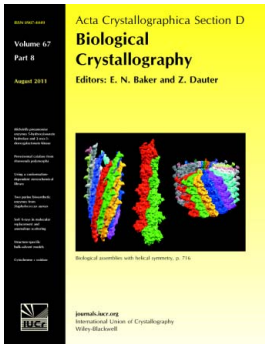
Acta Crystallographica Section D
**Biological Crystallography**
ISSN 0907-4449
Editors: **E. N. Baker and Z. Dauter**

# A conformation-dependent stereochemical library improves crystallographic refinement even at atomic resolution

**Dale E. Tronrud and P. Andrew Karplus**

*Acta Cryst.* (2011). D**67**, 699–706





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

# A conformation-dependent stereochemical library improves crystallographic refinement even at atomic resolution

**Dale E. Tronrud and P. Andrew Karplus***

Department of Biochemistry and Biophysics, Oregon State University, Corvallis, Oregon 97331, USA

Correspondence e-mail: karplusp@science.oregonstate.edu

Received 19 May 2011
Accepted 14 June 2011

To utilize a new conformation-dependent backbone-geometry library (CDL) in protein refinements at atomic resolution, a script was written that creates a restraint file for the *SHELXL* refinement program. It was found that the use of this library allows models to be created that have a substantially better fit to main-chain bond angles and lengths without degrading their fit to the X-ray data even at resolutions near 1 Å. For models at much higher resolution (∼0.7 Å), the refined model for parts adopting single well occupied positions is largely independent of the restraints used, but these structures still showed much smaller r.m.s.d. residuals when assessed with the CDL. Examination of the refinement tests across a wide resolution range from 2.4 to 0.65 Å revealed consistent behavior supporting the use of the CDL as a next-generation restraint library to improve refinement. CDL restraints can be generated using the service at http://pgd.science.oregonstate.edu/cdl_shelxl/.

## 1. Introduction

The geometric restraints used in macromolecular refinement have since the earliest restrained refinements (Dodson *et al.*, 1976; Ten Eyck *et al.*, 1976; Konnert, 1976) been defined as single fixed target values that depend only on the chemical nature of the group (Vijayan, 1976; Engh & Huber, 1991, 2001). Recently, however, we have introduced a new paradigm that treats protein geometry as ideal geometry functions that depend on conformation (Berkholz *et al.*, 2009) and thus parameterizes the restraint target values and standard deviations for protein main-chain bond lengths and angles as a function of the local $\varphi/\psi$ angles (Tronrud *et al.*, 2010).

This conformation-dependent library (CDL) was created by distilling the observed main-chain bond lengths and bond angles from a large collection of protein models determined at 1 Å resolution or better. It was found that the target values, particularly the bond-angle targets, were strong functions of the $\varphi/\psi$ angles. Explicitly modeling these variations allowed the standard deviations of the individual targets to be significantly reduced. The angle varying the most with conformation was the N—$C^{\alpha}$—C angle; its target values exhibited a spread of 6.5° and a drop in $\sigma$ from 2.8° to 1.4° (Berkholz *et al.*, 2009).

Use of the CDL concept, instead of the standard single-valued CSD-X dictionary developed by Engh & Huber (1991), has already been shown to improve modeling in a couple of applications. In a minimization routine used for homology

modeling, use of the CDL concept resulted in models closer to the known crystal structure (see Fig. 9 in Berkholz *et al.*, 2009). Also, the use of the CDL in crystallographic protein structure refinements at medium resolution (1.7 and 2.4 Å were specifically tested) resulted in models that had much lower residual differences from the library (see Fig. 1), without any degradation in *R* values (Tronrud *et al.*, 2010).

An analysis of the agreement of models in the Protein Data Bank (Berman *et al.*, 2000) with the CDL (Fig. 1) indicates that this library completely overcomes the major problem associated with the CSD-X library: that atomic resolution and ultrahigh-resolution protein models exhibit disturbingly large and increasing deviations from the library despite having been restrained to it. Indeed, at resolutions better than about 1.4 Å, even when the CSD-X restraint set was used to guide refinements the refined coordinates actually agree more closely with the CDL than they do with the CSD-X library (Fig. 1).

Nonetheless, an open question is how the CDL would impact atomic resolution refinements. Will they show similar improvements to those of lower resolution? Are they being negatively impacted by the inadequacies of single-value restraints? The tests reported in Tronrud *et al.* (2010) were performed by implementing the CDL restraints in the *TNT* refinement package (Tronrud *et al.*, 1987), but this package is not suitable for atomic resolution refinements owing to its inability to handle anisotropic temperature factors. The most widely used package for refinements at atomic resolution is *SHELXL* (Sheldrick, 2008). Fortunately, *SHELXL* is designed such that the geometric restraints are not part of the code itself but are included as part of the input along with the coordinates. Taking advantage of this organization, we here implement the CDL restraints in *SHELXL* and test the impact in refinements at resolutions between 1.3 and 0.65 Å.

## 2. Methods

The definition of target values in the *SHELXL* input file (*i.e.* the `.ins` file) is very simple and flexible (Sheldrick, 1997) as only four stereochemical restraints are defined for positional parameters: DFIX (bond length), DANG (1–3 atom distance), CHIV (chiral volume) and FLAT (planarity). The DFIX restraints match the form of the bond-length restraints in the CDL, making their conversion easy. The CDL does not currently include planarity, which means that the FLAT restraints are unchanged. More complicated is the creation of DANG and nonzero CHIV restraints, which must be mathematically derived from the bond-length and bond-angle values in the CDL.

