

# Automatic construction, implementation and assessment of Pettifor maps

Dane Morgan<sup>1</sup>, John Rodgers<sup>2</sup> and Gerbrand Ceder<sup>1</sup>

<sup>1</sup> Massachusetts Institute of Technology, Department of Materials Science and Engineering,  
77 Massachusetts Avenue, Cambridge, MA 02139, USA

<sup>2</sup> Toth Information Systems Inc., 2045 Quincy Avenue, Gloucester, K1J 6B2, Canada

E-mail: dmorgan@mit.edu, info@TothCanada.com (John Rodgers) and gceder@mit.edu

Received 11 November 2002

Published 13 June 2003

Online at [stacks.iop.org/JPhysCM/15/4361](http://stacks.iop.org/JPhysCM/15/4361)

## Abstract

The ability to predict the crystal structure of a material, given its constituent atoms, is one of the most fundamental problems in materials research. There exist a number of empirical methods which make predictions by clustering existing experimental data, generally using a few simple physical parameters. Although Pettifor maps are perhaps the best known and most successful of these empirical methods, the implementation and assessment of Pettifor maps has not been formalized. Here we propose well-defined algorithms for transforming data from a standard materials crystal structure database into a Pettifor map, using the map to predict the crystal structure for a new system, and assessing the predictive accuracy of the map. We introduce the idea of a candidate crystal structure list, demonstrating that by predicting more than one candidate for a new system the utility of the maps can be enhanced. We assess the accuracy of the maps by testing predictive accuracy using a cross-validation technique on all AB and A<sub>3</sub>B compounds in the CRYSTMET database. We show that for a new unknown alloy with a stable structure at the stoichiometry of the Pettifor map, a candidate list of five structures will contain the correct crystal structure for the alloy 86% of the time. The algorithms presented here can be used to automate Pettifor maps in materials crystal structure databases, making it possible for users to construct, apply and assess entirely new Pettifor maps quickly and easily.

## 1. Introduction

Predicting the stable crystal structures for materials is an unsolved problem of fundamental importance in materials science. First-principles approaches have made impressive progress [1–3] but are limited by the time it takes to explore the many possible structures for a new system. Historically, the problem of structure prediction has been addressed by

extracting rules from systems for which the experimental data is available and applying these rules to unknown systems. The Hume-Rothery [4] rules for solubility in metal alloys, or the Pauling rules for the structure of oxides [5], are well-known examples of this. A more formal data-mining approach has led to a number of heuristic methods to predict unknown alloy structures [6]. In such data-mining approaches one tries to correlate the stable structure to a small set of relatively easy to determine parameters, such as pseudopotential radii, electronegativities or electron density. If alloys with the same structure cluster together in this parameter space, some predictive capability can be obtained from these correlations. A well known, and remarkably simple, heuristic scheme is the one proposed by Pettifor [7]<sup>3</sup>. To construct a structure map Pettifor assigns a numerical 'Chemical Scale' value to each element, allowing binary alloys  $A_xB_y$  to be mapped onto a Cartesian coordinate system, with the abscissa the Chemical Scale for element A and the ordinate the Chemical Scale for element B. The predictive power of these Pettifor maps derives from the fact that alloys with similar stable structures will cluster together in the map. The unknown stable structure for a new alloy system can be predicted by placing the new alloy on the Pettifor map and examining the stable structures of the nearby neighbours.

In this paper we attempt to quantify the predictive power of Pettifor maps and provide some more details on the optimal methods of constructing and extracting information from them. This includes a description of which data should be used from the databases and making precise the intuitive 'neighbour lookup' procedure for structure predictions.

After this introduction, section 2 will discuss the preparation of the data set. Section 3 will discuss algorithms to formalize the intuitive 'neighbour lookup' procedure of making predictions with Pettifor maps. Section 4 will assess the accuracy of the Pettifor map and section 5 will give a discussion of the results. Finally, section 6 gives a summary of the main points.

