Outcome of the first wwPDB/CCDC/D3R Ligand Validation Workshop

Paul D. Adams[1], Kathleen Aertgeerts[2], Cary Bauer[3], Jeffrey A. Bell[4], Helen M. Berman[5,6], Talapady N. Bhat[7], Jeff Blaney[8], Evan Bolton[9], Gerard Bricogne[10], David Brown[11], Stephen K. Burley[5,6,12,*], David A. Case[6], Kirk L. Clark[13], Tom Darden[14], Paul Emsley[15], Victoria A. Feher[16,*], Zukang Feng[5,6], Colin R. Groom[17,*], Seth F. Harris[8], Jorg Hendle[18], Thomas Holder[4], Andrzej Joachimiak[19], Gerard J. Kleywegt[20,*], Tobias Krojer[21], Joseph Marcotrigiano[6,22], Alan E. Mark[23], John L. Markley[24,*], Matthew Miller[22], Wladek Minor[25], Gaetano T. Montelione[22,26], Garib Murshudov[15], Atsushi Nakagawa[27], Haruki Nakamura[27,*], Anthony Nicholls[14], Marc Nicklaus[28], Robert T. Nolte[29], Anil K. Padyana[30], Catherine E. Peishoff[29], Susan Pieniazek[31], Randy J. Read[32], Chenghua Shao[5], Steven Sheriff[33], Oliver Smart[20], Stephen Soisson[34], John Spurlino[35], Terry Stouch[36], Radka Svobodova[37], Wolfram Tempel[38], Thomas C. Terwilliger[39], Dale Tronrud[40], Sameer Velankar[20], Suzanna Ward[17], Gregory L. Warren[14], John D. Westbrook[5,6], Pamela Williams[41], Huanwang Yang[5,6], and Jasmine Young[5,6]

Author Affiliations

[1] Molecular Biophysics & Integrated Bioimaging Division, Lawrence Berkeley Laboratory, Berkeley, CA 94720-8235, USA, and Department of Bioengineering, UC Berkeley, Berkeley, CA 94720, USA

[2] DART NeuroScience, LLC, San Diego, CA 92131, USA

[3] Bruker AXS, Inc., Madison, WI 53711, USA

[4] Schrödinger, Inc., New York, NY 10036, USA

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

[6] Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

[7] Biosystems and Biomaterials Division, NIST, Gaithersburg, MD 20899, USA

[8] Genentech, Inc., South San Francisco, CA 94080, USA

[9] National Center for Biotechnology Information, U.S. National Library of Medicine, Bethesda MD, 20894, USA

[10] Global Phasing Ltd., Cambridge CB3 0AX, UK

[11] School of Biosciences, University of Kent, Canterbury CT2 7NH, UK and Charles River Ltd., Structural Biology and Biophysics, Cambridge CB10 1XL, UK

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

[13] Novartis Institutes for BioMedical Research, Cambridge, MA 02139, USA

[14] OpenEye Scientific, Cambridge, MA 02142, USA

[15] MRC Laboratory of Molecular Biology, Cambridge CB2 0QH, UK

[16] Drug Design Data Resource and Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, CA 92093, USA

[17] Cambridge Crystallographic Data Centre, Cambridge CB2 1EZ, UK

[18] Structural Biology, Lilly Biotechnology Center, San Diego, CA 92121, USA

[19] Structural Biology Center, Biosciences, Argonne National Laboratory, Argonne, IL 60439, USA

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

[21] Structural Genomics Consortium, University of Oxford, Oxford OX3 7DQ, UK

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

[23] School of Chemistry & Molecular Biosciences, University of Queensland, St Lucia, QLD 4072, Australia

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

[25] Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA 22908, USA

[26] Department of Molecular Biology and Biochemistry, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

[27] Protein Data Bank Japan, Institute for Protein Research, Osaka University, Osaka 565-0871, Japan

[28] Computer-Aided Drug Design Group, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Frederick, MD 21702, USA

[29] GlaxoSmithKline, Collegeville, PA 19426, USA

[30] Agios Pharmaceuticals, Inc., Cambridge, MA 02139, USA

[31] Bristol-Myers Squibb Research and Development, Pennington, NJ 08534, USA

[32] Department of Haematology, Cambridge Institute for Medical Research, University of Cambridge, Cambridge CB2 0XY, UK

[33] Bristol-Myers Squibb Research and Development, Princeton, NJ 08543, USA

[34] Merck Research Laboratories, West Point, PA 19486, USA

[35] Janssen Pharmaceuticals, Inc., Spring House, PA 19002, USA

[36] Science For Solutions, LLC, West Windsor, NJ 08550, USA

[37] CEITEC-Central European Institute of Technology and National Centre for Biomolecular Research, Masaryk University Brno, 625 00 Brno, Czech Republic

[38] Structural Genomics Consortium, University of Toronto, Toronto, ON M5G 1L7, Canada

[39] Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

[40] Department of Biochemistry and Biophysics, Oregon State University, Corvallis, OR 97331, USA

[41] Astex Pharmaceuticals, Cambridge CB4 0QA, UK

* Address correspondence to these authors (SKB: sburley@rcsb.org; VAF: vfeher@ucsd.edu; CRG: groom@ccdc.cam.ac.uk; GJK: gerard@ebi.ac.uk; JLM: markley@biochem.wisc.edu; HN: harukin@protein.osaka-u.ac.jp)

All registered Ligand Validation Workshop attendees are listed as authors in alphabetical order.

