## 18. REFINEMENT

### 18.5.6. The diffraction-component precision index

### 18.5.6.1. Statistical expectation of error dependence

From general statistical theory, one would expect the s.u. of an atomic coordinate determined from the diffraction data alone to show dependence on four factors:

$$
\begin{equation*}
\sigma(x) \propto(\mathcal{R})\left[\left(n_{\text {atoms }}\right) /\left(n_{\text {obs }}-n_{\text {params }}\right)\right]^{1 / 2}\left(1 / s_{\mathrm{rms}}\right) \tag{18.5.6.1}
\end{equation*}
$$

Here, $\mathcal{R}$ is some measure of the precision of the data; $n_{\text {atoms }}$ is the recognition that the information content of the data has to be shared out; $n_{\text {obs }}$ is the number of independent data, but to achieve the correct number of degrees of freedom this must be reduced by $n_{\text {params }}$, the number of parameters determined; and $1 / s_{\text {rms }}$ is a more specialized factor arising from the sensitivity $\partial|F| / \partial x$ of the data to the parameter $x$. Here $s_{\mathrm{rms}}$ is the r.m.s. reciprocal radius of the data. Any statistical error estimate must show some correspondence to these four factors.

### 18.5.6.2. A simple error formula

Cruickshank (1960) offered a simple order-of-magnitude formula for $\sigma(x)$ in small molecules. It was intended for use in experimental design: how many data of what precision are needed to achieve a given precision in the results? The formula, derived from a very rough estimate of a least-squares diagonal element in non-centrosymmetric space groups, was

$$
\begin{equation*}
\sigma\left(x_{i}\right)=(1 / 2)\left(N_{i} / p\right)^{1 / 2}\left[R / s_{\mathrm{rms}}\right] \tag{18.5.6.2}
\end{equation*}
$$

Here $p=n_{\text {obs }}-n_{\text {params }}, R$ is the usual residual $\sum|\Delta F| / \sum|F|$ and $N_{i}$ is the number of atoms of type $i$ needed to give scattering power at $s_{\mathrm{rms}}$ equal to that of the asymmetric unit of the structure, i.e., $\sum_{j} f_{j}^{2} \equiv N_{i} f_{i}^{2}$. [The formula has also proved very useful in a systematic study of coordinate precision in the many thousands of small-molecule structure analyses recorded in the Cambridge Structural Database (Allen et al., 1995a,b).]

For small molecules, the above definition of $N_{i}$ allowed the treatment of different types of atom with not-too-different $B$ 's. However, it is not suitable for individual atoms in proteins where there is a very large range of $B$ values and some atoms have $B$ 's so large as to possess negligible scattering power at $s_{\text {rms }}$.

Often, as in isotropic refinement, $n_{\text {params }} \simeq 4 n_{\text {atoms }}$, where $n_{\text {atoms }}$ is the total number of atoms in the asymmetric unit. For fully anisotropic refinement, $n_{\text {params }} \simeq 9 n_{\text {atoms }}$.

A first very rough extension of (18.5.6.2) for application in proteins to an atom with $B=B_{i}$ is

$$
\begin{equation*}
\sigma\left(x_{i}\right)=k\left(N_{i} / p\right)^{1 / 2}\left[g\left(B_{i}\right) / g\left(B_{\mathrm{avg}}\right)\right] C^{-1 / 3} R d_{\min } \tag{18.5.6.3}
\end{equation*}
$$

where $k$ is about $1.0, N_{i}=\sum Z_{j}^{2} / Z_{i}^{2}, B_{\text {avg }}$ is the average $B$ for fully occupied sites and $C$ is the fractional completeness of the data to $d_{\text {min }}$. In deriving (18.5.6.3) from (18.5.6.2), $1 / s_{\text {rms }}$ has been replaced by $1.3 d_{\min }$, and the factor $(1 / 2)(1.3)=0.65$ has been increased to 1.0 as a measure of caution in the replacement of a full matrix by a diagonal approximation. $g(B)=1+a_{1} B+a_{2} B^{2}$ is an empirical function to allow for the dependence of $\sigma(x)$ on $B$. However, the results in Section 18.5.4.2 showed that the parameters $a_{1}$ and $a_{2}$ depend on the structure.

