Outcome of the First wwPDB Hybrid / Integrative Methods Task Force Workshop

Andrej Sali[1], Helen M. Berman[2], Torsten Schwede[3], Jill Trewhella[4], Gerard Kleywegt[5], Stephen K. Burley[2,6], John Markley[7], Haruki Nakamura[8], Paul Adams[9], Alexandre M.J.J. Bonvin[10], Wah Chiu[11], Matteo Dal Peraro[12], Frank Di Maio[13], Thomas E. Ferrin[14], Kay Grünewald[15], Aleksandras Gutmanas[5], Richard Henderson[16], Gerhard Hummer[17], Kenji Iwasaki[18], Graham Johnson[19], Catherine L. Lawson[2], Jens Meiler[20], Marc A. Marti-Renom[21], Gaetano T. Montelione[22,23], Michael Nilges[24,25], Ruth Nussinov[26,27], Ardan Patwardhan[5], Juri Rappsilber[28,29], Randy J. Read[30], Helen Saibil[31], Gunnar F. Schröder[32], Charles Schwieters[33], Claus A. M. Seidel[34], Dmitri Svergun[35], Maya Topf[31], Eldon L. Ulrich[7], Sameer Velankar[5], and John D. Westbrook[2]

[1] Department of Bioengineering and Therapeutic Sciences, Department of Pharmaceutical Chemistry, California Institute for Quantitative Biosciences, Byers Hall Room 503B, University of California, San Francisco, 1700 4th Street, San Francisco, CA 94158-2330, USA

[2] Research Collaboratory for Structural Bioinformatics Protein Data Bank, Center for Integrative Proteomics Research, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA.

[3] SIB Swiss Institute of Bioinformatics; Biozentrum, University of Basel, Klingelbergstrasse 50-70, CH-4056 Basel, Switzerland

[4] Department of Molecular Bioscience, The University of Sydney, NSW 2006, Australia

[5] Protein Data Bank in Europe, European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom

[6] Skaggs School of Pharmacy and Pharmaceutical Sciences and San Diego Supercomputer Center, University of California, San Diego, La Jolla, CA 92093, USA.

[7] BioMagResBank, Department of Biochemistry, University of Wisconsin-Madison, Madison, WI 53706-1544, USA.

[8] Protein Data Bank Japan, Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan

[9] Physical Biosciences Division, Lawrence Berkeley Laboratory, Berkeley, CA 94720-8235, USA, and Department of Bioengineering, UC Berkeley, Berkeley, CA 94720, USA

[10] Bijvoet Center for Biomolecular Research, Faculty of Science - Chemistry, Utrecht University, Padualaan 8, Utrecht, 3584 CH, the Netherlands

[11] National Center for Macromolecular Imaging, Baylor College of Medicine, Houston, TX 77030, USA

[12] Institute of Bioengineering, School of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL) and Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland

[13] Department of Biochemistry, University of Washington, Seattle, WA 98195-7370

[14] Department of Pharmaceutical Chemistry, Department of Bioengineering and Therapeutic Sciences, California Institute for Quantitative Biosciences, University of California, San Francisco, 600 16th Street, San Francisco, CA 94158-2517, USA.

[15] Nuffield Department of Medicine, University of Oxford, Oxford, UK

[16] MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge CB2 0QH, UK

[17] Department of Theoretical Biophysics, Max Planck Institute of Biophysics, Max-von-Laue Str. 3, 60438 Frankfurt am Main, Germany

[18] Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan

[19] Department of Bioengineering and Therapeutic Sciences, California Institute for Quantitative Biosciences, University of California, San Francisco, 600 16th Street, San Francisco, CA 94158-2330, USA

[20] Department of Chemistry, Center for Structural Biology, Vanderbilt University, Nashville, TN 37235

[21] Genome Biology Group. Centre Nacional d'Anàlisi Genòmica (CNAG), Gene Regulation. Stem Cells and Cancer Program. Center for Genomic Regulation (CRG) and Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

[22] Center for Advanced Biotechnology and Medicine, and Department of Molecular Biology and Biochemistry, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

[23] Department of Biochemistry, Robert Wood Johnson Medical School, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

[24] Département de Biologie Structurale et Chimie, Unité de Bioinformatique Structurale, Institut Pasteur, F-75015 Paris, France

[25] Unité Mixte de Recherche 3258, Centre National de la Recherche Scientifique, F-75015 Paris, France

[26] Cancer and Inflammation Program, Leidos Biomedical Research Inc., Frederick National laboratory, National Cancer Institute, Frederick, MD 21702, USA

[27] Department of Human Molecular Genetics and Biochemistry, Sackler School of Medicine,Tel Aviv University, Tel Aviv 69978, Israel

[28] Wellcome Trust Centre for Cell Biology, Institute of Cell Biology, University of Edinburgh, Edinburgh EH9 3BF, UK

[29] Department of Bioanalytics, Institute of Biotechnology, Technische Universität Berlin, 13355 Berlin, Germany

[30] Department of Haematology, Cambridge Institute for Medical Research, University of Cambridge, Cambridge, UK.

[31] Institute of Structural and Molecular Biology, Department of Biological Sciences, Birkbeck College, Malet St, London WC1E 7HX, UK

[32] Institute of Complex Systems (ICS-6), Forschungszentrum Jülich, 52425 Jülich, Germany and Physics Department, Heinrich-Heine University Düsseldorf, 40225 Düsseldorf, Germany

[33] Division of Computational Bioscience, Center for Information Technology, National Institutes of Health, Bethesda, MD 20892-0520, USA

[34] Chair for Molecular Physical Chemistry, Heinrich-Heine-Universität, Universitätsstraße 1, 40225 Düsseldorf, Germany

[35] European Molecular Biology Laboratory, Hamburg Unit, Notkestrasse 85, 22607 Hamburg, Germany

All attendees of the Workshop are listed as authors.

Running Title: Integrative Workshop White Paper

Keywords: integrative modeling, hybrid modeling, integrative structural biology, Protein Data Bank

Summary

Structural models of biomolecular systems are increasingly determined by integrative modeling that relies on varied types of experimental data and theoretical information. We describe here the proceedings and conclusions from the first wwPDB Hybrid / Integrative Methods Task Force Workshop held at the European Bioinformatics Institute in Hinxton, UK, October 6 and 7, 2014. At the workshop, experts in various experimental fields of structural biology, experts in integrative modeling and visualization, and experts in data archiving addressed a series of questions central to the future of structural biology. How should integrative models be represented? How should the data and integrative models be validated? What data should be archived? How should the data and models be archived? What information should accompany the publication of integrative models?

1 Background

1.1 Historical rationale for the Workshop

The Protein Data Bank (PDB; wwpdb.org) was founded in 1971 with seven protein structures as its first holdings (Protein Data Bank, 1971). The global PDB archive now holds more than 100,000 atomic structures of biological macromolecules and their complexes, all of which are freely accessible. Most structures in the PDB archive (~90%) have been determined by X-ray crystallography, with the remainder contributed by two newer 3D structure determination methods, nuclear magnetic resonance (NMR) spectroscopy and electron microscopy (3DEM).

Considerable effort has gone into understanding how to best curate the structural models and experimental data produced with these methods. Over the past several years, the Worldwide Protein Data Bank (wwPDB; the global organization responsible for maintaining the PDB archive) (Berman et al., 2003) has established expert, method-specific Task Forces to advise on which experimental data and metadata from each method should be archived and how these data and the structure models therefrom should be validated. The wwPDB X-ray Validation Task Force (VTF) made detailed

recommendations on how to best validate structures determined by X-ray crystallography (Read et al., 2011). These recommendations have been implemented as a software pipeline used within the wwPDB Deposition and Annotation (D&A) system. Initial recommendations of the wwPDB NMR (Montelione et al., 2013) and Electron Microscopy (Henderson et al., 2012) VTFs have also been implemented. In addition, the wwPDB and, in later years, the Structural Biology Knowledgebase (SBKB), spearheaded three workshops focused on validation, archiving, and dissemination of comparative protein structure models (Berman et al., 2006; Schwede et al., 2009). It is anticipated that as new validation methods are developed and as more experience is gained with existing ones, additional validation procedures will be implemented in the wwPDB D&A system.

