JOURNAL OF APPLIED CRYSTALLOGRAPHY
ISSN 1600-5767

Mean weighted residuals reveal systematic overestimation of Bragg intensities in single-crystal diffraction

Julian Henn,[a]* Piero Macchi[b]* and Toms Rekis[c]

[a]DataQ Intelligence UG, Fichtelgebirgsstrasse 66, Germany, [b]Department of Chemistry, Materials and Chemical Engineering, Politecnico di Milano, Via Bassini 6, 20133, Milano, Italy, and [c]Institute of Inorganic and Analytical Chemistry, Goethe University Frankfurt, Max-von-Laue Str. 7, 60438 Frankfurt am Main, Germany. *Correspondence e-mail: julianhenn@web.de, piero.macchi@polimi.it

Received 27 November 2025
Accepted 23 March 2026

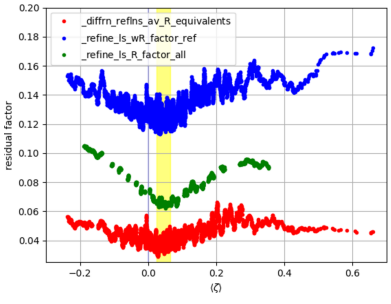
Edited by S. Moggach, The University of Western Australia, Australia

Keywords: systematic errors; rewarding errors; metrics; single-crystal diffraction.

The mean value of weighted residuals ($\langle\zeta\rangle$) was analysed for 8424 published single-crystal X-ray data sets of crystals containing only light elements (C, H, N, O). A striking asymmetry was observed: 71.5% of data sets exhibit positive $\langle\zeta\rangle$ values, occurring 2.5 times more often than negative values. This imbalance suggests systematic errors, with evidence pointing to a slight overestimation of observed intensities (I_{obs}). Simulations and theoretical analysis show that such overestimation artificially lowers common data-quality metrics, including the popular merging factor R_{merge}, the redundancy independent factor $R_{r.i.m.}$, the precision indicating factor $R_{p.i.m.}$, the weighted agreement factor $wR(F^2)$ and even atomic displacement parameters, creating a 'rewarding error' that may reinforce confirmation bias. Experimental data confirm these findings, as residual factors reach their minima for $\langle\zeta\rangle > 0$ rather than at zero. These results highlight the need for critical evaluation of data-processing strategies and caution against relying solely on conventional agreement factors as indicators of accuracy.

1. Introduction

Systematic errors in single-crystal diffraction are of general importance as the description and tracking of errors can be used to continuously improve the accuracy of the experiments, to validate data-acquisition and data-processing steps, to adjust parameter values in data-integration steps, to expose misconceptions, to validate new approaches and changes in hard- and software, and to improve correction procedures such as absorption and extinction models, as well as supporting modelling in challenging developing fields such as electron diffraction. In the recent past, the mean value of the weighted residuals $\langle\zeta\rangle$, $\zeta = (I_{obs} - I_{calc})/\sigma(I_{obs})$, has been found to be a helpful data descriptor. A significant deviation from zero indicates the presence of systematic errors, which is frequently the case (Henn, 2019). The significance of the deviation from zero is calculated by dividing $\langle\zeta\rangle$ by the standard deviation of the mean value $\sigma(\langle\zeta\rangle)$. The standard deviation of the mean value is given by the square root of the unbiased sample variance over the number of reflections, $\sigma(\langle I_{obs}\rangle) = [\mathrm{var}(\zeta)/N_{obs}]^{1/2}$, with the unbiased sample variance $\mathrm{var}(\zeta) = [1/(N_{obs} - 1)] \sum_{i=1}^{N_{obs}} (\zeta_i - \langle\zeta\rangle)^2$. This definition for the significance of the mean value of the weighted residuals is analogous to the definition of the significance of a redundantly measured observed reflection, where the standard deviation of the mean value (as opposed to the standard deviation of the sample) is obtained by dividing the unbiased sample variance by the redundancy and taking the square root.



It was previously found that $\langle\zeta\rangle$ tends to positive values. In a sample of 127 data sets published with *IUCrData* (https://iucrdata.iucr.org/x/), 52% of the data sets showed a significant positive deviation of the mean value of the weighted residuals from zero (Henn, 2019). These findings were later confirmed with an even larger sample of over 300 data sets from *IUCrData* (unpublished work). A positive shift of the residuals was connected earlier to several causes such as unrecognized low-energy contamination, unrecognized twinning and disorder problems (Domagala *et al.*, 2023).

It will be shown in this study how overestimation of observed intensities affects different metrics such as the merging R factor $R_{\rm merge}$, the redundancy independent merging factor $R_{\rm r.i.m.}$, the precision indicating merging factor, $R_{\rm p.i.m.}$ (Weiss, 2001) and the weighted agreement factor $wR(F^2)$: it leads to an artificial lowering in the studied metrics.

When aiming for high data quality, researchers may adjust various data-integration parameters. Success is then judged by monitoring commonly used metrics. These metrics can create the impression that quality is improving. In reality, after a certain point, the actual data quality begins to decline – even though the metrics continue to suggest improvement. Only after model refinement may these errors become apparent, for example by a systematic positive shift of the residuals, provided these traces of systematic errors are actively searched for. This systematic shift may be so small for individual data sets that it remains well below the noise level; however, with many data sets in the sample, the shift is clearly exposed.

When certain systematic errors (such as a slight overestimation of observed intensities) lead to seemingly higher data quality with respect to certain metrics [such as $R_{\rm r.i.m.}$, $R_{\rm p.i.m.}$, $R_{\rm merge}$, R, $wR(F^2)$, U_{ij} *etc.*], this is called a *rewarding error* with respect to the mentioned data quality metrics as it leads to an appreciated result. Rewarding errors are a particularly important class of errors since they meet the desired expectation of high data quality of the user. Therefore, they are less likely to be questioned (confirmation bias) and may occur frequently but may remain undetected for decades. When crystallographic software developers unintentionally and unknowingly fall for confirmation bias, this can also lead to undetected methodological issues.

The present work aims at confirming the tendency to positive residuals for a much larger sample of published data sets comprising only light elements (in order to minimize the impact of absorption correction errors discussed previously (Henn, 2025). Additionally, a possible explanation is offered by proposing that slight overestimation of $I_{\rm obs}$ on average is likely to be a cause of the shift of the residuals towards positive values.

