# Symmetry and random sampling of symmetry independent configurations for the simulation of disordered solids

**Philippe D'Arco[1,2], Sami Mustapha[3], Matteo Ferrabone[4], Yves Noël[1,2], Marco De La Pierre[4] and Roberto Dovesi[4]**

[1] UPMC (Université Pierre et Marie Curie) Université Paris 6, ISTEP UMR 7193, F-75005, Paris, France

[2] CNRS, ISTEP, UMR 7193, F-75005, Paris, France

[3] Institut de Mathématiques de Jussieu (UMR 7586 UPMC-CNRS), UPMC, Sorbonne Universités, Paris, France

[4] Dipartimento di Chimica, Università di Torino and NIS - Nanostructured Interfaces and Surfaces - Centre of Excellence, http://www.nis.unito.it, Via P. Giuria 7, 10125 Torino, Italy

E-mail: philippe.d_arco@upmc.fr

**Abstract.** A symmetry-adapted algorithm producing uniformly at random the set of symmetry independent configurations (SICs) in disordered crystalline systems or solid solutions is here presented. Starting from Pólya's formula, the role of the conjugacy classes of the symmetry group in uniform random sampling is shown. SICs can be obtained for all the possible compositions or for a chosen one, and symmetry constraints can be applied. The approach yields the multiplicity of the SICs and allows to operate configurational statistics in the reduced space of the SICs. The present low-memory demanding implementation is briefly sketched. The probability of finding a given SIC or a subset of SICs is discussed as a function of the number of draws and their precise estimate is given. The method is illustrated by application to a binary series of carbonates and to the binary spinel solid solution $Mg(Al,Fe)_2O_4$.

PACS numbers: 02.20.-a, 61.43.-j, 61.50.Ah

*keywords*: Disordered crystalline systems, configuration, symmetry exploitation, Pólya's theory, random sampling, CRYSTAL code

Submitted to: *J. Phys.: Condens. Matter*

## 1. Introduction

Nowadays total and formation energies, equilibrium geometries, vibrational spectra, dielectric and polarizability tensors as well as many other properties of ordered crystalline phases in their equilibrium state can be evaluated routinely for periodic systems with unit cell containing up to 1000 atoms [1, 2] and even more [3] using quantum mechanical computer simulation techniques. However, for disordered, non-stoichiometric systems (substitutional and/or occupational disorder), and/or non-equilibrium states which have a major importance for both Earth and materials science the availability of the same properties still remains an outstanding challenge.
In the last two decades, schemes based on the description of these systems as a weighted average of ordered configurations have provided the most promising results. Among these techniques, parametric ones are preferred to first-principles or *ab initio* schemes because they are less demanding in terms of computational resources. A further step forward consists in using the energies of some of these configurations (obtained from "accurate", possibly *ab initio* calculations) to estimate the coefficients of the so-called cluster expansion (CE) [4]. Then the energies of many other configurations can be estimated at very low cost using the obtained expansion. In a self-consistent manner this expansion can be refined and a few other low energy configurations to be investigated quantum-mechanically selected. Creating a CE for a system is however rather tedious because of formal reasons and the large number of possible configurations. The theoretically exact expansion needs to be truncated. Finding the appropriate number of terms in the truncated CE, their type and range and simultaneously selecting the training set of "accurately" calculated structures among the large number of configurations has been and yet remains a challenging difficulty. If the number of involved positions is $|D|$ (the positions being elements of the set $D$) and the number of species is $|R|$ (the species are the elements of the set $R$), then the total number of configurations over the complete range of compositions is $N = |R|^{|D|}$ (see Section 2 for complete notation). Interesting strategies have been developed to circumvent these difficulties avoiding largely a trial and error based research and empirically truncated expansions. An alternative implementation of the mixed-basis CE [5] has been devised to select interactions by variational procedure [6]. Van de Walle and Ceder [7] have proposed a formal statistical basis for the selection of both the coefficients and the structures to use, relying on the concepts of cross-validation and variance minimization. In the quest for "satisfactory" CE, the use of genetic algorithms has been proposed [8] to facilitate the elaboration of CE containing many-body interactions with three and more sites. More recently Ceder and Muller [9] have proposed to incorporate physical insights in the CE learning process, using a Bayesian framework.
In crystalline structures (of any dimensionality), symmetry partitions the $R^D$ set of configurations in equivalence classes or orbits. Thus, it is sufficient to know the number of equivalence classes, one class representative and the number of symmetry equivalent configurations, or multiplicity, per class. The full set of representatives is the *smallest*

*possible* subset of symmetry independent configurations (SICs) [10]. This may still be a quite large set; the enumeration of the SICs becomes rapidly a challenging problem as $|D|$ or $|R|$ increase.

In the past, simple systems with only one or a few atoms in the primitive cell, such as metallic alloys, have been considered. In that case the problem was to identify SICs within supercells derived from the parent structure, in which all atoms are translationally equivalent (see for example the excellent paper by Ferreira *et al.* [11], references therein and also Refs. [12, 13]). More recently, the major contribution by Hart and Forcade[14] showed how a very fast, linearly scaling algorithm can be designed to produce SICs, by using the so-called "Hermite normal form" of integer matrices. The method was subsequently extended to non-simple parent lattices by introducing a multilattice model [15].

Many crystalline compounds have a large unit cell that contains various atoms occupying positions either involved or not in the substitutions/disorder. For such compounds, the number of configurations becomes rapidly prohibitive. In these cases, space groups are the natural instruments to describe symmetry, as discussed in Ref. [10]. This point of view was developed by Grau-Crespo *et al.* [16] and applied to a set of large cell minerals [17, 18, 19]. In their enumeration algorithm, the space group operators play a crucial role. However, run-time scales quadratically with the number of SICs. The authors emphasize the importance of the multiplicity of the configurations to compute average properties. Multiplicity is obtained by applying all operators of the symmetry group to any new SIC.

In our previous paper [10] using space groups, we have shown that group action theory is a powerful framework to count and enumerate SICs and to derive their multiplicity for complex solid solutions or disordered systems, involving several independent sites, more than two atomic species and even spin distributions. The proposed scheme scales linearly with the number of SICs. Within the same framework, the complete set of all possible compositions or only a selected one can be analysed.

Whatever the efficiency of the algorithm is, the calculation of accurate average properties still remains a challenging task due to the rapid increase in the number of SICs as a function of both the cell size and compositions to be explored. Even for systems of modest size, a fully *ab initio* approach turns out to be inaccessible. For this reason till now either atomistic simulations including long-range electrostatic and short-range forces or cluster expansions, properly parametrized via effective cluster interactions, have been employed. Average properties can be computed by considering either the full set of SICs [17] (when not too large) or a subset of configurations chosen randomly via a Monte Carlo (MC) sampling [17] or by means of the Monte Carlo Metropolis (MCM) algorithm [20].

All these approaches suffer several drawbacks. MC or MCM coupled with a model Hamiltonian neither guarantee an accurate description of the atomic interactions nor allow relaxations. Then, effective properties other than energy are not avalaible as a configurational average [17]. Furthermore, in the case of very large simulation

cells, the convergence character of the MCM Markov chain cannot be unambiguously demonstrated, as mentioned by Metropolis *et al.* in 1953 [21].

The difficulty in reaching the most stable configuration by exchanging atoms is well illustrated in the studies of the Ca-(Mg,Mn) carbonates series [22, 17]. Vinograd *et al.* (2007)[22] used the (experimental) ordered dolomite structure, in which Ca and Mg alternate on parallel planes, as the initial configuration for supercell static structure energy calculations on a large set of randomly chosen structures at fixed composition $Mg_{0.5}Ca_{0.5}CO_3$. Wang *et al.*(2011) [17] obtained reasonable average properties through a MC sampling of the full configuration space for the composition $Mn_{0.5}Ca_{0.5}CO_3$, by using the "dolomite" configuration as the starting one (see their Figure 6). They encountered convergence difficulties for the composition $Mn_{\frac{1}{6}}Ca_{\frac{5}{6}}CO_3$, despite the fact that the corresponding configuration space is much smaller (see their Figure 1).

The choice of the starting configuration requires prior knowledge of the solid solution ordering and of the most stable configuration. It appears as a necessity because the random atom exchanges performed in MC schemes produce configurations belonging to an extremely large space. Besides that, very low energy structures tend to be symmetric and therefore to correspond to low multiplicity SICs [23]. For instance in Refs. [22, 17], whatever $n \times n \times 1$ ($n$ =2,3,4) supercells are used, the multiplicity of the fully ordered $(Mg,Mn)_{0.5}Ca_{0.5}CO_3$ configuration is two. In the smallest supercell (2x2x1) almost 3 million configurations exist. As far as we understand it, the adopted MC sampling scheme explores the full configuration space and does not take advantage of symmetry. Then, MC and MCM Markov processes have an extremely low probability of reaching the most stable "dolomite" configuration, despite both of them are ergodic. In such scheme, the partition function and the Boltzmann weighted averages are obtained exploiting the frequency of occurence of the symmetry equivalent configurations. Equivalent calculations are thus performed several times. Moreover, if low energy configurations are missing, average properties can be severly biased.

In order to totally or partially overcome these difficulties, the partition function and the Boltzmann weighted averages should be obtained by considering SICs and their multiplicity [16]. In other words, an improved MC able to hit SICs independently of their multiplicity (i.e. uniformly) would be highly desired. Once such an improved MC is available, the growth of computational resources is such that "the use of *ab initio* techniques to study disorder in systems, which previously were inaccessible" [16] may become practicable.

The main goal of the present work is to show how the SICs space can be sampled uniformly at random for both variable and constant composition, how the number of SICs of any composition can be calculated and how symmetry constraints can be imposed on the searched SICs. The case of two-species disorder will be considered in detail, while opportunities inspired by Hart *et al.* (2012) [25] for more species will be briefly outlined.

The method has been implemented in a development version of the CRYSTAL code [24]. However, it is not software-related and could easily be ported in other codes.

