Predicting Crystal Structures with Data Mining of Quantum Calculations

Stefano Curtarolo,[1] Dane Morgan,[1] Kristin Persson,[1] John Rodgers,[2] and Gerbrand Ceder[1,*]
[1]*Department of Materials Science and Engineering, MIT, Cambridge, Massachusetts 02139, USA*
[2]*Toth Information Systems Inc., Ottawa, Canada*

(Received 30 May 2003; published 24 September 2003)

Predicting and characterizing the crystal structure of materials is a key problem in materials research and development. It is typically addressed with highly accurate quantum mechanical computations on a small set of candidate structures, or with empirical rules that have been extracted from a large amount of experimental information, but have limited predictive power. In this Letter, we transfer the concept of heuristic rule extraction to a large library of *ab initio* calculated information, and we demonstrate that this can be developed into a tool for crystal structure prediction.

DOI: 10.1103/PhysRevLett.91.135503 PACS numbers: 61.50.Ah, 61.66.Dk, 61.68.+n

Ab initio methods, which predict materials properties from the fundamental equations of quantum mechanics, are becoming a ubiquitous tool for physicists, chemists, and materials scientists. These methods allow scientists to evaluate and prescreen new materials "in silico," rather than through time-consuming experimentation, and in some cases, even make suggestions for new and better materials [1–4]. One inherent limitation of most *ab initio* approaches is that they do not make explicit use of results of previous calculations when studying a new system. This can be contrasted with data-centered methods, which mine existing data libraries to help understand new situations. The contrast between data-centered and traditional *ab initio* methods can be seen clearly in the different approaches used to predict the crystal structure of materials. This is a difficult but important problem that forms the basis for any rational materials design. In heuristic models, a large amount of experimental observation is used in order to extract rules which rationalize crystal structure with a few simple physical parameters such as atomic radii, electronegativities, etc. The Miedema rules for predicting compound forming [5], or the Pettifor maps [6] which can be used to predict the structure of a new binary material by correlating the position of its elements in the periodic table to those of systems for which the stable crystal structure is known, are excellent examples of this. In contrast, *ab initio* approaches do not use data from previous studies but rather try to determine structure by optimizing from scratch the complex quantum mechanical description of the system, either directly (as in *ab initio* molecular dynamics) or in coarse-grained form (as in lattice models [7–9]). Here, we merge the ideas of data-centered methods with the predictive power of *ab initio* computation. We propose a new approach whereby *ab initio* investigations on new systems are informed with *knowledge* obtained from results already collected on other systems. We refer to this approach as data-mining of quantum calculations (DMQC) and demonstrate its efficiency in increasing the speed of predicting the crystal structure of new and unknown materials. Using a principal component analysis (PCA) on over 6000 *ab initio* energy calculations, we show that the energies of different crystal structures in binary alloys are strongly correlated between different chemical systems, and we demonstrate how this correlation can be used to accelerate the prediction of new systems. We believe that this is an interesting new direction to address in a practical manner the problem of predicting the structure of materials.

To determine the ground states of a system one needs to find, as a function of composition, the ordered compounds that have an energy lower than any other structure or any linear combination of structures that gives the proper composition. This set of ground-state structures forms a *convex hull* as all other structures have an energy that falls above the set of tie lines that connects the energy of the ground states. Using density functional theory we have calculated a *library* of *ab initio* energies for 114 different crystal structures in each of 55 binary metallic alloys. The alloys include all 45 binaries that can be made from row 4 transition metals, as well as ScAl, AgMg, AgTi, CdTi, MoTi, PdTi, RhTi, RuTi, TcTi, and TiZr. About 1/3 of the crystal structures in the library were chosen from the most common binary crystal structures in the CRYSTMET database for intermetallics [10]. The rest are superstructures of the fcc, bcc, and hcp lattices at various compositions.

