

Metrics for crystallographic diffraction- and fit-data: a review of existing ones and the need for new ones

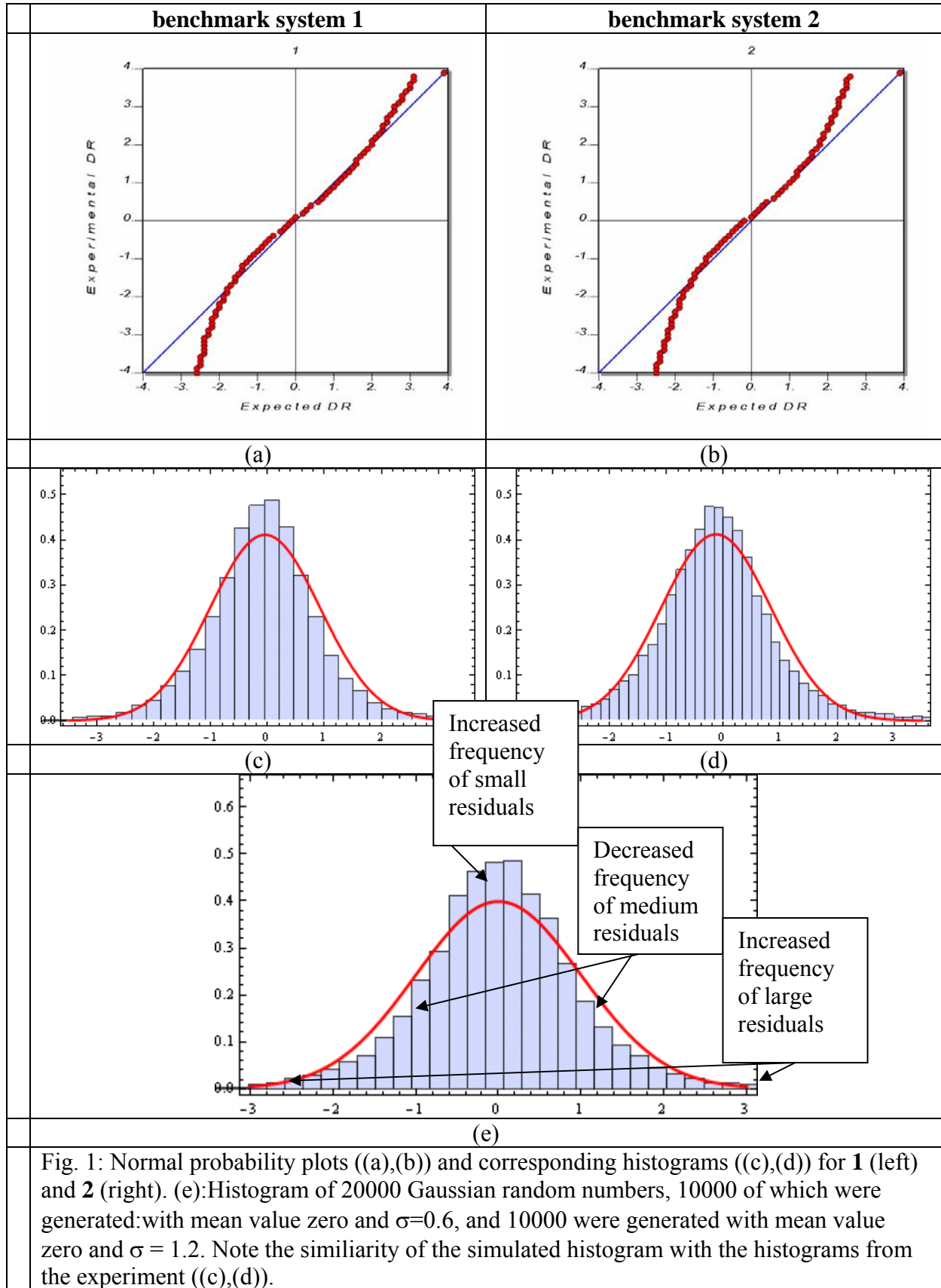
By Julian Henn

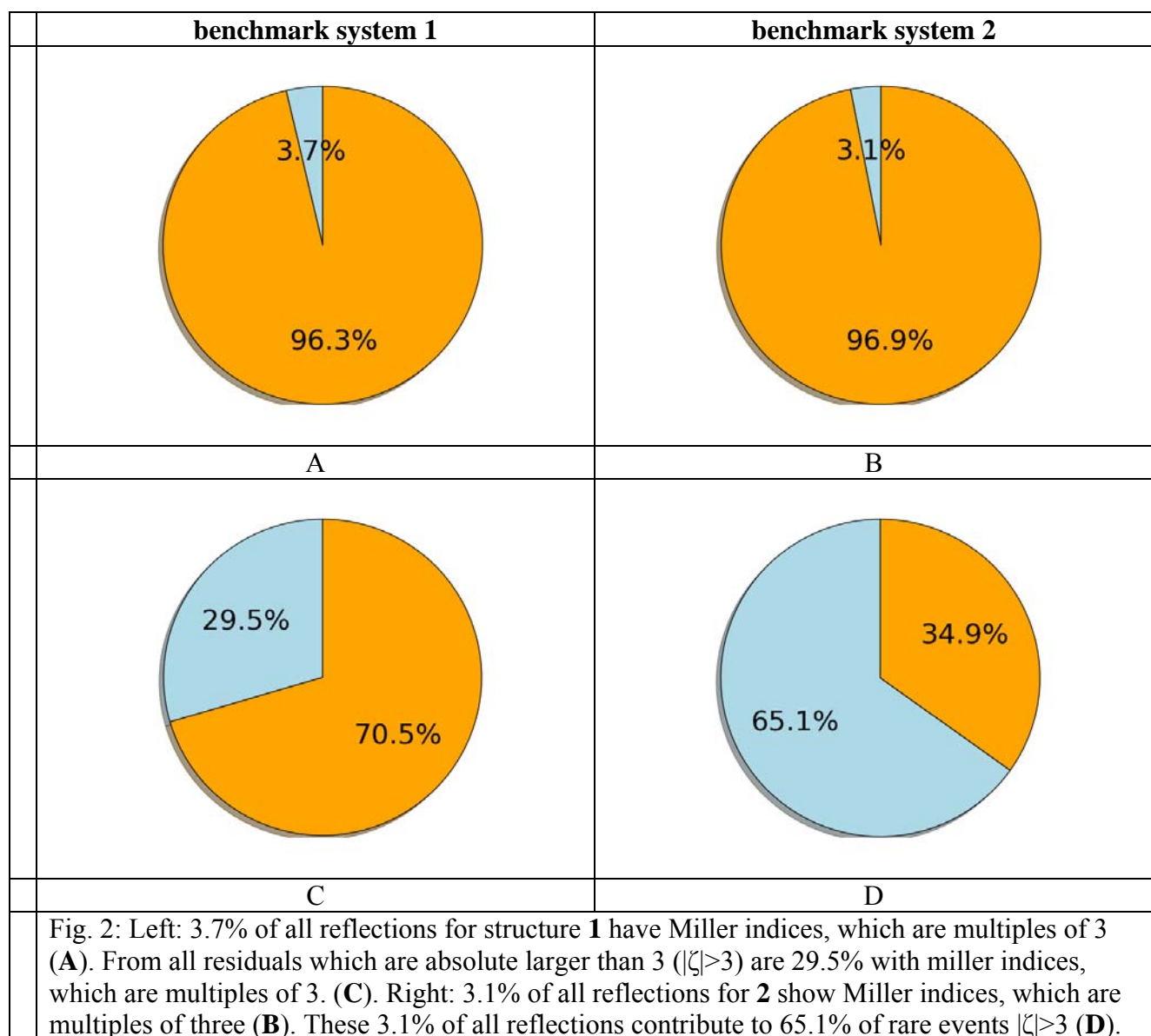
----Supplementary Material for Case Study 2---

