J Comput Aided Mol Des (2012) 26:169–183
DOI 10.1007/s10822-011-9538-6

The good, the bad and the twisted: a survey of ligand geometry in protein crystal structures

John Liebeschuetz · Jana Hennemann · Tjelvar Olsson · Colin R. Groom

Received: 10 November 2011 / Accepted: 21 December 2011 / Published online: 14 January 2012

Abstract The protein databank now contains the structures of over 11,000 ligands bound to proteins. These structures are invaluable in applied areas such as structure-based drug design, but are also the substrate for understanding the energetics of intermolecular interactions with proteins. Despite their obvious importance, the careful analysis of ligands bound to protein structures lags behind the analysis of the protein structures themselves. We present an analysis of the geometry of ligands bound to proteins and highlight the role of small molecule crystal structures in enabling molecular modellers to critically evaluate a ligand model's quality and investigate protein-induced strain.

Keywords X-ray refinement · Structure validation · Ligand strain · CSD · PDB · Conformation

Abbreviations

CSD	Cambridge Structural Database
PDB	Protein Data Bank
RMSD	Root mean square deviation
FDA	Food and drug administration

Electronic supplementary material The online version of this article (doi:10.1007/s10822-011-9538-6) contains supplementary material, which is available to authorized users.

J. Liebeschuetz (✉) · T. Olsson · C. R. Groom
The Cambridge Crystallographic Data Centre (CCDC), 12 Union Road, Cambridge CB2 1EZ, UK
e-mail: john@ccdc.cam.ac.uk

J. Hennemann
Institut für Pharmazeutische Chemie, Philipps-Universität Marburg, Marbacher Weg 6, 35032 Marburg, Germany

Introduction

Structures of protein–ligand complexes determined from single crystal X-ray diffraction data provide the clearest snapshots we have of the world of protein structure and function. Such structures have provided invaluable information in the understanding of biochemical processes and have been used extensively in successful structure-based drug design campaigns [1].

We referred above to *structures*—but when we talk about such *structures* we are, of course, really discussing *models*. It is easy to lose sight of this distinction, but it is an important one, as correctly interpreting the electron density in and around the protein binding site is far from being a straightforward task.

There is considerable evidence that the quality of ligand models within published X-ray derived protein structures has, in the past, been rather poor [1–3]. Despite this problem being well recognised, structures continue to be published which have obviously incorrect ligand geometry. For instance, scientists were recently treated to the long awaited structure of the potential drug target IKKβ (inhibitor of κB kinase β) (Protein Data Bank (PDB) code 3qad, ligand identifier XNM) [4]. However as originally published, the aminopyrimidine ring of the bound inhibitor did not have the planar geometry one might expect, raising doubts over the interpretation of the electron density of this bound ligand (Fig. 1). Subsequently, after being alerted to this error, the authors were able to re-refine the structure and generate the correct aminopyrimidine geometry (PDB code 3rzf).

The most obvious and understandable reason for errors in ligand geometry is the limited resolution to which data can be collected for some systems. The lower the resolution, the less well defined electron density maps become

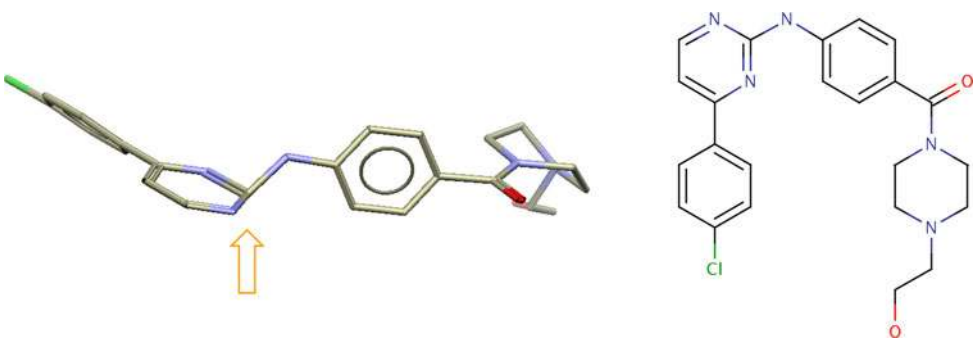

Fig. 1 Unusual aminopyrimidine geometry in the ligand in 3qad

and the more difficult it can be to model structures. However, another explanation is the failure of protein crystallographers to discard models with poor ligand geometry in favour of better ones.

There is a well known idiom in English that 'a bad workman blames his tools' and it is true that geometrical dictionaries for deriving bond, angle and torsional restraints for use in refinement do not exist for many substructures that are found within ligands. Protein crystallographers must therefore create their own tools; good geometries are unlikely to be obtained without good dictionaries. However, this brings another saying to mind—'It is not the tools we use which make us good, but rather how we employ them'.

It is worth at this point mentioning other factors that make the crystallographer's task more difficult. Most frequent is the partial occupancy of a binding site by a ligand. This problem presents itself with low affinity ligands, such as the small ligands favoured in fragment-based drug design, and those with low solubility. Alternatively a binding site might be fully occupied, but disorder occurs, the ligand taking up multiple different binding modes. In very high resolution structures it may be possible to resolve both binding modes. However this is not possible for low resolution structures and in such cases it can be difficult to position a ligand with a good fit to the electron density and plausible geometry.

Ligand structures with incorrect geometry have the potential to be highly misleading if used within a drug design context [5]. Moreover, such protein-ligand complexes may also find their way into test sets used for developing and evaluating molecular modelling software, for example protein-ligand docking packages [1, 6].

Errors in ligand refinement also confound analysis of the amount of strain that a protein can impart on a ligand in its bound state. It is undisputed that enzymes can strain substrates - this is, after all, the mechanism by which some operate, achieving an enzyme stabilised high energy state which can then facilitate the required chemical transformation. Synthetic enzyme inhibitors too can exhibit strain. Kuntz et al. [7] have observed that, as the size of ligands increases, affinity often reaches a plateau despite the larger number of ligand-protein interactions made. This might in part be due to the energetic compromises, in terms of internal strain energy, that a large ligand needs to make in order to still bind to a protein [8, 9]. It may also provide part of the explanation behind the observation that when small fragments of inhibitors are chemically linked, the potency gain achieved is seldom the theoretical maximum [10, 11].

So there is likely to be some justification to claims that ligands bind to proteins in conformations that are not their energetic minima [8], but how much strain is generally tolerated? A study by Perola and Charifson suggested that over 10% of ligands in a set of 150 abstracted from PDB models, had strain energies greater than 9 kcal mol^{-1} [12]. Hao et al. put an average value of 0.6 kcal mol^{-1} strain energy per torsional motif (i.e. per bond containing that motif) for nine different common torsional motifs again using structures drawn from the PDB [13] However, a more modest maximum strain energy of 3 kcal mol^{-1} per ligand has been suggested by others [14–16]. It also has been pointed out that current forcefields are unequal to the task of assessing relative energies for conformers of drug-size molecules [17]. The data of Perola and Charifson has recently been critically re-examined [15, 18], and it has been suggested that alternative low energy conformations can be found which fit the structural data as well or better for some of these high energy complexes. It seems perhaps, that it is not proteins which strain ligands, but protein crystallographers [1, 19, 20].

What methods exist therefore for the validation of ligand structures? First of all electron density maps can be used to visually examine how well a ligand fits the available data. In recent years deposition of structure factors has become more prevalent and, in a positive move, the world wide PDB stipulated in 2008 that structure factors must be deposited. So it is now in principle possible for any scientist to analyse ligand fit to electron density.

Use of prior knowledge of preferred geometry is another way to validate ligand geometry and the most comprehensive source of appropriate prior information is the

Cambridge Structural Database (CSD) [21] which contains structural information on over 500,000 small organic and metallo-organic molecules. The object of this study is to use the CSD to attempt an objective review of ligand structure quality in protein structures, and to establish whether it is possible to distinguish easily between those ligands which are likely to be correctly refined but are of strained geometry, and those ligands that simply exhibit incorrect geometry. Another object of this study is to see if the quality of ligand structures had increased over recent years. Use of a specific tool for validating the geometry of ligand models against experimental structures, Mogul [22], was to be a key component of this work.

Experimental section

Ligand-protein structures chosen for study

Three datasets were selected from the PDB, each containing approximately 100 ligand-protein structures. The first set was selected from structures published before the year 2000. A starting point was randomly chosen and then the next 100 PDB codes in alphabetic sequence, which corresponded to suitable ligand-protein complexes, were harvested. Those containing duplicate ligands and most of those containing co-factors were excluded. Also excluded were covalent ligands. Protein modifications (such as *N*-acetyl glucosamine), small co-solvent molecules and non-physiologically relevant co-crystal partners were not classed as ligands for this study. Lastly, if a series of structures from the same authors was encountered only the first two were used. The second and third sets of structures were selected in a similar manner from structures published in 2006 and 2009, respectively. These three lists of PDB codes are available as Supplementary Information.

CSD ligand structures chosen for study

A list of entries in the CSD for FDA approved drug molecules was searched to exclude structures with no common rotatable bonds and with $Z' > 1$ and number of chemical residues >1. Structures also had to have an R-factor ≤ 0.1, and not be disordered, or polymeric or present as an ion. The final list numbered 440 entries.