The formation energy for each structure is determined with respect to the most stable structure of the pure elements. Energy calculations were done using density functional theory, in the local density approximation, with the Ceperley-Alder form for the correlation energy as parametrized by Perdew-Zunger [11] with ultrasoft pseudopotentials, as implemented in VASP [12]. Calculations are at zero temperature and pressure, and without zero-point motion. The energy cutoff in an alloy was set to 1.5 times the larger of the suggested energy cutoffs of the pseudopotentials of the elements of the alloy (suggested energy cutoffs are derived by the method described in [12]). Brillouin zone integrations were done

0031-9007/03/91(13)/135503(4)$20.00

using 2000/(number of atoms in unit cell) $\mathbf{k}$ points distributed as uniformly as possible on a Monkhorst-Pack mesh. We verified that with these energy cutoffs and $\mathbf{k}$-point meshes the absolute energy is converged to better then 10 meV/atom. Energy differences between structures are expected to be converged to much smaller tolerances. Spin polarization was not used as no magnetic alloys were studied. All structures were fully relaxed.

For each alloy i, consider the 114 structural formation energies as the components of a vector $\mathbf{E}_i$ in a 114-dimensional space. If the energies of the structures are linearly dependent then the vectors for each alloy will not be distributed randomly in the 114 dimensional space, but confined in a subspace of reduced dimension. To look for such *approximate* linear dependencies we use PCA [13]. This allows us to express the energy vector of an alloy as an expansion in a basis of reduced dimension d, $\mathbf{E}_i = \sum_{j=1}^{d} \alpha_{ij}\mathbf{e}_i + \boldsymbol{\epsilon}_i(d)$, where $\boldsymbol{\epsilon}_i(d)$ is the error vector for the alloy i. PCA consists of finding the proper basis set $\{\mathbf{e}_j(d)\}$ that minimizes the remaining squared error $\sum_i \boldsymbol{\epsilon}_i^T \cdot \boldsymbol{\epsilon}_i$ for a given dimension d. These optimum basis vectors $\{\mathbf{e}_j(d)\}$ are called the principal components (PC's) and form a set of orthogonal vectors ordered by the amount of variation of the original sample they can explain. More intuitively, they are a new set of axes in the 114 dimensional space, ordered according to the fraction of the data lying along that axis. As an extreme example, if the energies of all 55 alloys were proportional to each other, then all the alloy vectors would lie along a single line, and the first PC would be a subspace that encompassed all the data ($d = 1$).

A principal components analysis of our *ab initio* data set (Fig. 1) shows that significant dimension reduction is possible in the space of structural energies. The solid curve, labeled "55," in Fig. 1 shows the remaining unexplained root mean square (rms) error (average error in the 114 structural energies of the 55 alloys), as a function of the number of principal components d. All quantities are given as energy per atom. The number of relevant dimensions depends on the error one can tolerate. For example, to describe the energies with a 50 meV rms error, only 9 dimensions are required, many fewer than the original 114. The implication is that it is possible to perform far fewer than 114 calculations to parametrize the 9 dimensional subspace, and then derive the other 105 energies through linear relationships given by the PCA.

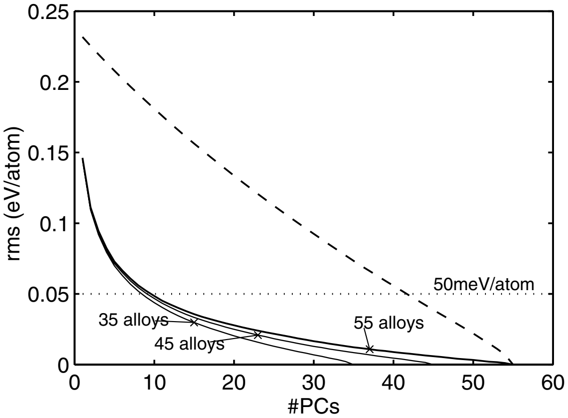

FIG. 1. The rms error as a function of the number of principal components. The solid lines show results for the libraries containing 35, 45, and 55 alloys. The dashed line shows results for the 55-alloy library where the energies for each alloy have been randomly permuted.

