electronic reprint

Journal of
Applied Crystallography
ISSN 0021-8898
Editor: **Gernot Kostorz**

Improved statistics for determining the Patterson symmetry from unmerged diffraction intensities

Nicholas K. Sauter, Ralf W. Grosse-Kunstleve and Paul D. Adams

Journal of
Applied Crystallography
ISSN 0021-8898

Received 6 August 2005
Accepted 16 December 2005

Improved statistics for determining the Patterson symmetry from unmerged diffraction intensities

Nicholas K. Sauter,* Ralf W. Grosse-Kunstleve and Paul D. Adams

Physical Biosciences Division, Lawrence Berkeley National Laboratory, One Cyclotron Road, Bldg 64R0121, Berkeley, CA 94720, USA. Correspondence e-mail: nksauter@lbl.gov

Procedures for detecting the point-group symmetry of macromolecular data sets are examined and enhancements are proposed. To validate a point group, it is sufficient to compare pairs of Bragg reflections that are related by each of the group's component symmetry operators. Correlation is commonly expressed in the form of a single statistical quantity (such as $R_{\rm merge}$) that incorporates information from all of the observed reflections. However, the usual practice of weighting all pairs of symmetry-related intensities equally can obscure the fact that the various symmetry operators of the point group contribute differing fractions of the total set. In some cases where particular symmetry elements are significantly under-represented, statistics calculated globally over all observations do not permit conclusions about the point group and Patterson symmetry. The problem can be avoided by repartitioning the data in a way that explicitly takes note of individual operators. The new analysis methods, incorporated into the program *LABELIT* (http://cci.lbl.gov/labelit), can be performed early enough during data acquisition, and are quick enough that it is feasible to pause to optimize the data collection strategy.

1. Introduction

Knowledge of the crystal symmetry is of fundamental importance to macromolecular crystallography. Experience has shown that even when it is possible to solve a structure under the wrong symmetry, the resulting atomic model can have subtle errors leading to incorrect biological conclusions (Kleywegt *et al.*, 1996). Here we focus on methods for identifying the Patterson symmetry from raw diffraction data [for definitions see §2 of the *International Tables for Crystallography;* (Hahn, 1996; referred to as IT96 hereafter)]. There are two components of this analysis. First, if the data are collected with the usual rotation method (Arndt *et al.*, 1973), a single oscillation image can reveal the metric symmetry of the unit-cell dimensions, permitting the crystal to be classified into one of 14 Bravais types. Secondly, once a complete or partial data set is acquired, the point-group symmetry of the reciprocal lattice (ignoring the effects of anomalous dispersion) produces an assignment into one of 11 Laue classes. The correct combination of Bravais type and Laue class, giving one of 24 Patterson symmetries, must be known before performing any subsequent steps, such as merging of symmetry-related measurements, detection of screw axes and phasing.

Correct identification of the Laue class is also crucial for calculating the optimal data collection strategy (Ravelli *et al.*, 1997; Dauter, 1999; Popov & Bourenkov, 2003) after a crystal has been transferred to the goniostat and the diffraction pattern has been indexed. In many cases the Laue class is known from previous studies of similar crystal samples. However, in other contexts, such as high-throughput crystal screening, the exact point group may not be known ahead of time. For this reason, it is desirable to have a quick robust method for analyzing the symmetry soon after data acquisition has begun. The data collection strategy can then be optimized before the remainder of the data set is collected.

There are several practical matters to consider if one is to determine the Patterson symmetry automatically and accurately. For low-symmetry crystal families (triclinic, monoclinic and orthorhombic) it is sometimes possible to learn the Patterson symmetry from a single oscillation image, since the point group follows immediately from the metric symmetry. However, with higher-symmetry families (tetragonal, trigonal, hexagonal and cubic), there is more than one possible Laue class for each Bravais lattice (*e.g.* a tetragonal cell can have either $4/m\,m\,m$ or $4/m$ point symmetry). The crystal must therefore be rotated through an angular range sufficient for the point-group symmetry to be tested. Furthermore, it is well known that experimental uncertainty in the unit-cell dimensions can preclude drawing conclusions about the Bravais lattice. It is common to find monoclinic lattices with β angles near 90°, which therefore appear to be orthorhombic; or orthorhombic crystals with two cell lengths nearly equal (Fig. 1), which therefore appear to be tetragonal. In these cases, again, one must collect enough data to determine the point-group symmetry before making a final conclusion about the Bravais lattice.

The issue of data completeness becomes central when an attempt is made to use an automated procedure to identify the

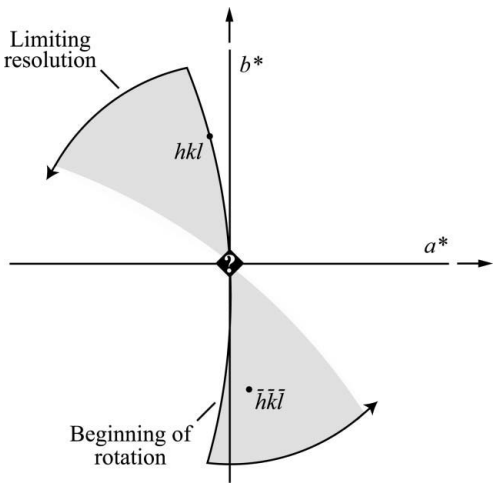

Figure 1
Ewald diagram (see Dauter, 1999, for a detailed explanation) showing a case where a global measure of data quality like $R_{\rm merge}$ cannot by itself establish the point-group symmetry. The diagram depicts a slice through the reciprocal-space a^*b^* plane for an orthorhombic crystal where $a \simeq b$ within the limits of experimental uncertainty. Diffraction data are collected from the shaded area of reciprocal space by rotating the crystal about the c^* axis, perpendicular to the incident beam. For total rotation angles that produce incomplete data sets (Table 1), there are relatively few reflection pairs available to test whether the c^* axis (indicated by a '?') is a fourfold rotation axis. If global statistics are calculated under various point groups, statistics for the incorrect group $4/m$ (containing a fourfold along c^*) are experimentally indistinguishable from statistics for the correct group mmm (containing twofolds along a^*, b^* and c^*). The particular rotation range shown was chosen to illustrate best how the under-representation of symmetry operators can interfere with symmetry identification.

point-group symmetry. A challenging example is illustrated in Fig. 1, where the point group of an apparently tetragonal crystal is in question. The algorithm must distinguish among the tetragonal Laue classes $4/mmm$ or $4/m$, as well as the lower-symmetry classes mmm, $2/m$ and $\bar{1}$. Fundamentally, point groups must be proven by observing correlated intensities for symmetry-related reflections. In common practice, data are merged under all possible Laue classes, and the correlation between related reflections is statistically quantified. A straightforward measure is the merging reliability (Weiss, 2001),

$$R_{\rm merge} = \frac{\sum_{hkl} \sum_{i=1}^{N_{hkl}} \left| I_i(hkl) - \langle I(hkl) \rangle \right|}{\sum_{hkl} \sum_i I_i(hkl)}, \qquad (1a)$$

where $I(hkl)$ is the intensity of a Bragg reflection, $\sum_{hkl}$ is a sum over all Miller indices in the merged data set, and $\sum_i$ is a sum over the N_{hkl} symmetry-related measurements for a given Miller index. One typically chooses the highest-symmetry Laue class with a reasonable $R_{\rm merge}$. However, for the incomplete data set illustrated in Fig. 1, $R_{\rm merge}$ cannot be directly used to determine if there is fourfold rotational symmetry along c, and therefore is problematic for automatically evaluating the $4/m$ Laue class. This can be made explicit by rewriting equation (1a) as

Table 1
Representation of fourfold symmetry in $R_{\rm merge}$ statistics resulting from the rotation geometry in Fig. 1.

