

### 3.5. Data reduction to $|F_{hkl}|$ values

A. LE BAIL

#### 3.5.1. Introduction

Collecting the structure-factor amplitudes  $|F_{hkl}|$  is the key step leading to structure solution from powder-diffraction data, just as from single-crystal data. However, there are specific difficulties and pitfalls associated with powder data, mainly because of diffraction-peak overlap. Once indexing is realized, data reduction to  $|F_{hkl}|$  is a fast process, using whole-powder-pattern decomposition (WPPD) methods. This comfortable situation was not attained without past efforts, which are reviewed in this chapter. The introduction of modern WPPD methods occurred slowly and progressively over the past 30 years, thanks to increases in computer power, improvements in graphical user interfaces, diffractometer data digitalization, the availability of synchrotron and neutron radiation, and last but not least, the proposition of new algorithms. Innovations were not instantly accepted, this also being true for the Rietveld (1969) method, or could not be applied immediately to every type of powder data. Predecessors of the current WPPD methods extracted peak intensities without restraining the cell, so that each peak position was a parameter to be refined (as well as the peak intensity, and the peak shape and width). This is still useful if the aim is to obtain peak positions for indexing, although simple derivative methods can make searching for peak positions faster. Taking advantage of the indexing (Bergmann *et al.*, 2004), new WPPD methods that applied cell restraints to the peak positions opened the door to a long list of new possibilities and applications (including first indexing confirmation and manual or automatic space-group estimation) which are detailed in this chapter. A partial review of the applications realized in thousands of published papers is given, and the evolution of the methods will be discussed. Additional information on the topic of reduction to  $|F_{hkl}|$  values can be found in the books by Young (1993), Giacovazzo (1998), David *et al.* (2002), Pecharsky & Zavalij (2003), Clearfield *et al.* (2008) and Dinnebier & Billinge (2008) or in selected reviews (Toraya, 1994; Langford & Louër, 1996; Le Bail, 2005).

#### 3.5.2. Algorithms

Whole-powder-pattern fitting (WPPF) is a general definition including WPPD as well as the Rietveld method (Rietveld, 1969). In the latter method, the atomic coordinates are required for the intensity calculations, and the sum of all the peak contributions produces a calculated powder pattern which is compared to the observed one, allowing the least-squares refinement of profile and structural parameters. The Rietveld method historically preceded modern WPPD methods, though the latter are applicable without atomic coordinates. Of course, one may use WPPD methods if the structure is known, but in some cases one does not want to use that knowledge (not wanting to restrain the peak intensities by the structural model, for instance, nevertheless believing in the indexing or wanting to confirm it, using the restraint of the cell parameters). Any WPPF approach should be able to model the peak shape and width variations with diffraction angle (complications not considered here may occur in the

case of anisotropic broadening intrinsic to the sample). This can be done by fitting some analytical profile shape and width parameters in a semi-empirical approach. The angular variation of these parameters is generally controlled by refining the  $U$ ,  $V$  and  $W$  terms in the Caglioti *et al.* (1958) expression  $[(\text{FWHM})^2 = U \tan^2 \theta + V \tan \theta + W]$  or a variation (where FWHM = full width at half maximum). The alternative is to use the fundamental parameter approach (FPA) (Cheary & Coelho, 1992). However, some of the original computer programs did not apply any cell restraint or even any restraint at all.

##### 3.5.2.1. Unrestrained cell

Without a cell hypothesis, no  $|F_{hkl}|$  values can be extracted; the intensity values collected will be noted by  $I(i)$  until Miller indices are attributed, enabling the multiplicity correction. Obtaining all the peak positions, areas, breadths and shape parameters as independent values for a whole powder pattern is limited to simple cases where there is not too much peak overlap. With such an approach (both cell and space group unknown or unused) one has to estimate the number of peaks to be fitted, so that the fit of a complex group of peaks will lead to large uncertainties. However, knowing the cell and space group provides at least the correct number of peaks and an estimate of their starting positions. Such calculations were made as an alternative to the Rietveld method, during the first stage of the so-called two-stage method for refinement of crystal structures (Cooper *et al.*, 1981). In the case of X-ray data, the profile shapes applied in the Rietveld method (Gaussian at the beginning for neutron data) evolved a great deal (Wiles & Young, 1981), and on the WPPD side happened to be described in these two-stage approaches by a sum of Lorentzian curves, or double Gaussians (Will *et al.*, 1983, 1987). The computer program *PROFIT* (Scott, 1987), derived from software for individual profile fitting (Sonneveld & Visser, 1975) and extended to the whole pattern, was applied to the study of crystallite size and strain in zinc oxide (Langford *et al.*, 1986) and for the characterization of line broadening in copper oxide (Langford & Louër, 1991). Studying a whole pattern can also be done in simple cases by using software designed for the characterization of single or small groups of peaks; an example is a ZnO study (Langford *et al.*, 1993) using the computer program *FIT* (Socobim/Bruker). WPPD on complex cases is mostly realized today by using peak positions controlled by the cell parameters, with the benefit of stronger accuracy of the  $|F_{hkl}|$  values, even if the lost degrees of freedom may lead to slightly worse fits, increasing the profile  $R$  factors. Before 1987, close to thirty structure determinations by powder diffractometry (SDPDs) were achieved using intensities extracted by using these old WPPD methods without cell constraints (see the SDPD database; Le Bail, 2007). It can be argued that freeing the peak positions allows one to take into account subtle effects in position displacement (in stressed samples, for example). But systematic discrepancy of observed peak positions with regard to the theoretical position, as expected from the cell parameters, can be modelled as well in modern WPPD methods or in the Rietveld method.

