The New wwPDB Deposition and Annotation System

John Westbrook

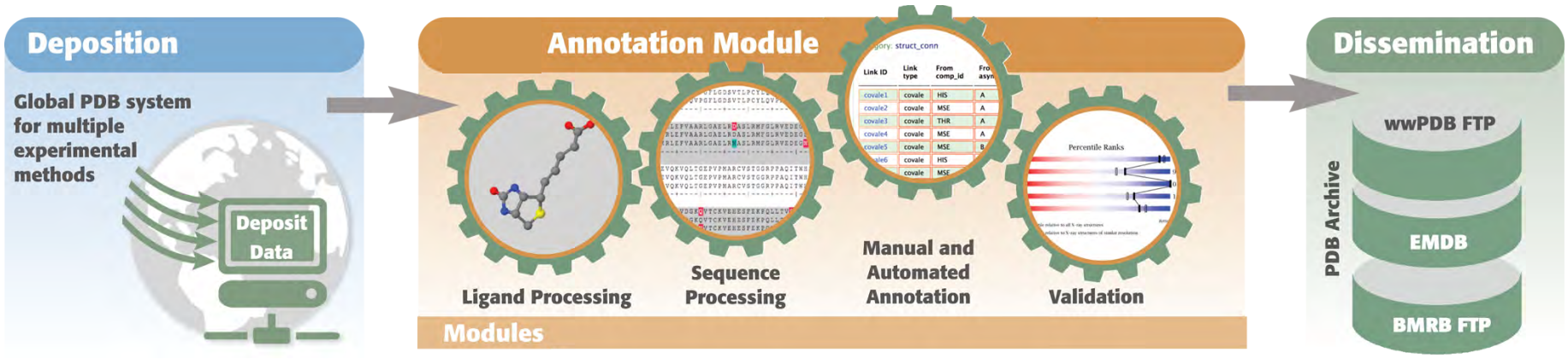
Rutgers, The State University of New Jersey

WORLDWIDE PDB
PROTEIN DATA BANK

www.wwpdb.org

wwPDB Deposition & Annotation System

- Enables depositions of structures of any size, determined using X-ray diffraction, EM, and/or NMR methods
- Implements format and validation protocols recommended by community Task Forces
- Will be used by all wwPDB Data Centers
- Will go into full production in early 2014



PDB Format Example

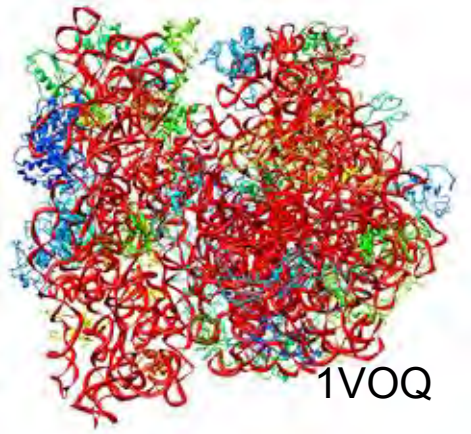
```
REMARK   3  DATA USED IN REFINEMENT.
REMARK   3   RESOLUTION RANGE HIGH (ANGSTROMS) : 1.57
REMARK   3   RESOLUTION RANGE LOW  (ANGSTROMS) : 23.00
REMARK   3   DATA CUTOFF            (SIGMA(F)) : 0.000
REMARK   3   COMPLETENESS FOR RANGE        (%) : NULL
REMARK   3   NUMBER OF REFLECTIONS             : 43316
REMARK   3
REMARK   3  FIT TO DATA USED IN REFINEMENT.
REMARK   3   CROSS-VALIDATION METHOD           : NULL
REMARK   3   FREE R VALUE TEST SET SELECTION   : NULL
REMARK   3   R VALUE     (WORKING + TEST SET) : NULL
REMARK   3   R VAL
REMARK   3   FREE
REMARK   3   FREE
REMARK   3   FREE
REMARK   3
```

- Record-oriented with fixed column format
- Metadata in semi-structured remarks
- Documentation by example

```
ATOM      1  N   VAL A 363
ATOM      2  CA  VAL A 363      21.557  -0.831  11.024  1.00 32.13           C
ATOM      3  C   VAL A 363      20.954  -1.757   9.943  1.00 31.73           C
ATOM      4  O   VAL A 363      19.737  -1.906   9.845  1.00 30.94           O
ATOM      5  CB  VAL A 363      21.883   0.552  10.391  1.00 33.45           C
```

Addressing the Format Challenges of Large Structures

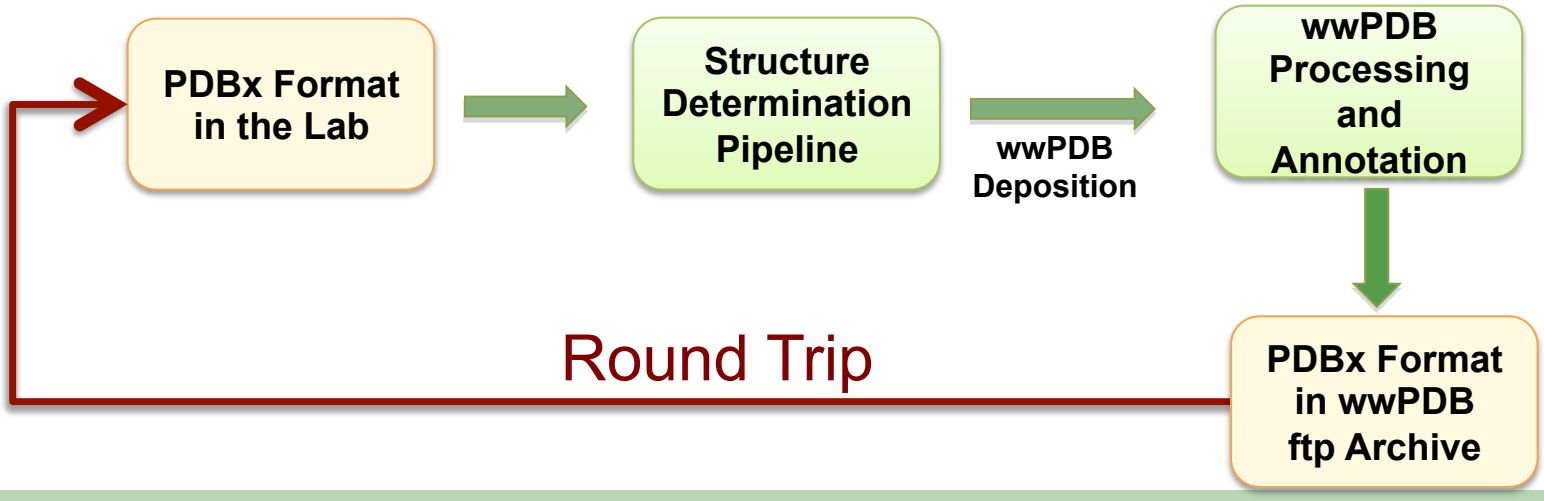
- PDB column format limitations
  - 1-character for polymer chain labels
  - 5-characters for atom serial numbers
  - 3-characters for monomers and ligand identifiers
  - 5-characters for atom names
  - F8.3 for model coordinates
- Implications –
  - Maximum of 62 chains (upper and lower case!)
  - Maximum of 99,999 atoms
    - Requires splitting structures across multiple entries (5 ribosomes in ASU stored in 10 PDB entries!)
    - Map and experimental validation are difficult for split entries
  - Cannot use standard monomer & ligand nomenclatures (e.g. carbohydrates & protonation variants)
  - Cannot use conventional atom names in large ligands
  - Limits molecular dimension (< 9999.999 Angstroms)



PDBx Deposition Working Group

PDBx Deposition Working Group
Refinement Developers Workshop 2011 - EBI

- 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



PDBx/mmCIF Format Example

- Name – value pairs

```
_exptl.entry_id          1XBB
_exptl.method            'X-RAY DIFFRACTION'
_exptl.crystals_number   1
```

- Table

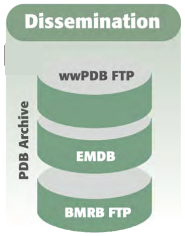
```
 loop
_data
_data
_data
_data
_database_PDB_rev.replaces
_database_PDB_rev.status
1 2004-11-02 2004-08-30 0 1XBB ?
2 2005-03-22 ?          1 1XBB ?
3 2009-02-24 ?          1 1XBB ?
```

