Federating Structural Models and Data: Outcomes from A Workshop on Archiving Integrative Structures

Helen M. Berman,1,2,3,* Paul D. Adams,4,5 Alexandre A. Bonvin,6 Stephen K. Burley,7,8,9,10 Bridget Carragher,11,12 Wah Chiu,13,14 Frank DiMaio,15 Thomas E. Ferrin,16 Margaret J. Gabanyi,8 Thomas D. Goddard,16 Patrick R. Griffin,17 Juergen Haas,18 Christian A. Hanke,19 Jeffrey C. Hoch,20 Gerhard Hummer,21,22 Genji Kurisu,23 Catherine L. Lawson,8 Alexander Leitner,24 John L. Markley,25 Jens Meiler,26 Gaetano T. Montelione,27,28,29 George N. Phillips, Jr.,30 (Author list continued on next page)

Structures of biomolecular systems are increasingly computed by integrative modeling. In this approach, a structural model is constructed by combining information from multiple sources, including varied experimental methods and prior models. In 2019, a Workshop was held as a Biophysical Society Satellite Meeting to assess progress and discuss further requirements for archiving integrative structures. The primary goal of the Workshop was to build consensus for addressing the challenges involved in creating common data standards, building methods for federated data exchange, and developing mechanisms for validating integrative structures. The summary of the Workshop and the recommendations that emerged are presented here.

Introduction

When the Protein Data Bank (PDB) (Protein Data Bank, 1971) was first established in 1971, X-ray crystallography was the only method for determining three-dimensional structures of biological macromolecules at sufficient resolution to build atomic models. A decade later, structures of biomolecules in solution could also be determined by nuclear magnetic resonance (NMR) spectroscopy (Williamson et al., 1985). Recently, three-dimensional cryo-electron microscopy (3DEM) (Henderson et al., 1990) began to achieve unprecedented near-atomic resolution for large complex assemblies. Increasingly, investigators are also modeling structures based on data from more than one method (Rout and Sali, 2019). These integrative/hybrid approaches to structure determination consist of collecting information about a system using multiple experimental and computational methods, followed by integrative/hybrid modeling that converts this information into integrative/hybrid structure models. For succinctness, we will use the term integrative hereafter to refer to integrative/hybrid approaches, modeling, and models.

The PDB has established a data-processing pipeline for depositing, validating, archiving, and disseminating structures determined by single methods, and to a limited extent structures based on data from two different experimental methods. Examples of the latter include structures derived from a combination of X-ray crystallography plus neutron diffraction data, NMR, or X-ray crystallography plus small-angle scattering (SAS) data. However, the processing of structures produced by integrating data from many different methods and/or those depicted by non-atomic, coarse-grained representations poses a greater challenge. Given the importance of integrative structures for advancing biological sciences and the significant investment made to determine them, the Worldwide Protein Data Bank (wwPDB) (Berman et al., 2003) initiated an effort to address the key challenges in enhancing its data-processing pipeline to accommodate integrative structures.

In 2014, the wwPDB convened an Integrative/Hybrid Methods (IHM) Task Force and sponsored a Workshop held at the European Bioinformatics Institute (EBI). The purpose of the Workshop was to engage a community of experts to make
recommendations for how to responsibly archive integrative structures. The five recommendations formulated by the Workshop participants were:

1. In addition to archiving the models themselves, all relevant experimental data and metadata as well as experimental and computational protocols should be archived; inclusivity is key.
2. A flexible model representation needs to be developed, allowing for multi-scale models, multi-state models, ensembles of models, and models related by time or other order.
3. Procedures for estimating the uncertainty of integrative models should be developed, validated, and adopted.
4. A federated system of model and data archives should be created.
5. Publication standards for integrative models should be established.

A white paper was published (Sali et al., 2015) and two working groups were established; the Federation Working Group was to address the issues of data federation (Figure 1) and the Model Working Group was tasked with helping set up the framework for model representation, validation, and visualization.

Over the last 5 years, steady progress has been made in implementing the IHM Task Force recommendations. Members of the Federation and Model Working Groups have met periodically in person and via video conferencing. One key challenge has been to develop common data standards for describing the multiple experimental and computational methods used to produce integrative structures. Thus, the PDB exchange/Macromolecular Crystallographic Information File (PDBx/mmCIF) dictionary (Fitzgerald et al., 2005; Westbrook, 2013) for describing structures has been extended to include the terms necessary for representing and archiving integrative structures (Valiati et al., 2018). Software support for these dictionary extensions has been developed, including software tools for visualizing integrative structures (Goddard et al., 2018) and a prototype archiving system called PDB-Dev (pdb-dev.wwpdb.org) (Burley et al., 2017; Vallat et al., 2018, 2019). Mechanisms that facilitate data exchange (e.g., transfer of restraints from an experimental

https://doi.org/10.1016/j.str.2019.11.002
At the center are the three structural biology model repositories: the PDB, ModelArchive of in silico structural models (www.modelarchive.org), and the PDB-Dev prototype system for archiving integrative structures (Burley et al., 2017; Vallat et al., 2018). The outer circle indicates experimental data that contribute to integrative structural biology. Existing data exchange mechanisms for X-ray, NMR, 3DEM, and SAS data are represented by black arrows. Ongoing and future projects aim to develop methods for data exchange with archives for other types of experimental data as well as among the existing structural model repositories (gray arrows).

