Correcting the public record of biological crystallography

Mariusz Jaskólski

Center for Biocrystallographic Research
Institute of Bioorganic Chemistry, Polish Academy of Sciences
Department of Crystallography, Faculty of Chemistry, A Mickiewicz University, Poznań

ECM-32 2019 • Vienna • 30 min • Workshop

Historical background

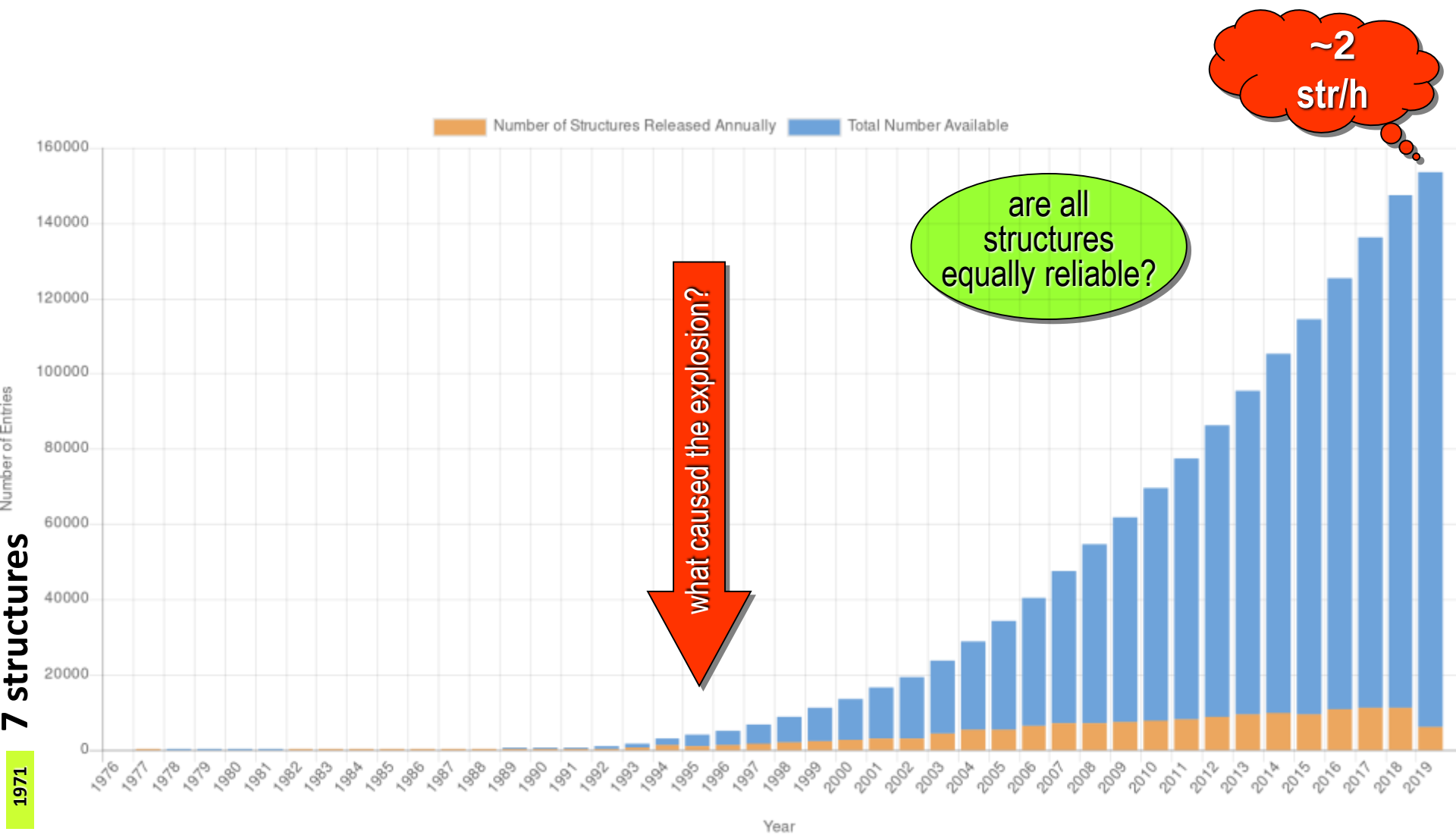
Macromolecular crystallography has been with us for 60+ years.

It has accumulated an enormous volume of structural biological information, key for the understanding of life and advancement of medicine.

It formed the gold standard in structural biology, and its results are viewed as almost error free.

Was that time and success story sufficient to learn how to do everything properly and avoid errors, temptations and traps?

Growth of the PDB



Valid concerns exist about invalid or irreproducible reserach

Why Most Published Research Findings Are False

Ioannidis JPA (2005) *PLoS Medicine* 2(8), 696-701.

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R+1)$. The probability of a study finding a true relationship

- Koehler JJ (1993) **The Influence of Prior Beliefs on Scientific Judgments of Evidence Quality**. *Org. Behavior Human Decision Proc.* **56**, 28-55.
- Frey BS (2003) **Publishing as Prostitution? Choosing Between One's Own Ideas and Academic Failure**. *Public Choice* **116**, 205-223.
- Simmons JP, Nelson LD and Simonsohn U (2011) **False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant**. *Psychological Science* **22**, 1359-1366.

Biomolecular structure models

"While certain discovered trends are very positive (e.g. clashscore markedly decreases with the year of structure publication), others are alarming (e.g. **ligand quality** stagnates with the year of structure publication)."

Macromolecular crystallography is a useful model science…

Crystallography is both data-rich (even millions of accurate experimental observations) and knowledge-rich (huge database of prior structures).
Ideal situation for Bayesian (1702-1761) analysis of Posterior Model Likelihood:

$$prob(M|D) \propto prob(D|M) \times prob(M)$$

Model Likelihood $\propto$ **Quality of Evidence** x **Prior probability**

There has to be a **balance** between the terms:
strong claim with little prior basis needs strong evidence !

However, the models of macromolecules are enormously huge, with hundreds of thousands of parameters, often outnumbering the observations

after: B. Rupp

Error types

- scientific fraud/fabricated data (**very rare**), e.g. complement proteins (Murthy case)
- totally wrong model (**rare**), e.g. ABC transporters, RuBisCO subunit
- wrong connections between secondary structure elements
- register error - sequence shift
- wrong residue assignment
- wrong side chain conformation
- wrong **metal**/water assignment
- unjustified solvent modeling
- **fictitious modeling of map noise ("ligands") at very low contour level**

mis-/over-interpretation of the data
R_{free} should be able to detect, but not necessarily pinpoint, this

Error sources

- **paucity of data** (reflections) - model "overinterprets" available data
- bad data quality
- cognitive bias = wishful thinking
- negligence of experimenter, lack of proper training
- lack of proper supervision

