Journal of
**Applied Crystallography**
ISSN 0021-8898

Received 4 March 2009
Accepted 25 August 2009

# *SMEPOC* – a computer program for the automatic generation of trial structural models for inorganic compounds with symmetry restriction

**Xiaodi Deng and Cheng Dong***

National Laboratory for Superconductivity, Institute of Physics, and Beijing National Laboratory for Condensed Matter Physics, Chinese Academy of Sciences, Beijing, People's Republic of China. Correspondence e-mail: chengdon@aphy.iphy.ac.cn

A new structure modelling algorithm that can automatically generate trial structure models based on a prior knowledge of the unit-cell content and space-group information is proposed. It can enumerate all possible equivalent position combination (EPC) models and eliminate unreasonable ones with symmetry restriction. Unlike other methods, it does not require the internal molecular connectivity or coordination polyhedron information and is mostly suitable for modelling inorganic crystals with independent atoms. Therefore, these EPC models can be used as input to global optimization procedures for inorganic crystal structure determination using powder diffraction data by the direct-space method. The efficiency of the direct-space method can be greatly improved using this EPC method because the global optimization problem in a $3N$-parameter space can be divided into a set of simpler ones. A program, *SMEPOC*, that can generate all possible EPC models for further global optimization procedures has been developed and is now available from the authors upon request.



## 1. Introduction

The direct-space method has been widely and successfully used to solve crystal structures from powder diffraction data in recent years (Černý, 2008; David & Shankland, 2008). It adopts the strategy of first generating trial models and then comparing calculated and observed diffraction patterns. In general terms, the direct-space method is a complex global optimization problem of finding the best agreement between calculated and observed diffraction patterns whose parameters have been defined at the trial model generating stage. In recent decades, the focus has mainly been on applying different kinds of global optimization algorithms such as Monte Carlo search (Harris *et al.*, 1994), simulated annealing (Newsam *et al.*, 1992; Andreev *et al.*, 1997), parallel tempering (Favre-Nicolin & Černý, 2002; Earl & Deem, 2005), genetic algorithm (Shankland *et al.*, 1997; Kariuki *et al.*, 1997; Harris *et al.*, 1998; Feng & Dong, 2007) and particle swarm (Csoka & David, 1999). Generation of trial models is often seen as a straightforward process for molecular crystals, and the connectivity of isolated molecular blocks makes it easy to parameterize a global optimization by defining internal coordinates like bond distances, bond angles and torsion angles (Andreev *et al.*, 1997). However, for inorganic materials without *a priori* knowledge of connectivity, little effort has been made to model trial structures. Although great progress has been made for framework (zeolite) crystals (Deem & Newsam, 1989; Facioni & Deem, 1999) and crystals with isolated molecules or polyhedra (Favre-Nicolin & Černý, 2002), the lack of a general-purpose modelling approach for the direct-space method means that to solve most inorganic materials one has to resort to reciprocal-space methods (Cascarano *et al.*, 1984, 1991); these require high quality diffraction data and accurately extracted diffraction intensities that are usually hard to obtain using the technique of powder diffraction.

The most straightforward description of an inorganic crystal structure without previously known isolated building blocks or connectivity consists of specifying $N$ independent atoms located with fractional coordinates $(x, y, z)$ in the asymmetric unit of a unit cell (David *et al.*, 2002). It is often computationally intractable when $N$ grows big as the complexity of the global optimization problem is exponentially related to the parameter number $3N$. Nevertheless, in this paper we present a modelling algorithm that can divide a $3N$-parameter problem into a set of simpler ones starting from unit-cell content and space-group information. Unlike the already known trial model generation algorithms, connectivity information is not needed for this algorithm. In general terms, it is an automation and generalization of the equivalent position combination (EPC) method used in some special cases by pioneer crystallographers when structure solution was still done manually (Brakken & Scholten, 1934; Lu & Liang, 1965*a*,*b*; Reddy *et al.*, 1965). The working principle of the EPC method is as follows: when the content of the unit cell and the space-group information are determined, element $A$ with $N_A$ atoms in a unit cell can only be placed at an equivalent position whose multiplicity equals $N_A$ or at several equivalent positions, the summation of whose multiplicities equals $N_A$. Different elements may have different equivalent position arrangements. When every element in the unit cell has been settled down, a combination of their equivalent position arrangements describes a structure model of the crystal. Mathematically, a complete list of all possible equivalent position combinations can be enumerated. In the early days, EPCs were usually listed manually for some crystals with low complexity and high symmetry. Then, the list would be checked and the improper EPC models would

be eliminated according to physical or chemical information such as the distribution of diffraction intensities, additional extinctions, cell parameters and bonding characteristics. This elimination process requires a wealth of knowledge about the structural chemistry and crystallography, and one or several trial structures could be picked out among the EPC models in favourable cases. Subsequently, atomic parameters of the chosen models would be solved based on the powder diffraction information (Brakken & Scholten, 1934; Lu & Liang, 1965*b*; Reddy *et al.*, 1965).