## 2. Data preparation

Pettifor maps for AB and  $A_3B$  compositions were constructed from all entries for these compositions in the CRYSTMET database [8]. These compositions were chosen since they have a large number of database entries and therefore should provide good statistics for assessment of the accuracy of Pettifor maps. Also, the AB composition was important in some of the early work establishing the Chemical Scale [9] and is therefore likely to provide something of a best case for its application. Pettifor maps apply at only one composition, so although we will examine both AB and  $A_3B$  systems and combine their statistics, the two compositions are completely independent.

The data set must be cleaned before it can be used. The as-received data set contained a total of 8019 distinct entries, where each entry contains at least some of the following: the constitutive elements, their concentrations in the alloy, the structure type, the space group number and the temperature and pressure. Entries without temperature and pressure were assumed to be at standard temperature and pressure, consistent with the conventions in CRYSTMET. In this paper, we will identify the structure type of an alloy by its structure type name (e.g. NaCl) and its space group number (e.g. 225). Two structure types will be considered equivalent if they share the same structure type name and space group.

All entries missing any of the constitutive elements, the structure type or the space group number were considered incomplete and removed from the data. When there are multiple

<sup>3</sup> To the best of our knowledge, there are no published updated versions of the binary alloy maps originally published by Pettifor. However, maps can be constructed from CRYSTMET and other databases within the MedeA InfoMaticA software package (see <http://www.materialsdesign.com/Pages/InfoMaticA.htm>).

studies on the same compound, where the same elements, structure type and space group number appear in both entries, only one copy was kept in the data. All entries not at standard temperature and pressure were removed from the data. After removing all incomplete, duplicate and non-standard temperature and pressure alloys, the database contained 2570 entries and 420 distinct structure types. These make up what we will call Data Set 1 (DS1). The most common five structure types and the number of times they appear in DS1 (in parentheses) are NaCl [ $Fm\bar{3}m$ ](274), Cu<sub>3</sub>Au [ $Pm\bar{3}m$ ](247), CsCl [ $Pm\bar{3}m$ ](243), CrB [ $Cmcm$ ](117) and Fe<sub>3</sub>C [ $Pnma$ ](98). These structure types make up 38% of the entries in DS1.

In DS1 there are a number of alloys which are identical in composition but do not have the same structure type and space group. For example, PSn has entries for structure types PSn (space group  $P\bar{3}m1$ ), GeAs (space group  $I4mm$ ) and NaCl (space group  $Fm\bar{3}m$ ). These cannot all be the true stable crystal structures for PSn at standard temperature and pressure. It is quite likely that, by careful examination of the original papers, the ambiguity of many of these multiple structure alloys could be resolved. However, this would be a very large undertaking, and therefore has not been done here. In general, if automated construction of Pettifor maps is to be possible, then an extensive literature search cannot be done each time the map is generated. These ambiguous data entries pose a clear problem for the Pettifor maps. If a new alloy has a neighbour with multiple different structures in the database, it is unclear how to use that neighbour to make a prediction. Since these multiple structure alloys are clearly not physical and create errors in a Pettifor map, we explore the effects of removing these data points. There are a total of 1041 entries with multiple crystal structures in DS1. When these are removed we are left with Data Set 2 (DS2), which contains 1540 entries and 205 distinct structure types. The most common five structure types and the number of times they appear in DS2 (in parentheses) are NaCl [ $Fm\bar{3}m$ ](228), Cu<sub>3</sub>Au [ $Pm\bar{3}m$ ](191), CsCl [ $Pm\bar{3}m$ ](182), Fe<sub>3</sub>C [ $Pnma$ ](93) and CrB [ $Cmcm$ ](80). These structure types make up 50% of the entries in DS2.

### 3. Formalizing predictions with Pettifor maps

The intuitive ‘neighbour lookup’ procedure used in Pettifor maps is to place a new alloy on the map and assume its structure type will be that of its nearest neighbours. This procedure does not specify how to deal with cases where the nearest neighbours have more than one structure type or whether more than just the nearest neighbours should be examined. To formalize the ‘neighbour lookup’ approach, we propose that the Pettifor maps should be used to predict not just a single structure, but a ranked list of candidate structures. Knowing that the true structure for an alloy is on a relatively short list of candidates is very useful, since it greatly reduces the space of possibilities one might have to consider in an experimental or first-principles computational investigation. The hope is that the correct structures will generally appear near the beginning of the candidate structure lists.

