International Tables for Crystallography (2018). Vol. H, Chapter 3.7, pp. 304–324.

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 ±0.1° 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

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

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 $0 \leq \mathrm{DMC} \leq 1$ for which $I_{\rm PDF}\mathrm{DMC} \leq I_{\rm unk}$ 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,

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

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 $\mu PDSM$ (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 materials (conventional superconductors, superconductor reaction products, superconductor-related and high-T_c superconductors), terpenes and thermoelectric materials. There is an educational package for classroom use, and the complete PDF is available for educational use on a time-limited basis. A primary purpose of the subfile system is to limit the size of the search universe by applying prior knowledge of the system being studied. This greatly reduces the number of false positives in a database that contains hundreds of thousands of materials. Field experts are consulted to guide the criteria for subfile selection, allowing novices to use the subfiles without being a subject expert.

3.7.2.1. *Sources and formats of the PDF*

The data incorporated into the Powder Diffraction File are acquired through contributions from individual scientists, corporate laboratories, literature surveys and a Grant-in-Aid programme. Approximately 200 leading scientific journals are searched manually for powder-diffraction data. Additional literature surveys covering patents, dissertations and the remaining open literature are performed using various online resources and search techniques.

Release 2019 (the current release as of this writing) contains more than 893 400 unique material data sets. The large size and comprehensive coverage of the PDF is achieved through the ICDD's historical sources of powder data (searches of the original literature, contributions and the Grant-in-Aid programme) as well as current and historic collaborations with crystallographic database organizations. Each PDF entry is assigned a unique identifying number of the format *ss-mmm-nnnn*. The integer *ss* indicates the source of the data: 00, ICDD location/generation of powder data; 01, Inorganic Crystal Structure Database; 02, Cambridge Structural Database; 03, NIST (a short-term collaboration focused on metals and alloys); 04, Pearson's Crystal Data; 05, ICDD extraction of atomic coordinates from published sources (including incommensurate/modulated structures). Powder-diffraction data for sources 01 through 05 are computed from the crystal structures provided by these sources.

The Powder Diffraction File is designed and produced in several different formats in order to serve different groups of users. The PDF-2 database is designed for phase identification of inorganic materials; many common organic materials have also been added to this database. The PDF-4+ database is the most advanced database and is designed for both phase identification and quantitative analysis. This database has comprehensive coverage of inorganic materials and contains numerous additional features such as digitized (raw) patterns, molecular graphics and atomic coordinates to facilitate Rietveld refinements. The PDF-4+ database is also available as a portable full-function WEBPDF-4+ version. The PDF-4/Minerals database is a subset of the PDF-4+ database, and is the most comprehensive collection of mineral diffraction data. The PDF-4/Organics database is designed for phase identification of organic and coordination compounds. It contains data from ICDD sources (both experimental powder patterns and extraction of coordinates) as well as patterns calculated from CSD entries.

Advances in hardware, software and computing power have led to the collection of higher-quality powder data, and thus have necessitated higher-quality reference data to perform more complex multiphase analyses and total-pattern analyses. The PDF now includes tools that permit users to evaluate different types of data collected using different types of detectors and different sources, including X-rays, neutrons and electrons. The goal is to include ideal specimen patterns in the PDF, patterns that can be modified by the user to correspond to the current experiment. The user can select the wavelength type and various instrumental parameters to simulate the whole diffraction pattern. A crystallite size calculation was added in 2007 and an orientation function in 2011.

Since 2006, the ICDD has begun to include several types of less-crystalline materials in the database, materials for which too much information is lost when reducing the raw data to a list of *d*-spacings and intensities. These materials include clays and other layered materials, mixed-crystallinity polymers, amorphous materials and nanomaterials.

Nanomaterials often contain crystalline and amorphous fractions, and their powder patterns are difficult to generate from an ideal crystal structure. The ICDD has developed quality-evaluation methods for noncrystalline materials, and has established two additional quality marks: 'good' (G) and 'minimal acceptable' (M). These marks reflect the quality of the supporting data used to characterize the material. An amorphous material with a G quality mark has been characterized by independent analyses verifying the stated composition or thermogravimetric/differential scanning calorimetry analyses confirming the physical stability or the presence of a glass transition. A G quality mark indicates that the editors are satisfied that the pattern is representative of both the diffraction conditions and the stated chemistry and have confidence that the user can reproduce the pattern using similar conditions. The quality mark M indicates that the ICDD received some supporting documentation but it was insufficient for structural interpretation and classification of the material.

Great care needs to be taken in interpreting the patterns of mixtures of crystalline and amorphous phases, particularly in the definition and subtraction of the background. Significant work is under way to develop and adapt numerical techniques for processing full patterns of low-crystallinity materials.

3.7.2.2. *Quality marks in the PDF*

All data are critically reviewed and evaluated by the PDF editorial staff. Each pattern must pass through a four-tiered editorial review process before it can be included in the PDF. As technology evolves, the quality requirements for reference data also evolve. As a result, the information in the PDF is continuously reviewed and upgraded for accuracy and quality.

For many years, a quality mark has been assigned to each experimental PDF entry. A Star (*/S) pattern represents high-quality diffractometer or Guinier data. Several criteria must be satisfied for a pattern to be assigned a Star quality mark:

(i) The chemical composition must be well characterized.

(ii) The intensities must have been measured objectively; no visual estimation is allowed.

(iii) The pattern has a good range and an even spread of intensities.

(iv) The completeness of the pattern is sensible.

(v) The *d*-spacing of each reflection with $d \leq 2.500$ Å is given to at least three decimal places. The *d*-spacings of reflections with $d \leq 1.2000$ Å are given to at least four decimal places.

(vi) No serious systematic errors exist.

(vii) The $|\Delta 2\theta|$ value (*i.e.* the difference between the observed peak position and the position calculated from the unit cell) of a qualifying reflection is $\leq 0.05^\circ$. In the case of multiply-indexed reflections, only the minimum absolute $\Delta 2\theta$ is considered.

(viii) The average $|\Delta 2\theta| \leq 0.03^\circ$ for qualifying reflections.

(ix) No unindexed, space-group-extinct or impurity reflections are present.

An Indexed (I) quality mark indicates that the pattern has been indexed; therefore, the material is almost certainly single-phase. There is a reasonable range and spread of intensities, and the completeness of the pattern is sensible. The d-spacings of reflections with $d \leq 2.000$ Å have at least three significant figures after the decimal point. No serious systematic errors exist. No qualifying reflection has $|\Delta 2\theta| \geq 0.20^\circ$ and the average $|\Delta 2\theta|$ is $\leq 0.06^\circ$. The maximum number of unindexed, space-group-extinct or impurity reflections is two, but none of these reflections are among the eight strongest lines.

A Blank (B) quality mark represents a mid-range quality. An O quality mark means that the data have been obtained from a poorly characterized material or that the data are known (or are suspected) to be of low precision and accuracy. Such patterns include those from multiphase mixtures or from a phase that is poorly characterized chemically. The O mark is commonly assigned to patterns for which no unit cell is reported, unless qualifying information indicates a single-phase material. Usually, the editor will have inserted a comment to explain why the O mark was assigned. For patterns with a unit cell, the following criteria are used to suggest the presence of two or more phases: the number of unindexed, space-group-extinct or impurity reflections is ≥ 3, or one of the three strongest peaks is unindexed.

Beginning with Release 2006, the quality-mark system was extended to patterns calculated from structural data supplied by ICDD partners. The focus of the quality mark is to determine the confidence level of the structural model used and its impact on the calculated pattern (especially for the purpose of phase identification). The major step involves several crystallographic and editorial checks by the ICDD, followed by extraction and flagging of the warnings/comments in the structural databases. The resulting calculated patterns are classified based on the significance and nature of the warnings. Any possible corrections that can be applied to resolve the errors are performed before publishing the calculated pattern.

The crystallographic checking rules are designed based on the expected quality of a contemporary crystal structure. An estimate of the missing electron density is made based on the difference between the reported composition and the structural composition. Transformations of nonstandard space groups are checked; the reported site multiplicities must match those generated by the symmetry operators. All of the eigenvalues of the anisotropic tensor matrix for each atomic displacement must be positive. All anisotropic tensor coefficients must be permitted by the site symmetry. Displacement coefficients should fall in the range $0.001 < U < 0.1$ Å^2. Isotropic displacement coefficients must be positive. Mixed displacement coefficients are converted to a standard type. The reported value of Z must be consistent with the sum of the site multiplicities. Lattice parameters are checked for missing decimal points, missing standard uncertainties and the magnitudes of the uncertainties. R factors close to the theoretical limits (0.83 for centrosymmetric structures and 0.59 for non-centrosymmetric structures) are signs of potential errors in the conversion to/from absolute/percentage values. Site occupancies cannot be greater than 1. Refining part of the structure as a group without locating the positions of the constituent atoms (for example, in C_{60}) will generate a warning. Possible typographical errors in element symbols are checked by comparing the chemical formula, atomic coordinate list and chemical name. When a measured density is available, the percentage difference between the measured and calculated density is determined.

Many warnings/comments from the collaborating databases are used in assignment of the quality mark. Editorial comments on unusually short or long bond lengths or questionable bond angles are considered; the comment needs to be very specific for structures exhibiting disorder or partial/mixed occupancies. A listing of other types of comments considered is contained in the PDF-4+ database help documentation. Entries are assigned a quality mark of * (no warning found during data evaluation), I (minor warning), B (significant warning found), O (major warning), P (the structure was assigned by the editor based on a prototype) or H (hypothetical) according to the criteria in Table 3.7.1.

3.7.2.3. *Features of the PDF*

Most users access the PDF through the software provided by their instrument manufacturer, but it is a powerful standalone database. The PDF is a large relational database consisting of many linked tables. The complete set of features can be accessed through the PDF front end supplied by the ICDD. It is possible to directly access a PDF entry by entering its PDF number. However, one can search for an entry or a class of entries through a series of search tabs. Queries from multiple tabs can be combined in a single search, or individual searches can be saved in a history and combined using Boolean operations. The results of such searches can be analysed as a group or can be used as subfiles for *SIeve*, the search/index phase-identification add-on for the PDF.

Selections on the main search screen permit selection by the source of data, quality mark, primary/alternate, ambient/non-ambient and subfile or subclass. The comprehensive nature of the PDF means that there are often many entries for an individual material. The ICDD editorial staff and volunteer task groups assign one experimental and one calculated entry (if present) as primary entries for each phase so that the user can avoid the duplication if desired. The other entries are designated as alternates. The subfiles and subclasses provide a convenient means for the user to limit the size of the search universe based on prior knowledge and result in faster searches and fewer false-positive matches.

Perhaps the most commonly used screen is the Periodic Table tab for chemistry searches. Individual elements, groups, periods and pre-defined selections (nonmetals, semimetals *etc.*) can be selected and combined in various ways. The ‘and’ operation requires that all selected elements be present in the entries in the selection set, but other elements can also be present. The ‘or’ operation requires at least one of the selected elements to be present. The ‘only’ operation requires that all of the selected elements, and only those elements, be present in the hit. The ‘just’ operation results in a hit list of entries that contain the selected elements in all combinations: elements, binaries, ternaries *etc.* The results of these four types of element searches can also be combined using Boolean operations. An alternative way of using periodic-table screening is through the labelling of each element with ‘yes’, ‘no’ or ‘maybe’ to indicate elements that are known to be present, absent or unsure in the specimen.

The Formula/Name tab facilitates searches on formula, empirical formula, structural formula and formula type ANX [as in the Inorganic Crystal Structure Database (ICSD)]. The formulae may be exact or contain individual elements or strings.

Table 3.7.1
Criteria for the assignment of quality marks to calculated patterns in the Powder Diffraction File

A Star (*/S) pattern has no warnings.

Minor warning (I)	Significant warning (B)	Major warning (O)
Density calculated from reported and calculated compositions differ (1% < x < 3%)	Density calculated from reported and calculated compositions differ (1% < x < 15%)	Density calculated from reported and calculated compositions differ (15% < x)
No e.s.d. reported/abstracted on cell dimensions	Lattice parameters taken from figure (approximated)	Incorrect lattice parameters
Magnitude of e.s.d.s on lattice parameters > 1000 p.p.m.	Missing decimal point in lattice parameter	Incorrect space group
0.07 < R < 0.12 (single crystal), 0.10 < R < 0.15 (powder)	0.12 < R (single crystal), 0.15 < R (powder)	Incommensurate/modulated structure. Only average structure of the subcell is given.
No R reported/abstracted	Anisotropic displacement tensor is non-positive definite	Published atomic coordinates are wrong
Reported Z is inconsistent with the sum of the site multiplicities	Anisotropic tensor coefficient not permitted by site symmetry	Structural database removed the entry corresponding to a published calculated pattern
Type of experiment (single crystal/powder) is not mentioned	Magnitude of displacement coefficients outside the range 0.001 < U < 0.1 Å^2	
Structure corrected by the editor	$U_{\rm iso}$ < 0.0	
Difference between measured and calculated density > 2%	Source-database warning on bond length/angle	
Misprint in original paper corrected in database	Average structure of a modulated structure	
Site occupation factor > 1.0	Probable site-occupation factor deduced from the nominal composition	
	Part of the structure was refined as a group without locating the constituent atoms	
	Comments containing a reference to a contradicting structure exist	
	Structure determined from projections	
	Structure determined using electron diffraction	

Searches on the number of elements present in the compound, as well as composition searches (by weight or atom per cent, *i.e.* wt% or at.%), are also possible.

The Formula/Name tab also permits searches on compound name, common name, mineral name and all names. It is also the screen from which searches on zeolite structure-type code (the International Zeolite Association codes are used) and mineral classification (according to the International Mineralogical Association) are performed.

Under the Reference tab, searches on author, journal name, CODEN, year, volume and title of the paper are possible. The titles were not originally included in PDF entries, but have been added to all entries in recent years. Also possible from this tab are searches on the Chemical Abstracts Service Registry Number (CASRN). Such searches are very powerful for organic compounds, with their complicated nomenclature. CASRNs are present for many, but not all, PDF entries.

The Classification and Crystallography tabs contain searches on Pearson symbol code, space group and space-group number, prototype structure, centrosymmetric/noncentrosymmetric and whether the entry contains atomic coordinates. Searches on the authors' cell, the Pearson's Crystal Data cell or the reduced cell are also possible. I find it useful to use fairly large tolerances (say 0.3–0.5 Å on edges) in such searches. I prefer to examine a longer list of potential matches which contains the correct phase, rather than risk missing an identification.

The Diffraction tab includes searches on the longest (lowest-angle or highest d-spacing) and strongest lines in the pattern. A line can be specified to be one of the three longest/strongest, or the first, second or third. This screen also includes searches on density, I/I_c (which is $I/I_{\rm corundum}$, a measure of the inherent scattering power of a phase and useful in quantitative phase analysis), melting point, R value, colour and Smith–Snyder figure of merit. There are check boxes to select whether the entries in the hit list include 'PD3' patterns (raw data) and property sheets. These property sheets are PDF documents embedded in an entry. These sheets are starting to be included for materials in subfiles that are defined by a particular property, such as superconductivity or thermoelectricity. These sheets are generated by groups of ICDD consulting editors.

Once a hit list has been generated, an individual entry can be selected (double clicked) to bring up the complete PDF entry. The results display can be customized using the Preferences menu (or by right clicking in the entry). The powder pattern can be plotted and additional PDF entries and/or raw data can be overlaid and scaled. In the Plot window, a PDF entry can be exported to several formats. The most useful is a CIF; the crystal structure described by the CIF can then be imported into to the user's graphics or Rietveld package.

Free-text searches of the comments are also included on this screen. These are particularly useful, as ICSD collection codes and CSD refcodes are included in the comments. If the user has the CSD installed on the same machine as the PDF, the PDF entry links live to the coordinates in the CSD entry.