Generating all possible EPC models manually is not a very easy task and is even impossible in some circumstances. The EPC-generation algorithm reported in this paper can do this job automatically with the following features:

(1) Generating a full list of all possible EPC models. Assuming that all site-occupation factors (SOF) of atoms in the unit cell are 1.0 (fractional occupation cases will be discussed in §3.2), the number of EPC models is finite and the mathematically reasonable EPC models are enumerable. The algorithm can generate all possible EPC models and avoid duplicated ones.

(2) Extendable interfaces to eliminate unreasonable EPC models. Mathematically, reasonability does not guarantee that EPC models are crystallographically or chemically acceptable. The most intuitive consideration is site-occupation conflict: when a Wyckoff position with fixed fractional coordinates has been occupied, it cannot be occupied a second time. Besides this, other physical and chemical restraints can also be employed to shorten the EPC list.

Global optimizations can be easily applied on these EPC models in SDPD (structure determination using powder diffraction) using the direct-space method. The parameters of an EPC model are the variable fractional coordinates of the corresponding Wyckoff positions. About 70% of inorganic crystals belong to crystal systems with symmetry higher than monoclinic and some atoms often occupy special equivalent positions. Therefore, the parameter number in an EPC model can be greatly reduced so that it is much less than $3N$ in most cases. The higher the symmetry, the fewer the parameters, and EPCs may even have no free parameter in some circumstances. Using EPC models, the global optimization problem in a $3N$-parameter space can be divided into a set of simpler ones. This 'divide and conquer' process can significantly increase the efficiency of direct-space global optimizations in SDPD, and may lead to a successful structure solution that is computationally intractable in $3N$-parameter space.

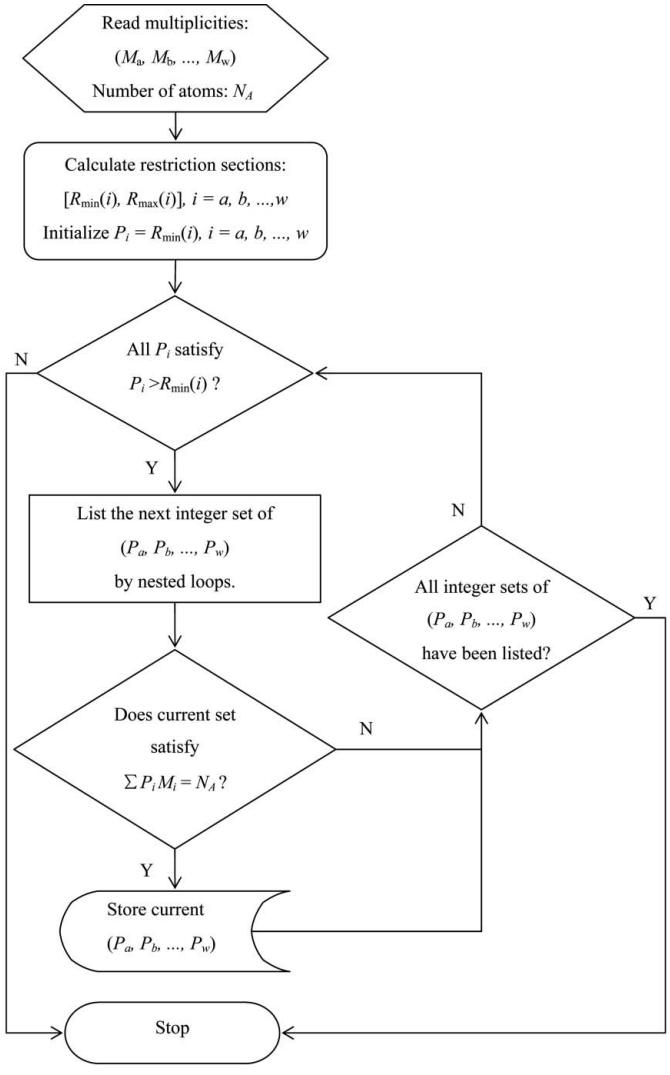


**Figure 1**
Algorithm flowchart of equivalent position combination for a single element.

## 2. Equivalent position combination enumeration algorithm

A full list of EPC models of a defined unit cell and symmetry can be enumerated by the following two steps: (1) enumerate EPCs for every single element in the unit cell; (2) with the results of the first step, enumerate EPC models for atoms in the whole unit cell.

Basically, the enumeration process is a pure mathematical process which depends only on the number of elements and atoms in the unit cell and the symmetry of the space group. However, physical and chemical restraints can also be used to eliminate irrational EPC models in this mathematical process.

### 2.1. Equivalent position combination for single element

Given Wyckoff positions $(a, b, c, \ldots, w)$ with multiplicities $(M_a, M_b, M_c, \ldots, M_w)$, respectively, for an element $A$ with $N_A$ atoms in the unit cell, all possible EPCs of element $A$ can be represented by the solution of the following equation:

$$N_A = \sum_{i=a}^{w} P_i M_i, \tag{1}$$

where $P_i$ are the numbers of times the corresponding Wyckoff positions have been occupied, and they are integers not less than zero. There is no mathematical formula for the solution of equation (1). However, a computer algorithm can be designed to enumerate all the $(P_a, P_b, \ldots, P_w)$ sets when we consider the upper limit of $P_i$. Assuming that $\lfloor x \rfloor$ represents the largest integer not greater than $x$, then the upper boundary of $(P_a, P_b, \ldots, P_w)$ can be written as $(\lfloor N_A/M_a \rfloor, \lfloor N_A/M_b \rfloor, \ldots, \lfloor N_A/M_w \rfloor)$. The enumeration process can be done by searching all integer sets of $P_i$ satisfying $(0 \le P_a \le \lfloor N_A/M_a \rfloor,\ 0 \le P_b \le \lfloor N_A/M_b \rfloor, \ldots,\ 0 \le P_x \le \lfloor N_A/M_w \rfloor)$ and equation (1). This is the basis of the algorithm for the single-element combination shown in Fig. 1.