Angular rotation (°)	Data set completeness to 3 Å resolution (%)†	Fraction of $(I_i - I_j)$ terms contributed by the fourfold (%)‡
15	27.8	0.7
30	47.0	1.1
45	63.5	1.6
60	79.3	2.7
75	94.2	6.6
90	98.4	15.2
105	98.4	32.7
120	98.4	43.3

† Assuming that the true Patterson symmetry is $Pmmm$, with unit cell a = 91.819, b = 92.358, c = 119.347 Å. Completeness never reaches 100% with the rotation depicted in Fig. 1 due to missing reflections along the c^*-axis spindle. ‡ In the context of an $R_{\rm merge}$ statistic [equation (1b)] calculated under $P4/m$ symmetry. Terms with $i = j$ are not counted.

$$R_{\rm merge} = \frac{\sum_{hkl} \sum_{i=1}^{N_{hkl}} \left| (1/N_{hkl}) \sum_{j=1}^{N_{hkl}} [I_i(hkl) - I_j(hkl)] \right|}{\sum_{hkl} \sum_i I_i(hkl)}. \qquad (1b)$$

If $R_{\rm merge}$, calculated under $4/m$ symmetry, is to be sensitive to the presence of a fourfold, then a significant fraction of the $[I_i(hkl) - I_j(hkl)]$ terms in equation (1b) must represent pairs of measured reflections I_i, I_j related by a fourfold rotation. But as shown in Table 1, there are vanishingly few such pairs when the angular rotation is small. Even as the data set nears completion (*e.g.* after 75° of angular rotation), fourfold symmetry is still poorly represented.

A conclusion from Table 1 is that globally defined statistics (those calculated over all measured reflections) cannot be relied upon universally to score candidate Laue classes. If the goal is to determine the point group during data collection, after completing only a small fraction of the total angular rotation, then individual symmetry operators may be drastically under-represented within the data set. This basic result is unchanged even if a different global indicator of data quality is used (see Weiss, 2001, for a general discussion of global indicators) such as the redundancy-independent $R_{\rm merge}$ or the $I/\sigma(I)$ ratio.

To overcome the limitations of global data quality indicators, we propose an algorithm that partitions the measurements so that individual symmetry operators can be assessed. Operators found to be valid are then re-grouped to give the correct point group. If particular operators are missing from the data set, this is duly noted. Our approach is analogous to that used in past years in which reflections on a single-layer precession photograph (Blundell & Johnson, 1976) were compared to distinguish, for example, between a twofold and a fourfold symmetry axis. Importantly, we provide a software framework within the *LABELIT* package (*Lawrence Berkeley Laboratory Indexing Toolbox*) where all possible Laue classes are efficiently analyzed by a single command, so it is not necessary for the user to submit multiple processing jobs to compare point groups.

2. Notation

Scalar quantities are represented in italic type. Vectors are denoted in lowercase bold, and matrices and tensors are written in uppercase bold. The superscript T represents the transpose of a vector or matrix.

The components of row vector **h** are given as $[hkl]$, while the corresponding column vector is written as $[hkl]^{\mathrm{T}}$. A unit-length vector is denoted $\hat{\mathbf{h}}$. Matrices are sometimes expressed as a sequence of column vector components, $\mathbf{A}^* = [\mathbf{a}^*, \mathbf{b}^*, \mathbf{c}^*]$, or written out fully as scalar components in brackets.

Sets are represented by non-serified symbols, with Z specifically meaning the set of integers, M a supergroup, and H a subgroup. Elements of a set are enumerated as $\{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$, while the form $\{hkl\}$ indicates a set of lattice planes.

The (**W, w**) formalism for symmetry operations from Fischer & Koch (1996) consists of a (3×3) matrix denoting the rotation part **W**, and a (3×1) translation part **w**. The rotation operator **W** can also be shown in Jones–Faithful notation (*e.g.* x, y, z).

The notation ${}^{[e_xe_ye_z]}N$ or ${}^{[e_xe_ye_z]}N^{-1}$ is used to describe a rotation operator of type $N \in \{1, 2, 3, 4, 6, \bar{1}, \bar{2}, \bar{3}, \bar{4}, \bar{6}\}$ with axis direction $[e_xe_ye_z]$. The superscript -1 is used for operators with a negative sense of rotation.

3. Computational approach

To determine the point-group symmetry accurately, we separately consider each rotational symmetry operator **W** (Fischer & Koch, 1996) that is consistent with the unit-cell parameters. A list is made of pairs of measured reflections related by **W** or its rotoinversion $\overline{\mathbf{W}}$, and the symmetry-operator reliability for **W** is defined as

$$R_{\mathrm{symop}}(\mathbf{W}) = \frac{\sum_{\mathrm{pairs}} \sum_i \left| I_i - \langle I \rangle \right|}{\sum_{\mathrm{pairs}} \sum_i I_i}, \qquad (2)$$

where $\sum_{\mathrm{pairs}}$ is a sum over the N_{W} pairs of Bragg spots related by **W** or $\overline{\mathbf{W}}$, and $\sum_i$ is a sum over both measurements in the pair. The definition of R_{symop} is intended to be analogous to that of R_{merge}, with the advantage that R_{symop} explicitly isolates the correlation of reflections related by the individual operator **W**, even if N_{W} is a relatively small number. For a candidate point group to be accepted, the maximum $R_{\mathrm{symop}}(\mathbf{W})$ value over all symmetry operators in the group must be reasonably low.

The procedures that lead to the R_{symop} calculation are outlined in Fig. 2, and are described below. The first part of the analysis rests upon measuring the positions of Bragg reflections, which are used to determine the unit cell; while the final section requires that reflection intensities be integrated and scaled.

3.1. Determining the metric symmetry

3.1.1. Data collection and indexing. The analysis of metric symmetry (the symmetry of the lattice, without regard for the physical contents of the unit cell) begins by identifying the repeating pattern in the diffraction image. This initial step of indexing (Steller *et al.*, 1997) can usually be performed with a single oscillation image. However, to determine the incident-beam position (Sauter *et al.*, 2004) and unit-cell parameters most accurately, it is often advantageous to utilize two images collected at rotational settings differing by 90° (where the rotational axis is perpendicular to the incident X-ray beam). It is best to obtain these two images before any other data are collected, so the metric analysis can be performed on data that are relatively unaffected by radiation damage. Starting with the positions of the Bragg reflections, one derives a set of three basis vectors that exactly cover the reciprocal lattice. Such vectors are said to be a primitive basis. Although vectors spanning the lattice can be chosen in an infinite number of ways, we employ a cell-reduction procedure (§6 of Grosse-Kunstleve *et al.*, 2004*a*) to assure that a standard set is selected with minimal vector lengths. These reduced vectors $\mathbf{A}^* = [\mathbf{a}^*, \mathbf{b}^*, \mathbf{c}^*]$ describe the unit-cell edges of the primitive lattice in reciprocal space. For relating this basis to the orthonormal $\hat{\mathbf{x}}\hat{\mathbf{y}}\hat{\mathbf{z}}$ laboratory coordinate system, it is equivalent to express the basis as an orientation matrix,

$$\mathbf{A}^* = \begin{pmatrix} a_x^* & b_x^* & c_x^* \\ a_y^* & b_y^* & c_y^* \\ a_z^* & b_z^* & c_z^* \end{pmatrix}, \qquad (3)$$

where the subscripts x, y and z refer to projections onto the laboratory axes at a crystal rotational setting $\varphi = 0°$. The direct-space basis $\mathbf{A} = [\mathbf{a}, \mathbf{b}, \mathbf{c}]^{\mathrm{T}}$ is obtained by inverting the reciprocal-space orientation matrix,

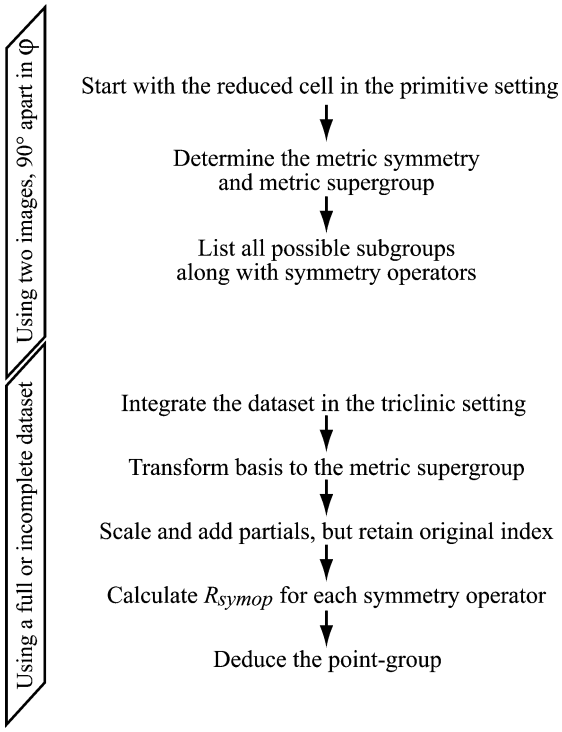

Figure 2
Computational steps for determining the Patterson symmetry. The crystal rotation axis φ is perpendicular to the incident beam.

