STRUCTURAL BIOLOGY
ISSN 2059-7983

Bulk-solvent and overall scaling revisited: faster calculations, improved results. Corrigendum.

P. V. Afonine,[a]* R. W. Grosse-Kunstleve,[a] P. D. Adams[a,b] and A. Urzhumtsev[c,d]

[a]Lawrence Berkeley National Laboratory, One Cyclotron Road, MS64R0121, Berkeley, CA 94720, USA, [b]Department of Bioengineering, University of California Berkeley, Berkeley, CA 94720, USA, [c]IGBMC, CNRS–INSERM–UdS, 1 Rue Laurent Fries, BP 10142, 67404 Illkirch, France, and [d]Université Nancy: Département de Physique – Nancy 1, BP 239, Faculté des Sciences et des Technologies, 54506 Vandoeuvre-lès-Nancy, France. *Correspondence e-mail: pafonine@lbl.gov

Received 23 May 2023
Accepted 2 June 2023

Edited by R. J. Read, University of Cambridge, United Kingdom

Keywords: *Phenix*; anisotropy; bulk solvent; scaling.

Equations in Sections 2.3 and 2.4 of the article by Afonine *et al.* [*Acta Cryst.* (2013). D**69**, 625–634] are corrected.

In the article by Afonine *et al.* (2013) some improper notations and errors in several equations in Sections 2.3 and 2.4 have been corrected. We note that the *Computational Crystallography Toolbox* (Grosse-Kunstleve *et al.*, 2002) has been using the correct version of these equations since 2013. Updated versions of Section 2.3 and equations (42), (43) and (45) are given below.

2.3. Bulk-solvent parameters and overall isotropic scaling

Assuming the resolution-dependent scale factors $k_{\rm mask}(\mathbf{s})$ and $k_{\rm isotropic}(\mathbf{s})$ to be constants $k_{\rm mask}$ and $k_{\rm isotropic}$ in each thin resolution shell, the determination of their values is reduced to minimizing the residual

$$\sum_{\mathbf{s}}\{|\mathbf{F}_{\rm calc}(\mathbf{s})+k_{\rm mask}\mathbf{F}_{\rm mask}(\mathbf{s})|^2 - [k_{\rm overall}\,k_{\rm anisotropic}(\mathbf{s})k_{\rm isotropic}]^{-2}F^2_{\rm obs}(\mathbf{s})\}^2, \quad (22)$$

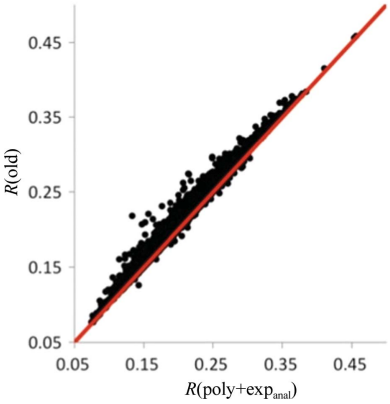
where the sum is calculated over all reflections $\mathbf{s}$ in the given resolution shell, and $k_{\rm overall}$ and $k_{\rm anisotropic}(\mathbf{s})$ are calculated previously and fixed. This minimization problem is generally highly over-determined because the number of reflections per shell is usually much larger than two.

Introducing $w_{\mathbf{s}} = |\mathbf{F}_{\rm mask}(\mathbf{s})|^2$, $v_{\mathbf{s}} = \frac{1}{2}[\mathbf{F}_{\rm calc}(\mathbf{s})\mathbf{F}^*_{\rm mask}(\mathbf{s}) + \mathbf{F}^*_{\rm calc}(\mathbf{s})\mathbf{F}_{\rm mask}(\mathbf{s})]$, $u_{\mathbf{s}} = |\mathbf{F}_{\rm calc}(\mathbf{s})|^2$, $I_{\mathbf{s}} = [k_{\rm overall}\,k_{\rm anisotropic}(\mathbf{s})]^{-2}F^2_{\rm obs}(\mathbf{s})$ and $K = k^{-2}_{\rm isotropic}$ and substituting them into (22) leads to the minimization of

$$\mathrm{LS}(K, k_{\rm mask}) = \sum_{\mathbf{s}}[(k^2_{\rm mask}w_{\mathbf{s}} + 2k_{\rm mask}v_{\mathbf{s}} + u_{\mathbf{s}}) - KI_{\mathbf{s}}]^2 \quad (23)$$

with respect to K and $k_{\rm mask}$. This leads to a system of two equations:

$$\begin{cases} \dfrac{\partial}{\partial K}\mathrm{LS}(K, k_{\rm mask}) = -2\sum\limits_{\mathbf{s}}[(k^2_{\rm mask}w_{\mathbf{s}} + 2k_{\rm mask}v_{\mathbf{s}} + u_{\mathbf{s}}) - KI_{\mathbf{s}}]I_{\mathbf{s}} \\ \qquad\qquad = 0, \\ \dfrac{\partial}{\partial k_{\rm mask}}\mathrm{LS}(K, k_{\rm mask}) = 4\sum\limits_{\mathbf{s}}[(k^2_{\rm mask}w_{\mathbf{s}} + 2k_{\rm mask}v_{\mathbf{s}} + u_{\mathbf{s}}) - KI_{\mathbf{s}}] \\ \qquad\qquad \times (k_{\rm mask}w_{\mathbf{s}} + v_{\mathbf{s}}) \\ \qquad\qquad = 0. \end{cases} \quad (24)$$



Developing these equations with respect to $k_{\rm mask}$,

$$\begin{cases} k_{\rm mask}^2 \sum\limits_{\mathbf{s}} w_{\mathbf{s}} I_{\mathbf{s}} + 2k_{\rm mask} \sum\limits_{\mathbf{s}} v_{\mathbf{s}} I_{\mathbf{s}} + \sum\limits_{\mathbf{s}} u_{\mathbf{s}} I_{\mathbf{s}} - K \sum\limits_{\mathbf{s}} I_{\mathbf{s}}^2 = 0, \\ k_{\rm mask}^3 \sum\limits_{\mathbf{s}} w_{\mathbf{s}}^2 + 3k_{\rm mask}^2 \sum\limits_{\mathbf{s}} w_{\mathbf{s}} v_{\mathbf{s}} + k_{\rm mask} \sum\limits_{\mathbf{s}} (2v_{\mathbf{s}}^2 + u_{\mathbf{s}} w_{\mathbf{s}} - K I_{\mathbf{s}} w_{\mathbf{s}}) \\ \qquad + \sum\limits_{\mathbf{s}} u_{\mathbf{s}} v_{\mathbf{s}} - K \sum\limits_{\mathbf{s}} I_{\mathbf{s}} v_{\mathbf{s}} = 0, \end{cases} \tag{25}$$

and introducing new notations for the coefficients, we obtain

$$\begin{cases} k_{\rm mask}^2 C_2 + k_{\rm mask} B_2 + A_2 - K Y_2 = 0, \\ k_{\rm mask}^3 D_3 + k_{\rm mask}^2 C_3 + k_{\rm mask} (B_3 - K C_2) + A_3 - K Y_3 = 0. \end{cases} \tag{26}$$

Multiplying the second equation by Y_2 and substituting KY_2 from the first equation into the new second equation, we obtain a cubic equation with fixed coefficients

$$\begin{aligned} &k_{\rm mask}^3 (D_3 Y_2 - C_2^2) + k_{\rm mask}^2 (C_3 Y_2 - C_2 B_2 - C_2 Y_3) \\ &\quad + k_{\rm mask} (B_3 Y_2 - C_2 A_2 - Y_3 B_2) + (A_3 Y_2 - Y_3 A_2) = 0. \end{aligned} \tag{27}$$

The senior coefficient in equation (27) satisfies the Cauchy–Schwarz inequality:

$$D_3 Y_2 - C_2^2 = \sum_{\mathbf{s}} w_{\mathbf{s}}^2 \sum_{\mathbf{s}} I_{\mathbf{s}}^2 - \sum_{\mathbf{s}} w_{\mathbf{s}} I_{\mathbf{s}} \sum_{\mathbf{s}} w_{\mathbf{s}} I_{\mathbf{s}} > 0. \tag{28}$$

Therefore, equation (27) can be rewritten as

$$k_{\rm mask}^3 + a k_{\rm mask}^2 + b k_{\rm mask} + c = 0 \tag{29}$$

and solved using a standard procedure.

The corresponding values of K are obtained by substituting the roots of equation (29) into the first equation in equation (26),

$$K = (k_{\rm mask}^2 C_2 + k_{\rm mask} B_2 + A_2)/Y_2. \tag{30}$$

If no positive root exists, $k_{\rm mask}$ is assigned a zero value, which implies the absence of a bulk-solvent contribution. If several roots with $k_{\rm mask} \geq 0$ exist then the one that gives the smallest value of LS(K, $k_{\rm mask}$) is selected.

If desired, one can fit the right-hand side of expression (10) to the array of $k_{\rm mask}$ values by minimizing the residual

$$\sum_{\mathbf{s}} [k_{\rm mask} - k_{\rm sol} \exp(-B_{\rm sol}\, s^2/4)]^2 \tag{31}$$

for all $k_{\rm mask} > 0$. This can be achieved analytically as described in Appendix *A*. Similarly, one can fit $k_{\rm overall} \exp(-B_{\rm overall}\, s^2/4)$ to the array of K values.