Information used to eliminate unreasonable sets can be embedded in this enumeration process. One of the essential and simplest considerations is the site-occupation conflict (SOC). Wyckoff positions with fixed coordinates (where $x$, $y$, $z$ are all constants) can be occupied only one time while others logically can be occupied unlimited times. When a site position of the first kind has been occupied, it cannot be occupied again or an SOC happens. We define

function $R(w)$ to represent restrictions introduced by these kinds of extra information. Then $R(w)$ for an SOC restriction can be written as

$$R_{\rm soc}(i) = \begin{cases} \leq 1 & \text{for sites } i \text{ with fixed coordinates} \\ \infty & \text{otherwise.} \end{cases} \quad (2)$$

Another simple consideration is that site occupation has been known *a priori*: either by analysing extra extinction of Wyckoff positions (Lu & Liang, 1965*a*) or by auxiliary experiments. The mathematical expression for this kind of restriction for the necessarily occupied Wyckoff position $w_0$ can be written as

$$R_{\rm prior}(i) = \begin{cases} 1 & i = w_0 \quad \text{with fixed coordinates} \\ \geq 1 & i = w_0 \quad \text{with variable coordinates} \\ \infty & i \neq w_0. \end{cases} \quad (3)$$

By defining $[R_{\rm min}(i), R_{\rm max}(i)]$ as the intersection of $[0, \lfloor N_A/M_i \rfloor]$ with all of the introduced restraints $R_{\rm soc}(i)$, $R_{\rm prior}(i), \ldots$, then the limitations of $P_i$ can be written as

$$R_{\rm min}(i) \leq P_i \leq R_{\rm max}(i), \quad i = a, b, \ldots, w. \quad (4)$$

Equation (4) can represent all possible restraints. Besides the simple and basic restraints discussed above, geometric and chemical restraints can also be applied based on the unit-cell dimensions, atomic radii and bond characteristics.

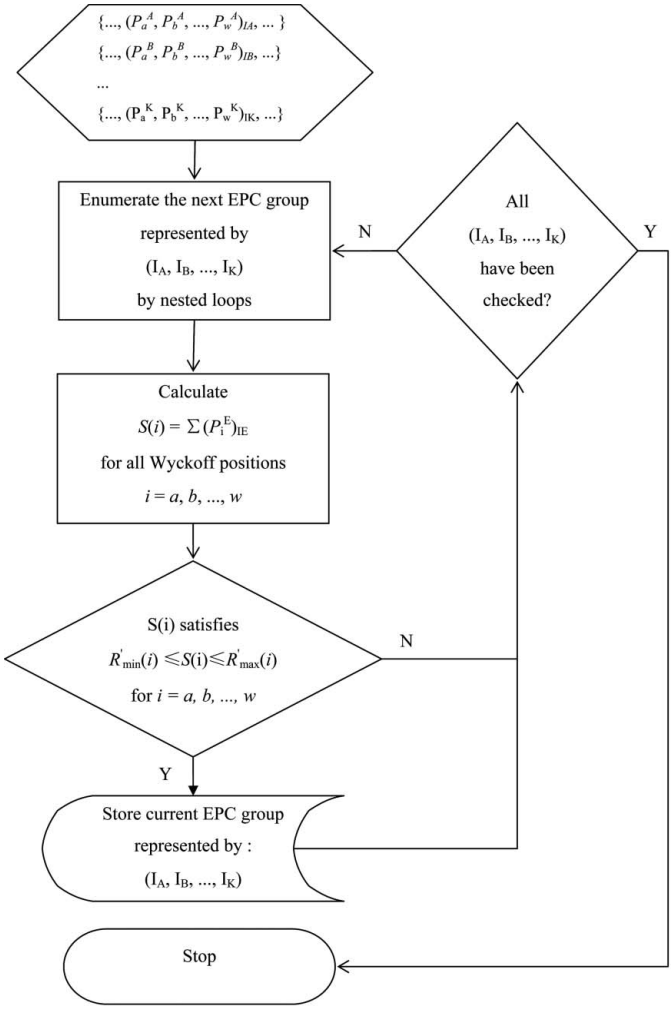


**Figure 2**
Algorithm flowchart of equivalent position combination for all elements in the unit cell.

The integer sets of $(P_a, P_b, \ldots, P_w)$ satisfying equations (1) and (4) are the reasonable EPCs for element $A$ with $N_A$ atoms in the unit cell. The enumeration algorithm flowchart is shown in Fig. 1.

