Evolutionary approach for determining first-principles hamiltonians

GUS L. W. HART[1]*, VOLKER BLUM[2†], MICHAEL J. WALORSKI[3] AND ALEX ZUNGER[2]
[1]Department of Physics and Astronomy, Northern Arizona University, Flagstaff, Arizona 86011-6010, USA
[2]National Renewable Energy Laboratory, Golden, Colorado 80401, USA
[3]Computer Science Department, Northern Arizona University, Flagstaff, Arizona 86011-5600, USA
†Present address: Fritz-Haber Institut der Max-Planck-Gesellschaft, Faradayweg 4–6, 14195 Berlin, Germany
*e-mail: gus.hart@nau.edu

Modern condensed-matter theory from first principles is highly successful when applied to materials of given structure-type or restricted unit-cell size. But this approach is limited where large cells or searches over millions of structure types become necessary. To treat these with first-principles accuracy, one 'coarse-grains' the many-particle Schrödinger equation into 'model hamiltonians'[1–3] whose variables are configurational order parameters (atomic positions, spin and so on), connected by a few 'interaction parameters' obtained from a microscopic theory[3,4]. But to construct a truly quantitative model hamiltonian, one must know just which types of interaction parameters to use, from possibly 10^6–10^8 alternative selections. Here we show how genetic algorithms[5], mimicking biological evolution ('survival of the fittest'), can be used to distil reliable model hamiltonian parameters from a database of first-principles calculations. We demonstrate this for a classic dilemma[6] in solid-state physics[7], structural inorganic chemistry[8] and metallurgy[9]: how to predict the stable crystal structure of a compound given only its composition. The selection of leading parameters based on a genetic algorithm is general and easily applied to construct any other type of complex model hamiltonian from direct quantum-mechanical results.

Even for a binary A_nB_m solid described by a unit cell with N sites, there can be $\sim 2^N$ possible structures (denoted as σ in the following). A direct quantum-mechanical evaluation of the total energy, $E_{QM}(\sigma) = \langle \Psi | H(\sigma) | \Psi \rangle$ (where Ψ is the wavefunction and $H(\sigma)$ is the electronic hamiltonian), of all 2^N configurations σ in search of one that minimizes the energy is practically impossible. Yet finding the stable structure among an astronomical number of possibilities is important because structure often decides function (for example, contrast the different properties of the graphite, diamond and fullerene structures of carbon) and because our notions of bonding are based on our understanding of the stable structure[7–9]. To access configurational energetics on a larger scale, one can renormalize away the explicit electronic coordinates of $E_{QM}(\sigma)$ and formulate instead a model hamiltonian with fewer and more transparent parameters, reproducing the dependence of $E_{QM}(\sigma)$ on σ with high accuracy. An example is a cluster expansion[3,4] of the configurational energy $E(\sigma)$ in terms of effective atom–atom interactions J,

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

Here, $\hat{S}_i$ denotes the fictitious spin variable which indicates if lattice site i is occupied by an A atom (in which case $\hat{S}_i = -1$) or a B atom ($\hat{S}_i = +1$). The variable $\{\sigma\}$ denotes collectively the set of all 2^N configurational possibilities (not to be confused with the continuous positional variables). The interaction energies $\{J\}$ represent the contribution of each group of atoms to $E(\sigma)$. Figure 1a shows the smallest such 'many-body interaction types' (MBITs) for a body-centred cubic (b.c.c.) lattice, up to a maximum site separation of third nearest neighbours, whereas Fig. 1b shows the increase of the MBITs with interaction range. If we knew which MBITs in equation (1) were most important for a given material, we could determine the numerical values of the corresponding coefficients $\{J\}$ by fitting equation (1) to a set of $\sim N$ directly calculated $E_{QM}(\sigma)$. Once we know the coefficients, we can then effortlessly compute $E_{CE}(\sigma)$ of many other configurations, thereby identifying the stablest 'ground-state structures' out of 2^N possibilities, or performing desired thermodynamic averages over many configurations. Furthermore, once the coefficients are known, one can subject equation (1) to finite temperature modelling (such as Monte Carlo), finding the phase diagram, thermodynamic properties, short-range order and so on. Note that the system is not restricted to cases where atomic positions are fixed to lattice sites, for we fit equation (1) to fully relaxed $E_{QM}(\sigma)$ corresponding to the deepest local energy minima of configuration σ.

