Predicting crystal structure by merging data mining with quantum mechanics

CHRISTOPHER C. FISCHER[1], KEVIN J. TIBBETTS[1], DANE MORGAN[2] AND GERBRAND CEDER[1]*

[1] Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

[2] Department of Materials Science and Engineering, University of Wisconsin, Madison, Wisconsin 53706, USA

*e-mail: gceder@mit.edu

Published online: 9 July 2006; doi:10.1038/nmat1691

Modern methods of quantum mechanics have proved to be effective tools to understand and even predict materials properties. An essential element of the materials design process, relevant to both new materials and the optimization of existing ones, is knowing which crystal structures will form in an alloy system. Crystal structure can only be predicted effectively with quantum mechanics if an algorithm to direct the search through the large space of possible structures is found. We present a new approach to the prediction of structure that rigorously mines correlations embodied within experimental data and uses them to direct quantum mechanical techniques efficiently towards the stable crystal structure of materials.

Crystal structure occupies a central and often critical role in materials science, particularly when establishing a correspondence between a material's performance and basic composition[1]. It has long been known that a diverse set of properties ranging from mechanical to electronic are intimately tied to the underlying symmetry of the crystal structure of interest[2]. Bearing this in mind, the ability to predict the structure of a material 'from scratch' prior to synthesis would be particularly useful.

Owing to significant advances in both computational power and basic materials theory[3–5], it is now possible to accurately predict the zero kelvin energy of a compound using quantum mechanical methods. However, this in itself does not enable structure prediction as quantum mechanics cannot suggest which structures to test for stability, and direct optimization of the system's energy over the coordinates of all atoms is not possible due to the presence of many local minima. As a simplifying measure, researchers often round up a set of candidate crystal structures, calculate their energies, and proclaim the lowest one as the true stable state[6], that is, the ground state. Therefore, the quality of such a prediction is strongly dependent on the initial set of candidate structures assumed; a set often determined in an uncontrolled and biased manner.

Alternatively, a material's structure may be guessed by gleaning information from past experiments or heuristic arguments. This process ranges from correlating the unknown structure with known compounds of chemically similar alloys (for example, looking at 'neighbouring' elements in the periodic table), to the remarkably successful structure mapping techniques[7–10]. All of these methods rely on a simplified representation of structural stability on the basis of physically motivated driving forces such as electronegativity difference and size mismatch between the constituent elements, or simply location in the periodic table. Although such heuristic methods can create considerable insight into the physical mechanisms that control structure selection, their accuracy is limited by assumptions made regarding what constitute relevant physical parameters. Moreover, heuristic methods lack a mechanism to systematically improve on prediction ability. In this article, we demonstrate that the knowledge of some stable

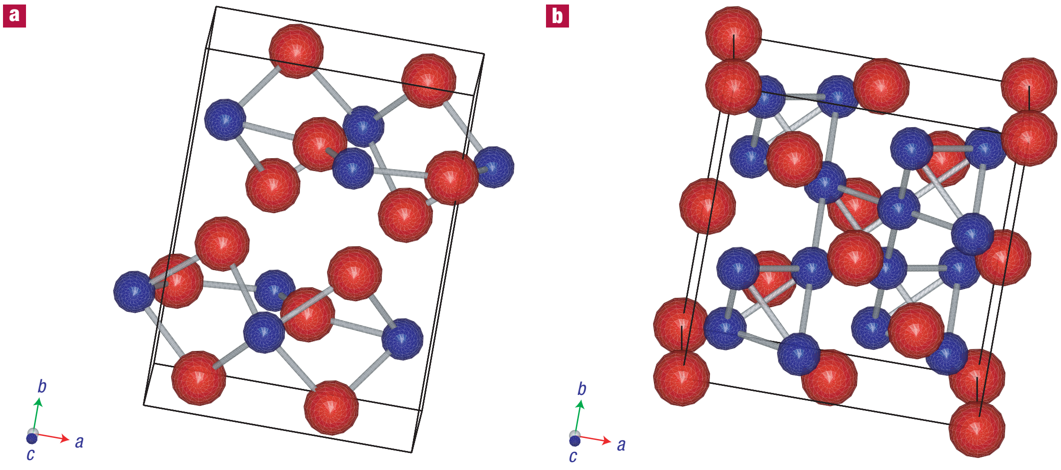

Figure 1 The Fe_3C and $MgCu_2$ structure types. a,b, In Fe_3C (**a**) and $MgCu_2$ (**b**) the coordination of large (red: Fe, Mg) and small (blue: C, Cu) atoms is quite different, leading to a high energy on interchange of their positions.

structures in an alloy provides important information regarding the true interatomic interactions present. As such, we leverage this knowledge to augment predictions at other compositions—a capability not available in previously developed heuristic methods.

