Published in final edited form as:

J Comput Aided Mol Des. 2015 September ; 29(9): 817–836. doi:10.1007/s10822-015-9833-8.

Models of protein–ligand crystal structures: trust, but verify

Marc C. Deller and

The Joint Center for Structural Genomics, San Diego, CA, USA

Department of Integrative Structural and Computational Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037, USA

Bernhard Rupp

k.-k. Hofkristallamt 991 Audrey Place, Vista, CA 92084, USA

Department of Genetic Epidemiology, Medical University of Innsbruck, Schöpfstr. 41, 6020 Innsbruck, Austria

Bernhard Rupp: br@ruppweb.org

Abstract

X-ray crystallography provides the most accurate models of protein–ligand structures. These models serve as the foundation of many computational methods including structure prediction, molecular modelling, and structure-based drug design. The success of these computational methods ultimately depends on the quality of the underlying protein–ligand models. X-ray crystallography offers the unparalleled advantage of a clear mathematical formalism relating the experimental data to the protein–ligand model. In the case of X-ray crystallography, the primary experimental evidence is the electron density of the molecules forming the crystal. The first step in the generation of an accurate and precise crystallographic model is the interpretation of the electron density of the crystal, typically carried out by construction of an atomic model. The atomic model must then be validated for fit to the experimental electron density and also for agreement with prior expectations of stereochemistry. Stringent validation of protein–ligand models has become possible as a result of the mandatory deposition of primary diffraction data, and many computational tools are now available to aid in the validation process. Validation of protein–ligand complexes has revealed some instances of overenthusiastic interpretation of ligand density. Fundamental concepts and metrics of protein–ligand quality validation are discussed and we highlight software tools to assist in this process. It is essential that end users select high quality protein–ligand models for their computational and biological studies, and we provide an overview of how this can be achieved.

Keywords

Crystal structure; Protein structure; Protein–ligand complex; Quality control; Structure validation; Structure-based drug design

Correspondence to: Bernhard Rupp, br@ruppweb.org.

Protein–ligand models

Models of biomolecular structures determined experimentally, by X-ray diffraction or Nuclear Magnetic Resonance (NMR) spectroscopy, are deposited in a world-wide public repository, the Protein Data Bank (PDB, http://www.pdb.org) [1–4]. Electron Microscopy (EM) models are deposited in a separate database, EMDB (http://www.emdatabank.org/) [5, 6]. As a basis for computational studies, and ultimately structure-based drug design (SBDD), atomic models of the highest quality are the most desirable [7, 8]. Atomic models determined by X-ray crystallography are often preferable to those generated by NMR spectroscopy, especially for use in SBDD, although both techniques have their own specific strengths and weaknesses [9, 10]. Accurate atomic models are also essential for computational methods such as ligand docking, active site identification, and in silico lead optimization.

Multiple structures of a protein target in complex with small molecule compounds (or fragments) [11], as well as complementary apo (ligand-free) structures are often needed in order to obtain a comprehensive atomic-level view of the ligand binding event. Ligand binding can often involve interesting rearrangements of the protein structure and changes in the features of the active site or ligand binding pocket. Such plasticity is often exploited, using SBDD techniques, in an attempt to increase both drug specificity and potency. Using the most accurate of the available atomic models allows computational methods to reveal relevant functional details with correspondingly high confidence. This review will focus on some common methods used to assess the *local* quality of ligands bound to protein structures. However, it is important to note that the *global* quality of the overall protein model must also be assessed and we will discuss these quality indicators where appropriate. Both *global* and *local* quality measures are closely coupled as a result of factors such as the resolution of the X-ray diffraction data, atomic displacement and dynamics, intrinsic flexibility, and disorder within the protein or the crystal (see “Local versus global validation measures” for more on *local* versus *global* measures of error).

In the following article, the fundamental concepts of protein–ligand validation will be discussed and some validation software tools will be reviewed. The aim of this article is to guide scientists with limited exposure to the practice of biomolecular crystallography, so that they can competently and confidently assess the quality of the protein–ligand complexes used in their research.

Assessment of protein–ligand model quality–the current state

One of the explicit strengths of X-ray crystallography is the unparalleled advantage of a clear mathematical formalism relating the experimental data to the primary evidence. In the case of protein crystallography, the primary evidence is the electron density of the protein, ligand, and solvent molecules that form the crystal. The ability to conduct strict evidence-based validation has revealed a number of instances in which the ligands reported as bound to the protein are insufficiently supported by the primary evidence [12–14]. It is important to note that there may be other lines of evidence (e.g. biochemical data) that support binding of the particular ligand to the protein under study, but it is essential to ensure that the electron density supports the placement of the ligand in the crystal structure model.

Electron density-based validation of protein–ligand models deposited in the PDB suggests that a small, but significant fraction, are only partially based on evidence and should be used with caution (Table 1) [12, 14, 15]. Given there are currently over 70,000 protein–ligand crystal structure models deposited in the PDB (as of Nov, 2014), a conservative estimate of 12 % 'bad' structures suggests that on the order of ~8,400 protein–ligand complexes may need very critical investigation. The true fraction of protein–ligand structures that are 'bad' is not trivial to determine. This is largely because no simple metric (or even complex composite measure) exists that accurately reflects all of the parameters of the atomic model and the chemical environment of both the ligand and the protein. These technical difficulties make the assessment of protein–ligand models non-trivial (see "Validation of protein–ligand models against the primary experimental evidence and prior expectations").

Furthermore, the definition of what constitutes a 'bad' protein–ligand structure clearly depends on the intended use of the model. For example, protein–ligand models used by a biologist to determine which residues of an active site to target for mutagenesis studies require a model with accurately positioned target residues in the active site. These target residues must be in good agreement with the electron density, because the overall reliability of the positioning of the residues and ligands in the active site is of interest. Measures used to assess the positioning against the electron density include the Real Space Correlation Coefficient (RSCC), Real Space R-Value (RSR) or more complex composite measures such as Local Ligand Density Fit (LLDF) (see "Electron density-based validation of protein–ligand models" for definitions). Less critical for the planning of mutagenesis studies are, for example, precise bond lengths and bond angles. Conversely, models used by a computational chemist for detailed Quantitative Structure–Activity Relationships (QSAR) calculations will require accurate positioning of the atoms in the active site in correspondingly good agreement with the electron density, and will additionally require a model with accurately determined stereochemistry. The stereochemical quality of the model is commonly assessed using parameters such as the root mean squared deviation (RMSD) of bond lengths and bond angles from known target or expectation values (see "Stereochemistry-based validation of protein–ligand models" for more on stereochemistry).

While test sets of pre-validated protein–ligand models are available for docking and other in silico studies, 'purpose-sensitive' validation of any protein–ligand model by the user is advisable [16, 17]. The suitably of protein–ligand models for docking and virtual screening depends on many factors including reliable pose prediction, bias in the ligand library, suitable scoring metrics and, fundamentally, errors in the protein–ligand crystal structure model [18]. This review focuses on error detection and validation of the protein–ligand models themselves.

The protein–ligand model and its parameters

A protein–ligand model is, in essence, an atom-by-atom listing of the refined positions of all atoms in the crystal structure. The Cartesian (orthogonal world) coordinates of each atom are recorded along with two additional parameters that model the atomic displacement (B-factor) and the occupancy of the atom. Additional header records contain the dimensions of the unit cell and define the symmetry of the crystal. These header records are necessary for

software programs to generate all of the symmetry-related neighboring molecules that make up the crystal lattice and often informs on the biological assembly.

Historically, coordinate files provided by the PDB are typically in a FORTRAN record-based format. However, several initiatives and software packages are encouraging the use of model coordinates and diffraction data that are stored in the more modern macromolecular crystallographic information file (mmCIF) format [19]. Regardless of the file format, the positional coordinates of each atom and its B-factor are recorded and refined during the process of model building and refinement (see "Refinement, stereochemical restraints and the data-to-parameter ratio" for more on refinement). However, the occupancy of each atom position is only modeled and refined when needed and justified by sufficient experimental data. Examples requiring occupancy refinement include models containing protein side chains with multiple conformations, where the sum of all modeled atom positions is constrained to a total occupancy of 1.0. Similarly, ligands bound to proteins can also undergo refinement of occupancy, for example, in cases where alternate conformations of the ligand are modeled by the crystallographer. However, assigning a default value of 1.0 to the occupancy of the ligand atoms, irrespective of the *actual* level of occupancy within the protein molecules of the crystal, is a common but not always justifiable practice (see "Ligand binding is rarely complete" for more on occupancy).

An important feature of crystallographic protein structure models is that hydrogen atoms are rarely listed in the coordinate file. Hydrogen atoms are not directly visible in electron density maps, at typical macromolecular resolutions, but they are included in refinement and stereochemical validation programs. The position of the hydrogen atoms can be computed, as so called 'riding' atoms, from known stereochemistry, and they can be easily added to the crystallographic model coordinates. The implicit incorporation of hydrogen atoms during the crystallographic refinement process improves both the fit of the model to the experimental data and also leads to better stereochemical repulsion restraints, resulting in a more accurate structure model.

There is no guarantee that a structure model, as deposited in the PDB by the crystallographer, is complete. It is possible that some of the atoms that are present in the crystallized molecules are not actually accounted for or interpretable in the electron density (see "The primary data and experimental evidence"). Unless some form of chemical reaction or proteolytic cleavage event has taken place, these atoms are obviously still present in the crystal, but the experimentalist has no data to support the exact positioning of the "missing" atoms. For example, solvent exposed side chains of proteins are often flexible and may adopt many conformations. In this instance, a single side chain will refine with correspondingly high B-factors that significantly exceed the average B-factors of the protein. The contribution of these atoms to the scattering of X-rays will be greatly reduced and little or no electron density will be evident for these atoms. Instead of accepting these high B-factors in the model, which is one method to indicate the high positional uncertainty for such atoms, some crystallographers prefer to simply omit these very high B-factor atoms from the model coordinates. These "missing" atoms can be a nuisance for the modeler or computational chemists, as many software packages require complete coordinate files and missing atoms must be rebuilt. Another practice, and perhaps the most problematic, is to

include the “missing” atoms in the model, but assign them an occupancy of zero. However, atoms with zero occupancy will not be included in crystallographic model refinement and many validation programs will exclude them from the analysis. Atoms modeled with an occupancy of zero are almost always simply ignored and no repulsive restraints are applied or close contacts analyzed. Such atoms are therefore often assigned a low or near zero occupancy (such as 0.01) by the crystallographer, to ensure they are included in the restraint generation for refinement (see “Refinement, stereochemical restraints and the data-to-parameter ratio”) and in the validation process. Zero (or near zero) occupancies can be misleading when not highlighted by a display program or when a computational program takes the presence of such atoms at face value. Additionally, the refined or assigned B-factor values of near-zero or zero occupancy atoms are unlikely to be meaningful. In summary, the end user of experimentally determined X-ray crystal structure models should be aware of the different practices used for annotating positional uncertainty of atomic coordinates and the effect that “missing” atoms may have on their display and modelling programs.

The primary data and experimental evidence

The electron density of a protein–ligand complex can be downloaded directly from the Electron Density Server (EDS) in a form that is suitable for display in most common graphics packages (http://bit.ly/1Ddnlkg) [20]. Additionally, electron density maps can be obtained directly from the PDB_REDO server (http://bit.ly/1pVYQQA) [21, 22]. The added advantage of obtaining electron density maps using this route is the ability to use plugins developed for PyMOL that automate many of the steps involved in the visualization of electron density maps (http://bit.ly/1tqK5uw and http://bit.ly/1u58hE2). In the absence of pre-computed electron density maps, structure factors can be downloaded from the PDB, and electron density maps created using the software tools discussed in the section titled “Structure validation software”.

None of these options are possible for PDB entries deposited without associated structure factors. Deposition of structure factors became mandatory in the PDB in 2008 (although voluntary compliance was in place much earlier). Still, for some PDB entries created prior to this date, structure factor data are absent and the generation of electron density maps is simply not possible. In these cases, the protein–ligand model cannot be validated against the primary data and the model should be used with caution.

Concepts and limitations of macromolecular crystallography

2014 is the centennial year of the award of the 1914 Nobel Prize in Physics to Max von Laue for recording the first X-ray diffraction pattern of a crystal (http://www.iycr2014.org/). Since then, X-ray crystallography has made countless technical advances that allow almost every reasonably trained scientist to determine—given sufficient patience and persistence—the structure of proteins and protein–ligand complexes. Many introductory protein crystallography texts are available including basic guides through to comprehensive textbooks [23, 24]. Here we will introduce a subset of important concepts, strengths and limitations of macromolecular crystallography, which the end user should appreciate in order to effectively validate the protein–ligand crystal structures that they use in their research.