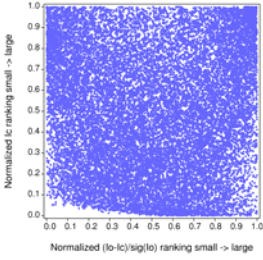
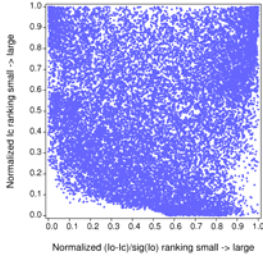
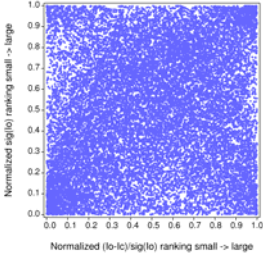
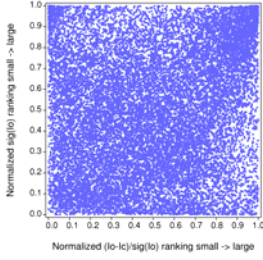
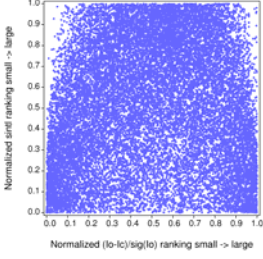
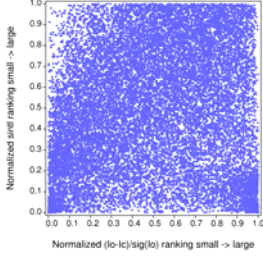
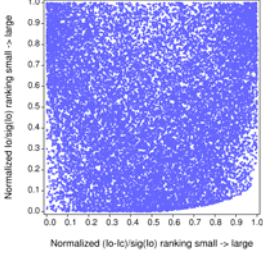
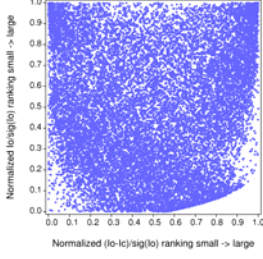
Overview:

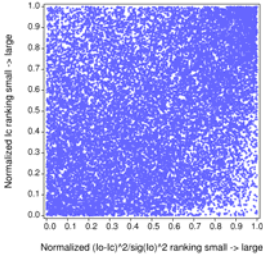
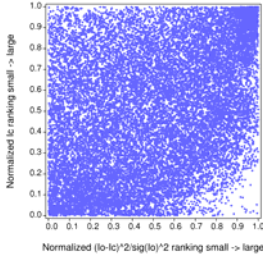
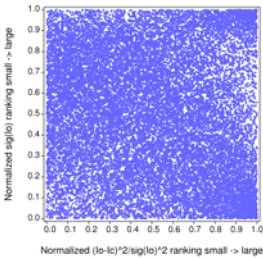
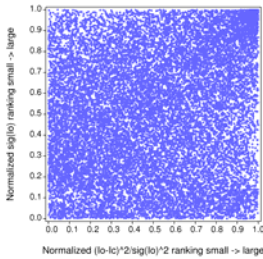
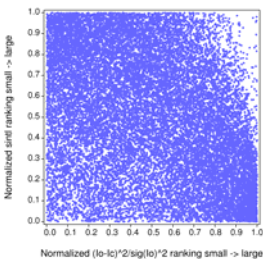
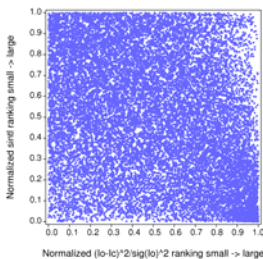
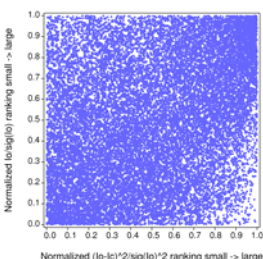
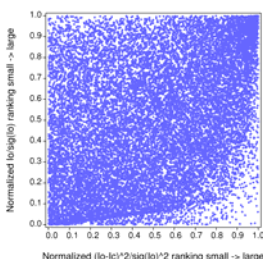
- Fig.1: Normal probability plots and corresponding histograms of the residuals for **1** and **2**. Additional simulation of the histogram of data generated from two Gaussians with different variances
- Fig.2: fraction of data with Miller triples being a multiple of 3 and fraction of rare events $|\zeta| > 3$ from these fractions for **1** and **2**
- Fig.3: BayCon plots (ζ, X) , $X = (I_c, \sigma, \text{sintl}, I_o / \sigma (I_o))$ and corresponding χ^2 values for **1** and **2**
- Fig.4: BayCon plots (ζ^2, X) , $X = (I_c, \sigma, \text{sintl}, I_o / \sigma (I_o))$ and corresponding χ^2 values for **1** and **2**
- Fig.5: Squared residuals in individual bins of the data sorted by significance $I_o / \sigma (I_o)$; observed intensity, I_o ; and resolution $(\sin \theta) / \lambda$ respectively for **1** and **2**
- Fig.6: Correlation coefficients $cc(\zeta^2, \sigma^2)$, $cc(\zeta^2, \Delta^2)$, $cc(\Delta^2, \sigma^2)$ for the data sorted in increasing order of the significance of the observed intensities for **1** and **2**
- Fig.7: Diagnostic plots for the neutron diffraction data set oxa14 from Kaminski et al