It is our belief that the problem of predicting crystal structure can be largely solved by combining modern quantum mechanical methods with machine learning techniques[11] into a common framework. A machine learning method captures the physics correlating crystal structures in nature, and quantum mechanics provides the final accuracy. In our approach, information extracted from a large body of historical data is used to suggest probable crystal structures for a new material, effectively steering detailed calculations to an informed set of candidates on the basis of historical knowledge. Our prescription for codifying correlations in historical data breaks from previously developed methods through the use of a mechanism-independent formalism, described later in this article. As such, our viewpoint provides a representation of historical data that remains tremendously informative, yet unbiased as to the important physics. This should give it broad applicability to diverse materials classes.

We first outline our general approach for constructing an informatics-based structure suggestion model, Data Mining Structure Predictor (DMSP), then show how it is 'informed' from large amounts of historical data, and finally demonstrate its effectiveness in suggesting probable ground-state structures with a specific prediction and a large-scale analysis of intermetallics.

At each of n discrete compositions, let the variable x_i indicate the crystal structure, or lack thereof, present in the alloy at composition i (that is, the domain of x_i extends over the set of non-equivalent crystal structure types that can occur at composition i, including a value representing the lack of a compound). Likewise, the variables $x_{E1}, x_{E2}, \ldots, x_{EC}$ identify C different constituent elements, and the $(n+C)$-tuple $\mathbf{X} = (x_{E1}, x_{E2}, \ldots, x_{EC}, x_{c_1}, \ldots, x_{c_n})$ fully specifies the ground states of an alloy system. Define furthermore a probability function, $p(\mathbf{X})$, giving the probability that $\mathbf{X}$ is the set of ground states in a binary alloy. To predict crystal structures in an alloy, the probability of unknown variables conditioned on available information about the system ($\mathbf{e}$) needs to be evaluated, that is, determine $p(\mathbf{X}|\mathbf{e})$. Available information typically consists of the constituent elements and their structure (for example, the gold–silver system would be represented with $\mathbf{e} = (x_{E1} = \text{Ag}, x_{E2} = \text{Au}, x_0 = \text{face-centred cubic}, x_1 = \text{face-centred cubic})$), but can naturally include knowledge regarding the structure of compounds at intermediate compositions. To explicitly construct $p(\mathbf{X})$, we use a method used by Morita[12] in his study of alloys expressing $p(\mathbf{X})$ as a generalized cumulant expansion starting with the formula:

$$p(\mathbf{X}) = \prod_i p(x_i) \prod_{j<k} g^{(2)}(x_j, x_k) \prod_{l<m<n} g^{(3)}(x_l, x_m, x_n) \cdots. \qquad (1)$$

In this expansion single variable terms, such as $p(x_i)$, capture variations in $p(\mathbf{X})$ due to variables behaving independently, whereas those such as $g^{(2)}(x_j, x_k)$ and $g^{(3)}(x_l, x_m, x_n)$, referred to as cumulant functions, reproduce correlation among pairs and triplets of structures respectively. The cumulant functions embed physics as they express which structures tend to occur together in systems because they satisfy similar underlying atomic interactions. Note that strong anticorrelation (cumulants close to zero indicate that structures never occur together) is just as useful in constructing a probability function.

It is through the cumulant functions that $p(\mathbf{X})$ will be 'informed', and the Methods section describes in more detail how cumulants are obtained from observations in the database. First, we show that substantial structure–structure correlation exists in the observations made for binary metallic alloys and demonstrate that such mathematical correlation implicitly reflects underlying physics.

