STRUCTURAL BIOLOGY
ISSN 2059-7983

Multivariate estimation of substructure amplitudes for a single-wavelength anomalous diffraction experiment

Navraj S. Pannu and Pavol Skubák*

Department of Infectious Diseases, Leiden University Medical Center, Albinusdreef 2, 2333 ZA, Leiden, The Netherlands.
*Correspondence e-mail: skubakp@gmail.com

Received 4 November 2022
Accepted 2 March 2023

Edited by A. Gonzalez, Lund University, Sweden

Keywords: substructure determination; experimental phasing; multivariate statistics; direct methods; single-wavelength anomalous diffraction; *Afro*.

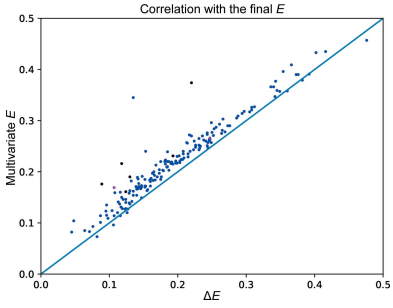
To determine a substructure from single-wavelength anomalous diffraction (SAD) data using Patterson or direct methods, the substructure-factor amplitude ($|F_a|$) is first estimated. Currently, the absolute value of the Bijvoet difference is widely used as an estimate of $|F_a|$ values for SAD data. Here, an equation is derived from multivariate statistics and tested that takes into account the correlation between the observed positive (F^+) and negative (F^-) Friedel pairs and F_a along with measurement errors in the observed data. The multivariate estimation of $|F_a|$ has been implemented in a new program, *Afro*. Results on over 180 test cases show that *Afro* provides a higher correlation to the final substructure-factor amplitudes (calculated from the refined, final substructures) than the Bijvoet differences and improves the robustness of direct-methods substructure detection.

1. Introduction

In determining a macromolecular crystal structure solely from its anomalous signal, the first step is to determine the position of the anomalous substructure that is present. The application of direct methods combined with Patterson techniques, as implemented, for example, in the programs *SHELXD* (Schneider & Sheldrick, 2002) and *HySS* (Grosse-Kunstleve & Adams, 2003), or the application of phase-retrieval techniques as implemented in *PRASA* (Skubák, 2018) have proven to be very powerful in detecting anomalous substructures, particularly when the anomalous substructure contains many atoms or the signal is very weak.

In all of these approaches, in order to detect the anomalous substructure an estimate of the substructure-factor amplitude $|F_a|$ is required. The absolute value of the Bijvoet difference ($\Delta F = ||F^+| - |F^-||$) is typically input to substructure-detection programs as an estimate for $|F_a|$.

To improve the methods further, here we propose new formulas and a new refinement strategy to calculate $|F_a|$ values. Previously, Terwilliger (1994) and Burla *et al.* (2002, 2003) employed Bayesian and multivariate approaches to obtain the probability distribution of $|F_a|$. Here, we expand on their work and derive a probability distribution for $P(|F_a|; |F^+|, |F^-|)$ that takes into account measurement errors in $|F^+|$ and $|F^-|$ and does not assume any relationship between the Friedel phases. We report that at least in our practical implementation, better results were obtained by using the approximation of Burla and coworkers, probably due to numerical stability issues of the more general equation. Furthermore, we propose the



maximum-likelihood refinement of errors and scale parameters to obtain the optimal values, given the distributions that we have obtained. Finally, we apply the newly implemented $|F_a|$ estimation to over 180 test cases and show the superior performance of these estimates compared with the ΔF values when used by the substructure-determination program *PRASA*.

2. Methods

To obtain an estimate of the substructure-factor amplitude $|F_a|$ from a SAD experiment, the expected value of $|F_a|$ given the observations $|F^+|$ and $|F^-|$ is required. Let F^+ denote a structure factor with Miller indices h, k, l; $(F^-)^*$ denote the complex conjugate of a structure factor with Miller indices $-h$, $-k$, $-l$; F_a denote a substructure factor with Miller indices h, k, l; and α^+ and α^- denote the phases of F^+ and $(F^-)^*$, which we will refer to as Friedel pair phases. Then, assuming a complex multivariate Gaussian distribution for $P[F_a, F^+, (F^-)^*]$, the following expression can be obtained:

