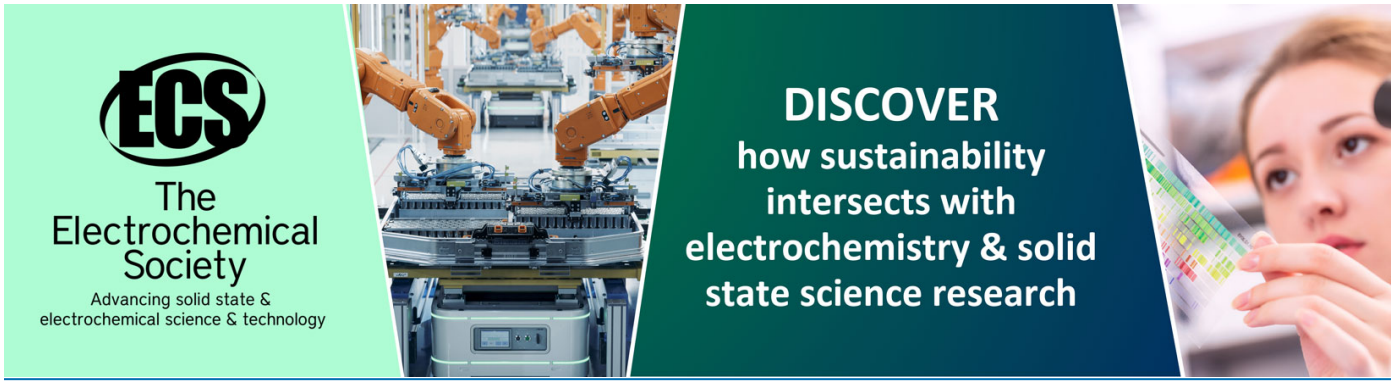
Multi-population evolutionary algorithm SPEA2 for crystal structure determination from X-ray powder diffraction data



IOP Conf. Series: Materials Science and Engineering **734** (2020) 012102 doi:10.1088/1757-899X/734/1/012102

Multi-population evolutionary algorithm SPEA2 for crystal structure determination from X-ray powder diffraction data

A N Zaloga[1], S V Burakov[2], I S Yakimov[1], K A Gusev[2] and P S Dubinin[1]

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

E-mail: zaloga@yandex.ru

Abstract. The problem of determining inorganic crystal structures with a large number of atoms from powder diffraction data is a relevant research task. In order to solve this problem, we made an algorithm of multi-objective optimization based on the SPEA2 approach implemented in a multi-threaded (parallel) version. This algorithm is a development of previously presented variants of single-criterion multi-population GA (MPGA). The article describes the algorithm and the results of applying to test crystal structures.

1. Introduction

In the development of materials with desired properties, knowledge of the structure and properties of substances at the atomic level is required. Many modern materials are created in the form of polycrystals and nanocrystals, and powder diffraction data are used to determine their atomic crystal structure. To process such data, there are methods that work in the direct, inverse, or dual space, and which optimize positions of atoms in the unit cell of the crystal.

Direct space methods for solving crystal structures from powder diffraction data are used in cases where it is not possible to obtain good quality diffraction patterns since the inverse 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. They include the Monte Carlo method [1, 2], genetic algorithms [3, 4] and the method of simulated annealing [5, 6]. Currently, the FOX [7], DASH [8], TOPAS [9] computer programs that use the simulated annealing method are widely spread. Genetic algorithms are implemented in the well-known programs EAGER [10], GEST [11], MAUD [12].

Simulated annealing methods are the most common among them and the easiest to use, but they have no opportunity to apply parallel optimization with the exchange of the best solutions. The essence of GA is an imitation of natural biological selection operations: pairwise crossing, mutations, and selection of the best test structural models to obtain new generations of evolution. This allows you to organize optimization by creating several parallel evolving populations that periodically exchange the best solutions among themselves. Such a property of GA allows achieving a nonlinear increase in efficiency on modern multicore computers.

