Meas. Sci. Technol. **16** (2005) 296–301 doi:10.1088/0957-0233/16/1/039

High-throughput and data mining with *ab initio* methods

Dane Morgan[1], **Gerbrand Ceder**[1] **and Stefano Curtarolo**[2]

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

[2] Department of Mechanical Engineering and Materials Science, Duke University, Durham, NC 27708, USA

Received 25 April 2004, in final form 20 August 2004
Published 16 December 2004
Online at stacks.iop.org/MST/16/296

Abstract
Accurate *ab initio* methods for performing quantum mechanical calculations have been available for many years, but their speed, complexity and instability have generally constrained researchers to studying only a few systems at a time. However, advances in computer speed and *ab initio* algorithms have now created fast and robust codes, where large numbers of calculations can be performed automatically, making it possible to do high-throughput *ab initio* computation. High-throughput computations can be used to efficiently screen and optimize for desired properties in broad classes of materials, as well as create large databases for data mining applications that can guide both experiments and further calculations. This paper discusses some of the challenges associated with preparing, running, collecting and assessing *ab initio* results in a high-throughput framework. An example application is given in the area of crystal structure prediction for binary alloys. The high-throughput results are in good agreement with known data, and suggest many possible new compounds not yet seen experimentally. Data mining techniques are used to find correlations among structural energies, and the correlations are then used to accelerate identification of stable crystal structures in new alloys.

Keywords: data mining, high throughput, *ab initio* computation

(Some figures in this article are in colour only in the electronic version)

1. Introduction

Over the last 40 years, *ab initio* methods have become ubiquitous tools in chemistry, physics and materials science [1, 2]. Here we define *ab initio* methods as those that accurately solve the fundamental quantum mechanical equations (Schrödinger or Dirac) for the electrons of a system. The prominence of these methods arises from their ability to accurately calculate many properties for a wide range of systems with no need for pre-existing experimental or empirical knowledge. Developments in computing power and *ab initio* techniques have made it possible to perform orders of magnitude more calculations than in the past, and there is increasing interest in generating large amounts of data through what we will call high-throughput *ab initio* computation.

The advent of high-throughput *ab initio* computation has been made possible by a number of advances. Most obviously, the increasing speed of computers has led to a proportional increase in the number of systems that can be calculated in a fixed time. Simultaneously, modern *ab initio* codes have become very robust, allowing many tasks to be done automatically, with minimal human intervention, thereby decreasing the amount of human time required for setting up, checking and possibly fixing each calculation. This combination of robust methods with ever increasing computing power makes high-throughput computational screening possible.

The promise of high-throughput *ab initio* computation is analogous to that of high-throughput (combinatorial) experiments. By studying large numbers of systems, one can screen combinatorial spaces for new systems with desired properties. Atomistic computation-based screening has been a tool for many years in drug design [3], but it has not been practical to utilize the full power of *ab initio* methods. The introduction of *ab initio* screening will allow exploration of many properties that cannot be reliably calculated without complete quantum mechanical methods (e.g., electronic structure properties). *Ab initio* screening will be particularly important in materials science, where it has proved very challenging to find acceptably accurate and broad non-quantum mechanical atomistic models.

Section 2 will discuss some of the challenges that must be overcome for successful high-throughput *ab initio* calculations, and section 3 will discuss using high-throughput methods for *ab initio* crystal structure prediction. Section 4 will summarize the results, and provide a discussion of some key issues and other work in the field.

2. High-throughput *ab initio* calculation methods

A high-throughput study can be divided into three parts—preparing the input files and run parameters, running the calculation and collecting and assessing the results. For high-throughput methods to be successful, all these steps must be automated. In most cases, automation consists of the relatively straightforward process of writing programs to manage what would otherwise be done by hand—moving files, changing formats, extracting result values, etc. Most of these tasks can be performed relatively easily with simple scripts in a UNIX or LINUX environment. However, there are a few areas where it can be surprisingly difficult to use automated scripts to match the scrutiny to which a human would subject the output. The problem we will discuss in more detail is one in which we generated a high-throughput *ab initio* database of structural parameters and energies for crystal structures in binary alloys (discussed further in section 3) An example of a particular calculation might be the determination of the fully relaxed atomic positions and energy of GaAs in the diamond cubic or face centred cubic structures. At this point, the database consists of 176 structures in each of 80 binary alloys.

