International Tables for Crystallography (2018). Vol. H, Section 3.7.1.2, pp. 304–305.

3.7. Crystallographic databases and powder diffraction

J. A. Kaduk

3.7.1. Introduction

Identifying compounds using powder-diffraction data requires a comparison of the current experimental pattern with essentially all previous crystallographic information. This information is incorporated into the Powder Diffraction File (Fawcett *et al.*, 2017), which is traditionally the primary tool for phase identification, but other databases are important to the process, as well as being the repositories of the atom coordinates necessary for Rietveld refinements. This chapter summarizes the characteristics of the various databases that the author has found useful in the practice of powder diffraction. It also provides several examples of the thought processes and capabilities which can be used to identify phases.

3.7.1.1. *History of the PDF/ICDD*

Although powder-diffraction experiments date from the beginning of the 20th century (Debye & Scherrer, 1916, 1917; Hull, 1919), what we now know as the Powder Diffraction File and the International Centre for Diffraction Data date from two papers from the Dow Chemical Company (Hanawalt & Rinn, 1936; Hanawalt *et al.*, 1938). The importance of these papers lies not only in the compilation of a database but also in a method for the identification of materials, and how the database was organized to work with the method. Discussion among industrial and academic scientists made the need for a central collection of powder-diffraction patterns apparent. The Joint Committee for Chemical Analysis by Powder Diffraction Methods was founded in 1941. It produced a primary reference of X-ray powder diffraction data, which became known as the Powder Diffraction File (PDF). This effort was supported initially by Committee E-4 of the American Society for Testing and Materials (ASTM). Over the next two decades, other professional bodies added their support, culminating in 1969 with the establishment of the Joint Committee on Powder Diffraction Standards (JCPDS). The JCPDS was incorporated as a separate nonprofit corporation to continue the mission of maintaining the PDF. In 1978 the name was changed to the International Centre for Diffraction Data to highlight the global nature of this scientific endeavour. Additional information on the history of the powder method is given in Parrish (1983) and on the early history of the Powder Diffraction File in Hanawalt (1983).

3.7.1.2. *Search/match*

What is now known as the Hanawalt search method (Hanawalt & Rinn, 1936; Hanawalt *et al.*, 1938) is an empirical scheme which was based on earlier ideas (Hull, 1919; Davey, 1922, 1934; Winchell, 1927; Waldo, 1935; Boldyrev *et al.*, 1938). The basic ideas behind the scheme have been summarized in more modern language by Hanawalt (1986). Other discussions of the method can be found in Jenkins & Rose (1990) and Hull (1983).

The patterns in the PDF are divided into 40 groups according to the *d*-spacing of the strongest peak and including error limits on the *d*-spacing. The entries within each group are sorted by the position of the second strongest peak. Because the peak intensities can be more difficult to measure than the positions and may vary from sample to sample, PDF entries appear in the index multiple times ('rotations' in the current nomenclature). All patterns appear at least once. Patterns appear twice when $I_2/I_1 > 0.75$ and $I_3/I_1 \leq 0.75$, three times when $I_3/I_1 > 0.75$ and $I_4/I_1 \leq 0.75$, and four times when $I_4/I_1 > 0.75$ (where I_1 is the strongest peak, I_2 is the second strongest and so on). There are four more rules, dealing with things such as low-angle peaks, rounding of *d*-spacings and closely spaced peaks.

The phase-identification process actually involves several steps. This was realized by Hanawalt, Rinn and Frevel even in 1938. The first step is to *search* the experimental data against an index (a structured subset of a database) to identify potential compounds. The printed *Hanawalt Search Manual* was such an index, and contemporary search/match programs all generate indices to enhance the speed of the phase-identification process. The second step is the *match* of the full PDF entry against the full experimental pattern to use all peaks in the identification process. Typically, the quality of the match is evaluated at this point to rank the potential candidate match among the others in the list; the hit list is sorted on goodness of match, similarity index, figure of merit or some similar quantity generated by the program. The third step is to *identify* the phase (generally by computer, but best with some human judgement). The pattern of the identified phase is then subtracted from the experimental pattern and the process is repeated to identify additional phases. The final step in the process is often quantification of the concentrations. Ultimately, the errors introduced during the subtractions limit the number of phases which can be identified in a mixture, and additional techniques are required to identify minor or trace phases. As specimen-preparation methods and equipment and standardized reference data have improved with time (over decades), the residual errors in the subtraction process have diminished, generally increasing the number of phases that can be identified when appropriate techniques are applied.