A common problem of all these methods is the deterioration of convergence with the increase of the task complexity, associated with a nonlinear increase in the probability of stagnation at numerous

local minima of the *R*-factor hypersurface. Therefore, in practice, their use is limited by the number of degrees of freedom of the crystal structures (as a rule, not more than 30–50 DoFs).

In all common programs that use direct space for crystal structures determining, optimization is performed according to one criterion: by the sum of the profile fitting criterion (Chi^2 or R_{wp}-factor) and the penalty for too small interatomic distances – each with its own weight coefficient. In previous works [13, 14], it was shown that the evolutionary approach for determining crystal structures is promising in terms of the possibilities of its modernization. In the process of developing this algorithm, its general concept included procedures that are specific to this problem. This led to an improvement in the quantitative and qualitative indicators of the algorithm's work (often simultaneously). However, it became obvious that the objective function based on the criteria convolution (the criterion of diffraction pattern difference and the criterion of physicality of the crystal structure) is one of the main limitations for of this method: when the dimension is increased, the criterion based on the convolution does not allow organize evolutionary selection efficiently. At the same time, the adjustment of the weighting coefficients of the criteria has a significant impact on the results of the algorithm; these coefficients can change significantly when moving from solving one problem to another one. Therefore, we propose to use the multi-objective optimization (MOO) algorithm to solve the problem. In particular, the evolutionary algorithm SPEA2. The article describes a multi-threaded version of this algorithm, which allows us to use the resources of modern computers efficiently.

2. Transition to the multi-objective evolutionary algorithm

A lot of multi-objective optimization algorithms are known; some of the most effective are SPEA2 [15], NSGA2 [16], MOEA/D [17]. The goal of the present work was to compare the single-criterion (based on convolution) approach and the multi-objective approach in optimizing the multi-objective problem of determining the crystal structure of powder substances by efficiency and prospects for further development. In connection with this, the main requirement when choosing the MOO algorithm was the highest similarity by genetic operators to the single-objective genetic algorithm we used to. Therefore, the SPEA2 algorithm has been chosen from the algorithms above. In the framework of this algorithm, similar genetic operators are used, the selection scheme is the closest to the single-criteria algorithm proposed earlier in [13].

2.1. The SPEA2 algorithm

There are two criteria we use for determining the crystal structure of a substance: the degree of correspondence between the calculated and experimental diffraction patterns, and the penalty parameter characterizing the deviation of interatomic distances in the model from physically acceptable ones. Therefore, we consider two-criteria optimization in the direction of decreasing the value of the criteria.

The main concept of MOO algorithms is the “Pareto set”. This term refers to a group of such individuals in a population that are better than others by at least one of the criteria. Such individuals are called "nondominated." The name “dominated” individual means that there is at least one individual that is better than the given one by all criteria. An explanation of these concepts is given in figure 1.

