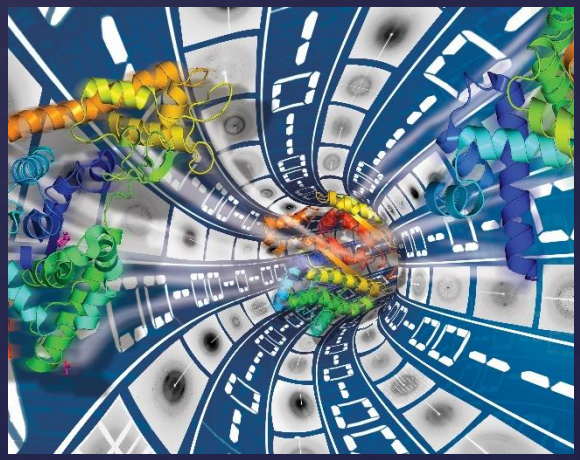
Integrated Resource for Reproducibility in Macromolecular Crystallography

Goal:

Protein Crystallography (Structural Biology) with Speed and Finesse



Wladek Minor

ACA, New Orleans 2017

Goals

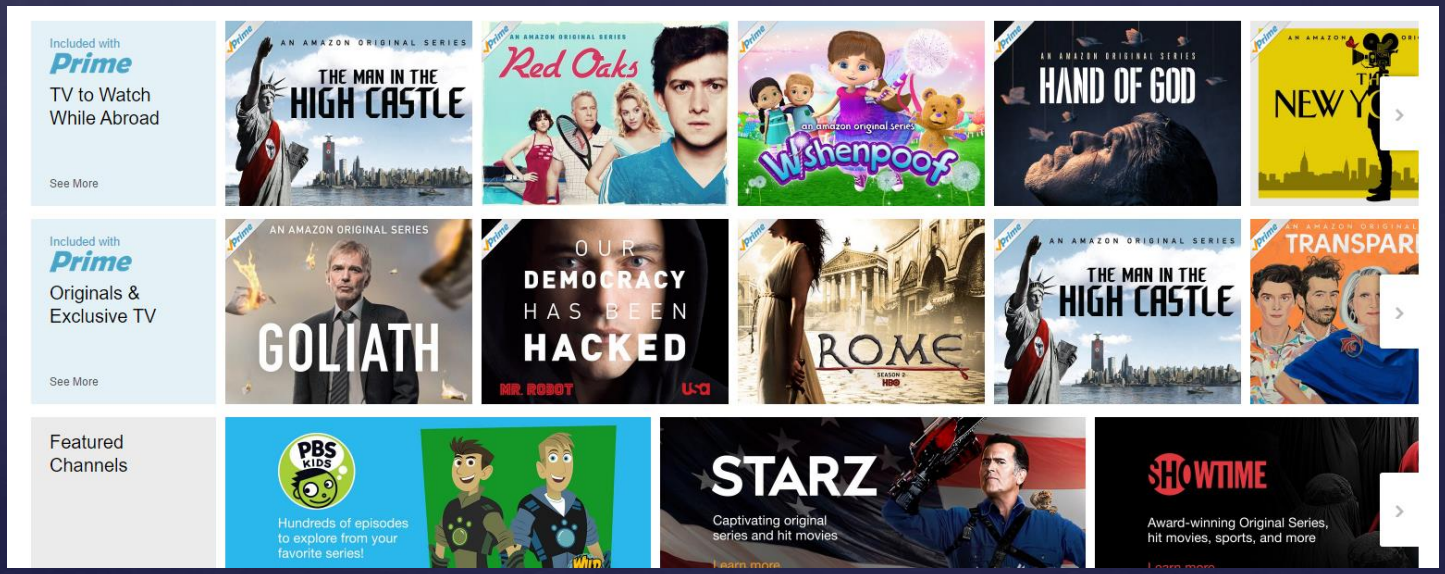
- Develop tools for automatically extracting and curating diffraction images and associated metadata, as well as producing detailed descriptions of all data needed for later reprocessing of the diffraction data as methods for structure determination improve.

- Create a web-based system for semantic searching, analysis, and data mining of appropriate subsets of diffraction images and associated metadata.

Goals

- Develop tools to automatically validate, preprocess, and score diffraction images, and to detect potential issues and errors.
- Creation of a repository for diffraction data that did not yield an X-ray structure with the currently available methods.
- Set up a pilot resource incorporating the tools developed in Aims 1-4 to collect a test set of data for development of tools and algorithms for validation and error detection.

Big Data *or not*



http://www.proteindiffraction.org

Home | About | Browse | Statistics | Submit data | News | Login

Search diffraction images

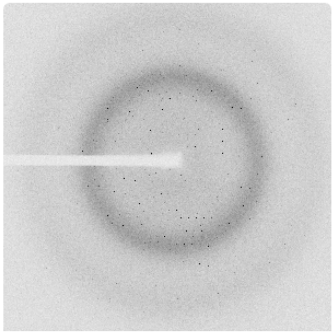
National Institutes of Health
Office of the Director
Data Science at NIH

Integrated Resource for Reproducibility in Macromolecular Crystallography

This project is being funded by the Targeted Software Development award 1 U01 HG008424-01 as part of the BD2K (Big Data to Knowledge) program of the National Institute of Health. The project is developing tools for "wrangling" data from protein diffraction experiments. We are also creating a growing repository of diffraction experiments used to determine protein structures in the PDB, contributed by the CSGID, SSGCID, JCSG, MCSG, SGC, and other large-scale projects, as well as individual research laboratories.

Currently indexed projects: **3300**

Currently indexed datasets: **6274**

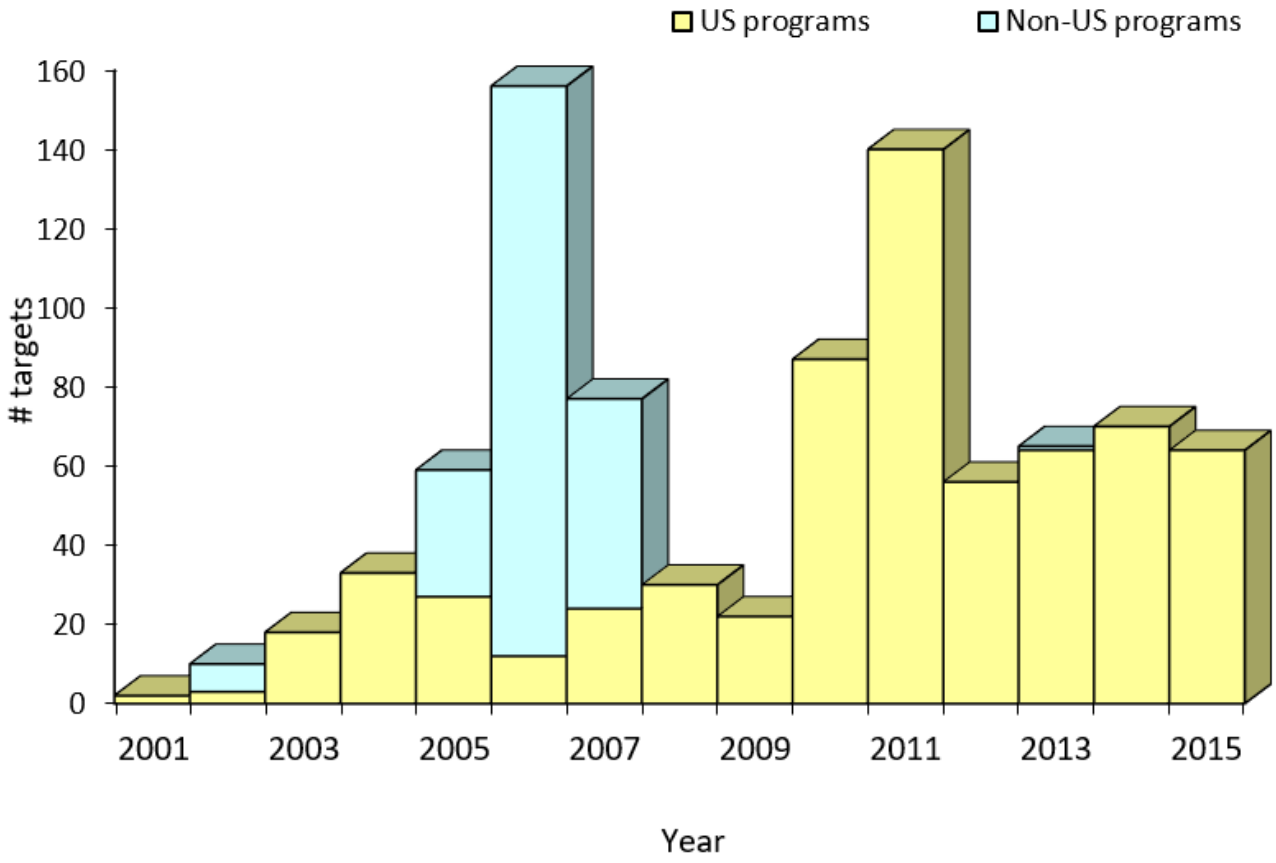


Browse & search | Statistics | Submit data | Publications | Beamlines | Educational resources

Search examples

Find a specific PDB ID: 4K6A

Data collected but unsolved ?

