Title

Of problems and opportunities—How to treat and how to not treat crystallographic fragment screening data

Journal

Protein Science, 31(9)

ISSN

0961-8368

Authors

Weiss, Manfred S
Wollenhaupt, Jan
Correy, Galen J
et al.

Publication Date

2022-09-01

DOI

10.1002/pro.4391

Peer reviewed

Received: 16 May 2022 | Revised: 17 June 2022 | Accepted: 11 July 2022

DOI: 10.1002/pro.4391

LETTER TO THE EDITOR

Of problems and opportunities—How to treat and how to not treat crystallographic fragment screening data

Manfred S. Weiss[1] | **Jan Wollenhaupt**[1] | **Galen J. Correy**[2] | **James S. Fraser**[2] | **Andreas Heine**[3] | **Gerhard Klebe**[3] | **Tobias Krojer**[4] | **Marjolein Thunissen**[4] | **Nicholas M. Pearce**[5]

[1]Macromolecular Crystallography, Helmholtz-Zentrum Berlin, Berlin, Germany

[2]Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, California, USA

[3]Institute of Pharmaceutical Chemistry, Philipps University Marburg, Marburg, Germany

[4]MAX IV Laboratory, Lund University, Lund, Sweden

[5]Department of Chemistry and Pharmaceutical Sciences, VU Amsterdam, Amsterdam, The Netherlands

Correspondence
Manfred S. Weiss, Macromolecular Crystallography, Helmholtz-Zentrum Berlin, D-12489 Berlin, Germany.
Email: msweiss@helmholtz-berlin.de

Review Editor: John Kuriyan

Abstract

In their recent commentary in Protein Science, Jaskolski *et al*. analyzed three randomly picked diffraction data sets from fragment-screening group depositions from the PDB and, based on that, they claimed that such data are principally problematic. We demonstrate here that if such data are treated properly, none of the proclaimed criticisms persist.

KEYWORDS

compositional heterogeneity, conformational heterogeneity, fragment-screening, group depositions, low-occupancy ligands, PanDDA

In their recent commentary in Protein Science,[1] M. Jaskolski and colleagues suggest that group depositions arising from large crystallographic screening campaigns to the Protein Data Bank wwPDB[2] are problematic and pose a serious threat to the integrity of the database. They mention that the group deposition models "*do not conform to the quality standards expected*", that they "*confuse most biomedical researchers*" and that they "*degrade the PDB integrity*". In order to overcome this, they postulate that such group depositions should either be "*clearly marked*" or they "*should be relocated from the PDB into a separate database*". Admittedly, some of the concerns of Jaskolski and colleagues are justified. However, PDB procedures and mechanisms, including optimized guidelines for group depositions are also rapidly evolving. Still, as with all new techniques, they will need some time to mature. In particular, standards for group depositions from fragment-screening campaigns are yet far from clear cut. Consequently, group depositions are often not adequately marked within the PDB and can even contain different data items. These attributes can make it difficult for the some PDB users, especially for those without an extensive structural biology background, to interpret these structures properly.

Here, we would like to caution that the arguments advanced by Jaskolski *et al.* are perhaps too one-sided.

They contrast refined data from single structure determinations with group depositions where the relevant information is spread over many individual data set and they conclude that the former are per se "better" and meet expected quality standards, particularly of placed ligands more precisely, whereas the latter are likely to suffer from reduced quality. We would like to assert, that these conclusions do not accurately reflect the complexity of these data and may, in fact, in our opinion hinder progress in the field of structural biology, in particular of drug discovery programs, in which these new methods are already widely applied. Commentaries that underestimate the knowledge of PDB users, that ignore the opportunities present in heterogeneous crystallographic data, and that miss out on chances for education on structure quality do more harm than good to structural biology. In the end, it may also sow distrust among end-users with respect to new methods to analyze valid primary data. Contrasting their rather conservative view of "a single dataset, a single structure, a single interpretation", we would like to inspire now a more positive and optimistic look at recent developments in particular in crystallographic screening campaigns. We envision a near future where many datasets can be collectively analyzed, bringing about an ensemble view of macromolecular structure that fully embraces both conformational and compositional heterogeneity in the underlying data.