$$\begin{aligned}
&\langle |F_a|; |F^+|, |F^-| \rangle \\
&= \frac{\int_0^\infty |F_a| \int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} P(|F_a|, \alpha_a, |F^+|, \alpha^+, |F^-|, \alpha^-)\, d\alpha^+\, d\alpha^-\, d\alpha_a\, dF_a}{\int_0^\infty \int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} P(|F_a|, \alpha_a, |F^+|, \alpha^+, |F^-|, \alpha^-)\, d\alpha^+\, d\alpha^-\, d\alpha_a\, dF_a} \\
&= \frac{1}{4(\pi a_{11})^{1/2} I_0\{2|F^+||F^-|[(a_{23} - \frac{a_{12}a_{13}+b_{12}b_{13}}{a_{11}})^2 + (b_{23} + \frac{a_{13}b_{12}-a_{12}b_{13}}{a_{11}})^2]^{1/2}\}} \\
&\times \int_{-\pi}^{\pi} \exp\left\{ -2|F^+||F^-| \left[\left(a_{23} - \frac{a_{12}a_{13} + b_{12}b_{13}}{a_{11}} \right) \cos(\beta) \right.\right. \\
&\left.\left. - \left(b_{23} + \frac{a_{13}b_{12} - a_{12}b_{13}}{a_{11}} \right) \sin(\beta) \right] \right\} \\
&\times \Phi(-\tfrac{1}{2}, 1, -\xi)\, d\beta, \qquad (1)
\end{aligned}$$

where

$$\begin{aligned}
\xi(|F^+|, |F^-|, \beta, \Sigma) = &\frac{(a_{12}^2 + b_{12}^2)|F^+|^2 (a_{13}^2 + b_{13}^2)|F^-|^2}{a_{11}} \\
&+ \frac{2|F^+||F^-|(a_{12}a_{13} + b_{12}b_{13})\cos(\beta)}{a_{11}} \\
&+ \frac{2|F^+||F^-|(a_{13}b_{12} - a_{12}b_{13})\sin(\beta)}{a_{11}}. \qquad (2)
\end{aligned}$$

The above expression is derived in Appendix *A*; it does not assume $\alpha^+ = \alpha^-$ as was required in earlier publications (Burla *et al.*, 2002, 2003), it incorporates the effect of measurement errors in the observed Friedel pair amplitudes and it can be calculated by a single numerical integration. In the above expression, Σ is the (Hermitian) covariance matrix of the complex Gaussian distribution $P[F_a, F^+, (F^-)^*]$, with the elements of its inverse denoted $z_{jk} = a_{jk} + ib_{jk}$, $\beta = \alpha^+ - \alpha^-$, $\Phi(x, y, z)$ is the Kummer confluent hypergeometric function and I_0 is the modified Bessel function of the first kind and of zero order. The covariance matrix Σ was calculated using the expressions derived previously (Pannu *et al.*, 2003) and the correlation between structure factors. To ensure that the matrix remains positive definite, the inverse of the covariance matrix was calculated from the eigenvalues and eigenvectors calculated from *LAPACK* routines (Anderson *et al.*, 1999) to remove negative eigenvalues.

We have implemented two equations based on equation (1) in a new program *Afro* for the multivariate estimation of $|F_a|$ values. One equation is equation (1) itself, while the other is a simplified form of equation (1) using the Friedel pair phase equality assumption as suggested by Burla *et al.* (2002, 2003):

$$\langle |F_a|; |F^+|, |F^-| \rangle = \frac{1}{2}\left(\frac{\pi}{a_{11}}\right)^{1/2} \Phi\left(-\frac{1}{2}, 1, -\frac{\xi}{a_{11}}\right). \qquad (3)$$

We have found that the simpler equation (3), *i.e.* assuming that the Friedel pair phases are equal, led to better performance in the test cases shown below, which is likely to be due to improved numerical stability. Thus, results from the implementation of this equation are shown below.

The covariance matrix Σ depends on both the number and the (overall) temperature factor of the substructure atoms. As these parameters are usually unknown, a likelihood estimate is obtained by *Afro*. Thus, after initial estimates of the number and the overall temperature factor of the substructure atoms have been input, the parameters are refined using the marginal distribution $P(|F^+|, |F^-|)$. The refinement of these parameters turned out to have a large radius of convergence, and better results were obtained when refined values were used compared with when unrefined values. We have previously discussed the procedure (Pannu, 2007) and a similar approach was recently reported by Hatti *et al.* (2021). After the refinement, the $|F_a|$ values are estimated using equation (1). Local scaling (Blessing, 1997) has been also implemented in *Afro* which scales $|F^+|$ to $|F^-|$ in local spheres.