By using the Results menu option when a hit list is displayed, ranges of cells in the spreadsheet can be selected and simple descriptive statistics (mean, median and estimated standard deviation) can be generated. Also under Results is a Graph Fields option. The variables used for the *x* and *y* axes of the plot can be selected and both scatter plots and histograms can be generated. Each of the points in such a plot is 'live' and can be clicked to display the full PDF entry. Fig. 3.7.1 shows a plot of the cubic lattice parameter with respect to at.% Fe in FeO (Fe and O only, space group No. 225) under ambient conditions. From such data it is easy to generate a correlation between the Fe stoichiometry and the lattice parameter.

An optional add-on module to the PDF is *SIeve* (Search Index). This is a peak-based search/match program which enables the use of a manually entered (or imported) peak list or derives a peak list from imported ASCII raw powder-diffraction data in several formats. It also has a flexible ASCII data-import module. Hanawalt, Fink, Long8 (the eight lowest-angle peaks in the pattern) or electron-diffraction searches can be carried out. Again, there is a Preferences option to customize the searches. A particularly useful (and easy-to-use) feature is the ability to apply a filter to the search/match. This filter can be selected from several pre-defined filters and/or any previous search in the session (stored in a history list). The combination of conventional search/match and Boolean searches can be very powerful, as illustrated in the next section.

3.7.2.4. *Boolean logic in phase identification*

Most phase identifications are carried out using the peak-based or full-pattern algorithms supplied by the instrument vendor. These often work well for major phases and can be customized to improve their success in identifying minor/trace phases. The native capabilities of the PDF (not all of which are accessible through some vendors' software) can be very powerful in identifying those extra peaks that result from a Rietveld difference plot (or any difference plot from pattern-fitting software) using the major phases. Below we use examples to illustrate several strategies.

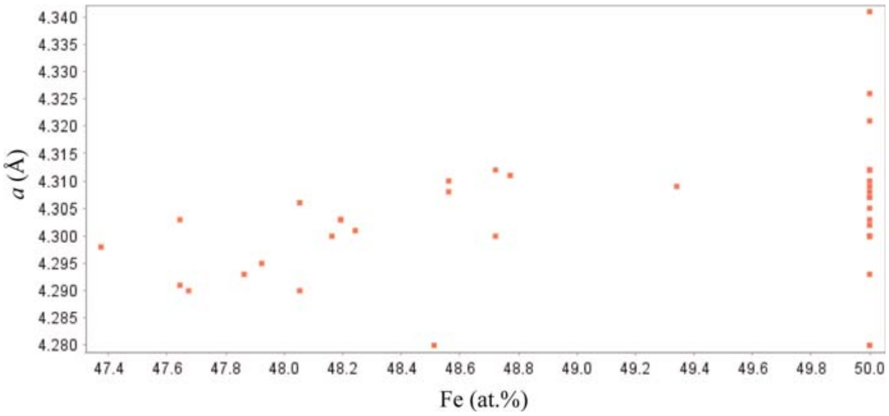


Figure 3.7.1
A plot generated by the Results/Graph Fields function in the Powder Diffraction File. The search was for entries containing only Fe and O and with space group No. 225 (resulting in FeO entries) measured under ambient conditions. One outlier was removed from the hit list manually. The trend in the cubic *a* lattice parameter with Fe content is apparent, as well as the large number of apparently stoichiometric FeO entries, some of which may not be correctly characterized.

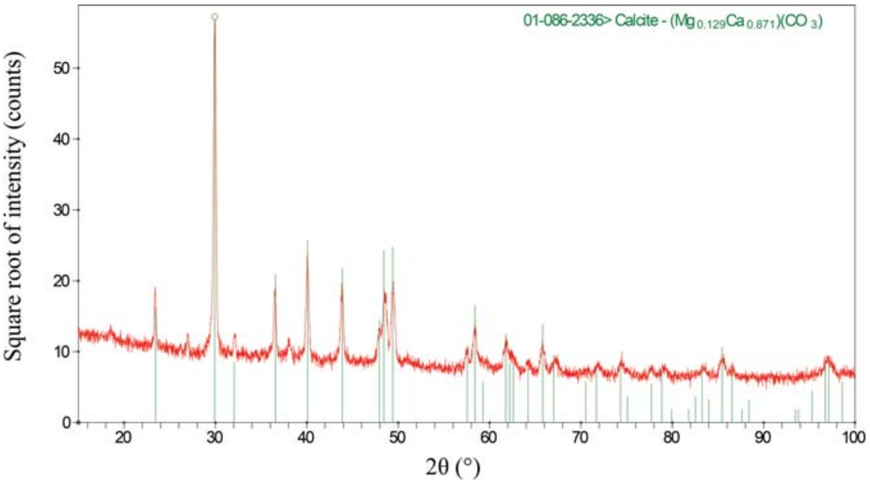


Figure 3.7.2
The result of applying a commercial search/match program (*Jade* 9.5; Materials Data, 2012) to the powder pattern of a water-still scale. Weak peaks not accounted for by the major magnesian calcite phase are apparent and additional tools in the Powder Diffraction File were needed to identify the additional phases.

3.7.2.4.1. *Water-still deposit*

A water still in my home eventually generates scale, much of which flakes off the walls, permitting easy analysis in the powder diffractometer. Any commercial search/match program will easily identify magnesian calcite (Fig. 3.7.2; files kadu1389.gsas, kadu1389.raw and iitd26_0510.prm, available in the supporting information). There are, however, three additional weak peaks at d/I = 4.788/38, 3.3089/51 and 2.3697/56. In Naperville, Illinois, the tap water comes from Lake Michigan. The bedrock underlying the Chicago region is the Racine Dolomite. Given the identity of the major phase in the scale and the source of the water, it seems likely that any minor phases will be mineral-related and contain some combination of the elements Ca, Mg, C, O and H (to include the possibility of hydrates and hydroxides). Accordingly, a search of mineral-related entries containing 'just' the elements Ca, Mg, C, O and H was performed and used as a filter in a Hanawalt search using these three peaks. This limits the search universe to 692 of the 328 660 entries in the PDF-4+ in 2012. The seven highest goodness-of-match entries in the hit list were brucite, $Mg(OH)_2$. This phase was added to the Rietveld refinement. Analysis of the difference plot indicated an unaccounted-for peak at a d-spacing of 3.3089 Å. A search for mineral-related entries with the same chemistry and having one of their three strongest peaks in the range 3.309 (30) Å yielded the vaterite polymorph of $CaCO_3$ as the hit with the highest goodness of match. This phase was added to the Rietveld refinement. The final quantitative phase analysis was: 94.7 (1) wt% $Ca_{0.84}Mg_{0.16}(CO_3)$, 5.2 (4) wt% $Mg(OH)_2$ and 0.2 (1) wt% vaterite.

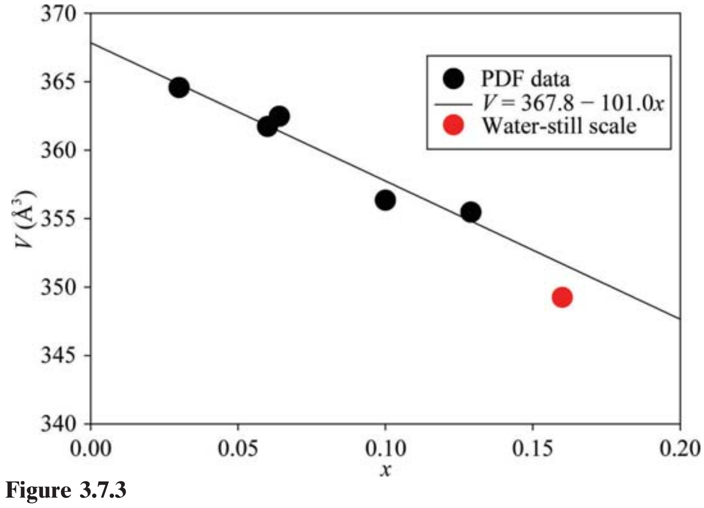


Figure 3.7.3
Variation of the unit-cell volume with the magnesium content in magnesian calcites in the Powder Diffraction File.

The composition of the major phase was refined, constraining the sum of the Ca and Mg site occupancies to equal 1.0. To understand how this fitted with previous magnesian calcites, a search for compounds containing only Ca, Mg, C and O, ambient conditions and space group No. 167 was carried out. Some manual editing of the hit list was required. Adjusting the preferences to include the display of composition in at.% and the unit-cell volume made it convenient to plot the variation in unit-cell volume as a function of Mg content x in $Ca_{1-x}Mg_x(CO_3)$ (Fig. 3.7.3). This magnesian calcite in the water-still scale has a higher Mg concentration than most, but falls close to the trend line. The flexibility and content of the Powder Diffraction File makes such data mining relatively straightforward.

3.7.2.4.2. *Vanadium phosphate butane-oxidation catalyst*

Vanadyl pyrophosphate [$(VO)_2P_2O_7$] catalysts are used commercially for the selective oxidation of butane to maleic anhydride. Modern third-generation search/match programs [using the background-subtracted, $K\alpha_2$-stripped data (files goed80.gsas, GOED80.raw and d8v3.prm); Fig. 3.7.4] had no trouble in identifying the desired major phase $(VO)_2P_2O_7$, but had difficulty with the minor phases that were clearly present. Unless the display of duplicate entries is turned off, most programs will yield several duplicate hits at the top of the list. Both 00-050-0380 and 04-009-2740 are Star quality, but only the Linus Pauling File (LPF) entry 04-009-2740 contains atom coordinates for a Rietveld refinement. Entry 01-070-8726 has the lower-quality B mark.

The native capabilities of the PDF proved helpful in identifying the minor phases. The lowest-angle peak not accounted for by the major phase is at a d-spacing of 7.2107 Å. A search for phases containing just the elements V, P, O and H (known from the synthesis procedure) and having one of their three strongest peaks in the range 7.21 ± 0.05 Å (an estimated range) yielded only the single hit 00-047-0967: $H_4V_3P_3O_{16.5}(H_2O)_2$. This is a low-precision (O quality mark) pattern from a US Patent (Harju & Pasek, 1983), and the pattern contains only four lines. The comments in the PDF entry indicate that this hydrated phase was formed by exposing a catalyst to ambient conditions, so it seems chemically reasonable but poorly defined.

To see whether this phase had been better characterized by a crystal structure, the four peaks were entered into *SIeve+* and a Hanawalt search using a wider than default tolerance of 0.3° on the peak positions and the 'just' chemistry filter V, P, O and H was carried out. As expected, PDF entry 00-047-0967 was at the top of the hit list, but close to the top was entry 04-017-1008 (Shpeizer *et al.*, 2001): $[H_{0.6}(VO)_3(PO_4)_3(H_2O)_3](H_2O)_4$. The article by Shpeizer *et al.* (2001) indicates that this phase was formed from an anhydrous precursor by exposing it to ambient conditions. The single-crystal structure was obtained at 173 K. The similarity of the two PDF entries (Fig. 3.7.5) and the difference in data-collection temperatures makes it clear that these correspond to the same phase, and that the structure of $[H_{0.6}(VO)_3(PO_4)_3(H_2O)_3]$-$(H_2O)_4$ could be used in a Rietveld refinement.

There were still unaccounted-for peaks at 3.5823 and 3.0760 Å. Under the assumption that these came from a single phase, two separate searches for phases containing just V, P, O and H and with one of their three strongest lines in the ranges 3.58 ± 0.03 and 3.08 ± 0.03 Å were carried out and then combined (using the History option) with a Boolean 'and' operation. All five of the entries on the hit list corresponded to α-$VOPO_4$. This yellow V^{5+} compound was consistent with the altered colour of the V^{4+}-based catalyst, and is a common impurity.

Close examination of the Rietveld difference plot from a refinement including these three phases indicated that there was a weak shoulder at a d-spacing of 3.985 Å. A search for phases containing just V, P, O and H and having a strong peak near this d-spacing yielded β-$(VO)(PO_3)_2$, another common catalyst impurity (Fig. 3.7.6). Including this compound as a fourth phase yielded a satisfactory Rietveld refinement and a quantitative analysis of 84.8 (1) wt% $(VO)_2P_2O_7$, 5.9 (1) wt% $[H_{0.6}(VO)_3$-$(PO_4)_3(H_2O)_3](H_2O)_4$, 5.6 (1) wt% α-$VOPO_4$ and 3.7 (1) wt% β-$VO(PO_3)_2$.

3.7.2.4.3. *Valve deposit from a piston aviation engine*

Applying a commercial search/match program to the diffraction pattern of a deposit from a valve in a gasoline-powered aircraft engine easily identified quartz and corundum. The specimen was scraped from the valve seat and micronized. The corundum represents abrasion from the elements of the micronizing mill, as it was not present in the pattern of the as-scraped sample. Metal particles were visibly present in the deposit, so one could reasonably guess the presence of both ferrite and austenite (Fig. 3.7.7; files maso04.gsas, maso04.rd and padv.prm). A Rietveld refinement using these four phases was carried out.

Six peaks picked from the difference plot were entered into *SIeve+* and a Hanawalt search was carried out. No chemically reasonable simple compounds were near the top of the hit list, so extra information was sought. An XPS analysis indicated the presence of Pb, Br, Fe, P, O and C (and H assumed). Aviation gasoline is still leaded, and ethylene dibromide is sometimes added as a lead scavenger. The result of a 'just' chemistry search using these seven elements (6543/328 660 entries) was applied as a filter to the Hanawalt search. Near the top of the hit list was $PbBr_2$. Although apparently surprising, this phase is reasonable given our chemical knowledge. Lead bromide was added to the Rietveld refinement. Further analysis of the difference pattern using the same techniques indicated the presence of cohenite, Fe_3C, from the steel, and $Fe_3Fe_4(PO_4)_6$, the reaction product of the steel with a phosphate fuel additive. The final Rietveld refinement yielded a quantitative analysis of 26.5 (4) wt% austenite (γ-Fe, stainless steel), 47.9 (4) wt% ferrite (α-Fe, carbon steel), 17.7 (4) wt% quartz (sand/dirt), 2.9 (2) wt% $PbBr_2$, 2.6 (2) wt% $Fe_3Fe_4(PO_4)_6$ and 2.2 (2) wt% cohenite (Fig. 3.7.8).

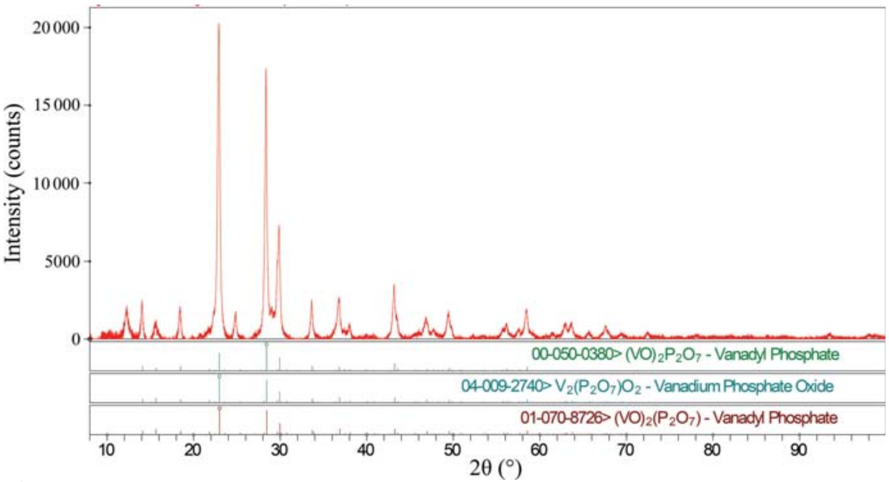


Figure 3.7.4
The results of applying a commercial search/match program (*Jade* 9.5; Materials Data, 2012) to the (background-subtracted, $K\alpha_2$-stripped) powder pattern of a butane-oxidation catalyst. The first three patterns in the hit list had equivalent figures of merit. The PDF entries 00-050-0380 and 04-009-2740 had Star quality marks and 04-009-2740 contained the atomic coordinates necessary for a Rietveld refinement. Additional peaks are apparent. The phases that give rise to them were identified using the native capabilities of the Powder Diffraction File.

