Acta Crystallographica Section D
**Biological Crystallography**
ISSN 0907-4449

# Techniques, tools and best practices for ligand electron-density analysis and results from their application to deposited crystal structures

**Edwin Pozharski,[a]* Christian X. Weichenberger[b] and Bernhard Rupp[c]***

[a]Department of Pharmaceutical Sciences, University of Maryland, Baltimore, Maryland, USA, [b]Center for Biomedicine, European Academy of Bozen/Bolzano (EURAC), Viale Druso 1, I-39100 Bozen/Bolzano, Italy, and [c]k.-k. Hofkristallamt, 991 Audrey Place, Vista, CA 92084, USA

Correspondence e-mail:
epozh001@umaryland.edu,
br@hofkristallamt.org

As a result of substantial instrumental automation and the continuing improvement of software, crystallographic studies of biomolecules are conducted by non-experts in increasing numbers. While improved validation almost ensures that major mistakes in the protein part of structure models are exceedingly rare, in ligand–protein complex structures, which in general are most interesting to the scientist, ambiguous ligand electron density is often difficult to interpret and the modelled ligands are generally more difficult to properly validate. Here, (i) the primary technical reasons and potential human factors leading to problems in ligand structure models are presented; (ii) the most common categories of building errors or overinterpretation are classified; (iii) a few instructive and specific examples are discussed in detail, including an electron-density-based analysis of ligand structures that do not contain any ligands; (iv) means of avoiding such mistakes are suggested and the implications for database validity are discussed and (v) a user-friendly software tool that allows non-expert users to conveniently inspect ligand density is provided.

Received 21 August 2012
Accepted 26 October 2012

'The human understanding is not composed of dry light, but is subject to influence from the will and the emotions, a fact that creates fanciful knowledge; man prefers to believe what he wants to be true . . . for what man had rather were true he more readily believes', Francis Bacon, *Novum Organum Scientiarum*, Aphorism 49 (1620).

'The author can be excused of dishonesty only on the grounds that before deceiving others he has taken great pains to deceive himself',
P. Medawar (1961)

## 1. Introduction

The extraordinary efforts of instrumentation and software developers have led to outstanding advances in automation, allowing non-experts to conduct previously daunting crystallographic studies of biomacromolecules with relative ease, resulting in accurate structure models that can be reliably interpreted in their biological context. Integrated validation procedures embedded in, and concurrent with, model building and refinement, and ultimately applied on deposition in the Protein Data Bank (PDB; Berman, 2008), have with few exceptions almost eliminated the occurrence of completely wrong or seriously flawed protein-structure models (Read *et al.*, 2011). The situation is not quite as encouraging as far as protein–ligand structures are concerned. As this investigation demonstrates, there are far too many protein–ligand structures in the PDB which either (i) clearly do not contain the purported ligand, (ii) provide only insufficient crystallographic evidence that such a ligand might be present or (iii) present an incorrect description of the ligand. In the following introductory subsections we examine the primary technical reasons leading to the problem of questionable protein–ligand complex structures and (admittedly more hesitantly) speculate that human factors may contribute to and amplify the problem.

Attempts to improve upon already deposited structure models as a result of the welcome enhancements of refinement programs and model parameterization have been conducted with success before (Joosten *et al.*, 2011), and ligand conformations have been data mined to detect implausible geometry (Kleywegt & Harris, 2007). The primary purpose of our publication is to raise awareness of the problem in general and to provide guidance to non-expert crystallographers as to how to critically examine the actual crystallographic evidence, primarily in the form of supporting electron density. In a few cases of obvious misplacement we were able to suggest specific corrections to the affected ligand structures, while in cases where there is no electron density there is nothing that can be corrected or improved. There is always the possibility that incorrect structure factors of apo structures or other ligand-free crystals have inadvertently been deposited, and in any of these cases we encourage the deposition of proper experimental structure factors that support electron density for the ligands in question.

From a top view, certain classes of common difficulties in the accurate interpretation of ligand density and the description of ligand models can be distinguished (§3). Given the clear indications that numerous structure models are not reliable as far as their ligand component is concerned, we provide suggestions as to how to present difficult situations, aiming to reduce the potential for misinterpretation, and provide a tool that allows even non-experts to quickly display ligand density and to annotate their own interpretation of it. It seems to be important to raise awareness of the inherent difficulties in the interpretation of ligand density, particularly with non-expert crystallographers, reviewers and journal editors. In addition, proper evaluation of ligand structures is particularly important to assure the integrity and usefulness of the publicly available structure databases.

### 1.1. Technical reasons for the difficulty in obtaining and modelling protein–ligand complex structures

**1.1.1. *R* values and target geometry are poor indicators of ligand model quality**. A high-resolution X-ray structure model of a protein–ligand complex generally provides more information (and associated impact) than a ligand-free (apo) structure. Most commonly, the ligand of interest (not necessarily the ligand actually bound) is a small-molecule moiety, often a drug lead molecule or a nonreactive substrate or transition-state analog. The scattering mass of such a ligand molecule of several 100 Da compares with the protein target mass of perhaps several tens to hundreds of kilodaltons and even at full occupancy the contribution of the ligand to the total scattering is in such cases only about 1/100 to 1/1000 of that of the protein partner. One immediately realises that the absolute values of global quality measures based on a linear residual between observed and calculated structure-factor amplitudes (that is, the *R* values) of the form

$$R = \frac{\sum_{\mathbf{h}} |F_{\rm obs} - kF_{\rm calc}|}{\sum_{\mathbf{h}} F_{\rm obs}} \qquad (1)$$

will not clearly indicate whether or not a ligand is present, nor whether such a ligand has been properly placed, modelled or refined. $F_{\rm obs}$ ($F_{\rm o}$) are the structure-factor amplitudes (scattering contributions) of each reflection from the actual diffraction data and $F_{\rm calc}$ ($F_{\rm c}$) are the structure-factor amplitudes computed from the model.

As expected from the small differences in the scattering mass of the protein alone *versus* the protein–ligand complex, the relative change in the *R* values of a refinement with and without a ligand is correspondingly small. In the example illustrated in §3.1, the *R* and $R_{\rm free}$ values for the ligand–free protein model are 17.35% and 20.18%, respectively, while for the deposited model (with all five sugar moieties included) the *R* values are 17.32% and 20.10%, respectively. For the correct model, with the last three sugar moieties (which show no density) removed, the *R* values fall to 17.31% and 19.96%, respectively. On an absolute scale, all three values are good and publishable. Although the correct model following the rule of parsimony, namely to model only what one can actually see in density, has the lowest *R* values, even a (nontrivial) Hamilton-type significance test (Hamilton, 1965; Merritt, 2012) based on *R* values may not provide a unique or decisive answer as to which model is better based on the global *R*-value statistic. The local evidence in the form of the electron density, however, distinctly and clearly favours the parsimonious model (*cf.* Fig. 3).

In addition, none of the commonly cited stereochemistry quality indicators for the protein part of the structure such as plausible backbone torsion-angle distribution (Ramachandran plot; Kleywegt & Jones, 1996) or the absence of any major deviations from stereochemical geometry target values (Engh & Huber, 2001) will indicate anything of relevance for quality assessment of the ligand part of the structure.

**1.1.2. Ligand binding is always incomplete**. A further complication arises from the fact that the occupancy of the

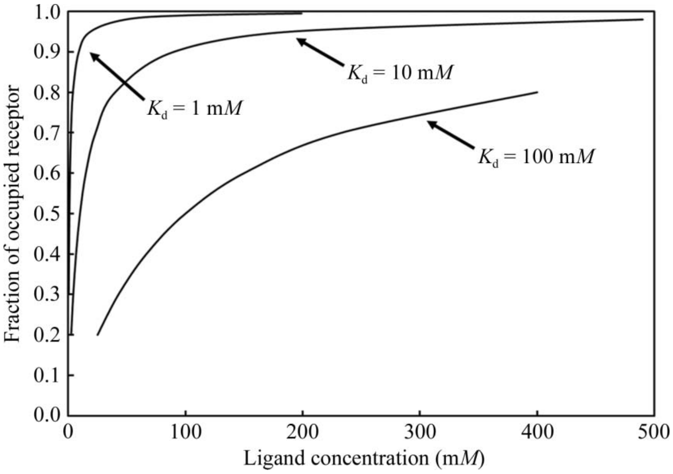


**Figure 1**
The fraction of occupied receptor sites plotted against ligand equilibrium concentration for three different binding constants. While in the millimolar and lower $K_d$ range small concentrations of ligand suffice to achieve reasonable binding-site occupancy (between 70 and 90%), quite impractical concentrations of ligand in the crystallization drop are required for poor binders. On the other hand, given a sufficiently high concentration, even weakly binding (and non-native) ligands can be forced into a binding site. From Rupp (2009), reproduced with the permission of Garland Science.

ligand site is a direct function of the binding affinity and ligand concentration. This can be simply derived from the definition of the dissociation constant $K_d$, which has been illustrated in great detail in this journal (Danley, 2006) and is reviewed in basic textbooks (Rupp, 2009). The simple fact, as illustrated in Fig. 1, is that if a ligand does not have a high binding affinity it will not have full occupancy and therefore even less of its scattering mass will contribute to the global refinement residuals. Similarly, any contributions to the corresponding ligand electron density will be reduced in proportion to the ligand occupancy.

The increase of occupancy with increasing ligand concentration, on the other hand, means that anything that is present in the crystallization mother liquor at a high enough concentration, whether a native ligand or not, may partly occupy the binding site. Binding sites have, by nature, evolved to attract ligand moieties, and while specifics assure that *in vivo* the correct substrate is processed, even remotely similar molecular moieties (*i.e.* anything from expression host cellular contents to purification buffers to crystallization-cocktail components) can be forced by high concentration into a binding site (and even partly replace or entirely compete out the desired ligand). This can be used to advantage in fragment-based drug lead discovery (Burley, 2004; Hajduk & Greer, 2007), but it also can lead to unexpected ligands such as buffers in the binding site (Gokulan *et al.*, 2005) and, somewhat more insidiously, produce some kind of obscure partial density in the binding site that beckons to be filled with a ligand of desire (Bacon, 1620).

**1.1.3. $R_{free}$-set selection and model bias**. Protein–ligand complex structures are often determined by molecular replacement from already known protein-structure models, and in the case of isomorphous structures simple rigid-body refinement followed by rebuilding and individual coordinate refinement may suffice. In such cases, the same reflections as selected for the $R_{free}$ set of the original data set should be kept for proper cross-validation (Brünger, 1997). In addition, if another isomorphous structure with a ligand has been used for re-refinement, spurious density resembling the original ligand might be reproduced as a result of model phase bias. Although modern maximum-likelihood methods are relatively robust against model bias, this possibility should be kept in mind. An initial round of simulated-annealing molecular-dynamics refinement (Adams *et al.*, 1997) with the ligand omitted can be used to eliminate bias issues if these are suspected.

Regarding model bias, it is also important to note that the EDS electron density (Kleywegt *et al.*, 2004) used in the evaluation and ranking of ligands by our *Twilight* script described in §2 is not ligand-omit density and therefore biases the density towards the presence of a ligand rather than its absence.

### 1.2. Primary evidence: electron density

The protein structure model, as interpreted by the crystallographer from the electron density and deposited in the PDB, is in principle nothing more than a listing of the Cartesian coordinates of all located atom positions, including a probabilistic measure (the *B* factor) indicating the amplitude of the variation in atomic position throughout the crystal lattice (which absorbs all types of contributions). This model is the end result of repeated cycles of model building and refinement, and the electron density is the primary crystallographic evidence for the presence and location of the model atoms. The fit of the model to minimally biased electron density is therefore also the primary indicator of local model quality, including ligands. Various statistical measures exist to quantify and visualize the correspondence between model and electron density, primarily the real-space *R* value (RSR), the real-space correlation coefficient (RSCC) and difference density measures. These were introduced decades ago (Brändén & Jones, 1990), their usefulness has been repeatedly reiterated (Rupp, 2006) and they are publicly available for deposited PDB structures through the Uppsala Electron Density Server (EDS; Kleywegt *et al.*, 2004). Despite their undisputed practicality for real-space model validation, the RSR and the RSCC have the flaw of not distinguishing between model accuracy and model precision. Newly developed statistical measures for real-space validation (Tickle, 2012) can distinguish between local model accuracy and model precision, which ultimately depends on data quality. Real-space validation scores and other quality indicators have been summarized by Weiss & Einspahr (2011).

The electron density is commonly presented in the form of $\sigma_A$-derived maximum-likelihood (ML) maps (Read, 1986; Pannu & Read, 1996) with Fourier coefficients of the form $(2mF_o - DF_c)\exp(i\varphi)$ suitable for initial building, and difference density maps of $(mF_o - DF_c)\exp(i\varphi)$ best suited for model correction. Here, $m$ is the figure of merit (directly related to the mean phase angle uncertainty as $m = \langle\cos\Delta\varphi\rangle$, $D$ is the Luzzati factor (Luzzati, 1953) and $\varphi$ is the phase angle calculated from the model. The derivation and meaning of the ML coefficients have been summarized, for example, in Rupp (2009).