**Progress on Archiving Integrative Structures**

**Archiving Requirements**

The requirements for archiving integrative structures include: (1) creating standard definitions for the experimental data used for structure determination and the structural features of the models; (2) developing methods for curation and validation of models and data; and (3) building the infrastructure for acquiring, archiving, and disseminating the models and the data. Because integrative structures are based on data derived from multiple experimental methods, the wwPDB IHM Task Force came up with the concept of Federated Resources, wherein structural models and experimental data could be seamlessly exchanged. Within the Federation model, expert communities are responsible for the creation of data standards in their respective areas. Experts in multiple domains contribute to multiple resources and provide coordination on common data standards among resources. The development of well-aligned data standards and efficient methods for data exchange among the different repositories as well as software platforms are key prerequisites for an effective Federation. An integrated federated system will provide a mechanism for archiving the experimental data, structural models, and associated metadata, such as citations, software, authors, workflows, sample, and data and model quality metrics. Furthermore, the availability of experimental data used for building structural models will facilitate the development of methods for building and validating integrative structures.

**Molecular Representation of Integrative Structures**

One of the fundamental requirements for all operations involving integrative structures, including computing, archiving, validating, visualizing, disseminating, and analyzing, is the creation of standards for representing these models. Integrative structures are often computed for large conformationally and compositionally heterogeneous systems, based on relatively sparse and potentially low-resolution datasets. Thus, a molecular representation of ensembles of multi-scale and/or multi-state structures is required. The first version of the prototype archiving system for integrative structures (Vallat et al., 2018) adopted the molecular representation developed as part of the open-source Integrative Modeling Platform (IMP) program (Russel et al., 2012).

**An Extensible Standard Dictionary of Terms**

During the 2000s, the wwPDB transitioned from using the PDB format (Callaway et al., 1996) to the mmCIF data representation (Fitzgerald et al., 2005; Westbrook, 2013) for archiving structural models. The PDBx/mmCIF standard provides a rich framework for defining macromolecular components, small-molecule ligands, polymeric sequences, and atomic coordinates. The PDBx/mmCIF data representation was extended by adding terms to accommodate the expanded molecular representation for integrative structures (see previous section, Molecular Representation of Integrative Structures) and the many experimental and computational methods used to determine them. These additional definitions are maintained as an extension dictionary called the IHM Dictionary (Vallat et al., 2018). The organization of the extension dictionary capturing these additional data definitions is depicted in Figure 2. Descriptions of starting structural models of the components used in integrative modeling of assemblies are also included, along with definitions of spatial restraints derived from multiple methods, including chemical...
crosslinking mass spectrometry (CX-MS), two-dimensional electron microscopy (2DEM), 3DEM, SAS, Förster resonance energy transfer (FRET), and electron paramagnetic resonance (EPR) spectroscopy. Generic methods for describing modeling workflows and for referencing data residing in external resources are also provided.

A software library called python-ihm (github.com/ihmwg/python-ihm) has been built to support reading, writing, and managing data files compliant with the IHM Dictionary. The library can be used as a stand-alone package or as part of an integrative modeling package. The IMP modeling program (Russel et al., 2012) and the ChimeraX visualization software (Goddard et al., 2018) already use the python-ihm library to support the IHM Dictionary.

**PDB-Dev: A Prototype Archiving System for Integrative Structures**

A prototype archiving system called PDB-Dev (Vallat et al., 2018) supporting integrative modeling was announced in 2017 (Burley et al., 2017). PDB-Dev (pdb-dev.wwpdb.org) currently contains ~35 structures and is growing rapidly. The structures in PDB-Dev range from small- and medium-size complexes (such as human Rev7 dimer [Rizzo et al., 2018], diubiquitin complex [Liu et al., 2018], 16S rRNA complexed with methyltransferase A [van Zundert et al., 2015], and human mitochondrial iron sulfur cluster core complex [Cai et al., 2018]), to large complexes (such as the yeast nuclear pore complex [Kim et al., 2018] and the RNF168-RING domain nucleosome complex [Horn et al., 2019]). The structures were determined based on data from experimental methods such as CX-MS, 2DEM, 3DEM, NMR, SAS, FRET, EPR, and other proteomics and biophysical techniques. Various modeling programs, such as IMP (Russel et al., 2012), HADDOCK (Dominguez et al., 2003; van Zundert et al., 2016), Rosetta (Leaver-Fay et al., 2011), XPLOR-NIH (Schwieters et al., 2018), TADbit (Trussart et al., 2015; Serra et al., 2017), iSPOT (Huang et al., 2016; Hsieh et al., 2017), FPS (Dimura et al., 2016), PatchDock (Schneidman-Duhovny et al., 2005), and BioEn (Hummer and Kofinger, 2015), have been used in building these structures.

**A Pipeline for Deposition, Curation, Validation, Visualization, and Dissemination**

Work is in progress to expand the PDB-Dev system into a pipeline that can handle deposition, curation, validation, and dissemination of integrative structures and associated data. A key objective is to integrate this PDB-Dev prototype system into the wwPDB OneDep system (Young et al., 2017) (Figure 3) and the integrative structures into the PDB archive.

**Resources for Computing and Visualizing Integrative Structures**

A variety of resources and approaches for integrative modeling exist (Table 2 in Rout and Sali, 2019), including programs developed specifically for integrative modeling and scripts that exploit programs originally developed for other types of modeling. Several modeling programs used to compute integrative structures deposited in PDB-Dev and software tools used to visualize these structures are outlined in the following subsections.

