tajkhorshid acknowledges funding from NIH (P41-GM104601, PI: Tajkhorshid; R24-GM145965, PI: Tajkhorshid). PDBj is supported by grants from the Database Integration Coordination Program from the department of NBDC program, Japan Science and Technology Agency (JPMJND2205, PI: G. Kurisu), and partially supported by Platform Project for Supporting Drug Discovery and Life Science Research (Basis for Supporting Innovative Drug Discovery and Life Science Research (BINDS)) from AMED under Grant Number 22ama121001. PDBe is supported by European Molecular Biology Laboratory-European Bioinformatics Institute. T. Schwede acknowledges funding from SIB Swiss Institute of Bioinformatics. J. Hoch acknowledges funding from NIH (R24GM150793) for BMRB. J. Meiler is supported by a Humboldt Professorship of the Alexander von Humboldt Foundation. J. Meiler acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG) through SFB1423 (421152132), SFB 1052 (209933838), and SPP 2363 (460865652). J. Meiler is supported by BMBF (Federal Ministry of Education and Research) through the Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI) and through DAAD project 57616814 (SECAI, School of Embedded Composite AI). Work in the Meiler laboratory is further supported through the NIH (R01 HL122010, R01 DA046138, U01 AI150739, R01CA227833, S10 OD016216, S10 OD020154, S10 OD032234). The Wellcome Centre for Cell Biology is supported by core funding from the Wellcome Trust (203149). T. Ferrin and T. Goddard acknowledge support from the NIH (R01GM129325, PI: T.E. Ferrin). A.M.J.J. Bonvin acknowledges support from the European Union Horizon 2020, projects BioExcel (675728, 823830 and 101093290) and EGI-ACE (101017567), from the Netherlands e-Science Center (027.020.G13) and from the Dutch Foundation for Scientific Research (NWO) (TOP-PUNT grant 718.015.001). C.A.M Seidel acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) CRC 1208 (ID 267205415, project A08) and SE 1195/17-1 as well as the European Research Council through the Advanced Grant 2014 hybridFRET (671208). EMDB is supported by funding from the Wellcome Trust [212977/Z/18/Z]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

FIGURE LEGENDS

Figure 1. Typical integrative modeling workflow. (A). First, all available information from experiments, prior experimental or computational models, physical theories, and/or statistical preferences, is gathered. A sample of such information is shown here. (B). Secondly, a suitable representation for the modeled system is chosen and the information gathered is translated into spatial restraints on the system. Some component representations may be coarse-grained by using spherical beads corresponding to multiple amino acid residues to reflect the lack of information and/or to increase efficiency of structural sampling. Four example representations and restraints are shown here corresponding to the information gathered in panel A. (C). The structure of the system is sampled to find those models that satisfy the spatial restraints as well as possible. The goal is to find a collection of representative models, each one of which satisfies the input data within acceptable thresholds. (D). The sampling is then assessed for convergence

and models are evaluated by the degree to which they satisfy the input information used to construct them, as well as omitted information. Iterations through this workflow may be used until the models are judged to be satisfactory, most often on the basis of their precision and the degree to which they satisfy the data.

Figure 2. Schematic representation of the data specifications in IHMCIF. Definitions reused from PDBx/mmCIF are identified using labels on white background (e.g., Polymeric Macromolecules and Atomic Coordinates) and the newly added definitions are identified using labels on gray background (e.g., Experimental Datasets and Localization Density). (A) Several data categories are added to describe the inputs used in integrative modeling, including datasets from a wide range of experimental methods and starting structural models, which can be experimentally determined or are the results of prior modeling. Sources of experimental datasets and starting models used are also captured. (B) Representations of molecular components and complexes are retained from PDBx/mmCIF. (C) Definitions for atomic coordinates are taken from PDBx/mmCIF. In addition, a model can be represented in a multi-scale fashion; it can describe more than one compositionally and/or structurally heterogeneous state; states can be ordered; and an entry can consist of a collection of representative models. (D) Definitions regarding how well the models fit the input data (e.g., crosslink restraints satisfied and violated) and the variability of models in a collection (e.g., localization densities) are included. (E) Several metadata definitions from PDBx/mmCIF are reused. New metadata definitions regarding modeling protocols, input or output files, as well as datasets accessible via DOIs or database accessions are added to IHMCIF.

