Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Experimental determination of optimal root-mean-square deviations of macromolecular bond lengths and angles from their restrained ideal values

Ian J. Tickle

Astex Therapeutics Ltd, 436 Cambridge Science Park, Milton Road, Cambridge CB4 0QA, England

Correspondence e-mail: i.tickle@astex-therapeutics.com

Received 11 September 2007
Accepted 12 October 2007

A number of inconsistencies are apparent in the recent research paper by Jaskolski *et al.* [(2007), *Acta Cryst.* D**63**, 611–620] concerning their recommendations for the values of the magnitude and resolution-dependence of the root-mean-square deviations (RMSDs) of bond lengths and angles from their restrained ideal values in macromolecular refinement, as well as their suggestions for the use of variable standard uncertainties dependent on atomic displacement parameters (ADPs) and occupancies. Whilst many of the comments and suggestions in the paper regarding updates for the ideal geometry values proposed by Engh and Huber are entirely reasonable and supported by the experimental evidence, the recommendations concerning the optimal values of RMSDs appear to be in conflict with previous experimental and theoretical work in this area [Tickle *et al.* (1998), *Acta Cryst.* D**54**, 243–252] and indeed appear to be based on a misunderstanding of the distinction between RMSD and standard uncertainty (SU). In contrast, it is proposed here that the optimal values of all desired weighting parameters, in particular the weighting parameters for the ADP differences and for the diffraction terms, be estimated by the purely objective procedure of maximizing the experiment-based log(free likelihood). In principle, this allows all weighting parameters that are not known accurately *a priori* to be scaled globally, relative to those that are known accurately, for an optimal refinement. The RMS Z score (RMSZ) is recommended as a more satisfactory statistic than the RMSD to assess the extent to which the geometry deviates from the ideal values and a theoretical rationale for the results obtained is presented in which the optimal RMSZ is identified as the calculated *versus* true Z-score correlation coefficient, the latter being a monotonic function of the resolution cutoff of the data. Regarding the proposal to use variable standard uncertainties, it is suggested that any departure from the current practice of using fixed weights for geometric restraints based on experimental values of standard uncertainties be subject to the same experiment-based validation.

1. Introduction

The three main recommendations of Jaskolski *et al.* (2007) can be summarized as follows: (i) some geometrical target values, as originally obtained by statistical analysis of small-molecule structures (Engh & Huber, 1991, 2001) extracted from the Cambridge Structural Database (CSD), should be adjusted (C—N bond length and N—C^{α}—C angle) or the weight reduced (peptide ω torsion angle); (ii) for many refinements at medium to high resolution the deviations from these target values are much too low because the default overall weighting of the X-ray diffraction data in some macromolecular refinement software in current use has been set too low and consequently should be substantially increased; and (iii) the individual geometry weights should be varied as a function of the atomic occupancies and atomic displacement parameters (ADPs), in order that the deviations from ideality for the poorly defined flexible regions are reduced whilst those for the well defined regions are allowed to increase.

Whilst it is recognized that there is clear experimental evidence for recommendation (i) above, the authors' findings in respect of their

other two main recommendations are disputed. Their suggestions in respect of (ii) above appear to conflict with earlier experimental results and theoretical considerations (Tickle *et al.*, 1998*a*) and also appear to be self-contradictory. A number of quotes from the paper that are relevant to the argument follow:

> This balance is often adjusted not quite correctly by giving too much weight to stereochemical data and in effect leading to regularization rather than optimization of the structure. This is illustrated by unusually small reported r.m.s.d. values, sometimes several times lower than the recommended target values.
>
> The effect is especially pronounced at 2 Å, where nearly 30% of the structures are reported with r.m.s. deviations from idealized bond lengths within 0.006 Å.
>
> Tight restraints are necessary at low resolution ... With increasing resolution, the weights should be progressively less tight for well defined parts of the model. ... As a practical guideline, we recommend to aim at an r.m.s.d. for bonds of 0.020 Å for models at 1.5 Å or lower resolution.
>
> At very high resolution ... it is normal in such situations for the model bond distances to deviate from the idealized targets by 0.02–0.03 Å.

Hence, on one hand the authors suggest that at low resolution 'tight restraints' are necessary (no numerical values for the target RMSDs are suggested in this situation) becoming 'progressively less tight' with increasing resolution, but then in apparent contradiction recommend a target RMSD for bond lengths of 0.02 Å independent of resolution cutoff, except at very high resolution (1.5 Å or higher) where the target RMSD may exceed 0.02 Å, up to around 0.03 Å.

Although the authors do point out that 'the overall level of the restraint weights can be validated by the use of the free *R* factor', they fail to perform any such validation of their recommended RMSD values. Minimization of cross-validation statistics such as $R_{\rm free}$ is currently the only completely objective practical method for obtaining reliable estimates of the restraint and X-ray weights (Brünger, 1992, 1993, 1997; Kleywegt & Brünger, 1996) that is implemented in currently available macromolecular refinement programs; importantly, it has also been demonstrated that this method minimizes the mean phase error and hence minimizes the standard uncertainty (SU) of the electron density. Lebedev *et al.* (2003) have proposed an alternative method based on maximum-likelihood estimation that uses all of the data to optimize the weights, rather than the 5–10% normally used for cross-validation, and thus has the potential to achieve much greater precision by reducing the sampling errors; however, their algorithm has not yet been implemented within the framework of currently available refinement software.

However, there are sound theoretical arguments (Tickle *et al.*, 1998*b*) as to why $R_{\rm free}$ is not the optimal statistic for this purpose and it is proposed that the log(free likelihood) ($\rm LL_{free}$) be used in its place. The essence of the argument is that it can be proved that in the case of least-squares refinement the optimal weights are those that are based on accurate estimates of the SUs of the restraints and X-ray observations. In the case that the SUs are not normally known accurately *a priori*, such as for the ADP restraints and the X-ray observations, it was proved that optimality is achieved when the free least-squares residual [the counterpart of the negative log(free likelihood) in maximum-likelihood refinement] is minimized with respect to the unknown weighting parameters, provided that the refinement with respect to the structure parameters has converged. Although these results are strictly applicable only to least-squares refinement, they are nevertheless still relevant to maximum-likelihood refinement because Murshudov *et al.* (1997) had previously shown that on convergence of structure refinement of a complete and as far as possible correct model when the phase errors are at a minimum, likelihood-based refinement tends asymptotically towards least-squares refinement.

