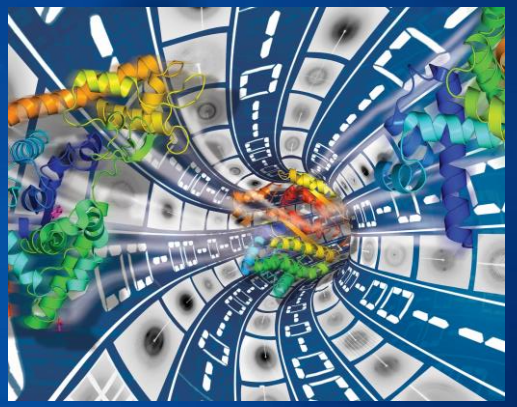
A subject specific repository for MX (proteindiffraction.org)



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

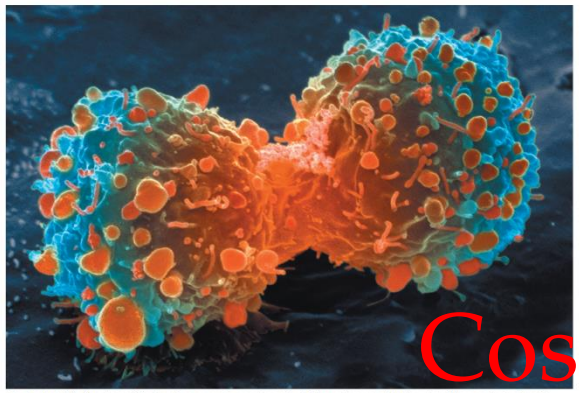
IUCr workshop: Raw diffraction data reuse: the good, the bad, and the challenging
Melbourne, August 2023

Disclosure

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 co-founder 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.



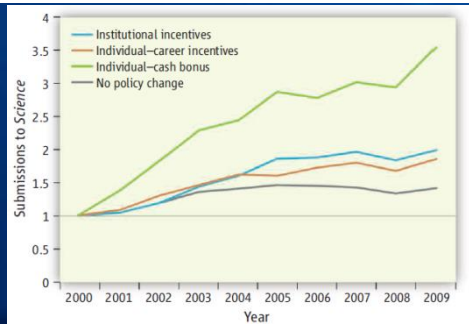
Many landmark findings in preclinical oncology research are not reproducible, in part because of inadequate cell lines and animal models.

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 nature of preclinical research.

Cost: $28 billion per year for US

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

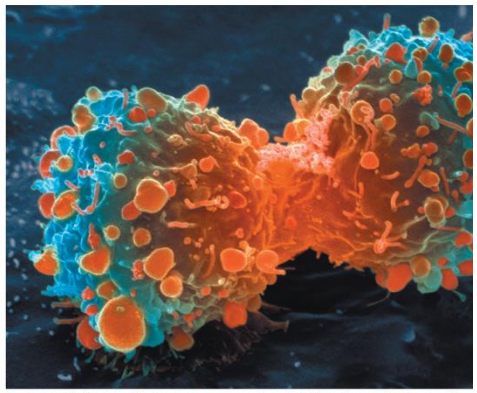
EDITORIAL

Raising the bar

Numbers. 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-

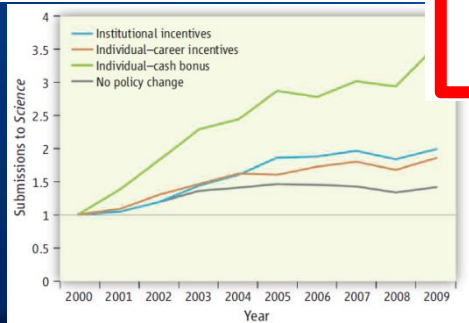
Reproducibility (2022)

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, ke
it's backed up by dc
academic findings. T
address it head on.



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

Raise standards fc
preclinical cancer rese



News in focus

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

GRANGER/BOTHER/ALAMY

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.

