Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Using a conformation-dependent stereochemical library improves crystallographic refinement of proteins

Dale E. Tronrud, Donald S. Berkholz‡ and P. Andrew Karplus*

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

‡ Current address: Department of Physiology and Biomedical Engineering and Department of Pediatric and Adolescent Medicine, Mayo Clinic College of Medicine, Rochester, Minnesota 55905, USA.

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

Received 9 March 2010
Accepted 21 May 2010

PDB References: ferredoxin:NADP$^+$ reductase at 100 K, 3lo8; at 295 K refined against all data, 1jb9; at 295 K refined withholding test set, 3lvb.

The major macromolecular crystallographic refinement packages restrain models to ideal geometry targets defined as single values that are independent of molecular conformation. However, ultrahigh-resolution X-ray models of proteins are not consistent with this concept of ideality and have been used to develop a library of ideal main-chain bond lengths and angles that are parameterized by the φ/ψ angle of the residue [Berkholz *et al.* (2009), *Structure*, **17**, 1316–1325]. Here, it is first shown that the new conformation-dependent library does not suffer from poor agreement with ultrahigh-resolution structures, whereas current libraries have this problem. Using the *TNT* refinement package, it is then shown that protein structure refinement using this conformation-dependent library results in models that have much better agreement with library values of bond angles with little change in the *R* values. These tests support the value of revising refinement software to account for this new paradigm.

1. Introduction

Traditional and current stereochemical libraries used in crystallographic refinement of proteins (Diamond, 1971; Vijayan, 1976; Hendrickson & Konnert, 1980; Tronrud *et al.*, 1987; Engh & Huber, 1991, 2001) have single fixed target values for bond lengths and angles that are independent of φ/ψ and other freely rotatable torsion angles. For protein structure refinement, the most recent major improvement in restraints was the introduction by Engh & Huber (1991) of a carefully selected restraint set, called the CSD-X library, based on the small-molecule structures from the Cambridge Structural Database (Allen, 2002). The replacement of the then commonly used *X-PLOR* param19x restraints (Brünger *et al.*, 1989) with the CSD-X library yielded little change in *R* values but a roughly 10% improvement in the agreement of structures with the restraints (Engh & Huber, 1991).

A recent series of *Letters to the Editor* in this journal has discussed the usage and limitations of such restraint libraries (Jaskolski *et al.*, 2007*a*,*b*; Stec, 2007; Tickle, 2007; Karplus *et al.*, 2008), focusing on the conflict between weighting the stereochemical restraints loosely enough to allow the model to exhibit real deviations from library values, yet weighting tightly enough to prevent nonsensical deviations. The key observation sparking concern was that compared with low-resolution structures, ultrahigh-resolution structures deviate more from the ideal library (Jaskolski *et al.*, 2007*b*). Karplus *et al.* (2008) suggested that the conflict largely occurs because the current restraint libraries do not account for real systematic variations in geometry that occur as a function of conformation. They further proposed that the dilemma can be avoided by creating a new kind of restraint library which accounts for the systematic conformation-dependent variation in the 'ideal geometry'.

Recently, Berkholz *et al.* (2009) have created such a conformation-dependent library (CDL) for the protein backbone. They used a large collection of ultrahigh-resolution protein models (≤1 Å resolution) to deduce target values and standard deviations for the main-chain bond lengths and angles as a function of the φ/ψ angles of the residue in question. They concluded that the bond-angle variations seen were well determined at these resolutions but that the bond-length variations seen would need higher resolution structures to be determined accurately; even then, Berkholz and coworkers expected that any trends in bond lengths would involve such small differences that the variations would have little impact on coordinate accuracy. For the bond angles, the target values were found to vary smoothly with conformation over a wide range for some angles (*e.g.* the $\mathrm{N-C^{\alpha}-C}$ angle having a spread of 6.5°) and to vary in ways consistent with the underlying nonbonded interactions that create the excluded zones in a Ramachandran plot (Ramachandran *et al.*, 1963).

The question then became how the use of the CDL would impact crystallographic refinement (Dauter & Wlodawer, 2009). Answering this question is nontrivial because current crystallographic refinement programs have been constructed using the paradigm of 'single ideal value' restraints and would require substantial modification to use the new CDL paradigm. Here, we assess the value of switching to the new paradigm by carrying out a series of test refinements using the *TNT* refinement package (Tronrud *et al.*, 1987; Tronrud, 1997), which has a flexible design that allows it to be used in the development of novel refinement techniques without modifying its source code (*e.g.* Chapman, 1995; Bricogne & Irwin, 1996). The tests show that the improvements brought by using the CDL in place of the CSD-X library are even larger than those that accompanied the adoption of the CSD-X restraint set 20 years ago.

2. Methods

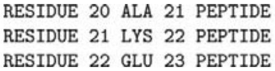
2.1. Implementation of the conformation-dependent library (CDL) in *TNT*

The CDL (Berkholz *et al.*, 2009) is a library of φ/ψ-dependent standard values for the bond lengths and angles of the protein backbone. Compared with the most widely used single-value library derived from the Cambridge Structural Database (Allen, 2002) (CSD-X; Engh & Huber, 1991), the target values for angles vary by as much as 3.5° from the single ideal values. In addition, the library is more precise in that the standard deviations associated with the target values in the CDL are generally smaller.

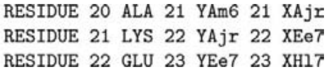
The particular CDL of Berkholz *et al.* (2009) parameterized five classes of residues each having a substantial number (>500) of observations. For our tests we have expanded the classification scheme to make it logically complete. The complete set of residue groupings are designated A–H, with A–D encompassing residues not preceding a proline and E–H containing otherwise equivalent residues followed by a proline. Class A contains all residues not included in the other, more specialized, classes. Class B contains isoleucine or valine. Class C contains glycine and class D contains proline. Within each of these eight categories the ideal values and standard deviations are tabulated for each 10° × 10° cell in φ/ψ space, resulting in 36 × 36 × 8 = 10 368 target values for each length and angle. For infrequent (<3) conformers, reliable target values could not be derived and these were set equal to the global average for that category. Because no φ angle is defined for an N-terminal residue and no ψ angle is defined for a C-terminal residue, these residues were restrained using CSD-X target values. All restraints not within the main chain were also based on the CSD-X target values.