REFERENCES

[1] Rout MP, Sali A. (2019). Principles for Integrative Structural Biology Studies. Cell. 177, 1384-1403.
[2] Sali A. (2021). From integrative structural biology to cell biology. J Biol Chem. 296, 100743.
[3] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature. 596, 583-589.
[4] Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. Science. 373, 871-876.
[5] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. (2000). The Protein Data Bank. Nucleic Acids Res. 28, 235-242.
[6] wwPDB consortium. (2019). Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res. 47, D520-D528.
[7] Berman HM, Henrick K, Nakamura H. (2003). Announcing the worldwide Protein Data Bank. Nature Structure Biology. 10, 980.
[8] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 3, 1-9.
[9] Sali A, Berman HM, Schwede T, Trewhella J, Kleywegt G, Burley SK, et al. (2015). Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop. Structure. 23, 1156-1167.
[10] Berman HM, Adams PD, Bonvin AA, Burley SK, Carragher B, Chiu W, et al. (2019). Federating Structural Models and Data: Outcomes from A Workshop on Archiving Integrative Structures. Structure. 27, 1745-1759.

[11] Leitner A, Bonvin A, Borchers CH, Chalkley RJ, Chamot-Rooke J, Combe CW, et al. (2020). Toward Increased Reliability, Transparency, and Accessibility in Cross-linking Mass Spectrometry. Structure. 28, 1259-1268.
[12] Masson GR, Burke JE, Ahn NG, Anand GS, Borchers C, Brier S, et al. (2019). Recommendations for performing, interpreting and reporting hydrogen deuterium exchange mass spectrometry (HDX-MS) experiments. Nature methods. 16, 595-602.
[13] Schiemann O, Heubach CA, Abdullin D, Ackermann K, Azarkh M, Bagryanskaya EG, et al. (2021). Benchmark Test and Guidelines for DEER/PELDOR Experiments on Nitroxide-Labeled Biomolecules. J Am Chem Soc. 143, 17875-17890.
[14] Lerner E, Barth A, Hendrix J, Ambrose B, Birkedal V, Blanchard SC, et al. (2021). FRET-based dynamic structural biology: Challenges, perspectives and an appeal for open-science practices. eLife. 10.
[15] Trewhella J, Vachette P, Bierma J, Blanchet C, Brookes E, Chakravarthy S, et al. (2022). A round-robin approach provides a detailed assessment of biomolecular small-angle scattering data reproducibility and yields consensus curves for benchmarking. Acta crystallographica Section D, Structural biology. 78, 1315-1336.
[16] Trewhella J, Duff AP, Durand D, Gabel F, Guss JM, Hendrickson WA, et al. (2017). 2017 publication guidelines for structural modelling of small-angle scattering data from biomolecules in solution: an update. Acta crystallographica Section D, Structural biology. 73, 710-728.
[17] Trewhella J, Jeffries CM, Whitten AE. (2023). 2023 update of template tables for reporting biomolecular structural modelling of small-angle scattering data. Acta crystallographica Section D, Structural biology. 79, 122-132.
[18] Westbrook JD, Fitzgerald PMD. (2009). Chapter 10 The PDB format, mmCIF formats, and other data formats. In Structural Bioinformatics, Second Edition, (Bourne PE, Gu J, eds), p. 271-291, John Wiley & Sons, Inc., Hoboken, NJ.
[19] Hall SR, Allen FH, Brown ID. (1991). The crystallographic information file (CIF): a new standard archive file for crystallography. Acta Crystallographica Section A Foundations of Crystallography. 47, 655-685.
[20] Fitzgerald PMD, Westbrook JD, Bourne PE, McMahon B, Watenpaugh KD, Berman HM. (2005). 4.5 Macromolecular dictionary (mmCIF). In International Tables for Crystallography G Definition and exchange of crystallographic data, (Hall SR, McMahon B, eds), p. 295-443, Springer, Dordrecht, The Netherlands.
[21] Westbrook J, Henrick K, Ulrich EL, Berman HM. (2005). 3.6.2 The Protein Data Bank exchange data dictionary. In International Tables for Crystallography, (Hall SR, McMahon B, eds), p. 195-198, Springer, Dordrecht, The Netherlands.
[22] Westbrook JD, Young JY, Shao C, Feng Z, Guranovic V, Lawson C, et al. (2022). PDBx/mmCIF Ecosystem: Foundational semantic tools for structural biology. J Mol Biol. 434, 167599.
[23] Westbrook JD, Berman HM, Hall SR. (2005). 2.6 Specification of a relational Dictionary Definition Language (DDL2). In International Tables for Crystallography, (Hall SR, McMahon B, eds), p. 61-72, Springer, Dordrecht, The Netherlands.
[24] Malfois M, Svergun DI. (2000). sasCIF: an extension of core Crystallographic Information File for SAS. Journal of Applied Crystallography. 33, 812-816.
[25] Kachala M, Westbrook J, Svergun D. (2016). Extension of the sasCIF format and its applications for data processing and deposition. J Appl Crystallogr. 49, 302-310.
[26] Vallat B, Tauriello G, Bienert S, Haas J, Webb BM, Zidek A, et al. (2023). ModelCIF: An Extension of PDBx/mmCIF Data Representation for Computed Structure Models. J Mol Biol. 168021.
[27] Vallat B, Webb B, Fayazi M, Voinea S, Tangmunarunkit H, Ganesan SJ, et al. (2021). New system for archiving integrative structures. Acta crystallographica Section D, Structural biology. 77, 1486-1496.