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 be required. 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 down information 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 Coughlan, 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

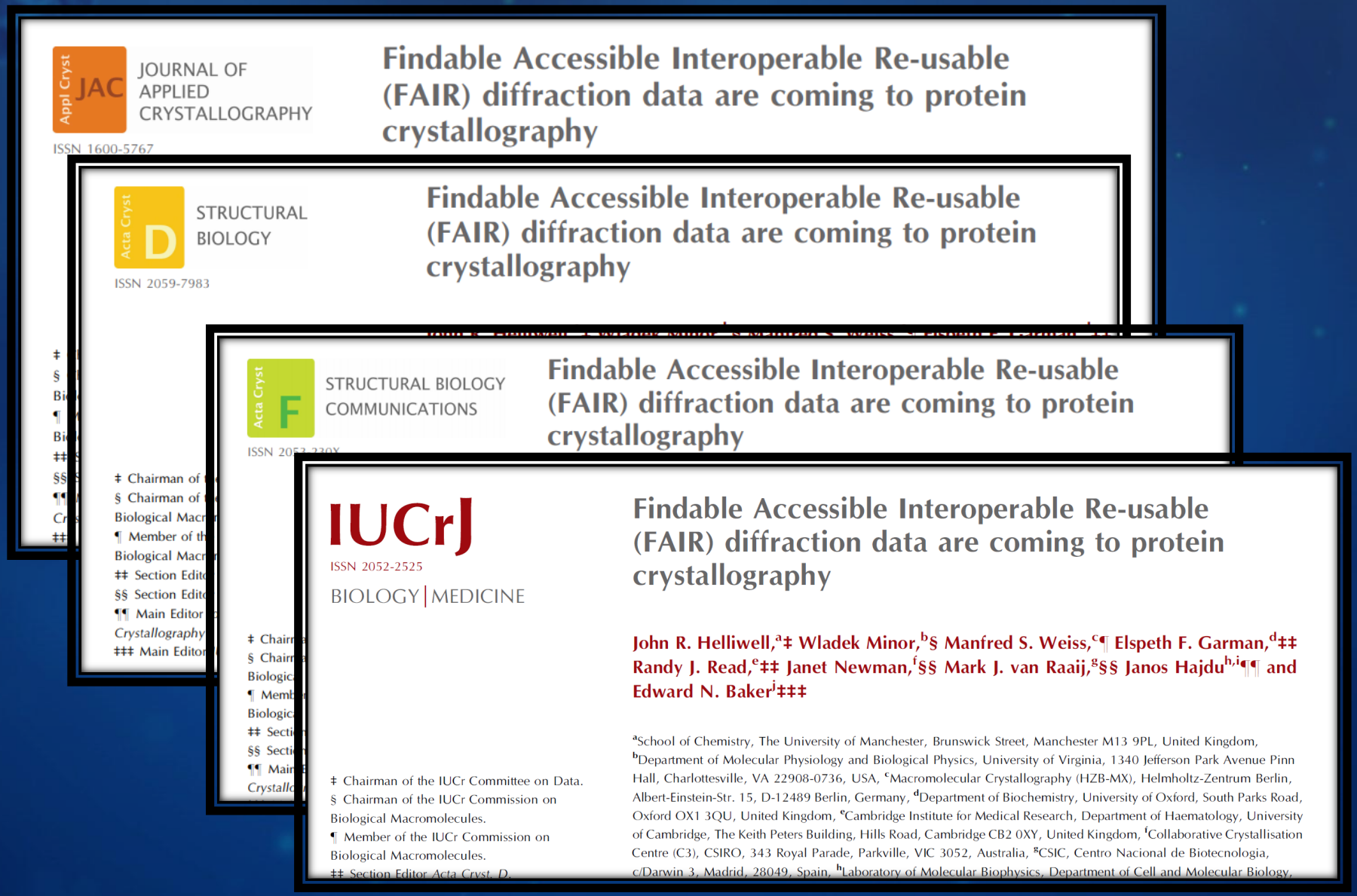
Believe it or not: how much can we rely on published data on potential drug targets?

