Structure-mining: screening structure models by automated fitting to the atomic pair distribution function over large numbers of models

Long Yang,[a] Pavol Juhás,[b] Maxwell W. Terban,[c] Matthew G. Tucker[d] and Simon J. L. Billinge[a,e]*

[a]Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY 10027, USA, [b]Computational Science Initiative, Brookhaven National Laboratory, Upton, NY 11973, USA, [c]Max Planck Institute for Solid State Research, Heisenbergstrasse 1, 70569 Stuttgart, Germany, [d]Neutron Scattering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37830, USA, and [e]Condensed Matter Physics and Materials Science Department, Brookhaven National Laboratory, Upton, NY 11973, USA. *Correspondence e-mail: sb2896@columbia.edu

A new approach is presented to obtain candidate structures from atomic pair distribution function (PDF) data in a highly automated way. It fetches, from web-based structural databases, all the structures meeting the experimenter's search criteria and performs structure refinements on them without human intervention. It supports both X-ray and neutron PDFs. Tests on various material systems show the effectiveness and robustness of the algorithm in finding the correct atomic crystal structure. It works on crystalline and nanocrystalline materials including complex oxide nanoparticles and nanowires, low-symmetry and locally distorted structures, and complicated doped and magnetic materials. This approach could greatly reduce the traditional structure searching work and enable the possibility of high-throughput real-time auto-analysis PDF experiments in the future.

1. Introduction

The development of science and technology is built on advanced materials, and new materials lie at the heart of technological solutions to major global problems such as sustainable energy (Moskowitz, 2009). However, the discovery of new materials still needs a lot of labor and time. The idea behind materials genomics (White, 2012) is to develop collaborations between materials scientists, computer scientists and applied mathematicians to accelerate the development of new materials through the use of advanced computation such as artificial intelligence (AI), for example, by predicting undiscovered materials with interesting properties (Jain *et al.*, 2013; Simon *et al.*, 2015; Curtarolo *et al.*, 2013).

The study of material structure plays a key role in the development of novel materials. Structure solution of well-ordered crystals is largely a solved problem, but for real materials, which may be defective or nanostructured, being studied under real conditions, for example in high-throughput *in situ* and *operando* diffraction experiments such as *in situ* synthesis (Cravillon *et al.*, 2011; Jensen *et al.*, 2012; Friščić *et al.*, 2013; Saha *et al.*, 2014; Shoemaker *et al.*, 2014; Katsenis *et al.*, 2015; Olds *et al.*, 2017; Terban *et al.*, 2018), determining structure can be a major challenge that could itself benefit from a genomics-style approach. Here we explore a data-mining methodology for the determination of inorganic material structures. The approach can rapidly screen large numbers of structures in a manner that is well matched to the
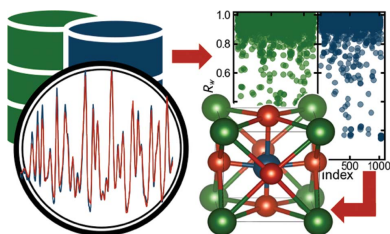
kinds of high-throughput experiments being envisaged in the materials genomics arena.

A number of structural databases are available for inorganic materials containing structures solved from experimental data such as the Inorganic Crystal Structure Database (ICSD) (Bergerhoff *et al.*, 1983; Belsky *et al.*, 2002), the American Mineralogist Crystal Structure Database (AMCSD) (Downs & Hall-Wallace, 2003), the Crystal Structure Database for Minerals (MINCRYST) (Chichagov *et al.*, 2001) and the Crystallography Open Database (COD) (Gražulis *et al.*, 2009). More recently, databases of theoretically predicted structures have begun to become available, such as the Materials Project Database (MPD) (Jain *et al.*, 2013), the Automatic Flow Library (AFLOWLIB) (Curtarolo *et al.*, 2012) and the Open Quantum Materials Database (OQMD) (Saal *et al.*, 2013; Kirklin *et al.*, 2015). Structural databases such as the International Centre for Diffraction Data (ICDD, 2019) have for some time been used for phase identification purposes. In phase identification studies no model fitting is carried out, but phases are identified in a powder diffraction pattern by matching sets of the strongest Bragg peaks from the database structures to peaks in the measured diffractogram (Hanawalt *et al.*, 1938; Marquart *et al.*, 1979; Gilmore *et al.*, 2004). Our goal is not just phase identification, but to speed up the process of finding structural candidates to unknown atomic pair distribution function (PDF) signals.

PDF analysis of X-ray and neutron powder diffraction data sets has been demonstrated to be an excellent tool for studying structures of many advanced materials, especially nanostructured materials (Zhang *et al.*, 2003; Neder & Korsunskiy, 2005; Masadeh *et al.*, 2007; Young & Goodwin, 2011; Beecher *et al.*, 2014; Terban *et al.*, 2017; Laveda *et al.*, 2018), but also bulk crystalline materials (Toby *et al.*, 1989; Billinge *et al.*, 1996; Billinge & Kanatzidis, 2004; Keen & Goodwin, 2015).

The PDF gives the scaled probability of finding two atoms in a material a distance r apart and is related to the density of atom pairs in the material. It does not presume periodicity so goes well beyond just well-ordered crystals (Egami & Billinge, 2012; Billinge, 2019). The experimental PDF, denoted $G(r)$, is the Q_{min} and Q_{max} truncated Fourier transform of powder diffraction data (Farrow & Billinge, 2009),

$$G(r) = \frac{2}{\pi} \int_{Q_{min}}^{Q_{max}} Q[S(Q) - 1] \sin(Qr) \, dQ, \qquad (1)$$

where Q is the magnitude of the scattering momentum. The total scattering structure function, $S(Q)$, is extracted from the Bragg and diffuse components of the X-ray, neutron or electron powder diffraction pattern.

$G(r)$ can be calculated from a given structure model (Egami & Billinge, 2012) and once the experimental PDFs are determined they can be analyzed through modeling. The PDF modeling is performed by adjusting the parameters of the structure model, such as the lattice parameters, atomic positions and atomic displacement parameters, to maximize the agreement between the calculated PDF from the structure model and the experimental PDF.

A number of PDF structure modeling programs are available for crystalline or nanocrystalline inorganic materials (Cranswick, 2008). Small box modeling programs use a small number of crystallographic parameters with a periodic structure model (Egami & Billinge, 2012). Three widely used examples are *pdfgui* (Farrow *et al.*, 2007), *TOPAS* (Coelho, 2018) and *diffpy-cmi* (Juhás *et al.*, 2015), among others (Petkov & Bakaltchev, 1990; Proffen & Billinge, 1999; Gagin *et al.*, 2014). Big box modeling programs, which move large numbers of atoms to minimize the difference between the observed and calculated PDFs, usually implement the reverse Monte Carlo (RMC) method (McGreevy & Pusztai, 1988; McGreevy, 2001), such as *RMCProfile* (Tucker *et al.*, 2007), *DISCUS* (Proffen & Neder, 1997; Page *et al.*, 2011) and *FullRMC* (Aoun, 2016). Other modeling programs use a hybrid approach where a large number of atoms are in the box, but the program refines only a small number of parameters, such as *EPSR* (Soper, 2005).

Though powerful for understanding the structure of complex materials, PDF modeling and structure refinement are difficult and present a steep learning curve for new users. There are two major challenges. The first is that PDF structure refinement requires a satisfactory plausible starting model to achieve a successful result. The second is that the refinement process is a nonlinear regression that is highly non-convex and generally requires significant user inputs to guide it to the best fit whilst avoiding overfitting. A more automated refinement program such as we propose here needs to address both issues.

Model selection traditionally requires significant chemical knowledge and experience, but can be quite challenging when unknown impurities or reaction products are present in the sample. To address the problem of phase identification, automated search–match algorithms for identifying phases in powder diffraction patterns have been developed and are widely used (Hanawalt *et al.*, 1938; Marquart *et al.*, 1979; Gilmore *et al.*, 2004). There are also programs for helping find candidate structures from structural databases (Barr *et al.*, 2004; Toby, 2005; Altomare *et al.*, 2008, 2015; Degen *et al.*, 2014). These search–match programs only work for reciprocal-space diffraction patterns, and in general do not allow for automated refinement of structures. Some attempts have been made to couple Rietveld refinement programs to structural databases such as *Full Profile Search Match* (Boullay *et al.*, 2014; Lutterotti *et al.*, 2019), though this is limited to refining structures from the COD database. Alternatively, programs that use scripting such as *TOPAS* (Coelho, 2018) have been used to automatically refine large numbers of candidate structures generated by symmetry-mode analysis from a given high-symmetry starting structure (Lewis *et al.*, 2016). Furthermore, a structure screening approach where large numbers of algorithmically generated small metal nanoparticle models were compared with PDF data was recently demonstrated (Banerjee *et al.*, 2020). This approach, called cluster-mining, was successful at obtaining significantly improved fits over standard approaches to nanoparticle PDF data from simple models with a small number of refinable

parameters. It also returned multiple plausible and well-performing structures rather than just one best-fit structure, allowing the user to choose a model based on more information than just the PDF data. We would like to combine these approaches (database searching, auto-refinement and screening of large numbers of structures) to the modeling of PDF data in general.

Here we describe an approach we call structure-mining to automate and manage structure model selection and PDF refinement. To make the whole procedure as high-throughput and automatic as possible, the required user inputs are kept to a minimum: simply the experimental PDF data and the searching criterion used to filter which structures to fetch from the databases. When finished, the best-fit candidate structures from the data mine are returned to the experimenter for further detailed investigations. Structure-mining currently supports both X-ray and neutron PDF data sets. This software enables high-throughput auto-refinement that may be used right after the PDF is obtained at a synchrotron X-ray or neutron beamline, unlike more traditional human-intensive approaches that typically take a large amount of time and effort after the experiment is over. It is designed to lighten the PDF modeling work after an experiment, but could also, in principle, be used for modeling PDF data sets in quasi-real time during the data acquisition at the beamline.

2. Approach

Structure-mining first obtains a large number of candidate structures from open structural databases. It then computes the PDFs of these structures and carries out structure refinements to obtain the best agreement between calculated PDFs and the measured PDF under study. The initial implementation uses two commonly utilized open structural databases: the Materials Project Database (MPD) (Jain *et al.*, 2013) and the Crystallography Open Database (COD) (Gražulis *et al.*, 2009). The structures are fetched directly from the databases using the *RESTful* API (Ong *et al.*, 2013, 2015). There are many rules that could be used for selecting candidate structures to try. In this initial implementation of structure-mining, we are using the following heuristics for filtering which structure models to fetch: (1) all the structures that have the same stoichiometry as prescribed by the experimenter, (2) all the structures containing a prescribed list of elements, (3) all the structures containing the prescribed list of elements plus a number of additional elements specified by a wild-card symbol, (4) all the structures containing a subset of the prescribed elements plus other elements if a wild-card symbol is specified. These heuristics go from more restrictive to less restrictive and may be selected as desired. The results on representative data sets are presented below.

After fetching the structures, structure-mining builds a list of candidate structures and loads their cif files from the database into the *diffpy-cmi* (Juhás *et al.*, 2015) PDF structure refinement program.

