Modelling Simul. Mater. Sci. Eng. **17** (2009) 055003 (19pp) doi:10.1088/0965-0393/17/5/055003

UNCLE: a code for constructing cluster expansions for arbitrary lattices with minimal user-input

D Lerch[1,2], O Wieckhorst[1], G L W Hart[3], R W Forcade[4] and S Müller[1,2]

[1] Lehrstuhl für Festkörperphysik, Universität Erlangen-Nürnberg, Staudtstrasse 7, D-91058 Erlangen, Germany
[2] Lehrstuhl für Theoretische Physik 2, Universität Erlangen-Nürnberg, Staudtstrasse 7, D-91058 Erlangen, Germany
[3] Department of Physics & Astronomy, Brigham Young University, Provo UT 84602, USA
[4] Department of Mathematics, Brigham Young University, Provo UT 84602, USA

E-mail: stefan.mueller@physik.uni-erlangen.de

Received 19 December 2008, in final form 5 May 2009
Published 4 June 2009
Online at stacks.iop.org/MSMSE/17/055003

Abstract
We present a new implementation of the cluster expansion formalism. The new code, UNiversal CLuster Expansion (UNCLE), consolidates recent advances in the methodology and leverages one new development in the formalism itself. As a core goal, the package reduces the need for user intervention, automating the method to reduce human error and judgment. The package extends standard cluster expansion formalism to the more complicated cases of ternary compounds, as well as surfaces, including adsorption and inequivalent sites.

(Some figures in this article are in colour only in the electronic version)

1. Introduction

One significant class of problems in materials science comprises those that can be characterized, at the atomic level, as *configurational*. Important configurational questions include order–disorder transitions, ground states (stable compounds in the solid part of the phase diagrams of crystalline metals and semiconductors), segregation of defects in alloy interfaces and surfaces, or adsorption at surfaces. The sheer size of configurational space makes it impractical to explore any of these questions directly from an electronic structure theory (first principles methods). But it is possible to map first-principles results onto a faster Hamiltonian. One example is the cluster expansion [1], which allows one to 'extract' the physics of atom–atom interactions from a set of first-principles, small-unit-cell calculations and then model, in a practical way, unit cells with millions of atoms (as in Monte Carlo modeling), or explore tens of million of different atomic configurations in different cell shapes (as in direct enumeration ground state searches).

Detailed discussions of the cluster expansion methodology can be found elsewhere [1–8]. Despite the usefulness of the method, its application has largely been restricted to specialists. The release of the Alloy Theoretic Automated Toolkit [9, 10, 11] partially remedies this situation. Our purpose for this paper is to discuss advances in the method, its formalism, and in its implementation that (1) simplify its use by non-specialists, (2) reduce the dependence on the user input as far as possible and (3) make its application more general. Though some of the discussion may seem banal to specialists, our aim is to present our approach to the non-specialist as well.

The concept of the cluster expansion is to express the configurational dependence of the energy (or some other observable) as a sum over bonds. A particular configuration (which kind of atom is sitting at each site) is given by a vector, σ, each component of which corresponds to one site. The energy depends on the configuration, $E(\sigma)$, and can be formally decomposed [1] into a sum over bonds (not just pairs, but triplets, quadruplets, etc). In the cluster expansion community, the bonds are referred to as *clusters* or *figures*, denoted F below.

Formally, the idea is expressed as

$$E(\sigma) = N \sum_F d_F J_F \overline{\Pi}_F(\sigma), \tag{1}$$

where each term in the sum is the product of the local structure $\overline{\Pi}_F$ and an expansion coefficient J_F (an effective cluster interaction, ECI). The factor d_F corresponds to the number of degeneracies of the figure-class F normalized to the number of sites N within the configuration σ. The configurational dependence of any observable can be represented *exactly* by the cluster expansion [1], but this is true only if every term is retained, which is impractical. In practice, however, drastically truncated expansions yield essentially the same answers, making the method practical and thus useful in a variety of contexts.

Though the formalism of the cluster expansion is conceptually simple, constructing an expansion that makes reliable predictions is not trivial. Potentially, there are many choices that the user faces. For example, which terms should be retained in a truncated cluster expansion? Which input structures should be used in assessing the reliability of the cluster expansion? Which structures should be chosen and calculated via first-principles to determine the ECIs via fitting? The predictions of the cluster expansion should be *independent* of choices made by the user, but since the predictions can depend sensitively on these choices, one desires an approach to constructing a cluster expansion which eliminates this 'user sensitivity'.

2. UNiversal CLuster-Expansion

In order to minimize the sensitivity of the cluster expansion to user choices, and also to make the cluster expansion applicable beyond simple bulk binary systems, we have developed a new program-package under the name of '*UN*iversal *CL*uster *E*xpansion' (UNCLE). An implementation that meets our objectives should

- be applicable to as many systems as possible, i.e. arbitrary lattices and multi-component systems.
- Require as little user input as possible in the construction of a reliable set of ECIs; i.e. it should automatically construct and select the figures, and the user input required for choosing input structures or scanning possible configuration spaces should be minimized.
- Contain efficient tools for the use of the ECIs, such as in kinetic or thermodynamic Monte Carlo simulation or ground state searches.

Figure 1. Illustration of the decomposition of a simple cubic lattice into the figures used in the CE. The sum has been truncated with respect to the number of vertices *and* possible shapes and orientations of figures with the same number of vertices.

- Reduce the burden on the user for generating input data, i.e. simplify the interaction between the cluster expansion and common DFT codes. UNCLE already provides an automatic communication with the codes VASP [12–16] and FLAIR [17–19].

For some of these objectives, suitable methods have already been developed by us and other researchers. The aim of this project was to combine them in a *single*, uniform program package and to make the code applicable to a broad variety of physical problems, such as the treatment of multi-component bulk alloys or semiconductors, multi-component surfaces and adsorbate-systems, as well as systems that include the coupled physics of both adsorbates and alloy surfaces. In the following subsections we present the detailed functionality of UNCLE and extensions of the cluster expansion formalism.

2.1. Definition of lattice and figures

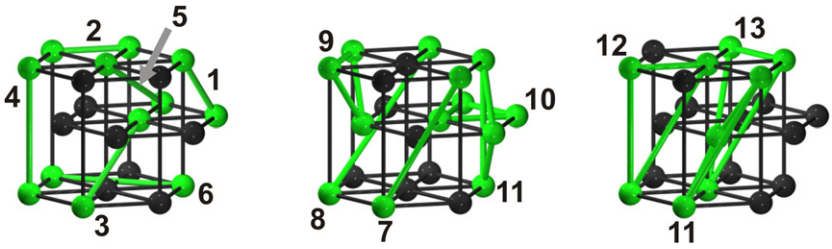
Up to now, typical fields of applications for cluster expansions are the modeling of special bulk phases in binary metal alloys, as e.g. precipitation (e.g. [20–22]), surface segregation [23–25], or adsorption on surfaces [26, 27]. The basic element of every such cluster expansion is the underlying lattice, the 'parent' lattice. With the general approach of UNCLE, the range of applications goes beyond *simple* lattices, applying equally well to cases where there is more than one site per unit cell. We refer to such cases as *multilattices*. Physical examples include hcp-based alloys (2 sites/cell), reconstructed surfaces, and adsorbate systems with inequivalent adsorption sites. In principle, the approach is applicable even to systems of lower dimensionality, such as free-standing clusters[5].

In simple terms, the cluster expansion is an expansion over *figures* (i.e. the 'clusters' or bonds in a system) where each figure has an associated expansion coefficient (see equation (1)). The decomposition of a lattice into figures is illustrated in figure 1. The first step in the UNCLE approach is to define the pool of figures. Although the number of symmetrically distinct figures is limited by the input structures, even this list may contain far more figures than is required to form an accurate cluster expansion. Our approach is to choose a 'cutoff radius,' taking all figures smaller than the cutoff. Because the importance of figures generally decreases as the number of vertices increases, we specify smaller cutoffs for figures with more vertices[6]. This approach requires little input from the user and the user's choice can easily be checked. The user specifies a cutoff radius, $R^{(k)}$, for each number of vertices k. The figure pool is then determined as follows:

(i) Find the set of all lattice points $\{P\}$ within a sphere with radius $r = R^{(k)}$ around the lattice's origin.
(ii) Construct all k-tuples out of this set which include the origin. Each k-tuple corresponds to one figure f.