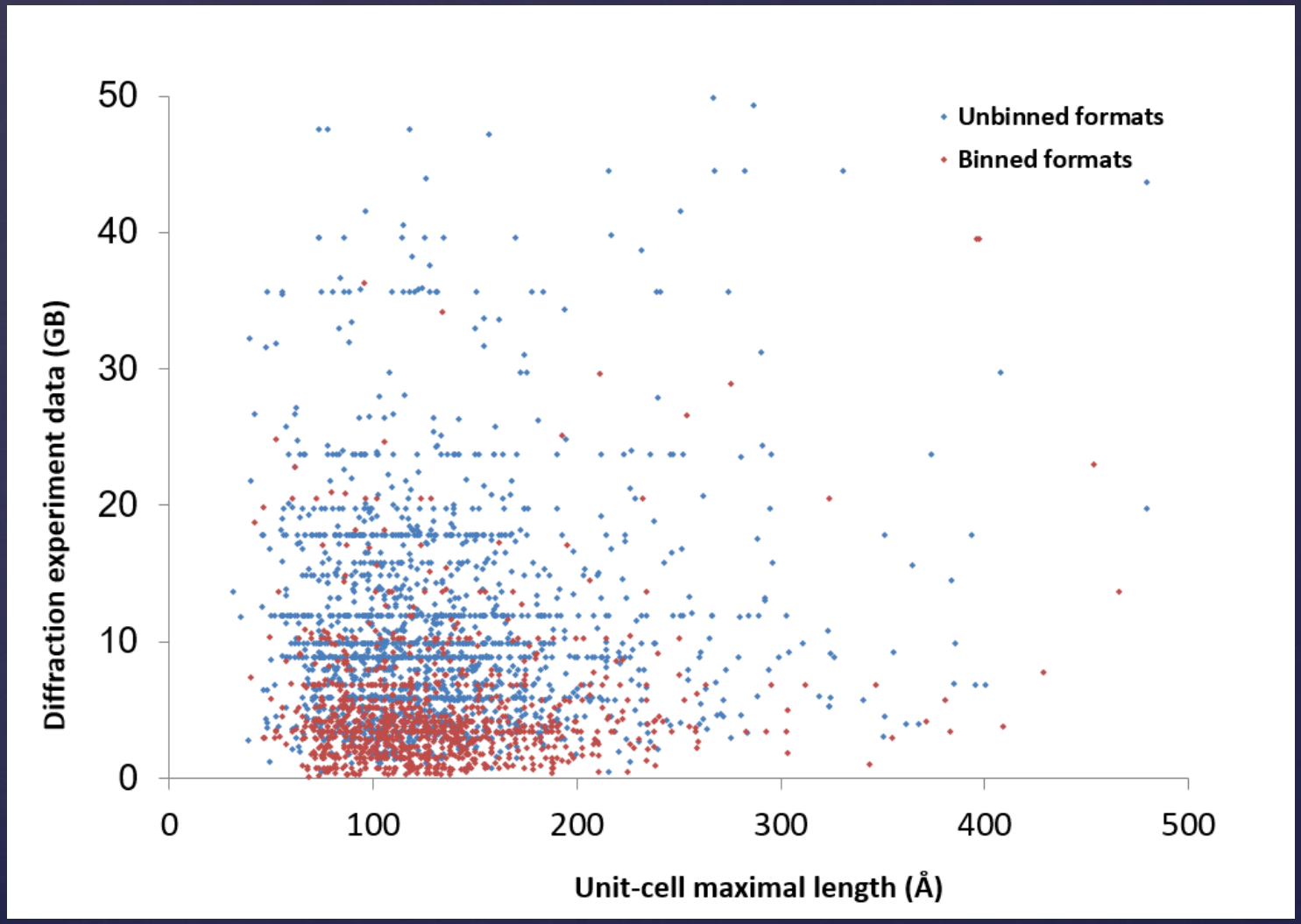


Zheng *et al*, *Expert Opinion on Drug Discovery(2014) 9: 125-37*

Metadata

Metadata source	Metadata parameters
User	Identity of the user Identity of people who collected the data Location and date of data collection (beamline, home source, etc.) Identity of the protein (e.g. GenBank, Uniprot identifiers) PDB identifier of solved structure (if deposited) Custom labels
Diffraction images	Detector type and serial numbers, and image format, Data collection parameters: number of frames, oscillation step size, experimental orientation angles (e.g. κ,φ,ω, and 2θ), detector distance
Structure factors/scaling logs	Integrated reflection data Nominal resolution cutoff Completeness, overall and highest resolution shell (HRS) Redundancy, overall and HRS Mean I/sigma I, overall and HRS Software used to process diffraction images R_{merge}, R_{meas}, R_{pim}
Automatic reprocessing	Validation of provided/extracted metadata Validation of spacegroup, Validation of merging statistics from deposited structure factors/scaling logs Estimation of radiation damage Estimation of crystal internal non-isomorphicity Presence/Strength of anomalous signal Diffraction image artifacts and other "features" (background scattering, ice rings, diffuse scattering, etc.)
Molecular models	Structure determination methods (SAD/MAD/MR etc.) Programs used to determine the structure Structure refinement methods Programs used to refine structure R/R_{free} Electron density maps (calculated or extracted from the Uppsala Electron Density Server)
SG databases/LIMS	Sample preparation data Target justification and selection criteria Crystallization conditions
External databases	PDB, GenBank, Uniprot, PubMed

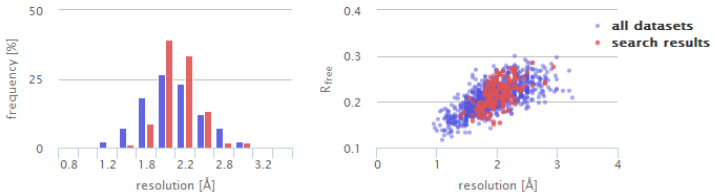
IRRMC: Dataset sizes 400 MB – 50 GB



http://www.proteindiffraction.org

Home | About | Browse | Statistics | Submit data | beamline=home_source

There were **170** results found for this search, out of a total 2719 diffraction experiments.

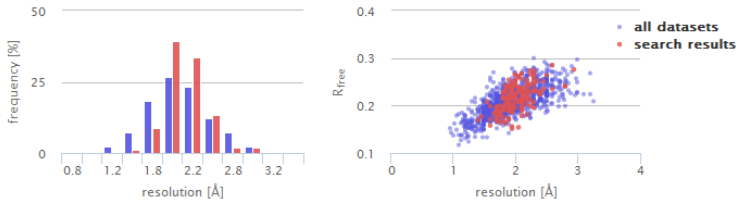


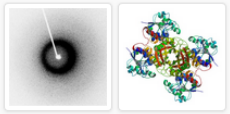
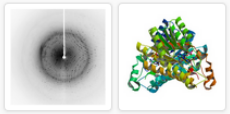
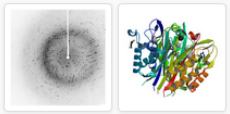
Project	Structure	Resolution	Beamline	Date
3DR6 (CSGID)	Ynca, a putative ACETYLTRANSFERASE from Salmonella typhimurium	1.75 Å	Home source	
3LNT (SSGCID)	Phosphoglyceromutase from Burkholderia Pseudomallei 1710B with bound maloni...	2.10 Å	Home source	
3DR3 (CSGID)	Idp00107, a potential N-acetyl-gamma-glutamylphosphate reductase from Shige...	2.00 Å	Home source	
3UJH (SSGCID)	Substrate-bound Glucose-6-phosphate isomerase from Toxoplasma gondii	2.10 Å	Home source	
4RGB (SSGCID)	A putative carveol dehydrogenase from Mycobacterium avium bound to NAD	1.95 Å	Home source	
3S99 (SSGCID)	A basic membrane lipoprotein from brucella melitensis, iodide soak	2.05 Å	Home source	
4XGI (SSGCID)	Glutamate dehydrogenase from Burkholderia thailandensis	2.00 Å	Home source	
3P10 (SSGCID)	2-c-methyl-d-erythritol 2,4-cyclodiphosphate synthase from Burkholderia pse...	1.70 Å	Home source	
3O2E (SSGCID)	A bol-like protein from babesia bovis	1.95 Å	Home source	
3LUZ (SSGCID)	Extragenic suppressor protein suhB from Bartonella henselae, via combined i...	2.05 Å	Home source	
3N5O (SSGCID)	Putative glutathione transferase from Coccidioides immitis bound to glutath...	1.85 Å	Home source	
4DDD (SSGCID)	An immunogenic protein from ehrlichia chaffeensis	1.90 Å	Home source	
3KM3 (SSGCID)	Eoxycytidine triphosphate deaminase from anaplasma phagocytophilum at 2.1A ...	2.10 Å	Home source	
4DDO (SSGCID)	3-oxoacyl-[acyl-carrier-protein] synthase ii from burkholderia vietnamiensi...	1.90 Å	Home source	
3S6L (SSGCID)	A YadA-like head domain of the trimeric autotransporter adhesin BoaA from B...	2.30 Å	Home source	
3S6M (SSGCID)	The structure of a Peptidyl-prolyl cis-trans isomerase from Burkholderia ps...	1.65 Å	Home source	

http://www.proteindiffraction.org

Home | About | Browse | Statistics | Submit data | beamline=home_source

There were **170** results found for this search, out of a total 2719 diffraction experiments.

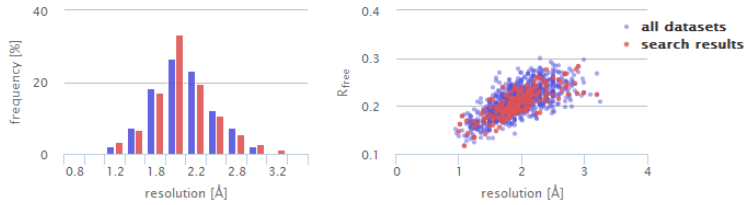


Project	Structure	Resolution	Beamline	Date
3DR6 (CSGID)	Ynca, a putative ACETYLTRANSFERASE from Salmonella typhimurium	1.75 Å	Home source	
3LNT (SSGCID)	Phosphoglyceromutase from Burkholderia Pseudomallei 1710B with bound maloni...	2.10 Å	Home source	
3DR3 (CSGID)	Idp00107, a potential N-acetyl-gamma-glutamylphosphate reductase from Shige...	2.00 Å	Home source	
	 First author: A.U. Singer Uniprot: P59310 Gene name: argC Space group: P 62 2 2 R/R_{free}: 0.17/0.22 I/σ in last shell: 10.1 View dataset details Download all images (1.1GB) CSGID website for IDP00107 PDB website for 3DR3			
3UJH (SSGCID)	Substrate-bound Glucose-6-phosphate isomerase from Toxoplasma gondii	2.10 Å	Home source	
4RGB (SSGCID)	A putative carveol dehydrogenase from Mycobacterium avium bound to NAD	1.95 Å	Home source	
	 First author: T.E. Edwards Uniprot: A0QDP5 Gene name: not available Space group: C 2 2 21 R/R_{free}: 0.21/0.24 I/σ in last shell: 2.8 View dataset details Download all images (1.0GB) SSGCID website for SSGCID-MyavA.01326.g PDB website for 4RGB			
3S99 (SSGCID)	A basic membrane lipoprotein from brucella melitensis, iodide soak	2.05 Å	Home source	
4XGI (SSGCID)	Glutamate dehydrogenase from Burkholderia thailandensis	2.00 Å	Home source	
3P10 (SSGCID)	2-c-methyl-d-erythritol 2,4-cyclodiphosphate synthase from Burkholderia pse...	1.70 Å	Home source	
	 First author: L. Baugh Uniprot: Q3JRA0 Gene name: not available Space group: C 1 2 1 R/R_{free}: 0.16/0.19 I/σ in last shell: 5.7 View dataset details Download all images (0.3GB) SSGCID website for BupsA.00122.a PDB website for 3P10			
3O2E (SSGCID)	A bol-like protein from babesia bovis	1.95 Å	Home source	

