Smart, O. S. and G. Bricogne (2015).

Achieving High Quality Ligand Chemistry in Protein-Ligand Crystal Structures for Drug Design.

doi 10.1007/978-94-017-9719-1_13

Chapter 13 from: Multifaceted Roles of Crystallography in Modern Drug Discovery. G. Scapin, D. Patel and E. Arnold, Springer Netherlands**.**

Chapter available to purchase from http://link.springer.com/chapter/10.1007/978-94-017-9719-1_13

Book available to purchase from Google

Achieving high quality ligand chemistry in protein-ligand crystal structures for drug design

Oliver S. Smart [a,b*] and Gérard Bricogne [a]

(a) Global Phasing Ltd.
Sheraton House
Castle Park
Cambridge CB3 0AX
United Kingdom

(b) SmartSci Limited
St John’s Innovation Centre
Cowley Road
Cambridge CB4 0WS
United Kingdom

*Address correspondence to *oliver@smartsci.uk*

Abstract

The production of an X-ray crystal structure for a protein-ligand complex involves many steps, encompassing experimental and computational crystallography as well as chemoinformatics and computational chemistry. Using examples taken from the PDB, we show how a mistake made in any of these steps adversely affects the quality of the resulting structure, including that of the ligand. Procedures to assess the reliability of a ligand in a protein-ligand crystal structure are described. The merits of different responses to the identification of a problematic ligand structure in the PDB are examined. It is proposed that the best course of action is to cooperate with authors of the PDB entry and to deposit a corrected structure to replace the original. Two detailed examples of this process are provided by the deposition of improvements to PDB entries 1BYK and 1PMQ with their original depositors.

Keywords

macromolecular crystallography, model validation, refinement, restraints, ligand strain, Protein Data Bank

1. Introduction

This chapter looks at the steps necessary to produce a crystal structure of a protein-ligand complex with high-quality ligand placement. We will also look at ways of assessing the reliability of the ligand in such structures, whether taken from the Protein Data Bank (PDB) (Berman et al., 2000) or supplied by a colleague. The chapter is accompanied by two workshop practical sessions given at the Erice International School of Crystallography (Smart, 2014).

Crystal structures of protein-ligand complexes play a crucial role in structure-guided drug design (Leach & Gillet, 2003): they are used to understand protein structure-function relationships as well as for training and/or validating ligand docking programs (Leach & Gillet, 2003; Morris et al., 2009; Nissink et al., 2002). Achieving reliable ligand placement in these structures is therefore of the utmost importance.

Producing a crystal structure for a ligand-soak experiment on a protein for which a complete X-ray structure of the ligand-free "apo" protein is already available typically involves:

- ***(a)*** Experimental X-ray data collection from a protein crystal with the ligand soaked or co-crystallized, using a synchrotron beamline or an in-house diffractometer.
- ***(b)*** Data processing and integration to give space group, unit cell and structure factor amplitudes (SF).
- ***(c)*** Molecular replacement to optimally reposition the protein model for the cell and SF from ***(b)*** by rigid-body movements.
- ***(d)*** Initial refinement of model from ***(c)*** without a ligand.
- ***(e)*** Assessment of whether the difference electron density (ED) for the model from ***(d)*** warrants attempting to place a ligand.
- ***(f)*** Produce a molecular model and a restraint dictionary for the ligand.
- ***(g)*** Fit the model of ligand ***(f)*** into difference density and protein model from ***(d)***.
- ***(h)*** Refinement of combined protein and ligand model.
- ***(i)*** Assessment of refined protein-ligand complex ***(h)***.
- ***(j)*** If assessment shows issues then rebuild/refit protein, ligand and/or solvent and back to step ***(h)***.
- ***(k)*** Deposition of the structure model, SF, maps and validation data to an in-house database (or the PDB).

Most of these steps can be automated into a structure determination pipeline - for instance steps ***(b)*** to ***(i)*** are tackled by the Global Phasing tool PIPEDREAM (Sharff et al., 2014). A mistake made in any of these steps will adversely affect the quality of the resulting structure, including that of the ligand. To exemplify this we will examine a number of structures taken from the PDB. The PDB (Berman et al., 2000) is a databank of "complete" structures and provides a great resource for looking at mistakes made in solving protein-ligand complexes and for improving procedures so as to avoid such issues in the future (Terwilliger & Bricogne, 2014). This is particularly important both for the developers and for the users of automated pipelines.

1.1 Validation of the ligand in the crystal structure of a protein-ligand complex

It is important for the user of a protein-ligand structure to be able to assess the reliability of its ligand(s). For this purpose we have developed the BUSTER-REPORT program ("buster-report,"). To use this tool BUSTER (Bricogne et al., 2014) is first run to produce ED maps or to refine the structure in question, and then BUSTER-REPORT will analyze results providing an HTML page that reports on:

- ○ The X-ray data using the BUSTER reciprocal space correlation coefficients (RecSCC) plot. The RecSCC plot allows the detection of problems such as ice-ring contamination, anisotropic diffraction and incomplete data collection. For details see: http://www.globalphasing.com/buster/wiki/index.cgi?BusterReport.

- The usual statistics Rwork and Rfree as indicators of the overall progress and final performance of the refinement process in fitting the experimental X-ray data.
- MolProbity evaluation of protein geometry, including Ramachandran plots (Chen et al., 2010).

In addition BUSTER-REPORT provides reports for each ligand in the model, giving:

- Pictures of the ED around the ligand. These are provided as animated GIFs to aid visualization. The presence of large amounts of difference density around a ligand is a matter of concern (Figure 1a).
- The real space correlation coefficient (CC) of the ligand which provides an overall measure of the agreement between the 2Fo-Fc ED and the molecular model of the ligand. CC values below 0.8 are a prompt to reconsider the ligand placement.
- The average and maximum B-factor for ligand atoms. The B-factors are adjusted in refinement and describe the degree to which the ED is spread out. High ligand B-factors are often an indication of problematic placement, unless a degree of local disorder is made plausible by the ligand's environment, e.g. its proximity to the solvent boundary.
- The results of MOGUL on the geometry of the ligand. The MOGUL (Bruno et al., 2004) program is a tool that facilitates searching the Cambridge Structural Database of small-molecule organic and metal-organic crystal structures (CSD)(Allen, 2002) for geometric information relevant to a given ligand. MOGUL will rapidly analyze bond lengths, bond angles and most dihedral angles by finding CSD entries that contain similar chemical groups. In addition it provides data for many five and six-membered rings checking whether the ring pucker is similar to that found in related CSD entries. BUSTER-REPORT presents the results of this evaluation of geometric quality by means of colored 2D diagrams of each ligand (Figure 1b). Dihedral angles and ring scores are the most useful as metrics for validation, particularly if a GRADE (Smart, Holstein, & Womack, 2014) restraint dictionary is used in the refinement.