Running Title: Ligand Structure Validation White Paper

Keywords: ligand structure validation, ligand structure, drug target structure, co-crystal structure, protein-ligand complex, Protein Data Bank

Summary

Crystallographic studies of ligands bound to biological macromolecules (proteins and nucleic acids) represent an important source of information concerning drug-target interactions, providing atomic level insights into the physical chemistry of complex formation between macromolecules and ligands. Of the more than 115,000 entries extant in the Protein Data Bank archive, ~75% include at least one non-polymeric ligand. Ligand geometrical and stereochemical quality, the suitability of ligand models for *in silico* drug discovery/design, and the goodness-of-fit of ligand models to electron density maps vary widely across the archive. We describe the proceedings and conclusions from the first Worldwide Protein Data Bank/Cambridge Crystallographic Data Centre/Drug Design Data Resource (wwPDB/CCDC/D3R) Ligand Validation Workshop held at the Research Collaboratory for Structural Bioinformatics at Rutgers University on July 30-31, 2015. Experts in protein crystallography from academe and industry came together with non-profit and for-profit software providers for crystallography and with experts in computational chemistry and data archiving to discuss and make recommendations on best practices, as framed by a series of questions central to structural studies of macromolecule-ligand complexes. What data concerning bound ligands should be archived in the Protein Data Bank? How should the ligands be best represented? How should structural models of macromolecule-ligand complexes be validated? What supplementary information should accompany publications of structural studies of biological macromolecules? Consensus recommendations on best practices developed in response to each of these questions are provided, together with some details regarding implementation. Important issues addressed but not resolved at the workshop are also enumerated.

1 Background

The Worldwide Protein Data Bank (wwPDB; wwpdb.org), the Cambridge Crystallographic Data Center (CCDC; www.ccdc.cam.ac.uk), and the Drug Design Data Resource (D3R; www.drugdesigndata.org) co-organized a Ligand Validation Workshop on July 30-31 2015 at Rutgers University. The workshop brought together academic and industrial protein crystallographers, providers of software for crystallography, computational chemists, and experts in data archiving. More than 50 participants from more than 40 organizations discussed and made recommendations on best practices for structural studies of macromolecule-ligand complexes and archiving of the resulting information.

1.1 Protein Data Bank (PDB) and Historical Context for the Workshop

The Protein Data Bank (PDB) was established in 1971 with just seven X-ray crystallographic structures of proteins as the first open access digital resource in the biological sciences (Protein Data Bank, 1971). In February 2016, some 44 years later, this *sui generis* global archive holds more than 115,000 experimentally determined 3D structural models of biological macromolecules and their complexes with a wide variety of ligands. In addition, descriptions of the chemistry of biopolymers and ligands are collected, as are metadata describing sample preparation, experimental methodology, structural model building and refinement statistics, literature references, *etc*. PDB data are made freely available without restrictions on usage. The vast majority of data in the PDB (~90%) come from X-ray, neutron, and combined X-ray/neutron 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 best to curate structural models and primary experimental data from X-ray, NMR, and 3DEM. Over the past decade, the Worldwide Protein Data Bank (wwPDB; the global organization responsible for managing the PDB archive; wwpdb.org) (Berman et al., 2003) has formed expert, method-specific Validation Task Forces to identify which experimental data and metadata from each structure determination method should be archived and how these data and the atomic level structural models therefrom should be validated. Initially, the wwPDB X-ray Validation Task Force (VTF) made recommendations on how to best validate crystallographic data (Read et al., 2011). These initial recommendations have been implemented as a validation pipeline used within the wwPDB Deposition and

Annotation (D&A) system. A wwPDB Validation Report accompanies every PDB deposition (ftp://ftp.wwpdb.org/pub/pdb/validation_reports/). Preliminary recommendations have also been made by wwPDB VTFs for NMR (Montelione et al., 2013) and 3DEM (Henderson et al., 2012). Implementation of NMR and 3DEM VTF recommendations within the wwPDB D&A validation pipeline is currently underway. It is anticipated that additional validation measures will be implemented within the wwPDB D&A system as new methods are developed and more experience is gained with existing procedures.

