JOURNAL OF APPLIED CRYSTALLOGRAPHY

ISSN: 1600-5767
journals.iucr.org/j

Bayesian approach to powder phase identification

Alexander Mikhalychev and Alex Ulyanenkov

J. Appl. Cryst. (2017). **50**, 776–786

JOURNAL OF APPLIED CRYSTALLOGRAPHY
ISSN 1600-5767

Bayesian approach to powder phase identification

Alexander Mikhalychev[a]* **and Alex Ulyanenkov**[b]

[a]Atomicus OOO, Minsk, Belarus, and [b]Atomicus GmbH, Karlsruhe, Germany. *Correspondence e-mail: alexander.mikhalychev@atomicus.by

Received 18 April 2016
Accepted 20 March 2017

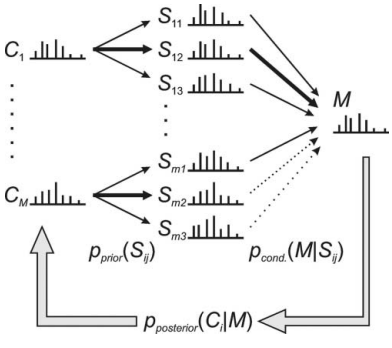
Edited by G. Renaud, CEA-Grenoble DSM/INAC/SP2M/NRS, Grenoble, France

Keywords: powder diffraction; phase identification; Bayesian approach.

Identification of unknown materials using X-ray powder diffraction patterns is a commonly used and well established technique with a number of proved implementations. Generally, qualitative phase analysis of X-ray diffraction data includes ranking of candidate phases on the basis of similarity of their diffraction patterns to the measured one. A standard strategy of such a ranking by algorithmization of manual search criteria may become inconvenient for modification and adaptation for problems that are not supported by our intuition. Here, the problem of providing physically grounded expressions for candidate phase ranking is addressed. The approach is based on calculation of Bayesian posterior probabilities of the phases' presence in the sample. The choice of the expressions for the prior probabilities for deviations of phases' diffraction patterns from database entries determines the degree of physical detailing and may be made according to the specifics of the problem being solved. It is shown that even for simple exponential expressions for prior probabilities the approach identifies the phases for IUCr round robin cases correctly, as well as ensuring sufficient robustness of the results with respect to diffraction peak shifts and intensity variations.

1. Introduction

Identification of investigated materials using powder X-ray diffraction measurements is a general tool which is commonly used in materials research and natural sciences (see *e.g.* Jenkins & Snyder, 1996; Mittemeijer & Scardi, 2004; Clearfield *et al.*, 2008; Pecharsky & Zavalij, 2009). Qualitative phase analysis of X-ray diffraction patterns originates from manual search techniques, based on comparing diffraction peaks of the investigated sample with a database of reference patterns (Winchell, 1927; Waldo, 1935; Hanawalt *et al.*, 1938; Bigelow & Smith, 1965; Hanawalt & Rinn, 1986). Each database entry contained several selected peaks of the known substance and was indexed in a particular way. Computer implementations of such search and match strategies enabled fast analysis of much larger databases (Frevel, 1965; Nichols, 1966; Snyder, 1981; Schreiner *et al.*, 1982; Jenkins, 1994; Langford & Louër, 1996 Faber *et al.*, 2004). Owing to improved computer performance, phase identification by comparing complete 'stick patterns' (pairs of positions and intensities of all the observed diffraction peaks) of investigated and reference materials became feasible (Caussin *et al.*, 1988; Nusinovici & Bertelmann, 1993; Nusinovici & Winter, 1994; Toby, 2005; Altomare *et al.*, 2008, 2015). An alternative approach, based on matching whole diffraction patterns, has also been proposed (Gilmore *et al.*, 2004; Barr *et al.*, 2004, 2009; Faber & Blanton, 2008). Elimination of the reduction of the whole X-ray diffraction pattern to a stick pattern leads to more accurate results, especially when certain peaks of the phases constituting the sample are overlapping. However, for implementation of this approach,



the whole diffraction patterns for the known phases must be either stored in the database or simulated 'on-the-fly', which may be memory or time consuming for analysis of hundreds of thousands of database entries.

Generally, the qualitative phase analysis of diffraction patterns includes a choice of the best candidate phases from a reference database according to some measure of their similarity to the measured data, quantified by a figure of merit (FOM). A standard strategy of FOM calculation (see *e.g.* Jenkins & Snyder, 1996; Mittemeijer & Scardi, 2004; Altomare *et al.*, 2008, 2015) is to combine several similarity criteria (correspondence of peak positions, intensities *etc.*) so that candidate ranking is performed according to the specifics of the problem to be solved. This approach, being quite natural as an algorithmization of manual search techniques, becomes less clear when modification and adaptation of the method are required for problems that are not intuitively formalized: for example, accounting for coinciding peaks of several phases, 'additive' search, solid solutions, combined analysis of X-ray diffraction and fluorescence data, *etc*.

There are two main questions to be answered in this context: (i) can the search and match strategy be formulated in a way suitable for modification and adaptation without referring to the intuition and experience of the scientist, and (ii) can the phase identification procedure be derived from certain basic physical assumptions, rather than being stated in its final form?

The phase identification problem is just a particular case of sample model reconstruction by optimal fitting of the measured signal, commonly encountered in physics. A well established approach to quantitative ranking of the suitability of possible models is calculation of their posterior probabilities, which take into account the measured signal by Bayes' equation. This approach has been shown to be effective for processing of X-ray intensity data, for deconvolution of reflections and background subtraction (Gilmore, 1996; David & Sivia, 2001), for solving crystal structures (Gilmore, 1996; Sivia & David, 2001; Marks *et al.*, 1999), for excluding biases from Rietveld refinement of powder diffraction data (Bergmann & Monecke, 2011), for accurate analysis of electron densities (Gilmore, 1996), and even for more exotic tasks such as, for example, 'data pattern' quantum tomography (Mikhalychev *et al.*, 2015).

In this paper, we formulate a Bayesian framework for deriving physically grounded expressions for FOM calculation. We consider only stick patterns, rather than whole X-ray diffraction patterns. Nevertheless, the basic ideas, outlined in §§2 and 3, are applicable to whole pattern analysis as well.

The starting point of the approach is to define the prior probabilities for deviations of the diffraction patterns of the phases that may be present in the investigated sample from the corresponding database entries. Together with the likelihood of the observed pattern, these prior probabilities are used for calculation of posterior probabilities for the presence of each phase in the sample. The obtained values quantify the correspondence between the diffraction patterns of the candidate phases and the observed ones and can be used as FOM values. The approach can include any desired level of physical detailing in prior probability specification and is applicable for multi-phase samples exhibiting coinciding peaks, for calculation of a 'collective' FOM for a combination of several reference substances, and for reliably finding intensity scales for subsequent quantitative analysis.

To demonstrate use of the approach, we provide an example of a search and match strategy built with simple exponential expressions for the prior probabilities. A statistical description of the database is used to take into account residual peaks. The derived expressions for FOM calculation have been applied to data from IUCr round robin examples (Madsen *et al.*, 2001; Scarlett *et al.*, 2002) and demonstrate good results both for identification of phases and for quantification of their abundance. The generality of the developed approach enables one to represent FOM values from other approaches as posterior probabilities of candidates for specially chosen prior probabilities.

The paper is organized as follows: in §2 we discuss the main idea of employing the Bayesian approach for the phase identification problem. Then, in §3, this idea is used for derivation of general equations describing the Bayesian search–match strategy. §4 introduces a simple model with exponential expressions for prior probabilities and a statistical description of residual peaks. Finally, we present the results of testing the discussed phase identification method on IUCr round robin data and of algorithm robustness analysis, and discuss the relation of the reported method to other approaches.

