Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

On effective and optical resolutions of diffraction data sets

Ludmila Urzhumtseva,[a] Bruno Klaholz[b] and Alexandre Urzhumtsev[b,c]*

[a]Architecture et Réactivité de l'ARN, UPR 9002 CNRS, IBMC (Institute of Molecular and Cellular Biology), 15 Rue René Descartes, 67084 Strasbourg, France; Université Louis Pasteur de Strasbourg, 67000 Strasbourg, France, [b]IGBMC (Institute of Genetics and of Molecular and Cellular Biology), Department of Integrative Structural Biology, Centre National de la Recherche Scientifique (CNRS) UMR 7104/ Institut National de la Santé de la Recherche Médicale (INSERM) U964/Université de Strasbourg, 1 Rue Laurent Fries, 67404 Illkirch, France, and [c]Physical Department, Université de Lorraine, 54506 Vandoeuvre-lès-Nancy, France

Correspondence e-mail: sacha@igbmc.fr

Received 4 May 2013
Accepted 15 June 2013

In macromolecular X-ray crystallography, diffraction data sets are traditionally characterized by the highest resolution $d_{\rm high}$ of the reflections that they contain. This measure is sensitive to individual reflections and does not refer to the eventual data incompleteness and anisotropy; it therefore does not describe the data well. A physically relevant and robust measure that provides a universal way to define the 'actual' effective resolution $d_{\rm eff}$ of a data set is introduced. This measure is based on the accurate calculation of the minimum distance between two immobile point scatterers resolved as separate peaks in the Fourier map calculated with a given set of reflections. This measure is applicable to any data set, whether complete or incomplete. It also allows characterizion of the anisotropy of diffraction data sets in which $d_{\rm eff}$ strongly depends on the direction. Describing mathematical objects, the effective resolution $d_{\rm eff}$ characterizes the 'geometry' of the set of measured reflections and is irrelevant to the diffraction intensities. At the same time, the diffraction intensities reflect the composition of the structure from physical entities: the atoms. The minimum distance for the atoms typical of a given structure is a measure that is different from and complementary to $d_{\rm eff}$; it is also a characteristic that is complementary to conventional measures of the data-set quality. Following the previously introduced terms, this value is called the optical resolution, $d_{\rm opt}$. The optical resolution as defined here describes the separation of the atomic images in the 'ideal' crystallographic Fourier map that would be calculated if the exact phases were known. The effective and optical resolution, as formally introduced in this work, are of general interest, giving a common 'ruler' for all kinds of crystallographic diffraction data sets.

1. Introduction

X-ray crystallography is the principal method for obtaining three-dimensional atomic structures of macromolecules and their complexes. The method starts with growing crystals of the macromolecule of interest, which are then exposed to an X-ray beam. These experimental steps result in diffraction intensities $\{I_{\mathbf{s}}\}$ measured for a set S of 'reflections', *i.e.* three-dimensional vectors $\mathbf{s} = (hkl)$ with integer indices (hkl). Each reflection is characterized by its resolution $d_{\mathbf{s}} = |\mathbf{s}|^{-1}$ calculated from its indices and the unit-cell parameters of the crystal. The larger the indices are, the smaller the value of $d_{\mathbf{s}}$ and the higher the resolution.

The electron-density distribution in the crystal, being a periodic function, can be presented as a Fourier series, and the experimental diffraction intensities allow the amplitudes

$\{F_{\mathbf{s}}\}$ for the corresponding complex Fourier coefficients $\{\mathbf{F}_{\mathbf{s}} = F_{\mathbf{s}}\exp(i\varphi_{\mathbf{s}})\}$ to be obtained. These are known in crystallography as 'structure factors'. Estimation of the associated phase values $\{\varphi_{\mathbf{s}}\}$ by one of the phasing methods allows one to calculate a three-dimensional Fourier synthesis $\rho(\mathbf{r})$ for the crystal under study, to present it as corresponding Fourier maps at various cutoff levels and to then interpret it by molecular model building.

Larger diffraction data sets give more detailed images and lead to better models. However, crystal quality, experimental conditions and experimental noise limit the amount of diffraction data that can be collected. There are many basic questions relevant to the choice of diffraction data for analysis of a given crystal structure, including the following.

(i) Which reflections contain structural information and 'how much'?

(ii) Which reflections should be used to obtain the most accurate, most detailed Fourier maps and with the maximum possible structural information (if the exact phase values are available)?

(iii) How accurate and detailed would these maps be?

(iv) How accurate and detailed are the particular calculated Fourier maps?

These well known questions are related to each other and have some apparent similarity, but in fact they are quite different and require separate analysis. In particular, while questions (i)–(iii) can be posed as soon as a set of diffraction amplitudes is available, the answer to question (iv) depends not only on the experimental data but also on the phase values and on the composition of the Fourier coefficients. Also, accuracy, the level of detail of the Fourier maps in questions (ii)–(iv) and the amount of structural information which they contain are different aspects of the problem and, for example, the choice of the goal may change the answer to question (ii).

Recently, Karplus & Diederichs (2012) have made significant progress in answering question (i). In this article, we do not analyze the accuracy of the data and that of the resulting maps, but concentrate on the second part of question (iii) of quantifying a set of selected reflections by the level of detail in the corresponding Fourier maps. At the same time, our analysis shows quantitatively that excluding some reflections from the selected data set can improve the detail of the maps, explicitly indicating the data to be excluded for this goal; this partially answers question (ii).

Sets S of selected reflections are traditionally quantified by their highest resolution, *i.e.* by the smallest value of $d_{\mathbf{s}}$ for the present reflections: $d_{\rm high} = \min\{d_{\mathbf{s}}, \mathbf{s} \in S\}$. In general, a smaller $d_{\rm high}$ indicates that smaller structural details are visible in the maps. The resolution of the data set may be a limiting parameter for a number of crystallographic procedures such as Fourier map improvement, model building or direct phasing (see, for example, Sheldrick, 1990; Caliandro *et al.*, 2005*a*,*b* and references therein); therefore, its exact measurement is important.

Unfortunately, the traditional measure $d_{\rm high}$ may change after the addition or removal of a few reflections with no apparent change in the synthesis; it does not reflect the incompleteness of the data set and reveals nothing about its eventual anisotropy. Therefore, a better quantitative characteristic of diffraction data sets is required. Other traditional characteristics of diffraction data sets are the overall data completeness, $C_{\rm overall}$, and the data completeness in the highest resolution shell, $C_{\rm high}$. Weiss (2001) suggested characterizing the 'true resolution' of a data set by a semi-empirical formula that combines $d_{\rm high}$ and $C_{\rm overall}$.

In this article, we propose a new accurate numerical characteristic showing how detailed the diffraction data set is. This measure is independent of outliers and reflects the data-set incompleteness. The same calculations allow characterization of the data-set anisotropy. Our analysis also illustrates its difference from and the relation between different meanings of the term 'resolution' used in crystallography.

2. Minimum distance and the interference function

2.1. Minimum distance: generalities

To quantify how detailed a given set of reflections is, some 'ruler' is required. The old and natural idea is to look at the size of the details in the Fourier synthesis calculated with these data, for example determining

> the minimum distance at which two features in the corresponding electron-density map can be resolved

(Weiss, 2001). Similar phrases may be also found in various crystallographic textbooks; for example, Blundell & Johnson (1976).

There are several studies that are relevant to the problem of the minimum distance. To the best of our knowledge, none of these specify which synthesis ('electron-density map') should be analyzed, while obviously the size of the details depends on the choice of the Fourier coefficients used.

Concerning the 'features', several authors have suggested looking at the images of scatterers calculated with the given data set. Electron density for crystallographic scatterers is described by a function with a positive maximum at their centre [we leave aside the particular case of artificial scatterers with negative values such as some IAS (Afonine *et al.*, 2007) and some other particular situations]. Their Fourier images also show a central peak, which, however, may be blurred and surrounded by Fourier ripples, *i.e.* by auxiliary peaks. When two scatterers are too close to each other their images merge and show a single common peak. A minimum distance between two scatterers exists when their principal peaks are observed separately in the given Fourier synthesis. Intuitively, this minimum distance can be used as a 'ruler' to describe how detailed the data set is, but it requires choice of the scatterers and the way to calculate the 'minimum distance' to be specified.

Each scatterer has an associated size and uncertainty in its localization. It is clear intuitively that the sharper the scatterer, the shorter the minimum distance when two of them are observed separately in the same Fourier synthesis. The immobile point scatterer is the sharpest one. James (1948)

studied its images at a resolution d_{high}. He estimated the minimum distance from the first zero of the image, obtaining $0.715d_{high}$. Stenkamp & Jensen (1984) discussed some confusing and contradicting observations and suggested estimating it from the first minimum for a C atom with atomic displacement factor B = 10 Å^2; this resulted in $0.917d_{high}$. In fact, there is no strict mathematical reason for the minimum distance to be equal to the distance either to the first zero or to the first minimum. Some further discussions about the separation of images of point scatterers and atoms, although at relatively high resolutions only, may be found in Altomare *et al.* (2008).

Vaguine *et al.* (1999) proposed another formal procedure to estimate the minimum distance (implicitly, they analyzed atoms with the atomic displacement parameter estimated from the diffraction intensities). Their main idea was based on the fact that the minimum distance is exactly known for two equal Gaussian peaks. For this reason, a Gaussian approximation for several functions involved in calculations was the principal tool in this approach.

In fact, knowledge of the minimum distance for two Gaussians is based on a trivial mathematical analysis. For a function describing the image of two peaks, peak separation means the presence of a local minimum along the line joining their centres; in other words, a point with a positive second derivative of the function. In contrast, when the peaks are merged such a point does not exist. Checking this condition provides a universal way to calculate the minimum distance for any kind of scatterer.

Moreover, one does not need to calculate a series of images for two equal scatterers at different distances. The critical situation is when they are exactly at the minimum distance.

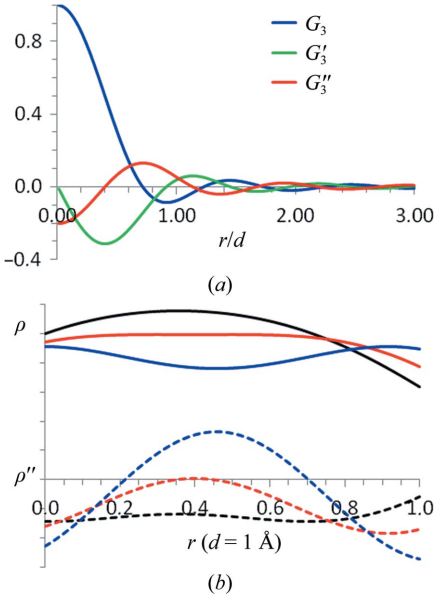

Figure 1
Three-dimensional interference function. (*a*) The function $G_3(2\pi rd^{-1}) = G_3(2\pi x)$ (blue curve) and its first (green) and second (red) derivatives. (*b*) Solid lines show the image of two point scatterers, in arbitrary units, at the resolution d_{high} = 1 Å when the distance between them is equal to $0.715d_{high}$ (black), $0.798d_{high}$ (this work; red) or $0.917d_{high}$ (blue). Dashed curves show the corresponding second derivatives. See text for details.

