THE PROTEIN SOCIETY

On the relationship between cumulative correlation coefficients and the quality of crystallographic data sets

Jimin Wang [1]*, Gary W. Brudvig,[1,2] Victor S. Batista,[2] and Peter B. Moore[1,2]

[1]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520-8114
[2]Department of Chemistry, Yale University, New Haven, Connecticut 06520-8107

Received 7 August 2017; Accepted 27 September 2017
DOI: 10.1002/pro.3314
Published online 29 September 2017 proteinscience.org

Abstract: In 2012, Karplus and Diederichs demonstrated that the Pearson correlation coefficient $CC_{1/2}$ is a far better indicator of the quality and resolution of crystallographic data sets than more traditional measures like merging R-factor or signal-to-noise ratio. More specifically, they proposed that $CC_{1/2}$ be computed for data sets in thin shells of increasing resolution so that the resolution dependence of that quantity can be examined. Recently, however, the $CC_{1/2}$ values of entire data sets, i.e., *cumulative correlation coefficients*, have been used as a measure of data quality. Here, we show that the difference in cumulative $CC_{1/2}$ value between a data set that has been accurately measured and a data set that has not is likely to be small. Furthermore, structures obtained by molecular replacement from poorly measured data sets are likely to suffer from extreme model bias.

Keywords: CC1/2; X-ray free-electron laser; femtosecond serial crystallography; photosystem II; PSII; model bias; cumulative correlation coefficients

Introduction

In his classic note on regression and the theory of evolution in 1896,[1] Karl Pearson introduced the statistic now known as the Pearson correlation coefficient (CC), and it has been widely used in the social and behavioral sciences ever since.[2] The Pearson CC of the data obtained when the same set of observations are made twice (x_j,y_j) is given by the following expression:

$$CC=\frac{\sum_j\left(x_j-\langle x\rangle\right)\left(y_j-\langle y\rangle\right)}{\sqrt{\sum_j\left(x_j-\langle x\rangle\right)^2\sum_j\left(y_j-\langle y\rangle\right)^2}} \quad (1)$$

The reflective CC, or CC_0, is obtained when the same expression is evaluated with the mean values replaced by zero [Eq. (2)].

$$CC_0=\frac{\sum_j\left(x_jy_j\right)}{\sqrt{\sum_j\left(x_j\right)^2\sum_j\left(y_j\right)^2}} \quad (2)$$

Both Pearson and reflective CCs can provide useful insights into the information content of experimental data sets and the reproducibility of measurements.

Although some crystallographic applications of reflective CC_0 were reported in the 1960s,[3,4] until recently, they were seldom used. Those early studies demonstrated that CC_0 of the amplitudes of crystallographic data sets are not good measures of overall data quality because the range of values possible for that statistic is so small. If two non-centrosymmetric data sets are being compared, the value of CC_0 will be 100% if the two are identical, but fall only to 78.5% ($=\pi/4$) if the two are completely unrelated (see Supporting Information Materials). These limits hold for the CC_0 values obtained within thin shells of constant resolution, and for cumulative (or

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: National Institutes of Health; Grant number: P01 GM022778 (JW); Grant sponsors: U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Division of Chemical Sciences, Geosciences, and Biosciences; Grant numbers: DESC0001423 (VSB), and DE-FG0205ER15646 (GWB).

*Correspondence to: Jimin Wang, Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520-8114. E-mail: jimin.wang@yale.edu

overall) CC_0 values. Another reason correlation coefficients were slow to gain a foothold is that data quality is seldom an issue in small molecule crystallography. The data are usually accurately measured, and the models derived from them often account well for them [e.g., Ref. 5].

In structural biology the situation can be quite different. The data are often harder to measure well, and the models inferred from them commonly fail to explain the data as accurately as they were measured.[6] As Karplus and Diederichs pointed out a few years ago, the Pearson CCs can play as useful a role in this arena,[7] as they do in electron microscopy (EM), where CC_0 (i.e., Fourier Shell Correlation) have been used for years to estimate the resolution of the EM maps.[4,8,9]