Evaluation tools and procedures

To evaluate structural geometry we used the program Mogul [22, 23], a module of the Cambridge Structural Database System (CSDS) (http://www.ccdc.cam.ac.uk). This is a knowledge-base of molecular geometry which, given an input structure, compares the bond lengths, bond angles and dihedral angles of that structure with parameter distributions derived from of similar substructures. Recently it has become possible to also analyse rings geometry in Mogul [33]. Using Mogul any geometric parameters which are clearly unusual can be located immediately. Mogul has also been used to help set up geometric constraints within the context of small molecule structure refinement in the CRYSTALS (http://cryst.chem.ox.ac.uk/mogul.html,) and DASH (http://www.ccdc.cam.ac.uk/support/documentation/dash/doc/portable_html/dash.1.138.html) packages. As most entries in the CSD are of atomic resolution they are much less reliant on interpretation and choice of restraints, and for the purposes of this analysis can be assumed to be correct.

We assume a premise that the statistical distributions derived from small molecule crystallographic information are generally applicable to protein-ligand complexes and other states. This needs some elaboration. The specific assumption is; given that a sufficiently large sample of relevant chemistry exists in the CSD, then the absence of experimental values close to the model value implies the model value represents an unusual geometry and suggests that significant internal or 'strain' energy would be incurred to achieve that geometry. It does not mean that the geometry is to all intents and purposes inaccessible under standard conditions as a zero incidence cannot be used to quantify the internal energy that would need to be incorporated. Allen et al have published comparisons of CSD distributions with quantum chemical calculations that provide evidence for the premise [24].

In this work ligands were first loaded into the CSDS module Mercury [25], hydrogen atoms were added where appropriate and bond and atom types assigned according to CSD conventions. The Mogul program was then called from within Mercury. Standard default settings were used. We examined bond lengths, bond angles, torsional dihedral angles and non-aromatic ring geometries, but were primarily concerned with identifying unusual torsional dihedral angles, for a number of reasons. First, the number of restraints needed to account for commonly found bond lengths and angles within ligands is far fewer than those required to account for commonly occurring torsional dihedrals, making it harder for refinement programs to correctly take account of all preferred torsions. Secondly, torsional dihedral angle distributions are frequently multimodal and have wide allowed ranges. Bond lengths and angles generally have much tighter tolerances and even if incorrect, will not usually lie too far from expected values. If one or more torsional dihedral angles are significantly wrong this will usually impact the overall molecular shape much more than having one or more deviant bond lengths or bond angles. Nevertheless, incorrect bond angles can be indicative of a poor ligand model, especially if the incorrect angle is near the centre of a large molecule.

Non-aromatic ring geometries can also take up unusual high strain conformations even if the individual bonds, bond angles and torsional dihedrals are accurate. A particular problem can occur when a ring with incorrect geometry in the initial model (e.g. a boat form cyclohexane) is not converted to the correct model (e.g chair form cyclohexane) during refinement, even if the data quality is good. So, non-aromatic ring geometries were also examined closely.

Mogul is able to identify all the unusual geometric parameters within a model, returning an assessment of how unusual a parameter is via an appropriate figure of merit. The torsional figure of merit in Mogul used in this work is termed the local density measure. This measures the ratio of incidences in the CSD within 10° of the torsional dihedral in question, to the number of total incidences of the torsional dihedral in the CSD. If this figure was less than 5% we consider the torsional dihedral to be 'unusual'. We describe a figure of 0% as a 'highly unusual' dihedral. It should be noted that a given torsion (say X1-X2) may be represented by several torsional dihedrals if X1 or X2 are multiply substituted. Numerical analyses are based on the number of incorrect torsional dihedrals. Figures of merit for bond lengths and angles are the numbers of standard deviations of the input query parameter from the mean value in the CSD (z-score). We classed an unusual bond-length or bond angle as one that is greater than two standard deviations from the mean (z-score $\geq$2). The figure of merit used for rings was the average angular RMSD of torsional angles in the ring, compared with similar average angular RMSDs over all the rings retrieved by the Mogul search. Ring conformations were classified as unusual when less than 5% of the average angular RMSDs values for the hits were within 10° of the average angular RMSD in the query.

Our ability to distinguish good from bad geometry is dependent on the number of examples available for the substructure under study. Where insufficient examples were available, no assessment was made. We decided that at least 5 examples were required for bonds and bond angles to be assessed and at least 15, for torsional dihedrals. Using these criteria less than 1% of all geometric features were insufficiently populated to allow an assessment. In most cases the number of hits available significantly exceeded the acceptance minima.

As the bond and angle criteria for 'unusualness' are based on a two standard deviation criterion, it could be argued there should be a natural error rate of 5% in an 'average' structure in the CSD. One could extend this reasoning to the torsional information as well, although perhaps less convincingly. To make conservative allowance for this we decided that an error rate of at least 10% for any single type of geometric feature (i.e. bond, angle or dihedral) was generally the minimum requirement to class a ligand as having unusual geometry. Exceptions were made for cases where a single geometric feature was very significantly non-standard; such ligands were also classed as having unusual geometry. Conversely, where several geometric features of a single type were borderline unusual, the ligand was not generally classified as having unusual geometry.

Other factors required evaluation before we could attempt to classify ligands with unusual geometry to either having been incorrectly refined, or to possibly being strained. The binding site of the protein in question was examined using the program Relibase+, [26] a tool designed for analysing protein-ligand interfaces. We noted (a) how well buried any substructure was within the protein, (b) how many stabilising interactions the ligand made with the protein and surrounding water networks, and (c) the contacts between protein and ligand or intramolecularly within the ligand. Whenever possible, we used the Electron Density Server at Uppsala University to visually check the fit of each ligand to the experimental electron density [27]. Whilst this could be done for many of the structures from the 2006 list and almost all of the structures from the 2009 list, only a few of the pre 2000 list had structure factors deposited.

With all these observations in hand, ligands were put into one of three classes; 'OK' (i.e. no unusual geometry that would alter the interpretation of the binding mode), 'Strained?' (i.e. unusual geometry but could be correctly refined) and 'Questionable' (i.e. unusual geometry likely due to poor refinement). Because this classification process is somewhat subjective, it was important to only categorise structures into the third category if the cumulative evidence was overwhelmingly strong that the structure was poorly refined. We collected additional information for the subset of structures submitted in 2006, viz: (a) the resolution of the structure, (b) the number of heavy atoms, (c) the number of non-terminal rotatable bonds, and (d) the proportion of 'unusual' geometric features found for each of the three geometric feature classes: bonds, bond angles and torsional dihedrals. A one-tailed, non-paired Student's T test was carried out for each pair of populations out of 'OK', 'Strained?' and 'Questionable' to identify whether they could be definitively separated according to criteria (a), (b) and (c).

Results and discussion

The result of the analysis process was to assign ligands into one of three classes; 'OK' (i.e. no unusual geometry that would alter the interpretation of the binding mode), 'Strained?' (i.e. unusual geometry but could be correctly refined) and 'Questionable' (i.e. unusual geometry likely due to poor refinement). It has already been stated that this

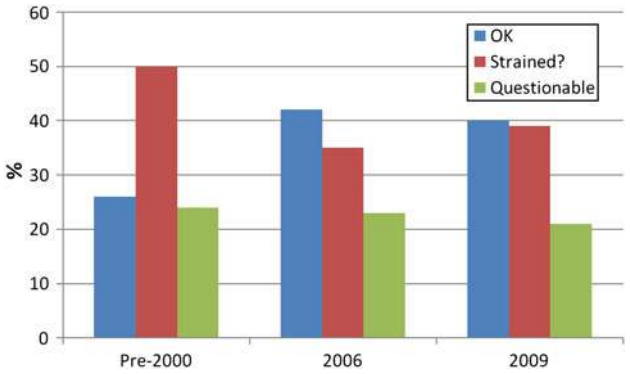

Fig. 2 Classification of ligands according to structural geometry for samples of protein/ligand structures submitted to the PDB at different times

classification process is somewhat subjective, and structures were only put into the third category if the cumulative evidence was overwhelmingly strong that the structure was poorly refined. A number of structures in the 'Strained?' category are likely to be incorrect, but we could not categorise them so with absolute certainty. In addition, we decided to class structures as 'OK' if serious deviations from normal geometry did not affect the overall interpretation of the binding mode, but we noted them and we will return to this set.

Simple statistics

The first notable result is that the number of structures deemed likely to be incorrectly refined, is a substantial proportion of all three subsets, at 20–25%, with a small, but statistically insignificant decline over time (Fig. 2). The second result of note is that the set of 'Strained?' structures submitted prior to 2000, is much larger than for the 2006 and 2009 subsets. This is entirely a consequence of the lack of electron density information available to us for most of these structures, which prevented a clearer classification being made. Had this information been available, it is likely that more of the pre-2000 ligand structures would have been classed as 'Questionable'. Consistent with this view is the observation that the number of structures having 'OK' geometry is much lower for the pre-2000 subset.

Ligand classification, structural complexity and geometric feature types

Table 1 gives means and standard deviations for a number of different ligand characteristics in the 2006 set divided into the three classifications of 'OK', 'Strained?' or 'Questionable'. A one-tailed, non-paired Student's T test was carried out for each pair of populations to identify whether they could be definitively separated according to the respective criteria. We also present histograms to reveal how classification depends on resolution and ligand complexity (Fig. 3). The T test indicates that no strong relationship exists between classification and the resolution of the structures. Figure 3a confirms this, although it is important to point out that ligands from extremely high resolution structures (<1.3 Å) usually have good ligand geometries, perhaps because the data is sufficient to refine to atomic resolution. Clear errors found in three slightly higher resolution structures (1.3–1.5 Å) are obvious and correctable mistakes that should have been picked up by inspection (i.e. very unusual bond lengths and angles for 2fgv and 2grb and a misinterpretation of electron density in 2cl2 that is discussed later).

