The PDB and *experimental data*

John Westbrook

Rutgers, The State University of New Jersey

IUCr

HRVATSKA
UDRUGA
KRISTALOGRAFA
CROATIAN
ASSOCIATION of
CRYSTALLOGRAPHERS

WORLDWIDE
wwPDB
PROTEIN DATA BANK

www.wwpdb.org

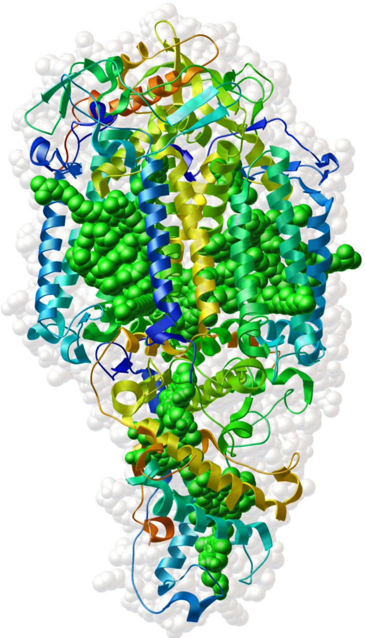
Workshop on Metadata for raw data from X-ray diffraction and other structural techniques

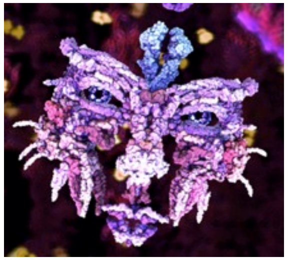
Overview

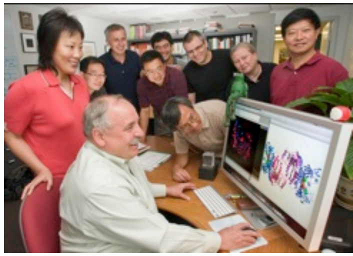
- Current PDB audience and growth
- Range of PDB primary *‘experimental metadata*’
- Challenges for data acquisition and deposition across the PDB view of the structural biology data pipeline
- How we are addressing these challenges

Protein Data Bank (PDB)

- Single global repository for macromolecular structure data (now >111K entries!)
- Archival database - Users depend directly on the PDB; Other Databases present PDB contents
- Our Users: Structural and Computational Specialists, Biophysicists, Biochemists, Biologists, Industrial Scientists, Educators, Students, and the General Public







Worldwide Protein Data Bank

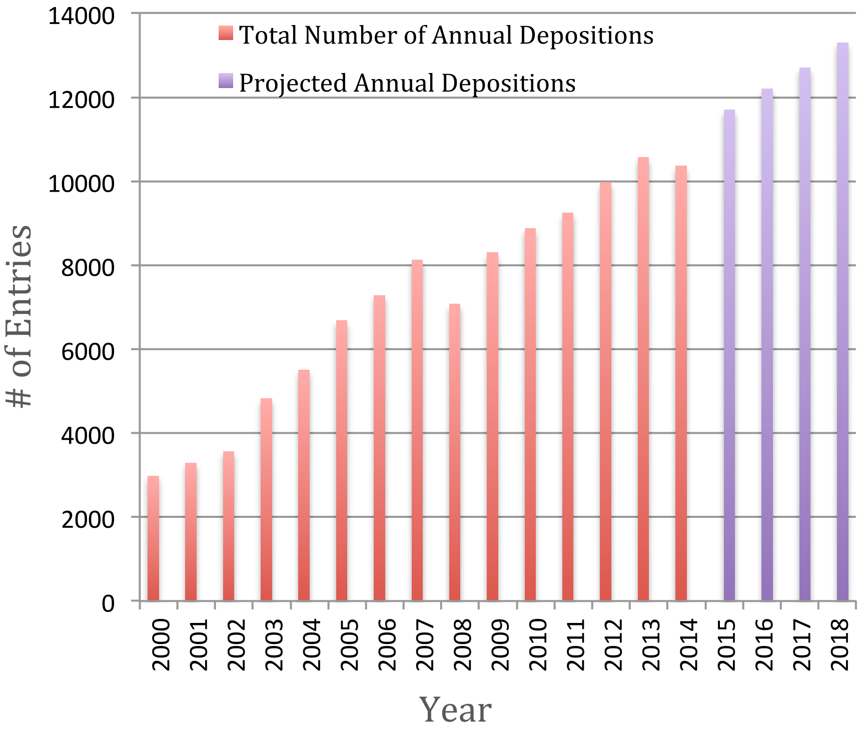
- Ensures data are freely available
- Data Centers
  - RCSB PDB (Research Collaboratory for Structural Bioinformatics)
  - PDBj (Osaka University)
  - PDBe (EMBL-EBI)
  - BioMagResBank (University Wisconsin, Madison)
- Institutional agreement
- Formalized procedures for deposition, validation, metadata representation, and annotation
- Each data center provides unique delivery services

WORLDWIDE wwPDB PROTEIN DATA BANK

RCSB PDB
PROTEIN DATA BANK
BMRB
PDBe
Protein Data Bank in Europe
PDBj
Protein Data Bank Japan