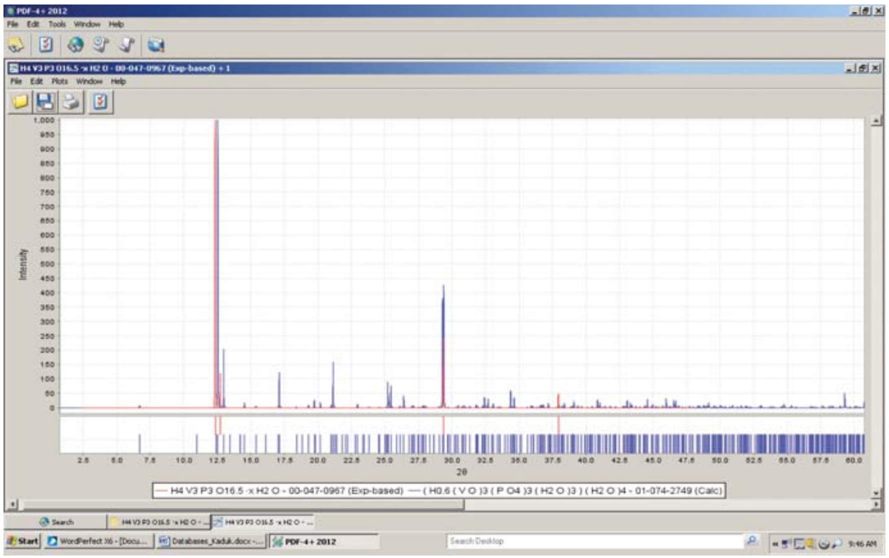

PDF-4+ 2012
File Edit Tools Window Help
H4 V3 P3 O16.5 ·x H2 O - 00-047-0967 (Exp-based) + 1
File Edit Plots Window Help
Intensity
1,000
950
900
850
800
750
700
650
600
550
500
450
400
350
300
250
200
150
100
50
0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5
25.0
27.5
30.0
32.5
35.0
37.5
40.0
42.5
45.0
47.5
50.0
52.5
55.0
57.5
60.0
2θ
— H4 V3 P3 O16.5 ·x H2 O - 00-047-0967 (Exp-based) — (H0.6 (V O)3 (P O4)3 (H2 O)3) (H2 O)4 - 01-074-2749 (Calc)
Search
Start
WordPerfect X6 - [Docu...
Databases_Kaduk.docx -...
PDF-4+ 2012
Search Desktop
9:46 AM

Figure 3.7.5
Comparison of the low-quality experimental PDF entry 00-047-0967 with the high-quality calculated pattern 01-074-2749 located by searching the experimental pattern against the rest of the PDF. The similarity in patterns and chemistry demonstrated that the two phases were the same and that the coordinates used to calculate entry 01-074-2749 could be used in a Rietveld refinement of a butane-oxidation catalyst.

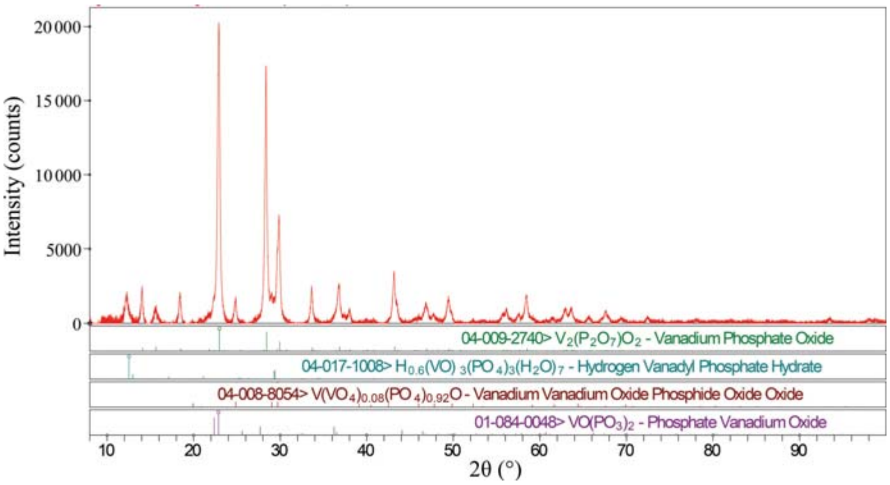


Figure 3.7.6
The four crystalline phases identified in a butane-oxidation catalyst.

3.7.2.4.4. *Isocracker sludge*

An isocracker is a refinery unit which simultaneously carries out cracking and isomerization reactions to produce more high-octane gasoline. A black deposit isolated from such a unit was surprisingly crystalline (Fig. 3.7.9; files NALK157.gsas, NALK157.raw and padv.prm). It was easy to identify small concentrations of elemental sulfur, pyrrhotite-4M (now called pyrrhotite-4C), haematite, lepidocrocite and dolomite, but the major peaks did not match well those of any entry in the PDF.

It seemed likely that a mineral-related phase would serve as a structural prototype for an apparently new phase, so two separate searches for mineral-related phases with one of their three strongest peaks in the *d*-spacing ranges 7.09 ± 0.03 and 5.57 ± 0.03 Å were combined. The two hits in the search list were both uranium minerals. These seemed unlikely in a refinery deposit(!). Widening the search ranges to 7.09 ± 0.10 and 5.57 ± 0.07 Å yielded rasvumite, KFe_2S_3 (PDF entry 00-033-1018), as the second entry in the hit list.

The fit to the major peaks in the deposit was reasonable, but there should not be any potassium in a refinery deposit and none was detected in a bulk chemical analysis. When the jar containing the deposit was opened, it smelled strongly of ammonia. Ammonium and potassium ions are about the same size and often form isostructural compounds. The infrared spectrum of the deposit was dominated by bands of ammonium ions.

The potassium in the structure of rasvumite (PDF entry 01-083-1322, used as a reference) was replaced by nitrogen. Analysis of potential hydrogen-bonding interactions yielded approximate hydrogen positions in the ammonium ion. These positions were refined using a density-functional geometry optimization. This model yielded a satisfactory Rietveld refinement (Fig. 3.7.10) and the quantitative analysis 45.7 (2) wt% $(NH_4)Fe_2S_3$, 12.8 (4) wt% S_8, 22.0 (6) wt% lepidocrocite (γ-FeOOH), 5.5 (5) wt% haematite (α-Fe_2O_3), 6.6 (3) wt% pyrrhotite-4C (Fe_7S_8) and 6.6 (3) wt% dolomite [$CaMg(CO_3)_2$; limestone environmental dust]. The powder pattern and crystal structure of $(NH_4)Fe_2S_3$ are now included in the PDF as entry 00-055-0533.

3.7.2.4.5. *Amoxicillin*

The amoxicillin powder from a commercial antibiotic capsule was highly crystalline. Its powder pattern (files kadu918.gsas, KADU918.raw, d8v3.prm and KADU921.rd) was matched well by the PDF entries 00-039-1832 and 00-033-1528 for amoxicillin trihydrate, but there was an additional peak at a *d*-spacing of 16.47 Å (5.37° 2θ). With such a low-angle peak, it seemed prudent to measure the pattern again

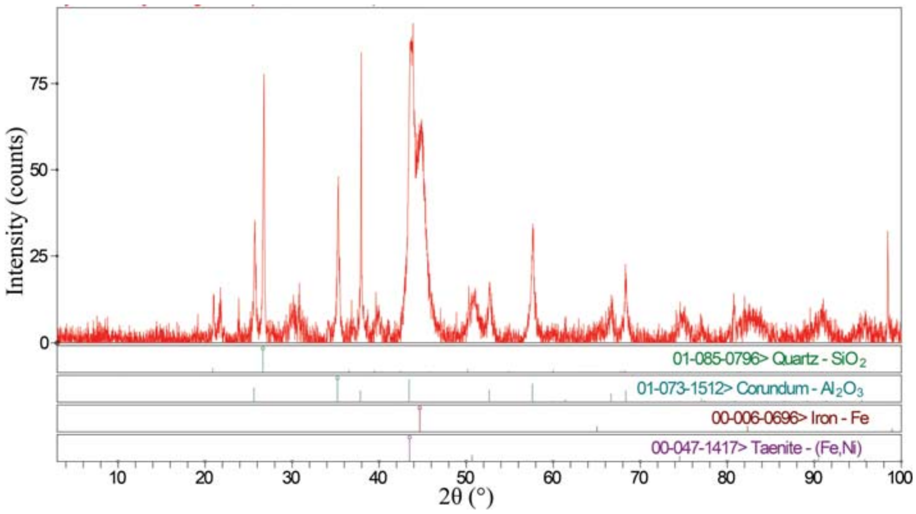


Figure 3.7.7
The four phases identified in a valve deposit from an aircraft engine by automated search/match methods and guessing based on the appearance of the sample. The pattern has had the background and $K\alpha_2$ peaks removed.

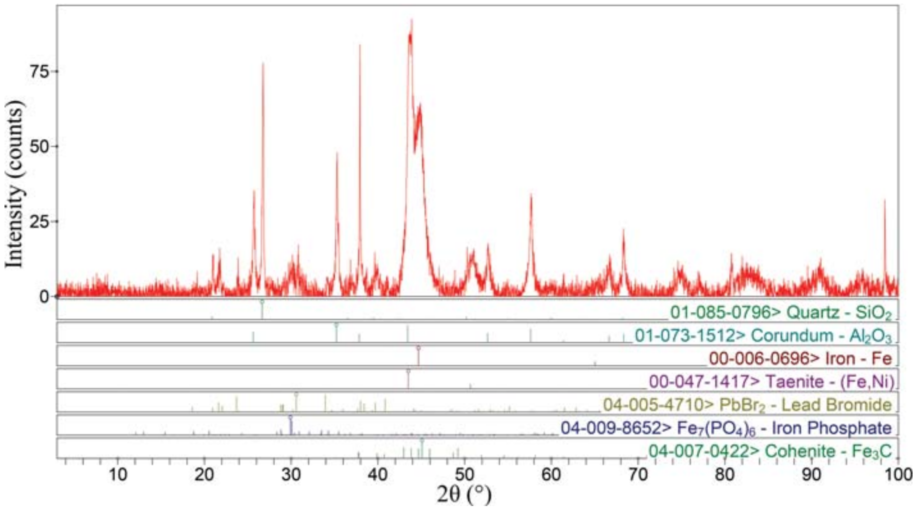


Figure 3.7.8
The seven phases identified in the valve deposit from an aircraft engine.

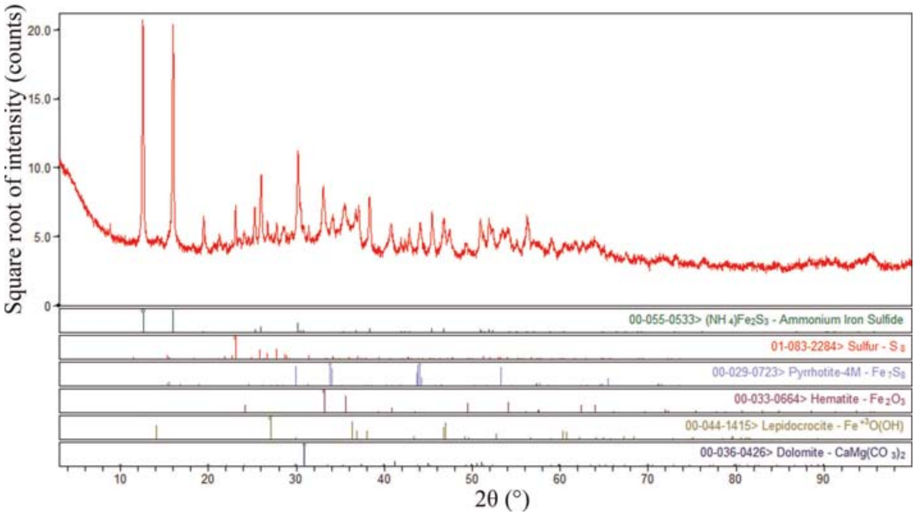


Figure 3.7.9
The phases identified in a deposit from a refinery isocracker. At the time, the (NH$_4$)Fe$_2$S$_3$ was a new phase, identified by analogy to KFe$_2$S$_3$, rasvumite.

starting at 3°, and another peak was observed at d = 24.80 Å (3.56° 2θ).

A search of the PDF-4/Organics 2013 for phases having two such peaks among their longest (lowest-angle) peaks yielded entry 00-005-0010 for calcium stearate at the top of the hit list, as well as two lead stearates. We can safely assume that lead stearate is not present in a pharmaceutical. Calcium stearate, however, has its strongest peak at 1.76°, so another pattern was measured starting at 1.5° 2θ. This peak is indeed present (Fig. 3.7.11).

The primary literature suggests that the compound in PDF entry 00-005-0010 is really calcium stearate monohydrate, and that its structure (like those of many other stearate salts) has not yet been determined. The CSD entry for amoxicillin trihydrate (AMOXCT10; Boles *et al.*, 1978) contained some incorrect H-atom positions and was missing an H atom, so these were corrected before a Rietveld refinement was carried out.

3.7.2.4.6. *Pseudoephedrine*

As P. W. Stephens was measuring the powder pattern of a commercial pseudoephedrine-based decongestant on beamline X16C at the National Synchrotron Light Source at Brookhaven National Laboratory, he noted that extra peaks were present. The lowest-angle peak was at a d-spacing of 12.73 Å, and other peaks occurred at 5.74, 4.62 (strongest) and 4.407 Å. A search in the PDF-4/Organics for compounds having the string ‘ephed’ in the name, a long line at 12.73 ± 0.05 Å and a strong line at 4.62 ± 0.02 Å yielded the single hit 00-041-1946, pseudoephedrine hydrochloride, a reasonable impurity in pseudoephedrine.

3.7.2.4.7. *Commercial multivitamin: Centrum A to Zn*

Commercial multivitamins are challenging phase-identification problems because they contain small concentrations of many different components. The application of a commercial search/match program to a pattern of Centrum A to Zn collected on beamline ID-32 at the Advanced Photon Source at Argonne National Laboratory using a wavelength of 0.495850 Å (files centrum.gsas and id320304.prm) easily identified brushite, CaHPO$_4$(H$_2$O)$_2$, and sylvite, KCl (Fig. 3.7.12).

To identify additional phases, 64 peaks with d > 1.91 Å were picked from the plot and entered into *SIeve+* in the PDF-4/Organics 2013 database. The PDF-4/Organics database was used to enhance the success in identifying organic compounds, and the relatively short d-spacing limit was used to ease the identification of the simple inorganic compounds which are often present in commercial vitamins.

A Hanawalt search using these peaks easily identified iron fumarate (00-062-1294), szmikite [$MnSO_4(H_2O)$; 00-033-0906], L-ascorbic acid (02-063-2295), monetite ($CaHPO_4$; 01-070-0359) and calcite ($CaCO_3$; 00-005-0586). Note that these hits come from four different data sources; searches based on just one source would not have identified all of these compounds.