The 1–3 atom distance is easily calculated from the 1–2 and 2–3 bond lengths and the inscribed angle using the law of cosines:

$$d_{13} = (d_{12}^2 + d_{23}^2 - 2d_{12}d_{23}\cos\theta)^{1/2}. \quad (1)$$

The nonlinearity of the law of cosines and the merging of uncertainties from three different sources, however, results in difficulties when estimating the propagation of errors. To reduce these difficulties, we made simplifying assumptions,

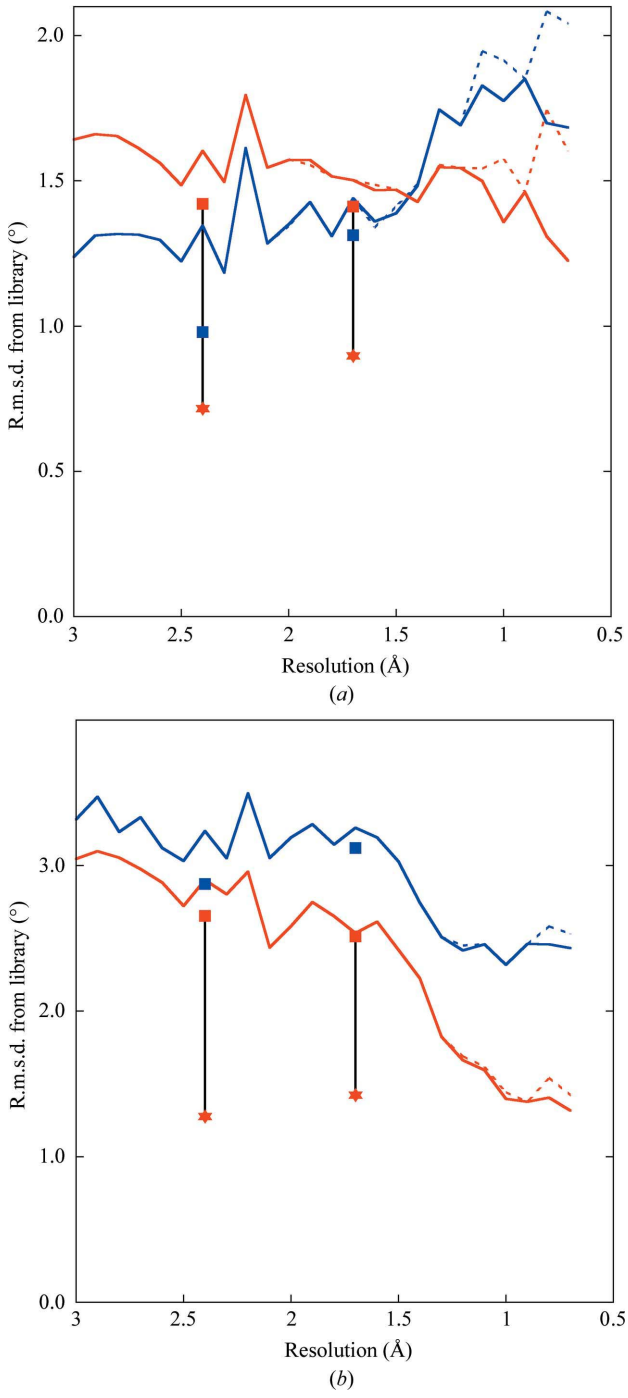



**Figure 1**
Agreement of existing models with geometry libraries and impact of refinement using the CDL at medium resolutions. For a set of 35 models selected from the PDB at each 0.1 Å bin of resolution the median r.m.s.d. value from the CSD-X library (blue lines) and the CDL (red lines) are shown. Dashed lines indicate the values when all atoms were used in the assessment and solid lines show the values leaving out atoms with alternative locations. Also shown are the results of test refinements at two resolutions reported by Tronrud *et al.* (2010). The squares indicate the r.m.s.d. agreement with the CSD-X library (blue) and with the CDL (red) when the models were restrained by the CSD-X library. The agreement with the CDL of a model restrained using the CDL is marked with a red star. Vertical black lines highlight the improvement in fit to the CDL that occurs on switching the refinement targets from the CSD-X library to the CDL. The values for the dashed lines between 3 and 1 Å resolution are from Tronrud *et al.* (2010).

most notably ignoring the contribution of the standard deviation of the two bond lengths. This will mean that the standard deviations of the 1–3 distances will be somewhat underestimated.

In our derivation of the relationship between the standard deviation of this distance [$\sigma(d_{13})$] and the standard deviation of $\theta$, we used the procedure outlined in Sivia (1996). The standard deviation of a random variable [here $\sigma(d_{13})$] is calculated from its probability distribution function by evaluating

$$\sigma(d_{13}) = \left[ -1 \Big/ \frac{\mathrm{d}^2 \log P_{d_{13}}(x)}{\mathrm{d}x^2} \right]^{1/2}, \qquad (2)$$

with $x$ equal to the most probable value for $d_{13}$. For uncertainties of the level expected for $\theta$ (*i.e.* <3°), test calculations show that the most probable value of $d_{13}$ is well estimated by (2). With the assumption that the uncertainty in $\theta$ is normally distributed and that all the uncertainty in $d_{13}$ originates in that angle, the probability distribution for $d_{13}$ can be calculated as

$$P_{d_{13}}(d_{13}) = P_\theta[\theta(d_{12}, d_{23}, d_{13})] \left| \frac{\mathrm{d}\theta(d_{12}, d_{23}, d_{13})}{\mathrm{d}d_{13}} \right| \qquad (3)$$

$$= P_\theta \left[ \cos^{-1} \left( \frac{d_{12}^2 + d_{23}^2 - d_{13}^2}{2 d_{12} d_{23}} \right) \right] \times \frac{d_{13}}{d_{12} d_{23} \left[ 1 - \frac{(d_{12}^2 + d_{23}^2 - d_{13}^2)^2}{4 d_{12}^2 d_{23}^2} \right]^{1/2}}. \qquad (4)$$

The derivation of $\sigma(d_{13})$ by substituting (4) into (2) results in an equation with two terms in its denominator. One term is a significant fraction of the other only when $\theta$ is smaller than 20° or larger than 160°. Since bond angles in these ranges do not occur in the backbone of proteins, this term was ignored. Further simplification leads to the equation used:

$$\sigma(d_{13}) = \sigma(\theta)[(-d_{12} + d_{13} + d_{23})(d_{12} + d_{13} - d_{23}) \times (d_{12} - d_{13} + d_{23})(d_{12} + d_{13} + d_{23})]^{1/2} / 2 d_{13}. \qquad (5)$$

The target value for the chiral volume of the $C^\alpha$ atom was calculated as

$$V_{C^\alpha} = d_{C^\alpha N} d_{C^\alpha C} d_{C^\alpha C^\beta} [1 - \cos\theta_{NC^\alpha C} - \cos\theta_{NC^\alpha C^\beta} - \cos\theta_{C^\beta C^\alpha C} + 2\cos\theta_{NC^\alpha C}\cos\theta_{NC^\alpha C^\beta}\cos\theta_{C^\beta C^\alpha C}]. \qquad (6)$$

For $\sigma(V_{C^\alpha})$ we used the *SHELXL* default value for nonzero chiral volumes since a proper calculation would require the contributions of uncertainties of the three bond lengths and three bond angles involved, making its calculation far more complex than that of $\sigma(d_{13})$.