The $CC_{1/2}$ is particularly useful for detecting the presence of weak signals in the high-resolution shells of crystallographic data sets. The measurements used to estimate the intensities of each reflection are partitioned into two non-overlapping sets of equal size that are then used to obtain two independent sets of intensity estimates. $CC_{1/2}$ is the Pearson correlation coefficient obtained by comparing these two sets of intensities. As mentioned earlier, the calculations are usually done after the two sets of intensities have been divided into thin shells of increasing resolution so that the dependence of $CC_{1/2}$ on resolution can be determined. At low resolution, the $CC_{1/2}$ of crystallographic data sets is usually close to 100%, and it tends to fall off quite rapidly as the resolution limit of the data is reached.[7,10] Strong reflections in a data set, which are almost always better measured than weak reflections, tend to dominate correlation coefficients, and that is why $CC_{1/2}$ is a better tool for detecting signals in noisy data than data quality indexes that treat each Bragg reflection equally, such as the fractional intensity R(diff), which is the average fractional intensity difference between the two half-data sets. More important, $CC_{1/2}$ can identify very weak signals in noisy data because it is reliable and does not involve a scaling issue (and even when scaling is involved, it is scaling independent). In contrast, traditional R-factors such as R_{sym}, R_{merge}, and R_{meas} (see Ref. 11 for definitions) may fail to do so because their values are sensitive to the linear and Wilson B scale factors applied to the individual images that contributed to the two sets of intensities that are being compared, and they can be hard to determine accurately when the data are noisy.

Results and Discussion

The R(diff) and Pearson $CC_{1/2}$ values of crystallographic data sets are related to one another. If the distribution of intensities in an X-ray diffraction data set obeys Wilson statistics,[12] which it will if properly measured, and the noise in those data is Gaussian [OSM & Refs. 7,11], it can be shown that:

$$CC_{1/2} = \frac{1}{1 + aR_{\mathrm{diff}}^2}, \tag{3}$$

where a is a constant that depends only on the symmetry-related multiplicity, and that, typically, has a value between 1.0 and 2.0. This relationship is independent of both the average resolution of the reflections in each resolution shell and the thickness of those shells, which means that it is as applicable to the $CC_{1/2}$ values of entire data sets (i.e., cumulative $CC_{1/2}$'s) as it is to the $CC_{1/2}$ values of individual resolution shells. R(diff) in this equation is a lower-bound estimate because it assumes that the intensities being compared can be, and have already been, properly scaled.

The data released recently for four X-ray free-electron laser (XFEL) structures of photosystem II (PSII) [5WS5, 5WS6, 5WS0, and 5GTI][13] are complete enough so that one can see whether Eq. (3) applies to real data. As Figure 1 shows, both the $CC_{1/2}$ values for thin shells of resolution from these data sets and their cumulative $CC_{1/2}$ values conform to the upper bound curve predicted by Eq. (3) (i.e., the a=1.0 curve). Analyses done on other data sets (data not shown), as well as studies published by Karplus and Diederichs[11] further support the conclusion that Eq. (3) is generally valid.

When cumulative $CC_{1/2}$ values are compared to the dependence of $CC_{1/2}$ on resolution from the same data sets, it becomes obvious that cumulative $CC_{1/2}$ values are insensitive measures of data quality. The cumulative $CC_{1/2}$ values of the four PSII data sets

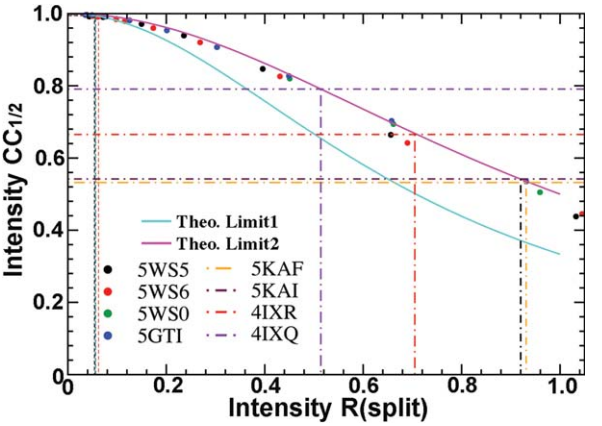

Figure 1. The relationship between intensity $CC_{1/2}$ and R(diff) values. The two theoretical limits for the relationship between these two quantities are shown using cyan and magenta solid lines. $CC_{1/2}$ and R(diff) values have been computed for four XFEL experimental data sets for PSII (5WS5, black spheres; 5WS6, red; 5WS0, green; 5GTI, blue) as a function of resolution. Those data follow the magenta curve, as do their cumulative $CC_{1/2}$ values (see the horizontal and vertical dotted lines having the same colors)

mentioned above range between 99.4% and 99.7%, even though $CC_{1/2}$ falls to $\sim$ 50% in the highest resolution shell in each of them. The cumulative $CC_{1/2}$ value of another, unrelated data set (4LR3),[10] which was obtained using conventional synchrotron methods, is only one percent smaller, 98.5%, even though the $CC_{1/2}$ value of that data set in its highest resolution shell is much worse, 13.5%.