1.2 Crystallographic Data in the PDB

For structural models determined *via* X-ray, neutron, and combined X ray/neutron crystallography methods, together with those determined using electron diffraction from 2D crystals, deposition of experimental data (i.e., diffracted intensities or structure factor amplitudes) into the PDB has been mandatory since 2008 (http://www.wwpdb.org/news/news?year=2007#29-November-2007). Validation against deposited structure factor amplitudes is carried out using procedures recommended by the wwPDB X-ray VTF (Read et al., 2011). wwPDB Validation Reports include graphical summaries of the quality of the overall structural model and residue-specific features. Detailed assessments of various aspects of the structural model, such as agreement with experimental data and chemical expectations, are also provided. In the near future, unmerged intensities will also be collected during PDB deposition, thereby enabling additional validation.

1.3 Chemical Component Dictionary

The Chemical Component Dictionary (CCD) was originally developed (Feng et al., 2004) to provide a more expressive alternative to the early PDB ligand descriptions, which were based purely on atom connectivity records. The CCD embraced the data representation for chemical components developed for the Macromolecular

Crystallographic Information Framework or mmCIF data dictionary (Fitzgerald et al., 2005). Following a major wwPDB undertaking to standardize nomenclature concluded in 2007 (Henrick et al., 2008), the global organization adopted a common dictionary of chemical definitions. The current Chemical Component Dictionary (Westbrook et al., 2015) is an extended reference file describing all polymer components and small molecules found in PDB archival entries. This dictionary contains detailed chemical descriptions for standard and modified amino acids/nucleotides, small molecule ligands, and solvent/solute molecules. Each chemical definition includes descriptions of chemical properties, such as stereochemical assignments, chemical descriptors [SMILES (Weininger, 1988), InChI, and InChIKeys (Heller et al., 2013)], and systematic chemical names. A set of atomic model coordinates from a selected experimental entry and a computed set of ideal atomic coordinates are provided for each entry in the CCD. Hydrogen atoms are computationally added to the experimental coordinates and unobserved heavy atoms, such as leaving groups specified by Depositors, are added to the ideal coordinates if they are not explicitly modeled in the experimental entry. Computed ideal coordinates are obtained from the software tools Corina (Gasteiger et al., 1990) or OpenEye/Omega (Hawkins et al., 2010). Cahn-Ingold-Prelog (CIP) stereochemical assignments (Cahn et al., 1966) and aromatic annotations are documented for each atom present in each CCD entry. The dictionary is organized by the 3-character alphanumeric code that the wwPDB assigns to each chemical component, and updated with each weekly release of the PDB archive (Sen et al., 2014).

A related PDB archive chemical reference dictionary is the Biologically Interesting molecule Reference Dictionary (BIRD) (Dutta et al., 2014; Young et al., 2013), which contains information about peptide-like antibiotic and inhibitor molecules present in the PDB archive. BIRD entries include molecular weight and chemical formula, polymer sequence and connectivity, descriptions of structural features and functional classification, natural source, and external references to corresponding UniProt (UniProt Consortium, 2015) or Norine (Caboche et al., 2008) reference sequences.

A BIRD molecule may be represented in a PDB archival entry as a polymer with sequence information or as a single ligand with chemical information. The preferred representation is specified in the BIRD file, with a representative PDB ID code. All PDB entries containing the same BIRD molecule or its analogue(s) are represented uniformly. An important feature of BIRD is to provide dual representation–both sequence and chemical information is provided, regardless of whether the molecule is represented as a polymer or as a ligand in the PDB archive.

1.4 Current Validation of Macromolecule-Ligand Complexes

The initial recommendations of the wwPDB X-ray VTF (Read et al., 2011) have been implemented in a software pipeline (Gore et al., 2012) embedded within the wwPDB D&A system. Officially watermarked wwPDB Validation Reports are provided to PDB contributors at the time of deposition. An increasing number of journals require that these reports accompany manuscripts reporting structural studies of biological macromolecules. Structural biologists can obtain a similar report using the wwPDB Validation Server (http://wwpdb-validation.wwpdb.org/) prior to deposition. For ligands, the wwPDB Validation Report includes both geometrical and model fit diagnostic information. Bond lengths and angles, acyclic torsion angles, and ring systems are assessed (Bruno et al., 2004) by comparison with preferred molecular geometries derived from high-quality, small-molecule structures in the Cambridge Structural Database (CSD) (Groom and Allen, 2014).

A Z-score is calculated for every bond length and bond angle in each ligand. Individual bond lengths or bond angles with a Z-score magnitude>2 are highlighted. The root-mean-square value of the Z-scores (RMSZ) of bond lengths (or angles) is calculated for the entire molecule. The EDS software (Kleywegt et al., 2004) is used to calculate density maps from deposited atomic coordinates and experimental data, which are compared to idealized map density with the difference reported as a Real Space R-value (RSR). This analysis is performed on an individual ligand basis. A Local Ligand Density Fit (LLDF; for a description of this calculation see http://www.wwpdb.org/validation/ValidationPDFNotes.html) then compares the RSR of a molecule to the mean and standard deviation of RSR for the neighbouring polymeric standard amino acids and/or nucleotides. Minimum, median, 95th percentile and maximum atomic displacement parameters (isotropic B-values) for all atoms in the molecule are presented along with the number of atoms in the ligand molecule with occupancies of less than 0.9.

