Acta Crystallographica Section F
Structural Biology and Crystallization Communications
ISSN 1744-3091

Detection and analysis of unusual features in the structural model and structure-factor data of a birch pollen allergen

Bernhard Rupp

k.-k. Hofkristallamt, San Marcos, CA 92978, USA

Correspondence e-mail: br@hofkristallamt.org

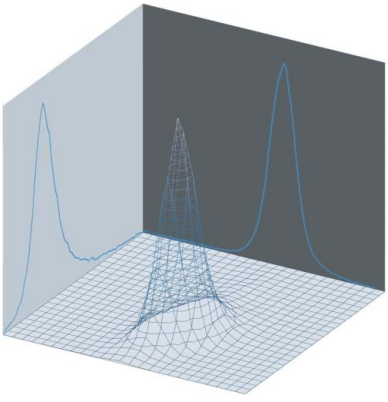
Received 12 January 2012
Accepted 24 February 2012

Physically improbable features in the model of the birch pollen structure Bet v 1d (PDB entry 3k78) are faithfully reproduced in electron density generated with the deposited structure factors, but these structure factors themselves exhibit properties that are characteristic of data calculated from a simple model and are inconsistent with the data and error model obtained through experimental measurements. The refinement of the 3k78 model against these structure factors leads to an isomorphous structure different from the deposited model with an implausibly small R value (0.019). The abnormal refinement is compared with normal refinement of an isomorphous variant structure of Bet v 1l (PDB entry 1fm4). A variety of analytical tools, including the application of Diederichs plots, $R\sigma$ plots and bulk-solvent analysis are discussed as promising aids in validation. The examination of the Bet v 1d structure also cautions against the practice of indicating poorly defined protein chain residues through zero occupancies. The recommendation to preserve diffraction images is amplified.

1. Introduction

During a routine search of the public *PDB_REDO* database (Joosten *et al.*, 2011) for a crystal structure model of birch pollen protein Bet v 1, a significant discrepancy between the originally reported R values (R_{free} = 0.298, R_{work} = 0.274) and the conservatively re-refined structure of PDB entry 3k78 (Bet v 1d) was detected (0.177, 0.126). These R values are unexpectedly low for a 2.8 Å structure. At the same time, the electron-density map provided by the Uppsala Electron Density Server, EDS (Kleywegt *et al.*, 2004), publicly accessible through the PDBe (Velankar *et al.*, 2010), shows numerous side chains that do not fit the experimental electron density. The EDS service also reported a negative bulk-solvent contribution B factor and a negligibly small bulk-solvent contribution scale factor, which is abnormal for an experimentally determined protein structure (Fokine & Urzhumtsev, 2002). Given the fact that the R values calculated by *PDB_REDO* from the data without refinement (0.265, 0.275; a new R_{free} set was calculated by *PDB_REDO*) agreed reasonably well with the values reported in the PDB header (0.298, 0.273), an accidental swap of experimentally observed structure factors F(obs) against the final calculated structure factors F(calc) when generating the deposited structure-factor file can be excluded (in that case also the reproduced R values without refinement would be improbably low). In view of these discrepancies it seemed sensible to re-examine the 3k78 model and the associated deposited diffraction data.

The crystal structure model of birch pollen hypoallergen Bet v 1d (Zaborsky *et al.*, 2010), PDB code 3k78, was reported as solved by molecular replacement (MR) from the nearly sequence identical model of the hypoallergenic isoform Bet v 1l (Marković-Housley *et al.*, 2003), PDB entry 1fm4. The model structures are isomorphous ($P2_1$) with cell constants identical within experimental error. 1fm4 itself was derived by MR from the $C222_1$ structure model of the clinically important inhalant major allergen, Bet v 1a (Gajhede *et al.*, 1996; PDB entry 1bv1). A sequence alignment including additional information relevant to the following discussion is provided in Fig. 1.



© 2012 International Union of Crystallography
All rights reserved

```
BETV1A P15494 1BV1 MGVFNYETET TSVIPAARLF KAFILDGDNL FPKVAPQAIS SVENIEGNGG PGTIKKISFP  60
BETV1L P43185 1FM4 MGVFNYETEA TSVIPAARMF KAFILDGDKL VPKVAPQAIS SVENIEGNGG PGTIKKINFP  60
BETV1D P43177 3K78 MGVFNYEIET TSVIPAARLF KAFILDGDNL VPKVAPQAIS SVENIEGNGG PGTIKKINFP  60
                   ******* *: ********:* *********:* .********* ********** *******.**

BETV1A P15494 1BV1 EGFPFKYVKD RVDEVDHTNF KYNYSVIEGG PIGDTLEKIS NEIKIVATPD GGSILKISNK 120
BETV1L P43185 1FM4 EGFPFKYVKD RVDEVDHTNF KYNYSVIEGG PVGDTLEKIS NEIKIVATPD GGCVLKISNK 120
BETV1D P43177 3K78 EGFPFKYVKD RVDEVDHTNF KYNYSVIEGG PVGDTLEKIS NEIKIVATPD GGCVLKISNK 120
                   *****.**** ********** *********** *:******** ***.****** **.:******

BETV1A P15494 1BV1 YHTKGDHEVK AEQVKASKEM GETLLRAVES YLLAHSDAYN  160
BETV1L P43185 1FM4 YHTKGNHEVK AEQVKASKEM GETLLRAVES YLLAHSDAYN  160
BETV1D P43177 3K78 YHTKGNHEVK AEQVKASKEM GETLLRAVES YLLAHSDAYN  160
                   *****:***. *.***.**.** ********** **********
```

Figure 1
Sequence alignment of Bet v 1 allergens. The yellow codes indicate sequence differences between search model 1fm4 and 3k78, while the red highlights indicate nine residues that contain zero occupancy atoms in both models, 1fm4 and 3k78, although at different atoms as detailed in the text and summarized in Fig. 8. Alignment by *ClustalW* (Larkin *et al.*, 2007).

The 3k78 model was refined against structure factors with 2.8 Å resolution, and 1fm4 was refined at 2.0 Å. Both structures appear unremarkable (in a technical sense, no insult to biological relevance intended), and the refinement statistics and protocols reported in the PDB entries are appropriate for the resolution. However, on closer inspection, both the model and the structure-factor data of 3k78 exhibit highly unlikely, physically improbable (if not impossible) features. For reference, the results of the 3k78 analysis and re-refinement are compared with those obtained for the isomorphous 1fm4 structure of good and reproducible quality. This comparison may provide useful reference for the aspiring crystallographer and can serve as teaching material.

2. Structure models and re-refinement

The two models were originally refined using different programs, *CNS* 1.0 (Brünger *et al.*, 1998), and *REFMAC*5 (Murshudov *et al.*, 1997, 2011; Winn *et al.*, 2001), with different refinement protocols. To aid comparison, a common isotropic *B*-factor refinement protocol with *REFMAC* was used in both cases, with parameters adjusted appropriate to each refinement.

The mmCIF structure-factor files and PDB coordinate files were downloaded from the PDBe (Velankar *et al.*, 2010). Structure-factor files were converted into mtz files using the programs of the *CCP*4 suite (Winn, 2003; Winn *et al.*, 2011) through the *CCP*4*i* user interface (Potterton *et al.*, 2003). The original $R_{\rm free}$ data sets were kept (except in an additional refinement of 3k78 for graphing purposes discussed in §3). Original maximum-likelihood maps were computed *via REFMAC* (zero cycles) with automated weighting from original coordinates and structure factors, and in case of 3k78 also the TLS parameters were read in from the deposited coordinate file. The procedures for analysis of the structure-factor data are provided in §3.

The common *REFMAC* protocol included isotropic individual *B* factors, flat bulk-solvent model (Jiang & Brünger, 1994), and riding H atoms were used in these refinements. The *REFMAC* X-ray matrix weight (Murshudov *et al.*, 2011) and *B*-factor restraint weights were manually adjusted by monitoring the negative cross-validation log-likelihood ($-{\rm LL}_{\rm free}$) minimum at convergence (Tickle, 2007).

2.1. Coordinates and model 1fm4

The coordinate file of the Bet v 1l search model, 1fm4, reveals no unusual features. The PDB file contains residues 2–160 of the sequence, but the residue numbers in the coordinate file are decremented by 1 compared to the aligned sequences in Fig. 1. As specified in REMARK 480, occupancies for the surface exposed, terminal side-chain atoms of Lys28, Lys65, Lys80, Lys103, Lys129, Glu131, Gln132, Lys134 and Lys137 are set to zero (§4, Fig. 8). Zero side-chain occupancies usually indicate that the side chains were poorly defined in electron density owing to displacement such as disorder or multiple conformations, and instead of accepting the correspondingly high displacement parameters or *B* factors from the refinement, the occupancies of such atoms are manually set to zero. While still common practice, such is not necessarily the best way to indicate the limited knowledge of their actual position (*c.f.* discussion in §4).

2.2. Re-refinement of 1fm4

Progress in the methodology of macromolecular refinement has led to steady improvements of the programs, and major efforts to re-refine already deposited PDB models have been undertaken in the *PDB_REDO* effort (Joosten *et al.*, 2011). In this work, the purpose of re-refining the already good 1fm4 structure is not to generate a better model (which ultimately would also require some minor rebuilding) but to provide a benchmark for the applied procedure and an example of the characteristics of a well refined model in order to appreciate the abnormal refinement of 3k78.

