

3.7. CRYSTALLOGRAPHIC DATABASES

The FOM was calculated as

$$\text{FOM} = d_R I_R^2 d_U, \quad (3.7.1)$$

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_{ref}) matched and d_U is the percentage of the unknown peaks (with intensities I_{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_{ref} and I_{unk} was carried out. Peaks with $I_{\text{calc}} < I_{\text{obs}} - 5$ 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 $\text{PSI} = (1000/d) \times 10 + 5 \log_{10} I_3$. An inverted PDF was created, an index which contained the characteristic integers of the strongest lines of the reference patterns (PSI, PDF No. pairs) sorted by decreasing PSI. The input d/I list was compared with the index. All patterns that contained the characteristic integers within the bandpass were considered as potential hits. The full PDF was used to compare observed and reference patterns. A Davey minimum concentration (DMC) was calculated; this was the largest value in the range $0 \leq \text{DMC} \leq 1$ for which $I_{\text{PDF}} \text{DMC} \leq I_{\text{unk}}$ for all peaks. The PDF entry was then subtracted from the unknown pattern and the process was repeated. Initially, there were no chemistry or user filters; these appeared in later versions.

The Johnson–Vand figure of merit,

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

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

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

Sometimes, references to 'generations' of search/match programs will be encountered. The first-generation programs include those of Johnson & Vand (1967, 1968), Nichols (1966), Frevel *et al.* (1976), Marquart *et al.* (1979) and O'Connor & Bagliani (1976). The distinction between first- and second-generation programs (Snyder, 1981; Jobst & Goebel, 1982; Huang & Parrish, 1982; Schreiner *et al.*, 1982; Goehner & Garbaskas, 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; Nusinovič & Bertelmann, 1993; Nusinovič

& Winter, 1994), *HighScore* (Degen *et al.*, 2014), *Match!* (Crystal Impact, 2012), *Crystallographica Search-Match* (Oxford Cryo-systems, 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 nano-material 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

3. METHODOLOGY

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 $d \leq 2.500 \text{ \AA}$ is given to at least three decimal places. The *d*-spacings of reflections with $d \leq 1.2000 \text{ \AA}$ 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

3.7. CRYSTALLOGRAPHIC DATABASES

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 $d \leq 2.000 \text{ \AA}$ 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 $0.001 < U < 0.1 \text{ \AA}^2$. 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.

3. METHODOLOGY

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 $0.001 < U < 0.1 \text{ \AA}^2$ | |
| Structure corrected by the editor | $U_{\text{iso}} < 0.0$ | |
| 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 $0.3\text{--}0.5 \text{ \AA}$ 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_{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

3.7. CRYSTALLOGRAPHIC DATABASES

structure described by the CIF can then be imported into 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.

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 \AA . A search for mineral-related entries with the same chemistry and having one of their three strongest peaks in the range $3.309 (30) \text{ \AA}$ 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.

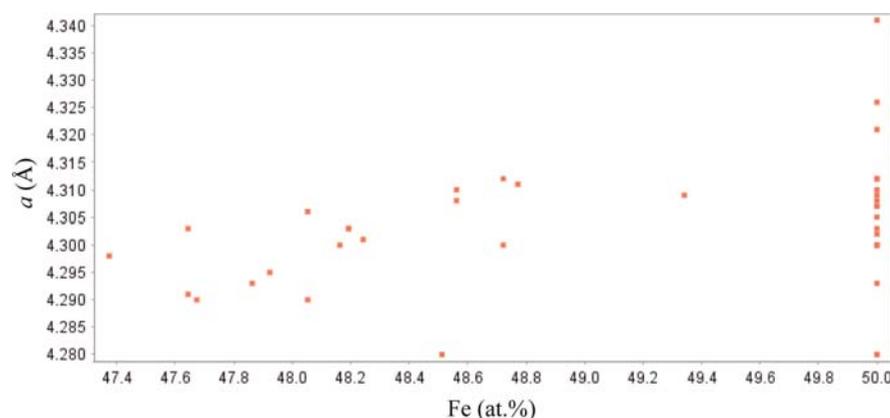


Figure 3.7.1

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.

