2005-10-26

Using genetic algorithms to map first-principles results to model Hamiltonians: Application to the generalized Ising model for alloys

Gus L. W. Hart
gus.hart@gmail.com

Volker Blum

Michael J. Walorski

Alex Zunger

Original Publication Citation

Volker Blum, Gus L. W. Hart, Michael Walorski*, and Alex Zunger, "Using genetic algorithms to develop model Hamiltonians: Applications to the generalized Ising model," Phys. Rev. B 72 165113 (26 Oct. 25). The original article may be found here: http://prb.aps.org/abstract/PRB/v72/i16/e165113

Using genetic algorithms to map first-principles results to model Hamiltonians: Application to the generalized Ising model for alloys

Volker Blum,[1,*] Gus L. W. Hart,[2] Michael J. Walorski,[3] and Alex Zunger[1]
[1]*National Renewable Energy Laboratory, Golden, Colorado 80401, USA*
[2]*Department of Physics and Astronomy, Northern Arizona University, Flagstaff, Arizona 86011-6010, USA*
[3]*Computer Science Department, Northern Arizona University, Flagstaff, Arizona 86011-5600, USA*
(Received 15 August 2005; published 26 October 2005)

The cluster expansion method provides a standard framework to map first-principles generated energies for a few selected configurations of a binary alloy onto a finite set of pair and many-body interactions between the alloyed elements. These interactions describe the energetics of all possible configurations of the same alloy, which can hence be readily used to identify ground state structures and, through statistical mechanics solutions, find finite-temperature properties. In practice, the biggest challenge is to identify the types of interactions which are most important for a given alloy out of the many possibilities. We describe a genetic algorithm which automates this task. To avoid a possible trapping in a locally optimal interaction set, we periodically "lock out" persistent near-optimal cluster expansions. In this way, we identify not only the best possible combination of interaction types but also any near-optimal cluster expansions. Our strategy is not restricted to the cluster expansion method alone, and can be applied to select the qualitative parameter types of any other class of complex model Hamiltonians.

DOI: 10.1103/PhysRevB.72.165113 PACS number(s): 61.50.Ah, 61.66.Dk, 71.15.Nc

I. INTRODUCTION: THE DILEMMA OF SELECTING INTERACTION TYPES IN CONFIGURATIONAL EXPANSIONS

The cluster expansion (CE) method[1–6] provides today's state-of-the-art framework to parametrize the energetics of multicomponent systems as a functional of configurational variables on an underlying lattice. It is routinely applied to binary alloys which form on an underlying primitive Bravais lattice (e.g., bcc or fcc), with applications to more demanding problems (complex basic lattices,[7–9] multicomponent systems,[10–15] and surfaces[16–24]) rapidly growing in number and success. The basic premise of the CE method is that the energetics of different configurations σ of a given element combination A-B can be described by an Ising-like framework of pair and multisite interactions J:

$$E_{\mathrm{CE}}(\sigma) = J_0 + \sum_{\text{sites}} J_i \hat{S}_i + \sum_{\text{pairs}} J_{ij} \hat{S}_i \hat{S}_j + \sum_{\text{triplets}} J_{ijk} \hat{S}_i \hat{S}_j \hat{S}_k + \cdots \tag{1}$$

[$\hat{S}_i = -1(+1)$ if lattice site i is occupied by $A(B)$]. In principle, this expansion is exact[1]—that is, if all inequivalent pair and many-body interaction types (MBIT) on the lattice are taken into account. In practice, the method relies on there being only a finite number of non-negligible interactions. For each given alloy A-B, Eq. (1) can then be truncated to the system-specific relevant interactions only. Their numerical values can be determined by a fit to a finite number of *ab initio* calculated formation enthalpies for different configurations. The result is a simple formula which describes the energetics of *any* configuration of A-B with the accuracy of the underlying *ab initio* method. This formula can then be used to scan many configurations σ in search for ground states[25] or configurational thermodynamic properties[26,27] such as phase transitions, short-range order, etc.

Establishing Eq. (1) requires addressing the following general issues.

A. Concentration-dependent or concentration-independent *J*'s

A successful cluster expansion $E_{\mathrm{CE}}(\sigma)$ is expected to reproduce the features of a direct quantum-mechanical energy $E_{\mathrm{QM}}(\sigma) = \langle \Psi | H | \Psi \rangle$. We thus have a choice of considering an equation-of-state $E_{\mathrm{QM}}(\sigma, V)$ description, where the energy is calculated at each volume. In this case the ensuing $\{J(V)\}$ will be volume dependent, and therefore also composition (x) dependent $J[V(x)]$. This approach was used extensively early on.[28–31] Alternatively, one may want to focus on the *equilibrium* quantum-mechanical energy $E_{\mathrm{QM}}[\sigma, V_{\mathrm{eq}}(\sigma)] \equiv E_{\mathrm{QM}}(\sigma)$, deducing the corresponding volume-independent interaction energies J. This is our choice here, and in earlier papers.[2,5,26,32–38] The set of volume-independent interactions is in principle complete:[1] since there are 2^N types of figures (=MBITs) on a lattice of N points, and 2^N possible configurations σ, the set of 2^N algebraic equations (1) uniquely defines a set of volume-independent J, and there is no mathematical need to add other variables. Indeed, the choice of representation is a matter of convenience. The two representations have different convergence properties, as discussed by Ferreira *et al.*,[39] and could be renormalized into each other.[40] Whichever representation is used, one must naturally demonstrate the series convergence to a give tolerance, e.g., by predicting the energies $E_{\mathrm{CE}}(\sigma)$ of additional input structures σ and verifying them against their directly calculated counterparts $E_{\mathrm{QM}}(\sigma)$ until the desired predictive accuracy is achieved. Indeed, this situation is analogous to a basis set expansion in electronic structure theory, where different

bases (plane waves; Gaussians; muffin-tin orbitals) do not have any particular physical meaning, have different convergence properties, but in the limit all produce the same variational total energy.

B. Obtain J directly or from a set of total energies $E_{\mathrm{QM}}(\sigma)$?

Our approach is based on the Connolly-Williams[41] suggestion to derive $\{J\}$ from a set of quantum-mechanically calculated total energies $\{E_{\mathrm{QM}}(\sigma, V_\sigma)\}$ of some ordered or disordered configurations $\{\sigma\}$. In principle, it is also possible to calculate the required interaction energies J directly,[42–46] rather than extracting them from the total energies of some configurations. The main advantage of the former approach is that presently it is possible to compute total energies of ordered structures $E_{\mathrm{QM}}(\sigma)$ without invoking computational approximations that characterize methods that obtain $\{J\}$ directly. Indeed, in practice, the latter often necessitate additional approximations at the electronic and/or structural level. For instance, linear response theory[43,47] gives *pair* interactions in SiGe[43] and GaInP[47] alloys, but higher-body interactions require higher-order linear response, which is much less tractable, and is often neglected. Similarly, the generalized perturbation method (GPM)[42,48] allows extraction of interaction values in lowest-order scattering theory from the coherent potential approximation (CPA). However, a number of computational compromises need to be made. For example, (i) because of the specific use of site-representation Green's functions, the underlying electronic structure approach must be restricted to a site-anchored representation such as the tight-binding atomic-sphere approximation in Korringa-Kohn-Rostoker (KKR) theory or the linear muffin tin orbital (LMTO) method. Thus, the variational flexibility and shape approximation of the basis set is restricted relative to more general bases (e.g., plane waves) used efficiently in contemporary methods for calculating E_{QM}. (ii) Until recently,[49–51] the existence of an electrostatic Madelung term in the energy of a random alloy was overlooked. This was pointed out in 1990.[52] As a result, the J's extracted before $\sim$2000 by the GPM for systems exhibiting some charge transfer are suspect. Recent remedies[49,50] fix the problem at the cost of introducing a parameter that cannot be determined by this theory itself, but must be borrowed from other types of calculations (e.g., supercells). (iii) The long-range strain field created by size mismatch in A_nB_m superstructures[6] has been neglected in the GPM. This "constituent strain" is included in our approach. (iv) The short-range relaxation is included via a model,[53] rather than by direct optimization of E_{QM} with respect to atomic positions. This model has been examined recently for Mo-Ta and found to be significantly deficient.[38] (v) The extraction of J from the total energy of the CPA is done fully for the sum of eigenvalues $\Sigma_i \epsilon_i$ part of the total energy, but only approximately for the interelectronic Coulomb and exchange term,[45,49,54] leading to unknown and uncontrolled errors. (vi) Finally, since in this approach the J's are extracted from a fictitious medium without configurational degrees of freedom, it remains unclear whether perturbation theory suffices to obtain J's that are appropriate for the description of configurationally ordered phases. Also, the number of relevant interactions can be underestimated in the perturbational limit to the random alloy, because all three- and higher-body contributions in Eq. (1) are weighted by third- and higher-order spin products.[55] For instance, Turchi *et al.*[56,57] conclude that only two pair interactions are needed to describe the configurational energetics of Mo-Ta and Ta-W, in stark contrast to our own results, based on O(50) first-principles configurational energies.[37,58,59] The interactions of Turchi *et al.* fail to predict almost all ground state configurations of Mo-Ta[37] and Ta-W,[58,59] and yield overestimated order-disorder transition temperatures in contradiction to calculations based on direct first-principles configurational energies[38,59] and experiment.[60–62]

We conclude that obtaining $\{J\}$ from a set of $\{E_{\mathrm{QM}}(\sigma)\}$ values is a more robust approach since such calculations can be done with (i) variationally unrestricted basis representations, (ii) full and unfitted Madelung energies, including both (iii) long-range strain[6] and (iv) accurate short-range atomic relaxation, while (v) keeping both the one-electron and the two-electron terms in the total energy on equal footing, and (vi) extracting $\{J\}$ nonperturbatively from reference systems which incorporate various explicit degrees of configurational order. Significantly, during the construction of a full CE (Sec. II), the expansion can be simply and directly tested to accurately reproduce $E_{\mathrm{QM}}(\sigma)$ for additional configurations not used in the fit, leaving no more doubt about convergence issues or numerical issues.