[28] Vallat B, Webb B, Westbrook JD, Sali A, Berman HM. (2018). Development of a Prototype System for Archiving Integrative/Hybrid Structure Models of Biological Macromolecules. Structure. 26, 894-904 e892.
[29] Burley SK, Kurisu G, Markley JL, Nakamura H, Velankar S, Berman HM, et al. (2017). PDB-Dev: a Prototype System for Depositing Integrative/Hybrid Structural Models. Structure. 25, 1317-1318.
[30] Shi Y, Fernandez-Martinez J, Tjioe E, Pellarin R, Kim SJ, Williams R, et al. (2014). Structural characterization by cross-linking reveals the detailed architecture of a coatomer-related heptameric module from the nuclear pore complex. Mol Cell Proteomics. 13, 2927-2943.
[31] Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, et al. (2008). BioMagResBank. Nucleic Acids Res. 36, D402-408.
[32] Tagari M, Newman R, Chagoyen M, Carazo JM, Henrick K. (2002). New electron microscopy database and deposition system. Trends Biochem Sci. 27, 589.
[33] Valentini E, Kikhney AG, Previtali G, Jeffries CM, Svergun DI. (2015). SASBDB, a repository for biological small-angle scattering data. Nucleic Acids Res. 43, D357-363.
[34] Kikhney AG, Borges CR, Molodenskiy DS, Jeffries CM, Svergun DI. (2020). SASBDB: Towards an automatically curated and validated repository for biological scattering data. Protein Sci. 29, 66-75.
[35] Deutsch EW, Csordas A, Sun Z, Jarnuczak A, Perez-Riverol Y, Ternent T, et al. (2017). The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. Nucleic Acids Res. 45, D1100-D1106.
[36] Varadi M, Bertoni D, Magana P, Paramval U, Pidruchna I, Radhakrishnan M, et al. (2024). AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. Nucleic Acids Res. 52, D368-D375.
[37] Peulen TO, Hengstenberg CS, Biehl R, Dimura M, Lorenz C, Valeri A, et al. (2023). Integrative dynamic structural biology unveils conformers essential for the oligomerization of a large GTPase. eLife. 12.
[38] Stahl K, Graziadei A, Dau T, Brock O, Rappsilber J. (2023). Protein structure prediction with in-cell photo-crosslinking mass spectrometry and deep learning. Nat Biotechnol.
[39] Vallat B, Webb B, Westbrook J, Sali A, Berman HM. (2019). Archiving and disseminating integrative structure models. J Biomol NMR. 73, 385-398.
[40] Sehnal D, Bittrich S, Velankar S, Koca J, Svobodova R, Burley SK, et al. (2020). BinaryCIF and CIFTools-Lightweight, efficient and extensible macromolecular data management. PLoS Comput Biol. 16, e1008247.
[41] Russel D, Lasker K, Webb B, Velazquez-Muriel J, Tjioe E, Schneidman-Duhovny D, et al. (2012). Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. PLoS Biol. 10, e1001244.
[42] Dominguez C, Boelens R, Bonvin AM. (2003). HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc. 125, 1731-1737.
[43] Meng EC, Goddard TD, Pettersen EF, Couch GS, Pearson ZJ, Morris JH, et al. (2023). UCSF ChimeraX: Tools for structure building and analysis. Protein Sci. 32, e4792.
[44] Sehnal D, Bittrich S, Deshpande M, Svobodova R, Berka K, Bazgier V, et al. (2021). Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. Nucleic Acids Res. 49, W431–W437.
[45] Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, et al. (2011). ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. 487, 545-574.
[46] Hummer G, Kofinger J. (2015). Bayesian ensemble refinement by replica simulations and reweighting. J Chem Phys. 143, 243150.

[47] Karakas M, Woetzel N, Staritzbichler R, Alexander N, Weiner BE, Meiler J. (2012). BCL::Fold--de novo prediction of complex and large protein topologies by assembly of secondary structure elements. PLoS One. 7, e49240.
[48] Kalinin S, Peulen T, Sindbert S, Rothwell PJ, Berger S, Restle T, et al. (2012). A toolkit and benchmark study for FRET-restrained high-precision structural modeling. Nature methods. 9, 1218-1225.
[49] Humphrey W, Dalke A, Schulten K. (1996). VMD: visual molecular dynamics. J Mol Graph. 14, 33-38.
[50] Singla J, McClary KM, White KL, Alber F, Sali A, Stevens RC. (2018). Opportunities and Challenges in Building a Spatiotemporal Multi-scale Model of the Human Pancreatic beta Cell. Cell. 173, 11-19.