Structure correlations were extracted from the Pauling File[13], one of the largest databases of binary crystal structures. We restricted our study to low-temperature and low-pressure entries in binary alloys of metallic elements. As only two constituents E1 and E2 are present, we hereafter refer to them as A and B to follow common nomenclature. Details of the data preparation can be found in the Methods section. All crystal structures occur in the database with a particular frequency. However, as might be suspected, some pairs of crystal structures occur together more or less often due to apt or conflicting compatibilities with the underlying interatomic interactions. We quantify correlation between two crystal structures through the ratio $f(x_i, x_j) = p(x_i, x_j)/p(x_i)p(x_j)$, where x_i and x_j are variables representing crystal structures occurring at compositions i and j in the alloy,

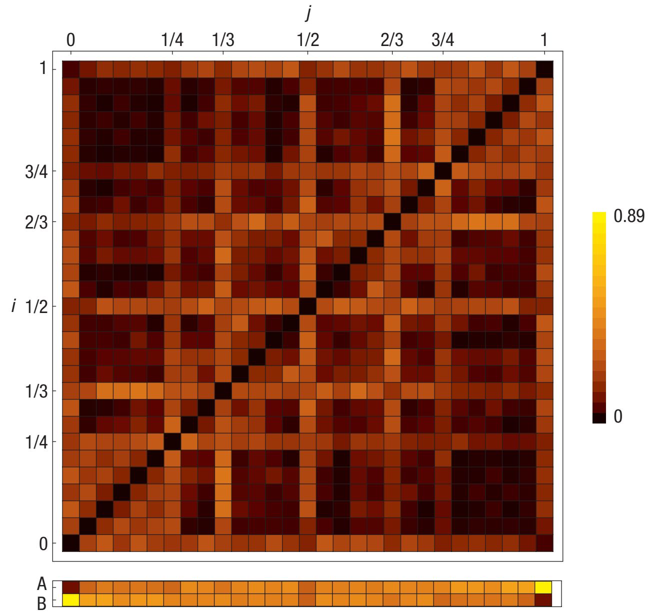

Figure 2 Mutual information between pairs of variables. Highly correlated variables appear as bright pixels, whereas uncorrelated pairs appear dark. Correlation is present in two distinct forms, occurring between structures at different compositions, as well as between structures and their constituent elements (bottom-most rows). Each value shown is scaled to fall in the range [0, 1] (see the Methods section for further details).

$p(x_i,x_j)$ is the probability that both will occur in the same binary system picked at random, and $p(x_i)$ is the probability for the structure occurring at composition i (in this article, as in ref. 13, structures are identified by their prototype name). Molecular level interactions driving some structures to frequently occur together, whereas others rarely together, manifest themselves as a correlation ratio strongly differing from 1. For example, we observe a strong correlation between the Fe_3C-type structure at a composition of AB_3 and the $MgCu_2$-type structure at A_2B. These two structures occur together in 52 of the 87 alloys in which Fe_3C is present, and using the fact that $MgCu_2$ occurs in 7.04% of the alloy systems, the correlation ratio f is 8.49. In other words, given that Fe_3C is present at AB_3, it is 8.49 times more likely that $MgCu_2$ will form at A_2B than if the structures were uncorrelated. In this case, the correlation can be easily understood as both structures (Fig. 1) form in systems where the two constituent elements A and B are of very different size. The Fe_3C and $MgCu_2$ structures are very unstable when the roles of small and large atoms are interchanged, and the correlation ratio for Fe_3C forming at the same composition, AB_3, but $MgCu_2$ at AB_2 (instead of A_2B) reveals strong anticorrelation ($f(x_{2/3}=\mathrm{MgCu}_2, x_{3/4}=\mathrm{Fe}_3\mathrm{C})\approx 0$), as the position of smaller and larger atoms cannot be interchanged in a size-effect stabilized crystal structure.