Although BUSTER-REPORT provides much useful information, it is best used together with direct visualization of the model and ED maps using COOT. This also gives an assessment of whether the ligand placement makes sense in terms of protein-ligand interactions. In general, correctly placed ligands will tend to form hydrogen-bond contacts to neighboring protein or solvent atoms as well as placing hydrophobic groups into hydrophobic environments.

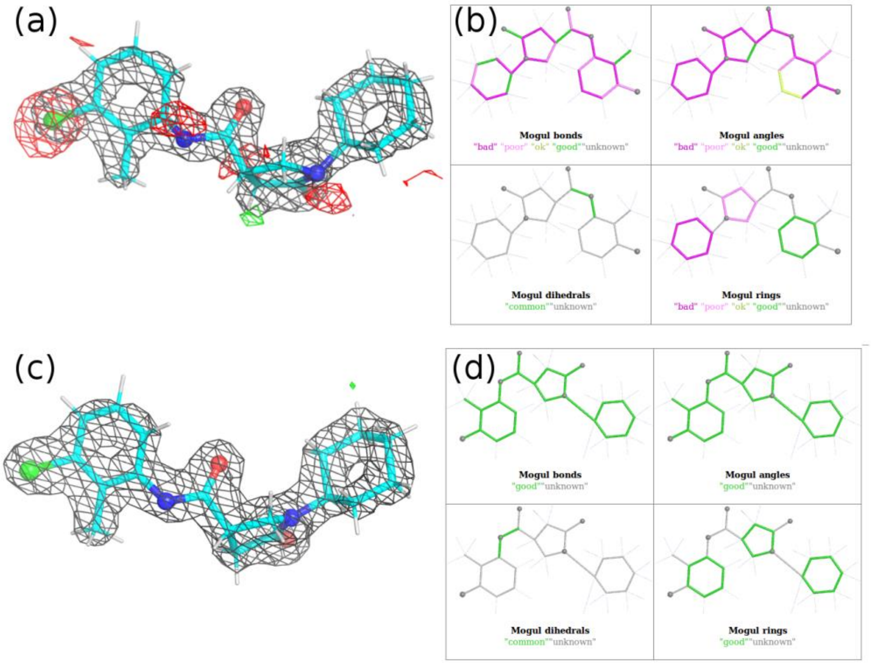

Figure 1 PDB entry 2H7P (He, Alian, Stroud, & de Montellano, 2006) (1.86 Å resolution). Panel (a) BUSTER (Bricogne et al., 2014) maps show considerable difference density for the pyrrolidine carboxamide ligand as modeled in the PDB (b) MOGUL (Bruno et al., 2004) validation measures show that the ligand has issues with bond lengths and angles and with ring puckers. After re-refinement with BUSTER using a refinement dictionary produced by GRADE (Smart et al., 2014) the fit to electron density is greatly improved (c) and no problems are found by MOGUL (d). All analysis and images are produced by BUSTER-REPORT ("BUSTER-REPORT,"). The 2Fo-Fc ED map is shown in grey at a contour-level of 1.3 rmsd. The Fo-Fc difference map is contoured at ±3.0 rmsd and shown in green for positive difference density and red for negative difference density. The full BUSTER-REPORT output is available from the introductory workshop practical available on-line (Smart, 2014). After seeing this analysis, Stroud and co-workers have deposited a corrected structure 4TZT into the PDB that has good ligand geometry and good fit to ED.

1.2 Electron Density

Examination of the electron density (ED) maps forms a crucial part of assessing whether a ligand in a protein-ligand complex can be relied upon. ED maps are produced by the program used to refine the structure. During the refinement process the maps will periodically be examined by the crystallographer using the COOT program (Emsley, Lohkamp, Scott, & Cowtan, 2010). Agreement between the experimental model of the protein, ligands and solvent molecules is assessed, and the model is adjusted as necessary, for instance by moving a protein side-chain or by placing water molecules into yet unmodelled density. Automated tools are increasingly used to help with the building process, but human examination and intervention are still normally necessary.

The ED maps at the end of refinement and model building can be seen to be as important as the refined model itself in reporting the result. It would be particularly useful to have access to the actual maps that the authors examined and interpreted in their work, in the concise form of their Fourier coefficients (i.e. amplitudes and phases). Unfortunately, these coefficients are not currently captured in a routine manner by the PDB deposition process and are seldom available from the archive itself. The Electron Density Server (EDS) at Uppsala (Kleywegt et al., 2004) provides maps recalculated with the REFMAC (Murshudov et al., 2011) refinement program for PDB entries where this is possible. EDS is a valuable resource for users of protein-ligand complexes from the PDB that enables rapid retrieval of the ED maps for most PDB depositions. Alternatively, BUSTER (Bricogne et al., 2014) includes tools that, for any given PDB code, will rapidly download data, calculate maps and provide a BUSTER-REPORT ("BUSTER-REPORT,") analysis of the structure. The BUSTER maps can then be inspected using BUSTER-REPORT or displayed using COOT (Emsley et al., 2010).

The 2Fo-Fc map indicates where ED is to be found according to the experimental X-ray data and the current refined atomic model. The Fo-Fc difference map indicates regions where the current model fails to place sufficient electrons (positive difference, normally shown in green) or places too many electrons (negative difference, normally shown in red). As shown in Figure 1c, re-refinement and/or rebuilding an incorrect model will tend to move atoms into the middle of 2Fo-Fc density and will reduce the amount of difference density. It should be noted that ED maps are not fixed: they generally improve as refinement and model building proceed. This is because as the model becomes more exact, the phases derived from it become more accurate, which in turn results in more accurate maps where more features become interpretable. The difference maps then become more sensitive and better able to highlight further unmodelled density or necessary corrections to the model.

1.3 The importance of the X-ray data resolution limit

The crystal structure of a protein-ligand complex is the result of an experiment where data are collected from a crystal of the protein soaked in, or co-crystallized with, the ligand compound. The resolution limit of the X-ray data has a great impact on the level of detail that will be revealed by the ED maps.

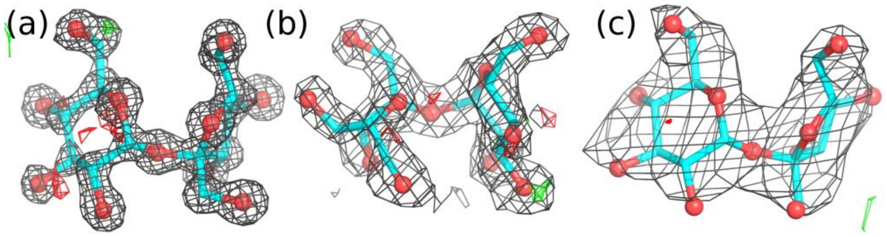

Figure 2 .The effect of X-ray data resolution limit on the level of detail available. Buster-report images of ED maps for the sucrose ligand in structures (a) 1YLT 1.2Å resolution, (b) 2PWE 2.0Å resolution and (c) 2QQV 3.0Å resolution. In all three cases the placement and refinement of the sucrose ligand is good.