2. The data

A total of 8424 crystallographic data sets containing only C, H, O and N were downloaded from the Crystallography Open Database (COD; Vaitkus *et al.*, 2023; Mesto *et al.*, 2013; Vaitkus *et al.*, 2021; Quirós *et al.*, 2018; Merkys *et al.*, 2016; Gražulis *et al.*, 2015; Gražulis *et al.*, 2012; Gražulis *et al.*, 2009; Downs & Hall-Wallace, 2003). The CIF tag `_exptl_absorpt_process_details` was used to determine the absorption correction processing software. Data processed with different releases of *SADABS* (Krause *et al.*, 2015), *SORTAV* (Blessing, 1987; Blessing, 1997; Blessing, 1995), Rigaku/Oxford Diffraction and Stoe & Cie software were included in the sample. The overwhelming majority of structure models use the independent atom model. It is known that the software versions may sometimes quote an older version of a program even when in fact the data were obtained with a newer version. No attempts were made to identify such cases. Most data sets were processed with one out of many releases of *SADABS* (N = 6781), with *SADABS* 1996 (N = 1919) having the largest share. Different releases of *CrysAlis PRO* [*CrysAlis PRO* Agilent releases from 2010 (N = 76), 2011 (N = 95), 2012 (N = 73), 2013 (N = 53) and 2014 (N = 80); *CrysAlis PRO* Oxford Diffraction releases from 2009 (N = 81) and 2010 (N = 71); *CrysAlis PRO* Rigaku OD 2015 (N = 72)], *CrysAlis RED* [*CrysAlis RED* Oxford Diffraction 2006 (N = 47), 2007 (N = 51), 2008 (N = 42) and 2009 (N = 83)] and *CrystalClear* [*CrystalClear* Rigaku MSC 2005 (N = 66) and *CrystalClear* Rigaku 2005 (N = 201)] add up to a total of 1091 Rigaku-associated data sets. Two releases of *SORTAV* follow [Blessing (1995) (N = 199) and Blessing (1997) (N = 33)], with a total of 232 data sets, and finally, *X-RED32* [Stoe & Cie 2002 (N = 235)].

The resolution limit as recalculated from $\theta_{\rm max}$ and the wavelength ranges between 0.4476 and 1.1744 Å^{-1}, with mean value 0.6402 Å^{-1} and median value 0.6276 Å^{-1}. Fifty per cent of all data sets have a maximum resolution in the range 0.6024–0.6601 Å^{-1}, with 25% of all data sets having a maximum resolution below and 25% above this range (see Table 1).

The number of observed intensities ranges between 314 from a low-temperature redetermination of metaldehyde (tetragonal space group $I4$) at 150 K [COD 2205445, Barnett *et al.* (2005)] and 38082 from a structure in monoclinic space group $P2_1/n$ containing a pyrene derivative $C_{34}H_{37}N$ [COD 2240967, Thekku Veedu & Techert (2015)], measured at 100 K and refined as a non-merohedral twin. Only two data sets show a zero weighting scheme parameter $a = 0$ [*N*,*N*′-bis(3-methylphenyl)succinamide dihydrate, monoclinic space group $P2_1/c$, COD 2230904, Saraswathi *et al.* (2011); a polymorph of butobarbital, monoclinic space group $P2_1/c$, COD 2016360, Gelbrich *et al.* (2007)]. A total of 1057 data sets show weighting scheme parameter $b > 1$; the largest values are 36.3175 [ethyl 4-butylamino-3-nitrobenzoate, monoclinic space group $C2/c$, COD 2222973, Narendra Babu *et al.* (2009)] and 38.7799 from a study of peptide nanotubes with flexible pores and disordered solvent [COD 2103455, Görbitz (2002)].

The maximum crystal dimension lies between 0.03 mm [COD 2230768, Ismiyev (2011); COD 2218750, Liang & Qu (2008); COD 2238399, Chetioui *et al.* (2013); COD 2022704, Aristov *et al.* (2023)] and 10.28 mm [benzohydrazide, $C_{25}H_{29}N_3O$, obtained with Cu $K\alpha$ radiation at 296 K, COD 2234318, Bhat *et al.* (2012)]. The weighted agreement factor $wR(F^2)$ lies between 0.0483 [polymorph of *myo*-inositol,

Table 1
Distribution of selected characteristics of the data set, including minimum, maximum, mean and median values for the number of reflections used in the least-squares minimization (N_{obs}), the weighting-scheme parameters a and b, the number of refined model parameters (N_{param}), the fraction $(N_{obs} - N_{param})/N_{obs}$, the maximum crystal size (in mm), the average R factor for equivalent reflections, the conventional R factor, and the weighted agreement factor.

Data sets processed using *SADABS*† (6781), Rigaku software‡ (1091), *SORTAV* (232), Stoe & Cie software (235) or unspecified software (85).

	Minimum	Maximum	Mean	Median
N_{obs}	314	38082	3534.62	3007.00
a	0.0000	0.2400	0.0633	0.0591
b	0.0000	38.7799	0.4993	0.2171
N_{param}	35	3345	243.58	214.00
$(N_{obs} - N_{param})/N_{obs}$	0.5617	0.9916	0.9253	0.9290
`_exptl_crystal_size_max`	0.0300	10.280	0.3607	0.3300
`_diffrn_reflns_av_R_equivalents`	0.0000	0.3610	0.0426	0.0352
`_refine_ls_R_factor_all`	0.0182	0.3024	0.0759	0.0685
`_refine_ls_wR_factor_ref`	0.0483	0.3664	0.1324	0.1260

† All releases including *SADABS* 1996, with the largest fraction of 1919 data sets. ‡ Includes releases from Agilent and Oxford Diffraction: *CrysAlis PRO* Agilent 2010, *CrysAlis PRO* Agilent 2011, *CrysAlis PRO* Agilent 2012, *CrysAlis PRO* Agilent 2013, *CrysAlis PRO* Agilent 2014, *CrysAlis PRO* Oxford Diffraction 2009, *CrysAlis PRO* Oxford Diffraction 2010, *CrysAlis PRO* Rigaku OD 2015, *CrysAlis RED* Oxford Diffraction 2006, *CrysAlis RED* Oxford Diffraction 2007, *CrysAlis RED* Oxford Diffraction 2008, *CrysAlis RED* Oxford Diffraction 2009, *CrystalClear* Rigaku MSC 2005, *CrystalClear* Rigaku 2005.

orthorhombic space group $Pna2_1$, $T = 180$ K, COD 2212154, Khan *et al.* (2007)] and 0.3664 [hexamethylenetetramine, $C_6H_{12}N_4{\cdot}2C_8H_8O_2$, monoclinic space group $P2_1/n$, $T = 173$ K, COD 2104946, Lemmerer (2011)], with in total 19 data sets with $wR(F^2) > 0.3000$ and 483 data sets with $wR(F^2) > 0.2000$.