How many of the possible 2^N MBITs should one use? Traditional inorganic chemistry[8] suggests that only a couple of atom–atom bonding and non-bonding terms are important, but phase-diagram characteristics and behaviour, needing millielectronvolt accuracy[3], may require considerably more interactions. Although working with a large set of MBITs is certainly feasible for a brute-force least-squares fit, this breaks with the very idea of model hamiltonian construction, as it fails to identify the important interactions from first principles. More importantly, it would require great effort to calculate the enormous number of $E_{QM}(\sigma)$

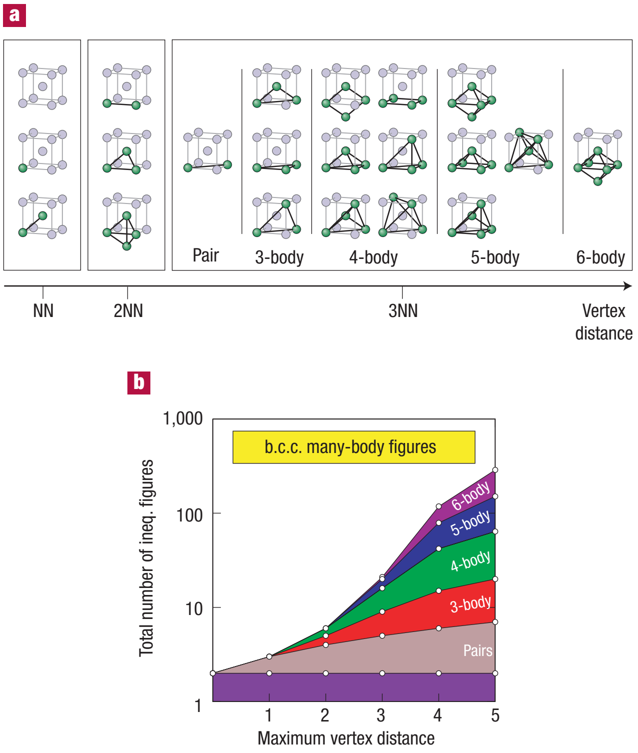

Figure 1 The first few MBITs for a b.c.c. lattice. a, Pair and multi-site 'figures' (MBIT) of equation (1) on a b.c.c. lattice up to third nearest-neighbour maximum vertex separation are shown. **b**, Combinatorial explosion of the number of inequivalent MBITs with increasing spatial extent (maximum vertex separation, *n*th nearest-neighbour).

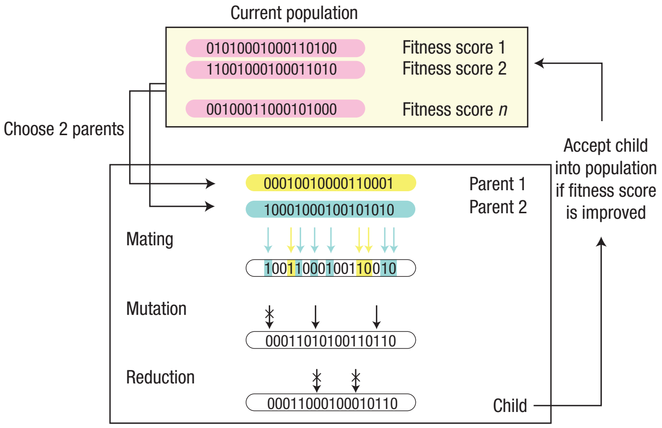

Figure 2 Flowchart of a genetic algorithm for choosing the terms to retain in a model hamiltonian. Each candidate model is represented by a series of zeros and ones, a one indicating that the corresponding term is included.

required for such a fit. If, on the other hand, we work with a smaller number p of MBITs of Fig. 1, we immediately face a 'combinatorial explosion' because selecting p MBITs from the possible 2^N corresponds to a difficult search problem. Thus, in practice the problem here is to decide exactly which terms are and which are not physically relevant. Chemical and magnetic interactions generally become weaker as the separation between atoms (or spins) increases[8], but popular oversimplified truncations to, for example, first nearest-neighbour distance[4,10,11] are generally unreliable. Traditionally, more distant interactions are introduced according to some intuitive or aesthetically appealing design principles[11–13], but there are well-established cases where such principles fail and where long-range interactions are important, even in simple alloys[3,14].