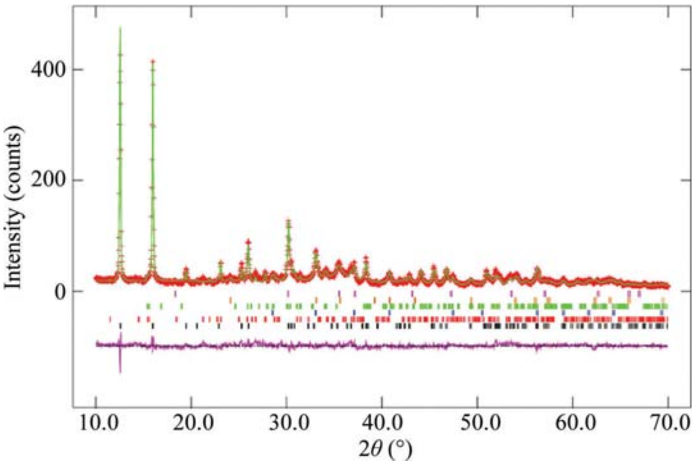


Figure 3.7.10
The final Rietveld plot from refinement of the isocracker deposit.

There were strong high-angle peaks that had not yet been accounted for at *d*-spacings of 2.4762, 2.1068, 1.4900 and 1.4783 Å. These four peaks were entered into a new Hanawalt search, which identified periclase (MgO; 01-071-3631) and zincite (ZnO; 01-075-9742).

Superimposing the peaks for all of these compounds onto the raw data made it clear that there were broad peaks in the pattern at *d*-spacings of approximately 5.8750, 5.3273, 4.3277 and 3.9217 Å. Since the lowest and highest angles of these four were the best defined, separate searches for compounds having each of these peaks as one of their three strongest lines were combined using a Boolean 'and'. Among the hit list was cellulose I_β (00-060-1502), which is a common constituent of pharmaceuticals. The structure model from PDF entry 00-056-1718 was added to the Rietveld refinement as a ninth phase.

One last peak at 5.9915 Å was unaccounted for. A search for pharmaceutical-related compounds with this peak as one of the three strongest included nicotinamide (02-063-5340; niacin or vitamin B_3). Ten phases were thus identified and these account for all of the peaks in the pattern.

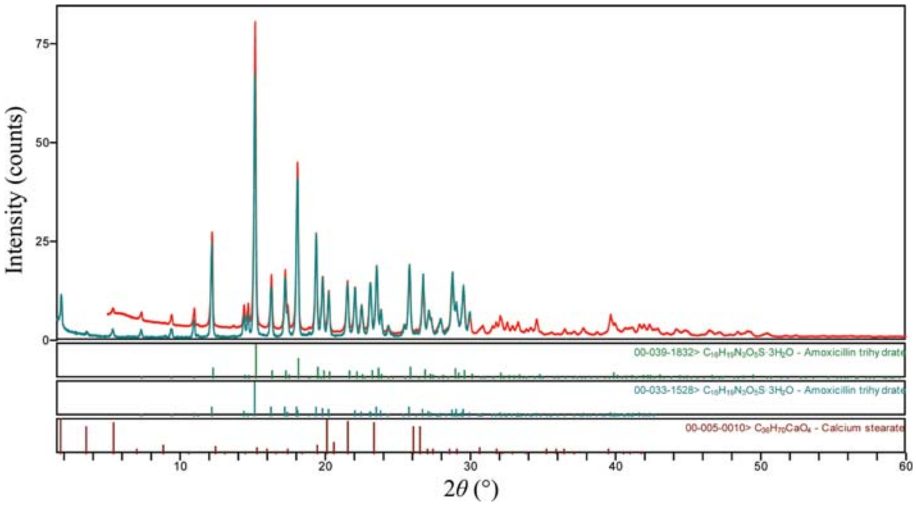


Figure 3.7.11
Phases identified in amoxicillin powder from a commercial capsule.

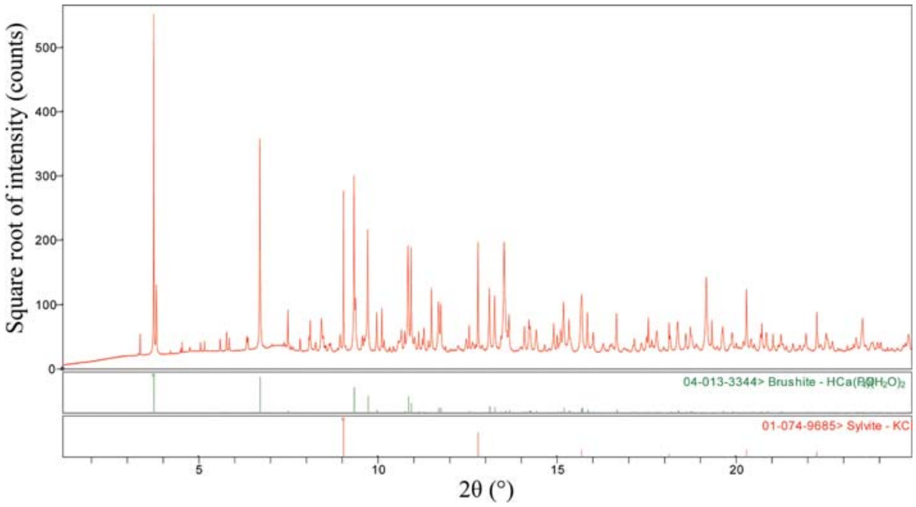


Figure 3.7.12
Phases identified by automated search/match in a Centrum A to Zn multivitamin tablet. Additional phases were identified using the native capabilities of the Powder Diffraction File.

3.7.3. Cambridge Structural Database (CSD)

Some features of the Cambridge Structural Database system (CSD; https://www.ccdc.cam.ac.uk; Groom *et al.*, 2016) are described in Chapter 22.5 of *International Tables for Crystallography* Volume F (Allen *et al.*, 2011). The CSD contains X-ray and neutron diffraction analyses of carbon-containing molecules with up to 1000 atoms (including hydrogens), including organic compounds, compounds of the main-group elements, organometallic compounds and metal complexes. The CSD covers peptides of up to 24 residues; higher polymers are covered by the Protein Data Bank. The CSD also covers mononucleotides, dinucleotides and trinucleotides; higher oligomers are covered by the Nucleic Acid Database (http://ndbserver.rutgers. edu). There is a small overlap between the CSD and the Inorganic Crystal Structure Database in the area of molecular inorganics.

Capabilities particularly useful for structure validation are covered in Chapter 4.9 of this volume. This discussion will not attempt a comprehensive description of the capabilities of the CSD, but will concentrate on features that are particularly relevant to powder diffraction.

The principal interface to the CSD is the program *ConQuest* (Bruno *et al.*, 2002). Its most distinctive feature is the ability to draw molecular structures and fragments and carry out substructure searches. Such searches eliminate the ambiguities that can arise when searching by compound name or other text-based properties. These chemical-connectivity searches can include the number of H atoms bonded to a particular atom, the charge, the number of bonded atoms and whether the atom is part of a ring. In addition, three-dimensional quantities can be defined, tabulated and analysed. These quantities can be analysed in *Mercury* (Macrae *et al.*, 2008) and *Mogul* (Bruno *et al.*, 2004). Such analyses are useful for defining geometrical restraints in a Rietveld refinement. A general practice is to use the mean and standard deviations directly output by *Mogul* for the restraints. It is important to understand the CSD conventions for defining bond types to obtain successful results.

In *ConQuest*, searches can be carried out on author and/or journal name, as well as the normal bibliographic characteristics. Compounds can be located by chemical and/or common names, but such searches should be complemented by chemical-connectivity searches. It is possible to limit the search universe by chemical class, including carbohydrates, nucleosides and nucleotides, amino acids, peptides and complexes, porphyrins, corrins and complexes, steroids, terpenes, alkaloids and organic polymers. Searches on elements and formulae are possible, as well as searches on space groups and crystal systems. Particularly useful in searching for structure analogues are reduced-cell searches. Queries on Z, Z' and density are useful in data mining. A wide variety of searches on experimental parameters are possible; there is an option to exclude powder structures. Searches on both pre-defined terms and general text searches are possible. Particularly convenient for users of the PDF-4/Organics database is the retrieval of individual refcodes; the refcode from the PDF entry can be input directly into *ConQuest*. Starting with the 2013 release of the PDF-4/Organics database this link is live; the display of a PDF entry will result in import of the coordinates from the CSD entry. Boolean operations can be used to combine search queries in many flexible ways.

In recent years, many (if not most) single-crystal structures have been determined at low temperatures, while most powder-diffraction measurements are made under ambient conditions. Thermal expansion (often anisotropic) can result in differences between the observed peak positions and those in a PDF-4/Organics entry calculated from a CSD entry. For successful phase identification, larger than default tolerances must often be used in the search/match process. Transparency effects in pure organic compounds can also lead to significant peak shifts to lower angles, as well as significant asymmetry, so wider search windows may be necessary for phase identification.

3.7.3.1. *Mercury*

The structure-visualization program *Mercury* (Macrae *et al.*, 2008) is available as a free version and as a version with the CSD which has additional capabilities that are useful for powder diffraction. *Mercury* reads and writes a variety of molecular and crystal structure file formats, but is most commonly used with CIFs. Structures can be edited, and among the edit options is Normalize Hydrogens. This is particularly useful to improve the approximate H-atom positions that are often used in the early and intermediate stages of a Rietveld refinement. It is always worth including the H atoms in the structure model (in at least approximate positions) because better residuals and improved molecular geometry are obtained.

The Display Symmetry Elements tool is particularly useful for teaching symmetry. The Display Voids tool is useful in validating structures after solution. For most materials (zeolites and metal–organic frameworks are notable exceptions) we do not expect empty spaces in the crystal structure, so the presence of voids suggests the presence of an incomplete structure model and/or errors.

Among the options in the Calculate menu is Powder Pattern. The calculation can be customized by the user to match the desired instrumental configuration. The calculated pattern can be saved in several formats for comparison in the user's instrument software. *Mercury* expects the displacement coefficients to be given as U values. CIFs can come from many sources and can use different conventions for the displacement coefficients or may be missing them entirely. Manual editing of the input CIF is often required, otherwise strange powder patterns can be calculated.

Among the CSD-Materials/Calculations options is BFDH morphology (Bravais, 1866; Friedel, 1907; Donnay & Harker, 1937). Although a simple calculation, it is often realistic enough to suggest the likelihood of profile anisotropy and preferred orientation, along with expected directions. The calculation can thus save guessing about preferred directions.

The Structure Overlay and Molecule Overlay options are very useful for comparing structures quantitatively. There is also an interface to the semi-empirical code *MOPAC*, which can also be a useful tool for assessing structural reasonableness. The H-Bonds and Short Contacts options are useful in completing a structure solved using powder data, as often the 'interesting' H-atom positions have to be deduced. There is a relatively new Solid Form menu, which contains several tools for analysing crystal structures.

3.7.4. Inorganic Crystal Structure Database (ICSD)

The Inorganic Crystal Structure Database (ICSD; https://icsd.fiz-karlsruhe.de; Bergerhoff & Brown, 1987; Belsky *et al.*, 2002; Hellenbrandt, 2004) strives to contain an exhaustive collection of inorganic crystal structures published since 1913, including their atomic coordinates. It is a joint project between FIZ Karlsruhe and NIST. The database is accessed through the online WebICSD or the locally-installed program *FINDIT*. Typical interatomic distances in inorganic compounds derived from the ICSD have been collected in Chapter 9.4 of *International Tables for Crystallography* Volume C (Bergerhoff & Brandenburg, 1999). Applications of the ICSD have been discussed by Kaduk (2002), Behrens & Luksch (2006) and Allmann & Hinek (2007).

The ICSD began as an inorganic crystal structure database of published structures with atomic coordinates. The scope was gradually extended to include intermetallic compounds. Since 2003, FIZ Karlsruhe has started to fill in the gaps, and the aim is for the ICSD to include all published intermetallic compounds. Originally the ICSD did not contain structures with C—H or C—C bonds. After 2003, this rule was modified so that new entries should not contain both C—H and C—C bonds; compounds containing tetramethylammmonium and oxalate ions are now included.

The ICSD contains fully determined structures with atomic coordinates. Coordinates of light atoms (such as H atoms) or extra-framework species (such as in zeolites) may be missing. Structures described as isotypic to other structures, but without determination of the atomic coordinates, are included using the coordinates from the corresponding structure-type prototype. Such entries get a special remark/comment: 'Cell and Type only determined by the author(s). Coordinates estimated by the editor in analogy to isotypic compounds.' Currently there are more than

26 000 entries with derived coordinates. At present, the ICSD contains more than 187 000 entries, including 2033 crystal structures of elements, 34 785 records for binary compounds, 68 730 records for ternary compounds and 68 083 records for quaternary and quintenary compounds. About 149 000 entries have been assigned a structure type; there are currently 9093 structure prototypes.

Most of the structures contained in the ICSD are from published journal articles, although private communications are also accepted. The entries are tested for formal errors, plausibility and logical consistency. The data are stored as published; the authors' settings of space groups are considered to be valuable information which should not be changed. Only some 'exotic' space groups are transformed. In addition, for each entry in the ICSD the structure is standardized using the program *STRUCTURE TIDY* by Gelato & Parthé (1987). The published cell, standardized cell and reduced cell are all searchable. Since 2003, FIZ Kalrsruhe has been assigning structure-type classifications (Allmann & Hinek, 2007). In the future, this feature will enable easier searches for compounds that are closely related in structure.

3.7.4.1. *General features of the ICSD*

The chemical name is given in English following IUPAC rules, with the oxidation state in roman numerals. The formula upon which the name is based is calculated from all atoms with defined coordinates. Phase (polymorph) designations are given after a hyphen. Mineral names and group names are given for all entries that correspond to minerals. Details of the origin are given after a hyphen. The formula is coded as a structural formula, which provides the opportunity to search for typical structure units (such as SiO_4). Such searches can be useful, but can easily miss structurally similar compounds, and should be used with caution.

The title of the publication is given in English, French or German. There can be several citations, but an author list is only given for the first reference. I have encountered truncated author lists. Authors' surnames can vary when the original publication uses a non-roman alphabet. In some cases, the first and last names of Chinese authors may be interchanged.

The Hermann–Mauguin space-group symbol is given according to the conventions of *International Tables for Crystallography* Volume A. If different origin choices are available, those space groups with the origin at a centre of symmetry (origin choice 2) are characterized by an additional '*z*', while an additional '*s*' is used for special origins (origin choice 1). Thus, the space group for magnetite may be reported as *Fd*-3*mz* or *Fd*-3*ms*, depending on which origin the authors used. Since all contemporary Rietveld programs use origin choice 2, care must be taken when importing coordinates.

Along with the fractional coordinates, atom identifiers are reported. These are principally running numbers and may differ from those reported by the authors. The oxidation state is given with a sign. When importing coordinates into a Rietveld program these oxidation states can influence which scattering factors are used, and so should be examined by the user. Both site multiplicities and Wyckoff positions are generated for all atoms.

The ICSD archives displacement coefficients (both isotropic and aniostropic) according to what the authors reported. Isotropic displacement coefficients can be given as either *B* or *U* values and anisotropic coefficients can be given as *β*, *B* or *U* values (or, in rare cases, using other conventions). Displacement coefficients imported into a Rietveld program should always be checked, as it is common for the program to interpret *B* as *U* and *vice versa*. Such wrong displacement coefficients can make Rietveld refinements hard to perform. There are a number of standard remarks and standard test codes; these text fields can be useful for limiting the universe of the search (such as for neutron-diffraction structures).

3.7.4.2. *Features particularly useful for powder crystallography*

A field which is particularly useful for identifying structural analogues is the ANX formula. This formula is generated according to the following rules:

(i) H^+ is not taken into account, even if coordinates are available.

(ii) The coordinates of all sites of all other atoms must be determined.

(iii) Different atom types on the same positions (for example, in solid solutions) are treated as a single atom type.

(iv) An exception: if cations and anions occupy the same site they will not be treated as one atom type.