In many situations, it may be known that a crystal structure forms at a particular composition, but what occurs at other compositions remains unknown. An important question regarding the use of correlation in a predictive setting is the degree to which knowledge that a particular structure occurs at one composition influences what should occur at other compositions. This concept is represented mathematically by the mutual information[14] between compositions i and j, given by $I_{i,j}=\sum_{x_i,x_j}p(x_i,x_j)\ln\left[p(x_i,x_j)/(p(x_i)p(x_j))\right]$, where the sum extends over all structures x_i and x_j that can occur at compositions i and j. In Fig. 2, we show the mutual information between compositions, as well as between a composition and a constituent element (the brighter the more mutually informative). Bright pixels for rows labelled A and B in Fig. 2 reiterate the fact that crystal structure in a binary alloy is strongly influenced by the identity of its constituent elements, giving substantial credence to methods empirically relating structure to an elemental property[7,10,15]. Correlation between the structure of $A_{1-c}B_c$ and element A (bottom-most row) is understandably very strong for dilute B compositions, but decreases only slowly with increasing B content. As we have suggested on the basis of physical arguments, structure–structure correlation should also be strong. The brighter columns and rows found along common compositions (1/4, 1/3, 1/2, and so on) are indicative of strong structure–structure correlation. For example, knowledge of what occurs at composition $c=1/2$ provides significant information as to what structures may form at other compositions—reflecting that interactions present at $c=1/2$ are indicative of interactions present at other compositions.

Given that we find substantial correlation information in historical crystal structure data, a DMSP should have substantial predictive capability. To illustrate the method and its effectiveness we describe a specific prediction, and summarize by showing the method's prediction accuracy over a large sample of alloys. Our specific prediction is made for the structure occurring at composition $c=0.75$ in the silver–magnesium system, summarized in Fig. 3. Experimental studies in this system indicate a stable structure type near 75% magnesium content[16], but the exact nature of the structure is undetermined. Additional information, available for the structure of the elements and for structures occurring at compositions AgMg and Ag_3Mg, is summarized in Fig. 3a. We have queried the DMSP model using this evidence to list the most probable structures for $AgMg_3$ (hereafter called the candidate list), of which the top five are shown in Fig. 3a. To verify this prediction, we proceeded to compute the formation enthalpies of the top 10 structures on the candidate list, as well as 26 other common structures[17] at $c=0.75$ in the generalized gradient approximation to density functional theory.

The $Cu_{2.82}P$ structure, suggested as the most likely candidate by our DMSP approach, is clearly the most stable of the structures calculated. Moreover, when related to a large-scale test of our DMSP method outlined in the next section, we estimate with 95% confidence that $Cu_{2.82}P$ is the true ground state among the 71 crystal structures occurring at AB_3 in our dataset. It is important to note that according to practices commonly found in the literature, $Cu_{2.82}P$ would probably not be tested for stability. This is primarily because $Cu_{2.82}P$ is an uncommon structure, occurring in a scant nine alloys in the Pauling File. Furthermore, the size of the $Cu_{2.82}P$ unit cell (24 atoms) alone would discourage calculation through quantum mechanical methods owing to the large computational overhead in calculating its energy (in fact, we know of no other *ab initio* calculation of this structure). In contrast, our method ranks $Cu_{2.82}P$ as a very likely, although non-obvious, candidate on the basis of available experimental correlations. The ability to predict rare, complicated structure types such as $Cu_{2.82}P$ is an important feature of our method and illustrates that structure correlation can significantly influence the order in which ground states should be searched for.

We have evaluated the predictive capability of our DMSP approach in a more general context, spanning a wide range of chemistries, by carrying out cross-validated[18] predictions of all 3,975 compounds occurring in two or more alloys in our dataset. For each prediction, we first remove data associated with the alloy of interest from the database so that no knowledge of that

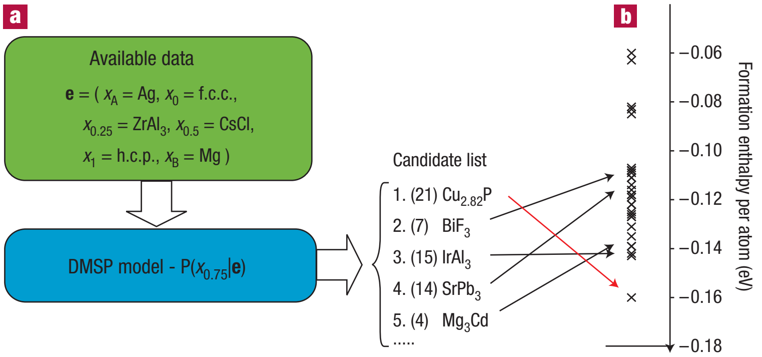
a
Available data
e = (x_A = Ag, x_0 = f.c.c.,
$x_{0.25}$ = $ZrAl_3$, $x_{0.5}$ = CsCl,
x_1 = h.c.p., x_B = Mg)
DMSP model - P($x_{0.75}$|e)
Candidate list
1. (21) $Cu_{2.82}P$
2. (7) BiF_3
3. (15) $IrAl_3$
4. (14) $SrPb_3$
5. (4) Mg_3Cd
.....
b
−0.06
−0.08
−0.10
−0.12
−0.14
−0.16
−0.18
Formation enthalpy per atom (eV)