The formalism is described in Section 2. Starting from Pólya's formula, a probability distribution on conjugacy classes is built according to Dixon and Wilf [26]. Selection of classes according to this distribution allows to generate uniformly at random the full set of SICs over all possible compositions. This approach is then refined for selecting SICs at a given composition. Implementation details are discussed in Section 3. In Section 4, the probability of finding a given SIC or a subset of SICs is calculated. This permits to estimate the number of steps required in order to obtain either a given SIC or a subset of SICs. Section 5 applies the proposed scheme to two examples that are of great geochemical interest: binary carbonates with calcite structure and Fe-Al exchange in the octahedral site of spinel. Finally, the main conclusions are drawn in Section 6.

## 2. Mathematical background

For the presentation of the method, we adopt the notation introduced in Ref. [10]. Let $D$, $R$ and $S$ be a finite set of objects, a set of colors $\{\text{color } 1, ..., \text{color } |R|\}$ and the set of all mappings $s$ from $D$ to $R$ ($S = R^D$) that associate colors to objects, respectively. Each $s$ is called a coloring of the set of objects, or a *configuration*. For $r \in R$ we shall denote $s^{-1}(r) = \{x \in D;\ s(x) = r\}$ the preimage of $r$, which associates to $r$ the set of objects of $D$ colored by this color. We shall refer to the $|R|$-tuple $(|s^{-1}(r_1)|, .., |s^{-1}(r_{|R|})|)$ as the color-pattern, or *composition*, of the configuration $s$. For a given $s$, it comes $\sum_{j=1}^{|R|} |s^{-1}(r_j)| = |D|$. If $G$ is a group of symmetry operators acting on the set $D$, then the isomorphism classes of configurations (SICs) are the orbits of the action induced by $G$ on the set $S$ (Pólya's action). This action is defined by:

$$(g \cdot s)(x) = s\left(g^{-1}x\right), \quad g \in G, \ \ x \in D. \tag{1}$$

An appropriate description of (1) is obtained once $G$ is identified as a subgroup of permutations of $D$ [27, 10]. Then, the cycle decomposition of each symmetry operator $g \in G$ provides a partition of $D$. We shall denote $Cyc_D(g)$ the set of elements of this partition and $|Cyc_D(g)|$ its length. The "type" of $g$ is the $|D|$-tuple $\mathcal{T}_D(g) = (l_1, l_2, .., l_{|D|})$, whose $i^{th}$ value $l_i$ indicates the number of cycles of length $i$ (obviously, $\sum_i l_i \cdot i = |D|$). Keep in mind that all operators belonging to the same conjugacy class have the same type and therefore the same number of cycles. This justifies the notation $\mathcal{T}_D(\mathcal{C}_j)$ to indicate the type of any element of the class $\mathcal{C}_j$. Once $G$, $R$ and $\mathcal{T}_D(g)$ for each symmetry operator are known, the Pólya's formula allows to count the number of orbits (or SICs):

$$|\Delta(S)| = \frac{1}{|G|} \sum_{g \in G} |R|^{|Cyc_D(g)|}. \tag{2}$$

where $\Delta(S)$ is the set of SICs. Note that this formula is not a tool to obtain representatives for the orbits. Instead, orderly generation combined with surjective resolution [10] produces efficiently representatives of the orbits. The corresponding algorithm scales as $\approx |R|^{|D|}$ [10]. Due to this scaling, the cost of an exhaustive enumeration can become prohibitive. An alternative approach is then required.

### *2.1. The Dixon-Wilf algorithm*

The Dixon-Wilf algorithm [26], originally developed for the random generation of all unlabelled graphs on $n$ vertices, proceeds by selecting conjugacy classes in $G$ with appropriate probabilities and then choosing a graph uniformly at random from those stabilized by a representative of the chosen class. In our case, this would correspond to the generation of all the configurations over the possible chemical compositions. One of our goals is the study of fixed compositions, so it is worthwhile to introduce the original Dixon-Wilf approach and then to present an alternative point of view, which allows for further refinements.

The starting point is the fact that Pólya's formula, Eq. (2), can be factorized by conjugacy classes, since two elements $g$ and $g'$ lying in the same conjugacy class satisfy $|Cyc_D(g)| = |Cyc_D(g')|$. Therefore:

$$|\Delta(S)| = \frac{1}{|G|}\sum_{j=1}^{|\mathcal{C}|}|\mathcal{C}_j||R|^{|Cyc_D(g_j)|},$$

where $g_j$ is a representative of $\mathcal{C}_j$. From this expression it follows that:

$$\sum_{j=1}^{|\mathcal{C}|}\frac{|\mathcal{C}_j||R|^{|Cyc_D(g_j)|}}{|\Delta(S)||G|} = 1. \tag{3}$$

Eq. (3) defines a probability distribution on the set of conjugacy classes: $\mathcal{C} = \{\mathcal{C}_1, ..., \mathcal{C}_{|\mathcal{C}|}\}$:

$$\mathbf{Prob}(\mathcal{C}_j) = \frac{|\mathcal{C}_j||R|^{|Cyc_D(g_j)|}}{|\Delta(S)||G|}, \quad j = 1, ..., |\mathcal{C}|. \tag{4}$$

In previous equations, $|R|^{|Cyc_D(g_j)|}$ is the number of configurations $s$ stabilized by $g_j$. The set of these configurations is denoted $S_{g_j}$. The reason is that, considering the cycle structure of $g_j$, $s$ is stabilized by $g_j$ if and only if every cycle of $g_j$ has all its elements mapped to one and only one color. Then, a natural correspondence between the set of fixed points $s \in S_g$ and the set of all mappings from $Cyc_D(g)$ to $R$ exists. It follows that two conjugate operators satisfy:

$$|S_g| = |S_{g'}|. \tag{5}$$

Within $S_{g_j}$, each configuration $s$ has the same probability $\frac{1}{|S_{g_j}|} = |R|^{-|Cyc_D(g_j)|}$. Then, the probability that an orbit $\omega$ contains a given configuration $s$ assuming that it is an element of $S_{g_j}$ is:

$$\mathbf{Prob}(\omega \ni s | s \in S_{g_j}) = \frac{|\omega \cap S_{g_j}|}{|S_{g_j}|}. \tag{6}$$

Based on Eqs. (4) and (6), Dixon and Wilf [26] showed that a uniform distribution on $\Delta(S)$ can be defined.

Here, we propose an alternative construction of the uniform distribution by considering the product structure $G \times S$ and projecting a canonical probability distribution on $\mathcal{C}$

and $\Delta(S)$. We will take advantage of a fundamental property of the fixed points sets $S_g$ used by the authors in Ref. [26]. This property is a refinement of the equality given in Eq. (5). The fact that two conjugate operators have the same cycle structure permits to put the two corresponding fixed point sets into a bijective mapping. Dixon and Wilf observed that this applies to the level of the orbits, too. More precisely, when $g$ and $g'$ lie in the same conjugacy class, for each orbit $\omega \in \Delta(S)$ we have:

$$|S_g \cap \omega| = |S_{g'} \cap \omega|. \tag{7}$$

A sum over all conjugacy classes gives:

$$\sum_{j=1}^{|\mathcal{C}|} |S_{g_j} \cap \omega||\mathcal{C}_j| = \sum_{g \in G} |S_g \cap \omega|,$$

where the r.h.s. can be interpreted as the cardinal of the subset $E_\omega \subset G \times \omega$ defined by:

$$E_\omega = \{(g, s);\ \ g \cdot s = s\} = \bigcup_{s \in \omega} G_s \times \{s\}. \tag{8}$$

Here $G_s$ denotes the stabilizer of the configuration $s$. Using the fact that all the stabilizers in Eq. (8) are conjugate and satisfy $|G_s| = |G|/|\omega|$, we deduce that $|E_\omega| = |\omega||G_s| = |G|$. This implies that

$$\sum_{j=1}^{|\mathcal{C}|} |S_{g_j} \cap \omega||\mathcal{C}_j| = |G|. \tag{9}$$

Now, a useful conceptual tool that accounts for the action of the symmetry group on the set of configurations $S$ is the action matrix as defined by Williamson [27]. This quantity allows us to provide a natural derivation for the distribution (4). The action matrix is a $|G| \times |S|$ matrix $M$ whose entries satisfy $M(g, s) = 1$ if $g \cdot s = s$ and $M(g, s) = 0$ if $g \cdot s \neq s$. An example is given in Figure 1a. The row sums $\sum_s M(g, s)$ and $\sum_s M(g', s)$ are equal when $g$ and $g'$ are conjugate (Eq. (5)). Similarly, if $s$ and $s'$ are on the same orbit $\omega$ the column sums $\sum_{g \in G} M(g, s) = \sum_{g \in G} M(g, s')$ because the stabilizers $G_s$ and $G_{s'}$ are conjugate. This matrix can be divided in $|\mathcal{C}| \times |\Delta(S)|$ blocks, each one corresponding to a class and an orbit. Starting from $M$, a new matrix $A$ can be defined, whose size is $|\mathcal{C}| \times |\Delta(S)|$ and whose entries are $A(j, l) = \sum_{g \in \mathcal{C}_j; s \in \omega_l} M(g, s)$. In other words, the value of $A(j, l)$ corresponds to the number of "1" in the corresponding block of $M$. By using Eq. (7), it comes that:

$$A(j, l) = |S_{g_j} \cap \omega_l||\mathcal{C}_j|. \tag{10}$$

$A$ is referred to as the contracted action matrix (see Figure 1b). All entries are positive, so that upon proper normalization $A$ induces a probability distribution $\{\mathbf{Prob}(\mathcal{C}_j, \omega_l);\ 1 \leq j \leq |\mathcal{C}|,\ 1 \leq l \leq |\Delta(S)|\}$ on the product $\mathcal{C} \times \Delta(S)$. The normalization is performed dividing each entry by the sum over all entries. Using Eq. (9):

$$\sum_{j,l} A(j, l) = \sum_l \sum_j |S_{g_j} \cap \omega_l||\mathcal{C}_j| = |\Delta(S)||G|, \tag{11}$$

so that:

$$\mathbf{Prob}(\mathcal{C}_j, \omega_l) = \frac{|S_{g_j} \cap \omega_l||\mathcal{C}_j|}{|\Delta(S)||G|}. \tag{12}$$