C. Truncating the expansion in J: Hierarchical approach vs selective approach

For practical use, in all CE approaches one must truncate Eq. (1) to a finite number of interaction types, but choosing exactly those MBIT which must be retained is not easy. For example, certain distant pairs or three-body figures may be more important than intimate pairs, and the set of "significant" MBIT varies for each different alloy. Essentially, two different approaches to this problem have been suggested in the literature.

1. Hierarchical approaches

Since, intuitively, the figures of smallest spatial range could be the most important, one might suggest to order all figures by their size, and declare a cutoff radius below which all figures will be included in the expansion. The simplest example is the nearest-neighbor fcc approximation widely used in early Ising Hamiltonians (Ref. 63 and references therein), the early cluster-variation method,[64] and the "Connolly-Williams" approximation.[41,65–68] Zarkevich and Johnson[69] have recently extended a hierarchical approach, legislating that if a given figure F is included, all other figures of same extent and vertex number as F and all subfigures of F should also be included. However, it is not clear, nor was it proven, that this restriction leads to better convergence or better predictions, and it is impractical to include all subfigures of a figure unless the series converges after very few terms. As an example, consider the fcc lattice and figures up to six vertices: There is a well-known group of only five

inequivalent figures that extend over a nearest-neighbor distance,[41] but already eleven if the maximum distance is second nearest-neighbors, and a total of 60 inequivalent figures that span a third-nearest-neighbor distance at most. In practice, it is well known that even third-nearest-neighbor distances may not be enough to capture the energetics of a binary alloy qualitatively,[37,70,71] and we have ourselves encountered many systems in the past where a hierarchy is not followed.[32,33,35,37,38]

Early truncation can be grossly inaccurate,[6,14,38] missing most (long-range) atomic relaxation effects and even qualitative features of a ground state hull and phase diagram. One may still attempt to fit all necessary figures impartially by including enough *ab initio* calculated input energies $E(\sigma)$, but this would lead to a brute-force approach of slow convergence. Van de Walle and Ceder[4] have shown how to make an automated hierarchy-based approach manageable by introducing leave-one-out cross-validation as a systematic criterion to assess the predictive power of a CE, but some computational overhead will be the price.

2. Selective approaches

An alternative approach, pursued, e.g., by Zunger *et al.*,[2,5,25,32,33,38] is to attempt to identify the leading interactions of Eq. (1) independent of hierarchical constraints, simply by comparing the predictive power of many different CE truncations for a given alloy system. In earlier papers, this was done by fitting the numerical values of J to only a subset of the input data and then predicting the rest, an approach more recently extended to leave-many-out cross-validation.[38,72,73] The set of input structures is split into two parts, one for fitting numerical values of J, and one to check predictions made with these numerical values. The procedure is repeated for different choices of fitting or prediction sets, and the average prediction error is the cross-validation score S_{cv}. In selective approaches, one sets up a pool of MBIT from which the leading interactions are selected without hierarchical constraints. We show in Fig. 1 some inequivalent MBIT (beyond pairs, as pairs can be reliably accounted for by a constrained fit method[5,6]) which we use as a standard pool of MBIT candidates on the body-centered cubic (bcc) lattice. Only a fraction of these MBIT are typically required, but it is not *a priori* clear which few must be kept. The overall pool is not designed according to any special principles. Instead, it is simply an exhaustive list of all MBIT up to a reasonable number of vertices and vertex distance, including all three-vertex MBIT up to fifth-nearest-neighbor distance, four-vertex MBIT up to fourth-nearest-neighbor distance, and five- and six-vertex MBIT up to third-nearest-neighbor distance. To ensure that the relevant physics of a given alloy system is not limited by the chosen pool of MBIT, the sufficient extent of the pool can be routinely tested by including additional figures as a convergence test, e.g., all three-body figures up to eighth-nearest-neighbor distance. Figure 1 also shows that the number of possible figures increases dramatically as longer distances and more vertices are added—for instance, there are only two bcc MBIT with a maximum vertex separation of 2, but already 14 bcc MBIT with a maximum vertex separation of three. In the past, the relevant MBIT were selected manually from the pool by minimizing the prediction error, but an exhaustive search is not feasible; e.g., searching for only five out of a pool of 45 possible MBIT leads to as many as 1.22 million different possibilities—a task beyond a brute-force search.

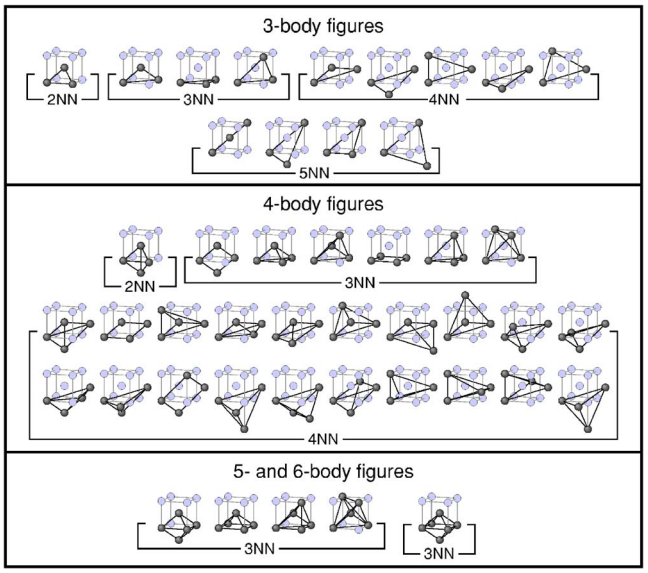


FIG. 1. (Color online) Pool of 45 MBIT on the bcc lattice. Figures are grouped by increasing number of vertices, and the largest vertex-vertex distance within a given figure (2NN, …, 5NN denote second- through fifth-nearest-neighbor separation).

We have recently pointed out[58] that the search for the "leading terms" of a model Hamiltonian can be efficiently performed using a genetic algorithm (GA).[74] In the present work, we show how this is done in particular for the choice of the MBIT which are relevant to reproduce local density approximation (LDA) energies in the approach (ii) above. The input information for a given alloy system is a set of first-principles calculated energies $\{E_{\mathrm{QM}}(\sigma)\}$ for selected configurations σ. The GA must then find the combination of MBIT with minimal cross-validation score, satisfying three criteria:

(1) It should converge significantly faster than a manual search.

(2) It should not get trapped in local minima.

(3) If there are multiple sets of MBIT which are almost equivalent to the best possible CE, the method should identify them all; a seemingly ambiguous CE for a given input set can then be unraveled by calculating selected additional input energies $E_{\mathrm{QM}}(\sigma)$.

II. DETERMINISTIC CONSTRUCTION OF A MIXED-BASIS CLUSTER EXPANSION

We employ the mixed-basis cluster expansion (MBCE) formalism[5,6] to determine the interaction types in Eq. (1), and their numerical values. Since the generality of Eq. (1) is fully preserved if different configuration-dependent reference terms are added to or subtracted from the total energy of a given alloy configuration, in the MBCE one improves the

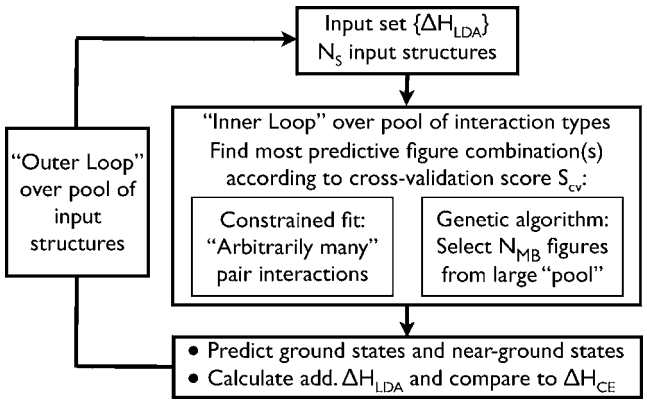


FIG. 2. Construction algorithm for a converged mixed-basis cluster expansion.

convergence of the cluster expansion by treating certain long-range contributions analytically.[6] The MBCE-expanded energy is written as

$$E(\sigma) = \Delta H_f(\sigma) - E_{\mathrm{CS}}(\sigma), \tag{2}$$

where ΔH_f denotes the enthalpy of formation of a given, fully relaxed alloy configuration σ $(A_{1-x}B_x)$ from the elemental solids A and B,

$$\Delta H_f(\sigma) = E_{\mathrm{tot}}(\sigma; A_{1-x}B_x) - (1-x)E_{\mathrm{tot}}(A) - xE_{\mathrm{tot}}(B) \tag{3}$$

(all total energies are per atom). $E_{\mathrm{CS}}(\sigma)$ is the configuration-dependent "constituent strain energy",[6] which can be calculated analytically from LDA data, and which removes a singularity from the Fourier transform of the real-space pair interactions, $J(k)$. Without subtracting E_{CS}, this singularity would arise because ΔH_f of a fully phase-separated configuration $[A_{1-x}B_x]^{\mathrm{phs}}$ on the same coherent underlying lattice is nonzero: $E_{\mathrm{tot}}([A_{1-x}B_x]^{\mathrm{phs}}) \neq (1-x)E_{\mathrm{tot}}(A) - xE_{\mathrm{tot}}(B)$, since the lattices of elemental A and B may relax independently while the coherent phase-separated limit remains constrained.