This means that a single point exists where the second derivative of the Fourier image is equal to zero. Since the scatterers are the same, this point is in the middle between them and the second derivative is equal to zero for each of them; this is the inflection point of the image. Therefore, such an inflection point can be calculated from the Fourier image of a single scatterer, for example placed at the origin. The minimum distance is then exactly twice as large as the distance to this inflection point.

Such an approach does not require any approximations, only the choice of the corresponding scatterer and calculation of its image using the diffraction data under study. As shown below, the image of an immobile point scatterer is fully described by the set of reflections $S = \{\mathbf{s}\}$ with no measured intensities required. Therefore, the minimum distance calculated for point scatterers describes the 'geometry' of the diffraction data set.

The set of available diffraction intensities corresponds to the atoms of the particular structure. These atoms are the 'physical entities' that we are interested in, with their shape and the uncertainties in their positions, and their analysis is complementary to that of point scatterers. Owing to the variety of atoms, some atom 'typical for the given structure' will be chosen as another particular type of scatterer for which the minimum distance is calculated.

In the following, we formally define the minimum distance and then accurately calculate it for a number of practical situations. This results in important quantitative and qualitative conclusions and also leads to a procedure applicable in structural projects.

2.2. Minimum distance for point scatterers

We start from the analysis of an isolated point scatterer, for which analytic calculations are possible. By definition, when the corresponding diffraction data are cut at a resolution d_{high}, an image of a point scatterer placed at the origin **O** of the coordinate system is described by the function

$$\upsilon(\mathbf{r}; d_{high}) = \upsilon(r; d_{high}) = \int_{|\mathbf{s}|\le d_{high}^{-1}} \exp[-2\pi i(\mathbf{rs})]\,d\mathbf{s} = \int_{\mathbf{s}} U(\mathbf{s}; d_{high}^{-1}) \exp[-2\pi i(\mathbf{rs})]\,d\mathbf{s}. \quad (1)$$

Here, $r = |\mathbf{r}|$ and $U(\mathbf{s}; D)$ is a function equal to one inside a sphere of a radius $D = d_{high}^{-1}$, *i.e.* for $|\mathbf{s}| \le D$, and equal to zero outside. The spherically symmetric function (1) can be expressed through the three-dimensional interference function $G_3(t)$,

$$G_3(t) = 3\frac{\sin(t) - t\cos(t)}{t^3}, \quad (2)$$

as

$$\upsilon(r; d_{high}) = \frac{4\pi d_{high}^{-3}}{3} G_3(2\pi r d_{high}^{-1}) = \frac{4\pi d_{high}^{-3}}{3} G_3(2\pi x) \text{ with } x = r d_{high}^{-1} \quad (3)$$

(Fig. 1*a*). The function name in (2) refers to that in Rossmann & Blow (1962) and its index '3' is a reminder that it is calculated for three-dimensional images. Reminders of some of the known features of $G_3(t)$ and $\upsilon(r; d_{high})$ and the derivation of (3) are given in the Supplementary Material.[1]

Without a loss of generality, let us consider two point scatterers on the **Ox** axis with coordinates $x = 0$ and $x = l$, respectively. Their image calculated at the resolution d_{high} and shown along the interatomic vector (*i.e.* along **Ox**) is

$$\rho(r; d_{high}, l) = \upsilon(r; d_{high}) + \upsilon(r - l; d_{high}). \quad (4)$$

Depending on l and d_{high}, the two principal peaks of (4) may be seen either separately, if $l >> d_{high}$, or be merged into a single peak, if $l << d_{high}$. More formally, they are seen as separate peaks if the interval $(0, l)$ contains at least one point $0 < r_0 < I$ where the second derivative of (4) along this direction is positive, $\rho''(r_0; d_{high}, l) > 0$; they are seen as a single peak if $\rho''(r; d_{high}, l) < 0$ for all points in the interval $(0, l)$ (Fig. 1*b*). In the intermediate situation $\rho''(r; d_{high}, l) < 0$ for all points of the interval except one for which

$$\rho''(r_0; d_{high}, l_{min}) = 0. \quad (5)$$

This represents a critical condition, determination of which provides information about the limit for resolving two neighbouring scatterers, *i.e.* the value that we are aiming for. Owing to the symmetry of the problem, this point $r_0 = l/2$ corresponds to the inflection point of both of the composing images. In other words, the distance from the position of the scatterer in the origin to the inflection point $l_{infl,\upsilon}$ of its image (equal to r_0 in the example above),

$$\upsilon''(l_{infl}, \upsilon; d_{high}) = 0, \quad (6)$$

defines the minimum distance between two such scatterers as $l_{min} = 2l_{infl,\upsilon}$; the derivative in (6) is calculated along the interatomic vector. Numerical solution of (6) gives

$$l_{infl,\upsilon} \simeq 0.399 d_{high} \quad (7)$$

(see Supplementary Material). Therefore,

$$l_{min} = 2l_{infl,\upsilon} \simeq 0.798 d_{high} \simeq 0.8 d_{high} \quad (8)$$

with an accuracy that is largely acceptable for practical calculations. This value differs from $0.715 d_{high}$ as obtained by James (1948), from $0.712 d_{high}$ as obtained by Vaguine *et al.* (1999) and from $0.917 d_{high}$ as obtained by Stenkamp & Jensen (1984), although formally speaking the last value was not given for point scatterers. Fig. 1(*b*) shows, at a resolution $d_{high} = 1$ Å, the image of two point scatterers at several key distances, confirming the correctness of the estimate (8).

[1] Supplementary material has been deposited in the IUCr electronic archive (Reference: LV5042). Services for accessing this material are described at the back of the journal.

3. Effective resolutions of a set of reflections

3.1. Effective resolution: definition

In crystallography, when working with periodic electron-density distributions, the image of a point scatterer at the origin at a resolution d_{high} is calculated as a Fourier series,

$$\upsilon(\mathbf{r}) = \sum_{\mathbf{s}\in S} \exp[-2\pi i(\mathbf{rs})], \quad (9)$$

and no longer through the Fourier integral (1). Here, the sum is calculated for a set S of reflections complete at the resolution d_{high}, *i.e.* such that it contains all reflections with $d_{\mathbf{s}} \geq d_{high}$ or $|\mathbf{s}| \leq d_{high}^{-1}$. Except for extreme cases in which too few reflections are taken, the series (9) calculated with a complete data set of resolution d_{high} coincides with the integral (1) very precisely. Note that the Fourier synthesis (9) is calculated with unit Fourier coefficients for the available set of reflections (all phases are equal to zero).

In practice, the set S is often incomplete at the resolution d_{high}, *i.e.* some reflections inside the sphere $|\mathbf{s}| \leq d_{high}^{-1}$ are missing. In particular, the highest resolution shell often lacks a significant percentage of the data. One may intuitively expect (as confirmed in §3.2) that for such a data set the minimum distance is larger than for a complete data set with the same d_{high}. For the moment, we consider only isotropic data sets for which the distance to the inflection point is the same in all directions.

In order to estimate the minimum distance corresponding to an incomplete data set, we calculate the series (9) and determine the coordinate $l_{infl,\upsilon}$ of its first inflection point in any direction from the origin. As previously, the minimum distance is $l_{min} = 2l_{infl,\upsilon}$. We know from (7) that the same value of $l_{infl,\upsilon}$ would be obtained if a synthesis were calculated with a *complete* data set S_0 such that

$$\begin{cases} 2.5 l_{infl,\upsilon}(S) > d_{high}(S) & S \text{ is incomplete} \\ 2.5 l_{infl,\upsilon}(S_0) = d_{high}(S_0), & S_0 \text{ is complete} \end{cases}, \quad (10)$$

where $l_{infl,\upsilon}(S_0) = l_{infl,\upsilon}(S)$. Moreover, we know that $l_{infl,\upsilon}$ is a monotonous (linear) function of d_{high} for *complete* data sets, *i.e.* that for a complete data set its d_{high} is unambiguously defined by $l_{infl,\upsilon}$. This leads us to a formal definition.

Definition. For a given crystal, let the image (9) of the point scatterer be calculated with a (isotropic) data set S of reflections. Let $l_{infl,\upsilon}$ be the distance to its inflection point along a direction from the origin. Let the same distance $l_{infl,\upsilon}$ correspond to a data set S_0 complete at its highest resolution $d_{high}(S_0)$. We say that $d_{high}(S_0)$ is the *effective resolution* of the set S: $d_{eff}(S) = d_{high}(S_0)$.

The term 'effective resolution' was adopted following Weiss (2001). By definition, for a complete set S of reflections the effective resolution coincides with the highest resolution, $d_{eff}(S) = d_{high}(S)$. Knowing (7), one can express the effective resolution d_{eff} directly by the distance to the inflection point as

$$d_{eff}(S) = 2.5 l_{infl,\upsilon}(S) \geq d_{high}(S). \quad (11)$$

The inequality in (11) follows from (10). The next sections show that indeed $d_{\rm eff}(S) > d_{\rm high}(S)$ for incomplete sets and that the difference between these two values may be significant and characterizes the degree of incompleteness of S.

In conclusion to this section, we note that the inflection point can be found by an explicit analysis of the second derivative of (9) along the chosen direction from the origin (or any direction if the data are isotropic). However, for complete data sets it may be also identified from the condition

$$\upsilon(\mathbf{r}_{\rm infl}) \simeq \tfrac{1}{2}\upsilon(\mathbf{0}) \qquad (12)$$

using the corresponding feature of the interference function G_3 (see Supplementary Material).

3.2. Data sets incomplete at high resolution

Figs. 2(*a*)–2(*c*) show the results of a test illustrating the influence of data-set incompleteness on the effective resolution. A unit cell with parameters $a = b = c = 50$ Å was taken. A complete data set was taken at a resolution $d_{\rm high} = 1$ Å; this particular value was taken purely to simplify the calculations. Some reflections of this data set were then removed randomly with a probability $p(d_{\rm s})$ depending only on their resolution $d_{\rm s}$. The probability of removing a reflection was of the form

$$p(d_{\rm s}) = \begin{cases} 0 & \text{if } d_{\rm s} > d_{\rm reduce} \\ \left(\dfrac{d_{\rm s}^{-3} - d_{\rm reduce}^{-3}}{d_{\rm high}^{-3} - d_{\rm reduce}^{-3}}\right)^{m} & \text{if } d_{\rm high} \le d_{\rm s} \le d_{\rm reduce} \end{cases}. \qquad (13)$$

The scale uniform with d^{-3} means that if one splits the interval ($d_{\rm high}$, $d_{\rm reduce}$) into shells then all of them will contain roughly the same number of reflections. The parameter m in (13) defines the rate at which data incompleteness increasess with resolution. For $m = 1$ the proportion of removed reflections increases linearly from the lowest resolution shell, $p(d_{\rm reduce}) = 0$, to the highest resolution shell, $p(d_{\rm high}) = 1$. Values of $m > 1$ mean a sharper fall in completeness at the higher resolution, $d_{\rm high}$, and values of $m < 1$ mean a sharper fall in completeness at the lower resolution, $d_{\rm reduce}$, as illustrated in Fig. 2(*c*). The overall completeness $C_{\rm overall}$ is lower for smaller m (Fig. 2*b*).