3. Mean and median of the weighted residuals increase simultaneously

Fig. 1(*a*) shows the histogram for the mean value of the weighted residuals $\langle\zeta\rangle$ from the above-described sample of $N = 8424$ structures with light atoms C, H, O and N only. The brackets indicate averaging of the weighted residuals over individual data sets such that for every crystallographic data set one mean value of the weighted residuals $\langle\zeta\rangle$ is obtained. The mean value over all data sets $\overline{\langle\zeta\rangle} = 0.0521$, where the overline indicates now averaging over all data sets in the sample, is slightly larger than zero. This may appear to be a small value; however, Fig. 1(*b*) shows the histogram of the significance of the deviation from zero for the 8424 analysed data sets. A minority of only 45.45% are within the boundaries of $\pm 3\sigma$, as indicated by the black vertical lines; 41.79% of all data sets show a significance larger than plus three, and 12.76% show a significance less than minus three. Positive outliers appear 3.27-fold more frequently than negative outliers and in total the 'outliers' are the majority. In Fig. 1(*c*), the median of the weighted residuals is plotted against the mean value of the weighted residuals for each data set. A large correlation between these entities is obvious from the plot. This observation is a first hint that for each individual data set the distribution of weighted residuals is shifted *as a whole*, *i.e.* the mean value is *not* mainly determined by a few strong positive outliers. The shift of the residual distribution as a whole can also be described by the fraction of positive excess residuals [Fig. 1(*d*)]. A well centred Gaussian distribution results in approximately 50% positive residuals $\zeta > 0 \coloneqq \zeta_+$ and 50% negative residuals $\zeta < 0 \coloneqq \zeta_-$. The difference between the integer numbers of positive and of negative residuals $\#\zeta_+ - \#\zeta_-$ divided by the total number of weighted residuals N_{obs} (called 'the fraction of positive excess residuals') is in this case a number close to zero and is accidentally sometimes slightly larger and sometimes slightly smaller than zero. But the number of positive excess residuals increases with increasing mean value of the weighted residuals $\langle\zeta\rangle$, which confirms and quantifies the observation from plot (*c*) that the shift of the distribution of weighted residuals is driven mainly by shifting the distribution as a whole, rather than by strong outliers.

When the 483 data sets with $wR(F^2) > 0.2000$ are excluded, the following values result: $\overline{\langle\zeta\rangle} = 0.0501$ (to be compared with 0.0521 for all data sets), median$\langle\zeta\rangle = 0.0384$ (to be compared with 0.0398 for all data sets), $\overline{\langle\zeta\rangle/\sigma(\langle\zeta\rangle)} = 2.8318$ (to be compared with 2.9477) and median$[\langle\zeta\rangle/\sigma(\langle\zeta\rangle)] = 2.0484$ (to be compared with 2.1287), *i.e.* the values are all slightly reduced, but the qualitative picture does not change.

3.1. Over- and underestimation of I_{obs} by δ

When I_{obs} denotes the reflection intensity in the reflection input file, I_{true} denotes the (unknown) true intensity excluding random noise, $\pm\Delta$ denotes random noise and δ denotes a constant systematic offset for all reflections, *e.g.* from a calibration error, and when no other errors apply, the resulting intensity is given by

$$I_{obs} = I_{true} \pm \Delta + \delta. \quad (1)$$

The symbol $\pm\Delta$ was used to emphasize the stochastic nature of noise with equal probabilities for positive and for negative fluctuations. Its amplitude is characterized by the estimated standard uncertainty of the observed reflection s.u.(I_{obs}) in the reflection input file. Stochastic noise is not affected by a constant shift of origin. With this notation, when $\delta = 0$ for all reflections, the observed intensity is unbiased with respect to the true intensity when averaging over the noise:

$$\langle I_{obs}\rangle = \langle I_{true} \pm \Delta\rangle = \langle I_{true}\rangle. \quad (2)$$

When, in contrast, a systematic offset $\delta \neq 0$ applies,

$$\langle I_{\rm obs}\rangle = \langle I_{\rm true} \pm \Delta + \delta\rangle = \langle I_{\rm true}\rangle + \delta \qquad (3)$$

and the observed intensity is not unbiased anymore. It is affected by a systematic shift δ.

A constant and for all reflections equal positive or negative offset δ may be seen as the extreme case of a whole class of systematic errors where the origin for only a subgroup of observed intensities, for example from resolution or exposure time batches, is shifted (origin drift) or where non-linearities in area detectors lead to spatial or intensity-dependent origin drifts. Distinct spatial inhomogeneities in detector responses were reported earlier (Paciorek *et al.*, 1999; Pflugrath, 1999; Dudka, 2018). Insufficient or missing absorption correction procedures may affect specifically those reflections with the longest path through the crystal, though depending on the linear absorption coefficient of the crystal. Time-dependent drifts of the origin may occur from decreasing or fluctuating beam intensity, crystal decay, or just insufficient or missing correction for changing irradiated crystal volume. These errors may lead to different individual shifts depending on the coordinates of the detecting pixel in the detector ($x_{\rm det.}$, $y_{\rm det.}$), resolution, intensity, beam profile or exposure time, $\delta = \delta(x_{\rm det.}, y_{\rm det.}, t, \sin\theta/\lambda, I)$, and on geometry parameters.

All of the mentioned errors may result in an average shift $\langle\delta\rangle$ of the reflection intensities. So all of these systematic errors may be kept in mind when only a constant value δ is discussed for simplification as it can be regarded as the total net effect of individual errors $\delta(x_{\rm det.}, y_{\rm det.}, t, \sin\theta/\lambda, I) \rightarrow \langle\delta\rangle$. To further