Preparing input files and run parameters consists of both setting up initial structures for all the alloys (used as a starting point to find the parameters of the unit cell that give the lowest energy), and setting optimal parameters for the *ab initio* codes. Because of the large number of crystal structures we wished to investigate, we needed a method to automatically choose the initial structures. This was done both by generating likely crystal structures as superstructures of face and body centred cubic lattices, and hexagonally closed packed lattices, and by extracting frequently occurring structure types from the CRYSTMET structure database [4]. The ability to interface with an experimental database to get structures that appear frequently in nature was very important for getting a relevant database. Initial volumes for each alloy structure were taken to be the concentration weighted volumes of the alloying elements, following Vegard's law [5].

Setting optimal *ab initio* run parameters is complicated by the fact that any tuning of the parameters to a specific alloy and structure has to be automated. As an example, consider the choice of how many *k*-points to use in each calculation. As part of a typical *ab initio* computation, electronic wavefunctions are determined at a set of *k*-points in the reciprocal space of the crystal. One can think of these *k*-points as a grid over which a numerical integration of the charge density and wavefunctions is performed. The more *k*-points used, the higher the numerical accuracy of the result. When working with just one system, one generally checks carefully for convergence with number of *k*-points, attempting to use enough for accurate results, but no more, since large *k*-point meshes slow down the calculations. However, careful checking cannot be done for all the systems in a high-throughput application. Instead, testing was done on a small subset of systems to find the minimum number of *k*-points times number of atoms in the unit cell that gave acceptably well-converged results. This number was then used for all the remaining calculations, so that approximately the same *k*-point sampling density was used for all different sizes of unit cells of the structures.

Running the calculations for high-throughput applications requires large amounts of computer power, but the calculations are usually independent and parallelize almost perfectly. A simple Linux cluster is therefore an effective computing environment for high-throughput *ab initio*. We developed special software to efficiently use free cycles and balance loads on our clusters, thereby completing many CPU years of calculations without disturbing other users of the cluster. At the heart of high-throughput *ab initio* are the readily available codes with robust implementations of *ab initio* techniques. We have used the Vienna *ab initio* simulation package (VASP), which is one of the fastest, most stable and most complete packages available for *ab initio* calculation within the standard density functional theory approximations [6].

Collecting and assessing results is a very important step, since large scale automated calculations can easily produce surprising errors that are difficult to catch. For example, in optimizing a very large number of crystal structures it was found that many of them were unstable, and ended up relaxing to entirely different structure types than that in which they began. Perhaps the simplest and best known class of these transformations is the relaxation of an unstable face centred cubic (fcc) phase into a body centred cubic (bcc) equivalent along a Bain strain path [7]. This type of relaxation can occur with different orderings, creating transformations between different fcc and bcc phases. For example, in $Cd_{0.5}Y_{0.5}$, an initially unstable $L1_0$ structure relaxed to the stable experimental ground state B2 structure during the calculation (see figure 1 for a schematic picture of this transformation).

In order to track changes in structure type, automated structure comparison codes were developed which can compare structures using such characteristics as volume, coordination shells, bond distances and space group symmetry (as calculated by the publicly available code PLATON [8]). Some stable structures that we obtained do not match any of the 176 structures in our database. These novel types have been automatically checked against all known structure types in the CRYSTMET database to see if they are known in *any* system, and some have been shown to be entirely new structure types [9].

In all steps every effort has been made to automate procedures frequently done by hand, so that calculations can

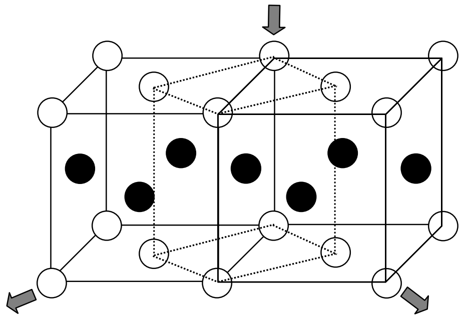

Figure 1. A schematic example of the fcc to bcc Bain strain path. The solid lines represent the conventional fcc cells and the dotted lines represent the conventional bcc cell. The arrows show how the $L1_0$ fcc structure becomes the B2 bcc structure (both are A1B1 superlattices along [100]).

be run with a minimum of human intervention. For the most recently assessed data of 176 structures in 80 alloys, we have performed 32 402 VASP calculations, and consumed a total of ~29 CPU years (estimated to be ~10^{18} total floating point operations). Details of the VASP calculations performed for this database can be found in [10, 11].

