Worldwide Protein Data Bank Advisory Committee Meeting

October 10th 2014

WORLDWIDE PDB
PROTEIN DATA BANK

wwpdb.org

Agenda

9:00 am	Welcome	Janet Thornton
9:10	RCSB PDB Leadership Transition	Helen Berman
9:15	Overview	Gerard Kleywegt
10:00	Common Tool for D&A	Stephen Burley
10:45	*Break*	
11:00	NMR	John Markley
11:30	Outreach	Haruki Nakamura
12:00 pm	*Lunch*	
1:00	Looking Ahead, Questions & Discussion	Gerard Kleywegt
2:30	*Group photo, followed by break*	
3:00	Executive Session & Feedback	
4:00	*Adjourn*	

Welcome

Janet Thornton

WORLDWIDE PDB
PROTEIN DATA BANK

wwpdb.org

RCSB PDB Leadership Transition

Helen Berman

WORLDWIDE PDB
PROTEIN DATA BANK

wwpdb.org

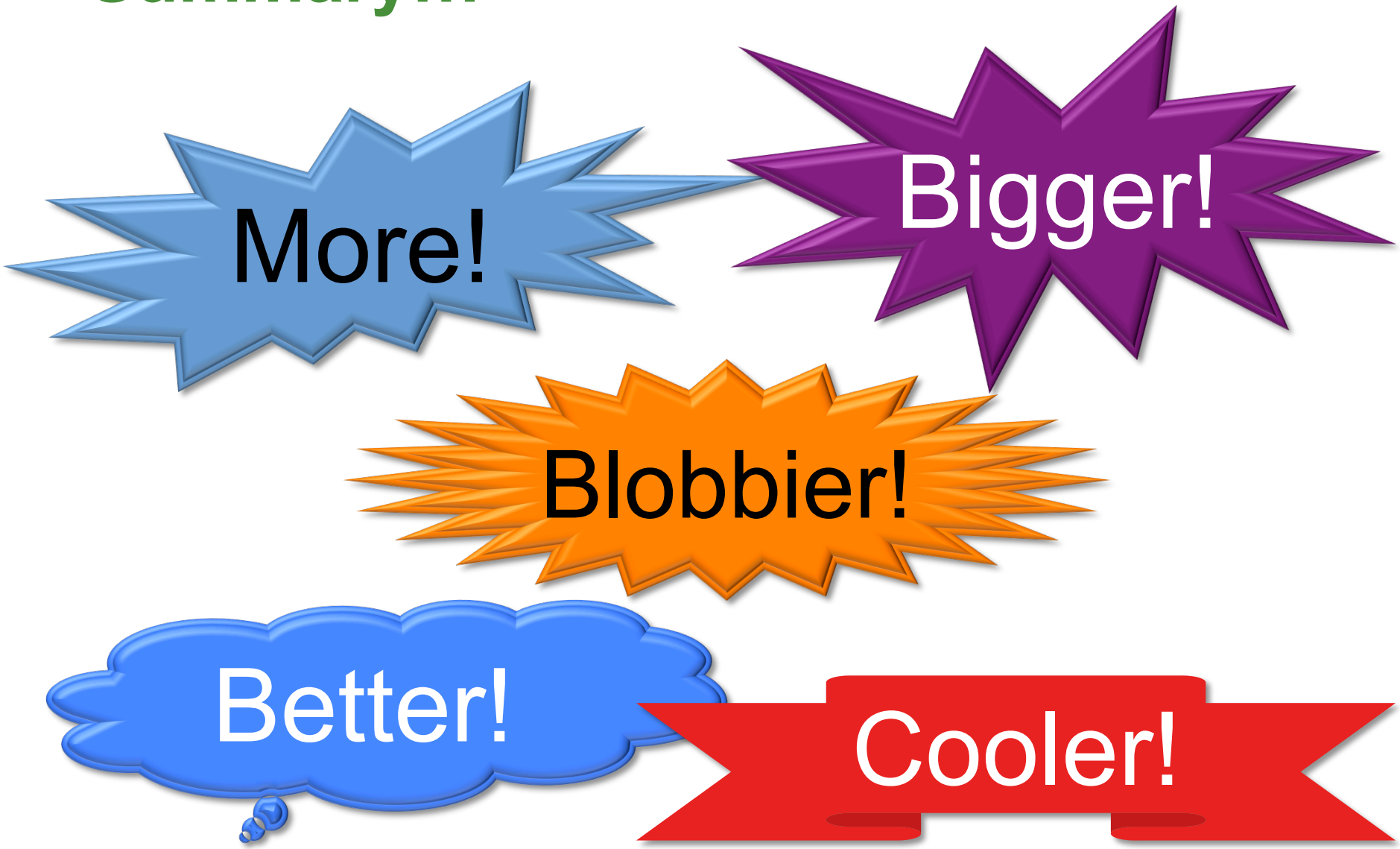
Overview

Gerard Kleywegt

WORLDWIDE PDB
PROTEIN DATA BANK

wwpdb.org

Summary…



wwPDB
September 2013 - October 2014

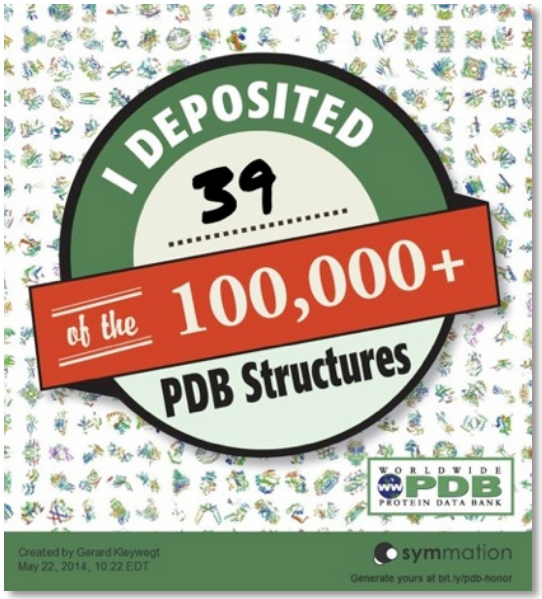
- Continued growth of archive: 100,000 structure milestone
- Increased use of data
- PDBx implemented in X-ray software packages
- Large structures released as “non-split” PDBx files
- Stand-alone X-ray validation server in production
- Validation reports for all X-ray entries released
- Common Deposition & Annotation System in production
- Workshops and meetings
  - mmCIF/PDBx Working Group
  - Hybrid Methods Task Force
  - SAS Task Force
  - NMR Restraints Format Meeting and Working Group
  - NMR Validation Task Force
  - mmCIF/PDBx Workshop for Programmers
  - IYCr and IUCr
- Funding stable (for now)
- wwPDB Foundation has new BoT Chair

PDB Depositions

Year	Total Depositions	Processed By		
		RCSB PDB	PDBj	PDBe
2000	2983	2297	158	528
2001	3287	2408	383	496
2002	3565	2401	657	507
2003	4830	3135	1026	669
2004	5508	3082	1614	812
2005	6678	3563	2110	1005
2006	7282	4252	1945	1085
2007	8130	4703	2299	1128
2008	7073	4106	1994	973
2009	8300	5069	2173	1058
2010	8878	5464	2041	1373
2011	9250	5938	1816	1496
2012	9972	6408	1888	1676
2013	10566	6653	2127	1786
2014	7521	4459	1342	1720
TOTAL	103823	63938	23573	16312

Last Updated: 24 Sep 2014

14 May 2014: 100,000+ PDB entries!



Hard data

It has been no small feat for the Protein Data Bank to stay relevant for 100,000 structures.

Sherlock Holmes understood: "It is a capital mistake," he said, "to theorise before one has data." Data are the lifeblood of science, the foundation of innovation. Behind every great discovery is a pile of data; but, crucially, it should not be too far behind.

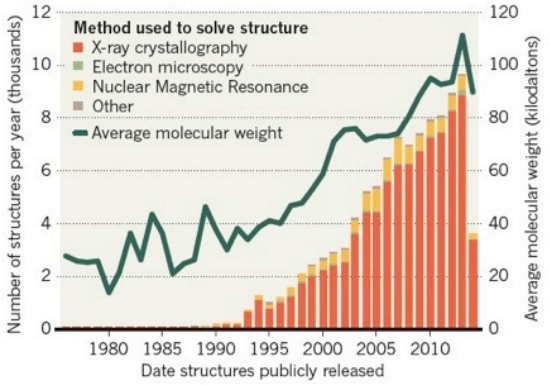
For more than four decades, the Protein Data Bank (PDB) has been where structural biologists keep their data close. Nearly every publishing journal in the world, *Nature* included, requires structures to be deposited in the PDB before publication.

TREND WATCH

A digital compendium of proteins and other biomolecules has surpassed 100,000 entries, with the release of 219 new structures on 14 May. The Protein Data Bank (PDB) was started in 1971 to store three-dimensional structural data down to the atomic level. Then and now, scientists mapped most proteins using X-ray crystallography, but they are increasingly using other tools, such as nuclear magnetic resonance and electron microscopy. See also page 260.

ONE HUNDRED THOUSAND PROTEIN STRUCTURES

Biomolecular structures stored in the Protein Data Bank are getting bigger and more complex.



FTP + Rsync Entry Downloads

RCSB PDB
2013: 312 million
2012: 298 million
2011: 282 million
2010: 159 million

PDBe
2013: 81 million
2012: 46 million
2011: 59 million
2010: 34 million

PDBj
2013: 40 million
2012: 21 million
2011: 38 million
2010: 16 million

PDBx File Format Milestones

- Limitations of PDB format necessitated “split” entries
- mmCIF/PDBx format does not have these limitations
- Workshop at EMBL-EBI in 2011 – decision to support PDBx in major refinement packages and to switch to PDBx as the distribution format for the PDB archive

PDBx File Format Milestones

- PDBx implemented in X-ray software packages
  - CCP4 (REFMAC 5.8)
  - Phenix (1.8.2)
- All large structures released as “non-split” PDBx files
  - Combined files available in separate FTP directory during a 6-month testing period
  - After review, combined files will move to the main PDB FTP archive
  - Multiple “split” entries will be taken out of the active archive (“*Formageddon*” – 10 December 2014)
  - Searches using a “split” PDB ID at all wwPDB member sites will return the intact entry

PDBx/mmCIF Working Group

Meeting at EMBL-EBI, Hinxton, 8 October 2014

PDBx/mmCIF Working Group

- Issues discussed:
  - Handling of reflection data and data-collection details
  - Ligands and chemical descriptions, restraints, linkages, etc.

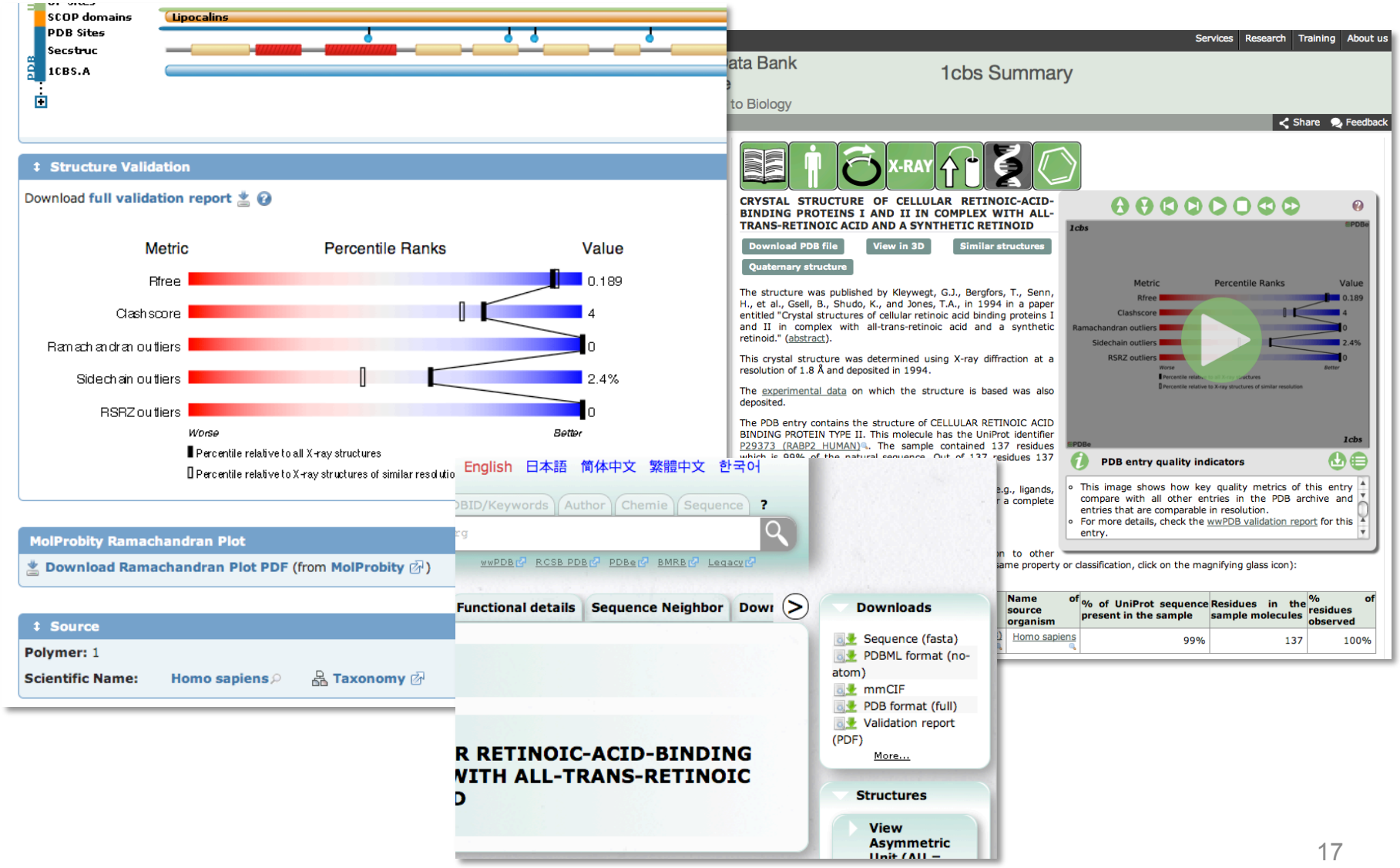
PDBx/mmCIF Programmer Workshop

EMBL-EBI, Hinxton, 20-21 November 2013

Validation

- wwPDB and EMDataBank have convened Validation Task Forces for X-ray, NMR, and EM
- Recommendations about validating new and existing structures
  - Implemented in software pipeline
  - Produces summary report (PDF) and XML file with detailed statistics
- Validation at different stages
  - While you determine the structure (preliminary)
  - During deposition (preliminary)
  - After annotation (official; can be sent to journals)
  - Upon release (publicly available; will be updated annually)

Public X-ray Validation Reports



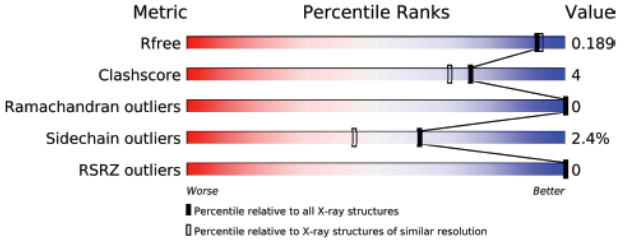
Validation Reports

- Summary
  - Quality vs. all PDB X-ray entries
  - Quality vs. entries at similar resolution
  - Overview of residue-based quality for every polymer
  - Table of ligands that may need attention

1 Overall quality at a glance

The reported resolution of this entry is 1.80 Å.