The candidate structure list for a new alloy with unknown structure is constructed by starting at the coordinates of the new alloy in the Pettifor map and collecting its closest neighbour alloys in order of increasing distance. Distance is measured by a usual Euclidean metric, so that for alloys with coordinates  $(a, b)$  and  $(c, d)$  the distance  $d$  is  $d = \sqrt{(c - a)^2 + (d - b)^2}$ . Each time a new structure is encountered it is added to the end of the list of candidate structures. This will be referred to as the *nearest-neighbour method*.

A variation on the nearest-neighbour method is to consider a fixed size set of nearest neighbours (the *range*) and build the candidate structure list by ordering all the different structure types according to their frequency in this set of neighbours. This will be referred to as the *dominant-neighbour method*.

For the same set of neighbours for a new alloy, the nearest-neighbour and dominant-neighbour methods will produce candidate structure lists with the same structures, but usually with a different ordering. The nearest-neighbour method will place structures close to the new alloy near the front of the list, while the dominant-neighbour method will place structures appearing more frequently in the neighbourhood near the front of the list.

As a benchmark for the above approaches we also construct a candidate structure list by simply ordering every structure type in the entire database according to frequency of appearance. This is a very simple and naive approach and does not make any use of the clustering patterns in the Pettifor maps. This naive approach will be called the *most-frequent method*. The dominant-neighbour method will converge to the most-frequent method as the set of neighbours is extended to include the whole database.

Note that for all these methods only structures at the same stoichiometry as the new alloy are used in the candidate structure list.

#### 4. Assessment of Pettifor map accuracy

The effectiveness of each of the methods proposed in section 3 can be assessed by measuring the percentage of the time the correct crystal structure for a new system is predicted within a candidate structure list of a given length. A cross-validation approach is used, where each alloy in the data set is in turn removed from the data and then predicted as if it were an unknown alloy. The accuracy of the prediction is assessed by comparing to the true crystal structure. This is done for every entry in the data set and the results are averaged. The cross-validation approach is important. Assessments of structure prediction methods are often done by drawing boundaries between structure types by hand and then examining the accuracy of the structural separation, *using all of the data*. This suffers from using an informal method to draw boundaries, but more importantly, it gives the accuracy of the method when boundaries are created already knowing all the data; the important question is the predictive accuracy of the method when the boundaries are drawn without knowing the structure of the new system. It is this latter, more correct assessment of the predictive accuracy that is provided by cross-validation.

The percentage of time the true structure for a new alloy is within the first  $L$  candidate structures is plotted for DS1 and each of the three prediction methods in figure 1.

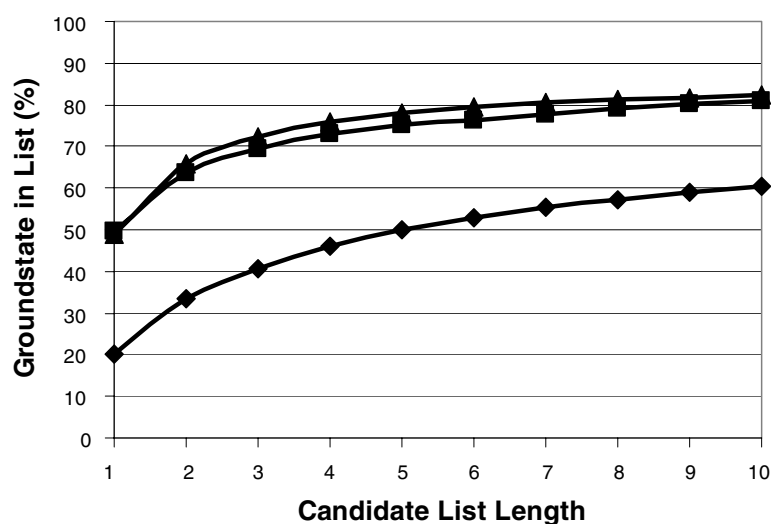
A similar plot is shown for DS2 in figure 2.