### 3.5. DATA REDUCTION TO $|F_{hkl}|$ VALUES

#### 3.5.2.2. *Restrained cell*

WPPD methods that strictly applied the peak positions calculated from a cell (hypothesized from indexing results) marked a great step forward in the quest for improving the SDPD success rate. This is essentially because the quality of the estimated intensities increased, and even if the main handicap of powder diffraction (peak overlap) could not be completely circumvented, it was at least more clearly delimited. Today two generic names are retained for such cell-restrained WPPD methods that can produce a set of extracted  $|F_{hkl}|$  values suitable for attempting a structure solution: the Pawley and Le Bail methods. Both were derived from the Rietveld (1969) method, so they share with it many equations which will not be restated here (see Chapter 4.7).

##### 3.5.2.2.1. *Pawley method*

The idea of removing the crystal structure refinement part in a Rietveld program and adding the potential to refine an individual intensity for every expected Bragg peak produced a new software package (named *ALLHKL*) allowing refinement of the cell parameters very precisely and extraction of a set of structure-factor amplitudes (Pawley, 1981). The process was much later called the ‘Pawley method’. Overcoming the least-squares ill conditioning due to peak overlap was achieved by using slack constraints (Waser, 1963). Pawley clearly insisted on the usefulness of the procedure for the confirmation of the indexing of a powder pattern of an unknown. Nevertheless, the structure of the  $C_6F_{10}$  (at 4.2 K) test case selected for demonstration purposes remained unsolved (but see Section 3.5.4.2 below). No SDPD of an unknown was realized using the Pawley method for several years (although successful tests were published corresponding to redeterminations of previously known structures). The first real SDPD of an unknown using the Pawley method seems to be that of  $I_2O_4$  (Lehmann *et al.*, 1987); its powder pattern had been previously indexed, but the structure not determined because of the lack of a suitable single crystal. During these pioneering years, *ALLHKL* could not extract the intensities for more than 300 peaks, so that, in more complex cases, it was necessary to subdivide the pattern into several parts. Moreover, it was rather difficult to avoid completely the ill conditioning due to overlapping peaks. Successful fits yielded equipartitioned intensities (*i.e.*, equal structure factors for those Bragg peaks with exact overlap). Unsuccessful fits could easily produce negative intensities which, combined with positive ones for other peak(s) at the same angle, reproduced the global positive value. Moreover, the first version to apply Gaussian peak shapes could not easily produce any SDPD because of the relatively poor resolution of constant-wavelength neutron data, so that it needed to be adapted to X-ray data, with the implementation of more complex peak shapes. Several programs were subsequently developed, based on the same principles as the original Pawley method. The first of them, by Toraya (1986), extended the use to X-ray data with non-Gaussian profile shapes, and introduced two narrow band matrices instead of a large triangular matrix, saving both computation time and memory space in a program named *WPPF*. Some programs were used to produce intensities in order to apply the so-called two-stage method (Cooper *et al.*, 1981) for structure refinement, such as *PROFIT* (Scott, 1987) and *PROFIN* (Will, 1988) (no slack constraints, but equal division of the intensity between expected peaks when the overlap was severe). There was intense continuing activity on Pawley-like software with other programs such as *FULFIT* (Jansen *et al.*, 1988), *LSQPROF* (Jansen *et al.*, 1992) and *POLISH* (Byrom & Lucas, 1993).

Estimation of intensities of overlapping reflections was improved in *LSQPROF* by applying relations between structure-factor amplitudes derived from direct methods, and the Patterson function was considered in the satellite program *DOREES* (Jansen *et al.*, 1992). The question of how to determine the intensities of completely (or largely) overlapping reflections (either systematic overlap due to symmetry or fortuitous overlap) from a single powder pattern cannot have a definite simple answer, but continues to be discussed, since it is essential for improving our ability to solve structures. An early view with a probabilistic approach was given by David (1987), later introducing Bayesian statistics (Sivia & David, 1994) into the Pawley method. Early detection of preferred orientation on the basis of analysis of the *E*-value distribution was another way (Peschar *et al.*, 1995) to improve the structure-factor-amplitude estimate. New computer programs based on the Pawley method continue to be written even today.