1fm4 was already well refined with *CNS*1.0 about a decade ago. During the multiple weight adjustment runs *REFMAC* reached stable convergence after about 30 cycles, with a resolution-typical X-ray matrix weight of 0.2 and restraint weight σs for *B*-factor main-chain 1–2, 1–3 neighbors and side-chain 1–2, 1–3 neighbors adjusted to 3, 5, 7 and 9 Å^2, which is reasonable given the empirical values (Tronrud, 1996). The re-refined *REFMAC* model differs very little from the original model. The overall coordinate r.m.s.d. between models on all atoms is 0.247 Å and on Cα is 0.078 Å, which is well below the historic value for 100% sequence identity expected from the Chothia and Lesk function (Chothia & Lesk, 1986). No significant geometry improvements resulted during re-refinement, and both 1fm4 and its re-refined model are of good quality. No attempts at model rebuilding were made, which probably could close the slightly increased R–$R_{\rm free}$ gap (Tickle *et al.*, 1998*b*, 2000) compared with the original refinement. A subset of refinement statistics relevant to the structure comparison are compiled in Table 1. Considering the different programs (*CNS*1.0 *versus REFMAC*5.6), the differences in protocol, as well as different X-ray and restraint weight optimization, this result is quite reassuring and attests to the reproducibility of crystallographic refinement.

The *B* factors of the previously 'unoccupied' side-chain atoms with reset occupancy refined as expected to high *B* factors, and the inspection of the electron density of these residues in *COOT* (Emsley *et al.*, 2010) shows the corresponding and increasing weakening of density along the side-chain terminals (§4, Fig. 9). Apart from polishing the model '*ad tedium*' (the term originally being coined by Phil Evans), the well refined 1fm4 model remains fully valid even

Table 1
Selected refinement statistics.

Statistics for 1fm4 and its re-refinement are normal. The values highlighted in bold for the 3k78 re-refinements are unusual or highly improbable given the 2.8 Å resolution. They include too low overall *B* factor; no bulk-solvent contributions; absurdly low *R* values; near perfect correlation between observed and calculated structure factors; and atypically high *REFMAC* X-ray matrix weights. n.r. not reported.

	1fm4 deposited	1fm4 re-refined isotropic *B*	3k78 deposited, hybrid TLS	3k78 re-refined, hybrid TLS	3k78 re-refined, isotropic *B*
Space group	$P2_1$	$P2_1$	$P2_1$	$P2_1$	$P2_1$
a (Å)	33.13	33.13	32.97	32.97	32.97
b (Å)	57.23	57.23	57.01	57.01	57.01
c (Å)	38.65	38.65	38.93	38.93	38.93
β (°)	91.94	91.94	92.27	92.27	92.27
Resolution (Å)	28.66–1.97	28.66–1.99†	32.95‡–2.80	25.56–-2.80	25.56–-2.80
Last resolution shell (Å)	2.09–1.97	2.04–1.99	2.87–2.80	2.87–2.80	2.87–2.80
No. of reflections	9658	8659	3184	3184	3184
Atoms of zero occupancy	29	0	29	29	29 at 0.01
Refinement program	*CNS* 1.0	*REFMAC* 5.6.0117	*REFMAC* 5.2.0019 TLS	*REFMAC* 5.6.0117 TLS	*REFMAC* 5.6.0117
Riding H atoms	n.r.	Yes	No	Yes	No
R_{free} set	10% random	10% random	9.8% random§	4.8% random	4.8% random¶
B Wilson (Å^2)	12.2	18.9	45.2	27.6††	27.6††
B mean overall (Å^2)	16.3	18.7	26.8	**3.67**‡‡	15.2
*B*_sol (Å^2), *k*_sol	66.1, n.r.	24.0, 0.37	−10.00, 0.01§§	**−10.00, 0.03**	**No bulk solvent**
R_{free}, overall (last shell)	0.240 (0.388)	0.213 (0.400)	0.298 (0.387)	**0.132 (0.250)**	**0.040 (0.062)**
R-work, overall (last shell)	0.197(0.359)	0.159(0.234)	0.273(0.350)	**0.069 (0.105)**	**0.019(0.048)**
Coordinate e.s.u. from R_{free} (Å)	0.160	0.187	0.379	0.235	**0.072**
Correlation between F_c and F_o		0.962	0.934	0.993	**0.999**
Correlation, F_c and F_o free		0.929	0.919	0.968	**0.997**
Ramachandran regions % (*COOT*)	97.5/2.5/0	97.5/2.5/0	92.2/2.0/5.8	92.2/2.0/5.8	91.0/6.5/2.6
R.m.s.d. bonds (Å)	0.009	0.011	0.017¶¶	0.015	0.011
R.m.s.d. angles (°)	1.30	1.62	1.54¶¶	1.82	1.69
R.m.s.d. all atoms (Å)	0.247	0.247	0.705†††	0.705†††	0.640†††
R.m.s.d. main chain (Å)	0.081	0.081	0.352†††	0.352†††	0.367†††
R.m.s.d. Cα (Å)	0.078	0.078	0.295†††	0.295†††	0.302†††
X-ray term matrix weight‡‡‡	n.r.	0.2	n.r.	Default	**0.6**
B-factor restraint weight§§§ (Å^2)	n.r.	3/5/7/9	n.r.	Default	5/7/9/11

† Deposited data extend only to 1.99 Å. ‡ This is a reporting error in the PDB header caused by *REFMAC*. Actual low resolution limit is 25.56 Å. § The deposited structure-factor file contains only a 5% cross-validation data set. ¶ A 10% *a posteriori* cross-validation set gives practically the same result. †† From *TRUNCATE*. ‡‡ Residual *B* factors, some atoms show the low *B*-factor cutoff of 2.0 Å^2. §§ From the EDS report. ¶¶ Not including the zero occupancy residues. With zero occupancy residues reset, 0.032 Å and 2.136°. ††† R.m.s.d. against the original 3k78 model. ‡‡‡ In *REFMAC*, the actual X-ray term weight (W_a in *CNS*/*X-PLOR*) is obtained as the product of the user-selectable X-ray matrix weight times the ratio of the trace of the geometry Hessian divided by the trace of the X-ray Hessian matrix. The *REFMAC* X-ray matrix weight is therefore not the same as W_a. Ian Tickle has kindly pointed me to the respective *REFMAC* source code for verification. §§§ *REFMAC B*-factor restraint weight σs (Å^2), for main-chain 1–2, 1–3 neighbors, and side-chain 1–2, 1–3 neighbors.

under different refinement protocols executed nearly a decade later. As stated above, setting the occupancies of side-chain atoms of residues with weak density to zero seems to be unnecessary and could probably be avoided.

2.3. Coordinates and model 3k78

Although the 3k78 Bet v 1d model has five backbone torsion angle outliers and numerous severe geometry deviations in the residues with zero occupancy atoms, it is otherwise unremarkable. The coordinate file of 3k78 contains residues 3–159 of the sequence, with the residue numbers matching the sequence alignment in Fig. 1 (*i.e.* incremented from 1fm4 by 1). However, for the residues containing zero occupancy atoms (Asn29, Lys66, Lys81, Lys104, Lys130, Glu132, Gln133, Lys135 and Lys138) an interesting pattern emerges: the zero occupancies are systematically shifted in atom number to lower values, *i.e.* it is not the terminal side-chain atoms that are unoccupied, but the zero occupancies move towards the Cβ, and even to the (in the PDB file but not physically) adjacent backbone O atoms of the respective residue, while the terminal atoms of the residues become occupied again (§4, Fig. 8). This pattern is physically highly improbable, but no explanation for this selection of zero occupancy atoms has been reported. These physically improbable model features do, however, lead to some interesting features in the electron density of the original refinement (§4, Fig. 9). The substantial bond distance deviations of most of the residues with zero occupancy atoms are listed in §4, Fig. 10. The remaining deviations can be found in the 3k78 PDB header REMARK 500 records or may be generated with RUN500 from *CCP4i*.

2.4. Original refinement of 3k78

The model was originally refined using the *REFMAC* hybrid TLS–isotropic *B*-factor refinement (Painter & Merritt 2006; Murshudov *et al.*, 2011) with a single TLS group. Given the 2.8 Å resolution, hybrid TLS refinement would not be unusual or unreasonable, although a rationale for the choice of protocol, parameterization, and analysis of the (small) TLS contributions is absent (Zaborsky *et al.*, 2010). Original density maps were calculated from unchanged deposited data and coordinates *via* a zero cycle refinement run in *REFMAC* (including the published TLS groups and matrices). The resulting *R* values (0.304, 0.269) were in reasonable agreement with those reported in the PDB header (0.298, 0.273) and by *PDB_REDO* (0.265, 0.275).