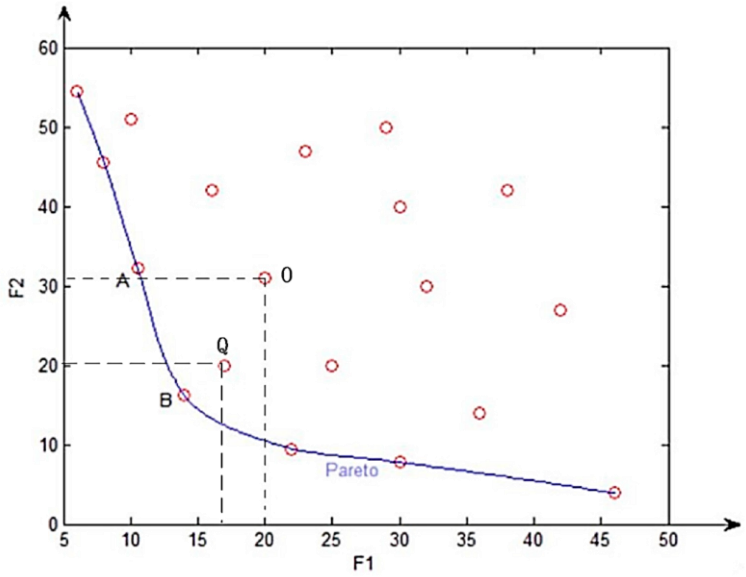


Figure 1. The Pareto front. The axes indicate the values of the two criteria by which minimization occurs. Solutions A and B are nondominated; solutions O and Q are dominated. The blue line is drawn over non-dominated individuals. It is the Pareto front.

The MOO algorithm called The Strength Pareto Evolutionary Algorithm (SPEA) was proposed for MOO problems by Zitzler and Thiele in 1999 [18]. In 2001, the same authors published an article on the improvements of the algorithm and called it SPEA2 [15]. Its main advantage compared with other MOO algorithms is that it is non-elitist, that is, it seeks to pass on to the next generation not only 1-2 best individuals but the entire Pareto front. For this purpose, an archive set is used. All individuals lying on the Pareto front are recorded into it every generation. Moreover, if the archive size is smaller than the number of individuals lying on the front, then solutions having close neighbors are removed from the archive (as shown in figure 2). When the next generation is generating, individuals for crossing are selected from a mixture of the main population and the archive.

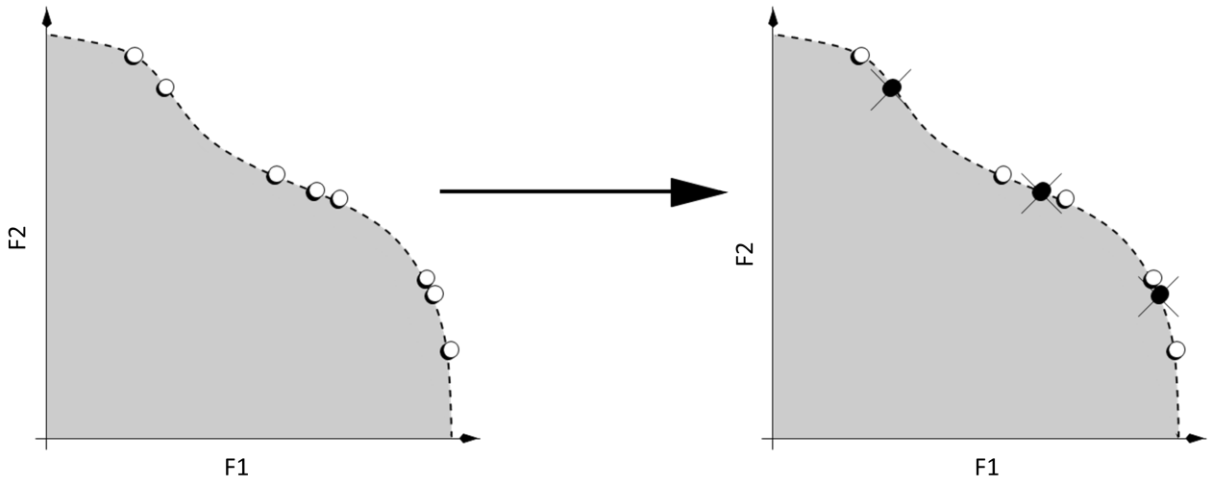


Figure 2. Scheme of solutions discarding in case the Pareto front size is larger than the specified size of the archive.

Here are the main stages of the SPEA2 algorithm in brief:

1. Initialization of the population and creation of an empty archive set (on each work process).
2. Calculation of the suitability of individuals in a population. Evaluation of the suitability of individuals of the archive.
3. Selection of non-dominated individuals into a new archive. If there are more such individuals than the specified size of the archive set, this set is truncated based on the criteria values and the distance between individuals in the criteria space. If such individuals are smaller than the size of the archive, this archive is filled by adding dominated individuals based on their suitability.
4. If the stop criterion is met, the archive is displayed as a result of the algorithm's launch.
5. Tournament-based selection.
6. Use of recombination and mutation operators.

