



# Model validation: local diagnosis, correction and when to quit

Jane S. Richardson,\* Christopher J. Williams, Bradley J. Hintze, Vincent B. Chen, Michael G. Prisant, Lizbeth L. Videau and David C. Richardson

Department of Biochemistry, Duke University, Durham, NC 27710, USA. \*Correspondence e-mail: jsr@kinemage.biochem.duke.edu

Received 2 May 2017

Accepted 3 July 2017

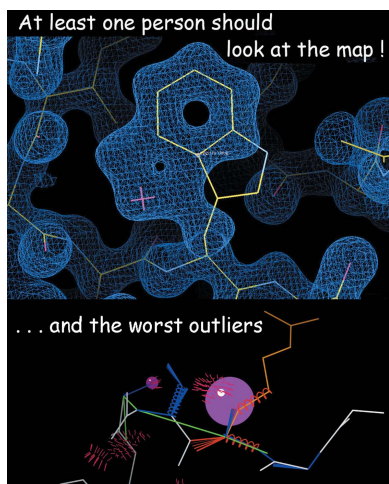
**Keywords:** structure validation; outlier correction; *MolProbity*; all-atom contacts; likelihood.

Traditionally, validation was considered to be a final gatekeeping function, but refinement is smoother and results are better if model validation actively guides corrections throughout structure solution. This shifts emphasis from global to local measures: primarily geometry, conformations and sterics. A fit into the wrong local minimum conformation usually produces outliers in multiple measures. Moving to the right local minimum should be prioritized, rather than small shifts across arbitrary borderlines. Steric criteria work best with all explicit H atoms. ‘Backrub’ motions should be used for side chains and ‘P-perp’ diagnostics to correct ribose puckers. A ‘water’ may actually be an ion, a relic of misfitting or an unmodeled alternate. Beware of wishful thinking in modeling ligands. At high resolution, internally consistent alternate conformations should be modeled and geometry in poor density should not be downweighted. At low resolution, *CaBLAM* should be used to diagnose protein secondary structure and *ERRASER* to correct RNA backbone. All atoms should not be forced inside density, beware of sequence misalignment, and very rare conformations such as *cis*-non-Pro peptides should be avoided. Automation continues to improve, but the crystallographer still must look at each outlier, in the context of density, and correct most of them. For the valid few with unambiguous density and something that is holding them in place, a functional reason should be sought. The expectation is a few outliers, not zero.

## 1. Background

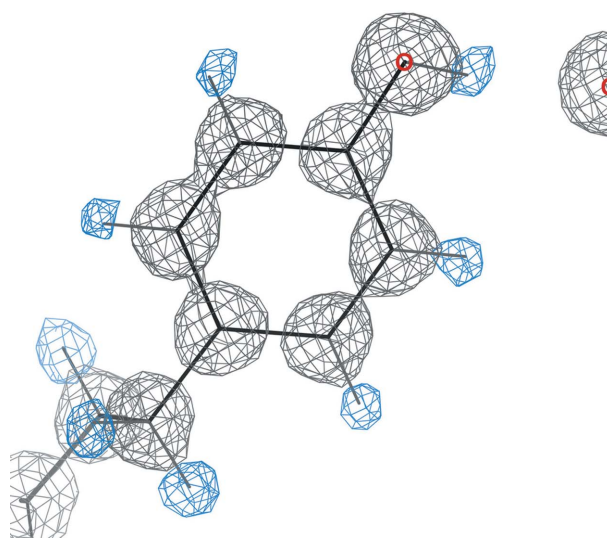
Structure validation highlights the good parts and identifies possible problems in macromolecular structures (initially for X-ray crystallography, but also for neutron, NMR and cryo-EM methods). It has aspects that address (i) the experimental data, (ii) the modeled coordinates and (iii) the model-to-data fit. It was spurred into existence around 1990 after two high-profile chain mis-tracings, and started with  $R_{\text{free}}$  (Brünger, 1992), bond lengths and angles (Engh & Huber, 1991), twinning (Yeates, 1997), and the multi-criterion systems of *PROCHECK* (Laskowski *et al.*, 1993), *OOPS* (Jones *et al.*, 1991) and *WHATCHECK* (Hoofst *et al.*, 1996). Here, we will emphasize validation of the model, from physical principles and prior experience, always within the requirement for a good model-to-data fit.

Finding problems is unrewarding unless they can be fixed, and our group has concentrated on techniques and criteria that can point the way for correction. This desire shifts the emphasis from global to local measures, and encourages cycles of validation and correction throughout the structure-solution process, not just at the end.

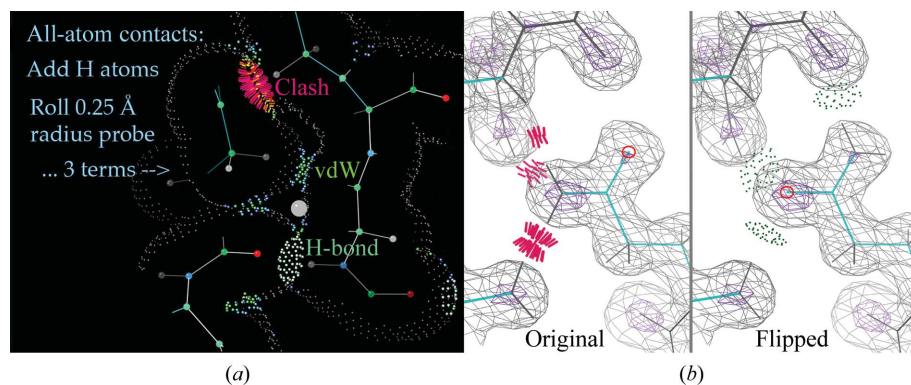


## 2. The foundation: all-atom contacts

The starting point for our *MolProbity* validation is the H atom. H atoms comprise about half of the atoms in biological macromolecules, as the ‘twigs’ at the outer edges of the covalent tree structure. About three quarters of all contacts within or between molecules have an H atom on one or both sides. Historically, H atoms were seldom included because they make calculations expensive and visualizations more cluttered, and because they are not directly observable in crystallography except at extremely high resolution, since they have only a single electron, and even this electron does not



**Figure 1**  
The H atoms are really there. Tyr13 in PDB entry 1yk4 for rubredoxin at 0.69 Å resolution, with  $2mF_o - DF_c$  map (black) contoured at  $1.5\sigma$  and  $mF_o - DF_c$  (blue) at  $2.8\sigma$ . All H atoms are visible, including the OH donor to a nearby backbone O.



**Figure 2**  
(a) The process of defining all-atom contacts. Small, pale dots are the van der Waals envelope of covalently bonded atoms, including H atoms. A probe sphere of 0.25 Å radius (gray) is rolled on this surface, leaving color-coded probe dots where it intersects with the surface of another atom. Favorable van der Waals contacts from just touching to 0.5 Å apart produce paired patches of green (close) and blue (more distant) contacts, while overlaps are either favorable hydrogen bonds between donor–acceptor atom pairs or repulsive overlaps in warm colors, with serious clashes of  $\geq 0.4$  Å shown by hot-pink spikes. (b) A definitive assignment of an amide ‘flip’ for Gln115 in PDB entry 1gk8. The larger  $\text{NH}_2$  group gives clashes and there are no hydrogen bonds in the original position, while the flip shows three hydrogen bonds and no clashes. The *REDUCE* flip is confirmed as correct by the higher electron density in the assigned O position.

diffract well. However, we know that they are really there, both from chemistry and from our very best structures (Fig. 1). All systems include some sort of ‘bump’ check, but most are simple center-to-center distance for the heavier atoms (C, N, O...) using the poor approximation of united atom radii that account for hydrogen volume but not directionality. We aim to convince the reader that explicit H atoms can be placed quite accurately from good heavier atom positions, and that considering the detailed geometry of their hydrogen bonds and steric contacts revolutionizes the ability to model and understand local structure.