Florian Prinz, Thomas Schlange and Khusru Asadullah

Raising the bar

Numbers. 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

I. 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

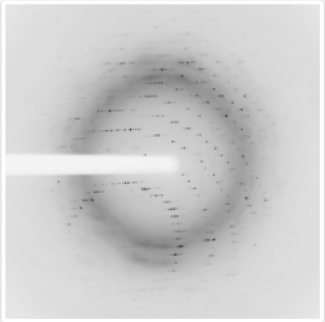
~29,000 visits
~21,000 sessions

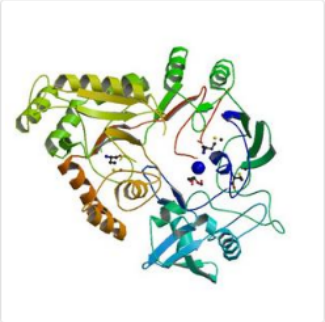
Harvesting metadata

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. κ,φ,ω, and 2θ), detector distance
Structure factors/scaling logs	Integrated reflection data, Nominal resolution cutoff Completeness, overall and highest resolution shell (HRS), R_{merge}, R_{meas}, R_{pim} 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

Download all images (1.6 GB)

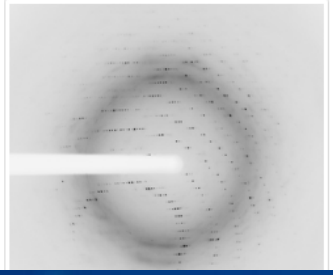
PDB website for 4WED

doi:10.18430/M3KG69

Project details

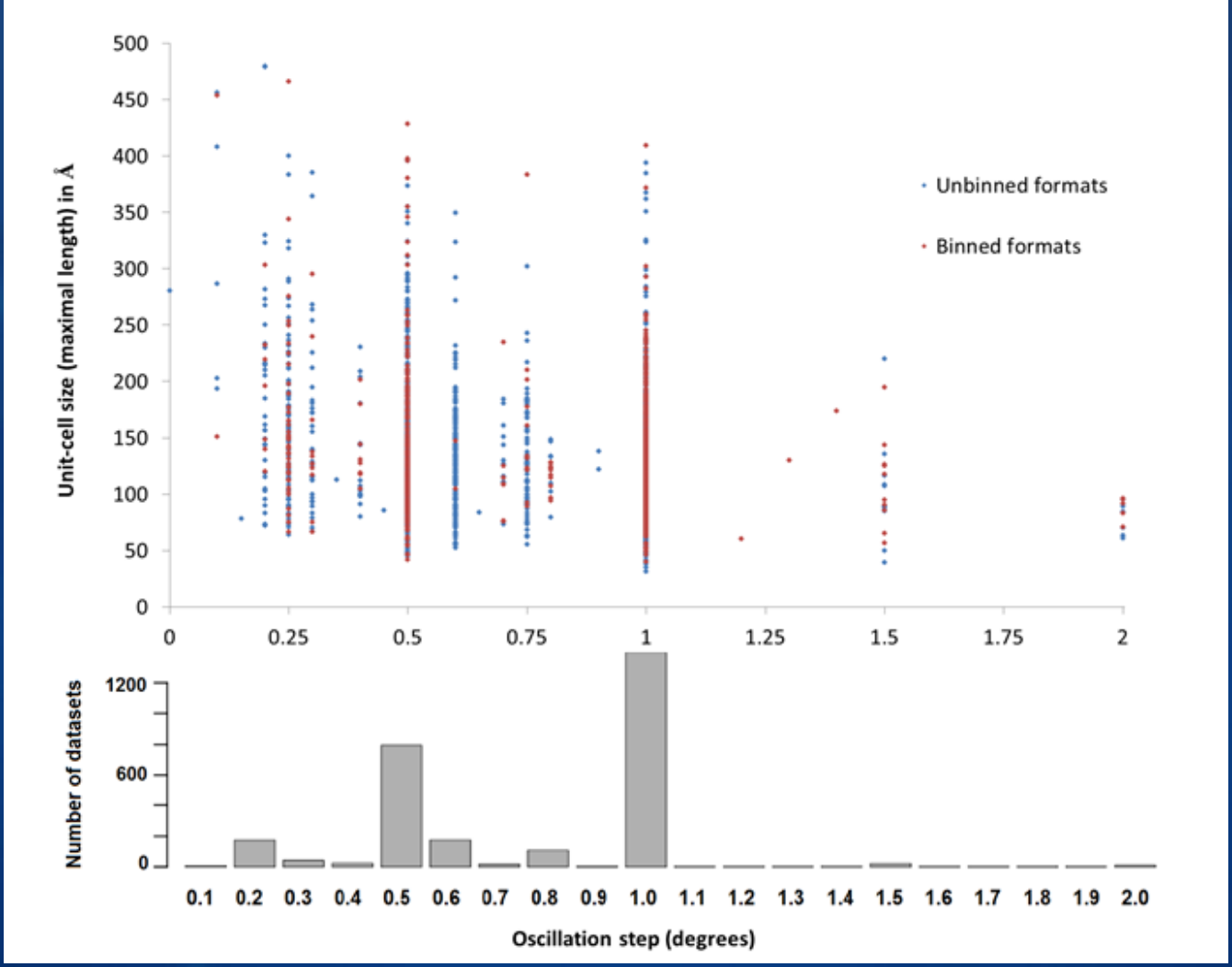
Title	Crystal structure of ABC transporter substrate-binding protein from Sinorhizobium meliloti
Authors	Shabalin, I.G., Handing, K.B., Minor, W.
R / R_{free}	0.16 / 0.23
Unit cell edges [Å]	57.33 x 57.33 x 132.06
Unit cell angles [°]	90.0, 90.0, 120.0