Increasingly, structures of very large macromolecular machines are being determined by combining observations from complementary experimental methods, including X-ray crystallography, NMR spectroscopy, 3DEM, small-angle scattering (SAS), crosslinking, and many others (**Figure 1, Table 1**). Data from these complementary methods are used to compute integrative or hybrid models (Ward et al., 2013). Atomic models produced in this fashion have been deposited into the PDB, but there is currently no mechanism available within the PDB framework for archiving the experimental data coming from methods other than X-ray crystallography, NMR spectroscopy, and 3DEM. The most recently established task force, the wwPDB SAS Task Force (Trewhella et al., 2013b) recommended creation of an SAS data and model repository that would interoperate with the PDB. The SAS Task Force also recommended that an international meeting be held to consider how best to deal with the archiving of data and models coming from integrative structure determination approaches.

In response, a Hybrid / Integrative Methods Task Force was assembled by the wwPDB organization. Its inaugural meeting was held at the EMBL European Bioinformatics Institute (EBI) October 6th and 7th 2014 (wwpdb.org/task/hybrid.php). In all, 38 participants from 35 academic and government institutions worldwide attended the workshop, which was co-chaired by Andrej Sali (University of California, San Francisco, USA), Torsten Schwede (SIB and University of Basel, Switzerland), and Jill Trewhella

(University of Sydney, Australia). Attendees included experts in relevant experimental techniques, integrative modeling, visualization, and data and model archiving.

The workshop began with plenary talks followed by focused discussions. Gerard Kleywegt introduced the workshop objectives. Andrej Sali outlined the current state of integrative modeling. Helen Berman gave an overview of the history and status of the wwPDB organization. Jill Trewhella described the increasing role of SAS in integrative structural modeling, the need for the development of community standards and validation tools for biomolecular modeling using SAS data, and how SAS data and modeling resources could interoperate with the PDB. Claus Seidel outlined state-of-the-art single molecule and ensemble Förster resonance energy transfer (FRET) spectroscopy (Kalinin et al., 2012), live cell imaging, as well as related label-based spectroscopic methods for measuring select interatomic distances in macromolecular systems. Torsten Schwede presented the Protein Model Portal (Haas et al., 2013) and its role in integrating large-scale databases of comparative protein structure models with experimental structure information from the PDB, and the Model Archive as a repository for all categories of *in silico* structural models.

1.2 Current archives for models and/or supporting data

In this section, we review the PDB and management of data derived from crystallography, NMR spectroscopy, 3DEM, and SAS, plus archives for models derived exclusively *via* computational methods.

1.2.1 Protein Data Bank

For more than four decades, the PDB has served as the single global archive for atomic models of biological macromolecules; first for those derived from crystallography, and subsequently for models from NMR spectroscopy and 3DEM. The PDB also archives experimental data necessary to validate the structural models determined using these three methods. In addition, descriptions of the chemistry of polymers and ligands are collected, as are metadata describing sample preparation, experimental methods, model building, refinement statistics, literature references, *etc*. For all structural models in the

PDB, geometric features are assessed with respect to standard valence geometry and intermolecular interactions, as recommended by the three wwPDB VTFs described above.

1.2.2 Crystallography: Models and Data

For structures derived using X-ray, Neutron and combined X-ray/Neutron methods, it has been mandatory to deposit structure factor amplitudes into the PDB since 2008 (http://www.wwpdb.org/news/news?year=2007#29-November-2007); until then, it was optional to submit these primary data. Additional validation against deposited structure factor amplitudes is carried out using procedures recommended by the X-ray VTF (Read et al., 2011). The resulting validation report includes graphical summaries of the quality of the overall model plus residue-specific features. Detailed assessments of various aspects of the model and its agreement with experimental and stereochemical data are also provided. In the near future, unmerged intensities will also be collected enabling further validation activities.

1.2.3 Nuclear Magnetic Resonance Spectroscopy: Models and Data

The Biological Magnetic Resonance Data Bank (BioMagResBank or BMRB; www.bmrb.wisc.edu) is a repository for experimental and derived data gathered from NMR spectroscopic studies of biological molecules. The BMRB archive contains quantitative NMR spectral parameters, including assigned chemical shifts, coupling constants, and peak lists together with derived data, including relaxation parameters, residual dipolar couplings, hydrogen exchange rates, pK_a values, *etc*. Other data contained in the BMRB include: NMR restraints processed from original author depositions available from the PDB; time-domain spectral data from NMR experiments used to assign spectral resonances and determine structures of biological macromolecules; chemical shift and structure validation reports; and a database of one- and two-dimensional ^{1}H and ^{13}C NMR spectra for over 1200 metabolites. The BMRB website also provides tools for querying and retrieving data.

Since 2006, BMRB has been a member of the wwPDB organization (Markley et al., 2008). Chemical shift and restraint data that accompany model data are housed in both the BMRB and PDB archives. Deposited NMR data without model coordinates reside exclusively in the BMRB archive. The wwPDB D&A system provides for deposition, annotation, and validation of NMR models and related experimental data. Depositors of chemical shift and other data sets without accompanying models are automatically redirected to BMRB to deposit their data. Data exchange between the BRMB and PDB archives is facilitated by software tools utilizing correspondences maintained between the PDB Exchange Dictionary (PDBx) and the BMRB NMRSTAR Dictionary. Validation methods for NMR-derived models, measured chemical shifts, and restraint data are currently under development, in response to recommendations of the NMR VTF (Montelione et al., 2013). A working group composed of the major biomolecular NMR software developers has created a common NMR exchange format (NEF) for structural restraints, similar to NMR-STAR. The adoption of this NEF by NMR software developers will simplify data exchange and the archiving of NMR structural restraints by the wwPDB.

1.2.4 Electron Microscopy: Models and Maps

Atomistic structural models determined using 3DEM methods were first archived in the PDB in the 1990s. In 2002, the EM Data Bank (EMDB) was created by the Macromolecular Structure Database (now PDBe) at the EBI. In 2006, the EMDataBank (www.EMDataBank.org) was established as the unified global portal for one-stop deposition and retrieval of 3DEM density maps, atomic models, and associated metadata (Lawson et al., 2011). EMDataBank is a joint effort among PDBe, the Research Collaboratory for Structural Bioinformatics (RCSB) at Rutgers, and the National Center for Macromolecular Imaging (NCMI) at Baylor College of Medicine. EMDataBank also serves as a resource for news, events, software tools, data standards, and validation methods for the 3DEM community. 3DEM model and map data are now stored in separate branches of the wwPDB ftp archive site.

As for NMR-based models, the wwPDB D&A system supports processing of atomistic models and map data from 3DEM structure determinations. 3DEM map data deposited without atomistic models are stored exclusively in EMDB. Again, as for NMR, a mapping

is maintained between the PDBx data dictionary and the EMDB XML-based data model. Validation methods for 3DEM maps and atomistic models are currently under development in response to recommendations from the EM VTF (Henderson et al., 2012).

1.2.5 Small-Angle Scattering: Data and Model Archiving

The report from the first meeting of the wwPDB SAS Task Force (Trewhella et al., 2013a) made the case for establishing "a global repository that holds standard format X-ray and neutron SAS data that is searchable and freely accessible for download" and that "options should be provided for including in the repository SAS-derived shape and atomistic models based on rigid-body refinement against SAS data along with specific information regarding the uniqueness and uncertainty of the model, and the protocol used to obtain it."

At present, there are two databases available for storing SAS data and models with associated metadata and analyses, both of which are freely accessible without limitations on data utilization *via* the Internet. As of March 2015, BIOISIS (http://www.bioisis.net/) contained 99 structures and is supported by teams at the Advanced Light Source and Diamond, while SASBDB (http://www.sasbdb.org/) (Valentini et al., 2015) contained 195 models and 114 experimental datasets and is supported by a team at EMBL-Hamburg.

Having evolved separately, these databases are distinctive in character. There was in principle agreement within the wwPDB SAS Task Force that BIOISIS and SASBDB will exchange datasets. Such exchange would be a step toward developing a federated approach to SAS data and model archiving, which in turn could ultimately be federated with the PDB, BMRB, and EMDB.

Further development of the sasCIF dictionary is required to permit full data exchange between the two SAS data repositories. sasCIF is a core Crystallographic Information File (CIF) developed to facilitate the SAS data exchange (Malfois and Svergun, 2000). As its name implies, sasCIF was implemented as an extension of the core CIF dictionary and has recently been extended to include new elements related to models, model

fitting, validation tools, sample preparation, and experimental conditions (M. Kachala, J. Westbrook and D.I. Svergun, in preparation). sasCIFtools were developed as a documented set of publicly available programs for sasCIF data processing and format conversion; currently, SASBDB supports both import and export of sasCIF files.

