JOURNAL OF APPLIED CRYSTALLOGRAPHY

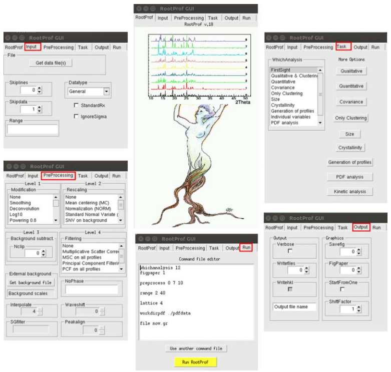
ISSN 1600-5767

Received 24 April 2023
Accepted 22 September 2023

Edited by H. Brand, Australian Synchrotron, ANSTO, Australia

Keywords: multivariate analysis; data analysis; X-ray powder diffraction; pair distribution functions; *in situ* experiments.

Supporting information: this article has supporting information at journals.iucr.org/j



```
whichanalysis 12
figpaper 1

preprocess 0 7 10

range 2 40

lattice 4

workdirpdf ./pdfdata

file nov.gr
```

New features of the *RootProf* program for model-free analysis of unidimensional profiles

Annamaria Mazzone,[a] Mattia Lopresti,[b] Benny Danilo Belviso[a] and Rocco Caliandro[a]*

[a]Institute of Crystallography, National Research Council of Italy, via Amendola 122/o, Bari 70126, Italy, and [b]Università degli Studi del Piemonte Orientale 'Amedeo Avogadro', Via Michel 11, Alessandria 15100, Italy. *Correspondence e-mail: rocco.caliandro@ic.cnr.it

The *RootProf* computer program applies multivariate model-free analysis to crystallographic data and to any *x*, *y* experimental data in general. It has been enhanced with several new features, including a graphical user interface, multithreading implementation and additional pre-processing options. The program also includes novel qualitative analysis methods, such as semiquantitative estimates derived from principal component analysis (PCA) and restrained PCA to extract the diffraction signal from active atoms. Additional quantitative analysis methods have been included, involving the combination of different datasets or the application of the standard addition method as well as tools for crystallinity analysis, kinetic analysis and extraction of free crystal cell parameters from a pair distribution function profile. The ROOT data analysis framework supports the program and can be installed on the current major platforms such as Windows, Linux and Mac OSX with detailed user documentation included. Applications of the new developments are presented and discussed in the paper, and related command files are provided as supporting information.

1. Introduction

The software package *RootProf* was launched in 2014 (Caliandro & Belviso, 2014) as a multi-purpose tool for processing unidimensional profiles. It is built upon the open-source data analysis framework ROOT (Brun & Rademakers, 1997), which was originally created at CERN as a tool for imaging and processing the huge quantities of data collected in high-energy physics experiments. ROOT offers a highly efficient data structure, based on the TTree class (https://root.cern.ch/doc/master/classTTree.html), enabling rapid access to massive data volumes (orders of magnitude faster than accessing a normal file). In addition, ROOT includes powerful mathematical and statistical tools, available as C++ applications, capable of interacting with data through parallel processing. Results can be displayed with histograms, scatter plots and three-dimensional plots, easily adjusted in real time by a few mouse clicks. High-quality figures can be saved as TIFF, PNG or other formats. ROOT can be run interactively or executed in batch mode at full speed.

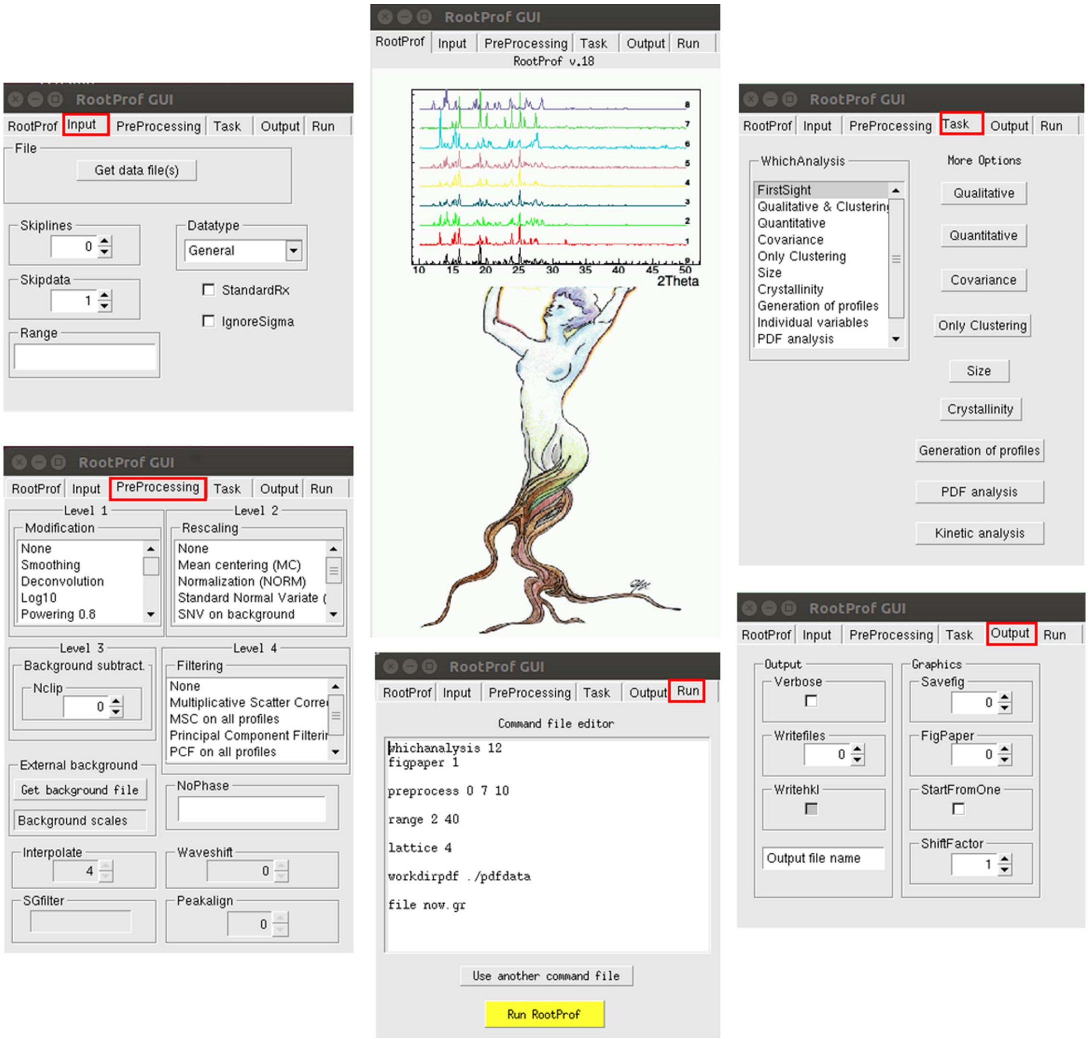
Feedback from users has highlighted several strengths of the program, including (i) multi-purpose processing that supports heterogeneous data, also collected through different experimental techniques, and the possibility to combine them; (ii) access to a plethora of multivariate techniques, through the ready-to-use ROOT class named TSpectrum (https://root.cern.ch/doc/master/classTSpectrum.html); (iii) advanced and interactive graphics tools that support visualization of the different processing steps and the final results; and (iv) the ability to develop new code as C++ scripts interpreted by ROOT. *RootProf* requires the pre-installation of ROOT on the user's local machine, which can be a significant drawback, especially when installing on the Windows operating system.

RootProf has been used in various applications spanning from crystallography to spectroscopy, many of which concern the use of principal component analysis (PCA) to extract trends in data (Olds *et al.*, 2017; Chernyshov *et al.*, 2020; Brennhagen *et al.*, 2021) or covariance analysis to combine data of different types, *i.e.* X-ray data with optical spectroscopy (Caliandro *et al.*, 2019), acoustic spectroscopy (Massara *et al.*, 2021) or differential scanning imaging (Lopresti *et al.*, 2023). Comparative analyses have been performed on profiles representing individual protein structural models (Liuzzi *et al.*, 2017; Miciaccia *et al.*, 2021), and specific structural determinants calculated from crystal structures (Belviso *et al.*, 2016) or monitored by molecular dynamics (Bolognino *et al.*, 2022). Faint signals from subtle structural changes originating from oxidation (Caliandro *et al.*, 2016) or induced by light (Colella *et al.*, 2018) have been highlighted for nanocrystals monitored *in situ* by synchrotron X-ray powder diffraction (PXRD) experiments. On the other hand, multivariate analysis has become increasingly popular in the treatment of crystallographic data (Guccione *et al.*, 2021), driving the demand for computational tools specifically designed to process this type of data.

In the following sections, we describe the key advances in the program features that were implemented to address specific user needs. These include the addition of a friendly graphical user interface (GUI), maximizing the information extracted from unidimensional profiles, adapting multivariate analysis to crystallographic data through the introduction of appropriate restraints, combining crystallographic information in real and reciprocal space, and performing kinetic analysis on time-resolved data. All the figures reported here were produced using *RootProf*.

2. Graphical user interface

RootProf operates through an input file, where the users can input specific commands, as detailed in the user guide on the website (https://users.ba.cnr.it//ic/crisrc25/RootProf/RootProf_help.html). To create the command file, users can utilize the GUI (Fig. 1) developed using the ROOT GUI classes,



```
whichanalysis 12
figpaper 1

preprocess 0 7 10

range 2 40

lattice 4

workdirpdf ./pdfdata

file now.gr
```

Figure 1
Main windows of the *RootProf* GUI.

inclusive of an extensive and rich set of widgets with a Windows-like look and feel. Once the ROOT interactive session is initiated, users can call *RootProf* by either executing a pre-compiled command file or accessing the *RootProf* GUI for creating the command file and running the program. Tooltip windows appear in the GUI when the mouse pointer is positioned over a number of boxes, providing users with a quick explanation of the command.

The GUI consists of different components, including an input manager, designated to data files uploads; a preprocessing manager, for specifying the data pre-processing; a task manager, for selecting the type of analysis to be performed; an output manager, for inserting options related to the files produced after execution; and a run manager, which checks the command file and runs the program. Additional windows can be opened via the task manager, for specifying commands related to each type of analysis. A detailed description of the *RootProf* GUI can be found at https://users.ba.cnr.it//ic/crisrc25/RootProf/GUIPage.htm.