$$\mathbf{A} = (\mathbf{A}^*)^{-1} = \begin{pmatrix} a_x & a_y & a_z \\ b_x & b_y & b_z \\ c_x & c_y & c_z \end{pmatrix}. \quad (4)$$

3.1.2. Location of rotational symmetry axes. The next step is to locate all of the twofold rotational symmetry axes of the lattice. To cast this problem in algebraic terms, points on the lattice are expressed as integer linear combinations of the basis vectors: direct-space lattice vectors are $\mathbf{t} = u_a\mathbf{a} + u_b\mathbf{b} + u_c\mathbf{c}$ (with $u_a, u_b, u_c \in \mathsf{Z}$) and reciprocal-space vectors are $\boldsymbol{\tau} = h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*$ (with Miller indices $h, k, l \in \mathsf{Z}$). Expressed in the laboratory frame, we have $\mathbf{t} = \mathbf{A}^{\mathrm{T}}\mathbf{u}$, with $\mathbf{u} = [u_a u_b u_c]^{\mathrm{T}}$; and $\boldsymbol{\tau} = (\mathbf{A}^*)\mathbf{h}$, with $\mathbf{h} = [hkl]^{\mathrm{T}}$. An algorithm given by Le Page (1982) allows us to select all combinations of $\mathbf{t}$ and $\boldsymbol{\tau}$ that line up to form twofolds. A compelling advantage of Le Page's approach is its ability to accommodate experimental uncertainty (Le Page, 1982; Grosse-Kunstleve *et al.*, 2004*a*; Sauter *et al.*, 2004). Because of measurement imperfections, $\mathbf{t}$ and $\boldsymbol{\tau}$ are unlikely to line up exactly on a symmetry axis, but the (small) angle between them,

$$\delta = \cos^{-1}\frac{|\mathbf{t}\cdot\boldsymbol{\tau}|}{|\mathbf{t}|\,|\boldsymbol{\tau}|}, \quad (5)$$

offers a convenient way for judging a candidate axis. Experience indexing several hundred data sets has led to the recommended cutoff of $\delta = 1.4^\circ$, used in *LABELIT*. If an axis has a larger δ value, it is certain not to be a twofold symmetry axis, while a lower value indicates that it is a plausible twofold candidate, subject to later verification when the point group is determined. It is instructive to consider the example of Fig. 1, taking the experimentally measured reduced-cell constants to be

$$\begin{aligned} &a = 91.80, b = 92.36, c = 119.37\ \text{Å}, \\ &\alpha = 89.996, \beta = 89.903, \gamma = 89.772^\circ. \end{aligned} \quad (6)$$

The resulting list of candidate twofold axes (Table 2) is sufficient to construct either an orthorhombic or a tetragonal cell, illustrating how the method can accommodate minor errors in the measured cell dimensions.

3.1.3. Expression of the rotation in matrix operator form. Lattice symmetry is completely determined by the twofold axes (Le Page, 1982). For example, the apparent fourfold axis along the [001] direction in Fig. 1 can be derived by combining the twofolds along [100] and [110]. This can be shown geometrically by considering space-group diagrams as in IT96, but for computational purposes it is most useful to combine symmetry operations through algebraic manipulation. A prerequisite for this is to rewrite the rotation in the form of a matrix operator $\mathbf{W}$ (Fischer & Koch, 1996). In laboratory (Cartesian) space, a twofold rotation about $\mathbf{t}$ is expressed by the operator

Table 2
Candidate twofold rotational symmetry axes from metric analysis.

Axis No.	δ (°)	**u**	**W**
1	0.097	[001]	$-x, -y, z$
2	0.228	[010]	$-x, y, -z$
3	0.248	[$\bar{1}$00]	$x, -y, -z$
4	0.355	[$\bar{1}$10]	$-y, -x, -z$
5	0.356	[110]	$y, x, -z$

Table 3
Symmetry operators in space group $P4/mmm$.

Operator pair†	Non-centrosymmetric operator	Centrosymmetric operator
a	1	$\bar{1}$
b	$[100]2$	$[100]\bar{2}$
c	$[010]2$	$[010]\bar{2}$
d	$[001]2$	$[001]\bar{2}$
e	$[110]2$	$[110]\bar{2}$
f	$[\bar{1}10]2$	$[\bar{1}10]\bar{2}$
g	$[001]4^1$	$[001]\bar{4}^1$
h	$[001]4^{-1}$	$[001]\bar{4}^{-1}$

† The letter symbols used here for identifying the operator pairs are not part of a standard nomenclature; they are simply intended to permit comparisons between Tables 3, 4, 7 and 8.

$$\mathbf{W}_{\mathrm{LAB}} = \begin{pmatrix} 2\hat{t}_x^2 - 1 & 2\hat{t}_x\hat{t}_y & 2\hat{t}_x\hat{t}_z \\ 2\hat{t}_x\hat{t}_y & 2\hat{t}_y^2 - 1 & 2\hat{t}_y\hat{t}_z \\ 2\hat{t}_x\hat{t}_z & 2\hat{t}_y\hat{t}_z & 2\hat{t}_z^2 - 1 \end{pmatrix}, \quad (7)$$

where we use components of the normalized vector $\hat{\mathbf{t}} = \mathbf{t}/(\mathbf{t}\cdot\mathbf{t})^{1/2}$ (Goldstein, 1980). The laboratory-frame matrix is converted to the crystallographic basis $\mathbf{A}$ with the transformation $\mathbf{W} = (\mathbf{A}^{\mathrm{T}})^{-1}\mathbf{W}_{\mathrm{LAB}}(\mathbf{A}^{\mathrm{T}})$, and the resulting matrix elements are rounded to the nearest integer. Rounding is necessary because of the non-zero value of angle δ, and because of the finite-precision representation of the matrix elements of $\mathbf{A}$. An example operator is the twofold along the $\mathbf{c}$ axis,

$$\mathbf{W} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \quad (8)$$

The corresponding Jones–Faithful expression $(-x, -y, z)$, is listed in Table 2.

3.1.4. The metric supergroup. With the twofold symmetry operators $\mathbf{W}$ constructed, the full symmetry group of the lattice can be derived by group multiplication (§3 of Grosse-Kunstleve, 1999). Continuing with the example of Table 2, the five listed rotations combine with the identity and inversion operators (1 and $\bar{1}$) to produce space-group $P4/mmm$, with symmetry operations listed in Table 3. This group represents the highest possible Patterson symmetry consistent with the measured unit cell, and indeed all possible Patterson symmetries are subgroups. The set is therefore denoted as the metric supergroup M. One metric supergroup corresponds to each Bravais type.

Table 4
Centrosymmetric subgroups of $P4/mmm$.

Conventional setting	Rotational part **C** of the change-of-basis matrix	Symmetry in the original, reduced basis†	Operator pairs of $P4/mmm$ (Table 3) present in the subgroup
$P4/mmm$	x, y, z	$P4/mmm$	*a b c d e f g h*
$P4/m$	x, y, z	$P4/m$	*a d g h*
$Cmmm$	$x/2-y/2, x/2+y/2, z$	Hall: $-C\,2\,2\,(x-y, x+y, z)$	*a d e f*
$C12/m1$	$x/2-y/2, x/2+y/2, z$	Hall: $-C\,2y\,(x+y, -x+y, z)$	*a e*
$C12/m1$	$x/2+y/2, -x/2+y/2, z$	Hall: $-C\,2y\,(x-y, x+y, z)$	*a f*
$Pmmm$	x, y, z	$Pmmm$	*a b c d*
$P12/m1$	$-y, -x, -z$	$P2/m\,1\,1$	*a b*
$P12/m1$	x, y, z	$P12/m1$	*a c*
$P\underline{1}2/m1$	$-x, -z, -y$	$P\underline{1}\,1\,2/m$	*a d*
$P\bar{1}$	x, y, z	$P\bar{1}$	*a*

† Hermann–Mauguin space-group symbols are used if possible; otherwise a Hall symbol is used (Hall, 1981; Hall & Grosse-Kunstleve, 2001).