All-atom contact analysis, as performed in *MolProbity* (Chen *et al.*, 2010), *PHENIX* (Adams *et al.*, 2010), *Coot* (Emsley *et al.*, 2010) and wwPDB validation (Berman *et al.*, 2003; Read *et al.*, 2011; Gore *et al.*, 2012), first requires *REDUCE* to be run (Word, Lovell, Richardson *et al.*, 1999). *REDUCE* adds all H atoms, at the electron-cloud center for X-ray data and at the nuclear position for NMR data (Deis *et al.*, 2013; Richardson *et al.*, 2013), and optimizes their hydrogen bonds, clashes and van der Waals contacts. It considers all possible combinations in complete hydrogen-bond networks, rotating OH, SH and  $\text{NH}_3$  groups but not methyl groups, and flipping Asn, Gln or His groups where needed. The first layer of waters is considered, but their H atoms are virtual and provide a donor or acceptor as needed by each interacting group. His protonation is assigned by avoiding clashes and optimizing hydrogen bonds. These strategies were chosen after extensive tests in maximizing overall accuracy by relying less on less reliably placed features, such as waters. The *PROBE* program (Word, Lovell, LaBean *et al.*, 1999) then looks at all noncovalent atom pairs whose van der Waals surfaces are  $\leq 0.5$  Å apart, calculating scores and producing dot surface visualizations, as shown in Fig. 2(a).

This methodology was originally tested (Arendall *et al.*, 2005), and is constantly being reaffirmed, by the near-perfect all-atom contacts seen in the well ordered parts of high-resolution structures. In less ideal circumstances the contacts act like a super-sized  $R_{\text{free}}$ , assessing the heavier atom positions by the fit of the unseen other half: the H atoms. An extremely useful byproduct of the *REDUCE* calculation is the automated correction of Asn, Gln and His ‘flips’ and assignment of His protonation (Richardson & Prisant, 2012), which are easy to get wrong just from a density fit, or even if only hydrogen-bonds are considered and not clashes, since for instance an  $\text{NH}_2$  group is much larger than an O atom (Fig. 2b). The right alternative is obvious to see and understand with this tool, but is very hard without it.

Since a hydrogen bond or a clash is between a specific atom pair and is highly directional, crystallographers can

see how to correct the problems that they diagnose. As plotted in Fig. 3, average all-atom clashscores have improved from 11 down to four clashes per 1000 atoms in mid-range PDB depositions worldwide since the start of *MolProbity* in 2002 (Davis *et al.*, 2004). The change has been remarkably steady, but with a few blips, including last year. It is possible that it has begun to level off: it cannot go below zero and should not actually reach zero, since our parameters are not perfect and there are usually a few small unexplained, unfixable clashes even in the best structures. The point is not to avoid all clashes by nudging something just across the 0.4 Å threshold; this does not improve biological understanding, and refinement is quite likely to nudge it right back. Instead, use clashes to find and fix places where a group was fitted into the wrong local minimum, which often does affect an interpretation. If enough of these are corrected, the map will improve elsewhere, perhaps in the active site.

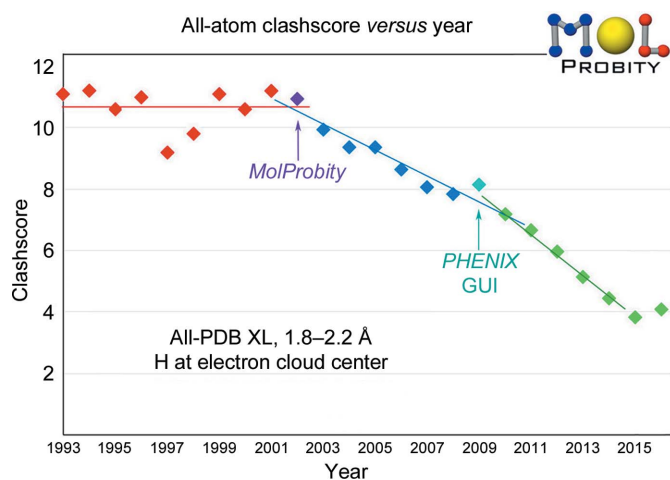
### 3. Updates of traditional validation measures

Since serious problems usually show up in multiple criteria, and since it is possible to ‘game’ any one measure at the expense of the others, model validation should be as comprehensive as feasible. Fig. 4 shows a key to all of the three-dimensional graphical outlier markup in *MolProbity*, as seen on the website and in the subsequent figures. Outliers are also listed in chart or table form, with their parameters and scores. All-atom contact dots and spikes have already been explained in Fig. 2(a). Bond-length and angle parameters were pioneered by Engh & Huber (1991) for proteins and by Berman and Olson (Gelbin *et al.*, 1996) for nucleic acids; they are shown here as springs or fans, in red if too broad and in blue if too short. They have been updated primarily by making them context-dependent: on  $\varphi$ ,  $\psi$  by Karplus and Dunbrack (Berkholtz *et al.*, 2009), on ribose pucker by us (Jain *et al.*, 2015) or by combining angles into the  $C^\beta$  deviation, which flags a  $C^\beta$  position forced to be nontetrahedral by an incorrect local fit of

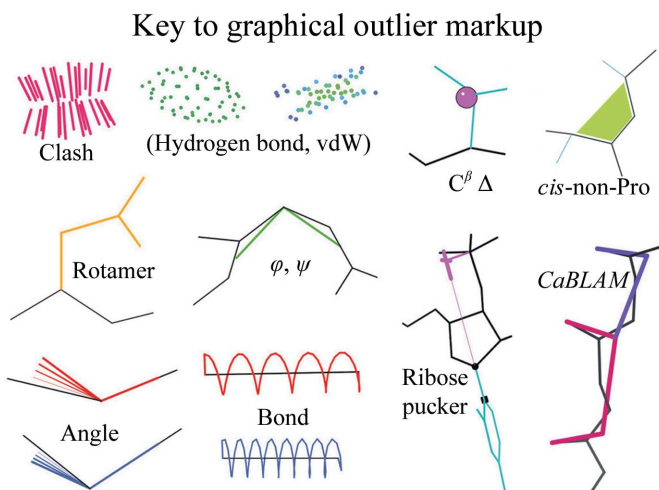
either the side chain or the backbone (Lovell *et al.*, 2003). Outliers in conformation (combinations of rotatable dihedrals) are shown in gold for side-chain rotamers, in green for Ramachandran and as magenta crosses for ribose pucker. Newer validation types are designed to be more robust at low resolution (*CaBLAM*) or to flag systematic errors such as too many *cis*-non-Pro peptides. Each of the last five will be discussed further below.

The Ramachandran plot was first used for model validation by Thornton and coworkers (Morris *et al.*, 1992) and has since been updated by many groups. Current Ramachandran criteria as used in *MolProbity*, *PHENIX* and at the wwPDB (Read *et al.*, 2011; Richardson *et al.*, 2013) divide the amino-acid types into six categories, each with its own  $\varphi$ ,  $\psi$  plot (Fig. 5). The use of reference data that have been quality-filtered both at the file level (resolution, redundancy, *MolProbity* score,  $\leq 5\%$  of residues with various bad outliers) and at the residue level (in this case, just no alternate conformations or backbone  $B$  factors  $> 30 \text{ \AA}^2$ ) gives plots with extremely few data points outside a smooth, reproducible outlier contour: for the general case, only one in 5000 reference examples is in this outer 60% of the plot area. These data points are mostly valid, but in poor density or at lower resolution most Ramachandran outliers are wrong. In either case, though, each outlier should be looked at. These conformations are rare because they are strained, so they are not believable without both good density and something holding them in place. The few valid outliers are very likely to be functionally important and are conserved because it is worthwhile for the molecule to spend energy stabilizing them.

Note that each of the six plots in Fig. 5 has very differently shaped contours; the other 16 amino acids show nearly identical outlier and 2% contours and so are combined in the general distribution (see the figures in the Supplementary Information for Read *et al.*, 2011). Nearly all amino acids differ significantly in their highly populated  $\varphi$ ,  $\psi$  regions, but validation does not use this part of the distribution because we



**Figure 3**  
Plot of average clashscore *versus* year for mid-resolution PDB depositions worldwide, showing a steady improvement since the introduction of *MolProbity*.



