

Harvesting Data Collection Information for the RCSB PDB Depositions

RCSB PDB PROTEIN DATA BANK www.pdb.org • info@rcsb.org

Harvesting Data Collection Information for RCSB PDB Depositions

John Westbrook
Rutgers, The State University of New Jersey

RCSB PDB PROTEIN DATA BANK www.pdb.org • info@rcsb.org

Data Management at the RCSB PDB

- Dictionary-driven architecture
- Pipeline of data collection spanning experimental structural biology
- Tools to automate data harvesting
- Continuous focus of data standardization and data quality
- Distribution and delivery deeply integrated with biological and medical resources

RCSB PDB PROTEIN DATA BANK www.pdb.org • info@rcsb.org

Dictionary Resources

<http://mmcif.pdb.org/>

An Information Portal to Biological Macromolecular Structures

Dictionary Resources

The Protein Data Bank (PDB) uses macromolecular Crystallographic Information File (mmcif) data dictionaries to describe the information content of PDB entries. The PDB Exchange data dictionary consolidates content from a variety of crystallographic dictionaries including the IUCr Core, mmCIF, Image and Symmetry dictionaries. The PDB Exchange Dictionary also includes extensions describing NMR, Cryo-EM, and protein production data. PDB data processing, data exchange, annotation, and database management operations all make heavy use of the data format and the content of the PDB Exchange Dictionary. Software tools are used to convert mmCIF data files to the older PDB format and to PDBML/XML.

- Data files in mmCIF format can be downloaded from the RCSB PDB website or by ftp.
- Software tools are available for processing and editing dictionaries.
- Software tools are available for converting mmCIF data files to PDB and PDBML formats.
- A complete list of PDB software tools for managing PDB data in mmCIF format can be found [here](#).

Dictionary Content and Representation

- Background and Introduction about mmCIF
- The Macromolecular Crystallographic Information File (mmCIF) Math. Enzymol. (1997) 277, 571-590.
- 574mmCIF: An Extensive Challenge for Macromolecular Structure and Beyond (PDF). Bioinformatics (2000) 16(2), 159-166.
- mmCIF Software Developers Workshop 1997
- mmCIF Dictionary Templates
- mmCIF Examples
- References

Data Dictionaries

- PDB mmCIF Exchange Dictionary (ASCI) (compressed) (HTML) | XML Schema
Data dictionary developed as a collaboration between MSD-EBI and RCSB and used by wwPDB members for data exchange.
- mmCIF Dictionary (ASCI) (compressed) (HTML)
IUCr Standard Dictionary
Original working group members: Paula M. Fitzgerald, Helen Berman, Phil Bourne, Brian McMahon, Keith Waterspaugh, and John Westbrook

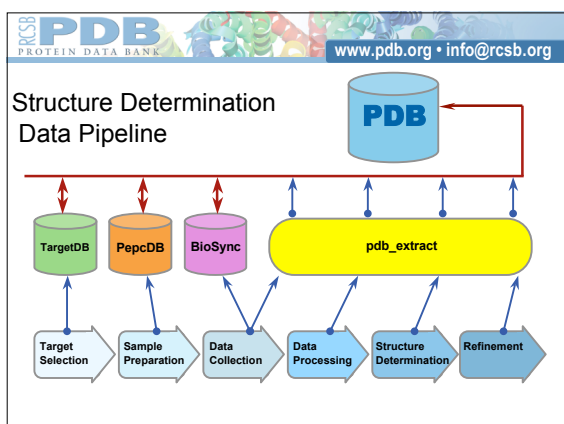
RCSB PDB PROTEIN DATA BANK www.pdb.org • info@rcsb.org

PDBML Schemas

<http://pdbml.pdb.org/>

- New PDBML website
- Schemas provided for the PDB Exchange dictionary and component dictionaries
- Schemas and data dictionaries are updated synchronously

PDBML: the representation of archival macromolecular structure data in XML. John Westbrook, Nobutoshi Ito, Haruki Nakamura, Kim Henrick and Helen M. Berman, *Bioinformatics*, 21(7), 988-992, 2005.