**Integrative Modeling Platform**

IMP is an open-source software package that provides programmatic support for implementing and distributing integrative modeling protocols (Russel et al., 2012). Building a structural model is cast as a computational optimization problem, whereby knowledge about the modeled system can be used in five different ways, guided by maximizing the accuracy and precision of the model while remaining computationally feasible: (1) representing components of a model, (2) scoring a model for its consistency with input information, (3) searching for good-scoring...
models, (4) filtering models based on input information, and (5) validating the resulting models (Rout and Sali, 2019). IMP is designed to allow mixing and matching of different molecular representations, scoring functions, and sampling schemes. It has been used mainly for structural modeling of macromolecular complexes by assembling subunits of known structure based on data from 3DEM, CX-MS, FRET, SAS, hydrogen-deuterium exchange mass spectrometry (HDX-MS), and various proteomics and bioinformatics methods. Integrative structures of several complexes determined using IMP have been deposited in PDB-Dev, including the nuclear pore complex (Kim et al., 2018) and various of its subcomplexes (Kim et al., 2014; Shi et al., 2014; Fernandez-Martinez et al., 2016; Upla et al., 2017), exosome (Shi et al., 2015), mediator (Robinson et al., 2015), 26S proteasome (van Zundert et al., 2015), the human mitochondrial iron sulfur cluster core complex (Cai et al., 2018), the human Rev7 dimer (Rizzo et al., 2018), and the nucleosome complex with RNF168-RING domain and Ubiquitin (Horn et al., 2019). Work is in progress to support automated deposition of files created by HADDOCK into PDB-Dev.

**Rosetta**

Rosetta (Leaver-Fay et al., 2011) is a comprehensive software suite for macromolecular modeling and design. Rosetta provides a wide range of functionalities, including de novo structure prediction, protein design, small-molecule and protein docking, and modeling based on restraints derived from a variety of experimental techniques such as X-ray crystallography, NMR, 3DEM, SAS, HDX-MS, CX-MS, and EPR. Restraints can be combined in flexible forms. RosettaScripts (Fleishman et al., 2011) and PyRosetta (Chaudhury et al., 2010) allow for the development of problem-tailored protocols in a plug-and-play fashion, allowing incorporation of multiple sources of experimental data in a single computational experiment. It has been demonstrated

---

**Figure 3. Schematic Representation of the Pipeline for Archiving Integrative Structures (Top Panel) and the Future wwPDB OneDep Pipeline (Bottom Panel)**

The blue boxes in the top panel show the past and ongoing development projects for archiving integrative structures. These projects include creation of the data representation, development of specific methods for annotation and validation of integrative structures, and creation of a prototype deposition and archiving system, called PDB-Dev (Valat et al., 2018). The green boxes show current and future components of the wwPDB OneDep pipeline (Young et al., 2017). The methods developed for processing and archiving integrative structures in the top panel will be transferred into the wwPDB OneDep pipeline in the bottom panel to provide support for integrative structures within OneDep.
that Rosetta can refine integrative structures and accurately add atomic details not present in the experimental data (Wang et al., 2016). The Rosetta software package is open-source, free for academic use, and developed by the RossettaCommons consortium that new developers can join readily. Rossetta-based integrative structures that have been deposited into PDB-Dev include structures of the serum albumin domains in human blood serum (Belsom et al., 2016), the peptide Ghrelin bound to its G-protein-coupled receptor (Bender et al., 2019), HCN voltage-gated ion channel (Dai et al., 2019), and the native BBsomer (Chou et al., 2019). Work is in progress to implement support within Rosetta for creating data files that can be archived in PDB-Dev.

**Bayesian Inference of Ensembles**

BioEn (Bayesian Inference of Ensembles) is a modeling application that integrates data from diverse experiments with reference ensemble information obtained from simulation or modeling using a Bayesian framework (Hummer and Kofinger, 2015). It enables assessment of the quality and consistency of the experimental data as well as the reference ensemble. The method has been successfully applied to model structures based on EPR data, such as the dimeric SLC26 transporter (Chang et al., 2019), which has been deposited into PDB-Dev. In addition, ensemble refinement based on SAS data has been used to determine the solution structures of the Atg1-Atg13 and Atg17-Atg31-Atg29 subcomplexes and the Atg1 complex (Kofinger et al., 2015). Ongoing research is focused on the development of mechanisms to deal with inconsistent data, automated assessment of model and data quality, and designing a formalism to assess error estimates (Kofinger et al., 2019).

**Integrative Modeling with CNS and X-plor**

The flexibility of general-purpose structure refinement programs, such as X-plor (Brünger, 1992) and CNS (Brünger et al., 1998), made it possible to generate protocols for integrative structure modeling. For example, the complex between single-stranded DNA and single-stranded DNA binding protein of a filamentous bacteriophage was modeled based on stoichiometry and data from low-resolution electron microscopy (EM) and NMR spectroscopy (Folmer et al., 1994); the complex of multi-functional hexameric arginine repressor with DNA was modeled based on chemical footprinting (Sunnerhagen et al., 1997); and structures of bacterial pili were modeled based on symmetry derived from low-resolution 3DEM data, crosslinking, and double charge-inversion mutations (Campos et al., 2010, 2011). Similarly, a coarse-grained model of RNA polymerase Pol III was sampled by a Bayesian, ISD-like method implemented in CNS, based on restraints from crosslinking mass spectrometry (Ferber et al., 2016).

**Biochemical Library**

The Biochemical Library (BCL) program models proteins as assemblies of secondary structure elements (Karakas et al., 2012). The BCL can simultaneously use experimental restraints from 3DEM (Lindert et al., 2009), NMR (Weiner et al., 2014), EPR (Fischer et al., 2015), CX-MS (Hofmann et al., 2015), and SAS (Putnam et al., 2015) experiments. The rationale for replacing flexible loop regions with a loop closure constraint is to substantially reduce the conformational space of a protein, correspondingly reducing the sampling challenge. As many experimental data points relate to secondary structure elements, sampling can often be simplified without substantially reducing the experimental data used for structure determination. The strength of BCL lies in modeling proteins that are rich in secondary structure, such as membrane proteins (Weiner et al., 2013). It has been used, for example, to compute a structural model for the phage T4 recombination mediator protein UvsY (Gajewski et al., 2016).