Problems can arise when cumulative $CC_{1/2}$ values fall below $\sim$ 95%, as is the case for some of the other XFEL data sets reported for PSII.[14–17] For example, the 3.0-Å resolution data set for 5KAF has a cumulative $CC_{1/2}$ value of only 53.2%, and the cumulative $CC_{1/2}$ value for the 2.8-Å resolution data set that corresponds to 5KAI is only 54.2%.[17] These cumulative $CC_{1/2}$ values imply that the lower-bound value of R(diff) for these two data sets, *as a whole*, is likely to be 92% to 93% (Fig. 1)! In two of the other XFEL data sets (4IXR and 4IXQ) that have been reported for PSII the cumulative $CC_{1/2}$ values are 66.5% and 79.1%, respectively,[15] which suggests R(diff) values that are only slightly better, $\sim$ 71% and 51%. Thus by normal crystallographic standards, the quality of all of these data sets is very poor.

The four data sets just discussed are similar in quality to many of the other XFEL data sets that have been deposited in the PDB. Diederichs and colleagues have expressed concern about the quality of some of them, and about the wisdom of using cumulative $CC_{1/2}$ values as measure of data set quality.[7,18,19] It has been recommended that individual values at both low and high resolution bins should be reported, instead of just an overall value.[11] Nevertheless, cumulative $CC_{1/2}$ values are sometimes the only data quality statistics reported [e.g., Refs. 17,20]. It should be noted that the overall CC value is also being used as a measure of the correspondence between cryo-EM maps and the atomic models derived from them [e.g., Ref. 21]. This practice too is suspect.

How much structural information can there really be in the data sets that have quality statistics like those of the four data sets mentioned above? Rossmann has asked the same question about other XFEL data sets.[22,23] The reply sometimes given is that the quality of data cannot be as bad as it seems because the electron density (ED) maps obtained from them by molecular replacement "look good," and the model R-factors appear acceptable. We have explored this issue computationally in two different ways. First, starting with the 5KAI model for PSII,[17] a 2.80-Å resolution ED map was calculated using model phases, and a set of amplitudes obtained by Fourier transformation of a molecular model that was generated from 5KAI by changing the positions of all protein atoms at random so that the distribution of their displacements is a Gaussian function with $\sigma_{|\Delta \mathbf{r}|} = 1.038$ Å (see OSM). This shifted-coordinate model is unphysical, and its transform differs from that of its parent structure, on average, by 42.5% in amplitude and by 67.6% in intensity, which is a lot. Nevertheless, the cumulative intensity Pearson CC between these two data sets is 75%, consistent with the arguments made above about the insensitivity of that statistic to data quality. Furthermore, the ED map computed using these coordinate-shifted amplitudes and model phases is strikingly similar to the one obtained using the measured amplitudes and model phases [Fig. 2(A,B), top panels]. In fact, the fit of the original protein model to the ED map calculated using both the amplitudes and phases obtained from the coordinate-shifted model is reasonably good, and not obviously worse than its fit to the ED map computed using coordinate-shifted amplitudes but the original model phases (data not shown).

In the second test, the 5KAI experimental data were divided into 100 thin shells of resolution, each containing the same number of reflections. A second data set was obtained from the first by assigning measured amplitudes to Bragg indices at random within each shell (see OSM). This procedure is called random permutation, and it has used in the past to investigate the statistical properties of small molecule crystallographic data.[24] The average Pearson CC between the parent data and the randomized data in each shell was zero for all intents and purposes: $\sim -0.001 \pm 0.021$. Furthermore, on average, amplitudes had changed by 55.7% and intensities by 101.9%, which is close to what would be seen if diffraction data sets were compared that had been obtained from two unrelated, non-centrosymmetric crystals that happened to have the same symmetry and unit cell dimensions (58.6% and 100%).[25,26] Nevertheless, the cumulative Pearson CC between the permuted data set and its parent was 22%, rather than the 0% one might have anticipated (see below). Figure 2(A,C) (top panels) show the same small region of the ED map (ρ_{obs}) one computed using the original amplitudes and model phases [Fig. 2(A), Supporting Information Fig. S1A], and the other computed using these randomized amplitudes and the original model phases ($\rho_{chimeric}$) [Fig. 2(C), Supporting Information Fig. S1B]. Once again, the similarity of the two is striking. This observation should not come as a surprise because similar results were reported for computations carried out using small molecule data over half a century ago.[27–29]

What these computations illustrate is the impact that model bias can have on the outcomes of a structure determination that depend on molecular replacement, a problem that has long been recognized, but may still be underappreciated. It should also serve as a warning that the claim that an ED map "looks good" is no guarantee that the data on which it is based are meaningful.

