Island model with genetic algorithm for solution of crystal structure from X-ray powder diffraction data

IOP Conf. Series: Materials Science and Engineering **537** (2019) 022002 doi:10.1088/1757-899X/537/2/022002

Island model with genetic algorithm for solution of crystal structure from X-ray powder diffraction data

Aleksandr N Zaloga[1,3], Sergey V Burakov[2], Igor S Yakimov[1], Konstantin A Gusev[2] and Petr S Dubinin[1]

[1] Siberian Federal University, Krasnoyarsk, Russia
[2] Reshtnev Siberian State University of Science and Technology, Krasnoyarsk, Russia

[3] E-mail: zaloga@yandex.ru

Abstract. In this paper, we consider the problem of the study of polycrystalline substances: restoration of a substance atomic structure by full-profile analysis of powder diffraction data. This task is specific since it is not necessary to find very good solutions on average, but it is necessary to find the best one at least sometimes. To solve this problem, it is proposed to use an evolutionary algorithm based on the cooperative island model. The article describes the main stages and features of the algorithm and notes the qualitative advantages of this model in comparison with other methods (including evolutionary). The description of innovations proposed and the results of computational experiments are given. Conclusions from the experimental results are given, and further prospects for improving the efficiency of this method were noted.

1. Introduction

To develop materials with specified properties, we need to have fundamental knowledge that determines the structure and properties of matter at the atomic level. Most modern materials are synthesized in the form of polycrystals and nanocrystals. To determine their atomic-crystal structure, various methods of powder diffraction are used. These methods work in direct, inverse or dual space. The determination of the crystal structure of a substance from powder diffraction data suggests the optimization of atomic positions in the independent part of the unit cell of a crystal.

Direct space methods for solving crystal structures from powder data are usually used in cases where it is not possible to capture a good quality diffraction pattern since the reverse and dual space methods are sensitive to the quality of the input data. Direct space methods began their development with the advent of high-speed computers in the 1990s, including the Monte-Carlo method [1, 2], genetic algorithms [3, 4], and simulated annealing [5, 6]. Currently, computer programs FOX [7], DASH [8], and TOPAS [9], which use the simulated annealing method, are widely used. Genetic algorithms (GA) are implemented in the well-known programs EAGER [10], GEST [11], MAUD [12].

The essence of the GA is to simulate the operations of natural biological selection: pairwise crossing, mutations, and selection of the best test structural models for obtaining new generations of evolution. A common problem with these methods is the deterioration of convergence with increasing complexity of the structures being solved, which is associated with a nonlinear increase in the probability of stagnation in the numerous local minima of the hypersurface of the the objective function. Therefore, in practice, their application is limited by the number of degrees of freedom of the determined atomic coordinates

(as a rule, no more than 30-50). Simulated annealing methods are the most common and easy to use among them. In [13], it was noted that the disadvantage of the existing implementations of genetic algorithms is the need for the user to configure numerous parameters that control the evolution process.

This paper suggests ways to improve the genetic algorithm itself, which can make it less dependent on non-optimal settings. To search for the coordinates of atoms of a substance, it is proposed to use an evolutionary algorithm with an island model of cooperation of parallel operating evolutionary processes with the exchange of part of the solutions among the islands. This approach leads not only to a quantitative improvement in the efficiency of the algorithm due to the wide coverage of the search space in populations on different islands but also to the ability to find solutions to such complex structures that could not be solved using evolutionary processes without the exchange. A new local search (LS) operator is also proposed, which shows greater efficiency for the problem of crystal structure determination than the classical local search operator.

The article is structured as follows: Chapter 2 discusses the specifics of the problem and the proposed approach to its solution; Chapter 3 describes the essence of the proposed cooperative exchange among evolutionary processes; Chapter 4 describes the new local search operator used to optimize the crystal structure; Chapter 5 presents the results of the experiments performed; and then there are conclusions, acknowledgment, and references.

2. Methods

In the problems of determining the atomic crystal structure, the known initial data are the chemical formula, the parameters of the crystal unit cell, the symmetry group, and the experimental diffractogram. It is required to find such optimal positions of atoms in the crystal cell at which the differences between the experimental and calculated diffraction patterns would be minimal. Figure 1 shows an example of a powder diffraction pattern, its model and a difference between them.