It can be shown that in theory $R_{\rm free}$ is also minimized by the optimal choice of weights but only if some fairly drastic assumptions are made (*e.g.* assuming that all diffraction data have the same uncertainty), otherwise it is clear that $R_{\rm free}$ is only approximately minimized by the optimal choice of weights. No-one has ever contemplated using $R_{\rm work}$ as an optimization target function (or as a component of one) and there are very good reasons for this: unlike the least-squares or the log-likelihood functions, the absolute value function $|F_{\rm obs} - F_{\rm calc}|$ has discontinuous gradients which preclude the use of any standard optimization algorithm such as conjugate gradient that relies on the gradient of the function being optimized having an asymptotic limit of zero as a criterion for convergence; in any case, no existing theory requires that $R_{\rm work}$ be an optimization target. For exactly the same reasons crystallographers should not be contemplating the use of $R_{\rm free}$ as a target function!

One further practical problem with the use of $R_{\rm free}$ as a validation statistic is that unlike $\rm LL_{free}$ it is not uniquely defined. Some programs, *e.g. REFMAC* (Murshudov *et al.*, 1997), define it in the 'traditional' manner based on the calculated structure amplitude $F_{\rm calc}$,

$$R_{\rm free} = \sum_{\mathbf{h}\in\text{test set}} |F_{\rm obs}(\mathbf{h}) - F_{\rm calc}(\mathbf{h})| / \sum_{\mathbf{h}\in\text{test set}} F_{\rm obs}(\mathbf{h}). \qquad (1)$$

Other programs, *e.g. BUSTER* (Blanc *et al.*, 2004), use a version of (1) that is arguably more appropriate in the context of maximum-likelihood methodology based on the expected value of the structure amplitude $F_{\rm expect}$ (Pannu & Read, 1996),

$$R'_{\rm free} = \sum_{\mathbf{h}\in\text{test set}} |F_{\rm obs}(\mathbf{h}) - F_{\rm expect}(\mathbf{h})| / \sum_{\mathbf{h}\in\text{test set}} F_{\rm obs}(\mathbf{h}) \qquad (2)$$

The aim of this commentary on the paper by Jaskolski *et al.* (2007) is firstly to show how optimal weighting parameters and target RMSD values can be obtained by a completely objective optimization procedure by using the log(free likelihood) as a cross-validation statistic for four test structures with a range of high-resolution cutoffs and secondly to rationalize the values obtained by means of rigorous statistics-based theoretical arguments.

2. Methods

The method used for establishing optimal overall restraint weights and the overall X-ray weight by means of cross-validation is described in detail in this section. The method has been tested on four test data sets spanning high-resolution cutoffs from 2.55 to 1.33 Å: (i) human β-secretase (hydroxyethylamine derivative inhibitor complex; PDB code 1w51; Patel *et al.*, 2004), (ii) maize β-glucosidase mutant (dhurrin complex; PDB code 1e55; Czjzek *et al.*, 2000), (iii) human AMPK γ1 (R299Q mutant and AMP complex; PDB code 2uv6; Day *et al.*, 2007) and (iv) human AMPK γ1 (AMP complex; PDB code 2uv4; Day *et al.*, 2007).

The macromolecular restrained maximum-likelihood refinement program *REFMAC* (a locally modified version of the distributed version 5.3.0040; Murshudov *et al.*, 1997) was used throughout for the test refinements. The negative logarithm of the posterior probability function of the structure parameters, that is the function whose global minimum is sought by restrained refinement in *REFMAC*, can be written in the form

$$M = \sum_{\text{bonds}} (d_{\text{calc}} - d_{\text{ideal}})^2/\sigma_d^2 + \sum_{\text{angles}} (\theta_{\text{calc}} - \theta_{\text{ideal}})^2/\sigma_\theta^2 + \sum_{\text{torsions}} (\tau_{\text{calc}} - \tau_{\text{ideal}})^2/\sigma_\tau^2 + \sum_{\text{planar atoms}} \Delta_{\text{plane}}^2/\sigma_{\text{plane}}^2 + W_{B\text{iso}} \sum_{\text{bonded IDPs}} \Delta_{B\text{iso}}^2/\sigma_{B\text{iso}}^2 + W_{\text{X-ray}} \sum_{\text{reflections}} -\log\{L[F_{\text{obs}}(\mathbf{h})]\} + \text{optional terms}. \quad (3)$$

Here, the standard uncertainties (SUs) σ_d and σ_θ for the bond-length (d_{calc}) and angle (θ_{calc}) terms are equated to the sample standard deviations of the corresponding ideal bond length (d_{ideal}) or angle (θ_{ideal}) extracted from the CSD (Engh & Huber, 1991, 2001); for the calculated torsion angles (τ_{calc}), the corresponding ideal values τ_{ideal} (taking the angular periodicity into account) and their SUs σ_τ are derived from an analysis of torsion angles in the PDB (MacArthur & Thornton, 1996); for the plane deviations (Δ_{plane}) the SUs σ_{plane} are set to a suitable uniform arbitrary value (0.02 Å); finally, $L[F_{\text{obs}}(\mathbf{h})]$ is the likelihood function of the observed amplitudes. Note that 'maximum-likelihood refinement' is actually a misnomer in this context since it is the posterior probability $\exp(-M)$ that is being maximized not the likelihood L.

For this test, the SUs $\sigma_{B\text{iso}}$ for the bonded atomic isotropic displacement parameter (IDP) differences ($\Delta_{B\text{iso}}$) were set to 5 Å^2 for main-chain bonded atoms and 7 Å^2 for side chains and ligands; the *REFMAC* source code was modified so that no restraints were applied to IDP differences related by bond angles on the grounds that these are already sufficiently restrained by (and in any case are not independent of) the bonded IDP difference restraints. The form of the $\Delta_{B\text{iso}}$ term in (3) implies that the errors in the IDP differences are normally distributed with constant uncertainty; however, it is clear that this cannot be a realistic error model in this case. For example, an error of 10 Å^2 in a B_{iso} of 10 Å^2 would clearly be much more significant in terms of the contributions to the structure factors than the same error in a B_{iso} of 100 Å^2; also, deviations in B_{iso} do not have a symmetrical effect as would be implied by the normal error distribution, *i.e.* negative deviations have a greater effect than positive ones. This implies that the uncertainty in B_{iso} should be proportional to some low power of B_{iso} (*e.g.* between 1 and 2); this value could in principle be determined (Tickle, unpublished work) by the same method of optimizing LL_{free} as is proposed here for determining the other unknown weighting parameters. Nevertheless, a normal error model for the $\Delta_{B\text{iso}}$ term was assumed for the purpose of the tests because it is the only one available in the current version of the *REFMAC* program. Thus, the overall weighting parameters $W_{B\text{iso}}$ and $W_{\text{X-ray}}$ are the only two unknowns to be optimized with respect to LL_{free}.