The similarity between ρ_{obs} and $\rho_{chimeric}$ maps is easy to understand.

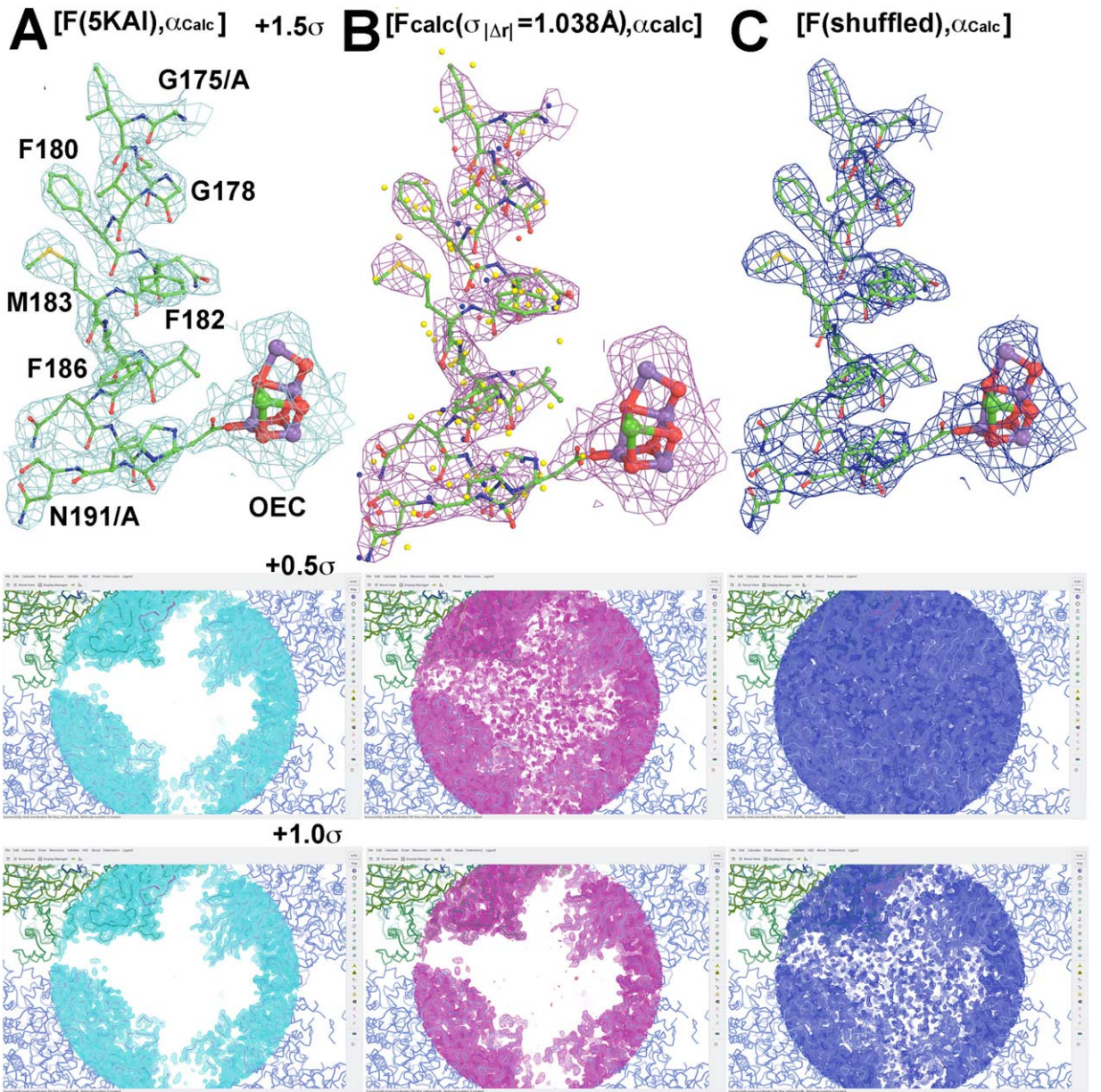

Figure 2. Electron density (ED) maps. (A) A portion of the ED map computed for 5KAI using observed amplitudes and model phases with the 5KAI model superimposed. (B) The ED map for the same region as (A) computed by combining amplitudes obtained from a model that was derived from 5KAI by altering the locations of all protein atoms at random by the standard deviation of 1.038 Å with (unaltered) 5KAI phases. The position of the atoms in the partially randomized model used to compute the amplitudes used are shown as colored spheres. (C) The ED obtained for the same region shown in (A) using random permutated amplitudes and phases from the original model. The 5KAI model is again superimposed. The first row shows map features contoured at +1.5σ for an α-helix next to the oxygen-evolving complex (OEC) of photosystem II. The lower two rows contoured at +0.5σ for and +1.0σ show map features in solvent channels

$$\rho_{\mathrm{obs}}(\boldsymbol{r}) = FT(\ F_{\mathrm{obs}}, \alpha_{\mathrm{obs}}); \tag{4}$$

$$\rho_{\mathrm{chimeric}}(\boldsymbol{r}) = FT(\ F_{\mathrm{alt}}, \alpha_{\mathrm{obs}})$$

$$= FT(\ F_{\mathrm{obs}}, \alpha_{\mathrm{obs}}) + FT\left[(F_{\mathrm{alt}} - F_{\mathrm{obs}}), \alpha_{\mathrm{obs}}\right]$$

$$= \rho_{\mathrm{obs}}(\boldsymbol{r}) + FT[(\Delta F), \alpha_{\mathrm{obs}}]; \tag{5}$$

where FT indicates Fourier transformation, experimentally observed structure factors are F_{obs} and α_{obs}, F_{alt} are the randomly altered amplitudes, and ΔF are differences between the experimental and altered amplitudes. The second term in the last line of Eq. (5) is a difference Fourier map. It is well known, of course, that if the amplitudes identified here as F_{alt} derive from crystals that are isomorphous to those used to produce the F_{obs} data, and the structures of the molecules in the two crystals are closely related, but not identical, the $FT[\Delta F,\alpha_{\mathrm{obs}}]$ map obtained will reveal the specific differences between the two