- Simple syntax
- Named data items
- Data semantics defined in the PDBx data dictionary
- Dictionary metadata are easily extended
- Software support in most popular languages

Support for PDBx/mmCIF

- Deposition ready PDBx files from CCP4 and Phenix
- Reference data (e.g. symmetry data and PDB & CCP4 chemical libraries)
- Visualization tool support (e.g. Coot, Jmol, Chimera)
- Archiving image data (e.g. imgCIF/CBFlib)
- Parsing libraries and tools (C++, Java, Python, Perl)
- Data harvesting tools for deposition (e.g. PDB_Extract)
- mmCIF/PDBx is the PDB archival/exchange format
- PDBx dictionaries and examples at http://mmcif.pdb.org/

PDBx Deposition Working Group Recommendations

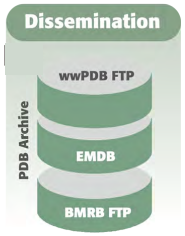


Announced: 22-May-2013

Format extensions for large structures:

- Atom serial numbers (1 to the number atoms)
- Chain identifiers up to 4 characters
- Cartesian coordinates with field widths as required and 3 decimal places
- B-factors and occupancies with 3 decimal places
- Extensions used only as required

Similar Format Styles



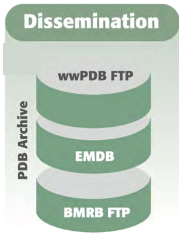
Original PDB

```
ATOM      1  N   GLN A  39      24.690 -27.754  24.275  1.00 60.76           N
ATOM      2  CA  GLN A  39      23.581 -26.768  24.416  1.00 60.98           C
ATOM      3  C   GLN A  39      23.990 -25.379  23.905  1.00 59.98           C
ATOM      4  O   GLN A  39      25.070 -25.209  23.330  1.00 60.25           O
ATOM      5  CB  GLN A  39      23.136 -26.685  25.878  1.00 60.69           C
```

PDBx/mmCIF

```
loop_
_atom_site.group_PDB
_atom_site.id
_atom_site.auth_atom_id
_atom_site.type_symbol
_atom_site.auth_comp_id
_atom_site.auth_asym_id
_atom_site.auth_seq_id
_atom_site.Cartn_x
_atom_site.Cartn_y
_atom_site.Cartn_z
_atom_site.pdbx_PDB_model_num
_atom_site.occupancy
_atom_site.pdbx_auth_alt_id
_atom_site.B_iso_or_equiv
 ATOM          20405  N    N  VAL  S2AA  121          163.443      59.452     136.546 1  1.000  .  144.800
 ATOM          20406  CA   C  VAL  S2AA  121          163.065      60.895     136.531 1  1.000  .  144.210
 ATOM          20407  CB   C  VAL  S2AA  121          164.255      61.798     136.939 1  1.000  .  144.790
 ATOM          20408  CG1  C  VAL  S2AA  121          163.766      63.172     137.376 1  1.000  .  144.210
 ATOM          20409  CG2  C  VAL  S2AA  121          165.057      61.156     138.066 1  1.000  .  144.080
 ATOM          20410  C    C  VAL  S2AA  121          162.543      61.279     135.142 1  1.000  .  143.210
```

Transitional Home for Large Structures

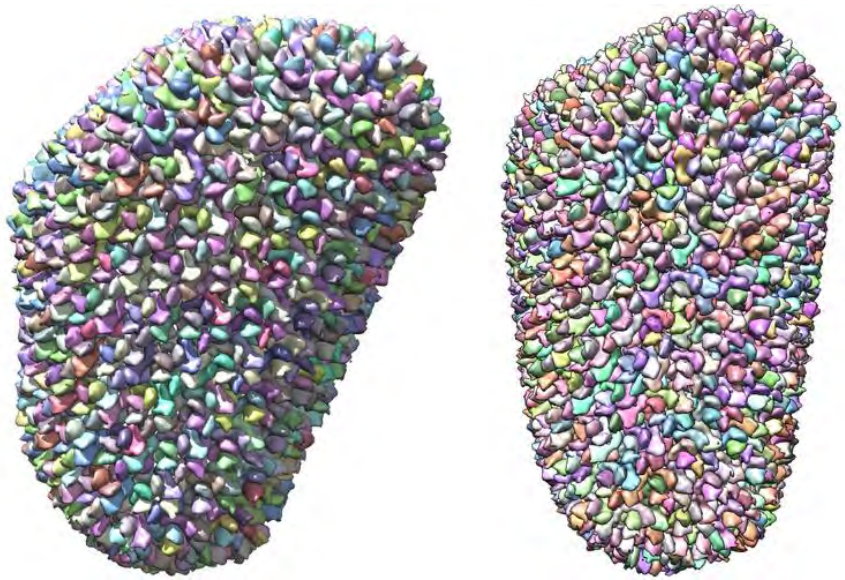


Large single entries are now stored separately on the wwPDB ftp site, and PDB internally produces *divided/split* PDB format files.

ftp://ftp.wwpdb.org/pub/pdb/data/large_structures/mmCIF/
ftp://ftp.wwpdb.org/pub/pdb/data/large_structures/XML/