The canonical projections $\mathcal{C} \times \Delta(S) \to \Delta(S)$ and $\mathcal{C} \times \Delta(S) \to \mathcal{C}$ can be used to define two probability distributions on $\Delta(S)$ and $\mathcal{C}$, respectively:

$$\mathbf{Prob}(\omega_l) \ = \sum_j \mathbf{Prob}(\mathcal{C}_j, \omega_l) = \frac{\sum_j A(j,l)}{|\Delta(S)||G|}, \tag{13a}$$

$$\mathbf{Prob}'(\mathcal{C}_j) = \sum_l \mathbf{Prob}(\mathcal{C}_j, \omega_l) = \frac{\sum_l A(j,l)}{|\Delta(S)||G|}. \tag{13b}$$

For any orbit $\omega_l$ we have, according to Eq. (9): $\sum_j A(j,l) = \sum_j |S_{g_j} \cap \omega_l||\mathcal{C}_j| = |G|$. Therefore:

$$\mathbf{Prob}(\omega_l) = \frac{1}{|\Delta(S)|}, \quad l = 1, \ldots, |\Delta(S)|. \tag{14}$$

The last formula shows that a uniform distribution on the orbits exists in natural relation with the probability distribution on $\mathcal{C}$ defined by Eq. (13*b*). It turns out that the probability distribution (13*b*) is equal to the Dixon-Wilf distribution defined by (4):

$$\mathbf{Prob}'(\mathcal{C}_j) = \frac{\sum_l |S_{g_j} \cap \omega_l||\mathcal{C}_j|}{|\Delta(S)||G|} = \frac{|\mathcal{C}_j||S_{g_j}|}{|\Delta(S)||G|} = \frac{|\mathcal{C}_j||R|^{Cyc_D(g_j)}}{|\Delta(S)||G|}.$$

From this result it follows that, if one proceeds as indicated at the begining of this paragraph: 1) select at random a set of conjugacy classes according to $\mathbf{Prob}(\mathcal{C}_j)$; 2) build at random a stabilized configuration for each selected class, then orbits are found with same probability $\frac{1}{|\Delta(S)|}$.

We observe that the probability of the conjugacy classes can be redefined for any subset of orbits considering the sub-matrix of the matrix action corresponding to the chosen subset. This remark will be used in Section 2.2 where the random selection of SICs at fixed composition is described in the case of two colors. The set $\Delta(S)$ must be replaced by the set $\Delta(S_\alpha)$ of SICs corresponding to composition $\alpha$. Furthermore, it is possible to unbalance the distribution probability $\mathbf{Prob}(\omega_l)$ by considering only some conjugacy classes. This technique can be used to enhance the visibility of specific orbits, e.g. high symmetry ones.

### *2.2. The random selection at fixed composition*

The previously stated restriction of the approach to some orbits or classes will now be illustrated. This will provide the refinement of the Dixon-Wilf algorithm proposed by Kerber et al. [28] to generate uniformly at random orbits with a fixed composition. Furthermore, it will offer new opportunities. This derivation is presented in the case of two colors:

$$R = \{\text{color } 1, \text{color } 2\}.$$

Extension to more than two colors is briefly discussed in the Appendix, based on the ideas exposed by Hart *et al.* (2012) [25]. We refer the reader to paragaph 2.2 of Ref. [10], with particular focus on Eq. (17). Since all conjugated operators have the same type, this equation can be factorized by conjugacy classes:

$$\sum_{s'\in\Delta(S)} W(s') = \frac{1}{|G|}\sum_{j=1}^{|\mathcal{C}|} |\mathcal{C}_j|(x+y)^{b_1^j}(x^2+y^2)^{b_2^j}\dots(x^{|D|}+y^{|D|})^{b_{|D|}^j}, \qquad (15)$$

where $b_i^j$ are the elements of the $|D|$-tuple $\mathcal{T}_D(\mathcal{C}_j)$ and where the variables $z_1$ and $z_2$ of Eq. (17) in Ref. [10] have been replaced by $x$ and $y$, respectively. By expanding the product terms of the sum $(x+y)^{b_1^j}(x^2+y^2)^{b_2^j}\dots(x^{|D|}+y^{|D|})^{b_{|D|}^j}$, the r.h.s. of Eq. (15) becomes:

$$\frac{1}{|G|}\sum_{j=1}^{|\mathcal{C}|}|\mathcal{C}_j|\sum_c \frac{b_1^j!\dots b_{|D|}^j!}{c_1!(b_1^j-c_1)!\dots c_{|D|}!(b_{|D|}^j-c_{|D|})!}x^{\sum_i ic_i}\cdot y^{|D|-\sum_i ic_i},$$

where the sum $\sum_c$ runs over all the non-negative integer vectors $c = \big(c_1,\dots,c_{|D|}\big)$, satisfying $0 \le c_1 \le b_1^j,\dots,0 \le c_{|D|} \le b_{|D|}^j$. By rearranging in agreement with composition, $\alpha$, the r.h.s. of Eq. (15) becomes:

$$\sum_{\alpha=0}^{|D|}\frac{1}{|G|}\Big[\sum_{j=1}^{|\mathcal{C}|}|\mathcal{C}_j|\sum_{c\in\Lambda_{j,\alpha}}\frac{b_1^j!\dots b_{|D|}^j!}{c_1!(b_1^j-c_1)!\dots c_{|D|}!(b_{|D|}^j-c_{|D|})!}\Big]x^\alpha y^{|D|-\alpha}, \qquad (16)$$

where $\Lambda_{j,\alpha}$ is the set of vectors $c$ satisfying

$$0 \le c_1 \le b_1^j,\dots,0 \le c_{|D|} \le b_{|D|}^j,\ \sum_{i=1}^{|D|} ic_i = \alpha. \qquad (17)$$

The polynomial in Eq. (16) calls for some comments. The sum of products of binomial coefficients on the right should be interpreted as the number of ways of forming configurations stabilized by one representative of the class $\mathcal{C}_j$ and possessing composition $\alpha$. More precisely, for each $i = 1,\dots,|D|$ there are $\begin{pmatrix} b_i^j \\ c_i \end{pmatrix}$ ways of mapping $c_i$ cycles among the $b_i^j$'s of length $i$ on color 1. The last condition in Eq. (17) guarantees to obtain the chosen composition $\alpha$.

From Eqs. (15) and (16) it follows that the number of orbits with composition $(\alpha, |D|-\alpha)$ is:

$$|\Delta(S_\alpha)| = \sum_{j=1}^{|\mathcal{C}|}\sum_{c\in\Lambda_{j,\alpha}}\frac{|\mathcal{C}_j|}{|G|}\prod_{i=1}^{|D|}\begin{pmatrix} b_i^j \\ c_i \end{pmatrix}. \qquad (18)$$

Here, $|\Delta(S_\alpha)|$ replaces $k_{(\alpha,|D|-\alpha)}$ in Eq. (18) of Ref. [10]. The sum contains as many terms as the number of possible decompositions of $\alpha$ compatible with Eq.(17). The contributing products are those in which all $c_i \le b_i^j$. This equation offers a convenient

way for computing the coefficients of the Pólya's polynomial.
By rewriting Eq.(18) it turns out that:

$$|\Delta(S_\alpha)||G| = \sum_{j=1,\dots,|\mathcal{C}|} \sum_{c\in\Lambda_{j,\alpha}} |\mathcal{C}_j| \prod_{i=1}^{|D|} \begin{pmatrix} b_i^j \\ c_i \end{pmatrix}. \tag{19}$$

In Eq.(19), $\sum_{c\in\Lambda_{j,\alpha}}$ is the number of configurations with composition $\alpha$ that are stabilized by an element of $\mathcal{C}_j$. So:

$$\sum_{c\in\Lambda_{j,\alpha}} |\mathcal{C}_j| \prod_{i=1}^{|D|} \begin{pmatrix} b_i^j \\ c_i \end{pmatrix} = \sum_{\omega\in\Delta(S_\alpha)} |S_{g_j} \cap \omega||\mathcal{C}_j|. \tag{20}$$

Eq. (20) is analoguous to Eq. (11), thus Eqs. (13$a$)-(13$b$) can be applied to chosen composition by substituting $|\Delta(S_\alpha)|$ for $|\Delta(S)|$. Therefore, a probability distribution adapted to the composition $\alpha$ can be defined on the set of conjugacy classes $\mathcal{C} = \{\mathcal{C}_1, \dots, \mathcal{C}_{|\mathcal{C}|}\}$:

$$\mathbf{Prob}_\alpha(\mathcal{C}_j) = \frac{\sum_{c\in\Lambda_{j,\alpha}} |\mathcal{C}_j| \prod_{i=1}^{|D|} \begin{pmatrix} b_i^j \\ c_i \end{pmatrix}}{|\Delta(S_\alpha)||G|}. \tag{21}$$

Each decomposition of $\alpha$ corresponds to $\prod_{i=1}^{|D|} \begin{pmatrix} b_i^j \\ c_i \end{pmatrix}$ $g_j$-stabilized configurations and contributes to the set of the configurations proportionally. It has a weight equal to:

$$\mathbf{Prob}[(c_1, \dots, c_{|D|})] = \frac{\prod_{i=1}^{|D|} \begin{pmatrix} b_i^j \\ c_i \end{pmatrix}}{\sum_{c\in\Lambda_{j,\alpha}} \prod_{i=1}^{|D|} \begin{pmatrix} b_i^j \\ c_i \end{pmatrix}}. \tag{22}$$

The same approach can be employed on a set of compositions instead of only one. For example, in the case of two compositions ($\alpha$ and $\alpha'$), in Eq. (21) the sum $\sum_{c\in\Lambda_{j,\alpha}}$ must be replaced by $\sum_{c\in\Lambda_{j,\alpha}} + \sum_{c'\in\Lambda_{j,\alpha'}}$ and $|\Delta(S_\alpha)|$ by $|\Delta(S_\alpha)| + |\Delta(S_{\alpha'})|$.