**1.2.1. Proper use of difference electron-density maps**. An $F_o - F_c$-type difference map of a ligand structure that actually contains a ligand will show distinct positive difference density for the omitted ligand, while a difference map calculated with a placed ligand where no ligand exists will show equally distinct negative difference density for the ligand. It is therefore imperative to declare in electron-density figures the type of (difference) density that is being displayed, the contour level, and the procedure through which the electron density was generated. As an example, without specifying the electron density in Fig. 9(*a*) accurately as negative difference density, it could be interpreted, particularly in a greyscale or black-and-white figure, by a non-expert as a $2mF_o - DF_c$ map actually demonstrating the presence of the last two (in fact absent) saccharide moieties. Care must be taken not to mistake clear negative difference density indicating an absent ligand for normal $2mF_o - DF_c$ density or positive difference density.

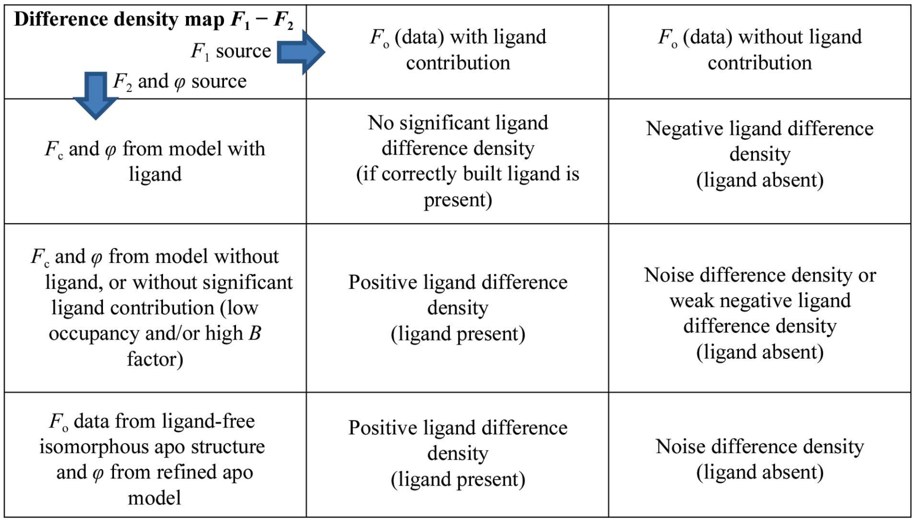
Alternatively, if an already refined isomorphous apo structure (a target protein without bound ligand) and data for the corresponding ligand-bound structure are available, an

$F_o - F_{o(apo)}$ difference map using model phases from the refined apo structure can be generated. Of importance in this case is that a high level of isomorphism is necessary, in general indicated (but not guaranteed) by a change in unit-cell parameters of only a few percent.

Difference density analysis is meaningful only when (i) either the model contributes to the calculated scattering factors $F_c$ and/or (ii) there is an actual experimental contribution of a ligand to the observed structure factors $F_o$. If the ligand model does not contribute to $F_c$ and the $F_o$ data do not contain any ligand contribution, then a Fourier synthesis based on the $F_o - F_c$ difference becomes a noise-level subtraction and meaningless (Fig. 2). This is the case with either extreme $B$ factors refined for an absent ligand and/or when using very low (partial) ligand occupancies.

**1.2.2. Contouring of electron-density maps**. In the search for a desired ligand, it may be tempting to contour the electron density down until some noise features start to appear in a binding site and to place the ligand in some plausible pose into the binding site. Occasionally, a $2mF_o - DF_c$ density figure with a reasonable appearance may even be produced, but the absence of clear positive omit difference density will serve well as cross-validation. Noise density levels are reached in normal $2mF_o - DF_c$ maps below approximately 0.7$\sigma$, but in very clear (omit) maps of quality models obtained from excellent data inspection at lower levels may be useful.

While the choice of contouring level in the $2mF_o - DF_c$ maps that one considers to be acceptable as evidence of ligand binding is somewhat flexible (*e.g.* it may depend on the data/model quality and the solvent content), the interpretation of difference maps ($mF_o - DF_c$) for the nearly final model is more rigid. With almost all of the ordered structural elements and bulk solvent accounted for, variation in the remaining electron density simply reflects the noise level in the underlying data. In practice, the 3$\sigma$ level is generally accepted, in part because it is the default contouring level in the popular display program *Coot* (Emsley *et al.*, 2010), as the initial point in difference map inspection. This means that one is expected to identify some electron-density 'blob' that is consistently above the 3$\sigma$ level as a starting point for ligand placement. In somewhat simplified terms, it is expected that individual peaks outside ±3$\sigma$ will appear on average for every 370 points of the grid on which the electron-density map is calculated. (It is assumed here that the noise in the difference density map obeys a normal distribution). Thus, on average a peak at the 3$\sigma$ level is expected for ~7 × 7 × 7 grid elements. For a 0.5 Å grid the average distance between such noise-level peaks is about 6 Å ($0.5 \times 7 \times 3^{1/2}$). This means that an occasional peak may be ignored, but extended 'blobs' in the difference density maps should be investigated. Importantly, the predicted distance between noise-level peaks rapidly decreases with lower contouring levels: it is only ~3.5 Å at 2.5$\sigma$ and ~1.75 Å at 2$\sigma$.



| **Difference density map $F_1 - F_2$**<br>$F_1$ source<br>$F_2$ and $\varphi$ source | $F_o$ (data) with ligand contribution | $F_o$ (data) without ligand contribution |
|---|---|---|
| $F_c$ and $\varphi$ from model with ligand | No significant ligand difference density (if correctly built ligand is present) | Negative ligand difference density (ligand absent) |
| $F_c$ and $\varphi$ from model without ligand, or without significant ligand contribution (low occupancy and/or high $B$ factor) | Positive ligand difference density (ligand present) | Noise difference density or weak negative ligand difference density (ligand absent) |
| $F_o$ data from ligand-free isomorphous apo structure and $\varphi$ from refined apo model | Positive ligand difference density (ligand present) | Noise difference density (ligand absent) |

**Figure 2**
Difference density map results with Fourier coefficients $(F_1 - F_2)\exp(i\varphi)$ based on the presence or absence of scattering contributions in the data and/or model. If neither an $F_o$ nor an $F_c$ contribution is present, the difference map becomes meaningless. Interpretable positive omit difference density (*i.e.* a difference density map calculated from the model with the ligand omitted) is therefore the primary evidence providing distinct proof for the presence of a ligand.

The electron-density maps published for the peptide ligand in PDB entry 1f83 (Hanson & Stevens, 2000) and discussed elsewhere (Rupp, 2010) provide instructive examples of how at first glance reasonable-looking electron-density maps can be obtained by cutting the noise-level density around model atoms by (ab)using the 'blob' features of model-building software (Rupp & Segelke, 2001).

### 1.3. Prior expectations: ligand geometry and contacts

In a Bayesian model of inductive inference (reviewed, for example, in Rupp, 2009, 2010), the likelihood of a model being correct can be expressed as a joint conditional probability of how well it reproduces the data (*e.g.* its match to the electron density) and how well it complies with independently acquired prior knowledge about its intrinsic properties (*e.g.* reasonable stereochemistry). Many ligand models suffer from improbable stereochemistry, largely as a result of improper or incomplete stereochemical restraints (Kleywegt & Harris, 2007; http://eds.bmc.uu.se/eds/valligurl.php). The observation of improbable conformations often coincides with the absence of restraining electron density or experimental scattering contributions. It is imperative that a ligand has reasonable stereochemistry and makes reasonable contacts to its binding partner, thus providing a strong positive and amplifying prior probability term to the likelihood that the ligand may exist, provided that clear primary evidence in the form of positive omit difference density is also present.

Proper ligand stereochemistry practically always requires correct geometry-restraint files for refinement. Several tools exist to generate proper restraint files such as the *Grade Server* (Smart *et al.*, 2011), *PRODRG* (Schüttelkopf & van Aalten, 2004), *JLigand* (Lebedev *et al.*, 2012) or the *PDB Ligand Expo* (http://ligand-expo.rcsb.org), which provides a collection of tools to access ligand structures already in PDB files. For the examination of ligand binding, visualization tools such as *LigPlot+* (Laskowski & Swindells, 2011) or the validation tools within the model-building program *Coot* can be used. The macromolecular program package *PHENIX* (Adams *et al.*, 2010) provides a module *eLBOW* (*electronic Ligand Builder and Optimization Workbench*) for proper restraint-file generation (Moriarty *et al.*, 2009).

### 1.4. Human factors: the ligands of desire

Basic scientific epistemology requires that a strong claim must be supported by equally convincing specific evidence and that the claim should not violate independently acquired and established prior knowledge (while at the same time, in exceptional cases of extraordinarily powerful evidence, even prior expectations or beliefs are subject to revision). This guiding principle of empirical inductive reasoning, which originally evolved during the Enlightenment in the 15th century (Bacon, 1620), was put into a framework of formal logic about a century later by Bayes (1763). Probabilistic Bayesian models were adopted early on in protein crystallography, ranging from first applications to intensity statistics (French, 1978), to overarching acceptance of a comprehensive probabilistic approach towards protein crystallography (Read, 1986; Bricogne, 1988, 1997; Rupp, 2009) and to many specific applications such as density modification (Terwilliger, 2003) and geometric restraint implementation (Roversi *et al.*, 2000).

A recent educational paper describes the use of a simplified qualitative Bayesian reasoning model allowing practitioners of crystallography to assess their beliefs and expectations against the (sometimes painful) necessity of balancing experimental evidence against desired outcomes (Rupp, 2010). The type of cognitive bias creating the tendency to find what one seeks and to ignore contradictory evidence (or the absence of evidence) is well documented in psychology literature as confirmation bias (see, for example, Koehler, 1993). An additional deep-rooted problem seems to be the resistance to correction of errors, as documented by the persistence of false positives in the scientific literature (Simmons *et al.*, 2011). Once a finding has passed review and is in print (or is in the PDB), it becomes very hard to correct.

The point to reiterate here is very simple: publications of protein–ligand complexes usually carry exciting fundamental information such as elucidating the mechanism of an enzymatic reaction, or they provide high-impact, also commercially valuable, information about drug–target interactions. This potential of significant intellectual and pecuniary rewards carries with it a responsibility to ensure that the asserted claim is sound by providing a valid protein–ligand structure model that is supported by crystallographic evidence; that is, (positive omit) electron density for the ligand.

In the following, we examine how well a small subset of ligand structures fare in view of the necessity for strong experimental evidence.

## 2. Identification and ranking of ligand structures

### 2.1. The *Twilight* script

The *Twilight* script used to identify, extract and rank ligand structures from the PDB is described in detail in an accompanying publication (Weichenberger *et al.*, 2013). *Twilight* allows the inspection of a sorted list of ligands ranked by real-space correlation (Brändén & Jones, 1990), the addition of comments for each single case analysed, and viewing of the three-dimensional protein–ligand complex and its associated electron-density maps with *Coot* or through direct links to the EDS (Kleywegt *et al.*, 2004). In addition, journal links, PMID codes and links to the PDBe are provided.

We analysed ligands including covalently bound sugars from glycosylations, but omitted any common buffer or solvent molecules (Weichenberger *et al.*, 2013). The program is freely available for download from http://www.ruppweb.org/twilight/. We encourage readers to install the tool and examine the evidence themselves in the form of electron density and provide annotations as to their personal preferences. 93 false positives were identified in 1259 ligands distributed in roughly 528 PDB entries. Given the possibility that some of the problematic PDB entries may result from the deposition of incorrect experimental data sets, authors of annotated entries are invited to examine the evidence and to use the opportunity to provide correct experimental data files that reproduce the appropriate ligand density.

### 2.2. Electron-density map calculations and density figures

If not stated otherwise, in the figures the model obtained directly from the PDB is shown with a $2mF_o - DF_c$ maximum-likelihood omit map contoured at $1\sigma$ (blue) plus the corresponding $mF_o - DF_c$ map contoured at $+3\sigma$ (green) and $-3\sigma$ (red). Omit maps were calculated by removing the ligands in question followed by refinement in *REFMAC* (Murshudov *et al.*, 2011). For the initial ligand inspection, electron-density maps were in most cases downloaded from the EDS (http://eds.bmc.uu.se/eds/; Kleywegt *et al.*, 2004). If unavailable *via* EDS, the map coefficients were calculated from the deposited model and structure factors without refinement but after the inclusion of bulk-solvent correction using *REFMAC* and were rendered in *PyMOL* (http://www.pymol.org). For electron-density inspection $2mF_o - DF_c$ map levels of $0.8\sigma$ were used and additional contouring down to noise level was used to detect (or exclude) minor contributions that might justify the proposed ligand poses. In all specific cases discussed below, electron-density maps were calculated using model phases calculated directly from the deposited models. Ligands were omitted from calculations as indicated in the captions.