3. Application to crystal structure prediction

3.1. The problem of crystal structure prediction

Most materials properties, from band gaps to brittle fracture, melting temperature to magnetism, depend strongly on the structure of the materials involved. For this reason, crystal structure prediction has long been a fundamental problem in materials science, and identifying structure is still one of the key steps in many materials development projects. In addition, the growing interest in high-throughput *ab initio* methods and attendant materials screening efforts creates a new need for crystal structure prediction. Without a detailed knowledge of a material's structure, *ab initio* predictions of properties will often not be relevant to the real material. High-throughput calculations of properties are therefore constrained to materials with already known crystal structures, which greatly limits the potential of these computational methods to efficiently look for new materials with improved properties. Structure prediction is therefore a major challenge in high-throughput *ab initio* screening and a prerequisite for rational materials design with computational methods.

Ab initio methods are probably the best available tool for general structure prediction, since they can accurately calculate the relevant energies to determine stable structures. Unfortunately, there are so many possible structures that practically one cannot blindly enumerate every relevant structure for every alloy of interest. The goal of our study is therefore to combine high-throughput and data mining approaches to perform rapid, smart searches through the space of possible structures to find the most stable ones. As a first step in this process, we have used high-throughput approaches to construct a database which now contains fully relaxed electronic structure calculations for 176 structures of 80 binary alloys. There are a number of valuable things that can be done with this database. First, direct comparison to experiments can be used to make an unprecedented assessment of the accuracy and completeness of both the computational and experimental results. Furthermore, the database makes it possible to use data mining methods to establish patterns within the structural energies, providing guidance to efficiently predict structures for new alloys for which one does not want to calculate the energies of all 176 structures.

3.2. Comparing the database to experiment

We have performed comparison between *ab initio* predicted and experimental crystal structure types for 80 different alloy systems. In the experimental comparison, as with many high-throughput applications, some compromises must be made to make the large amounts of data tractable. In this case, the experimental data were taken almost entirely from two compilations, the Binary Alloy Phase Diagrams books [12] and the Pauling Files [13] (although more recent references were included when we were aware of them). The detailed comparison between the database and experiments is given in [11], and here we give a summary of the key results. To illustrate how much can be learned from the computed database, consider the cases of the Ag–Au and Ag–Pd binary alloy systems. Experimentally, no compound formation has been reported, and both systems are disordered (possibly with some short-range-order) at the high temperatures where they have been studied [12, 13]. However, calculated energies show that, for both systems, the elements do in fact have a significant ordering tendency, and that a rich series of ordered phases are to be expected if the alloys can be equilibrated at low temperatures. This can be seen clearly in the energy versus composition plots in figure 2, where the convex hulls show the predicted stable structures at zero temperature[3].

Restricting consideration to compounds for which both experimental and computational results were available made for 236 cases where the computational structural predictions could be compared with experimental measurements. In regions of composition that had not been studied experimentally, or that were experimentally assigned as two-phase or solid solution regions, we often predicted compounds that had not been seen by experiments. These cases created an impressive 96 *ab initio* predictions of compounds that might be observed through more detailed and/or effectively lower-temperature experiments. *Ab initio* results also provided predictions for an additional 21 compounds which were observed experimentally, but not identified reliably. For example, in the important electrocatalyst Pt–Ru, an ordered face centred cubic (fcc) phase was recently seen experimentally at $Pt_{0.5}Ru_{0.5}$, but the detailed atomic structure has not been identified [14]. The *ab initio* calculations suggest that $Pt_{0.5}Ru_{0.5}$ forms an fcc superlattice of A2B2 ordering along the [100] direction, as shown in figure 3.