3. New implementations

3.1. Pre-processing

Pre-processing dedicated to unidimensional profiles is a unique strength of *RootProf*. Since its original implementation, we realized that optimizing the input data could impact the results of subsequent multivariate analysis significantly. Several pre-processing options have been developed to enhance signal over background, minimizing unwanted differences between profiles and eliminating features that may hinder information extraction. New pre-processing options include a cubic spline interpolation, which allows users to adjust the sampling points of profiles, and the conversion of a PXRD profile measured at an arbitrary energy of the primary beam to that measured using Cu $K\alpha$ radiation, the latter being typically used for comparing the characteristic peaks of the profile. Two other important options, described below, complement the existing *RootProf* suite of pre-processing tools, aimed at maximizing data quality prior to further analysis.

3.1.1. Savitzky–Golay filter. Savitzky–Golay filtering (Savitzky & Golay, 1964) involves a convolution operation that fits a subset of points (w) from the equidistant raw data with a polynomial of degree p. The central point of the subset, which must be odd, is used to calculate the new value of the ordinate y'_j and the window moves to the next point for the next iteration. The polynomial fitting has well defined coefficients based on the size of the window (w) and the degree of the polynomial (p) (Guest, 2012). This convolution procedure is known as a moving average because the central point is estimated by weighting the surrounding points for the convolution coefficients. By selecting an appropriate window of points w, the smoothed central point y'_j of the window can be treated with convolution coefficients C_i according to

$$y'_j = \sum_{i=(1-w)/2}^{(w-1)/2} C_i \, y_{j+i}, \qquad (w+1)/2 \le j \le n-(w-1)/2. \quad (1)$$

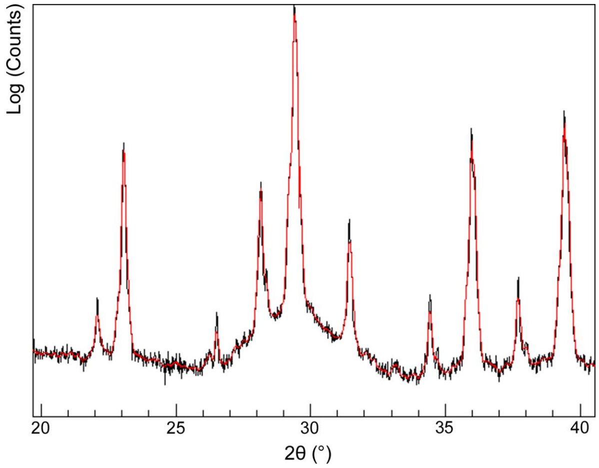


Figure 2
Example application of Savitzky–Golay filtering of the PXRD profile of calcite. The profile before (black line) and after (red line) application of the filtering with the following parameters: order of derivation = 0, polynomial degree = 2, window = 9.

The result of this operation can be observed in Fig. 2, where the red curve is obtained using the Savitzky–Golay smoothing on the points of the black curve with the coefficients described in equation (1). An additional parameter m is included in the algorithm to establish the order of derivation of the polynomial fitting function, enabling the user to calculate the numerical derivative of order m of the smoothed spectrum while smoothing the data profile. The Savitzky–Golay filtering can be included in the software in two ways: (i) by creating tables with the convolution parameters for each combination of m, p and w; or (ii) by calculating the fitting polynomial on the points at each step of the iteration starting from zero and using ordinary least squares in the matrix form:

$$\mathbf{a} = (\mathbf{M}^{\rm T}\mathbf{M})^{-1}\mathbf{M}^{\rm T}\mathbf{y}, \quad (2)$$

where $\mathbf{a}$ is the vector containing the coefficients C_i of the interpolation as they would be obtained from equation (1), $\mathbf{M}$ is a Vandermonde matrix suited for building the model of the desired p degree and $\mathbf{y}$ is the vector of length w containing the original intensities to be interpolated. The latter method involves a slow step such as the matrix inversion, but the ROOT framework fitting procedure is very fast and efficient, and enables polynomial coefficient refinement and the fitting of the points of the window w without using the matrix notation of equation (2) relying on the Cholesky decomposition method (Benoit, 1924) included in the *MINUIT* minimizer. The polynomial is stored by means of the ROOT `TF1` class (https://root.cern.ch/doc/master/classTF1.html) and its coefficients are available for further analysis, such as derivation.

3.1.2. Peak alignment. During *in situ* PXRD experiments, external stimuli such as temperature or pressure variations could result in unit-cell expansion or shrinkage, thus altering the cell parameters. However, these changes may not be the focus of investigation, which may instead be directed towards structural rearrangements due to ion adsorption or the onset

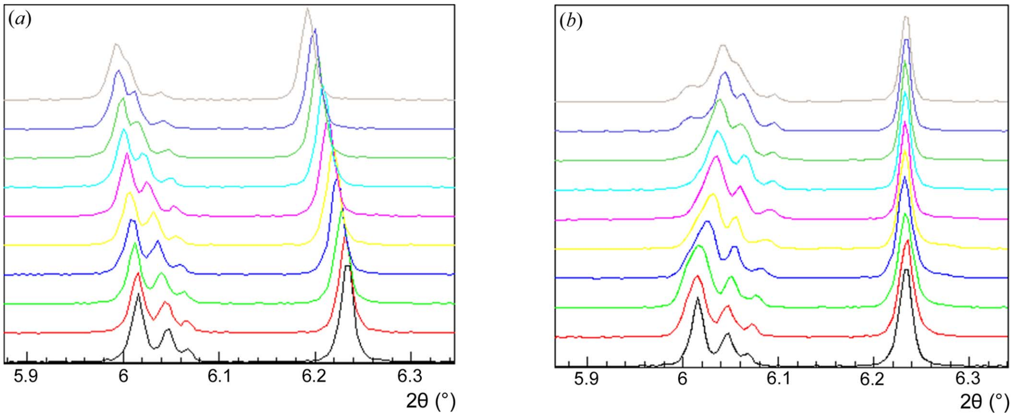


Figure 3
PXRD patterns collected during a temperature-dependent *in situ* experiment (*a*) before and (*b*) after application of the peak-alignment procedure consisting of peak correspondence followed by harmonic time stretch, zoomed-in on a restricted 2θ interval.

of new interactions. In this context, peak shifts resulting from crystal cell variations could hide interesting data trends in the diffraction signal related to the structural changes. Transforming diffraction patterns so that these peak shifts are reduced is a challenging task that involves stretching or deforming the independent variable (2θ) axis in each profile. Two different approaches were developed by Guccione *et al.* (2018*b*) as MATLAB scripts to solve the problem.

The first approach, called peak correspondence, entails detecting the peaks in various profiles and tracing of the corresponding peaks through the profiles. The shifts among corresponding peaks are then used to estimate the optimal 2θ stretch model by means of a spline interpolation.

The second approach, 'harmonic time stretch', is inspired by dynamic time warping (Rabiner *et al.*, 1978). This approach assumes that any change in the 2θ axis can be represented by a generic function. To extend the 2θ stretch model, the function is allowed to be composed of a superposition of harmonic functions, whose number, amplitude and phase are unknown. To determine the optimal unknown values, a cost function is minimized, which is the correlation coefficient of all the profiles with a reference profile that coincides with the first uploaded file.

The two approaches have been implemented in *RootProf* and can be used separately or in sequence. The efficiency of the peak-alignment procedure is evaluated with a figure of merit (FOM) that measures the degree of peak misalignment. This is defined as the average of the absolute difference between the peak position and the reference position, weighted by the peak height. This evaluation is carried out for all selected peaks and profiles.

Fig. 3 depicts the effect of the peak-alignment procedure on PXRD data for the first steps of a solid-state non-isothermal reaction between tetracyanoquinodimethane and fluorine which leads to co-crystal formation in the final state (Guccione *et al.*, 2018*b*).

3.2. Qualitative analysis

The qualitative analysis session of *RootProf* is based on PCA. It has proven to be a very versatile tool, capable of providing a great amount of information regarding the trends in data, identifying outlier samples and clustering of experimental data. PCA results include scores, *i.e.* the coordinates representing individual profiles and loadings (*i.e.* the coordinates representing individual variables of the profiles). Both scores and loadings are determined for each of the principal components (PCs) into which the dataset has been decomposed. Thus, individual input profiles can be compared using scores, while the contribution of given peaks to the profile separation can be assessed using loadings. Important extensions of such capabilities are described in the following.

3.2.1. Semi-quantitative analysis by PCA. A fast semi-quantitative analysis that can be applied to PCA results has been implemented to directly quantify components in mixtures. It is based on the barycentric coordinate system, which is a coordinate system defined within a simplex, typically useful when describing mixtures (Mobius, 1827). Each point inside the simplex can be described by a set of q coordinates that is equal to the order of the simplex minus one. The sum of all the coordinates of a point is always equal to 1, *i.e.* $\sum_{i=1}^{q} x_i = 1$. The barycentric system is closely linked to the Euclidean coordinate system, and points can be transformed from one system to the other relatively easily. The book *Experiments with Mixtures* (Cornell, 2011) describes a convenient method for projecting the points belonging to a q-dimensional simplex-centroid design in a $q - 1$ independent coordinate system. The inverse problem, which involves passing from a set of points represented by Cartesian coordinates to the space of a simplex, can be a relatively simple operation if the set is arranged optimally in space. However, if the group of points does not correspond to a simplex projected in Euclidean space, with the right orientation along the bisector of the axes, the problem can be more complex and can be addressed using PCA. The orthonormal basis of the PC space guarantees that the Euclidean metric can be used to transform simplexes that may appear among the scores or between the loadings to a barycentric system for the estimation of the relative distances between one point and the others.

Code has been developed for the case q = 3 and has been successfully applied to the results reported by Lopresti *et al.*

(2022). The algorithm uses the $q - 1$ generated scores as coordinates for the points, as shown in Fig. 4. The scores are all range-scaled in [0, 1], and one vertex, the nearest one, is centred at (0; 0). The q vertices of the simplex are identified, and the $q - 1$ PC scores are range-scaled for the proper interval indicated by Cornell (2011) as $a + b$. An iterative loop then searches for the correct orientation of the simplex in space by multiplying the $q - 1$ scores matrix by a rotation matrix. The loop refines the rotation angle by a dichotomic search until the difference between the coordinates of the vertices reaches the stop condition, which is met when the difference between the coordinates of the vertices is lower than 10^{-10}. Finally, the estimated weight fractions are calculated by multiplying the PC scores, in the form of a matrix, by the inverse of the $\mathbf{A}$ matrix, defined as

$$\mathbf{A} = \mathbf{I} + \frac{1}{1 + \sqrt{q}}\mathbf{J}, \tag{3}$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{J}$ is a $(q - 1)$-dimensional square matrix with all elements equal to 1. The results of this operation are reported in Fig. 4.