The construction of a verifiably predictive cluster expansion for $E(\sigma)$ consists of two iterative loops, as visualized in Fig. 2. The *inner loop* identifies the most predictive set of interaction types to describe a given set of first-principles calculated energies $\{E_{\mathrm{LDA}}(\sigma)\}$ for N_s input structures. The measure for the predictive power of a given set of interaction types is a leave-many-out cross-validation score[72,73] S_{cv}, as defined in Ref. 38. The N_s input structures are subdivided into a group of $N_f < N_s$ structures to *fit* the numerical values of the selected interaction types, and a group of $N_v = N_s - N_f$ structures which are not fitted, so that their *predicted* energies $E_{\mathrm{CE}}(\sigma)$ can be compared to the known energy $E_{\mathrm{LDA}}(\sigma)$ after the fit. This process is then repeated for b independent subdivisions into N_f fitting and N_v prediction structures, until each of the N_s input energies $\{E_{\mathrm{LDA}}(\sigma)\}$ was predicted at least twice. The average overall prediction errors from this process define

$$S_{\mathrm{cv}} = \frac{1}{bN_v} \sum_{(b \text{ sets})} \sum_{(N_v \ \sigma \text{ in set})} |E_{\mathrm{CE}}(\sigma) - E_{\mathrm{LDA}}(\sigma)|^2. \tag{4}$$

The goal of the inner loop, then, is to identify the combination(s) of interaction types (candidate CEs) with minimal S_{cv}.

The *outer loop* acts as a feedback loop to ensure that a CE, identified in the inner loop for the fixed subset of N_S structures, really possesses good predictive power for *all* 2^N configurations. Each candidate CE is used to search all 2^N structures for additional ground states or near-ground-state structures σ_{new}. Their energies $E_{\mathrm{LDA}}(\sigma_{\mathrm{new}})$ are then evaluated by direct LDA calculations and compared to the predicted $E_{\mathrm{CE}}(\sigma_{\mathrm{new}})$, giving an objective estimate of the predictive power of each candidate cluster expansion. The newly calculated $\{E_{\mathrm{LDA}}(\sigma_{\mathrm{new}})\}$ are added to the previous input set, and the inner loop is repeated. The outer loop iterations are converged when no more significant new ground-state structures are predicted, and all verified predicted energies agree with their direct LDA counterparts to within a few meV. For bulk alloys, $\gtrsim$50 LDA input structures[38,59] are usually enough to achieve convergence. The complete iterative procedure guarantees the identification of a well-converged truncated expansion Eq. (1), and additionally acts as a prediction engine for important candidate structures for ground states whose energy must be calculated directly in LDA.

The inner loop is where the difficult search problem for the most relevant interaction types arises, as outlined in the introduction. This problem is manageable for pairs, whose number increases relatively slowly with distance, and which can therefore be treated by the constrained fit method of Ref. 6, but the number of MBIT with three or more vertices increases much more rapidly with distance. The present paper concentrates on the selection of MBIT. We thus assume a fixed set of input structures, and always use the constrained fit method for pair interactions. Our goal is to select the best set of MBIT to minimize S_{cv} using a genetic algorithm. The rest of the paper explains how this task is done.

III. GENETIC ALGORITHM SELECTION OF MBIT

Genetic algorithms[74] use the biological idea of "survival of the fittest" to find the optimum solution to a given problem. GA's are particularly helpful when faced with strongly correlated search spaces, where other algorithms such as the sequential optimization of individual parameters, or methods based on individual, random parameter "flips" (Monte Carlo) would end up in local minima, or even fail to converge at all. GA's have been applied in many different settings, e.g., in computational condensed matter physics to find the optimal numerical values of given physical parameters such as geometric structure[75–78] or tight-binding parameters.[79] Our present application is different in that we aim to find the actual shape of a cluster expansion Hamiltonian, i.e., its interaction types rather than only their numerical values.

Generally, the trial solutions in a GA are encoded as *binary sequences* (the "genomes") of 0's and 1's (the "genes"). Here, the objective is to pick, from a large pool, a handful[5–10] of MBIT to be included in a trial CE, i.e., a truncation of Eq. (1). A natural encoding of trial CE is a

genome "…01110100011…" with one gene for each candidate MBIT in the pool, and a one (zero) denoting whether that figure is (is not) included. Over the course of the GA, a set of genomes is monitored over many *iterations* ("generations"). From one iteration to the next, "child" genomes are created by a *cross-over* ("mating") of two selected "parent" genomes of the earlier iteration. Each gene of a child genome takes on the value of that gene in either the first or the second parent. If this strategy were strictly implemented, only preexisting "genetic" information could be proliferated in a mating step. So, if a certain MBIT (or combination) were eliminated from the entire population of trial CE's in any one generation, this MBIT could never return later. A GA might lose a vital piece of the optimal solution at an early stage by accident and would later be doomed to remain stuck in a local (but not global) optimum forever. Nature's solution to this dilemma is *mutation*. To prevent a starvation of the diversity of possible trial solutions, individual genes can randomly be turned on or off in a newly created child genome, similar to the random mutations of evolutionary biology. We make the following choices [Sec. III A–III F below] to control the convergence of our particular GA.

A. Maximum number of "active" genes per genome

The "genomes" in our problem represent sets of MBIT (i.e., figure *types* as opposed to numerical values J) which are used to construct a CE. The optimized quantity is the cross-validation score S_{cv}, which measures the ability of a given CE to predict E_{QM} for structures not used in the fit. One additional measure is taken as a safeguard against overoptimization of S_{cv}: we impose a deliberate limit on the number N_{MB} of active MBIT per CE, i.e., we cap the number of active genes ("ones") in each genome. The development of S_{cv} as a function of N_{MB} may be studied to determine to what degree an increase in the number of CE parameters still helps improve predictive accuracy significantly.

B. Population size

The number of genomes per generation, N_{pop}, determines the amount of "genetic diversity" which is available to spawn subsequent generations. For optimum genetic diversity, we choose N_{pop} based on the number of MBIT in each CE, N_{MB}, with the requirement that each MBIT appear at least twice (possibly more often) in the initial generation.

C. Survival rate

A fraction r_{s} of the original N_{pop} candidate genomes with the momentary optimum fitness is retained from one generation to the next. The other genomes are replaced with children mated from the preceding generation. For instance, from a generation of 20 genomes with a survival rate $r_{\mathrm{s}} = 1/2$, the ten best individuals would be carried over unmodified. Ten children would be created to fill the remaining slots.

D. Mating favoritism

To create a child, two parents are randomly selected from the existing generation. Then, one by one the genes (zeroes and ones) of the child genome are selected from parent 1 or parent 2. The parent with better fitness has a higher probability of passing its genes on to the child than the less fit parent. In this way, the preferred proliferation of "better" genetic information is ensured.

E. Mutation rate

After each mating step, we allow each gene to be "flipped" from zero to one or vice versa with a certain (relatively low) probability. In fact, we choose this probability so as to obtain a certain number of flips N_{flips} *per genome on average*. Of course, we might accidentally end up with more MBIT in a CE than allowed by the maximum number N_{MB} after this step. In that case, we randomly pick some of these "ones" and turn them off [i.e., we remove figures from the corresponding truncation of Eq. (1)], until their number is reduced to the prescribed target number.

F. "Lock-out" strategy

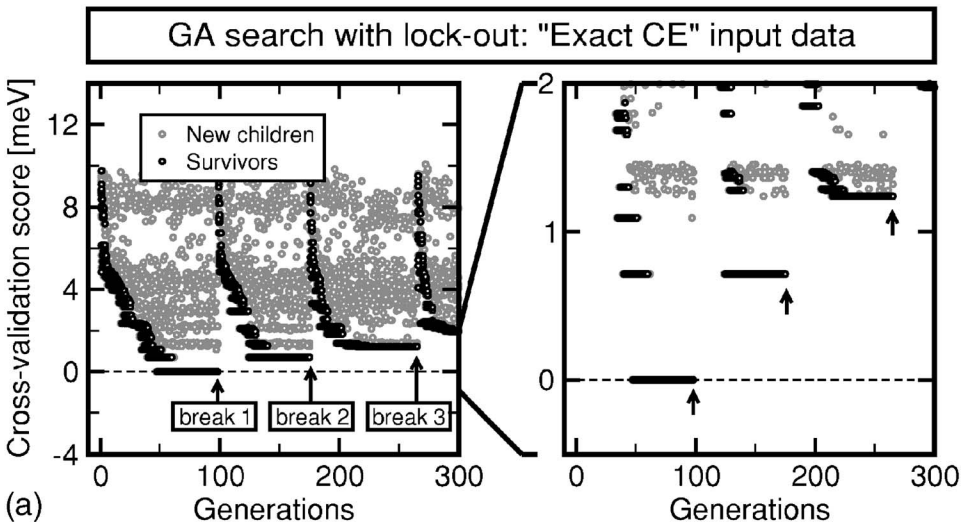
Even with significant initial genetic diversity and mutations, the problem of local optima—which exists in any global optimization scheme, not just a GA—is not fully resolved. If the GA first reaches a locally optimal CE that differs from the global one by several MBIT, the probability to progress by random mutations alone may become hopelessly small. As a result, the prospective alloy researcher may easily spend thousands of generations waiting for the correct minimum to be found. Even worse, in an actual application the best answer is not known, and hence it is impossible to be sure whether or not a persistent solution is already the best possible CE or not. To overcome this "locking" of the algorithm into a local minimum, we implement the idea of "locking out" any persistent solutions after progress has stopped for a certain number of generations (50–100). The persistent CE is recorded on a blacklist, and barred from ever occurring again. The algorithm is then reinitialized with a momentarily increased mutation rate in the next generation. The benefit is twofold. First, the algorithm is forced to look for another CE, which may or may not be better than the first. Second, the result of a GA run is a list of several near-optimal CEs in addition to the actual optimum. This gives direct insight into the degree of degeneracy of the search space explored.