[5] Though the computational overhead of such systems can be prohibitive due to the low symmetry, resulting in an expansion with many terms.
[6] This intuitive but arbitrary choice has been used for years among CE practioners. Reference [33] was the first work to formally justify the choice.

Table 1. Pair and triplet figure classes found for $R^{(2)} = 1.7$ and $R^{(3)} = 1.1$ on the hcp lattice. k is the total number of vertices of the figure, d corresponds to the figure's degeneracy (i.e. the number of symmetry equivalent representatives of each figure) and $\boldsymbol{r}_F$ is the average vertex distance relative to the 'center of mass' (see section 2.6). The vertices of one representative of each figure are given in Cartesian coordinates.



#	k	d_F	$\boldsymbol{r}_F$	Vertex #1	Vertex #2	Vertex #3
1	2	6	0.500	(0.0,0.0,0.0)	(0.5,0.289,0.815)	—
2	2	6	0.500	(0.0,0.0,0.0)	(1.0,0.0,0.0)	—
3	2	6	0.707	(0.0,0.0,0.0)	(1.0,–0.577,0.815)	—
4	2	2	0.815	(0.0,0.0,0.0)	(0.0,0.0,1.63)	—
5	2	12	0.866	(0.0,0.0,0.0)	(1.0,1.155,0.815)	—
6	2	6	0.866	(0.0,0.0,0.0)	(1.5,0.866,0.0)	—
7	2	12	0.956	(0.0,0.0,0.0)	(1.0,0.0,1.63)	—
8	2	12	1.189	(0.0,0.0,0.0)	(1.5,0.866,1.63)	—
9	3	12	0.577	(0.0,0.0,0.0)	(0.5,0.289,–0.815)	(0.5,0.866,0.0)
10	3	2	0.577	(0.0,0.0,0.0)	(0.5,0.866,0.0)	(1.0,0.0,0.0)
11	3	6	0.687	(0.0,0.0,0.0)	(0.0,0.0,1.63)	(0.5,0.289,0.815)
12	3	12	0.705	(0.0,0.0,0.0)	(0.5,0.289,0.815)	(1.0,0.0,1.63)
13	3	24	0.878	(0.0,0.0,0.0)	(0.0,0.0,1.63)	(1.0,0.0,1.63)
14	3	12	0.939	(0.0,0.0,0.0)	(0.5,0.866,1.63)	(1.0,0.0,1.63)

(iii) Reduce each figure class $F = \{f | f = \hat{R}_i \mathrm{f}\}$ to a single representative f by eliminating all other symmetry-equivalent representatives of f. In the case of a binary system there will be one ECI for each symmetrically distinct figure class F.

This is done up to a maximum number of vertices $2 \leqslant k \leqslant k_{\max}$. As an example of the construction of figures for a multilattice, table 1 presents a set of pair and triplet figures for the hcp lattice (using $R^{(2)} = 1.7$ a_0 and $R^{(3)} = 1.1$ a_0).

While the parent lattice is a basic element that is determined by the system of study, the choice of $k_{\max}$ and $R^{(k)}$ is arbitrary and needs to be checked as illustrated in section 2.5. In practice, the check is simple; increase both values to the point where the predictive value of the expansion stops improving. Regarding the last point, it should be mentioned that a complete automatization of this step can easily lead to a waste of computer time. Because the check is simple, it is left to the user.

2.2. Checking the input database

The coefficients of the cluster expansion (i.e. the ECIs) are determined by fitting to an input database. This input database consists of a set of atomic configurations, whose energy has been

determined using *ab initio* methods. An efficient cluster expansion method will facilitate the exchange of structural information between the fitting routines and the first-principles code. This decreases the amount of user time required and reduces the chances for human error[7].

Currently, UNCLE's format for structural information has been designed to match that of the first-principles code VASP [12–16] and to adapt to the input of the FLAPW code FLAIR [17–19]. It should be mentioned, though, that the source of the input values in the database can be arbitrary, and do not necessarily have to originate from first-principles calculations. For every input value in the database, the corresponding structural information is given as follows: real-space coordinates of the supercell B, the number of each chemical atomic species in the cell, and positions of the basis atoms within the superstructure. The latter is given either in direct or Cartesian coordinates. Following the structural information, the corresponding value of the observable to be expanded is given.

After the input structures have been read in, the validity of the input is checked. UNCLE checks whether all their basis atoms lie on the lattice and whether there are symmetry-equivalent structures within the input list. As trivial as this step may seem, in practice this becomes an extremely useful feature; converged cluster expansions typically require around 50–150 input structures, which tend to contain subsets of similar, though symmetrically distinct, atomic configurations. Unintentional duplication of input structures can easily occur without this check, which not only wastes calculation time, but also falsely over weights the structure during the fitting.

2.3. Choice of input structures

The choice of atomic configurations, from which the ECIs are extracted, affects the ECIs. To avoid biasing the input database, and thus the ECIs, we systematically increase the database. We begin with a hand-chosen set $\{\sigma\}$ of usual suspects, small-unit-cell structures derived from the parent lattice (following the recipe of [28]), and some quasi-random structures. The first cluster expansion determined from this initial set makes predictions, perhaps not accurately, for the ground states and other structures with enthalpies of formation near the convex hull. These predictions are in turn calculated via first principles and added to the input database. This cycle is repeated, as shown in figure 2, letting the current cluster expansion itself pick new structures to add to the database.

2.4. Calculation of correlation functions

We apply a new mathematical formalism to the cluster expansion that considerably simplifies aspects of the cluster expansion implementation. Two places where this is particularly useful is in calculating the correlations (needed to perform the sum in equation (1)) and in Monte Carlo simulations. The new formalism works in the 'space' of 3×3 *integer* matrices and provides an alternative representation for structures and figures.

Any supercell of the parent lattice is an integer multiple of the parent cell. So if the vectors of the parent lattice are the column vectors of a matrix A, there exists a matrix N, with all integer elements, such that $B = AN$. The columns of B are the lattice vectors of the supercell and the determinant of N will be the multiplicative factor; that is, if the supercell has twice the volume of the parent cell, then $|N| = 2$.

[7] This idea is exemplified in van de Walle's excellent package, Alloy Theoretic Automated Toolkit (ATAT) [9–11], which runs practically automatically. We prefer our package for a number of other reasons. It allows more flexibility in choosing the clusters (but it is still automatic), it offers a more efficient approach to cluster expansions for surfaces, it can run much larger MC simulations, etc.

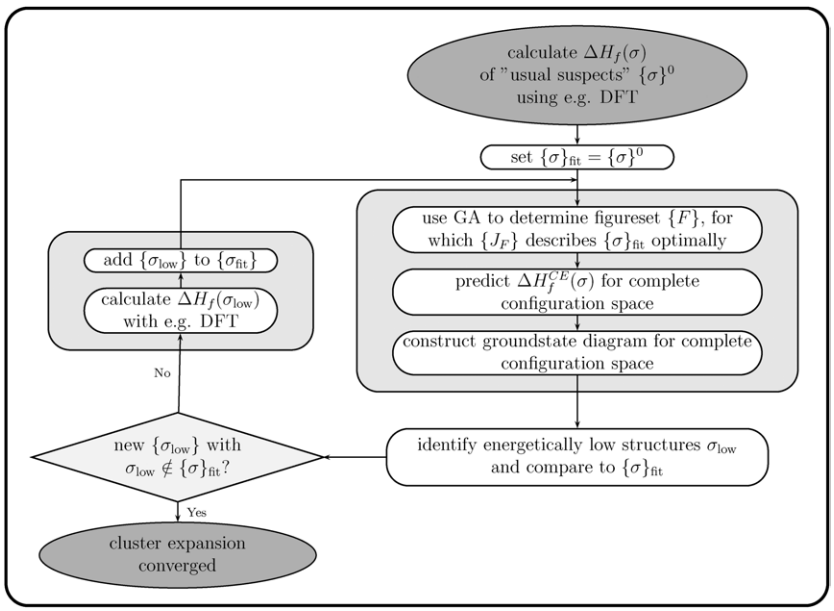


Figure 2. Illustration of the self-consistent 'outer loop', which chooses the input structures of the cluster-expansion.