http://www.proteindiffraction.org

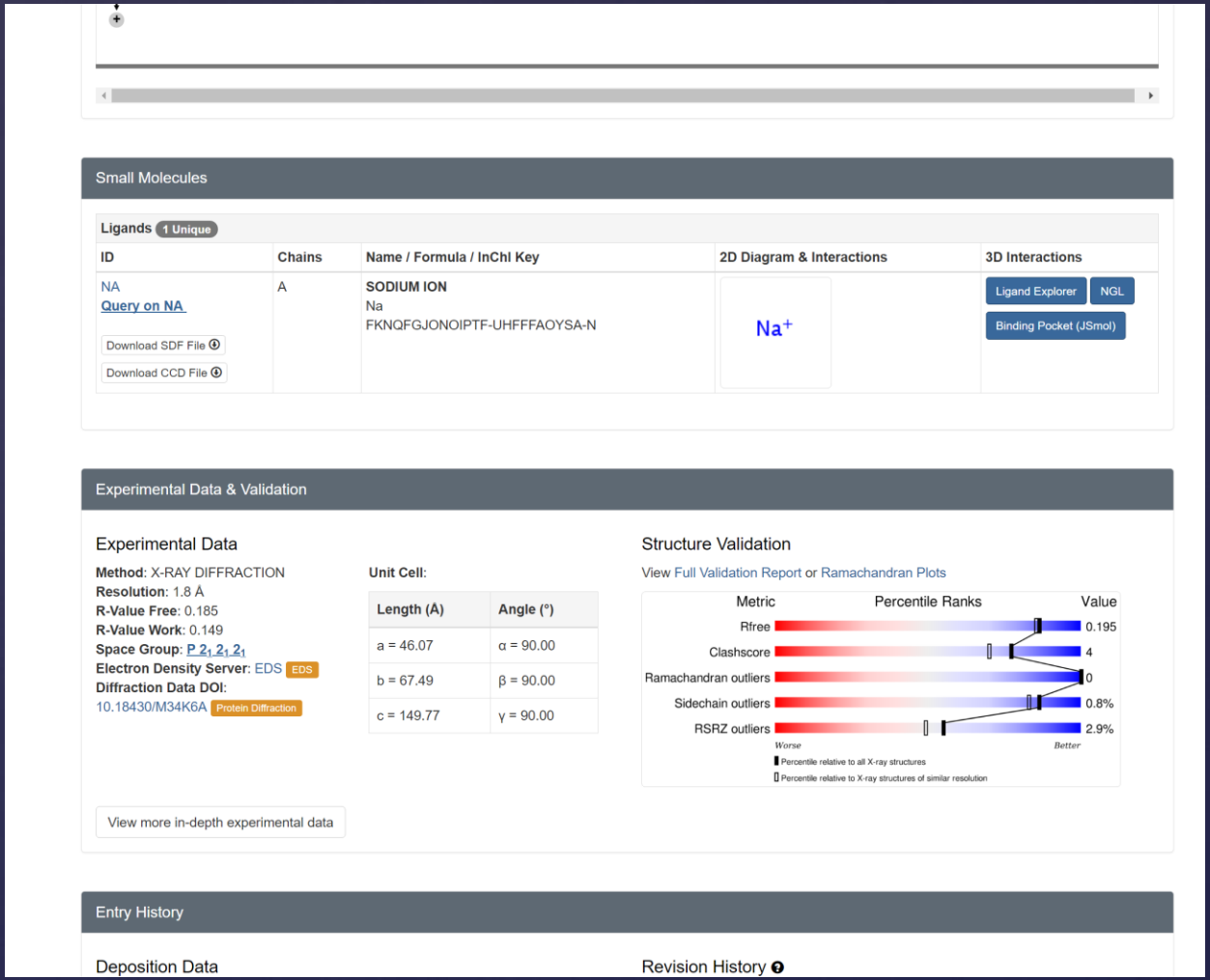
Home | About | Browse | Statistics | Submit data | beamline=21-ID-G

There were **180** results found for this search, out of a total 2719 diffraction experiments.



Project	Structure	Resolution	Beamline	Date
4EQ9 (CSGID)	1.4 Angstrom Crystal Structure of ABC Transporter Glutathione-Binding Prote...	1.40 Å	APS / 21-ID-G	
4X9K (CSGID)	Beta-ketoacyl-acyl carrier protein synthase III-2 (FabH2)(C113A) from Vibri...	1.61 Å	APS / 21-ID-G	
4EAQ (MCSG)	Thymidylate Kinase from Staphylococcus aureus in complex with 3'-Azido-3'-D...	1.85 Å	APS / 21-ID-G	
3QTB	Universal stress protein from Archaeoglobus fulgidus in complex with dAMP	2.10 Å	APS / 21-ID-G	
4JBE (MCSG)	1.95 Angstrom Crystal Structure of Gamma-glutamyl phosphate Reductase from ...	1.95 Å	APS / 21-ID-G	
4KWT (CSGID)	Unliganded anabolic ornithine carbamoyltransferase from Vibrio vulnificus a...	1.86 Å	APS / 21-ID-G	
4GIB (CSGID)	2.27 Angstrom Crystal Structure of beta-Phosphoglucomutase (pgmB) from Clos...	2.27 Å	APS / 21-ID-G	
4JG9 (CSGID)	X-ray Crystal Structure of a Putative Lipoprotein from Bacillus anthracis	2.42 Å	APS / 21-ID-G	
4OC9 (CSGID)	2.35 Angstrom resolution crystal structure of putative O-acetylhomoserine (...	2.35 Å	APS / 21-ID-G	
4HVN (MCSG)	Hypothetical protein with ketosteroid isomerase-like protein fold from Cate...	1.95 Å	APS / 21-ID-G	
3NNT (CSGID)	K170m Mutant of Type I 3-Dehydroquinate Dehydratase (aroD) from Salmonella ...	1.60 Å	APS / 21-ID-G	
3M07 (CSGID)	1.4 Angstrom Resolution Crystal Structure of Putative alpha Amylase from Sa...	1.40 Å	APS / 21-ID-G	
3LAY (CSGID)	Alpha-Helical barrel formed by the decamer of the zinc resistance-associate...	2.70 Å	APS / 21-ID-G	
4OJ7 (SSGCID)	Chorismate Mutase from Burkholderia thailandensis	2.15 Å	APS / 21-ID-G	
3TMQ (SSGCID)	A 2-dehydro-3-deoxyphosphooctonate aldolase from Burkholderia pseudomallei ...	2.10 Å	APS / 21-ID-G	
3IJ3 (CSGID)	1.8 Angstrom Resolution Crystal Structure of Cytosol Aminopeptidase from Co...	1.80 Å	APS / 21-ID-G	

http://www.proteindiffraction.org



Small Molecules

Ligands 1 Unique

ID	Chains	Name / Formula / InChI Key	2D Diagram & Interactions	3D Interactions
NA Query on NA Download SDF File Download CCD File	A	SODIUM ION Na FKNQFGJONOIPTF-UHFFFAOYSA-N	Na^+	Ligand Explorer NGL Binding Pocket (JSmol)

Experimental Data & Validation

Experimental Data

Method: X-RAY DIFFRACTION
Resolution: 1.8 Å
R-Value Free: 0.185
R-Value Work: 0.149
Space Group: $P\,2_1\,2_1\,2_1$
Electron Density Server: EDS EDS
Diffraction Data DOI:
10.18430/M34K6A Protein Diffraction

Unit Cell:

Length (Å)	Angle (°)
a = 46.07	α = 90.00
b = 67.49	β = 90.00
c = 149.77	γ = 90.00