Equations (42), (43) and (45) in Section 2.4 of Afonine *et al.* (2013) are also updated as follows

$$\mathbf{b} = \left[\sum_{\mathbf{s}} I(\mathbf{s}) I_1(\mathbf{s}_1), \ldots, \sum_{\mathbf{s}} I(\mathbf{s}) I_N(\mathbf{s}_N), 1 \right]^t, \tag{42}$$

$$\mathrm{LS}(K, k_{\rm mask}) = \sum_{\mathbf{s}} \left\{ \left[\sum_{j=1}^{N} \alpha_j |\mathbf{F}_{\rm calc}(\mathbf{s}_j) + k_{\rm mask} \mathbf{F}_{\rm mask}(\mathbf{s}_j)|^2 \right] - K I_{\mathbf{s}} \right\}^2, \tag{43}$$

$$\mathrm{LS}(K, k_{\rm mask}) = \sum_{\mathbf{s}} \left[(k_{\rm mask}^2 w_{\mathbf{s}} + 2k_{\rm mask} v_{\mathbf{s}} + u_{\mathbf{s}}) - K I_{\mathbf{s}} \right]^2. \tag{45}$$

References

Afonine, P. V., Grosse-Kunstleve, R. W., Adams, P. D. & Urzhumtsev, A. (2013). *Acta Cryst.* D**69**, 625–634.

Grosse-Kunstleve, R. W., Sauter, N. K., Moriarty, N. W. & Adams, P. D. (2002). *J. Appl. Cryst.* **35**, 126–136.

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Bulk-solvent and overall scaling revisited: faster calculations, improved results

P. V. Afonine,[a]* R. W. Grosse-Kunstleve,[a] P. D. Adams[a,b] and A. Urzhumtsev[c,d]

[a]Lawrence Berkeley National Laboratory, One Cyclotron Road, MS64R0121, Berkeley, CA 94720, USA, [b]Department of Bioengineering, University of California, Berkeley, Berkeley, CA 94720, USA, [c]IGBMC, CNRS–INSERM–UdS, 1 Rue Laurent Fries, BP 10142, 67404 Illkirch, France, and [d]Université de Lorraine: Département de Physique – Nancy 1, BP 239, Faculté des Sciences et des Technologies, 54506 Vandoeuvre-lès-Nancy, France

Correspondence e-mail: pafonine@lbl.gov

Received 13 December 2012
Accepted 5 January 2013

A fast and robust method for determining the parameters for a flat (mask-based) bulk-solvent model and overall scaling in macromolecular crystallographic structure refinement and other related calculations is described. This method uses analytical expressions for the determination of optimal values for various scale factors. The new approach was tested using nearly all entries in the PDB for which experimental structure factors are available. In general, the resulting R factors are improved compared with previously implemented approaches. In addition, the new procedure is two orders of magnitude faster, which has a significant impact on the overall runtime of refinement and other applications. An alternative function is also proposed for scaling the bulk-solvent model and it is shown that it outperforms the conventional exponential function. Similarly, alternative methods are presented for anisotropic scaling and their performance is analyzed. All methods are implemented in the *Computational Crystallography Toolbox* (*cctbx*) and are used in *PHENIX* programs.

1. Introduction

Macromolecular crystals typically contain a substantial amount of disordered solvent, ranging from approximately 20 to 90% of the crystal volume, with a mean of 55%, in the Protein Data Bank (PDB; Bernstein *et al.*, 1977; Berman *et al.*, 2000). Anisotropy in the diffracted intensities is another common feature of macromolecular crystals that arises from various sources including crystal lattice vibrations (Shakked, 1983; Sheriff & Hendrickson, 1987). When modelling diffracted intensities, for example in structure refinement or automated model building, it is therefore critical to account for these two phenomena (see, for example, Jiang & Brünger, 1994; Urzhumtsev & Podjarny, 1995; Kostrewa, 1997; Badger, 1997; Urzhumtsev, 2000; Fokine & Urzhumtsev, 2002*a*; Fenn *et al.*, 2010). The flat bulk-solvent model (Phillips, 1980; Jiang & Brünger, 1994) combined with overall anisotropic scaling in either exponential (Sheriff & Hendrickson, 1987) or polynomial (Usón *et al.*, 1999) forms is a well established and computationally efficient approach. Alternatives have been proposed (Tronrud, 1997; Vassylyev *et al.*, 2007), but are not currently in wide use.

In the commonly used approach, the total structure factor is defined as

$$\mathbf{F}_{\rm model} = k_{\rm total}(\mathbf{F}_{\rm calc} + k_{\rm mask}\mathbf{F}_{\rm mask}), \qquad (1)$$

where $k_{\rm total}$ is the overall Miller-index-dependent scale factor, $\mathbf{F}_{\rm calc}$ and $\mathbf{F}_{\rm mask}$ are the structure factors computed from the atomic model and the bulk-solvent mask, respectively, and $k_{\rm mask}$ is a bulk-solvent scale factor. The mask can be computed

efficiently using exact asymmetric units as described in Grosse-Kunstleve *et al.* (2011).

The overall scale factor $k_{\rm total}$ can be thought of as the product

$$k_{\rm total} = k_{\rm overall}\, k_{\rm isotropic}\, k_{\rm anisotropic}, \quad (2)$$

where $k_{\rm overall}$ is the overall scale factor and $k_{\rm isotropic}$ and $k_{\rm anisotropic}$ are the isotropic and anisotropic scale factors, respectively.

$k_{\rm overall}$ is a scalar number that can be obtained by minimizing the least-squares residual

$$\mathrm{LS} = \sum (F_{\rm obs} - k_{\rm overall}|\mathbf{F}'_{\rm model}|)^2, \quad (3)$$

where $F_{\rm obs}$ are the observed structure factors and

$$\mathbf{F}'_{\rm model} = k_{\rm isotropic}\, k_{\rm anisotropic}(\mathbf{F}_{\rm calc} + k_{\rm mask}\mathbf{F}_{\rm mask}). \quad (4)$$

The sum is over all reflections. Solving $\partial \mathrm{LS}/\partial k_{\rm overall} = 0$ leads to

$$k_{\rm overall} = \sum F_{\rm obs}|\mathbf{F}'_{\rm model}| / \sum |\mathbf{F}'_{\rm model}|^2. \quad (5)$$

In the exponential model the anisotropic scale factor is defined as

$$k_{\rm anisotropic} = \exp(-2\pi^2 \mathbf{s}^t \mathbf{U}_{\rm cryst}\mathbf{s}), \quad (6)$$

where $\mathbf{U}_{\rm cryst}$ is the overall anisotropic scale matrix equivalent to $\mathbf{U}^*$ defined in Grosse-Kunstleve & Adams (2002); $\mathbf{s}^t = (h, k, l)$ is the transpose of the Miller-index column vector $\mathbf{s}$.

Usón *et al.* (1999) define a polynomial anisotropic scaling function that can be rewritten in matrix notation as follows:

$$k_{\rm anisotropic} = \mathbf{s}^t\mathbf{V}_0\mathbf{s} + (\mathbf{s}^t\mathbf{V}_1\mathbf{s})s^2, \quad (7)$$

where $\mathbf{V}_0$ and $\mathbf{V}_1$ are symmetric 3×3 matrices, $s^2 = \mathbf{s}^t\mathbf{G}^*\mathbf{s}$ and $\mathbf{G}^*$ is the reciprocal-space metric tensor. Expression (7) is equivalent to the first terms in the Taylor series expansion of the exponential function (6),

$$\exp(-2\pi^2\mathbf{s}^t\mathbf{U}_{\rm cryst}\mathbf{s}) \simeq 1 - 2\pi^2\mathbf{s}^t\mathbf{U}_{\rm cryst}\mathbf{s} + 2\pi^4(\mathbf{s}^t\mathbf{U}_{\rm cryst}\mathbf{s})(\mathbf{s}^t\mathbf{U}_{\rm cryst}\mathbf{s}), \quad (8)$$

with the constant term omitted. The omission of the constant 1 means that $k_{\rm anisotropic}$ is equal to zero for the reflection $\mathbf{F}_{000}$, as follows from (7). Therefore, in this work we modify (7) by adding the constant

$$k_{\rm anisotropic} = 1 + \mathbf{s}^t\mathbf{V}_0\mathbf{s} + (\mathbf{s}^t\mathbf{V}_1\mathbf{s})s^2. \quad (9)$$

The bulk-solvent scale factor is traditionally defined as

$$k_{\rm mask} = k_{\rm sol}\exp(-B_{\rm sol}s^2/4), \quad (10)$$

where $k_{\rm sol}$ and $B_{\rm sol}$ are the flat bulk-solvent model parameters (Phillips, 1980; Jiang & Brünger, 1994; Fokine & Urzhumtsev, 2002*b*).

Depending on the calculation protocol, $k_{\rm isotropic}$ may be assumed to be a part of $k_{\rm anisotropic}$ or it can be assumed to be exponential: $k_{\rm isotropic} = \exp(-Bs^2/4)$, where B is a scalar parameter. Alternatively, it may be determined as described in §2.3 below.

The determination of the anisotropic scaling parameters ($\mathbf{U}_{\rm cryst}$ or $\mathbf{V}_0$ and $\mathbf{V}_1$) and the bulk-solvent parameters $k_{\rm sol}$ and $B_{\rm sol}$ requires the minimization of the target function (3) with respect to these parameters. Despite the apparent simplicity, this task is quite involved owing to a number of numerical issues (Fokine & Urzhumtsev, 2002*b*; Afonine *et al.*, 2005*a*). Previously, we have developed a robust and thorough procedure (Afonine *et al.*, 2005*a*) to address these issues. This procedure is used routinely in *PHENIX* (Adams *et al.*, 2010). However, owing to its thoroughness the procedure is relatively slow and may account for a significant fraction of the execution time of certain *PHENIX* applications (for example, *phenix.refine*).

