

# A subject specific repository for MX (proteindiffraction.org)



Wladek Minor, Marcin Cymborowski, David Cooper Department of Molecular Physiology and Biological Physics University of Virginia

*IUCr workshop: Raw diffraction data reuse: the good, the bad, and the challenging Melbourne, August 2023* 



WM notes that he has been involved in the development of state-of-the-art software, data management and mining tools; some of them were commercialized by HKL Research and are mentioned in this presentation. WM is the cofounder of HKL Research and a member of the board. The author(s) have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

## Reproducibility (2014-15)

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.



#### Many landmark findings in preclinical oncology research are not reproducible, in part because of inadequate cell lines and animal models.

## NIH plans to enhance reproducibility

Francis S. Collins and Lawrence A. Tabak discuss initiatives that the US National Institutes of Health is exploring to restore the self correcting fatore of for US preclinical research.

# Raise standards for preclinical cancer research



CORRESPONDENCE

Believe it or not: how much can we rely on published data on potential drug targets?

Florian Prinz, Thomas Schlange and Khusru Asadullah

# Raising the bar

**EDITORIAL** 

umbers. Lots and lots of numbers. It is hard to find a paper published in *Science* or any other journal that is not full of numbers. Interpretation of those numbers provides the basis for the conclusions, as well as an assessment of the con-

# Reproducibility (2022)

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, keg it's backed up by do academic findings. T address it head on.



Many landmark findings in preclinical oncology research are not reproducible, in part because of inadequate cell li

## Raise standards fc preclinical cancer rese



#### News in focus



The US National Institutes of Health is located in Bethesda, Maryland.

## NIH ISSUES A SEISMIC Mandate: Share Data Publicly

Policy could set a standard for research, scientists say, but they have questions about logistics and equity.

Believe it or not: how much can we rely on published data on potential drug targets?

Florian Prinz, Thomas Schlange and Khusru Asadullah

software or tools needed to analyse the data, when and where the raw data will be published and any special considerations for accessing or distributing those data.

Such a seismic shift in practice has left some researchers worried about the amount of work that the mandate will require when it becomes effective.

Jenna Guthmiller, an immunologist at the University of Chicago in Illinois, can attest that more work will probably berequired. She is one of a handful of researchers funded through a US National Institute of Allergy and Infectious Diseases programme that has enacted a policy similar to the NIH-wide plan, she says. For Guthmiller, that meant tracking downinformation on long-gone reagents and experimental conditions for a project that's been running for four years. That took 15 hours, she says, "and I was fortunate enough to work with a data manager".

Because the vast majority of laboratories and institutions don't have data managers who organize and curate data, the policy although well-intentioned — will probably put a heavy burden on trainees and early-career principal investigators, says Lynda Coughian, a vaccinologist at the University of Maryland School of Medicine in Baltimore, who has been leading a research team for fewer than two years and is worried about what the policy will mean for her.

Jorgenson says that, although the policy might require researchers to spend extra time

> umbers. Lots and lots of numbers. It is hard to find a paper published in *Science* or any other journal that is not full of numbers. Interpretation of those numbers provides the basis for the conclusions, as well as an assessment of the con-

## **FAIR** Journals



## Data dump vs resource



## Repositories for raw data: IRRMC, SBGrid, EMPIAR

## Data dump vs resource

- 1. Long-term availability of data
- **II.** Persistent identifiers are assigned for all datasets
- III. The status of datasets is trackable via persistent identifiers
- IV. All datasets are accessible via persistent identifiers
- v. Restricted access datasets are discoverable via persistent ids
- vi. Bidirectional links exist between datasets and the scientific publications that use them
- vii. Archive is searchable by a wide variety of criteria
- viii. Datasets are validated

Gus, J.M. & McMahon, B. (2014). Acta Crystallogr D Biol Crystallogr 70, 2520-2532

## https://proteindiffraction.org

NIH) National Institutes of Health Crystallography Mithows at We want we want we want we want water water

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 labroatories.

Currently indexed projects: 6206

Currently indexed datasets: 9648

Data downloaded from IRRMC may be freely used under the Creative Commons license CC0 (Public Comain Dadication Waiver), IRRMC strongly urges users who download data to credit the source data by using the DOI in any publications and/or derived data that make use of the downloaded data.