We propose here instead to choose the leading parameters of a model hamiltonian directly by means of a genetic algorithm[5]. We start by constructing a pool of MBITs from which the genetic algorithm will be required to select the few most important ones. The pool is not subject to special postulated design principles[11,13], but instead simply consists of a list of all figures, with no omissions, up to a reasonably large cut-off value for number of vertices and vertex distance, and includes larger MBITs than we ever expect to be selected. We find after that fact that the final MBITs selected by the genetic algorithm do not obey any simple rule[11,13] that we could have used to design our pool.

Genetic algorithms mimic the ideas of biological evolution, roughly 'survival of the fittest'. In materials research, successful applications have aimed to determine the values of given physical parameters from a known underlying hamiltonian. Such applications include the structure of small clusters[15–17], the grouping of point charges[18], the best components for superalloys[19] or the magnitude of the interactions of a tight-binding electronic structure hamiltonian[20]. In contrast, we use a genetic algorithm to assemble the relevant physical pieces forming the model hamiltonian that describes $E(\sigma)$. Also, unlike many optimization problems, instead of optimizing a number of continuous numerical variables, we face discrete 'yes–no' decisions regarding the inclusion of each MBIT. The search space quickly becomes astronomic (see below) and is naturally correlated. For instance, two MBITs together may provide a particularly good cluster expansion even if each one on its own does not. In this situation, traditional optimization schemes are either not applicable (gradient methods), or will not perform well (simulated annealing). In contrast, genetic algorithms juxtapose entire segments of 'genetic information' (here, binary sequences) and are therefore naturally adapted to this correlated problem.

Our approach is based on an 'outer loop' in which we iteratively select a set of input structures $\{\sigma\}_{\text{input}}$ and an 'inner loop', where the genetic algorithm is used to search for the best MBITs to fit the current set of structures. The number of structures used in the outer loop is increased iteratively by inspecting whether the current cluster expansion, when applied to 2^N configurations, suggests new structures with lower energies than those already included. The energies $E(\sigma)$ of these new structures are computed and the process is repeated until no new deep ground states are identified. Thus, we can view the cluster expansion as a 'driver,' guiding us in a space of 2^N structures to those that need to be computed by first-principles methods. In the inner loop, we use a genetic algorithm (Fig. 2). Each candidate cluster expansion is represented by a string ('genome') of N_p ones and zeroes ('genes'), representing whether a particular MBIT is or is not included in equation (1). The algorithm is iterative, improving a small population of trial genomes over many successive 'generations.' From one generation to the next, new trial genomes are generated from previous ones by 'mating': Two parent genomes—that is, two different sets of MBITs—are chosen first. Then, one by one, each MBIT (represented by zero or one) of the new child is chosen either from parent 1 or from parent 2. Next, random mutations are introduced into the child genome, and finally, the mutated child genome is adjusted to satisfy external conditions (for example, the maximum number of MBITs in a cluster expansion). A new child