IV. APPLICATION TO MO-TA

The criteria Sec. III A–III F determine our algorithm completely. Once the key parameters are set, the mating process can be repeated for an arbitrary number of generations until a target value of S_{cv} has been achieved.

A. Successful retrieval of the leading interactions

We first demonstrate the GA's ability to successfully retrieve the leading interactions from an input set $\{E_{\mathrm{exact}}(\sigma)\}$ whose underlying interactions are exactly known. To that end, we use Eq. (1) itself to calculate $E_{\mathrm{exact}}(\sigma)$ for 60 bcc input configurations σ, inserting the set interactions retrieved



(b)

Generation	Genome	CV score
97	0100100010010000000000010000000000000000000000	0.000
174	0000110010010000000000010000000000000000000000	0.714
443	0100100010010000000000010000000000000000000000	1.093
264	0000100010110000000000010000000000000000000000	1.239
520	0000100010010000000000010000000000000000000100	1.252
350	0000111000010000000000010000000000000000000000	1.973

FIG. 3. Identification of the five optimum MBIT out of a pool of 45 for the input set $\{E_{\text{exact}}(\sigma)\}$. (a) Development of S_{cv} as a function of GA generation number for all trial CEs. Persistent solutions are locked out after 50 generations. The optimum combination of MBIT is locked out in generation 97. (b) List of the first six locked-out "persistent" CEs, encoded as genomes.

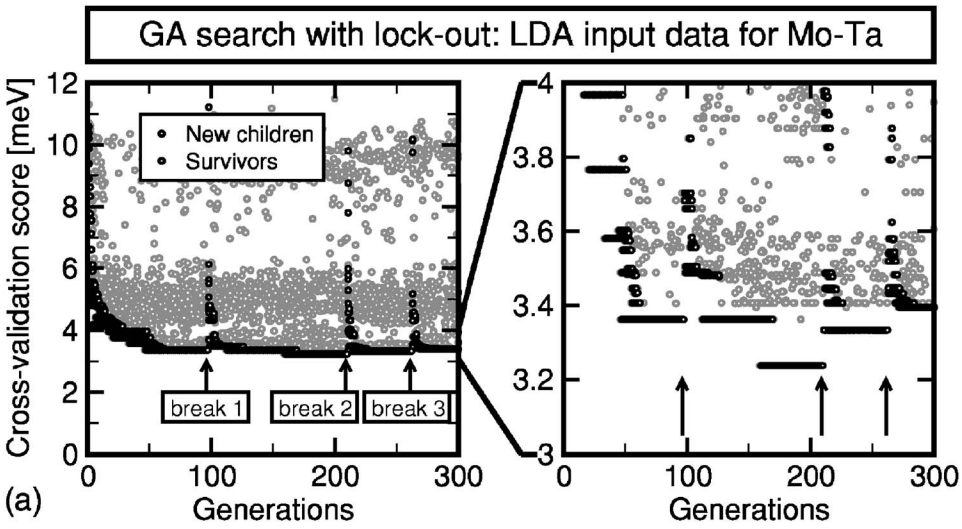
in an earlier study of the alloy system Mo-Ta.[37,38] (for details see Appendix A). This choice is advantageous because the underlying cluster expansion describes a real alloy system. In Refs. 37 and 38, the cluster expansion was constructed manually and tested thoroughly, predicting physical ground states, order-disorder transition temperatures T_c, short-range order, and the random alloy enthalpy of mixing of Mo-Ta.

Figure 3(a) shows the development of S_{cv} as a function of generation number in a typical GA run. The GA picks the optimum five MBIT out of a pool of 45 candidates (Fig. 1), using $N_{\text{pop}}=27$ trial CEs to truncate Eq. (1). The 13 fittest CEs of each generation are allowed to survive into the next generation. The mutation rate is chosen to flip one gene per newly mated child on average, meaning that the mutation probability is 1/45 to switch a particular MBIT off or on at random. Since the input energies $E_{\text{exact}}(\sigma)$ are constructed from the known interactions of Table I, the search must select these precise MBIT, with $S_{\text{cv}}=0$. This optimum solution is indeed obtained after 46 generations. To arrive at this result, only 657 individual combinations of MBIT were probed, less than 1/1000 of the total space which contains of $\binom{45}{5}\approx 1.22$ million distinct possible CEs.

After the optimum CE is identified, it persists through the subsequent iterations of the GA, and is therefore "locked out" after 96 generations. The algorithm then continues to probe the search space for a next best CE, and so forth. Figure 3(b) lists the six CE's which were locked out within 600 GA generations of this run. All six candidates share two specific MBIT, but differ in the remaining three. In terms of S_{cv}, the best solution is clearly separated from the competing possible truncations of Eq. (1). It is noteworthy that for the selected lock-out criterion (exclude persistent solutions after

TABLE I. Interaction types and (symmetry-weighted) numerical interaction values for bcc Mo-Ta according to Refs. 37 and 38, used here to generate the set of configurational energies $\{E_{\text{exact}}(\sigma)\}$.

Figure	Vertices [excl. (0,0,0)]	Numerical value (meV)
	Empty and point interaction	
$J0$		−144.7
$J1$		+12.8
	Pair interactions	
1	(0.5,0.5,0.5)	+108.1
2	(1,0,0)	−15.7
3	(1,1,0)	+23.0
4	(1.5,0.5,0.5)	−3.7
5	(1,1,1)	+12.0
6	(2,0,0)	+3.7
7	(1.5,1.5,0.5)	+6.3
8	(2,1,0)	+21.2
	Three-body interactions	
$M1$	(0.5,0.5,0.5), (1,1,0)	−3.7
$M2$	(0.5,0.5,0.5), (1.5,0.5,0.5)	−21.8
M3	(0,1,1),(1.5,0.5,0.5)	−5.2
$M4$	(1,0,0),(1,1,1)	+18.1
	Four-body interactions	
$M5$	(0.5,0.5,0.5),(1,1,0), (1.5,0.5,0.5)	−9.8



(b)

Generation:	Genome	CV score
209	0100100010010000000000010000000000000000000000	3.238
261	0100100000010000000000010000000000000000001000	3.334
96	0000111000010000000000010000000000000000000000	3.362
388	0100110000010000000000010000000000000000000000	3.363
321	0100100000010000000000010001000000000000000000	3.395
581	0100100000010000000100010000000000000000000000	3.405
444	0000110000010000000000010000000000000000000100	3.429
504	0000110001010000000000010000000000000000000000	3.469

FIG. 4. Identification of the five optimum MBIT out of a pool of 45 for $\{E_{\mathrm{LDA}}(\sigma)\}$ of bcc Mo -Ta. (a) Development of S_{cv} of all trial solutions as a function of GA generation number. Persistent solutions are locked out after 50 generations. The optimum CE is found second, in generation 209, after a persistent local minimum has been removed. (b) List of the first eight locked-out "persistent" CEs, encoded as genomes.

50 generations), the six optimum CE's are not found precisely in order of increasing S_{cv}. Without locking out, the third and fourth identified CEs (in generations 264 and 350) could have significantly delayed the algorithm's convergence to the actual third-best solution (S_{cv}=1.09 meV, locked out in generation 443).

Next, we show that the GA performs just as well for actual LDA input data for Mo-Ta. The input set $\{E_{\mathrm{LDA}}(\sigma)\}$ consists of the 56 structures used in Refs. 37 and 38, and is described in Appendix B. To construct the optimum CE for $\{E_{\mathrm{LDA}}(\sigma)\}$, we again pick the five MBIT out of the pool of 45 candidates (Fig. 1), using the same basic GA settings as for $\{E_{\mathrm{exact}}(\sigma)\}$. The GA run shown in Fig. 4(a) demonstrates a case where the algorithm is first trapped in a local minimum, which is then locked out after 96 generations total (50 generations after it first appears) according to criterion in Sec. III