All restraints for atoms other than protein main chain are unchanged, being restrained to CSD-X target values. In addition, if either $\varphi$ or $\psi$ cannot be calculated, for instance owing to the residue being at a chain terminus or lacking certain atoms, CSD-X target values are used.

As described in Tronrud *et al.* (2010), the CDL is very detailed, with distinct backbone restraint sets depending on the residue type and its $\varphi/\psi$ values. To produce this customized information we created a Python script which reads a PDB-format file and writes the specific target values and standard deviations based on CDL v.1.2 in the appropriate format. The content of this file is then pasted into the *SHELXL* `.ins` file to replace the CSD-X restraints placed there by *SHELXPRO* (Fig. 2).

In our previous tests at medium resolution using the *TNT* refinement package (Tronrud *et al.*, 2010), we varied the weights that balance the X-ray data and the CDL geometric restraints, but found that use of the CDL required no change in weighting compared with use of the CSD-X library. In accord with this result, the refinement tests here were performed using the weighting scheme recommended in the *SHELX* manual (Sheldrick, 1997): each restraint was given a standard deviation based on the values in the CDL and the value on the WGHT card was set to 0.1 unless a previous run of *SHELXL* had recommended a different value. For each test case a sufficient number of cycles of conjugate-gradient least-squares refinement was performed to achieve convergence.

The script that can be used to prepare a *SHELXL* refinement using the CDL, `cdl_shelxl.py`, can be accessed online at http://pgd.science.oregonstate.edu/cdl_shelxl/ or downloaded from http://sourceforge.net/projects/proteingeometry/. The CDL v.1.2 can be downloaded from http://dunbrack.fccc.edu/nmhrcm/.

## 3. Results

### 3.1. Test cases and assessment strategy

For our test refinements we chose six models that had previously been refined with *SHELXL* at resolutions from 1.3 to 0.65 Å. They are maize root ferredoxin–NADP$^+$ reductase at 1.05 Å resolution (rFNR; PDB entry 3lo8), which was used in our lower resolution test refinements (Tronrud *et al.*, 2010);

```
                                        REM   ALA VAL Phi = -91 Psi = 106
DFIX_* 1.329 C_- N                      DFIX_116 1.331 0.014 C_- N
DANG_* 2.425 CA_- N                     DANG_116 2.428 0.017 CA N_+
DANG_* 2.250 O_- N                      DANG_116 2.253 0.013 O N_+
DANG_* 2.435 C_- CA                     DANG_116 2.446 0.016 C_- CA
FLAT_* 0.5 O_- CA_- N C_- CA            FLAT_116 0.5 O CA N_+ C CA_+

CHIV_ALA C                              CHIV_116 C
CHIV_ALA 2.477 CA                       CHIV_116 2.581 CA
DFIX_ALA 1.231 C O                      DFIX_116 1.235 0.012 C O
DFIX_ALA 1.525 C CA                     DFIX_116 1.525 0.013 C CA
DFIX_ALA 1.521 CA CB                    DFIX_116 1.527 0.013 CA CB
DFIX_ALA 1.458 N CA                     DFIX_116 1.459 0.012 N CA
DANG_ALA 2.462 C N                      DANG_116 2.421 0.025 C N
DANG_ALA 2.401 O CA                     DANG_116 2.402 0.013 O CA
DANG_ALA 2.503 C CB                     DANG_116 2.512 0.025 C CB
DANG_ALA 2.446 CB N                     DANG_116 2.449 0.024 CB N
```

**Figure 2**
Comparison of input for *SHELXL* using the CSD-X library and the CDL. On the left are the definitions for residue 116 of rFNR used in a CSD-X-based *SHELXL* refinement. Note that these target values are defined solely in terms of the amino-acid type (*e.g.* all alanine residues have the same values). On the right are the target values from the CDL for the same residue. In this case the target values for residue 116 are specific and based on its $\varphi/\psi$ angles and residue type. The atoms listed in the peptide restraints differ because the CDL implementation groups a residue with the following peptide bond, whereas the *SHELXL* CSD-X implementation groups a residue with the preceding peptide bond.

**Table 1**
Test cases and the agreement of the resulting models with the X-ray data and the library used.

| Data-set name | Resolution (Å) | Published $R/R_{free}$ (%) | Target library | Refined $R/R_{free}$ (%) | R.m.s. shift† (Å) | MC bonds‡ (Å) | MC angles‡ (°) | N—$C^{\alpha}$—C angles‡ (°) |
|---|---|---|---|---|---|---|---|---|
| aFMO | 1.30 | 13.5/16.1 | CSD-X | 13.53/16.13 | | 0.0124 | 1.92 | 2.63 |
| | | | CDL | 13.59/16.10 | 0.017 | 0.0069 | 1.12 | 1.45 |
| T4Lmut | 1.20 | 15.1/17.4 | CSD-X | 15.14/17.51 | | 0.0120 | 1.86 | 1.93 |
| | | | CDL | 15.18/17.42 | 0.021 | 0.0072 | 1.07 | 1.09 |
| rFNR | 1.05 | 12.5/15.5 | CSD-X | 12.49/15.52 | | 0.0140 | 1.95 | 2.33 |
| | | | CDL | 12.54/15.51 | 0.012 | 0.0086 | 1.16 | 1.16 |
| hGR | 0.95 | 12.3/15.2 | CSD-X | 12.30/15.41 | | 0.0164 | 1.80 | 2.34 |
| | | | CDL | 12.33/15.12 | 0.007 | 0.0110 | 1.20 | 1.20 |
| PDZ2 | 0.73 | 7.4§/8.7 | CSD-X | 7.70/8.59 | | 0.0102 | 1.49 | 2.35 |
| | | | CDL | 7.68/8.60 | 0.001 | 0.0078 | 0.97 | 0.98 |
| HEWL | 0.65 | 8.48/9.52 | CSD-X | 8.44¶ | | 0.0113 | 1.43 | 2.18 |
| | | | CDL | 8.46¶ | 0.002 | 0.0087 | 1.01 | 1.12 |