Description of a simulation for testing the standard deviations $\sqrt{1/N_{\text{ref}}}$ of the correlation coefficients with the help of random numbers.





	benchmark system 1	benchmark system 2
	<p>(lo-ζ)/sig(lo) vs ζ</p> 	<p>(lo-ζ)/sig(lo) vs ζ</p> 
	$\chi^2 = 2051.82$	$\chi^2 = 4340.90$
	<p>(lo-ζ)/sig(lo) vs sig(lo)</p> 	<p>(lo-ζ)/sig(lo) vs sig(lo)</p> 
	$\chi^2 = 1805.47$	$\chi^2 = 2289.91$
	<p>(lo-ζ)/sig(lo) vs sintl</p> 	<p>(lo-ζ)/sig(lo) vs sintl</p> 
	$\chi^2 = 3229.75$	$\chi^2 = 2685.92$
	<p>(lo-ζ)/sig(lo) vs lo/sig(lo)</p> 	<p>(lo-ζ)/sig(lo) vs lo/sig(lo)</p> 
	$\chi^2 = 2434.31$	$\chi^2 = 3663.24$
	<p>Fig. 3: BayCoN plots (ζ,X) for 1 (left) and 2 (right) with corresponding χ^2 test against uniformity of the plot. X = (ζ, σ, sintl, ζ/σ (lo)) from top to bottom. A value $\chi^2 < 149$ indicates a uniform distribution. A uniform distribution indicates a uniform joint probability distribution between residuals ζ and .X, hence no systematic connection of the residuals with the property X. Residuals, which are true random numbers as for a fit with no systematic errors whatsoever show no systematic connections.</p>	

	benchmark system 1	benchmark system 2
	<p>$(I_o - I_c)^2 / \text{sig}(I_o)^2$ vs I_c</p> 	<p>$(I_o - I_c)^2 / \text{sig}(I_o)^2$ vs I_c</p> 
	$\chi^2 = 2008.61$	$\chi^2 = 3514.61$
	<p>$(I_o - I_c)^2 / \text{sig}(I_o)^2$ vs $\text{sig}(I_o)$</p> 	<p>$(I_o - I_c)^2 / \text{sig}(I_o)^2$ vs $\text{sig}(I_o)$</p> 
	$\chi^2 = 1559.08$	$\chi^2 = 1406.77$
	<p>$(I_o - I_c)^2 / \text{sig}(I_o)^2$ vs sintl</p> 	<p>$(I_o - I_c)^2 / \text{sig}(I_o)^2$ vs sintl</p> 
	$\chi^2 = 3614.00$	$\chi^2 = 2434.81$
	<p>$(I_o - I_c)^2 / \text{sig}(I_o)^2$ vs $I_o / \text{sig}(I_o)$</p> 	<p>$(I_o - I_c)^2 / \text{sig}(I_o)^2$ vs $I_o / \text{sig}(I_o)$</p> 
	$\chi^2 = 2409.84$	$\chi^2 = 3532.21$
	<p>Fig. 4: BayCoN plots (ζ^2, X) for 1 (left) and 2 (right) with corresponding χ^2 test against uniformity of the BayCoN plot. $X = (I_c, \sigma, \text{sintl}, I_o / \sigma(I_o))$ from top to bottom. A value $\chi^2 < 149$ indicates a uniform distribution. A uniform distribution indicates a uniform joint probability distribution between squared residuals ζ^2 and X, hence no systematic connection of the squared residuals (strength of the residuals) with the property X.</p>	