### *2.3. Symmetry-enhanced selection*

In Section 2.1 it has been shown that the probability distribution of conjugacy classes and orbits are naturally related. So, if the probability distribution on classes is modified, the probability distribution on orbits will deviate from the uniform distribution. This point can be well illustrated by considering only the identity class, i.e. the first line of the reduced action matrix, and ignoring all other classes. The probability related to an orbit is proportional to its length (or multiplicity):

$$\mathbf{Prob}(\omega) = \frac{|\omega|}{|S|}. \tag{23}$$

If multiplicities of different orbits widely differ with each other, shortest orbits will be sampled with very low probability, so that they may require a very large number of tries

before being picked up.
On the other hand, identity could be removed from the conjugacy classes available for selection. In this way, the probability distribution on orbits is modified, favoring the shortest ones against the longest ones. In other words, the orbits whose elements have largest stabilizers become the most promoted, while the asymmetric orbits become unaccesible, their probability being zero. More precisely, the corresponding probability is given by:

$$\mathbf{Prob}(\omega) = \frac{|G| - |\omega|}{|\Delta(S)||G| - |S|}. \tag{24}$$

Removing the identity might be useful for large systems in which the asymmetric orbits are by far the most numerous ones.
Note that Eqs. (23) and (24) have been derived assuming no chemical restriction. They apply also to fixed composition, by replacing $S$ by $S_\alpha$.

## 3. Details of the implementation

The implementation is divided in two parts. The first one is devoted to the evaluation of the probability distribution of the conjugacy classes, according to Eq. (21). It requires the decomposition for composition $\alpha$ according to Eq. (17) and the calculation of $\Delta(S_\alpha)$ according to Eq. (18). The probability distribution is scaled over an integer range by multiplying the probabilities by $|\Delta(S_\alpha)||G|$. Thus, classses are selected by drawing integer numbers in the range $[1, |\Delta(S_\alpha)||G|]$.
Composition is described by one integer $\alpha$ in a 2-color problem and by $|R| - 1$ integers in a $|R|$-color one; its decompositions, Eq. (17), are found recursively. The largest value of $i$, denoted $m$, is smaller or equal to $|D|$; in practice, it is equal to the length of the longest cycle among all cycles of all operators (see comments in text). Then, $c_m$ takes any of the $\alpha \div m + 1$ values $0, \ldots, \alpha \div m$. For a chosen value of $c_m$, the residual $\alpha_{m-1} = \alpha - m \cdot c_m$ can be or not equal to zero. In the first case, all $c_{1,\ldots,m-1}$ are equal to 0, no other decomposition exists. In the second case, some entries among $\{c_{1,\ldots,m-1}\}$ are not equal to zero. Then, $c_{m-1}$ is obtained as $c_m$ by substituting $\alpha_{m-1}$ and $m-1$ for $\alpha$ and $m$, respectively. As previously, $c_{1,\ldots,l<m} = 0$ as soon as $\alpha_l = 0$.
Orbits are generated in the second part of the implementation. Five steps are required:

- a conjugacy class is randomly selected;
- within the class a representative is selected at random;
- a decomposition of $\alpha$ is chosen at random according to Eq. (22);
- the various subsets of cycles of different length required according to the selected decomposition of $\alpha$ are picked up randomly;
- the selected subsets are mapped to the appropriate color.

These steps are repeated either for a predefined number of times or until a predefined number of orbits is obtained. In the first case, the loop stops as soon as $|\Delta(S_\alpha)|$ orbits

are found.

The quality of the random number generator is crucial for an appropriate implementation, as random numbers are required at least at three levels, i.e. drawing the conjugacy class, the decomposition of $\alpha$ and the set of cycles. Note that in our implementation, we have chosen to select the representative element of conjugacy classes at random in order to reduce any biais of the random generator.

This algorithm suffers of the so-called isomorphism desease [29]: a given class of configurations can be sampled several times. If heavy calculations have to be performed, one would like to get a unique representative for each class. However, orbits are not orderly generated, so that no simple canonicity test can be performed.

To solve this issue, a logical vector of some length is introduced, which has to be initialized to 0. Once a configuration is obtained, its rank is calculated using some ranking function. If the corresponding logical element is equal to 0, the configuration belongs to an unknown (new) orbit, and the logical element is set to 1. Then, the orbit of the configuration is generated by applying all the operators of the group, the rank of its elements is calculated and the corresponding elements in the logical vector are set to 1. Every time the obtained configuration does not belong to a new orbit, a new draw is performed. Two ranking functions have been tested. On the one hand, a lexicographical rank function can be used, which however is not a perfect minimal hash function in the case of fixed composition. So, the length of the boolean vector has to be larger than $|S_\alpha|$ and equal to $|R|^{|D|}$. On the other hand, a formally better choice is based on the perfect hash function proposed by Hart *et al.* (2012) [25]. In this case, the logical vector exactly contains $|S_\alpha|$ element. Both functions have been implemented, the former being about 3 times faster than the latter, but requiring a much larger boolean vector. Note however that up-to-date computer memory ressources are such that, in the case of a few million configurations, this solution requires a reasonably small fraction of the total memory.

The process illustrated above provides a number of by-products, namely the stabilizer of the configuration, its multiplicity (obtained as the quotient between $|G|$ and the cardinal of the stabilizer) and the canonical representative of the orbit containing the configuration (chosen as the lowest rank equivalent configuration).

## 4. How many draws are required to find a given subset of SICs?

For convenience, all the discussion is made in the case we consider all possible compositions. The special case of a chosen composition can be covered just by replacing $\Delta(S_\alpha)$ for $\Delta(S)$.

Any series of $t$ draws yields a sequence of $t$ SICs belonging to $\Delta(S)$. This sequence is one of the $|\Delta(S)|^t$ equiprobable possibilities. Its probability is $\frac{1}{|\Delta(S)|^t}$. Then one can wonder how many tries must be performed to obtain a fixed number of SICs or orbits with some minimal probability $\rho$.

Let us select one orbit and denote by $\mathrm{P}_t^{(1)}$ the probability that this orbit has been

observed at least once in the $t$ tries. Let $E_1$ denote the corresponding event. The complementary event ($E_{\bar{1}}$) corresponds to a series of $t$ tries in which this orbit has not been seen. There are $(|\Delta(S)|-1)^t$ such possibilities. Then, $|E_1| = |\Delta(S)|^t-(|\Delta(S)|-1)^t$ and:

$$\mathrm{P}_t^{(1)} = 1 - \left(1 - \frac{1}{|\Delta(S)|}\right)^t. \tag{25}$$

The condition $P_t^{(1)} > \rho$ yields:

$$t > \frac{\ln(1-\rho)}{\ln\left(1-\frac{1}{|\Delta(S)|}\right)}. \tag{26}$$

If $t$ satisfies Eq. (26) then the chosen orbit is obtained at least once with a probability greater than $\rho$.

More generally $\mathrm{P}_t^{(m)}$ denote the probability that $m$ chosen orbits $1,\ldots,m$ have been observed at least once in a series of $t$ draws. $\mathrm{P}_t^{(m)}$ is the probability of the intersection of the events $E_1,\ldots,E_m$. That is $\mathrm{P}_t^{(m)} = \mathbf{Prob}(E_1 \cap \ldots \cap E_m)$. We first consider the case $m = 2$:

$$\mathrm{P}_t^{(2)} = \mathbf{Prob}(E_1 \cap E_2) = \mathbf{Prob}(E_1|E_2)\mathbf{Prob}(E_2). \tag{27}$$

$\mathbf{Prob}(E_2)$ is given by Eq. (25). $\mathbf{Prob}(E_1|E_2)$ relates to the fraction of elements of $E_2$ containing 1. The number of such elements is $|E_2|$ minus the number of elements of the sample space not containing 1, but 2 ; i.e. $|E_{1|2}| = |E_2| - (|E_{\bar{1}}| - |E_{\bar{1},\bar{2}}|)$. One should have in mind that the number of elements of the sample space not containing $n$ selected orbits is simply given by:

$$|E_{\bar{1},\ldots,\bar{n}}| = (|\Delta(S)| - n)^t. \tag{28}$$

Then, it follows that:

$$\mathbf{Prob}(E_1|E_2) = \frac{|E_2| - [(|\Delta(S)|-1)^t - (|\Delta(S)|-2)^t]}{|E_2|}. \tag{29}$$

By combining Eqs. (27), (25) and (29):

$$\mathrm{P}_t^{(2)} = 1 - 2\left(1 - \frac{1}{|\Delta(S)|}\right)^t + \left(1 - \frac{2}{|\Delta(S)|}\right)^t.$$

By proceeding with similar arguments and making a systematic use of Eq. (28) as well as of well-known properties of the binomial coefficient, the following formula can be proved:

$$\mathrm{P}_t^{(m)} = 1 - m\left(1 - \frac{1}{|\Delta(S)|}\right)^t + \sum_{k=2}^{m}(-1)^k \frac{m!}{k!(m-k)!}\left(1 - \frac{k}{|\Delta(S)|}\right)^t.$$

This exact formula makes to obtain a lower estimate of $t$ difficult. However, one can note that:

$$\mathrm{P}_t^{(m)} = \mathbf{Prob}(E_1 \cap \ldots \cap E_m) = 1 - \mathbf{Prob}(E_{\bar{1}} \cup \ldots \cup E_{\bar{m}}) \geq 1 - [\mathbf{Prob}(E_{\bar{1}}) + \ldots + \mathbf{Prob}(E_{\bar{m}})],$$

so that using Eq. (25) it comes:

$$\mathrm{P}_t^{(m)} \geq 1 - m\left(1 - \frac{1}{|\Delta(S)|}\right)^t.$$

Despite the fact that the r.h.s. of the previous equation can take negative values, imposing it to be greater than $\rho$ (thus positive) permits to estimate $t$ for a given set of $m$ SICs with the guarantee that $\mathrm{P}_t^{(m)} \geq \rho$:

$$t > \frac{\ln\left(\frac{m}{1-\rho}\right)}{\ln\left(\frac{|\Delta(S)|}{|\Delta(S)|-1}\right)}. \tag{30}$$

Finally, in order to find all $\Delta(S)$ SICs, Eq. (30) can be approximated as:

$$t \gtrsim |\Delta(S)|[\,\ln|\Delta(S)| - \ln(1-\rho)\,]. \tag{31}$$

## 5. Applications

### *5.1. Counting example: binary carbonates with calcite structure*

Here, the counting capability and the stability of the method are illustrated by applying it to the extensively studied binary carbonate $\mathrm{X}_x\mathrm{Ca}_{1-x}\mathrm{CO}_3$, with $0 \leq x \leq 1$ and X = Mn, Mg, Cd. A calcite crystal structure and a supercell of index 4 containing 24 sites involved in the mixing will be considered; the results will be compared with data obtained with a direct sampling [17].