(v) All sites occupied by the same atom type are combined, unless the oxidation state is different. Thus, $Fe^{2+}(Fe^{3+})_2O_4$ yields AB2X4, while $(Fe^{2.667+})_3O_4$ yields A3X4.

(vi) For each atom type, the multiplicities are multiplied by the site-occupancy factors and the products are added. The sums are rounded and divided by the greatest common divisor.

(vii) If the rounded sum equals zero, all sums are multiplied by a common factor so that the smallest sum equals unity, so no element will be omitted.

(viii) Cations are assigned the symbols A–M, neutral atoms are assigned N–R and anions are assigned X, Y, Z and S–W.

(ix) The symbols are sorted alphabetically and the characters are assigned according to ascending indices: AB2X4, not A2BX4.

(x) All ANX symbols with more than four cation symbols, three neutral atom symbols or three anion symbols are deleted.

The utility of these symbols is illustrated by the fact that the three garnets $Mg_3Al_2(SiO_4)_3$, $Ca_3(Al_{1.34}Fe_{0.66})Si_3O_{12}$ and $(Mg_{2.7}Fe_{0.3})(Al_{1.7}Cr_{0.3})Si_3O_{12}$ all yield ANX = A2B3C3X12.

Reduced-cell searches [see *International Tables for Crystallography* Volume A, Section 3.1.3 (de Wolff, 2016)] are particularly easy to carry out in the 'Cell' section of Advanced Searches. Once a unit cell has been determined by indexing the powder pattern, it is always worth carrying out a reduced-cell search to identify potential isostructural compounds using lattice-matching techniques. It is often wise to first carry out such a search using relatively narrow tolerances (say, 1% on the lattice parameters) and then carry out additional searches using larger tolerances. Systematic searches of the subcells and supercells of a given unit cell, as could be carried out using the *NBS*LATTICE* program with the NIST Crystal Data Identification File (Mighell & Himes, 1986; Mighell, 2003), are not yet implemented.

Under the 'Crystal Chemistry' section it is possible to search for crystal structures that contain bonds between particular atom types in a distance range. Such searches are particularly valuable in assessing the chemical reasonableness of crystal structures, such as the study by Sidey (2013) on the shortest $B^{III}-O$ bonds.

Because the ICDD Powder Diffraction File '01' entries contain the ICSD collection code in the comments, searching for the collection code of a hit in a search/match is particularly easy in the 'DB Information' section. In this way, the relevant ICSD entry can be located without any ambiguity and the best structure for the problem at hand can be used to start the Rietveld refinement.

3.7.5. Pearson's Crystal Data (PCD/LPF) (with Pierre Villars and Karen Cenzual)

3.7.5.1. *General information*

The Pearson's Crystal Data database (PCD; Villars & Cenzual, 2013) is an outgrowth of the (Linus) Pauling File (LPF; Villars *et al.*, 1998; http://www.paulingfile.com), which was designed to combine crystal structures, phase diagrams and physical properties under the same computer framework to form a tool useful for materials design. PCD is the result of a collaboration between Material Phases Data Systems (Vitznau, Switzerland) and ASM International (Materials Park, Ohio, USA). The retrieval software was developed by Crystal Impact (Bonn, Germany). As suggested by the name, Pearson's Crystal Data is a follow-up product to *Pearson's Handbook: Crystallographic Data for Intermetallic Phases* (Villars & Calvert, 1985, 1991; Villars, 1997). However, in contrast to the latter, it also covers oxides and halides, which represent about 80% of the compounds with more than four chemical elements.

The 2016/2017 release of Pearson's Crystal Data contains more than 288 000 data sets for more than 165 300 different chemical formulae, representing over 53 000 distinct chemical systems. To achieve this, the editors have processed over 93 500 original publications; recent literature is surveyed in a cover-to-cover approach, including about 250 journal titles. Over 153 000 database entries contain refined atom coordinates, as well as isotropic and/or anisotropic displacement parameters when published, whereas more than 72 000 data sets contain atom coordinates corresponding to the structure prototype assigned by the authors of the original publication or by the database editors. Approximately 15 000 data sets contain only crystallographic data such as the lattice parameters and possibly a space group.

When available in the original publications, each data set contains comprehensive information on the sample-preparation and experimental procedure, as well as on the stability of the phase with respect to temperature, pressure and composition. The presence of plots (cell parameters or diffraction patterns) in the original paper is indicated, and over 30 000 descriptions of the variation of the cell parameters as a function of temperature, pressure or composition are proposed. Roughly 18 300 experimental diffraction patterns are reported.

The Linus Pauling File was designed as a phase-oriented, fully relational database system. This required the creation of a 'distinct phases' table, with internal links between the three parts of the database. In practice, this means that the senior editors have evaluated the distinct phases existing in the system for every chemical system using all information available in the LPF. Each structure entry in Pearson's Crystal Data has been linked to such a distinct phase, which allows a rapid overview of a particular chemical system.

3.7.5.2. *Evaluation procedure*

Extensive efforts have been made to ensure the quality and reliability of the crystallographic data. Pearson's Crystal Data is checked for consistency by professional crystallographers, assisted by an original software package, *ESDD* (*Evaluation, Standardization and Derived Data*), containing more than 60 different modules (Cenzual *et al.*, 2000). The checking is carried out progressively, level by level. The following checks are made.

Individual database fields:

(i) order of magnitude of numerical values;
(ii) Hermann–Mauguin symbols, Pearson symbols;
(iii) consistency of journal CODEN, year, volume, first page, last page;
(iv) formatting of chemical formulae;
(v) neutrality of oxides and halides;
(vi) spelling.

Consistency within individual data sets:

(i) atom coordinates, Wyckoff letters, site multiplicities;
(ii) chemical elements in different database fields;
(iii) computed, published values (cell volume, density, absorption coefficient, *d*-spacings);
(iv) Pearson symbol, space group, cell parameters;
(v) Bravais lattice, Miller indices;
(vi) site symmetry, anisotropic displacement parameters.

Particular crystal-structure checks:

(i) interatomic distances, sum of atomic radii;
(ii) geometry of functional groups;
(iii) search for overlooked symmetry elements;
(iv) composition from refinement, chemical formula.

Consistency within the database:

(i) comparison of cell-parameter ratios for isotypic entries;
(ii) comparison of atom coordinates for isotypic entries with refined coordinates;
(iii) comparison of densities;
(iv) thorough search for duplicates, also considering translated references.

Wherever possible, misprints have been corrected based on arguments explained in remarks; as a result, more than 13 000 crystallographic data sets are accompanied by at least one erratum. In other cases remarks drawing the attention to discrepancies or unexpected features have been added.

The *ESDD* software package also produces derived data such as the Niggli reduced cell, equivalent isotropic displacement parameters, density and formula weight.

3.7.5.3. *Standardized crystallographic data*

The crystallographic data in Pearson's Crystal Data are presented as published, respecting the original site labels, but are also standardized following the method proposed by Parthé and Gelato (Parthé & Gelato, 1984, 1985; Parthé *et al.*, 1993). This second presentation of the same data has been further adjusted so that compounds crystallizing with the same prototype structure (isotypic compounds) can be easily compared. It is prepared in a three-step procedure as follows.

(i) The crystallographic data are checked for the presence of overlooked symmetry elements. Whenever it is possible to describe the structure in a higher-symmetry space group, or with a smaller unit cell, without any approximations, this is performed.
(ii) In the next step, the crystallographic data are standardized using the program *STRUCTURE TIDY* (Gelato & Parthé, 1987).
(iii) The resulting data are compared with the standardized data of the type-defining data set and, if relevant, adjusted using an *ESDD* module based on the program *COMPARE* (Berndt, 1994).

For data sets with no published coordinates, the cell parameters are standardized following the criteria defined for the unit-cell and space-group setting. For data sets with unknown space group, the cell parameters have been standardized assuming the space group of lowest symmetry in agreement with the Pearson symbol, *e.g.* *P*222 for *oP** or *o***.

Standardized data are described with respect to the standard settings described in *International Tables for Crystallography* Volume A, with the following additional restraints: inversion centre at the origin, unique *b* axis and 'best' cell for monoclinic structures (Parthé & Gelato, 1985), triple-hexagonal cell for rhombohedral structures or Niggli reduced cell for triclinic structures. As a consequence, they can easily be incorporated into any program handling crystallographic data. The systematic standardization of the crystallographic data also greatly simplifies the classification of crystal structures into different prototypes.

A conversion tool to standardize cell parameters and/or compute the Niggli reduced cell is included in the software of Pearson's Crystal Data.

3.7.5.4. *Consequent prototype assignment*

The prototype is a well known concept in inorganic chemistry, where a large number of compounds often crystallize with very similar atom arrangements. The compilation *Strukturbericht* started to catalogue crystal structures into types named by codes such as A1, B1 or A15. These notations are still in use; however, today prototypes are generally referred to by the name of the compound for which this particular kind of atom arrangement was first identified, *i.e.* Cu, NaCl and Cr_3Si for the types enumerated above. Pearson's Crystal Data uses a longer notation which also includes the Pearson symbol and the space-group number: Cu,cF4,225, NaCl,cF8,225 and Cr_3Si,cP8,223. In a few cases several prototypes correspond to the same code, for example several polytypes of CdI_2 have the same notation. A similar situation occurs for the wrong and the correct structure proposals for FeB, which have the same Pearson code and space group. In these cases a letter is added after the type-defining compound, for example the correct FeB type will be referred to as FeB-b,oP8,62.

Each prototype is defined on a particular PCD database entry. In principle, this data set represents a recent refinement of the structure of the type-defining compound, but no effort has been made to find or use the most recent determination.

All of the data sets with published coordinates in Pearson's Crystal Data have been classified into prototypes following the criteria defined in *TYPIX* (Parthé *et al.*, 1993, 1994). According to this definition, isotypic compounds must crystallize in the same space group and have similar cell-parameter ratios; the atoms should occupy the same Wyckoff positions in the standardized description and have similar positional coordinates. If all of these criteria are fulfilled, the atomic environments should be similar. Note that H^+ (protonic hydrogen) is ignored in the assignment of the prototype as well as in the Wyckoff sequence, Pearson symbol/code and atomic environments. Isopointal substitution variants are usually distinguished; however, no distinction is made between structures with fully and partly occupied atom sites. At present, 29 470 prototypes are represented.

When possible, a prototype has also been assigned to data sets without published atom coordinates. The prototype is often stated in the publication; in other cases the editors have assigned it. The editor will have added the exact space-group setting to which the cell parameters refer when this was not published. It is important to note that a prototype has been assigned at two different levels. The first is intimately related to the published data (entry level), whereas the second is assigned at the phase level and may, in some cases, be inconsistent with the crystallographic data listed below.

For partly investigated structures, the available structural information is given using a similar way, for example the complete Pearson symbol may be replaced by *t*** (tetragonal) or *cI** (cubic body-centred) and the place of the type-defining compound is occupied by an asterisk.

3.7.5.5. *Assigned atom coordinates*

In order to give an approximate idea of the actual structure, a complete set of positional coordinates and site occupancies is proposed for data sets where a prototype could be assigned but the atom coordinates were not determined. The coordinates of the type-defining entry are proposed as a first approximation. The atom distribution is inserted by an *ESDD* module that compares the chemical formula of the type-defining entry with the chemical formula of the isotypic compound where the chemical elements have been reordered by the editor so that the first element is expected to occupy the same atom sites as the first element in the type-defining formula, and so on. Depending on the character of the prototype, substitutions and/or vacancies are either distributed over all atom sites occupied by the corresponding element or are expected to occur selectively on particular atom sites.

For this category of database entries, structure drawings, diffraction patterns and interatomic distances have also been computed. The structural portion of the database is thus more extensive than the primary literature.

3.7.5.6. *External links*

When relevant, the database entries contain links to external data sources, including ASM International Alloys Phase Diagrams Centre Online, SpringerMaterials (The Landolt–Börnstein Database incorporating Inorganic Solid Phases PAULING FILE Multinaries Edition – 2010 in SpringerMaterials) and the original publication (through https://www.crossref.org/). A (static) reference to the Powder Diffraction File entry number is provided for database entries that are included in the PDF4+ product.

3.7.5.7. *Retrievable database fields*

In addition to bibliographic (*e.g.* a particular institute) and chemical (*e.g.* sulfates) searches, many characteristics of the experiment and data processing (*e.g.* single crystal, neutron diffraction, range of temperature or reliability factors) or additional studies (*e.g.* pressure-dependence studies, magnetic structure) can be used as search criteria. Published crystal data, standardized crystal data and the Niggli reduced cells can be searched, as well as crystallographic classifications such as crystal class, Pearson symbol, Pearson code, Wyckoff sequence, structure prototype or structure class. Such searches can be very valuable in identifying a structural model for a new composition and saving the work of an *ab initio* structure determination.

The Quick Search pane includes commonly used searches on chemical elements (including cations in a particular oxidation state for oxides and halides), the number of elements and functional groups. The chemical selection (and/or/not) can be combined with selection on structure prototypes, space-group numbers and symbols or the crystal system. Retrieval on cell parameters (with ranges) and bibliographic information is also possible and the desired level of structural studies (*e.g.* complete) may be specified.

Many more searches can be carried out in the complete Search Dialog. Particularly useful are searches on atomic environments and interatomic distances. The atomic environment is defined by the coordination number, the geometry of the coordination polyhedron and the identities of the central and peripheral atoms. Searches on the number of different atomic environment types in the same structure can also be carried out. Specifying a pair of elements makes it possible to select a range of interatomic distances to be included in the search. These histograms are also useful in assessing the reasonableness of a particular distance in a Rietveld refinement.

3.7.5.8. *Particular software features*

All searches use the 'Perpetual Restraining' feature, which updates the selection set in real time as a new query is introduced, so that the progress of the search scheme can easily be monitored. The complete Search Dialog offers a large variety of features that make the retrieval and presentation of information extremely flexible.

The Chemical System Matrix View makes it easy to locate phases in binary, ternary, quaternary and pseudo-quaternary systems. The Phases List View collects a selection set into its 'distinct' phases. From the individual database entry it is particularly easy to find all database entries with the same prototype structure and plot the unit-cell volume as a function of selected atomic radii. The standard display of an entry includes a short summary about the phase, structural, bibliographic, experimental and editorial data, as well as a structure drawing, a powder pattern and a table of interatomic distances.

The software for producing structure drawings offers the visualization of atomic environments (coordination polyhedra), the statistics of interatomic distances and the calculation of selected distances and angles. Four different models are available (ball and stick, wires, sticks and space-filling), with on-the-fly rotation controlled by the mouse. The nearest-neighbour histogram of a selected atom is compared with a statistical plot containing all distances in the database involving the same chemical elements, and the atomic environments can be instantaneously modified by clicking on the nearest-neighbour histogram.

Powder-diffraction patterns can be computed for any user-defined wavelength and the visualization includes a tool for zoom-in/out tracking. Patterns based on published lists of interplanar spacings can also be visualized. It is further possible to export database entries as CIF files, tables (*e.g.* powder-diffraction pattern, distances and angles) or graphics (*e.g.* structure drawings; BMP, GIF, JPG, PNG, TIFF or Diamond documents), and individually tailored dossiers can be designed and printed.

3.7.6. Metals data file (CRYSTMET)

CRYSTMET (White *et al.*, 2002) began as a database of critically evaluated crystallographic data for metals, including alloys, intermetallics and minerals, and has grown to include inorganic compounds in general. It was started in 1960 by Cromer and Larson at Los Alamos National Laboratory, and its development was continued by the National Research Council of Canada. In 1996, the production and dissemination was transferred to Toth Information Systems.

CRYSTMET contains chemical, crystallographic and bibliographic data, together with comments regarding experimental details for each study. Using these data, a number of associated data files are generated, with the major one being a file of calculated powder patterns. Entry into CRYSTMET is *via* a number of search screens, including chemistry, bibliographic information, unit cell and reduced cell, powder patterns (using the positions of the strongest peaks as input), formula, structure type, Pearson symbol and space group. The results of queries reside in sets, which can be further manipulated using logical operations.