Figure 2 shows that, at a resolution 1.2 Å or better, individual atoms can be distinguished in the map, thus providing often exquisite amount of detail for a ligand (such as indicating its exact chemistry). At around 2.0 Å resolution the map is less detailed but ligand placement will still be good with the data generally determining torsion angles well and often revealing details about ring pucker. At 3.0 Å resolution or worse, much less detail is available. Ligands can still normally be positioned with confidence, but it becomes increasingly essential to have prior knowledge of the chemistry of the ligand as the resolution worsens. However, at low resolution many details are not available, and it must be borne in mind that features such as ring pucker may eventually be set as a consequence of the restraint dictionary and fitting procedures used, rather than on the basis of the X-ray data.

1.4 Data collection problems

The importance of collecting data correctly cannot be overstated. An example of a PDB entry where poor data collection directly affects the result is 1T0O (Golubev et al., 2004). BUSTER reports that the data are incomplete (Figure 3a), and further analysis with the CCP4 program HKLVIEW shows that little data has been collected along the k axis (Figure 3b). This results in a map with artefacts along the y axis, causing the ED for the ligand to join up with that of the protein (Figure 3c). Although the nominal data resolution limit of 1T0O is 1.96 Å (Golubev et al., 2004), this systematic data incompleteness makes interpretation difficult. The only way to tackle data collection problems is to collect more and/or better data in the course of the experiment itself.

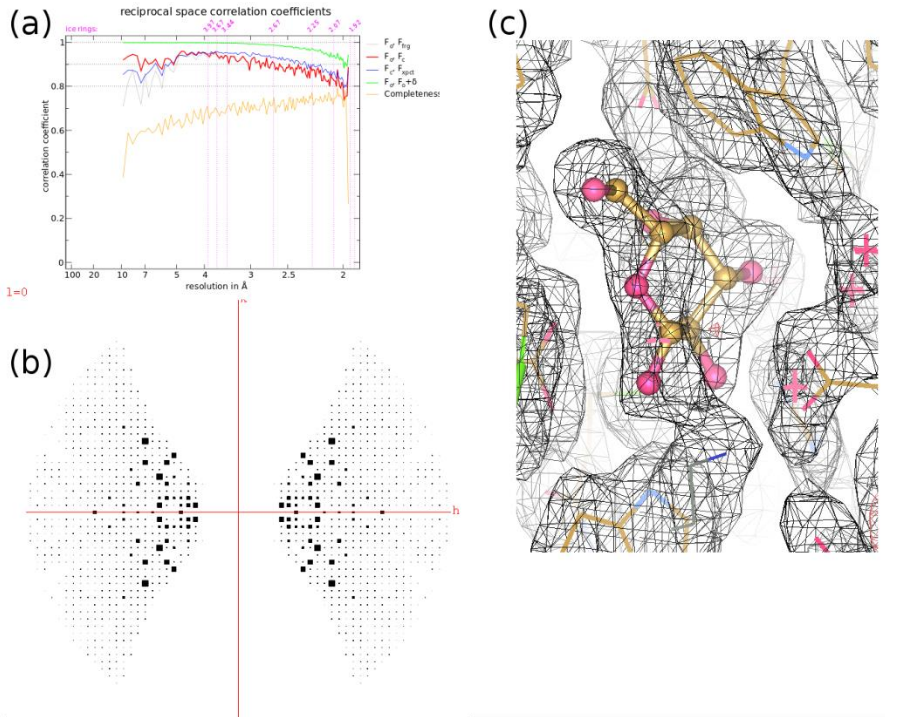

Figure 3 . PDB entry 1T0O (a) BUSTER (Bricogne et al., 2014) RecSSC shows the data is incomplete across the entire resolution range (b) HKLVIEW shows incomplete data is because no data were collected along k axis (c) BUSTER map after refinement shows density is poor along y direction, with the 2Fo-Fc ED for the galactose ligand (ball and stick) merging with that for protein side-chains at top and bottom. This merging along the y axis happens throughout the structure (the water molecule on the left provides another example).

Global Phasing is currently helping a number of synchrotrons to provide users with strategies to collect better data for a given crystal.

1.5 Data processing problems

Once the X-ray data are collected, the resulting diffraction images must be processed and the Bragg diffraction spots integrated. There are a number of programs to do this and the topic is outside the scope of this presentation except to note that data processing must be correctly done to obtain meaningful results.

Many mistakes can be made at the data integration and other stages during the processing. A common error is to not properly tackle "ice rings" in the diffraction images (Rupp, 2010; Vonrhein et al., 2011). These are caused by the build-up of ice microcrystals on the protein crystal during data collection, and result in rings at characteristic resolutions. The affected resolution ranges should be excluded from all processing steps. Failure to do so has a detrimental effect on the internal scaling of the data, resulting in poor refinement and in ED map artefacts.

1.6 Is there electron density for the ligand?

Given successful data processing, molecular replacement and initial refinement of the ligand-free protein model, the next step will be to assess the resulting ED maps for the presence of bound ligand, either at a known binding site or elsewhere. Pozharski, Weichenberger, and Rupp (2013) emphasise that an unfortunately very common error is to believe that, because a ligand compound has been soaked, it must necessarily bind, and to model the ligand despite there being no evidence of its presence in the ED maps. For example, Pozharski et al. (2013) classify the diclofenac ligand in PDB entry 3IB0 (Mir et al., 2009) as "absent". The BUSTER (Bricogne et al., 2014) map supports this classification (Figure 4a). Revising the model by removing the diclofenac and refining with BUSTER including automated water molecule placement shows the ED into which the ligand had been placed can be well modeled by three water molecules (Figure 4b).