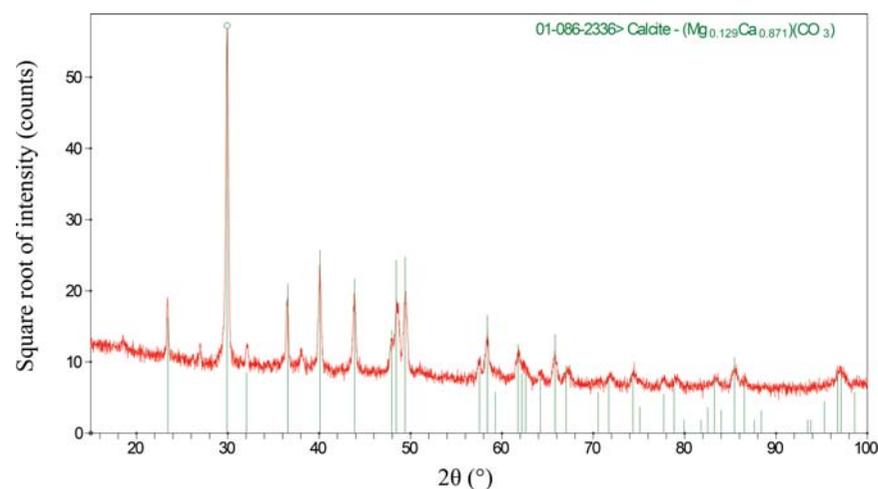


Figure 3.7.2

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. METHODOLOGY

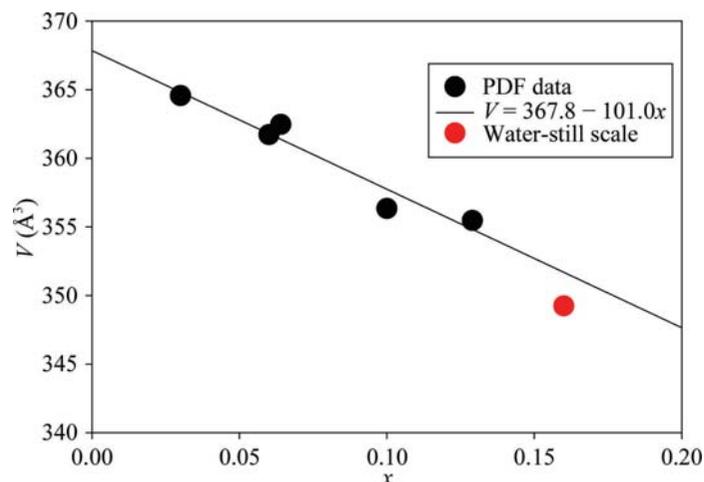


Figure 3.7.3
Variation of the unit-cell volume with the magnesium content in magnesian calcites in the Powder Diffraction File.

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 $\text{Ca}_{1-x}\text{Mg}_x(\text{CO}_3)$ (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 $[(\text{VO})_2\text{P}_2\text{O}_7]$ 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 $(\text{VO})_2\text{P}_2\text{O}_7$, 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: $\text{H}_4\text{V}_3\text{P}_3\text{O}_{16.5}(\text{H}_2\text{O})_2$. 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): $[\text{H}_{0.6}(\text{VO})_3(\text{PO}_4)_3(\text{H}_2\text{O})_3](\text{H}_2\text{O})_4$. 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 $[\text{H}_{0.6}(\text{VO})_3(\text{PO}_4)_3(\text{H}_2\text{O})_3](\text{H}_2\text{O})_4$ 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 -VOPO₄. 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)(PO₃)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% $(\text{VO})_2\text{P}_2\text{O}_7$, 5.9 (1) wt% $[\text{H}_{0.6}(\text{VO})_3(\text{PO}_4)_3(\text{H}_2\text{O})_3](\text{H}_2\text{O})_4$, 5.6 (1) wt% α -VOPO₄ and 3.7 (1) wt% β -VO(PO₃)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 PbBr₂. 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, Fe₃C, from the steel, and Fe₃Fe₄(PO₄)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% PbBr₂, 2.6 (2) wt% Fe₃Fe₄(PO₄)₆ and 2.2 (2) wt% cohenite (Fig. 3.7.8).