In this paper, we describe a new procedure which is approximately two orders of magnitude faster than the approach described in Afonine *et al.* (2005*a*) and often leads to a better fit of the experimental data. The speed gain is the result of an analytical determination of the optimal bulk-solvent and scaling parameters. The better fit to the experimental data is partially the result of employing a more detailed model for $k_{\rm mask}$ compared with the exponential model in equation (10) and is partially a consequence of the new analytical optimization method. Analytical optimization eliminates the possibility of becoming trapped in local minima, which exists in all iterative local optimization methods, including the procedure used previously.

2. Methods

2.1. Anisotropic scaling: exponential model

To obtain the elements of the anisotropic scaling matrix (6), the minimization of (3) is replaced by the minimization of

$$\mathrm{LSL} = \sum_{\mathbf{s}} [\ln(F_{\rm obs}) - \ln(|\mathbf{F}_{\rm model}|)]^2. \quad (11)$$

For this, we assume that $F_{\rm obs}$ and $|\mathbf{F}_{\rm model}|$ are positive. We also assume that the minima of (3) and (11) are at similar locations. This assumption is not obvious and, as discussed below, may not always hold (see §3.3 and Table 2). Expression (11) can be rewritten as

$$\mathrm{LSL} = (2\pi^2)^2 \sum_{\mathbf{s}} (Z + \mathbf{s}^t\mathbf{U}_{\rm cryst}\mathbf{s})^2. \quad (12)$$

Here, $Z = [1/(2\pi^2)]\ln[F_{\rm obs}(k_{\rm overall}k_{\rm isotropic}|\mathbf{F}_{\rm calc} + k_{\rm mask}\mathbf{F}_{\rm mask}|)^{-1}]$. Defining

$$\widetilde{\mathrm{LSL}} = \mathrm{LSL}/(2\pi^2)^2 \quad (13)$$

and using

$$\mathbf{U}_{\rm cryst} = \begin{pmatrix} U_{11} & U_{12} & U_{13} \\ U_{12} & U_{22} & U_{23} \\ U_{13} & U_{23} & U_{33} \end{pmatrix}, \quad (14)$$

the target function determining the optimal $\mathbf{U}_{\rm cryst}$ is

$$\widetilde{\mathrm{LSL}} = \sum_{\mathbf{s}} (Z + U_{11}h^2 + U_{22}k^2 + U_{33}l^2 + 2U_{12}hk + 2U_{13}hl + 2U_{23}kl)^2. \quad (15)$$

The $\mathbf{U}_{\rm cryst}$ values that minimize (15) are determined from the condition $\nabla_{\mathbf{U}}\widetilde{LSL} = 0$, which gives a system of six linear equations

$$\mathbf{M}\,\mathbf{U}_{\rm cryst} = \mathbf{b}. \tag{16}$$

where $\mathbf{M} = \sum_{\mathbf{s}} \mathbf{V} \otimes \mathbf{V}$, $\mathbf{V} = (h^2, k^2, l^2, 2hk, 2hl, 2kl)^t$, $\otimes$ denotes the outer product and $\mathbf{b} = -\sum_{\mathbf{s}} Z\mathbf{V}$.

The desired $\mathbf{U}_{\rm cryst}$ matrix is determined by solving the system (16):

$$\mathbf{U}_{\rm cryst} = \mathbf{M}^{-1}\mathbf{b}. \tag{17}$$

Crystal-system-specific symmetry constraints can be incorporated *via* a constraint matrix (**C**), which we derive from first principles by solving the system of linear equations $\mathbf{R}^t\mathbf{U}\mathbf{R} = \mathbf{U}$ for all rotation matrices **R** of the crystal-system point group. Alternatively, symmetry constraints are often derived manually and tabulated (Nye, 1957; Giacovazzo, 1992). For example, the constraint matrix for the tetragonal crystal system is

$$\mathbf{C} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}. \tag{18}$$

The number of rows in **C** determines the number of independent coefficients of $\mathbf{U}_{\rm cryst}$. Let $\mathbf{U}_{\rm ind}$ be the column vector of independent coefficients; the (redundant) set of six coefficients $\mathbf{U}_{\rm cryst}$ is then obtained *via*

$$\mathbf{U}_{\rm cryst} = \begin{pmatrix} U_{11} & U_{22} & U_{33} & U_{12} & U_{13} & U_{23} \end{pmatrix} = \mathbf{C}^t\,\mathbf{U}_{\rm ind}. \tag{19}$$

The constraint matrix **C** is introduced into equations (16) and (17) above as follows:

$$\mathbf{M}_{\rm C}\mathbf{U}_{\rm ind} = \mathbf{b}_{\rm C} \tag{20}$$

with $\mathbf{M}_{\rm C} = \sum_{\mathbf{h}} \mathbf{V}_{\rm C} \otimes \mathbf{V}_{\rm C}$, $\mathbf{V}_{\rm C} = \mathbf{C}\mathbf{V}$, $\mathbf{b}_{\rm C} = -\sum_{\mathbf{h}} Z\mathbf{V}_{\rm C}$ and

$$\mathbf{U}_{\rm ind} = \mathbf{M}_{\rm C}^{-1}\mathbf{b}_{\rm C}. \tag{21}$$

The full $\mathbf{U}_{\rm cryst}$ is then determined *via* equation (19).

2.2. Anisotropic scaling: polynomial model

The polynomial model (Usón *et al.*, 1999) for anisotropic scaling allows the direct use of the residual (3) to find the optimal coefficients for $\mathbf{V}_0$ and $\mathbf{V}_1$ in equation (9). An advantage of this model is that no assumptions about the similarity of the location of the minima of targets (3) and (11) are required. Conceptually, a disadvantage of equation (9) is that it is only an approximation of equation (6), as was shown above. However, the number of parameters is doubled in equation (9) compared with equation (6), since $\mathbf{V}_0$ and $\mathbf{V}_1$ are treated independently. The increased number of degrees of freedom may therefore compensate for approximation inaccuracies.

Similarly to §2.1, the optimal coefficients for $\mathbf{V}_0$ and $\mathbf{V}_1$ are determined by the condition $\nabla_{\mathbf{V}}\mathrm{LS} = 0$ and can be obtained by solving a system of 12 linear equations. We follow the arguments of Usón *et al.* (1999) for not using symmetry constraints in this case.

2.3. Bulk-solvent parameters and overall isotropic scaling

Defining $K = k_{\rm total}^{-2} = (k_{\rm overall}k_{\rm isotropic}k_{\rm anisotropic})^{-2}$, the determination of the desired scaling parameters $k_{\rm isotropic}$ and $k_{\rm mask}$ is reduced to minimizing

$$\mathrm{LS}_s(K, k_{\rm mask}) = \sum_s (|\mathbf{F}_{\rm calc} + k_{\rm mask}\mathbf{F}_{\rm mask}|^2 - KI)^2 \tag{22}$$

in resolution bins, where $k_{\rm overall}$ and $k_{\rm anisotropic}$ are fixed. This minimization problem is generally highly overdetermined because the number of reflections per bin is usually much larger than two.

Introducing $w = |\mathbf{F}_{\rm mask}|^2$, $v = (\mathbf{F}_{\rm calc}, \mathbf{F}_{\rm mask})$ and $u = |\mathbf{F}_{\rm calc}|^2$ and substitution into (22) leads to

$$\mathrm{LS}_s(K, k_{\rm mask}) = \sum_s [(k_{\rm mask}^2 w + 2k_{\rm mask}v + u) - KI]^2. \tag{23}$$

Minimizing (23) with respect to K and $k_{\rm mask}$ leads to a system of two equations:

$$\begin{cases} \dfrac{\partial}{\partial K}\mathrm{LS}_s(K, k_{\rm mask}) = -\sum_s [(k_{\rm mask}^2 w_s + 2k_{\rm mask}v_s + u_s) - KI_s]I_s \\ \qquad\qquad = 0 \\ \dfrac{\partial}{\partial k_{\rm mask}}\mathrm{LS}_s(K, k_{\rm mask}) = 2\sum_s [(k_{\rm mask}^2 w_s + 2k_{\rm mask}v_s + u_s) - KI_s] \\ \qquad\qquad \times (k_{\rm mask}w_s + v_s) = 0. \end{cases} \tag{24}$$

Developing these equations with respect to $k_{\rm mask}$,

$$\begin{cases} k_{\rm mask}^2 \sum_s w_s I_s + 2k_{\rm mask}\sum_s v_s I_s + \sum_s u_s I_s - K\sum_s I_s^2 = 0 \\ k_{\rm mask}^3 \sum_s w_s + 3k_{\rm mask}^2 \sum_s w_s v_s + k_{\rm mask}\sum_s (2v_s^2 + u_s w_s - KI_s w_s) \\ \qquad + \sum_s u_s v_s - K\sum_s I_s v_s = 0, \end{cases} \tag{25}$$

and introducing new notations for the coefficients, we obtain

$$\begin{cases} k_{\rm mask}^2 C_2 + k_{\rm mask}B_2 + A_2 - KY_2 = 0 \\ k_{\rm mask}^3 D_3 + k_{\rm mask}^2 C_3 + k_{\rm mask}(B_3 - KC_2) + A_3 - KY_3 = 0. \end{cases} \tag{26}$$

Multiplying the second equation by Y_2 and substituting KY_2 from the first equation into the new second equation, we obtain a cubic equation

$$\begin{aligned} &k_{\rm mask}^3(D_3Y_2 - C_2^2) + k_{\rm mask}^2(C_3Y_2 - C_2B_2 - C_2Y_3) \\ &\quad + k_{\rm mask}(B_3Y_2 - C_2A_2 - Y_3B_2) + (A_3Y_2 - Y_3A_2) = 0. \end{aligned} \tag{27}$$

The senior coefficient in (27) satisfies the Cauchy–Schwarz inequality:

$$D_3Y_2 - C_2^2 = \sum_s w_s^2 \sum_s I_s^2 - \sum_s w_s I_s \sum_s w_s I_s > 0. \tag{28}$$

Therefore, equation (27) can be rewritten as

$$k_{\rm mask}^3 + ak_{\rm mask}^2 + bk_{\rm mask} + c = 0 \tag{29}$$

and solved using a standard procedure.