Percentile scores (ranging between 0-100) for global validation metrics of the entry are shown in the following graphic. The table shows the number of entries on which the scores are based.



Metric	Whole archive (#Entries)	Similar resolution (#Entries, resolution range(Å))
R_{free}	66092	3513 (1.80-1.80)
Clashscore	79885	4461 (1.80-1.80)
Ramachandran outliers	78287	4404 (1.80-1.80)
Sidechain outliers	78261	4403 (1.80-1.80)
RSRZ outliers	66119	3515 (1.80-1.80)

The table below summarises the geometric issues observed across the polymeric chains and their fit to the electron density. The red, orange, yellow and green segments on the lower bar indicate the fraction of residues that contain outliers for >=3, 2, 1 and 0 types of geometric quality criteria. The upper red bar (where present) indicates the fraction of residues that have poor fit to the electron density.

Mol	Chain	Length	Quality of chain
1	A	137	

Mol	Chain	Length	Quality of chain
1	A	735	
1	B	735	
1	C	735	
1	D	735	

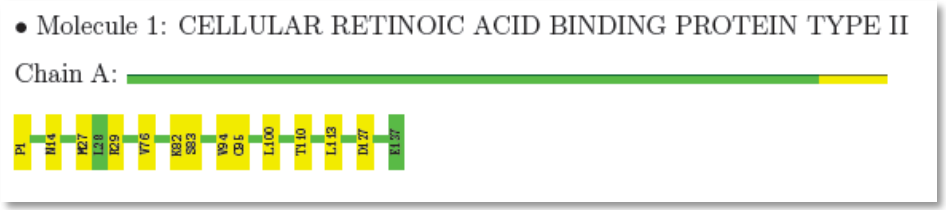
The following table lists non-polymeric compounds that are outliers for geometric or electron-density-fit criteria:

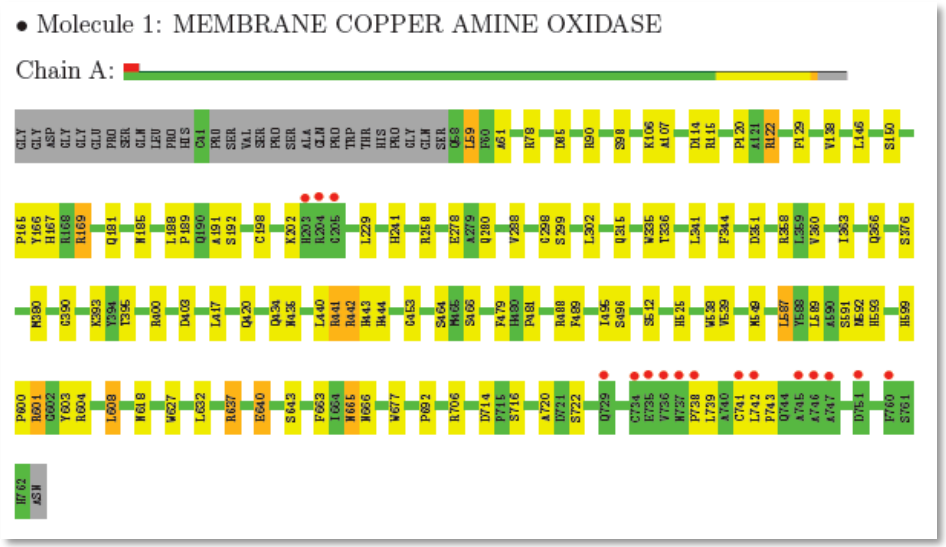
Mol	Type	Chain	Res	Geometry	Electron density
5	NAG	A	1768	-	X
5	NAG	B	1768	-	X
5	NAG	C	1768	-	X
5	NAG	D	1768	-	X

Continued on next page...

Validation Reports (cont.)

- Residue quality
  - One plot per polymer
  - Coloured by number of ***types*** of geometric outliers
  - Grey if not modelled
  - Red dots: poor density (RSR-Z > 2, as in EDS)

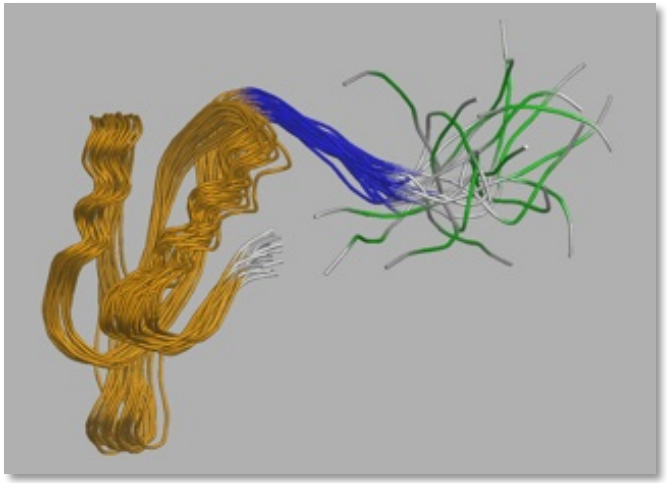




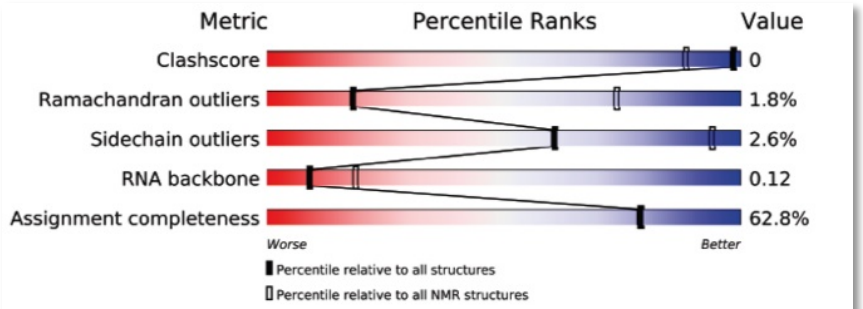
Other Methods?

- Model validation using same criteria as X-ray
  - MolProbity, Mogul
  - Later: WhatCheck
- Some special model-related issues per technique
  - X-ray: alternative conformations
  - NMR: ensemble of models; well-defined regions
  - 3DEM: clashes of rigid-body fitted models; difference in species of model and sample sequence
- Data quality and model/data-fit assessment will be different for each technique

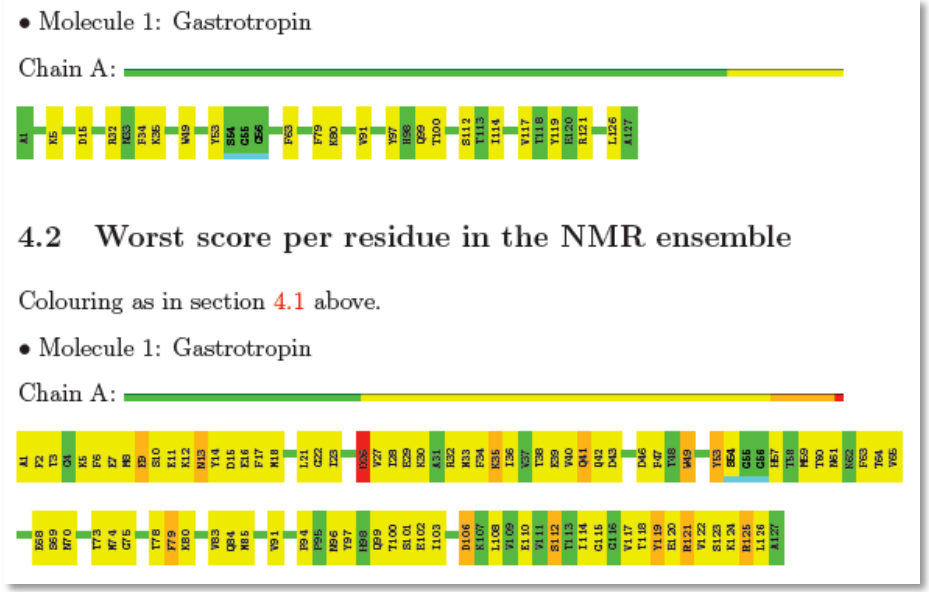
NMR Validation



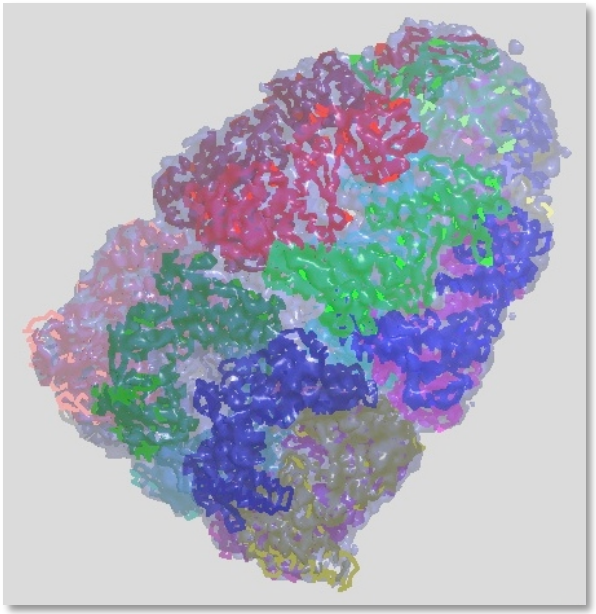
- NMR VTF recommendations published
- Global quality scores reported for “well-defined residues” only
  - As averages over the ensemble
  - Worst-case instance in the ensemble



Metric	Whole archive (#Entries)	NMR archive (#Entries)
Clashscore	99129	10081
Ramachandran outliers	96105	8982
Cα geometry	96347	8988
Sidechain outliers	96047	8965
RNA backbone	2807	549
Assignment completeness	1540	1532



3DEM Validation



- Model validation
  - Clashes?
  - Taxonomy?
  - Homology models?
  - Non-atomistic models?
  - Cα-only models?
  - Rigid-body *vs*. flexible fitting *vs*. de novo modelling?
- Data and map validation
  - Per technique and resolution regime
  - Tilt-pair analysis; handedness; projections vs. raw data
- Map + model
  - Depending on resolution regime and model-building method?

EM Validation Reports

- Metrics relevant for EM models
- Define “Table 1” for EM

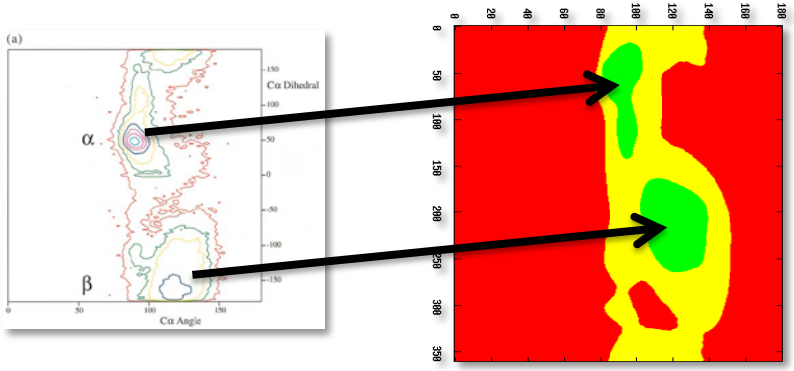
wwPDB EM Map/Model Validation Report (i)

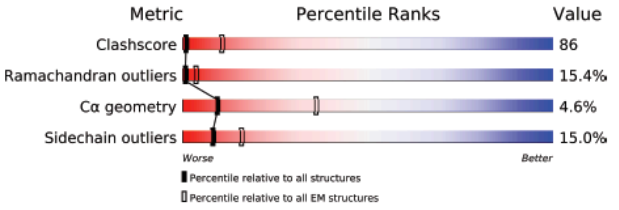
Jun 19, 2014 – 03:44 PM BST

PDB ID	:	3IZX
Title	:	3.1 Angstrom cryoEM structure of cytoplasmic polyhedrosis virus
Authors	:	Yu, X.; Ge, P.; Jiang, J.; Atanasov, I.; Zhou, Z.H.
Deposited on	:	2011-01-15
Resolution	:	3.10 Å(reported)

4 Experimental information (i)

Property	Value	Source
Reconstruction method	Not provided	Depositor
Imposed symmetry	I	Depositor
Number of images	28993	Depositor
Resolution determination method	FSC at 0.143 cut-off	Depositor
CTF correction method	Each particle	Depositor
Microscope	OTHER	Depositor
Voltage (kV)	300	Depositor
Electron dose ($e^-/\AA^2$)	25	Depositor
Minimum defocus (nm)	1000	Depositor
Maximum defocus (nm)	2500	Depositor
Magnification	59000	Depositor
Image detector	Kodak SO 163 film	Depositor



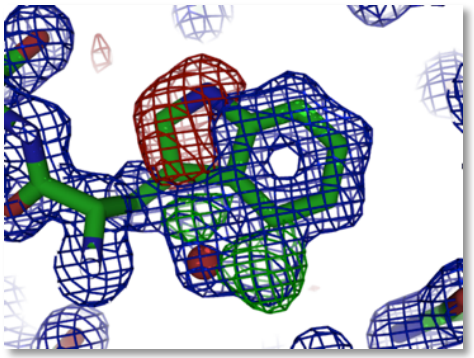


Metric	Whole archive (#Entries)	EM structures (#Entries)
Clashscore	99129	735
Ramachandran outliers	96105	539
Cα geometry	96347	682
Sidechain outliers	96047	526

The table below summarises the geometric issues observed across the polymeric chains. The red, orange, yellow and green segments on the lower bar indicate the fraction of residues that contain outliers for >=3, 2, 1 and 0 types of geometric quality criteria.