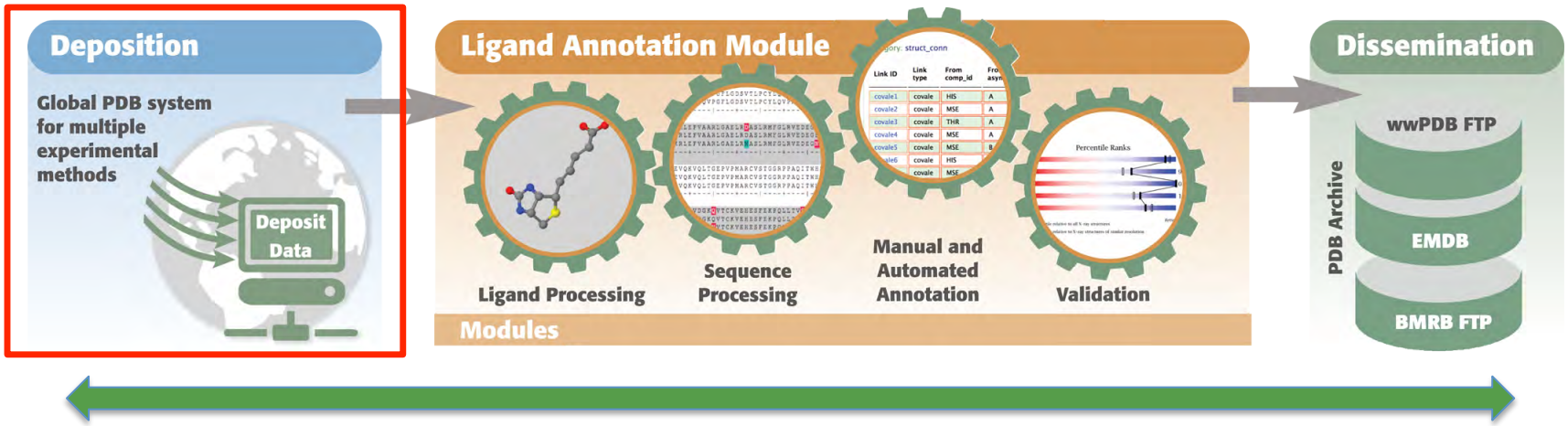
HIV-1 Capsid 3J3Q –

- 1356 chains
- >2M atoms
- 25 – PDB format entries



3J3Y 3J3Q

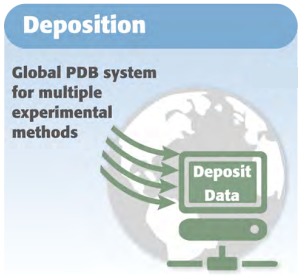
New Deposition System



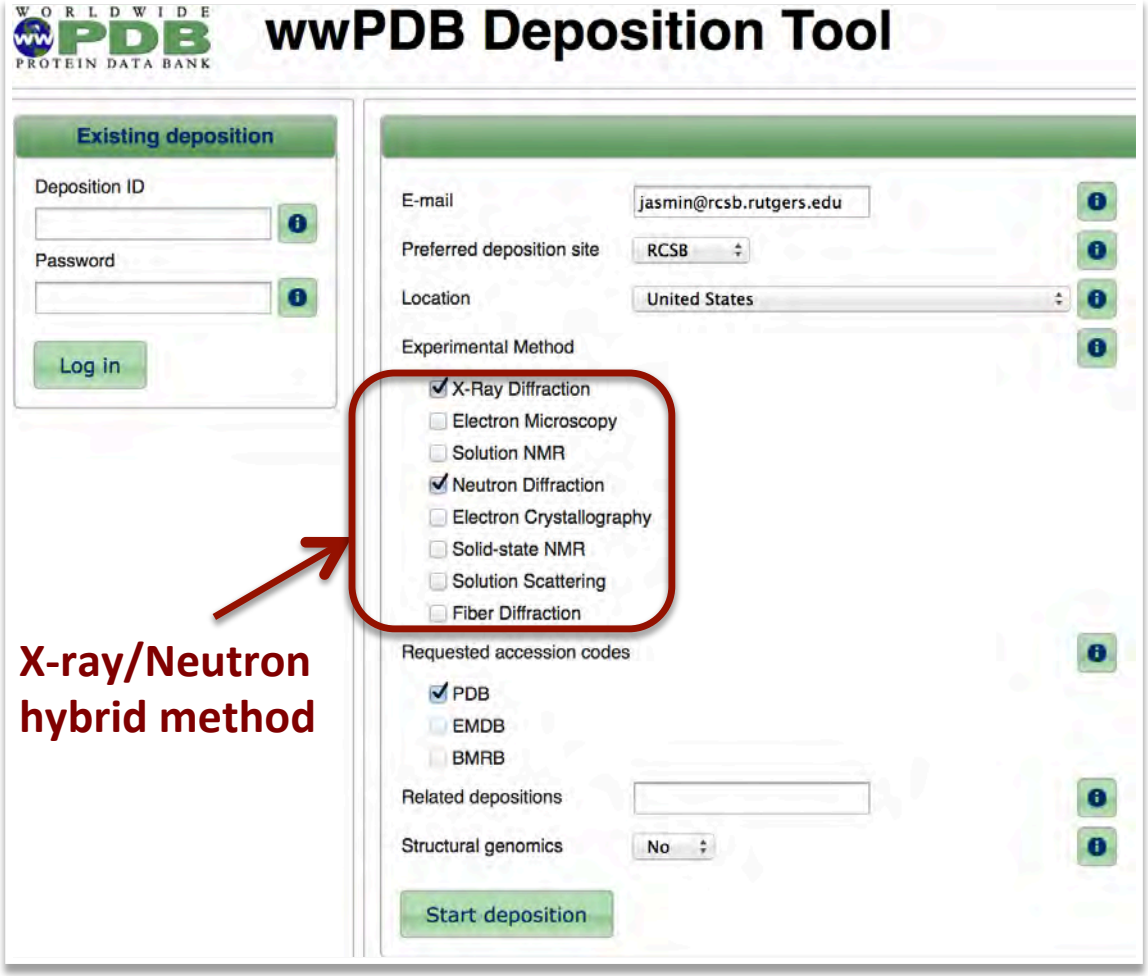
mmCIF/PDBx

End-to-end support for PDBx/mmCIF

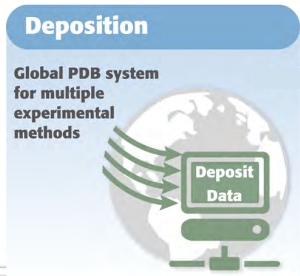
One Worldwide Entry Point for Deposition



- Single point of entry
- Automatic redirection to one of the wwPDB processing sites
- Support for multiple and hybrid methods
- New depositions based on prior depositions



Interface Screen Layout



wwPDB Deposition Tool - D_123763

SUBMIT

Data-Supplied

Content/Object Listing

- Welcome
- Replacement Upload
- Upload Summary
- Admin
  - Contact information
  - Grant information
  - Release status
  - Entry Title & author
  - Citation information
- Macromolecules
  - Molecule 1
  - Molecule 2
  - Add molecule
- Data collection
  - X-ray
- Refinement
  - X-ray refine
  - X-ray TLS
- Ligands
- Biological assembly
- Model & data validation
- Summary & conditions
- Downloads & View
  - Download Files
  - Output Cif
  - View

Log out

Load succesfu

Navigation Panel

Welcome to the wwPDB

This page contains a summary of the uploaded data. You must check that the data content here is correct before proceeding to other pages as data problems here are likely to require new data to be uploaded resulting in lost of data input on any other page.

▸ Upload file check

▾ Molecule

Entities we found in your upload. If data has not been provided in your uploaded data then you will have the opportunity to provide this later in the deposition interface:

Entity type	Description	Number
polymer	PUTATIVE HEXOSE OXIDASE	4
polymer	TEICOPLANIN	8
non-polymer	FLAVIN-ADENINE DINUCLEOTIDE	4
non-polymer	CITRIC ACID	5
non-polymer	SUGAR (ALPHA-D-MANNOSE)	8
non-polymer	SUGAR (N-ACETYL-D-GLUCOSAMINE)	8
non-polymer	SUGAR ((2R,3R,4S,5R,6R)-3-azanyl-6-[bis(oxidanyl)methyl]oxane-2,4,5-triol)	6
non-polymer	8-METHYLNONANOIC ACID	6
water	water	1400

Report & Data-entry Panel

▾ Sequence

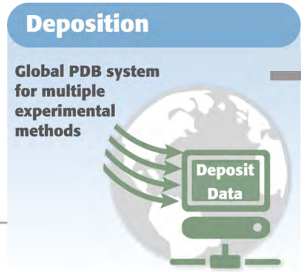
Sequence data provided in your upload, if you did not provide sequence data then you will have the opportunity to enter this later wihitn the Molecule pages; discrepancies between the coordinate sequence and biological sequence are not allowed.

```
Sample sequence aligned with Coordinates - all chains :

Molecule : 1
Chain : A
--------10--------20--------30--------40--------50--------60--------70--------80--------90-------100
MTGGTGADAASAGASSTRPELRGERCLPPAGPVKVTPDDPRYLNLKLRGANSRFNGEPDYIHLVGSTQQVADAVEETVRTGKRVAVRSGGHCFEDFVDNP
---------------|----------||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
..............S.........LPPAGPVKVTPDDPRYLNLKLRGANSRFNGEPDYIHLVGSTQQVADAVEETVRTGKRVAVRSGGHCFEDFVDNP

-------110-------120-------130-------140-------150-------160-------170-------180-------190-------200
DVKVIIDMSLLTEIAYDPSMNAFLIEPGNTLSEVYEKLYLGWNVTIPGGVCGGVGVGGHICGGGYGPLSRQFGSVVDYLYAVEVVVVNKQGKARVIVATR
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
DVKVIIDMSLLTEIAYDPSMNAFLIEPGNTLSEVYEKLYLGWNVTIPGGVCGGVGVGGHICGGGYGPLSRQFGSVVDYLYAVEVVVVNKQGKARVIVATR
```