Undoubtedly, the last decade has seen tremendous changes in macromolecular crystallography at synchrotron beamlines. Improvements to sample handling automation, detector speed and sensitivity, and on-the-fly data processing have made it possible to record and analyze large numbers of diffraction data sets.[3,4] These advances have been foundational to the use of "crystal soaking with small fragments followed by crystallography" as an extremely sensitive "binding assay" in fragment screens.[4–8] With further medicinal chemistry, the identified binders in such fragment screens may then be developed by rational design concepts into stronger binding compounds and ultimately into drugs.[9] While it was plainly impossible to record several hundreds of diffraction data sets within a manageable amount of time 10 years ago, it is now almost routine within just 24 hr of beam time[4] at several synchrotron sites around the world.[4–7] Industrial pharmaceutical research has heralded such experiments much earlier than the academic sector, however, due to intellectual property concerns, few of those results have been made publicly available. As academic efforts have intensified, fragment screens are reported more frequently in the recent literature,[10–12] and their results are deposited as a batch in the wwPDB.[2] Without any doubt, these data will have tremendous impact on health sciences and at the same time help pave the ground for new and essential techniques in data handling and evaluation.

Crucially, these developments have been enabled by diverting from the traditional paradigm of "a single dataset, a single structure, a single interpretation". A typical crystallographic fragment screening campaign comprises hundreds of individual diffraction data sets. In order to efficiently analyze the entire screen as one overall data set, Pearce and von Delft developed the Pan-Dataset Density Analysis (PanDDA) procedure.[13] PanDDA exploits the fact that the hundreds of structures are all nearly identical and exhibit correlated signals and errors. PanDDA works by aligning individual density maps and identifying spatially contiguous voxels with signals outside the background distribution. The background distribution can then be subtracted, resulting in an "event map", which is the most crucial read-out of a fragment screening experiment. The event map represents the primary evidence for the presence of a ligand, allowing the identification and modelling of low-occupancy ligands that are often elusive to classical (mFo-DFc)-difference-density based approaches. By inherently embracing the idea that the contents of the crystal are *compositionally* heterogeneous, in this case possibly as a result of the applied soaking procedure, this procedure is much more sensitive than treating the individual datasets in isolation and analyzing each for differences relative to a single "apo" dataset. As a consequence, individual fragments are often identified, despite having such low occupancy in the crystal that they are imperceptible in both the original electron density maps and by traditional reciprocal space difference map approaches.

Jaskolski *et al.*[1] are right to note that such a new approach with reporting and analyzing batch-data does not conform to some of the classical PDB expectations and that they cannot and should not be treated in the same way. However, what they do is that they pull out three individual coordinate sets from three different group deposition entries and look at them using the classical difference electron density map approach. It is therefore not surprising that they encounter the problems they reported. In the first case (PDB-Id 5RTL[12]) they do not observe a good agreement of low-occupancy ligand density with the associated (mFo-DFc) difference electron density map (Figure 1a) and question the presence of the ligand at all. However, if Figure 1a of Jaskolski *et al.* is contrasted by the PanDDA generated event map of PDB-Id 5RTL, the evidence for the presence of the ligand is clearly there. In their second example (PDB-Id 5RDH[10]), Jaskolski *et al.* report that the ultra-high resolution of the data set does not match the expected quality indicators of the structure (Figure 1b). They note that the ultra-high resolution may confuse non-expert users of the model and potentially lure them into using it as a reference model. However, a closer look at the data processing statistics reveals that the ultra-high resolution is just a consequence of a hitherto undiscovered