Conversely, ligand complexity, as determined simply by number of heavy atoms or number of rotatable bonds, does depend on whether a ligand is classed as good or not. The 'OK' class of ligands is clearly separated from the 'Strained?' and the 'Questionable' classes with >95% confidence in both cases. It is not possible to separate the 'Strained?' and the 'Questionable' classes in this way. Ligands with less than 20 heavy atoms or with no rotatable bonds are usually refined without significant geometric error.

Table 1 Statistical data on each ligand class for attributes of the ligand structure and the resolution of the protein model; and *P* values for rejecting the null hypothesis, that the attribute does not distinguish between pairs of classes (significant comparisons in bold)

Category	n	Resol. (Å)	Heavy atoms	No. torsions	%Bonds 'unusual'	%Angles 'unusual'	%Dihedrals 'unusual'
Mean (SD)							
OK	43	2.1 (0.5)	22.1 (12	4.7 (5)	22.0 (24)	16.9 (14)	7.5 (17)
Strained?	38	2.04 (0.37)	34.5 (21)	8.9 (6)	20.4 (20)	19.7 (13)	22.7 (20)
Questionable	24	2.14 (0.42)	30.1 (18)	8.3 (7)	26.5 (19)	24.5 (14)	21.0 (17)
P value							
OK/Strained?		0.3975	**0.0006**	**0.0004**	0.3755	0.1749	**0.0003**
OK/Quest.		0.2692	**0.0149**	**0.0112**	0.2165	**0.0194**	**0.0016**
Strained?/Quest.		0.1577	0.2009	0.3337	0.1203	0.0852	0.3605

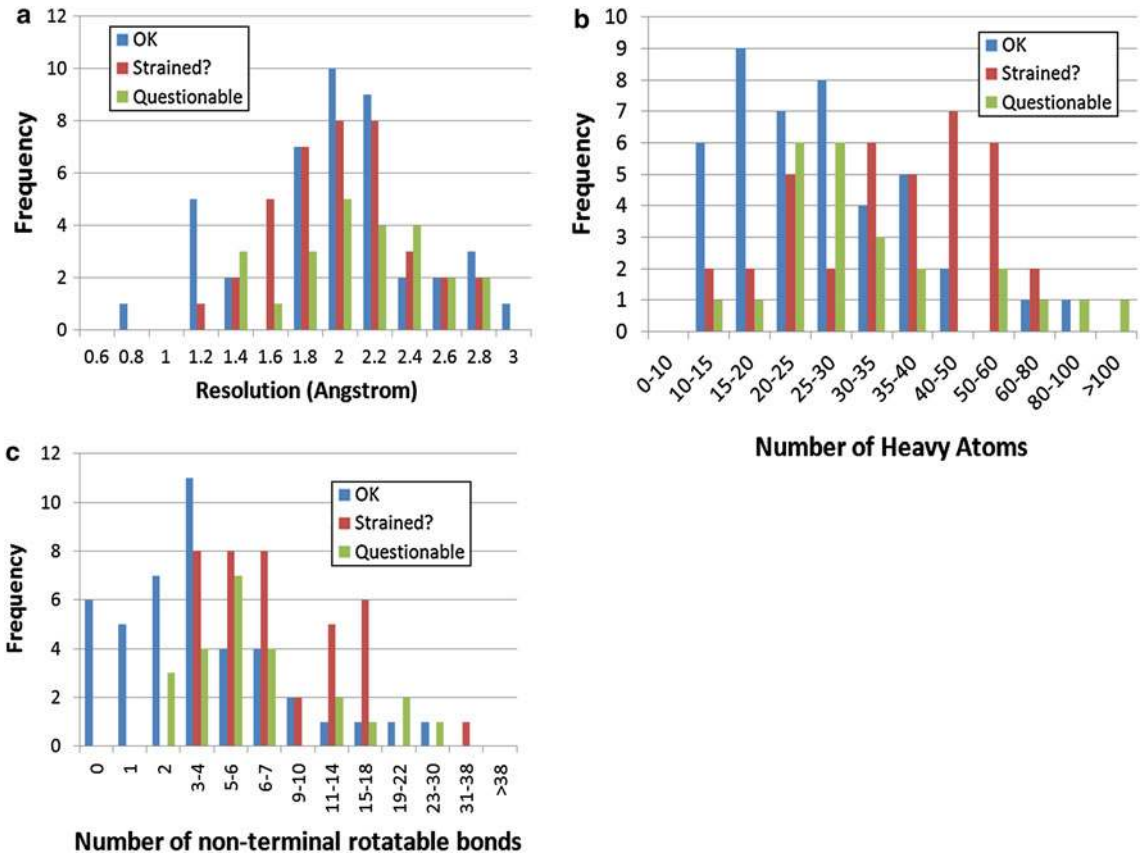

Fig. 3 Breakdown of each class **a** according to resolution of the protein structure; **b** according to the number of heavy atoms in the ligand, **c** according to the number of non-terminal rotatable bonds in the ligand

The histograms in Fig. 3b, c confirm this statistical analysis. The histograms also appear to show that there is a significant and counterintuitive difference between the 'Strained?' and the 'Questionable' sets in that a much greater proportion of the 'Strained?' set, are of high complexity. This is an artefact of the conservative application of the 'Questionable' classification in this work. Many of the larger ligands in the 'Strained?' set may in fact be non-optimally refined and therefore in principal assignable to the 'Questionable' category. However, because it is hard to say with certainty that this is so for individual cases, many of these structures were given the benefit of the doubt. There is an inference from this that the actual number of poorly modelled ligands is significantly more than the 25% conservatively claimed here.

For completeness we now look at how our ligand classification depends on geometric feature classification. None of the three classes can be separated from each other with reference to bond geometry only. By contrast, the torsion error rate can clearly be used to distinguish 'OK' from both 'Questionable' and from 'Strained?' structures. This is entirely expected as it was the major criterion used to identify the 'OK' structures in the first place. Of more interest is that a low bond angle error rate can also be used to identify 'OK' structures from 'Questionable' structures. It cannot be used to identify 'OK' structures from 'Strained?' structures.

Errors in incorrectly refined ligands

Bonds and angles

One clear observation is that many bond length and bond angle errors are found in structures that were classified as 'OK'. However, as already noted, substitution of the correct bond lengths or bond angles would not, on minimisation of the structure, lead to a different interpretation of the binding mode. Nevertheless they do represent true refinement errors and point to incorrect ligand dictionaries being employed in the refinement. Over 70% of the ligands in the 2006 set were found to have errors over bond length and bond angle features of greater than 10%. Of the 30 or so ligands which did not show such errors, four were peptidic ligands, and three were saccharides. It is perhaps unsurprising that good dictionaries do exist for these types of ligand.

In some of cases errors are extraordinarily large, sufficiently so as to put these models in the 'Questionable' category. For instance the pyrollidine carboxamide inhibitor of enoyl reductase (PDB 2h7 m, ligand ID 641) has a