1.5 Quality of Macromolecule-Ligand Complexes in the PDB

Of the more than 115,000 entries in the PDB today, some 76% include at least one non-polymeric small molecule ligand. While some of these ligands are almost certainly crystallization solutes, many were intentionally included in the experimental sample or co-purified with the structure determination target and are of considerable biological, biochemical, or medical interest. Recently published review articles assessing the quality of macromolecule-ligand complexes in the PDB can be usefully broken down into three categories, including

i) assessments of geometrical and stereochemical quality (Liebeschuetz et al., 2012; Sehnal et al., 2015; Zheng et al., 2014);

ii) the suitability of ligand models for *in silico* drug discovery/design (Davis et al., 2008; Smart and Bricogne, 2015; Warren et al., 2012); and

iii) general issues with ligand atomic model fit to the electron density map (Malde and Mark, 2011; Pozharski et al., 2013; Sitzmann et al., 2012; Weichenberger et al., 2013).

It has been emphasized by some that a non-negligible number of structural biologists err by interpreting weak density map features as indicating the presence of a bound small molecule that has been included in the crystallization process or soaked into a pre-formed crystal [e.g., (Rupp, 2010)]. Current validation and journal refereeing policies and practices do not always prevent such cases from entering either the PDB archive or the scientific literature. Other explanations of problems with macromolecule-ligand complexes in the PDB include the following:

i) some ligands undergo chemical transformation upon binding, which may not be reflected in the atomic model used for refinement;
ii) the ligand may be present, but was modeled incorrectly or refinement was performed with incorrect restraint targets; and
iii) the ligand does bind, but the experimentalist does not provide an accurate chemical descriptor.

1.6 Workshop Format and Charge to Participants

Catherine E. Peishoff (GlaxoSmithKline) gave the keynote address emphasizing the value of atomic level structural information for pharmaceutical discovery research and the growing opportunities for pre-competitive engagement and data sharing. She stressed the importance of data and structural model quality and the need for data archived in the PDB to be fit for purpose. Finally, she suggested a move away from the historical view of the PDB as an archival database towards an increased emphasis on data provisioning, which would shift the focus from any single structure to the structures as a collective. Increased attention to data standards, governance, and quality, together with improving tools to analyze the collective data, will significantly help researchers derive insight from this valuable scientific resource.

Stephen K. Burley (RCSB Protein Data Bank) and Gerard J. Kleywegt (Protein Data Bank in Europe) then introduced the workshop rationale/objectives and charged the participants with dividing among smaller breakout groups and addressing five questions regarding best practices for macromolecule-ligand complex data deposition and validation and journal editorial, refereeing, and publication practices. Breakout group members were selected on the bases of interest and expertise as follows: Group A, Academic and Industrial Crystallographers; Group B, Crystallographic Software Specialists; Group C, Computational Chemistry Software Specialists; and Group D, Academic and Industrial Crystallographers. After lengthy and lively discussions, the four breakout groups reconvened to report their findings and develop consensus recommendations. Each group independently approached the same set of questions.

2 Workshop Deliberations and Recommendations

2.1 Charge to the Workshop

To address some of the myriad challenges facing PDB Depositors and Users and Editors and Referees of scientific journals that publish the results of structural studies of macromolecule-ligand complexes, the community stakeholders assembled at Rutgers considered the following five questions:

1) What are current best practices for selecting an initial ligand atomic model(s) for co-crystal structure refinement against diffraction data?

2) What are current best practices for validating the ligand(s) coming from such a co-crystal structure refinement?

3) What new data pertaining to co-crystal structures should be required for PDB depositions going forward?

4) What information should accompany journal submissions reporting co-crystal structure determinations? What supplementary materials should accompany publication of co-crystal structure determinations?

5) What do you recommend be done with existing co-crystal structures in the PDB archive?

Towards the close of the meeting, the groups reconvened to compare findings, identify areas of commonality and divergence, and determine how best to move forward. This document reflects the resulting consensus.

2.2 Workshop Recommendations

Recommended Best Practices for PDB Archive Deposition of Co-crystal Structure Data:

Depositors should

1. Provide unambiguous chemical definitions for ligands present in the crystal mother liquor and in the refined structural model, including hydrogen atoms and covalent modifications.

2. Provide the geometry of the starting model of the refined ligand(s), ligand-related refinement restraints, and their provenance.

3. Use the PDBx/mmCIF dictionary _atom_site.calc_flag to identify non-experimentally modeled atoms. Non-experimentally modeled atoms, for the purposes of this recommendation, are defined as those atoms whose positions are not adequately localized by experimental data (e.g., electron density map) to be assigned (x,y,z) positional coordinates, but whose presence is deduced by chemical knowledge of the crystal content and other information. This flag will usually be applicable to the hydrogen atom records for ligands. It is intended for use as an alternative to zero occupancy, which would be a less accurate indicator of the status of these atoms.