2. Basic idea

The phase identification technique would be trivial in the case when the measured pattern for any phase always coincides with the corresponding database entry. The search and match procedure then would consist just of selection of all the phases for which all the expected peaks are present in the measured pattern. In reality, however, the measured patterns even for single-phase materials differ from the reference ones. Such

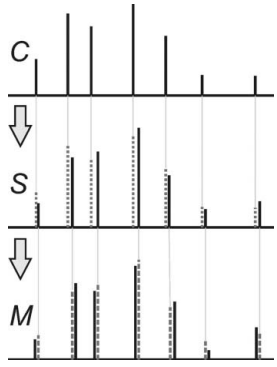


Figure 1
Difference between reference and measured patterns: the reference pattern, determined by database entry *C*, is transformed into the phase pattern *S* in the considered sample owing to parameter deviations; the pattern *S* differs from the measured pattern *M* because of measurement and data processing inaccuracies.

differences between the measured and the reference patterns (Fig. 1) are caused by

(*a*) the difference between reference and investigated substances for the same pattern (preferred orientation, modifications of crystal lattice *etc.*) and changeable measurement conditions and

(*b*) measurement effects, caused by the instrumental function and statistical noise, and data processing inaccuracies (for example, background removal and peak search procedure).

To solve the phase-identification problem under such non-ideal conditions, certain assumptions have to be made about the probabilities of the above-mentioned pattern deviations. This step implicitly or explicitly exists in any search–match approach: it may be based on the user-defined tolerance Δd of peak positions or on the peak's association window, estimated from the peak width, or introduced in any other way. In opposition to these techniques, the Bayesian approach enables explicit incorporation of the probabilities into the equations for phase identification.

Bayes' formula, applied to the problem of an object's identification by its pattern, links the posterior probability of an object with its prior probability (initial guess) and the conditional probability (likelihood) of the observed pattern if this object was the true one. For the problem of powder phase identification, Fig. 2 schematically shows the possible ways to obtain the observed pattern M. For the sake of simplicity, we focus here on a single-phase sample. The phase, described by the ith reference database pattern C_i, may have some deviations of its parameters in the observed sample and produce one of the possible patterns S_{ij} (where index j enumerates possible sets of deviated parameters), each characterized by its prior probability $p_{\rm prior}(S_{ij})$. Having some information or, at least, assumptions about measurement and data processing

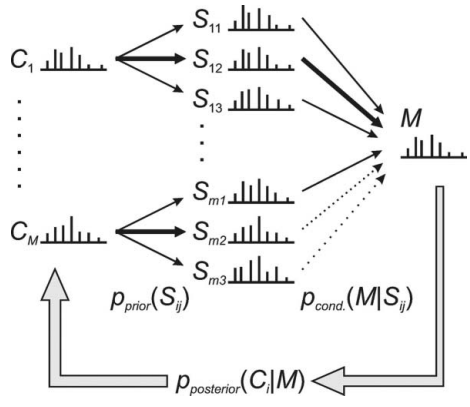


Figure 2
The possible ways to obtain the measured pattern M: for each database entry C_i there exist several possible phase patterns S_{ij}, corresponding to different sets of pattern deviation parameters. For each phase pattern S_{ij} there exists a certain likelihood (sometimes equal to zero) of observation of the realized pattern M, shown by the thickness of the arrows connecting the patterns. The thickness of the arrows connecting the database entries C_i with the changed phase patterns S_{ij} indicates the prior probabilities $p_{\rm prior}(S_{ij})$ of the corresponding deviations. The posterior probabilities of the patterns S_{ij} and, consequently, the probabilities of the presence of reference phases described by database entries C_i in the sample are determined from Bayes' formula.

operations, the likelihood $p_{\rm cond.}(M \mid S_{ij})$ of observing pattern M can be estimated, if the real pattern of the sample was S_{ij}. The posterior probability of the investigated single-phase sample being represented by the ith phase with the pattern S_{ij} equals

$$p_{\rm posterior}(S_{ij} \mid M) = \frac{p_{\rm prior}(S_{ij})\, p_{\rm cond.}(M \mid S_{ij})}{\sum_{i',j'} p_{\rm prior}(S_{i'j'})\, p_{\rm cond.}(M \mid S_{i'j'})}. \quad (1)$$

The posterior probability of the presence of phase C_i in the sample in any of its possible variations S_{ij} is described by the following expression:

$$\begin{aligned} p_{\rm posterior}(C_i \mid M) &= \sum_j p_{\rm posterior}(S_{ij} \mid M) \\ &= \frac{\sum_j p_{\rm prior}(S_{ij})\, p_{\rm cond.}(M \mid S_{ij})}{\sum_{i',j'} p_{\rm prior}(S_{i'j'})\, p_{\rm cond.}(M \mid S_{i'j'})}. \end{aligned} \quad (2)$$

The value obtained characterizes the degree of similarity between the database entry C_i, describing the considered phase, and the measured pattern M and, therefore, can be used as a figure of merit for ranking of the candidate phases.

This formula can also be interpreted in terms of the Bayesian approach to the problem of model selection (see *e.g.* Jaynes, 2003). Each database entry C_i defines a model for the single-phase sample, and the pattern S_{ij}, provided by the model for the investigated sample, depends on some model parameters $\mathbf{\Theta}_j^{(i)}$ with prior probabilities $p_{\rm prior}(\mathbf{\Theta}_j^{(i)})$. The likelihood for the observation of the pattern M for the model C_i with parameters $\mathbf{\Theta}_j^{(i)}$ is equal to $p_{\rm cond.}(M \mid S_{ij})$. Therefore, the evidence of this model is

$$\begin{aligned} E(C_i \mid M) &= \sum_j p_{\rm prior}(S_{ij})\, p_{\rm cond.}(M \mid S_{ij}) \\ &= p_{\rm posterior}(C_i \mid M) \sum_{i',j'} p_{\rm prior}(S_{i'j'})\, p_{\rm cond.}(M \mid S_{i'j'}), \end{aligned} \quad (3)$$

which is just the posterior probability described by equation (2) multiplied by a constant normalization factor, having the meaning of the total prior probability for observation of the pattern M. This is the evidence commonly used for the quantification of the model applicability, which is another argument for using this value as the figure of merit.

The discussed idea can easily be generalized for multi-phase samples. Replacing the single-phase patterns S_{ij} by the multi-phase patterns $S_{ij\ldots;k}$, formed by the combination of several database entries C_i, $C_j, \ldots$, the posterior probabilities $p_{\rm posterior}(S_{ij\ldots;k} \mid M)$ and $p_{\rm posterior}(C_i, C_j, \ldots \mid M) = \sum_k p_{\rm posterior}(S_{ij\ldots;k} \mid M)$ can be calculated. The following issues have to be addressed in such a generalization:

(*a*) Any reasonable finite combination of phases may explain the measured pattern only partially; the posterior probability of such a combination depends, generally speaking, on our ability to explain the residual pattern by the residual part of the database.

(*b*) Different phases may have almost completely coinciding positions of certain peaks.

In §4 we show that the former problem can be solved by using an approximate statistical description of the residual part of the database instead of direct calculation of all possible

combinations of minor phases, which is not feasible. The problem of providing correct association of the measured peaks with the coinciding peaks of several candidates is solved by maximizing the posterior probabilities of the phases' combination.

3. Derivation of phase identification equations

3.1. Parameterization

To provide a strict mathematical description of the ideas briefly discussed in §2, the following notations are used. Let C_i be the reference stick pattern for the ith database entry. Such a pattern is represented by a set of pairs (d, I) of positions d and intensities I of the powder diffraction peaks: $C_i = \{(d_{ik}, I_{ik})\}_k$. The deviation of the considered phase properties in the sample relative to the reference ones is described by a vector of parameters $\mathbf{\Theta}^{(i)}$. Each vector $\mathbf{\Theta}_j^{(i)}$ of these deviation parameters corresponds to a certain phase pattern $S_{ij} = S_i(\mathbf{\Theta}_j^{(i)})$. The prior probability of these values of the deviation parameters is further denoted by $p_i(\mathbf{\Theta}_j^{(i)})$. For unification of the description, we consider the scale factor ρ_i of the phase C_i in a multi-phase sample, proportional to the phase concentration, as one of the deviation parameters in the vector $\mathbf{\Theta}^{(i)}$.