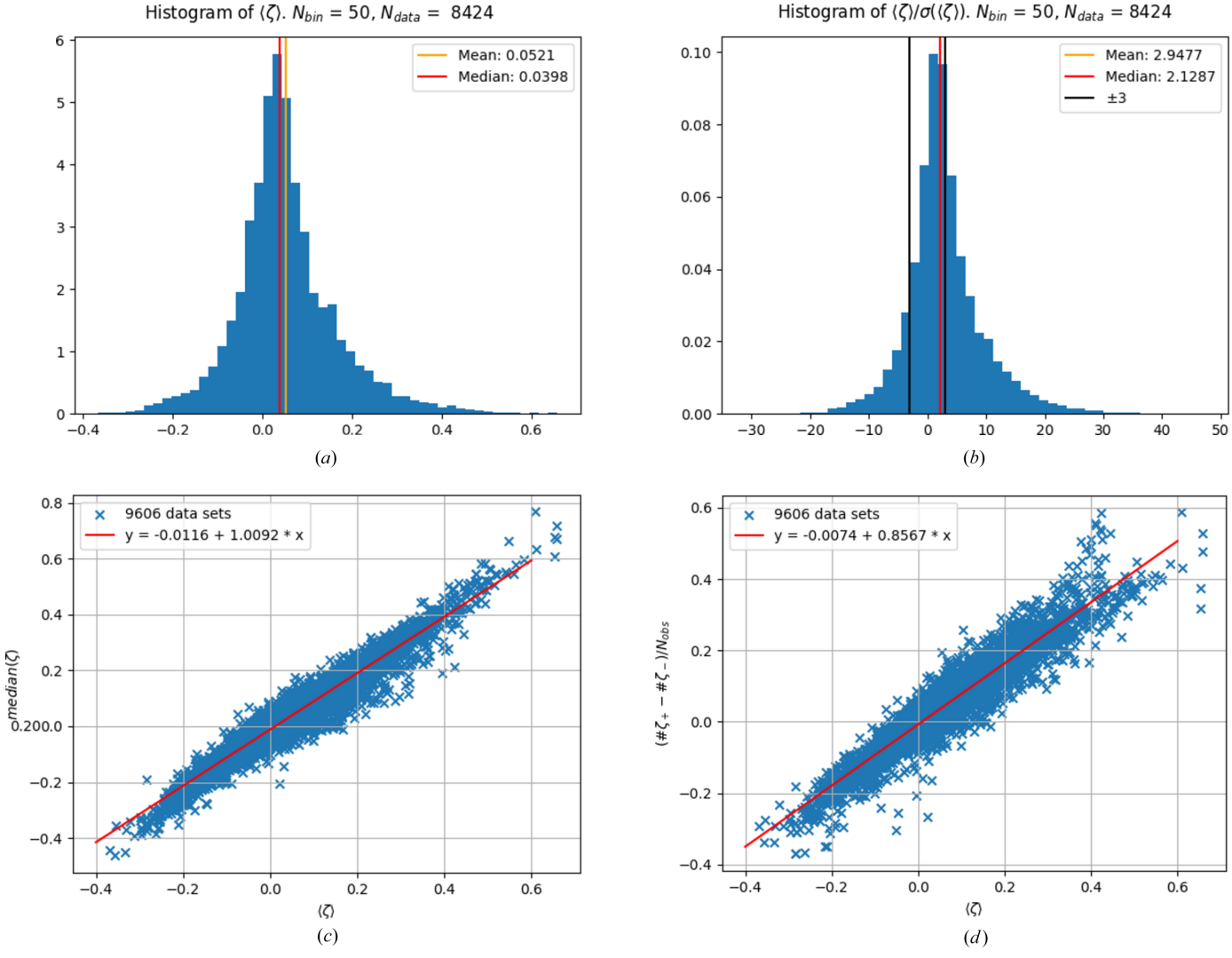


Figure 1
(*a*) The mean value of the weighted residuals is not at zero but at a slightly larger value of 0.0521. This appears to be a small value; however, plot (*b*) shows the histogram of the significance of the deviation from zero for 8424 analysed data sets. (*c*) The mean value of the weighted residuals is strongly correlated with the median of the weighted residuals. The 95% confidence intervals for the fit parameters are given by [−0.012, −0.011] for the constant and [1.003, 1.015] for the slope. (*d*) The shift of the residual distribution as a whole can also be described by the fraction of positive excess residuals. When this fraction is multiplied by 100 it gives the percentage of positive excess residuals. When the distribution of residuals for an individual data set is well centred at zero, the percentage of positive excess residuals is close to zero. The fitted parameters are given in the plot with the respective estimated standard deviations. The 95% confidence intervals for the fitted parameters are given by [−0.008, −0.007] for the constant and [0.850, 0.863] for the slope. Plots (*c*) and (*d*) together may indicate data-processing or -reduction problems. For details, see text.

simplify the discussion, all of these errors are summarized under the keywords 'overestimation' and 'underestimation' of the observed intensity. Fig. 1 indicates a dominance of overestimation of $I_{\rm obs}$, so the focus is on overestimation.

A slight – but systematic – overestimation from, say, data-integration steps would not necessarily be visible in the individual data set, where other systematic errors may overlie and mask this specific small error. Additionally, a slight over- or underestimation $I_{\rm obs} = I_{\rm true} \pm \Delta + \delta$ would easily explain the strong correlation between median(ζ) and $\langle\zeta\rangle$ as depicted in Fig. 1(*c*), as even small errors $|\delta_{hkl}| < {\rm s.u.}(I_{{\rm obs},hkl})$ in the abundant weak data would immediately lead to small linear positive and negative changes in the median of the residuals, just as observed in Fig. 1(*c*).

The working hypothesis is therefore from here on that small but systematic errors in $I_{\rm obs}$ are an important factor in explaining Fig. 1(*d*). Overestimation of $I_{\rm obs}$ on average explains the overall appearance of the plots in Figs. 1(*c*) and 1(*d*) by accounting for the linear increase in $(\#\zeta_+ - \#\zeta_-)/N_{\rm obs}$ with increasing $\langle\zeta\rangle$.

But overestimation of $I_{\rm obs}$ on average also raises further questions: (i) Why would small errors more frequently lead to overestimation, $\langle\zeta\rangle > 0$, than to underestimation (negative shifts $\langle\zeta\rangle < 0$)? And (ii) would regular overestimation of $I_{\rm obs}$ not increase the residual factors and call for corrections in this way?

4. How overestimation of $I_{\rm obs}$ affects the residual factors

The last two questions have a surprising answer: The residual factors *decrease* when the observed intensities are slightly *over*estimated – and they tend to *increase* when the observed intensities are slightly *under*estimated. This may unconsciously incentivize overestimation of $I_{\rm obs}$ rather than underestimation in detector calibration experiments or when setting default data-integration parameter values. In the following section an example will be used to briefly discuss how and why the residuals factors are lower when the observed intensities $I_{\rm obs}$ are overestimated. In order to give more evidence that the rewarding behaviour of the agreement factors is not just a theoretical idea but a very tangible thing, a simulation is performed with artificial data to prove this point, and traces in data from experiments are presented to further substantiate the working hypothesis.

4.1. Overestimation of $I_{\rm obs}$ reduces $R_{\rm merge}$, $R_{\rm r.i.m.}$ and $R_{\rm p.i.m.}$

As an example, the merging R factor, $R_{\rm merge}$, which is also called $R_{\rm sym}$, is briefly discussed. The importance of the merging R factor lies in its availability, as it is very frequently given in published data sets. As a metric for data quality it has severe weaknesses (Diederichs & Karplus, 1997). These were solved with the redundancy independent merging R factor $R_{\rm r.i.m.}$ and with the precision indicating merging R factor $R_{\rm p.i.m.}$ (Weiss, 2001). These important descriptors should be included as standard.

The merging R factor is defined according to

$$R_{\rm merge} = \frac{\sum_i \sum_{j=1}^{n_j} |I_j(hkl) - \langle I(hkl)\rangle|}{\sum_i \sum_{j=1}^{n_j} I_j(hkl)}, \qquad (4)$$

where n_j is the redundancy of the unique reflection i and where $\langle I(hkl)\rangle$ indicates the mean value over the redundantly measured reflection (Arndt *et al.*, 1968). The contribution from one arbitrarily chosen unique reflection i with redundancy n_j in the numerator is the sum $\sum_{j=1}^{n_j} |I_j(hkl) - \langle I(hkl)\rangle|$. Suppose each individual measurement carries the same small constant error $\delta > 0$. How does this affect the sum in the numerator? Obviously, it increases each individual reflection by the same amount $I_j(hkl) \rightarrow \tilde{I}_j(hkl) = I_j(hkl) + \delta$, and as a consequence also the average value by the same amount: $\langle I(hkl)\rangle \rightarrow \langle\tilde{I}(hkl)\rangle = \langle I(hkl) + \delta\rangle$. As these terms are subtracted, the sum remains unchanged:

$$\sum_{j=1}^{n_j} |I_j(hkl) + \delta - \langle I(hkl) + \delta\rangle| = \sum_{j=1}^{n_j} |I_j(hkl) - \langle I(hkl)\rangle|. \qquad (5)$$