Overall quality of protein–ligand models in the PDB

Overall, the vast majority of protein–ligand models deposited in the PDB are of good quality. This particularly holds true for structures determined automatically using high-throughput methods, such as those utilised by structural genomics consortia including the JCSG [25]. A common metric used to measure the overall quality of a protein structure is the R-value, a linear residual measuring the difference between the observed and the calculated diffraction data. Using such metrics, it can be shown that automatically determined crystal structure models, are generally *on par* with, or of better quality than conventionally built models (Fig. 1a). This is largely because automated high-throughput methods dictate a strict workflow and a set of standard operating procedures, which ensure that high-quality data collection and structure refinement practices are upheld. Such standard operating procedures are not enforced within the crystallography community at large and, therefore, the quality of such protein–ligand models must be carefully assessed by the end user on a case-by-case basis. To help address this issue, we provide the user with a list of validation software and recommended practices towards the end of the review (see "Structure validation software" and "Recommended practices for interpretation of crystallographic data and validation of protein–ligand models").

Resolution—As with every experimental method, the amount and quality of data that can be collected determines the quality and detail of the results that can be obtained. In the case of protein crystallography, the better the crystal diffracts X-rays, the more details can be discerned in the crystal structure (Fig. 1b–d). A common *global* indicator of detail in an atomic model is the 'resolution' of the structure, expressed in Ångströms (Å, 10^{-10} m, 0.1 nm). This measure is derived from the extent to which useful diffraction data are observed when the protein crystal is exposed to X-rays. Just like the R-values discussed above, resolution is a *global* indicator of quality and does not provide any information about the *local* quality of the protein–ligand model (see "Local versus global validation measures" for more on *local* versus *global* measures of error). The resolution of a protein–ligand model is nevertheless an important metric, as it gives a first impression of the achievable detail in the structure model and can be used as a provisional indicator of how useful the model is for further computational studies. As illustrated in Fig. 1, approximately 1.5–2.5 Å (or better) resolution is desirable to reliably determine the position and conformation of a ligand. At increasingly lower resolution, the placement of the ligand becomes less reliable as the positional accuracy of the atoms in the electron density is reduced. It is important to note that *global* measures of overall quality, such as R-values and resolution, are closely coupled; low resolution structures will have higher residual errors when compared with a similarly well-refined structure at a higher resolution (circled data points in Fig. 1a). Therefore, it is important that these measures of overall quality and error are considered together and *not* in isolation.

Effects of the crystalline state

Protein crystals form by the self-assembly of large and irregularly shaped protein molecules into a regular periodic lattice. Such crystals are best described as a loose network of protein molecules held together by a few weak, but specific, non-covalent intermolecular interactions. As a result of these weak interactions, protein crystals are very fragile and

sensitive to environmental change. Furthermore, the loose arrangement of protein molecules in the crystal often results in large solvent channels surrounding the protein molecules of the crystal. The solvent content of most crystals is between 30 and 70 % [26]. The solvent channels between individual protein molecules of the crystal make it possible to diffuse small molecules into pre-formed protein crystals. These “soaking” experiments, and the co-crystallization of proteins incubated with ligands, allow the growth of crystals of protein–ligand complexes and enable methods such as fragment-based screening and SBDD [11].

Crystal packing—Although significant portions of the protein molecule are in contact with solvent, some parts of the protein molecule must also be in contact with neighboring protein molecules in order to form a crystal lattice. Protein residues in these interfaces are said to be involved in “crystal packing”. It is important to check if any residues required for protein function or ligand binding are involved in crystal packing, as such residues may be in conformations different to their native solution state. Additionally, protein residues in the binding sites may be blocked or perturbed by neighboring protein molecules and ligand binding may be adversely affected. Therefore, it is essential that the entire local environment of the ligand binding region is inspected, in the context of the crystal symmetry, before comprehensive validation of the ligand can be carried out. Inspection of crystal packing requires the generation of all symmetry-related molecules in the vicinity of the protein–ligand complex. Tools are available to automate symmetry mate generation in graphics viewers, such as Coot and PyMOL [27, 28].

Biological unit—Protein–ligand models deposited in the PDB represent the components observed in the asymmetric unit (ASU) of the protein crystal. The ASU is the fundamental building unit of the crystal, and by application of the symmetry operations determined by the space group of the crystal, the entire structure of the crystal can be generated from the ASU. Therefore, the model contained within a PDB file does not necessarily represent the biologically relevant quaternary assembly. The complete local environment of the molecule or sub-assembly must be assessed for the presence of a plausible oligomeric state. For example, a ligand may bind at the interface of a dimer assembly, and it is therefore essential that validation of the ligand is carried out in the context of a dimer assembly. If the dimer is contained within the ASU, then the model as deposited in the PDB may be used directly for validation; however, it is possible that the dimer assembly spans two ASUs and the symmetry mates must be generated in order to correctly validate the ligand model. The PDB provides models of the biological assembly and software tools such as PISA are available to further assess the possibility of higher-order biological assemblies (http://bit.ly/13u0zbv) [29, 30]. Often, further biophysical studies under conditions approximating the native state, such as size exclusion chromatography (SEC) or dynamic light scattering (DLS), are required to confirm the most probable quaternary structure in vitro.

Disorder and mobility—Structure models determined by crystallography provide some (but limited) information about the dynamics of the protein molecule in the form of an atomic displacement parameter or B-factor (see “The protein–ligand model and its parameters”). For high resolution structures, the B-factor is refined for each atom of the model. It is important to note that while the B-factor can be expressed in terms of a mean

atomic displacement (in units of $Å^2$), it is in essence only a statistical measure indicating the probability of finding the atom at the defined coordinate position. For example, if the B-factor of an atom is higher than the average B-factor of the other atoms in the molecule, it suggests that the atom is statistically less likely to be present in its stated position. Reasons for this displacement include thermal motion of residues or the polypeptide chain, dynamic disorder, static disorder and long-range disorder between the protein molecules forming the crystal: the experimental diffraction data originate from an averaged scattering contribution of about 10^8–10^{11} molecules in a typical protein crystal. Therefore, the atomic model is essentially a snapshot of an averaged ensemble of molecules. If portions of the protein molecule (or even entire protein chains) deviate significantly from this averaged ensemble, this will manifest itself as higher than average B-factors and limited diffraction. Large atomic displacements parameters can also be a warning sign that the atomic model is incorrect and the refinement program does not support any significant scattering matter at the specified location.

Abnormally high B-factors for ligands are a concern, and ligands with B-factors significantly higher than the average of the local environment should be cause for alarm. For example, some of the lowest resolution protein–ligand structures deposited in the PDB have ligands with B-factors in excess of 200 $Å^2$. Such large B-factors correspond to a mean isotropic atomic displacement of ~1.6 Å (which is significantly larger than any of the bond lengths in the ligand the model is attempting to portray). If such ligand B-factors are significantly higher than the average B-factor of their protein environment, problems with the model or low occupancy can often be the cause (see “Ligand binding is rarely complete”).

Ligand binding is rarely complete—Binding of a ligand to a protein receptor reaches full occupancy only asymptotically, and for ligands that display tight binding affinities or at extremely high concentrations of ligand [12, 31]. For example, protein–ligand systems with a dissociation constant (K_d) in the 10–100 mM range will only achieve a maximal occupancy of between 70 and 90 % at normal working concentrations of ligand (<500 mM). Concentrations of ligand above this value are often impractical to achieve, particularly for hydrophobic ligands, and are potentially damaging to the crystal or protein sample. Optimal occupancies of >90 % are, in general, only obtained for ligands with K_d values <1 mM. Unfortunately, the dissociation constants of many natural and biologically relevant ligands are in the low to high millimolar range, thus making their capture in crystallographic models extremely challenging. A weakly binding ligand, at practically achievable concentrations (often in the low millimolar range), simply cannot have full occupancy, and consequently, less of its mass will be contributing to the overall scattering of the crystal. This results in a reduced contribution of the ligand to the electron density, in proportion to the occupancy, which reduces the experimental evidence available for the placement of a weakly binding ligand in the model. In addition, even remotely chemically similar compounds like precipitants, buffer molecules or additives, which are present in high concentrations in the crystallization cocktail, can compete with the targeted low solubility ligand, resulting in a binding site that either appears empty or contains an incorrect ligand [24, 31–34].

Refinement, stereochemical restraints and the data-to-parameter ratio

Model refinement is the process of improving the parameters of the initial approximate model until a best fit is achieved with the experimental data. In the case of protein crystallography, the primary experimental data are the intensities of the individual reflections of the diffraction data collected from the crystal. The intensities of the reflections are commonly reduced to their corrected amplitude values, the structure factor amplitudes. The process of refinement involves adjusting the parameters of the trial model until the calculated amplitudes generated by the model most closely match those of the diffraction pattern (for detailed theory see [24]).

The initial trial model built into the electron density often contains many small, highly correlated errors in its stereochemistry. However, the stereochemistry and geometries of small molecules are usually known, with high accuracy, and it has been shown that these parameters are similar in macromolecules [35, 36]. Therefore, these stereochemical parameters can be 'restrained' within a certain probable range that is based on prior expectations. Common so-called "strong" restraints used by crystallographers include bond distances (1–2 restraints), bond angles (1–3 restraints), planarity restraints, and chiral restraints in the form of chiral volumes (computational chemists often rely on angular restraints to prevent chiral inversion). Even though the bond lengths and bond angles of proteins are tightly restrained, proteins still have considerable flexibility and conformational freedom. The conformational freedom of proteins is largely a result of the flexibility in torsions between the planar peptide bond units and the Cα atoms of the polypeptide chain (backbone torsion angles) and the multiple rotamer conformations that side chains of the amino acid can assume. Torsions (1–4 restraints) are generally much less restrained than bond distances and bond angles and display a considerable degree of freedom. Restraints applied during the refinement of a crystallographic model have the important effect of increasing the data-to-parameter ratio.

In so-called restrained reciprocal space refinement, the four adjustable parameters of the atomic model are refined (3 positional coordinates and the B-factor). Several other parameters are used to describe the overall scaling, anisotropic displacements, and the bulk solvent of the model. In order to reach a minimal level of determinacy, the number of datapoints, n, needs to be at least equal to the number of parameters, p. For proteins with an average solvent content of approximately 50 %, the minimal level of experimental determinacy is reached at a resolution of about 2.5 Å. However, in order for any meaningful refinement to occur, n/p needs to be $\gg 1$. For structures determined to moderate resolution, values of $n/p > 1$ are only possible because the stereochemical restraints provide additional data points beyond those collected from the diffraction experiment. At the same time, these restraints assure that the model maintains physically reasonable stereochemistry.

True unrestrained refinement requires a high data-to-parameter ratio ($n/p \approx 10$) and is generally only feasible at true atomic resolutions better than about 1.2 Å. Therefore, it follows that the less data that are available, the more a structure model relies on the stereochemical restraints. The same is also true for the ligand model, and incorrect restraint files are often the source of poor ligand stereochemistry. Ligands have an infinitely more

diverse chemical composition and conformational freedom than the protein residues themselves. Therefore, to ensure that correct ligand stereochemistry is maintained throughout the course of refinement, it is essential that suitable restraints are applied to the ligand [37] (see "Stereochemistry-based validation of protein–ligand models" for more on ligand restraints).

Summary of important factors for validation of protein–ligand crystal structure models

The protein–ligand model is a complex system and many factors need to be critically assessed during the validation process, including but not limited to:

- The protein–ligand model is typically a set of refined atomic coordinates that does not contain any direct measure of uncertainty for the positions of the individual atoms.
- The protein–ligand model is determined using the electron density as the primary experimental evidence.
- The actual protein–ligand model deposited in the PDB requires automated or human interpretation of the electron density, which may be ambiguous. Interpretation by individual crystallographers can be subjective or biased.
- For most models, the electron density can be downloaded or quite easily computed using standard programs.
- Without inspection of its electron density, validation of the quality of the protein–ligand model is incomplete.
- The resolution of the data is a *global* indicator of the amount of diffraction data used in determining the model.
- High-resolution protein–ligand models are generally preferential because they reveal a higher level of detail and are likely to be more accurate.
- As a *global* parameter, resolution alone does *not* inform about the *local* model quality of the ligand or the surrounding residues.
- One of the limitations of the crystalline state is that one must *always* check for the effects of crystal packing.
- Parts of the protein–ligand model that exhibit any form of dynamic motion or disorder will typically have less clear electron density and are modeled with greater uncertainty. These regions of the model typically refine to higher B-factors.
- Generation of accurate protein–ligand models typically depends on stereochemical restraints derived from prior expectations of bond angles and bond lengths.
- Only accurate restraint files will give correctly refined protein–ligand models with plausible stereochemistry.
- The general dependence on restraints also demonstrates that macromolecular crystallography is not a suitable method for determining novel small molecule (ligand) structures.