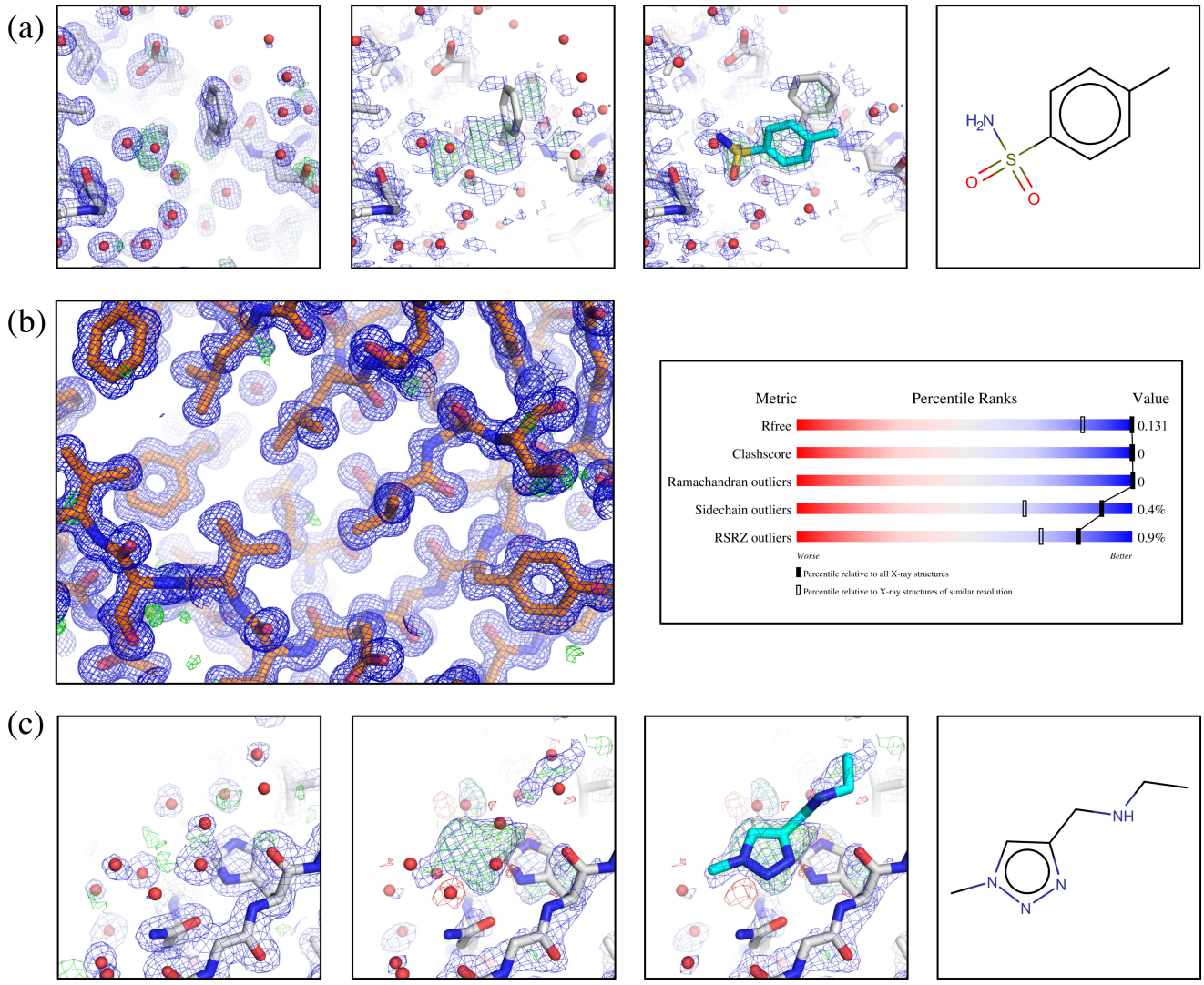

FIGURE 1 (a) Ligand identification for PDB-Id 5RTL. Left panel: the auto-refined model from the intermediate DIMPLE[18] step is shown along with the corresponding electron density maps: the (2mFo-DFc) map contoured at 1.0σ (blue) and the (mFo-DFc)-difference map contoured at 3.0σ (green/red). This panel is similar to fig. 1a of Jaskolski *et al.*[1] Middle panels: PanDDA Z-map (contoured at $Z = 3$, green/red) and PanDDA event map contoured at 1.0σ (blue), along with the auto-refined model and the ligand placed, respectively. The PanDDA event map coefficients are available from the PDB in the deposition's structure factor CIF. Right panel: chemical structure of the ligand, ZINC388056. (b) PDB-Id 5RDH after reprocessing. Left panel: auto-refined model and (2mFo-DFc) map, contoured at 1.0σ (blue), (mFo-DFc)-difference map contoured at 3.0σ (green/red). (c) Ligand identification for PDB-Id 5RFB. Left panel: the auto-refined model from the intermediate DIMPLE[18] step (D Fearon and F von Delft, personal communication) is shown along with the corresponding electron density maps: the (2mFo-DFc) map contoured at 1.0σ (blue) and the (mFo-DFc)-difference electron density map contoured at 3.0σ (green/red). This panel is similar to fig. 1c of Jaskolski *et al.*[1] Middle panels: The PanDDA Z-map (contoured at $Z = 3$, green/red) (D Fearon and F von Delft, personal communication) and the PanDDA event map (available from the PDB under PDB-Id 5RFB) contoured at 1.0σ (blue), along with the auto-refined model and the ligand placed (=bound state) model, respectively. Right panel: chemical structure of the ligand, Z1271660837

problem, leading to a faulty resolution cutoff during the automated data processing step.[14] Simple reprocessing yielded a data set to 0.93 Å resolution and a free R-factor after auto-refinement of 13%, which is perfectly in the expected range (Figure 1c). The third case that is reported (PDB-Id 5RFB[11]) is similarly as in the first case where the ligand is in the focus (Figure 1c). Also here, the presence of the ligand is clearly supported by the respective event map.