The corresponding values of K are obtained by substituting the roots of equation (29) into the first equation in (26):

$$K = (k_{\rm mask}^2 C_2 + k_{\rm mask}B_2 + A_2)/Y_2. \tag{30}$$

If no positive root exists $k_{\rm mask}$ is assigned a zero value, which implies the absence of a bulk-solvent contribution. If several roots with $k_{\rm mask} \geq 0$ exist then the one that gives the smallest value of $\mathrm{LS}_s(K, k_{\rm mask})$ is selected.

If desired, one can fit the right-hand side of expression (10) to the array of $k_{\rm mask}$ values by minimizing the residual

$$\mathrm{LS} = \sum_{\mathbf{s}} [k_{\rm mask} - k_{\rm sol} \exp(-B_{\rm sol} s^2/4)]^2 \qquad (31)$$

for all $k_{\rm mask} > 0$. This can be achieved analytically as described in Appendix *A*. Similarly, one can fit $k_{\rm overall}\exp(-B_{\rm overall}s^2/4)$ to the array of K values.

2.4. Presence of twinning

In case of twinning with N twin-related domains, the total model intensity is

$$I_{\rm model}(\mathbf{s}) = \sum_{j=1}^{N} \alpha_j I_j(\mathbf{T}_j\mathbf{s}), \qquad (32)$$

where α_j is the twin fraction of the jth domain, $\mathbf{T}_j$ is the corresponding twin operator (a 3×3 rotation matrix) and

$$I_j(\mathbf{T}_j\mathbf{s}) = k_{\rm total}(\mathbf{T}_j\mathbf{s})|\mathbf{F}_{\rm calc}(\mathbf{T}_j\mathbf{s}) + k_{\rm mask}(\mathbf{T}_j\mathbf{s})\mathbf{F}_{\rm mask}(\mathbf{T}_j\mathbf{s})|^2. \qquad (33)$$

$k_{\rm total}$ includes all scale factors (overall, isotropic and anisotropic). We make the reasonable assumption that $k_{\rm total}$ and $k_{\rm mask}$ are identical for all twin domains.

Finding the twin fractions α_j can be achieved by solving the minimization problem

$$\mathrm{LS}(\alpha_1, \ldots, \alpha_N) = \sum_{\mathbf{s}} \left[\sum_{j=1}^{N} \alpha_j I_j(\mathbf{s}_j) - I(\mathbf{s})\right]^2, \qquad (34)$$

with the constraint condition

$$C(\alpha_1, \ldots, \alpha_N) = \sum_{j=1}^{N} \alpha_j - 1 = 0, \qquad (35)$$

where $I(\mathbf{s}) = F_{\rm obs}^2$ and $\mathbf{s}_j = \mathbf{T}_j\mathbf{s}$. This constrained minimization problem can be reformulated as an unconstrained minimization problem by the standard technique of introducing a Lagrange multiplier:

$$\mathrm{LS}(\alpha_1, \ldots, \alpha_N, \lambda) = \mathrm{LS}(\alpha_1, \ldots, \alpha_N) + \lambda C(\alpha_1, \ldots, \alpha_N). \qquad (36)$$

The values $\{\alpha_1, \ldots, \alpha_N, \lambda\}$ that minimize (36) are the solution of the system of $N + 1$ linear equations with $N + 1$ variables:

$$\begin{cases} \partial \mathrm{LS}(\alpha_1, \ldots, \alpha_N, \lambda)/\partial\alpha_1 = 0 \\ \quad \ldots \\ \partial \mathrm{LS}(\alpha_1, \ldots, \alpha_N, \lambda)/\partial\alpha_N = 0 \\ \partial \mathrm{LS}(\alpha_1, \ldots, \alpha_N, \lambda)/\partial\lambda = 0 \end{cases} \qquad (37)$$

or

$$\begin{cases} \sum\limits_{\mathbf{s}} \left[\sum\limits_{j=1}^{N} \alpha_j I_j(\mathbf{s}_j) - I(\mathbf{s})\right] I_1(\mathbf{s}_1) + \lambda = 0 \\ \quad \vdots \\ \sum\limits_{\mathbf{s}} \left[\sum\limits_{j=1}^{N} \alpha_j I_j(\mathbf{s}_j) - I(\mathbf{s})\right] I_N(\mathbf{s}_N) + \lambda = 0 \\ \sum\limits_{j=1}^{N} \alpha_j - 1 = 0. \end{cases} \qquad (38)$$

The solution of this system is

$$(\tilde{\alpha}_1, \ldots, \tilde{\alpha}_N, \tilde{\lambda})^t = \mathbf{M}^{-1}\mathbf{b} \qquad (39)$$

with the $(N + 1) \times (N + 1)$ matrix

$$\mathbf{M} = \begin{pmatrix} \sum\limits_{\mathbf{s}} \mathbf{V} \otimes \mathbf{V} & \mathbf{1} \\ \mathbf{1} & 0 \end{pmatrix}, \qquad (40)$$

and

$$\mathbf{V} = [I_1(\mathbf{s}_1), \ldots, I_N(\mathbf{s}_N)]. \qquad (41)$$

Here, $\mathbf{1}$ is a row or column containing N unit elements to complete the matrix $\mathbf{M}$ and

$$\mathbf{b} = \left[\sum_{\mathbf{s}} I(\mathbf{s}) I_1(s_1), \ldots, \sum_{\mathbf{s}} I(\mathbf{s}) I_N(s_N), 1\right]^t. \qquad (42)$$

The values of ε are expected to be between 0 and 1, and λ is proportional to the sum of squared intensities. Therefore, it is numerically beneficial to multiply the $\lambda C(\alpha_1, \ldots, \alpha_N)$ term in (36) by a constant $\sum_{\mathbf{s}} I^2(\mathbf{s})$ in order to make the value for λ numerically similar to the values for the twin fractions α.

Once the twin fractions have been found, the procedure described in §2.3 can be used to obtain the overall and bulk-solvent scale factors. Similarly to (23), we can write

$$\mathrm{LS}_s(K, k_{\rm mask}) = \sum_{s} \left[\sum_{j=1}^{N} \alpha_j |\mathbf{F}_{\rm calc}(\mathbf{s}_j) + k_{\rm mask}\mathbf{F}_{\rm mask}(\mathbf{s}_j)|^2 - KI\right]^2, \qquad (43)$$

where α_j are known twin fractions and K and $k_{\rm mask}$ are the scale factors to be determined. Similarly to §2.3, we obtain

$$\begin{aligned} \sum_{j=1}^{N} \alpha_j |\mathbf{F}_{\rm calc}(\mathbf{s}_j) + k_{\rm mask}\mathbf{F}_{\rm mask}(\mathbf{s}_j)|^2 &= \sum_{j=1}^{N} \{\alpha_j |\mathbf{F}_{\rm calc}(\mathbf{s}_j)|^2 \\ &+ 2k_{\rm mask}\alpha_j [\mathbf{F}_{\rm calc}(\mathbf{s}_j)\mathbf{F}_{\rm mask}(\mathbf{s}_j)] + k_{\rm mask}^2 \alpha_j |\mathbf{F}_{\rm mask}(\mathbf{s}_j)|^2\}. \end{aligned} \qquad (44)$$

Introducing new variables as before for equation (23) leads to

$$\mathrm{LS}_s(K, k_{\rm mask}) = \sum_{s} [(k_{\rm mask}^2 w + 2k_{\rm mask} v + u) - KI]^2. \qquad (45)$$

The determination of the twin fractions α and scales $k_{\rm total}$ and $k_{\rm mask}$ are iterated several times until convergence. The determination of α does not guarantee that the individual twin fractions α_j are in the range 0–1. For any α_j outside this range the corresponding twin operation is ignored for the current iteration and the new smaller set of twin fractions and scales are redetermined. However, in the next iteration the full set of α is tried again.

3. Results

3.1. Implementation of the new protocol

The scale factors involved in the calculation of $\mathbf{F}_{\rm model}$ according to equation (1) are highly correlated. Therefore, the order of their determination is important. Empirically, we found that the determination of $k_{\rm isotropic}$ and $k_{\rm mask}$ followed by the determination of $k_{\rm anisotropic}$ works optimally in most cases. The determination of ($k_{\rm mask}$, $k_{\rm isotropic}$) and $k_{\rm anisotropic}$ is repeated several times until the R factor decreases by less than 0.01% between cycles. The number of cycles required to reach convergence is typically between 1 and 5.

Table 1
Comparison of binning schemes performed with d^{-3} and ln(d) spacing for three selected PDB data sets: 1kwn, 3hay and 3gk8.

All three data sets have very low completeness in the lowest resolution bin, which d^{-3} binning obscures while ln(d) binning makes clear even when using approximately half the number of bins. Completeness in the high-resolution region is similar in the two binning schemes. For each binning method three columns of data are presented: resolution range (Å), completeness and number of reflections.