When the original coordinate file is loaded into *COOT* (Emsley *et al.*, 2010), difference density peaks > 5σ clearly indicate that several residues such as Ile8, Gln37, Glu43, Gly52, Lys56, Glu61, Arg71, Asp110, Glu128, Tyr151 and His155 should be modeled with different conformations (Fig. 2), in agreement with the findings of the EDS service (Kleywegt *et al.*, 2004) which can be readily accessed *via* the PDB validation links. While such modeling errors are not unusual, they can easily be corrected. There was no support for the claim of unidentified density in the core of the molecule made in the 3k78 publication (Zaborsky *et al.*, 2010). Instead, two chemically plausible water molecules included in the model can be discerned in the

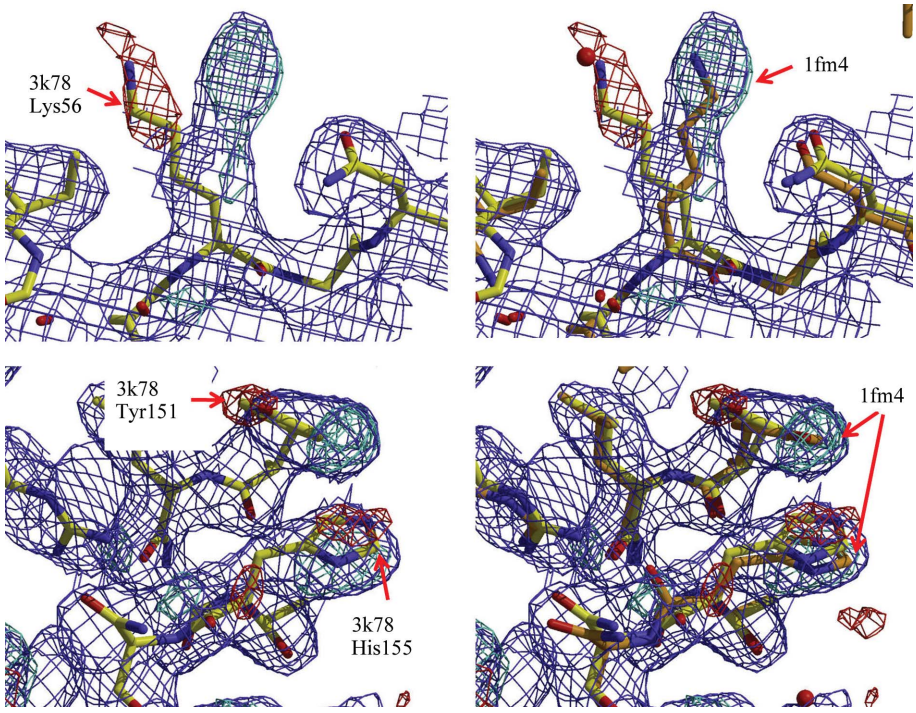


Figure 2
Electron density of original 3k78 model. $2mF_o - DF_c$ electron density contoured at 0.8σ (blue), 5σ $mF_o - DF_c$ difference density (positive light green, negative red). The left panel shows the misplaced residues in the original 3k78 model (yellow carbon stick model) and in the original electron density, reconstructed as described in the text. No refinement has been conducted, but the correct placement of the residues can be easily recognized. The right panels show the same electron density, but now additionally with the starting model 1fm4 (not a re-refined 3k78 model) loaded into *COOT*. The starting model 1fm4 (orange carbon stick model) fits the electron density better than the deposited model, which indicates that the 3k78 model has not been properly refined (or that the structure factors do not match the model).

electron density. Given the relatively high R values and poor geometry of the side chains with zero occupancy atoms in the published model, rebuilding and re-refinement of 3k78 appeared promising.

2.5. Isotropic *B*-factor refinement of 3k78

The original 3k78 coordinates were used without rebuilding (only the zero occupancies were reset to 0.01) for isotropic B-factor refinement. Initially a resolution-appropriate low X-ray matrix weight of 0.1 was used to keep the geometry tight and repair the originally distorted zero-occupancy residues. The same B-factor restraint weights as for 1fm4 (3/5/7/9 Å^2) were used for 30 cycles. The refinement did not reach convergence, but the R values already dropped unexpectedly quickly to 0.131 and 0.068. Inspection of the model geometry showed that the model overall had in fact improved, and maps showed that the misplaced residues Ile8, Gln37, Glu43, Gly52, Lys56, Glu61, Arg71, Asp110, Glu128, Tyr151 and His155 all had assumed correct positions practically identical to those in 1fm4 with good geometry in the remarkably noiseless density map. Nine water atoms from 1fm4 that also occupied density in the 3k78 map were added to the new model by a simple cut and paste.

At that point of the refinement the R values had already reached values typical for atomic resolution structures. Given the negative bulk-solvent B factor of -10 Å^2 and small bulk-solvent scale factor of 0.026 e$^-$ Å^{-3}, no sensible bulk-solvent scattering contribution seemed to be present, and the assumption of calculated structure factors was made. As a consequence, (*a*) the bulk-solvent correction was turned off, (*b*) no riding H atoms were included, (*c*) X-ray matrix weights were increased to 0.6, (*d*) B-factor restraint weights were loosened up to their physically reasonable limit (5/7/9/11 Å^2) as established by empirical values (Tronrud, 1996).

The refinement, with its atypical protocol for any experimental protein structure, reached stable convergence at R values of 0.040 and 0.019, with stable geometry and practically the same target r.m.s.d. values as 1fm4 (Table 1). The resulting density maps were practically noiseless, with the only remaining significant difference density features in the vicinity of the residues with unoccupied side-chain atoms. According to *PROCHECK* (Laskowski, 2001) or RUN500, the entire model had excellent geometry quality. Tedium was declared and no manual rebuilding of the side chains with unoccupied atoms was attempted.

At this point it was clearly established that (*a*) the deposited structure factors are calculated structure factors, (*b*) the resulting re-refined model resembles in most details the mutated search model, (*c*) that the original model has not, or not properly, been refined against these structure factors (or had been altered from a model essentially similar to the re-refined model and after the structure factors had been calculated).

3. Analysis of structure factors

Given the highly improbable refinement results inconsistent with experimental data at 2.8 Å resolution, a closer examination of the deposited structure-factor data was undertaken.

3.1. Intensity statistics and *R*-value analysis

The data for 1fm4 and for 3k78 were collected in-house on rotating anode sources and recorded on imaging plate detectors, with reported