**Modeling of Genomes Using Hi-C Data**

Data obtained from chromosome conformation capture (Hi-C) experiments can be used to model the three-dimensional structures of genomes (Oluwadare et al., 2019). TADBir (Serra et al., 2017) and Population-based Genome Structure (PGS) (Hua et al., 2018) are two software packages that model 3D genome structures from Hi-C data. TADBir relies on IMP, using modeling by satisfaction of spatial restraints to build 3D structures of genomes from chromatin interaction frequencies obtained through Hi-C experiments. PGS uses a population-based probabilistic approach to model 3D genome structures that are consistent with chromatin-chromatin interaction probabilities obtained from Hi-C data. The multi-scale 3D Chromatin model of the first 4.5 Mb of Chromosome 2L from the Drosophila melanogaster genome (Trussart et al., 2015) obtained using TADBir has been deposited in PDB-Dev. Work is in progress to archive three-dimensional models of the human genome obtained using PGS.

**ChimeraX**

ChimeraX (Goddard et al., 2018) is a new software application for the visualization and analysis of molecular structures and associated data built using the extensive code base, knowledge, and experience gained from Chimera (Pettersen et al., 2004). It can be used to visualize the integrative structures archived in PDB-Dev. Correspondingly, ChimeraX enables the visualization of multi-scale ensembles composed of atomic and coarse-grained bead representations, input spatial restraints such as distances from CX-MS experiments, 2DEM images, and 3DEM maps, as well as preliminary validation information regarding satisfaction of input restraints. Satisfied and violated crosslinks are displayed in different colors in ChimeraX, thus facilitating the visualization of preliminary validation information.

**Visual Molecular Dynamics**

Visual Molecular Dynamics (VMD) is a rapidly evolving modeling and visualization platform that provides tools for simulation preparation, visualization, and analysis (Humphrey et al., 1996). In particular, it is applicable to large-scale systems and datasets. VMD uses advanced technologies to enable cell-scale modeling and visualization using all-atom and coarse-grained molecular representations. It can also integrate experimental data, such as cryo-EM density maps. Work is in progress to support visualization of integrative structures archived in PDB-Dev and to create new graphical interfaces to query and interact with the data. The current focus is on visualizing multi-scale ensembles, restraint information from experiments, statistical inferences, and associated model uncertainties.

**Standards for Representing, Validating, and Archiving Experimental Data**

Data standards are required to build stable databases and to exchange data among different software programs. The various levels of data standards include data definitions for the experimental and computational methods as well as descriptions of
the chemistry and structures. As validation methods are developed, clear definitions for the relevant terms must be created for these methods. The process of creating generally adopted standards requires participation among community stakeholders. These stakeholders include experimentalists, software developers, and the stewards of databases. Once the standards are created and codified into dictionaries, there needs to be cooperation by the journals and funders in enforcing the standards.

We next describe standards for structures derived from traditional single experimental methods followed by emerging standards for experimental and computational methods contributing to integrative structural biology.

**Standards for Models Derived by Single Methods**

Following the establishment of the PDB and the enforcement of data deposition into the PDB as a requirement for publication in journals, efforts to further standardize the data began. A data dictionary for macromolecular crystallography was created as an International Union of Crystallography (IUCr)-sponsored community effort (Bourne et al., 1997). The dictionary, called mmCIF, contained more than 3,000 definitions for many aspects of the X-ray crystallography experiments, as well as definitions for the chemistry and the three-dimensional structures. Over time, extensions have been added for the other methods used for structure determination. The extended dictionary is called PDBx. A resource site contains the dictionary, software, and general information about mmCIF (mmcif.wwpdb.org). The Master Format for the PDB Core Archive is now PDBx/mmCIF (Fitzgerald et al., 2005).

After the community demanded to require structure factors as part of data deposition in 2008, an X-ray Validation Task Force was established with the goal of creating standards for validation of structures determined using X-ray crystallography data. Their recommendations were published in 2011 (Read et al., 2011) and were implemented as part of the wwPDB OneDep system (Gore et al., 2012; Young et al., 2017).

Biomolecular NMR data are deposited into the BMRB (Ulrich et al., 2008) and the structural models into the PDB. An NMR Data Exchange Format (NEF) for representation of chemical shift and restraint data with future extensions to various other data, as well as relevant metadata, has been created (Gutmanas et al., 2015). NEF is a subset of the more comprehensive NMR-STAR format employed for the BMRB Core Archive (Ulrich et al., 2018). The wwPDB NMR Validation Task Force (NMR VTF) was established and published recommendations in 2013 (Montelione et al., 2013). The first set of recommendations was implemented in the wwPDB NMR validation pipeline using existing software. The NMR VTF has worked with the NMR community to develop standards for designating representative structures from a set of deposited models, and for defining well-defined versus ill-defined regions of protein structures. It has recommended that the depositor be allowed to also provide a depositor-designated representative structure. This structural representation information is essential for users of models generated from NMR data. Longer-term goals include handling of all aspects of dynamic processes, including multi-conformers, multi-model ensembles, partially and completely unfolded proteins, as well as all types of biomolecules studied by NMR, including proteins, nucleic acids, polysaccharides, and small molecules.

The 3DEM community has developed a common metadata standard for archiving both experimental maps and map-derived structural models (Lawson et al., 2011; Patwardhan and Lawson, 2016). Incorporation of the standard into the PDBx/mmCIF dictionary enables joint deposition of 3DEM maps into the EMDB (Tagari et al., 2002) and 3DEM models into the PDB (wwPDB consortium, 2019). Raw two-dimensional image datasets may be archived separately in the EMPIAR (Iudin et al., 2016). A 3DEM Validation Task Force that met in 2010 emphasized the need to develop and standardize validation practices and metrics for evaluation and comparison of maps and models (Henderson et al., 2012). Subsequent workshops and community challenges are helping to advance this effort (Patwardhan et al., 2012, 2014; Baker, 2018; Editorial, 2018; Lawson and Chiu, 2018). A follow-up meeting focused on 3DEM map/model validation is planned for 2020.