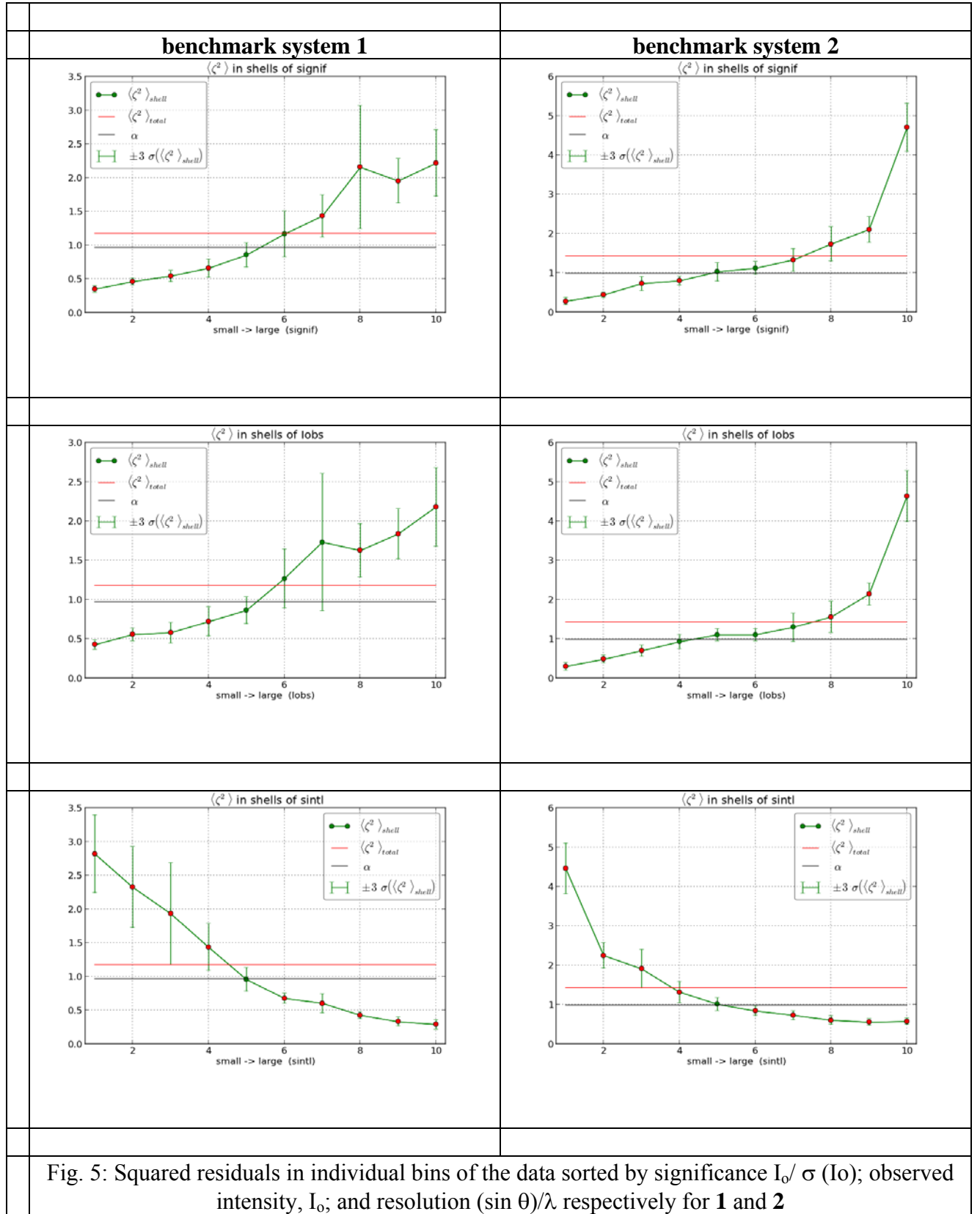