Dataset VA7-1_s.### details



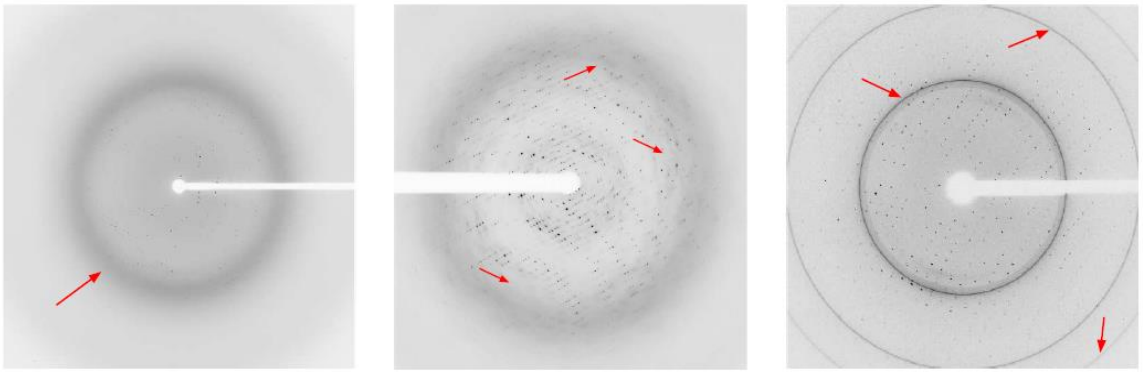
Number of frames	200 (1 - 200)
Distance [mm]	320.0
Oscillation width [°]	1.00
Phi [°]	-90.0
Wavelength [nm]	1.07819
Equipment	21-ID-D at APS (Advanced Photon Source)

Insights from the metadata: data collection strategy



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

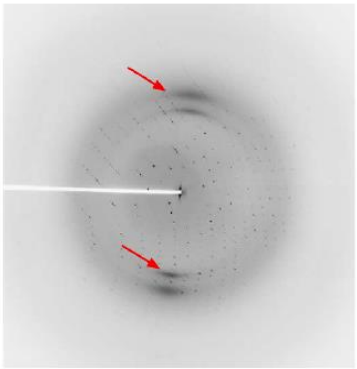
IRRMC as a training set for machine learning