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

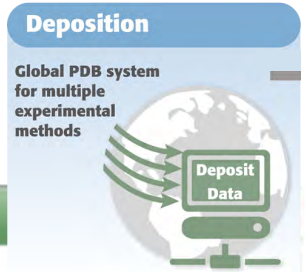
▾ Overall data quality

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

PDB EXTRACT

Chemical Assignment



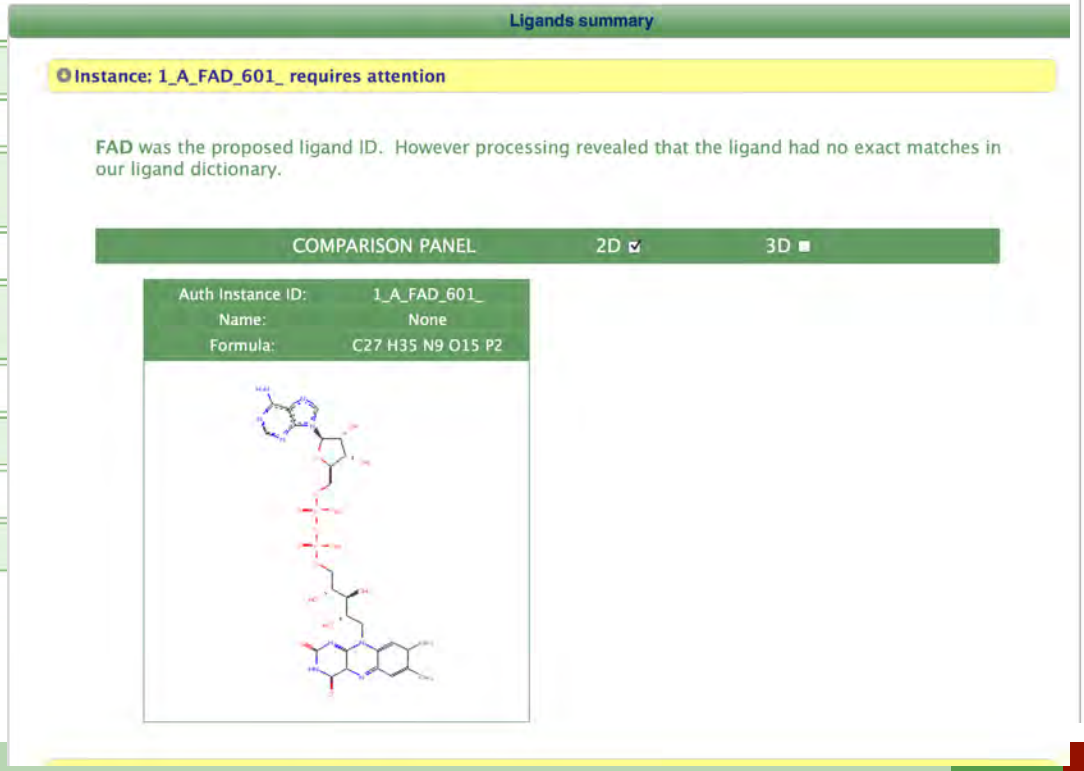
Ligands summary

Save Finish

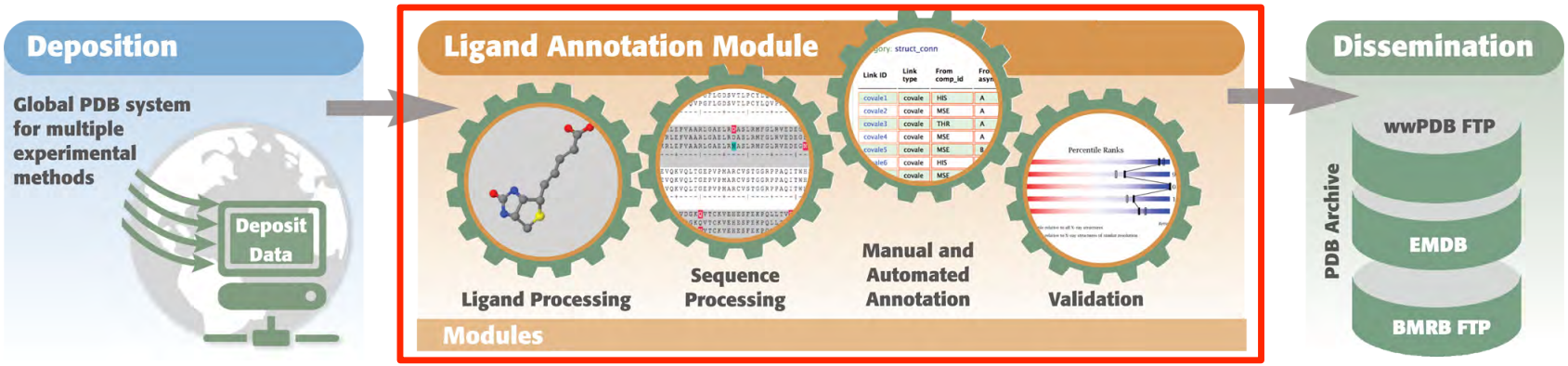
Summary of ligands identified in coordinate file provided for dataset: D_123763