### 2.2. Equivalent position combination for all elements in the unit cell

Assuming that the unit-cell content is $A_{N_A}B_{N_B}\ldots K_{N_K}$, generally, EPCs for a single element $E$ are $C_E$ $n_w$-dimensional arrays, where $C_E$ is the number of reasonable EPCs for element $E$, and $n_w$ is the number of Wyckoff positions of the space group. Then, the result of the first step can be written as

$$\begin{aligned} A &= \left[(P_a^A, P_b^A, \ldots P_w^A)_1, (P_a^A, P_b^A, \ldots P_w^A)_2, \ldots, (P_a^A, P_b^A, \ldots P_w^A)_{C_A}\right] \\ B &= \left[(P_a^B, P_b^B, \ldots P_w^B)_1, (P_a^B, P_b^B, \ldots P_w^B)_2, \ldots, (P_a^B, P_b^B, \ldots P_w^B)_{C_B}\right] \\ &\ldots \\ K &= \left[(P_a^K, P_b^K, \ldots P_w^K)_1, (P_a^K, P_b^K, \ldots P_w^K)_2, \ldots, (P_a^K, P_b^K, \ldots P_w^K)_{C_K}\right]. \end{aligned} \quad (5)$$

When EPCs for every element $A, B, \ldots, K$ have been obtained, we can take the second step of combining among elements to obtain the final EPC models for this crystal.

Mathematically, there are $C_A C_B \cdots C_K$ possible combinations according to equation (5). Just as in the first step, we must check the crystallographic rules to eliminate the unreasonable combinations. The restraints defined in the first step are based on the intrinsic properties of Wyckoff positions and can also be applied here. The algorithm is straightforward: we check all combinations and keep the reasonable ones. Given an EPC model represented by $(I_A, I_B, \ldots, I_K)$:

$$\begin{aligned} &A : (P_a^A, P_b^A, \ldots P_w^A)_{I_A} \\ &B : (P_a^B, P_b^B, \ldots P_w^B)_{I_B} \\ &\ldots \\ &K : (P_a^K, P_b^K, \ldots P_w^K)_{I_K}, \end{aligned} \quad (6)$$

where $I_E, E = A, \ldots, K$, are indices of the EPCs of the corresponding element. The Wyckoff position occupation state can be defined by the summation of the corresponding site-occupation numbers $P_i$ of all elements in the unit cell. For Wyckoff site $i$

$$S(i) = \sum_{E=A}^{K} (P_i^E)_{I_E}. \quad (7)$$

Summations for all Wyckoff positions $S(i), i = a, b, \ldots, w$, must be checked by the restraints [equation (4)] to make sure that the EPC model is reasonable. However, there is a minor difference here; $R_{\rm min}(i)$ and $R_{\rm max}(i)$ are no longer an intersection of $[0, \lfloor N_A/M_i \rfloor]$ with the restraints but an intersection of these restraints themselves. We define this new intersection as $R'_{\rm min}(i)$ and $R'_{\rm max}(i)$; then EPC models satisfying $R'_{\rm min}(i) \leq S(i) \leq R'_{\rm max}(i)$ for all $i = a, b, \ldots, w$ will be kept. The flowchart in Fig. 2 illustrates this idea.

The final EPC models represented by sets of $(I_A, I_B, \ldots, I_K)$ can be further converted into the format that can be used in the global optimization stage of SDPD by direct-space strategies.

### 2.3. Example and modelling process of EPC

In this section, we take well known $La_2CuO_4$ (a parent compound of high-$T_c$ superconductors) as an example to describe the equivalent position combination algorithm and we will also discuss why enormous reduction of computational complexity can be achieved using the generated EPC models in SDPD by the direct-space method.

**Table 1**
Wyckoff position list of space group *Fmmm*.

| Wyckoff position | | Site symmetry | Coordinate (partial) |
|---|---|---|---|
| 32 | *p* | 1 | *x*, *y*, *z* |
| 16 | *o* | ..*m* | *x*, *y*, 0 |
| 16 | *n* | .*m*. | *x*, 0, *z* |
| 16 | *m* | *m*.. | 0, *y*, *z* |
| 16 | *l* | 2.. | *x*, $\frac{1}{4}$, $\frac{1}{4}$ |
| 16 | *k* | .2. | $\frac{1}{4}$, *y*, $\frac{1}{4}$ |
| 16 | *j* | ..2 | $\frac{1}{4}$, *y*, $\frac{1}{4}$ |
| 8 | *i* | *mm*2 | 0, 0, *z* |
| 8 | *h* | *m*2*m* | 0, *y*, 0 |
| 8 | *g* | 2*mm* | *x*, 0, 0 |
| 8 | *f* | 222 | $\frac{1}{4}$, $\frac{1}{4}$, $\frac{1}{4}$ |
| 8 | *e* | ..2/*m* | $\frac{1}{4}$, $\frac{1}{4}$, 0 |
| 8 | *d* | .2/*m*. | $\frac{1}{4}$, 0, $\frac{1}{4}$ |
| 8 | *c* | 2/*m*.. | 0, $\frac{1}{4}$, $\frac{1}{4}$ |
| 4 | *b* | *mmm* | 0, 0, $\frac{1}{2}$ |
| 4 | *a* | *mmm* | 0, 0, 0 |

Wyckoff positions *a*, *b*, *c*, *d*, *e* and *f* can be occupied only once, while Wyckoff positions *g*, *h*, *i*, *j*, *k*, *l*, *m*, *n*, *o* and *p* can be occupied unlimited times. According to the definition of $R_{soc}$ in equation (2), we have $R_{soc}(i') = 1$, $w = a, b, \ldots, f$; $R_{soc}(i') = \infty$, $w = g, h, \ldots, p$.