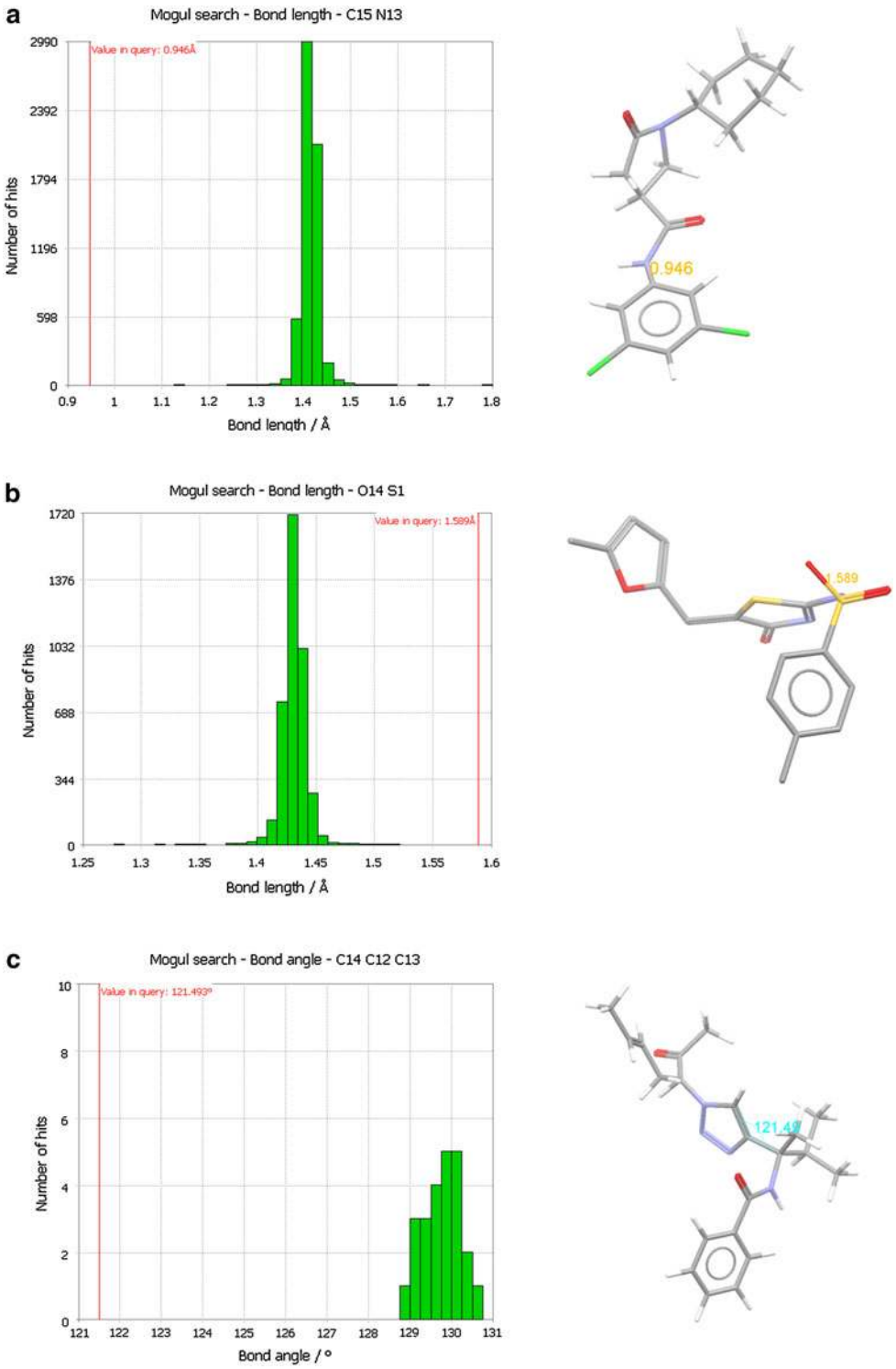

Fig. 4 Severe deviations from normal bond and angle geometry: **a** extremely short bond in 2h7 m ligand; **b** extremely long S=O bonds in 2hwh ligand; **c** Very tight C=C–C bond angle in 2hxz ligand

C–N bond length 0.946 Å, 23 standard deviations from the bond length in similar structures (Fig. 4a). Sulphur-oxygen and phosphorous-oxygen bonds appear to be particularly poorly treated, e.g. in 2hwh (ligand ID RNA), the sulphonamide S=O bonds are elongated, such that they lie 14 standard deviations from normal geometry (Fig. 4b). Although in most cases bond and angle errors do not change the interpretation of the ligand binding, this is not always true. A highly abnormal bond angle (C12–C13–N) associated with the 1,2,3-triazole in the ligand (H7J, chain A) in 2hxz is 14 standard deviations from ideal (Fig. 4c). The chain B and C ligand models in 2hxz have more reasonable C12–C13–N bond angles though they are still 3 SD from ideal. Large errors in bond angle assignment in a

central part of the molecule, as here, will often lead to difficulty in finding reasonable placements for the peripheral parts, because of leverage.

Torsion angles

Highly unusual bonds and bond angles can easily be identified by reference to the distributions in the CSD. The distributions of torsion angles are often more complex and our definition of 'unusual' needs a little more thought. However, there are many simple cases where a dihedral angle distribution may have a small number of very sharp maxima, implying that the torsional energy barrier is too high to allow non-standard values. Perhaps the best example is that of amide bonds, where clear preferences exist in the CSD. Despite this we often see unusual amide torsions in ligands bound to proteins. A striking example is the glycosidic ligand in endo-1,3-beta glucanase (2cl2, ligand ID BMA). This structure is of very high resolution, 1.35 Å, and visual inspection of electron density maps would suggest that is has been well refined, with the large, flexible ligand fitting well to the electron density (Fig. 5). However two amide bonds in the ligand are set as *cis* rather than *trans* for no obvious reason. This results in a short internal $CH_3\cdots OH$ contact of 2.86 Å for one of them. The alternative *trans* orientations of the amides are almost isoelectronic, would fit the electron density equally well, and would replace this unlikely short contact with a good C=O···OH hydrogen bond.

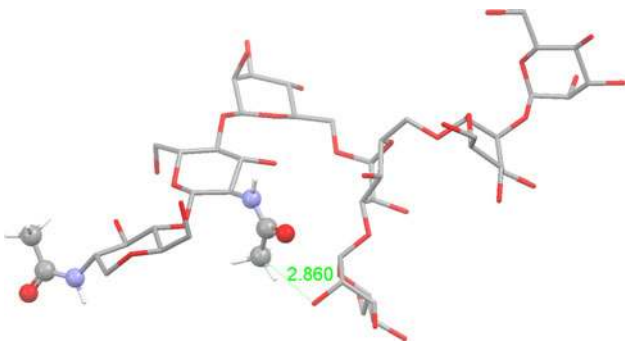

Fig. 5 Incorrect *cis* amide bond geometries (*highlighted*) in an otherwise high quality ligand structure (2cl2)

Amide bonds and similar torsions aside, many propensity histograms of torsional dihedrals show broad and multimodal maxima, implying a relatively low energy difference between rotamers. Observations that a small proportion of dihedrals for a given ligand, are lying in the less populated regions of the distribution might well be an indication of conformational strain. However for ligands where a large proportion of dihedrals are classified as unusual it is a strong indication that refinement error exists. Figure 6a shows the breakdown of each class of ligands according to percentage of torsional errors. The proportion of unusual torsional dihedrals in 'Questionable' structures is usually in excess of 20%. The proportion of unusual dihedrals in the 'Strained' set is also high, suggesting once more that a number of these do in fact represent poor models. Figure 6b, by contrast shows the breakdown of unusual torsional dihedrals in a set of 440 FDA approved molecules (i.e druglike molecules) for which crystal structures exist in the CSD. 90% of the structures show no unusual dihedrals at all and the overall distribution is similar to that of the 'OK' set in Fig. 6a.

Ring systems

All ligands containing rings of unusual geometry automatically were classified as 'Strained?' or 'Questionable', irrespective of other geometric features. Of the thirteen ligands in the 'Strained?' set that contained non-aromatic rings, six showed ring geometry errors. There are also thirteen such ligands in the 'Questionable' class, of which five have unusual ring geometry.

Highly unusual ring geometries, like torsions, are not necessarily indicative of an error, however further

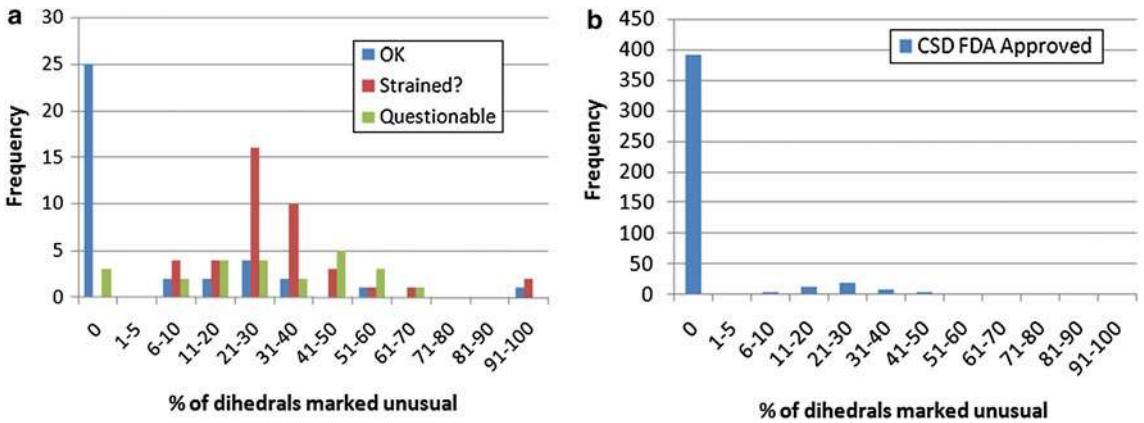

Fig. 6 **a** Breakdown of each ligand class according to percentage of unusual dihedrals. **b** Breakdown for a set of 440 CSD structures of FDA approved molecules