As previously reported, the most stable configuration at $x = 0.5$ is the ordered dolomite structure. In general, at each composition $n/6$ with $n = 1, 2, 3$, the lowest energy configuration is characterized by homogeneous (0001) layers of cations equally spaced by the maximum possible distance [17] and by the lowest possible multiplicity: 6, 3, and 2, respectively. Among these compositions, $x = 1/6$ corresponds to the smallest configurational space with 10626 configurations partitioned in 102 SICs. In this favorable case, with direct sampling neither the most symmetric (and most stable) configuration was hitted nor a satisfactory partition function was obtained after 500 draws and 500 energy calculations [17]. At other compositions, the probability of the most stable configuration decreases further more, for instance being in the order of $10^{-6}$ for $x = 3/6$. Then, at this composition average properties were calculated by randomly sampling the configurational space but starting from the most symmetric SIC [17].

Figure 2 shows the average number of tries $\langle t \rangle$ necessary to find the full set of SICs at different compositions as a function of the number of SICs. On the basis of Eq. (31), this quantity is expected to vary as:

$$< t >= |\Delta(S)|[\ln|\Delta(S)| + \epsilon(|\Delta(S)|)], \tag{32}$$

where $\epsilon(|\Delta(S)|)$ is a corrective term that, for the considered case, is close to 0.6. At $x = 1/6$ the full set of SICs is found after 530 tries on average. So the probability

$\rho$ of finding the most symmetric SIC is very high after 500 tries, in agreement with (26). Using formula (26) it is possible to estimate the number of tries $t(\rho)$ required in order to obtain with some probability $\rho$ a predefined SIC (Figure 3a). In Figure 3b, the hitting or observed frequency $\nu_{hit}$ of such a predefined SIC after $t(\rho)$ tries, defined as the ratio of the number of occurences the predefined SIC to $t(\rho)$, is plotted as a function of the predefined probability $\rho$ for $x = 1/6$. The agreement between $\nu_{hit}$ and the expected probability $\rho$ is excellent. The present method allows to get the most symmetric configuration with a probability close to 0.99 after 500 tries. Note that this finding is in contrast with the conclusion by Wang *et al.* [17] about the risk of random configurational sampling methods. Eq. (25) can be used to estimate the probability of occurrence of a given SIC in the case of direct configurational sampling [17], by replacing $|\omega|/|S|$ for $1/|\Delta(S)|$. At $x = 1/6$ the multiplicity $\omega$ of the most symmetric configuration is 6, $|S|$ is 10626, so after 500 tries the probability of finding this SIC does not exceed 0.25. To reach a probability of 0.99, more than 8000 draws are required.

For a closer comparison with the data published by Wang *et al.* [17], in Figure 3c the hitting frequency $\nu_{hit}$ is plotted as a function of the probability $\rho$ in the case $x = 1/2$. After about 13000 tries, the most symmetric configuration is obtained with a probability greater or equal to 0.5.

As previously demonstrated, the probability of finding symmetric SICs can be enhanced by ignoring the identity operator. In this way we focus on the subset $\Delta(S^*)$ of symmetric SICs. The observed frequencies of various SICs belonging to $\Delta(S^*)$ are shown as a function of the orbit length $|\omega|$ in Figure 4a and 4b for $x$ equal to 1/6 and 1/2, respectively. The observed trends reflect the increased probability of finding symmetric, low multiplicity SICs predicted by Eq. (24). The efficient sampling of symmetric SICs is illustrated by Figure 5, that shows the average number of tries $\langle t \rangle$ required to find the full subset of symmetric SICs for various compositions. For example, at $x = 1/2$ only 824 out of 19219 SICs are symmetric; the corresponding $\langle t \rangle$ is around 6100. Note that the dependence of $\langle t \rangle$ on $|\Delta(S^*)|$ is the same as found for the general case (see Figure 2 and discussion above). We can state that this symmetry-enhanced selection offers a naive counting mean for symmetric SICs.

### *5.2. Monte Carlo sampling example: cation exchange in spinels*

As a test for the proposed uniform random sampling scheme, the spinel solid solution $Mg(Al_x,Fe^{3+}_{1-x})_2O_4$ has been considered, where $x$ is the atomic fraction of Al. We are aware that in this series the two independent cationic sites of the spinel structure are involved in the mixing [30]. However, site occupancies involve large uncertainties. Then, in this preliminary investigation, we have chosen a normal spinel model, where Al and Fe substitute each other only on the 16 octahedral positions. We think that the possible stability of normal compounds with respect to the two end members deserves some interest.

Modeling has been performed using the conventional cell. Three compositions have been

considered, corresponding to $x$ values of 0.25, 0.50 and 0.75. The cubic symmetry breaks the full configuration spaces in a few SICs. At $x = 0.50$, configurations are partitioned into 97 SICs, among which 49 have a non-trivial space group. At $x = 0.25$ and 0.75, only 22 SICs exist.

First, we investigate the efficiency of the proposed MC sampling method (not MCM), without and with symmetry-enhancing, by analysing the convergence of average energy at fixed composition $x = 0.50$. For each newly found SIC, a full geometry optimization was performed on its representative using the CRYSTAL code. Calculations were performed on the ferromagnetic solution (due to the presence of $Fe^{3+}$ cations) by using the hybrid spin polarized B3LYP functional from the density functional theory (DFT) [31] and an all-electron Gaussian-type basis set of triple zeta plus polarizations quality. Fractional atomic coordinates and unit cell parameters were optimized within a quasi-Newton scheme using analytic energy gradients combined with the BFGS algorithm for Hessian updating. In order to deal with the large number of configurations, an automated, multitasking, massively parallel version of the CRYSTAL code was developed. This implementation allows to launch automatically the geometry optimizations for the selected set of configurations on a large number of processors and is explicitly designed to run on large supercomputing facilities. The multitasking technique permits to perform independent tasks (such as geometry optimizations of different structures) in parallel, to take advantage of the large processor counts available on modern High Performance Computing (HPC) machines, avoiding the performance degradation occurring when the number of allocated CPUs is too large compared to the size of a single problem/task. The calculations on the set of 97 configurations were carried out over 14 tasks by using 3584 CPUs (256 CPUs per task) at the SuperMUC HPC facility; the full set of optimized structures was provided in about 16 hours.

During the exploration of the SICs space, the average energy per cell $\langle E \rangle$ has been calculated, each SIC being assigned a Boltzmann-like probability:

$$\langle E_n \rangle = \frac{\sum_i^n |\omega_i| E_i \exp\left(-\frac{E_i}{k_B T}\right)}{\sum_i^n |\omega_i| \exp\left(-\frac{E_i}{k_B T}\right)},$$

where $n$, $|\omega_i|$, $E_i$, $k_B$ and $T$ are the number of considered SICs, the length or multiplicity of the $i^{th}$ SIC, the energy of a representative of the $i^{th}$ SIC, the Boltzmann constant and the absolute temperature, respectively.

Without symmetry enhancement, the 97 SICs have been randomly generated according to the algorithm here proposed. The running Boltzmann average energy is plotted as a function of the number of MC-generated SICs in Figure 6 (circle data points) at different temperatures: 300 K, 1500 K and infinite limit. In both cases $\langle E \rangle$ converges quite rapidly, i.e. after about 30-40 configurations, to the exact average energy (calculated with all 97 configurations). For comparison, a second Boltzmann average energy has been *a posteriori* calculated, upon reordering of the full set of data by increasing energy (triangle data points in Figure 6). This average can be considered the best possible approximation of the Boltzmann average energy for a given number of SICs; we have

called it the limit average. The MC sampling converges to the exact average more rapidly than the limit average, except at 300 K. This is connected to the fact that at low temperature, low energy configurations significantly contribute to the average. As temperature rises, their contribution reduces.

Severals authors [32, 33, 34, 23] suggested that both the most stable and the less stable configurations are expected to have some symmetry. So, to increase the probability of a rapid sampling of the most symmetric and possibly most stable configurations, a two-step generation of the SICs has been performed according to Section 2.3. First, the identity class has been removed, so that asymmetric SICs become unreachable. Then, once the full subset of symmetric SICs is complete, the identity class has been recovered, and only asymmetric SICs have been obtained. Note that neither the symmetric nor the asymmetric SICs are ordered by energy, as they are generated at random. The corresponding running symmetry-enhanced (SE) Boltzmann average is compared with the MC average in Figure 6 (square data points). As expected, at high temperature the average is weakly dependent on the adopted procedure, either symmetry-enhanced or not. However at low temperature, the SE average converges much more slowly than the MC average. The relatively "small" number of SICs (97) does not allow to draw general conclusions from this preliminary experience. As the size of the system increases, the proportion of symmetric SICs decreases. So, the SE sampling might become useful to reach high symmetry, and possibly low energy, configurations within a reasonable number of draws.

At the two other compositions, $x = 0.25$ and 0.75, the number of SICs is not large enough to perform a study as detailed as the one just presented for $x = 0.50$. In both cases, the average energy has been calculated at 300 K only. The few data (not reported here) suggest that energies of the intermediate compounds are larger than the ones obtained upon linear interpolation between the end members. However, the corresponding excess energies are littler than 2.5 mHa per primitive cell. At this stage, vibrational frequencies are not yet available, so that the Gibbs free energy cannot be estimated. However, the configurational entropy times temperature could be of the same order than this excess energy. A systematic and more complete study of this system seems worthwhile and is currently under progress.