In the standard *TNT* library, the PEPTIDE linkage defines all of the bond-length and bond-angle restraints for the backbone. In the CDL, for a peptide bond connecting residue *i* to residue *i* + 1 most backbone target values depend on the classification of residue *i*, while the C—N length and $\mathrm{C-N-C^{\alpha}}$ angle depend more heavily on that of residue *i* + 1. Thus, CDL information was implemented by introducing 20 736 distinct peptide-linkage groups: 10 386 having the letter X as the first character of their name and defining those target values determined by the classification of residue *i* + 1 and another 10 368 starting with the letter Y with target values defined by residue *i*. For the names of the linkages, the second symbol was the amino-acid category code (*i.e.* A–H) and the third and fourth symbols specify the φ and ψ bins using the letters a–z as well as the digits 0–9 to provide the 36 required symbols. The letter a, for example, corresponds to an angle between −185° and −175°. A Python program was written to convert the CDL to this *TNT* geometry library.

A second program created a special *TNT* sequence file including the proper links based on φ/ψ angles and residue type. Normally sequence files are named with the `.seq` extension, but the new versions are named with a `.kseq` extension to make them distinct. As an example, a few residues of a `.seq` file and a `.kseq` file for rFNR (PDB code 3lo8) are shown here on the left and the right, respectively:



```
RESIDUE 20 ALA 21 PEPTIDE
RESIDUE 21 LYS 22 PEPTIDE
RESIDUE 22 GLU 23 PEPTIDE
```



```
RESIDUE 20 ALA 21 YAm6 21 XAjr
RESIDUE 21 LYS 22 YAjr 22 XEe7
RESIDUE 22 GLU 23 YEe7 23 XHl7
```

To refine with the CSD-X library one simply includes the `.seq` file in the *TNT* control file since CSD-X is the default library. To use the CDL, one includes the `.kseq` file and the CDL data file. The only change made to the *TNT* code was to increase the array size for storing linkages.

2.2. Refinement protocol

Refinement involved preconditioned conjugate-gradient (Tronrud, 1992) least-squares minimization with no manual intervention. As use of the new library might change the optimal overall weight for the diffraction data relative to the geometric restraints, refinements were run with overall weights of 0.25, 0.5, 1, 2 and 4, where the higher numbers enforce the diffraction data more strongly. This set of weights bracketed the lowest free *R* in all of the test cases performed.

Table 1
List of test-case data sets and study models.

Data-set name†	Resolution (Å)	X-ray source	Completeness (%)	$R_{\rm meas}$ (%)	Starting model
$\rm LT_{1.05}$	1.05	APS 14c	99.7	10.0	No new refinement
$\rm LT_{1.7}$	1.7	APS 14c	99	9.7	3lo8; anisotropic B factors removed
$\rm LT_{2.4}$	2.4	R-AXIS IV	93	6.8	3lo8; waters with $B > 30$ Å^2 deleted
$\rm RT_{1.7}$	1.7	R-AXIS IV	98.6	6.8	3lvb
$\rm RT_{2.3}$	2.3	R-AXIS IV	83	11.5	3lvb; waters with $B > 50$ Å^2 deleted

† LT and RT data sets were collected at 100 and 295 K, respectively. The $\rm LT_{1.7}$ data set was simply derived by truncating the $\rm LT_{1.05}$ data set. The others were all independently collected from crystals of rFNR.

We did not attempt to fine-tune the weight. Because initial tests showed that the R values did not always vary smoothly with the overall weight, 20 refinements were run for each test condition using starting models 'jiggled' by random coordinate and B-factor shifts with an r.m.s. magnitude of 0.2 Å and 2 Å^2, respectively. The statistics reported for each overall weight are the averages and the standard deviations of the mean based on each set of 20 runs.

2.3. Test-case data sets and models

For the test refinements, we chose the maize-root ferredoxin:NADP$^+$ reductase (rFNR) system, for which we have access to an ultrahigh-resolution reference structure as well as data sets in three resolution ranges and at two temperatures. Data sets used in the test cases are denoted as $\rm LT_{1.05}$, $\rm LT_{1.7}$, $\rm LT_{2.4}$, $\rm RT_{1.7}$ and $\rm RT_{2.3}$ to indicate the temperature of data collection and the resolution of each (Table 1). The reference ultrahigh-resolution model (3lo8) was originally refined against $\rm LT_{1.05}$ using *SHELXL* (Sheldrick, 2008) and has an $R_{\rm work}$ and $R_{\rm free}$ of 0.125 and 0.155, respectively (Faber & Karplus, unpublished data). Since we did not have a data set from a frozen rFNR crystal that only diffracted to 1.7 Å resolution, the data set $\rm LT_{1.7}$ is simply $\rm LT_{1.05}$ truncated to 1.7 Å. $\rm RT_{1.7}$ and the accompanying models, PDB entries 3lvb and 1jb9, have been described previously (Aliverti *et al.*, 2001). 1jb9 is the final model, refined against all data, and has an R value of 0.167. 3lvb is the penultimate model produced before the final round of refinement against the full data set. It has an $R_{\rm work}$ and $R_{\rm free}$ of 0.164 and 0.223, respectively. Data sets $\rm RT_{2.3}$ and $\rm LT_{2.4}$ were collected at our laboratory source using the same protocol as described for the $\rm RT_{1.7}$ data set (Aliverti *et al.*, 2001), but using a very short exposure time of 20 s per frame to ensure lower quality. Their resolution cutoffs reflect the resolution at which their intensities naturally fall to the noise level.

We included test refinements against both room-temperature and low-temperature data sets for two reasons. Firstly, because the CDL was derived from models based on LT data, including both assesses to what extent these restraints are also useful in refinements against RT data. Secondly, including the RT data sets allowed us to perform test refinements using a true 1.7 Å resolution data set rather than only one derived by the truncation of a higher resolution data set. Including the LT refinements was important as these refinements had a reference ultrahigh-resolution structure to compare against.

