Data Mining for Structure type Prediction

by

Kevin Tibbetts

B.S., Engineering Physics (1998)

University of Maine

Submitted to the Department of Materials Science and Engineering
In Partial Fulfillment of the Requirements for the Degree of
Masters of Engineering in Materials Science

At the

Massachusetts Institute of Technology

September 2004

Signature of Author ....
Kevin Tibbetts
Department of Materials Science and Engineering
August 13, 2004

Certified By .... 8/12/04 ....
Gerbrand Ceder
R.P. Simmons Professor of Materials Science and Engineering
Thesis Advisor

Accepted by ....
Carl V. Thompson II
Stavros Salapatas Professor of Materials Science and Engineering
Chair, Departmental Committee on Graduate Students

Data Mining for Structure type Prediction

By

Kevin Tibbetts

Submitted to the Department of Materials Science and Engineering
on August 13, 2004 in Partial Fulfillment of the
Requirements for the Degree of
Masters of Engineering in Materials Science

Abstract

Determining the stable structure types of an alloy is critical to determining many properties of that material. This can be done through experiment or computation. Both methods can be expensive and time consuming. Computational methods require energy calculations of hundreds of structure types. Computation time would be greatly improved if this large number of possible structure types was reduced. A method is discussed here to predict the stable structure types for an alloy based on compiled data. This would include experimentally observed stable structure types and calculated energies of structure types.

In this paper I will describe the state of this technology. This will include an overview of past and current work. Curtarolo et al. showed a factor of three improvement in the number of calculations required to determine a given percentage of the ground state structure types for an alloy system by using correlations among a database of over 6000 calculated energies. I will show correlations among experimentally determined stable structure types appearing in the same alloy system through statistics computed from the Pauling File Inorganic Materials Database Binaries edition. I will compare a method to predict stable structure types based on correlations among pairs of structure types that appear in the same alloy system with a method based simply on the frequency of occurrence of each structure type. I will show a factor of two improvement in the number of calculations required to determine the ground state structure types between these two methods.

This paper will examine the potential market value for a software tool used to predict likely stable structure types. A timeline for introduction of this product and an analysis of the market for such a tool will be included. There is no established market for structure type prediction software, but the market will be similar to that of materials database software and energy calculation software. The potential market is small, but the production and maintenance costs are also small. These small costs, combined with the potential of this tool to improve greatly over time, make this a potentially promising investment. These methods are still in development. The key to the value of this tool lies in the accuracy of the prediction methods developed over the next few years.

Thesis Supervisor: Gerbrand Ceder

Title: R.P. Simmons Professor of Materials Science and Engineering

Acknowledgements

I would first like to thank my advisor, Professor Ceder. In addition to guiding me throughout the project, he has also helped me stay focus and improved my skills as a researcher.

Dr. Dane Morgan and Chris Fischer were very helpful answering any questions I had and filling in the many gaps in my knowledge of the subject.

I would also like to thank Dr John Rodgers of Toth Systems, Inc for taking the time to help me understand the many commercial aspects to consider.

In addition I would like to thank Kathleen Farrell, who always knows the answer, and all my fellow Masters of Engineering students for support and laughter along the way.

1.0 Introduction:

Determining the stable crystal structures that can form in an alloy is critical to determining most properties of the material. We refer to the low temperature stable structures as ground states of the system. The structure type of a material can be determined through experiment, but this is not an affordable way to scan several possible alloys in search of a particular property. The ground state structure types of an alloy can be predicted through computation. This potentially requires *ab initio* energy calculations for thousands of structure types. Someone could spend several years calculating the energies for one alloy.

There are several databases of known ground state structure types. The Pauling File [14] has documented 80,000 structure entries. There are methods to use this experimental data to predict structure types. Pettifor maps have been used for years to predict stable structure types in binary alloys through correlations with known data [1]. Many structure types have computed energies available as well. It has been shown by Curtarolo et al. [2] that there is a correlation among the computed energies of various alloys. A method to use these correlations among experimental data and among computational data to predict ground state structure types would be very useful.

In this paper I will first give a background of this technology. I will explain the process of determining ground state structure types through computation. I will also discuss the results of Curtarolo et al., who presented a study of correlations among a database of 6,000 calculated energies [2].

I have extracted and processed data from the Pauling File Inorganic Materials Database Binaries Edition. I will explain how the data was processed and show some statistics determined from the database. This includes tests to assess the predictive power of the data.

A second goal of this paper is to discuss the commercial potential of this technology. A software tool to predict ground state structure types would be very useful. This tool would predict the most likely stable structure types for a particular alloy. This could be used to predict what alloys are likely to have a desired material property.

There is no known structure type prediction software available. There is a moderate market for energy calculation software and materials database software. The market for this tool would be similar. This paper will discuss the approximate size of this market. The resources necessary to develop this tool will be evaluated and compared to the potential market value. I will give a timeline describing various phases of production of this tool.

2.0 Background:

The key to understanding and predicting the properties of a material is knowledge of the structure type of the material. There are thousands of known crystal structure types. Traditionally the structure type of a material is determined experimentally through x-ray diffraction or a similar technique. The material must be synthesized before this can be done. This is an expensive way to scan several materials for a desired property. There are heuristic models, where experimental observation is used to extract rules that rationalize crystal structures based on a few physical parameters, such as atomic radii and electronegativities. The Miedema rules are one such technique [3].

A widely used structure type prediction method is Pettifor maps [1, 4]. They have been used for many years to predict likely stable structures. These maps predict stable structure types for binary alloys by comparing them to binary alloys with known stable structure types. Pettifor assigned a 'Chemical Scale' value to each element. Known stable structure types are indicated on a two dimensional plot of the chemical indices of the constituent elements. By locating the alloy of interest on this plot it is possible to determine what structure types are stable for alloys with similar indices. Figure 1 shows a Pettifor map for AB binary alloys. Each symbol represents a different structure type. Notice that the appearances of many structure types are clustered on this map. This shows a tendency for alloys of similar elements to have the same structure type.

Pettifor maps are useful, but have limitations. They apply at only one composition. Only data at the composition of interest is used for prediction. This makes it difficult to predict structure types at compositions for which little data is known. Pettifor map methods have been expanded to ternary alloys by Villars [5], but this requires known structure information about ternary alloys to compare. The fraction of ternary alloys with known structure types is much less than the fraction of binary alloys. More binary structures are known and there are fewer possible binary alloys than ternary alloys. It would be useful to use binary alloy information to predict stable structure types for ternary alloys, and to use structure type information at one composition to predict stable structure types at another composition.

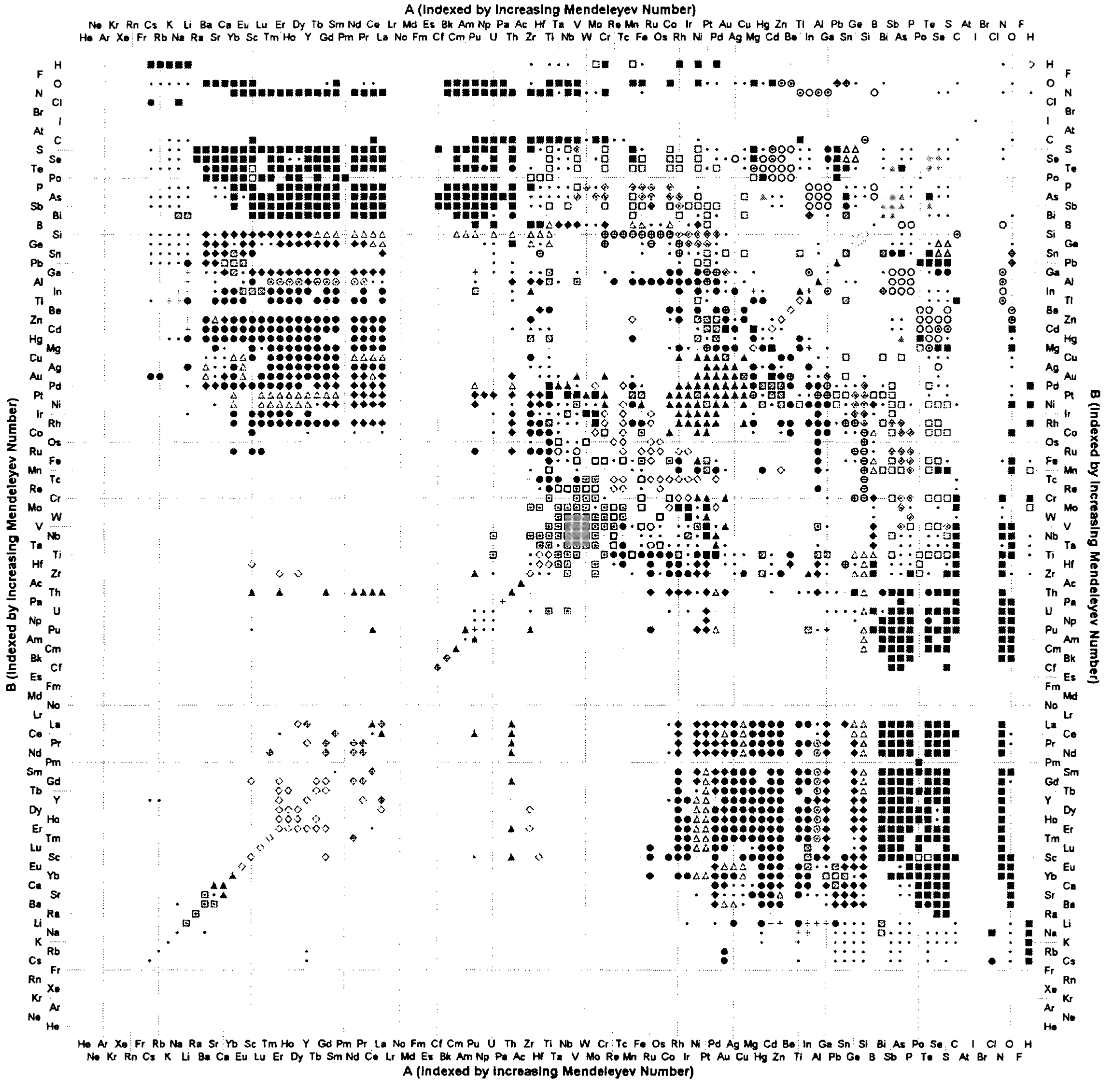

Figure 1: AB Binary alloy Pettifor Map

There is growing interest in using computational methods to predict structure types of materials [6-7]. First principles approaches have made impressive progress, but are limited by the time it takes to explore the many possible structures for a new system. This becomes easier as computing speeds increase, but it is still a time consuming process. The energy for all known structure types must be computed for the alloy of interest. From these energies a convex hull is created for this alloy. The convex hull is the set of stable structures, or combination of structures, as a function of composition that has lower energy than any other structure. Figure 2 shows the convex hull for the AgAu alloy system [8], based on 173 calculated structure types. The energy points on the hull are the ground state structure types. The composition of the alloy and this convex hull are used to determine the structure type or types that will be stable. The energy calculation for one structure type will take somewhere between several hours and days. This means it would take on the order of several months to a year to determine the convex hull and the stable structure types for one alloy, assuming one has to explore 170+ structures. If a method could predict the most likely stable structure types the computation time would be greatly decreased.