3.7. CRYSTALLOGRAPHIC DATABASES

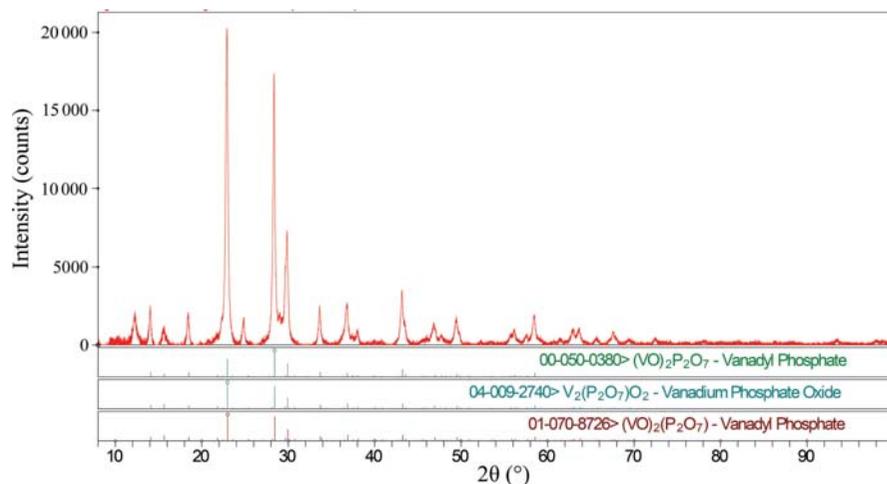


Figure 3.7.4

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.

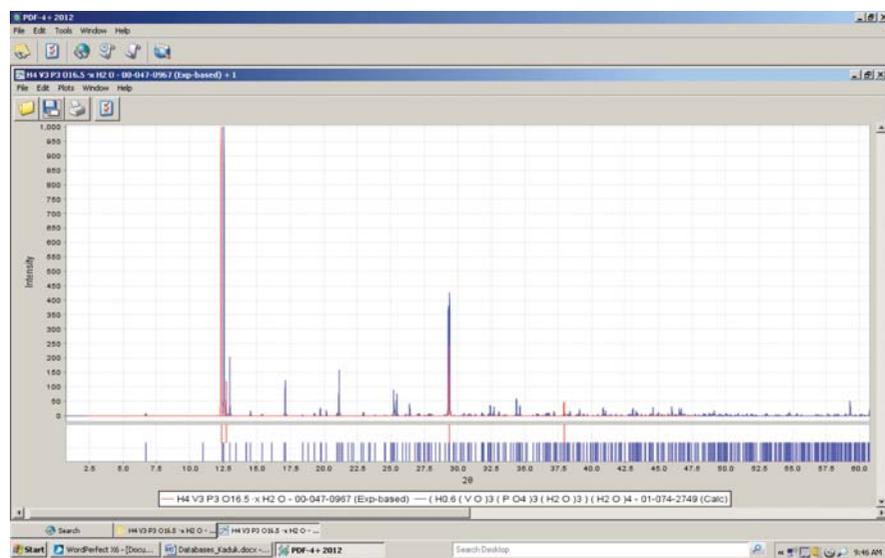


Figure 3.7.5

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.