1.2.6 Protein Model Portal

Comparative or homology protein structure modeling is routinely used to generate structural models of proteins for which experimentally determined structural models are not yet available (Schwede et al., 2009). Until 2006, such *in silico* models could be archived in the PDB, albeit in the absence of clear policies and procedures for validation thereof. Following recommendations from a stakeholder workshop convened in November 2005 (Berman et al., 2006), depositions to the PDB archive are limited to structural models substantially determined by experimental measurements from a defined physical sample (effective date October 15, 2006). The workshop also recommended that a central, publicly available archive or portal should be established for exclusively *in silico* models, and that methodology for estimating the accuracy of such computational models should be developed.

Thereafter, the Protein Model Portal (PMP) (Arnold et al., 2009; Haas et al., 2013) was developed at the Swiss Institute of Bioinformatics (SIB) at the University of Basel as a component of the Structural Biology Knowledgebase (SBKB) (Berman et al., 2009; Gabanyi et al., 2011). Today, the SBKB integrates experimental information provided by the PDB with *in silico* models computed by automated modeling resources. In addition, the PMP provides access to several state-of-the-art model quality assessment services (Schwede et al., 2009). Since 2013, the Model Archive (http://modelarchive.org) resource has also served as a repository for individually generated *in silico* models of macromolecular structures, primarily those described in peer-reviewed publications. Finally, the Model Archive hosts all legacy models that were available from the PDB archive prior to 2006.

Each model in the PMP is assigned a stable, unique accession code (and digital object identifier or DOI) to ensure accurate cross-referencing in publications and other data

repositories. Unlike experimentally determined structural models, *in silico* models are not the product of experimental measurements of a physical sample. They are generated computationally using various molecular modeling methods and underlying assumptions. Examples include comparative modeling, virtual docking of ligand molecules to protein targets, virtual docking of one protein to another, simulations of molecular dynamics and motions, and *de novo* (*ab initio*) protein models.

Effective archival storage of such models depends critically on capturing sufficient detail regarding underlying assumptions, parameters, methodology, and modeling constraints, to allow for assessment and, if necessary, faithful re-computation of the model. It is also essential that these models be accompanied by reliable estimates of uncertainty and, possibly, error. In October 2013, a workshop on "Theoretical Model Archiving, Validation and PDBx/mmCIF Data Exchange Format" (http://www.proteinmodelportal.org/workshop-2013/) was hosted at Rutgers University to launch development of community standards for theoretical model archiving.

2 Integrative / Hybrid Structure Modeling

2.1 Motivation

Samples of many biological macromolecules prove recalcitrant to mainstream structural biology methods (*i.e.*, crystallography, NMR, or 3DEM), because they are not crystallizable, are insoluble, are not of adequate purity, are conformationally heterogeneous, are too large or small, or do not remain intact during the course of the experiment. In such cases, integrative modeling is increasingly being used to compute structural models based on complementary experimental data and theoretical information (**Figures 1 and 2; Table 1**) (Alber et al., 2007; Alber et al., 2008; Sali et al., 2003; Sali et al., 1990; Schneidman-Duhovny et al., 2014; Ward et al., 2013). Structural biology is no stranger to integrative models. Insights into the molecular detail of the B-DNA double helix (Watson and Crick, 1953), the α-helix, and the β-sheet (Pauling et al., 1951) all depended on constructing structural models that encompassed data derived from multiple sources (albeit without the benefit of digital computation). Integrative

structure modeling of today has its origins in attempts to fit X-ray derived substructures into an EM density map of a larger assembly (Rayment et al., 1993); other early examples include models based on NMR and SAS data (Sunnerhagen et al., 1996) together with X-ray and 3DEM data (Olah et al., 1995).

Beyond overcoming sample limitations, the integrative approach has several additional advantages (Alber et al., 2007). First, synergy among the input data minimizes the drawbacks of sparse, noisy, and ambiguous data obtained from compositionally and structurally heterogeneous samples. Each individual piece of data may contain relatively little structural information, but by simultaneously fitting a model to all data derived from independent experiments, the uncertainty of the structures that fit the data can be markedly reduced. Second, the integrative approach can be used to produce all structural models consistent with available data, instead of myopically focusing on just one. Third, comparison of an ensemble of structural models permits estimation of precision and, sometimes, the accuracy of both the experimental data and the model. Fourth, the integrative approach can make structural biologists more efficient by identifying which additional measurements are likely to have the greatest impact on integrative model precision and accuracy. Finally, integrative modeling provides a framework for considering the perturbations of the system that are often required to collect the data; for example, spin labels are required for EPR experiments, membrane proteins are often reconstituted in micelles for NMR spectroscopy, and point mutations or even entire domains are introduced to stabilize preferred conformations for crystallization. While such perturbations complicate structural analysis, integrative modeling may allow us to distinguish biologically relevant states from artifacts of any individual approach. In summary, integrative structure determination maximizes the accuracy, precision, completeness, and efficiency of the structural coverage of biomolecular systems.

2.2 Experimental and computational methods for generating structural information

Input information for integrative modeling can come from various experimental methods, physical theories, and statistical analyses of databases of known structures, biopolymer sequences, and interactions. These methods probe different structural aspects of the system (**Table 1**). In addition to information about average structures, a number of methods provide dynamical insights, which can also be incorporated into integrative modeling procedures (Russel et al., 2009). For example, both NMR spectroscopy and X-ray crystallography provide access to various measures of conformational dynamics; FRET, time-dependent Double Electron-Electron Resonance (DEER) spectroscopies, and even quantitative cross-linking/mass spectrometry (qCLMS) (Fischer et al., 2013) can map distance changes in time; SAXS can provide time-resolved information on the structures and processes with the temporal resolution of a millisecond; molecular dynamics simulations can map the dynamics of an atomic structure up to the millisecond time scale; and High-speed AFM imaging can detect the dynamic live images of single molecules (Ando, 2014).

Table 1. Types of structural data used in integrative modeling. Example methods that are informative about a variety of structural aspects of biomolecular systems are listed.

Structural information	Method
Atomic structures of parts of the studied system	X-ray and neutron crystallography, NMR spectroscopy, 3DEM, comparative modeling, and molecular docking
3D maps and 2D images	Electron microscopy and tomography
Atomic and protein distances	NMR, FRET and other fluorescence techniques, DEER, EPR, and other spectroscopic techniques; chemical crosslinks detected by mass spectrometry and disulfide bonds detected by gel electrophoresis
Binding site mapping	NMR spectroscopy, mutagenesis, FRET
Size, shape, and pairwise atomic distance distributions	SAS
Shape and size	Atomic force microscopy, ion mobility mass spectrometry, fluorescence correlation spectroscopy and fluorescence anisotropy

Component positions	Super-resolution optical microscopy, FRET imaging
Physical proximity	Co-purification, native mass spectrometry, genetic methods, and gene/protein sequence covariance
Solvent accessibility	Footprinting methods, including H/D exchange assessed by mass spectrometry or NMR, and even functional consequences of point mutations
Proximity between different genome segments	Chromosome Conformation Capture and other data
Propensities for different interaction modes	Molecular mechanics force fields, potentials of mean force, statistical potentials, and sequence co-variation

2.3 Approach

All structural characterization approaches correspond to finding models that best fit input information, as judged by use of a scoring function quantifying the difference between the observed data and the data computed from the model. Thus, any information about a structure determination target must always be converted to an explicit structural model through computation. Integrative approaches explicitly combine diverse experimental and theoretical information, with the goal of increasing accuracy, precision, coverage, and efficiency of structure determination. Input information can vary greatly in terms of resolution (*i.e.*, precision, noise, uncertainty), accuracy, and quantity. All structure determination methods are integrative, albeit with differences in degree. At one end of the spectrum, even structure determination using predominantly crystallographic, NMR, or high-resolution single particle EM data also generally requires a molecular mechanics force field description of atomic structure. At the other end of the spectrum, integrative methods rely more evenly on different types of information, often resulting in coarser models with higher uncertainty (**Figure 1**). Examples of such integrative methods include docking of comparative models of subunits into a 3DEM density map of the macromolecular assembly (Lasker et al., 2009); rigid-body model fitting of domain structures experimentally determined by crystallography or NMR into an overall molecular envelope derived from SAS data (Petoukhov and Svergun, 2005); and use of conformational sampling methods with low-density NMR data (Lange et al., 2012; Mueller et al., 2000), chemical cross links (Young et al., 2000), or even chemical shift data alone (Shen et al., 2008). It is not difficult to appreciate how integrative methods

blur distinctions between models based primarily on theoretical considerations and those based primarily on experimental measurements from a physical sample.