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	

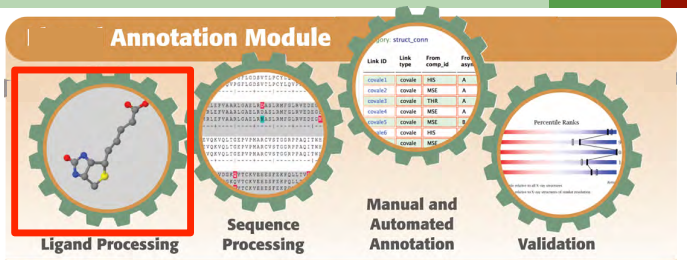
Instance Inspection View



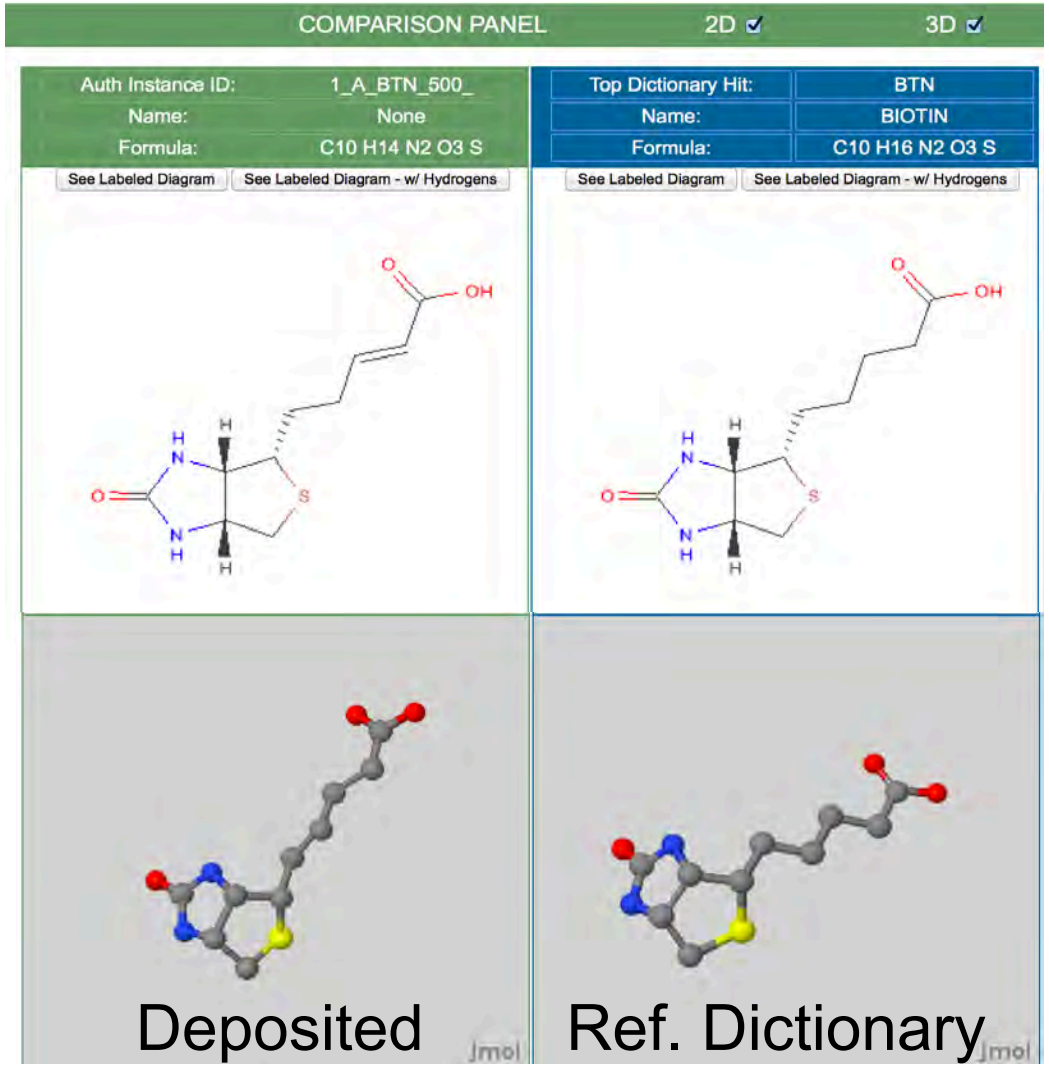
New Annotation System



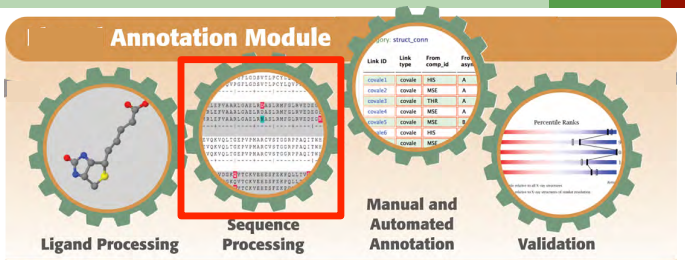
Chemical Features



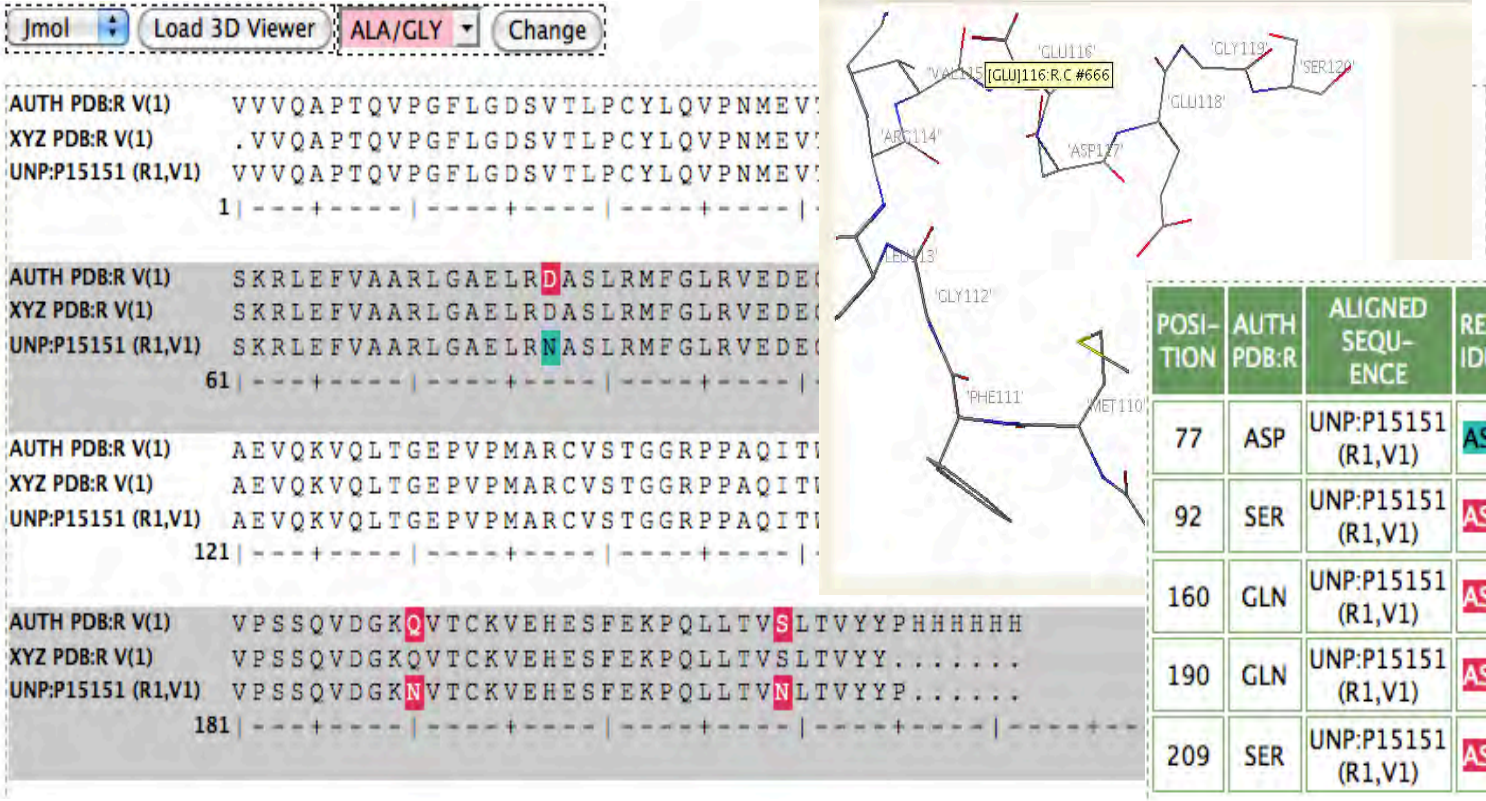
- Chemical structure search against our Chemical Component and BIRD dictionaries
- Integrates author-provided chemical information
- Comparison Panel
  - 2D and 3D views of ligand for review
- Creates and edits new and existing chemical definitions



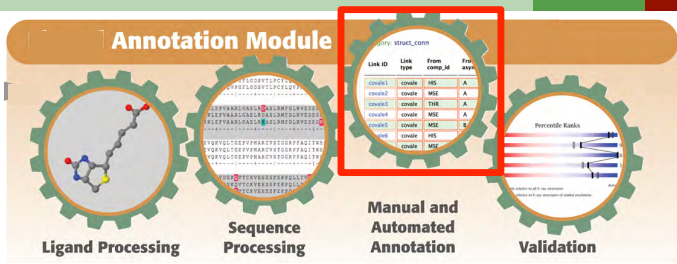
Sequence Annotation



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

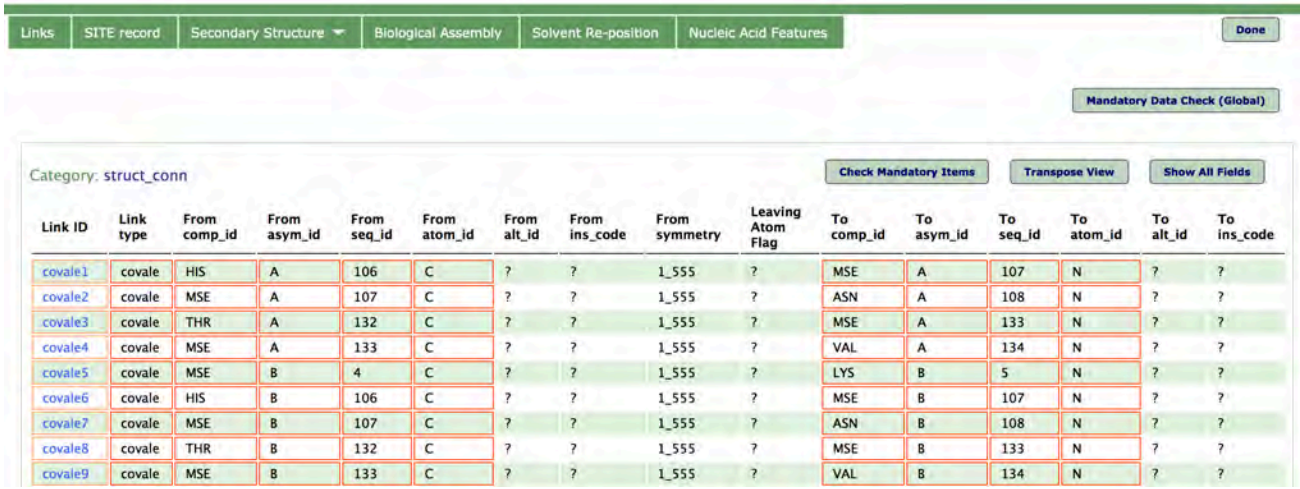


Structure Annotation



Site & Linkage Details

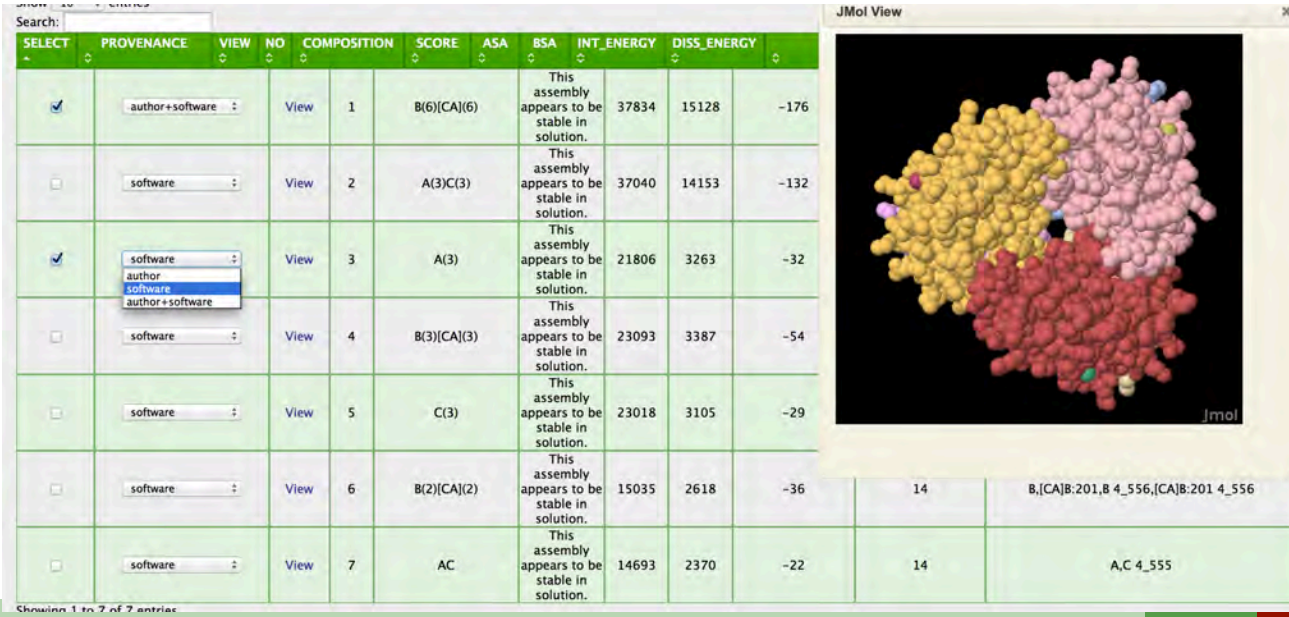
Annotators review automatically derived data and edit as needed using this annotation module.