- Every atom in the model contributes to each calculated data point during refinement. Therefore, it is important that the entire structure model, not just the ligand, is refined and validated as accurately as possible.
- For ligands with a K_d of >10 mM, the occupancy is likely incomplete and reduced contribution of the ligand to the electron density should be expected.

Validation of protein–ligand models against the primary experimental evidence and prior expectations

Protein–ligand models derived from crystallographic data must always be checked against both the primary experimental evidence, and their agreement with established prior expectations. The primary experimental evidence is the electron density and the prior expectations are the known distributions of stereochemical descriptors such as bond lengths, bond angles, and general stereochemistry [12, 14, 15, 35]. A multitude of diverse measures have been introduced to assess the quality of atomic models generated using protein crystallography (see Table 2 for a summary). Many of these parameters are used routinely by the crystallographer during the building of an atomic model. However, it is important that even the end user of atomic coordinates is aware of the meaning of these quality indicators.

Local versus *global* validation measures

It is important to distinguish between *reciprocal* space quality metrics and *real* space quality metrics. *Reciprocal* space metrics are those that relate directly to the diffraction data (i.e. the structure factor amplitudes), while *real* space metrics relate to the electron density and the atomic model. A second important classification of metrics distinguishes between *global* versus *local* validation measures. Given our inability to separate the contributions of individual atoms to each experimental observed diffraction data point, *reciprocal* space measures are always *global* in nature, and conversely, *real* space measures are always *local* in nature. Therefore, *global* and *reciprocal* space quality measures, such as resolution and R-value, are measures indicative of the overall quality of the model; they do *not* inform us about the validity of the specific position of an individual atom. Conversely, *real* space *local* measures, such as RSCC, RSR and LLDF, are measures of the fit of the model to the electron density. Although they inform us about the validity of the position of an individual atom, certain *local* measures can also be averaged into a regional or *global* measure. For example, one can trivially calculate an average RSCC for the entire ligand, the protein, or both together. Therefore, *real* space measures are useful for both *local* and *global* model validation, whereas, reciprocal measures are only applicable for *global* model quality assessment.

Common *global* validation measures of overall protein structure quality and their experimental data include the R-value, the cross validation measure R-free [38], CC* [39] and resolution. Useful *local* validation measures include RSCC, RSR, LLDF and Occupancy-Weighted Average B-factors (OWAB) [15, 20] (see Table 2 for summary and section “Electron density-based validation of protein–ligand models” for more on *local* quality measures). Certain *local* and *global* quality measures are correlated. For example, both the *local* RSCC measure and the overall *global* resolution of the structure correlate

with the OWAB of the ligand (Fig. 2a, b, respectively). As a general rule of thumb, we would expect a high-resolution protein–ligand model to have a high RSCC and low B-factors, as a result of generally low levels of thermal, dynamic and static disorder in the crystal (Fig. 2c). Conversely, we would expect a low-resolution protein–ligand model to have a lower RSCC value and higher B-factors, as a result of higher overall disorder in the crystal (Fig. 2e).

For the purpose of this review, *global* validation measures will not be discussed in detail and we will focus on *local* validation measures used to analyze the quality of the ligands bound to the protein model. G*lobal* validation measures, although correlated with the *local* quality of the ligand as shown in Fig. 2, can only be used to assess the overall quality of the atomic model and further *local* validation checks are required.

Electron density-based validation of protein–ligand models

For a comprehensive review on the statistics of electron density quality, the reader is referred to Tickle, 2012 [40]. In general, these methods rely on a comparison of the "observed" electron density computed from the diffraction data (ρ_{obs} in the examples below) with a "calculated" electron density map computed from the model coordinates (ρ_{calc} in the examples below). Electron density maps are typically σ_A-weighted $2mF_o–DF_c$ (maximum likelihood) maps, although other map types can be used [41].

It is important to realize that the publicly available electron density maps computed by EDS or PDB_REDO are calculated using phases based on the entire protein–ligand complex model. These maps are therefore biased towards the *presence*, rather than the *absence*, of ligand density. Therefore, problems with ligand density in these maps are a strong indication that careful inspection of the ligand density and the model is warranted. During model building, positive difference density omit-maps [42] calculated without the ligand model are used as 'proof positive' for the presence and placement of a ligand [12]. To obtain actual omit maps, it is therefore necessary to re-compute the electron density using crystallographic software following recommended procedures [12, 24].

One of the most commonly used metrics for assessing the fit of a model to the electron density is the RSCC, which is a standard linear sample correlation coefficient defined as,

$$\mathrm{RSCC}=\frac{\sum(\rho_{\mathrm{obs}}-\rho_{\mathrm{obs}})\cdot(\rho_{\mathrm{calc}}-\rho_{\mathrm{calc}})}{\left[\sum(\rho_{\mathrm{obs}}-\rho_{\mathrm{obs}})^2\cdot\sum(\rho_{\mathrm{calc}}-\rho_{\mathrm{calc}})^2\right]^{\frac{1}{2}}}$$

where ρ_{obs} and ρ_{calc} are the observed and calculated electron densities, respectively. RSCC values range from 0 ('Bad'), which suggests that the electron density for the ligand is essentially absent, through to 1 ('Good'), which suggests that the ligand fits the electron density perfectly. The spectrum of fits to the electron density between 'Good' and 'Bad' are summarized in Table 1 and Fig. 3. RSCC values are provided by the EDS server which is accessible through the PDB web sites, and tabulated or easily computed for protein–ligand models using modelling or validation software such as Coot, Twilight or VHELIBS (see "Electron density validation software tools").

Protein–ligand models with an RSCC >0.9 have well defined electron density for the ligand and the ligand fits the electron density well (Fig. 3b). After diligent inspection of the electron density maps for other possible errors in the model, such as present but unmodelled ligand electron density, it is generally safe to use such models.

Protein–ligand models with an RSCC of between 0.9 and 0.8 often have less well modeled ligands and portions of the model that do not show clear electron density (Fig. 3c). Ligands in this portion of the distribution are often simply over-modeled ('Dubious' classification according to VHELIBS). Over-modelled structures have parts of the ligand that are correct, but electron density for other parts of the model is missing. Unless some type of cleavage or molecular rearrangement event has taken place, the over-modeled portions of the ligand still exist; the problem arises from the fact that the electron density does not support the specific pose of the ligand as modeled. For this reason, we prefer the Twilight classification, which flags such ligands as "fits density partially" (Table 1). Models in this category require further inspection and can often be corrected with further rounds of refinement or by downloading the fully optimized version of the coordinates from the PDB_REDO server [21, 22].

Protein–ligand models with an RSCC of <0.8 are generally poorly modeled (or over-modeled) with significant portions of the ligand outside of the electron density (Fig. 3d). The placement and conformation of the ligand in such structures is not fully supported by the experimental evidence and these models should be used with caution. It is predicted that ~12 % of the protein–ligand structures deposited in the PDB contain ligand models that are not fully supported by the experimental electron density (Table 1 and Fig. 3).

Another related real space metric used for assessing the fit of the ligand to the electron density is the RSR value [43, 44],

$$\mathrm{RSR}=\frac{\sum|\rho_{\mathrm{obs}}-\rho_{\mathrm{calc}}|}{\sum|\rho_{\mathrm{obs}}+\rho_{\mathrm{calc}}|}$$

where ρ_{obs} and ρ_{calc} are the observed and calculated electron densities, respectively. One of the main disadvantages of the RSR metric is that the "observed" and "calculated" electron density maps must be scaled together. Scaling of the maps is not required for RSCC metrics making the calculations easier. However, RSR values have been adopted by the PDB on the basis of the Validation Task Force recommendations [45]. Validation reports are available for each PDB entry, and a per-residue analysis of RSR values is given, so that the user can directly assess the quality of the fit to the electron density (http://bit.ly/1si1ZeL).

Although metrics such as RSCC and RSR are powerful statistics for assessing the fit of a ligand to the electron density, these metrics have a significant disadvantage in that they do not have the ability to distinguish between the *accuracy* of the model and the *precision* (or uncertainty) of the underlying crystal and diffraction data. Therefore, it is important to note that both of these metrics are sensitive to the B-factor and occupancy of the ligand, the resolution of the diffraction data, and the size of the ligand. To address these issues, an Real

Space Observed Density Z-score (RSZO) metric has been proposed that directly reports the model precision [40],

$$\mathrm{RSZO}=\frac{\mathrm{mean}(\rho_{\mathrm{obs}})}{\sigma(\Delta\rho)}$$

where ρ_{obs} is the observed electron density of the ligand and $\sigma(\Delta\rho)$ is a measure of the uncertainty in the electron density of the ligand. The RSZO can be considered as a substitute for other precision metrics such as the B-factor. Improved B-factor metrics, such as the Occupancy-Weighted Average B-factor (OWAB), can also been used as a precision metric for assessing the quality of the ligand (Fig. 2) [15],

$$\mathrm{OWAB}=\frac{\sum(\mathrm{B}_{\mathrm{ligand}}\times\mathrm{Q}_{\mathrm{ligand}})}{\sum(\mathrm{B}_{\mathrm{ligand}})}$$

where B_{ligand} and Q_{ligand} are the B-factors and occupancies of the ligand atoms, respectively. Unreasonably low occupancies are not meaningful in a macromolecular structure model, simply evidenced by the fact that the occupancy factor reduces the scattering factors of an atom proportionally. For example, a carbon atom with 6 scattering electrons and an occupancy factor of 0.15 would correspond to less than one electron in scattering power—certainly not visible above the noise level in a typical macromolecular electron density map.

A recent addition to the PDB validation arsenal is the LLDF metric (http://bit.ly/1si1ZeL). LLDF is currently under development as a means for validating ligands. The LLDF metric compares the RSR of a ligand with the mean and standard deviation of the RSR values of neighboring amino acids within a radius of 5 Å of the ligand,

$$\mathrm{LLDF}=\frac{\mathrm{RSR}_{\mathrm{ligand}}-\sum\mathrm{RSR}_{\mathrm{site}}}{\sigma(\mathrm{RSR}_{\mathrm{site}})}$$

where RSR_{ligand} and RSR_{site} are the RSR values computed for the ligand and protein residues within 5 Å, respectively. The LLDF metric reflects the quality of the fit to the electron density for the ligand with respect to the fit of neighboring protein atoms in the ligand binding site. At present, LLDF values of greater than 2 are flagged in the PDB validation report as worthy of further investigation.

Poor scores for electron density-based validation metrics can have multiple causes. A very common practice leading to poor real space measures is overzealous inclusion of parts of the ligand that are not supported by the electron density. This is particularly true for ligands refined using tight B-factor restraints, where the B-factors are not allowed to vary significantly from one atom to the next. In such examples, the restraints on the ligand are not loose enough to correctly model the atom displacements in the ligand, and the refinement program is unable to raise the B-factors sufficiently high to eliminate the scattering contributions of the incorrectly placed atoms. In such scenarios, the parts of the model that

fit the electron density are not necessarily bad despite a 'suspect' overall real space measure for the ligand. It is also often the case that a ligand is simply not bound to the protein, or is in a low occupancy state, despite an abundance of supporting biochemical data suggesting that the ligand is bound (see "Ligand binding is rarely complete" for more on occupancy). For example, structures of Botulinum neurotoxin type B, purportedly containing a bound inhibitor, were inspected using electron-density-based validation methods [46]. It was concluded that the models, as deposited in the PDB, had little experimental electron density in support of a bound ligand and the structures were subsequently retracted [47, 48]. Ultimately, poor electron density-based metrics for a ligand are generally the result of overzealous modeling. Psychological factors such as expectation bias or confirmation bias can also be the motivation for placing fancy ligands into spurious or non-existent electron density [12, 49].