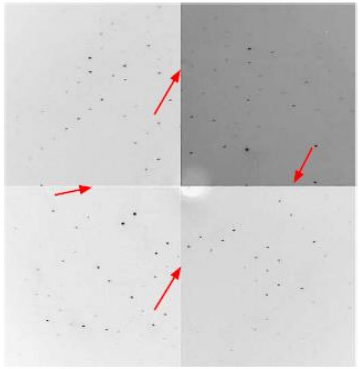
(a) Background Ring

(b) Diffuse Scattering

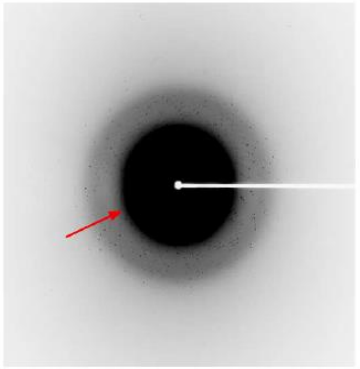
(c) Ice Ring



(d) Loop Scattering



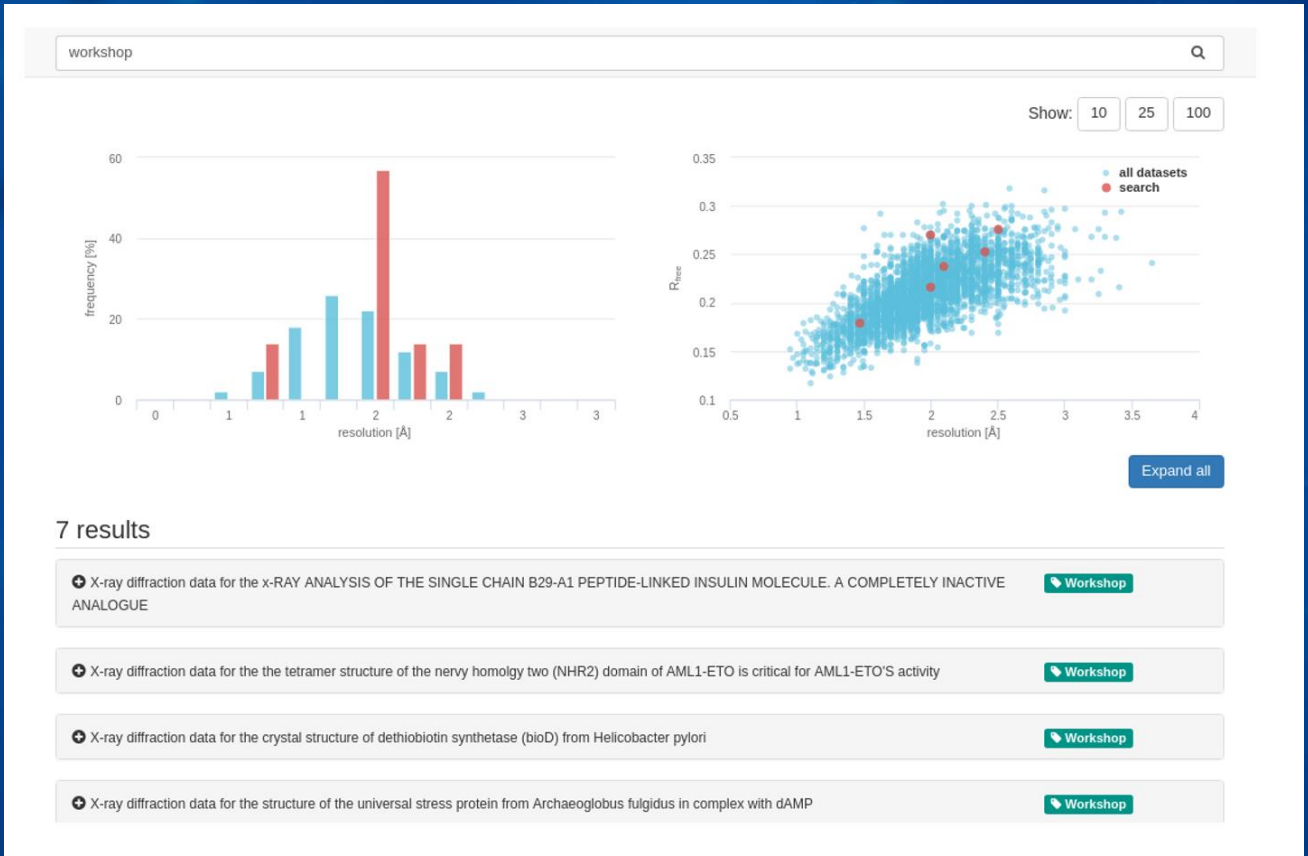
(e) Non-uniform Detector



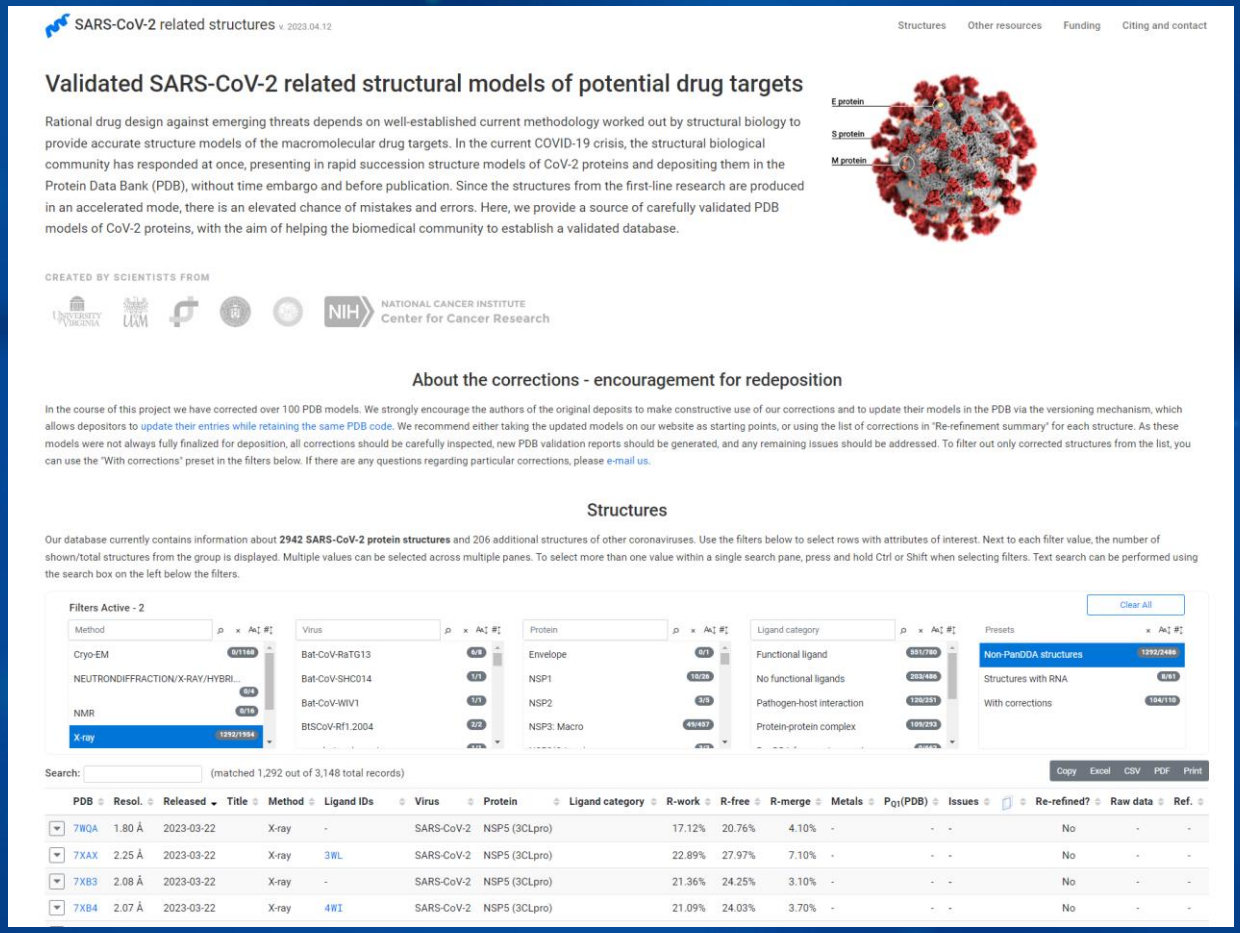
(f) Strong Background

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

IRRMC hosts data used for workshops…



https://covid-19.bioreproducibility.org



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	Re-refinement
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>	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 (Å²)	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
Reflections for Rfree	5504, 5.1%	5649, 4.9%
Bond lengths rmsd (Å)	0.011	0.009
Bond angles rmsd (°)	1.18	1.56
Mean B value (Å²)	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)
Mean B value for water atoms (Å²)	31	39
Number of ligands/ions atoms	58	74
Mean B value for ligands/ions atoms (Å²)	21	25
Clashscore	4.64	3.95
Clashscore percentile (100)	52.6	62.5
Rotamer outliers (<1%)	0.19	2.15
Ramachandran outliers (<0.2%)	0.00	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	0.74
MolProbity score	1.24	1.43

FIGURE 3 Report showing the re-processing and re-refinement of PDB structure 7BRR. This structure will be re-deposited to the PDB under a new PDB ID, due to significant changes in comparison with the original deposit. The second part of the report is presented in Figure S1

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

Example of use

research papers

Acta Cryst D STRUCTURAL BIOLOGY

ISSN 2059-7983

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

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.