The *diffpy-cmi* program works by first building a fit recipe which is the set of information needed to run a model refinement to PDF data, and then executing it. The PDF fit recipe for each structure is generated automatically. By default, the fits are carried out over the range of $1.5 < r < 20$ Å on the Nyquist–Shannon sampling grid (Farrow *et al.*, 2011); however, a different fit range may be specified by the user. The following phase-related parameters are initialized and refined: a single scale factor uses the initial value 1.0; lattice parameters are constrained according to the crystal systems using the initial lattice parameter values of the structures; the isotropic atomic displacement parameter (ADP), U_{iso}, for each element atom of the structure is applied with initial value 0.005 Å^2; the spherical particle diameter (SPD) parameter can be used if the PDF data are from nano-sized objects, by having the experimenter specify an initial value (in the unit of Å). The instrument resolution parameters, Q_{damp} and Q_{broad}, which are the parameters that correct the PDF envelope function for the instrument resolution (Proffen & Billinge, 1999; Farrow *et al.*, 2007), are preferably obtained by measuring a standard calibration material in the same experimental setup geometry as the measured sample, and are fixed in the subsequent structure refinements of the measured sample PDF. They are applied according to the following strategy. If the experimenter specifies Q_{damp} and Q_{broad} values, the experimenter's values are used and they are fixed during the structure refinement. If they are not specified by the experimenter, the program will make a best-effort attempt to allocate meaningful values. This is done currently by storing a table of reasonable values by instruments. So far, we have established reasonable values for the XPD X-ray instrument and the NOMAD and NPDF neutron instruments. If the program cannot find reasonable values in its lookup table for a specified instrument, or if no instrument can be determined, standard global default values are selected. These are $Q_{damp} = 0.04$ Å^{-1} for rapid-acquisition X-ray PDF (RAPDF) experiments (Chupas *et al.*, 2003) and 0.02 Å^{-1} for time-of-flight (TOF) neutron PDFs. Similarly, $Q_{broad} = 0.01$ and 0.02 Å^{-1} are the global defaults for RAPDF X-ray and TOF neutron measurements, respectively. In all the cases where the user does not specify values for Q_{damp} and Q_{broad}, these parameters are allowed to vary in the refinement process.

Different regression algorithms may be used to perform the structure refinement minimizing the fit residual, with the goodness-of-fit R_w given by

$$R_w = \left\{ \frac{\sum_{i=1}^n [G_{obs}(r_i) - G_{calc}(r_i, P)]^2}{\sum_{i=1}^n G_{obs}(r_i)^2} \right\}^{1/2}, \qquad (2)$$

where G_{obs} and G_{calc} are the observed and calculated PDFs and P is the set of parameters refined in the model.

Initially we use the widely applied damped least-squares method (Levenberg–Marquardt algorithm) (Levenberg, 1944; Marquardt, 1963), which is deployed in the Python programming package *Scipy* (Jones *et al.*, 2001), to vary the adjustable parameters to achieve the best agreement between the calculated and measured PDFs, since none of the algorithms for nonlinear least-squares problems has been proved to be superior to this standard solution (Young, 1993; Floudas &

research papers

Pardalos, 2001), such as the Gauss–Newton method (Gauss, 1809), modified Marquardt method (Fletcher, 1971) and conjugate direction method (Powell, 1964). However, *diffpy-cmi* supports the use of different minimizers and the implementation with different optimizers will be tested in the future. During the structure refinement different types of parameters have quite different characteristic behaviors. A systematic parameter turn-on sequence is important to achieve convergence because turning on unstable parameters too early can result in divergent fits or getting trapped at local false minima. To make the structure-mining highly automatic without any human intervention during the whole procedure, here we tested an automatic turn-on sequence that was suggested for conventional full-profile Rietveld refinement (Young, 1993) as well as considering the difference between PDF and Rietveld refinement procedures. The current structure-mining deploys the following parameter turn-on sequence: (i) scale factor and lattice parameters are allowed to vary for up to ten iterations, (ii) isotropic ADPs are allowed to vary for up to 100 iterations, (iii) if selected, the instrument resolution parameters, Q_{damp} and Q_{broad}, are turned on for up to 100 iterations, and finally (iv) if SPD is specified, it will then be turned on for up to 100 iterations. When the whole procedure is finished, if the refinement cannot converge, the refinement will stop, record the latest goodness-of-fit parameter R_w value, and continue with the next structure. If the resulting $R_w > 1.0$ (unconverged fit), it would be marked as 1.0.

This process is repeated for every structure fetched from databases. When the program has looped over all the structures it returns a plot of best-fit goodness-of-fit parameters R_w of each model. We call this plot the structure-mining map (see a representative plot in Fig. 1).

The program also returns a detailed formatted table that is suitable for inserting into a manuscript summarizing the results of the structure-mining process. The experimenter can also select one or multiple structure model entries to save the corresponding results, figures of the data and the fit, the
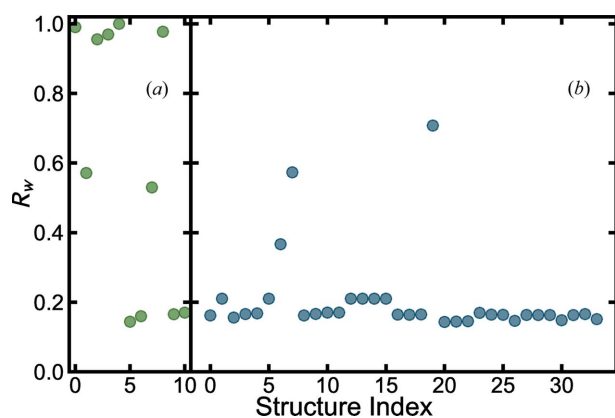

Figure 1
R_w values for each of the structures fetched from the databases for the BaTiO$_3$ nanoparticle X-ray data using heuristic 1, filtering for all the structures with composition BaTiO$_3$ from (*a*) the MPD (green) and (*b*) the COD (blue). The R_w parameter represents the goodness-of-fit for each structure.

Table 1
The experimental PDF data sets for testing the structure-mining approach.

The reference describing the experiments is given, except for the Ti$_4$O$_7$ data which are unpublished.

Composition	Scatterer	Beamline
BaTiO$_3$†	X-ray	XPD
Ti$_4$O$_7$†	X-ray	XPD
NaFeSi$_2$O$_6$‡	X-ray	XPD
Ba$_{0.8}$K$_{0.2}$(Zn$_{0.85}$Mn$_{0.15}$)$_2$As$_2$§	Neutron	NOMAD
CuIr$_2$S$_4$¶	X-ray	XPD
MnO††	Neutron	NPDF
V$_2$N + VN‡‡	X-ray	XPD

† Lombardi *et al.* (2019). ‡ Lewis *et al.* (2018). § Frandsen, Gong *et al.* (2016). ¶ Božin *et al.* (2019). †† Frandsen & Billinge (2015). ‡‡ Urbankowski *et al.* (2017).

calculated and difference PDF data files, the initial and refined structures in cif format, and the values of initial and refined parameters in a formatted table.

Structure-mining will be made available on a cloud-based platform at https://pdfitc.org.

3. Testing the approach

3.1. Testing methodology

To test the method, we selected PDFs of seven different materials from X-ray and neutron total scattering data, as listed in Table 1.

The total scattering measurements were conducted at one synchrotron X-ray facility, the XPD beamline (28-ID-2) at the National Synchrotron Light Source II (NSLS-II), Brookhaven National Laboratory, and two neutron time-of-flight facilities, the NOMAD beamline (BL-1B) (Neuefeind *et al.*, 2012) at the Spallation Neutron Source (SNS) at Oak Ridge National Laboratory and the NPDF beamline (Proffen *et al.*, 2002) at the Manuel Lujan Jr Neutron Scattering Center at Los Alamos Neutron Science Center (LANSCE), Los Alamos National Laboratory. All of the data sets are from previously published work, indicated in the table, except for that of Ti$_4$O$_7$, which is unpublished.

For the XPD beamline the samples were loaded in 1 mm inner diameter polyimide capillaries mounted perpendicular to the beam and the X-ray data sets were collected at room temperature, except for the vanadium nitride sample which was collected at 100 K (Urbankowski *et al.*, 2017) and the CuIr$_2$S$_4$ sample at 500 K (Božin *et al.*, 2019), using the RAPDF method (Chupas *et al.*, 2003). A large-area 2D Perkin Elmer detector was mounted behind the samples. The collected data frames were summed, corrected for detector and polarization effects, and masked to remove outlier pixels before being integrated along arcs of constant Q, where $Q = 4\pi \sin\theta/\lambda$ is the magnitude of the momentum transfer on scattering, to produce 1D powder diffraction patterns using the *fit2d* program (Hammersley, 2016). Standardized corrections and normalizations were applied to the data to obtain the total scattering structure function, $S(Q)$, which was Fourier

Table 2
Structure-mining results for the $BaTiO_3$ nanoparticle X-ray data using heuristic 1 from the MPD.

Here No. refers to the structure index [Fig. 1(a)], which is the order fetched from the database, and s.g. represents the space group of the structure model. The initial isotropic atomic displacement parameter (U_{iso}) of all atoms in each structure is set to 0.005 $Å^2$ to start the structure refinements. The a, b and c are the lattice parameters of the structure model. The subscript 'i' indicates an initial value before refinement and the subscript 'r' indicates a refined value. DB ID represents the database ID of the structure model. Q_{max} = 24.0 $Å^{-1}$, Q_{damp} = 0.037 $Å^{-1}$ and Q_{broad} = 0.017 $Å^{-1}$ were set and not varied in the refinements (see Section 2 for details).

No.	R_w	s.g.	Ba U_{iso} ($Å^2$)	Ti U_{iso} ($Å^2$)	O U_{iso} ($Å^2$)	a_i (Å)	a_r (Å)	b_i (Å)	b_r (Å)	c_i (Å)	c_r (Å)	DB ID
5	0.144	$Amm2$	0.0021	0.0070	0.0126	5.81	5.67	5.86	5.76	3.99	3.99	mp-5777
6	0.160	$P4mm$	0.0027	0.0074	0.0116	4.00	4.00	—	—	4.22	4.07	mp-5986
9	0.165	$R3m$	0.0027	0.0074	0.0123	5.75	5.66	—	—	7.11	7.05	mp-5020
10	0.170	$P4/mmm$	0.0026	0.0105	0.0174	4.03	4.00	—	—	4.04	4.07	mp-2998
7	0.530	$C222_1$	0.0047	0.0023	0.0373	5.84	5.69	10.02	9.84	14.14	13.98	mp-558125
1	0.571	$P6_3/mmc$	0.0070	0.0041	0.0468	5.79	5.69	—	—	14.10	13.97	mp-5933
2	0.956	$P4/mmm$	0.0172	0.0011	0.0884	4.11	4.16	—	—	5.04	4.73	mp-19990
3	0.969	$Amm2$	0.0003	0.0941	0.0090	5.31	5.26	5.33	5.44	8.88	8.80	mp-1076932
8	0.977	$Amm2$	0.0075	0.0006	0.0010	6.64	6.76	8.63	8.60	3.75	3.86	mp-644497
0	0.990	$Amm2$	0.0017	0.0031	0.0000	5.81	6.00	5.85	5.98	5.03	4.84	mp-995191
4	1.000	$Pm\overline{3}m$	0.0115	0.0104	0.0003	4.65	4.78	—	—	—	—	mp-504715

Table 3
Structure-mining results for the $BaTiO_3$ nanoparticle X-ray data using heuristic 1 from the COD.

See the caption of Table 2 for an explanation of the entries.