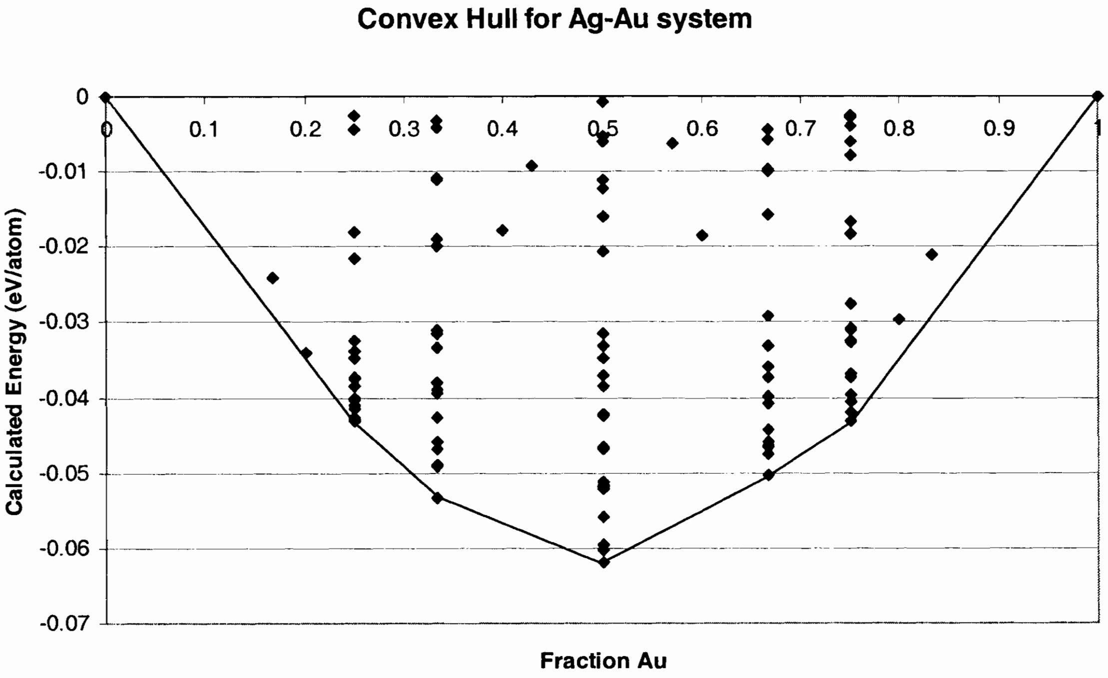

Figure 2: Convex Hull for Ag-Au alloy system

3.0 Data Mining:

Methods are being developed to use data mining to predict stable structure types. The goal is to reduce the number of calculations required to produce the convex hull by predicting the most likely stable structure types. Curtarolo et al. [2] have shown that there are correlations among the *ab initio* energies of different structure types and these correlations can be used to predict the most likely stable structure types.

There is a growing amount of structure type data available. This includes databases of experimentally determined structure types, as well as databases of calculated energies for various structure types and alloys. I will show in this paper that there are correlations among experimentally known stable structure types. By mining available databases of structure types it is possible to predict likely ground state structure types of unknown alloy systems. A robust method of combining the computational and experimental data to predict structure types would be a very useful tool. Such a tool would continually improve as more data is available.

3.1 Past Work:

Curtarolo et al. studied correlations among a database of energy calculations of 114 structure types for each of 55 binary alloys. The formation energies were calculated using density functional theory in the local density approximation with ultra-soft pseudopotentials. Calculations were at zero temperature and pressure and without zero point motion. The number of k-points used for Brillouin zone integrations was 2000 divided by the number of atoms in the unit cell. The absolute energy of these calculations was converged to better than 10 meV per atom. This study was later expanded to a library of 154 structure types for each of 82 binary alloys [9].

A principal component analysis (PCA) was performed on this database of energies [10]. Consider the 114 structural energies for an alloy as an energy vector. The goal of the PCA was to express this energy vector as an expansion in a basis of reduced dimension, *d*. PCA consists of finding the proper basis set that minimizes the remaining squared error for a given dimension, *d*. This regression was done using a Partial Least Squares method implemented with the SIMPLS algorithm [11-13].

Figure 3 shows the results of this PCA analysis for the larger library of 82 alloys and 154 structure types. This plot shows the root mean squared error, in eV per atom, as a function of the reduced dimension d, the number of principal components used. The dashed line in the plot shows the rms error if the structure energies in the database are randomly permuted. This is what we would see if the structural energies were not correlated. The plot shows that for an acceptable error of 50 meV/atom, less than 20 principal components are required.

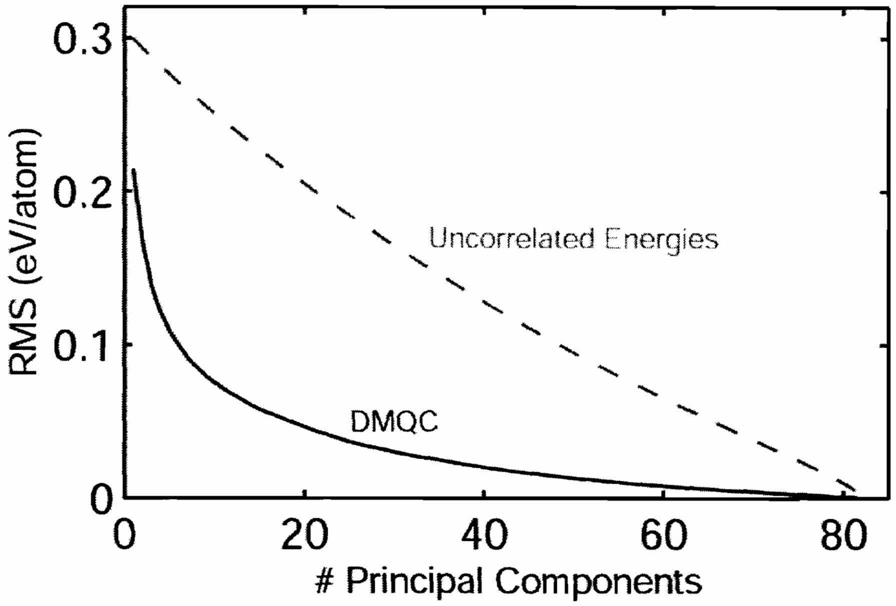

Figure 3: RMS error as a function of number of principal components used for DMQC compared to uncorrelated energies [9]

A test was defined to assess the predictive ability of this data. This is done through an iterative process. This test started with the energies for the bcc, fcc and hcp structures for the pure elements of an alloy system. The energies of the remaining structure types were predicted using the partial least squares regression. The structure type with an energy value the furthest below the convex hull, based on the least squares fit, is computed using *ab initio* methods. At the start of the test this convex hull is a straight line connecting the lowest energy structures of each of the two pure elements. If no energies fall below the convex hull the structure with an energy value nearest to the hull is computed. This energy is then added to the database and the PLS regression is performed again to predict the next candidate structure type. This method is referred to as Data Mining of Quantum Calculations (DMQC).

This method was tested on the database of energies with a leave one out validation method. One alloy was left out of the library. The partial least squares regression was performed using the remaining alloys. From this regression the structure type with the predicted energy farthest below the convex hull for the new alloy system is determined. The *ab initio* energy for this structure type is added to the list of energies for the new alloy system and the process is repeated. This method was repeated for several alloy systems.

Figure 4 shows the number of calculations required as a function of the percentage of ground states predicted correctly using this DMQC method, and the larger library of 82 alloy systems and 154 structure types. The dashed line shows the number of calculations required using random structure selection to choose each candidate structure type. This shows a great improvement in the calculation time required through the DMQC method compared to random structure selection. One hundred percent of the ground state structures were predicted with 80 calculations, much less than the 154 calculations required with random structure selection. Overall this method shows approximately a factor of three improvement in the calculations required for a given degree of accuracy.

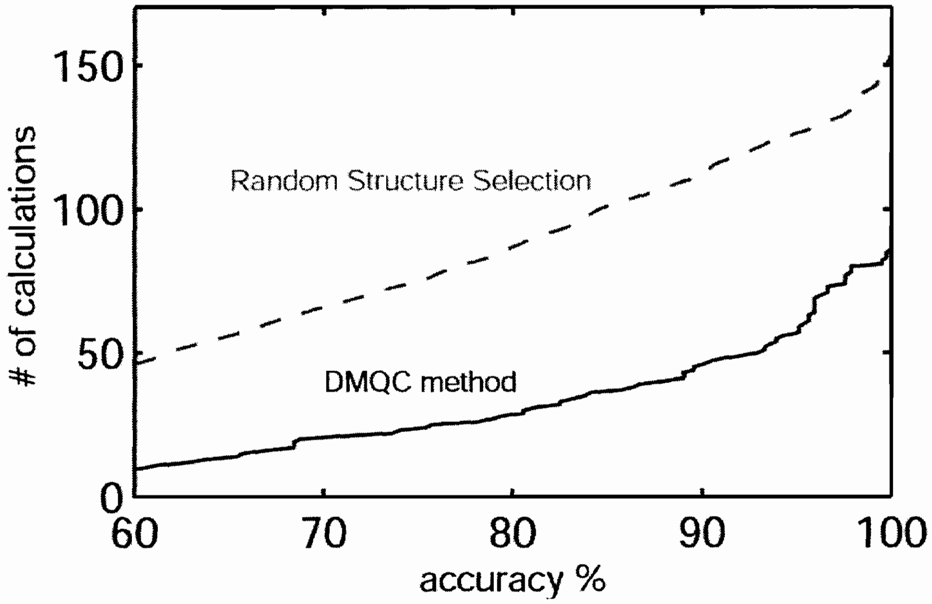

Figure 4: Number of calculations required as a function of the percentage of ground states predicted correctly for DMQC method compared to Random Structure Selection [9]

This method was also able to predict whether the alloy system left out was compound forming. This was done using the smaller library of 55 alloy systems and 114 structure types. It was determined if an alloy system was compound forming with 13 calculations using DMQC, compared to 98 calculations required with random structure selection.

3.2 Current Work:

The work by Curtarolo et al. showed approximately a factor of three improvement between random structure selection and the DMQC method in the amount of computation required for a given accuracy. It would be useful to use experimental structure type data in addition to the *ab initio* energies to predict likely stable structure types. This way it would be possible to predict structure types from experimental data that are not in the library of structures used in the calculated energy database. In order to assess the feasibility of this it is necessary to investigate the correlations among experimentally observed structure types.

I have used the Pauling File Inorganic Materials Database Binaries edition for this study. This contains 27,395 structure type entries. The Pauling File is compiled from 150,000 original publications taken from over 1,000 scientific journals since 1900 [14]. I have extracted and processed the data from the Pauling File to determine correlations among the structure type entries of various alloy systems.

3.2.1 Preparing the Data:

In order to obtain the most accurate statistics from the database, we must determine what data should be included. Some of the entries are for high temperature or high pressure phases. These must be studied separately. We will only consider standard temperature and pressure entries. There are many systems that have been extensively studied. As a result there are many duplicate entries. There are also entries that are similar enough to be considered duplicate entries. For example, there is an entry for CuTi and $Cu_{0.96}Ti_{1.04}$, both with the same structure type. Entries such as this must be removed along with the duplicate entries.