RCSB PDB PROTEIN DATA BANK www.pdb.org • info@rcsb.org

Target Registration Database

TargetDB • <http://targetdb.pdb.org/>

- Targets downloadable in XML (~120K targets)
- Targets downloaded from 21 worldwide centers weekly
- Target search by:
 - Sequence (FASTA), project target ID, project site, status (selected, cloned, expressed, ... in PDB), update date, protein name, source organism
- Report output in HTML, FASTA, and XML
- Integrates PDB entry sequences (~90K sequences)
- Includes PDB pre-release sequence data
- Provides links to related project sequence databases
- Summary reports of target and project progress
- 3-Tier architecture (Apache/Tomcat/MySQL)

Harvesting Data Collection Information for the RCSB PDB Depositions

RCSB PDB
PROTEIN DATA BANK
www.pdb.org • info@rcsb.org

Protein Expression Purification and Crystallization Database

PepeDB • <http://pepdb.pdb.org/>

- Extends content of TargetDB
- Includes all TargetDB sequences
- All protocols for cloning, expression, purification are stored and are searchable
- Reports provide links to status history, related protocols, project, sequence and domain databases

RCSB PDB
PROTEIN DATA BANK
www.pdb.org • info@rcsb.org

Data Collection - BioSync

<http://biosync.rcsb.org/>

BioSync
Structural Biology Synchrotron Users Organization

BioSync: A Structural Biologist's Guide to Synchrotron Facilities

Current Synchrotron Information: Updated by synchrotron personnel

Baseline Descriptions Contain: Wavelength ranges, flux and optical configurations, detector, hardware/software

Baseline Descriptions: Baseline statistics for synchrotrons around the world based on PDB release data

Coming Soon: Baselines outside the US, more small angle x-ray scattering and x-ray spectroscopy, including cross-links to the RCSB PDB website

Read History: BioSync was formed in 1998 as a grassroots organization that would provide access to synchrotron facilities for scientists whose primary research is in the field of structural biology. The members were leaders of North American research groups that used synchrotron radiation for experiments in structural biology. The members of the original Steering Committee for BioSync were: Keith Wilson, Robert Sweet, Richard Kohn, Keith Moffat, Janet Smith, Robert Sweet, and John Westbrook

The activities of BioSync included documenting the needs of structural biologists for synchrotron radiation, evaluating the status of existing and planned facilities, recommending funding and access policies, lobbying in Washington for improved support of synchrotron rings and beamlines, and organizing a web-based clearinghouse of baseline information.

Read More...

RCSB PDB
PROTEIN DATA BANK
www.pdb.org • info@rcsb.org

Data Collection - BioSync

BioSync
Structural Biology Synchrotron Users Organization

Home | US Synchrotrons | International Synchrotrons | Structural Genomics

US Synchrotrons

Current Baselines

Synchrotron Statistics

National Synchrotron Light Source (NSLS)