The multivariate $|F_a|$ calculation using the Friedel pair phase equality assumption as implemented in *Afro* was tested on a sample of 182 SAD data sets as specified in Appendix *B* containing a large number of anomalous scatterers (selenium, sulfur, iodine, zinc, gold, copper, platinum, krypton, manganese, iron, cadmium, nickel, calcium and mercury) and a large range of data resolutions from 0.94 to 3.9 Å. For each data set, a complete *Crank*2 (Skubák & Pannu, 2013) structure-solution run was performed, with *Afro* being used for the calculation of $|F_a|$ and E (normalized $|F_a|$), *PRASA* being used for substructure determination and *REFMAC*5 (Nicholls *et al.*, 2018), *Parrot* (Cowtan, 2010), *Buccaneer* (Cowtan, 2008) and *SHELXE* (Usón & Sheldrick, 2018) being used in the subsequent combined phasing, density modification and model building. Versions of the programs corresponding to *CCP*4 (Winn *et al.*, 2011) version 8.0.002 were used, except for *Crank*2, where the more recent version 2.0.325 was used, and a bug fix in *REFMAC*5 implemented by us to prevent the program from crashing for very large data sets.

The input to *Crank*2 consisted of the SAD data set, the protein sequence and a specification of the anomalously scattering atom type with anomalous scattering coefficients. For five data sets, a value of the solvent content corresponding to the correct number of monomers in the asymmetric unit was specified, otherwise the default options were used. An incorrect solvent-content estimate would not affect the $|F_a|$

estimation as it is not used in it; however, since it is an important phase-improvement parameter, it would lead to 'randomly' incomplete models for data sets that could otherwise be automatically built, thus making the model-building analysis less relevant.

For each data set we calculated the overall correlation of the estimated E values with the 'final' substructure E values in the following way. The final anomalously scattering substructure (either deposited or, if not available, determined from the anomalous difference maps) was input to *REFMAC*5 using 0 refinement cycles. The calculated amplitudes from *REFMAC*5 were then input to *ECALC* from *CCP*4 (Ian Tickle, unpublished work), providing the final substructure E values. The correlation between the estimated and final E values was calculated using the *SFTOOLS* utility from *CCP*4 (Bart Hazes, unpublished work), which divided the data-set reflections into 20 resolution bins and calculated the correlations per resolution bin. Finally, an average of the bin correlations up to 'anomalous resolution' was calculated. The anomalous resolution was determined once for each data set, corresponding to the lowest resolution (the largest number) included in those resolution bins in which the correlation between the multivariate E values and the final E values was smaller than 0.05 and an average of correlations from three consecutive resolution bins was smaller than 0.05.

Estimation of E values from Friedel pair differences (ΔE) was also implemented in *Afro* and was tested on the 182 SAD data sets to compare its performance against the multivariate estimation. Complete structure solution from ΔE was attempted with *Crank*2 using the same pipeline and default options as used in the runs from multivariate *Afro*.

The anomalous substructure obtained by *PRASA* is considered to be 'correctly determined' if at least one third of the atoms in the final anomalous substructure had a matching atom (within 2 Å distance) in the substructure obtained after transformation by *SITCOM* (Dall'Antonia & Schneider, 2006). Similarly to as in Skubák (2018), we have observed that typically if approximately 1/3 of the substructure atoms have been correctly identified in substructure determination, the remaining significant anomalous scatterers can either be added by *Crank*2 from the anomalous maps or their absence does not affect the success of model building.

The model-building performance is judged by the fraction of the PDB-deposited model backbone that is 'correctly built'. A residue is considered to be correctly built if its C^{α} position is at a distance of at most 2 Å from a deposited model C^{α} ('C^{α}-deposited') position and a neighbouring C^{α} position is at a distance of at most 2 Å from a neighbour of the C^{α}-deposited position (sequence identity or directionality is not checked). A custom script evaluating the model-building performance using these criteria was used.

For all data sets where one of the pipelines failed to determine the substructure, a 'thorough' substructure-determination protocol was tested: the number of *PRASA* trials was increased to 100 000 trials from the default maximum of 2000 trials, more high-resolution cutoffs were tested (the high-resolution cutoff step was decreased to 0.1 from the default of 0.25) and the initial high-resolution cutoff was set to be identical to the anomalous resolution. The thorough protocol aims to estimate whether it is possible to determine the substructure by *PRASA* from the input E values at all.

Table 1
Number of data sets for which the substructure was determined and the majority of the model was built by the two tested pipelines: starting from E calculated as Friedel pair differences and by multivariate *Afro*.

The first number in each cell denotes the number of successes using the default substructure-determination protocol and the second number that using the 'thorough' substructure-determination protocol with a substantially larger number of trials and a larger number of high-resolution cutoffs.