3.2. Subgroup algebra needed for point-group determination

3.2.1. Enumeration of the Patterson subgroups. Since the metric supergroup is based merely on the unit-cell measurements, it is realised ahead of time that the symmetry must be validated by examining the Bragg spot intensities. In some cases (including the example of Fig. 1) some symmetry operations of the metric supergroup will be disproved and will have to be discarded. Anticipating this possibility, the next task is to list all possible ways that subsets of M can form a space group. In principle, a subset of M is a group if it is closed under the operation of multiplication. It would be computationally expensive to identify all subgroups by brute force: in the example of Table 3, there are 65535 non-empty subsets of M. However, the problem is greatly simplified by mathematical theorems asserting that at most three symmetry operators are required to generate any space group (for example, Boisen & Gibbs, 1990), if the operations are referred to a primitive basis. Furthermore, for the present purpose of examining the point symmetry, anomalous dispersion will be ignored, so the reciprocal-space pattern will be centrosymmetric. This means that one of the three space-group generators can be assumed to be the inversion operator. Our algorithm to generate the subgroups of interest therefore consists of adding $\bar{1}$ to all possible subsets of M containing two noncentrosymmetric operators (only 28 in the example of Table 3). These sets are expanded algebraically by group multiplication (§3 of Grosse-Kunstleve, 1999; §3 of Grosse-Kunstleve *et al.*, 2004*b*) to derive full subgroups, and duplicate subgroups are removed. Table 4 lists the resulting ten centrosymmetric subgroups for the example.

Producing the list of subgroups in this manner is extremely valuable in organizing the search for symmetry. Among the useful features are: (*a*) it is made clear (Table 3) which pairs of symmetry operators need to be validated, and how these operators combine to form groups (Table 4); (*b*) all the relevant Laue classes are automatically generated in cases where several point groups are possible for the crystal family (*e.g.* $4/m\,m\,m$ and $4/m$ for the tetragonal family); and (*c*) the listing makes clear that subgroups can have the same space-group type (*e.g.* $P12/m1$) while possessing symmetry operators that are oriented differently with respect to the unit cell (twofold axes along [100], [010] or [001]). Since commonly used software does not make this information explicit at the time of indexing, it may be surprising that certain lattice types have as many as 30 centrosymmetric subgroups that need to be considered when assigning the symmetry (Table 5). However, we must emphasize that this subgroup listing is required if the goal is to be sure that the true symmetry is not overlooked.

Table 5
Number of centrosymmetric subgroups.

Bravais lattice†	Metric supergroup	Number of non-centrosymmetric operators in the supergroup (N_P)	Number of centrosymmetric subgroups
aP	$P\bar{1}$	1	1
mP	$P12/m1$	2	2
mC	$C12/m1$	2	2
oP	$Pmmm$	4	5
oC	$Cmmm$	4	5
oI	$Immm$	4	5
oF	$Fmmm$	4	5
tP	$P4/mmm$	8	10
tI	$I4/mmm$	8	10
hP	$P6/mmm$	12	16
hR	$R\bar{3}m$: H‡	6	6
cP	$Pm\bar{3}m$	24	30
cI	$Im\bar{3}m$	24	30
cF	$Fm\bar{3}m$	24	30

† Crystal families: *a*, anorthic (triclinic); *m*, monoclinic; *o*, orthorhombic; *t*, tetragonal; *h*, hexagonal; *c*, cubic. Centering types: *P*, primitive; *C*, C-centered; *I*, body-centered; *F*, face-centered; *R*, rhombohedral. ‡ The hexagonal setting of space group $R\bar{3}m$.

3.2.2. Interconversion between subgroups. All derivations presented so far have been performed relative to the basis set **A** [equation (4)], which defines the primitive, reduced unit cell obtained from indexing. However, for expressing crystallographic quantities in a particular subgroup H it is common practice to adopt the particular basis $\mathbf{A}_H$ conventionally used for that subgroup, as tabulated in IT96. We have previously described the derivation of the change-of-basis matrix (**C**, **c**) needed for this transformation (§5 of Grosse-Kunstleve, 1999; **C** is a rotational part and **c** is a translational part). After following this published procedure, we make the following minor adjustments to produce results identical with established convention. For monoclinic and orthorhombic

Table 6
Conversion formulae for expressing crystallographic quantities in the conventional setting.

Quantity to be transformed	Transformation rule†
A, the matrix of direct-space basis components as defined in equation (4)	$\mathbf{A}_\mathsf{H} = \mathbf{C}_\mathsf{H}^{-1}\mathbf{A}$
A*, the matrix of reciprocal-space basis components as defined in equation (3)	$\mathbf{A}^*_\mathsf{H} = \mathbf{A}^*\mathbf{C}_\mathsf{H}$
u, an (atomic) position in direct space expressed in fractional coordinates	$\mathbf{u}_\mathsf{H} = \mathbf{C}_\mathsf{H}\mathbf{u}$
h, Miller indices	$\mathbf{h}_\mathsf{H} = \mathbf{h}\mathbf{C}_\mathsf{H}^{-1}$
W, a symmetry operator expressed in the direct-space basis	$\mathbf{W}_\mathsf{H} = \mathbf{C}_\mathsf{H}\mathbf{W}\mathbf{C}_\mathsf{H}^{-1}$

† In the equations shown, $\mathbf{C}_\mathsf{H}$ is the change-of-basis matrix for transforming quantities expressed in the primitive, reduced basis into the conventional setting for subgroup H. The subscript H on the left-hand side of the equation indicates quantities expressed in the subgroup's conventional setting.

subgroups, the affine normalizer (IT96, §15) is used to enumerate possible alternative settings for the unit cell. The best monoclinic setting is defined as the one having its unique cell angle nearest to 90°. For orthorhombic settings, the basis with the smallest cell length *a* is preferred. If this does not lead to a unique choice, the smallest length *b* is preferred, followed by the smallest length *c*. Table 4 shows the resulting rotational parts of the transformations to each subgroup of $P\,4/m\,m\,m$ (the translational parts are identically zero throughout this paper). As detailed elsewhere (Giacovazzo *et al.*, 1992), the **C** matrices can be used to transform atomic coordinates, Miller indices and symmetry operations. Sample transformation rules are shown in Table 6.

The indexing process (§3.1.1) establishes the six unit-cell measurements (three cell lengths and three cell angles) for the primitive reduced lattice. Establishing the correct Patterson symmetry will impose additional metric constraints on the cell dimensions, so that not all six of these dimensions will remain independent. Formulae for expressing the constraints are found in most crystallography texts (Blundell & Johnson, 1976), but are commonly shown in the conventional setting $\mathbf{A}_\mathsf{H}$. In contrast, the derivation of the symmetry operators in §§3.1.2–3.1.4 is most conveniently performed in the original reduced basis **A**. To work with metric constraints in the **A** basis, we therefore use the following framework, which can be used to derive the constraints in any basis, either in direct or reciprocal space. The introduction of the metrical matrix **G**,

$$\mathbf{G} = \mathbf{A}\mathbf{A}^\mathrm{T} = \begin{pmatrix} \mathbf{a}\cdot\mathbf{a} & \mathbf{a}\cdot\mathbf{b} & \mathbf{a}\cdot\mathbf{c} \\ \mathbf{a}\cdot\mathbf{b} & \mathbf{b}\cdot\mathbf{b} & \mathbf{b}\cdot\mathbf{c} \\ \mathbf{a}\cdot\mathbf{c} & \mathbf{b}\cdot\mathbf{c} & \mathbf{c}\cdot\mathbf{c} \end{pmatrix} = \begin{pmatrix} g_{11} & g_{12} & g_{13} \\ g_{21} & g_{22} & g_{23} \\ g_{31} & g_{32} & g_{33} \end{pmatrix}, \tag{9}$$

