Mat. Res. Soc. Symp. Proc. Vol. 804 © 2004 Materials Research Society JJ9.25.1

Data Mining Approach to Ab-Initio Prediction of Crystal Structure

Dane Morgan, Gerbrand Ceder
Department of Materials Science and Engineering, Massachusetts Institute of Technology
Cambridge, Massachusetts 02139, USA
Stefano Curtarolo
Department of Mechanical Engineering and Materials Science, Duke University,
Durham, NC 27708

ABSTRACT

Predicting crystal structure is one of the most fundamental problems in materials science and a key early step in computational materials design. *Ab initio* simulation methods are a powerful tool for predicting crystal structure, but are too slow to explore the extremely large space of possible structures for new alloys. Here we describe ongoing work on a novel method (Data Mining of Quantum Calculations, or DMQC) that applies data mining techniques to existing *ab initio* data in order to increase the efficiency of crystal structure prediction for new alloys. We find about a factor of three speedup in *ab intio* prediction of crystal structures using DMQC as compared to naïve random guessing. This study represents an extension of work done by Curtarolo, et al. [1] to a larger library of data.

INTRODUCTION

Predicting the stable crystal structures for new alloys is a challenging problem, and only a few paths lead to practical solutions. One important class of techniques, generally called structure maps [2], clusters known crystal structures based on properties of the alloying elements. The crystal structure of new alloys can then be predicted by looking at known structures in the relevant clusters. These methods have recently been shown to be quite accurate [3], but are severely limited by requiring very large amounts of data to map out well-defined clusters.

Other approaches predict the relative stability of different structures using total energy models. These approaches are limited by the very large space of possible structures and presence of many local minimum, making direct optimization to find the minimum energy very difficult. By using coarse-grained Hamiltonians that can be evaluated very quickly (for example, empirical rules and potentials [4] or cluster expansions [5]) it is often possible to optimize structures over some portion of the possible structural space. However, these coarse-grained models can be difficult to construct, are often inaccurate, and usually cannot explore the whole structure space to obtain a global minimum.

Problems of model construction and accuracy can be largely overcome by avoiding coarse graining and using a full *ab initio* Hamiltonian to calculate the energies. Unfortunately, *ab initio* methods can only examine a very limited number of candidate structures because of computational limitations. Directly optimizing the energy over all of the space of possible structures with full *ab initio* simulations is not possible, and ground state searches have been limited to very restricted structure subsets [6]. For *ab*

initio prediction the problem is to find a list of candidate structures that is as short as possible, but still likely to contain the true crystal structure.

In this paper we discuss a method for generating a sensible list of candidate structures for *ab initio* prediction of crystal structures for new alloys. The method will be referred to as Data Mining of Quantum Calculations (DMQC). The basic idea of the approach is to mine a library of known crystal structure energies to find patterns that can help generate a sensible candidate structure list for a new alloy. The results presented here are similar to those presented in Ref. [1], but updated to an extended database about twice as large.

DATABASE

The DMQC method relies on mining a library of previously calculated data to predict properties of new data. Therefore, the size and details of the library can be important. The library from Ref. [1] consisted of structural energies for 115 different structures for each of 55 alloys. We have now extended this library to contain 154 structures and 82 alloys, for a total of 154 x 82 = 12,628 *ab initio* energies. The structures in the present library are discussed in Ref. [7] and the alloys are all binary systems, primarily 3d and 4d transition metals:
AgAu,AgCd,AgMg,AgMo,AgNa,AgNb,AgPd,AgPt,AgRh,AgRu,AgTc,AgTi,AgY,AgZr,
AlSc,AuCd,AuMo,AuNb,AuPd,AuPt,AuRh,AuRu,AuSc,AuTc,AuTi,AuY,AuZr,CdMo,
CdNb,CdPd,CdPt,CdRh,CdRu,CdTc,CdTi,CdY,CdZr,CrMg,GeSi,MoNb,MoPd,MoPt,M
oRh,MoRu,MoTc,MoTi,MoY,MoZr,NbPd,NbPt,NbRh,NbRu,NbTc,NbY,NbZr,PdPt,Pd
Rh,PdRu,PdTc,PdTi,PdY,PdZr,PtRh,PtRu,PtTc,PtTi,PtY,PtZr,RhRu,RhTc,RhTi,RhY,Rh
Zr,RuTc,RuTi,RuY,RuZr,TcTi,TcY,TcZr,TiZr,YZr