##### 3.5.2.2.2. *Le Bail method*

In order to be able to estimate *R* factors related to integrated intensities, Rietveld (1969) stated [see also the book *The Rietveld Method* edited by Young (1993)]: ‘a fair approximation to the observed integrated intensities can be made by separating the peaks according to the calculated values of the integrated intensities,’ *i.e.*

$$I_{hkl}(\text{obs}) = \sum_j [w_{j,hkl} S_{hkl}^2(\text{calc}) y_j(\text{obs}) / y_j(\text{calc})], \quad (3.5.1)$$

where  $w_{j,hkl}$  is a measure of the contribution of the Bragg peak at position  $2\theta_{hkl}$  to the diffraction profile  $y_j$  at position  $2\theta_j$  [corresponding to equation 7 in Rietveld (1969)].  $S_{hkl}^2$  is the sum of the nuclear and magnetic contributions for neutron diffraction, or is more simply  $F_{hkl}^2$  for X-rays. The sum is over all  $y_j(\text{obs})$  that can theoretically contribute to the integrated intensity  $I_{hkl}(\text{obs})$ . Bias is introduced here by apportioning the intensities according to the calculated intensities; this is why the observed intensities are said to be ‘observed’, in quotation marks, in the Rietveld method. These ‘observed’ intensities are used in the  $R_B$  and  $R_F$  calculations (residuals on intensities and structure-factor amplitudes, respectively). They are also required for Fourier-map estimations, which, as a consequence, are less reliable than those from single-crystal data.

A process using the Rietveld decomposition formula iteratively for WPPD purposes was first applied in 1988 (Le Bail *et al.*, 1988) and much later was called the ‘Le Bail method’ or ‘Le Bail fit’, or ‘pattern matching’ as well as ‘profile matching’ in the *FULLPROF* Rietveld program (Rodriguez-Carvajal, 1990). In the original computer program (named *ARITB*), arbitrarily all equal  $S_{hkl}^2(\text{calc})$  values are first entered in the above equation, instead of using structure factors calculated from the atomic coordinates, resulting in ‘ $I_{hkl}(\text{obs})$ ’ which are then re-entered as new  $S_{hkl}^2(\text{calc})$  values at the next iteration, while the usual profile and cell parameters (but not the scale) are refined by least squares (*ARITB* used profile shapes represented by Fourier series, either analytical or learned from experimental data, providing an easy way to realize convolution by broadening functions modelling size–strain sample effects, possibly anisotropic). Equipartition of exactly overlapping reflections comes from the strictly equal result from equation (3.5.1) for Bragg peaks at the same angles which would have equal starting calculated intensities. Not starting from a set of all equal  $S_{hkl}^2(\text{calc})$  values avoids equipartition for the exactly overlapping reflections but produces  $I_{hkl}(\text{obs})$  keeping the same original ratio

### 3. METHODOLOGY

as the  $S_{hkl}^2(\text{calc})$  ones. It is understandable that such an iterative process requires starting cell and profile parameters as good as the Rietveld method itself. The process is easier to incorporate within an existing Rietveld code than the Pawley method, so that most Rietveld codes now include structure-factor amplitudes extraction as an option (generally multiphase), with the possibility of combining Rietveld refinement(s) together with Le Bail fit(s).

A non-exhaustive list of programs applying this method (either exclusively or added within a Rietveld code) includes *MPROF* (Jouanneaux *et al.*, 1990), later renamed *WinMPROF*; *FULLPROF* (Rodriguez-Carvajal, 1990); *EXTRACT* (Baerlocher, 1990); *EXTRA* (Altomare *et al.*, 1995); *EXPO* (Altomare *et al.*, 1999), which is the integration of *EXTRA* and *SIRPOW.92* for solution and refinement of crystal structures; and *RIETAN* (Izumi & Ikeda, 2000). Then followed most well known Rietveld codes (*BGMN*, *GSAS*, *MAUD*, *TOPAS* *etc.*) or standalone programs (*AJUST* by Rius *et al.*, 1996). In the work of the Giacobozzo group, many modifications of the  $|F_{hkl}|$  values for SDPD purposes were applied before or after the extraction and were integrated in *EXPO2011* (Altomare *et al.*, 2011): obtaining information about the possible presence of preferred orientation by statistical analysis of the normalized structure-factor moduli; using the positivity of the Patterson function in the decomposition process, this having been considered previously (David, 1987; Estermann & Gramlich, 1993); characterization of pseudotranslational symmetry used as prior information in the pattern-decomposition process; multiple Le Bail fits with random attribution of intensity to the overlapping reflections, instead of equipartition, followed by application of direct methods to large numbers of such data sets; use of a located structure fragment for improving the pattern-decomposition process; and use of probability (triplet-invariant distribution functions) integrated with the Le Bail algorithm. Another approach for solving the overlapping problem was proposed by using maximum-entropy coupled with likelihood evaluation (Dong & Gilmore, 1998). The list of structure solutions made from intensities extracted by using the Pawley and Le Bail methods is too long to be given here; a partial list (>1000 first cases, including those using  $|F_{hkl}|$  values extracted by other methods) can be found on the web (Le Bail, 2007). The first application of the Le Bail method was to the structure solution of  $\text{LiSbWO}_6$  (Le Bail *et al.*, 1988) using the *ARITB* software.