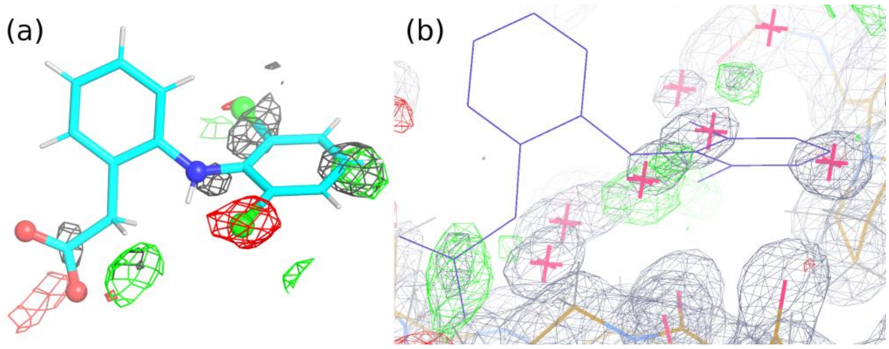

Figure 4 Pozharski et al. (2013) classify the diclofenac ligand in PDB entry 3IB0 (Mir et al., 2009) (1.4 Å resolution) as "absent". BUSTER-REPORT supports this classification: the ligand has high B-factors and a CC with the (2Fo-Fc) map of 0.57, and as shown in panel (a) there is only a small amount of disconnected ED around it. Panel (b) shows a Coot image of the result of BUSTER re-refinement of the protein after the diclofenac (shown here as a thin purple "ghost") has been removed. The re-refinement included automated water placement, and shows that the ED can be well modeled by three water molecules (red crosses) that form good hydrogen bonds.

By contrast, it is also possible to misinterpret ligand density as bound solvent. An interesting example of this is provided by PDB entry 2GWX as discussed in a review by Andrew M. Davis, St-Gallay, and Kleywegt (2008). In the original structure, ED in the ligand binding site was interpreted as being due to bound water molecules. Re-evaluation of the structure using the original SF by Fyffe et al. (2006) led to the conclusion that this ED was actually due to a fatty acid ligand. In addition, clear density was found for n-heptyl-b-D-glucopyranoside (an additive in crystallization) in four sites. The revised structure is available as PDB entry 2BAW.

1.7 Producing the restraint dictionary for the ligand

Given evidence in electron density to place the ligand, the next step is to fit a model into that density. Before this, it is necessary to produce an initial molecular model for the ligand, together with a restraint dictionary comprising a complete set of ideal bond distances and bond angles as well as listing chiral atoms and planar groups. Such a dictionary describes, typically using a CIF format, the chemical nature of the ligand, its molecular connectivity and its flexibility. That information is required not only to define the degrees of freedom available in fitting the ligand into its target ED and for manipulating ligands in COOT, but also to provide additional stereochemical information to packages such as REFMAC (Murshudov et al., 2011), BUSTER and phenix.refine (Afonine et al., 2012) to maintain good molecular geometry during structure refinement in spite of the limited resolution of the X-ray data. Molecular mechanics force fields provide an alternative to simple restraint dictionaries [Wlodek et al, 2006]. BUSTER has recently been extended to allow the use of the MMFF94s force field for ligands (and force field conformational strain energy may provide and additional ligand validation metric).

GRADE (Smart et al., 2014) is the Global Phasing restraint dictionary generator. It takes a SMILES string or "mol2" file containing 3D coordinates of all atoms as input. Like BUSTER-REPORT, GRADE uses the CSD structures as the primary source of restraint information by invoking the MOGUL (Bruno et al., 2004) program. MOGUL will rapidly analyze bond lengths, bond angles and many dihedral angles by finding CSD structures that contain similar chemical groups. Where MOGUL cannot provide information quantum chemical procedures are invoked. As well as being distributed with the BUSTER package GRADE can be used through the Grade Web Server ("Grade Web Server,").

A mistake in describing the stereochemistry of the ligand can result in the wrong ligand being fit and refined. Chiral inversions in carbohydrates are a good example. Smart et al. (2012) describe how re-refinement of PDB entry 1DET using BUSTER and a GRADE dictionary corrected a chiral inversion in the ribose ring of the ligand: the re-refined model has been deposited as PDB entry 3SYU. Liebeschuetz, Hennemann, Olsson, and Groom (2012) mention PDB entry 2EVS (Malinina et al., 2006) as a similar example, where the hexyl-beta-D-glucoside ligand has been refined and deposited with a chiral inversion of the anomeric carbon atom. This inversion can be corrected by re-refinement, but re-deposition has not yet been performed. An additional example is described in section 2.1 where re-refinement is used to correct an inverted chiral atom in trehalose-6-phosphate in 1byk.

Figure 1 shows how BUSTER re-refinement of PDB entry 2H7P (He et al., 2006) with a GRADE (Smart et al., 2014) dictionary for the ligand markedly improves its fit to the ED. As shown in the first workshop practical given at the Erice School (now available online (Smart, 2014)) re-refinement also deals with stereochemistry issues raised by MOGUL. Most notably it alters the pucker of the 5-membered lactam ring and cyclohexyl rings to conformations seen in the CSD (Figure 1d). A corrected structure 4TZT that has both a good fit to ED and ligand geometry has now been deposited into the PDB to replace 2H7P.

1.8 Ligand fitting

Given suitable difference density and a restraint dictionary for the ligand, the next step is to exploit the flexibility of the ligand, as implicitly defined by that dictionary, to fit it into ED (either difference density, or 2Fo-Fc density). This can be done by hand using the COOT program, or by entrusting the task to an automated ligand fitter such as Global Phasing's RHOFIT or OpenEye's AFITT (Wlodek, Skillman, & Nicholls, 2006).

Ligand fitting becomes increasingly difficult as the data resolution limit worsens, because the ED will necessarily cease to reflect aspects of the ligand shape that have a decisive role in the selection of the correct ligand pose. An extreme example is the location of an extra copy of the 12-residue cyclic peptide in PDB entry 1OSG by Smart et al. (2012) where knowledge of the conformation of the peptide was essential to be able to interpret the difference ED. The re-interpreted model including that extra copy of the ligand is available as PDB entry 3V56.

The importance of achieving a good ligand fit for structure-guided drug discovery is illustrated by the example of the inhibitor DDR1-IN-1 bound to DDR1 kinase domain by Kim et al. (2013). In the original published structure [28] and the associated PDB deposition 4BKI, the indolin-2-one ring of the inhibitor was positioned according to the inhibitor design so as to form two hydrogen bonds to the protein. BUSTER re-refinement of the structure with GRADE restraints (Smart et al., 2014) and evaluation of the ligand geometry with BUSTER-REPORT revealed to us that this ring positioning resulted in geometrical strain as well as in a strengthening of the difference ED, indicating that the ring should be flipped (Figure 5a). The ligand placement after a ring flip and re-refinement is significantly better, with a good fit to the 2Fo-Fc density (Figure 5b). After seeing this analysis, Canning, Bullock and co-workers deposited a corrected structure 4CKR and published a correction (Kim et al., 2014). Given that the indolin-2-one ring fails to form the anticipated hydrogen bond contact and instead packs with the hydrophobic side of the ring adjacent to the main chain carbonyl of residue 702, there is a clear scope for revising the initial approach to designing a ligand that would form optimal interactions with the protein at that site.

