Title

Pattern-matching indexing of Laue and monochromatic serial crystallography data for applications in materials science

Permalink

https://escholarship.org/uc/item/2kf2105k

Journal

Journal of Applied Crystallography, 53(3)

ISSN

0021-8898

Authors

Dejoie, Catherine
Tamura, Nobumichi

Publication Date

2020-06-01

DOI

10.1107/s160057672000521x

Peer reviewed

Pattern matching indexing of Laue and monochromatic serial crystallography data for applications in Materials Science

Authors

XCatherine Dejoie[a]* and Nobumichi Tamura[b]*

[a] European Synchrotron Radiation Facility, 71 Avenue des Martyrs, Grenoble, 38000, France

[b]Advanced Light Source, Lawrence Berkeley National Lab, 1 Cyclotron road, Berkeley, CA, 94720, USA

Correspondence email: catherine.dejoie@esrf.fr; ntamura@lbl.gov

ynopsis An algorithm, based on the matching of **q**-vectors pairs, is combined with three-dimensional pattern matching using a nearest-neighbors approach to index Laue and monochromatic serial crystallography data recorded on small unit cell samples.

bstract Serial crystallography data could be challenging to index, as each frame is processed individually, rather than processed as a whole like in conventional X-ray single-crystal crystallography. We developed an algorithm to index still diffraction patterns arising from small unit cell samples. The algorithm is based on the matching of **reciprocal lattice vector** pairs, as developed for Laue microdiffraction **data indexing, combined with three-dimensional pattern matching using a nearest-neighbors approach. As a result, large bandpass data (e.g. 5-24 keV energy range) as well as monochromatic data can be processed, the main requirement being the *prior* knowledge of the unit cell. Angles calculated in the vicinity of a few theoretical and experimental reciprocal lattice** vectors are compared, and only **vectors** with the highest number of common angles are selected as candidates to obtain the orientation matrix. **Global matching on the entire pattern is then checked. Four indexing options are available, two for the ranking of the theoretical reciprocal lattice vectors**, and two for reducing the number **of possible candidates. The algorithm was used to index several datasets collected under different experimental conditions on a series of model samples. Knowing the crystallographic structure of the sample and using this information to rank the theoretical reflections based on the structure factor helps the**

indexing of large bandpass data for the largest unit cell samples. For small bandpass data, shortening the candidate list to determine the orientation matrix should be based on pairs of reciprocal lattice vectors matching instead of triplet matching.

eywords: Indexing; Energy bandpass; Laue microdiffraction; Serial Crystallography.

1. Introduction

With the emergence of high-energy X-ray free electron laser (XFEL) sources generating ultra-fast X-ray pulses of high brilliance, serial crystallography has become a method of choice to collect single-crystal X-ray diffraction data (Chapman *et al.*, 2011; Boutet *et al.*, 2012). By exposing a crystal to a single monochromatic X-ray pulse, a diffraction pattern is collected before radiation damage occurs. By combining a series of single shot diffraction patterns obtained from randomly oriented crystals, a complete dataset can be retrieved, and the structure of complex systems studied, with a focus on macromolecular structural biology applications (Johansson *et al.*, 2017). The sudden rise of serial crystallography has triggered the development of dedicated analytical tools, to analyze the huge number of diffraction patterns collected, index the individual diffraction patterns, and reconstruct useable reflection intensities (White *et al.*, 2016; Hattne *et al.*, 2014; Kabsch 2014; Liu & Spence, 2016). Owing to the nature of the XFEL beam (each X-ray pulse has its own energy and intensity spectrum), the sample variability (the crystallites exposed to the beam may vary in size and crystallinity), and the measurement strategy (one or several crystals randomly oriented may diffract simultaneously), data processing can be very challenging.

In recent years, new strategies to index such complex data have been proposed, with a focus on sparse data and smaller unit cell samples, in particular making use of a *prior* knowledge of the unit cell (Brewster *et al.*, 2015; Ginn *et al.*, 2016; Li *et al.*, 2019). For example, in order to obtain the crystal orientation matrix of still images with sparse data, Brewster *et al.* (2015) used a powder-like diffraction pattern reconstructed from the aggregate of thousands of still images to derive accurate cell information, before using the model powder pattern to assign initial Miller indices to reflections. Li *et al.* (2019) have developed an auto-indexing algorithm for sparse and small unit cell diffraction data, comparing the length and angles of paired scattering vectors with referenced values derived from *prior* knowledge of the unit cell. Finally, dedicated tools have been developed to take into account the non-monochromatic nature of the XFEL beam in the indexing process (Gevorkov *et al.*, 2019).

The term of “serial crystallography” to define the concept of collecting single shot data and processing each frame individually was used at an earlier stage with XFEL, especially in the

case of Laue diffraction. When a crystal is exposed to a broad energy bandpass (polychromatic, pink or white beam), a reasonably large number of reflections can be recorded simultaneously in a single exposure. Because of this, the Laue method is a good alternative to the monochromatic one for *in situ* time-resolved studies of macromolecules (Moffat & Helliwell, 1989; Bourgeois *et al.*, 2003; Yorke *et al.*, **2014). If combined with a micron or sub-micron size beam, Laue diffraction can also be used to map crystal orientation and strain in materials (Chen *et al.*, 2016). In addition, a complete structure characterization only requires a few random orientations to be combined (Cornaby** *et al.*, **2010; Dejoie, McCusker, Baerlocher, Kunz & Tamura, 2013). The indexing of individual patterns collected using the Laue microdiffraction technique is generally based on a matching process of pairs of reciprocal lattice vectors** (Chung & Ice, 1999; Tamura, 2014), which requires *prior* **knowledge of the unit cell. In short, each measured reflection is converted into normalized reciprocal lattice vectors (they are normalized, as their length is not directly accessible in Laue diffraction), and the angle between the vectors are matched with a list of expected angles calculated from the known unit cell. Data processing for structure solution based on a set of Laue reflection integrated intensity measurements tends to be difficult, mainly due to the energy dependence of the various correction factors, and the overlap of harmonic reflections (Helliwell *et al.*, 1989).**

The non-monochromatic nature of the XFEL beam attracted our attention a few years ago. With a small energy bandpass (a few percent in ΔE/E, E being the energy of the incident beam), more reflections are in diffraction condition, and the probability a reflection intensity to be truncated is also reduced. The possibility to measure more Bragg peaks in a single shot is particularly interesting for samples with small unit cells. As a 4% energy bandpass beam had been planned at the Swiss free electron laser (SwissFEL) (Patterson *et al.*, 2014), we developed a methodology to simulate such data and implement the data processing appropriate for small unit cell samples. A first indexing of the simulated data was carried out using a Laue microdiffraction approach, showing that such Laue indexing algorithm could be adapted to index data collected over smaller energy bandpass (Dejoie, McCusker, Baerlocher, Abela *et al.*, **2013). A short description of the indexing strategy was published in Dejoie *et al.* (2015). The "classic" Laue pair of reciprocal lattice vectors** matching approach (Chung & Ice, 1999; Tamura, 2014) was combined with a three-dimensional pattern matching approach based on nearest neighbor**s initially developed for fast 2D pattern matching of fingerprints (Van Wamelen *et al.*, 2004), which appeared to be efficient to index these 4% bandpass data.**

Keeping a similar combined approach, the code has been revised and optimized for the indexing of different types of data collected with varying energy bandpass, from Laue (5-24 keV range) to

monochromatic. A complete description of the indexing algorithm is presented here, along with the results of the indexing tests carried out on five small unit cell model samples (cell volume ranging from 722 to 6640 Å^3). First, angles in the vicinity of selected theoretical and experimental reciprocal lattice vectors are calculated, and only vectors with the highest number of common angles are kept. The list of candidates from nearest-neighbors matching can be further reduced before checking for global match. Four indexing strategies are available, depending on whether or not the crystallographic structure is (at least partially) known (this will affect the ranking of theoretical reflections), and on the type of matching process chosen (based on pair or triplet of reciprocal lattice vectors matching). Our objective is to identify the main parameters influencing the indexing process, to check the limitations of the indexing algorithm, and to propose the best indexing strategy depending on the type of sample and the type of diffraction data. We show that knowing the crystallographic structure of the sample helps the indexing of large bandpass data for the largest unit cell samples. On the other hand, for the indexing of small bandpass data, a matching process based on pairs of reciprocal lattice vectors should be favored. The current indexing algorithm uses routines written in the XMAS software (Tamura, 2014), but can be used as a stand-alone program.