	No. of data sets (default/thorough)	
	Delta	Multivariate
Substructures determined	162/163	168/170
Models built	156/157	161/162

3. Results and discussion

The correlation of the multivariate E values estimated by *Afro* with the final substructure E is typically significantly larger than that for ΔE, as demonstrated by Fig. 1. In tests on the 182 SAD data sets, the average correlation improved by 13% (from 0.197 to 0.223) and an improved correlation was observed for 94% of the data sets.

The overall better quality of the E estimates calculated by *Afro* allowed successful substructure determination by *PRASA* for six data sets that did not work using ΔE. As summarized in Table 1, the total number of data sets with the substructure correctly determined increased from 162 (89.0%) using ΔE to 168 (92.3%) using multivariate *Afro*. If these six

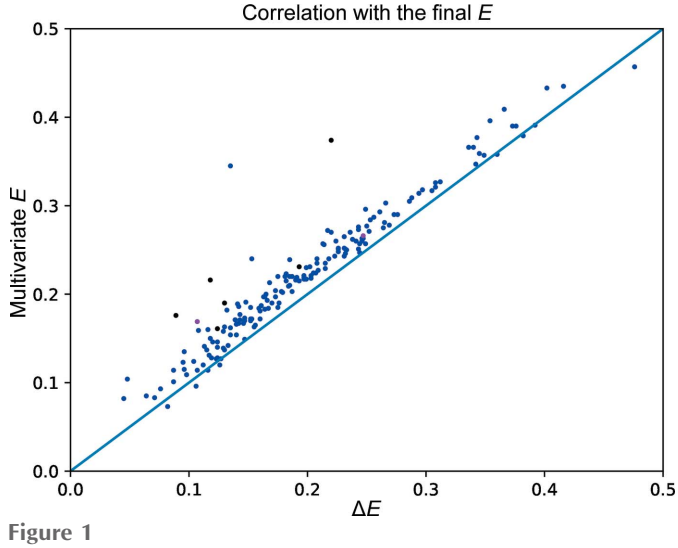


Figure 1
The correlation of ΔE (x axis) and multivariate E from *Afro* (y axis) with the 'final' substructure E for each of the 182 tested data sets. The data sets for which the substructure was correctly determined from the multivariate E but not from ΔE are displayed in black (comparing the results of the default substructure-determination protocol) and magenta (the 'thorough' substructure-determination protocol).

data sets were removed from the comparison, the average fraction of the substructure that was correctly determined remained similar (0.774 versus 0.760). This indicates that the improvement in the quality of the multivariate E values from *Afro* may not be of great practical importance if the substructure can be obtained using the ΔE values; however, it may allow successful substructure determination for data sets where the substructure could not be determined using the ΔE values.

A majority of the model was correctly built for 156 data sets (85.7%) starting from substructure determination using ΔE and for 161 data sets (88.5%) starting from substructures determined by multivariate E from *Afro*.

Using the 'thorough' substructure-determination protocol with a large number of substructure trials and resolution cutoffs for the data sets where substructure determination failed led to the determination of another two substructures starting from the multivariate *Afro*. Similarly, one more substructure could be determined using the thorough protocol starting from ΔE; this substructure was obtained starting from the multivariate E using the default protocol.

In total (default + thorough protocol), seven substructures were determined from the multivariate E values that were not determined from the ΔE values. Furthermore, determination of one other substructure required the thorough protocol starting from ΔE, while the default protocol was sufficient if multivariate *Afro* was used. Analysis of the success rates for this data set (PDB entry 2pgc) shows that this was not a coincidence: only four solutions were obtained in 100 000 trials from ΔE (a success rate of 1 in 25 000) and 27 solutions were obtained using the multivariate *Afro* (1 in 3704).

The data sets used in this paper may not be fully representative of user data. In particular, a large fraction (almost 45%) of the data sets come from the automated JCSG pipeline (Elsliger *et al.*, 2010), which may differ from more recent data-collection methods. Furthermore, a limited number of data sets for which the structure could not be solved are included in the sample used for the paper; such data sets are typically neither deposited nor shared. Thus, the differences in results between the pipelines should not be considered as a quantitative estimate of success-rate improvement for user data but rather as qualitative evidence that the improved $|F_a|$ and E estimates by *Afro* may lead to successful substructure determination and model building for data sets where it failed using ΔE.

The multivariate $|F_a|$ estimation by *Afro* has been integrated into the *Crank*2 pipeline for automated structure solution from experimental phases and is distributed as part of the *CCP*4 package, which is available as a binary and as open source.