For a combination of phases $\Omega = \{C_1, \ldots, C_N\}$ with deviation parameters $\mathbf{\Theta}_\Omega = \{\mathbf{\Theta}^{(1)}, \ldots, \mathbf{\Theta}^{(N)}\}$, the expected pattern $J(\Omega, \mathbf{\Theta}_\Omega)$ represents the union of the individual patterns $S_1(\mathbf{\Theta}^{(1)}), \ldots, S_N(\mathbf{\Theta}^{(N)})$:

$$J(\Omega, \mathbf{\Theta}_\Omega) = \bigcup_{i:C_i\in\Omega} S_i(\mathbf{\Theta}^{(i)}). \quad (4)$$

The likelihood of observing the measured pattern M for this combination of phases is described by the conditional probability $p(M \mid \Omega, \mathbf{\Theta}_\Omega)$.

3.2. Posterior probabilities

According to Bayes' formula, a combination of phases Ω with deviation parameters $\mathbf{\Theta}_\Omega$ provides the correct sample model for the observed pattern M with the posterior probability

$$p(\Omega, \mathbf{\Theta}_\Omega \mid M) = \frac{p_\Omega(\mathbf{\Theta}_\Omega)\, p(M \mid \Omega, \mathbf{\Theta}_\Omega)}{\sum_{\Omega', \mathbf{\Theta}_{\Omega'}} p_{\Omega'}(\mathbf{\Theta}_{\Omega'})\, p(M \mid \Omega', \mathbf{\Theta}_{\Omega'})}, \quad (5)$$

where $p_\Omega(\mathbf{\Theta}_\Omega) = \prod_{i:C_i\in\Omega} p_i(\mathbf{\Theta}^{(i)})$ is the total prior probability for the deviation parameters; the summation in the denominator is performed over all of the phase and deviation parameter combinations, providing nonzero likelihood of observing the pattern M. The methods for effective calculation of these sums are discussed in §4.

The total posterior probability of complete sample description by the phase combination Ω can be found as the sum of the probabilities given by equation (5) over all deviation parameter values:

$$p(\Omega \mid M) = \sum_{\mathbf{\Theta}_\Omega} p(\Omega, \mathbf{\Theta}_\Omega \mid M). \quad (6)$$

Finally, the probability of the phase described by the database entry C_i being present in the investigated sample equals the sum of posterior probabilities of all of the combinations Ω containing the database entry C_i:

$$p(C_i \mid M) = \sum_{\Omega:C_i\in\Omega} p(\Omega \mid M). \quad (7)$$

This probability (or any of its monotonic functions) can be used as the figure of merit for the search and match procedure, if no additional information about the sample is available. If such information is available, for example the elemental composition of the sample from X-ray fluorescence (XRF) data, it can be taken into consideration by modifying the expressions for the prior probabilities. Because of instrumental limitations, XRF data may give certain estimates (probabilities) of the presence of elements in the sample instead of their accurate discrimination, especially for light elements. For high-quality XRF data with completely resolved elemental composition, these probabilities can be set as 0 or 100% for absent and present elements, respectively. Then, the prior probability w_i of the reference phase C_i can be estimated as a product of the XRF-based probabilities of all the elements composing the phase. The total prior probability $p_\Omega(\mathbf{\Theta}_\Omega)$ is modified to the following expression:

$$p_\Omega(\mathbf{\Theta}_\Omega) = \prod_{i:C_i\in\Omega} p_i(\mathbf{\Theta}^{(i)}) w_i. \quad (8)$$

3.3. Multi-phase search strategies

The most straightforward approach for the analysis of multi-phase samples is to consider all possible combinations of phases Ω and to select the best combination, characterized by maximal posterior probability $p(\Omega \mid M)$. For a typical database with 10^5 entries, there exist 10^{10} possible combinations even for a two-phase sample.

There are two main approaches for the analysis of multi-phase samples: (i) subtractive search: the iterative search of single phases with update of information about the residual peaks after identification of each new phase (see *e.g.* Clearfield *et al.*, 2008); and (ii) additive search: the search for a combination from some preselected list of phases providing the best combined explanation of the observed pattern (Schreiner *et al.*, 1982).

The former approach requires calculation of posterior probabilities for the presence of phases in the sample, conditioned by the previously identified phases Ω_0, which are assumed to be present in the sample, as well as by the measured pattern M:

$$p(C_i \mid M, \Omega_0) = \frac{\sum_{\Omega:C_i\in\Omega,\, \Omega_0\subset\Omega} p(\Omega \mid M)}{\sum_{\Omega:\Omega_0\subset\Omega} p(\Omega \mid M)}. \quad (9)$$

The summation in the numerator is performed over all possible phase combinations Ω that include both the phase C_i and the phases Ω_0, which have already been identified, while the sum in the denominator includes the posterior probabilities of all the combinations Ω containing the subset Ω_0 of identified phases but not necessarily containing the phase C_i.

For formulation of the Bayesian additive search approach it is necessary to take into account the possible presence of minor phases. The figure of merit for each combination of phases Ω_0 can be calculated as the total posterior probability of all combinations Ω that include Ω_0 as a subset:

$$p'(\Omega_0 \mid M) = \frac{\sum_{\Omega:\Omega_0\subset\Omega} p(\Omega \mid M)}{\sum_{\Omega} p(\Omega \mid M)}. \quad (10)$$

A typical way of using this expression is to select the set of candidate phases described by sufficiently high values of individual figures of merit $p(C_i \mid M)$, and then to construct different combinations of the selected phases by iterative addition of phases, with the aim of identifying those combinations that provide maximal growth of the total figure of merit.

Equations (7), (9) and (10) specify the general formalism for a Bayesian search and match procedure for different phase identification strategies. To construct a practical phase identification scheme on their basis, the following definitions have to be provided:

(*a*) A model for pattern deviations and prior probabilities $p_i(\mathbf{\Theta}^{(i)})$ for the deviation parameters used.

(*b*) Conditional probabilities $p(M \mid \Omega, \mathbf{\Theta}_\Omega)$ corresponding to a certain model of data acquisition.

(*c*) A method for approximate calculation of sums over all combinations of phases in equations (5), (7), (9) and (10).

The models selected at this stage of strategy construction determine the trade-off between accuracy and calculation complexity in accordance with the specifics of the problem to be solved.

4. Simple model

4.1. Conditional probabilities

To illustrate the applicability of the reported approach to phase identification, we consider a simple model and demonstrate the sufficiency of the method to solve all the basic problems. Note that the presented model is just an example of a search and match strategy that can be constructed by the above-discussed method. Finding the optimal Bayesian strategy for a particular class of phase identification tasks is a separate problem not considered here, with its solution depending on the specifics of the investigated samples and the measurement conditions.

Assuming sufficiently accurate experimental data, one can model the likelihood of observing the measured pattern M by the following expression:

$$p(M \mid \Omega, \mathbf{\Theta}_\Omega) = (\Delta d \Delta I)^K \delta^{(2K)}[M - J(\Omega, \mathbf{\Theta}_\Omega)], \quad (11)$$

where Δd and ΔI are the inaccuracies of the experimental determination of peak positions and intensities, respectively; K is the number of measured pattern peaks exceeding some cutoff intensity I_0 (determined by the level of background signal); and the delta function on the right-hand side of the equation describes the coincidence requirement for the positions and intensities of all the peaks higher than I_0 in the pattern M and in the predicted pattern $J(\Omega, \mathbf{\Theta}_\Omega)$ for the set of phases Ω.

The condition (11) implies that the main contribution to the sums in equations (5) and (6) is made by the term with optimal vector of deviation parameters $\mathbf{\Theta}_\Omega(M)$, providing a zero value of the delta-function argument in equation (11) and maximizing the prior probability $p_\Omega(\mathbf{\Theta}_\Omega)$:

$$p(\Omega \mid M) \simeq p[\Omega, \mathbf{\Theta}_\Omega(M) \mid M] \simeq \frac{p_\Omega[\mathbf{\Theta}_\Omega(M)]}{\sum_{\Omega'} p_{\Omega'}[\mathbf{\Theta}_{\Omega'}(M)]}. \quad (12)$$

The expression obtained contains the prior probabilities of optimal deviation parameters for the considered combinations of phases. The sum in the denominator includes all of the combinations of phases Ω' providing the explanation of the observed pattern M.