Mol	Chain	Length	Quality of chain
1	A	1058	
2	B	1333	
2	C	1333	
3	D	448	
3	E	448	

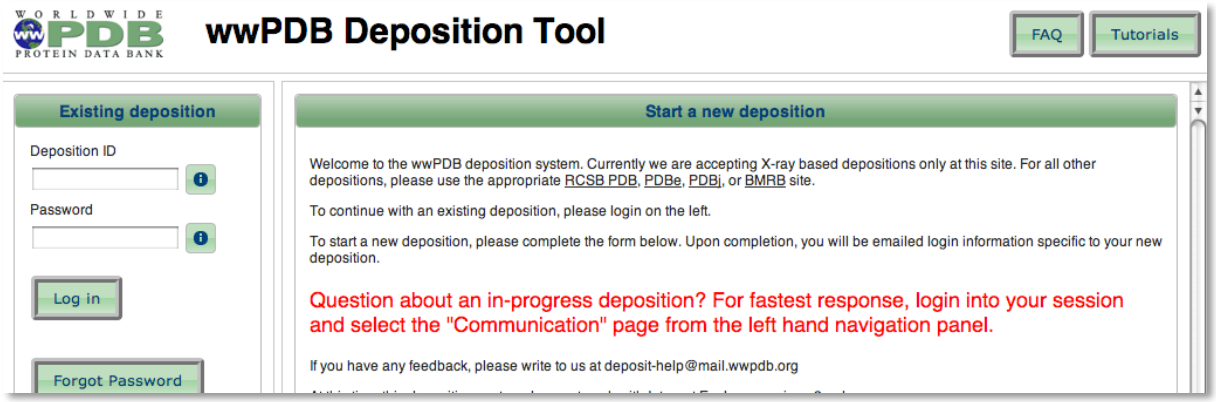
Validation by wwPDB



- By no means the end of the story!
  - Room for extension and improvement
    - Ligands, nucleic acids, carbohydrates, NCS, spacegroup errors, …
    - wwPDB ligand-validation workshop in 2015
  - X-ray
    - Re-convene X-ray VTF in 2015 to evaluate and update recommendations
  - NMR
    - Further development in progress
  - EM
    - Rudimentary at present, lots more work needed
  - All methods: annual re-compute of distributions
  - User feedback welcome at validation@mail.wwpdb.org

Common Deposition & Annotation: 2014 Project Milestones

- January 2014: X-ray production testing began
- Fall 2014: 3DEM and NMR testing
- From early 2015: Parallel deposition systems (new and legacy) available to depositors during transition period



http://deposit.wwpdb.org/deposition

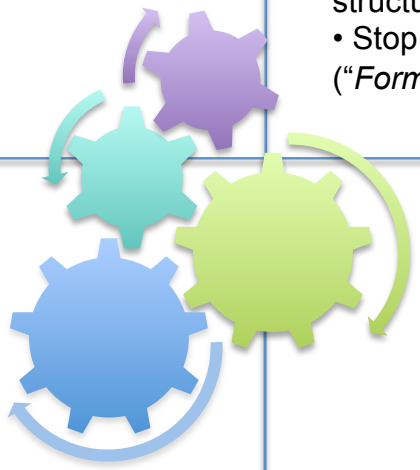
Transition Status

Deposition & Annotation

✓New and old annotation systems in production
✓New and old deposition systems in production
✓Weekly update supporting transition
✓~2000 structures deposited in new system
•Retire legacy systems

Format

✓PDBx Working Group initial implementations
✓Ribosome example: new style mmCIF files
✓Community outreach
✓Start accepting new style format depositions
✓Created & released combined large structures
• Provide best-effort PDB-format files for structures that don't fit PDB-format requirements
• Stop supplying PDB files in ftp archive ("*Formageddon*")



wwPDB Website(s)

✓ Announced D&A
• Unveil new redesigned website at wwpdb.org

Archive

✓ Archive updated to support "combined" large structures
• Reformat PDBx/mmCIF data files in archive to conform to new style guidelines

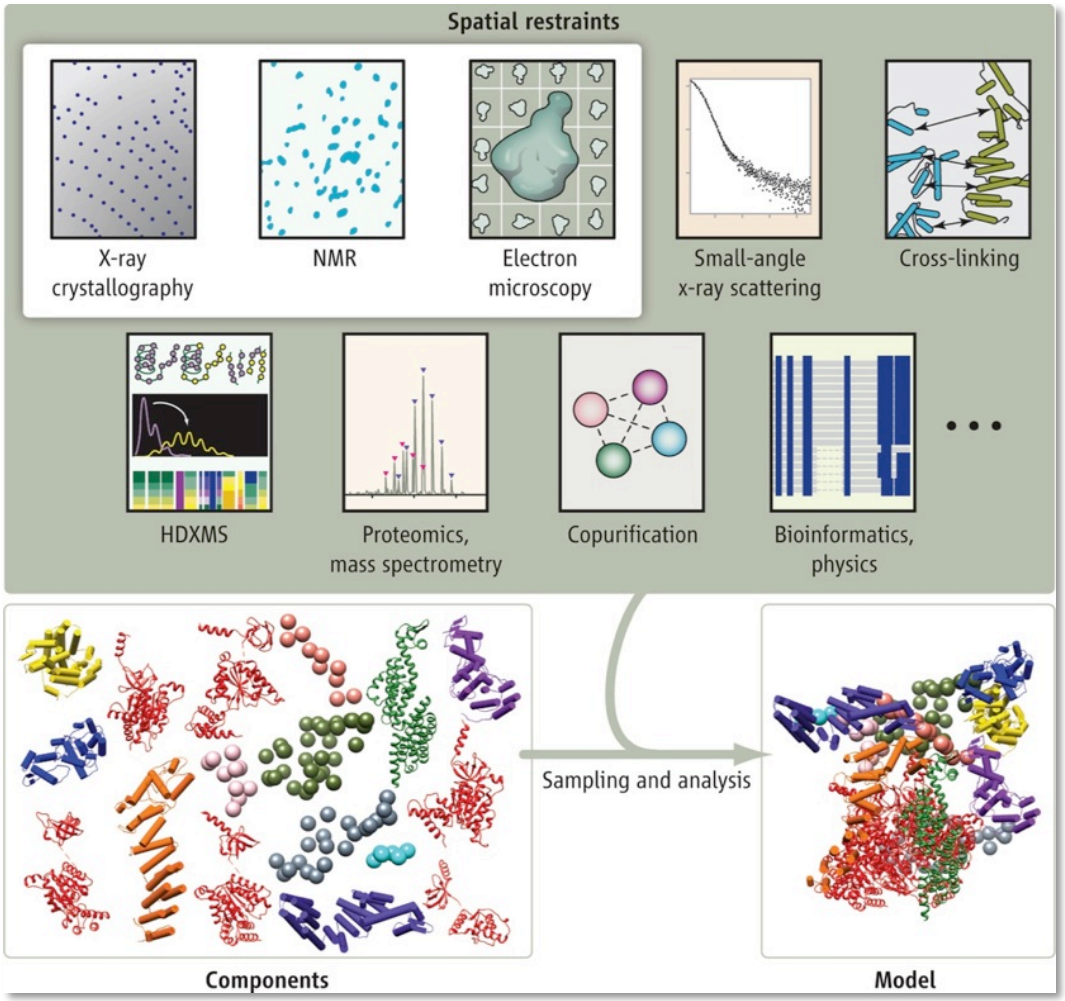
wwPDB Task Forces

Method-specific (Validation) Task Forces have been convened to collect recommendations and develop consensus on additional validation that should be performed, and to identify software applications to perform validation tasks, and to discuss archival needs and opportunities for non-traditional techniques.

Task Force	Meeting/ Workshop	Chair(s)/Membership	Outcome
X-ray Validation Task Force	2008 (2015)	Randy Read (Univ of Cambridge) 17 members	(2011) *Structure* 19: 1395-1412
NMR Validation Task Force	2009, 2011, 2013 (x2) (2014)	Gaetano Montelione (Rutgers) Michael Nilges (Institut Pasteur) 10 members	(2013) *Structure*, 21: 1563-1570
3DEM Validation Task Force	2010	Richard Henderson (MRC-LMB) Andrej Sali (UCSF) 21 members	(2012) *Structure* 20: 205-214
Small-Angle Scattering Task Force	2012, 2014	Jill Trewhella (Univ Sydney) 6 members	(2013) *Structure* 21: 875-881
Hybrid Methods Task Force	2014	Andrej Sali (UCSF), Torsten Schwede (Univ Basel), Jill Trewhella (Univ Sydney) 27 members	(To be published)

Hybrid Methods

- Task Force met at EMBL-EBI this week
- Representatives of existing task forces, other methods, integrative modellers, and wwPDB
- Questions about what to archive where, what data and meta-data, how to validate



wwPDB Hybrid Methods Task Force

EMBL-EBI, Hinxton, 6-7 October, 2014

X-ray NMR 3DEM/ET SAS FRET EPR MS …
Modelling Docking Validation Visualisation Archiving …

Key Outcomes of Discussion

- Be as inclusive as possible in collecting data from many different experimental methods
- Accommodate many types of structural representations
- Create a federated system to collect/curate data
- Use a common interface to collect data
- wwPDB should play a leadership role
- Whitepaper to describe vision

wwPDB Organization

- New wwPDB charter (July 1, 2013)
  - Updated from 2003 text and circumstances
  - Signed by PIs and heads of host institutions
  - Covers 2013-2023, with review possible in 2018
  - Technical details in appendix
    - Conventions for PDB accession codes
    - Conventions for file names
    - Guidelines for redistribution
    - Current and future guidelines for data exchange
    - Guidelines for PDBx data dictionary maintenance
    - Guidelines for maintenance of reference data
- New appointments and terms of reference for wwPDB AC

Funding

- RCSB PDB non-competitive renewal funded by NSF (2014-2018)
- PDBe main funding from EMBL and Wellcome Trust
  - EMBL-EBI: core funding for ~15 posts
  - Wellcome Trust: new competitive grant awarded (4 posts, 2015-2019)
  - Additional new project grants from BBSRC, CCP4, and MRC
- PDBj competitive renewal funded by JST (Japan Science & Technology Agency) for April 2014 - March 2017
- BMRB is now funded by NIH-NIGMS (2014-2019)

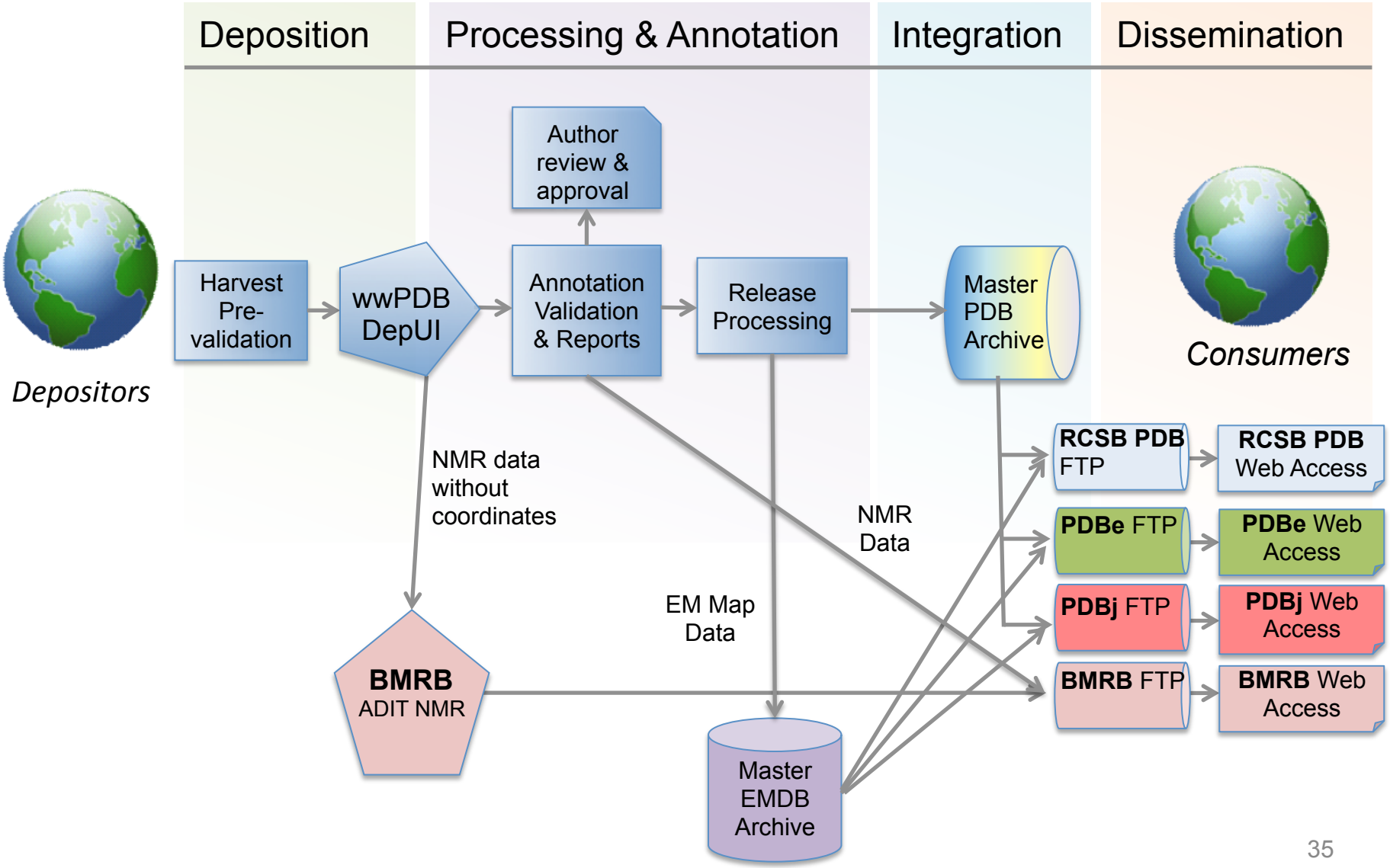
Worldwide Protein Data Bank Foundation

- Established to support specific wwPDB activities
  - Advisory committee meetings
  - Outreach and education activities, including seminars and workshops
- 501(c)3 organization
  - American, tax-exempt association dedicated to scientific, literary, charitable, and educational purposes
- Fundraising on-going
- New BoD Chair: Anthony Nicholls
- 2014: Supported IYCr calendar, Hybrid Methods TF
- 2015: Will support ligand-validation workshop

Common Deposition & Annotation (D&A) Tool: Summary Status

Stephen Burley

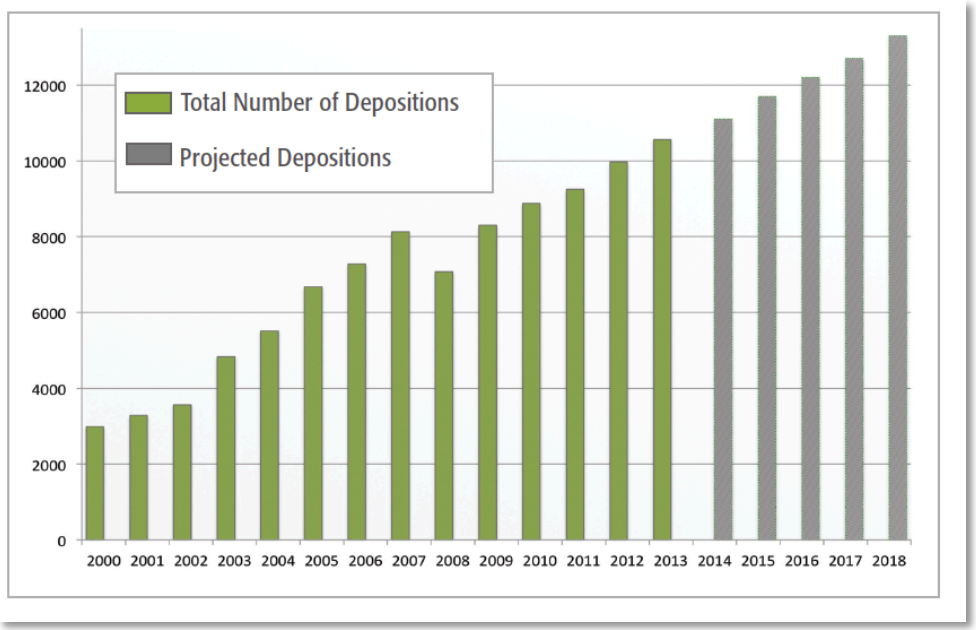
Common Deposition and Annotation Pipeline for PDB, EMDB and BMRB Data