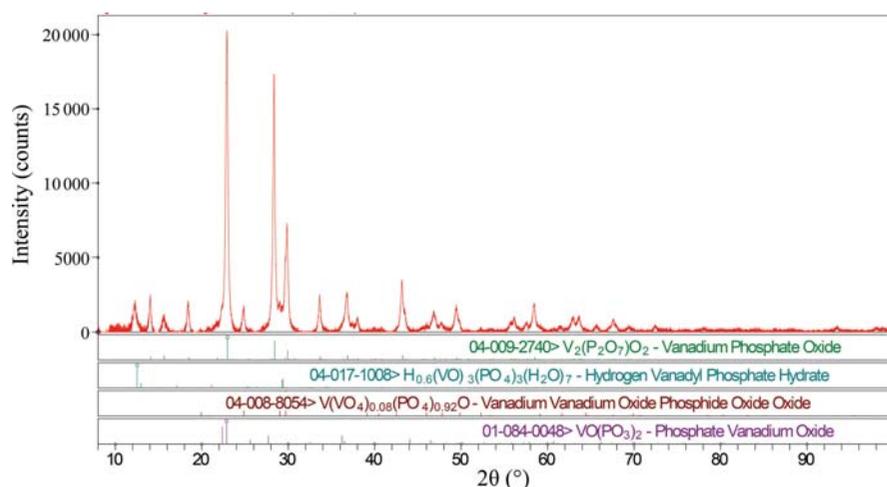


Figure 3.7.6

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% $(\text{NH}_4)\text{Fe}_2\text{S}_3$, 12.8 (4) wt% S_8 , 22.0 (6) wt% lepidocrocite ($\gamma\text{-FeOOH}$), 5.5 (5) wt% haematite ($\alpha\text{-Fe}_2\text{O}_3$), 6.6 (3) wt% pyrrhotite-4C (Fe_7S_8) and 6.6 (3) wt% dolomite [$\text{CaMg}(\text{CO}_3)_2$; limestone environmental dust]. The powder pattern and crystal structure of $(\text{NH}_4)\text{Fe}_2\text{S}_3$ 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 ($5.37^\circ 2\theta$). With such a low-angle peak, it seemed prudent to measure the pattern again

3. METHODOLOGY

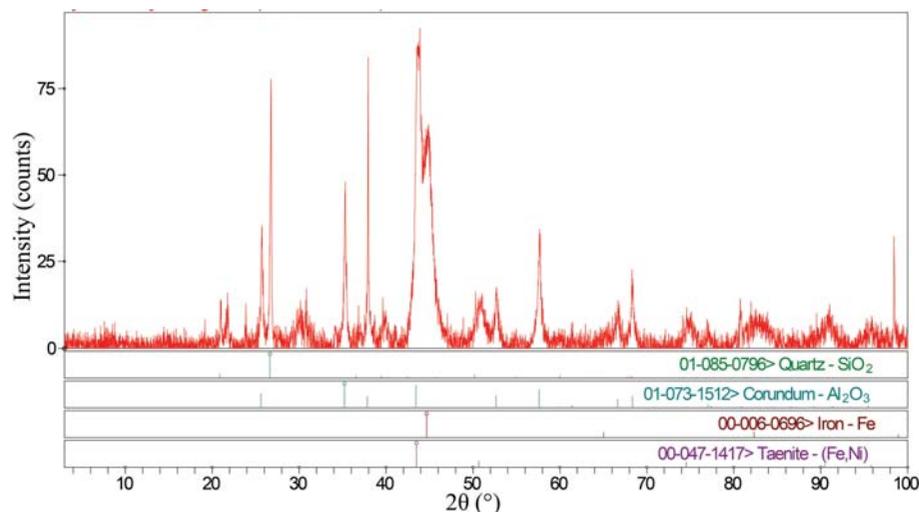


Figure 3.7.7

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.