Because $B = AN$, the integer matrix N is an alternative representation for the superlattice. Realizing this, we can then map the superlattice and its atomic sites to this alternate representation, the g-representation, where the calculation of correlations is greatly simplified. In the g-representation, the atomic sites lie on an integer lattice, $\mathbb{Z}^3$, and the shape of the supercell is always orthorhombic. This simplifies the algorithm and thus makes the code much more efficient, both in time and memory.

Mapping to the g-representation is accomplished by reducing N into its Smith normal form (SNF). The SNF is a diagonal form with property that each diagonal divides the next one down. Specifically, a 3×3 matrix in SNF looks like

$$\begin{pmatrix} D_{1,1} & 0 & 0 \\ 0 & D_{2,2} & 0 \\ 0 & 0 & D_{3,3} \end{pmatrix},$$

where $D_{3,3}$ is divisible by $D_{2,2}$ which is in turn divisible by $D_{1,1}$, i.e. $D_{1,1}|D_{2,2}|D_{3,3}$. The decomposition to SNF, D, is accomplished by two unimodular transformation matrices with integer entries: $LNR = D$.

To illustrate the idea, figure 3 shows two superlattices of the hexagonal lattice. The matrices N depict how the parent lattice is multiplied to yield superlattices. In this example, both superlattices are four times larger than the parent cell; $|N| = 4$. Then, these superlattices are reduced to their corresponding SNFs D.

The transformation matrix L provides a map for the new basis (g-representation) in which the superlattice vectors are orthogonal and the parent lattice is a simple cubic lattice. In this representation, each vector $\boldsymbol{g} \in \mathbb{N}^3$ with $0 \leqslant g_i < D_{i,i}$ represents a lattice point *within* the supercell. The occupation (atom type), s_i, of each atomic site in the supercell can thus be stored in an integer table G, whose extent in each dimension is one of the diagonal elements of D. Because each of the basis atoms corresponds to exactly one element in G, G is a minimal

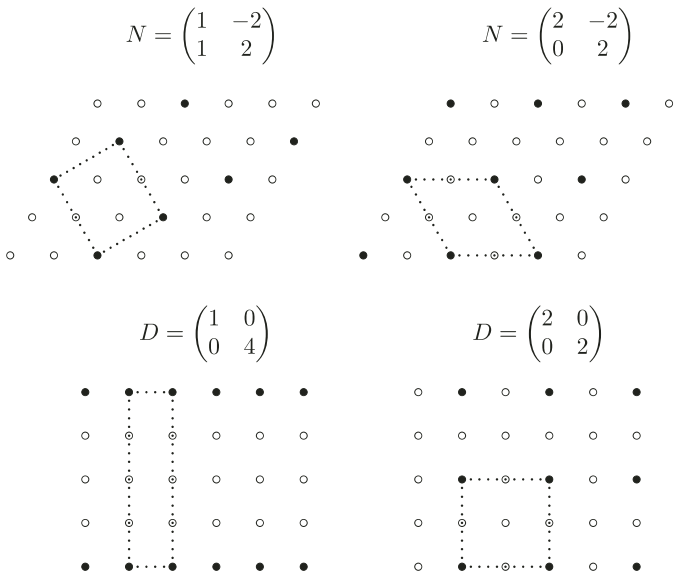


Figure 3. Two superlattices of the hexagonal lattice and their reductions to SNF. The superlattice vectors are defined via $B = AN$, where A is the matrix of parent lattice vectors (as columns), N is the transformation matrix and B is the matrix of superlattice vectors. If B is reduced to its corresponding SNF D, then new representation of the superlattice is a superlattice of a square parent lattice.

hash table[8] for the occupation of the supercell sites, and any vector $\boldsymbol{g}$ is a perfect hash function into the table. (For a more in-depth discussion of the group-theoretic approach to representing superstructures, see [28, 29]).

This mapping from a real-space representation to a simple cubic integer lattice representation is the key for the efficient computation of correlations. The lattice vectors and lattice points are represented by integers rather than floating point variables. No logic statements in the loops are required; no comparison of floating point numbers are needed. This improves both the efficiency and the robustness of the implementation. Explicitly, to calculate the correlations, one needs to calculate the product of occupations (atom types), s_i, over every vertex of a figure and then sum over all symmetry-equivalent figures. In this g-representation, a figure comprises k integer vectors $\boldsymbol{g}$ which are simply locations in the G table. One simply loops over the k locations in the G table, forming the product, and then adds these products together for each representative of the figure. In other words, for a given figure,

$$\Pi_f = \prod_{\text{vertices } j} G(\boldsymbol{g}^{f,j}). \tag{2}$$

The averaged correlation function $\overline{\Pi}_F$, as it is used in the cluster-expansion Hamiltonian, equation (1), is defined as the expectation value of Π_f over all symmetry-equivalent figures, all $f \in F$:

$$\overline{\Pi}_F = \frac{1}{d_F} \sum_{f \in F} \Pi_f. \tag{3}$$

[8] A hash table and its associated hash function constitute a data structure in elementary computer science. It associates keys with values. Its advantage is that it allows very efficient lookup. In our case, the hash table is minimal (no empty locations), and the hash function is perfect (keys and values have a one-to-one correspondence).

This general concept applies equally well for the computation of correlations on multilattices. As a multilattice of n lattice points per unit cell is an n-fold copy of the lattice, each shifted relative to the others, we expand the dimensionality of our problem from $\mathbb{N}^3$ to $\mathbb{N}^4$, where the fourth integer dimension denotes to which sublattice of the multilattice a given vector $\boldsymbol{g}$ belongs. Once the figures of a multilattice have been defined in $\mathbb{N}^4$, the correlations and CE-sum is treated in the same way as for a *simple* parent lattice.

The necessity of allowing for parent lattices which are multilattices can be seen in multiple contexts: hcp alloys (parent lattice is a 2-lattice with crystallographically equivalent sites), surface alloys [23, 24, 25], adsorbate systems with crystallographically inequivalent adsorption sites [27], or surface adsorbing systems with a multi-component substrate. In the example of an alloy surface, the loss of translational symmetry in the third dimension leads to layer-dependent ECIs, which increases their number tremendously. The same holds for the case of an adsorbate-system where more than one site is occupied: in general, the occupation of each symmetry-inequivalent site will be energetically different. These must be included in the CE by site-dependent figures, resulting in site-dependent ECIs.

One simple example of a multilattice situation is the hexagonal closed-packed lattice, which has two sites per unit cell. Table 2 presents a few of the correlations of several simple hcp-based structures, which are obtained for the figures shown in table 1.

Our implementation of UNCLE can be generalized to treat multinary systems. The treatment of ternary compounds has already been implemented and used. The extension beyond ternary systems is relatively simple and will be made as soon as required. To handle multinary expansions, the correlations must be calculated over a set of cluster functions. Formally there is also a set of cluster functions for a binary expansion, but there is only one function in the set and it can be taken to be the occupation itself, that is $\theta(s_i) = s_i$.