Both the nearest-neighbour and dominant-neighbour methods do significantly better than the most-frequent method, demonstrating just how much can be gained by making use of the geometric clustering in the Pettifor maps. It is also clear that the nearest-neighbour method is the most effective approach to prediction, which is gratifying, since it most closely resembles the intuitive use of Pettifor maps.

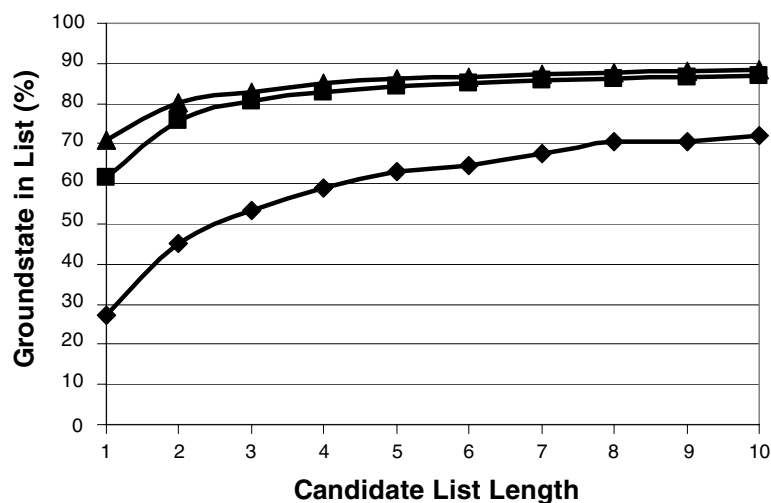
Some of the more technical details associated with the assessment of Pettifor maps are discussed in the appendix.

#### 5. Discussion

The idea of using a candidate structure list, rather than just predicting a single structure, seems to be quite useful. For example, in the nearest-neighbour method applied to DS2, the percentage of time the true structure is in the candidate list increases from 71 to 86% in going from 1 to 5 structures. However, adding an additional five structures to the list increases the percentage accuracy only to 89%, showing that the marginal gain for increasing the length of the candidate structure list is minimal after about five structures. This 'shortlist' of candidate



**Figure 1.** The percentage of times the true structure for a new alloy is within the first  $L$  candidate structures using the most-frequent (diamonds ♦), dominant-neighbour (squares ■) and nearest-neighbour (triangles ▲) methods. Analysis performed with DS1.



**Figure 2.** The percentage of times the true structure for a new alloy is within the first  $L$  candidate structures using the most-frequent (diamonds ♦), dominant-neighbour (squares ■) and nearest-neighbour (triangles ▲) methods. Analysis performed with DS2.

structures can then be used as trial structures in first-principles calculations or in experimental refinements in order to identify the ground state.

The difference between DS1 and DS2 is dramatic. For the nearest-neighbour method, the percent accuracy for the first neighbour increases by 22% in going from DS1 to DS2. This demonstrates the importance of the step of removing compositions with multiple structures when making assessments. However, this cleaning must be approached with some caution, since it is possible that multiple structures are found for alloys that have a greater than chance probability of being in the boundary regions between clusters in the Pettifor map. Preferential

removal of systems in boundary regions biases the data by only keeping the systems more susceptible to clustering, thereby overestimating the effectiveness of the Pettifor map.

There is very little difference between the accuracy of prediction for the AB and A<sub>3</sub>B datasets. For example, when using the nearest-neighbour method with a five-structure candidate list on DS2, the percentage of time the true structure is in the candidate list is 87% for AB and 85% for A<sub>3</sub>B. This shows quantitatively that the Pettifor scheme is transferable to different compositions, although more extensive testing is required before these accuracies can be assumed in general.