In 110 cases, the *ab initio* and experimental results agreed on the low-temperature compound structure types, or that there were none (a phase-separating system). The agreements range from simple cases, such as predicting the single ordered C15 Ag_2Na phase in the Ag–Na system, to more complex phase diagrams, such as predicting all four

[3] In an energy versus composition plot, the convex hull is the set of lines that connect alloy phases such that all other phases are above the convex hull lines. The convex hull represents the free energy of the alloy at zero degrees Kelvin.

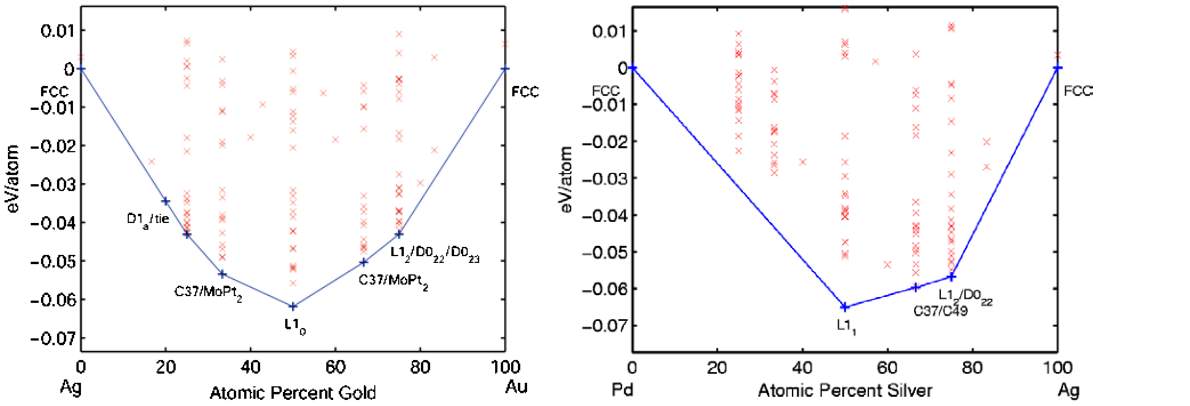

Figure 2. Computed energies of the Ag–Au and Ag–Pd alloy systems. The ‘x’ symbols denote energies of metastable structures and the ‘+’ symbols and connecting lines represent the stable convex hull. All results are for zero temperature. These figures are given, with additional details, in [11].

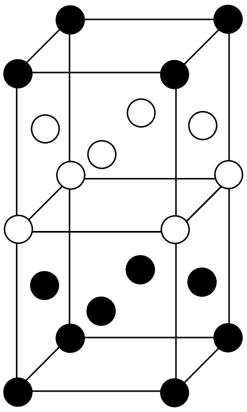

Figure 3. A2B2 ordering along [100] in a face centred cubic parent lattice. This is the stable structure suggested for $Pt_{0.5}Ru_{0.5}$ from the *ab initio* database.

known phases in the Al–Sc system [11, 12]. In only nine cases was there large disagreement between the experimental measurements and *ab initio* predictions of stable compounds. After careful analysis of these discrepancies, in only three cases were the experimental results complete enough that one could claim unambiguously that either the calculations or experiments were significantly in error. These include a failure to predict the $L1_2$ Cd_3Nb structure (calculations predict $L1_2$ to be unstable by ∼100 meV atom^{-1} with respect to phase separation into Cd and Nb), a failure to predict the B27 PtY structure (calculations predict B33 to be lower than B27 by ∼60 meV atom^{-1}) and a failure to predict the $D8_8$ Pt_3Zr_5 structure (calculations predict $D8_8$ to be ∼26 meV atom^{-1} above the nearest tie-line) [11, 12]. It should be noted that we have not explicitly included as errors calculated predictions of ordered phases in experimental two-phase regions, since it is difficult to know to what extent thermal effects are causing the difference between the zero-temperature calculations and the much higher temperature experiments. However, some cases are almost certainly problematic, such as Mo–Ti, where we obtain some calculated formation energies more negative than –200 meV atom^{-1}, while the experimental phase diagram shows a miscibility gap. These cases are discussed in more detail in [11]. Although the failures warrant further investigation, the fact that there are so few is very encouraging. These detailed comparisons with experiment demonstrate both the impressive accuracy of high-throughput *ab initio* methods, and their utility in providing a wealth of new predictions for alloy systems.