APPENDIX *A*
Derivation of the expected value of $|F_a|$

The expected value of $|F_a|$ is calculated, by definition, from the conditional probability distribution $P(|F_a|; |F^+|, |F^-|)$,

$$\langle |F_a|; |F^+|, |F^-| \rangle = \frac{\int_0^\infty |F_a| \int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} P(|F_a|, \alpha_a, |F^+|, \alpha^+, |F^-|, \alpha^-)\, d\alpha^+\, d\alpha^-\, d\alpha_a\, dF_a}{\int_0^\infty \int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} P(|F_a|, \alpha_a, |F^+|, \alpha^+, |F^-|, \alpha^-)\, d\alpha^+\, d\alpha^-\, d\alpha_a\, dF_a}. \quad (4)$$

The top and bottom integrals are the first and zeroth moments of the distribution $P(|F_a|; |F^+|, |F^-|)$, which can be obtained from the joint distribution of structure factors F_a, F^+, $(F^-)^*$, which can be approximated by a complex multivariate normal of mean zero and covariance Σ,

$$\begin{aligned} P(|F_a|, \alpha_a, |F^+|, \alpha^+, |F^-|, \alpha^-) &= \frac{|F_a||F^+||F^-|}{\pi^3 \det(\Sigma)} \\ &\times \exp(-a_{11}|F_a|^2 - a_{22}|F^+|^2 - a_{33}|F^-|^2) \\ &\times \exp\{-2|F_a||F^+|[a_{12}\cos(\alpha_a - \alpha^+) - b_{12}\sin(\alpha_a - \alpha^+)]\} \\ &\times \exp\{-2|F_a||F^-|[a_{13}\cos(\alpha_a - \alpha^-) - b_{13}\sin(\alpha_a - \alpha^-)]\} \\ &\times \exp\{-2|F^+||F^-|[a_{23}\cos(\alpha^+ - \alpha^-) - b_{23}\sin(\alpha^+ - \alpha^-)]\}, \end{aligned} \quad (5)$$

where a_{jk} and b_{jk} denote the real and imaginary components of the inverse covariance matrix. The zeroth, first and second moments of $|F_a|$ can be obtained by integrating out the unknown phase angles (α_a, α^+ and α^-) and averaging over $|F_a|$:

$$\begin{aligned} \langle |F_a|^n \rangle = \int_0^\infty \int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} &\frac{|F_a|^{n+1}|F^+||F^-|}{\pi^3 \det(\Sigma)} \\ &\times \exp(-a_{11}|F_a|^2 - a_{22}|F^+|^2 - a_{33}|F^-|^2) \\ &\times \exp\{-2|F^+||F^-|[a_{23}\cos(\alpha^+ - \alpha^-) - b_{23}\sin(\alpha^+ - \alpha^-)]\} \\ &\times \exp\{-2|F^+||F_a|[a_{12}\cos(\alpha^+ - \alpha_a) - b_{12}\sin(\alpha^+ - \alpha_a)]\} \\ &\times \exp\{-2|F^-||F_a|[a_{13}\cos(\alpha^- - \alpha_a) - b_{13}\sin(\alpha^- - \alpha_a)]\} \\ &d|F_a|\, d\alpha_a\, d\alpha^+\, d\alpha^-. \end{aligned} \quad (6)$$

Changing variables ($\beta = \alpha^+ - \alpha^-$, $\varphi = \alpha^- - \alpha_a$) leaves only an expression in β and φ; thus, α_a can be integrated out:

$$\begin{aligned} \langle |F_a|^n \rangle = &\frac{2|F^+||F^-|}{\pi^2 \det(\Sigma)} \\ &\times \exp(-a_{22}|F^+|^2 - a_{33}|F^-|^2) \int_0^\infty |F_a|^{n+1} \exp(-a_{11}|F_a|^2) \\ &\times \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \exp\{-2|F^+||F^-|[a_{23}\cos(\beta) - b_{23}\sin(\beta)]\} \\ &\times \exp\{-2|F^+||F_a|[a_{12}\cos(\beta + \varphi) - b_{12}\sin(\beta + \varphi)]\} \\ &\times \exp\{-2|F^-||F_a|[a_{13}\cos(\varphi) - b_{13}\sin(\varphi)]\}\, d|F_a|\, d\beta\, d\varphi, \end{aligned} \quad (7)$$

$$\langle|F_a|^n\rangle = \frac{2|F^+||F^-|}{\pi^2 \det(\Sigma)} \times \exp(-a_{22}|F^+|^2 - a_{33}|F^-|^2) \int_0^\infty |F_a|^{n+1} \exp(-a_{11}|F_a|^2) \times \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \exp\{-2|F^+||F^-|[a_{23}\cos(\beta) - b_{23}\sin(\beta)]\} \times \exp\left(-2|F_a|\{\cos(\varphi)[|F^+|a_{12}\cos(\beta) - |F^+|b_{12}\sin(\beta) + |F^-|a_{13}] + \sin(\varphi)[|F^+|a_{12}\sin(\beta) - |F^+|b_{12}\sin(\beta) + |F^-|b_{13}]\}\right) \mathrm{d}|F_a|\,\mathrm{d}\beta\,\mathrm{d}\varphi. \quad (8)$$