2. Sample tests and data acquisition

The indexing algorithm has been tested on X-ray diffraction data collected on five model samples. The chemical composition and main crystallographic information for each of the five samples are given in Table 1. A 30 µm thin section of feldspar (sanidine) (Ackermann *et al.*, 2004) was provided by Professor H. R. Wenk (UC Berkeley, USA). Hydrated caesium cyanoplatinate (CsPt) (Johnson *et al.*, 1977) and ZSM-5 zeolite crystals (Olson *et al.*, 1981; van Koningsveld *et al.*, 1987) were provided by Dr. P. Pattison (EPFL Lausanne, Switzerland) and Professor Henri Kessler (Université de Haute-Alsace, Mulhouse, France), respectively. The zirconium phosphate (ZrPOF) (Liu *et al.*, 2009) and the magnesium acetate (MgAc) samples were provided by Dr. L. B. McCusker (ETH Zurich, Switzerland). The magnesium acetate structure was refined using single crystal data collected at the ALS-11.3.1 beamline. Results agree with the published structure (Scheurell *et al.*, 2015).

X-ray single crystal diffraction data were collected using monochromatic ($\Delta E/E \sim 10^{-4}$) and non-monochromatic X-ray incident beams: $\Delta E/E \sim 4$ %; E ~ 11-17 keV (e.g. $\Delta E/E \sim 50$ %); and E ~ 5-24 keV. Information about all the datasets collected varying both the energy range and the experimental setup (setups (1) to (5)) are summarized in Table 2.

Conventional monochromatic data (setup (1)) and 4% bandpass data ($\Delta E/E \sim 4$ %, setup (2)) were collected at the Swiss–Norwegian Beamline (SNBL/BM01A) at the European Synchrotron

Radiation Facility (ESRF). To do so, single crystals of sanidine, CsPt, ZSM-5 and ZrPOF were mounted on MiTeGen MicroMeshes. A two-dimensional DECTRIS Pilatus 2M detector was positioned at a distance of 224 mm from the sample. The broad bandpass mode was achieved by collecting a diffraction pattern while the monochromator was scanned over a 4% energy bandpass (average energy 17.34 keV, or 0.7153 Å). The shape of the X-ray incident spectrum achieved in such a way was extracted using the 'reverse method' from sanidine data (Dejoie *et al.*, 2011) (Fig. 1a). The monochromatic datasets and 4% bandpass datasets were collected by rotating single crystals using 0.25° rotation and 1° rotation step, respectively. Geometry calibration (sample-detector distance, normal incidence position of the detector, tilt angle of the detector) was carried out with the XMAS program (Tamura, 2014), using a LaB_6 reference powder pattern.

Laue diffraction (E ~ 11-17 keV and E ~ 5-24 keV) experiments were conducted on Beamline 12.3.2 at the Advanced Light Source (ALS) of the Lawrence Berkeley National Laboratory. A polychromatic X-ray beam (5–24 keV) was focused down to about 1 × 1 µm using a pair of Kirkpatrick-Baez (KB) mirrors (Tamura *et al.*, 2003; Kunz *et al.*, 2009). Laue microdiffraction patterns were collected using a two-dimensional DECTRIS Pilatus 1M X-ray detector. An exposure time of 1s per pattern was used for all samples. The sample–detector distance, the center channel of the detector and the tilt of the detector relative to the sample surface were calibrated using a Laue pattern obtain from a strain-free Si single-crystal.

Two different setups were used to collect full range Laue data (5–24 keV). In the first setup (DET-90, setup (5)), the two-dimensional Pilatus detector was positioned at 90° with respect to the incident beam and the sample at 45° (reflection geometry). The distance from the sample to the center of the detector was ~141 mm. This configuration is optimized to collect more data in a single shot by favoring the high-resolution (low *d*-spacing) range and is used in a routine way for strain and stress mapping (Tamura *et al.*, 2003). Single shot patterns were collected on single crystals of ZSM-5 and MgAc randomly dispersed on a glass slide, and on the thin section of sanidine. In the second setup (setup (4)), both reflection and transmission geometries were used. First, the two-dimensional detector was placed at 60° (sample-detector distance ~147mm) and the sample at 15° relative to the incident X-ray beam (reflection geometry, DET-60), and single crystal data were collected on single crystals of ZSM-5 and MgAc randomly dispersed on a glass slide. For CsPt, ZrPOF and sanidine, data were collected in transmission geometry with the 2D detector positioned at 50° (sample-detector distance ~167mm, DET-50). The crystals were spread over Mitegen Micromounts covered with a 10nm Au layer. Data were collected using the methodology described in Dejoie, McCusker, Baerlocher, Kunz & Tamura (2013). Both DET-60 and DET-50 setups give access to lower resolution data. The shape of the X-ray incident spectrum corresponding to setup (4) and setup (5) can be seen in Fig. 1c.

The reduced Laue range (E ~ 11-17 keV, setup (3)) was achieved in the following way: the high energy part of the beam was cut by increasing the pitch angle of the vertically focusing KB mirror, and the low energy part was restricted using the low-energy threshold of the Pilatus detector. As a result, the vertical size of the beam increases to ~4µm, the overall intensity of the incident flux decreases by about 20%, and the number of reflections of different orders (harmonics) decreases from 25% (Dejoie, McCusker, Baerlocher, Kunz & Tamura, 2013) to 5%. The resulting incident flux as a function of energy is shown on Fig. 1b. Some remaining intensity can be seen between 17 and 19 keV, with limited impact on data processing. The detector was positioned at 50° relative to the incident X-ray beam, for a sample-detector distance of ~167mm. Sanidine, ZrPOF and MgAc crystals were spread on Mitegen Micromounts and measured in transmission geometry. ZSM-5 data were measured using a geometry in reflection (sample positioned at 15° relative to the incident beam).

Powder diffraction data were collected at the high-resolution powder diffraction beamline (ID22) at the European Synchrotron Radiation Facility (Grenoble, France). The ZrPOF, ZSM-5 and MgAc samples were packed in 0.7mm-sized borosilicate capillaries, and measured using a wavelength of 0.399963 Å over 40° (2θ) at a speed of 2° min^{-1}. The ZrPOF sample suffering radiation damage, several fast 2θ scans (15°/min) have been collected on a fresh zone of the sample at 100K before averaging. Pawley fits were carried out using the TOPAS software (Coelho, 2018).

3. Description of the indexing algorithm

The indexing algorithm combines the pair of reciprocal lattice vectors matching strategy first proposed by Chung & Ice (1999) for the indexing of Laue microdiffraction data, with a nearest neighbor ranking approach initially proposed for pattern matching of 2D data such as fingerprint matching (Van Wamelen *et al.*, 2004). A first version of the algorithm has been written for the indexing of 4% bandpass data (Dejoie, McCusker, Baerlocher, Abela *et al.*, 2013; Dejoie *et al.*, 2015). The current version can be used to index several types of serial crystallographic data, from monochromatic to Laue. There are two important aspects in relation with the indexing process: i) it requires *prior* knowledge of the unit cell parameters of the sample; ii) it does not require knowledge of the energy of the diffraction peaks, which is retrieved during the indexing process. The main steps of the indexing process are described hereafter.

3.1. Step 1: data pre-processing

Prior to indexing, experimental patterns were processed (background subtraction, peak search) using the routines developed in the XMAS software (Tamura, 2014). The positions of the experimental peaks (X_{exp}, Y_{exp}) in pixels on the 2D diffraction images were obtained, and corresponding integrated intensities (I_{int}) extracted using a simple box method, before ranking them by decreasing values. The current version of the indexing algorithm uses a simple text file per frame as input, in which the positions of the experimental peaks in pixels and their integrated intensities for a given frame of interest are listed. Next step is to convert the reflection positions measured on the area detector into lattice vectors of the reciprocal space. Knowing the experimental setup (sample-detector distance, center channel and tilt of the detector), the experimental peaks are then converted into normalized reciprocal lattice vectors (q-vectors), as defined by the following equation:

$$q=\frac{k_{out}-k_{¿}}{||k_{out}-k_{¿}||}$$

with $k_{¿}$ the incident wave vector and k_{out} the wave vector pointing from the diffracting volume towards the reflection on the detector (Fig. 2a). The lengths of the q-vectors are normalized to unity because the wavelengths of the reflections are not *a priori* known for non-monochromatic data (Fig. 2b).