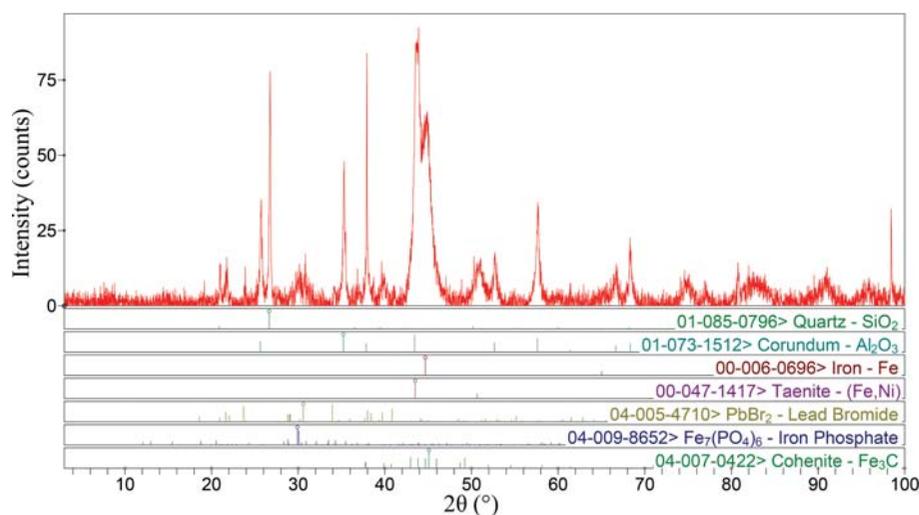


Figure 3.7.8

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

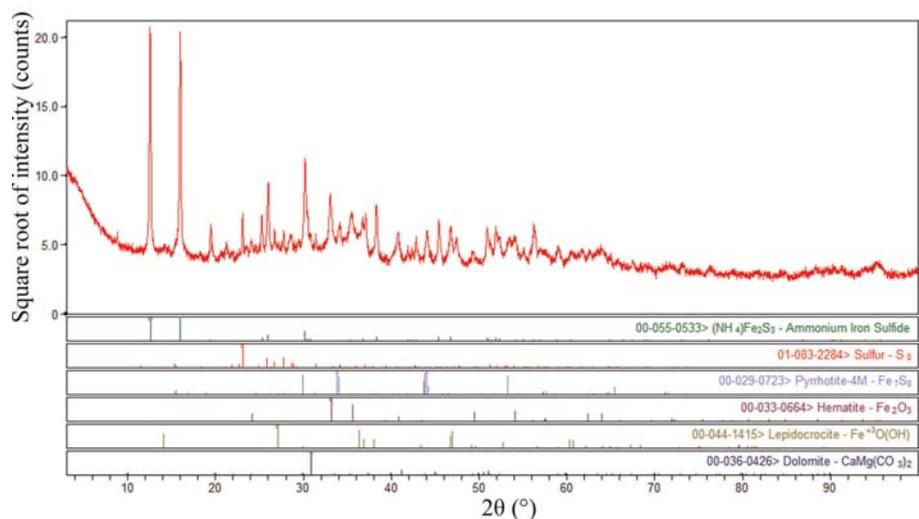


Figure 3.7.9

The phases identified in a deposit from a refinery isocracker. At the time, the $(\text{NH}_4)\text{Fe}_2\text{S}_3$ was a new phase, identified by analogy to KFe_2S_3 , rasvumite.

starting at 3° , and another peak was observed at $d = 24.80 \text{ \AA}$ ($3.56^\circ 2\theta$).

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 $1.5^\circ 2\theta$. 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 \AA , and other peaks occurred at 5.74 , 4.62 (strongest) and 4.407 \AA . A search in the PDF-4/Organics for compounds having the string 'ephed' in the name, a long line at $12.73 \pm 0.05 \text{ \AA}$ and a strong line at $4.62 \pm 0.02 \text{ \AA}$ 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 \AA (files centrum.gas and id320304.prm) easily identified brushite, $\text{CaHPO}_4(\text{H}_2\text{O})_2$, and sylvite, KCl (Fig. 3.7.12).

To identify additional phases, 64 peaks with $d > 1.91 \text{ \AA}$ 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 identifi-

3.7. CRYSTALLOGRAPHIC DATABASES

cation 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 [$\text{MnSO}_4(\text{H}_2\text{O})$; 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.

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.

