

Detect, correct, retract: How to manage incorrect structural models

Alexander Wlodawer¹, Zbigniew Dauter², Przemyslaw J. Porebski³, Wladek Minor³, Robyn Stanfield⁴, Mariusz Jaskolski^{5,6}, Edwin Pozharski^{7,8}, Christian X. Weichenberger⁹ and Bernhard Rupp^{9,10}

1 Protein Structure Section, Macromolecular Crystallography Laboratory, National Cancer Institute, Frederick, MD, USA

2 Synchrotron Radiation Research Section, Macromolecular Crystallography Laboratory, National Cancer Institute, Argonne National Laboratory, IL, USA

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

4 Department of Structural and Computational Biology, BCC206, The Scripps Research Institute, La Jolla, CA, USA

5 Department of Crystallography, Faculty of Chemistry, A. Mickiewicz University, Poznan, Poland

6 Center for Biocrystallographic Research, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland

7 Department of Biochemistry and Molecular Biology, University of Maryland School of Medicine, Baltimore, MD, USA

8 Institute for Bioscience and Biotechnology Research, Rockville, MD, USA

9 CVMO, k.-k.Hofkristallamt, Vista, CA, USA

10 Department of Genetic Epidemiology, Medical University Innsbruck, Austria

Keywords

electron density; error detection; evidence-based scientific discovery; Protein Data Bank; structure validation

Correspondence

B. Rupp, CVMO, k.-k.Hofkristallamt, 991 Audrey Place, Vista, CA 92084, USA
E-mail: br@hofkristallamt.org

(Received 7 October 2017, accepted 1 November 2017)

doi:10.1111/febs.14320

The massive technical and computational progress of biomolecular crystallography has generated some adverse side effects. Most crystal structure models, produced by crystallographers or well-trained structural biologists, constitute useful sources of information, but occasional extreme outliers remind us that the process of structure determination is not fail-safe. The occurrence of severe errors or gross misinterpretations raises fundamental questions: Why do such aberrations emerge in the first place? How did they evade the sophisticated validation procedures which often produce clear and dire warnings, and why were severe errors not noticed by the depositors themselves, their supervisors, referees and editors? Once detected, what can be done to either correct, improve or eliminate such models? How do incorrect models affect the underlying claims or biomedical hypotheses they were intended, but failed, to support? What is the long-range effect of the propagation of such errors? And finally, what mechanisms can be envisioned to restore the validity of the scientific record and, if necessary, retract publications that are clearly invalidated by the lack of experimental evidence? We suggest that cognitive bias and flawed epistemology are likely at the root of the problem. By using examples from the published literature and from public repositories such as the Protein Data Bank, we provide case summaries to guide correction or improvement of structural models. When strong claims are unsustainable because of a deficient crystallographic model, removal of such a model and even retraction of the affected publication are necessary to restore the integrity of the scientific record.

Abbreviations

CDR, complementarity determining region; ED, electron density; NAG, *N*-acetyl-D-glucosamine; NSAIDs, nonsteroidal anti-inflammatory drugs; PDB, Protein Data Bank.

Success is its own enemy

In a sense, macromolecular crystallography has become a victim of its own spectacular success. Technical progress allows any sufficiently trained structural biologist to contribute to the avalanche of new results, evidenced by the nonlinear growth of the number of structural models deposited in the Protein Data Bank (PDB) [1]. The power of macromolecular crystallography as a branch of science originates from two major strengths: (a) Each experiment itself is extraordinarily rich in data – every single diffraction experiment may consist of hundreds of gigabytes of raw data and hundreds of thousands of reduced observations. Powerful statistical methods combined with mathematical rigor make, in principle, the path from diffraction data to reconstructed electron density (ED) map quite robust. (b) Biomolecular crystallography is founded upon the rich body of prior knowledge, which provides highly reliable guidance regarding plausible stereochemistry of new macromolecular models and expected interactions. Yet, despite these strengths, potential danger lies in putting the powerful technology in inexperienced hands. This, when combined with lack of scientific rigor, which is indispensable in crystallographic practice, can lead to lamentable consequences.

The source of problems and general preventive measures

The ready access to powerful tools by untrained or poorly supervised users might point towards technical ineptitude as the primary reason for errors. However, examination of recent publications either highlighting some ‘bad apples’ [2], or simply ranking models by numeric quality indicators [3–5], indicates that poor models are almost always associated with the main Achilles heel of biomolecular crystallography: the interpretation of the electron density. In contrast to small-molecule crystallography, the macromolecular electron density maps are rarely at atomic resolution, are sometimes of poor quality, and are frequently compromised as a result of – hard to deconvolute – complex molecular disorder and heterogeneity. The step of electron density interpretation allows the subjective element of the human mind, which is always present, to influence the process of model building.

Two major elements associated with the human mind threaten the robustness of the process: the well-documented cognitive expectation- and confirmation bias [6,7], and the neglect of rigorous discipline in empirical reasoning. There is nothing new to this insight: Cognitive bias was already recognized by the great minds of early

Enlightenment, cf. [8]. Subsequently, it was realized that prior knowledge can restrain expectations [9,10] and that empirical reasoning demands a strong claim to be supported by correspondingly strong experimental evidence. In addition, falsifiability is the fundamental requirement of a scientific hypothesis [11,12]. Protein crystallography very early adopted Bayesian and Likelihood concepts [13–18] to counteract the pervasive wish to find what one seeks [19,20], and proposed better epistemological training as a systemic remediation, e.g. [21,22].

The mighty burden of proof: weak or nonexistent electron density versus new biomedical findings

Problems with structure models are either discovered by individual scientists with an interest in the particular subject of study, e.g. [23], or by algorithms that assess and rank models by certain quality indicators, such as fit to electron density [3,24,25], stereochemistry [26,27], chemical plausibility [5,28], or some combinations thereof [4,29]. As repeatedly pointed out, the weakness of the automated, unsupervised data mining and validation programs (for review see [30,31]) is that they are unaware of the specific claim or hypothesis that the crystallographic model intends to support [4,22,32]. Precisely herein lies what separates a legitimate freedom of interpretation from epistemological heresy: without evaluation of the context – that is, without reading the publication where the claim associated with the questioned feature is proposed – and without careful local inspection of the electron density maps, it is not possible to fully appreciate the level of evidence necessary to support a given claim. For instance, some disordered features of a macromolecule such as glycans, or components of the crystallization cocktail and unidentified solvent components are often characterized by a poor score in some quality metric, yet may be an indication – still informative and based on evidence – that some entity is present at a given location. Inclusion of such features may, for example, restrict the accessible molecular space in modelling and docking of functional ligands, as discussed by [33], or affect the binding of biologically relevant molecules [34,35]. Specific claims or hypotheses, however, must acknowledge the speculative nature of such evidential observations.

The situation is entirely different when crystallographic evidence becomes the basis, or provides critical support, for a strong scientific statement or hypothesis. The presence of a key ligand, whether a small-molecule drug lead or an immunologically relevant peptide, in a specific location and in a specific conformation must be supported by correspondingly convincing evidence. It is

generally expected that a combination of bias-minimized $2mF_o - DF_c$ electron density and positive $mF_o - DF_c$ omit difference density at an adequate level provides such evidence [22,36].

In simple words, the electron density should look, without undue stretch of imagination, like the outline of the ligand model. Reliable tools for electron density-based validation exist, for example as summarized in [30,37], and – if applied reasonably and with caution – advanced electron density construction and refinement methods can compensate for solvent intrusion and provide improved model fit [38,39].

Absence of supporting evidence is indeed proof of absence

Why is it that weak electron density still beckons to be filled with at least a fragment of a desired ligand? In addition to expectation and confirmation bias, which are known and have been discussed almost *ad nauseam* even in the crystallographic literature [21,40,41], an epistemological problem surfaces here as well. In defence of unsustainable claims not supported by adequate evidence, arguments are voiced which reduce to the point that a valid critique must prove the absence of a ligand with certainty. Such response fails on two grounds. Firstly, and fundamentally, the absence of something nonexistent cannot be proven by empirical science. Therefore, the best the critic can do is to show the absence of convincing evidence. Secondly, from a simple estimate, it can be understood why there will be, invariably and always, spurious evidence that can be abused to ‘prove’ that proof of absence has failed. This is so because electron density reconstruction relies on noisy and incomplete diffraction data originating from an average of billions of flexible molecules self-assembled into an imperfect crystal lattice. The resulting electron density map is correspondingly noisy and thus contains peaks and continuous fragments that could be interpreted as the desired density if displayed at a level at which noise is predominant. Under the conservative assumption of a random distribution with a zero mean for difference density, a positive difference density level of more than 2.5σ will appear about once in 160 density voxels (corresponding to a volume of about $5 \times 5 \times 5$ grid points), which at oversampling at the Nyquist limit of $d_{\max}/2$ corresponds to once in every volume of $\sim 5 \times 5 \times 5 \text{ \AA}^3$ at 2 Å resolution.

Classification of severe errors and possible remedies

Severe errors generally fall in two primary categories. One of them involves direct, evidence-linked problems

of electron density misinterpretation including absent electron density, leading to a low data likelihood term in a Bayesian likelihood model. Density misinterpretation often goes together with the second category of problems, namely that the proposed model, often because of nonexistent evidence, violates basic prior expectations. Violation of stereochemistry, implausible chemical environment, or occurrence of severe steric clashes, are typical examples (cf. the implausible peptide in the 36–65 germline antibody [42]). Large numbers of close contacts between atoms are unlikely. Here, the prior probability of the model is already low, and combined with poor evidence, the joint posterior probability of the model, the model likelihood, becomes vanishingly small. Such models are often detected by automated validation software and appear as worthy of inspection.

Models that are not supported by convincing evidence, such as ligands placed in the noise of electron density maps, can only be improved by removing the sources of the noise to clearly demonstrate that evidence for the suspicious ligand is lacking. This can be done by removing the spurious ligand and/or correcting other errors of the model that are usually unrelated to the ligand presence, and may also require better processing of the diffraction data. While such procedures minimize the noise, the situation remains unsatisfying because no improvement of the most crucial part of the model – the nonexistent ligand – can be achieved. Nonetheless, correction of other possible errors and replacing the model in the database does contribute to restoring the structural database integrity. Whether the corresponding publication needs an erratum or to be retracted in its entirety, depends on the scientific importance of the unsupported claims which must be thoroughly judged on an individual basis.

Selection of examples

In the following discussion, we present several case studies with errors of different severity, and propose corresponding remedial action. We follow what has been previously termed case-controlled validation [32], emphasizing that the context of the proposed new scientific finding plays a crucial role in determining the optimal corrective action. Many trivial errors and mistakes can be simply corrected thanks to improved contemporary validation and refinement tools, and they are not the subject of our discussion. Efforts such as PDB_REDO [43] automatically re-refine and improve all structural models using the newest version of refinement software; however, this advance is necessarily

limited to models that can actually be improved, and the procedure is also incognizant of case-specific context.

A recently generated validation database, Validator^{DB} [44], was mined for trends in structure quality (ValTrends^{DB}) and revealed that, whereas some metrics such as clash scores or geometry improved over recent years, ligand model quality alarmingly did not improve, emphasizing our focus on protein-ligand complex models. The examples are provided in sequence of decreasing severity of the errors as determined by unsupervised data mining. Small-molecule ligand structure models were selected based on real space fit ranking in the TWILIGHT-1 data base [22] or for peptide ligands by TWILIGHT-2, based on a combined real space and geometry score [4]. One model [45] was flagged by CHECKMYMETAL [28] as containing an unusual metal coordination. A study [46] flagged by Salunke and Nair [47] as a case of an electron density map for a ligand contoured at very low level, revealed amongst other errors a different ligand (buffer) in the electron density, which unsupervised methods failed to detect due to an acceptable real space fit and the fact that free amino acid ligands are annotated by the PDB with the same identifier as an amino acid residue and not recognized as a ligand. The example from a recent study [48] was identified based on the correlation emerging between detected errors and certain author clusters. The proposed implausible ligand model was not detected by unsupervised TWILIGHT data mining due to real space fit above the error cutoff, and became obvious only by careful map inspection. Finally, some selected models also revealed suspect features of the journal presentation, such as contouring electron density maps of unclear provenance and/or at inappropriate levels. It needs to be stressed, though, that our selection of examples for this paper cannot be all-inclusive, but rather attempts to cover a range of commonly encountered problems.