There are a number of systems with structures that are unrepeated, i.e. appear nowhere else in the database. These systems cannot be predicted correctly with a Pettifor map, since no neighbour of the uniquely structured alloy will ever share its structure. It is interesting to consider how the unrepeated structures affect the results, since they represent a source of error intrinsic to the data set for any empirical approach that relies on clustering like structures. There are 232 entries with an unrepeated crystal structure in DS1 and 120 in DS2. These unrepeated structures therefore make up 9 and 8% of the entries in DS1 and DS2, respectively, and have a proportional effect. By removing all the unrepeated structures the results for the nearest-neighbour method improve by 8–10% for DS1 and 8–9% for DS2. For five structures this gives a percentage accuracy of 95% for DS2 without the unrepeated structures. Given that the data will contain some errors and the simplicity of the Pettifor maps, this is an extremely impressive result.

Mendeleev numbers [7] are a slightly different way to represent each element other than the Chemical Scale. Mendeleev numbers follow the same ordering of the elements as the Chemical Scale, but map the elements to uniformly spaced integers. Since the Chemical Scale was directly optimized for structural separation, a Pettifor map built from the Chemical Scale is likely to be more accurate than one built from the Mendeleev numbers [7]. This is confirmed by our assessment method, where, with the nearest-neighbour method, we see up to a 9% (5%) improvement when using the Chemical Scale as compared to Mendeleev numbers for DS1 (DS2).

Examining the alloys with incorrectly predicted structures can provide clues for improving the accuracy of the Pettifor maps. For example, consider the prediction errors when using the nearest-neighbour method with a five-structure candidate list on DS2, where we ignore all errors that are associated with structure types that appear only once. It is to be expected that the Pettifor maps will have more difficulty predicting less common structures, since their geometrical regions will be poorly delineated in the map. The errors clearly show that the less common structures are more likely to be predicted incorrectly. For example, 38% of the entries in DS2 with structure types that appear only twice are predicted incorrectly, compared to only 0.4% of entries with the most common NaCl [*Fm* $\bar{3}$ *m*] structure type. By keeping only entries with structure types appearing more than five times, the accuracy of the nearest-neighbour method with a five-structure candidate list increases to 98% on DS2. Significant improvements might be obtained by treating these uncommon structures types by other methods.

The geometric distribution within the Pettifor map of the incorrectly predicted structures may also provide clues to more accurate approaches. For example, the methods used here were all isotropic, in that they considered neighbours in a circle surrounding the alloy being predicted. Other methods, where the search for relevant neighbours is extended anisotropically along optimal directions, may be more effective. Guidance in choosing effective anisotropic methods may be obtained from the distribution of the errors in the isotropic approaches.

The methods employed in this work test only for the ability of a Pettifor map to predict the correct structure for an alloy known to have an ordered phase at the stoichiometry of the map. In other words, the probabilities calculated here are all conditional on first establishing that the alloy has an ordered phase at stoichiometry *s*. For an unknown alloy, if  $p_{str}$  is the probability

of correctly predicting the structure type *conditional* on there being an ordered phase at  $s$ , and  $p_{ord}$  is the probability of correctly predicting if there is any ordered structure at  $s$ , then the probability that the predicted ordered phase actually occurs in the alloy is given by  $p_{ord} * p_{str}$ . Therefore, the usefulness of the Pettifor maps also depends on the ability to accurately predict the existence of some ordered phase, which is a limitation that is rarely acknowledged. In some experimental applications of Pettifor maps it may be obvious that an ordered phase exists, but this is not true in general. There are a number of methods that can help predict the presence of an ordered phase (for example, Miedema's [10] and first-principles methods [1, 3]), and perhaps the Pettifor map can itself be useful, but to our knowledge the accuracy of different approaches for predicting the presence of some ordered phase at a given stoichiometry has not been established quantitatively. The prediction of the presence of an ordered phase remains an open problem in the application of Pettifor maps.