structures.[30–32] However, for the computations just described, the amplitude differences are completely uncorrelated with the structure that gave rise to the model phases. Thus, this component of the chimeric maps should consist of random features that extend throughout the unit cell, including its solvent channels, and bear no relationship to $\rho_{obs}(\mathbf{r})$ (Fig. 2). That this is in fact the case is evident in the lower panels in Figure 2 that are centered on the solvent region in the crystal of interest, which show a much larger portion of the unit cell than the top panels. The solvent region is almost featureless in the map computed using model phases and the observed amplitudes [Fig. 2(A), lower panels]. It is significantly noisier in the map computed using model phases, but amplitudes derived from a randomly distorted model [Fig. 2(B), lower panels], and it is as feature-filled as the parts of the unit cell that contain protein in the map obtained with random permutated intensities and model phases [Fig. 2(C), lower panels]. There are other indications that all is not well with the two randomized data sets. First, the R factors that measure the correspondence between the observed amplitudes and the computed data are poor: 42.5% and 55.7%, and the density histograms within the protein-containing regions of the two maps are obviously abnormal.

It is straightforward to convert intensity R(diff) estimates into estimates of intensity and amplitude R(sigma) values (OSM),[4] and when this is done one finds that the intensity R(sigma) values for the 5KAF and 5KAI data sets[17] should be 65.1% to 65.8%, and their amplitude R(sigma) values should be 37.6% to 38.0%. In light of what has just been said, it is quite surprising that the amplitude free R-factors reported for the models derived from these two data sets are significantly lower than these lower bound estimates of amplitude R(sigma) values of the data from which they derive: 30.30% (5KAF) and 29.97% (5KAI).[17] How can a model derived from a set of data explain those data more accurately than the data explain themselves? These observations suggest that both models are over-fitted, and one has to ask why the cross validation procedures used, which are supposed to prevent over-fitting,[33] failed to do so in both cases.