Crystallography

Initial evaluation of the electron density maps was based on data downloaded from EDS [24] and displayed with the program COOT [49,50]. For re-refinement, coordinates and structure factors were downloaded from the PDB and maps were reconstructed using HKL3000 [51] coupled with REFMAC5 [52,53]. If properly marked in the deposited structure factors, the same set of reflections was used for cross-validation via R_{free} calculations [54]; otherwise, a new set of R_{free}

reflections was generated *a posteriori* by applying a suitable parameter-annealing procedure to erase the memory of the test reflections. Most models were rebuilt using the modules of the HKL3000 package coupled with COOT [50] and were re-refined with REFMAC5 [52,53]. The structure 2A6I (§3.4) was refined with PHENIX [55] and rebuilt with COOT. In all cases, maximum-likelihood σ_A [16] weighted $2mF_o - DF_c$ and $mF_o - DF_c$ electron density maps were calculated, with m being the figure of merit and D the Luzzati coefficient; cf. p. 619 ff. of [56] for a review of maximum-likelihood map coefficients. Difference electron density maps were routinely displayed at the 3σ level, and $2mF_o - DF_c$ maps at 1σ . However, different contour levels were used when a need arose to address or compare the levels of previously published figures. Final model statistics, such as the percentage of Ramachandran outliers and clashscores, were generated with the MOLPROBITY on-line server [57].

Model comparisons using MOLSTACK

Some models were refined only superficially to verify the presence or absence of certain features; the models resulting from such abbreviated refinement were not resubmitted to the PDB. However, selected structures were fully re-refined and in these cases the resulting coordinates were submitted to the PDB. MOLSTACK [58], a new cloud-based platform for structural data that allows the presentation of structural models alongside electron density maps, was employed to present original and corrected structures in a side-by-side interactive fashion within a standard web browser. Table 1 includes selected statistics for the original PDB entries and for the re-refined models.

Effect and pitfalls of unrefined B-factors and near-zero occupancies

Unrefined B-factors

In a paper postulating the structural basis for the prevention of nonsteroidal anti-inflammatory drug-induced gastrointestinal tract damage by the C-lobe of bovine colostrum lactoferrin [59], the authors describe four crystal structures of this protein (PDB ID 3IAZ, 3IB0, 3IB1, 3IB2; these are replacements for 2G5J, 2B6D, 2ALT, 3HWQ) determined using crystals prepared from samples that contained four different nonsteroidal anti-inflammatory drugs (NSAIDs). The binding affinities were inferred from tryptophan fluorescence quenching with apparent binding constants in

Table 1. Statistics for the structures re-refined in this project (right column for each entry). The left column for each entry gives the refinement statistics corresponding to the original PDB deposition. Ramachandran analysis and clashscores were calculated with the MolProbity server.

	Mouse kynurenine aminotransferase apo form		Mouse kynurenine aminotransferase with glutamine		Mouse kynurenine aminotransferase with kynurenine		<i>Aedes aegypti</i> kynurenine aminotransferase	
Resolution (Å)	29.2–2.59		29.6–2.26		29.6–2.81		39.3–1.55	
No. of reflections	29 069		45 037		24 006		123 988	
$R_{\text{work}}/R_{\text{free}}$	0.186/0.263	0.194/0.264	0.177/0.221	0.173/0.221	0.194/0.237	0.167/0.251	0.254/0.279	0.186/0.215
No. of atoms protein/ligand or ion/water	6536/60/140	6536/84/162	6506/76/411	6408/113/405	6536/60/140	6535/91/101	6660/5/440	6515/18/801
 factors (Å ²) protein/ligand or ion/water	30.4/51.0/30.3	36.8/48.0/34.4	27.6/25.2/31.4	33.8/37.4/37.2	30.4/51.0/30.3	30.5/36.9/24.5	20.8/3.0/24.0	19.6/22.9/28.1
R.m.s. deviations from ideal bonds (Å)/angles (°)	0.025/2.21	0.013/1.77	0.020/1.83	0.014/1.74	0.024/2.21	0.013/1.88	0.010/2.41	0.017/1.91
Ramachandran analysis (%) most favoured/ allowed/outliers	92.2/6.6/1.2	94.7/4.6/0.7	95.8/3.7/0.5	97.0/2.9/0.1	89.2/9.7/1.1	92.6/6.3/1.1	92.2/5.6/2.2	98.1/1.8/0.1
Clashscore/ percentile	17.9/82nd	5.9/99th	12.0/88th	6.1/99th	22.2/86th	7.0/99th	28.7/3rd	4.3/96th
PDB entry	3E2F	5VEP	3E2Y	5VEQ	3E2Z	5VER	1YIZ	5VEH
	Fab 36–65 in complex with peptide KLA		Eggplant vicilin		scFv 2D10		Phospholipase A2 in complex with designed pentapeptide	
Resolution (Å)	57.8–2.50		22.7–1.49		35.1–1.55		19.51–2.00	
No. of reflections	15 460		70 495		41 534		17 899	
$R_{\text{work}}/R_{\text{free}}$	0.245/0.264	0.203/0.250	0.199/0.210	0.131/0.163	0.167/0.193	0.153/0.180	0.187/0.198	0.155/0.227
No. of atoms protein/ligand or ion/water	3309/67/55	3360/10/131	2912/14/256	3252/20/353	1829/69/291	1910/100/280	1888/57/283	1888/0/268
 factors (Å ²) protein/ligand or ion/water	21.0/55.0/21.8	27.1/32.8/23.3	20.5/35.1/30.6	18.2/28.4/33.4	22.9/16.6/37.4	23.5/35.0/37.2	36.2/42.3/53.3	40.2/-/49.2
R.m.s. deviations from ideal bonds (Å)/angles (°)	0.008/1.68	0.003/0.625	0.009/1.97	0.014/1.69	0.010/1.59	0.015/1.71	0.007/1.34	0.016/1.79
Ramachandran analysis (%) most favoured/ allowed/ outliers	89.6/5.3/5.1	96.5/3.5/0.0	95.5/3.7/0.8	98.1/1.9/0.0	97.0/2.6/0.4	97.0/3.0/0.0	97.5/2.1/0.4	95.8/3.8/0.4
Clashscore/ percentile	38.0/25th	0.3/100th	9.8/58th	3.4/97th	3.5/97th	2.5/99th	24.1/20th	3.8/99th
PDB entry	2A6I	5VGA	5CAD	5VF5	5I4F	5VF2	1JQ8	5VET

the submillimolar range. The authors report that for four NSAIDs they observe ‘reasonably characteristic electron densities at the ligand-binding sites’. This

finding is contradicted by the electron density maps produced by the EDS server and our own calculations, as illustrated in fig. 13 in [22]. The real space

correlation coefficients (RSCC) place all of these drug ligands into the poorest density fit category in TWILIGHT-1 [3], and the ‘definitely bad’ category of VHELIBS analysis [5] for the complexes with indomethacin (PDB ID 3IB1, IMN 701A RSCC 0.45, 2.2 Å), diclofenac (3IB0, DIF 701A RSCC 0.29, 1.4 Å), aspirin (3IAZ, AIN 1202 RSCC 0.28, 2.0 Å) and ibuprofen (3IB2, IBP A3960 RSCC 0.62, 2.29 Å). Refinement of these models would simply involve removing the ligands. Redeposition would likely provide few new insights beyond the automated PDB_REDO re-refinement [43]. Inspection of the models and data do, however, allow the discussion of a few very relevant points.

Figure 2 in [59] shows clear electron density for the drug ligands, but inspection of the ligand density of each entry through the PDBe ligand tab as well as fig. 13 in [22] shows that this electron density cannot be produced by any acceptable means. Interestingly, the aspirin molecule (AIN 1202A) shows quite clear negative electron density for the ligand (Fig. 1A,B). The reason is that, as deposited, the aspirin has relatively low B-factors, comparable to those of the protein. Once the B-factors are refined, the difference density disappears and no more evidence for the aspirin

molecule can be found (Fig. 1C,D) in any type of electron density maps.

Near-zero occupancies

Another means of self-deception leading to biased map generation is the questionable practice of setting the occupancy of disordered parts of a molecule to very low values, say 0.02, instead of full or near-full occupancy approaching 1.0. When this practice is extended to ligands, two concurrent events happen: The refinement program finds the atoms placed at whatever their position, and consequently, prevents the solvent mask from intruding into this space. At the same time, there is practically no scattering contribution to F_c left from a ligand molecule with 2% occupancy. When there is no ligand present, the actual solvent density occupying this place in the crystal structure will invariably fill this void resulting in a low but distinct F_o contribution. The result is positive $F_o - F_c$ difference electron density in the shape carved out by the ligand. The difference in appearance relative to a correctly generated omit difference map is striking (Fig. 2). A feature of such ‘low-occupancy’ $mF_o - DF_c$ difference electron density

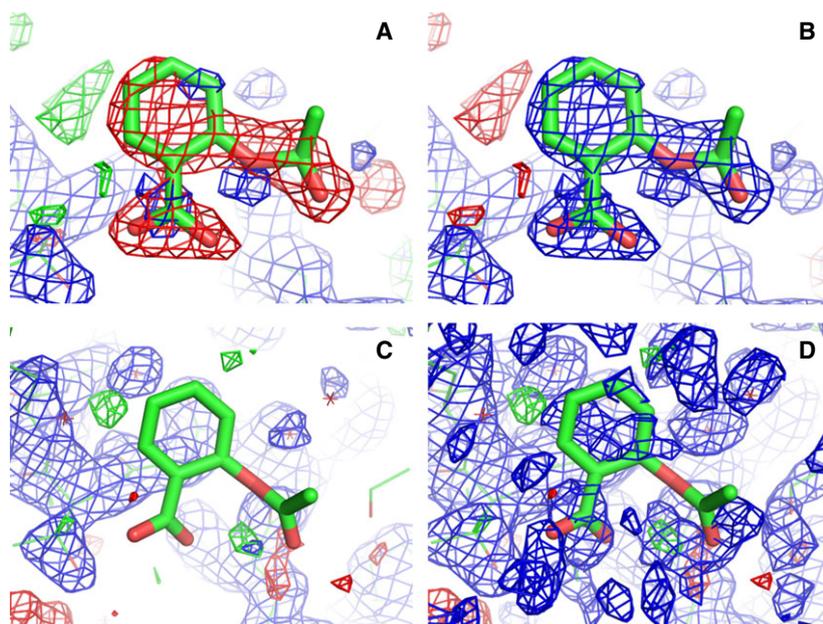


Fig. 1. The effect of unrefined B-factors. (A) $mF_o - DF_c$ negative difference electron density contoured in red at -3σ for aspirin 1202A in 3IAZ, calculated from deposited model and data. A simple colour change of that same map to blue contours (B) imports the appearance of an $2mF_o - DF_c$ electron density map comparable to that of the protein. Once the B-factors are refined (C), they climb from ~ 30 to $> 100 \text{ \AA}^2$ and no more $mF_o - DF_c$ difference electron density at the 3σ level nor any $2mF_o - DF_c$ electron density at 1σ can be discerned. Contouring down to 0.4σ , close to noise levels (D) shows that there is perhaps a water molecule in the vicinity, but evidence for the aspirin molecule is absent.

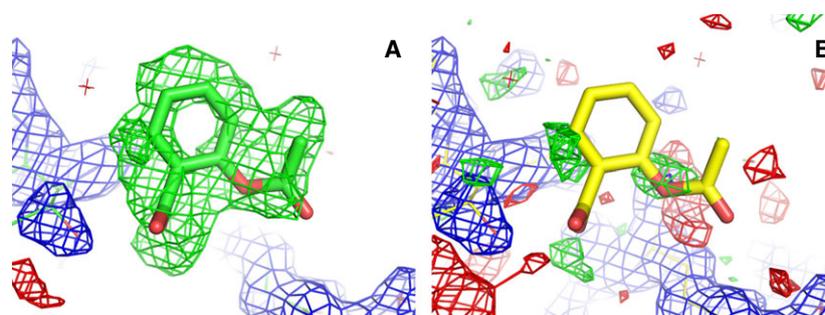


Fig. 2. Comparison of a 'low-occupancy' ligand difference map and a correctly generated omit difference map. When the solvent is excluded from the map calculation due to the placement of a low-occupancy ligand model, then the contribution of the actually present solvent cannot be calculated in that carved-out region, and will show up as positive difference electron density. Panel (A) shows such a 'low-occupancy' mF_o-DF_c map (green) for the aspirin molecule in 3IAZ contoured at 3.5σ , while (B) shows a 'true' mF_o-DF_c omit difference map (that is, calculated with the ligand completely omitted from the model) of the same region, contoured at 2.5σ .

maps is often the almost complete absence of reasonably contoured $2mF_o-DF_c$ electron density overlapping the ligand.