No.	R_w	s.g.	Ba U_{iso} ($Å^2$)	Ti U_{iso} ($Å^2$)	O U_{iso} ($Å^2$)	a_i (Å)	a_r (Å)	b_i (Å)	b_r (Å)	c_i (Å)	c_r (Å)	DB ID
20	0.143	$Amm2$	0.0021	0.0080	0.0181	5.67	5.67	5.69	5.76	3.98	3.99	9014492
21	0.144	$Amm2$	0.0020	0.0087	0.0178	5.67	5.67	5.69	5.76	3.98	3.99	9014627
22	0.145	$Amm2$	0.0020	0.0091	0.0175	5.67	5.67	5.69	5.76	3.99	3.99	9014645
26	0.146	$Amm2$	0.0020	0.0094	0.0168	5.67	5.67	5.68	5.76	3.99	3.99	9014774
30	0.148	$Amm2$	0.0020	0.0100	0.0179	5.67	5.67	5.69	5.76	3.98	3.99	9016084
33	0.151	$Amm2$	0.0025	0.0083	0.0070	5.68	5.68	5.69	5.75	3.99	3.99	9016638
2	0.156	$P4mm$	0.0027	0.0064	0.0116	3.99	4.00	—	—	4.04	4.07	1513252
8	0.162	$P4mm$	0.0026	0.0086	0.0162	4.00	4.00	—	—	4.02	4.07	2100858
0	0.162	$P4mm$	0.0026	0.0086	0.0162	4.00	4.00	—	—	4.02	4.07	1507756
31	0.163	$R3m$	0.0027	0.0076	0.0163	5.65	5.66	—	—	6.96	7.05	9016152
29	0.163	$Amm2$	0.0028	0.0054	0.0029	5.62	5.63	5.64	5.70	4.01	4.06	9015715
28	0.163	$R3m$	0.0027	0.0077	0.0161	5.65	5.66	—	—	6.95	7.05	9015616
27	0.163	$R3m$	0.0027	0.0079	0.0160	5.65	5.66	—	—	6.95	7.05	9015236
25	0.164	$R3m$	0.0027	0.0082	0.0160	5.66	5.66	—	—	6.95	7.05	9014756
17	0.164	$R3m$	0.0027	0.0083	0.0158	5.66	5.66	—	—	6.96	7.05	9014179
24	0.164	$R3m$	0.0026	0.0084	0.0153	5.65	5.66	—	—	6.95	7.05	9014743
16	0.164	$R3m$	0.0026	0.0085	0.0157	5.66	5.66	—	—	6.95	7.05	9014074
18	0.165	$R3m$	0.0026	0.0087	0.0150	5.65	5.66	—	—	6.95	7.05	9014230
32	0.166	$R3m$	0.0026	0.0091	0.0149	5.65	5.66	—	—	6.96	7.05	9016624
3	0.166	$P4mm$	0.0026	0.0096	0.0151	3.99	4.00	—	—	4.03	4.07	1525437
9	0.166	$P4mm$	0.0026	0.0097	0.0158	4.00	4.00	—	—	4.02	4.07	2100859
4	0.168	$Pmm2$	0.0026	0.0095	0.0151	3.98	3.99	4.01	4.01	4.02	4.07	1540757
23	0.169	$P4mm$	0.0026	0.0103	0.0163	4.00	4.00	—	—	4.02	4.07	9014668
11	0.170	$P4/mmm$	0.0026	0.0105	0.0174	4.00	4.00	—	—	4.02	4.07	2100861
10	0.170	$P4/mmm$	0.0026	0.0105	0.0174	4.00	4.00	—	—	4.02	4.07	2100860
15	0.210	$Pm\overline{3}m$	0.0046	0.0132	0.0172	3.97	4.02	—	—	—	—	5910149
1	0.210	$Pm\overline{3}m$	0.46	0.0132	0.0172	4.01	4.02	—	—	—	—	1507757
13	0.210	$Pm\overline{3}m$	0.0046	0.0132	0.0172	4.01	4.02	—	—	—	—	2100863
12	0.210	$Pm\overline{3}m$	0.0046	0.0132	0.0172	4.01	4.02	—	—	—	—	2100862
5	0.210	$Pm\overline{3}m$	0.0046	0.0132	0.0172	4.00	4.02	—	—	—	—	1542140
14	0.210	$Pm\overline{3}m$	0.0046	0.0132	0.0172	4.03	4.02	—	—	—	—	4124842
6	0.367	$Pm\overline{3}m$	0.0058	0.0126	0.0799	4.08	4.02	—	—	—	—	1542189
7	0.573	$P6_3/mmc$	0.0070	0.0041	0.0469	5.72	5.69	—	—	13.96	13.97	2009488
19	0.708	$P4mm$	0.0042	0.2490	0.0479	3.99	4.04	—	—	4.03	3.98	9014273

transformed to obtain the PDF, using *pdfgetx3* (Juhás *et al.*, 2013) within *xpdfsuite* (Yang *et al.*, 2015). The incident X-ray wavelengths and the calibrated sample-to-detector distances are listed in Appendix *A* (Table 6).

For the NOMAD and NPDF beamlines, the samples were loaded in vanadium cans. The $Ba_{0.8}K_{0.2}(Zn_{0.85}Mn_{0.15})_2As_2$ data

from the NOMAD beamline were collected at room temperature (Frandsen, Gong *et al.*, 2016) and the data were reduced and transformed to the PDF using the automated data reduction scripts at the NOMAD beamline. The MnO data from the NPDF beamline were collected at 15 K (Frandsen & Billinge, 2015) and the data were reduced and

transformed to the PDF using the *pdfgetn* program (Peterson *et al.*, 2000).

The full experimental details may be found in Lombardi *et al.* (2019), Lewis *et al.* (2018), Frandsen, Gong *et al.* (2016), Božin *et al.* (2019), Frandsen & Billinge (2015) and Urbankowski *et al.* (2017). The maximum range of data used in the Fourier transformation, Q_{max}, was chosen to give the best trade-off between statistical noise and real-space resolution. The instrument resolution parameters, Q_{damp} and Q_{broad}, which are relevant parameters for our structure-mining activity, were obtained by calibrating the experimental conditions in each case using a well-crystallized standard sample. The values are reproduced in Appendix *A* (Table 6).

3.2. Results

We first apply this approach to the measured PDF from barium titanate (BTO) nanoparticles, $BaTiO_3$. BTO is one of the best studied perovskite ferroelectric materials (Frazer *et al.*, 1955; Kwei *et al.*, 1993). Heuristic 1 is applied, fetching all structures that have the same composition as input $BaTiO_3$. The structure-mining results from the MPD and COD are shown in Figs. 1(*a*) and 1(*b*), and Tables 2 and 3, respectively.

The best-fit structures from each data mine were MPD structure No. 5 (Shirane *et al.*, 1957) and COD structure No. 20 (Kwei *et al.*, 1993) with $R_w = 0.144$ and 0.143, respectively. The calculated and measured PDFs are shown in Figs. 2(*a*) and 2(*b*), respectively. Unlike the traditional manual PDF structure refinement methodology, the structure-mining approach followed by the automated fitting resulted in satisfactory and
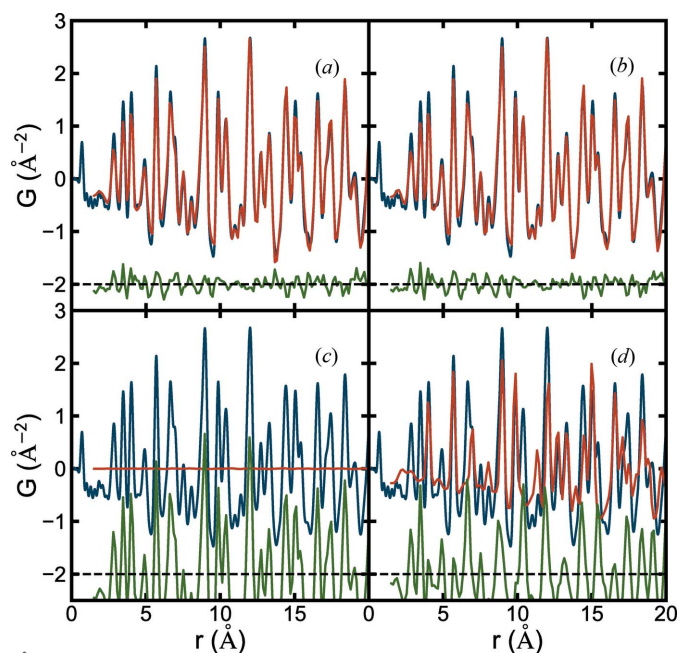
reasonable fits without any human intervention. These structures may be investigated in more detail by traditional manual fitting approaches.

Some structures retrieved from the mine also resulted in very poor fits, as shown in Figs. 2(*c*) and 2(*d*), which are the automatically determined fits of MPD structure No. 4 (Jain *et al.*, 2013) and COD structure No. 19 (Shirane *et al.*, 1957), respectively. We expect that this will be due to the fact that the structure fetched from the database is different from that of our sample, and it is this automated screening of database structures to find the most plausible candidates that is the goal of structure-mining. However, we investigate this in more detail below.

The structure of this measured $BaTiO_3$ nanoparticle data set has been carefully studied before (Lombardi *et al.*, 2019). In that work, it was reported that the structure of this nanoparticle sample was noncentrosymmetric at room temperature and had one of the ferroelectric forms of the $BaTiO_3$ structures (Kwei *et al.*, 1993), among one of the distorted structures with space groups *Amm*2, *P4mm* and *R3m*. All these structures gave a somewhat comparable fit to the data and it was difficult to distinguish which among them was definitively the correct structure because of the Bragg peak broadening and the relatively weak X-ray scattering of the oxygen sublattice. Nearby centrosymmetric space groups also performed well based on R_w but could be ruled out by careful consideration of refined ADPs of Ti ions [we note that in Table V of Kwei *et al.* (1993) there is a typo where the s.g. (space group) *P4mm* structure is described as *P4/mmm*, which is the centrosymmetric parent, but in the body of the table the Ti ion is shown as displacing off the center of the unit cell, breaking centrosymmetry].

The MPD result, as shown in Table 2, clearly reveals that the top three best-fit structures are exactly the noncentrosymmetric ferroelectric forms of $BaTiO_3$ structures with space groups *Amm*2, *P4mm* and *R3m*. In addition, the closely similar centrosymmetric perovskite model with space group *P4/mmm* (No. 10, ranked 4) (Srilakshmi *et al.*, 2016) gives a slightly worse but comparable R_w. The heuristic 1 has therefore found the correct candidate structure models from the MPD, as well as returning nearby structures for a more detailed manual comparison.

The COD contained many more candidate structures for this composition (Table 3). Again, the structure-mining shows that the best three noncentrosymmetric perovskite models are found as expected, along with the similar general barium titanate perovskite models (with slightly worse R_w) with space groups *P4/mmm* and *Pm$\bar{3}$m*.

The COD result also returned a space group *Pmm*2 structure (No. 4) (Zeng & Jiang, 1991) with a reasonable fit ($R_w = 0.168$) which turns out to be a general perovskite structure having two half-filled Ti ions at (0.5, 0.5, 0.509) and (0.5, 0.5, 0.491) sites, similar to a doubled unit cell of the tetragonal barium titanate perovskite model with space group *P4mm*, albeit with a small orthorhombic distortion. This illustrates the power of this structure-mining approach as it does a good job of finding all plausible structures in the

Figure 2
PDFs from representative satisfactory and unsatisfactory structures from (*a*), (*c*) the MPD and (*b*), (*d*) the COD. Blue curves are the measured PDF of $BaTiO_3$ nanoparticles. Red curves are the calculated PDFs after retrieving from the databases using heuristic 1 and automatically fitting to the data (see Section 2 for details). Offsets below in green are the difference curves.

database. These can then be considered and ruled out by researchers based on other criteria.

There is also a hexagonal structure (space group $P6_3/mmc$) in the databases for BaTiO$_3$, and this gives a very poor fit to the BaTiO$_3$ nanoparticle data from both the MPD (No. 1) (Akimoto *et al.*, 1994) and COD (No. 7) (Akimoto *et al.*, 1994), showing that the approach is capable of finding true positive and true negative results.

The structure-mining gives the COD structure No. 19 (space group $P4mm$) (Shirane *et al.*, 1957) a bad fit because the model is wrong, with the Ti ion sitting at 1b (0.5, 0.5, 0.265) and the O2 ion sitting at 2c (0.5, 0, 0.236), which is significantly offset from the correct position such that the Ti ion is at or near the center of the unit cell. We checked the reference for this database entry (COD ID: 9014273), and it turned out to be correct in the paper but a wrong entry in the database because the reference reported that the Ti ion was at 1b (0.5, 0.5, 0.515) and the O2 ion was at 2c (0.5, 0, 0.486) (Shirane *et al.*, 1957). This indicates that this structure-mining approach may actually help to find errors in the database, but at worst will not return incorrect structures as candidate models.