In order to determine if two entries are duplicates we want to determine which compositions are valid compositions for any given structure. This was done through an iterative process for each structure type. Initially all compositions, binned to the nearest 1%, that contained at least 5% of the entries for that structure type were considered valid. All other entries were moved to the nearest composition. The cutoff value was then increased from 5% to 30% in 1% intervals. After each increase, a new set of valid compositions was determined for the new cutoff value and all other entries were binned to the nearest valid composition. A set of 29 compositions with rational fractions were determined based on the distribution of data in the Pauling file. Any of the valid structure compositions that did not occur at one of these rational fractions was moved to the nearest rational fraction for this study. Table 1 shows these composition values and the number of structure types at each composition. I only show the 15 values up to 0.5. The numbers have been symmetrized so the number of structure types at 0.1 is actually the number of structure types at 0.1, 0.9, or both.

This set of rational fractions is included to accommodate the tests that I will discuss later. We want to be able to consider the structure types that could be seen at a certain composition. In order to do this the set of compositions must be defined.

Composition	Fraction	# Structure Types
0	0	25
0.1	1/10	20
0.143	1/7	27
0.167	1/6	31
0.2	1/5	35
0.222	2/9	24
0.25	1/4	66
0.286	2/7	33
0.3	3/10	28
0.333	1/3	85
0.375	3/8	49
0.4	2/5	60
0.429	3/7	21
0.444	4/9	26
0.5	1/2	65

Table 1: Number of Structure types at each allowed composition

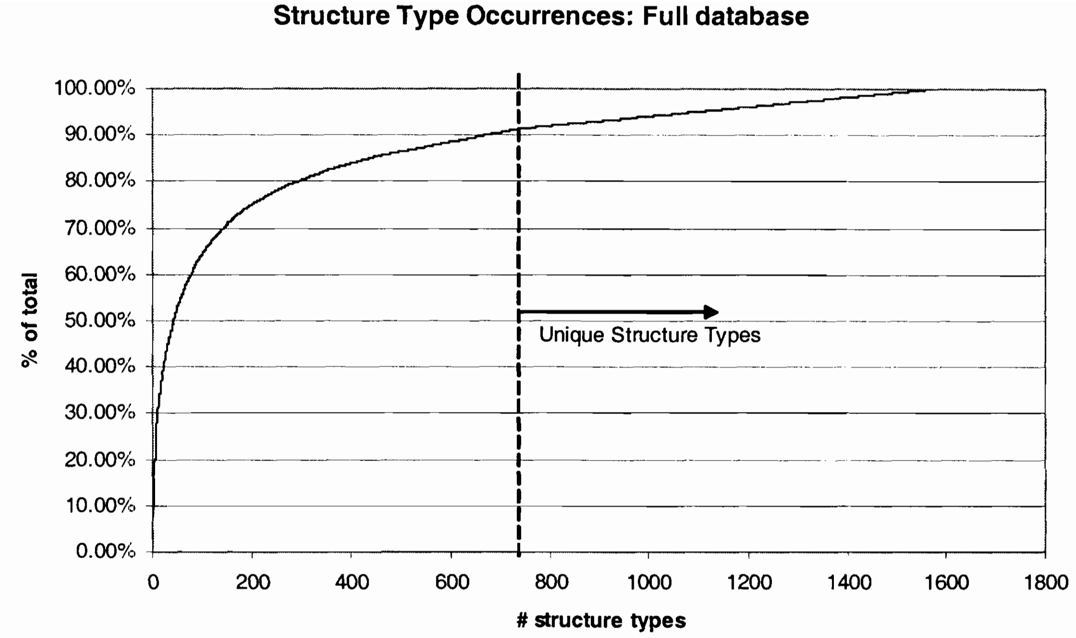

Figure 5: Percentage of total entries in Pauling File as a function of the number of most frequent structure types included

There were 3,436 entries for high temperature phases and high pressure phases. 14,316 of the entries were duplicate entries. After removing these unwanted entries, 9,643 entries are left. From these entries there are 1520 different structure types. It is interesting to look at the distribution of these structure types. I ordered the structure types by the frequency of occurrence. Figure 5 shows the percentage of entries included by the top *n* structure types from this list. The 200 most frequent structure types contain 75% of the total entries. 761 of the structure types only appear once.

For further analysis of the data, a structure type is defined by the prototype name and the composition at which it is being considered. It is important to have a consistent method for defining the composition. For any binary alloy system, AB, I have considered the first element alphabetically as A. So the compound Au_3Cu has a composition of 0.75 while $AuCu_3$ has a composition of 0.25. This is required to determine the proper correlations among structure types. I will discuss in the next section how the data was symmetrized. This was done so that when a structure type occurs at 0.25 and 0.75, for example, both compositions are considered together.

3.2.2 Statistics:

We would like to investigate the correlations among structure types seen in the same alloy system, at different compositions. We know from previous work with structure maps that there are correlations among structure types seen at one composition in similar alloy systems. It is essential to the approach discussed here that experimental data shows a tendency for certain structure types to appear in the same alloy system.

I am making comparisons of pairs of structure types appearing together in the same system. When a structure type is seen at two symmetric compositions, such as 0.25 and 0.75, both entries are instances of the same structure type, but they must be considered differently for this comparison. Consider the compositions and structure types shown in table 2. Structure types α_1 and α_2 are the same structure type appearing at symmetric compositions, as are β_1 and β_2. The appearance of α_1 and β_1 in the same alloy system is equivalent to the appearance of α_2 and β_2 in the same alloy system, as the difference is only in how the composition variable is defined. On the other hand, α_1 and β_1 appearing in the same system is not equivalent to α_1 and β_2 appearing together in the same system. This is handled in this study by considering symmetric structure types separately when counting statistics, but combining the results later.

Composition	0.25	0.33	0.67	0.75
Structure Type	α_1	β_1	α_2	β_2

Table 2: Example of structure type pair correlations

Index of Terms:

N_{sys} : Number of systems in the data set

N_{c_i} : Number of systems with a structure at composition c_i

$N_{\alpha(c_i)}$: Number of systems where structure α appears at composition c_i

$N_{c_i c_j}$: Number of systems with a structure at composition c_i and composition c_j

$N_{\alpha(c_i)\beta(c_j)}$:Number of systems with structure α at composition c_i and structure β at composition c_j

In order to see the correlations among the data I have defined two enhancement factors, $\tilde{p}_{\alpha(c_i)\beta(c_j)}$, and $\tilde{p}^{*}{}_{\alpha(c_i)\beta(c_j)}$ that show correlations among pairs of structure types occurring in the same alloy system. The *pair cumulant* is the cumulant for a pair of structure types. This is defined as

$$\tilde{p}_{\alpha(c_i)\beta(c_j)} = \frac{p(\alpha(c_i)\beta(c_j))}{p(\alpha(c_i)) * p(\beta(c_j))} = \frac{N_{\alpha(c_i)\beta(c_j)} \big/ N_{sys}}{N_{\alpha(c_i)} \big/ N_{sys} * N_{\beta(c_j)} \big/ N_{sys}} = \frac{N_{\alpha(c_i)\beta(c_j)}}{N_{\alpha(c_i)} * N_{\beta(c_j)}} * N_{sys}$$

The *pair cumulants* I will show in this paper are the average of the *pair cumulant* and the *pair cumulant* when both structure types are at the symmetric composition. If there is no correlation between the occurrences of the two structure types the value of this factor will be one. Values larger than one indicate positive correlation; values less than one indicate negative correlation. The largest possible value for this factor will occur when $N_{\alpha(c_i)\beta(c_j)}$ is equal to the smaller of $N_{\alpha(c_i)}$ and $N_{\beta(c_j)}$. This value will be N_{sys} divided by the larger of $N_{\alpha(c_i)}$ and $N_{\beta(c_j)}$.

The *composition restricted cumulant*, $\tilde{p}_{\alpha(c_i)\beta(c_j)|c_ic_j}$, is similar to the *pair cumulant*, with the condition that we know there is a structure at both compositions, c_i and c_j. This is defined as

$$\tilde{p}_{\alpha(c_i)\beta(c_j)|c_ic_j} = \frac{p(\alpha(c_i)\beta(c_j) \mid c_ic_j)}{p(\alpha(c_i) \mid c_ic_j) * p(\beta(c_j) \mid c_ic_j)} = \frac{N_{\alpha(c_i)\beta(c_j)} \big/ N_{c_ic_j}}{\frac{N_{\alpha(c_i)}}{N_{c_i}} * \frac{N_{\beta(c_j)}}{N_{c_j}}} = \frac{N_{\alpha(c_i)\beta(c_j)}}{N_{\alpha(c_i)} * N_{\beta(c_j)}} * \frac{N_{c_i} * N_{c_j}}{N_{c_ic_j}}$$

There is one assumption made in the above equation. I am making the approximation that $p(\alpha(c_i) \mid c_ic_j) \approx p(\alpha(c_i) \mid c_i)$. This is saying the probability of seeing a particular structure type at composition c_i in a system where we know there is a structure type at compositions c_i and c_j, is approximately the same as seeing that structure type at composition c_i in a system where we only know there is a structure type at composition c_i. For some structure types this may not be accurate, especially if c_i and c_j are close in value.

This factor is intended to account for the fact that some compositions are more likely to have a stable structure type than others. Pairs of structure types at compositions that are unlikely to both have stable structure types in the same system will have values for the *composition restricted cumulant* that are larger than the pair *cumulant* for that pair of structure types. If there are two compositions that very seldom appear in the same system, but two structure types from these compositions appear together a lot, this will be reflected by a large value for the *composition restricted cumulant*. As with the *pair cumulant*, the values I will show for this cumulant are an average with the symmetric compositions.

I have also calculated *conditional pair probabilities* for each pair of structure types. This represents the probability of finding structure β at composition c_j given that structure α appears at composition c_i in the same system. This value is averaged with it's symmetric equivalent, with each individual probability weighed by the number of instances of structure type α at that composition.

$$p_{\beta(c_j)|\alpha(c_i)} = \frac{\frac{N_{\alpha(c_i)\beta(c_j)}}{N_{\alpha(c_i)}} * N_{\alpha(c_i)} + \frac{N_{\alpha(1-c_i)\beta(1-c_j)}}{N_{\alpha(1-c_i)}} * N_{\alpha(1-c_i)}}{N_{\alpha(c_i)} + N_{\alpha(1-c_i)}} = \frac{N_{\alpha(c_i)\beta(c_j)} + N_{\alpha(1-c_i)\beta(1-c_j)}}{N_{\alpha(c_i)} + N_{\alpha(1-c_i)}}$$

I have also calculated the probability of no structure type appearing at composition c_j given that structure α appears at composition c_i, the *conditional null structure probability*, $p_{\emptyset(c_j)|\alpha(c_i)}$. The importance of this variable will be made clear later.

I have calculated and analyzed these factors for the set of data that I am calling the metallics. This is all entries not containing the non-metals, He, B, C, N, O, F, Ne, Si, P, S, Cl, Ar, As, Se, Br, Kr, Te, I, Xe, At, or Rn. This subset of the data was chosen because it still contains a majority of the data, 4,836 entries, and the remaining alloy systems are expected to have more similarities than the entire database.

I have sorted the data from this statistical analysis in several different ways in order to see the important correlations and anti correlations. Each list I am including here contains twelve columns. These columns are the name and composition of each structure type, the number of systems each structure type appears in, the number of times the two structure types appear in the same system, the number of systems with a structure type at each composition, the two cumulants, the conditional pair probability, and the conditional null structure probability. The numbers shown all include the statistics for the symmetric equivalent. Each list shows the top 50 results.