The results are displayed as a series of screens, which include crystallographic data, distances and angles, and the powder pattern. There is some ability to customize the calculation of the powder pattern of an entry; the calculation is performed for Debye–Scherrer geometry. Included on the Results tabs is a direct interface to the *MISSYM* program (Le Page, 1987, 1988), which searches the reported structure for additional symmetry elements. This is a very useful tool for detecting missed symmetry.

3.7.7. Protein Data Bank (PDB)

The Protein Data Bank is described in Chapter 24.1 of *International Tables for Crystallography* Volume F (Berman *et al.*, 2011). Current information is available on the web at https://www.wwpdb.org/.

3.7.7.1. *Powder diffraction by proteins*

Although powder-diffraction techniques had been applied to proteins as long ago as 1936 (Wyckoff & Corey, 1936; Corey & Wyckoff, 1936), and proof-of-principle experiments had been carried out (Rotella *et al.*, 1998, 2000), real progress in protein powder crystallography began with the work of Von Dreele (Von Dreele, 1998, 1999, 2003; Von Dreele *et al.*, 2000).

Progress in powder crystallography on macromolecules has been reviewed by Margiolaki & Wright (2008) and is also discussed in Chapter 7.1 of this volume. Notable studies include the characterization of the binding of *N*-acetylglucosamine oligosaccharides to hen egg-white lysozyme (Von Dreele, 2007*a*) and determination of the second SH3 domain of ponsin (Margiolaki *et al.*, 2007).

As with all powder diffraction, peak overlap ultimately limits the information available. Multi-pattern strategies to overcome the overlap problem have been investigated by Von Dreele (2007*b*). Multiple-pattern resonant-diffraction experiments have enabled study of the binding of $PtBr_6^{2-}$ ions to lysozyme (Helliwell *et al.*, 2010). A bootstrap approach has been used to determine the structure of bacteriorhodopsin to 7 Å resolution (Dilanian *et al.*, 2011). Parametric resonant-scattering experiments have been used to determine the secondary structures of lysozyme derivatives (Basso *et al.*, 2010). Powder-diffraction experiments have also been used to gain insight into the general features of a nonstructural protein 3 (nsp3) macro domain (Papageorgiou *et al.*, 2010).

The structure of a five-residue peptide has been determined *ab initio* using laboratory powder data (Fujii *et al.*, 2011). We can expect further useful results at this interface between small-molecule and protein powder crystallography.

As is typical in other areas of science, powder diffraction has proven to be useful in more practical features of protein processing. It has been used to identify insulin (Norrman *et al.*, 2006) and GB1 (Frericks Schmidt *et al.*, 2007) polymorphs and lot-to-lot variations in lyophilized protein formulations (Hirakura *et al.*, 2007), and has been explored for use in structure-based generic assays (Allaire *et al.*, 2009).

3.7.7.2. *Calculation of protein powder patterns (with Kenny Ståhl)*

The Powder Diffraction File contains a few experimental powder patterns of proteins. These include silk fibroin protein (00-054-1394), tubulin (00-036-1547 and 00-036-1548), insulin (00-060-1360 through 00-060-1368), tomato bushy stunt virus (00-003-0001) and tobacco mosaic virus (00-003-0003 and 00-003-0004). Patterns have not yet been calculated from the structures in the Protein Data Bank because the calculated intensities generally fit poorly to those in experimental patterns.

Protein structures in the PDB do not generally contain H-atom positions, and the contributions from the disordered solvent in the solvent channels (which is the major source of the discrepancy) is not described (Hartmann *et al.*, 2010). The conventional Lorentz factor tends to infinity when approaching $2\theta = 0°$. Differences in data-collection temperatures and solvent content between powder and single-crystal specimens often mean that the lattice parameters differ. The relatively poor scattering from the protein and the large scattering from the mother liquor and sample holder result in significant background contributions to experimental powder patterns.

Optimization of the lattice parameters is generally straightforward and is important because most protein crystal structures are determined at low temperatures, while powder data are collected under ambient conditions. Protein crystals contain 30–80% disordered solvent. The solvent contribution to the diffraction pattern is most important for the low-angle powder data. In conventional protein crystallography several correction models have been developed (Moews & Kretsinger, 1975; Phillips, 1980; Jiang & Brünger, 1994), but the flat bulk-solvent model is the simplest one which yields a realistic correction (Jiang & Brünger, 1994; Hartmann *et al.*, 2010). This model includes two parameters: $k_{\rm sol}$, which defines the level of electron density in the solvent region, and $B_{\rm sol}$, which defines the steepness of the border between the solvent and macromolecular regions. These parameters are typically refined in contemporary software and cluster around $k_{\rm sol}$ = 0.35 e Å^{-3} and $B_{\rm sol}$ = 46 Å^2 (Fokine & Urzhumtsev, 2002).

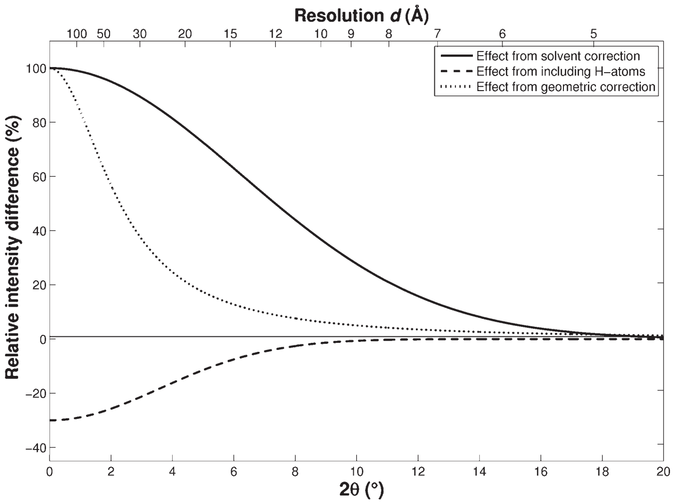


Figure 3.7.13
Overview of the trends from the different corrections. The effects are shown as the relative intensity difference $(I_{\rm non\text{-}corr} - I_{\rm corr})/I_{\rm non\text{-}corr}$ plotted as functions of the scattering angle 2θ (using Cu $K\alpha_1$) and resolution $d = \lambda/(2\sin\theta)$. The curves are based on average corrections of lysozyme and insulin data. $I_{\rm non\text{-}corr}$ is the raw intensity from a calculated pattern which has only been Lorentz corrected. The geometric correction curve was calculated using η = 0.045 Å^{-1}. From Hartmann *et al.* (2010).

The flat bulk-solvent correction can be applied using *phenix.pdbtools* (Adams *et al.*, 2010), which requires a PDB coordinate file and values of $k_{\rm sol}$ and $B_{\rm sol}$ as input. Average values can be used, but refined values or values from the Electron Density Server (EDS; Kleywegt *et al.*, 2004) can improve the results. The bulk-solvent correction is highly anisotropic, and both parameters affect the anisotropy.

The ideal H-atom positions can be calculated using *phenix.pdbtools*. The solvent and hydrogen contributions to the pattern can be significant (Fig. 3.7.13).

The Lorentz factor L describes the fraction of a reflection that is in the diffracting condition. For Bragg–Brentano and Debye–Scherrer geometries it is given by

$$L = \frac{1}{\sin 2\theta}\frac{1}{\sin\theta}. \tag{3.7.3}$$

This equation assumes ideal crystals, resulting in infinitesimally small reciprocal-lattice points. The true size of the lattice points depends on the crystallite size and imperfections (strain). This smearing needs to be included in the Lorentz factor at low angles. A revised Lorentz factor for protein powder diffraction has been derived (Hartmann *et al.*, 2010),

$$L_{\rm rev} = \frac{1}{\sin 2\theta}\frac{1}{\sin\theta}\frac{\sin^2\theta}{(\sin^2\theta + \lambda^2\eta^2/12)}, \tag{3.7.4}$$

in which η reflects the distribution of scattering-vector amplitudes. For Guinier geometry these equations become more complex (Hartmann *et al.*, 2010). Fig. 3.7.14 shows that the Lorentz factor has a smaller effect than the solvent and H atoms, but that it is still significant. By applying these corrections it should be possible for the ICDD editorial staff to calculate useful powder patterns from PDB entries that could be included in the Powder Diffraction File.

Separating the background from the diffraction pattern is not straightforward (Frankaer *et al.*, 2011). Estimation of the background is greatly assisted by a correct calculated pattern. The calculated pattern can be scaled to the experimental data using *PROTPOW* (http://www.kemi.dtu.dk/english/Research/PhysicalChemistry/Protein_og_roentgenkrystallografi/Protpow).

Ståhl *et al.* (2013) have demonstrated that existing search/match procedures can be used to identify proteins using their powder patterns, and that powder patterns calculated from Protein Data Bank coordinates with proper care can be added to a database and included in the search/match procedure. Several problems can be foreseen when including large amounts of protein data into the Powder Diffraction File. It may be worthwhile including powder patterns with several levels of solvent correction, rather than just an average value. Asymmetry from instrumental effects and specimen transparency, which can affect the peak positions, needs to be taken into account. The use of an average thermal expansion coefficient may be sufficient to account for the differences in lattice parameters between low-temperature single-crystal structures and powder patterns measured under ambient conditions.

3.7.8. Crystallography Open Database (COD) (with Saulius Gražulis)

The Crystallography Open Database (COD) project (http://www.crystallography.net/cod/; Gražulis *et al.*, 2009, 2012) aims at collecting in a single open-access database all organic, inorganic and organometallic structures, except for the structures of biological macromolecules, which are available in the Protein Data Bank (Berman *et al.*, 2003, 2011). The database was founded by Armel Le Bail, Lachlan Cranswick, Michael Berndt, Luca Lutterotti and Robert M. Downs in February 2003 as a response to Michael Berndt's letter published on the Structure Determination by Powder Diffractometry (SDPD) mailing list (Berndt, 2003). Since December 2007, the main database server has been maintained and new software has been developed by Saulius Gražulis and Andrius Merkys at the Institute of Biotechnology of Vilnius University (VU). Currently, the database includes more than 376 000 entries describing structures of small molecules and small-to-medium-sized unit-cell materials as published in IUCr journals and other major crystallographic and peer-reviewed journals, as well as contributions by crystallographers from major laboratories. Most of the mineral data are obtained from the American Mineralogist Structure Database (Rajan *et al.*, 2006) and are donated by its maintainer and COD co-founder Robert M. Downs.

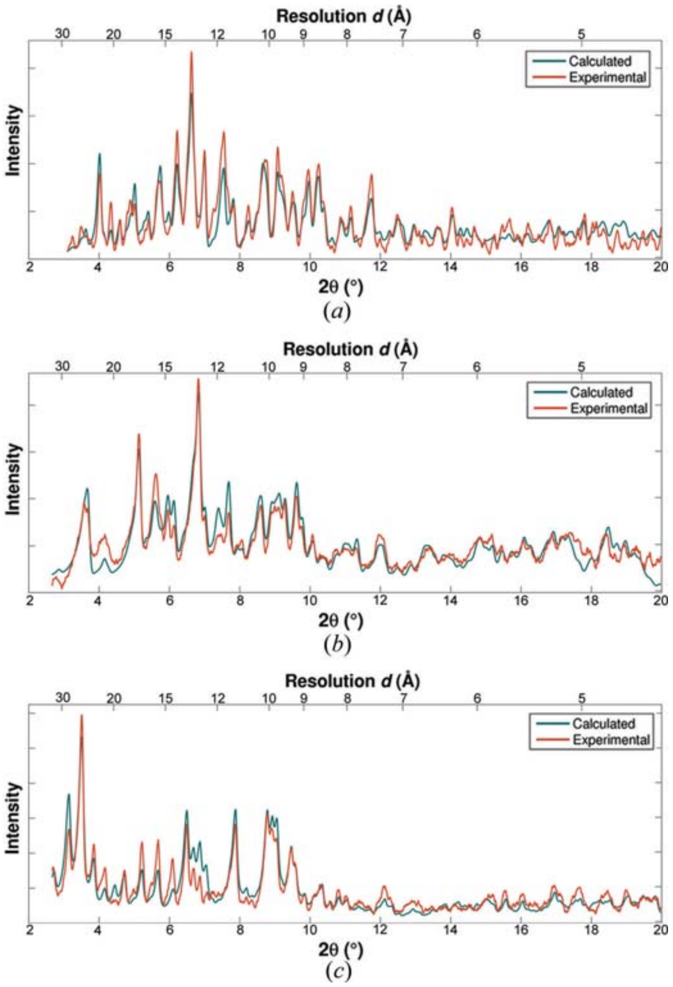
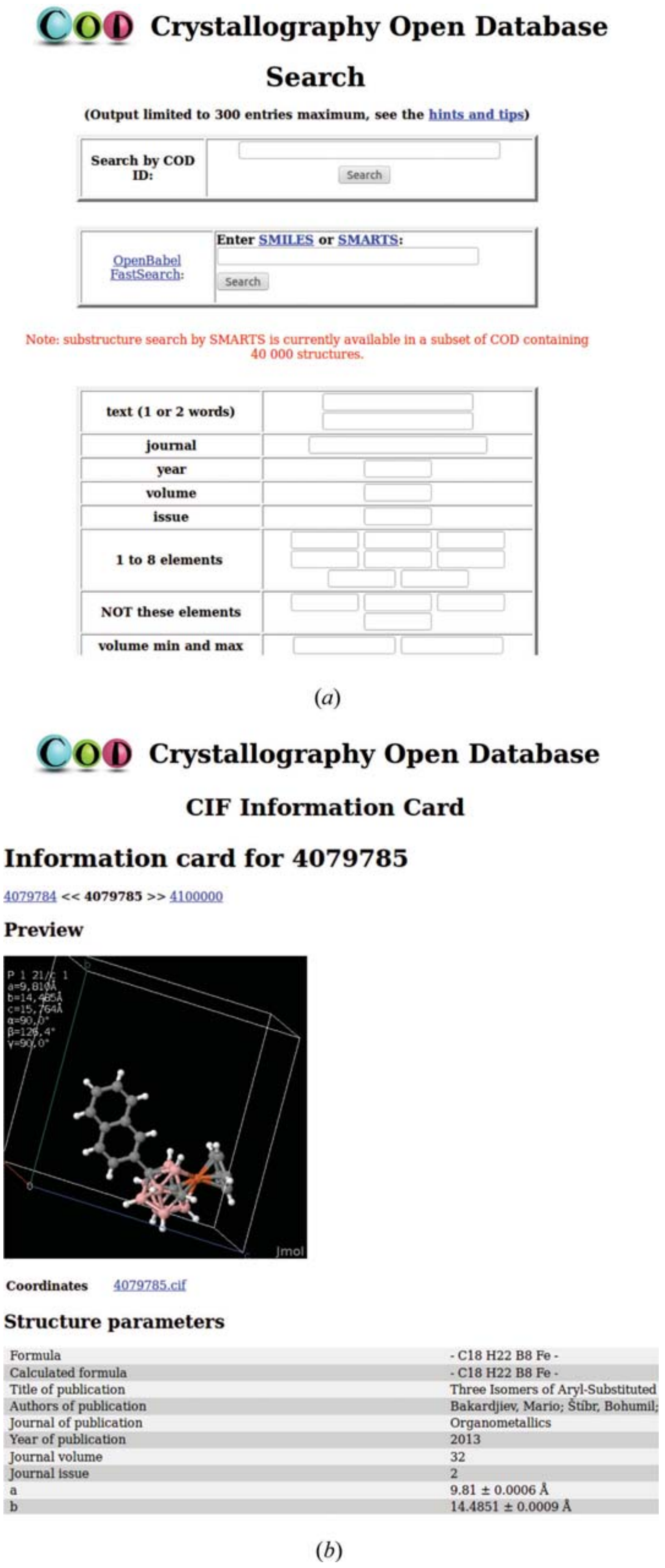


Figure 3.7.14
Calculated and experimental powder patterns for (*a*) lysozyme, (*b*) trigonal insulin and (*c*) cubic insulin. The calculated patterns (blue) are corrected for bulk-solvent and geometrical effects using the revised Lorentz factor. From Hartmann *et al.* (2010).