In the binary case, the correlation is computed merely by taking the product of each occupation value (± 1) over each vertex of a figure:

$$\Pi = \prod_{i=1}^{k} s_i \tag{4}$$

and there is one ECI (that is, one J) for each figure. But in the case of a n-ary system (n-components represented by n spin values), the complete description of the correlations requires $(n-1)$ cluster functions θ_l. Therefore, a figure with k vertices is no longer connected with a single correlation function, but instead $(n-1)^k$ correlation functions $\Pi^{(j)}$. The ith entry of the superscript vector (j), which contains k entries, defines the cluster function θ_l, which is to be applied to the i^{th} vertex of the figure

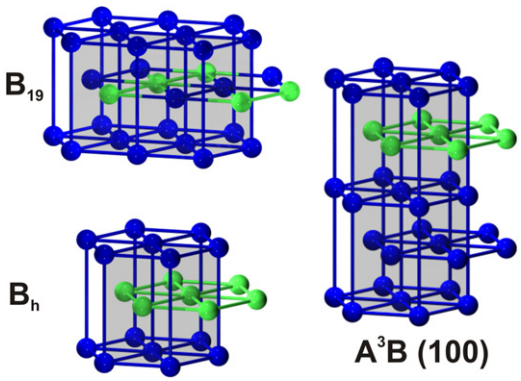
$$\Pi^{(j)} = \prod_{i=1}^{k} \theta_{l=(j_i)}(s_i) \tag{5}$$

The full set of correlation functions of a figure consists of all the 2^k possible vectors (j). This number can be reduced according to the symmetry of a figure. The general multinary formalism was laid out by Sanchez *et al* in [1], and later applied by Wolverton and de Fontaine [30]. Along the lines of these papers, we use Chebychev polynomials as cluster functions in the ternary case, an example for their application is given in section 3.1.

2.5. Identifying relevant figures

The cluster expansion approach is exact only when all possible figures are included in the cluster expansion sum, equation (1). But including such an (astronomic) number of terms in

Table 2. Correlations for five different structures based on the hcp lattice: (1) a full occupation of the complete lattice with all $s_i = -1$, (2) a full occupation with all $s_i = +1$, (3) an AB-superlattice in the [1 0 0]-direction (Strukturbericht name B_h), (4) an AAAB stacking of the (0 0 0 1)-planes and (5) the structure with Strukturbericht name B_{19}. In addition to the correlations of the figures in table 1, the 'onsite-correlation' (the average of the occupation over all the sites of a given structure) is shown and denoted as figure # 0.



#	pure "−1"	pure "+1"	B_h	A_3B (0001)	B_{19}
0	−1.000	+1.000	0.000	+0.500	+0.500
1	+1.000	+1.000	−1.000	0.000	0.000
2	+1.000	+1.000	+1.000	+1.000	+0.333
3	+1.000	+1.000	−1.000	0.000	0.000
4	+1.000	+1.000	+1.000	0.000	+1.000
5	+1.000	+1.000	−1.000	0.000	0.000
6	+1.000	+1.000	+1.000	+1.000	+0.333
7	+1.000	+1.000	+1.000	0.000	+0.333
8	+1.000	+1.000	+1.000	0.000	+0.333
9	−1.000	+1.000	0.000	+0.500	−0.166
10	−1.000	+1.000	0.000	+0.500	+0.500
11	−1.000	+1.000	0.000	−0.500	+0.500
12	−1.000	+1.000	0.000	−0.500	−0.166
13	−1.000	+1.000	0.000	+0.500	+0.500
14	−1.000	+1.000	0.000	+0.500	+0.500

the expansion is impractical. To be useful, the expansion must be truncated to a relatively small number of terms without losing the expansion's predictive power. Choosing which figures to retain is the most critical step of the cluster expansion method. Nevertheless, finding a good selection of figures is a formidable task—even when the cutoffs discussed in section 2.1 are relatively short, there may be thousands of figures to choose from. Selecting a few dozen interactions from a pool of thousands is impossible to do exhaustively.

So far an evolutionary approach based on a genetic algorithm (GA) has proven to be the most effective[9] way to choose the figures. The set of figures chosen by the GA results in a

[9] In the cluster expansion community, the problem of truncating the expansion (that is, selecting the figures) has generated debate and several proposed solutions. The two main competing methods are hierarchal approaches (van de Walle [9, 10] and Zarkevich and Johnson [33]) and the Variational Cluster Expansion (Diaz-Ortiz and Drautz [34, 35]). One reason that this question does not have consensus in the community yet is that the community does not have a

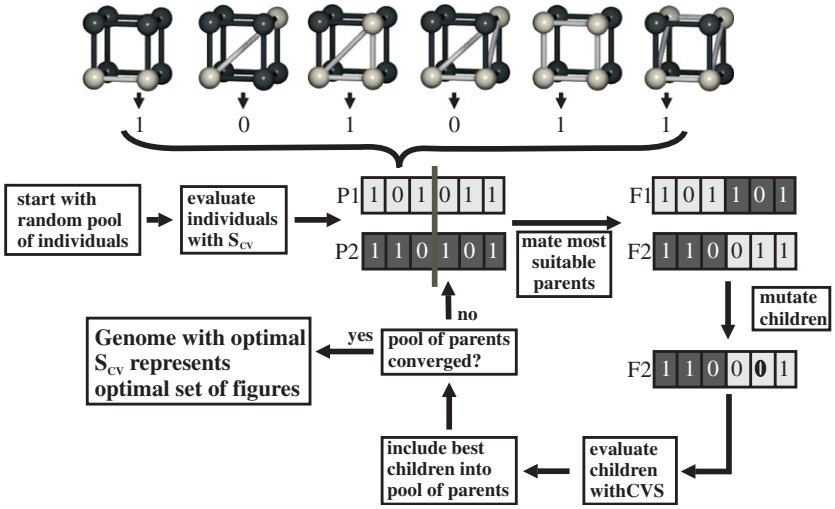


Figure 4. Illustration of the GA, which helps to safely identify the relevant figures that need to be included in the CE-sum.

cluster expansion that has better predictive power than if chosen using other approaches. The details of the algorithm, which is implemented in UNCLE, have been described in [7, 31]. Its basic principle is illustrated in figure 4.

The fitness criterion for the selection of figures is a *leave-many-out* cross-validation-score (see e.g. [9, 32]). This fitness score S_{CV} is a measure of the predictive power for a given choice of figures. Its value is calculated by the following scheme.

(i) Randomly choose $\mathcal{N}$ sets $\{\sigma\}^i_{\mathrm{prediction}}$, $(i \in \{1...\mathcal{N}\})$ of n different structures out of the total pool of input structures.

(ii) For each of the $\mathcal{N}$ prediction-sets $\{\sigma\}^i_{\mathrm{prediction}}$, perform a cluster expansion based on all input structures except for those contained in $\{\sigma\}^i_{\mathrm{prediction}}$. The resulting ECIs are not influenced by the energetics of $\{\sigma\}^i_{\mathrm{prediction}}$.

(iii) Use the resulting ECIs to predict the energy of every member of $\{\sigma\}^i_{\mathrm{prediction}}$ and compare it with the energy calculated by density functional theory.

(iv) Calculate the expectation value of the root-mean-square error for all the predicted structures:

$$S_{\mathrm{CV}} = \sqrt{\frac{1}{\mathcal{N}\cdot n}\sum_{\mathcal{N}}\sum_{n}|E_{\mathrm{DFT}}(\sigma) - E_{\mathrm{CE}}(\sigma)|^2}. \tag{6}$$

Other successful applications of the GA within a cluster expansion can be found, e.g. in [25]. For a large pool of figures, the GA has already been compared with the hierarchal method proposed in [33] and the variational cluster expansion [34, 35] and proved to be the most reliable in finding the choice of figures with the best S_{CV}[9].

common metric for determining the robustness of a fit. In our opinion, the most robust way to measure the goodness of a fit, and the method that seems to be most accepted for testing predictive models, is leave-many-out cross validation. We have tested the hierchal and variational approaches and found that *under this metric* our evolutionary approach is more robust.

2.6. Determination of ECIs

For a given choice of figures and a set of $\mathcal{N}$ input structures $\{\sigma\}$, the ECIs J are extracted by minimizing

$$\sum_{\mathcal{N}} \left(E_{\mathrm{DFT}}(\sigma) - \sum_{F} D_F J_F \overline{\Pi}_F(\sigma) \right)^2 + \sum_{F} t_F J_F \stackrel{!}{=} \min , \tag{7}$$

where the last term is a *damping* term, which penalizes figures with large spatial extent (the spatial extent is determined as the average distance of the vertices from a figure's center of mass) $\boldsymbol{r}_F$:

$$t_F = c \cdot (\boldsymbol{r}_F)^{\lambda} \tag{8}$$

The scaling variables c and λ are set independently for pair figures and higher-order figures. They are *not* chosen by the user, but optimized within the GA.