Using the formula $\int_{-\pi}^{\pi} \exp[a\cos(x) + b\sin(x)]\,\mathrm{d}x = 2\pi I_0[(a^2+b^2)^{1/2}]$, the following equation results:

$$\langle|F_a|^n\rangle = \frac{4|F^+||F^-|}{\pi \det(\Sigma)} \exp(-a_{22}|F^+|^2 - a_{33}|F^-|^2) \times \int_0^\infty |F_a|^{n+1} \exp(-a_{11}|F_a|^2) \times \int_{-\pi}^{\pi} \exp\{-2|F^+||F^-|[a_{23}\cos(\beta) - b_{23}\sin(\beta)]\} \times I_0(2|F_a|\xi^{1/2})\,\mathrm{d}|F_a|\,\mathrm{d}\beta, \quad (9)$$

where

$$\begin{aligned}\xi &= [|F^+|a_{12}\cos(\beta) - |F^+|b_{12}\sin(\beta) + |F^-|a_{13}]^2 + [|F^+|a_{12}\sin(\beta) - |F^+|b_{12}\sin(\beta) + |F^-|b_{13}]^2 \\ &= (a_{12}^2 + b_{12}^2)|F^+|^2 + (a_{13}^2 + b_{13}^2)|F^-|^2 + 2|F^+||F^-|[(a_{12}a_{13} + b_{12}b_{13})\cos(\beta) + (a_{12}b_{13} - a_{13}b_{12})\sin(\beta)]. \end{aligned} \quad (10)$$

The integral over $|F_a|$ has an analytical solution:

$$\langle|F_a|^n\rangle = \frac{2|F^+||F^-|\Gamma(\frac{n+2}{2})}{\pi a_{11}^{\frac{n+2}{2}} \det(\Sigma)} \exp(-a_{22}|F^+|^2 - a_{33}|F^-|^2) \times \int_{-\pi}^{\pi} \exp\{-2|F^+||F^-|[a_{23}\cos(\beta) - b_{23}\sin(\beta)]\} \times \Phi\left(\frac{n+2}{2}, 1, \frac{\xi}{a_{11}}\right)\mathrm{d}\beta. \quad (11)$$

To derive the equation that assumes that the phases of the Friedel pairs are equal while considering measurement errors in the observed structure-factor amplitudes, we assume that $\beta = 0$ and the equations reduce to the following:

$$\langle|F_a|^n\rangle = \frac{2|F^+||F^-|\Gamma(\frac{n+2}{2})}{\pi a_{11}^{\frac{n+2}{2}} \det(\Sigma)} \exp(-a_{22}|F^+|^2 - a_{33}|F^-|^2) \times \exp(-2a_{23}|F^+||F^-|) \times \Phi\left(\frac{n+2}{2}, 1, \frac{\xi}{a_{11}}\right), \quad (12)$$

$$\xi = (a_{12}^2 + b_{12}^2)|F^+|^2 + (a_{13}^2 + b_{13}^2)|F^-|^2 + 2|F^+||F^-|[(a_{12}a_{13} + b_{12}b_{13})]. \quad (13)$$

Substituting equation (11) for $n = 1$ in the numerator of equation (3) and for $n = 0$ in its denominator and using the formulas $\Phi(\frac{3}{2}, 1, x) = \Phi(-\frac{1}{2}, 1, -x)\exp(x)$ and $\Phi(1, 1, x) = \exp(x)$, we obtain

$$P(|F_a|, \alpha_a, |F^+|, \alpha^+, |F^-|, \alpha^-) = \frac{1}{2}\left(\frac{\pi}{a_{11}}\right)^{1/2} \Phi\left(-\frac{1}{2}, 1, -\frac{\xi}{a_{11}}\right). \quad (14)$$