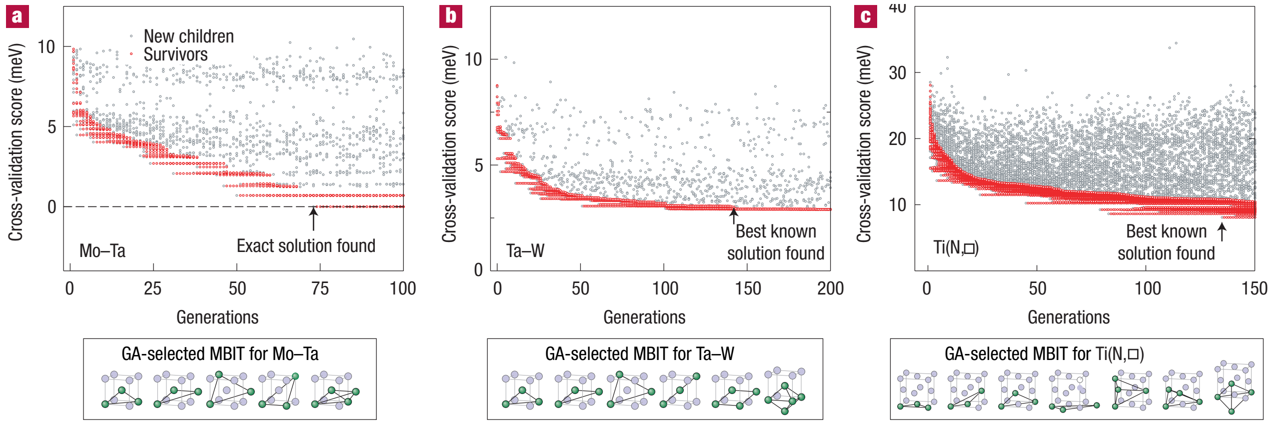

Figure 3 Genetic algorithm-based identification of the optimally predictive sets of MBITs. a, Cluster-expansion-generated input data for the Mo–Ta alloy so that the underlying MBITs are known exactly, **b**, LDA input data for Ta–W and **c**, LDA input data for the pseudo-binary system TiN$_{1-x}\square_x$, which forms vacancy-ordered compounds. Each case shows the development of cross-validation scores of the entire population of trial solutions for each case as a function of genetic algorithm (GA) generation, and the set of MBITs that constitutes the optimum solution.

is accepted on the basis of its fitness, representing the ability of this combination of MBITs to predict (through equation (1)) the energy of new configurations as determined from $E_{\mathrm{QM}}(\sigma)$.

In Fig. 3a–c, we demonstrate the ability of the method to find the correct MBITs for three examples. We first tackle a case for which the underlying interactions $\{J\}$ are known exactly. We accomplish this by substituting for $\{E_{\mathrm{QM}}(\sigma)\}$ a set of 60 energies $\{E_{\mathrm{CE}}(\sigma)\}$ generated from a known set of MBITs[21] (Fig. 3). The energies $E(\sigma)$ are passed to the genetic algorithm to recover the five underlying MBITs from a pool of 45 possible candidates. Figure 3a shows the convergence to and correct identification of the exact MBITs for which both cross-validation score and fitting error are zero. We recover the right set of MBITs in 75 genetic algorithm steps (generations), after evaluating only 975 actual cluster expansions among 1.22 million possible combinations of MBITs.

In Fig. 3b, we turn to the application to actual first-principles data $\{E_{\mathrm{QM}}(\sigma)\}$. We focus on the b.c.c. refractory alloy system Ta–W. As the measured phase diagram does not extend to low temperatures[22], it is not known whether Ta–W has ordered phases and what their structures might be. A theoretical prediction of ground-state structures could provide exciting new insights into this alloy system. We calculated the total energies $E_{\mathrm{QM}}(\sigma)$ for 59 different atomic configurations σ in the local density approximation (LDA) to density functional theory (see Methods). Based on these, we used the generic algorithm to choose six MBITs from the same pool of 45 as before (8.1 million possible combinations). They were retrieved after 140 genetic algorithm generations. The selected MBITs (Fig. 3b, lower panel) are certainly not amenable to an intuitive guess. For instance, only two are among the already large group of MBITs shown in Fig. 1a.

Figure 3c illustrates the effectiveness of the cluster expansion construction based on the genetic algorithm for a very different chemical system, namely the configurational energetics of TiN$_{1-x}\square_x$. The compound TiN crystallizes in the NaCl structure, with a tendency to form vacancies ('□') on the N sublattice. So the configurational degree of freedom is the ordering of vacancies on the N sublattice, which by itself is of face-centred cubic (f.c.c.) type (with the Ti atoms acting only as 'spectators'). The chemistry of this system is substantially different from the Ta–W system and illustrates the broad applicability of the genetic algorithm approach. For this cluster expansion, the genetic algorithm is used to select 7