	1kwn		3hay		3gk8	
Bin No.	d^{-3}	ln(d)	d^{-3}	ln(d)	d^{-3}	ln(d)
1	19.96–3.25 0.967 4363	19.96–7.87 0.860 301	44.86–13.44 0.932 715	44.86–17.61 0.852 300	22.18–5.00 0.906 1938	22.18–8.16 0.610 300
2	3.25–2.58 0.997 4280	7.87–6.33 0.971 300	13.43–10.71 1.000 716	17.58–14.23 1.000 301	5.00–3.98 0.994 2052	8.15–7.00 0.993 300
3	2.58–2.26 0.999 4214	6.33–5.10 0.966 564	10.71–9.37 1.000 688	14.22–11.51 1.000 556	3.98–3.48 0.997 2060	7.00–6.01 0.996 452
4	2.26–2.05 1.000 4218	5.10–4.10 0.961 1037	9.37–8.52 1.000 693	11.51–9.31 1.000 1011	3.48–3.16 0.995 2051	6.01–5.16 0.994 700
5	2.05–1.90 0.990 4135	4.10–3.30 0.986 1987	8.52–7.91 1.000 679	9.31–7.53 1.000 1853	3.16–2.93 0.976 1988	5.16–4.43 0.993 1087
6	1.90–1.79 0.993 4133	3.30–2.66 0.997 3772	7.91–7.45 1.000 673	7.53–6.10 1.000 3448	2.93–2.76 0.968 1973	4.43–3.81 0.996 1735
7	1.79–1.70 0.992 4119	2.66–2.14 0.999 7177	7.45–7.08 1.000 675	6.10–4.99 0.997 5905	2.76–2.62 0.958 1902	3.81–3.27 0.996 2716
8	1.70–1.63 0.989 4070	2.14–1.72 0.993 13453	7.08–6.77 1.000 657		2.62–2.51 0.952 1961	3.27–2.81 0.979 4149
9	1.63–1.57 0.988 4094	1.72–1.38 0.990 25516	6.77–6.51 1.000 672		2.51–2.41 0.954 1941	2.81–2.41 0.955 6410
10	1.57–1.51 0.990 4093	1.38–1.20 0.989 28106	6.51–6.29 1.000 671		2.41–2.33 0.941 1876	2.41–2.07 0.931 9748
11	1.51–1.46 0.987 4036		6.28–6.09 1.000 657		2.33–2.26 0.933 1897	2.07–1.85 0.827 9681
12	1.46–1.42 0.990 4073		6.09–5.92 1.000 655		2.26–2.19 0.940 1881	
13	1.42–1.39 0.993 4088		5.91–5.76 1.000 666		2.19–2.13 0.931 1876	
14	1.39–1.35 0.992 4057		5.76–5.62 1.000 656		2.13–2.08 0.914 1838	
15	1.35–1.32 0.992 4077		5.62–5.49 1.000 667		2.08–2.03 0.897 1834	
16	1.32–1.29 0.995 4052		5.49–5.38 1.000 653		2.03–1.99 0.891 1766	
17	1.29–1.27 0.991 4047		5.38–5.27 1.000 635		1.99–1.95 0.865 1765	
18	1.27–1.24 0.991 4045		5.27–5.17 1.000 663		1.95–1.92 0.825 1645	
19	1.24–1.22 0.988 4026		5.17–5.08 1.000 660		1.91–1.88 0.767 1537	
20	1.22–1.20 0.972 3993		5.08–4.99 0.973 623		1.88–1.85 0.732 1497	

To determine $k_{\rm anisotropic}$, our protocol can make use of three available scaling methods: polynomial (poly; §2.2), exponential with analytical calculation of the optimal parameters ($\exp_{\rm anal}$; §2.1) and exponential with the optimal parameters obtained *via* L-BFGS (Liu & Nocedal, 1989) minimization ($\exp_{\rm min}$; Afonine *et al.*, 2005*a*). The three methods can be tested independently, in which case the result with the lowest R factor is accepted. However, because $\exp_{\rm min}$ is up to an order of magnitude slower than the other two methods it is not expected to be used routinely.

The calculation of $k_{\rm isotropic}$ and $k_{\rm mask}$ requires dividing the data into resolution bins (§3.2). If oscillation of $k_{\rm mask}$ between bins occurs, smoothening (Savitzky & Golay, 1964) is applied to the bin-wise determined values of $k_{\rm mask}$ such that it reduces the oscillations without altering the monotonic behavior of $k_{\rm mask}$ as a function of resolution (see Fig. 1). Finally, the smoothed values are assigned to individual reflections using linear interpolation. The $k_{\rm isotropic}$ scales are updated using equation (5) in order to account for the changed $k_{\rm mask}$.

As illustrated in §3.2, the minimum of the R-factor function

$$R = \sum \left| F_{\rm obs} - |\mathbf{F}_{\rm model}| \right| / \sum |F_{\rm obs}| \qquad (46)$$

and the minimum of the least-squares function (22) can be at significantly different locations in the ($k_{\rm mask}$, $k_{\rm isotropic}$) parameter space. To assure that the final ($k_{\rm mask}$, $k_{\rm isotropic}$) values correspond to the lowest R factor, a fast grid search is performed around the optimal values of the least-squares function.

3.2. Binning

The goal of binning is to group data by common features to characterize each group by a set of common parameters. Here, the key parameter is the resolution d of reflections. Binning schemes with bins containing an approximately equal number of reflections (*i.e.* the resolution range is uniformly sampled in d^{-3}) or a predefined number of bins are typically used. Since the low-resolution region of the data is sparse, such binning

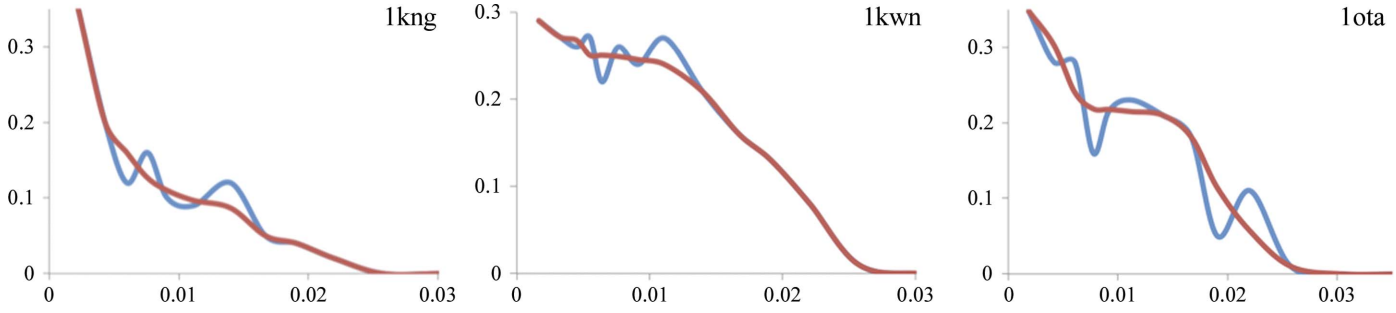


Figure 1
Examples of smoothening of $k_{\rm mask}$. The original $k_{\rm mask}$ (blue; obtained as the solution of equation 29) and that after smoothening (red) are shown for three PDB entries with the PDB codes shown on the plots.

Table 2
Comparison of $\mathbf{U}_{cryst}$ corresponding to the minima of the functions LS (3), LSL (11) and R factor (46).

To improve readability, the $\mathbf{U}_{cryst}$ are shown as B values with respect to a Cartesian basis (Grosse-Kunstleve & Adams, 2002). To reduce the runtimes for the systematic parameter searches (see text), we have selected examples with symmetry constraints leading to all-zero off-diagonal elements.

PDB code	Optimization target	B_{11}, B_{22}, B_{33}	R factor
2fih	R factor	−2.15, −1.85, −1.60	0.1935
	LS	−4.20, −3.90, −3.35	0.2179
	$\widetilde{\text{LSL}}$	−2.65, −1.95, −1.60	0.1939
2fih (data cut at 2.5 Å)	R factor	−9.30, −10.20, −10.35	0.2417
	LS	−18.35, −19.65, −20.75	0.2599
	$\widetilde{\text{LSL}}$	−38.25, −42.15, −46.20	0.3769
1ous (data cut at 6.5 Å)	R factor	9.25, −2.20, 4.35	0.2082
	LS	2.90, −2.45, 8.60	0.2086
	$\widetilde{\text{LSL}}$	19.55, 6.55, 12.85	0.2088

schemes tend to produce only one or very few low-resolution bins, which is insufficient to best model the bulk-solvent contribution. Unfortunately, decreasing the number of reflections per bin will disproportionally increase the number of bins (N_{bins}) at higher resolution and may still provide insufficient detail for the low-resolution data (Table 1).

An alternative approach which divides the resolution range uniformly on a logarithmic scale $\ln(d)$ (Urzhumtsev *et al.*, 2009) efficiently solves this problem. The flowchart of the algorithm is shown in Fig. 2. This scheme allows the higher resolution bins to contain more reflections than the lower resolution bins and more detailed binning at low resolution without increasing the total number of bins. An additional reason for using logarithmic binning is that the dependence of the scales on resolution is approximately exponential (see previous sections), which makes the variation of scale factors more uniform between bins when a logarithmic binning algorithm is used. Table 1 compares binning performed uniformly in d^{-3} and in $\ln(d)$ spacing for three data sets (PDB entries 3hay, 1kwn and 3gk8). Note the data completeness of the low-resolution bins.