The *REFMAC* source code was also modified to allow input of a resolution-independent X-ray weight related to the usual 'matrix weight',

$$W_{\text{X-ray}} = W_{\text{Matrix}} \text{Tr}(\mathbf{H}_{\text{Restraints}})/\text{Tr}(\mathbf{H}_{\text{X-ray}}), \quad (4)$$

where Tr(**H**) is the trace of a Hessian matrix of second derivatives, and also to print out a table of −log(free likelihood) values ($-\text{LL}_{\text{free}}$) and RMS Z scores (RMSZ) for bond-length and angle restraints.

For each of the four PDB structures used for the test, $W_{\text{X-ray}}$ was fixed at a suitable approximate initial value (2.5) and 50 iterations of standard restrained refinement using default values for the control parameters (except where previously otherwise stated) for each of a range of values of $W_{B\text{iso}}$ were performed in order to ensure convergence; $W_{B\text{iso}}$ was then fixed at its optimum value obtained at the maximum of LL_{free}. Further runs of 50 iterations were then performed, this time varying $W_{\text{X-ray}}$, and the final optimum value of $W_{\text{X-ray}}$ was determined at the new maximum of LL_{free}. The optimizations of $W_{B\text{iso}}$ and $W_{\text{X-ray}}$ were each repeated using the updated values each time in order to check for convergence of the weighting parameter optimization. Additional runs for a 14 × 14 matrix of input value pairs of $W_{B\text{iso}}$ and $W_{\text{X-ray}}$ centred near the optimum pair were performed for one of the test data sets (2uv4) in order to obtain an indication of the local bivariate dependence of R_{free} and $-\text{LL}_{\text{free}}$ on $W_{B\text{iso}}$ and $W_{\text{X-ray}}$.

3. Results

Fig. 1 shows, for the refinement of 1w51 (2.55 Å resolution data), R_{free} and $-\text{LL}_{\text{free}}$ plotted against $W_{B\text{iso}}$. Fig. 2 shows, for the same data, R_{free}, $-\text{LL}_{\text{free}}$, $\text{RMSD}_{\text{bonds}}$ and $\text{RMSZ}_{\text{bonds}}$ plotted against $W_{\text{X-ray}}$. Figs. 3 and 4 show the same plots for 2uv4 (1.33 Å resolution data). The 'noisy' appearance of these plots is likely to be a consequence of rounding errors in the calculation of geometry and structure-factor derivatives and also in the conjugate-gradient optimization algorithm itself. In Figs. 2 and 4 it appears that the minimum of R_{free} occurs at a

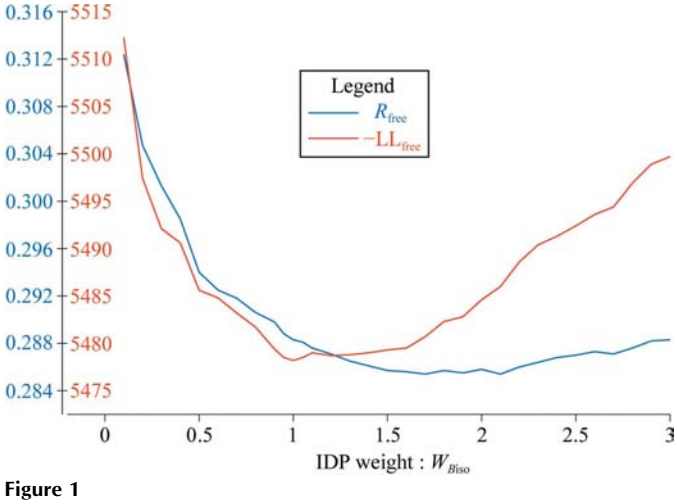

Figure 1
Plot of R_{free} (tic labels and line coloured blue) and $-\text{LL}_{\text{free}}$ (red) at convergence of refinement *versus* overall isotropic displacement parameter (IDP) weighting parameter $W_{B\text{iso}}$ for the β-secretase structure (PDB code 1w51) with $W_{\text{X-ray}}$ fixed at 2.5.

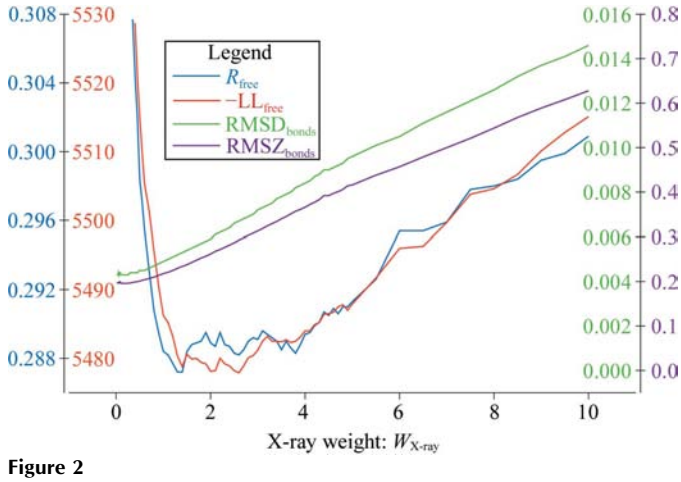

Figure 2
Plot of R_{free} (tic labels and line coloured blue), $-\text{LL}_{\text{free}}$ (red), $\text{RMSD}_{\text{bonds}}$ (green) and $\text{RMSZ}_{\text{bonds}}$ (purple) at convergence of refinement *versus* overall reflection weighting parameter $W_{\text{X-ray}}$ for the β-secretase structure (PDB code 1w51) with $W_{B\text{iso}}$ fixed at its optimum value.

Table 1
Summary of refinement statistics for the four test structures.