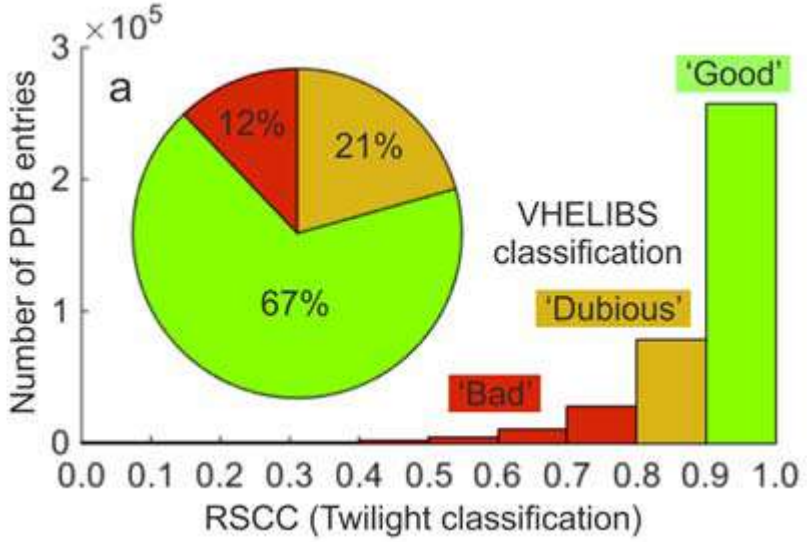
PDB data mining consistently shows:

1. Most ligand models have resonably good quality/electron density fit
2. Some interpretations qualify as generously optimistic
3. **Some are blatantly wrong**

source: B. Rupp

Table 1 Electron density-based validation of protein–ligand models

Scores			Classification	
RSCC	% of structures	Predicted number of PDB	Twilight Rupp et al.	VHELIBS Pujadas et al.
1.0–0.9	67	~46,900		'Good'
<0.9–0.8	21	~14,700		'Dubious'
<0.8–0.7	7	~4,900		'Bad'
<0.7–0.6	3	~2,100		
<0.6–0.5	1	~700		
<0.5	1	~700		



The FEBS Journal

Volume 282 Number 22 November 2015 | ISSN 1742-464X www.febsjournal.org

Structural biology: are a few rotten apples spoiling the barrel?

B Rupp, A Wlodawer, W Minor, JR Helliwell, M Jaskolski (2016) Correcting the record of structural publications requires joint effort of the community and journal editors. FEBS J 283, 4452-4457

FEBS PRESS science publishing by scientists

WILEY

12 years to investigate fraud…

H.M. Krishna Murthy, Ph.D., University of Alabama at Birmingham: Based on evidence and findings of an investigation conducted by the University of Alabama at Birmingham (UAB), the Office of Research Integrity's (ORI's) review of UAB's investigation, and additional evidence obtained and analysis conducted by ORI in its oversight review of UAB's investigation, ORI found that Dr. H.M. Krishna Murthy (Respondent), former Research Associate Professor, Department of Vision Sciences, UAB, **committed research misconduct** in research supported by PHS grants, specifically NIAID, NIH, grants R01 AI051615, R01 AI032078, and R01 AI045623; NHLBI, NIH, grants P01 HL034343 and R01 HL064272; and NIDDK, NIH, grant R01 DK046900.

Falsified and/or fabricated research was reported in:

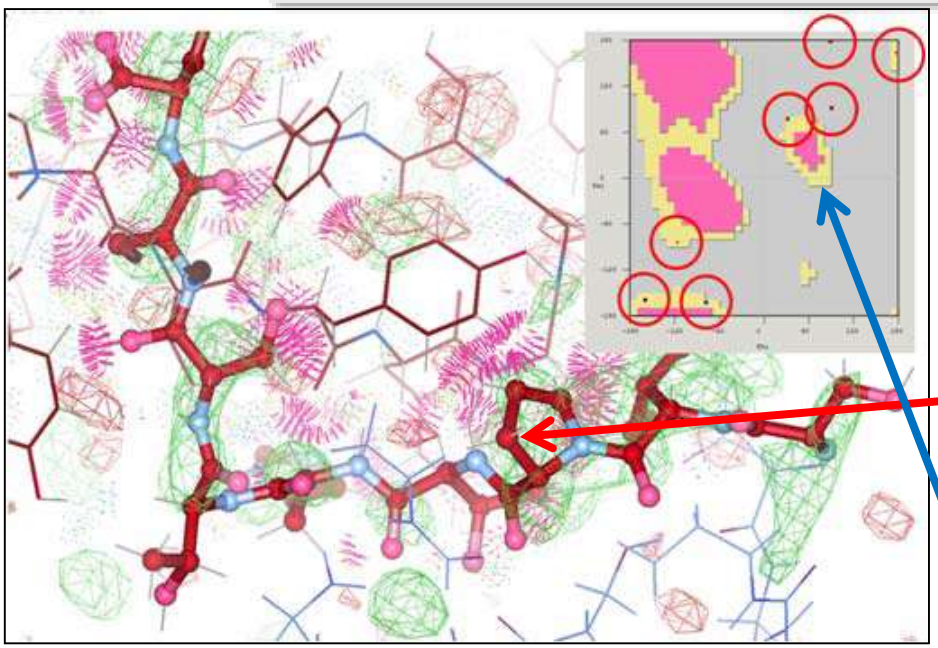
Nature 444:221-225, **2006**; retracted in: Nature 532:268, **2016**
JBC 274:5573-5580, **1999**; retracted in: J. Biol. Chem. 284:34468, **2009**
PNAS 101:8924-8929, **2004**; Editorial Expression of Concern in: PNAS 107:6551, **2010**
Biochem. 44:10757-10765, **2005**
PNAS 103:2126-2131, **2006**; Editorial Expression of Concern in: PNAS 107:6551, **2010**
Acta Cryst. D55:1971-1977, **1999**; retracted in: Acta Cryst. D66:222, **2010**
JMB 301:759-767, **2000**; retracted in: J. Mol. Biol. 397:1119, **2010**
Cell 104:301-311, **2001**
Biochem. 41:11681-11691, **2002**
PDB deposits 2HR0, 1BEF, 1RID, 1Y8E, 2A01, 1CMW, 2QID, 1DF9, 1G40, 1G44, 2OU1, 1L6L

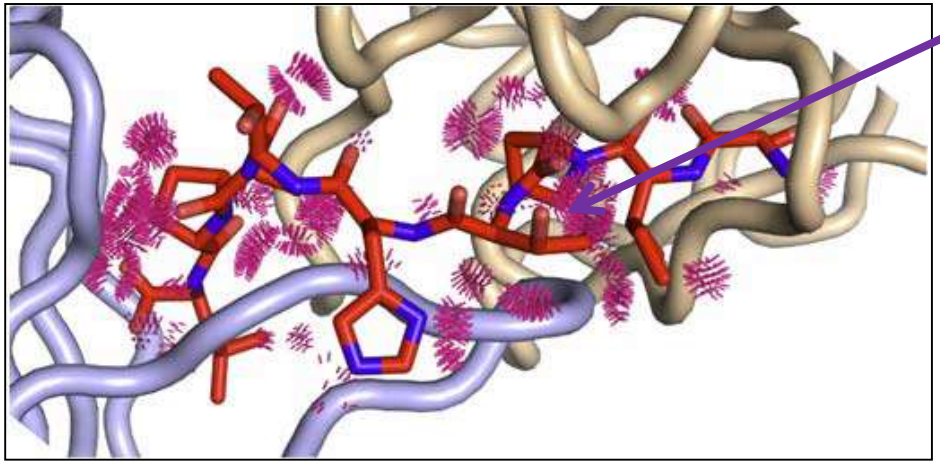
Falsified and/or fabricated research results also were referenced in the following PHS grant applications:

1 R21 AI056224-01 submitted to NIAID, NIH
1 R01 AI064509-01 submitted to NIAID, NIH
1 R01 AI64509-01A1 submitted to NIAID, NIH
1 R01 AI051615-01A1 submitted to NIAID, NIH
1 R03 TW006840-01 submitted to Fogarty International Center (FIC), NIH

Office of Research Integrity, April 4, 2018

Claim vs evidence and prior expectations





Claim: a dodecapeptide KLASIPTHTSPL bound to Fab 36-65 '**provides mechanistic insights into the generation of antibody diversity**' (Salunke et al. *Immunity* 2006)

(1) Evidence: absent: parts of Fab CDR loop modeled as peptide

(2) Prior expectations I: high energy backbone conformation **implausible**

(3) Prior expectations II: 69 severe steric clashes of 67 atoms, 26 clashes within peptide. 87 clashes when CDR H138-H140 properly built. Physically impossible

Posterior model likelihood = zero.
What can be done? Request retraction?