4. Provide the Fourier coefficients of the density map(s) used for ligand(s) structure interpretation.

5. Identify
   a) any ligand that is a focus of the study, where appropriate;
   b) any other biologically important ligand(s);
   c) adventitiously bound ligand(s) (i.e., co-purified) and ligands added for
   experimental convenience (e.g., crystallization additives or cryo-protectants); and
   d) the experimental method (crystal soaking *versus* co-crystallization) for (a) and
   (b).

6. As applicable, communicate other experimental findings, judgment calls, and perceived ambiguities regarding tautomers and protonation states of ligands not determined conclusively from the crystallographic data and chemical environment of the ligand by either (a) using the existing alternate conformation mechanism with partial occupancies or (b) providing the chemical descriptions recommended in Item 1 above.

7. Where appropriate, include comments explaining outliers, etc. identified in the wwPDB Validation Report.

Recommended Best Practices for wwPDB Validation of Co-crystal Data:

Building on the framework of the current wwPDB Validation Report, the following new items should be included:

1. Informative images of ligand pose(s) plus nearby density map features using Fourier coefficients endorsed by the wwPDB X-ray VTF [e.g., 2m|Fo|-D|Fc|, m|Fo|-D|Fc|, and omit map (Bhat, 1988; Bhat and Cohen, 1984)] and those provided by the Depositor. The presentation style in the Buster Report tool (Smart and Bricogne, 2015) exemplifies the diagnostic utility of such representations (Figure 1).

2. Stick-figure representations of ligand(s) with non-hydrogen atom labels annotated with geometric validation findings.

3. Identification of atoms modeled but not interpreted from density maps.

4. Quality assessment metrics for each study compound and biologically important ligand(s).

5. Identification of ligands capable of tautomerism or alternate protonation states within the pH range typical of protein crystals, nominally 4-10.

6. The wwPDB D&A Validation pipeline should be described in full in peer-reviewed publications and continue to be publicly available for use in improving models prior to PDB archive deposition. The reference data used to calculate quality metrics/percentile scores should also continue to be publicly available. All the details describing the wwPDB Validation pipeline should be made available so that it can be implemented in an external environment. Specifically, details related to wwPDB Validation pipeline script(s), versions of the publicly available and commercial programs used therein, and input parameters and any other details necessary for reproducibility should be made public as soon as possible.

Recommendations regarding Editorial/Refereeing/Publication Standards for Co-crystal Structure Publications:

Journals should

1. Require submission of officially watermarked PDF wwPDB Validation Reports as Supplementary Materials accompanying manuscripts describing macromolecular structure determinations.

2. Ensure that at least one of the Referees selected for manuscript review has the technical expertise to evaluate in full the content of the wwPDB Validation Report.

2.3 Response of the wwPDB X-ray Validation Task Force

Following the conclusion of the Workshop, the recommendations outlined herein were presented to the membership of the wwPDB X-ray Validation Task Force (Chair: Randy Read, Cambridge University) when the group reconvened at the European Bioinformatics Institute in November 2015. The recommendations received strong support from the Task Force. The wwPDB partner organizations (RCSB PDB, PDBj, PDBe, and BMRB) are currently developing an implementation plan for recommendations relating to data requirements and updates of the PDBx/mmCIF dictionary and will finalize the plan in due course with the benefit of further advice from the Task Force and the PDBX/mmCIF Working Group (Chair: Paul Adams, Lawrence Berkeley Laboratory).

2.4 Implementation Details

Implementation of the recommendations regarding additional archival content will require extension of the PDBx/mmCIF dictionary to capture details of the starting ligand model and the Depositors' identification of the role of the ligand in each study. The PDBx/mmCIF Working Group (Chair: Paul Adams, Lawrence Berkeley Laboratory) is currently working on developing deposition standards for ligand refinement restraints and delivery of additional supporting data in the form of density map Fourier coefficients and unmerged intensities. Further extensions of the PDBx/mmCIF dictionary can be made as needed. With the requisite PDBx/mmCIF dictionary items in place, the wwPDB D&A system can be modified to ensure efficient capture of these new data during deposition.

An enhanced version of the wwPDB Validation Report will furnish the recommended depictions for ligand fits to map density and the annotated stick-figure models, with geometrical, stereochemical, and absence annotations. Development of a summary indicator of ligand quality for inclusion within the wwPDB Validation Report summary graphic requires additional research.

The wwPDB Validation Report also provides a convenient vehicle for delivering the recommended depictions of ligand density map to improve publication practices. Some scientific journals already require that wwPDB Validation Reports accompany structure manuscripts. Further community lobbying of Editors is needed to expand the number of journals requiring submission of the wwPDB Validation Report. Finally, it is incumbent on the scientific community that experts continue to undertake rigorous review of manuscripts describing structural studies of macromolecule-ligand complexes.