~29,000 visits ~21,000 sessions



# Harvesting metadata

Metadata source	Metadata parameters
User	Identity of the user, 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)
Diffraction images	Detector type and serial numbers, and image format, Data collection parameters: number of frames, oscillation step size, experimental orientation angles (e.g. $\kappa$ , $\phi$ , $\omega$ , and 2 $\theta$ ), detector distance
Structure factors/scaling logs	Integrated reflection data, Nominal resolution cutoff Completeness, overall and highest resolution shell (HRS), R <sub>merge</sub> , R <sub>meas</sub> , R <sub>pim</sub> Redundancy, overall and HRS, Mean I/sigma I, overall and HRS Software used to process diffraction images
Automatic reprocessing	Validation of provided/extracted metadata, spacegroup,, merging statistics Estimation of radiation damage, crystal internal non-isomorphicity Presence/Strength of anomalous signal Diffraction image artifacts and other "features" (background scattering, ice rings, diffuse scattering, etc.)
Molecular models	Data from PDB 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 extraction metadata/annotations from image headers, PDB, SG databases ...

Diffraction project datasets 021011\_4wed





Method: SAD Resolution: 2.35 Å Space group: P 31

Lownload all images (1.6 GB)

PDB website for 4WED

🗹 doi:10.18430/M3KG69

#### Project details

Title	Crystal structure of ABC transporter substrate-binding protein from Sinorhizobium meliloti
Authors	Shabalin, I.G., Handing, K.B., Minor, W.
R / R <sub>free</sub>	0.16 / 0.23
Unit cell edges [Å]	57.33 x 57.33 x 132.06
Unit cell angles [°]	90.0, 90.0, 120.0

#### Dataset VA7-1\_s.### details



Number of frames	200 (1 - 200)
Distance [mm]	320.0
Oscillation width [°]	1.00
Phi [°]	-90.0
Wavelength [nm]	1.07819
Equipment	21-ID-D at APS (Advanced Photon Source)

# Insights from the metadata: data collection strategy



Grabowski et al Acta Cryst. (2016). D72 1181–1193

# IRRMC as a training set for machine learning



#### Czyzewski et al (2021), Expert Syst. Appl. 174: 114740

# IRRMC hosts data used for workshops...



## https://covid-19.bioreproducibility.org

SARS-CoV-2 related structures v 2023.04.12

Structures Other resources Funding Citing and contact

#### Validated SARS-CoV-2 related structural models of potential drug targets

Rational drug design against emerging threats depends on well-established current methodology worked out by structural biology to provide accurate structure models of the macromolecular drug targets. In the current COVID-19 crisis, the structural biological community has responded at once, presenting in rapid succession structure models of CoV-2 proteins and depositing them in the Protein Data Bank (PDB), without time embargo and before publication. Since the structures from the first-line research are produced in an accelerated mode, there is an elevated chance of mistakes and errors. Here, we provide a source of carefully validated PDB models of CoV-2 proteins, with the aim of helping the biomedical community to establish a validated database.



CREATED BY SCIENTISTS FROM



#### About the corrections - encouragement for redeposition

In the course of this project we have corrected over 100 PDB models. We strongly encourage the authors of the original deposits to make constructive use of our corrections and to update their models in the PDB via the versioning mechanism, which allows depositors to update their entries while retaining the same PDB code. We recommend either taking the updated models on our website as starting points, or using the list of corrections in "Re-refinement summary" for each structure. As these models were not always fully finalized for deposition, all corrections should be carefully inspected, new PDB validation reports should be generated, and any remaining issues should be addressed. To filter out only corrected structures from the list, you can use the With corrections "prevent in the filters below. If there are any questions regarding particular corrections, please e-mail us.

#### Structures

Our database currently contains information about 2942 SARS-CoV-2 protein structures and 206 additional structures of other coronaviruses. Use the filters below to select rows with attributes of interest. Next to each filter value, the number of shown/total structures from the group is displayed. Multiple values can be selected across multiple panes. To select more than one value within a single search pane, press and hold Ctrl or Shift when selecting filters. Text search can be performed using the search box on the left below the filters.