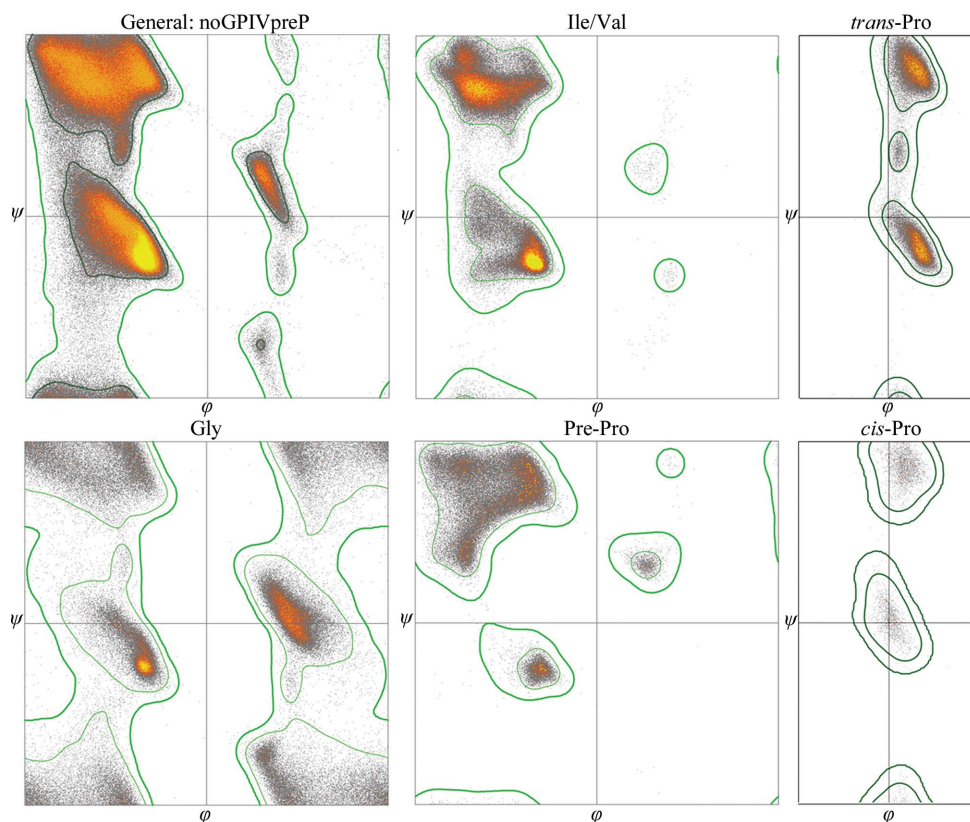
**Figure 4**  
Key to the markup for various categories of validation outliers in *MolProbity*'s three-dimensional graphics, as also seen in the figures here.

should not penalize a good loop or  $\beta$ -strand for not lying in the tall, bright peak of regular  $\alpha$ -helix. For the rather different needs of prediction or design, the highly populated regions of  $\varphi$ ,  $\psi$  or rotamer space are more important than the edges, so

more narrowly divided libraries are used in that work.

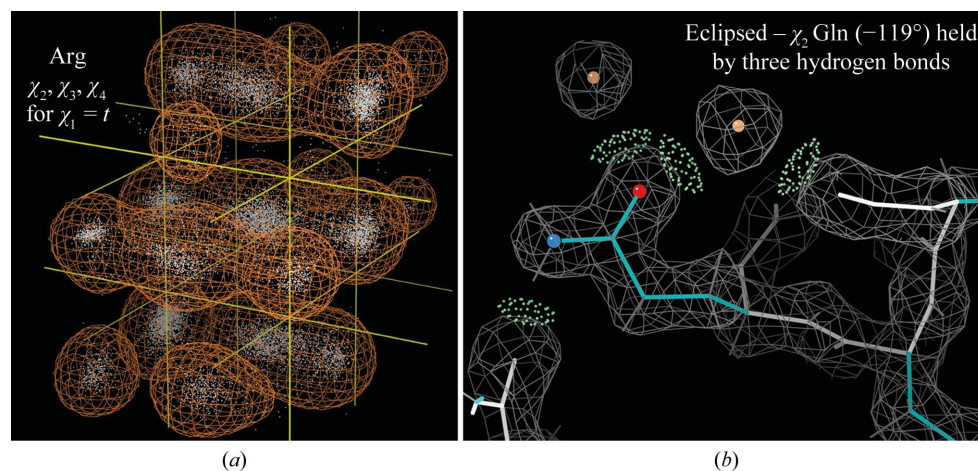
Side-chain rotamers are another traditional (Ponder & Richards, 1987) but recently updated protein validation criterion. Side chains have between one and four  $\chi$  dihedral angles, and their rotamer distributions must be treated as multidimensional because the occurrence probabilities are not just a product of the individual  $\chi$  preferences. Fully tetrahedral  $\chi$  values are quite cleanly trimodal, while  $\chi$  values between a tetrahedral and a planar group are clustered but more continuous and often unevenly spaced, as seen for Arg  $\chi_4$  in Fig. 6(a). Even for the all-tetrahedral case the barriers to rotation are only about  $3 \text{ kcal mol}^{-1}$ , so two or three hydrogen bonds can hold an eclipsed  $\chi$  angle in place, as for the Gln in Fig. 6(b), or tight packing can hold an eclipsed  $\chi$  angle for an aromatic ring. However, a side chain out on the surface without strong interactions should be fitted as a good rotamer, or better yet as two or three alternates in good rotamers.

Our most recent rotamer update (Hintze *et al.*, 2016) is in *MolProbity* and *PHENIX* but not yet in wwPDB validation. It took advantage not only of more high-resolution structures but also of the PDB requirement since 2008 for data deposition, in order to improve the residue-level quality filters. A combination of  $2mF_o - DF_c$  electron-density values, real-space correlation and *B*-factor measures produced a large reference data set with very few false positives (that is, data points that were unconvincing on manual inspection). This allowed us to define well behaved outlier contours that exclude only 0.3% rather than the previous 1% of the reference data, and to add a 2% contour defining allowed *versus* favored rotamers, analogous to standard practice for Ramachandran measures.



**Figure 5**

The six plots of data distributions and contours used for current Ramachandran validation. The million residues of quality-filtered data from the Top8000 data set are color-coded in  $0.1^\circ$  pixels, from gray for one data point to bright yellow for the highest density (30–45 data points per pixel in the general distribution). For Gly the outer contours are symmetric but the data are not, since Gly serves different functional roles in the  $\alpha$  and the  $L\alpha$  regions.



**Figure 6**

(a) Data-point distribution and 2% contour in the  $\chi_2$  (into page),  $\chi_3$  (vertical) and  $\chi_4$  (near-horizontal) dimensions for quality-filtered Arg residues with  $\chi_1$  *trans*. (b) A valid rotamer outlier: Gln321 in PDB entry 1n83, with its eclipsed  $\chi_2$  ( $-119^\circ$ ) held by three side-chain hydrogen bonds and clearly validated by the electron density.

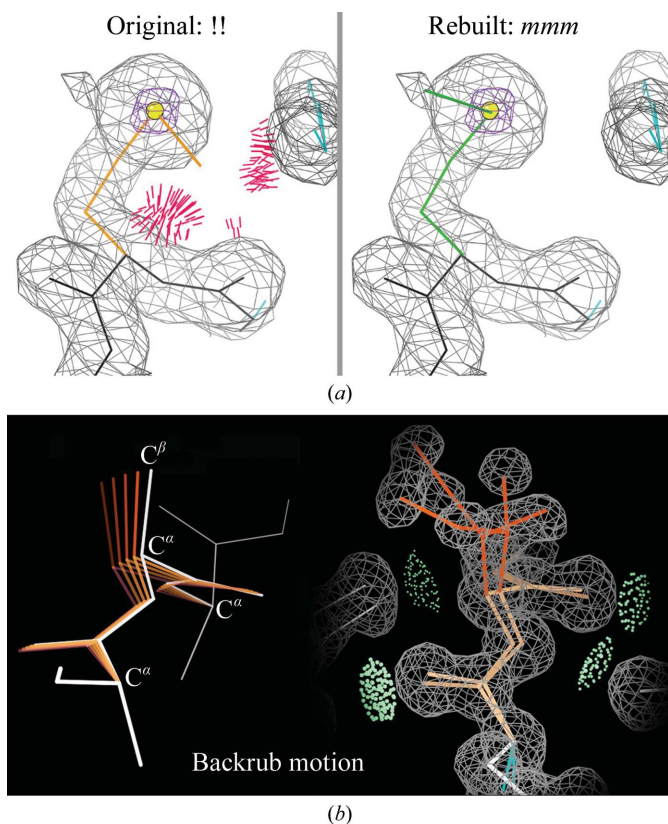
#### 4. Typical corrections

A misfitting inside a molecule usually produces outliers in more than one criterion. A bad clash means that one or both of the clashing atoms is in the wrong position, but it is seldom the case that just moving the two atoms apart is the right answer. For instance, for the selenomethionine (Mse351) in Fig. 7(a) all of the clashes have the methyl group in common, and the side chain is a rotamer outlier. Among rotamers that keep the Se atom centered in its clear density, the *mmm* rotamer ( $-60, -60, -60^\circ$ ) fits with no clashes and even places the methyl in a tiny bit of electron density, providing a win-win answer.