## 3. Classification of problematic ligands

Visual inspection of the electron density of ligands flagged as outliers by the above scoring uncovered several common and typical situations revealing ligands whose presence, placement and/or refined structure appear to be questionable. While we discuss several examples in detail below, we emphasize that our findings and annotations are limited to the critique of crystallographic evidence for any given hypothesis and that other evidence may or may not exist that justifies the proposed biological hypotheses. As always, our annotations are our interpretation of the evidence, and naturally in view of spurious or ambiguous density interpretations may differ. However, in accordance with accepted scientific epistemology, strong evidence, which is required in support of strong claims such as a specific pose of a ligand, generally does not lead to ambiguous interpretations.

We wish to reiterate that the possibility of the deposition of incorrect structure factors or intermediate or incomplete structure models cannot be excluded in any of the discussed or annotated cases. In all such cases inadvertent errors can be readily corrected by the submission of structure factors and models that allow the reconstruction of electron density supporting the strong claim of the presence of the ligand in question in the specified pose(s).

The following classification of ligand structures emerged from analysis of the results of visual inspection. The frequency of occurrence is given in parentheses for every class.

(i) Ligands incorrectly identified as questionable (7.4%). Occasionally, a low RSCC is obtained for a ligand that presents a good fit to the electron density. It must be noted that this classification is somewhat subjective and that a more stringent analysis may place some of these cases into the third category described below. Other examples of 'false positives' include ultrahigh-resolution structures (for these, the RSCC presently reported by the EDS is more sensitive to minor shifts in atomic positions) and obvious deposition problems [*e.g.* PDB entry 2ny2 (Zhou *et al.*, 2007) contains several ligand molecules that are clearly present in electron density but for some reason have their corresponding *B* factors set to 200–300 Å$^2$]. In addition, EDS density is not ligand-omit density and therefore rather biases the density in favor of the presence of a ligand instead of its absence. Some weak $2mF_o - DF_c$ density at levels of about $0.6\sigma$ in clean high-resolution EDS maps may in fact be the result of model bias.

(ii) Incorrectly modelled ligands (5.2%). While electron density is present and appears to resemble the actual ligand, the latter is not correctly placed.

(iii) Ligands with partially missing density (29.2%). Part of the ligand molecule is supported by the electron density, but a significant fraction of the total number of atoms appear to be missing.

(iv) Glycosylation sites (31.3%). Entries falling into this class could also be classified as (iii) or (vii), but we separate them because of the high frequency of their occurrence.

(v) Ligands placed into electron density that is likely to originate from mother-liquor components (10.4%). In these cases, the electron-density 'blob' is clearly present but is not easily interpretable.

(vi) Incorrect ligand (4.7%). In some cases, the electron density clearly resembles a commonly encountered mother-liquor or stock component (*e.g.* glycerol, sulfate, buffer molecules, polyethylene glycol chains *etc.*).

(vii) Ligands that are entirely unjustified by the electron density (11.9%). No electron density is observed above the noise level in the area where the ligand is placed.

We discuss examples of these specific classes in more detail below.

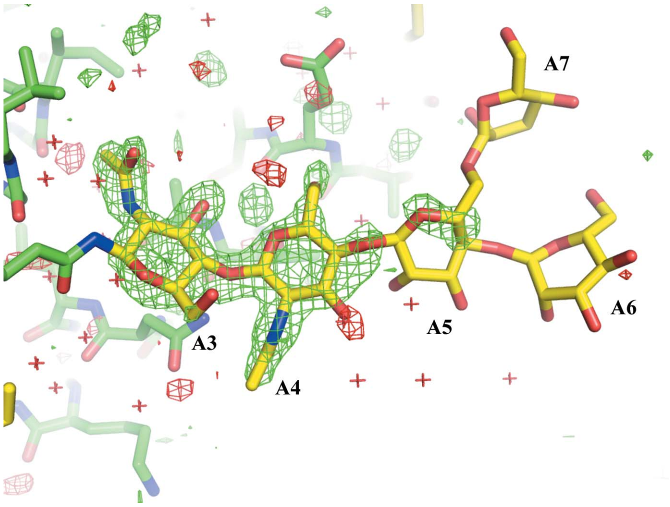


**Figure 3**
Missing density: extended glycosylations. The specific conformation of the last three α-D-mannose moieties (A5–A7) of the extended branched glycosylation in PDB entry 3ib0 (Mir *et al.*, 2009) is unsupported by electron density in the structure of bovine lactotransferrin, while the first two sugar moieties (A3–A4) are clearly present. The 1.4 Å resolution $mF_o - DF_c$ map contoured at $+3\sigma$ (green) was calculated after refining a model omitting the sugar moieties of the glycosylation site.

### 3.1. Glycosylation sites

Covalently attached sugar moieties are the most common form of post-translational modification of protein molecules. Anecdotal evidence suggests that these decorations impede protein crystallization, thus making structures of glycosylated proteins more difficult to obtain. Positive results with (conformationally homogenous) glycosylations confirm that, on the other hand, they can play a role in the formation of crystal contacts (Garcia *et al.*, 1996; Fusetti *et al.*, 2002). With respect to the refinement of such models, it should be expected that in most cases the sugar moieties will be disordered since they do not generally form extended specific interactions with the protein (except for the covalent linkage) or specific crystal contacts to symmetry mates in the crystal. Hence, there is a great potential that the corresponding electron density will become progressively featureless, decreasing in level and clarity with increasing distance from the linkage site and consequently making the exact placement of an extended single conformation problematic. In addition to such occurrences, where the glycosylation site appears to be disordered to some extent yet present, in a significant number of cases the extended modelled carbohydrate chains were obviously not present and are characterized by an absence of positive omit difference density in the difference density map (Fig. 3).

### 3.2. Lipid and detergent molecules

Given the hydrophobic nature of the transmembrane surface of membrane proteins, lipid/detergent molecules are routinely added to the crystallization medium to make them soluble and to promote crystallization. It is expected that these additives will be present in the crystal structure. Yet, given the lack of specific interactions with the protein, these moieties rarely show completely interpretable electron density. Often only a part of the lipid/detergent molecule reveals any density, yet it is added to the model in its entirety in a random orientation that can be misleading when inspected outside the context of displayed electron density (Fig. 4).

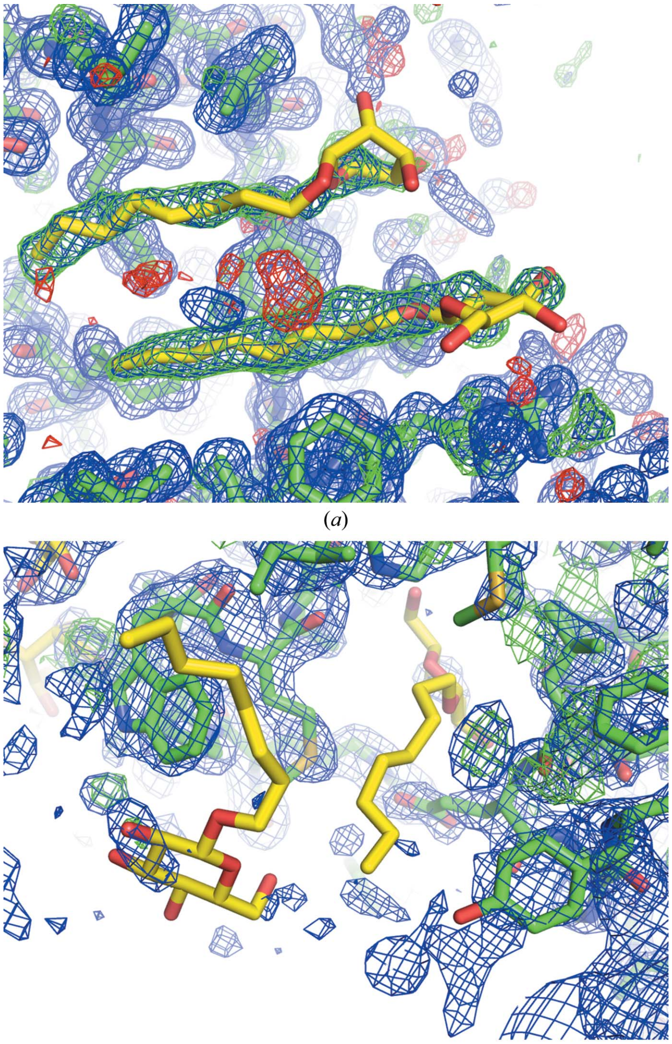


**Figure 4**
Missing density: detergents. Two examples of detergent molecules placed into models of membrane proteins. (*a*) The plant SLAC1 anion channel structure, PDB entry 3m73 (Chen *et al.*, 2010), shows two molecules (BOG A317/A318) that have clear density for the hydrophobic acyl chain but not for the head groups. (*b*) The presence of detergent molecules (BOG A700/A801 is shown) in the crystal structure of the *Escherichia coli* membrane enzyme glycerol-3-phosphate dehydrogenase, PDB entry 2qcu (Yeh *et al.*, 2008), appears to be entirely unsupported by the electron density. The 1.15 Å (*a*) and 1.75 Å (*b*) resolution $mF_o - DF_c$ maps contoured at $\pm 3\sigma$ (green/red) and $2mF_o - DF_c$ maximum-likelihood omit maps contoured at $1\sigma$ (blue) are shown. Maps were calculated for a refined model with the ligand atoms omitted.

### 3.3. Partially disordered ligands

A significant fraction of the identified problematic ligand molecules in crystal structures deposited in the PDB appear to have at least some part of the ligand placed in density that is in fact consistent with the chemical structure of the ligand molecule, yet the remainder of the ligand is not supported by electron density (Fig. 5). In some cases there is strong evidence that the problem arises from partial disorder, similar to the cases in which side chains of protein molecules or extended glycosylations appear to be missing in electron density. Ligand molecules may also break into fragments owing to chemical degradation, but crystallographic evidence alone is insufficient to distinguish such cases from partial disorder. Irrespective of the actual cause of the absence of ligand density, presenting the entire ligand molecule in one specific conformation is misleading unless the consumer of the structure has the knowledge and skill to interpret the likely elevated *B* factors or to directly inspect the electron-density maps. In practice, none of this seems to be routinely performed, thus making it highly desirable that a publicly available database is created that contains a list of ligands in protein structures that require further inspection.

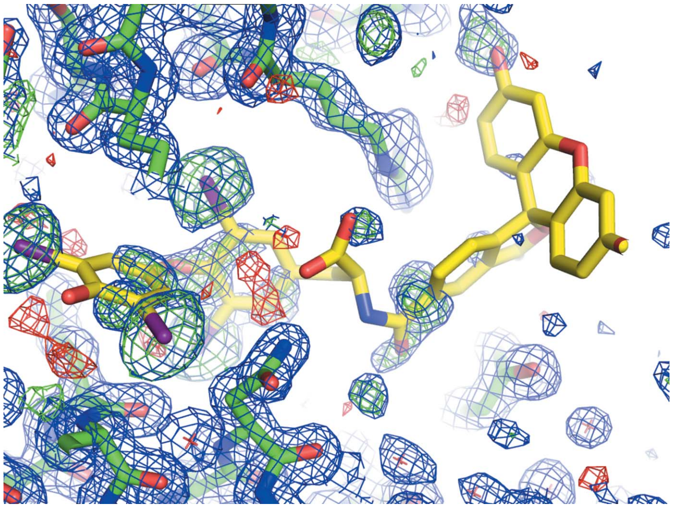


**Figure 5**
Partially disordered ligand. The fluorescein moiety of the ligand molecule (F6Z A1356) is missing in the electron density of the thyroxine-binding globulin, PDB entry 2xn7 (Qi *et al.*, 2011), even in $0.4\sigma$ noise-level $2mF_o - DF_c$ density. The 1.43 Å resolution $mF_o - DF_c$ map contoured at $\pm 3\sigma$ (green/red) and $2mF_o - DF_c$ maximum-likelihood omit map contoured at $1\sigma$ (blue) are shown. Maps were calculated for a refined model with the ligand atoms omitted.

### 3.4. Ligands placed in uninterpretable density likely to originate from mother-liquor components

The vast majority of proteins only crystallize in the presence of a highly concentrated precipitant cocktail. A high concentration of the cryoprotecting agent can be another source of unintended ligands. Consequently, components of the crystallization cocktail are often the source of some electron density that is visible in a known binding site or elsewhere. Occasionally, specific interactions are formed in these sites and the identity of the unexpected ligand is easily revealed (Gokulan *et al.*, 2005). In the case of unexpected ligands that are disordered and appear in or near the predicted target binding sites, it may be rather tempting to place the ligand of interest in an arbitrary or even a plausible pose into such uninterpretable density (Fig. 6). The poor fit may then be explained by invoking the possibility that the ligand binds in multiple conformations, as detailed in §3.3. For instance, in PDB entry 3qd1 (Pyburn *et al.*, 2011) a disaccharide was positioned into difference density that can be readily identified as originating from a HEPES molecule (Yves Muller, personal communication). The resulting misinterpretations and the conclusions drawn from them may make significant portions of the corresponding publications invalid.