Owing to the isotropy of the removed reflections, the effective value of the resolution $d_{\rm eff}$ (11) was the same for all three axes, as the calculation confirms. Fig. 2(*a*) shows $d_{\rm eff}$ as a function of m for $d_{\rm reduce}$ values of 1.2, 1.5 and 2.0 Å. This

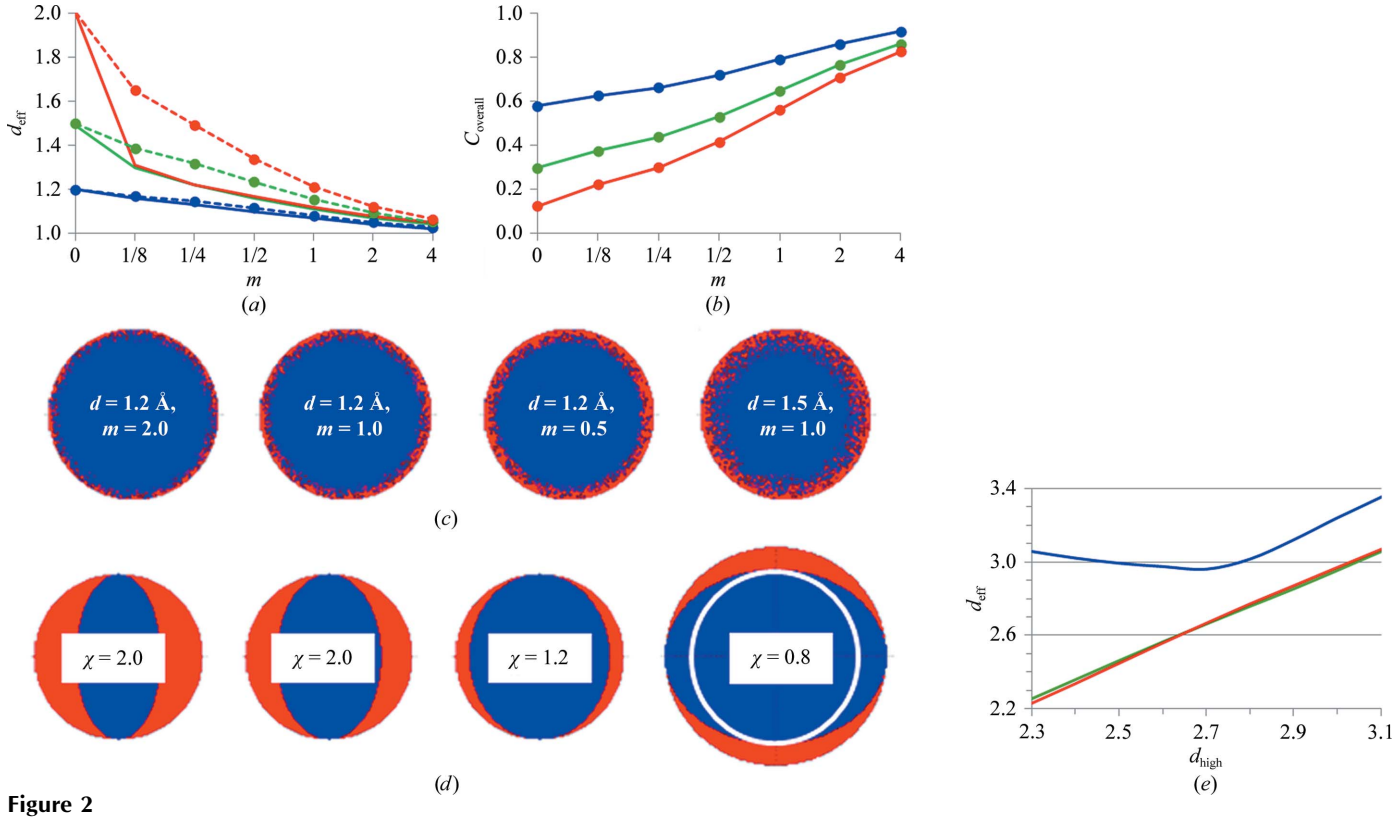

Figure 2
Analysis of data sets that are incomplete at high resolution. (*a*) Lines without markers show the effective resolution $d_{\rm eff} = 2.5 l_{{\rm infl},\upsilon}$ as a function of the incompleteness parameter m (13) for data sets with high resolution $d_{\rm high} = 1$ Å and that are incomplete in the shell ($d_{\rm high}$, $d_{\rm reduce}$). The blue curve is for $d_{\rm reduce} = 1.2$ Å, the green curve is for $d_{\rm reduce} = 1.5$ Å and the red curve is for $d_{\rm reduce} = 2.0$ Å. Larger m values correspond to a sharper fall at higher resolution. Dashed lines with markers show the resolution estimates $d_{\rm W} = d_{\rm high}{\rm C}^{-1/3}$ from Weiss (2001). (*b*) Overall completeness $C_{\rm overall}$ corresponding to the data sets in (*a*). (*c*) Present (in blue) and excluded (in red) reflections corresponding to several data sets from (*a*) are shown; *d* in the images stands for $d_{\rm reduce}$. (*d*) $l = 0$ section of reciprocal space showing present (in blue) and excluded (in red) reflections corresponding to anisotropic data sets. The white circle in the right image indicates the Ewald sphere for $d_{\rm high} = 1$ Å. (*e*) Effective resolution $d_{\rm eff}$ for PDB entry 4b47 calculated along the **a** (blue), **b** (green) and **c** (red) coordinate axes for different cutoff values $d_{\rm high}$. The program *FOBSCOM* (Urzhumtseva & Urzhumtsev, 2011) was used to prepare (*c*) and (*d*).

Table 1
The effective resolution $d_{\rm eff}$ (Å) along the coordinate axes calculated for different values of the parameter χ describing the anisotropy of the data set along the **a** axis.

See §3.3 for details. $C_{\rm overall}$ gives the ratio of the number of reflections in the data sets to the total number of reflections at a resolution of 1 Å.

	Elliptic			Segment		
χ	$d_{\rm eff}$, **a**	$d_{\rm eff}$, **b**, **c**	$C_{\rm overall}$	$d_{\rm eff}$, **a**	$d_{\rm eff}$, **b**, **c**	$C_{\rm overall}$
0.60	0.60	1.00	1.67	—	—	—
0.80	0.80	1.00	1.25	—	—	—
1.00	1.00	1.00	1.00	1.00	1.00	1.00
1.20	1.20	1.00	0.81	1.08	0.99	0.96
1.40	1.39	1.00	0.71	1.23	0.98	0.89
1.60	1.59	1.00	0.63	1.36	0.97	0.82
1.80	1.79	1.00	0.56	1.54	0.96	0.74
2.00	1.99	1.00	0.50	1.66	0.95	0.70
2.20	2.19	1.00	0.45	1.87	0.95	0.63
2.40	2.39	1.00	0.41	2.05	0.94	0.58

confirms that $d_{\rm eff} > d_{\rm high}$ for data sets incomplete at the higher resolution end and that the difference between these two values increases with the incompleteness and also depends on the manner in which the reflections are missing. Fig. 2(*a*) also shows the resolution estimates of Weiss (2001) using the overall completeness $C_{\rm overall}$ of the data,

$$d_{\rm Weiss} = d_{\rm high} C_{\rm overall}^{-1/3}. \qquad (14)$$

This estimate is correct for a sharp cutoff level, for which it was probably derived: $d_{\rm Weiss} = d_{\rm eff} = d_{\rm reduce}$ for $m = 0$. It also works well, $d_{\rm Weiss} \simeq d_{\rm eff}$, when the incomplete shell is relatively narrow, $d_{\rm reduce} \leq 1.2 d_{\rm high}$. However, for larger shells, *e.g.* $d_{\rm reduce} \geq 1.5 d_{\rm high}$, and for nontrivial incompleteness, $m > 0$, formula (14) significantly underestimates the effective resolution.

For the values of $d_{\rm high}$ that differ from 1 Å, equivalent results may be obtained by changing $d_{\rm reduce}$ proportionally.

The next sections analyze a different situation in which some reflections are missing not randomly but systematically.

3.3. Anisotropic data sets

If the diffraction data set is anisotropic, its effective resolution may be different in different directions. To simulate such a situation, a unit cell with parameters $a = b = c = 50$ Å was taken as in the previous test and a full set of reflections $\mathbf{s} = (hkl)$ with integer indices was calculated. Several anisotropic data sets shrunk along the **a*** axis in reciprocal space were obtained (the directions of **a*** and **a** coincide in this example). To do this, we removed the reflections that did not satisfy the condition

$$\frac{h^2}{(a/\chi)^2} + \frac{k^2}{b^2} + \frac{l^2}{c^2} \leq 1, \qquad (15)$$

taking various values of $\chi > 1$. A value of $\chi = 1$ corresponds to a complete 1 Å resolution data set. Fig. 2(*d*) shows the $l = 0$ section for several χ values. We also generated several data sets extended along the **a*** axis, completing the original set by reflections satisfying (15) with values $\chi < 1$ (right image in Fig. 2*d*). For each of these data sets we calculated the function $\upsilon(\mathbf{r})$ (9), which is no longer spherically symmetrical for $\chi \neq 1$. For each $\upsilon(\mathbf{r})$, we calculated the distance from the origin to its inflection point along the coordinate axes and then calculated the effective resolution $d_{\rm eff}$. As one would intuitively expect, this gave $d_{\rm eff} = 1$ Å for the **b** and **c** coordinate axes and $d_{\rm eff} = \chi$ for the **a** axis (Table 1). Analysis of the Fourier syntheses calculated with these sets of Fourier coefficients and with two scatterers placed varying the distance confirmed (results not shown) the obtained values of the minimum distance (8).

In the second test, we simulated a different type of anisotropy/incompleteness. We took the same data set complete at a resolution of 1 Å as above and removed the spherical segments $|h| > a/\chi$ for various values of $\chi > 1$. Table 1 shows several interesting observations. Firstly, for these data sets the conserved reflections somehow interpolate missing diffraction information along the **a*** axis, giving a corresponding effective resolution significantly higher than χ. Secondly, such a loss of data along **a*** and extending the image of the point scatterer in the **a** direction results in some shrinkage in the normal plane, slightly decreasing the corresponding minimum distance and improving the effective resolution for the **b** and **c** axes.

Obviously, apart from test examples such as that presented above, the principal directions of anisotropy are *a priori* unknown and will be defined first.