In the case of the marginal distribution $P(F^+, F^-)$ without the Friedel pair phase equality assumption, substituting $n = 0$ and using the relation $\Phi(1, 1, x) = \exp(x)$ gives

$$\begin{aligned} P(|F^+|, |F^-|) &= \frac{2|F^+||F^-|}{\pi a_{11}\det(\Sigma)} \exp(-a_{22}|F^+|^2 - a_{33}|F^-|^2) \times \int_{-\pi}^{\pi} \exp\{-2|F^+||F^-|[a_{23}\cos(\beta) - b_{23}\sin(\beta)]\} \times \exp\left(\frac{\xi}{a_{11}}\right)\mathrm{d}\beta \\ &= \frac{2|F^+||F^-|}{\pi a_{11}\det(\Sigma)} \exp\left[-\left(a_{22} - \frac{a_{12}^2 + b_{12}^2}{a_{11}}\right)|F^+|^2\right] \times \exp\left[-\left(a_{33} - \frac{a_{13}^2 + b_{13}^2}{a_{11}}|F^-|^2\right)\right] \\ &\quad \times \int_{-\pi}^{\pi} \exp\left\{-2|F^+||F^-|\left[\left(a_{23} - \frac{a_{12}a_{13} + b_{12}b_{13}}{a_{11}}\right)\cos(\beta) - \left(b_{23} - \frac{a_{12}b_{13} - a_{13}b_{12}}{a_{11}}\right)\sin(\beta)\right]\right\}\mathrm{d}\beta \\ &= \frac{4|F^+||F^-|}{a_{11}\det(\Sigma)} \exp\left[-\left(a_{22} - \frac{a_{12}^2 + b_{12}^2}{a_{11}}\right)|F^+|^2\right] \times \exp\left[-\left(a_{33} - \frac{a_{13}^2 + b_{13}^2}{a_{11}}|F^-|^2\right)\right] \\ &\quad \times I_0\left\{2|F^+||F^-|\left[\left(a_{23} - \frac{a_{12}a_{13} + b_{12}b_{13}}{a_{11}}\right)^2 + \left(b_{23} - \frac{a_{12}b_{13} - a_{13}b_{12}}{a_{11}}\right)^2\right]^{1/2}\right\}. \end{aligned} \quad (15)$$

The covariance matrix Σ is calculated as follows:

$$\Sigma = \begin{pmatrix} \varepsilon\Sigma_H & \varepsilon(\Sigma_H - i\Sigma f' f'') & \varepsilon(\Sigma_H + i\Sigma f' f'') \\ \varepsilon(\Sigma_H + i\Sigma f' f'') & k^2(\varepsilon\Sigma_N + \sigma_{F^+}^2) & \varepsilon(\Sigma_N - \Sigma f''^2 - 2i\Sigma f' f'') \\ \varepsilon(\Sigma_H - i\Sigma f' f'') & \varepsilon(\Sigma_N - \Sigma f''^2 + 2i\Sigma f' f'') & k^2(\varepsilon\Sigma_N + \sigma_{F^-}^2) \end{pmatrix}, \quad (16)$$

where σ_{F^+} and σ_{F^-} denote the measurement errors of $|F^+|$ and $|F^-|$, respectively, $\Sigma_N = (\langle|F^+|^2 + |F^-|^2\rangle)/2$, $\Sigma_H = \Sigma(f_o + f')^2 + f''^2$, f_o is the non-anomalous scattering factor of the anomalously scattering atom type, f' and f'' are the real and imaginary anomalous scattering factors, k is a refinable local scale factor and ε is a symmetry-related statistical weight of reflection counting how many times the symmetry operations map the reflection to itself. All of the covariance matrix terms and summations are calculated per resolution bin, except for σ_{F^+}, σ_{F^-} and ε which are applied per reflection.