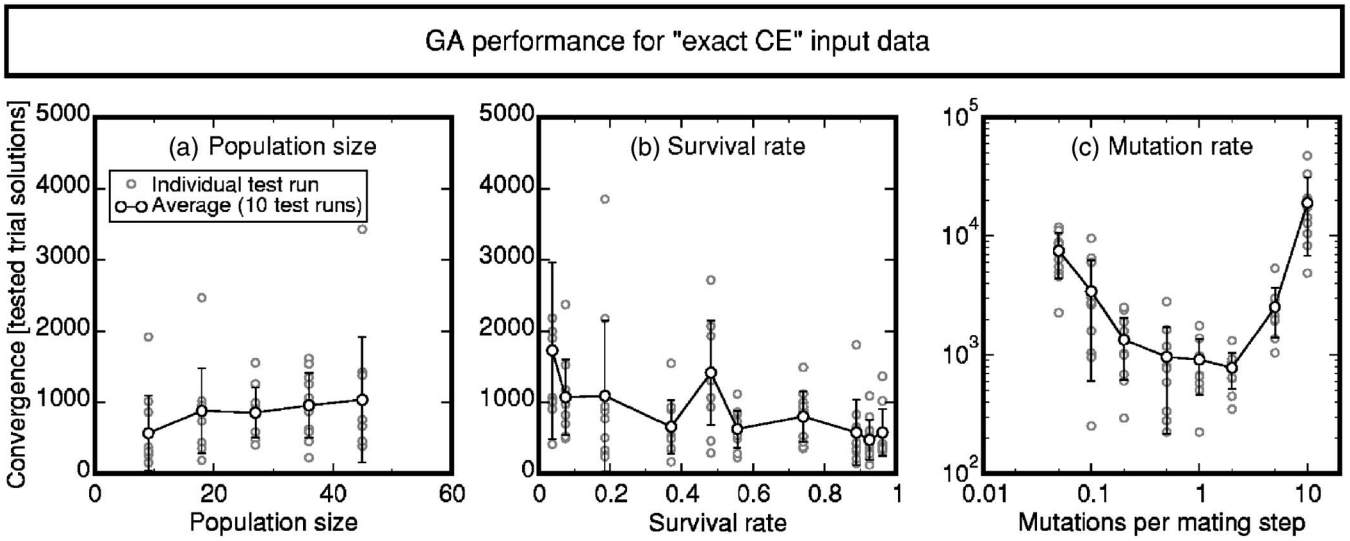


FIG. 5. Number of trial CEs evaluated by the GA as a function of (a) population size of each generation, (b) survival rate between two generations, and (c) mutation rate in the mating step. Ten different GA runs were evaluated in each step (open symbols). The full line represents the average number trial solutions for each setting of GA parameters, including their standard deviation (error bars).

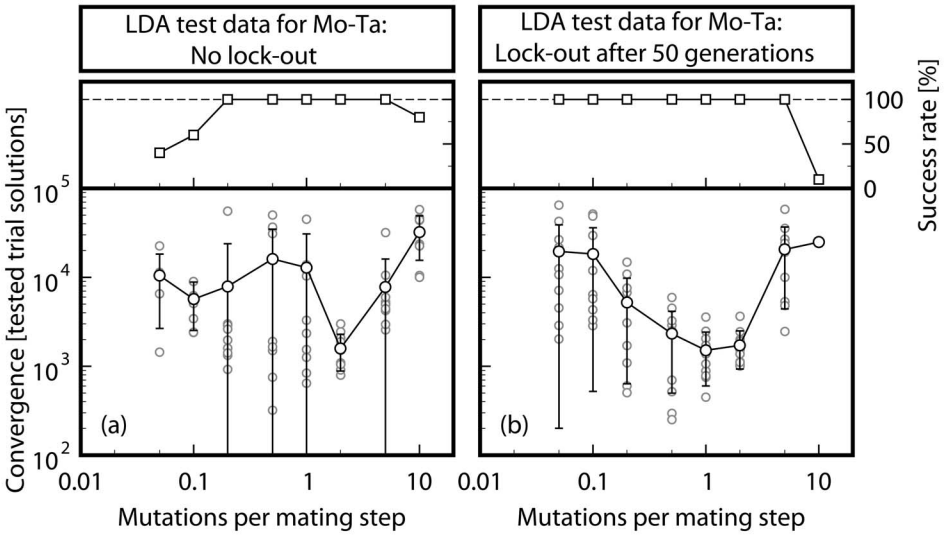


FIG. 6. (a) Number of trial CEs evaluated by the GA as a function of mutation rate in the mating step to find the optimum five MBIT for $\{E_{\mathrm{LDA}}(\sigma)\}$. All settings are the same as for Fig. 5(c). (b) Same input data and parameters, except persistent candidate CEs are now locked out after 50 generations without improvement.

F above. (That these numbers are the same as for the first lock-out in Fig. 3 is pure coincidence.) The actual optimum solution is found second, after 159 generations, and locked out in generation 209. Compared to the total space of $\binom{45}{5}$ ≈ 1.22 million possibilities, again only $\approx 1/1000$ of the solution space was explored.

Figure 4(b) shows the list of locked-out trial CEs after 600 generations. Since, for actual LDA input data, there is no exact solution, the optimum selected individuals are much closer together in terms of S_{cv} than in the case of $\{E_{\mathrm{exact}}(\sigma)\}$ (Fig. 3). Still, the best solution is relatively clearly separated from the competing possible CEs. Indeed, it coincides with the result of our previous, much more tedious search "by hand"[38] (Table I), yet this time with certainty that no correlations between the MBIT are missed. All further locked-out CEs share three of the optimum MBIT. It is instructive to note that the nonoptimal solution which was locked out first differs from the actual optimum in *both* remaining MBIT. Its relative persistence is thus explained by the lower probability of a correlated switch of two MBIT, required to reach the actual best solution.

B. Optimizing the algorithm's efficiency

We examine the impact of the three major scalable parameters, population size, survival rate, and mutation rate, on the convergence efficiency of our algorithm. This first set of tests is based on the input set $\{E_{\mathrm{exact}}(\sigma)\}$ as described in Appendix A. For clarity, the lock-out criterion was not applied when generating these results.

Figures 5(a)–5(c) show the performance of the GA as a function of (a) population size N_{pop}, (b) survival rate r_{s}, and (c) mutation rate in the mating step. As we aim to visualize the actual computational effort, we plot the total number of trial CEs that the GA explored before the solution was found, i.e., the number of child CEs per generation $N_{\mathrm{pop}} \times r_{\mathrm{s}}$ multiplied by the number of generations needed to find the correct solution. For each choice of parameters, ten different GA runs were evaluated, shown as small open circles. Also plotted are their averages and standard deviations, represented by larger symbols including lines and error bars. It is obvious that the scatter of results is relatively large, but several trends are nevertheless apparent.

(a) *The effect of population size.* We use a probability of one mutation on average per newly mated child and $r_{\mathrm{s}} = 1/2$. The impact of N_{pop} on the overall computational effort of our algorithm is relatively small. As we increase N_{pop}, the number of new trial CEs *per generation* increases. However, the average number of generations needed to find the actual

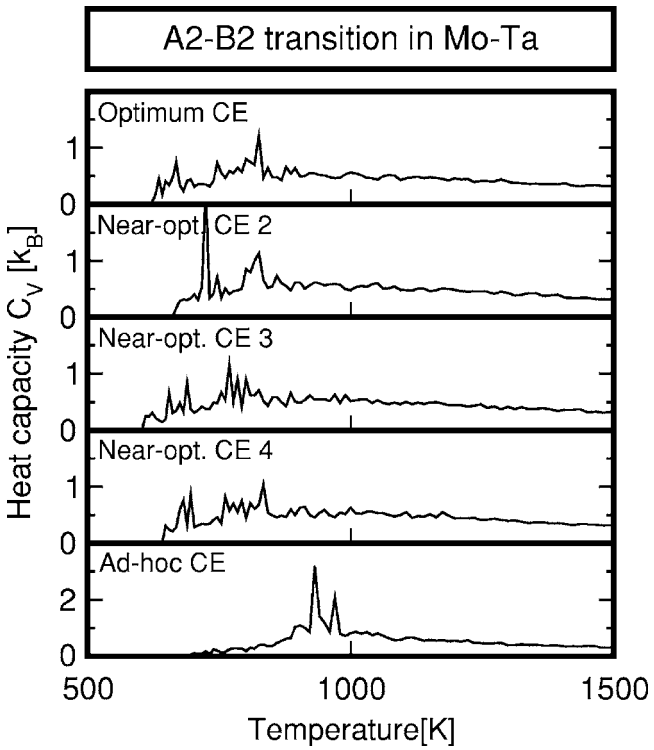


FIG. 7. Configurational heat capacity $C_v(T)$ from Monte Carlo simulations of the *A*2-*B*2 phase transition in $Mo_{0.5}Ta_{0.5}$ by stepwise cooling. $C_v(T)$ is shown for the optimum CE selected in Fig. 4(b), three near-optimal CE candidates, and an *ad hoc* hierarchy-based expansion which contains the five shortest-ranged MBIT of Fig. 1.

solution decreases almost as fast with N_{pop}, leaving the total number of required trial CEs almost constant. So, while it seems slightly beneficial to sample fewer rather than more new trial solutions per generation, the overall effect is not dramatic.

(b) *The effect of the survival rate.* We set a probability of one mutation on average per newly mated child, and $N_{pop}=27$. The scatter of results is again larger than any actual trend, but it does seem that high survival rates (down to only one newly created CE per generation) give somewhat better results. The GA then makes the most efficient use of the previously acquired genetic information, since each child is generated almost exclusively from previously accepted survivors, rather than from a parent which was itself a child in the preceding generation, with potentially high S_{cv}.

(c) *The effect of the mutation rate*. This governs the child-mating process, and shows the clearly strongest effect of all the adjustable quantities. Tested for $N_{pop}=27$ and $r_s=13/27$, a logarithmic plot is needed to display the full results. It is evident that the fastest results are reached for 0.5–2 mutations per mating step. Lower mutation rates slow down the algorithm because not enough fresh genetic information is introduced, causing the algorithm to dwell in local minima over many generations. In contrast, mutation rates that are too high lead to an almost random search pattern, drowning out the useful information that the algorithm has already collected in preceding search generations.

For the simple test case only around 1/1000 of the available search space must be scanned to find the best possible CE. While the algorithm does not fail for any of the tested settings, an appropriate mutation rate is the key to its efficient functioning.