3.3. Systematic tests

We evaluated the performance of the new scaling protocol by applying it to approximately 40 000 data sets selected from the PDB. The structures were selected by evaluating all PDB entries using *phenix.model_vs_data* (Afonine *et al.*, 2010) and excluding all entries for which the recalculated R_{work} was greater than the published value by five percentage points.

To score the test results three crystallographic R factors (46) were computed using all reflections, using only low-resolution reflections and using only high-resolution reflections. Low-resolution reflections were selected using the condition $d_{min} >$ 8 Å but selecting at least the 500 lowest resolution reflections. High-resolution reflections were taken from the highest resolution bin. Each of the three anisotropic scaling methods (poly, $\exp_{anal}$ and $\exp_{min}$) was tested independently within each run. Additionally, two other tests were performed: one combining poly and $\exp_{anal}$ as described in §3.1 (referred to as poly+$\exp_{anal}$) and the other using the protocol of Afonine *et al.* (2005*a*) (referred to as old).

Fig. 3 shows a comparison of the alternative methods for determining $k_{anisotropic}$ (see §3.1). Comparing the polynomial model (poly) *versus* the analytical exponential model ($\exp_{anal}$), with a few minor exceptions poly results in slightly lower R factors overall and for the low-resolution reflections, while $\exp_{anal}$ results in lower R factors for the high-resolution reflections. Comparing poly *versus* the original exponential model using minimization ($\exp_{min}$), the R factors are very similar overall and for the high-resolution reflections, while poly often results in lower R factors for the low-resolution reflections. Comparing the two different exponential models, $\exp_{min}$ results in lower R factors overall and nearly identical results for low-resolution reflections, but $\exp_{anal}$ results in lower R factors for the high-resolution reflections. Fig. 4 compares the new protocol combining poly and $\exp_{anal}$ with the old protocol. With very few exceptions, the new protocol performs better for all three resolution groups.

As described above, occasionally the minima of the R-factor function (46) and the LS function (22) are at significantly different locations in the (k_{mask}, $k_{isotropic}$) parameter space (see Fig. 5). For example, considering $k_{isotropic}$ to be a single-value

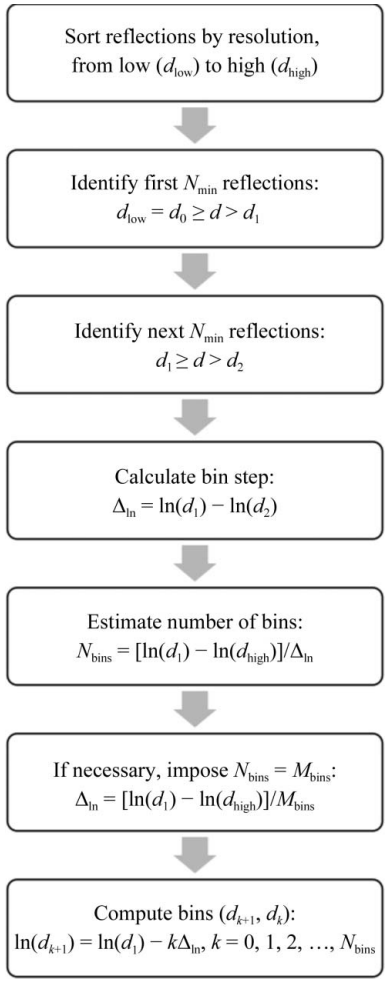


Figure 2
Flowchart of the logarithmic resolution-binning algorithm.

scalar the pair (k_{mask}, $k_{isotropic}$) that minimizes the R factor in the low-resolution range of PDB data set 1kwn is (0.2913, 0.0961), while the pair (0.3218, 0.0863) minimizes the LS function. The corresponding R factors are 0.3073 and 0.3372, respectively. The data for PDB entry 1hqw lead to an even more dramatic difference, in which the pairs (k_{mask}, $k_{isotropic}$) that minimize the R factor and the LS function are (0.25, 0.0131) and (0.6166, 0.0151), respectively, and the corresponding R factors are 0.2924 and 0.5046. We made a similar observation for the overall anisotropic scale $k_{anisotropic}$, as

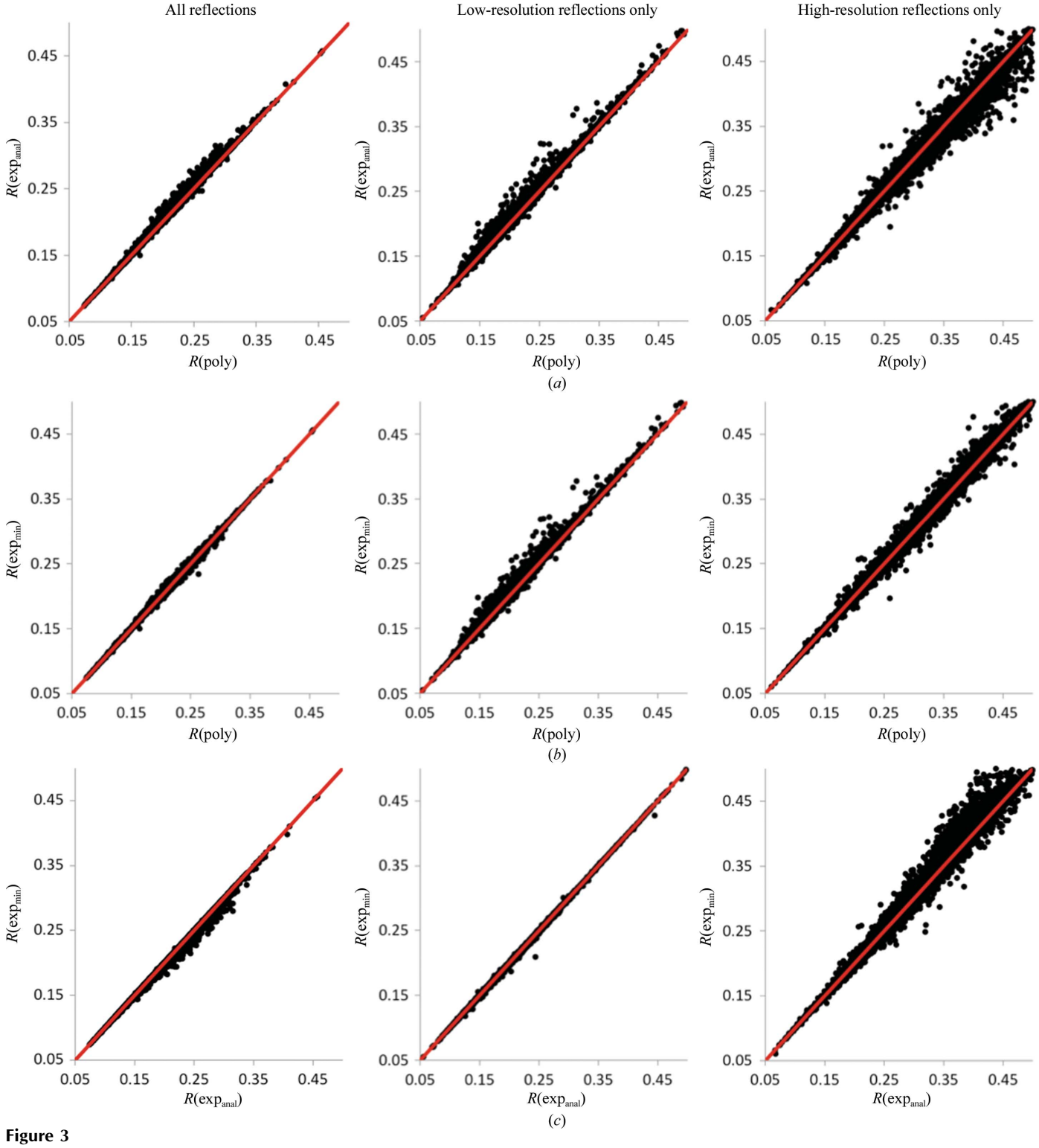


Figure 3
A comparison of the new scaling protocol using different models for the anisotropic scale factor. R *versus* R factor scatter plots for (*a*) poly *versus* $\exp_{anal}$, (*b*) poly *versus* $\exp_{min}$ and (*c*) $\exp_{anal}$ *versus* $\exp_{min}$ R factors were computed using all reflections (left), low-resolution reflections only (middle) and high-resolution reflections only (right). See §3.3 for details.