For the fitting of the interactions according to equation (7), a set of constraints is introduced as proposed by Garbulsky and Ceder [36]. These constraints maintain the energetic hierarchy of the input structures within the hierarchy of the predicted energetics:

$$|\Delta H^{\mathrm{DFT}}(\sigma) - \Delta H^{\mathrm{CE}}(\sigma)| \quad < \delta_1(\sigma), \tag{9}$$

$$|\Delta H^{\mathrm{DFT}}_{\mathrm{GSL}}(\sigma) - \Delta H^{\mathrm{CE}}_{\mathrm{GSL}}(\sigma)| \quad < \delta_2(\sigma), \tag{10}$$

$$|\Delta H^{\mathrm{DFT}}_{\mathrm{lowest}}(\sigma) - \Delta H^{\mathrm{CE}}_{\mathrm{lowest}}(\sigma)| < \delta_3(\sigma). \tag{11}$$

The first constraint simply requires that the enthalpy $\Delta H(\sigma)$ of every structure σ, as calculated by DFT and predicted by the CE, matches within the error bars $\delta_1(\sigma)$. Independent error bars $\delta_i(\sigma)$ are set up for the energy distance of the enthalpy of a structure to the value of the ground state line at the respective concentration $\Delta H_{\mathrm{GSL}}(\sigma)$, as well as for the energy distance between a structure's enthalpy and the enthalpy of the energetically lowest structure at this concentration $\Delta H_{\mathrm{lowest}}(\sigma)$. For the actual fitting of equation (7) within the constraints of equation (9), an algorithm proposed by Goldfarb and Idnani [37] is implemented.

In some cases it may be more important to conserve the energy hierarchy for low energy input structures than for less stable structures. Thus, the error bars $\delta_i(\sigma)$ defined in equation (9) depend on each structure's energy difference to the lowest structure at the respective concentration $\Delta H^{\mathrm{DFT}}_{\mathrm{lowest}}(\sigma)$, determined from first principles:

$$\delta_{\{1,2,3\}}(\sigma) = \delta^{\mathrm{const}}_{\{1,2,3\}} \cdot \exp\left(-\frac{\Delta H^{\mathrm{DFT}}_{\mathrm{lowest}}(\sigma)}{k_B \cdot T} \right), \tag{12}$$

The constant part $\delta^{\mathrm{const}}_{\{1,2,3\}}$ is specified at runtime. The Boltzmann-like energy-dependence can be varied through the term $k_B T$, and effectively turned off if desired.

While the fitting process is automatic, it introduces a set of new parameters for the fit itself (c and λ) as well as the Garbulsky–Ceder constraints. While the variables c and λ are optimized automatically within the GA, $\delta_i(\sigma)$ and $k_B T$ have to be specified by the user. Nevertheless, it is simple to make sure that the constraints are set correctly by checking whether the hierarchy predicted by the cluster expansion correctly reflects the hierarchy as determined by density functional theory. If this is not the case, then the constraints are lowered until the energetic hierarchy is preserved. In practice, this is a simple step for the user.

It should be mentioned that, indeed, in practice the use of δ_1 alone for the fitting of the ECIs may change the energetic hierachy of the formation enthalpies compared with the *ab initio* calculations, and consequently lead to a wrong ground state prediction. One example for such a scenario is the ternary system 'Ni–Al–Vacancy' discussed in section 3.1. For this system, it is possible to reach a CVS smaller than 6 meV by use of the Conolly–Williams

approach; however the resulting convex hull is incorrect due to the existence of many defect configurations around the ground state line with nearly identical formation enthalpies.

2.7. Treatment of alloy surfaces

The selection and determination of the ECIs becomes challenging for low-symmetry systems such as surfaces. In the case of a surface, there is a loss of *translational* symmetry in one dimension. Consequently, the number of independent figures increases significantly because the ECIs become *layer dependent*. Compared with a bulk case, a larger number of input structures is necessary in order to determine the ECIs. However, it is possible to circumvent a part of this problem by treating the surface interactions as 'correction' of the bulk interactions.

Because energies are additive, we may write

$$\Delta H_f^{\mathrm{CE}} = \Delta H_f^{\mathrm{vol}} + \Delta H_f^{\mathrm{surf}}. \tag{13}$$

This ansatz was first applied by Drautz *et al* to study the energetics of Ni-rich Ni–Al surfaces [23]. The advantage in treating the surface interactions as a *correction* of the bulk interactions comes from the fact that the DFT calculations for different surface terminations and segregation profiles do not have to account for an infinite bulk reservoir. We only have to make sure that the DFT slab model is thick enough that the center layer of the slab is bulk-like. The energy of a structure σ can then be written as

$$E(\sigma) = \sum_{i=1}^{N} \left\{ \sum^{N_F} d_F \overline{\Pi}_F(\sigma) J_F + \sum^{N'_F} d'_F(\boldsymbol{R}_i) \overline{\Pi}'_F(\boldsymbol{R}_i) \delta J_F(\boldsymbol{R}_i) \right\}. \tag{14}$$

The first term in the brackets is for the bulk energetics and the second is for the surface. We see that for the surface part the interactions become site dependent. Here, $\boldsymbol{R}_i$ defines the position of the atom i with respect to the alloy surface. Hence, for an atom i within the segregation profile, every individual interaction J_F to neighboring atoms will be corrected to $J_F + \delta J_f(\boldsymbol{R}_i)$. Naturally, with increasing distance from the alloy surface, $\delta J_F \rightarrow 0$ and consequently the surface term (second term) in equation (14) becomes zero. In the case of, e.g. a $Pt_{25}Rh_{75}(1\,1\,1)$ surface, the convergence was rapid. By the fourth layer the δJ_F had vanished [25, 38].

In practice, for more complex surface problems, even this partition of the energy may be an insufficient strategy. In some cases, finding a sufficiently predictive set of ECIs may still require an unreasonably large number of DFT calculations. We are currently developing an additional concept to be implemented in UNCLE that will provide an improved reference energy as starting point for surface investigation. In this extended approach, the mixed space cluster expansion [3, 4] is applied to incorporate strain effects into the reference energy part. Next, the energies of individual surface configurations are built from fully relaxed 1×1 surface structures, and, again, added to the reference energy part. We call this the concept of 'structural bricks'. After its implementation, it will be described in detail in [39].

2.8. Exhaustive searching of configuration spaces

The prediction of the energy (or any other observable) over a system's configuration space (e.g. ground state searches) also requires only minimal user input. We have implemented an algorithm [28] that automatically generates all possible atomic configurations within all geometrically possible supercells for an arbitrary number of basis atoms on a given lattice. The algorithm removes all symmetry-equivalent structures and still scales linearly with the number of unique configurations.

For a ground state search based on the cluster-expansion Hamiltonian, equation (1), the user only has to provide (i) the maximum number of basis atoms up to which configurations are to be considered and (ii) the figure set chosen by a previous GA run, along with the corresponding ECIs J. With this input, UNCLE automatically generates all possible superstructures (configurations) and determines their energy as predicted by the cluster expansion. The resulting ground state diagram and convex hull essentially constitute the phase diagram of the system at $T = 0\,\mathrm{K}$.

2.9. Application to Monte Carlo simulations

Within UNCLE we have also implemented a Monte Carlo simulation [40] based upon the Metropolis algorithm [41]. The code performs simulations either within the canonical or the grand canonical ensemble.

Similar to the calculation of the input structures' correlations for the cluster expansion, the determination of the starting energy of the Monte Carlo cell is done within the g-representation provided by the SNF. The Monte Carlo cell is thus represented by the tensor G, as introduced in section 2.4. Changing the atomic occupation of a site corresponds to changing the corresponding integer value of one element of G. In a Monte Carlo simulation, the calculation of the energy changes due to changes in the occupation (atom swaps) can be computed efficiently as only the energy contribution of those interactions 'touched' by the swapped sites needs to be evaluated.