4.2. Prior probabilities

In the considered model, the deviation parameters vector $\mathbf{\Theta}^{(i)}$ for the ith database entry includes the phase scale factor ρ_i, which is the fixed scaling multiplier for the intensities of all the peaks of the pattern; the relative displacements of each of the peaks $\delta_{ij} = (\overline{d}_{ij} - d_{ij})/d_{ij}$, where d_{ij} is the jth peak position in the pattern C_i and $\overline{d}_{ij}$ is the position of the corresponding measured peak; and the relative intensity changes for each peak $\varepsilon_{ij} = \overline{I}_{ij}/(\rho_i I_{ij})$, where I_{ij} and $\overline{I}_{ij}$ are the intensities of the jth peak in the reference pattern C_i and in the measured pattern M, respectively.

For simplicity, we assume the deviations of the positions and the intensities are independent for different peaks and, therefore, the prior probabilities of deviation vectors have the following form:

$$p_i(\mathbf{\Theta}^{(i)}) = \prod_j p_d(\delta_{ij}) p_I(\varepsilon_{ij}). \quad (13)$$

Several examples of suitable probability distribution functions are shown in Fig. 3. The factorized form of equation (13) implies that logarithms of the probabilities are additive and have a simple interpretation, where the quality of match between measured and reference patterns is characterized by

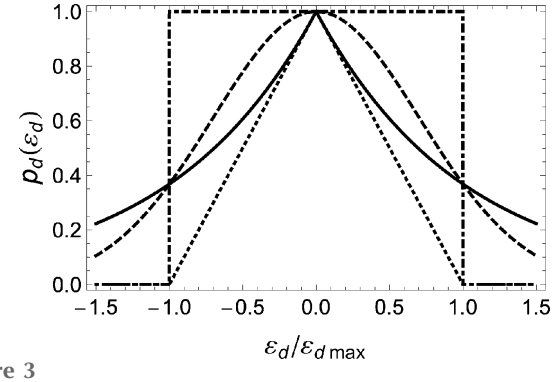


Figure 3
Examples of probability distribution functions for the prior probabilities of relative deviations of peak positions $\varepsilon_d = \delta/d$: exponential $p_d(\varepsilon_d) = \exp(-|\varepsilon_d|/\varepsilon_{d\,\max})$ (solid line), Gaussian $p_d(\varepsilon_d) = \exp(-\varepsilon_d^2/\varepsilon_{d\,\max}^2)$ (dashed line), triangular (linear) $p_d(\varepsilon_d) = \max(0, 1 - |\varepsilon_d|/\varepsilon_{d\,\max})$ (dotted line), and rectangular $p_d(\varepsilon_d) = 1$ for $|\varepsilon_d| < \varepsilon_{d\,\max}$ and 0 otherwise (dot–dashed line).

a sum of 'fines' for mismatches of peak positions and intensities. The most simple final expressions are obtained when the logarithms of the probabilities depend linearly on the deviations. For this reason, the following simple exponential form is used in the constructed model:

$$\log p_d(\delta_{ij}) = -\frac{|\delta_{ij}|}{d_{ij}\varepsilon_{d\max}} \tag{14}$$

and

$$\log p_I(\varepsilon_{ij}) = -\frac{|\log \varepsilon_{ij}|}{\log \varepsilon_{I\max}} = -\frac{|\log \varepsilon_{ij}|}{\beta}. \tag{15}$$

The constants $\varepsilon_{d\max}$ and $\beta = \log \varepsilon_{I\max}$ determine the sensitivity of the resulting figure of merit to the peak displacements and to the intensity deviations, respectively.

4.3. Statistical description of residual peaks

Both the normalization factor, represented by the sum over all combinations of phases Ω' in the denominator of equation (12), and the sum over all combinations Ω containing the phase C_i in equation (7) can be estimated using statistical treatment of the database by applying equations (13)–(15) instead of performing the explicit analysis of all available combinations.

The substitution of equation (12) into equation (7) results in the following expression for the posterior probability of the presence of phase C_i in the sample:

$$p(C_i \mid M) = \frac{\sum_{\Omega: C_i \in \Omega} p_\Omega[\mathbf{\Theta}_\Omega(M)]}{\sum_{\Omega'} p_{\Omega'}[\mathbf{\Theta}_{\Omega'}(M)]}. \tag{16}$$

The sum in the denominator utilizes both types of combinations: the ones including the phase C_i and the ones not including it. Therefore, the posterior probability can be represented as

$$p(C_i \mid M) = \frac{\overline{p}(C_i \mid M)}{\overline{p}(C_i \mid M) + 1}, \tag{17}$$

where

$$\overline{p}(C_i \mid M) = \frac{\sum_{\Omega: C_i \in \Omega} p_\Omega[\mathbf{\Theta}_\Omega(M)]}{\sum_{\Omega': C_i \notin \Omega'} p_{\Omega'}[\mathbf{\Theta}_{\Omega'}(M)]} \tag{18}$$

is the ratio of the posterior probabilities of representing the investigated sample by combinations of phases including and not including C_i, respectively. The denominator describes the total probability $P_{\rm DB}(M)$ of explaining the measured pattern M by the residual part of the database (after excluding the phase C_i). Neglecting the dependence of the optimal deviation parameters $\mathbf{\Theta}^{(i)}$ on the other phases in the combination Ω, the numerator can be represented as a product of the probability $p_i[\mathbf{\Theta}^{(i)}(M)]$ of the phase C_i and the total probability of explaining the residual part of the measured pattern by the residual part of the database:

$$\overline{p}(C_i \mid M) = \frac{p_i[\mathbf{\Theta}^{(i)}(M)] P_{\rm DB}\{M \setminus S_i[\mathbf{\Theta}^{(i)}(M)]\}}{P_{\rm DB}(M)}, \tag{19}$$

where the residual pattern $M \setminus S_i[\mathbf{\Theta}^{(i)}(M)]$ is obtained by subtracting the peaks of the phase C_i from the measured pattern M.

The expression for the probability of interpreting a certain pattern J by the statistically described residual part of the database follows from simple basic statements. Let Ω be one of the most suitable combinations of phases to describe the considered pattern J, and let C_k be one of the phases from this combination. N_k peaks of the reference pattern C_k are associated with certain peaks of the pattern J, while $\bar{N}_k$ peaks of this pattern do not have any measured counterparts. Equation (11) implies that any peak of the considered reference pattern must either be associated with some measured peak or have an intensity that is not greater than the background level I_0. According to equations (13)–(15), the logarithm of the prior probability of the optimal deviation parameters for the phase C_i equals

$$\log p_k(\mathbf{\Theta}^{(k)}) = \sum_{j=1}^{N_k} \left[\log p_d(\delta_{kj}) + \log p_I(\varepsilon_{kj})\right] - \sum_{j=N_k+1}^{N_k+\bar{N}_k} \log p_I\left(\rho_k I_{kj}/I_0\right), \tag{20}$$

where the first and the second sums correspond to the associated and non-associated peaks, respectively. By averaging over the database and introducing the parameters $\alpha_d = -\langle \log p_d(\delta) \rangle$ and $\alpha_I = -\langle \log p_I(\varepsilon) \rangle$ for the mean log probability of the optimal associated peak deviations, the following relation is derived:

$$\log p_k(\mathbf{\Theta}^{(k)}) \simeq -N_k(\alpha_d + \alpha_I) - \bar{N}_k \log p_I(\rho_k \langle I_k \rangle / I_0), \tag{21}$$

where $\langle I_k \rangle$ is the average intensity of non-associated peaks of the phase C_k.