APPENDIX *B*
Complete list of PDB codes of the data sets used for testing

A total of 182 SAD data sets for 170 macromolecular structures were used for testing. The sample consisted of 169 data sets for 157 structures used by Skubák (2018): PDB entries 1c8u, 1djl, 1dpx, 1dtx, 1dw9, 1e3m, 1e42, 1e6i, 1fj2, 1fse, 1ga1, 1hf8, 1h29, 1i4u, 1lvy, 1lz8, 1m32, 1mso, 1ocy, 1of3, 1rgg, 1rju, 1vjn, 1vjr, 1vjz, 1vk4, 1vkm, 1vlm, 1vqr, 1z82, 1zy9, 1zyb, 2a3n, 2a6b, 2ahy, 2aml, 2avn, 2b78, 2b79, 2b8m, 2etd, 2etj, 2ets, 2etv, 2evr, 2f4p, 2fdn, 2fea, 2ffj, 2fg0, 2fg9, 2fna, 2fqp, 2fur, 2fzt, 2g42, 2g4h, 2g4j, 2g4k, 2g4l, 2g4m, 2g4n, 2g4o, 2g4p, 2g4q, 2g4r, 2g4s, 2g4t, 2g4u, 2g4v, 2g4w, 2g4x, 2g4y, 2g4z, 2g51, 2g52, 2g55, 2gc9, 2hba, 2ill, 2nlv, 2nuj, 2nwv, 2o08, 2o0h, 2o1q, 2o2x, 2o2z, 2o3l, 2o62, 2o7t, 2o8q, 2obp, 2oc5, 2od5, 2od6, 2oh3, 2okc, 2okf, 2ooj, 2opk, 2osd, 2otm, 2ozg, 2ozj, 2p10, 2p4o, 2p7h, 2p7i, 2p97, 2pg3, 2pg4, 2pgc, 2pim, 2pn1, 2ppv, 2pr7, 2prr, 2prv, 2prx, 2pv4, 2pw4, 2q2l, 2rkk, 2v0o, 3bpj, 3fki, 3gyv, 3k9g, 3km3, 3lmt, 3lmu, 3men, 3njb, 3o2e, 3oib, 3p96, 3s6l, 4us7, 4xvz, 4xxt, 4yf1, 5b82, 5gwd, 5ifg, 5irr, 5j4r, 5kjh, 5lg6, 5llw, 5loi, 5lsq, 5sus and 5suu, and three undeposited data sets. Furthermore, 13 more recent data sets for 13 different structures deposited in the previous few years were randomly chosen from the PDB and added to the sample: PDB entries 6kvr, 6tke, 6xjn, 6xqi, 6ygu, 6yrl, 7cdw, 7eiv, 7fad, 7fi4, 7lt1, 7oc3 and 7yx8.

Funding information

Funding for this work was provided by NWO (https://www.nwo.nl) Applied Sciences and Engineering Domain and CCP4 (https://www.ccp4.ac.uk; grant No. 16219).

References

Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A. & Sorensen, D. (1999). *LAPACK Users' Guide*, 3rd ed. Philadelphia: Society for Industrial and Applied Mathematics.
Blessing, R. H. (1997). *J. Appl. Cryst.* **30**, 176–177.
Burla, M. C., Carrozzini, B., Cascarano, G. L., Giacovazzo, C. & Polidori, G. (2003). *Acta Cryst.* D**59**, 662–669.
Burla, M. C., Carrozzini, B., Cascarano, G. L., Giacovazzo, C., Polidori, G. & Siliqi, D. (2002). *Acta Cryst.* D**58**, 928–935.
Cowtan, K. (2008). *Acta Cryst.* D**64**, 83–89.
Cowtan, K. (2010). *Acta Cryst.* D**66**, 470–478.
Dall'Antonia, F. & Schneider, T. R. (2006). *J. Appl. Cryst.* **39**, 618–619.
Elsliger, M.-A., Deacon, A. M., Godzik, A., Lesley, S. A., Wooley, J., Wüthrich, K. & Wilson, I. A. (2010). *Acta Cryst.* F**66**, 1137–1142.
Grosse-Kunstleve, R. W. & Adams, P. D. (2003). *Acta Cryst.* D**59**, 1966–1973.
Hatti, K. S., McCoy, A. J. & Read, R. J. (2021). *Acta Cryst.* D**77**, 880–893.
Nicholls, R. A., Tykac, M., Kovalevskiy, O. & Murshudov, G. N. (2018). *Acta Cryst.* D**74**, 492–505.
Pannu, N. S. (2007). *Acta Cryst.* A**63**, s80.
Pannu, N. S., McCoy, A. J. & Read, R. J. (2003). *Acta Cryst.* D**59**, 1801–1808.
Schneider, T. R. & Sheldrick, G. M. (2002). *Acta Cryst.* D**58**, 1772–1779.
Skubák, P. (2018). *Acta Cryst.* D**74**, 117–124.
Skubák, P. & Pannu, N. S. (2013). *Nat. Commun.* **4**, 2777.
Terwilliger, T. C. (1994). *Acta Cryst.* D**50**, 11–16.
Usón, I. & Sheldrick, G. M. (2018). *Acta Cryst.* D**74**, 106–116.
Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., Keegan, R. M., Krissinel, E. B., Leslie, A. G. W., McCoy, A., McNicholas, S. J., Murshudov, G. N., Pannu, N. S., Potterton, E. A., Powell, H. R., Read, R. J., Vagin, A. & Wilson, K. S. (2011). *Acta Cryst.* D**67**, 235–242.