For the inter-metallics dataset there are 4,836 entries spread over 29 compositions in each of 1408 systems. The probability that any of these system-composition pairs has no structure type is 88%. This will be the average value of the *conditional null structure probability*. I have listed the *conditional null structure probability*, because it is significant when the *conditional pair probability* and the *conditional null structure probability* sum to 1. In these cases every time a structure was seen at composition c_j in a system with structure α at composition c_i, it was structure β. Situations when the *conditional null structure probability* is 1 can also be significant. This signifies for every system that structure α appeared in at composition c_i, no structure was seen at composition c_j. This is especially significant when $\overline{N}_{\alpha(c_i)}$ and $\overline{N}_{c_i c_j}$ are large.

Highest Enhancement Factor:

The first list I will show here is sorted by the highest *pair cumulant*. Half of the structure types only appear once. Anytime two of these unique structure types appear in the same system they have an extremely large *pair cumulant* due to their low frequency of occurrence. These structures could show accurate correlations, but the large number of unique structure types in the Pauling File causes them to dominate lists such as this. I have not included unique structure types in this list.

The first two entries on this list are cases where every time structure type α appeared at composition c_i, either structure type β appeared at composition c_j or no structure appeared at composition c_j. Correlations such as this are of note when considering the predictive power of the data. Knowing that every time structure α appears at composition c_i, either structure β appears or no structure is found greatly reduces the number of energy calculations required to construct the complex hull. It is possible to understand the reason for some of these correlations by examining the pairs of structure types more closely. Both entries Rb_5Hg_{19} and Rb_3Hg_{20} only appear in Mercury systems. Relations such as this could be useful in predicting candidate structure types.

Structure type α	c_i	Structure type β	c_j	$N_{\alpha(ci)}$	$N_{\beta(cj)}$	$N_{\alpha(ci)\beta(cj)}$	N_{cicj}	***Pair cumulant***	***Composition restricted cumulant***	***Conditional pair probability***	***Conditional null structure probability***
Rb5Hg19	0.2	Rb3Hg20	0.143	2	2	2	34	1338	350.07	1	0
Ti5Te4	0.444	OsGe2	0.667	3	3	2	87	836.25	539.2	0.667	0.333
Ni3P	0.25	TiAs2	0.667	5	3	2	287	752.62	610.2	0.4	0.2
Cu6Ce	0.143	Cu6La	0.143	5	4	2	96	724.75	19.12	0.4	0
Ca2Cu	0.333	CaCu	0.5	2	2	1	524	669	589.15	0.5	0
Li9Ge4	0.3	Li7Ge2	0.222	2	2	1	7	669	435	0.5	0.5
LiGe	0.5	Li7Ge2	0.222	4	2	2	37	669	524.06	0.5	0.5
Mg2Ga	0.333	MgGa	0.5	2	2	1	524	669	566.38	0.5	0.5
Pt3Ga	0.25	Pt2Ga	0.333	2	2	2	115	669	622.72	1	0
CeH3	0.25	CeH2.1	0.3	2	2	2	23	669	424.91	1	0
Li2Ga	0.333	Li5Ga4	0.444	2	2	2	23	669	591.3	1	0
Ba7Cd31	0.833	K2Hg7	0.778	2	2	2	26	669	102.13	1	0
DyGe3	0.714	YGe1.82	0.667	2	2	2	14	669	1001.57	1	0
DyGe3	0.25	DyGe1.85	0.333	2	2	2	115	669	622.72	1	0
Er3Ge4	0.429	DyGe1.85	0.333	2	2	2	23	669	563.59	1	0
Er3Ge4	0.429	DyGe3	0.25	2	2	2	30	669	342.62	1	0
KGe	0.5	Cs4Ge9	0.3	6	2	2	68	446	321.06	0.333	0.667
Li3Al2	0.4	Li5Ga4	0.444	3	2	2	10	446	247.47	0.667	0.333
Li3Al2	0.4	Li2Ga	0.333	3	2	2	43	446	382.17	0.667	0.333
CuAl	0.5	Au4Al	0.2	2	2	1	36	334.5	275.42	0.5	0.5
Fe6Ge5	0.556	Fe3Ga4	0.429	2	2	1	7	334.5	215.68	0.5	0.5
Ir3Si	0.25	Fe2P	0.333	2	2	1	115	334.5	311.36	0.5	0
KHg	0.5	K2Hg7	0.778	2	2	1	26	334.5	298.72	0.5	0.5
LiGe	0.5	Li5Sn2	0.714	2	2	1	34	334.5	277.04	0.5	0.5
LiGe	0.5	Li9Ge4	0.3	4	2	1	68	334.5	239.61	0.25	0.5
NbPt3	0.75	Au2V	0.667	2	2	1	172	334.5	257.44	0.5	0
Pt2Ga	0.333	Ir3Si	0.25	2	2	1	115	334.5	311.36	0.5	0
Pt3Ga	0.25	Ir3Si	0.25	2	2	1	337	334.5	84.25	0.5	0
Rb3Hg20	0.143	KHg	0.5	2	4	1	69	334.5	320.78	0.5	0.5
Rb5Hg19	0.2	KHg	0.5	2	4	1	120	334.5	275.42	0.5	0.5
TaH0.5	0.444	Ta2H	0.333	2	2	1	23	334.5	295.65	0.5	0.5
Tl2Pt3	0.4	CuAl	0.5	2	4	1	133	334.5	269.99	0.5	0
V2H	0.333	NbH0.95	0.429	2	2	1	23	334.5	281.79	0.5	0.5
VAl10	0.9	V7Al45	0.857	2	2	1	13	334.5	175.67	0.5	0.5
IrGe4	0.2	Co5Ge7	0.429	2	2	1	2	334.5	457.5	0.5	0.5
La2Ni3	0.6	Ce24Co11	0.7	2	2	1	16	334.5	57.66	0.5	0.5
OsGe2	0.667	Ru2Ge3	0.6	2	2	1	27	334.5	410	0.5	0.5
Pt8Al21	0.714	Pt2Ga	0.333	2	2	1	31	334.5	195.36	0.5	0
Pt8Al21	0.714	Pt3Ga	0.25	2	2	1	16	334.5	300.14	0.5	0.5
Pt8Al21	0.286	PdAl	0.5	3	2	1	59	334.5	277.04	0.333	0
Zn6.3Sb4.7	0.556	CdSb	0.5	2	2	1	66	334.5	247.88	0.5	0.5

Ba7Cd31	0.167	Rb3Hg20	0.143	4	2	1	18	334.5	118.65	0.25	0.75
Ba7Cd31	0.167	Rb5Hg19	0.2	4	2	1	41	334.5	158.2	0.25	0.5
Cr9.5Al16	0.625	V7Al45	0.857	2	2	1	11	334.5	197.59	0.5	0
Cr9.5Al16	0.571	Cr9.5Al16	0.625	2	2	1	1	334.5	276	0.5	0.5
Li22Pb5	0.833	Li5Sn2	0.714	4	2	2	9	334.5	178.92	0.5	0.25
Li22Pb5	0.833	LiGe	0.5	4	2	2	77	334.5	242.51	0.5	0
LiRh	0.5	LiIr3	0.25	8	2	2	423	334.5	272.44	0.25	0.75
Mg3In	0.25	PuGa	0.5	4	2	1	423	334.5	269.19	0.25	0
RbGa3	0.75	Rb2In3	0.6	2	2	1	41	334.5	197.56	0.5	0.5

Table 3: Structure type pairs sorted by Enhancement Factor 1

Highest number of occurrences together:

This next list is sorted by the number of times the structures appear in the same system. This shows correlations among frequently occurring structure types. The beginning of this list is dominated by the elemental structures, because they appear in many more systems, but there are other structure types on the list. The first pair of two non elemental structure types is the 16th entry on the list, Fe3C and MgCu2. The pair cumulant for this pair is 8.95. This is a strong correlation for two common structure types. There are 20 pairs of structure types that appeared in more than 50 systems together.

Structure type α	c_i	**Structure type β**	c_j	$N_{\alpha(ci)}$	$N_{\beta(cj)}$	$N_{\alpha(ci)\beta(cj)}$	N_{cicj}	***Pair cumulant***	***Composition restricted cumulant***	***Conditional pair probability***	***Conditional null structure probability***
Mg	0	Mg	1	892	892	214	2482	0.76	0.76	0.24	0.035
CsCl	0.5	Mg	0	448	892	197	1301	1.32	1.29	0.44	0
Cu	0	Mg	1	805	892	186	2482	0.68	0.68	0.231	0.017
Mg	0	W	1	892	542	131	2482	0.72	0.73	0.147	0.035
Cu	0	W	1	805	542	127	2482	0.79	0.79	0.158	0.017
Cu	0	Cu3Au	0.25	805	241	125	671	1.78	1.78	0.155	0.47
Cu	0	MgCu2	0.333	805	198	112	862	2.09	2.14	0.139	0.412
Cu	0	CsCl	0.5	805	448	93	1301	0.66	0.65	0.116	0.317
Cu	0	Cu	1	805	805	88	2482	0.38	0.38	0.109	0.017
Mg	0	MgZn2	0.333	892	121	85	862	2.18	2.23	0.095	0.589
MgCu2	0.333	Mg	1	198	892	73	906	1.06	1.03	0.369	0.02
CsCl	0.5	W	0	448	542	65	1301	0.71	0.7	0.145	0
Cu3Au	0.25	Mg	1	241	892	65	688	0.8	0.78	0.27	0.008
Cu	0	CaCu5	0.167	805	83	57	177	2.42	2.34	0.071	0.637

Mg	0	MgZn2	0.667	892	121	55	906	1.36	1.33	0.062	0.572
MgCu2	0.333	Fe3C	0.75	198	87	54	287	8.95	7.38	0.273	0.601
Cu	0	TlI	0.5	805	174	54	1301	1.07	1.05	0.067	0.317
Mg	0	Mn5Si3	0.375	892	118	53	312	1.28	1.23	0.059	0.751
Cu3Au	0.25	CsCl	0.5	241	448	52	423	1.29	1.05	0.216	0.32
Cu3Au	0.25	Cu3Au	0.75	241	241	50	232	2.3	1.8	0.207	0.614
MgCu2	0.333	W	1	198	542	49	906	1.25	1.22	0.247	0.02
Cu	0	Fe3C	0.75	805	87	47	688	1.89	1.86	0.058	0.547
MgZn2	0.333	W	1	121	542	46	906	1.93	1.88	0.38	0.008
Cu3Au	0.25	W	1	241	542	46	688	0.94	0.92	0.191	0.008
Cu3Au	0.25	Mn5Si3	0.625	241	118	45	137	4.4	2.56	0.187	0.656
NaCl	0.5	As	0	142	111	44	1301	7.39	7.23	0.31	0.106
W	0	W	1	542	542	42	2482	0.38	0.38	0.077	0.052
CsCl	0.5	CuTi	0.5	448	76	42	1322	3.3	1.63	0.094	0
Cu	0	Nd	1	805	120	42	2482	1.26	1.26	0.052	0.017
Mn5Si3	0.375	Sm5Ge4	0.444	118	56	41	80	17.31	4.1	0.347	0.492
Cu	0	FeB-b	0.5	805	90	41	1301	1.48	1.44	0.051	0.317
MgCu2	0.333	CaCu5	0.167	198	83	39	128	7.04	3.07	0.197	0.783
Cu	0	Th7Fe3	0.7	805	45	39	97	3.06	3.01	0.048	0.675
Cu	0	Cu3Au	0.75	805	241	39	688	0.53	0.52	0.048	0.547
Mg	0	TlI	0.5	892	174	38	1301	0.64	0.63	0.043	0.476
Mn5Si3	0.375	W	0	118	542	37	312	1.66	1.59	0.314	0
Mg	0	CuTi	0.5	892	76	37	1301	1.45	1.42	0.041	0.476
Cu3Au	0.25	CuAu	0.5	241	86	37	423	4.79	3.89	0.154	0.32
Mg	0	Fe3C	0.25	892	87	35	671	1.23	1.23	0.039	0.691
CsCl	0.5	MgCu2	0.333	448	198	35	524	1.02	0.88	0.078	0.592
Cu	0	Mn5Si3	0.625	805	118	35	312	1.04	1.01	0.043	0.641
Mg	1	NaZn13	0.9	349	43	33	144	2.94	2.8	0.095	0.693
CsCl	0.5	KHg2	0.333	448	63	33	524	3.34	2.87	0.074	0.592
Cu3Au	0.25	Mg	0	241	892	33	671	0.37	0.37	0.137	0
Mg	0	As	1	892	111	32	2482	0.83	0.83	0.036	0.035
Fe3C	0.25	Mn5C2	0.286	87	35	31	52	27.16	13.94	0.356	0.632
MgCu2	0.667	PuNi3	0.75	124	47	31	172	7.12	5.48	0.25	0.476
MgCu2	0.333	Th2Ni17	0.1	198	58	31	96	6.86	4.71	0.157	0.763
Cu	0	CuAu	0.5	805	86	31	1301	1.25	1.23	0.039	0.317
Cu	0	Sm5Ge4	0.556	805	56	31	160	1.94	1.91	0.039	0.67