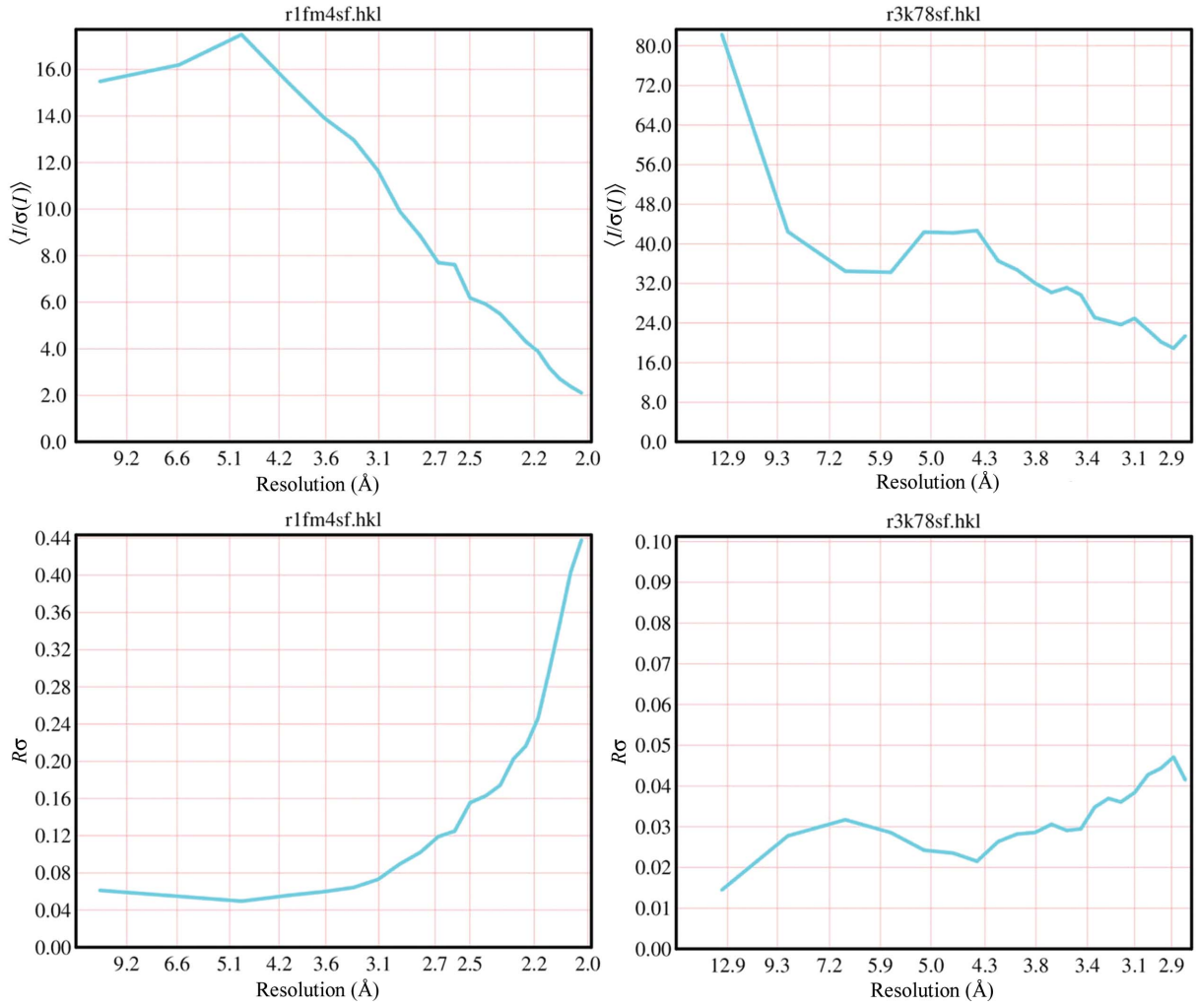


Figure 3
Mean $I/\sigma(I)$ and $R\sigma$ *versus* resolution for 1fm4 and 3k78. The left column shows what can be considered representative statistics for experimental diffraction data (1fm4). The $I/\sigma(I)$ *versus* resolution graphs generally reproduce the trend of the Wilson plots, which are readily available *via TRUNCATE* from the *CCP*4 suite. Note for 3k78 (left column) the abnormally high values of $I/\sigma(I)$ as well as the sharp increase at low resolution, normally not observed with protein structures containing bulk-solvent contributions which supress the strong high-resolution scattering contributions. In the second row, 1fm4 intensities display the normal increase of $R\sigma$ *versus* resolution, and its values are representative of what is expected for a data set that is useful to a mean $I/\sigma(I)$ level of about 2.0 in the highest resolution shell. 3k78 data in contrast show absurdly low values for $R\sigma$ corresponding to the extremely high mean $I/\sigma(I)$ values, with a mean $I/\sigma(I)$ of over 20 in the last resolution shell (*c.f.* Table 2). Figure panels are PostScript plots generated by *XPREP*.

redundancies of 3.3 and 2.1 respectively, and should be comparable. In absence of unmerged intensity data, a *SHELX* (Sheldrick, 2008) format data file was generated from the mtz structure-factor amplitudes, read into *XPREP* (George Sheldrick, Bruker AXS) with HKL3 format option, and converted to intensities following the basic, error-propagation-based F to I conversion (see *e.g.* Rupp, 2009, pp. 328), *i.e.* $I = F^2$, $\sigma(I) = 2F\sigma(F)$.

Table 2
Comparison of key intensity statistics of 1fm4 *versus* 3k78.

Unusual or improbable values are shown in bold. The overall mean $I/\sigma(I)$ of 3k78 is more representative of strong synchrotron data (not in-house data), while the mean $I/\sigma(I)$ in the last (highest) resolution shell is atypically high, indicating that the noise level in the highest resolution shell is improbably low. The maximum $I/\sigma(I)$ is unreasonably high, and the $R\sigma$ is again improbably and atypically low. See also Fig. 3 and Supplementary Table 3(*a*).

XPREP analysis	1fm4	3k78
Unique reflections	9658	3346
$\lvert E^2 - 1\rvert$	0.755	0.773
Resolution range (Å)	28.66–1.99	25.56–2.80
Last resolution shell (Å)	2.09–1.99	2.90–2.80
Redundancy from PDB (all, last)	3.3 (1.9)	2.1 (1.5)
Completeness (all, last)	96.2 (75.9)	92.5 (76.6)
Mean I (all, last)	170.9 (30.3)	59.7 (21.0)
Mean $I/\sigma(I)$ (all, last)	8.16 (2.29)	**31.29 (20.34)**
Max $I/\sigma(I)$	35.9	**615.1**
$R\sigma$ (all, last)	0.092 (0.412)	**0.026 (0.044)**

While the mean I, mean $I/\sigma(I)$, and $R\sigma$ (Schneider & Sheldrick, 2002) values for 1fm4 are typical, the 3k78 data show highly unusual features (Table 2, Supplementary Table 3*b*[1], Fig. 3). The value of $R\sigma$ for validation is based on the fact that it allows computation and assessment of an *a posteriori* $R_{\rm merge}$-like data-quality indicator when unmerged data or images for proper reprocessing are not available owing to the unfortunate absence of a formal obligation to deposit unmerged intensity data or diffraction images. $R\sigma = \sum_{\mathbf{h}} \sigma_{(\mathbf{h})i} / \sum_{\mathbf{h}} I_{(\mathbf{h})i}$ tends to be somewhat lower than the corresponding linear $R_{\rm merge}$. For a discussion of the various merging R values see Diederichs &

[1] Supplementary materials have been deposited in the IUCr electronic archive (Reference: WD5176).

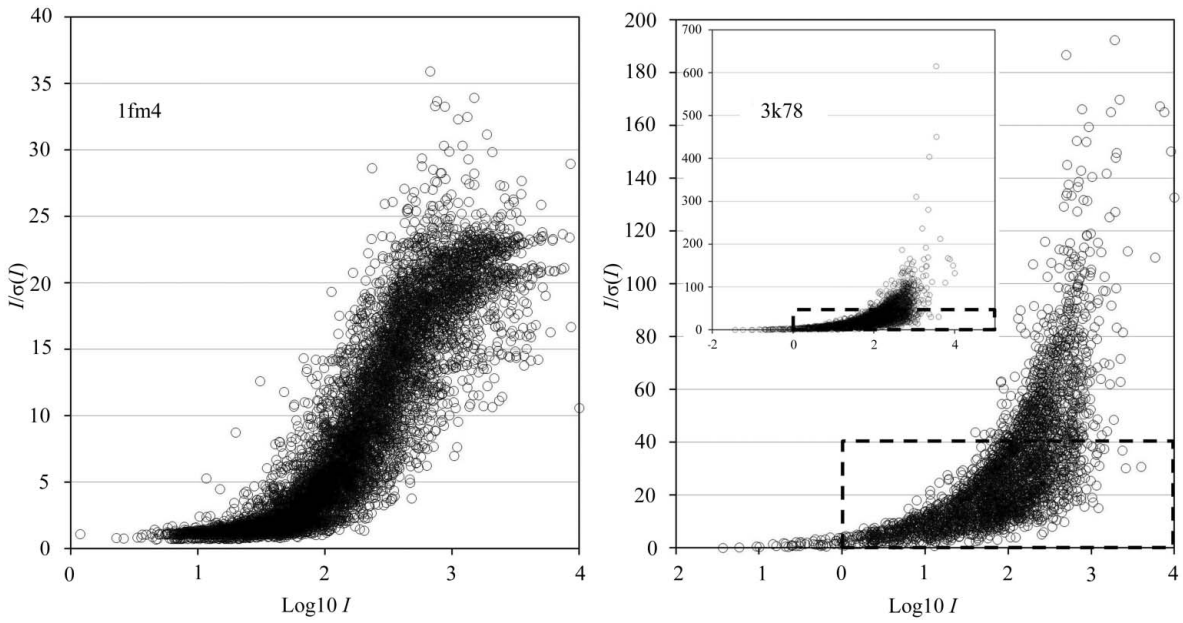


Figure 4
Diederichs plots for 1fm4 and 3k78. The left panel depicts the graph of $I/\sigma(I)$ *versus* $\log(I)$ for each unique reflection in the 1fm4 data set. It can be clearly seen that the sigmoid shape of the distribution levels off at around 20 to 30 $I/\sigma(I)$, as established and expected for normal data sets (Diederichs, 2010). In contrast, data for 3k78 show a steady increase to improbable $I/\sigma(I)$ values, indicating that they are not influenced by or do not contain any instrumentation-related measurement errors. The dashed boxes show how the 1fm4 graphs would scale into the 3k78 plots. The insert includes the extreme values for 3k78 which are omitted in the main panel. Note that the original Diederichs plots are based on unmerged intensities (which are not available, but redundancies are comparable between 1fm4 and 3k78). Merged reflections will have $I/\sigma(I)$ values higher by approximately the square root of the redundancy (K. Diederichs, personal communication).

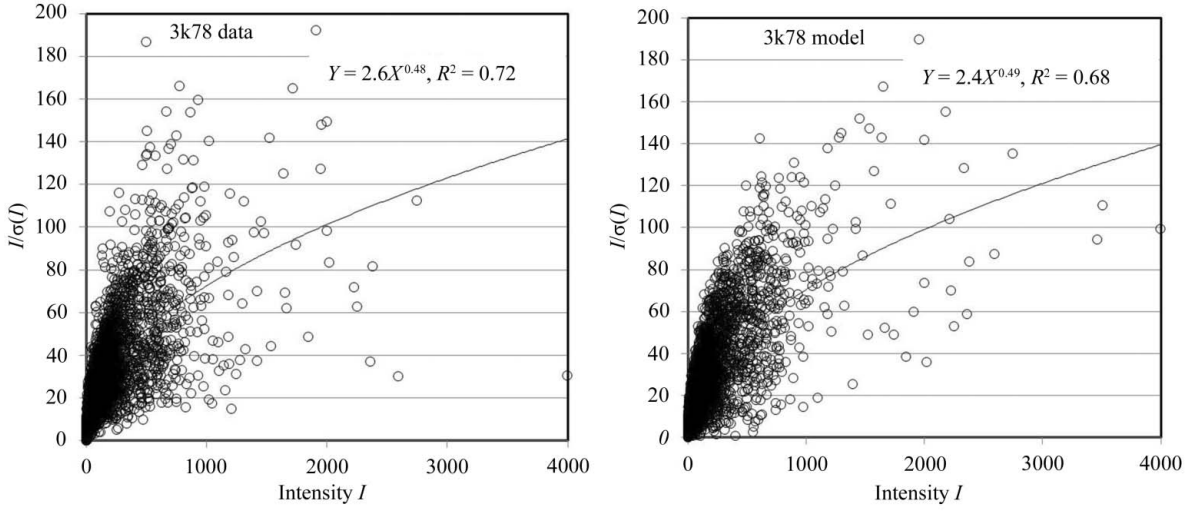


Figure 5
Model of the experimental uncertainties. The left panel depicts the graph of $I/\sigma(I)$ *versus* (I) for the 1fm4 data set (*i.e.*, a subsection of a non-log Diederichs plot). The distribution follows the $I^{1/2}$ *versus* I parabola (a.k.a. power law), indicating that the σs are derived without limiting experimental errors from I(calc) or F(calc). Adding random noise as described in the text yields an error distribution (right panel) that closely resembles that of the deposited data (left panel).