The $La_2CuO_4$ crystal belongs to the orthorhombic crystal system with space group *Fmmm*. The number of formula units contained in an $La_2CuO_4$ unit cell is four, which means that there are eight La atoms, four Cu atoms and 16 O atoms in a unit cell. Table 1 lists all Wyckoff positions of the space group *Fmmm*. With this information we can obtain all reasonable EPCs using algorithms discussed in §§2.1 and 2.2.

Combined with the boundaries $[0, \lfloor N_A/M_w \rfloor]$, the $R_{min}$ and $R_{max}$ values of the corresponding Wyckoff positions for the La, Cu and O elements can be obtained and the values are listed in Table 2. It can be seen in Table 2 that only Wyckoff positions below '*i*' for the La element have $R_{max}$ values greater than zero. By applying the algorithm in Fig. 2, La atoms can be positioned at one of the Wyckoff positions *i*, *h*, *g*, *f*, *e*, *d*, *c*, or at both Wyckoff positions *a* and *b*. Eight EPCs can be obtained for La; for Cu, only Wyckoff positions *a* and *b* have non-zero $R_{max}$ values, which indicates that the Cu atom has only two choices, either Wyckoff position *a* or *b*. Two EPCs can be obtained for Cu; the situation for O is more complex and 37 reasonable EPCs can be obtained. The results for each element are listed in Table 3.

With EPCs for elements, we can get the final EPC models for the crystal. In the example described above, we considered only the SOC restraint $R_{soc}$. The elimination process described in the algorithm in §2.2 is then quite intuitive. Take EPC La 4*a* + 4*b*, Cu 4*a*, O 4*a* + 4*b* + 8*c* (the first EPC of each element in Table 3) as an example: the occupation summation of Wyckoff position *a* is *S*(*a*) = 1 (La) + 1 (Cu) + 1 (O) = 3 which is beyond the upper limit $R_{max}(a) = R_{soc}(a) = 1$; an SOC happens, and similarly an SOC also happens for Wyckoff position *b* whose occupation summation is 2 for this EPC model. This kind of EPC model must be eliminated. Mathematically, there are $8 \times 2 \times 37 = 592$ EPC models that need to be checked, an SOC happens in about 30% of them and only 372 reasonable EPC models are kept.

EPC models can easily be used to parameterize global optimization processes. Two kinds of typical EPC models, fixed-coordinate models and variable-coordinate models, are shown in Figs. 3(*a*) and 3(*b*), respectively. Apparently, global optimization procedures are not needed for the fixed-coordinate models, only for the variable-coordinate models. These models have much fewer parameters than 3*N*. We consider a grid-search algorithm for simplicity. The complexity of the global optimization problem increases exponentially with parameter number, while it increases only linearly with the number of EPC models. As a result, the decrease in parameter number reduces drastically the scale of the global optimization problem. The $La_2CuO_4$ example discussed here is a 21-parameter conventional global optimization problem. There are 372 EPC models for this crystal, and the maximum parameter number for these models is three. If we divide every search direction into ten parts, then the global optimization problem can be reduced from a scale of $10^{21}$ to $372 \times 10^3$, which means a big performance improvement. The number of variable coordinates in these models depends on the symmetry of the crystal: the higher the symmetry the smaller the variable-coordinate number.

Generally speaking, this method is a process of 'divide and conquer' which is a universal technique used to solve difficult problems. As we can see in the $La_2CuO_4$ example, a troublesome (21 parameters) global optimization problem is divided into 372 simpler ones (at most three parameters). It is much easier to solve a set of three-parameter problems than a single 21-parameter problem and the result of the former is often more reliable than that of the latter. More importantly, these subset problems can be solved by parallel programming technology using multi-processor machines. In fact, parallel programming is one of the biggest advantages of this method compared with other modelling methods.