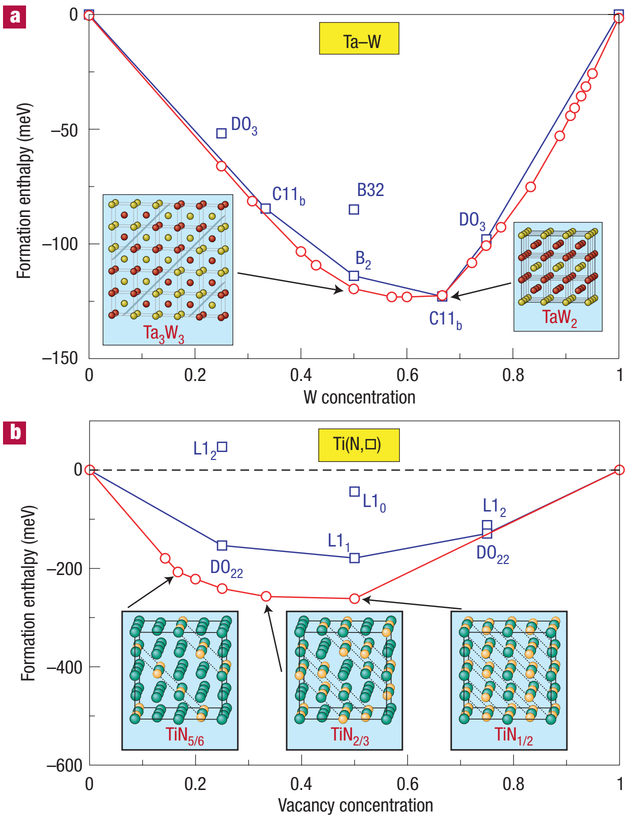

Figure 4 'Usual suspect' structures and actual ground state lines for Ta–W and Ti–N. a, Energetics of some common ground-state structures on the b.c.c. lattice (blue), and actual ground-state structures of Ta–W of the cluster expansion of Fig. 3c constructed through the genetic algorithm (red line and corresponding structures). **b**, Energetics of f.c.c. compound TiN$_{1-x}\square_x$ as a function of x. Several common ground states for f.c.c. alloys are shown as well as the ground states predicted by cluster expansion. (The Ti atoms are not shown, only the N atoms and vacancies on the non-metal sublattice.) The dashed lines indicate the (011) planes mentioned in the text that contain vacancies on every other site.

MBITs from a pool of 58 possible MBITs. The size of the search space is $\binom{58}{7} \approx 3 \times 10^8$. The solution is found in under 200 generations, less than 10,000 individual guesses (that is, only 1/30,000 of possible cases). Extensive testing indicates that the same solution is found with all randomly generated initial populations, thus giving us confidence in its correctness.

Having constructed a model hamiltonian $E_{\mathrm{CE}}(\sigma)$ by selecting a few interaction parameters from an enormous number of possible combinations, we are now in the position to tackle another selection problem of similar scale: finding the most stable ground-state structures by scanning $E(\sigma)$ for ~3 million distinct configurations σ. We do so by selecting for each composition the lowest-energy configuration and then eliminating those that are less stable than its disproportionation products. This gives the 'ground-state line' of Fig. 4, showing the stablest structures as 'breaking points' of the convex hull of E versus x. Monte Carlo simulations of cells bigger than 20 sites reveal no new ground states. That some of the structures shown in Fig. 4a for Ta–W are unsuspected can be judged by comparing with the blue symbols, showing the b.c.c.-based 'usual suspect' stable structures that are frequently encountered: $\mathrm{D0_3}$ ($\mathrm{Ta_3W}$ and $\mathrm{TaW_3}$), $\mathrm{C11_b}$ ($\mathrm{Ta_2W}$ and $\mathrm{TaW_2}$), B2 and B32 (both TaW). Previously[23,24], the B2 and $\mathrm{D0_3}$ structures have been predicted to be the low-temperature stable phases in Ta–W, a prediction contradicted by our results. The predicted stable structures of the pseudo-f.c.c. compound $\mathrm{TiN}_{1-x}\square_x$ (Fig. 4b) are equally surprising. Three particularly deep ground states are found at concentrations $x = 1/6$, 1/3 and 1/2. The predicted $\mathrm{TiN}_{1/2}$ ground state is the NbP-type structure, a (201) AABB superlattice. $\mathrm{TiN}_{2/3}$ is the 20-atom structure known as $\mathrm{Sc_2S_3}$. The ground state of $\mathrm{TiN}_{5/6}$ is identical to $\mathrm{Sc_2S_3}$ except with one more occupied N site per cell. All of these structures contain a common motif: (011) planes with vacancies on alternate sites (Fig. 4b).