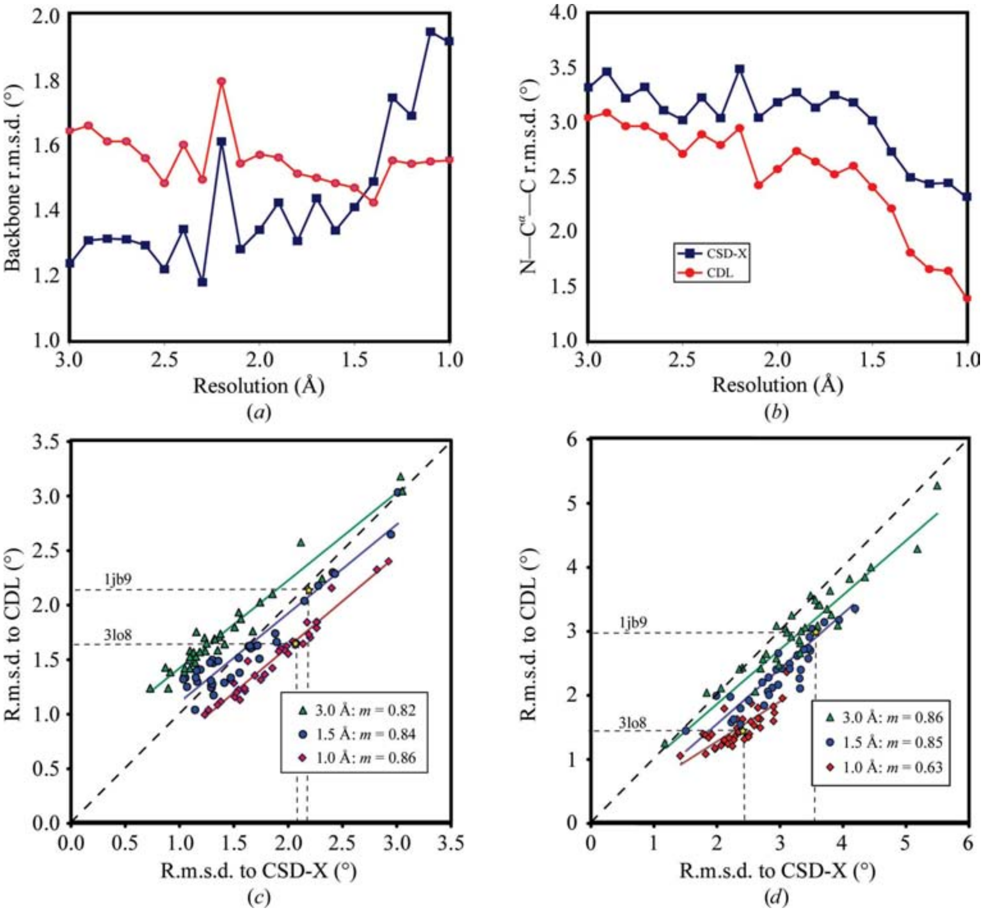


Figure 1
Agreement of existing models with the CSD-X and the CDL geometries as a function of resolution. (*a*) The median r.m.s.d. of the main-chain angles from the CSD-X library (blue squares) and the CDL (red circles) for 35 protein chains in each 0.1 Å resolution shell are plotted as a function of resolution. (*b*) Same as (*a*) but only the N—C$^\alpha$—C bond angles are considered. (*c*) Scatter plot comparing the r.m.s.d. of the backbone angles to the CDL for each chain as a function of its fit to the CSD-X library. The resolution bins shown are 3.0 Å (green triangles), 1.5 Å (blue circles) and 1.0 Å (red diamonds). (*d*) Same as C but for the N—C$^\alpha$—C bond-angle statistics. The 35 structures in each 0.1 Å resolution shell were selected from the PDBselect 25 list of Protein Data Bank entries (Griep & Hobohm, 2010) and are listed in the supplementary material.

Table 2
Agreement of the two test-case models with the libraries.

PDB code	Resolution (Å)	Library	MC lengths† (Å)	MC angles‡ (°)	$N-C^{\alpha}-C$ angles (°)
1jb9	1.7	CSD-X	0.0119	2.17	3.49
		CDL	0.0138	2.13	2.92
		Perfect	0.0180	2.41	2.61
3lo8	1.05	CSD-X	0.0145	2.08	2.39
		CDL	0.0145	1.68	1.42
		Perfect	0.0035	0.82	0.38

† R.m.s.d. for the five main-chain bond lengths. ‡ R.m.s.d. for the seven main-chain angles.

2.4. A 'perfect' restraint library

As a control, we constructed a 'perfect' *TNT* library by creating a sequence file where each residue and linkage had a unique type and the target values (backbone, side chain and cofactor) for these types and linkages were simply calculated from the 1.05 Å model 3lo8. The standard deviations were set to be equal to those from the equivalent entries in the CDL. The planarity restraints were left identical to those in the CSD-X library. When atoms occurred in alternative conformations the 'perfect' bond lengths and angles were calculated from the *A* conformer alone. Since the other conformers have different values, the overall r.m.s.d. (root-mean-square deviation) of 3lo8 from the 'perfect' library is not zero, but 0.004 Å and 1.00° (0.004 Å and 0.82° for main-chain restraints; see Table 2). Although 3lo8 is only an approximation to the true structure, it does match the diffraction data better than any other model. In principle, this library is perfect for guiding a lower resolution refinement to produce a structure that matches the reference 1.05 Å rFNR model. Although impractical for any other use, this library represents the extreme of a library that accounts for all the fluctuations in geometry caused by local context and conformation.