**Table 2**
Search limits of La, Cu and O elements in the $La_2CuO_4$ unit cell.

| | | 8 La | | 4 Cu | | 16 O | |
|---|---|---|---|---|---|---|---|
| Wyckoff position | | $R_{min}$ | $R_{max}$ | $R_{min}$ | $R_{max}$ | $R_{min}$ | $R_{max}$ |
| 32 | *p* | 0 | 0 | 0 | 0 | 0 | 0 |
| 16 | *o* | 0 | 0 | 0 | 0 | 0 | 1 |
| 16 | *n* | 0 | 0 | 0 | 0 | 0 | 1 |
| 16 | *m* | 0 | 0 | 0 | 0 | 0 | 1 |
| 16 | *l* | 0 | 0 | 0 | 0 | 0 | 1 |
| 16 | *k* | 0 | 0 | 0 | 0 | 0 | 1 |
| 16 | *j* | 0 | 0 | 0 | 0 | 0 | 1 |
| 8 | *i* | 0 | 1 | 0 | 0 | 0 | 2 |
| 8 | *h* | 0 | 1 | 0 | 0 | 0 | 2 |
| 8 | *g* | 0 | 1 | 0 | 0 | 0 | 2 |
| 8 | *f* | 0 | 1 | 0 | 0 | 0 | 1 |
| 8 | *e* | 0 | 1 | 0 | 0 | 0 | 1 |
| 8 | *d* | 0 | 1 | 0 | 0 | 0 | 1 |
| 8 | *c* | 0 | 1 | 0 | 0 | 0 | 1 |
| 4 | *b* | 0 | 1 | 0 | 1 | 0 | 1 |
| 4 | *a* | 0 | 1 | 0 | 1 | 0 | 1 |

**Table 3**
EPCs for every element in the $La_2CuO_4$ unit cell in readable format.

| 8 La atoms (8) | 2 Cu atoms (2) | 16 O atoms (37) | | |
|---|---|---|---|---|
| 4*a* + 4*b* | 4*a* | 4*a* + 4*b* + 8*c* | 8*e* + 8*g* | 8*e* + 8*i* |
| 8*c* | 4*b* | 4*a* + 4*b* + 8*d* | 8*f* + 8*g* | 8*f* + 8*i* |
| 8*d* | | 8*c* + 8*d* | 8*g* + 8*g* | 8*g* + 8*i* |
| 8*e* | | 4*a* + 4*b* + 8*e* | 4*a* + 4*b* + 8*h* | 8*h* + 8*i* |
| 8*f* | | 8*c* + 8*e* | 8*c* + 8*h* | 8*i* + 8*i* |
| 8*g* | | 8*d* + 8*e* | 8*d* + 8*h* | 16*j* |
| 8*h* | | 4*a* + 4*b* + 8*f* | 8*e* + 8*h* | 16*k* |
| 8*i* | | 8*c* + 8*f* | 8*f* + 8*h* | 16*l* |
| | | 8*d* + 8*f* | 8*g* + 8*h* | 16*m* |
| | | 8*e* + 8*f* | 8*h* + 8*h* | 16*n* |
| | | 4*a* + 4*b* + 8*g* | 4*a* + 4*b* + 8*i* | 16*o* |
| | | 8*c* + 8*g* | 8*c* + 8*i* | |
| | | 8*d* + 8*g* | 8*d* + 8*i* | |

## 3. Discussion of the algorithm

### 3.1. Estimation of the combination scale

The computation time required to generate all reasonable EPCs is not the main problem for most inorganic crystals. For some simple

problems (crystals with only a few atoms in the unit cell and high symmetry), we only need to check a few EPC models which requires little computation time. However, when the complexity of the problem grows, computation time grows accordingly and we may want to estimate the scale of the problem before applying this method. This is helpful to avoid wasting time on tackling problems with extreme complexity using this method.

The number of combinations to be checked for each element can be written as

$$C_E = \left\lfloor \frac{N_E}{M_a} \right\rfloor \left\lfloor \frac{N_E}{M_b} \right\rfloor \cdots \left\lfloor \frac{N_E}{M_w} \right\rfloor, \tag{8}$$

where $\lfloor N_E/M_x \rfloor = 0$ should be discarded. Generally, the value of $C_E$ can be calculated in advance. The number of reasonable EPCs is often reduced after the elimination process as discussed in §2.1. This number varies with the problems we are tackling and we can use its upper limit $C_E$ instead:

$$N_S = \prod_{E=A}^{K} C_E. \tag{9}$$

The value of $N_S$ can be used as an indicator for the computation scale of the combination problem. Therefore, the execution time of the algorithm can be estimated in advance when the computing speed is known.