3.4. Anisotropy determination

In order to identify and characterize the anisotropy of a set of reflections, the asymmetry of the central peak of the function $\upsilon(\mathbf{r})$ can be studied. Firstly, we calculate the function (9) on a three-dimensional grid around the origin. Then, all points $\mathbf{r}_m = (x_m, y_m, z_m)$, $m = 1, \ldots, M$ with $\upsilon(\mathbf{r}) \geq \upsilon(\mathbf{0})/2$ are selected, referring to relation (12).

In many situations the shape of this set may be well approximated by an ellipsoid. For such data sets, the symmetrical matrix calculated over the selected points is

$$\begin{pmatrix} \sum_{m=1}^{M} x_m x_m & \sum_{m=1}^{M} x_m y_m & \sum_{m=1}^{M} x_m z_m \\ \sum_{m=1}^{M} x_m y_m & \sum_{m=1}^{M} y_m y_m & \sum_{m=1}^{M} y_m z_m \\ \sum_{m=1}^{M} x_m z_m & \sum_{m=1}^{M} y_m z_m & \sum_{m=1}^{M} z_m z_m \end{pmatrix}. \qquad (16)$$

Matrix (16) always has three real positive eigenvalues $0 < \lambda_1 \leq \lambda_2 \leq \lambda_3$. The corresponding eigenvectors $\mathbf{t}_1$, $\mathbf{t}_2$, $\mathbf{t}_3$ of (16) are orthogonal to each other and correspond to the principal directions of anisotropy. When $\lambda_1 = \lambda_2 = \lambda_3$ the data set is isotropic and $\upsilon(\mathbf{r})$ is the same for all directions. Otherwise, the highest and the lowest effective resolution $d_{\rm eff}$ will be for $\mathbf{t}_1$ and for $\mathbf{t}_3$, respectively. The functions $\upsilon_1(\mathbf{r})$ and $\upsilon_3(\mathbf{r})$ (9) are calculated along these directions, defining the respective distances $l_{\rm infl,min}$ and $l_{\rm infl,max}$ to their inflection points. The smallest and largest values of the effective resolution are $d_{\rm eff,min} = 2.5 l_{\rm infl,min}$ and $d_{\rm eff,max} = 2.5 l_{\rm infl,max}$.

In a more general situation, the surface of the selected set of points is studied (the technical details and corresponding algorithms will be described elsewhere) and the points with

the smallest and largest distance $l_{\rm infl,min}$ and $l_{\rm infl,max}$ from the surface point to the origin define the two extreme effective resolutions as above. The anisotropy of the data set can be characterized numerically by the ratio

$$R_{\rm aniso} = l_{\rm infl,min} : l_{\rm infl,max}. \qquad (17)$$

§5 below gives several practical examples of such calculations for isotropic and anisotropic cases.

§§3.3 and 3.4 show that technically there is no principal difference in analyzing anisotropic cases except that the effective resolution will vary from one direction to another. Below, for simplicity we suppose that the data sets are isotropic unless mentioned explicitly. In the next section we analyze another typical example in which some diffraction data are systematically missing.

3.5. Data sets that are incomplete at low resolution

In practice, most experimental data sets lack low-resolution reflections: those below some cutoff level $d \geq d_{\rm low}$. The value of $d_{\rm low}$ may vary from 6 to 10 Å (mostly for old data sets) to very large values. It is well known that excluding such data increases the apparent resolution. The formal definition of the effective resolution, as introduced here, allows measurement of the practical impact of such a cutoff on three-dimensional crystallographic syntheses.

Indeed, excluding these data modifies the image of a point scatterer to

$$\begin{aligned}\rho(r) &= \upsilon(r; d_{\rm high}) - \upsilon(r; d_{\rm low}) \\ &= \frac{4\pi d_{\rm high}^{-3}}{3} G_3(2\pi r d_{\rm high}^{-1}) - \frac{4\pi d_{\rm low}^{-3}}{3} G_3(2\pi r d_{\rm low}^{-1}) \\ &= \frac{4\pi d_{\rm high}^{-3}}{3}\left[G_3(2\pi r d_{\rm high}^{-1}) - \left(\frac{d_{\rm high}}{d_{\rm low}}\right)^3 G_3(2\pi r d_{\rm low}^{-1})\right]. \qquad (18)\end{aligned}$$

For the $d_{\rm low}/d_{\rm high}$ values usual nowadays in crystallography, *e.g.* about 10 or larger, the correcting term in brackets in (18) is negligibly small (Fig. 3*a*). As a consequence, the distance to the inflection point of (18) and the corresponding effective resolution $d_{\rm eff}$ are quite close to those calculated for the complete data set. For example, for $d_{\rm low} \geq 4d_{\rm high}$ the difference between $d_{\rm eff}$ and $d_{\rm high}$ is below 0.5% and may be neglected in practice. In fact, even when the correcting term is more significant its influence on the coordinate of the inflection point is weak (Fig. 3*c*). When $d_{\rm cut} \rightarrow d_{\rm high}$, (18) approaches $\cos(2\pi r d_{\rm high}^{-1})$, $l_{{\rm infl},\upsilon} \rightarrow 0.25 d_{\rm high}$ and $d_{\rm eff} \rightarrow 2.5 l_{{\rm infl},\upsilon} = 0.625 d_{\rm high}$.

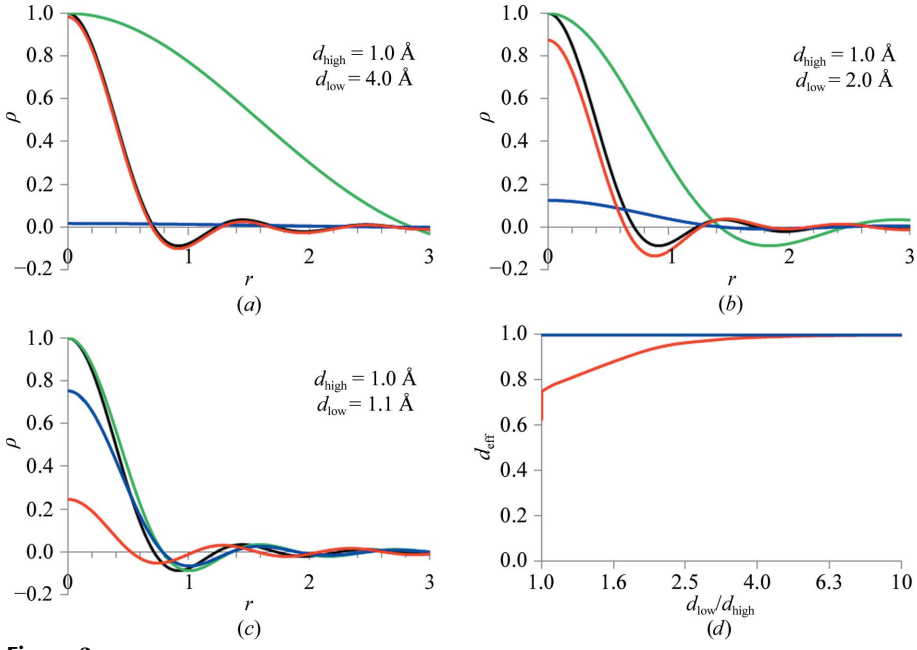

Figure 3
Tests with data sets that are incomplete at low resolution. The image of a point scatterer at $d_{\rm high}$ = 1 Å when all data are included is shown in black. Its image at $d_{\rm low}$, nonscaled is shown in green and that scaled by $(d_{\rm high}/d_{\rm low})^3$ is shown in blue. The resulting image (18) at $d_{\rm high}$ = 1 Å when the data below the resolution $d_{\rm low}$ are excluded is shown in red. Images are shown for (*a*) $d_{\rm low}$ = 4.0 Å, (*b*) $d_{\rm low}$ = 2.0 Å and (*c*) $d_{\rm low}$ = 1.1 Å. While the image changes significantly in (*c*), the coordinate of the inflection point changes much less. (*d*) Variation of the effective resolution $d_{\rm eff}$ (red curve) shown for $d_{\rm high}$ = 1 Å as a function of the low-resolution cutoff $d_{\rm low}$ (taken on a logarithmic scale; Urzhumtsev *et al.*, 2009). For reference, the blue curve shows the $d_{\rm eff}$ = 1 Å value when all low-resolution data are present.

This transition is very sharp, and even for $d_{\rm low} = 1.001 d_{\rm high}$ the value of $d_{\rm eff} \simeq 0.74 d_{\rm high}$ (Fig. 3*d*).

While such exclusion of low-resolution reflections formally increases the resolution, it destroys the crystallographic image (see, for example, Urzhumtsev, 1991) and is counterproductive. This remark refers to the questions about accuracy, level of details and the amount of structural information in the Fourier syntheses as mentioned in §1.

4. Minimum distance for non-point scatterers

4.1. Atomic structures and *B* factors

The effective resolution $d_{\rm eff}$ defined through the minimum distance for point scatterers characterizes the set of reflections S and does not refer to the corresponding intensity values $\{I_{\mathbf{s}}, \mathbf{s} \in S\}$. These intensities describe a particular structure composed from atoms of particular shape and with some uncertainties in their positions. These uncertainties are modelled by the atomic displacement factor, also known as the *B* factor. In fact, these values characterize not only uncertainties in atomic positions owing to thermal motion but also statistical uncertainty over the different unit cells in the crystal and different kinds of experimental errors including those in the data. A faster decrease in the mean intensity $\langle I_{\mathbf{s}} \rangle$ with resolution reflects a larger value of the *B* factor typical for the structure (Wilson, 1949). The Wilson factor $B_{\rm Wilson}$ may be

referred to as a typical value for the given structure, $B_{type} = B_{Wilson}$.

A given structure may contain atoms with different B values. Owing to atomic shape and positional uncertainties, the minimum distance for atoms of the structure is larger than $0.8d_{eff}$ for the point scatterers and increases with B. The values of individual B factors are unknown before the structure is solved; also, changing B for a couple of atoms will change the diffraction intensity values only very marginally. Therefore, calculating the minimum distance for atoms with a particular B value, for example B close to 0, is not specific for a given set of intensities and is statistically not representative. In contrast, the minimum distance calculated for the value B_{type}, which is typical for the structure and represents the overall fall of intensities with resolution, may be a data-set characteristic complementary to d_{eff}. Below, we analyze this characteristic accurately and show a practical way to calculate it.