A short way to state this fact is 'The sum in the numerator of equation (4) remains unchanged under a transformation $I_j(hkl) \rightarrow I_j(hkl) + \delta$' – it is *invariant* under such a transformation. This is actually a trivial statement; it just means that the difference between numbers that are increased by the same amount remains unchanged. This holds also for each individual term in the numerator and thus for the numerator of equation (4) in total. The denominator of equation (4), however, is *not* invariant under such a transformation. It increases by $n_j\delta$ for the unique reflection i:

$$\sum_{j=1}^{n_j} I_j(hkl) < \sum_{j=1}^{n_j} \left[I_j(hkl) + \delta\right] = \sum_{j=1}^{n_j} I_j(hkl) + \sum_{j=1}^{n_j} \delta. \qquad (6)$$

The last two equations taken together state that the merging R factor decreases when the observed intensities are overestimated by $\delta > 0$ as the numerator is unchanged and the denominator increases in this case. The merging agreement factor responds 'rewardingly' to the systematic error of overestimated intensities when a low value is perceived as desirable (confirmation bias). This holds also when not all of the redundantly measured intensities are overestimated by the exact same value δ, which was only assumed to simplify the discussion; it also holds when the observed intensities are overestimated by different amounts δ_j.[1]

So far, the discussion has assumed small errors δ. However, the conclusions also apply to large errors: indeed, the larger δ is, the smaller the merging R factor. In other words, the more the observed intensities are overestimated, the smaller $R_{\rm merge}$ gets.

[1] Within certain limits: The more uniform the distribution of δ_j values for a given set of redundantly measured intensities $\{I_1(hkl), I_2(hkl), \ldots, I_j(hkl)\}$ is, the more exactly is the sum in equation (5) invariant. Equation (5) holds exactly in the case of a calibration error, where basically the origin has the same offset δ for all reflections. For larger differences between individual values of δ_j, the described cancellation with the mean value does not work fully anymore, but the equation holds approximately when the set of δ_j is sufficiently uniform.

The demonstrated behaviour of the merging R factor to reward overestimation of I_{obs} with lower values holds also for $R_{r.i.m.}$ and $R_{p.i.m.}$, as these differ from R_{merge} only in the factors $[n_i/(n_i-1)]^{1/2}$ and $1/n_i^{1/2}$, respectively, in the numerator.

4.2. Overestimation of I_{obs} reduces $wR(F^2)$, R_1 and atomic displacement parameters

For the weighted agreement factor

$$wR(F^2) = \frac{\sum_{i=1}^{N_{ref}} w_i (I_{obs,i} - I_{calc,i})^2}{\sum_{i=1}^{N_{ref}} w_i I_{obs,i}^2} \quad (7)$$

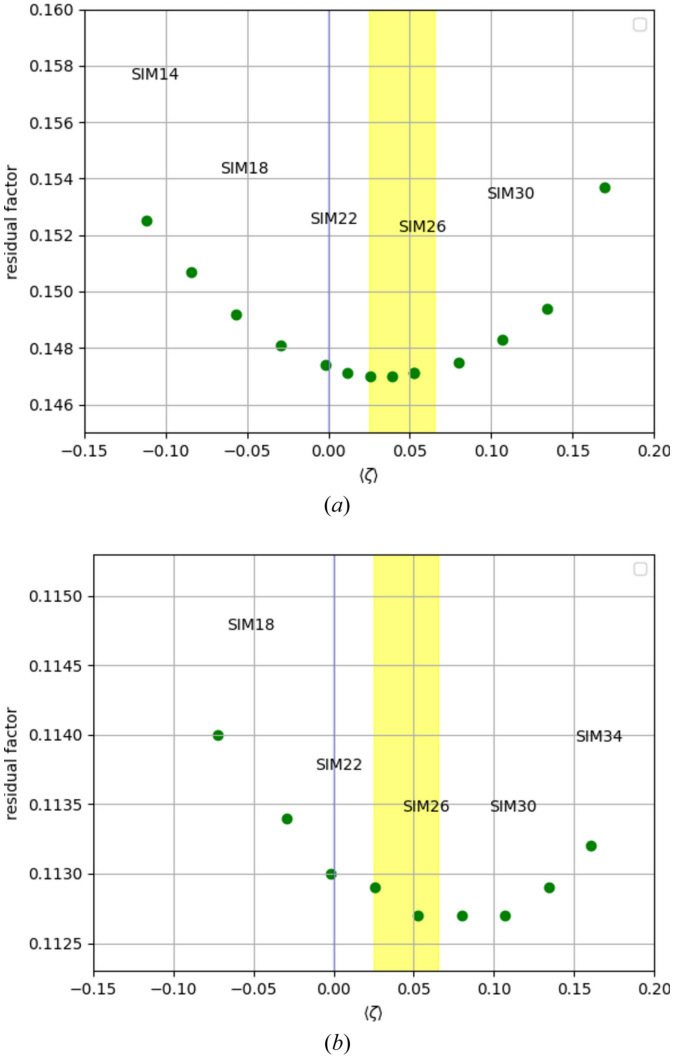


Figure 2
Simulations with artificial data. A calibration error is simulated by adding for each simulation a different constant amount δ to all observed intensities. The added amount is small in the sense that (i) it is smaller than 2 times the smallest value of $\sigma(I_{obs})$ and (ii) model refinement against the simulated data results in weighting scheme parameters $a = 0$ and $b = 0$ for SIM 14–SIM 30. For the exact values of δ for each simulated data set see Table 2 in Appendix A. (a) $wR(F^2)$ shows a minimum for slightly overestimated I_{obs}. The minimum is at approximately $\langle\zeta\rangle \approx 0.03$. ($b$) R_1 shows a minimum at a larger positive value $\langle\zeta\rangle \approx 0.08$. Both residual factors 'reward' overestimation of I_{obs}.

(with weights w_i and where N_{ref} is the number of reflections included in the refinement) and for the conventional R factor

$$R_1 = \frac{\sum_{i=1}^{N_{ref}} |F_{obs,i} - F_{calc,i}|}{\sum_{i=1}^{N_{ref}} |F_{obs,i}|} \quad (8)$$

one cannot expect exactly the same results, as the structure model is involved in these cases in the form of I_{calc} or structure factor F_{calc}. This is in contrast to R_{merge}, $R_{r.i.m.}$ and $R_{p.i.m.}$ where the observed entities are not compared with calculated entities but only with other observed entities. For the weighted agreement factor and the conventional agreement factor it is reasonable to assume that overestimation of the observed intensity initially also lowers the respective residual factor, until, at some point, the difference between the weakest observed and calculated entities starts to increase the residual sum at a rate that is larger than the rate of increase of the denominator.