Examples of poor quality data sets can also be found in the electron crystallographic literature. For example, the overall data merging R-factor[34] reported for a data set obtained from proteinase K crystals at a resolution between 21.91 and 1.30 Å was 62.9%. Yet, the authors were able to produce what appears to be an outstanding electrostatic potential (ESP) map of that molecule by molecular replacement, even though the model they used for molecular replacement took no account of the impact that partial charges have on electron scattering factors, which are known to be large.[35–38] In addition, a recent analysis done by Spence and colleagues has demonstrated that a number of electron crystallographic structures are computational artifacts because the influence of model bias had on the ESP density functions on which they are based.[39]

In the past, structural biologists were more concerned about the quality of their diffraction data than the quantity, and often adopted a very conservative approach[40] to this problem by, for example, rejecting all the data they had measured beyond the resolution at which the average signal-to-noise ratio dropped to 2.0. The introduction of the $CC_{1/2}$ method for evaluating resolution limits has encouraged structural biologists to use high resolution data they would otherwise have ignored, and by so doing, they have been able to arrive at better models.[7,10,11,41]

Nevertheless, the fact that it is a good idea to use high-resolution data that have a $CC_{1/2}$ value of, say, 50%, does not mean that it is a good idea to use data sets that have a cumulative $CC_{1/2}$ value of 50%. The cumulative Pearson $CC_{1/2}$ statistic is not very sensitive to the overall quality, as we have demonstrated above, and while we have not found a rigorous way to estimate a lower bound value for this parameter, that bound is clearly greater than zero, and it is easy to understand why. If the average intensities of the reflections in two data sets have the same dependence on resolution, then, even if the two data sets are completely unrelated, their cumulative Pearson CC will be greater than zero because (i) strong reflections predominate at low resolution and weak reflections predominate at high resolution, and (ii) the two uncorrelated data sets are likely to have similar mean intensity values within each resolution shell. It appears from the results described above that the cumulative $CC_{1/2}$ of a data set that has been well measured by traditional reproducibility criteria, such as R(diff), R(split), R(sigma), R(merging), R(meas), or R(pim) (see Ref. 11, for definitions), will certainly exceed 90%.

Over the last few decades, molecular replacement has become the preferred method for solving crystal structures in structural biology. It is usually the case that the molecular model used is not a fully accurate representation of the molecules in the crystals of interest, and the hope is that the ED map that ultimately emerges will not be identical to that anticipated for the starting model, and that new insights will emerge when those differences are interpreted. This expectation is unlikely to be satisfied if the quality of the data set under consideration is low because of the overwhelming impact that model-phase bias will have on outcomes, as discussed above. Indeed, the recent literature provides an example of a failure of just this kind that resulted when an XFEL data set having a resolution of 1.35 Å, and a cumulative $CC_{1/2}$ of 81.3% was used to compute a difference map [Figure 6A in Ref. 20].

In conclusion, the $CC_{1/2}$ method Karplus and Diederichs devised for assessing data quality was an

important advance in macromolecular crystallography.[11] By contrast, although related, the cumulative $CC_{1/2}$, is a statistic best avoided because it is ill behaved, and less informative than more traditional statistics like R(merge).[11] When it is used, it must be supplemented with other relevant statistics.

Acknowledgment

Authors thank Drs. P.A. Karplus, M. Rossmann, J. Spence, and B.W. Matthews for critical comments on this manuscript.

Conflict of Interest Statement

The authors declare no conflict of interest in publishing results of this study.