Table 4: Structure type pairs sorted by occurrences in the same alloy system

Lowest Enhancement Factor

This list is the anti correlations. This shows structures that never appear together. This list is sorted by the minimum of $N_{\alpha(c_i)}$ and $N_{\beta(c_j)}$, in descending order. The first several entries in this list are some of the most frequently occurring structure types. This indicates that the frequency of occurrence of structure types is not enough information to effectively predict stable structure types.

Structure type α	**c_i**	**Structure type β**	**c_j**	**$N_{\alpha(ci)}$**	**$N_{\beta(cj)}$**	**$N_{\alpha(ci)\beta(cj)}$**	**N_{cicj}**	***Pair cumulant***	***Composition restricted cumulant***	***Conditional pair probability***	***Conditional null structure probability***
CsCl	0.5	TlI	0.5	448	174	0	1322	0	0	0	0
Cu3Au	0.25	NaCl	0.5	241	142	0	423	0	0	0	0.32
CsCl	0.5	NaCl	0.5	448	142	0	1322	0	0	0	0
MgCu2	0.333	NaCl	0.5	198	142	0	524	0	0	0	0.551
TlI	0.5	NaCl	0.5	174	142	0	1322	0	0	0	0
MgCu2	0.333	MgZn2	0.667	198	121	0	226	0	0	0	0.798
MgZn2	0.333	MgZn2	0.667	121	121	0	226	0	0	0	0.934
MgZn2	0.333	NaCl	0.5	121	142	0	524	0	0	0	0.769
MgCu2	0.333	Nd	0	198	120	0	862	0	0	0	0
MgZn2	0.333	Nd	0	121	120	0	862	0	0	0	0
W	0	Nd	0	542	120	0	2578	0	0	0	0
Nd	0	Nd	1	120	120	0	2482	0	0	0	0.067
MgCu2	0.333	Mn5Si3	0.375	198	118	0	128	0	0	0	0.985
MgZn2	0.333	Mn5Si3	0.375	121	118	0	128	0	0	0	0.983
Mn5Si3	0.375	Mn5Si3	0.625	118	118	0	94	0	0	0	0.771
Mn5Si3	0.375	W	1	118	542	0	312	0	0	0	0
Mn5Si3	0.375	Nd	1	118	120	0	312	0	0	0	0
Cu	0	As	0	805	111	0	2578	0	0	0	0
Cu3Au	0.25	As	1	241	111	0	688	0	0	0	0.008
MgCu2	0.333	As	1	198	111	0	906	0	0	0	0.02
Mg	0	As	0	892	111	0	2578	0	0	0	0
MgZn2	0.333	As	0	121	111	0	862	0	0	0	0
MgZn2	0.333	As	1	121	111	0	906	0	0	0	0.008
Mn5Si3	0.375	As	0	118	111	0	312	0	0	0	0
W	0	As	0	542	111	0	2578	0	0	0	0
TlI	0.5	As	0	174	111	0	1301	0	0	0	0
Nd	0	As	0	120	111	0	2578	0	0	0	0
As	0	As	1	111	111	0	2482	0	0	0	0.018
MgZn2	0.333	FeB-b	0.5	121	90	0	524	0	0	0	0.769
NaCl	0.5	FeB-b	0.5	142	90	0	1322	0	0	0	0
FeB-b	0.5	As	0	90	111	0	1301	0	0	0	0
Cu3Au	0.25	Fe3C	0.25	241	87	0	697	0	0	0	0
MgCu2	0.333	Fe3C	0.25	198	87	0	287	0	0	0	0.586
MgZn2	0.333	Fe3C	0.25	121	87	0	287	0	0	0	0.959
Mn5Si3	0.375	Fe3C	0.75	118	87	0	137	0	0	0	0.449
W	0	Fe3C	0.75	542	87	0	688	0	0	0	0.613
Fe3C	0.25	Fe3C	0.75	87	87	0	232	0	0	0	0.552

Fe3C	0.25	NaCl	0.5	87	142	0	423	0	0	0	0.517
Fe3C	0.25	Nd	1	87	120	0	688	0	0	0	0
Fe3C	0.25	As	0	87	111	0	671	0	0	0	0
Fe3C	0.25	As	1	87	111	0	688	0	0	0	0
MgCu2	0.333	CuAu	0.5	198	86	0	524	0	0	0	0.551
MgZn2	0.333	CuAu	0.5	121	86	0	524	0	0	0	0.769
TlI	0.5	CuAu	0.5	174	86	0	1322	0	0	0	0
Fe3C	0.25	CuAu	0.5	87	86	0	423	0	0	0	0.517
NaCl	0.5	CuAu	0.5	142	86	0	1322	0	0	0	0
FeB-b	0.5	CuAu	0.5	90	86	0	1322	0	0	0	0
CuAu	0.5	Nd	0	86	120	0	1301	0	0	0	0
Cu3Au	0.25	CaCu5	0.833	241	83	0	36	0	0	0	0.992
MgCu2	0.333	CaCu5	0.833	198	83	0	16	0	0	0	0.98

Table 5: Structure type pairs sorted by lowest Enhancement Factor 1

Sum of conditional Probabilities is 1:

This is a list of structure type pairs for which the sum of the conditional pair probability and the conditional null structure probability is 1. For this list I have only included pairs of structure types from among the 100 most frequently occurring structure types. Consider the ninth pair on this list. There are only 24 systems where a structure appears at 0.833 and 0.3. 20 of these times the two structures are $CaCu_5$ and Th_7Fe_3. Both of these structure types have hexagonal primitive cells.

Structure type α	c_i	**Structure type β**	c_j	$N_{\alpha(ci)}$	$N_{\beta(cj)}$	$N_{\alpha(ci)\beta(cj)}$	N_{cicj}	***Pair cumulant***	***Composition restricted cumulant***	***Conditional pair probability***	***Conditional null structure probability***
Cu3Au	0.25	CaCu5	0.167	241	83	6	59	0.82	0.64	0.025	0.975
Cu3Au	0.25	Th2Ni17	0.9	241	58	4	39	0.73	0.8	0.017	0.983
Sm5Ge4	0.444	CaCu5	0.833	56	83	2	7	1.16	1.7	0.036	0.964
CaCu5	0.167	Sm5Ge4	0.556	83	56	2	7	1.16	1.7	0.024	0.976
PbCl2	0.333	Sm5Ge4	0.444	48	56	10	64	10.02	8.87	0.208	0.792
PbCl2	0.333	CaCu5	0.833	48	83	1	16	0.51	1.67	0.021	0.979
Fe3C	0.25	PbCl2	0.333	87	48	9	287	6.57	5.3	0.103	0.897
FeB-b	0.5	Th7Fe3	0.3	90	45	7	68	4.61	3.32	0.078	0.922
Th7Fe3	0.3	CaCu5	0.833	45	83	20	24	14.47	3.86	0.444	0.556
Th7Fe3	0.3	Fe3C	0.25	45	87	19	41	13.06	8.13	0.422	0.578
PbCl2	0.333	Th7Fe3	0.3	48	45	3	22	3.61	5.52	0.062	0.938

Th2Ni17	0.1	Th7Fe3	0.7	58	45	1	4	0.97	3.05	0.017	0.983
PuNi3	0.75	CaCu5	0.167	47	38	8	11	5.99	9.38	0.17	0.83
Mn5C2	0.286	PbCl2	0.333	35	48	1	25	1.28	1.63	0.029	0.971
Mn5C2	0.286	Fe3C	0.25	35	87	31	52	27.16	13.94	0.886	0.114
Th7Fe3	0.3	Mn5C2	0.286	45	35	3	15	4.85	0.84	0.067	0.933
PbCl2	0.333	Mn5C2	0.286	48	35	1	25	1.28	1.63	0.021	0.979
Sm5Ge4	0.444	Mn5C2	0.286	56	35	2	7	2.74	2.61	0.036	0.964
Th2Ni17	0.1	Mn5C2	0.714	58	35	1	3	1.24	3.94	0.017	0.983
KHg2	0.333	Mn5C2	0.714	63	35	1	62	0.95	0.5	0.016	0.984
CaCu5	0.167	Mn5C2	0.714	83	35	7	9	6.2	5.02	0.084	0.916
Fe3C	0.75	Ag51Gd14	0.8	42	35	1	50	0.91	0.69	0.024	0.976
MgZn2	0.333	Ag51Gd14	0.8	56	35	1	38	0.68	0.8	0.018	0.982
Pu3Pd5	0.375	Sm5Ge4	0.556	33	56	4	35	5.33	2.95	0.121	0.879
Zn17Th2	0.9	PuNi3	0.75	32	47	16	48	14.23	11.57	0.5	0.5
Zn17Th2	0.9	CaCu5	0.833	32	45	12	39	11.15	3.5	0.375	0.625
CuAl2	0.333	Zn17Th2	0.9	41	32	2	25	2.04	3.76	0.049	0.951
MgZn2	0.333	Zn17Th2	0.9	56	32	1	25	0.75	1.38	0.018	0.982
PuNi3	0.75	Sm5Ge4	0.556	47	30	1	36	0.95	0.7	0.021	0.979
W5Si3	0.375	Mn5C2	0.286	28	35	1	32	2.46	1.11	0.036	0.964
W5Si3	0.375	CaCu5	0.833	28	83	1	13	1.1	1.82	0.036	0.964
CuAl2	0.333	Th2Ni17	0.1	41	28	1	34	1.17	0.52	0.024	0.976
PuNi3	0.75	Th2Ni17	0.1	47	28	6	6	6.1	13.13	0.128	0.872
Th2Ni17	0.1	Zn17Th2	0.9	28	32	10	13	14.93	5.98	0.357	0.643
Th2Ni17	0.1	PuNi3	0.75	28	47	6	6	6.1	13.13	0.214	0.786
Co2Si-b	0.333	Th2Ni17	0.9	27	30	4	25	6.61	12.17	0.148	0.852
Th2Ni17	0.9	Co2Si-b	0.333	30	27	4	25	6.61	12.17	0.133	0.867
CuZr2	0.333	Sm5Ge4	0.444	28	26	1	23	1.84	1.62	0.036	0.964
AlB2	0.667	CaCu5	0.833	25	45	5	74	5.95	3.34	0.2	0.8
Zn58Gd13	0.167	Cr5B3	0.625	25	28	1	13	4.05	6.62	0.04	0.96
Zn58Gd13	0.167	Fe3C	0.75	25	87	1	36	1.35	1.54	0.04	0.96
Zn58Gd13	0.167	CsCl	0.5	25	448	24	127	5.76	3.92	0.96	0.04
NaZn13	0.9	AlB2	0.667	43	25	1	62	1.24	1.07	0.023	0.977
NaZn13	0.9	KHg2	0.333	43	24	6	25	7.78	14.33	0.14	0.86
Cd2Ce	0.667	Mg3Cd	0.25	23	27	4	145	8.62	7.37	0.174	0.826
Cd2Ce	0.667	CsCl	0.5	23	224	22	276	5.71	5.03	0.957	0.043
Zn17Th2	0.9	Th7Fe3	0.3	32	23	3	4	5.45	8.57	0.094	0.906
MoSi2	0.667	Th7Fe3	0.3	34	23	1	33	1.71	1.11	0.029	0.971
PuNi3	0.75	Th7Fe3	0.3	47	23	5	20	6.19	4.83	0.106	0.894
Be5Au	0.833	Th7Fe3	0.7	22	22	1	4	2.76	2.39	0.045	0.955

