R-factor definitions in BUSTER

Snapshot release dated 11 December 2020

The calculation of R-factors in BUSTER has until now remained the same as described in the original 2004 paper by Blanc *et al.* (https://scripts.iucr.org/cgi-bin/paper?ba5067), namely an R-value, denoted Rxpct, based on the comparison of the observed amplitude Fobs to the expectation value Fxpct of the amplitude of the model-derived complex structure factor (i.e. the first moment of the Rice distribution for that amplitude). This is spelled out on p. 2216:

> **5.1. *R* factors**
>
> The *R* factors (both overall and in resolution bins) are computed using the expectation of the model structure amplitude rather than its calculated value, on the grounds that the expectation is the current model prediction for an observation. For reflections around resolution d^*,
>
> $$R[d^*] = \langle |F_{\mathbf{h}}^{\mathrm{xpct}} - F_{\mathbf{h}}^{\mathrm{obs}}| \rangle_{d^*} / \langle |F_{\mathbf{h}}^{\mathrm{obs}}| \rangle_{d^*}. \qquad (28)$$

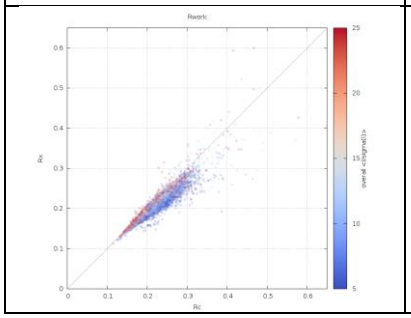
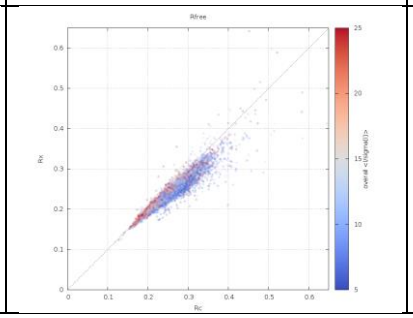
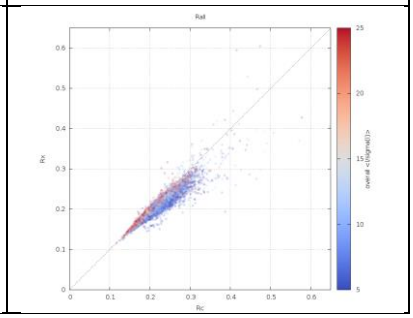
This definition differed from that of the conventional R-factor given in the IUCR Dictionary (https://dictionary.iucr.org/R_factor) and we initially took care of calling attention to this difference in early papers using BUSTER that we co-authored – see for instance the 2004 *Structure* paper (https://www.cell.com/structure/pdf/S0969-2126(04)00202-3.pdf) in which the then traditional footnote defining the R-factor at the bottom of Table 2 (on p. 1196) reads:

> [a] R factor $= \dfrac{\sum hkl||Fobs| - <|Fcalc|>|}{\sum hkl|Fobs|}$,
>
> where $<|Fcalc|>$ is the expectation of $|Fcalc|$ under the error model used in maximum-likelihood refinement.

Generally speaking, numerical values of Rxpct differed little from those of the conventional R-values (denoted Rconv in the sequel), so we continued using them, especially as Fxpct was used in the computation of one of the curves in the Reciprocal Space Correlation Coefficient (RecSCC) plot, an exclusive feature of BUSTER since its inception in 1994 (see material in https://www.globalphasing.com/buster/wiki/plugin/attachments/BRrecipCCplot/NewBusterCCplotmaterial.pdf).

The use of Fxpct was a tentative intermediate step towards a planned re-examination of the appropriateness of the conventional R-factor as an agreement factor between model and data in cases where there are known error models for both Fobs and Fcalc; unfortunately this re-examination was swept to one side when the new wave of BUSTER developments (improved optimiser, restraint dictionaries, LSSR, ligand detection and fitting …) took over. We have however made available, since the release of 31st October 2012, the "rvalue" tool that enables users to compare, whenever desired, the values of Rxpct (calculated internally and reported by BUSTER) with those of Rconv (computed from the output mtz file).

Perusal of this tool, as well as many earlier regular spot-checks, showed an essential equivalence between Rxpct and Rconv for refinements against the typical diffraction data collected by the rotation method on macro-crystals, with the cut-off criteria applied to them at the time. However, new experimental approaches have since appeared that produce much weaker data. Such is the case with many serial datasets collected at XFELs and synchrotrons, where average I/sig(I) values over the whole dataset may not exceed a few units. In such situations the Rxpct values become systematically much lower than those of Rconv, which is of course problematic. This trend is clearly illustrated by the following scatter plots, compiled from a collection of weekly re-refinement carried out in 2019-2020, encompassing a total of 20777 refinements for 20001 PDB identifiers:

Work	Free	All
		

showing that Rxpct becomes systematically lower than Rconv for weaker (blue) data.

We are aware that this behaviour of Rxpct is made worse by the "variance inflation" that occurs for weak noisy data when the estimated observational variance for amplitude Fobs is combined sub-optimally with that for the model-based complex structure factor, a matter that we will be revisiting in relation with its impact on refinement itself. This can however only be done at some stage in the future rather than right away.

We have therefore modified all internal calculations, reporting and graphing of R-values so as to compute and display both Rxpct and Rconv on as equal a footing as possible.

We have also made Rconv the default R-value used in the mmCIF files for deposition into the PDB and in the REMARK 3 section of the final PDB file.