Pettifor maps are a simple method of data mining for structure prediction. It is an important question to ask if a research program for significantly better methods is worthwhile. If we have any data-mining method that can predict only from structures already identified, then unrepeated structures will always be predicted incorrectly. If we exclude these unrepeated structures, the 95% predictive accuracy which can be obtained using five-candidate structures sets a very high bar for any other method. However, approaches that could push that accuracy close to 100%, or that give comparable accuracy with a candidate list of fewer than five structures, would be quite useful. More importantly, the restriction to structures and alloys that have been measured experimentally renders Pettifor maps ineffective for many key problems, since in the more interesting space of multicomponent materials only a small percentage of systems have been studied [6]. However, first-principles methods may provide an efficient method to fill in holes in the databases, and data mining these results is an exciting new area [3].

Perhaps the most important impact of having formalized the process of constructing, using and assessing Pettifor maps is that it allows their application to be completely automated, making it relatively simple to go far beyond the standard Pettifor maps available today. By interfacing with materials structure databases it is possible to rapidly construct, apply and assess entirely new Pettifor maps. These can be restricted to reduced sets of binary data, thereby perhaps increasing accuracy, or can be extended to new binary or multicomponent systems. In addition, other mapping schemes, e.g. using different Chemical Scales or more complex three-dimensional parametrizations [6], can also be implemented. In all these cases, the algorithms discussed here can be used to assess the accuracy of the new Pettifor maps self-consistently, thereby giving immediate feedback on whether new maps are likely to be effective.

## 6. Summary

We have developed a method to automate the construction and testing of Pettifor maps based on data from a standard materials crystal structure database. We describe in detail the cleaning of the database entries, which is an essential step in constructing a useful Pettifor map. We propose a nearest-neighbour method for implementing predictions with the Pettifor map, demonstrating that it is more accurate than a dominant-neighbour approach. In the process we confirm that the clustering properties of the Pettifor map greatly increase the predictive accuracy over a more naive approach based on the most frequently appearing structures. We introduce the idea of generating a candidate structure list for predicting a new structure and show that this adds significant predictive accuracy to the Pettifor maps compared to just predicting a single structure. Using a cross-validation approach we show that the predictive accuracy of Pettifor

maps for the AB and A<sub>3</sub>B alloys is about 86% for a candidate list of five structures. Without unrepeatable structures this predictive accuracy increases to 95%, demonstrating that there is only marginal room for improvement with data-mining methods that are restricted to predicting experimentally known structures. It is now possible to implement the automated construction, use and testing of Pettifor maps in materials databases, giving the opportunity to develop Pettifor maps in new alloy spaces quickly and easily.

### Acknowledgments

We gratefully acknowledge support from the Department of Energy under grant no DE-FG02-96ER 45571 and the Singapore-MIT Alliance.

### Appendix

There are a few important ambiguities with the assessment of Pettifor maps that must be considered. First, there are two choices for how to map an A<sub>n</sub>B<sub>m</sub> alloy onto a point (*x*, *y*) in the map. One can map A → *x*, B → *y*, or A → *y*, B → *x*. It is important to make this choice in a consistent manner, or alloys that should be close to each other might end up being widely separated. For *m* ≠ *n* we always made *x* the minority constituent. For *m* = *n* we always made *x* the constituent with the lowest value on the Chemical Scale.

Another issue is that, in DS1, there are a number of alloy systems identical in composition but with different crystal structures (these are removed in DS2, as discussed in section 2), and it is not clear exactly how to treat these. We chose to treat every entry in the data set as a different alloy. There may be other methods to deal with the problem of multiple structures for a single alloy, but we believe this to be a logical and transparent approach.