Archive Growth

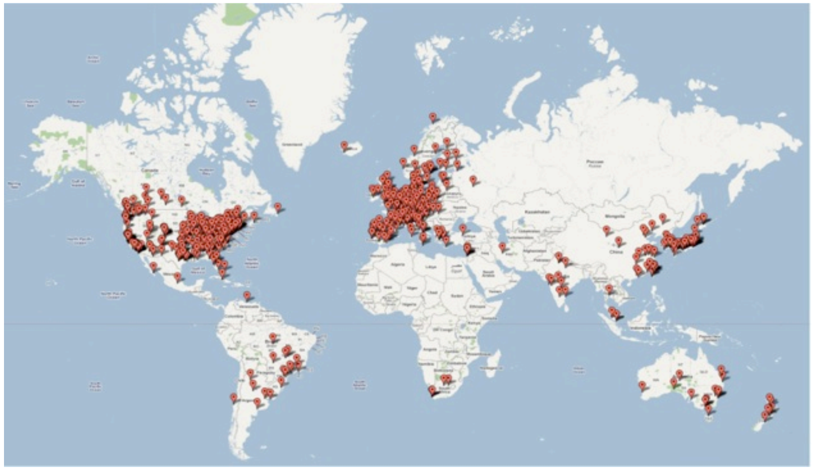
Growing Number of PDB Depositions



As of 2014, ~50% increase in the number of global depositions since 2008

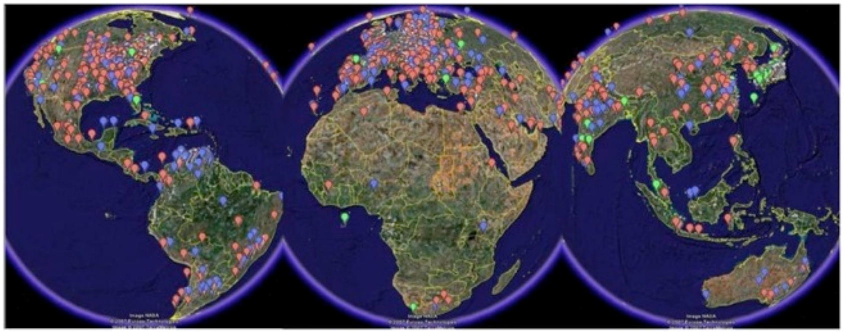
PDB Depositors

>800 new entries/month



PDB Users

FTP and RSYNC Download Traffic in 2014:
505 million downloads



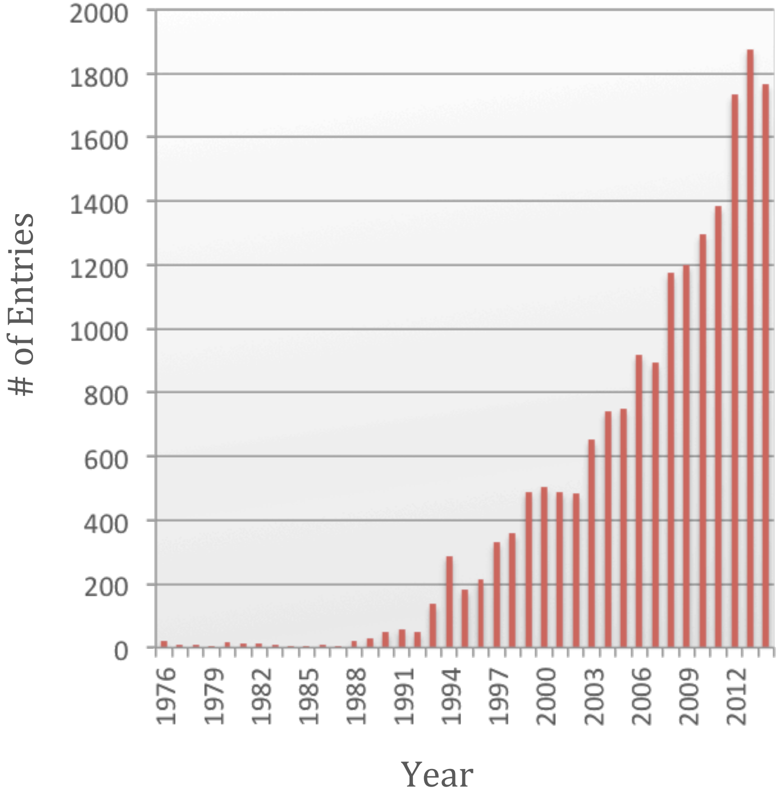
🟥 RCSB PDB
347 million

🟦 PDBe
100 million

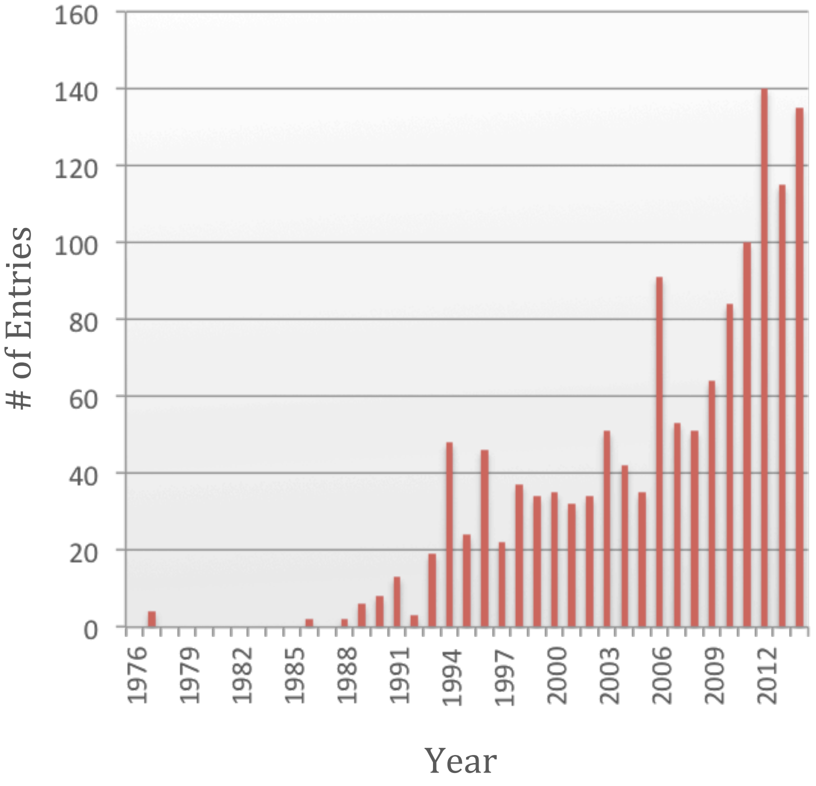
🟩 PDBj
58 million

Increasing Complexity

Number of new ligands present in PDB entries released per year



Number of entries with peptide-like inhibitors/antibiotics released per year

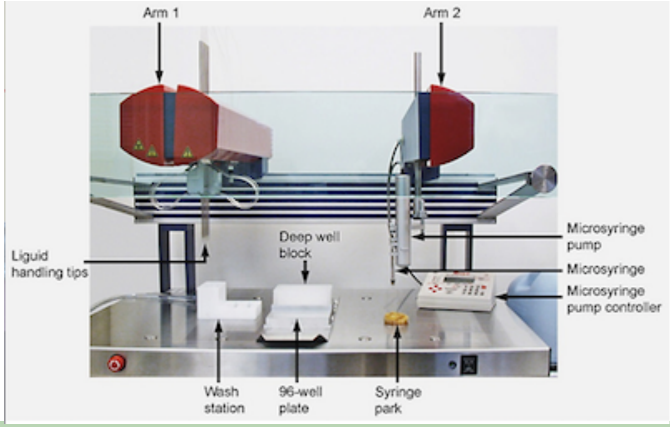


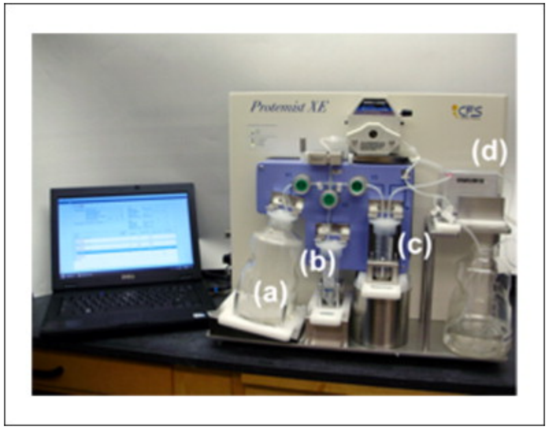
Primary Experimental Content

…at the beginning of the PDB pipeline

Sample Details

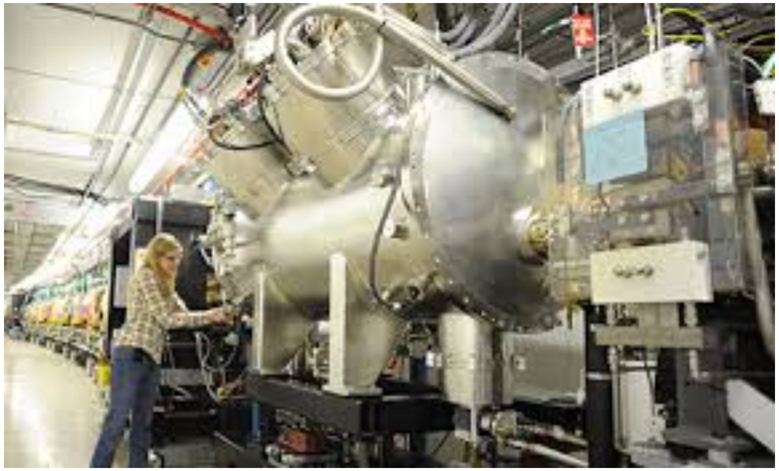
- Sample composition
- Chemical and molecular descriptions
- Source, production details, biological role
- Crystallization conditions
- Specimen preparation details