In our analysis, we found that a significant fraction of the problematic ligands belonged to the class of misinterpreted crystallization-cocktail components. Naturally, such a classification is somewhat subjective and we used our best judgment to recognize the familiar patterns of common cocktail components in electron density, examples of which include

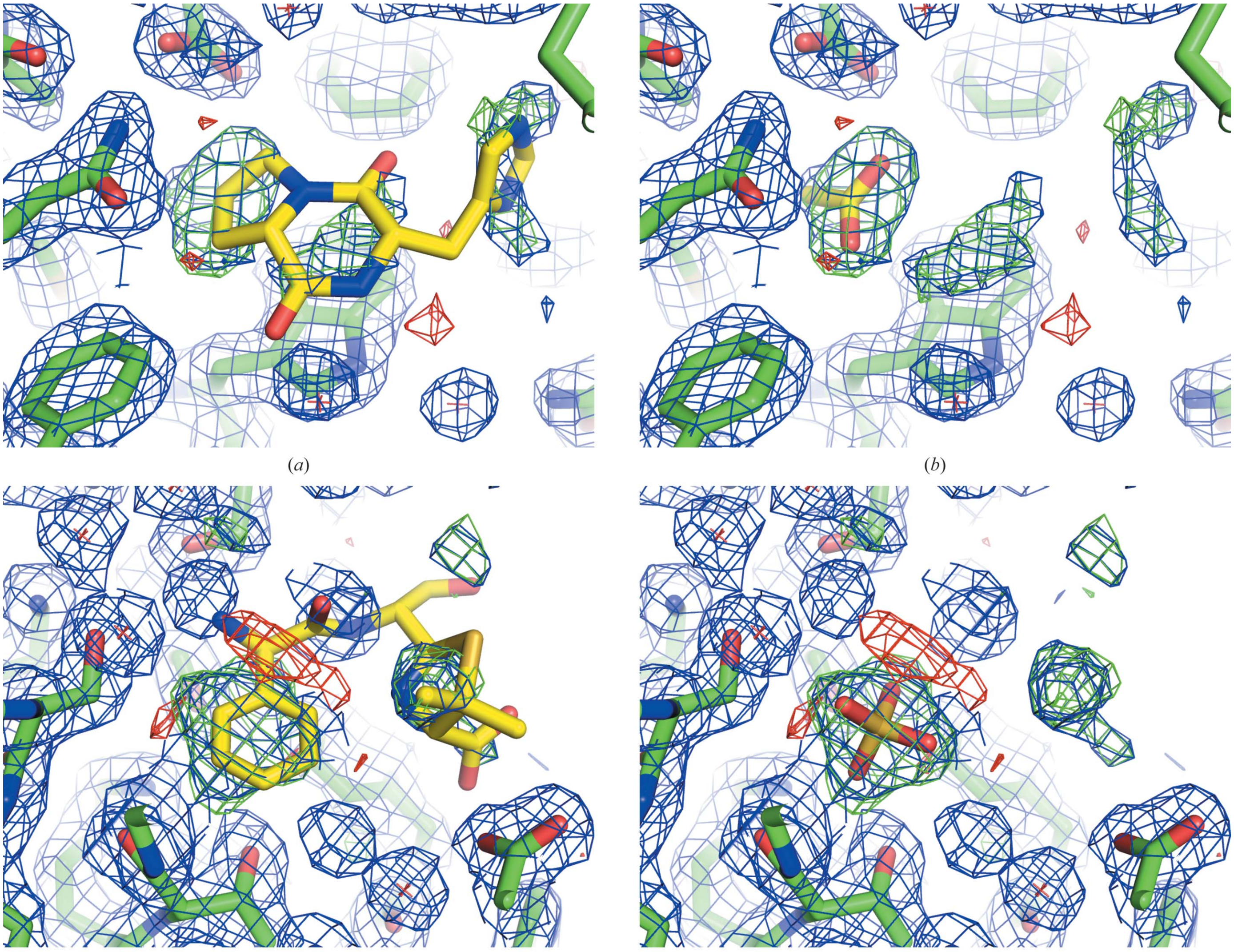


**Figure 6**
Ligands placed into mother-liquor density. In the structure of *Bacillus cereus* chitinase, PDB entry 3n1a (Hsieh *et al.*, 2010), the cyclo-(L-His-L-Pro) molecule (CHQ A1514) is placed into low-level electron density that is difficult to interpret (*a*) and which may be plausibly interpreted as an acetate molecule present in the crystallization cocktail at 200 m*M* supported by a newly formed hydrogen bond between Asp143 and the suggested acetate (*b*). In the structure of penicillin-binding protein 4 from *Staphylococcus aureus*, PDB entry 3hun (Navratna *et al.*, 2010), the phenyl moiety of the ampicillin (ZZ7 B501) is placed in a region of the electron density that based on difference density analysis could be better interpreted as a sulfate ion (*c*). The re-refined model that includes sulfate ion is shown in (*d*). The 2.0 Å resolution $mF_o - DF_c$ maps contoured at $\pm 3\sigma$ (green/red) and $2mF_o - DF_c$ maximum-likelihood omit maps contoured at $1\sigma$ (blue) are shown. Maps were calculated for a refined model with the ligand atoms omitted.

glycerol molecules, sulfate ions, short polyethylene glycol chains *etc.* (Fig. 6).

### 3.5. Incorrectly modelled ligands

In a minority of cases, we identified problematic structures in which the correct ligand was inaccurately placed. These cases are easy to correct using modern model-building software (Fig. 7).

### 3.6. Incorrectly identified ligands

Occasionally, the electron density surrounding the ligand was well defined but obviously represented a different ligand. In most cases, this appears to be owing to incorrectly identified components of the crystallization cocktail. Many of these cases constitute incorrect interpretation of the nonspecific interactions with the off-target solute at an off-target site and are thus generally inconsequential to the conclusions derived from the corresponding structures. For the sake of brevity, we do not provide any specific examples here. Readers can readily annotate such entries in the ligand table distributed with *Twilight*.

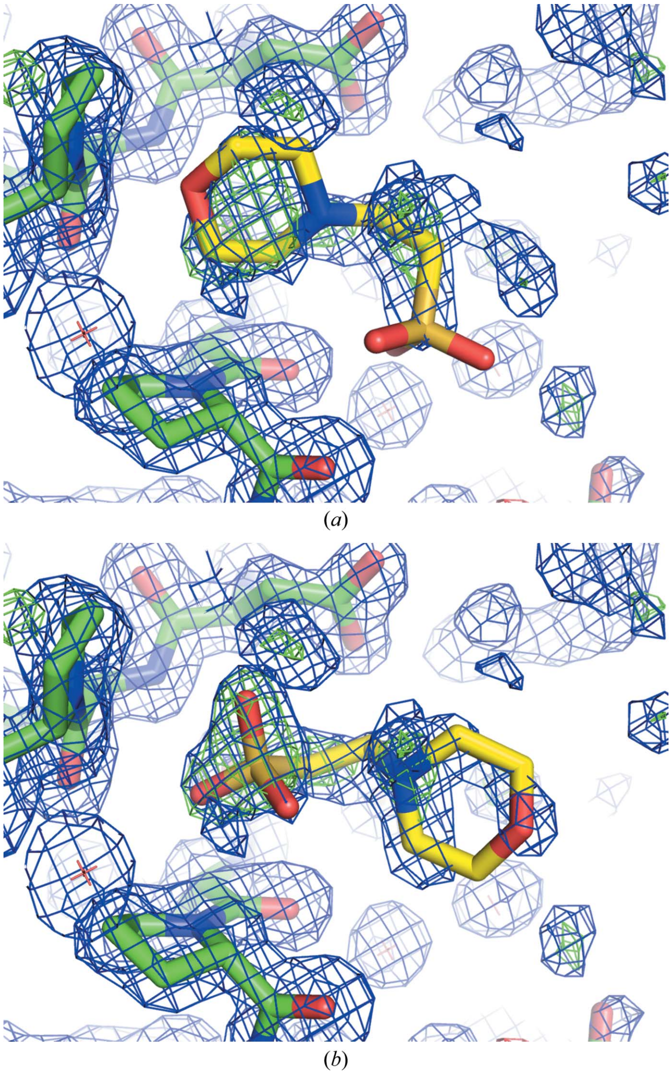


**Figure 7**
Misplaced but correct ligand. In the structure of *Streptomyces coelicolor* cytochrome P450 158A2, PDB entry 1se6 (Zhao *et al.*, 2005), the MES A632 molecule is correctly identified but is placed in the electron density in an incorrect pose (*a*), as clearly confirmed by manual rebuilding and real-space refinement in *Coot* after rotating the buffer molecule by 180° (*b*). The 1.75 Å resolution $mF_o - DF_c$ maps contoured at $\pm 3\sigma$ (green/red) and $2mF_o - DF_c$ maximum-likelihood omit maps contoured at $1\sigma$ (blue) are shown. Maps were calculated for a refined model with the ligand atoms omitted.

### 3.7. Entirely unjustified ligands

With alarmingly high frequency, ligands appeared to have been placed into the deposited structural models without any justification through evidence in the form of electron density (Fig. 8). Technically, all of the scenarios discussed above in which the ligands are (incorrectly) placed into the electron density may be considered in the same category of unjustified ligands. We separate the cases discussed here because with the possible exception of the deposition of wrong structure factors or incorrect model coordinates (which could and should be easily corrected) one can hardly find any explanation whereby some common and understandable error in electron-density interpretation might have led to these structures. While it is expected that the electron density for some ligand molecules that are either partially occupied, disordered or degraded either chemically or by radiation will be weak and/or imperfect, we find it nearly impossible to justify a model that produces neither $2mF_o - DF_c$ electron density nor positive omit difference density above the noise level. Disturbingly, in several cases that we present in the next section it is claimed in the corresponding publications that clear electron density was

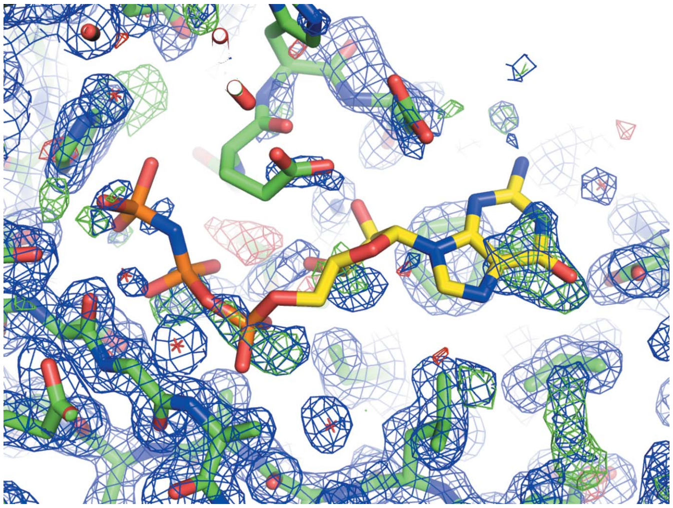
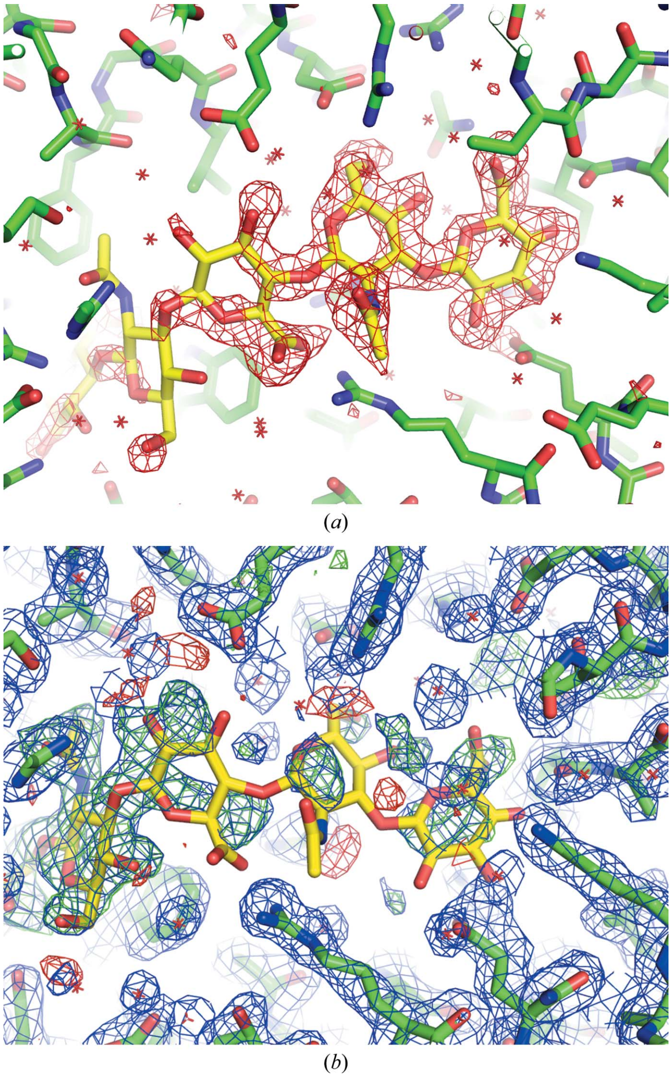


**Figure 8**
Absent ligand density in the omit map. In the structure of the Nudix hydrolase DR1025 from *Deinococcus radiodurans*, PDB entry 1sz3 (Ranatunga *et al.*, 2004), the nonhydrolyzable GDP analogue (GNP 3030A) is placed in a conformation and position entirely unsubstantiated by $2mF_o - DF_c$ electron density. The 1.6 Å resolution $mF_o - DF_c$ map contoured at $\pm 3\sigma$ (green/red) and $2mF_o - DF_c$ maximum-likelihood omit map contoured at $1\sigma$ (blue) are shown. Maps were calculated for a refined model with the ligand atoms omitted.

visible in difference maps, often accompanied by a figure presentation.