Vision and Delivery

Standardization, Quality and Efficiency

- ✓ Larger and more complex biological molecules
- ✓ Expanded annotation
- ✓ Increased throughput: Automation and validation of routine submissions

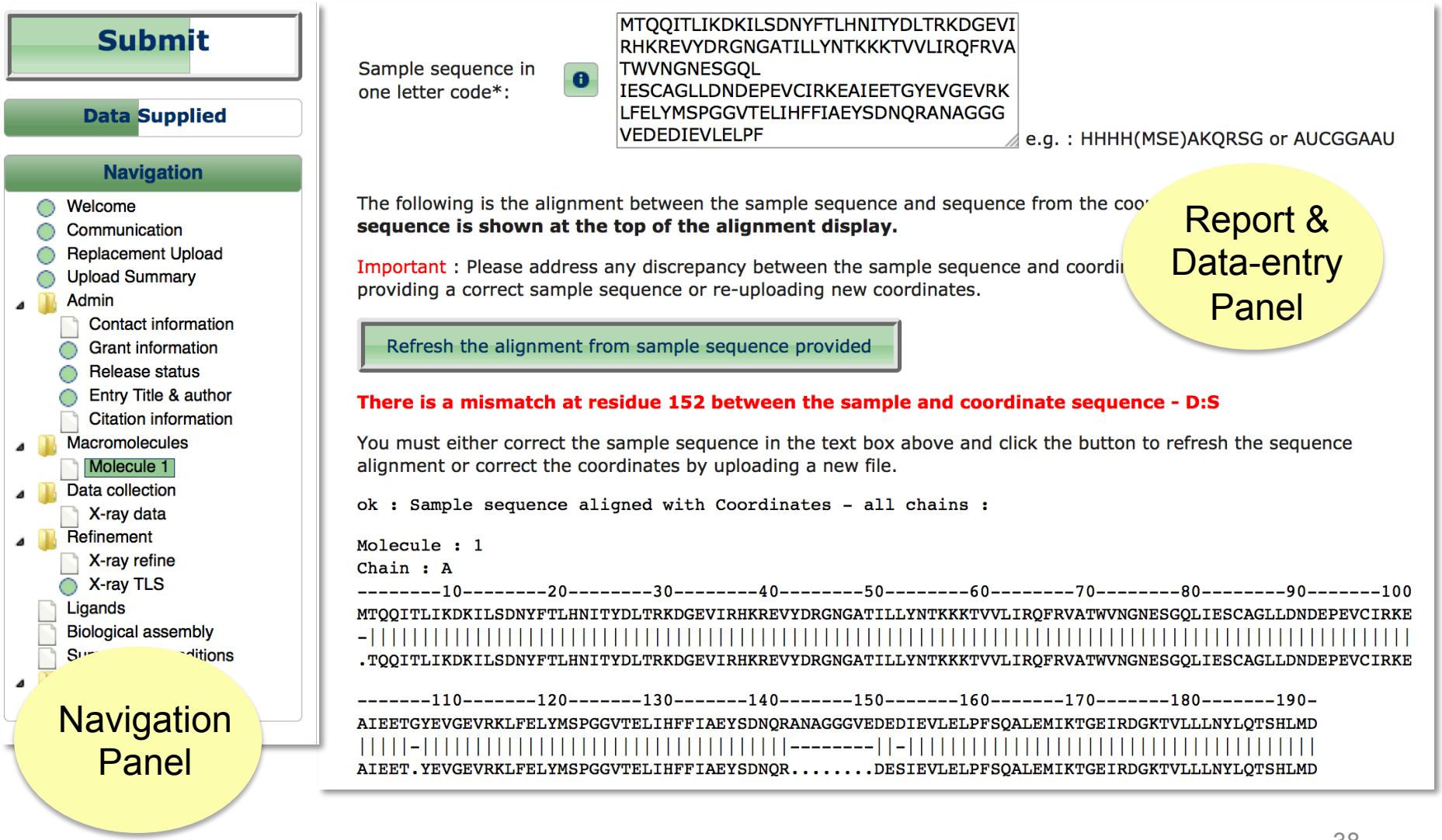


Deposition Pipeline

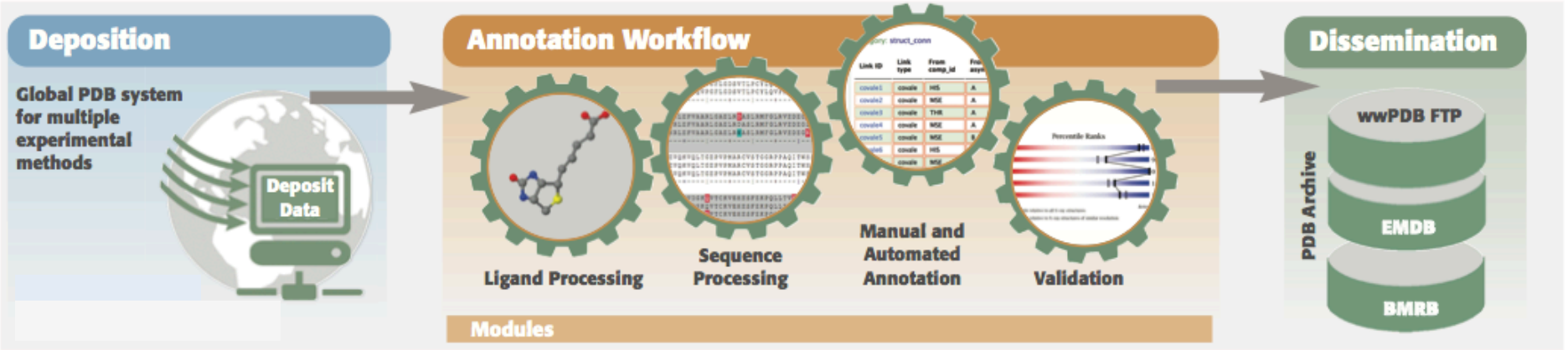
Depositors can

- Base new entries on previous depositions
- Upload replacement files mid-deposition (e.g., a re-refined model)
- Provide more complete data with required mandatory data items
- View community-defined validation report prior to submission
- Review and download annotated files post-deposition
- Communicate with expert annotators during deposition using web-based tools

Interactive Deposition User Interface (UI)



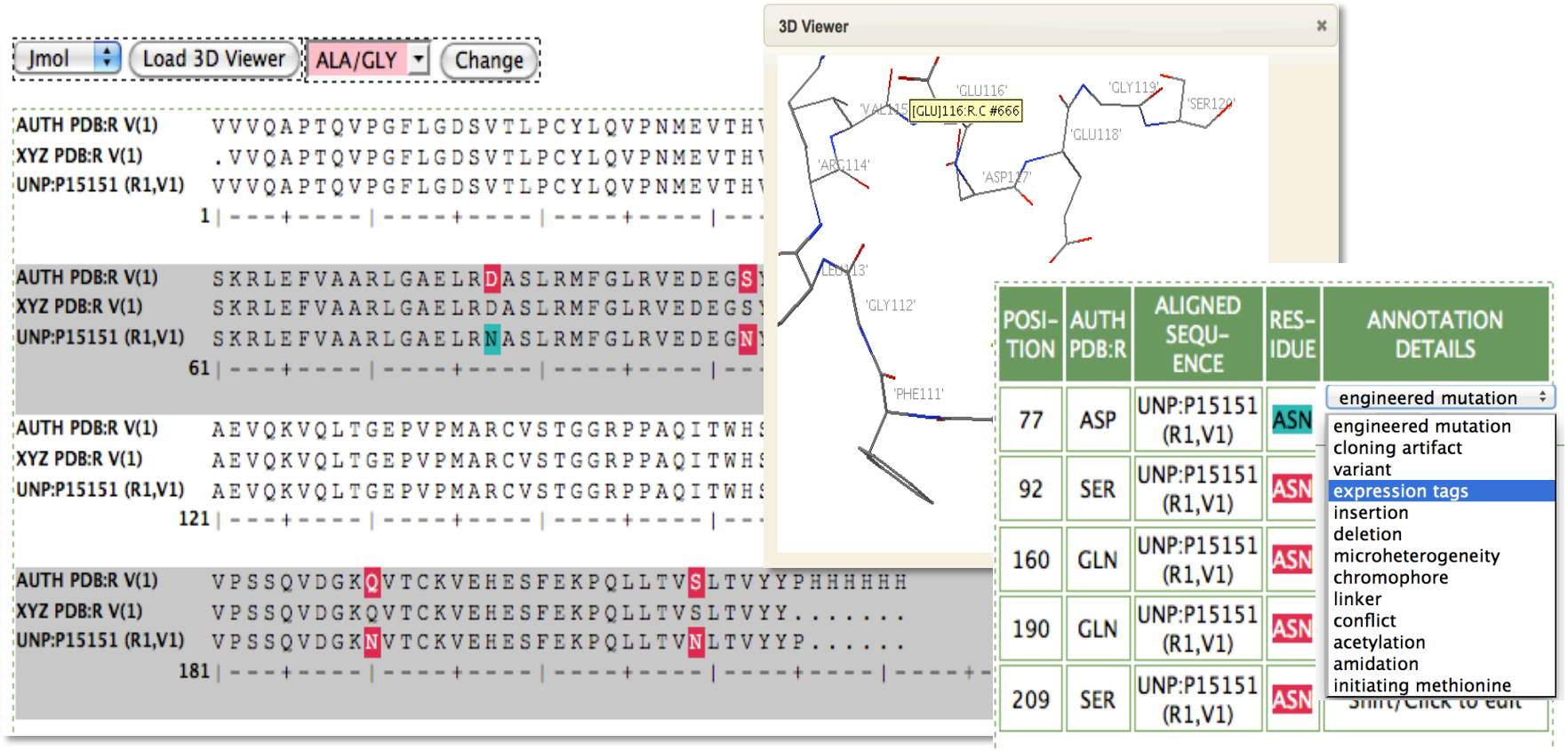
Annotation System



- Enables workload balancing and increased productivity
- Better quality assurance of polymer sequences and ligand chemistry
- PDBx/mmCIF is the master file format
- Validation suites based on recommendations from expert task forces; X-ray validation pipelines available as a stand-alone server
- System will support all accepted experimental methods

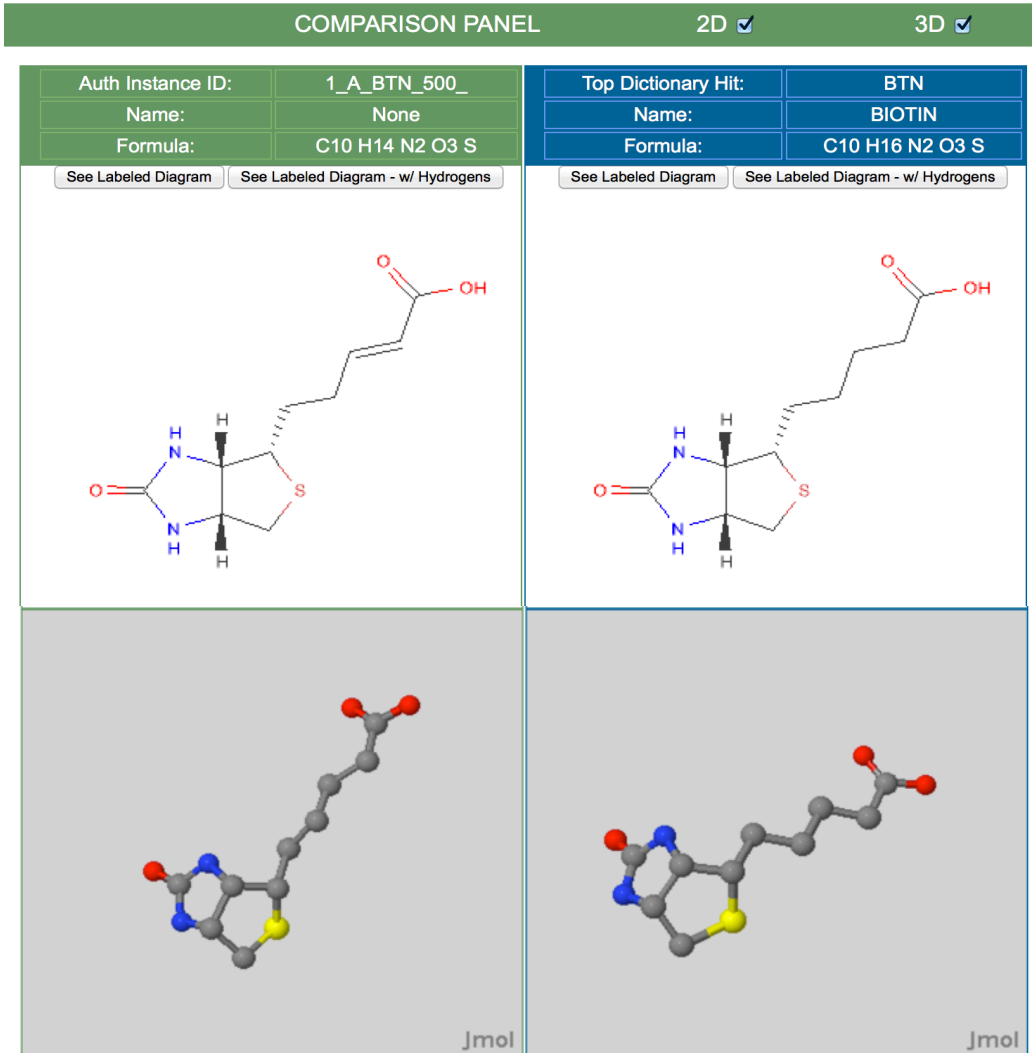
Sequence Module: Improved Data Consistency

- Biological sequence checked against atomic coordinate sequence and cross-referenced to UniProt/GenBank
- 3D structure view
- Sequence discrepancy annotation