C. Impact of the lock-out criterion

Figure 6 shows the performance (number of trial CEs required to find the actual optimum set of MBIT, as previously determined in a tedious search by hand[38]) of the GA as a function of mutation rate for actual LDA data $\{E_{LDA}(\sigma)\}$ of Mo-Ta (Appendix B). In Fig. 6(a), all settings are exactly the same as for Fig. 5(c); in particular, persistent solutions were never locked out. Again, we averaged over ten GA runs for each setting, and also show the scatter of individual runs. The scatter of the number of required trial CE evaluations is much larger for $\{E_{LDA}(\sigma)\}$ than for $\{E_{exact}(\sigma)\}$ in Fig. 5(c). Moreover, a minimum develops only at two mutations per child genome on average, which appears as a sharp spike. The reason for this behavior can also be seen in Fig. 6(a). A number of individual test runs shows exactly the same behavior as observed for $\{E_{exact}(\sigma)\}$ in Fig. 5(c), namely a parabolalike distribution with a minimum around 0.5–2 mutations per genome. However, another group of runs takes disproportionately longer (data points between 10 000 and 100 000 trial solutions), driving up both the average and the standard deviation of our search. The origin of this population of outliers is that, in these cases, the GA encounters a local optimum CE that differs by *several* MBIT from the actual one. The actual optimum can now only be reached by several random mutations in the same step, which must all be simultaneously correct. The probability for this correlated switch is low, and the algorithm remains trapped for some time. This problem is particularly grave for small mutation rates, where a large number of test runs do not find the correct CE at all within 5000 generations, as shown by the success rate in the upper panel of Fig. 6(a).

This behavior is mended by the "lock-out" strategy described in the preceding section. In Fig. 6(b), the lock-out threshold is set to 50 generations, with otherwise the same parameters as Fig. 6(a). The success of this strategy is convincing; the outlier population is eliminated entirely, and the qualitative behavior is now the same as that of Fig. 5. In particular, the success rate is now 100% even in the previously difficult cases of very low mutation rates. It is also worth noting that the lock-out strategy does not improve the

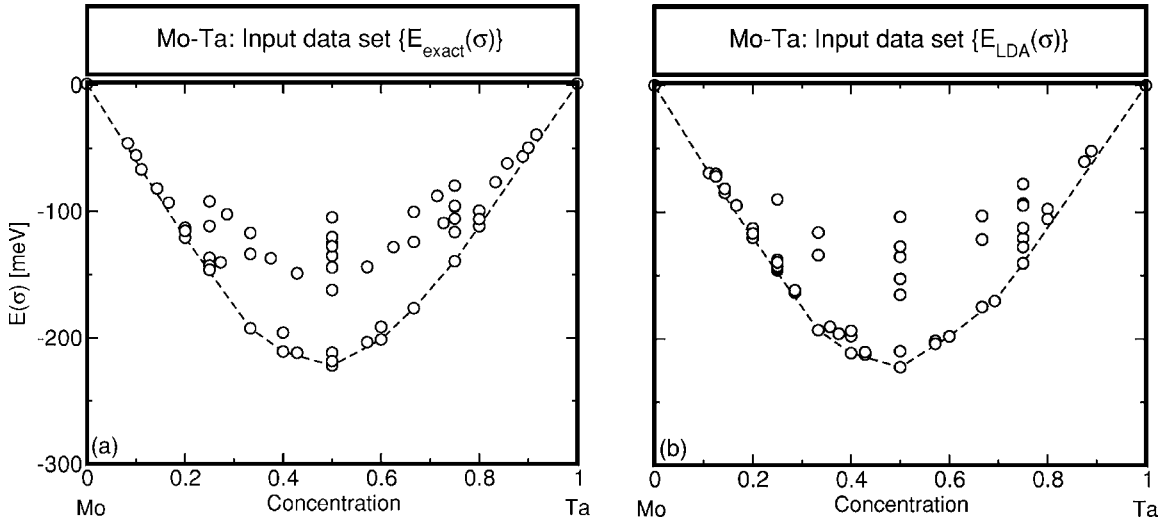


FIG. 8. Physical data $E(\sigma)$ as a function of composition x of each configuration σ, based on which the GA selects the optimum MBIT in the present work. Dashed lines serve as guides to the eye. (a) $E_{exact}(\sigma)$ for 60 bcc configurations, calculated directly from Eq. (1) using the interaction parameters tabulated in Table I. (b) LDA-calculated input set $E_{LDA}(\sigma)=\Delta H_{LDA}(\sigma)-E_{CS}(\sigma)$ for 56 configurations of bcc Mo-Ta.

TABLE II. $E_{\text{exact}}(\sigma)$ for 60 input configurations, generated from Eq. (1) using the interactions of Table I. See text for details of the structure notation used.

Composition	Structure	$E(\sigma)$ (meV)	Composition	Structure	$E(\sigma)$ (meV)
Mo	$A2$	0.9	MoTa	(310) A_2B_2 SL	−211.7
$Mo_{11}Ta$	(211) $A_{11}B$ SL	−46.1		(311) A_3B_3 SL	−218.4
Mo_9Ta	(521) A_9B SL	−55.6		(221) $A_4B_3A_2B_3$ SL	−120.4
Mo_8Ta	"A_8B"	−66.8		(221) $A_3B_2A_3B_4$ SL	−127.7
Mo_6Ta	(111) A_6B SL	−81.9		SQS-16	−144.3
Mo_5Ta	(332) A_5B SL	−93.2	Mo_3Ta_4	(111) $A_2B(AB)_2$ SL	−203.4
Mo_4Ta	(100) A_4B SL	−120.8		SQS-14	−143.9
	(310) A_4B SL	−113.0	Mo_2Ta_3	(100) A_2BAB SL	−201.3
	(332) A_6BA_2B	−115.5		(111) A_2BAB	−191.3
Mo_3Ta	$D0_3$	−139.9	Mo_3Ta_5	"A_5B_3"	−128.4
	$L6_0$	−136.8	$MoTa_2$	$C11_b$	−176.5
	(100) A_3B SL	−143.0		(110) $A-2B$ SL	−100.5
	(110) A_3B SL	−92.2		(111) A_2B SL	−124.3
	"A_4B_{12}"	−146.2	Mo_2Ta_5	(111) A_5B_2 SL	−87.8
	SQS-16	−111.7	Mo_3Ta_8	(111) $(A_3B)_2A_2B$ SL	−109.3
Mo_8Ta_3	(111) $(A_3B)_2A_2B$ SL	−140.4	$MoTa_3$	$D0_3$	−96.2
Mo_5Ta_2	(111) A_5B_2 SL	−102.4		$L6_0$	−95.7
Mo_2Ta	$C11_b$	−192.4		(100) A_3B SL	−116.4
	(110) A_2B SL	−117.1		(110) A_3B SL	−79.7
	(111) A_2B SL	−133.8		"A_4B_{12}"	−139.5
Mo_5Ta_3	"A_5B_3"	−137.2		SQS-16	−105.8
Mo_3Ta_2	(100) A_2BAB SL	−210.8	$MoTa_4$	(100) A_4B SL	−99.6
	(111) A_2BAB SL	−195.8		(310) A_4B SL	−111.8
Mo_4Ta_3	(111) $A_2B(AB)_2$ SL	−211.8		(332) A_6BA_2B	−106.1
	SQS-14	−148.9	$MoTa_5$	(332) A_5B SL	−76.9
MoTa	A_1	−135.2	$MoTa_6$	(111) A_6B SL	−62.0
	$B2$	−221.9	$MoTa_8$	"A_8B"	−56.5
	$B11$	−162.1	$MoTa_9$	(521) A_9B SL	−49.6
	$B32$	−125.6	$MoTa_{11}$	(211) $A_{11}B$ SL	−39.3
	(110) A_2B_2 SL	−104.7	Ta	$A2$	1.1

behavior for unreasonably high mutation rates [e.g., 10 mutations per genome in Fig. 6(b)]. Here, the convergence is slowed down not by trapping in local minima but by the noise of random mutations drowning out the valuable genetic information—the lock-out solution does not apply. For reasonable mutation rates, the algorithm is now completely reliable.

V. PHYSICAL IMPACT

We have shown how a GA can be employed to solve a decisive step in the construction of a CE Hamiltonian of the form Eq. (1). Based on a set of sufficiently many configurational energies $\{E(\sigma)\}$, identify those interaction types which promise the greatest power to predict energies of further, as yet unknown energies for the same alloy system. During the construction process of a CE, one may test predictions made with these MBIT after the fact, and increase the number of structures σ for which first-principles input is available. A completed CE then provides the ability to assess the energies of literally millions of configurations within minutes, enabling both the identification of ground-state structures by exhaustive search,[25] and the evaluation of configurational averages, e.g., in Monte Carlo simulations,[26,27] for finite-T thermodynamics.

In addition, the rigorous application of the lock-out criterion provides physical information beyond that contained in the optimum set of MBIT alone. With a rigorous list of near-optimal cluster expansions, it is now possible to assess how sensitive the *physical* target quantities of a cluster expansion are against the final choice of MBIT, i.e., how reliable the information is that we can extract from a given set of input structures $\{\sigma\}_{\text{input}}$. As an example, we examine the $A2$-$B2$ phase transition in bcc $Mo_{0.5}Ta_{0.5}$ using canonical Monte Carlo simulations (cell size: $16\times16\times16$, 4000 flips per lattice site and T step). Figure 7 shows the development of the configurational heat capacity C_v with decreasing simulation temperature for the optimum selected set of MBIT in Fig.