Interestingly, the mining operation did report one false negative. It missed one of the plausible perovskite structure models in the MPD database, the cubic model with space group $Pm\bar{3}m$ (MPD No. 4) (Jain *et al.*, 2013), which was correctly found in the COD database. The reason why this did not give a good refinement was that the starting lattice parameters taken from the database were much too large (a = 4.65 Å) and the automated refinement could not converge to the correct minimum (a = 4.02 Å) due to the 55% cell volume mismatch from the correct one, resulting in a poor fit. Although we refine the lattice parameter during the process, if the starting value is too far away from the correct one, it is possible that the refinement program will not be able to find the right solution in the parameter space and result in a poor fit and a false negative result. In some respect it is a success of the program because we actually hope that incorrect models in the database will fit the data poorly, and if the value of the lattice parameter recorded in the database is far from being correct for the measured sample, in some sense this constitutes a bad model. Similar lattice parameter situations happen for MPD No. 0 (Xiao *et al.*, 2008), 2 (Donohue *et al.*, 1958), 3 (Xiao *et al.*, 2008) and 8 (Hayward *et al.*, 2005). The entries in the MPD that are taken from the ICSD have gone through an energy relaxation step using density functional theory (DFT) (Hohenberg & Kohn, 1964; Kohn & Sham, 1965) before the crystal structures are deposited in the MPD. For some reason, the DFT relaxation took some of the lattice parameters somewhat far away from the experimental values in the original structure reports (Xiao *et al.*, 2008; Donohue *et al.*, 1958; Hayward *et al.*, 2005). Overall the heuristic 1 approach already returned the correct structures for BaTiO$_3$ nanoparticles. The complete mining operation took 29.3 s when searching with the MPD and 47.8 s for the COD search to complete, using a general laptop.

We would like to further test the more loosely filtered heuristic 2 approach on the BaTiO$_3$ nanoparticle data. The
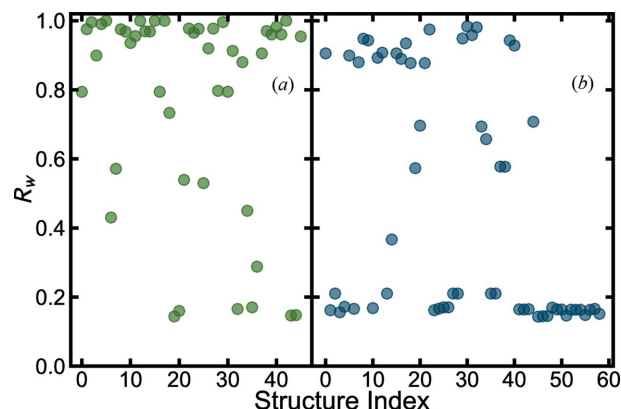

Figure 3
R_w values for each of the structures fetched from the databases for the BaTiO$_3$ nanoparticle X-ray data using heuristic 2, filtering for all the structures with Ba, Ti and O elements from (*a*) the MPD (green) and (*b*) the COD (blue).

structure-mining results from the MPD and COD, fetching all structures that contain just Ba, Ti and O elements with any stoichiometry, are shown in Figs. 3(*a*) and 3(*b*), respectively. More details about the results can be found in the supporting information CSV files.

Heuristic 2 found all the structures that were found with heuristic 1, as expected. This approach also found a number of additional good structural candidates. The MPD returned three more that were within $\Delta R_w \simeq 0.1$ from the best-fit R_w (approximately 0.14), *i.e.* MPD No. 43 (Ba$_{12}$Ti$_{12}$O$_{27}$) (Jain *et al.*, 2013), 44 (Ba$_3$Ti$_3$O$_8$) (Woodward *et al.*, 2004), 36 (Ba$_4$Ti$_4$O$_{10}$) (Jain *et al.*, 2013) and COD returned one, No. 4 model (Ba$_{0.92}$Ti$_{0.9}$O$_{2.89}$) (Wada *et al.*, 2000), where ΔR_w is the deviation in R_w of a structure from the R_w of the best-fit structure. Close inspection of these models indicates that they have a stoichiometry that is approximately the Ba:Ti:O = 1:1:3 ratio and that structure-mining found some nearby defective structures in addition to the standard 113 perovskite structures. This will allow the experimenters to further investigate the defective models to find any physical or chemical insights that they might provide.

The heuristic 2 structure-mining operation also, as expected, returned some structures from the databases for which the atomic composition ratio was not close to 1:1:3. None of these additional structures gave reasonable fits to the PDF, resulting in poor R_w values larger than 0.4 for the MPD [such as Ba$_2$Ti$_3$O$_8$ MPD No. 6 (Jain *et al.*, 2013)] and 0.6 for the COD [such as Ba$_{11}$Ti$_{28}$O$_{66.48}$ COD No. 34 (Vanderah *et al.*, 2004)]. The entire search process took 493.7 s for the MPD and 469.5 s for the COD.

The heuristic 3 approach was also tested on the BaTiO$_3$ nanoparticle data by fetching all structures that contain Ba, Ti, O elements and one additional element with any stoichiometry. More details about the results can be found in the supporting information CSV files. It took about 10.3 and 41.0 min for the MPD (in total 57 structures) and COD (in total 103 structures) to finish, respectively. Of these new structures that were found, most of the best-fit structures have slightly worse R_w than those in heuristics 1 ($R_w \simeq 0.2$) and 2

Table 4
The top seven structure-mining results for the Ti_4O_7 experimental X-ray PDF using heuristic 2 on data from the MPD and COD.

See the caption of Table 2 for an explanation of the entries. The full table can be found in the supporting information CSV files. The initial lattice parameters and refined ADPs are listed. The refined lattice parameters are not listed because they are close to initial values.

DB No.	R_w	Formula	s.g.	Ti U_{iso} ($\mathring{A}^2$)	O U_{iso} ($\mathring{A}^2$)	a_i ($\mathring{A}$)	b_i ($\mathring{A}$)	c_i ($\mathring{A}$)	α_i (°)	β_i (°)	γ_i (°)	Reference
COD 20	0.168	Ti_4O_7	$P\bar{1}$	0.0051	0.0076	5.60	7.13	12.47	95.1	95.2	108.7	Marezio & Dernier (1971)
COD 1	0.169	Ti_4O_7	$P\bar{1}$	0.0050	0.0104	5.59	6.91	7.13	64.1	71.0	75.3	Hodeau & Marezio (1979)
COD 21	0.170	Ti_4O_7	$P\bar{1}$	0.0050	0.0104	5.59	6.91	7.13	64.1	71.1	75.5	Marezio et al. (1973)
COD 0	0.173	Ti_4O_7	$P\bar{1}$	0.0048	0.0108	5.59	6.90	7.12	64.1	71.2	75.7	Hodeau & Marezio (1979)
MPD 38	0.174	Ti_5O_9	$P\bar{1}$	0.0046	0.0065	5.62	7.18	8.56	69.5	75.2	71.3	Marezio et al. (1977)
MPD 49	0.183	Ti_4O_7	$P\bar{1}$	0.0048	0.0108	5.64	6.96	7.18	64.2	71.1	75.1	Hodeau & Marezio (1979)
COD 36	0.225	Ti_5O_9	$P1$	0.0053	0.0088	5.57	7.12	8.49	69.8	75.0	71.5	Andersson (1960)

$\vdots$

($R_w \simeq 0.14$). The new structures are mostly substituting the Ba or Ti site by another element and they also have an approximate stoichiometry 113, such as MPD No. 43 ($Ba_3Sr_5Ti_8O_{24}$) (Jain et al., 2013) and COD No. 22 ($Ba_{0.93}Ti_{0.79}Mg_{0.21}O_{2.97}$) (Wada et al., 2000), which agrees with what has been found in heuristic 2.

Finally we tested the very loose heuristic 4 approach. Here the experimenter can freely choose any searching criteria, such as Ba-Ti-*, Ba-*-O, or even *-*-*, in which an * represents an arbitrary element. In our test case we set the search to be that where the structure contains three elements, including Ba and two other elements, i.e. Ba-*-*. The structure-mining map plot is shown in Fig. 4. This search took much longer, 174.3 and 205.2 min on a single CPU core for the MPD and COD, respectively. This may be sped up by running on more cores. In total, 1833 structures were fetched from the MPD and 1046 from the COD. More details about the results are available in the supporting information CSV files. The less restrictive heuristic 4 found all the structures that were found with heuristics 1 and 2, as expected. The normal $BaTiO_3$ perovskite structures are still ranked at the top. Following that, it additionally returns some perovskite structures that have Ti replaced by other species with similar X-ray scattering power as Ti, such as MPD No. 1660 ($BaVO_3$) (Nishimura et al., 2014), MPD No. 1268 ($BaMnO_3$) (Jain et al., 2013) and COD

No. 683 ($BaFeO_3$) (Erchak et al., 1946). These gave agreements of $R_w \gtrsim 0.2$ compared with 0.14 for the best-fit structures ($BaTiO_3$). Thus the structure-mining is able to distinguish these nearby but incorrect structures from the ones with correct atom species. The perovskite structures with the B-site element replaced by one with a significantly different X-ray scattering power than Ti resulted in a significantly poorer R_w, away from the best-fit structures by $\Delta R_w \simeq 0.15$, such as MPD No. 1482 ($BaRhO_3$) (Balachandran et al., 2017) and COD No. 431 ($BaNbO_3$) (Grin et al., 2014).

Overall we achieved a satisfactory result for the barium titanate nanoparticle data set using all the four structure-mining heuristics.

We now test structure-mining for some different materials, for example, the low-symmetry Ti_4O_7 system. Its published room-temperature crystal structure is a triclinic model (space group $P\bar{1}$) with all the atoms sitting on (x, y, z) general positions (Marezio & Dernier, 1971). We used the structure-mining heuristic 2 approach, fetching all the structures that contain Ti and O elements with any stoichiometry. The structure-mining map plot is shown in Fig. 5 and the detailed results are available in the supporting information CSV files. The top seven structure-mining results are also summarized in Table 4. The titanium oxides have many different structures,
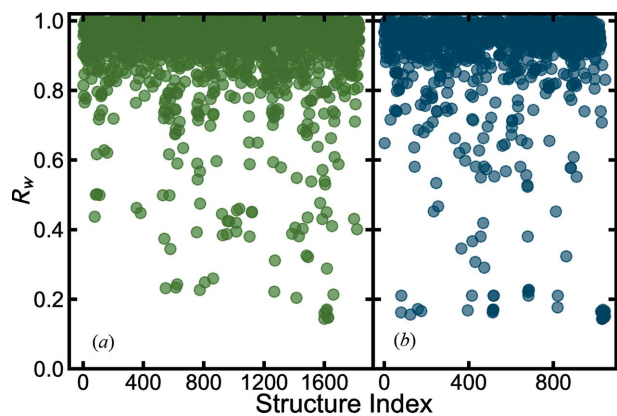

Figure 4
R_w values for each of the structures fetched from the databases for the $BaTiO_3$ nanoparticle X-ray data using heuristic 4, filtering for all the structures with Ba, and two other arbitrary elements from (a) the MPD (green) and (b) the COD (blue).
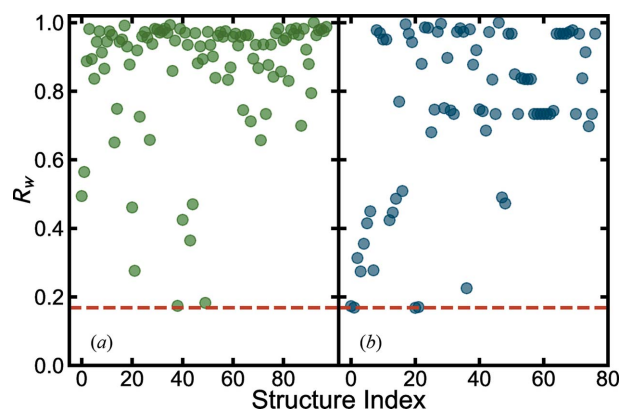

Figure 5
R_w values for each of the structures fetched from the databases for the Ti_4O_7 X-ray data using heuristic 2, filtering for all the structures with Ti and O elements from (a) the MPD (green) and (b) the COD (blue). The horizontal dashed line represents the lowest R_w entry found, COD No. 20 (Ti_4O_7, s.g.: $P\bar{1}$).

largely depending on the stoichiometry (98 structures from the MPD and 77 from the COD), but structure-mining returned the published structure for Ti_4O_7 on the top, *i.e.* COD No. 20 (Marezio & Dernier, 1971).