Strong sentiments expressed both in the literature [e.g., (Terwilliger and Bricogne, 2014)] and during the workshop favored revision of the current wwPDB policy requiring issuance of new PDB ID codes following update of deposited atomic coordinates. Indeed, some Depositors report being reluctant to update atomic coordinates, because issuance of the new PDB ID code is thought to weaken the connection between the revised PDB archival entry and prior publications describing the structure. It was agreed that the wwPDB leadership, in consultation with the wwPDB Advisory Committee, should come to closure on the matter of versioning of atomic coordinates and other archival data as soon as possible.

Binding of ligands to macromolecules can also be studied using NMR spectroscopy. Members of the wwPDB NMR VTF present at the workshop volunteered the services of their task force to develop recommendations regarding data deposition and validation standards for structural models of macromolecule-ligand complexes determined by NMR.

2.5 Issues Addressed but Not Resolved at the Workshop

Workshop participants discussed three additional topics without reaching consensus.

First, some participants strongly advocated mandatory journal submission of processed diffraction data and atomic coordinates to accompany manuscripts describing crystallographic studies of biological macromolecules. This practice is the norm for small-molecule crystallography publications. With the benefit of full and frank discussion, it was recognized that author sensitivities regarding providing primary data and atomic coordinates in advance of publication to reviewers, who may also be competitors, precluded consensus on this matter. The wwPDB leadership in consultation with the wwPDB Advisory Committee will revisit this issue.

Second, some participants strongly advocated mandatory PDB deposition of all-atom structural models, including computed positions of hydrogen atoms (properly identified with the _atom_site.calc_flag). As inclusion of explicit hydrogen atoms will impact the entire PDB archive, it was agreed that

i) technical recommendations on this front should be made by the wwPDB X-ray Validation Task Force; and

ii) wwPDB leadership, in consultation with the wwPDB Advisory Committee, should make further policy recommendations as necessary.

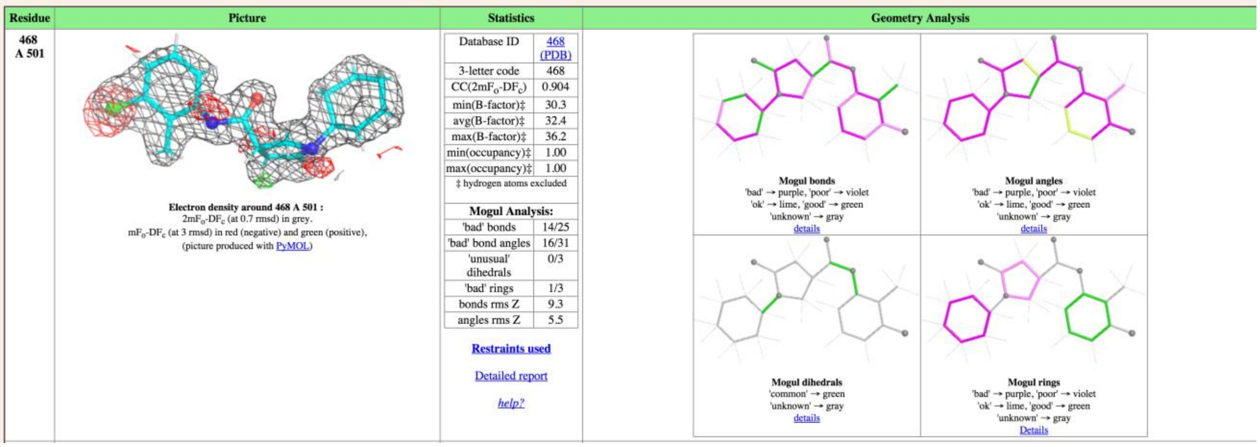
Finally, workshop participants identified a number of challenges that will come to the fore once enhanced validation of macromolecule-ligand complexes already archived in the PDB is concluded and updated wwPDB Validation Reports are made publicly available for every entry. Simply put, what should be done with existing PDB entries found wanting by the validation procedures recommended herein?

Workshop participants believe that the majority of Depositors would be motivated to correct entries identified as not meeting minimal standards for enhanced ligand validation. However, it was also recognized that, over time, increasing numbers of Depositors would not be in a position to make corrections. To ensure the integrity of the database, workshop participants propose that, following a reasonable interval for self-correction, community experts could be mobilized to apply targeted corrections to any remaining PDB archival entries with poor validation outcomes, particularly for bound ligands of significant biological and/or medical interest.

Acknowledgements

The workshop was supported by funding to RCSB PDB by the National Science Foundation (DBI 1338415); PDBe by the Wellcome Trust (104948); PDBj by JST-NBDC; BMRB by the National Institute of General Medical Sciences (GM109046); D3R by the National Institute of General Medical Sciences (GM111528); registration fees from industrial participants; and tax-deductible donations to the wwPDB Foundation by the Genentech Foundation and the Bristol-Myers Squibb Foundation.