3. Results

3.1. Validation of the CDL

Given that a major symptom of the limitations of the CSD-X library is its poorer agreement with ultrahigh-resolution structures, we first tested how it and the CDL compared in this regard. As seen in Fig. 1(*a*), the r.m.s.d. from the CSD-X library for representative structures in the PDB shows the expected trend that agreement is good ($\sim$1.3°) at lower resolutions where restraints are dominant, but the agreement systematically worsens as the resolution gets better; finally, for structures better than 1.5 Å resolution there is a sharp degradation in agreement. In contrast, the fit to the CDL is largely independent of resolution, hovering near 1.5–1.6°. These trends mean that starting at near 1.5 Å resolution the agreement with the CDL becomes much better than that with the CSD-X library even though these structures have been generated using restraints to the CSD-X library. A similar analysis that only includes the highly variable $N-C^{\alpha}-C$ bond angles (Fig. 1*b*) remarkably shows that the CDL fits these angles better than the CSD-X library at all resolutions, with an increased margin of improvement at higher resolutions.

The analysis for bond lengths shows that the agreement with both libraries follows the same trend, with agreement becoming worse at higher resolutions (Fig. S1 in supplementary material[1]). This supports the conclusion that even for the $\leq$1 Å resolution structures used to create the CDL the diffraction data do not contain sufficient information to overcome bias from the CSD-X restraints.

To see how individual structures behave and how the CSD-X library and the CDL deviations correlate with each other, for each structure in the 3.0, 1.5 and 1.0 Å shells we compared how well it agreed with each of the two libraries (Figs. 1*c* and 1*d*). For both backbone angles and the $N-C^{\alpha}-C$ angle, within each shell the fit to the CDL increases in lock-step with the improvement in fit to the CSD-X library. In all cases the slopes are less than 1, showing that as restraints are loosened the agreement with the CDL decreases less than the agreement with the CSD-X library.

3.2. Assessing the suitability of the test cases

For the test refinements we used several data sets and models of maize-root ferredoxin:NADP$^+$ reductase (rFNR; see Table 1). Mapping the rFNR models (3lo8 and 1jb9) onto the plots of the sets of structures from the above analysis (Figs. 1*c* and 1*d*) shows that their behaviour fits well into the distributions, indicating they are representative.

We also compared these two rFNR models with all three libraries (Table 2). Both models agree with the target values from the CDL better than the targets from the CSD-X library, although the difference is smaller for the lower resolution model 1jb9. The differences are particularly striking for the $N-C^{\alpha}-C$ angles, where the fit is 20% better for 1jb9 and 40% better for 3lo8. We also compared the two models against the 'perfect' library derived from 3lo8, the 1.05 Å refined rFNR structure. For 3lo8 the backbone and $N-C^{\alpha}-C$ angles are of course very close to the perfect library; the residual is not exactly zero only because the model includes alternative conformations (see §2). For 1jb9, the backbone angles as a whole deviate more from the perfect library than from the other two libraries, but the $N-C^{\alpha}-C$ angles are substantially closer to the perfect library.

3.3. Head-to-head refinements comparing geometry libraries

The test data sets representing room temperature and cryoconditions at 1.7 Å and at 2.3 Å resolution (Table 1) were each used for parallel refinements with restraints either from the CSD-X library, the CDL or the 'perfect' library. A family of 20 refinements from different ('jiggled') starting models (see §2) was run for each condition. As monitors of refinement quality, we report the family average of $R_{\rm work}$ and $R_{\rm free}$ values, the r.m.s.d. from ideality of backbone angles as a whole and the $N-C^{\alpha}-C$ angles in particular. We do not report bond-

[1] Supplementary material has been deposited in the IUCr electronic archive (Reference: DZ5202). Services for accessing this material are described at the back of the journal.

length r.m.s.d.s because as noted above their variations owing to φ/ψ are small and in all our tests these did not change significantly when the library was changed (*e.g.* Table 2).

The general rule for choosing the weight for the diffraction data relative to the stereochemical restraints is to run a series of refinements with a selection of weights and to choose the weight that resulted in the lowest free R (Brünger, 1997). When the results of these tests are plotted as the $R_{\rm free}$ *versus* the logarithm of the weight, the resulting curve is expected to have roughly the shape of a parabola. Of the 12 sets of test refinements (four data sets with three different libraries each; Figs. 2 and 3), eight of them showed the lowest $R_{\rm free}$ at a weight of 1.0. In the other four cases the bottom of the $R_{\rm free}$ *versus* weight parabola was rather flat so that the lowest $R_{\rm free}$ and the $R_{\rm free}$ achieved using a weight equal to 1.0 were very close, with the difference never being greater than 0.0002. This difference is well below the precision of the $R_{\rm free}$ calculation, which can be estimated from the spread of the $R_{\rm free}$ values calculated for the 20 refinements performed for each trial (Figs. 2*a* and 3*a*). We can conclude that the $R_{\rm free}$ test has not indicated a significant difference between the weight at the minimal $R_{\rm free}$ and a weight of 1.0 in these cases. Since the other indicators monitored in these tests are strong functions of the weight, analysis of the differences between the three libraries is greatly simplified if we consider a weight equal to 1.0 to be representative of the performance for all of the test refinements.