Energy calculations were done using density functional theory, in the local density approximation, with the Ceperley-Alder form for the correlation energy as parameterized by Perdew-Zunger [8] with ultrasoft pseudopotentials, as implemented in VASP [9]. 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 [9]). Brillouin zone integrations were done using 2000/(number of atoms in unit cell) k-points distributed as uniformly as possible on a Monkhorst-Pack mesh. We verified that with these energy cutoffs and k-points meshes the absolute energy is converged to better than 10 meV/atom. Energy differences between structures are expected to be converged to even smaller tolerances. Spin polarization was not used as no magnetic alloys were studied. All structures were fully relaxed.

METHODS

DMQC can use many methods to find patterns in the library data. The DMQC approach discussed here uses linear regression on the library of known structural energies to help predict structural energies in a new alloy. 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…n} in the library. Define **y** as the *n*-component vector of known energies for the new alloy and **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 **y** on **X** using the Partial Least Squares method [10] [11] implemented with the SIMPLS algorithm [12]. The resulting regression coefficients are used to predict y' from **X'**.

This is done for every structure of the new alloy for which the energy has not yet been calculated. Using these predicted energies an iterative approach can be developed whereby the next structure to calculate can be chosen so that it has a much better than random chance of being a ground state of the system. The details of the iterative approach are given in Ref. [1].

RESULTS

For a linear regression based technique to be successful it is necessary that there is some linear dependence among the structural energies. If we view each of the alloys as a vector with 154 components, one for each structural energy, then our database can be seen as 82 points in a 154 dimensional space. The geometric interpretation of linear dependence is that these 82 points are located in a subspace of fewer than 82 dimensions. The optimal subspace of any dimension can be identified by applying Principal Component Analysis [13], where the number of principal components is the dimension of the subspace. For any subspace we can find the distance between each true alloy vector and its projection into the subspace. The RMS of these distances provides a measure of how well the points are localized in the subspace. This RMS distance is plotted against the number of principal components in Fig. 1. For comparison, we also show the RMS distance associated with the same data but where the energies have been randomly permuted for each alloy. It is clear that the errors are dramatically lower for the true data compared to the randomized data, which shows that the structural energies have strong linear correlations.

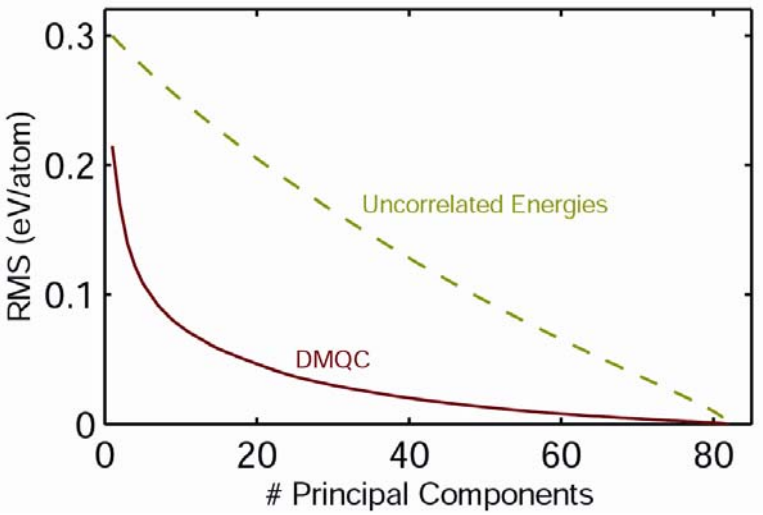

Figure 1: The RMS distances in the data as a function of the number of principal components used. Solid line is true data, dashed line is randomized data.