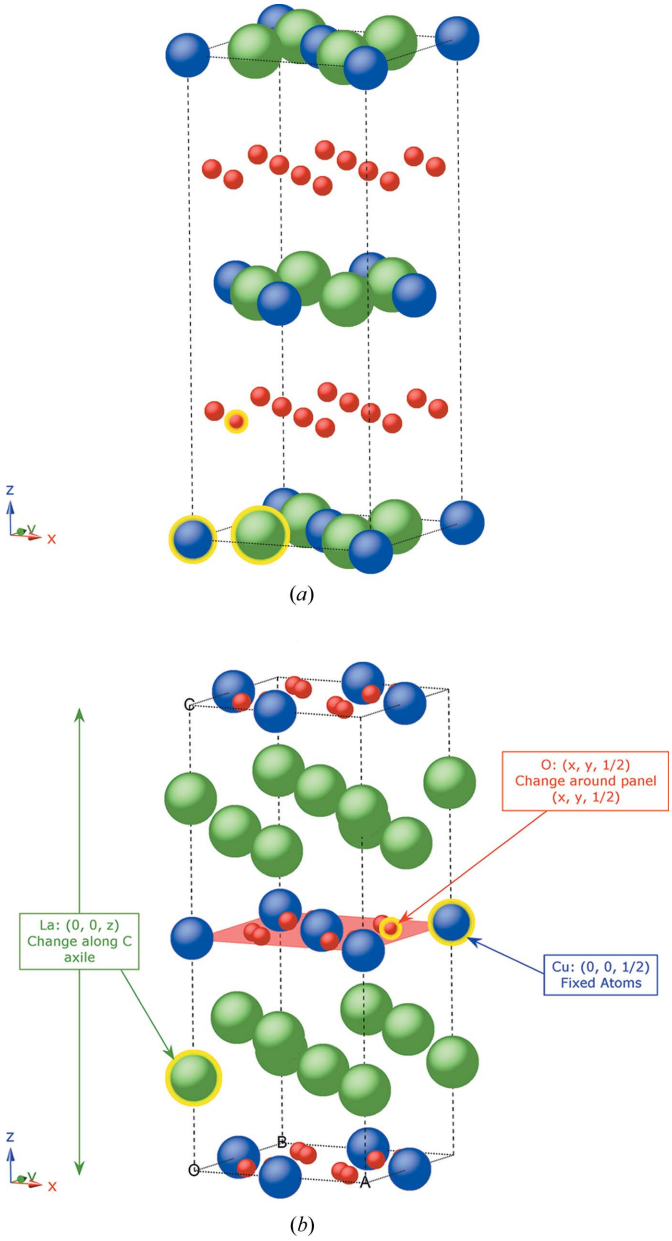


**Figure 3**
Parameterization example of $La_2CuO_4$ EPCs. (*a*) A fixed-coordinate model: Cu 4*a* + La 8*e* + O 8*c* + O 8*d*, coordinates of this model are constant values according to Table 1. No parameters are needed to describe this model. (*b*) A variable-coordinate model: Cu 4*b* + La 8*i* + O 16*o*, Cu atoms are at a fixed Wyckoff position *b*, La atoms are settled at position *i* and have one variable coordinate along the *c* axis. O atoms can change all around a plane which is defined by Wyckoff position *o* and have two variable coordinates. Altogether, only three parameters are needed to describe this model.

### 3.2. Generation of models with fractional occupation

We began our introduction of the EPC algorithm with an assumption that the site-occupation factors of all atoms in the unit cell are 1.0 which means that no fractional occupation happens. Unfortunately, fractional occupation often happens in inorganic crystals, for example, in solid solutions and non-stoichiometric compounds. These problems can be solved consistently with integer SOF situations by adopting the concept of 'imaginary' atoms. We first take solid-solution crystals as an example. The formula of a solid solution can be written as $A_{n-x}B_x \ldots K_m$, where $n$ and $m$ are integer numbers and $x$ is a real number. We can replace the $A_{n-x}B_x$ part with imaginary atoms $D'_n$, then the algorithm can be applied to enumerate all possible EPCs for $D'_n \ldots K_m$. The value of SOF for both $A$ and $B$ can be obtained in the subsequent structure solution and refinement processes. In the case of fractional occupation of equivalent positions by a single element, we can consider $A_{n-x}$ as imaginary atoms and the EPC models can be generated assuming the SOF of the imaginary atoms are 1.0.

### 3.3. Application to organic compounds

This EPC modelling method is mostly suitable for inorganic crystals with independent atoms. However, it is also a helpful tool for the generation of trial structure models of organic crystals in cases when the symmetries of the crystals are higher than triclinic and the molecule has symmetry other than identity.

By considering a molecule as an imaginary atom, *i.e.* ignoring the rotational asymmetry of the molecule, EPC models for a molecular crystal can be generated easily using the above algorithm. Proper crystal structure models can be chosen among these EPC models by a further check of symmetry agreement between the molecule and site positions.

### 3.4. Potential applications in structure prediction

The EPC models generated from the unit-cell content and space-group information may also be used as the starting point for structure predictions because symmetry is a fundamental property of all crystals. Although there is still a long way to go to achieve ideal structure predictions and it is not easy to combine the EPC model generation with other methods, the EPC generation method has potential applications in crystal structure predictions based on first principle calculation, energy minimization and so on.

Model 233:

==========

Parameter Number: 2

Parameters: P1[z] P2[z]

| Atoms: | Element | OX | Wyckoff | X | Y | Z |
|---|---|---|---|---|---|---|
| | Cu | 0 | 4a | 0.000(const0) | 0.000(const0) | 0.000(const0) |
| | La | 0 | 8i | 0.000(const0) | 0.000(const0) | -1.000(P1) |
| | O | 0 | 8e | 0.250(const0) | 0.250(const0) | 0.000(const0) |
| | O | 0 | 8i | 0.000(const0) | 0.000(const0) | -1.000(P2) |

EPC: Cu(4a)+La(8i)+O(8e)+O(8i)

**Figure 4**
A typical model in the output file of *SMEPOC*.

## 4. The *SMEPOC* program

Based on the algorithm discussed in §2, we have developed a computer program, *SMEPOC* (named as a shortened form of 'structure modelling by equivalent position combination' method). *SMEPOC* has been written using the C programming language and can run on both Windows and Linux platforms. Only one input file (.in) is needed. The input file must contain the formula of the material, the number of formula units in the unit cell and the space-group number of the crystal. Optional information such as unit-cell dimensions and element oxidation states can also be given in the input file. With this information, the program will automatically parse the formula, read the Wyckoff position information and list all reasonable combination models under the restraints of equation (4).