Table 6: Structure type pairs sorted by the sum of conditional pair probability and conditional null structure probability

Most Frequent Structure types:

This last list shows the most frequently occurring structure types. The list is ordered by the minimum of $\overline{N}_{\alpha(c_i)}$ and $\overline{N}_{\beta(c_j)}$, in descending order. This is interesting to see correlations among the most common structure types, where the statistics are the best.

Structure type α	**c_i**	**Structure type β**	**c_j**	**$N_{\alpha(ci)}$**	**$N_{\beta(cj)}$**	**$N_{\alpha(ci)\beta(cj)}$**	**N_{cicj}**	***Pair cumulant***	***Composition restricted cumulant***	***Conditional pair probability***	***Conditional null structure probability***
Mg	0	Mg	1	892	892	214	2482	0.76	0.76	0.24	0.035
Cu	0	Mg	1	805	892	186	2482	0.68	0.68	0.231	0.017
Cu	0	Mg	0	805	892	15	2578	0.06	0.06	0.019	0
Cu	0	Cu	1	805	805	88	2482	0.38	0.38	0.109	0.017
W	0	W	1	542	542	42	2482	0.38	0.38	0.077	0.052
Mg	0	W	1	892	542	131	2482	0.72	0.73	0.147	0.035
Mg	0	W	0	892	542	11	2578	0.06	0.06	0.012	0
Cu	0	W	1	805	542	127	2482	0.79	0.79	0.158	0.017
Cu	0	W	0	805	542	6	2578	0.03	0.03	0.007	0
CsCl	0.5	W	0	448	542	65	1301	0.71	0.7	0.145	0
CsCl	0.5	Mg	0	448	892	197	1301	1.32	1.29	0.44	0
Cu	0	CsCl	0.5	805	448	93	1301	0.66	0.65	0.116	0.317
Cu3Au	0.25	W	1	241	542	46	688	0.94	0.92	0.191	0.008
Cu3Au	0.25	W	0	241	542	8	671	0.16	0.16	0.033	0
Cu3Au	0.25	Mg	1	241	892	65	688	0.8	0.78	0.27	0.008
Cu3Au	0.25	Mg	0	241	892	33	671	0.37	0.37	0.137	0
Cu3Au	0.25	CsCl	0.5	241	448	52	423	1.29	1.05	0.216	0.32
Cu3Au	0.25	Cu3Au	0.75	241	241	50	232	2.3	1.8	0.207	0.614
Cu	0	Cu3Au	0.75	805	241	39	688	0.53	0.52	0.048	0.547
Cu	0	Cu3Au	0.25	805	241	125	671	1.78	1.78	0.155	0.47
MgCu2	0.333	W	1	198	542	49	906	1.25	1.22	0.247	0.02
MgCu2	0.333	W	0	198	542	27	862	0.64	0.65	0.136	0
MgCu2	0.333	Mg	1	198	892	73	906	1.06	1.03	0.369	0.02
MgCu2	0.333	Mg	0	198	892	25	862	0.4	0.41	0.126	0
MgCu2	0.333	MgCu2	0.667	198	198	2	226	0.15	0.2	0.01	0.798
CsCl	0.5	MgCu2	0.333	448	198	35	524	1.02	0.88	0.078	0.592
Cu3Au	0.25	MgCu2	0.667	241	198	10	287	0.47	0.4	0.041	0.614
Cu3Au	0.25	MgCu2	0.333	241	198	29	287	1.6	1.35	0.12	0.639
Cu	0	MgCu2	0.667	805	198	5	906	0.08	0.08	0.006	0.548
Cu	0	MgCu2	0.333	805	198	112	862	2.09	2.14	0.139	0.412
W	0	TlI	0.5	542	174	17	1301	0.48	0.47	0.031	0.461
Mg	0	TlI	0.5	892	174	38	1301	0.64	0.63	0.043	0.476
MgCu2	0.333	TlI	0.5	198	174	23	524	1.97	1.69	0.116	0.551
CsCl	0.5	TlI	0.5	448	174	0	1322	0	0	0	0
Cu3Au	0.25	TlI	0.5	241	174	29	423	1.84	1.51	0.12	0.32
Cu	0	TlI	0.5	805	174	54	1301	1.07	1.05	0.067	0.317
TlI	0.5	NaCl	0.5	174	142	0	1322	0	0	0	0
W	0	NaCl	0.5	542	142	21	1301	0.73	0.71	0.039	0.461
Mg	0	NaCl	0.5	892	142	16	1301	0.32	0.31	0.018	0.476
MgCu2	0.333	NaCl	0.5	198	142	0	524	0	0	0	0.551
CsCl	0.5	NaCl	0.5	448	142	0	1322	0	0	0	0

Cu3Au	0.25	NaCl	0.5	241	142	0	423	0	0	0	0.32
Cu	0	NaCl	0.5	805	142	10	1301	0.27	0.27	0.012	0.317
MgZn2	0.333	NaCl	0.5	121	142	0	524	0	0	0	0.769
MgZn2	0.333	TlI	0.5	121	174	2	524	0.24	0.21	0.017	0.769
MgZn2	0.333	W	1	121	542	46	906	1.93	1.88	0.38	0.008
MgZn2	0.333	W	0	121	542	14	862	0.57	0.58	0.116	0
MgZn2	0.333	MgZn2	0.667	121	121	0	226	0	0	0	0.934
Mg	0	MgZn2	0.667	892	121	55	906	1.36	1.33	0.062	0.572
Mg	0	MgZn2	0.333	892	121	85	862	2.18	2.23	0.095	0.589

Table 7: Structure type pairs sorted by frequency of occurrence

3.2.3 Prediction Tests:

We want to test the ability to predict stable structure types at one composition based on structure types seen in that system at a different composition. I have defined and run some tests for this purpose. These tests are intended to compare different methods of creating candidate lists from which to choose possible ground state structure types. For these tests I leave out all structure entries from the system I am testing when calculating statistics. Unique structures are ignored as it is impossible to predict them when they are left out. Compositions where no structure type appears in the Pauling File for a given system, null structure entries, are not considered. These cases are very common. As a result, when they are included, the null structure prediction is the correct choice most of the time so the results are similar with every method.

The four methods tested are random structure selection, ordering by frequency of occurrence, ordering by pair probabilities, and ordering by a cumulant expansion. For each method I record the fraction of structure types checked before finding the stable structure type for that system.

With random structure selection structure types are selected at random from a list of all structure types appearing at the composition of interest. On average half of the possible structure types will have to be checked before finding the correct one with this method.

With frequency of occurrence, all structure types seen at the composition of interest are ordered by the number of times that structure type appears at that composition in the database. This is the simplest way of using the experimental data to predict possible structure types.

For the pair probability method I order the possible structure types by the conditional pair probabilities. For each system I try to predict each stable structure type that appears in the system based on the conditional pair probabilities from each of the other structure types that appear in the system. For any system with more than one structure type a series of tests are run for each structure type α in the system. This series of tests consists of looking at each composition other than that of structure type α where a structure type appears. A list of candidate structure types for this composition is compiled based on the conditional pair probabilities from structure type α. If there are n_i structure types that appear in alloy system i. There will be $n_i*(n_i-1)$ tests for that system. Structure types with the same conditional pair probability are ordered by frequency of occurrence.

For the cumulant expansion method I order the possible structure types by the conditional probability based on all other structure types that appear in the system. This is $P(\beta \mid \alpha_1 ... \alpha_n)$, where β is the structure I am trying to predict and $\alpha_1 ... \alpha_n$ are all other structure types that appear in the system. I calculate this using a cumulant approach.

$$P(\beta \mid \alpha_1 ... \alpha_n) = \frac{P(\beta \alpha_1 ... \alpha_n)}{P(\alpha_1 ... \alpha_n)}$$

$$P(\beta \alpha_1 ... \alpha_n) = \prod_{\gamma \subseteq (\beta \alpha_1 ... \alpha_n)} \tilde{p}_\gamma$$

$$P(\beta \mid \alpha_1 ... \alpha_n) = \frac{\prod_{\gamma \subseteq (\beta \alpha_1 ... \alpha_n)} \tilde{p}_\gamma}{\prod_{\gamma \subseteq (\alpha_1 ... \alpha_n)} \tilde{p}_\gamma}$$

The term on the right of the second equation represents the product of the cumulants of all subsets of the set of structure types that appear in a system. I am making the approximation that the cumulant for all subsets with more than two elements is 1, so I only use two event cumulants.

$$\tilde{p}(\beta \alpha) = \frac{p(\beta \alpha)}{p(\beta) p(\alpha)}$$

All subsets not containing β will appear in both the numerator and denominator, canceling out.

$$P(\beta \mid \alpha_1 ... \alpha_n) = p(\beta) \prod_{i=1}^{n} \tilde{p}(\beta \alpha_i) = p(\beta) \prod_{i=1}^{n} \frac{p(\beta \alpha_i)}{p(\beta) p(\alpha_i)} = \left(\frac{1}{p(\beta)} \right)^{n-1} \prod_{i=1}^{n} p(\beta \mid \alpha_i)$$

Using this method to predict the correct structure type, there is one test for each structure type entry, based on all other entries for that system. Structure types with the same conditional probability are ordered by frequency of occurrence.

3.2.4 Results:

I recorded the results of the tests as the fraction of structure types that must be tested with each method before finding the correct structure type. Table 8 shows the average and standard deviation for the four methods. For the conditional pair probability method there were 19,701 tests. One test involves compiling an ordered list of structure types and determining where on this list the known stable structure type for the alloy system and composition of interest appears. For the frequency of occurrence method and cumulant approach there were 4096 tests.