The database is an internet resource (Fig. 3.7.15) with data-search and download capabilities designed by Armel Le Bail and Michael Berndt. In addition, registered users may deposit new data, whether from previous publications or as personal communications, using the deposition web site designed at VU by



COD Crystallography Open Database

Search

(Output limited to 300 entries maximum, see the hints and tips)

Search by COD ID:	Search
OpenBabel FastSearch:	Enter SMILES or SMARTS: Search

Note: substructure search by SMARTS is currently available in a subset of COD containing 40 000 structures.

text (1 or 2 words)	
journal	
year	
volume	
issue	
1 to 8 elements	
NOT these elements	
volume min and max	

(*a*)

COD Crystallography Open Database

CIF Information Card

Information card for 4079785

4079784 << 4079785 >> 4100000

Preview

Coordinates 4079785.cif

Structure parameters

Formula	- C18 H22 B8 Fe -
Calculated formula	- C18 H22 B8 Fe -
Title of publication	Three Isomers of Aryl-Substituted
Authors of publication	Bakardjiev, Mario; Štíbr, Bohumil;
Journal of publication	Organometallics
Year of publication	2013
Journal volume	32
Journal issue	2
a	9.81 ± 0.0006 Å
b	14.4851 ± 0.0009 Å

(*b*)

Figure 3.7.15
(*a*) The website and search interface of the Crystallography Open Database (COD) permits searches of crystallographic data by a range of parameters and unrestricted retrieval of the found data. (*b*) Data can be viewed online in the interactive *Jmol* applet (Hanson, 2010, 2013) or downloaded for further processing either one record at a time or in bulk.

Saulius Gražulis, Justas Butkus and Andrius Merkys. The deposition software performs rigorous checks of syntax and semantics.

The COD website allows searching on COD numerical identifier, unit-cell parameters, chemical composition and bibliographic data. Substructure searches using SMILES and SMARTS strings have been implemented. The free software package *OpenBabel* (O'Boyle *et al.*, 2011; Hutchison, 2007) is used for both the CIF-to-SMILES transformation and the actual search.

The retrieved records can be viewed online or downloaded for further processing. For massive data mining, COD permits downloads and updates of the whole database using Subversion, Rsync or http protocols. The ease of access to the COD data and its open nature has spurred the use of this resource for software testing (Grosse-Kunstleve & Gildea, 2011), teaching (Moeck, 2004) and research (First & Floudas, 2013). Multiple mirrors around the globe (Quirós-Olozábal, 2006; Gražulis, 2007; Moeck, 2007*a*; Chateigner, 2010) ensure data preservation, provide off-site backups, offer improved search interfaces (Moeck, 2007*b*) and increase reliability.

For the powder-diffraction community, the COD is interesting not only as an archive of structures solved by powder-diffraction methods, but also as a possibility for use in search/match procedures to identify crystalline compounds. Recently, the development of an open full-pattern search/match internet tool was launched by the COD developers. It allows phase quantifications from X-ray, neutron and electron powder patterns (with high- or medium-resolution instruments) provided that the structures are already in the COD. This tool is particularly suited to nanocrystalline powders, in which severe line broadening appears, precluding phase identification from only peak positions (Lutterotti *et al.*, 2012). COD-derived databases are also offered for software produced by several diffractometer vendors (Rigaku, 2011; PANalytical, 2012*a*,*b*; Bruker, 2013). In addition to the COD, searches and matches can be performed against its sister database, the PCOD, which contains structures predicted by the *GRINSP* program (Le Bail, 2005) and hypothetical zeolites (Pophale *et al.*, 2013). The power of such an approach is demonstrated by PCOD entry 3102887 (formulated as SiO_2). It was recently identified as corresponding structurally to a new phosphorus(V) oxonitride polymorph δ-PON (Baumann *et al.*, 2012).

3.7.9. Other internet databases

Other useful databases include the following:

(i) The American Mineralogist Crystal Structure Database (http://rruff.geo.arizona.edu/AMSamcsd.php).
(ii) The Mineralogy Database (http://webmineral.com).
(iii) MinCryst (http://database.iem.ac.ru/mincryst/index.php).
(iv) The International Zeolite Association Database of Zeolite Structures (http://www.iza-structure.org/databases).
(v) The Incommensurate Structures Database (http://webbdcrista1.ehu.es/incstrdb/).
(vi) The Nucleic Acid Database (http://ndbserver.rutgers.edu).

I thank Timothy G. Fawcett and Cyrus Crowder of the International Centre for Diffraction Data, Colin Groom of CCDC, Stephan Ruehl of FIZ Karlsruhe, Pierre Villars and Karen Cenzual of Material Phases Data Systems, and Saulius Gražulis of Vilnius University Institute of Biotechnology for their valuable comments and additions.

References

Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L.-W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C. & Zwart, P. H. (2010). *PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Cryst.* D**66**, 213–221.

Allaire, M., Moiseeva, N., Botez, C. E., Engel, M. A. & Stephens, P. W. (2009). *On the possibility of using polycrystalline material in the development of structure-based generic assays. Acta Cryst.* D**65**, 379–382.

Allen, F. H., Cole, J. C. & Verdonk, M. L. (2011). *The relevance of the Cambridge Structural Database in protein crystallography. International Tables for Crystallography*, Vol. F, 2nd ed., edited by E. Arnold, D. M. Himmel & M. G. Rossmann, pp. 736–748. Chichester: Wiley.

Allmann, R. & Hinek, R. (2007). *The introduction of structure types into the Inorganic Crystal Structure Database ICSD. Acta Cryst.* A**63**, 412–417.

Barr, G., Dong, W. & Gilmore, C. J. (2009). *PolySNAP3: a computer program for analysing and visualizing high-throughput data from diffraction and spectroscopic sources. J. Appl. Cryst.* **42**, 965–974.

Barr, G., Gilmore, C. J. & Paisley, J. (2004). *SNAP-1D: a computer program for qualitative and quantitative powder diffraction pattern analysis using the full pattern profile. J. Appl. Cryst.* **37**, 665–668.

Basso, S., Besnard, C., Wright, J. P., Margiolaki, I., Fitch, A., Pattison, P. & Schiltz, M. (2010). *Features of the secondary structure of a protein molecule from powder diffraction data. Acta Cryst.* D**66**, 756–761.

Baumann, D., Sedlmaier, S. J. & Schnick, W. (2012). *An unprecedented AB_2 tetrahedra network structure type in a high-pressure phase of phosphorus oxonitride (PON). Angew. Chem. Int. Ed.* **51**, 4707–4709.

Behrens, H. & Luksch, P. (2006). *A bibliometric study in crystallography. Acta Cryst.* B**62**, 993–1001.

Belsky, A., Hellenbrandt, M., Karen, V. L. & Luksch, P. (2002). *New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design. Acta Cryst.* B**58**, 364–369.

Bergerhoff, G. & Brandenburg, K. (1999). *Typical interatomic distances: inorganic compounds. International Tables for Crystallography*, Vol. C, edited by E. Prince, pp. 770–781. Dordrecht: Kluwer Academic Publishers.

Bergerhoff, G. & Brown, I. D. (1987). *Crystallographic Databases*, edited by F. H. Allen, G. Bergerhoff & R. Sievers. Chester: International Union of Crystallography.

Berman, H. M., Henrick, K., Kleywegt, G., Nakamura, H. & Markley, J. (2011). *The Worldwide Protein Data Bank. International Tables for Crystallography*, Vol. F, 2nd ed., edited by E. Arnold, D. M. Himmel & M. G. Rossmann, pp. 827–832. Chichester: Wiley.

Berman, H., Henrick, K. & Nakamura, H. (2003). *Announcing the Worldwide Protein Data Bank. Nature Struct. Mol. Biol.* **10**, 980.

Berndt, M. (1994). Thesis, University of Bonn, Germany. Updates by O. Shcherban, SCC Structure-Properties Ltd, Lviv, Ukraine.

Berndt, M. (2003). *Open crystallographic database – a role for whom?* http://www.cristal.org/SDPD-list/2003/msg00025.html.

Bigelow, W. C. & Smith, J. V. (1964). *Two new indexes to the Powder Diffraction File. ASTM Spec. Tech. Publ.* STP372, 54. https://doi.org/10.1520/STP48334S.

Boldyrev, A. K., Mikheev, V. I., Dubinina, V. N. & Dovalev, G. A. (1938). *X-ray determination tables for minerals, Ft. I. Ann. Inst. Mines Leningrad*, **11**, 1–157.

Boles, M. O., Girven, R. J. & Gane, P. A. C. (1978). *The structure of amoxycillin trihydrate and a comparison with the structures of ampicillin. Acta Cryst.* B**34**, 461–466.

Bravais, A. (1866). *Etudes Cristallographiques*. Paris: Gathier Villars.

Bruker-AXS (2013). Crystallography Open Database for DIFFRAC.EVA. https://www.bruker.com/products/x-ray-diffraction-and-elemental-analysis/x-ray-diffraction/xrd-software/eva/cod.html.

Bruno, I. J., Cole, J. C., Edgington, P. R., Kessler, M., Macrae, C. F., McCabe, P., Pearson, J. & Taylor, R. (2002). *New software for searching the Cambridge Structural Database and visualizing crystal structures. Acta Cryst.* B**58**, 389–397.

Bruno, I. J., Cole, J. C., Kessler, M., Luo, J., Motherwell, W. D. S., Purkis, L. H., Smith, B. R., Taylor, R., Cooper, R. I., Harris, S. E. & Orpen, A. G. (2004). *J. Chem. Inf. Comput. Sci.* **44**, 2133–2144.

Caussin, P., Nusinovici, J. & Beard, D. W. (1988). *Using digitized X-ray powder diffraction scans as input for a new PC-At search/match program. Adv. X-ray Anal.* **31**, 423–430.

Caussin, P., Nusinovici, J. & Beard, D. W. (1989). *Specific data handling techniques and new enhancements in a search-match program. Adv. X-ray Anal.* **32**, 531–538.

Cenzual, K., Berndt, M., Brandenburg, K., Luong, V., Flack, E. & Villars, P. (2000). *ESDD* software package. Updates by O. Shcherban, SCC Structure-Properties Ltd, Lviv, Ukraine.

Chateigner, D. (2010). *Crystallography Open Database Mirror at ENSICAEN*. http://cod.ensicaen.fr.

Cherukuri, S. C., Snyder, R. L. & Beard, D. W. (1983). *Comparison of the Hanawalt and Johnson–Vand computer search/match strategies. Adv. X-ray Anal.* **26**, 99–104.

Chipera, S. J. & Bish, D. L. (2002). *FULLPAT: a full-pattern quantitative analysis program for X-ray powder diffraction using measured and calculated patterns. J. Appl. Cryst.* **35**, 744–749.

Corey, R. B. & Wyckoff, R. W. G. (1936). *Long spacings in macromolecular solids. J. Biol. Chem*. **114**, 407–414.

Crystal Impact (2012). *Match!* v.2. Crystal Impact, Bonn, Germany.

Davey, W. P. (1922). *A new X-ray diffraction apparatus. Gen. Elec. Rev.* **25**, 565.

Davey, W. P. (1934). *Study of Crystal Structure and Its Applications*. New York: McGraw-Hill.

Debye, P. & Scherrer, P. (1916). *Interference on inordinate orientated particles in Roentgen light. Phys. Z.* **17**, 277–283.

Debye, P. & Scherrer, P. (1917). *Interference on inordinate orientated particles in X-ray light. III. Phys. Z.* **18**, 291–301.

Degen, T., Sadki, M., Bron, E., König, U. & Nénert, G. (2014). *The HighScore suite. Powder Diffr.* **29**, S13–S18.

Dilanian, R. A., Darmanin, C., Varghese, J. N., Wilkins, S. W., Oka, T., Yagi, N., Quiney, H. M. & Nugent, K. A. (2011). *A new approach for structure analysis of two-dimensional membrane protein crystals using X-ray powder diffraction data. Protein Sci.* **20**, 457–464.

Donnay, J. D. H. & Harker, D. (1937). *A new law of crystal morphology extending the law of Bravais. Am. Mineral.* **22**, 463–467.

Fawcett, T. G., Kabekkodu, S. N., Blanton, J. R. & Blanton, T. N. (2017). *Chemical analysis by diffraction: the Powder Diffraction File. Powder Diffr.* **32**, 63–71.

First, E. L. & Floudas, C. A. (2013). *MOFomics: computational pore characterization of metal–organic frameworks. Microporous Mesoporous Mater.* **165**, 32–39.

Fokine, A. & Urzhumtsev, A. (2002). *Flat bulk-solvent model: obtaining optimal parameters. Acta Cryst.* D**58**, 1387–1392.

Frankaer, C. G., Harris, P. & Ståhl, K. (2011). *A sample holder for in-house X-ray powder diffraction studies of protein powders. J. Appl. Cryst.* **44**, 1288–1290.

Frericks Schmidt, H. L., Sperling, L. J., Gao, Y. G., Wylie, B. J., Boettcher, J. M., Wilson, S. R. & Rienstra, C. M. (2007). *Crystal polymorphism of protein GB1 examined by solid-state NMR spectroscopy and X-ray diffraction. J. Phys. Chem. B*, **111**, 14362–14369.

Frevel, L. K. (1965). *Computational aids for identifying crystalline phases by powder diffraction. Anal. Chem.* **37**, 471–482.

Frevel, L. K., Adams, C. E. & Ruhberg, L. R. (1976). *A fast search-match program for powder diffraction analysis. J. Appl. Cryst.* **9**, 199–204.

Friedel, G. (1907). *Etudes sur la loi de Bravais. Bull. Soc. Fr. Miner.* **30**, 326–455.

Fujii, K., Young, M. T. & Harris, K. D. M. (2011). *Exploiting powder X-ray diffraction for direct structure determination in structural biology: the P2X4 receptor trafficking motif YEQGL. J. Struct. Biol.* **174**, 461–467.

Gelato, L. M. & Parthé, E. (1987). *STRUCTURE TIDY – a computer program to standardize crystal structure data. J. Appl. Cryst.* **20**, 139–143.

Gilmore, C. J., Barr, G. & Paisley, J. (2004). *High-throughput powder diffraction. I. A new approach to qualitative and quantitative powder diffraction pattern analysis using full pattern profiles. J. Appl. Cryst.* **37**, 231–242.

Goehner, R. P. & Garbauskas, M. F. (1984). *PDIDENT – a set of programs for powder diffraction phase identification. X-ray Spectrom.* **13**, 172–179.

Gražulis, S. (2007). *COD Mirror in Vilnius*. http://cod.ibt.lt/cod/.

Gražulis, S., Chateigner, D., Downs, R. T., Yokochi, A. F. T., Quirós, M., Lutterotti, L., Manakova, E., Butkus, J., Moeck, P. & Le Bail, A. (2009). *Crystallography Open Database – an open-access collection of crystal structures. J. Appl. Cryst.* **42**, 726–729.

Gražulis, S., Daškevič, A., Merkys, A., Chateigner, D., Lutterotti, L., Quirós, M., Serebryanaya, N. R., Moeck, P., Downs, R. T. & Le Bail, A. (2012). *Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. Nucleic Acids Res.* **40**, D420–D427.

Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. (2016). *The Cambridge Structural Database. Acta Cryst.* B**72**, 171–179.