In Figure 7 the volume of the configurations is plotted as a function of the configuration energy, the different symbols relating to the different SIC multiplicities or, equivalently, symmetries. Energy scales have been biased to illustrate the fact that the functional relation between volume and energy does not depend on composition. For all compositions the most stable (and the less stable) configurations are the most compact (and the less compact) ones and possess the lowest multiplicities (6 and 12) and therefore the highest symmetries. This test case does not allow to draw definite conclusion about the energy-volume relation. However, it tends to support the possible connection between energy and symmetry suggested by Wales [33, 34] : the highest and lowest energy configurations would correspond to most symmetric structures. This last suggestion has been exploited in the random searching strategy of stable structures

by Pickard *et al.* [23]. This possible energy-symmetry connection deserves further investigations. Considering the partially ordered lattice of subgroups of the group ($G$), the proposed method can be modified, so that SICs are generated by decreasing the index stabilizer class. Such approach would offer the opportunity to find rapidly the stablest configuration.
Over all, and despite the relatively small number of SICs, the proposed MC sampling scheme proves to be satisfactorily efficient.

## 6. Conclusions

Grau-Crespo *et al.* [16] have called for a symmetry-adapted Monte Carlo method, in order to enhance the application of first-principles methods to the modeling of disordered or solid solution materials. In the present work we have shown how a group action theory derived method can be applied to sample uniformly at random the set of symmetry independent configurations for both varying and fixed compositions. The proposed approach overcomes unbalanced direct sampling of the symmetry independent configurations (SICs) in the full configuration space. The so-emphasized visibility of low multiplicity (high symmetry) SICs can be renforced. This is a major feature of the method because high symmetry configurations often correspond to the most stable ones, which implies they contribute significantly to average properties especially at low temperature. The implementation is extremely straightforward, and does not require the use of large memory. The method also allows to calculate in a very convenient way the number of configurations for a given composition. Extension of the method to more than two colors has been sketched.

## Appendix

The considerations exposed in Section 2.2 can be extended to handle more than two colors by means of the very convenient notation introduced by [25].
Eq. (18) can be recast as:

$$|\Delta(S_\alpha)| = \sum_{j=1}^{|\mathcal{C}|} \frac{|\mathcal{C}_j|}{|G|} \sum_{c\in\Lambda_{j,\alpha}} \prod_{i=1}^{|D|} \frac{b_i^j}{c_i!(b_i^j - c_i)!}. \tag{33}$$

The sum $\sum_{c\in\Lambda_{j,\alpha}}$ is the number of configurations or colorings stabilized by $g_j$. It is equivalent to the general counting formula given in appendix A2 in Ref. [25] for the 2-color case. Eq. (33) permits the generalization to the $|R|$-color case. Here, composition is a $|R|$-tuple $(\alpha_1, \ldots, \alpha_{|R|})$, whose sum of elements is equal to $|D|$. The counting formula (18) becomes:

$$|\Delta(S_{(\alpha_1,\ldots,\alpha_{|R|})})| = \sum_{j=1}^{|\mathcal{C}|} \frac{|\mathcal{C}_j|}{|G|} \sum_{\mathbb{M}\in\Gamma_j} \prod_{i=1}^{|D|} \frac{b_i^j}{c_{i,1}!\ldots c_{i,|R|}!},$$

where $\Gamma_j$ is the set of all $|D| \times |R|$ matrices $\mathbb{M} = [c_{i,k}]$ of non-negative integers satisfying:

$$\sum_{k=1}^{|R|} c_{i,k} = b_i^j \ ; \ \sum_{i=1}^{|D|} i c_{i,k} = \alpha_k \ ; \ 0 \leq \alpha_k \leq |R|. \tag{34}$$

In the 2-color case, these simultaneous conditions are equivalent to Eq. (17).
As matrices $\mathbb{M}$ depend only on the type of the class $\mathcal{C}_j$, Eq. (21) generalizes as:

$$\mathbf{Prob}_{(\alpha_1,\ldots,\alpha_{|R|})}(\mathcal{C}_j) = \frac{|\mathcal{C}_j| \sum_{\mathbb{M}\in\Gamma_j} \prod_{i=1}^{|D|} \frac{b_i^j}{c_{i,1}!\ldots c_{i,|R|}!}}{|\Delta(S_{(\alpha_1,\ldots,\alpha_{|R|})})||G|}. \tag{35}$$

It is sufficient to replace Eq. (22) with:

$$\mathbf{Prob}[\mathbb{M}] = \frac{\prod_{i=1}^{|D|} \frac{b_i^j}{c_{i,1}!\ldots c_{i,|R|}!}}{\sum_{\mathbb{M}\in\Gamma_j} \prod_{i=1}^{|D|} \frac{b_i^j}{c_{i,1}!\ldots c_{i,|R|}!}}, \tag{36}$$

to define the appropriate weight for the $g_j$-stabilized configurations corresponding to the decomposition of the $|R|$-tuple $(\alpha_1, \ldots, \alpha_{|R|})$ induced by the matrix $\mathbb{M}$.

## Acknowledgments

This work was supported by CALSIMLAB under the French funds "Investissements d'Avenir" under reference ANR-11-IDEX-0004-02. The authors also acknowledge Compagnia di San Paolo for financial support (Progetti di Ricerca di Ateneo-Compagnia di San Paolo-2011-Linea 1A, progetto ORTO11RRT5). Improvements of the CRYSTAL09 code in its MPP version was made possible thanks to the PRACE proposal no. 2011050810 and to the CINECA Award No. HP10BLSOR4-2012. The authors would like to express special thanks to the anonymous referees, for thoroughly reading the manuscript and providing useful and sound comments.

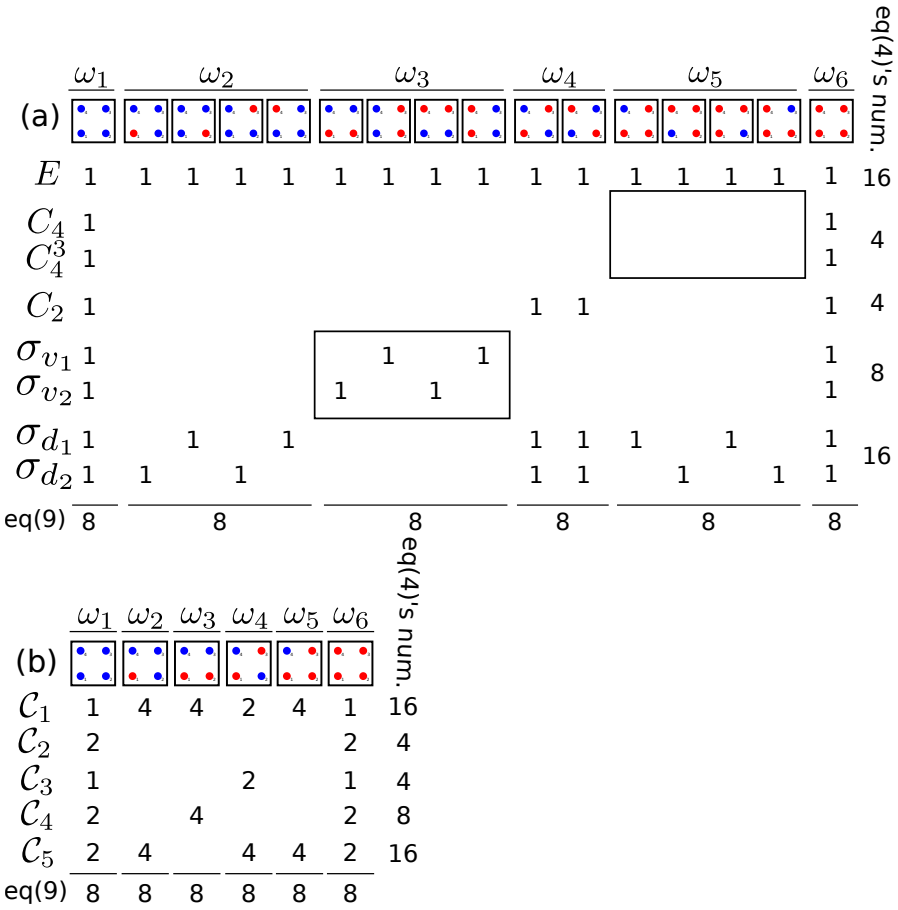


**Figure 1.** Action matrices for the symmetry group $C_{4v}$ (8 symmetry operators) acting on a set of 4 equivalent positions mapped on two colors. (a) and (b) correspond to the action matrix and the contracted action matrix, respectively. In (a) the 16 possible configurations grouped by symmetry classes or orbits are sketched on the top line, orbits being labeled as $\omega_l$. Symmetry operators are listed in the first column. When a configuration is stabilized by an operator, the corresponding entry is set to 1, otherwise it is set to 0 (not reported here, for sake of clarity). The bottom line and the rightmost column illustrate Eq. (9) and the numerator of Eq. (4), respectively. Two matrix blocks are outlined, as defined in the text: $A(2,5)$ and $A(4,3)$. In (b), the top line and the first column correspond to orbits and conjugacy classes, respectively; a representative is shown for each orbit.

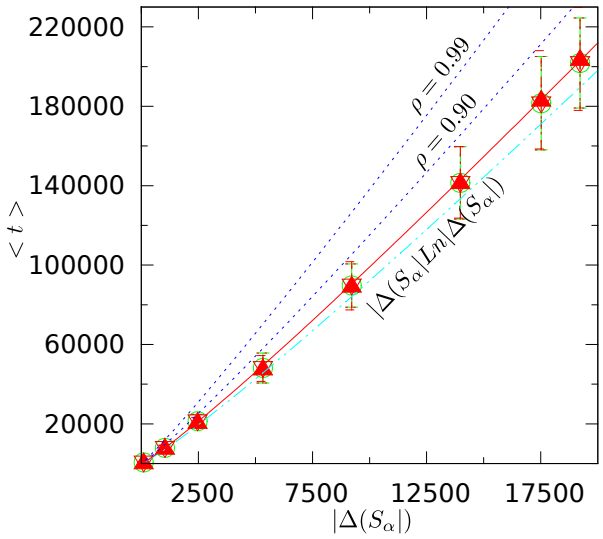


**Figure 2.** Carbonate case study: average number of tries $\langle t \rangle$ required to find the full set $\Delta(S_\alpha)$ as a function of $|\Delta(S_\alpha)|$. Studied compositions are $\alpha = \frac{4}{24}, \frac{6}{24}, \frac{7}{24}, \frac{8}{24}, \frac{9}{24}, \frac{10}{24}, \frac{11}{24}, \frac{12}{24}$. For each composition, three sets of 100 (circle), 500 (up triangle) and 1000 (down triangle) runs have been performed independently. Bars represent the standard deviation. Additional lines are drawn for sake of reference: two dotted lines according to Eq. 31 with $\rho$ equal to 0.90 and 0.99; a dash-dotted line according to Eq. 32 with $\epsilon(|\Delta(S)|) = 0$.