3.3. Data mining the database

Data mining methods are becoming increasingly prevalent in materials science applications [15]. High-throughput *ab initio* methods are now generating computed data on a scale where data mining algorithms can be used to determine useful patterns. This creates many opportunities for testing out traditional empirical ideas from materials science, as well as establishing new relationships that have not been guessed previously. We have used data mining methods to help in the structure prediction problem.

A data mining approach to structure prediction is particularly attractive, since patterns in structure formation among similar alloys are well known. In fact, many traditional empirical methods of structure prediction, such as structure maps [16], can be considered data mining methods. High-throughput *ab initio* has now made it possible to extend some of these crystal structure related data mining approaches beyond experimentally known data to calculated results.

The crystal structure prediction problem can be thought of in the following way: given a new alloy and a very large list of candidate structures, how can one order the search through the candidates so that the ones more likely to be stable are calculated first. The idea behind a data mining approach is to use results on previously calculated alloy systems to extract patterns that can guide the choice of good new candidate structures. The details of this method can be found in [10]. The following analysis will be done within the restricted space of the database we have developed, so that testing against the full calculations can be performed.

We make use of two data mining approaches to order the candidate structures for a new alloy. The first is to order candidate structures based on the frequency with which they are ground states for the alloys in the database (frequency ordering). This simply guarantees that very common structure types are first on the list of candidates to try. A more elaborate analysis is also employed, which uses partial least squares (PLS) linear regression [17] to establish correlations among the

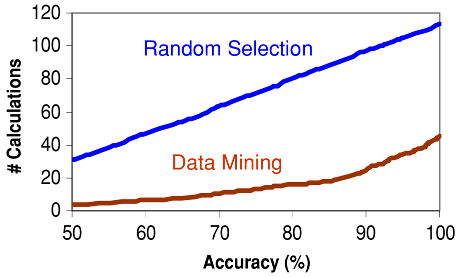

Figure 4. Comparison of the number of calculations required to reach a given predictive accuracy using random structure selection (top line) and data mining methods (bottom line).

structural energies within an alloy. These linear relationships can be used to predict energies for structures that have not been calculated, and these predicted energies are then used to order the candidate structures (PLS ordering). An iterative structure prediction algorithm is used which makes use of both these methods for ordering candidates, where the more elaborate PLS approach is used only when enough data have been collected to make it effective. For a new alloy, not yet calculated in the database, the structure prediction algorithm iteratively proposes likely stable structures, which are then calculated with the *ab initio* methods. The hope is that the data mining guided search through the structures will be much more efficient than just guessing.

The iterative data mining algorithm has been tested on all the alloys in the database, removing each alloy in succession and treating it as if it were new. The utility of the data mining algorithm is assessed by determining how few calculations are needed to predict correct ground states compared to simply choosing structures at random. The results are shown in figure 4, where it can be seen that, for the same average accuracy of ground state prediction, the data mining algorithm requires approximately four times fewer *ab initio* calculations than randomly guessing structures. Further refinement of the techniques and enhancement of the database could lead to significantly greater speed-up. The results demonstrate that high-throughput *ab initio* databases can be used in combination with data mining methods to attack practical problems such as crystal structure prediction.

4. Summary and conclusions

Advances in *ab initio* methods and computational speed have recently made high-throughput *ab initio* methods a reality. It is now possible to consider tens of thousands of calculations to construct large databases for screening compounds. The key ingredients needed to enable high-throughput *ab initio* are robust methods and automating tasks, as described in section 2.