As also mentioned in Section 18.5.4.2, Sheldrick has found that the $Z_{i}$ in $N_{i}$ is better replaced by $Z_{i}^{\#}$, the scattering factor at $\sin \theta / \lambda=0.3 \AA^{-1}$. Hence, $N_{i}$ may be taken as

$$
\begin{equation*}
N_{i}=\left(\sum Z_{j}^{\# 2} / Z_{i}^{\# 2}\right) \tag{18.5.6.4}
\end{equation*}
$$

A useful comparison of the relative precision of different structures may be obtained by comparing atoms with the respective $B=B_{\text {avg }}$ in the different structures. (18.5.6.3) then reduces to

$$
\begin{equation*}
\sigma\left(x, B_{\text {avg }}\right)=1.0\left(N_{i} / p\right)^{1 / 2} C^{-1 / 3} R d_{\min } \tag{18.5.6.5}
\end{equation*}
$$

The smaller the $d_{\min }$ and the $R$, the better the precision of the structure. If the difference between oxygen, nitrogen and carbon atoms is ignored, $N_{i}$ may be taken simply as the number of fully occupied sites. For heavy atoms, (18.5.6.4) must be used for $N_{i}$.

Equation (18.5.6.5) is not to be regarded as having absolute validity. It is a quick and rough guide for the diffraction-data-only error component for an atom with Debye $B$ equal to the $B_{\text {avg }}$ for the structure. It is named the diffraction-component precision index, or DPI. It contains none of the restraint data.

### 18.5.6.3. Extension for low-resolution structures and use of $R_{\text {free }}$

For low-resolution structures, the number of parameters may exceed the number of diffraction data. In (18.5.6.3) and (18.5.6.5), $p=n_{\mathrm{obs}}-n_{\text {params }}$ is then negative, so that $\sigma(x)$ is imaginary. This difficulty can be circumvented empirically by replacing $p$ with $n_{\mathrm{obs}}$ and $R$ with $R_{\text {free }}$ (Brünger, 1992). The counterpart of the DPI (18.5.6.5) is then

$$
\begin{equation*}
\sigma\left(x, B_{\text {avg }}\right)=1.0\left(N_{i} / n_{\mathrm{obs}}\right)^{1 / 2} C^{-1 / 3} R_{\text {free }} d_{\text {min }} \tag{18.5.6.6}
\end{equation*}
$$

Here $n_{\text {obs }}$ is the number of reflections included in the refinement, not the number in the $R_{\text {free }}$ set.

It may be asked: how can there be any estimate for the precision of a coordinate from the diffraction data only when there are insufficient diffraction data to determine the structure? By following the line of argument of Cruickshank's (1960) analysis, (18.5.6.6) is a rough estimate of the square root of the reciprocal of one diagonal element of the diffraction-only least-squares matrix. All the other parameters can be regarded as having been determined from a diffraction-plus-restraints matrix.

Clearly, (18.5.6.6) can also be used as a general alternative to (18.5.6.5) as a DPI, irrespective of whether the number of degrees of freedom $p=n_{\text {obs }}-n_{\text {params }}$ is positive or negative.

Comment. When $p$ is positive, (18.5.6.6) would be exactly equivalent to (18.5.6.5) only if $R_{\text {free }}=R\left[n_{\text {obs }} /\left(n_{\text {obs }}-n_{\text {params }}\right)\right]^{1 / 2}$. Tickle et al. (1998b) have shown that the expected relationship in a restrained refinement is actually

$$
\begin{equation*}
R_{\text {free }}=R\left\{\left[n_{\mathrm{obs}}+\left(n_{\text {params }}-h\right)\right] /\left[n_{\mathrm{obs}}-\left(n_{\text {params }}-h\right)\right]\right\}^{1 / 2}, \tag{18.5.6.7}
\end{equation*}
$$

where $h=n_{\text {restraints }}-\sum w_{\text {geom }}(\Delta Q)^{2}$, the latter term, as in (18.5.3.1), being the weighted sum of the squares of the restraint residuals.

### 18.5.6.4. Position error

Often an estimate of a position error $|\Delta \mathbf{r}|$, rather than a coordinate error $|\Delta x|$, is required. In the isotropic approximation,

$$
\begin{equation*}
\sigma\left(r, B_{\text {avg }}\right)=3^{1 / 2} \sigma\left(x, B_{\text {avg }}\right) \tag{18.5.6.8}
\end{equation*}
$$

Consequently, the DPI formulae for the position errors are

$$
\begin{equation*}
\sigma\left(r, B_{\text {avg }}\right)=3^{1 / 2}\left(N_{i} / p\right)^{1 / 2} C^{-1 / 3} R d_{\min } \tag{18.5.6.9}
\end{equation*}
$$

with $R$ and

$$
\begin{equation*}
\sigma\left(r, B_{\text {avg }}\right)=3^{1 / 2}\left(N_{i} / n_{\text {obs }}\right)^{1 / 2} C^{-1 / 3} R_{\text {free }} d_{\text {min }} \tag{18.5.6.10}
\end{equation*}
$$

with $R_{\text {free }}$.