is located at Brookhaven National Laboratory in Upton, NY

NSLS Structural Biology Baselines Overview

Beamline	Operational (general use)	Current	Wavelength (Å)
X1A	Yes	MACS	0.7-1.9
X2A	Yes	MACS	0.7-1.9
X3A	Yes	MACS	0.7-1.9
X4A	Yes	MACS	0.7-1.9
X5A	Yes	MACS	0.7-1.9
X6A	Yes	MACS	0.7-1.9
X7A	Yes	MACS	0.7-1.9
X8A	Yes	MACS	0.7-1.9
X9A	Yes	MACS	0.7-1.9
X10A	Yes	MACS	0.7-1.9
X11A	Yes	MACS	0.7-1.9
X12A	Yes	MACS	0.7-1.9
X13A	Yes	MACS	0.7-1.9
X14A	Yes	MACS	0.7-1.9
X15A	Yes	MACS	0.7-1.9
X16A	Yes	MACS	0.7-1.9
X17A	Yes	MACS	0.7-1.9
X18A	Yes	MACS	0.7-1.9
X19A	Yes	MACS	0.7-1.9
X20A	Yes	MACS	0.7-1.9
X21A	Yes	MACS	0.7-1.9
X22A	Yes	MACS	0.7-1.9
X23A	Yes	MACS	0.7-1.9
X24A	Yes	MACS	0.7-1.9
X25A	Yes	MACS	0.7-1.9
X26A	Yes	MACS	0.7-1.9
X27A	Yes	MACS	0.7-1.9
X28A	Yes	MACS	0.7-1.9
X29A	Yes	MACS	0.7-1.9
X30A	Yes	MACS	0.7-1.9
X31A	Yes	MACS	0.7-1.9
X32A	Yes	MACS	0.7-1.9
X33A	Yes	MACS	0.7-1.9
X34A	Yes	MACS	0.7-1.9
X35A	Yes	MACS	0.7-1.9
X36A	Yes	MACS	0.7-1.9
X37A	Yes	MACS	0.7-1.9
X38A	Yes	MACS	0.7-1.9
X39A	Yes	MACS	0.7-1.9
X40A	Yes	MACS	0.7-1.9
X41A	Yes	MACS	0.7-1.9
X42A	Yes	MACS	0.7-1.9
X43A	Yes	MACS	0.7-1.9
X44A	Yes	MACS	0.7-1.9
X45A	Yes	MACS	0.7-1.9
X46A	Yes	MACS	0.7-1.9
X47A	Yes	MACS	0.7-1.9
X48A	Yes	MACS	0.7-1.9
X49A	Yes	MACS	0.7-1.9
X50A	Yes	MACS	0.7-1.9
X51A	Yes	MACS	0.7-1.9
X52A	Yes	MACS	0.7-1.9
X53A	Yes	MACS	0.7-1.9
X54A	Yes	MACS	0.7-1.9
X55A	Yes	MACS	0.7-1.9
X56A	Yes	MACS	0.7-1.9
X57A	Yes	MACS	0.7-1.9
X58A	Yes	MACS	0.7-1.9
X59A	Yes	MACS	0.7-1.9
X60A	Yes	MACS	0.7-1.9
X61A	Yes	MACS	0.7-1.9
X62A	Yes	MACS	0.7-1.9
X63A	Yes	MACS	0.7-1.9
X64A	Yes	MACS	0.7-1.9
X65A	Yes	MACS	0.7-1.9
X66A	Yes	MACS	0.7-1.9
X67A	Yes	MACS	0.7-1.9
X68A	Yes	MACS	0.7-1.9
X69A	Yes	MACS	0.7-1.9
X70A	Yes	MACS	0.7-1.9
X71A	Yes	MACS	0.7-1.9
X72A	Yes	MACS	0.7-1.9
X73A	Yes	MACS	0.7-1.9
X74A	Yes	MACS	0.7-1.9
X75A	Yes	MACS	0.7-1.9
X76A	Yes	MACS	0.7-1.9
X77A	Yes	MACS	0.7-1.9
X78A	Yes	MACS	0.7-1.9
X79A	Yes	MACS	0.7-1.9
X80A	Yes	MACS	0.7-1.9
X81A	Yes	MACS	0.7-1.9
X82A	Yes	MACS	0.7-1.9
X83A	Yes	MACS	0.7-1.9
X84A	Yes	MACS	0.7-1.9
X85A	Yes	MACS	0.7-1.9
X86A	Yes	MACS	0.7-1.9
X87A	Yes	MACS	0.7-1.9
X88A	Yes	MACS	0.7-1.9
X89A	Yes	MACS	0.7-1.9
X90A	Yes	MACS	0.7-1.9
X91A	Yes	MACS	0.7-1.9
X92A	Yes	MACS	0.7-1.9
X93A	Yes	MACS	0.7-1.9
X94A	Yes	MACS	0.7-1.9
X95A	Yes	MACS	0.7-1.9
X96A	Yes	MACS	0.7-1.9
X97A	Yes	MACS	0.7-1.9
X98A	Yes	MACS	0.7-1.9
X99A	Yes	MACS	0.7-1.9
X100A	Yes	MACS	0.7-1.9

RCSB PDB
PROTEIN DATA BANK
www.pdb.org • info@rcsb.org

Data Collection - BioSync

Table 6: Structural Genomics structures in the PDB for which data were collected at NSLS

Synchrotron Beamline	Total Synchrotron Structures	Total SG Structures	PDB Structures	Non-PDB Structures in North America	SGC Structures	MPPEFA Structures	European SG Structures
X12B	185	7	3	3	1	0	0
X12C	401	67	63	4	0	0	0
X25	530	36	30	3	3	0	0
X26C	213	6	5	1	0	0	0
X28A	160	85	78	8	1	0	0
X3A	5	4	4	0	0	0	0
X5A	675	214	214	0	0	0	0
X6C	9	3	3	0	0	0	0
X8A	31	5	3	2	0	0	0
X8C	338	40	10	28	1	0	1
X9A	132	61	60	1	0	0	0
X10C	276	12	11	1	0	0	0
UNKNOWN	9	0	0	0	0	0	0
Total	2981	521	465	49	6	0	1