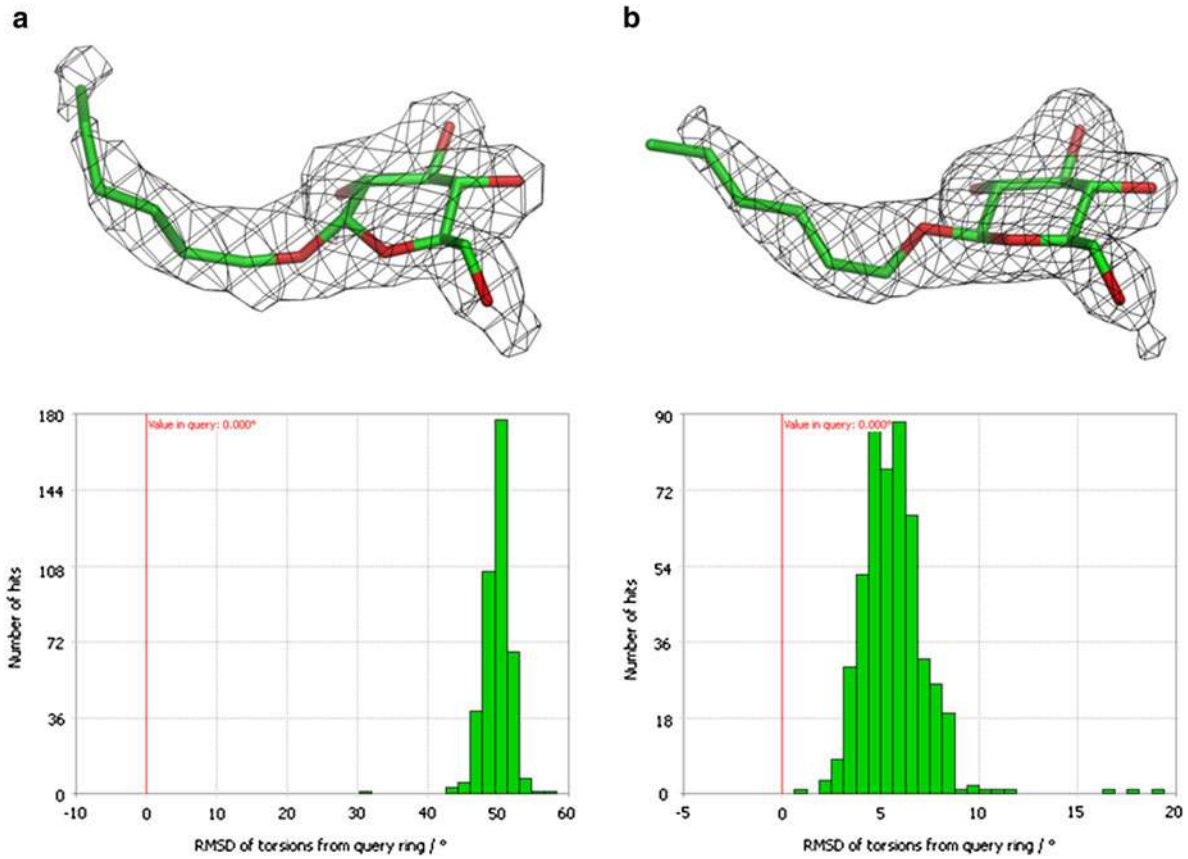

Fig. 7 Although the ligand in 2evs (2.2 Å Resolution) apparently fits the e.d. the ring geometry is poor (**a**). Re-refinement leads to similar quality structure with much improved ring geometry (**b**)

examination may reveal that an error exists. The sugar of hexyl-beta-D-glucoside bound to human glycolipid transfer protein 2evs (ligand ID GLC) (2.2 Å resolution), differs significantly from normal geometry (Fig. 7a, left). As this ring also has several implausible bond angles and there are other unusual torsions within this ligand it is clearly in error. However, this is not evident from assessing the fit of this ligand to electron density maps alone. This example is a particularly good illustration of the value of tools for assessing ligand geometry, especially in cases where the structure is only determined at modest resolution. This structure was found to be originally generated using the incorrect α-anomer. and has since been re-refined using the correct β-anomer, with a more plausible geometry (Fig. 7b, right) [33].

Treatment of disorder and multiple occupancy

It is common to see only a partial occupancy of a binding site by a ligand in a protein crystal structure [1, 3]. This may reflect disorder or even two equally low energy binding modes for the ligand [9]. Difficulties in appropriately modelling the complex electron density, which may represent a mixture of solvent and ligand molecules, potentially in multiple conformations, do make it easy to allow geometrical errors to creep in.

We identified many implausible geometric features in regions of ligands that did not produce easily interpretable electron density. Most often errors are seen in pendant groups attached to the rest of the ligand by a single bond. A good example is the thrombin inhibitor ligand in 1ae8 (ligand ID AZL) (resolution 2.00 Å) (Fig. 8a). The N9-C23-O24-C24 carbamate dihedral is well represented in the CSD and the distribution is sharp, implying distinct preferences for this system. This ligand is, however, modelled with an angle not observed in the small molecule structures in the CSD. There is no observable electron density for this group, which does not contact the protein and it is difficult to provide scientific justification for such unusual geometry. Clearly, any arguments relating to 'proteins straining ligands', are not appropriate here.

Where the resolution of crystal structure is reasonably high it is possible to identify multiple binding modes for a ligand to a protein. Our selected structures contain seven such cases (PDB codes 2hs1, 2i0h, 2b60, 2f3k, 2ij7, 3dtx, 3fwa), all in the 2006 and 2009 sample sets. If we make the assumption that this proportion is representative of structures in the entire PDB then at least 2% of all protein-

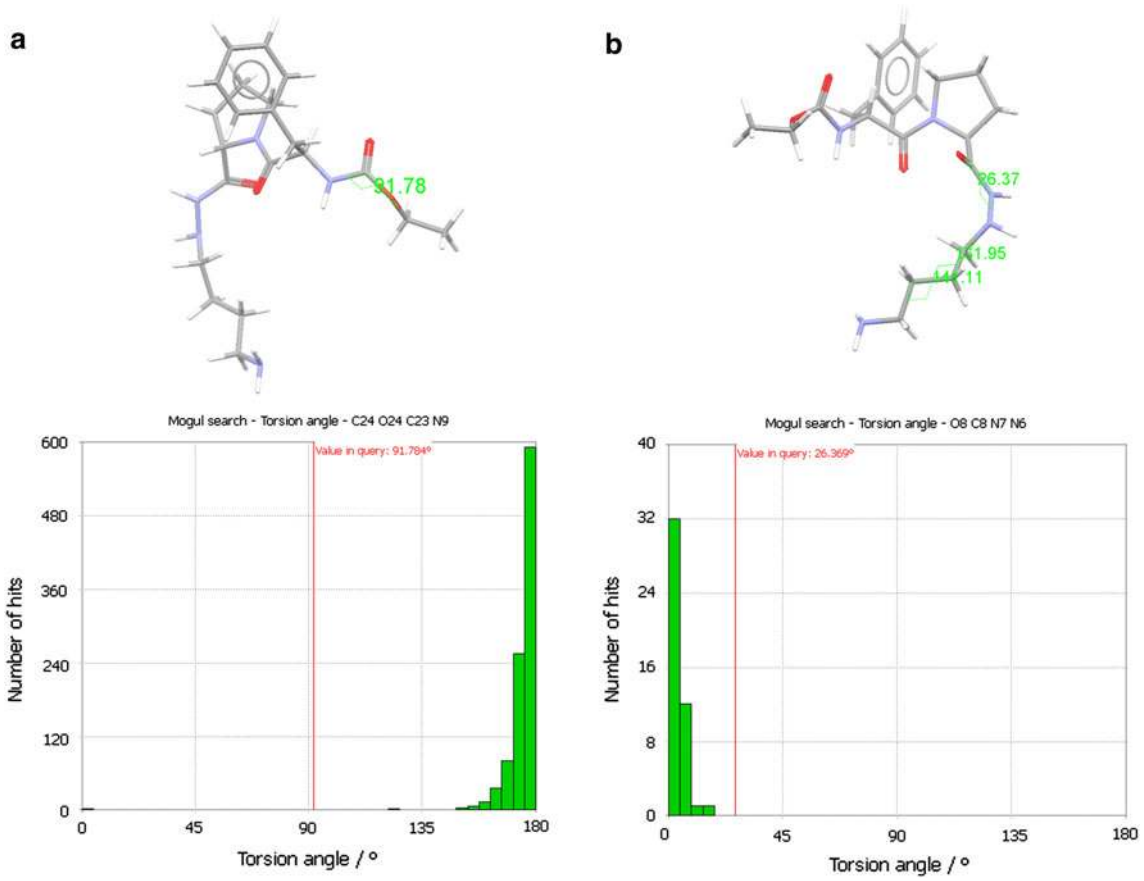

Fig. 8 Thrombin inhibitor structure 1ae8 **a** analysis of the carbamate geometry, **b** analysis of O=C–N–N (O8 C8 N7 N6) dihedral. This and two other dihedral angles of unusual geometry are *highlighted* on the 1ae8 ligand structure

ligand complexes might be expected to show multiple binding modes. In fact the figure is likely to be larger than this because we could reasonably argue that multiple-occupancy has been missed in the structures with low resolution, and may well not have been picked up in structures submitted pre-2000.

Even when electron density is interpreted as representing ligands binding in multiple modes it remains difficult to model ligands. The structure 2i0 h (ligand ID 222) (resolution 2.0 Å) illustrates this well (Fig. 9a). Although two slightly different binding modes have been modelled there are many implausible bond angles and torsional dihedrals; neither of the ligand models is likely to be correct. One might speculate that an alternative ligand binding mode might fit the electron density equally well, without the need to model a ligand with unusual geometry and to postulate two distinct binding possibilities.

The structure of the berberine bridge enzyme, 3fwa (ligand ID REN) (resolution 1.5 Å) is an interesting case for which the resolution does appear to be high enough to resolve multiple alternative binding modes of the ligand reticuline (Fig. 9b). We classify this ligand as 'Strained?' because the aromatic OMe torsion is unusual in both placements. Such systems are invariably planar in structures of small molecules yet are often found out-of-plane in PDB ligand structures. We concur with Brameld et al. [28] that most of these occurrences are electron density fitting errors, but in this case a closer examination is warranted. The methyl group fits well to a defined pocket of the protein and it is difficult to postulate an alternative conformation. This may well be a true example of a ligand that can only bind to this target in a conformation that is strained; although the possibility that alternative binding modes could be fitted to the electron density cannot be ruled out.

Can we identify torsional strain?