† The r.m.s. shift of the main-chain atoms between the models created by refining against each library. ‡ R.m.s. deviations from restraints; MC means main chain. § The *R* value is calculated over all data in both the working and test sets. ¶ Test-set flags were not deposited.

human glutathione reductase at 0.95 Å resolution (hGR; PDB entry 3dk9; Berkholz *et al.*, 2008); T4 lysozyme mutant D72A/R96H at 1.2 Å resolution (T4Lmut; PDB entry 3fad; Mooers *et al.*, 2009); the FMO protein from *Prosthecochloris aestuarii* 2K at 1.3 Å resolution (aFMO; PDB entry 3eoj; Tronrud *et al.*, 2009); the PDZ2 domain of syntenin at 0.73 Å resolution (PDZ2; PDB entry 1r6j; Kang *et al.*, 2004) and hen egg-white lysozyme at 0.65 Å resolution (HEWL; PDB entry 2vb1; Wang *et al.*, 2007). The latter two models had been refined with no stereochemical restraints on their ordered portions, giving a view of protein structure unbiased by either library. Of these models, only the PDZ2 domain was used in the creation of the CDL. We chose to use this structure anyway because it is the protein structure refined to the lowest $R_{free}$ value (Kang *et al.*, 2004), making it an excellent benchmark structure.

To ensure that these structures are suitable representatives of structures in general, we extended to beyond 1 Å resolution our previous analysis (Tronrud *et al.*, 2010) of how the root-mean-square deviations (r.m.s.d.) of models based on X-ray data of varying resolution compared with either the CSD-X library or the CDL. This extension yielded the surprising result that the highest resolution structures (near 0.7 Å resolution) had increased deviation from the CDL (Fig. 1*a*, dotted line). By examining the HEWL test case we traced the source of this behavior to the stretches of polypeptide where the main chain adopts two conformations. [This problem had been noted by Jaskolski *et al.* (2007) in their analysis of the HEWL model.] The fit of these regions to either library is far worse than the fit of the full occupancy regions. For example, for HEWL the backbone-angle r.m.s.d. from the CSD-X library for atoms with alternative locations is 3.75° *versus* 1.44° for the ordered atoms alone. Among residues with multiple conformations, those with low occupancy behave worst, with the '*A*' conformers (average occupancy 0.7) having an r.m.s.d. of 2.52° and the '*B*' conformers (average occupancy 0.3) having an r.m.s.d. of 4.66°.

Even at atomic resolution, poorly occupied atoms have weak density and their locations are not reliably determined. So that these more poorly defined atoms would not skew our results, we repeated our previous survey with all multiply modeled atoms removed (Fig. 1, continuous lines). Since alternative conformations for main-chain atoms are rare in lower resolution models, those r.m.s.d. values change little. At the highest resolutions, the removal of the multiply modeled atoms leads to a decrease in the deviations from both the CSD-X library and the CDL. All remaining comparisons reported in this paper are carried out using only the atoms modeled in single locations.

### 3.2. Test refinements with *SHELXL*

The results of the *SHELXL* refinements for the six test cases are reported in Table 1 and Fig. 3. For each test case, we carried out independent refinements using the CSD-X and the CDL restraints and in all cases except the PDZ2 domain (7.7% *versus* 7.4%) the *R* factors obtained were within 0.2% of those published (Table 1). Upon refinement against the CDL, the *R* factors changed very little, with the working *R* values moving slightly higher and the free *R* values slightly lower. Comparing the CDL refined structures with the CSD-X refined structures, the root-mean-square (r.m.s.) shifts for backbone atoms vary with resolution, being near 0.02 Å for the refinements at ∼1.2–1.3 Å, near 0.01 Å for refinements at ∼1 Å and near 0.002 Å for refinements at ∼0.7 Å resolution (Table 1).

In each case, refinement against the CDL resulted in a 40–50% improvement in the fit to the bond-angle targets (Table 1). Fig. 3 illustrates these numbers and also shows a second way to quantify the improvements: comparing how well the two models agree with the CDL targets (comparing the red squares with the red stars). For backbone bond angles this comparison shows smaller but still substantial (∼20%) improvements for the 1.3–0.95 Å resolution test cases and little to no change for the two test cases at better than ∼0.8 Å resolution. For the N—$C^{\alpha}$—C bond angles minimal improvements occur for all test cases (Fig. 3*b*).

Bond-length r.m.s.d.s dropped substantially for all six test cases with the use of the CDL target values (Table 1). The test

cases near 1 Å resolution showed an ∼40% drop in the r.m.s.d. (from 0.0140 to 0.0086 Å for the rFNR test case, for example) and even the test cases near 0.7 Å resolution showed an ∼25% or an ∼10% drop depending on which measure of improvement was used. These are surprising large drops since the survey of models in the PDB showed only small differences in deviations from the CSD-X library *versus* the CDL targets (Fig. 3*c*, blue and red lines).

## 4. Discussion

Building on the previously reported medium-resolution test refinements (Tronrud *et al.*, 2010), the atomic resolution refinements reported here strengthen the conclusion that the CDL (Berkholz *et al.*, 2009) in specific, and the paradigm of context-dependent ideal geometry functions in general, represents a substantial step forward in the ability to accurately refine and describe protein structures.