References

1. Pearson K (1896) Mathematical contributions to the theory of evolution. III. Regression, Heredity and Panmixia. Philos Trans of Royal Society London 187: 253–318.
2. Evans JD. (1996). Straightforward statistics for the behavior sciences. Pacific Grove: Brooks/Cole Publishing Co.
3. Srinivasan R, Chandrasekaran R (1966) Correlation functions connected with structure factors and their application to observed and calculated structure factors. Indian J Pure Appl Phys 4:178–182.
4. Srinivasan R, Parthasarathy S. (1976). Some statistical applications in X-ray crystallography. Oxford, New York: Pergamon Press.
5. Koritsanszky T, Flaig R, Zobel D, Krane H, Morgenroth W, Luger P (1998) Accurate experimental electronic properties of dl-proline monohydrate obtained within 1 Day. Science 279:356–358.
6. Holton JM, Classen S, Frankel KA, Tainer JA (2014) The R-factor gap in macromolecular crystallography: an untapped potential for insights on accurate structures. FEBS J 281:4046–4060.
7. Karplus PA, Diederichs K (2012) Linking crystallographic model and data quality. Science 336:1030–1033.
8. Rosenthal PB, Henderson R (2003) Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. J Mol Biol 333:721–745.
9. Vanheel M (1987) Similarity measures between Images. Ultramicroscopy 21:95–99.
10. Wang J, Wing RA (2014) Diamonds in the rough: a strong case for the inclusion of weak-intensity X-ray diffraction data. Acta Cryst D70:1491–1497.
11. Karplus PA, Diederichs K (2015) Assessing and maximizing data quality in macromolecular crystallography. Curr Opin Struct Biol 34:60–68.
12. Wilson AJC (1949) The probability distribution of X-ray intensities. Acta Cryst 2:318–321.
13. Suga M, Akita F, Sugahara M, Kubo M, Nakajima Y, Nakane T, Yamashita K, Umena Y, Nakabayashi M, Yamane T, Nakano T, Suzuki M, Masuda T, Inoue S, Kimura T, Nomura T, Yonekura S, Yu LJ, Sakamoto T, Motomura T, Chen JH, Kato Y, Noguchi T, Tono K, Joti Y, Kameshima T, Hatsui T, Nango E, Tanaka R, Naitow H, Matsuura Y, Yamashita A, Yamamoto M, Nureki O, Yabashi M, Ishikawa T, Iwata S, Shen JR (2017) Light-induced structural changes and the site of O=O bond formation in PSII caught by XFEL. Nature 543:131–135.
14. Kern J, Alonso-Mori R, Hellmich J, Tran R, Hattne J, Laksmono H, Glockner C, Echols N, Sierra RG, Sellberg J, Lassalle-Kaiser B, Gildea RJ, Glatzel P, Grosse-Kunstleve RW, Latimer MJ, McQueen TA, DiFiore D, Fry AR, Messerschmidt M, Miahnahri A, Schafer DW, Seibert MM, Sokaras D, Weng TC, Zwart PH, White WE, Adams PD, Bogan MJ, Boutet S, Williams GJ, Messinger J, Sauter NK, Zouni A, Bergmann U, Yano J, Yachandra VK (2012) Room temperature femtosecond X-ray diffraction of photosystem II microcrystals. Proc Natl Acad Sci USA 109:9721–9726.
15. Kern J, Alonso-Mori R, Tran R, Hattne J, Gildea RJ, Echols N, Glockner C, Hellmich J, Laksmono H, Sierra RG, Lassalle-Kaiser B, Koroidov S, Lampe A, Han G, Gul S, Difiore D, Milathianaki D, Fry AR, Miahnahri A, Schafer DW, Messerschmidt M, Seibert MM, Koglin JE, Sokaras D, Weng TC, Sellberg J, Latimer MJ, Grosse-Kunstleve RW, Zwart PH, White WE, Glatzel P, Adams PD, Bogan MJ, Williams GJ, Boutet S, Messinger J, Zouni A, Sauter NK, Yachandra VK, Bergmann U, Yano J (2013) Simultaneous femtosecond X-ray spectroscopy and diffraction of photosystem II at room temperature. Science 340:491–495.
16. Kern J, Tran R, Alonso-Mori R, Koroidov S, Echols N, Hattne J, Ibrahim M, Gul S, Laksmono H, Sierra RG, Gildea RJ, Han G, Hellmich J, Lassalle-Kaiser B, Chatterjee R, Brewster AS, Stan CA, Glockner C, Lampe A, DiFiore D, Milathianaki D, Fry AR, Seibert MM, Koglin JE, Gallo E, Uhlig J, Sokaras D, Weng TC, Zwart PH, Skinner DE, Bogan MJ, Messerschmidt M, Glatzel P, Williams GJ, Boutet S, Adams PD, Zouni A, Messinger J, Sauter NK, Bergmann U, Yano J, Yachandra VK (2014) Taking snapshots of photosynthetic water oxidation using femtosecond X-ray diffraction and spectroscopy. Nat Commun 5:4371.
17. Young ID, Ibrahim M, Chatterjee R, Gul S, Fuller FD, Koroidov S, Brewster AS, Tran R, Alonso-Mori R, Kroll T, Michels-Clark T, Laksmono H, Sierra RG, Stan CA, Hussein R, Zhang M, Douthit L, Kubin M, de Lichtenberg C, Vo Pham L, Nilsson H, Cheah MH, Shevela D, Saracini C, Bean MA, Seuffert I, Sokaras D, Weng TC, Pastor E, Weninger C, Fransson T, Lassalle L, Brauer P, Aller P, Docker PT, Andi B, Orville AM, Glownia JM, Nelson S, Sikorski M, Zhu D, Hunter MS, Lane TJ, Aquila A, Koglin JE, Robinson J, Liang M, Boutet S, Lyubimov AY, Uervirojnangkoorn M, Moriarty NW, Liebschner D, Afonine PV, Waterman DG, Evans G, Wernet P, Dobbek H, Weis WI, Brunger AT, Zwart PH, Adams PD, Zouni A, Messinger J, Bergmann U, Sauter NK, Kern J, Yachandra VK, Yano J (2016) Structure of photosystem II and substrate binding at room temperature. Nature 450:453–457.
18. Assmann G, Brehm W, Diederichs K (2016) Identification of rogue datasets in serial crystallography. J Appl Crystallogr 49:1021–1028.
19. Diederichs K (2017) Dissecting random and systematic differences between noisy composite data sets. Acta Cryst D73:286–293.
20. Uervirojnangkoorn M, Zeldin OB, Lyubimov AY, Hattne J, Brewster AS, Sauter NK, Brunger AT, Weis WI (2015) Enabling X-ray free electron laser crystallography for challenging biological systems from a limited number of crystals. Elife 4:05421–05429.
21. Hryc CF, Chen DH, Afonine PV, Jakana J, Wang Z, Haase-Pettingell C, Jiang W, Adams PD, King JA, Schmid MF, Chiu W (2017) Accurate model annotation