makes explicit the fact that there are initially six independent quantities $\mathbf{g} = [g_{11}\ g_{12}\ g_{13}\ g_{22}\ g_{23}\ g_{33}]$. As shown in various references (Table 2.E.1 of Giacovazzo *et al.*, 1992), the metrical matrix transforms under a general change-of-basis as

$$\mathbf{G}' = \mathbf{C}\mathbf{G}\mathbf{C}^\mathrm{T}, \tag{10}$$

where $\mathbf{G}'$ is the modified metrical matrix resulting from the transformation **C**. Metric constraints arise when we realise that the metrical matrix must remain invariant under all of the symmetry operations in the group. Focusing on just one symmetry operation, **W**, and recognizing that the symmetry operator can be treated as a change-of-basis matrix, we find that

$$\mathbf{W}\mathbf{G}\mathbf{W}^\mathrm{T} - \mathbf{G} = 0. \tag{11}$$

Expanding equation (11) in (3×3) matrix notation initially gives nine separate equations. Accounting for the symmetric nature of the metrical matrix ($g_{pq} = g_{qp}$) reduces this to just six equations in the six components of **g**:

$$\sum_{p=1}^{3}\sum_{q=p}^{3} g_{pq}(1-\delta_{pq}/2)(w_{pi}w_{qj} + w_{qi}w_{pj} - \delta_{pi}\delta_{qj} - \delta_{qi}\delta_{pj}) = 0; \quad i = 1, 2, 3;\ \ i \le j \le 3. \tag{12}$$

w_{xy} is a matrix element of **W**, and δ_{xy} is 1 if $x = y$, and 0 otherwise. To incorporate symmetry information from all the operators of the space group, this enumeration of constraint conditions is iterated over all N_P non-centrosymmetric operators. Separately applying equation (12) for each operator **W**, we arrive at the final set of $6N_P$ conditions expressed in the form

$$\mathbf{M}\mathbf{g}^\mathrm{T} = 0, \tag{13a}$$

where **M** is a $(6N_P \times 6)$ matrix completely describing the metric constraints. Since there are N_P times as many rows as independent variables in this system of equations, it is useful to employ standard Gaussian elimination techniques to rearrange equation (13*a*) into an equivalent but much simpler expression. We thus compute the row echelon form **U** (*e.g.* Strang, 1988; Boisen & Gibbs, 1990) containing zeros in all but the first *i* rows ($0 \le i \le 5$), with the result that

$$\mathbf{U}\mathbf{g}^\mathrm{T} = 0 \tag{13b}$$

describes the *i* metric constraints in the most compact fashion. In the example of Table 4, symmetry operators for the $P\,4/m\,m\,m$ subgroup produce the row echelon form

$$\mathbf{U} = \begin{pmatrix} 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 \end{pmatrix}. \tag{14}$$

This substitutes into equation (13*b*) to give the system of equations

$$\begin{aligned} \mathbf{a}\cdot\mathbf{a} - \mathbf{b}\cdot\mathbf{b} &= 0, & 2(\mathbf{a}\cdot\mathbf{b}) &= 0, \\ \mathbf{a}\cdot\mathbf{c} + \mathbf{b}\cdot\mathbf{c} &= 0, & 2(\mathbf{b}\cdot\mathbf{c}) &= 0, \end{aligned} \tag{15}$$

reducing to the expected tetragonal metric constraints $a = b$; α, β, $\gamma = 90°$.

For computational purposes, we never take this last step to express the constraints explicitly in terms of *a*, *b*, *c*, α, β and γ, as would be done for publication. Instead, the coefficients **U** are used directly in a round of quasi-Newton refinement of the **A*** basis vectors (§2.9 of Sauter *et al.*, 2004). This allows us to

take the initial unit cell from indexing and adapt it to the symmetry of any subgroup of interest.

3.3. Statistical detection of the point group

3.3.1. Scripted operation of standard programs. *LABELIT* delegates certain calculations to programs in the *CCP4* suite (Collaborative Computational Project, Number 4, 1994), version 5.0.2. Briefly, the orientation matrix **A** (§3.1.1) is provided to *MOSFLM* (Leslie, 1999) for integration of Bragg reflection intensities. A change-of-basis matrix $\mathbf{C}_\mathrm{H}$ is applied with the program *REINDEX*, and Miller indices are re-sorted with *SORT*. The program *SCALA* (written by Phil Evans, MRC Laboratory of Molecular Biology, Cambridge, UK) is then used for batch-to-batch scaling (Hamilton *et al.*, 1965), summation of partial intensities, and optionally (see below) the merging of symmetry-related reflections. Command files for running the *CCP4* programs are automatically generated and executed, and the output files are parsed to collect the results. Bragg reflection data contained in MTZ-formatted files (the standard interchange format for the *CCP4* suite) are manipulated through the `iotbx.mtz` library (Grosse-Kunstleve *et al.*, 2005).

3.3.2. Integration of the Bragg reflections. We must be careful to distinguish between the normal process of data integration used for structure solution (Leslie, 1999) and the integration of reflections for determination of the Patterson symmetry (Fig. 2). For structure solution, it is assumed that the Bravais lattice is known ahead of time. The positions of Bragg reflections can be predicted with high accuracy because (*a*) the initial **A** basis is symmetry-adapted with the proper metric constraints, and (*b*) the metric constraints are used again in a cycle of postrefinement to improve the model further (Rossmann *et al.*, 1979). For symmetry determination, we can assume no prior knowledge about the metric symmetry. Since the unit-cell model is not optimal, we do not know if weak Bragg measurements in the highest-resolution shells of reciprocal space are due to truly weak intrinsic diffraction, or to inaccurate spot positions. We therefore set an unusually stringent criterion for accepting data, truncating the diffraction pattern at a resolution limit where the average $I/\sigma(I)$ ratio is 5.0 (treating both partially and fully recorded reflections as a single set). Once the Patterson symmetry is determined, the data are reprocessed for structure solution using the proper Bravais lattice.

3.3.3. Track 1: traditional calculation of merging reliability. To determine if global indicators of data quality can be used to distinguish among possible point groups, the R_merge statistic [equation (1)] was calculated for each centrosymmetric subgroup. For this calculation, the data are integrated only once, but are then reindexed, scaled and merged between 1 and 30 separate times depending on the apparent Bravais lattice (see Table 5). R_merge values calculated with *SCALA* indeed reveal the correct symmetry in many cases, but the results (§4) make it clear that global indicators fail to give an accurate picture when some of the subgroup operators are poorly represented in the data set. The *Track 1* approach is therefore incorporated into *LABELIT* for research purposes only, not as a recommended feature.

3.3.4. Track 2: novel calculation of symmetry-operator reliability. When calculating operator-specific statistics, we wish to avoid the computationally intensive step of executing the program *SCALA* separately for each possible subgroup. Furthermore, experience has shown that there may not be enough information to scale the data in low-symmetry subgroups when the diffraction data in question are from a small incomplete data set (*i.e.* data collected from a small angular rotation range). For these reasons, batch-to-batch scaling and summation of partial intensities is performed once only, in the metric supergroup (Fig. 2). Many cases were examined in which the metric supergroup contained symmetry operations not present in the actual data, and it was found that the resulting batch-to-batch scale factors were still suitable for deriving the true point group.[1]

In contrast to the *Track 1* procedure above, *Track 2* does not ask *SCALA* to merge symmetry-related reflections. Instead, the command `output unmerged original` is issued, causing *SCALA* to list the scaled, summed unmerged intensities with their original Miller indices. It is therefore possible to evaluate the expression

$$(\mathbf{h}_b = \mathbf{h}_a\mathbf{W}) \quad \text{or} \quad (\mathbf{h}_b = \mathbf{h}_a\overline{\mathbf{W}}) \tag{16}$$

to test whether two Miller indices $\mathbf{h}_a$ and $\mathbf{h}_b$ are related by a symmetry operation **W**. Note that the 'original' Miller indices referred to in this context have already been transformed during the *REINDEX* step into the conventional setting of the metric supergroup. Therefore, the proper change-of-basis matrix $\mathbf{C}_\mathrm{M}$ for this transformation must also be applied to the set of possible symmetry operations (§3.1.4) using the rule listed in Table 6.