In order to prove that the weighted agreement factor is initially decreasing as a response to a slight overestimation of I_{obs}, a simulation was performed with artificial data (for more details about the simulations see Appendix A). In the simulation we know the exact true values of the observed intensities, which are never known in an experiment. Additionally, the exact error δ is known, as are the true model parameter values. For the simulation, the calculated intensities were extracted after convergence of a refinement and written to a reflection input file. Gaussian random noise was added in proportion to the s.u.(I_{obs}) values from the experiment. In order to perform the simulation at a level of the weighted agreement factor that approximately compares with average experimental values, the noise was chosen to be 4 times s.u.(I_{obs}). A number of such reflection input files were generated. Different values δ were added in incremental steps in addition to the noise for different artificial data sets. The same starting model was refined against each of these resulting artificial reflection input files. The resulting residuals factors $wR(F^2)$ and R_1 are plotted in Figs. 2(a) and 2(b), respectively, against the mean value of the weighted residuals $\langle\zeta\rangle$. The lowest residual values are not attained for $\langle\zeta\rangle = 0$, as one might expect naively, but for $\langle\zeta\rangle > 0$. This proves that $wR(F^2)$ and R_1 are rewarding overestimation of I_{obs} similarly to R_{merge}, $R_{r.i.m.}$ and $R_{p.i.m.}$. The main effect of overestimation of I_{obs} on the structure model as obtained from the simulation is a systematic reduction in the atomic displacement parameter U_{eqiv}. As an example, the true value for the chlorine atom in the simulation described in Appendix A is $U_{eqiv} = 0.05793$. Increasing I_{obs} until $\langle\zeta\rangle = 0.1071$ leads to a reduction to $U_{eqiv} = 0.05729$, which corresponds to 2.56 standard deviations. The reduction in U_{eqiv} is systematic for all atoms and corresponds on average to 0.92 standard deviations.

5. Are these findings in accordance with the experimental data?

The theoretical considerations and the simulations from the previous sections showed that the residual factors R_{merge},

$R_{\rm r.i.m.}$ and $R_{\rm p.i.m.}$, $wR(F^2)$, and R_1 are lower when the true intensity is slightly overestimated compared with the case where the observed intensity is unbiased with respect to the true intensity. For $R_{\rm merge}$, $R_{\rm r.i.m.}$ and $R_{\rm p.i.m.}$ (and all other residual factors with a similar structure), this is easy to demonstrate theoretically, as in these cases only observed entities enter the defining equations. The numerators in all of these equations are composed of sums of absolute differences from observed intensities and mean values thereof. These differences remain unchanged when the observed intensities are overestimated, whereas the denominators in all of these equations increase. This leads to a reduction of the residual factors in response. In the case of $wR(F^2)$ and R_1, model-derived entities are involved. This changes the situation slightly, but not qualitatively, when only a small overestimation of $I_{\rm obs}$ is considered. These residual factors also decrease initially for overestimation of $I_{\rm obs}$. A qualitative difference is that they finally start to indicate the systematic error if the overestimation is sufficiently distinct, whereas for $R_{\rm merge}$, $R_{\rm r.i.m.}$ and $R_{\rm p.i.m.}$ this is not the case.

The theoretically derived and simulation-confirmed rewarding behaviour of the residual factors with respect to overestimation of $I_{\rm obs}$ may explain why there are so many more data sets with a positive shift of the mean value of the residuals compared with those with a negative shift, where, ideally, positive and negative shifts should be equally distributed and equally strong.

The answer to this riddle could lie in the fact that overestimation of $I_{\rm obs}$ is rewarded such that it remains undetected in the data (confirmation bias). As a consequence, this should be visible from the experimental data themselves.

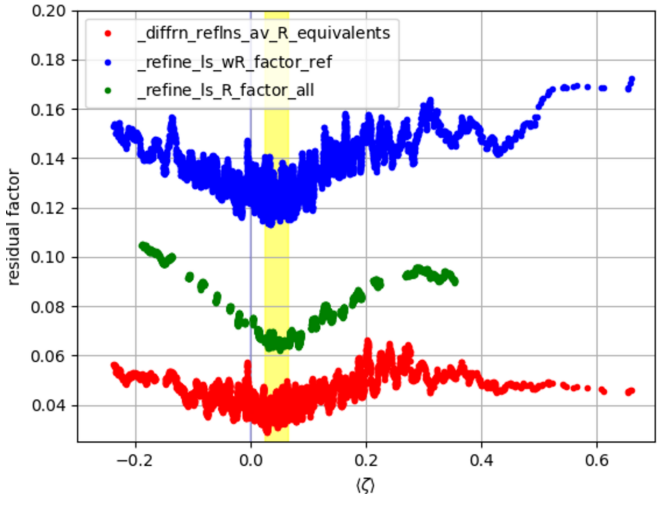


Figure 3
Residual factors (y axis) plotted as moving averages over a window of 50 consecutive data points as a function of $\langle\zeta\rangle$ (x axis). Blue: `_refine_ls_wR_factor_ref`. Red: `_diffrn_reflns_av_R_equivalents`. Green: `_refine_ls_R_factor_all`. The plot contains empty spaces where individual data points were not available. The averaging continues only when 50 or more data sets were available in a row with the desired value due to the chosen window. Different residual factors show their minimum values also for $\langle\zeta\rangle > 0$. The common minimum area for the three residual factors is shown as a yellow stripe.

Fig. 3 shows the residual factors $R_{\rm merge}$ (`_diffrn_reflns_av_R_equivalents`, red), $wR(F^2)$ (`_refine_ls_wR_factor_ref`, blue) and R_1 (`_refine_ls_R_factor_all`, green), plotted as moving averages (with a window of 50 consecutive data points) as a function of the mean value of the weighted residuals. The common area in which all three residual factors reach their respective minimum value is shown as a yellow stripe in the range $0.025 \leq \langle\zeta\rangle \leq 0.065$. The same yellow stripe is also depicted in Fig. 2. This confirms again the overall slight overestimation of $I_{\rm obs}$.

6. Discussion and outlook

It was shown that the shift of the mean value of the residuals correlates strongly with the median of the residuals. This strong correlation is interpreted as a sign that the mean value of the weighted residuals is determined by a shift of the distribution of the residuals *as a whole* rather than by a small number of strong outliers. In other words, the abundant weak data in each data set may have a stronger influence on the mean value of the residuals than individual large outliers from other, more conventional errors, such as a slight unmodelled disorder or neglect of bonding density. This holds in particular when the abundant weak intensities are slightly over- or underestimated. The individual under- or overestimation of $I_{\rm obs} = I_{\rm true} \pm \Delta \pm \delta$ may well be within the limits of noise $|\delta| \ll |\Delta| \approx \sigma(I_{\rm obs})$ – it is virtually invisible on the level of the individual reflection – and may nevertheless influence the mean value of the weighted residuals as these many small but systematic errors δ accumulate. It was shown with the help of artificial data that slight overestimation of $I_{\rm obs} = I_{\rm true} \pm \Delta + |\delta|$ leads to a *lower* weighted agreement factor than would be obtained with the unbiased, true values of $I_{\rm obs} = I_{\rm true} \pm \Delta$. The true intensity is available in a simulation in contrast to an experiment. Note that overestimation of $I_{\rm obs}$ may already occur at the step of data integration and data processing.