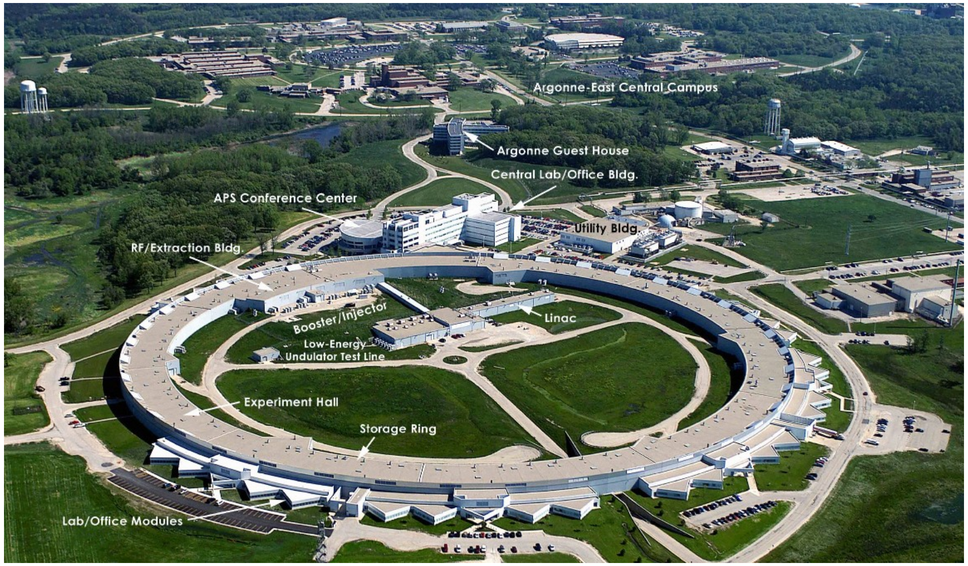


Data Collection Details

- Instrumentation details
- Sample handling and collection conditions
- Data collection protocol



Skip a few steps …



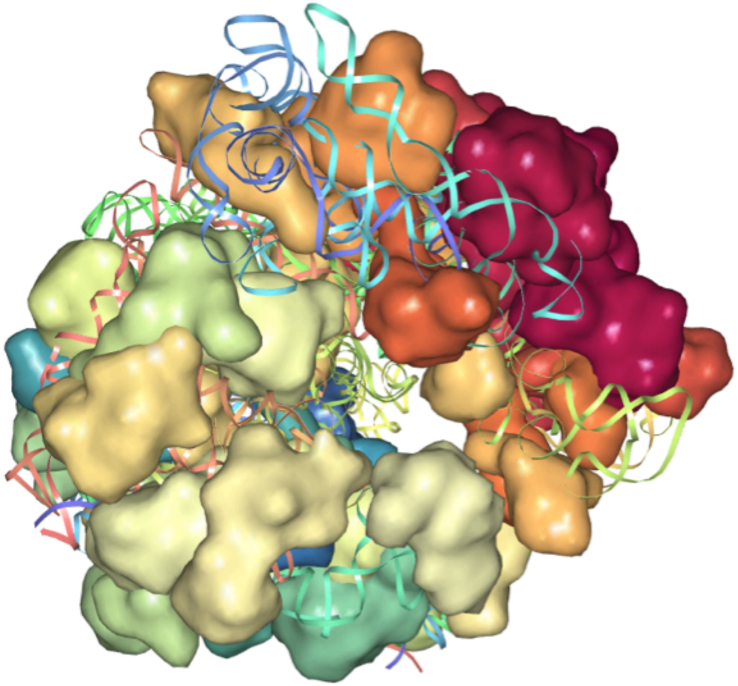
Data Processing Steps

- Summary statistics
- Software tools
- Processed data

Processed Data in the PDB Repository

- Structure factor data sets for ~89% of current X-ray entries
- ~10K additional data sets containing
  - Derivative and multiple wavelength data sets
  - Intermediate phasing data
  - Map coefficients
  - Unmerged intensities

PDB Data Acquisition

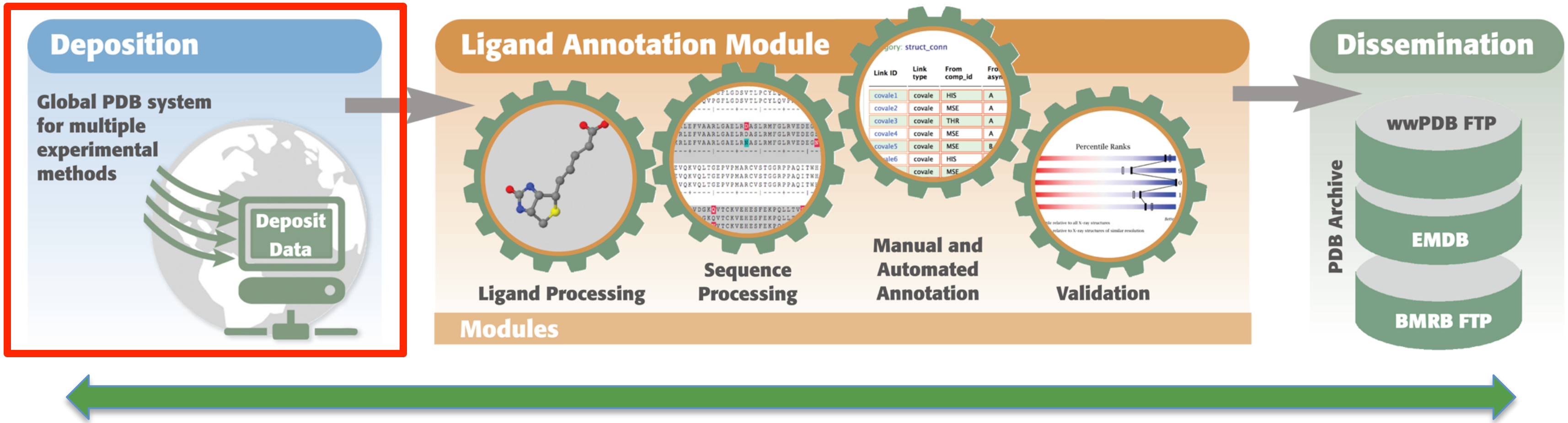


Model Easy

Sample Harder

Go figure….