In this context, we also refute the statement of Jaskolski *et al.* that "*No useful conclusions can be derived by PDB users from this ligand ...*". Again, if the data are looked at properly,[15] accounting for the compositional heterogeneity and the fact that the ligands are low occupancy, none of the described criticisms persists (Figure 1).

Unfortunately, based on these three cases, Jaskolski *et al.* come to the general conclusion that models from

such group depositions will contaminate the PDB and that they should be deposited differently in a distinct resource, potentially even outside the PDB. Obviously, we agree on the notion that group depositions ought to be represented differently from single crystal data sets. Instead of demanding such "offending" models to be removed from the data bases such as the PDB, it seems that the real question to be asked is: "*how* should low-occupancy ligand structures and multi-state crystal structures be best presented in the PDB?" At the moment, there seems to be no satisfactory way to deposit full multi-state ensembles. With proper data management and curation tools and data sets accompanied by metadata for handling multi-state models in place in the PDB, however, this issue could be immediately solved allowing continued deposition of full sets of crystallographic data as well as clear presentation of the "states of interest" to the (non-specialist) user. Simply banning structures with low-occupancy ligands would almost certainly negatively interfere with future developments in structural biology.

The idea that depositing data beyond what might be of biological interest today is reflected in the instructive example of the surprising usefulness of Structural Genomics. About 25 years ago, Structural Genomics engaged in the massive determination of macromolecular structures with no particular biological question associated with the vast majority of the targets. Back then, some people in the field argued: "*These semi-automatically determined structural genomics structures are worse than the handmade structures*", that "*Structural Genomics is like stamp collecting*[16] *with no scientific meaning*" and so forth. Irrespective of that, Structural Genomics pushed forward and thousands of these structures ended up in the PDB, many of them without an associated publication. What is more, Structural Genomics triggered major advances not only in algorithms and automation, but also in solutions to less obvious problems such as data standardization, paving the way for next-generation developments like crystallographic fragment screening. But the benefits appear even far more wide-ranging: two decades later, clever scientists from way outside structural biology, leveraged these diverse coordinate sets to solve the long sought after sequence-to-structure prediction problem. It is quite likely that, without Structural Genomics, there would be no AlphaFold2.[17]

So here we are again... "*Thou shalt not contaminate the PDB*", we can hear the gatekeepers of the holy structure roar, "*with your fragment-screening data sets*". But then we may simply counter this with "why not?", as there will always be scientists who can make use of our data for novel developments and methods in a much cleverer way than we can currently imagine. Likely people will soon invent smart ways to efficiently extract all aspects of *conformational* as well as of *compositional* heterogeneity out of all these data sets. In doing so they might even "solve" protein conformational dynamics or protein-ligand binding prediction, much the way AlphaFold2 has "solved" protein structure prediction. As long as the data is there, let us embrace it and make it available!

AUTHOR CONTRIBUTIONS

Manfred S. Weiss: Conceptualization (equal); methodology (equal); writing – original draft (lead); writing – review and editing (lead). **Jan Wollenhaupt:** Methodology (supporting); visualization (lead); writing – original draft (supporting); writing – review and editing (supporting). **Galen J. Correy:** Methodology (supporting); writing – original draft (supporting). **James S. Fraser:** Conceptualization (equal); methodology (equal); writing – original draft (equal); writing – review and editing (supporting). **Andreas Heine:** Methodology (supporting); writing – original draft (supporting); writing – review and editing (supporting). **Gerhard Klebe:** Methodology (supporting); writing – original draft (supporting); writing – review and editing (supporting). **Tobias Krojer:** Methodology (supporting); writing – original draft (supporting); writing – review and editing (supporting). **Marjolein Thunissen:** Methodology (supporting); writing – original draft (supporting); writing – review and editing (supporting). **Nicholas M. Pearce:** Methodology (supporting); writing – original draft (supporting); writing – review and editing (supporting).