The survey of the agreement of existing models with either geometry library (Fig. 1, blue and red dashed lines) showed that the observations of Jaskolski *et al.* (2007) in relation to the poor geometry of HEWL regions modeled as alternate conformations apply more generally. While they recommended that regions with multiple conformations or high temperature factors be more tightly restrained to the target values, Tickle (2007) disagreed. He noted that the time-averaged positions of atoms experiencing large-amplitude motions cannot be expected to match geometry libraries which model motionless structures. This argument, however, does not apply to models with multiple but well ordered conformations, *i.e.* those with low mobility in addition to their low occupancy. In the future such regions, at least, should be more tightly restrained.

Also, at the highest resolutions, when it is possible to release restraints completely for well ordered fully occupied atoms, it

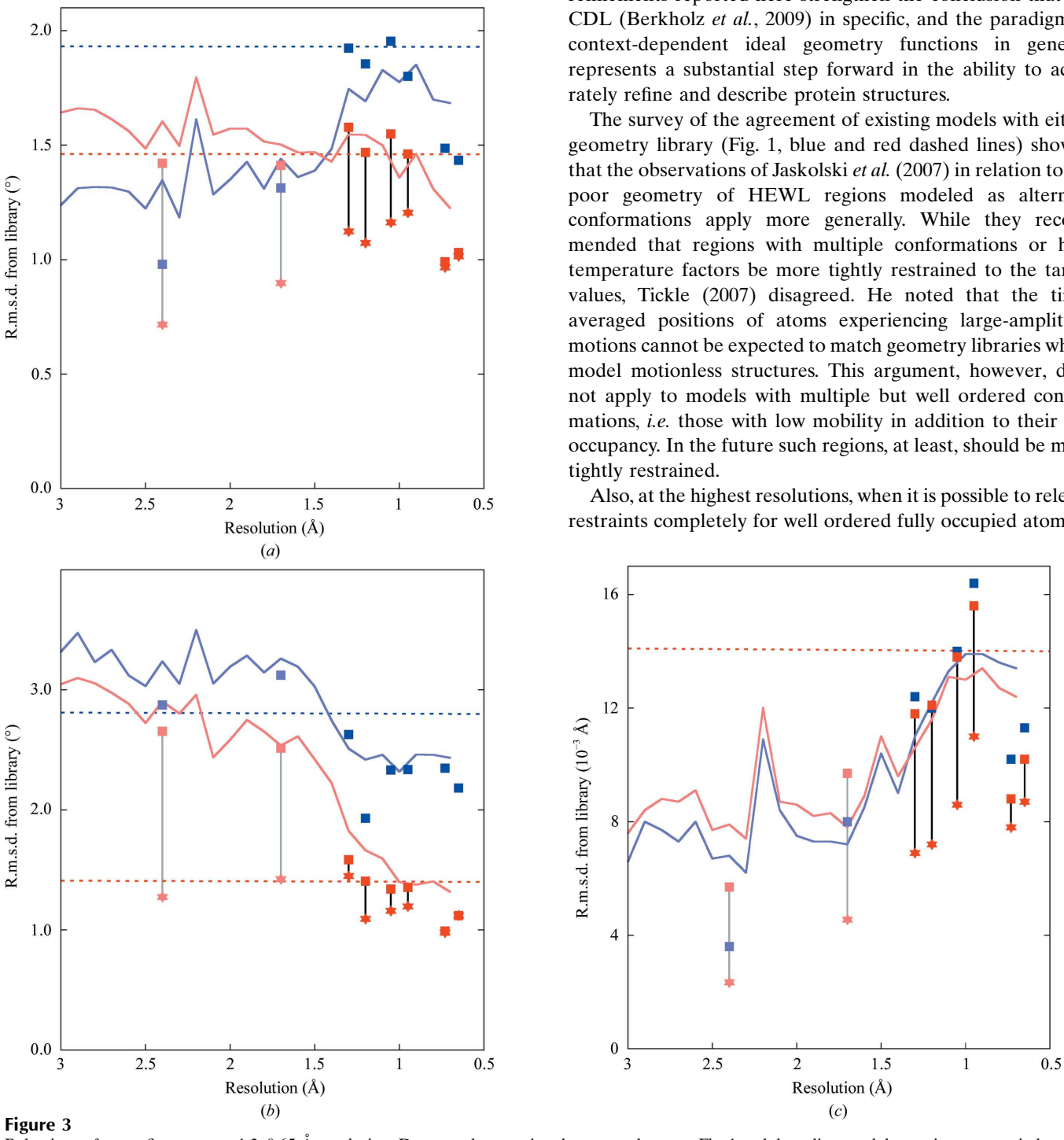



**Figure 3**
Behaviour of test refinements at 1.3–0.65 Å resolution. Data are shown using the same scheme as Fig. 1 and these lines and data points are carried over to provide context. (*a*) Backbone bond-angle r.m.s.d.s from ideality are shown for each of the six *SHELXL* test cases restrained by either the CSD-X library (blue and red squares) or the CDL (red stars). The horizontal dashed lines indicate the average standard deviation of the target values for each library (CSD-X, blue; CDL, red). (*b*) Same as (*a*) except showing N—$C^{\alpha}$—C bond-angle deviations from ideality. (*c*) Same as (*a*) except showing backbone bond-length deviations. The average standard deviation for these bond lengths in the CSD-X library is 0.019 Å and falls outside the limits of this plot.

is essential that the restraints be maintained for residues with alternate conformations. Such a carefully considered restraint scheme was actually used in the original refinements of both the HEWL (Jaskolski *et al.*, 2007) and PDZ2 (Kang *et al.*, 2004) test cases; however, the weights used were apparently insufficient to bring the sections with alternative conformations up to a conventional level of agreement with the library target values.