2.2. Features of multi-thread implementation of the SPEA2 algorithm

We have made a variant of the parallel multi-population SPEA2, in which there are one control process and several work processes. The exchange between them is implemented according to the "star" scheme. Work processes are evolving according to the usual SPEA2 algorithm. The “main archive” is stored in the control process. Its size is equal to the number of work processes multiplied by their archive size. The exchange of individuals between processes occurs as follows.

1) Each generation, the archival sets of work processes are sent to the control process. The archival set of the control process is filled out from a mixture of individuals of its archival set from the previous generation and newly-arrived individuals.

2) Every 100 generations, the control process sends its entire archive set to all work processes.

Such a scheme makes it possible to simultaneously exchange the best individuals between populations and provide independence for the evolution of work processes between exchange acts. figure 3 shows a flowchart that illustrates how this algorithm works.

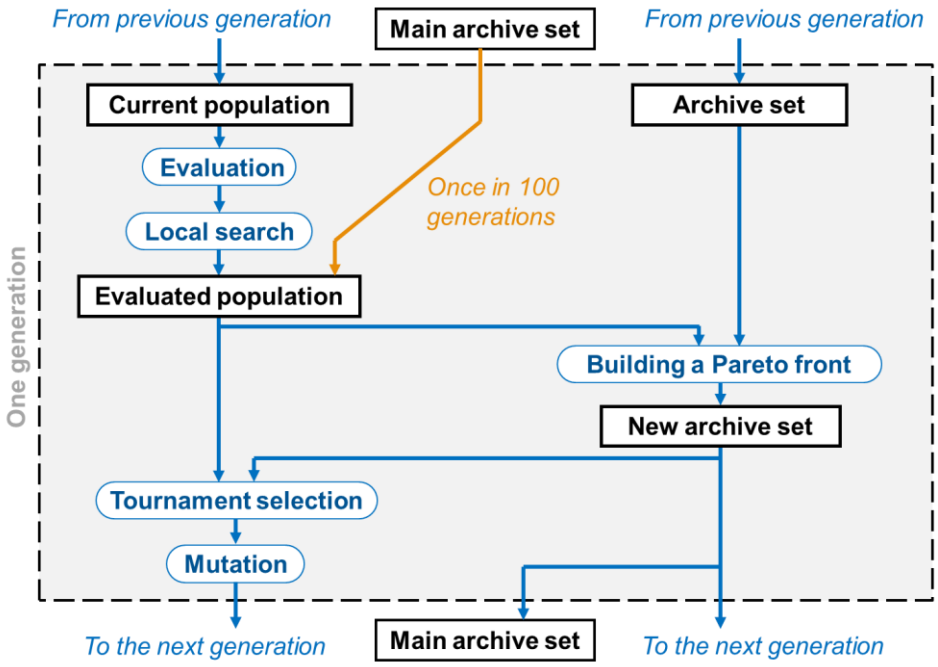


Figure 3. The flowchart of the parallel multi-population algorithm SPEA.

In previous single-objective versions of MPGA, we also used the control process to accumulate better solutions. The difference is that in previous versions only one the best solution by a single criterion was sent to the control process. Also, the control process accumulated all the solutions received. And the control process sent to the work processes several random solutions the suitability of which was lower than the average suitability of the accumulated solutions. In the present version of the algorithm, we use an exchange scheme that is more suitable for multi-objective optimization.

3. Description of the objective function

The proposed algorithm is used to restore an atomic crystal structure from a powder diffraction pattern. Such patterns are obtained by irradiating the powder sample with X-rays in special devices called powder diffractometers. A diffraction pattern represents the dependence of the intensity of x-ray reflected from the sample on the scattering angle. It is a one-dimensional convolution of a three-dimensional distribution of electron density in the crystal.