An early computer version of the Hanawalt search algorithm was implemented by Frevel (Frevel, 1965; Frevel *et al.*, 1976). This program used a 300-phase microfile database of common phases resulting from empirical work performed over decades at the Dow Chemical Company. Another early computer implementation of the Hanawalt search algorithm was developed by Snyder (1981). The index file stored the *d*/*I* pairs in the PDF as 16-bit integers: 11 bits for 1/*d* and five bits for *I*. The index file was an indexed sequential file with the PDF entries sorted on d_1 (the *d*-spacing of the strongest peak). Each PDF set was indexed separately, and smaller MICRO (300 phases) and MINI (2500 common phases) index files were also generated to permit faster searches on the slow computers of the day. After input of the *d*/*I* list for the experimental pattern, the program located PDF entries whose d_1 values lay within $\pm 0.1^\circ$ 2θ (copper radiation) of the observed d_1. If the reference pattern passed three tests – it was a member of a user-specified subfile, all PDF entry peaks with $I \geq 50$ were present in the unknown pattern and user-specified chemical constraints were satisfied – a figure of merit (FOM) was calculated. The pattern with the highest FOM was saved for the match step, and the process was repeated for d_2 and d_3. If no hits were obtained, larger error windows and then weaker peaks were used.

The FOM was calculated as

$$\text{FOM} = d_R I_R^2 d_U, \tag{3.7.1}$$

where d_R is the percentage of the reference peaks which match the unknown (within the error window) and have I greater than that of the lowest-intensity matched peak, I_R is the percentage of the reference intensity ($I_{\rm ref}$) matched and d_U is the percentage of the unknown peaks (with intensities $I_{\rm unk}$) matched.

PDF hits were considered for the match step if the FOM was >10. For the hit with the highest FOM, an I-weighted linear regression between $I_{\rm ref}$ and $I_{\rm unk}$ was carried out. Peaks with $I_{\rm calc} < I_{\rm obs} - 5$ were assigned as overlapped, and the least-squares scale factor was recalculated using only the non-overlapped peaks. The scaled PDF entry was subtracted from the unknown pattern and the residual was sent back to the search step.

Several commercial search/match programs have been developed, not from Snyder's implementation of the Hanawalt algorithm, but from the Johnson–Vand algorithm (Johnson & Vand, 1967, 1968; Cherukuri *et al.*, 1983). This algorithm used constant error windows in $1/d$ and $\log(I)$ and used integer arithmetic. The d/I pairs were packed into characteristic integers PSI = $(1000/d) \times 10 + 5 \log_{10} I_3$. An inverted PDF was created, an index which contained the characteristic integers of the strongest lines of the reference patterns (PSI, PDF No. pairs) sorted by decreasing PSI. The input d/I list was compared with the index. All patterns that contained the characteristic integers within the bandpass were considered as potential hits. The full PDF was used to compare observed and reference patterns. A Davey minimum concentration (DMC) was calculated; this was the largest value in the range $0 \leq \text{DMC} \leq 1$ for which $I_{\rm PDF}\text{DMC} \leq I_{\rm unk}$ for all peaks. The PDF entry was then subtracted from the unknown pattern and the process was repeated. Initially, there were no chemistry or user filters; these appeared in later versions.

The Johnson–Vand figure of merit,

$$\text{FOM} = A\left[1 - \frac{\sum_N |\Delta D|}{(IW)N}\right]\left[1 - \frac{\sum_N |\Delta I| - K}{\sum_N I}\right], \tag{3.7.2}$$

was calculated, in which A is the percentage of peak match in the d-space range considered (above the background), $\Delta D = d_{\rm unk} - d_{\rm ref}$ (integer), N is the number of peaks under consideration, $\Delta I = I_{\rm unk} - I_{\rm ref}$, K is a scale factor and $IW = d$ is the error window (integer).