The practice of integrative structure determination is iterative, consisting of four stages (**Figure 2**): data gathering; choosing the representation and encoding of all data within a numerical scoring function consisting of spatial restraints; configurational sampling to identify structural models with good scores; and analyzing the models, including quantifying agreement with input spatial restraints and estimating model uncertainty. Input information about the system can be used (i) to select the set of variables that best represent the system (system representation), (ii) to rank the different configurations (scoring function), (iii) to search for good scoring solutions (sampling); and (iv) to further filter good-scoring solutions produced by sampling.

2.4 Types of integrative models

A structural model of a macromolecular assembly is defined by the relative positions and orientations of its components (*e.g.*, atoms, pseudo-atoms, residues, secondary structure elements, domains, subunits, and subcomplexes). While traditional structural biology methods usually produce a single atomistic model, integrative models tend to be more complex in at least four respects. First, a model can be multi-scale, representing different levels of structural detail by a collection of geometrical primitives (*e.g.*, points, spheres, tubes, 3D Gaussians, or probability densities) (Grime and Voth, 2014). Thus, the same part of a system can be described with multiple representations or different parts of a system can be represented differently. An optimal representation facilitates accurate formulation of spatial restraints together with efficient and complete sampling of good-scoring solutions, while retaining sufficient detail (without over fitting) such that the resulting models are maximally useful for subsequent biological analysis (Schneidman-Duhovny et al., 2014). Second, a model can be multi-state, specifying multiple discrete states of the system required to explain the input information (each state may differ in structure and/or composition) (Molnar et al., 2014; Pelikan et al., 2009). Third, a model can also specify the order of states in time and/or transitions between the states. This feature allows representation of a multi-step biological process, a functional cycle (Diez

et al., 2004), a kinetic network (Pirchi et al., 2011), time evolution of a system (*e.g.*, a molecular dynamics trajectory) (Bock et al., 2013), or FRET trajectories; for a comprehensive description of biomolecular function, it is essential to register state lifetimes, characteristic relaxation times, and direct rate constants. Finally, an ensemble of models may be provided to underscore the uncertainty in the input information, with each individual model satisfying the input information within an acceptable threshold (*e.g.*, NMR-derived ensembles currently available in the PDB (Clore and Gronenborn, 1991; Snyder et al., 2005; Snyder et al., 2014) and the ensembles generated from SAXS (Tria et al., 2015)). This aspect of the representation allows us to describe model uncertainty and assess the completeness of input information; such ensembles are distinct from multiple structural states that represent actual variations in the structure, as implied by experimental information that cannot be accounted for by a single representative structure (Schneidman-Duhovny et al., 2014; Schroder, 2015).

3 Task Force Deliberations and Recommendations

3.1 Charge to the Task Force

As previously discussed, virtually every structural model of a macromolecule deposited into the PDB archive is an integrative model that has been derived both from experimental measurements involving a physical sample of a biological macromolecule and prior knowledge of the underlying stereochemistry. It is, therefore, difficult if not impossible to draw a definitive line between structural models on the spectrum ranging from very well-determined ultra-high resolution crystallographic structures (>40 experimental observations per non-hydrogen atom in the crystallographic asymmetric unit) and structural models based on a single or even no experimental observation. A major reason for concern about the data-to-parameter ratio is that satisfactory general methods to assess the precision and accuracy of each model do not currently exist.

Among structural biologists, a healthy debate is underway as to how best classify structural models and partition them among distinct, publicly accessible model/data repositories. In principle, there are several possibilities, including (i) a single mega

archive that serves as the repository for every type of structural model and data; (ii) independent, free-standing repositories that house distinct types of models and data; and (iii) a federated system of inter-operating repositories with "spheres of influence" based on community consensus that archive models and data.

To address some of the challenges ahead and make recommendations about how best to proceed, the community stakeholders who assembled at the October 2014 meeting of the wwPDB Hybrid / Integrative Methods Validation Task Force were divided into three discussion groups, each tasked with considering a series of related questions. What experimental data (beyond crystallography, NMR, and 3DEM) should be archived? Where and how should it be validated? What kinds of non-atomistic models can we expect and how should they be validated? What are the criteria for deciding where models should be archived? How should non-atomistic and mixed atomistic/non-atomistic models be archived? Should there be a separate archive for integrative (mixed) models (and data)? Should we establish a federated system of data and model archives to support integrative structural biology? The three breakout groups were asked to address these questions, report back with their findings, and make recommendations for the future. Each group independently approached the same set of questions. At the close of the meeting, the teams converged to compare notes, identify areas of commonality and diversity, and determine how best to move forward. The resulting consensus is reflected in this document.

3.2 Recommendations

Recommendation 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.**

Ideally, structural models of any kind, derived by any method, should be archived.

Models are of greatest value when they are independently tested, potentially improved, and serve to further our understanding of how the function of a biological system is determined by its 3D structure(s). Therefore, models must be freely available to the

research community, with the necessary annotations, including demonstrated reproducibility of the modeling process. Information concerning all aspects of a model should be deposited, including input data, corresponding spatial restraints, output models, and protocols used to convert input data into models. In addition to the input experimental data, the archival deposition should specify or include theoretically derived restraints used to compute the model (*e.g.*, a statistical potential and a molecular mechanics force field). In practice, frequently used data types (*e.g.*, distance information) should be prioritized for early attention. Uncertainty in the input data needs to be well documented; some data uncertainty estimates may require modeling (*e.g.*, Bayesian error estimates (Rieping et al., 2005)). Consistency between input data and the structural model should be documented as part of model validation.

Each expert community should drive decisions as to how much raw data, processed data, and metadata to deposit, subject to the minimal requirement that the spatial restraints used for modeling must be derivable from the deposited information. Attention needs to be paid to annotating measurement conditions, such as temperature (Fenwick et al., 2014), sample concentration, environmental conditions (*e.g.*, buffer), construct definition, and identification of all assembly components, all of which can significantly influence the experimental outcome. Cost benefit analyses should be used to help guide which data should be archived. As much data as practical should be deposited, to facilitate model validation, future improvements of the model, and methods development (*e.g.*, benchmarking sets). Of particular importance will be availability of some raw data to help drive improvement of data processing methods and for use by methods developers, who are often not generating the experimental data themselves.

Recommendation 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.**

Model representation should allow for as many types of “structural” models as possible, thereby encouraging collaboration among developers of integrative modeling software (Russel et al., 2012). At a minimum, the model representation should allow encoding of an ensemble of multi-scale multi-state time-ordered models (**Section 2.4)**. Uncertainty of

the model coordinates should be tightly associated with the model coordinates in the model representation. Any model resident within an archive should be "self-contained" to facilitate utilization (*e.g.*, for visualization). A common representation and format for models are useful for reasons of software interoperability; particle-based representations/primitives will be prioritized. Non-particle-based model representations (*e.g.*, continuum representations) merit further consideration by appropriate community stakeholders.

Recommendation 3: **Procedures for estimating the uncertainty of integrative models should be developed, validated, and adopted.**

Assessment of both an integrative model and the information on which it is based is of critical importance for guiding subsequent use of the model. For atomistic models, extant standard validation criteria from X-ray crystallography should be used. Beyond this test, validation of integrative models and data is a major research challenge that must be addressed and overcome. The following represent promising considerations (Alber et al., 2007; Schneidman-Duhovny et al., 2014): convergence of conformational sampling, fit of the model to the input information, test for clashes between geometrical primitives comprising the model, precision of the ensemble of solutions (visualized with, for example, ribbon plots), cross-validation and statistical bootstrapping based on available data, tests based on data determined after the model was computed, and sensitivity analysis of the model to input data. Bayesian approaches may be particularly well suited to describe model uncertainty by computing posterior model densities from a forward model, noise model, and priors (Muschielok et al., 2008; Rieping et al., 2005). Tools for visualizing model validation should be developed.