RCSB PDB
PROTEIN DATA BANK
www.pdb.org • info@rcsb.org

Data Collection - BioSync

The Advanced Photon Source (APS)

is located at Argonne National Laboratory in Argonne, IL

APS Structural Biology Baselines Overview

Beamline	Operational (general use)	Current	Wavelength (Å)
19-ID	Yes	MACS	0.7-1.9
19-ID-B	Yes	MACS	0.7-1.9
19-ID-C	Yes	MACS	0.7-1.9
19-ID-D	Yes	MACS	0.7-1.9
19-ID-E	Yes	MACS	0.7-1.9
19-ID-F	Yes	MACS	0.7-1.9
19-ID-G	Yes	MACS	0.7-1.9
19-ID-H	Yes	MACS	0.7-1.9
19-ID-I	Yes	MACS	0.7-1.9
19-ID-J	Yes	MACS	0.7-1.9
19-ID-K	Yes	MACS	0.7-1.9
19-ID-L	Yes	MACS	0.7-1.9
19-ID-M	Yes	MACS	0.7-1.9
19-ID-N	Yes	MACS	0.7-1.9
19-ID-O	Yes	MACS	0.7-1.9
19-ID-P	Yes	MACS	0.7-1.9
19-ID-Q	Yes	MACS	0.7-1.9
19-ID-R	Yes	MACS	0.7-1.9
19-ID-S	Yes	MACS	0.7-1.9
19-ID-T	Yes	MACS	0.7-1.9
19-ID-U	Yes	MACS	0.7-1.9
19-ID-V	Yes	MACS	0.7-1.9
19-ID-W	Yes	MACS	0.7-1.9
19-ID-X	Yes	MACS	0.7-1.9
19-ID-Y	Yes	MACS	0.7-1.9
19-ID-Z	Yes	MACS	0.7-1.9
19-ID-AA	Yes	MACS	0.7-1.9
19-ID-AB	Yes	MACS	0.7-1.9
19-ID-AC	Yes	MACS	0.7-1.9
19-ID-AD	Yes	MACS	0.7-1.9
19-ID-AE	Yes	MACS	0.7-1.9
19-ID-AF	Yes	MACS	0.7-1.9
19-ID-AG	Yes	MACS	0.7-1.9
19-ID-AH	Yes	MACS	0.7-1.9
19-ID-AI	Yes	MACS	0.7-1.9
19-ID-AJ	Yes	MACS	0.7-1.9
19-ID-AL	Yes	MACS	0.7-1.9
19-ID-AM	Yes	MACS	0.7-1.9
19-ID-AN	Yes	MACS	0.7-1.9
19-ID-AO	Yes	MACS	0.7-1.9
19-ID-AP	Yes	MACS	0.7-1.9
19-ID-AQ	Yes	MACS	0.7-1.9
19-ID-AR	Yes	MACS	0.7-1.9
19-ID-AS	Yes	MACS	0.7-1.9
19-ID-AT	Yes	MACS	0.7-1.9
19-ID-AU	Yes	MACS	0.7-1.9
19-ID-AV	Yes	MACS	0.7-1.9
19-ID-AW	Yes	MACS	0.7-1.9
19-ID-AX	Yes	MACS	0.7-1.9
19-ID-AY	Yes	MACS	0.7-1.9
19-ID-AZ	Yes	MACS	0.7-1.9
19-ID-BA	Yes	MACS	0.7-1.9
19-ID-BB	Yes	MACS	0.7-1.9
19-ID-BC	Yes	MACS	0.7-1.9
19-ID-BD	Yes	MACS	0.7-1.9
19-ID-BE	Yes	MACS	0.7-1.9
19-ID-BF	Yes	MACS	0.7-1.9
19-ID-BG	Yes	MACS	0.7-1.9
19-ID-BH	Yes	MACS	0.7-1.9
19-ID-BI	Yes	MACS	0.7-1.9
19-ID-BJ	Yes	MACS	0.7-1.9
19-ID-BL	Yes	MACS	0.7-1.9
19-ID-BM	Yes	MACS	0.7-1.9
19-ID-BN	Yes	MACS	0.7-1.9
19-ID-BO	Yes	MACS	0.7-1.9
19-ID-BP	Yes	MACS	0.7-1.9
19-ID-BQ	Yes	MACS	0.7-1.9
19-ID-BR	Yes	MACS	0.7-1.9
19-ID-BS	Yes	MACS	0.7-1.9
19-ID-BT	Yes	MACS	0.7-1.9
19-ID-BU	Yes	MACS	0.7-1.9
19-ID-BV	Yes	MACS	0.7-1.9
19-ID-BW	Yes	MACS	0.7-1.9
19-ID-BX	Yes	MACS	0.7-1.9
19-ID-BY	Yes	MACS	0.7-1.9
19-ID-BZ	Yes	MACS	0.7-1.9
19-ID-CA	Yes	MACS	0.7-1.9
19-ID-CB	Yes	MACS	0.7-1.9
19-ID-CC	Yes	MACS	0.7-1.9
19-ID-CD	Yes	MACS	0.7-1.9
19-ID-CE	Yes	MACS	0.7-1.9
19-ID-CF	Yes	MACS	0.7-1.9
19-ID-CG	Yes	MACS	0.7-1.9
19-ID-CH	Yes	MACS	0.7-1.9
19-ID-CI	Yes	MACS	0.7-1.9
19-ID-CJ	Yes	MACS	0.7-1.9
19-ID-CL	Yes	MACS	0.7-1.9
19-ID-CM	Yes	MACS	0.7-1.9
19-ID-CN	Yes	MACS	0.7-1.9
19-ID-CO	Yes	MACS	0.7-1.9
19-ID-CP	Yes	MACS	0.7-1.9
19-ID-CQ	Yes	MACS	0.7-1.9
19-ID-CR	Yes	MACS	0.7-1.9
19-ID-CS	Yes	MACS	0.7-1.9
19-ID-CT	Yes	MACS	0.7-1.9
19-ID-CU	Yes	MACS	0.7-1.9
19-ID-CV	Yes	MACS	0.7-1.9
19-ID-CW	Yes	MACS	0.7-1.9
19-ID-CX	Yes	MACS	0.7-1.9
19-ID-CY	Yes	MACS	0.7-1.9
19-ID-CZ	Yes	MACS	0.7-1.9
19-ID-DA	Yes	MACS	0.7-1.9
19-ID-DB	Yes	MACS	0.7-1.9
19-ID-DC	Yes	MACS	0.7-1.9
19-ID-DD	Yes	MACS	0.7-1.9
19-ID-DE	Yes	MACS	0.7-1.9
19-ID-DF	Yes	MACS	0.7-1.9
19-ID-DG	Yes	MACS	0.7-1.9
19-ID-DH	Yes	MACS	0.7-1.9
19-ID-DI	Yes	MACS	0.7-1.9
19-ID-DJ	Yes	MACS	0.7-1.9
19-ID-DL	Yes	MACS	0.7-1.9
19-ID-DM	Yes	MACS	0.7-1.9
19-ID-DN	Yes	MACS	0.7-1.9
19-ID-DO	Yes	MACS	0.7-1.9
19-ID-DP	Yes	MACS	0.7-1.9
19-ID-DQ	Yes	MACS	0.7-1.9
19-ID-DR	Yes	MACS	0.7-1.9
19-ID-DS	Yes	MACS	0.7-1.9
19-ID-DT	Yes	MACS	0.7-1.9
19-ID-DU	Yes	MACS	0.7-1.9
19-ID-DV	Yes	MACS	0.7-1.9
19-ID-DW	Yes	MACS	0.7-1.9
19-ID			