4.2. Finite-resolution image of a Gaussian peak

We start from an analysis of artificial structures composed of isotropic Gaussian scatterers, which is the same as point scatterers with harmonic uncertainties in their positions characterized by the parameter B. Fourier coefficients for such a scatterer placed at the origin obey spherical symmetry and can be presented as

$$\mathbf{F}(\mathbf{s}) = \mathbf{F}(|\mathbf{s}|) = \mathbf{F}(s) = \exp(-\tfrac{1}{4}Bs^2). \qquad (19)$$

The exact shape of the corresponding Gaussian with no resolution cutoff is then

$$\rho_{G,0}(\mathbf{r}; B) = \left(\frac{4\pi}{B}\right)^{3/2} \exp\left[-\frac{1}{B}(2\pi|\mathbf{r}|)^2\right]. \qquad (20)$$

It may be noted that its image $\rho_G(\mathbf{r}; B, d_{high})$ at a resolution d_{high},

$$\begin{aligned}\rho_G(\mathbf{r}; B, d_{high}) &= \int_{|\mathbf{s}|\le d_{high}^{-1}} \exp(-\tfrac{1}{4}Bs^2)\exp[-2\pi i(\mathbf{rs})]\,\mathrm{d}\mathbf{s} \\ &= \int_{|\mathbf{t}|\le 1} \exp\left(-\frac{1}{4}\frac{B}{d_{high}^2}t^2\right)\exp\left[-2\pi i\left(\frac{\mathbf{r}}{d_{high}}\mathbf{t}\right)\right]\frac{\mathrm{d}\mathbf{t}}{d_{high}^3} \\ &= d_{high}^{-3}\rho_G(d_{high}^{-1}\mathbf{r}; d_{high}^{-2}B, 1) \qquad (21)\end{aligned}$$

can be expressed through its image $\rho_G(\mathbf{r}; B, 1)$ at resolution 1 Å by linear rescaling. [In (21), an internal variable $\mathbf{t} = \mathbf{s}d_{high}$ was introduced for integration]. The last formula makes it sufficient to analyze the Gaussian images only at 1 Å and then transfer the results to any resolution d_{high} by rescaling (21).

Fig. 4(*a*) shows that at $d_{high} = 1$ Å for relatively small B (up to 1 Å^2) the image of a Gaussian practically coincides with that of a point scatterer, $\upsilon(\mathbf{r}; d_{high})$; this happens because the decrease in the atomic factor (19) owing to harmonic disorder is insignificant in comparison with a sharp resolution cutoff (Fig. 4*e*). Inversely, for $B \simeq 25$–30 Å^2 or larger, the image of (20) practically coincides with itself (Fig. 4*d*) since the interference function is very sharp in comparison with the Gaussian and the resolution cutoff makes practically no contribution (Fig. 4*e*). The same result leads to the fact that, for example, an image of a Gaussian with $B = 30$ Å^2 is practically not affected by a resolution cutoff higher than 1.2 Å (Fig. 4*f*). For intermediate B values (Figs. 4*b* and 4*c*) the two factors, a sharp

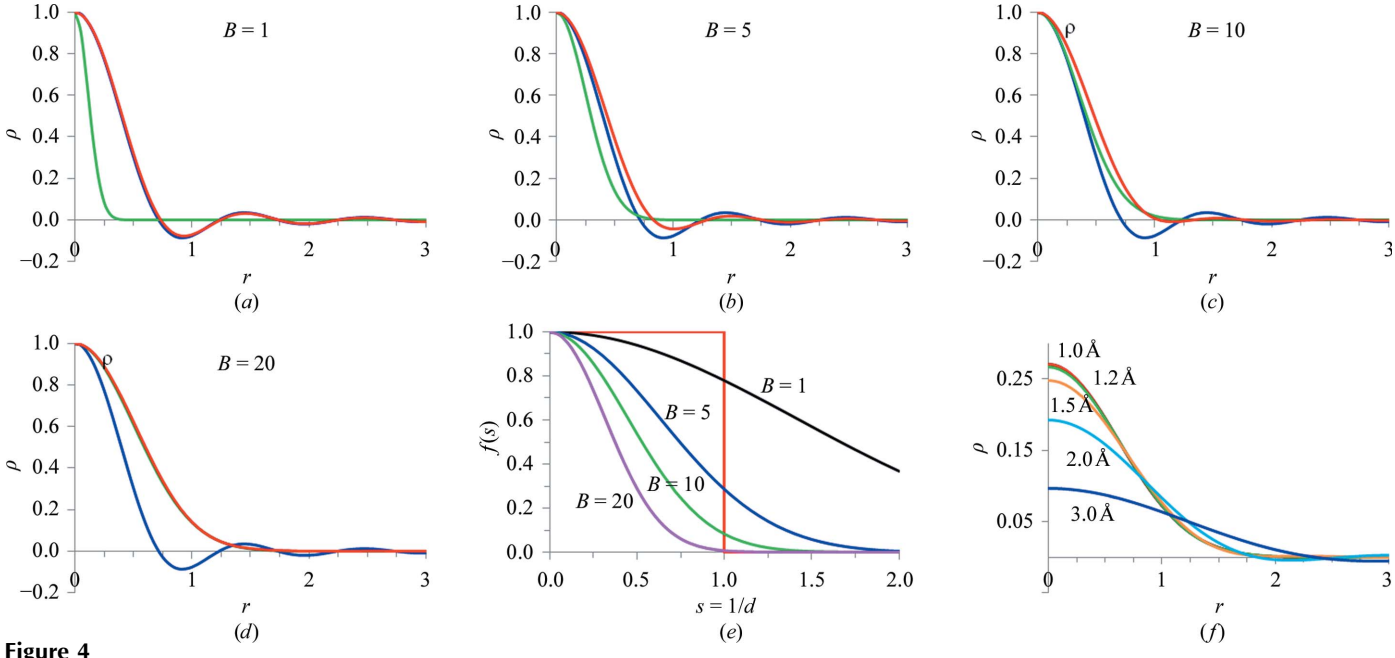

Figure 4
Fourier images of a Gaussian peak. A Gaussian peak (in green) with B value equal to 1 Å^2 (*a*), 5 Å^2 (*b*), 10 Å^2 (*c*), 20 Å^2 (*d*) and its Fourier image (red) in the synthesis of a resolution $d_{high} = 1$ Å. The blue curve shows the corresponding interference function. The image of the Gaussian practically coincides with the interference function in (*a*) and with the Gaussian itself in (*d*). (*e*) Sharp resolution cutoff at $d_{high} = 1$ Å (red) and the function $\exp(-\frac{1}{4}Bs^2)$ corresponding to different B values (black for 1 Å^2, blue for 5 Å^2, green for 10 Å^2 and magenta for 20 Å^2). (*f*) Fourier image of a Gaussian with $B = 30$ Å^2 at several resolutions d_{high}. For resolutions higher than 1.2 Å the image practically coincides with the peak itself shown in black; the peak is practically hidden behind the curves for $d_{high} = 1.0$ Å (red) and $d_{high} = 1.2$ Å (green).

Fourier coefficient resolution cutoff and a smooth decay owing to harmonic disorder (Fig. 4*e*), interact in a more complicated way.

Knowing the image of a Gaussian at a given resolution, we may analyze the corresponding minimum distance.

4.3. Minimum distance for equal Gaussians

Let us consider an artificial crystal structure composed of Gaussians at positions $\mathbf{r}_n$, $n = 1, \ldots, N$ and with the same B values. When no resolution cutoff is applied, $l_{\mathrm{infl,G}}(B; 0)$ for two Gaussians is known analytically (see Supplementary Material) to be

$$l_{\mathrm{infl,G}}(B; 0) = \left(\frac{B}{8\pi^2}\right)^{1/2} \simeq 0.1125 B^{1/2}. \qquad (22)$$

Fig. 5 is an image of such a structure at resolution $d_{\mathrm{eff}} = d_{\mathrm{high}}$ (let us consider only complete data sets) is calculated as the Fourier sum,

$$\rho(\mathbf{r}; B, d_{\mathrm{high}}) = \sum_{\mathbf{s}} \mathbf{F}_{\mathbf{s}} \exp[-2\pi i(\mathbf{rs})], \quad |\mathbf{s}| \leq d_{\mathrm{high}}^{-1}. \qquad (23)$$

Here, the Fourier coefficients $\mathbf{F}_{\mathbf{s}}$ reflect not only the positions of the scatterers but also their uncertainty as described as above by B,

$$\begin{aligned}\mathbf{F}_{\mathbf{s}} &= \sum_{n=1}^{N} \exp(-\tfrac{1}{4}Bs^2) \exp(2\pi i \mathbf{r}_n \mathbf{s}) \\ &= \exp(-\tfrac{1}{4}Bs^2) \sum_{n=1}^{N} \exp(2\pi i \mathbf{r}_n \mathbf{s}). \end{aligned} \qquad (24)$$

As previously, the minimum distance for a Gaussian scatterer is expressed through the distance $l_{\mathrm{infl,G}}(B; d_{\mathrm{high}})$ from the centre to the inflection point of its image,

$$l_{\mathrm{min,G}}(B; d_{\mathrm{high}}) = 2 l_{\mathrm{infl,G}}(B; d_{\mathrm{high}}). \qquad (25)$$

For the data set complete at resolution $d_{\mathrm{high}} = 1$ Å, the value of $l_{\mathrm{infl,G}}(B; 1)$ can be calculated numerically; Fig. 5(*b*) shows it as a function of $B^{1/2}$. For d_{high} different from 1 Å, the property (21) leads to

$$l_{\mathrm{infl,G}}(B; d_{\mathrm{high}}) = d_{\mathrm{high}} l_{\mathrm{infl,G}}(d_{\mathrm{eff}}^{-2} B; 1). \qquad (26)$$

In a general case, the effective geometric resolution d_{eff} should be used in (26) instead of d_{high} if eventually incomplete data sets are used.

As shown above, for large $d_{\mathrm{eff}}^{-2}B$ starting from ~25 the resolution cutoff does not play a role, resulting in

$$l_{\mathrm{infl,G}}(d_{\mathrm{eff}}^{-2} B; 1) \simeq l_{\mathrm{infl,G}}(d_{\mathrm{eff}}^{-2} B; 0). \qquad (27)$$

Inversely, for $d_{\mathrm{eff}}^{-2}B$ below approximately 1 the distance to the inflection point practically coincides with that for point scatterers, $l_{\mathrm{infl,G}} \simeq l_{\mathrm{infl},\upsilon}$. For intermediate B values, which are most typical for macromolecular structures, the function is more general, showing that the contributions from the interference function and from the Gaussian are non-additive, disagreeing with the basic proposition of Vaguine *et al.* (1999).

One may expect the situation to be even more complicated for other scatterers such as atoms.

4.4. Minimum distance for macromolecular atoms

In order to use a similar procedure to study the minimum distance for atomic structures, we need to address several features. Firstly, atomic scattering factors (*i.e.* the Fourier transform of an isolated atom placed at the origin with $B = 0$) can be accurately approximated by a Gaussian function only at relatively low resolutions of ~3 Å or lower (Agarwal, 1978). Therefore, the results obtained above cannot be used directly to estimate the minimum distance between two atoms with a given B value. Secondly, one might expect the distance l_{infl} to be different for different types of atoms.

In the absence of a resolution cutoff, $d_{\mathrm{high}} = 0$, the second derivative of the atomic electron density can be calculated analytically using a multi-Gaussian approximation of scattering factors (Brown *et al.*, 2006),

$$f(s) = \sum_{j=1}^{5} C_j \exp(-\tfrac{1}{4}B_j s^2), \quad \text{with } B_5 = 0. \qquad (28)$$

Let us introduce as previously the distance $l_{\mathrm{infl}}(B, 0)$ to the inflection point of the electron density of some atom with atomic displacement factor B; the value '0' stands for $d_{\mathrm{high}} = 0$. In spite of the difference in the coefficients in (28) for different atoms, the distance $l_{\mathrm{infl}}(B, 0)$ is fairly similar for the principal macromolecular atomic types, with the exception of hydrogen.