PDB code	Resolution cutoff (Å)	Mean B_{iso} (Å^2)	Final W_{Biso}	Final $W_{X\text{-}ray}$	Optimal $RMSD_{bonds}$ (Å)	Optimal $RMSZ_{bonds}$	Optimal $RMSD_{angles}$ (°)	Optimal $RMSZ_{angles}$
1w51	2.55	42.4	1.0	2.55	0.0066	0.29	1.01	0.44
1e55	2.00	24.7	1.8	2.50	0.0084	0.35	1.15	0.52
2uv6	2.00	22.5	1.3	1.31	0.0067	0.28	1.08	0.46
2uv4	1.33	13.6	0.17	1.40	0.0114	0.47	1.30	0.62

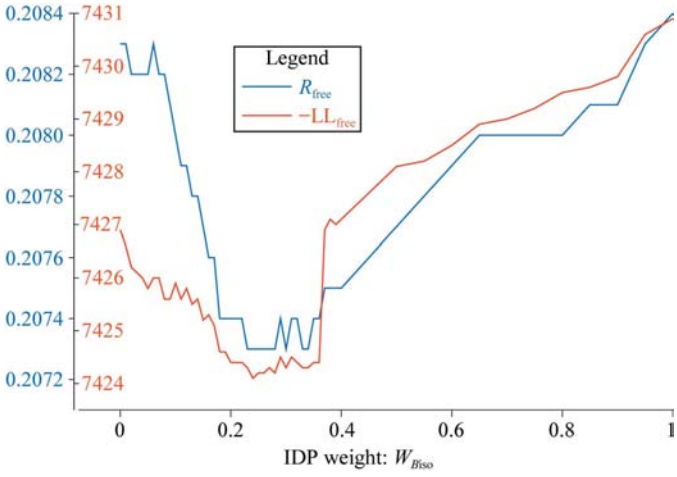

Figure 3
Plot of R_{free} (tic labels and line coloured blue) and $-LL_{free}$ (red) at convergence of refinement *versus* overall IDP weighting parameter W_{Biso} for the AMPK structure (PDB code 2uv4) with $W_{X\text{-}ray}$ fixed at 2.5.

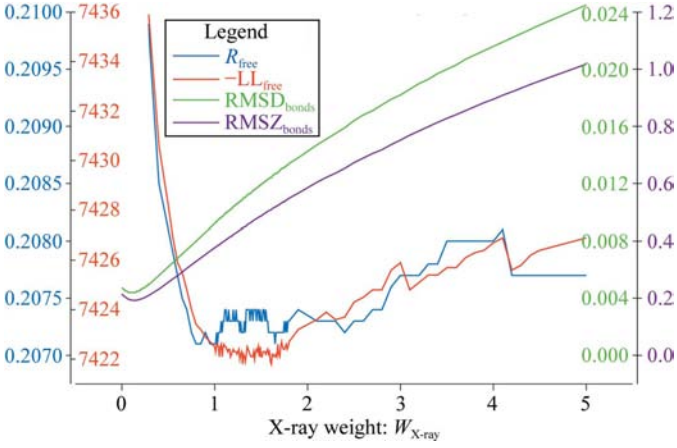

Figure 4
Plot of R_{free} (tic labels and line coloured blue), $-LL_{free}$ (red), $RMSD_{bonds}$ (green) and $RMSZ_{bonds}$ (purple) at convergence of refinement *versus* overall reflection weighting parameter $W_{X\text{-}ray}$ for the AMPK structure (PDB code 2uv4) with W_{Biso} fixed at its optimum value.

somewhat lower value of $W_{X\text{-}ray}$ than that of $-LL_{free}$; however, this effect is not seen consistently for all the test cases.

Table 1 shows a summary of the results for all four structures used in this test. For each structure, the values found after repetition of the optimizations of W_{Biso} and $W_{X\text{-}ray}$ were both very close to the values found after the initial optimization. Table 2 (see supplementary material[1]) shows relative values of R_{free} and $-LL_{free}$ obtained for a 14×14 matrix of input value pairs of W_{Biso} and $W_{X\text{-}ray}$ centred near the pair giving the minimum $-LL_{free}$ for the 2uv4 data; Fig. 5 shows the same data as a contoured map of R_{free} and $-LL_{free}$, demonstrating that there is no significant correlation between W_{Biso} and $W_{X\text{-}ray}$. Taken together, these observations indicate that one iteration of the whole process (*i.e.* two sets of refinements to locate the optimal values of W_{Biso} and $W_{X\text{-}ray}$, respectively) should normally be sufficient.

Also in Table 1, it can be seen that the range of values of $RMSD_{bonds}$ determined by this procedure is between 0.0066 Å for the 2.55 Å 1w51 structure and 0.0114 Å for the 1.33 Å 2uv4 structure, in line with the earlier results of Tickle *et al.* (1998*a*). These values are also much more in line with those usually obtained using the recommended weighting scheme in the *CNS* program (Brünger *et al.*, 1998) and are significantly lower than the target values recommended by Jaskolski *et al.* (2007). This is of course not surprising since *CNS* uses cross-validation (though only by minimization of R_{free} with respect to $W_{X\text{-}ray}$) to determine the X-ray weight. It was also noted that in all cases the target RMSD of 0.02 Å suggested by the latter corresponded to significantly higher than optimal values of both R_{free} and $-LL_{free}$, thus potentially giving a lower phase accuracy and hence poorer map quality than the optimum. The optimal values of W_{Biso} and $W_{X\text{-}ray}$ are no doubt dependent both on the resolution cutoffs and on the overall isotropic displacement parameter (shown in Table 1; see Tickle *et al.*, 1998*a*, for details of calculation of the latter). Table 1 also shows that the optimized value of W_{Biso} was exactly 1.0 for 1w51, so in this case by chance the initial choices of SUs for bonded Δ_{Biso} were already optimal; however, in general this is unlikely to be true, hence it is important to optimize W_{Biso} before attempting to optimize $W_{X\text{-}ray}$, as is shown for the other three test cases.