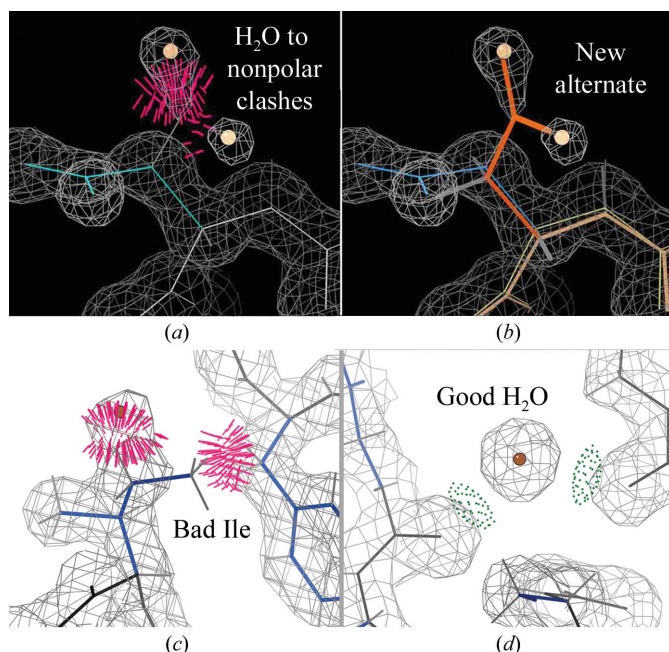
For rotamer problems that shift more of the side chain, the subtle but powerful ‘backrub’ motion (Davis *et al.*, 2006) is usually needed. It is approximated by a rigid rotation around an axis between the  $i \pm 1$   $C^\alpha$  atoms, with small reverse rotations of the two individual peptides if needed to maintain their hydrogen bonds or to keep the carbonyl O atoms in density (Fig. 7b). A backrub is the most common movement used by protein backbone to accommodate alternate side-chain conformations that change the rotamer or hydrogen bonding (or even mutations; Keedy *et al.*, 2012), using the leverage of a

very small backbone shift in the  $C^\alpha - C^\beta$  bond-vector direction to make substantial changes in side-chain contacts. This same motion is amazingly effective in correcting a misfitted side chain. The backrub correction can be directly controlled in the *KiNG* graphics and modeling program (Chen *et al.*, 2009; Richardson & Headd, 2012), providing a better understanding of how it works within the automated side-chain rebuilding in *Coot* or *PHENIX*.

Another very common type of correction that is needed is finding the right model interpretation for a problematic density peak fitted as a water (Headd & Richardson, 2013). If the ‘water’ is really an ion it will show all-atom clashes rather than hydrogen bonds to polar groups, and the charge of these groups will indicate the complementary charge of the ion. If several ‘waters’ are too close together within continuous, interestingly shaped density, they probably represent an unidentified small-molecule ligand. If it proves impossible to figure out what the ligand is, it would be a great help to future users of the structure to give the group the UNL (UNidentified Ligand) designation provided by the PDB. If a ‘water’ with low or no density clashes with an outlier residue, it may be left over from a difference peak next to a misfitting such as a backward-fitted  $C^\beta$ -branched side chain, and correcting the side chain and deleting the water may give an excellent answer. If the ‘water’ has clashes with non-outlier nonpolar atoms, as in Fig. 8(a), it is almost always the next atom in an unmodeled alternate conformation; trying additional rotamers, with backrub as needed, may identify a perfect fix such



**Figure 7**  
(a) An easy outlier correction. The methyl of Mse351 in PDB entry 1j58, as deposited, shows several bad clashes and the side chain is a rotamer outlier. The favorable *mmm* rotamer fixes both problems. (b) The backrub motion, shown as a schematic of the small-amplitude backrub rotation around the  $C^\alpha$   $i - 1$  to  $i + 1$  axis with leverage on the  $C^\alpha - C^\beta$  direction and on side-chain contacts, and an example of alternate conformations in Ile47 of the 1n9b  $\beta$ -sheet, where the backrub shift allows good packing in two distinct rotamers.



**Figure 8**  
Interpreting ‘water’ peaks. (a) When they clash with nonpolar atoms then they are likely to be the next atoms in an unmodeled alternate (b), as shown for Asp9 in PDB entry 1eb6. (c) Do not let a ‘water’ push a side-chain atom out of its density, as happened for Ile195 in PDB entry 3js8. (d) A good water peak should show density separated from other atoms, with at least one polar interaction at good hydrogen-bond distance; here, HOH 543 in PDB entry 3js8 makes two good hydrogen bonds. Contours are at  $1.2\sigma$ .

as the Asp in Fig. 8(b). As also noted below, a water clash with an alternate conformation means that the water needs a lower occupancy, and preferably an alternate flag as well. The case, one hopes, that will not occur in your more careful structure is where a ‘water’ was fitted early on into the density of a side-chain atom and forced this side-chain branch out of place. As in Fig. 8(c), this type is immediately clear once anyone actually looks at it. Finally, the peak may of course really be a water molecule, in which case it should have a peak well separated from the macromolecule, with at least one polar interaction at good hydrogen-bonding distance, as for the water in Fig. 8(d).

## 5. High resolution

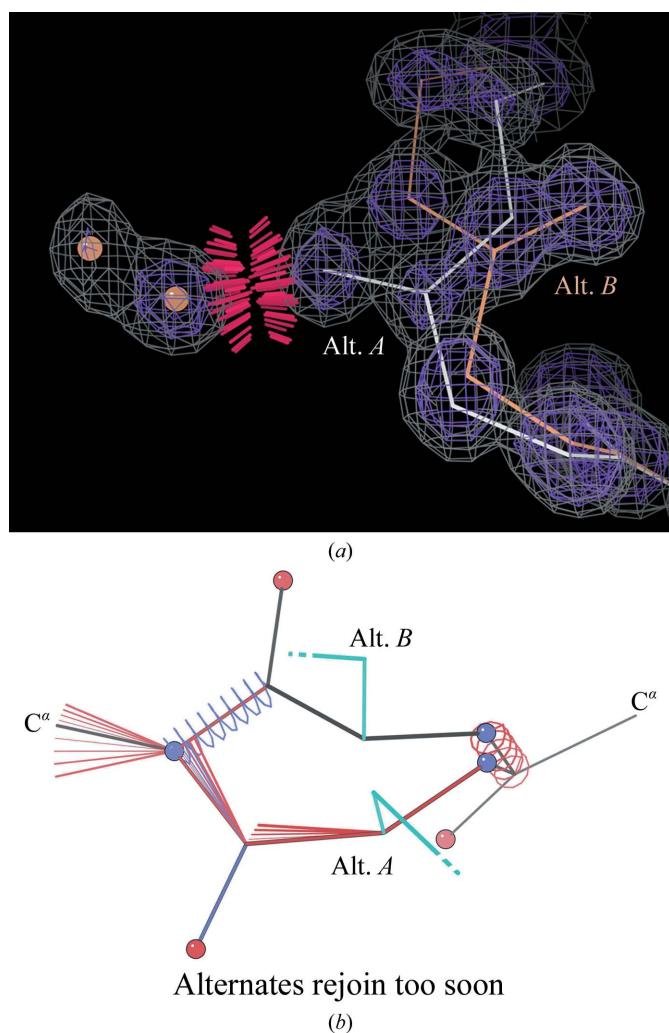
High resolution is wonderful, but complex, and requires much more work than one might expect. The hardest part is correctly sorting out the many alternate conformations that

are visible. Even when each atom position shows a distinct peak, they often cross back and forth confusingly, making it hard to trace more than one valid, self-consistent model through them. Historically, the problems are worst in alternates *B* and higher, since most validation systems have not evaluated these at all. *MolProbity* has always given some assessment of multiple alternates, and we are now in the process of making this functionality more complete and easy to use. Some advice in the meantime is that if the occupancies are discernably different, use the relative peak heights to join up atoms. Clashes with neighboring groups often mean that they also need modeled alternates, perhaps not differing enough to have been obvious. Nearby waters are the most frequent offenders, as in Fig. 9(a), since they are really part of the alternate conformation network but are typically only given high *B* factors rather the partial occupancy and alternate flags that they really deserve. Watch out for deviant bond-length, angle and  $C^\beta$  deviations, which signal that the alternates are either incorrectly mixed or that separate alternates were not defined widely enough along the covalent structure. If alternates for a side chain are fitted with  $C^\beta$  atoms  $>0.2 \text{ \AA}$  apart then alternates should also be defined in the backbone, and if any alternate backbone atoms are widely separated then the alternates should not rejoin at the peptide bonds (as is typically performed), but only at the flanking  $C^\alpha$  atoms. *PHENIX* can now extend alternates in this manner (Deis *et al.*, 2013) far enough out to avoid the sort of dire geometry seen in Fig. 9(b). In another program, try duplicating and alternate flagging the few extra atoms that need to separate slightly.