Ligand Module: Improved Chemistry Assessment

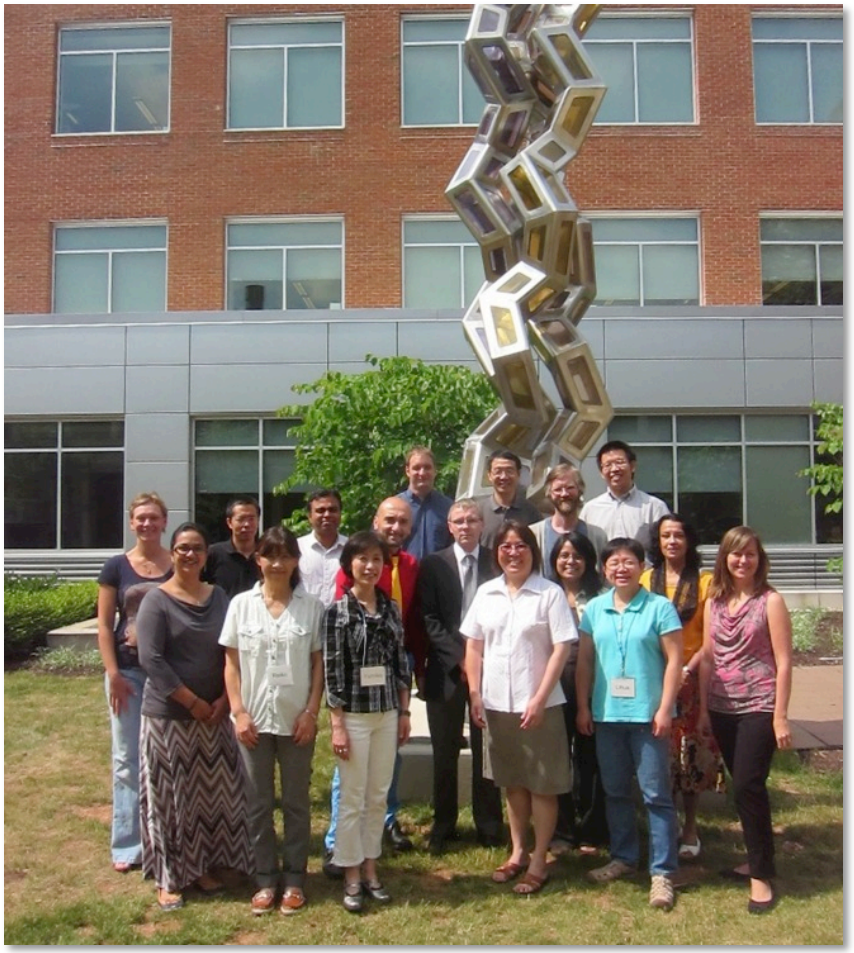
- Batch search against Chemical Component Dictionary with automated ligand ID assignment
- Captures and displays author-provided chemical information
- Comparison panel
  - 2D and 3D views of ligand for review
  - ID assignment



Author instance from coordinates (left) and the closest match in the dictionary (right)

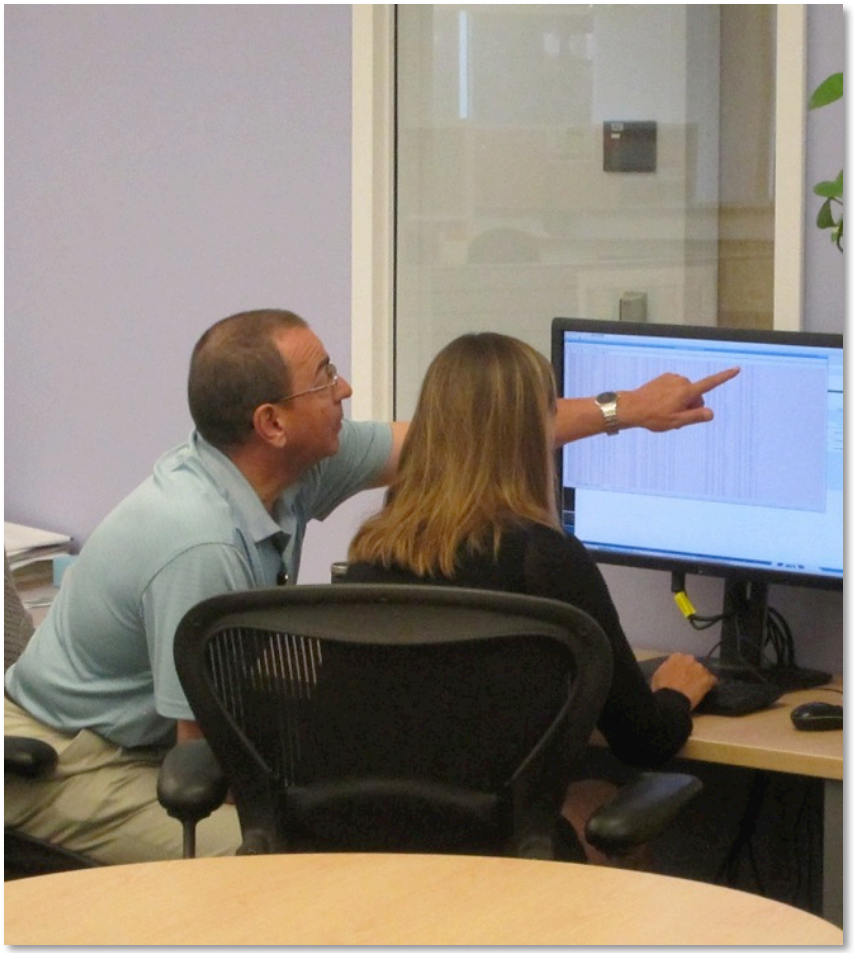
wwPDB Annotator Summit

- June 23-25 at RCSB PDB-Rutgers
- Annotators in attendance
  - 3 PDBe
  - 2 PDBj
  - 5 RCSB PDB
- Goals
  - Prioritize existing issues
  - Develop future requirements



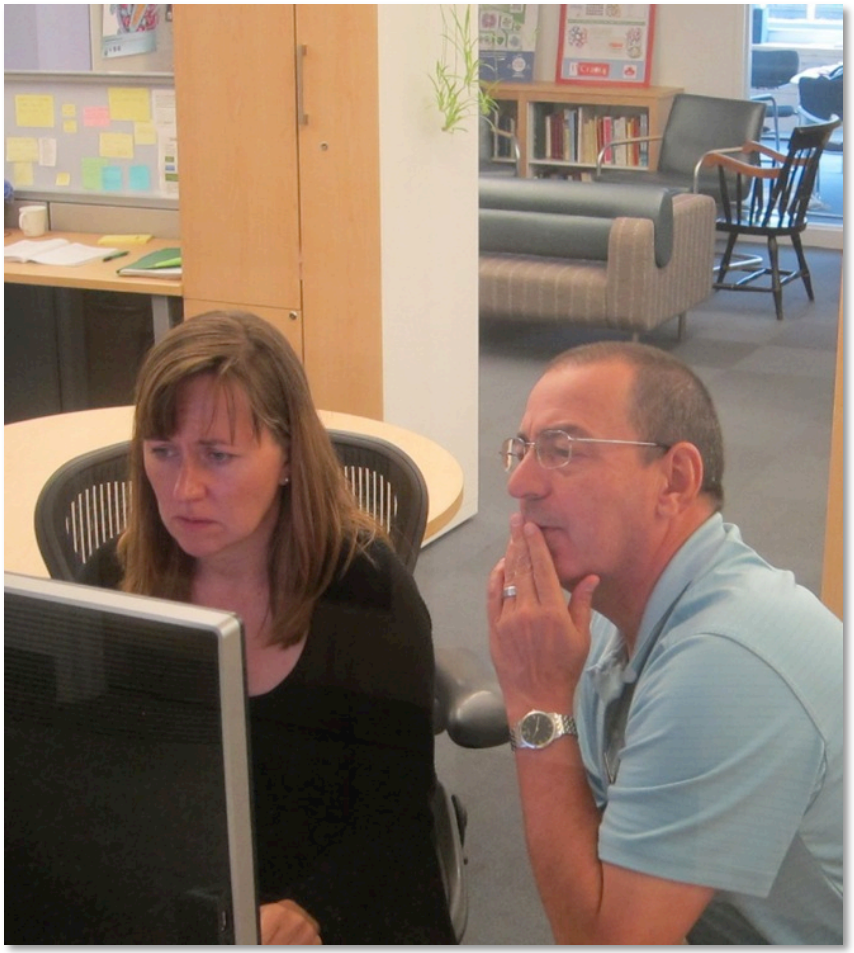
Reflections of a Newly-Trained wwPDB Annotator

- Surprisingly good for a system “designed by committee”
- Sequence Processing module easy to use
- Ligand Processing module excellent!
- Looking forward to full implementation



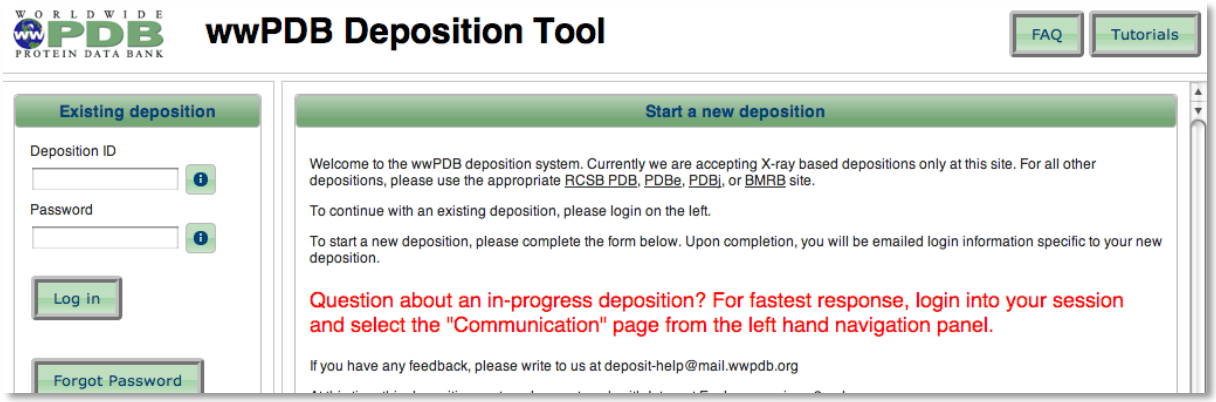
Reflections of a Newly-Trained wwPDB Annotator (cont.)

- Co-crystal structures
- “It was the best of times, it was the worst of times”-Dickens (1859)
- Ligand Processing module lays bare the wide variation in quality of co-crystal structures
- Looking forward to the Ligand-Validation workshop



Common Deposition & Annotation: 2014 Project Milestones

- January 2014: X-ray production testing began
- Fall 2014: 3DEM and NMR testing
- From early 2015: Parallel deposition systems (new and legacy) available to depositors during transition period



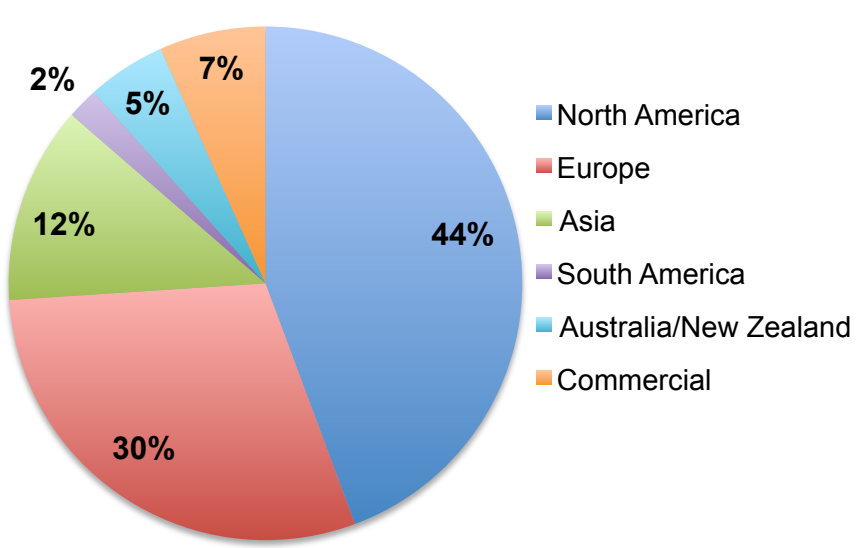
http://deposit.wwpdb.org/deposition

2014 Depositions by Continent

January 1 – August 31, 2014

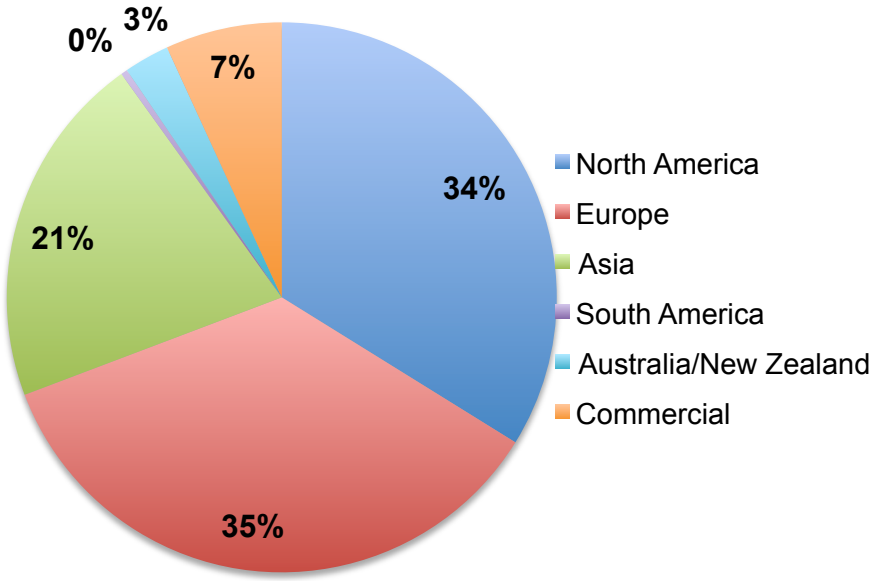
D&A System

(2197 entries, 30% of total depositions)