New Deposition System



Deposition

Global PDB system for multiple experimental methods

Deposit Data

Ligand Annotation Module

Ligand Processing

Sequence Processing

Manual and Automated Annotation

Validation

Modules

Dissemination

PDB Archive

wwPDB FTP

EMDB

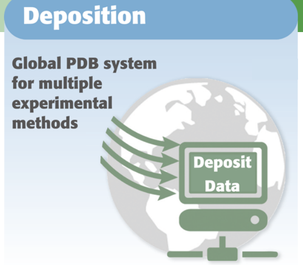
BMRB FTP

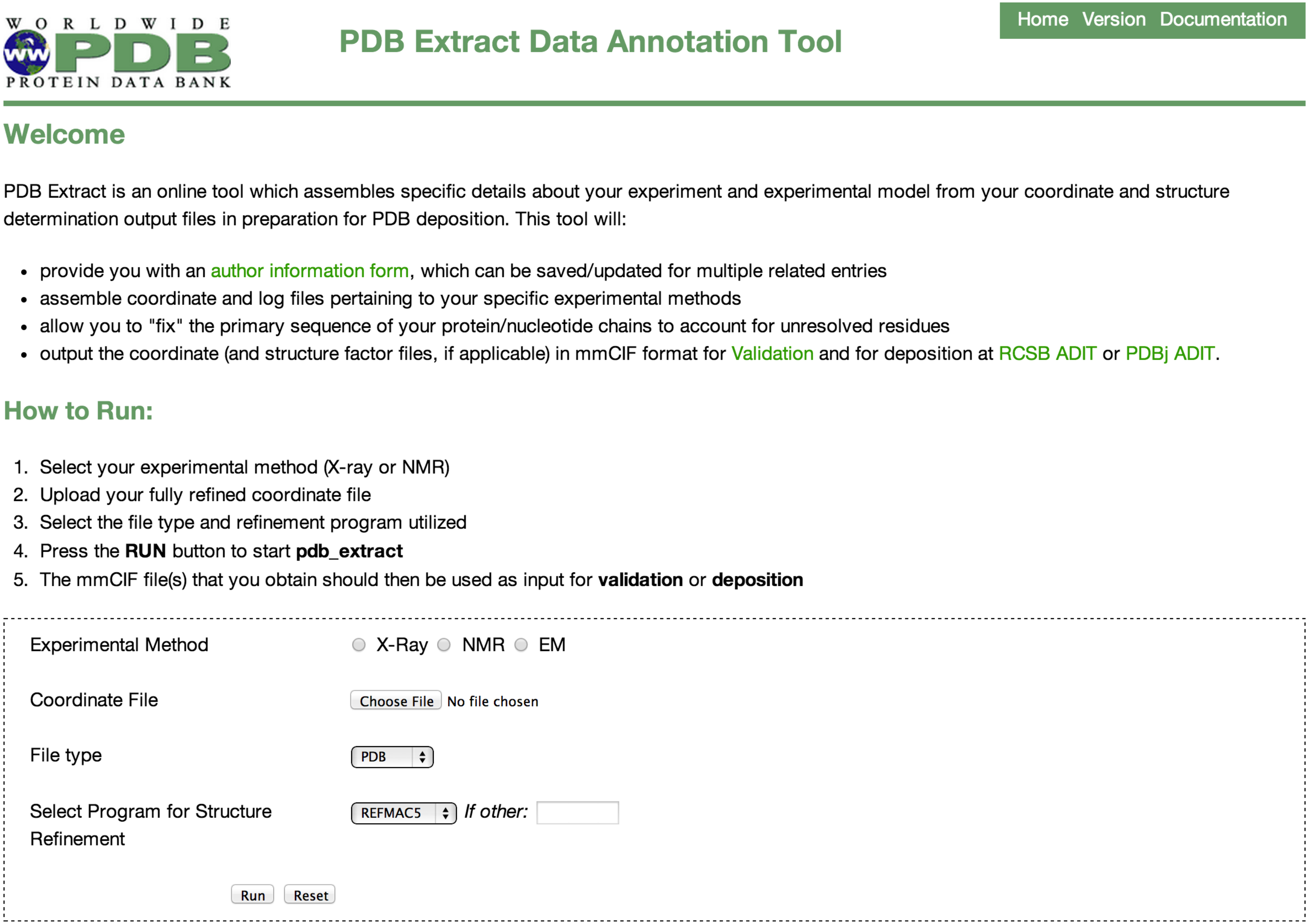
mmCIF/PDBx

End-to-end support for PDBx/mmCIF metadata

Data Harvesting Site

pdb-extract.wwpdb.org/





PDB Extract Data Annotation Tool

Welcome

PDB Extract is an online tool which assembles specific details about your experiment and experimental model from your coordinate and structure determination output files in preparation for PDB deposition. This tool will:

- provide you with an author information form, which can be saved/updated for multiple related entries
- assemble coordinate and log files pertaining to your specific experimental methods
- allow you to "fix" the primary sequence of your protein/nucleotide chains to account for unresolved residues
- output the coordinate (and structure factor files, if applicable) in mmCIF format for Validation and for deposition at RCSB ADIT or PDBj ADIT.

How to Run:

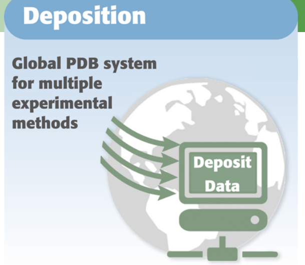
1. Select your experimental method (X-ray or NMR)
2. Upload your fully refined coordinate file
3. Select the file type and refinement program utilized
4. Press the **RUN** button to start **pdb_extract**
5. The mmCIF file(s) that you obtain should then be used as input for **validation** or **deposition**

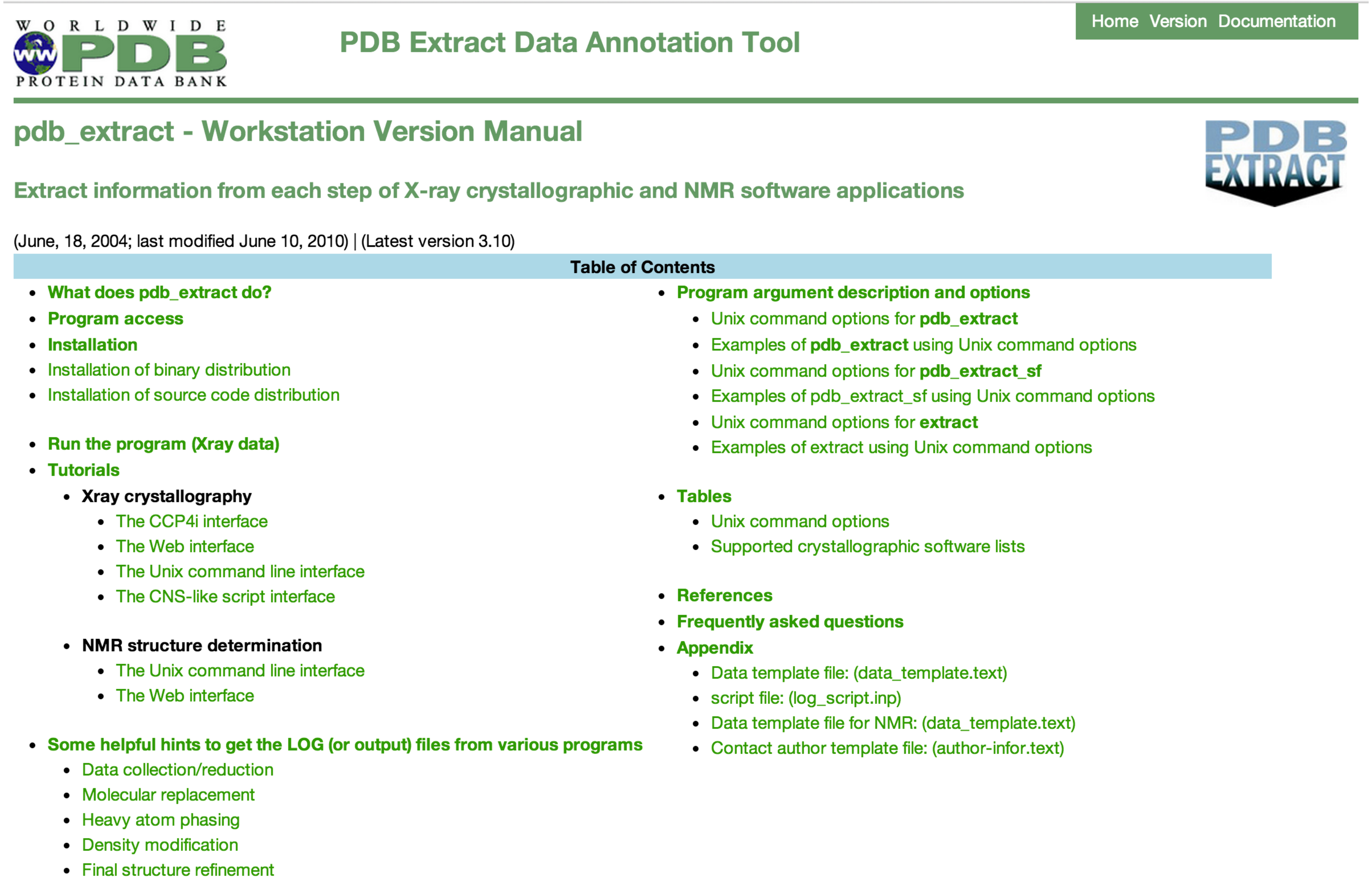
Experimental Method	○ X-Ray ○ NMR ○ EM
Coordinate File	Choose File No file chosen
File type	PDB
Select Program for Structure Refinement	REFMAC5 *If other:*

Run Reset

Data Harvesting Site

pdb-extract.wwpdb.org/





WORLDWIDE PDB PROTEIN DATA BANK

PDB Extract Data Annotation Tool

Home Version Documentation

pdb_extract - Workstation Version Manual

PDB EXTRACT

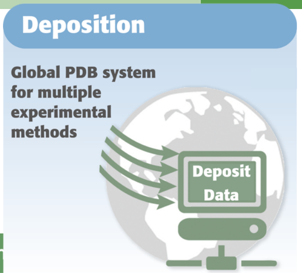
Extract information from each step of X-ray crystallographic and NMR software applications