A problem that is easier to avoid is globally downweighting the geometry term at high resolution, so that loops or termini in poor density can end up with near random-like coordinates,  $>1 \text{ \AA}$  clashes and  $>20\sigma$  bond-length and angle deviations (Chen *et al.*, 2011). Always look at your very worst outliers before deposition!

## 6. Tackling low resolution

There is no question that low resolution ( $\geq 3 \text{ \AA}$ ) is a truly difficult challenge. Discernable bumps for carbonyl O atoms mostly disappear, giving ill-defined backbone conformations. Some side-chain atoms should genuinely lie outside the density, not only confusing rotamer choice but also tempting both people and software to scrunch them back in. Electron-density connectivity is part way along in its change from following the atomic connectivity at  $2 \text{ \AA}$  to being a slab for  $\beta$ -sheet and a solid tube for  $\alpha$ -helix by  $5\text{--}6 \text{ \AA}$ , and it makes this change through inconsistent, misleading intermediate forms. Crystallographic methods were developed at resolutions where one can first trace the chain and then deal with side chains, but at low resolution they mix together, with the size and position of local side chains causing backbone density to break or coalesce in the wrong patterns. The best overall advice is always to fit structure as much more regular and ideal than it looks, which essentially always turns out to be the right answer if a higher resolution structure is solved later. For

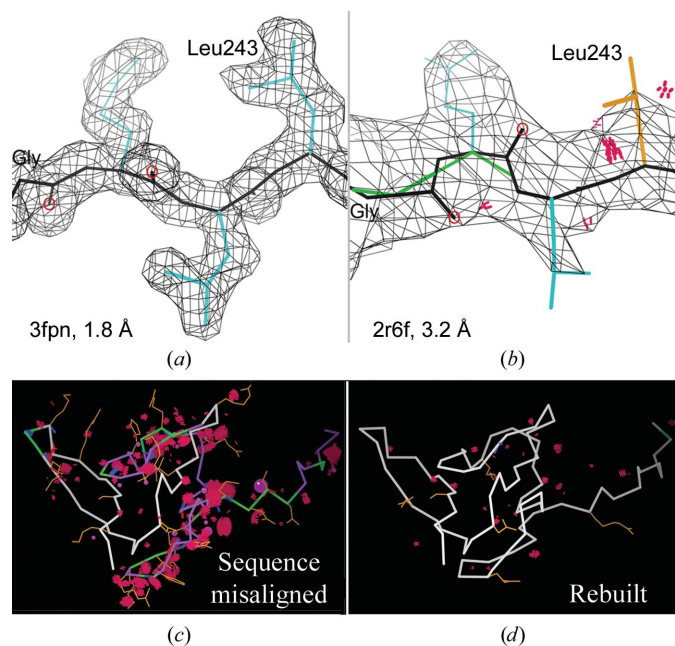


**Figure 9**  
Handling alternate conformations. (a) Use peak heights to assign consistent alternate IDs, including partial occupancies for interacting waters, which was not performed here. Leu105 from PDB entry 1gwe at 1.2 and  $3\sigma$ . (b) When any alternate backbone atom is widely separated, do not rejoin alternates until the flanking  $C^\alpha$  atoms in order to avoid bad geometry. Here, for Asp42 from PDB entry 1w0n, there are bond-length outliers up to  $8\sigma$  and bond-angle outliers up to  $12\sigma$ .

instance, Fig. 10(a) shows that the blobby bit of structure in Fig. 10(b) is actually a very regular  $\beta$ -strand with full hydrogen bonding, good rotamers and no Ramachandran or clash outliers. Once regular stretches of secondary structure have been fitted, they can be bent or shifted somewhat better into the overall density by refinement tools such as *DEN* (Schröder *et al.*, 2010), jelly body (Murshudov *et al.*, 2011) or morphing (Terwilliger *et al.*, 2012), but preferably with the help of judicious hydrogen-bond restraints to minimize distortion.

Beware of sequence misalignments, which hurt biological interpretation by misplacing functional side chains large distances from where they should be, and which are very much more common at low resolution than anyone would like to believe. Figs. 10(c) and 10(d) show a before-and-after portrait for the longest of about 20 misalignments that we have worked on (Dunkle *et al.*, 2011). When you find the right offset, the improvement is sometimes magical.

To more easily and reliably locate helices and strands at low resolution, there is a new tool called *CaBLAM* (*C $\alpha$ -Based Low-resolution Annotation Method*; Williams *et al.*, 2013). It defines a novel parameter space of  $C^\alpha$ - $C^\alpha$  and CO-CO virtual dihedrals, where the CO dimension diagnoses fitting problems caused by the frequent large distortions of peptide orientation at low resolution and the two  $C^\alpha$  dimensions identify the probable secondary structure disguised by these problems. The  $C^\alpha$  trace is the most reliable feature in low-resolution models, whereas peptide orientations are often very incorrect, destroying the use of  $\varphi$ ,  $\psi$  Ramachandran values or backbone

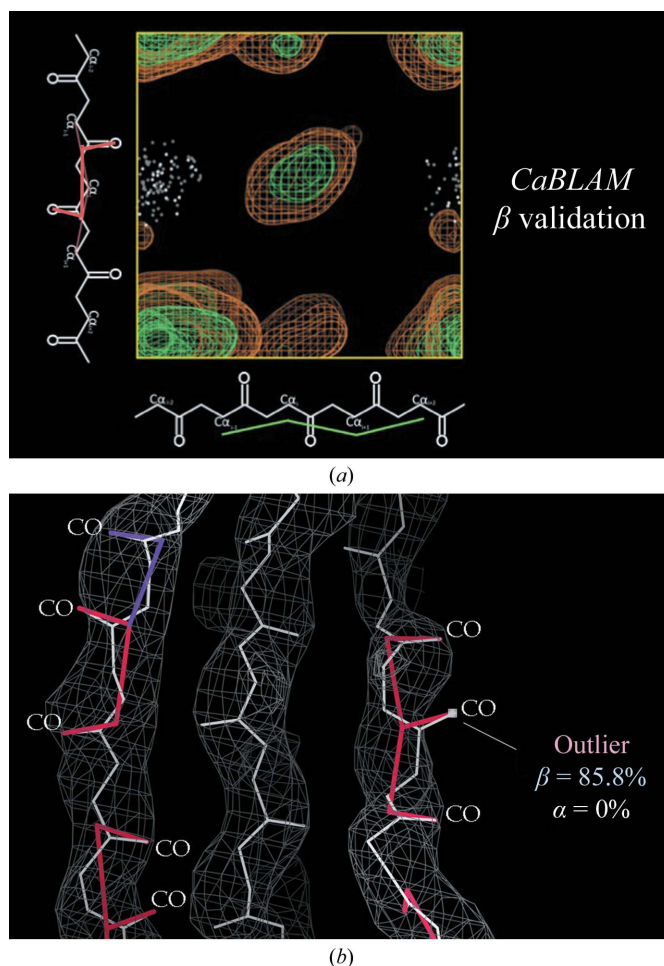


**Figure 10** Difficulties at low resolution. (a) Higher resolution shows that this is a regular  $\beta$ -strand with no outliers or clashes, but in (b) the backbone CO directions are misoriented because they are not observed, and the side chains are pulled inwards towards density nubbins. Contours are at  $1.2\sigma$ . (c) Localized outliers for a long sequence misalignment in 70S ribosomal protein L27 at 3.2 Å resolution (PDB entry 3i1n). (d) After rebuilding of the one- to three-residue sequence shifts in the improved PDB entry 4gd1 (Dunkle *et al.*, 2011).

hydrogen bonds for identifying secondary structure. Fig. 11(a) shows a plot of the CO *versus*  $C^\alpha$ -in virtual dihedrals with contours for  $\beta$  structure in the good reference data. The white data points, which are well outside the 99% contour, represent >100 cases of a systematic error that is fairly common at low resolution, where three CO bonds in a row point in the same direction rather than alternating. Fig. 11(b) shows a ribosomal protein example of this, where *CaBLAM* diagnoses outliers and an 85% likelihood of disguised  $\beta$  structure, guiding corrections that lowered all types of outliers and improved the hydrogen bonds and *R* factors. *CaBLAM* validation is reported in *MolProbity*, and is available in *PHENIX* as one of the diagnostics for secondary structure.