**Standards for Other Experimental Methods Providing Information for Integrative Modeling**

The experimental methods that can contribute to integrative structure determination include traditional 3D structure determination methods (X-ray crystallography, NMR spectroscopy, and 3DEM) as well as many other methods that provide restraints on, for example, solvent exposure, regions of interaction, and shapes and relative dispositions of components (Table 1 in Rout and Sali, 2019). The heterogeneity of input information presents a significant challenge not only for archiving the final model, as is addressed above, but for making the input information available for validation and potentially further refinement as new data emerge. The challenges are manifold. First, individual communities have to agree on standards for their data and criteria to ensure quality and reliability. Next, these communities must communicate with each other to ensure that data exchange is facilitated. Various communities are at different stages of this coordination.

SAS was one of the first methods to be combined with the PDB standard bearers (X-ray crystallography, NMR spectroscopy, and 3DEM) in computing integrative structures. With the rapid increase in the number of non-expert users, the field saw a wide variability in reporting of data and results. Thus, experts in SAS recognized the need for quality assurance regarding sample provenance, measurement, and processing of data, underpinned by standard tools for assessing the data and models. With sustained community input, preliminary guidelines were developed (Jacques et al., 2012a, 2012b), followed by their adoption by the International Union of Crystallography (IUCr) Commission Journals in 2012. In 2014, the wwPDB SAS Validation Task Force (SAS VTF) was established (Trewhella et al., 2013) and expanded the guidelines to provide additional recommendations for archiving SAS data. One of the key recommendations of the SAS VTF was to bring together structural biology leaders to address the challenges involved in archiving integrative structures. The 2014 wwPDB IHHM Task Force meeting (Sali et al., 2015) was the realization of this recommendation.

A universal exchange dictionary for SAS named sasCIF was established in 2000 (Malfois and Svergun, 2000). The sasCIF Data Dictionary was then extended to describe the experimental information, results, and models, including relevant metadata for analysis and validation of the data and models (Kachala et al., 2016). Processing tools for these files have been developed...
and made available as open-source programs. The SASBDD repository (Valentini et al., 2015) was established as a searchable public repository for SAS data and models; it currently contains more than 1,100 released entries with more than 350 additional entries on hold. In 2017, the biomolecular SAS publication guidelines were updated (Trewella et al., 2017). Recently, a community project was initiated to generate SAS datasets for benchmarking different approaches to predicting SAS profiles from atomic coordinates (sas.wwpdb.org). Finally, a proof-of-concept software system for bidirectional data exchange between SABDB and the PDB is currently under development.

The CX-MS community has recommended proteomics data standards established by the Proteomics Standard Initiative (www.psidev.info [Deutsch et al., 2017b]). These standards include mzML (Martens et al., 2011) as a standard format for raw data and mzIdentML for search results (crosslink identifications). Support for crossinglinking data has been established in mzIdentML 1.2 (Vizzaino et al., 2017), but at this point not all workflows used by the community are supported. Data are increasingly archived in repositories of the ProteomeXchange consortium (Deutsch et al., 2017a) and ChorusProject (chorusproject.org). Work is in progress to reach agreement on minimal metadata standards, to expand crossinglink support in mzIdentML, and to develop reporting standards for publication. A definition for reporting crossinglink restraints is already available in the new extension dictionary for integrative modeling; the development of tools for the seamless integration of MS and modeling data is therefore an obvious next step.

An extension of the PDBx/mmCIF dictionary with terms for fluorescence-based experiments with a current focus on FRET has been created recently (github.com/ihmwg/FLR-dictionary). This extension includes the description of fluorescent probes and resulting FRET-derived inter-dye distances. These extensions can also be applied to other probe-based spectroscopies, such as paramagnetic relaxation enhancement in NMR and spin labels for double electron-electron resonance in EPR. A recent multi-laboratory FRET benchmark study demonstrated the precision and accuracy of FRET measurements for double-stranded DNA rulers (estimated uncertainty in relative distance measurement deviation of less than 0%–5% is well within the expected error) as well as documented measurement and analysis procedures (Hellenkamp et al., 2018). The FRET community (www.FRET.community) was founded to enhance dissemination, community-driven development of analysis tools, and sharing of data and tools. Even though the starting point and scientific focus of this community is FRET spectroscopy and imaging, it is open to members of other communities, including those that use other types of fluorescence techniques. Currently, researchers in the FRET community perform benchmark FRET studies for proteins with the aim to find the best tool for extracting kinetic information from single-molecule traces (kinSoftChallenge, 2019). In addition to these community-driven experimental and computational challenges, work is in progress to achieve agreement on minimal metadata, establish a standard file format to provide workflow support, establish guidelines for documentation and validation of experiments, analysis, and simulations, and create reporting standards for publication. A key goal is to standardize methods for the validation of fluorescence-based structural models. A proposal to create a Fluorescence Biological Data Bank is in progress, aiming to archive data from fluorescence experiments. A number of workshops have been held to discuss FRET and issues of standards and reproducibility in the FRET community. A yearly Workshop is planned as a satellite meeting to MAF (Methods and Applications of Fluorescence) conferences (in 2019 at UC San Diego and in 2020 at Chalmers University of Technology, Gothenburg).