	Average	Standard Deviation
Random Structure Selection	0.5	0.333
Frequency of Occurrence	0.2458	0.2688
Conditional Pair Probability	0.1110	0.1859
Cumulant Approach	0.1726	0.2502

Table 8: Statistical data for structure type prediction using four different methods. Values are in fraction of possible structure types checked

There is a greater than two times improvement in the average fraction of structure types tested before finding the ground state structure type between the frequency of occurrence method and the conditional pair probability method. The average fraction from the cumulant approach is higher than that from the conditional pair probability method. This can be explained by a lack of adequate data for the cumulant method. This method orders structure types by the products of the two point cumulants for all structure types appearing in the system at a composition other than the composition being tested. Many of the two point cumulants are zero because many pairs of structure types never appear together in the same system, or only appear together in the system being studied, which is left out. As a result 92 % of the conditional probabilities determined with this method are zero. This effect is less of a problem with the conditional pair probability method because we are only looking at one conditional pair probability at a time. With the cumulant approach many conditional pair probabilities are multiplied together, increasing the chance of a zero value. Structure types with a cumulant expansion of zero are ordered by frequency of occurrence. The improvement in average fractions, shown in table 8, from 0.2458 to 0.1726 is large considering only 8% of the data is affected.

In addition to the average and standard deviation from each method, it is of interest to see the percentage of correct structure types found with each method for a specific fraction of possible structure types checked. This is shown in figure 6. The cumulant approach is similar to the frequency of occurrence method at low and high percentages. The major deviation is between 60 and 85%. All three methods perform fairly well up to about a 60% chance of finding the correct structure type. The frequency of occurrence method begins to get worse here. The cumulant approach continues to do well until about 70%. The conditional pair probability method holds on until about 80%. The frequency of occurrence method does well at predicting common structures. It is not expected that this method would improve much with more data. The results for this method are a reflection of the fraction of the database that is contained by the most frequently occurring structure types. The other two methods are able to predict some of the less frequent structures well. More data would improve the ability of these methods to predict the less frequent structure types.

I have compared the structure types in the Pauling File with the structure types in the database Curtarolo et al. used in the DMQC work. I was able to match 61 of the structures in the Curtarolo database with structures in the Pauling File. These 61 structure types contain 57% of the entries from the Pauling File, after the duplicates and non standard temperature and pressure phases were removed. The remaining 43 % of the entries represent cases where experimental data can be used to predict structures that are not currently in the database of structures used to create the library of computational energies. This is one of the reasons it is desired to use experimental data in conjunction with computed structural energies to predict candidate structure types.

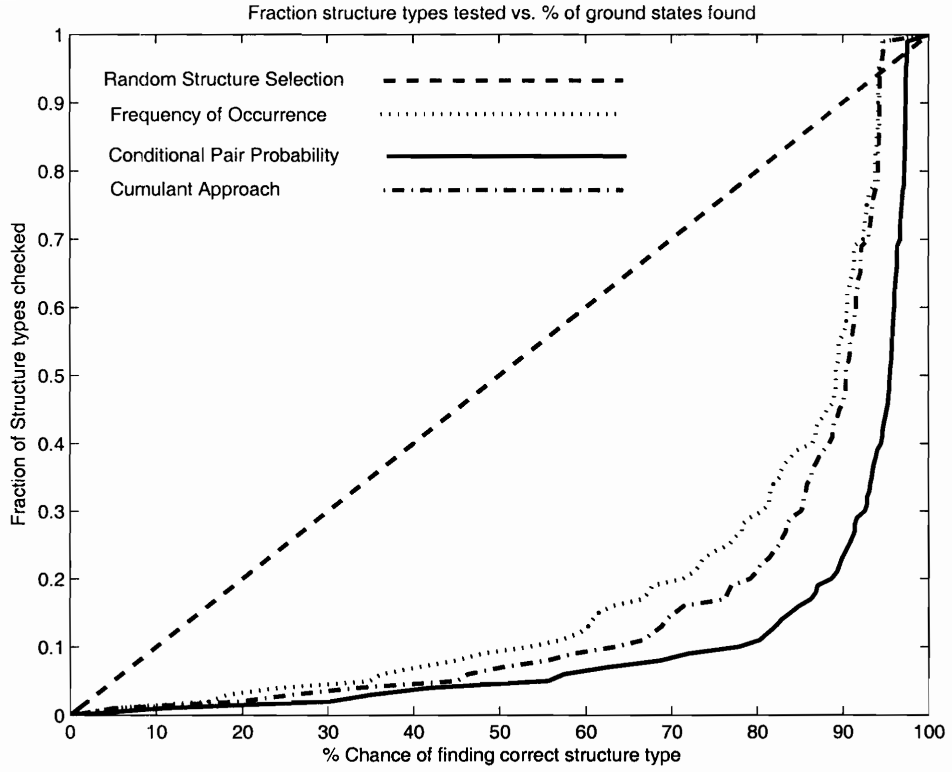

Figure 6: Comparison of fraction of structure types that must be checked to identify a given percentage of stable ground states using the four different methods

3.3 Future Work:

There are several important obstacles to overcome before this technology is ready for commercial use. This paper and the work by Curtarolo et al. have shown there are correlations among both experimental and calculated structure type data. Tests in both cases have shown the ability to use these correlations to predict the most likely stable structure types. The correlations among the computed energies must be combined with the correlations among the experimentally known structure types in a robust predictive algorithm.

The work so far has used only binary alloys. It is important that these methods apply to other systems as well. More testing is required to determine how well these correlations can be used to predict structure types in alloys of more than two elements.

In addition to predicting the stable structure types, it is also important to predict when there is no stable structure type. Curtarolo et al. were able to show great improvement in the calculations required to determine if an alloy system is compound forming using the DMQC method. This is an area that must be researched further.

4.0 Commercial Analysis:

4.1 Potential commercial application:

Part of the goal of this project is to examine commercial applications for this technology. The product considered here is a software tool that could predict stable structure types of an alloy or predict alloys with a given structure type. This tool would be of use to many scientists and engineers. Such a tool would be useful in improving the efficiency of research in both academic and commercial settings. If someone is interested in investigating the properties of a particular alloy they could use this tool to predict the stable structure types. Also if a particular property is desired the tool could predict alloys like to have a structure type with this desired property.

There is work to be done before such a tool would be ready, but previous and current work shows that such a prediction tool is possible. The confidence level of these predictions is the key factor to be determined. This confidence level would improve over time as more data is available. This is an important detail about this technique. Any prediction methods will improve as more structure type data is made available. This is sort of a snowball effect. The tool will improve the efficiency of calculations to determine structure types. The results of these calculations will improve the prediction tool. This allows continued revenue as updates to the tool are required.

4.2 Intellectual property:

When considering the commercial potential of a technology it is important to understand any intellectual property associated with the technology. I have found no patents related to predicting structure types [16]. Professor Ceder's group has made a technology disclosure for this technology. It has yet to be determined if it is worthwhile to apply for a patent for this technology.

There are numerous patents on techniques to correlate data [17-19]. These patents will not affect the commercial potential of the data mining approach to structure type prediction discussed here. The correlation techniques being used are developed in house. The methods are specific to structure type prediction, which is not covered by any patents I have found.

One other intellectual property consideration is database protection. Currently databases only have copyright protection in the United States. This does not sufficiently protect the work done to compile and organize the data. In some cases entire databases can be extracted and reproduced. The Database and Collections of Information Misappropriation act [20] has been in congress for several years. It is currently in the House of Representatives. It has been passed by the House of Representatives before, but the Senate has rejected it. This act is intended to prevent misappropriation of data in this manner. It would allow the producer of a database to prohibit someone from using their database in a product that competes in the same market. An alternative to this act was introduced to congress in March of 2004. This is the Consumer Access to Information Act. This act more narrowly defines misappropriation of a database.

These acts only apply to the United States. The European Union has much more strict regulations in place protecting the creator of a database [21]. The best approach to allow worldwide sales would be to market the product as part of a database package. This would require working with a database producer, such as Pauling File, or ICSD.

4.3 Market:

There are no structure prediction tools available so there is no direct competition for this product or an established market. Much work has been done using structure maps like the Pettifor maps discussed earlier, to predict stable structure types. Some companies use this sort of method internally for structure type prediction. Rational Discovery has published a paper discussing the use of structure maps for structure type prediction [22]. There is no known software on the market for this purpose. There is an established market for energy calculation software and materials databases. The potential market for the product described here would be similar to the market for these products.

The most widely used energy calculation software packages available commercially are Gaussian and VASP [23-24]. Gaussian is a software package used to predict energies, molecular structure and vibrational frequencies through computation. This is the most popular such product for chemists. The exact customer data for Gaussian is protected [25-26], but they have thousands of customers worldwide. VASP is the most common energy calculation software package for crystals. They have 400 academic licenses and 40 industry licenses [27]. There are also several energy computation software packages available for no charge.

The Inorganic Crystal Structure Database (ICSD) and the Pauling File are the most popular crystal structure databases. The exact customer numbers for ICSD are protected, but they are of the same order as the energy calculation software packages, approximately 1000 - 2000 [21]. The Pauling File is new to the market. Their free online database has 1918 registrations as of June 2004, almost double there June 2003 total of 1038 [35]. Ninety percent of the customer base for crystal structure databases is academia [32].

The prices for these types of software package are vastly different for academia and industry [28-30]. The price for energy calculation software for academics ranges from $500 to $5000. The prices for industry vary from $1000 to $30,000. The high end of these ranges is for Gaussian and VASP. The prices for materials database software range from $300 for a single user CD version of ICSD for academia to $3000 for multiple user industry packages [36, 37]. These are well established and developed software packages. Typical add on packages are $100 to $200 for academia, and twice this for industry. The software product I am proposing should be sold as an additional package in conjunction with a database. The exact price would depend on the effectiveness of the software and inflation before the product is ready for market.

The International Technology Research Institute did a study between 1998 and 2000 on the use of computational modeling in industry [31]. This study investigated modeling work at companies in Japan, Europe and the US. Automobile manufacturers are interested in predicting the optimum alloys for various automotive applications. There is also increasing interest in studying hydrogen fuel technology. Semiconductor manufacturers are interested in determining the optimum material for various applications. Pharmaceutical companies and research laboratories also constitute a large number of the companies involved in this study.

I have used this study to compile a list of the number of companies doing modeling research organized by the year they started this research. This is shown in figure 7. The numbers for the last two years are somewhat low because this was the same time the study was taking place so the study is less complete for companies starting modeling in these years. Over the last 10 years approximately five companies have begun research in modeling each year. This trend should continue or increase as new uses for modeling are discovered and computing speeds increase. This shows that the market will increase greatly in the 6-8 years before this product would be ready for sale.