The tensor G is the only large entity stored at runtime, requiring only *one byte* per site within the Monte Carlo cell; the correlations do *not* have to be stored at runtime. The minimal memory footprint allows for Monte Carlo cells of billions of sites, cpu time, rather than memory, becoming the limiting factor. A parallel implementation is underway to take advantage of this approach.

3. Application

Due to its very general concept, UNCLE can be applied to different types of material problems such as multi-component systems, multi-site adsorption on surfaces or surface segregation phenomena. In this section we give three representative examples of the successful application of UNCLE to different classes of problems.

3.1. Ternary bulk system: NiAl with explicit treatment of vacancies

For the Ni–Al system, it is well known that the ordering of defects plays a fundamental role to understand the structure and stability within the NiAl B2 and the Ni_2Al_3 phase. As shown earlier [42–45], on the simple cubic Ni sublattice of the B2-NiAl phase, vacancies are the dominant defect type in Ni-poor NiAl. In the sense of the cluster expansion, both phases can be described as bcc-based superstructures. For studying individual domains of B2-NiAl and Ni_2Al_3 on one common lattice, the latter is not be modeled as a bcc-like structures without any vacancies. Instead, the Ni_2Al_3 is described as a decoration of Ni- and Al-atoms on the B2 lattice with $\frac{1}{6}$ of the lattice sites left vacant. Therefore, for the simulation of defect order with Ni_xAl_{1-x} in the concentration range $0.4 \leqslant x \leqslant 0.5$, the cluster expansion needs to explicitly treat vacancies as a third component.

In order to obtain a converged cluster expansion for this system, 129 structures were calculated using VASP. Based on a total number of 711 figures with up to six vertices, the GA chose a set of figures with a total of 82 ECI J_F. The cross-validation-score was calculated

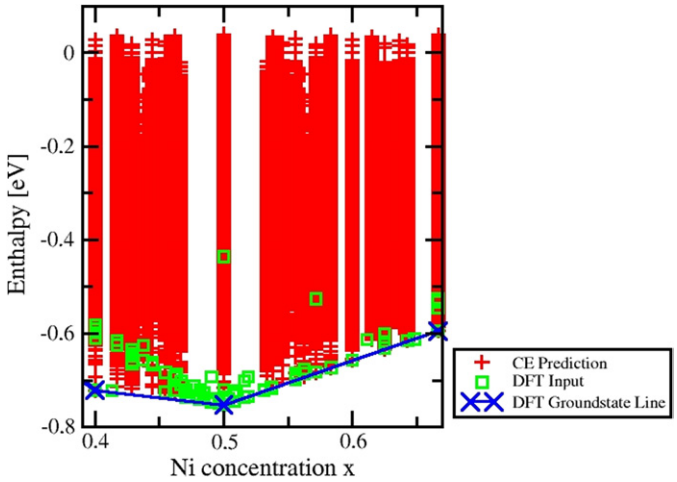


Figure 5. Calculated ground state diagram of all bcc-based superstructures with up to 16 sites occupied by Ni, Al and less than 21% vacancies. The points to which the enthalpies give reference are bcc-Ni and bcc-Al.

using two hundred 'leave-out' sets, each with 10 predicted structures. The resulting CV score was $S_{\mathrm{CV}} = 6.0$ meV, just a few per cent of the enthalpies in the input set.

Figure 5 shows that part of the resulting ground state diagram to which the constructed CE is applicable, as outside of the concentration regime fcc-based superstructures are energetically stable. This is also the regime where we made sure that the CE is fully converged. In detail, the ground state diagram has been limited to $\mathrm{Ni}_x\mathrm{Al}_{(1-x)}$ concentrations $0.4 < x < 0.\overline{6}$. This is the only concentration regime within which bcc-based superstructures are observed experimentally [42, 46]. Furthermore our investigation exclusively focused on the description of point defects *within* this concentration regime. This is why the cluster expansion only required convergence for this concentration range. The configuration space search included all ternary bcc superstructures with 16 sites or less and with less than 21% vacancies. The number of unique configurations is nearly 13 million, which can be computed quickly on a single cpu.

Each '□' in figure 5 indicates the enthalpy of a structure that was calculated by the DFT and included in the cluster expansion to extract the ECIs. Every '+' in the figure corresponds to the cluster expansion prediction for one atomic configuration. Consistent with the observed phase diagram, $\mathrm{Ni}_2\mathrm{Al}_3$ and B2 are predicted to be stable at $x = 0.4$ and $x = 0.5$ respectively. The third stable ground state within the converged part of the cluster expansion is $\mathrm{Ni}_2\mathrm{Al}$ at $x = 0.\overline{6}$, which can be observed experimentally to be a metastable state [46].

The cluster expansion Hamiltonian corresponding to the ground state diagram of figure 5 can also be applied to study the defect order at finite temperature. More than the ground state search (which holds no surprises), it is in this context that the cluster expansion is useful for the Ni-Al system. As one example, figure 6 provides a view into the ordering of B2-NiAl for $T \approx 4900$ K (left) and room temperature (right) resulting from Monte Carlo modeling. The (1 0 0) plane shown in the figure is one layer of a Monte Carlo cell consisting of one million lattice sites. The concentrations of the three constituents have been fixed to 50% Ni, 45% Al and 5% vacancies.

In full global thermodynamic equilibrium, a Monte Carlo cell containing 50% of both Ni and Al should exhibit a single B2-ordered domain. Thus a cut along a (1 0 0) plane would only

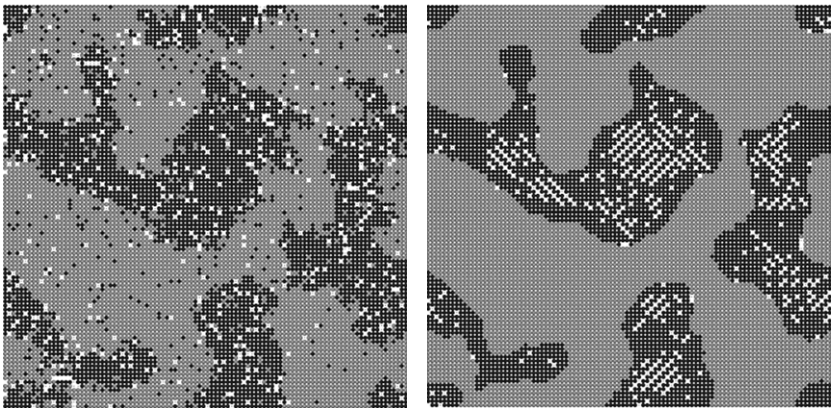


Figure 6. B2-NiAl with 5% vacancies: Cut along the (1 0 0) orientation through a $100 \times 100 \times 100$ Monte Carlo cell for $T = 4936$ K (left) and $T = 294$ K. It can be seen that for lower temperatures all vacancies (white) short range order in the Ni-domains (black).

contain either Ni- or Al-atoms, depending on whether it lies within the Ni- or the Al-sublattice. Figure 6 shows that the (1 0 0)-plane consists of both Ni- and Al-domains. These regions of Ni and Al belong to different B2-domains, which coexist within the Monte Carlo cell. By changing the external parameters of the simulation, the Monte Carlo cell can be brought into thermal equilibrium and the different domains visible in figure 6 merge into a single B2 domain. While strongly increasing the required calculation time, this does not add any new scientific insight, as the presence of the anti-phase boundaries between different B2-domains does not interfere with the observation of short-ranged vacancy order in the bulk of the respective B2-domains. The important point is to notice that figure 6 shows B2-domains with a stacking fault in between them and *not* domains of pure Ni or Al. This allows us to observe Al-subplanes of the B2-structure (light-gray domain) and the Ni-subplanes (dark-gray domains) within a single cut along (1 0 0).

For the high temperature case the formation of different B2-domains (dark and light gray) on the lattice can already be observed. The vacancies (white) occupy nearly random sites within Ni- and Al-subplanes. At room temperature the formation of the different B2-domains is complete and the vacancies form diagonal chains within the Ni-subplanes of the B2-phase only. These chains can be interpreted as the starting growth of the Ni_2Al_3 phase, where the vacancies are ordered in the same way. Such simulations allow for a quantitative analysis of the phase stability of these alloy phases. A detailed interpretation and evaluation of the structural properties can be found in [47].