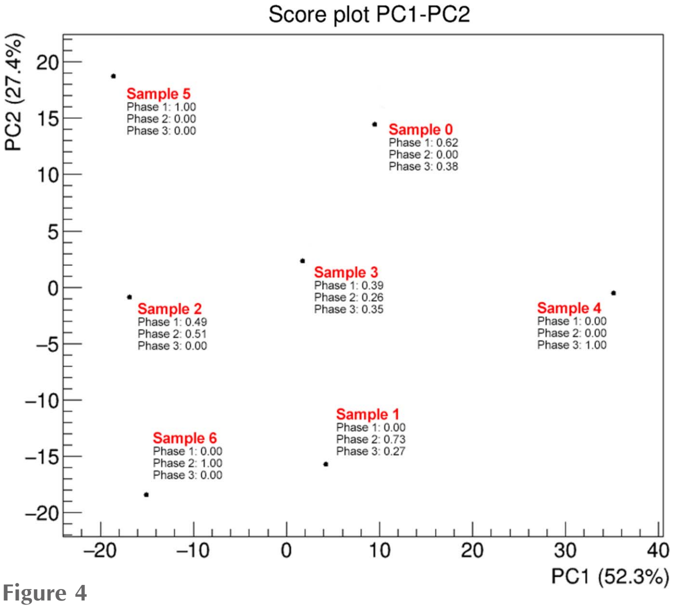


Figure 4
Score plot of the first two PCs for ternary mixtures composed as described by Lopresti *et al.* (2022). Weight fraction estimations for the three crystal phases obtained by semi-quantitative analysis are reported close to each representative point. The percentages of the total data variance explained by the first two PCs (PC1 and PC2) are reported on the axes.

3.2.2. Optimal constrained component rotation. PCA is a data-reduction method used to extract general trends in large datasets. When applied to PXRD profiles collected from *in situ* experiments, it allows the user to describe the whole data matrix in terms of a few PCs. The score values of these PCs capture the trends over time, while the loading values indicate the contribution of each individual peak in the profile to the particular component (Guccione *et al.*, 2021). An example is given in Fig. 5, where a data matrix composed of 25 PXRD profiles [Fig. 5(*a*)] is described using three PCs, accounting for 99.9% of the total data variance. Their scores [Fig. 5(*b*)] define the main trends in the data, which are distinct among the

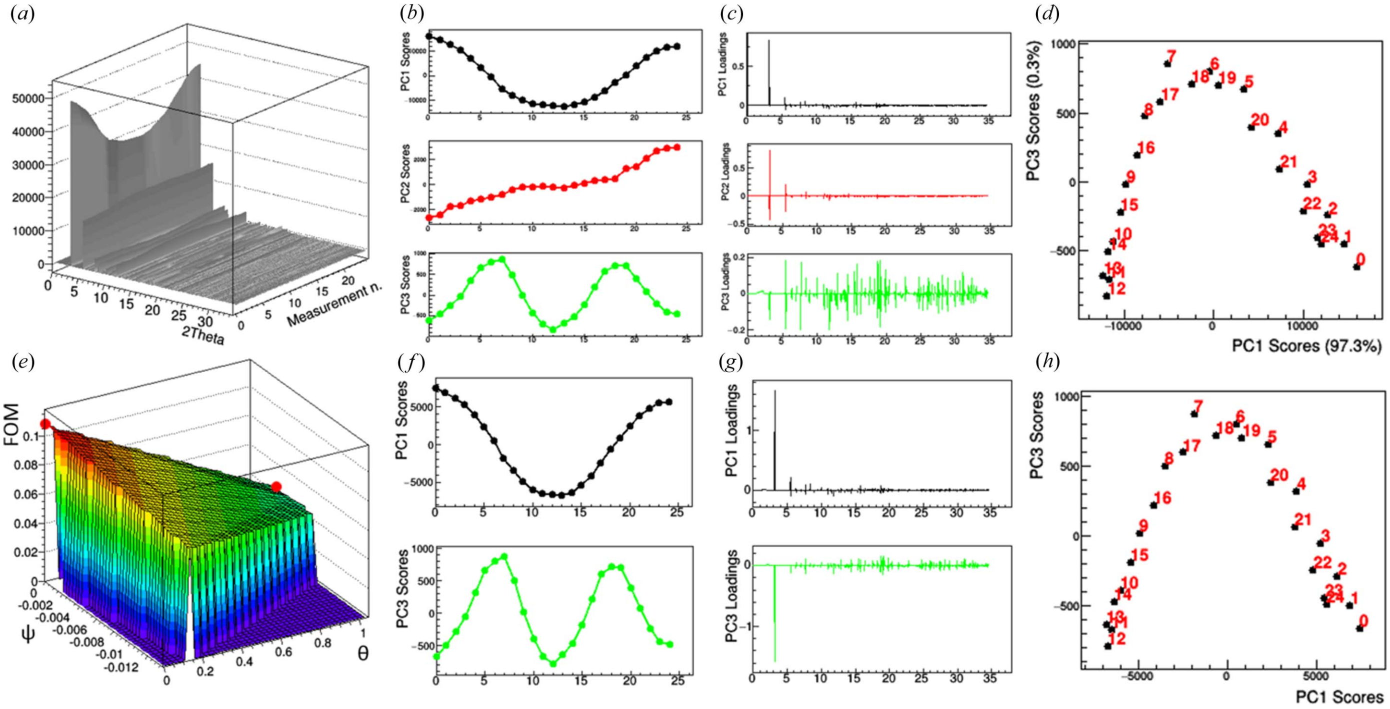


Figure 5
Steps involved in the OCCR applied to *in situ* PXRD measurements. The data matrix formed by the collected PXRD profiles (*a*) is processed by PCA, obtaining (*b*) the scores and (*c*) the loadings of the first three PCs. (*d*) Scatter plot of the scores of the first and third PCs, with the percentages of the total data variance explained by the two PCs reported on the axes. (*e*) FOM explored by the OCCR procedure, as a function of the θ and ψ rotation angles, where the initial and final rotations are indicated by red dots. (*f*) Scores and (*g*) loadings plots of the first and third PCs, recalculated by applying the best rotation found by OCCR. (*h*) Scatter plot of the recalculated values of the scores of the first and third PCs.

different components (since they are orthogonal). The scores of the first PC (PC1, which accounts for 97.3% of the total data variance) exhibit a symmetric behaviour that mainly depends on the intensity variations of the basal peak at $2\theta = 3.1°$, and in fact the same peak is prominent in the PC1 loadings [Fig. 5(*c*)]. On the other hand, the scores of the second PC (PC2), which describes 2.3% of the total data variance, show a monotonic trend. By analysing the corresponding loadings, we can deduce that this is due to a systematic shift of the peak position during the *in situ* experiment. In fact, the PC2 loadings have non-zero values corresponding to the 2θ values of the main PXRD peaks, and for each peak, they exhibit a negative peak slightly shifted at lower 2θ values followed by a positive peak slightly shifted at higher 2θ values. This typical shape arises from the convolution between two peaks translating with respect to each other. The third PC (PC3), which only accounts for 0.3% of the total data variance, could be considered noise given the loading values, which have similar fluctuating contributions from all the peaks of the PXRD profiles, regardless of their height. However, the PC3 scores exhibit a well defined trend, which is approximately the square of the PC1 scores. Correlating PC1 and PC3 scores produce a nice parabola [Fig. 5(*d*)], although its axis is not perfectly parallel to the *y* axis. A theory developed by Chernyshov *et al.* (2011) can interpret these findings in terms of kinematic diffraction arising from the active atoms, *i.e.* atoms varying in phase with the stimulus received in the course of the *in situ* experiment, while the bulk atoms in the crystal sample remain silent. The signature for this phenomenon is a relationship between two trends in the data, where one is squared with respect to the other. According to this theory, called modulation enhanced diffraction (MED), the diffraction signal related to the squared component should only have contributions from active atoms (Caliandro *et al.*, 2012). Following these arguments, a procedure to constrain PCA on the basis of the MED theory was developed, where two orthogonal PCs previously identified are individually rotated to maximize the FOM ($\text{FOM}_{\text{scores}}$) defined by the Pearson correlation coefficient between the scores of the second component and the square of the scores of the first (Caliandro *et al.*, 2015). This algorithm was named optimal constrained component rotation (OCCR) and its application is illustrated in Fig. 5(*e*). Once the standard PCA protocol is completed, by activating the MED theory option, *RootProf* searches for the pair of components related by the quadratic relationship in the scores. In the present case study, PC1 and PC3 are rotated by the angles θ and ψ identified by OCCR. The new scores and loading values are shown in Figs. 5(*f*) and 5(*g*), respectively. The correlation between the rotated values of the PC1 and PC3 scores produces a parabola with the axis perfectly aligned with the *y* axis [Fig. 5(*h*)]. Notably, a dominant peak at $2\theta = 3.1°$ appears in the rotated PC3 loadings, which is the supposed diffraction