## 7. RNA

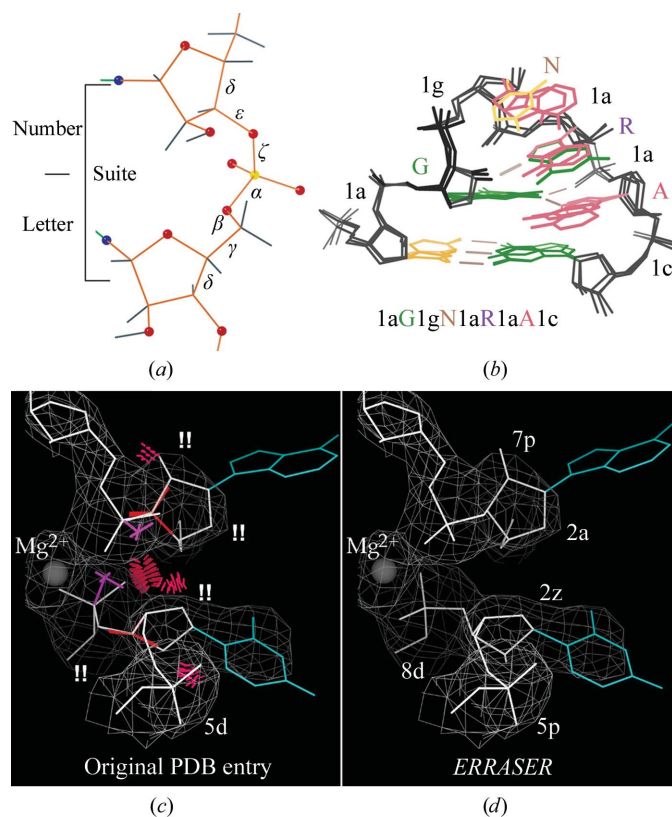
RNA structure has quite different properties than either DNA or protein, and its complex tertiary structures, catalytic and binding functions, and roles in large dynamic molecular machines make it highly important. RNA backbone confor-



**Figure 11** *CaBLAM* low-resolution diagnosis. (a) Plot of CO *versus*  $C^\alpha$ -in virtual dihedrals, with contours for the good  $\beta$ -sheet reference data and white data points for examples that have three adjacent CO bonds parallel rather than alternating. (b) *CaBLAM* scoring for three such outliers, showing definitively that they should be fitted as regular  $\beta$  structure. PDB entry 3i1n; contours at  $1.2\sigma$ .

mation is crucial to all of these functions, but has too many variables to model straightforwardly at the usual resolutions attainable, where the backbone between ribose and phosphate is a featureless tube. Fortunately, there are some tools to help with this problem (Jain *et al.*, 2015).

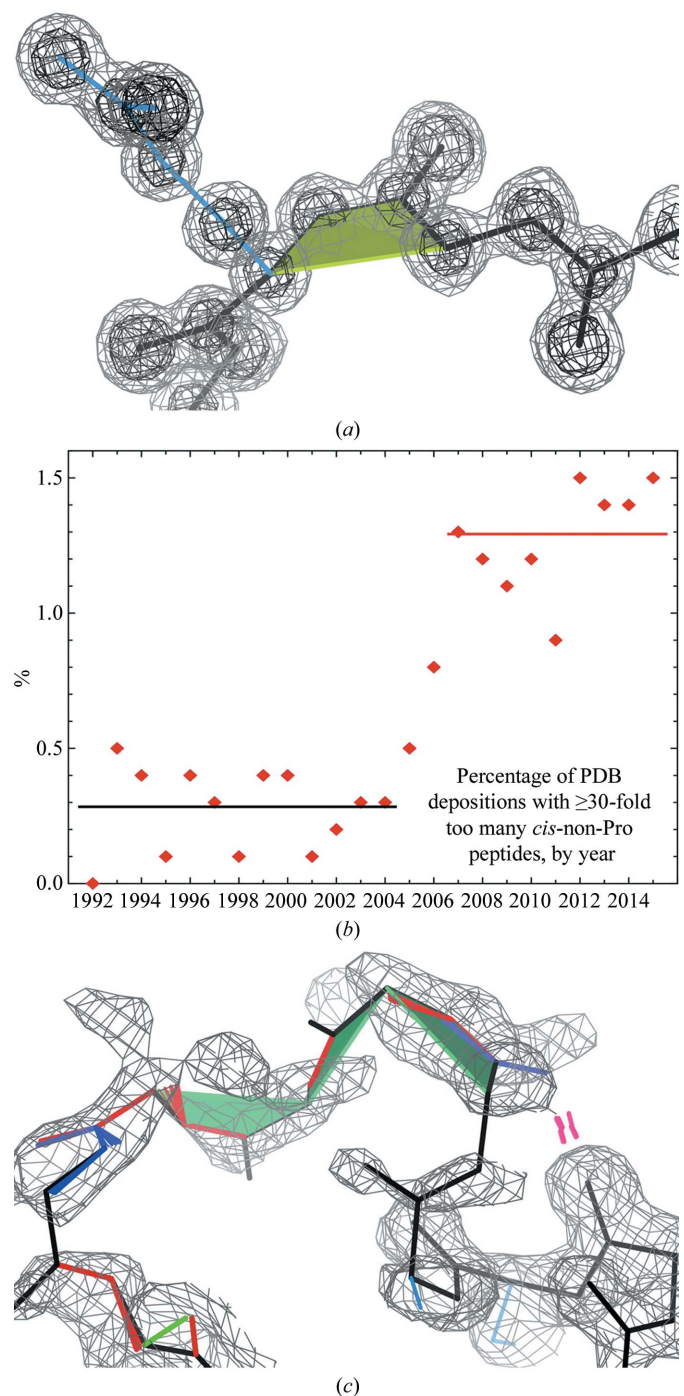
Dividing RNA backbone into sugar-to-sugar units (suites) works better than using chemical nucleotides and shows that the RNA backbone is 'rotameric' (Murray *et al.*, 2003), while that of DNA is not. As an international collaborative effort (Richardson *et al.*, 2008), we defined a set of about 50 distinct, valid RNA backbone conformers across the seven variable dihedrals in a suite, each with a two-character name (Fig. 12*a*). The GNRA tetraloop conformation is described in this system as 1aG1gN1aR1aA1c (Fig. 12*b*), and the conformers can be used in model building. Even simpler and more powerful is the 'P-perp' test for 2'-endo versus 3'-endo ribose puckers. At most resolutions the ribose pucker cannot be seen. Fortunately, however, it turns out to be determined by two features that can be seen in the density: the phosphates and the direction between the ribose and base, which is the line of the glycosidic bond. The P-perp test drops a perpendicular to this line from the 3' P; if the perpendicular is long ( $\geq 3$  Å) the ribose pucker is 3' and if it is short the pucker is 2'. This difference is easy to



**Figure 12** RNA backbone conformers and corrections. (a) Definition of the suite divisions (sugar-to-sugar) of RNA backbone. (b) Using two-character suite names to describe the backbone conformation of the GNRA tetraloop. (c) Original conformation of two touching loops in the riboswitch (PDB entry 2gis) at 2.9 Å resolution, with clashes, bad ribose puckers and four out of five outlier suite conformers (!!). (d) After correction using *ERRASER*.