This is a challenging problem because there are similar structures belonging to the Ti_nO_{2n-1} Magnéli homologous series (Andersson & Magnéli, 1956; Andersson *et al.*, 1957). Among the top seven entries, the other four Ti_4O_7 structures are very similar to COD No. 20. COD 20 is reported in a different structural setting than the other four (Setyawan & Curtarolo, 2010), which explains the rather different values for the lattice parameters, but the only real difference in structure between COD 20 and the other Ti_4O_7 structures reported in Table 4 is that one oxygen position is shifted by about 0.7 Å along the *b* axis compared with the other four. This is a significant structural difference yet does not result in a very large difference in R_w and so differentiating these two structures probably deserves some additional consideration by the experimenter. Atomic positions are not refined independently during this structure-mining process and it is possible that this discrepancy may be resolved by a full refinement of the best performing models, as well as suggesting to the user the oxygen *b*-axis position as a possibly relevant variable. Structure-mining also returned some results with slightly different stoichiometry with similar R_w values, for example the MPD No. 38 (Ti_5O_9) (Marezio *et al.*, 1977), which belongs to a different variant in the Magnéli series. The Magnéli phases are constructed from similar TiO_6 octahedral motifs, containing rutile-like slabs extending infinitely in the *a–b* plane, but the TiO_6 octahedra are stacked along the *c* axis in slabs of different widths depending on the composition (Andersson & Magnéli, 1956; Andersson *et al.*, 1957; Marezio *et al.*, 1977). In Ti_4O_7, every oxygen atom connects four octahedra, but in Ti_5O_9 (MPD 38), oxygen atoms link three octahedra. Despite these differences, the MPD 38 model performs similarly, albeit somewhat worse, than some of the well-performing Ti_4O_7 models, suggesting that it at least warrants being explicitly ruled out as a candidate in a more careful modeling. This

Figure 6
R_w values for each of the structures fetched from the databases for the $NaFeSi_2O_6$ nanowire X-ray data using heuristic 4, filtering for all the structures with Si, O, and two other arbitrary elements from (*a*) the MPD (green) and (*b*) the COD (blue). The horizontal dashed line represents the lowest R_w entry found, MPD No. 1021 ($NaGaSi_2O_6$, s.g.: *C2/c*).

illustrates how the structure-mining approach, beyond just automatically finding the 'right' structure, additionally can add value by suggesting alternative nearby models to the experimenter. We also note that, from Table 4, COD No. 36 (Ti_5O_9, s.g.: *P1*) (Andersson, 1960) performs worse ($R_w > 0.2$), and it is the first model that has a significantly different structure, where some Ti atoms are tetrahedrally coordinated by oxygen rather than octahedrally. This model can probably be ruled out on the basis of structure-mining alone.

Now let us turn to a challenging data set, nanowire bundles of a pyroxene compound with a generic composition of $XYSi_2O_6$ (where *X* and *Y* refer to metallic elements such as but not limited to Co, Na and Fe). This example is particularly challenging because the samples formed as nanowires that were reported to be ~3 nm in width (Lewis *et al.*, 2018). In that work, a series of candidate structures were tried manually and the best-fit model was found to be monoclinic $NaFeSi_2O_6$ with a space group *C2/c* (Clark *et al.*, 1969).

The structure-mining heuristic 1 approach was first tested. The MPD found one structure (Clark *et al.*, 1969) and the COD found six non-duplicated structures (Sueno *et al.*, 1973; Thompson & Downs, 2004; Redhammer *et al.*, 2000, 2006; Nestola, Tribaudino *et al.*, 2007; McCarthy *et al.*, 2008), all having a quite similar structure, $NaFeSi_2O_6$ (s.g.: *C2/c*). The returned structure-mining results have $R_w \simeq 0.35$. These are poor fits overall, but comparable with the fits reported in prior work (Lewis *et al.*, 2018). Although the R_w is not ideal, possibly due to the sample's complicated geometry, structural heterogeneity and defects, the structure-mining approach seems still to be working. Using heuristic 2 (Na-Fe-Si-O) and 3 (Na-Fe-Si-O-*) approaches found similar results, with heuristic 3 finding some Ca- and Li-doped compounds albeit with the same structure.

The least restrictive heuristic 4 approach was also tried. Here we show the result of fetching all the structures that contain Si and O elements and two other arbitrary elements with any stoichiometry, *i.e.* *-*-Si-O (Fig. 6). The mining operation took about 12 h for the MPD (in total 1700 structures) and 122 h for the COD (3187 structures) to finish. The COD is significantly more time-consuming because many of the COD structures have large numbers of hydrogen atoms, which could be neglected for X-ray PDF calculation to shorten the running time in future work. More details about the results are available in the supporting information CSV files. However, the top ten entries across the MPD and COD are listed here for convenience in Table 5.

The returned $NaGaSi_2O_6$ entries (s.g.: *C2/c*) (Ohashi *et al.*, 1983, 1995; Nestola, Rotiroti *et al.*, 2007) have a similar structure to $NaFeSi_2O_6$ (s.g.: *C2/c*). They both fit experimental data comparably well with $NaGaSi_2O_6$ slightly preferred. The $NaGaSi_2O_6$ solution can be ruled out on the basis that no Ga was in the synthesis. The X-ray scattering powers of Fe and Ga are similar with that of Ga being slightly higher [$Z(Fe) = 26$, $Z(Ga) = 31$]. The fact that structure-mining prefers to put a slightly higher atomic number, *Z*, element at this position suggests that we have the right structure, but some details of the refinement need to be worked out by the experimenter.
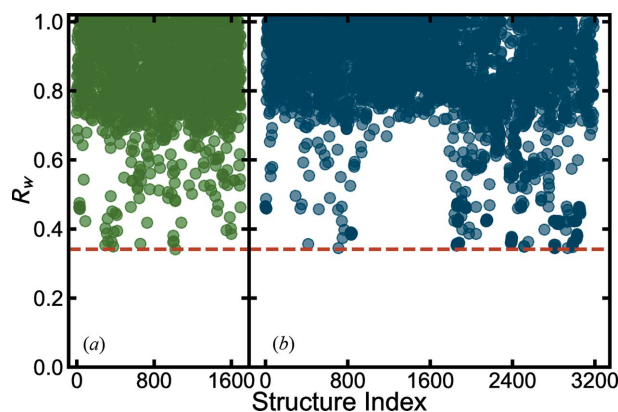
Table 5
The top ten structure-mining results for the $NaFeSi_2O_6$ nanowire experimental X-ray PDF using heuristic 4 on data from the MPD and COD, fetching all the structures that contain Si and O elements and two other arbitrary elements with any stoichiometry, *i.e.* *1-*2-Si-O (where the *1 and *2 represent the first and the second atoms in the formula, respectively).

See the caption of Table 2 for an explanation of the entries. The full table can be found in the supporting information CSV files. The refined lattice parameters and ADPs are listed. The initial lattice parameters are available in the supporting information but are not reproduced here. They are all very close to refined values.

DB No.	R_w	Formula	s.g.	*1 U_{iso} (Å^2)	*2 U_{iso} (Å^2)	Si U_{iso} (Å^2)	O U_{iso} (Å^2)	a_r (Å)	b_r (Å)	c_r (Å)	β_r (°)	SPD (Å)	DB ID
MPD 1021	0.341	$NaGaSi_2O_6$	$C2/c$	0.0193	0.0053	0.0048	0.0118	9.69	8.81	5.32	107.5	34.0	mp-6822
COD 709	0.345	$NaGaSi_2O_6$	$C2/c$	0.0174	0.0054	0.0049	0.0112	9.68	8.81	5.32	107.5	33.8	2004306
COD 2935	0.345	$NaGaSi_2O_6$	$C2/c$	0.0174	0.0054	0.0049	0.0112	9.68	8.81	5.32	107.5	33.8	9011383
COD 2809	0.345	$NaGaSi_2O_6$	$C2/c$	0.0173	0.0054	0.0048	0.0112	9.68	8.81	5.32	107.5	33.8	9010186
COD 2983	0.348	$NaFeSi_2O_6$	$C2/c$	0.0249	0.0033	0.0088	0.0129	9.68	8.82	5.32	107.5	34.1	9013274
COD 2513	0.348	$NaFeSi_2O_6$	$C2/c$	0.0214	0.0035	0.0070	0.0144	9.68	8.82	5.32	107.5	34.7	9005439
MPD 377	0.349	$Ca_{0.5}NiSi_2O_6$	$C2$	0.0118	0.0041	0.0052	0.0136	9.68	8.81	5.31	107.4	32.9	mvc-12761
COD 1856	0.352	$NaFeSi_2O_6$	$C2/c$	0.0221	0.0033	0.0079	0.0137	9.69	8.81	5.32	107.6	34.6	9000327
COD 2805	0.353	$NaFeSi_2O_6$	$C2/c$	0.0227	0.0032	0.0082	0.0135	9.69	8.81	5.32	107.6	34.7	9010095
MPD 294	0.353	$Ca_{0.5}CoSi_2O_6$	$C2$	0.0277	0.0042	0.0050	0.0231	9.68	8.82	5.32	107.3	34.8	mvc-11818
⋮													

This example illustrates how careful interrogation of the fits to the database models compared with the original parameters can highlight possible defects or impurities and guide the experimenter towards what things to search for.

The MPD also returned some computed theoretical structures with space group $C2$, MPD No. 377 ($Ca_{0.5}NiSi_2O_6$, s.g.: $C2$) and MPD No. 294 ($Ca_{0.5}CoSi_2O_6$, s.g.: $C2$) (Jain *et al.*, 2013). These perform slightly worse than the fully stoichiometric $NaGaSi_2O_6$ and $NaFeSi_2O_6$ structures. Inspection of these structures indicates that they are very similar in nature but with a lowered symmetry due to missing Ca ions and can probably be ruled out, though the fact that structure-mining finds them may suggest trying sub-stoichiometry models on the *A* site.

Overall, the heuristic 4 returned a number of isostructural structures, but with different composition. For this system, it is possible that the correct structure is not limited to the pure $NaFeSi_2O_6$ (s.g.: $C2/c$) stoichiometry only and substituting impurity ions or atom deficiencies may be occurring for such a complicated synthesis (Lewis *et al.*, 2018). These candidate structures found by structure-mining are valuable to resolve the ambiguity. Furthermore, taking the structure-mining approach yields different but similarly fitting models which can also give meaningful information about uncertainty estimates on refined parameters such as metal or oxygen ion positions. This test again shows the huge potential of structure-mining on PDF data to help experimenters be aware of some possible structural solutions that were overlooked or not realized in the traditional workflow.