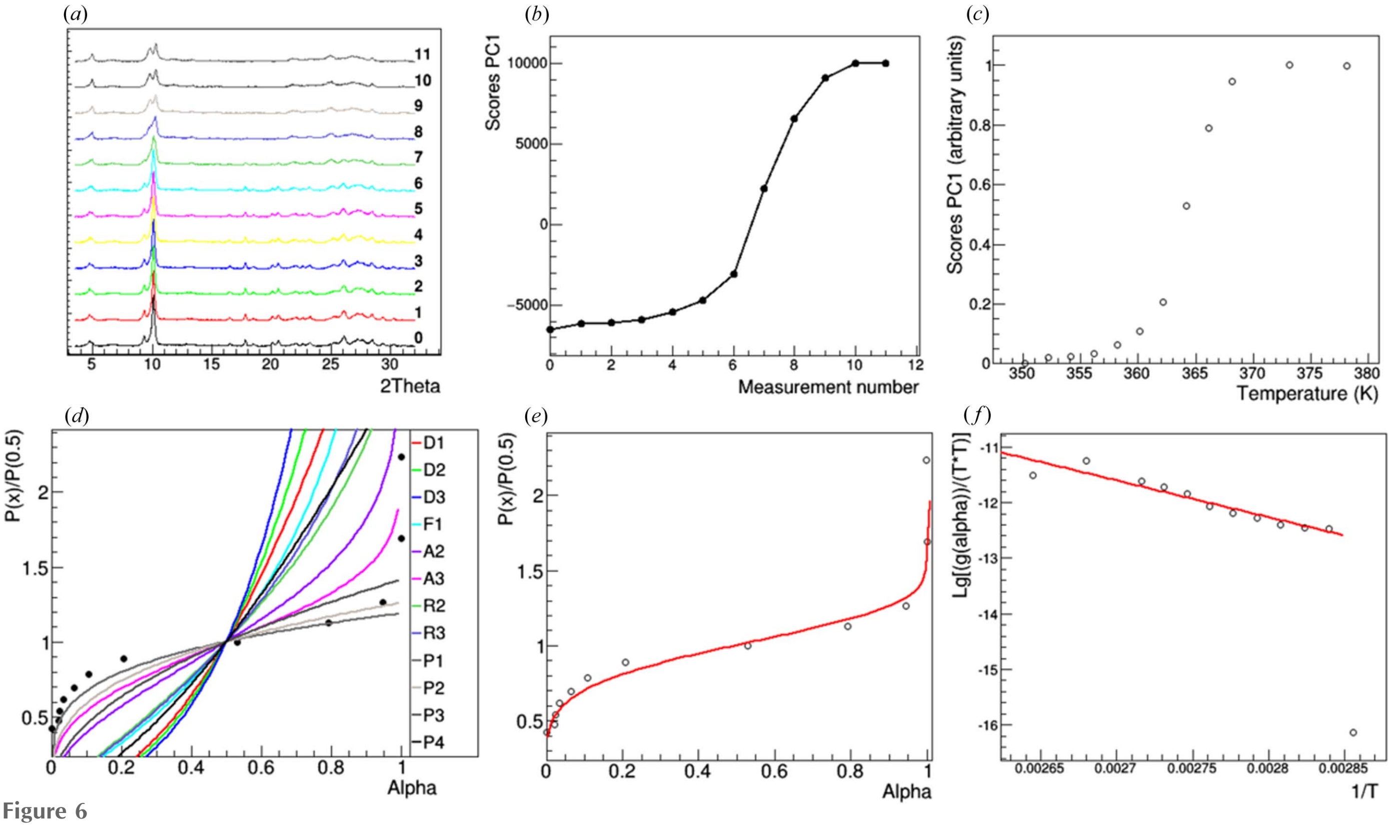


Figure 6

Steps involved in the kinetic analysis of *in situ*/*operando* measurements. (*a*) The data matrix formed by the collected PXRD profiles is processed by PCA, (*b*) obtaining the PC1 scores. (*c*) Scores are rescaled to range between 0 and 1 and are plotted as a function of the temperature on the sample. (*d*) A masterplot is generated to compare experimental data with a set of kinetic models and (*e*) the reaction order parameter for the chosen model is determined by fitting with experimental data. (*f*) The Coats & Redfern (1964) plot allows us to determine the activation energy and frequency factor parameters by fitting with experimental data.

pattern related to active atoms (note that a sign ambiguity is present in the PCA loadings). At the end of the MED procedure implemented in *RootProf*, two figures of merit are calculated: $\mathrm{FOM}_{\mathrm{scores}}$ and the normalized difference between the positive and the negative parts of the area underlying the second loading [denoted by $\mathrm{FOM}_{\mathrm{loadings}}$ by Caliandro *et al.* (2015)]. Their values go from 0.89 and 0.08 before OCCR to 0.96 and 0.25 after OCCR, respectively (both figures of merit have a maximum value of 1, which indicates the best performance). The loadings of the rotated second component can be written in an output file and used as input for the crystal structure solution of the substructure, which is composed of the active atoms, as was demonstrated by Palin *et al.* (2015).

3.2.3. Kinetic analysis. A common issue encountered when processing data collected from *in situ*/*operando* experiments is identifying the reaction coordinate, *i.e.* a one-dimensional coordinate that signifies the advancement of the sample on a reaction pathway. One approach to resolving this issue is through PCA, where the scores of the first PC can closely approximate the reaction coordinate as it represents the main trend in the data. *RootProf* offers a solution for visualizing and writing scores of PCs in output files using the `writescore` command. The scores are sorted on the basis of the total data variance they explain. Once the user has assessed that the scores of a specific PC (usually the one with the highest eigenvalue, referred to as PC1) can be used to represent the reaction coordinate of the studied sample, such scores are stored in the output file (named ScoresPC1) and *RootProf* can be re-run to perform the kinetic analysis. To better visualize the steps involved in this process, we show in Fig. 6 the most relevant *RootProf* plots resulting from a kinetic analysis applied to PXRD profiles collected by a laboratory diffractometer equipped with a system for measurements at a controlled temperature on a Pd-based vapochromic compound [see Belviso *et al.* (2021) for further details, referring to the compound Pd(ppy)]. The set of PXRD patterns collected at different temperatures [Fig. 6(*a*)] is subjected to PCA, which determines the PC1 scores. In this case, the first PC explains 96% of the total data variance [Fig. 6(*b*)]. These are rescaled in the range [0, 1] and plotted against sample temperature [Fig. 6(*c*)]. The masterplot technique (Goiss, 1963) is used to choose the best kinetic model that most accurately describes the experimental data [Fig. 6(*d*)]. A least-squares procedure [Fig. 6(*e*)] is applied to fine-tune the best kinetic model, which in this case is A3 (*i.e.* the Avrami–Eroféev model), and determine the order of reaction. Finally, the kinetic parameters are estimated from the plot reported by Coats & Redfern (1964) [Fig. 6(*f*)]. For more information on the kinetic analysis, refer to Guccione *et al.* (2018*a*). The graphics in Figs. 6(*c*)–6(*f*) are generated by the second run of *RootProf* using the command file reported in Section S5 of the supporting information.

3.3. Quantitative analysis

The quantitative analysis tool is another key feature of *RootProf*. The input profiles collected from mixtures are evaluated in terms of a linear superposition of reference profiles, *i.e.* containing individual components of the mixture. Multicomponent analysis is performed using the least-squares method, either by sequential fitting of each input profile or by applying an unfolding procedure that processes the input data matrix as a whole and decomposes it into a specific set of component profiles (Jandel *et al.*, 2004). Although the unfolding procedure is much faster, it is often less precise. Neither method requires any specific structural model and either can be applied to various sources of data. In the latest updates, the least-squares procedure called MultiFit by Caliandro & Belviso (2014) has been sped up by introducing the possibility to employ multiple cores (Section 3.3.1), in addition to integrating with additional information (Section 3.3.2) or using different datasets (Section 3.3.3). The unfolding procedure is instead used for pair distribution function (PDF) calculations (Section 3.5).

3.3.1. Multithreading. Parallelization refers to the process of breaking down software operations into smaller tasks and distributing them across multiple processors or cores. This can be achieved either implicitly through multithreading or explicitly by dividing the tasks between processors. A recent study by Piparo *et al.* (2017) examined the differences between these two approaches to parallelization within the context of the ROOT framework and showed how the two methods can be complementary. The main goal of parallelization is to improve the performance of computer systems and reduce the execution time of an algorithm. The *RootProf* parallelization implementation currently utilizes implicit parallelism, and it is not yet optimized for computer clusters. However, the computational speed has improved significantly on the host computer architecture. This is evident in the quantitative analysis module which achieved a performance boost of almost a tenth of the previously required time on a machine powered by an AMD Ryzen 5 3600X processor with 6 cores, where each core can handle 12 threads.

3.3.2. Standard addition method. Quantitative analysis of the weight amount of an individual crystal phase in a complex mixture of unknown composition can be challenging. In the work of Zappi *et al.* (2019), the quantification of the active pharmaceutical compound (API) paracetamol (polymorphic form I) in a solid formulation of the commercial drug Tachifludec was considered as a case study. In the first instance, the commercial sample was measured by PXRD in three replicates and the known crystal phase was analysed using the MultiFit procedure of *RootProf*, as shown in Fig. 7(*a*) (samples No. 0, 1, 2 and inset on the left). The weight fraction averaged on the three replicates was 0.137 ± 0.012, which provided a preliminary estimate of the amount of paracetamol present in the commercial drug. To obtain a more accurate measurement, the standard addition method (SAM) was employed, whereby Tachifludec samples with known amounts of the API were measured. MultiFit was applied to these additional profiles, and the weight fraction estimates are shown in Fig. 7(*a*) (samples 3–11, inset on the right). This information was then used in combination with the known additions to produce the calibration plot shown in Fig. 7(*b*). The SAM determination of the API concentration was found to be 0.141 ± 0.010 by