It was furthermore shown that these ideas are not merely theoretical but are confirmed by experimental data. The minimum values for different residual factors are found in the sample of experimental data sets again for a slightly positive mean value of the weighted residuals – they are *not* centred around $\langle\zeta\rangle = 0$. This is taken as confirmation (i) that these residual factors respond in a rewarding way to slight overestimation of $I_{\rm obs}$ and (ii) that the rewarding behaviour may constitute an (unconscious) incentive for overestimation.

In the simulation presented in this study, under- and overestimation of $I_{\rm obs}$ was modelled by adding the same small amount $\pm n\delta$ to each reflection. Different increments n resulted in different simulated data sets. This error is of course a simplified model of systematic errors in real experiments, where errors may be much more complicated. For discussion purposes, however, and for working out the consequences, it is a valid model. Also, detectors need to be calibrated. The calibration itself comes with an error, even if it is small. A calibration error offsets the origin for all reflections by the same value. Therefore, such an error may describe a real-

Table 2
Summary of the simulated data sets, including the increment n for the applied offset leading to the total offset δ_{SIM}; maximum, minimum and average increments relative to $\sigma(I_{obs})$; the total change relative to F_{000}^2; the resulting weighting-scheme parameters a and b; minimum and maximum simulated intensities; mean weighted residuals; and the significance of the mean weighted residuals.

The weighting-scheme parameters remain zero even for highly significant deviations, as the applied offset results in many small residuals rather than large residuals.

	n	δ_{SIM}	$[\delta_{SIM}/\sigma(I_{obs})]_{max}$	$[\delta_{SIM}/\sigma(I_{obs})]_{min}$	$[\delta_{SIM}/\sigma(I_{obs})]_{av}$	$N_{obs}\delta_{SIM}/F_{000}^2$	a	b	$(I_{SIM})_{min}$	$(I_{SIM})_{max}$	$\langle\zeta\rangle$	$\langle\zeta\rangle/\sigma(\langle\zeta\rangle)$
SIM 14	−8	$-0.0032F_{000}$	−0.0001	−1.8420	−0.1929	− 0.0163	0.00	0.00	−5.81	9899.20	−0.1117	−7.7025
SIM 18	−4	$-0.0016F_{000}$	−0.0001	−0.9210	−0.0964	− 0.0081	0.00	0.00	−5.69	9899.31	−0.0566	−3.9511
SIM 22	0	$0.0000F_{000}$	0.0000	0.0000	0.0000	0.0000	0.00	0.00	−5.58	9899.42	−0.0019	−0.1321
SIM 26	4	$0.0016F_{000}$	0.9210	0.0001	0.0964	0.0081	0.00	0.00	−5.47	9899.53	0.0527	3.6798
SIM 30	8	$0.0032F_{000}$	1.8420	0.0001	0.1929	0.0163	0.00	0.00	−5.36	9899.64	0.1071	7.3924

world case quite accurately, even if it seems a little idealized and artificial at first glance.

A calibration error explains the simultaneous increase of the mean of the residuals with the number of positive excess residuals. The artificial lowering of agreement factors and of atomic displacement parameters by overestimation of I_{obs} explains why this error was overlooked for such a long time (confirmation bias). Overestimation of weak data in small-molecule single-crystal experiments was found and discussed earlier in the context of measurement strategies (Williams *et al.*, 2019) for low-exposure-time and high-resolution data. The current results indicate that the problem might be much more widespread.

6.1. Factors contributing to overestimation of I_{obs} from the perspective of metrics

Unintentionally allowing for a slight overestimation $I_{obs} > I_{Bragg}$ of the observed intensities may unknowingly be supported by leading to more desirable results. A typical situation for confirmation bias is (i) lower residual factors R_{merge}, $R_{r.i.m.}$, $R_{p.i.m.}$, $wR(F^2)$, R_1 and similar and (ii) the atomic displacement parameters indicate smaller amplitudes, as briefly mentioned in Section 4.2. It is intuitive that small atomic displacement parameters are associated with high data quality as systematic errors tend to accumulate in displacement parameters by increasing them. For this reason, atomic displacement parameters from X-ray diffraction experiments are sometimes compared with results from neutron diffraction experiments [as an example see Chodkiewicz & Woźniak (2025)] in order to give evidence for the accuracy of the X-ray refinement. Neutron diffraction experiments have a reputation of being of a higher accuracy; however, they may also be affected by slight systematic over- or underestimation of intensities.

For either X-ray or neutron diffraction experiments, small atomic displacement parameters may arise from effects that artificially increase $I_{obs} > I_{Bragg}$ and thus may just be an artefact. Systematic errors that artificially reduce atomic displacement parameters may need to be excluded in order to ensure that small atomic displacement parameters are physically meaningful and not an artefact in X-ray and neutron diffraction experiments.

APPENDIX *A*
The simulations

The data set collected by Shraddha *et al.* (2020) was selected from the sample. It describes a structure crystallizing in the monoclinic space group $P2_1/n$ and contains a moiety belonging to the imidazoles ($C_{29}H_{23}ClN_2O$). The measurement was conducted on a Bruker diffractometer with Mo $K\alpha$ radiation at 297 K.

For preparing the simulation, first the refinement was repeated with the command `OMIT -100` but no other changes. This ensures that all reflections are taken into account, including large negative intensity observations. From the resulting reflection list file, the calculated intensities were extracted. They correspond to the true Bragg intensity without any noise. For the reference simulation SIM 22, noise was added to each reflection in proportion to the s.u.(I_{obs}) from the experimental data. The model parameters resulting from this reference refinement are defined to be the true model parameters and correspond to a refinement against observed intensities unbiased with respect to the true intensities as described in equation (2).

For refinements including a systematic shift δ of the origin like for a calibration error, equation (3), n increments of $\delta_{SIM} = \pm 0.0004F_{000}$ were added to all I_{true} and exactly the same noise was added. For example, when the noise was +1.234 for reflection 234 in SIM 22, it was also +1.234 in all other simulations (SIM 14, SIM 18, SIM 26, SIM 30). This is to make sure that the results do not depend on the set of random numbers. In order to monitor the effect on the individual reflection, the maximum, minimum and mean value for $\delta_{SIM}/\sigma(I_{obs})$ was calculated for each individual reflection in each simulated data set. The aim was to make the added systematic error insignificant with respect to $\sigma(I_{obs})$ such that it would not give rise to large residuals. SIM 22 is the reference simulation with no systematic error, $\delta_{SIM\ 22} = 0$. Simulations with numbers >22 have the increment added, while the others have it subtracted. For example, in SIM 26, $\delta_{SIM\ 26} = +0.0016F_{000}$ for each reflection ($n = 4$ as $26 - 22 = 4$), and in SIM 30 $\delta_{SIM\ 30} = +0.0032F_{000}$. The largest change for an individual reflection in SIM 30 due to the systematic error was $(\delta_{SIM\ 30})_{max} = 1.8420\sigma(I_{obs})$, well below three standard deviations of the observed intensity. Similarly, for SIM 14, the largest change was negative, $\delta_{SIM\ 14} = -1.8420\sigma(I_{obs})$.