From the known unit cell parameters of the crystal, the theoretical reciprocal lattice reflections are calculated. Three options for the ranking of these reflections are available: i) if the structure (i.e. atomic decoration of the unit cell) or part of it is known, structure factors are also calculated, and theoretical reflections are ranked by decreasing structure factors. This is the strategy used in the Laue indexing algorithm implemented in the XMAS software (Tamura, 2014); ii) in the case of a fully unknown structure (but known unit cell parameters), theoretical reflections are ranked by decreasing *d*-spacing; iii) a third mode is available, in which a user-defined list is used, and extended if necessary by calculated theoretical reflections ranked by *d*-spacing. For example, the user-defined list can be generated from a powder pattern, after extraction of the strongest intensity reflections using a Le Bail or Pawley fit. The ranking of the theoretical reflections according to the “likelihood” to be found in the actual dataset is a way to increase the speed efficiency of the indexing process, especially for low symmetry materials, by considerably decreasing the number of calculations. Once generated following one of the three options described above, theoretical reflections are converted into normalized q-vectors. The number of calculated theoretical q-vectors can be restricted, either to a particular threshold *d*-

spacing if the *d*-spacing ranking strategy has been chosen, or by imposing a minimum structure factor value in the case of the structure factor ranking strategy.

3.2. Step 2: nearest neighbour matching

The indexing starts at this step. The q-vectors corresponding to the brightest (highest integrated intensities) experimental reflections are first considered, and the angles between these selected q-vectors and the other experimental q-vectors in the first vicinity are calculated. The first vicinity/neighborhood is defined by a limiting maximum threshold angle. Similar strategy is applied around the q-vectors corresponding to a selection of theoretical unique reflections ranked following one of the three options described above (Fig. 3a).

Then, the neighborhood of experimental q-vectors and of theoretical q-vectors are compared in terms of the number of common angles. At the end of this process, a list of theoretical unique reflection with similar neighborhood is obtained for each experimental peak, these theoretical reflections being ranked depending on the number of common q-vector angles. This is the "Nearest neighbors list". For example, as can be seen on Fig . 3a, the experimental **q**-vector $q_{exp\_start}(1)$ (corresponding to the experimental reflection with the highest intensity) has 5, 5, 3 and 4 similar neighbors with the theoretical unique reflections *m*, *n*, *o*, and *p*, respectively. These four theoretical matching candidates will then be ranked as *m*, *n*, *p*, *o* or *n, m, p, o*.

3.3. Step 3: global matching

To verify global matching (i.e. reflection matching not limited to the nearest neighbors), a candidate orientation matrix has to be generated. The indexing is considered to be successful if more than a minimum number of experimental peaks are indexed. The orientation matrix is built out of 3 vectors (or1, or2, or3), with the first vector or1 selected from the "Nearest neighbors list", and the vector or2 selected from a non-restricted list of theoretical q-vectors.

The theoretical unique q-vectors with the highest number of common neighbors obtained from step 2 can be used directly to build the first vector or1. This approach was implemented and tested in Dejoie *et al.* (2015). Nevertheless, most of the time, these reflections are not the best candidates to build the first vector of the orientation matrix. In particular, when the energy range and/or the volume of the cell increase, the number of calculated reflections increases, and the number of common neighbors is becoming a less discriminating parameter. As a result, indexing time increases. To overcome this, additional conditions were introduced in order to select the first vector or1, and two additional options are currently available.

The first additional option to retrieve candidates for or1 relies on a "pair matching" selection among the reflections of the "Nearest neighbors list". If the angle between the q-vectors of two experimental peaks is α, then the angle between the corresponding matching theoretical q-vectors has to be α (within an angular resolution). This is illustrated in Fig. 3b, where two out of four vectors ($q_{th\_match}(1)(1)$ and $q_{th\_match}(1)(3)$) fulfil the condition. By applying the "pair matching" conditions to all of the reflections part of the "Nearest neighbors list", a reduced list of candidates for the vector or1 is generated.

The second option is an extension of the previous concept, implemented over three q-vectors ("triplet matching"). If an experimental q-vector is making an angle α with a second experimental q-vector and an angle β with a third one, then, same should apply to the corresponding theoretical q-vectors from the "Nearest neighbors list", as illustrated in Fig. 3c.

In order to determine the second vector of the orientation matrix or2, a pair of q-vectors matching strategy is used, over experimental and theoretical q-vectors. Candidates are generated by looking for any experimental/theoretical couples with similar angles (within an angular resolution). In a similar way as for or1, an additional "pair matching" restriction in the selection process applies: if the angle between two experimental q-vectors is α, then the angle between the two corresponding selected theoretical q-vectors should also be α within a given angular resolution.

The third vector or3 of the orientation matrix is deduced from or1 and or2, with only two possibilities, allowing for right-handed and left-handed coordinate systems. From the candidate orientation matrix, the positions of all the expected reflections can be devised and compared to the positions of the experimental peaks on the diffraction pattern. The experimental pattern is indexed if the number of experimental peaks indexed is higher than a defined minimum.

3.4. Step 4: post-calculations

Once a satisfactory indexing is obtained, the successful orientation matrix is refined using the entire set of indexed reflections, and the number of matching reflections is calculated again. The output is a text file, either using XMAS indexing file format, or ShelX (Sheldrick, 2008) format. For each reflection in the diffraction pattern, the Miller indices (h, k, l), the integrated intensity of the experimental peak (I_{int}) and corresponding wavelength are given in the output.

3.5. Additional features

An option for indexing several orientations (several crystal grains) in a single frame is available. This is performed sequentially as implemented in the XMAS software (Tamura, 2014). A maximum number of orientations to look for per frame is provided, and the algorithm simply

loops over the previously described steps 1-4, removing each time the newly indexed peaks from the experimental peak list.

A second option has been introduced, taking into account possible disorientation of the crystal. This is for example the case when the crystal is slightly rotated while taking an exposure. As a result, additional reflections will be measured. This rotation effect can be specified by introducing a rotation angle. This option will not be discussed further in this paper.

4. Results and discussion

4.1. Indexing results

The indexing algorithm has been implemented in Fortran 90, using existing routines from the XMAS software (Tamura, 2014). Indexing trials have been carried out on a DELL Optiplex 9020 computer equipped with a 3.6GHz Intel Core i7-4790 processor. The indexing results for the datasets presented in Table 2 are shown in Table 3. Four indexing strategies have been tested, using either *d*-spacing (dsp) or structure factor (strf) ranking of the theoretical reflections (see 3.1. Pre-calculations part), and either pair (pm) or triplet (tm) matching when selecting candidates for the or1 vector (see 3.3. Global matching part). The percentage of successfully indexed patterns as well as the average time per pattern processed are given in each case. The complete set of parameters used for the indexing of each dataset is given in the Table SI1. The set of parameters necessary to obtain a successful indexing may not be unique, and parameter values different from the ones indicated in this paper may also provide a successful indexing. Moreover, the total indexing time will vary depending on the computing machine used.

A key step is to find the successful candidate to build the first vector or1 of the orientation matrix (step3). A first selection is done through nearest neighbor matching (step 2), and a second one at step 3 through pair or triplet matching. In order to have a successful indexing, the first requirement is to have the correct solution as candidate in the selection list. Then, this candidate should also be among the first to be checked. Selecting a large number of candidates may increase the indexing success rate, but also the processing time if the solution is ranked too far down in the list. The 15 patterns of the ZSM-5 sample collected using setup (4) and indexed using a *d*-spacing ranking strategy and triplet matching strategy to obtain or1 candidates (dsp-tm) can be used as an example. For each of the 15 patterns in the dataset, the processing time as a function of the ranking of the successful candidate has been plotted in Fig. 4. One pattern could not be indexed, probably due to the too limited number of possible candidates generated (15). In all the other cases, even if the indexing was successful, a fast indexing could only be achieved with the solution occupying one of the first two positions in the selection list.