Examples of high-throughput *ab initio* computation are still relatively few, are largely limited to the fastest and most robust *ab initio* methods (primarily density functional theory approaches) and are focused on the most straightforward properties for calculation (e.g., electronic structure, energies, bulk moduli). However, a range of interesting studies is emerging, and we give some selected examples here. Franceschetti and Zunger developed specialized high-throughput *ab initio* methods to predict band gaps in semiconductors, and combined these with simulated annealing optimization methods to find the ordered phase with the largest band gap in $Al_{0.25}Ga_{0.75}As$ [18]. Smithson *et al* have performed *ab initio* calculations of formation energies on almost 200 different transition metal hydrides (as well as a smaller number of electronic densities of states) in order to better understand metal-hydride formation [19]. Johannesson *et al* and Bligaard *et al* have used the very rapid *ab initio* linear muffin tin orbital method in the atomic sphere approximation (LMTO-ASA) in order to predict the stability, bulk moduli and lattice parameters of a very large set of alloys (they have over 64 000 calculations of different alloy structures). Data mining methods such as genetic algorithms and Pareto-optimal set approaches were used to optimize for low compressibility, high stability and low cost [20, 21]. Some of the same authors have also used *ab initio* methods to calculate the impurity surface segregation energies of 24 3d–5d transition metals, where each metal can be both the impurity and the host. This created a database of 552 host and impurity pairs, and allowed the authors to identify key factors governing surface segregation energetics [22].

As described in section 3, we have used plane wave pseudopotential methods to calculate formation energies and structural parameters for 80 binary alloys in 176 different structures, creating a database with over 14 000 alloy structures. Detailed comparisons to experiments have been made, proposing structures for a number of previously unidentified ordered compounds, and allowing perhaps the most comprehensive assessment to date of the accuracy of *ab initio* density functional methods for predicting structural energies [11]. We have also shown how the database can be combined with partial least squares data mining methods to accelerate future *ab initio* predictions [10, 23]. Widom and Mihalkovic have also assembled a database of *ab initio* structural and energetic information for assorted structures on over 200 binary, ternary and quaternary alloys (these data can be conveniently searched and visualized through a helpful website) [24]. A somewhat different example of what is made possible by high-throughput *ab initio* is the work of van de Walle *et al*, in constructing the alloy theory automated toolkit (ATAT) [25]. This toolkit is concerned with the prediction of alloy thermodynamic properties, but relies on being able to automatically perform hundreds or more *ab initio* calculations to extract energies and force constants. Such higher level automated tools are an exciting area enabled by high-throughput *ab initio* technology.

High-throughput *ab initio* screening has an enormous potential to impact design of compounds in materials science, chemistry and biology. Impact will come in two forms. The simplest is direct calculation of properties, such as structural parameters, energetics, elastic and optical properties, etc. For example, in section 3 we showed how direct comparison of high-throughput structural energetics to experiments suggested over a 100 possible new compounds that had not been identified previously. However, for many properties of interest, *ab initio* methods cannot efficiently or accurately directly calculate the desired information (e.g., materials properties such as hardness, melting temperature and corrosion resistance). In these cases, *ab initio* methods may be able to provide accurate descriptors upon which correlations to the desired properties can be built. This follows the spirit of quantitative structure–activity relationships (QSAR) and quantitative structure–property relationships (QSPR) methods, which attempt to predict complex properties from atomistic ones. An elegant example along these lines is the work of Chalk *et al*, who built a neural network to predict compound boiling points from their structures, as calculated with atomistic methods [26]. Chalk *et al* used potential models to determine three-dimensional molecular structures, but the general approach could easily be used with *ab initio* methods. In a similar spirit, Vitos *et al* used correlations between elastic constants (which can be calculated *ab initio*) and hardness and ductility (which are very difficult to calculate *ab initio*) to perform *ab initio* screening of mechanical properties of different stainless steels [27].

As large databases of *ab initio* data are produced, more data management and mining tools will need to be developed. For example, in section 3 we demonstrated how one can use a large database and data mining techniques to accelerate structure prediction in new alloys. The above-mentioned approach of correlating atomic scale descriptors with macroscopic properties will also require data mining methods. Further work in data mining tools, particularly in combining *ab initio* and experimental databases, will greatly extend the power of the data being developed.

Finally, we have stressed that, for materials design, high-throughput *ab initio* methods will require input crystal structures, and that crystal structure prediction has therefore become a newly pressing problem. As a beginning step in solving this problem, we have built a very large structural energy database, and used it both to directly predict crystal structures and develop data mining methods of accelerating structure prediction in the future. We hope that further development along these lines will produce an efficient general crystal structure prediction tool, which will enable high-throughput calculations to explore and discover in many new materials systems.

Acknowledgments

This work was supported by the Department of Energy, under contract number DE-FG02-96ER45571, and by the National Science Foundation, Information Technology Program, under contract number DMR-031253. Additional computing resources were provided by the National Science Foundation, National Partnership for Advanced Computing Infrastructure (NPACI).