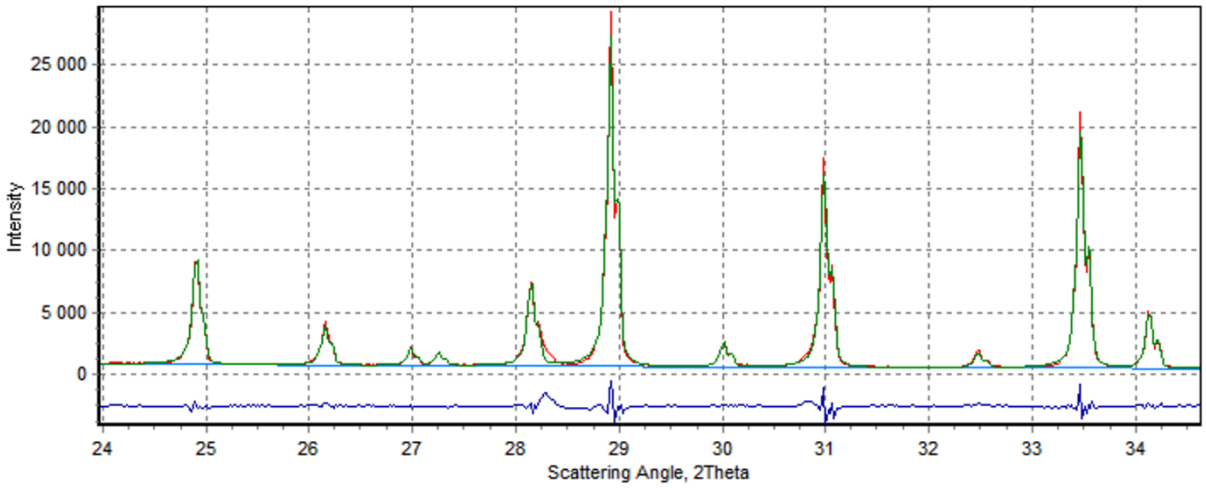

Figure 1. Fragment of powder diffraction pattern of the substance $Ca_2Al_3F_6O$ and its well-refined model. The colors of the lines indicate: red – experimental profile, green – model profile, blue – difference line between experimental and model profiles.

An experimental diffraction pattern is the result of a diffraction experiment in which a polycrystalline material (in other words, a powder) is irradiated with a beam of X-rays, electrons or neutrons from different angles, and the reflection of the beam is collected by the detector. That is, the powder diffraction pattern is the dependence of the intensity of the reflected radiation on beam incidence angle. The position of the diffraction maxima on the diffractogram correspond to the interplanar distances in the crystal and is determined by the Wulff-Bragg's condition:

$$2d_{hkl}\, sin\theta = n\lambda \quad (1)$$

where d_{hkl} is an interplanar spacing along the crystallographic direction (h, k, l), θ is a diffraction angle, λ is the wavelength of the radiation used (for example, for common laboratory instruments, a wavelength of ≈ 1.56 Å that corresponds to an X-ray tube with a copper anode is often used).

The Rietveld method is used To optimize the crystal model. The following functional is minimized in this method:

$$\Phi(\mathbf{P}) = \sum_j w_j \left(Y_o(\overline{\mathbf{P}}, \theta_j) - Y_c(\theta_j)\right)^2 \to \min_{\mathbf{P}}, \tag{2}$$

where j is the number of the diffraction profile point; Y_o and Y_c are observed and calculated from structural models profile intensities, θ is the diffraction angle, w_i is the weight; P is the crystal model parameter vector; the background of the diffraction pattern is deleted by differentiation.

The intensity at each point of the diffraction pattern is calculated by the following expression:

$$Y_c(\overline{P}, \theta_j) = S\sum_{hkl} I_{hkl}(F_{hkl}(\overline{X})) \cdot \Omega_i(\overline{P}_{hkl}, \theta_j), \tag{3}$$

where S is scale factor, I_{hkl} is repeatability factor for this family of crystallographic planes (h, k, l), Ω is profile function depending on the profile parameters P_{hkl} and the diffraction angle θ, F_{hkl} is structural amplitude, which depends on the coordinates of the atoms in the crystal X.

The structural amplitude is calculated as follows:

$$F_{hkl} = \sum_j f_j e^{-2\pi i(hx_j + ky_j + lz_j)} \tag{4}$$

where f_j is scattering energy of the j-th atom, depending on its electron density; x_j, y_j, z_j are coordinates of the j-th atom in the cell in the range from 0 to 1. For given parameters of the peak shape, the intensities of the diffraction maxima depend on the positions of atoms in the crystal cell.