Stereochemistry-based validation of protein–ligand models

The X-ray methods used to determine small molecule structures (<900 daltons) differ substantially from those used for macromolecular crystallography. Small molecule crystals generally diffract to a much higher resolution and the additional data allow small molecule structures to be determined with much greater accuracy and precision. It is important to note that macromolecular crystallography generally involves lower resolution data, and is therefore, not an appropriate method for determining, *de novo*, the precise structure of a small molecule. However, in a macromolecular complex, particularly when an enzyme is involved, additional and unexpected chemistry can occur. High resolution protein–ligand structures can indeed reveal that small molecule ligands have undergone significant changes or modifications. The transformation from a prodrug to an active compound is exemplified by the conversion of the anti-tubercular drug isoniazid (isonicotinic acid hydrazide) into isonicotinic acyl-NADH by the mycobacterial enzyme KatG. In this case the active drug targets InhA, an enoyl-acyl-carrier protein reductase essential for the biosynthesis of mycobacterial cell walls [50].

The application of complete and accurate restraints for the ligand allows macromolecular crystallography to generate accurate and precise protein–ligand models. The geometry and stereochemistry of small molecule compounds do not differ substantial to those observed in protein structures [35, 36]. Therefore, standard libraries of small molecule structures, such as the Cambridge Structural Database (CSD), can be used to define restraints for both proteins and ligands alike [51]. For ligands that do not have representatives in the CSD, tools such as Mogul can be used to derive restraint parameters from small molecule structures containing similar fragments or sub-structures [52] (see "Refinement, stereochemical restraints and the data-to-parameter ratio" for more on restraints).

Several measures are available to assess the stereochemistry of ligands bound to protein models, including the deviation of an individual atom from an ideal target value, or for multiple atoms, the root mean squared deviation (RMSD) from an ideal value (Table 2). Deviations of bond lengths or bond angles larger than 4σ (about one in 10,000) are typically flagged as outliers and justify further inspection. Deviations from the experimental target

value can be conveniently expressed in multiples of the standard deviation from the respective target value using statistics such as the *Z*-score,

$$Z_j = \frac{x_j - x}{\sigma_x}$$

where <x> is the target value and σ is the standard deviation of the distribution. For multiple atoms, the RMSZ value indicates the number of standard deviations the experimentally determined values differ from their established mean,

$$RMSZ = \sqrt{\frac{1}{N}\sum_{j=1}^{N} Z_j^2}$$

Global deviations in RMSZ values for bond lengths and bond angles can be valuable for identifying problems with the selection of the weight of the restraints. In restrained reciprocal space refinement, the weight assigned to the restraint term varies and is used to balance the experimental X-ray data with the restraint parameters. In extreme cases involving atomic (<1.2 Å) or subatomic-resolution data (<1 Å) unrestrained refinement can be carried out, and the RMSZ for bond lengths and bond angles will approach the ideal values for small molecules (approximately 0.02 Å and 2.0°, respectively). Conversely, at lower resolution, with fewer data available, a higher weight of restraints is necessary to keep the model within physically realistic bounds. In such cases, the RMSZ of the refined atoms can be lower (but never higher) than the RMSZ of the restraint target values.

Individual, local bond length and bond angle deviations and outliers can be useful for identifying interesting biology or ligand chemistry, but often stereochemistry outliers identify a region of poor local fit. RMSD and RMSZ values are reported by several software packages including ValLigURL [13], wwPDB (http://bit.ly/1si1ZeL), Molprobity [53, 54], WHAT_CHECK [55] and Coot [27, 28]. The PDB Validation Task Force recommends bond lengths, bond angles and planarities, with a percentile ranking lower than 0.1 % and individual outliers with an RMSZ >5, are flagged in the wwPDB validation reports [45] (http://bit.ly/1si1ZeL).

In contrast to poor electron density-based metrics, which are generally a result of inadequate or overzealous modeling, poor ligand stereochemistry metrics are generally a result of a poor starting model or incorrectly defined restraints. It is essential that high quality ligand structures, and correspondingly reliable restraint files, are used as starting models for the crystallographic refinement process. The restraints file of a ligand must contain a complete specification of the ideal stereochemistry of the ligand. If an initial ligand model is far away from these ideal specifications, the refinement software will unlikely have a sufficiently large radius of convergence to bring these parameters closer to the ideal values. Similarly, if the chemistry of the ligand is incorrectly defined—which can happen more often than expected—or if individual restraints contain incorrect values, the refinement software will produce a chemically implausible model of the ligand. Errors such as these are largely avoidable with the use of automated ligand and restraint generation software. Many

resources are available for the automated generation of ligand models and ligand restraint files, including the Grade Web Server (Global Phasing Ltd.), HIC-UP [56], PRODRG [57], CORINA [58], A La Mode [59], AFITT [60], E-MSD [61], ChemDB [62], RESID [63], SWEET [64], Hess2FF [65], CSD [51], Inorganic Crystal Structure Database (ICSD), Crystallographic Open Database (COD), Ligand Depot [66], Zinc [67], MOGUL [52], RELIBASE [68] and PURY [69].

Structure validation software

Software packages available to assist the user in the validation of atomic models generated by crystallography are listed in Tables 3 and 4. Many of these packages are designed to assess the quality of the entire protein and ligand model and to guide the expert crystallographer during the course of model building and refinement (Table 4). To aid the end user of protein–ligand models, special software packages have been developed to specifically assess the local quality of the ligand model (Table 3). We discuss some ligand-specific validation tools that enable assessment of the electron density and stereochemistry including Twilight [15], VHELIBS [14], ValLigURL [13] and MotiveValidator/ValidatorDB [70, 71]. The use of open source crystallography-specific molecular graphics packages, such as PyMOL and Coot, are central to much of the validation process given the focus on electron density and crystallographic data. However, it is important to note that several other commercial molecular graphics, analysis and modeling packages are available, including Discovery Studio (http://bit.ly/1CtV084), Maestro (http://bit.ly/1Glvpo5), Molecular Operating Environment (http://bit.ly/1zslk53) and Molsoft ICM (http://bit.ly/1BiuNdn). The latter packages can also display electron density for inspection, but place more of an emphasis on the computational chemistry aspects of ligand validation, and are therefore, beyond the scope of this review and are not included in our tables.

Electron density validation software tools—Twilight is a standalone Python script useful for highlighting protein–ligand models which are not sufficiently supported by the experimental electron density [15]. The software can directly access models from the PDB and also includes a partially annotated database of protein–ligand models with an RSCC <0.6 [12]. Ligands can be sorted using various measures, including the *S* score,

$$\mathrm{SScore}=\frac{2}{(\mathrm{RSCC}/0.6)+(\mathrm{RESOL}/1.3)}$$

where RESOL is the resolution of the crystal structure. The RSCC is used to identify problematic ligands and the S score can be used for ranking of the ligands in a resolution dependent manner. The software is tightly coupled to the EDS [20] and PDB_REDO server [22] and electron density maps are automatically downloaded and loaded for visualization in Coot [28].

Validations Helper for Ligands and Binding Sites (VHELIBS) shares some features with Twilight, but additionally assesses the quality of the fit to the electron density for the protein residues in the ligand binding site (protein residues within 4.5 Å are defined as belonging to the active site) [14]. Well-modeled ligand binding sites are essential for modern drug

discovery techniques, making it possible to produce larger and more diverse libraries of protein–ligand models. Curated libraries of validated protein–ligand models, such as the Astex Diverse Set [16] and OpenEye Iridium [17], form a “gold standard” for many in silico methods such as ligand docking. Both VHELIBS and Twilight simplify and automate many of the steps required for the inspection of a ligand and its electron density and are intuitive to use for non-expert users. Therefore, such tools are recommended for the initial inspection of ligands and their fit to the electron density.

Stereochemistry validation software tools—Tools such as ValLigURL are available to automate the analysis of ligand stereochemistry [13]. VaLigURL is a webserver that compares the stereochemistry of a ligand of interest with other instances of the ligand in the PDB. A wealth of information is output by the server, including standard RSR metrics and links to the EDS server for automated visualization of the electron density in Astex Viewer [72]. Additionally, RMSD values are provided for atom displacements, bond lengths, bond angles, improper torsions and dihedral angles [13]. A rough overall measure of the quality of the ligand is also provided, the Q score, which takes into account the resolution of the structure, the quality of the fit to the electron density and the quality of the stereochemistry,

$$\mathrm{Qscore}=\mathrm{d}^2\rho(10\beta+0.1\alpha)$$

where, d is the resolution, ρ is the RSR value, β is the RMSD of the bond lengths and α is the RMSD of the bond angles. In addition to its use as a validation tool, ValLigURL can also be used as a rapid way of data mining the PDB for similar ligands and carrying out superposition of the ligands for further analysis. ValLigURL can also be used to investigate if the conformation of a particular ligand is unusual and worthy of further investigation.

Finally, ligands that are incorrectly annotated or have errors in their nomenclature are likely to be handled incorrectly during the refinement process as a result of missing or incorrectly applied restraints. ValidatorDB/MotiveValidator is a tool that automates nomenclature checks and also checks the ligand model for completeness [70, 71]. ValidatorDB is a database of pre-calculated validation results for non-standard protein residues and ligands in the PDB. The server reports atoms with incorrect chirality, atom substitutions, missing atoms, missing rings, alternate conformations, foreign atoms and atoms with different naming. Naming of ligand atoms is compared to standard ligand and protein residue models stored in the wwPDB Chemical Component Dictionary [73]. The use of standard ligand naming and nomenclature is strongly encouraged and can eliminate many of the common errors in ligand stereochemistry [74].

Recommended practices for interpretation of crystallographic data and validation of protein–ligand models

Many of the common errors in protein–ligand models can largely be mitigated through the diligent use of automated model building and refinement methods (see “Overall quality of protein-ligand models in the PDB”). Software such as ARP/wARP [75], SOLVE/RESOLVE [76] and Bucaneer [77] automate much of the protein model building process and reduce the potential for the introduction of errors. However, automated model building and refinement

of protein structures containing bound ligands is more of a challenge for several reasons. First, ligands often have more conformational flexibility than proteins and accurate restraints that reflect this flexibility must be applied. Second, ligands have more chemical variability than proteins. Correct definition of the atom types of the ligand is therefore essential; at typical macromolecular resolutions it is not always obvious from the electron density levels alone which atom type to assign. In a similar situation, inability to distinguish between nitrogen and oxygen atoms requires accounting for the chemical environment and hydrogen bonding to correctly assign the orientation of Asn, Gln, and His side chains [78].

In addition to automated atom typing and restraint file generation for ligands (see "Stereochemistry-based validation of protein-ligand models"), several methods are available to enable accurate ligand identification [79, 80], ligand building [81, 82], ligand placement and ligand refinement [37, 83, 84]. Such automated methods help enforce correct atom typing and correctly applied restraints and, in turn, result in better quality protein–ligand models as assessed by stereochemical validation methods. Additionally, automated ligand placement methods rely on the electron density for placement, and therefore, largely eliminate human errors such as overambitious ligand placement.

While it is not possible to anticipate every specific use of protein–ligand models, a number of key questions need to be asked when evaluating a crystallographic protein–ligand model (Fig. 4, Table 5). These questions fall into three main categories covering validation of the ligand electron density (Fig. 4a), validation of the ligand stereochemistry (Fig. 4b), and validation of the protein and ligand environment (Fig. 4c):

Validation of the ligand electron density

- Is the resolution of the diffraction data compatible with the level of detail claimed in the protein–ligand model description? Higher resolution data are generally preferable, but high resolution alone does not necessarily inform about the *local* quality of the ligand model. For low resolution structures, full validation is essential.
- Are the structure factors of the model available in the PDB? If not, validation against the primary evidence is not possible. Only validation of the stereochemistry and the protein environment is possible. Use the model with corresponding caution.
- Inspect the fit of the ligand model to the electron density using tools such as EDS [20], Twilight [15], VHELIBS [14], ValLigURL [13], wwPDB [45], Coot [27, 28] and PyMOL (Table 3) (see "Structure validation software" for more details). If the RSCC is >0.9, the electron density supports the ligand as modeled.
- If the RSCC scores are poor, more detailed questions about the ligand model must be answered. Is the ligand in the electron density? Are parts of the ligand model conjectural? Are the parts of the ligand model you are interested in still correctly modeled and trustworthy? Is the stereochemistry of the ligand model plausible? Is the protein environment possibly responsible for perturbations in the area of the ligand?