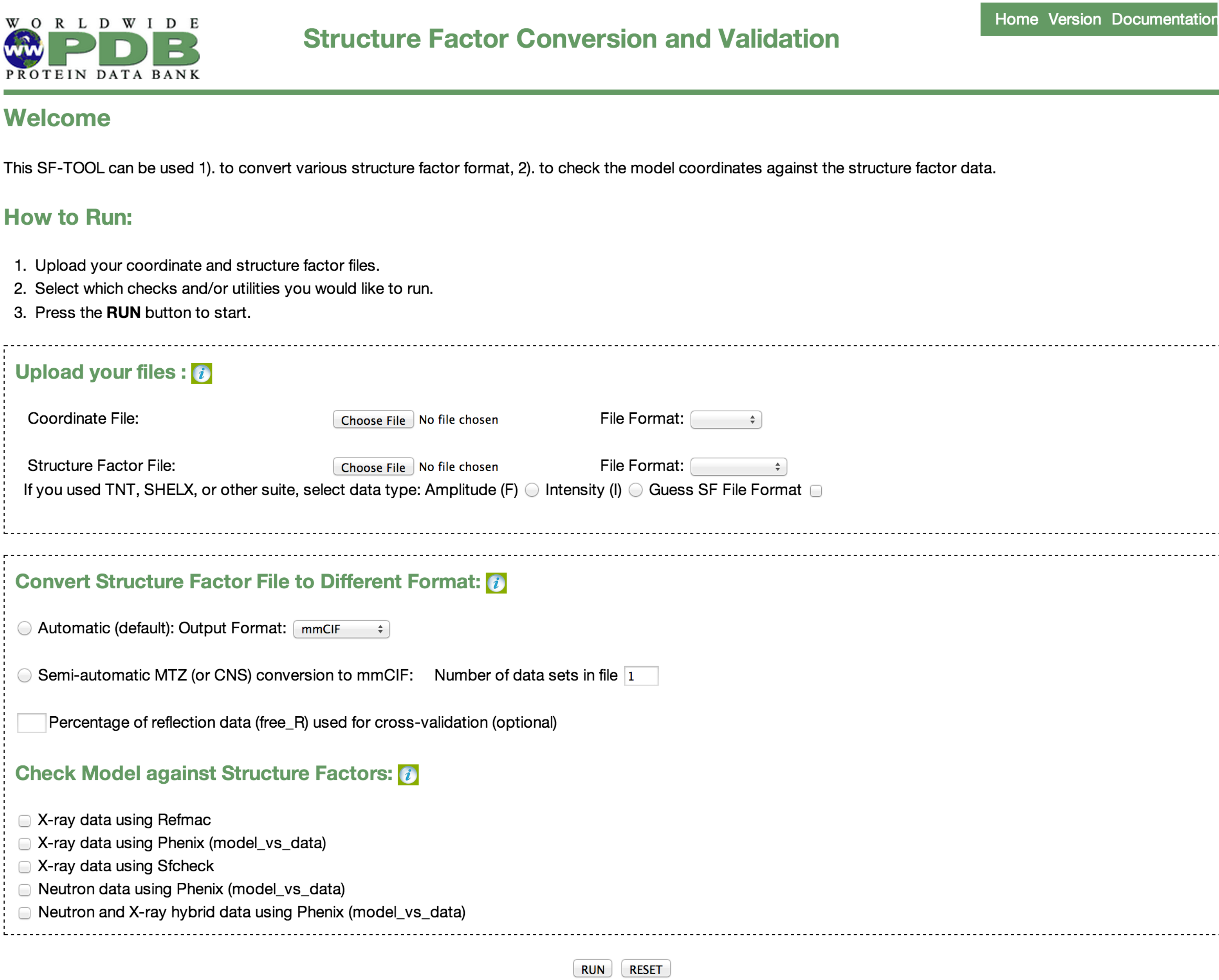
(June, 18, 2004; last modified June 10, 2010) | (Latest version 3.10)

Table of Contents

Structure Factor Utilities

sf-tool.wwpdb.org/





Structure Factor Conversion and Validation

Welcome

This SF-TOOL can be used 1). to convert various structure factor format, 2). to check the model coordinates against the structure factor data.

How to Run:

1. Upload your coordinate and structure factor files.
2. Select which checks and/or utilities you would like to run.
3. Press the **RUN** button to start.

Upload your files :

Coordinate File: Choose File No file chosen File Format:

Structure Factor File: Choose File No file chosen File Format:

If you used TNT, SHELX, or other suite, select data type: Amplitude (F) ○ Intensity (I) ○ Guess SF File Format □

Convert Structure Factor File to Different Format:

○ Automatic (default): Output Format: mmCIF

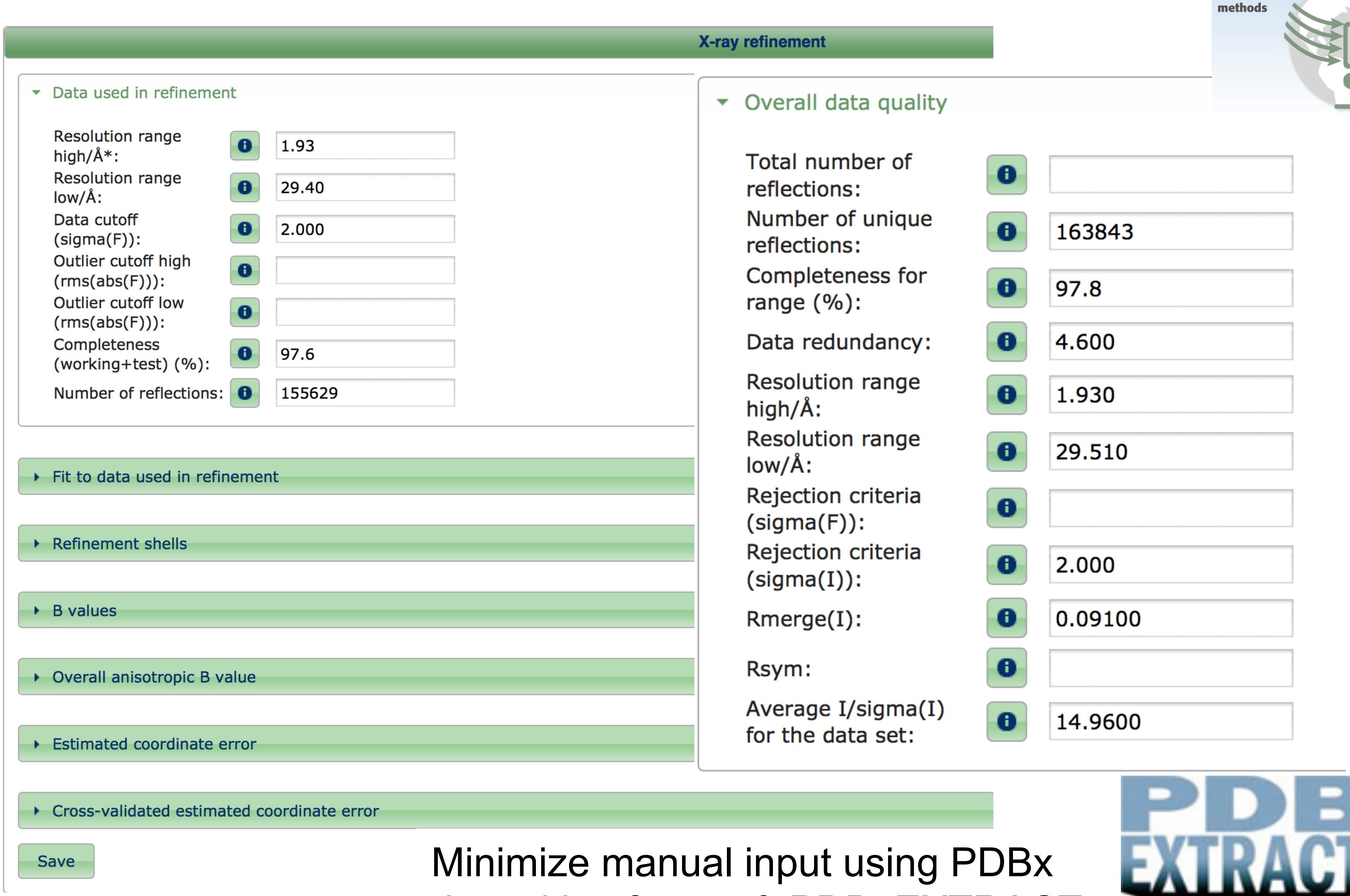
○ Semi-automatic MTZ (or CNS) conversion to mmCIF: Number of data sets in file 1