3.2. Multi-site adsorption: H on the Ir(1 0 0) − (5 × 1) − *H surface*

One important system class in materials science (whose modeling demands both site-dependent interactions and arbitrary lattice types) is multi-site adsorption at surfaces. Here, the self-organization of nanostructures on surfaces has potential technological importance. Such self-organization has been observed when adsorbing iron on the iridium (1 0 0) surface [48, 49]. In its ground state, the Ir(1 0 0) surface exhibits a (5 × 1) quasi-hexagonal reconstruction (figure 7(*a*)). The unit cell contains six atoms within a (5 × 1) unit cell, that is, it contains one extra atom per unit cell relative to the layers below. As can be seen from the side view (figure 7(*a*)), the reconstruction goes several atomic layers deep into the crystal. Upon H

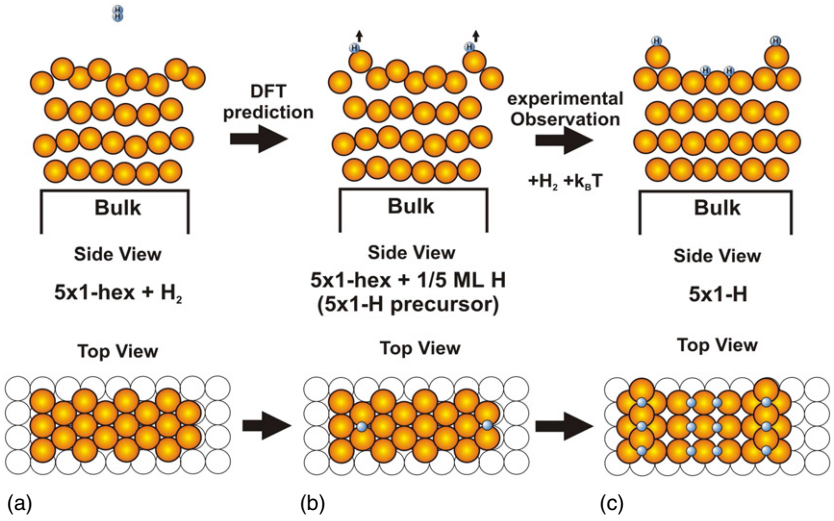


Figure 7. Schematic side view of the restructuring of the Ir(1 0 0) surface to the so-called Ir(1 0 0)-5 × 1-H phase by hydrogen adsorption.

adsorption, the Ir(1 0 0) surface reconstructs (figure 7(*b*)) to the so-called $\mathrm{Ir}(1\,0\,0)-(5\times1)-H$ phase (figure 7(*c*)). This phase can be characterized as single atom Ir chains, spaced five nearest-neighbor distances away, residing on an otherwise unreconstructed fcc (1 0 0) surface [48]. Because of its reconstruction, this phase is a potential template for the stabilization of magnetic nanostructures. Such an application requires a quantitative understanding of the geometrical and electronic properties to predict the stability of the template and the nanostructures grown on it. A prerequisite to such an understanding is the knowledge of the ground state configurations as a function of the hydrogen coverage.

Experimental and theoretical studies [27] find that at least six relevant adsorption sites for the adsorption of H exist, out of which 4 are inequivalent by symmetry and thus exhibit different energetic behavior. For this set of six sites within the (5 × 1) unit cell a cluster expansion has been performed with UNCLE leading to the ground state diagram displayed in figure 8. Top views of the resulting ground state configurations are shown below the diagram. In total, the search finds five stable ground states between 0 and 1.2 ML coverage. Two of them mark especially pronounced kinks in the ground-state line and therefore deserve special attention: at $\theta = 0.2$ ML the stable ground state is the single occupation of the site B1 on the Ir chains as already expected from density functional calculations [27]. At $\theta = 0.8$ ML the stable ground state is the full occupation of the B1, B2 and both equivalent B4 sites (configuration 'B1B4B2B4'). The coverage of this ground state coincides with the saturation coverage and therefore the structure of this adsorbate configuration is easily accessible by low-energy electron diffraction (LEED) experiments, which corroborate the ground state predicted by the cluster expansion. A detailed discussion can be found in [27].

3.3. Surface segregation: the system $Pt_{25}Rh_{75}$(1 1 1)

As known from experimental studies on $\mathrm{Pt_{25}Rh_{75}}$(1 1 1) [50, 51], this surface possesses a characteristic segregation profile: while the top layer shows a Pt enrichment, Pt depletion is found for the layer underneath. The existence of an equilibrium segregation profile is manifested by chemically resolved STM images [50], low energy ion scattering (LEIS) and

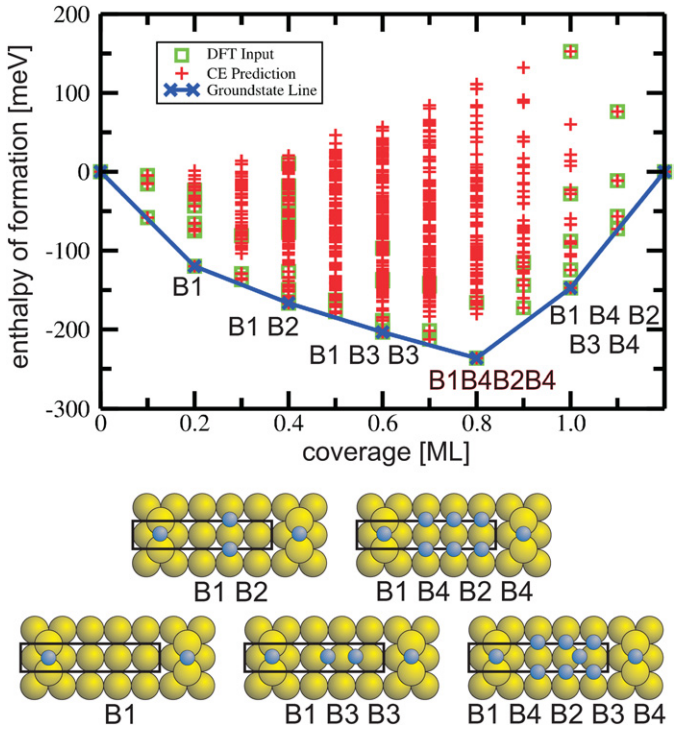


Figure 8. Ground state diagram of the system H/Ir(1 0 0)-5 × 1-H. The diagram considers all possible combinations of the energetically stable adsorption sites. The ground states found are displayed in top view below the ground state diagram.

quantitative LEED analyses [51]. These studies unambiguously show that above ∼1000 K the observed segregation profile does not depend on the annealing temperature of the sample.

Considering the energetics of the alloy system Pt–Rh, the pronounced segregation profile appears to be a surprise, because formation enthalpies of intermetallic compounds are all between 0 and about −20 meV per atom, i.e. smaller than kT at room temperature. This is in agreement with the bulk phase diagram of this binary system which does not show any long-range ordered structures in the experimentally accessible temperature regime. Instead, a fcc-based solid solution is stable for all concentrations. As a consequence of this small heterogeneous bonding, all constructed ECIs J_F for bulk and surface are unusually small, possessing energy values much smaller than 20 meV per atom, and cannot explain the characteristic segregation profile found for the (1 1 1) surface. However, there is one relevant deviation between the energetic properties of the bulk and the surface: due to the symmetry break the onsite energies of individual atomic sites which are defined by J_0 and J_1 in equation (1) are different for the near-surface layers compared with the bulk. For only weakly ordering systems as the $Pt_{25}Rh_{75}$(1 1 1) surface these onsite energies represent a good measure for the segregation behavior. Actually, it turns out that the top layer shows a tremendous tendency for an enrichment with Pt atoms reflected by an energy gain of about 0.2 eV per atom! Interestingly the opposite is true for the layer underneath: here, the onsite energy speaks for a Pt depletion and clustering of Rh atoms.