Communities generating data used in integrative modeling should agree on the standard set of descriptors for data quality, as has been done for crystallography, NMR, and 3DEM.

Recommendation 4: **A federated system of model and data archives should be created.**

Integrative models can be based on a broad array of different experimental and computational techniques. While the specific spatial restraints implied by the data and used to construct an integrative model should be deposited with the model itself, the underlying experimental data often contain much richer information. This information should be captured in a federated system of domain-specific model and data archives. These individual member archives should be developed by community experts, based on method-specific standards for data archiving and validation. A federated system of model and data archives implies the need for a seamless exchange of information between independent archives. This seamless exchange requires a common dictionary of terms, agreed data formats, persistent and stable data object identifiers, and close synchronization of policies and procedures. Federated model and data archives need to develop efficient methods for data exchange to allow for transparent data access across the enterprise.

A single interface for the deposition of all data and models into the federated system is highly desirable. Such an interface would greatly facilitate the task of the depositor, and, thereby, maximize compliance with deposition standards and requirements. In addition, reliance on a single entry point will help to ensure consistency across the federation at the time of deposition. Following successful deposition, individual data sets can be transferred to member databases for data curation and archiving if domain-specific databases exist. There should also be provision for collecting unstructured information in a “data commons”, as proposed by the data science initiative at the NIH (Margolis et al., 2014).

Access to the contents of the federated database through a single portal is also most desirable, to facilitate dissemination of data, models, and experimental/computational protocols.

Of particular importance for integrative modeling will be the option to modify or update any aspect of the modeling procedure, for example, by adding new data. The federated archive should allow versioning for each deposited model. Such capabilities will facilitate the cycle of experiment and modeling, and accelerate production of more accurate, precise, and complete models (Russel et al., 2012).

Recommendation 5: **Publication standards for integrative models should be established.**

Over the past decade, the wwPDB organization has worked with relevant scientific journals to help establish publication standards for structural models coming from crystallography, NMR spectroscopy, and 3DEM. Community standards now include requiring authors to make their validation reports available to reviewers and editors. Through the International Union of Crystallography (IUCr) Small Angle Scattering and Journals Commissions, the SAS community developed and agreed upon publication guidelines for structural modeling of biomolecules therefrom (Jacques et al., 2012). A set of standards for publishing integrative models should be developed along similar lines.

3.3 Implementation

Implementation of ***Recommendation 1*** poses a host of cultural and technical challenges. Experimentalists and modelers need to provide the data, models, and protocols, thus at least partly addressing increasing concerns regarding reproducibility of scientific results. From a technical perspective, interoperating data dictionaries for all methods need to be created. In addition, potential storage bottlenecks need to be addressed.

Implementation of ***Recommendations 2*** **and** ***3*** will require significant research as to how best to represent and validate the many different kinds of integrative models. In addition, the community will need to agree on a common set of standards that are sufficiently mutable to allow for future innovation. Efforts such as the “Cryo-EM Modeling Challenge” may facilitate this process (ncmi.bcm.edu/challenge).

Implementation of ***Recommendation 4*** will require agreement on a common data exchange system among member repositories. Based on past accomplishments, the wwPDB is well positioned to play a leadership role in establishing the proposed federated system, including provision of common deposition and access interfaces. The wwPDB should begin this process by providing training and advice on data archiving and curation to contributing domin-specific member repositories.

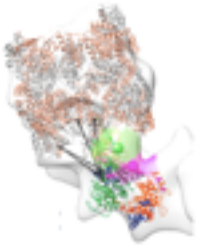
Implementation of ***Recommendation 5*** will require continued work with the journals that publish structural models of biological macromolecules.

Significant resources will be required to implement these recommendations, including grants for research, infrastructure, and workshops. These efforts are international by their very nature and will require funding from multiple public and private sources in North America, Europe, Asia, and elsewhere.

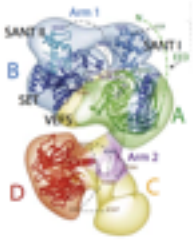
Acknowledgements

The workshop was supported by funding to PDBe by Wellcome Trust 088944; RCSB PDB by NSF DBI 1338415; PDBj by JST-NBDC; BMRB by NLM P41 LM05799; EMDataBank by NIH GM079429, and tax-deductible donations made to the wwPDB Foundation in support of wwPDB outreach activities.

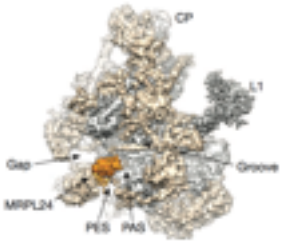
Figures



INO80

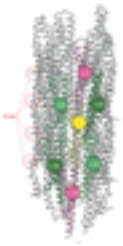


PRC2

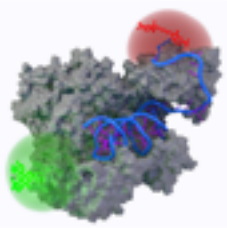


39S ribosome subunit

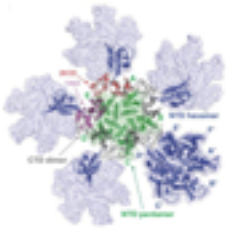
RNA pol II



type III secretion system needle



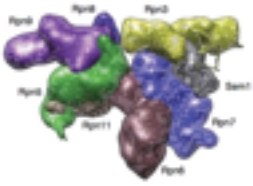
HIV-1 reverse transcriptase:DNA



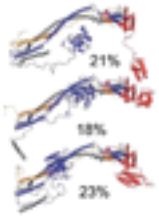
HIV capsid

genome architecture

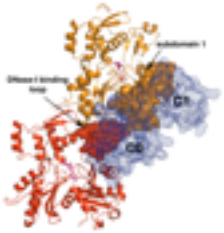
α-globin gene domain



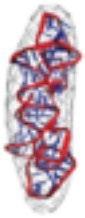
proteosomal lid



ESCRT-I complex

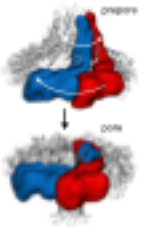


actin – myosin binding protein C



$[\Psi^{CD}]_2$

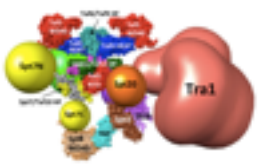
KaiB-KaiC complex



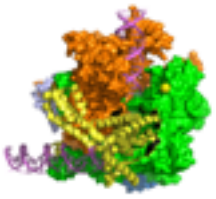
pore-forming toxin aerolysin

pleurotolysin

desulfurase-Isu-frataxin



SAGA transcriptional coactivator



bacterial RNA polymerase-promoter open complex

RNA ribosome-binding element

RNA polymerase II - transcription factor IIF

Figure 1. Examples of recently determined integrative structures. The molecular architecture of INO80 was determined with a 17 Å resolution cryo-EM map and 212 intra-protein and 116 inter-protein crosslinks (Russel et al., 2009). The molecular architecture of Polycomb Repressive Complex 2 (PRC2) was determined with a 21 Å resolution negative-stain EM map and ~60 intra-protein and inter-protein crosslinks (Shi et al., 2014). The molecular architecture of the large subunit of the mammalian mitochondrial ribosome (39S) was determined with a 4.9 Å resolution cryo-EM map and ~70 inter-protein cross-links (Ward et al., 2013). The molecular architecture of the RNA polymerase II transcription pre-initiation complex was determined with a 16 Å resolution cryo-EM map plus 157 intra-protein and 109 inter-protein crosslinks (Alber et al., 2008). The atomic model of Type III secretion system needle was determined with a 19.5 Å resolution cryo-EM map and solid-state NMR data (Loquet et al., 2012). Molecular architecture of the productive HIV-1 reverse transcriptase:DNA primer-template complex in the open educt state was determined by FRET positioning and screening (FPS) using a known HIV-1 reverse transcriptase structure (Kalinin et al., 2012). The structure of HIV-1 capsid protein was determined using residual dipolar couplings (RDC) and SAXS data (Deshmukh et al., 2013). The human genome architecture was determined based on tethered chromosome conformation capture (TCC) and population-based modeling (Kalhor et al., 2012). The structural model of α-globin gene domain was determined based on Chromosome Conformation Capture Carbon Copy (5C) experiments (Bau et al., 2011). The molecular architecture of the proteosomal lid was determined using native MS and 28 cross-links (Politis et al., 2014). Atomic resolution conformations of ESCRT-I complex were determined with SAXS, double electron-electron transfer (DEER), and FRET (Boura et al., 2011). Integrative model of actin and the cardiac myosin binding protein C was developed from a combination of crystallographic and NMR structures of subunits and domains, with positions and orientations optimized against SAXS and SANS data to reveal information about the quaternary interactions (Whitten et al., 2008). The ensemble of $[\Psi^{CD}]_2$ NMR structures were fitted into the averaged cryo-electron tomography map (Miyazaki et al., 2010). Integrative model of the cyanobacterial circadian timing KaiB-KaiC complex was obtained based on H/D exchange and collision cross section data from MS (Snijder et al., 2014). The prepore