A derivative of the Johnson–Vand program was *μPDSM* (Marquart *et al.*, 1979; Marquart, 1986). This program also used the integer $1000/d$ internally, and considered the probability of the occurrence of a d-spacing in calculating its figure of merit. It used the 15 strongest peaks of the reference patterns in the search step and was the first to make extensive use of pre-screens (especially chemistry) to speed up the search. In addition to the similarity index, other measures of the quality of a match were the numbers of matched and missing lines.

Sometimes, references to 'generations' of search/match programs will be encountered. The first-generation programs include those of Johnson & Vand (1967, 1968), Nichols (1966), Frevel *et al.* (1976), Marquart *et al.* (1979) and O'Connor & Bagliani (1976). The distinction between first- and second-generation programs (Snyder, 1981; Jobst & Goebel, 1982; Huang & Parrish, 1982; Schreiner *et al.*, 1982; Goehner & Garbauskas, 1984; Toby *et al.*, 1990; Caussin *et al.*, 1988) is fuzzy, and is partially a matter of timing and features. Contemporary third-generation programs such as *Jade* (Materials Data, 2016), *EVA* (Caussin *et al.*, 1989; Nusinovici & Bertelmann, 1993; Nusinovici & Winter, 1994), *HighScore* (Degen *et al.*, 2014), *Match!* (Crystal Impact, 2012), *Crystallographica Search-Match* (Oxford Cryosystems, 2012) and *Siroquant* (Sietronics, 2012) are distinguished mainly by the ability to use raw data in addition to peak lists. The presence and absence of peaks in particular regions are both considered in the calculation of the figure of merit. The width of the peak profiles serves as an error window. After the mid-1990s, there is virtually nothing in the open literature about search/match programs, and we are forced to rely on the help documentation of the commercial programs. Occasionally, one will encounter references to a fourth-generation program such as *SNAP* (Barr *et al.*, 2004; Gilmore *et al.*, 2004), *PolySNAP* (Barr *et al.*, 2009) or *FULLPAT* (Chipera & Bish, 2002). There is current development in using similarity indices as a complementary method for the analysis of noncrystalline materials, as these methods depend on whole-pattern fitting instead of peak location and intensity. These methods also cluster isotypical and isostructural crystalline materials, and can be applied to nanomaterial analyses, where there is frequently severe peak overlap.

Originally developed for use with both electron and/or X-ray diffraction data, the Fink search (Bigelow & Smith, 1964) uses the d-spacings of the eight strongest peaks in the pattern, but does not otherwise use the intensities. The justification for not using the intensities was that electron-diffraction intensities were not very reliable, often as a result of poor counting statistics in the small areas analysed in a typical electron-diffraction attachment to a scanning or transmission electron microscope coupled with the effects of dynamical scattering and sample decomposition in the electron beam. The search was named in honour of William H. Fink, a long-time chairman of the JCPDS/ICDD. In the current *SIeve+* module of the PDF, all eight rotations (considering each of the eight peaks as the strongest in turn) are commonly used. *SIeve+* also incorporates a 'Long 8' search, which uses the eight lowest-angle peaks. Fundamentally, searches using electron-diffraction data have deviated from traditional powder-diffraction searches because of the unreliability of both the intensities and the peak locations often brought about by the limited space within an electron microscope. Most modern electron-diffraction searches incorporate elemental data as an integral part of the method. As for X-ray diffraction, there are various generations that integrate elemental composition data, d-spacings or crystallographic data into a search/match process. The *SIeve+* program can also incorporate composition data into the search process.

3.7.2. Powder Diffraction File (PDF)

The PDF is a collection of single-phase X-ray powder patterns in the form of tables of characteristic interplanar spacings and corresponding relative intensities, along with other pertinent physical, chemical and crystallographic properties. The PDF contains various subfiles, which include alkaloids, amino acids, peptides and complexes, battery materials, bioactive compounds, carbohydrates, cement materials, ceramics (bioceramics, ferroelectrics, microwave materials, perovskites and semiconductors), common phases, education, explosives, forensic, hydrogen-storage materials, inorganics, intercalates, ionic conductors, Merck Index compounds, metals and alloys, meso- and microporous (clathrates, metal–organic frameworks and zeolites), mineral-related (minerals, gems, natural and synthetic), modulated structures, nucleosides and nucleotides, organics, pharmaceuticals, pigments and dyes, polymers, porphyrins, corrins and complexes, steroids, superconducting