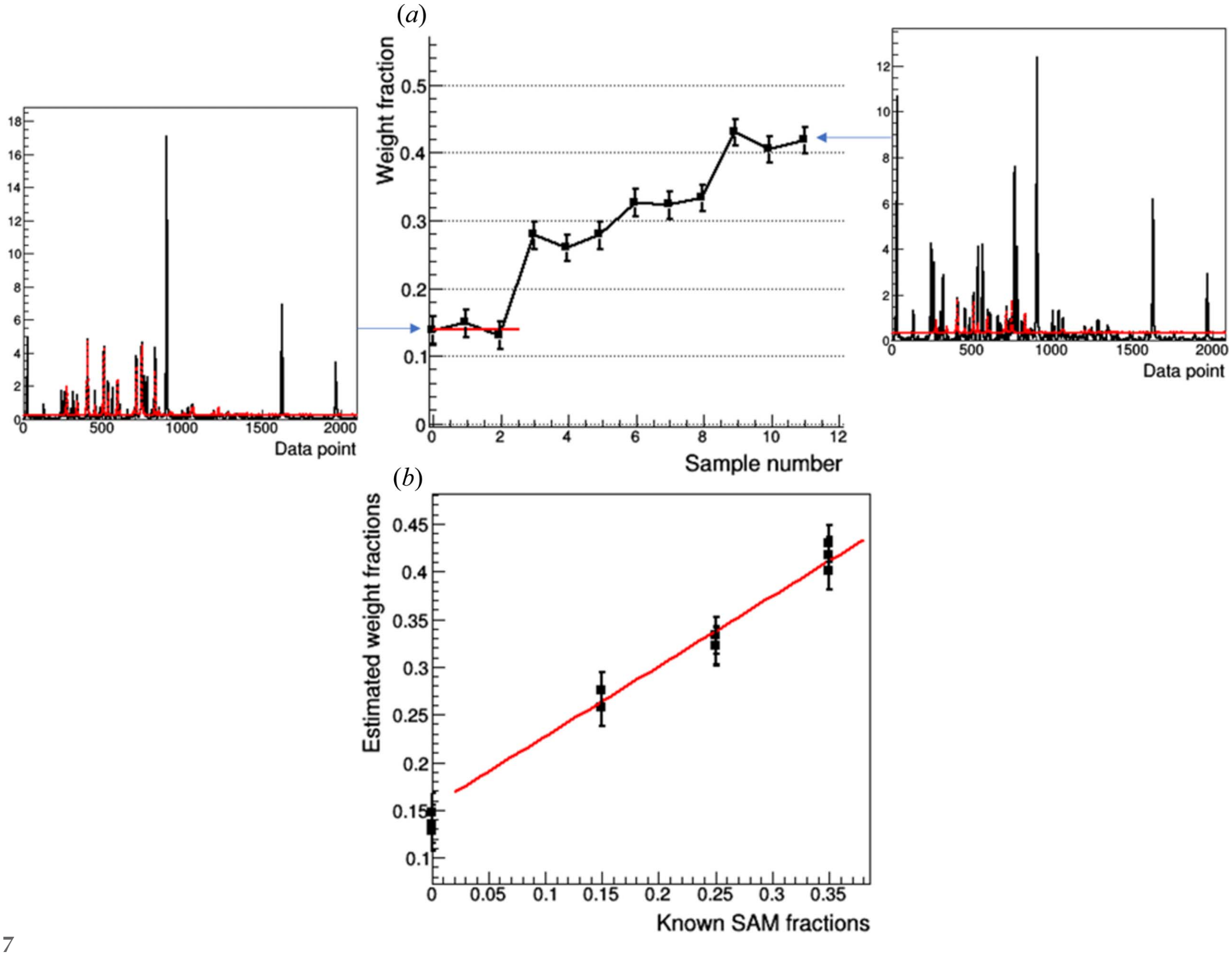


Figure 7
Quantitative analysis to assess the amount of paracetamol present in Tachifludec. (*a*) Weight fractions estimated from powder samples of Tachifludec, where paracetamol has been added at concentrations of 0, 15, 25, 35%(*w*/*w*) with three replicates per sample. A linear fit has been carried out on the samples without standard added (samples 0, 1, 2). Insets show the fit of the pure paracetamol profile (red line) on the first (left) and last (right) samples (black line). (*b*) Calibration plot obtained by the SAM procedure, where a linear fit has been carried out on the samples with the standard added (samples 3–11). Errors are calculated from the whole-profile fitting procedure on single mixtures, and regression lines are superimposed in red.

calculating the intercept of the regression line on the *y* axis, using samples 3–11. Alternatively, the weighted mean between the intercept of the regression line on the *y* axis and the absolute value of its intercept on the *x* axis was considered, which was found to be 0.152 ± 0.009. Notably, the nominal concentration of the API in Tachifludec is 0.145. *RootProf* performs these calculations by providing the known added concentration of the standard as input, as reported in Section S7 of the supporting information.

3.3.3. Combined fitting of different datasets.

Quantitative estimates by PXRD may be difficult to obtain when reference profiles have limited defining features. In fact, achieving a reliable quantitative phase analysis is quite challenging for poorly crystalline, nanosized or quasi-amorphous mixtures. This limitation has a notable impact on the pharmaceutical field, especially given that amorphous compounds can potentially offer higher bioavailability or may be the only option when crystal products cannot be obtained (Kavanagh *et al.*, 2012). However, the PDF approach provides a convenient means of investigating these types of compounds. By interpreting total scattering in direct space and defining the probability $G(r)$ of finding any two atoms at a given interatomic distance r, the PDF allows for more accurate qualitative analysis (Egami & Billinge, 2012). The MultiFit approach of *RootProf* has been adapted to improve the quantitative analysis of low-crystallinity compounds, by combining the results obtained in direct and reciprocal space. The approach, as shown in Fig. 8, involves collecting PXRD patterns on nine binary mixtures containing microcrystalline cellulose and the Eudragit L100 polymer (Mahdi, 2015) using a PDF setup that required the sample to be placed very close to the detector to acquire data at higher momentum transfer. A linear combination of the two pure-phase profiles was used to fit the data separately in reciprocal space [Figs. 8(*a*) and 8(*b*)] and in direct space [Figs. 8(*c*) and 8(*d*)]. By comparing Figs. 8(*b*) and 8(*d*) we can observe that Eudragit L100, which represents the quasi-amorphous phase (phase 1), was better quantified in reciprocal space, whereas MCC, which has higher crystallinity, was better quantified in direct space. Quantitative estimations in the direct and reciprocal spaces were then averaged, resulting in the improved estimates shown in Fig. 8(*e*), where the combined contribution from the dual spaces provided

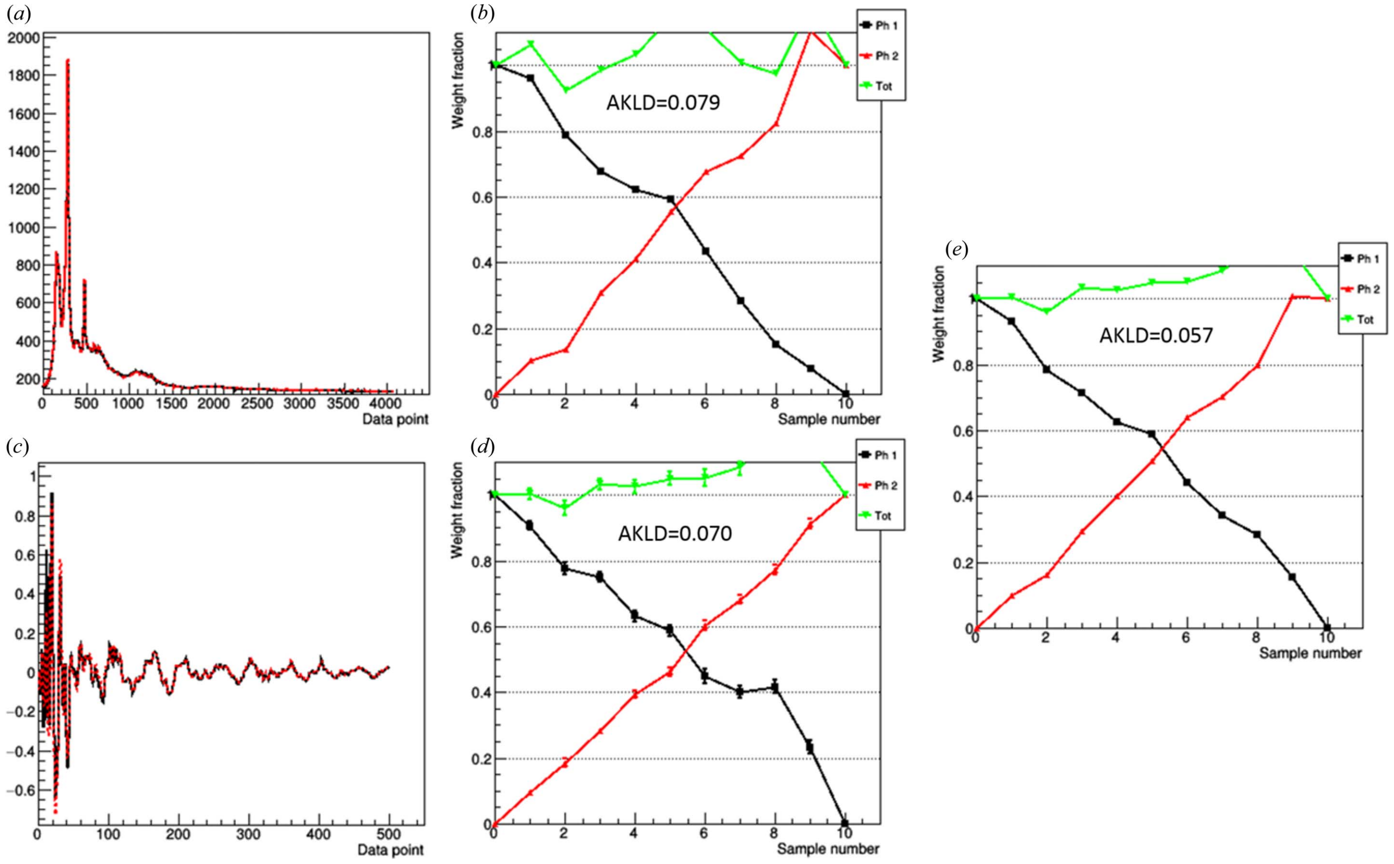


Figure 8
Quantitative analysis of a poorly crystalline pharmaceutical mixture, processing PXRD data in reciprocal and direct space. Least-square fits in (*a*) reciprocal and (*c*) direct space for sample number 8, are shown, with observed and calculated profiles shown in black and red, respectively. Estimated weight fractions of Eudragit L100 (black, Ph 1) and MCC (red, Ph 2) determined in (*b*) reciprocal and (*d*) direct space, and (*e*) obtained as the mean value between the two spaces. Error bars indicate fitting errors in (*b*) and (*d*), and propagated errors in (*e*). The sum of the weight fractions for each sample (in green) and the values of the Kullback–Leibner distance (AKLD) are reported.

equally precise estimations for both phases. The Kullback–Leibner distance (Kullback & Leibler, 1951) was used to quantify the improvement obtained, as reported in Figs. 8(*b*), 8(*d*) and 8(*e*).

A step forward in combining direct and reciprocal space datasets was attempted by implementing a procedure for combined fitting, but worse results were obtained, possibly due to the difficulty in properly weighting the profiles across the two spaces.