Link ID	Link type	From comp_id	From asym_id	From seq_id	From atom_id	From alt_id	From ins_code	From symmetry	Leaving Atom Flag	To comp_id	To asym_id	To seq_id	To atom_id	To alt_id	To ins_code
covale1	covale	HIS	A	106	C	?	?	1_555	?	MSE	A	107	N	?	?
covale2	covale	MSE	A	107	C	?	?	1_555	?	ASN	A	108	N	?	?
covale3	covale	THR	A	132	C	?	?	1_555	?	MSE	A	133	N	?	?
covale4	covale	MSE	A	133	C	?	?	1_555	?	VAL	A	134	N	?	?
covale5	covale	MSE	B	4	C	?	?	1_555	?	LYS	B	5	N	?	?
covale6	covale	HIS	B	106	C	?	?	1_555	?	MSE	B	107	N	?	?
covale7	covale	MSE	B	107	C	?	?	1_555	?	ASN	B	108	N	?	?
covale8	covale	THR	B	132	C	?	?	1_555	?	MSE	B	133	N	?	?
covale9	covale	MSE	B	133	C	?	?	1_555	?	VAL	B	134	N	?	?

Biological Assembly

Annotators assign biological assembly information to an entry after visualizing in 3D and reviewing software suggestion and depositor input.



SELECT	PROVENANCE	VIEW	NO	COMPOSITION	SCORE	ASA	BSA	INT_ENERGY	DISS_ENERGY	
☑	author+software	View	1	B(6)[CA](6)	This assembly appears to be stable in solution.		37834	15128	-176	
☐	software	View	2	A(3)C(3)	This assembly appears to be stable in solution.		37040	14153	-132	
☑	software (author / software / author+software)	View	3	A(3)	This assembly appears to be stable in solution.		21806	3263	-32	
☐	software	View	4	B(3)[CA](3)	This assembly appears to be stable in solution.		23093	3387	-54	
☐	software	View	5	C(3)	This assembly appears to be stable in solution.		23018	3105	-29	
☐	software	View	6	B(2)[CA](2)	This assembly appears to be stable in solution.		15035	2618	-36	14 B,[CA]B:201,B 4_556,[CA]B:201 4_556
☐	software	View	7	AC	This assembly appears to be stable in solution.		14693	2370	-22	14 A,C 4_555

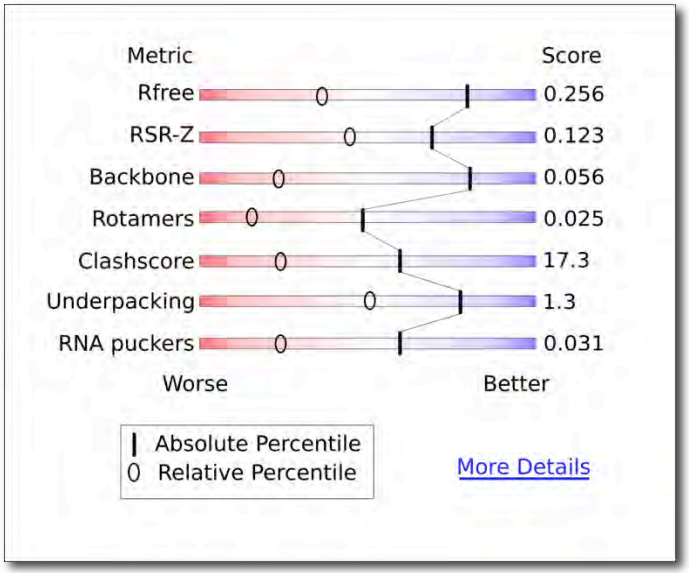
Showing 1 to 7 of 7 entries

wwPDB Validation 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.

Group	Meeting/ Workshop	Chair/ Membership	Outcome
X-ray Validation Task Force	2008	Randy Read (Univ of Cambridge) 17 members	(2011) *Structure* 19: 1395-1412
NMR Validation Task Force	2009, 2011	Gaetano Montelione (Rutgers) Michael Nilges (Institut Pasteur) 10 members	*Structure*, in press
3DEM Validation Task Force	2010	Richard Henderson (Maps, MRC-LMB) Andrej Sali (Models, UCSF) 21 members	(2012) *Structure* 20: 205-214
Small-Angle Scattering Task Force	2012	Jill Trewhella (Univ Sydney) 6 members	(2013) *Structure* 21: 875-881

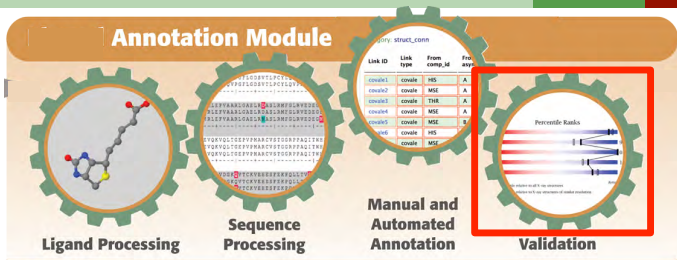
wwPDB X-ray VTF



- "Relative" = compared to structures of similar resolution
- Reference values and distributions will be recomputed annually

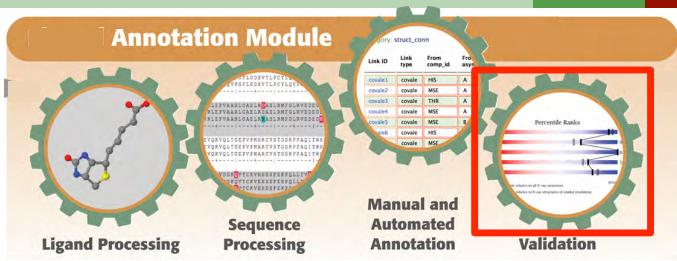
Report: Read *et al.*, *Structure* **19**, 1395 (2011)

Validation Report Content

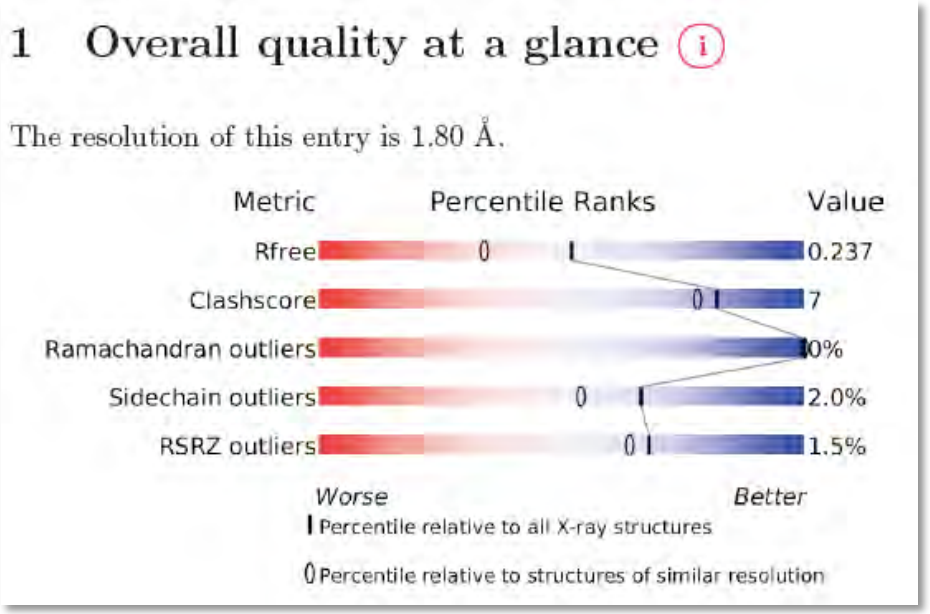


- Model quality
  - Bond lengths and angles (outlier info, RMS-Z)
  - Chirality, planarity
  - Close contacts (incl. clashscore, worst clashes)
  - Torsion angles (Ramachandran, rotamers for proteins)
  - Ligand geometry (Mogul analysis)
- Model/data fit
  - Macromolecules: RSR, RSR-Z, B-factors, partial occupancies
  - Ligands: same, but RSR-Z undefined
- Residue plots
  - Residues with model-quality outliers (0, 1, 2, >2)
  - Residues with RSR-Z > 2 are highlighted
  - Unmodeled residues