Harvesting Data Collection Information for the RCSB PDB Depositions

RCSB PDB PROTEIN DATA BANK www.pdb.org • info@rcsb.org

PDB EXTRACT  **CCP4**

pdextract
<http://pdextract.pdb.org/>
<http://sw-tools.pdb.org/>

- Data collection and reduction
 - HKL, SCALEPACK, d'TREK, SAINT, SCALA
- Molecular replacement
 - CNS, CNX, Amore, Morep, EPMR
- Heavy atom phasing
 - CNS, CNX, SOLVE, MLPHARE, SHARP/autosharp, SHELXD, SHELX, PHASES, SnB, BnP, Phaser
- Density modification
 - CNS, CNX, DM, Solomon, RESOLVE, SHELXE
- Structure refinement
 - CNS, CNX, REFMAC, RESTRAIN, SHELXL, TNT, WARP, PHENIX

NMR - CNS, CYANNA, DYANNA

RCSB PDB PROTEIN DATA BANK www.pdb.org • info@rcsb.org

Current Data Item Coverage

Data Item	PDB_EXTRACT		ALL	
	# pop	# total %	# pop	# total %
reflns.number_all	1049	2239 46.8	16540	37906 43.6
reflns.number_obs	2229	2239 99.5	34154	37906 90.1
reflns.pdbx_Rsym_value	693	2239 30.9	11225	37906 29.6
reflns.percent_possible_obs	2166	2239 96.7	32683	37906 86.2
reflns.pdbx_redundancy	1694	2239 75.6	26547	37906 70.0
reflns.pdbx_Rmerge_I_obs	1848	2239 82.5	23562	37906 62.1
reflns.pdbx_Rsym_value	693	2239 30.9	11225	37906 29.6
reflns.pdbx_netI_over_av_sigmaI	1782	2239 79.5	24282	37906 64.0