### 4.1. Output and usage of *SMEPOC*

The output file of *SMEPOC* is designed carefully so that it can be easily parsed and adopted for further global optimization usage. A typical model in the output is given in Fig. 4.

In the output models, the number of parameters and the variables are specified firstly, and then an atom list of the combination model follows. Coordinates of atoms are either a constant value denoted by a 'const0' token or an initial value −1.0 with variable names in the format of '*Pn*', where *n* is an integer sequence starting from 1. All variable names are listed in the row of 'Parameters'.

More information and examples of the input and output file formats are available at the web page http://sites.google.com/site/smepoc/examples.

Combination models listed in the .mdl output file can be easily converted into the input format of existing software for SDPD such as *Fox* (Favre-Nicolin & Černý, 2002), *Topas* (Bruker, 2000) and *DASH* (David *et al.*, 2006).

### 4.2. Availability

Both Windows and Linux versions are now available from the authors upon request (email dengxiaodi@ssc.iphy.ac.cn). For detailed information about installing and running the program, please refer to the readme.txt file contained in the distribution package.

### 4.3. Future development

Enumerating all possible EPC models is the first essential step for a new computer program for inorganic crystal structure determination using powder diffraction data. Our further research plan includes (1) solving the EPC models automatically by applying suitable global optimization algorithms for models with a different number of parameters, (2) adaptation of the programs for execution on a multi-processor machine by parallel programming, (3) utilization of the geometric and chemical information, such as size, charge and valence of the atoms, to reduce the number of EPC models. The computer program development is now well underway in our laboratory to solve inorganic crystal structures from powder diffraction data using models generated by *SMEPOC*.

This project was supported by the National Natural Science Foundation of China under grant Nos. 20271083 and 20871119.

## References

Andreev, Yu. G., Lightfoot, P. & Bruce, P. G. (1997). *J. Appl. Cryst.* **30**, 294–305.
Brakken, H. & Scholten, W. (1934). *Z. Kristallogr.* **89**, 448–455.
Bruker (2000). *TOPAS*. Bruker AXS Inc., Madison, Wisconsin, USA.
Cascarano, G., Giacovazzo, C., Camalli, M., Spagna, R., Burla, M. C., Nunzi, A. & Polidori, G. (1984). *Acta Cryst.* A**40**, 278–283.
Cascarano, G., Giacovazzo, C. & Guagliardi, A. (1991). *Acta Cryst.* A**47**, 698–702.
Černý, R. (2008). *Chem. Met. Alloys*, **1**, 120–127.
Csoka, T. & David, W. I. F. (1999). *Acta Cryst.* A**55**, Suppl., Abstract No. P08.03.012.
David, W. I. F. & Shankland, K. (2008). *Acta Cryst.* A**64**, 52–64.
David, W. I. F., Shankland, K., McCusker, L. B. & Baerlocher, Ch. (2002). *Structure Determination from Powder Diffraction Data*, pp. 256–261. Oxford University Press.
David, W. I. F., Shankland, K., van de Streek, J., Pidcock, E., Motherwell, W. D. S. & Cole, J. C. (2006). *J. Appl. Cryst.* **39**, 910–915.
Deem, M. W. & Newsam, J. M. (1989). *Nature (London)*, **342**, 260–262.
Earl, D. J. & Deem, M. W. (2005). *Phys. Chem. Chem. Phys.* **7**, 3910–3916.
Facioni, M. & Deem, M. W. (1999). *J. Chem. Phys.* **110**, 1754–1766.
Favre-Nicolin, V. & Černý, R. (2002). *J. Appl. Cryst.* **35**, 734–743.
Feng, Z. J. & Dong, C. (2007). *J. Appl. Cryst.* **40**, 583–588.
Harris, K. D. M., Johnston, R. L. & Kariuki, B. M. (1998). *Acta Cryst.* A**54**, 632–645.
Harris, K. D. M., Tremayne, M., Lightfoot, P. & Bruce, P. G. (1994). *J. Am. Chem. Soc.* **116**, 3543–3547.
Kariuki, B. M., Serrano-González, H., Johnston, R. L. & Harris, K. D. M. (1997). *Chem. Phys. Lett.* **280**, 189–195.
Lu, H.-S. & Liang, C.-K. (1965*a*). *Acta Phys. Sin.* **21**, 849–857.
Lu, H.-S. & Liang, C.-K. (1965*b*). *Acta Phys. Sin.* **21**, 997–1007.
Newsam, J. M., Deem, M. W. & Freeman, C. M. (1992). *Accuracy in Powder Diffraction II*, NIST Special Publication No. 846, pp. 80–91. Washington, DC: US Government Printing Office.
Reddy, J. M., Storm, A. R. & Knox, K. (1965). *Z. Kristallogr.* **121**, 441–448.
Shankland, K., David, W. I. F. & Csoka, T. (1997). *Z. Kristallogr.* **212**, 550–552.