TABLE III. Full input set $\{E_{\mathrm{LDA}}(\sigma)\}$ [Eq. (2)] for 56 bcc Mo-Ta input configurations. See Ref. 38 for details.

Composition	Structure	$E(\sigma)$ (meV)	Composition	Structure	$E(\sigma)$ (meV)
Mo	$A2$	0.0	MoTa	A_1	−135.4
Mo_8Ta	"A_8B"	−69.2		$B2$	−222.4
Mo_7Ta	(210) A_7B SL	−69.8		$B11$	−165.1
	"A_7B"	−71.9		$B32$	−127.3
Mo_6Ta	(100) A_6B SL	−84.9		(110) A_2B_2 SL	−103.8
	(111) A_6B SL	−81.6		(310) A_2B_2 SL	−209.8
Mo_5Ta	(433) A_8BA_2B SL	−94.8		"A_8B_8"	−152.6
Mo_4Ta	(111) A_4B SL	−112.9	Mo_3Ta_4	(100) $A_2B(AB)_2$ SL	−201.6
	(100) A_4B SL	−120.4		(111) $A_2B(AB)_2$ SL	−203.8
	(310) A_4B SL	−116.9	Mo_2Ta_3	(100) A_2BAB SL	−198.0
Mo_3Ta	$D0_3$	−140.3	$MoTa_2$	$C11_b$	−174.8
	$L6_0$	−140.8		(110) AB_2 SL	−103.0
	(100) A_3B SL	−145.8		(111) AB_2 SL	−121.7
	(110) A_3B SL	−90.0	Mo_4Ta_9	"A_4B_9"	−170.1
	(310) A_3B SL	−144.1	$MoTa_3$	$D0_3$	−93.4
	"$A_{12}B_4$-I"	−137.8		$L6_0$	−94.9
	"A_4B_{12}"	−142.6		(100) AB_3 SL	−120.7
	"$A_{12}B_4$-II"	−139.7		(110) AB_3 SL	−77.9
Mo_5Ta_2	(100) A_3BA_2B SL	−163.3		(310) AB_3 SL	−127.7
	(111) A_4BAB SL	−161.6		"A_4B_{12}"	−140.5
Mo_2Ta	$C11_b$	−193.1		"$A_{12}B_4$-II"	−112.6
	(110) A_2B SL	−116.2	$MoTa_4$	(100) A_4B SL	−97.6
	(111) A_2B SL	−134.1		(310) A_4B SL	−105.5
Mo_9Ta_5	(710) $A_4B_3A_4BAB$ SL	−190.4	$MoTa_7$	(210) A_7B SL	−60.2
Mo_5Ta_3	(210) $A_3B(AB)_2$ SL	−195.9	$MoTa_8$	"A_8B"	−51.9
Mo_3Ta_2	(210) $A_3B(AB)_3$ SL	−197.8	Ta	$A2$	0.0
	(111) A_2BAB SL	−193.6			
	(100) A_2BAB SL	−211.3			
Mo_4Ta_3	(100) $A_2B(AB)_2$ SL	−212.4			
	(111) $A_2B(AB)_2$ SL	−210.4			

4(b), and the three best near-optimal candidates of Fig. 4(b). As a contrast, the result for an *ad hoc* hierarchy-based CE is also shown; this CE also contains five MBIT, but they are now the four shortest-ranged three-body interaction types and the shortest-ranged four-body interaction type of Fig. 1. As shown in Ref. 38 for the optimum CE, the $A2$-$B2$ transition occurs for $T_c \approx 600-1000$ K. $C_v(T)$ is quantitatively very similar to the optimum CE for all three near-optimal CE's, as expected at the end of a well-converged CE construction process, which is based on a large enough input database $\{\sigma\}_{\mathrm{input}}$. In contrast, $C_v(T)$ from the shorter-ranged *ad hoc* CE differs clearly from the other four curves, and would falsely suggest a clearly higher T_c than all others, close to 1000 K. However, this *ad hoc* CE is safely ruled out by the GA, since it is characterized by $S_{cv} \approx 7.0$ meV, more than twice the prediction error estimated for the GA-determined near-optimal MBIT combinations.

VI. CONCLUSION

We show how a genetic algorithm removes most human guesswork from the construction of a cluster expansion, where otherwise a select few combinations of MBIT (e.g., the shortest) would have to be favored over millions of other possible combinations by some intuition. The algorithm converges fast both for the test case where the correct solution is known analytically, and for realistic first-principles input data to a cluster expansion. The algorithm is easy to use, since its performance is almost exclusively controlled by the mutation rate alone, and it is robust against getting stuck in apparent local optima by strictly "locking out" persistent solutions. The resulting list of near-optimal solutions can be used to verify directly the reliability of all CE-predicted physical alloy properties (ground states, phase transitions, short-range order). The procedure is not restricted to the cluster expansion method which we emphasize here, and we

expect the same benefits in the construction of any general model Hamiltonian where a system-dependent choice of parameter types must be made.

ACKNOWLEDGMENTS

The following support is gratefully acknowledged: NREL financial support by contract DOE-SC-BES-DMS; M.J.W. and G.L.W.H. supported through the Intramural Grant Program at Northern Arizona, and the NSF through DMR-0224183. M.J.W. is also grateful for partial funding from Research Corporation, CC5944.

APPENDIX A: INPUT SET $\{E_{\mathrm{exact}}(\sigma)\}$ FOR MA-TA: CONFIGURATIONS AND ENERGIES

To test our algorithm using an input database for which an exact solution is known, we selected 60 bcc-based configurations σ. We calculated $\{E_{\mathrm{exact}}(\sigma)\}$ for each σ according to Eq. (1) using the interactions of Table I, which were found to describe the Mo-Ta alloy system in Refs. 37 and 38. It is evident that the MBIT do not follow from a simple scheme of selection among the 45 interactions displayed in Fig. 1: they include four three-body-figures, one of them extending to the fifth-nearest-neighbor vertex (1,1,1), and one four-body figure with a fourth-nearest-neighbor vertex (1.5,0.5,0.5). In a hierarchy-based approach, this choice would mandate a large number of additional unrelated figures.

The distribution of $\{E_{\mathrm{exact}}(\sigma)\}$ is shown in Fig. 8(a) as a function of the atomic concentration x of each configuration σ. There is some energetic asymmetry with regard to equiatomic composition, with lower $E(\sigma)$ towards the Mo-rich side, and it is precisely this asymmetry which is captured by three-body MBIT (pair interactions alone would produce a symmetric distribution of configurations). The chosen input configurations include the bcc configurations of elemental Mo and Ta, the "usual suspect" configurations B2 MoTa, B32 Mo_2Ta_2, $D0_3$ Mo_3Ta and $MoTa_3$, and $C11_b$ Mo_2Ta and $MoTa_2$, as well as 14 other structures with four or fewer atoms per unit cell, and five special quasirandom structures with 14 or 16 atoms per unit cell (as described in the appendix of Ref. 38). The remaining 33 structures are all relatively low in energy, spanning unit cell sizes between 5 and 16 atoms across a broad range of intermediate concentration values; a full listing is given in Table II. Wherever possible, the structures are described in a short way as superlattices (SLs) of pure atomic planes (e.g., the "(100) A_2BAB SL" is a sequence of two pure (100) planes of element A, followed by one pure B plane, another A and another B plane). Where such a notation is not possible, a description of the structure is referred to Ref. 38. There is one structure which neither fits a superlattice notation nor has been described previously—this is the structure labeled "A_5B_3." It is a sequence of three mixed (100) planes of $c(2\times2)$ type AB occupation, followed by one plane of pure A.

APPENDIX B: INPUT SET $\{E_{\mathrm{LDA}}(\sigma)\}$ FOR MO-TA: CONFIGURATIONS AND ENERGIES

A description of all Mo-Ta input structures σ and a listing of their formation enthalpies $\Delta H_{\mathrm{LDA}}(\sigma)$ can be found in Ref. 38. In the application of the GA above, we do not use $\Delta H_{\mathrm{LDA}}(\sigma)$ directly, but rather $E_{\mathrm{LDA}}(\sigma)=\Delta H_{\mathrm{LDA}}(\sigma)-E_{\mathrm{CS}}(\sigma)$ [Eq. (2)]. $E_{\mathrm{CS}}(\sigma)$ denotes the constituent strain energy, calculated according to Eq. (5) and Fig. 5 of Ref. 38. We tabulate $E_{\mathrm{LDA}}(\sigma)$ for all 56 Mo-Ta input structures in Table III. $E_{\mathrm{LDA}}(\sigma)$ as a function of a configuration's concentration x is also displayed in Fig. 8(b).

*Present address: Abteilung Theorie, Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, 14195 Berlin-Dahlem, Germany.