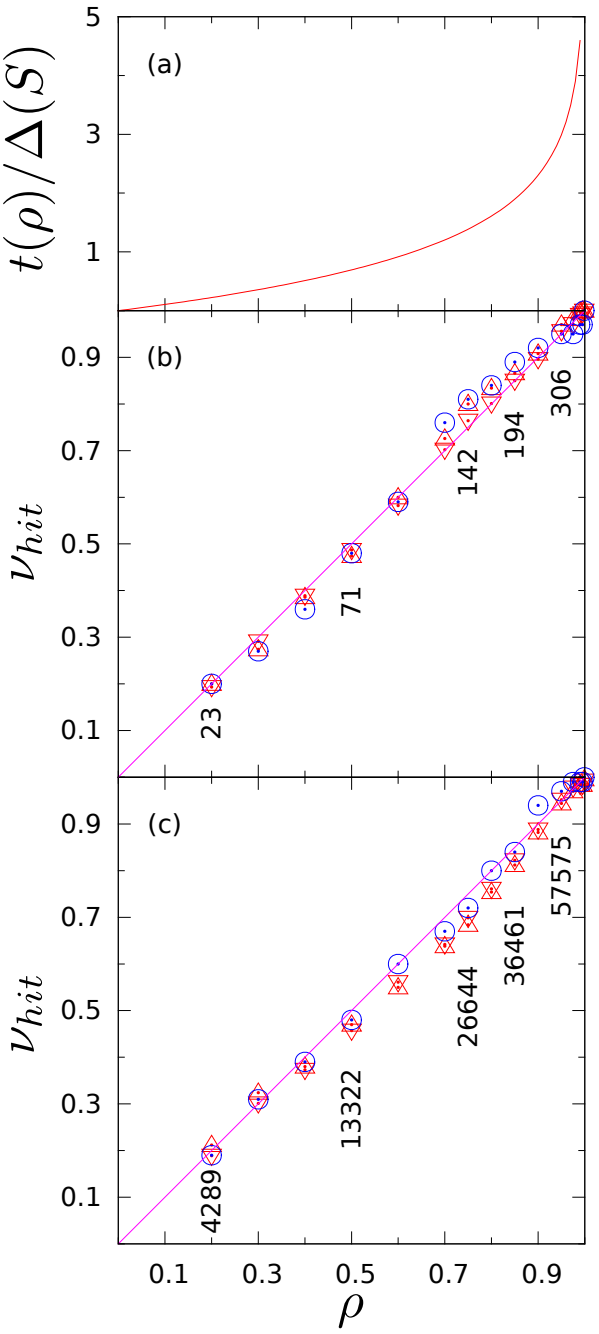


**Figure 3.** Carbonate case study: (a) Theoretical number of tries $t(\rho)$, normalized by $\Delta(S)$, required to find a predefined SIC as a function of the given probability $\rho$ (see Eq. (26)). (b) Frequency $\nu_{hit}$ of tries hitting a predefined SIC at composition 1/6 among $n$ ($n$=100, 500, 1000) tests composed by $t(\rho)$ tries each; the number of SICs $|\Delta(S_\alpha)|$ is 102. (c) Same than (b) in the case of composition 1/2 ($|\Delta(S_\alpha)| = 19219$). Some $t(\rho)$ values are reported in the plots. For symbols, see caption to Figure 2.

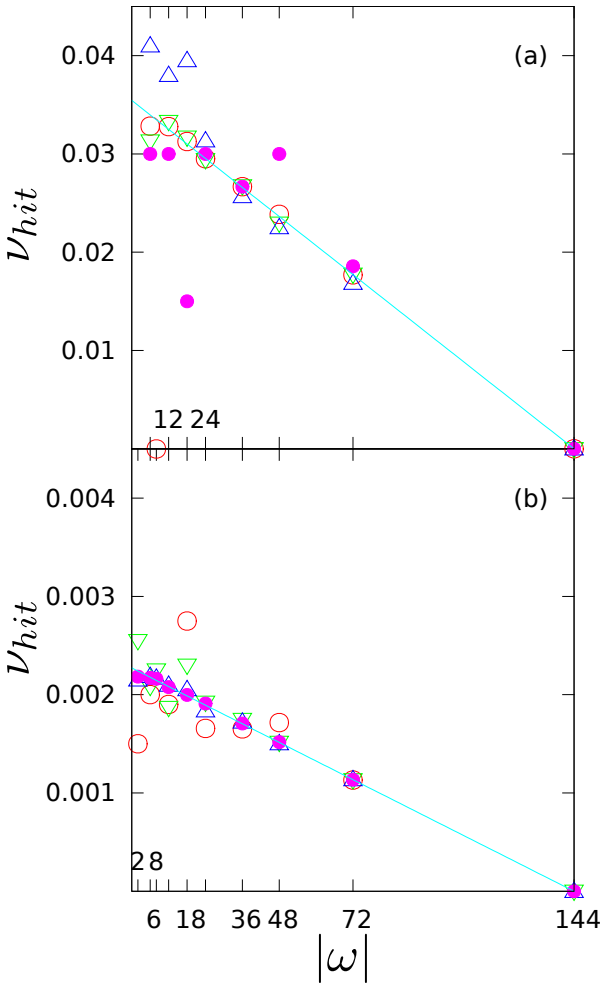


**Figure 4.** Carbonate case study: frequency of orbit finding $\nu_{hit}$ as a function of orbit length $|\omega|$. (a) Chosen composition is 1/6; circle, green down triangle, blue up triangle and dot correspond to 100, 1000, 10000 and 10000 tries, respectively. (b) Chosen composition is 1/2; circle, green down triangle, blue up triangle and dot correspond to 2000, 20000, 200000 and 200000 tries, respectively. Absciss ticklabel are indicated outside and inside to improve the readibility.

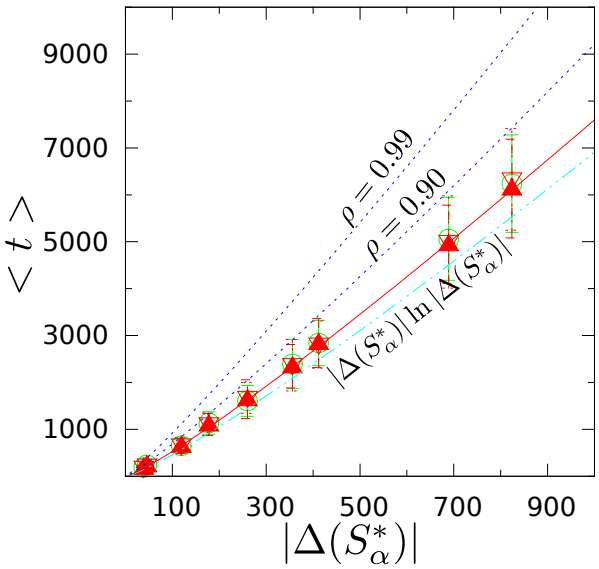


**Figure 5.** Carbonate case study: average number of tries $\langle t \rangle$ required to find the full subset $\Delta(S^*_\alpha)$ of symmetric SICs as a function of $|\Delta(S^*_\alpha)|$. For studied compositions and other details, see caption to Figure 2.

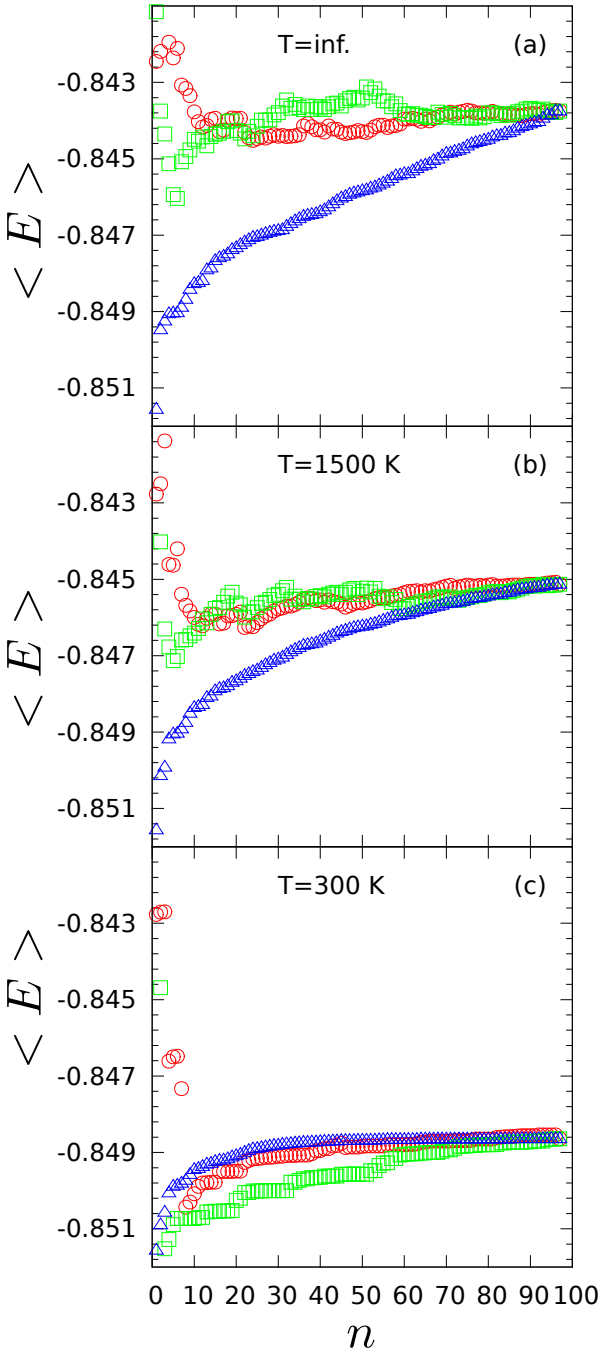


**Figure 6.** Spinel case study: Boltzmann average energy $\langle E \rangle$, in a.u., as a function of the number of considered configurations $n$. (a), (b) and (c) correspond to the infinite temperature limit, 1500 K and 300 K, respectively. For each temperature, the Boltzmann average energy has been calculated according to three procedures over the full set of configurations: circle: MC sampling; triangle: limit average; square: symmetry-enhanced sampling. For more details, see Section 5.2.