The contribution to the log probability $\log P_{\rm DB}(J)$ provided by the right-hand side of equation (21) corresponds to the interpretation of N_k measured peaks by the phase pattern C_i. When divided by the number of explained peaks N_k and averaged over the possible near-optimal interpretations of the measured pattern J, this value provides the contribution of a single measured peak to the total log probability of pattern interpretation by the residual part of the database:

$$\log P_{\rm DB}(J) = \sum_{j:\mathrm{peak}_j \in J, I_j > I_0} \left[-\alpha_d - \alpha_I - \eta \log p_I(I_j/I_0)\right], \tag{22}$$

where η is the average ratio of the number of non-associated peaks of added phases to the number of measured peaks correctly interpreted by these phases. In practice, the parameters α_d, α_I and η are chosen rather from the condition of good performance of the constructed phase identification algorithm than from collection of the real statistics delivered by the database and various samples.

4.4. Final expressions for single-phase posterior probability

Expression (22) for the probability of successful explanation of the residual peaks by the remaining part of the database consists of independent terms, each one corresponding to a single peak of the pattern J. Therefore, the numerator and

denominator of equation (19) can be canceled by equal contributions of the peaks that are not present in the reference pattern C_i. The final expression, therefore, takes the following form:

$$\log \overline{p}(C_i \mid M) = \sum_{j:\mathrm{peak}_j \in C_i, I_j > I_0} F_{ij}. \quad (23)$$

Here, the single-peak terms differ for the peaks associated and not associated with certain peaks of the measured pattern M:

$$F_{ij} = \begin{cases} F_{ij}^{(d)} + F_{ij}^{(I)}, & \text{for associated peaks;} \\ -(1/\beta)\log(\rho_i I_{ij}/I_0), & \text{for not associated peaks.} \end{cases} \quad (24)$$

The peak displacement term is written as

$$F_{ij}^{(d)} = -\frac{|\overline{d}_{ij} - d_{ij}|}{d_{ij}\varepsilon_{d\,\max}} + \alpha_d, \quad (25)$$

where $\overline{d}_{ij}$ is the position of the measured peak associated with the considered reference peak. The intensity term is

$$F_{ij}^{(I)} = \frac{\eta}{\beta}\log\frac{\overline{I}_{ij}}{I_0} - \frac{1}{\beta}\left|\log\frac{\overline{I}_{ij}}{\rho_i I_{ij}}\right| + \alpha_I, \quad (26)$$

where $\overline{I}_{ij}$ is the intensity of the measured peak associated with the considered peak.

The scale factor ρ_i for the considered phase C_i is chosen from the condition of maximality of the probability ratio $\overline{p}(C_i \mid M)$, which also provides the maximality of the posterior probability $p(C_i \mid M)$, the latter being an increasing monotonic function of $\overline{p}(C_i \mid M)$.

4.5. Final expressions for multi-phase identification strategies

For the subtractive search strategy, equation (9) can be rewritten in the following form, similar to equations (17) and (19):

$$p(C_i \mid M, \Omega_0) = \frac{\overline{p}(C_i \mid M, \Omega_0)}{\overline{p}(C_i \mid M, \Omega_0) + 1}, \quad (27)$$

$$\overline{p}(C_i \mid M, \Omega_0) = \frac{p_i[\boldsymbol{\Theta}^{(i)}(M)]\, P_{\mathrm{DB}}[M \setminus J(\Omega_0 \cup \{C_i\})]}{P_{\mathrm{DB}}[M \setminus J(\Omega_0)]}, \quad (28)$$

where $M \setminus J(\Omega_0)$ is the set of residual peaks obtained by subtracting the peaks of the deviated reference patterns for the phases Ω_0 from the measured pattern M; $M \setminus J(\Omega_0 \cup \{C_i\})$ is the set of residual peaks after the pattern of the phase C_i is subtracted too. The probability ratio $\overline{p}(C_i \mid M, \Omega_0)$ is calculated in the same way as $\overline{p}(C_i \mid M)$ [see equations (23)–(26)], but with the residual pattern $M \setminus J(\Omega_0)$ used instead of pattern M.

Expression (10) for the additive search strategy can be rewritten as

$$p'(\Omega_0 \mid M) = \frac{\overline{p}'(\Omega_0 \mid M)}{\overline{p}'(\Omega_0 \mid M) + 1}, \quad (29)$$

where

Table 1
Iterations of the subtractive search procedure for Sample 2.

Iteration	Phase	FOM
1	**Zincite ZnO**	1.000
	MoNi	0.999
	Fluorite CaF_2	0.984
	Corundum Al_2O_3	0.965
	Silicon Si	0.922
	Ce–Gd mixed oxide	0.879
	Brucite $Mg(OH)_2$	0.870
	Sodium yttrium fluoride	0.818
	Other phases	<0.800
2	**Zincite ZnO**	**Accepted**
	Fluorite CaF_2	0.979
	Silicon Si	0.922
	Brucite $Mg(OH)_2$	0.828
	Corundum Al_2O_3	0.819
	Ce–Gd mixed oxide	0.757
	Other phases	<0.700
3	**Zincite ZnO**	**Accepted**
	Fluorite CaF_2	**Accepted**
	Corundum Al_2O_3	0.818
	Brucite $Mg(OH)_2$	0.786
	Ce–Sm mixed oxide	0.712
	Ce–Gd mixed oxide	0.540
	Other phases	<0.500
4	**Zincite ZnO**	**Accepted**
	Fluorite CaF_2	**Accepted**
	Corundum Al_2O_3	**Accepted**
	Brucite $Mg(OH)_2$	0.720
	Ce–Gd mixed oxide	0.711
	MoNi	0.689
	Ce–Sm mixed oxide	0.544
	Other phases	<0.500
5	**Zincite ZnO**	**Accepted**
	Fluorite CaF_2	**Accepted**
	Corundum Al_2O_3	**Accepted**
	Brucite $Mg(OH)_2$	**Accepted**
	MoNi	0.625
	Ce–Gd mixed oxide	0.541
	Other phases	<0.500

$$\log \overline{p}'(\Omega_0 \mid M) = \sum_{j:\mathrm{peak}_j \in J(\Omega_0)} F_{ij}, \quad (30)$$

and the single-peak terms are described by equations (25) and (26). The scale factors for the phases are chosen from the condition of the posterior probability maximality.

5. Discussion

5.1. Subtractive search examples

The phase identification procedures based on the above-introduced model were tested using data from IUCr round robin examples (Madsen *et al.*, 2001; Scarlett *et al.*, 2002). The open-access data, provided by the IUCr Commission on Powder Diffraction, were analyzed for the following four samples:

Sample 1g (simple): corundum, fluorite, zincite.

Sample 2 (preferred orientation): corundum, fluorite, zincite, brucite.

Table 2
Iterations of the subtractive search procedure for Sample 4.

Iteration	Phase	FOM
1	**Corundum Al_2O_3**	1.000
	Zircon $ZrSiO_4$	0.997
	Magnetite Fe_3O_4	0.979
	Magnesioferrite $MgFe_2O_4$	0.976
	Cuprospinel $CuFe_2O_4$	0.976
	Other phases	<0.975
2	**Corundum Al_2O_3**	**Accepted**
	Zircon $ZrSiO_4$	0.992
	Magnetite Fe_3O_4	0.957
	Magnesioferrite $MgFe_2O_4$	0.956
	Other phases	<0.950
3	**Corundum Al_2O_3**	**Accepted**
	Zircon $ZrSiO_4$	**Accepted**
	Magnetite Fe_3O_4	0.731
	Magnesioferrite $MgFe_2O_4$	0.730
	Other phases	<0.680
4	**Corundum Al_2O_3**	**Accepted**
	Zircon $ZrSiO_4$	**Accepted**
	Magnetite Fe_3O_4	**Accepted**
	Magnesioferrite $MgFe_2O_4$	0.493
	Other phases	<0.450

Sample 3 (amorphous content): corundum, fluorite, zincite, glass.

Sample 4 (microabsorption): corundum, magnetite, zircon.

The purpose of the performed analysis was rather to illustrate an application of the derived expressions than to provide an accurate description of the data. For this reason, the only correction taken into account was the subtraction of the amorphous constituent signal for Sample 3. The Crystallography Open Database (COD; Grazulis *et al.*, 2012), containing about 346 000 entries, was used as the reference database.