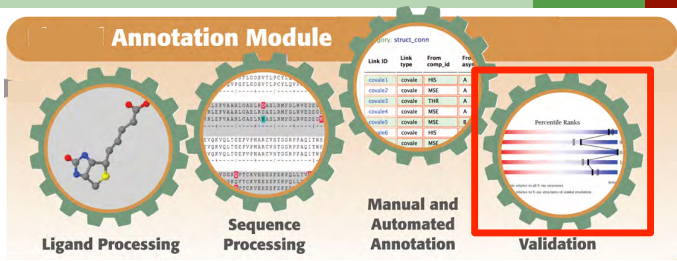
Report Presentation

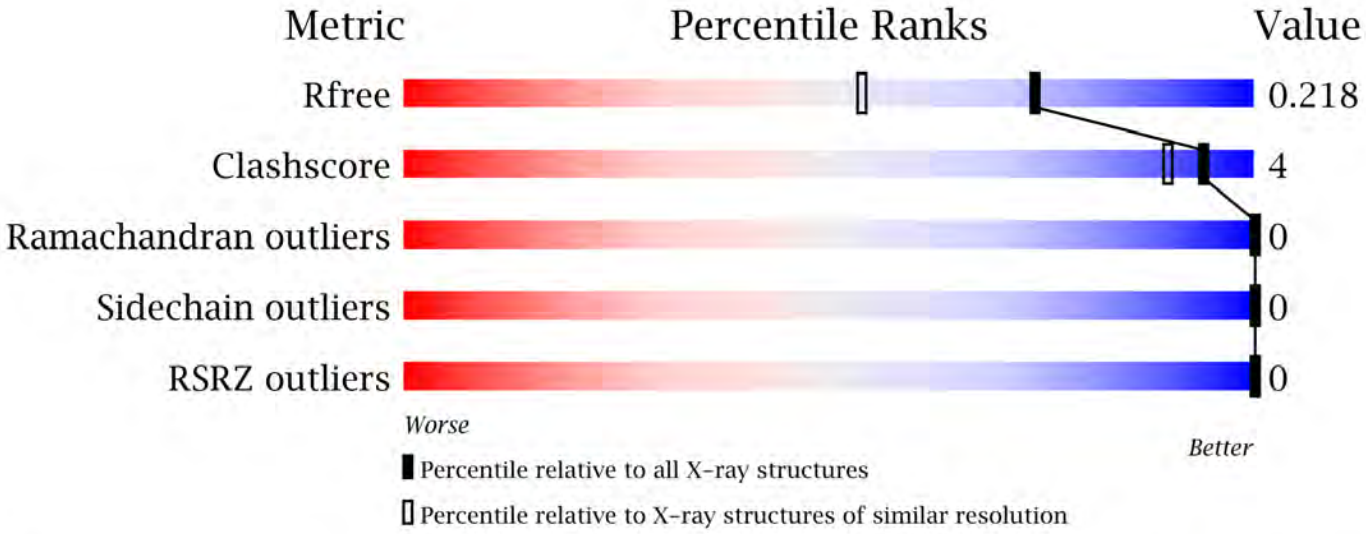


- Summarizes key issues
  - All details in XML file
  - Supporting help details
- Title page
  - Authors, title, accession
- Overall quality at-a-glance
  - Slider plots of key statistics
- “Table 1” summary
  - Key data and refinement stats
- Tabulations of diagnostics for entry components
  - Macromolecules
  - Ligands



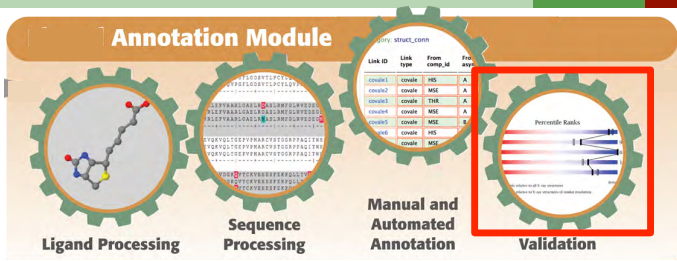
Overall Quality Summary

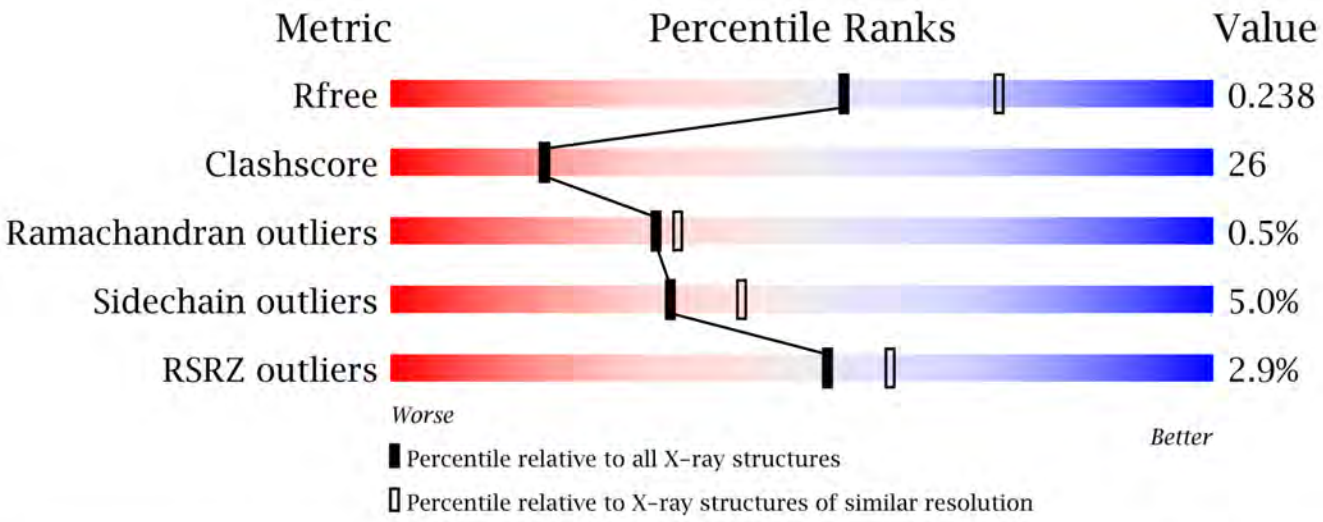




Metric	Whole archive (#Entries)	Similar resolution (#Entries, resolution range(Å))
R_{free}	65580	5415 (1.76-1.64)
Clashscore	76988	6143 (1.76-1.64)
Ramachandran outliers	75395	6037 (1.76-1.64)
Sidechain outliers	75377	6036 (1.76-1.64)
RSRZ outliers	65576	5415 (1.76-1.64)

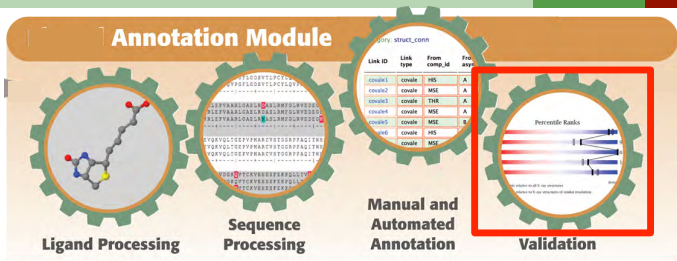
Overall Quality Summary





Metric	Whole archive (#Entries)	Similar resolution (#Entries, resolution range(Å))
R_{free}	65580	5027 (2.36-2.24)
Clashscore	76988	5057 (2.34-2.26)
Ramachandran outliers	75395	5827 (2.36-2.24)
Sidechain outliers	75377	5827 (2.36-2.24)
RSRZ outliers	65576	5027 (2.36-2.24)