Filters Ac	tive - 2													Clear All	
Method		, a	Aaţ #ţ	Virus	x م	AA: #: Prot	ein	,0 x A41	#1 L	igand category		,o x Aa‡#ĭ	Presets	× A	A] #]
Cryo-EM		(0/1	160	Bat-CoV-RaTG13		Enve	Норе	97	F	unctional ligan	1	551/780	Non-PanDDA structures	1292/2	486)
NEUTRON	DIFFRACT	ION/X-RAY/HYBRI	_	Bat-CoV-SHC014		M NSP	1	10/20	N	lo functional lig	ands	203/486	Structures with RNA	C	161
NMD				Bat-CoV-WIV1		NSP	2	3/5	P	athogen-host ir	nteraction	120/251	With corrections	(104/	110
X-ray		(1292/1	954	BtSCoV-Rf1.2004		NSP	3: Macro	(45)(457)	• P	rotein-protein c	omplex	109/293			
Search:		(match	ned 1,292 o	ut of 3,148 total reco	ds)								Сору	Excel CSV PC	)F Pri
PDB 😄	Resol. 🗢	Released - Titl	e 🌼 Meth	od    Ligand IDs	• Virus	Protein	Ligand category	R-work ©	R-free #	R-merge 😄	Metals 0	P <sub>Q1</sub> (PDB)   Iss	ues 🌣 🎁 🗘 Re-refined?	Raw data	Ref.
▼ 7WQA	1.80 Å	2023-03-22	X-ray	1	SARS-CoV-2	NSP5 (3CLpro	)	17.12%	20.76%	4.10%	÷		No		3
<b>7</b> 7XAX	2.25 Å	2023-03-22	X-ray	3WL	SARS-CoV-2	NSP5 (3CLpro	)	22.89%	27.97%	7.10%			No		÷
▼ 7XB3	2.08 Â	2023-03-22	Х-гау	2	SARS-CoV-2	NSP5 (3CLpro	)	21.36%	24.25%	3.10%	2		No	19. I	÷.
▼ 7XB4	2.07 Å	2023-03-22	X-ray	4WI	SARS-CoV-2	NSP5 (3CLpro	)	21.09%	24.03%	3.70%	<b>7</b> 2	5.8	No		







A., Wlodawer *et al.*, (2020) *FEBS J.*, **287**, 3703–3718
I., Shabalin *et al.*, (2020). *IUCrJ.*, **7**, 1048–1058.
D. Brzezinski *et al.*, (2021) *Protein Sci.*, 30, 115–124
M. Grabowski *et al.*, (2021) *IUCrJ*, **8**, 395–407
M. Kowiel *et al.*, (2019) *Bioinformatics*, **35**, 452–461

## Re- refinement

Data collection					
	7BRR	<b>Re-refinement</b>			
Resolution (Å)	50.00 - 1.40 (1.45 - 1.40)	56.49 - 1.35			
Wavelength (Å)	0.979	0.97919			
Space group	P21	P21			
a, b, c (Å)	55.45, 99.02, 59.58	55.45, 99.02, 59.58			
α, β, γ (°)	90, 108.54, 90	90, 108.54, 90			
Completeness (%)	100	98.7 (86.8)			
Reflections used	120024	131802			
<i> / <sigma i=""></sigma></i>	43.2	36.5 (1.0)			
Redundancy	6.8	6.4 (4.1)			
Rmerge		0.048 (1.090)			
Rpim		0.020 (0.553)			
CC1/2 last shell	0.71	0.51			
Wilson B factor (Å <sup>2</sup> )	14.5	14.5			

#### Refinement Rwork / Rfree 0.184 / 0.197 0.117/0.157 Resolution (Å) 29.02 - 1.40 49.56 - 1.35 Reflections all 108002 108915 5649, 4.9% Reflections for Rfree 5504, 5.1% Bond lengths rmsd (Å) 0.011 0.009 Bond angles rmsd (°) 1.18 1.56 Mean B value (Å<sup>2</sup>) 20 24 Number of protein atoms 4713 4878 Mean B value for protein atoms (Å2) 19 22 Number of water atoms (expected) 667 (829) 684 (829) 39 Mean B value for water atoms (Å2) 31 Number of ligands/ions atoms 58 74 Mean B value for ligands/ions atoms (Å2) 21 25 Clashscore 4.643.95 Clashscore percentile (100) 52.6 62.5 Rotamer outliers (<1%) 0.19 2.15 0.00 Ramachandran outliers (<0.2%) 0.00 Ramachandran favored (>98%) 98.99 98.68 Residues with bad bonds (<0%) 0.08 0.00 Residues with bad angles (<0.1%) 1.33 0.74MolProbity score 1.241.43