Structure Validation

View Full Validation Report or Ramachandran Plots

View more in-depth experimental data

Entry History

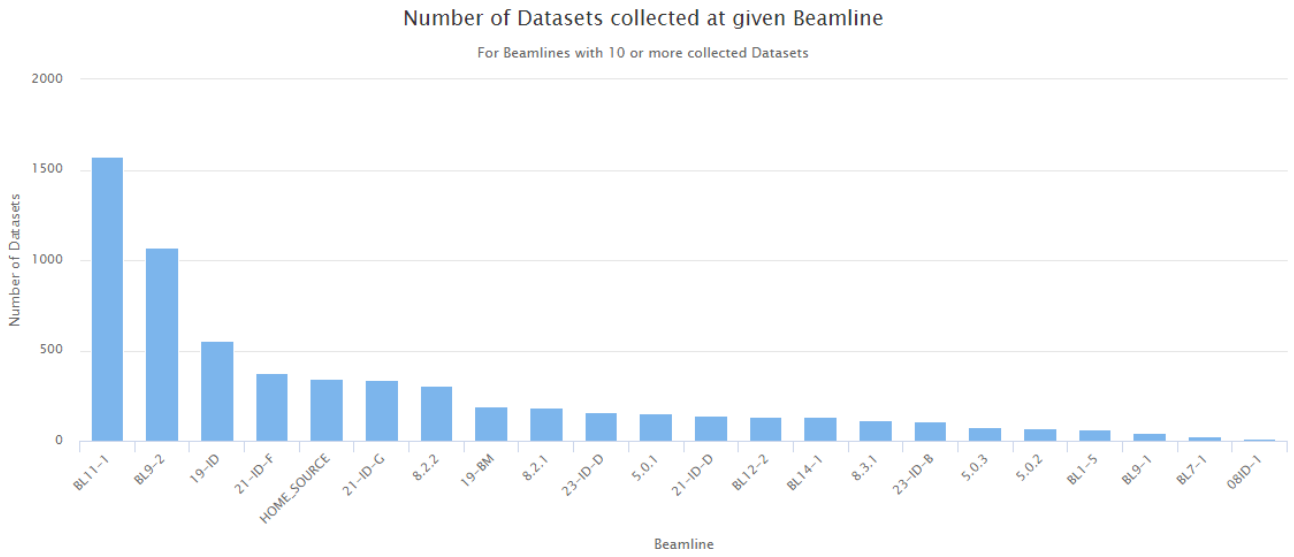
Deposition Data

Revision History

http://www.proteindiffraction.org

Home About Browse Statistics Submit data News Login

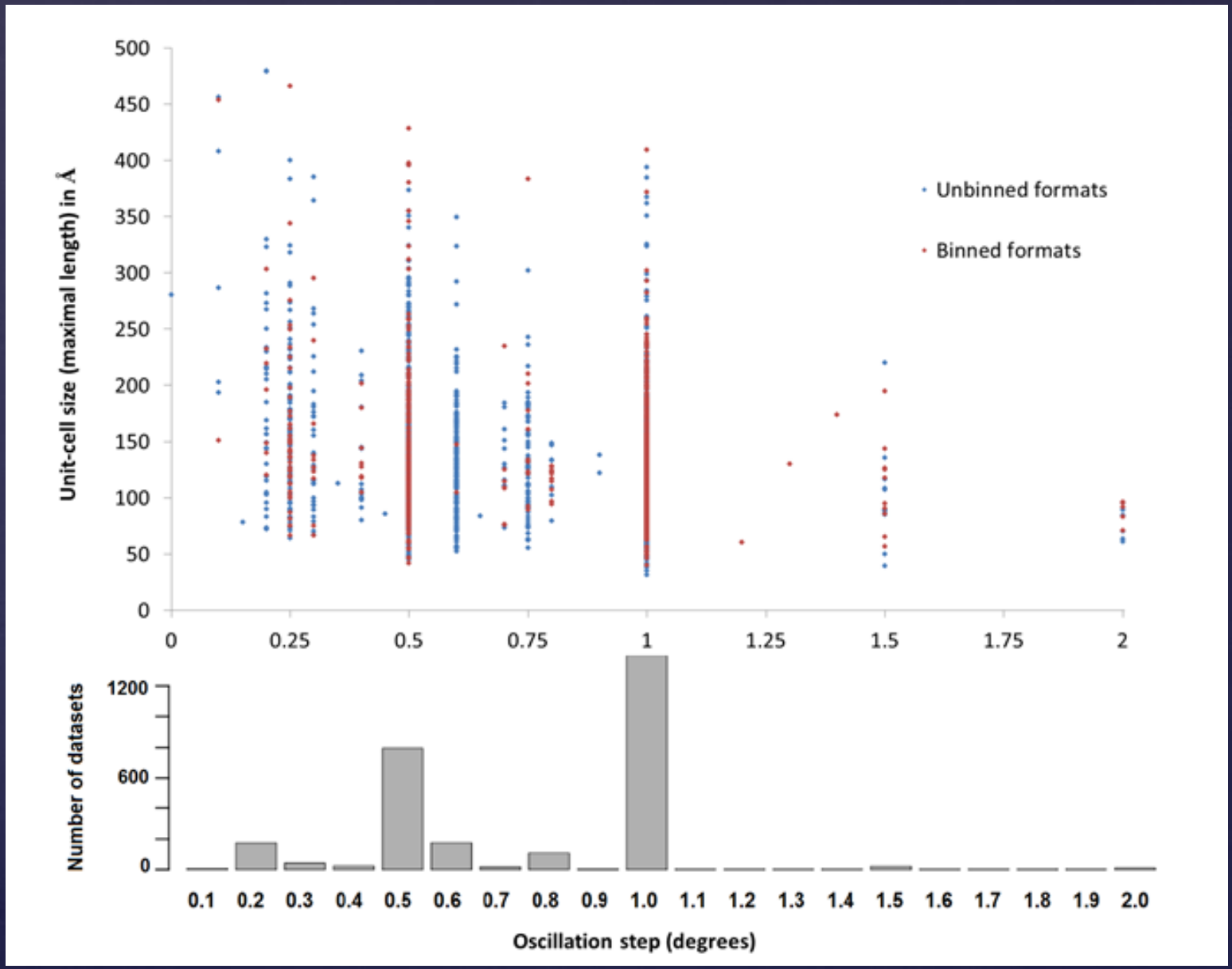
Search diffraction images



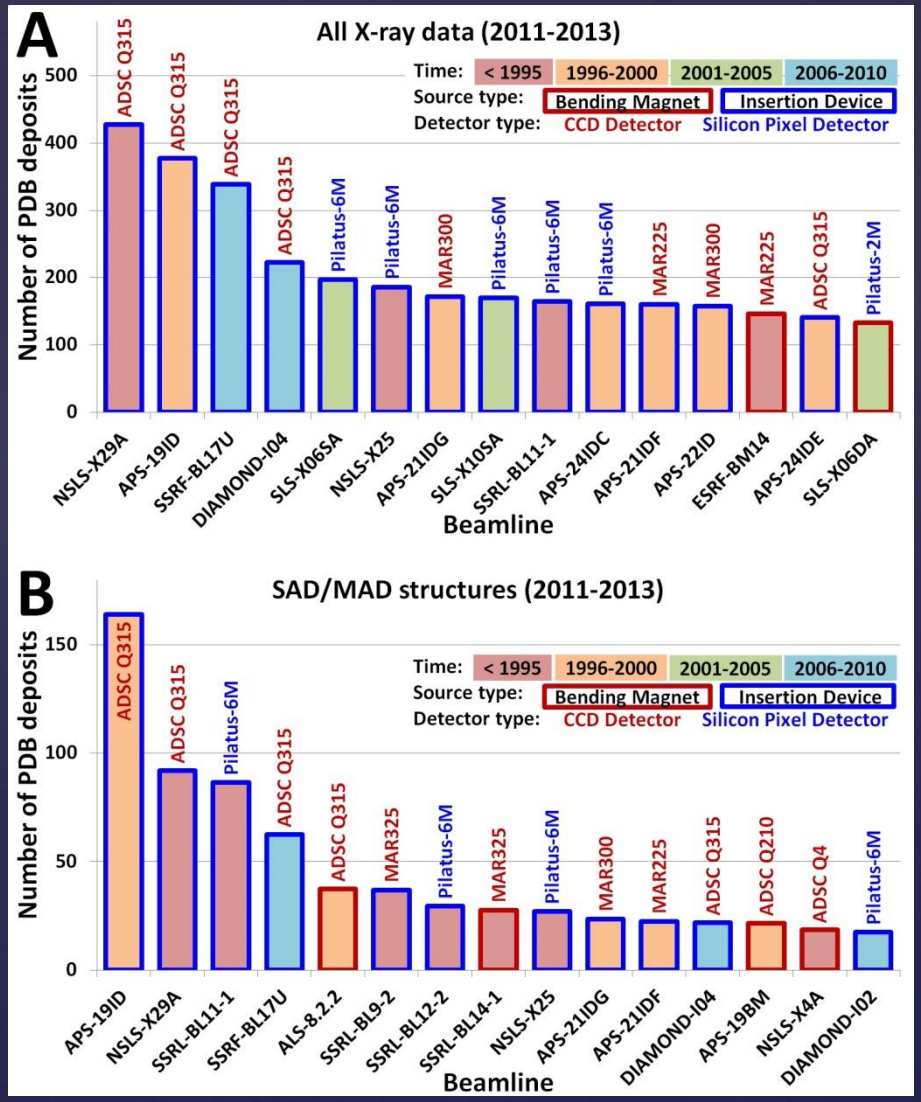
Beamlines (43)

Beamline	Synchrotron	Datasets
BL11-1	SSRL (Stanford Synchrotron Radiation Laboratory)	1576
BL9-2	SSRL (Stanford Synchrotron Radiation Laboratory)	1071
19-ID	APS (Advanced Photon Source)	552
21-ID-F	APS (Advanced Photon Source)	375

Data collection strategy

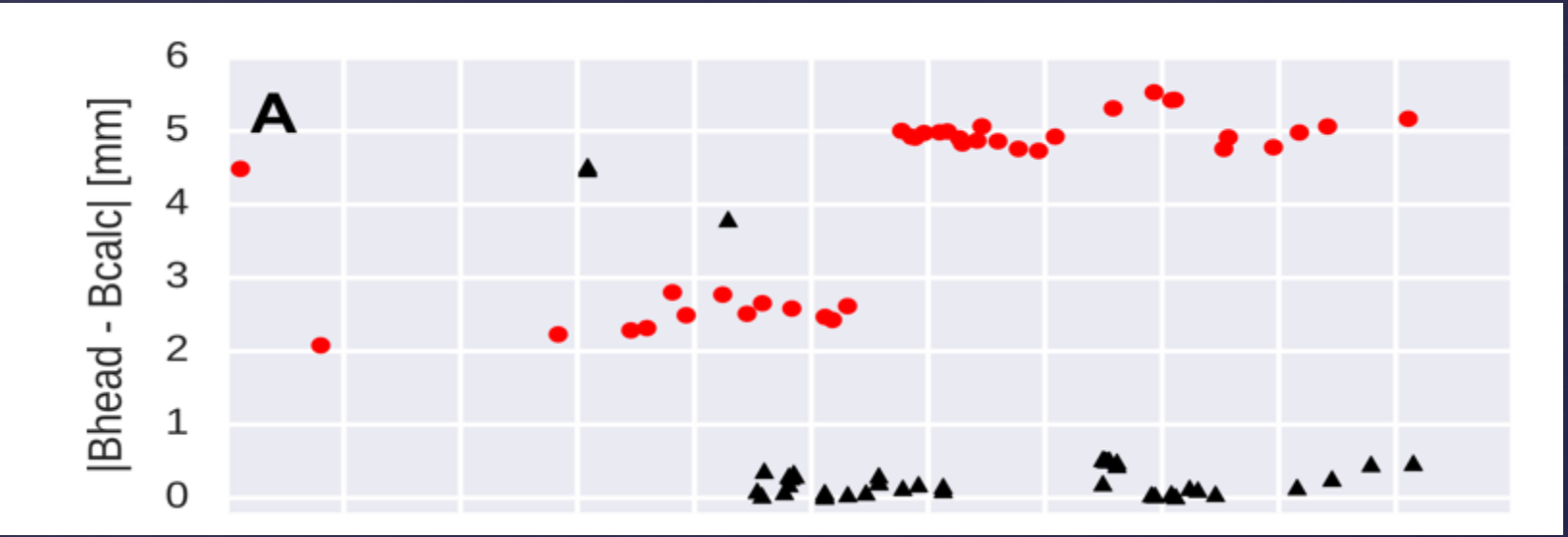


Where we should collect data ?