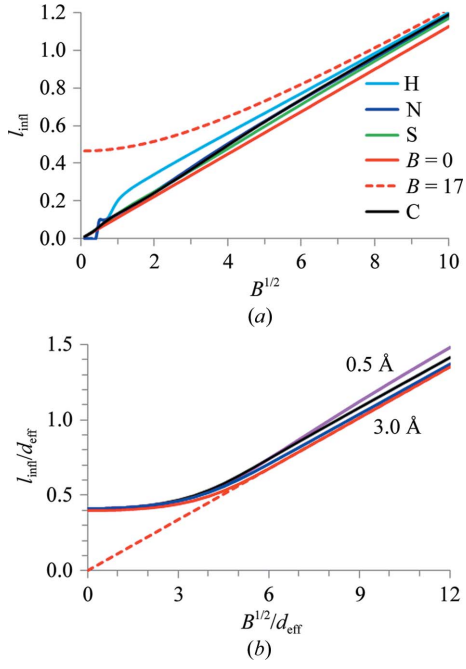

Figure 5
Distance to the inflection point of atomic images for different B values. (*a*) Distance from the atomic centre to the inflection point of the atomic electron density; no resolution cutoff was applied. Curves for hydrogen (light blue), nitrogen (blue), carbon (black) and sulfur (green) are shown. For comparison, red curves show the distance to the inflection point for a Gaussian scatterer with $B^0 = 0$ (solid line) and with $B^0 = 17$ Å^2 (dashed line). (*b*) Normalized distance $l_{\mathrm{infl}}/d_{\mathrm{eff}}$ to the inflection point of the Fourier image shown, as a function of $B^{1/2}/d_{\mathrm{eff}}$, for a Gaussian scatterer (red) and for a C atom at resolutions d_{eff} equal to 0.5 Å (magenta), 1.0 Å (black) and 2.0 Å (blue). The dashed red line shows the asymptote $(B/8\pi^2)^{1/2}$.

The variation is less than 0.01 Å for carbon, oxygen and nitrogen (except for B values below $\sim$0.5 Å^2, which do not exist in practical macromolecular studies), with a maximal difference of $\sim$0.03 Å for some particular values between carbon and phosphorus or sulfur (Fig. 5*a*). Similarly, when the resolution cutoff $d_{\rm high}$ was applied (we checked the range from 0.5 to 10 Å), the difference in $l_{\rm infl}(B, d_{\rm high})$ between different atomic types was very small. For hydrogen, the minimum distance is larger than for other atoms (Fig. 5*a*) and is less characteristic for a crystal. In contrast, a C atom can be considered as a typical scatterer for our goals and the corresponding minimum distance is studied in the next sections.

4.5. Minimum distance for a C atom

We started by calculating the distance $l_{\rm infl,C}(B, d_{\rm eff})$ to the inflection point of the images of carbon for different B-factor values and at different resolutions considering the complete data sets, $d_{\rm eff} = d_{\rm high}$. We referred to (26), which suggests presenting the distance to the inflection point as a function of $B^{1/2}$,

$$L_{\rm G}(B^{1/2}) = l_{\rm infl,G}(B; 1). \quad (29)$$

With (29), the normalized distance

$$d_{\rm eff}^{-1} l_{\rm infl,G}(B; d_{\rm eff}) = l_{\rm infl,G}(d_{\rm eff}^{-2} B; 1) = L_{\rm G}(d_{\rm eff}^{-1} B^{1/2}) \quad (30)$$

to the inflection point for Gaussian scatterers is independent of the resolution when taken as a function of $B^{1/2}/d_{\rm eff}$. This suggests also analyzing the normalized distance $d_{\rm eff}^{-1} l_{\rm infl,C}(B; d_{\rm eff})$ for C atoms as a function of $B^{1/2}/d_{\rm eff}$,

$$L_{\rm C}(d_{\rm eff}^{-1} B^{1/2}; d_{\rm eff}) = d_{\rm eff}^{-1} l_{\rm infl,C}(B; d_{\rm eff}). \quad (31)$$

Fig. 5(*b*) shows that the function (31) resembles (30) but differs in that it depends on the resolution $d_{\rm eff}$ when it varies from high values up to $d_{\rm eff} \simeq 3$ Å. The difference between the curves may be significant and may reach $\sim$0.15 Å/$d_{\rm eff}$. For large $B^{1/2}/d_{\rm eff}$ values the curves for carbon approach the curve for the Gaussian with $B_0 \simeq 17$ Å^2. Therefore, the function (31) tabulated for several resolutions, say from 0.5 to 3 Å, and interpolated when necessary allows the minimum distance for non-H atoms of a macromolecular structure, in particular for carbon, to be recalculated with any B value and at any resolution $d_{\rm eff}$.

4.6. Optical resolution for a crystal structure: revisited

Obviously, except for some test cases, the atoms of a given structure have different B values. Defining some 'typical' value $B_{\rm type}$ and associating it with a C atom allows a typical scatterer to be defined and used as a 'ruler' to calculate the minimum distance which characterizes the set of corresponding intensities. The goal is to perform this *a priori* and independently of any particular atomic model built using these data (*i.e.* independently of the structure-solution method, of the phase values *etc.*).

The set of intensities $\{I_{\mathbf{s}}, \mathbf{s} \in S\}$ defines a particular value of the Wilson B factor $B_{\rm Wilson}$ which may be referred to as a typical value for the given structure, $B_{\rm type} = B_{\rm Wilson}$, even when its accurate definition is sometimes difficult, as mentioned in particular by Vaguine *et al.* (1999). Let $l_{\rm infl,C}(B_{\rm type}; d_{\rm eff})$ be the distance to the inflection point for C atoms with $B_{\rm type}$. The value $2l_{\rm infl,C}(B_{\rm type}; d_{\rm eff})$ defines the minimum distance at which two such atoms are still seen as separate peaks in the 'ideal' synthesis corresponding to the given set of reflections. This value is similar to the minimum distance for other non-H atoms with the same $B_{\rm type}$. In other words, this is a characteristic value for a synthesis with the Fourier amplitudes corresponding to the given intensities and with the ideal phase values $\{\varphi_{\mathbf{s}}\}$, if they were known. In some way, it shows the 'potential' of this set of diffraction intensities.

Previously, Vaguine *et al.* (1999) discussed 'the expected minimum distance between two resolved peaks in the electron density maps'. Their considerations are essentially based on a Gaussian approximation of both the atomic shape and the central peak of the interference function and on the additivity of the contributions from these two sources to blurring the atomic images. As we showed in §2.2, approximation of the central peak of the interference function by a Gaussian leads to an incorrect estimate of the minimum distance between two point scatterers. Also, we illustrated 'non-Gaussian behaviour' of the distance to the inflection point for the atoms of a macromolecular structure. All together, this shows that the idea of additivity of these two contributions is incorrect. These authors had the very nice idea of using the Patterson peak at the origin to estimate the B factor for the structure instead of $B_{\rm Wilson}$. This approach is very promising, but its current realisation is inaccurate (in particular by also being based on the 'Gaussian' hypothesis discussed above). All of these inconsistencies may lead to some surprising results when, for example, the resolution calculated with an incomplete data set is higher than that for the same data set but complete (see Fig. 2 of Vaguine *et al.*, 1999).

Considerations of the previous section and in particular (31) allow an accurate definition and calculation of the minimal distance for the 'typical' (C) atom for the given set of intensities as

$$l_{\rm min,C}(B_{\rm type}) = 2l_{\rm infl,C}(B_{\rm type}; d_{\rm eff}). \quad (32)$$

Both this work and that of Vaguine *et al.* (1999) address the same goal. To avoid introducing new terms and in spite of the differences between the two approaches and inaccuracies in the previous definition, we suggest referring to the minimum distance which we define above (32) as the *optical resolution* (Vaguine *et al.*, 1999) $d_{\rm opt} = l_{\rm min,C}(B_{\rm type})$.

Naturally, the particular structure may contain some atoms with B close to 0 for which the minimum distance is smaller than (32) and is only slightly larger than $0.8d_{\rm eff}$ for the point scatterers. However, this shorter distance will be a feature of this particular structure and not of the set of intensities, which will not practically change if we replace these two small B values by larger values, for example by $B_{\rm Wilson}$.

4.7. The procedure to define the effective and optical resolutions

Summarizing the results above, this procedure is suggested to characterize how detailed the set of intensities $\{I_{\mathbf{s}}, \mathbf{s} \in S\}$ is. This procedure consists of several steps.

(i) The function (9) is calculated along the coordinate axes; for each of them the distance to the inflection point of (9) is calculated.

(ii) The maximum of the values calculated in the previous step is used as an estimate of the radius of the sphere inside which the function (9) is calculated in a three-dimensional relatively fine grid.

(iii) All points around the origin such that $\upsilon(\mathbf{r}) \geq \frac{1}{2}\upsilon(\mathbf{0})$ are selected; the points on the surface of this set with the shortest, $l_{\rm infl,min}$, and the longest, $l_{\rm infl,max}$, distance to the origin are determined.

(iv) The minimum and maximum effective resolutions of the set of reflections are defined as $d_{\rm eff,min} = 2.5 l_{\rm infl,min}$ and $d_{\rm eff,max} = 2.5 l_{\rm infl,max}$; the anisotropy factor is calculated as $R_{\rm aniso} = l_{\rm infl,min}{:}l_{\rm infl,max}$.

(v) A value $B_{\rm type}$ of the B factor typical for the given structure is defined for the set of intensities; the Wilson factor $B_{\rm Wilson}$ can be taken as such a value.

(vi) The optical resolution (32) is calculated with $B_{\rm type}$ for the minimum and maximum effective resolutions.

(vii) Referring to (8), the value $d_{\rm high}$ may be compared with $d_{\rm eff}$ to determine the loss of resolution owing to data incompleteness and with $1.25 d_{\rm opt}$ to determine its further loss owing to the various disorders characterized by the B factor.

The procedure can be applied as soon as a set of intensities is selected for structure solution. §5 gives several practical examples of such calculations. We reiterate that this procedure is not aimed at the selection of intensities from raw data but to characterize the selected data. Also, this procedure does not estimate the resolution of particular Fourier syntheses, in which several other factors play a role, making this a separate problem and project.

This procedure has been implemented as a command-line test program which can be obtained by request from the authors. A GUI version of the program is under development.