The importance of optimizing W_{Biso} was demonstrated by fixing the IDP restraint weighting parameters at their default values for the *REFMAC* program, *i.e.* $W_{Biso} = 1$, $\sigma_{Biso} = 1.5$ and 3 Å^2 for main-chain and side-chain/ligand bonded atoms, respectively, and $\sigma_{Biso} = 2$ and 4.5 Å^2 similarly for angle-related atoms: many users of the software perhaps do not appreciate that these default values are totally arbitrary and certainly do not represent the optimal values for the majority of refinements. With the IDP restraint weights thus fixed, the $W_{X\text{-}ray}$ optimization runs with the 2uv4 data were repeated. The results are shown in Fig. 6: apparent optimum values for the RMSDs and RMSZs are obtained that are wildly different from the true optima found previously; in this example, it can be seen that the new minimum in LL_{free} occurs at $W_{X\text{-}ray} = 7.1$ (previously at 1.4) with $RMSD_{bonds} = 0.0298$ (was 0.0114) and $RMSZ_{bonds} = 1.25$ (was 0.47). Also, it can be seen that the minimum found for R_{free} is at an even greater value of $W_{X\text{-}ray}$ (around 15) with completely unrealistic corresponding values of $RMSD_{bonds}$ and $RMSZ_{bonds}$.

One further notable observation is that if one extrapolates the $RMSZ_{bonds}$ plot (Figs. 2 and 4) linearly to zero $W_{X\text{-}ray}$, the intercept value of $RMSZ_{bonds}$ is not zero as one might expect from Bayes' theorem; rather, it always appears to attain a limiting minimum value of about 0.2 (corresponding in the test cases used to $RMSD_{bonds} \simeq 0.004$ Å); similarly $RMSZ_{angles}$ attains a minimum value of about 0.25 ($RMSD_{angles} \simeq 0.6°$). A likely explanation is that rounding errors in the calculation of geometry derivatives or high parameter correlations prevent the conjugate-gradient optimization procedure from converging to the true minimum; a Newton optimization procedure using the full Hessian matrix with derivatives computed in double precision (Tickle *et al.*, 1998*a*) might help to resolve this problem.

[1] Supplementary material has been deposited in the IUCr electronic archive (Reference: GX5119). Services for accessing this material are described at the back of the journal.

This is supported by an observation that the limiting value obtained is dependent on the refinement software used even though the sets of ideal values used are identical: derivatives computed at higher precision should improve the convergence properties. It is also possible though unlikely that internal inconsistencies in the geometric restraints may contribute to the nonzero RMSDs; for example, the set of restraints for the bond angles at an atom or in a ring may not exactly obey the rules of three-dimensional geometry.

4. Discussion

The recommendations of Jaskolski *et al.* (2007) concerning relative weighting of diffraction and geometric terms appear to have arisen from a failure to make the important distinction between on the one hand the standard uncertainty (SU, σ) of an individual geometric restraint (or its RMS value), which is related to the weight of the restraint (weight = $1/\sigma^2$), and on the other the root-mean-square deviation (RMSD) of the values calculated from the parameters of the refined atomic model for a set of geometric restraints from their respective ideal values. The sole rationale for their recommendation of a target RMSD for bond lengths of around 0.02 Å (apparently independent of resolution cutoff, except at very high resolution, *i.e.* 1.5 Å or higher, where they suggest that the target RMSD may be increased to around 0.03 Å) is that this is the RMS value of the SUs, taken over the same set of ideal values for which the RMSDs have been calculated. They identify RMSDs several times smaller than the 'recommended library values' (*i.e.* presumably referring to the RMS values of the SUs) as a problem to be corrected. In reality no such problem exists because the RMS SU and the RMSD are in general only very loosely related quantities: the RMS SU is only the approximate upper limit of the RMSD, the lower limit being zero. The results presented here show that the apparently low values for the bond-length RMSD obtained by some refinement programs, particularly those that make optimal use of cross-validation statistics, are actually much closer to the optimal values based on an objective analysis of the data, although the authors' conclusion that there is much room for improvement of refinement software in this area is clearly apposite.

In the fitting of parameters to a functional model by maximization of the posterior probability (3), it is well known (Cruickshank, 1999, 2001) that the overall mean-square deviation is always a biased (under-) estimate of the overall mean-square SU value by the factor $N_{\rm dof}/N_{\rm obs}$, where $N_{\rm obs}$ is the total number of observations including restraints and $N_{\rm dof}$ is the number of degrees of freedom (= $N_{\rm obs} - N_{\rm var}$, where $N_{\rm var}$ is the number of fitted variable parameters). In the limit when $N_{\rm dof}$ tends to zero (*i.e.* when $N_{\rm obs}$ becomes less than $N_{\rm var}$ in the limiting case where there is no diffraction data), so will the RMSD, whereas of course the SUs remain constant at the values fixed *a priori*, independent of resolution cutoff. Furthermore, Tickle *et al.* (1998*a*) gave a formal proof that the mean-square deviation of any individual component of the posterior probability function [*e.g.* the bond-length restraint term in (3)] must also underestimate the mean-square SU value of that component by a factor that depends on the inverse of the Hessian (second derivative) matrix; however, a quantitative estimate of the magnitude of the effect requires calculation of the full Hessian matrix (*i.e.* including all components of the posterior probability function) and this is not yet a routine operation in macromolecular refinement.

The RMSD is not in any case the most appropriate statistic to use in this context, simply because as already indicated the SU values vary significantly according to the bond type so that deviations can only be sensibly compared or summed if divided first by the corresponding SU. More formally, the prior probability is a function of the ratio (deviation from ideal value)/SU (also known as '*Z* score' or 'normal score'), so the appropriate statistic to use is the root-mean-square *Z* score (RMSZ). This will have asymptotic values of zero (low resolution) and unity (high resolution) and will naturally vary monotonically between these limits as a function of resolution. Use of RMSZ in place of RMSD enables a meaningful comparison of statistics for structures containing different proportions of bond and angle types. Hooft *et al.* (1996) were the first to implement RMSZ for this purpose in their *WHATCHECK* structure-validation software; unfortunately, it is used there in a completely inappropriate manner by generating warning or error messages if an RMSZ value falls significantly below unity (*WHATCHECK* does claim that this is only a problem at high resolution).