## 4. Case studies

In contrast to the previous categories, in which the presence or absence of a ligand does not have much relevance to the underlying biology, the following case studies involve multiple ligand structures intended to support biological hypotheses. Again, we argue that, depending on the specific case, the crystallographic evidence for these hypotheses is weak if not absent. We are not commenting on other evidence for or against the proposed biological relevance.


(*a*)

(*b*)

**Figure 9**
Negative difference density. The difference density map from the EDS (*a*) shows negative density at the $-3\sigma$ level (indicating absence of the model) for saccharide units 5–6 (HA3) in PDB entry 1loh (Jedrzejas *et al.*, 2002). Saccharide unit 4 of HA2 is present but is partly displaced from density owing to the incorrect placement of the subsequent units 5 and 6. Omit maps (*b*) provide no electron density consistent with placement of the two HA3 units. The 2.0 Å resolution $mF_o - DF_c$ map contoured at $\pm3\sigma$ (green/red) and $2mF_o - DF_c$ maximum-likelihood omit map contoured at $1\sigma$ (blue) are shown. Maps were calculated for a refined model with the ligand atoms omitted.

### 4.1. Hyaluronate lyases from *Streptococcus pneumoniae* and *S. agalactiae*

A series of four articles report the crystal structures of two enzymes in complex with modelled substrates, including disaccharides, tetrasaccharides and hexasaccharides. The first paper describes the structure determination of the complex of the Y408F mutant of *S. pneumoniae* hyaluronate lyase with tetrasaccharide and hexasaccharide substrates (Jedrzejas *et al.*, 2002). One of the structure models included in this study, namely that of the hexasaccharide complex, shows positive omit difference electron density that largely corresponds to the tetrasaccharide. The authors label the consecutive disaccharides HA1, HA2 and HA3, and there is no electron density supporting the positioning and/or presence of the HA3 unit. Upon subsequent refinement of the model containing the full hexasaccharide, the *B* factors of the atoms corresponding to the HA3 unit increase to values of higher than 100 Å$^2$, while those for HA1/HA2 remain around 30 Å$^2$. Combined with the negative difference electron density observed for the HA3 unit (Fig. 9), this leaves no doubt that the hexasaccharide model is erroneous. Nevertheless, a $2F_o - F_c$ electron-density figure contoured at the $1\sigma$ level without specification of Fourier coefficients is provided in the publication (Jedrzejas *et al.*, 2002). The hexasaccharide in the deposited model has low *B* factors, indicating that it has not been properly refined because the absence of electron density for the saccharide moieties 5–6 is not reflected in the expected increase in *B* factors for these absent parts. The expected significantly higher *B* factors ($\sim$100 Å$^2$) are in fact observed upon proper refinement of the complex structure. Saccharide unit 4 shows weak density but is misplaced, while units 5 and 6 show no distinct density (Fig. 9).

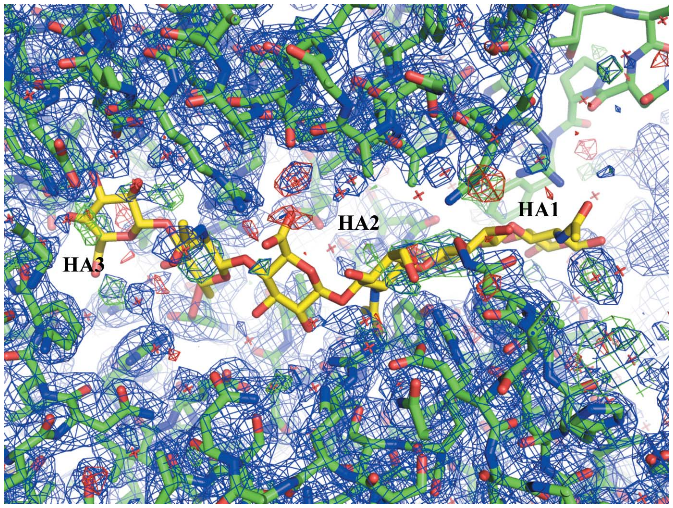


**Figure 10**
Missing ligand density, ligand omit map. The active site of *S. pneumoniae* hyaluronate lyase shows no meaningful density for the entire hexasaccharide molecule in the structure (PDB entry 1n7q; Nukui *et al.*, 2003). Disaccharide unit labels are consistent with the original publication. The 2.3 Å resolution $mF_o - DF_c$ map contoured at $\pm3\sigma$ (green/red) and $2mF_o - DF_c$ maximum-likelihood omit map contoured at $1\sigma$ (blue) are shown. Maps were calculated for a refined model with the ligand atoms omitted.

In subsequent work, Nukui and coworkers studied a set of mutations that affect the 'hydrophobic patch' which is presumed to be responsible for the positioning of the HA1 unit of the full substrate (Nukui *et al.*, 2003). The mutant enzyme structures (PDB entries 1n7n, 1n7o and 1n7p) were presented and, apart from some modelling errors affecting remote parts of the molecule, they are correct. In addition, two further structures (PDB entries 1n7q and 1n7r) are presented of the hexasaccharide complexes with the W291A/W292A and W291A/W292A/F343V mutants. While the similar complex structure described above was only missing the HA3 unit of the substrate molecule, there is no electron density at all in the complex structures 1n7q and 1n7r that would support the presence of any ligand in the active site of the enzyme. Several very basic crystallographic tests consistently lead to this unequivocal conclusion. Specifically, no significant electron density or positive omit difference density can be reconstructed and no significant correlation between observed and calculated model density exists. In addition and as expected, when both purported complex structures (PDB entries 1n7q and 1n7r) are used to calculate the $mF_o - DF_c$ difference electron-density maps, the entire substrate-molecule density coincides with negative density peaks, indicating the absence of the substrate (*cf.* Fig. 2). Despite the high *B*-factor values of the ligand molecules in these models already indicating an absence of meaningful density, negative difference density is still observed at levels that are significantly higher than expected for a ligand that actually contributes to the observed structure factors. Upon refinement, the *B* factors of the hexasaccharide molecules (~130 Å$^2$) substantially exceed those of the surrounding protein atoms (~35 Å$^2$). In fact, this difference is also present in the deposited models (although it is smaller: 110 *versus* 50 Å$^2$). When the hexasaccharide is removed from the model and the enzyme model is refined, the resulting electron-density maps contain no discernible positive difference density indicative of the ligand (Fig. 10). In agreement with vanishing real-space correlation, bias-minimized electron-density maps reveal no ligand density. Consistent with this observation, the original publication provided no electron-density figure.

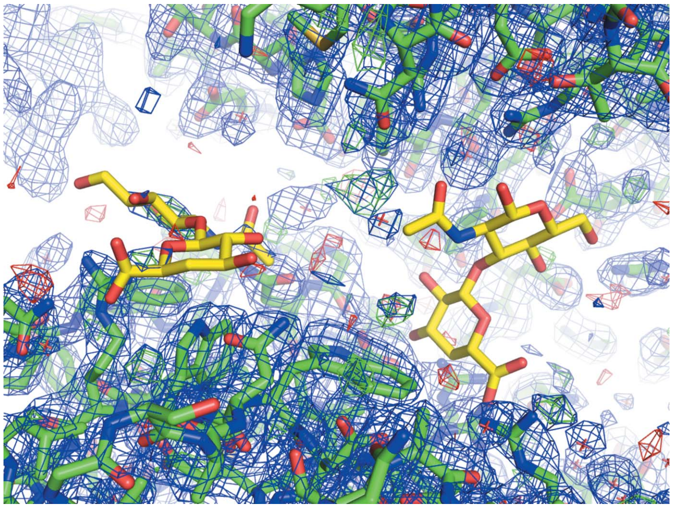


**Figure 11**
Missing ligands. Two disaccharide molecules in the structure of hyaluronate lyase from *S. agalactiae* (PDB entry 1i8q; Li & Jedrzejas, 2001) are not supported by the omit electron-density maps. The 2.2 Å resolution $mF_o - DF_c$ map contoured at ±3σ (green/red) and $2mF_o - DF_c$ maximum-likelihood omit map contoured at 1σ (blue) are shown. Maps were calculated for a refined model with the ligand atoms omitted.

Another earlier paper published by the same group describes hyaluronate lyase from *S. agalactiae* (Li & Jedrzejas, 2001). This protein is ~50% sequence identical to the *S. pneumoniae* ortholog and it is therefore not surprising that the structural models are also similar. Given the high degree of similarity between the two proteins, the authors anticipated

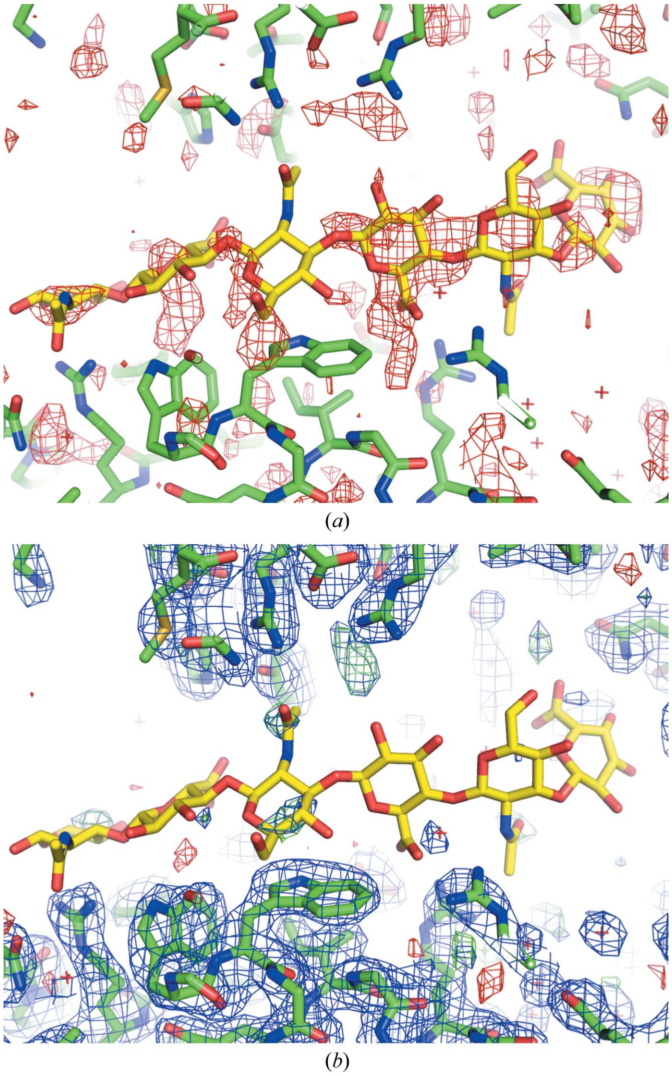


**Figure 12**
Absent ligands. The active site of hyaluronate lyase from *S. agalactiae*, PDB entry 1lxm (Mello *et al.*, 2002), does not contain the electron density needed to justify the presence of the hexasaccharide ligand. (*a*) shows difference density from the EDS contoured at −2σ, coinciding with the location of the ligand in the deposited model. (*b*) shows the omit maps and no electron density is observed that would allow ligand placement. (*b*) shows the 2.2 Å resolution $mF_o - DF_c$ map contoured at ±3σ (green/red) and $2mF_o - DF_c$ maximum-likelihood omit map contoured at 1σ (blue). Maps were calculated for a refined model with the ligand atoms omitted.

corresponding similarity of the catalytic mechanism. To verify the role of the corresponding residues and regions of the structure, the disaccharide-complex structure (PDB entry 1i8q) is presented. The structure of the protein part of the model alone (PDB entry 1f1s) appears to be refined correctly in general, except for a few disordered loops. It is claimed that two product molecules were found in the electron density. We have recalculated the electron density consistent with the EDS density using the deposited structure and experimental data with the disaccharide molecules removed. No significant positive density is observed beyond solvent within the active site to justify the placement of these molecules. With the disaccharides placed in their reported positions, despite the corresponding $B$ factors (~110 Å$^2$ upon refinement) being much higher than the surrounding atoms (~30 Å$^2$) and indicating the absence of significant model density contributions, negative density peaks are still observed in the difference map, while the omit maps show no evidence that the ligand molecules are present (Fig. 11).