and pore conformations of the pore-forming toxin aerolysin were obtained combining cryo-EM data and molecular dynamics simulations (Degiacomi and Dal Peraro, 2013; Degiacomi et al., 2013). Segment of a pleurotolysin pore map (~11 Å resolution) with an ensemble of conformations showing the trajectory of β-sheet opening during pore formation (Lukoyanova et al., 2015). A SAXS-based rigid body model of a ternary complex of the iron-sulphur cluster assembly proteins desulfurase (orange) and scaffold protein Isu (blue) with bacterial orthologue of frataxin (yellow) was validated by NMR chemical shifts and mutagenesis (Prischi et al., 2010). The molecular architecture of the SAGA transcription coactivator complex was determined with 199 inter- and 240 intra-subunit crosslinks, several comparative models based on X-ray crystal structures, and a TFIID core EM map at 31 Å resolution (Han et al., 2014). Structural Organization of the bacterial (*T. aquaticus*) RNA polymerase-promoter open complex obtained by FRET (Mekler et al., 2002), subsequently validated by a crystal structure (Zhang et al., 2012). The RNA ribosome-binding element from turnip crinkle virus genome, determined using NMR, SAXS, and EM data (Gong et al., 2015). The molecular architecture of the complex between RNA polymerase II and transcription factor IIF was determined using a deposited crystal structure of RNA polymerase II, homology models of some domains in transcription factor IIF, as well as 95 intra-protein and 129 inter-protein cross links (Chen et al., 2010).

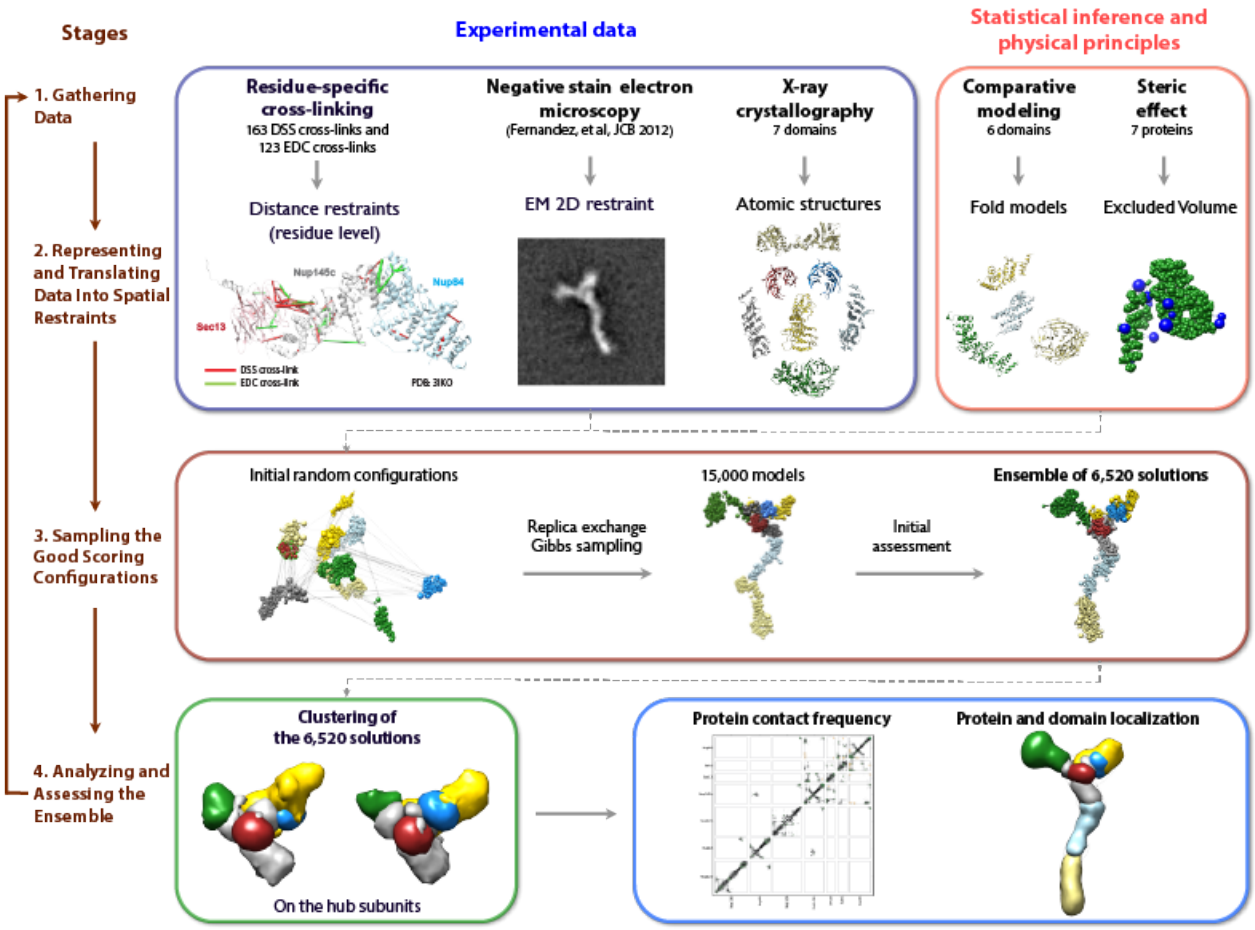


Figure 2. The four stages of integrative structure determination. The approach is illustrated by its application to the heptameric Nup84 subcomplex of the Nuclear Pore Complex (Shi et al., 2014).

References

Alber, F., Dokudovskaya, S., Veenhoff, L., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B.*, et al.* (2007). Determining the architectures of macromolecular assemblies. Nature *450*, 683-694.

Alber, F., Forster, F., Korkin, D., Topf, M., and Sali, A. (2008). Integrating diverse data for structure determination of macromolecular assemblies. Annu. Rev. Biochem. *77*, 443-477.

Ando, T. (2014). High-speed AFM imaging. Curr Opin Struct Biol *28*, 63-68.

Arnold, K., Kiefer, F., Kopp, J., Battey, J.N., Podvinec, M., Westbrook, J.D., Berman, H.M., Bordoli, L., and Schwede, T. (2009). The Protein Model Portal. J. Struct. Funct. Genomics *10*, 1-8.

Bau, D., Sanyal, A., Lajoie, B.R., Capriotti, E., Byron, M., Lawrence, J.B., Dekker, J., and Marti-Renom, M.A. (2011). The three-dimensional folding of the alpha-globin gene domain reveals formation of chromatin globules. Nat. Struct. Mol. Biol. *18*, 107-114.

Berman, H.M., Burley, S.K., Chiu, W., Sali, A., Adzhubei, A., Bourne, P.E., Bryant, S.H., Dunbrack Jr., R.L., Fidelis, K., Frank, J.*, et al.* (2006). Outcome of a workshop on archiving structural models of biological macromolecules. Structure *14*, 1211-1217.

Berman, H.M., Henrick, K., and Nakamura, H. (2003). Announcing the worldwide Protein Data Bank. Nat. Struct. Biol. *10*, 980.

Berman, H.M., Westbrook, J.D., Gabanyi, M.J., Tao, W., Shah, R., Kouranov, A., Schwede, T., Arnold, K., Kiefer, F., Bordoli, L.*, et al.* (2009). The protein structure initiative structural genomics knowledgebase. Nucleic Acids Res. *37*, D365-368.