Salunke's response: 1. The burden of **proof of the absence** is on the critic; 2. Relativism: scientists have the right to **alternative interpretation** of experimental observations (electron density); 3. **Others have done it** before;

after: B. Rupp

Solution: redeposit correct Fab-only model

	Original deposit (2a6i)	*PDB_REDO* Calculated	*PDB_REDO* Conservative	*PDB_REDO* Optimised	Manually rebuilt (5vga)
R	0.245	0.244	0.246	0.242	0.203
R_{free}	0.264	0.267	0.285	0.287	0.250
Clashscore/ Percentile	36/26th		1.8/100th	2.6/100th	0.3/100th
Ramachandran outliers	22/5.1%		7/1.6%	8/1.8%	0
Poor rotamers	31/8.1%		20/5.2%	17/4.4%	4/1.0%

Unsupervised automated refinement cannot (yet?) correct such models

Manual intervention and rebuilding is necessary and can be successfully done

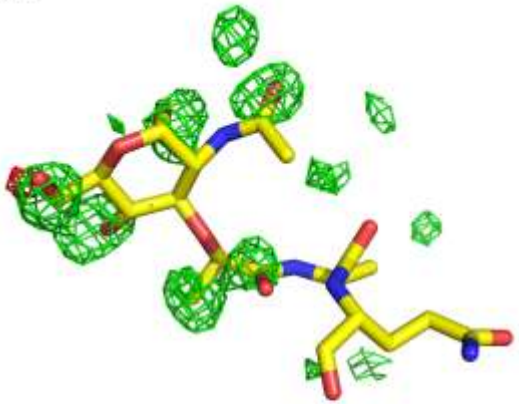
The corrected model has been deposited

source: B. Rupp

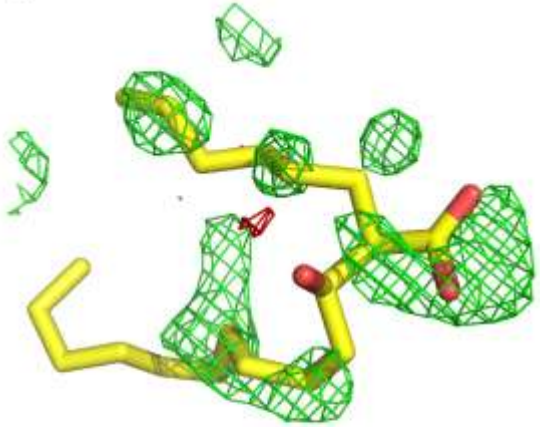
Ligands from fantasyland "found" in ribosome inactivating protein