Summary: The crystal structures of the C-lobe of bovine colostrum lactoferrin do not support drug binding. In this series of structures, present in the PDB and the scientific literature, no evidence exists from crystallography that the purported NSAIDs bind where stated. Significant discrepancies exist between the electron density maps displayed in the original publications versus maps that can be generated by accepted standard procedures. Caution must be exercised when maps are constructed in unconventional ways. A possible explanation for the absence of ligands in this entire series of models could be the deposition of structure factor amplitudes corresponding to ligand-free (apo) structure. In this case, the authors would need to produce the correct original data. Here and in any similar cases, the journals where the original work was published should be informed of the problem, and an erratum linked to the new PDB deposit (if any), should be published.

Vanishingly small evidence for ligand molecules in entries without primary publication

The case of the ribosome-inactivating protein from *Momordica balsamina* illustrates the problems caused by the proliferation in the PDB of multiple coordinate sets of the same protein in complex with many ligands, some of them of doubtful validity and with no reference to the published literature. The structure of a very closely related enzyme, α -momorcharin from *Momordica charantia*, was originally determined in

1994 in the free form at 2.0 Å resolution, as well as complexed with adenine and formycin phosphate, each at 2.2 Å [60]. While quite old, these structures appear to be of reasonable quality, although the lack of structure factors in the PDB precludes independent evaluation of the electron density for the ligands. The amino acid sequence of α -momorcharin is 94% identical to the sequence of the ribosome-inactivating protein from *Momordica balsamina* and the crystals of both proteins are isomorphous. The PDB presently contains 60 isomorphous structures of the latter enzyme, without and with diverse ligands, all determined more recently by another laboratory (Table 2). The structures were apparently solved by molecular replacement [61], although this was not necessary given the isomorphism. Most of these structures have not been published, thus it is not clear why, for example, six separate sets of coordinates of the apoenzyme, all isomorphous, have been deposited. Only 12 structures of the complexes with various ligands have been published so far [61–63]. Most of the structures have been refined at a resolution in the range of 1.6 to 2.0 Å, and only 11 at lower resolution, within 2.65 Å.

The ribosome-inactivating protein interacts with nucleic bases of RNA and DNA. For 24 structures that contain various nucleic bases and their substituted derivatives, the EDS electron density supports the presence of these ligands. The remaining 30 structures contain diverse ligands modelled in the active site of the enzyme, such as various sugars, amines and other small molecules. In all structures, one of the asparagines is N-linked to mono- or di-NAG (*N*-acetyl-D-glucosamine) and most structures list one or more molecules of glycerol.

Inspection of the electron density of these crystal structures reveals that many ligand molecules, other

Table 2. Crystal structures of the ribosome-inactivating protein and their complexes with various ligands in the PDB.

PDB code	Resolution (Å)	Ligand	Ligand quality	Reference
Original PDB deposits				
1AHA	2.20	Adenine	? – OK in paper	[60]
1AHB	2.20	Formycin-P	? – OK in paper	[60]
1AHC	2.00	–		[60]
Subsequent PDB deposits				
3S9Q	1.67	–		[61]
4L66	1.70	Water		
3MRW	1.70	–		
4KMK	1.65	–		
4KWN	1.80	–		
5GM7	1.78	–		
3N1N	2.23	Guanine	OK	[61]
3R19	1.90	Adenine	OK	[61]
3U6Z	1.70	Adenine	OK	[61]
3V2K	2.07	Adenosine-PPP	OK	[61]
3SJ6	1.60	Ribose	Dubious	[61]
4I47	2.65	Me-guanine	OK	[62]
4EMF	1.77	Me-hydroguanosine-PP	OK	[62]
4EMR	1.75	Me-guanosione-PPP	OK	[62]
4ZT8	1.98	Cytidine	OK	[63]
5CSO	1.78	Cytidine	OK	[63]
5CST	1.78	Cytidine-PP	OK	[63]
4ZU0	1.80	Cytidine-P	OK	
4ZZ6	2.00	Cytidine-PPP	Not supported	
3U70	2.00	Adenine	OK	
4O4Q	1.81	Uridine	OK	
4O8E	2.00	Uridine-P	OK	
5ILW	1.98	Uridine	OK	
5ILX	1.70	Uracil	OK	
3N3X	1.70	Guanine	Not hexapeptide, as stated in PDB	
4Q9F	1.75	Guanosine-P	OK	
3MY6	2.65	Me-guanine	OK	
4F9N	2.65	Me-guanine	OK	
3QJI	1.75	Me-guanosine-PPP	OK	
3Q4P	1.80	Me-hydroguanosine-PP	OK	
3MRY	2.00	6-aminopurine	OK	
4O0O	2.59	5-fluorouracil	OK	
4KPV	2.57	Pyrimidine-2.4-dione	OK	
3N1D	1.70	Ribose	Not supported, spurious PEG 251	
3N31	2.11	Fucose	Dubious	
3N5D	1.90	Glucose	Not supported, GOL 249	
3NFM	2.50	Fructose	Not supported, GOL 249	
3NJS	2.10	Lactose	Not supported, GOL 3968	
3NX9	1.70	Maltose	Not supported	
3V14	1.70	Trehalose	Not supported	
4HOA	2.00	Lactose	Not supported	
4JTP	1.85	Ascorbic acid	Not supported	
3U6T	1.85	Kanamycin	Not supported	
4FZ9	1.70	Mannose, NAG	Not supported, GOL 302	
4RZJ	1.98	NAG	Not supported, GOL 303	
4H0Z	2.00	<i>N</i> -acetyl-muramic acid	Not supported	
4LWX	1.78	Peptidoglycan	Not supported, GOL 302	
3U8F	1.55	Mycolic acid	Not supported	
4GUW	1.60	Lipopolysaccharide	Not supported	
5CIX	1.88	Triethanolamine	OK	

Table 2. (Continued).

PDB code	Resolution (Å)	Ligand	Ligand quality	Reference
4K2Z	1.80	Methylethylamine	OK	
4LRO	1.98	Spermidine	Not supported	
4LT4	1.69	Arginine	OK	
4M5A	1.70	Dimethylarginine	OK	
4FXA	1.70	N-acetylarginine	OK	
4DWM	1.70	NAG	No NAG in complex	
4XY7	2.50	NAG	No NAG in complex	
4KL4	1.90	PEG	Dubious	
5GZ7	1.95	Glycol	OK	
4JTB	1.71	Phosphate ion	OK	

than nucleic bases and some amines, are not supported by the $2mF_o - DF_c$ maps or by the omit maps recalculated from the diffraction data. Several examples of dubious ligands are listed in Table 2 and the corresponding maps are presented in Fig. 3. Apart from the ligands in the active site, the presence of several other small molecules, such as glycerol or glycol, is quite doubtful, since their corresponding electron density is also not convincing.

There are also some bookkeeping errors in several of the PDB deposits, e.g. the structure 3N3X contains guanine instead of the hexapeptide declared in the title

of that entry. The reference given for the PDB structures 4ZZ6, 4ZT8, 4ZU0, 5CSO and 5CST [63] (cited with the wrong year of publication), describes only three of them (4ZT8, 5CSO and 5CST). It is also not clear why three of these structures (4ZZ6, 5CSO and 5CST) supersede the earlier submitted and obsoleted structures 4KGS, 4ZTW and 4ZW4. Additionally, these 60 models are placed in 11 different locations in the unit cell, making their comparison cumbersome and nonintuitive. We did not feel that re-refinement of any of the structures with book-keeping errors or placement inconsistencies was warranted.

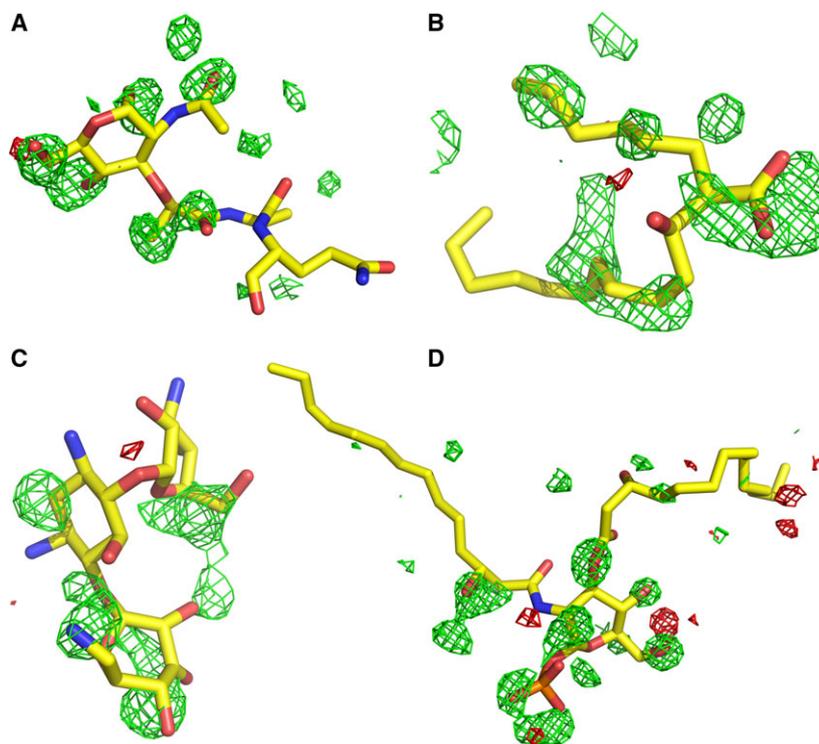


Fig. 3. Putative ligands of the ribosome-inactivating protein from *Momordica balsamina*. Ligand models in stick representation are superposed on $mF_o - DF_c$ omit difference electron density maps. Each map was calculated after 10 cycles of REFMAC5 [53] refinement using the diffraction data deposited in the PDB, with the ligand in question removed from the coordinate set. Each map was contoured at $\pm 3\sigma$ (green/red). The putative ligands are (A) peptidoglycan (PDB ID 4LWX), (B) mycolic acid (3U8F), (C) kanamycin (3U6T) and (D) lipopolysaccharide (4GUW). None of these PDB entries are associated with a publication.

Summary: Multiple models of the ribosome-inactivating protein from *Momordica balsamina* do not contain ligands. Numerous related models with nonexistent ligands (Table 2) contaminate the scientific literature and the PDB. When such cases are identified, the journals where the original work was published should be made aware of this fact and, in turn, request from the authors either corrections or full retractions. However, there seems to be no accepted way to deal with questionable structures that have been deposited in the PDB but are not associated with a primary publication. This aspect of the problem requires serious discussion by the community.

Discrepancies between electron density figures in publications and electron density maps generated from deposited data

Chandra *et al.* [64] describe the structure of a complex of phospholipase A₂ with a designed peptide Leu-Ala-Ile-Tyr-Ser (PDB ID 1JQ8). According to the authors, the peptide was identified as bound to only one of the two molecules of the enzyme in the asymmetric unit (A), whereas only a few water molecules were found in the corresponding locations in the second molecule (B). The presence of the peptide was inferred from the appearance of the $F_o - F_c$ difference Fourier map, supposedly calculated before any ligand modelling, and contoured at 2.5σ (Fig. 4A). As in the first example, the $mF_o - DF_c$ map downloaded from the EDS server, contoured at -2.5σ (Fig. 4B), shows a very comparable negative density, very clearly indicating that the diffraction data do not support the presence of the modelled ligand. After refinement, the B-factors of the peptide again climb to $\sim 150 \text{ \AA}^2$, indicating no scattering contribution from its stated position. The omit map calculated after 10 cycles of maximum-likelihood refinement with the pentapeptide deleted from the deposited coordinates to remove phase bias is shown in Fig. 4C. The structure was subsequently refined with only a few water molecules placed in the area originally occupied by the pentapeptide. It should be noted that although the final R_{free} of the re-refinement is higher than its counterpart reported in the original PDB file (Table 1), it is considerably lower than the corresponding value present in the validation report that actually accompanies the structure 1JQ8. Without exact knowledge of the underlying refinement parameterization, the absolute numbers of R -values are not comparable.

The resulting map is very different from the originally published one shown in Fig. 4A and does not support the presence or the claimed pose of the pentapeptide. A similar problem also affects the analogous coordinate set 3JTI, which does not have an associated publication. In addition to no evidence of electron density, per PDB report, the stereochemistry of the peptides is in the zeroth percentile for backbone torsion angles and side chain conformers. Difference electron density also indicates that the entire binding site in 3JTI is poorly modelled (not shown).