The conclusions provided in Li & Jedrzejas (2001) are based exclusively on the purported complex structure 1i8q. Given the absence of the saccharide, the conclusions are unsubstantiated extrapolations from what is known about the *S. pneumoniae* homolog and are unsupported by crystallographic evidence. In the paper, a figure of a $2F_o - F_c$ electron density without an indication of $\sigma$ levels or specification of Fourier coefficients is provided. We were unable to reproduce this figure.

This work was further extended in a subsequent publication (Mello *et al.*, 2002). Here, the authors present the structure of the hyaluronate lyase complex with the hexasaccharide (PDB entry 1lxm), followed by a discussion of the role played by different regions of the enzyme and modelling of its dynamics. Similar to the case of the *S. pneumoniae* homolog, structural

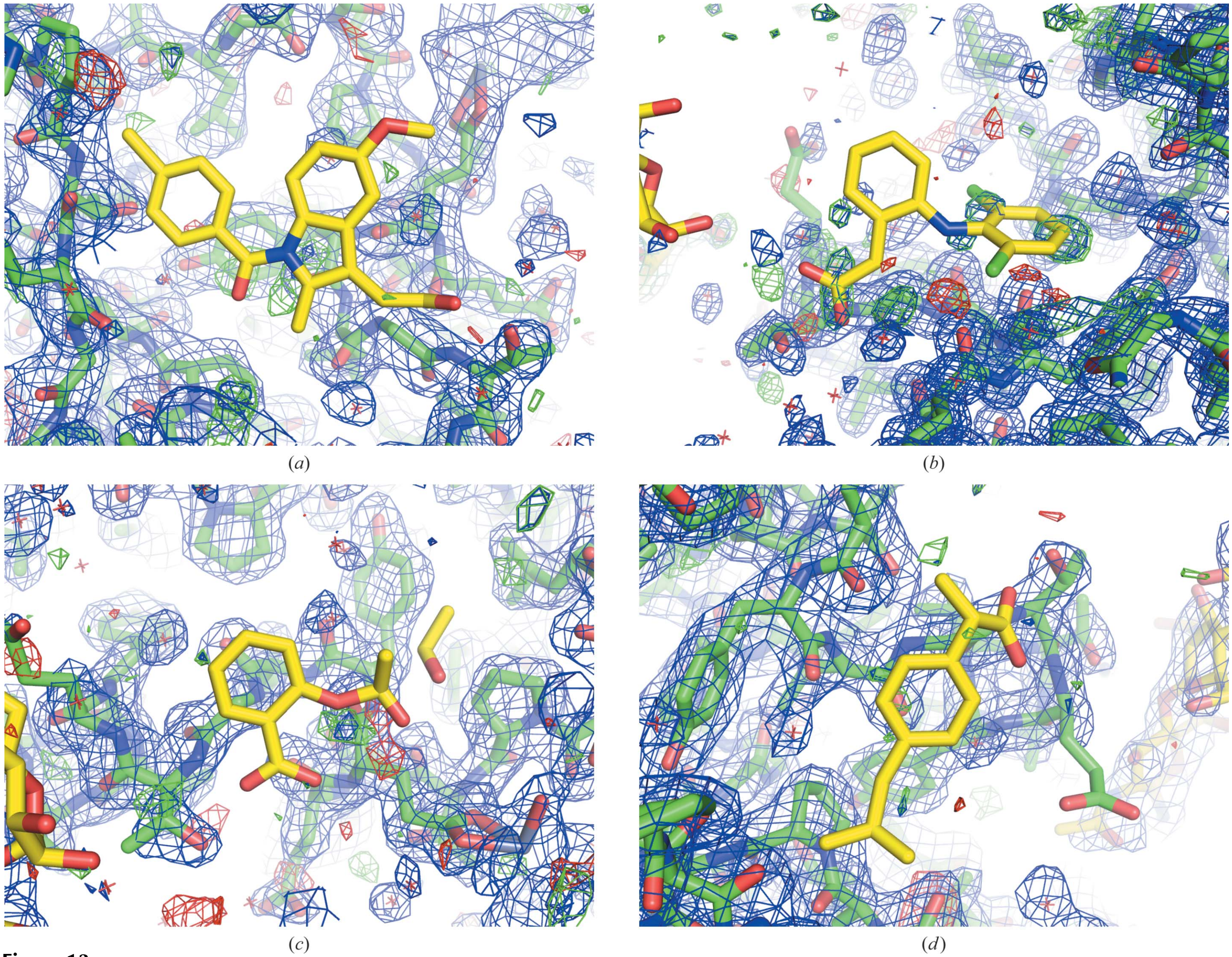


**Figure 13**
Absent ligands. Four protein–ligand complex structures presented in Mir *et al.* (2009) include ligands that are not supported by electron density. All panels show the omit maps for complex structures with the following ligands: (*a*) indomethacin (PDB entry 3ib1), (*b*) diclofenac (PDB entry 3ib0), (*c*) aspirin (PDB entry 3iaz) and (*d*) $\alpha$-methyl-4-(2-methylpropyl)benzeneacetic acid (PDB entry 3ib2). The 2.2 Å (*a*), 1.4 Å (*b*), 2.0 Å (*c*) and 2.29 Å (*d*) resolution $mF_o - DF_c$ maps contoured at $\pm 3\sigma$ (green/red) and $2mF_o - DF_c$ maximum-likelihood omit maps contoured at $1\sigma$ (blue) are shown. Maps were calculated for a refined model with the ligand atoms omitted.

findings supposedly explain the specific details of how the enzyme achieves the processivity of the substrate degradation.

The problems that we have found with the experimental data in this case are identical to all the other examples and are dramatic. If the hexasaccharide molecule is removed from the model (PDB entry 1lxm), the recalculated electron density shows a completely empty active site. With the substrate in place, negative density is observed in the difference map (Fig. 12) and the *B* factors of the substrate molecule (~135 Å$^2$) significantly exceed those of the surrounding atoms (~30 Å$^2$) (this is also evident from the statistics table included in the paper, although the *B* factors of the ligand are somewhat lower at ~100 Å$^2$). We thus conclude that the experimental data in this case once again do not support the presence of any hexasaccharide molecule in the active site. No electron-density figure is provided in Li & Jedrzejas (2001), consistent with the experimental evidence of low real-space correlation and high *B* factors.

Curiously, the structure presented in the paper is that of the wild-type enzyme and not an enzymatically inactive mutant. The authors do not discuss the fact that it would therefore be expected that upon long incubation the substrate would have been digested, even considering the reduced activity of the enzyme in the crystalline form.

### 4.2. Protein–ligand complex structures with unjustified ligands

The examples of problematic ligands in the crystal structures described below all share the same set of features. Namely, (i) no electron density is found in the ($2mF_o - DF_c$) maps that would justify the ligand placement, (ii) when the derivative model with ligands excluded is subjected to refinement no interpretable positive difference electron density is found in the ($mF_o - DF_c$) omit maps and, finally, (iii) when the ligands in the original locations are re-refined their *B* factors increase to levels that are incompatible with the surrounding protein atoms. In the cases where the deposited models do have *B* factors of ligand atoms that match the surrounding protein, there is distinct negative difference electron density overlapping with ligands in the maps calculated from the deposited models without modification. All of these observations indicate a lack of experimental evidence for the proposed ligands. What the presented examples have in common is that they originate from the same laboratory. We clearly cannot specifically address what may have led to the erroneous models. Possible explanations are consistent erroneous deposition of incorrect structure factors and/or incorrect structure models. Given the evidence, one cannot exclude the possibility of a pattern of flawed methodological approaches routinely used by the researchers working at the laboratory in question.

**4.2.1. 'The structural basis for the prevention of nonsteroidal anti-inflammatory drug-induced gastrointestinal tract damage by the C-lobe of bovine colostrum lactoferrin'**. In this paper (Mir *et al.*, 2009), the authors described four crystal structures of bovine lactoferrin determined using crystals prepared from samples that contained four different nonsteroidal anti-inflammatory drugs (NSAIDs). The authors inferred the binding constants from tryptophan fluorescence quenching and found apparent binding affinities in the submillimolar range. They further report that for four NSAIDs they observe 'reasonably characteristic electron densities at the ligand binding sites'. This finding is directly contradicted by the electron-density maps produced by the EDS and our own calculations (Fig. 13). In fact, only the complex with indomethacin shows any electron density at all near the reported ligand-binding site, yet this density is uninterpretable. Even assuming that this density can be attributed to the ligand, it only covers part of the indomethacin molecule and hence placing the latter in the structure is problematic (there is also no density above the 3$\sigma$ level in the difference omit map). Only the aspirin molecule has relatively low *B* factors as deposited; the model however overlaps with negative difference density, confirming its potential status as an artifact.

**4.2.2. 'Polysaccharide binding sites in hyaluronate lyase – crystal structures of native phage-encoded hyaluronate lyase and its complexes with ascorbic acid and lactose'**. Mishra and coworkers presented the crystal structures of hyalouronate lyase from *Streptococcus pyogenes* in its apo form and in complex with lactose and ascorbic acid (Mishra *et al.*, 2009). These crystal structures were determined at a resolution of ~3 Å, which in itself makes the reliable identification of a small-molecule ligand difficult. At the time of writing, the experimental data were available only for the ascorbic acid complex; therefore, we cannot evaluate the validity of the complex with lactose. As far as the complex with ascorbic acid is concerned, the placement of the ligand does not appear to be justified by the experimental data (Fig. 14).

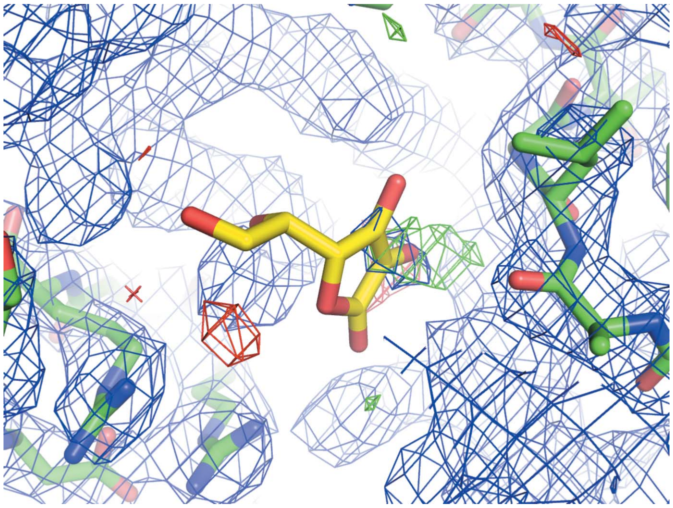


**Figure 14**
Absent ligand. The ascorbic acid molecule as placed in the structure of hyalouronate lyase from *S. pyogenes*, PDB entry 3eka (Mishra *et al.*, 2009), is unsubstantiated by the difference omit map. The 3.1 Å resolution $mF_o - DF_c$ map contoured at ±3$\sigma$ (green/red) and $2mF_o - DF_c$ maximum-likelihood omit map contoured at 1$\sigma$ (blue) are shown. Maps were calculated for a refined model with the ligand atoms omitted.