mFo-DFc
omit maps
±3σ green/red

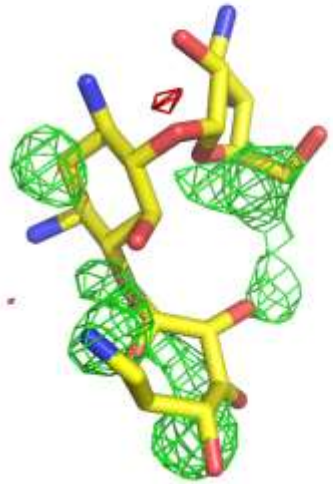
A peptidoglycan 4LWX



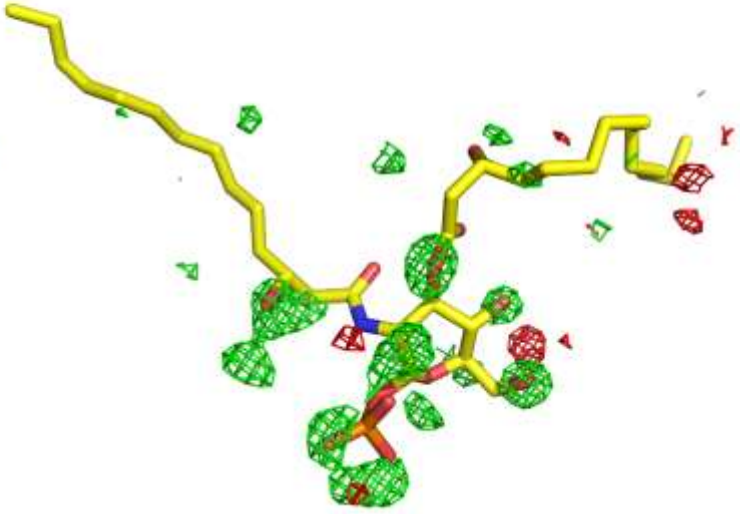
B mycolic acid 3U8F



C kanamycin 3U6T



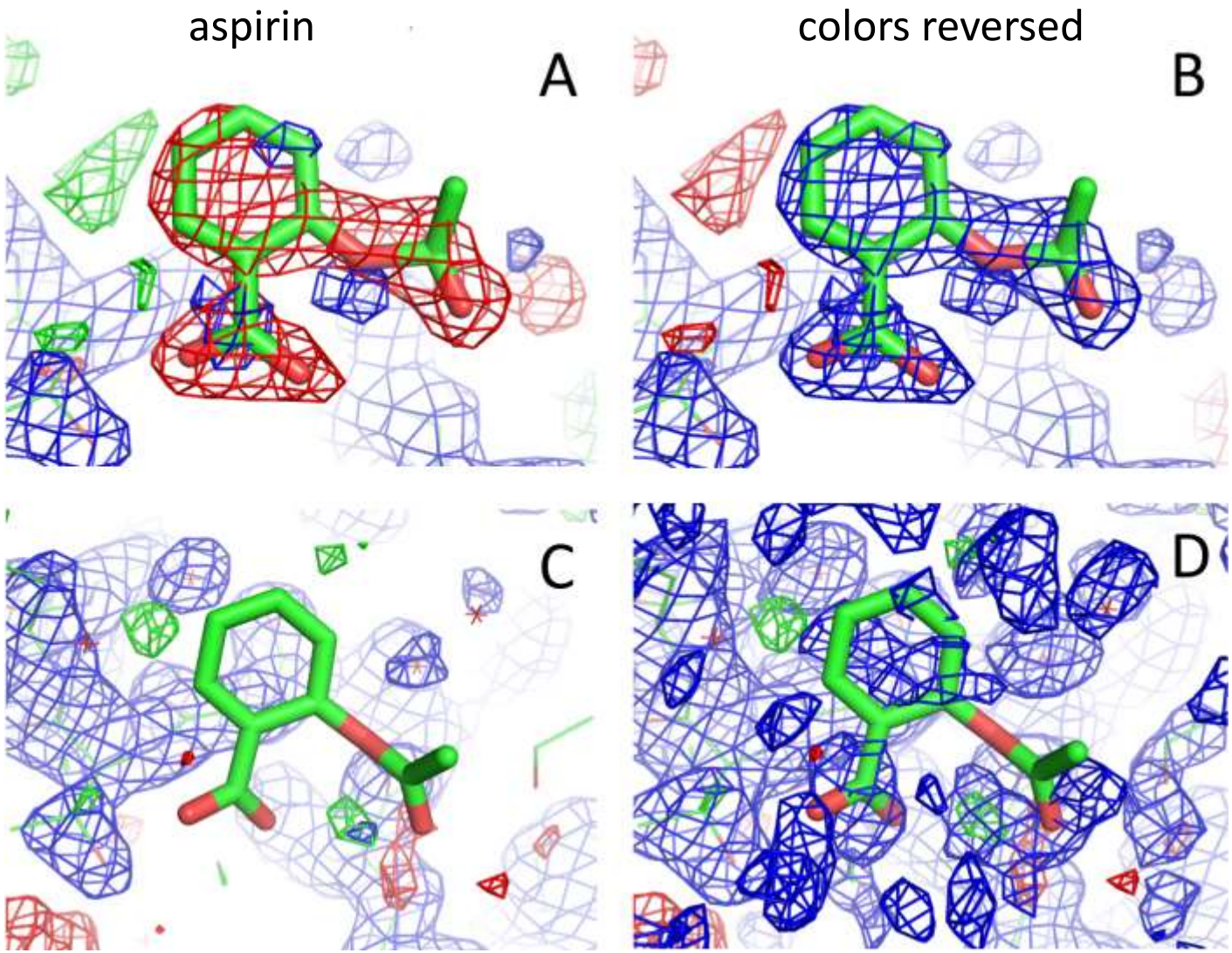
D lipopolyscaccharide 4GUW



Structures deposited, but not published

Aspirin may give you a headache if…B-factors not refined

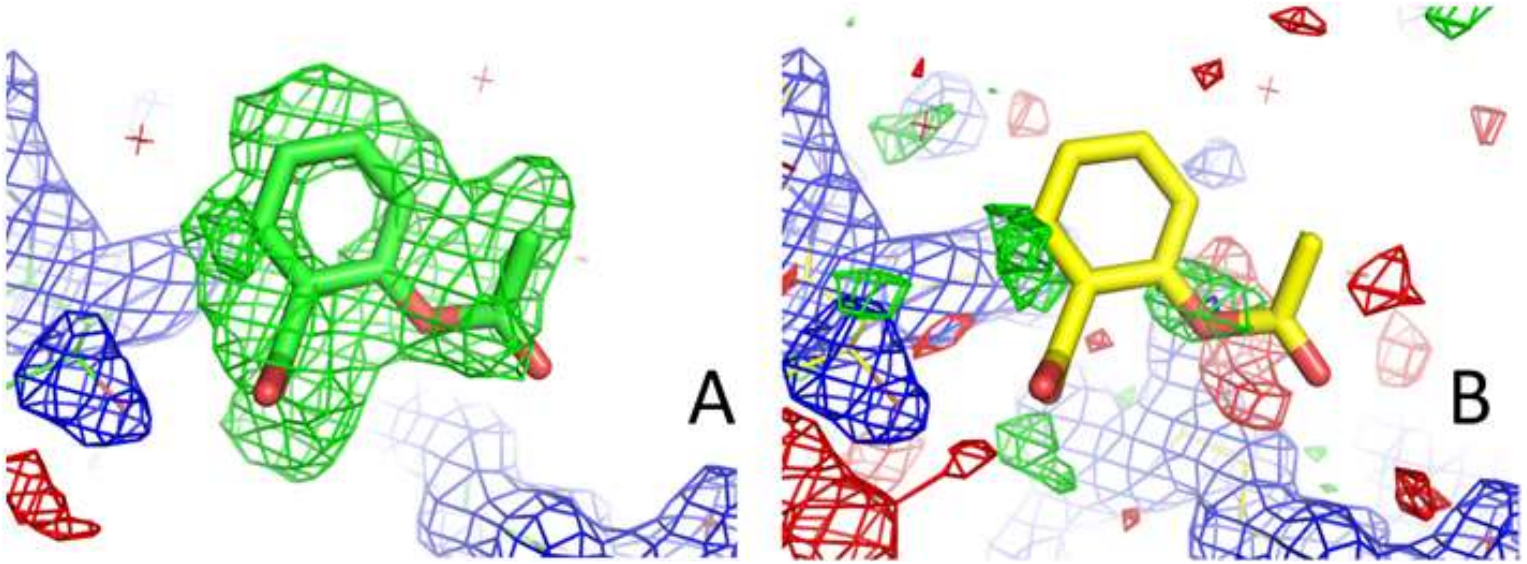
3IAZ



Singh et al. (2009): lactoferrin complexes relevant to gastrointestinal inflammation

Aspirin may give you a headache if…at absurd occupancy

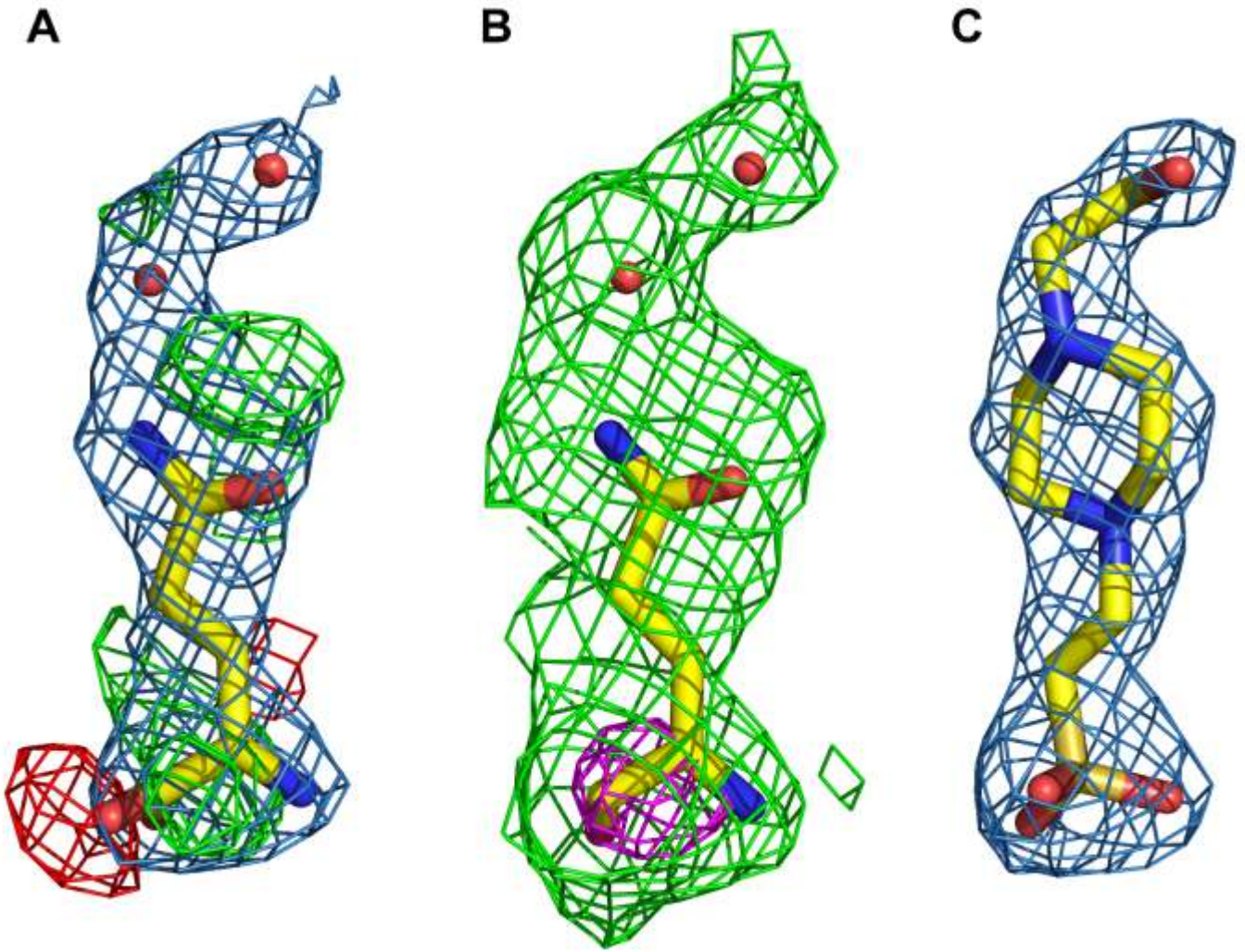
3IAZ



occ = 0.02
excluded solvent reappears in the shape of the low-occupancy ligand

true *mF*o-*DF*c omit map calculated with the ligand completely omitted from the model, contoured at 2.5σ