A few parameters have a strong influence on the ranking of the or1 candidates, and consequently on the indexing time and indexing success. A detailed discussion of these parameters is given in SI1 and SI2, using the tests performed on ZSM-5 as examples. The requirements are different depending on whether small bandpass (monochr. or 4%) or larger bandpass (50%, DET-50/60, DET90) data are being indexed. In the latter case, a subset of experimental and theoretical reflections/q-vectors should be considered at step 2, when for small bandpass data, it is recommended to exploit as many experimental data as possible. Two main limitations have been identified. On the small bandpass side, the indexing may fail if not enough data per frame are present (e.g. CsPt monochr. data, with in average less than 10 reflections per frame). On the other hand, on the large bandpass side, the number of required expected reflections may increase drastically, preventing the indexing to be successful in a reasonable amount of time (e.g. ZSM-5 DET-90 data using *d*-spacing ranking).

4.2. Indexing modes

Four possible indexing strategies are available, depending on the theoretical reflections ranking mode (*d*-spacing- or structure factor-based) and on the selection mode of the first vector or1 of the orientation matrix (pair matching or triplet matching). The efficiency of these four indexing strategies will be discussed next for the various samples and setups.

As the indexing process is based on matching experimental q-vectors with theoretical ones, we expected that the ranking of theoretical reflections by structure factor would provide the best results. Indeed, this ranking mode has favored the indexing of the datasets of ZrPOF, ZSM-5 and MgAc collected with a large bandpass beam (50pc, DET-50 and DET90) (Table 3). For the datasets of CsPt and Sanidine collected with similar setups, this trend is less clear, and similar or even better results have been obtained using *d*-spacing ranking (Table 3). Reflection intensities from a single Laue diffraction pattern are usually fully measured (except for the reflections lying at the ends of the energy range), and even when affected by the presence of harmonics (reflections of different orders overlapping), our results show that this does not hinder a good match and the patterns can be indexed. The fact that a *d*-spacing ranking is also providing good results when indexing Sanidine and CsPt datasets is more difficult to interpret. We assume that, because fewer theoretical reflections are expected (smaller cell volume), the matching process is converging faster.

The results for datasets collected with a large bandpass beam show that the choice of using a pair matching or triplet matching strategy follows a binary distribution (Table 3). Indeed, in order to index the ZrPOF, ZSM-5 and MgAc datasets, the triplet matching process associated to structure factor ranking gives the best results. On the other hand, when indexing Sanidine and CsPt, similar results are obtained using either pair matching or triplet matching options. By imposing a matching among three q-vectors, the triplet matching process provides a higher degree of discrimination among potential candidates to build the first vector of the orientation matrix, which seems to be what is required to successfully index samples with larger cell volumes. Such a degree of discrimination appears less crucial to index samples with smaller cells, and both pair matching and triplet matching approaches can give acceptable results.

In the case of the indexing of datasets obtained using a smaller bandpass beam (monochr. and 4%), both structure factor and *d*-spacing ranking can be used, with nevertheless slightly better results with the second option (Table 3). When using a small bandpass beam, mainly partial reflection intensities are measured, and this may affect the intensity ranking of the experimental peaks. Consequently, the matching process will not be strongly affected by the chosen ranking strategy of theoretical reflections. On the other hand, the triplet matching option seems to be much less efficient than the pair matching one to index such small bandpass data. As previously mentioned, triplet matching requires a match between three experimental/theoretical q-vectors, this requirements being more difficult to achieve when fewer reflections/q-vectors are available. As less data per frame are measured with a small bandpass beam and with less theoretical reflections expected, a pair matching strategy is giving better results.

When searching for the most appropriate indexing strategies, three main tendencies emerge. These main indexing modes are presented in Fig. 5. For small bandpass data, both ranking by *d*-spacing or by structure factor can be used, combined to a pair matching approach. For larger bandpass data, there are two main strategies, depending on the dimension of the sample cell volume. In the case of a large cell, combining structure factor ranking with triplet matching ensure good indexing results. For smaller cell, any strategy can be used.

4.3. Alternative ranking of theoretical reflections

As shown in the previous section, the theoretical reflections ranking strategy plays an important role to obtain a successful indexing. However, the fact that a structure factor ranking seems to be required to index patterns measured with a large bandpass beam on large cell samples (Fig. 5) is an issue in the case of samples with unknown crystallographic structure. To cope with this, an alternative ranking strategy may be desirable. As mentioned when describing the pre-calculation part (step 1) of the indexing algorithm, a user-defined theoretical reflections list can be imposed, and we have tested the possibility of using the strongest reflection intensities extracted from a powder diffraction pattern. This approach has been tested on three datasets obtained with large bandpass beams: setup (4) (Laue DET-50) for ZrPOF and setup (5) (Laue DET-90) for both ZSM-5 and MgAc.

A Pawley refinement requires the cell parameters of a particular sample to be known, and this is indeed the case here, as it is also a requirement for the indexing algorithm. The refinements of the powder patterns measured on ZrPOF, ZSM-5 and MgAc are shown in Fig. 6. The resolution at which the powder diffraction signal vanishes (d_{max} powder) as well as the three resolution limits (lowest *d*-spacing, d_{min} single-crystal) of the three relevant single-crystal datasets are indicated. We can see that single-crystal data for ZSM-5 (Fig. 6b) and MgAc (Fig. 6c) mainly cover high 2θ range (low *d*-spacing), as imposed by the DET-90 configuration. This is also true for ZrPOF (Fig. 6a), even if the DET-50 configuration allows higher *d*-spacing reflections to be measured. A resolution of 1 Å (23° 2θ), 0.665 Å (35° 2θ) and 0.888 Å (26° 2θ) has been reached with powder data for ZrPOF, ZSM-5 and MgAc, respectively, which is still far away from the resolution obtained with single-crystal data (0.313 Å, 0.230 Å and 0.229 Å, respectively, Table 2).

The indexing results using as theoretical reflections the strongest reflection intensities extracted from the refined powder patterns (*p_int*) are shown in Table 4. For comparison, results obtained using *d*-spacing ranking and structure factor ranking (Table 3) have also been reported. The complete set of parameters to index the three datasets can be found in supplementary information. The number of unique reflections chosen at step 2 of the indexing process corresponds to the number of unique reflections extracted from the powder patterns, ranked in decreasing intensities. Following the indications given in Fig. 5, only the triplet matching strategy (step 3) has been used. We can see that the indexing using powder diffraction intensities as theoretical reflections gives intermediate results, with a better score than using the *d*-spacing ranking method, but still not as good as when using the structure factor ranking method. Using reflections ranked by decreasing intensities obtained from a powder pattern or reflections ranked by decreasing structure factors calculated from a known structure should give similar results. This is not the case yet, which means that the intensities extracted from the powder patterns may not be fully accurate. We attribute this to the low diffraction signal and the strong overlapping of reflections in the three relevant 2θ ranges (mainly high 2θ), preventing an optimal measure of the integrated intensities. Nevertheless, as we were looking for an improvement of the indexing score when a structure factor ranking cannot be used, the powder ranking strategy is indeed providing better results.

4.4. Indexing efficiency

A good indication of the appropriate indexing method to choose depending on the bandpass of the beam and on the volume of the unit cell of a particular sample has been given in Fig. 5. However, using only these two parameters may be a bit restrictive. In an attempt to **better assess the results, the different datasets used in the present study have been ranked depending on their "complexity". To do so, we have identified three main parameters that may play a significant role: the volume of the crystallographic unit cell (*Volume*), the number of unique reflections expected in the relevant energy range (*Unique refl.*), and the number of actually measured reflections per frame. If the first parameter is only related to the dimension of the crystallographic cell of the sample, the second one is linked to its symmetry and to the experimental setup (e.g. the energy range). The first two parameters can be calculated for a given sample and a particular setup. On the other hand, the third parameter is less predictable, and may fluctuate depending for instance on the brilliance of the incident beam, the quality of the crystal, or possible radiation damage. As a result, a "complexity" parameter *comp* has been calculated as follow: *comp = Observe refl. / Unique refl. / Volume*, with *Observe refl.* being the average number of measured reflections per frame for a particular dataset.**

The indexing is considered to be successful when a maximum number of frames are indexed in a minimum of time. Within the average indexing time per frame (Table 3), the time spent to successfully index a frame may be much shorter than for a non-successful indexing. To take that into account, an indexing efficiency coefficient *Eff* has been calculated:

$$Eff = t_N * (N_{success} * t_{success}) / (N_{tot} * t_{tot})$$

with $N_{success}$ being the number of patterns successfully indexed, $t_{success}$ the time spend to successfully index the patterns, N_{tot} the total number of patterns in the dataset, and t_{tot} the total time taken to index the dataset. In order to obtain a meaningful comparison between the different datasets, a time normalization t_N has been introduced:

$$t_N = -0.001 * Av\_time + 1$$

with *Av_time* the average indexing time per frame for a particular dataset, given in Table 3. In such a way, an indexing efficiency of 1 (best efficiency) and an efficiency of 0 (worse efficiency) can only be reached for an average indexing time of 0s per frame, and of 1000s per frame, respectively.