So it is becoming evident that separating 'strain' from 'error' is a non-trivial task. For all but the highest resolution structures, there is a significant probability that one or more unusual torsion angles or other features could be mutually tweaked to more plausible values by careful re-refinement. Let's return to the peptidic thrombin inhibitor,

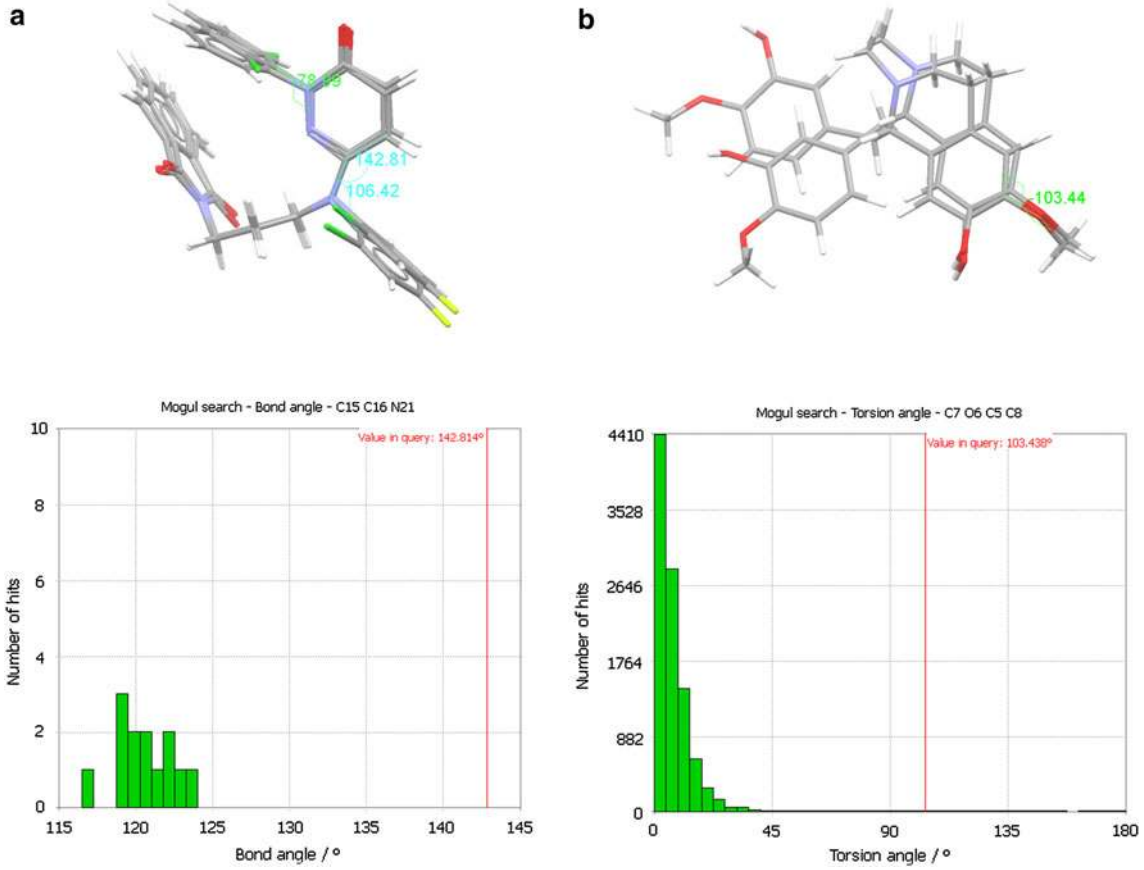

Fig. 9 Cases where there is double occupancy: **a** The CSD distribution of the C-Car-Nar bond angle (C15 C16 N21) in the 2i0 h ligand modelled for double occupancy. This, at 142.8°, is extremely distorted from normal geometry. Other unusual geometric features are labelled in the *left-hand image*; **b** the CSD distribution of the CC-OMe (C7 O6 C5 C8) dihedral *highlighted* on the ligand from 3fwa. This, at 103.44°, is significantly distorted from normal in both ligand placements

1ae8 (ligand ID AZL) (2.00 Å). This has an O8–C8–N7–N6 hydrazide dihedral angle value which is not represented by the small molecule crystal structures CSD (Fig. 8b). The hydrazide torsion is central to the ligand and appears within a well defined area of electron density. It might be considered that there is no need to further improve the geometry of this part of the molecule, however altering this angle by a mere 12° places it within a well populated region of the CSD dihedral distribution. There seems no reason why corrected dihedral would not be well accommodated by the electron density and it would also allow the pendant alkyl amino chain, which also has two unusual dihedral angles, to be modelled rather differently (Fig. 8b). This structure has in fact been recently re-refined to generate a ligand structure with fewer geometry errors (Personal communication: Dr Oliver Smart.)

This leads to the conclusion that we should consider only the highest resolution structures if we wish to clearly identify strain. Tellingly, almost all of the very highest quality structures of our 2006 test set were classified as 'OK'. Six out of seven have no unusual dihedral angles, indicating that no significant torsional strain is present in these ligands. Let's look at some of these structures in more detail. The structure of a mutant of HIV protease, with the bound ligand darunavir, 2hs1 (ligand ID O17) has the highest resolution (0.84 Å) of all structures examined [29]. The ligand is a peptomimetic with 13 rotatable bonds, so has ample chance to exhibit torsional strain (Fig. 10a). Two binding modes with opposite orientations are identified in the symmetrical binding site, with a 3:2 preference for one orientation. It has already been noted that multiple binding modes creates problems for good ligand model building. Despite this it is notable that whilst 9% of bonds and 14% of bond angles are marked unusual (a failure of the ligand dictionaries used perhaps?) *not a single dihedral angle* in either structure is so identified. Moreover the geometry found in the models often corresponds to a maximum in the dihedral frequency histogram. It is to the credit of the authors of this work that it has been possible to create models for two binding modes of this highly flexible ligand, neither of which are significantly strained.

There do exist dihedrals in 2hs1 which have non-optimal values. Where sufficient data in the CSD exists (>700 hits) it is possible to make a crude estimate of strain for a given

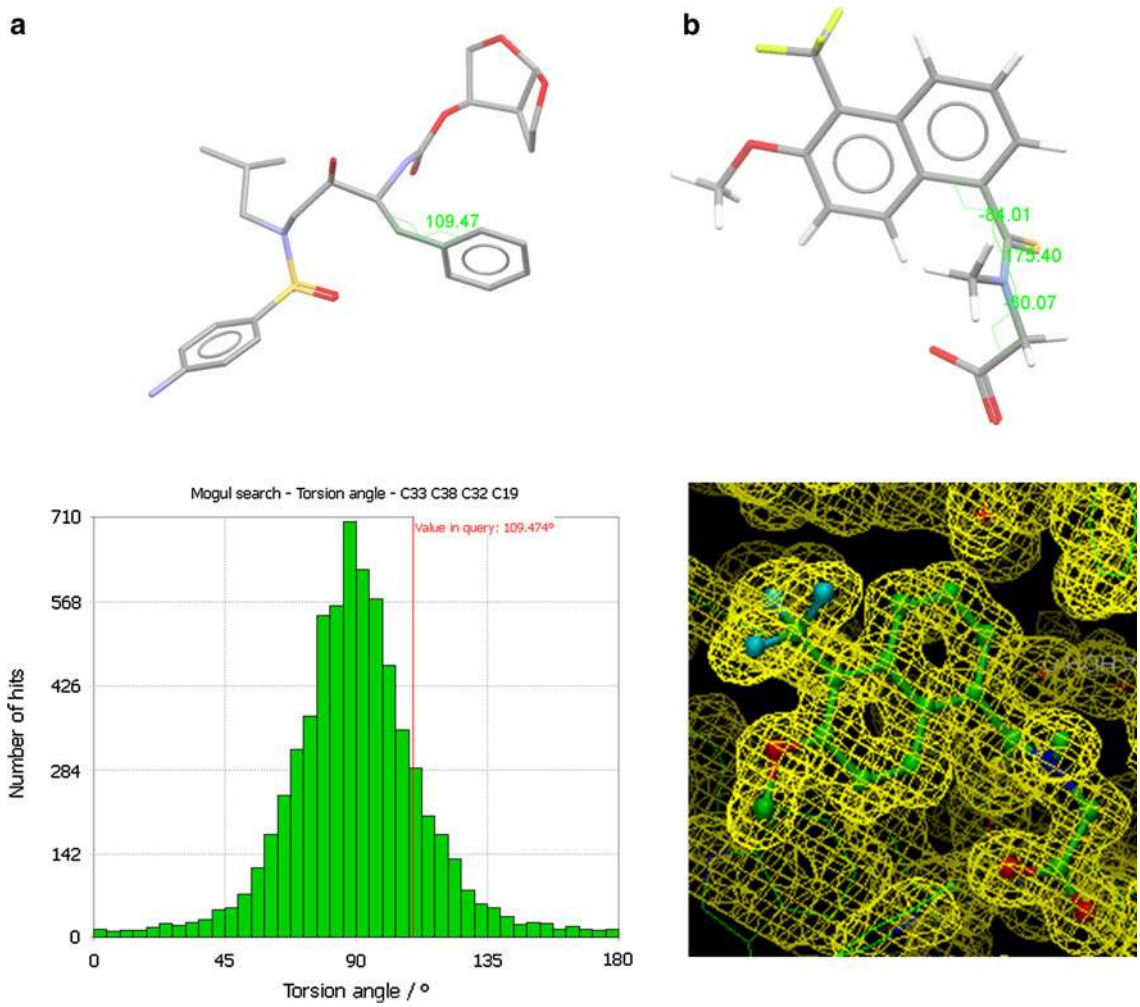

Fig. 10 High resolution structures: **a** The CSD distribution for a typical non-optimum dihedral (C33 C38 C32 C19) in the ligand from the ultra-high resolution structure 2hs1; **b** the electron density fit and molecular structure for tolrestat in 1zua. The three dihedrals *highlighted* have

torsion, from the dihedral frequency histograms by taking the ratio of histogram bar heights at the maximum (h_{max}) and at the query values (h_{query}) and then using these in the Gibbs free energy equation, $\Delta G = -RTln(h_{max}/h_{query})$. The dihedral which shows the most pronounced deviation from standard geometry is C33–C32–C38–C19 (Fig. 10a). The estimate of strain energy in this torsion comes to about 0.5 kcal mol^{-1}; an energy penalty to binding that would reduce the affinity of the compound by about twofold. If this is an estimate of the torsional strain for one of the most strained torsions in this ligand then it follows that the strain per bond over the whole ligand is considerably lower.