The convergence criterion is the profile R_{wp} factor of the Rietveld method:

$$R_{wp} = \left(\frac{\sum_j w_j \left(Yo_j - Yc_j\right)^2}{\sum_j w_j \left(Yo_j\right)^2} \right)^{1/2} \tag{5}$$

where j is the number of the diffraction profile point; w_j is weighting factor reflecting the reliability of the measurement, usually equal to $1/Yo_j$.

Thus, we can consider the task of restoring the crystal structure of a substance as an optimization problem with some objective function, as a rule, of the form (2). Due to the symmetry of many crystal lattices, the solution of such a problem will often not be the only one. That is, the task in question is multiextreme (having a large number of local, and usually several global minima).

The deterministic Rietveld method is based on the nonlinear method of least squares, and the disadvantage of this method is in rapid convergence to a true solution only with a fairly good initial approximation. However, the dimension of the search space (especially for complex substances with a major number of atoms) is very large, and the region of convergence is very narrow – it is necessary to get the point with a radius of less than 0.03 of the variation range of each parameter [14]. In [15], it is noted that the ambitious goal of powder method development is the structural determination of more than 200 parameters. However, at this time, the determination of complex structures usually requires the use of additional methods, such as NMR spectroscopy data and the computationally-capacious method of DFT (density functional theory).

Using evolutionary algorithms to solve the problem of structural analysis allows generating a large number of solutions. Then each of the solutions can be subjected to both evolutionary and deterministic

search for the minimum of the objective function. This allows to save resources, in contrast to the algorithms for a complete enumeration of promising solutions, and eventually allows to solve more complex problems.

The algorithm scheme is standard and includes the following main steps:

1. Initialization of the population at random – at the first stage, a given number of points in the search hyperspace are randomly generated.
2. Estimation of population suitability – calculating the value of the objective function for the entire solution set. As well as a local search for a certain number of promising solutions.
3. Execution of genetic operators: selection, recombination, mutation.
4. Go to step 2.

Evolutionary algorithms have shown their effectiveness in solving many optimization (especially multi-extremal) problems. Such algorithms allow running parallel execution of the algorithm by distributing the most "expensive" operations (calculating the value of the objective function) across several computational nodes. Thus, it can significantly reduce the runtime of the program. A natural extension of this approach is the implementation of the implementation of the evolutionary algorithm on each of the computing nodes of a computer, that is, the parallel operation of several evolutionary algorithms. In this case, the efficiency of the algorithm is increased multiple of the number of computational nodes.

As shown in [16], the implementation of additional exchange (migration) among parallel-working populations gives a synergistic effect, and such an algorithm allows solving crystal structures that could not be solved without exchange. In this case, the scheme of the algorithm remains the same, except that each population exchange a part of the solutions with one of the other populations after point 2, and then the genetic operators are executed. This approach makes it possible to use the resources of modern multi-core PCs and computing clusters efficiently. It should be noted that the implementation of such a scheme of the GA work is possible on one computing node: the logic of the algorithm will remain the same.

3. Island model of an evolutionary algorithm

To implement migration between different populations, we use an oriented connected graph, in which each vertex represents a certain population (or computational node), and each edge of the graph determines the migration direction of a given number of individuals from one population to another. Moreover, each vertex of the graph must be finite (at least for one edge), and it may turn out that some vertices will not be initial at all. This model is also called the complete network topology; it corresponds to the natural processes in wildlife. An example of the exchange scheme for the proposed island model is presented in figure 2.

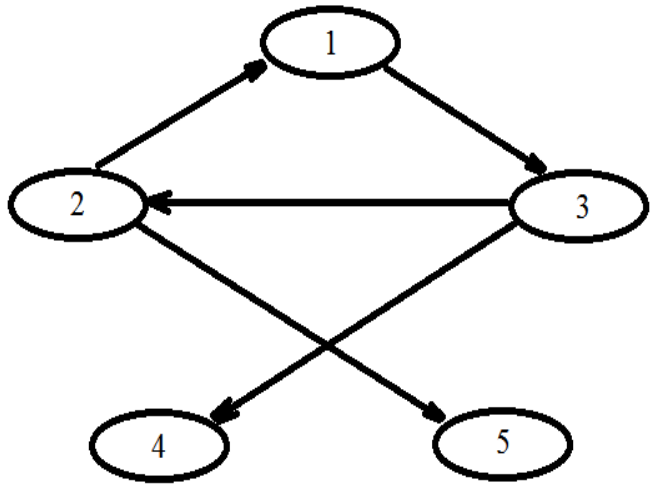

Figure 2. An example of an exchange between five islands. It shows that in one act of exchange, some islands may not be donors of solutions, while others may send their best individuals multiple times.