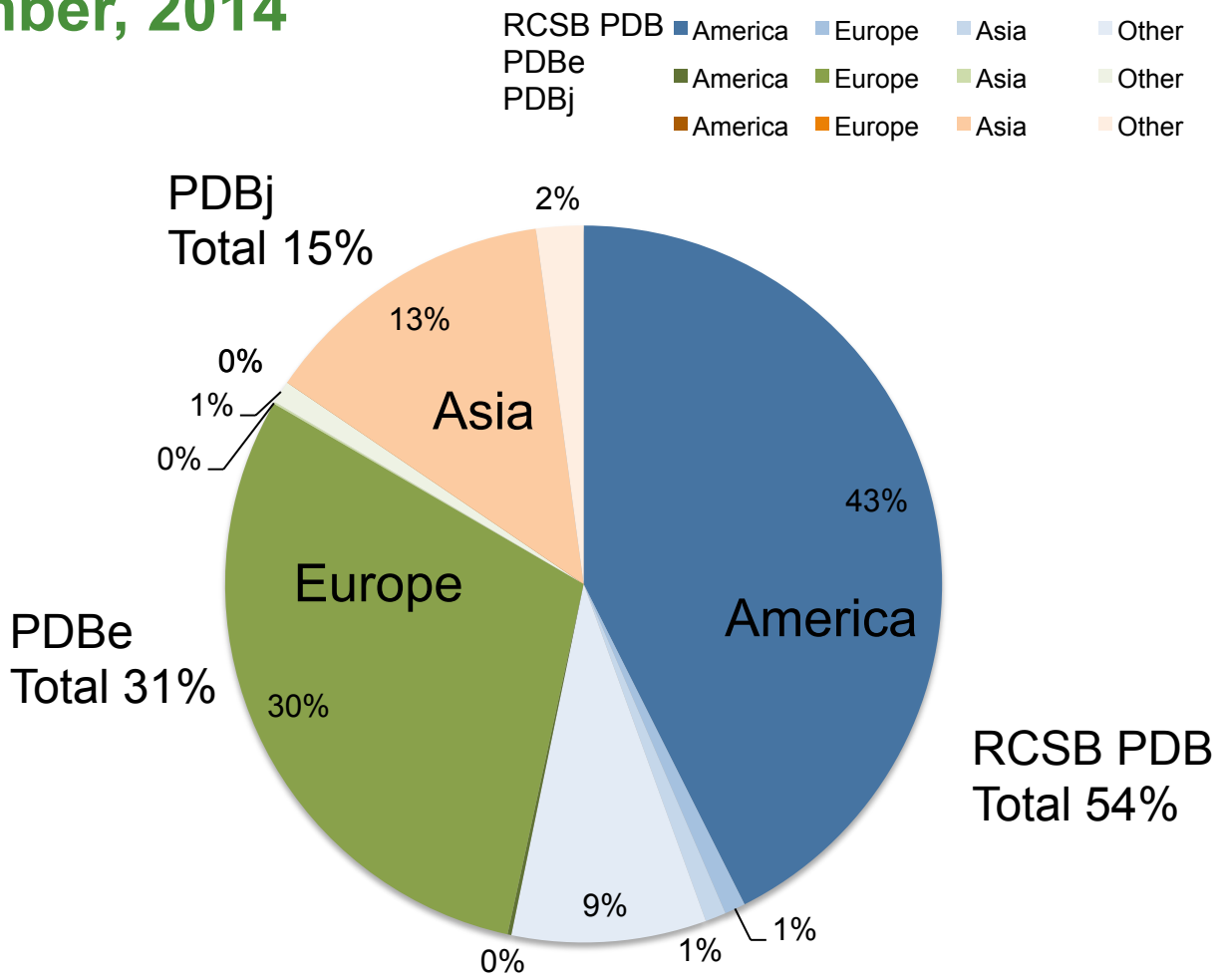
Legacy System

(4713 entries, 70% of total depositions)



D&A Annotation Distribution by Processing Site and Geography

June-September, 2014



NMR - BMRB

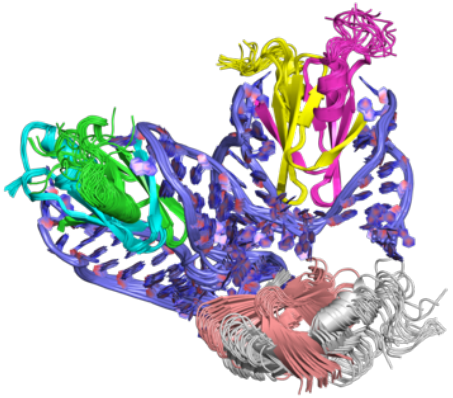
John Markley

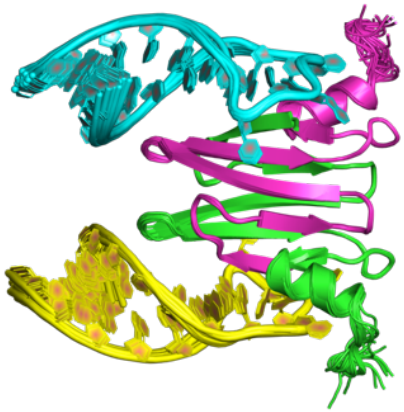
WORLDWIDE PDB
PROTEIN DATA BANK

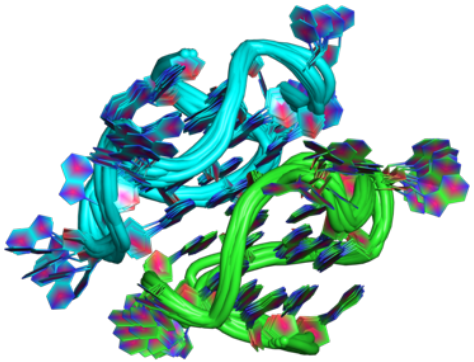
wwpdb.org

NMR Depositions (Sep 2013 – Aug 2014)

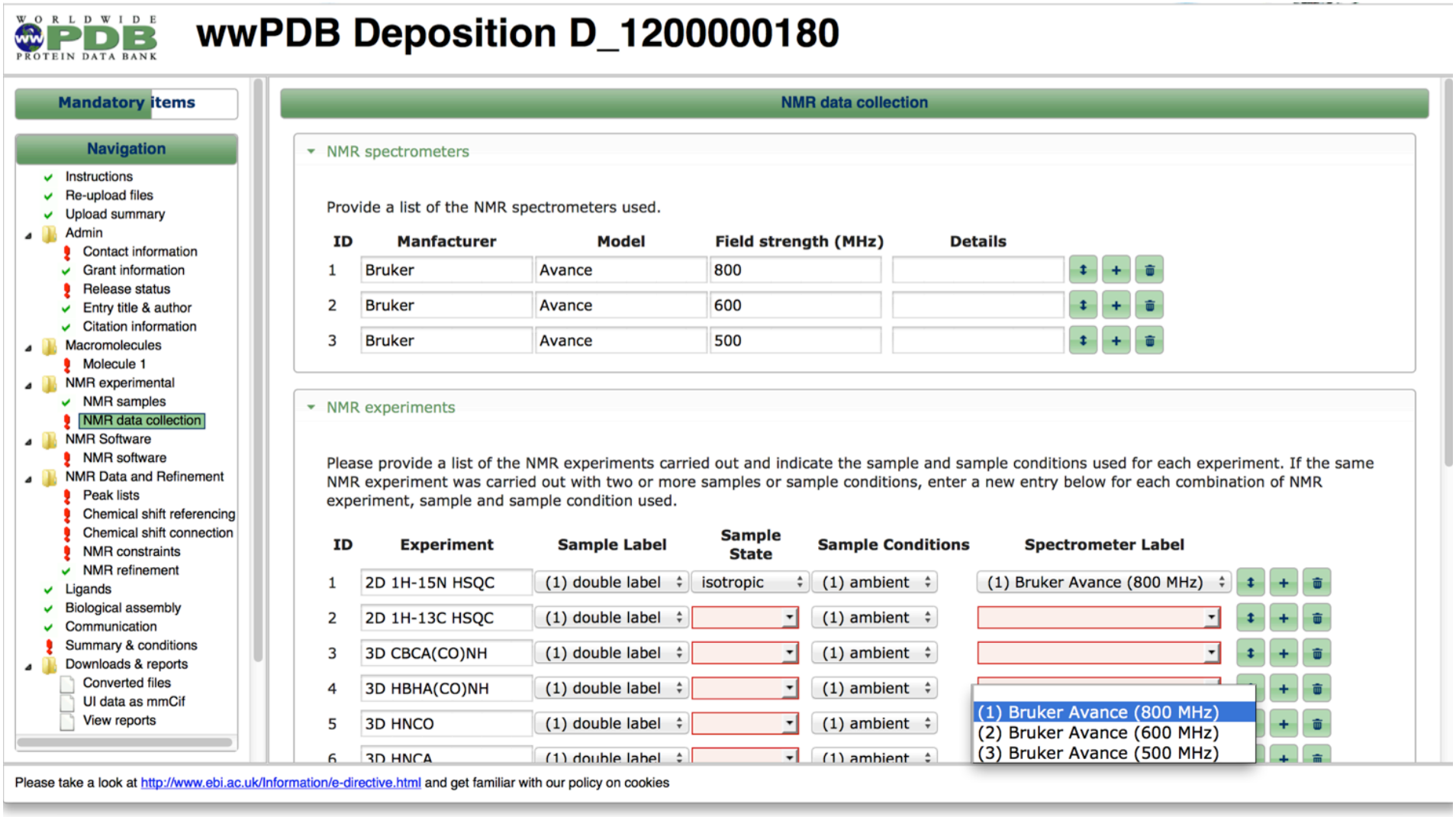
Site	Structures	Experimental data without structures	Total
ADIT-NMR	538	226	764
AutoDep	12	-	12
PDBj-BMRB	25	14	39



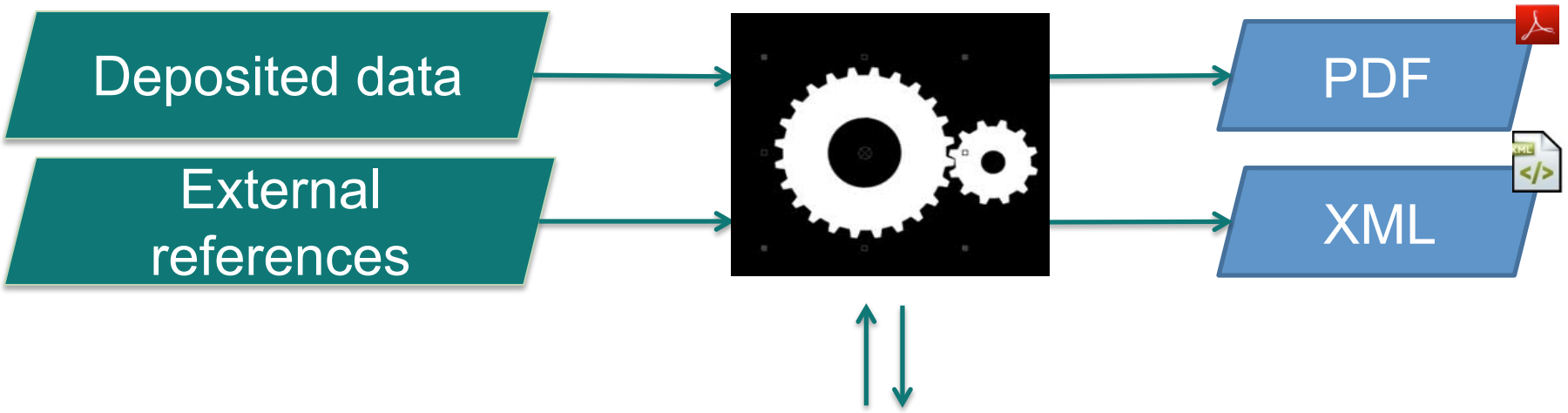




Deposition & Annotation System: NMR



Validation Pipeline Overview



Mogul
MolProbity
Percentiles

EDS
Xtriage

Cyrange
BMRB stats
PANAV
RCI

Common
X-ray
NMR

NMR Validation: Work in Progress

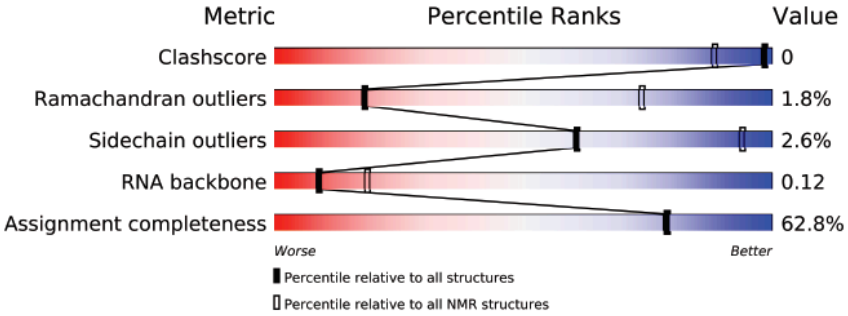
- Version 1 works with real data
- Includes ensemble analysis, model validation, chemical shifts, annotation information
- Include in D&A system

- Next steps
  - Testing on the whole archive
  - Incorporate restraint validation and peak lists
  - Provide as stand-alone server

NMR Validation Reports

1 Overall quality at a glance

Percentile scores (ranging between 0-100) for global validation metrics of the entry are shown in the following graphic. The table shows the number of entries on which the scores are based.



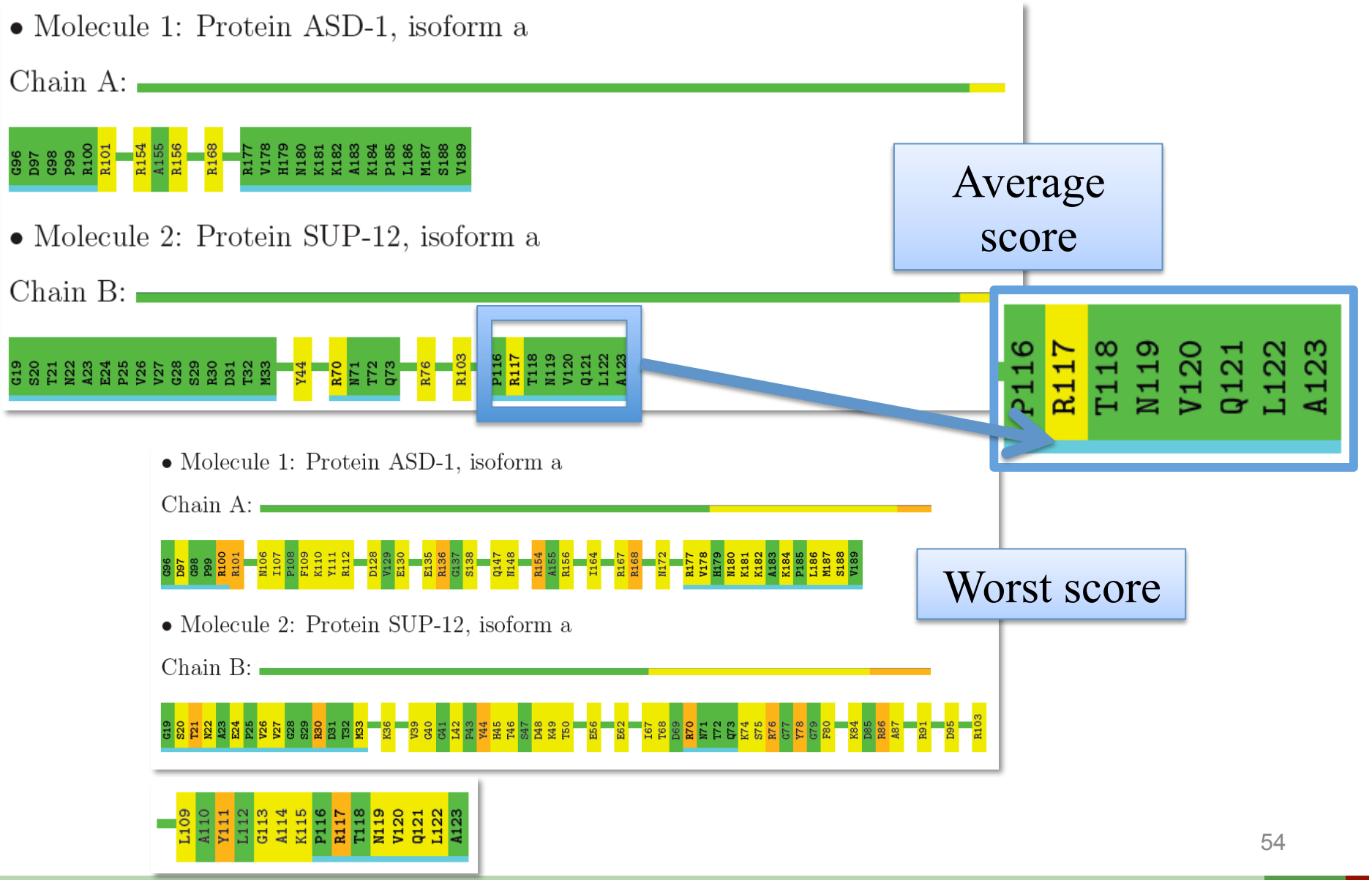
Metric	Whole archive (#Entries)	NMR archive (#Entries)
Clashscore	99129	10081
Ramachandran outliers	96105	8982
Cα geometry	96347	8988
Sidechain outliers	96047	8965
RNA backbone	2807	549
Assignment completeness	1540	1532

The table below summarises the geometric issues observed across the polymeric chains and their fit to the experimental data. The red, orange, yellow and green segments indicate the fraction of residues that contain outliers for >=3, 2, 1 and 0 types of geometric quality criteria.