□ Percentage of reflection data (free_R) used for cross-validation (optional)

Check Model against Structure Factors:

- □ X-ray data using Refmac
- □ X-ray data using Phenix (model_vs_data)
- □ X-ray data using Sfcheck
- □ Neutron data using Phenix (model_vs_data)
- □ Neutron and X-ray hybrid data using Phenix (model_vs_data)

Data Entry Forms



X-ray refinement

▾ Data used in refinement

Resolution range high/Å*:	1.93
Resolution range low/Å:	29.40
Data cutoff (sigma(F)):	2.000
Outlier cutoff high (rms(abs(F))):	
Outlier cutoff low (rms(abs(F))):	
Completeness (working+test) (%):	97.6
Number of reflections:	155629

▸ Fit to data used in refinement

▸ Refinement shells

▸ B values

▸ Overall anisotropic B value

▸ Estimated coordinate error

▸ Cross-validated estimated coordinate error

Save

▾ Overall data quality

Total number of reflections:	
Number of unique reflections:	163843
Completeness for range (%):	97.8
Data redundancy:	4.600
Resolution range high/Å:	1.930
Resolution range low/Å:	29.510
Rejection criteria (sigma(F)):	
Rejection criteria (sigma(I)):	2.000
Rmerge(I):	0.09100
Rsym:	
Average I/sigma(I) for the data set:	14.9600

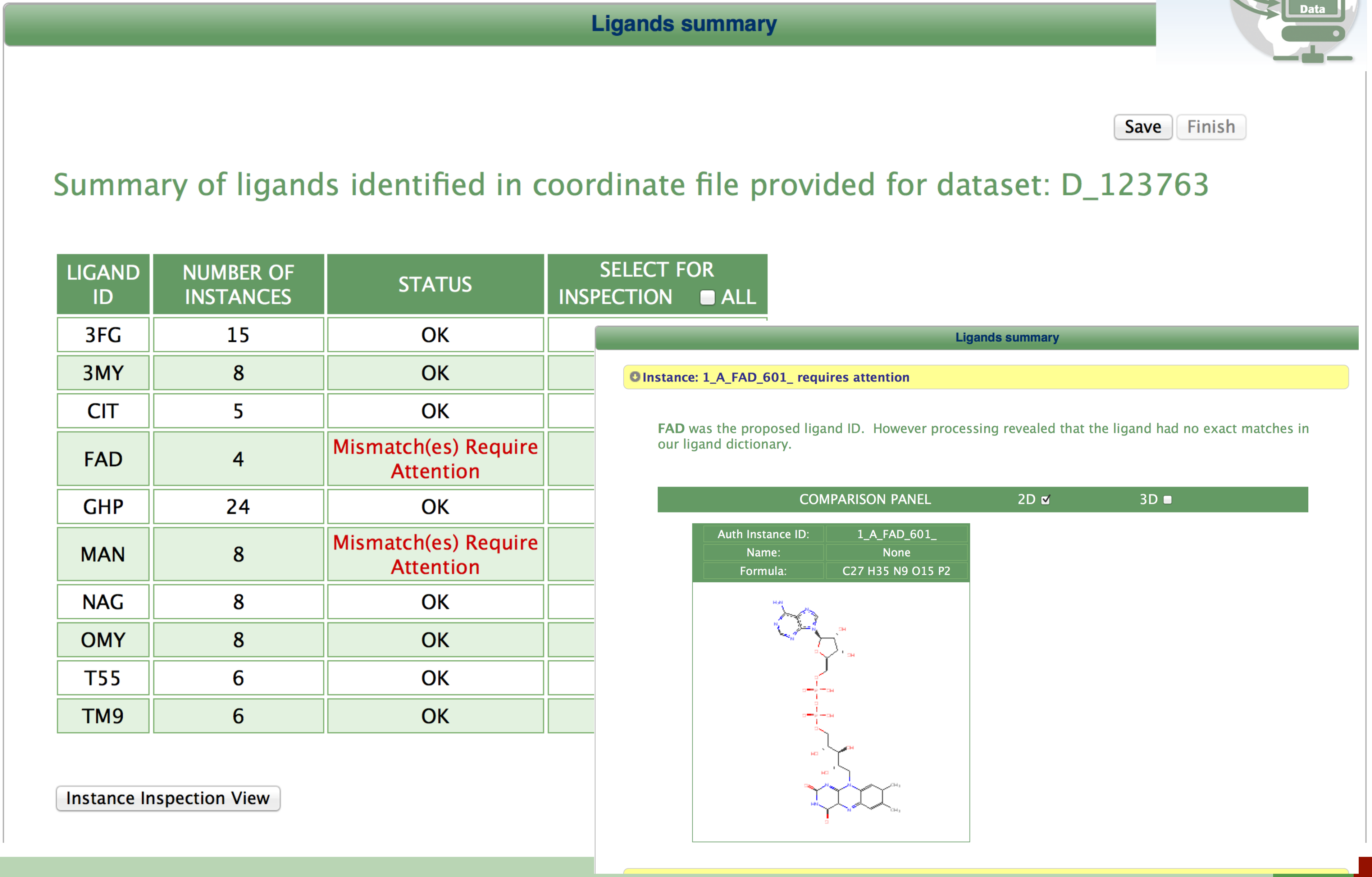
Minimize manual input using PDBx deposition format & PDB_EXTRACT

Chemical Assignment

Deposition

Global PDB system for multiple experimental methods

Deposit Data



Ligands summary

Save | Finish

Summary of ligands identified in coordinate file provided for dataset: D_123763

LIGAND ID	NUMBER OF INSTANCES	STATUS	SELECT FOR INSPECTION □ ALL
3FG	15	OK	
3MY	8	OK	
CIT	5	OK	
FAD	4	Mismatch(es) Require Attention	
GHP	24	OK	
MAN	8	Mismatch(es) Require Attention	
NAG	8	OK	
OMY	8	OK	
T55	6	OK	
TM9	6	OK	

Instance Inspection View

Ligands summary

Instance: 1_A_FAD_601_ requires attention

FAD was the proposed ligand ID. However processing revealed that the ligand had no exact matches in our ligand dictionary.

COMPARISON PANEL 2D ☑ 3D □

Auth Instance ID:	1_A_FAD_601_
Name:	None
Formula:	C27 H35 N9 O15 P2

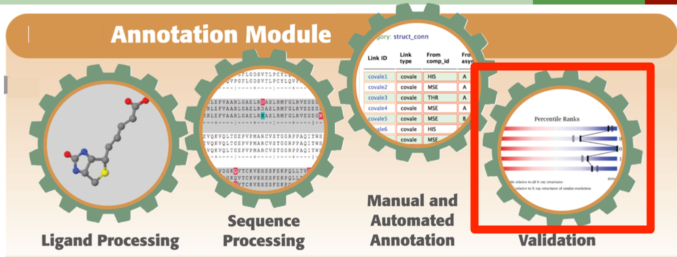
Chemical Reference Data

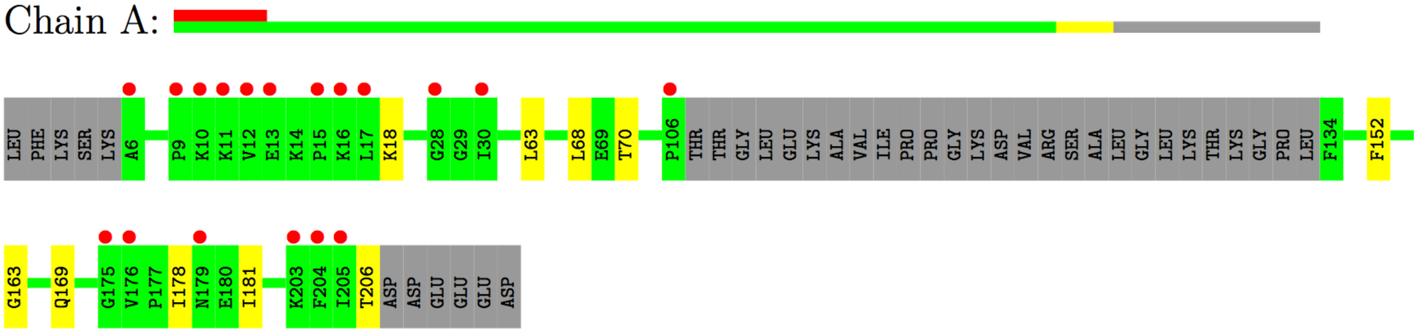
Chemical Component Dictionary

- Library of all polymer and non-polymer chemical components in PDB
  - >20,000 chemical component definitions
  - 400 additional definitions of amino acid protonation variants
- ~700 new components released this year
- ~1700 component definitions updated this year
- Complimentary to the CCP4 monomer library