At the second stage of the evolutionary algorithm, a suitability estimation is made, after which a predetermined number of individuals in the population is selected. The index of the island from which migration will take place is randomly generated for each island (this value should not be equal to the current island). Then a certain number of the best individuals of one population migrate to another one by adding this group of individuals to receiving population.

The migration interval, generally speaking, is an additional parameter for the work of the island evolutionary algorithm. The migration interval is a parameter interdependent with some other parameters (for example, with a mutation, the "depth" of the local search and the frequency of its launch). So, with large migration intervals, each of the evolutionary processes works longer in isolation, and this allows to find more local/global minima, but increases the runtime of the algorithm and reduces the accuracy of the result. Otherwise, the algorithm converges to some solution of the problem more quickly, but the probability of capturing local minima and the stagnation of the algorithm increases.

Previously, we used an approach to the exchange of solutions in a multi-population genetic algorithm (MPGA) based on the accumulation of the best solutions on the managing process and their periodic distribution to working processes. A detailed description of it is given in [16]. The 5th part of the article describes a comparison of the results obtained in different versions of MGPA.

4. Modification of the local search operator for crystal structure optimization

The specificity of the problem is that the value of the objective function is the sum of the profile R_{wp} factor of the Rietveld method (which depends on the proximity of atomic coordinates to true positions in the unit cell) and the value of the penalty for violating minimal interatomic distances (i.e., the relative positions of the atoms). The classic local search operator minimizes only by the R_{wp} factor. Its application to solutions with good interatomic distances and with the wrong atoms arrangement relative to true positions often leads to the situation when both the correct positions of atoms are not found (getting into a local minimum) and correct interatomic distances are destroyed.

Therefore, we propose an operator which cyclically shift all atoms of the current best solution by the same random value along each of the three directions of the crystal cell (we selected 30 shifts: 10 along each of the directions). The reason for this is the fact that crystalline substances consist of repeating identical cells. When simulating a crystal, it suffices to describe the atomic positions in only one of such cells. If at varying the atomic coordinate goes beyond the cell (say, it becomes equal to 1.25 of the cell size), then it will be fair to discard the integer part of this coordinate and take into account only the fractional part - the position inside the cell (in this example, the 1.25 coordinate will be equivalent to the 0.25 coordinate). This is called translational symmetry. The displacement of all atoms along one of the directions of the cell by the same value will mean the preservation of interatomic distances due to the effect of translational symmetry. In some cases, it can lead to a configuration of atoms, which will be closer to the correct solution than before the displacement. In order to verify this, the resulting configuration is refined after each shift by the Rietveld method (with two LS cycles). If the fitness of the obtained solution is better than the fitness of the initial solution, it is remembered. Otherwise, the original solution is stored in the population. Thus, solutions that are promising from interatomic distances will have a chance to be transformed into good solutions and from the correct arrangement of atoms in the crystal.

5. Computational experiments and results

The work of the island model of the evolutionary algorithm was tested on some samples with various features (depth and character of local and global minima). Two samples of medium complexity and one sample of high complexity were considered:

1) K_2PbO_2 – 10 atoms (30 degrees of freedom) [17]. This is a sample with a low symmetry group *P-1*, which means that local minima are quite narrow and relatively small.

2) $Ca_2Al_3O_6F$ – 9 atoms (25 degrees of freedom, since one of the atoms was symmetrically fixed on the edge c) [18]. This is a sample with a higher symmetry group *R-3:H*, which means more local minima.

3) $Er_{10}W_2O_{21}$ – 28 atoms (54 degrees of freedom) [19]. This is a sample with a symmetry group *Pbcn*. It has an average number of symmetry elements, but the solution of this structure is complicated by the number of atoms.

The first and third structures were taken from the ICSD database [20]; diffractograms calculated from these structures with the following parameters: angle range 5-80º, step 0.015º, *CuKα1* radiation, were taken as experimental diffractograms. The second sample was analyzed by an experimental diffraction pattern obtained from one of the authors of this compound.