4. Application to experimental data

The public availability of data from the NIH-funded Protein Structure Initiative projects provides a testing ground for new processing methodologies like those outlined above. We evaluated the two symmetry determination methods (*Track 1* and *Track 2*) using 326 data sets collected by the Joint Center for Structural Genomics (JCSG; Lesley *et al.*, 2002), provided to us by Ashley Deacon. Data sets were included from both published structures and from studies terminated for various reasons. Each data set was indexed based on two images collected at rotational settings 90° apart. Reliability statistics were calculated on nested subsets of images, beginning with the data set's first image. One experiment is highlighted below because it illustrates a situation where the *Track 2* approach is

[1] For all *SCALA* runs, outlier measurements were rejected if they deviated by more than four standard deviations from the weighted mean $I(hkl)$; however, it is not strictly necessary to reject outliers in order to use supergroup scale factors, even when the supergroup symmetry is higher than the true symmetry. We speculate that the problem of Hamilton *et al.* (1965) batch-to-batch scaling is so overdetermined (*e.g.* in Table 7, ∼30000 independent measurements are used to determine ∼30 scale factors) that scaling is rather insensitive to the presence of extra symmetry operators.

superior for timely discovery of the Patterson symmetry. Additional results will be presented elsewhere.

4.1. Cases where symmetry-operator analysis helps resolve symmetry ambiguity

JCSG collected diffraction data from two crystal forms of hypothetical protein 29342463 (EF0366) from *Enterococcus faecalis* V583. The first form was solved at 2.52 Å resolution in space group $P6_5$, with one protomer per asymmetric unit (Protein Data Bank accession code 1VPY). The second data set was not initially amenable to analysis. The diffraction pattern could be indexed and integrated, but the choice between the Patterson symmetries $P4/m$ and $Pmmm$ was not clear. Scaling and merging under $P4/m$ resulted in large numbers of measurements being rejected as outliers (data not shown). However, the fact that the $R_{\rm merge}$ for $P4/m$ (Table 7*b*) was of intermediate value between those of obviously correct ($P\bar{1}$) and incorrect ($P4/mmm$) symmetries left open the question of whether there were other systematic problems with the data set. Because of this ambiguity and the fact that the crystal diffracted only to a limit of 3.10 Å, JCSG's initial efforts were focused on the better-diffracting crystal form #1.

Since a goal of our work is to develop automated procedures for choosing among candidate Patterson symmetries, it was of interest to us to investigate in more detail the symmetry of the poorer-diffracting crystal form #2. The evidence from outlier rejections and merging reliability statistics argues against $P4/m$, but it is difficult to combine this information into a suitable scoring function. After developing equation (2), we realised that the analysis based on individual symmetry operators (Table 7*a*) offers the clearest indication that the correct Patterson symmetry is $Pmmm$. Once the data are integrated and scaled, it only takes an additional 4 s of CPU time (on a 2.8 GHz Xeon processor) to calculate Table 7(*a*), so $R_{\rm symop}$ is a very practical addition to existing methods.

Knowing the correct Patterson symmetry allowed us to solve the structure of crystal form #2 in space group $P2_12_12_1$, using the method of molecular replacement. The 1VPY structure was used as a replacement model, and two protomers were located in the asymmetric unit, forming a non-crystallographic dimer. The solved structure shows that the near equality of the *a* and *b* axis lengths is pure coincidence; the unit-cell contents do not display pseudo fourfold symmetry. Full model building and refinement was conducted in collaboration with JCSG (details to be published elsewhere), and the model deposited under PDB accession code 1ZTV. Importantly, 97% of the protomer's residues are present in the 1ZTV model, in contrast to only 87% of residues in 1VPY, which lacks two hairpin turns and the C-terminal. Thus in hindsight, the additional structural detail observed warrants the extra effort to identify the space group, despite the poorer diffraction limit of crystal form #2.

Another JCSG structure was solved from a primitive orthorhombic crystal with a pseudo-tetragonal lattice, which caused initial difficulty identifying the symmetry (PDB accession code 1VR6). The structure was solved before the symmetry-operator analysis was developed for point-group assignment. Afterwards, it was confirmed that $R_{\rm symop}$ analysis easily gives the correct symmetry.

Table 7
Symmetry analysis of the full 1ZTV data set out to a 3.37 Å limit.

The full data set consists of 90 1° oscillation images; the incident wavelength is 0.979 Å. The $Pmmm$ diffraction pattern is 99.4% complete out to the limiting resolution of 3.10 Å. Indexing, integration and scaling took 366 s, single-threaded on a 2.8 GHz 32-bit Xeon machine with 2.0 Gbyte memory running under RedHat Linux 8.0. The images contained 3072 × 3072 pixels and were recorded in 18 Mbyte files.

(*a*) Correlation of measured intensities.

Operator pair(s)	Representative operator(s), **W**	$N_{\rm W}$	$R_{\rm symop}$ (%)
a	1	19618	5.5
b	$[100]2$	4720	7.9
c	$[010]2$	12422	9.2
d	$[001]2$	29497	5.9
e	$[110]2$	30337	42.8
f	$[\bar{1}10]2$	36525	43.7
g, *h*†	$[001]4^{1,-1}$	18717	44.9

(*b*) Quality indicators for possible subgroups.

Subgroup	Operators present	Maximum $R_{\rm symop}$ (%)	$R_{\rm merge}$ (%) (with outlier rejection‡)	$R_{\rm merge}$ (%) (no outlier rejection§)
$P4/mmm$	*a b c d e f g h*	44.9	44.0	42.2
$P4/m$	*a d g h*	44.9	9.7	17.5
$Cmmm$	*a d e f*	43.7	41.9	40.1
$C12/m1$	*a e*	42.8	40.6	35.6
$C12/m1$	*a f*	43.7	31.0	37.9
$Pmmm$¶	*a b c d*	9.2	7.4	7.5
$P12/m1$	*a b*	7.9	5.7	5.8
$P12/m1$	*a c*	9.2	6.1	6.2
$P12/m1$	*a d*	5.9	6.9	7.0
$P\bar{1}$	*a*	5.5	5.4	5.5

† Rotational operators $[001]4^1$ and $[001]4^{-1}$ are grouped together because they are mutual inverses, respectively representing counterclockwise and clockwise quarter-turn rotations about [001]. The operators must be considered together in order to construct the complete list of $N_{\rm W}$ pairs of Miller indices related by the fourfold axis. Similar groupings must be made in other metric supergroups containing mutually inverse pairs of three- and sixfold operators. This is not an issue for one- and twofold operators because they are self-inverses. ‡ Scaling and merging performed by *SCALA* separately for each subgroup. § Supergroup scaling performed exactly as described in §3.3.4, with $R_{\rm merge}$ then calculated strictly with equation (1), with no further outlier rejections. ¶ Correct subgroup.

4.2. Early identification of the point group through symmetry-operator analysis

A goal of many beamline software developers is to establish automated protocols to optimize the data collection strategy. The result of §4.1 clearly shows that incorporation of $R_{\rm symop}$ analysis in these protocols will increase the reliability of point-group determination. We now ask whether an answer can be obtained quickly: before the crystal has undergone too much radiation exposure, which could produce radiation damage, and before too much time has elapsed, which would make the process inefficient. Fig. 3 explores how well $R_{\rm symop}$ can reveal the point group of 1ZTV as the data collection progresses. In the interval shown (data frames 4–90), it is always possible to distinguish incorrect subgroups like $P4/mmm$ and $P4/m$

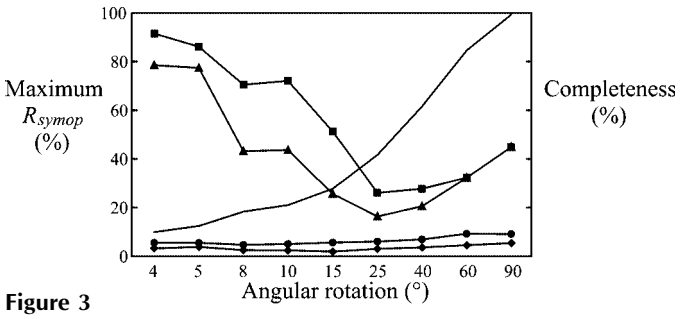

Figure 3
Ability to establish the point group of the 1ZTV crystal from nested subsets of the full data. The maximum R_{symop} value is plotted for key subgroups of interest: $P4/mmm$ (squares); $P4/m$ (triangles); $Pmmm$ (circles); $P\bar{1}$ (diamonds). Data completeness computed in the true Patterson group ($Pmmm$) is shown as a line without symbols.

from subgroups that are consistent with the data like $Pmmm$ and $P\bar{1}$. Remarkably, the symmetry can be determined after only 4° of angular rotation, at a point where only 10% of the unique reflections have been measured. Based on both criteria, namely radiation exposure and required CPU time, it would be feasible to pause after frame 4 to optimize the data collection strategy.