Mouse kynurenine aminotransferase

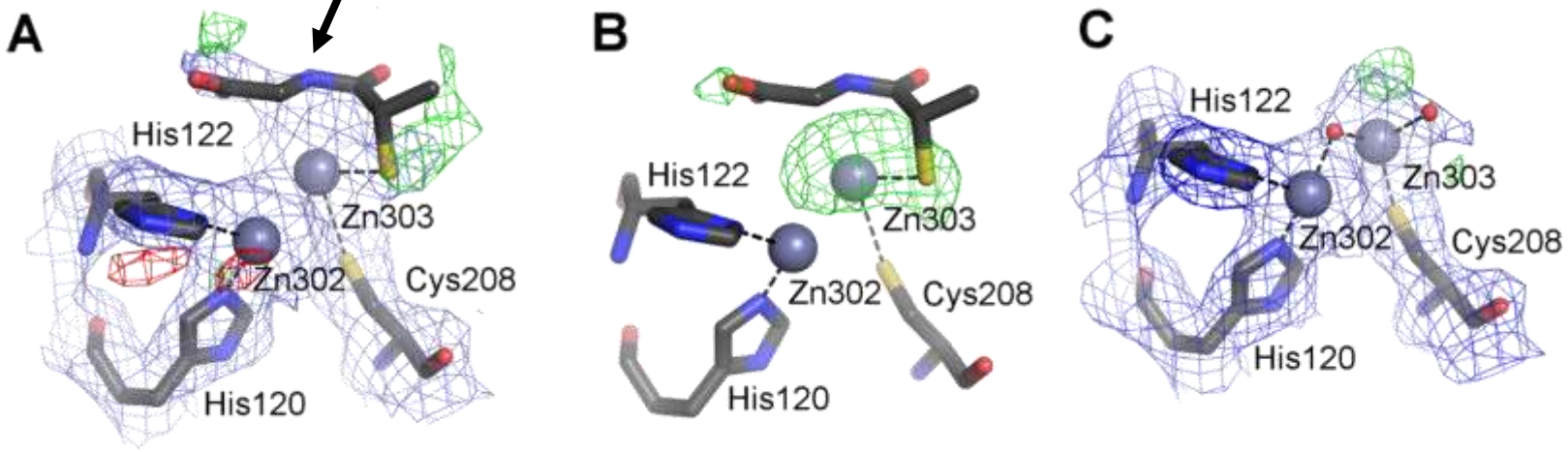


"glutamine" omit map HEPES

Han et al. Mol Cell Biol 29, 784–793, 2009

Correction of many PDB structures of metallo-β-lactamases

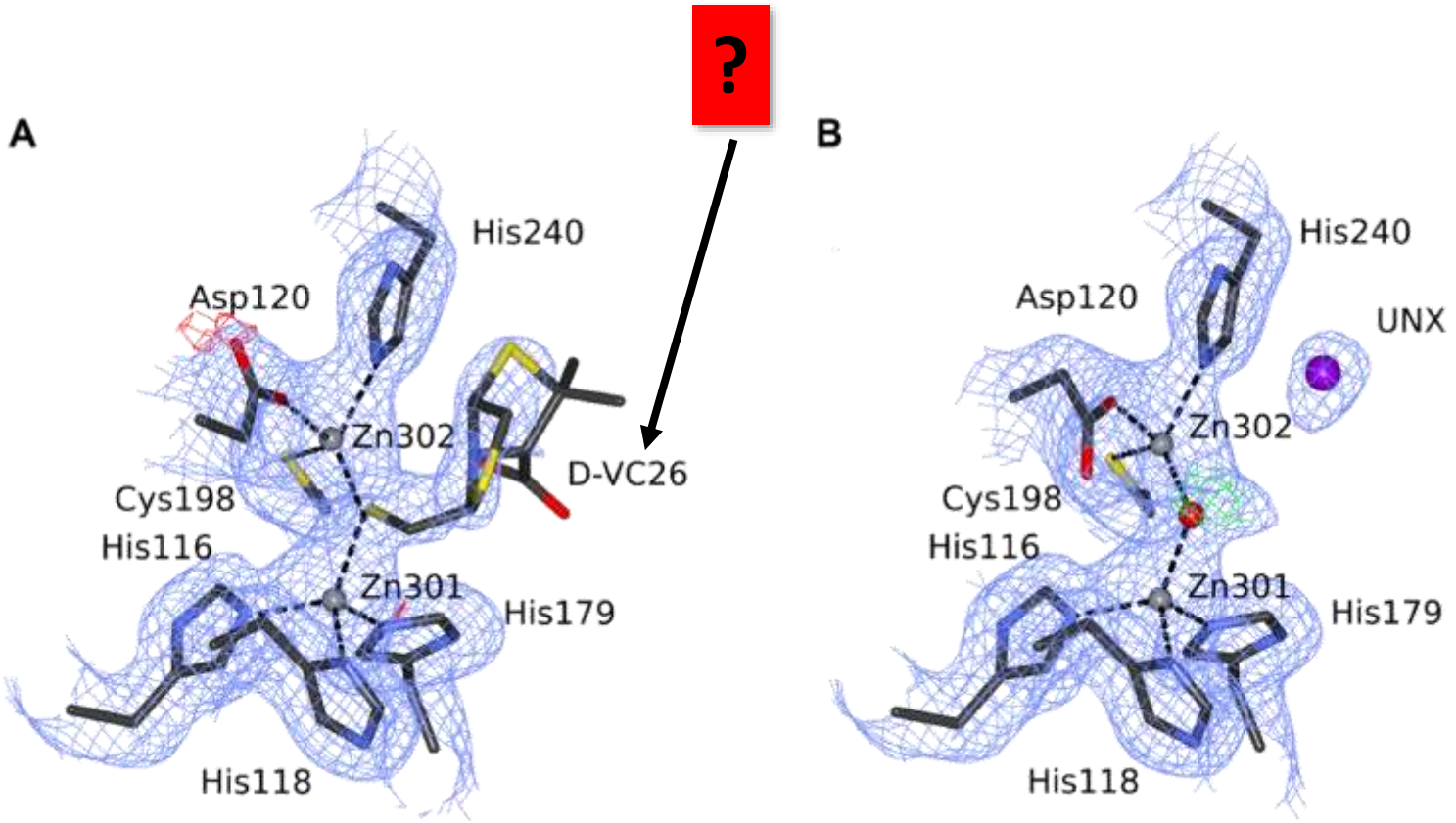
tiopronin?