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.* D**73**, 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

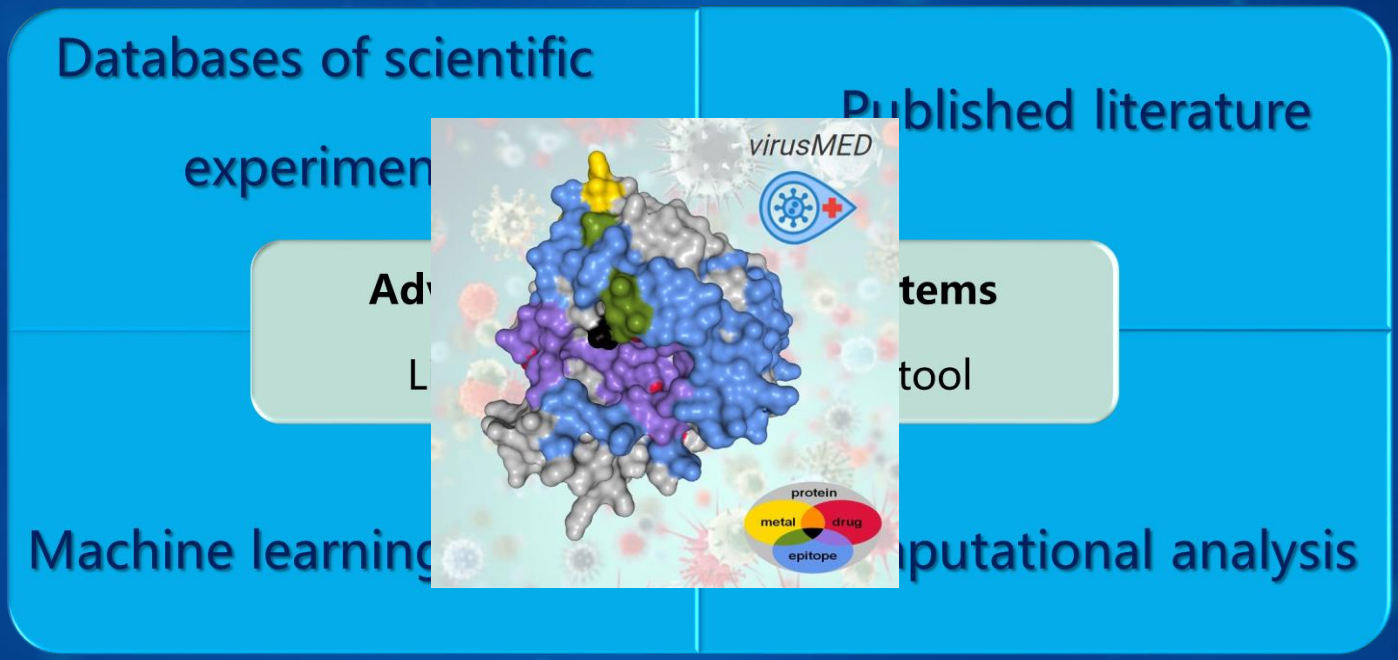
CheckMyMetal	CMM
CheckMyBlob	CMB
Molstack	
Fitmunk	
covid-19.bioreproducibility.org	
Integrated Resource for Reproducibility Macromolecular Crystallography	IRRMC
Protein Purification and Crystallization Artefacts	

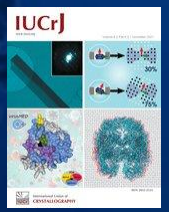
Tower of Babel

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

Health

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 coronavirus booster shots has exposed a fundamental weakness in the United States' ability to respond to a public health crisis: The data is a mess.

'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.

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

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

Health

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 coronavirus booster shots has exposed a fundamental weakness in the United States' ability to respond to a public health crisis: The data is a mess.

'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.

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

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

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

Acknowledgments

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

Grants:
NIH HG008424
GM132595
NIAID HHSN272201200026C

Depositors of data
JCSG
CSGID
SSGCID
MCSG
SGC
NYSGRC
Many individual crystallographers…

Marek Grabowski