5. Application to calculated and experimental data sets

5.1. Test data

To verify the suggested approach, we used the data for the recently determined crystal structures of translation initiation factor 2 (IF2; Simonetti *et al.*, 2013) and its complexes. The crystals were isomorphous to each other but diffracted to different resolutions and differed from each other in the level of anisotropy, in the data completeness and in some other characteristics (Table 2). The crystals belonged to space group $P2_12_12_1$, with unit-cell parameters $a \simeq 45$, $b \simeq 62$, $c \simeq 160$–162 Å and contained a single protein molecule (fragment 1–363 of IF2) in the asymmetric unit. The crystal packing of the IF2 molecules showed a large cavity around the origin. The corresponding atomic models were refined using *phenix.refine* (Afonine *et al.*, 2012); the corresponding PDB entries are 4b3x, 4b47 and 4b48. The estimated factor $B_{\rm Wilson}$ was very close to the most frequent B values of the IF2 models as shown by *PHENIX* (Adams *et al.*, 2010).

Table 2
Examples of the effective and optical resolutions.

Experimental diffraction data were collected by the authors (PDB entries 4b3x, 4b47 and 4b48; IF2 data) or taken from the PDB. $C_{\rm overall}$ and $C_{\rm high}$ stand for the data completeness overall and in the highest resolution shell. $d_{\rm eff}$ shows the smallest and largest effective resolutions in different directions; the anisotropy index $R_{\rm aniso}$ is their ratio as defined in the text. $d_{\rm opt}$ shows the minimum distances for the C atoms calculated with $B_{\rm Wilson}$. The last column is given for comparison with $d_{\rm high}$ and $d_{\rm eff}$ (§4.7). The calculations for PDB entry 1a00 were performed at two different resolutions. The optical resolution for IF2 was calculated both with isotropic $B_{\rm Wilson}$ and with $B_{\rm anisotropic}$ after addition of the anisotropic corrections obtained after model refinement [indicated by #; the minimal and maximal values of the correcting coefficients in (33) are given].

PDB code	$d_{\rm high}$ (Å)	$C_{\rm overall}$	$C_{\rm high}$	$d_{\rm eff}$ (Å)	$R_{\rm aniso}$	$B_{\rm Wilson}$ ($B_{\rm anisotropic}$) (Å^2)	$d_{\rm opt}$ (Å)	$1.25 \times d_{\rm opt}$ (Å)
4b3x	1.95	1.00	1.00	1.94/1.94	1.00	29	1.75/1.75	2.19/2.19
4b3x#						(−5/+10)	1.72/1.82	2.16/2.28
4b48	2.80	1.00	1.00	2.79/2.79	1.00	68	2.51/2.52	3.14/3.15
4b48#						(−27/+14)	2.40/2.59	3.00/3.23
2gc1	1.95	0.91	0.88	1.93/1.93	1.00	26	1.72/1.73	2.15/2.16
1yeq	2.61	0.91	0.32	2.67/2.69	0.99	22	2.24/2.25	2.80/2.82
2wit	3.35	0.91	0.34	3.40/3.58	0.95	64	2.92/3.05	3.65/3.81
3a9y	1.85	0.91	0.88	1.83/1.99	0.92	24	1.64/1.75	2.05/2.19
2whx	2.20	0.89	0.60	2.19/2.48	0.88	21	1.88/2.09	2.35/2.61
2ntc	2.40	0.90	0.60	2.35/2.70	0.87	53	2.17/2.39	2.71/2.99
1a00	2.03	0.94	0.71	2.01/2.22	0.91	22	1.75/1.90	2.19/2.38
1a00	1.73	0.67	0.06	1.73/2.26	0.77	22	1.56/1.94	1.95/2.42
4b47	2.30	0.76	0.50	2.23/3.06	0.73	45	2.04/2.60	2.55/3.26
4b47#						(−24/+19)	1.91/2.68	2.44/3.39

Additionally, several examples of experimental diffraction data were obtained from the PDB (Bernstein *et al.*, 1977; Berman *et al.*, 2000) according to the completeness statistic (P. Afonine, personal communication; Urzhumtseva *et al.*, 2009; Urzhumtseva & Urzhumtsev, 2011). For one of these data sets, 1a00, the structure was reported at a resolution of 2 Å while the data set deposited contained reflections with a resolution up to 1.73 Å; however, the completeness in the higher resolution shells was very low.

5.2. Effective resolution

Analysis of the effective resolution for several practically complete data sets showed that the suggested method estimates their effective resolution quite well (Table 2). These examples also illustrate the accuracy of the practical estimates as $\sim 0.01 d_{\rm eff}$.

For isotropic data sets that are incomplete in the highest resolution shell the estimated $d_{\rm eff}$ is close to $d_{\rm high}$ if the shell is very narrow and relatively complete. It becomes lower otherwise; for example, it decreased by about $\sim$0.1 Å for PDB entry 1yeq.

For anisotropic data sets the $d_{\rm eff,max}$ value may be significantly larger than $d_{\rm eff,min}$ (see, for example, PDB entries 2whx, 2ntc and 4b47). In most cases $d_{\rm eff,min}$ practically coincides with $d_{\rm high}$ even for highly anisotropic data.

The experimental data set for PDB entry 4b47 was analyzed in more detail. This data set failed to be completed owing to technical reasons, so that a large continuous region roughly corresponding to a cone around the **a*** axis of reciprocal space was unmeasured. Calculation showed that the effective resolution for the **b** and **c** axes was practically the same, with a marginally better value for the **c** axis. Naturally, the worst effective resolution, $d_{\rm eff,max}$, corresponded to the **a** axis. Two particular effects were observed.

Firstly, such very strong anisotropy made $d_{\rm eff,min}$ slightly higher than $d_{\rm high}$; a similar effect can also be observed for the 2ntc data (Table 2). Such behaviour has been discussed in §3.3.

Secondly, removing the highest resolution shell 2.3–2.7 Å makes the data set less anisotropic and improves $d_{\rm eff,max}$ by more than 0.1 Å (Fig. 2*e*). Obviously, this decreases the $d_{\rm eff,min}$ value and for practical studies syntheses at resolutions of both 2.3 and 2.7 Å can be helpful in such situations. A similar effect was observed for another highly anisotropic data set, 1a00.

5.3. Optical resolution: IF2 model data

As a first check, a series of Fourier syntheses was calculated with the IF2 model data (PDB entry 4b3x) at various resolutions. Each time, two C atoms with B = 35 Å^2 corresponding roughly to $B_{\rm Wilson}$ were artificially added to the model. They were placed into a cavity in the crystal lattice such that the contribution of neighbouring atoms was negligible and no bias from these neighbours was introduced (Fig. 6*a*). The distance $d = 2l_{\rm atoms}$ between the added C atoms varied with a step of 0.1 Å ($l_{\rm atoms}$ varied by 0.05 Å) and the Fourier syntheses were calculated with the model data for each such model. Two types of model were considered: with and without a bulk-solvent contribution added to the structure factors. This contribution was calculated using the flat mask solvent model (Jiang & Brünger, 1994) with scale factors $k_{\rm solv}$ = 0.31, $B_{\rm solv}$ = 31 Å^2 as obtained by *phenix.refine* (Afonine *et al.*, 2012) for the experimental data.

The Fourier syntheses were calculated using complete data sets at various resolutions $d_{\rm eff} = d_{\rm high}$ and were visualized using *PyMOL* (DeLano, 2002). Two distances were noted: a smaller distance when the two artificial atoms were still seen as a single peak and a larger distance when the two peaks were already clearly separated (Figs. 6*b*, 6*c* and 6*d*). The calculated values, which were practically independent of the orientation of the atomic pair in this case of isotropic data, were compared with the theoretical values $l_{\rm infl,C}(B = 35; d_{\rm eff})$.

Visual analysis of the model without bulk-solvent contribution confirmed the analytical estimates, with errors of the order of 0.01 Å/$d_{\rm eff}$ or below (Table 3). The observed values did not change significantly when the bulk-solvent contribution was added; some small changes can be explained by a tiny variation of the atomic environment in the cavity. This means that the addition of a special contribution such as that from the bulk solvent does not change the general scheme developed for structures composed of individual scatterers.

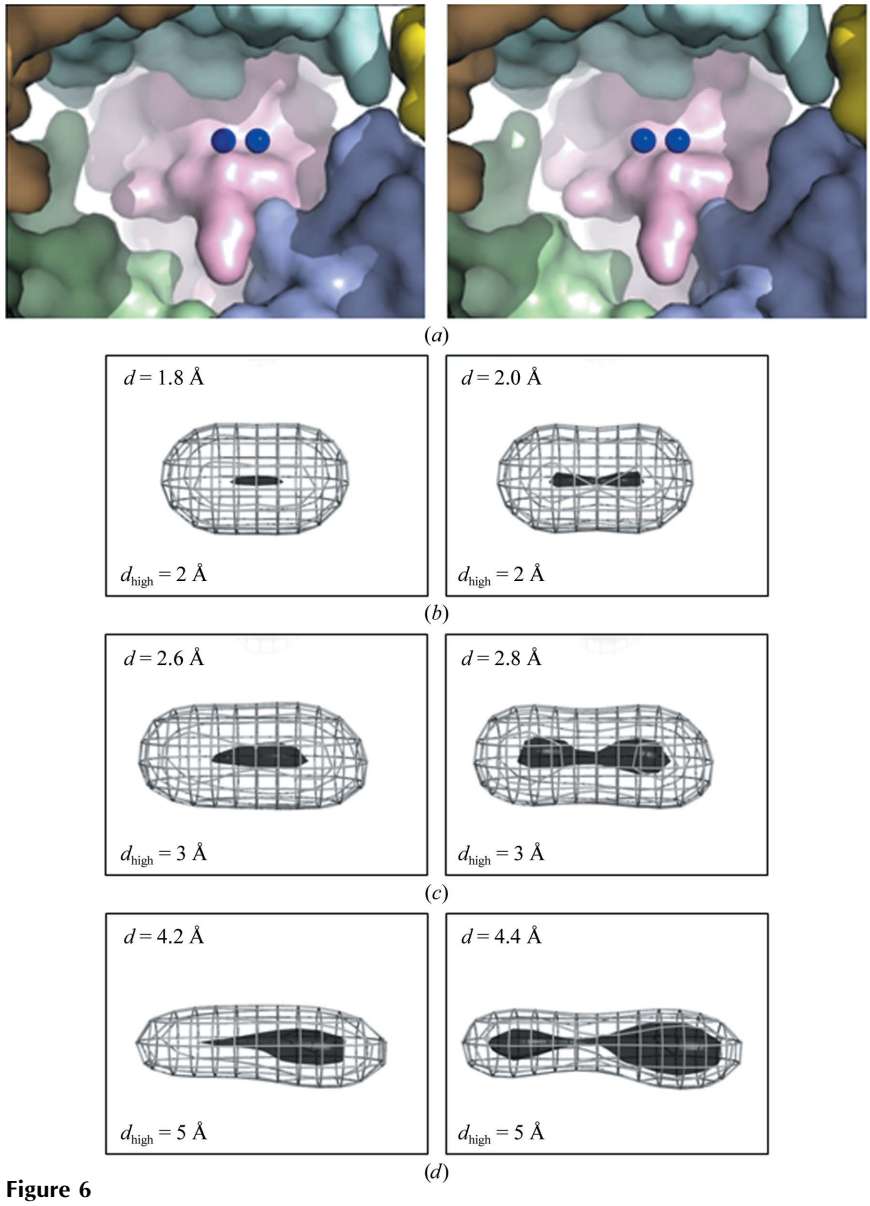

Figure 6
Separation of atomic peaks in crystallographic maps. (*a*) Stereoview of the IF2 intermolecular cavity with two artificial C atoms (blue spheres) with B = 35 Å^2 at a distance d equal to 2 Å; different molecules are in different colours. (*b*)–(*d*) Fourier maps corresponding to these atoms separated by the distance d given in the figures; the syntheses were calculated with the complete model data sets at resolutions $d_{\rm high}$ equal to 2 Å (*b*), 3 Å (*c*) and 5 Å (*d*). The left maps correspond to the interatomic distance when the atomic peaks are merged; the right maps are for the distance when the peaks are observed separately. The peak asymmetry in (*d*) is owing to the contribution of neighbouring atoms which becomes significant at this resolution.