Next, we test structure-mining on a complicated doped material, $Ba_{1-x}K_x(Zn_{1-y}Mn_y)_2As_2$. We used the neutron PDF data with composition $(x, y) = (0.2, 0.15)$, with both *A*-site and *B*-site dopings. Its published room-temperature crystal structure is a tetragonal structure with the space group $I4/mmm$ (Frandsen, Gong *et al.*, 2016). First we applied heuristic 2 specifying all the elements including the dopants, *i.e.* fetching Ba-Zn-As-K-Mn structures regardless of stoichiometry. This returned no structures from the MPD or the COD. We next tested a heuristic 4 approach with Ba-Zn-As-*-*. This did result in two structures being returned, but they were both incorrect compounds, $Ba_2MnZn_2As_2O_2$ (Ozawa *et al.*, 1998) and $BaZn_2As_3HO_{11}$ (Jain *et al.*, 2013), with R_w values close to 1, as shown in Fig. 7(*b*). We then looked for structures with doping on one site. The 'Ba-Zn-As-*' searches the databases for compositions containing four elements, including Ba, Zn, As and one other element. But it still only found incorrect structures, as shown in Fig. 7(*c*). Finally, we resorted to a heuristic 2 approach but only giving the composition of the undoped endmember, Ba-Zn-As. This did find the correct structure, tetragonal phase MPD No. 1 ($BaZn_2As_2$, s.g.: $I4/mmm$) (Hellmann *et al.*, 2007), as marked by the red circle in Fig. 7(*d*), even though we were fitting to the doped data. This suggests a feasible strategy for doped systems if they are not represented in the databases, which is to try searching for the parent undoped structure, on the basis that the doped structure may be still close to its parent phase, regardless of possible local structure distortions introduced by doping (Frandsen, Gong *et al.*, 2016). Starting from this
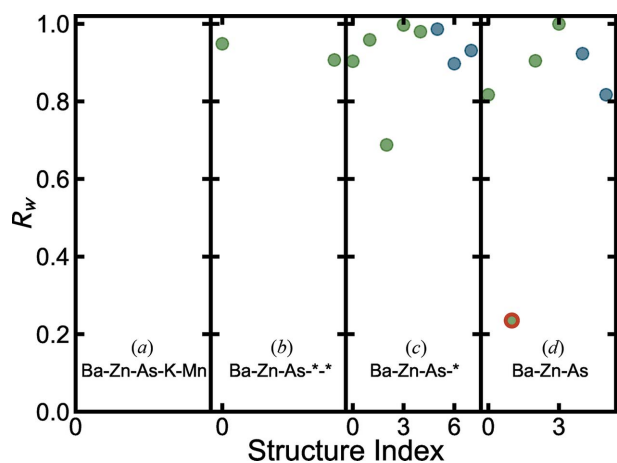

Figure 7
R_w values for each of the structures fetched from the databases for the $Ba_{0.8}K_{0.2}(Zn_{0.85}Mn_{0.15})_2As_2$ neutron data using the heuristics of (*a*) Ba-Zn-As-K-Mn, (*b*) Ba-Zn-As-*-*, (*c*) Ba-Zn-As-* and (*d*) Ba-Zn-As from the MPD (green) and the COD (blue). The best-fit model MPD No. 1 ($BaZn_2As_2$) in (*d*) is marked by a red circle.

success, the experimenter could then easily change the occupancy of the *A* site or *B* site, which was also how structural analysis was previously performed on this doped material (Zhao *et al.*, 2013; Rotter *et al.*, 2008). Thus, even for the case of doped structures, structure-mining found the correct geometric structure which was from the nearest undoped variant in the database (in this case, there were no structures in the databases that had the same composition as the measured sample). The experimenter can take this structure model and introduce dopants with the known composition.

Note that the PDF measures the *local* structure of materials, which is not necessarily identical to the long-range ordered crystal structure in some materials (Keen & Goodwin, 2015; Božin *et al.*, 2019). As a result, structure-mining will search for the closest structure in the crystal structure databases to the local atomic arrangement. Here we present the case of the $CuIr_2S_4$ system, which has a tetragonal local structure (s.g.: $I4_1/amd$) but a cubic long-range ordering (s.g.: $Fd\bar{3}m$) above the metal–insulator transition temperature (T_{MIT} = 226 K) (Božin *et al.*, 2019). The X-ray data measured at 500 K were tested. When fitting over the narrow range of $1.5 < r < 5$ Å, $1.5 < r < 10$ Å or $1.5 < r < 20$ Å, structure-mining ranks the tetragonal model above the cubic model. However, when fitting over the broad range of $1.5 < r < 50$ Å, the cubic model fits slightly better than the tetragonal model. The representative results for the $1.5 < r < 5$ Å and $1.5 < r < 50$ Å fit ranges can be found in the supporting information CSV files. As a result, structure-mining returns the symmetry-broken and non-symmetry-broken structural candidates whether it is fitted over a narrow or broad range, but it also ranks them correctly depending on the *r*-range fit used in this test case. We note that it is possible in structure-mining for a user to specify a custom fit range, which would allow the researcher to search for structures that are relevant for the measured PDF on different length scales.

We would also like to test the robustness of the structure-mining approach when the structural data also include non-structural signals, such as the magnetic PDF (mPDF) signal (Frandsen *et al.*, 2014; Frandsen & Billinge, 2015; Frandsen, Brunelli *et al.*, 2016) in a neutron diffraction experiment of a magnetic material. To test this we consider the MnO neutron PDF data, measured at 15 K, which have a strong mPDF signal. Early neutron diffraction studies reported that MnO has a cubic structure in space group $Fm\bar{3}m$ at high temperature and undergoes an antiferromagnetic transition with a Néel temperature of T_N = 118 K, which results in a rhombohedral structure in space group $R\bar{3}m$ (Shull *et al.*, 1951; Roth, 1958). More recently it has been suggested that, at low temperature, the local structure is even lower symmetry, *e.g.* monoclinic in s.g. *C*2 (Goodwin *et al.*, 2006; Frandsen & Billinge, 2015). Here we see which of these structural results are returned by the structure-mining process.

The heuristic 2 approach is applied, *i.e.* fetching all the atomic structures with Mn and O elements. The rhombohedral MnO model is the best-performing model [MPD No. 41 (Jain *et al.*, 2013) with $R_w = 0.236$, Fig. 8]. The second best fit is the cubic MnO model [COD No. 56 (Zhang, 1999) with
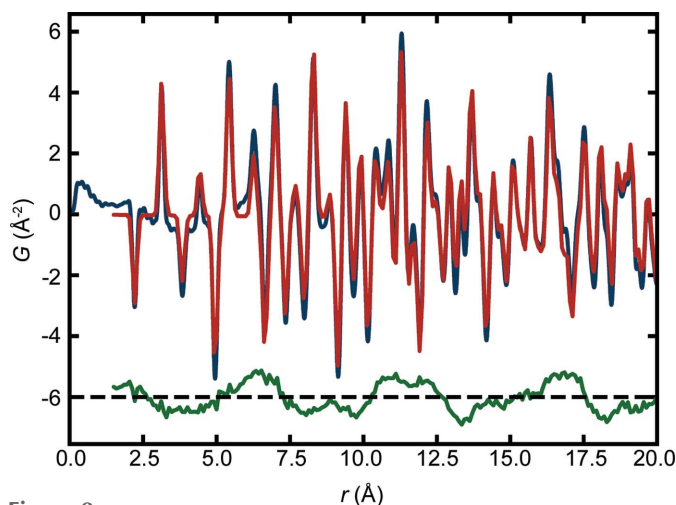

Figure 8
The neutron PDF of the MnO data (blue curve) measured at 15 K with the best-fit calculated atomic PDF (red) for the MPD No. 41, rhombohedral MnO model from heuristic 2. The difference curve is shown offset below (green). Notice the strong magnetic PDF signal in the difference curve, which did not confuse structure-mining.

$R_w = 0.310$]. This correctly reflects the fact that at 15 K the material is expected to be in the rhombohedral phase. The monoclinic s.g. *C*2 model was not returned by structure-mining but this is because it is not in any of the databases. The fit agreements are similar to those reported by Frandsen & Billinge (2015) when the magnetic model is not included in the fit (as is the case here). Therefore, even in the presence of significant magnetic scattering, structure-mining is able to find the correct solution.

Structure-mining was conceived as a structure selection approach and not for finding multiple phases in a sample. However, it is interesting to establish how well it performs when the PDF signal consists of more than one phase. For this test we use the X-ray PDF data set of a vanadium nitride sample (Urbankowski *et al.*, 2017). In the original publication (Urbankowski *et al.*, 2017) it had manually been assigned as consisting of a majority (64%) of V_2N with a structure in space group $P\bar{3}1m$, and a minority (36%) of VN (s.g.: $Fm\bar{3}m$). First we applied the heuristic 2 procedure searching for V-N structures on the measured data. The structure-mining found the correct V_2N structure successfully with $R_w = 0.29$ and with other structures being $R_w > 0.7$. The procedure did not find VN as a candidate structure. We then subtracted the calculated V_2N structure from the measured PDF and carried out structure-mining on the difference. All of the returned structures resulted in values of R_w that were large (0.66 and higher) which is presumably because of the low signal-to-noise ratio in the subtracted data, and the fact that our definition of R_w [equation (2)] does not account for measurement noise. Nonetheless, the top best-fit structure returned by structure-mining was exactly the correct cubic VN phase ($R_w = 0.66$). More details about the results can be found in the supporting information CSV files. Structure-mining was, therefore, successful at finding both the majority phase and the secondary phase. This shows that, at least in favorable

Table 6
The experimental PDF data sets for testing the structure-mining approach with relevant parameters.

Here L is the sample-to-detector distance and Q_{damp} and Q_{broad} are standard fitting parameters for the PDF that come primarily from instrumental resolution effects. The instrument resolution parameters and L of the CuI_2S_4 data are not available.

Composition	Scatterer	Beamline	Q_{damp} (Å^{-1})	Q_{broad} (Å^{-1})	Q_{max} (Å^{-1})	X-ray wavelength (Å)	L (mm)
$BaTiO_3$[†]	X-ray	XPD	0.037	0.017	24.0	0.1867	202.8031
Ti_4O_7	X-ray	XPD	0.041	0.009	25.0	0.1866	202.9990
$NaFeSi_2O_6$[‡]	X-ray	XPD	0.035	0.016	22.0	0.18288	204.2825
$Ba_{0.8}K_{0.2}(Zn_{0.85}Mn_{0.15})_2As_2$[§]	Neutron	NOMAD	0.018	0.019	20.0	—	—
$CuIr_2S_4$[¶]	X-ray	XPD	—	—	25.0	0.183	—
MnO[††]	Neutron	NPDF	0.0198	0.0195	35.0	—	—
V_2N+VN[‡‡]	X-ray	XPD	0.0369	0.0131	25.0	0.1847	205.3939

[†] Lombardi *et al.* (2019).　[‡] Lewis *et al.* (2018).　[§] Frandsen, Gong *et al.* (2016).　[¶] Božin *et al.* (2019).　[††] Frandsen & Billinge (2015).　[‡‡] Urbankowski *et al.* (2017).

circumstances, multi-phase samples may be successfully structure-mined.

We note that it should be straightforward to extend the structure-mining methodology to study the PDFs of organic materials. However, this is not done in the current version. First, for organic material data, the PDF peaks are sharp at low r (intramolecular range) and broad at high r (intermolecular range). To handle this correctly, different ADPs should be applied for two separate regions. In particular the sharp intramolecular peaks cannot be sufficiently fitted, which usually requires some special treatments on parameters such as the correlated atomic motion parameter δ, sratio and ADPs (Prill *et al.*, 2015, 2016). This requires a separation of the molecule from its neighbors which is currently done manually and an automated approach needs to be developed for structure-mining to work. Second, the current heuristics, searching by compositions and elements, are not suitable for organic materials. For example, searching 'C-H-O' would return too many candidates, about 16 000 entries from COD and MPD databases. Development of new search heuristics, such as searching by organic molecule name, is possible but needs some future work. Third, there are limited organic material entries in the currently supported COD and MPD databases. Supporting some more comprehensive organic structural databases, such as the Cambridge Structural Database (CSD) (Allen, 2002), is necessary for finding organic compounds.