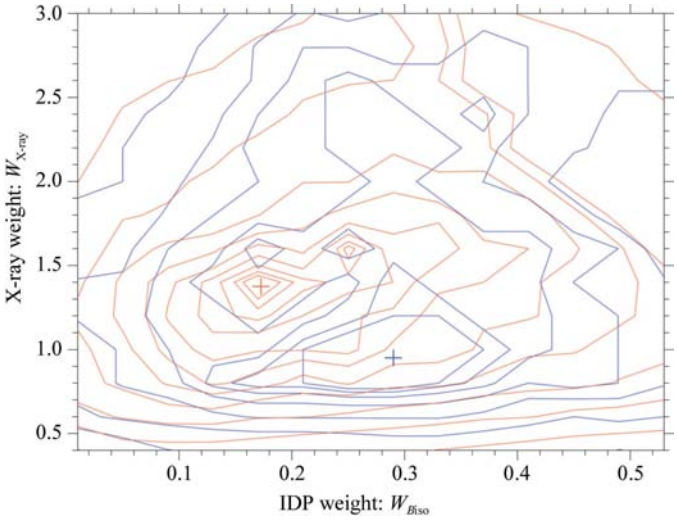

Figure 5
Contoured maps of $R_{\rm free}$ (blue contours) and $-{\rm LL}_{\rm free}$ (red) showing bivariate dependence on $W_{\rm Biso}$ and $W_{\rm X\text{-}ray}$ for the AMPK structure (PDB code 2uv4). For $R_{\rm free}$, the contour levels are 0.00010, 0.00016, 0.00025, 0.00040, 0.00063, 0.0010 and 0.0016 above the global minimum value (0.2071, marked with a blue cross at $W_{\rm Biso}$ = 0.29, $W_{\rm X\text{-}ray}$ = 0.95); for $-{\rm LL}_{\rm free}$ the contour levels are 0.10, 0.16, 0.25, 0.40, 0.63, 1.0, 1.6, 2.5, 4.0, 6.3 and 10 above the global minimum value (7421.9, marked with a red cross at $W_{\rm Biso}$ = 0.17, $W_{\rm X\text{-}ray}$ = 1.40).

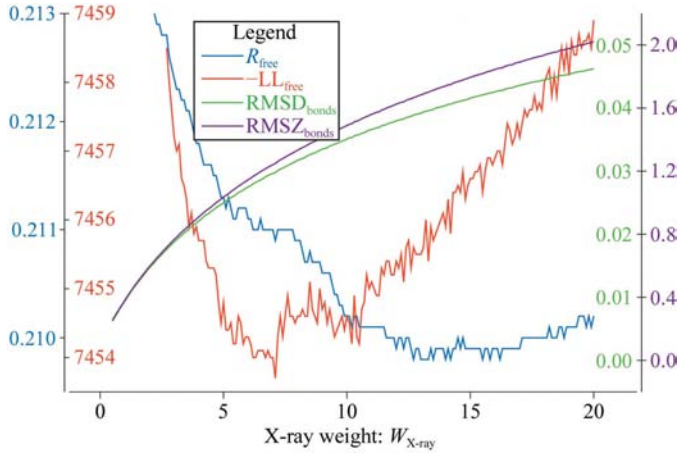

Figure 6
Plot of $R_{\rm free}$ (tic labels and line coloured blue), $-{\rm LL}_{\rm free}$ (red), ${\rm RMSD}_{\rm bonds}$ (green) and ${\rm RMSZ}_{\rm bonds}$ (purple) at convergence of refinement *versus* overall reflection weighting parameter $W_{\rm X\text{-}ray}$ for the AMPK structure (PDB code 2uv4), with all the IDP weighting parameters fixed at the *REFMAC* default values.

Another way of looking at the question of optimal RMSD or RMSZ values is in terms of Bayes' theorem, upon which of course restrained refinement crucially depends for its justification (Bricogne, 1997; Murshudov *et al.*, 1997; Blanc *et al.*, 2004). The restraints define the prior probabilities of the derived geometrical parameters (bonds, angles *etc.*); hence, at low resolution where the diffraction terms cease to significantly influence the local geometry the latter will be determined almost exclusively by the restraints and thus the optimal RMSD or RMSZ values should tend exactly to zero.

At low resolution, where the observation-to-parameter ratio barely exceeds unity (the count of 'observations' includes both restraints and X-ray data), the diffraction terms have little or no influence on the geometry, so in the absence of restraints the deviations of the calculated geometry values from the ideal ones would be large and essentially random. The goal of refinement must always be to obtain the most accurate possible model of the structure; this is a model with the minimum mean-square error (MSE), *i.e.* the minimum mean-square Z score of the calculated geometry values relative to the true (but unknowable) geometry values. In practice, this is achieved by maximizing both the posterior probability and the free likelihood: the first with respect to the parameters that determine the structure factors and the second with respect to the parameters defining various weights that determine the relative magnitudes of the summed contributions to the posterior probability function shown in (3) and which are not known with sufficient accuracy *a priori*, such as weights for ADP differences and X-ray terms.

Some insight can be gained by considering the relationship of the calculable (calculated − ideal) deviations to the hypothetical (true − ideal) deviations. If the (calculated − ideal) deviations are completely random then there will be no correlation with the (true − ideal) deviations. One cannot assume in general that the (calculated − ideal) deviations have even the same signs as the (true − ideal) deviations, let alone the same magnitudes. This is in essence the simplistic assumption that Jaskolski *et al.* (2007) are making when they equate RMSD with RMS SU. In the situation of zero correlation of the two sets of deviations, it can be shown (see Appendix A) that the desired minimum mean-square (calculated − true) Z score is obtained when the (calculated − ideal) Z scores are also all exactly zero, *i.e.* when RMSD and RMSZ are exactly zero.

Furthermore, it can be shown (Appendix A) in general that the optimal (calculated − ideal) RMSZ value of a given set of restraints should equal the correlation coefficient of the (calculated − ideal) and (true − ideal) Z scores. Of course, the value of this correlation coefficient can never actually be determined since it depends on the unknowable true geometry values and so it has no practical worth; however, the concept is useful because it helps to rationalize the experimental results obtained above. Hence, at very high resolution in the theoretical limit of infinite observation-to-parameter ratio, the diffraction terms will dominate and the deviations from the ideal values will accurately reflect the true deviations, so the correlation coefficient will be near its maximum value of 1 and the X-ray weight should be set so that the (calculated − ideal) RMSZ value should also be near 1. At intermediate resolutions the correlation between the (calculated − ideal) and (true − ideal) deviations will be less than perfect, *i.e.* some fraction of the calculated values will be on the wrong side of their corresponding ideal value relative to their true value, and so to minimize the effect of these errors the RMSZ value must be adjusted accordingly by an appropriate choice of weights. The correlation coefficient measures, on average, the fraction of the SU that equals the optimal value of the absolute (calculated − ideal) deviation.