Table 3
The distance (Å) from the peak to the inflection point of a typical scatterer calculated for different IF2 models at resolution d_{high} compared with the experimentally observed value.

The two values given for a visual analysis correspond to the half-distance $l_{atoms} = \frac{1}{2}d$ between the two atoms when they are observed to be merged and when they are clearly separated. See the text for further details.

	Theory, B = 35 Å^2	Map, visual, no bulk solvent	Map, visual, with bulk solvent
d_{high} (Å)	$l_{infl,C}$	l_{atoms}	l_{atoms}
2	0.90	0.90/0.95	0.95/1.00
3	1.25	1.25/1.35	1.30/1.40
4	1.65	1.65/1.70	1.65/1.70
5	2.10	2.15/2.20	1.90/1.95

The most significant change in the critical interatomic distance l_{atoms} is its decrease at d_{eff} = 5 Å by 0.25 Å. This may include both inaccuracies in the visual analysis and weak changes in the intermolecular cavity since the contribution from neighbouring atoms becomes more significant at a longer distance. This is also consistent with the fact that the low-resolution bulk-solvent structure factors are similar to those from the atomic model but have the opposite sign (see, for example, Urzhumtsev, 2000). This decreases the total structure-factor amplitudes at resolutions $d_{bulk} \simeq$ 6–7 Å or lower (Phillips, 1980), thus 'softly' suppressing them; this may slightly increase the minimum distance, similar to sharp low-resolution data suppression (§3.5). The variation is expected to be more significant for smaller d_{bulk}/d_{eff} ratios, as is found to be the case.

5.4. Optical resolution with experimental data

After verifying the physical significance of the optical resolution (32) using the IF2 model data, d_{opt} was calculated for several experimental data sets including those for IF2 and its complexes. While d_{opt} is a convenient measure to characterize the data set, as was performed in the previous section, the value $1.25d_{opt}$ is convenient for comparison with d_{high} and d_{eff}. Table 2 shows that for $B_{Wilson} \simeq 20$ Å^2 the difference $1.25d_{opt} - d_{eff}$ is of the order of 0.2 Å and for $B_{Wilson} \simeq 60$ Å^2 it may approach 0.3–0.4 Å (for example, PDB entries 2wit and 4b48). For the analyzed isotropic but incomplete data sets the maximum value of the difference $d_{eff} - d_{high}$ reached a smaller value of 0.08 Å. This reflects the fact that the high-resolution shells of diffraction data are traditionally excluded from structural analysis if they have low data completeness. Naturally, such a cutoff avoids confusing values of d_{high} but leads to a loss of diffraction information. We believe that these data could be used for structure analysis independently of their completeness if they satisfy other selection conditions. This might eventually improve the results and correctness of the model, while estimates of the 'real' effective resolution d_{eff} will not be influenced by these reflections.

Crystal disorder may be anisotropic, as reflected by the anisotropic correction to the overall B factor (see, for example, Afonine *et al.*, 2013). For our example of the IF2 structure such a correction was obtained *a posteriori* from model refinement. Here, for all three data sets the principal anisotropy axis coincided with the mutually orthogonal coordinate axes, reducing the anisotropic scale factor to a simplified form

$$K_{aniso} = \exp[-(B_{hh}h^2 + B_{kk}k^2 + B_{ll}l^2)/4] \qquad (33)$$

convenient for comparison with B_{Wilson}.

This diffraction anisotropy is complementary to and superimposes with the eventual anisotropy of the set of reflections, as the IF2 examples show (Table 3). Indeed, for the isotropic data sets 4b3x and 4b48 with different effective resolution d_{eff} and different B_{Wilson} values, this correction led to a significant variation in d_{opt}. It decreased in the **c** direction and increased following **a** compared with the values for the isotropic B_{Wilson}.

For the incomplete and anisotropic set of reflections (4b47) introduction of the anisotropic correction (33) amplified the already existing difference between d_{opt} in the **a** and **c** directions.

6. Discussion

The goal of this project was to introduce a strictly defined and physically meaningful characteristic to compare crystallographic diffraction data sets that are complete and incomplete. The suggested approach is based on an accurate calculation of the minimum distance between two scatterers when their images are still separated in the synthesis calculated with the given data set. Obviously the minimum distance depends on the scatterers, and two particular types are chosen: point immobile scatterers and C atoms with a B factor typical of the given structure.

Comparison of the calculated minimum distance for point scatterers with the theoretical values for the complete data sets defines the effective resolution d_{eff} of a data set. For each complete data set this value coincides with its highest resolution d_{high}, while for incomplete data sets it is larger and depends on the data-set completeness and on the shape of the regions in which data are lost. Such a minimum distance depends on the set of reflections and not on the corresponding intensities. One may say that d_{eff} characterizes the 'geometry' of the data set.

For anisotropic sets of reflections d_{eff} depends on the direction in which the minimum distance is calculated. The ratio of the smallest to the largest of these values gives an unambiguous numerical characteristic of anisotropy.

A similar analysis of the minimum distance for C atoms is based on the observation that such a distance depends only very weakly on the type of atoms composing the macromolecules (except hydrogen). This minimum distance, called the optical resolution d_{opt} following Vaguine *et al.* (1999), can be calculated as soon as a diffraction data set is processed. It does not require the knowledge of an atomic model and can characterize various data sets in a common way for the same or for different structures. This characteristic describing how detailed these data sets are is complementary to the common

characteristics describing the accuracy of the data (Karplus & Diederichs, 2012, and references therein).

The current approach considers the Wilson factor $B_{\rm Wilson}$ as a typical value for these atoms. As a consequence, the minimum distance estimate is based on the decrease of the isotropically averaged intensities with the resolution. In the future, a more detailed analysis may be developed defining the typical B values, especially for low-resolution data sets, as indicated by Vaguine *et al.* (1999). An accurate estimate of anisotropic typical B values directly from the set of intensities and prior to model building may be another important future development.

We reiterate that this approach works well with experimental data, as §5 illustrates, but it is not aimed at checking their correctness or that of corresponding syntheses or at selecting the data. Nevertheless, the method may suggest removing some higher resolution reflections in order to improve the expected details in the Fourier maps, as some examples in the text show.

In the future, the proposed measures may find other applications apart from use in describing how detailed the diffraction data sets are. For example, for a given pair of atoms a minimum distance between them may be estimated using the same approach. If this minimum distance is shorter than the actual distance between the atoms, this would mean that the diffraction data contain sufficient information so that the stereochemical restraints for these atoms may be omitted during atomic model refinement. This and other suggestions may be the subject of a separate project.

We thank the European Research Council (ERC Starting Grant), the French Infrastructure for Integrated Structural Biology (FRISBI) and Instruct as part of the European Strategy Forum on Research Infrastructures (ESFRI) for their support, A. Ben-Shem for discussing the problem and for critical reading of the manuscript, P. V. Afonine for useful discussions, and the referees and the Co-editor for their very constructive comments which helped to improve the presentation of the results.

References

Adams, P. D. *et al.* (2010). *Acta Cryst.* D**66**, 213–221.
Afonine, P. V., Grosse-Kunstleve, R. W., Adams, P. D., Lunin, V. Y. & Urzhumtsev, A. (2007). *Acta Cryst.* D**63**, 1194–1197.
Afonine, P. V., Grosse-Kunstleve, R. W., Adams, P. D. & Urzhumtsev, A. (2013). *Acta Cryst.* D**69**, 625–634.
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.
Agarwal, R. C. (1978). *Acta Cryst.* A**34**, 791–809.
Altomare, A., Cuocci, C., Giacovazzo, C., Kamel, G. S., Moliterni, A. & Rizzi, R. (2008). *Acta Cryst.* A**64**, 326–336.
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.
Blundell, T. L. & Johnson, L. N. (1976). *Protein Crystallography*. London: Academic Press.
Brown, P. J., Fox, A. G., Maslen, E. N., O'Keefe, M. A. & Willis, B. T. M. (2006). *International Tables for Crystallography*, Vol. *C*, edited by E. Prince, pp. 555–595. Dordrecht: Kluwer Academic Publishers.
Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C. & Siliqi, D. (2005*a*). *Acta Cryst.* D**61**, 556–565.
Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C. & Siliqi, D. (2005*b*). *Acta Cryst.* D**61**, 1080–1087.
DeLano, W. L. (2002). *PyMOL*. http://www.pymol.org.
James, R. W. (1948). *Acta Cryst.* **1**, 132–134.
Jiang, J.-S. & Brünger, A. T. (1994). *J. Mol. Biol.* **243**, 100–115.
Karplus, P. A. & Diederichs, K. (2012). *Science*, **336**, 1030–1033.
Phillips, S. E. (1980). *J. Mol. Biol.* **142**, 531–554.
Rossmann, M. G. & Blow, D. M. (1962). *Acta Cryst.* **15**, 24–31.
Sheldrick, G. M. (1990). *Acta Cryst.* A**46**, 467–473.
Simonetti, A., Marzi, S., Fabbretti, A., Hazemann, I., Jenner, L., Urzhumtsev, A., Gualerzi, C. O. & Klaholz, B. P. (2013). *Acta Cryst.* D**69**, 925–933.
Stenkamp, R. E. & Jensen, L. H. (1984). *Acta Cryst.* A**40**, 251–254.
Urzhumtsev, A. G. (1991). *Acta Cryst.* A**47**, 794–801.
Urzhumtsev, A. (2000). *CCP4 Newsl. Protein Crystallogr.* **38**, 38–49.
Urzhumtsev, A., Afonine, P. V. & Adams, P. D. (2009). *Acta Cryst.* D**65**, 1283–1291.
Urzhumtseva, L., Afonine, P. V., Adams, P. D. & Urzhumtsev, A. (2009). *Acta Cryst.* D**65**, 297–300.
Urzhumtseva, L. & Urzhumtsev, A. (2011). *J. Appl. Cryst.* **44**, 865–872.
Vaguine, A. A., Richelle, J. & Wodak, S. J. (1999). *Acta Cryst.* D**55**, 191–205.
Weiss, M. S. (2001). *J. Appl. Cryst.* **34**, 130–135.
Wilson, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.