Summary: Published electron density maps for the pentapeptide ligand in phospholipase A₂ cannot be reconstructed from deposited data. The situation presented here is similar to the one described in the first example and possible reasons for the discrepancy of the published maps and the actual ones generated from data and model are given there. We did conduct model rebuilding and re-refinement to demonstrate that in similar cases the model refined without the ligand should be deposited in the PDB and a comment, linked to the original publication, should be published in an appropriate journal. Whether a retraction of the original publication should be initiated or an erratum published requires editorial involvement, and at present is an unresolved and contentious issue [41,65,66].

Lack of evidence, implausible geometry and abundant collisions, combined

36-65 is a germline antibody that has been used to evaluate the role of conformational flexibility in germline antibody recognition of haptens. Using a phage-displayed random peptide library and screening against antibody 36-65, [67] discovered three peptides that bound (as phage clones) to the Fab with low micromolar affinity. The crystal structure of the unliganded Fab (2A6J) and complexes with three different dodecapeptides (Kla, RII and Sig; PDB IDs 2A6I, 2A6D, and 2A6K, respectively) were then reported [42]. Unfortunately, there is no discernible electron density for a peptide molecule in any of the purported complex structures, as reported by EDS and TWILIGHT-2. We have re-refined one of these complexes (2A6I) with the Kla peptide coordinates removed (Table 1). Some minor changes included mutating the heavy chain residues ProH196 and ArgH197 to Thr and Trp, respectively, as supported by very clear electron density. Mouse IgG1 C_{H1} regions have been found with either

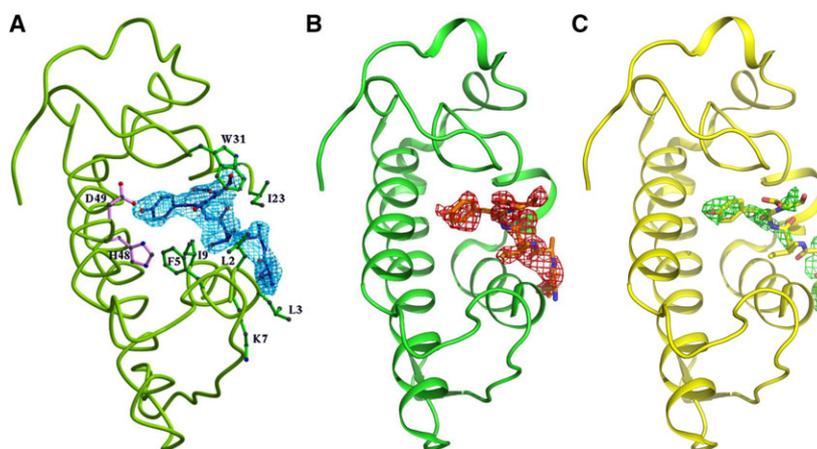


Fig. 4. A putative pentapeptide bound to phospholipase A₂. The backbone of the enzyme (PDB ID 1JQ8) is traced as a ribbon and the peptide is shown in stick representation. (A) A copy of fig. 1 of [64], showing the F_o-F_c map contoured at the 2.5σ level. (B) Negative mF_o-DF_c residual electron density based on a map downloaded from EDS and contoured at the -2.5σ level. (C) mF_o-DF_c omit map contoured at 2.5σ after the removal of the pentapeptide and 10 cycles of maximum-likelihood refinement. The original (1JQ8) and re-refined (5VET) models and maps can be interactively inspected using MOLSTACK at: <http://molstack.bioreproducibility.org/project/view/TOGXJSJ2A9GGGWHXVX0K/>.

Pro-Arg or Thr-Trp at these positions. A register-shift error in the complementarity determining region (CDR) L2 residues L51–L57 was also corrected. This register shift is not found in the search model (1JFQ; Fab 36–71 with anti-*p*-azophenylarsonate; [68]) used for molecular replacement. However, this error could easily be propagated into any structures using 2A6I as a molecular replacement probe. Additional residues missing in 2A6I were also modelled into weak density around H138–H140 and two additional residues were added at the C terminus of the heavy chain. The resulting Ramachandran plot [27] has no outliers compared to 21 found in the original structure, and all-atom clashscore (number of bad steric collisions remaining with a 0.4 \AA grace margin, per 1000 atoms) of 0 compared to 36 in the original structure. Because of the extensive manual corrections, the statistics for unsupervised automated refinement by PDB_REDO were significantly poorer (Table 3).

In the 2A6I crystal, the bottom of the C_{H1} region of one Fab packs closely against the antigen binding site of a neighbouring Fab, leaving very little space for any peptide ligand to fit. Indeed, the originally modelled peptide, with 67 atoms, has 69 bad steric clashes with neighbouring Fab atoms as reported by MOLPROBITY [27]. These collisions are with both the Fab molecule to which the peptide is purportedly bound, and with the neighbouring Fab (Fig. 5). There are also 26 bad steric clashes within the peptide itself. As the re-refined coordinates include two extra residues at the C-terminal end of the heavy chain, and additional residues previously missing in the H138–H140 segment, this situation worsens if the model peptide is placed into the rebuilt coordinates, resulting in 87 steric clashes with the Fab. In addition, seven out of the nine originally modelled peptide residues (P3–P9) are Ramachandran outliers. We can thus conclusively demonstrate that, based on steric clashes and the Ramachandran plot,

Table 3. Selected statistics comparing automated and manual re-refinement of PDB entry 2A6I. The manual improvement over the automatically refined PDB_REDO models is significant, emphasizing the frequent need of skilled human-expert ‘polishing’ of macromolecular structural models.

	Original deposit (2A6I)	PDB_REDO		PDB_REDO Optimized	PDB_REDO Manually rebuilt (5VGA)
		Calculated	PDB_REDO Conservative		
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%

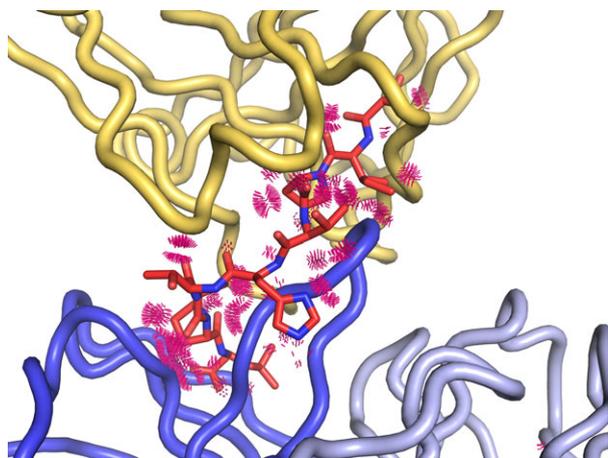


Fig. 5. Steric clashes around the dodecapeptide in the PDB 2A6I model. The peptide (red model) in the original 2A6I structure is shown bound to Fab 36–65 (light and dark blue C α trace) with the neighbouring Fab molecule shown in beige/brown. MOLPROBITY steric clashes are indicated by red spikes that are proportional to the severity of the clash. Clearly, there is not enough space between the Fab molecules to accommodate the purported peptide. The original (2A6I) and re-refined (5VGA) models and maps can be interactively inspected using MOLSTACK at: <http://molstack.bioreproducibility.org/project/view/DJ2PD0QFBGKCOUDVZKY/>.

the peptide as modelled by [42] (a) does not fit into the crystal lattice; (b) has physically impossible geometry; and (c) has no meaningful electron density. The two other peptide complexes, 2A6D and 2A6K, crystallize in a related crystal form with roughly the same *a* and *c* axes and the *b* axis doubled, and their crystal packing is very similar to that found in 2A6I. Therefore, these two crystal forms also have no space for a ligand as large as a dodecapeptide. The protein sample used for the 2A6I crystallization experiment contained the germline Fab 36–65 and 25-fold molar excess of the dodecapeptide (KLASIPTHTSPL). While the affinity of Fab for this peptide on phage was measured by surface plasmon resonance to be 0.12 μM (K_d) [69], the affinity for the peptide used in crystallization was not reported. Cocrystallization of proteins with low-affinity ligands is quite difficult, and interpretation of any weak electron density resulting from such experiments must be undertaken with extreme caution. So far, the deposited model and the related primary citation [42] have not been corrected, because the journal harbouring the primary publication has maintained that the indisputable absence of the peptide does not warrant retraction of the publication because the data have not been shown to be fabricated. This example illustrates the difficulty of removing clearly invalid models and correcting scientific literature.

Summary: The 36–65 germline antibody Fab fragment structure contains no dodecapeptide. The case of 2A6I illustrates the combination of absent evidence with vanishingly small prior probability, resulting in a practically zero posterior likelihood. The crystallographic model provides no support for the claim of a bound peptide. On the other hand, re-refinement of this structure in the unliganded state has yielded valuable information regarding corrections to the amino acid sequence of the protein and its CDR L2 conformation, and thus re-refinement and reposition are worthwhile. Also in this case, automated re-refinement could not completely correct the model and manual rebuilding was necessary.

Crystallographic resequencing, implausible metals and incorrectly assigned water molecules meet absent ligands

A paper by Jain *et al.* [45], further abbreviated J2016, describes the crystal structure of *Solanum melongena* vicilin, a protein with a known fold related to canavalin, phaseolin and other proteins isolated from seeds. The structure was determined by single-wavelength anomalous scattering of intrinsically present sulphur atoms (S-SAD) from data collected at 1.77 Å wavelength, whereas structure refinement used a data set measured to 1.5 Å resolution at 0.95 Å wavelength. The coordinates and structure factors corresponding to the final model were deposited in the PDB (5CAD), while the data set used for structure determination is not available. Our letter to the depositors requesting the original anomalous diffraction data has remained unanswered.

The vicilin used by J2016 was isolated from *S. melongena* seeds and its previously unknown amino acid sequence was determined as part of the published work. The methods used for sequence determination included (a) fragmentation of the protein with enzymes such as trypsin, chymotrypsin and Glu V8 protease, followed by chemical sequencing; (b) Edman degradation sequencing of the N-terminal residues and (c) mass-spectrometric analysis. Those chemical experiments, however, left several sequence positions unassigned. The identity of several unknown residues was divined from the electron density maps. The putative sequences were aligned with the purportedly related sequence of vicilin from *S. lycopersicum*, for which no crystal structure had been determined, and the results of the final sequence assignment were shown in fig. 1 of J2016. Surprisingly, however, the sequence shown