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

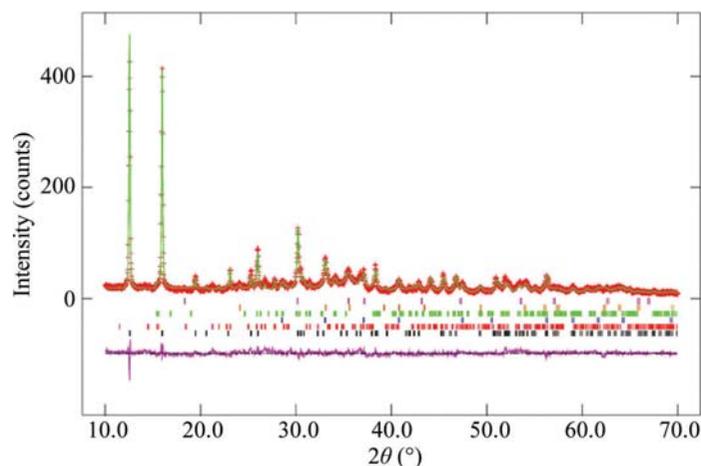


Figure 3.7.10
The final Rietveld plot from refinement of the isocracker deposit.

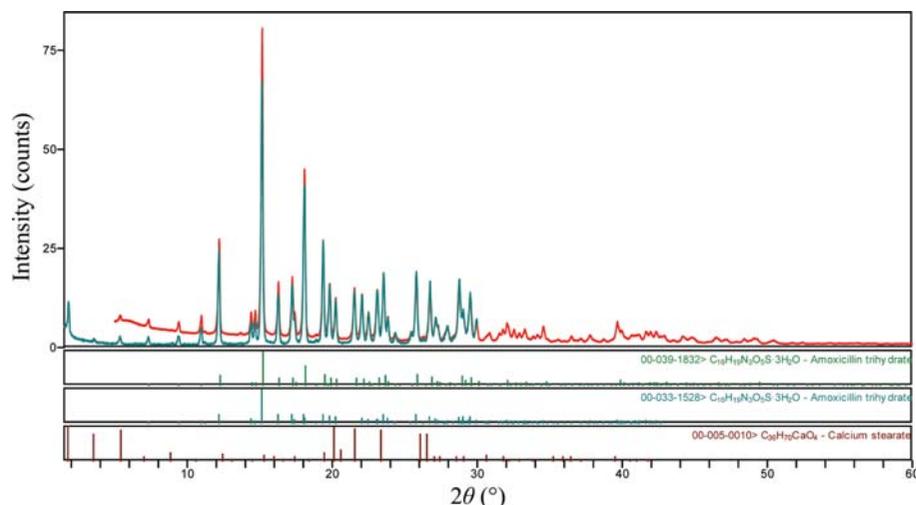


Figure 3.7.11
Phases identified in amoxicillin powder from a commercial capsule.

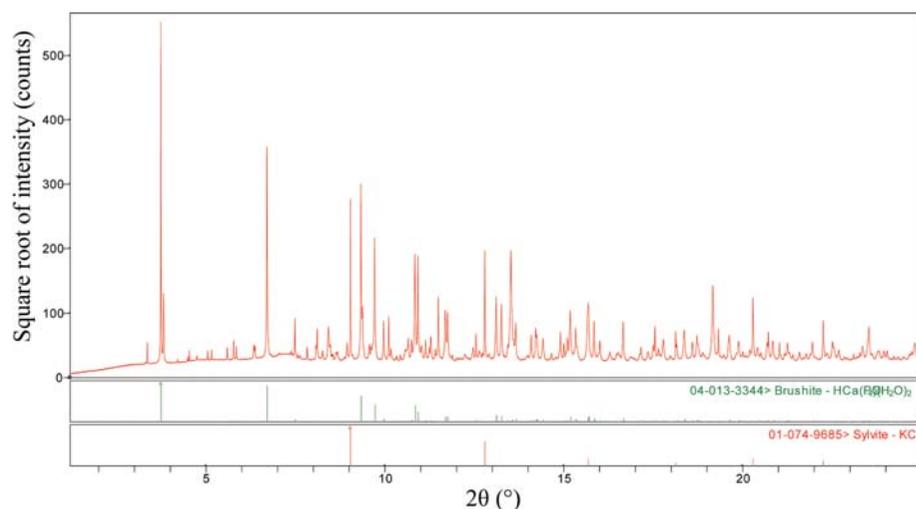


Figure 3.7.12
Phases identified by automated search/match in a Centrum A to Zn multivitamin tablet. Additional phases were identified using the native capabilities of the Powder Diffraction File.