We have shown that structure-mining is able to find the desired structures from the mine on a range of test cases. We now consider its robustness against factors that might prove problematic, specifically data collected at a different temperature to the data in the mine and data measured under a range of different experimental conditions. Structure-mining seems to work well on data collected at different temperatures and so is robust against differences in lattice parameters and ADPs due to temperature effects. This assertion is supported by the $BaTiO_3$ example described above where structure-mining found all the barium titanate structural variants, which were measured at a range of temperatures from 15 K (COD No. 24 in Table 3) (Kwei *et al.*, 1993) to 1000 K (COD No. 14 in Table 3) (Edwards *et al.*, 1951) when compared with the data measured at room temperature. Structure-mining also

performed well in tests where data were coming from a wide range of different instruments and measurement conditions. In these tests the Q_{max} values for the data varied between 18.6 $\text{Å}^{-1} < Q_{max} < 25.0$ Å^{-1}, and the ranges of instrument resolution parameters are $0.038 < Q_{damp} < 0.058$ Å^{-1} and $0.0 < Q_{broad} < 0.048$ Å^{-1} (Yang *et al.*, 2013; Frandsen, Gong *et al.*, 2016; Quinson *et al.*, 2018). All of them worked well in structure-mining which successfully found the correct structures regardless of the fact that different Q_{max} and instrument resolution parameters were in effect. Thus the method should work in general for many other instruments even when Q_{damp} and Q_{broad} vary from one instrument to another.

4. Conclusion

In this paper, we have demonstrated a new approach, called structure-mining, for automated screening of large numbers of candidate structures to the atomic pair distribution function (PDF) data, by automatically fetching candidate structures from structural databases and automatically performing PDF structure refinements to obtain the best agreement between calculated PDFs of the structures and the measured PDF under study. The approach has been successfully tested on the PDFs of a variety of challenging materials, including complex oxide nanoparticles and nanowires, low-symmetry structures, complicated doped, magnetic, locally distorted and mixed-phase materials. This approach could greatly speed up and extend the traditional structure searching workflow and enable the possibility of highly automated and high-throughput real-time PDF analysis experiments in the future.

APPENDIX A
See Table 6.

References

Akimoto, J., Gotoh, Y. & Oosawa, Y. (1994). *Acta Cryst.* C**50**, 160–161.

Allen, F. H. (2002). *Acta Cryst.* B**58**, 380–388.

Altomare, A., Corriero, N., Cuocci, C., Falcicchio, A., Moliterni, A. & Rizzi, R. (2015). *J. Appl. Cryst.* **48**, 598–603.

Altomare, A., Cuocci, C., Giacovazzo, C., Moliterni, A. & Rizzi, R. (2008). *J. Appl. Cryst.* **41**, 815–817.

Andersson, S. (1960). *Acta Chem. Scand.* **14**, 1161–1172.

Andersson, S., Collén, B., Kuylenstierna, U. & Magnéli, A. (1957). *Acta Chem. Scand.* **11**, 1641–1652.

Andersson, S. & Magnéli, A. (1956). *Naturwissenschaften*, **43**, 495–496.

Aoun, B. (2016). *J. Comput. Chem.* **37**, 1102–1111.

Balachandran, J., Lin, L., Anchell, J. S., Bridges, C. A. & Ganesh, P. (2017). *J. Phys. Chem. C*, **121**, 26637–26647.

Banerjee, S., Liu, C.-H., Jensen, K. M. Ø., Juhás, P., Lee, J. D., Tofanelli, M., Ackerson, C. J., Murray, C. B. & Billinge, S. J. L. (2020). *Acta Cryst.* A**76**, 24–31.

Barr, G., Gilmore, C. J. & Paisley, J. (2004). *J. Appl. Cryst.* **37**, 665–668.

Beecher, A. N., Yang, X., Palmer, J. H., LaGrassa, A. L., Juhás, P., Billinge, S. J. L. & Owen, J. S. (2014). *J. Am. Chem. Soc.* **136**, 10645–10653.

Belsky, A., Hellenbrandt, M., Karen, V. L. & Luksch, P. (2002). *Acta Cryst.* B**58**, 364–369.

Bergerhoff, G., Hundt, R., Sievers, R. & Brown, I. D. (1983). *J. Chem. Inf. Comput. Sci.* **23**, 66–69.

Billinge, S. J. L. (2019). *International Tables for Crystallography*, Vol. H, edited by C. Gilmore, J. Kaduk & H. Schenk, pp. 649–672. Chester: International Union of Crystallography.

Billinge, S. J. L., DiFrancesco, R. G., Kwei, G. H., Neumeier, J. J. & Thompson, J. D. (1996). *Phys. Rev. Lett.* **77**, 715–718.

Billinge, S. J. L. & Kanatzidis, M. G. (2004). *Chem. Commun.* pp. 749–760.

Boullay, P., Lutterotti, L., Chateigner, D. & Sicard, L. (2014). *Acta Cryst.* A**70**, 448–456.

Božin, E. S., Yin, W. G., Koch, R. J., Abeykoon, M., Hor, Y. S., Zheng, H., Lei, H. C., Petrovic, C., Mitchell, J. F. & Billinge, S. J. L. (2019). *Nat. Commun.* **10**, 1–7.

Chichagov, A. V., Varlamov, D. A., Dilanyan, R. A., Dokina, T. N., Drozhzhina, N. A., Samokhvalova, O. L. & Ushakovskaya, T. V. (2001). *Crystallogr. Rep.* **46**, 876–879.

Chupas, P. J., Qiu, X., Hanson, J. C., Lee, P. L., Grey, C. P. & Billinge, S. J. L. (2003). *J. Appl. Cryst.* **36**, 1342–1347.

Clark, J. R., Appleman, D. E. & Papike, J. J. (1969). *Miner. Soc. Am. Spec. Pap.* **2**, 31–50.

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

Cranswick, L. M. D. (2008). *Powder Diffraction: Theory and Practice*, edited by R. E. Dinnebier & S. J. L. Billinge, pp. 494–570. Cambridge: Royal Society of Chemistry.

Cravillon, J., Schröder, C. A., Nayuk, R., Gummel, J., Huber, K. & Wiebcke, M. (2011). *Angew. Chem. Int. Ed.* **50**, 8067–8071.

Curtarolo, S., Hart, G. L. W., Nardelli, M. B., Mingo, N., Sanvito, S. & Levy, O. (2013). *Nat. Mater.* **12**, 191–201.

Curtarolo, S., Setyawan, W., Wang, S., Xue, J., Yang, K., Taylor, R. H., Nelson, L. J., Hart, G. L. W., Sanvito, S., Buongiorno-Nardelli, M., Mingo, N. & Levy, O. (2012). *Comput. Mater. Sci.* **58**, 227–235.

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

Donohue, J., Miller, S. J. & Cline, R. F. (1958). *Acta Cryst.* **11**, 693–695.

Downs, R. T. & Hall-Wallace, M. (2003). *Am. Mineral.* **88**, 247–250.

Edwards, J. W., Speiser, R. & Johnston, H. L. (1951). *J. Am. Chem. Soc.* **73**, 2934–2935.

Egami, T. & Billinge, S. J. L. (2012). *Underneath the Bragg Peaks: Structural Analysis of Complex Materials*, 2nd ed. Amsterdam: Elsevier.

Erchak, M., Fankuchen, I. & Ward, R. (1946). *J. Am. Chem. Soc.* **68**, 2085–2093.

Farrow, C. L. & Billinge, S. J. L. (2009). *Acta Cryst.* A**65**, 232–239.

Farrow, C. L., Juhás, P., Liu, J., Bryndin, D., Božin, E. S., Bloch, J., Proffen, T. & Billinge, S. J. L. (2007). *J. Phys. Condens. Matter*, **19**, 335219.

Farrow, C. L., Shaw, M., Kim, H.-J., Juhás, P. & Billinge, S. J. L. (2011). *Phys. Rev. B*, **84**, 134105.

Fletcher, R. (1971). *Modified Marquardt Subroutine for Nonlinear Least Squares.* Technical Report AERE-R-6799. Atomic Energy Research Establishment, Harwell, UK.

Floudas, C. A. & Pardalos, P. M. (2001). *Encyclopedia of Optimization.* Boston, Massachusetts, USA: Springer Science & Business Media.

Frandsen, B. A. & Billinge, S. J. L. (2015). *Acta Cryst.* A**71**, 325–334.

Frandsen, B. A., Brunelli, M., Page, K., Uemura, Y. J., Staunton, J. B. & Billinge, S. J. L. (2016). *Phys. Rev. Lett.* **116**, 197204.

Frandsen, B. A., Gong, Z., Terban, M. W., Banerjee, S., Chen, B., Jin, C., Feygenson, M., Uemura, Y. J. & Billinge, S. J. L. (2016). *Phys. Rev. B*, **94**, 094102.

Frandsen, B. A., Yang, X. & Billinge, S. J. L. (2014). *Acta Cryst.* A**70**, 3–11.

Frazer, B. C., Danner, H. R. & Pepinsky, R. (1955). *Phys. Rev.* **100**, 745–746.

Friščić, T., Halasz, I., Beldon, P. J., Belenguer, A. M., Adams, F., Kimber, S. A., Honkimäki, V. & Dinnebier, R. E. (2013). *Nat. Chem.* **5**, 66–73.

Gagin, A., Allen, A. J. & Levin, I. (2014). *J. Appl. Cryst.* **47**, 619–629.

Gauss, C. F. (1809). *Theoria motus corporum coelestium in sectionibus conicis solem ambientium.* Perthes et Besser.

Gilmore, C. J., Barr, G. & Paisley, J. (2004). *J. Appl. Cryst.* **37**, 231–242.

Goodwin, A. L., Tucker, M. G., Dove, M. T. & Keen, D. A. (2006). *Phys. Rev. Lett.* **96**, 047209.

Gražulis, S., Chateigner, D., Downs, R. T., Yokochi, A. F. T., Quirós, M., Lutterotti, L., Manakova, E., Butkus, J., Moeck, P. & Le Bail, A. (2009). *J. Appl. Cryst.* **42**, 726–729.

Grin, Y., Müller-Buschbaum, H. & Schnering, H. G. (2014). *Z. Naturforsch. B*, **52**, 153–156.

Hammersley, A. P. (2016). *J. Appl. Cryst.* **49**, 646–652.

Hanawalt, J. D., Rinn, H. W. & Frevel, L. K. (1938). *Ind. Eng. Chem. Anal. Ed.* **10**, 457–512.

Hayward, S. A., Redfern, S. A. T., Stone, H. J., Tucker, M. G., Whittle, K. R. & Marshall, W. G. (2005). *Z. Kristallogr.* **220**, 735–739.

Hellmann, A., Löhken, A., Wurth, A. & Mewis, A. (2007). *Z. Naturforsch. B*, **62**, 155–161.

Hodeau, J. L. & Marezio, M. (1979). *J. Solid State Chem.* **29**, 47–62.

Hohenberg, P. & Kohn, W. (1964). *Phys. Rev.* **136**, B864–B871.

ICDD (2019). *The Powder Diffraction File.* International Centre for Diffraction Data, Newton Square, Pennsylvania, USA.

Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G. & Persson, K. A. (2013). *APL Mater.* **1**, 011002.

Jensen, K. M. O., Christensen, M., Juhás, P., Tyrsted, C., Bøjesen, E. D., Lock, N., Billinge, S. J. L. & Iversen, B. B. (2012). *J. Am. Chem. Soc.* **134**, 6785–6792.

Jones, E., Oliphant, T. & Peterson, P. (2001). *SciPy: Open Source Scientific Tools for Python.* http://www.scipy.org/.

Juhás, P., Davis, T., Farrow, C. L. & Billinge, S. J. L. (2013). *J. Appl. Cryst.* **46**, 560–566.

Juhás, P., Farrow, C. L., Yang, X., Knox, K. R. & Billinge, S. J. L. (2015). *Acta Cryst.* A**71**, 562–568.

Katsenis, A. D., Puškarić, A., Štrukil, V., Mottillo, C., Julien, P. A., Užarević, K., Pham, M. H., Do, T. O., Kimber, S. A., Lazić, P., Magdysyuk, O., Dinnebier, R. E., Halasz, I. & Friščić, T. (2015). *Nat. Commun.* **6**, 6662.

Keen, D. A. & Goodwin, A. L. (2015). *Nature*, **521**, 303–309.

Kirklin, S., Saal, J. E., Meredig, B., Thompson, A., Doak, J. W., Aykol, M., Rühl, S. & Wolverton, C. (2015). *NPJ Comput. Mater.* **1**, 15010.