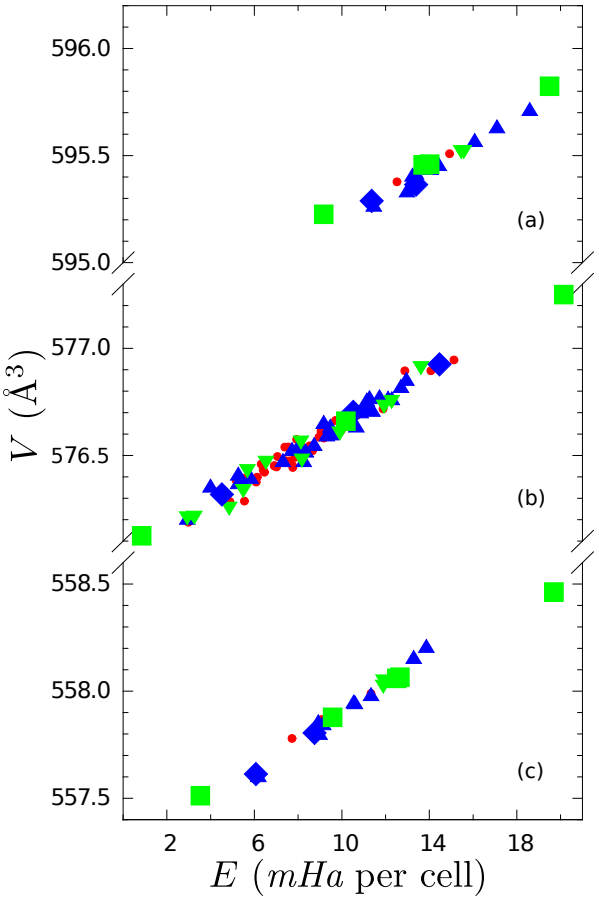


**Figure 7.** Spinel case study: volume $V$ of the SIC representative as a function of energy $E$ in the case of three studied compositions $x$: (a) 0.25; (b) 0.50; (c) 0.75. Energy has been shifted by some quantity depending on the composition. Different symbols relate to different SIC multiplicities: red dot: 192; blue up triangle: 96; green down triangle: 48; blue diamond: 24; green square: 12 and 6.

## References


[1] R. Dovesi, B. Civalleri, R. Orlando, C. Roetti, and V. R. Saunders. *Ab initio* Quantum Simulation in Solid State Chemistry. volume 21 of *Reviews in Computational Chemistry*, pages 1–125. John Wiley and Sons, 2005.

[2] P. Ugliengo, M. Sodupe, F. Musso, I. J. Bush, R. Orlando, and R. Dovesi. Realistic Models of Hydroxylated Amorphous Silica Surfaces and MCM-41 Mesoporous Material Simulated by Large-scale Periodic B3LYP Calculations. *Adv. Mater.*, 20:4579–4583, 2008.

[3] J. VandeVondele, U. Borštnik, and J. Hutter. Linear Scaling Self-Consistent Field Calculations with Millions of Atoms in the Condensed Phase. *J. Chem. Theory Comput.*, 8:3565–3573, 2012.

[4] J. M. Sanchez, F. Ducastelle, and D. Gratias. Generalized cluster description of multicomponent systems. *Physica A*, 128:334–350, 1984.

[5] D. B. Laks, L. G. Ferreira, S. Froyen, and A. Zunger. Efficient cluster expansion for substitutional systems. *Phys. Rev. B*, 46:12587–12605, 1992.

[6] A. Zunger, L. G. Wang, G. L. W. Hart, and M. Sanati. Obtaining Ising-like expressions for binary alloys from first principles. *Modelling Simul. Mater. Sci. Eng.*, 10:685–706, 2002.

[7] A. van de Walle and G. Ceder. Automating First-principles phase diagram calculations *J. Phase Equilib.*, 23:348–359, 2001.

[8] V. Blum, G. L. W. Hart, M. J. Walorski, and A. Zunger. Using genetic algorithm to map first principles results to model Hamiltonians : Application to the generalized Ising model for alloys. *Phys. Rev. B*, 72:165113–165125, 2005.

[9] T. Mueller and G. Ceder. Bayesian approach to cluster expansions. *Phys. Rev. B*, 80:24103–24115, 2009.

[10] S. Mustapha, Ph. D'Arco, M. De La Pierre, Y. Noel, M. Ferrabone, and R. Dovesi. On the use of symmetry in configurational analysis for the simulation of disordered solids. *J. Phys.: Condens. Matter*, 25:105401, 2013.

[11] L. G. Ferreira, S.-H. Wei, and A. Zunger. Stability, Electronic Structure, and Phase Diagrams of Novel Inter-Semiconductor Compounds. *Int. J. High Perform. Comput. Appl.*, 5:34–56, 1991.

[12] J. M. Sanchez and D. de Fontaine. The fcc Ising model in the cluster variation approximation. *Phys. Rev. B*, 17:2926–2936, 1978.

[13] R. Magri, J. E. Bernard, and A. Zunger. Predicting structural energies of atomic lattices. *Phys. Rev. B*, 43:1593–1597, 1991.

[14] G. L. W. Hart and R. W. Forcade. Algorithm for generating derivative structures. *Phys. Rev. B*, 77:224115, 2008.

[15] G. L. W. Hart and R. W. Forcade. Generating derivative structures from multilattices: Algorithm and application to hcp alloys. *Phys. Rev. B*, 80:014120, 2009.

[16] R. Grau-Crespo, S. Hamad, C.R.A. Catlow, and N. H. de Leeuw. Symmetry-adapted configurational modelling of fractional site occupancy in solids. *J. Phys.: Condens. Matter*, 19:256201, 2007.

[17] Q. Wang, R. Grau-Crespo, and N.H. de Leeuw. Mixing Thermodynamics of the Calcite-Structured (Mn,Ca)CO3 Solid Solution: A Computer Simulation Study. *J. Phys. Chem. B*, 115:13854–13861, 2011.

[18] S. Haider, R. Grau-Crespo, A. J. Devey, and N. H. de Leeuw. Cation distribution and mixing thermodynamics in Fe/Ni thiospinels. *Geochim. Cosmochim. Acta.*, 88:275–282, 2012.

[19] M. Habgood, R. Grau-Crespo, and S. L. Price. Substitutional and orientational disorder in organic crystals: a symmetry-adapted ensemble model. *Phys. Chem. Chem. Phys.*, 13:9590–9600, 2011.

[20] V. L. Vinograd, M. H. F. Sluiter, and B. Winkler. Subsolidus phase relations in the $CaCO_3$-$MgCO_3$ system predicted from the excess enthalpies of supercell structures with single and double defects. *Phys. Rev. B*, 79:104201–104209, 2009.

[21] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, and A. H. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092, 1953.

[22] V. L. Vinograd, B. P. Burton, J. D. Gale, N. L. Allan, and B. Winkler. Activity-composition relations in the system $CaCO_3$-$MgCO_3$ predicted from static structure energy calculations and Monte Carlo simulations. *Geochim. Cosmochim. Acta.*, 71:974–983, 2007.

[23] C. J. Pickard and R. J. Needs. *Ab initio* random structure searching. *J. Phys.: Condens. Matter*, 23:53201–53223, 2013.

[24] R. Dovesi, V. R. Saunders, C. Roetti, R. Orlando, C. M. Zicovich-Wilson, F. Pascale, B. Civalleri, K. Doll, N. M. Harrison, I. J. Bush, Ph. D'Arco, and M. Llunell. *CRYSTAL 2009 User's Manual*. University of Torino, Torino, 2009.

[25] G. L. W. Hart, L. J. Nelson, and R. W. Forcade. Generating derivative structures at a fixed concentration. *Comp. Mater. Sci.*, 59:101–107, 2012.

[26] J. D. Dixon and H. S. Wilf. The random selection of unlabeled graphs. *J. Algorithms*, 4:205–213, 1983.

[27] S. G. Williamson. *Combinatorics for computer science*. Computer Science Press, New York, 1985.

[28] A. Kerber, R. Laue, R. Hager, and W. Weber. Cataloging graphs by generating them uniformly at random. *J. Graph Theory*, 14:559–563, 1990.

[29] A. Kerber. *Applied finite group action*. Springer-Verlag, New York, 1999.

[30] A. Nakatsuka, H. Ueno, N. Nakayama, T. Mizota, and H. Maekawa. Single-crystal X-ray diffraction study of cation distribution in $MgAl_2O_4$-$MgFe_2O_4$ spinel solid solution. *Phys. Chem. Miner.*, 31:278–287, 2004.

[31] A. D. Becke. Density functional theochemistry. III The role of exact exchange. *J. Chem. Phys.*, 98:5648–5652, 1993.

[32] L. Pauling. The principles determining the structure of complex ionic crystals. *J. Am. Chem. Soc.*, 51:1010–1026, 1929.

[33] D. J. Wales. Symmetry, near-symmetry and energetics. *Chem. Phys. Lett.*, 285:330–336, 1998.

[34] D. J. Wales. Erratum to "Symmetry, near-symmetry and energetics". *Chem. Phys. Lett.*, 294:262, 1998.