Validation of the ligand stereochemistry

- Inspect the ligand stereochemistry using a ligand validation tool such as ValLigURL [13], wwPDB [45], Molprobity [53, 54], WHAT_CHECK [55] or Coot [27, 28] (Table 3). If the scores are good then the ligand model has plausible stereochemistry supported by prior expectations.
- If the stereochemistry scores are bad, the chemistry of the ligand is likely implausible and further validation of the ligand and protein environment is advised. The primary cause of errors in stereochemistry is an incorrect starting model for the ligand or incorrect restraints.
- Trying to refine a ligand, with correct starting stereochemistry, into spurious electron density can also give rise to distorted ligand geometry. Restraints support the chemical integrity of the ligand during refinement but are not intended as a substitute for the absence of electron density.

Validation of the protein and ligand environment

- Inspect the environment of the binding site and the ligand to ensure that crystal packing artifacts are not affecting the modeling of the ligand. Are there any symmetry-related molecules forming contacts that could affect the ligand binding site? Tools such as PyMOL or Coot can be used to generate the symmetry-related molecules of the protein–ligand model.
- Check the occupancy of the ligand by inspection of the PDB file. If the occupancy is low, alternate conformations of the ligand may have been modeled, or the binding may be incomplete or partial.
- If the K_d of the ligand is in the high mM range, its solubility is low, or there are high concentrations of competing crystallization reagents present in the crystallization cocktail, then low occupancies of the ligand may result.
- Check the B-factors of the ligand using metrics such as the OWAB parameter reported by Twilight [15] (see “Refinement, stereochemical restraints and the data-to-parameter ratio”). Is the B-factor of the ligand similar to the neighboring protein atoms? If not, the ligand is probably disordered or partially occupied, and its placement in the electron density may be suspect.
- Inspect the environment of the ligand for correct chemistry including appropriately modeled double bonds, resonance structures and tautomerizations. Corrections can be applied to the restraints file as appropriate, but may require re-refinement.
- Does the ligand interact reasonably with the protein residues of the ligand binding pocket? Are the hydrogen bonds satisfied and reasonable? Does the ligand form other interactions with solvent molecules, ions, metals or stacking interactions with protein residues? Currently, no tools exist to fully automate such analysis and further analysis using molecular graphics visualization is required. Display and analysis tools such as LigPlot [85] and Molprobity [53, 54] can be used to assist in this process.

- Are multiple structures of the same ligand available? Superimpose the models and check for conformational similarities. Tools such as ValLigURL [13] can be used to find similar ligands in the PDB and carry out comparative analysis of the stereochemistry.

Trust, but verify…—Since the first protein structures were solved some 50 years ago, a gradual shift has taken place in macromolecular crystallography. At the dawn of protein crystallography, the main object of interest was the structure of the protein itself. Nowadays, the structure of the protein can often be somewhat secondary, and the real interest is placed in the ligand, co-factor, inhibitor, product, drug or other small-molecule bound to the protein. This presents a number of unique problems for the crystallographer including, (1) identification of an appropriate starting conformation of the ligand model that correctly fits the experimental electron density, (2) selection of appropriate restraints to ensure the ligand maintains plausible stereochemistry during refinement, (3) correct modeling of the protein environment including disorder, displacements, crystal packing and occupancies, and (4) validation of the final model to ensure that (1) through (3) remain correct. Validation of protein structures is relatively straightforward, as the repertoire of conformations adopted by proteins and individual residues is relatively limited [86]. Conversely, small-molecule ligands have an almost unlimited variability in chemical character and conformational freedom, and are extremely difficult to generalize; individual restraints must be generated and applied for each ligand under study.

In an effort to reassure the end user of crystallographic protein–ligand models, many of the problems discussed in this review have been highlighted in the crystallographic community, and efforts are underway to minimize such errors [9, 12, 37, 45, 87, 88]. Best practices such as mandatory deposition of structure factor amplitudes is now enforced and further discussions are underway to allow deposition of the primary diffraction images [89]. Furthermore, a PDB validation task force has been assembled and several of their recommendations are now being applied to new structures deposited in the PDB [45]. Both the end user and expert crystallographer alike, now have an unprecedented plethora of tools available for the validation of protein–ligand models, and they are strongly encouraged to use them (Tables 2, 3 and 4). While protein–ligand models, in general, can be *trusted*, it is good practice that each specific ligand model is *verified—Trust, but verify*.

Acknowledgments

MCD acknowledges support from the NIH, National Institute of General Medical Sciences, Protein Structure Initiative under Grant Number U54 GM094586. BR acknowledges support from the European Union under a FP7 Marie Curie People Action, Grant PIIF-GA-2011–300025 (SAXCESS).

References

1. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF Jr, Brice MD, et al. The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol. 1977; 112:535–542. [PubMed: 875032]
2. Berman H. The Protein Data Bank: a historical perspective. Acta Crystallogr A. 2008; 64:88–95. [PubMed: 18156675]

3. Berman H, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. Nat Struct Biol. 2003; 10:980. [PubMed: 14634627]
4. Gutmanas A, Alhroub Y, Battle GM, Berrisford JM, Bochet E, et al. PDBe: protein Data Bank in Europe. Nucleic Acids Res. 2014; 42:D285–D291. [PubMed: 24288376]
5. Henderson R, Sali A, Baker ML, Carragher B, Devkota B, et al. Outcome of the first electron microscopy validation task force meeting. Structure. 2012; 20:205–214. [PubMed: 22325770]
6. Dutta, S.; Burkhardt, K.; Swaminathan, GJ.; Kosada, T.; Henrick, K., et al. Data deposition and annotation at the Worldwide Protein Data Bank. In: Kobe, B.; Guss, M.; Huber, T., editors. Structural proteomics: high-throughput methods. New York, NY: Humana Press/Springer; 2008.
7. Carvalho AL, Trincao J, Romao MJ. X-ray crystallography in drug discovery. Methods Mol Biol. 2009; 572:31–56. [PubMed: 20694684]
8. Zheng H, Hou J, Zimmerman MD, Wlodawer A, Minor W. The future of crystallography in drug discovery. Expert Opin Drug Discov. 2014; 9:125–137. [PubMed: 24372145]
9. Davis AM, St-Gallay SA, Kleywegt GJ. Limitations and lessons in the use of X-ray structural information in drug design. Drug Discov Today. 2008; 13:831–841. [PubMed: 18617015]
10. Krishnan VV, Rupp B. Macromolecular structure determination: comparison of X-ray crystallography and NMR. Spectroscopy. eLS. 2012 doi: 10.1002/9780470015902.a9780470002716.pub9780470015902.
11. Davies TG, Tickle IJ. Fragment screening using X-ray crystallography. Top Curr Chem. 2012; 317:33–59. [PubMed: 21678136]
12. Pozharski E, Weichenberger CX, Rupp B. Techniques, tools and best practices for ligand electron-density analysis and results from their application to deposited crystal structures. Acta Crystallogr D. 2013; 69:150–167. [PubMed: 23385452]
13. Kleywegt GJ, Harris MR. ValLigURL: a server for ligand-structure comparison and validation. Acta Crystallogr. 2007; 63:935–938.
14. Cereto-Massague A, Ojeda MJ, Joosten RP, Valls C, Mulero M, et al. The good, the bad and the dubious: VHELIBS, a validation helper for ligands and binding sites. J Cheminform. 2013; 5:36. [PubMed: 23895374]
15. Weichenberger CX, Pozharski E, Rupp B. Visualizing ligand molecules in twilight electron density. Acta Crystallogr. 2013; F69:195–200.
16. Hartshorn MJ, Verdonk ML, Chessari G, Brewerton SC, Mooij WT, et al. Diverse, high-quality test set for the validation of protein-ligand docking performance. J Med Chem. 2007; 50:726–741. [PubMed: 17300160]
17. Warren GL, Do TD, Kelley BP, Nicholls A, Warren SD. Essential considerations for using protein-ligand structures in drug discovery. Drug Discov Today. 2012; 17:1270–1281. [PubMed: 22728777]
18. Hawkins PCD, Warren GL, Skillman AG, Nicholls A. How to do an evaluation: pitfalls and traps. J Comput Aided Mol Des. 2008; 22:179–190. [PubMed: 18217218]
19. Westbrook JD, Fitzgerald PM. The PDB format, mmCIF, and other data formats. Methods Biochem Anal. 2003; 44:161–179. [PubMed: 12647386]
20. Kleywegt GJ, Harris MR, Zou JY, Taylor TC, Wahlby A, et al. The uppsala electron-density server. Acta Crystallogr. 2004; D60:2240–2249.
21. Joosten RP, Joosten K, Murshudov GN, Perrakis A. PDB_REDO: constructive validation, more than just looking for errors. Acta Crystallogr D. 2012; 68:484–496. [PubMed: 22505269]
22. Joosten RP, Long F, Murshudov GN, Perrakis A. The PDB_REDO server for macromolecular structure model optimization. IUCrJ. 2014; 1:213–220.
23. Rhodes, G. Crystallography made crystal clear. London, UK: Academic Press; 2006.
24. Rupp, B. Biomolecular crystallography: principles, practice, and application to structural biology. New York: Garland Science; 2009.
25. Elsliger MA, Deacon AM, Godzik A, Lesley SA, Wooley J, et al. The JCSG high-throughput structural biology pipeline. Acta Crystallogr. 2010; F66:1137–1142.

26. Weichenberger CX, Rupp B. Ten years of probabilistic estimates of biocrystal solvent content: new insights via nonparametric kernel density estimate. Acta Crystallogr D. 2014; 70:1579–1588. [PubMed: 24914969]
27. Debreczeni JE, Emsley P. Handling ligands with coot. Acta Crystallogr. 2012; D68:425–430.
28. Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of coot. Acta Crystallogr D. 2010; 66:486–501. [PubMed: 20383002]
29. Krissinel E. Crystal contacts as nature's docking solutions. J Comput Chem. 2010; 31:133–143. [PubMed: 19421996]
30. Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. J Mol Biol. 2007; 372:774–797. [PubMed: 17681537]
31. Danley D. Crystallization to obtain protein-ligand complexes for structure-aided drug design. Acta Crystallogr D. 2006; 62:569–575. [PubMed: 16699182]
32. Muller Y. Unexpected features in the Protein Data Bank entries 3qd1 and 4i8e: the structural description of the binding of the serine-rich repeat adhesin GspB to host cell carbohydrate receptor is not a solved issue. Acta Crystallogr. 2013; F69:1071–1076.
33. Tronrud D, Allen J. Reinterpretation of the electron density at the site of the eighth bacteriochlorophyll in the FMO protein from Pelodictyon phaeum. Photosynthesis Res. 2012; 112:71–74.
34. Gokulan K, Khare S, Ronning D, Linthicum SD, Sacchettini JC, et al. Co-crystal structures of NC6.8 Fab identify key interactions for high potency sweetener recognition: implications for the design of synthetic sweeteners. Biochemistry. 2005; 44:9889–9898. [PubMed: 16026161]
35. Engh RA, Huber R. Accurate bond and angle parameters for X-ray protein structure refinement. Acta Crystallogr A. 1991; 47:392–400.
36. Engh, RA.; Huber, R. International tables for crystallography. Arnold MGRE. , editor. Dordrecht: Kluwer; 2001. p. 382-392.
37. Kleywegt GJ. Crystallographic refinement of ligand complexes. Acta Crystallogr D. 2007; 63:94–100. [PubMed: 17164531]
38. Brunger AT. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. Nature. 1992; 355:472–475. [PubMed: 18481394]
39. Karplus PA, Diederichs K. Linking crystallographic model and data quality. Science. 2012; 336:1030–1033. [PubMed: 22628654]
40. Tickle IJ. Statistical quality indicators for electron-density maps. Acta Crystallogr D. 2012; 68:454–467. [PubMed: 22505266]
41. Read RJ. Improved Fourier coefficients for maps using phases from partial structures with errors. Acta Crystallogr A. 1986; 42:140–149.
42. Hodel A, Kim S-H, Brunger AT. Model bias in macromolecular structures. Acta Crystallogr D. 1992; 48:851–858.
43. Branden C-I, Alwyn Jones T. Between objectivity and subjectivity. Nature. 1990; 343:687–689.
44. Jones TA, Zou JY, Cowan SW, Kjeldgaard M. Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Crystallogr A. 1991; 47:110–119. [PubMed: 2025413]
45. Read Randy J, Adams Paul D, Arendall Iii WB, Brunger Axel T, Emsley P, et al. A new generation of crystallographic validation tools for the protein data bank. Structure. 2011; 19:1395–1412. [PubMed: 22000512]
46. Rupp B, Segelke BW. Questions about the structure of the botulinum neurotoxin B light chain in complex with a target peptide. Nat Struct Biol. 2001; 8:643–664.
47. Hanson MA, Oost TK, Sukonpan C, Rich DH, Stevens RC. Structural basis for BABIM inhibition of botulinum neurotoxin type B protease. J Am Chem Soc. 2002; 124:10248.
48. Hanson MA, Stevens RC. Retraction: cocrystal structure of synaptobrevin-II bound to botulinum neurotoxin type B at 2.0 A resolution. Nat Struct Mol Biol. 2009; 16:795. [PubMed: 19578378]
49. Rupp B. Scientific inquiry and inference in macromolecular crystallography. Acta Crystallogr A. 2008; 64:C81.