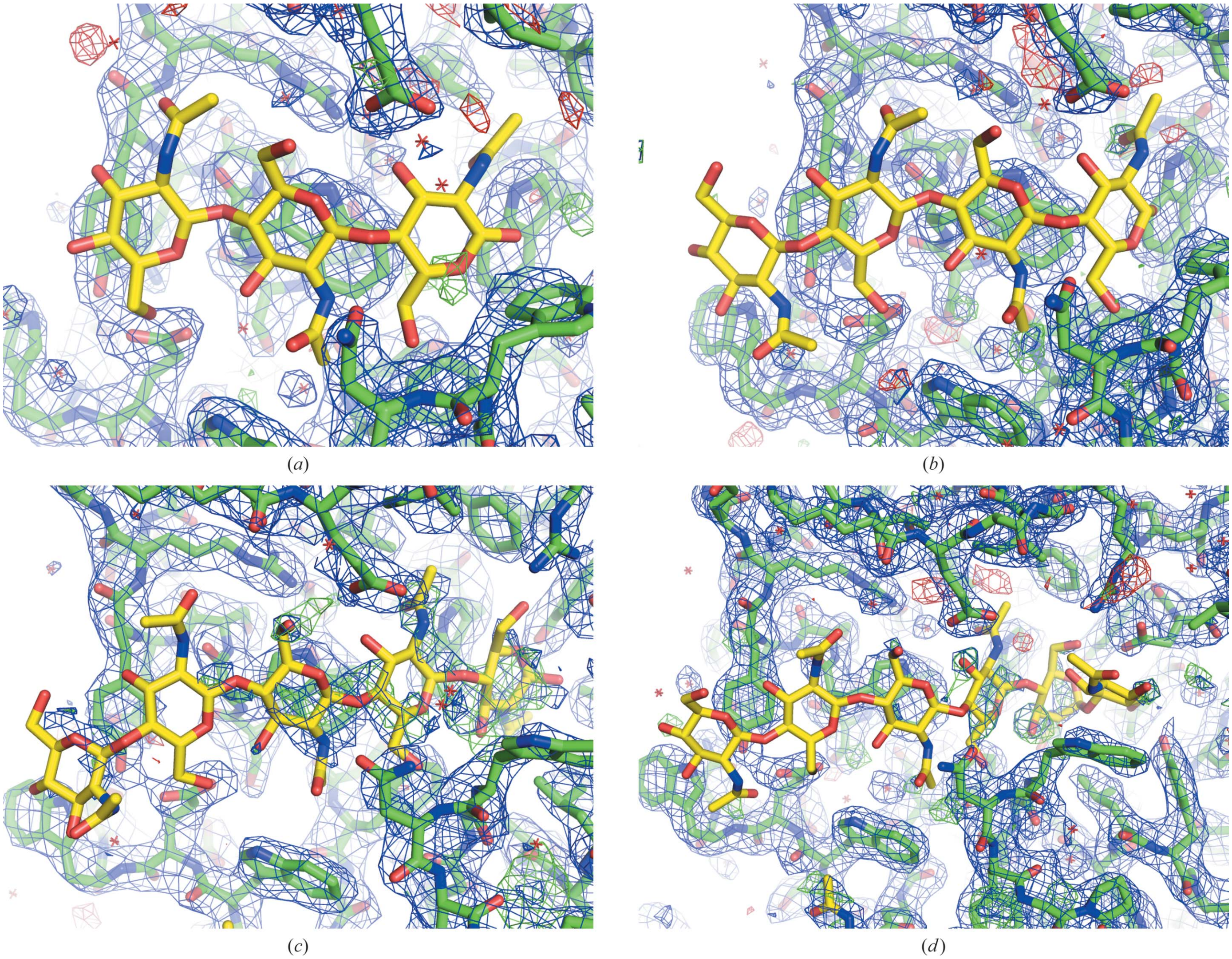


**Figure 15**
Absent oligosaccharides. The omit maps of the four oligosaccharides from Kumar *et al.* (2007) are shown for the protein complexed with the following oligosaccharides: (*a*) trisaccharide (PDB entry 2dt0), (*b*) tetrasaccharide (PDB entry 2dt1), (*c*) pentasaccharide (PDB entry 2dt2) and (*d*) hexasaccharide (PDB entry 2dt3). The 2.45 Å (*a*), 2.09 Å (*b*), 2.9 Å (*c*) and 2.28 Å (*d*) resolution $mF_o - DF_c$ maps contoured at $\pm 3\sigma$ (green/red) and $2mF_o - DF_c$ maximum-likelihood omit maps contoured at $1\sigma$ (blue) are shown. Maps were calculated for a refined model with the ligand atoms omitted.

**4.2.3. ‘Carbohydrate-binding properties of goat secretory glycoprotein (SPG-40) and its functional implications: structures of the native glycoprotein and its four complexes with chitin-like oligosaccharides’.** Kumar and coworkers presented structures of SPG-40 in complex with four oligosaccharides of different lengths (Kumar *et al.*, 2007). The authors reported that the ‘ligands were included only when they were well defined by unbiased $|F_o - F_c|$ maps’. This is not confirmed by the electron-density maps calculated with oligosaccharides excluded from the model (Fig. 15). Only in the pentasaccharide complex is there a continuous electron density that appears to cover three of the five sugars of the ligand, although this particular structure was determined at a significantly lower resolution (2.9 Å) than the others (2.1–2.5 Å).

**4.2.4. ‘Design of specific peptide inhibitors of phospholipase A$_2$: structure of a complex formed between Russell's viper phospholipase A$_2$ and a designed peptide Leu-Ala-Ile-Tyr-Ser (LAIYS)’.** In this paper, Chandra and coworkers presented the structure of Russell's viper phospholipase A$_2$ in complex with an inhibitory pentapeptide (Chandra *et al.*, 2002). The presence of the ligand is not supported by the experimental data (Fig. 16), although in this case some density is present that is clearly interpretable as discrete water molecules. Curiously, the authors report this interpretation for the second protein molecule present in the asymmetric unit.

## 5. Conclusions and suggestions

Macromolecular crystallography provides a direct view of the microscopic structure of biological molecules, and its ability to elucidate the nature of biological phenomena at an atomic level of detail has an enormous influence on biological research and drug development. The tremendous methodological advances of recent years have created a perception of ease about the way in which crystal structures can be determined, often with very little direct input from the researcher. As the detailed examples given here demonstrate, it is our fear that this view may be overly optimistic and that no amount of automation and clever data-interpretation algorithms can or even should replace the need for direct validation of a crystallographic structure model in the context of the primary evidence in the form of electron density. In the following, we provide suggestions of how direct improvements or clearer presentation of certain classes of ligand structures that we have identified as problematic can be achieved. We hope that some of our suggestions may initiate some movement towards a community consensus which will eventually provide clear and generally accepted guidelines for the presentation and validation of ligand structures.

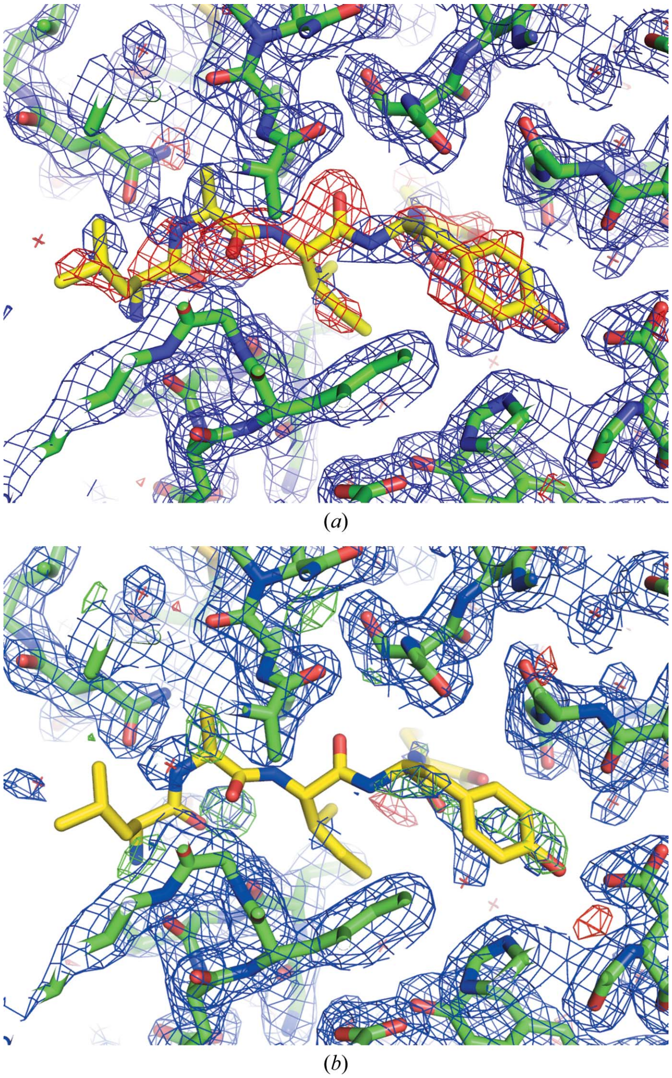


**Figure 16**
Absent inhibitor: the peptide inhibitor in the structure of phospholipase $A_2$ (PDB entry 1jq8; Chandra *et al.*, 2002). The electron-density maps downloaded from the EDS show that the placed ligand overlaps with negative difference density below the $-3\sigma$ level (*a*), while the omit maps do not support the presence of the ligand in the active site of the enzyme (*b*). (*b*) shows the 2.0 Å resolution $mF_o - DF_c$ map contoured at $\pm3\sigma$ (green/red) and $2mF_o - DF_c$ maximum-likelihood omit map contoured at $1\sigma$ (blue). Maps were calculated for a refined model with the ligand atoms omitted.

The problem of false positives in the form of ligands that are not there, or are at best the result of unjustifiably optimistic interpretation, should not be taken lightly. As has been pointed out, false positives waste resources by inspiring fruitless research programs and, ultimately, a field that is perceived as producing a large degree of false positives is at risk of losing its credibility (Simmons *et al.*, 2011).

### 5.1. Proof positive: omit difference density

The proposition that a ligand in a complex structure is in a specific position and exhibits a unique conformation (*i.e.* is present in a specific pose) is a very strong and powerful statement, and accepted scientific epistemology requires that strong claims are backed by correspondingly strong evidence. It is quite acceptable that the authors of a structural model are free to interpret the electron density derived from the experimental data to their liking, yet consistency with the experimental data is expected. Necessary self-imposed limits include an expectation that most experts would concur with the proposed interpretation. Some of the examples that we have presented in §3 and §4 included deposited structural models in which the ligand molecules (i) have severalfold higher *B* factors, sometimes combined with partial occupancy, than surrounding protein atoms, (ii) do not correlate with the ($2mF_o - DF_c$) electron-density map or (iii) do not coincide with any significant electron density in ($mF_o - DF_c$) difference density maps calculated from the models with the ligand removed, *i.e.* show no positive omit difference density. In our judgment, clear positive omit difference density (preferably supported by a properly annotated figure stating the exact type of Fourier coefficients used and density levels) would be the minimum requirement to justify ligand placement based on experimental crystallographic data. Very high *B* factors and low partial occupancies, particularly when combined, in general do not provide sufficient scattering contributions (*cf.* §1.1), so that a convincing positive difference density cannot be calculated. Problems of structures with entirely absent ligand density (as presented in §4.2) could be easily prevented by simple map inspection, even by a non-expert.

In cases where ligands contain heavier atoms, anomalous difference data can provide clues towards their identification and evidence for their presence. While this is quite obvious for ligands that contain phosphate moieties or metal ions, an educational example of identification of the S atom in HEPES buffer by means of anomalous difference Fourier density is provided in a publicly available tutorial collection (Faust *et al.*, 2008).

### 5.2. Overinterpretation and missing parts

A large pool of examples consist of ligand molecules resulting from, in our opinion, overly enthusiastic interpretation of the underlying experimental data. Given the importance of ligand placement in a protein–ligand complex for the purposes of 'biological interpretation', it appears to be unacceptable to place a ligand of interest into any blob of density that is not protein. It cannot be overemphasized that the vast majority of consumers of crystal structures still remain unacquainted with the methods of macromolecular crystallography and therefore consider deposited structural models to be akin to factual truth written in stone. Hence, a crystallographer must be careful not to include questionable or projected elements of a crystal structure. In the absence of a community consensus or better methods of stating positional uncertainty, the exclusion of model parts that cannot be verified by electron density seems to be preferable over random guesses. It should be remembered that in contrast to (hopefully past) reviewers' opinion, the absence of density is not a sign of a lack of crystallographic ability on the part of the investigator but is inherent to the underlying properties of the material and the nature of the method.

In the case of a protein model with disordered side chains, one knows that the side chain has to be somewhere but that it is probably split over multiple conformations that could conceivably be populated with occupancies (with their sum constrained to 1.0) according to empirical distributions of the possible conformers. This is not as simple in most ligands, because owing to the lack of covalent bonds to the protein occupancies lower than 1.0 are the norm rather than the exception (*cf.* §1.1) and often literally hundreds of poses (conformations and locations) in a binding site are more or less plausible *a priori*, as also shown in virtual ligand screening (An *et al.*, 2005). There may in fact not be a unique pose in a given crystal structure, which is a likely reason for poorly defined density or density where only certain parts of a ligand or, with luck, one or two major conformations are present. Unfortunately, the scattering contributions decrease with lower occupancy, and if additional positional uncertainty for whatever reason contributes to high *B* factors, no clear and unambiguous density can be observed out of principle. Again, it is not the absence of clear density that blames the crystallographer, but the desire to model what one fancies.

One possibility to reconcile the fact of indetermination with the often expressed concern that providing incomplete models (*e.g.* by excluding disordered side chains) may be just as misleading could be the establishment of the practice of depositing two types of models. One would be a minimalistic model that only includes the elements firmly supported by the underlying electron density. The secondary model would be just that: a model reflecting the opinion of the researcher as to where the disordered side chains are and which blobs of density may be interpreted as a glimpse of disordered and/or partially occupied ligand sites. It is important to emphasize that difficult-to-interpret electron density should not always be discarded, since it may contain important information, but presenting it in a structural model that noncrystallographers and crystallographers alike will tend to interpret as a hard fact seems to be misleading at best.