The HDX community is in the early stages of developing its standards for reporting and data deposition. The International Society for HDX Mass Spectrometry was formed (www.hdxms.net) in 2017, in part to address the high degree of variability in methods, data reporting, and interpretation employed within this rapidly growing field. The community recently published the “Gothenburg Guidelines” describing best practices for performing and reporting HDX-MS experiments (Masson et al., 2019). A recent Workshop engaged the wider structural community to learn from experiences in establishing durable community standards. As a result of these efforts, the international society formed a task group to develop a position on the adoption of a data-exchange dictionary, the creation of data standards, and an open archive for HDX-MS data. Discussions are under way with the proteomics community at the EBI for archiving data in the PRIDE database (Vizzaino et al., 2013, 2016) as well as in ChorusProject (chorusproject.org). In addition to standardization of HDX-MS data reporting and deposition, the HDX community has also been engaged in interpretation of HDX-MS data. Despite being complementary to structure-based methods, the current role of HDX in integrative structural analysis is only qualitative; although solvent exchange is generally correlated with protein dynamics, the structure-rate relationship of protein solvent exchange remains ambiguous (Skinner et al., 2012a, 2012b).

EPR spectroscopy, also known as electron spin resonance (ESR) spectroscopy, in combination with site-directed spin labeling, generates long-range distance restraints (in the 1.5- to 8-nm range) for macromolecular characterization. Recently, different paramagnetic labels have been developed and optimized for such applications. Several software tools to obtain distance distributions from time-domain EPR data are available (e.g., DeerAnalysis [Jeschke et al., 2006]). The EPR community is currently working on a white paper with recommendations for experimental procedures and data standards for pulsed dipolar spectroscopy. In a first step, the German Research Society will initiate an international EPR Expert Workshop at the end of 2019. This initiative results from the strong interactions between the German Priority Program New Frontiers in Sensitivity for EPR Spectroscopy (spp1601.de) and the NSF-funded USA-based sharedEPRnetwork (sharedEPR.org). Expected outcomes of this meeting are recommendations for experimental procedures, data standards for publications, and quality assessments of EPR data. A task force will describe the final protocols in a white paper. The International EPR (ESR) Society (www.ieprs.org) has committed to supporting and hosting an open database for original EPR time traces and the resulting distance restraints.

Standards for Computational Methods Providing Information for Integrative Modeling

In addition to experimental information, prior models, such as computationally derived structural models of components, secondary structure predictions, disorder region predictions, and predicted residue-residue contacts, can also be used in integrative structure modeling.
Following a decision reached at a Workshop held in 2006 (Berman et al., 2006), the PDB archive is restricted to structural models derived from experimental methods. Based on community recommendations (Schwede et al., 2009), the macromolecular ModelArchive (www.modelarchive.org) has been built to archive structural models that are not based on experimental information about the modeled system, such as homology models, ab initio predictions, and models based on contact distances predicted by co-evolutionary analysis and deep learning approaches (Ovchinnikov et al., 2015; Koscielak and Jones, 2016; Hou et al., 2019). About 1,500 models have been made publicly accessible in ModelArchive so far. An extension of the PDBx/mmCIF dictionary for representing computational models was developed recently (github.com/ihmwg/MA-dictionary), aiming to facilitate the development of methods for efficient data exchange among the structural model repositories (PDB, ModelArchive, and PDB-Dev; Figure 1). Work is in progress to support the new dictionary within the SWISS-MODEL repository (Bienert et al., 2017; Waterhouse et al., 2018) and ModelArchive.

The Critical Assessment of Protein Structure Prediction (CASP) has been exploring modeling methods based in part on sparse experimental data, including data from SAS, NMR, cross-linking, and FRET. This Integrative CASP Experiment was highlighted at the recent CASP13 meeting (www.predictioncenter.org/casp13), and the resulting manuscripts are currently in review. In particular, CASP has catalyzed continued development of methods for contact prediction from evolutionary co-variance data (Schaarschmidt et al., 2018). Several of the fully automated structure prediction methods participating within the Continuous Automated Model Evaluation (CAMEO) platform infer and subsequently integrate contact predictions in their pipelines. Such contact predictions have already been combined with sparse experimental NMR data for integrative modeling of protein structures (Tang et al., 2015).

Standards for Validating Integrative Structures
A structural model of any type must be validated to evaluate how it can be interpreted. Standardized validation of integrative structures will ultimately be part of depositions into the PDB, as is already the case for structures derived using traditional methods (Read et al., 2011; Henderson et al., 2012; Montellone et al., 2013; Trewella et al., 2013, 2017; Gore et al., 2017). Thus, an effort to build a validation pipeline for integrative structures and incorporate it into the OneDep (Young et al., 2017) deposition system was initiated under the auspices of the wwPDB. The input for validation will be the integrative structure and the data used to compute it, represented in the standard format. The output will be a validation report listing validation criteria, presented graphically in a pdf file or on a web page, relying heavily on the extensive experience of the wwPDB working with the structural biology community. The validation report will facilitate reviewing, publishing, and using the results of integrative structural biology studies. A standardized table will report key parameters of a study, similar to such tables used for other structure determination methods (Read et al., 2011; Trewella et al., 2017).

The proposed wwPDB validation pipeline for integrative structures borrows from the validation implemented in IMP (Rout and Sali, 2019). In addition, it is informed by feedback from the members of the Model Working Group of the wwPDB IHM Task Force and members of the broader integrative structural biology community. The validation pipeline will leverage existing software developed by the structural biology community (e.g., wwPDB [Gore et al., 2017], MolProbity [Williams et al., 2018], BMRB [Ulrich et al., 2008], EMDB [Tagari et al., 2002; Lawson et al., 2016; Patwardhan and Lawson, 2016], SASDBB [Valentini et al., 2019], PHENIX [Adams et al., 2010], and PDBStat [Tejero et al., 2013]). For the time being, the proposed wwPDB validation criteria for integrative structures are organized into five broad categories, described in the following five subsections.