original maps | ligand omit maps | final model

5a5z → 5nbk

Correction of many PDB structures of metallo-β-lactamases



original model

final model

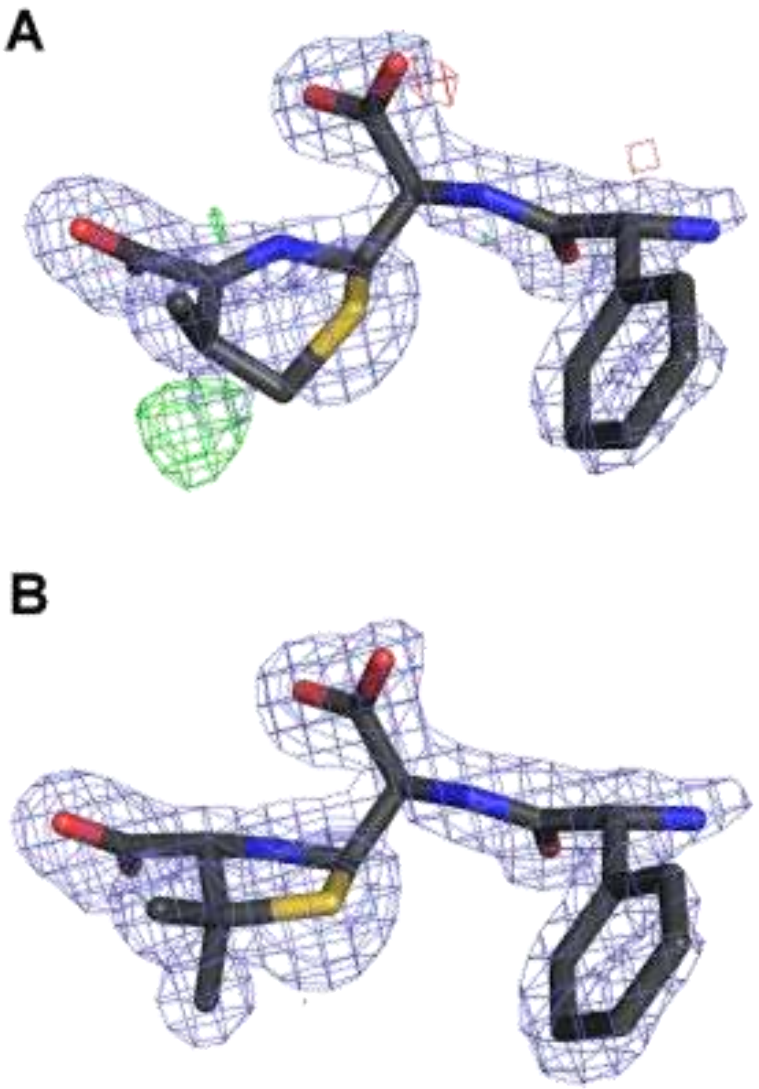
4nq7

5w8w

Correction of many PDB structures of metallo-β-lactamases

4rl2

5o2f



purported cephalexin hydrolysis intermediate

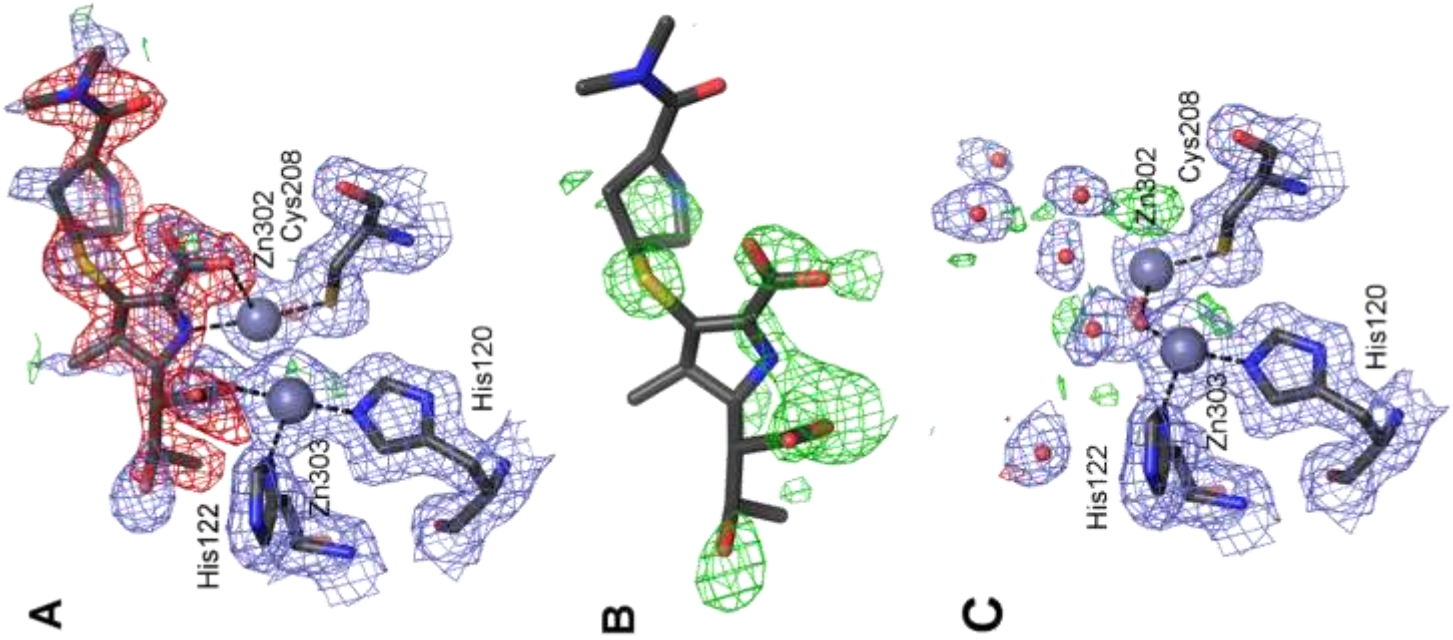
hydrolyzed cephalexin

Correction of many PDB structures of metallo-β-lactamases

original model with meropenem

meropenem in omit map

corrected model



4eyl → 5n0h

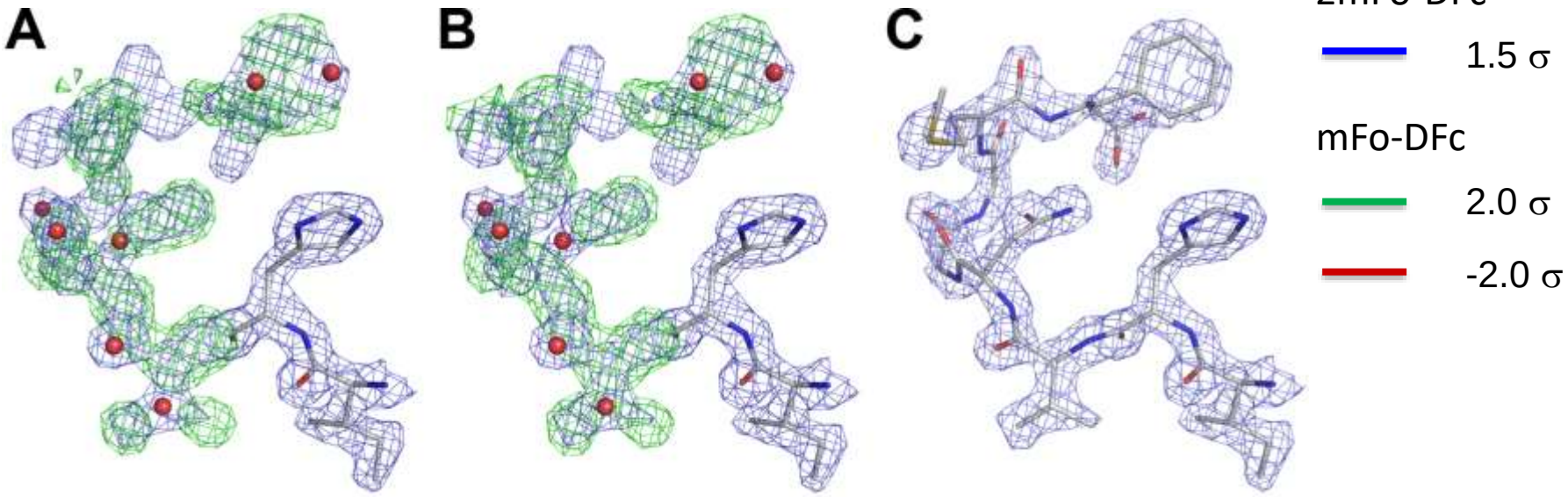
Reaction of the corrected authors (MBL structures)

original PDB ID	corrected	response of original authors
4rl2	5o2f	All communication attempts failed;
5rl0	5o2e	All communication attempts failed;
5a5z	5nbk	Complete disagreement with changes; author insists there is enough experimental evidence to support claims;
4exy	5n0i	Disagreement about glycol to mercaptoethanol change; all other changes agreed upon with author;
4eyl	5n0h	Disagreement about ligand sidechain conformation; all other changes agreed upon with author;
1k07	5wck	All changes agreed upon with author;
4nq7	5w8w	All changes agreed upon with author;
1jt1	5w90	All changes agreed upon with author;
4hky	6ex7	All changes agreed upon with author;
3m8t	5wcm	All changes agreed upon with author;

Paper >7 months in review; reviewer requested ALL e-mail correspondence with criticized authors; editors get cold feet; finally accepted by DRU

Forgotten part of the structure

2P68 (R/R_{free} 0.183/0.223)



EDS PDB_REDO manual rebuilding

Deposited 2007
"to be published"

Are the conclusions supported? – By what?

research papers

STRUCTURAL BIOLOGY

ISSN 2059-7983

Hydrogen bonds are a primary driving force for *de novo* protein folding

Schuyler Lee,[a,b] Chao Wang,[a] Haolin Liu,[a,b] Jian Xiong,[c] Renee Jiji,[c] Xia Hong,[a] Xiaoxue Yan,[a] Zhangguo Chen,[b] Michal Hammel,[d] Yang Wang,[a,b] Shaodong Dai,[a,b] Jing Wang,[b] Chengyu Jiang[e]* and Gongyi Zhang[a,b]*

[a]Department of Biomedical Research, National Jewish Health, Denver, CO 80206, USA, [b]Department of Immunology and Microbiology, School of Medicine, University of Colorado Denver, Aurora, CO 80206, USA, [c]Department of Chemistry, University of Missouri, Columbus, Mississippi, USA, [d]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA, and [e]Department of Biochemistry and Molecular Biology, Peking Union Medical College, Beijing 100005, People's Republic of China. *Correspondence e-mail: chengyujiang@gmail.com, zhangg@njhealth.org