In order to predict the segregation profile quantitatively, Monte Carlo simulations were performed. As displayed in figure 9(left) our constructed cluster expansion is able to reproduce

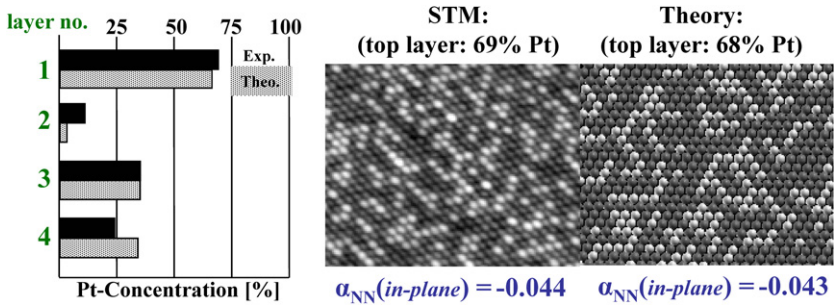


Figure 9. Left: experimentally determined and predicted segregation profile for $Pt_{25}Rh_{75}$ (1 0 0) ($T_{anneal} = 1400$ K). Right: corresponding short range order behavior as found by STM and predicted by our CE approach.

the experimental segregation profile determined via quantitative LEED analysis [51]. It turns out that for this surface system a 40 × 40 atom cell per layer was sufficient for a quantitative description of the segregation profile as well as the substitutional ordering. Figure 9(right) compares an STM image with atomic and chemical resolution [50] with our predicted one. It can be seen that there is an excellent (quantitative) agreement between experiment and theory.

4. Conclusions

With the program package UNCLE, we present a tool that makes the cluster expansion more accessible to non-specialists and applicable to a wide variety of physical problems. Several extensions of the formalism were presented: use of the *g*-representation simplifies and automates the 'chores' of setting up and constructing a cluster expansion, performing ground state searches and using the ECIs in Monte Carlo simulations. By automating much of the cluster expansion construction and use, problems arising from user errors are less likely, resulting in more robust predictions. The treatment of surface alloys and related systems is made possible through the separation of the cluster expansion Hamiltonian into a bulk and a surface part.

The possibility to apply UNCLE to arbitrary lattice types will extend the application field of the cluster expansion method to much larger and more complex systems.

Acknowledgments

The authors gratefully acknowledge support by DFG, MU 1648/2 and Mu 1648/3 (Germany), Studienstiftung des deutschen Volkes, and by the NSF DMR-0650406 (USA).

References

[1] Sanchez J M, Ducastelle F and Gratias D 1984 *Physica* A **128** 334–50
[2] de Fontaine D 1994 Cluster approach to order-disorder transformations in alloys *Solid State Phys.* **47** 33–176
[3] Laks D B, Ferreira L G, Froyen S and Zunger A 1992 *Phys. Rev.* B **46** 12587–605
[4] Zunger A 1994 ed P Turchi and A Gonis *Statics and Dynamics of Alloy Phase Transformations (NATO ASI)* (New York: Plenum) pp 361–419
[5] Müller S 2003 *J. Phys.: Condens. Matter* **15** R1429
[6] Zunger A, Wang L, Hart G L W and Sanati M 2002 *Model. Simul. Mater. Sci. Eng.* **10** 3105
[7] Blum V, Hart G L W, Walorski M J and Zunger A 2005 *Phys. Rev.* B **72** 165113

[8] van de Walle A 2008 *Nature Mater.* **7** 455–8
[9] van de Walle A and Ceder G 2002 *J. Phase Equilib.* **23** 348
[10] van de Walle A, Asta M and Ceder G 2002 *CALPHAD* **26** 539
[11] van de Walle A and Asta M 2002 *Model. Simul. Mater. Sci. Eng.* **10** 521
[12] Kresse G and Hafner J 1993 *Phys. Rev.* B **47** 558
[13] Kresse G and Furthmüller J 1996 *Phys. Rev.* B **54** 11169
[14] Kresse G and Joubert D 1999 *Phys. Rev.* B **59** 1758–75
[15] Kresse G and Furthmüller J 1996 *Comput. Mat. Sci.* **6** 15
[16] Kresse G and Hafner J 1994 *J. Phys.: Condens. Matter* **6** 8245
[17] Wimmer E, Krakauer H, Weinert M and Freeman A J 1981 *Phys. Rev.* B **24** 864–75
[18] Weinert M 1981 *J. Math. Phys.* **22** 2433
[19] Weinert M, Wimmer E and Freeman A J 1982 *Phys. Rev.* B **26** 4571–8
[20] Wolverton C 1999 *Phil. Mag. Lett.* **79** 683
[21] Müller S, Wolverton C, Wang L-W and Zunger A 2001 *Europhys. Lett.* **55** 33
[22] Müller S, Wang L-W and Zunger A 2002 *Model. Simul. Mater. Sci. Eng.* **10** 131
[23] Drautz R, Reichert H, Fähnle M, Dosch H and Sanchez J M 2001 *Phys. Rev. Lett.* **87** 236102
[24] Wieckhorst O, Müller S, Hammer L and Heinz K 2004 *Phys. Rev. Lett.* **92** 195503
[25] Müller S, Stöhr M and Wieckhorst O 2005 *Appl. Phys.* A **82** 415
[26] Marcel H F Sluiter and Kawazoe Y 2005 *Phys. Rev.* B **71** 212201
[27] Lerch D, Wieckhorst O, Hammer L, Heinz K and Müller S 2008 *Phys. Rev.* B **78** 121405
[28] Hart G L W and Forcade R W 2008 *Phys. Rev.* B **77** 224115
[29] Hart G L W and Forcade R 2009 *Phys. Rev.* B submitted
[30] Wolverton C and de Fontaine D 1994 *Phys. Rev.* B **49** 8627–42
[31] Hart G L W, Blum V, Walorski J and Zunger A 2005 *Nature Mater.* **4** 391
[32] Baumann K 2003 *Trends Anal. Chem.* **22** 395
[33] Zarkevich N A and Johnson D D 2004 *Phys. Rev. Lett.* **92** 255702
[34] Diaz-Ortiz A and Dosch H 2007 *Phys. Rev.* B **76** 012202
[35] Díaz-Ortiz A, Dosch H and Drautz R 2007 *J. Phys.: Condens. Matter* **19** 406206
[36] Garbulsky G D and Ceder G 1994 *Phys. Rev.* B **49** 6327–30
[37] Goldfarb D and Idnani A 1983 *Math. Prog.* **27** 1
[38] Müller S 2006 *Surf. Interface Anal.* **38** 1158–63
[39] Kerscher T, Wieckhorst O and Müller S in preparation
[40] Binder K and Heermann D W 1978 *Monte Carlo Simulation in Statistical Physics* (Berlin: Springer)
[41] Metropolis N, Rosenbluth A W, Rosenbluth M N, Teller A H and Teller E 1953 *J. Chem. Phys.* **21** 1087
[42] Taylor A and Doyle N J 1972 *J. Appl. Crystallogr.* **5** 201–9
[43] Fu C L, Ye Y-Y, Yoo M H and Ho K M 1993 *Phys. Rev.* B **48** 6712–15
[44] Meyer B and Fähnle M 1999 *Phys. Rev.* B **59** 6072–82
[45] Lechermann F and Fähnle M 2000 *Phys. Rev.* B **63** 012104
[46] Reyaud F 1976 *J. Appl. Crystallogr.* **9** 263–8
[47] Lerch D and Müller S in preparation
[48] Hammer L, Meier W, Klein A, Landfried P, Schmidt A and Heinz K 2003 *Phys. Rev. Lett.* **91** 156101
[49] Klein A, Schmidt A, Hammer L and Heinz K 2004 *Europhys. Lett.* **65** 830–6
[50] Hebenstreit E L D, Hebenstreit M, Schmid M and Varga P 1999 *Surf. Sci.* **441** 441
[51] Platzgummer E, Sporn M, Koller R, Forsthuber S, Schmid M, Hofer W and Varga P 1999 *Surf. Sci.* **419** 236