Notably, performing pre-processing of data could lead to fundamental improvement in quantitative analysis, and this could differ for the two datasets. With this aim, the *RootProf* procedures for supervised quantitative analysis, employing all possible pre-processing options sorted by a proper FOM, can be used to select the best pre-processing option.

3.4. Crystallinity and crystal size

The ability to calculate the average size of the crystallite domain in a powder sample from the width of selected peaks in the PXRD pattern by means of the Scherrer equation (Patterson, 1939) was already included in the first version of *RootProf* (Caliandro & Belviso, 2014). In the current version the procedure has been updated to include the Lorentzian and the pseudo-Voigt shape functions in addition to the Gaussian one. An example of the application is given in Fig. 9, where a peak at 2θ = 14.06° is fitted by a pseudo-Voigt [Fig. 9(*a*)] and the corresponding average size of the crystalline domain is determined for each of the 34 measurements taken on the same sample every 30 s [Fig. 9(*b*)]. The case study involves measuring a composite resin for dental restoration *in situ* for 17 min after exposing it to light for 40 s. The results showed no significant dependence on time of the crystallite size. However, a new option in *RootProf*, which analyses crystallinity [Fig. 9(*c*)], showed a clear dependence on time [Fig. 9(*d*)]. The time evolution of the polymerization process of the dental resin can be traced with this new feature, demonstrating an increase in the number of crystallite domains rather than in their size. The procedure for estimating crystallinity involves a peak search in the selected 2θ range of the profile, followed by fitting the profile with a sum of Gaussians centred on each peak, and estimating the background using the SNIP algorithm. The crystallinity is calculated using equation (4):

$$\text{Crystallinity} = \frac{\int_{x_{\min}}^{x_{\max}} f(x)\,\mathrm{d}x - \int_{x_{\min}}^{x_{\max}} b(x)\,\mathrm{d}x}{\int_{x_{\min}}^{x_{\max}} f\,\mathrm{d}x}, \qquad (4)$$

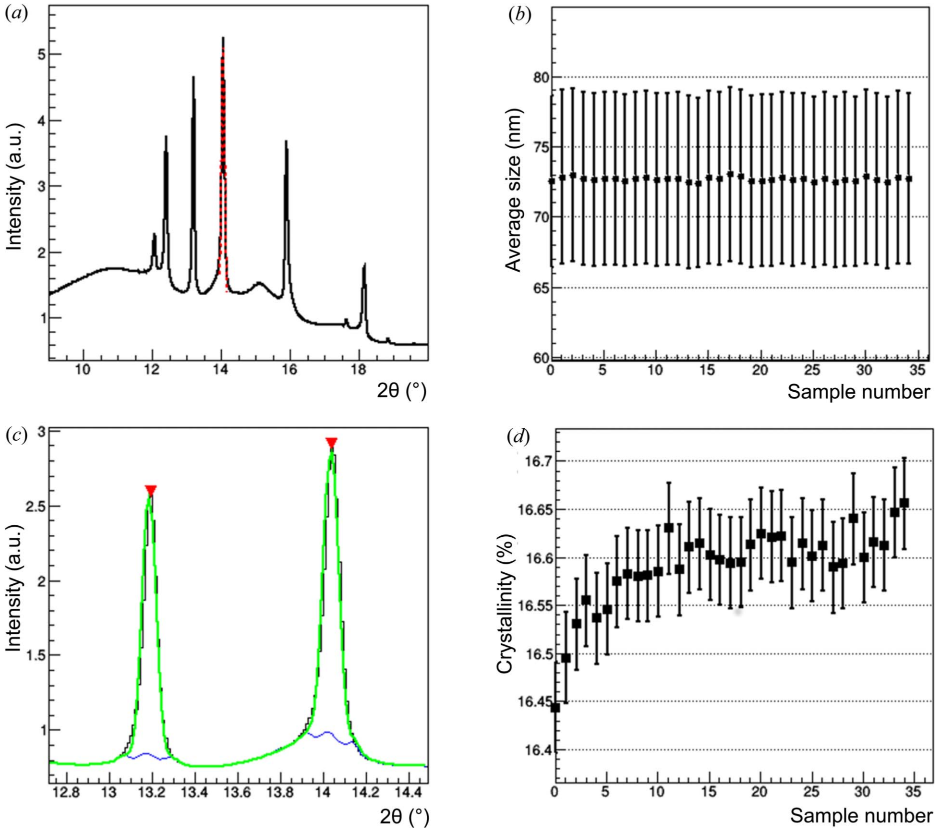


Figure 9
(*a*) Fit of the peak at 2θ = 14.06° of a PXRD profile (black line) by a pseudo-Voigt function (dashed red line) for profile No. 0; (*b*) values of the average size of crystalline domains determined for all the profiles; (*c*) fit of two peaks in the PXRD profile (black line) identified by a peak-search procedure (red triangles) by the sum of two Gaussian functions (green line) and background estimated by the SNIP algorithm (blue line) for profile No. 0; (*d*) values of the crystallinity determined for all the profiles.

where the f and b functions describe the Gaussian signal and the background, respectively, and $x_{\min}$ and $x_{\max}$ define the 2θ range of the profile considered.

3.5. Crystal cell parameter extraction from PDF profiles

A recent option developed for *RootProf* is the processing of PDF profiles to extract the crystal cell parameters (Guccione *et al.*, 2023). This model-free analysis is particularly useful for gaining information about the crystal phase and initiating the structure determination of nanocrystals or poorly crystalline samples in direct space. The procedure requires knowledge of the type of lattice, which could be, for example, acquired from the cloud platform *PDFitc* (Yang *et al.*, 2021). Three different strategies are followed according to the type of crystal cell related to the assumed crystal lattice. In the case of cells with a unique free length parameter (monometric cells), possible cell candidates are sought among the peaks of the PDF profile. As shown in Fig. 10, the peak-search procedure is facilitated by

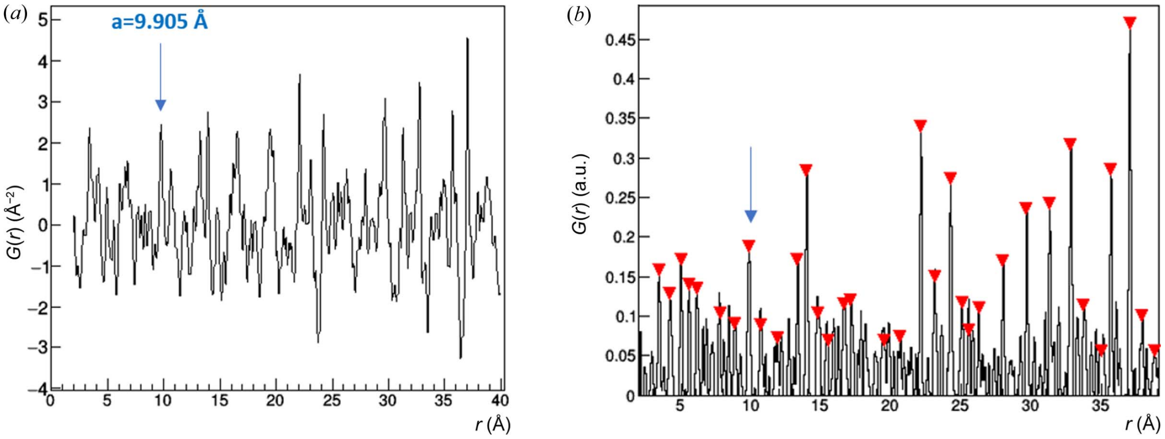


Figure 10
PDF profiles calculated from the cubic langbeinite $K_2Mg_2O_{12}S_3$ (Gajda *et al.*, 2022), (*a*) with the crystal cell parameter a = 9.905 Å, and (*b*) pre-processed to initiate the extraction of the crystal cell parameter, with the peaks considered for the calculation highlighted by a red triangle.

pre-processing the PDF profile so that it is rescaled in the range [0, 1] and background-subtracted by the SNIP algorithm (Ryan *et al.*, 1988) with a window of 10. Each candidate cell is then refined by a least-squares procedure where the pre-processed PDF profile is fitted against a synthetic profile calculated as the sum of Gaussians centred on the length of all the interatomic vectors generated by a putative monoatomic cell with the same parameters as the candidate cell. The refined cells are then clustered to reduce their number and the representative solutions of each cluster are sorted according to the FOM, which is the χ^2 of the fit. The output of the procedure is reported in Section S5.1 of the supporting information. It can be noted that a total of 12 solutions are generated and the first one (with minimum χ^2) has the cell parameter a = 9.9 Å, which corresponds to the true value (a = 9.905 Å).

In the case of cells with two free length parameters (dimetric cells), possible cell candidates are determined by applying a special unfolding procedure, borrowed from the quantitative analysis procedures implemented in *RootProf* (Section 3.3). The idea is to use a set of monoatomic cells generated by sampling the free cell parameters with a given step as components for the decomposition of the given PDF profile. The latter, shown in Fig. 11(*a*), is pre-processed as the cubic case [Fig. 11(*b*)]. The PDF profiles for each monoatomic cell have been pre-calculated and stored in a ROOT `TTree` structure and are supplied as external files when downloading *RootProf*. They contain the interatomic distances due to lattice translation, which represent a subset of the interatomic distances present in the given PDF profile. As a result of the unfolding procedure, the 'weight fractions' of each monoatomic cell are estimated, representing their affinity with the true cell. These are shown in Fig. 11(*c*) as a function of the sampled values of the free parameters a = b and c. A detrend processing is applied to these values by fitting the 2D histogram with a third-order bidimensional polynomial to eliminate the dependence on the number of interatomic distances considered (the affinity could be higher for monoatomic cells of small parameters simply because their PDF profiles have more peaks). The resulting affinities, shown in Fig. 11(*d*), undergo a peak-search procedure in two dimensions, which supplies the list of possible cell candidates. It is then narrowed down by checking that each free cell parameter of the cell candidates is present in the list of peaks of the input PDF profile, with a tolerance of 0.5 Å. The remaining cell