Figure 3 Predicting the structure of $AgMg_3$. a, DMSP prediction (candidate list) of the crystal structure of $AgMg_3$ on the basis of the limited data available at other compositions (green box). The structures are ordered by decreasing probability within the DMSP model. This ordering is compared with a ranking on the basis of the frequency with which these structures occur in the experimental database (parenthesized value in candidate list). **b**, *Ab initio* formation enthalpy (with respect to the pure elements) of the top five structures along with 26 additional structure types calculated to aid in verifying the prediction.

specific alloy is embedded in the correlations. Then we attempt to predict each compound using knowledge of the elements and the structure of the other compounds in the system. As before, the prediction consists of generating a candidate list of structures at each composition of interest, ranked by the probability of occurrence; providing a natural order in which structures should be tested for stability. In Fig. 4, we show the number of structures that would have to be investigated on the ordered candidate list to find the true structure with a particular probability. Our new DMSP method (blue curve in Fig. 4) performs remarkably well, capturing the observed compound 90% of the time if we are willing to investigate the top five structures at each composition. We have compared the DMSP method with two simpler data-mining schemes. An obvious approach would be to pick trial ground states on the basis of the frequency with which they occur in nature (red curve in Fig. 4). Investigating five structures with such a frequency-based method obtains only 62% accuracy. As a point of reference, the result of guessing structures with a uniform probability (at random) is also shown (green curve in Fig. 4). These results demonstrate that a considerable amount of historical information is used when making predictions with our DMSP method.

Our results confirm that correlation captured from experimental data substantially constrains the structures that may occur together in alloy systems (by significantly updating the probability, $p(\mathbf{X}|\mathbf{e})$, assigned to the possible ground-state combinations $\{\mathbf{X}\}$). Correlations between structures at different compositions in an alloy are ultimately the result of similar atomic interactions between constituent atoms, and as such, we expect DMSP to be effective in identifying ground states in multicomponent alloys where the potential use of structure prediction is great.

Predicting crystal structure with absolute certainty is currently an intractable problem in materials science. Pragmatic solutions will require both an accurate energy model, which can be provided by modern quantum mechanical methods, as well as an efficient algorithm to search through the space of possible crystal structures. By constructing a mathematical representation of historical knowledge we have been able to combine the suggestive nature of heuristic methods with the accuracy of quantum mechanics into a highly efficient structure prediction algorithm. This combined data-mining and quantum mechanics approach stands in contrast to the computational procedures used until now where historical information only enters through the experience of the user. The statistical nature of the method also shows a route to attaching a level of confidence to first-principles-based structure predictions. Interested readers can find an online version of the method presented here and additional data available at http://datamine.mit.edu.

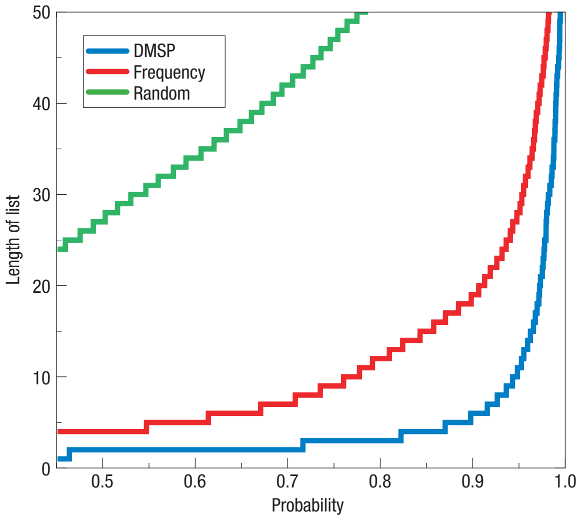

Figure 4 Large-scale prediction capability. Length of candidate list required to contain the true structure with a certain probability with the DMSP method, 'random' structure guessing, or suggesting structures on the basis of the 'frequency' with which they occur in nature.