References

[1] Jones R O and Gunnarson R O 1989 Density functional formalism, its applications and prospects *Rev. Mod. Phys.* **61** 689

[2] Lewars E G 2003 *Computational Chemistry: Introduction to the Theory and Applications of Molecular and Quantum Mechanics* (Boston: Kluwer)

[3] Marrone T J, Briggs J M and McCammon J A 1997 Structure-based drug design: computational advances *Ann. Rev. Pharmacol. Toxicol.* **37** 71

[4] White P S, Rodgers J and Le Page Y 2002 CRYSTMET: a database of structures and powder patterns of metals and intermetallics *Acta Cryst.* B **58** 343

[5] Vegard L 1921 Die Konstitution der Mischkristalle und die Raümfullung der Atome *Z. Phys.* **5** 17

[6] Kresse G and Furthmüller J 1996 Efficiency of *ab-initio* total energy calculations for metals and semiconductors using a plane-wave basis set *Comput. Mater. Sci.* **6** 15

[7] Bain E C 1924 *Trans. AIME* **70** 25

[8] Spek A L 2003 Single-crystal structure validation with the program PLATON *J. Appl. Crystallogr.* **36** 7

[9] Rodgers J 2004 Private communication

[10] Curtarolo S *et al* 2003 Predicting crystal structures with data mining of quantum calculations *Phys. Rev. Lett.* **91** 135503

[11] Curtarolo S, Morgan D and Ceder G 2004 Accuracy of *ab-initio* methods in predicting the crystal structures of metals: review of 80 binary alloys Submitted

[12] Massalski T B 1990 *Binary Alloy Phase Diagrams* (Materials Park, OH: ASM International)

[13] Villars P *et al* 2002 *Pauling File* (Materials Park, OH: ASM International)

[14] Camara G A *et al* 2002 Correlation of electrochemical and physical properties of PtRu alloy electrocatalysts for PEM fuel cells *J. Electroanal. Chem.* **537** 21

[15] Morgan D and Ceder G 2005 Data mining in materials development *Handbook of Materials Modeling* (vol 1: *Methods and Models of Materials Modeling*) ed R Catlow, H Shercliff and S Yip (Dordrecht: Kluwer)

[16] Villars P 1994 Factors governing crystal structures *Intermetallic Compounds, Principle and Practice* vol 1 ed J H Westbrook and R L Fleischer (New York: Wiley) chapter 11 p 227

[17] Wold S *et al* 1984 The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses *SIAM J. Sci. Stat. Comput.* **5** 735

[18] Franceschetti A and Zunger A 1999 The inverse hand-structure problem of finding an atomic configuration with given electronic properties *Nature* **402** 60

[19] Smithson H *et al* 2002 First-principles study of the stability and electronic structure of metal hydrides *Phys. Rev.* D **66** 144107

[20] Johannesson G H *et al* 2002 Combined electronic structure and evolutionary search approach to materials design *Phys. Rev. Lett.* **88** 255506

[21] Bligaard T *et al* 2003 Pareto-optimal alloys *Appl. Phys. Lett.* **83** 4527

[22] Ruban A V, Skriver H L and Norskov J K 1999 Surface segregation energies in transition-metal alloys *Phys. Rev.* B **59** 15990

[23] Morgan D, Curtarolo S and Ceder G 2004 Data mining approach to *ab-initio* prediction of crystal structure *Materials Research Society Symp. Proc.* vol 804 JJ.9.25.1

[24] Widom M and Mihalkovic M 2004 *Alloy Database* http://alloy.phys.cmu.edu

[25] van de Walle A, Asta M and Ceder G 2002 The alloy theoretic automated toolkit: a user guide *Calphad, Comput. Coupling Phase Diagr. Thermochem.* **26** 539

[26] Chalk A J, Beck B and Clark T 2001 A quantum mechanical/neural net model for boiling points with error estimation *J. Chem. Inform. Comput. Sci.* **41** 457

[27] Vitos L, Korzhavyi P A and Johansson B 2003 Stainless steel optimization from quantum mechanical calculations *Nature Mater.* **2** 25