Details of the symmetry analysis using frames 1–4 are presented in Table 8. The number of Miller index pairs available for comparison is quite small: for example $N_{\mathbf{W}}$ is only 1 for operator *f*, 3 for operator *a*, and 6 for operator *e*. With small populations, there is a danger that the resultant $R_{\rm symop}$ values may be biased by sampling error. Yet we do not wish to be too cautious, since it may be possible to prove or disprove a point group from a few observations. Based on results from all 326 data sets considered, the following heuristic was established: for most Bravais lattices, no $R_{\rm symop}$ is calculated if $N_{\mathbf{W}} < 5$ (but for operator pair *a*, the cutoff is $N_{\mathbf{W}} < 3$). Therefore, in Table 8(*a*), no $R_{\rm symop}$ is listed for operator pair *f*. However, if the Bravais lattice is cubic or primitive hexagonal, it is more difficult to obtain the requisite sample sizes for the large number of symmetry operators, so a more permissive cutoff of $N_{\mathbf{W}} < 2$ is used. In the future we plan to replace these simple heuristics with a more rigorous statistical treatment so that an error estimate can be placed on $R_{\rm symop}$.

Point-group determination depends on the ability to conclude from the $R_{\rm symop}(\mathbf{W})$ statistics that certain symmetry operators (but not others) describe true correlations in the observed reflections. In the case illustrated in Table 8(*a*) there is an obvious distinction between those operators permitted by the data (*a*, *b*, *c* and *d*), and those ruled out (*e*, *g* and *h*). In the context of beamline automation, a simple clustering procedure can perform this analysis. Heuristic rules were implemented to divide the $R_{\rm symop}$ values into one or two clusters: one cluster if all values are acceptably small, and two clusters otherwise. Parameters in the rules were adjusted in such a way as to maximize the ability to select the correct point group in the JCSG data sets, with as few frames as possible taken from the beginning of each data set. The resulting code from this trial-and-error process is distributed with *LABELIT* in the file `track2_cluster.py`. A regression test is also included to allow interested users to experiment with alternative criteria. Under some conditions, it is not possible to divide the $R_{\rm symop}$ values into two clear groups, and consequently no conclusions can be made about the point group. Such cases arise when the incomplete angular rotation of a crystal gives an insufficient sample size ($N_{\mathbf{W}}$) of Miller index pairs. Furthermore, of the 326 successfully integrated data sets we considered, about 1% contained data of such poor quality that the symmetry could not be identified. A much larger concern for beamline automation (but not within the scope of this paper) is to recognize systematically cases where indexing and integration are not successful. We found several dozen additional data sets where various pathologies such as high mosaicity, crystal splitting, or radiation damage rendered the data useless.

Once the sets of permitted and disallowed operators are listed ($\{a, b, c, d\}$ and $\{e, g, h\}$ respectively, in this case), *LABELIT* selects the correct subgroup by logical inference. In this example, Table 8(*b*) shows that $Pmmm$ is the only possible subgroup. It is not true that a subgroup can be selected simply if it has a low maximum $R_{\rm symop}$ value; this is a necessary but not sufficient condition. For example, even though the second $C12/m1$ subgroup has a maximum $R_{\rm symop}$ of 3.3% (due to the identity operator *a*), the fact that it lacks the permitted operators *b*, *c* and *d* excludes it from consideration. Interestingly, even though the status of operator *f* is

Table 8
Symmetry analysis of the incomplete 1ZTV data set out to a 3.37 Å limit.

Analysis of the first four 1° oscillation images, at which point the $Pmmm$ diffraction pattern is 10% complete. The computations shown (including indexing, integration and scaling) took 24 s under the same conditions noted in Table 7.

(*a*) Correlation of measured intensities.

Operator pair(s)	Representative operator(s), **W**	$N_{\mathbf{W}}$	$R_{\rm symop}$ (%)
a	1	3	3.3
b	$^{[100]}2$	131	5.6
c	$^{[010]}2$	24	4.2
d	$^{[001]}2$	115	4.2
e	$^{[110]}2$	6	91.5
f	$^{[\bar{1}10]}2$	1	–
g, *h*	$^{[001]}4^{1,-1}$	12	78.5

(*b*) Quality indicators for possible subgroups.

Subgroup	Operators unknown	Operators permitted	Operators ruled out	Maximum $R_{\rm symop}$ (%)
$P4/mmm$	*f*	*a b c d*	*e g h*	91.5
$P4/m$		*a d*	*g h*	78.5
$Cmmm$	*f*	*a d*	*e*	91.5
$C12/m1$		*a*	*e*	91.5
$C12/m1$	*f*	*a*		3.3
$Pmmm$ †		*a b c d*		5.6
$P12/m1$		*a b*		5.6
$P12/m1$		*a c*		4.2
$P12/m1$		*a d*		4.2
$P\bar{1}$		*a*		3.3

† Correct subgroup.

not revealed directly by the experiment, it can be inferred: knowing that $\{a, b, c, d\}$ are allowed, $\{e, g, h\}$ are disallowed, and that the subgroup must be in the list in Table 8(b), implies that f is disallowed too. In a similar fashion, the subgroup could still be uniquely identified even if $R_{\rm symop}$ is known for only two of the three operators b, c and d.

4.3. Correlation of Friedel pairs

The heuristic process of dividing the symmetry operators into one or two clusters treats all operator pairs the same with one important exception: a, the operator pair consisting of the identity 1 and inversion $\bar{1}$. It is clear that operator pair a must be categorized with the 'permitted' cluster: $R_{\rm symop}(1)$ reflects the experimental uncertainty in making repeat measurements of the same reflection, while $R_{\rm symop}(\bar{1})$ includes additional differences due to anomalous dispersion. If the value of $R_{\rm symop}(a)$ (*i.e.* all repeat measurements and Friedel pairs taken together) clusters with the 'disallowed' operators, or if it exceeds a certain absolute value (*e.g.* 20%), we must conclude either that the diffraction pattern is misindexed or that there is some other serious measurement problem.

In the 1ZTV data set, the $R_{\rm symop}(a)$ value is slightly smaller than that for any other operator (see Tables 7*a*, 8*a*, and Fig. 3). Indeed, this is generally true throughout the collection of JCSG data sets analyzed. These data sets all have angular rotation ranges of less than 180°, as is the normal protocol for collecting native data when there is no intention of analyzing anomalous differences. Consequently, there are no repeat measurements of a given Bragg spot; the only contribution to $R_{\rm symop}(a)$ is from Friedel pairs. A consideration of Bragg diffraction geometry helps to explain why Friedel pairs are better correlated than other types of symmetry-related pairs.

To illustrate, Fig. 1 shows a Miller index hkl in diffracting position on the trailing (concave forwards) edge of the Ewald sphere. The Friedel mate $\bar{h}\bar{k}\bar{l}$ reaches the diffracting position on the sphere's leading (convex forwards) edge after the crystal has been rotated by an angle

$$\Delta\varphi = 2\sin^{-1}(\lambda/2d), \qquad (17)$$

where λ is the X-ray wavelength and d is the $\{hkl\}$ lattice spacing. Equation (17) implies that Friedel pairs are highly correlated for two distinct reasons, depending on whether the total angular rotation producing the data is large or small. For large angular rotations (*e.g.* the nearly complete 1ZTV data set of Table 7), the Friedel pairs always come from close φ settings; specifically $\Delta\varphi \leq 16.7°$ in Table 7, even though the full range of crystal rotation is 90°. It is reasonable that the Friedel pairs have the best correlation because sources of error that depend on the crystal orientation, such as absorption of the incident beam, vary the least. For small angular rotations (*e.g.* the incomplete data from a very thin wedge in Table 8), the Friedel pairs all come from low resolution shells, in particular $d \geq 14.0$ Å for this 4° rotation. The low-resolution Friedel pairs are therefore likely to consist of brighter spots, having relative measurement errors that are smaller than average.