Each of the samples was launched several dozen times with the same settings by different versions of the MPGA. The basic settings were the same; they are listed in table 1. Genetic operators were the same on all the islands. Tournament selection with a tournament size of 3 was used as a selection operator. The recombination was a single point. The mutation coefficients in table 1 indicate the probability of changing each of the parameters to a random value (within the limits of variation). The initial population on each of the islands was created from solutions filled with random values of parameters. The atomic coordinates were represented by real numbers without coding to binary strings.

A local search in all modes was carried out as follows: every five generations, the current best solution in the population and several random solutions were refined according to the LS of the Rietveld method. In mode # 4 (in table 2), a modified local search operator was additionally applied: every 50 generations, the best current solution of the population was refined according to the scheme described in section 4.

Differing settings and calculation results are shown in table 2. The calculation time of samples in different versions of the algorithm varied by no more than 5% and the average time required for each sample is also indicated in table 2.

Table 1. Settings of the MPGA used for testing.

Setting Name	Value
Number of parallel evolution processes	3
Population size	50
Frequency of use of the LS	5
Number of LS cycles	3
Number of individuals for LS	10 (1 the best + 9 random)
Frequency of use of the modified LS (for mode # 4 in Table 2)	50
Number of individuals for the modified LS (for mode # 4 in Table 2)	1 (the best)
Migration interval	10
Number of individuals to send	2
Generations of local elitism	3
Tournament size	3
Mutation coefficient, % (different for each process)	0.75, 1.0, 1.25

As a comparison, the solution of the same samples was carried out in the popular program FOX [7]. It solves the same problem using the simulated annealing method. The number of calculations of the objective function was increased by three times in order to compensate for the more frequent launch of a local search in the MPGA. A local search was carried out for the best model every 150 000 cycles. The results of these calculations are shown in table 3. It can be seen that the proposed changes in the MPGA make it possible to find solutions more efficiently using the same number of objective function calculations.

Table 2. Convergence comparison of different versions of the MPGA.

Sample	Degrees of freedom	Launch settings	Number of objective function calculations in one run	Average time for one launch, min.	Ratio of successful launches and the average value of objective function of solutions found			
					Mode #1 Migration through the managing process	Mode #2 Direct exchange between populations	Mode #3 Mixed exchange (in the ratio of 1 to 1)	Mode #4 Mixed exchange + new local search operator
K_2PbO_2	30	300 generations, 50 launches	About 50 000	5	55% 39.96	60% 27.73	58% 26.19	72% 19.34
$Ca_2Al_3O_6F$	25	1000 generations, 50 launches	About 160 000	15	22% 18.22	26% 15.48	24% 15.15	40% 13.49
$Er_{10}W_2O_{21}$	54	3000 generations, 10 launches	About 480 000	95	0% 4.96	0% 3.81	0% 4.17	10% 4.69

Table 3. Convergence of the FOX program with the same test samples.

Sample	Degrees of freedom	Launches	Number of objective function calculations in one run	Average time for one launch, min.	Ratio of successful launches and the average value of an objective function of solutions found
K_2PbO_2	30	50	150 100	1	62% 14.33
$Ca_2Al_3O_6F$	25	50	500 000	3	34% 12.38
$Er_{10}W_2O_{21}$	54	10	2 000 000	24	0% 4.32

6. Conclusion

According to the results of the experiments, the following conclusions can be made:

1. The new model with the direct exchange between populations is better than the old model with the dispatch through the managing process in terms of reliability (that is, in terms of the probability of finding the right solution for the same problem).
2. The option of combining these models has also been tested. Its convergence is slightly lower, but most of the solutions it finds are closer to the correct one than the solutions of the version with the direct exchange.
3. The use of an algorithm for randomly displacing the entire structure along the directions of a crystal cell allowed to improve the reliability of crystal structures determining significantly.
4. The use of a new island model allowed us to achieve results comparable to the results of the FOX program with the same computing resources. The use of a modified local search operator allowed to exceed these results.
5. In the future, we plan to conduct a more thorough study of the options for optimal (in terms of computational resources expended) simultaneously displacement of all atoms in the cell.
6. Since the settings can greatly influence the work of the GA, it is planned to collect data for suitable variation ranges of the GA control parameters for the problem of crystal structures solving in order to realize self-tuning of these parameters in the future.

Acknowledgments
The reported study was funded by Russian Foundation for Basic Research, Government of Krasnoyarsk Territory, Krasnoyarsk Regional Fund of Science to the research project № 18-43-243009 «Coevolutionary modelling of an atomic-crystal structure of new substances by diffraction data on basis of parallel genetic algorithms and supercomputer computations».