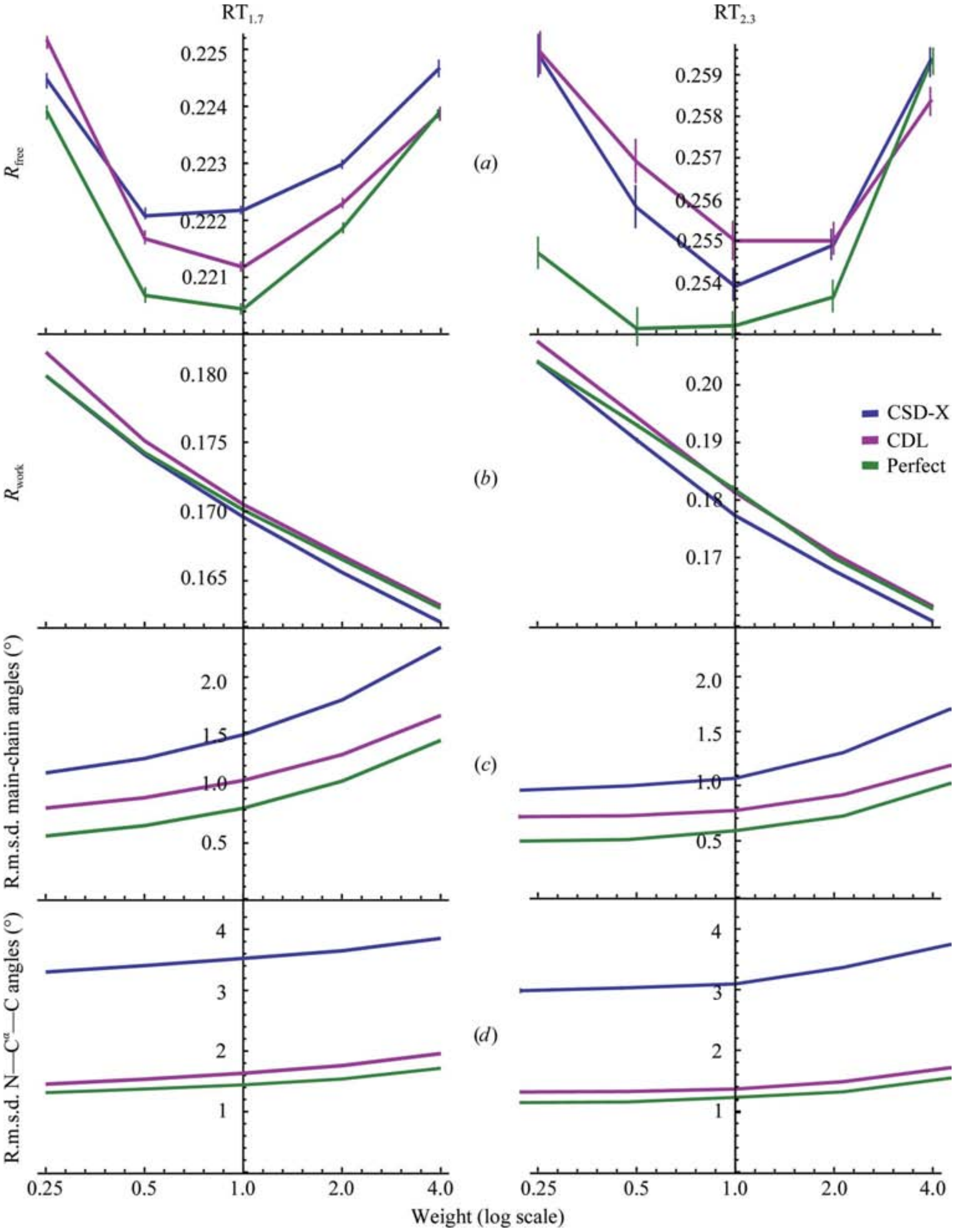


Figure 2
Behavior of test refinements against the RT$_{1.7}$ and RT$_{2.3}$ data sets. Each panel plots the mean values (with $\pm 1\sigma$ of the mean shown as error bars) of one refinement statistic as a function of the weight on the crystallographic data used for test refinements performed using the CSD-X library (blue), the CDL (violet) and the 'perfect' library (green). (*a*) $R_{\rm free}$. (*b*) $R_{\rm work}$. (*c*) R.m.s.d. for main-chain angles. (*d*) R.m.s.d. for N—C$^\alpha$—C angles. For many points error bars are smaller than the line thickness and are not visible.

3.3.1. Room-temperature rFNR test cases. Test refinements against RT$_{1.7}$ and RT$_{2.3}$ used PDB entry 3lvb as the starting model. The statistics for the 1.7 Å and the 2.3 Å refinements show very similar patterns (Fig. 2). The $R_{\rm free}$ differences between the CSD-X library and the CDL are very small (~0.001), with the $R_{\rm free}$ being slightly lower at 1.7 Å with the CDL and at 2.3 Å with the CSD-X library. The $R_{\rm work}$ value is strongly dependent on the weight, decreasing with increasing weight, but at both resolutions use of the CDL is associated with a consistent very small increase in $R_{\rm work}$ at any given weight (Fig. 2*b*).

In terms of geometric ideality, the impact is much more notable. For both refinements, use of the CDL led to drops in the r.m.s.d. from ideality of about 25% for backbone bond angles in general and of near 50% for the N—C$^\alpha$—C angles in particular (Figs. 2*c* and 2*d*). Interestingly, the perfect library performs better than the CDL, but only incrementally so.

3.3.2. Low-temperature rFNR test cases. The test refinements against the low-temperature data sets at 1.7 and 2.4 Å (Fig. 3) behaved similarly to the room-temperature tests. Changes in $R_{\rm work}$ were again very small but in the same directions and $R_{\rm free}$ changed very little, with the CDL giving the lower value against $LT_{1.7}$ but the CSD-X library being better against $LT_{2.4}$. Also, using the CDL, at both resolutions the r.m.s.d.s for the backbone bond angles and the $N—C^{\alpha}—C$ angles dropped by about 25 and 50%, respectively. As with the room-temperature test cases, the r.m.s.d.s dropped only a little further with use of the 'perfect' library. For both the room-temperature and low-temperature refinements Fig. 4 summarizes the improvement in main-chain and $N—C^{\alpha}—C$ bond-angle ideality obtained using the CDL or perfect libraries and a weight equal to 1.0.

3.3.3. Comparison of refined models to 3lo8. An estimate of the accuracy of the models resulting from the low-temperature refinements can also be obtained by comparing their main-chain atomic positions with those of the ultrahigh-resolution model 3lo8 (Fig. 5). The models created using the CDL were better at matching the 3lo8 standard that those created using the CSD-X library by about 10% at 1.7 Å and 15% at 2.4 Å. Using this measure, models created using the perfect library were only another 8% better.

4. Discussion

For over 35 years protein crystallographers have been refining models using the paradigm of a single 'ideal' target value. For a given chemical type (considering protonation state, *cis versus trans* isomers and glycine and proline backbones as distinct chemical types), bond lengths and angles were defined as quantities that did not vary with conformation. The availability of a new library in which the ideal values for the main-chain bond lengths and angles are parameterized by φ/ψ angles (*i.e.* the CDL; Berkholz *et al.*, 2009) provides the opportunity to explore its utility in protein refinement.