Zheng *et al*, *Expert Opinion on Drug Discovery(2014) 9: 125-37*

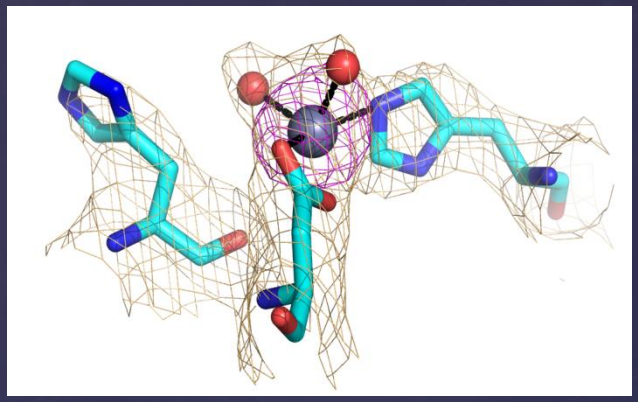
Header metadata – two beamlines ?



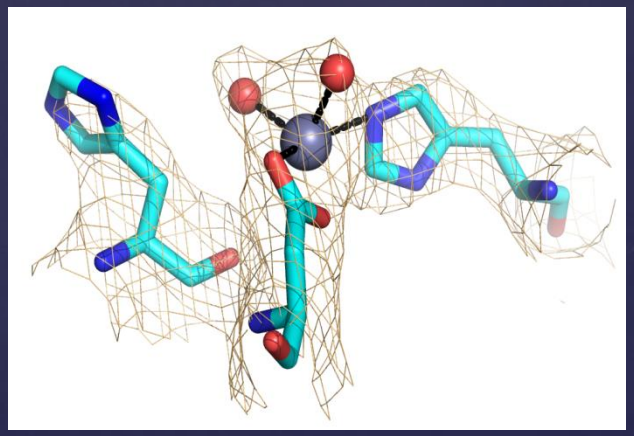
PX Synchrotron Station

Detector
Flux
Goniostat stability and reliability
Sample movement
Time needed for wavelength change
Location of the crystal
Size of the beam

The way to prove presence of some metals



Metal binding site A
9668eV



Metal binding site A
9618eV

Data collected below and above zinc absorption edge - APS 19BM

Handing et al. Chem. Sci. 7: 6635-6648

Protein purification and crystallization artifacts: The tale usually not told

Table I. *Known Structures of Proteins that Have Been Identified as Common Purification and Crystallization Artifacts*

	Name of the protein	Molecular weight (kDa)	PDB ID
Affinity, solubility, anti-aggregation tags	Maltose-binding protein (MBP)	43	1LLS, 1MPB, 3PUW, 3SEU, 4KYC
	Glutathione-S-transferase (GST)	24	4ECB
	Thioredoxin (Trx)	11	1F6M, 2AJQ, 2H73, 4HU9, 4X43
	N-Utilization substance (NusA)	55	1U9L,[a] 1WCN,[b] 2KWP,[c] 4MTN[d]
	Small ubiquitin related modifier 1 (SUMO1)	12	2UYZ, 1Z5S, 4WJQ, 2IO2
	Haloalkane dehalogenase	33	4E46
***E. coli* native proteins**	Metal-binding lipocalin (YodA)	25	1OEJ, 4TNN
	Carbonic anhydrase (YadF)	25	2ESF
	Ferric uptake regulator (Fur)	16	2FU4
	cAMP-regulatory protein (CRP)	24	1CGP, 2CGP, 2GZW, 3FWE, 3HIF, 3N4M, 3QOP, 4FT8, 4HZF, 4I0A, 4I0B, 4N9H, 4N9I
	Glucosamine-6-phosphate synthase (GlmS)	67	4AMV, 1JXA, 3OOJ, 2J6H
	Glycogen synthase (GlgA)	53	2QZS
	Component 1 of the 2-oxoglutarate dehydrogenase complex (ODO1)	105	2JGD
	Component E2 of dihydrolipoamide succinyltransferase (ODO2)	44	1C4T
	Formyl transferase (YfbG, ArnA)	46	1U9J, 1YRW, 1Z7E, 2BLN, 4WKG
	Cu/Zn-superoxide dismutase (Cu/Zn-SODM)	16	1ESO
	Chloramphenicol-O-acetyl transferase (CAT)	26	1Q23
	Host factor-I protein (Hfq)	11	3VU3
Proteases	Tobacco etch virus (TEV)	28	1LVM
	Rhinovirus 3C protease	48	1CQQ
	SUMO protease C-terminal domain	26	2HL9
	Enterokinase	26	1EKB
	Trypsin	26	3UY9
	Chymotrypsin	26	1GGD
	Thrombin (active form)	36	3SQE, 1MH0, 4H6T
	Thermolysin	60	4D9W
	Proteinase K	40	3DVS
	Pepsin	41	5PEP
	Neutrophil elastase	29	5ABW
	LysN Peptidyl-Lys metalloendopeptidase	44	1GE7
	Lysyl endopeptidase	28	4NSY
	Factor Xa	55	1KIG
Exogenous proteins	Lysozyme	16	4TWS, 4PRQ, 1AKI
	DNase protein	31	2A40

Niedzialkowska *et al*, *Protein Sci. 2016 Mar;25(3):720-33*

Reproducibility

The **unspoken rule is that at least 50% of the studies published even in top tier academic journals – *Science, Nature, Cell, PNAS,* etc… – can't be repeated** with the same conclusions by an industrial lab. In particular, key animal models often don't reproduce. This 50% failure rate isn't a data free assertion: it's backed up by dozens of experienced R&D professionals who've participated in the (re)testing of academic findings. This is a huge problem for translational research and one that won't go away until we address it head on.

REPRODUCIBILITY OF RESEARCH FINDINGS

Preclinical research generates many secondary publications, even when results cannot be reproduced.

Journal impact factor	Number of articles	Mean number of citations of non-reproduced articles*	Mean number of citations of reproduced articles
>20	21	248 (range 3–800)	231 (range 82–519)
5–19	32	169 (range 6–1,909)	13 (range 3–24)

Results from ten-year retrospective analysis of experiments performed prospectively. The term 'non-reproduced' was assigned on the basis of findings not being sufficiently robust to drive a drug-development programme.
*Source of citations: Google Scholar, May 2011.

Nature (2012) **483**, 531-533

The Magnitude of Reproducibility problem

Table 1. Examples of Some Reported Reproducibility Concerns in Preclinical Studies

Author	Field	Reported Concerns
Ioannidis et al (2009)[22]	Microarray data	16/18 studies unable to be reproduced in principle from raw data
Baggerly et al (2009)[23]	Microarray data	Multiple; insufficient data/poor documentation
Sena et al (2010)[24]	Stroke animal studies	Overt publication bias: only 2% of the studies were negative
Prinz (2011)[1]	General biology	75% to 80% of 67 studies were not reproduced
Begley & Ellis (2012)[2]	Oncology	90% of 53 studies were not reproduced
Nekrutenko & Taylor(2012)[25]	NGS data access	26/50 no access to primary data sets/software
Perrin (2014)[26]	Mouse, in-vivo	0/100 reported treatments repeated positive in studies of ALS
Tsilidis et al (2013)[27]	Neurological studies	Too many significant results, overt selective reporting bias
Lazic & Essioux (2013)[28]	Mouse VPA model	Only 3/34 used correct experimental measure
Haibe-Kains et al (2013)[29]	Genomics/cell line analysis	Direct comparison of 15 drugs and 471 cell lines from 2 groups revealed little/no concordant data
Witwer (2013)[30]	Microarray data	93/127 articles were not MIAME compliant
Elliott et al (2006)[31]	Commercial antibodies	Commercial antibodies detect wrong antigens
Prassas et al (2013)[32]	Commercial ELISA	ELISA Kit identified wrong antigen
Stodden et al (2013)[33]	Journals	Computational biology: 105/170 journals noncompliant with National Academies recommendations
Baker et al (2014)[34]	Journals	Top tier fail to comply with agreed standards for animal studies
Vaux (2012)[35]	Journals	Failure to comply with their own statistical guidelines

ALS indicates amyotrophic lateral sclerosis; MIAME, minimum information about a microarray experiment; NGS, next generation sequencing; and VPA, valproic acid (model of autism).

Big brother?