At the ultrahigh-resolution end of the spectrum, the behavior of the HEWL (at 0.65 Å resolution) and PDZ2 (at 0.73 Å resolution) test cases provide very valuable points of reference. Both of the deposited models had been refined with no stereochemical restraints on the ordered parts of the proteins to very low $R$ factors and so provide examples of unbiased bond lengths and angles. In fact, refining these two models against either the CSD-X library or the CDL resulted in shifts of only 0.001–0.002 Å in the main chain and no noteworthy change in their fit to the CDL bond-angle targets (as shown by the close overlap of the red squares and stars in Figs. 3*a* and 3*b*). It appears that the default weight in *SHELXL* has the same result as a zero weight for bond angles at these resolutions.

In contrast, the bond lengths of these two models improve when refined with either set of target values, with the greater improvement resulting from the CDL. The r.m.s.d. from the CSD-X bond-length targets for the deposited PDZ2 and HEWL models is 0.0104 and 0.0125 Å, respectively. In contrast to the bond-angle results, refining against the CDL does actually improve the fit of the bond lengths to the CDL by an average of about 10% (Table 1) without a worsening of the $R$ factors, showing that restraining the bond lengths in these refinements has a positive impact.

The consistent and robust behavior of these two structures allows us to consider them as standards that provide unbiased estimates of the level of variability of bond lengths and angles in proteins compared with current CDL targets. The observation that the levels of deviation from the CDL are very similar to each other supports their use in this manner and also allays concerns that the level of agreement of PDZ2 with the CDL has been biased by its inclusion in the creation of the CDL.

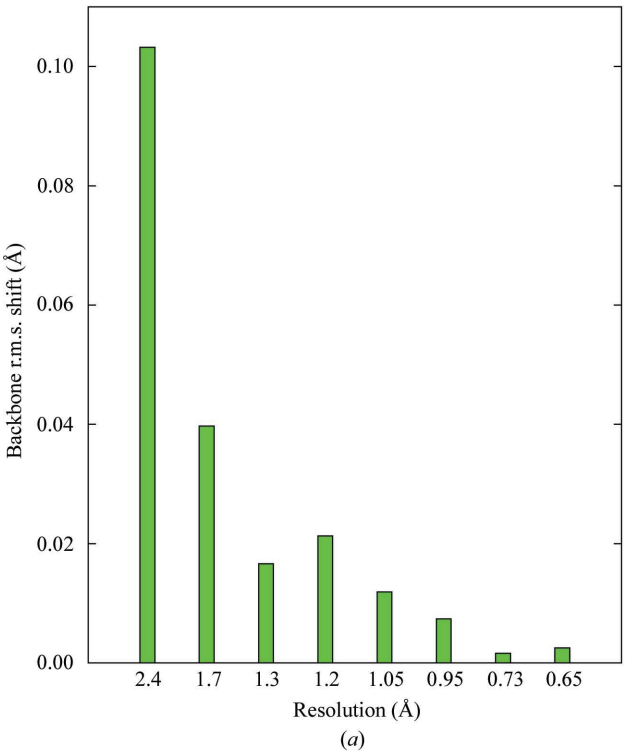
Taking these structures as standards leads to some concrete conclusions. Firstly, the CDL is more accurate than the CSD-X library as it matches these models with backbone bond-angle deviations at $\sim1^\circ$ *versus* $\sim1.5^\circ$ and N$-$C$^\alpha-$C angle deviations at $\sim1^\circ$ *versus* $\sim2.2^\circ$ and bond-length deviations about 0.001 Å less (with the fit to the CDL dropping further after refinement). Secondly, the r.m.s.d.s to the CDL just noted are not simply the r.m.s. deviations of these two structures, but can be taken as estimates of the standard deviations that are appropriate to guide weighting in the application of the library. According to the principle that modeling accuracy is enhanced by allowing less variation for lower resolution structures (Tickle, 2007), these numbers are maximal deviations that should be allowed for any structure, *i.e.* for structures at lower resolutions the targeted deviations from the library should be smaller.

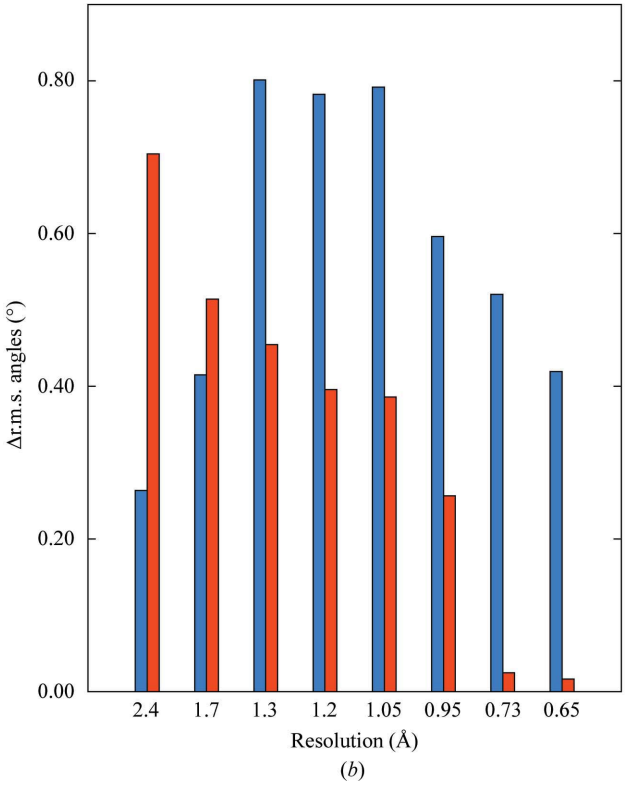
With this knowledge that the CDL is more accurate than the CSD-X library, we can now conclude that (considering just the atoms modeled in single positions) the results as a whole prove that the models refined at between 1.3 and 0.95 Å resolution benefit substantially by refinement against the CDL. The ∼40% improvements in backbone bond-length and bond-angle r.m.s.d.s seen on comparing the CSD-X r.m.s.d.s with the CDL r.m.s.d.s overestimate the true improvement because they are heavily influenced by the lower accuracy of the CSD-X target values. This is proven by considering the PDZ2 case, for which the atoms move only 0.001 Å and the agreement with the CDL library does not improve at all, yet a comparison of the backbone bond-angle CSD-X residual with that of the CDL gives rise to an apparent ∼30% improvement (1.49° to 0.97°). Thus, a much better assessment of true model improvement comes from comparing deviations from the CDL before and after refinement using CDL targets.