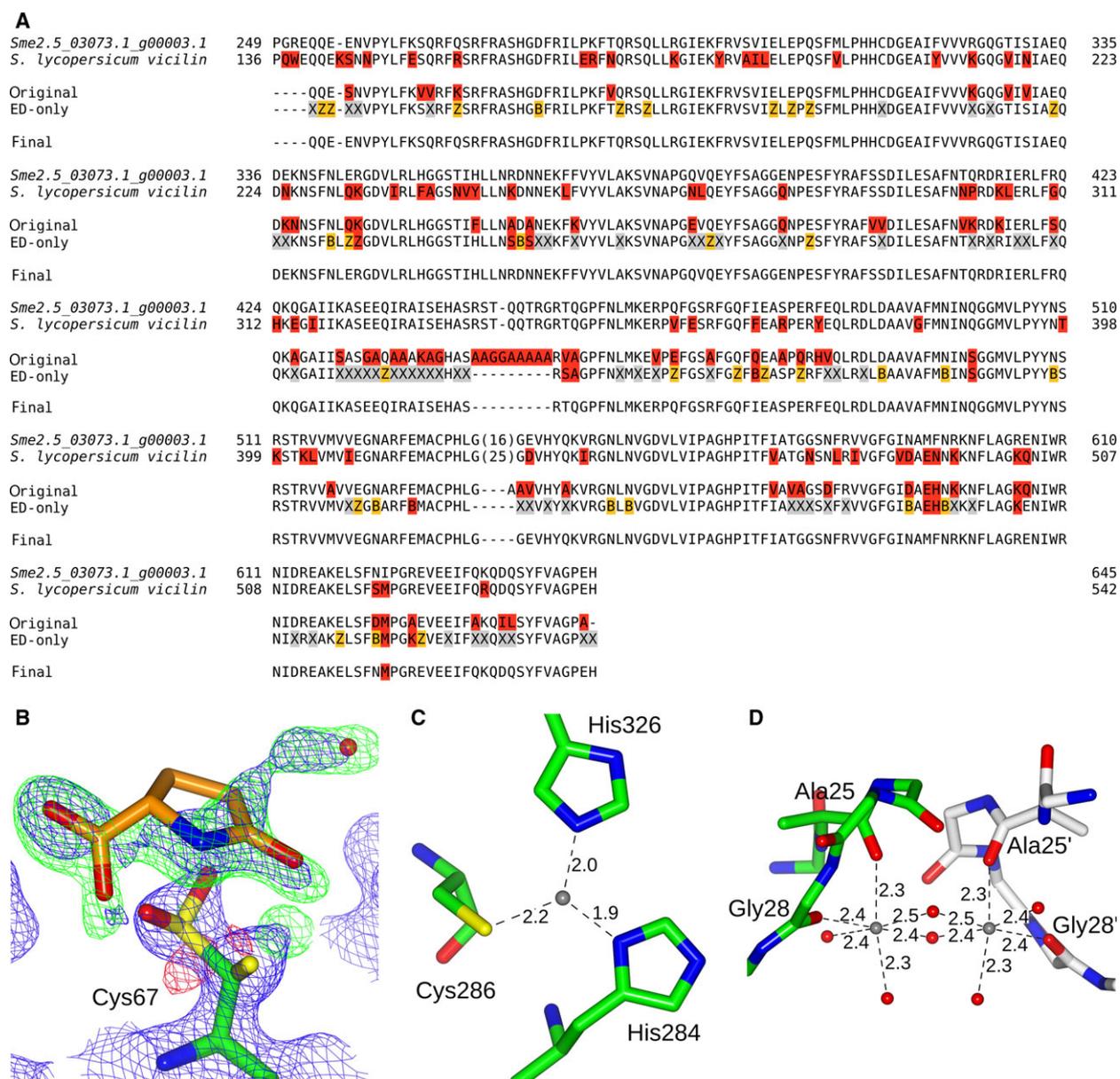


Fig. 6. Results of the re-refinement of the PDB entry 5CAD. (A) Comparison of the following sequences: predicted for vicilin from eggplant draft genome (*Sme2.5_03073.1_g00003.1*), predicted for vicilin from tomato (*S. lycopersicum*, Uniprot ID: B0JEU3), deposited as 5CAD (original), conservatively assigned according to electron density only (ED-only); with the final sequence of the re-refined structure, reconciled with the genomic data. The residues highlighted red are different from the eggplant vicilin sequence predicted from the draft genome; residues highlighted orange are ambiguous Asp/Asn (marked as B) or Glu/Gln (marked as Z) pairs from the crystal structure that agreed with genomic data; and residues highlighted grey could not be identified based on electron density alone (marked as X). (B) Omit electron density ($2mF_o - DF_c$ contoured at 1σ in blue, $mF_o - DF_c$ contoured at $\pm 3\sigma$ in green/red) for the pyroglutamate as calculated for the re-refined structure. The oxidized Cys67 from the re-refined structure is presented as green sticks and the pyroglutamate from the original deposit is shown as orange sticks. The discontinuous electron density does not support unambiguous assignment of any ligand in this place. (C) Coordination of a copper ion (grey sphere) modelled and validated in the re-refined structure as described in the text. (D) Coordination of two symmetry-related sodium ions (grey spheres) modelled in the re-refined structure in place of magnesium; the protein monomer from the asymmetric unit is shown in green and a symmetry-related copy in grey. In panels C and D, the interatomic distances in the final re-refined structure are in Å. The original (5CAD) and re-refined (5VF5) models and maps can be interactively inspected using MOLSTACK at: <http://molstack.bioreproducibility.org/project/view/DEM47929PNOILFSGS5SW/>.

as identified from the electron density (the first line in fig. 1b of J2016) is at several positions different from either the chemically determined sequence or from the sequence deposited with the coordinates in the PDB. For example, the sequence at the N terminus resulting from Edman degradation is PGREQQEENVPYLF, the sequence in fig. 1b of J2016 is —GGEGAVPYLF, and in the deposited coordinates —QQESNVPYLF.

Partial assignment of protein sequence from electron density maps is feasible even at medium resolution and from noisy maps [70–73], particularly when supported by independent prior knowledge, as exemplified by including Kabat propensities [74] in the assessment of the posterior likelihood of sequence assignments in antibody structures [33]. Electron density is often weak for side chains located at the protein surface or in partially disordered fragments, and some residues are almost impossible to be distinguished based on the electron density shape. Glutamine/glutamate, asparagine/aspartate or valine/threonine are indistinguishable based on electron density alone, although careful analysis of H-bonding interactions can often help. The identification can be further complicated by coexistence of multiple conformations, resembling a different residue in electron density. For example, a dual conformation serine will resemble a threonine. Some, but not all, of these cases (e.g. serine vs. valine vs. threonine) can be resolved by carefully considering intermolecular interactions and the B-factors. The unresolved ambiguities can be flagged by denoting them as unidentified residues (UNK), which is an accepted procedure in the PDB. An unresolved glutamine/glutamate or asparagine/aspartate ambiguity can be denoted as GLX or ASX.

In view of the discrepancies in the sequences published in J2016, we decided to carefully reanalyse the sequence solely based on electron density and reconcile it with the sequences from the draft eggplant genome [75] (Fig. 6A). We used BLAST searches provided by the authors of the eggplant genome (<http://eggplant.kazusa.or.jp>) to match the putative density-based protein sequence to the predicted protein sequences (ID: Sme2.5_03073.1_g00003.1). Figure 6A presents a comparison of the sequence of vicilin translated from the draft eggplant genome, of the original sequence assigned in 5CAD, of the re-refined putative electron density-based sequence, and of the reconciled sequence from our reinvestigation. Unfortunately, the authors of J2016 did not make their MS/MS data publicly available, providing only the sequence of the closest matching peptides from *S. lycopersicum*. It is, therefore, not possible to validate their results in the

context of genome sequencing. Our letter to the depositors requesting original data has remained unanswered.

Further comparisons of the deposited and re-refined models indicate that the electron density for loop 198–207 is unconvincing in the original map and the removal of this segment from refinement did not bring back any meaningful electron density. We thus consider the original tracing of the 198–207 segment as not supported by the available experimental data and omitted this segment from the re-refined model. This illustrates the situation when we know that some element of the structure must be present (there is no indication that the chain was proteolytically cleaved), but do not have convincing electron density for its modelling. Whether such missing elements should be still (somehow) modelled or left uninterpreted, is an open question. In the present case, we chose to build an ‘amputated’ model that is consistent with the experimental evidence. It should be noted, however, that this case of ‘we know it must be there’, related to near certainty of the continuity of the protein chain, is fundamentally different from the case of ‘we wish it were there’ illustrated by the absent ligands in our other examples.

The reduced diffraction data deposited in the PDB were collected at $\lambda = 0.95$ Å where the anomalous scattering of atoms relevant to this structure is $f''(\text{S}) = 0.22$ e, $f''(\text{Cu}) = 2.11$ e, $f''(\text{Na}) = 0.046$ e and $f''(\text{Mg}) = 0.067$ e, as estimated by the CCP4 program CROSSEC [76]. According to the text and supplementary table 1 of J2016, the structure contained only a single Mg^{2+} metal ion located in the intermolecular space and coordinated by two carbonyl oxygen and four water molecules, in a typical octahedral configuration. Surprisingly, the supplementary table 1 of J2016 claims the presence of an anomalous peak at the magnesium ion position, although, as shown above, the anomalous signal of Mg^{2+} should be close to zero. Indeed, no significant signal is found at this site in the anomalous difference map calculated based on the deposited data and the model refined by us. The distinctive octahedral coordination sphere (Fig. 6D) precludes the modelling of a water molecule at this position, but the observed metal-oxygen distances (2.3–2.5 Å) strongly suggest the isoelectronic sodium ion instead of magnesium [77]. The placement of Na^+ is consistent with the crystallization conditions, which included 1.5 M sodium malonate, but no magnesium ions. The final validation of metals in CHECKMYMETAL [28] significantly favoured assignment of this ion as Na^+ .

Conversely, fig. 4B of J2016 shows a site which in the closely related structure of adzuki bean 7S vicilin [78] contains a copper ion. Although the corresponding site in the 5CAD crystal of the *S. melongena* protein is virtually identical, it does not contain a metal. A water molecule (Wat501) placed at that position has a B-factor of 1.05 \AA^2 and its distances from the N δ 1 atom of His284 and N ϵ 2 of His320 are 2.0 \AA each. With His284 being poorly ordered, we placed a Cu^{2+} ion at half occupancy at this site and its refined B-factor matches almost exactly the B-factor of the well-ordered N ϵ 2 atom of His320 (Fig. 6C). This interpretation was confirmed by the presence of a $\sim 30\sigma$ peak in the anomalous difference map, at least five times higher than for the strongest sulphur peak, and copper assignment and coordination was validated using CHECKMYMETAL.

The electron density for one of the proposed anionic ligands of vicilin, namely an acetate ion, was completely unambiguous in the map calculated using the deposited data, but the geometry of the molecule in the 5CAD coordinate set was distorted and nonplanar, and some negative difference electron density appeared. Clearly, this was due to the application of incorrect restraints during the 5CAD refinement, since the corresponding difference omit map based on our re-refined coordinates has no such peaks and the molecule is perfectly planar.

The presence of the purported small-molecule ligand pyroglutamate is not supported by electron density (Fig. 6B). The ligand omit $mF_o - DF_c$ map in supplementary fig. 3b of J2016 is contoured at an unrealistically low 1.8σ level, and no density beyond what could be assigned as water molecules within an unclearly delineated hydrogen bond network is found in the recalculated map. The interpretation of the electron density in this region is further complicated by the sequence assignment. Although the sequence of the eggplant vicilin inferred from the draft genome has a cysteine at this position, the crystallographic data alone do not allow for a convincing identification of this residue. Moreover, cysteine residues in crystal structures can be observed in multiple oxidized forms, such as S-oxy cysteine, cysteinyl-S-sulphinic or -sulphonic acid. Cysteine can also be covalently modified or desulphurized by radiation damage [79]. The peaks that would correspond to a sulphur atom in the anomalous map calculated using the deposited data are at noise level, and this residue is present in more than one orientation. Based on the indication from the draft genome, we tentatively assigned it as a mixture of cysteine (in one orientation) and cysteinyl-S-sulphinic acid (in the second orientation). It cannot be ruled

out that the residual density, which was originally interpreted as pyroglutamate, covers part of some covalent modification of the cysteine residue. The presence of multiple oxidation states and/or modifications of the cysteine would change the hydrogen bond network and significantly contribute to the observed disorder in this area. The data available to us and the adjacent weak electron density do not allow an unambiguous interpretation of this region.

We were, however, able to identify an additional ligand that was not included in the original structure – a malonate ion present at a crystal contact between residues Ser97 and Arg367. The assignment of this density as a partially occupied malonate is consistent with the high concentration of malonate (1.5 M) in the reported crystallization conditions. The nature of the corrections necessary for this model, resulting from the inclusion of an additional source of information about the sequence, made it necessary to manually rebuild and ‘resequence’ the model. Comparison with a model automatically re-refined by PDB_REDO (Table 4) shows the importance of metadata for automatic reinterpretation of crystal structures. In this case, the incomplete sequence presented in the PDB entry header prevents automatic tools from improving the model.

Summary: The model of vicilin from *Solanum melongena* could be corrected in many aspects and was redeposited. The identification of the protein sequence, metal ions and ligands could be more reliable if all data sets were deposited either in the PDB as unmerged structure factors or, preferably, as raw diffraction images to repositories such as proteindiffraction.org [80]. The identification of ligands supposedly retained during protein purification should be carried out carefully, with support of strong experimental evidence.

Density from crystallization cocktail components misinterpreted as ligand density

Three structures of mouse kynurenine aminotransferase III (unliganded and complexed with glutamine or kynurenine) were determined at the resolution of 2.59 \AA (PDB ID 3E2F), 2.26 \AA (PDB ID 3E2Y) and 2.81 \AA (PDB ID 3E2Z), respectively [46]. None of them would have raised much concern in view of their acceptable validation parameters (which disregard, however, free amino acid ligands in EDS and MOLPROBITY validation analysis), were it not for the fact that the presence of glutamine, a ligand that had been soaked into the crystals, was proposed in the highest-resolution structure 3E2Y based on an $mF_o - DF_c$ omit map

Table 4. Selected statistics comparing automated and manual re-refinement of the PDB model 5CAD. The manual improvement over the automatically refined, PDB_REDO models results from better sequence assignment and emphasizes the importance of additional information during model building, refinement, and analysis.