[1] J. M. Sanchez, F. Ducastelle, and D Gratias, Physica A **128**, 334 (1984).
[2] A. Zunger, *Statics and Dynamics of Alloy Phase Transformations*, edited by P. E. A. Turchi and A. Gonis (Plenum Press, New York, 1994), pp. 361–419.
[3] D de Fontaine, Solid State Phys. **47**, 33 (1994).
[4] A. van de Walle and G. Ceder, J. Phase Equilib. **23**, 348 (2002).
[5] A. Zunger, L. G. Wang, G. L. W. Hart, and M. Sanati, Modell. Simul. Mater. Sci. Eng. **10**, 685 (2002).
[6] D. B. Laks, L. G. Ferreira, S. Froyen, and A. Zunger, Phys. Rev. B **46**, 12587 (1992).
[7] M. H. F. Sluiter, K. Esfarjani, and Y. Kawazoe, Phys. Rev. Lett. **75**, 3142 (1995).
[8] C. Berne, M. Sluiter, Y. Kawazoe, T. Hansen, and A. Pasturel, Phys. Rev. B **64**, 144103 (2001).
[9] C. Berne, M. Sluiter, Y. Kawazoe, and A. Pasturel, J. Phys.: Condens. Matter **13**, 9433 (2001).
[10] R. Osório, S. Froyen, and A. Zunger, Phys. Rev. B **43**, 14055 (1991).
[11] R. Osório, S. Froyen, and A. Zunger, Solid State Commun. **78**, 249 (1991).
[12] R. Osório, Z.-W. Lu, S.-H. Wei, and A. Zunger, Phys. Rev. B **47**, 9985 (1993).
[13] C. Wolverton and D. de Fontaine, Phys. Rev. B **49**, 8627 (1994).
[14] G. Rubin and A. Finel, J. Phys.: Condens. Matter **7**, 3139 (1995).
[15] R. McCormack and D. de Fontaine, Phys. Rev. B **54**, 9746 (1996).
[16] R. Osório, J. E. Bernard, S. Froyen, and A. Zunger, Phys. Rev. B **45**, 11173 (1992).
[17] S. Froyen, J. E. Bernard, R. Osório, and A. Zunger, Phys. Scr., T **T45**, 272 (1992).
[18] R. Drautz, H. Reichert, M. Fähnle, H. Dosch, and J. M. Sanchez, Phys. Rev. Lett. **87**, 236102 (2001).
[19] R. Drautz, R. Singer, and M. Fähnle, Phys. Rev. B **67**, 035418 (2003).
[20] M. H. F. Sluiter and Y. Kawazoe, Phys. Rev. B **68**, 085410 (2003).

[21] S. Müller, J. Phys.: Condens. Matter **15**, R1429 (2003).
[22] R. Singer, R. Drautz, and M. Fähnle, Surf. Sci. **559**, 241 (2004).
[23] H. R. Tang, A. van der Ven, and B. L. Trout, Mol. Phys. **102**, 273 (2004).
[24] H. Tang, A. Van der Ven, and B. L. Trout, Phys. Rev. B **70**, 045420 (2004).
[25] L. Ferreira, S.-H. Wei, and A. Zunger, Int. J. Supercomput. Appl. **5**, 34 (1991).
[26] Z. W. Lu, D. B. Laks, S.-H. Wei, and A. Zunger, Phys. Rev. B **50**, 6642 (1994).
[27] A. van de Walle and M. Asta, Modell. Simul. Mater. Sci. Eng. **10**, 521 (2002).
[28] S.-H. Wei, A. A. Mbaye, L. G. Ferreira, and A. Zunger, Phys. Rev. B **36**, 4163 (1987).
[29] J. E. Bernard, L. G. Ferreira, S.-H. Wei, and A. Zunger, Phys. Rev. B **38**, 6338 (1988).
[30] S.-H. Wei, L. G. Ferreira, and A. Zunger, Phys. Rev. B **41**, 8240 (1990).
[31] S.-H. Wei, L. G. Ferreira, and A. Zunger, Phys. Rev. B **45**, 2533 (1992).
[32] V. Ozolins, C. Wolverton, and A. Zunger, Phys. Rev. B **57**, 6427 (1998).
[33] S. Müller, L.-W. Wang, A. Zunger, and C. Wolverton, Phys. Rev. B **60**, 16448 (1999).
[34] G. L. W. Hart and A. Zunger, Phys. Rev. Lett. **87**, 275505 (2001).
[35] S. Müller and A. Zunger, Phys. Rev. B **63**, 094204 (2001).
[36] M. Sanati, G. L. W. Hart, and A. Zunger, Phys. Rev. B **68**, 155210 (2003).
[37] V. Blum and A. Zunger, Phys. Rev. B **69**, 020103(R) (2004).
[38] V. Blum and A. Zunger, Phys. Rev. B **70**, 155108 (2004).
[39] L. G. Ferreira, A. A. Mbaye, and A. Zunger, Phys. Rev. B **37**, 10547 (1988).
[40] L. G. Ferreira, S.-H. Wei, and A. Zunger, Phys. Rev. B **40**, 3197 (1989).
[41] J. W. D. Connolly and A. R. Williams, Phys. Rev. B **27**, R5169 (1983).
[42] F. Ducastelle and F. Gautier, J. Phys. F: Met. Phys. **6**, 2039 (1976).
[43] S. de Gironcoli, P. Giannozzi, and S. Baroni, Phys. Rev. Lett. **66**, 2116 (1991).
[44] J. B. Staunton, D. D. Johnson, and F. J. Pinski, Phys. Rev. B **50**, 1450 (1994).
[45] A. V. Ruban, S. Shallcross, S. I. Simak, and H. L. Skriver, Phys. Rev. B **70**, 125115 (2004).
[46] R. Drautz, M. Fähnle, and J. M. Sanchez, J. Phys.: Condens. Matter **16**, 3843 (2004).
[47] N. Marzari, S. de Gironcoli, and S. Baroni, Phys. Rev. Lett. **72**, 4001 (1994).
[48] A. Gonis, X.-G. Zhang, A. J. Freeman, P. Turchi, G. M. Stocks, and D. M. Nicholson, Phys. Rev. B **36**, 4630 (1987).
[49] A. V. Ruban and H. L. Skriver, Phys. Rev. B **66**, 024201 (2002).
[50] A. V. Ruban, S. I. Simak, P. A. Korzhavyi, and H. L. Skriver, Phys. Rev. B **66**, 024202 (2002).
[51] F. J. Pinski, J. B. Staunton, and D. D. Johnson, Phys. Rev. B **57**, 15177 (1998).
[52] R. Magri, S.-H. Wei, and A. Zunger, Phys. Rev. B **42**, 11388 (1990).
[53] A. V. Ruban, S. I. Simak, S. Shallcross, and H. L. Skriver, Phys. Rev. B **67**, 214302 (2003).
[54] R. Monnier, Philos. Mag. B **75**, 67 (1997).
[55] D. D. Johnson, A. V. Smirnov, J. B. Staunton, F. J. Pinski, and W. A. Shelton, Phys. Rev. B **62**, R11917 (2000).
[56] P. E. A. Turchi, A. Gonis, V. Drchal, and J. Kudrnovsky, Phys. Rev. B **64**, 085112 (2001).
[57] P. E. A. Turchi, V. Drchal, J. Kudrnovsky, C. Colinet, L. Kaufman, and Z.-K. Liu, Phys. Rev. B **71**, 094206 (2005).
[58] G. L. W. Hart, V. Blum, M. Walorski, and A. Zunger, Nat. Mater. **4**, 391 (2005).
[59] V. Blum and A. Zunger, Phys. Rev. B **72**, 020104(R) (2005).
[60] R. Hultgren, P. Desai, D. Hawkins, M. Gleiser, and K. Kelley, *Selected Values of the Thermodynamic Properties of Binary Alloys* (Am. Soc. of Metals, Metals Park, OH, 1973).
[61] *Pearson's Handbook of Crystallographic Data for Intermetallic Phases*, 2nd ed., edited by P. Villars and L. Calvert (ASM International, Materials Park, OH, 1991).
[62] *Phase Equilibria, Crystallographic and Thermodynamic data of Binary Alloys of Landolt-Börnstein, New Series, Group IV*, edited by B. Predel, Vol. 5H (Springer, Berlin, 1997).
[63] D. F. Styer, Phys. Rev. B **32**, 393 (1985).
[64] R. Kikuchi, J. Chem. Phys. **60**, 1071 (1974).
[65] A. E. Carlsson, Phys. Rev. B **35**, 4858 (1987).
[66] A. E. Carlsson and J. M. Sanchez, Solid State Commun. **65**, 527 (1988).
[67] A. E. Carlsson, Phys. Rev. B **40**, 912 (1989).
[68] T. Mohri, K. Terakura, S. Tazikawa, and J. M. Sanchez, Acta Metall. Mater. **39**, 493 (1991).
[69] N. A. Zarkevich and D. D. Johnson, Phys. Rev. Lett. **92**, 255702 (2004).
[70] G. D. Garbulsky and G. Ceder, Phys. Rev. B **51**, 67 (1995).
[71] R. Drautz, A. Diáz-Ortiz, M. Fähnle, and H. Dosch, Phys. Rev. Lett. **93**, 067202 (2004).
[72] J. Shao, J. Am. Stat. Assoc. **88**, 486 (1993).
[73] K. Baumann, TrAC, Trends Anal. Chem. **22**, 395 (2003).
[74] Z. Michalewicz and D. B. Fogel, *How to Solve It: Modern Heuristics* (Springer-Verlag, Berlin, 2000).
[75] D. M. Deaven and K. M. Ho, Phys. Rev. Lett. **75**, 288 (1995).
[76] K. M. Ho, A. Shvartzburg, B. C. Pan, Z. Y. Lu, C. Z. Wang, J. Wacker, J. Fye, and M. Jarrold, Nature **392**, 582 (1998).
[77] G. H. Johannesson, T. Bligaard, A. V. Ruban, H. L. Skriver, K. W. Jacobsen, and J. K. Norskov, Phys. Rev. Lett. **88**, 255506 (2002).
[78] D. P. Stucke and V. H. Crespi, Nano Lett. **3**, 1183 (2003).
[79] G. Klimeck and R. C. Bowen, Superlattices Microstruct. **27**, 77 (2000).