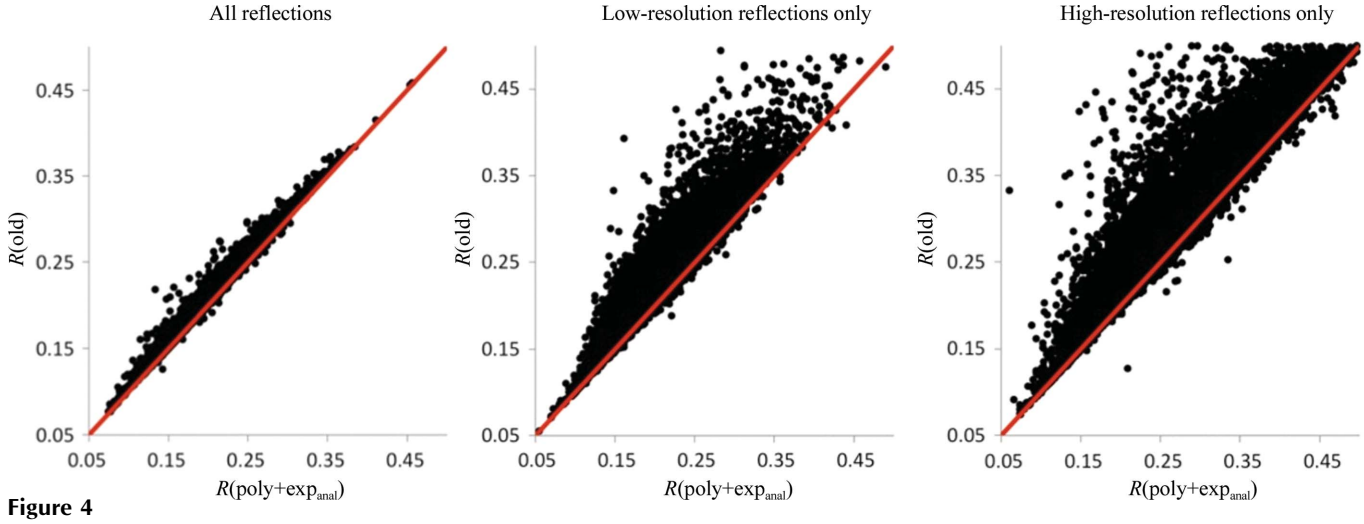


Figure 4
R versus R factor scatter plots comparing the new scaling protocol using poly+exp$_{anal}$ for the anisotropic scale factor with the old protocol. For each structure the full set of structure factors available from the PDB was used to calculate scale factors and to calculate *R* factors (left). Using the same scale-factor values the *R* factors were calculated separately for the low-resolution reflections (middle) and high-resolution reflections (right). A large spread of points in the vertical direction above the diagonal (red line) in these latter plots indicates that in many cases the scale factors produced by the old protocol resulted in a poorer fit to the data at low and high resolutions, while the new protocol generates scale factors with a good fit across all resolution ranges. See §3.3 for details.

illustrated in Table 2. For this, the best values for $\mathbf{U}_{cryst}$ were determined *via* a systematic search for the minima of the functions (3), (11) and (46) for three combinations of structures and high-resolution cutoffs. Note the difference in the optimal $\mathbf{U}_{cryst}$ values and the corresponding *R* factors.

The parameterization of the total model structure factor (1) does not make any assumption about the shape of k_{mask}; for example, it does not assume it to be exponential (10). This provides an opportunity to explore the behavior of k_{mask} as a function of resolution and compare it with k_{mask} obtained *via* (10). Fig. 6 illustrates the differences between the two methods of determining k_{mask} for six representative PDB entries selected from approximately 40 000 entries after inspection of the k_{mask} values. We observe that the plots of the values obtained using our new approach are in general significantly different from the exponential function. This observation is in line with Fig. 1 of Urzhumtsev & Podjarny (1995).

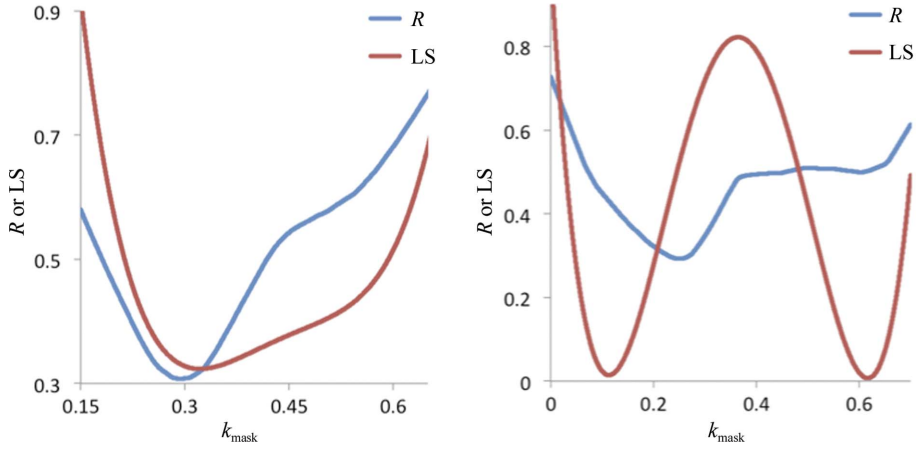


Figure 5
Plots of *R* factors (with $k_{isotropic} = 0.0961$) and the LS function (with $k_{isotropic} = 0.0863$) for PDB entry 1kwn (left) and *R* factors (with $k_{isotropic} = 0.0131$) and the LS function (with $k_{isotropic} = 0.0151$) for PDB entry 1hqw (right), illustrating that the minima of the *R*-factor function (46) and the LS function (22) can be at significantly different locations in parameter space. In such cases, a line search around the value of k_{mask} obtained by minimization of the LS function is necessary in order to obtain a value that minimizes the *R* factor. For plotting purposes, the values of the LS function were scaled to be similar to the *R* factors.

At very low resolution the structure factors computed from the atomic model are approximately anticorrelated to the structure factors computed from the bulk-solvent mask:

$$\mathbf{F}_{mask} \simeq -p\mathbf{F}_{calc}. \quad (47)$$

Here, *p* is a scale factor (Urzhumtsev & Podjarny, 1995). Relation (47) is the basis for alternative bulk-solvent scaling methods that employ the Babinet principle (Moews & Kretsinger, 1975; Tronrud, 1997). Substitution of relation (47) into equation (1) yields

$$\mathbf{F}_{model} \simeq k_{total}(1 - p\,k_{mask})\mathbf{F}_{calc}. \quad (48)$$

Obviously, $\mathbf{F}_{model}$ is invariant for any combination of scale factors k_{total} and k_{mask} satisfying the condition

$$k_{total}(1 - p\,k_{mask}) = \text{const}. \quad (49)$$

Since our new scaling procedure determines k_{mask} and $k_{isotropic}$ (which are part of k_{total}) simultaneously, without imposing constraints on their values, these scale factors may assume unusual values in the low-resolution range. However, we

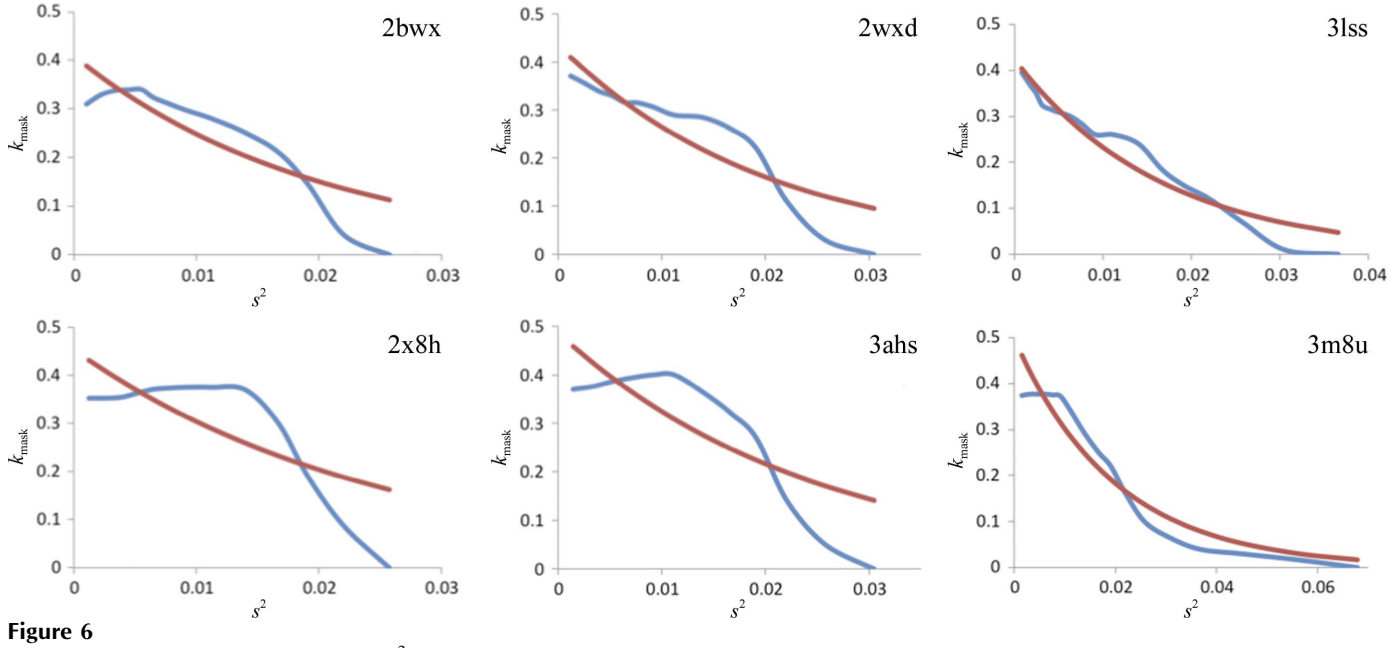


Figure 6
Plots of k_{mask} as a function of resolution (s^2) for six selected PDB entries. The blue lines show k_{mask} as determined using the new method. The red lines show k_{mask} based on the exponential function (10) using optimized k_{sol} and B_{sol} parameters.

Table 3
Runtime comparison for selected PDB entries.

Absolute runtimes for the new protocol range from a few hundredths of a second to a second.

PDB code	Resolution (Å)	No. of atoms	No. of reflections	Speed gain
1us0	0.66	3679	511265	105
1akg	1.10	136	4471	132
1ous	1.20	3784	104889	86
1yjp	1.80	66	495	64
1f8t	1.95	3593	28288	104
1av1	4.00	6588	16201	110
1jl4	3.99	4474	7428	78
2i07	4.0	12157	20412	126
2gsz	4.2	16344	17131	166

observe that in practice this only happens for a very small number of the test cases.