The structure of the human aldo-keto reductase, 1zua (1.25 Å), with tolrestate (ligand ID TOL) bound, is another high resolution structure with good ligand geometry (Fig. 10b) [30]. The thioamide chemistry is unusual and is represented hardly at all in the CSD. However, the small molecule crystal structure of a close analogue of tolrestate is available (Refcode CAKWAM). The match of the values of the dihedrals for the three torsions of thioamide is almost exact between the PDB and CSD models (maximum deviation over all possible dihedrals is 6.3°, minimum deviation is 0.6°, Fig. 10b). The authors do not state that they referred to this structure in setting appropriate refinement restraints or building an appropriate starting model for tolrestate, presumably because the data quality was good enough that this was not necessary. The agreement in thioamide geometries is therefore remarkable. Clearly there is negligible protein-induced torsional strain in this ligand.

Finally, we'll look at two examples of ligands bound to proteins where the deviations from standard geometry *are* likely to be because of conformational strain. The ligand) in the aminotransferase structure 1ajs (ligand ID PLA, is a

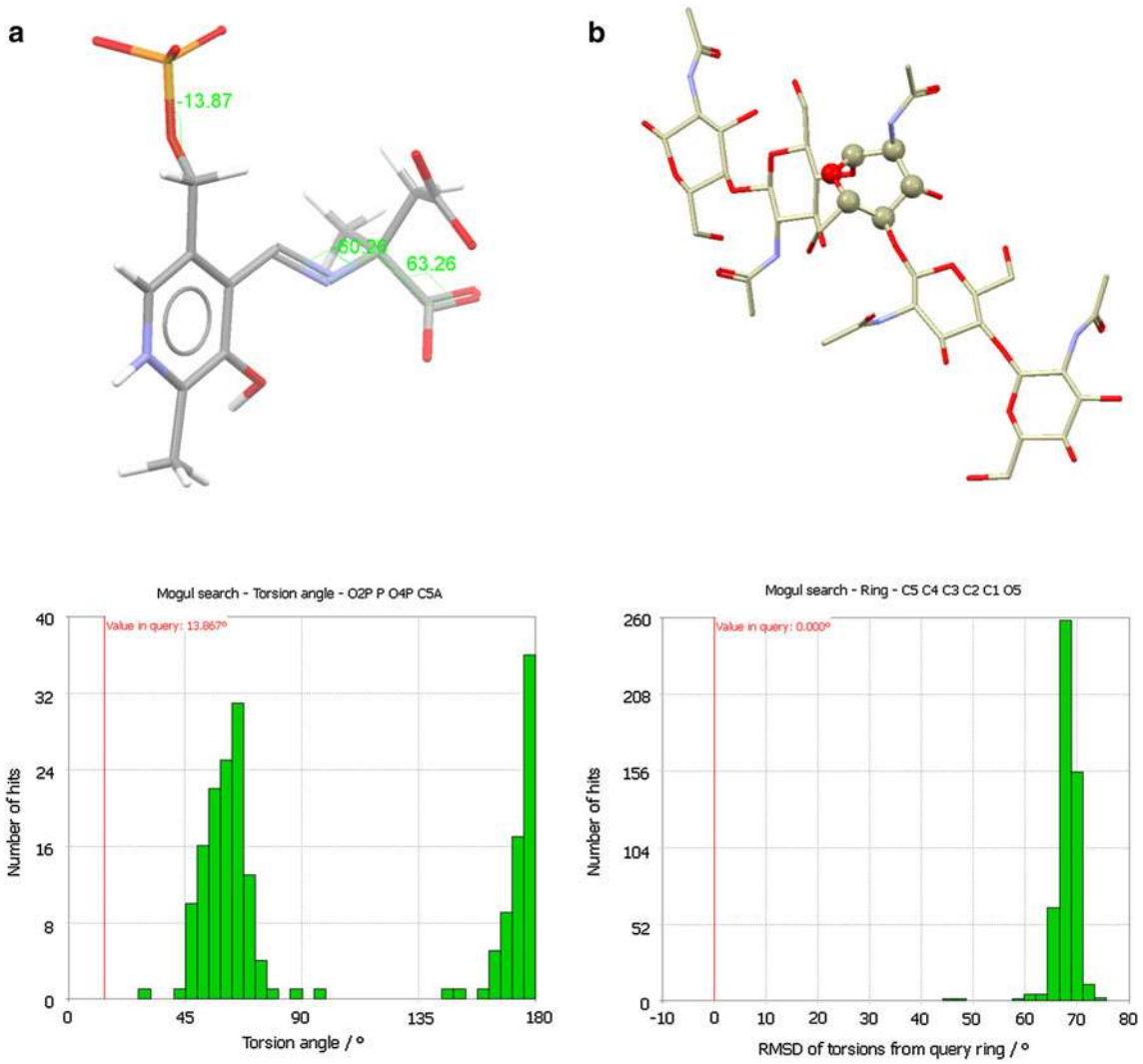

Fig. 11 Structures of substrates or analogues: **a** The NAD-methylaspartate conjugate from 1ajs. Three dihedrals are highly strained (CSD distribution is shown only for the C–O–P–O dihedral); **b** Penta-NAG bound in chitinase. The central unit takes up an unusual boat form

covalent adduct of the cofactor NAD and 2-methylaspartate (determined to 1.6 Å resolution) (Fig. 11a). The adduct has no fewer than 9 unusual dihedrals according to our analysis, representing three torsions in total; we therefore classified it as 'Strained?'. The unusual dihedrals are highlighted in Fig. 11a alongside the CSD distribution for one of the well represented dihedral angles. Visual inspection of electron density maps shows a good fit of the ligand.

In this particular case the six polar groups in the ligand all make numerous hydrogen bonding interactions with residues in the active site. No unfavourable contacts are made. The interactions with the aspartate carboxylates and the phosphate of the ligand are likely to contribute significantly to the enthalpy of binding. Clearly there appears to be torsional strain in this structure but the mechanism by which this can be offset is also evident. The authors of this structure point out that structure of the complex differs from unliganded structure, with a surface loop of hydrophobic groups on the N-terminal helix becoming buried. This entropy-driven process and the interactions between ligand and protein therefore provide sufficient free energy to compensate for the unfavourable conformation of the ligand and drive the formation of the cofactor—methylaspartate aldimine complex [31].

Our second example is a penta-*N*-acetyl glucosamine bound within a thermostable chitinase from pyrococcus furiosus (3a4w (ligand ID NAG) (1.8 Å). This structure was not one of our three hundred randomly selected complexes, but is included as it illustrates a number of points very well. The central six-membered ring of the ligand is in boat form and our Mogul[22] generated histogram clearly indicates that similar geometry is not found within the CSD (Fig. 11b). The fit of the ligand to the electron

density is good. It seems that in this case, the four *N*-acetylglucosamine units all make favourable steric interactions and hydrogen bonds with the protein and that the free energy this provides brings about the chair-to-boat conversion in the central ring.

The ever present possibility of non-optimal refinement means that a rigorous analysis of strain energy in bound ligands is challenging. Careful analysis is required to reveal how much ligand strain energy can be compensated for by interactions with a protein. Nevertheless, our analysis suggests that the average torsional strain in ligands that are inhibitors rather than substrates (or substrate mimics) is probably much less than previous estimates of 0.6 kcal $\cdot$ mol^{-1} per torsion [13]. Substantial strain is only likely to be seen in structures where an enzymatic intermediate has been captured within the protein binding site.

Conclusions

This analysis set out with a primary aim of assessing model quality for ligand structures in protein crystal structures; and a secondary aim of investigating the amount of torsional strain energies in protein bound ligands that can be compensated for by the other components of free energy of binding.

We found that the geometries of a large proportion of a sample of 300 recently determined protein-bound ligands are not consistent with geometries seen within small molecule crystal structures and that some of this inconsistency is due to refinement error.

An estimated 70% of the recently determined structures of protein bound ligands have bond length and bond angle errors that could be removed by use of better restraint dictionaries. At least 25% of recently determined structures, it is estimated, have geometric errors that could have been caught during crystallographic refinement and are large enough to potentially lead to a misleading interpretation of the binding interactions.