In practice, the main reason for collecting data to very high resolution is to be able to fit more variable parameters (*e.g.* multipole and anharmonic displacement parameterizations *etc.*) in order to describe the structure more accurately; hence, it is likely that in practice even at very high resolution the optimal (calculated − ideal) RMSZ values will still be somewhat less than the theoretical asymptotic value of unity and therefore the optimal RMSD values will correspondingly be less than the asymptotic value given approximately by the RMS SU. One cannot predict what the optimal RMSD value at a given resolution will be without knowing something about the observation-to-parameter ratio for the particular parameterization in use; of course there may well be an optimal parameterization defined by the maximum of the log(free likelihood).

There are some additional effects that Jaskolski *et al.* (2007) have not taken into account in their analysis: the SUs of the ideal values obtained from the CSD data are just the sample standard deviations (Engh & Huber, 1991, 2001). As they point out, this will include real variations arising from the chemical environment and from random sampling errors in cases where only a limited number of samples of a given bond or bond-angle type are available. However, there will also be contributions arising from the experimental errors of the diffraction measurements obtained from the crystals of the small molecules (Evans, 2007) and also from the neglect of libration corrections for the geometry calculated from the CSD coordinates. For bond lengths in small-molecule structures the experimental SUs are typically in the range 0.005–0.01 Å, *i.e.* somewhat smaller than the range of sample standard deviations (0.01–0.06 Å); the contribution to the uncertainty from neglect of libration corrections is impossible to quantify, if only because the CSD does not hold the atomic displacement parameters needed to compute these corrections. In most cases the experimental error will probably be a relatively small contribution, but in some particular cases it could be significant. The overall effect will be to further increase the random error and thus reduce the correlation of (calculated − ideal) and (true − ideal) Z scores; hence, the theoretical result proved in Appendix A allows us to conclude that the effect of this will be to lead to lower optimal RMSZs or RMSDs than would otherwise be expected.

The third main proposal of Jaskolski *et al.* (2007) concerns the use of variable restraint weights dependent on the ADP values and the occupancies, the intention being on the one hand to reduce the large deviations from ideality often observed in mobile regions and on the other to allow increased deviations for the well defined regions of the structure. In the former case the problem is that the large deviations are in many cases caused by libration; the bond-length deviations arise because X-ray diffraction always determines the time- and lattice-averaged structure, for which the geometry may differ significantly from the instantaneous structure, so that in these cases larger deviations than normal from ideal geometry are entirely to be expected. Hence, restraining them tightly to the ideal values would be inconsistent with the diffraction data; however, the effect is only likely to be significant for the more flexible parts of the structure such as loops and side chains. The formally correct procedure would be to use the libration-corrected calculated geometry values in (3). However, this is very difficult to deal with in practice because correction for this systematic bias in the bond lengths can only be reliably (and then only partially) performed in the case of the completely correlated curvilinear atomic motion (Busing & Levy, 1964) of large pseudo-rigid groups such as secondary-structure elements (*e.g.* helices) and whole domains and also, if high-resolution data are available, for the smaller rigid planar groups in side chains of *e.g.* His, Phe, Tyr and Trp. The correction for libration in the case of extended pseudo-rigid groups is generally not found to be significant

because the correction factor is proportional to the square of the amplitude of libration, which for large groups of atoms tends to be severely limited by nonbonded contacts, the lateral displacement of an atom being equal to the product of the libration amplitude and the distance of the atom from the centre of motion.

Hence, correction for libration of the geometry of small flexible groups is generally not feasible even for small molecules, let alone macromolecules, and so the next best alternative is to make an allowance for the expected increased uncertainty of the observed geometry values relative to their ideal values. The effect is somewhat ameliorated because the ideal geometry values obtained from the CSD coordinates are not corrected for libration either (it is very unfortunate that the ADP values that would be needed to repeat these corrections are not held in the CSD); nevertheless, because typical values of ADPs and hence of libration amplitudes are generally much higher for macromolecules than for small molecules, one should not be tempted to conclude that the effects cancel! The implication therefore is that the SUs of bond lengths and angles should if anything be increased to reflect this greater uncertainty in the ideal geometry values, thus requiring that the individual geometry restraint weights in the more flexible regions be reduced and thus contradicting the suggestion by Jaskolski *et al.* (2007) to *increase* the geometry restraint weights relatively in the flexible regions. The authors do point out that where the geometry is poorly determined by the X-ray data owing to a low contribution to the scattering at high resolution the restraint terms will dominate and the refined geometry will tend to reproduce the ideal geometry both in value and in SU; however, in such cases the standard geometry weights will be sufficient to maintain the geometry and there is no reason to increase their values above the defaults.

It may be feasible to deal with the effects of libration in a completely objective manner, as has been demonstrated above for the other weighting parameters, by parameterizing the geometry weights as a function of the ADPs of the atoms involved and then determining the weighting parameters experimentally, *i.e.* without making arbitrary assumptions about their values, although an *ad hoc* functional form for the parameterization would still have to be postulated.

Jaskolski *et al.* (2007) also argue for reduced restraint weights in the well determined regions on the grounds that high restraint weights may bias some geometry values that truly differ from the ideal values owing to an unusual chemical environment. The problem with this suggestion is that clearly one does not know *a priori* which are the normal geometries and which are the abnormal ones and so the weights have to be reduced equally for all; however, the theory presented above shows that reducing the overall level of restraint weights in well determined regions is performed at the cost of introducing unwanted random errors into the vast majority of normal geometry values in these regions, since for these cases the optimal weights are still those that are based on the standard SU values. In any case, truly unusual geometries will stand out best as outliers if the random errors in the majority of normal geometry values are kept to a minimum: increasing the overall random error can only serve to reduce the signal-to-noise ratio for the aberrant geometry. This requires that the standard weights are used throughout.

APPENDIX *A*

The aim is to show that the optimal choices for the (calculated − ideal) RMSZ (and hence RMSD) values are equal to the respective correlation coefficients of the (calculated − ideal) and (true − ideal) Z scores for the 'most correct' model of the structure, *i.e.* one that has the minimum mean-square normalized error of the geometry values.