Figure



Residue	Picture	Statistics	Geometry Analysis
468 A 501	Electron density around 468 A 501 : $2mF_o$-DF_c (at 0.7 rmsd) in grey. mF_o-DF_c (at 3 rmsd) in red (negative) and green (positive), (picture produced with PyMOL)	see below	see below

Database ID	468 (PDB)
3-letter code	468
CC($2mF_o$-DF_c)	0.904
min(B-factor)‡	30.3
avg(B-factor)‡	32.4
max(B-factor)‡	36.2
min(occupancy)‡	1.00
max(occupancy)‡	1.00
‡ hydrogen atoms excluded	

Mogul Analysis:	
'bad' bonds	14/25
'bad' bond angles	16/31
'unusual' dihedrals	0/3
'bad' rings	1/3
bonds rms Z	9.3
angles rms Z	5.5

Restraints used

Detailed report

help?

Mogul bonds
'bad' → purple, 'poor' → violet
'ok' → lime, 'good' → green
'unknown' → gray
details

Mogul angles
'bad' → purple, 'poor' → violet
'ok' → lime, 'good' → green
'unknown' → gray
details

Mogul dihedrals
'common' → green
'unknown' → gray
details

Mogul rings
'bad' → purple, 'poor' → violet
'ok' → lime, 'good' → green
'unknown' → gray
Details

Figure 1. Example highlighting the value of presenting ligand electron density model fit and geometrical analysis from CCDC Mogul from the Global Phasing Buster Report (PDB ID: 2H7P, later superseded by entry 4TZT (He et al., 2006); CCD ID: 468).

References

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

Bhat, T.N. (1988). Calculation of an OMIT map. J. Appl. Cryst. *21*, 279-281.

Bhat, T.N., and Cohen, G.H. (1984). OMITMAP - An Electron-Density Map Suitable For The Examination Of Errors In a Macromolecular Model. J. Appl. Cryst. *17*, 244-248.

Bruno, I.J., Cole, J.C., Kessler, M., Luo, J., Motherwell, W.D., Purkis, L.H., Smith, B.R., Taylor, R., Cooper, R.I., Harris, S.E.*, et al.* (2004). Retrieval of crystallographically-derived molecular geometry information. J. Chem. Inf. Comput. Sci. *44*, 2133-2144.

Caboche, S., Pupin, M., Leclere, V., Fontaine, A., Jacques, P., and Kucherov, G. (2008). NORINE: a database of nonribosomal peptides. Nucleic Acids Res. *36*, D326-331.

Cahn, R.S., Ingold C.K., and Prelog V. (1966). Specification of molecular chirality. Angew. Chem. Int. Edition *5*, 385–415.

Davis, A.M., St-Gallay, S.A., and Kleywegt, G.J. (2008). Limitations and lessons in the use of X-ray structural information in drug design. Drug Discovery Today *13*, 831-841.

Dutta, S., Dimitropoulos, D., Feng, Z., Persikova, I., Sen, S., Shao, C., Westbrook, J., Young, J., Zhuravleva, M.A., Kleywegt, G.J.*, et al.* (2014). Improving the representation of peptide-like inhibitor and antibiotic molecules in the Protein Data Bank. Biopolymers *101*, 659-668.

Feng, Z., Chen, L., Maddula, H., Akcan, O., Oughtred, R., Berman, H.M., and Westbrook, J. (2004). Ligand Depot: a data warehouse for ligands bound to macromolecules. Bioinformatics *20*, 2153-2155.

Fitzgerald, P.M.D., Westbrook, J.D., Bourne, P.E., McMahon, B., Watenpaugh, K.D., and Berman, H.M. (2005). 4.5 Macromolecular dictionary (mmCIF). In International Tables for Crystallography G. Definition and exchange of crystallographic data, S.R. Hall, and B. McMahon, eds. (Dordrecht, The Netherlands: Springer), pp. 295-443.

Gasteiger, J., Rudolph, C., and Sadowski, J. (1990). Automatic generation of 3D-atomic coordinates for organic molecules. Tetrahedron Comp. Method. *3*, 537-547.

Gore, S., Velankar, S., and Kleywegt, G.J. (2012). Implementing an X-ray validation pipeline for the Protein Data Bank. Acta Crystallogr. D Biol. Crystallogr. *68*, 478-483.

Groom, C.R., and Allen, F.H. (2014). The Cambridge Structural Database in retrospect and prospect. Angew. Chem. Int. Edition *53*, 662-671.

Hawkins, P.C., Skillman, A.G., Warren, G.L., Ellingson, B.A., and Stahl, M.T. (2010). Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. J. Chem. Inf. Model. *50*, 572-584.

He, X., Alian, A., Stroud, R., and Ortiz de Montellano, P.R. (2006). Pyrrolidine carboxamides as a novel class of inhibitors of enoyl acyl carrier protein reductase from Mycobacterium tuberculosis. J. Med. Chem. *49*, 6308-6323.

Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D., and Pletnev, I. (2013). InChI - the worldwide chemical structure identifier standard. J Cheminform *5*, 7.

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.

Henrick, K., Feng, Z., Bluhm, W.F., Dimitropoulos, D., Doreleijers, J.F., Dutta, S., Flippen-Anderson, J.L., Ionides, J., Kamada, C., Krissinel, E.*, et al.* (2008). Remediation of the protein data bank archive. Nucleic Acids Res. *36*, D426-433.

Kleywegt, G.J., Harris, M.R., Zou, J.Y., Taylor, T.C., Wahlby, A., and Jones, T.A. (2004). The Uppsala Electron-Density Server. Acta Crystallogr. D Biol. Crystallogr. *60*, 2240-2249.

Liebeschuetz, J., Hennemann, J., Olsson, T., and Groom, C.R. (2012). The good, the bad and the twisted: a survey of ligand geometry in protein crystal structures. J. Comput. Aided. Mol. Des. *26*, 169-183.

Malde, A.K., and Mark, A.E. (2011). Challenges in the determination of the binding modes of non-standard ligands in X-ray crystal complexes. J. Comput. Aided. Mol. Des *25*, 1-12.

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

Pozharski, E., Weichenberger, C.X., and Rupp, B. (2013). Techniques, tools and best practices for ligand electron-density analysis and results from their application to deposited crystal structures. Acta Crystallogr. D Biol. Crystallogr. *69*, 150-167.

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

Read, R.J., Adams, P.D., Arendall, W.B., 3rd, 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.

Rupp, B. (2010). Scientific inquiry, inference and critical reasoning in the macromolecular crystallography curriculum. J. Appl. Cryst. *43*, 1242-1249.

Sehnal, D., Svobodova Varekova, R., Pravda, L., Ionescu, C.M., Geidl, S., Horsky, V., Jaiswal, D., Wimmerova, M., and Koca, J. (2015). ValidatorDB: database of up-to-date validation results for ligands and non-standard residues from the Protein Data Bank. Nucleic Acids Res. *43*, D369-375.

Sen, S., Young, J., Berrisford, J.M., Chen, M., Conroy, M.J., Dutta, S., Di Costanzo, L., Gao, G., Ghosh, S., Hudson, B.P.*, et al.* (2014). Small molecule annotation for the Protein Data Bank. Database *2014*, bau116.

Sitzmann, M., Weidlich, I.E., Filippov, I.V., Liao, C., Peach, M.L., Ihlenfeldt, W.D., Karki, R.G., Borodina, Y.V., Cachau, R.E., and Nicklaus, M.C. (2012). PDB ligand conformational energies calculated quantum-mechanically. J. Chem. Inf. Model. *52*, 739-756.

Smart, O., and Bricogne, G. (2015). Achieving High Quality Ligand Chemistry in Protein-Ligand Crystal Structures for Drug Design. In Multifaceted Roles of Crystallography in Modern Drug Discovery, P.D. Scapin G, Arnold E., ed. (Netherlands: Springer), pp. 165–181.

Terwilliger, T.C., and Bricogne, G. (2014). Continuous mutual improvement of macromolecular structure models in the PDB and of X-ray crystallographic software: the dual role of deposited experimental data. Acta Crystallogr. D Biol. Crystallogr. *70*, 2533-2543.

UniProt Consortium. (2015). UniProt: a hub for protein information. Nucleic Acids Res. *43*, D204-212.

Warren, G.L., Do, T.D., Kelley, B.P., Nicholls, A., and Warren, S.D. (2012). Essential considerations for using protein-ligand structures in drug discovery. Drug Discovery Today *17*, 1270-1281.

Weichenberger, C.X., Pozharski, E., and Rupp, B. (2013). Visualizing ligand molecules in Twilight electron density. Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun. *69*, 195-200.

Weininger, D. (1988). SMILES 1. Introduction and encoding rules. J. Chem. Inf. Comput. Sci. *28*, 31-36.

Westbrook, J.D., Shao, C., Feng, Z., Zhuravleva, M., Velankar, S., and Young, J. (2015). The chemical component dictionary: complete descriptions of constituent molecules in experimentally determined 3D macromolecules in the Protein Data Bank. Bioinformatics *31*, 1274-1278.

Young, J.Y., Feng, Z., Dimitropoulos, D., Sala, R., Westbrook, J., Zhuravleva, M., Shao, C., Quesada, M., Peisach, E., and Berman, H.M. (2013). Chemical annotation of small and peptide-like molecules at the Protein Data Bank. Database *2013*, bat079.

Zheng, H., Chordia, M.D., Cooper, D.R., Chruszcz, M., Muller, P., Sheldrick, G.M., and Minor, W. (2014). Validation of metal-binding sites in macromolecular structures with the CheckMyMetal web server. Nature Protocols *9*, 156-170.