Tables 1 and 2 show the results of the iterative subtractive search approach, applied to the test data for Samples 2 and 4 (the performance of the method for two other samples is similar). At the ith iteration of the search procedure, the $i-1$ phases are already identified (marked as 'accepted' in the tables). The correct phases of the test samples are printed in bold to provide better visual understanding of the results. For all four samples, the uppermost (*i.e.* most probable) candidate phase at each iteration was one of the correct and so far not identified phases from the known sample model. All the correct phases fell into at least the top ten of the candidates at the first step of the phase identification procedure and moved toward the top of the list after a few iterations. The total search time (including queries to the database) was about 1–2 s for the above-mentioned samples, analyzed on a standard PC with an Intel Core i5-2300 processor. Therefore, the introduced Bayesian approach, coupled with factorized exponential expressions for the prior probabilities, provides a reliable phase identification algorithm with good performance.

5.2. Additive search examples

The same test data were also analyzed by the additive search approach. The combinations of the phases characterized by $p(C_i \mid M) > 0.01$ were analyzed to provide the best description of the measured data. Fig. 4 shows the dependence of the probability ratio $\overline{p}(\Omega_k \mid M)$, which is a monotonic increasing function of the posterior probability $p(\Omega_k \mid M)$, for the best combination Ω_k of k phases as a function of the number of used phases k. This quantity grows when the number of phases is small, but saturates when the sample model contains three to five phases. Fig. 4 demonstrates that the saturation occurs approximately when the expected sample model is constructed. The saturation condition can be used as a stopping criterion for the additive search procedure; however, further investigation, falling out of the scope of this paper, is required to formulate such a condition in a more rigorous way.

5.3. Application to quantitative analysis

As follows from the comments before equation (12), finding the optimal deviation parameters vector $\boldsymbol{\Theta}_\Omega(M)$ includes the optimization of the scale factors ρ_i of the phases C_i present in the considered combination Ω. This information is useful for quantitative phase analysis and provides better stability than, for example, when using the intensity of the highest peak, especially for the situation with almost coinciding intense peaks of several phases. To illustrate this idea, we estimated the mass concentrations of the components for the four analyzed test samples using the reference intensity ratio (RIR) method (Hubbard *et al.*, 1976; Hubbard & Snyder, 1988) with the calculated RIR values for the database entries.

For a sample consisting of crystalline phases with known RIR values, a 'standardless' quantitative analysis can be performed by determining the weight fraction of the ith phase as the ratio of the intensity I_i of the highest peak of the phase, normalized by its RIR value RIR_i, to the sum of similarly normalized intensities over all of the present phases (Dinnebier & Billinge, 2008; Clearfield *et al.*, 2008; Chung, 1974*a*,*b*):

$$w_i = \frac{I_i/\mathrm{RIR}_i}{\sum_j I_j/\mathrm{RIR}_j}. \tag{31}$$

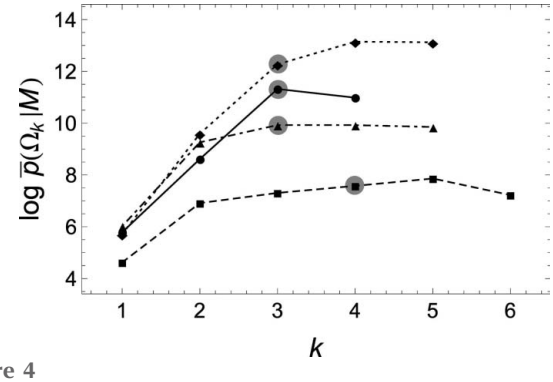


Figure 4
Dependence of the probability ratio $\overline{p}(\Omega_k \mid M)$ for the best combination Ω_k of k phases as a function of the number of used phases k. The solid line stands for Sample 1g, the dashed line for Sample 2, the dotted line for Sample 3 and the dot–dashed line for Sample 4. The gray circles indicate the construction of the sample model that is expected to be correct.

Table 3
Results of quantitative analysis of Samples 1g–4.

Sample	Phase	Estimated concentration (%)	Correct concentration (%)	Standard deviation (%)
1g	Zincite ZnO	33.35	34.21	5.21
	Fluorite CaF_2	32.62	34.42	6.35
	Corundum Al_2O_3	34.03	31.37	4.50
2	Zincite ZnO	19.43	19.94	5.21
	Fluorite CaF_2	24.09	22.53	2.81
	Corundum Al_2O_3	20.92	21.27	5.18
	Brucite $Mg(OH)_2$	35.56	36.26	7.35
3	Zincite ZnO	30.44	27.90	5.04
	Fluorite CaF_2	26.84	28.44	5.81
	Corundum Al_2O_3	42.72	43.65	6.51
4	Corundum Al_2O_3	64.61	50.46	15.56
	Zircon $ZrSiO_4$	21.46	29.90	9.99
	Magnetite Fe_3O_4	13.93	19.64	14.80

Let us consider the case when all the peak intensities of the database entries are normalized in such a way that the highest peak of each phase has an integrated intensity equal to 1. Then, if the effects of preferred orientation and peak overlap are negligible, the observed intensities I_j of the highest peaks of the sample constituents C_j must be close to the corresponding scale factors ρ_j: $I_j = 1 \times \rho_j$. In such an ideal situation, equation (31) can be rewritten in the following way:

$$w_i = \frac{\rho_i/\mathrm{RIR}_i}{\sum_j \rho_j/\mathrm{RIR}_j}. \qquad (32)$$

When preferred orientation of crystallites is present, equations (31) and (32) may yield different results. We suggest using equation (32), including the scale factors ρ_j, instead of equation (31) in this case. Calculation of the scale factors ρ_j during the phase identification procedure takes into account all of the peaks of the reference database entry and, therefore, is affected by preferred orientation to a lesser extent than the intensity of the highest peak (see §5.4 for quantitative comparison). Another advantage of the proposed approach is its robustness with respect to overlap of the peaks, ensured by the introduced Bayesian search and match algorithm.

The results of quantitative analysis are listed in Table 3, where the correct values (Madsen *et al.*, 2001; Scarlett *et al.*, 2002) and standard deviations of the data obtained by other approaches are also shown for comparison. For Sample 3 with an amorphous component, the concentrations were normalized by the total content of the crystalline phases. The significant deviations of the estimated concentrations from the correct ones for Sample 4 can be explained by microabsorption, which has a much stronger influence on the results of quantitative analysis than on the procedure of phase identification.

5.4. Robustness analysis

To test the robustness of the constructed phase identification method, we also applied it to simulated X-ray diffraction patterns of a mixture of two organic phases. Using simulated data instead of measured patterns enabled the introduction of controlled peak shifts and intensity variations, as well as separation of the sample preparation and measurement effects from additional inaccuracies, caused by differences of internal structure between the corresponding investigated and reference phases. The modeled sample consisted of two randomly chosen substances from the COD database with overlapping peaks: acetamidoxime (database entry 2013664; Olmstead & Sahbari, 2003) and lactosyl acetamide (entry 2012076; Lakshmanan *et al.*, 2001). During simulation of diffraction patterns we took into account resolution effects (each peak was modeled by a 0.14° wide Lorenzian distribution), background signal (25 counts on average, while the maximal diffraction peak was 10^4 counts high) and statistical noise (characterized by a Poisson distribution of counts). Shifts of the peak positions 2θ were modeled by assuming a nonzero vertical displacement Δh of the sample (Clearfield *et al.*, 2008):

$$\Delta(2\theta) = -\frac{2\Delta h \cos\theta}{R}, \qquad (33)$$

where R is the radius of the goniometer in Bragg–Brentano geometry. To model intensity variations, preferred orientation of crystallites, as described by the March–Dollase model