METHODS

MUTUAL INFORMATION

In Fig. 2, the mutual information, $I_{i,j}$, between each pair of variables is scaled by $\min(H_i, H_j)$, where $H_i = -\sum_{x_i} p(x_i)\ln p(x_i)$. This convention is adopted so that a value of 1 will consistently mean that the pair of variables, x_i and x_j, are perfectly correlated. Mutual information is greater than or equal to zero, symmetric in its indices $I_{i,j} = I_{j,i}$, and can be written as $I_{i,j} = H_j - H_{j|i} = H_i - H_{i|j}$, where $H_{i|j} = -\sum_{x_i,x_j} p(x_i,x_j)\ln p(x_i|x_j)$. Noting that $0 \leq H_{i|j} \leq H_i$, the mutual information will fall in the range $0 \leq I_{i,j} \leq \min(H_i, H_j)$.

DATA PREPARATION

The experimental data for this study was obtained from the Pauling File[13] database containing 28,457 entries of experimentally determined structure types in 2,600 binary alloys. High-temperature and pressure entries were removed for a representation closer to the true set of ground states. This study was restricted to 'metallic' alloys by removing all alloy systems containing at least one of the elements: He, B, C, N, O, F, Ne, Si, P, S, Cl, Ar, As, Se, Br, Kr, Te, I, Xe, At, or Rn.

COMPOSITION BINNING

Structure entries in ref. 13 are recorded at any composition, in some cases reflecting more the actual composition of the material, than the stoichiometry of the structure prototype. However, the majority of entries in the Pauling File occur at rational fraction-valued compositions which were used to discretize composition space. This set of compositions consists of the following values and their symmetric, $(1-c_i)$, counterparts: 0, 1/10, 1/7, 1/6, 1/5, 2/9, 1/4, 2/7, 3/10, 1/3, 3/8, 2/5, 3/7, 4/9, 1/2. Each entry in our dataset was binned in an iterative procedure to one of the aforementioned compositions. Following binning, multiple entries in the same alloy with the same composition and prototype were considered duplicates and reduced to a single entry. After data preparation, 4,836 entries remained distributed in 1,335 alloy systems.

DETAILS OF CUMULANT EXPANSION

The full expansion (equation (1)) contains no approximation, as the product extends over all possible groupings of variables. However, to represent a joint probability distribution over many variables in a practical way, equation (1) is truncated after including all single- and two-variable terms. Our approximation for $p(\mathbf{X})$ is written as:

$$p(\mathbf{X}) = \frac{1}{Z}\prod_i p(x_i)\prod_{j<k} g^{(2)}(x_j,x_k), \qquad (2)$$

where the factor Z^{-1} is included to ensure normalization. Morita identified the cumulant functions[12], expressing each in terms of the marginals of $p(\mathbf{X})$. The two-variable terms are identified as

$$g^{(2)}(x_i,x_j) = \frac{p(x_i,x_j)}{p(x_i)p(x_j)}.$$

PARAMETER ESTIMATION

In equation (2) the set of 'point' and 'pair' probabilities, denoted $\{p(x_i)\}$ and $\{p(x_j,x_k)\}$, encode the correlation information relevant for the DMSP model. These probabilities are estimated from available data using Bayesian parameter estimation[19] outlined below.