50. Vilcheze C, Wang F, Arai M, Hazbon MH, Colangeli R, et al. Transfer of a point mutation in Mycobacterium tuberculosis inhA resolves the target of isoniazid. Nat Med. 2006; 12:1027–1029. [PubMed: 16906155]
51. Allen FH. The Cambridge structural database: a quarter of a million crystal structures and rising. Acta Crystallogr B. 2002:380–388. [PubMed: 12037359]
52. Bruno IJ, Cole JC, Kessler M, Luo J, Motherwell WDS, et al. Retrieval of crystallographically-derived molecular geometry information. J Chem Inf Comput Sci. 2004; 44:2133–2144. [PubMed: 15554684]
53. Chen VB, Arendall WB III, Headd JJ, Keedy DA, Immormino RM, et al. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D. 2010; 66:12–21. [PubMed: 20057044]
54. Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, et al. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Res. 2007; 35:W375–W383. [PubMed: 17452350]
55. Hooft RW, Vriend G, Sander C, Abola EE. Errors in protein structures. Nature. 1996; 381:272. [PubMed: 8692262]
56. Kleywegt GJ, Jones TA. Databases in protein crystallography. Acta Crystallogr. 1998; D54:1119–1131.
57. van Aalten DM, Bywater R, Findlay JB, Hendlich M, Hooft RW, et al. PRODRG, a program for generating molecular topologies and unique molecular descriptors from coordinates of small molecules. J Comput Aided Mol Des. 1996; 10:255–262. [PubMed: 8808741]
58. Gasteiger J, Rudolph C, Sadowski J. Automatic generation of 3D-atomic coordinates for organic molecules. Tetrahedron Comput Methodol. 1990; 3:537–547.
59. Clowney L, Westbrook JD, Berman HM. CIF applications. XI. A la mode: a ligand and monomer object data environment. I. Automated construction of mmCIF monomer and ligand models. Appl Cryst. 1999; 32:125–133.
60. Peat TS, Christopher J, Schmidt K. AFITT- working with good chemistry. Acta Crystallogr A. 2005; 61:C165.
61. Golovin A, Oldfield TJ, Tate JG, Velankar S, Barton GJ, et al. E-MSD: an integrated data resource for bioinformatics. Nucleic Acids Res. 2004; 32:D211–D216. [PubMed: 14681397]
62. Chen J, Swamidass SJ, Dou Y, Bruand J, Baldi P. ChemDB: a public database of small molecules and related chemoinformatics resources. Bioinformatics. 2005; 21:4133–4139. [PubMed: 16174682]
63. Garavelli JS. The RESID database of protein modifications as a resource and annotation tool. Proteomics. 2004; 4:1527–1533. [PubMed: 15174122]
64. Bohne A, Lang E, von der Lieth CW. SWEET: WWW-based rapid 3D construction of oligo- and polysaccharides. Bioinformatics. 1999; 15:767–768. [PubMed: 10498779]
65. Nilsson K, Lecerof D, Sigfridsson E, Ryde U. An automatic method to generate force-field parameters for heterocompounds. Acta Crystallogr D. 2003; 59:274–289. [PubMed: 12554938]
66. Feng Z, Chen L, Maddula H, Akcan O, Oughtred R, et al. Ligand depot: a data warehouse for ligands bound to macromolecules. Bioinformatics. 2004; 20:2153–2155. [PubMed: 15059838]
67. Irwin JJ, Shoichet BK. ZINC–a free database of commercially available compounds for virtual screening. J Chem Inf Model. 2005; 45:177–182. [PubMed: 15667143]
68. Hendlich M, Bergner A, Günther J, Klebe G. Relibase: design and development of a database for comprehensive analysis of protein-ligand interactions. J Mol Biol. 2003; 326:607–620. [PubMed: 12559926]
69. Andrejasic M, Praaenikar J, Turk D. PURY: a database of geometric restraints of hetero compounds for refinement in complexes with macromolecular structures. Acta Crystallogr D. 2008; 64:1093–1109. [PubMed: 19020347]
70. Sehnal D, Svobodová Vařeková R, Pravda L, Ionescu C-M, Geidl S, et al. ValidatorDB: database of up-to-date validation results for ligands and non-standard residues from the Protein Data Bank. Nucleic Acids Res. 2015; 43:D369–D375. [PubMed: 25392418]

71. Varekova RS, Jaiswal D, Sehnal D, Ionescu CM, Geidl S, et al. MotiveValidator: interactive web-based validation of ligand and residue structure in biomolecular complexes. Nucleic Acids Res. 2014; 42:W227–W233. [PubMed: 24848013]
72. Hartshorn MJ. AstexViewer: a visualisation aid for structure-based drug design. J Comput Aided Mol Des. 2002; 16:871–881. [PubMed: 12825620]
73. Henrick K, Feng Z, Bluhm WF, Dimitropoulos D, Doreleijers JF, et al. Remediation of the protein data bank archive. Nucleic Acids Res. 2008; 36:D426–D433. [PubMed: 18073189]
74. Jaskolski M. On the propagation of errors. Acta Crystallogr D. 2013; 69:1865–1866. [PubMed: 24100306]
75. Langer G, Cohen SX, Lamzin VS, Perrakis A. Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. Nat Protoc. 2008; 3:1171–1179. [PubMed: 18600222]
76. Terwilliger T. SOLVE and RESOLVE: automated structure solution, density modification and model building. J Synchrotron Radiat. 2004; 11:49–52. [PubMed: 14646132]
77. Cowtan K. Completion of autobuilt protein models using a database of protein fragments. Acta Crystallogr D. 2012; 68:328–335. [PubMed: 22505253]
78. Weichenberger CX, Sippl MJ. NQ-Flipper: recognition and correction of erroneous asparagine and glutamine side-chain rotamers in protein structures. Nucleic Acids Res. 2007; 35:W403–W406. [PubMed: 17478502]
79. Carolan CG, Lamzin VS. Automated identification of crystallographic ligands using sparse-density representations. Acta Crystallogr D. 2014; 70:1844–1853. [PubMed: 25004962]
80. Terwilliger TC, Adams PD, Moriarty NW, Cohn JD. Ligand identification using electron-density map correlations. Acta Crystallogr D. 2007; 63:101–107. [PubMed: 17164532]
81. Aishima J, Russel DS, Guibas LJ, Adams PD, Brunger AT. Automated crystallographic ligand building using the medial axis transform of an electron-density isosurface. Acta Crystallogr D. 2005; 61:1354–1363. [PubMed: 16204887]
82. Evrard GX, Langer GG, Perrakis A, Lamzin VS. Assessment of automatic ligand building in ARP/wARP. Acta Crystallogr D. 2007; 63:108–117. [PubMed: 17164533]
83. Wlodek S, Skillman AG, Nicholls A. Automated ligand placement and refinement with a combined force field and shape potential. Acta Crystallogr D. 2006; 62:741–749. [PubMed: 16790930]
84. Klei HE, Moriarty NW, Echols N, Terwilliger TC, Baldwin ET, et al. Ligand placement based on prior structures: the guided ligand-replacement method. Acta Crystallogr D. 2014; 70:134–143. [PubMed: 24419386]
85. Laskowski RA, Swindells MB. LigPlot + : multiple ligand-protein interaction diagrams for drug discovery. J Chem Inf Model. 2011; 51:2778–2786. [PubMed: 21919503]
86. Kleywegt GJ. Validation of protein crystal structures. Acta Crystallogr D. 2000; 56:249–265. [PubMed: 10713511]
87. Dauter Z, Wlodawer A, Minor W, Jaskolski M, Rupp B. Avoidable errors in deposited macromolecular structures: an impediment to efficient data mining. IUCrJ. 2014; 1:179–193.
88. Liebeschuetz J, Hennemann J, Olsson T, Groom CR. The good, the bad and the twisted: a survey of ligand geometry in protein crystal structures. J Comput Aided Mol Des. 2012; 26:169–183. [PubMed: 22246295]
89. Baker E, Dauter Z, Guss M, Einspahr H. Deposition of diffraction images to be discussed at the Open Meeting of the Commission on Biological Macromolecules of the IUCr in Osaka. Acta Crystallogr. 2008; F64:231–232.
90. Cruickshank DW. Remarks about protein structure precision. Acta Crystallogr D. 1999; 55:583–601. [PubMed: 10089455]
91. Laskowski RA, Macarthur MW, Moss DS, Thornton JM. {PROCHECK}: a program to check the stereochemical quality of protein structures. Appl Cryst. 1993; 26:283–291.
92. Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. J Mol Biol. 1963; 7:95–99. [PubMed: 13990617]

93. Sheffler W, Baker D. RosettaHoles: rapid assessment of protein core packing for structure prediction, refinement, design, and validation. Protein Sci. 2009; 18:229–239. [PubMed: 19177366]

94. Sheffler W, Baker D. RosettaHoles2: a volumetric packing measure for protein structure refinement and validation. Protein Sci. 2010; 19:1991–1995. [PubMed: 20665689]

95. Debye P. Interferenz von Röntgenstrahlen und Wärmebewegung. Ann Phys. 1913; 348:49–92.

96. Waller I. Zur Frage der Einwirkung der Wärmebewegung auf die Interferenz von Röntgenstrahlen. Zeitschrift für Physik. 1923; 17:398–408.

97. Lutteke T, Frank M, von der Lieth CW. Carbohydrate Structure Suite (CSS): analysis of carbohydrate 3D structures derived from the PDB. Nucleic Acids Res. 2005; 33:D242–D246. [PubMed: 15608187]

98. Lutteke T, von der Lieth CW. pdb-care (PDB carbohydrate residue check): a program to support annotation of complex carbohydrate structures in PDB files. BMC Bioinformatics. 2004; 5:69. [PubMed: 15180909]

99. Collaborative Computational Project, Number 4. Acta Cryst. 1994; D50:760–763. http://dx.doi.org/10.1107/S0907444994003112.

100. Smart OS, Womack TO, Flensburg C, Keller P, Paciorek W, et al. Exploiting structure similarity in refinement: automated NCS and target-structure restraints in BUSTER. Acta Crystallogr D. 2012; 68:368–380. [PubMed: 22505257]

101. Vriend G. WHAT IF: a molecular modeling and drug design program. J Mol Graph. 1990; 8(52–56):29.

102. Adams PD, Afonine PV, Bunkoczi G, Chen VB, Davis IW, et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D. 2010; 66:213–221. [PubMed: 20124702]

103. Vaguine AA, Richelle J, Wodak SJ. SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. Acta Crystallogr D. 1999; 55:191–205. [PubMed: 10089410]

104. Luthy R, Bowie JU, Eisenberg D. Assessment of protein models with three-dimensional profiles. Nature. 1992; 356:83–85. [PubMed: 1538787]

105. Urzhumtseva L, Afonine PV, Adams PD, Urzhumtsev A. Crystallographic model quality at a glance. Acta Crystallogr D. 2009; 65:297–300. [PubMed: 19237753]

106. Bhattacharya A, Tejero R, Montelione GT. Evaluating protein structures determined by structural genomics consortia. Proteins. 2007; 66:778–795. [PubMed: 17186527]

107. Sippl MJ. Recognition of errors in three-dimensional structures of proteins. Proteins. 1993; 17:355–362. [PubMed: 8108378]

108. Wiederstein M, Sippl MJ. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res. 2007; 35:W407–W410. [PubMed: 17517781]

109. Colovos C, Yeates TO. Verification of protein structures: patterns of nonbonded atomic interactions. Protein Sci. 1993; 2:1511–1519. [PubMed: 8401235]

110. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. The protein data bank. Nucleic Acids Res. 2000; 28:235–242. [PubMed: 10592235]