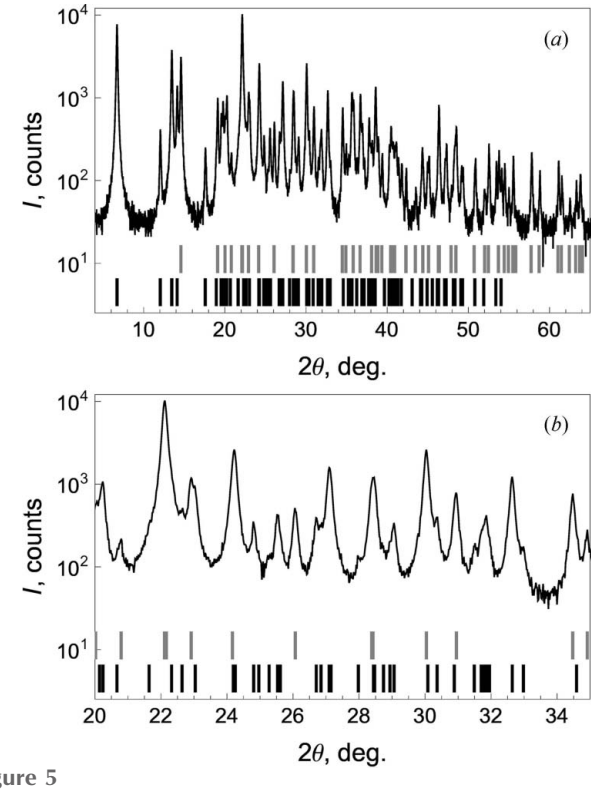


Figure 5
Simulated diffraction pattern for a mixture of acetamidoxime and lactosyl acetamide with equal weight fractions and preferred orientation along direction (100), characterized by March coefficient $r = 2.2$: the whole scan range (*a*) and an enlarged region with overlapping peaks (*b*) are shown. Stick plots below the diffraction patterns indicate peaks of the reference database entries (gray for acetamidoxime and black for lactosyl acetamide).

(March, 1932; Dollase, 1986), was assumed. According to the model, the intensity correction factor for a reflection depends on the angle α between the corresponding lattice plane normal and the preferred direction:

$$P = \left(r^2 \cos^2 \alpha + r^{-1} \sin^2 \alpha\right)^{-3/2}, \qquad (34)$$

where r is the dimensionless March coefficient (the value $r = 1$ corresponds to the absence of preferred orientation). An example of a simulated diffraction pattern is shown in Fig. 5.

The simulated whole X-ray diffraction patterns were transformed into stick patterns by using the second derivative peak search method (Pecharsky & Zavalij, 2009) and then used for testing qualitative and quantitative phase analysis techniques.

Fig. 6 shows the region of peak shifts and intensity variations for which the constructed search and match algorithm

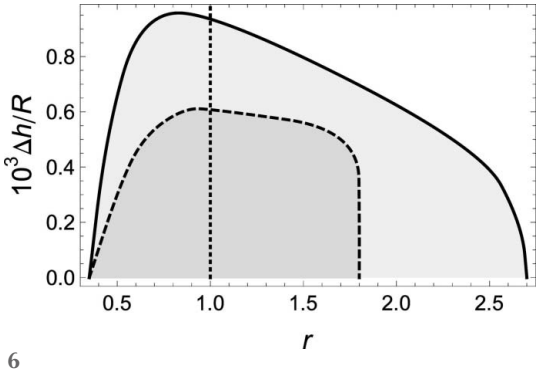


Figure 6
Region of robust phase identification for a mixture of acetamidoxime and lactosyl acetamide with weight fractions 50%:50% (light-gray region, bounded by solid line) and 80%:20% (darker-gray region, bounded by dashed line). The dimensionless parameter $\Delta h/R$ characterizes peak shifts, modeled by vertical displacement Δh of the sample. Preferred orientation in the (100) direction is characterized by the March coefficient r.

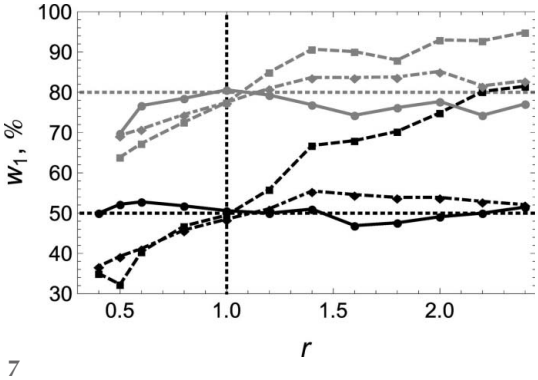


Figure 7
Results of quantitative analysis (weight fraction of acetamidoxime) for simulated diffraction patterns with different degrees of preferred orientation. The sample was modeled as mixture of acetamidoxime and lactosyl acetamide with weight fractions 50%:50% (black lines) and 80%:20% (gray lines). The weight fractions were estimated on the basis of equations (31) and (32) using scale factors obtained at the stage of phase identification (solid lines), single highest peaks of the phases (dashed lines) and groups of the three highest peaks for each phase (dot–dashed lines). Horizontal dotted lines indicate the weight fractions of acetamidoxime used for simulation of the patterns.

successfully identified the two phases (*i.e.* the iterative subtractive search found acetamidoxime and lactosyl acetamide as the best candidate phases at two subsequent iterations). The results of quantitative analysis for the samples containing 50 and 80% of acetamidoxime are shown in Fig. 7. For comparison, dashed lines show the weight fractions calculated from equation (31) by using single highest peaks. Dot–dashed lines correspond to RIR-based analysis, using the three strongest lines for each phase. The results obtained prove the usefulness of the scale factors, being in fact a by-product of phase identification, for robust quantitative analysis.

5.5. Relations with other methods

The generality of the introduced approach enables us to represent the expressions for calculation of the figure of merit from other phase identification methods in terms of Bayesian probabilities. The methods of FOM calculation based on summing 'fines' and 'bonuses' for separate matched (or not matched) peaks can be described similarly to the the model introduced in §4. The 'fines' for peak displacements and intensity differences correspond to $\log p_d(\delta)$ and $\log p_I(\varepsilon)$, respectively. The 'bonuses' for the interpretation of measured peaks are provided by the single-peak terms in $-\log P_{\rm DB}(J)$. For this parametrization, the quantity $\log \overline{p}(C_i \mid M)$ represents the sum of the 'fines' and 'bonuses' for all the peaks and, therefore, corresponds to the figure of merit. For example, the goodness of match (Faber *et al.*, 2004), $\mathrm{GOM} = 1000 \sum_i (1 - |\delta d_i|/\mathrm{SW})^2$, where δd_i is the difference of d spacings for the ith peak and SW is the search window, can be represented as a renormalized sum of logarithms of the prior probabilities of peak displacements: $\mathrm{GOM} = 1000 \times [\sum_i \log p(\delta d_i) + N]$, where $p(\delta d_i) = \exp[(1 - |\delta d_i|/\mathrm{SW})^2 - 1]$, $|\delta d_i| \leq \mathrm{SW}$ and N is the number of analyzed peaks.

Another example is applying the so-called normalized R index (Hofmann & Kuleshova, 2005; Faber & Blanton, 2008) as a measure of the similarity of intensities to the associated peaks:

$$R_{\rm s} = \sum_{i=1}^{N} \left| \frac{I_i^{\rm exp}}{\sum_{j=1}^{N} I_j^{\rm exp}} - \frac{I_i^{\rm DB}}{\sum_{j=1}^{N} I_j^{\rm DB}} \right|, \qquad (35)$$

where intensities $I_i^{\rm exp}$ and $I_i^{\rm DB}$ correspond to the measured and reference patterns' peaks, respectively. The R index can be represented as a sum of logarithms of prior probabilities of normalized peak intensities: $-R_{\rm s} = \sum_{i=1}^{N} \log p_I(\delta \overline{I}_i) = \sum_{i=1}^{N} \log[\exp(-|\delta \overline{I}_i|)]$, where $\delta \overline{I}_i = \overline{I}_i^{\rm exp} - \overline{I}_i^{\rm DB}$ is the difference of normalized intensities $\overline{I}_i^{\rm exp} = I_i^{\rm exp} / \sum_{j=1}^{N} I_j^{\rm exp}$ and $\overline{I}_i^{\rm DB} = I_i^{\rm DB} / \sum_{j=1}^{N} I_j^{\rm DB}$. The normalization can be included in the discussed Bayesian approach by setting the scale factor of the reference pattern to be equal to $\rho = \sum_{j=1}^{N} I_j^{\rm exp} / \sum_{j=1}^{N} I_j^{\rm DB}$.