	Original deposit (5CAD)	PDB_REDO		PDB_REDO	
		Calculated	PDB_REDO Conservative	Optimized	Manually rebuilt (5VF5)
<i>R</i>	0.199	0.191	0.146	0.146	0.131
<i>R</i> _{free}	0.210	0.200	0.175	0.175	0.163
Clashscore/Percentile	10.0/58th		9.1/65th	7.4/79th	3.4/97th
Ramachandran outliers	3/0.8%		2/0.5%	3/0.8%	0
Poor rotamers	7/2.4%		6/2.0%	6/2.0%	0

contoured at the extremely low 1.5σ level. The $mF_o - DF_c$ map downloaded from the EDS server [24] indicated, however, a strong positive peak near the putative carboxylate of the glutamine molecule, as well as a negative peak nearby (Fig. 7A). An omit map based on the original PDB model with additional 10 cycles of refinement carried out after removal of the glutamine ligand, and contoured at the 1.5σ level, indeed showed an extensive electron density at the position proposed for the glutamine, but additionally it showed a 10σ peak within that area (Fig. 7B). Subsequent re-refinement of the structure indicated,

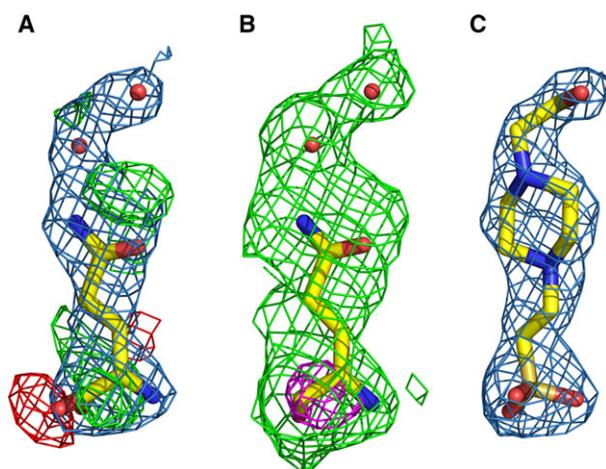


Fig. 7. A putative glutamine ligand bound to kynurenine aminotransferase III. (A) $2mF_o - DF_c$ map (blue) contoured at 1σ , and $mF_o - DF_c$ map contoured at $\pm 3\sigma$ (green/red) for the putative glutamine modelled as bound to the enzyme in the PDB deposit 3E2Y. (B) $mF_o - DF_c$ difference Fourier omit map calculated after 10 cycles of refinement of the original coordinates with the marked glutamine (sticks) and two water molecules (red spheres) removed, contoured at 1.5σ (green) and 10σ (magenta). (C) A HEPES molecule modelled and refined at the same site, with the superposed maps contoured as in panel A. The original (3ef2, 3E2Y and 3E2Z) and re-refined (5VEP, 5VEQ and 5VER) models and maps can be visually inspected at: <http://molstack.bioreproducibility.org/project/view/U5IGUUW0BNKLBE8P1L07/>.

without any doubt, that the purported glutamine was a HEPES molecule from the crystallization buffer (Fig. 7C).

Additionally, we located four Ca^{2+} ions, also present in the crystallization solution, as well as several molecules of glycerol and polyethylene glycol. Re-refinement of the putative complex with kynurenine yielded a model that was virtually identical to that corresponding to the ‘glutamine complex’. Whereas a HEPES molecule was also clearly present in the unliganded structure (its electron density had been modelled as two glycerol molecules), only the sulphate group of the HEPES molecule was located adjacent to the guanidinium moiety of Arg430, while the rest of the molecule was rotated $\sim 45^\circ$ compared to its orientation in the two other structures. The change of HEPES orientation required a considerable rearrangement of several active site residues (Asn52, Trp54, Phe57, Gln71 and Tyr312). Several trivial errors were also corrected; for example, all three coordinate sets had chirality errors of some residues. An unfortunate inconsistency between the three original structures was their presentation in different parts of the unit cell, making comparisons far from straightforward, especially for nonspecialists. The inconsistency was remediated during the re-refinement by placing all models in the same location, as suggested by the ACHESYM server [81].

The cases presented in this section illustrate several mistakes that are easy to avoid, but which otherwise may lead to significant difficulties. First of all, the risk of interpreting omit maps (or any electron density maps, for that matter) contoured at unreasonably low level leads to overinterpretation of results. Whereas such maps may indeed suggest the presence of a ligand, they must be considered with extreme caution and their interpretation must not be biased by the assumption that the electron density at the expected binding site is always that of the desired ligand. Secondly, if high difference density peaks are present after

refinement, they could represent ions rather than water molecules. Similarly, connected patches of electron density assigned as water molecules might indicate the presence of other solvent or buffer molecules. Finally, these three structures raise a perplexing question, impossible for us to answer, namely how soaking glutamine or kynurenine into the apo crystals could change the vicinity of the active site, when none of these ligands could be detected in these structures.

Two other structures originating from the same laboratory, of kynurenine aminotransferase from the mosquito *Aedes aegypti*, in the PLP and PMP forms (PDB codes 1YIZ and 1YIY, respectively) are also available [82]. The 1YIZ model disagrees with the electron density maps from the EDS server in several places. The asymmetric unit of 1YIZ contains two protein molecules as chains A and B, each consisting of 418 residues numbered from 12 to 429. The EDS maps calculated at the resolution of 1.55 Å very clearly showed that the fragments 12–24 and 352–357 of both molecules had been placed wrongly, and unambiguously showed the correct tracing. In addition, there were 138 residues with wrong side chain rotamers, constituting 16.5% of all residues or ~ 20% of non-Gly, non-Ala residues (which do not have rotamers). We have reinterpreted the 1YIZ structure and re-refined it with REFMAC5 to R/R_{free} values of 0.185/0.215, whereas the corresponding values reported in the PDB are 0.254/0.279 (Table 1). As shown in Table 5, automated refinement with PDB_REDO led to significant improvement, especially in the optimized mode, but it was not able to correct the severe tracing errors of the main chain.

The EDS maps for the 1YIY model indicate the same problems, confirming that neither of these models was carefully checked against the electron density maps. The 1YIY model has not been re-refined by us. The original and re-refined models and maps can be interactively inspected using MOLSTACK at: <http://molstack.bioreproducibility.org/project/view/KQK1P9AJV57TIM4MCD1M/>

Summary: Manual model rebuilding and refinement is still necessary. Some poorly refined PDB models can be improved by automated re-refinement such as PDB_REDO (Table 5). However, to correct major tracing errors that lie outside the convergence radius of the automated refinement, visual inspection of the electron density maps followed by manual model rebuilding is almost always necessary. In addition, contouring electron density at noise level almost always indicates problems with the model. In cases where the re-refined structures contradict the conclusions of the originally published paper, it is necessary to alert the authors, redeposit (preferably jointly) the structure, and inform the journal where it was reported.

High resolution electron density suggests a different ligand

Several high-resolution structures of antibody fragment 2D10 complexed with different ligands have been determined, but the complex with α -1,6-mannobiose (PDB ID 5I4F, 1.55 Å) is the sole subject of a recent publication that described it in considerable detail [48]. Although the reported global refinement statistics do not indicate any obvious problems ($R = 0.167$, $R_{\text{free}} = 0.193$), even a cursory look at the electron density map reveals that the carbohydrate ligands do not fit it well (Fig. 8A). The asymmetric α -1,6-mannobiose disaccharide was modelled bound to the antibody at two sites, in two overlapping 0.5-occupancy orientations at each site. One of these sites is on a general position and consists of two independent half-occupancy moieties. The other site is located on a crystallographic dyad, which automatically generates the alternative orientation. However, the presence of both positive and negative peaks in the $mF_o - DF_c$ map suggests that the disaccharide has been misidentified. Indeed, the fully symmetric trehalose fits the electron

Table 5. Selected refinement statistics for the PDB entry 1YIZ. Provided are statistics for the published, automatically refined, and manually rebuilt/refined models. PDB_REDO did not use the same set of R_{free} reflections as the original refinement, but the refinement of the manually rebuilt model did. The values of clashscore, Ramachandran outliers and poor rotamers were reported by the MolProbity server. The improvements in global refinement statistics and geometry over the original model are significant.

	Original deposit (1YIZ)	PDB_REDO		PDB_REDO	
		Calculated	PDB_REDO Conservative	Optimized	Manually rebuilt (5VEH)
R	0.254	0.278	0.237	0.217	0.185
R_{free}	0.279	0.282	0.263	0.236	0.215
Clashscore/percentile	28.7/3rd		14.7/32 nd	3.0/98th	4.3/96th
Ramachandran outliers	18/2.18%		14/1.69%	18/2.18%	1/0.13%
Poor rotamers	144/20.3%		110/15.5%	19/2.7%	0

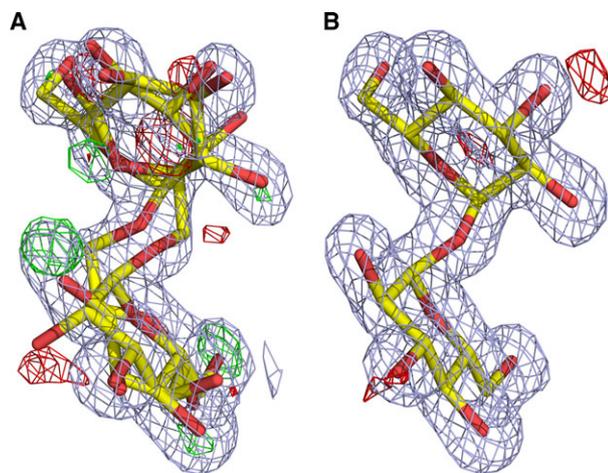


Fig. 8. Model and electron density map of trehalose vs α -1,6-mannobiose in the PDB entry 5I4F. (A) The original α -1,6-mannobiose, modelled in two overlapping general-position orientations. (B) Trehalose bound in single orientation at the same location in the reinterpreted electron density of the re-refined model. The $2mF_o - DF_c$ map (grey) is contoured at the 1σ level, and the $mF_o - DF_c$ map (green/red) at $\pm 3\sigma$. The differences between 5I4F and re-refined structure 5VF2 can be examined interactively at <http://molstack.bioreproducibility.org/project/view/MPYO83KA6I78W8HZS1C0/>.

density in both locations much better than α -1,6-mannobiose, with no residual density remaining after refinement (Fig. 8B). A molecule from the crystallization buffer (MES) and an Mg^{2+} cation could also be placed elsewhere in the electron density with a high degree of confidence. Whereas several other connected densities suggest possible presence of some additional ligands, their identity could not be ascertained as their shapes do not correspond to any components of the crystallization solutions, and we simply marked them in the re-refined model (PDB ID 5VF2) using unknown atoms (UNX with seven electrons) as placeholders at the electron density peaks.

We cannot offer any explanation why the principal ligand was misidentified. However, clear problems with electron density fit of the modelled ligand should have alerted the authors to this problem and encouraged them to trace back possible mistakes. Instead, this clear misinterpretation did not prevent them from discussing in great detail the interactions of the protein with the misidentified carbohydrate, which are substantially different from the real ones (see fig. 2 of [48]). The presented disaccharide binding energy calculations for the antibody are therefore misleading, if not meaningless, in view of the misidentified ligand.

Summary: High resolution is no safeguard against overmodelling. Authors sometimes expend considerable effort trying to model a desired ligand or components expected to be present in the crystallization media while disregarding clear disagreement with electron density. If the wrong ligand is modelled, the discussion of its interactions, even if followed by sophisticated theoretical calculations, is meaningless and this conclusion needs to be communicated to the journal that published the original paper. At present, no validation program can detect such errors and manual rebuilding and re-refinement after careful inspection of the electron density maps are necessary. When electron density suggestive of an unknown chemical entity is present, it is best to mark the atoms as UNX.

Concluding remarks and suggestions

The Protein Data Bank is the most important repository of structure model and data. Maintaining its high standard is exceedingly important. Quality and even veracity problems related to individual entries can mislead scientists that rely on the correctness of not only the deposited atomic models themselves, but also on the validity of their interpretation in the resulting publications. Contamination of this database with ‘bad apples’ will bias any meta-analyses based on all structures in the PDB or on a selected subset, as has been frequently discussed in the past, for example in the context of using protein structures for drug design [2,40,83]. The assumption that the structures in the PDB are correct in all details may also lead to problems such as the discovery of unusual antigen-antibody complexes [42,84,85] or ‘novel’ Zn-binding sites [86], although in the latter case misinterpretation of the PDB data was the main offense [87] and the authors of the original publication have already acknowledged that [88]. Thus, constant watchfulness and awareness of potentially flawed models must become part of the way of conduct. At the same time, improvement of outlier detection, development of methods for their correction or elimination, and evaluation of their impact on the associated publications, are never-ending tasks.