Our implementation of DMQC utilizes the linear relationships among the structural energies to help choose sensible structures (one's likely to be the ground state) to calculate. As a test of the methods effectiveness, we compare results to a naïve random guessing approach. The comparison is made based on the ability to predict the stable structures (convex hull) for each alloy in the library. In the following, the convex hull will be based on the total library data, not the true experiments. Here is how the test works. We remove one alloy from the library, which will be considered the new alloy. We then step through all 154 structures for the removed alloy in a random order, imagining we are calculating the structural energies for the first time in that order. This creates a list of calculated energies that grows at each step. After each step we use the calculated energies from all the completed steps to predict the convex hull, keeping track of what percentage of the ground states we identify correctly. We then do this for every alloy in turn and average the results. The number of calculations required to reach a given percentage accuracy is shown in Fig. 2, where the dashed line represents the naïve approach of random structure selection. We then repeat this whole calculation, but now choose the structures to calculate at each step using our DMQC method, rather than choosing structures at random. Note that the removed alloy is not included in the library, so it cannot bias the DMQC results. The percentage of correctly identified ground states using the DMQC method is shown as a solid line in Fig. 2. Fig. 2 shows that *DMQC can provide accuracy comparable to random guessing with about three times fewer calculations.*

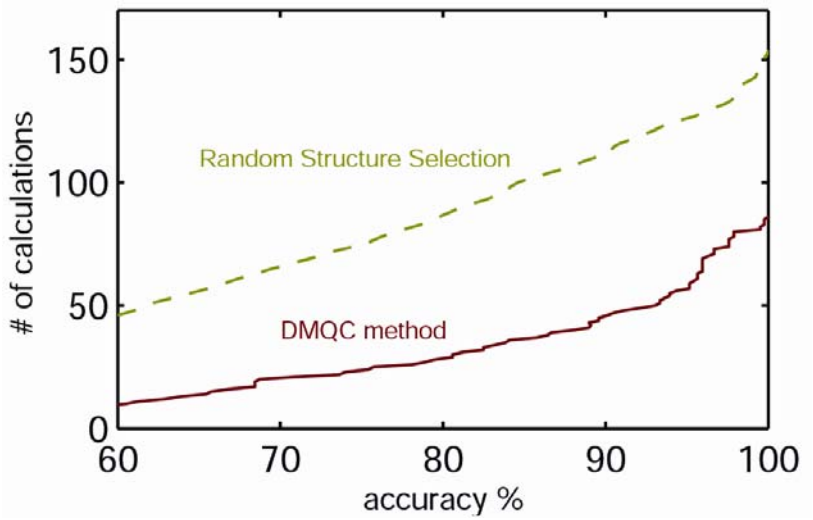

Figure 2: Comparison of the number of calculations required to reach a given predictive accuracy using random structure selection (dashed line) and DMQC method (solid line).

Although a factor of three improvement is very useful, it is our hope that for larger libraries the relative speedup (the fraction of the possible structures that DMQC allows us to avoid calculating) will increase significantly. To test scaling with our present data we have studied how many calculations are required for a specific predictive accuracy as a function of the number of structures in the library. The results for a few different predictive accuracies are shown if Fig. 3. In each case, the results are averaged over 50 randomly chosen subsets of the structures. A linear slope implies that the relative