Quality of the Data
The quality of an integrative structure clearly depends on the quality of the data used to compute it (cf., garbage in, garbage out). Thus, it is essential to annotate integrative structures with data-quality measures. These measures are best established by the communities generating the data, illustrating one benefit of the wwPDB Federation model. Importantly, the data-quality criteria need to be computable only from the deposited data and its annotations, without requiring non-deposited information or the structural model itself. Examples include the resolution of the EM map, the false-positive rate of chemical crosslinks, and the adequacy of the measurement range and signal-to-noise ratio of a SAS profile.

Standard Criteria for Assessing Atomic Models
Some integrative structures or their parts may be represented at atomic resolution. In such cases, all criteria for assessing the quality of atomic structures already implemented in OneDep (Young et al., 2017) (e.g., clash score, Ramachandran plot outliers, and side-chain outliers) will be adopted, as provided by the MolProbity program (Williams et al., 2018). This assessment may result in annotating some regions as well-defined versus ill-defined, similar to the annotation of structural ensembles determined by NMR spectroscopy (cf. section Uncertainty of the Model). Use of tools developed in the CAMEO project (Haas et al., 2018) will also be explored.

Fit of a Model to Information Used to Compute It
A model must sufficiently satisfy the data used to compute it. We will adopt standard validation criteria for assessing the fit of a model to these data; for example, cross-correlation coefficient between the model and the EM map, the fraction of chemical crosslinks satisfied by the model, and the discrepancy $\chi^2$ value between the computed and experimental SAS profiles combined with the goodness-of-fit test for the correlation map (e.g., the p value from Franke et al., 2015). We may need to improve these validation criteria; for example, the threshold on the cross-correlation coefficient between an EM map and a model may depend on the degree of coarse-graining of the model. We will also ensure that all criteria are compatible with the richness of the molecular representations available for integrative structures (i.e., ensembles of multi-scale and multi-state structures) (see the section Molecular Representation of Integrative Structures). Because both integrative structure modeling and NMR-based modeling involve satisfaction of spatial restraints, lessons will be learned from quantifying spatial restraint satisfaction in NMR-based modeling (Tejero et al., 2013; Gutmannas et al., 2015).

Violations of input data by the model occur when the data are more uncertain than assumed (e.g., the false-positive rate of chemical crosslinks is higher than the presumed threshold), the representation of a model is incorrect (e.g., a subunit structure
in the modeled complex is not rigid or the system exists in multiple states instead of a single state), the scoring is incorrect (e.g., a crosslink restraint does not consider the ambiguity resulting from multiple copies of a crosslinked subunit in the modeled system), and/or the sampling is not sufficient (i.e., a model that satisfies all the data does exist but was simply not found by the sampling scheme). Thus, this test provides immediate feedback for improving the modeling protocol.

**Fit of a Model to Information Not Used to Compute It**

A particularly informative test is a comparison of a model against the data that were not used to compute the model. Validation criteria described in the previous section apply, except perhaps with more lenient thresholds. We will encourage deposition of such additional unused data with the model so that the corresponding standard tests can be performed during deposition.

Resampling tests (e.g., jack-knifing and bootstrapping) consist of repetitively omitting a random subset of the input data, recomputing the model, and comparing the models against the omitted data, to validate both the model and the data. Such tests are the basis for the $R_{	ext{free}}$ criterion in X-ray crystallography (Brünger, 1993) and the use of half-maps in modeling based on 3DEM data (van Heel and Schatz, 2005; Chen et al., 2013; Afonine et al., 2018). An example from integrative structure modeling is using random multiple subsets of chemical crosslinks to assess the Nup84 heptamer model (Fernandez-Martinez et al., 2012; Shi et al., 2014). Unfortunately, these resampling tests can only be performed by the depositors themselves, because the wwPDB validation pipeline cannot reproduce a modeling protocol used for each deposited structure. Accordingly, the authors will be encouraged to perform resampling tests before the deposition and report the results in a standardized manner during model deposition.

**Uncertainty of the Model**

One of the most useful assessments of a model is quantification of its uncertainty. Model uncertainty is most explicitly described by the set of “all” models that are sufficiently consistent with the input information (i.e., the model ensemble; correspondingly, the entire ensemble, not just a single representative member, is in fact the model). In practice, computing such an ensemble requires sufficient structural sampling, which is often neither performed nor tested (Viswanath et al., 2017). If an ensemble is available, model precision can be assessed by analyzing the variability among the models constituting the ensemble. The ensemble can optionally be described by one or more representative models and their uncertainties (e.g., when an ensemble consists of multiple clusters of models, each cluster can be represented by its centroid model). Importantly, the uncertainty is generally not distributed evenly across a model. Only those model features that are coarser than model uncertainty can be interpreted. Thus, the model needs to be annotated by its uncertainty, and tools for visualizing this uncertainty need to be further developed. The model uncertainty reflects the actual heterogeneity of the physical sample(s) used to obtain the data as well as the uncertainties in the input information, representation of the model, and scoring of the alternative models. It is generally difficult to deconvolute the effects of these different uncertainties on the model uncertainty.

Because of the importance of estimating model uncertainty, the authors will be encouraged to develop and apply modeling methods that compute a complete ensemble of models consistent with input information and estimate sampling precision for their method (Viswanath et al., 2017). However, not all useful methods for computing integrative structures are able to produce a representative ensemble of models (e.g., when models are constructed by hand or a single model computation is performed). Therefore, we will allow for the following three deposition scenarios.