Our genetic algorithm can thus be used to select the best components of model hamiltonians from a vast set of possibilities using a handful of first-principles calculations as a guide. We have successfully predicted the hitherto unknown ground states for two chemically distinct systems with the accuracy of the first-principles hamiltonian. Their very nature, showing both simple known types of order and unexpectedly complex arrangements, demonstrates the power of our approach.

METHODS

FITNESS CRITERION

For our particular choice of model hamiltonian (the cluster expansion), the fitness of a particular cluster expansion (that is, set of MBITs) is quantified by a form of leave-many-out cross-validation[25]: The set of structures σ for which we possess first-principles energies E_{QM}, $\{\sigma\}_{\mathrm{input}}$, is subdivided into a fitting set $\{\sigma\}_{\mathrm{fit}}$ and a prediction set $\{\sigma\}_{\mathrm{pred}}$. Because $E_{\mathrm{QM}}(\sigma)$ includes in it all the proper chemical bonding information, we determine the interaction parameters $\{J\}$ by fitting to $E_{\mathrm{QM}}(\sigma)$ in $\{\sigma\}_{\mathrm{fit}}$. The predicted values $E_{\mathrm{CE}}(\sigma)$ for each structure $\sigma \epsilon \{\sigma\}_{\mathrm{pred}}$ are then compared with their (known) first-principles counterparts $E_{\mathrm{QM}}(\sigma)$. We repeat the process M times for different subdivisions of $\{\sigma\}_{\mathrm{input}}$ into fitting and prediction sets $\{\sigma\}_{\mathrm{fit}}^{(m)}$ and $\{\sigma\}_{\mathrm{pred}}^{(m)}$ ($m = 1 \ldots N$) and define a cross-validation (CV) score[25] as the average prediction error of this procedure,

$$S_{\mathrm{CV}} = \frac{1}{M} \frac{1}{N_{\mathrm{pred}}} \sum_{m=1}^{M} \sum_{\sigma \in \{\sigma\}_{\mathrm{pred}}^{(m)}} \left| E_{\mathrm{CE}}^{(m)}(\sigma) - E_{\mathrm{QM}}(\sigma) \right|^2 . \tag{2}$$

The optimum combination of MBITs is characterized by minimal S_{cv}, a resilient criterion for sufficiently large $\{\sigma\}_{\mathrm{input}}$.

FIRST-PRINCIPLES INPUT DATA E(σ)

We apply our genetic algorithm to first-principles results $E(\sigma)$ for $\mathrm{Ta}_{1-x}\mathrm{W}_x$ and $\mathrm{TiN}_{1-x}\square_x$ compounds σ. All energies were calculated for geometrically fully relaxed compounds in the local density approximation to density functional theory, using projector-augmented wave potentials and other standard algorithms as implemented in the VASP computer program package[26].

CLUSTER-EXPANSION FORMALISM

Specifically, we use the mixed-basis cluster-expansion formalism[3], which simplifies the selection problem by determining all pair terms explicitly in a constrained fit. (This is possible because the number of pair terms rises relatively slowly with distance.) In Fig. 1b, the true combinatorial problem is due to MBITs with three or more vertices, from which our genetic algorithms chooses those relevant to achieve quantitative accuracy.

Received 13 October 2004; accepted 1 March 2005; published 17 April 2005.