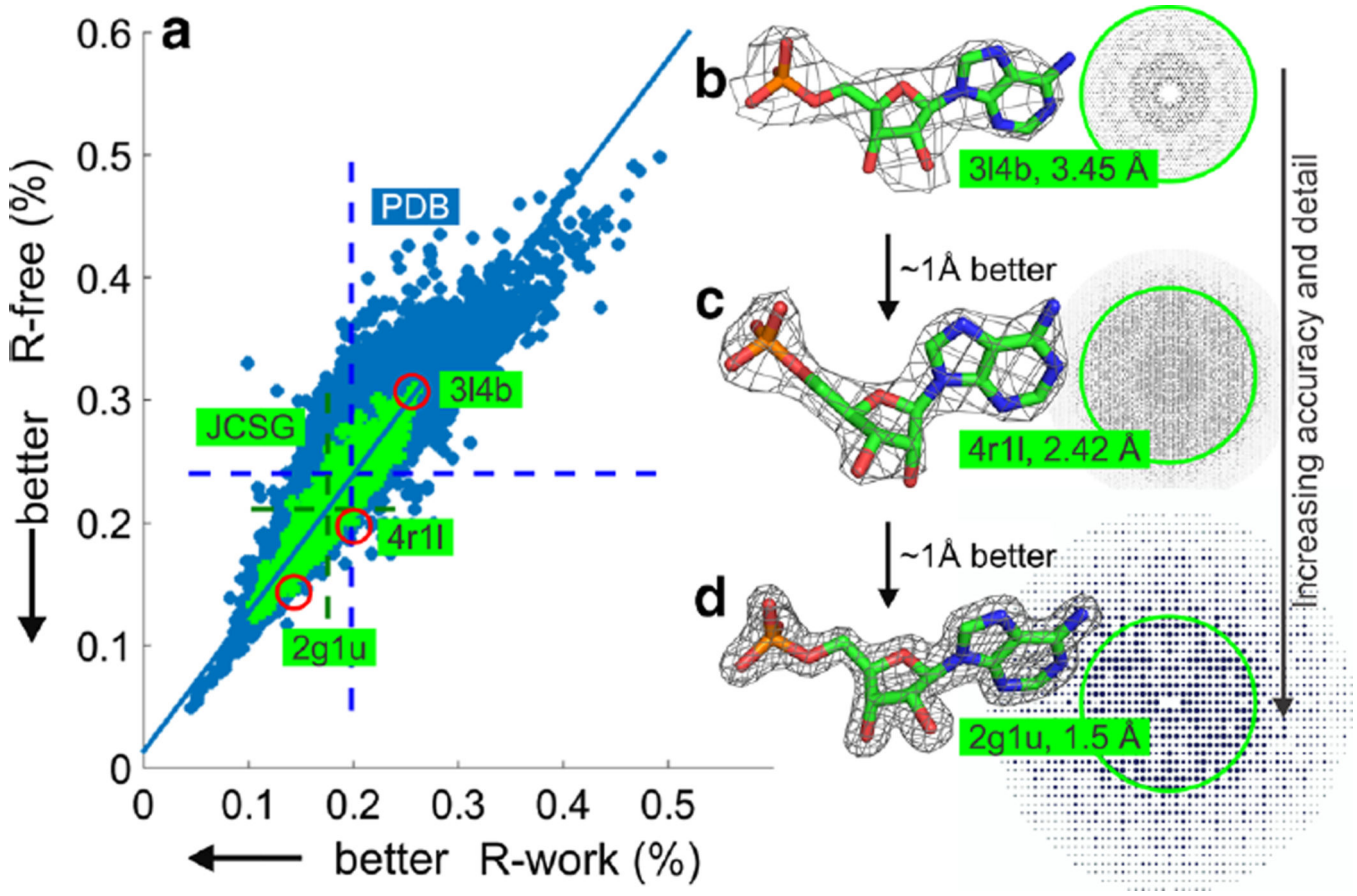

Fig. 1.

Overall quality of protein–ligand models **a** *Global* assessment of protein–ligand model quality based on R-work (a linear residual measuring the difference between the observed and the calculated diffraction data) and the corresponding cross-validation measure R-free. 85,623 data points are shown for all models deposited in the PDB (*blue*) and 1466 models deposited by the JCSG (*green*). Data are available from the PDB (http://www.rcsb.org). *Dotted lines* show the distribution of the data points for the entire PDB dataset (*blue dotted line*) and the JCSG (*green dotted line*). **b** through **d** demonstrate the increasing accuracy and resolution of the diffraction data for a series of protein–ligand models of adenosine monophosphate (AMP, shown as sticks) determined by the JCSG. Carbon atoms are shown in *green*, nitrogen in *blue*, oxygen in *red* and phosphorus in orange. The diffraction data are shown, with the *green circles* highlighting a resolution of 3.45 Å. The σA-weighted $2mF_o–DF_c$ electron density map is shown as a grey mesh contoured at 2σ. The corresponding R-free and R-work values for these models are highlighted by *red circles* in panel (**a**)

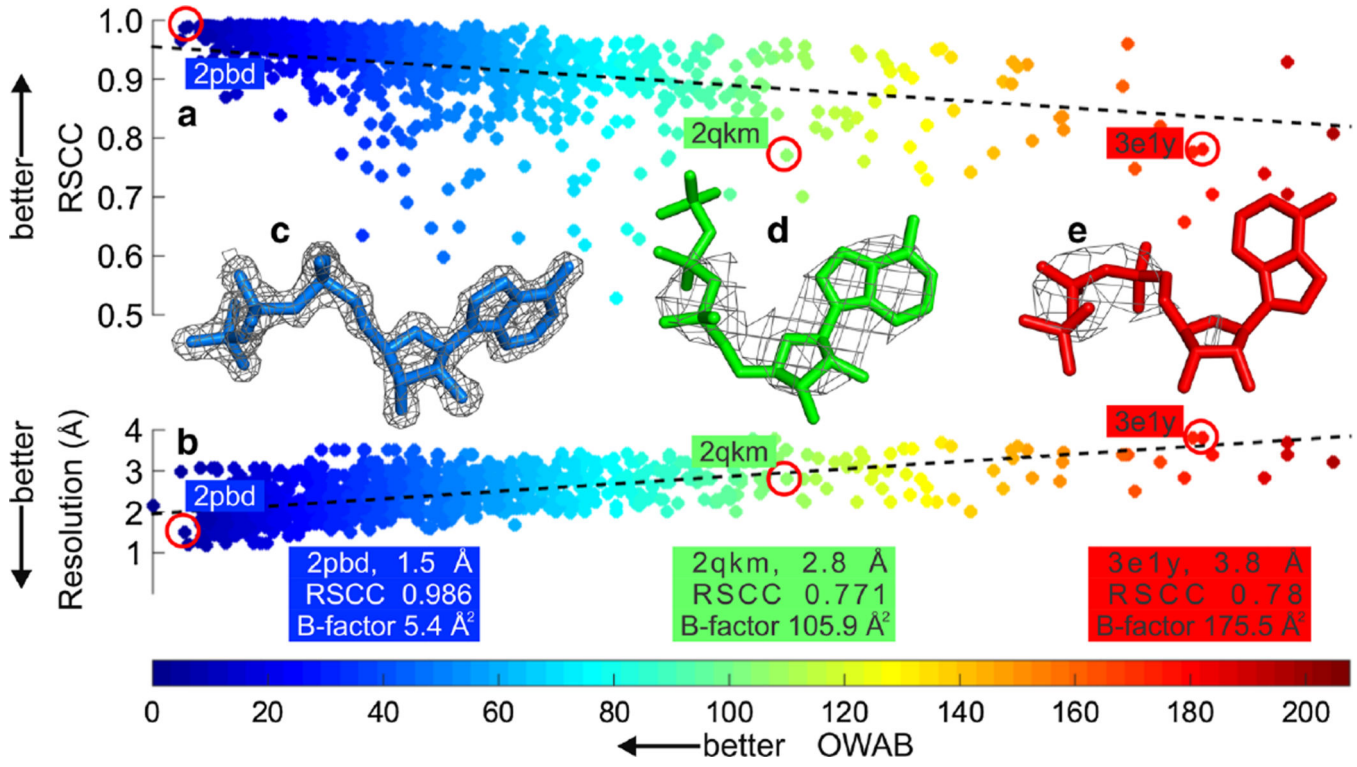

Fig. 2.

Local versus *global* measures of error in protein–ligand models **a** Plot of RSCC against OWAB and **b** resolution against OWAB. 1,183 data points are shown in each plot for adenosine triphosphate (ATP). RSCC is a metric used to determine the *local* measure of ligand fit to the electron density and OWAB is a *local* measure of the displacement of the atoms of the ligand. RSCC and OWAB values were calculated using Twilight [12, 15]. The latest pre-calculated Twilight data are available at http://bit.ly/1shcwu4. **c** through **e** demonstrate the sensitivity of the *local* RSCC and OWAB metrics to *global* metrics such as resolution. The ATP is shown as sticks and a σA-weighted $2mF_o–DF_c$ electron density map is shown as a grey mesh contoured at 2σ. The corresponding RSCC and resolution values for the ATP ligands are highlighted by red circles in **a** and **b**, respectively

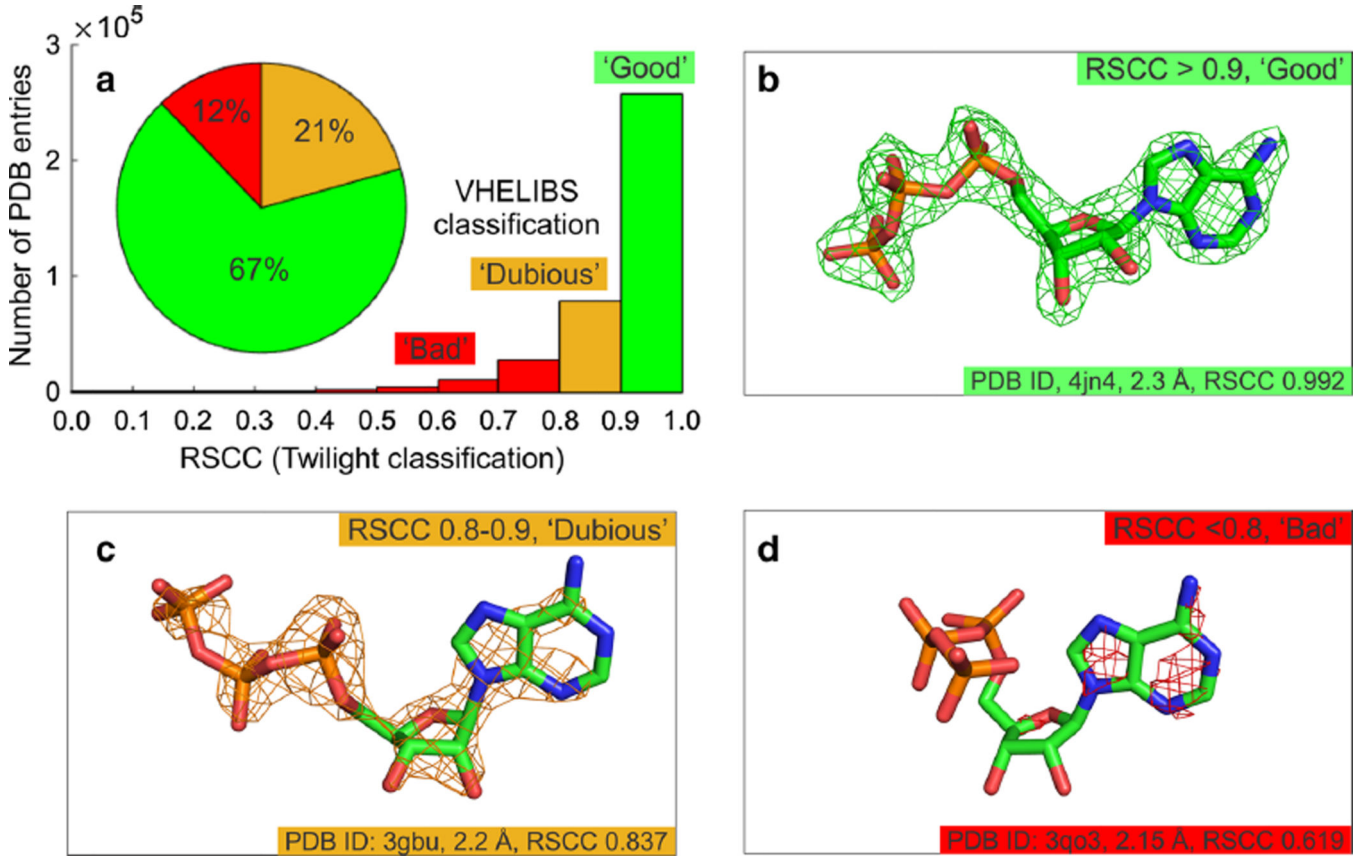

Fig. 3.