For simplicity, all the calculated, ideal and true geometry values (*e.g.* bond length or angle) are first normalized, *i.e.* divided by the SU of the corresponding ideal value (*e.g.* those obtained as the sample standard deviations of the values extracted from the CSD). Cruickshank (1999, 2001) and Parisini *et al.* (1999) showed that the SU of the calculated value will always be less than but close to that of the ideal value, except for very high resolution data, where the effect of the data is reduce the SU significantly below that of the ideal value, so we can assume that in most cases the SUs of the calculated and ideal geometry values will be approximately equal.

Let g be a normalized calculated geometry value and let $g_{\rm ideal}$ and $g_{\rm true}$ be the corresponding normalized ideal and true geometry values, the former being a known constant and the latter constant but unknowable. Then let ζ be the desired (calculated − ideal) RMSZ value

$$\zeta = [\langle (g - g_{\rm ideal})^2 \rangle]^{1/2}, \tag{5}$$

where $\langle \ldots \rangle$ implies the arithmetic mean throughout.

The maximum absolute calculated deviation $|g - g_{\rm ideal}|$ will occur when $\zeta = 1$; let the value of g in this situation be g',

$$[\langle (g' - g_{\rm ideal})^2 \rangle]^{1/2} = 1. \tag{6}$$

Therefore, ζ is the scale factor which scales the maximum values (in absolute terms) of the calculated deviations down to the actual values of the calculated deviations,

$$(g - g_{\rm ideal}) = \zeta (g' - g_{\rm ideal}) \tag{7}$$

or

$$g = g_{\rm ideal} + \zeta (g' - g_{\rm ideal}). \tag{8}$$

The assumption here is that terms higher than first order in ζ in the Taylor series expansion may be ignored.

The mean-square error MSE is the mean-square deviation of the calculated geometry value from the true value,

$$\begin{aligned} {\rm MSE} &= \langle (g - g_{\rm true})^2 \rangle \\ &= \langle [g_{\rm ideal} + \zeta (g' - g_{\rm ideal}) - g_{\rm true}]^2 \rangle \\ &= \langle (g_{\rm ideal} - g_{\rm true})^2 \rangle + 2\zeta \langle (g_{\rm ideal} - g_{\rm true})(g' - g_{\rm ideal}) \rangle \\ &\quad + \zeta^2 \langle (g' - g_{\rm ideal})^2 \rangle \\ &= \langle (g_{\rm true} - g_{\rm ideal})^2 \rangle - 2\zeta\rho + \zeta^2, \end{aligned} \tag{9}$$

where (9) is simplified by substitution first from (8) and then from (6) and where ρ is the correlation coefficient of the (calculated − ideal) and (true − ideal) Z scores given by

$$\rho = \langle (g' - g_{\rm ideal})(g_{\rm true} - g_{\rm ideal}) \rangle. \tag{10}$$

We wish to find the value of ζ that minimizes the MSE given by (9), so differentiating this with respect to ζ and equating to zero we obtain the very simple result

$$\zeta = \rho, \tag{11}$$

i.e. the optimal value of the RMSZ ζ is just the correlation coefficient ρ. This means that when there is no correlation between the (calculated − ideal) and (true − ideal) Z scores in the limiting case of no X-ray data then $\zeta = \rho = 0$ and when there is complete correlation in the limiting case of infinite X-ray data then $\zeta = \rho = 1$. At intermediate resolutions ρ will lie somewhere between 0 and 1 and the optimal RMSZ (and hence the optimal RMSD) value will be determined accordingly.

I should like to thank Dr Tom Davies (Astex) for critical reading of the manuscript and the referees for constructive comments.

References

Blanc, E., Roversi, P., Vonrhein, C., Flensburg, C., Lea, S. M. & Bricogne, G. (2004). *Acta Cryst.* D**60**, 2210–2221.

Bricogne, G. (1997). *Methods Enzymol.* **276**, 361–423.

Brünger, A. T. (1992). *Nature (London)*, **355**, 472–475.

Brünger, A. T. (1993). *Acta Cryst.* D**49**, 24–36.

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

Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.

Busing, W. R. & Levy, H. A. (1964). *Acta Cryst.* **17**, 142–146.

Cruickshank, D. W. J. (1999). *Acta Cryst.* D**55**, 583–601.

Cruickshank, D. W. J. (2001). *International Tables for Crystallography*, Vol. *F*, edited by M. G. Rossmann & E. Arnold, pp. 403–414. Dordrecht: Kluwer Academic Publishers.

Czjzek, M., Cicek, M., Zamboni, V., Bevan, D. R., Henrissat, B. & Esen, A. (2000). *Proc. Natl Acad. Sci. USA*, **97**, 13555–13560.

Day, P., Sharff, A., Parra, L., Cleasby, A., Williams, M., Hörer, S., Nar, H., Redemann, N., Tickle, I. & Yon, J. (2007). *Acta Cryst.* D**63**, 587–596.

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.

Evans, P. R. (2007). *Acta Cryst.* D**63**, 58–61.

Hooft, R. W. W., Vriend, G., Sander, C. & Abola, E. E. (1996). *Nature (London)*, **381**, 272.

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

Kleywegt, G. J. & Brünger, A. T. (1996). *Structure*, **4**, 897–904.

Lebedev, A. A., Tickle, I. J., Laskowski, R. A. & Moss, D. S. (2003). *Acta Cryst.* D**59**, 1557–1566.

MacArthur, M. W. & Thornton, J. M. (1996). *J. Mol. Biol.* **264**, 1180–1195.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.

Pannu, N. S. & Read, R. J. (1996). *Acta Cryst.* A**52**, 659–668.

Parisini, E., Capozzi, F., Lubini, P., Lamzin, V., Luchinat, C. & Sheldrick, G. M. (1999). *Acta Cryst.* D**55**, 1773–1784.

Patel, S., Vuillard, L., Cleasby, A., Murray, C. W. & Yon, J. (2004). *J. Mol. Biol.* **343**, 407–416.

Tickle, I. J., Laskowski, R. A. & Moss, D. S. (1998*a*). *Acta Cryst.* D**54**, 243–252.

Tickle, I. J., Laskowski, R. A. & Moss, D. S. (1998*b*). *Acta Cryst.* D**54**, 547–557.