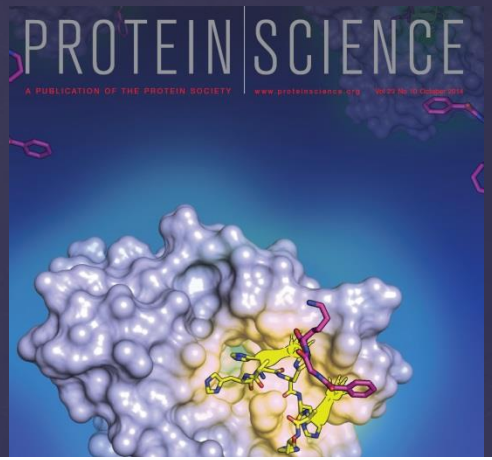
Statistics / Progress in Minor Lab LIMS by researcher

Last week (17 Apr 2015 - 24 Apr 2015)

Person	Clones	Exprs	Purifs	Macro preps	Plates	Drops	Crystals	Datasets processed	Structure refs	Kinetic assays	Thermal shift assays
Cooper, David	0	0	0	0	0	0	23	18	0	0	0
Handing, Katarzyna	0	0	0	0	0	0	51	53	13	0	0
Hou, Jing	0	0	0	0	1	30	0	1	1	0	0
Kowiel, Marcin	0	0	0	0	0	0	1	8	3	0	0
Shabalin, Ivan	0	0	0	0	0	0	125	14	9	0	0
Shumilin, Igor	0	0	0	0	0	0	0	3	2	0	0
Szlachta, Karol	0	0	0	0	0	0	34	20	3	0	0

Last month (25 Mar 2015 - 24 Apr 2015)

Papers:



CrossMark

research papers

IUCrJ

ISSN 2052-2525

BIOLOGY | MEDICINE

Avoidable errors in deposited macromolecular structures: an impediment to efficient data mining

Zbigniew Dauter,[a]* Alexander Wlodawer,[b] Wladek Minor,[c,d,e,f,g] Mariusz Jaskolski[b,i] and Bernhard Rupp[j,k]

Data Management in the Modern Structural Biology and Biomedical Research Environment

thew D. Zimmerman, Marek Grabowski, Marcin J. Domagalski,
abeth M. MacLean, Maksymilian Chruszcz, and Wladek Minor

The future of crystallography in drug discovery

Heping Zheng, Jing Hou, Matthew D Zimmerman, Alexander Wlodawer & Wladek Minor†

† *University of Virginia, Department of Molecular Physiology and Biological Physics, Charlottesville, VA, USA*

PROTOCOL

Validation of metal-binding sites in macromolecular structures with the CheckMyMetal web server

Heping Zheng[1,2], Mahendra D Chordia[1,2], David R Cooper[1,2], Maksymilian Chruszcz[1–3], Peter Müller[4], George M Sheldrick[5] & Wladek Minor[1,2]

the FEBS Journal

REVIEW ARTICLE

Protein crystallography for aspiring crystallographers or how to avoid pitfalls and traps in macromolecular structure determination

Alexander Wlodawer[1], Wladek Minor[2,3,4,5], Zbigniew Dauter[6] and Mariusz Jaskolski[7]

Nucleic Acids Research Advance Access published March 23, 2015

Nucleic Acids Research, 2015 1
doi: 10.1093/nar/gkv225

The Quality and Validation of Structures from Structural Genomics

Marcin J. Domagalski, Heping Zheng, Matthew D. Zimmerman, Zbigniew Dauter, Alexander Wlodawer, and Wladek Minor

Magnesium-binding architectures in RNA crystal structures: validation, binding preferences, classification and motif detection

Heping Zheng[1,2,3,4,†], Ivan G. Shabalin[1,2,3,4,5,†], Katarzyna B. Handing[1,3,4,6], Janusz M. Bujnicki[6,7] and Wladek Minor[1,2,3,4,5,*]

IUCrJ

ISSN 2052-2525

BIOLOGY | MEDICINE

Data to knowledge: how to get meaning from your result

Helen M. Berman,[a]‡ Margaret J. Gabanyi,[a]‡ Colin R. Groom,[b]‡ John E. Johnson,[c]‡ Garib N. Murshudov,[d]‡ Robert A. Nicholls,[d]‡ Vijay Reddy,[c]‡ Torsten Schwede,[e,f]‡ Matthew D. Zimmerman,[g]‡ John Westbrook[a] and Wladek Minor[g]*

High throughput SB

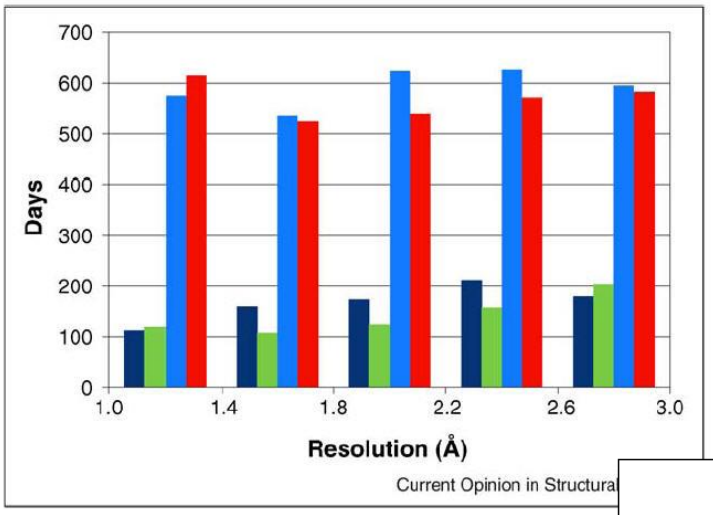
- Automatic cloning
- HT automatic expression
- HT automatic purification
- HT automatic crystallization
- HT automatic data collection
- HT automatic structure solution/refinement

High throughput SB

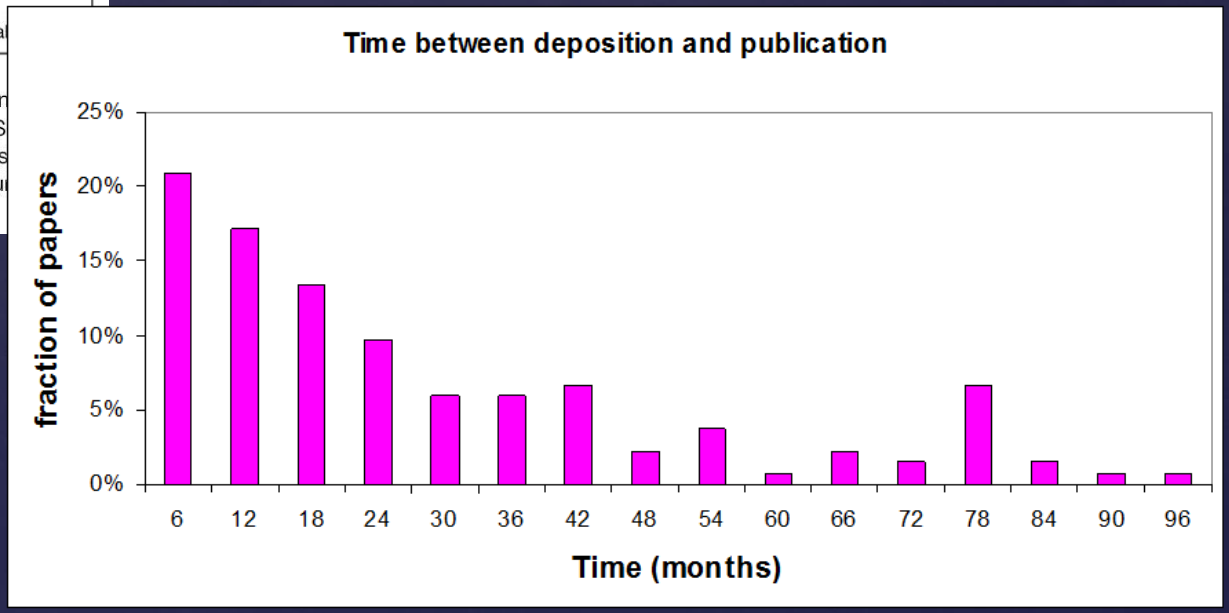
- Automatic cloning
- HT automatic expression
- HT automatic purification
- HT automatic crystallization
- HT automatic data collection
- HT automatic structure solution/refinement

- **Automatic paper writing**

High throughput .. or high output



Average time (in days) between data collection and deposition and non-SG structures. Dark blue and green bars represent S structures, whereas light blue and red bars represent non-SG s deposited in 2000–2004 and 2005–2009, respectively. Structu binned by reported resolution limit (0.4 Å bin width).



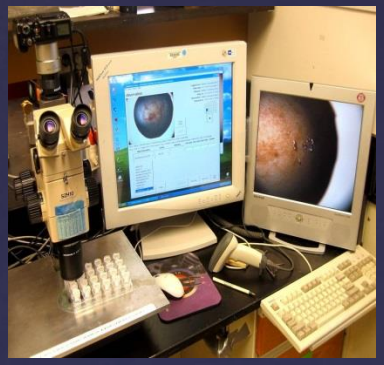
Curr. Opinion in Struct. Biology (2010) **20**: 587-597