RCSB PDB PROTEIN DATA BANK www.pdb.org • info@rcsb.org

Current Data Item Coverage

Data Item	PDB_EXTRACT		ALL	
	# pop	# total %	# pop	# total %
reflns_shell.meanI_over_sigI_obs	1348	2239 60.2	19285	37906 50.8
reflns_shell.number_measured_all	265	2239 11.8	283	37906 0.7
reflns_shell.number_measured_obs	450	2239 20.1	490	37906 1.2
reflns_shell.number_unique_all	1136	2239 50.7	9646	37906 25.4
reflns_shell.number_unique_obs	286	2239 12.7	296	37906 0.7
reflns_shell.percent_possible_all	1893	2239 84.5	28230	37906 74.4
reflns_shell.percent_possible_obs	553	2239 24.7	700	37906 1.8
reflns_shell.Rmerge_I_obs	1812	2239 80.9	19318	37906 50.9
reflns_shell.pdbx_redundancy	1506	2239 67.2	17988	37906 47.4
reflns_shell.pdbx_Rsym_value	654	2239 29.2	9664	37906 25.4
_exptl_crystal.density_Matthews	2201	2239 98.3	19680	37906 51.9
_exptl_crystal.density_percent_sol	2200	2239 98.2	36750	37906 96.9

RCSB PDB PROTEIN DATA BANK www.pdb.org • info@rcsb.org

Current Data Item Coverage

Data Item	PDB_EXTRACT		ALL	
	# pop	# total %	# pop	# total %
diffn.ambient_temp	2139	2239 95.5	32538	37906 85.8
diffn.detector.details	680	2239 30.3	10345	37906 27.2
diffn.detector.detector	2142	2239 95.6	33018	37906 87.1
diffn.detector.type	2112	2239 94.3	32289	37906 85.1
diffn.detector.pdbx_collection_date	2187	2239 97.6	34362	37906 90.6
diffn.radiation.monochromator	1206	2239 53.8	16397	37906 43.2
diffn.radiation.rcsb_diffn_protocol	2239	2239 100.0	33669	37906 88.8
diffn.radiation.wavelength.wavelength	2164	2239 96.6	33233	37906 87.6
diffn.reflns.number	88	2239 3.9	103	37906 0.2
diffn.reflns.av_sigmaI_over_netI	11	2239 0.4	14	37906 0.0
diffn_source.type	2218	2239 99.0	32563	37906 85.9
diffn_source.pdbx_synchrotron_beamline	1835	2239 81.9	22849	37906 60.2
diffn_source.pdbx_synchrotron_site	1835	2239 81.9	23005	37906 60.6
diffn_source.source	2228	2239 99.5	32333	37906 85.3

RCSB PDB PROTEIN DATA BANK www.pdb.org • info@rcsb.org

pdextract - Online
 A MEMBER OF THE PDB
 An Information Portal to Biological Macromolecular Structures

pdextract

The **pdextract** online tool extracts key information from X-ray crystallographic and NMR software applications in preparation for PDB deposition to:

- Reduce the human effort required to assemble complete and validated protein structure entries ready for PDB deposition.
- Prepare an mmCIF file for deposition quickly.

HOW TO RUN:

- Select experimental method (X-Ray or NMR).
- Specify your workstation for the name of the coordinate file obtained from final structure refinement.
- Select the file type and program name.
- Press the RUN button to start this operation.

Coordinate File **File type**

Run **Reset**

NOTE:

- If the file size is large, it is recommended to upload gzipped (.gz) file for faster loading.
- After assembling you data, you will get two mmCIF files for X-Ray or one mmCIF file for NMR. Please load them to ADIT for a complete deposition.