This comparison shows an improvement of about 25% for backbone bond angles, as the 0.95–1.3 Å resolution test cases restrained to the CSD-X library fit the CDL targets with an $\sim1.5^\circ$ r.m.s.d. but improve to $\sim1.1^\circ$ when restrained to the CDL (Fig. 3*a*). This improvement reveals potential for iteratively improving the CDL because CDL target values, which are derived mainly from models in the 1.0–0.9 Å resolution range and restrained using CSD-X target values, have standard deviations (dashed red line in Fig. 3) that match the median r.m.s.d. from the CDL of the models in this resolution range (the value of the solid red line near 0.9 Å resolution). After refinement using the CDL targets the bond angles in these structures become more tightly grouped around the targets, while the agreement with the diffraction data as measured by $R_{\rm free}$ is maintained or even very slightly improved. This implies that an updated revision of the CDL based on rerefined structures will have smaller standard deviations (near 1.1°) for the backbone angles and could in turn lead to still further improvements in performance. It also implies that additional analyses of refined models may reveal new and unexpected patterns of variability.

In general, it is interesting that these striking improvements in fit to the ideal geometry target values are achieved with remarkably small changes in the positions of the main-chain atoms. Apparently, the usage of the CDL targets results in small but coordinated shifts of the atoms to produce significant improvements in the model geometry. We note that even though only the main-chain atoms experienced changed restraints, accompanying the small shifts in backbone atoms many other atoms have shifts three to ten times larger despite experiencing exactly the same target values.

With regard to main-chain bond lengths, models refined against the CSD-X targets fit similarly to the CDL and the CSD-X libraries (blue and red squares in Fig. 3*c*). From this observation we infer, consistent with the expectations of Berkholz *et al.* (2009), that the two sets of bond-length target values are rather similar. Refinement against the CDL nevertheless resulted in a drop of about 40% in bond-length r.m.s.d. from the CDL (Table 1). Because the libraries are not very distinct, we suspected that the improvement was largely a




consequence of the smaller standard deviations of the CDL targets: ∼0.014 Å (red dotted line in Fig. 3*c*) compared with the 0.02 Å standard deviation used by *SHELXL* for all CSD-X bond-length targets. This hypothesis was confirmed both by noting that the models refined with the CDL targets show a similar improved agreement with the CSD-X targets and by showing that test refinements using the CSD-X bond-length targets but with artificially increased weighting can produce models with bond-length r.m.s.d.s that are similarly low to those produced by CDL refinement.

As noted above, the improvement in refinement brought by use of the CDL can be measured in two ways: comparing the quality of models refined with CSD-X and assessed using the CSD-X library to models refined with the CDL and assessed using the CDL or assessing both models with the CDL. The former measures the apparent improvement, while the latter measures the actual improvement. Stepping back to view the behavior of refinements across the whole resolution range from 2.4 to 0.65 Å, we can see to what extent the improved performance of the CDL arises from each of the criteria. The balance between them depends on resolution in an interesting manner (Fig. 4).

Firstly, we note that the backbone coordinate shifts arising from using the CDL (*versus* CSD-X) targets decrease progressively from 0.10 to 0.001 Å as the resolution improves from 2.4 to ∼0.7 Å (Fig. 4*a*). Associated with these shifts is a

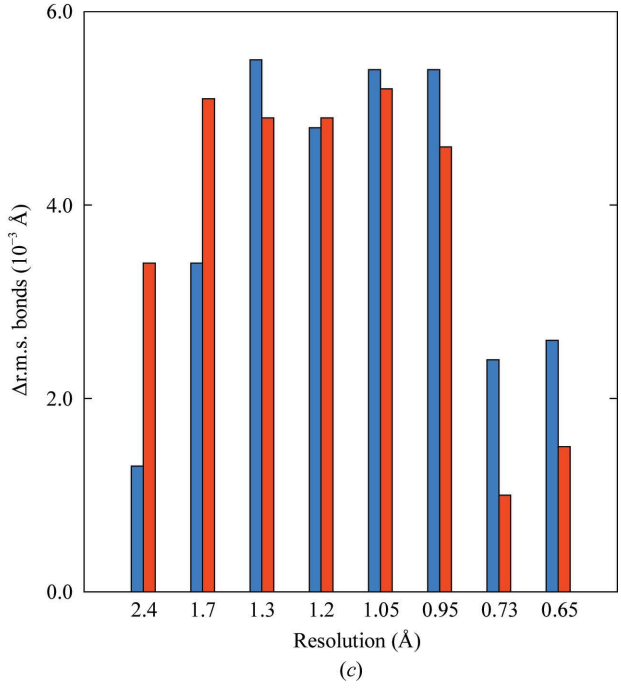



**Figure 4**
Changes in refinement behavior as a function of resolution. For each test case between 2.5 and 0.65 Å resolution a parameter resulting from refinement against the CDL is contrasted with that resulting from refinement against the CSD-X library. (*a*) The r.m.s. positional shift of the main-chain atoms between the models refined using CSD-X targets and those refined using the CDL. Main-chain atoms with alternative locations are ignored. (*b*) Improvements in main-chain bond-angle r.m.s.d. for each test case are shown. The blue bars show the decrease in r.m.s.d. when a model refined and assessed using the CSD-X library is compared with a model refined and assessed using the CDL. The red bars show the decrease when a model refined using the CSD-X library but assessed using the CDL is compared with one refined and assessed using the CDL. (*c*) Improvements in main-chain bond-length r.m.s.d. for each test case is shown. The color-coding is the same as in (*b*).

real improvement in bond-angle ideality (based on deviation from the CDL values), which also decreases at higher resolutions from a differential of ∼0.7° to near-zero (red bars in Fig. 4*b*). The ability of the CDL to 'truly' improve models decreases to near zero, because at near 0.7 Å resolution, but not yet at ∼1 Å resolution, the X-ray data are apparently able to determine them rather well regardless of the restraint library used.