### 5.3. Improved review and validation

It is our hope that in the long term better teaching and training, as outlined in the next paragraph, will reduce the occurrence of overinterpreted ligand structures. In the immediate future, it is only through more competent and better enabled review and validation that this problem can be addressed. We see no technical reason why the PDB or another independent and trusted organization could not provide a tool similar to our *Twilight* script (Weichenberger *et al.*, 2013), allowing confidential analysis as soon as a structure is deposited in the PDB. Almost all respectable journals already demand the deposition of coordinates with the PDB, and since February 2008 deposition of the associated structure factors in the PDB has also been mandatory. Journal editors could give reviewers (or themselves) confidential access codes to pre-computed electron densities, which would allow even nonspecialists to reach a judgment as to whether or not the electron density supports the ligand pose. We therefore have provided in the *Twilight* distribution a simple review script that allows any PDB entry with EDS density to be fetched (after the provision of an at this point inactive password). The review script automatically displays the electron density and model beginning with the first bona fide ligand, as summarized in §2 and detailed in Weichenberger *et al.* (2013).

In addition, having structure entries reviewed for technical accuracy by an independent technical reviewer or body disinvested in their biological importance might likewise be feasible. We also wish to point out that the PDB policy that allows structure factors to be put on hold for longer than coordinate entries should be reconsidered. Suppression of primary evidence is inherently unscientific.

Validation and further review of small-molecular-weight ligands in protein crystal structures may be greatly facilitated if the wwPDB required the deposition of the dictionaries used, corrected atom names where possible, checked these against expected chemistry and returned a report to the depositor as part of the deposition screening. It is also important that these dictionaries are part of the deposition and should be retrieved by anyone who wishes to download the structure.

### 5.4. Training and teaching

Another significant concern lies with the ever-diminishing training requirements in biomolecular crystallography. We are not proposing that the structural scientists of today spend years learning every intricate detail of the technique before being allowed to deposit their first structure in the PDB. It is difficult to gauge how widespread cases are in which a young researcher is directed to perform protein crystallography without receiving adequate training and expert advice. However, we extrapolate from personal experience and through informal polling that the number of cases resulting from poor supervision by busy and result-oriented principal investigators may not be small. It is undeniable that with the ever-increasing number of structural studies performed today, and with the progressively increasing number of crystal structures deposited in the PDB, delivery of proper training is absolutely necessary to assure that the technique retains its trustworthy status. Efforts at remediation of the deposited crystal structures and quality control from the crystallographic community at large may be helpful, but to reduce errors in the first place by better training seems to be a more desirable long-term solution. The necessary teaching material including data and documented exercises are certainly available, for example, at http://www.helmholtz-berlin.de/bessy-mx-tutorial/ (Faust *et al.*, 2008).

One concerning observation made by the authors during recent poster-viewing sessions at crystallography conferences and structural biology meetings is that only few protein–ligand structure posters do actually show any electron-density figure in support of the ligand pose. This clearly seems to indicate that no emphasis is put on primary evidence, again pointing towards neglect of adequate training and expert advice. Despite all the diagnostics and validation tools available during model building, model refinement, and ultimately upon PDB deposition, one needs to realise and to impress that not the PDB but the individual crystallographer bears the final (and sometimes far-reaching; Petsko, 2007) responsibility for the correctness of the deposited model.

We thank Gerard Bricogne and Oliver Smart, Global Phasing Ltd, Cambridge, England for critical comments and validating our findings on PDB entries 1loh, 1n7q, 1i8q and 1lxn with the *BUSTER* software suite. We are also grateful to the anonymous reviewers for their helpful suggestions on improving the manuscript.

## References

Adams, P. D. *et al.* (2010). *Acta Cryst.* D**66**, 213–221.
Adams, P. D., Pannu, N. S., Read, R. J. & Brünger, A. T. (1997). *Proc. Natl Acad. Sci. USA*, **94**, 5018–5023.
An, J., Totrov, M. & Abagyan, R. (2005). *Mol. Cell. Proteomics*, **4**, 752–761.
Bacon, F. (1620). *Novum Organim Scientarium*, Aphoprism 49.
Bayes, T. (1763). *Philos. Trans. R. Soc.* **53**, 370–418.
Berman, H. M. (2008). *Acta Cryst.* A**64**, 88–95.
Brändén, C.-I. & Jones, T. A. (1990). *Nature (London)*, **343**, 687–689.
Bricogne, G. (1988). *Acta Cryst.* A**44**, 517–545.
Bricogne, G. (1997). *Methods Enzymol.* **276**, 361–423.
Brünger, A. T. (1997). *Methods Enzymol.* **277**, 366–396.
Burley, S. (2004). *Modern Drug Discov.* **7**, 53–56.
Chandra, V., Jasti, J., Kaur, P., Dey, S., Srinivasan, A., Betzel, C. & Singh, T. P. (2002). *Acta Cryst.* D**58**, 1813–1819.
Chen, Y., Hu, L., Punta, M., Bruni, R., Hillerich, B., Kloss, B., Rost, B., Love, J., Siegelbaum, S. A. & Hendrickson, W. A. (2010). *Nature (London)*, **467**, 1074–1080.
Danley, D. E. (2006). *Acta Cryst.* D**62**, 569–575.
Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* D**66**, 486–501.
Engh, R. A. & Huber, R. (2001). *International Tables for Crystallography*, Vol. *F*, edited by M. G. Rossmann & E. Arnold, pp. 382–392. Dordecht: Kluwer Academic Publishers.
Faust, A., Panjikar, S., Mueller, U., Parthasarathy, V., Schmidt, A., Lamzin, V. S. & Weiss, M. S. (2008). *J. Appl. Cryst.* **41**, 1161–1172.
French, S. (1978). *Acta Cryst.* A**34**, 728–738.
Fusetti, F., Schröter, K. H., Steiner, R. A., van Noort, P. I., Pijning, T., Rozeboom, H. J., Kalk, K. H., Egmond, M. R. & Dijkstra, B. W. (2002). *Structure*, **10**, 259–268.
Garcia, K. C., Degano, M., Stanfield, R. L., Brunmark, A., Jackson, M. R., Peterson, P. A., Teyton, L. & Wilson, I. A. (1996). *Science*, **274**, 209–219.
Gokulan, K., Khare, S., Ronning, D. R., Linthicum, S. D., Sacchettini, J. C. & Rupp, B. (2005). *Biochemistry*, **44**, 9889–9898.
Hajduk, P. J. & Greer, J. (2007). *Nature Rev. Drug Discov.* **6**, 211–219.
Hamilton, W. C. (1965). *Acta Cryst.* **18**, 502–510.
Hanson, M. A. & Stevens, R. C. (2000). *Nature Struct. Biol.* **7**, 687–692.
Hsieh, Y.-C., Wu, Y.-J., Chiang, T.-Y., Kuo, C.-Y., Shrestha, K. L., Chao, C.-F., Huang, Y.-C., Chuankhayan, P., Wu, W., Li, Y.-K. & Chen, C.-J. (2010). *J. Biol. Chem.* **285**, 31603–31615.
Jedrzejas, M. J., Mello, L. V., de Groot, B. L. & Li, S. (2002). *J. Biol. Chem.* **277**, 28287–28297.
Joosten, R. P., te Beek, T. A., Krieger, E., Hekkelman, M. L., Hooft, R. W., Schneider, R., Sander, C. & Vriend, G. (2011). *Nucleic Acids. Res.* **39**, D411–D419.
Kleywegt, G. J. & Harris, M. R. (2007). *Acta Cryst.* D**63**, 935–938.
Kleywegt, G. J., Harris, M. R., Zou, J., Taylor, T. C., Wählby, A. & Jones, T. A. (2004). *Acta Cryst.* D**60**, 2240–2249.
Kleywegt, G. J. & Jones, T. A. (1996). *Structure*, **4**, 1395–1400.
Koehler, J. J. (1993). *Organ. Behav. Hum. Decis. Process.* **56**, 28–55.
Kumar, J., Ethayathulla, A. S., Srivastava, D. B., Singh, N., Sharma, S., Kaur, P., Srinivasan, A. & Singh, T. P. (2007). *Acta Cryst.* D**63**, 437–446.
Laskowski, R. A. & Swindells, M. B. (2011). *J. Chem. Inf. Model.* **51**, 2778–2786.
Lebedev, A. A., Young, P., Isupov, M. N., Moroz, O. V., Vagin, A. A. & Murshudov, G. N. (2012). *Acta Cryst.* D**68**, 431–440.
Li, S. & Jedrzejas, M. J. (2001). *J. Biol. Chem.* **276**, 41407–41416.
Luzzati, V. (1953). *Acta Cryst.* **6**, 142–152.
Mello, L. V., De Groot, B. L., Li, S. & Jedrzejas, M. J. (2002). *J. Biol. Chem.* **277**, 36678–36688.
Merritt, E. A. (2012). *Acta Cryst.* D**68**, 468–477.
Mir, R., Singh, N., Vikram, G., Kumar, R. P., Sinha, M., Bhushan, A., Kaur, P., Srinivasan, A., Sharma, S. & Singh, T. P. (2009). *Biophys. J.* **97**, 3178–3186.
Mishra, P., Prem Kumar, R., Ethayathulla, A. S., Singh, N., Sharma, S., Perbandt, M., Betzel, C., Kaur, P., Srinivasan, A., Bhakuni, V. & Singh, T. P. (2009). *FEBS J.* **276**, 3392–3402.
Moriarty, N. W., Grosse-Kunstleve, R. W. & Adams, P. D. (2009). *Acta Cryst.* D**65**, 1074–1080.
Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* D**67**, 355–367.
Navratna, V., Nadig, S., Sood, V., Prasad, K., Arakere, G. & Gopal, B. (2010). *J. Bacteriol.* **192**, 134–144.
Nukui, M., Taylor, K. B., McPherson, D. T., Shigenaga, M. K. & Jedrzejas, M. J. (2003). *J. Biol. Chem.* **278**, 3079–3088.
Pannu, N. S. & Read, R. J. (1996). *Acta Cryst.* A**52**, 659–668.
Petsko, G. A. (2007). *Genome Biol.* **8**, 103.
Pyburn, T. M., Bensing, B. A., Xiong, Y. Q., Melancon, B. J., Tomasiak, T. M., Ward, N. J., Yankovskaya, V., Oliver, K. M., Cecchini, G., Sulikowski, G. A., Tyska, M. J., Sullam, P. M. & Iverson, T. M. (2011). *PLoS Pathog.* **7**, e1002112.
Qi, X., Loiseau, F., Chan, W. L., Yan, Y., Wei, Z., Milroy, L. G., Myers, R. M., Ley, S. V., Read, R. J., Carrell, R. W. & Zhou, A. (2011). *J. Biol. Chem.* **286**, 16163–16173.
Ranatunga, W., Hill, E. E., Mooster, J. L., Holbrook, E. L., Schulze-Gahmen, U., Xu, W., Bessman, M. J., Brenner, S. E. & Holbrook, S. R. (2004). *J. Mol. Biol.* **339**, 103–116.
Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.
Read, R. J. *et al.* (2011). *Structure*, **19**, 1395–1412.
Roversi, P., Blanc, E., Vonrhein, C., Evans, G. & Bricogne, G. (2000). *Acta Cryst.* D**56**, 1316–1323.
Rupp, B. (2006). *Nature (London)*, **444**, 817.
Rupp, B. (2009). *Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology*, 1st ed. New York: Garland Science.
Rupp, B. (2010). *J. Appl. Cryst.* **43**, 1242–1249.
Rupp, B. & Segelke, B. W. (2001). *Nature Struct. Biol.* **8**, 643–664.
Schüttelkopf, A. W. & van Aalten, D. M. F. (2004). *Acta Cryst.* D**60**, 1355–1363.

Simmons, J. P., Nelson, L. D. & Simonsohn, U. (2011). *Psychol. Sci.* **22**, 1359–1366.

Smart, O. S., Womack, T. O., Flensburg, C., Keller, P., Paciorek, W., Sharff, A., Vonrhein, C. & Bricogne, G. (2011). *Acta Cryst.* A**67**, C134.

Terwilliger, T. C. (2003). *Acta Cryst.* D**59**, 1688–1701.

Tickle, I. J. (2012). *Acta Cryst.* D**68**, 454–467.

Weichenberger, C. X., Pozharski, E. & Rupp, B. (2013). *Acta Cryst.* F**69**, doi:10.1107/S1744309112044387.

Weiss, M. S. & Einspahr, H. (2011). *International Tables for Crystallography*, Vol. *F*, edited by M. G. Rossmann & E. Arnold, pp. 64–74. Dordecht: Kluwer Academic Publishers.

Yeh, J. I., Chinte, U. & Du, S. (2008). *Proc. Natl Acad. Sci. USA*, **105**, 3280–3285.

Zhao, B. *et al.* (2005). *J. Biol. Chem.* **280**, 11599–11607.

Zhou, T. *et al.* (2007). *Nature (London)*, **445**, 732–737.