ACKNOWLEDGEMENTS

We would like to thank Daren Fearon and Frank von Delft (Diamond Light Source, UK) for providing the map and model files used to generate Figure 1c. We are also grateful for Hans Wienk (NKI Amsterdam, The Netherlands) for critically reading the manuscript and for many textual improvements. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Manfred S. Weiss https://orcid.org/0000-0002-2362-7047
Jan Wollenhaupt https://orcid.org/0000-0002-3418-5213
Galen J. Correy https://orcid.org/0000-0001-5155-7325
James S. Fraser https://orcid.org/0000-0002-5080-2859

Andreas Heine https://orcid.org/0000-0002-5285-4089
Gerhard Klebe https://orcid.org/0000-0002-4913-390X
Tobias Krojer https://orcid.org/0000-0003-0661-0814
Nicholas M. Pearce https://orcid.org/0000-0002-6693-8603

REFERENCES

1. Jaskolski M, Wlodawer A, Dauter Z, Minor W, Rupp B. Group depositions to the Protein Data Bank need adequate presentation and different archiving protocol. Protein Sci. 2022;31:784–786.
2. Burley SK, Berman HM, Bhikadiya C, et al. Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res. 2019;47:D520–D528.
3. Bowler MW, Svensson O, Nurizzo D. Fully automatic macromolecular crystallography: the impact of MASSIF-1 on the optimum acquisition and quality of data. Crystallogr Rev. 2016;22:233–249.
4. Douangamath A, Powell A, Fearon D, et al. Achieving efficient fragment screening at XChem facility at diamond light source. J Vis Exp. 2021;e62414. https://doi.org/10.3791/62414
5. Wollenhaupt J, Barthel T, Lima GMA, et al. Workflow and tools for crystallographic fragment screening at the Helmholtz-Zentrum Berlin. J Vis Exp. 2021;2021:1–19.
6. Lima GMA, Talibov VO, Jagudin E, et al. FragMAX: The fragment-screening platform at the MAX IV Laboratory. Acta Crystallogr Sect D Struct Biol. 2020;76:771–777.
7. Schiebel J, Krimmer SG, Röwer K, et al. High-throughput crystallography: reliable and efficient identification of fragment hits. Structure. 2016;24:1398–1409.
8. Schiebel J, Radeva N, Krimmer SG, et al. Six biophysical screening methods miss a large proportion of Crystallographically discovered fragment hits: a case study. ACS Chem Biol. 2016;11:1693–1701.
9. Erlanson D. Fragments in the clinic: 2021 edition, available under: https://practicalfragments.blogspot.com/2021/11/fragments-in-clinic-2021-edition.html
10. Wollenhaupt J, Metz A, Barthel T, et al. F2X-universal and F2X-entry: structurally diverse compound libraries for crystallographic fragment screening. Structure. 2020;28:694–706.e5.
11. Douangamath A, Fearon D, Gehrtz P, et al. Crystallographic and electrophilic fragment screening of the SARS-CoV-2 main protease. Nat Commun. 2020;11:5047.
12. Schuller M, Correy GJ, Gahbauer S, et al. Fragment binding to the Nsp3 macrodomain of SARS-CoV-2 identified through crystallographic screening and computational docking. Sci Adv. 2021;7:eabf8711.
13. Pearce NM, Krojer T, Bradley AR, et al. A multi-crystal method for extracting obscured crystallographic states from conventionally uninterpretable electron density. Nat Commun. 2017;8:15123.
14. Sparta KM, Krug M, Heinemann U, Mueller U, Weiss MS. XDSAPP2.0. J Appl Cryst. 2016;49:1085–1092.
15. Pearce NM, Krojer T, Von Delft F. Proper modelling of ligand binding requires an ensemble of bound and unbound states. Acta Crystallogr Sect D Struct Biol. 2017;73:256–266.
16. Fletcher L. Efforts to commercialize structural genomics may be limited. Nat Biotechnol. 2000;18:1036.
17. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–589.
18. Winn MD, Ballard CC, Cowtan KD, et al. Overview of the CCP4 suite and current developments. Acta Crystallogr. 2011;D67:235–242.

How to cite this article: Weiss MS, Wollenhaupt J, Correy GJ, Fraser JS, Heine A, Klebe G, et al. Of problems and opportunities—How to treat and how to not treat crystallographic fragment screening data. Protein Science. 2022;31(9):e4391. https://doi.org/10.1002/pro.4391