Residue Plots

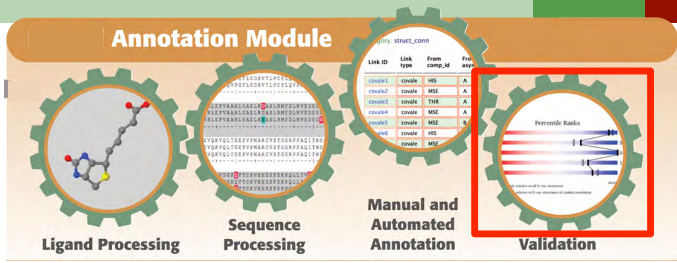


Gray – not modeled

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

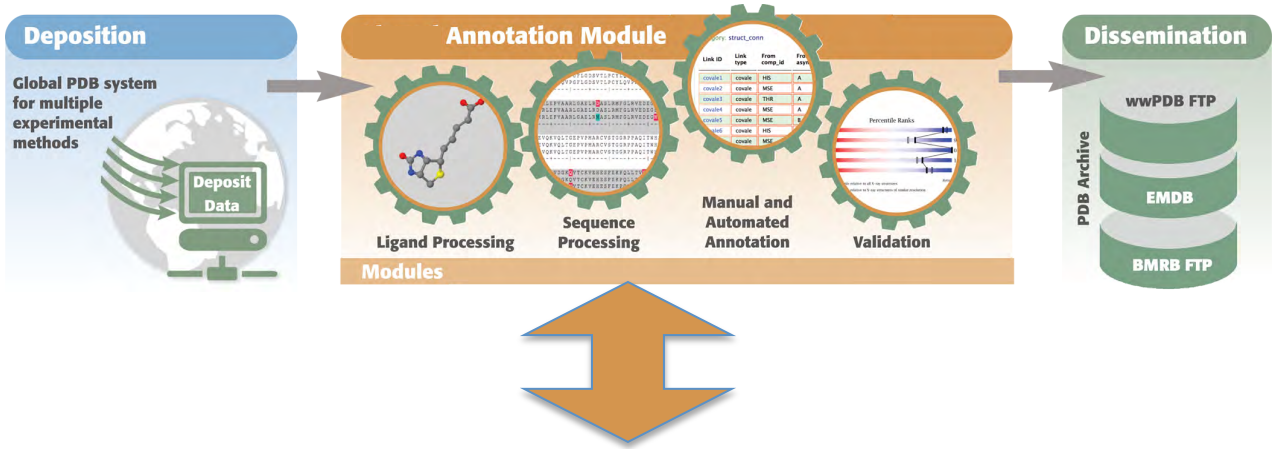
Red dot – poor fit to electron density

Table One - Details

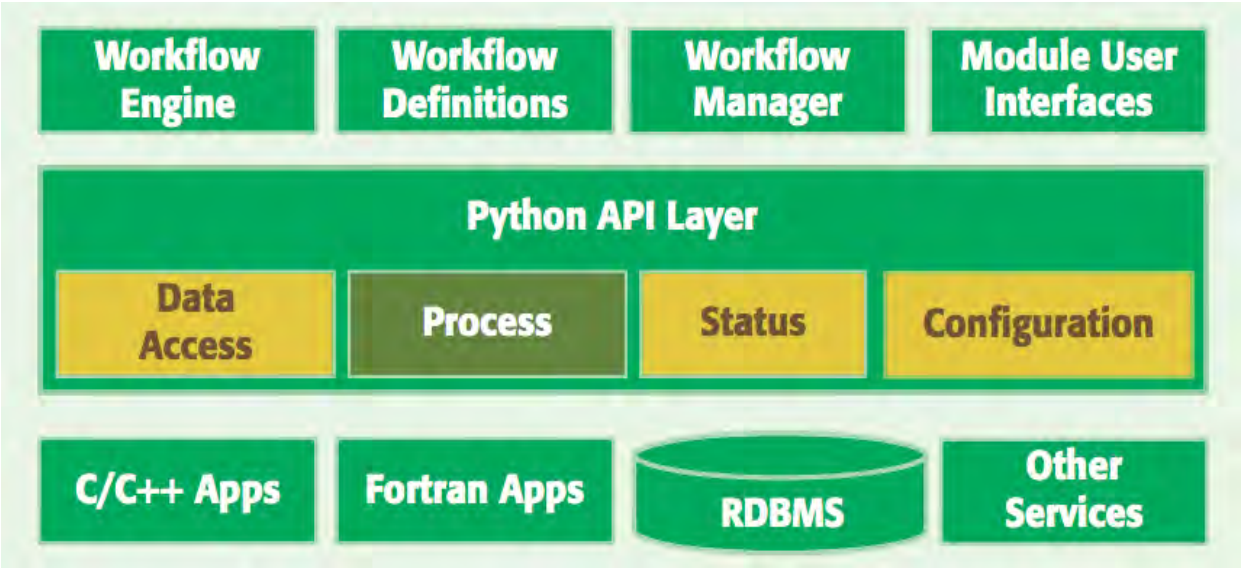


Property	Value	Source
Space group	C 1 2 1	Depositor
Cell constants a, b, c, α, β, γ	153.85Å 97.90Å 101.02Å 90.00° 128.53° 90.00°	Depositor
Data completeness (%)	92.5 92	Depositor EDS
R_{merge}	0.10	Depositor
R_{sym}	0.10	Depositor
$< I/\sigma(I) >$	3.14 (2.29Å)	Xtriage
Resolution (Å)	15.00 – 2.30 70.27 – 2.30	Depositor EDS
Refinement program	CNS 1.2	Depositor
R, R_{free}	0.211 , 0.240 0.207 , 0.238	Depositor DCC
Wilson B-factor (Å^2)	27.2	Xtriage
Anisotropy	0.172	Xtriage
Bulk solvent K_{sol}(e/Å^3), B_{sol}(Å^2)	0.334 , 35.116	EDS
Estimated twinning fraction	0.019 for -h+k-l,-l,-k 0.000 for -h-k-l,l,k 0.042 for -h-2*l,-k,l	Xtriage Xtriage Xtriage
L-test for twinning	$< \|L\| > = 0.49$, $< L^2 > = 0.32$	Xtriage
Outliers	0 of 48182 reflections	Xtriage
F_o,F_c correlation	0.92	EDS
Total number of atoms	12401	Calculated
Average B, all atoms (Å^2)	26.0	Calculated

Annotation and Deposition Workflow System



Workflow Management System



Deployment Plan

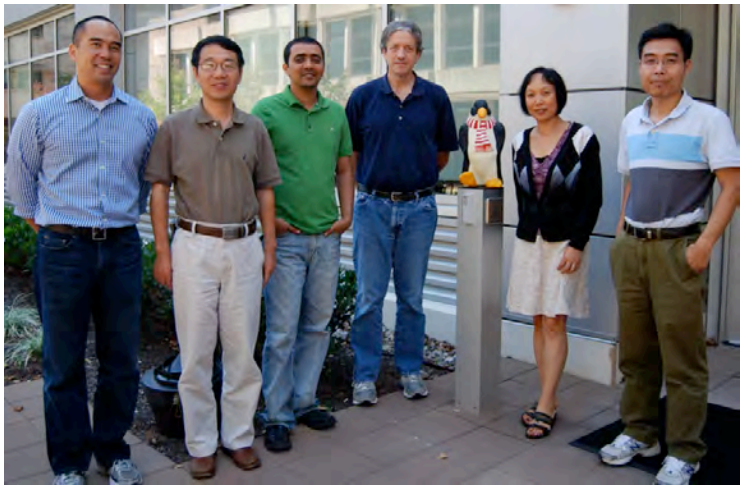
- New validation reports to be provided for all depositions
- A standalone validation server will be available for testing
- Both old and new deposition systems available during transition
- Beginning in 2014 -
  - All NEW depositions will use the new system.
  - Depositions initiated in old deposition system will have 6 months for completion

Beta testing begins in September 2013!

The wwPDB is looking for beta testers to deposit data using the new system

Interested in beta testing the deposition interface?
Drop by the PDB Booth (#101) or poster 13.05.6,
or e-mail: **deposit-request@mail.wwpdb.org**

The wwPDB Deposition and Annotation Development Team



Acknowledgements

WORLDWIDE PDB 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