Bock, L.V., Blau, C., Schroder, G.F., Davydov, II, Fischer, N., Stark, H., Rodnina, M.V., Vaiana, A.C., and Grubmuller, H. (2013). Energy barriers and driving forces in tRNA translocation through the ribosome. Nat Struct Mol Biol *20*, 1390-1396.

Boura, E., Rozycki, B., Herrick, D.Z., Chung, H.S., Vecer, J., Eaton, W.A., Cafiso, D.S., Hummer, G., and Hurley, J.H. (2011). Solution structure of the ESCRT-I complex by small-angle X-ray scattering, EPR, and FRET spectroscopy. Proc. Natl. Acad. Sci. USA *108*, 9437-9442.

Chen, Z.A., Jawhari, A., Fischer, L., Buchen, C., Tahir, S., Kamenski, T., Rasmussen, M., Lariviere, L., Bukowski-Wills, J.C., Nilges, M.*, et al.* (2010). Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry. The EMBO journal *29*, 717-726.

Clore, G.M., and Gronenborn, A.M. (1991). Structures of larger proteins in solution: three- and four-dimensional heteronuclear NMR spectroscopy. Science *252*, 1390-1399.

Degiacomi, M.T., and Dal Peraro, M. (2013). Macromolecular symmetric assembly prediction using swarm intelligence dynamic modeling. Structure *21*, 1097-1106.

Degiacomi, M.T., Iacovache, I., Pernot, L., Chami, M., Kudryashev, M., Stahlberg, H., van der Goot, F.G., and Dal Peraro, M. (2013). Molecular assembly of the aerolysin pore reveals a swirling membrane-insertion mechanism. Nat Chem Biol *9*, 623-629.

Deshmukh, L., Schwieters, C.D., Grishaev, A., Ghirlando, R., Baber, J.L., and Clore, G.M. (2013). Structure and dynamics of full-length HIV-1 capsid protein in solution. J. Am. Chem. Soc. *135*, 16133-16147.

Diez, M., Zimmermann, B., Börsch, M., König, M., Schweinberger, E., Steigmiller, S., Reuter, R., Felekyan, S., Kudryavtsev, V., Seidel, C.A.M., and Gräber, P. (2004). Proton-powered subunit rotation in single membrane-bound F0 F1 -ATP synthase. Nature Structural & Molecular Biology *11*, 135-141.

Fenwick, R.B., van den Bedem, H., Fraser, J.S., and Wright, P.E. (2014). Integrated description of protein dynamics from room-temperature X-ray crystallography and NMR. Proceedings of the National Academy of Sciences of the United States of America *111*, E445-E454.

Fischer, L., Chen, Z.A., and Rappsilber, J. (2013). Quantitative cross-linking/mass spectrometry using isotope-labelled cross-linkers. Journal of proteomics *88*, 120-128.

Gabanyi, M.J., Adams, P.D., Arnold, K., Bordoli, L., Carter, L.G., Flippen-Andersen, J., Gifford, L., Haas, J., Kouranov, A., McLaughlin, W.A.*, et al.* (2011). The Structural Biology Knowledgebase: a portal to protein structures, sequences, functions, and methods. J. Struct. Funct. Genomics *12*, 45-54.

Gong, Z., Schwieters, C.D., and Tang, C. (2015). Conjoined Use of EM and NMR in RNA Structure Refinement. Plos One 10 *e0120445*.

Grime, J.M.A., and Voth, G.A. (2014). Highly Scalable and Memory Efficient Ultra-coarse-grained Molecular Dynamics Simulations. J. Chem. Theor. Comp. *10*, 423–431.

Haas, J., Roth, S., Arnold, K., Kiefer, F., Schmidt, T., Bordoli, L., and Schwede, T. (2013). The Protein Model Portal--a comprehensive resource for protein structure and model information. Database : the journal of biological databases and curation *2013*, bat031.

Han, Y., Luo, J., Ranish, J., and Hahn, S. (2014). Architecture of the Saccharomyces cerevisiae SAGA transcription coactivator complex. The EMBO journal *33*, 2534-2546.

Henderson, R., Sali, A., Baker, M.L., Carragher, B., Devkota, B., Downing, K.H., Egelman, E.H., Feng, Z., Frank, J., Grigorieff, N.*, et al.* (2012). Outcome of the first electron microscopy validation task force meeting. Structure *20*, 205-214.

Jacques, D.A., Guss, J.M., Svergun, D.I., and Trewhella, J. (2012). Publication guidelines for structural modelling of small-angle scattering data from biomolecules in solution. Acta Crystallogr. D: Biol. Crystallogr. *68*, 620-626.

Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F., and Chen, L. (2012). Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat. Biotechnol. *30*, 90-98.

Kalinin, S., Peulen, T., Sindbert, S., Rothwell, P.J., Berger, S., Restle, T., Goody, R.S., Gohlke, H., and Seidel, C.A.M. (2012). A toolkit and benchmark study for FRET-restrained high-precision structural modeling. Nature Methods *9*, 1218-1225.

Lange, O.F., Rossi, P., Sgourakis, N.G., Song, Y., Lee, H.W., Aramini, J.M., Ertekin, A., Xiao, R., Acton, T.B., Montelione, G.T., and Baker, D. (2012). Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. Proc Natl Acad Sci U S A *109*, 10873-10878.

Lasker, K., Topf, M., Sali, A., and Wolfson, H.J. (2009). Inferential optimization for simultaneous fitting of multiple components into a CryoEM map of their assembly. J Mol Biol *388*, 180-194.

Lawson, C.L., Baker, M.L., Best, C., Bi, C., Dougherty, M., Feng, P., van Ginkel, G., Devkota, B., Lagerstedt, I., Ludtke, S.J.*, et al.* (2011). EMDataBank.org: unified data resource for CryoEM. Nucleic Acids Res. *39*, D456-464.

Loquet, A., Sgourakis, N.G., Gupta, R., Giller, K., Riedel, D., Goosmann, C., Griesinger, C., Kolbe, M., Baker, D., Becker, S., and Lange, A. (2012). Atomic model of the type III secretion system needle. Nature *486*, 276-279.

Lukoyanova, N., Kondos, S.C., Farabella, I., Law, R.H., Reboul, C.F., Caradoc-Davies, T.T., Spicer, B.A., Kleifeld, O., Traore, D.A., Ekkel, S.M.*, et al.* (2015). Conformational changes during pore formation by the perforin-related protein pleurotolysin. PLoS Biol *13*, e1002049.

Malfois, M., and Svergun, D. (2000). sasCIF: an extension of core Crystallographic Information File for SAS. J. App. Cryst. *33*, 812-816.

Margolis, R., Derr, L., Dunn, M., Huerta, M., Larkin, J., Sheehan, J., Guyer, M., and Green, E.D. (2014). The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. J Am Med Inform Assoc *21*, 957-958.

Markley, J.L., Ulrich, E.L., Berman, H.M., Henrick, K., Nakamura, H., and Akutsu, H. (2008). BioMagResBank (BMRB) as a partner in the Worldwide Protein Data Bank (wwPDB): new policies affecting biomolecular NMR depositions. J. Biomol. NMR *40*, 153-155.

Mekler, V., Kortkhonjia, E., Mukhopadhyay, J., Knight, J., Revyakin, A., Kapanidis, A.N., Niu, W., Ebright, Y.W., Levy, R., and Ebright, R.H. (2002). Structural organization of bacterial RNA polymerase holoenzyme and the RNA polymerase-promoter open complex. Cell *108*, 599-614.

Miyazaki, Y., Irobalieva, R.N., Tolbert, B.S., Smalls-Mantey, A., Iyalla, K., Loeliger, K., D'Souza, V., Khant, H., Schmid, M.F., Garcia, E.L.*, et al.* (2010). Structure of a conserved retroviral RNA packaging element by NMR spectroscopy and cryo-electron tomography. J Mol Biol *404*, 751-772.

Molnar, K.S., Bonomi, M., Pellarin, R., Clinthorne, G.D., Gonzalez, G., Goldberg, S.D., Goulian, M., Sali, A., and DeGrado, W.F. (2014). Cys-scanning disulfide crosslinking and bayesian modeling probe the transmembrane signaling mechanism of the histidine kinase, PhoQ. Structure *22*, 1239-1251.