The ‘Appeal to normalcy’

Efforts at correcting the record should be undertaken, and responded to, in good faith, with the benefit of science in mind, and not on a personal level. Occasionally, however, the critiqued authors respond by

pointing out that similar procedures and results were reported before, sometimes by their critics. For example, table 1 of [47] lists a dozen structures that were accompanied by publications in which F_o-F_c omit maps showing putative ligand density were contoured at levels lower than 3σ . We analysed the six structures with maps contoured at 1.5σ and concluded that such low level was completely unnecessary for 3R6R [89] and 1UJJ [90], since appropriate contour levels show unambiguously the presence of modelled ligands. Contouring maps on a very low level may raise some unnecessary concerns, while on the other hand, analysis of the low-level density in 3E2Y and 3E2Z [46] led us to question and re-refine the structures of kynurenine aminotransferase [46]. Detachment from context and failure to properly analyse the models only emphasize the need for careful examination of each individual case instead of superficially relying on statistical metrics. We have pointed out the need for contextual evaluation already in our own analysis of PDB metrics [4,22]. The contradiction between claim stated and evidence required is particularly concerning, for example, when the stringent necessity of evidence for an immunogenic peptide in a complex structure [42] is presented on par with modelling of unknown ambiguous solvent density [47].

Context should guide corrective action

In several of the examined PDB structures also the protein component had gross deficiencies, even for models refined with high resolution data. An example is provided by several structures of kynurenine aminotransferase from the mosquito *Aedes aegypti* [82] in which parts of the polypeptide chain were clearly mistraced and which included a very large number of unlikely rotamers of the side chains. The hazard of such models is that they may be later used as molecular replacement probes, leading to further, practically uncontrolled, propagation of errors. In this case, we re-refined and redeposited one of the models (1YIZ), but we are aware that other related structures, such as 1YIY, may also be defective. Since the problematic parts of these structures do not extend to the vicinity of the active site, we suggest that it would be highly desirable if the authors of the original deposits revisited the structures in question, re-refined and redeposited them, as well as submit an erratum to the journal in which the original results were presented. We do not feel, however, that in cases like these, full retraction is necessary.

A much more serious situation is encountered when questionable structures form the basis of an extensive

analysis focused exactly on the interpretation of the questionable fragments. Several of the cases described herein fall into this category. For example, misidentified ligands of mouse kynurenine aminotransferase III [46] make further interpretations of their mode of binding to the enzyme moot. Similarly, basing a detailed description of the mode of binding of a misidentified carbohydrate and following it by a sophisticated computational analysis [48] is simply meaningless and not acceptable. In our opinion, papers based on fundamentally flawed premises should be retracted or at least followed by an extensive erratum describing the results of a follow-up study.

Many problems could be avoided if all isomorphous structures were systematically placed in the same location in the unit cell, a task greatly simplified by the existing standardization rules [81] and server (<http://ac.hesym.ibch.poznan.pl/>). Whereas atomic coordinates residing in different locations may be easily superimposed, superposition of the corresponding electron density maps is not as straightforward. Thus, the fact that the electron density for the putative glutamine and for kynurenine in the coordinate sets 3E2Y and 3E2Z was virtually identical, may have escaped the notice of even the original authors. Some isomorphous structures are submitted to the PDB in so many ways as to completely stymie any attempts at their comparison. As an illustration, 60 fully isomorphous structures of the same ribosome inactivating protein from *Momordica balsamina* are presented in the PDB in 11 different ways. Involvement of the PDB in better standardization of the coordinate sets is necessary to improve this Babylonian multitude.

Those 60 structures of the same protein complexed with different ligands bring up a number of additional points. Only 12 of them are mentioned in the scientific literature [61–63], whereas the remaining ones have not been described in any publication. In this case, contextual evaluation of unpublished complexes with ligands is virtually impossible.

Better policies for model replacement

Since several structures discussed here have now been re-refined by us and the new coordinates have been deposited in the PDB, we are aware of the difficulty in tracking such corrections. According to the current PDB policies, obsolete structures are linked to the deposits that supersede them. However, while deposits correcting structures that remain in the PDB are back-referenced, there is no indication whatsoever in the original (i.e. uncorrected) entry that a potentially more correct interpretation has been deposited. We propose

that in such cases a note be placed in the original deposit that a different version is also available. The lack of such cross-referencing is glaringly visible even in the example in the PDB deposition instructions that refer to the original model 1T3N [91], re-refined and redeposited as 1ZET by a different author [92]. Whereas a back-reference to 1T3N is present in the 1ZET deposit, there is no forward-reference in the previous deposit that could alert the users to a potential problem with it. Additionally, we strongly recommend that any re-refined data set, whether replacing an obsolete entry or correcting one that remains in the PDB, should carry a REMARK record clearly stating why the structure was re-refined and what has changed compared to the previous deposit. In cases where a re-refinement does not affect the claims of the associated publications, a simple process of replacing the original entries per authors' request, for example with PDB_REDO refinements, could be implemented. However, proliferation of poorly linked models in the PDB based on the same diffraction data should be avoided. It is too early to judge the success of efforts to achieve this goal by using, for example, deposit version numbers.

Acknowledgements

This work was funded by the Austrian Science Fund (FWF) under project P28395-B26, by the Polish National Science Centre (NCN) through grant No. 2013/10/M/NZ1/00251, by the Intramural Research Program of the National Institutes of Health (NIH), National Cancer Institute, Center for Cancer Research, and by NIH grants U01HG008424, R01GM117080, R01GM117325.

Author contributions

All authors jointly contributed to data mining, example selection, model refinement, analysis, and manuscript preparation.

References

- Dutta S, Burkhardt K, Swaminathan GJ, Kosada T, Henrick K, Nakamura H & Berman HM (2008) Data deposition and annotation at the Worldwide Protein Data Bank. In *Structural Proteomics: High-Throughput Methods* (Kobe B, Guss M & Huber T, eds), pp. 81–101. Humana Press/Springer, New York, NY.
- Minor W, Dauter Z, Helliwell JR, Jaskolski M & Wlodawer A (2016) Safeguarding structural data repositories against bad apples. *Structure* **24**, 216–220.
- Weichenberger CX, Pozharski E & Rupp B (2013) Visualizing ligand molecules in Twilight electron density. *Acta Crystallogr* **F69**, 195–200.
- Weichenberger C, Pozharski E & Rupp B (2017) Twilight reloaded: the peptide experience. *Acta Crystallogr* **D73**, 211–222.
- Cereto-Massague A, Ojeda MJ, Joosten RP, Valls C, Mulero M, Salvado MJ, Arola-Arnal A, Arola L, Garcia-Vallve S & Pujadas G (2013) The good, the bad and the dubious: VHELIBS, a validation helper for ligands and binding sites. *J Cheminform* **5**, 36.
- Koehler JJ (1993) The influence of prior beliefs on scientific judgments of evidence quality. *Organ Behav Hum Decis Process* **56**, 28–55.
- Simmons JP, Nelson LD & Simonsohn U (2011) False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci* **22**, 1359–1366.
- Bacon F (1620) *Novum Organum Scientiarum; Partis Secundae Summa, Digesta in Aphorismos*, Aphorismus XLIX.
- Bayes T (1763) An essay towards solving a problem in the doctrine of chances. *Phil Trans Roy Soc* **53**, 370–418.
- Laplace PS (1814) *Essai philosophique sur les probabilités*. Paris Bachelier, Paris.
- Popper K (1982) *Logic der Forschung*. J.C.B. Mohr, Tuebingen.
- Popper K (2002) *The Logic of Scientific Discovery*, 14th, printing edn. Routledge, New York, NY.
- Bricogne G (1974) Geometric sources of redundancy in intensity data and their use for phase determination. *Acta Crystallogr A* **30**, 395–405.
- French S & Wilson K (1978) On the treatment of negative intensity observations. *Acta Crystallogr A* **34**, 517–525.
- Main P (1979) A theoretical comparison of the α , γ and $2F_o - F_c$ syntheses. *Acta Crystallogr* **35**, 779–785.
- Read RJ (1986) Improved Fourier coefficients for maps using phases from partial structures with errors. *Acta Crystallogr A* **42**, 140–149.
- Bricogne G (1997) Bayesian statistical viewpoint on structure determination: basic concepts and examples. *Methods Enzymol* **276**, 361–423.
- Otwinowski Z (1991) Maximum Likelihood refinement of heavy atom parameters. In *Proceedings of the 1991 CCP4 study weekend* (Wolf W, Evans PR & Leslie AGW, eds), pp. 80–86. CLRC Daresbury Laboratory, Warrington, UK.
- Brändén CI & Jones TA (1990) Between objectivity and subjectivity. *Nature* **343**, 687–689.
- Kleywegt GJ & Jones TA (1995) Where freedom is given, liberties are taken. *Structure* **3**, 535–540.
- Rupp B (2010) Scientific inquiry, inference and critical reasoning in the macromolecular crystallography curriculum. *J Appl Crystallogr* **43**, 1242–1249.

- 22 Pozharski E, Weichenberger CX & Rupp B (2013) Techniques, tools and best practices for ligand electron-density analysis and results from their application to deposited crystal structures. *Acta Crystallogr D* **69**, 150–167.
- 23 Muller Y (2013) Unexpected features in the Protein Data Bank entries 3qd1 and 4i8e: the structural description of the binding of the serine-rich repeat adhesin GspB to host cell carbohydrate receptor is not a solved issue. *Acta Crystallogr F* **69**, 1071–1076.
- 24 Kleywegt GJ, Harris MR, Zou J-Y, Taylor TC, Wahlby A & Jones TA (2004) The Uppsala electron-density server. *Acta Crystallogr D* **60**, 2240–2249.
- 25 Tickle IJ (2012) Statistical quality indicators for electron-density maps. *Acta Crystallogr D* **68**, 454–467.
- 26 Laskowski RA, MacArthur MW, Moss DS & Thornton JM (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Crystallogr* **26**, 283–291.
- 27 Chen VB, Arendall WB 3rd, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS & Richardson DC (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D* **66**, 12–21.
- 28 Zheng H, Chordia MD, Cooper DR, Chruszcz M, Muller P, Sheldrick GM & Minor W (2014) Validation of metal-binding sites in macromolecular structures with the CheckMyMetal web server. *Nature Protoc* **9**, 156–170.
- 29 Badger J (2003) An evaluation of automated model-building procedures for protein crystallography. *Acta Crystallogr D* **59**, 823–827.
- 30 Deller MC & Rupp B (2015) Models of protein–ligand crystal structures: trust, but verify. *J Comp Aided Mol Des* **29**, 817–836.
- 31 Adams Paul D, Aertgeerts K, Bauer C, Bell Jeffrey A, Berman Helen M, Bhat Talapady N, Blaney Jeff M, Bolton E, Bricogne G, Brown D *et al.* (2016) Outcome of the first wwPDB/CCDC/D3R ligand validation workshop. *Structure* **24**, 502–508.
- 32 Read RJ & Kleywegt GJ (2009) Case-controlled structure validation. *Acta Crystallogr D* **65**, 140–147.
- 33 Naschberger A, Fürnrohr B, Lenac Rovis T, Malic S, Scheffzek K, Dieplinger H & Rupp B (2016) The N14 anti-afamin antibody Fab: a rare VL1 CDR glycosylation, crystallographic re-sequencing, molecular plasticity, and conservative versus enthusiastic modelling. *Acta Crystallogr D* **72**, 1267–1280.
- 34 Majorek KA, Kuhn ML, Chruszcz M, Anderson WF & Minor W (2014) Double trouble-buffer selection and His-tag presence may be responsible for nonreproducibility of biomedical experiments. *Protein Sci* **23**, 1359–1368.
- 35 Dym O, Song W, Felder C, Roth E, Shnyrov V, Ashani Y, Xu Y, Joosten RP, Weiner L, Sussman JL *et al.* (2016) The impact of crystallization conditions on structure-based drug design: a case study on the methylene blue/acetylcholinesterase complex. *Protein Sci* **25**, 1096–1114.
- 36 Kleywegt G (2007) Crystallographic refinement of ligand complexes. *Acta Crystallogr D* **63**, 94–100.
- 37 Read RJ, Adams PD, Arendall WB 3rd, Brunger AT, Emsley P, Joosten RP, Kleywegt GJ, Krissinel EB, Lutteke T, Otwinowski Z *et al.* (2011) A new generation of crystallographic validation tools for the protein data bank. *Structure* **19**, 1395–1412.
- 38 Smart OS, Womack TO, Flensburg C, Keller P, Paciorek W, Sharff A, Vornrhein C & Bricogne G (2011) Better ligand representation in BUSTER protein-complex structure determination. *Acta Crystallogr A* **67**, C134.
- 39 Liebschner D, Afonine PV, Moriarty NW, Poon BK, Sobolev OV, Terwilliger TC & Adams PD (2017) Polder maps: improving OMIT maps by excluding bulk solvent. *Acta Crystallogr D* **73**, 148–157.
- 40 Dauter Z, Wlodawer A, Minor W, Jaskolski M & Rupp B (2014) Avoidable errors in deposited macromolecular structures: an impediment to efficient data mining. *IUCr J* **1**, 179–193.
- 41 Rupp B, Wlodawer A, Minor W, Helliwell JR & Jaskolski M (2016) Correcting the record of structural publications requires joint effort of the community and journal editors. *FEBS J* **283**, 4452–4457.
- 42 Sethi DK, Agarwal A, Manivel V, Rao KV & Salunke DM (2006) Differential epitope positioning within the germline antibody paratope enhances promiscuity in the primary immune response. *Immunity* **24**, 429–438.
- 43 Joosten RP, Long F, Murshudov GN & Perrakis A (2014) The PDB_REDO server for macromolecular structure model optimization. *IUCr J* **1**, 213–220.
- 44 Svobodova Varekova R, Horsky V, Sehnal D, Bendova V, Pravda L & Koca J (2017) Quo Vadis. Biomacromolecular structure quality. *Biophysical J* **112**, 346a–347a.
- 45 Jain A, Kumar A & Salunke DM (2016) Crystal structure of the vicilin from *Solanum melongena* reveals existence of different anionic ligands in structurally similar pockets. *Sci Rep* **6**, 23600.
- 46 Han Q, Robinson H, Cai T, Tagle DA & Li J (2009) Biochemical and structural properties of mouse kynurenine aminotransferase III. *Mol Cell Biol* **29**, 784–793.
- 47 Salunke DM & Nair DT (2017) Macromolecular structures: quality assessment and biological interpretation. *IUBMB Life* **69**, 563–571.
- 48 Vashisht S, Kumar A, Kaur KJ & Salunke DM (2016) Antibodies can exploit molecular crowding to bind new