Suppose that in a set of N_{sys} alloy systems, k different structures, denoted $\{\alpha_1,\alpha_2,\ldots,\alpha_k\}$, occur at composition i. Let $N_{x_i=\alpha_j}$ indicate the number of times structure α_j occurs at composition i, and $\mathbf{N}_i = (N_{x_i=\alpha_1}, N_{x_i=\alpha_2},\ldots,N_{x_i=\alpha_k})$ denote the set of such counts for all structures occurring at composition i. Let the parameter $\theta_{x_i=\alpha_j}$ denote the true probability that an alloy chosen at random will contain the structure α_j at composition i, and $\boldsymbol{\theta}_i = (\theta_{x_i=\alpha_1},\ldots,\theta_{x_i=\alpha_k})$ denote the set of such parameters for composition i. Given incomplete, and possibly noisy data $(\mathbf{N}_i)$, all values of $\theta_{x_i=\alpha_j} \in [0,1]$ are possible, but occur with different probabilities depending on both the data and any relevant prior information. The function $p(\boldsymbol{\theta}_i|\mathbf{N}_i)$ represents the probability density of the true parameter vector given the available data, and is determined with Bayes' rule

$$p(\boldsymbol{\theta}_i|\mathbf{N}_i) = p(\mathbf{N}_i|\boldsymbol{\theta}_i)\frac{p(\boldsymbol{\theta}_i)}{p(\mathbf{N}_i)}. \qquad (3)$$

In equation (3), the term $p(\mathbf{N}_i|\boldsymbol{\theta}_i)$ is the well-known multinomial distribution, whereas $p(\boldsymbol{\theta}_i)$ and $p(\mathbf{N}_i)$ are referred to as prior probabilities (that is, probabilities assigned prior to collecting data using only information available about the problem at hand). For $p(\boldsymbol{\theta}_i)$, we have used a 'uniform minimum information Dirichlet' distribution[20] over allowed values of $\boldsymbol{\theta}_i$. Values of the point probability function, $p(x_i)$, are found by averaging $\boldsymbol{\theta}_i$ over all possible values under the distribution $p(\boldsymbol{\theta}_i|\mathbf{N}_i)$.

$$\langle\boldsymbol{\theta}_i\rangle = \int \boldsymbol{\theta}_i \times p(\boldsymbol{\theta}_i|\mathbf{N}_i)\times\delta\left(\sum_{j=1}^{k}\theta_{x_i=\alpha_j}-1\right)\times \mathrm{d}\boldsymbol{\theta}_i.$$

It is important to note that the average above is influenced by both the peak and the spread of the function $p(\boldsymbol{\theta}_i|\mathbf{N}_i)$ (refs 11,19). As more data become available, the function $p(\boldsymbol{\theta}_i|\mathbf{N}_i)$ becomes more-strongly peaked around the maximum likelihood estimate for $\boldsymbol{\theta}_i$. This averaging process leads to the estimate (in this case for structure α_j at composition i where k different structures can occur):

$$p(x_i=\alpha_j) = \langle\theta_{x_i=\alpha_j}\rangle = \frac{N_{x_i=\alpha_j}+1/k}{N_{\mathrm{sys}}+1}.$$

Note that the estimate, $\langle\theta_{x_i=\alpha_j}\rangle$, will be equivalent to the frequency that α_j occurs at composition i in the limit of large datasets and that a probability of 0 or 1 is only obtained in the same limit. An analogous process for structures α_l and β_m occurring at compositions i and j yields:

$$p(x_i=\alpha_l,x_j=\beta_m) = \langle\theta_{x_i=\alpha_l,x_j=\beta_m}\rangle = \frac{N_{x_i=\alpha_l,x_j=\beta_m}+1/(np)}{N_{\mathrm{sys}}+1},$$

where n and p different structures occur at compositions i and j respectively, and $N_{x_i=\alpha_l,x_j=\beta_m}$ is the number of times that both structure α_l occurs at composition i and β_m occurs at j. Bayesian estimates differ from maximum likelihood estimates in a minor way numerically, but allow our method to make predictions even for cases where $N_{x_i=\alpha_l,x_j=\beta_m}=0$.

CALCULATIONS

Ab initio electronic-structure calculations were carried out using density functional theory in the generalized gradient approximation, using the projector augmented-wave method[21], as implemented in the Vienna *ab initio* Simulation Package[22], using the Perdew–Wang parameterization (PW91) of the exchange correlation energy. Kohn–Sham orbitals were expanded in a plane wave basis up to an energy cutoff of 405 eV. Brillouin zone integrations were carried out using at least 2,500/(number of atoms in unit cell) k-points distributed as uniformly as possible on a Monkhorst–Pack mesh[23] (Γ-centred for hexagonal cells) and a Methfessel–Paxton[24] smearing of the occupied states. Calculations are at zero temperature and pressure and all structures were fully relaxed.