Karplus (1997); Weiss (2001); Rupp (2009); and Einspahr & Weiss (2012).

3.2. Diederichs plots

The improbably low $R\sigma$ values in 3k78 data are caused by a discrepancy between the intensities and their exceptionally low standard uncertainties. In addition to Poisson-statistics-derived counting errors, multiple other sources of instrumental errors limit the achievable signal to noise ratio, that is, $I/\sigma(I)$. This has been investigated in detail (Diederichs, 2010), and Diederichs notes that even with good crystals the $I/\sigma(I)$ ratio of the strongest (unmerged) observations is rarely above 30 even in the lowest resolution shell. It is obvious then, that 'counting statistics are not the limiting factor, as individual reflections may well have many more than 10 000 counts, which would allow $I/\sigma(I)$ ratios of more than 100 and low-resolution R factors of better than 1%' (Diederichs, 2010). The paper also provides multiple plots of $I/\sigma(I)$ *versus* $\log(I)$ which show distinct plateaux at around $I/\sigma(I)$ values of about 20 to 30.

In absence of original unmerged intensity data and to account for possible effects of redundancy, the 1fm4 data with a reported overall redundancy of 3.3 and of 3k78 with a redundancy of 2.1 were

compared with the aid of Diederichs plots (Fig. 4). 1fm4 shows the behavior expected for a normal data set, while 3k78 shows extremely high $I/\sigma(I)$ values and completely atypical behavior, and are apparently unlimited by any instrument measurement errors.

The resulting improbably high signal-to-noise ratios in turn indicate that these standard uncertainties are not based on any experimental variances. Some analysis of a possible origin can be provided by examining a non-logarithmic version of the Diederichs plot. A simple power law fit of the deposited data reveals that the signal-to-noise ratio $I/\sigma(I)$ is essentially proportional to the square root of I, which is expected if the $\sigma(I)$ is computed from $I^{1/2}$. An error model closely reproducing the deposited standard uncertainties can be obtained by generating a random error from the absolute inverse cumulative normal distribution around mean zero with a σ of 3.0 *via* the *Excel* NORMINV function, and forming the square root of the product of this random error with I. From these $I/\sigma(I)$ values (Fig. 5), F and $\sigma(F)$ follow again by basic error propagation, with an atypical $\sigma(F)$ distribution very similar to the deposited standard uncertainties. Spreadsheets including the calculations and additional graphs are included in the supplementary material.

3.3. Bulk-solvent content analysis

Proteins contain large fractions of disordered solvent, whose bulk-solvent scattering contributions supress the low-resolution intensities in an experimentally collected protein diffraction data set. The low-resolution structure factors calculated without bulk-solvent contributions should be significantly higher than the observed structure factors, while at the same time the R values for a refinement of a not bulk-solvent-corrected structure should be much higher than for a properly bulk-solvent-corrected structure. Representative graphs and a review of bulk-solvent scattering models can be found in Fokine & Urzhumtsev (2002) and in basic textbooks (*e.g.* Rupp, 2009).

The original cross-validation data set contained only 4.8% of the data (162 reflections), and in the two lowest resolution shells the

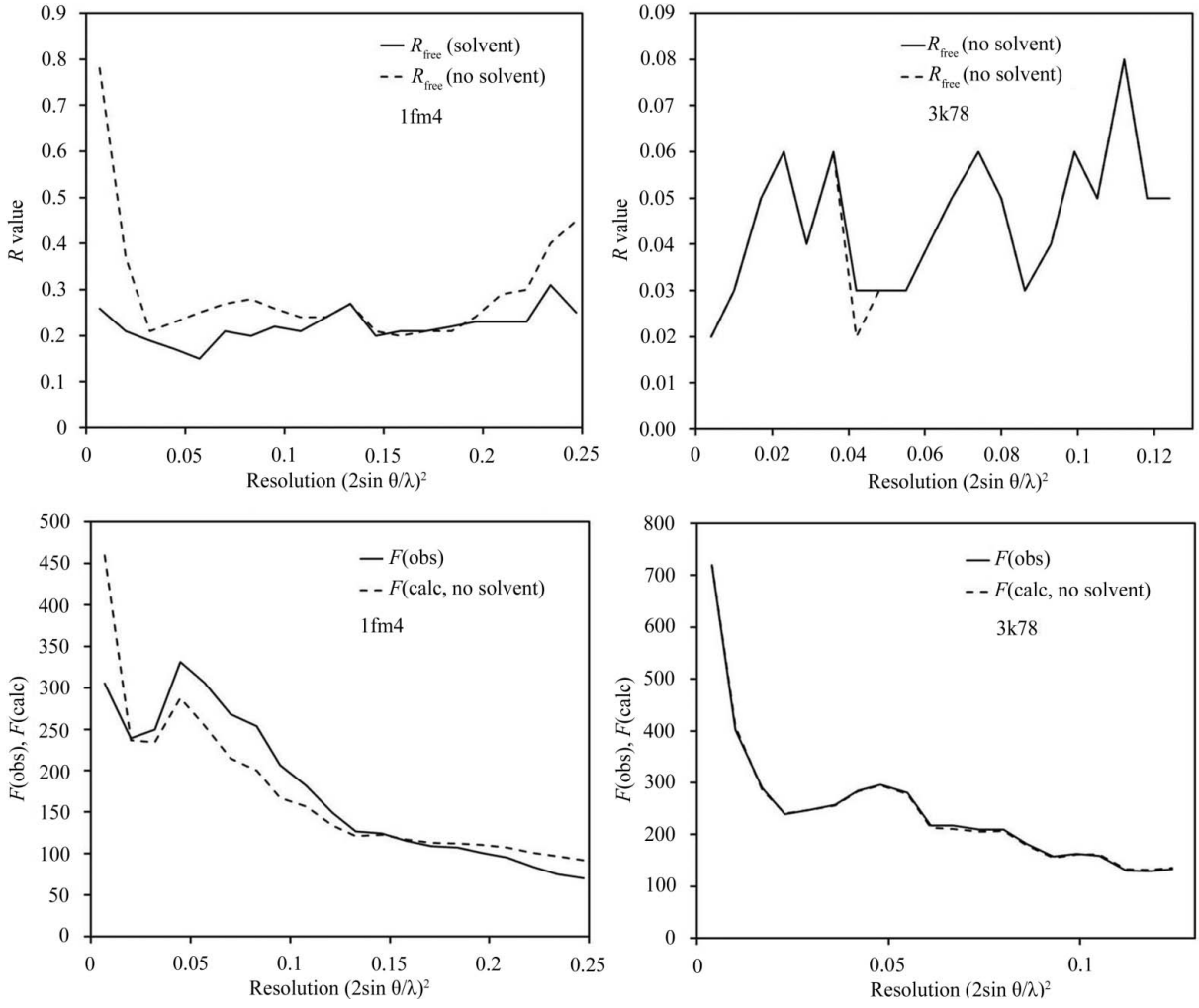


Figure 6
Bulk-solvent contribution analysis for 1fm4 and 3k78. The left panels depict the expected, nearly textbook-like behavior of a normal crystal structure like 1fm4. The top row shows the resolution-dependent behavior of R_{free} when the bulk-solvent correction is included (solid lines) and when it is not included (dashed lines) in the R-value calculation. 1fm4 shows the expected increase of low the resolution R values in the absence of bulk-solvent correction, indicating that bulk-solvent scattering contributions are present in the observed data. Such is not the case for 3k78. Bottom row: the presence of bulk-solvent contributions also causes the low-resolution calculated structure factors (dashed line) to be higher that the observed ones (solid), which are appropriately attenuated by the disordered bulk scattering contributions in 1fm4. There is no difference between F(obs) and F(calc) for 3k78, again indicating the absence of bulk-solvent scattering in the structure-factor data.

original 3k78 data contained no or only one cross-validation reflection, respectively. For the overall data range, the uncertainty in R_{free} (Kleywegt & Brünger, 1996; Tickle *et al.*, 1998*a*) is still acceptable with the low number of crossvalidation reflections, but for plotting in shells the R_{free} count is too low to be of practical value. For plotting, new *a posteriori* R_{free} data (Brünger, 1997) were obtained from new cross-validation data sets with 10% random selection against which the coordinate-perturbed starting model from the first 3k78 isotropic refinement was refined. Even with this suboptimal cross-validation procedure, the isotropic *B*-factor refinements reproduced the same *R* values of around 0.04/0.02. The R_{free} *versus* resolution plots for 3k78 were still noisy but show the same trend as plots from the original cross-validation set, and these data were used in the following analysis.