Another choice that must be made is the range of the neighbour environment to use for the dominant-neighbour method. A very small range, say only a few nearest neighbours, will capture clustering well, but produce only a very short list of candidate structures, possibly missing the correct one. A large range will produce a large list of candidate structures, but the space sampled may be too far from the new alloy to make good use of the local clustering in the Pettifor maps. The optimal range is dependent somewhat on the needs of the user. Here we chose a range of 20 neighbours, primarily because it gave about ten structures on the candidate structure list, making for an easy comparison with the other methods, and provided fairly representative accuracy. More than 20 neighbours increases the length of the candidate list, but there is little improvement in having a longer list than ten entries, and there is a reduction in the accuracy for the earlier parts of the list. For DS2, a range of ten neighbours gives seven structures in the candidate structure list but improves the percentages by 0–5% (5% for the first candidate structure) over using a range of 20 neighbours, for equivalent length candidate lists. A range of only three neighbours gives a list of only four candidate structures but the accuracy is increased by 2–9% (9% for the first candidate structure) compared to 20 neighbours, for equivalent length candidate lists. However, the total accuracy when the full candidate list is used is 7% higher for a range of 20 as compared to three neighbours, since the candidate structure list is significantly longer. Therefore, as the range increases, there is a trade-off between accuracy for the first few structures and total accuracy over the whole list. However, we have found no range that yields better results than the nearest-neighbour method for either dataset.

A somewhat more complicated problem is that of degeneracy, where two or more different structures are located at the same distance or have the same frequency. When structures are

degenerate the ordering of candidate structure lists is ambiguous. For example, in the dominant-neighbour and most-frequent methods, when multiple structures have the same frequency, it is not clear which should come first in the candidate structure list. Similarly, in the nearest-neighbour method there is ambiguity introduced by the fact that there are often multiple neighbours, with multiple structure types, all at the same distance.

Consider a case where the candidate structure list has  $P$  well-ordered elements, and then there is a set of  $Q$  elements that are all degenerate in their order due to the prediction method being used. If the true structure of the alloy is in those  $Q$  elements then it is not clear how many candidate structures had to be examined in order to predict the correct structure. If we are being very charitable to the prediction method, we might assume that the correct structure can be put first and that our candidate list gave the correct prediction with length  $P + 1$ . At the opposite extreme we might assume that the correct structure should be put last and that our candidate list gave the correct prediction with length  $P + Q$ . In order not to bias the results, the choice was made to choose the length randomly from a uniform distribution of integral values in the interval  $[P + 1, P + Q]$ .

These ambiguities introduced by degeneracy seem to have a very small effect for DS2. For example, using the nearest-neighbour method, the largest change in percentage accuracy was only 3% between using the two extremes of  $P + 1$  and  $P + Q$ . For DS1 the effect was much larger, giving a maximal change in percentage accuracy of 23% between the two extremes. There are methods one could explore to treat this degeneracy. One idea is to combine some of the proposed prediction methods to break degeneracies. For example, if a degeneracy is encountered using the nearest-neighbour method, it might be broken by ordering the degenerate structures by frequency, as is done in the most-frequent method. Because the effect of degeneracy is only significant due to the unphysical entries in DS1 we have not pursued these more involved approaches.

## References

- [1] Springborg M 2000 *Methods of Electronic Structure Calculations: From Molecules to Solids* (New York: Wiley)
- [2] Ceder G 1998 *Science* **280** 1099
- [3] Curtarolo S, Morgan D, Persson K, Rodgers J and Ceder G 2002 submitted
- [4] Hume-Rothery W 1988 *The Structure of Metals and Alloys* (London: Institute of Metals)
- [5] Kingery W D, Bowen H K and Uhlman D R 1976 *Introduction to Ceramics* (New York: Wiley)
- [6] Villars P 1994 *Intermetallic Compounds, Principle and Practice* vol 1, ed J H Westbrook and R L Fleischer (New York: Wiley) ch 11, p 227
- [7] Pettifor D G 1986 *J. Phys. C: Solid State Phys.* **19** 285
- [8] White P S, Rodgers J and Le Page Y 2002 *Acta Crystallogr. B* **58**
- [9] Pettifor D G 1984 *Solid State Commun.* **51** 31
- [10] de Boer F R, Boom R, Matten W C M, Miedema A R and Niessen A K 1988 *Cohesion in Metals: Transition Metal Alloys* (Amsterdam: North-Holland)