Mol	Chain	Length	Quality of chain
1	A	94	
2	B	105	
3	C	12	

Summary page

NMR Validation Reports



NMR Validation Reports

7 Chemical shift validation (i)

7.1 Chemical shift list 1

File name: 2mgz_cs.str

The following table describes results of parsing the sift list and outlier detection.

Total number of shifts	949
Number of unparsed shifts	0
Number of shifts mapped to atoms	949
Number of shifts with mapping errors	0
Number of shifts with mapping warnings	0
Number of shift outliers (ShiftChecker)	6

The following table shows the suggested chemical shift referencing corrections (values are in *ppm*).

Cα	Cβ	C$'$	N
-0.27 $\pm$ 0.05	0.18 $\pm$ 0.05	0.00 $\pm$ 0.00	0.41 $\pm$ 0.15

NMR Validation Reports

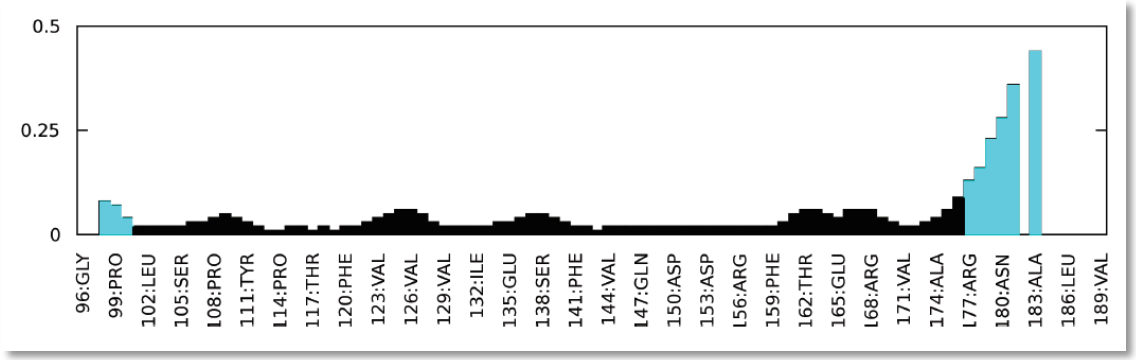
The following table shows the completeness of the chemical shift assignments. The overall completeness is 31%, i.e. 848 atoms were assigned a chemical shift out of possible 2729. 1 out of 27 assigned methyl groups (LEU and VAL) were assigned stereospecifically.

	Total	H	C	N
Backbone	318/973 (33%)	159/387 (41%)	81/398 (20%)	78/188 (41%)
Sidechain	456/1268 (36%)	279/724 (39%)	163/466 (35%)	14/78 (18%)
Aromatic	74/263 (28%)	44/155 (28%)	30/104 (29%)	0/4 (0%)
Overall	848/2729 (31%)	482/1395 (35%)	274/1047 (26%)	92/287 (32%)

The following table lists the chemical shift outliers reported.

Mol	Chain	Res	Type	Atom	Shift	Expected value	Z-Score
1	A	115	ASP	HB3	1.15	2.67	-5.44
1	A	154	ARG	HG3	123.78	84.68	24.59
1	A	159	PHE	HD1	1.95	4.26	-5.24
1	A	123	VAL	HG22	0.18	2.32	-7.94
1	A	170	GLU	CB	125.14	84.68	25.44
1	A	168	ARG	NE	110.36	84.68	16.15

Random coil index (RCI) for chain A:



NMR VTF on Validation of Restraints

- Initially, simple counts and violations
- Later, more sophisticated measures
  - Metric for information content
  - Dealing with novel potentials
  - Analysis of violations
- **HOWEVER**
  - This must be fully automatic
  - Not feasible to continue to accept restraints in any old (or new) format
    - *e.g.,* 3 ways to represent a logical OR in one popular software system

Workshop on Representation of NMR Restraints

EMBL-EBI, Hinxton, 18-19 November, 2013

Workshop on Representation of NMR Restraints (cont.)

- All major NMR structure determination, refinement and validation packages represented, mostly by PIs:
  - D. Case (AMBER)
  - P. Güntert (CYANA)
  - T. Herrmann (UNIO)
  - O. Lange (CS-Rosetta)
  - M. Nilges (CNS/ARIA)
  - Ch. Schwieters (Xplor-NIH)
  - W. Vranken (CCPN)
  - G. Vuister (CCPN/CING)
  - D. Wishart (PROSESS/PANAV/SHIFTX2)
  - G. Montelione (Autostructure/PSVS)
- Observers
- wwPDB representatives

Workshop Agreements

- mmCIF/STAR-like format
- Will include
  - Polymer-sequence specification
  - Chemical shifts used in structure determination
  - Restraints lists (priority for distance, angular, RDC and SAXS/SANS following SAS TF recommendations)
  - Peak lists, if available
- Commitment from all participants to write and read such files
- Ambitious timeline of 12 months

```
save_distance_restraint_list_l1
  sf_category                   distance_restraint_list
  sf_framecode                  distance_restraint_list_l1
  restraint_averaging_type      r-6
  potential_type                Log-normal

  loop_
      _Distance_restraint.restraint_ID
      _Distance_restraint.chain_code_1
      _Distance_restraint.res_number_1
      _Distance_restraint.res_type_1
      _Distance_restraint.atom_name_1
      _Distance_restraint.chain_code_2
      _Distance_restraint.res_number_2
      _Distance_restraint.res_type_2
      _Distance_restraint.atom_name_2
      _Distance_restraint.weight
      _Distance_restraint.target_val
      _Distance_restraint.target_val_err

    1   A 11  Ala  MB  A 77  Val  HN  1.00 3.7  0.1
    1   A 11  Ala  MB  A 87  Leu  HN     .   .    .
    1   A 13  Thr  MG2 A 77  Val  HN  1.00 4.7  0.1
    1   A 13  Thr  MG2 A 87  Leu  HN     .   .    .
    5   A  7  Ser  HB2 A 19  Asp  HBx 1.00 2.8  0.1
    8   A  7  Ser  HB2 A 19  Asp  HBy 1.00 4.4  0.1
  stop_
```

Workshop Agreements (cont.)

- Dictionary versioning
- Namespaces for individual developers
- Ownership lies with the community
- Working group to develop a detailed format proposal
  - G. Vuister (chair)
  - M. Nilges, Ch. Schwieters, P. Güntert, T. Herrmann, D. Case
  - E. Ulrich, J. Westbrook, A. Gutmanas, R. Fogh
- Publish brief workshop report
- Authoritative publication once format specified and implemented

Workshop Follow-up and Status

- Advanced draft proposal
  - Detailed - e.g., recommends using uppercase for atom names in standard residues, includes mechanism for versioning and namespaces
- Currently discussing three issues:
  - STAR or CIF syntax
  - Amino-acid variants
  - Dealing with RDCs, tensors, etc.
- Follow-up workshop at Rutgers in January 2015
  - Aim: developers to try and read in files produced by each other's software
- Publication in preparation

Outreach

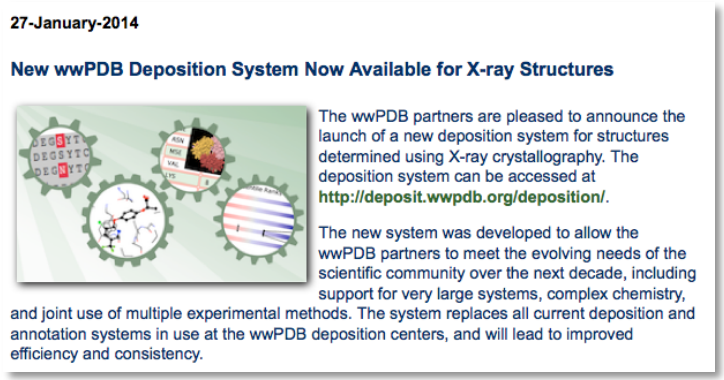
Haruki Nakamura

WORLDWIDE PDB
PROTEIN DATA BANK

wwpdb.org

Communicating with PDB Users

- Online
  - wwPDB.org
  - Help desk: info@wwpdb.org
  - Facebook, mailing list
- In person
  - Conferences, workshops
  - Publications, posters
- Recent news
  - Improved Representation of Large Structures
  - PDB Reaches a New Milestone: 100,000+ Entries
  - wwPDB X-ray Validation Reports Added to PDB Archive
  - New wwPDB Deposition System Now Available for X-ray Structures



Redesign of wwPDB.org

2014: International Year of Crystallography

This calendar was distributed to classrooms and at scientific and educational meetings worldwide, including the IYCr Opening Ceremony

The images are available, free for use, at wwpdb.org

May 14: PDB Reaches 100,000+ Entries

Hard data

It has been no small feat for the Protein Data Bank to stay relevant for 100,000 structures.

Sherlock Holmes understood: "It is a capital mistake," he said, "to theorise before one has data." Data are the lifeblood of science, the foundation of innovation. Behind every great discovery is a pile of data; but, crucially, it should not be too far behind.

For more than four decades, the Protein Data Bank (PDB) has been where structural biologists keep their data close. Nearly every biology-publishing journal in the world, *Nature* included, requires protein structures to be deposited in the PDB before publication.

So there was considerable worry at the database when *Nature* accepted a molecular map of HIV's capsid protein shell last year (G. Zhao *et al. Nature* **497**, 643–646; 2013). The multimillion-atom complex was larger than anything then in the PDB, and the database's team had to devise a way to make the data dump available (and useful) at short notice.

Thus it goes at the PDB — whose trove surpasses 100,000 structures this week (see page 265) — and other long-running archives that have managed to stay relevant and essential. It is not easy. Just ask the scientists, funders, technicians and others who shepherd them.

Money is often the limiting factor. Computer storage and processing power may be getting cheap as chips, but much of the expense is in paying the people (many of them highly trained scientists) who organize and verify data entries, and engage scientific communities.

There are many ways for a database to stay in the black. The three-decades-old GenBank, a clearing house for DNA sequences, is funded directly by the US government's support of the National Center for Biotechnology Information (NCBI). By contrast, the 50-year-old Cambridge Structural Database, which stores 700,000 small-molecule structures, gets by on support from industry and around 1,300 institutes.

The PDB is actually hosted by several organizations that provide access to the same data trove, each funded independently. Gerard Kleywegt, who heads the European franchise at the European Bioinformatics Institute (EBI) in Hinxton, UK, says that healthy competition between his portal and others in the United States and Japan helps him to get grants, and keeps the database pertinent. Scientists "vote with their mouse clicks", he says. "They go to the place where they get the best answer for their questions."

In the 1970s, protein structures were consumed by a small community of X-ray crystallographers interested in the nitty-gritty of individual enzymes. Now scientists use a range of techniques to determine structures, and researchers of many stripes want to know how proteins behave in a larger context, such as in a malignant cancer cell. A database must change with the times, or face extinction.

The closure of a database is not so awful — as long as its useful information remains available elsewhere. In 2011, NCBI announced that it was mothballing a database that collected information about protein fragments used in proteomics experiments. A competing database run by the EBI has since swallowed up those data. But with 100,147 structures (as *Nature* went to press), and growing at about 200 per week, the PDB, at least, shows no sign of folding. ■

NATURE.COM
To comment online, click on Editorials at: go.nature.com/xhmqv

© 2014 Macmillan Publishers Limited. All rights reserved

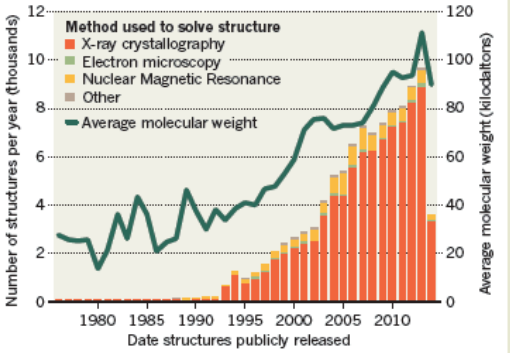
SOURCE: PDBE

TREND WATCH

A digital compendium of proteins and other biomolecules has surpassed 100,000 entries, with the release of 219 new structures on 14 May. The Protein Data Bank (PDB) was started in 1971 to store three-dimensional structural data down to the atomic level. Then and now, scientists mapped most proteins using X-ray crystallography, but they are increasingly using other tools, such as nuclear magnetic resonance and electron microscopy. See also page 260.

ONE HUNDRED THOUSAND PROTEIN STRUCTURES

Biomolecular structures stored in the Protein Data Bank are getting bigger and more complex.



NIGMS Feedback Loop Blog

A catalyst for interaction with the scientific community

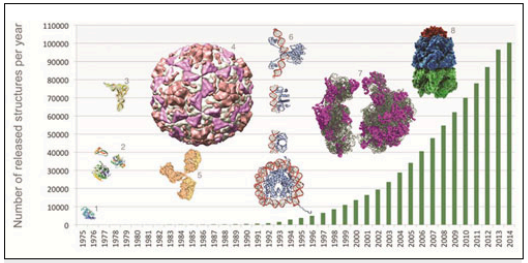
Protein Data Bank Passes 100,000-Structure Mark

Posted by Dr. Ward Smith on May 20, 2014
Post a Comment | View Comments (1) ↓

The Protein Data Bank (PDB) just passed a major threshold—the release of its 100,000th entry. This free online repository of experimentally determined protein and nucleic acid structures, which NIGMS and other parts of NIH have helped fund since 1978, facilitates atomic-level insight into protein structure and function. PDB is widely used by the scientific community to study basic biological processes like transcription, translation, enzymology, bioenergetics and metabolism and also for more medically oriented investigations into disease mechanisms and drug design.

The latest update brings the total number of PDB entries to 100,147.

In addition to scientists, students and educators use the digital resource for their own explorations of protein structure, function and interactions as well as to gain greater knowledge about biology.



View larger image

Number of structures available in the PDB per year, with selected examples. For details, see http://www.eurekalert.org/multimedia/pub/73206.php?from=267554.