of a near-atomic resolution cryo-EM map. Proc Natl Acad Sci U S A 114:3103–3108.
22. Gati C, Bourenkov G, Klinge M, Rehders D, Stellato F, Oberthur D, Yefanov O, Sommer BP, Mogk S, Duszenko M, Betzel C, Schneider TR, Chapman HN, Redecke L (2014) Serial crystallography on in vivo grown microcrystals using synchrotron radiation. IUCrJ 1:87–94.
23. Rossmann MG (2014) Serial crystallography using synchrotron radiation. IUCrJ 1:84–86.
24. Srinivasan R, Srikrishnan T (1966) New statistical tests for distinguishing between centrosymmetric and noncentrosymmetric structures. Acta Cryst 21:648–652.
25. Wilson AJC (1950) Largest likely values for the reliability index. Acta Cryst 3:397–398.
26. - (1974) Largest likely values of residuals. Acta Cryst A30:836–838.
27. Ramachandran G, Srinivasan R (1961) An apparent paradox in crystal structure analysis. Nature 190:159–161.
28. Srinivasan R (1961) The significance of the phase synthesis. Proc Indian Acad Sci 53A:252–264.
29. Lattman E, DeRosier D (2008) Why phase errors affect the electron function more than amplitude errors. Acta Cryst A64:341–344.
30. Stryer L, Kendrew JC, Watson HC (1964) The Mode of Attachment of the Azide Ion to Sperm Whale Metmyoglobin. J Mol Biol 8:96–104.
31. Kraut J (1965) Structural Studies with X-Rays. Annu Rev Biochem 34:247–268.
32. Wang J, Mauro M, Edwards SL, Oatley SJ, Fishel LA, Ashford VA, Xuong NH, Kraut J (1990) X-ray structures of recombinant yeast cytochrome c peroxidase and three heme-cleft mutants prepared by site-directed mutagenesis. Biochemistry 29:7160–7173.
33. Brunger AT (1992) Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. Nature 355:472–475.
34. Hattne J, Shi D, de la Cruz MJ, Reyes FE, Gonen T (2016) Modeling truncated pixel values of faint reflections in MicroED images. J Appl Crystallogr 49:1029–1034.
35. Wang J, Moore PB (2017) On the interpretation of electron microscopic maps of biological macromolecules. Protein Sci 26:122–129.
36. Wang J (2017) On the appearance of carboxylates in electrostatic potential maps. Protein Sci 26:396–402.
37. Wang J (2017) Experimental charge density from electron microscopic maps. Protein Sci 26:1619–1629
38. Wang J, Videla PE, Batista VS (2017) Effects of aligned alpha-helix peptide dipoles on experimental electrostatic potentials. Protein Sci 26:1692–1697.
39. Subramanian G, Basu S, Liu H, Zuo JM, Spence JC (2015) Solving protein nanocrystals by cryo-EM diffraction: multiple scattering artifacts. Ultramicroscopy 148:87–93.
40. Wang J (2015) Estimation of the quality of refined protein crystal structures. Protein Sci 24:661–669.
41. Diederichs K, Karplus PA (2013) Better models by discarding data? Acta Cryst D69:1215–1222.