2. Results

2.1 Achieving correct ligand geometry in trehalose receptor structure

The re-refinement of 1BYK provides an informative example of how a wrong assignment of chirality in a ligand can produce clear knock-on effects that are sensed by the metrics for both the ED fit and the ligand geometry. The structure is that of the *E. coli* Trehalose Receptor in complex with trehalose-6-phosphate, solved in 1998 at 2.5Å resolution structure by Hars, Horlacher, Boos, Welte, and

Diederichs (1998). Trehalose is a natural alpha-linked disaccharide formed by an α,α-1,1-glycosidic bond between two α-glucose units. However, the original 1BYK deposition used β-glucopyranose instead of the α-anomer for one of the sugar rings. This error propagated to the PDB chemical components dictionary (Dimitropoulos, Ionides, & Henrick, 2006; Feng et al., 2004), giving rise to a definition of trehalose-6-phosphate T6P that specified the incorrect anomer, whereas the entry for trehalose itself was correct. BUSTER re-refinement of the structure with a GRADE (Smart et al., 2014) restraints dictionary for that incorrect anomer results in a structure with strong difference ED next to the inverted atom (Figure 5a). Furthermore, MOGUL geometry validation through BUSTER-REPORT (Figure 5b and Table 1) show that the geometry of the molecule is forced to be “unusual” because of the strain induced by fitting to ED that is not compatible with the model density for the ligand with its incorrect geometry.

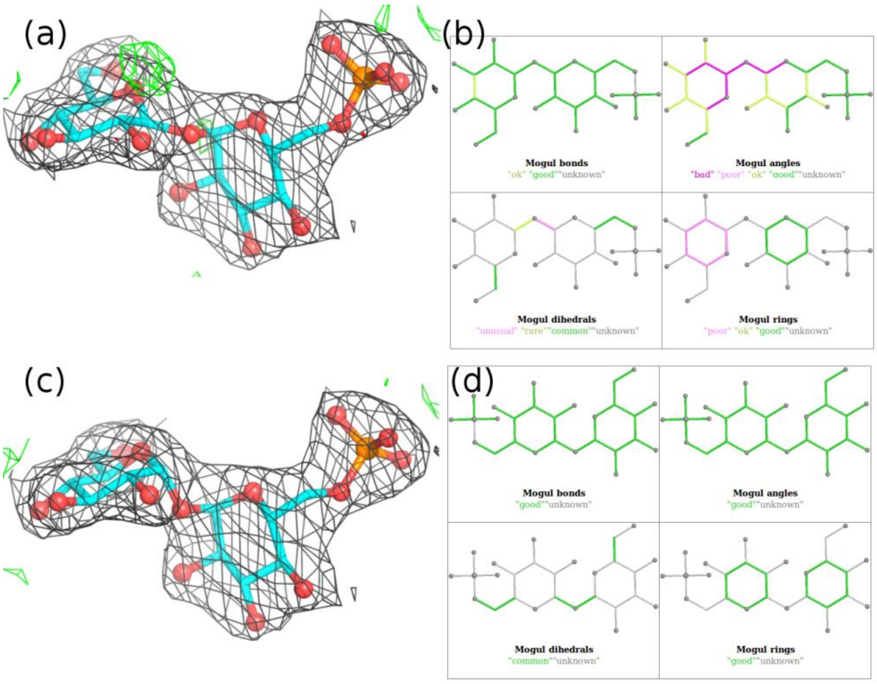

Figure 5 Panel (a) Buster re-refinement of 1BYK with a restraint dictionary for trehalose-6-phosphate specifying an incorrect anomer results in a structure with strong difference ED next to the inverted C1 atom. Panel (b) shows MOGUL results, indicating that refining with the inverted C1 atom produces a conformation with poor geometry. Re-refinement with a corrected trehalose-6-phosphate yields a much better fit to density (c) and alters MOGUL metrics to “good” or “common” (d).

Table 1 re-refinement of 1byk correcting ligand geometry

	1byk.pdb	1byk.pdb re-refined‡ using T6P dictionary with incorrect trehalose	1byk.pdb re-refined‡ using corrected T6P dictionary	After multiple rounds of re-building and re-refinement‡
BUSTER R_{work}	0.1935	0.1617	0.1604	0.1510
BUSTER R_{free}	0.1976	0.1871	0.1866	0.1730
100*(R_{free} - R_{work})	0.4%	2.5%	2.6%	2.2%
T6P ED fit CC 2Fo-Fc†	0.961	0.978	0.985	0.988
T6P Mogul "bad" angles (#|Z|>4) †	6	4	0	0
T6P Mogul "unusual" dihedrals/rings†	2 / 0	1 / 1	0 / 0	0 / 0
Number of water molecules placed	44	44	44	79
MolProbity Ramachandran outliers	0.8%	0.4%	0.2%	0%
MolProbity Ramachandran favored	94.9%	97.0%	96.8%	98.4%
MolProbity side chains with poor rotamers	6.6%	7.1%	7.4%	1.3%
MolProbity Overall Score / Percentile	2.13 / 92nd	1.64/99th	1.56/99th	0.73/100th

† Figure given for A chain copy only and the B chain values are similar

‡ BUSTER –autoncs option used (Smart et al., 2012)

Kay Diederichs and colleagues at the University of Konstanz asked for our assistance in correcting the structure. The raw diffraction data were re-processed with the current version of XDS (Kabsch, 2010) resulting in a dataset that had a completeness of 99.4% compared to 67.5% for the original. This demonstrates the importance of the retention of diffraction images (Terwilliger & Bricogne, 2014). Care was taken to ensure that the set of reflections used for Rfree (Brunger, 1992) was kept consistent with original structure factors. The next task was to produce a GRADE (Smart et al., 2014) restraints dictionary for T6P with the correct chirality. BUSTER re-refinement of the structure using this dictionary flipped the incorrect chiral centre without any further intervention. Following this re-refinement, the trehalose-6-phosphate fits the ED well with no difference density (figure 5c). In addition, all MOGUL metrics are altered to "good", showing that the trehalose-6-phosphate stereochemistry is now in complete agreement with that expected from related saccharides in the CSD (Table 1 and Figure 5d). The model was improved by rounds of rebuilding using coot and MolProbity (Chen et al., 2010) to assess geometry and ED fit. Table 1 shows how modern tools can achieve a structure that has improved interpretation and much better "quality metrics" than in 1998. This is a good example of the process of the mutual improvement of X-ray crystallographic software and structure models in the PDB (Terwilliger & Bricogne, 2014). It should be noted that the conclusions drawn by Hars et al. (1998) from the original structure are unaffected. The corrected structure has been deposited in the PDB as entry 4XXH obsoleting the original 1BYK entry. The PDB chemical components dictionary (Dimitropoulos et al., 2006; Feng et al., 2004) definition of trehalose-6-phosphate T6P has also been updated.