Complexity values obtained for each datasets as well as the indexing efficiency coefficients for the different indexing methods are reported in Table 5. The indexing efficiency as a function of sample complexity has been plotted in Fig. 7. In the case of hopeless indexing (e.g. ZSM-5, DET90, dsp-pm), the efficiency coefficient has been set to 0.

In Table 5, the datasets have been ranked by decreasing complexity, the smallest *comp* value corresponding to the higher degree of complexity. With such classification, the datasets collected on ZSM-5 crystals with a monochromatic beam and on MgAc with a large bandpass beam appear to be most complex, when the one obtained on CsPt using the DET-50 configuration the least. Indeed, indexing of the ZSM-5 monochromatic dataset and of the MgAc DET-90 dataset have been demanding, with a highest score below 90% (Table 3), no matter the indexing method used. On the other hand, the CsPt DET-50 dataset is part of these datasets that can be easily indexed, with any of the methods chosen (Table 3). This shows that the complexity of a dataset is not correlated to an increase of the bandpass of the beam.

We have chosen a *comp* value of 6.10^{-6} to separate the datasets into two categories (see the vertical dash line in Fig. 7). When looking at the less complex datasets (comp > 6.10^{-6}), an efficiency coefficient higher than 80% is achieved most of the time, no matter the indexing method. The only exception concerns the CsPt monochromatic dataset, with a drop of the indexing efficiency when using the triplet matching method, most probably due to a lack of data per frame (Table 2), as mentioned earlier. For the most complex datasets (*comp* < 6.10^{-6}), the indexing efficiency is clearly dropping, irrespective of the indexing methods used. One of the most affected is the dsp-tm method, with an efficiency never reaching 80%. On the other hand, the structure factor ranking methods are the most robust, in agreement with the results shown in Fig. 5.

5. Conclusion

The indexing algorithm presented in the current paper has been tested on a series of datasets obtained from five different samples with variable experimental conditions. Four indexing strategies can be used, with the calculated theoretical reflections ranked by *d*-spacing or by structure factors, and the matching process based on pair or triplet of q-vectors. The main parameters to tweak and the best indexing mode to choose to obtain a successful indexing differ depending on whether small bandpass (monochr. or 4%) or larger bandpass (50%, DET-50, DET90) are considered. The calculation of a complexity parameter for each dataset reveals that the most complex datasets are not simply correlated to a particular bandpass, and that the most robust indexing methods are the ones based on structure factor ranking. An additional feature has been added, allowing a user-defined theoretical reflection list to be provided. This is particularly useful when only the lattice parameters of a sample are known, and a *d*-spacing ranking strategy has to be used. Using the reflection intensities extracted from a powder patterns have shown that the indexing can indeed be improved. The idea of using crystallographic information coming from powder diffraction to index serial crystallography data has been proposed previously, the powder pattern being in that case directly build from single crystal data (Brewster *et al.*, 2015). Combining methods and practices from different communities is always a good approach to solve challenging crystallographic problems.

The indexing program can be downloaded at https://sites.google.com/a/lbl.gov/bl12-3-2/user-resources/.

Figure 1 Incident flux for the four non-monochromatic setups extracted from sanidine single-crystal data using the reverse method (Dejoie *et al.*, 2011). Setup 2: 4% bandpass; setup 3: 10-17 keV range (50% bandpass); setup 4 and 5: 5-24 keV range.

Figure 2 a) Schematic representation of an experimental setup using a geometry in reflection; b) Ewald construction for non-monochromatic (Laue) diffraction. Nodes of the reciprocal lattice in the blueish zone are in diffraction condition. Two normalized q-vectors, q_1 and q_2, corresponding to the h_1 k_1 l_1 and the h_2 k_2 l_2 reflections, respectively, are shown. The wavelength of these two reflections, defining the length of the q-vectors, is not known *a priori*, and lies on the solid line crossing the reflections. Note that the reflection h_1 k_1 l_1 has a harmonic reflection (h_{1n} k_1n l_{1n}).

Figure 3 Indexing algorithm. a) Step 2, nearest neighbors matching. The red, green and blue cones represent the limiting maximum threshold angle around the experimental q-vectors 1, 2 and 3, respectively. On the experimental side, 6, 4 and 4 nearest neighbor angles can be calculated around the q-vectors 1, 2, and 3, respectively. On the theoretical side, four q-vectors having similar nearest neighbors as the first experimental q-vector are shown. The theoretical q-vectors m, n, o, and p have 5, 5, 3 and 4 similar neighbors (purple arrows) with $q_{exp\_start}(1)$, respectively. Additional q-vectors (dashed black arrows) may also be present; b) Step 3, pair matching. If two experimental q-vectors form an angle α, then, the same applied for theoretical q-vectors.

Among the 4 potential candidates with similar nearest neighbors as the first experimental reflection, only two ($q_{th\_match}(1,1)$ and $q_{th\_match}(1,3)$) fulfil the requirement; c) Step 3, triplet matching. Three q-vectors are involved in the matching process, and only one theoretical q-vector ($q_{th\_match}(1,1)$) fulfils the requirement.

Figure 4 Indexing of the dataset collected on ZSM-5 using the DET-60 configuration and the dsp-tm strategy, showing the indexing time as a function of the position of the solution within the candidate list to determine the or1 vector of the orientation matrix.

Figure 5 Schematic representation of the most successful indexing modes depending on the volume of the crystallographic cell and on the energy bandpass of the beam.

Figure 6 Pawley refinement of a) ZrPOF (Rp=4.2%, Rwp=6.6%, Rexp=1.4%), b) ZSM-5 (Rp=5.9%, Rwp=9.8%, Rexp=1.3%) and c) MgAc (Rp=4.1%, Rwp=5.9%, Rexp=0.7%).

Figure 7 Efficiency of the indexing methods as a function of the complexity of the datasets (dsp: d-spacing ranking; strf: structure factor ranking; pm: pair matching; tm: triplet matching).

Table 1 Main crystallographic information for the samples of the present study (SG: space group).

Name	Formula	SG	Volume (Å^3)	a (Å)	b (Å)	c (Å)	α (°)	β (°)	γ (°)
Sanidine	$KAlSi_3O_8$	C2/m	721.79	8.58320	13.0076	7.1943	90	116.023	90
CsPt	$Cs_2[Pt(CN)_4].H_2O$	$P6_5$	1619.73	9.791	9.791	19.510	90	90	120
ZrPOF	$\vert(C_9H_8N)_4(H_2O)_4\vert[Zr_8P_{12}O_{40}(OH)_8F_8]$	$P\bar{1}$	1977.53	10.7567	13.8502	14.8995	109.6	101.1	100.5
ZSM-5	$(SiO_2)_{96}$	Pnma	5343.32	20.022	19.899	13.383	90	90	90
MgAc	$Mg_5(C_2H_3O_2)_8(OH)_2$	$I4_1/a$	6640.68	23.3126	23.3126	11.9855	90	90	90

Table 2 Single crystal datasets collected for the five samples of this study, using five different setups ((1) to (5)) – The number of diffraction patterns per dataset, the resolution range (d-spacing), the average number of experimental peaks per frame, and the number of independent reflections expected in the resolution range are indicated.