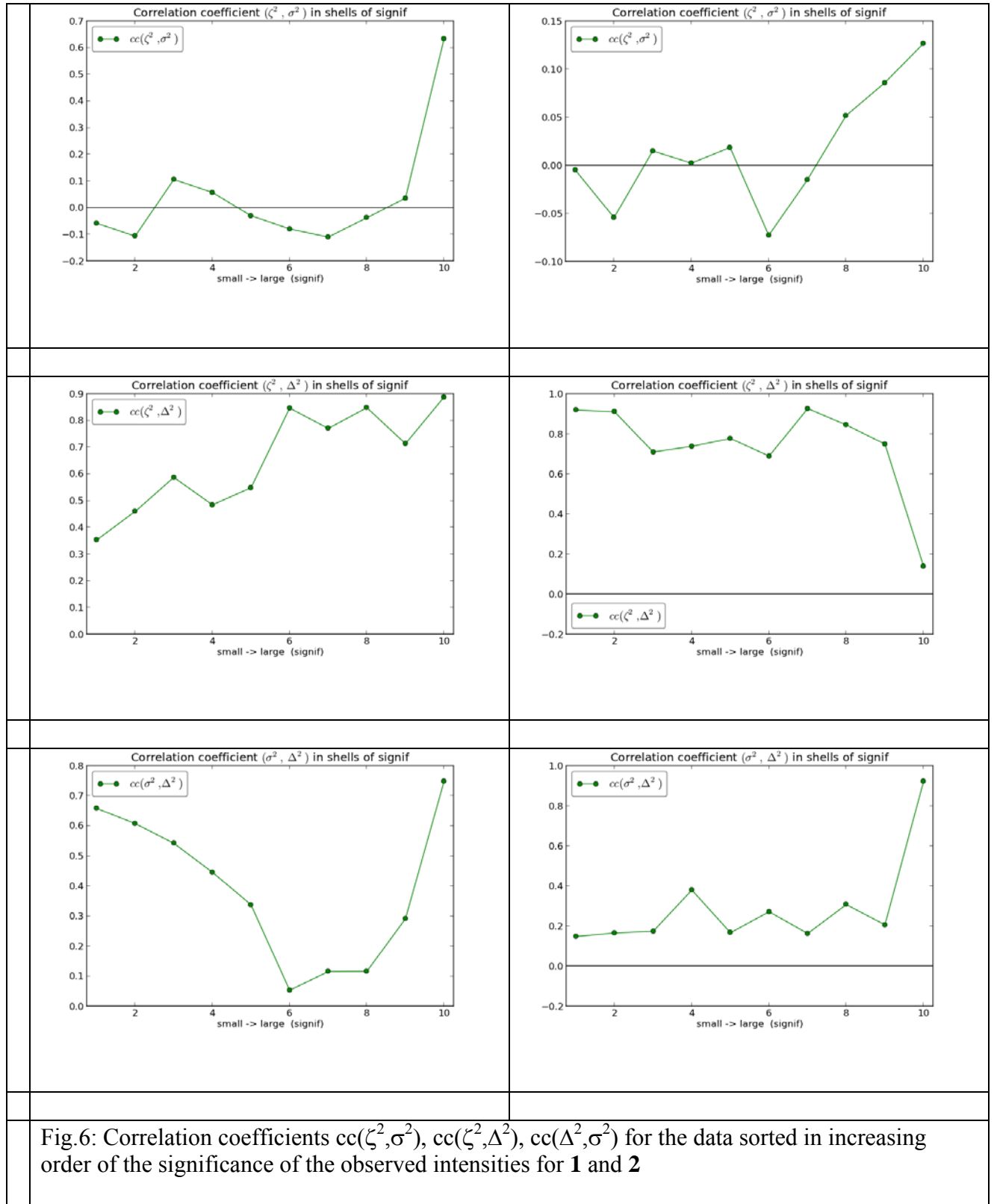