Kohn, W. & Sham, L. J. (1965). *Phys. Rev.* **140**, A1133–A1138.

Kwei, G. H., Lawson, A. C., Billinge, S. J. L. & Cheong, S.-W. (1993). *J. Phys. Chem.* **97**, 2368–2377.

Laveda, J. V., Johnston, B., Paterson, G. W., Baker, P. J., Tucker, M. G., Playford, H. Y., Jensen, K. M. O., Billinge, S. J. L. & Corr, S. A. (2018). *J. Mater. Chem. A*, **6**, 127–137.

Levenberg, K. (1944). *Q. Appl. Math.* **2**, 164–168.

Lewis, C. S., Moronta, D., Terban, M. W., Wang, L., Yue, S., Zhang, C., Li, Q., Corrao, A., Billinge, S. J. L. & Wong, S. S. (2018). *CrystEngComm*, **20**, 223–236.

Lewis, J. W., Payne, J. L., Evans, I. R., Stokes, H. T., Campbell, B. J. & Evans, J. S. O. (2016). *J. Am. Chem. Soc.* **138**, 8031–8042.

Lombardi, J., Yang, L., Pearsall, F. A., Farahmand, N., Gai, Z., Billinge, S. J. L. & O'Brien, S. (2019). *Chem. Mater.* **31**, 1318–1335.

Lutterotti, L., Pillière, H., Fontugne, C., Boullay, P. & Chateigner, D. (2019). *J. Appl. Cryst.* **52**, 587–598.

Marezio, M. & Dernier, P. D. (1971). *J. Solid State Chem.* **3**, 340–348.

Marezio, M., McWhan, D. B., Dernier, P. D. & Remeika, J. P. (1973). *J. Solid State Chem.* **6**, 213–221.

Marezio, M., Tranqui, D., Lakkis, S. & Schlenker, C. (1977). *Phys. Rev. B*, **16**, 2811–2821.

Marquardt, D. W. (1963). *J. Soc. Ind. Appl. Math.* **11**, 431–441.

Marquart, R. G., Katsnelson, I., Milne, G. W. A., Heller, S. R., Johnson, G. G. Jr & Jenkins, R. (1979). *J. Appl. Cryst.* **12**, 629–634.

Masadeh, A. S., Božin, E. S., Farrow, C. L., Paglia, G., Juhas, P., Billinge, S. J. L., Karkamkar, A. & Kanatzidis, M. G. (2007). *Phys. Rev. B*, **76**, 115413.

McCarthy, A. C., Downs, R. T., Thompson, R. M. & Redhammer, G. J. (2008). *Am. Mineral.* **93**, 1829–1837.

McGreevy, R. L. (2001). *J. Phys. Condens. Matter*, **13**, R877.

McGreevy, R. L. & Pusztai, L. (1988). *Mol. Simul.* **1**, 359–367.

Moskowitz, S. L. (2009). *The Advanced Materials Revolution: Technology and Economic Growth in the Age of Globalization.* Hoboken, New Jersey, USA: John Wiley & Sons.

Neder, R. B. & Korsunskiy, V. I. (2005). *J. Phys. Condens. Matter*, **17**, S125–S134.

Nestola, F., Rotiroti, N., Bruno, M., Tribaudino, M., van Smaalen, S., Ohashi, H. & Redhammer, G. J. (2007). *Am. Mineral.* **92**, 560–569.

Nestola, F., Tribaudino, M., Boffa Ballaran, T., Liebske, C. & Bruno, M. (2007). *Am. Mineral.* **92**, 1492–1501.

Neuefeind, J., Feygenson, M., Carruth, J., Hoffmann, R. & Chipley, K. K. (2012). *Nucl. Instrum. Methods Phys. Res. B*, **287**, 68–75.

Nishimura, K., Yamada, I., Oka, K., Shimakawa, Y. & Azuma, M. (2014). *J. Phys. Chem. Solids*, **75**, 710–712.

Ohashi, H., Fujita, T. & Osawa, T. (1983). *J. Jpn. Assoc. Min. Petr. Econ. Geol.* **78**, 159–163.

Ohashi, H., Osawa, T. & Sato, A. (1995). *Acta Cryst.* C**51**, 2476–2477.

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.

Ong, S. P., Cholia, S., Jain, A., Brafman, M., Gunter, D., Ceder, G. & Persson, K. A. (2015). *Comput. Mater. Sci.* **97**, 209–215.

Ong, S. P., Richards, W. D., Jain, A., Hautier, G., Kocher, M., Cholia, S., Gunter, D., Chevrier, V. L., Persson, K. A. & Ceder, G. (2013). *Comput. Mater. Sci.* **68**, 314–319.

Ozawa, T., Olmstead, M. M., Brock, S. L., Kauzlarich, S. M. & Young, D. M. (1998). *Chem. Mater.* **10**, 392–396.

Page, K., Hood, T. C., Proffen, Th. & Neder, R. B. (2011). *J. Appl. Cryst.* **44**, 327–336.

Peterson, P. F., Gutmann, M., Proffen, Th. & Billinge, S. J. L. (2000). *J. Appl. Cryst.* **33**, 1192.

Petkov, V. & Bakaltchev, N. (1990). *J. Appl. Cryst.* **23**, 138–140.

Powell, M. J. D. (1964). *Comput. J.* **7**, 155–162.

Prill, D., Juhás, P., Billinge, S. J. L. & Schmidt, M. U. (2016). *Acta Cryst.* A**72**, 62–72.

Prill, D., Juhás, P., Schmidt, M. U. & Billinge, S. J. L. (2015). *J. Appl. Cryst.* **48**, 171–178.

Proffen, Th. & Billinge, S. J. L. (1999). *J. Appl. Cryst.* **32**, 572–575.

Proffen, T., Egami, T., Billinge, S. J. L., Cheetham, A. K., Louca, D. & Parise, J. B. (2002). *Appl. Phys. Mater. Sci. Process.* **74**, s163–s165.

Proffen, Th. & Neder, R. B. (1997). *J. Appl. Cryst.* **30**, 171–175.

Quinson, J., Kacenauskaite, L., Christiansen, T. L., Vosch, T., Arenz, M. & Jensen, K. M. Ø. (2018). *ACS Omega*, **3**, 10351–10356.

Redhammer, G. J., Amthauer, G., Lottermoser, W. & Treutmann, W. (2000). *Eur. J. Mineral.* **12**, 105–120.

Redhammer, G. J., Amthauer, G., Roth, G., Tippelt, G. & Lottermoser, W. (2006). *Am. Mineral.* **91**, 1271–1292.

Roth, W. L. (1958). *Phys. Rev.* **110**, 1333–1341.

Rotter, M., Tegel, M. & Johrendt, D. (2008). *Phys. Rev. Lett.* **101**, 107006.

Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. (2013). *JOM*, **65**, 1501–1509.

Saha, D., Jensen, K. M. O., Tyrsted, C., Bøjesen, E. D., Mamakhel, A. H., Dippel, A.-C., Christensen, M. & Iversen, B. B. (2014). *Angew. Chem. Int. Ed.* **53**, 3667–3670.

Setyawan, W. & Curtarolo, S. (2010). *Comput. Mater. Sci.* **49**, 299–312.

Shirane, G., Danner, H. & Pepinsky, R. (1957). *Phys. Rev.* **105**, 856–860.

Shoemaker, D. P., Hu, Y.-J., Chung, D. Y., Halder, G. J., Chupas, P. J., Soderholm, L., Mitchell, J. F. & Kanatzidis, M. G. (2014). *Proc. Natl Acad. Sci. USA*, **111**, 10922–10927.

Shull, C. G., Strauser, W. A. & Wollan, E. O. (1951). *Phys. Rev.* **83**, 333–345.

Simon, C. M., Kim, J., Gomez-Gualdron, D. A., Camp, J. S., Chung, Y. G., Martin, R. L., Mercado, R., Deem, M. W., Gunter, D., Haranczyk, M., Sholl, D. S., Snurr, R. Q. & Smit, B. (2015). *Energy Environ. Sci.* **8**, 1190–1199.

Soper, A. K. (2005). *Phys. Rev. B*, **72**, 104204.

Srilakshmi, C., Saraf, R., Prashanth, V., Rao, G. M. & Shivakumara, C. (2016). *Inorg. Chem.* **55**, 4795–4805.

Sueno, S., Cameron, M., Papike, J. J. & Prewitt, C. T. (1973). *Am. Mineral.* **58**, 649–664.

Terban, M. W., Shi, C., Silbernagel, R., Clearfield, A. & Billinge, S. J. L. (2017). *Inorg. Chem.* **56**, 8837–8846.

Terban, W. M., Banerjee, D., Ghose, S., Medasani, B., Shukla, A., Legg, B. A., Zhou, Y., Zhu, Z., Sushko, M. L., Yoreo, J. J. D., Liu, J., Thallapally, P. K. & Billinge, S. J. L. (2018). *Nanoscale*, **10**, 4291–4300.

Thompson, R. M. & Downs, R. T. (2004). *Am. Mineral.* **89**, 614–628.

Toby, B. H. (2005). *J. Appl. Cryst.* **38**, 1040–1041.

Toby, B. H., Dmowski, W., Egami, T., Jorgensen, J. D., Subramanian, M. A., Gopalakrishnan, J., Sleight, A. W. & Parise, J. B. (1989). *Physica C*, **162–164**, 101–102.

Tucker, M. G., Keen, D. A., Dove, M. T., Goodwin, A. L. & Hui, Q. (2007). *J. Phys. Condens. Matter*, **19**, 335218.

Urbankowski, P., Anasori, B., Hantanasirisakul, K., Yang, L., Zhang, L., Haines, B., May, S. J., Billinge, S. J. L. & Gogotsi, Y. (2017). *Nanoscale*, **9**, 17722–17730.

Vanderah, T. A., Siegrist, T., Roth, R. S., Ramirez, A. P. & Geyer, R. G. (2004). *Eur. J. Inorg. Chem.* **2004**, 2434–2441.

Wada, S., Yano, M., Suzuki, T. & Noma, T. (2000). *J. Mater. Sci.* **35**, 3889–3902.

White, A. (2012). *MRS Bull.* **37**, 715–716.

Woodward, D. I., Reaney, I. M., Yang, G. Y., Dickey, E. C. & Randall, C. A. (2004). *Appl. Phys. Lett.* **84**, 4650–4652.

Xiao, C. J., Jin, C. Q. & Wang, X. H. (2008). *Mater. Chem. Phys.* **111**, 209–212.

Yang, X., Juhás, P., Farrow, C. & Billinge, S. J. L. (2015). arXiv:1402.3163.

Yang, X., Masadeh, A. S., McBride, J. R., Božin, E. S., Rosenthal, S. J. & Billinge, S. J. L. (2013). *Phys. Chem. Chem. Phys.* **15**, 8480–8486.

Young, C. A. & Goodwin, A. L. (2011). *J. Mater. Chem.* **21**, 6464–6476.

Young, R. A. (1993). *The Rietveld Method. IUCr Monographs on Crystallography*, No. 5. Oxford: IUCr/Oxford University Press.

Zeng, L. & Jiang, Y. (1991). *Beijing Gongye Daxue Xuebao*, **17**, 32–37.

Zhang, H., Gilbert, B., Huang, F. & Banfield, J. F. (2003). *Nature*, **424**, 1025–1029.

Zhang, J. (1999). *Phys. Chem. Miner.* **26**, 644–648.

Zhao, K., Deng, Z., Wang, X. C., Han, W., Zhu, J. L., Li, X., Liu, Q. Q., Yu, R. C., Goko, T., Frandsen, B., Liu, L., Ning, F., Uemura, Y. J., Dabkowska, H., Luke, G. M., Luetkens, H., Morenzoni, E., Dunsiger, S. R., Senyshyn, A., Böni, P. & Jin, C. Q. (2013). *Nat. Commun.* **4**, 1442.