Structure factors and *R* values were calculated by *REFMAC* with and without bulk-solvent correction from the respective re-refined models of 1fm4 and 3k78. The R_{free} *versus* resolution plots as well as *F*(calc) and *F*(obs) *versus* resolution show expected behavior for 1fm4 consistent with bulk-solvent scattering contributions (Fig. 6). The same plots for 3k78 indicate absence of bulk-solvent scattering contributions in the structure factors, consistent with the negative bulk-solvent correction and trivially small bulk-solvent scale factor reported by *REFMAC* and the EDS report. The R_{free} plot for 3k78 shows the same lack of the strong increase in low resolution *R* value that would be expected for the refinement in the absence of a bulk-solvent correction and resembles the findings for the fabricated C3b structure (Janssen *et al.*, 2007). Given identical *F*(obs) and *F*(calc) without bulk-solvent contribution, logarithmic intensity ratio data plots (not shown) again replicate the situation demonstrated for the C3b structure.

For the purpose of validation, bulk-solvent parameters need to be calculated reliably from the original data. The EDS data at present suffer from some divergences, leading to a multimodal distribution probably caused by certain threshold or limit values for the bulk-solvent parameters. A consistent calculation using the flat bulk-solvent contribution (Afonine *et al.*, 2005; Afonine 2012) model using *phenix.refine* (Adams *et al.*, 2010) provides ~40 000 valid bulk-solvent contribution *B*-factor–scale-factor pairs. The probability distribution function represented in Fig. 7 is consistent with the earlier published smaller set of data (Fokine & Urzhumtsev, 2002). Entry 3k78, the fabricated entry 2hr0 (Janssen *et al.*, 2007), and two entries that are now updated (1n0q and 1n0r) but contained erroneously deposited calculated structure factors (Mosavi *et al.*, 2002), could be clearly identified as outliers given the distribution in Fig. 7.

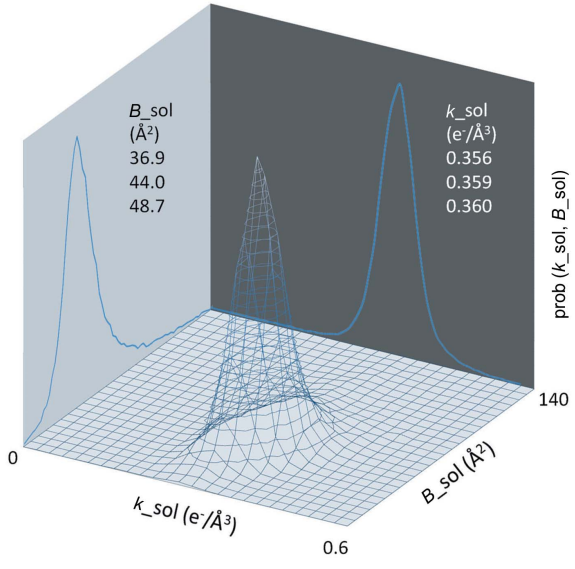


Figure 7
Probability distribution function of bulk-solvent correction parameters. The plot shows the distribution of bulk-solvent parameter pairs (scale factor and *B* factor) calculated from 40 000 PDB entries where valid parameters could be refined using *phenix.refine*. The walls of the plot show the separate distributions of *k*_sol and *B*_sol, with mode, median and mean listed next to the respective graphs. Raw data are included in the supplementary material.

4. Improbable model features caused by zero occupancies

The pattern that the zero occupancy atoms of 3k78 residues (Asn29, Lys66, Lys81, Lys104, Lys130, Glu132, Gln133, Lys135 and Lys138) display seems to be caused by a shift of zero occupancies to atoms with atom numbers decremented consistently by 2. This shift causes the backbone O atoms of the respective residue to become unoccupied, while the terminal atoms of the residues become occupied again (Fig. 8 and Supplementary Table 4*a*). Such errors could be introduced during the preparation of molecular replacement models. In case of experimental structure factors, the electron-density map will indicate the error by positive difference density peaks in place of the atoms missing in the model. In case of 3k78, however, the atom absences propagate into the electron density.

Quite unexpected is that in original 3k78 maps (§2.4) no $2mF_o - DF_c$ density for the unoccupied missing atoms down to near-noise levels below 0.5σ nor difference density the $mF_o - DF_c$ maps is visible for unoccupied atoms, including the backbone O atoms in Lys130, Glu132 and Gln133 (Fig. 9). The weak difference density for Lys135 probably results from incorrect placement. Given the reported typical main-chain *B* factors (~30–35 Å^2) of the adjacent, covalently connected backbone atoms, this behavior is very unusual and improbable. Following the lysine side chains towards the solvent, there is again clear density for the solvent-exposed C^{ε} and N^{ζ} atoms of the lysine residues, but they are untethered by hydrogen bonds or other contacts. These observations are characteristic of data calculated from a model with zero occupancy atoms.

Setting occupancies of protein atoms that are poorly defined or absent in electron density to zero has very little effect on the overall model quality or refinement itself: zero occupancy as well as a very high *B* factor both lead to respectively zero or negligible scattering contributions, and either will have an insignificant effect on the rest of the model. Inspection of the electron density of the side-chain atoms of residues with reset occupancy in the re-refined 1fm4 model illustrate the fact that such atoms simply refine to high *B* factors and

```
RES C SSEQ ATOMS 1FM4          RES C SSEQ ATOMS 3K78
LYS A 28 CE NZ                 ASN A 29 CB CG
LYS A 65 CE NZ                 LYS A 66 CG CD
LYS A 80 CD CE NZ              LYS A 81 CB CG CD
LYS A 103 CD CE NZ             LYS A 104 CB CG CD
LYS A 129 CG CD CE NZ          LYS A 130 O CB CG CD
GLU A 131 CG CD OE1 OE2        GLU A 132 O CB CG CD
GLN A 132 CG CD OE1 NE2        GLN A 133 O CB CG CD
LYS A 134 CG CD CE NZ          LYS A 135 O CB CG CD
LYS A 137 CD CE NZ             LYS A 138 CB CG CD
```

Figure 8
Zero occupancy atoms in 1fm4 and 3k78. Condensed REMARK 480 from PDB headers. The atoms in 3k78 (right-hand columns) are shifted towards lower atom numbers compared to 1fm4, causing the zero occupancies to progress towards the main chain including the backbone O, and the terminal atoms of the side chain to become occupied again. This situation is physically improbable. See also Supplementary Table 4*a*.

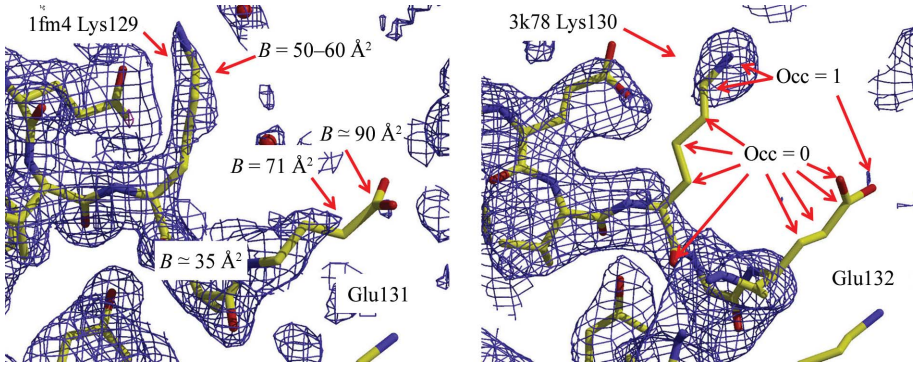


Figure 9
Normal and pathological side-chain density. $2\,mF_o - DF_c$ electron density contoured at 0.8σ. The left panel shows the progressive weakening of electron density owing to displacement of the side-chain atoms, after re-refinement with the originally zero occupancies reset to 1. The *B* factors are restrained against unreasonable increases between subsequent adjacent atoms, and in normal situations show a continuous increase along the side chain. The right panel shows an improbable scenario where atoms that had previously zero occupancies assigned refine to extreme *B* factors at the limit of what the restraints allow and the electron density abruptly disappears, and, in the case of Lys130, abruptly reappears for the terminal C^{ε} and N^{ζ} side chain atoms. This is also true but less visible owing to the stronger main-chain *B*-factor restrains for the Lys130 backbone O atom. These observations provide a first indication that the deposited structure factors do not to contain any contributions from the unoccupied atoms. Note that in some real scenarios the terminal lysine N^{ζ} for example can be tethered through non-covalent interactions with inter- or intra-molecular neighboring residues and become better defined than the remaining hydrophobic side-chain atoms. This is however not the case for Lys130 of 3k78. All density figures were prepared with *XtalView* (McRee, 1999) and rendered by *Raster3d* (Merritt & Bacon, 1997).