Fig. 5: Squared residuals in individual bins of the data sorted by significance $I_0 / \sigma(I_0)$; observed intensity, I_0 ; and resolution $(\sin \theta) / \lambda$ respectively for **1** and **2**



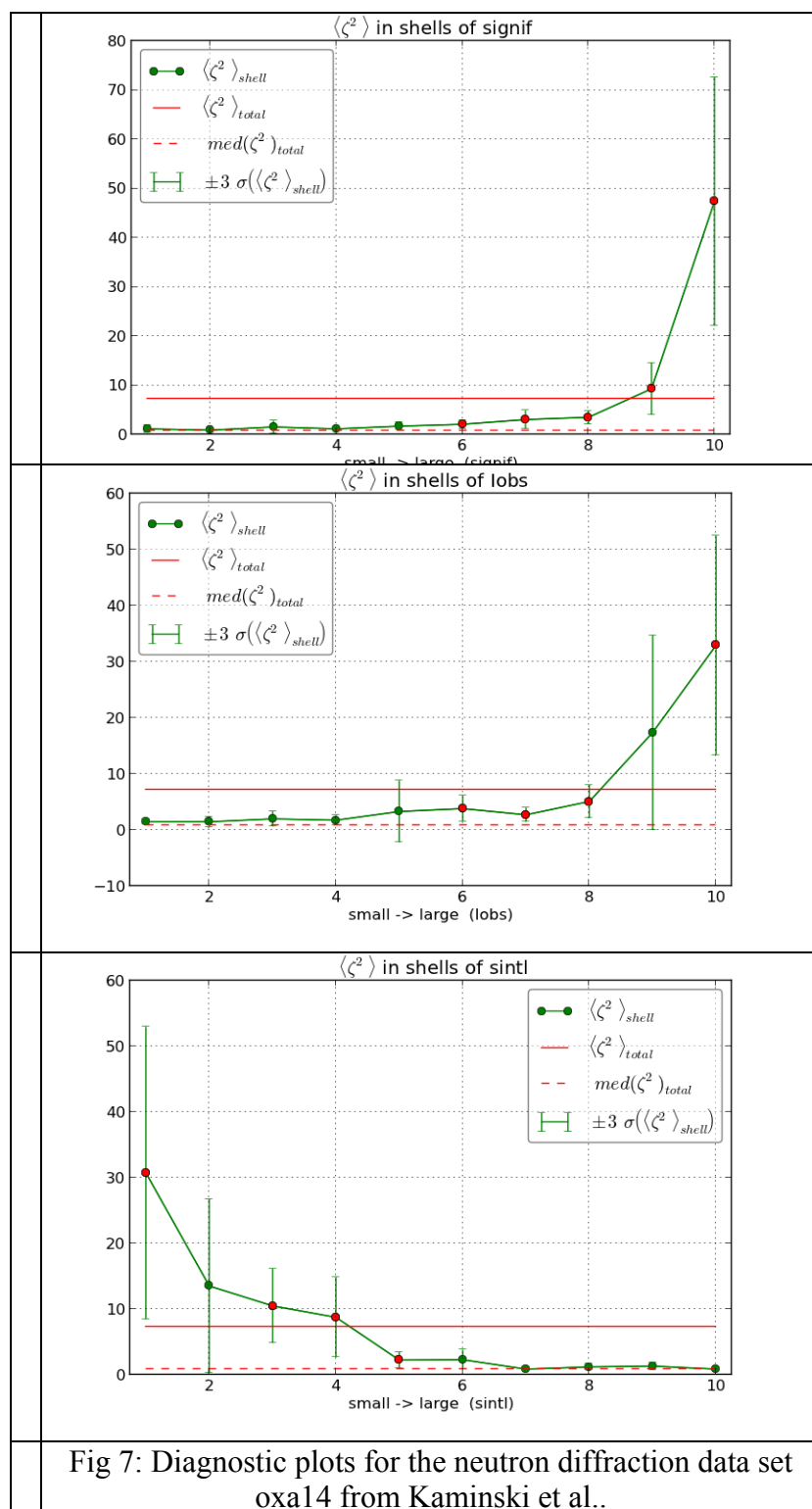


Fig 7: Diagnostic plots for the neutron diffraction data set oxa14 from Kaminski et al..

To investigate into the significance of the correlation coefficients between the squared residuals and the variances of the observed intensities, $cc(\zeta^2, \sigma^2)$, the following experiment was conducted for **1** and **2**:

- 1) The observed intensities I_o were extracted from the fco file together with their respective $\sigma(I_o)$ values
- 2) For each observed intensity a random number was generated with mean value zero and variance corresponding to the variance of the observed intensity.
- 3) This random number was added to the observed intensity and called “calculated intensity”, I_c
- 4) The squared residuals $(I_o - I_c)^2 / \sigma^2(I_o)$ were calculated and written to a list
- 5) In another list the corresponding $\sigma^2(I_o)$ values were written
- 6) A correlation coefficient between these two lists was calculated and written to a list of correlation coefficients $[cc(\zeta^2, \sigma^2)]$
- 7) Steps 2)-6 were repeated in total 500 times
- 8) The list of resulting correlation coefficients $[cc(\zeta^2, \sigma^2)]$ with 500 entries showed the following mean values and standard deviations:

	1	2
Nref	20711	18435
Mean $[cc(\zeta^2, \sigma^2)]$	0.000319477	-0.00011532
Variance $[cc(\zeta^2, \sigma^2)]$	0.0000481589	0.0000511962
Sqrt(Variance $[cc(\zeta^2, \sigma^2)]$)	0.00693966	0.00715515
Sqrt(1/Nref)	0.00694863	0.0073651

The following observations are made:

The mean correlation coefficient is very close to zero indeed in both cases

The corresponding square root of the variance of the list of correlation coefficients (the standard deviation) is indeed close to $\sqrt{1/Nref}$.

It is concluded that $\sqrt{1/Nref}$ is indeed a good estimator for the standard deviation of the correlation coefficient.