There is a mathematical model of the powder diffraction, and some of the parameters of this model is coordinates of atoms in the crystal. Since the arrangement of atoms in the crystal is periodically repeated, it is sufficient to describe only unique atomic positions to describe the entire crystal. The minimum volume of the crystal that contains unique atomic positions is called the unit cell. The intensity of diffraction reflexes is determined by the following formula:

$$Yc = f(F_{hkl}(\overline{X})),$$

$$F_{hkl} = \sum_k f_k e^{-2\pi i(hx_k + ky_k + lz_k)}$$

where Yc is the calculated profile intensity; F_{hkl} is structural amplitude, which depends on the atomic coordinates X; f_k is scattering energy of the k-th atom, depending on its electron density; x_k, y_k, z_k are relative coordinates of the k-th atom in the unit cell in the range from 0 to 1; h, k, l are indexes of given crystallographic plane (integer numbers).

In the problem under consideration, the coordinates of the atoms are variable parameters of the model. The difference between the simulated diffraction pattern and the experimental one is used as the objective function. This criterion is called the weighted profile R-factor (R_{wp}-factor) of the Rietveld method and is calculated by the formula:

$$R_{wp} = \sqrt{\frac{\sum_j w_j \left(Yo_j - Yc_j\right)^2}{\sum_j w_j \left(Yo_j\right)^2}}$$

where j is the number of the profile point; Yo_j and Yc_j are the observed and calculated profile intensities; w_j is the weight of j-th point.

A penalty for violation of interatomic distances in the solution is an additional objective function. The user sets the minimum allowable distance for each pair of atomic types that are contained in the substance (for example, O-O, O-Fe, Fe-S). When calculating the fitness value of a solution, the distances between all atoms are compared. Then the penalty value is calculated as the sum of these violations.

4. Verification of the algorithm on test samples

As test samples, we used the same three samples on which previous versions of the MPGA were tested. These were two samples of medium complexity and one sample of high complexity:

- K_2PbO_2 – 10 atoms (30 degrees of freedom) [19]. This is a sample with a low symmetry group *P*-1, which means that local minima are quite narrow and relatively small.
- $Ca_2Al_3O_6F$ – 9 atoms (25 degrees of freedom, since one of the atoms was symmetrically fixed on the edge *c*) [20]. This is a sample with a higher symmetry group *R*-3:*H*, which means more local minima.
- $Er_{10}W_2O_{21}$ – 28 atoms (54 degrees of freedom) [21]. 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 [22]; diffractograms calculated from these structures with the following parameters: angle range 5-80º, step 0.015º, Cu$K\alpha$1 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 GA. The basic settings were the same; they are listed in table 1. Genetic operators were the same for all the work processes. 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 in each of the processes 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 (LS) was carried out as follows: every few generations, the best solution in the population and several random solutions are selected, and their parameters are refined using the Rietveld method.

Table 1. The common settings of the SPEA2 algorithm.

Parameter	Value
Number of work processes	3
Population size	50
Archive size in every work process	7 (15% of population size)
Archive size in the control process	21 (number of work processes * archive size in the work processes)
LS launch interval	10
Number of individuals for LS	5 (the best one + 4 random ones)
The number of LS cycles	3
The interval of sending individuals from the control process	100
The number of individuals to be sent	Equal to the archive size of the control process
Tournament size	3
Mutation coefficients (different for work processes)	0.5, 1.0, 1.5

Table 2 shows the results of the launches. Each sample was run in a series of several starts with the same settings. Table 2 also contains a comparison with the results obtained by single-criterion versions of MPGA for the same samples.

Table 2. Results of the algorithm launches and comparison with the results obtained by single-criterion versions of MPGA.