#### 3.5.3. Pitfalls in the extraction of accurate $|F_{hkl}|$ values using the Pawley and Le Bail methods

In the Rietveld refinement guidelines published by the IUCr Commission on Powder Diffraction (McCusker *et al.*, 1999), it is said that the Rietveld  $R_{\text{wp}}$  value should approach the value obtained in a structure-free refinement (*i.e.* using WPPD methods). Such a refinement is recommended for the estimation of initial values of the Rietveld profile parameters. Consequently,  $|F_{hkl}|$  values extracted by WPPD can be used to make a good reproduction of the experimental powder pattern if the cell is correct (which is ultimately only proven if the structure is solved and refined). Pitfalls can occur during post-treatment and application of the  $|F_{hkl}|$  data if one neglects the possible errors that are inherently present due to exact or accidental overlap, preferred orientation effects or wrong background estimations, citing only three of the main possible causes of errors in these  $|F_{hkl}|$  values.

#### 3.5.3.1. Consequences of (exact or accidental) overlap

The uncertainties of the  $|F_{hkl}|$  values of overlapped reflections cannot be overcome in a single powder-diffraction experiment. This problem has led to various approaches, all being more or less inefficient: equipartition, non-equipartition by random distribution *etc.* If direct methods are applied, the trend is to multiply the number of solution attempts, trying to identify the most convincing one by using structural arguments (such as atoms in chemically reasonable positions). When applying real-space methods (which require chemical knowledge, such as the three-dimensional molecular structure or the presence of definite polyhedra) one generally chooses to work either directly on the raw powder pattern or on a pseudo pattern built from the extracted  $|F_{hkl}|$  values, so that wrong individual values are less of a problem, since only the sums of the contributions in overlapping regions are checked during the search for the molecule, polyhedra or atom positions. Indeed, working on the raw powder pattern does not need reduction to  $|F_{hkl}|$  values in theory, but in practice either the Pawley or Le Bail methods are applied first in order to fix the zero point, background, cell and profile parameters which will then be applied during the structure model checking, and to speed the calculations. The extracted  $|F_{hkl}|$  values can be used in mathematical expressions defining correlations induced by the overlap. These equations were developed by David *et al.* (1998) for the Pawley method in the real-space structure solution program *DASH* and by Pagola *et al.* (2000) for the Le Bail method in *PSSP*. Regenerating a powder pattern from the extracted  $|F_{hkl}|$  values was carried out in the *ESPOIR* real-space computer program (Le Bail, 2001) using a simple Gaussian peak shape whose width follows the Caglioti relation established from the raw pattern. With such a pseudo powder pattern (without profile asymmetry, background *etc.*), the calculations are much faster than if the raw pattern is used. When using direct methods instead of real-space methods, the approaches are different, because direct methods require a more complete data set (up to  $d = 1 \text{ \AA}$ ) of accurate  $|F_{hkl}|$  values. However, removing up to half of them (those with too much overlap, *i.e.* where the overlap is greater than half the FWHM, for instance) can lead to success with direct methods. One can even remove up to 70–80% of the data if the Patterson method is applied and if only a small number of heavy atoms are to be located.

#### 3.5.3.2. Preferred-orientation effects

One has to ensure that preferred orientation is minimized during the recording of the powder pattern if the extracted  $|F_{hkl}|$  values are to be used for structure solution. In transmission geometry with a capillary specimen (provided that it is not composed of long needle-shaped particles that are all aligned), there is generally no problem. But in reflection geometry with samples pressed on the holder, preferred orientation is not rare, even if it is not obvious in the data. Collecting a second pattern from a sample dusted onto the holder through a fine sieve can be informative. However, some WPPD applications may not be sensitive to such a problem. If only the cell parameters have to be accurately estimated for thermal-expansion studies (Evans *et al.*, 1996), it can be much faster to use WPPD rather than the Rietveld method. However, it is not recommended to do this systematically, especially if the structure is complex and the resolution is low [see the warnings in Peterson (2005)].