[illegible]

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:

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.

[illegible] $\boldsymbol{J}$ [illegible] $\boldsymbol{E}$ $(\boldsymbol{\sigma})$?

Our approach is based on the Connolly-Williams[41] suggestion to derive $\{J\}$ from a set of quantum-mechanically calculated total energies $\{E_{\text{QM}}(\sigma, V_\sigma)\}$ 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 $E_{\text{QM}}(\sigma)$

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

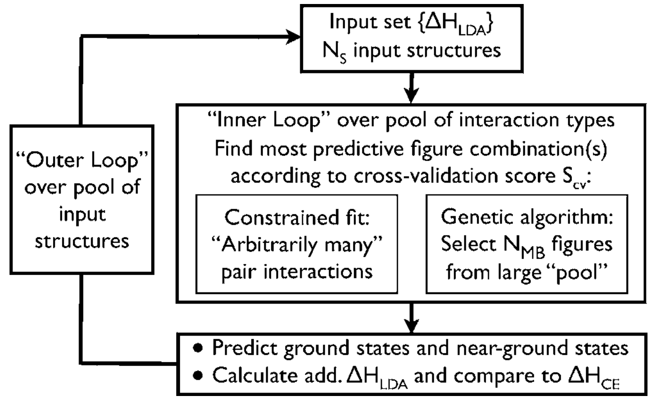
Input set {ΔH$_{LDA}$}
N$_S$ input structures
"Inner Loop" over pool of interaction types
Find most predictive figure combination(s)
according to cross-validation score S$_{cv}$:
Constrained fit:
"Arbitrarily many"
pair interactions
Genetic algorithm:
Select N$_{MB}$ figures
from large "pool"
"Outer Loop"
over pool of
input
structures
• Predict ground states and near-ground states
• Calculate add. ΔH$_{LDA}$ and compare to ΔH$_{CE}$

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

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

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,

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

(all total energies are per atom). $E_{\text{CS}}(\sigma)$ 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 $[A_{1-x}B_x]^{\text{phs}}$ on the same coherent underlying lattice is nonzero: $E_{\text{tot}}([A_{1-x}B_x]^{\text{phs}}) \neq (1-x)E_{\text{tot}}(A) - xE_{\text{tot}}(B)$, 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 $\{E_{\text{LDA}}(\sigma)\}$ 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 $E_{\text{CE}}(\sigma)$ can be compared to the known energy $E_{\text{LDA}}(\sigma)$ 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 $\{E_{\text{LDA}}(\sigma)\}$ was predicted at least twice. The average overall prediction errors from this process define

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

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 $E_{\text{LDA}}(\sigma_{\text{new}})$ are then evaluated by direct LDA calculations and compared to the predicted $E_{\text{CE}}(\sigma_{\text{new}})$, giving an objective estimate of the predictive power of each candidate cluster expansion. The newly calculated $\{E_{\text{LDA}}(\sigma_{\text{new}})\}$ 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.

.

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.

[illegible]

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 over-optimization 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.

[illegible]

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.

[illegible]

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.

[illegible]

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.

[illegible]

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 Eeinzero -22.7682-1.15T

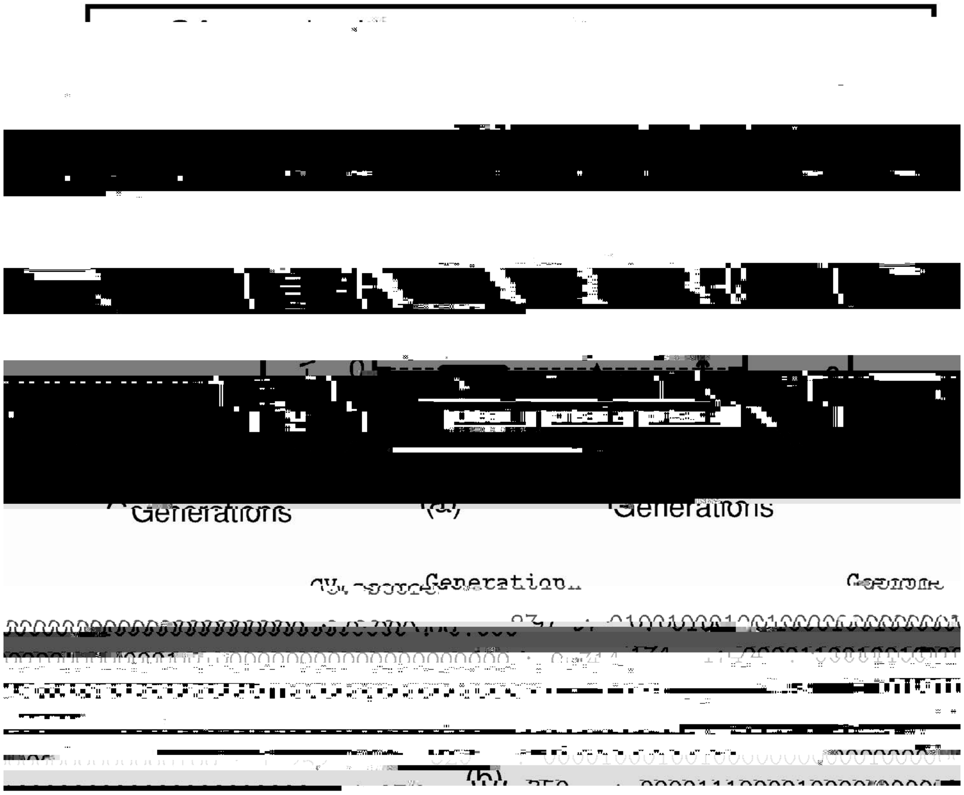

FIG. 3. Identification of the five optimum MBIT out of a pool of 45 for the input set $\{E_{\rm exact}(\sigma)\}$. (a) Development of $S_{\rm cv}$ as a function of GA generation number for all trial CEs. Persistent solutions are locked out after 50 generations. The optimum combination of MBIT is locked 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 $S_{\rm cv}$ 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 $N_{\rm pop}=27$ 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 $E_{\rm exact}(\sigma)$ are constructed from the known interactions of Table I, the search must select these precise MBIT, with $S_{\rm cv}=0$. 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 $S_{\rm cv}$, 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 $\{E_{\rm exact}(\sigma)\}$.

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

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.

[illegible]

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 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 $N_{\mathrm{pop}} = 27$. 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 $N_{\mathrm{pop}} = 27$ and $r_s = 13/27$, a logarithmic plot is needed

TABLE II. $E_{\mathrm{exact}}(\sigma)$

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.

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 $\{\sigma\}_{\mathrm{input}}$. As an example, we examine the $A2$-$B2$ phase transition in bcc $\mathrm{Mo}_{0.5}\mathrm{Ta}_{0.5}$ 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.

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

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.

[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]