Dimension reduction holds only because the energy differences of structures are strongly correlated between alloys with different chemistry. In fact, if we perform a PCA analysis in which the structural energies for each alloy are randomly permuted, and hence destroy their relations, there is little opportunity for dimension reduction, as the dashed curve in Fig. 1 shows. Given an acceptable accuracy, dimension reduction does not depend on the dimension of the library, once the library is bigger than a certain size. Figure 1 shows the PCA analysis for 35, 45, and 55 alloys. For subspaces defined by up to ≈ 20 PC's (27 meV rms accuracy) the variance is essentially independent of the number of alloys, indicating that the dimension reduction we obtain can be expected to apply to new alloy systems.

These correlations are further confirmation that the success of heuristic methods is not accidental and that with relatively few parameters it can be possible to predict the structure of a binary alloy. In fact, these correlations can be used to develop an *ab initio data-mining* algorithm that rapidly searches through the available space of possible structures. It is important to emphasize that the correlations do not lead to exact linear dependencies, but only approximate ones. The number of degrees of freedom that have to be retained is determined by choosing the level of approximation.

Given a library of N_a alloys, N_s structures, and a new alloy where the first n energies have been calculated, we predict the energy for structure $i > n$ of the new alloy as follows. Define $\mathbf{X}$ as the (n, N_a) matrix of energies for structures $\{1, \ldots, n\}$ in the library. Define $\mathbf{y}$ as the n-component vector of known energies for the new alloy and $\mathbf{X}'$ as the N_a-component vector of energies for structure i for all alloys in the library. The scalar y' represents the unknown energy of structure i for the new alloy. We regress $\mathbf{y}$ on $\mathbf{X}$ using the partial least squares method [14,15] implemented with the SIMPLS algorithm [16]. The resulting regression coefficients are used to predict y' from $\mathbf{X}'$. This is done for every structure of the new alloy for which the energy has not yet been calculated.

The ground states for an alloy are found through an iterative scheme. At each step, the PLS regression is used to find the most probable ground state, which is then calculated with quantum mechanics and added to the data. Although the PLS regression is not accurate enough to substitute for the *ab initio* energy, it is sufficient for selecting candidate structures efficiently. The algorithm is started with only the pure element energies for the two elements of the alloy in the bcc, fcc, and hcp structures, and then proceeds as follows.

Step 1 (prediction).—The regression algorithm given above is used to predict all unknown structural energies in the new system. We found that for early iterations (< 10) the rms error can be reduced by preclustering the library into ordering and phase-separating systems and regressing only within the library subcluster in which the system is predicted to fall. Physically, this means that for the early stage of the iterative procedure, new alloys regress better with similar alloys than with the complete library.

Step 2 (suggestion).—With the available *ab initio* calculated energies we determine the ground-state energy versus composition curve (convex hull). The structure with *data-mined* predicted energy farthest below the convex hull of calculated energies is calculated with quantum mechanics and added to the database. If no structure breaks the hull, we look for the structure predicted to be closest to the hull. For early iterations (< 13 in Fig. 2), if no such structure can be found within 80 meV of the ground-state hull, we consider the prediction to have failed in this step, and instead we add the most frequent and not yet calculated ground-state structure of the database.

Step 3 (calculation).—The candidate suggested structure is then *calculated* with quantum mechanics and added to the list, and the entire process is iterated (*prediction* $\Rightarrow$ *suggestion* $\Rightarrow$ *calculation*). With each step, more energetic information for the new alloy is incorporated and a better prediction of the ground state can be expected.

Any structure in the library can be predicted and there are no preconceived biases as to the symmetry or underlying superlattice of the structure as is the case for methods that work with lattice model approaches. For example, in the Ti-Pt alloy, our method correctly finds the A15 [17,18] structure to be a ground state for Ti_3Pt after only 20 steps in the algorithm, even though this structure is not a superstructure of fcc (the structure of Pt) or of hcp (the structure of Ti) and is therefore not an obvious structure to investigate for this system. To study in a more statistically significant way how this iterative scheme converges we tested how well the library minus one alloy can perform predictions on the alloy left out. A key property is whether the alloy is immiscible (no ordered compounds) or has intermediate compounds (compound forming). Empirical schemes such as the one developed by Miedema have been particularly successful in classifying this difference [5]. We find that DMQC can predict whether an alloy excluded from the library is compound forming with 95%, 98%, and 100% accuracy using 3, 6, and 13 calculations, respectively. Note that here and below, we do not count the initial pure element calculations, since these are performed only once for each element. For comparison, if one randomly picked trial structures from the list of 114 structures, predictions with 95%, 98%, and 100% accuracy require 7, 21, and 98 calculations, respectively. The DMQC method performs extremely well, far better than a naive random choice of structures, and gives almost a perfect prediction with a small amount of computation.