Received 18 January 2006; accepted 1 June 2006; published 9 July 2006.

References

1. Olson, G. Designing a new material world. *Science* **288,** 993–998 (2000).
2. Nye, J. F. *Physical Properties of Crystals* (Oxford Univ. Press, Oxford, 1985).
3. Wolverton, C., Yan, X.-Y., Vijayaraghavan, R. & Ozoliņš, V. Incorporating first-principles energetics in computational thermodynamics approaches. *Acta Mater.* **50,** 2187–2197 (2002).
4. Asta, M., Ozoliņš, V. & Woodward, C. A first-principles approach to modeling alloy phase equilibria. *J. Mater.* **53,** 16–19 (2001).
5. Curtarolo, S., Morgan, D. & Ceder, G. Accuracy of *ab initio* methods in predicting the crystal structures of metals: A review of 80 binary alloys. *CALPHAD* **29,** 163–211 (2005).
6. Ceder, G. Predicting properties from scratch. *Science* **280,** 1099–1100 (1998).
7. Pettifor, D. G. The structures of binary compounds: I. Phenomenological structure maps. *J. Phys. C* **19,** 285–313 (1986).
8. Villars, P. A three-dimensional structural stability diagram for 998 binary AB intermetallic compounds. *J. Less Common Met.* **92,** 215–238 (1983).
9. Morgan, D. & Ceder, G. in *Handbook of Materials Modeling* Vol. 1 (eds Catlow, R., Shercliff, H. & Yip, S.) 395–421 (Kluwer Academic, Dordrecht, 2005).
10. Morgan, D., Rodgers, J. & Ceder, G. Automatic construction, implementation and assessment of Pettifor maps. *J. Phys. C* **15,** 4361–4369 (2003).
11. Russell, S. & Norvig, P. *Artificial Intelligence: A Modern Approach* (Prentice Hall, Upper Saddle River, 1995).
12. Morita, T. Cluster variation method of cooperative phenomena and its generalization I. *J. Phys. Soc. Japan* **12,** 753–755 (1957).
13. Villars, P. *The Pauling File Inorganic Materials Database and Design System—Binaries Edition (CD-ROM)* (ASM International, Ohio, 2002).
14. Cover, T. M. & Thomas, J. A. *Elements of Information Theory* (Wiley, New York, 1991).
15. De Boer, F. R. *Cohesion in Metals* (North-Holland, Amsterdam, 1988).
16. Prokof'ev, M. V., Kolesnichenko, V. E. & Karonik, V. V. Composition and structure of alloys in the Mg-Ag system near Mg_3Ag. *Inorg. Mater.* **21,** 1168–1170 (1985).
17. Curtarolo, S., Morgan, D., Persson, K., Rodgers, J. & Ceder, G. Predicting crystal structures with data mining of quantum calculations. *Phys. Rev. Lett.* **91,** 135503 (2003).

18. Stone, M. Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. B* **36,** 111–147 (1974).
19. Jaynes, E. T. *Probability Theory: The Logic of Science* (Cambridge Univ. Press, New York, 2003).
20. Cheeseman, P. & Stutz, J. in *Advances in Knowledge Discovery and Data Mining* (ed. Fayyad, U. M. *et al.*) 61–83 (AAAI Press, Menlo Park, California, 1996).
21. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50,** 17953 (1994).
22. 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).
23. Monkhorst, H. J. & Pack, J. D. Special points for Brillouin-zone integrations. *Phys. Rev. B* **13,** 5188–5192 (1976).
24. Methfessel, M. & Paxton, A. T. High-precision sampling for Brillouin-zone integration in metals. *Phys. Rev. B* **40,** 3616–3621 (1989).

Acknowledgements

This work was funded by NSF-ITR grant DMR-0312537, the Institute for Soldier Nanotechnologies (ISN) grant DAAD 19-02-D-0002, and the DOE, Office of Basic Energy Sciences under Contract No. DE-FG02-96ER45571. We would also like to acknowledge the National Science Foundation and the San Diego Supercomputer Center for additional computational resources.
Correspondence and requests for materials should be addressed to G.C.

Competing financial interests

The authors declare that they have no competing financial interests.