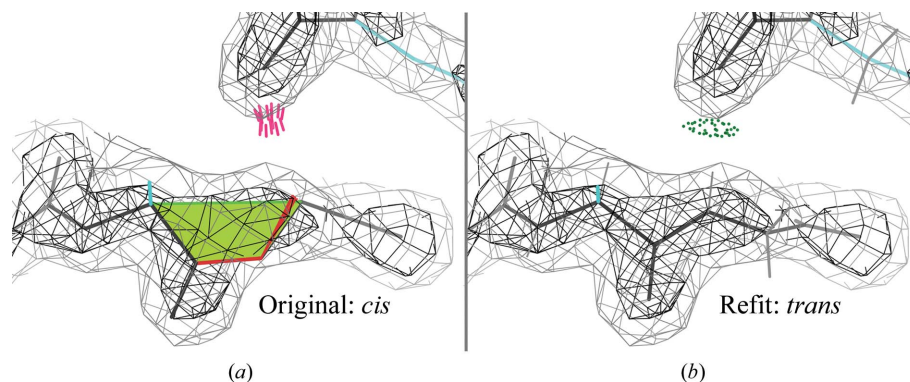
see by eye when manually fitting, and it is performed on the fly in *PHENIX* refinement to allow use of pucker-specific target parameters.

Correcting the whole backbone conformation is a very hard search problem. Rhiju Das, with our encouragement, developed an automated protocol called *ERRASER* (Chou *et al.*,



**Figure 13** Use and overuse of very rare *cis*-non-Pro peptides. (a) A clear, genuine *cis*-non-Pro in PDB entry 2ddx at 0.86 Å resolution, flagged by the seagreen trapezoid. (b) Time-course plot of the epidemic overuse of *cis*-non-Pro peptides. (c) An example of how *cis* peptides can fit better than *trans* peptides into patchy, poor electron density at  $1.2\sigma$ . PDB entry 2j82, 1092 loop.





**Figure 14**

Likelihood-based choice of *cis* versus *trans* peptides. (a) Original *cis* model of Lys–Gly270 in PDB entry 2cn3, with a decent fit to the contours at 1.2 and  $3\sigma$ . (b) Model rebuilt as *trans*, with even better fit, a hydrogen bond rather than a clash, and a log-likelihood gain of 75.6, plus eight units better log prior probability.

2012; Adams *et al.*, 2013). It diagnoses clashes and incorrect puckers with *MolProbity*, uses *relax* in *Rosetta* and a special-purpose brute-force search within *Rosetta*, followed by *Rosetta* and *PHENIX* refinement, and repeats this cycle. It requires the installation of *Rosetta* and takes a large amount of computer time, but produces excellent results and truly beats weeks of rebuilding and usually failing. As an example, Figs. 12(c) and 12(d) show a noncovalent contact of loops in a riboswitch with an Mg atom (currently *ERRASER* does not see non-RNA atoms); it fixes all clashes and pucker outliers, all suite conformer outliers and even improves the Mg coordination. Its corrections are not always quite this thorough, but it is amazingly effective. If one is having serious problems with an RNA structure, *ERRASER* should really be tried.

## 8. Avoiding systematic errors by model building with likelihood

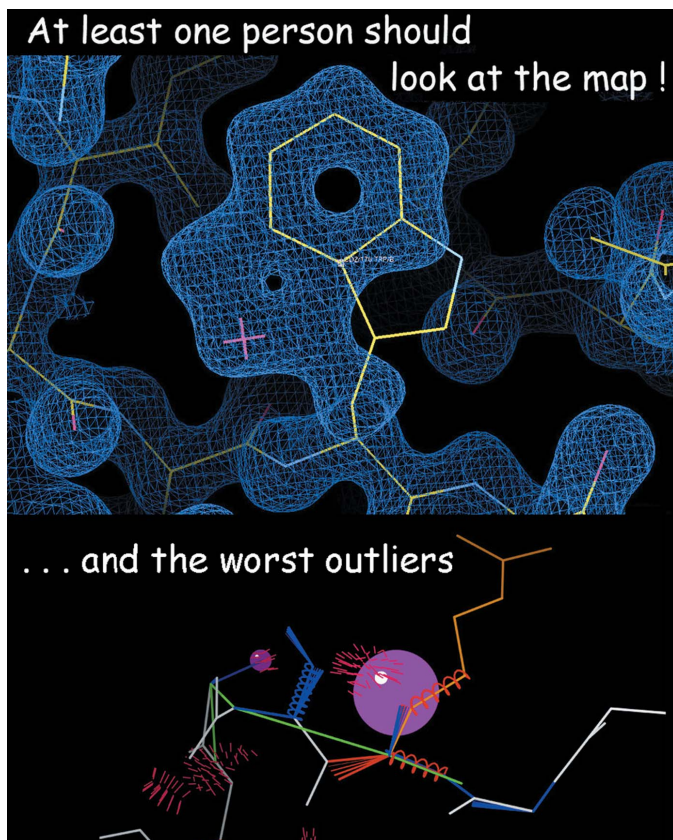
Most of this validation has been around for quite a long time, and one might think that all one needs to do is check out the sliders and worst specifics on the PDB report. However, new issues keep arising from new methodologies or from unusual things in new macromolecules. Much of validation is now built into automated procedures, which is great, but occasionally this makes something new go wrong. The most notable current example is with *cis*-non-Pro peptides. As seen in Fig. 13(a) there are indeed genuine examples, which are almost always functionally important, but they are very rare: only about one in 3000 nonproline peptides are *cis*. However, as first pointed out by Croll (2015), in the last ten years there has been an epidemic of their overuse, with >10% of structures containing  $\geq 30$  times too many *cis*-non-Pro peptides (Fig. 13b), probably without the depositors realising. Once a conformation is in lists or fragment libraries, it becomes used whenever it happens to fit just a little better, and overall usage is not tracked. The *cis*-non-Pro case is even worse than random, since a *cis* peptide is more compact than a *trans* peptide and fits better into patchy density or contracted density in a low-resolution loop (Fig. 13c). The first, easy stopgap is that *cis*-

non-Pro peptides are now prominently flagged (see Fig. 4; Williams & Richardson, 2015) in *MolProbity*, *PHENIX* and *Coot* so that people will be aware of them. Highly twisted peptides ( $\omega > 30^\circ$  nonplanar) are also now flagged; they are essentially never correct, since only two clear examples of  $>30^\circ$  have been found in good reference data (Berkholz *et al.*, 2012).

However, we really want to prevent these in the first place, rather than needing to correct them later. Such overuse occurs because conformations or fragments are chosen one at a time, only considering local density fit, with no weighting by probability of occurrence. For low-resolution crystallography or poor electron density, we need a Bayesian likelihood approach to model building. This means balancing the relative prior probability (frequency of occurrence) against the relative likelihood, for each alternative, that it could produce the observed density or data.

Our test implementation of this strategy (i) defines prior probability as the log of occurrence frequency in our laboratory's reference data set, currently the Top8000 data set, filtered by clashes, electron density and other quality criteria (Hintze *et al.*, 2016) and (ii) defines data likelihood by the reciprocal-space likelihood function (Read, 1986, 2001) that has been successful in refinement and molecular replacement. Different proposed conformations (for example *cis* versus *trans*) are compared by running carefully equivalent parallel refinements and balancing the resulting log-likelihood scores with the prior probabilities.

The results so far are surprisingly sensitive. Cases where the correct answer is truly obvious give log-likelihood differences in the hundreds, so no priors are needed. The more subtle example of the Lys–Gly *cis*-peptide in Fig. 14(a), which is a decently acceptable fit in its density, gives a worse LLG than the *trans* rebuild (Fig. 14b) by 75.6. Adding the prior-probability disadvantage of 8 (log of 3000:1) prefers the *trans* form by 83.6, while *R* and *R*<sub>free</sub> are unhelpful ( $\pm 0.0001$ ). This unambiguous answer is confirmed by the correction of a bond-angle outlier and the change of a clash to a hydrogen bond, neither of which are seen by the Bayesian calculation. In even more demanding tests at 3.5–4 Å resolution, the  $\Delta$ LLG was in



**Figure 15**  
A take-home message.

the range of  $-2$  to  $+10$  in favor of what we believe to be the correct answer, allowing the prior probability to make the choice clear. In contrast, real-space fit measures can show strong but incorrect preferences (such as for squeezing all atoms into density). After this proof-of-concept, we are now working on implementing this method in practical, production-level model building, so we hope that this type of validation outlier will become avoidable with less effort.

## 9. Conclusions: when to stop

Here are the main take-home points for your own work.

- (i) At least one person should look at the map and at each of the worst outliers (Fig. 15).
- (ii) The goal is few outliers, not zero.
- (iii) Follow the Zen of model anomalies.
  - (1) Correct most of them.
  - (2) Treasure the genuine few.
  - (3) Then rest serenely.