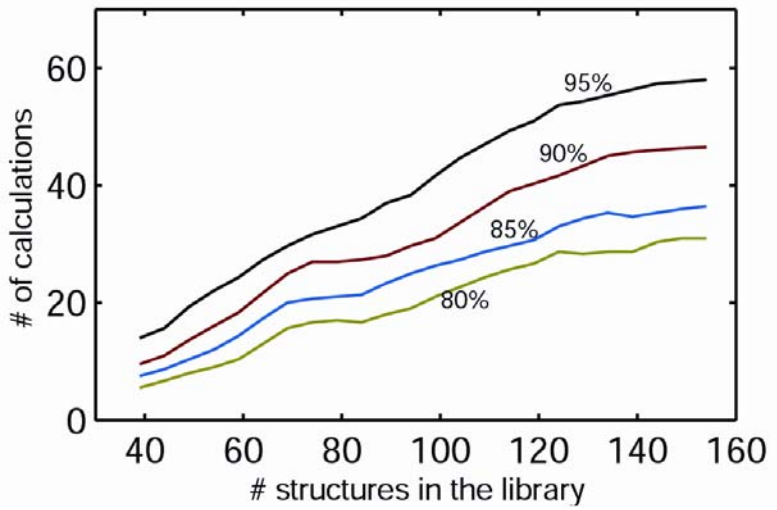

Figure 3: Number of calculations required to reach a specific predictive accuracy as a function of the number of structures in the library.

speedup is remaining constant with the number of structures in the library. It is encouraging to see that the curves seem to be leveling off as the library approaches 140 structures, suggesting an increase in the relative speedup. In our previous work on a smaller library [1], similar curves seemed to be leveling out at only 100 structures, but this previous library had fewer alloys. It is likely that the leveling out of the curves depends on the number and type of structures in the library, as well as the number and type of alloys. Therefore, at this point it is hard to predict at what size library the relative speedup provided by the present implementation of DMQC will begin to improve.

CONCLUSIONS

This paper gives a brief description of a novel method (DMQC) of accelerating crystal structure prediction with *ab initio* methods. We have established a large test database of 12,628 structural energies (154 structures and 82 alloys). The DMQC method is here implemented by using linear correlations among structural energies to guide sensible structure choices for a new alloy. We use principal component analysis to show that linear correlations among the structural energies do exist. Our tests show that DMQC can predict stable crystal structures about three times as fast as naïve random guessing. We demonstrate that the relative speedup compared to naïve guessing is likely to improve with a larger library, although it is not clear at what size this will occur.

ACKNOWLEDGEMENTS

This work was performed with the support from the National Science Foundation Information Technology Research (NSF-ITR) grant, number DMR-0312537, and the Department of Energy grant, number DE-FG02-96ER45571. This work was also

partially supported by computing resources from the National Partnership for Advanced Computational Infrastructure (NPACI).

REFERENCES

1. S. Curtarolo, D. Morgan, K. Persson*, et al.*, *Phys. Rev. Lett.* **91** (2003).
2. P. Villars, in *Intermetallic Compounds, Principle and Practice*, edited by J. H. Westbrook and R. L. Fleischer (John Wiley & Sons, New York, 1994), Vol. 1, Chapter 11, p. 227.
3. D. Morgan, J. Rodgers, and G. Ceder, *J. Phys.-Condes. Matter* **15**, 4361 (2003).
4. S. M. Woodley, P. D. Battle, J. D. Gale*, et al.*, *Phys. Chem. Chem. Phys.* **1**, 2535 (1999).
5. D. de Fontaine, in *Solid State Physics*, edited by H. Ehrenreich and D. Turnbull (Academic Press, 1994), Vol. 47, p. 33.
6. G. H. Johannesson, T. Bligaard, A. V. Ruban*, et al.*, *Phys. Rev. Lett.* **88**, art. no. (2002).
7. S. Curtarolo, *Coarse-Graining and Data Mining Approaches to the Prediction of Structures and their Dynamics,* PhD. Thesis (Massachusetts Institute of Technology, Cambridge, 2003).
8. J. P. Perdew and A. Zunger, *Phys. Rev. B* **23**, 5048 (1981).
9. G. Kresse and J. Furthmüller, *Comput. Mat. Sci.* **6**, 15 (1996).
10. S. Wold, A. Ruhe, W. H.*, et al.*, *SIAM J. Sci. Stat. Comput.* **5**, 735 (1984).
11. R. Kramer, *Chemometric Techniques for Quantitative Analysis* (Dekker, New York, 1998).
12. S. d. Jong, *Chemometrics and Intelligent Laboratory Systems* **18**, 251 (1993).
13. J. E. Jackson, *A User`s Guide to Principal Components* (John Wiley & Sons, New York, 1991).