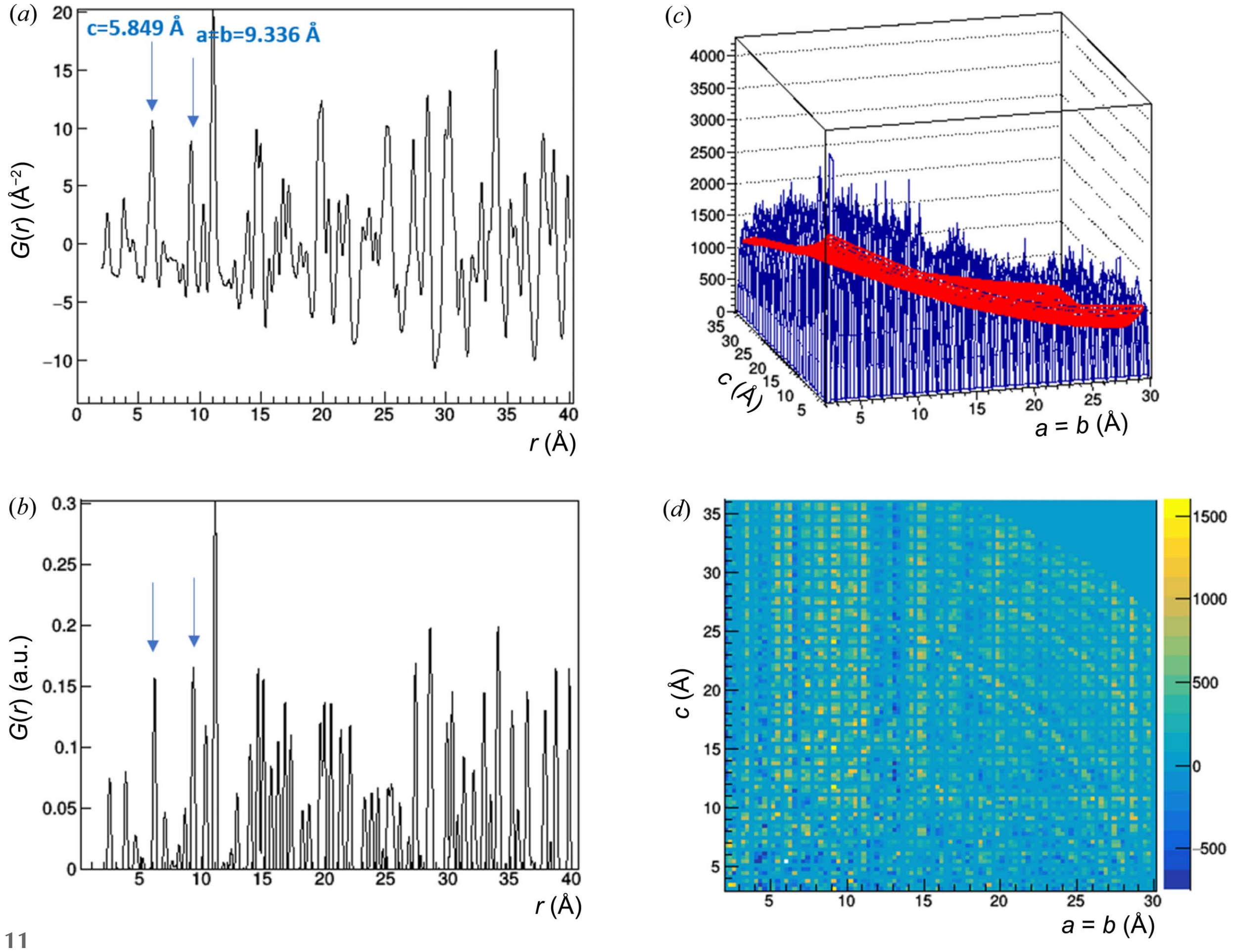


Figure 11

(*a*) Original and (*b*) pre-processed PDF profiles calculated from the hexagonal anhydrous rare-earth perchlorate $Pr(ClO_4)_3$ (Wickleder & Schäfer, 1999), with the crystal cell parameters a = b = 9.336 Å and c = 5.849 Å. (*c*) Weight fractions determined by the unfolding procedure applied to monoatomic crystal cells and fitted with a third-order bidimensional polynomial (surface in red) and (*d*) scatter plot showing the fit residuals.

candidates are then refined by a least-squares procedure similar to that described for the cubic case, where now two cell parameters are refined at the same time. The refined cells are finally listed in the output, sorted by the χ^2 of the fit. An example is given in Section S5.2 of the supporting information. It can be noted that a total of 127 solutions are reported, and the first one (with minimum χ^2) has the cell parameters $a = b =$ 9.34 Å and c = 5.85 Å, which correspond to those of the true cell ($a = b$ = 9.336 Å, c = 5.849 Å).

Similarly, in the case of trimetric cells, the parameters are extracted by combining the unfolding procedure for a global search of cell candidates with a least-squares fitting for their local optimization. An example is shown in Fig. 12 where the original [Fig. 12(*a*)] and pre-processed [Fig. 12(*b*)] PDF input profiles are shown. Given the higher number of free cell parameters, the heterogeneity of the possible interatomic vectors increases with respect to the previous cases, as well as the overlap of peaks in the PDF profile. The 'weight fractions' of each monoatomic cell estimated by the unfolding procedure are reported in Fig. 12(*c*), and in Fig. 12(*d*) after de-trending was carried out by fitting the 3D histogram with a third-order three-dimensional polynomial. A peak-search procedure in three dimensions supplies the list of possible cell candidates, which are narrowed down by checking that each free cell parameter of the cell candidates is present in the list of peaks of the input PDF profile, with a tolerance of 0.5 Å. The remaining cell candidates are then refined by a least-squares procedure, where three cell parameters are refined at the same time. The refined cells are finally listed in the output, sorted by the FOM, which for trimetric cells is defined as the intersection between the observed and calculated PDF (Guccione *et al.*, 2023). In Section S9.3 of the supporting information, an example is presented where a total of 500 solutions are reported. The first solution, with the maximum FOM, has the cell parameters a = 6.48 Å, b = 28.33 Å and c = 13.10 Å, which correspond to those of the true cell (a = 6.538 Å, a = 13.087 Å, c = 28.301 Å), but with b and c swapped. As the orthorhombic lattice maintains the set of interatomic vectors of its monoatomic cell invariant for cell length parameter permutations, such parameter warps are possible. The 18th cell solution, with parameters a = 6.48 Å, b = 13.10 Å and c = 25.15 Å, is the next closest solution to the true cell. The 103rd solution has the cell parameters a = 6.48 Å, b = 13.43 Å and c = 28.32 Å, which also correspond to those of the true cell.

Note that the basis set of pre-calculated monoatomic cells samples the cell parameters with a step of 0.4 Å for dimetric cells and 1.7 Å for trimetric cells. This selection was made to mantain a reasonable number of generated PDF profiles in both cases, with around 5000 profiles, and to limit the memory and computational capacity requirements of the machine

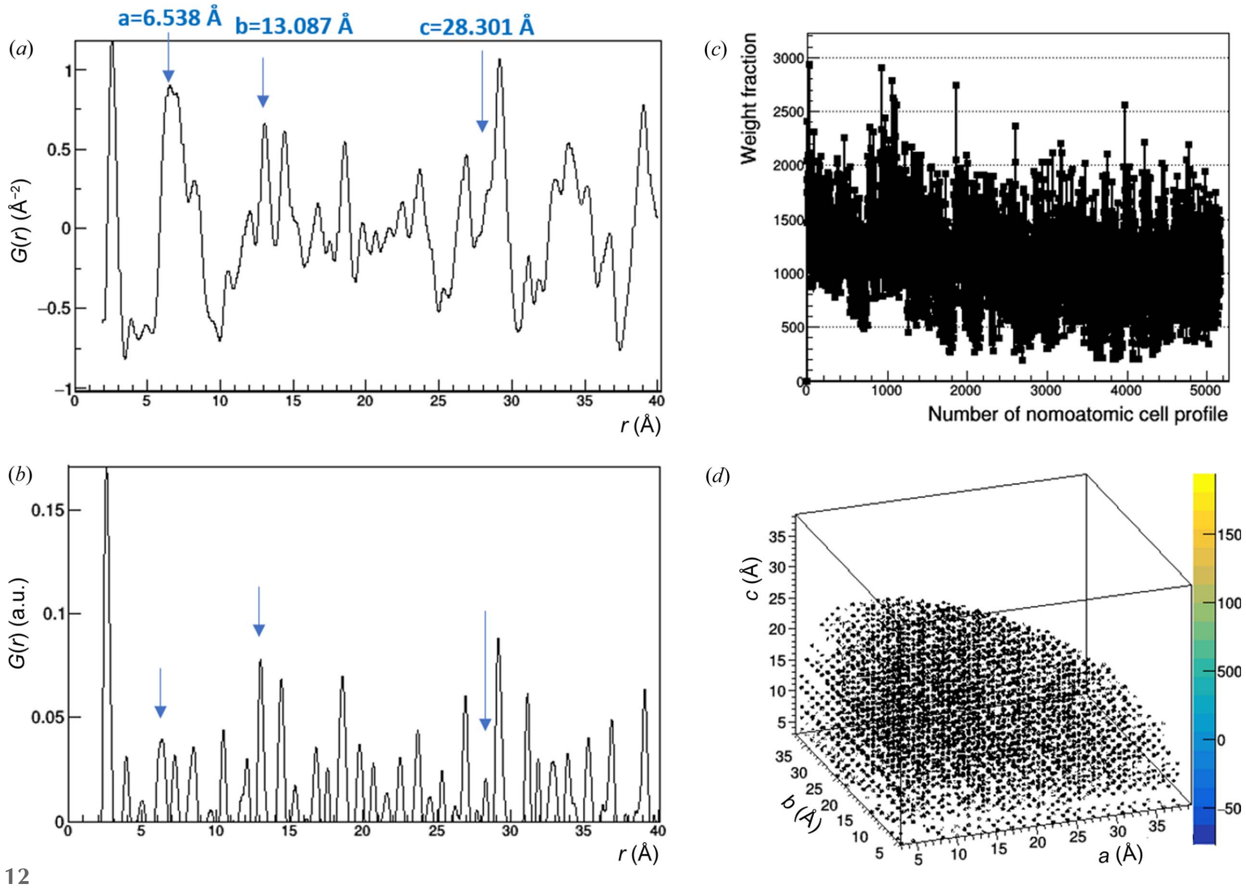


Figure 12
(*a*) Original and (*b*) pre-processed PDF profiles calculated from the orthorhombic maoecrystal V carbon skeleton $C_{25}H_{40}O_5Si$ (Krawczuk *et al.*, 2009), with the crystal cell parameters a = 6.538 Å, b = 13.087 Å and c = 28.301 Å. Weight fractions determined by the unfolding procedure applied to monoatomic crystal cells, (*c*) shown sequentially and (*d*) in 3D representation, after detrending.

running the program. However, the subsampling approach for trimetric cells could potentially impact the accuracy of the cell parameter estimations.