In addition to the CDL's incorporation of φ/ψ variability, it differs from the most commonly used library of the previous paradigm, the CSD-X library of Engh & Huber (1991), both in the source of structural models and the number of residue categories. Although the CDL is based on ultrahigh-resolution protein models as opposed to small-molecule peptides, these sources have been shown to be roughly equivalent when constructing CSD-X style dictionaries (*e.g.* Jaskolski *et al.*, 2007*b*). Also, the increased number of residue categories in the CDL (eight *versus* three) has little practical effect, as most of the new categories are based on few structural examples and were included only for the sake of logical completeness. The vast majority of the new information present in the CDL arises from its φ/ψ variability.

An analysis of the agreement of models in the Protein Data Bank with the new CDL (Fig. 1) indicates that this library completely overcomes the major

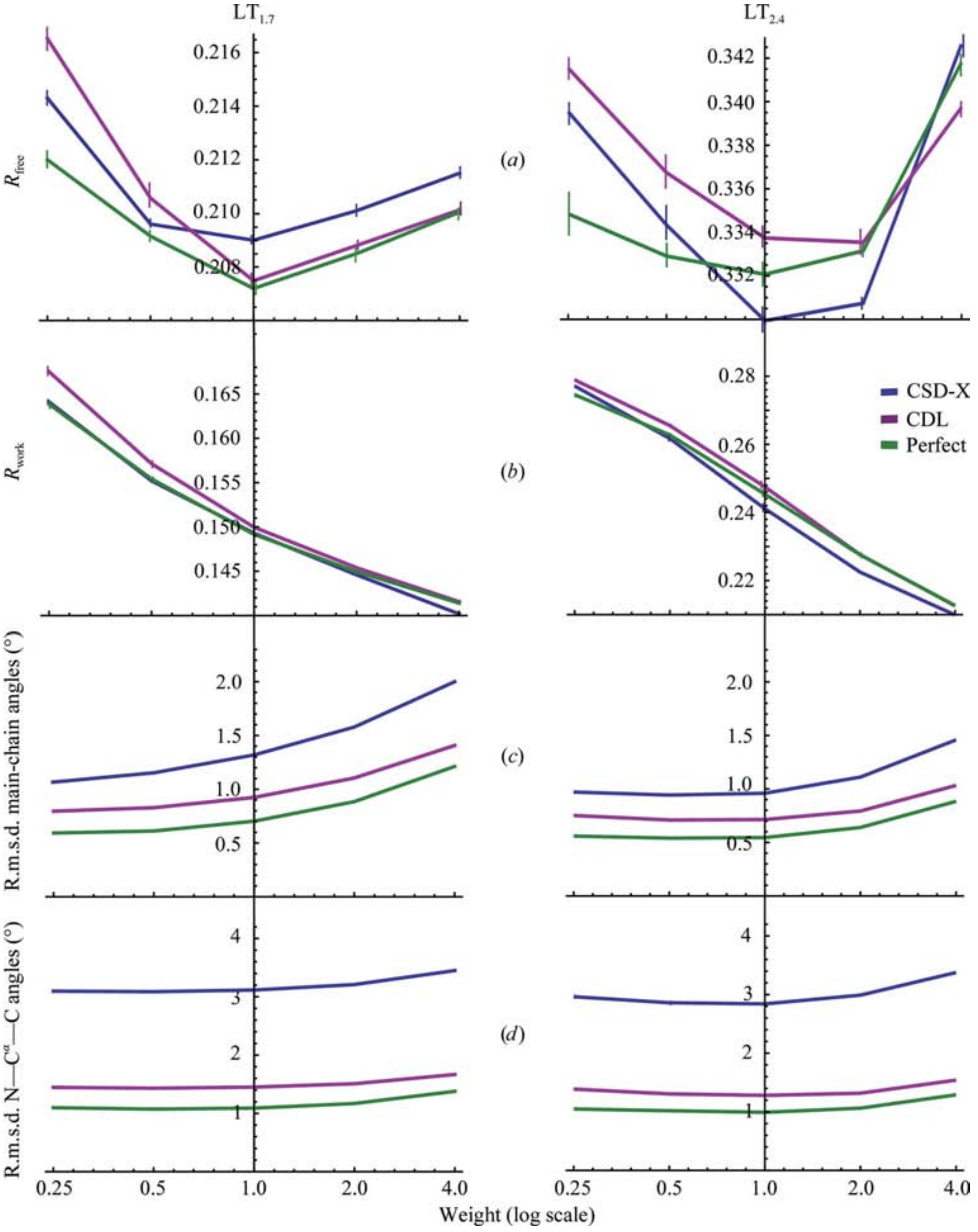


Figure 3
Behavior of test refinements against the $LT_{1.7}$ and $LT_{2.4}$ data sets. Panel contents (including error bars) and colors are as in Fig. 2. (*a*) $R_{\rm free}$. (*b*) $R_{\rm work}$. (*c*) R.m.s.d. for main-chain angles. (*d*) R.m.s.d. for $N—C^{\alpha}—C$ angles.