HKL-3000 at SBC

Database-controlled pipeline

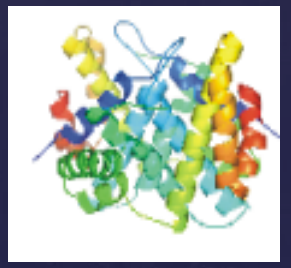
lab e-book

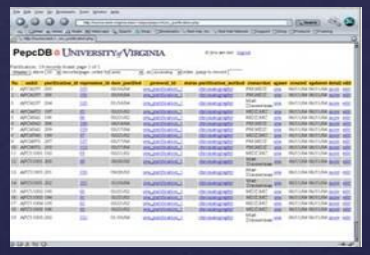
FPLC



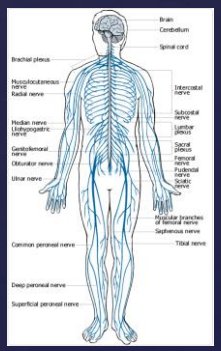
EQUIPMENT

HKL-3000

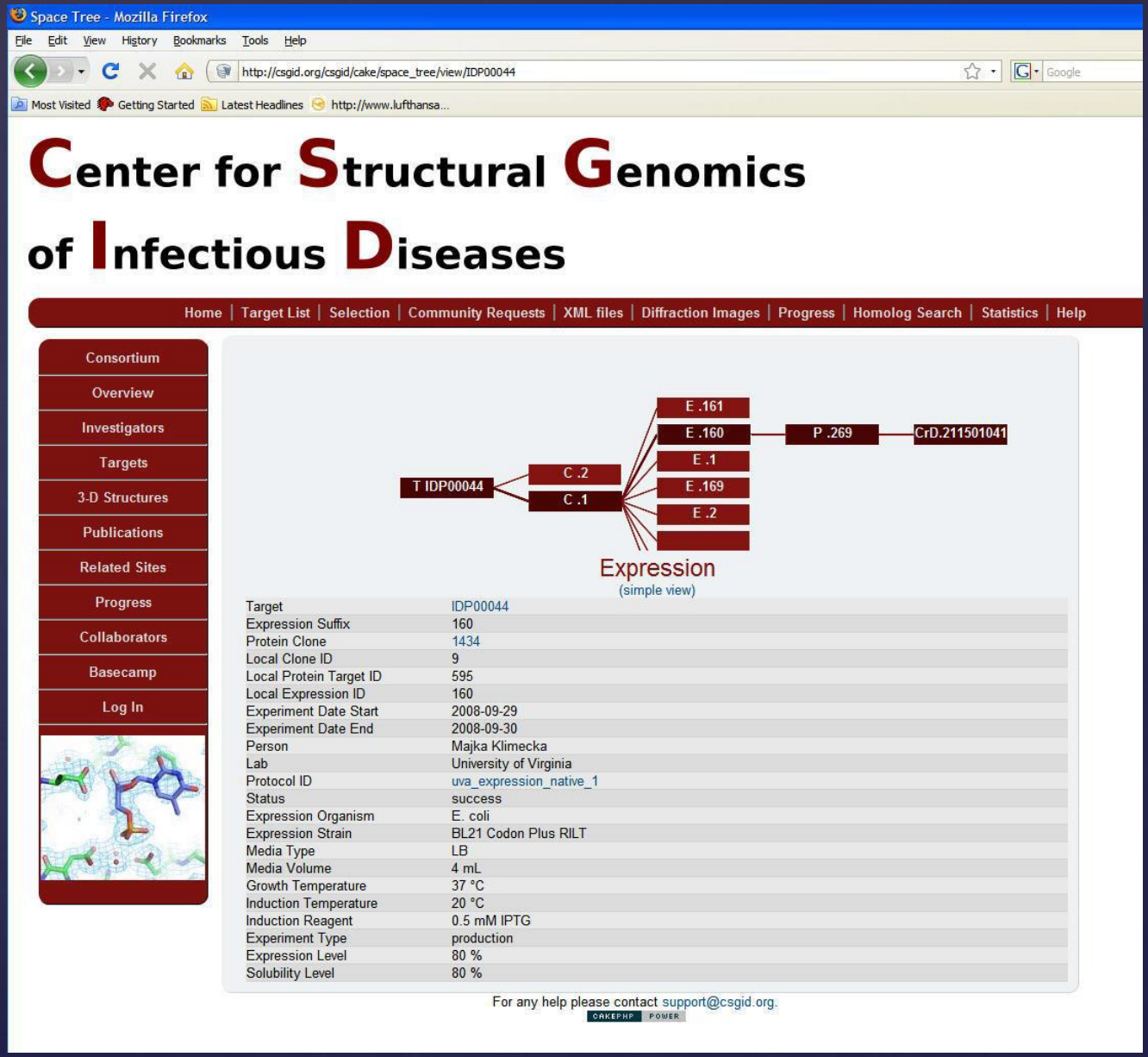




DATABASES



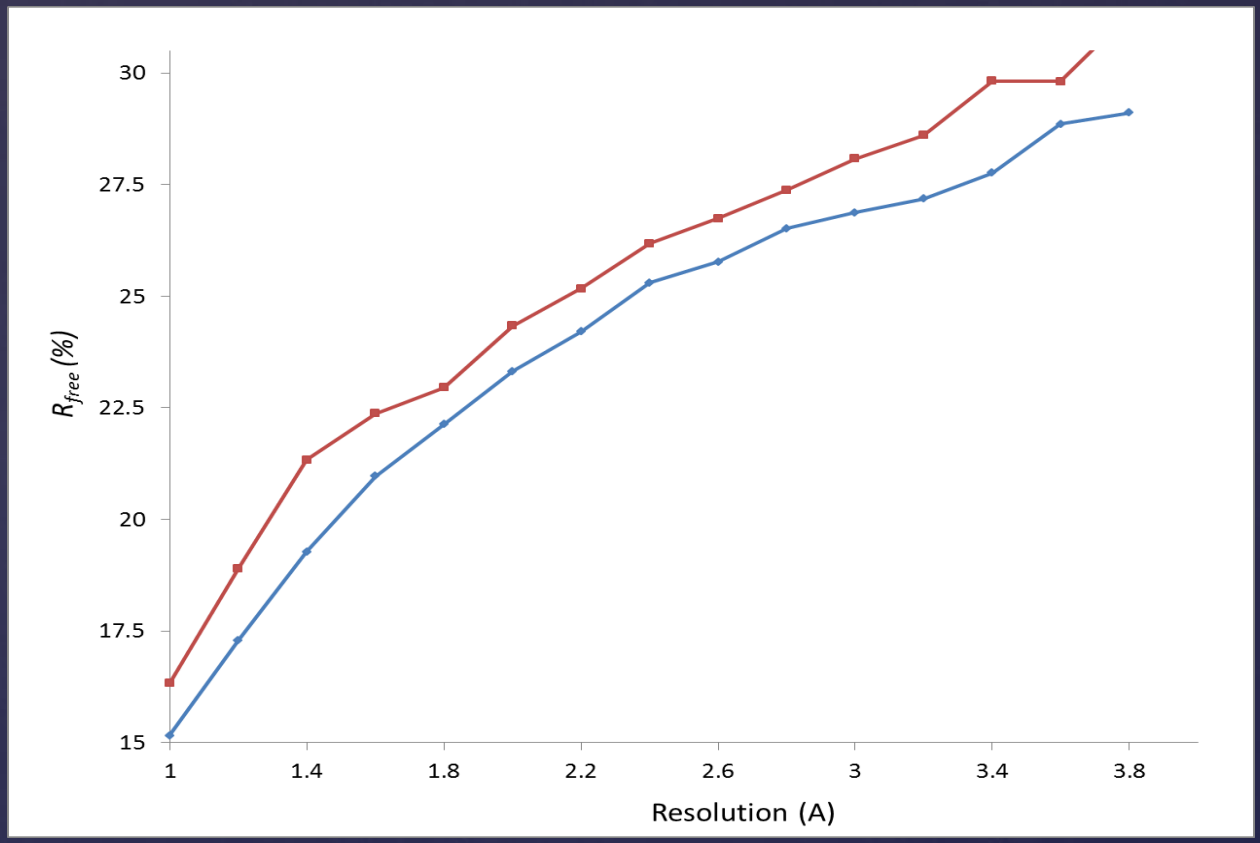
Target status and path to success



What experimenters know about data collection ?

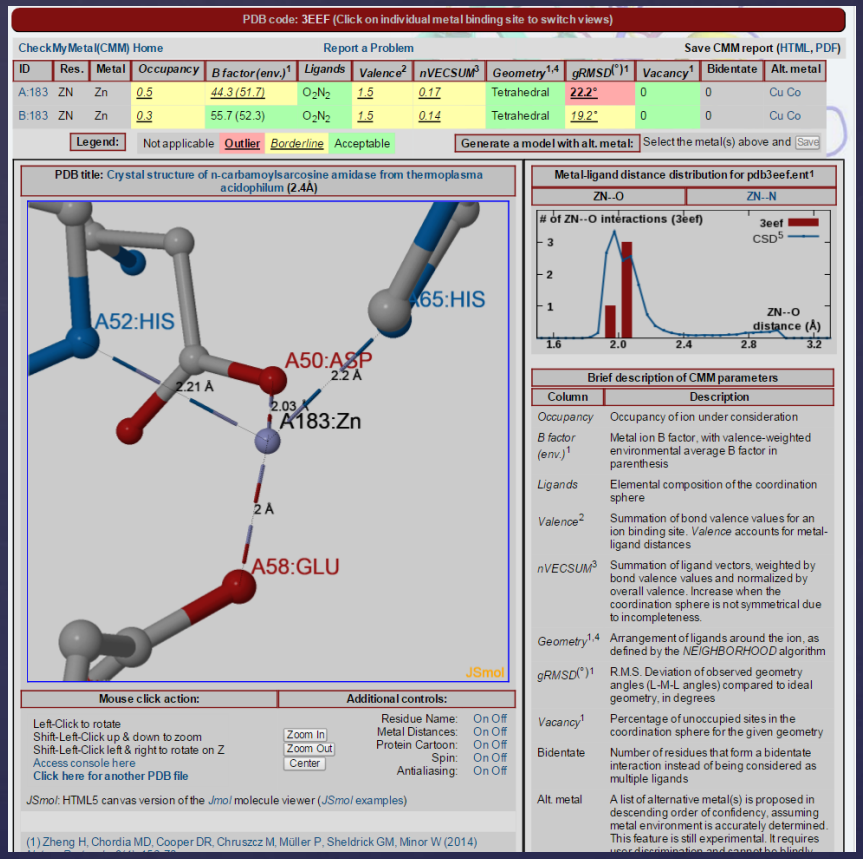
```
REMARK   3  ESTIMATED OVERALL COORDINATE ERROR.
REMARK   3   ESU BASED ON R VALUE                            (A): NULL
REMARK   3   ESU BASED ON FREE R VALUE                       (A): NULL
REMARK   3   ESU BASED ON MAXIMUM LIKELIHOOD                 (A): NULL
REMARK   3   ESU FOR B VALUES BASED ON MAXIMUM LIKELIHOOD (A**2): NULL
REMARK   3
REMARK   3  RMS DEVIATIONS FROM IDEAL VALUES.
REMARK   3   DISTANCE RESTRAINTS.                      RMS    SIGMA
REMARK   3    BOND LENGTH                     (A) : NULL  ; NULL
REMARK   3    ANGLE DISTANCE                  (A) : NULL  ; NULL
REMARK   3    INTRAPLANAR 1-4 DISTANCE        (A) : NULL  ; NULL
REMARK   3    H-BOND OR METAL COORDINATION    (A) : NULL  ; NULL
REMARK   3
REMARK   3   PLANE RESTRAINT                  (A) : NULL  ; NULL
REMARK   3   CHIRAL-CENTER RESTRAINT       (A**3) : NULL  ; NULL
REMARK   3
REMARK   3   NON-BONDED CONTACT RESTRAINTS.
REMARK   3    SINGLE TORSION                  (A) : NULL  ; NULL
REMARK   3    MULTIPLE TORSION                (A) : NULL  ; NULL
REMARK   3    H-BOND (X...Y)                  (A) : NULL  ; NULL
REMARK   3    H-BOND (X-H...Y)                (A) : NULL  ; NULL
REMARK   3
REMARK   3   CONFORMATIONAL TORSION ANGLE RESTRAINTS.
REMARK   3    SPECIFIED                 (DEGREES) : NULL  ; NULL
REMARK   3    PLANAR                    (DEGREES) : NULL  ; NULL
REMARK   3    STAGGERED                 (DEGREES) : NULL  ; NULL
REMARK   3    TRANSVERSE                (DEGREES) : NULL  ; NULL
REMARK   3
REMARK   3  ISOTROPIC THERMAL FACTOR RESTRAINTS.     RMS    SIGMA
REMARK   3   MAIN-CHAIN BOND                (A**2) : NULL  ; NULL
REMARK   3   MAIN-CHAIN ANGLE               (A**2) : NULL  ; NULL
REMARK   3   SIDE-CHAIN BOND                (A**2) : NULL  ; NULL
```

Unexpected correlation?