Montelione, G.T., Nilges, M., Bax, A., Güntert, P., Herrmann, T., Markley, J.L., Richardson, J., Schwieters, C., Vuister, G.W., Vranken, W., and Wishart, D. (2013). Recommendations of the wwPDB NMR Structure Validation Task Force. Structure *21*, 1563-1570.

Mueller, G.A., Choy, W.Y., Yang, D., Forman-Kay, J.D., Venters, R.A., and Kay, L.E. (2000). Global folds of proteins with low densities of NOEs using residual dipolar couplings: application to the 370-residue maltodextrin-binding protein. J Mol Biol *300*, 197-212.

Muschielok, A., Andrecka, J., Jawhari, A., Bruckner, F., Cramer, P., and Michaelis, J. (2008). A nano-positioning system for macromolecular structural analysis. Nature Methods *5*, 965-971.

Olah, G.A., Gray, D.M., Gray, C.W., Kergil, D.L., Sosnick, T.R., Mark, B.L., Vaughan, M.R., and Trewhella, J. (1995). Structures of fd gene 5 protein.nucleic acid complexes: a combined solution scattering and electron microscopy study. J. Mol. Biol. *249*, 576-594.

Pauling, L., Corey, R.B., and Branson, H.R. (1951). The structures of proteins. Proc. Natl. Acad. Sci. USA *37*, 205.

Pelikan, M., Hura, G.L., and Hammel, M. (2009). Structure and flexibility within proteins as identified through small angle X-ray scattering. Gen Physiol Biophys *28*, 174-189.

Petoukhov, M.V., and Svergun, D.I. (2005). Global rigid body modeling of macromolecular complexes against small-angle scattering data. Biophys J *89*, 1237-1250.

Pirchi, M., Ziv, G., Riven, I., Cohen, S.S., Zohar, N., Barak, Y., and Haran, G. (2011). Single-molecule fluorescence spectroscopy maps the folding landscape of a large protein. Nature Communications *2*.

Politis, A., Stengel, F., Hall, Z., Hernandez, H., Leitner, A., Walzthoeni, T., Robinson, C.V., and Aebersold, R. (2014). A mass spectrometry-based hybrid method for structural modeling of protein complexes. Nat. Methods *11*, 403-406.

Prischi, F., Konarev, P.V., Iannuzzi, C., Pastore, C., Adinolfi, S., Martin, S.R., Svergun, D.I., and Pastore, A. (2010). Structural bases for the interaction of frataxin with the central components of iron-sulphur cluster assembly. Nat Commun *1*, 95.

Protein Data Bank (1971). Protein Data Bank. Nature New Biol *233*, 223.

Rayment, I., Holden, H.M., Whittaker, M., Yohn, C.B., Lorenz, M., Holmes, K.C., and Milligan, R.A. (1993). Structure of the actin-myosin complex and its implications for muscle contraction. Science *261*, 58-65.

Read, R.J., Adams, P.D., Arendall, W.B., III, Brunger, A.T., Emsley, P., Joosten, R.P., Kleywegt, G.J., Krissinel, E.B., Lutteke, T., Otwinowski, Z*., et al.* (2011). A new generation of crystallographic validation tools for the Protein Data Bank. Structure *19*, 1395-1412.

Rieping, W., Habeck, M., and Nilges, M. (2005). Inferential structure determination. Science *309*, 303-306.

Russel, D., Lasker, K., Phillips, J., Schneidman-Duhovny, D., Velazquez-Muriel, J., and Sali, A. (2009). The structural dynamics of macromolecular processes. Curr. Opin. Cell Biol. *21*, 97-108.

Russel, D., Lasker, K., Webb, B., Velazquez-Muriel, J., Tjioe, E., Schneidman-Duhovny, D., Peterson, B., and Sali, A. (2012). Putting the pieces together: integrative structure determination of macromolecular assemblies. PLoS Biol *10*, e1001244.

Sali, A., Glaeser, R., Earnest, T., and Baumeister, W. (2003). From words to literature in structural proteomics. Nature *422*, 216-225.

Sali, A., Overington, J.P., Johnson, M.S., and Blundell, T.L. (1990). From comparisons of protein sequences and structures to protein modelling and design. Trends in biochemical sciences *15*, 235-240.

Schneidman-Duhovny, D., Pellarin, R., and Sali, A. (2014). Uncertainty in integrative structural modeling. Curr. Opin. Struct. Biol. *28*, 96-104.

Schroder, G.F. (2015). Hybrid methods for macromolecular structure determination: experiment with expectations. Curr Opin Struct Biol *31*, 20-27.

Schwede, T., Sali, A., Honig, B., Levitt, M., Berman, H.M., Jones, D., Brenner, S.E., Burley, S.K., Das, R., Dokholyan, N.V*., et al.* (2009). Outcome of a workshop on applications of protein models in biomedical research. Structure *17*, 151-159.

Shen, Y., Lange, O., Delaglio, F., Rossi, P., Aramini, J.M., Liu, G., Eletsky, A., Wu, Y., Singarapu, K.K., Lemak, A*., et al.* (2008). Consistent blind protein structure generation from NMR chemical shift data. Proc Natl Acad Sci U S A *105*, 4685-4690.

Shi, Y., Fernandez-Martinez, J., Tjioe, E., Pellarin, R., Kim, S.J., Williams, R., Schneidman, D., Sali, A., Rout, M.P., and Chait, B.T. (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.

Snijder, J., Burnley, R.J., Wiegard, A., Melquiond, A.S.J., Bonvin, A.M.J.J., Axmann, I.M., and Heck, A.J.R. (2014). Insight into cyanobacterial circadian timing from structural details of the KaiB-KaiC interaction. Proc. Natl. Acad. Sci. USA *111*, 1379-1383

Snyder, D.A., Bhattacharya, A., Huang, Y.J., and Montelione, G.T. (2005). Assessing precision and accuracy of protein structures derived from NMR data. Proteins *59*, 655-661.

Snyder, D.A., Grullon, J., Huang, Y.J., Tejero, R., and Montelione, G.T. (2014). The expanded FindCore method for identification of a core atom set for assessment of protein structure prediction. Proteins *82 Suppl 2*, 219-230.

Sunnerhagen, M., Olah, G.A., Stenflo, J., Forsen, S., Drakenberg, T., and Trewhella, J. (1996). The relative orientation of Gla and EGF domains in coagulation factor X is altered by Ca2+ binding to the first EGF domain. A combined NMR-small angle X-ray scattering study. Biochemistry *35*, 11547-11559.

Trewhella, J., Hendrickson, W.A., Kleywegt, G.J., Sali, A., Sato, M., Schwede, T., Svergun, D.I., Tainer, J.A., Westbrook, J., and Berman, H.M. (2013a). Report of the wwPDB Small-Angle Scattering Task Force: data requirements for biomolecular modeling and the PDB. Structure *21*, 875-881.

Trewhella, J., Hendrickson, W.A., Sato, M., Schwede, T., Svergun, D., Tainer, J.A., Westbrook, J., Kleywegt, G.J., and Berman, H.M. (2013b). Meeting Report of the wwPDB Small-Angle Scattering Task Force: Data Requirements for Biomolecular Modeling and the PDB Structure *21*, 875-881.

Tria, G., Mertens, H.D.T., Kachala, M., and Svergun, D.I. (2015). Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering. IUCrJ *2*, 207-217.

Valentini, E., Kikhney, A.G., Previtali, G., Jeffries, C.M., and Svergun, D.I. (2015). SASBDB, a repository for biological small-angle scattering data. Nucleic Acids Res. *43*, D357-363.

Ward, A., Sali, A., and Wilson, I. (2013). Integrative structural biology. Science *339*, 913-915.

Watson, J.D., and Crick, F.H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature *171*, 737-738.

Whitten, A.E., Jeffries, C.M., Harris, S.P., and Trewhella, J. (2008). Cardiac myosin-binding protein C decorates F-actin: implications for cardiac function. Proc. Natl. Acad. Sci. USA *105*, 18360-18365.

Young, M.M., Tang, N., Hempel, J.C., Oshiro, C.M., Taylor, E.W., Kuntz, I.D., Gibson, B.W., and Dollinger, G. (2000). High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. Proc Natl Acad Sci U S A *97*, 5802-5806.

Zhang, Y., Feng, Y., Chatterjee, S., Tuske, S., Ho, M.X., Arnold, E., and Ebright, R.H. (2012). Structural Basis of Transcription Initiation. Science *338*, 1076-1080.