Error rates are not found to correlate straightforwardly with the resolution of the structure. However the highest resolution structures (resolution <1.3 Å) do not usually have unusual ligand geometries. One might have expected ligands modelled in lower resolution structures to have near ideal geometry in regions not contacting the protein; the less data there are available, the less justification there is for postulating unusual ligand geometry. As this is not the case it is clear that appropriate restraints are not being applied during the refinement of some ligands.

Although ligands in structures submitted prior to 2000 appear to be the most error prone, we found little evidence to suggest that quality of ligand geometry has improved since 2006. Whilst this may be so we point out the Protein Databank organisations well recognise the issues reported here and are currently developing tools that will allow users to check their ligand structures against CSD standard geometries prior to submission [34].

In addition a bulk re-refinement of PDB structures has been carried out in an effort to correct many refinement errors by Joosten et al. [32]. A number of the unusual geometries explicitly mentioned above, which appear correctable via rather small atomic position adjustments (e.g. 2h7m, 1ae8, 2i0h) are in fact significantly improved in the PDB_REDO database. Cases where there is an intrinsic error in the ligand model which requires a significant and concerted atomic movement to correct (as in 2cl2 and 2evs) are not improved.

Frequent incidence of errors in ligand geometry mean that efforts to evaluate the ligand strain that can be accommodated when binding to a protein are fraught with difficulty. We recommend that only the very highest quality protein-ligand structures be used for such studies. Where this criterion is met, we see evidence that, for a ligand that is a highly potent and is not a substrate or substrate derivative, ligand strain energy is considerably lower than some commentators have previously suggested [12, 13]. More work is needed however.

This study highlights the need for greater attention to be directed towards ensuring good ligand geometry during electron density interpretation and structure refinement. The message for the molecular modeller is that careful validation of ligand geometry and fit to electron density is an essential preparation for any modelling work based on a protein-ligand structure. Tools such as the Electron Density Server [27] and Mogul [25] are available to make this a relatively simple task.

Acknowledgments We thank Matthew Lightfoot (CCDC) for generating the original lists of PDB structures. We would like thank Oliver Smart, Thomas Womack and Andrew Sharff at Global Phasing Ltd for re-refining the 2evs structure. We thank Thomas Womack and Chun-Wa Chung (GSK) for contributing useful examples. We thank Frank Allen (CCDC) and Oliver Smart for helpful comments during the writing of the manuscript.

References

1. Davis AM, Teague SJ, Kleywegt GJ (2003) Applications and limitations of X-ray crystallographic data in structure-based ligand and drug design. Angew Chem Int Ed 42:2718–2736
2. Nissink JWM, Murray C, Hartshorn M, Verdonk VL, Cole JC, Taylor R (2002) A new test set for validating predictions of protein-ligand interaction. Proteins 49:457–471

3. Kleywegt GJ (2007) Crystallographic refinement of ligand complexes. Acta Crystallogr D 63:94–100
4. Xu G, Lo Y-C, Li Q, Napolitano G, Wu X, Jiang X, Dreano M, Karin M, Wu H (2011) Crystal structure of inhibitor of κB kinase β. Nature 472:325–330
5. Davis AM, St-Galley SA, Kleywegt GJ (2008) Limitations and lessons in the use of X-ray structural information in drug design. Drug Discovery Today 13:831–841
6. Hartshorn MJ, Verdonk ML, Chessari G, Brewerton SC, Mooij WTM, Mortenson PN, Murray CW (2007) Diverse high-quality test set for the validation of protein-ligand docking performance. J Med Chem 50:726–741
7. Kuntz ID, Chen Z, Sharp KA, Kollman PA (1999) The Maximum affinity of ligands. Proc Natl Acad Sci USA 96:9997–10002
8. Sharp KA (2005) Important considerations impacting molecular docking. In: Shoichet BK, Alvarez J (eds) Virtual screening in drug discovery. CRC Press, Boca Raton
9. Mobley DL, Dill KA (2009) Binding of small-molecule ligands to proteins: "What you see" is not always "What You Get". Structure 17:489–497
10. Chung S, Parker JB, Bianchet M, Amzel LM, Stivers ST (2009) Impact of linker strain and flexibility in the design of a fragment-based inhibitor. Nat Chem Biol 5(6):407–413
11. Borsi V, Calderone V, Fragai M, Luchinat C, Sarti N (2010) Entropic contribution to the linking coefficient in fragment based drug design: a case study. J Med Chem 53:4285–4289
12. Perola E, Charifson PS (2004) Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganisation upon binding. J Med Chem 47:2499–2510
13. Hao MH, Haq O, Muegge I (2007) Torsion angle preference and energetic of small-molecule ligands bound to proteins. J Chem Inf Model 47:2242–2252
14. Vieth M, Hirst JD, Brooks CL (1998) Do active site conformations of small ligand structures correspond to low free-energy solution structures? J Comput-Aided Mol Des 12:563–572
15. Butler KT, Luque FJ, Barril X (2008) Toward accurate relative energy predictions of the bioactive conformations of drugs. J Comput Chem 30(4):601–610
16. Boström J, Norrby P-O, Liljefors T (1998) Conformational energy penalties of protein-bound ligands. J Comput-Aided Mol Des 12:383–396
17. Tirado-Rives J, Jorgensen WL (2006) Contribution of conformer focusing to the uncertainty in predicting free energies for protein-ligand binding. J Med Chem 49:5880–5884
18. Nicholls A, McGaughey GB, Sheridan RP, Good AC, Warren G, Mathieu M, Muchmore SW, Brown SP, Grant JA, Haigh JA, Nevins N, Jain AN, Kelley B (2010) Molecular shape and medicinal chemistry: a perspective. J Med Chem 53:3862–3886
19. Kleywegt GJ (2009) On vital aid: the why, what and how of validation. Acta Crystallogr D 65:134–139
20. Fu Z, Li X, Merz KM (2011) Accurate assessment of the strain energy in a protein bound drug using QM/MM X-ray refinement and converged quantum chemistry. J Comput Chem 32:2587–2597
21. Allen FH (2002) The Cambridge structural database: a quarter of a million structures and rising. Acta Crystallogr B 58:380–388
22. Bruno IJ, Cole JC, Kessler M, Luo J, Motherwell WDS, Purkis LH, Smith BR, Taylor R, Cooper RI, Harris SE, Orpen AG (2004) Retrieval of crystallographically-derived molecular geometry information. J Chem Inf Comput Sci 44(6):2133–2144
23. http://www.ccdc.cam.ac.uk/products/csd_system/mogul/
24. Allen FH, Harris SE, Taylor R (1996) Comparison of conformer distributions in the crystalline state with conformational energies calculated by ab initio techniques. J Comput-Aided Mol Des 10:247–254
25. Macrae CF, Bruno IJ, Chisholm JA, Edgington PR, McCabe P, Pidcock E, Rodriguez-Monge L, Taylor R, van de Streek J, Wood PA (2008) Mercury CSD 2.0—new features for the visualization and investigation of crystal structures. J Appl Cryst 41:466–470
26. Hendlich M, Bergner A, Gunther J, Klebe G (2003) Relibase—design and development of a database from comprehensive analysis of protein-ligand interactions. J Mol Biol 326:607–620
27. Kleywegt GJ, Harris MR, Zou JY, Taylor TC, Wählby A, Jones TA (2004) The Uppsala electron-density server. Acta Crystallogr D 60:2240–2249
28. Brameld KA, Kuhn B, Reuter DC, Stahl M (2008) Small molecule conformational preferences derived from crystal structure data. A medicinal chemistry focused analysis. J Chem Inf Mod 48(1):1–24
29. Kovalevsk AY, Liu F, Leshchenko S, Ghosh AK, Louis JM, Harrison RW, Weber IT (2006) Ultra-high resolution crystal structure of HIV-1 protease mutant reveals two binding sites for clinical inhibitor TMC114. J Mol Biol 363:161–173
30. Gallego O, Ruiz FX, Arde A, Domı′nguez M, Alvarez R, de Lera AR, Rovira C, Farres J, Fita I, Pares X (2007) Structural basis for the high all-trans-retinaldehyde reductase activity of the tumor marker AKR1B10. Proc Natl Acad Sci USA 104:20764–20769
31. Rhee S, Silva MM, Hyde CC, Rogers PH, Metzler CM, Metzler DE, Arnone A (1997) Refinement and comparisons of the crystal structures of pig cytosolic aspartate aminotransferase and its complex with 2-methyltransferase. J Biol Chem 272:17293–17302
32. Joosten RP, Salzemann J, Bloch V, Stockinger H, Berglund A-C, Blanchet C, Bongcam-Rudloff E, Combet C, Da Costa AL, Deleage G, Diarena M, Fabbretti R, Fettahi G, Flegel V, Gisel A, Kasam V, Kervinen T, Korpelainen E, Mattila K, Pagni M, Reichstadt M, Breton V, Tickle IJ, Vriend G (2009) PDB_REDO: automated re-refinement of X-ray structure models in the PDB. J Appl Cryst 42:376–384
33. Cottrell S, Olsson T, Bowden S, Taylor R, Korb O, Cole J, Liebeschuetz JW. Understanding and modelling ring conformations using small molecule crystallographic data. J Chem Inf Mod (submitted)
34. Gore S, Velankar S, Kleywegt GJ. Implementing an X-ray validation pipeline for the protein data bank. Acta Cryst D (submitted)