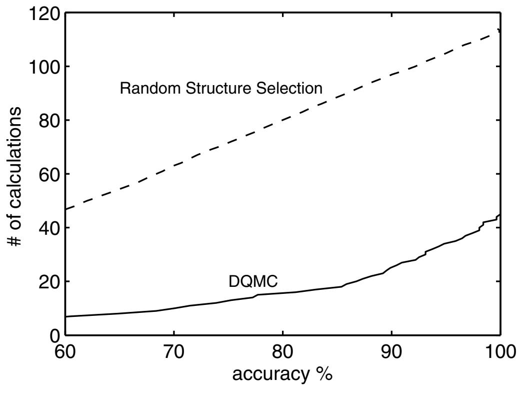

FIG. 2. The number of calculations as a function of the percentage of ground states predicted correctly, with the DMQC method (solid line) and with random structure selection (dashed line). 90% accuracy can be achieved with DMQC with 26 calculations, many fewer than the 98 calculations necessary for random structure selection.

A more stringent evaluation is whether the correct stable crystal structures are predicted for the system left out. Figure 2 (solid line) shows the number of calculations required as a function of the percentage of ground states predicted correctly (averaged over all alloys). For our purpose, "correct" is what would be obtained from the direct quantum mechanical calculations on all 114 structures. 90% accuracy can be achieved with less than 26 calculations for an alloy. To achieve the same confidence level with random structure selection (dashed line) one needs to calculate almost the complete database (98 calculations).

Even though it is generally believed that the binary alloys are well characterized experimentally, our approach can be used to quickly predict previously unidentified stable structures in some systems. For example, with only 26 calculations we predict Ag_3Cd and Ag_2Cd, respectively, to have the DO_{24} and C37 structures. In addition, we predict the previously unidentified structure for $CdZr_3$ to be A15 (Cr_3Si-type). This prediction takes only 21 iterations and is particularly interesting since A15 does not share the hcp parent lattice of Cd and Zr. These predictions were confirmed by calculation of all the prototype structures in the library. A more detailed analysis of the predictions made from our database in a large number of systems will be published elsewhere [19].

More structures will need to be added to the library to give the method better applicability to many unknown systems. It is therefore important to assess how the number of required calculations scales with the number of structures in the library. This scaling is shown in Fig. 3

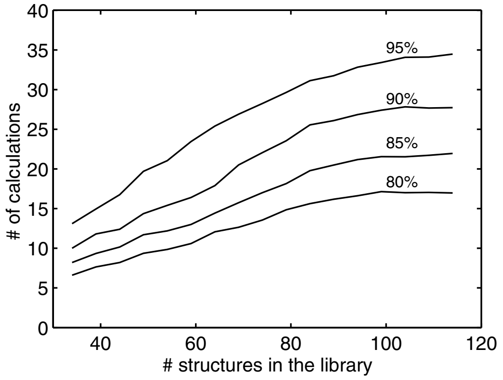

FIG. 3. The average number of calculations needed to obtain a given accuracy of predicted crystal structures, as a function of the number of structures in the library. Results are given for 80%, 85%, 90%, and 95% accuracies. The number of calculations increases less than linearly with the number of structures in the database, demonstrating that efficiency increases as the library grows.

for various required confidence levels. As the library grows, more calculations are needed to select between the increasing number of possibilities. Fortunately, the number of calculations increases less than linearly with the number of structures in the database, demonstrating that efficiency increases as the library grows.