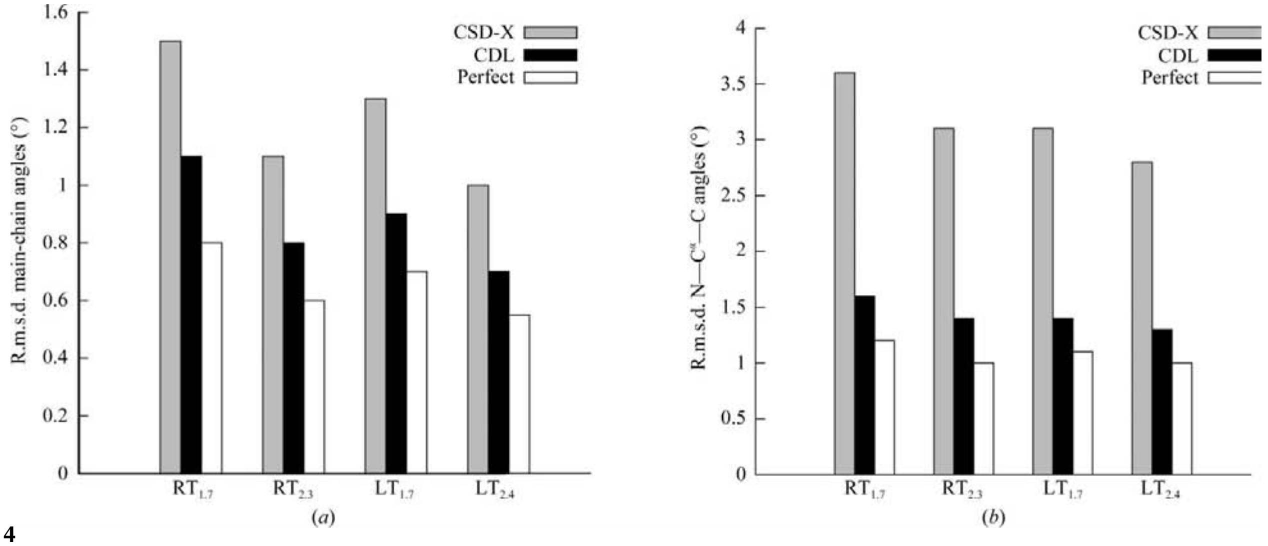


Figure 4
Summary of comparisons of bond-angle agreements for test refinements. Plotted here are the mean bond-angle r.m.s.d. from those refinements where the weight was equal to 1.0. (*a*) Average r.m.s.d. from main-chain angle restraints for models refined against the three libraries. (*b*) Average r.m.s.d. from N—C$^\alpha$—C angle restraints for the same models.

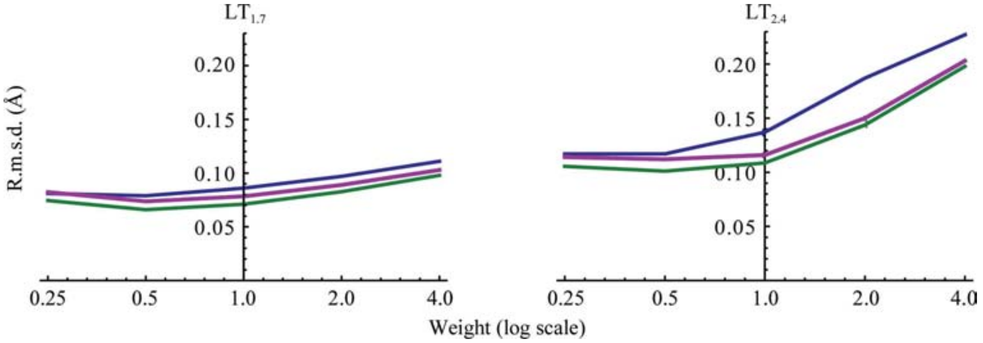


Figure 5
Assessing the accuracy of models from the low-temperature test refinements. Plotted as a function of refinement weight are the average backbone r.m.s.d. between 3lo8 and the models resulting from trial refinements against LT$_{1.7}$ and LT$_{2.4}$. Colors and error bars are as in Fig. 2. Since 3lo8 is a highly accurate determination of the protein structure within this crystal form, agreement with the 3lo8 structure can be used as a measure of the accuracy of the structures resulting from lower resolution refinements.

problem associated with the CSD-X library (Engh & Huber, 1991): that ultrahigh-resolution protein models exhibit disturbingly large and increasing deviations from the library despite having been restrained to it. This in itself is a compelling argument for adopting the CDL for use in crystallographic refinement.

The relationships shown in Figs. 1(*c*) and 1(*d*) reveal an interesting very strongly positive correlation between how well models agree with the two libraries. This implies that, all other things being equal, restraining a model more tightly to the CSD-X library leads to a model that also agrees better with the CDL. Where the lines drop below the diagonal the models agree with the CDL even better than they do with the CSD-X library, despite having been restrained to the CSD-X library. For backbone angles as a whole this becomes consistently the case at about 1.5 Å resolution (Fig. 1*c*), but for the N—C$^\alpha$—C angles this already is true at 3 Å resolution (Figs. 1*b* and 1*d*). That the N—C$^\alpha$—C angle is already more robustly determined at 3 Å resolution makes sense because it represents not just the angle between three atoms, but also between two peptide planes. What is somewhat surprising is that for backbone bond angles in general the correlation in r.m.s.d.s from the two libraries is equally as strong at 1.0 Å resolution as at the lower resolutions.

The concomitant improvement in fit to the restraints of the CSD-X library and the CDL makes sense because of their strong correlation (*i.e.* if the φ/ψ variability within the CDL were averaged away the resulting restraints roughly match the CSD-X library). The slope of the line fitting the distribution of models at any particular resolution is about 0.8, showing that restraining to the CSD-X library does not cause an equal incidental improvement in fit to the CDL. (If the two libraries were equivalent the slope would be one and if they were uncorrelated the slope would be zero.) For any particular r.m.s.d. from the CSD-X library an increase in resolution of the X-ray data set will result in a better fit to the CDL, which is further proof that the additional variability of the CDL is reflective of true variability of the protein main chain. For 1 Å resolution models the fit of the N—C$^\alpha$—C angles approaches the limit of the CSD-X library. For these models the slope of only 0.63 indicates that further tightening of the weight on the CSD-X restraints is less effective. All this evidence together supports the conclusion that even at resolutions as low as 3 Å there is information in the diffraction data sensitive to φ/ψ variability which cannot be fitted by simply increasing the weight on the CSD-X restraints. The CDL brings advantages for improving structures at all resolutions.

The refinements of rFNR models against diffraction data sets at varying resolution produced two conclusions. Firstly, the geometry quality improved substantially at all resolutions, with little impact on the working and free R values. Interestingly, these results are qualitatively similar to the improvements in refinement seen on the introduction of the CSD-X library (Engh & Huber, 1991). In that case, which was before the introduction of the free R, when the CSD-X library replaced the P19X library of *X-PLOR* (Brünger *et al.*, 1987) tests showed only a very small 0.002 drop in R value for a 1.2 Å refinement and a 0.001 drop for a 1.66 Å refinement. The accompanying drops in r.m.s.d. bond angles for the two refinements were considered to be substantial at ~13% and ~8%, respectively. Apparently, real improvements in ideal geometry libraries do not necessarily improve the overall fit to diffraction data (as measured by R values) even though they do improve the overall quality of the model as seen by the improvement in geometry ideality at an equivalent R value. The larger 25% improvements in bond-angle ideality seen here suggest that the adoption of the CSD paradigm will be a more substantial step forward than the adoption of the CSD-X library. Furthermore, the refinements against the RT diffraction data sets (Fig. 2) show that the CDL, a library primarily derived from frozen crystals, is equally effective for models derived from cryotemperature and room-temperature studies.