2.2 New insights into the ligand geometry in a JNK3 kinase structure

The PDB entry 1PMQ for the JNK3 kinase complex (Scapin, Patel, Lisnock, Becker, & LoGrasso, 2003) provides an interesting example of how advances in methodology can lead to improvements in the modeling of ligands. Note that the material here forms the basis for the second workshop practical session given at the Erice School, now available online (Smart, 2014).

1PMQ is the structure of JNK3 in complex with an imidazole-pyrimidine inhibitor, solved in 2003 by Giovanna Scapin and colleagues at Merck (Scapin et al., 2003). The ligand has been assigned the three-letter code 880 in the PDB chemical components dictionary (Dimitropoulos et al., 2006; Feng et al., 2004) . Visual inspection of the deposited PDB entry together with ED maps from BUSTER shows that the model for ligand 880 fits the density well (Smart, 2014). However, MOGUL analysis as provided by BUSTER-REPORT ("buster-report,") shows that it would be expected from CSD structures that the atom C55 of the cyclohexyl ring should be coplanar with the pyrimidine ring in the ligand, but that this is not the case in 1PMQ. Simple re-refinement with BUSTER using a GRADE (Smart et al., 2014) restraints dictionary for 880 cannot fix the problem, but once the cyclohexyl ring is flipped manually re-refinement achieves a good fit to density, good ligand geometry and improved geometry for the protein-ligand hydrogen bonds (Smart, 2014).

As the data set has a high degree of anisotropy the Diffraction Anisotropy Server (Strong et al., 2006) was used, producing a noticeable improvement in map quality. Inspection of the ED enables further improvements to the structure. The dichlorophenyl ring in the 880 ligand shows difference density near atom CL45, indicating that the ring has two alternate conformations that can be modelled (Smart, 2014). The improved maps and model support the identification by Scapin et al. (2003) of the “accidental” second ligand AMP-PCP as well as a subtle improvement in its ED fit and geometry. After additional rounds of rebuilding the protein/solvent, the corrected structure (Smart, 2014). has now been deposited in the PDB as entry 4Z9L, obsoleting the original 1PMQ entry. Once again improvements to the structure are limited and only add support to the conclusions drawn byScapin et al. (2003).

3. Discussion

Many researchers have pointed out that there are problems in the chemistry, placement and fit of ligands in the PDB (A. M. Davis, Teague, & Kleywegt, 2003; Joosten, Womack, Vriend, & Bricogne, 2009; Kleywegt, 2007; Liebeschuetz et al., 2012; Malde & Mark, 2011; Reynolds, 2014; Warren, Do, Kelley, Nicholls, & Warren, 2012; Wlodek et al., 2006) and have asked what can be done to improve matters. The implementation of the recommendations of wwPDB X-ray crystallographic task force (Read et al., 2011), and in particular the use of MOGUL analysis as part of the deposition process, will hopefully contribute to avoiding problems in current and future depositions. A critical factor in this context is “the urgent need to provide adequate training to next-generation crystallographers” as noted by Dauter, Wlodawer, Minor, Jaskolski, and Rupp (2014). It is hoped that this chapter, together with the accompanying workshop practical material (Smart, 2014), will make a positive contribution, however small it may be, towards this goal.

Improving matters for the future is essential, but what can be done about problems with existing PDB entries? One solution is to produce secondary databases containing re-refined, corrected and/or curated structures, as has been done as part of the PDB-REDO (Joosten, Joosten, Murshudov, & Perrakis, 2012) and IRIDIUM (Warren et al., 2012) projects. This is a valuable approach but its usefulness is likely to be restricted to a small number of specialist users. It is the PDB (Berman et al., 2000) itself that is the vital resource for many non-specialists, and it is a regrettable that problematic entries very often persist in the database.

On reflection, what is unacceptable is to criticize PDB entries to the extent of proposing alternative models, without taking any corrective action nor being required to do so. In many respects this is fundamentally unfair to both the original depositors and the users of the PDB. Journals now all require deposition into the PDB of any structures reported in a paper. Although this criterion is universally applied to the reports of *new* structures, it appears not to be applied to publications pointing out errors in *existing* (i.e. already deposited) structures, even when an alternative, purportedly improved structure is shown in a figure. Matters are made worse by the fact that the allegedly problematic structures are only referred by their PDB code, without citing the original reference. The practical upshot of this is that a researcher relying on such a PDB entry has little chance of ever becoming aware that it has been called into question in a published paper, as a literature survey would fail to find a reference to the latter. It would be particularly annoying for any researcher who used a PDB result to find out that a correction was available but had gone unrecorded.

On occasion, complete deposition may not be straightforward because the result originates from a molecular modeling method. In such cases, the best option in the first instance is to contact the authors of the deposition(s) and suggest that revision and re-deposition is in order. Failing this, it may be possible to find a friendly protein crystallographer to produce a re-refined corrected structure and deposit this as a “REMARK 0 alternative interpretation”, including the methodology used as part of the publication. Such an entry is given a separate PDB code and does not replace the original entry in the PDB. If it is not possible to achieve a deposition in the actual PDB, then at the minimum the coordinates of the proposed alternative model should be included as Supplementary Material that will thus be available with the publication. It is important to include a citation of the original publication to ensure that users of a structure will more easily be able to find relevant information about its amended versions. Paper reviewers should encourage deposition whenever possible.

We hasten to add that in the past we have been guilty of exactly the behavior that we criticize above. However, our intention is to ensure that corrected entries appear in the PDB wherever possible. This should further invigorate the recently analysed process of continuous mutual improvement of macromolecular structure models in the PDB and of X-ray crystallographic software (Terwilliger & Bricogne, 2014).

To conclude, we would strongly recommend that users of ligand complexes from the PDB take a cautious approach and make full use of the critical assessment tools available at the time when they wish make use of an existing entry, however recently it may have been deposited, as those tools may themselves have improved.

4. Acknowledgements

Our warmest thanks go to our colleagues at Global Phasing for discussions, provision of examples and help with these notes. This work was supported by members of the Global Phasing Consortium and by the SILVER FP7-HEALTH Integrated Project of the European Commission under grant agreement no. 260644. We thank all the crystallographers who deposited into the PDB the structures that have been analyzed here. Special thanks to Kay Diederichs, Giovanna Scapin and colleagues. Critical assessment of the structures allows for improvements to be made in methods, so that problems can be corrected in existing structures and avoided in future structures.