Sample	DoF	Launch method	A fraction of successful launches and the mean fitness value of the solutions found			
			MPGA with star-type exchange	MPGA with island model exchange	MPGA with island model exchange + periodic structure shift	SPEA2
K_2PbO_2	30	300 gen., 50 launches	55% 39.96	60% 27.73	72% 19.34	84% 12.91
$Ca_2Al_3O_6F$	25	1000 gen., 50 launches	22% 18.22	26% 15.48	40% 13.49	54% 11.58
$Er_{10}W_2O_{21}$	54	3000 gen., 10 launches	0% 4.96	0% 3.81	10% 4.69	50% 1.31

5. Discussion

In order to obtain the results shown in table 2, many tests were conducted with different values of the SPEA2 control parameters. On average, all parameters in table 1 were tested in the range from -50% to +100% of the values in the table. And in table 1 are the values that allowed to achieve optimal results in terms of efficiency/time spent. However, studies on the effect of different types of selection or crossover operators have not been conducted. The given number of populations and the number of individuals in each population was chosen in order to compare efficiency with previous versions of the algorithm. In general, these two parameters depend on the performance of the computer on which the algorithm is launched.

As can be seen from table 2, the effectiveness of the application of the multi-population SPEA2 algorithm for the first two test structures did not show a considerable increase. At the same time, for the third structure, the efficiency turned out to be significantly more than for single-objective optimization algorithms. It is also noteworthy that according to the convergence charts, for the third structure, the correct solution was found up to 300th generation in 7 successful launches, and up to 600th generation in 3 successful launches from 10. The convergence of the solutions of the first two structures did not have such a strongly-pronounced dependence of convergence on the stage of evolutionary search.

It can be assumed that for the first two simpler structures, the SPEA2 algorithm does not add new ways to get out of local minima in comparison with single-objective algorithms. For the third, more complex structure, the number of local minima is many times (nonlinear) larger and these algorithms begin to cope poorly with the getting out of them. The SPEA2 algorithm continues to provide this capability by maintaining population diversity by the archive set, which ultimately leads to convergence to a global minimum.

6. Conclusion

A multi-population algorithm of multi-objective optimization SPEA2 is proposed, which is applied to the problem of solving inorganic crystal structures from powder diffraction data. Our experiments show that this algorithm is more efficient than the previously used single-objective optimization algorithms. In the future, this will allow us to solve the problem of determining more complex crystal structures.

Acknowledgment

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* ***2*80** 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 M J, 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] Zaloga A, Yakimov I and Dubinin P 2018 Multi-population genetic algorithm for crystal structures solution from X-ray powder diffraction data *International Journal on Information Technologies and Security* **10(2)** 119-28

[14] Zaloga A N, Burakov S V, Yakimov I S, Gusev K A and Dubinin P S 2019 Island model with genetic algorithm for solution of crystal structure from X-ray powder diffraction data *IOP Conf. Ser.: Mater. Sci. Eng* **537** 1-9

[15] Zitzler E, Laumanns M and Thiele L 2001 SPEA2: Improving the strength pareto evolutionary algorithm, Technical report *Swiss Federal Institute of Technology* 1-21

[16] Deb K, Pratap A and Agarwal S 2002 A fast and elitist multi-objective genetic algorithm: NSGA-II *IEEE Transactions on Evolutionary Computation* **6(2)** 182-97

[17] Zhang Q and Li H 2007 MOEA/D: a multiobjective evolutionary algorithm based on decomposition *IEEE Transactions on Evolutionary Computation* **11(6)** 712-31

[18] Zitzler E and Thiele L 1999 Multiobjective evolutionary algorithms: A comparative case study and the strength pareto approach *IEEE Transactions on Evolutionary Computation* **3(4)** 257-71

[19] 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

[20] 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
[21] 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* **B38** 2991-7
[22] Inorganic Crystal Structure Database. Available online: https://www.fiz-karlsruhe.de/en