Electron density-based validation of protein–ligand models **a** Plots showing the distribution of the fits of ligands to their experimental electron density. Classifications were determined using Twilight [12, 15] and VHELIBS [14]. **b** through **d** demonstrate the fit of the ligand to the experimental electron density for a series of protein–ligand models of adenosine triphosphate (ATP, shown as sticks). Carbon atoms are shown in green, nitrogen in blue, oxygen in red and phosphorus in orange. The σA-weighted $2mF_o–DF_c$ electron density map is shown as a mesh contoured at 2σ. Data are shown in Table 1

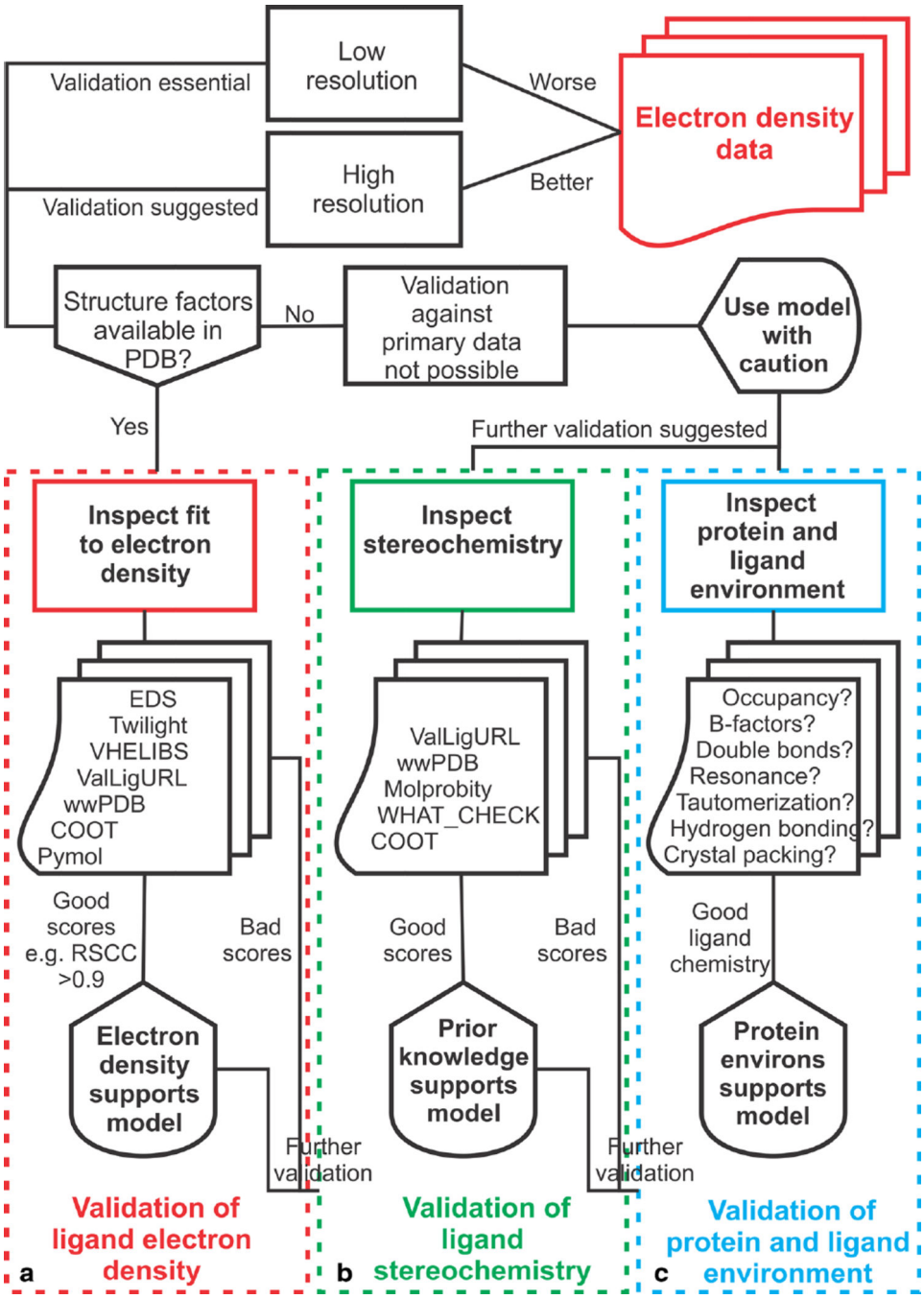

Fig. 4.

Recommended practices for interpretation of crystallographic data and validation of protein–ligand models. Flowchart detailing a pathway of recommended practices starting with the electron density data (*top right*). Important validation steps include, **a** Inspect the electron density and validate that the electron density supports the ligand model. **b** Inspect the stereochemistry of the ligand and validate that the ligand model is supported by prior

expectations and finally, **c** inspect the protein and ligand environment and validate that the environment supports the ligand model

Table 1

Electron density-based validation of protein–ligand models

Scores			Classification		Remedy
RSCC	**% of structures**	**Predicted number of PDB**	**Twilight**	**VHELIBS**	
1.0–0.9	67	~46,900	Ligand fits density	‘Good’	Ligand model good to use
<0.9–0.8	21	~14,700	Ligand fits density partially	‘Dubious’	Ligand over-modeled or may benefit from further refinement
<0.8–0.7	7	~4,900	Significant parts of ligand not in density	‘Bad’	Use ligand model with caution!
<0.7–0.6	3	~2,100	Very poor fit of ligand to density		
<0.6–0.5	1	~700	Improbably poor fit of ligand to density, density almost absent		
<0.5	1	~700	Catastrophic fit of ligand to density, density completely absent		

Real Space Correlation Coefficient (RSCC) values were calculated using Twilight [12, 15] and classifications determined using Twilight and VHELIBS [14]. The percentage of structures was determined using a curated set of 382,588 ligands from the PDB with electron density available in the EDS [20]. Pre-calculated data available for download from http://bit.ly/1shcwu4. The predicted number of PDB entries was calculated using the percentage of ligands in each category in the curated set and applying to the total PDB count (just over 70,000 protein–ligand models in the PDB as of Nov, 2014). This is likely a conservative estimate based on the assumption that each PDB files contains just one ligand

J Comput Aided Mol Des. Author manuscript; available in PMC 2016 September 01.

Table 2

Local and *global* structure validation measures

Validation measure		*Local*	*Global*	Reference
Quality and fit of ligand	Real Space Correlation Coefficient (RSCC)	●	●	N/A
	Real Space R-value (RSR)	●	●	[43, 44]
	Real Space Observed Density Z-Score (RSZO)	●	●	[40]
	Occupancy-Weighted Average B-factor (OWAB)	●		[15, 20]
	S score	●		[15]
	Q score	●		[13]
	Local Ligand Density Fit (LLDF)	●		http://bit.ly/1si1ZeL
	deviation of bond lengths	●		[35]
	deviation of bond angles	●		
Quality of model and data	R-value		●	N/A
	R-free		●	[38]
	CC*		●	[39]
	Real Space R-value (RSR)	●	●	[44]
	Diffraction-data Precision Indicator (DPI)		●	[90]
Quality of model and stereochemistry	RMSD/Z bond lengths	●	●	[35]
	RMSD/Z bond angles	●	●	
	G-factor	●		[91]
	Clashscore	●	●	[53, 54]
	Ramachandran outliers	●		[92]
Other measures of model and data quality	Volumetric packing scores	●	●	[93, 94]
	B-factor	●	●	[95, 96]
	Resolution		●	N/A

Local validation measures are metrics used to validate the quality of the ligand model, and *global* validation measures are used to assess the overall quality of the protein–ligand model. Many *local* measures can also be averaged to provide a (less informative) *global* measure

Table 3

Protein-ligand validation software

Validation	Package	Description	URL	Reference
Quality and fit of ligand	Twilight	Visual analysis of ligand density	http://bit.ly/1shcwu4	[12, 15]
	VHELIBS	Fit of ligands and binding sites	http://bit.ly/1t5szxl	[14]
	ValLigURL server	Compare conformations of ligands in the PDB	http://bit.ly/1v80yoS	[13]
	MotiveValidator and ValidatorDB	Interactive web-based validation of ligands and residues	http://bit.ly/1tNd7Vs	[70]
	PDB_REDO	Updated and optimised X-ray structure models and maps	http://bit.ly/1pVYQQA	[21, 22]
	wwPDB	wwPDB Validation Server	http://bit.ly/1xdfoZN and http://bit.ly/1si1ZeL	[45]
	PDB-CARE and CARP	Checks glycan nomenclature and stereochemistry	http://bit.ly/1pW9gQ0	[97, 98]
	LIGPLOT	Ligand–protein interaction diagrams	http://bit.ly/1qwZ7cS	[85]
	EDS	Electron density server	http://bit.ly/1Ddnlkg	[20]
	EDSTATS	Statistical quality indicators of electron density maps	N/A	[40]
	OVERLAPMAP	Average of two maps	http://bit.ly/ZZUSk3	[99]
	BUSTER	Refinement of proteins and ligands	http://bit.ly/1w8NX1Q	[100]
	WHAT_IF/WHAT_CHECK	Protein and ligand verification tools	http://bit.ly/1F0j19Y	[55, 101]

Software that is particularly useful for ligand validation is highlighted. For general protein structure validation software see Table 4

Table 4

Protein structure validation software

Validation	Package	Description	URL	Reference
Overall fit of model to electron density	PHENIX	Realspace map correlations and geometry outliers	http://bit.ly/ZukF31	[102]
	Coot	Molecular graphics package for model building and validation	http://bit.ly/1wJNWBy	[27, 28]
	SFCHECK	Check of structure factors	http://bit.ly/1neH8fk	[103]
Overall quality of model and stereochemistry	MolProbity	All-atom structure validation for macromolecules	http://bit.ly/1o1PuHW	[53, 54]
	PDB-CARE and CARP	Checks glycan nomenclature and stereochemistry	http://bit.ly/1pW9gQ0	[97, 98]
	PROCHECK	Stereochemical quality checks	http://bit.ly/1tGOJjU	[90]
Other servers and validation suites	PDB_REDO	Updated and optimised X-ray structure models and maps	http://bit.ly/1pVYQQA	[21, 22]
	Verify3D	Comparison of atomic model to its sequence	N/A	[104]
	POLYGON	Compares model quality with others in the PDB	N/A	[105]
	JCSG QC server	Automated quality control check	http://stanford.io/1xun5u2	N/A
	Protein Structure Validation Suite (PSVS)	Assessment of protein structures generated by NMR and X-ray methods	http://bit.ly/1vx2TKa	[106]
	ProSA-web	Knowledge-based potentials for assessing atomic models	http://bit.ly/1sU7hRx	[107, 108]
	ERRAT structure verification server	Verification of atomic model using patterns of non-bonded interactions	N/A	[109]
	ADIT validation server	PDB pre-deposition validation suite	http://bit.ly/1rd87Dy	[110]
	wwPDB	wwPDB validation server	http://bit.ly/1xdfoZN and http://bit.ly/1si1ZeL	[45]

Software that is useful for overall protein structure validation is highlighted. For specialized ligand validation software see Table 3

Table 5

Important validation decisions

Decision	Available tools	Validation parameter
Does the electron density support the ligand model?	Inspect electron density using Twilight, PyMOL, Coot or EDS	RSCC < 0.8 suggest significant disagreement with electron density
Is the ligand in a region affected by crystal packing?	Inspect symmetry of model using PyMOL or Coot	Clashes of symmetry mates with ligand binding region suggest possible perturbations of conformation
Are the B-factors of the ligand significantly higher than the overall average?	Inspect B-factors of the ligand using PDB record or Twilight	B-factor >200 $Å^2$ corresponds to ~1.6 Å displacement
Is the resolution of the structure sufficiently high to support the model of the ligand?	Inspect reflection count and resolution in PDB record	Resolutions >2.5 Å likely have a lower data-to-parameter ratio
Is the K_d of the ligand binding high in the crystallization conditions employed?	Further biophysical or biochemical studies required	Ligand occupancy is likely less than 100 %

Questions that any user of protein–ligand models should ask in order to use the ligand model with confidence