Bibliography

Afonine, P. V., Grosse-Kunstleve, R. W., Echols, N., Headd, J. J., Moriarty, N. W., Mustyakimov, M., Terwilliger, T. C., Urzhumtsev, A., Zwart, P. H., & Adams, P. D. (2012). Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallographica Section D-Biological Crystallography, 68,* 352-367. doi: 10.1107/s0907444912001308

Allen, F. H. (2002). The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallographica Section B-Structural Science, 58*, 380-388. doi: 10.1107/s0108768102003890

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research, 28*(1), 235-242. doi: 10.1093/nar/28.1.235

Bricogne, G., Blanc, E., Brandl, M., Flensburg, C., Keller, P., Paciorek, W., Roversi, P., Sharff, A., Smart, O. S., Vonrhein, C., & Womack, T. O. (2014). BUSTER version 2.13.0. Cambridge, United Kingdom.: Global Phasing Ltd.

Brunger, A. T. (1992). Free R-value - a novel statistical quantity for assessing the accuracy of crystal-structures. *Nature, 355*(6359), 472-475. doi: 10.1038/355472a0

Bruno, I. J., Cole, J. C., Kessler, M., Luo, J., Motherwell, W. D. S., Purkis, L. H., Smith, B. R., Taylor, R., Cooper, R. I., Harris, S. E., & Orpen, A. G. (2004). Retrieval of crystallographically-derived molecular geometry information. *Journal of Chemical Information and Computer Sciences, 44*(6), 2133-2144. doi: 10.1021/ci049780b

buster-report. from https://www.globalphasing.com/buster/wiki/index.cgi?BusterReport

Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., & Richardson, D. C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D-Biological Crystallography, 66*, 12-21. doi: 10.1107/s0907444909042073

Dauter, Z., Wlodawer, A., Minor, W., Jaskolski, M., & Rupp, B. (2014). Avoidable errors in deposited macromolecular structures: an impediment to efficient data mining. *IUCrJ, 1*(Pt 3), 179-193. doi: 10.1107/s2052252514005442

Davis, A. M., St-Gallay, S. A., & Kleywegt, G. J. (2008). Limitations and lessons in the use of X-ray structural information in drug design. *Drug Discovery Today, 13*(19-20), 831-841. doi: 10.1016/j.drudis.2008.06.006

Davis, A. M., Teague, S. J., & Kleywegt, G. J. (2003). Application and limitations of X-ray crystallographic data in structure-based ligand and drug design. *Angewandte Chemie-International Edition, 42*(24), 2718-2736. doi: 10.1002/anie.200200539

Dimitropoulos, D., Ionides, J., & Henrick, K. (2006). Using MSDchem to search the PDB ligand dictionary. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.], Chapter 14*, Unit14.13. doi: 10.1002/0471250953.bi1403s15

Emsley, P., Lohkamp, B., Scott, W. G., & Cowtan, K. (2010). Features and development of Coot. *Acta Crystallographica Section D-Biological Crystallography, 66*, 486-501. doi: 10.1107/s0907444910007493

Feng, Z. K., Chen, L., Maddula, H., Akcan, O., Oughtred, R., Berman, H. M., & Westbrook, J. (2004). Ligand Depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics, 20*(13), 2153-2155. doi: 10.1093/bioinformatics/bth214

Fyffe, S. A., Alphey, M. S., Buetow, L., Smith, T. K., Ferguson, M. A. J., Sorensen, M. D., Bjorkling, F., & Hunter, W. N. (2006). Reevaluation of the PPAR-beta/delta ligand binding domain model reveals why it exhibits the activated form. *Molecular Cell, 21*(1), 1-2. doi: 10.1016/j.molcel.2005.12.001

Golubev, A. M., Nagem, R. A. P., Neto, J. R. B., Neustroev, K. N., Eneyskaya, E. V., Kulminskaya, A. A., Shabalin, K. A., Savel'ev, A. N., & Polikarpov, I. (2004). Crystal structure of alpha-galactosidase from Trichoderma reesei and its complex with galactose: Implications for catalytic mechanism. *Journal of Molecular Biology, 339*(2), 413-422. doi: 10.1016/j.jmb.2004.03.062

Grade Web Server. from http://grade.globalphasing.org/

Hars, U., Horlacher, R., Boos, W., Welte, W., & Diederichs, K. (1998). Crystal structure of the effector-binding domain of the trehalose-repressor of Escherichia coli, a member of the LacI family, in its complexes with inducer trekalose-6-phosphate and noninducer trehalose. *Protein Science, 7*(12), 2511-2521. doi: 10.1002/pro.5560071204

He, X., Alian, A., Stroud, R., & de Montellano, P. R. O. (2006). Pyrrolidine carboxamides as a novel class of inhibitors of enoyl acyl carrier protein reductase from Mycobacterium tuberculosis. *Journal of Medicinal Chemistry, 49*(21), 6308-6323. doi: 10.1021/jm060715y

Joosten, R. P., Joosten, K., Murshudov, G. N., & Perrakis, A. (2012). PDB_REDO: constructive validation, more than just looking for errors. *Acta Crystallographica Section D-Biological Crystallography, 68*, 484-496. doi: 10.1107/s0907444911054515

Joosten, R. P., Womack, T., Vriend, G., & Bricogne, G. (2009). Re-refinement from deposited X-ray data can deliver improved models for most PDB entries. *Acta Crystallographica Section D-Biological Crystallography, 65*, 176-185. doi: 10.1107/s0907444908037591

Kabsch, W. (2010). XDS. *Acta Crystallographica Section D-Biological Crystallography, 66*, 125-132. doi: 10.1107/s0907444909047337

Kim, H. G., Tan, L., Weisberg, E. L., Liu, F. Y., Canning, P., Choi, H. G., Ezell, S., Zhao, Z., Wu, H., Wang, J. H., Mandinova, A., Bullock, A. N., Liu, Q. S., Lee, S. W., & Gray, N. S. (2014). Discovery of a Potent and Selective DDR1 Receptor Tyrosine Kinase Inhibitor (vol 8, pg 2145, 2013). *Acs Chemical Biology, 9*(3), 840-840. doi: 10.1021/cb5000949

Kim, H. G., Tan, L., Weisberg, E. L., Liu, F. Y., Canning, P., Choi, H. G., Ezell, S. A., Wu, H., Zhao, Z., Wang, J. H., Mandinova, A., Griffin, J. D., Bullock, A. N., Liu, Q. S., Lee, S. W., & Gray, N. S. (2013). Discovery of a Potent and Selective DDR1 Receptor Tyrosine Kinase Inhibitor. *Acs Chemical Biology, 8*(10), 2145-2150. doi: 10.1021/cb400430t

Kleywegt, G. J. (2007). Crystallographic refinement of ligand complexes. *Acta Crystallographica Section D-Biological Crystallography, 63*, 94-100. doi: 10.1107/s0907444906022657

Kleywegt, G. J., Harris, M. R., Zou, J. Y., Taylor, T. C., Wahlby, A., & Jones, T. A. (2004). The Uppsala Electron-Density Server. *Acta Crystallographica Section D-Biological Crystallography, 60*, 2240-2249. doi: 10.1107/s0907444904013253

Leach, A. R., & Gillet, V. J. (2003). *An introduction to chemoinformatics*. Dordrecht ; Boston: Kluwer Academic Publishers.