In contrast, the apparent improvement in bond angles (blue bars) is small at low resolution and increases at higher resolutions. The low apparent improvement owing to the CDL at the lower resolutions occurs because the restraints of the library dominate over the limited diffraction data and these models show good agreement with whatever target library is used, CDL or CSD-X, although the more precise target values of the CDL lead to the model being more tightly restrained. One would expect the apparent improvement (blue bars) to increase as resolution improves, being maximal at the highest resolution, but this is not the case. Instead, it rises until about 1.4 Å and then decreases again. We suggest that this occurs because the standard deviations associated with the CSD-X target values for bond angles overestimate the true range of bond-angle variation, so at the resolutions in the 1.4–1.0 Å range the models end up deviating from the restraints much more than is the case for the unrestrained models at 0.7 Å resolution.

The equivalent analysis for measures of the improved agreement with bond-length restraints shows similar patterns for the apparent improvement (blue bars in Fig. 4*c*), implying that in the CSD-X library the bond-length restraints also have standard deviations that are larger than the true level of the variation and so end up allowing too much freedom in refinements at near-atomic resolution. In this case, however, because the target libraries are rather similar the real improvements (red bars in Fig. 4*c*) for the most part match the apparent improvements, as would be expected for improvements that are in a large part the consequence of the tighter intrinsic nature of the CDL restraints (*i.e.* their lower standard deviations). It is worth noting though that because for the two test cases at 0.7 Å resolution the real improvement is less than the apparent improvement (*i.e.* the blue bars exceed the red bars), we can conclude that the CDL bond-length targets themselves are also somewhat more accurate (*i.e.* in better agreement with the unrestrained structures). Overall, the use of the CDL bond-length restraints provides substantial improvement in models at all resolutions worse than ∼0.9 Å and even provides some improvement at the highest resolutions.

These tests of the CDL using structures near and beyond atomic resolution extends our earlier results to show that, compared with the CSD-X library the CDL substantially improves refinement behavior at all resolutions. The tests have further allowed us to show that the improved behavior arises from both the enhanced information content (*i.e.* the greater accuracy) of the conformation-dependent target values and also the greater precision (*i.e.* lower sigmas) associated with these targets. The behavior of the two reference structures near 0.7 Å resolution, with r.m.s.d.s of ∼0.008 Å for bond lengths and ∼1.0° for bond angles, provides upper-limit residuals that can guide future weighting of restraints when using this CDL at any resolution. With an ever-increasing number of models in the Protein Data Bank and the potential for the CDL to improve the quality of existing models, we expect updated versions of the CDL will allow even lower r.m.s.d.s and will also be extended to include side-chain geometry and additional contextual aspects such as the secondary structure as defined by hydrogen bonding (Touw & Vriend, 2010). As such, to allow refinement results to be considered with a proper perspective, reporting the particular version of the CDL used in a structure determination will also be required, with the version used here being designated CDL v.1.2.

This work was supported by National Institutes of Health (NIH) grant R01-GM083136. We would also like to thank Professor Blaine Mooers and Professor Beom Sik Kang for supplying their *SHELXL* input files for two of the test cases.

## References

Berkholz, D. S., Faber, H. R., Savvides, S. & Karplus, P. A. (2008). *J. Mol. Biol.* **382**, 371–384.
Berkholz, D. S., Shapovalov, M. V., Dunbrack, R. L. J. & Karplus, P. A. (2009). *Structure*, **17**, 1316–1325.
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. (2000). *Nucleic Acids Res.* **28**, 235–242.
Dodson, E. J., Isaacs, N. W. & Rollett, J. S. (1976). *Acta Cryst.* A**32**, 311–315.
Engh, R. A. & Huber, R. (1991). *Acta Cryst.* A**47**, 392–400.
Engh, R. A. & Huber, R. (2001). *International Tables for Crystallography*, Vol. *F*, edited by M. G. Rossmann & E. Arnold, pp. 382–392. Dordrecht: Kluwer Academic Publishers.
Jaskolski, M., Gilski, M., Dauter, Z. & Wlodawer, A. (2007). *Acta Cryst.* D**63**, 611–620.
Kang, B., Devedjiev, Y., Derewenda, U. & Derewenda, Z. (2004). *J. Mol. Biol.* **338**, 483–493.
Konnert, J. H. (1976). *Acta Cryst.* A**32**, 614–617.
Mooers, B. H., Tronrud, D. E. & Matthews, B. W. (2009). *Protein Sci.* **18**, 863–870.
Sheldrick, G. M. (1997). *The SHELX-97 Manual*. University of Göttingen, Germany.
Sheldrick, G. M. (2008). *Acta Cryst.* A**64**, 112–122.
Sivia, D. S. (1996). *Data Analysis: A Bayesian Tutorial*, pp. 77–80. Oxford University Press.
Ten Eyck, L. F., Weaver, L. H. & Matthews, B. W. (1976). *Acta Cryst.* A**32**, 349–350.
Tickle, I. J. (2007). *Acta Cryst.* D**63**, 1274–1281.
Touw, W. G. & Vriend, G. (2010). *Acta Cryst.* D**66**, 1341–1350.
Tronrud, D. E., Berkholz, D. S. & Karplus, P. A. (2010). *Acta Cryst.* D**66**, 834–842.
Tronrud, D. E., Ten Eyck, L. F. & Matthews, B. W. (1987). *Acta Cryst.* A**43**, 489–501.
Tronrud, D. E., Wen, J., Gay, L. & Blankenship, R. E. (2009). *Photosynth. Res.* **100**, 79–87.
Vijayan, M. (1976). *CRC Handbook of Biochemistry and Molecular Biology*, 3rd ed., *Proteins*, Vol. III, edited by G. D. Fasman, pp. 742–759. Cleveland: CRC Press.
Wang, J., Dauter, M., Alkire, R., Joachimiak, A. & Dauter, Z. (2007). *Acta Cryst.* D**63**, 1254–1268.