	Dataset name	Monochr.	4%	50%	LaueDET-50/60	LaueDET-90
	Energy range (keV)	17.75	16.9-17.7	10-17	5-24	5-24
	Setup no.	(1)	(2)	(3)	(4)	(5)
Sanidine	No. patterns		100	85	80	10
	Resolution range (Å-1)		Inf-0.450	3.599-0.406	5.872-0.312	1.563-0.227
	Average no. peaks / frame		16	67	98	201
	No. independent reflections		4337	5809	12674	32313
CsPt	No. patterns	100	100		12	
	Resolution range (Å-1)	Inf-0.574	Inf-0.560		5.950-0.315	
	Average no. peaks / frame	9	16		353	
	No. independent reflections	523	551		2715	
ZrPOF	No. patterns	21			17	
	Resolution range (Å-1)	0-12.311			1.069-20.085	
	Average no. peaks / frame	34			170	
	No. independent reflections	7767			33884	
ZSM-5	No. patterns	100	100	17	15	19
	Resolution range (Å-1)	Inf-0.552	Inf-0.559	3.471-0.387	4.155-0.287	1.685-0.230
	Average no. peaks / frame	12	26	148	472	517
	No. independent reflections	4604	4427	12884	31063	59358
MgAc	No. patterns			29		37
	Resolution range (Å-1)			3.599-0.406		1.670-0.229
	Average no. peaks / frame			118		213
	No. independent reflections			7150		37989

Table 3 Indexing results for the different datasets tested (dsp: d-spacing ranking; strf: structure factor ranking; pm: pair matching; tm: triplet matching). For each indexing option,

the percentage of successfully indexed frames and the average indexing time per pattern (s) is given.

	Dataset name	Monochr.	4%	50%	LaueDET-50/60	LaueDET-90
	Energy range (keV)	17.75	16.9-17.7	10-17	5-24	5-24
	Setup no.	(1)	(2)	(3)	(4)	(5)
Sanidine	dsp-pm		96, 13.2	100, 0.7	100, 0.7	100, 1.6
	dsp-tm		65, 25.1	100, 1.8	100, 1.1	100, 2.0
	strf-pm		99, 12.4	100, 0.7	100, 0.7	100, 1.3
	strf-tm		79, 25.1	100, 1.5	100, 1.4	100, 0.8
CsPt	dsp-pm	87, 3.0	100, 2.8		100, 1.4	
	dsp-tm	55, 6.3	93, 6.8		100, 1.6	
	strf-pm	85, 3.2	97, 2.8		100, 5.3	
	strf-tm	55, 6.2	89, 5.8		100, 5.4	
ZrPOF	dsp-pm	75, 38.8 – 53, 19.4**			76, 432	
	dsp-tm	40, 191.1 – 20, 95.6**			76, 78	
	strf-pm	95, 31.3 – 80, 15.7**			76, 33.5	
	strf-tm	40, 178.2 – 28, 89.1**			82, 24.9	
ZSM-5	dsp-pm	85, 4.8	98, 8.6	-	73, 378.5	-
	dsp-tm	37, 15.5	78, 85.7	65, 301.7	93, 56.7	79, 672.5
	strf-pm	82, 5.2	92, 8.2	100, 17.4	100, 8.4	95, 68.0
	strf-tm	42, 15.2	71, 108.2	100, 3.6	100, 7.2	100, 35.7
MgAc	dsp-pm			-		-
	dsp-tm			90, 138.0		57, 550.0
	strf-pm			90, 72.6		65, 271.1
	strf-tm			97, 10.2		86, 108.4

* For ZSM-5, the solutions where the a and b axes are reversed were accepted as correct (the flipped solution can be checked in an additional step, and correct indexing is usually the one where more reflections are indexed)

** For the monochromatic dataset of ZrPOF, the algorithm is looking for 2 orientations in each patterns sequentially. The first two numbers correspond to the percentage of success if at least one orientation per pattern was indexed and the indexing time per pattern (20 in total), and the last two numbers to the percentage of success per orientation indexed and the indexing time per orientation (40 in total)

Table 4 Indexing results using reflection intensities extracted from a powder pattern (dsp: d-spacing ranking; strf: structure factor ranking; p_int: powder diffraction intensities ranking; tm: triplet matching).

Name	Indexing method	Laue datasets
ZrPOF	dsp-tm	76, 78.0
	p_int-tm	82, 50.8
	strf-tm	82, 24.9
ZSM-5	dsp-tm	79, 672.5
	p_int-tm	84, 50.6
	strf-tm	100, 35.7
MgAc	dsp-tm	57, 550.0
	p_int-tm	76, 169.8
	strf-tm	86, 108.4

Table 5 Efficiency of the four indexing modes (*Eff1* to *Eff4*) depending on the complexity (*Comp*) of the datasets.

Sample	Dataset	Comp (*10^{-6})	Eff1	Eff2	Eff3	Eff4
			dsp-pm	dsp-tm	strf-pm	strf-tm
ZSM5	Monochr.	0.50	0.758	0.166	0.692	0.243
MgAc	DET-90	0.85	0	0.091	0.260	0.621
ZSM5	4%	1.11	0.946	0.546	0.765	0.469
ZSM5	DET-90	1.63	0	0.206	0.719	0.964

ZSM5	50%	2.16	0	0.143	0.983	0.996
ZrPOF	Monochr.	2.25	0.566	0.115	0.869	0.141
MgAc	50%	2.48	0	0.546	0.642	0.900
ZrPOF	DET-50	2.53	0.194	0.266	0.416	0.457
ZSM5	DET-60	2.84	0.155	0.746	0.992	0.993
San	4%	4.95	0.932	0.477	0.973	0.656
San	DET-90	8.62	0.998	0.998	0.999	0.998
San	DET-50	10.68	0.999	0.999	0.999	0.999
CsPt	Monochr.	10.90	0.809	0.370	0.778	0.401
San	50%	16.01	0.999	0.998	0.999	0.998
CsPt	4%	18.00	0.997	0.865	0.959	0.833
CsPt	DET-50	80.21	0.999	0.998	0.995	0.995

cknowledgements The authors would like to thank Dr. L.B. McCusker and Dr. C. Baerlocher (ETH Zurich, Switzerland), Dr. P. Pattison (EPFL Lausanne, Switzerland and SNBL beamline, ESRF, France) and Dr. A. N. Fitch (ID22 beamline, ESRF, France), for their support and their advices. The Advanced Light Source at the Lawrence Berkeley National Laboratory is supported by the Office of Science, Office of Basic Energy Sciences, Scientific User Facilities Division of the US Department of Energy under contract No. DE-AC02-05CH11231.

References

Ackermann, S., Kunz, M., Armbruster, T., Schefer, J. & Hänni, H. (2004). *Schweiz. Miner. Petro. Mitt.* **84**, 345–354.

Bourgeois, D., Vallone, B., Schotte, F., Arcovito, A., Miele, A. E., Sciara, G., Wulff, M., Anfinrud, P. & Brunori, M. (2003). *Proc. Natl Acad. Sci. USA*, **100**, 8704–8709.

Boutet, S. *et al.* (2012). *Science*, **337**, 362–364.

Brewster, A. S., Sawaya, M. R., Rodriguez, J., Hattne, J., Echols, N., McFarlane, H. T., Cascio, D., Adams, P. D., Eisenberg, D. S. & Sauter, N. K. (2015). *Acta Cryst.* D**71**, 357–366.

Chapman, H. *et al.* (2011). *Nature*, **470**, 73–77.

Chen, X., Dejoie, C., Jiang, T., Ku, C.S. & Tamura, N. (2016). *MRS Bulletin*, **41**, 1-20.

Chung, J. S. & Ice, G. E. (1999). *J. Appl. Phys.* **86, 5249–5255.**

Coelho A. A. J. (2018). *J. Appl. Cryst.* 51, 210-218.

Cornaby, S., Szebenyi, D. M. E., Smilgies, D.-M., Schuller, D. J., Gillilan, R., Hao, Q. & Bilderback, D. H. (2010). *Acta Cryst.* D**66**, 2–11.

Dejoie, C., Kunz, M., Tamura, N., Bousige, C., Chen, K., Teat, S., Beavers, C. & Baerlocher, C. (2011). *J. Appl. Cryst.* **44**, 177–183.