Original *WHAT_CHECK* report of 3k78 bond distance violations

102	LYS	(104)	A	CE	NZ	1.64	4.9
119	TYR	(121)	A	CB	CG	1.60	4.1
122	LYS	(124)	A	CB	CG	1.66	4.7
144	ARG	(146)	A	CB	CG	1.65	4.2

WHAT_CHECK report of 3k78 with occupancies reset

27	ASN	(29)	A	CA	CB	1.41	-6.2
27	ASN	(29)	A	CG	OD1	1.01	-11.0
27	ASN	(29)	A	CG	ND2	1.67	16.4
64	LYS	(66)	A	CD	CE	1.68	5.2
79	LYS	(81)	A	CA	CB	1.41	-6.0
102	LYS	(104)	A	CA	CB	1.39	-7.2
102	LYS	(104)	A	CD	CE	1.40	-4.1
102	LYS	(104)	A	CE	NZ	1.64	4.9
119	TYR	(121)	A	CB	CG	1.60	4.1
122	LYS	(124)	A	CB	CG	1.66	4.7
128	LYS	(130)	A	CD	CE	1.32	-6.8
130	GLU	(132)	A	C	O	1.31	4.1
130	GLU	(132)	A	CD	OE2	1.69	23.1
131	GLN	(133)	A	CD	OE1	1.45	11.1
131	GLN	(133)	A	CD	NE2	1.45	5.8
133	LYS	(135)	A	C	O	1.54	15.6
133	LYS	(135)	A	CA	CB	1.35	-8.9
133	LYS	(135)	A	CD	CE	1.37	-5.1
136	LYS	(138)	A	CD	CE	1.24	-9.3
144	ARG	(146)	A	CB	CG	1.65	4.2

Figure 10
WHAT_CHECK report of bond distance violations for 3k78. The last column contains the deviation from known r.m.s. values, expressed in σ levels. Setting atoms to zero occupancies can lead to missing them during model validation and correction. In the case of 3k78, even a backbone atom distance violation of 15.6σ would go undetected (but the PDB validation reports it in REMARK 500).

display correspondingly weak electron density (Fig. 9). Nevertheless, it should be kept in mind that for many cases of local disorder, large isotropic displacement (*B*) factors are not a physically correct description either (Merritt, 2012). A number of other inconsistencies and problems however can be introduced by zero occupancy atoms in the chain of a protein model.

(i) Despite the fact that these unoccupied atoms are not included in the refinement, they do remain in the model but may not be included in the calculation of the r.m.s. deviation from geometry restrain target values listed in the PDB header. Table 1 lists such a discrepancy for 3k78.

(ii) An additional problem caused by the zero occupancies is that geometry validation programs may be misled. For example, *WHAT_CHECK* (Hooft *et al.*, 1996) properly warns of zero occupancy atoms but does not compute their geometry deviations, leaving the corresponding errors unlisted. Fig. 10 demonstrates this scenario for entry 3k78. *MolProbity* (Davis *et al.*, 2007) also excludes atoms with occupancies below 0.02 and also does not report side-chain bond distance and angle violations (J. Richardson, personal communication). However, the PDB validation does include zero occupancy atoms in the preparation of geometry violation statistics for REMARK 480 and 500 (available as RUN500 from the *CCP4i* interface).

(iii) Not all display programs recognise zero occupancies, while at the same time the *B* factors of those atoms can be set to an arbitrary, non-representative (often low) value which again may be misinterpreted, or missed in *B*-factor analysis.

5. Conclusions

The findings surfacing during model refinement in §2 and amplified during the structure factor analysis in §3 and the feature propagation discussed in §4 provide consistent and very convincing evidence that (*a*) the structure-factor data deposited for 3k78 are calculated structure factors, (*b*) the resulting re-refined model resembles in most details the mutated search model, (*c*) that the original model has not, or not properly, been refined against these structure factors (or had been altered from a model essentially similar to the re-refined model and after the structure factors had been calculated). Being not refined against the deposited structure factors, the 3k78 model at present at least lacks experimental basis. The findings leading to the above conclusions are summarized below.

(i) The deposited structure factors do not contain any bulk-solvent contribution.

(ii) The noise level of the data is abysmally small and nearly constant over the entire resolution range, consistent with a truncated calculated data set with inappropriate error model.

(iii) The Diederichs plots show almost orders of magnitude higher signal-to-noise ratios than expected for real data, indicative of absence of instrumentation errors in calculated structure factors and in the error model.

(iv) The structure factors deposited for the PDB entry 3k78 are in fact calculated structure factors, and their standard uncertainties are not based on experimental errors.

(v) Because the original refinement against these structure factors gives the same R values as reported or calculated by *PDB_REDO* and in this work, a simple error of swapping the F(obs) and F(calc) columns during data deposition can be excluded.

(vi) The refinement statistics reported in the PDB header are inconsistent with actual refinement against the structure-factor data.

(vii) The model refines against the deposited 2.8 Å data without the need for bulk-solvent correction, no H atoms, atypical X-ray matrix weights, to near-zero R values, compatible only with calculated structure factors.

(viii) The model obtained by re-refinement does not correspond to the deposited model, but is in details closer to the molecular replacement starting model.

(ix) The non-physical zero occupancy residues in the model are faithfully reproduced in the electron density calculated from the deposited structure-factor data, which is inconsistent with experimental data obtained from a real protein structure.

(x) Numerous residues of the original model are not located in their electron density, but return to the exact position of the density when refined. This is consistent with these parts of the re-refined model being manipulated after the structure factors were generated from it.

Each of these points alone is reason for concern, and when combined and evaluated against prior expectations, they leave no doubt that model and data of 3k78 are incompatible and that the deposited structure factors are not based on actual experiments, and their standard uncertainties are not based on experimental errors.

Following basic scientific epistemology, strong and convincing evidence would have to be provided to overcome these doubts (Rupp, 2010). In case of an error during deposition, this should be trivial to achieve, and database integrity could be easily restored. At least an experimental data set which refines to the deposited structure, or unmerged intensity data reprocessed from the original images should be supplied. Most convincing and irrefutably, the presentation of actual diffraction images which produce data representing the deposited model would establish the facts.

6. A few recommendations

Considerable efforts by the PDB validation task force (Read, 2011) will make it much less likely that poorly refined models, models inconsistent with data, or implausible data will enter the public databases. Nevertheless, it remains a fact that – irrespective of the cause of the problem – in the case of 3k78 a calculated data set also incompatible with the associated coordinate entry has been successfully deposited. The example of 3k78 provides a few additional suggestions that might be useful not just for *a posteriori* validation during deposition but also particularly for the aspiring crystallographer during structure refinement.

(i) Diffraction image deposition and archival. The need for preserving diffraction images for scientific reasons has been officially suggested by the IUCr in 2008 (Baker *et al.*, 2008) and a standing IUCr committee on data deposition has been formed in 2011. Although matters of policies and technical issues remain to be resolved, there is little doubt that image deposition is a timely and beneficial practice for scientific reasons. As an additional side-effect, image deposition allowing reprocessing would immediately resolve any questions of data provenance. Successful redeposition of the correct observed structure factors of entries 1n0q and 1n0r (Mosavi *et al.*, 2002), reprocessed form original diffraction images collected a decade ago, clearly demonstrates the value of proper image data archiving.

(ii) Bulk-solvent correction. It would be useful if all refinement programs consistently report the bulk-solvent B factor and also the bulk-solvent scale factor in the REMARK 3 section of the PDB header. Implausible values could be readily detected and corrective action taken already during refinement. The bulk-solvent scale factor actually becomes a more useful measure than the bulk-solvent B factor, particularly at the spurious solvent contents refined from calculated structure factors.

(iii) Setting the occupancy of protein chain atoms to zero as an indication of positional uncertainty is physically not correct. Accepting high B factors (which are not necessarily a correct physical description of substantial disorder either) causes less problems, such as geometry validation programs not including unoccupied atoms in the validation statistics. Isolated backbone zero occupancies are physically not meaningful and should be correspondingly flagged as a serious problem. Side-chain atoms may be absent owing to radiation damage, and in such cases the use of zero occupancies as an indicator could be arguably justified.

(iv) The Diederichs plot (§3.2) seems to be a valuable tool in spotting anomalies in diffraction data, particularly as far as the signal-to-noise ratios, *i.e.* $I/\sigma(I)$ and the instrumentation error model is concerned. Potential for abuse by fitting calculated error models to the sigmoid distribution does exist.

(v) $R\sigma$ (§3.1) can serve as a useful *a posteriori* measure for the plausibility of the error model and signal-to-noise levels in the absence of any merging R values.

(vi) *A posteriori*, the *PDB_REDO* database can be examined for improbably high discrepancies between the originally reported R values and the conservatively re-refined structure of a PDB entry.

(vii) In the absence of image deposition, and as an option requiring no special effort, more refinement data could be deposited. At least the F(calc) set could be submitted in addition to F(obs) to allow easy detection of simple column swapping or other possible deposition mistakes. Even better, the Fourier coefficients for the final electron-density map should be deposited, because this map ultimately represents what the crystallographer was interpreting during model building. EDS can only reconstruct maps from what it is provided with, which presently are only the deposited structure-factor amplitudes and the model coordinates.