Received 16 September 2017
Accepted 20 October 2017

Edited by Q. Hao, University of Hong Kong

Keywords: hydrogen bonds; *cis*/*trans*-proline; protein folding.

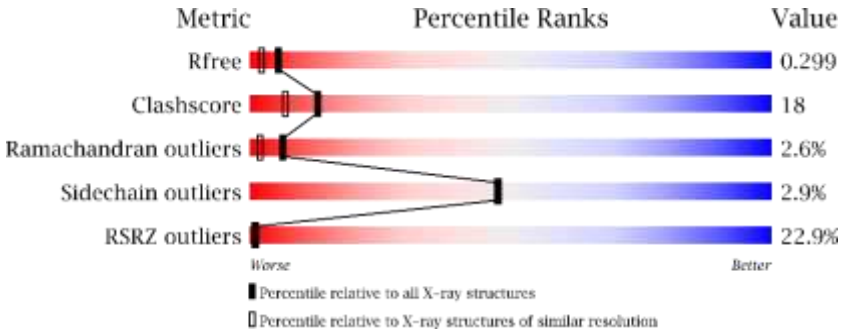
PDB reference: twinned human AID, 5w09

Supporting information: this article has supporting information at journals.iucr.org/d

The protein-folding mechanism remains a major puzzle in life science. Purified soluble activation-induced cytidine deaminase (AID) is one of the most difficult proteins to obtain. Starting from inclusion bodies containing a C-terminally truncated version of AID (residues 1–153; AID^{153}), an optimized *in vitro* folding procedure was derived to obtain large amounts of AID^{153}, which led to crystals with good quality and to final structural determination. Interestingly, it was found that the final refolding yield of the protein is proline residue-dependent. The difference in the distribution of *cis* and *trans* configurations of proline residues in the protein after complete denaturation is a major determining factor of the final yield. A point mutation of one of four proline residues to an asparagine led to a near-doubling of the yield of refolded protein after complete denaturation. It was concluded that the driving force behind protein folding could not overcome the *cis*-to-*trans* proline isomerization, or *vice versa*, during the protein-folding process. Furthermore, it was found that successful refolding of proteins optimally occurs at high pH values, which may mimic protein folding *in vivo*. It was found that high pH values could induce the polarization of peptide bonds, which may trigger the formation of protein secondary structures through hydrogen bonds. It is proposed that a hydrophobic environment coupled with negative charges is essential for protein folding. Combined with our earlier discoveries on protein-unfolding mechanisms, it is proposed that hydrogen bonds are a primary driving force for *de novo* protein folding.

PDB reference: twinned human AID, 5w09

PDB Validation Report



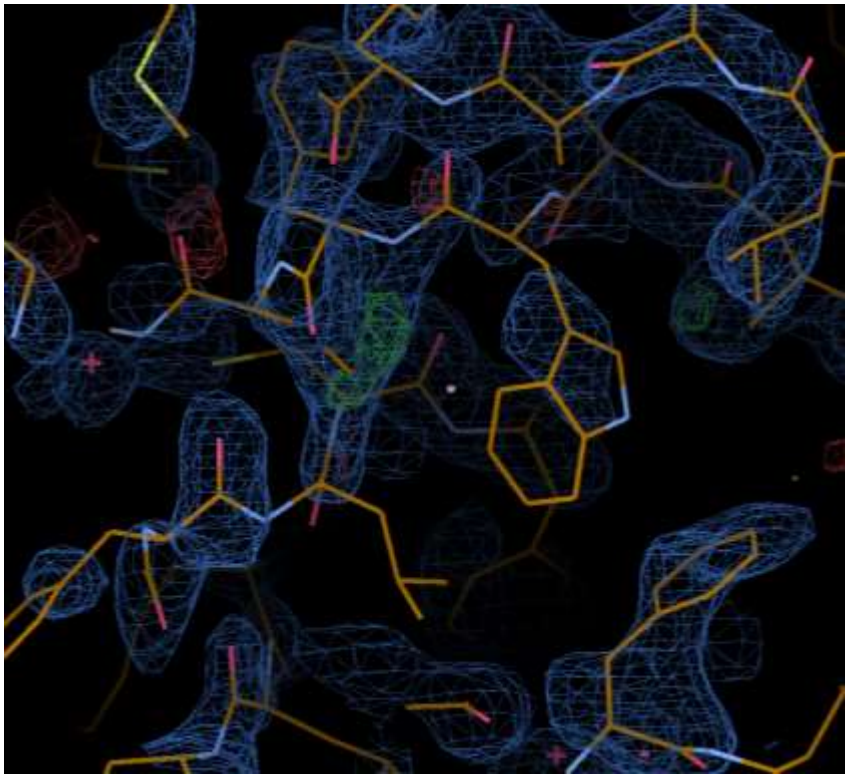
Experimental Data

·Method: X-RAY DIFFRACTION

- **Resolution**: 2.0 Å
- **R-Value Free**: 0.291
- **R-Value Work**: 0.267
- **Space Group**: $P2_1$