Another important feature of the Bayesian approach, already mentioned above, is the ability to naturally incorporate any additional information about the composition of the investigated sample. This information can be taken into account by modifying the prior probabilities, as described by equation (8).

6. Conclusions

To summarize, a Bayesian approach to the problem of phase identification from X-ray powder diffraction patterns was proposed and proved using round robin data. The results obtained demonstrate that the expressions for ranking the candidate phases can be consistently derived from physical assumptions, instead of being guessed on the basis of experience in powder data analysis.

Any analysis technique has certain limits of its applicability and may require changes if inadequate results are obtained. The introduced Bayesian approach provides a clearer understanding of the underlying physics of the phase identification methods, enables tracing of all the assumptions used during the derivation of the final expressions and makes the process of finding the sources of encountered problems easier than in the case of traditional techniques for calculation of the figure of merit. The designed method provides a unified implementation of different strategies for multi-phase sample analysis and is suitable for the incorporation of additional information about the composition of the investigated sample, for example, X-ray fluorescence data.

The application of the reported Bayesian method has been illustrated by a model with factorized prior probabilities of the deviations of peak positions and intensities and an approximate statistical description of the database. The derived expressions for calculation of the figure of merit on the basis of this model have been tested on several IUCr round robin samples. Good performance of both subtractive and additive search strategies has been shown and a stopping criterion for the additive approach to the sample model construction has been suggested. The scale factors for the reference diffraction patterns of the identified phases, which are a by-product of the matching procedure, have been used for a quantitative analysis of the samples and showed sufficiently good agreement with the known values and robustness with respect to peak shifts and intensity variations.

The Bayesian approach provides a phase identification technique whose performance is at least comparable to that of other existing search and match methods. The universality of the introduced technique provides better adaptability to various types of investigated samples and an ability to use different search strategies and to combine powder diffraction methods with other analytical techniques.

References

Altomare, A., Corriero, N., Cuocci, C., Falcicchio, A., Moliterni, A. & Rizzi, R. (2015). *J. Appl. Cryst.* **48**, 598–603.
Altomare, A., Cuocci, C., Giacovazzo, C., Moliterni, A. & Rizzi, R. (2008). *J. Appl. Cryst.* **41**, 815–817.
Barr, G., Cunningham, G., Dong, W., Gilmore, C. J. & Kojima, T. (2009). *J. Appl. Cryst.* **42**, 706–714.
Barr, G., Gilmore, C. J. & Paisley, J. (2004). *J. Appl. Cryst.* **37**, 665–668.
Bergmann, J. & Monecke, T. (2011). *J. Appl. Cryst.* **44**, 13–16.
Bigelow, W. & Smith, J. V. (1965). *ASTM Spec. Tech. Publ.* **STP 372**, 54–89.
Caussin, P., Nusinovici, J. & Beard, D. W. (1988). *Adv. X-ray Anal.* **31**, 423–430.
Chung, F. H. (1974*a*). *J. Appl. Cryst.* **7**, 519–525.
Chung, F. H. (1974*b*). *J. Appl. Cryst.* **7**, 526–531.
Clearfield, A., Reibenspies, J. & Bhuvanesh, N. (2008). *Principles and Applications of Powder Diffraction.* Chichester: Wiley-Blackwell.
David, W. I. F. & Sivia, D. S. (2001). *J. Appl. Cryst.* **34**, 318–324.
Dinnebier, R. E. & Billinge, S. J. L. (2008). *Powder Diffraction: Theory and Practice.* Cambridge: RSC Publishing.
Dollase, W. A. (1986). *J. Appl. Cryst.* **19**, 267–272.
Faber, J. & Blanton, J. (2008). *Adv. X-ray Anal.* **51**, 183–189.
Faber, J., Weth, C. & Bridge, J. (2004). *Adv. X-ray Anal.* **47**, 166–173.
Frevel, L. K. (1965). *Anal. Chem.* **37**, 471–482.
Gilmore, C. J. (1996). *Acta Cryst.* A**52**, 561–589.
Gilmore, C. J., Barr, G. & Paisley, J. (2004). *J. Appl. Cryst.* **37**, 231–242.
Grazulis, S., Daskevic, A., Merkys, A., Chateigner, D., Lutterotti, L., Quiros, M., Serebryanaya, N. R., Moeck, P., Downs, R. T. & Le Bail, A. (2012). *Nucleic Acids Res.* **40**(D1), D420–D427.
Hanawalt, J. D. & Rinn, H. W. (1986). *Powder Diffr.* **1**, 2–6.
Hanawalt, J. D., Rinn, H. W. & Frevel, L. K. (1938). *Ind. Eng. Chem. Anal. Ed* **10**, 457–512.
Hofmann, D. W. M. & Kuleshova, L. (2005). *J. Appl. Cryst.* **38**, 861–866.
Hubbard, C. R., Evans, E. H. & Smith, D. K. (1976). *J. Appl. Cryst.* **9**, 169–174.
Hubbard, C. R. & Snyder, R. L. (1988). *Powder Diffr.* **3**, 74–77.
Jaynes, E. T. (2003). *Probability Theory: The Logic of Science.* Cambridge University Press.
Jenkins, R. (1994). *Adv. X-ray Anal.* **37**, 117–121.
Jenkins, R. & Snyder, R. L. (1996). *Introduction to X-ray Powder Diffractometry.* New York: John Wiley and Sons.
Lakshmanan, T., Sriram, D. & Loganathan, D. (2001). *Acta Cryst.* C**57**, 825–826.
Langford, J. I. & Louër, D. (1996). *Rep. Prog. Phys.* **59**, 131–234.
Madsen, I. C., Scarlett, N. V. Y., Cranswick, L. M. D. & Lwin, T. (2001). *J. Appl. Cryst.* **34**, 409–426.
March, A. (1932). *Z. Kristallogr. Cryst. Mater.* **81**, 285–297.
Marks, L. D., Sinkler, W. & Landree, E. (1999). *Acta Cryst.* A**55**, 601–612.
Mikhalychev, A., Mogilevtsev, D., Teo, Y. S., Řeháček, J. & Hradil, Z. (2015). *Phys. Rev. A*, **92**, 052106.
Mittemeijer, E. J. & Scardi, P. (2004). *Diffraction Analysis of the Microstructure of Materials*, Springer Series in Materials Science, Vol. 68. Berlin, Heidelberg: Springer.
Nichols, M. C. (1966). Report UCRL-70078. Lawrence Livermore Laboratory, Livermore, California, USA.
Nusinovici, J. & Bertelmann, D. (1993). *Adv. X-ray Anal.* **36**, 327–332.
Nusinovici, J. & Winter, M. J. (1994). *Adv. X-ray Anal.* **37**, 59–66.
Olmstead, M. M. & Sahbari, J. J. (2003). *Acta Cryst.* C**59**, o719–o720.
Pecharsky, V. K. & Zavalij, P. Y. (2009). *Fundamentals of Powder Diffraction and Structural Characterization of Materials*. New York: Springer.
Scarlett, N. V. Y., Madsen, I. C., Cranswick, L. M. D., Lwin, T., Groleau, E., Stephenson, G., Aylmore, M. & Agron-Olshina, N. (2002). *J. Appl. Cryst.* **35**, 383–400.
Schreiner, W. N., Surdukowski, C. & Jenkins, R. (1982). *J. Appl. Cryst.* **15**, 513–523.
Sivia, D. S. & David, W. I. F. (2001). *J. Phys. Chem. Solids*, **62**, 2119–2127.
Snyder, R. L. (1981). *Adv. X-ray Anal.* **24**, 83–90.
Toby, B. H. (2005). *J. Appl. Cryst.* **38**, 1040–1041.
Waldo, A. W. (1935). *Am. Mineral.* **20**, 575.
Winchell, A. N. (1927). *Am. Mineral.* **12**, 261.