First, a single structural model is deposited. In such a case, not much can be inferred about the uncertainty of the model from the model itself, although some empirical methods for estimating uncertainty based on a single model may yet be developed (cf., the accuracy of a comparative model is correlated with the sequence similarity to the template structure on which it is based or with a structure-dependent statistical potential score). To encourage quantification of uncertainty, the IHM Dictionary will provide terms for specifying the uncertainty of each part of an integrative structure, similarly to the atomic B factors in the crystallographic structure files.

Second, a small ensemble of structural models is deposited, potentially representing more than one cluster of solutions. Here, we will consider adopting the best practices of the NMR community (Montelione et al., 2013), as follows. The total uncertainty of a model, resulting from both the lack of information and sample heterogeneity, is represented approximately by a relatively small ensemble of 20–30 structures, which is often selected from a larger ensemble of 50–100 structures. The deposited ensemble is annotated by identifying the medoid structure that is most similar to all the other structures. Furthermore, well- and ill-defined regions within the ensemble are identified, using domain identification and local superposition to eliminate artifacts that can result from global superposition (Kirchner and Guntert, 2011).

Third, a large ensemble of structural models is deposited, again potentially representing more than one cluster of solutions. For example, IMP routinely generates thousands of structural models that represent as completely as possible all structures that satisfy the input information (Rout and Sali, 2019). The ensemble is used to estimate the sampling precision (Viswanath et al., 2017), cluster these models based on their structural similarity, and represent the resulting clusters with their localization densities (i.e., the probability of any model component at any grid point [Alber et al., 2007]). These clusters and localization densities are a useful representation of model uncertainty. The corresponding visualization will be implemented in the validation pipeline by relying on the programs such as Chimerax [Goddard et al., 2018] and VMD (Humphrey et al., 1996) as well as the Molstar web application (molstar.org).

Finally, care will be taken to expand the representation of integrative structures in the IHM Dictionary to allow for deposition of all commonly used ensemble depictions (e.g., ensemble modeling of intrinsically disordered proteins or regions based on SAS data).

**Remarks**

While the validation pipeline proposed above will certainly be helpful, it does not include all useful tests, because some criteria cannot be easily applied during deposition at this time. As mentioned above, examples include an estimate of sampling precision, which requires extensive stochastic sampling, and...
data resampling tests, which require repeated modeling with subsets of data. Therefore, describing such validations will by necessity be limited to original papers, contributed by the authors during deposition. It is expected that the validation pipeline will mature over time as more advanced methods are developed and adopted by the community.

Similarly, the validation of structural models entirely within the Bayesian framework will eventually be explored. Such a formulation promises the most rigorous and general validation, especially if the models are also computed within the Bayesian framework in the first place. The current proposal does not reflect these future advances; even if they were in hand, many existing useful criteria are not Bayesian. However, we expect that our validation pipeline will eventually be informed by the Bayesian view of computing, assessing, and using models.

Recommendations
To address the challenges involved in archiving integrative structures, the Workshop participants were divided into two discussion groups that focused on (1) standards and data exchange and (2) validation of integrative models. Their collective recommendations are summarized as follows.

1. Continue to develop the IHM Dictionary for integrative structures with standard definitions for the experimental and computational methods used for integrative modeling. This dictionary-based approach will allow for maximum interoperability among the experimental and computational methods used for structure determination and ultimately facilitate deposition of integrative structures into the PDB.

2. Develop new tools that will facilitate dictionary development in the PDBx/mmCIF framework. Such tools are critical to accelerate the development of resources needed to archive structures.

3. Promote the development of common data standards that will enable efficient data exchange among scientific repositories contributing to structural biology.

4. Create a validation pipeline for integrative structures, including measures of the quality of the data on which the structures were based, the standard criteria for assessing atomic models, the fit of a model to information used to compute it, the fit of a model to information not used to compute it, and uncertainty of the model.

5. Raise awareness among journal editors regarding the new standards being developed for structure determination and the emergence of new data repositories, and advocate for depositing structures and data prior to publication.

6. Raise awareness broadly, including at funding agencies, of the critical need for support of the underlying hardware, software, and personnel with expert knowledge, that together form the infrastructure essential for the archiving of integrative structures.

ACKNOWLEDGMENTS
Funding for the BPS Satellite meeting was provided by an NSF OAC-1838628 to H.M.B. Additional funding for the development of the pipeline for integrative archiving are: NSF DBI-1756248 (H.M.B.), NSF DBI-1756250 (A.S.), NSF DBI-1519158 (H.M.B., A.S.), DBI-1832184 (S.K.B., A.S.), R01GM083960 (A.S.), and P41GM109824 (A.S.).

AUTHOR CONTRIBUTIONS
Writing—Original Draft, H.M.B., A.S., J.T., and B.V.; Writing—Review and Editing, all authors.

DECLARATION OF INTERESTS
The authors declare no competing interests.

REFERENCES


tion guidelines for structural modelling of small-angle scattering data from bio-

pretation of small-angle scattering data from bio-
molecules in solution—the importance of quality control and a standard reporting framework. BMC Struct. Biol. 12, 9.

hensive software package for analyzing pulsed ELDOR data. Appl. Magn. Re-
son. 30, 473–498.


tallogr. 49, 302-310.


ution structure of the Atg1 complex: implications for the architecture of the phagophore assembly site. Structure 23, 809–818.

put. 15, 3390–3401.


Lawson, C.L., Baker, M.L., Best, C., Bi, C., Dougherty, M., Peng, P., van Gin-


Lawson, C.L., Baker, M.L., Best, C., Bi, C., Dougherty, M., Peng, P., van Gin-


**wwPDB consortium (2019). Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res. 47, D520–D528.**