## Acknowledgements

We thank the other *PHENIX* development teams, especially Randy Read for his reciprocal-space likelihood function and general advice on likelihood, Pavel Afonine for the paired-refinement script and Tom Terwilliger for testing probability-weighted fragments in *RESOLVE*. We thank

crystallographers for all of the beautiful structures that made these elegant plots and accurate criteria possible, and the wwPDB for making data deposition mandatory.

## Funding information

The following funding is acknowledged: National Institutes of Health, National Institute of General Medical Sciences (award No. R01-GM073919; award No. P01-GM063210).

## References

- Adams, P. D. *et al.* (2010). *Acta Cryst.* **D66**, 213–221.
- Adams, P. D., Baker, D., Brunger, A. T., Das, R., DiMaio, F., Read, R. J., Richardson, D. C., Richardson, J. S. & Terwilliger, T. C. (2013). *Annu. Rev. Biophys.* **42**, 265–287.
- Arendall, W. B. III, Tempel, W., Richardson, J. S., Zhou, W., Wang, S., Davis, I. W., Liu, Z.-J., Rose, J. P., Carson, W. M., Luo, M., Richardson, D. C. & Wang, B.-C. (2005). *J. Struct. Funct. Genomics*, **6**, 1–11.
- Berkholz, D. S., Driggers, C. D., Shapovalov, M. V., Dunbrack, R. L. Jr & Karplus, P. A. (2012). *Proc. Natl Acad. Sci. USA*, **109**, 449–453.
- Berkholz, D. S., Shapovalov, M. V., Dunbrack, R. L. Jr & Karplus, P. A. (2009). *Structure*, **17**, 1316–1325.
- Berman, H. M., Henrick, K. & Nakamura, H. (2003). *Nature Struct. Biol.* **10**, 980.
- Brünger, A. T. (1992). *Nature (London)*, **355**, 472–475.
- Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S. & Richardson, D. C. (2010). *Acta Cryst.* **D66**, 12–21.
- Chen, V. B., Davis, I. W. & Richardson, D. C. (2009). *Protein Sci.* **18**, 2403–2409.
- Chen, V., Williams, C. & Richardson, J. (2011). *Comput. Crystallogr. Newsl.* **2**, 86. [http://phenix-online.org/newsletter/CCN\\_2011\\_07.pdf](http://phenix-online.org/newsletter/CCN_2011_07.pdf).
- Chou, F.-C., Sripakdeevong, P., Dibrov, S. M., Hermann, T. & Das, R. (2012). *Nature Methods*, **10**, 74–76.
- Croll, T. I. (2015). *Acta Cryst.* **D71**, 706–709.
- Davis, I. W., Arendall, W. B. III, Richardson, J. S. & Richardson, D. C. (2006). *Structure*, **14**, 265–274.
- Davis, I. W., Murray, L. W., Richardson, J. S. & Richardson, D. C. (2004). *Nucleic Acids Res.* **32**, W615–W619.
- Deis, L. N., Verma, V., Videau, L. L., Prisant, M. G., Moriarty, N. W., Headd, J. J., Chen, V. B., Adams, P. D., Snoeyink, J., Richardson, J. S. & Richardson, D. C. (2013). *Comput. Crystallogr. Newsl.* **4**, 9–10. [http://phenix-online.org/newsletter/CCN\\_2013\\_01.pdf](http://phenix-online.org/newsletter/CCN_2013_01.pdf).
- Dunkle, J. A., Wang, L., Feldman, M. B., Pulk, A., Chen, V. B., Kapral, G. J., Noeske, J., Richardson, J. S., Blanchard, S. C. & Cate, J. H. D. (2011). *Science*, **332**, 981–984.
- Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* **D66**, 486–501.
- Engh, R. A. & Huber, R. (1991). *Acta Cryst.* **A47**, 392–400.
- Gelbin, A., Schneider, B., Clowney, L., Hsieh, S.-H., Olson, W. K. & Berman, H. M. (1996). *J. Am. Chem. Soc.* **118**, 519–529.
- Gore, S., Velankar, S. & Kleywegt, G. J. (2012). *Acta Cryst.* **D68**, 478–483.
- Headd, J. & Richardson, J. (2013). *Comput. Crystallogr. Newsl.* **4**, 2–5. [http://phenix-online.org/newsletter/CCN\\_2013\\_01.pdf](http://phenix-online.org/newsletter/CCN_2013_01.pdf).
- Hintze, B. J., Lewis, S. M., Richardson, J. S. & Richardson, D. C. (2016). *Proteins*, **84**, 1177–1189.
- Hooft, R. W. W., Vriend, G., Sander, C. & Abola, E. E. (1996). *Nature (London)*, **381**, 272.
- Jain, S., Richardson, D. C. & Richardson, J. S. (2015). *Methods Enzymol.* **558**, 181–212.
- Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* **A47**, 110–119.
- Keedy, D. A., Georgiev, I., Triplett, E. B., Donald, B. R., Richardson, D. C. & Richardson, J. S. (2012). *PLoS Comput. Biol.* **8**, e1002629.

- Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.
- Lovell, S. C., Davis, I. W., Arendall, W. B. III, de Bakker, P. I. W., Word, J. M., Prisant, M. G., Richardson, J. S. & Richardson, D. C. (2003). *Proteins*, **50**, 437–450.
- Morris, A. L., MacArthur, M. W., Hutchinson, E. G. & Thornton, J. M. (1992). *Proteins*, **12**, 345–364.
- Murray, L. J. W., Arendall, W. B. III, Richardson, D. C. & Richardson, J. S. (2003). *Proc. Natl Acad. Sci. USA*, **100**, 13904–13909.
- Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* **D67**, 355–367.
- Ponder, J. W. & Richards, F. M. (1987). *J. Mol. Biol.* **193**, 775–791.
- Read, R. J. (1986). *Acta Cryst.* **A42**, 140–149.
- Read, R. J. (2001). *Acta Cryst.* **D57**, 1373–1382.
- Read, R. J. *et al.* (2011). *Structure*, **19**, 1395–1412.
- Richardson, J. & Headd, J. (2012). *Comput. Crystallogr. Newsl.* **3**, 29–30. [http://phenix-online.org/newsletter/CCN\\_2012\\_07.pdf](http://phenix-online.org/newsletter/CCN_2012_07.pdf).
- Richardson, J. & Prisant, M. (2012). *Comput. Crystallogr. Newsl.* **3**, 3–4. [http://phenix-online.org/newsletter/CCN\\_2012\\_01.pdf](http://phenix-online.org/newsletter/CCN_2012_01.pdf).
- Richardson, J. S., Prisant, M. G. & Richardson, D. C. (2013). *Curr. Opin. Struct. Biol.* **23**, 707–714.
- Richardson, J. S., Schneider, B., Murray, L. W., Kapral, G. J., Immormino, R. M., Headd, J. J., Richardson, D. C., Ham, D., Hershkovits, E., Williams, L. D., Keating, K. S., Pyle, A. M., Micallef, D., Westbrook, J. & Berman, H. M. (2008). *RNA*, **14**, 465–481.
- Schröder, G. F., Levitt, M. & Brunger, A. T. (2010). *Nature (London)*, **464**, 1218–1222.
- Terwilliger, T. C., Read, R. J., Adams, P. D., Brunger, A. T., Afonine, P. V., Grosse-Kunstleve, R. W. & Hung, L.-W. (2012). *Acta Cryst.* **D68**, 861–870.
- Williams, C. J., Hintze, B. J., Richardson, D. C. & Richardson, J. S. (2013). *Comput. Crystallogr. Newsl.* **4**, 33–35. [http://phenix-online.org/newsletter/CCN\\_2013\\_07.pdf](http://phenix-online.org/newsletter/CCN_2013_07.pdf).
- Williams, C. J. & Richardson, J. S. (2015). *Comput. Crystallogr. Newsl.* **6**, 2–6. [http://phenix-online.org/newsletter/CCN\\_2015\\_01.pdf](http://phenix-online.org/newsletter/CCN_2015_01.pdf).
- Word, J. M., Lovell, S. C., LaBean, T. H., Taylor, H. C., Zalis, M. E., Presley, B. K., Richardson, J. S. & Richardson, D. C. (1999). *J. Mol. Biol.* **285**, 1711–1733.
- Word, J. M., Lovell, S. C., Richardson, J. S. & Richardson, D. C. (1999). *J. Mol. Biol.* **285**, 1735–1747.
- Yeates, T. O. (1997). *Methods Enzymol.* **276**, 344–358.