Our current DMQC approach has the limitation that structure types must already be in the database to be predicted. However, a concerted effort to develop a public database, analogous to those used in biology, may make this limitation less important. Our work has also focused on a simple test library of binary alloys. The real payoff will come with the inclusion of multicomponent systems, where fewer than 10% of all intermetallic systems have been characterized [6,20]. A library of ternary structures can be integrated with the binary libraries and extensions of the formalism are not required, besides adding an extra composition variable. Although the data-mining methods discussed here are centered around dimension reduction and linear correlation, other approaches, including nonlinear methods (e.g., neural nets, clustering, learning machines, etc.) will certainly be more effective in extracting information from the library.

In summary, by data-mining quantum mechanical calculations (DMQC) we have established that there exist significant correlations among *ab initio* energies of different structures in different materials. The correlations we found can be seen as a formal extension of the heuristic structure-properties selection rules that have been established in the past on the basis of large amounts of *experimental* structure information [20–22]. Our approach differs from the previous classifications in that we correlate on calculated information (structural energies in our particular example), and hence our description can be used when there is limited experimental data, and can be extended to arbitrary accuracy. We believe that the integration of data-mining techniques with *ab initio* methods is a promising development towards the practical prediction of crystal structure.

The research was supported by the Department of Energy, Office of Basic Energy Science under Contract No. DE-FG02-96ER45571.

---

*Corresponding author.
Electronic address: gceder@mit.edu

[1] R. H. Victora, C. F. Brucker, T. K. Hatwar, J. E. Hurst, B. Uryson, and D. Karns, J. Appl. Phys. **81**, 3833 (1997).
[2] G. Ceder, Y. M. Chiang, D. R. Sadoway, M. K. Aydinol, Y. I. Jang, and B. Huang, Nature (London) **392**, 694 (1998).
[3] G. B. Olson, Science **288**, 993 (2000).
[4] G. H. Johannesson, T. Bligaard, A. V. Ruban, H. L. Skriver, K. W. Jacobsen, and J. K. Norskov, Phys. Rev. Lett. **88**, 255506 (2002).
[5] F. R. De Boer, *Cohesion in Metals* (North-Holland, Amsterdam, 1998).
[6] D. G. Pettifor, J. Phys. C **19**, 285 (1986).
[7] G. Ceder, Comput. Mater. Sci. **1**, 144 (1993).
[8] D. de Fontaine, in *Solid State Physics*, edited by H. Ehrenreich and D. Turnbull (Academic Press, New York, 1994), Vol. 47, pp. 33–176.
[9] A. Zunger, in *Statics and Dynamics of Alloy Phase Transformations*, edited by P. E. A. Turchi and A. Goniss (Plenum, New York, 1994), pp. 361–419.
[10] P. S. White, J. R. Rodgers, and Y. Le Page, Acta Crystallogr. Sect. B **58**, 343 (2002).
[11] J. P. Perdew and A. Zunger, Phys. Rev. B **23**, 5048 (1981).
[12] G. Kresse and J. Furthmuller, Comput. Mater. Sci. **6**, 15 (1996).
[13] J. E. Jackson, *A User's Guide to Principal Components* (Wiley, New York, 1991).
[14] S. Wold, A. Ruhe, H. Wold, and W. J. Dunn, SIAM J. Sci. Stat. Comput. **5**, 735 (1984).
[15] R. Kramer, *Chemometric Techniques for Quantitative Analysis* (Dekker, New York, 1998).
[16] S. De Jong, Chemom. Intell. Lab. Syst. **18**, 251 (1993).
[17] E. C. Van Reuth and R. M. Waterstrat, Acta Crystallogr. Sect. B **24**, 186 (1968).
[18] *Binary Alloy Phase Diagrams*, edited by T. B. Massalski (ASM International, Metals Park, OH, 1992).
[19] S. Curtarolo, D. Morgan, and C. Ceder, "High Throughput *Ab Initio* Computing: Transition Metal Binary Alloys" (to be published).
[20] P. Villars, in *Intermetallic Compounds*, edited by J. H. Westbrook and R. L. Fleisher (Wiley, New York, 1994), pp. 227–275.
[21] W. Hume-Rothery, *The Structure of Metals and Alloys* (Institute of Metals, London, 1988).
[22] M. F. Ashby, *Materials Selection in Mechanical Design* (Butterworth-Heinemann, Oxford, 1999).