### D. Brzezinski et al., (2021) Protein Sci., 30, 115–124

## Example of use

## research papers



ISSN 2059-7983

Received 15 June 2020 Accepted 1 February 2021

Edited by M. Schiltz, Fonds National de la Recherche, Luxembourg

**Keywords:** protein crystallography; ice; stacking disorder; structure-factor error.

## Ice in biomolecular cryocrystallography

#### David W. Moreau, Hakan Atakisi and Robert E. Thorne\*

Physics Department, Cornell University, Ithaca, NY 14853, USA. \*Correspondence e-mail: ret6@cornell.edu

Diffraction data acquired from cryocooled protein crystals often include diffraction from ice. Analysis of ice diffraction from crystals of three proteins shows that the ice formed within solvent cavities during rapid cooling is comprised of a stacking-disordered mixture of hexagonal and cubic planes, with the cubic plane fraction increasing with increasing cryoprotectant concentration and increasing cooling rate. Building on the work of Thorn and coworkers [Thorn *et al.* (2017), *Acta Cryst.* D73, 729–727], a revised metric is defined for detecting ice from deposited protein structure-factor data, and this metric is validated using full-frame diffraction data from the Integrated Resource for Reproducibility in Macromolecular Crystallography. Using this revised metric

Servers in Minor lab

CheckMyMetal

CheckMyBlob

Molstack

Fitmunk

covid-19.bioreproducibility.org

Integrated Resource for Reproducibility Macromolecular Crystallography

**Protein Purification and Crystallization Artefacts** 

CMB

CMM





## **Tower of Babel** we need to understand each other



Pieter Bruegel the Elder

# A Prototype of Advanced Information System





Zhang, H. et al. (2021) IUCrJ 8:1-12

## The Washington Post

## Democracy Dies in Darkness

## Messy, incomplete U.S. data hobbles pandemic response

The nation's decentralized, underfunded reporting system hampers efforts to combat the coronavirus.



The U.S. Capitol dome seen reflected in the window of a medical vehicle. (Jabin Botsford/The Washington Post)

By Joel Achenbach and Yasmeen Abutaleb September 30, 2021 at 9:30 a.m. EDT

The contentious and confusing debate in recent weeks over <u>coronavirus booster shots</u> has exposed a fundamental weakness in the United States' ability to respond to a public health crisis: The data is a mess.

MOST READ HEALTH >



- As covid persists, nurses are leaving staff jobs and tripling their salaries as travelers
- 2 Over half of young adults are obese or overweight, study says



3 U.S. coronavirus cases approach 50 million as New York City imposes new vaccine mandate



#### 'A largely 19th-century system'

The CDC compiles national statistics by collecting data from every state and locality, but these jurisdictions often have different ways of counting tests, infections and even deaths. The data may not be submitted to the CDC for days or weeks. Many smaller jurisdictions still share that data via fax, an outdated technology.

5 How scary is omicron? Scientists are racing to find answers.



Joel Achenbach, Yasmeen Abutaleb Washington Post, Sept 30, 2021

## The Washington Post

### Democracy Dies in Darkness

## Messy, incomplete U.S. data hobbles pandemic response

The nation's decentralized, underfunded reporting system hampers efforts to combat the coronavirus.



The U.S. Capitol dome seen reflected in the window of a medical vehicle. (Jabin Botsford/The Washington Post)

By Joel Achenbach and Yasmeen Abutaleb September 30, 2021 at 9:30 a.m. EDT

The contentious and confusing debate in recent weeks over <u>coronavirus booster shots</u> has exposed a fundamental weakness in the United States' ability to respond to a public health crisis: The data is a mess.

MOST READ HEALTH >



- 1 As covid persists, nurses are leaving staff jobs and tripling their salaries as travelers
- 2 Over half of young adults are obese or overweight, study says



3 U.S. coronavirus cases approach 50 million as New York City imposes new vaccine mandate



'A largely 19th-century system'

The CDC compiles national statistics by collecting data from every state and locality, but these jurisdictions often have different ways of counting tests, infections and even deaths. The data may not be submitted to the CDC for days or weeks. Many smaller jurisdictions still share that data via fax, an outdated technology.

5 How scary is omicron? Scientists are racing to find answers.



Joel Achenbach, Yasmeen Abutaleb Washington Post, Sept 30, 2021

Wladek Minor lab: Marcin Cymborowski David R. Cooper Przemek Porebski Heping Zheng

Acknowledgments



Grants: NIH HG008424 GM132595 NIAID HHSN272201200026C Depositors of data JCSG CSGID SSGCID MCSG SGC NYSGRC Many individual crystallographers..



## Marek Grabowski