4. Discussion

A new method for overall anisotropic and bulk-solvent scaling of macromolecular crystallographic diffraction data has been developed which is an improvement over the existing algorithm of flat (mask-based) bulk-solvent modeling and overall anisotropic scaling, versions of which are routinely used in various refinement packages such as *CNS* (Brunger, 2007), *REFMAC* (Murshudov *et al.*, 2011) and *phenix.refine* (Afonine *et al.*, 2012). In the process of developing this method, we concluded that the bulk-solvent scale factor k_{mask} deviates quite significantly from the exponential model that has traditionally been used. This new method is approximately two orders of magnitude faster than the previous implementation and yields similar or often better R factors. Table 3 compares runtimes for a number of selected cases covering a broad range of resolutions and atomic model sizes. Therefore, the computational speed of the new method makes it possible to robustly compute bulk-solvent and anisotropic scaling parameters even as part of semi-interactive procedures.

An inherent feature of the mask-based bulk-solvent model is that it relies on the existing atomic model to compute the mask. This in turn implies that any unmodeled (as atoms) parts of the unit cell are considered to belong to the bulk-solvent region. This may obscure weakly pronounced features in residual maps such as partially occupied solvent or ligands. This is common to all mask-based bulk-solvent modeling methods, leading to the development of algorithms to account for missing atoms (Roversi *et al.*, 2000). In the future, improved maps may be obtained by combining this latter approach with the new fast overall anisotropic and bulk-solvent scaling method that we have presented.

The new method is implemented in the *cctbx* project (Grosse-Kunstleve *et al.*, 2002) and is used in a number of *PHENIX* applications since v.1.8 of the software, most notably *phenix.refine* (Afonine *et al.*, 2005*b*, 2012), *phenix.maps* and *phenix.model_vs_data* (Afonine *et al.*, 2010). The *cctbx* project is available at http://cctbx.sourceforge.net under an open-source license. The *PHENIX* software is available at http://www.phenix-online.org.

All results presented are based on *PHENIX* v.1.8.1.

APPENDIX *A*
Analytical derivation of a one-Gaussian approximation of a one-dimensional discrete data set

Our goal is to approximate a set of data points $\{Y(x)\}_j^N = 1$ with a Gaussian function,

$$a\exp(-bx^2). \tag{50}$$

For this, we use the standard approach of minimizing a least-squares (LS) function,

$$\mathrm{LS} = \sum_{j=1}^{N} [Y(x_j) - a\exp(-bx_j^2)]^2. \quad (51)$$

If $Y(x_j) \geq 0 \ \forall \ x_j$, $j = 1, N$, the minimization of LS can be replaced by the minimization of

$$\mathrm{LSL} = \sum_{j=1}^{N} \{\ln[Y(x_j)] - \ln[a\exp(-bx_j^2)]\}^2. \quad (52)$$

The minimum of this LSL function can be determined analytically,

$$\begin{aligned}\mathrm{LSL} &= \sum_{j=1}^{N} \{[\ln(Y(x_j)] - \ln[a\exp(-bx_j^2)]\}^2 \\ &= \sum_{j=1}^{N} \{\ln(a) - bx_j^2 - \ln[Y(x_j)]\}^2. \end{aligned} \quad (53)$$

Defining $u = \ln(a)$, $v_j = x_j^2$, $d_j = \ln[Y(x_j)]$, we obtain

$$\mathrm{LSL} = \sum_{j=1}^{N} (u - bv_j - d_j)^2. \quad (54)$$

The variables $\{a, b\}$ minimizing the LSL function are determined by the condition

$$\begin{cases} \dfrac{\partial \mathrm{LSL}}{\partial u} = 0 \\ \dfrac{\partial \mathrm{LSL}}{\partial b} = 0. \end{cases} \quad (55)$$

This leads to

$$\begin{cases} -2\sum_{j=1}^{N} (u - bv_j - d_j) = 0 \\ -2\sum_{j=1}^{N} (u - bv_j - d_j)v_j = 0 \end{cases} \quad (56)$$

and

$$\begin{cases} uN - b\sum_{j=1}^{N} v_j - \sum_{j=1}^{N} d_j = 0 \\ u\sum_{j=1}^{N} v_j - b\sum_{j=1}^{N} v_j^2 - \sum_{j=1}^{N} v_j d_j = 0. \end{cases} \quad (57)$$

Defining $p = \sum_{j=1}^{N} d_j$, $q = \sum_{j=1}^{N} v_j$, $r = \sum_{j=1}^{N} v_j^2$ and $s = \sum_{j=1}^{N} v_j d_j$, we obtain

$$\begin{cases} uN - bq - p = 0 \\ uq - br - s = 0 \end{cases} \quad (58)$$

and

$$\begin{cases} u = \dfrac{1}{N}(bq + p) \\ b = \dfrac{1}{r}(uq - s). \end{cases} \quad (59)$$

From this, we obtain

$$u = \frac{p - \dfrac{sq}{r}}{N - \dfrac{q^2}{r}}, \quad b = \frac{1}{r}(uq - s) \quad (60)$$

and finally

$$a = \exp(u), \quad b = \frac{1}{r}(uq - s). \quad (61)$$

The authors thank the NIH (grant GM063210) and the PHENIX Industrial Consortium for support of the *PHENIX* project. This work was supported in part by the US Department of Energy under Contract No. DE-AC02-05CH11231.

References

Adams, P. D. *et al.* (2010). *Acta Cryst.* D**66**, 213–221.
Afonine, P. V., Grosse-Kunstleve, R. W. & Adams, P. D. (2005*a*). *Acta Cryst.* D**61**, 850–855.
Afonine, P. V., Grosse-Kunstleve, R. W. & Adams, P. D. (2005*b*). *CCP4 Newsl. Protein Crystallogr.* **42**, contribution 8.
Afonine, P. V., Grosse-Kunstleve, R. W., Chen, V. B., Headd, J. J., Moriarty, N. W., Richardson, J. S., Richardson, D. C., Urzhumtsev, A., Zwart, P. H. & Adams, P. D. (2010). *J. Appl. Cryst.* **43**, 669–676.
Afonine, P. V., Grosse-Kunstleve, R. W., Echols, N., Headd, J. J., Moriarty, N. W., Mustyakimov, M., Terwilliger, T. C., Urzhumtsev, A., Zwart, P. H. & Adams, P. D. (2012). *Acta Cryst.* D**68**, 352–367.
Badger, J. (1997). *Methods Enzymol.* **277**, 344–352.
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.
Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.
Brunger, A. T. (2007). *Nature Protoc.* **2**, 2728–2733.
Fenn, T. D., Schnieders, M. J. & Brunger, A. T. (2010). *Acta Cryst.* D**66**, 1024–1031.
Fokine, A. & Urzhumtsev, A. (2002*a*). *Acta Cryst.* A**58**, 72–74.
Fokine, A. & Urzhumtsev, A. (2002*b*). *Acta Cryst.* D**58**, 1387–1392.
Giacovazzo, C. (1992). *Fundamentals of Crystallography.* Oxford University Press.
Grosse-Kunstleve, R. W. & Adams, P. D. (2002). *J. Appl. Cryst.* **35**, 477–480.
Grosse-Kunstleve, R. W., Sauter, N. K., Moriarty, N. W. & Adams, P. D. (2002). *J. Appl. Cryst.* **35**, 126–136.
Grosse-Kunstleve, R. W., Wong, B., Mustyakimov, M. & Adams, P. D. (2011). *Acta Cryst.* A**67**, 269–275.
Jiang, J. S. & Brünger, A. T. (1994). *J. Mol. Biol.* **243**, 100–115.
Kostrewa, D. (1997). *CCP4 Newsl. Protein Crystallogr.* **34**, 9–22.
Liu, D. C. & Nocedal, J. (1989). *Math. Program.* **45**, 503–528.
Moews, P. C. & Kretsinger, R. H. (1975). *J. Mol. Biol.* **91**, 201–225.
Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* D**67**, 355–367.
Nye, J. F. (1957). *Physical Properties of Crystals.* Oxford: Clarendon Press.
Phillips, S. E. (1980). *J. Mol. Biol.* **142**, 531–554.
Roversi, P., Blanc, E., Vonrhein, C., Evans, G. & Bricogne, G. (2000). *Acta Cryst.* D**56**, 1316–1323.
Savitzky, A. & Golay, M. J. E. (1964). *Anal. Chem.* **36**, 1627–1639.
Shakked, Z. (1983). *Acta Cryst.* A**39**, 278–279.
Sheriff, S. & Hendrickson, W. A. (1987). *Acta Cryst.* A**43**, 118–121.
Tronrud, D. E. (1997). *Methods Enzymol.* **277**, 306–319.
Urzhumtsev, A. G. (2000). *CCP4 Newsl. Protein Crystallogr.* **38**, 38–49.
Urzhumtsev, A., Afonine, P. V. & Adams, P. D. (2009). *Acta Cryst.* D**65**, 1283–1291.
Urzhumtsev, A. G. & Podjarny, A. D. (1995). *Jnt CCP4/ESF–EACMB Newsl. Protein Crystallogr.* **31**, 12–16.
Usón, I., Pohl, E., Schneider, T. R., Dauter, Z., Schmidt, A., Fritz, H. J. & Sheldrick, G. M. (1999). *Acta Cryst.* D**55**, 1158–1167.
Vassylyev, D. G., Vassylyeva, M. N., Perederina, A., Tahirov, T. H. & Artsimovitch, I. (2007). *Nature (London)*, **448**, 157–162.