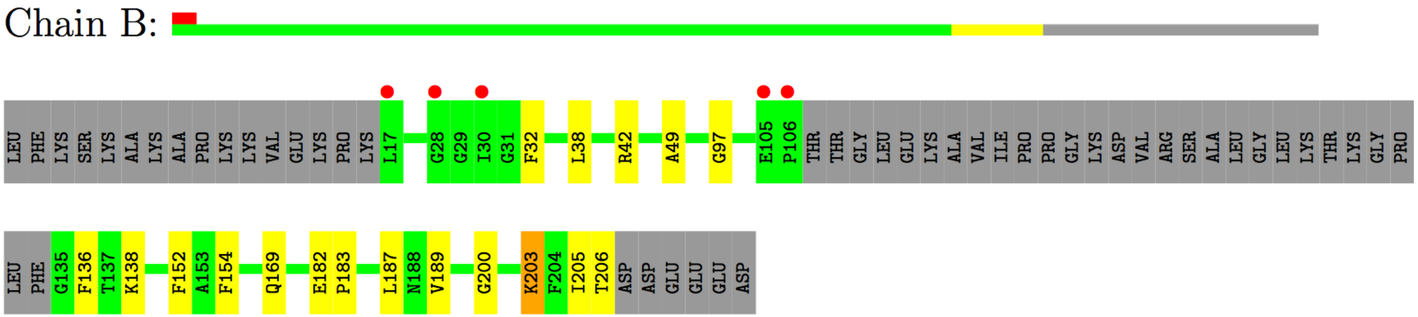
Using the experimental data we collect…

Leveraging Exp. Data in Quality Assessment





- Molecule 1: Photosystem II 22 kDa protein, chloroplastic

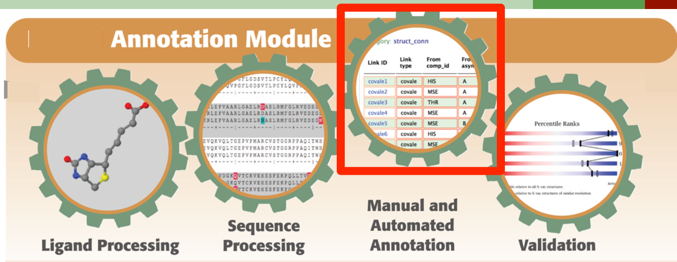


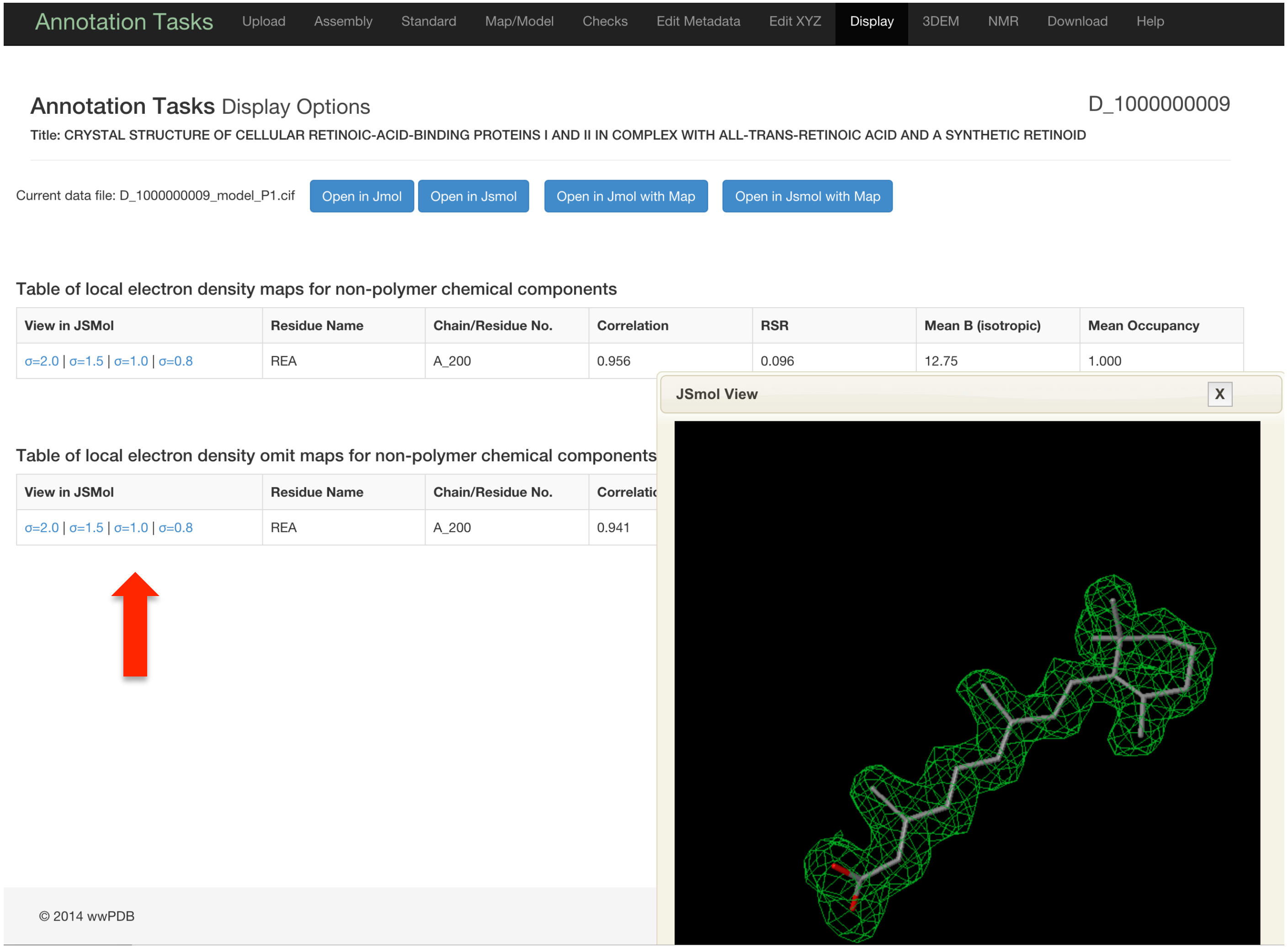
Gray – not modeled

Green, yellow, orange, red – 0,1,2, 3 or more issues

Red dot – poor fit to electron density

Map/Model Fit





Annotation Tasks Display Options

D_1000000009

Title: CRYSTAL STRUCTURE OF CELLULAR RETINOIC-ACID-BINDING PROTEINS I AND II IN COMPLEX WITH ALL-TRANS-RETINOIC ACID AND A SYNTHETIC RETINOID

Current data file: D_1000000009_model_P1.cif

Table of local electron density maps for non-polymer chemical components

View in JSMol	Residue Name	Chain/Residue No.	Correlation	RSR	Mean B (isotropic)	Mean Occupancy
σ=2.0 | σ=1.5 | σ=1.0 | σ=0.8	REA	A_200	0.956	0.096	12.75	1.000