References

1. Landau, L. D. & Lifshitz, E. M. *Statistical Physics* (Pergamon, Oxford, 1980). Transl. from the Russian by J. B. Sykes & M. J. Kearsley.
2. Rado, G. T. & Suhl, H. (eds) *Magnetism* Vol. 2B (Academic, New York, 1965).
3. Zunger, A. In *Statics and Dynamics of Alloy Phase Transformations* (eds Turchi, P. & Gonis, A.) 361–419 (Plenum, New York, 1994).
4. de Fontaine, D. Cluster approach to order-disorder transformations in alloys. *Solid State Phys.* **47,** 33–176 (1994).
5. Michalewicz, Z. & Fogel, D. B. *How to Solve it: Modern Heuristics* (Springer, Berlin, 2000).
6. Maddox, J. Crystals from first principles. *Nature* **335,** 201 (1988).
7. Phillips, J. C. *Bonds and Bands in Semiconductors* (Academic, New York, 1973).
8. Pauling, L. *The Nature of the Chemical Bond* (Cornell Univ. Press, Ithaca, 1960).
9. Hume-Rothery, W. & Raynor, G. *The Structure of Metals and Alloys* (Institute of Metals, London, 1954).
10. Connolly, J. & Williams, A. Density-functional theory applied to phase transformations in transition-metal alloys. *Phys. Rev. B* **27,** 5169–5172 (1983).
11. Kikuchi, R. In *Noble Metal Alloys* (eds Massalski, T. B., Pearson, W. B., Bennet, L. H. & Chang, Y. A.) (The Metallurgical Society, Warrendale, PA, 1986).
12. van der Walle, A. & Ceder, G. Automating first-principles phase diagram calculations. *J. Phase Equil.* **23,** 348–359 (2002).
13. Zarkevich, N. A. & Johnson, D. D. Reliable first-principles alloy thermodynamics via truncated cluster expansions. *Phys. Rev. Lett.* **92,** 255702 (2004).
14. de Gironcoli, S. & Baroni, S. Effects of disorder on the vibrational properties of SiGe alloys: failure of mean-field approximations. *Phys. Rev. Lett.* **69,** 1959–1962 (1992).
15. Deaven, D. M. & Ho, K. M. Molecular geometry optimization with a genetic algorithm. *Phys. Rev. Lett.* **75,** 288–291 (1995).
16. Ho, K. M. *et al.* Structures of medium-sized silicon clusters. *Nature* **392,** 582 (1998).
17. Stucke, D. P. & Crespi, V. H. Predictions of new crystalline states for assemblies of nanoparticles: perovskite analogues and 3-D arrays of self-assembled nanowires. *Nano Lett.* **3,** 1183–1186 (2003).
18. Morris, J. R., Deaven, D. M. & Ho, K. M. Genetic algorithm energy minimization for point charges on a sphere. *Phys. Rev. B* **53,** R1740 (1996).
19. Johannesson, G. H. *et al.* Combined electronic structure and evolutionary search approach to materials design. *Phys. Rev. Lett.* **88,** 255506 (2002).
20. Klimeck, G. & Bowen, R. C. Si tight-binding parameters from genetic algorithm fitting. *Superlattices Microstruct.* **27,** 77–88 (2000).
21. Blum, V. & Zunger, A. Structural complexity in binary bcc ground states: the case of bcc MoTa. *Phys. Rev. B* **69,** 020301 (2004).
22. Predel, B. (ed.) *Phase Equilibria, Crystallographic and Thermodynamic Data of Binary Alloys. Landolt-Börstein, New Series, Group IV* Vol. 5H (Springer, Berlin, 1997).
23. Sigli, C. & Sanchez, J. M. Electronic Structure calculation of ordering and segregation energies of transition metal alloys. *Acta Metall. Mater.* **36,** 367–375 (1988).
24. Turchi, P. E. A., Gonis, A., Drchal, V. & Kurdnovsky, J. First-principles study of stability and local order in substitutional Ta–W alloys. *Phys. Rev. B* **64,** 085112 (2001).
25. Baumann, K. Cross validation as the objective function for variable selection. *Trends Anal. Chem.* **22,** 395–406 (2003).
26. Kresse, G. & Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **6,** 15–50 (1996).

Acknowledgements

We acknowledge financial support from the NSF through DMR-0244183, and from DOE-SC-BES-DMS, as well as the Intramural Grant Program at Northern Arizona University.
Correspondence and request for materials should be addressed to G.L.W.H.

Competing financial interests

The authors declare that they have no competing financial interests.