The average change due to systematic errors was a reduction of the individual reflections by $0.1929\sigma(I_{obs})$ for SIM 14 and correspondingly an increase by $0.1929\sigma(I_{obs})$ for SIM 30. This corresponds for SIM 30 to adding to all reflections in a uniform way the total amount of additional scattering mass of 1.63% of F_{000}^2, which leaves the relative maximum simulated intensities virtually unaffected when compared with the true value from SIM 22. This leads to maximum changes in the weakest reflections by approximately only 4% when compared with the true value of SIM 22. Overall it can be said that the changes induced by the simulation are all small and most likely much smaller than other errors, for example due to a fluctuating beam intensity. The weighting scheme parameters *a* and *b* remain consequently zero after invoking the weighting scheme. Despite each individual distortion from the true intensity being small, the overall effect of these many small but one-sided distortions adds up to a measurable effect (Table 2).

A shift of the weighted residuals by 0.0527 like in SIM 26 may appear small; however, it is already significant at 3.6798σ. The significance is given by dividing the mean value by the standard deviation of the mean value, $\sigma(\langle\zeta\rangle) = [\mathrm{var}(\zeta)/N_{obs}]^{1/2}$, where $\mathrm{var}(\zeta) = [\sum(\zeta - \langle\zeta\rangle)^2]/(N-1)$ is the unbiased sample variance. The larger N_{obs}, the smaller the standard deviation of the mean value. The simulations SIM 14–SIM 30 all have the same $N_{obs} = 4807$. $\langle N_{obs}\rangle = 3534.62$ for the 8424 data sets from the sample described in Section 2.

Acknowledgements

JH thanks S. Mebs for bringing 'confirmation bias' to his attention. Open access publishing facilitated by Politecnico di Milano, as part of the Wiley–CRUI-CARE agreement.

References

Aristov, M. M., Geng, H., Harris, J. W. & Berry, J. F. (2023). *Acta Cryst.* C**79**, 133–141.
Arndt, U., Crowther, R. & Mallett, J. (1968). *J. Phys. E Sci. Instrum.* **1**, 510–516.
Barnett, S. A., Hulme, A. T. & Tocher, D. A. (2005). *Acta Cryst.* E**61**, o857–o859.
Bhat, M. A., Abdel-Aziz, H. A., Ghabbour, H. A., Hemamalini, M. & Fun, H.-K. (2012). *Acta Cryst.* E**68**, o1135.
Blessing, R. H. (1987). *Crystallogr. Rev.* **1**, 3–58.
Blessing, R. H. (1995). *Acta Cryst.* A**51**, 33–38.
Blessing, R. H. (1997). *J. Appl. Cryst.* **30**, 421–426.
Chetioui, S., Boudraa, I., Bouacida, S., Bouchoul, A. & Bouaoud, S. E. (2013). *Acta Cryst.* E**69**, o1322–o1323.
Chodkiewicz, M. & Woźniak, K. (2025). *IUCrJ* **12**, 74–87.
Diederichs, K. & Karplus, P. A. (1997). *Nat. Struct. Mol. Biol.* **4**, 269–275.
Domagala, S., Nourd, P., Diederichs, K. & Henn, J. (2023). *J. Appl. Cryst.* **56**, 1200–1220.
Downs, R. T. & Hall-Wallace, M. (2003). *Am. Mineral.* 88, 247–250.
Dudka, A. P. (2018). *Crystallogr. Rep.* **63**, 1051–1056.
Gelbrich, T., Zencirci, N. & Griesser, U. J. (2007). *Acta Cryst.* C**63**, o751–o753.
Görbitz, C. H. (2002). *Acta Cryst.* B**58**, 849–854.
Gražulis, S., Chateigner, D., Downs, R. T., Yokochi, A. F. T., Quirós, M., Lutterotti, L., Manakova, E., Butkus, J., Moeck, P. & Le Bail, A. (2009). *J. Appl. Cryst.* **42**, 726–729.
Gražulis, S., Daškevič, A., Merkys, A., Chateigner, D., Lutterotti, L., Quirós, M., Serebryanaya, N. R., Moeck, P., Downs, R. T. & Le Bail, A. (2012). *Nucleic Acids Res.* **40**, D420–D427.
Gražulis, S., Merkys, A., Vaitkus, A. & Okulič-Kazarinas, M. (2015). *J. Appl. Cryst.* **48**, 85–91.
Henn, J. (2019). *Crystallogr. Rev.* **25**, 83–156.
Henn, J. (2025). *Crystals* **15**, 898.
Ismiyev, A. I. (2011). *Acta Cryst.* E**67**, o1863.
Khan, U., Qureshi, R. A., Saeed, S. & Bond, A. D. (2007). *Acta Cryst.* E**63**, o530–o532.
Krause, L., Herbst-Irmer, R., Sheldrick, G. M. & Stalke, D. (2015). *J. Appl. Cryst.* **48**, 3–10.
Lemmerer, A. (2011). *Acta Cryst.* B**67**, 177–192.
Liang, W.-X. & Qu, Z.-R. (2008). *Acta Cryst.* E**64**, o1198.
Merkys, A., Vaitkus, A., Butkus, J., Okulič-Kazarinas, M., Kairys, V. & Gražulis, S. (2016). *J. Appl. Cryst.* **49**, 292–301.
Mesto, E., Scordari, F., Lacalamita, M., De Cola, L., Ragni, R. & Farinola, G. M. (2013). *Acta Cryst.* C**69**, 480–482.
Narendra Babu, S. N., Abdul Rahim, A. S., Abd Hamid, S., Balasubramani, K. & Fun, H.-K. (2009). *Acta Cryst.* E**65**, o2070–o2071.
Paciorek, W. A., Meyer, M. & Chapuis, G. (1999). *J. Appl. Cryst.* **32**, 11–14.
Pflugrath, J. W. (1999). *Acta Cryst.* D**55**, 1718–1725.
Quirós, M., Gražulis, S., Girdzijauskaitė, S., Merkys, A. & Vaitkus, A. (2018). *J. Cheminform* **10**, 23.
Saraswathi, B. S., Foro, S. & Gowda, B. T. (2011). *Acta Cryst.* E**67**, o1591.
Shraddha, K. N., Devika, S. & Begum, N. S. (2020). *IUCrData* **5**, x191690.
Thekku Veedu, S. & Techert, S. (2015). *Acta Cryst.* E**71**, o629–o630.
Vaitkus, A., Merkys, A. & Gražulis, S. (2021). *J. Appl. Cryst.* **54**, 661–672.
Vaitkus, A., Merkys, A., Sander, T., Quirós, M., Thiessen, P. A., Bolton, E. E. & Gražulis, S. (2023). *J. Cheminform.* **15**, 123.
Weiss, M. S. (2001). *J. Appl. Cryst.* **34**, 130–135.
Williams, A. E., Thompson, A. L. & Watkin, D. J. (2019). *Acta Cryst.* B**75**, 657–673.