The second main conclusion from the rFNR test refinements is that the improvement added by the use of the CDL (compared with that of the CSD-X library) is a substantial fraction of that achievable by the use of a 'perfect' library. The trial refinements using the 'perfect' library provide an upper bound to the amount of improvement achievable through the use of better and better geometry libraries. We note that since the 1.05 Å resolution model (3lo8) was the source of this 'perfect' library there may be bias toward these atomic positions, so that these particular tests may overestimate but should not underestimate the power of a truly perfect library. As such, these tests clearly show that even a perfect library is no panacea. While its usage does cause the free R values to drop in all but one case (Figs. 2*a* and 3*a*), these drops are still very small. In the most direct indicator of impact, the comparisons to the 1.05 Å reference model *via* r.m.s. positional error, it is seen that for the 1.7 Å refinement the CDL provides half of the improvement provided by the perfect library (Fig. 5*a*). Remarkably, for the 2.4 Å refinement the CDL achieves >75% of the improvements provided by the perfect library (Fig. 5*b*). This proves that incorporating target-value variability into a stereochemical library can have tangible impacts that are important even at such lower resolutions. In complete agreement with this is the summary observation that in comparison with the perfect library the CDL captures the majority of the variability of bond angles, meaning that any other 'contextual' information is of secondary importance (Fig. 4).

We conclude that it is well worth reconfiguring geometry libraries to be able to implement φ/ψ dependencies in refinement because they improve behavior at all resolutions. The strategy for incorporating the CDL into refinement software should look beyond the information in this first-generation CDL and allow the future creation of broader context dependencies, including effects that have already been shown to occur for peptide planarity (Karplus, 1996) and as yet uncharacterized correlations that we presume exist between side-chain torsion-angle values and side-chain bond angles.

This work was supported in part by National Institutes of Health grant R01-GM083136, National Science Foundations grant MCB-9982727 and National Institute of Environmental Health Sciences grant P30 ES00210. M. V. Shapovalov and R. L. Dunbrack created the version of the CDL used in these tests. H. R. Faber collected the $LT_{1.05}$ data set and created the model (PDB code 3lo8) of rFNR based on it. C. J. Karplus collected the $LT_{2.4}$ and $RT_{2.3}$ data sets. Use of the Advanced Photon Source was supported by the US Department of Energy, Basic Energy Sciences, Office of Science, under Contract No. DE-AC02-06CH11357. Use of the BioCARS Sector 14 was supported by the National Institutes of Health, National Center for Research Resources, under grant number RR007707.

References

Aliverti, A., Faber, R., Finnerty, C. M., Ferioli, C., Pandini, V., Negri, A., Karplus, P. A. & Zanetti, G. (2001). *Biochemistry*, **40**, 14501–14508.

Allen, F. H. (2002). *Acta Cryst.* B**58**, 380–388.

Berkholz, D. S., Shapovalov, M. V., Dunbrack, R. L. Jr & Karplus, P. A. (2009). *Structure*, **17**, 1316–1325.

Bricogne, G. & Irwin, J. J. (1996). *Proceedings of the CCP4 Study Weekend. Macromolecular Refinement*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey, pp. 85–92. Warrington: Daresbury Laboratory.

Brünger, A. T. (1997). *Methods Enzymol.* **277**, 366–396.

Brünger, A. T., Karplus, M. & Petsko, G. A. (1989). *Acta Cryst.* A**45**, 50–61.

Brünger, A. T., Kuriyan, K. & Karplus, M. (1987). *Science*, **235**, 458–460.

Chapman, M. S. (1995). *Acta Cryst.* A**51**, 69–80.

Dauter, Z. & Wlodawer, A. (2009). *Structure*, **17**, 1278–1279.

Diamond, R. (1971). *Acta Cryst.* A**27**, 436–452.

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.

Griep, S. & Hobohm, U. (2010). *Nucleic Acids Res.* **38**, D318–D319.

Hendrickson, W. A. & Konnert, J. H. (1980). *Computing in Crystallography*, edited by R. Diamond, S. Ramaseshan & K. Venkatesan, pp. 13.01–13.26. Bangalore: Indian Academy of Sciences.

Jaskolski, M., Gilski, M., Dauter, Z. & Wlodawer, A. (2007*a*). *Acta Cryst.* D**63**, 1282–1283.

Jaskolski, M., Gilski, M., Dauter, Z. & Wlodawer, A. (2007*b*). *Acta Cryst.* D**63**, 611–620.

Karplus, P. A. (1996). *Protein Sci.* **5**, 1406–1420.

Karplus, P. A., Shapovalov, M. V., Dunbrack, R. L. & Berkholz, D. S. (2008). *Acta Cryst.* D**64**, 335–336.

Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. (1963). *J. Mol. Biol.* **7**, 95–99.

Sheldrick, G. M. (2008). *Acta Cryst.* A**64**, 112–122.

Stec, B. (2007). *Acta Cryst.* D**63**, 1113–1114.

Tickle, I. J. (2007). *Acta Cryst.* D**63**, 1274–1281.

Tronrud, D. E. (1992). *Acta Cryst.* A**48**, 912–916.

Tronrud, D. E. (1997). *Methods Enzymol.* **277**, 306–319.

Tronrud, D. E., Ten Eyck, L. F. & Matthews, B. W. (1987). *Acta Cryst.* A**43**, 489–501.

Vijayan, M. (1976). *CRC Handbook of Biochemistry and Molecular Biology*, 3rd ed., Vol. III, edited by G. D. Fasman, pp. 742–759. Cleveland: CRC Press.