Liebeschuetz, J., Hennemann, J., Olsson, T., & Groom, C. R. (2012). The good, the bad and the twisted: a survey of ligand geometry in protein crystal structures. *Journal of Computer-Aided Molecular Design, 26*(2), 169-183. doi: 10.1007/s10822-011-9538-6

Malde, A. K., & Mark, A. E. (2011). Challenges in the determination of the binding modes of non-standard ligands in X-ray crystal complexes. *Journal of Computer-Aided Molecular Design, 25*(1), 1-12. doi: 10.1007/s10822-010-9397-6

Malinina, L., Malakhova, M. L., Kanack, A. T., Lu, M., Abagyan, R., Brown, R. E., & Patel, D. J. (2006). The liganding of glycolipid transfer protein is controlled by glycolipid acyl structure. *Plos Biology, 4*(11), 1996-2011. doi: 10.1371/journal.pbio.0040362

Mir, R., Singh, N., Vikram, G., Kumar, R. P., Sinha, M., Bhushan, A., Kaur, P., Srinivasan, A., Sharma, S., & Singh, T. P. (2009). The Structural Basis for the Prevention of Nonsteroidal Antiinflammatory Drug-Induced Gastrointestinal Tract Damage by the C-Lobe of Bovine Colostrum Lactoferrin. *Biophysical Journal, 97*(12), 3178-3186. doi: 10.1016/j.bpj.2009.09.030

Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., & Olson, A. J. (2009). AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *Journal of Computational Chemistry, 30*(16), 2785-2791. doi: 10.1002/jcc.21256

Murshudov, G. N., Skubak, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F., & Vagin, A. A. (2011). REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallographica Section D-Biological Crystallography, 67*, 355-367. doi: 10.1107/s0907444911001314

Nissink, J. W. M., Murray, C., Hartshorn, M., Verdonk, M. L., Cole, J. C., & Taylor, R. (2002). A new test set for validating predictions of protein-ligand interaction. *Proteins-Structure Function and Genetics, 49*(4), 457-471. doi: 10.1002/prot.10232

Pozharski, E., Weichenberger, C. X., & Rupp, B. (2013). Techniques, tools and best practices for ligand electron-density analysis and results from their application to deposited crystal structures. *Acta Crystallographica Section D-Biological Crystallography, 69*, 150-167. doi: 10.1107/s0907444912044423

Read, R. J., Adams, P. D., Arendall, W. B., Brunger, A. T., Emsley, P., Joosten, R. P., Kleywegt, G. J., Krissinel, E. B., Lutteke, T., Otwinowski, Z., Perrakis, A., Richardson, J. S., Sheffler, W. H., Smith, J. L., Tickle, I. J., Vriend, G., & Zwart, P. H. (2011). A New Generation of Crystallographic Validation Tools for the Protein Data Bank. *Structure, 19*(10), 1395-1412. doi: 10.1016/j.str.2011.08.006

Reynolds, C. H. (2014). Protein-Ligand Cocrystal Structures: We Can Do Better. *Acs Medicinal Chemistry Letters, 5*(7), 727-729. doi: 10.1021/ml500220a

Rupp, B. (2010). *Biomolecular crystallography : principles, practice, and application to structural biology*. New York: Garland Science.

Scapin, G., Patel, S. B., Lisnock, J., Becker, J. W., & LoGrasso, P. V. (2003). The structure of JNK3 in complex with small molecule inhibitors: Structural basis for potency and selectivity. *Chemistry & Biology, 10*(8), 705-712. doi: 10.1016/s1074-5521(03)00159-5

Sharff, A., Keller, P., Vonrhein, C., Smart, O., Womack, T., Flensburg, C., Paciorek, W., & Bricogne, G. (2014). Pipedream documentation, version 1.0.0. from http://www.globalphasing.com/buster/manual/pipedream/manual/index.html

Smart, O. S. (2014). Achieving high quality ligand chemistry in protein X-ray structures. Workshop prepared for the "Structural Basis of Pharmacology: Deeper Understanding of Drug Discovery through Crystallography", Meeting Erice June 2014., from http://grade.globalphasing.org/tut/erice_workshop/

Smart, O. S., Holstein, J., & Womack, T. (2014). Grade documentation. version 1.2.8., from http://www.globalphasing.com/buster/manual/grade/manual/index.html

Smart, O. S., Womack, T. O., Flensburg, C., Keller, P., Paciorek, W., Sharff, A., Vonrhein, C., & Bricogne, G. (2012). Exploiting structure similarity in refinement: automated NCS and target-structure restraints in BUSTER. *Acta Crystallographica Section D-Biological Crystallography, 68*, 368-380. doi: 10.1107/s0907444911056058

Strong, M., Sawaya, M. R., Wang, S. S., Phillips, M., Cascio, D., & Eisenberg, D. (2006). Toward the structural genomics of complexes: Crystal structure of a PE/PPE protein complex from Mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences of the United States of America, 103*(21), 8060-8065. doi: 10.1073/pnas.0602606103

Terwilliger, T. C., & Bricogne, G. (2014). Continuous mutual improvement of macromolecular structure models in the PDB and of X-ray crystallographic software: the dual role of deposited experimental data. *Acta Crystallogr D Biol Crystallogr, 70*(Pt 10), 2533-2543. doi: 10.1107/S1399004714017040

Vonrhein, C., Flensburg, C., Keller, P., Sharff, A., Smart, O., Paciorek, W., Womack, T., & Bricogne, G. (2011). Data processing and analysis with the autoPROC toolbox. *Acta Crystallographica Section D-Biological Crystallography, 67*, 293-302. doi: 10.1107/s0907444911007773

Warren, G. L., Do, T. D., Kelley, B. P., Nicholls, A., & Warren, S. D. (2012). Essential considerations for using protein-ligand structures in drug discovery. *Drug Discovery Today, 17*(23-24), 1270-1281. doi: 10.1016/j.drudis.2012.06.011

Wlodek, S., Skillman, A. G., & Nicholls, A. (2006). Automated ligand placement and refinement with a combined force field and shape potential. *Acta Crystallographica Section D-Biological Crystallography, 62*, 741-749. doi: 10.1107/s0907444906016076