Average R_{free} by resolution bin (with a width of 0.2 Å for X-ray crystallography PDB structures deposited after January 1, 2001, divided into two groups by the number of missing data items ("NULLs") in the PDB file. The means for "high-completion" deposits (20 NULLs or less) are shown in blue, and the means for "low-completion" deposits (50 or more NULLs) are shown in red.

CheckMyMetal (CMM): Validation of metal-binding sites in macromolecular structures



PROTOCOL

Validation of metal-binding sites in macromolecule structures with the CheckMyMetal web server

Heping Zheng[1,2], Mahendra D Chordia[1,2], David R Cooper[1,2], Maksymilian Chruszcz[1–3], Peter Müller[4], George M Sheldrick[5] & Wladek Minor[1,2]

[1]Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, Virginia, USA. [2]Center for Structural Genomics of Infectious Diseases (CSGID). [3]Department of Chemistry and Biochemistry, University of South Carolina, Columbia, South Carolina, USA. [4]Department of Chemistry and Institute for Soldier Nanotechnologies, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [5]Lehrstuhl für Strukturchemie, Universität Göttingen, Göttingen, Germany. Correspondence should be addressed to W.M. (wladek@iwonka.med.virginia.edu).

Published online XX XX 2013; doi:10.1038/nprot.2013.172

Metals have vital roles in both the mechanism and architecture of biological macromolecules. Yet structures of metal-containing macromolecules in which metals are misidentified and/or suboptimally modeled are abundant in the Protein Data Bank (PDB). This shows the need for a diagnostic tool to identify and correct such modeling problems with metal-binding environments. The CheckMyMetal (CMM) web server (http://csgid.org/csgid/metal_sites/) is a sophisticated, user-friendly web-based method to evaluate metal-binding sites in macromolecular structures using parameters derived from 7,350 metal-binding sites observed in a benchmark data set of 2,304 high-resolution crystal structures. The protocol outlines how the CMM server can be used to detect geometric and other irregularities in the structures of metal-binding sites, as well as how it can alert researchers to potential errors in metal assignment. The protocol also gives practical guidelines for correcting problematic sites by modifying the metal-binding environment and/or redefining metal identity in the PDB file. Several examples where this has led to meaningful results are described in the ANTICIPATED RESULTS section. CMM was designed for a broad audience—biomedical researchers studying metal-containing proteins and nucleic acids—but it is equally well suited for structural biologists validating new structures during modeling or refinement. The CMM server takes the coordinates of a metal-containing macromolecule structure in the PDB format as input and responds within a few seconds for a typical protein structure with 2–5 metal sites and a few hundred amino acids.

INTRODUCTION

Metals are essential in many biological processes. They are present in many macromoleculres, and serve in structural and/or catalytic roles. Structural information about metal-binding environments is often used to understand the molecular mechanism of macromolecule-based or knowledge-assisted structure determination[11]. In addition, the scientific community needs to be aware of the abundance of errors in metal-binding sites in macromolecular structures and of the need to critically assess the quality of each

So far, CMM server has validated 11574 structures from 2669 distinct computer addresses from 42 different countries

NMR results

Structure, Vol. 8, R213–R220, November, 2000, ©2000 Elsevier Science Ltd. All rights reserved. PII S0969-2126(00)00524-4

Does NMR Mean "Not for Molecular Replacement"? Using NMR-Based Search Models to Solve Protein Crystal Structures

Ways & Means

Yu Wai Chen,*§ Eleanor J. Dodson,†
and Gerard J. Kleywegt‡
*Centre for Protein Engineering and Cambridge
University Chemical Laboratory
MRC Centre

difficult MR problem [6]. It was clear that this method could be extended to using a whole NMR ensemble as an MR search model. In 1995, Müller et al. [7] confirmed that use of an ensemble of models led to better results in MR than the use of single models. Simultaneously, and independently, the structure of bovine acyl-

LETTER

doi:10.1038/nature09964

Improved molecular replacement by density- and energy-guided protein structure optimization

Frank DiMaio[1], Thomas C. Terwilliger[2], Randy J. Read[3], Alexander Wlodawer[4], Gustav Oberdorfer[5], Ulrike Wagner[5], Eugene Valkov[6], Assaf Alon[7], Deborah Fass[7], Herbert L. Axelrod[8], Debanu Das[8], Sergey M. Vorobiev[9], Hideo Iwai[10], P. Raj Pokkuluri[11] & David Baker[1]

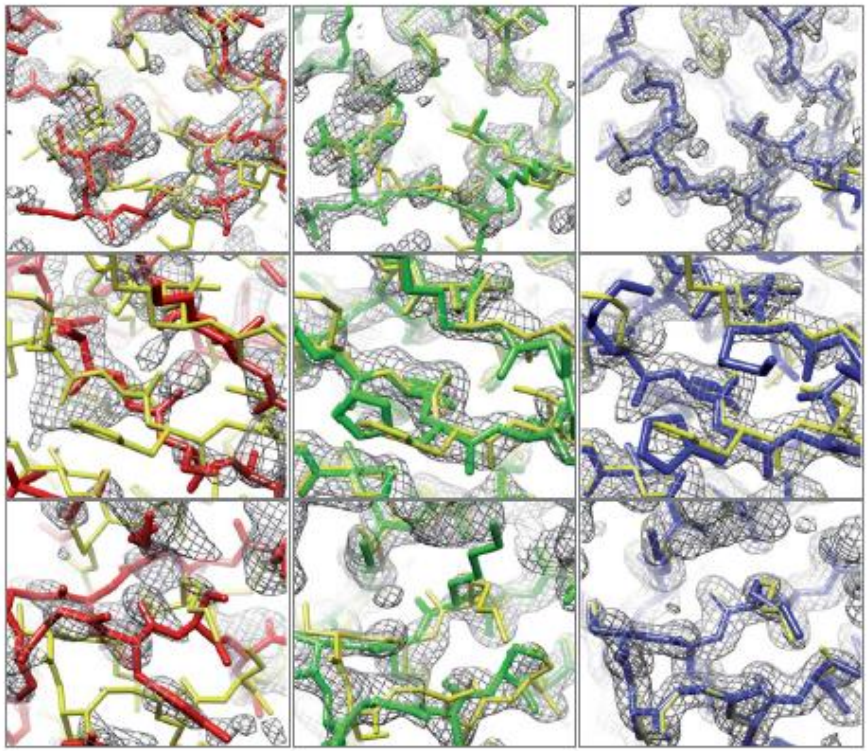


Figure 1 | Examples of improvement in electron density and model quality. Each row corresponds to one of the entries in Table 1. First row: 6 (2.0 Å resolution); second row: 7 (2.1 Å resolution); third row: 12 (1.7 Å resolution). Left column: correct initial molecular replacement solution (not necessarily identifiable at this stage) using starting model and corresponding density. Middle column: energy-optimized model and corresponding density. Right column: model and density following automatic building using the energy-optimized model as the source of phase information. The final deposited structure is shown in yellow in each panel; the initial model, energy-optimized model, and model after chain rebuilding are in red, green and blue, respectively. The sigma-A-weighted $2mF_o - DF_c$ density contoured at 1.5σ is shown in grey.

Acknowledgments

Wladek Minor
- Matt Zimmerman
- Marek Grabowski
- Heping Zheng
- Marcin Cymborowski
- Karol Langner (Google)
- Przemek Porebski
- Piotr Sroka
- Ivan Shabalin
- Katherine Handing
- Ewa Niedzialkowska

Zbyszek Otwinowski
- Dominika Borek

Andrzej Joachimiak
MCSG and SBC staff

Wayne Anderson and CSGID staff

Steve Almo and NYSGRC Staff

Ian Wilson, Marc Elslinger and JCSG staff

Steven Burley, John Westbrook and PDB staff

Tom Terwilliger
Zbyszek Dauter

Grants:
U01-HG008424
NIH GM53163, GM62414, GM74942
GM093342, GM094585, GM094662
NIAID HHSN272200700058C
NIAID HHSN272201200026C
HKL Research. Inc.

Acknowledgments

Wladek Minor
- Matt Zimmerman
- Marek Grabowski
- Heping Zheng
- Marcin Cymborowski
- Karol Langner (Google)
- Przemek Porebski
- Piotr Sroka
- Ivan Shabalin
- Katherine Handing
- Ewa Niedzialkowska
- Piotr Sroka

Zbyszek Otwinowski
- Dominika Borek

Andrzej Joachimiak
MCSG and SBC staff

Wayne Anderson and CSGID staff

Steve Almo and NYSGRC Staff

Ian Wilson, Marc Elslinger and JCSG staff

Steven Burley, John Westbrook and PDB staff

Tom Terwilliger
Zbyszek Dauter

Grants:
U01-HG008424
NIH GM53163, GM62414, GM74942
GM093342, GM094585, GM094662
DOE, NCI
NIAID HHSN272200700058C
NIAID HHSN272201200026C
HKL Research. Inc.