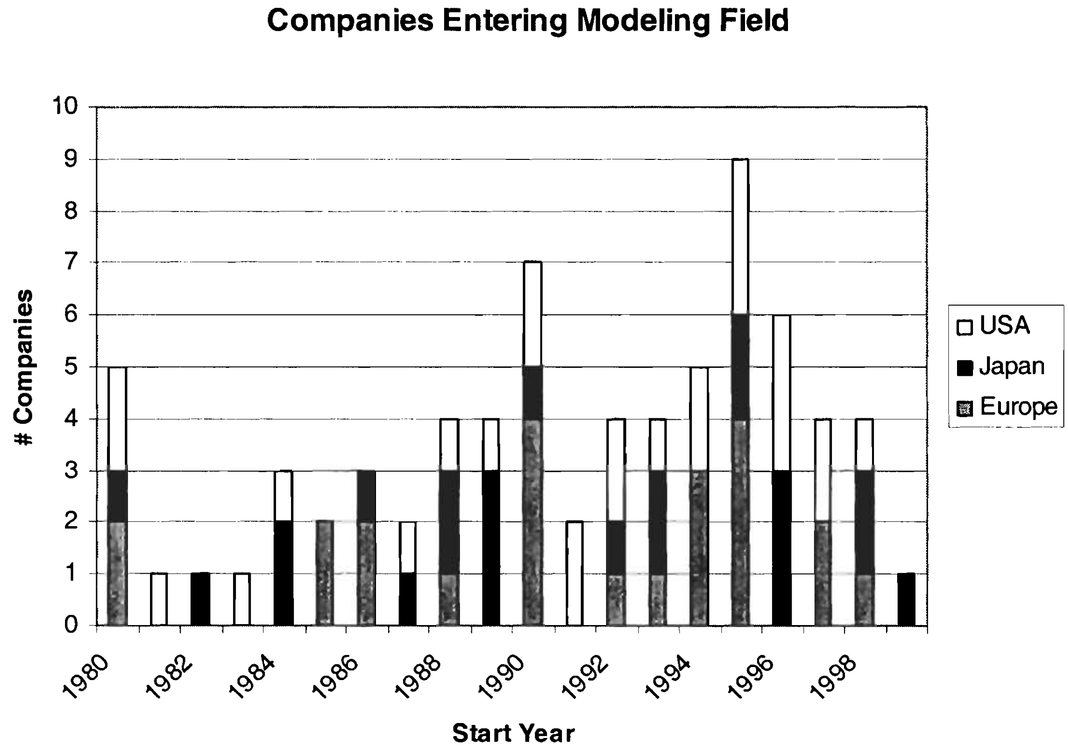

Figure 7: Histogram of companies starting computational modeling work by year started

The potential market for the proposed software tool is on the order of 2,000 customers. Approximately 90% of the customers are academic [27, 32], but the industry numbers appear to be growing. I have constructed a table of the projected annual revenue for different market shares and pricing schemes. I have also shown how these results would change if the industry percentage of the market grew from 10% to 15% of the overall market. Table 4 shows these results. The projected revenue ranges from $88,000 per year to $575,000 per year.

Market Share Pricing	20% Market Share 10% Industry	50% Market Share 10% Industry	20% Market Share 15% Industry	50% Market Share 15% Industry
$200 Academic $400 Industry	$88,000 /year	$220,00 /year	$92,000 /year	$230,000 /year
$500 Academic $1000 Industry	$220,000 /year	$550,000 /year	$230,000 /year	$575,000 /year

Table 4: Projected Potential Revenue for various pricing schemes and market shares

4.4 Business plan:

I have developed a business plan for development of this product and introduction into the market. There are four phases of this plan, preliminary work, development, limited release and commercial release. Estimating software development times is a difficult task, especially at this early stage. A range of from 2 to 6 years can be expected for this software tool [33-34]. This work will be started in phase 1 and completed in phase 2.

We are currently in the preliminary work phase. During this phase we will gain a better understanding of the correlations among the data. In this phase the obstacles discussed in the future work section will need to be overcome. Structures of more than two elements must be investigated and included in the prediction methods. Prediction methods must be developed that combine the computed and experimental data. I estimate this phase will take 2 to 4 years. This is a rough estimate based on the work I have done this year and past work done by Professor Ceder's research group.

During the development phase the software will be developed for internal use based on the methods determined in phase 1. This software will be tested and used to fine tune the prediction methods during this phase. This phase should take approximately

2 years. There is some overlap between phase 1 and phase 2. An early version of the software tool would make testing the methods much easier.

When the software has been thoroughly tested internally, it should be released in a limited fashion for beta testing and to increase interest in the tool. This could be accomplished through a free version released through the database producer. I would expect this phase to last 2 or 3 years. Beta testing can be accomplished in one or two years. If there is enough interest in the software and there has been sufficient beta testing, this time frame can change. Marketing this product with an established database provider could greatly decrease the time to develop a customer base for this tool. During this phase a final determination should be made on the commercial value of the product. This will depend on the interest in the product and the estimated price that could be charged based on the confidence level of the predictions.

Once beta testing is complete and there is enough interest in the prediction tool it could be released into the market. The revenue for the product would be through an annual license fee charged to users in association with the database manufacturer. The license fee is to receive updates to the tool, when new data is amassed. This is one of the important selling points of the tool. The accuracy and usefulness increases greatly over time as more data is included in the system.

Phase	Time Estimate	Objectives
Preliminary Work	4-6 years	Develop prediction methods
Development	2 years	Develop Software
Initial Release/Beta Testing	2-3 years	Beta Test / Increase interest in tool
Commercial Release		Support and Update software

Table 9: Business Plan for structure type prediction software tool

4.5 Funding:

The development of this tool would have to be done with research funding. It is unreasonable to expect outside investment until a proven tool exists, especially with the small market size. There is a good possibility of receiving research funding for this project. Regardless of the market value of the product such a tool would be useful and it has been a hot topic in computational modeling. Current work is done with funds from a National Science Foundation Information Technology Research (NSF-ITR) grant and Department of Energy grant. The Pauling File project [14] is a joint project between Japan Science and Technology Corporation (JST) and Material Phases Data System (MPDS). This is being funded by the National Institute for Materials Science (NIMS). The goal of the project is to compile a comprehensive materials database which covers all non-organic solid state materials.

Once a tool has been developed and tested, outside investment could be sought. I would consider this a new product in an existing market, so it is in the comfort zone for investors. The potential returns are not large, but only minimal investment would be required. There are no materials needed, just computation time and man hours. During the development and beta testing period the work could be done by graduate students so the man hours are cheap. When the tool reaches the market, two employees would be required, one person for support and one for development. This development would include adding new data to the database as it is available.

The ability of this tool to improve with use would cause interest among investors. The value of this tool will increase over time for multiple reasons. As more data is included the prediction ability of the tool would improve. As computational methods increase in popularity, the market for this tool will also increase.

5.0 Conclusions:

Past work has shown that there are correlations among the computed structural energies over different alloy systems. Curtarolo showed a factor of three improvement in calculations required using data mining on quantum calculations compared to random structure selection. Current work and the use of Pettifor maps has shown that there are correlations among experimentally observed stable structure types. Certain pairs of structure types have a tendency to appear together in the same alloy system, while other pairs of structure types show a strong tendency not to appear together. These correlations have also been shown to improve the ability to determine ground state structure types of an alloy system. A factor of two improvement was shown using structure type pair correlations compared to choosing structure types based on their frequency of occurrence.

A tool to predict stable structure types for an alloy of interest or to predict alloys likely to have a desired structure would be of use to many people. This could greatly improve efficiency of experimental and computational work in materials. It will be useful in many different academic departments as well as to commercial companies interested in computational modeling.

There are several tasks remaining before such a tool could be developed. The correlations among computed energies and the correlations among experimental data must be combined into a robust prediction tool. These correlations must be extended to multi-component alloy systems. Any method must also be able to predict whether or not an alloy system will be compound forming and at what compositions no structure type is expected.

The value of this tool in the market is yet to be determined. This will depend largely on the confidence levels that can be achieved by such a tool. The growing interest in computational modeling in industry will also play a large role in this decision. It is expected that the number of companies involved in computational modeling will greatly increase in the next 10 years during which this tool will be developed. This increase in modeling in industry is important because the price that can be charged in industry is much larger than the academic prices. I estimate the potential market value on the order of $150,000 annually.

The estimated time to market is 8 to 11 years. The business plan involves four phases. Preliminary work must be done before methods are ready for such a prediction tool to be developed. The tool must be developed and internally tested. Next the tool should be released in a limited fashion to academia. This is important for beta testing and to determine interest. If this third phase shows enough interest in the product it could be released in the market.

Funding required for production of this tool is limited. Research funding could allow for development and testing of the tool. If and when the tool is ready for a commercial release it could be supported by two people, a software developer and a scientist, possibly part time. An updated market assessment and cost analysis should be done during the beta testing phase to determine if this is a potentially profitable application. At this time outside investors could be contacted. The low investment required and the improvement expected over time would make this product interesting to investors.

6.0 References:

[1] D Morgan, J Rodgers, G Ceder 2003 Journal of Physics: Condensed Matter **15** 4361-4369.

[2] S. Curtarolo, D. Morgan, K, Perrson, J. Rodgers, G. Ceder 2003 Physical Review Letters **91** 135503.

[3] F.R. De Boer, *Cohesion in Metals* (North-Holland, Amsterdam, 1988).

[4] Data Mining in Materials Development, D. Morgan, G. Ceder, submitted to Handbook of Materials Modeling, Vol. 1. Methods and Models. 1.

[5] P. Villars, *Intermetallic Compounds, Principle and Practice*, edited by J. H. Westbrook and R.L. Fleisher (John Wiley & Sons, New York, 1994).

[6] G Ceder, Y M Chiang, D R Sadoway, M K Aydinol, Y I Jang, and B Huang 1998 Nature **392** 694-696.

[7] R H Victoria, C F Brucker, T K Hatwar, J E Hurst, B Uryson, D Karns 1997 Journal of Applied Physics **81** 3833-3837.

[8] http://burgaz.mit.edu:8099/jsp-examples/jsp2/miner/alloys.jsp

[9] Data Mining Approach to Ab Initio Prediction of Crystal Structure, D. Morgan, G. Ceder, S. Curtarolo, Mat. Res. Soc. Symp. Proc. 804, JJ9.25.1 (2004).

[10] J. E. Jackson, *A User's Guide to Principal Components* (John Wiley & Sons, New York, 1991).

[11] S. Wold, A. Ruhe, W. H., et al., SIAM Journal on Scientific Computing **5**, 735 (1984)

[12] R. Kramer, *Chemometric Techniques for Quantitative Analysis* (Dekker, New York, 1998).

[13] S. d. Jong, *Chemometrics and Intelligent Laboratory Systems* **18**, 251 (1993).

[14] http://crystdb.nims.go.jp/top.html

[15] D Hankerson, G Harris, P Johnson, *Introduction to Information Theory and Data Compression* (CRC Press, New York)

[16] http://www.uspto.gov/

[17] L Choi, *U. S. Patent* 6745184 (2004)

[18] P Vishnubhotla, *U. S. Patent* 6718338 (2004)

[19] J Han, *U. S. Patent* 6665669 (2003)

[20]http://www.aipla.org/Content/ContentGroups/Legislative_Action/108th_Congress1/House/hr3261.pdf

[21] Telephone Interview with Dr. John Rodgers, Toth Information Systems

[22] G. A. Landrum, Prediction of Structure Types for Binary Compounds, (Rational Discovery, Inc., Palo Alto, 2001)

[23] http://www.gaussian.com

[24] http://cms.mpi.univie.ac.at/vasp/

[25] Telephone interview with Gaussian, Inc.

[26] E-mail from Maria Temple of Gaussian, Inc.

[27] E-mail from George Kresse of VASP

[28] http://www.scm.com/Sales/pricesDollar.html

[29] E-mail from Jürgen Hafner, VASP

[30] E-mail from Ian Clements, Accelrys

[31] Westmoreland, Phillip et al. 2002. Applying Molecular and Materials Modeling, (Kulmer Academic Press, Norwell)
[32] E-mail from Mariette Hellenbrandt, ICSD
[33] S McConnell, Rapid Development: Taming Wild Software Schedules, (Microsoft Press, Washington)
[34] C Jones, 1995 IEEE Computer **28** 2 73-75
[35] E-mail from Yamazaki Masayoshi, Materials Information Technology Station (MITS)
[36] http://www.fiz-informationsdienste.de/en/DB/icsd/preise.html
[37] http://www.asminternational.org/Template.cfm?Section=BrowsebyTopic&template=Ecommerce/ProductDisplay.cfm&ProductID=12104