Questions, comments, and suggestions? [Contact Us](#)

Versions

- Latest Release Notes
- [pdextract Online](#)
- [pdextract Workstation](#)

Documentation

- [pdextract Online Tutorial X-ray](#)
- [pdextract Online Tutorial NMR](#)
- [pdextract Manual HTML](#)
- [pdextract Manual PDF](#)

RCSB PDB PROTEIN DATA BANK www.pdb.org • info@rcsb.org

Deposition Services
 A MEMBER OF THE PDB
 An Information Portal to Biological Macromolecular Structures

RCSB PDB Data Validation and Deposition Services

- Beta-ADIT** at RCSB | ADIT at RCSB or PDB | [pdextract](#)
- Beta-ADIT NMR** | Validation Server at RCSB or PDB | [Ligand Depot](#)

The following **data deposition** tools and instructions can make your structure deposition quick, easy, complete and accurate:

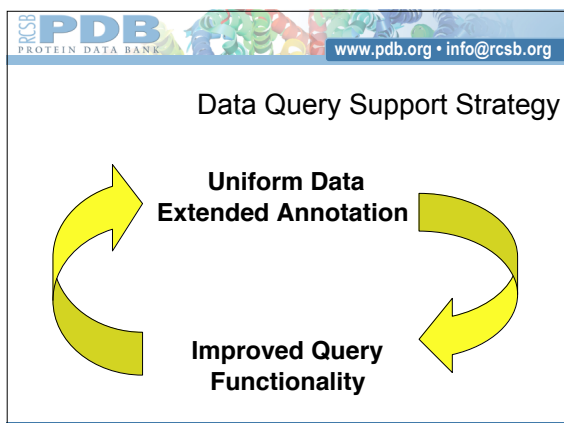
- Prepare** your structural data for deposition using **pdextract** and/or the desktop version of ADIT
- Validate** your structure using the **Validation Server** at RCSB PDB or the **Validation Server** at PDB
- Deposit** your structure using the structure deposition tool **Beta-ADIT** at RCSB PDB or ADIT at RCSB PDB or ADIT at PDB, or using **AutoDep** at MSD-EBI

The RCSB PDB, PDB, and MSD-EBI are members of the **wwPDB**.

- Instructions for X-ray crystallography structure depositions.
- Instructions for NMR structure depositions.
- Instructions for EM structure depositions.
- More information on deposition of structures determined by other methods (including Electron diffraction, Fiber diffraction, Theoretical modeling).

- Search for**
 - Your ligand using [Ligand Depot](#)
 - Appropriate sequence database references for proteins or nucleic acids in your structure (e.g. using [BLAST](#))

Harvesting Data Collection Information for the RCSB PDB Depositions



wwPDB Remediation Project

The wwPDB has collaborated to remediate the PDB archive and create a more consistent set of files.

E-MSD is supported by grants from the Wellcome Trust, the EU (TEMBLOR, NMRQUAL and IIMS), CCP4, the BBSRC, the MRC and EMBL.

PDBj is supported by grant-in-aid from the Institute for Bioinformatics Research and Development, Japan Science and Technology Agency (BIRD-JST), and the Ministry of Education, Culture, Sports, Science and Technology (MEXT).

The **BMRB** is supported by NIH grant LM05799 from the National Library of Medicine.

The **RCSB PDB** is supported by grants from the National Science Foundation, National Institute of General Medical Sciences, the Office of Science-Department of Energy, the National Library of Medicine, the National Cancer Institute, the National Center for Research Resources, the National Institute of Biomedical Imaging and Bioengineering, the National Institute of Neurological Disorders and Stroke, and the National Institute of Diabetes & Digestive & Kidney Diseases.

Worldwide Protein Data Bank
www.wwpdb.org

Data Remediation
<http://remediation.wwpdb.org>

- Sequence and taxonomy
 - Resolved anomalies relative to UniProt (61K sequences)
 - Resolved anomalies between chemical and macromolecular sequence
- Atom nomenclature
 - Conforms to IUPAC for standard amino acids and nucleotides
- Ligands and monomers
 - New dictionary with full chemical description (7800+ definitions)
 - All files annotated relative to this dictionary (163K non-polymers + 101K polymer residues)
- Biological assembly defined
 - Viruses now uniformly annotated (250 entries)

Worldwide Protein Data Bank
www.wwpdb.org

Nomenclature Standardization

- IUPAC-compliant H-atom names for standard amino acids and nucleotides (exceptions: OXT and HXT)
- DNA and RNA now differentiated (Adenine is DA for DNA; A for RNA)
- Modified nucleotides expressed as 3-letter codes
 - (replacing +T, +C ... etc.)
- PDB asterisks replaced by single quotes in atom labels (O2* becomes O2')
- Unusual names made more conventional (AC8 becomes C8A)
- White-space characters removed (N 2 becomes N2)

Worldwide Protein Data Bank
www.wwpdb.org

Chemical Dictionary

- Stereochemical assignments
- Aromatic bond assignments
- Nomenclature
 - IUPAC nomenclature for standard amino acids and nucleotides (Pure & Applied Chem., 70, 117-142, 1998)
 - All atom labels begin with element type symbol
 - Retention of all prior names as an alternate identifier
- Model and idealized coordinates
- Chemical descriptors (SMILES & InChI)
- Systematic chemical names
- Redundant chemical components obsoleted
- Additional definitions for protonated forms