References

[1] Kariuki B M, Zin D M S, Tremayne M and Harris K D M 1996 Crystal structure solution from powder X-ray diffraction data: The development of Monte Carlo methods to solve the crystal structure of the gamma-phase of 3-chloro-trans-cinnamic acid *Chem. Mater.* **8** 565
[2] Andreev Y G, Lightfoot P and Bruce P G 1997 A new Monte Carlo approach to structure solution from powder data *J. Appl. Crystallogr.* **30** 294-305
[3] Kariuki B M, Serrano-González H, Johnston R L and Harris K D M 1997 The application of genetic algorithm for solving crystal structures from powder diffraction data *Chem. Phys. Lett.* **280** 189-95
[4] Shankland K, David W I F and Csoka T 1997 Crystal structure determination from powder diffraction data by the application of a genetic algorithm *Z. Kristallogr.* **212** 550-2
[5] David W I F, Shankland K and Shankland N 1998 Routine determination of molecular crystal structures from powder diffraction data *Chem. Commun.* **8** 931-2
[6] Andreev Y G and Bruce P G 1988 Solving crystal structures of molecular solids without single crystals: a simulated annealing approach *J. Chem. Soc. Dalton Trans.* **24** 4071-80
[7] Favre-Nicolin V and Cerny R 2002 FOX, 'free objects for crystallography': a modular approach to ab initio structure determination from powder diffraction *J. Appl. Cryst.* **35** 734-43
[8] David W I F, Shankland K, Van de Streek J, Pidcock E, Motherwell W D S and Cole J C 2006 DASH: a program for crystal structure determination from powder diffraction data *J. Appl. Crystallogr.* **39** 910-5
[9] Whitfield P S, Davidson I J, Mitchell L D, Wilson S A and Mills S J 2010 Problem solving with the TOPAS macro language: corrections and constraints in simulated Annealing and Rietveld refinement *Materials Science Forum* **651** 11-25
[10] Harris K D M, Johnston R L and Kariuki B M 1998 The genetic algorithm: Foundations and applications in structure solution from powder diffraction data *Acta Crystallogr., Sect. A: Found. Crystallogr.* **54** 632-45
[11] Feng Z J and Dong C 2007 GEST: A program for structure determination from powder diffraction data using a genetic algorithm *J. Appl. Crystallogr.* **40** 583-8
[12] Lutterotti L and Bortolotti M 2003 Object oriented programming and fast computation techniques in Maud, a program for powder diffraction analysis written in java *IUCr: Compcomm Newsletter* **1** 43-50
[13] Shankland K, Spillman M J and Kabova E A 2013 The principles underlying the use of powder diffraction data in solving pharmaceutical crystal structures *Acta Crystallogr.* C **69** 1251-59
[14] Yakimov Y I, Kirik S D, Semenkin E S, Solovyov L A and Yakimov I S 2013 The evolutionary method of modeling a crystal structure by powder diffraction data *Journal of Siberian Federal University. Chemistry* **2** 180-91
[15] David W I F and Shankland K 2008 Structure determination from powder diffraction data *Acta Cryst.* A **64** 52-64
[16] Zaloga A N, Yakimov I S and Dubinin P S. 2017 Determination of the $[Pt(NH_3)_5Cl]Br_3$ crystal structure from X-ray powder diffraction data using multi-population genetic algorithm *Powder Diffraction* **32** 110-7
[17] Martens K-P and Hoppe R. 1978 Neue Oxoplumbate(II) A_2PbO_2 (A = K, Rb, Cs) mit zweikernigen Gruppen $[OPbO_2PbO]$ *Z. Anorg. Allg. Chem.* **440** 81-104
[18] Xia Z, Molokeev M S, Oreshonkov A S, Atuchin V V, Liu R-S and Dong C 2014 Crystal and local structure refinement in $Ca_2Al_3O_6F$ explored by X-ray diffraction and Raman

spectroscopy *Phys. Chem. Chem. Phys.* **16** 5952-7
[19] Bevan D J M, Drennan J and Rossell H J 1982 Structure determination of the fluorite-related superstructure phases $Er_{10}W_2O_{21}$ and $Y_{10}W_2O_{21}$ *Acta Crystallogr.* B **38** 2991-7
[20] Inorganic crystal structure database Available online: http://www2.fiz-karlsruhe.de/icsd_publications.html