Dejoie, C., McCusker, L. B., Baerlocher, C., Abela, R., Patterson, B. D., Kunz, M. & Tamura, N. (2013). *J. Appl. Cryst.* **46, 791–794.**

Dejoie, C., McCusker, L. B., Baerlocher, C., Kunz, M. & Tamura, N. (2013). *J. Appl. Cryst.* 46, 1805–1816.

Dejoie, C., Smeets, S., Baerlocher, C., Tamura, N., Pattison, P., Abela, R. & McCusker, L. B. (2015). *IUCrJ.* **2, 361–370.**

Gevorkov, Y., Barty, A., Brehm, W., White, T. A., Tolstikova, A., Wiedorn, M. O., Meents, A., Grigat, R. R., Chapman, H. N. & Yefanov, O. (2020). *Acta Cryst. A*76, 121-131.

Ginn, H. M., Roedig, P., Kuo, A., Evans, G., Sauter, N. K., Ernst, O. P., Meents, A., Mueller-Werkmeister, H., Miller, R. J. D. & Stuart, D. I. (2016). *Acta Cryst. D*72, 956-965.

Hattne, J. *et al.* (2014). *Nat. Methods*, **11**, 545–548.

Helliwell, J. R., Habash, J., Cruickshank, D. W. J., Harding, M. M., Greenhough, T. J., Campbell, J. W., Clifton, I. J., Elder, M., Machin, P. A., Papiz, M. Z. & Zurek, S. (1989). *J. Appl. Cryst.* **22**, 483–497.

Johansson, L.C., Stauch, B., Ishchenko, A. & Cherezov, V. (2017). *Trends Biochem. Sci.* 42, 749-762.

Johnson, P. L., Koch, T. R. & Williams, J. M. (1977). *Acta Cryst.* B**33, 1293–1295.**

Kabsch, W. (2014). *Acta Cryst. D*70, 2204-2216.

Kunz, M. *et al.* (2009). *Rev. Sci. Instrum.* **80**, 035108.

Li, C., Li, X., Kirian, R., Spence, J. C. H., Liu, H. & Zatsepin, N. A.(2019). *IUCrJ.* **6**, 72-84.

Liu, L., Li, J., Dong, J., Sisak, D., Baerlocher, C. & McCusker L. B. (2009). *Inorg. Chem.* 48, 8947-8954.

Liu, H. & Spence C. H. (2016). *Quant. Biol.* 4, 159-176.

Moffat, K. & Helliwell J.R. (1989). Synchrotron Radiation in Chemistry and Biology III, p. 61.

Olson, D. H., Kokotailo, G. T., Lawton, S. L. & Meier, W. M. (1981). *J. Phys. Chem.* **85, 2238–2243.**

Patterson, B. D., Beaud, P., Braun, H. H., Dejoie, C., Ingold, G., Milne, C., Patthey, L., Pedrini, B., Szlachentko, J. & Abela, R. (2014). *CHIMIA Int. J. Chem.* **68**, 73–78.

Scheurell, K., Troyanov, S.I. & Kemnitz, E. (2015). *Z. Anorg. Allg. Chem.* 641, 1106-1109.

Sheldrick, G. M. (2008). *Acta Cryst.* A**64**, 112–122.

Tamura, N., MacDowell, A. A., Spolenak, R., Valek, B. C., Bravman, J. C., Brown, W. L., Celestre, R. S., Padmore, H. A., Batterman, B. W. & Patel, J. R. (2003). *J. Synchrotron Rad.* **10**, 137–143.

Tamura, N. (2014). *XMAS: a versatile tool for analyzing synchrotron X-ray microdiffraction data* **In** *Strain and Dislocation Gradients from Diffraction. Spatially Resolved Local Structure and Defects*, **edited by R. Barabash & G. Ice, pp. 125–155. London: Imperial College Press.**

Van Koningsveld, H., van Bekkum, H. & Jansen, J. C. (1987). *Acta Cryst.* B**43**, 127–132.

Van Wamelen, P., Li, Z. & Iyengar, S. (2004). *Pattern Recognit.* **37, 1699–1711.**

White, T.A., Mariani, V., Brehm, W., Yefanov, O., Barty, A., Beyerlein, K.R., Chervinskii, F., Galli, L., Gati, C., Nakane, T., Tolstikova, A., Yamashita, K., Yoon, C.H., Diederichs, K. & Chapman, H.N. (2016). ***J. Appl. Cryst.*** **49, 680-689.**

Yorke, B. A., Beddard, G. S., Owen, R. L. & Pearson, A. R. (2014). *Nat. Methods*, **11**, 1131–1134.

Supporting information

S1. User-defined parameters

n_exp_peaks: total number of experimental peaks (imposed by the data)

n_th_refl: total number of theoretical reflections (imposed by the unit cell and space group of the sample)

<u>Indexing process, step 2:</u>

exp_start_set: first set of experimental peaks/q-vectors (*exp_start_set* ≤ *n_exp_peaks*)

exp_set: second set of experimental peaks/q-vectors (*exp_start_set* ≤ *exp_set* ≤ *n_exp_peaks*)

th_set_uniques: set of theoretical unique reflections/q-vectors (*th_set_uniques* ≤ *n_th_refl*)

th_set: set of theoretical reflections/q-vectors (*th_set_uniques* < *th_set* ≤ *n_th_refl*)

ineid: angle defining the first vicinity/neighbourhood around experimental and theoretical q-vectors

dmax-dmin: limiting resolution range for nearest neighbor matching

<u>Indexing process, step 3:</u>

thresh_neighbors: for each experimental q-vectors, number of candidates unique q-vectors from the "Nearest neighbors list" to consider for pair matching or triplet matching

ang_dev_or1: angle deviation acceptance to select candidates for the first vector of the orientation matrix or1

ang_dev_or2: angle deviation acceptance to select candidates for the first vector of the orientation matrix or2

ang_dev: deviation angle acceptance during the final matching process

min_ind_peak: minimum number of experimental peaks to be indexed to consider the indexing successful

S2. Description of the main indexing parameters

The role of the main parameters are discussed in the next few paragraphs, using the tests performed on ZSM-5 as examples. Even if all these parameters are user-defined, default values have been identified and can be used depending on the beam bandpass and the cell volume of the sample (see input files examples in https://sites.google.com/a/lbl.gov/bl12-3-2/user-resources/).

The first parameters to consider are the number of experimental q-vectors *exp_start_set* and *exp_set* parameters. Both parameters play an important role in the nearest neighbors matching process (step 2). In step 3, *exp_start_set* and *exp_set* also have an influence on the selection of the candidates for the or1 and or2 vectors, respectively. In the case of large energy bandpass data (10-17 keV and 5-24keV, setup (3), (4) and (5)), using all the experimental q-vectors obtained from all the peaks found in a diffraction pattern did not give appropriate results, both in term of indexing time and success, and the use of a subset is preferable. After a series of tests, for most samples, the best indexing results were obtained with a starting set (*exp_start_set*) of 8 to 15 experimental q-vectors. ZrPOF is the only sample (DET-50 datasets) for which a larger starting set had to be selected (15 to 20 peaks). Concerning the choice of *exp_set*, which corresponds to the number of experimental q-vectors taking part to the neighborhood calculation (step 2), 40 and 50 have been the values commonly used. The choice of the *exp_start_set* parameter can greatly influence the indexing process, and increasing or decreasing its value by one or two elements can sometimes make a huge difference. On the other hand, the *exp_set* parameter was found to be less critical, as long as enough data are taken into account (a value below 30 may not be recommended). In the case of smaller bandpass data (monochromatic and 4%, setup (1) and (2)), the full experimental set has been used to define both *exp_start_set* and *exp_set*. Using subsets of data in that case only resulted in a decrease of the indexing success. In average, for small bandpass data, less than 30 experimental peaks per frame have been recorded (Table 2). Being slightly short in term of experimental data per frame means that the neighborhood of the q-vectors corresponding to the first brightest reflections may not be well defined. By using all available data in a frame and setting both *exp_start_set* and *exp_set* to *n_exp_peaks* (total number of experimental peaks/q-vectors), we increase the possibility to obtain a better nearest neighbor matching around at least one of the experimental q-vector, and the diffraction pattern is more likely to be indexed.