4. Conclusions

The computer program *RootProf* is designed for the general-purpose analysis of unidimensional profiles. It is based on the advanced processing capabilities of the ROOT software toolkit and has been optimized to handle crystallographic data, but when tested the results showed that it was also well suited for spectroscopic data. Users can utilize this program to have a clear view of (hidden) trends in data, which can be very useful for quick assessments during *in situ* experiments, fast quantitative analyses, combining data from the same sample taken from different techniques, and performing more general model-free analysis, such as extracting crystal cell parameters from a PDF profile or recognizing the diffraction signals from active atoms among those from bulk atoms. This analysis is propaedeutic to structural analysis, which can be performed with dedicated programs using input data generated by *RootProf*.

The new version of the program includes different algorithms originally developed using other packages, such as R or MATLAB, providing crystallographers with a complete platform able to explore the various crystallographic aspects, while extracting information from any type of data. Multivariate applications are specifically addressed, and the new version of the program is unique in providing PCA-based tools adapted to crystallography. Specific constraints derived from the MED theory have been applied to PCA to enable extraction of the time trend and diffraction signal from the part of the sample responding to the stimulus applied during the diffraction experiment. This is particularly important in dealing with complex samples of unknown structures, where we are interested in characterizing only the structural changes occurring during *in situ* measurements.

All applications presented can run on a laptop and require a limited amount of CPU time. Most of them can be executed in less than 1 min, with the extraction of crystal cell parameters from the PDF taking approximately 10 min for dimetric cells and 30 min for trimetric cells.

The program is available at https://www.ic.cnr.it/software/rootprof/ (free for academics), where a web-based user guide and tutorials are available.

Acknowledgements

The authors thank Marco Milanesio (Università del Piemonte Orientale) for enlightening suggestions and discussions, Pietro Guccione (Politecnico di Bari) for developing and testing some of the new *RootProf* algorithms in MATLAB, and Fabia Gozzo and Mathilde Reinle-Schmitt (Excelsus Structural Solutions) for providing the PXRD and PDF data for quantitative estimates of pharmaceuticals in dual spaces.

References

Belviso, B. D., Marin, F., Fuertes, S., Sicilia, V., Rizzi, R., Ciriaco, F., Cappuccino, C., Dooryhee, E., Falcicchio, A., Maini, L., Altomare, A. & Caliandro, R. (2021). *Inorg. Chem.* **60**, 6349–6366.

Belviso, B. D., Tangorra, R. R., Milano, F., Omar, O. H., la Gatta, S., Ragni, R., Agostiano, A., Farinola, G., Caliandro, R. & Trotta, M. (2016). *MRS Adv.* **1**, 3789–3800.

Benoit (1924). *Bull. Géodésique*, **2**, 66–67.

Bolognino, I., Carrieri, A., Purgatorio, R., Catto, M., Caliandro, R., Carrozzini, B., Belviso, B. D., Majellaro, M., Sotelo, E., Cellamare, S. & Altomare, C. D. (2022). *Separations*, **9**, 7.

Brennhagen, A., Cavallo, C., Wragg, D. S., Sottmann, J., Koposov, A. Y. & Fjellvåg, H. (2021). *Batteries Supercaps*, **4**, 1039–1063.

Brun, R. & Rademakers, F. (1997). *Nucl. Instrum. Methods Phys. Res. A*, **389**, 81–86.

Caliandro, R. & Belviso, D. B. (2014). *J. Appl. Cryst.* **47**, 1087–1096.

Caliandro, R., Chernyshov, D., Emerich, H., Milanesio, M., Palin, L., Urakawa, A., van Beek, W. & Viterbo, D. (2012). *J. Appl. Cryst.* **45**, 458–470.

Caliandro, R., Guccione, P., Nico, G., Tutuncu, G. & Hanson, J. C. (2015). *J. Appl. Cryst.* **48**, 1679–1691.

Caliandro, R., Sibillano, S., Belviso, B. D., Scarfiello, R., Hanson, J. C., Dooryhee, E., Manca, M., Cozzoli, P. D. & Giannini, C. (2016). *ChemPhysChem*, **17**, 699–709.

Caliandro, R., Toson, V., Palin, L., Conterosito, E., Aceto, M., Gianotti, V., Boccaleri, E., Dooryhee, E. & Milanesio, M. (2019). *Chem. A Eur. J.* **25**, 11503–11511.

Chernyshov, D., van Beek, W., Emerich, H., Milanesio, M., Urakawa, A., Viterbo, D., Palin, L. & Caliandro, R. (2011). *Acta Cryst.* A**67**, 327–335.

Chernyshov, D., Dovgaliuk, I., Dyadkin, V. & van Beek, W. (2020). *Crystals*, **10**, 581.

Coats, A. V. & Redfern, J. P. (1964). *Nature*, **201**, 68–69.

Colella, S., Todaro, M., Masi, S., Listorti, A., Altamura, D., Caliandro, R., Giannini, C., Carignani, E., Geppi, M., Meggiolaro, D., Buscarino, G., De Angelis, F. & Rizzo, A. (2018). *ACS Energy Lett.* **3**, 1840–1847.

Cornell, J. (2011). *Experiments with Mixtures: Designs, Models, and the Analysis of Mixture Data*, Wiley Series in Probability and Statistics. Hoboken: Wiley.

Egami, T. & Billinge, S. (2012). *Underneath the Bragg Peaks: Structural Analysis of Complex Materials*, Vol. 16. Amsterdam: Elsevier Science.

Gajda, R., Zhang, D., Parafiniuk, J., Dera, P. & Woźniak, K. (2022). *IUCrJ*, **9**, 146–162.

Goiss, E. A. (1963). *J. Am. Ceram. Soc.* **46**, 274.

Guccione, P., Diacono, D., Toso, S. & Caliandro, R. (2023). *IUCrJ*, **10**, 610–623.

Guccione, P., Lopresti, M., Milanesio, M. & Caliandro, R. (2021). *Crystals*, **11**, 12.

Guccione, P., Palin, L., Belviso, B. D., Milanesio, M. & Caliandro, R. (2018*a*). *Phys. Chem. Chem. Phys.* **20**, 19560–19571.

Guccione, P., Palin, L., Milanesio, M., Belviso, B. D. & Caliandro, R. (2018*b*). *Phys. Chem. Chem. Phys.* **20**, 2175–2187.

Guest, P. G. (2012). *Numerical Methods of Curve Fitting*. Cambridge University Press.

Jandel, M., Morháč, M., Kliman, J., Krupa, L., Matoušek, V., Hamilton, J. H. & Ramayya, A. V. (2004). *Nucl. Instrum. Methods Phys. Res. A*, **516**, 172–183.

Kavanagh, A., McConvey, I., McCabe, J., Blade, H. & Cosgrove, S. (2012). *Am. Pharm. Rev.* 119521.

Krawczuk, P. J., Schöne, N. & Baran, P. S. (2009). *Org. Lett.* **11**, 4774–4776.

Kullback, S. & Leibler, R. (1951). *Ann. Math. Stat.* **22**, 79–86.

Liuzzi, V. C., Mirabelli, V., Cimmarusti, M. T., Haidukowski, M., Leslie, J. F., Logrieco, A. F., Caliandro, R., Fanelli, F. & Mulè, G. (2017). *Toxins*, **9**, 45.

Lopresti, M., Mangolini, B., Conterosito, E., Milanesio, M. & Palin, L. (2023). *Cryst. Growth Des.* **23**, 1389–1402.

Lopresti, M., Mangolini, B., Milanesio, M., Caliandro, R. & Palin, L. (2022). *J. Appl. Cryst.* **55**, 837–850.

Mahdi, H. J. (2015). *Issues Biol. Sci. Pharm. Res.* **3**, 14–20.

Massara, N., Boccaleri, E., Milanesio, M. & Lopresti, M. (2021). *HardwareX*, **10**, e00231.

Miciaccia, M., Belviso, B. D., Iaselli, I., Cingolani, G., Ferorelli, S., Cappellari, M., Loguercio Polosa, P., Perrone, M. G., Caliandro, R. & Scilimati, A. (2021). *Sci. Rep.* **11**, 4312.

Mobius, A. F. (1827). *Der Barycentrische Calcül*. Leipzig: Verlag von Johann Ambrosius Barth.

Olds, D., Peterson, P. F., Crawford, M. K., Neilson, J. R., Wang, H.-W., Whitfield, P. S. & Page, K. (2017). *J. Appl. Cryst.* **50**, 1744–1753.

Palin, L., Caliandro, R., Viterbo, D. & Milanesio, M. (2015). *Phys. Chem. Chem. Phys.* **17**, 17480–17493.

Patterson, A. (1939). *Phys. Rev.* **56**, 978–982.

Piparo, D., Tejedor, E., Guiraud, E., Ganis, G., Mato, P., Moneta, L., Valls Pla, X. & Canal, P. (2017). *J. Phys. Conf. Ser.* **898**, 072022.

Rabiner, L., Rosenberg, A. & Levinson, S. (1978). *IEEE Trans. Acoust. Speech Signal. Process.* **26**, 575–582.

Ryan, C. G., Clayton, E., Griffin, W. L., Sie, S. H. & Cousens, D. R. (1988). *Nucl. Instrum. Methods Phys. Res. B*, **34**, 396–402.

Savitzky, A. & Golay, M. J. E. (1964). *Anal. Chem.* **36**, 1627–1639.

Wickleder, M. S. & Schäfer, W. (1999). *Z. Anorg. Allg. Chem.* **625**, 309–312.

Yang, L., Culbertson, E. A., Thomas, N. K., Vuong, H. T., Kjær, E. T. S., Jensen, K. M. Ø., Tucker, M. G. & Billinge, S. J. L. (2021). *Acta Cryst.* A**77**, 2–6.

Zappi, A., Maini, L., Galimberti, G., Caliandro, R. & Melucci, D. (2019). *Eur. J. Pharm. Sci.* **130**, 36–43.