Worldwide Protein Data Bank
www.wwpdb.org

Protonation Variants

- Companion dictionary of 220 definitions
- Additional definitions provide nomenclature for protonation states of:
 - N, C-terminal and free amino f
 - common side chain variants
- IUPAC compliant names
- Covers BMRB & CCPN variants

The diagram shows three chemical structures of a peptide backbone segment. The first structure shows a neutral state with an amine group (NH2) and a carboxylate group (COO-). The second structure shows a protonated amine group (NH3+). The third structure shows a protonated carboxyl group (COOH). The structures are connected by a peptide bond.

Harvesting Data Collection Information for the RCSB PDB Depositions

Worldwide Protein Data Bank
www.wwpdb.org

Virus Structures

Worldwide Protein Data Bank
www.wwpdb.org

<http://remediation.wwpdb.org>
for details and dictionaries

Remediation Test Site

Component definition ATP

wwPDB Remediation Test Site

The full remediated archive is now available at <http://remediation.wwpdb.org> for testing. This site will production use at <http://www.pdb.org>. Both sites share the same organization structure. The access site <http://www.wwpdb.org/remediation-downloads.html>.

The Chemical Component Dictionary and smaller test sets of data files are also available for testing.

PDB users are encouraged to test the remediated data files between April and July 2007. As of July, if site containing remediated data files. The final transition date will be announced on this website.

Comments about the files should be sent to info@wwpdb.org. Major announcements will be made at 5 websites.

About the wwPDB Remediation Project

The evolution of experimental methods, functional knowledge of proteins, and methods used to process archive. The wwPDB has remediated these data to create a more uniform archive.

The results of these efforts can be found in the remediated coordinate files. Highlights include:

Sequence	Updated references to databases and taxonomies
Classion	Resolved differences between chemical and macromolecular sequences
Assembly and virus information	Verified and updated primary citation assignments
Nucleic acid labeling	Improved representation of deposited and experimental coordinate frame
Baseline data	Density and ribbon nucleotides assigned separate chemical definitions. The RNA forms remain labeled as A, G, C, U, T.
Atom nomenclature	Standardized to reflect changes in Chemical Component Dictionary (see below)

Properties

Name	Formula	Formal charge	Chemical features
ADENOSINE-5'-TRIPHOSPHATE	C ₁₀ H ₁₆ N ₅ O ₁₃ P ₃	0	Atom count: 47 Disal atom count: 5 Disal atom: PB PA GA GA GP GP Bond count: 40 Aromatic bond count: 10

Worldwide Protein Data Bank
www.wwpdb.org

Testing and Roll-out

- Remediated data files and chemical dictionary previewed by 60 software developers and database maintainers (November 2006)
- Example files files and chemical components dictionary released for public review (January 2007)
- Full remediated archive released for public review in April 2007
- Planned review period continues through July 2007
- Further details of the review updated at <http://www.pdb.org>, <http://remediation.wwpdb.org> and wwPDB sites.

Images of remediated PDB entries 1tk0, 2bvv, and 407d obtained using Jmol, Chimera & OpenRasMol

RCSB PDB
www.pdb.org • info@rcsb.org

Access

RCSB PDB

- <http://www.pdb.org/>

wwPDB

- <http://www.wwpdb.org/>

wwPDB Remediation

- <http://remediation.wwpdb.org/>

Dictionary & Schema Resources

- <http://mmcif.pdb.org/> & <http://pdml.pdb.org>

TargetDB & PepcDB

- <http://targetdb.pdb.org> & <http://pepcdb.pdb.org>

RCSB Software Download Site

- <http://sw-tools.pdb.org>
- CVS server rcsb-cvs.rcsb.org - anonymous access

RCSB PDB
www.pdb.org • info@rcsb.org

The RCSB PDB Team

RCSB PDB
www.pdb.org • info@rcsb.org

Acknowledgements

Operated by two members of the RCSB:

RUTGERS SDSC UCSD
SARGAS SCHOOL OF PHARMACY and PHARMACEUTICAL SCIENCES

The RCSB PDB is a member of the

WORLDWIDE PDB PROTEIN DATA BANK

Supported by:

NST NIGMS NIDDK
NATIONAL CANCER INSTITUTE NATIONAL CENTER for Research Resources NIDDK