Finally, despite all the diagnostics and validation tools available during model building, refinement, and ultimately upon PDB deposition, one needs to recollect that not the PDB but the individual crystallographer bears the final – and sometimes far reaching (Petsko, 2007) – responsibility for the correctness of the deposited model.

I wish to anonymously acknowledge several colleagues who provided critical comments and detailed information about the refinement and data analysis programs used in this work. Ed Pozharski extracted raw data from the EDS database. P. Afonine computed bulk-solvent contributions with an improved bulk-solvent parameter implementation in *phenix.refine*. Reviewers have pointed out a number of didactical and presentational improvements to the

manuscript. The *REFMAC* command script, the input files, and the results for the isotropic *B*-factor refinement of 3k78 as well as the *XPREP* data analysis and bulk-solvent data are deposited as supplementary materials. The hyperlink to *PDB_REDO* of 3k78 is http://www.cmbi.ru.nl/pdb_redo/k7/3k78/index.html, for the EDS report http://eds.bmc.uu.se/cgi-bin/eds/uusfs?pdbCode=3k78, and the electron density can be loaded *via* the EDS link to the *ASTEX Viewer* at http://eds.bmc.uu.se/cgi-bin/eds/eds_astex.pl?infile=3k78&centre=A61.

References

Adams, P. D. *et al.* (2010). *Acta Cryst.* D**66**, 213–221.

Afonine, P. V., Grosse-Kunstleve, R. W. & Adams, P. D. (2005). *Acta Cryst.* D**61**, 850–855.

Afonine, P. V., Grosse-Kunstleve, R. W., Echols, N., Headd, J. J., Moriarty, N. W., Mustyakimov, M., Terwilliger, T. C., Urzhumtsev, A., Zwart, P. H. & Adams, P. D. (2012). *Acta Cryst.* D**68**, 352–367.

Baker, E. N., Dauter, Z., Guss, M. & Einspahr, H. (2008). *Acta Cryst.* D**64**, 337–338.

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.

Chothia, C. & Lesk, A. M. (1986). *EMBO J.* **5**, 823–826.

Davis, I. W., Leaver-Fay, A., Chen, V. B., Block, J. N., Kapral, G. J., Wang, X., Murray, L. W., Arendall, W. B. III, Snoeyink, J., Richardson, J. S. & Richardson, D. C. (2007). *Nucleic Acids Res.* **35**, W375–W383.

Diederichs, K. (2010). *Acta Cryst.* D**66**, 733–740.

Diederichs, K. & Karplus, P. A. (1997). *Nature Struct. Biol.* **4**, 269–275.

Einspahr, H. M. & Weiss, M. S. (2012). *International Tables for Crystallography*, Vol. F, 2nd ed., edited by E. Arnold, D. M. Himmel & M. G. Rossmann, pp. 64–74. Chichester: Wiley.

Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* D**66**, 486–501.

Fokine, A. & Urzhumtsev, A. (2002). *Acta Cryst.* D**58**, 1387–1392.

Gajhede, M., Osmark, P., Poulsen, F. M., Ipsen, H., Larsen, J. N., Joost van Neerven, R. J., Schou, C., Løwenstein, H. & Spangfort, M. D. (1996). *Nature Struct. Biol.* **3**, 1040–1045.

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

Janssen, B. J., Read, R. J., Brünger, A. T. & Gros, P. (2007). *Nature (London)*, **448**, E1–E2.

Jiang, J.-S. & Brünger, A. T. (1994). *J. Mol. Biol.* **243**, 100–115.

Joosten, R. P., te Beek, T. A., Krieger, E., Hekkelman, M. L., Hooft, R. W., Schneider, R., Sander, C. & Vriend, G. (2011). *Nucleic Acids. Res.* **39**, D411–D419.

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

Kleywegt, G. J., Harris, M. R., Zou, J., Taylor, T. C., Wählby, A. & Jones, T. A. (2004). *Acta Cryst.* D**60**, 2240–2249.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J. & Higgins, D. G. (2007). *Bioinformatics*, **23**, 2947–2948.

Laskowski, R. A. (2001). *Nucleic Acids. Res.* **29**, 221–222.

Marković-Housley, Z., Degano, M., Lamba, D., von Roepenack-Lahaye, E., Clemens, S., Susani, M., Ferreira, F., Scheiner, O. & Breiteneder, H. (2003). *J. Mol. Biol.* **325**, 123–133.

McRee, D. E. (1999). *J. Struct. Biol.* **125**, 156–165.

Merritt, E. A. (2012). *Acta Cryst.* D**68**, 468–477.

Merritt, E. A. & Bacon, D. J. (1997). *Methods Enzymol.* **277**, 505–524.

Mosavi, L. K., Minor, D. L. & Peng, Z. Y. (2002). *Proc. Natl Acad. Sci. USA*, **99**, 16029–16034.

Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* D**67**, 355–367.

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

Painter, J. & Merritt, E. A. (2006). *Acta Cryst.* D**62**, 439–450.

Petsko, G. A. (2007). *Genome Biol.* **8**, 103.

Potterton, E., Briggs, P., Turkenburg, M. & Dodson, E. (2003). *Acta Cryst.* D**59**, 1131–1137.

Read, R. J. *et al.* (2011). *Structure*, **19**, 1395–1412.

Rupp, B. (2009). *Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology*, 1st ed. New York: Garland Science.

Rupp, B. (2010). *J. Appl. Cryst.* **43**, 1242–1249.

Schneider, T. R. & Sheldrick, G. M. (2002). *Acta Cryst.* D**58**, 1772–1779.

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

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

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.

Tickle, I. J., Laskowski, R. A. & Moss, D. S. (2000). *Acta Cryst.* D**56**, 442–450.

Tronrud, D. E. (1996). *J. Appl. Cryst.* **29**, 100–104.

Velankar, S. *et al.* (2010). *Nucleic Acids Res.* **38**, D308–D317.

Weiss, M. S. (2001). *J. Appl. Cryst.* **34**, 130–135.

Winn, M. D. (2003). *J. Synchrotron Rad.* **10**, 23–25.

Winn, M. D., Isupov, M. N. & Murshudov, G. N. (2001). *Acta Cryst.* D**57**, 122–133.

Winn, M. D. *et al.* (2011). *Acta Cryst.* D**67**, 235–242.

Zaborsky, N., Brunner, M., Wallner, M., Himly, M., Karl, T., Schwarzenbacher, R., Ferreira, F. & Achatz, G. (2010). *J. Immunol.* **184**, 725–735.

Acta Crystallographica Section F
Structural Biology and Crystallization Communications
ISSN 1744-3091

Response to *Detection and analysis of unusual features in the structural model and structure-factor data of a birch pollen allergen*

Nadja Zaborsky, Marietta Brunner, Michael Wallner, Martin Himly, Tanja Karl, Robert Schwarzenbacher, Fatima Ferreira* and Gernot Achatz‡

Department of Molecular Biology, University of Salzburg, Hellbrunnerstrasse 34, Salzburg, A-5020, Austria

‡ Professor Gernot Achatz unexpectedly passed away on 14 August 2011.

Correspondence e-mail: fatima.ferreira@sbg.ac.at

Received 24 February 2012
Accepted 24 February 2012

The authors of *J. Immunol.* **184**, 725–735 respond to the article by Rupp (2012), *Acta Cryst.* F**68**, 366–376.

On 9 February 2012, the Editors of *Acta Crystallographica Section F* contacted the authors of a paper published in *J. Immunol.* in 2010 (Zaborsky *et al.*, 2010) concerning a manuscript received for publication on 12 January 2012 (Rupp, 2012). This manuscript presents strong evidence that the diffraction data of Bet v 1d (PDB code 3k78; published in the *J. Immunol.* paper) are not derived from a diffraction experiment and that the model of 3k78 contains some highly improbable features.

The University of Salzburg immediately informed and commissioned the Austrian Agency for Research Integrity (OeAWI) to carry out an investigation into possible data fraud on the part of author Robert Schwarzenbacher, the co-author solely responsible for the Bet v 1d structure and the crystallographic section of the *J. Immunol.* paper. The OeAWI is presently preparing a report of this investigation.

Author Schwarzenbacher admits to the allegations of data fabrication and deeply apologizes to the co-authors and the scientific community for all the problems this has caused. The Protein Data Bank has been informed and has been asked to retract the 3k78 entry, which was obsoleted in February 2012. The main body of the *J. Immunol.* publication concerns the immunological study and the retraction of the crystallographic section does not affect the major conclusions. The Editors of *J. Immunol.* have been informed about the problem with the structure 3k78. Co-authors Zaborsky, Brunner, Wallner, Himly, Karl, Ferreira and Achatz were in no way involved in the generation of the crystallographic data.

Note added in proof: subsequent to the acceptance of this article for publication, author Schwarzenbacher withdrew his admission of the allegations.

References

Rupp, B. (2012). *Acta Cryst.* F**68**, 366–376.
Zaborsky, N., Brunner, M., Wallner, M., Himly, M., Karl, T., Schwarzenbacher, R., Ferreira, F. & Achatz, G. (2010). *J. Immunol.* **184**, 725–735.

© 2012 International Union of Crystallography
All rights reserved