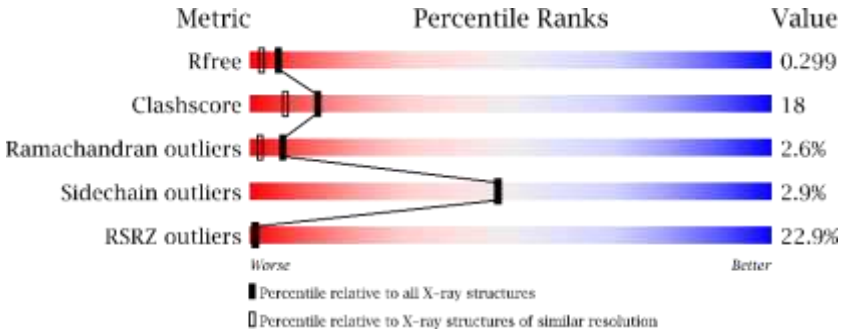
Unit Cell:

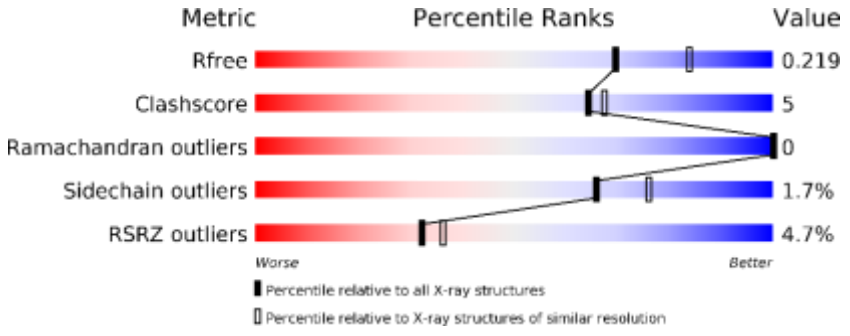
Length (Å)	Angle (°)
a = 61.46	α = 90.00
b = 28.36	β = 119.99
c = 61.51	γ = 90.00



5w09

PDB Validation Reports





5w09

$P2_1$

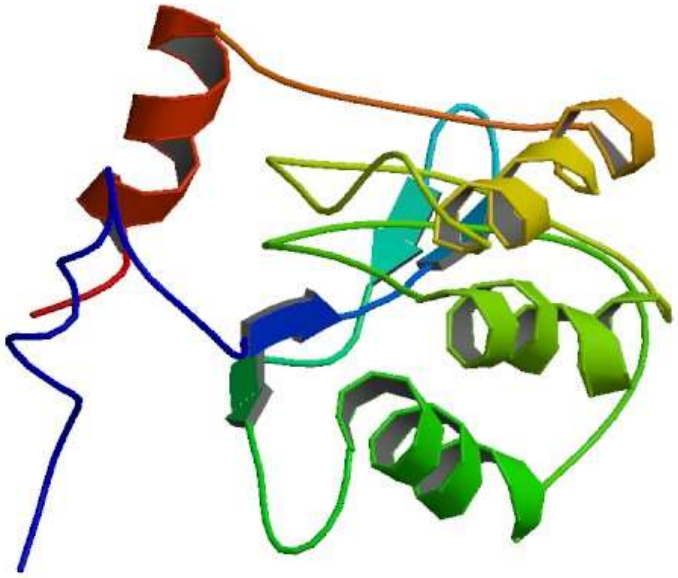
Length (Å)	Angle (°)
a = 61.46	α = 90.00
b = 28.36	β = 119.99
c = 61.51	γ = 90.00

2y90

*P*6

Length (Å)	Angle (°)
a = 61.50	α = 90.00
b = 61.50	β = 90.00
c = 28.25	γ = 120.00

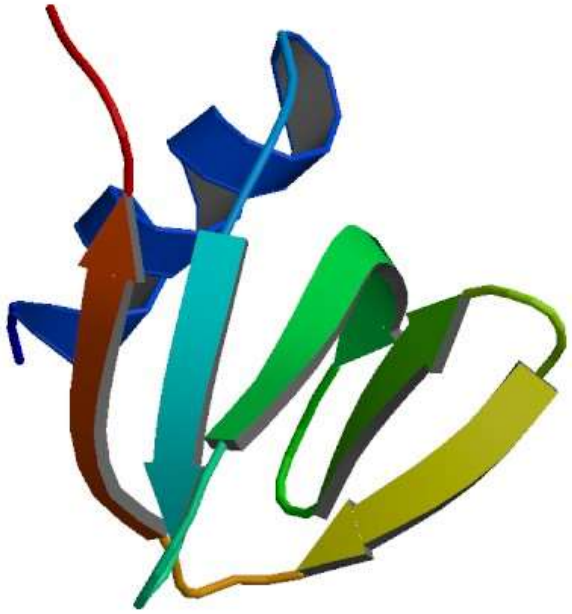
Wrong protein modeled!



5w09

$P2_1$

"AID protein"



2y90

*P*6

E. coli riboregulator Hfq protein

What should we do?

- ***trust but verify*** approach highly recommended
- structural publications should contain electron density maps supporting critical claims (ligand OMIT mFo-DFc electron density maps)
- key experimental data should be in the main text, not in Supplement
- deposition of raw diffraction images should be required
- referees should do a better job identifying suspicious structural models/claims
- journals (editors) should be more responsi(v/bl)e with retraction of papers based on fraudulent/erroneous data
- organizations like *RetractionWatch* or *PubPeer* form grassroot movement to protect science integrity
- PDB Validation Reports/protocols need revision, especially for ligand validation
- automatic remediation by PDB_REDO not very successful in difficult case
- better mechanisms of retraction/obsoleting of wrong PDB entries
- better mechanisms for linking corrected (old) PDB entries to new ones, not only NEW → OLD
- new rules for redeposition by other authors of corrected models based on original data

What to do - even more important

Training! Training! Training! Not just technical but based on sound epistemology

- Focus on Bayesian (skeptical) reasoning: How likely is it in view of established priors, that a proposition is meaningful?
- Emphasize the need to back up extraordinary claims with extraordinary proof: Do I have the necessary clear evidence?
- Understand cognitive bias: expectation bias and confirmation bias: am I deceiving myself (and others?)
- Understand logical fallacies: Appeal to Normalcy: others have done!; alternative interpretation; or demanding 'Proof of absence'

Acknowledgment

Asia Raczynska	CBB Poznan
Alex Wlodawer Zbyszek Dauter	NCI Frederick Argonne
Wladek Minor Ivan Shabalin	Univ Virginia
Bernhard Rupp	k.-k. Hofkristallamt