日本経済新聞

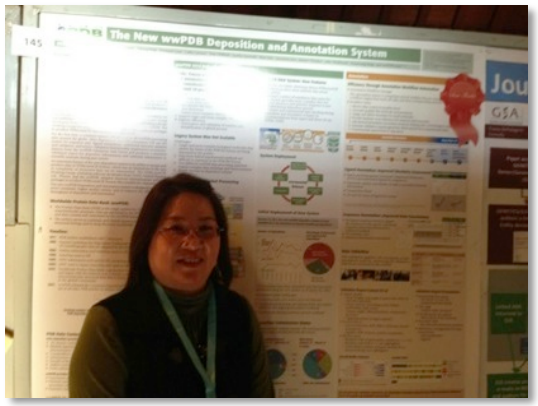
ネットが開く科学研究

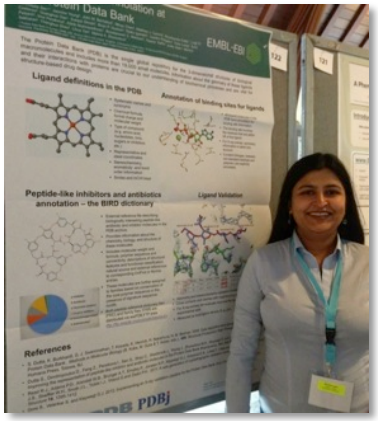
先端データ公開／東大授業無料で

Nikkei Newspaper 1st July 2014

Meetings

- Asian Crystallographic Association (AsCA) 2013
  - Dec 7-10, Hong Kong
- International Biocuration Conference 2014
  - April 6-9, Toronto, Canada
- American Crystallographic Association 2014
  - May 24-28, Albuquerque, NM
  - D&A System Workshop
- IUCr 2014
  - August 5-12, Montreal, Canada
  - wwPDB Exhibit Stand





IYCr Events

Training

Classroom Outreach

Science Festivals

PDBx/mmCIF Programmer Workshop

EMBL-EBI, Hinxton, 20-21 November, 2013

Publications

Structure
Perspective

How Community Has Shaped the Protein Data Bank

Helen M. Berman,[1,*] Gerard J. Kleywegt,[2] Haruki Nakamura,[3] and John L. Markley[4]
[1]RCSB PDB, Center for Integrative Proteomics Research and Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, NJ USA 08854
[2]PDBe, European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK
[3]PDBj, Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan
[4]BioMagResBank, Department of Biochemistry, University of Wisconsin-Madison, Madison, WI 53706 USA
*Correspondence: berman@rcsb.rutgers.edu
http://dx.doi.org/10.1016/j.str.2013.07.010

Following several years of community discussion, the Protein Data Bank (PDB) was established in 1971 as a public repository for the coordinates of three-dimensional models of biological macromolecules. Since then, the number, size, and complexity of structural models have continued to grow, reflecting the productivity of structural biology. Managed by the Worldwide PDB organization, the PDB has been able to meet increasing demands for the quantity of structural information and of quality. In addition to providing unrestricted access to structural information, the PDB also works to promote data standards and to raise the profile of structural biology with broader audiences. In this perspective, we describe the history of PDB and the many ways in which the community continues to shape the archive.

Structure
Meeting Review

Report of the wwPDB Small-Angle Scattering Task Force: Data Requirements for Biomolecular Modeling and the PDB

Jill Trewhella,[1,*] Wayne A. Hendrickson,[2] Gerard J. Kleywegt,[3] Andrej Sali,[4] Mamoru Sato,[5] Torsten Schwede,[6,7] Dmitri I. Svergun,[8] John A. Tainer,[9,10] John Westbrook,[11] and Helen M. Berman[11]
[1]School of Molecular Bioscience, The University of Sydney, NSW 2006, Australia
[2]Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032, USA
[3]European Molecular Biology Laboratory–European Bioinformatics Institute, Cambridge CB10 1SD, UK
[4]Departments of Bioengineering and Therapeutic Sciences, and Pharmaceutical Chemistry, California Institute for Quantitative Biosciences, University of California, San Francisco, San Francisco, CA 94143, USA
[5]Graduate School of Medical Life Science, Yokohama City University, Yokohama, Kanagawa Prefecture 236-0027, Japan
[6]Biozentrum, Universität Basel, University of Basel, 4003 Basel, Switzerland
[7]SIB Swiss Institute of Bioinformatics, 4056 Basel, Switzerland
[8]European Molecular Biology Laboratory, Hamburg Outstation, 22603 Hamburg, Germany
[9]Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94704, USA
[10]Department of Integrative Structural and Computational Biology, The Skaggs Institute for Chemical Biology, The Scripps Research Institute, LaJolla, CA 92037, USA
[11]Department of Chemistry and Chemical Biology, Rutgers University, New Brunswick, NJ 07102, USA
*Correspondence: jill.trewhella@sydney.edu.au
http://dx.doi.org/10.1016/j.str.2013.04.020

Improving the Representation of Peptide-Like Inhibitor and Antibiotic Molecules in the Protein Data Bank

Shuchismita Dutta,[1] Dimitris Dimitropoulos,[2] Zukang Feng,[1] Irina Persikova,[1] Sanchayita Sen,[3] Chenghua Shao,[1] John Westbrook,[1] Jasmine Young,[1] Marina A. Zhuravleva,[1] Gerard J. Kleywegt,[3] Helen M. Berman[1]
[1] *RCSB Protein Data Bank, Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, NJ 08854-8076*
[2] *RCSB Protein Data Bank, San Diego Supercomputer Center and Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, CA 92093-0537*
[3] *Protein Data Bank in Europe (PDBe), European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK*

Received 18 October 2013; accepted 27 October 2013
Published online 30 October 2013 in Wil

ABSTRACT:
With the accumulation of a large numbe of molecules in the Protein Data Bank (I need on occasion to review and improve tation. The Worldwide PDB (wwPDB) p periodically updated various aspects of s representation to improve the integrity a of the archive. The remediation effort de was focused on improving the representa like inhibitor and antibiotic molecules sc be easily identified and analyzed. Peptid

Correspondence to: Shuchismita Dutta, RCSB Protein of Chemistry and Chemical Biology, Rutgers, The S Jersey, 174 Frelinghuysen Road, Piscataway, NJ 088 sdutta@rcsb.rutgers.edu
Contract grant sponsor: NSF DBI
Contract grant number: 0829586 (to RCSB PDB)
Contract grant sponsors: NIGMS, DOE, NLM, NCI, N RCSB PDB)
Contract grant sponsor: EMBL-EBI (to PDBe)
Contract grant sponsor: Wellcome Trust
Contract grant number: 088944 (to PDBe)
Contract grant sponsor: BBSRC
Contract grant numbers: BB/J007471/1, BB/I02576X/1, and
Contract grant sponsor: NIGMS
Contract grant number: 1RO1 GM079429-01A1 (to PDB
Contract grant sponsor: EU
Contract grant number: 284209 (to PDBe)
©2013 The Authors Biopolymers Published by Wiley Peri
This is an open access article under the terms of the Cre tion License, which permits use, distribution and repro provided the original work is properly cited.

Biopolymers Volume 101 / Number 6

J Comput Aided Mol Des
DOI 10.1007/s10822-014-9770-y

The Protein Data Bank archive as an open data resource

Helen M. Berman · Gerard J. Kleywegt · Haruki Nakamura · John L. Markley

Received: 30 April 2014 / Accepted: 23 June 2014
© The Author(s) 2014. This article is published with open access at Springerlink.com

Abstract The Protein Data Bank archive was established in 1971, and recently celebrated its 40th anniversary (Berman et al. in Structure 20:391, 2012). An analysis of interrelationships of the science, technology and community leads to further insights into how this resource evolved into one of the oldest and most widely used open-access data resources in biology.

Keywords Protein Data Bank · Protein structure · Biomacromolecules · Data archive

Early history of protein crystallography

In 1934, Dorothy Crowfoot (Hodgkin) together with John D. Bernal at Cambridge University obtained the first diffraction pattern of the protein pepsin [1]. Bernal had trained in crystallography at the Royal Institution in London with Sir William Henry Bragg, who with his son Sir

William Lawrence Bragg founded the field of X-ray crystallography. Bernal went on to establish his own research group in Cambridge. He was a visionary figure in the field earning the nickname "Sage" while still an undergraduate at Cambridge. He had strong views about the interactions of science and society and felt that science had to be useful, in opposition to others who voiced that science should be pure and separated from societal needs [2]. His philosophies continue to influence how crystallographers work and collaborate today. Dorothy Hodgkin went on to Oxford and determined structures of biologically important small molecules as well as proteins, most notably insulin [3, 4]. Max Perutz arrived in Cambridge from Austria in 1936 and began his study of hemoglobin which led to its structure determination in 1959 [5]. Both Hodgkin and Perutz trained large numbers of crystallographers who set up laboratories around the world. John Kendrew arrived at Cambridge's newly formed Medical Research Council Laboratory of Molecular Biology and determined the

Structure
Ways & Means

Recommendations of the wwPDB NMR Validation Task Force

Gaetano T. Montelione,[1,2,*] Michael Nilges,[4,5] Ad Bax,[6] Peter Güntert,[8,9] Torsten Herrmann,[10,11] Jane S. Richardson,[12] Charles D. Schwieters,[7] Wim F. Vranken,[13,14] Geerten W. Vuister,[15] David S. Wishart,[16,17] Helen M. Berman,[3] Gerard J. Kleywegt,[18] and John L. Markley[19]

Publications

letters to the editor

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Comment on *On the propagation of errors* by Jaskolski (2013)

Helen Berman,[a]* Gerard J. Kleywegt,[b] Haruki Nakamura[c] and John L. Markley[d]

[a]Department of Chemistry & Chemical Biology, Center for Integrative Proteomics Research, Rutgers, The State University of New Jersey, 174 Frelinghuysen Road, Piscataway, NJ 08854, USA, [b]PDBe, European Molecular Biology Laboratory–European Bioinformatics Institute, Cambridge CB10 1SD, UK, [c]PDBj, Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka, 565-0871, Japan, and [d]BioMagResBank, Department of Biochemistry, University of Wisconsin-Madison, Madison, WI 53706, USA

Correspondence e-mail: berman@rcsb.rutgers.edu

Received 18 October 2013
Accepted 22 October 2013

The wwPDB responds to the article by Jaskolski [(2013), *Acta Cryst.* D**69**, 1865–1866].

In his paper *On t*
several points abou
Bank. Since 2007,
undertaken several
in an effort to imp
data. This is an ong
paper into account
Our response to
Atom naming. W
naming of the first
as well as it shoul
distribution format
officially sanctioned
by the depositor.
CCDC with whom
Protonation stat
well as the microen
the depositor is con
compounds, residu
collect this informa
Restraint diction
Working Group for
structure-determini
addressing the issu

References

Jaskolski, M. (2013). .

letters to the editor

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Comment on *Timely deposition of macromolecular structures is necessary for peer review* by Joosten *et al.* (2013)

Helen Berman,[a]* Gerard J. Kleywegt,[b] Haruki Nakamura[c] and John L. Markley[d]

[a]Department of Chemistry & Chemical Biology, Center for Integrative Proteomics Research, Rutgers, The State University of New Jersey, 174 Frelinghuysen Road, Piscataway, NJ 08854, USA, [b]PDBe, European Molecular Biology Laboratory–European Bioinformatics Institute, Cambridge CB10 1SD, UK, [c]PDBj, Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka, 565-0871, Japan, and [d]BioMagResBank, Department of Biochemistry, University of Wisconsin-Madison, Madison, WI 53706, USA

Correspondence e-mail: berman@rcsb.rutgers.edu

Received 18 October 2013
Accepted 22 October 2013

The wwPDB responds to the article by Joosten *et al.* [(2013), *Acta Cryst.* D**69**, 2293–2295].

The Worldwide Protein Data Bank (wwPDB) strongly agrees with the overall views expressed by Joosten *et al.* (2013) in their article about timely deposition of macromolecular structures in the Protein Data Bank. In 2010, *Acta Crystallographica Section D* began to require validation reports as part of the manuscript-submission process. In that same year, the wwPDB sent letters to the key journals that publish structures requesting that they require authors to submit wwPDB validation reports at the same time as their manuscripts. In this way, reviewers are able to better evaluate the work. The *Journal of Biological Chemistry*, which is currently the journal that publishes the largest number of papers per year about structures of biological macromolecules, began requiring these reports in 2012.

Joosten *et al.* suggest that it would be helpful to have an option to suppress entry titles at the time of submission to the PDB until the structure is released. Policy matters such as this are regularly reviewed by the wwPDB partners and its Advisory Committee (wwPDB AC). The issue was discussed at our 2013 meeting, and it was agreed that we will make this option available in the new wwPDB Deposition Tool that will be launched early in 2014.

References

Joosten, R. P., Soueidan, H., Wessels, L. F. A. & Perrakis, A. (2013). *Acta Cryst.* D**69**, 2293–2295.

letters to the editor

CrossMark

Acta Crystallographica Section D
Biological Crystallography
ISSN 1399-0047

Response to *On prompt update of literature references in the Protein Data Bank*

Helen M. Berman,[a]* Stephen K. Burley,[a] Gerard J. Kleywegt,[b] Haruki Nakamura[c] and John L. Markley[d]

[a]Department of Chemistry and Chemical Biology, Center for Integrative Proteomics Research, Rutgers, The State University of New Jersey, 174 Frelinghuysen Road, Piscataway, NJ 08854, USA, [b]PDBe, European Molecular Biology Laboratory–European Bioinformatics Institute, Cambridge CB10 1SD, UK, [c]PDBj, Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka, 565-0871, Japan, and [d]BioMagResBank, Department of Biochemistry, University of Wisconsin-Madison, Madison, WI 53706, USA

Correspondence e-mail: berman@rcsb.rutgers.edu

Received 7 June 2014
Accepted 12 September 2014

The wwPDB responds to the article *On the prompt update of literature references in the Protein Data Bank* [Wlodawer (2014), *Acta Cryst.* D**70**, 2779].

The wwPDB receives publication dates and citation information from authors, several journals, the user community, and additionally scans the literature for publications. The journals that currently provide related citation information either in advance or at the time of publication include IUCr Journals (*Acta Crystallographica Sections A–F, IUCrJ, Journal of Applied Crystallography, Journal of Synchrotron Radiation*); *Journal of Biological Chemistry*; *Journal of Molecular Biology*; *Proceedings of the National Academy of Sciences*; *Nature* journals; *Science*; *Proteins: Structure, Function, and Bioinformatics*; and *Protein Science*.

For each weekly update of the PDB archive, all files scheduled for release or modification are checked and validated one final time. Authors may be contacted to resolve any issues that may arise while preparing the entries for release. If this is not accomplished in time for that release cycle, the entry may be scheduled for release at a later date.

The preparation and packaging of the weekly PDB update begins Thursday afternoon (local time at each wwPDB member site), and the files are publicly released on the following Wednesday at 00:00 UTC (Coordinated Universal Time).

The wwPDB encourages journals to provide citation information [paper title, author list, related PDB ID(s), DOI, and publication date] to deposit@wwpdb.org, ideally at least two weeks in advance of publication of a structure report so that the release of the corresponding PDB entry can be timed accordingly.

References

Wlodawer, A. (2014). *Acta Cryst.* D**70**, 2779.

© 2014 International Union of Crystallography