A different phenomenon arises when the crystal is rotated through an angle greater than 180°, as is done for 'inverse-beam' measurement of Bijvoet pairs (Hendrickson & Ogata, 1997). Pairings arise from the same edge (either leading or trailing) of the Ewald sphere, and the relationship between the incident beam and the crystal lattice is nearly identical when each mate is in diffracting position. If such data are included, this further reinforces the observation that the inversion operator produces the lowest $R_{\rm symop}$ values.

5. Conclusion

When selecting the proper Laue class and Patterson symmetry, it can be misleading to compare statistics that aggregate all reflections, particularly when the data set is incomplete. If potential symmetry relations are poorly sampled in the experimental data, there may be little hint from global statistics such as $R_{\rm merge}$ and $I/\sigma(I)$ that a high-symmetry group is incorrect. The clearest method is to consider whether the data support the individual symmetry operators that comprise the symmetry group. While existing programs implicitly use symmetry during the data-merging process, available software does not generally present this information in a way that allows particular symmetry operators to be clearly accepted or rejected. In contrast, the calculation of $R_{\rm symop}(\mathbf{W})$ [equation (2)], along with the enumeration of all subgroups that are metrically possible (Table 4), together give a straightforward framework for deducing the symmetry unambiguously. *LABELIT* generates the required output (Tables 7 and 8) in response to a single command, without the necessity for separate jobs to evaluate each possible symmetry.

For future beamline automation efforts, it will be important to be able to identify the symmetry as early as possible in the data acquisition process. The $R_{\rm symop}$ method appears to be useful in this regard: an analysis of 326 data sets from the JCSG suggests that the Patterson symmetry can generally be determined after a narrow angular wedge of data has been collected. For the monoclinic crystal family, an average rotation of 9° is required; for the orthorhombic crystal family, 6°; and for higher-symmetry families, 5°.

The tools presented above can be used to confirm a Patterson symmetry assignment with high confidence. In order to substantiate the published symmetry, it would be most advantageous to archive unmerged intensity data with original Miller indices, in a public database such as the PDB (Berman *et al.*, 2000). We favor this as an addition to, not as a substitute for, the more common practice of archiving merged data.

In the future, we plan to test whether the $R_{\rm symop}$ method can be applied to the detection of pseudosymmetry; *e.g.* in situations where an apparent symmetry breaks when investigated at high resolution.

6. Software availability

Non-commercial users may download *LABELIT* at the URL http://cci.lbl.gov/labelit. Metric symmetry and Patterson

symmetry can be determined, respectively, with the commands `labelit.index` and `labelit.rsymop`. As presently implemented, symmetry is analyzed without considering the possibility of merohedral twinning. Detailed instructions are posted on the Web site. Code to reproduce the tables in this paper is given in the subdirectory labelit/labelit/publications/rsymop.

We thank Ashley Deacon, Herb Axelrod and Guenter Wolf (Joint Center for Structural Genomics and the Stanford Synchrotron Radiation Laboratory) for making raw data sets available for testing *LABELIT*; and Tom Terwilliger (Los Alamos National Laboratory) for helpful discussions. Our work was supported by the US Department of Energy under Contract No. DE-AC02-05CH11231, and by funding from the NIH/NIGMS under grant number 1P50GM62412.

References

Arndt, U. W., Champness, J. N., Phizackerley, R. P. & Wonacott, A. J. (1973). *J. Appl. Cryst.* **6**, 457–463.
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.
Blundell, T. L. & Johnson, L. N. (1976). *Protein Crystallography*. London: Harcourt Brace.
Boisen, M. B. Jr & Gibbs, G. V. (1990). *Mathematical Crystallography, Reviews in Minerology*, Vol. 15 (revised ed.). Washington, DC: Mineralogical Society of America.
Dauter, Z. (1999). *Acta Cryst.* D**55**, 1703–1717.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Fischer, W. & Koch, E. (1996). In *International Tables for Crystallography, Volume A: Space-Group Symmetry*, 4th revised ed., edited by T. Hahn, pp. 793–798. Dordrecht: Kluwer.
Giacovazzo, G., Monaco, H. L., Viterbo, D., Scordari, F., Gilli, G., Zonotti, G. & Catti, M. (1992). *Fundamentals of Crystallography*. IUCr/Oxford University Press.
Goldstein, H. (1980). *Classical Mechanics*, 2nd ed., pp. 164–166. Reading, MA: Addison-Wesley.
Grosse-Kunstleve, R. W. (1999). *Acta Cryst.* A**55**, 383–395.
Grosse-Kunstleve, R. W., Sauter, N. K. & Adams, P. D. (2004*a*). *Acta Cryst.* A**60**, 1–6.
Grosse-Kunstleve, R. W., Sauter, N. K. & Adams, P. D. (2004*b*). *Newslett. IUCr Commission Crystallogr. Comput.* **3**, 22–31.
Grosse-Kunstleve, R. W., Afonine, P. V., Sauter, N. K. & Adams, P. D. (2005). *Newslett. IUCr Commission Crystallogr. Comput.* **5**, 69–91.
Hahn, T. (1996) Editor. *International Tables for Crystallography, Volume A: Space-Group Symmetry*, 4th revised ed. Dordrecht: Kluwer.
Hall, S. R. (1981). *Acta Cryst.* A**37**, 517–525.
Hall, S. R. & Grosse-Kunstleve, R. W. (2001). *International Tables for X-ray Crystallography, Volume B: Reciprocal Space*, 2nd ed., edited by U. Shmueli, pp. 112–119. Dordrecht: Kluwer.
Hamilton, W. C., Rollett, J. S. & Sparks, R. A. (1965). *Acta Cryst.* **18**, 129–130.
Hendrickson, W. A. & Ogata, C. M. (1997). *Methods Enzymol.* **276**, 494–523.
Kleywegt, G. J., Hoier, H. & Jones, T. A. (1996). *Acta Cryst.* D**52**, 858–863.
Le Page, Y. (1982). *J. Appl. Cryst.* **15**, 255–259.
Lesley, S. A., Kuhn, P., Godzik, A., Deacon, A. M., Mathews, I., Kreusch, A., Spraggon, G., Klock, H. E., McMullan, D., Shin, T., Vincent, J., Robb, A., Brinen, L. S., Miller, M. D., McPhillips, T. M., Miller, M. A., Scheibe, D., Canaves, J. M., Guda, C., Jaroszewski, L., Selby, T. L., Elsliger, M.-A., Wooley, J., Taylor, S. S., Hodgson, K. O., Wilson, I. A., Schultz, P. G. & Stevens, R. C. (2002). *Proc. Natl Acad. Sci. USA*, **99**, 11664–11669.
Leslie, A. G. W. (1999). *Acta Cryst.* D**55**, 1696–1702.
Popov, A. N. & Bourenkov, G. P. (2003). *Acta Cryst.* D**59**, 1145–1153.
Ravelli, R. B. G., Sweet, R. M., Skinner, J. M., Duisenberg, A. J. M. & Kroon, J. (1997). *J. Appl. Cryst.* **30**, 551–554.
Rossmann, M. G., Leslie, A. G. W., Abdel-Meguid, S. S. & Tsukihara, T. (1979). *J. Appl. Cryst.* **12**, 570–581.
Sauter, N. K., Grosse-Kunstleve, R. W. & Adams, P. D. (2004). *J. Appl. Cryst.* **37**, 399–409.
Steller, I., Bolotovsky, R. & Rossmann, M. G. (1997). *J. Appl. Cryst.* **30**, 1036–1040.
Strang, G. (1988). *Linear Algebra and its Applications*, 3rd ed. San Diego: Harcourt Brace Jovanovich.
Weiss, M. S. (2001). *J. Appl. Cryst.* **34**, 130–135.