Grosse-Kunstleve, R. & Gildea, R. (2011). *Computational Crystallography Initiative: COD stats.* http://cci.lbl.gov/cod_stats.

Hanawalt, J. D. (1983). *History of the Powder Diffraction File (PDF). Crystallography in North America*, edited by D. McLachlan & J. P. Glusker, pp. 215–219. Buffalo: American Crystallographic Association.

Hanawalt, J. D. (1986). *Manual search/match methods for powder diffraction in 1986. Powder Diffr.* **1**, 7–13.

Hanawalt, J. D. & Rinn, H. W. (1936). *Identification of crystalline materials. Ind. Eng. Chem. Anal. Ed.* **8**, 244–247.

Hanawalt, J. D., Rinn, H. W. & Frevel, L. K. (1938). *Chemical analysis by X-ray diffraction. Ind. Eng. Chem. Anal. Ed.* **10**, 457–512.

Hanson, R. M. (2010). *Jmol – a paradigm shift in crystallographic visualization. J. Appl. Cryst.* **43**, 1250–1260.

Hanson, R. M. (2013). *Jmol: an open-source Java viewer for chemical structures in 3D.* http://www.jmol.org.

Harju, P. & Pasek, P. (1983). *Vanadium–hydrogen–phosphorus–oxygen catalytic material.* US Patent 4374756; PDF entry 00-047-0967.

Hartmann, C. G., Nielsen, O. F., Ståhl, K. & Harris, P. (2010). *In-house characterization of protein powder. J. Appl. Cryst.* **43**, 876–882.

Hellenbrandt, M. (2004). *The Inorganic Crystal Structure Database (ICSD) – present and future. Crystallogr. Rev.* **10**, 17–22.

Helliwell, J. R., Bell, A. M. T., Pryant, P., Fisher, S. J., Habash, G., Helliwell, M., Margiolaki, I., Kaenket, S., Watier, Y., Wright, J. P. & Yalamanchilli, S. (2010). *Time-dependent analysis of K_2PtBr_6 binding to lysozyme studied by protein powder and single crystal X-ray analysis. Z. Kristallogr.* **225**, 570–575.

Hirakura, Y., Yamaguchi, H., Mizuno, M., Miyanishi, H., Ueda, S. & Kitamura, S. (2007). *Detection of lot-to-lot variations in the amorphous microstructure of lyophilized protein formulations. Int. J. Pharm.* **340**, 34–41.

Huang, T. C. & Parrish, W. (1982). *A new computer algorithm for qualitative X-ray powder diffraction analysis. Adv. X-ray Anal.* **25**, 213–219.

Hull, A. W. (1919). *A new method of chemical analysis. J. Am. Chem. Soc.* **41**, 1168–1175.

Hull, A. W. (1983). *An account of early studies at Schenectady. Crystallography in North America*, edited by D. McLachlan & J. P. Glusker, p. 32. Buffalo: American Crystallographic Association.

Hutchison, G. R. (2007). *OpenBabel: The Open Source Chemistry Toolbox.* http://openbabel.org.

ICDD (2016). *PDF-4+ 2016 (Database)*. Newtown Square: International Centre for Diffraction Data. http://www.icdd.com.

Jenkins, R. & Rose, R. N. (1990). *Don Hanawalt – early days and his contribution to qualitative powder diffractometry. Powder Diffr.* **5**, 70–75.

Jiang, J.-S. & Brünger, A. T. (1994). *Protein hydration observed by X-ray diffraction. Solvation properties of penicillopepsin and neuraminidase crystal structures. J. Mol. Biol.* **243**, 100–115.

Jobst, B. A. & Goebel, H. E. (1982). *IDENT – a versatile microfile-based system for fast interactive XRPD phase analysis. Adv. X-ray Anal.* **25**, 273–282.

Johnson, G. G. Jr & Vand, V. (1967). *A computerized powder diffraction identification system. Ind. Eng. Chem.* **59**, 19–31.

Johnson, G. G. Jr & Vand, V. (1968). *Computerized multiphase X-ray powder diffraction identification system. Adv. X-ray Anal.* **11**, 376–384.

Kaduk, J. A. (2002). *Use of the Inorganic Crystal Structure Database as a problem solving tool. Acta Cryst.* B**58**, 370–379.

Kleywegt, G. J., Harris, M. R., Zou, J., Taylor, T. C., Wählby, A. & Jones, T. A. (2004). *The Uppsala Electron-Density Server. Acta Cryst.* D**60**, 2240–2249.

Le Bail, A. (2005). *Inorganic structure prediction with GRINSP. J. Appl. Cryst.* **38**, 389–395.

Le Page, Y. (1987). *Computer derivation of the symmetry elements implied in a structure description. J. Appl. Cryst.* **20**, 264–269.

Le Page, Y. (1988). *MISSYM1.1 – a flexible new release. J. Appl. Cryst.* **21**, 983–984.

Lutterotti, L. (2012). *Qualitative Phase Analysis: Method Developments.* In *Uniting Electron Crystallography and Powder Diffraction*, pp. 233–242. Dordrecht: Springer.

Macrae, C. F., Bruno, I. J., Chisholm, J. A., Edgington, P. R., McCabe, P., Pidcock, E., Rodriguez-Monge, L., Taylor, R., van de Streek, J. & Wood, P. A. (2008). *Mercury CSD 2.0 – new features for the visualization and investigation of crystal structures. J. Appl. Cryst.* **41**, 466–470.

Margiolaki, I. & Wright, J. P. (2008). *Powder crystallography on macromolecules. Acta Cryst.* A**64**, 169–180.

Margiolaki, I., Wright, J. P., Wilmanns, M., Fitch, A. N. & Pinotsis, N. (2007). *Second SH3 domain of ponsin solved from powder diffraction. J. Am. Chem. Soc.* **129**, 11865–11871.

Marquart, R. G. (1986). *μPDSM: mainframe search/match on an IBM PC. Powder Diffr.* **1**, 34–39.

Marquart, R. G., Katsnelson, I., Milne, G. W. A., Heller, S. R., Johnson, G. G. & Jenkins, R. (1979). *A search-match system for X-ray powder diffraction data. J. Appl. Cryst.* **12**, 629–634.

Materials Data (2016). *Jade* 9.6. Livermore: Materials Data Inc. https://materialsdata.com/.

Mighell, A. D. (2003). *The normalized reduced form and cell mathematical tools for lattice analysis – symmetry and similarity. J. Res. Natl Inst. Stand. Technol.* **108**, 447–452.

Mighell, A. D. & Himes, V. L. (1986). *Compound identification and characterization using lattice-formula matching techniques. Acta Cryst.* A**42**, 101–105.

Moeck, P. (2004). *EDU-COD: Educational Subset of COD.* http://nanocrystallography.research.pdx.edu/search/edu.

Moeck, P. (2007*a*). *Crystallography Open Database Mirror in North America.* http://nanocrystallography.org.

Moeck, P. (2007*b*). *Alternative COD search interface at Portland State University.* http://nanocrystallography.research.pdx.edu/search/codmirror.

Moews, P. C. & Kretsinger, R. H. (1975). *Refinement of the structure of carp muscle calcium-binding parvalbumin by model building and difference Fourier analysis. J. Mol. Biol.* **91**, 201–225.

Nichols, M. (1966). *A Fortran Program for the Identification of X-ray Powder Diffraction Patterns*. Report UCRL-70078, Lawrence Livermore National Laboratory, USA.

Norrman, M., Ståhl, K., Schluckebier, G. & Al-Karadaghi, S. (2006). *Characterization of insulin microcrystals using powder diffraction and multivariate data analysis. J. Appl. Cryst.* **39**, 391–400.

Nusinovici, J. & Bertelmann, D. (1993). *Practical determination of the acceptable 2θ error for non-ambiguous identification of pure phases with DIFFRAC-AT. Adv. X-ray Anal.* **36**, 327–332.

Nusinovici, J. & Winter, M. J. (1994). *Diffrac-At search: search/match using full traces as input. Adv. X-ray Anal.* **37**, 59–66.

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T. & Hutchison, G. R. (2011). *Open Babel: an open chemical toolbox. J. Cheminf.* **3**, 33.

O'Connor, B. H. & Bagliani, F. (1976). *A semi-automated system for identifying crystalline materials with powder diffraction data. J. Appl. Cryst.* **9**, 419–423.

Oxford Cryosystems (2012). *Crystallographica Search-Match (CSM)*. Oxford: Oxford Cryosystems Ltd. http://www.oxcryo.com/software/.

PANalytical (2012*a*). *HighScore Plus*. Almelo: PANalytical B.V.

PANalytical (2012*b*). *The COD database files for the PANalytical HighScore or HighScore Plus software packages*. http://www.crystallography.net/archives/2012/PANalytical.

Papageorgiou, N., Watier, Y., Saunders, L., Coutard, B., Lantez, V., Gould, E. A., Fitch, A. N., Wright, J. P., Canard, B. & Margiolaki, I. (2010). *Preliminary insights into the non structural protein 3 macro domain of the Mayaro virus by powder diffraction. Z. Kristallogr.* **225**, 576–580.

Parrish, W. (1983). *History of the X-ray powder method in the USA*. *Crystallography in North America*, edited by D. McLachlan & J. P. Glusker, pp. 201–214. Buffalo: American Crystallographic Association.

Parthé, E., Cenzual, K. & Gladyshevskii, R. (1993). *Standardization of crystal structure data as an aid to the classification of crystal structure types. J. Alloys Compd*, **197**, 291–301.

Parthé, E. & Gelato, L. M. (1984). *The standardization of inorganic crystal-structure data. Acta Cryst.* A**40**, 169–183.

Parthé, E. & Gelato, L. M. (1985). *The 'best' unit cell for monoclinic structures consistent with b axis unique and cell choice 1 of International Tables for Crystallography (1983). Acta Cryst.* A**41**, 142–151.

Parthé, E., Gelato, L., Chabot, B., Penzo, M., Cenzual K. & Gladyshevskii, R. (1993, 1994). *TYPIX – Standardized Data and Crystal Chemical Characterization of Inorganic Structure Types.* Heidelberg: Springer. https://doi.org/10.1007/978-3-662-10641-9.

Phillips, S. E. V. (1980). *Structure and refinement of oxymyoglobin at 1.6 Å resolution. J. Mol. Biol.* **142**, 531–554.

Pophale, R., Daeyaert, F. & Deem, M. W. (2013). *Computational prediction of chemically-synthesizable organic structure directing agents for zeolites. J. Mater. Chem. A*, **1**, 6750–6760.

Quirós-Olozábal, M. (2006). *COD Mirror of Granada University.* http://qiserver.ugr.es/cod.

Rajan, H., Uchida, H., Bryan, D., Swaminathan, R., Downs, R. M. & Hall-Wallace, M. (2006). *Building the American Mineralogist Crystal Structure Database: A recipe for construction of a small internet database. Geoinformatics: Data to Knowledge*, edited by A. Sinha. McLean: Geological Society of America. https://doi.org/10.1130/2006.2397(06).

Rigaku (2011). *COD for PDXL: Integrated Powder X-ray Diffraction Software.* http://www.crystallography.net/archives/2011/Rigaku.

Rotella, F. J., Duke, N. & Kaduk, J. A. (1998). *X-ray powder diffraction from biological macromolecules. What do we see? What can we tell? American Crystallographic Association Meeting Abstracts, Washington DC, 18–23 July*. Buffalo: American Crystallographic Association.

Rotella, F. J., Duke, N. & Kaduk, J. A. (2000). *Powder diffraction from biological macromolecules using synchrotron X-rays. American Crystallographic Association Meeting Abstracts, St Paul MN, 22–27 July*. Buffalo: American Crystallographic Association.

Schreiner, W. N., Surdukowski, C. & Jenkins, R. (1982). *A new minicomputer search/match/identify program for qualitative phase analysis with the powder diffractometer. J. Appl. Cryst.* **15**, 513–523.

Shpeizer, B., Ouyang, X., Heising, J. M. & Clearfield, A. (2001). *Synthesis and crystal structure of a new vanadyl phosphate $[H_{0.6}(VO)_3(PO_4)_3(H_2O)_3].4H_2O$ and its conversion to porous products. Chem. Mater.* **13**, 2288–2296.

Sidey, V. (2013). *On the shortest $B^{III}-O$ bonds. Acta Cryst.* B**69**, 86–89.

Sietronics (2012). *Siroquant* v.4. Canberra: Sietronics. http://www.siroquant.com.

Snyder, R. L. (1981). *A Hanawalt type phase identification procedure for a minicomputer. Adv. X-ray Anal.* **24**, 83–90.

Ståhl, K., Frankaer, C. G., Petersen, J. & Harris, P. (2013). *Monitoring protein precipitates by in-house X-ray powder diffraction. Powder Diffr.* **28**, S448–S457.

Toby, B. H., Harlow, R. L. & Holomany, M. A. (1990). *The POWDER SUITE: computer programs for searching and accessing the JCPDS-ICDD powder diffraction database. Powder Diffr.* **5**, 2–7.

Villars, P. (1997). *Pearson's Desk Edition*, Vols. 1–2. Materials Park: ASM International.

Villars, P. & Calvert, L. D. (1985). *Pearson's Handbook of Crystallographic Data for Intermetallic Phases*, 1st Ed., Vols. 1–3. Materials Park: ASM International.

Villars, P. & Calvert, L. D. (1991). *Pearson's Handbook of Crystallographic Data for Intermetallic Phases*, 2nd Ed., Vols. 1–4. Materials Park: ASM International.

Villars, P. & Cenzual, K. (2013). *Pearson's Crystal Data: Crystal Structure Database for Inorganic Compounds* (CD-ROM). Materials Park: ASM International.

Villars, P., Onodera, N. & Iwata, S. (1998). *The Linus Pauling File (LPF) and its application to materials design. J. Alloys Compd*, **279**, 1–7.

Von Dreele, R. B. (1998). *Protein structures by powder diffraction? Abstr. Pap. Am. Chem. Soc.* **215**, U759.

Von Dreele, R. B. (1999). *Combined Rietveld and stereochemical restraint refinement of a protein crystal structure. J. Appl. Cryst.* **32**, 1084–1089.

Von Dreele, R. B. (2003). *Protein crystal structure analysis from high-resolution X-ray powder-diffraction data. Methods Enzymol.* **368**, 254–267.

Von Dreele, R. B. (2007*a*). *Binding of N-acetylglucosamine oligosaccharides to hen egg-white lysozyme: a powder diffraction study. Acta Cryst.* D**61**, 22–32.

Von Dreele, R. B. (2007*b*). *Multipattern Rietveld refinement of protein powder data: an approach to higher resolution. J. Appl. Cryst.* **40**, 133–143.

Von Dreele, R. B., Stephens, P. W., Smith, G. D. & Blessing, R. H. (2000). *The first protein crystal structure determined from high-resolution X-ray powder diffraction data: a variant of T_3R_3 human insulin–zinc complex produced by grinding. Acta Cryst.* D**56**, 1549–1553.

Waldo, A. W. (1935). *Identification of the copper ore minerals by means of X-ray powder diffraction patterns. Am. Mineral.* **20**, 575–597.

White, P. S., Rodgers, J. R. & Le Page, Y. (2002). *CRYSTMET: a database of the structures and powder patterns of metals and intermetallics. Acta Cryst.* B**58**, 343–348.

Winchell, A. N. (1927). *Further studies in the mica group. Am. Mineral.* **12**, 267–279.

Wolff, P. M. de (2016). *International Tables for Crystallography*, Vol. A, 6th ed., edited by M. I. Aroyo, pp. 709–714. Chichester: Wiley.

Wyckoff, R. W. G. & Corey, R. B. (1936). *X-ray diffraction patterns of crystalline tobacco mosaic proteins. J. Biol. Chem.* **116**, 51–55.