The second set of parameters (*th_set_uniques* and *th_set*) concerns the calculated theoretical reflections, to limit the number of theoretical q-vectors taken into account. In a similar way as *exp_start_set* and *exp_set* parameters, *th_set_uniques* and *th_set* parameters both have an influence on the nearest neighbors matching process, and later on the selection of the or1 and or2 vectors, respectively. The values of *th_set_uniques* and *th_set* used to index the datasets collected on ZSM-5 have been reported in Fig. S1a and S1b, respectively. For both parameters, there is a clear difference depending on the indexing method chosen and how the theoretical reflections have been ranked. Using structure factor ranking, the number of theoretical reflections needed to obtain a successful match tends to decrease when trying to index large bandpass data compared to small bandpass data. On the other hand, when theoretical

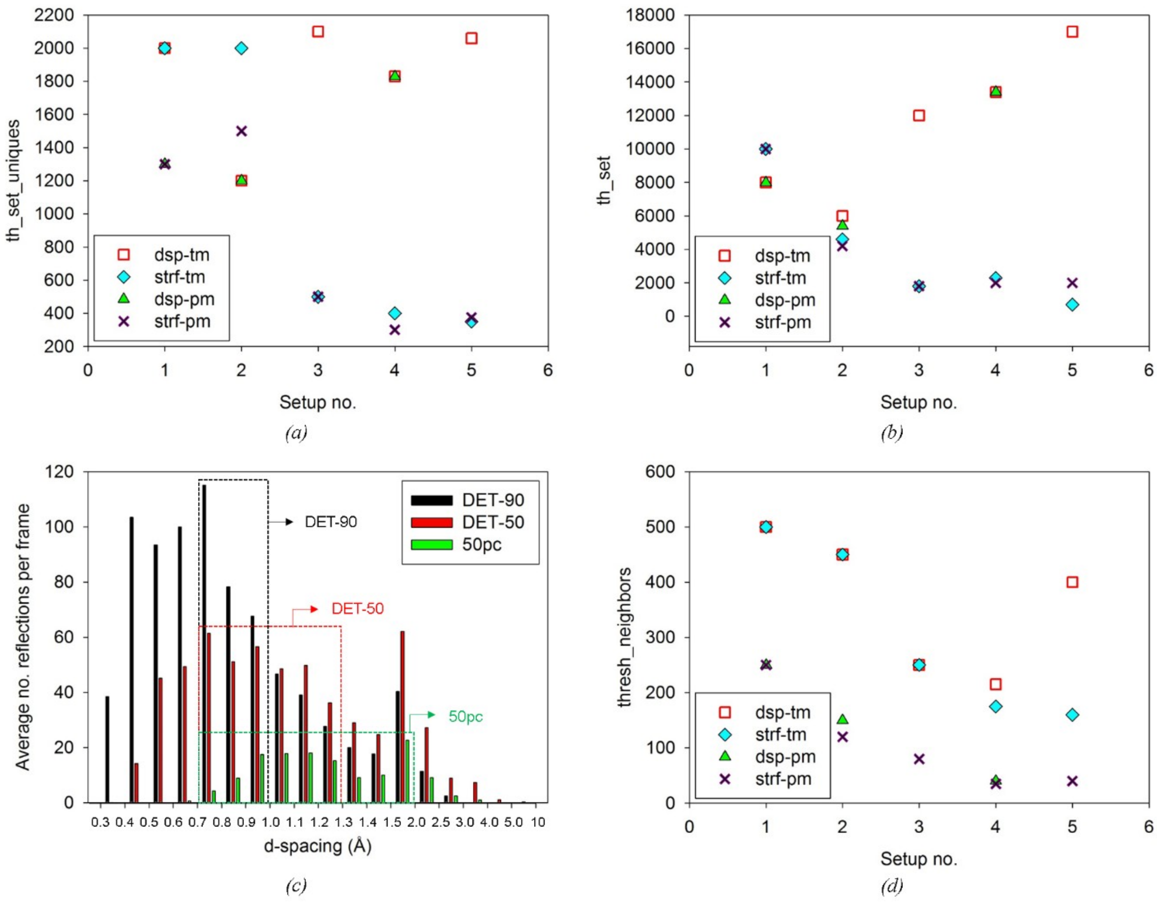
reflections are ranked by decreasing *d*-spacing, both *th_set_uniques* and *th_set* stay at a high value, with even an increase for *th_set*. In most of the cases tested in this project, for large bandpass datasets (setups (3), (4), (5)), a *d*-spacing ranking induces the use of more theoretical reflections (see Table S1). A ranking by structure factor is then more efficient to index large bandpass data. For small bandpass data, this effect is not so clear anymore.

Two additional parameters play an important role in the nearest neighbor matching process (step 2). The first one is the *ineid* parameter, which defines the angular limit to calculate angles with neighboring q-vectors of a particular q-vector. An angular value of 30° has been used for large bandpass data, and additional tests carried out with different values did not bring better results (not shown). A larger angular value, up to 50°, was found more adequate in the case of smaller bandpass datasets (see Table S1). This is correlated to the fact that less data per frame are available, as mentioned when discussing the *exp_start_set* and *exp_set* parameters. Increasing the angular range helps in defining the nearest neighbors around a particular q-vector. The second parameter to consider is the limited resolution range (*d*-spacing range) *dmax-dmin* in which the nearest neighbors matching is performed. If the full resolution range (imposed by the setup) can be used in the case of small bandpass data, using a restricted range for larger bandpass data helps to obtain a faster indexing. The average number of indexed reflections as a function of 18 different resolution ranges has been calculated, and the resulting plot for three ZSM-5 datasets (setups (3), (4) and (5)) is shown in Fig. S1c. In each case, the restricted resolution range used in the indexing process has been indicated. These chosen ranges correspond to those where the majority of the reflection lies. As the number of measured reflections decreases when going from DET-90 (setup (5)), to DET-60 (setup (4)) to 50% (setup (3)) configurations, the resolution range can be increased towards higher *d*-spacing (low *q*) data. Using only low-resolution data (d > 1.256 Å, or q < 5 Å^{-1}) did not give any good results, probably because not enough reflections are present for the nearest neighbors matching process in that range. On the other hand, too many reflections being expected on the high-resolution side (low *d*-spacing or high *q*), the matching process is becoming quite inefficient, and this explains why the use of a restricted range is necessary.

The last parameter to mention is *thresh_neighbors*, which corresponds, for each experimental q-vectors, to the number of unique q-vectors with the highest number of common neighbors (step 3). Its evolution has been plotted in Fig. S1d. The number of unique q-vectors to take into account globally decreases when going from small bandpass data to larger bandpass data. What is important to note is that the *thresh_neighbors* parameter clearly reflects the difference between pair matching and triplet matching procedures for the determination of or1

candidates. If a relatively small number can be chosen for pair matching, a higher number is always required when choosing triplet matching. The latter involves three q-vectors (instead of only two for pair matching), which means that enough possibilities have to exist in order to find potential candidates.

Table S1 **Complete set of parameters used for the indexing of each dataset (see excel file attached).**



(a)

(b)

(c)

(d)

Figure S1 **a) Number of theoretical unique reflections used to index ZSM-5 datasets as a function of the setup. A clear difference can be seen depending on the ranking strategy chosen. b) Number of theoretical reflections used to index ZSM-5 datasets as a function of the setup. As for the unique reflections, the behavior between *d*-spacing and structure factor ranking is different. c) Average number of reflections per resolution range (*d*-spacing) for three ZSM-5 datasets, displayed left-to-right for each resolution range in the following order:** setup 5 (5-24 keV energy range**, DET-90), setup 4 (5-24 keV energy range, DET-50), and setup 3 (10-17 keV energy range, 50pc). The dashed rectangle are indicating where the maximum number of reflection for a particular setup lies. d) Evolution of the *thresh_neighbors* parameter used to index ZSM-5 datasets as a function of the setup. This parameter reflects the difference between triplet matching and pair matching strategies (dsp: *d*-spacing ranking; strf: structure factor ranking; pm: pair matching; tm: triplet matching, setup 1:monochromatic beam; setup 2: 4% bandpass beam; setup 3: 10-17 keV energy range; setup 4: 5-24 keV energy range, DET-60; setup 5: 5-24 keV energy range, DET-90).**