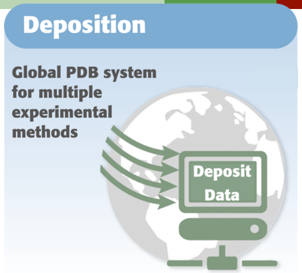
Table of local electron density omit maps for non-polymer chemical components

View in JSMol	Residue Name	Chain/Residue No.	Correlatic
σ=2.0 | σ=1.5 | σ=1.0 | σ=0.8	REA	A_200	0.941

© 2014 wwPDB

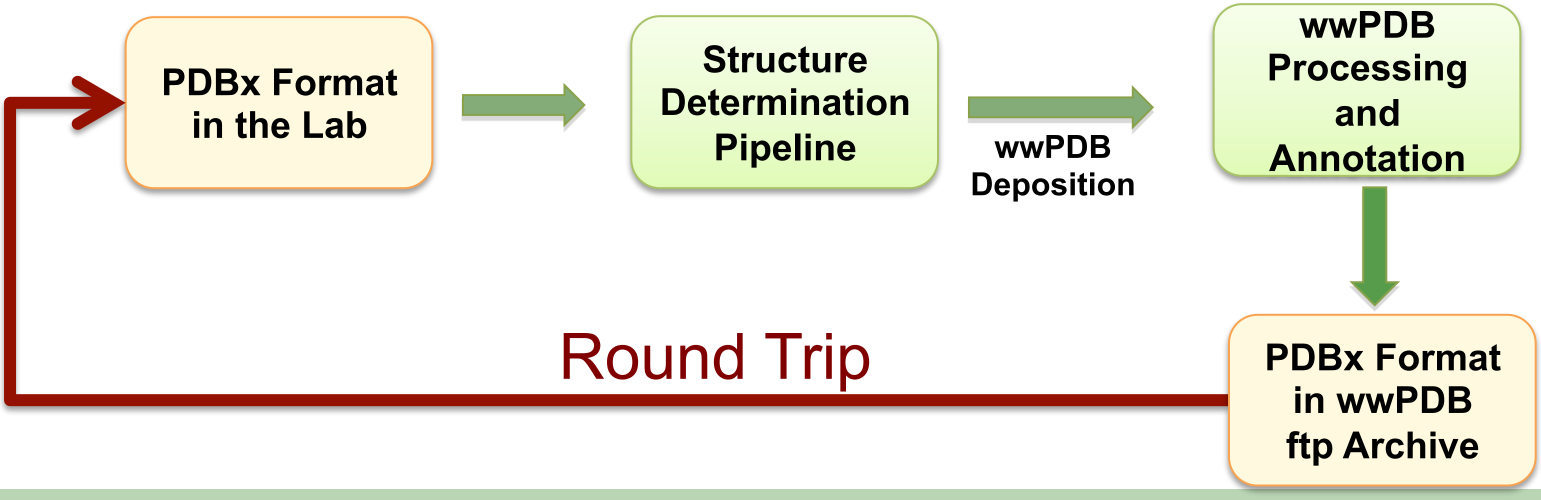
Improving data acquisition …

PDBx Deposition Working Group



PDBx Deposition Working Group
Oct 8 2014 Workshop – EBI/Hinxton

- In 2011, charged with finding a “round trip” single format that can handle complex data not supported by the PDB file format
- Consensus reached on using dictionary-driven PDBx format
- Implementations delivered in January 2013
- Currently working on recommendations for delivery of non-standard chemistry and reflection/intensity data in later 2015



PDBx Format in the Lab → Structure Determination Pipeline → wwPDB Deposition → wwPDB Processing and Annotation → PDBx Format in wwPDB ftp Archive → Round Trip → PDBx Format in the Lab

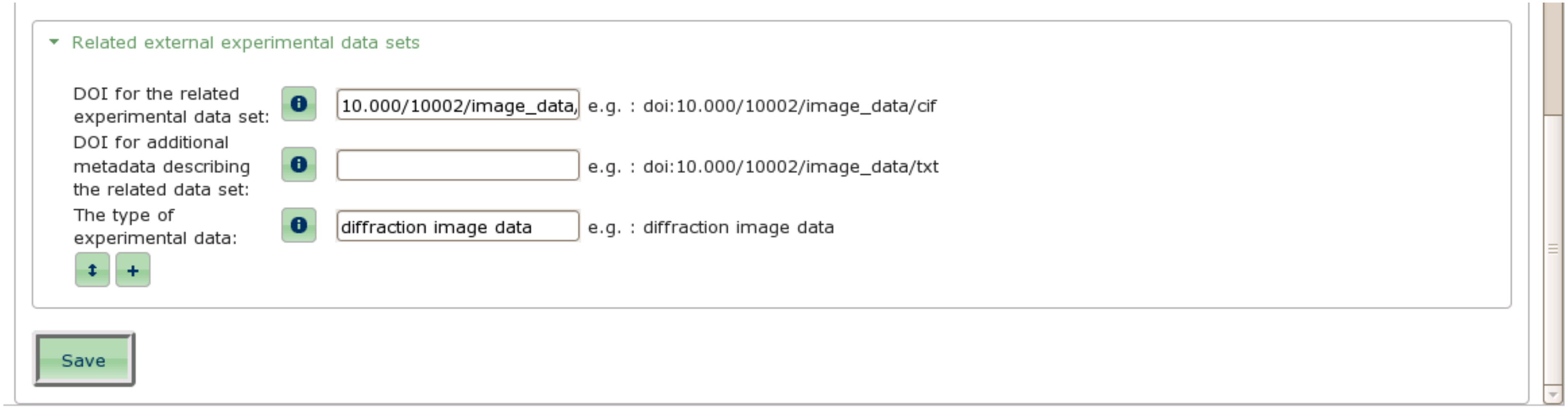
Recommendations Under Development

- Improved organization and packaging of structure factor, intensity, phasing and map data
- Controlled vocabulary of data set content types
- Standard specifications for each data category
- Improved linking between related data sets, crystal samples, and refined models
- Incorporation of unmerged intensities and map coefficients
- Comprehensive representation of chemical topology and restraints

Restoring the missing bits …

Identifying *Related Experimental Data*

- References to hosted data sets
  - DOIs for data sets
  - DOIs for related metadata
  - Text descriptions of data and metadeta



Related external experimental data sets

DOI for the related experimental data set: 10.000/10002/image_data, e.g. : doi:10.000/10002/image_data/cif

DOI for additional metadata describing the related data set: e.g. : doi:10.000/10002/image_data/txt

The type of experimental data: diffraction image data e.g. : diffraction image data

We Live in a Distributed World

… just a few examples

http://proteindiffraction.org/

http://tardis.edu.au/deposit/

BMRB

http://www.bmrb.wisc.edu/

EMBL

http://www.sasbdb.org/

PDBe
Protein Data Bank in Europe

http://www.ebi.ac.uk/pdbe/emdb/

EMDataBank
Unified Data Resource for 3DEM

http://www.ebi.ac.uk/pdbe/emdb/empiar/

EMPIAR

First Step Data Identification - Next Step Federation

Federation of loosely coupled resources with well defined data exchange protocols based on shared metadata standards.

Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop Structure 23: 1156–1167 doi: 10.1016/j.str.2015.05.013

Acknowledgements

WORLDWIDE
wwPDB
PROTEIN DATA BANK

RCSB PDB
PROTEIN DATA BANK

PDBe
Protein Data Bank in Europe

PDBj
Protein Data Bank Japan

BMRB

RCSB PDB
PROTEIN DATA BANK

RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

UCSD
SDSC
SKAGGS SCHOOL *of* PHARMACY *and* PHARMACEUTICAL SCIENCES

Funding: NSF NIGMS Office of Science U.S. DEPARTMENT OF ENERGY NATIONAL LIBRARY OF MEDICINE NATIONAL CANCER INSTITUTE® NATIONAL INSTITUTE OF NEUROLOGICAL DISORDERS AND STROKE NIDDK NATIONAL INSTITUTE OF DIABETES AND DIGESTIVE AND KIDNEY DISEASES