- antigens at noncanonical paratope positions. *Chem Select* **1**, 6287–6292.
- 49 Emsley P & Cowtan K (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr D* **60**, 2126–2132.
- 50 Emsley P, Lohkamp B, Scott WG & Cowtan K (2010) Features and development of Coot. *Acta Crystallogr D* **66**, 486–501.
- 51 Minor W, Cymborowski M, Otwinowski Z & Chruszcz M (2006) HKL-3000: the integration of data reduction and structure solution—from diffraction images to an initial model in minutes. *Acta Crystallogr D Biol Crystallogr* **62**, 859–866.
- 52 Murshudov GN, Vagin AA & Dodson ED (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D* **53**, 240–255.
- 53 Murshudov GN, Skubak P, Lebedev AA, Pannu NS, Steiner RA, Nicholls RA, Winn MD, Long F & Vagin AA (2011) REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr D* **67**, 355–367.
- 54 Brunger AT (1992) Free R value: A novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355**, 472–475.
- 55 Afonine PV, Grosse-Kunstleve RW, Echols N, Headd JJ, Moriarty NW, Mustyakimov M, Terwilliger TC, Urzhumtsev A, Zwart PH & Adams PD (2012) Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr D* **68**, 352–367.
- 56 Rupp B (2009) *Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology*, 1st edn. Garland Science, New York.
- 57 Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, Murray LW, Arendall WB 3rd, Snoeyink J, Richardson JS *et al.* (2007) MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucl Acids Res* **35**, W375–W383.
- 58 Porebski PJ, Sroka P, Zheng H, Cooper DR & Minor W. Molstack—Interactive visualization tool for presentation, interpretation, and validation of macromolecules and electron density maps. *Protein Sci*, in press, <https://doi.org/10.1002/pro.3272>.
- 59 Mir R, Singh N, Vikram G, Kumar RP, Sinha M, Bhushan A, Kaur P, Srinivasan A, Sharma S & Singh TP (2009) The structural basis for the prevention of nonsteroidal antiinflammatory drug-induced gastrointestinal tract damage by the C-lobe of bovine colostrum lactoferrin. *Biophys J* **97**, 3178–3186.
- 60 Ren J, Wang Y, Dong Y & Stuart DI (1994) The N-glycosidase mechanism of ribosome-inactivating proteins implied by crystal structures of alpha-momorcharin. *Structure* **2**, 7–16.
- 61 Kushwaha GS, Pandey N, Sinha M, Singh SB, Kaur P, Sharma S & Singh TP (2012) Crystal structures of a type-1 ribosome inactivating protein from *Momordica balsamina* in the bound and unbound states. *Biochim Biophys Acta* **1824**, 679–691.
- 62 Kushwaha GS, Yamini S, Kumar M, Sinha M, Kaur P, Sharma S & Singh TP (2013) First structural evidence of sequestration of mRNA cap structures by type 1 ribosome inactivating protein from *Momordica balsamina*. *Proteins* **81**, 896–905.
- 63 Yamini S, Pandey SN, Kaur P, Sharma S & Singh TP (2015) Binding and structural studies of the complexes of type 1 ribosome inactivating protein from *Momordica balsamina* with cytosine, cytidine, and cytidine diphosphate. *Biochem Biophys Rep* **4**, 134–140.
- 64 Chandra V, Jasti J, Kaur P, Dey S, Srinivasan A, Betzel C & Singh TP (2002) Design of specific peptide inhibitors of phospholipase A2: structure of a complex formed between Russell's viper phospholipase A2 and a designed peptide Leu-Ala-Ile-Tyr-Ser (LAIYS). *Acta Crystallogr D Biol Crystallogr* **58**, 1813–1819.
- 65 Fink PJ (2016) Comments from the Editor-in-Chief. *J Immunol* **196**, 521.
- 66 Rupp B (2016) Only seeing is believing: the power of evidence and reason. *Adv Biochemistry* **62**, 250–256.
- 67 Manivel V, Sahoo NC, Salunke DM & Rao KV (2000) Maturation of an antibody response is governed by modulations in flexibility of the antigen-combining site. *Immunity* **13**, 611–620.
- 68 Parhami-Seren B, Viswanathan M, Strong RK & Margolies MN (2001) Structural analysis of mutants of high-affinity and low-affinity p-azophenylarsonate-specific antibodies generated by alanine scanning of heavy chain complementarity-determining region 2. *J Immunol* **167**, 5129–5135.
- 69 Manivel V, Bayiroglu F, Siddiqui Z, Salunke DM & Rao KV (2002) The primary antibody repertoire represents a linked network of degenerate antigen specificities. *J Immunol* **169**, 888–897.
- 70 Pietrzyk AJ, Panjekar S, Bujacz A, Mueller-Dieckmann J, Lochynska M, Jaskolski M & Bujacz G (2012) High-resolution structure of Bombyx mori lipoprotein 7: crystallographic determination of the identity of the protein and its potential role in detoxification. *Acta Crystallogr D Biol Crystallogr* **68**, 1140–1151.
- 71 Pietrzyk AJ, Bujacz A, Mueller-Dieckmann J, Lochynska M, Jaskolski M & Bujacz G (2013) Crystallographic identification of an unexpected protein complex in silkworm hemolymph. *Acta Crystallogr D Biol Crystallogr* **69**, 2353–2364.
- 72 Niedzialkowska E, Gasiorowska O, Handing KB, Majorek KA, Porebski PJ, Shabalin IG, Zasadzinska E, Cymborowski M & Minor W (2016) Protein purification and crystallization artifacts: the tale usually not told. *Protein Sci* **25**, 720–733.

- 73 Porebski PJ, Cymborowski M, Pasenkiewicz-Gierula M & Minor W (2016) Fitmunk: improving protein structures by accurate, automatic modeling of side-chain conformations. *Acta Crystallogr D Struct Biol* **72**, 266–280.
- 74 Martin AC (1996) Accessing the Kabat antibody sequence database by computer. *Proteins* **25**, 130–133.
- 75 Hirakawa H, Shirasawa K, Miyatake K, Nunome T, Negoro S, Ohyama A, Yamaguchi H, Sato S, Isobe S, Tabata S *et al.* (2014) Draft genome sequence of eggplant (*Solanum melongena* L.): the representative solanum species indigenous to the old world. *DNA Res* **21**, 649–660.
- 76 Cromer DT (1983) Calculation of anomalous scattering factors at arbitrary wavelengths. *J Appl Crystallogr* **16**, 437–438.
- 77 Zheng H, Cooper DR, Porebski PJ, Shabalin IG, Handling KB & Minor W (2017) CheckMyMetal: a macromolecular metal-binding validation tool. *Acta Crystallogr D Struct Biol* **73**, 223–233.
- 78 Fukuda T, Maruyama N, Salleh MR, Mikami B & Utsumi S (2008) Characterization and crystallography of recombinant 7S globulins of Adzuki bean and structure-function relationships with 7S globulins of various crops. *J Agric Food Chem* **56**, 4145–4153.
- 79 Garman E (2010) Radiation damage in macromolecular crystallography: what is it and why should we care? *Acta Crystallogr* **D66**, 339–351.
- 80 Grabowski M, Langner KM, Cymborowski M, Porebski PJ, Sroka P, Zheng H, Cooper DR, Zimmerman MD, Elsliger MA, Burley SK *et al.* (2016) A public database of macromolecular diffraction experiments. *Acta Crystallogr D Struct Biol* **72**, 1181–1193.
- 81 Kowiel M, Jaskolski M & Dauter Z (2014) ACHESYM: an algorithm and server for standardized placement of macromolecular models in the unit cell. *Acta Crystallogr* **D70**, 3290–3298.
- 82 Han Q, Gao YG, Robinson H, Ding H, Wilson S & Li J (2005) Crystal structures of *Aedes aegypti* kynurenine aminotransferase. *FEBS J* **272**, 2198–2206.
- 83 Zheng H, Hou J, Zimmerman MD, Wlodawer A & Minor W (2014) The future of crystallography in drug discovery. *Expert Opin Drug Discov* **9**, 125–137.
- 84 Stanfield RL, Pozharski E & Rupp B (2016) Additional comment on three X-ray crystal structure papers. *J Immunol* **196**, 528–530.
- 85 Stanfield RL, Pozharski E & Rupp B (2016) Comment on three X-ray crystal structure papers. *J Immunol* **196**, 521–524.
- 86 Yao S, Flight RM, Rouchka EC & Moseley HN (2015) A less-biased analysis of metalloproteins reveals novel zinc coordination geometries. *Proteins* **83**, 1470–1487.
- 87 Raczynska JE, Wlodawer A & Jaskolski M (2016) Prior knowledge or freedom of interpretation? A critical look at a recently published classification of “novel” Zn binding sites. *Proteins* **84**, 770–776.
- 88 Yao S, Flight RM, Rouchka EC & Moseley HNB (2017) Perspectives and expectations in structural bioinformatics of metalloproteins. *Proteins* **85**, 938–944.
- 89 Bazzicalupi C, Ferraroni M, Bilia AR, Scheggi F & Gratteri P (2013) The crystal structure of human telomeric DNA complexed with berberine: an interesting case of stacked ligand to G-tetrad ratio higher than 1:1. *Nucleic Acids Res* **41**, 632–638.
- 90 Shiba T, Kametaka S, Kawasaki M, Shibata M, Waguri S, Uchiyama Y & Wakatsuki S (2004) Insights into the phosphoregulation of beta-secretase sorting signal by the VHS domain of GGA1. *Traffic* **5**, 437–448.
- 91 Nair DT, Johnson RE, Prakash S, Prakash L & Aggarwal AK (2004) Replication by human DNA polymerase- ϵ occurs by Hoogsteen base-pairing. *Nature* **430**, 377–380.
- 92 Wang J (2005) DNA polymerases: Hoogsteen base-pairing in DNA replication? *Nature* **437**, E6–E7.